Shaun Bell-Gibson

Data Analytics Report – Data Processing Pipeline and Evaluation

Wales is a relatively small country that has access to transport by land, sea and air. Wales has one of the lowest per capita spend on public transport in the UK, which decreased from £74.7 million in 2012-13 to £45.4 million in 2016-17 (The Future Generations Report, 2020). In Wales the dominant form of transport in terms of total distance travelled and total amount of journeys undertaken is via personal use of cars. There are many locations on the road network that regularly experience congestion, most of which are on or directly connected to the M4 motorway which has far reaching consequences on accessibility (Welsh Government, 2021).

Travel by train has increased in recent years, but public satisfaction is a major issue (The Future Generations Report, 2020). Rail within Wales frequently experiences overcrowding which can lead to users being forced into cramped situations or being left at the station. Around 80% of trains arrive within 3 minutes of the scheduled time which is similar to the average around Britain (Welsh Government, 2021). The rail network is limited by its east-west nature greatly restricting travel to the north or south (Fry et al., 2013).

The National Survey for Wales (2019) showed that people felt that using public transport after dark felt less safe than any other mode of transport. There was a 13.3% increase in offences in 2019 compared to 2017/2018, the majority of which were public order offences, violence against the person and theft (Welsh Government, 2021). Populated areas within Wales can reach NHS facilities within an hour using public transport however service availability is greatly reduced outside of major towns, which greatly increases travel times (Fry et al., 2013). Bus travel has decreased between 2003-2017. Around 57% of people that were surveyed said that they would be willing to use their cars less and public transport more, if the quality of public transport improved (The Future Generations Report, 2020).

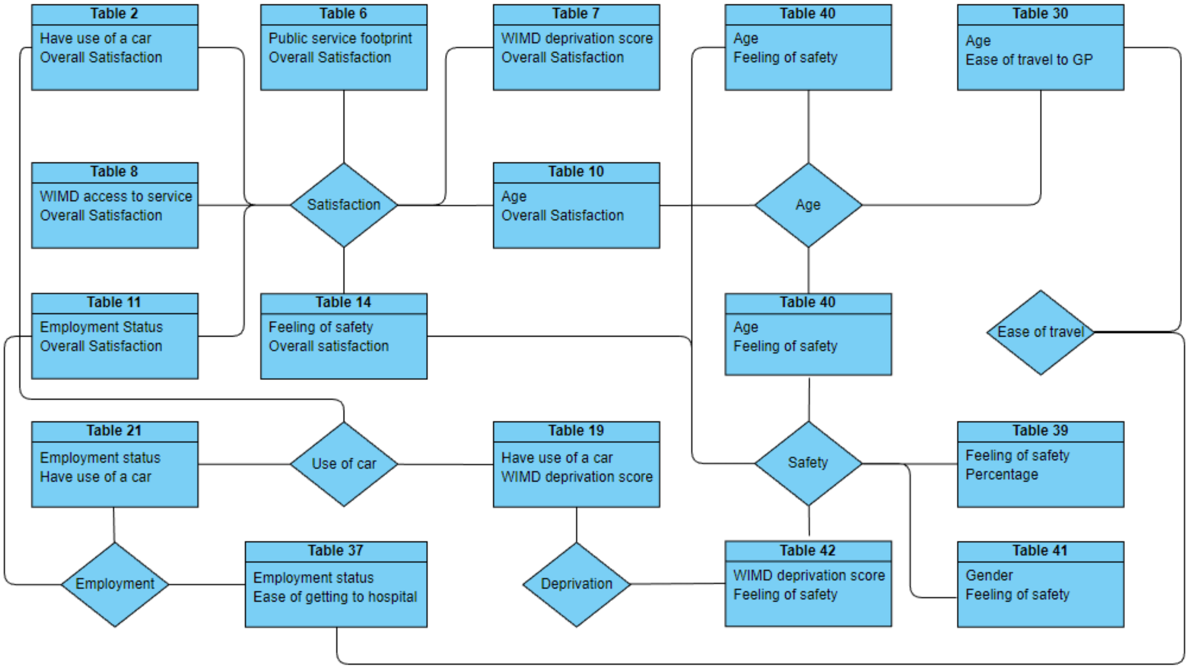
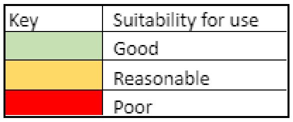


Figure 1: UML diagram of relationships between selected tables

The sample sizes for the survey (approx. 12,300) are a small fraction of the total population of Wales which was approximately 3.1 million in 2014. This means that only 0.4% of the population was surveyed. The survey data set is limited by only showing the sample size for how many people were asked the question. There is no breakdown of the number of people in each category that answered the question. This means that there could have been a small percentage of male responses being compared to a large percentage of female responses which could greatly skew the results. Tables 27,35, 36 and 38 contain values that have been suppressed due to less than 30 participants responding to the question. A major limitation of small sample sizes in relation to a large population is that they can produce false-positive results, or over-estimate the magnitude of an association (Hackshaw, 2008).

Individual data tables don't provide much data individually therefore need to be concatenated to do this the tables will be categorised by a common theme (e.g. household types, age) in order to enhance the usability of the tables providing richer analysis. A significant proportion of the data collected is subjective in nature which is known to contain many cognitive biases (Jahedi & Méndez, 2014), the answers that the participant provided could significantly alter depending on a number of factors outside the control of the data collection process. This subjectivity may lead to inaccuracies due to certain answer options that may be interpreted differently by respondents (DeFranzo, 2022). Only data from one year is provided therefore change over time can not be observed and the results can not be compared to other years to check for anomalies in the data.



Key	Suitability for use
🟩	Good
🟨	Reasonable
🟥	Poor

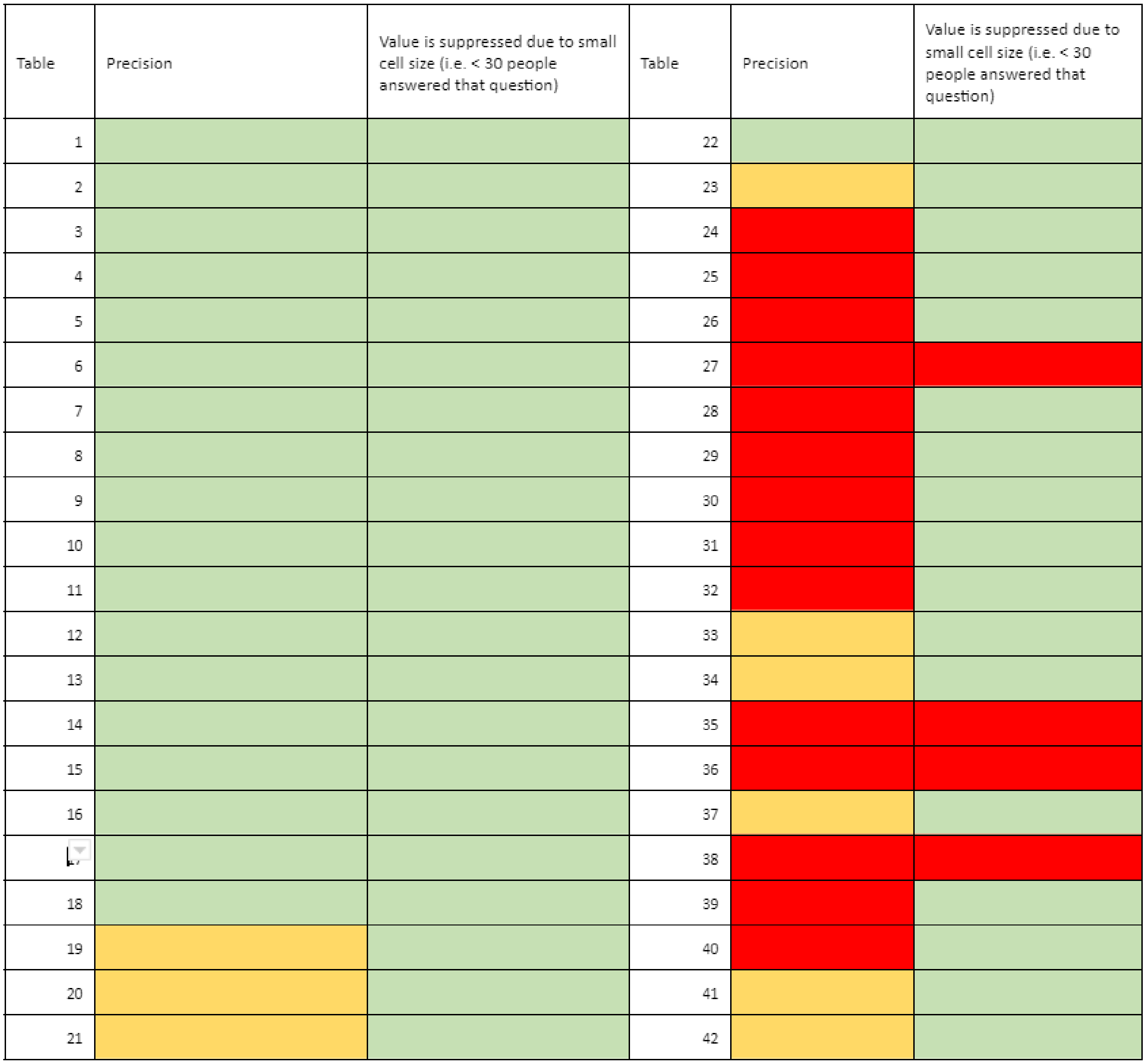


Table	Precision	Value is suppressed due to small cell size (i.e. < 30 people answered that question)	Table	Precision	Value is suppressed due to small cell size (i.e. < 30 people answered that question)
1	🟩	🟩	22	🟩	🟩
2	🟩	🟩	23	🟨	🟩
3	🟩	🟩	24	🟥	🟩
4	🟩	🟩	25	🟥	🟩
5	🟩	🟩	26	🟥	🟩
6	🟩	🟩	27	🟥	🟥
7	🟩	🟩	28	🟥	🟩
8	🟩	🟩	29	🟥	🟩
9	🟩	🟩	30	🟥	🟩
10	🟩	🟩	31	🟥	🟩
11	🟩	🟩	32	🟥	🟩
12	🟩	🟩	33	🟨	🟩
13	🟩	🟩	34	🟨	🟩
14	🟩	🟩	35	🟥	🟥
15	🟩	🟩	36	🟥	🟥
16	🟩	🟩	37	🟨	🟩
17	🟩	🟩	38	🟥	🟥
18	🟩	🟩	39	🟥	🟩
19	🟨	🟩	40	🟥	🟩
20	🟨	🟩	41	🟨	🟩
21	🟨	🟩	42	🟨	🟩

Figure 2: Analysis of dataset tables

The key performance indicators (KPIs) have been selected to assess the problems identified in the background research. If these factors are improved then it is likely that the overall performance of the transport network is more efficient.

The KPIs that have been selected are:

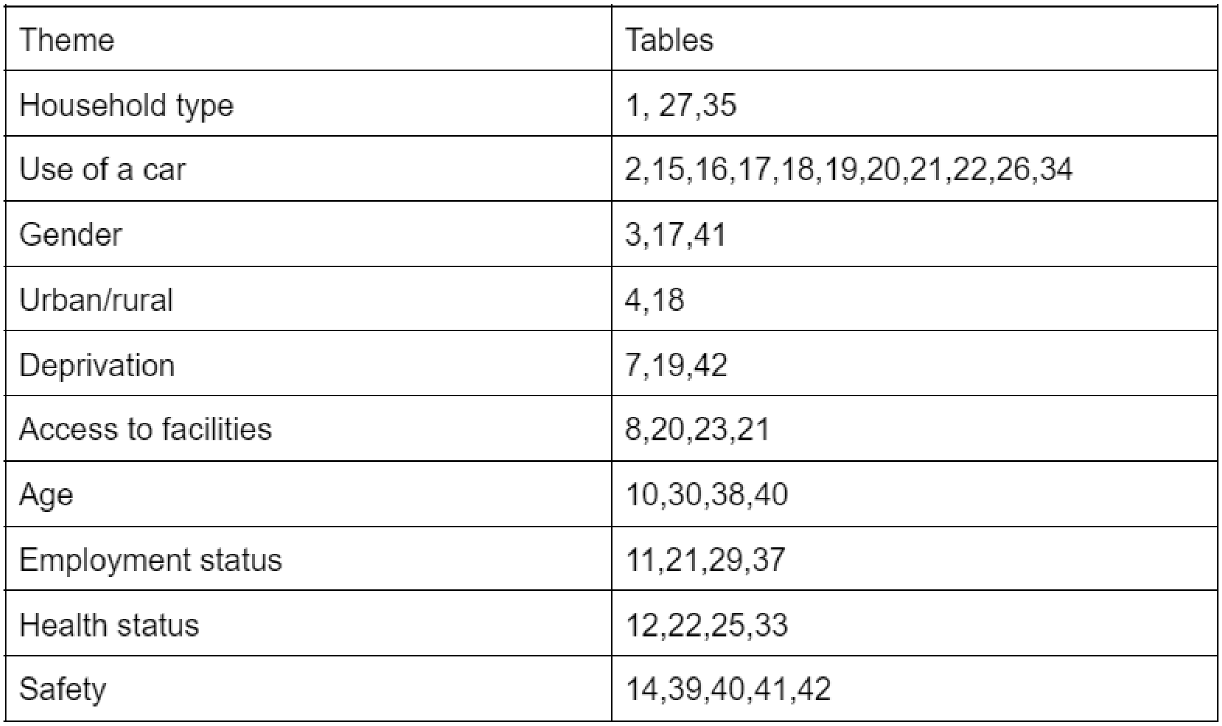
1. The public's satisfaction levels,

2. How safe members of the public feel while using transport
3. How accessible health services are to the public.

The main questions that have been identified are:

1. Which groups of the population are the least satisfied with the transport system?
2. Does car ownership affect views on the transport system?
3. What effects does safety have on transport satisfaction?
4. How accessible are health care services via the transport system?

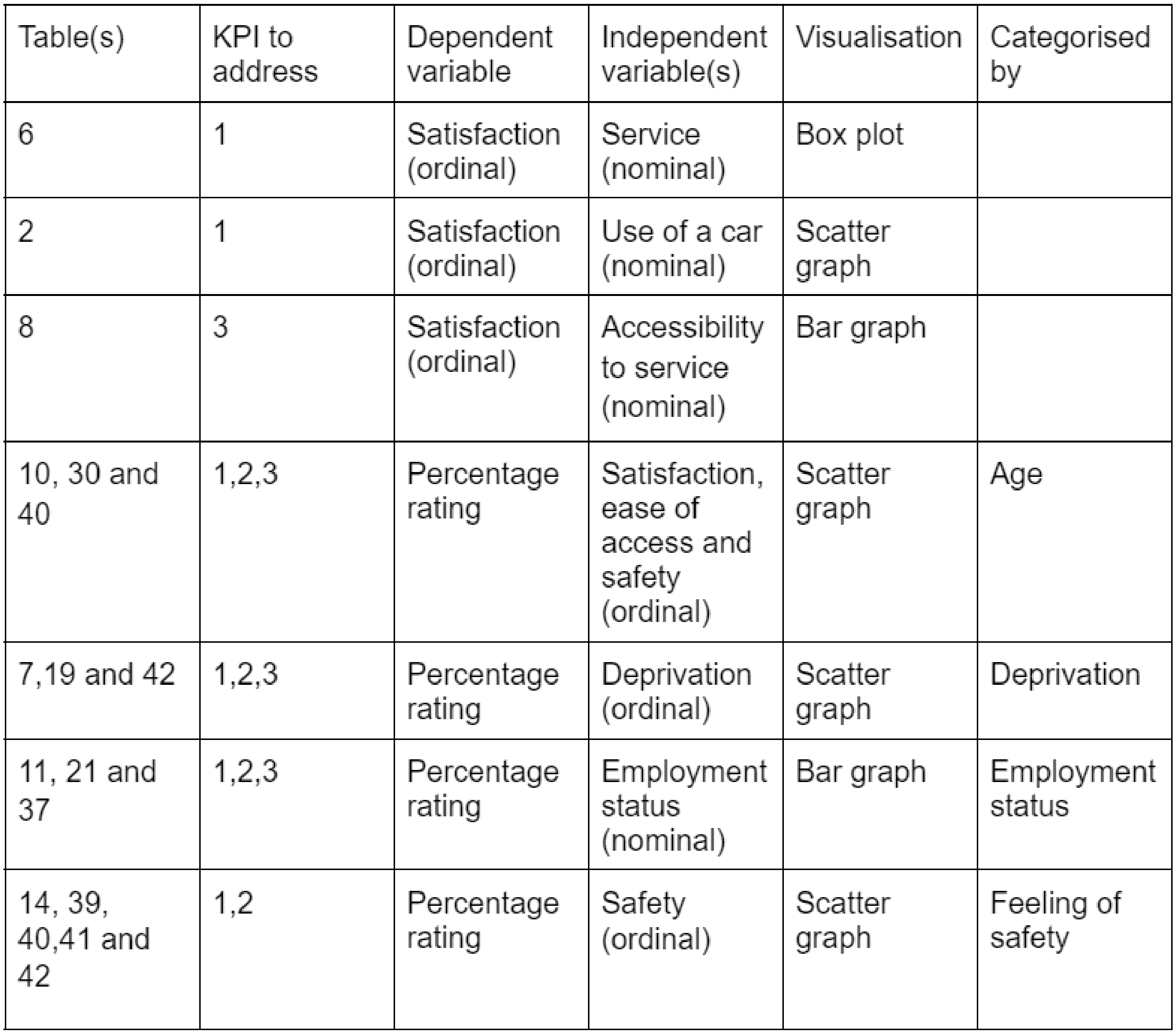
The tables of the data set are sorted into categories based on certain themes that their data relates to, these themes are shown in Figure 3 (below).



Theme	Tables
Household type	1, 27,35
Use of a car	2,15,16,17,18,19,20,21,22,26,34
Gender	3,17,41
Urban/rural	4,18
Deprivation	7,19,42
Access to facilities	8,20,23,21
Age	10,30,38,40
Employment status	11,21,29,37
Health status	12,22,25,33
Safety	14,39,40,41,42

Figure 3: Theme of tables

A pandas dataset will be created to store and manipulate the data. Pandas allows for simple methods of preprocessing and transforming data as well as allowing easy integration with other libraries which will be important for this assignment (Kupferschmidt, 2020). The isnull method will be used to detect null values. Within data tables missing values are caused by a small sample size therefore instead of replacing the missing values, the columns with the missing values will be omitted from the dataset. To create visualisations seaborn has been selected over matplotlib. Seaborn is more efficient at handling pandas data frames whilst using comparatively simple syntax which offers more functionality (Hacksight, 2022). A limitation of seaborn is that multiple figures may lead to out of memory related issues (Hacksight, 2022).



Table(s)	KPI to address	Dependent variable	Independent variable(s)	Visualisation	Categorised by
6	1	Satisfaction (ordinal)	Service (nominal)	Box plot	
2	1	Satisfaction (ordinal)	Use of a car (nominal)	Scatter graph	
8	3	Satisfaction (ordinal)	Accessibility to service (nominal)	Bar graph	
10, 30 and 40	1,2,3	Percentage rating	Satisfaction, ease of access and safety (ordinal)	Scatter graph	Age
7,19 and 42	1,2,3	Percentage rating	Deprivation (ordinal)	Scatter graph	Deprivation
11, 21 and 37	1,2,3	Percentage rating	Employment status (nominal)	Bar graph	Employment status
14, 39, 40,41 and 42	1,2	Percentage rating	Safety (ordinal)	Scatter graph	Feeling of safety

Figure 4: Breakdown of tables to create visualisations

Satisfaction by service will be shown using table 6. This table will show which of the services in Wales have significantly high or low levels of satisfaction. This allows for the customer to pinpoint problem areas that need improvement or to discover what variables set the services with high satisfaction apart. The independent variable is nominal data and the dependent is ordinal which means for this a box plot will be created displaying the mean satisfaction as well as the range of values for easy comparison.

A box plot was selected to display the satisfaction by service table. This type of chart was selected for its ability to display and easily compare the distribution of data between independent variables. Therefore a box plot will allow the customer to easily interpret which of the services have users with high satisfaction levels and which services have low satisfaction. The main strength of a box plot is its ability to clearly display information in a clear summary that can compare different results (Ladkin, 2019). A histogram is another chart that could have been selected to show the distribution of data. A box plot shows less data than a histogram, which can make it clear and easy to interpret for the customer (Statistics Kingdom, nd). A histogram requires a larger sample size than a box plot to be useful (Krzywinski & Altman, 2014). A limitation of the box plot is that the exact values of the data points are not retained meaning that it can not be used to show a detailed analysis of the data. However, this limitation can be overcome by using the box plot alongside other more detailed visual methods (Ladkin, 2019).

Satisfaction by use of car will be shown using table 2. This table will show whether there is a higher level of satisfaction within car owners or people without a car, who are more likely to use public transport. If the level of satisfaction is lower in car owners then that may indicate that there is an issue within the road network. A scatter chart will be created displaying the mean satisfaction for easy comparison.

In order to assess whether accessibility to services has an effect on the levels of satisfaction table 8 will be displayed as a bar chart. A bar chart will also be created for tables in the employment theme, very safe and safe columns and easy/very easy will be combined to create a single easy column. A bar chart was selected as a suitable method of displaying qualitative variables that are either nominal or ordinal in nature. A strength of bar charts is that it is a simple way of displaying differences between variables; however, it only allows for a small number of variables to be described at a time and may become difficult to interpret with a large number of categories (Biderbost & Carrasquero, 2021).

The tables categorised in the age theme will be concatenated into a dataframe. Table 38 will not be used due to its unreliable estimates and missing values. Within the new dataframe columns easy and very easy will be combined with the values added together to display the total percentage of everyone that found getting to the hospital easy. The same will occur with the very safe and safe columns of table 40. The difficult and unsafe columns will be dropped due to unreliable data being present. Once the values have been normalised to a range of 0-1 to allow the multiple variables to be plotted on the same axis for easy comparison, a scatter chart will be created. The same process will be used for tables categorised into the deprived theme and the safety theme.
Scatter graphs are an effective method of visualising the potential correlation between variables and how strong the correlation is (Royal Geographical Society, nd). A strength of a scatter graph is that data can be categorised by another variable to provide further comparison.

References:

Biderbost, P. and Carrasquero, G. (2021) “Learn to create bar charts in RStudio with data from a study of sleep patterns for college students (2012).” Available at: https://doi.org/10.4135/9781529777819.

DeFranzo, S.E. (2022) Advantages and disadvantages of surveys, Snap Surveys Blog. Available at: https://www.snapsurveys.com/blog/advantages-disadvantages-surveys/ (Accessed: November 11, 2022).

Fry, R., Rodgers, S. and Lyons, R. (2013) Public Transport Overview, ABMU Changing for the Better: Public Transport . Centre for Health Information, Research and evaluation (CHIRAL), Swansea University. Available at: https://sbuhb.nhs.wales/files/freedom-of-information-disclosure-log-2020/august/20-h-041-public-transport-report-2013/ (Accessed: November 3, 2022).

Hackshaw, A. (2008) “Small studies: Strengths and Limitations,” European Respiratory Journal, 32(5), pp. 1141–1143. Available at: https://doi.org/10.1183/09031936.00136408.

Hacksight (2022) *Difference between matplotlib vs Seaborn*, *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/difference-between-matplotlib-vs-seaborn/ (Accessed: November 17, 2022).

Jahedi, S. and Méndez, F. (2014) “On the advantages and disadvantages of subjective measures,” Journal of Economic Behavior & Organization, 98, pp. 97–114. Available at: https://doi.org/10.1016/j.jebo.2013.12.016.

Krzywinski, M. and Altman, N. (2014) “Visualizing samples with box plots,” Nature Methods, 11(2), pp. 119–120. Available at: https://doi.org/10.1038/nmeth.2813.

Kupferschmidt, K. (2020) *Spark vs Pandas, part 1 — Pandas - towardsdatascience.com*, *Spark vs Pandas, part 1 — Pandas*. Available at:

https://towardsdatascience.com/spark-vs-pandas-part-1-pandas-10d768b979f5 (Accessed: November 14, 2022).

Ladkin, A. (2019) How to create a box-plot chart, Sciencing. Available at: https://sciencing.com/create-boxplot-chart-5858878.html (Accessed: November 14, 2022).

National Survey for Wales (2019) National Survey for Wales 2019, GOV.WALES. National Survey for wales. Available at: https://gov.wales/national-survey-wales-results-viewer (Accessed: November 3, 2022).

Royal Geographical Society (nd) “A Guide to Scatter and Line Graphs.” Available at: https://www.rgs.org/CMSPages/GetFile.aspx?nodeguid=227c9dd8-b2e5-4081-a67f-8a59641207c2&lang=en-GB

Statistics Kingdom (nd) Box plot maker. Available at: https://www.statskingdom.com/boxplot-maker.html (Accessed: November 14, 2022).

The Future Generations Report (2020) Executive Summaries Transport - exec summary - futuregenerations.wales. Available at: https://www.futuregenerations.wales/wp-content/uploads/2020/07/Transport-Exec-Summary.pdf (Accessed: November 3, 2022).

Welsh Government (2021) A New Wales Transport Strategy Consultation Draft , GOV.WALES. Available at: https://gov.wales/sites/default/files/consultations/2020-11/supporting-information-transport-data-and-trends.pdf (Accessed: November 3, 2022).