Data Analytics Report

Matthew Bowyer

Breakdown of Project:

Transportation of Wales has conducted a detailed survey to understand its citizen's view on the public transport system. As a data scientist, you are tasked to write a report that will facilitate the executives to take necessary measures to improve the transportation systems in Wales:

- Carry out some background research on the transportation systems in Wales, UK. The aim of this step is to explore and understand the transportation systems in Wales and current issues related to transportation in the area.
- Critically evaluate all 42 tables of the National Survey for Wales results, 2013-14: Transport.
- Design a data analysis strategy for creating the data analytics report. Provide a rationale for your design choice (i.e., which tables to choose and/or concatenate, and why). Use the Data Pipeline and the Master Data Management Model for designing your strategy.
- Define your data representation strategies (i.e., what type of graphs/charts you would use and why).

Your report (1,500 words) should contain the following aspects. Note that the associated grading criteria are highlighted in the requirements below, to be reviewed alongside the criteria grid (Module Resources):

- A description of transportation in Wales, focusing primarily on citizen's view and current issues. This section should reflect your knowledge and understanding on transportation in Wales (15% weight).
- An overview of the survey dataset and an UML diagram showing relationship within the tables. This section should reflect your knowledge and understanding on the dataset (25% weight).
- A critical discussion on data analysis design choice and methodology (30% weight).
- A critical discussion on data representation choices (10% weight).
- Detailed discussion on the strengths and limitations of different choices (20% weight).

Some useful links to help you in your report are:

- https://www.gov.wales/transport
- https://www.walesonline.co.uk/all-about/transport

My work:

Unit 9

Data Analytics Report

Since the national survey used for this report is from 2013-14, all background information will be from that time period too. Later in the report, I will dive into current findings. The Welsh Government is heading initiatives to integrate public transport across Wales. With emphasis on taking a national approach to planning and execution of the project for increased cohesive, across-the-board planning. This will also assist with a statement from the government about how the people who use public transport are not often taken into consideration. (Priestley, 2013) Integrating the publics needs and wants into the planning and execution of the project reflects Wales drive towards inclusivity and accessibility to all. (Public Transport Users' Advisory Panel, 2013) The entire initiative has the potential to increase and support jobs and the economic stability, promote equality, and reduce poverty while decreasing environmental damage. (Webster, 2013) Projects like the Cardiff Captial Region Metro are larger projects that could only be completed by 2030, but have the opportunity to create 7000 jobs and contribute 4 Billion pounds towards the economy. This will connect over 70% of the public transportation users to the Cardiff City Region. (Barry, 2013) Currently, the public transport system is not run at full efficiency and the government is looking to step in and assist. With many social and economic benefits predicted, Improvements to the transportation systems in Wales looks to benefit all.

The national survey used for this report is the National Survey for Wales, 2013-14 – Transport. Wales nationals were asked questions on the transport system. Questions in the context of personal usage of cars, overall satisfaction with the public transport system in Wales, ease of getting to and from the hospital or GP surgery and feeling of safety while travelling. The interviews were carried out between April 2013 to March 2014. The questions created 42 tables of feedback. The vast majority of the tables are the questioned I listed above but group by variables like gender, satisfaction with life and many more. The grouping of the data creates a significant advantage for analysis as we can compare the multiple variables to each other in order to approach a potential solution. Noting that as the survey states, relationships between two variables does not prove correlation. More technical notes by the survey include refusal to answer or “Don’t know” answers were removed from the survey. If too few people answered an opinion-based question, the results were grouped, like “Very” and “Fairly” anything was grouped as one. E.g. “Very satisfied” and “Fairly satisfied” were grouped as “Satisfied”. A coefficient of variation (CV) value has been added to assist with knowing the solidity of the estimates given. With lower CV values showing more precision, larger showing less precision and no value showing too small of sample size to display CV. 95% confidence intervals, to assist with calculating if there is a real difference between two variables. Sample sizes have been given below the results, rounded to the nearest 100, except for samples below 100 which show the actual value.

The 42 tables could provide deep insight into the views of the public. But to analyse the data, I need to store, clean, and present the data in a way that can be analysed. Following the data management life cycle, we have already completed step 1, collection of data (The survey results). The next step is to process the data, this involves cleaning and manipulation of the data. A great method of doing this is through Python Pandas. With excellent integration with excel and methods of skipping lines and getting the exact inputs you require. Python can clean and prepare the data for the next step, which is store the data. I will opt for a relational database. Using a relational database is often an excellent choice for structured data that fits well into tables. With the structured data in tables, querying and further analysis becomes significantly easier. Other options included a non-relational data base which is possible with the data given and would provide a single table to extract information from but would become complex with the sub-grouping of headers to fit the different layers of depth some tables have in the data. Complicated and hence limiting the analytical potential. Relational databases allows for a Master table that follows Master Data Modelling (MDM) practises and focuses on data quality and standardization. The Master table will have the ID's that make up the complex data from the survey. From there, you can dive into the tables that contain the ID values. Question_ID stores the main question, Group_ID stores the group and Result_ID is why this data is so complex. The Result_ID also needs to be linked with a Sub_Result_ID, which stores questions that had multiple choice. This creates rigorous structure and less chance of contamination of data. Notice that I add a column called "Table" in the Result_ID table to have the option to go back to the Survey and check if the data in the database matches with what is in the Survey excel sheet.

Master_Table

Question_ID	Group_ID	Result_ID
1	1	1
1	1	2
2	3	3

Question_table:

Question_ID	Question
1	Overall satisfaction with state of transport system in Wales
2	Have use of a car
3	Ease of getting to and from GP surgery

Group_table:

Group_ID	Group
1	Household type
2	Have use of a car
3	Limiting long-term illness

Result_table:

Result_ID	Result	Sub_Result_ID	Mean	%	Lower_CI	Upper_CI	Table
1	Single pensioner (no children)	Null	6.2	Null	6.0	6.4	1
2	Married couple pensioner (no children)	Null	6.0	Null	5.9	6.2	1
3	Limiting long term illness	1	Null	69	67	70	22

Sub_results_table:

Sub_Result_ID	Sub_Result
1	Yes
2	No
1	Very Easy

Now that the data is in a structured format. I can use the data which is step 4. Use the data to extract analysis and mine for corelations that may be able to answer insightful questions. Step 5 will be where I share and communicate the results, using visualizations and statistics from the analysis to provide facts and correlations found. The final step is the decide on the fate of the data. In this case, the data will be stored to use for comparison with the next year's survey results. To see if the improvements made have made significant impact on the public opinion. Difficulties with this decision include the time it will take to individually code Python Pandas to extract all tables from the national survey excel sheet. The sheet needs to be more standardized, where the tables all have the same format. The database will also require background knowledge of the national survey and the storage method as relational databases are rarely self-explanatory.

The types of visualizations used can make a significant difference to the way results are interpreted. Visualizations that are difficult to interpret and confusing can be looked over and force your viewers to miss key points in the results. I believe a lot of my visualizations will come in the form of bar charts and histograms. Comparing the results against one another to portray the publics views. Bar graphs can emphasize differences in opinions between multiple variables/groups. A more complex way of demonstrating the publics views will be with a scatterplot matric, which demonstrates correlations between variables/groups.

In 2024 current affairs, Maesteg residents were angered by limited bus timetables and disappointing bus turnaround times. executives are listening to Maesteg residents views and planning changes like connecting towns to improve the issues. (Gavaghan, 2024) This is a great way of demonstrating how listening to the public can improve productivity and eventually profit

margins. Not all situations require executives to make changes, sometimes the public can see where businesses need to improve. In a historic north Wales village, due to the road being originally built for horse and cart, residents who do not take extra precautions cause prolonged waits for buses unable to squeeze between the buildings and the parked cars. (Summer & Forgrave ,2024) In this case, residents know it is car-owning residents causing the issues and they need to take accountability and not the bus executives. In both cases, public views are being used to figure out the problem to increase the chances of solving it efficiently.

The transport system can be an asset for all. A carefree transport system for the public and a profit gainer for executives. Using the instrumental tool, that is public views, the transport system as an asset can grow. Hence analysing the national survey and making sure the public views are kept on top of annually, strategic measure can be taken to improve the transportation system to benefit all.

(1445/1500 words)

References

Priestley, M. (2013) North East Wales Integrated Transport Task Force Report to Edwina Hart AM OStJ MBE Minister for Economy, Science and Transport. Available from: north-east-wales-integrated-transport-report.pdf (gov.wales) [Accessed 13 March 2024]

Public Transport Users' Advisory Panel. (2013) INTEGRATED TRANSPORT IN WALES. Available from: Integrated transport in Wales report 2013 | GOV.WALES [Accessed 13 March 2024]

Webster, H. (2013) North East Wales Integrated Transport Task Force Technical Report. Available from: north-east-wales-integrated-transport-technical-report.pdf (gov.wales) [Accessed 13 March 2024]

Barry, MD. (2013) A Cardiff Capital Region Metro: Impact Study - Executive Summary. Available from: a-cardiff-capital-region-metro-impact-study-executive-summary.pdf (gov.wales) [Accessed 13 March 2024]

Gavaghan, B. (2024) Town's residents feel failed and isolated amid public transport problems. Available from: Town's residents feel failed and isolated amid public transport problems - Wales Online [Accessed 18 March 2024]

Summer, B., Forgrave, A. (2024) Bus and car in hour-long incident in narrow street leaving passengers furious. Available from: Bus and car in hour-long incident in narrow street leaving passengers furious - Wales Online [Accessed 18 March 2024]