Design and In Pandemic Validation of Correlation Visualisation for Sleep Data Analytics

by

Amal Vincent

B.Tech, College of Engineering, Trivandrum, 2014

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Interactive Arts and Technology
Faculty of Communication Art and Technology

SIMON FRASER UNIVERSITY
Summer 2021

Declaration of Committee

Name:	**Amal Vincent**
Degree:	**Master of Science**
Thesis title:	**Design and In Pandemic Validation of Correlation Visualisation for Sleep Data Analytics**
Committee:	**Chair:** Diane Gromala Professor, Interactive Arts and Technology
	Chris Shaw Supervisor Professor, Interactive Arts and Technology
	Marek Hatala Committee Memeber Professor, Interactive Arts and Technology
	Steve DiPaola Examiner Professor, Interactive Arts and Technology

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

Sleep plays an important role in the overall health and well-being of a child. The relationship between sleep and daytime behaviours of children with sleep disorders is understood poorly; different aspects of a child's routine may interact with each other to contribute to sleep disorders. To diagnose, monitor and successfully treat many medical conditions pertaining to sleep, it becomes imperative to analyse the many aspects of a child's daytime and sleep behaviours. We built a visual analytic tool for studying the correlation between different variables pertaining to the daily life of the child. The tool allows clinicians to explore how the different aspects of a child's behaviour and activities affect their sleep and overall well-being. This tool is developed as an extension of an existing tool SWAPP, which allows caregivers and clinicians to log and monitor the child's everyday data. Later, we performed a remote usability study on the tool to demonstrate the efficacy of the tool. Finally, we generated actionable guidelines for improving the tool from the results of the study.

Keywords: data analytics; health informatics; visual analytics

Dedication

For my mother, who always believed in me. For my father because he survived for me. For my grandparents because they made me smile.

Acknowledgements

There are far too many people I am thankful to for the memorable four years at SFU. I will try to capture many of them here. But if I miss your name, please understand that I would always be grateful for everything you did for me.

Firstly I would like to thank my Thesis Committee, Chris and Marek. It took a lot of effort behind the scenes from them to complete the thesis work. Secondly, I would like to thank my colleague and mentor Ankit Gupta, for his supervision and critique in the early stages of the work. Next, I would like to thank my colleagues, Ruoyu Li and Nazanin Kadaivar, for all their encouragement and support over the years. I would also like to acknowledge the support from Diane and the Pain Lab team over the years. Diane always treated me as one of her students and, I always enjoyed working with her. As for the Pain Lab team, I am particularly thankful to Bhairavi and Pegah, who were always supportive of my work. I hope I get to work together with all these amazing people again soon.

I would like to thank Tiffany Taylor for all her help and support over the years. Tiffany has always been there to help me through many difficult situations since my first day in SIAT. Her positive influence on the graduate program and the students is seriously underrated. I am pretty sure, without her, I probably wouldn't have completed the program. Once again, I would like to thank you for everything you did for me. I would also like to thank the rest of the SIAT office team - Desiree, Lisa, Faria, Naomi, Kim and Marion. Without them, my graduate life would be very different.

Any acknowledgements I make would be incomplete without mentioning Helmine Serban. I started working as TA for Helmine back in the Spring of 2018. Since then, I have worked with her for most of my time in SIAT. I will miss working with her. She is an amazing mentor, leader and a lot of my teaching philosophy comes from her. She turned me from a timid graduate student who did not like teaching to a confident instructor who enjoys every bit of the time spent with students.

I made far too many friends during my time at SFU. Mentioning them all would be impossible. But I will mention a few names that come to my mind. I would like to thank Abbin and Noel for being wonderful roommates over the years. Next, I would like to thank my SIAT buddies - Nafiz, Shuvro, Dash, and Amy, for surviving the pointless graduate coursework at SIAT with me. I also thank Ivan for helping me find participants for my study. Last but not least, I would like to thank Jade, Sujit, Mirette, Ed, Lea, Lillian,

Katie, Mona, Arash and the rest of the ‘Safe Space’ friends for their never-ending love and emotional support in everything I do. They were very supportive through a lot of hard times.

Finally, I would like to thank the TSSU family for making the university a better place for workers and students. I got to work with some wonderful people during my time at TSSU, something I will remember for a long time.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Sleep is an essential part of a child's health. Research shows that children in the age group 6-12 years should sleep 9-12 hours a day to promote optimal health [48]. Sleep disorders among children impede their proper "physical, cognitive, emotional and social development" [67]. Also, sleep disorders and sleep deprivation can lead to adverse health outcomes such as "obesity, diabetes and impaired glucose tolerance, cardiovascular disease and hypertension, anxiety symptoms, depressed mood and alcohol use" [15] later in life.

Several studies [24, 21, 42, 58] indicate that sleep disorders may be under-diagnosed in paediatrics. This may have a negative impact on the day to day life of those undiagnosed children, which is corroborated by previous medical research [29, 32, 51, 42]. Research shows that the majority of parents are unable to produce a faithful account of the insomnia episodes and sleep history to paediatricians [22, 58, 61], which would aid the diagnosis and treatment. Chervin et al. discovered that less than 15% of the children with parent-reported sleep disorders had no recorded sleep history data [25]. Wise et al. [67] substantiate this by stating that, "because parents are generally asleep during the night" they are unable to come up with a more polished record of the sleep data. Note that such studies involved paper-based logging methods.

Kluwer [43] indicates that the process of interpreting the sleep history itself is highly intellectually and analytically taxing for the medical practitioners, as the datasets involved are considerably large (typically spanning for over 6 months) and their dependencies on the medical condition of the person may well be intricate.

To summarise, there is a considerable portion of the child population who are suffering from sleep-related disorders, which may lead to numerous health hazards. Sleep history plays a significant role in the diagnosis and treatment of sleep disorders. Parents face significant difficulty in logging such sleep history data. Finally, the doctors interpreting the results, have significant difficulty doing so. Our work SWAPP (Sleep Wake-Behaviour Application) is aimed at overcoming these difficulties. SWAPP is a web-based health informatics application that allows logging of sleep-related data of a child by the child's parent/caretaker, through

the caretaker account on the application. This data is accessible to the clinicians via a clinician account linked to the child's data.

Research has shown that sleep-related variables (many studies refer to them as sleep factors, e.g., sleep duration, sleep mood, etc.) can be closely correlated to behaviour problems in school-aged children [62] or preschoolers [39]. Numerous studies have outlined the importance of analysing the various drug interactions with sleep-related behaviours/disorders [27, 44, 59]. Shrivastava et al. point out that analysis of sleep diaries/sleep logs is required for "referring physician to review the sleep study report and correlate patient's presenting sleep complaints to the results" [57] and also corroborates that multiple factors of the sleep logs are involved in the analysis of the disorders, similar results are presented by other researchers [49, 20]. In order to cater to the data analysis needs of the clinicians, we developed a tool to explore the relationship between variables such as sleep duration, sleep quality, behaviours, medication, moods etc. We use correlation visualisations to support exploration of the relationship between variables, which is also the prime focus of this thesis. The work important for clinicians and who operate under high level of work related stress and time constraints [66]. Our work is centred around a prototype built for the purpose. In this work, we analyse the ability of the tool to mitigate difficulties in correlation analysis and the overall usability of the tool.

1.1 Research Questions

Our extensive literature review and project discussions with two clinicians revealed that often the correlation or dependency of various sleep-related variables/factors need to be analysed with each other [65]. As the data spans over multiple days the task of analysis was cumbersome for the clinicians. The Correlation Visualisation Tool was designed as a module in the SWAPP to ease the analysis.

This work revolves around the design, development and evaluation of a data Visualisation Tool to overcome the above difficulties. The following are my research questions:

1. **RQ1: How to best visualise sleep data for analysing correlations, given the following requirements?**

   Requirements for the visualisation:

   - Minimise complexity of the data interpretation
   - Quick summary of Correlations
   - Ability to visualise individual data points
   - Total number of variables assumed to be 15-50

   The work makes use of existing methodologies in the development of Correlation Visualisation Toolkit in order to answer the research question. The novelty of the

work comes from the unique problem the tool is designed for. Details on the design and development of the data visualisation and the interaction can be found in Chapters 3 and 4.

2. **RQ2: How to run a user study on the Correlation Visualisation Toolkit during a global pandemic?**

   Midway through the process of Ethics approval for the planned user-study, the world was struck by COVID 19, a global pandemic. SFU went into full lockdown in March, 2020. As a result, we had to adapt the user study to be delivered online over Zoom video conferencing. In doing so, we developed new knowledge on performing data Visualisation studies online with ethical guidelines from SFU. The details about the adaptation of the user-study can be found in Chapter 5.

3. **RQ3: Does the Correlation Visualisation Toolkit perform better than Microsoft Excel in terms of overall workload and time taken for task completion?**

   Or from a Quantitative perspective:

   (a) *Hypothesis 1 (H1): Mean Overall Workload for Correlation Visualisation Toolkit is less than that for Microsoft Excel.*

   (b) *Hypothesis 2 (H2):Mean time taken for completing correlation analysis tasks on Correlation Visualisation Toolkit is less than that for Microsoft Excel.*

   In order to validate the performance of the Correlation Visualisation Toolkit, we performed a comparative analysis of the tool with Microsoft Excel. This was done through a controlled experiment, where the participants were asked to complete data analysis tasks using both Correlation Visualisation Toolkit and Microsoft Excel. On completion of the tasks we evaluated the overall workload arising from the tasks using NASA-TLX questionnaire. More details about the results of the study can be found in Chapter 6.

4. **RQ4: How can we improve the Correlation Visualisation Toolkit?**

   While, the Correlation Visualisation Toolkit mitigates some of the problems caused by state of the art technology, it is in no way perfect. Hence, we solicited feedback from the participants in our user-study to improve the tool. Recommendations to improve the Toolkit can be found in Chapter 6.

1.2 Previously Published Works

A significant portion of the dissertation has been featured in two of our previous publications:

1. Amal Vincent, Ankit Gupta, Chris Shaw, and Ruoyu Li. Correlation Visualisation for Sleep Data Analytics in SWAPP (Sleep Wake Application). Electronic Imaging, Visualization and Data Analysis, 2019, page 682-1-682–10. Society for Imaging Science and Technology.

2. Amal Vincent, Ankit Gupta, Ruoyu Li, Chris Shaw, and Saba Akhyani. Data acquisition and visual analytic tool-set for paediatric sleep data. In Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth'19, page 320–326, New York, NY, USA, 2019. Association for Computing Machinery.

1.3 Research Contributions

My thesis revolves around the design, development and validation of the Correlation Visualisation Toolkit. My contributions include:

- Analysis of extant data visualisation techniques and their efficacy in addressing the problem at hand; ultimately establishing the design of the Correlation Visualisation Toolkit.

- Validation of the Correlation Visualisation Toolkit, which involved the development of a novel user-study technique for remote in-pandemic validation of the tool.

- Framing recommendations to make the Correlation Visualisation Toolkit more user friendly, from the above user-study.

Chapter 2

Sleep Wake-Behaviour Application(SWAPP)

While the focus of the the dissertation is on a specific module in SWAPP, it is important to understand SWAPP before studying the Correlation Visualisation Toolkit.

SWAPP is a web-based health informatics application developed by our team [64], that allows logging of sleep-related data of a child by the child's parent/caretaker, through the caretaker account on the application. This data is accessible to the clinicians via the clinician account linked to the child's data. SWAPP is also equipped with a clinician dashboard which provides the summary and other insights on the entered data and previous data assessments made. The dashboard exploits various visual analytic primitives to aid the doctors to better interpret the data.

Our design of SWAPP began with analysing the conventional technique of sleep diaries and observing how the caregivers and clinicians interacted with them. This provided insights to overcoming some of the limitations of the conventional paper based sleep diary.

2.1 Web Interaction

A caregiver using the SWAPP application can record periods of sleep (preparation before sleeping, sleeping, and awakenings during the night), daytime activities, and medication. For the sleep, the caregiver/parent records the start and end time of the activity (like sleep, swimming, studying, etc.), the mood of the child during or after the activity, any disruptive behaviours during the activity. The process flow for creating the activity log is depicted in Figures 2.1, 2.2 and 2.3. For medication, the caregiver records the time of medication, dosage, and any side-effects.

When a log is created, it gets added to the timeline visualisation (Figure 2.4) which shows the logs on a vertical timeline. The caregiver can update or delete the logs by selecting them from the timeline visualisation.

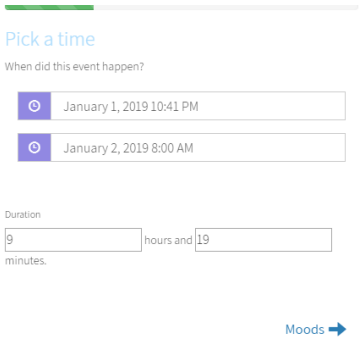


Figure 2.1: The user enters the start time and end time.

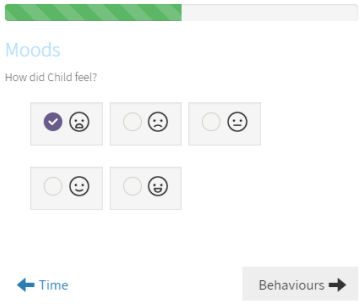


Figure 2.2: The user enters the child's mood.

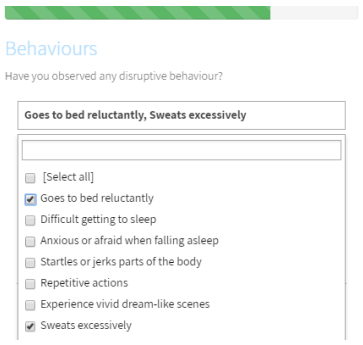


Figure 2.3: The user records any disruptive behaviours.

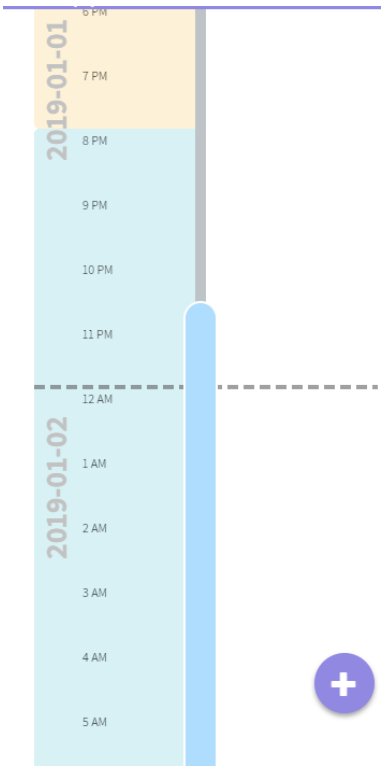


Figure 2.4: The sleep log gets added and visualised on the timeline

The application also consists of a dashboard which visualises the different metrics such as average sleep duration, quality, moods, and doctor's assessments.

2.2 The visual analytic toolset

One of the prominent features of the SWAPP is the visual analytic tool-set that enables the users (clinicians and caregivers) to summarise and analyse the data. The toolset is comprised of the timeline visualisation, dashboard and correlation analysis tool.

2.2.1 The dashboard

Both the clinicians and caregivers have access to a dashboard that allows selective filtering of a child's data for a range of dates. For the caregivers, this data acts as self-management, while on the clinician side the information presented provides insights to diagnosis and treatment of the associated medical conditions. The dashboard for clinicians is shown on figure 2.5.

2.2.2 Correlation analysis tool

As discussed above, the Correlation Visualisation Toolkit (see Figure 2.6) helps clinicians to understand and explore the correlation or dependency of various sleep-related variables/-

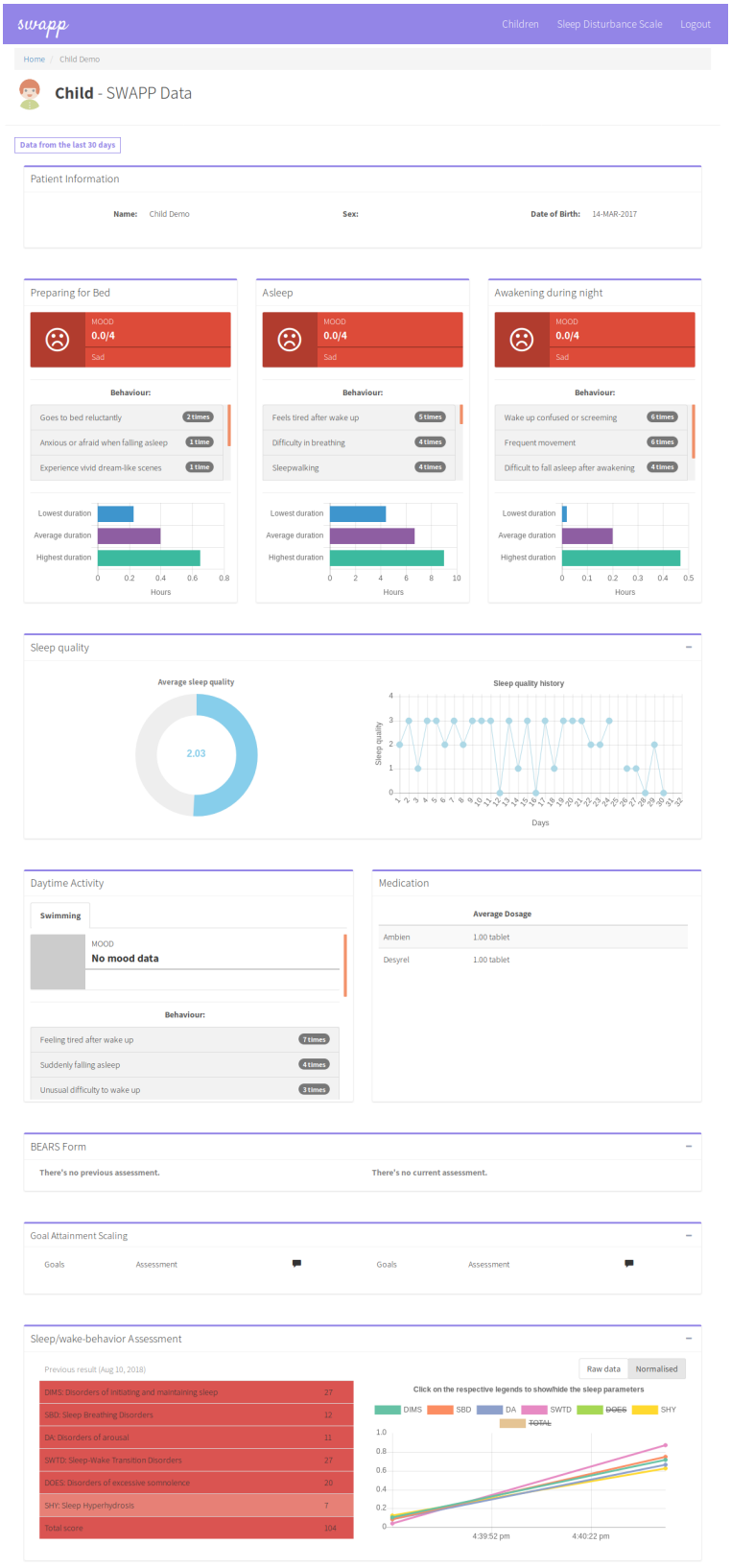


Figure 2.5: Dashboard of SWAPP

factors (pertaining to the data collected using SWAPP) with each other [65]. The Tool is the prime focus of my thesis, and will be further discussed in the subsequent chapters.

Figure 2.6: Correlation Visualisation Toolkit

Chapter 3

Related Work

The review of literature began with exploring possible visualisations for representing correlation. Our application aims to minimise the complexity of the visualisation and allow effortless interpretations. Past research has shown that 3d plots are significantly more difficult to interpret than their 2d counterparts for large datasets [50], so we will be limiting our discussion to two-dimensional visualisations.

When it comes to visualising correlations there are two basic approaches. The first technique involves visualising the individual data sample points (e.g., scatterplots, parallel coordinates etc.). These visualisations can then be used to determine correlations between variables and functional relationships. The alternative to this is to make use of parameters like Pearson's / Spearman's coefficient for correlation, which are a measure of correlation. These parameters for multiple variables can then be conveniently visualised to infer correlations (correlation matrix heatmap, Venn diagrams, solar plots etc.). Various visualisations falling into each of the two categories is discussed in this chapter.

3.1 Design Requirements

The proposed interaction aims at allowing users to exercise selective control over a subset of variables to be analysed for correlation from a larger pool /set of variables. In SWAPP, we log over 20 variables currently, which makes it imperative that the chosen visualisation method is optimal for analysing correlations for numerous variables. The clinicians are assumed to have minimal knowledge of the statistical primitives involved in correlation/relationship analysis between the variables. If a user with significant exposure to the statistics and sufficient knowledge of the sleep data itself wants to apply the expertise into the analysis, provision should exist to provide the user with an option to visualise individual samples in the data.

3.2 Review of Visualisations for Correlation Analysis

3.2.1 Scatterplots

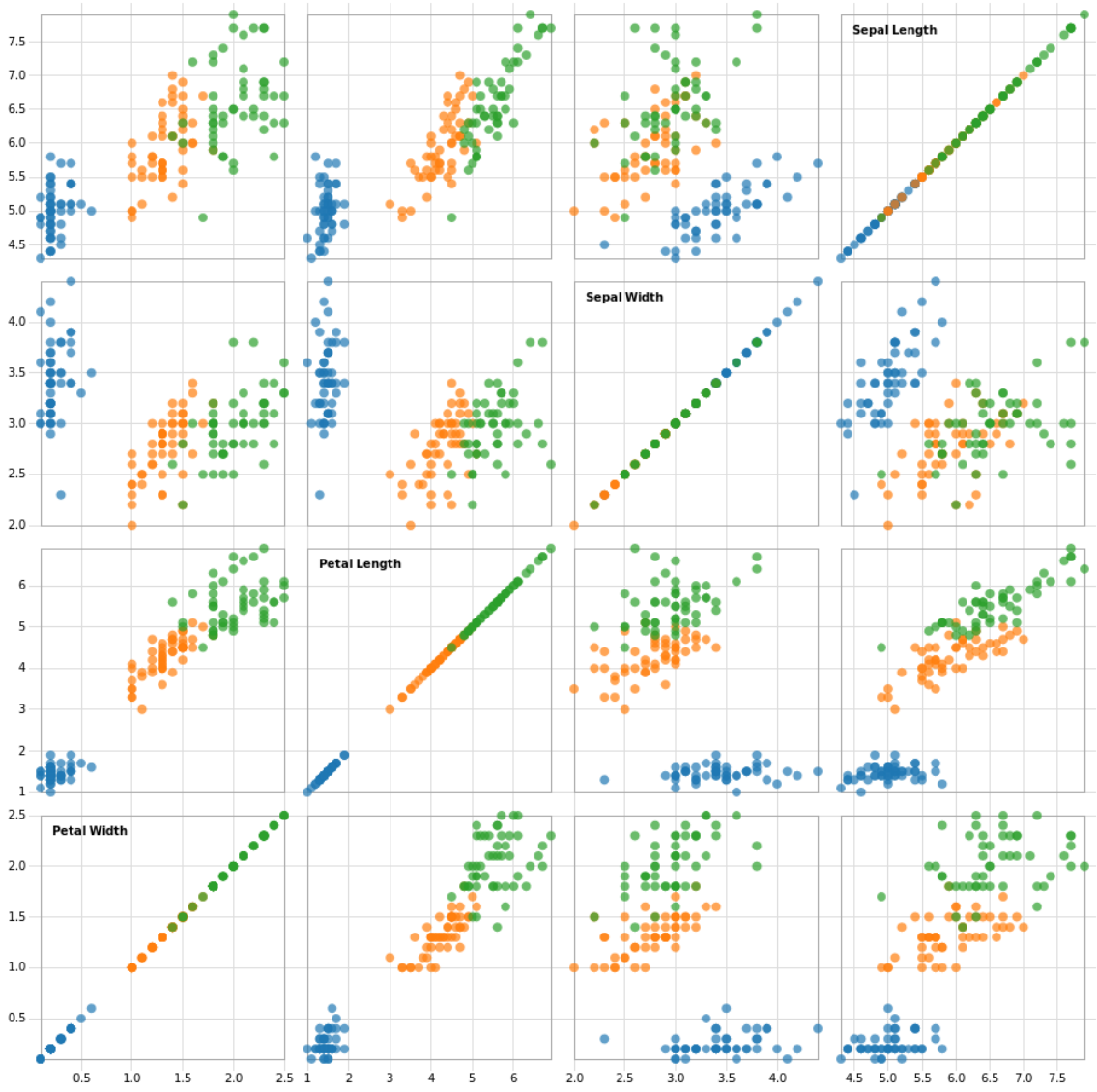


Figure 3.1: A scatter plot matrix example of petals and sepals of a plant [8]

The research literature documents numerous visualisations for depicting correlations, the most obvious of these being the scatter plot [45]. Scatter plots are useful tools in interpreting the "direction, form and strength" of the relationship between two variables [45]. This may be extended to analysis for multivariate relationships by using scatter plot matrices [28].

Scatter plots use points to represent values for two different variables. The position of each point on the x-axis and y-axis indicate values for an individual data sample. The

scatter plot shows the relationship between two variables, while the scatter plot matrix shows all pairwise scatter plots for many variables (see Figure 3.1). If the variables are observed to increase and decrease together, the correlation is positive (e.g., petal width vs petal length in Figure 3.1). If one variable tends to increase as the other decreases, the association correlation is negative. If there is no pattern, the association is zero.

Scatter plots offer the advantage that they can easily be generated from raw data without any analytical computation, they provide insight on each and every sample point in the dataset. Interpreting the scatter plot may not be straightforward and may require very stringent analysis of the plots, especially with increasing number of dimensions/variables or for non-linear functions [60]. And scatter plots as such are not suitable for discrete variables. While, some of the characteristics of the scatter plot were desirable for the project it did not meet all the requirements. Nonetheless, scatter plot was part of the tool we designed, as we required a visualisation that allowed display of individual data points and functional relationship between variables.

3.2.2 Parallel Coordinates

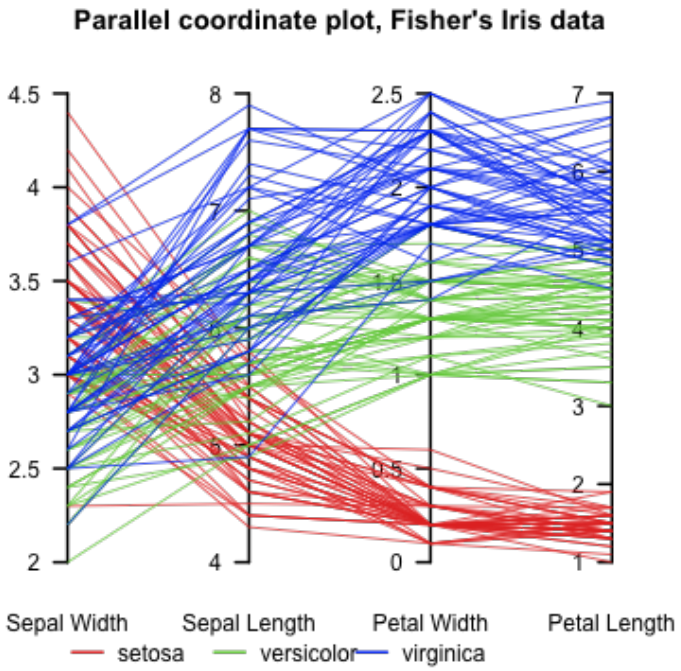


Figure 3.2: A parallel coordinate plot example of petals and sepals of a plant [6]

Some previous literature [35] also refer to parallel coordinate graphs for interpreting correlations. Parallel coordinates plot each variable to an axis/plane, and the axes are

organised as uniformly spaced M vertical lines. A data element in M-dimensional space is represented as a connected set of points, one on each axis (see Figure 3.2). In Figure 3.2 there is a negative correlation between sepal width and sepal length, since higher values of sepal width correspond to lower values of sepal length (which results in intersecting lines). There is also a positive correlation between petal width and petal length since the lines are approximately parallel.

Despite being a powerful visualisation they tend to take a longer duration to interpret with increasing number of variables. They only allow analysis of adjacent variables and take up a lot of space along the horizontal direction. Moreover, they cause cluttering for large datasets [56].

3.2.3 Contingency Wheel

Contingency Wheel is a visualisation mostly used in the analysis of categorical data [19]. The categories are visualised as sectors of a ring-chart and the individual data points are depicted as dots in these sectors (see Figure 3.3). Arcs are drawn between two sectors if one or more rows have dots in both sectors. The arc is thicker if there is higher correlation between the two sectors. It is a relatively more complex visual analytic solution. Contingency wheel works best when there are myriads of variables/dimensions compared to what is required in SWAPP. The visualisation is also limited to applications with categorical variables. Interpretation of the contingency wheel also tends to be more complex than the techniques discussed above and analysts who use contingency wheel needs some training to make effective use of it.

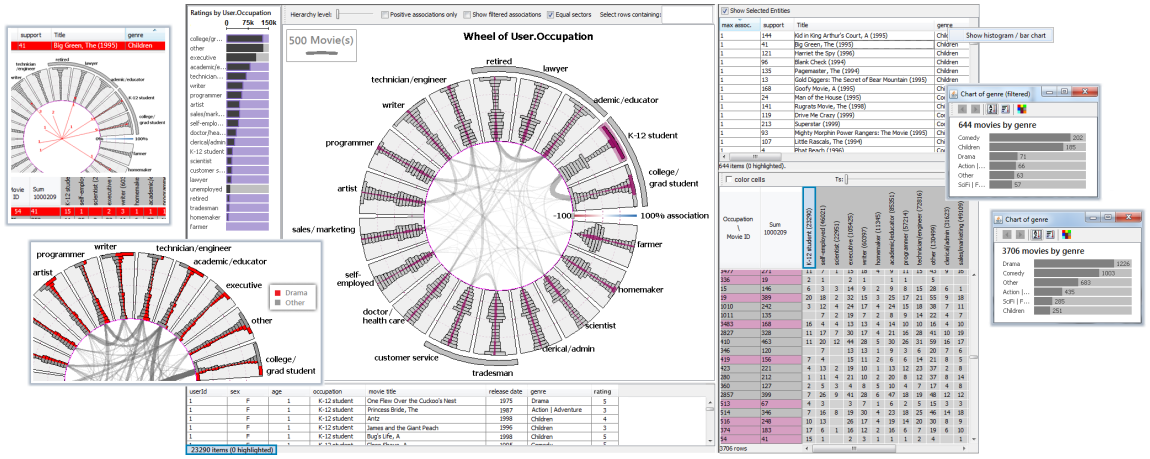


Figure 3.3: Contingency wheel [19]

3.2.4 Parallel Sets

Parallel set is a categorical version of the parallel coordinates which may be used to visualise quantifiable relationships between variables/categories [38]. The technique is based on the axis layout of parallel coordinates. The boxes representing the categories and parallelograms between the axes showing the relations between categories. The width of the parallelogram between the categories can be used to represent the level of correlation (see Figure 3.4). But as with the parallel coordinates, with an increasing number of variables, the interpretations from the visualisation may not be straightforward. Furthermore, the applications are limited to categorical variables.

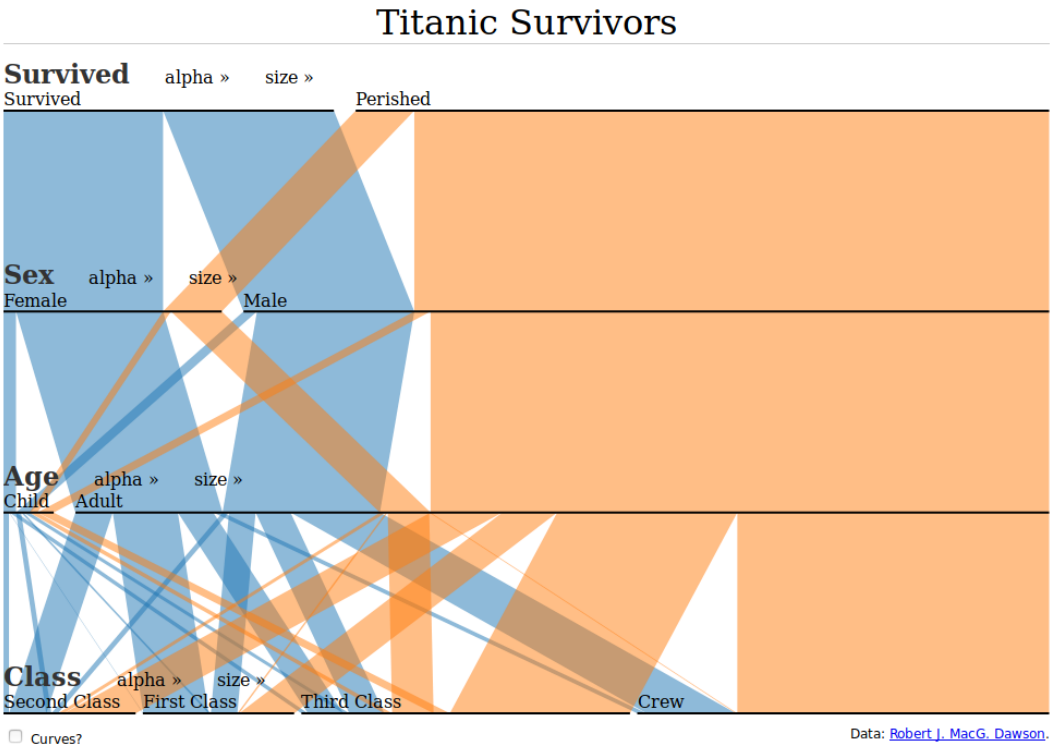


Figure 3.4: Parallel set visualisation [7]

3.2.5 Venn Diagrams

Other conventional visualisations for depicting relationships, including Venn diagrams [37]. Each variable in a Venn diagram is represented by a circle, the amount of shared variance or correlation can be represented by amount of overlap between the circles in the Venn (see Figure 3.5). However, Venn diagrams do not work too well for larger dimensions/variables and quantifying the inferences tend to be cumbersome and erroneous in such cases as circle/ellipse intersections are difficult to interpret and compare.

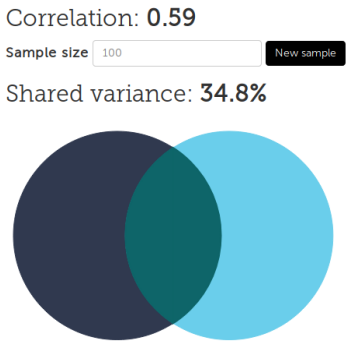


Figure 3.5: Venn Diagrams [1]

3.2.6 Connected Graphs and Dendrograms

Connected graphs [55, 54] have also been employed in visualising correlation, however as with the case of contingency wheel, such approaches work best for a very large number of variables. In connected graphs the nodes represent different variables while the edge between the nodes represent the correlation between different variables. Longer the edges, the less correlated the variables are. (e.g., see Figure 3.6)

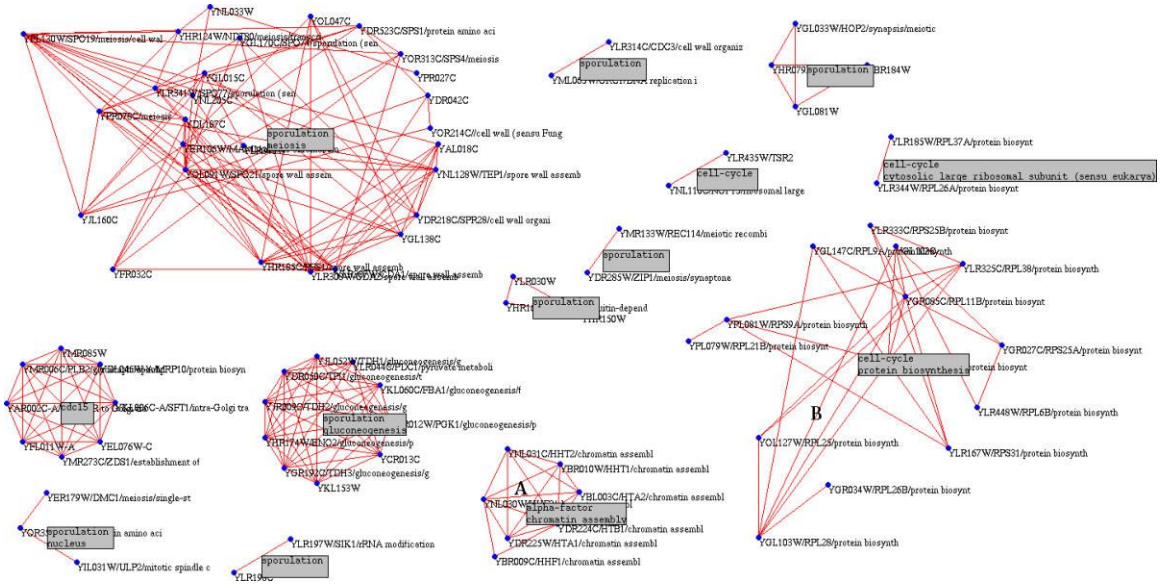


Figure 3.6: Connected Graphs [54]

Variants of this method include, approaches such as clustering/ dendrograms successfully applied in multiple disciplines [52]. Dendrograms have a tree like architecture. Each variable in the dendrogram is represented by the leaf nodes on the tree. The correlation between

any two variables is determined by the height of the most recent ancestor node of the two variables (e.g., Figure 3.7).

Such methods, though, assume a connected relationship or correlation between all the variables. Both connected graphs and dendograms do not allow visualisation of the individual data points.

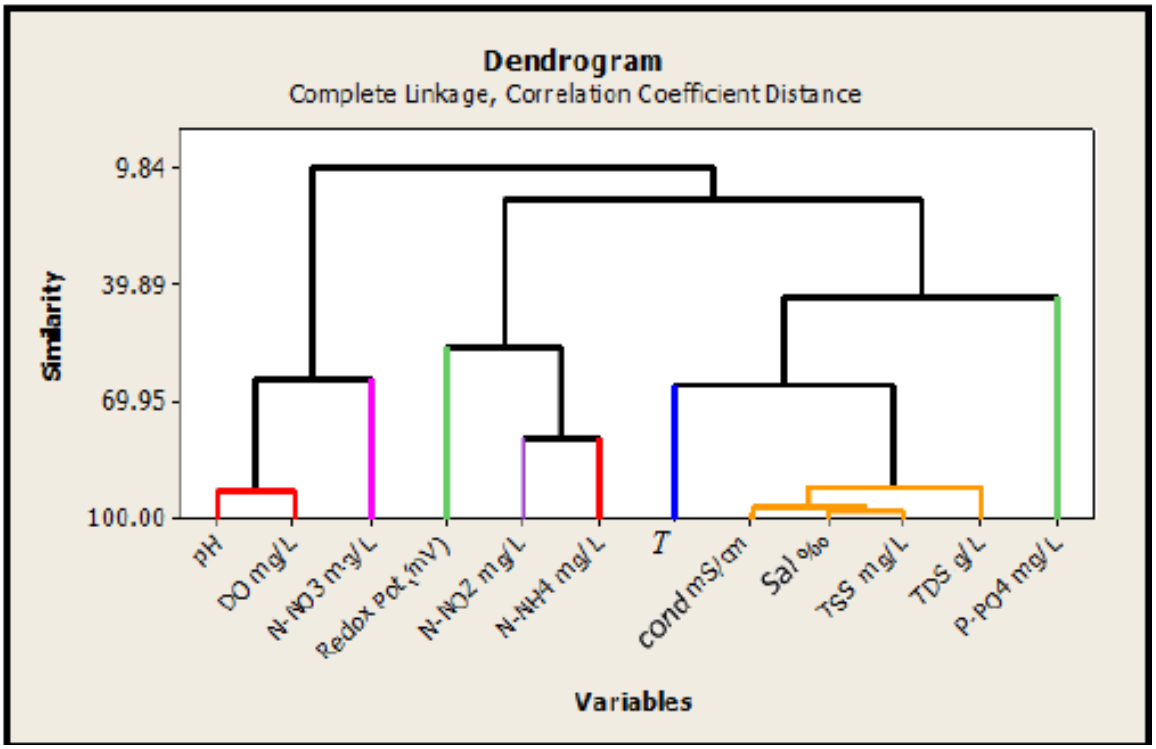


Figure 3.7: Dendrogram [13]

3.2.7 Correspondence Analysis Plots and Moon Plots

Correspondence analysis Plots [36, 63] and its modern variant - moon plots [23], have been deployed successfully in correlation analysis in multiple disciplines.

An example of correspondence plot is given in Figure 3.8. The rows (brands), the individual data samples of the dataset used is represented by blue points. While columns (personality), variables of the dataset are represented by orange points. The distance between any rows or columns in the plot, indicates their correlation. Rows with similar profile are close on the plot. The same is true for columns. As for the moon plots (see Figure 3.9), the variables are arranged in a circle, and the individual data samples are plotted inside the circle. The distance between the attributes and data points indicate the correlation.

However, analysis of literature revealed that there is some confusion introduced by the visualisation while a novice user interprets them [23]. Although moonplots attempt to rectify

the shortcomings, the method suffers from the same drawbacks. Moreover, the methods are infeasible for large number of data points.

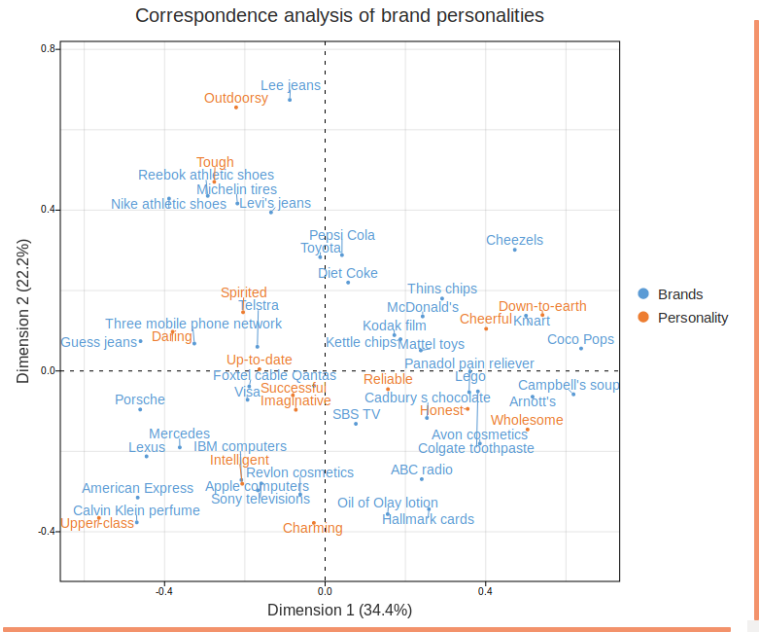


Figure 3.8: Correspondence Plot [17]

3.2.8 Solar Correlation Plot

Other methods like solar correlation plots [68] are confusing when it comes to interpreting negative correlations, and have found very little application in the research domain. In a solar correlation plot, each circle around the sun is an orbit. The variables are represented as planets, and moons are variables that are inter-correlated with the planet variable they orbit. The closer the orbit, the stronger the correlation (see example in Figure 3.10). However, solar plots tend to be confusing for large number of variables and also the lack the ability to throw light on the functional relationship between variables.

3.2.9 Heatmap of Correlation Matrix

Heatmaps of the correlation matrix have been successfully applied in correlation/relationship analysis/ display in multiple disciplines including health informatics [33] and could possibly be the most used visualisation in correlation applications. Heatmaps are visualisations of the correlation matrix between all the variables in a dataset (see Figure 3.11). The correlation between two different variables in the dataset is indicated by the colour of the corresponding matrix element. The heatmap is an option that came closest to satisfying the primary requirements of the SWAPP correlation analysis tool. Unfortunately, even the

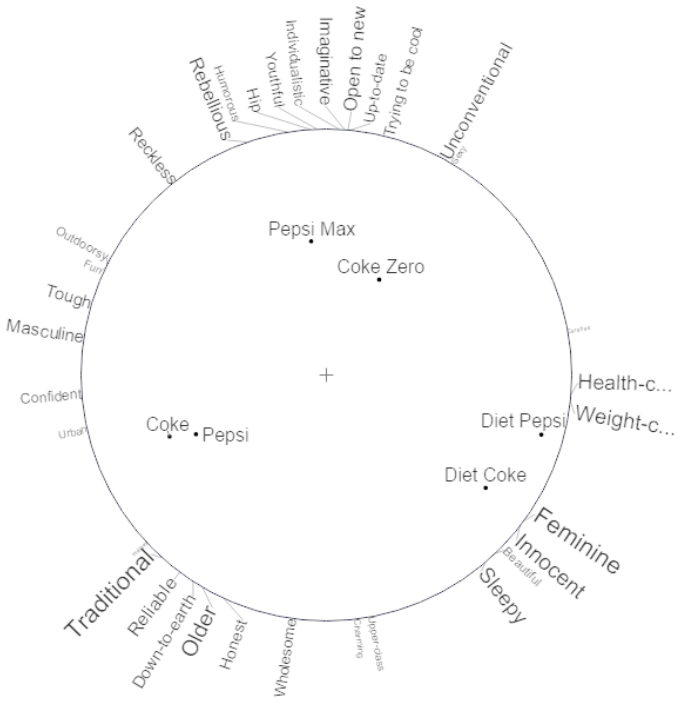


Figure 3.9: Moon Plots [18]

heatmaps on its own did not provide the option to visualise individual samples in the data. We also noted that information displayed on the heatmap can be overwhelming when the number of variables are over 10. Nonetheless heatmaps met part of the requirements, so we did incorporate a modified version of heatmap in our proposed tool.

3.3 Going forward

Summary of our findings from the review of literature can be found in table 3.1. From our analysis it was apparent that the scatterplots by far provided the best functional representation for continuous variables. This could easily be extended to discrete variables by using Jittering. Jittering is a data visualisation technique that involves adding random noise to data points, this is to prevent overplotting of data points in statistical graphs. However, the major downsides with scatterplots were the lack of quick summary and difficulty in analysing non linear functions. Heatmaps on the other hand provided a well organised summary that was easy for the user read and comprehend. However, the actual data and the functional representation is lost in heatmaps, since they are a representation of the correlation coefficient and significance. So our challenge now was to design a system that gave us the best of both worlds. This is described in the next chapter.

Visualisation	Types of variables supported	View data points	Quick Summary of Correlation	Remarks
Scatterplot	Continuous only without modifications	Yes	No	Good for viewing functional relationships
Parallel Coordinates	Continuous, Discrete/ Categorical	Yes	No	Causes cluttering with large number of data points
Contingency Wheel	Categorical	Yes	No	Application limited to categorical variables
Parallel Sets	Categorical	No	Yes	Application limited to categorical variables
Venn Diagrams	Continuous, Discrete/Categorical	No	Yes	Only suitable for correlations involving very few variables
Connected Graphs	Continuous, Discrete/Categorical	No	Yes	Suitable for very large number of variables
Dendrograms	Continuous, Discrete/Categorical	No	Yes	Assumes Connected relationship
Correspondence analysis plots	Continuous, Discrete/Categorical	Yes	Yes	Not suitable for novice user
Moonplots	Continuous, Discrete/Categorical	Yes	Yes	Plot can get cluttered for large datasets
Solar Correlation Plot	Continuous, Discrete/ Categorical	No	Yes	Can be confusing with negative correlations
Heatmaps	Continuous, Discrete/ Categorical	No	Yes	Best of the bunch for quick summary of correlations

Table 3.1: Summary of Visualisations analysed

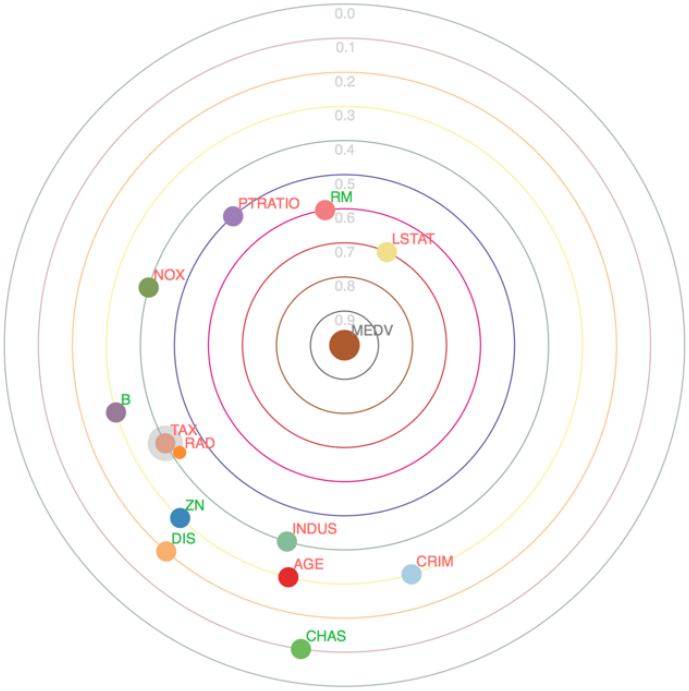


Figure 3.10: Solar Plot [68]

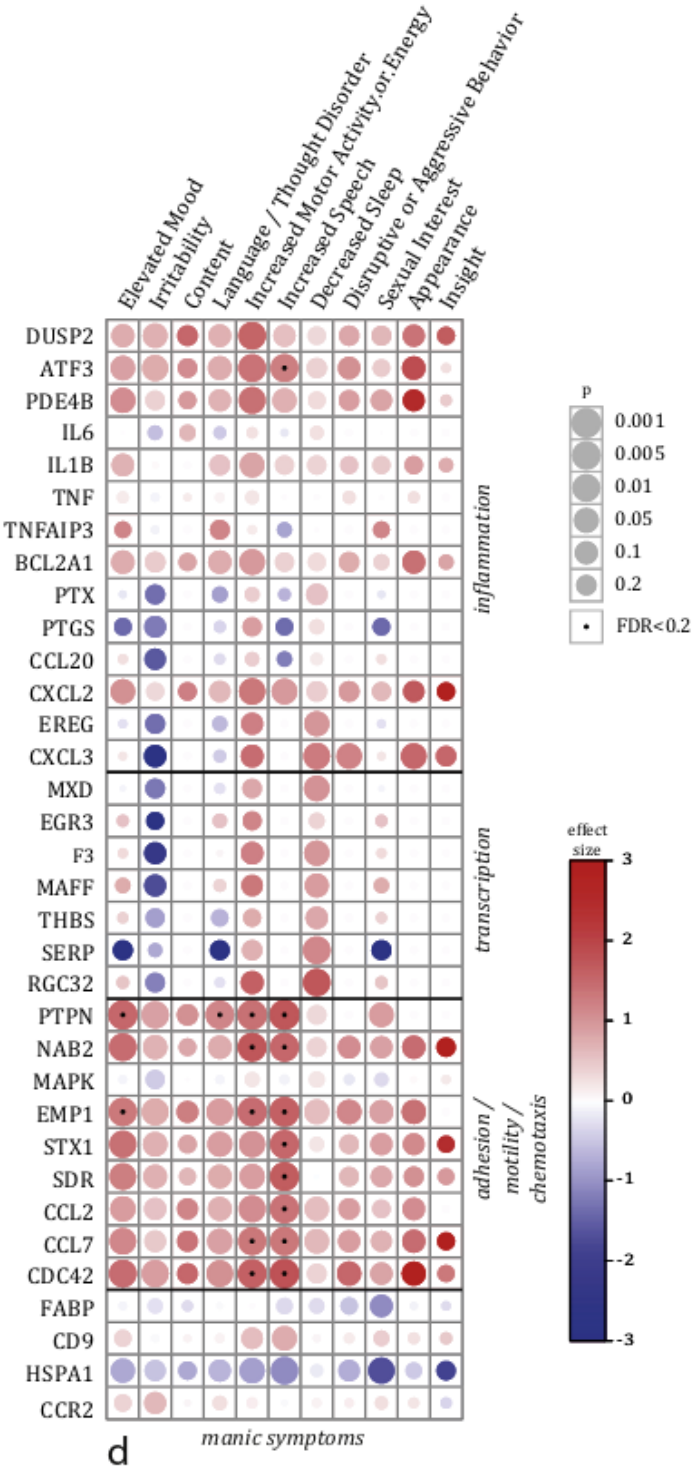


Figure 3.11: A modified heatmap, showing effect size and significance [33]

Chapter 4

Correlation Visualisation Toolkit

The idea of visualising correlations between variables of interest is not unique. Many statistical software products like Microsoft Excel [4], SPSS [9] and JMP [3] allow users to find correlations between two variables. However, the use of such software assumes knowledge about statistical primitives and the software packages. Since the tool we built is integrated into the data acquisition or sleep logging system there is also no need to export, clean up or format the data for correlation assessments. Furthermore, the procedure involved in generating the correlation reports from such statistical software packages tend to progress in a tedious and algorithmic manner, thus work-flow can be reused for a given set of data variables. Moreover, opting for a visual analytic solution gives the experienced clinicians more power, control and flexibility in the data interpretation.

4.1 The interaction

Clinicians are typically required to analyse correlations between variables pertaining to sleep and other daily activities of the child. For instance, a clinician may be interested to know how a particular drug/treatment (in some cases multiple drugs) interacts with the sleep quality of the child. Another example would be, where the clinician is required to analyse if exercising reduces disorders such as sleepwalking. We noticed that in most scenarios, the clinicians worked with a selective subset of variables. Based on the observation we decided to build an interaction that would help filter variables.

The design of the interaction was inspired by the drag and drop interaction of the data visualisation software Tableau (see Figure 4.1) [11]. The available sleep-related variables/-parameters/events for visualisation are stacked on the left-hand side of the visualisation in categorical boxes (figure 4.2 (a)). Note that the variables are in the form of the small rectangular capsules (figure 4.2 (b)) in the larger categorical boxes (figure 4.2 (a)). The categorical classifications of the sleep variables, i.e., headings of the boxes holding the sleep variables are designed to closely resemble the categorical classification and boxes that can be found on the clinician dashboard of the SWAPP web application, this ensures recogni-

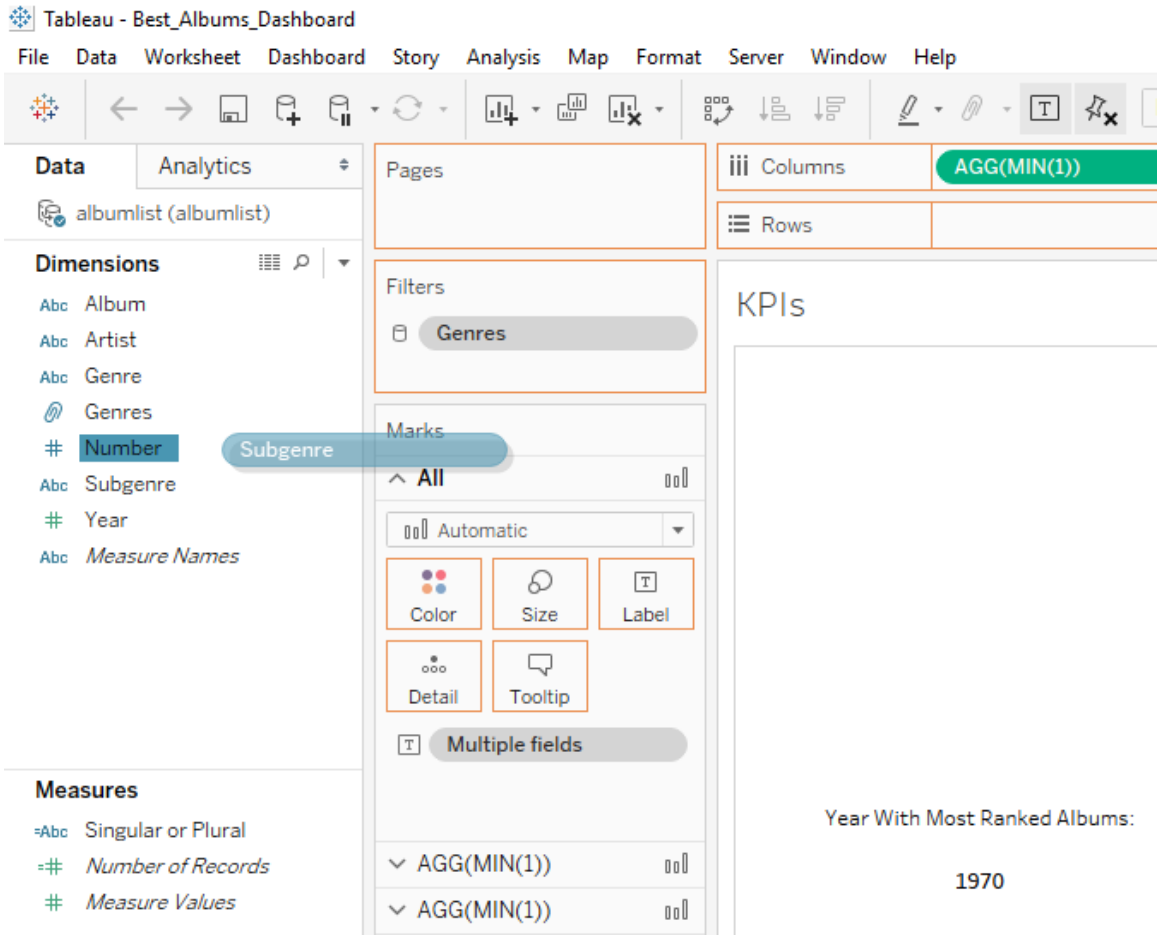


Figure 4.1: Tableau drag and drop

tion of the variable setup from the dashboard which the users are already familiar with, rather than contrived recall of the setup, which is one of the key heuristics of interaction evaluation according to Nielsen [47].

There is a 'horizontal box'(figure 4.2 (d)) and a 'vertical box' (figure 4.2 (c)), to the right of the 'variable/parameter boxes' (figure 4.2 (a)). These horizontal and vertical boxes act as input for the x-axis and y-axis of the visualisation respectively. Initially when the page loads, the area where the visualisation originally appears is depicted by a white space to the right of the 'y-axis box' (figure 4.2 (c)) and above the 'x-axis box' (figure 4.2 (d)). The user can initialise the visualisation by dragging and dropping 'individual variables'(figure 4.2 (b)) into the 'x-axis box' and 'y-axis box'. There is a direct correspondence between the order of variables along the respective axes in the visualisation and the order of variables in axes boxes, the interaction allows reordering and even moving variables between the two axes boxes according to the user's needs by further dragging and dropping. The user can remove unwanted variables from the x-axis and y-axis boxes by clicking on the 'x' button (figure 4.2 (e)) on the right of the rectangle containing the variables.

4.2 The visualisation

As discussed in the related work section, the other existing methods were inadequate on their own to deliver to the requirements of the application of sleep data analytics, where the users are typically doctors. The alternative was to combine two of the visualisations cited in the previous research section, by exploiting the best features from 'two worlds' for the convenience of the user. We also identified that the effect size is as important as the significance value/ correlation when it comes to assessing relationships between variables, in light of previous scientific literature [33]. Careful analysis of the possible combinations of the visualisations suggested that combining the heatmaps with the scatter plots satisfied the requirements of the application. The primary visualisation is a heatmap plot of the correlation matrix arising from the variables dropped in the x-axis and y-axis boxes. There should be at least one item in both x-axis box and y-axis box to initialise the visualisation. Pearson correlation coefficients are used between two continuous variables and Spearman's coefficient is used if a discrete variable is involved. As discussed above there will be columns in the visualisation corresponding to the variables dropped in the x-axis box and rows corresponding to the variables dropped in the y-axis box. The intersection of the rows and columns i.e, the individual cells in the visualisation depict the correlation between those two variables. The visualisation was designed using the Plotly, a JavaScript library. Colour is used to represent the value of the individual correlations or the effect size; a red-blue scale is used for mapping correlations (figure 4.2(f)). The colour scale was chosen based on similar research in the past [33]. The red region on the scale represents a high positive correlation while the blue signifies a high negative correlation.

Figure 4.2: Annotated interaction for the visualisation

The size/radius of the circle represents the significance value. Mathematically, where R is the radius of the circle and $p \in [0, 1]$ is the significance value. For some constants $c, k \in \mathbb{R}^+$ and $c > k$

$$R = \begin{cases} c \times (1 - p), & \text{if } p \leqslant 0.05; \\ k \times (1 - p), & \text{elsewhere;} \end{cases}$$

Thus, the radius of the circle increases as significance value decreases. Also, the rate of increase is more when the significance was less than or equal to 0.05. This ensures that insignificant relationships (p $>$ 0.05) are represented by much smaller circles.

The transparency of the circle in the heatmap was adjusted to represent the product of the magnitude of correlation and the difference of the significance value from unity. Thus, the circles would mostly be transparent or invisible unless both the significance and Correlation conditions are satisfied. So where significance is represented by $p \in [0, 1]$, correlation coefficent is represented by $r \in [-1, 1]$ and transparency is represented by $\alpha \in [0, 1]$

$$\alpha \propto (1 - p) \times |r|$$

One of the requirements of the system was to simplify the primary analysis of the sleep variables for correlation. Meticulous filtering from trials of the possible visualisation designs discussed in the related works section suggested that the heat map visualisations provide a quick overview when analysing numerous variables. Once, the user has identified the pair of variables that need to be examined, the user can gain further insights into the variable pair from the scatter plot between the two variables. This is achieved by clicking on the respective circles on heat map visualisation, which would then display the scatter plot between the two variables as a popover. This would allow the user to have a deeper view of the sample points in the data. Jittering was introduced in the scatter plot to accommodate visualisation of discrete variables using a scatter plot (see Figure 4.2). Jittering is a data visualisation technique that involves adding random noise to data points, this is to prevent overplotting of data points in statistical graphs.

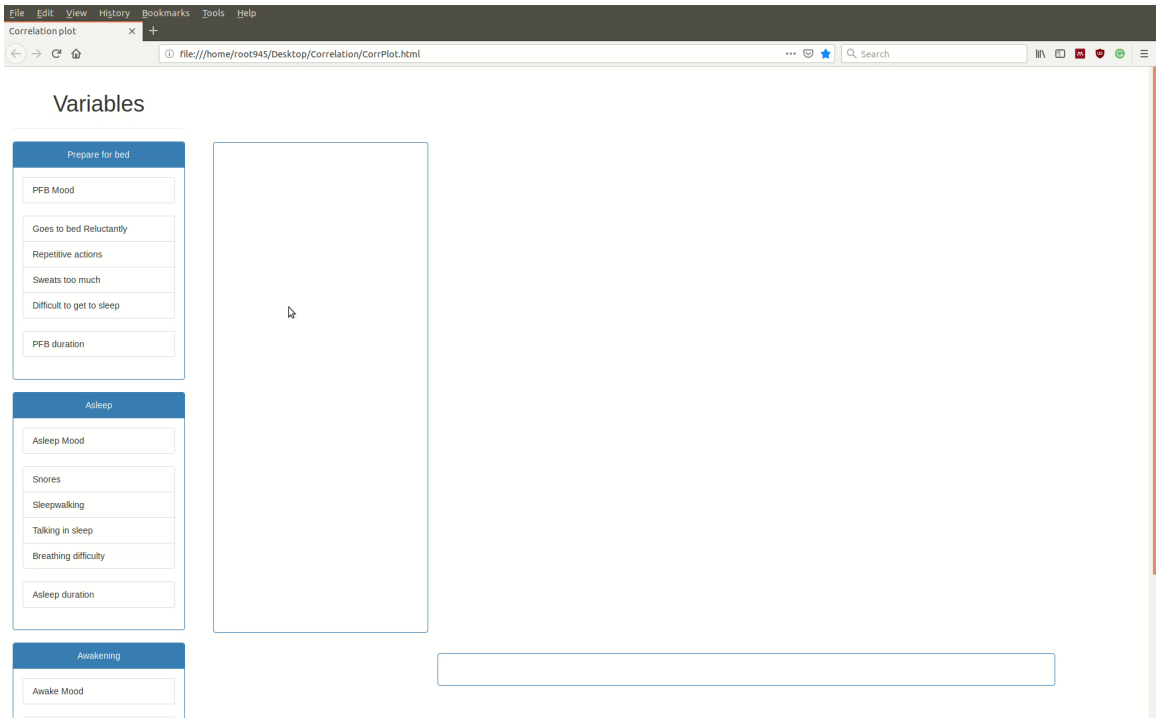


Figure 4.3: Initial view when the correlation visualisation tool loads in the browser

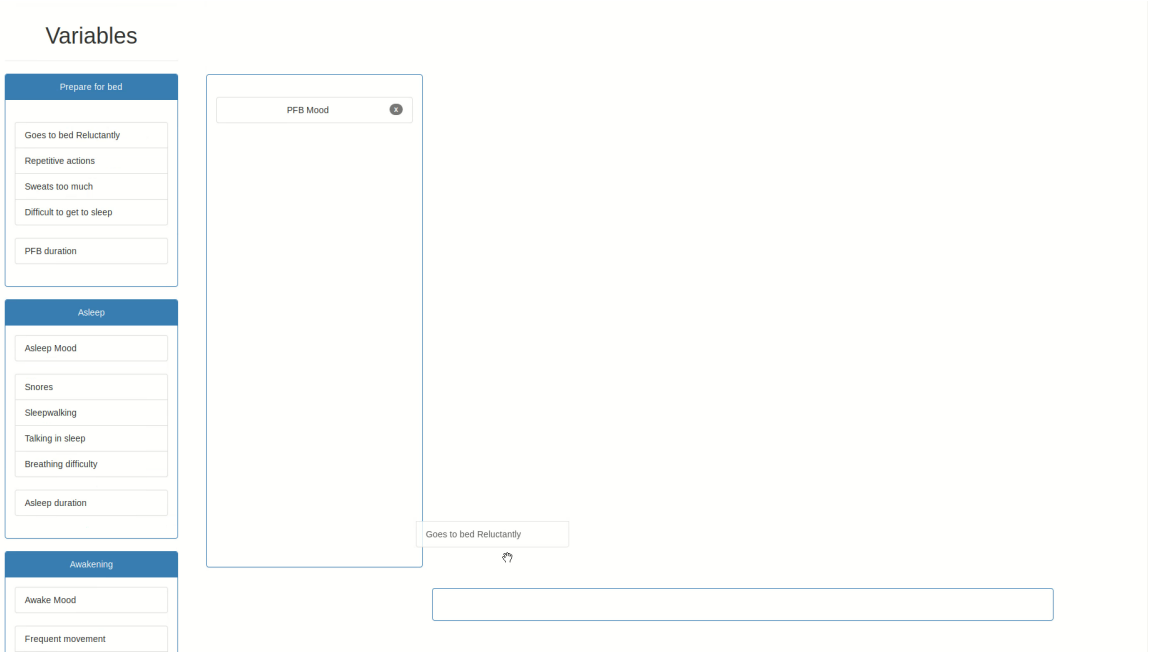


Figure 4.4: Drag and drop items into the X-axis and Y-axis boxes to initialise the heatmap visualisation

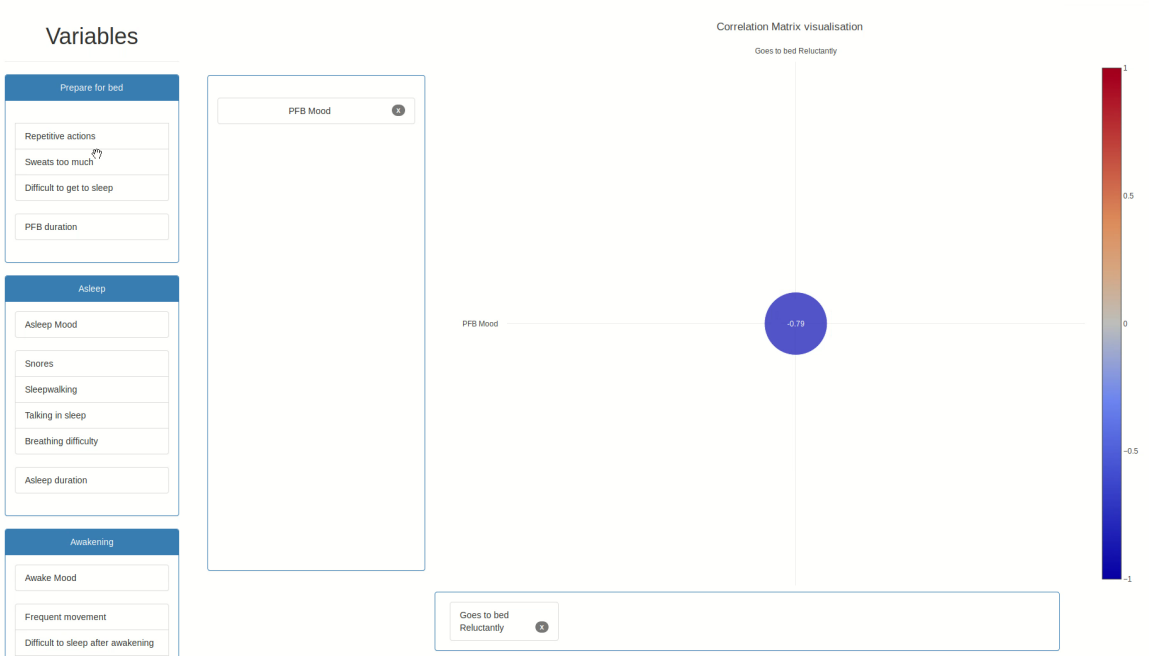


Figure 4.5: View after drag and drop. The colour of the circle reveals the correlation between corresponding variables with regards to the heatmap scale, correlation value is also written in text on the circle (-0.79 in this case). The radius of the circle shows the significance of the correlation. Finally, the transparency of the circle is a combination of the significance and correlation (by multiplication).

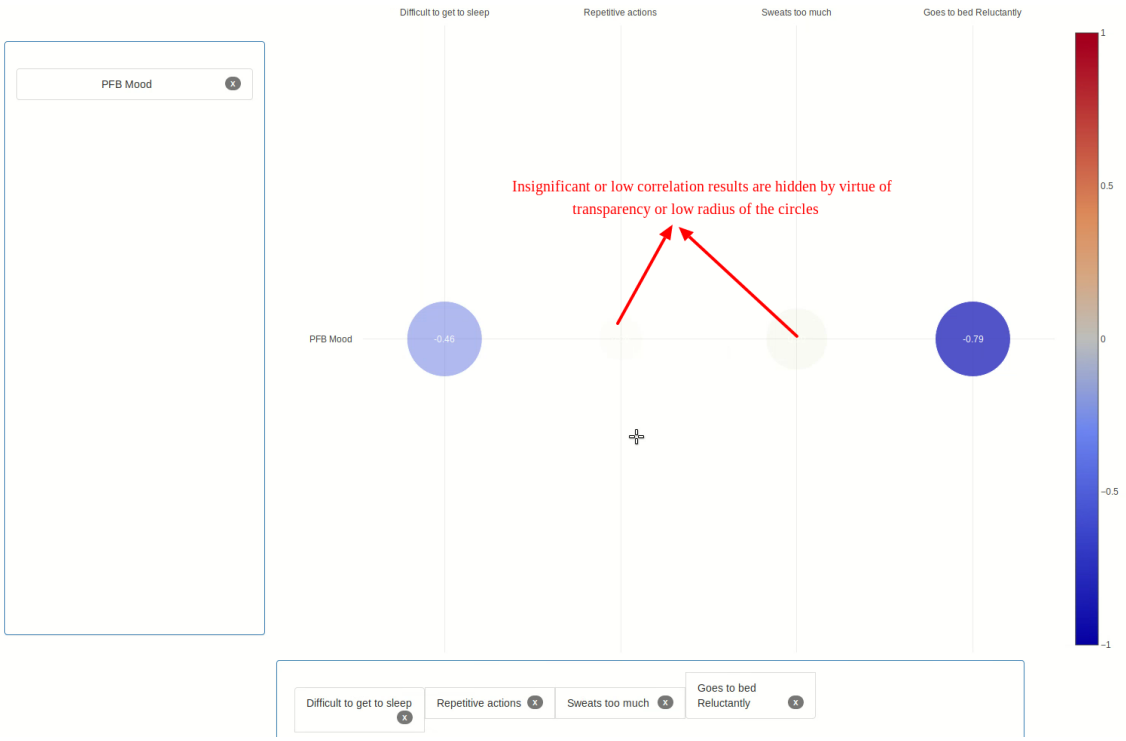


Figure 4.6: One vs many search in the variable space using drag and drop reveals all the variables that correlate with the variable of interest (PFB Mood in this case).

Figure 4.7: Scatter plot integrated into the visualisation, revealed by clicking the circle in the heatmap, this example shows the scatterplot of two discrete variables with positive correlation

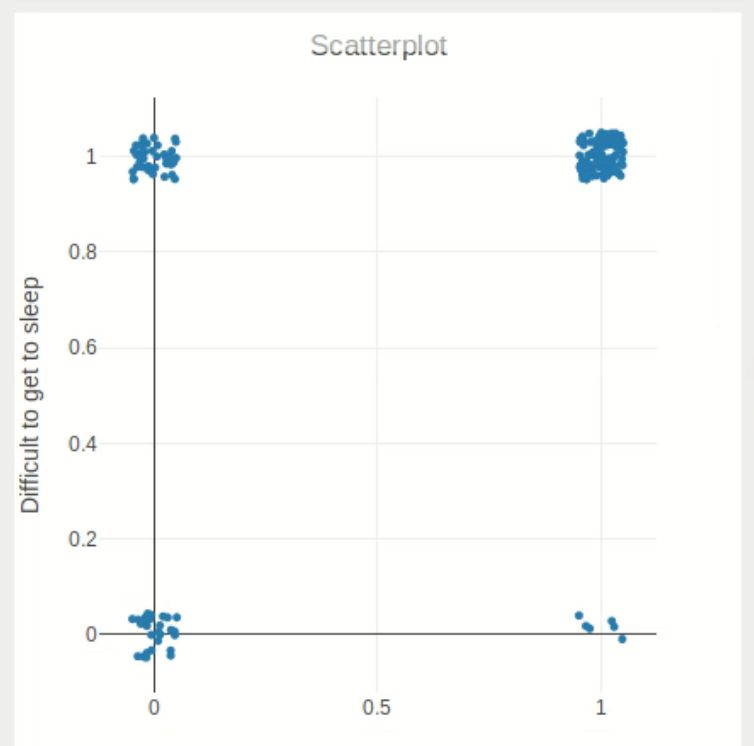


Figure 4.8: Jittering was added to the scatterplot popover where discrete variables were involved

Chapter 5

Study Design for Validation

In order to determine the efficacy of the Correlation Visualisation Toolkit, we decided to do a comparative analysis of the Toolkit with one of the ubiquitous Data Analysis Tools for Correlation. Microsoft Excel in the Office 365 Suite is reportedly one of the most popular data analysis software in market [30]. Our consultations with the clinicians also revealed that Excel was popular among their colleagues. Thus Microsoft Excel was chosen for the comparison. As stated earlier, the objectives of the Correlation Visualisation Toolkit is to reduce the workload, time demands and error-rate involved while analysing correlations in the sleep-related data. So we decided to design the experiment around measuring these three key parameters.

I used a mixed-method study design [26] as the strategy of inquiry, since some of the components of the research could not be represented exclusively by qualitative or quantitative data. The study design followed a pragmatic worldview [26]. I used embedded mixed methods design, the study may be viewed as quantitative strand embedded in case study.

The study was conducted on 20 adult participants over the age of 19 years. Originally the study was intended to be conducted in-person at the SFU Surrey campus. Unfortunately, the study was approved by SFU Office of Research Ethics in the midst of COVID-19 lockdown. As a result, we made changes to the study in order to deploy it online. Thus this chapter answers the following research question (RQ2 from 1.1):

> *How to run a User Study on the Correlation Visualisation Toolkit during a global Pandemic?*

As the University was completely shutdown, the participants were recruited through online means including social networking platforms and announcements in undergraduate virtual classrooms.

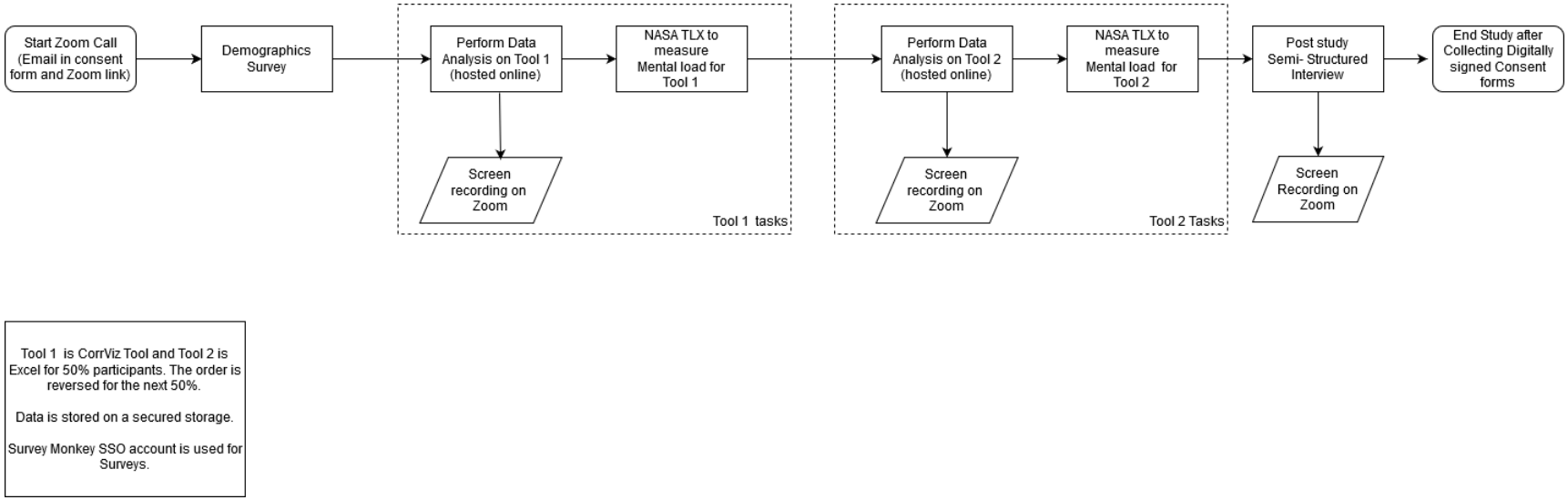


Figure 5.1: Overview of the study

During the study the participants were asked to use Correlation Visualisation Toolkit and Microsoft Excel. At the start of the study, they had to complete a pre-study questionnaire meant to collect demographic information. They were then asked to perform three different types of tasks using each of the software applications - the Correlation Visualisation Toolkit and Microsoft Excel. Their approach of carrying out the task, the accuracy of the analysis performed was observed and noted over screenshare. The screenshare was also recorded. Once they completed the tasks on each software, they were asked to complete the NASA Task Load Index questionnaire (NASA- TLX). This was followed by short interview. The overview of the actual study can be seen in Figure 5.1.

5.1 Study Description

5.1.1 Study Setup

During the initial email correspondence with the participants, I explained the basic outline of the study and send the Excel datasheet file (containing the dataset) that would be used for the study. Having a Microsoft Excel installation on the participant's computer was advertised as a requirement for the study. Figure 5.2 illustrates the dataset being loaded on Microsoft Excel. The Zoom link for the study, the training document for using the tools, and the consent form for the study were also included in the email.

The Correlation Visualisation Toolkit was hosted on the SFU web-space. The software was configured to be a standalone web-page to ensure the privacy of the users. You can see the hosting on Figure 5.3. The dataset to be used for the study was preloaded into the Correlation Visualisation Toolkit.

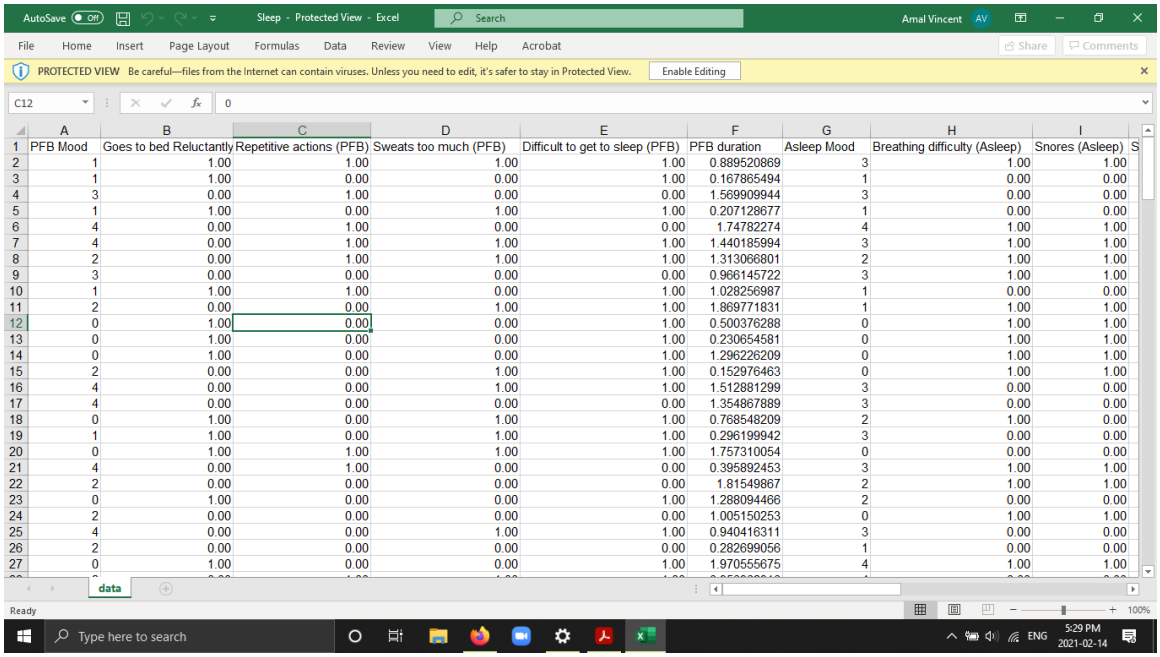
The data used in the Correlation Visualisation Toolkit was identical to that in Microsoft Excel, except for the file format. Excel used a xls file while, Correlation Visualisation Toolkit used a JSON file.

The dataset used for the study was prepared synthetically using mockaroo.com. Mockaroo is a a simple web app that allows you to create datasets from scratch [5]. The variables in the datasets could be configured to be random numbers or custom functions. I deliberately introduced functional relationships between some of the variables present in the synthetic dataset. In the real-world scenario actual patient data would be used for analysis. Particularly the data logged using SWAPP would be used for this analysis.

5.1.2 Study Procedure

The study procedure was completed over a Zoom call. Before the start of the study the participants were requested to fill in the pre-study questionnaire.

During the study, I went over the goals of the study and then proceeded to train the participants to use the first tool to complete the tasks. Half of the participants used the Correlation Visualisation Toolkit first, while the other half used Microsoft Excel first.



	A	B	C	D	E	F	G	H	I
1	PFB Mood	Goes to bed Reluctantly	Repetitive actions (PFB)	Sweats too much (PFB)	Difficult to get to sleep (PFB)	PFB duration	Asleep Mood	Breathing difficulty (Asleep)	Snores (Asleep)
2	1	1.00	1.00	1.00	1.00	0.889520869	3	1.00	1.00
3	1	1.00	0.00	0.00	1.00	0.167865494	1	0.00	0.00
4	3	0.00	1.00	0.00	0.00	1.569909944	3	0.00	0.00
5	1	1.00	0.00	1.00	1.00	0.207128677	1	0.00	0.00
6	4	0.00	1.00	0.00	0.00	1.74782274	4	1.00	1.00
7	4	0.00	1.00	1.00	1.00	1.440185994	3	1.00	1.00
8	2	0.00	1.00	1.00	1.00	1.313066801	2	1.00	1.00
9	3	0.00	0.00	0.00	0.00	0.966145722	3	1.00	1.00
10	1	1.00	1.00	0.00	1.00	1.028256987	1	0.00	0.00
11	2	0.00	0.00	1.00	1.00	1.869771831	1	1.00	1.00
12	0	1.00	0.00	0.00	1.00	0.500376288	0	1.00	1.00
13	0	1.00	0.00	0.00	1.00	0.230654581	0	1.00	1.00
14	0	1.00	0.00	0.00	1.00	1.296226209	0	1.00	1.00
15	2	0.00	0.00	1.00	1.00	0.152976463	0	1.00	1.00
16	4	0.00	0.00	1.00	1.00	1.512881299	3	0.00	0.00
17	4	0.00	0.00	0.00	0.00	1.354867889	3	0.00	0.00
18	0	1.00	0.00	1.00	1.00	0.768548209	2	1.00	0.00
19	1	1.00	0.00	1.00	1.00	0.296199942	3	0.00	0.00
20	0	1.00	1.00	1.00	1.00	1.757310054	0	0.00	0.00
21	4	0.00	1.00	0.00	0.00	0.395892453	3	1.00	1.00
22	2	0.00	0.00	0.00	0.00	1.81549867	2	1.00	1.00
23	0	1.00	0.00	0.00	1.00	1.288094466	2	0.00	0.00
24	2	0.00	0.00	0.00	0.00	1.005150253	0	1.00	1.00
25	4	0.00	0.00	1.00	1.00	0.940416311	3	0.00	0.00
26	2	0.00	0.00	0.00	0.00	0.282699056	1	0.00	0.00
27	0	1.00	0.00	0.00	1.00	1.970555675	4	1.00	1.00

Figure 5.2: Microsoft Excel with the dataset loaded

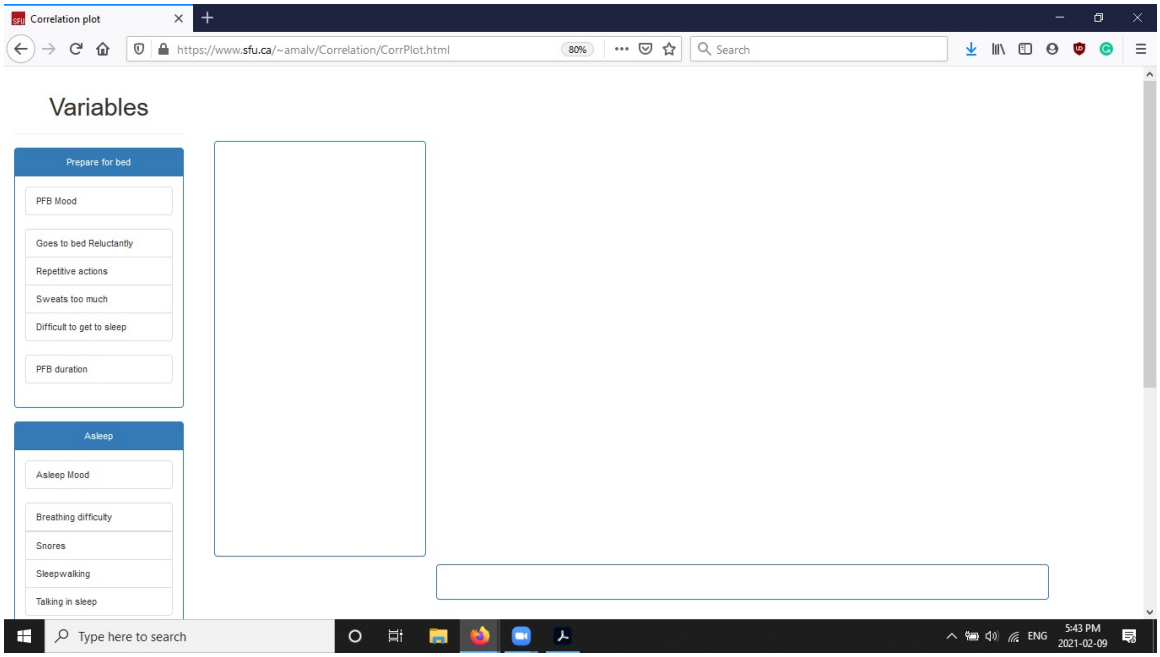


Figure 5.3: Correlation Visualisation Toolkit hosted on SFU Webspace, with dataset loaded

The tasks

There were two sets of tasks used, Set A and Set B. Half of the participants (50% split on each software) used Correlation Visualisation Toolkit for solving Set A and Excel for solving Set B. The combination was flipped for the other half. Set A and Set B were mostly similar, barring different target variables in the tasks. In any case, the procedure to complete the tasks remained the same irrespective of the task set. The participants were requested to share the software tool window over Zoom as they completed the tasks. The screenshare was recorded locally through Zoom. I also observed the actions on the participant's screenshare as they completed the tasks and took notes. The tasks in each set were classified into three types:

1. **Simple Correlation Analysis:** These tasks had two target variables in each question and the participants were required to analyse the correlation between the two. The participants completed a total of three questions of this category for each tool. The participants were required to choose from 'yes', 'no' or 'do not know' options, after analysing the data using the tool(either Excel or Correlation Visualisation Toolkit). *E.g., Is "PFB mood" correlated to "Difficult to get to sleep"?*

2. **Correlation Search Problems:** This task involved analysing one vs many correlation. The participant was given a single target variable and they were asked to analyse its correlation with all other variables. This task can be viewed as an expanded version of the Simple Correlation Analysis task. *E.g., Which "Asleep Behaviour(s)" is "Cough" correlated with?*

3. **Functional Relationship Analysis:** The goal of this task was to understand the functional relationship between two given target variables. For this, the participant had to produce the scatter plot between the two variables. The participants were then required to identify which one of the following functions best described the scatter plot - linear(decreasing/increasing), parabolic and exponential. Where the participants felt that the relationship did not fall into any of the above functions, they were asked to indicate the functional relationship as 'other'. *E.g., What is the functional relationship between 'Asleep duration' and 'Awake duration'?*

For completing the correlation analysis tasks the participants were required to analyse the significance and correlation coefficient of the relationship between the two variables. Where the significance, $p < 0.05$ and the absolute value of correlation coefficient, $|r| > 0.5$, the participants were instructed to assume correlation. Detailed procedural instructions for each tool can be found in the training document in Appendix A.2.

After completing the tasks on the first tool, the participants were asked to complete the NASA-TLX questionnaire. The whole process was repeated for the second tool.

5.1.3 Post-study interview

At the end of the study I conducted a semi-structured interview to understand the participant's experience with the two software tools in a qualitative manner. The interview was recorded locally on my computer through Zoom. The guiding questions could be grouped broadly into four categories.

1. Questions reviewing participant's familiarity with data analysis tools.
2. Questions focused on the experience of the participant with the Correlation Visualisation Toolkit
3. Questions comparing Microsoft Excel with Correlation Visualisation Toolkit
4. Questions based on observations from the screenshare while the participant was completing the tasks.

These were aimed to help answer RQ3 and RQ4 defined in section 1.1. More details on the guiding questions can be found in appendix section A.4.

5.1.4 Questionnaires

The study made use of two Questionnaires, namely the Pre-Study Questionnaire and the NASA-TLX Questionnaire. Data from all questionnaires were collected using SFU's Survey Monkey hosting.

Pre-Study Questionnaire

The pre-study questionnaire collected demographics information about the participants including, but not limited to age, information about disabilities affecting vision and experience with data analysis tools. See Appendix A.1 for the questionnaire.

NASA Task Load Index (NASA-TLX) Questionnaire

NASA-TLX is a tool used to evaluate workload in various human-machine systems [34]. Using multidimensional ratings, NASA-TLX provides the Overall Workload Score for the task. The Overall Workload score is a number in [0, 100]. The six sub-ratings are:

1. Mental Demand
2. Physical Demand
3. Temporal Demand
4. Performance
5. Effort

6. Frustration

The questionnaire was used to capture the workloads associated with using Correlation Visualisation Toolkit and Microsoft Excel. The participants were asked to fill in the form immediately after completing the tasks on each of the above softwares. See Appendix A.3 for the questionnaire.

5.1.5 Data Collected

The data collected during the study came from the following sources

- **Data from Questionnaires**:
  - Pre-Study Questionnaire
  - NASA-TLX Questionnaire
- **Observational Data and Screen Recording**: As the participants completed the tasks on the software tools, I observed over the Zoom screenshare. The data was also recorded locally on my computer to facilitate further analysis (if required) in the future. I made note of observations relevant for validating the Correlation Visualisation Toolkit. Some of the observations noted were also followed up during the interview. The recordings were also used to calculate task completion times for participants.
- **Interview transcripts**: As discussed earlier the data from the interviews were recorded locally. After the study, these were transcribed and analysed by me.

5.1.6 Participants

For the study we recruited students and workers at SFU who were over the age of 19. The participants were recruited through online means including social networking platforms and announcements in Undergraduate virtual classrooms. They were compensated 10 CAD for their time.

The participants were between the age group of 19 and 34, with the average age of 25 years. All of the participants indicated that they had previous experience with Excel at some point in their life. None of the participants had any conditions that affected their vision.

5.2 Ethical Considerations

The following Ethical considerations were made to ensure the welfare of the participants and their data.

- **Zoom Protocol**: The whole study was conducted over a Zoom call. Zoom is a video conferencing solution. Zoom was chosen because it provides video and audio encryption using 'using AES-256 encryption, and optional end-to-end encryption' [12]. This allowed us to ensure the privacy of the data transmitted over the zoom call. Each participant was sent a unique Zoom link, this was to ensure that 'Zoombombing' [40] did not happen. The study followed relevant guidelines from the Ethical Considerations and best practices set by the SFU Office of Research Ethics [2].

- **Informed Consent**: At the start of the Zoom call I went over the goals of the study, the study procedure and the concept of informed consent. The participants were then requested to provide informed consent before proceeding with the study. I took care in explaining the participant that consent is an ongoing process and, they can opt out at any point. The consent form was converted into a digitally signable forms using Adobe Acrobat DC to suit the online study.

- **Data Storage**: In order to ensure privacy of the participants, names of the participants were anonymised and only the unique identifier (such as P1, P2) was listed for each participant. All the data from the participants were stored in SFU based servers to ensure that the data would not be breached due to legislations such as the Patriot Act [14].

5.3 Data Analysis

5.3.1 Quantitative data

Two dependent variables (DVs) were used in the quantitative data analysis:

1. Overall Workload from NASA-TLX Questionnaire

2. Time taken for completing all tasks in seconds

The total Time for task completion for each participant was carefully extracted from the screen recording of them completing the tasks. The independent variable (IV) was the software tool used. So, the IV can take two values, either Microsoft Excel or Correlation Visualisation Toolkit. I then ran Paired Samples t-Test on each of the DVs to understand the influence of the software tool on the DVs.

Paired Samples t-Test/ Dependent t-Test/ Repeated Measures t-Test

Paired Samples t-Test or paired t-Test is a statistical test used to compares two means that are from the same source (participant, object, or related units) [53, 10]. Our study involves making the DV measurements with both Microsoft Excel and Correlation Visualisation Toolkit. The test can be used to determine if there is statistical evidence that the mean

difference between paired observations on a particular outcome is significantly different from zero. The study is also known as Dependent t-Test or Repeated Measures t-Test.

For the results from the Paired t-Test to hold, the following conditions need to be satisfied:

- DVs should be continuous: Clearly the Total Time taken for all tasks is continuous. And NASA-TLX Overall Workload as discussed earlier occurs in a continuous interval of [0,100].
- Participants in each sample or group are same: We have already mentioned that all participants used both the software tools.
- The distribution of the difference between the two paired values is approximately normal: In order to verify the condition we ran the Shapiro-Wilk statistical test for normality [41] on the difference between paired samples (samples taken on each of the tools). This was further verified using QQ-plot (Quantile to Quantile plot) [16], which is a visual-analytic technique that can be used in verifying normality. See Appendix B for results of normality tests. These tests suggested that the data was normal.
- No outliers in the difference between the two paired values: In order to verify that no outliers existed in the difference between paired samples (samples taken on each of the tools) we inspected the box plot of the difference. See Appendix C for results of outlier detection. No outliers were observed for either variables.

After verifying the above conditions we proceeded to the Paired t-Test for each of the DVs, with the IV remaining the same for both. The results from the two paired t-Tests (one for each DV) were used to validate RQ3 and sub-questions defined in section 1.1.

5.3.2 Qualitative data

The interviews were transcribed by me and I then proceeded to do thematic analysis on the transcripts. I used open coding to generate the themes. Followed by axial coding to categorise them. The reliability of the codes were verified by two experts independently. The analysis was guided by RQ3 and RQ4 defined in section 1.1:

> RQ3: *Does the Correlation Visualisation Toolkit perform better than Microsoft Excel in terms of overall workload and time taken for task completion?*
>
> RQ4: *How can we improve the Correlation Visualisation Toolkit?*

For RQ3, the qualitative data from the interviews was used to provide complementary and contextual information about the results from the quantitative data analysis. As for RQ4 the qualitative data was exclusively used to generate recommendations to improve the Correlation Visualisation Toolkit.

Chapter 6

Results

This chapter focuses on the results from the data analysis. The results from the data analysis are organised into the following three sections:

1. **Comparison of Workloads and Time taken for completing tasks Correlation Visualisation Toolkit and Microsoft Excel**: In this section we discuss the results from the two Paired t-Tests discussed in 5.3.1. Qualitative data from the interviews was then used to add complementary and contextual information to the results.

2. **Improving the Correlation Visualisation Toolkit**: The section focuses on results from the thematic analysis of the qualitative data from the interviews. The section provides recommendations to improve the Correlation Visualisation Toolkit

3. **The Curious Case of P05**: One of the participants in the study, P05 faced significant distress while using Microsoft Excel. This is discussed in the section.

6.1 Comparison of Workloads and Time taken for completing tasks Correlation Visualisation Toolkit and Microsoft Excel

In this section we start with results from the quantitative data analysis. Then we use qualitative data to reaffirm the conclusions from the quantitative data and to explain why the quantitative data results are as such. In this section we answer RQ3 defined in Section 1.1.

> RQ3: *Does the Correlation Visualisation Toolkit perform better than Microsoft Excel in terms of Overall Workload and time taken for task completion?*

Or from quantitative perspective we can formulate the following hypotheses from the above question:

- H1: *Mean Overall Workload for Correlation Visualisation Toolkit is less than that for Microsoft Excel.*
- H2: *Mean time taken for completing correlation analysis tasks on Correlation Visualisation Toolkit is less than that for Microsoft Excel.*

6.1.1 Comparison of Workload

Our analysis of the results for comparison of Workloads between Microsoft Excel and Correlation Visualisation Toolkit began by looking into the quantitative data. We specifically analysed the Overall Workload parameter obtained from the NASA-TLX questionnaire completed by the participants. As mentioned in the previous chapter, all participants worked on both the tools in a counterbalanced manner. Half of the participants worked initially with the Correlation Visualisation Toolkit, while the other half worked with Microsoft Excel. I applied Paired t-Test in order to analyse the data. I used hypothesis H1 for analysing the data.

As mentioned earlier, a Paired Samples t-Test was conducted to compare the NASA-TLX Overall workload for Correlation Visualisation Toolkit and Microsoft Excel. From the results we can conclude that there was a significant average difference between NASA-TLX Overall Workload scores for Correlation Visualisation Toolkit and Microsoft Excel.

Variables	Description
CorrViz_Overall	NASA-TLX Overall Workload on Correlation Visualisation Toolkit
Excel_Overall	NASA-TLX Overall Workload on Excel

Table 6.1: Variables in analysis of Workload

The mean NASA-TLX Overall Workload for Correlation Visualisation Toolkit (M = 21.93, SD = 13.183) was lower than that for Microsoft Excel (M = 62.77 , SD = 17.932) (see table 6.2). And this was significant, ($t_{19} = -11.442, p < 0.001$) (see table 6.3).

Variable	Mean	N	Std.Dev.	Std. Error Mean
CorrViz_Overall	21.93	20	13.183	2.948
Excel_Overall	62.77	20	17.932	4.010

Table 6.2: Paired Samples statistics for NASA-TLX Overall Workload

Paired Difference	Mean	Std. dev.	Std. error	t	df	Sig (2 tailed)
CorrViz_Overall - Excel_Overall	-40.833	15.959	3.569	-11.442	19	.000

Table 6.3: Paired Samples t Test results on NASA-TLX Overall Workload

These results suggest that, **Mean Overall Workload for Correlation Visualisation Toolkit is less than that for Microsoft Excel**.

6.1.2 Comparison of Time Duration

Next we analysed the DV, total time taken to complete all the tasks. The total Time (in seconds) taken for completing tasks on Correlation Visualisation Toolkit and Microsoft Excel is represented by CorrViz_Time and Excel_Time respectively.

Variables	Description
CorrViz_Time	Total Time for tasks on Correlation Visualisation Toolkit
Excel_Time	Total Time for tasks on Excel

Table 6.4: Variables in analysis of Time

As discussed earlier, Paired Samples t-Test was conducted to compare the total time taken for completing tasks on Correlation Visualisation Toolkit and Microsoft Excel. From the results we can conclude there was a significant difference in total time taken for completing tasks on Correlation Visualisation Toolkit compared to that on Microsoft Excel.

Variable	Mean	N	Std.Dev.	Std. Error Mean
CorrViz_Time	281.8	20	57.76140	12.91584
Excel_Time	1094.3	20	310.55436	69.44207

Table 6.5: Paired Sample Statistics for Total Time

Paired Difference	Mean	Std.Dev.	Std. Error Mean	t	df	Sig.(2 tailed)
CorrViz_Time - Excel_Time	-812.5	281.08409	62.85231	-12.927	19	0.000

Table 6.6: Paired Samples t Test result for Total Time

The mean for total time taken to complete tasks on Correlation Visualisation Toolkit (M = 281.8, SD = 57.76140) was lower than that on Microsoft Excel (M = 1094.3 , SD = 310.55436) (see table 6.5). And this was significant ($t_{19} = -12.927, p < 0.001$) (see table 6.6).

These results suggest that, **Mean time taken for completing correlation analysis tasks on Correlation Visualisation Toolkit is less than that for Microsoft Excel**.

6.1.3 Reaffirming and understanding why Correlation Visualisation Toolkit performs better using qualitative data

The qualitative data from the interview transcripts reaffirmed the conclusion that Microsoft Excel presents excessive burden in terms of workload and time taken to complete tasks.

To our question '*If you have a choice, which one of the two tools would you choose for correlation analysis in the future?*', all the participants unanimously picked correlation Visualisation Toolkit. Detailed analysis of interview transcripts revealed why they came to that conclusion. This is discussed in detail below.

Correlation Visualisation Toolkit is Easy, Intuitive and Straightforward

Participants found Correlation Visualisation Toolkit more easy, intuitive and straightforward compared to Microsoft Excel. The analysis showed that following were the major reasons for this:

- **Effortless Interaction**: The Correlation Visualisation Toolkit's amicable interaction (discussed in section 4.1). The features of the interaction that were favourite among the participants were the Drag and Drop and the Scatterplot option.

  > P4 says, *'I think compared to using Excel it is much easier, you just drag and drop and then you just compare. And using the scatter plot you can see the correlation pretty easy and it is straightforward.'*

  The Drag and Drop interaction for the Correlation Visualisation Toolkit simplified the process of variable selection for analysis. Most of the participants found the process of selecting variables by using the drag and drop feature largely intuitive. Most of the participants feel they would have been able to use the drag and drop feature with minimal supervision/training.

  Coming into the Scatterplot functionality of the Correlation Visualisation Toolkit, the participants found the process easier compared to Excel. The Scatterplot was displayed only on click on the Correlation Visualisation Toolkit, as a result the participants felt they were not confused by excessive information. Finally, it was easier to choose the variables to be plotted, since the circles corresponding to the pairs of variables was largely apparent.

- **Good Visual Vocabulary**: The analysis suggested that Correlation Visualisation Toolkit had good visual vocabulary that simplified the tasks to a large extend. In Excel the participants had to peruse through rows, columns and numbers, and then go through intricate process flow to get to the results. The Correlation Visualisation on the other hand Toolkit presented the results graphically. In the Correlation Visualisation Toolkit there well defined visual metaphors, including the colour, transparency and size of the circle as discussed in 4.2. These visual metaphors largely simplified the analysis of correlation. This is largely evident in what P9 says:

  > *'I simply just put there and I get the result and I can interpret the results and it is visualised. Unlike Excel it [Correlation Visualisation Toolkit] has*

colours and tells me how you said that the parts won't show up if the p is not there. So it is much more easier to look for what I am looking for, so I don't have to worry about all those things. But I have to do it for excel.'

- **Lower learning curve**: The Correlation Visualisation Toolkit had a lower learning curve compared to Microsoft Excel, as use of many of the features were largely apparent (e.g., Drag and drop feature described above). Many of the participants (7 out of 20) felt they would have required minimal (or no) training/ supervision to complete tasks on Correlation Visualisation Toolkit. As discussed above, the Visual Vocabulary of the Correlation Visualisation Toolkit was easier to learn than the process flow on Excel. P16 says,

  I think the toolkit was much easier to learn. I think you explained first time once and I really didn't have to ask a question. But Excel you explained, but couple of times I had to ask the question there were couple of different elements to it. So learning curve for the toolkit was easier.

  Also, the components of the Correlation Visualisation Toolkit interaction, hinted the user what to do next. For example, the x-axis and y-axis boxes hinted the user on how to use the drag and drop feature to select variables. Furthermore, the participants felt largely frustrated while learning to find correlations on Microsoft Excel, compared to the Correlation Visualisation Toolkit. Finally, some of the participants pointed out that the interaction on Microsoft Excel required more background knowledge compared to Correlation Visualisation Toolkit. For instance, Microsoft Excel requires knowledge of row and column selection and manipulation. All our participants had previous experience with Microsoft Excel, so this was not an issue during our study.

Correlation Visualisation Toolkit is Effective, Useful and Convenient

The major motivation for the development of the Correlation Visualisation Toolkit came from the necessity reducing the workload and time required to analyse the correlations in the sleep data (discussed in Chapter 1). The participants found the Correlation Visualisation toolkit effective in terms of reducing the time to complete tasks. The participants also note that the tool required far less effort from them. We identified that the following were the reasons for this.

- **Fewer steps to goal**: As warranted by the results from the quantitative data, Correlation Visualisation Toolkit allowed users to finish their tasks quicker when compared to Microsoft Excel. This was mostly because Correlation Visualisation Toolkit had fewer steps to get to the result. Fewer steps also meant lesser mental effort and workload on the participants. On the Correlation Visualisation Toolkit, the selection of

variables and the analysis of correlation is largely simplified by the Effortless Interaction and the Visual Vocabulary. In order to analyse the correlation, all the participant needed to do was drag and drop the desired variables, and analyse the colour, radius and/or transparency of the circle. This procedurally required fewer steps than Excel. P14 says,

> *'So for excel I felt like there was a lot of steps to remember it can be a bit too much, it like a lot of numbers are spinning at you. It can it be a bit hectic. Like the other one [Correlation Visualisation Toolkit] it was easier than number seeing, it's just labels. And ok colour, I mean it is correlated and it is way easier for me to understand than look at the number and compare the number.'*

- **Confident about results**: The participants felt they were far more confident about the results for the tasks they completed on the Correlation Visualisation Toolkit, than on Microsoft Excel. The procedural complexity on Excel made the participant wary of the results, as they were not sure if they followed the steps accurately. Furthermore, the participants accidentally selected the wrong column for data analysis occasionally. Therefore, they had to double check their actions throughout the study on Excel. This took time, and also required a significant workload from the participant. Thus, the participants felt they were under less pressure while concluding results on Correlation Visualisation Toolkit than on Excel. P18 says,

  > *'I found on your tool [Correlation Visualisation Toolkit], I didn't have to double check if I did it correctly. While with excel I kept having to double check.'*

  Coming into the scatterplot option, the participants noted that the Jittering feature allowed them to understand the functional relationships better on the Correlation Visualisation Toolkit, than on Excel which had regular scatterplots. This was another factor that contributed into the participants' confidence in concluding results. P3 says,

  > *'In the first one in the tools [Correlation Visualisation Toolkit] you had more points in the middle line. So I knew that kind of was linearly related. But in excel file all dots had had the same size and so it didn't give you sense of correlation.'*

- **Less Mental Effort**: The participants confirmed that they experienced less mental effort on the Correlation Visualisation Toolkit, than on Microsoft Excel. As discussed above, Correlation Visualisation Toolkit had fewer steps to complete tasks when compared to Excel. Also, the participants were not distracted by redundant information on

Correlation Visualisation Toolkit unlike Excel. Furthermore, the process flow on Excel required the user to double-check each and every step. All these factors contributed to reducing the participants' mental effort and in turn the workload. P8 says,

> *'And when you want to select the variable, it takes much effort because you have to go through more steps to reach your goal.'*

Correlation Visualisation Toolkit is Pleasant, Accessible and less Frustrating

The qualitative analysis revealed that the participants found the Correlation Visualisation Toolkit more pleasant and accessible and less frustrating. Following are their rationale for this.

- **Simultaneously handle multiple variables**: Correlation Visualisation Toolkit allows users to drag and drop multiple variables to both x-axis and y-axis boxes. This proved particularly handy in completing the Correlation Search Problems (see 5.1.2). The interaction also allowed users to obtain both p-value and effect size simultaneously. P1 says,

  > *'And also having the ability to put multiple variables that is also handy when I am thinking about excel that is a lot of pain.'*

- **Hides redundant information**: Participants also pointed out that the display of redundant information on Excel was distracting. This was one of the most frustrating aspects of Excel for the participants. On the Correlation Visualisation Toolkit the dataset itself was hidden from the user and only the names of the variables and the correlation between the variables were displayed using visual vocabulary in a pleasant manner. Even inside the visualisation uncorrelated relationships were obfuscated by means of transparency (discussed in 4.2). This allowed the participants to focus on the task rather the procedure. Thus, unlike Excel Correlation Visualisation Toolkit does not overwhelm or frustrate the user with information. P10 says,

  > *'For Excel, all the data is at your face, so it is hard to find stuff. I have been used to it for multiple things so kind of have a familiarity there. But there was so much data there so slower and harder there.'*

- **Better presentation**: Participants found the categorical organisation of variables in the interaction of Correlation Visualisation Toolkit very convenient. This allowed them to locate and select variables using the drag and drop interaction easily. Two of the participants felt, the presence of colours in the data visualisation also improved the experience for them. P13 says,

'It's a lot more intuitive than Excel for sure. Since, the values are kind of... are structured, they are not categorised in the way on Excel. In the tool you have boxes that say what happens when you are asleep, you are preparing for bed [PFB] and you are awake, and you can easily go to ,"I need to check PFB vs asleep" that is really easy compared to Excel which I have to ... almost go through text.'

6.1.4 Summary of findings on Correlation Visualisation Toolkit vs Microsoft Excel

Thus, the results from the quantitative data analysis suggest that the Correlation Visualisation Toolkit performs better than Microsoft Excel in terms of Overall Workload and Total Time taken for task completion. Furthermore, the qualitative data reaffirmed the conclusions from the quantitative data and adds richer detail on the merits of the Correlation Visualisation Toolkit. This includes details on why Correlation Visualisation Toolkit works better than Microsoft Excel for the Given tasks. Finally, the qualitative data also showed how the Interaction and the Visualisation of the Correlation Visualisation Toolkit improved the experience of correlation analysis for the participants.

6.2 Improving the Correlation Visualisation Toolkit

We recognised that the Correlation Visualisation Toolkit is far from a finished product and we needed suggestions to improve the tool. This section is driven by the RQ4 from 1.1:

RQ4: *How can we improve the Correlation Visualisation Toolkit?*

To achieve this, we had questions in the interview, that were tailored to give us information about the shortcomings of the tool. The qualitative data also gave us suggestions about new features and improving the existing ones. Here are the conclusions.

6.2.1 Desired features from Excel

While, the Correlation Visualisation Toolkit outperforms Microsoft Excel in many things, this comes with a caveat. Unlike the Correlation Visualisation Toolkit which is tailored to a specific task. Excel on the other hand can cater to multiple tasks with it's wide range of features. In that regard Excel is far more powerful. Following are the major talking points:

Viewing datasets and more

Unlike Excel, the Correlation Visualisation Toolkit did not have a provision to display raw data. While, SWAPP itself had options that allowed users to download data, this was not apparent for the participants in our study. Some of the participants were therefore keen on

having an option to display raw data. A few participants also felt it would be nice to see the p-value numerically at times. P12 says,

> *'For Excel you can see all the values. After you calculate the R value you can see all these in table. For the t test you can see values other than the p value, if that is useful that would be good.'*

We are however concerned whether participants maybe distracted by the raw spreadsheet data. A reasonable trade-off maybe to add a small button that brings up the spreadsheet option allowing the user to view, manipulate and download the data, at will. As for the p-value we feel a better way to go about it would be to display it in a tooltip when the user hovers over the circle. That way there would not be immense amount of information on the user's window by default. But these features would only be included if clinicians find it necessary. So further consultation with clinicians are required before inclusion of such features.

Desirable functions from Excel

The qualitative analysis revealed that improving the tool may involve, adding the possibility of running more statistical tests on the Correlation Visualisation Toolkit. The other direction for improvement is in terms or adding support other types of data visualisations. P19 says,

> *'Excel has more function than drag and drop. You know how excel can do more than one scatter plot, it can do a box plot it can do different charts. Oh, maybe an improvement is that you can make different charts instead of the scatter plot, we can select a bar chart or make other charts. So it's better for analysing data if you have more charts it's useful.'*

However, once again we are wary about adding additional options to the Correlation Visualisation Toolkit as it could potentially confuse the users. And this may even lead to misinterpretation of the data, especially if the user is not skilled enough to work with such options.

6.2.2 Issues with current features

The other direction for improving the tool is by polishing the existing features. During the study we noticed that the Correlation Visualisation Toolkit suffers from the following problems.

Software implementation issues

Two of the participants experienced resolution problems while using the Correlation Visualisation toolkit. And subsequently reported it during the interview. While, we had designed

for responsiveness, our software implementation seems to break on some of the screen sizes participants used. A more rigorous responsive-design testing routine should fix the problem. P1 says,

> *'Overall it was good. The low spaces for the fields, it looks a bit disproportionate.'*

The only other isolated (for one participant) software issue was delay in loading the visualisation. I believe this was probably a result of poor internet bandwidth. Throttled bandwidth testing may help understand the problem.

Web element issues

The major problems with regards to Web elements was the size of icons (e.g., clear button for variables) and fonts. Two of the participants indicated that the size of the icons in the interaction were significantly small. The problem probably a result of poor responsiveness with regards to changing screen sizes. As with the resolution problem, a rigorous responsive-web design testing should resolve the problem. P7 says,

> *'Another thing was I did not quite realise what those categories were for. The labels of the category seemed real small. If it were a bit bigger it would be a lot more easier.'*

One of the participants also pointed out that the contrast between font colour for the writing on the coloured circles, and the colour of the circle was poor. As a result, it was difficult to read the numbers. So the font colours used need to be revisited. Finally, one of the participants felt that the borders of the x-axis and y-axis boxes were thin, hence were difficult to spot. One way around this would be to add solid colour to boxes and labelling the boxes.

Aesthetics

Another largely apparent aspect for future improvement, which was also brought up by participants was the aesthetics of the tool. Aesthetics of the tool are far from polished and there is significant room for improvement in this area. P20 says,

> *'Also, the visual design is also not that good.'*

6.2.3 Uncertainty in the current setup

Another important shortcomings in the Correlation Visualisation Toolkit came in the form of various uncertainties in the interaction and the visualisation. Those are discussed here.

Unclear, how to produce Scatterplot

This was by far the most important problem to be dealt with in the next version of the Correlation Visualisation Toolkit. A significant number of participants felt there was nothing in the interaction implying that clicking on the circles in the Correlation Visualisation Toolkit would produce Scatterplot. P13 says,

> *'It seems a bit weird to me that you had to click on the circle to get the scatter plot, maybe a button that says scatter plot for various axes might work better. Though I guess in the cases where you have multiple circles that might be a bit more problematic. Nothing told me I had to click it . Everything else makes sense.'*

One straightforward way to resolve the issue would be to include a short caption in the interaction about how to produce a scatter plot. The other way would be to improve the training documentation for the tool.

Confusing transparency of circle

The other major feature the participants were confused about was the transparency of the circle. As discussed in 4.2, the transparency of the circle was adjusted according to both correlation coefficient and significance. While, the transparency did prevent the participants from clicking or analysing the transparent circles, this was distracting to some degree. P15 says,

> *'Maybe, all I can think of is probably the... maybe the white colour in the circle. That's like the only thing I can really think of.'*

Tweaking the the constant used to adjust the transparency of the circles, maybe one of the ways to improve the experience for the users. The other obvious way would be to improve the training documentation for the Correlation Visualisation Toolkit.

Shortcomings in training

Many of the uncertainties experienced, stemmed from shortcomings in the training documentation. Also, given enough time to get used to the tool, the users may perform much better. This is probably the easiest way to address many of the problems. P4 says,

> *'Yeah maybe you could give clear instructions, about how to do it and then there is toolbar for doing multiple including charts and diagrams, scatterplots.'*

Other problems

There were other isolated issues faced by individual participants. Following are the issues:

- A participant pointed out that, they were not particularly clear that they could drag and drop multiple items into the same box. However, this was directly contradicted by many of the other participants.
- One of the participants felt the need for a reference size for circles to compare the p-value. However, the problem was resolved once they started seeing non correlated relationships. Giving examples in the training documentation maybe a way to get around this.
- Lastly, one participant was confused initially by the simultaneous use of colour and number on top of the circle to indicate correlation. Once again this problem is best resolved by improving the training documentation.

6.2.4 Upgrades for existing features

All suggestions for upgrading existing features came from individual participants, and no other participant requested for the same. Those are summarised below.

Option to zoom in/out the Scatterplot

Currently the Scatter plot functionality in the Correlation Visualisation Toolkit produces a static view. A participant pointed out that it would be handy to have the option to zoom in. P18 says,

> *'It [Scatter plot on Correlation Visualisation Toolkit] was really convenient to access it, with the clickable button. If I were to click on it again will I be able to get a bigger picture? I didn't test it.'*

Drag multiple variables simultaneously

One of the participants requested the ability to drag multiple variables simultaneously in a single drag. This is another feature to be explored. P11 says,

> *' It will be kind of nice if we could select a bunch of them at the same time a lot easier, just drag and hold them at the same time.'*

Improve categorisation of variables

There was also a request to indicate the the nature of the variable, i.e., whether it is continuous, discrete or binary, in the the categorical boxes. P1 says,

> *' Maybe you can be indicated with colour so you can know it is a different type of variable. So then you can know what is happening so overall it looks good.'*

More than two variables in scatterplot

A participant thought that the scatterplot can represent up to 4 at the same time, i.e., the third and fourth discrete/binary variable would be represented by colour and shape of the points in the plot. However, how the interaction would work in such a scenario needs to be pondered. P2 says,

> *' And then it was good if I can change the points. Maybe if they had multiple scatter plots. If it had multiple variables if we wanted to check such as to see their scatter plot. It was good if we could find different colours and different shapes of points.'*

Discrete Colour scale

The Correlation Visualisation Toolkit currently has a colour scale for correlations, with continuous colour gradients. P1 felt, a discrete colours scale might work better.

6.2.5 Suggested new features

As with 6.2.4, all suggestions for new features also came from individual participants, and no other participant requested for the same.

Clear all button

A participant pointed out that, it was cumbersome to clear multiple variables at once from the x-axis and y-axis box, after dragging and dropping a lot of variables. This could easily be resolved by a clear-all button.

Search bar for variables

One of the participants felt it took some time to scan through the variable list, and find the desired variable. A straightforward way to resolve this would be to incorporate a search bar in the next iteration of the Correlation Visualisation Toolkit.

Option to save notes

While, it didn't affect the experience of using the Correlation Visualisation Toolkit, P20 thought that an option to save notes would be a great upgrade.

6.3 The curious case of P05

While, all of the other participants had no issue completing all of the given tasks on both the tools, P05 had to forfeit one of the data analysis tasks on Microsoft Excel due to anxiety issues. P05 took a break between the study, lasting about 15-20 minutes before

they were able to resume the study. While we have very little evidence in order generalise the phenomenon, we are deeply concerned by the effect of state of the art tools like Microsoft Excel on people with anxiety issues. One of the problems that particularly disturbed was the amount of information the person saw on Excel. So, one of the directions of future research would be to analyse accessibility of the Correlation Visualisation Toolkit for people with anxiety.

The excerpts from interview is below. From the excerpts it is evident that the anxiety reaction was exclusively from Microsoft Excel. About Excel P05 says the following:

> *'Yeah, I was in physical pain. I was physically or physiologically more sedate. Internally my nervous system and respiratory system were really, horribly affected by that [Excel]. The first visual of all those numbers and just them being so scattered without seeing the correlation of them at the very beginning, that was overwhelming and frustrating. Having to navigate the data and analytic tools was a bit, major nuisance. And I found that those that gather overall like, eventually I got into a bit of rhythm but like immediate accessibility was almost non-existent and yeah I like it actually just over-accentuated a lot of my own, like responses when I felt like I was kind of shutdown.'*

In contrast the participant found Correlation Visualisation Toolkit far more accessible. The accessibility can be ascribed to the pleasant and simplistic interaction, Correlation Visualisation Toolkit has. The participant felt that the Correlation Visualisation Toolkit was a convenient solution for someone with little or no coding experience. P05 says the following about Correlation Visualisation Toolkit:

> *'Just being able to kind of actually play around with this and work through it without feeling like it's overwhelming. Yeah, it's like just from a simple piece of not someone who deals with coding in general. And in my work, I don't have to go through coding even though I should. It was the very least okay to play around and go and now I am trying to understand what it is. So I think the tool is relatively accessible. So long as you are given the time to just play around and just understand it a little bit.'*

6.4 Limitations

Firstly, the User Study for the Correlation Visualisation Toolkit was conducted within a narrow demographic. All the participants were within the age group of 19 years and 34 years with average age of 25 years. Also, all the participants students and workers at SFU. Therefore generalising the results beyond the given demographic can be moot. Furthermore, since the participants are not clinicians, some of the results and suggested improvements

need to be approached with wariness. These improvements may not be a reflection of the needs of the clinicians. Further consultations with clinicians and critical analysis of the suggested improvements and features may help us identify desirable changes. Secondly, a confirmation bias [46] could possibly be introduced since I validated the tool I designed . As I conducted all studies and interviews, participant responses may also be affected by response bias [31].

Excel was chosen as the benchmark for comparison purely based on the popularity of the tool, both among clinicians and in society. However, we cannot eliminate the possibility of other tools that perform better than the Correlation Visualisation Toolkit for correlation analysis. We also understand that the quality of training instructions provided for each tool could also determine the performance of the participants on each tool. While, all the participants unanimously agreed that the training instructions were not a deciding factor on their performance, we are sceptical about ignoring the possible confound.

Finally, the synthetic dataset and tasks given to the users during the study tried to emulate scenarios that may occur during the actual deployment of the tool. However, it is not possible to eliminate possibility of unprecedented scenarios and outcomes while the toolkit is used. Real world tasks and real data may be significantly different from what we had envisioned while designing the study.

Chapter 7

Conclusions

The Correlation Visualisation Toolkit provides a convenient platform for the clinicians in analysing the sleep-related data for the correlation between various sleep-related variables. Need for such a system that aids identification correlations between sleep variables was recognised both from existing scientific literature and informal discussions with the clinicians, who are the primary users of the system. The work is especially important for clinicians who operate under high level of work related stress and time constraints [66]. Our work is centred around a prototype built for the purpose. We analyse the ability of the tool to mitigate difficulties in correlation analysis and the overall usability of the tool.

The design of the visualisation achieved by a comparative review of the possible visualisation techniques presented in existing scientific literature. We evaluated the merits and demerits of each visualisation with respect to the requirements posed by the specific application of the tool and the users of the tool. Development of the proposed system was done by the assumption of a user model for the application; the user is presumed to be a practising clinician with minimal knowledge of statistics. The categorical arrangement of the variables is designed to be closely related to the original dashboard of the SWAPP web application. Drag and drop scheme for control of the visualised variables provide users with a convenient and intuitive method of manipulating the variables of interest.

The principle of operation of the interaction and the visualisation may be summarised as a two-step process:

1. A primary visualisation aims at fast, accurate filtering of the relevant variables based on their correlations, achieved by the heat map.

2. A secondary visualisation that allows exploratory data analysis of two variables of interest, by means of a scatter plot.

The heatmap provides the user with a quick summary of the variable correlations. The proposed system carefully integrates the use of significance and effect size parameters into the heatmap, as both were found to be relevant to the interpretation of data. Once the variables of interest have been identified, the user can initiate the scatter plot by clicking

on the respective circles corresponding the set of variables. The scatter plot provides a powerful tool to gain a deep understanding of the relationship between the variables of the data for users with necessary data analytic skills, knowledge and experience. We believe that the design knowledge resulting from our research may be applied to other data analysis problems both inside and outside health informatics while dealing with correlations.

We also designed a novel 'in-pandemic' mixed method user study over Zoom, to perform a comparative analysis of the Correlation Visualisation Toolkit against Microsoft Excel. The study validates that the Correlation Visualisation Toolkit is far more efficient in terms of Overall Workload and task completion time compared to that of Excel. This was validated from both qualitative and quantitative data collected during the study. The qualitative data also provided insights into the reasons why the participants thought the Correlation Visualisation Toolkit performed better. Finally, the qualitative data also provided information on areas to improve the Correlation Visualisation Toolkit.

7.1 Future Work

The future steps for the tool apparently includes improving the tool based on feedback solicited during the user-study. Also, more features need to be integrated into the tool before the tool is ready for deployment. One of the things that are of interest to us is means for analysis of temporal correlations. Finally, a more scalable and diverse user-study needs to be conducted on the improved Correlation Visualisation Toolkit.

Bibliography

[1] Interpreting correlations. http://rpsychologist.com/d3/correlation/.

[2] Interviews on zoom 5.0ethical considerations + best practices. https://www.sfu.ca/research/sites/default/files/2020-06/Infographic_Zoom%20v9.pdf.

[3] Jmp software. https://www.jmp.com/en_us/home.html.

[4] Microsoft excel. https://www.microsoft.com/en/microsoft-365/excel.

[5] Mockaroo. https://www.mockaroo.com/help/about.

[6] Parallel coordinates. https://en.wikipedia.org/wiki/Parallel_coordinates.

[7] Parallel sets. https://www.jasondavies.com/parallel-sets/, journal=Word Cloud Generator.

[8] Scatterplot matrix brushing. https://bl.ocks.org/mbostock/4063663.

[9] Spss software. https://www.ibm.com/analytics/spss-statistics-software.

[10] Spss tutorials: Paired samples t test.

[11] Tableau software. https://www.tableau.com/.

[12] Video conferencing, web conferencing, webinars, screen sharing. https://zoom.us/security.

[13] What does the dendrogram show, or what is correlation analysis? http://www.nonlinear.com/support/progenesis/comet/faq/v2.0/dendrogram.aspx.

[14] What is the usa patriot web. https://www.justice.gov/archive/ll/highlights.htm.

[15] *Sleep Disorders and Sleep Deprivation*. National Academies Press, Washington, D.C., sep 2006.

[16] *Q-Q Plot (Quantile to Quantile Plot)*, pages 437–439. Springer New York, New York, NY, 2008.

[17] How to interpret correspondence analysis plots (it probably isn't the way you think). https://www.displayr.com/interpret-correspondence-analysis-plots-probably-isnt-way-think/, Aug 2018.

[18] Moonplots: A better visualization for brand maps. https://www.displayr.com/correspondence-analysis-moonplots-a-better-visualization-for-brand-mapping/, Nov 2020.

[19] B. Alsallakh, W. Aigner, S. Miksch, and M. E. Gröller. Reinventing the contingency wheel: Scalable visual analytics of large categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2849–2858, Dec 2012.

[20] Jessica M. Blaxton, Cindy S. Bergeman, Brenda R. Whitehead, Marcia E. Braun, and Jessic D. Payne. Relationships among nightly sleep quality, daily stress, and daily affect. *The Journals of Gerontology: Series B*, 72(3):363–372, 2017.

[21] S Blunden, K Lushington, B Lorenzen, T Ooi, F Fung, and D Kennedy. Are sleep problems under-recognised in general practice? *Archives of disease in childhood*, 89(8):708–12, aug 2004.

[22] S Blunden, K Lushington, B Lorenzen, T Ooi, F Fung, and D Kennedy. Are sleep problems under-recognised in general practice? *Archives of Disease in Childhood*, 89(8):708–712, aug 2004.

[23] Tim Bock. Improving the display of correspondence analysis using moon plots. *International Journal of Market Research*, 53(3):307–326, 2011.

[24] R D Chervin, K H Archbold, P Panahi, and K J Pituch. Sleep problems seldom addressed at two general pediatric clinics. *Pediatrics*, 107(6):1375–80, jun 2001.

[25] R D Chervin, K H Archbold, P Panahi, and K J Pituch. Sleep problems seldom addressed at two general pediatric clinics. *Pediatrics*, 107(6):1375–80, jun 2001.

[26] J.W. Creswell and V.L.P. Clark. *Designing and Conducting Mixed Methods Research*. SAGE Publications, 2011.

[27] Tirosh E, Sadeh A, Munvez R, and Lavie P. Effects of methylphenidate on sleep in children with attention-deficit hyperactivity disorder: An activity monitor study. *American Journal of Diseases of Children*, 147(12):1313–1315, 1993.

[28] N. Elmqvist, P. Dragicevic, and J. Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1539–1148, Nov 2008.

[29] Gahan Fallone, Judith A Owens, and Jennifer Deane. Sleepiness in children and adolescents: clinical implications. *Sleep medicine reviews*, 6(4):287–306, aug 2002.

[30] Mary Jo Foley. Microsoft office 365 now has 120 million business users, Oct 2017.

[31] Adrian Furnham. Response bias, social desirability and dissimulation. *Personality and Individual Differences*, 7(3):385–400, 1986.

[32] Alice M Gregory, Avshalom Caspi, Terrie E Moffitt, and Richie Poulton. Sleep problems in childhood predict neuropsychological functioning in adolescence. *Pediatrics*, 123(4):1171–6, apr 2009.

[33] Bartholomeus C.M. (Benno) Haarman, Rixt F. Riemersma-Van der Lek, Willem A. Nolen, R. Mendes, Hemmo A. Drexhage, and Huibert Burger. Feature-expression heat maps – a new visual method to explore complex associations between two variable sets. *Journal of Biomedical Informatics*, 53:156 – 161, 2015.

[34] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139 – 183. North-Holland, 1988.

[35] H. Hauser, F. Ledermann, and H. Doleisch. Angular brushing of extended parallel coordinates. In *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.*, pages 127–130, Oct 2002.

[36] Donna L. Hoffman and George R. Franke. Correspondence analysis: Graphical representation of categorical data in marketing research. *Journal of Marketing Research*, 23(3):213–227, 1986.

[37] Hans A. Kestler, André Müller, Thomas M. Gress, and Malte Buchholz. Generalized venn diagrams: a new method of visualizing complex genetic set relations. *Bioinformatics*, 21(8):1592–1595, 2005.

[38] R. Kosara, F. Bendix, and H. Hauser. Parallel sets: interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568, July 2006.

[39] John V. Lavigne, Richard Arend, Diane Rosenbaum, Andy Smith, Marc Weissbluth, Helen J. Binns, and Katherine Kaufer Christoffel. Sleep and behavior problems among preschoolers. *Journal of Developmental and Behavioral Pediatrics*, 20(3):164–169, 1 1999.

[40] Taylor Lorenz and Davey Alba. 'zoombombing' becomes a dangerous organized effort. `https://www.nytimes.com/2020/04/03/technology/zoom-harassment-abuse-racism-fbi-warning.html`, Apr 2020.

[41] Albert Madansky. *Testing for Normality*, pages 14–55. Springer New York, New York, NY, 1988.

[42] Lisa J Meltzer, Courtney Johnson, Jonathan Crosette, Mark Ramos, and Jodi A Mindell. Prevalence of diagnosed sleep disorders in pediatric primary care practices. *Pediatrics*, 125(6):e1410–8, jun 2010.

[43] Jodi A. Mindell and Judith A. Owens. *A clinical guide to pediatric sleep : diagnosis and management of sleep problems.* Lippincott Williams & Wilkins, 2003.

[44] Jaime M. Monti. Serotonin control of sleep-wake behavior. *Sleep Medicine Reviews*, 15(4):269 – 281, 2011.

[45] David S Moore. *Chapter 4, The basic practice of statistics / David S. Moore, William I. Notz, Michael A. Fligner.* 7th ed. edition, 2015.

[46] Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998.

[47] Jakob Nielsen. Usability inspection methods. chapter Heuristic Evaluation, pages 25–62. John Wiley & Sons, Inc., New York, NY, USA, 1994.

[48] Shalini Paruthi, Lee J. Brooks, Carolyn D'Ambrosio, Wendy A. Hall, Suresh Kotagal, Robin M. Lloyd, Beth A. Malow, Kiran Maski, Cynthia Nichols, Stuart F. Quan, Carol L. Rosen, Matthew M. Troester, and Merrill S. Wise. Recommended amount of sleep for pediatric populations: A consensus statement of the American Academy of Sleep Medicine. *Journal of Clinical Sleep Medicine*, 2016.

[49] Michael J. Peterson and Ruth M. Benca. Chapter 130 - mood disorders. In Meir H. Kryger, Thomas Roth, and William C. Dement, editors, *Principles and Practice of Sleep Medicine (Fifth Edition)*, pages 1488 – 1500. W.B. Saunders, Philadelphia, fifth edition edition, 2011.

[50] H. Piringer, R. Kosara, and H. Hauser. Interactive focus+context visualization with linked 2d/3d scatterplots. In *Proceedings. Second International Conference on Coordinated and Multiple Views in Exploratory Visualization, 2004.*, pages 49–60, July 2004.

[51] J. Quach, H. Hiscock, L. Canterford, and M. Wake. Outcomes of Child Sleep Problems Over the School-Transition Period: Australian Population Longitudinal Study. *PEDIATRICS*, 123(5):1287–1292, may 2009.

[52] Aaron J. Quigley and Peter Eades. Proveda : A scheme for progressive visualization and exploratory data analysis of clusters . work in progress report. 1999.

[53] Amanda Ross and Victor L. Willson. *Paired Samples T-Test*, pages 17–19. SensePublishers, Rotterdam, 2017.

[54] Jacques Rougemont and Pascal Hingamp. Dna microarray data and contextual analysis of correlation graphs. *BMC Bioinformatics*, 4(1):15, Apr 2003.

[55] N. Sauber, H. Theisel, and H. Seidel. Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):917–924, Sept 2006.

[56] Muhaafidz Saufi, Zainura Idrus, Sharifah Aliman, and Nur Atiqah Sia Abdullah. *Clutter-Reduction Technique of Parallel Coordinates Plot for Photovoltaic Solar Data: 4th International Conference, SCDS 2018, Bangkok, Thailand, August 15-16, 2018, Proceedings*, pages 337–349. 01 2019.

[57] Deepak Shrivastava, Syung Jung, Mohsen Saadat, Roopa Sirohi, and Keri Crewson. How to interpret the results of a sleep study. *Journal of community hospital internal medicine perspectives*, 4(5):24983, 2014.

[58] H Smedje, J E Broman, and J Hetta. Parents' reports of disturbed sleep in 5-7-year-old Swedish children. *Acta paediatrica (Oslo, Norway : 1992)*, 88(8):858–65, aug 1999.

[59] MARCEL G. SMITS, HENK F. van STEL, KRISTIAAN van der HEIJDEN, ANNE MARIE. MEIJER, ANTON M.L. COENEN, and GERARD A. KERKHOF. Melatonin improves health status and sleep in children with idiopathic chronic sleep-onset insomnia: A randomized placebo-controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, 42(11):1286 – 1293, 2003.

[60] Robert Spence. *Representation*, pages 41–110. Springer International Publishing, Cham, 2014.

[61] M A Stein, J Mendelsohn, W H Obermeyer, J Amromin, and R Benca. Sleep and behavior problems in school-aged children. *Pediatrics*, 107(4):E60, apr 2001.

[62] Mark A. Stein, Janis Mendelsohn, William H. Obermeyer, Julie Amromin, and Ruth Benca. Sleep and behavior problems in school-aged children. *Pediatrics*, 107(4):e60–e60, 2001.

[63] Cajo J. F. Ter Braak. The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, 69(1):69–77, Apr 1987.

[64] Amal Vincent, Ankit Gupta, Ruoyu Li, Chris Shaw, and Saba Akhyani. Data acquisition and visual analytic tool-set for paediatric sleep data. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth'19, page 320–326, New York, NY, USA, 2019. Association for Computing Machinery.

[65] Amal Vincent, Ankit Gupta, Chris Shaw, and Ruoyu Li. Correlation visualisation for sleep data analytics in swapp (sleep wake application). *Electronic Imaging*, 2019(1):682–1, 2019.

[66] Kathryn Wilkins. Work stress among health care providers. *Health Reports*, 18(4):82–003, 2007.

[67] Merril S Wise and Daniel G Glaze. Assessment of sleep disorders in children, 2017.

[68] Stefan Zapf and Christopher Kraushaar. A new visualization to beautifully explore correlations. `https://www.oreilly.com/learning/a-new-visualization-to-beautifully-explore-correlations`, Jan 2017.

Appendix A

Study instruments and training document

Please find the documents in the subsequent pages.

A.1 Pre-Study Questionnaire

SFU Ethics Application **2019s0220** Screening Questionnaire, version 2

School of Interactive Arts + Technology (SIAT), Simon Fraser University Surrey

250 -13450 102 Avenue, Surrey, BC V3T 0A3 Canada

Tel: +1 778 782 8013 Web: **https://www.swapp.iat.sfu.ca/**

Screening Questionnaire

We will use this questionnaire to determine your eligibility for the Correlation Visualisation Toolkit Study.

Name: _____

Email: _____

What is your age?

- 0 – 18 years old
- 19 – 29 years old
- 30 – 49 years old
- 50 – 64 years old
- 65 years and over

What is your Occupation (if a student please mention your discipline as well)?

_____

Are you suffering from any disability or any condition that affects your vision?

- Yes
- No

Are you suffering from any disability or any condition that prevents you from using a desktop computer, that has a mouse and keyboard?

- Yes
- No

How many hours a day do you spend in front of a laptop or desktop computer (not including mobile phone)?

- Less than 1 hour
- Between 1 and 3 hours
- More than 3 hours

Have you used Excel, SPSS or Tableau or similar tools for data analysis?

- Yes
- No

A.2 Training Document

Correlation analysis

In this experiment we will be using the Pearson Correlation Coefficient or the **Effect Size(r)** along with the **significance value(p-value).** The inferences from the two values are shown below.

P (between 0 and 1)	**Inference**
0.1 > p	Very weak significance
0.05 < p < 0.1	Weak significance
p < 0.05	Strong significance

r (between -1 to 1)	**Inference**
0.1 < | r | < 0.3	Small effect
0.3 < | r | < 0.5	Medium size effect
| r | > 0.5	Large effect

Positive **r** between two variables means that when one variable increases other decreases. Negative **r** means that when one variable increases other decreases.

You should be looking for strong significance and medium/large effect size, which means you need to analyse both these values before making a conclusion.

Correlation analysis - Excel training

Finding the p value (compiled from https://www.excel-easy.com/examples/t-test.html)

For the sake of this experiment we will be using the output of the t-Test as the p-value.

Below you can find the study hours of 6 female students and 5 male students.

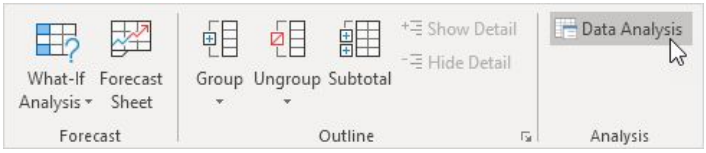
H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 \neq 0$

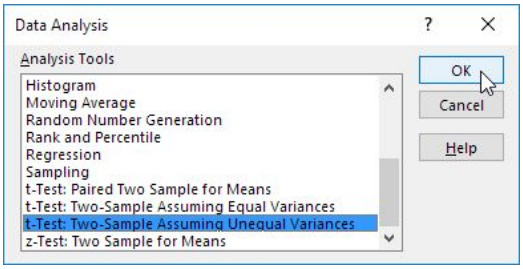
	A	B	C
1	Female	Male	
2	26	23	
3	25	30	
4	43	18	
5	34	25	
6	18	28	
7	52		
8			

In the datasets you have been provided the variances of the two populations are not equal. So you should do a T-test with unequal variances.

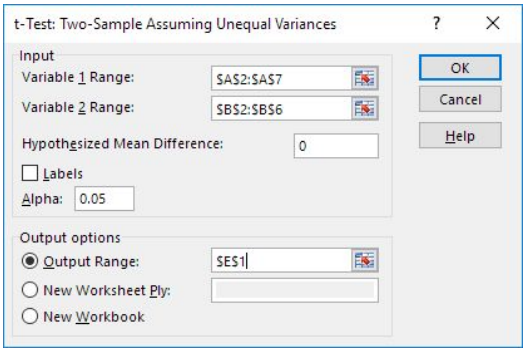
1. On the Data tab, in the Analysis group, click Data Analysis.



2. Select t-Test: Two-Sample Assuming Unequal Variances and click OK.



3. Click in the Variable 1 Range box and select the range A2:A7.
4. Click in the Variable 2 Range box and select the range B2:B6.
5. Click in the Hypothesized Mean Difference box and type 0 (H_0: $\mu_1 - \mu_2 = 0$).
6. Click in the Output Range box and select cell E1. Click ok.

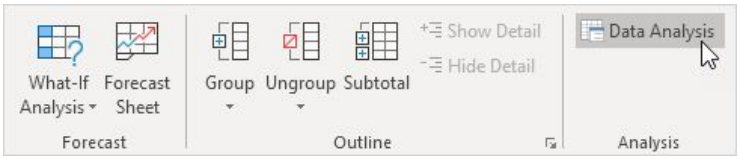


7. Result

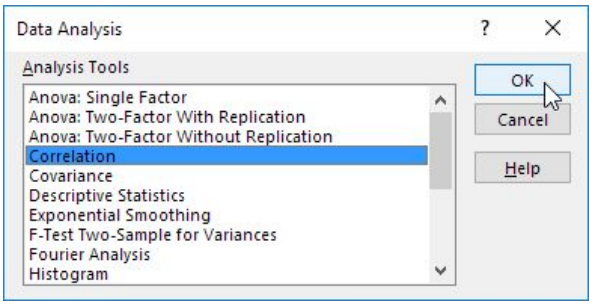
E	F	G
t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	*Variable 2*
Mean	33	24.8
Variance	160	21.7
Observations	6	5
Hypothesized Mean Difference	0	
df	7	
t Stat	1.47260514	
P(T<=t) one-tail	0.092170202	
t Critical one-tail	1.894578605	
P(T<=t) two-tail	0.184340405	
t Critical two-tail	2.364624252	

Finding r in Excel

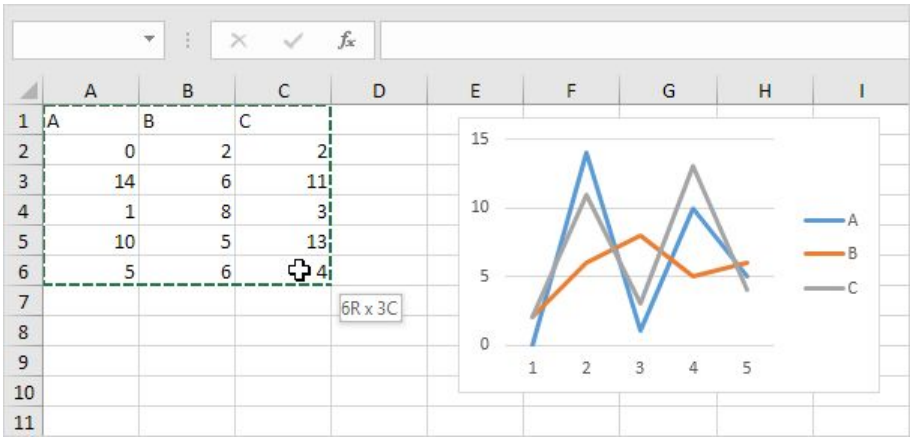
1. On the Data tab, in the Analysis group, click Data Analysis.



2. Select Correlation and click OK.

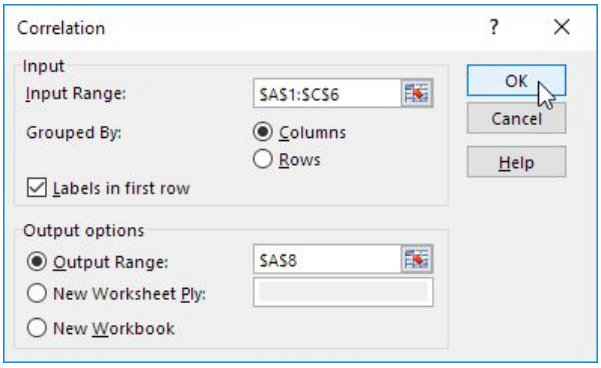


3. For example, select the range A1:C6 as the Input Range.

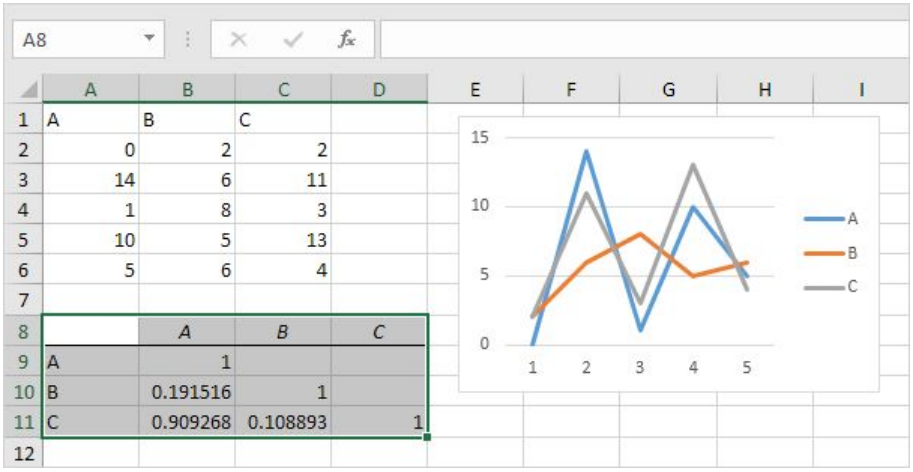


4. Check Labels in the first row.

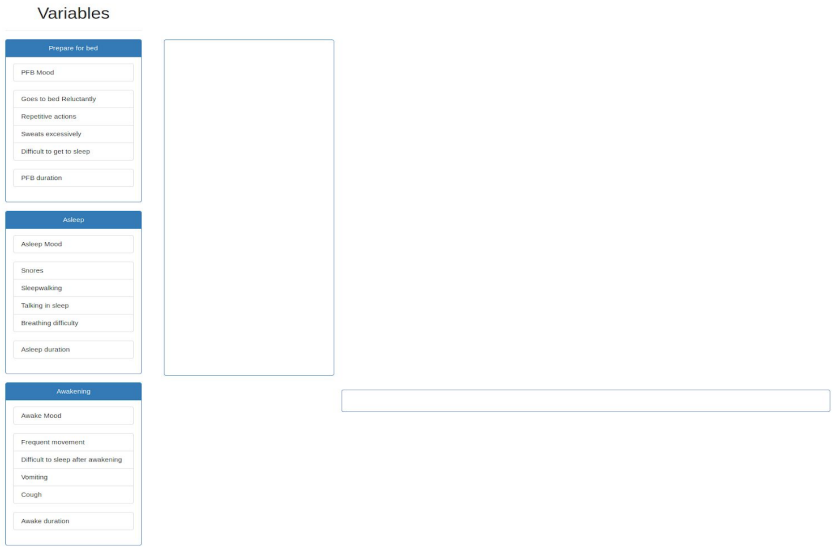
5. Select cell A8 as the Output Range. Click ok.



6. Result



Correlation analysis - Using Viz tool



The available sleep-related variables/parameters/events for visualisation are stacked on the left-hand side of the visualisation in categorical boxes.

The visualisation can be initialised by dragging and dropping the variables from the categorical boxes.

Interpretation of the visualisation:

- You are trying to look for significance (**p-value)** < 0.05 and correlation coefficient (**r)** greater than 0.3. So look for colour of circles in the visualisation close to red or blue.
- Red circle indicates correlation close to one and Blue correlation indicates correlation close to negative one. **The scale on the right of the visualisation will help you interpret the data.**
- If the radius of the variable is not high enough it means that the significance of the correlation between the two variables.
- Ignore circles which are transparent or close to being transparent.

Further details on the correlation matrix visualisation:

- Colour: correlation coefficient/effect size (pearson/spearman's coefficient, r-value)
- Radius: Significance, p-value
- Transparency: (1-p) x |r|

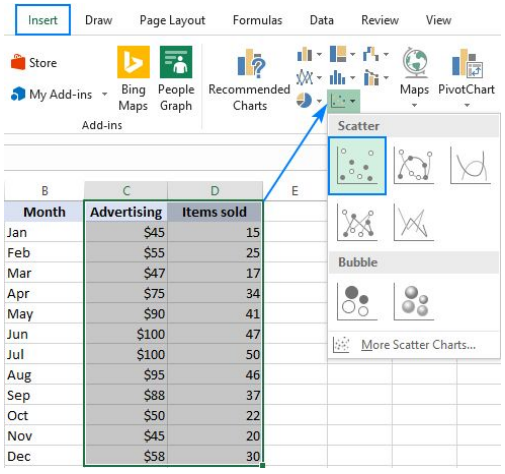
Scatter plot using - viz tool

Click on top of the circle to generate the scatter plot, since we have a lot of discrete variables random noise (jitter) is added to the numerical data value to see all points separately.

Scatter Plot on Excel

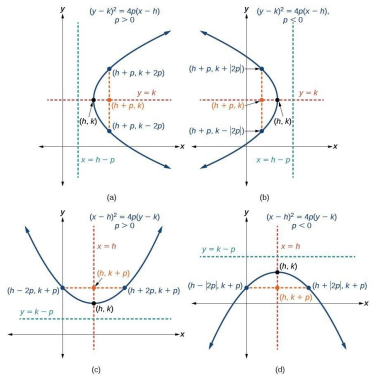
With the source data correctly organized, making a scatter plot in Excel takes these two quick steps:

1. Select two columns with numeric data, including the column headers. In our case, it is the range C1:D13. Do not select any other columns to avoid confusing Excel.
2. Go to the Inset tab > Chats group, click the Scatter chart icon, and select the desired template. To insert a classic scatter graph, click the first thumbnail:

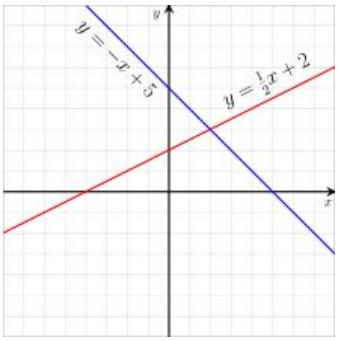


Types of functions:

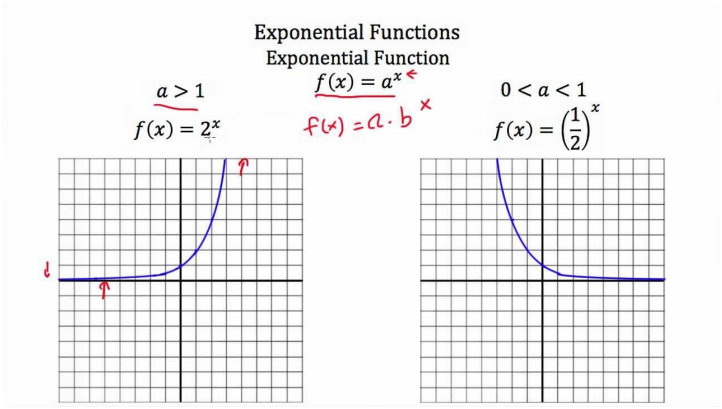
Parabola:



Linear:



Exponential:



A.3 NASA-TLX Questionnaire

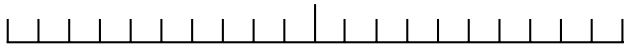
SFU

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

ID	Task	Date

Mental Demand — How mentally demanding was the task?

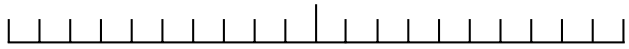


Very Low — Very High

Physical Demand — How physically demanding was the task?

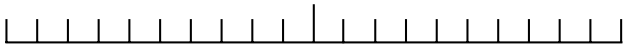
Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?



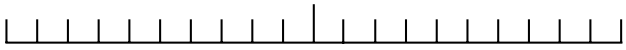
Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?



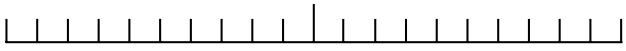
Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?



Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?



Very Low — Very High

A.4 Semi Structured Interview Guide

SFU Ethics Application **2019s0220** Interview script

School of Interactive Arts + Technology (SIAT), Simon Fraser University Surrey

250 -13450 102 Avenue, Surrey, BC V3T 0A3 Canada

Tel: +1 778 782 8013 Web: https://www.swapp.iat.sfu.ca

SFU

Correlation Visualisation Toolkit Study - Interview Script

Review of experience using data analysis tools (based on pre-study questionnaire)

1. You have indicated that you have experience using data analysis tools, did you happen to work with correlation analysis?
2. How long have you worked with data analysis tools?
3. What was the data analysis operations and principles that you are familiar with?

Experience using Correlation Visualisation Toolkit

1. How was your experience with the visualization?
2. What do you think of the colour scale for the Correlation Visualisation Toolkit?
3. Do you have any comments about the scatter plot option in the correlation visualisation toolkit, did you find it useful?
4. What were the things that you liked most about the toolkit?
5. Did you face any issues in using Correlation Visualisation Toolkit?
6. What were the things you did not like about the toolkit?
7. Do you have any suggestions or feedback for improving the correlation visualisation toolkit?

Comparing MS Excel with Correlation Visualisation Toolkit

1. What were the features you liked about correlation analysis on MS Excel compared to Correlation Visualisation Toolkit and vice versa?
2. How was the learning experience using both the tools?
3. Overall which tool would you prefer for correlation analysis of data and why?

Questions based on observation of the data analysis experiments

The following questions are framed based on the approaches/steps taken by the user to complete the given tasks. It is difficult to anticipate all possible scenarios or questions arising from those scenarios as there are infinitely many possibilities. However, a few examples are below.

1. I see that you spent a lot of time in choosing the highest correlated item in the analysis of “snoring’ VS all other sleep variables, what was it that confused you?
2. You concluded that the mood on sleeping is not correlated to time asleep, what was your rational behind this?

Appendix B

Tests for Normality

I ran tests for Normality on the difference between NASA-TLX Overall scores, (CorrViz_Overall) and Excel_Overall, I called the difference Difference_Overall. The Shapiro-Wilk test gave a significance of 0.692, thus we cannot reject the null hypothesis that the distribution is normal (see table B.1). The approximate normality is also apparent from the QQ-plots (see figure B.2).

Next I ran tests for Normality on the difference between total times, CorrViz_Time and Excel_Time, we called the difference Difference_Time. This was to meet the the assumptions of the paired t-test. Considering the sample size, we looked at the Shapiro-Wilk test. The test gave a significance of 0.217, thus we cannot reject the null hypothesis that the distribution is normal (see table B.2). The approximate normality is also apparent from the QQ-plots (see figure B.3).

	Kolmogorov-Smirnova			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Difference_Overall	.123	20	.200*	.967	20	.692

Table B.1: Test of Normality on Difference between NASA-TLX Overall Workloads for Correlation Visualisation Toolkit and Microsoft Excel

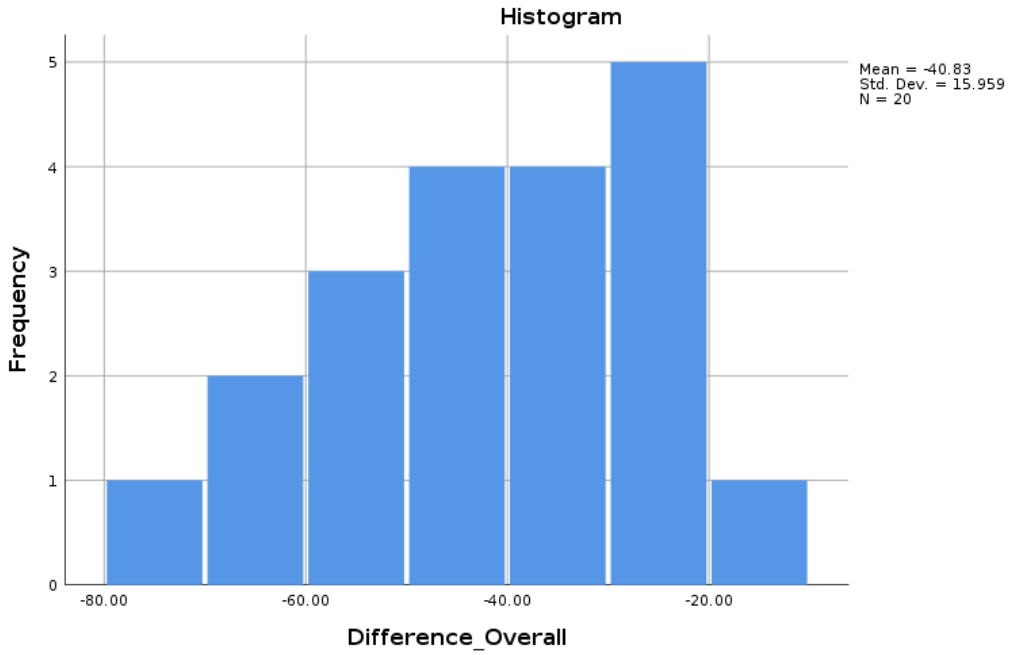


Figure B.1: Histogram of difference between NASA-TLX Overall Workload for Correlation Visualisation Toolkit and Microsoft Excel

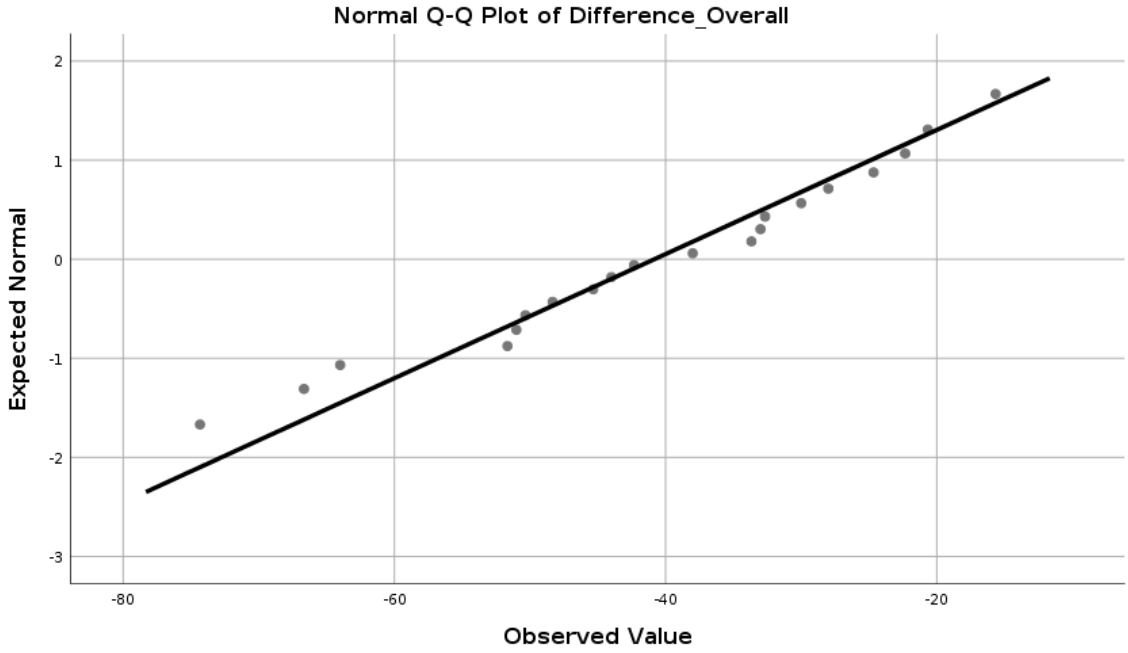


Figure B.2: QQ-Plot of difference between NASA-TLX Overall Workload for Correlation Visualisation Toolkit and Microsoft Excel

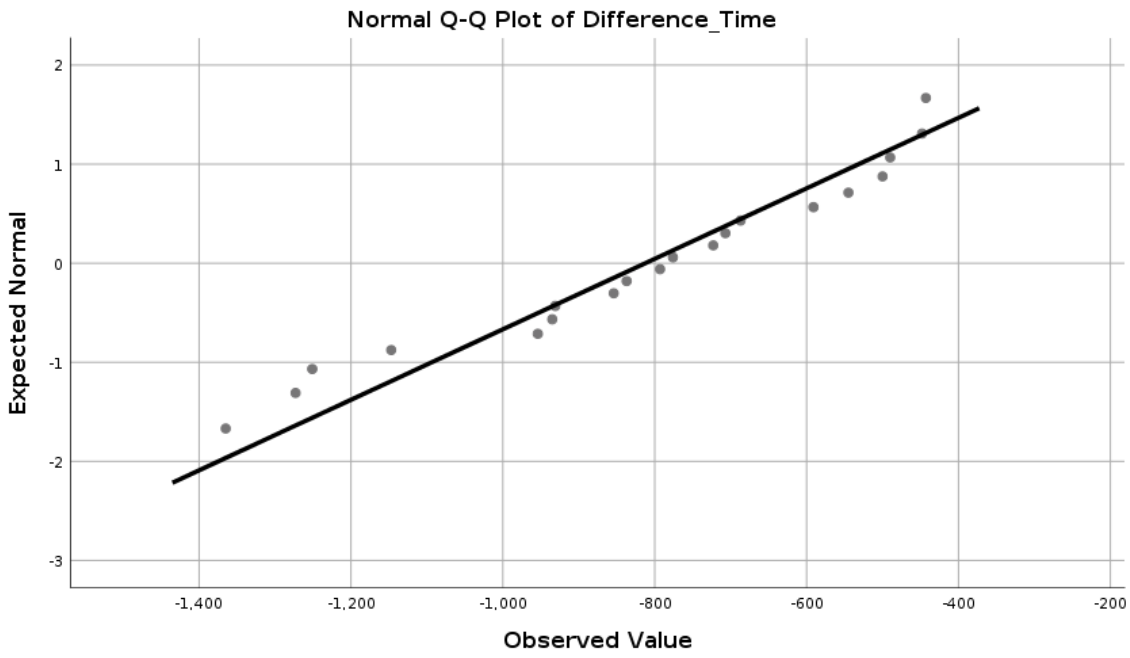


Figure B.3: QQ Plot of difference between Time Duration for Correlation Visualisation Toolkit and Microsoft Excel

	Kolmogorov-Smirnova			Shapiro-Wilk			
	Statistic	df	Sig.	Statistic	df	Sig.	
Difference_Time	.107	20	.200*	.938	20	.217	

Table B.2: Test of Normality on Difference between Time duration for Correlation Visualisation Toolkit and Microsoft Excel

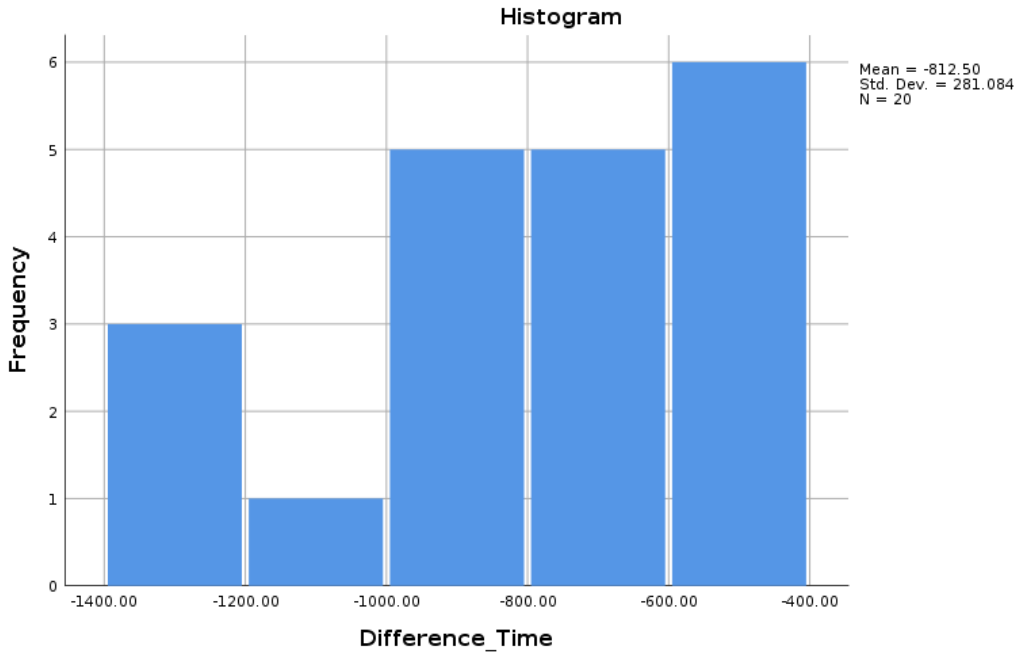


Figure B.4: Histogram of difference between Time Duration for Correlation Visualisation Toolkit and Microsoft Excel

Appendix C

Box Plots for outlier detection

For the results from the Paired t-Tests to hold, the difference between the variables should not have any significant outliers and the distribution should be approximately normal [10]. From the box-plots for the Difference_Overall and Difference_Time it is apparent that there are no outliers in either (see figures C.1 and C.2).

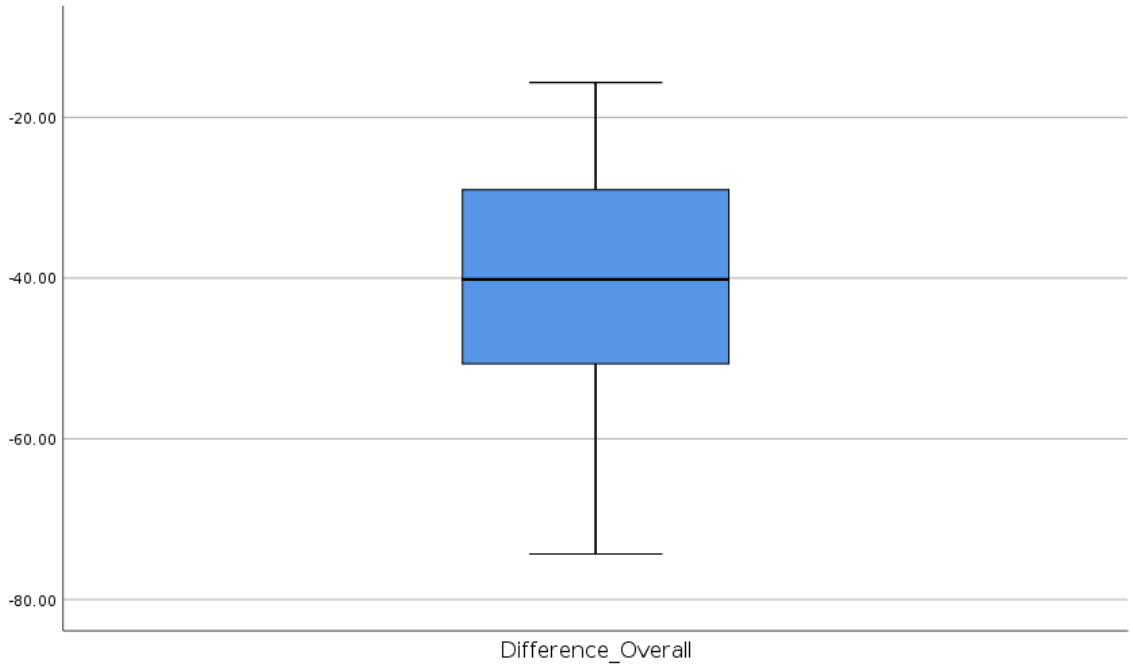


Figure C.1: Box Plot of difference between NASA-TLX Overall Workload for Correlation Visualisation Toolkit and Microsoft Excel

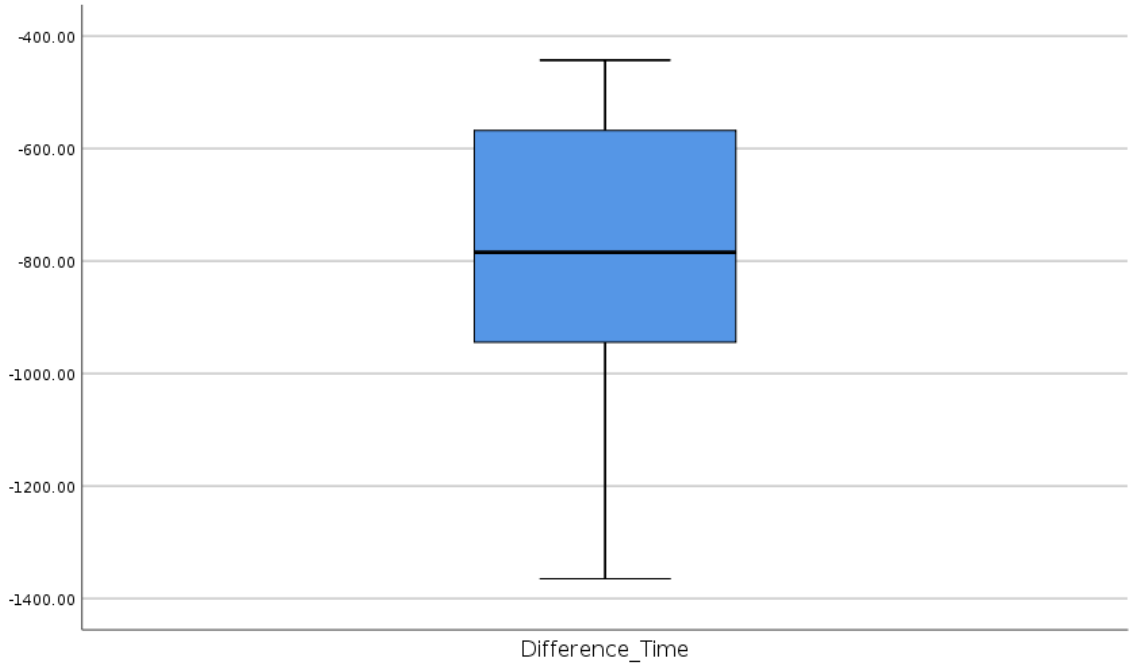


Figure C.2: Box Plot of difference between Time Duration for Correlation Visualisation Toolkit and Microsoft Excel