COMP 1800
Data Visualisation Report.

Student ID - 001002629

26^{th} March 2021

Contents

1 Introduction

Data Visualisation is a process of taking data or information and representing it in a graph or other format. This allows faster interpretation of data which would otherwise just be in raw formats which are difficult to understand and decipher to those who are not used to looking at raw data and information.

In 2014 a survey conducted on data visualisation concluded that there are four main categories *"ranging from network visualization and text visualization, to map visualization and multivariate data visualization"*(Liu et al. 2014) From this survey Chotisarn et al. (2020) builds on this classification and produce their own categorical list which are *"empirical methodologies, interactions, frameworks, and applications."*

Regardless of the naming convention that is used to categorise what type of visualisation is being produced, the end goal is the same, to represent data or information in a easy to read manner that allow insights that could have otherwise remained hidden. Not all information will suit a scatter graph or a 3d plot, so choosing what data to represent in what format is extremely important.

2 Visualisations

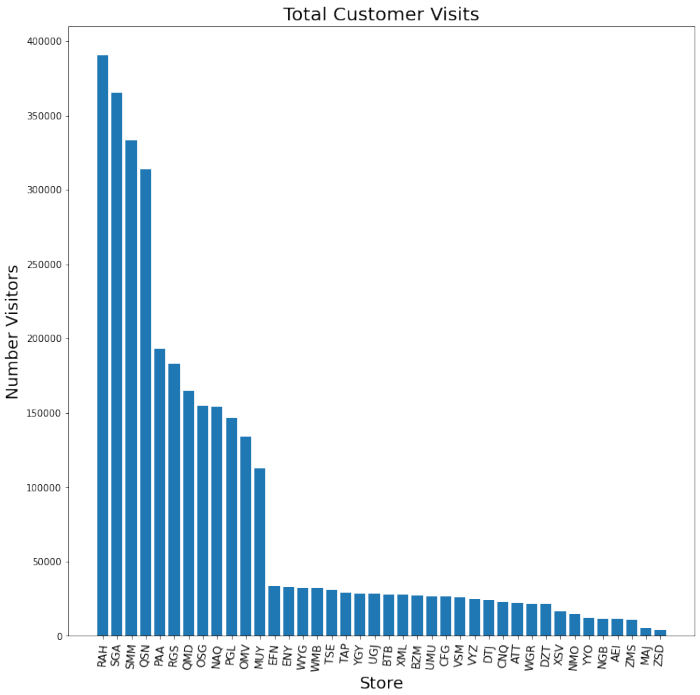
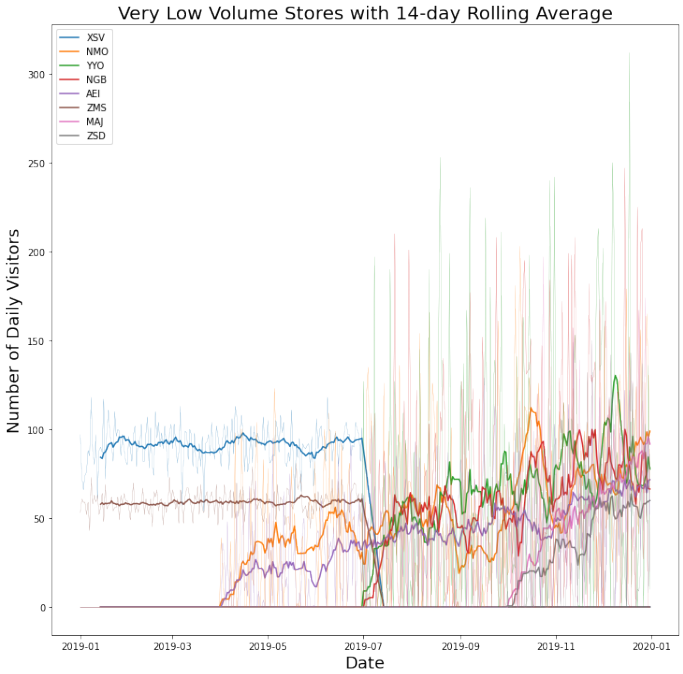


Figure 1: Total Visitors for all stores

Bar charts are simple graphs that are very good at displaying data that is grouped by category, with Figure 1 displaying all stores in the dataset along the x axis and the total number of visitors to the store set along the y axis. Although there is a lot of data contained in Figure 1 it is still simple to read and interpret without needing to create multiple smaller charts to get the same data.

As data that is used for Figure 1 has been ordered highest to lowest with respect to total customer visits we can see 3 clear groups emerging which could be classified as high, medium and low volume stores. With RAH, SGA, SMM, and QSN stores having the highest volume of visitors, with the next set of stores, PAA, RGS, QMD, OSG, NAQ, PGL, OMV and MUY being medium volume stores with the remaining stores being low volume stores.

Figure 2: Line Graph with 14 day rolling average



Interesting data revealed itself while exploring time series data for daily customer visits for the year 2019 using Line Plot graphs. These graphs show daily visitors represented with the faded line that has sharp peaks and troughs and a 14 day rolling average line represented by a solid smoother line that is the same colour as its daily counterpart and a legend at the top left of the graph that lets the reader know what line is representing what store using colour coordination. Axis for this graph show that this x axis as the date and the y axis as the total number of daily visitors over this period. With Figure 2 we can see that all the stores that have been classified as very low volume stores had either closed or opened during the year of 2019. XSV and ZMS are two stores out of eight that appear to have closed during this period with the rest of the stores opening during this period.

Figure 2 has been included as this shows very interesting data in regards to these eight stores that have been selected, a deeper insight then a bar graph can show is uncovered and shows possible reasons to why some stores have very low daily visitors. The six stores NMO, AEI, YYO, NGB MAJ and ZSD, that opened at different points over the year are all trending upwards with respect to daily customer visits, however trend lines were not included in this graph as it does not add anymore information as it would make reading this particular graph harder and with the rolling average line, we can reasonably discern an upward trend for the newly opened stores.

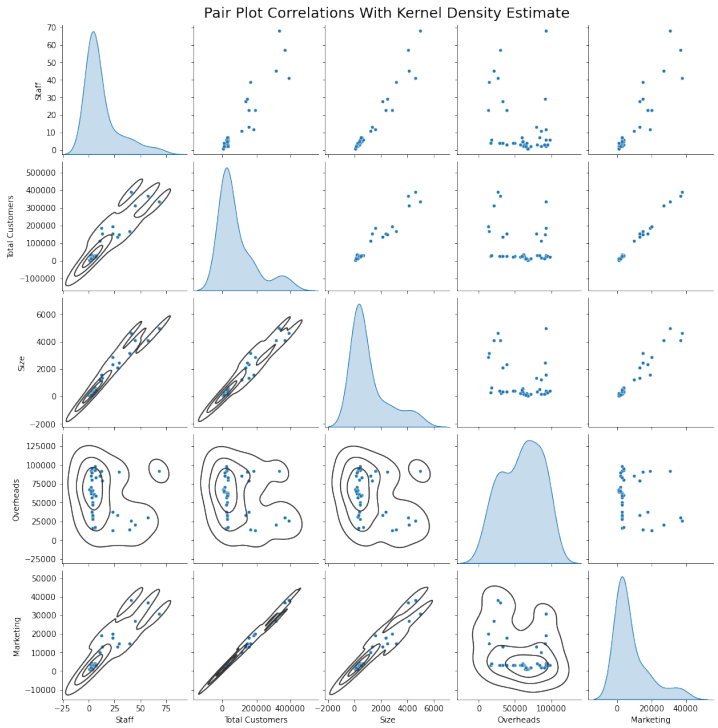
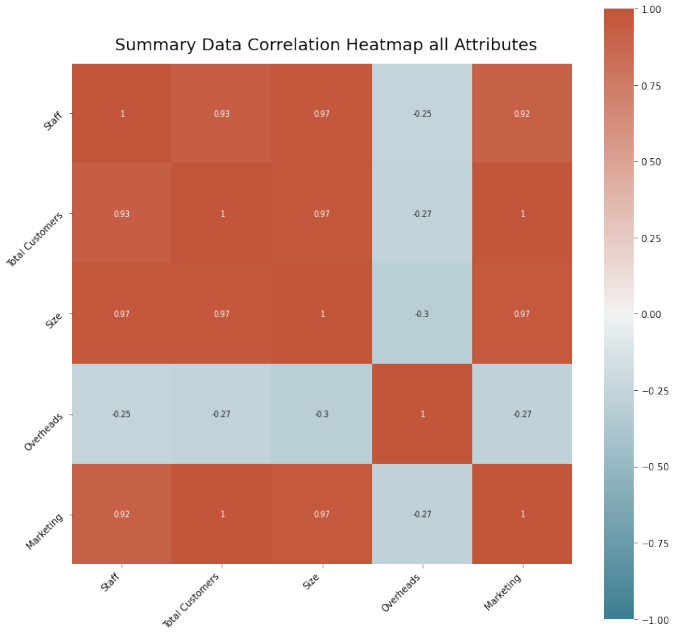


Figure 3: Pair Plot With Correlations, Histogram and Kernel Density Plot

Pair plot graphs are very useful for comparing two variables in the dataset and and visualising their relationship, Figure 3 has three different types of comparison in a single graph. Top right corner of this pair plot has correlations between each element, though the middle of the graph is a histogram showing single variable distributions and bottom left corner has a Kernel Density Estimate which helps to visualise the shape of the data. A strong correlation will show up as a close scatter graph that goes from bottom left of a graph to top right.

From Figure 3 there are strong correlations between all pairs of variables except for overheads, which from this graph does not have correlations between any other variable. From this graph there is extremely strong correlation between total customers and marketing as the scatter graph in the top right is tightly distributed along a rising diagonal line which can also been seen even more clearly with the Kernel Density Estimate shape for these variables. When observing the histogram for overheads, the distribution across this is very wide and does not show a single peak as the other variables do, which could mean that overheads for stores vary wildly and that the elements that have been used to generate this graph have little to no effect on the overhead.

Figure 4: Summary Data Heatmap all Attributes



Heatmap graphs show the relationship between variables again, but display the values with a colour gradient and a numeric value displayed. For this particular graph the colour gradient ranges from blue which represents no correlation, too red which indicates high correlations. The numeric value ranges from -1.0 to 1.0 with the negative being no or little correlation and positive being high correlations.

This heatmap in Figure 4 confirms what had been indicated in the scatter charts from Figure 3, the overheads for each store have no or very little correlation between any other element that has been provided in the data. The least correlated elements in Figure 4 are Overheads and Size with a value of -0.3 which would be an oddity of shop floor space as one would assume that larger stores would have higher overheads due to running costs being higher, but without data to support this line of thinking it is pure conjecture. Total Customers and Marketing have the highest correlations with a value of 1.0 which does make sense as spending money on marketing of a store should correlate highly with the number of customers that have knowledge of what kind of shop it is.

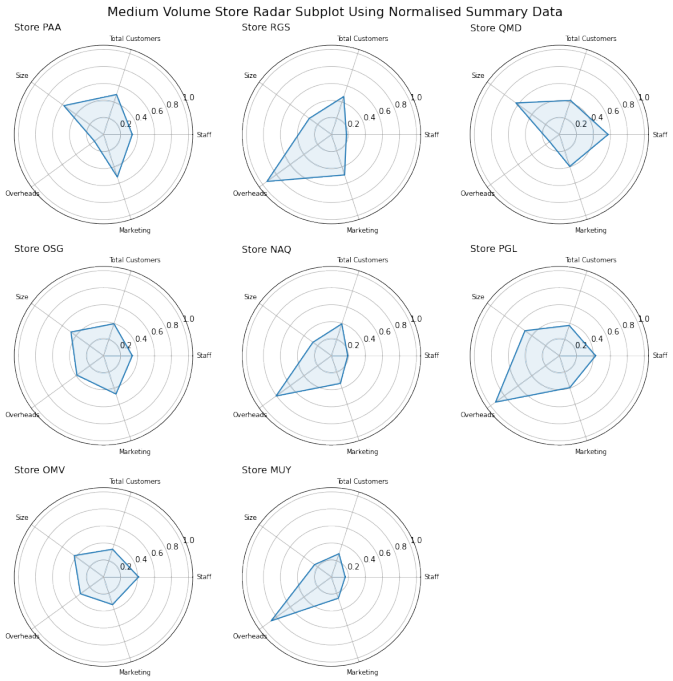


Figure 5: Medium Sized Store Radar Plot With Normalised Data

As the data that has been provided is multivariate a radar plot is a good way to visualise each stores data. For Figure 5 five variables are shown for eight stores which have been identified as the medium volume visitors. The scale is from 0.0 to 1.0 as this data has been normalised which is where the scale of data is reduced using $X' = \frac{x - x_{min}}{x_{max} - x_{min}}$ with x being the dataset in this case. This scale of reduction is helpful as the number of staff would be lost in a radar plot against the total customers values, with these particular stores having a total visitor count between 50,000 and 200,000.

Figure 5 attributes have been arranged so values that are assumed to be important are closer to the top of the radar and the attributes that we assume the company wants to be as small as possible pointing to the bottom of the radar. This allows radars plot to be interpreted quickly as large points at the top are see as favourable attributes with stores PAA and RGS having the highest total customers with regards to the medium volume stores, disadvantageous variables such as overheads and marketing are points would in a perfect would be minimal in size for Figure 5. Using Stores PAA and RGS we can see that the overheads for RGS are much higher then for PAA, this could be due to numerous reasons but the dataset that has been provided does not allude to the reasons why. What this radar plot does show is that four out of eight stores RGS, NAQ, PGL and MUY have very high overheads in comparison to the size and total customers while PAA and QMD stores have extremely low overheads.

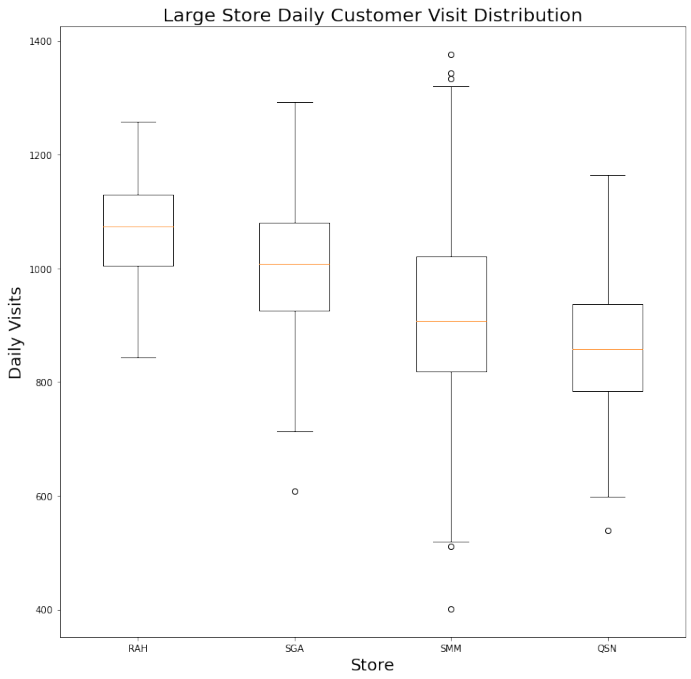


Figure 6: Large Sized Store Box Plot for Daily Customer Visits

Box plots a seen in Figure 6 have six components to them, these are shown from bottom to top, minimum, first quartile(Q1), median values, third quartile(Q3), maximum and outliers represented by small circles. The minimum and the maximum values are not true maximum values, they are derived by using an interquartile range (IQR) of the 25^{th} to 75^{th}, so minimum value would be $Q1 - 1.5 * IQR$ and the maximum $Q3 + 1.5 * IQR$. Box plots allow visualisations of distributions, and for this case high volume store customer visits along with any outliers hidden in the data.

Considering Figure 6 we can see that SGA and QSN both have a single outlier that falls below the derived minimum value where SMM has two outliers that fall below the minimum and three outliers that fall above the derived maximum value. Store SMM also has very long tails which show the distribution of daily visits range from a value of around 550 too 1300 visits per day. Store RAH however has no outliers and comparatively short tails which alludes to this particular store having a steady daily customer visit frequency where SMM has the opposite and a frequency that could vary massively on a day to day basis, this could be due to location but again without that data it is unknown and would require more research into the reasons.

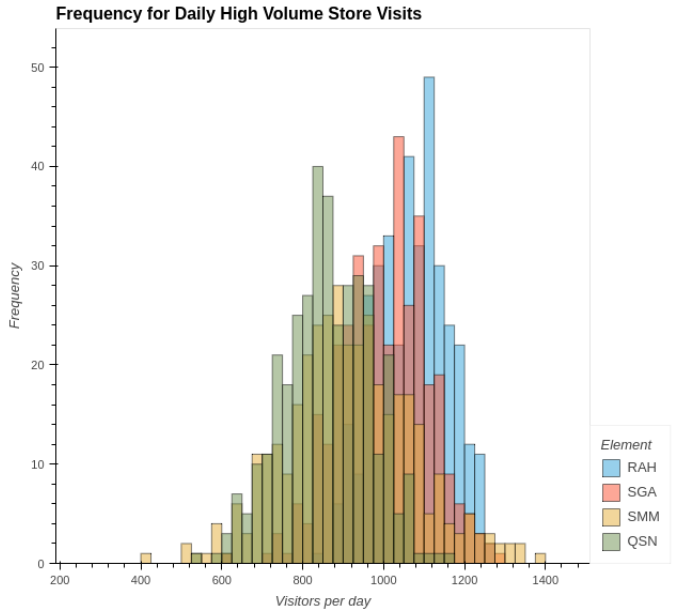


Figure 7: High Volume Store Frequency Graph

Figure 7 builds from data in Figure 6 and shows the frequency for visits per day to each high volume store in a bar graph format. As this bar chart overlays elements on the chart it is not a good candidate to visualise data where there is a large set of data to visualise. However as Figure 7 only has four elements to display it can convey information quickly. With the x axis showing the number of visitors per day and the y axis showing the frequency we can see the number of times one of these stores has had a particular number of visits.

As mentioned previously store SMM has a larger range of visitors to it when compared to RAH, SGA or QSN, we can see the outliers that have been identified from Figure 6 but with a value that is attached to the number of visits, RAH has a tight cluster of bars, displayed in blue, towards the higher end of visits and also the highest frequency of visits alluding to this store being very popular with loyal customers who return on a regular basis. Again this is conjecture without relevant data to further explore this avenue further.

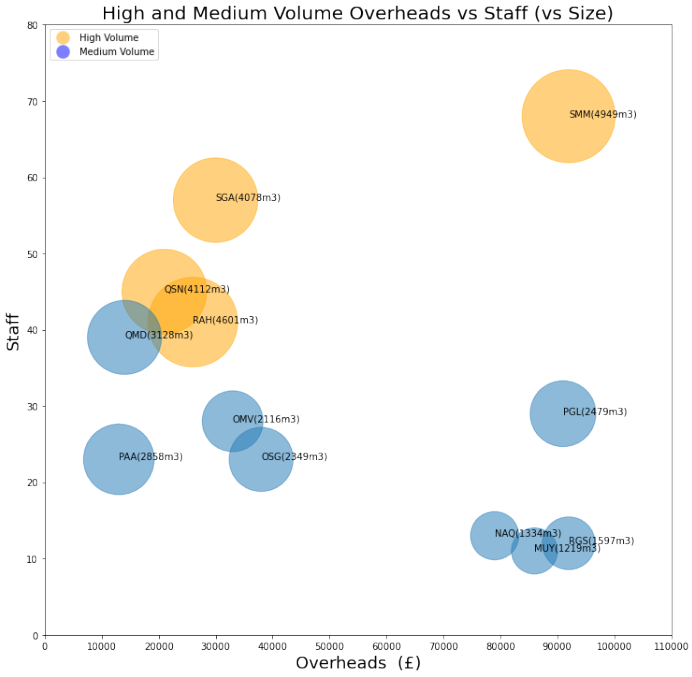


Figure 8: Bubble Plot For high and medium volume stores comparing Overheads, Staff and Size

Bubble plots are excellent graphs when three elements of data need to be compared, for this particular bubble plot the x axis has the overheads for each store with tick values increasing by 10,000 and the y axis has the number of staff at each store with these tick values increasing by 10, the third variable is the size of the bubble, and for this graph refers to the physical size of the store where a larger bubble means a larger store size. In this case there is also both high and medium volume stores that are identified by the colour of the bubble itself which can be identified by the legend located at the top left of this graph, so that comparisons can be made to these three variables across high and medium volume stores.

The interesting information that can be extracted from this graph concern stores NAQ, MUY and RGS, these stores have the fewset members of staff according to the data, all have relatively small floor area as they range from $1219m^3$ to $1597m^3$ but have huge overheads. Store SMM has the most members of staff and largest size of all these stores that are shown in this graph, it would make sense that the overheads for this particular store are high, but when compared to SGA, which is also a high volume store with a similar number of staff assigned to it, the overheads are more then double for SMM.

3 Critical Review

Each of the visualisations has a main title and the axis titles where applicable which makes these graphs clear for the reader to interpret, best practice for the radar plot seen in Figure 5 has also been applied with elements that a company might see as important to grow at the top of the graph and elements they may wish to minimise facing down. Legends have been added to Figures 2, 7 and 8 so that the reader knows exactly what different colours represent when extracting information that is presented. Figure 3 could have been more refined without x and y tick labels for each pair plot as each of the pairs are small so reader accuracy for these particular graphs is not top priority when reading information.

Figure 7 is a second bar chart that is a little messy with overlaying bars, but after trying different alpha values for this chart this plot is the best that could be achieved. There are no seasonality plots in this document as this particular area of statistics is still fairly alien, and explaining something that I do not have a full understanding on how to interpret information in a manner it deserves.

Overall these plots are well built and show the information in a clear and concise manner with labelled axis and legends where they are needed. Subplots also have sub titles to allow readers to know what data relates to which store.

4 Summary

With six stores being opened during the time that this data had been collected, all of these stores are on an upward trend with regard's to customer visit and with two stores closing in the second quarter of the year the business has ended the year with 38 stores opened for business.

Marketing spends correlates with customers visiting stores very closely as does marketing and size. Only overheads do not correlate with any other variable, but to understand why would require more data than is present. Larger sized stores do have more staff assigned to them pertaining to high and medium customer volume stores which would make sense in a real world situation.

Three out of four high customer volume stores have a tight box plot with one or no outliers which would lean towards these stores being stable with customer numbers, but one has a large spread for minimum and maximum values seen in Figure 6 and five outlier values concluding that this particular store has a very varied day to day customer visit count.

5 Conclusion

This particular course has been very enjoyable, having not had any formal teaching on data visualisation before starting has lead to some frustration while building the visualisations that I would like to have have made but turned to joy when these particular graphs worked. Further to this frustration was how to interpret seasonality, which is beyond the remit of this module and was something that I could not dissect myself and ended up not including in the report due to my own lack of knowledge.

Lessons learnt throughout this module will be extremely useful when presenting any type of data in a clear and readable manner using the best practices taught to us and applied in this report, in the near term with regards to the Masters Projects and in the long term career wise.

Reference

Chotisarn, Noptanit et al. (Aug. 2020). "A systematic literature review of modern software visualization". In: *Journal of Visualization* 23.4. Accessed 1 March 2021, pp. 539–558. ISSN: 1875-8975. DOI: 10.1007/s12650-020-00647-w. URL: https://doi.org/10.1007/s12650-020-00647-w.

Liu, Shixia et al. (Dec. 2014). "A survey on information visualization: recent advances and challenges". In: *The Visual Computer* 30.12. Accessed 1 March 2021, pp. 1373–1393. ISSN: 1432-2315. DOI: 10.1007/s00371-013-0892-3. URL: https://doi.org/10.1007/s00371-013-0892-3.

Bibliography

Alnjar, Hanan (2020). "Data visualization metrics between theoretic view and real implementations: A review". In: *DYSONA - Applied Science* 1.2. "Accessed 1 March 2021, pp. 43–50. ISSN: 2708-6283. DOI: 10.30493/das.2020.216111. URL: http://applied.dysona.org/article_107485.html.

Harris, Charles R. et al. (Sept. 2020). "Array programming with NumPy". In: *Nature* 585.7825, pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

Hunter, J. D. (2007). "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

Rudiger, Philipp et al. (Feb. 2020). "holoviz/hvplot: Version 0.5.2". In: DOI: 10.5281/zenodo.3634719.

Seabold, Skipper and Josef Perktold (2010). "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*.

team, The pandas development (Feb. 2020). *pandas-dev/pandas: Pandas*. Version latest. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

Waskom, Michael and the seaborn development team (Sept. 2020). *mwaskom/seaborn*. Version latest. DOI: 10.5281/zenodo.592845. URL: https://doi.org/10.5281/zenodo.592845.

6 Appendix A

List of Figures