Data Analytics implementation

Matthew Bowyer

Breakdown of Project:

This assignment follows on from your work in the Unit 9 submission, where you outlined your approach to analyse the National Survey for Wales results, 2013-14: Transport dataset.

For this assignment, you will undertake the analysis according to your chosen UML design using Python libraries to map up your UML implementation. Ensure you review the feedback you received from that submission before attempting this assignment.

Instructions

Your assignment should contain the following aspects:

1. A brief description of transportation of Wales. (Knowledge and Understanding weighted at 5%).
2. A detailed description of the data pre-processing steps undertaken. (Knowledge and Understanding weighted at 5%, Application of Knowledge and Understanding weighted at 10%).
3. The step-by-step review of the data analysis methods that were undertaken. If there is any deviation from the original design provided in Unit 9, please provide a rationale for the change. (Knowledge and Understanding weighted at 5%, Application of Knowledge and Understanding weighted at 5%).
4. Sequential presentation of the data analysis results in the form of graphs/charts. (Knowledge and Understanding weighted at 10%, Application of Knowledge and Understanding weighted at 10%).
5. Provide your own interpretation of the result for each graph/chart. From this interpretation the issues related to transportation in Wales should be clearly identified and should be substantiated by the graphs/charts, and
6. Provide solution(s) to overcome the identified issues with possible limitations/risks involved. (Criticality weighted at 25%).

Presentation and Structure of your work (weighted at 25%) includes spelling, style, evidence of proofreading, correct use (and format) of citations and references.

Deliverables

- Your deliverables/submission for this assessment are:
- An implementation model capturing data management and deployment concepts in in Python submission and a Rationale document explaining the content above.

My work:

Report:

Brief Description of Transportation in Wales

The Welsh transportation system, highlighted in my previous report, underpins substantial social and economic benefits such as job creation and environmental sustainability. The Cardiff Capital Region Metro, anticipated to be operational by 2030, exemplifies the long-term strategic planning undertaken to enhance public transport infrastructure, aiming to connect over 70% of public transportation users within the Cardiff City Region. (Barry, 2013).

Current research underscores the pivotal role of accessibility in public transport, which can significantly influence unemployment rates, mode choice, property prices, public health, and social inequity. For instance, improved transportation access could help break the poverty cycle by facilitating easier access to education for underprivileged children (Torres & McArthur, 2024). While technological advances offer potential enhancements, their adoption must be judicious and based on thorough research to ensure they meet the practical needs of users (Merkert & Nelson, 2024). The COVID-19 pandemic has further demonstrated the necessity for preparedness and adaptable planning in public transportation to avoid rash decisions that fail to benefit the populace, as seen in the transportation delays in Maesteg (Preston & Wreststrand, 2024; Gavaghan, 2024). Leveraging historical data can guide us in making strategic reliable decisions, like knowing where underprivileged children struggle the most to find transport, to refine and advance our transportation framework.

Data Pre-processing Steps

Data pre-processing is a fundamental step in ensuring the quality and usability of data for analysis. In this project, I utilized Python libraries, notably pandas, NumPy and Itertools, to perform data cleaning and preparation. Pandas is a great way to store and manipulate data. NumPy is used to work through or generate NaN values. Itertools is used to iterate between the grouping done in the excel sheet.

Once loaded, the data underwent several cleaning procedures. I removed unnecessary or completely blank rows and columns using dropna(). This step was crucial to eliminate noise and focus only on meaningful data. Additionally, specific rows beneath the "Total" entries in all tables were discarded, as data below that was unnecessary for my analysis. Noting, I found no use for sample size, so I left it out. The noise and blank data were due to the excel sheet being in a more human readable structure. This made the code difficult to generalize, having to work around the different types of tables and groupings throughout the excel document. Generalizing code is usually used in documents like CSV files. In documents like the excel document, you will see a lot of step-by-step cleaning had to be done.

Data transformation included renaming columns for better clarity and accessibility during analysis. For instance, initial column names derived directly from the file were replaced with more descriptive titles based on the content of each column. Like “%” becoming “Percentage”. This is why I have a Pandas step between the excel document and the SQLite database. Using the easier to manipulate Panda’s library before inputting into a database.

Integration with a SQL database was the final step in my data pre-processing routine. After cleaning and structuring the data, it was formatted for database insertion into SQLite. I created several tables in the database to categorically store questions, groups, and results from the processed data, facilitating efficient data retrieval and management.

In conclusion, the data pre-processing performed serves as a foundation for reliable data analysis. With all the tables linked, we can compare each Table from the excel document to each other or call multiple surveys for graphing. Making the data completely accessible.

Data Analysis Methods

The data analysis approach for this project is derived from the SQL relational database outlined in the previous assignment's Unified Modelling Language (UML) diagram, which served as a blueprint for structuring the database and guiding the analysis. This UML framework was instrumental in maintaining data integrity and ensuring the consistency of analysis methodologies across the different survey data sets. No significant deviations from the original UML design were necessary because the initial model effectively accommodated the complexity and scope of the data involved.

The SQL database has a Master table which acts as a central hub, linking questions, groups, and results through foreign keys (Question_ID, Group_ID, and Result_ID). This structure enables flexible querying and cross-referencing, allowing you to analyse data in multiple dimensions.

The Question table provides the source of the main questions like “Feeling of safety travelling by public transport after dark” from Table 42. Its link with the Master Table allows associating questions with specific groups and results.

The Group Table categorizes data into various segments, which can be referenced to understand different perspectives or sub-groups within the survey. Groups like “by WIMD deprivation” from Table 42.

The Result Table represents the primary data points, with a connection to the Sub Results Table via Sub_Result_ID, allowing for more granular analysis where needed. Results include choices like “20% most deprived” from table 42 and then links the “%”, “Lower CI” and “Upper CI” to the result and the Result.

The Sub Results Table provides additional context and detail for results, supporting complex hierarchical relationships within the data. Like “Very safe” from table 42.

Benefits of this Structure allows for flexibility and scalability, enabling me to use foreign keys ensures data integrity and consistency across the database. The interconnected tables allow for complex queries, facilitating comprehensive analysis and reporting.

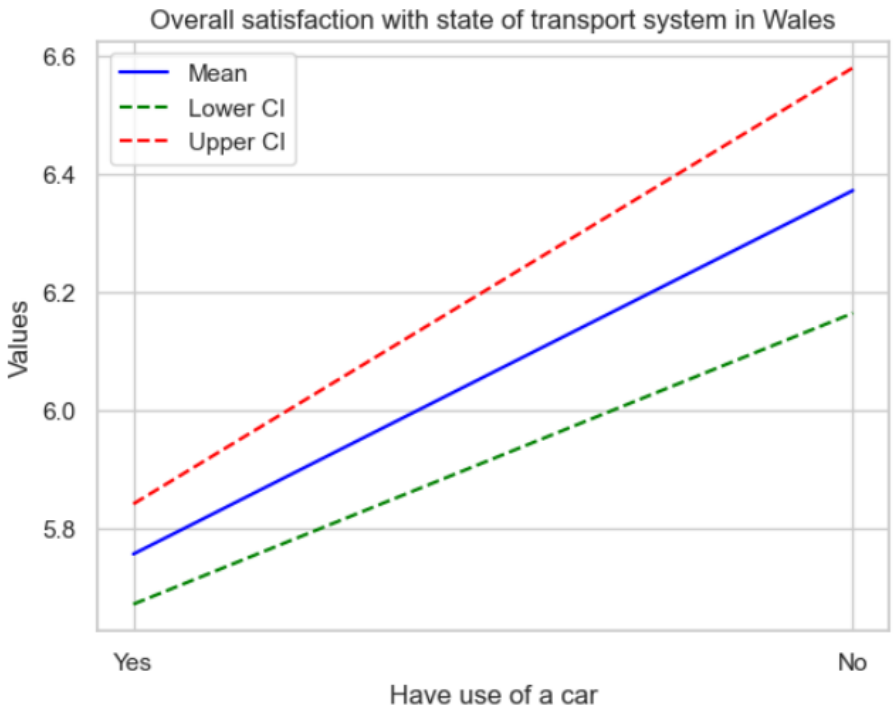
Though I cannot access the links on the contents page to learn more about the Survey questionnaire or the other survey details. I looked into some of the terms I do not know off hand. Like “WIMD”, which as per (StatsWales), Welsh Index of Multiple Deprivation (WIMD) provides a score from 1 (Most deprived) to 1 909 (Least deprived) locations in Wales. In the excel document, 5 percentage ranges are used, 20% most deprived, 20%-40% most deprived, 40%-60% most deprived, 20%-40% least deprived and 20% least deprived. “ACORN classification”, explained in (Limbu, 2023), is a concatenation of data sources to provide a deeper

understanding of an area. In the Excel document, they are categorized as Wealthy Achievers, Urban Prosperity, Comfortably Off, Moderate Means and Hard Pressed.

Multiple factors also need to be considered. Correlation between tables does not prove causation, just because the data may show correlation between tables does not mean this is fact. This relationship could just be coincidence. Mean is used when a score out of 10 is given, demonstrating the average answer between the survey takers. Percentage is used when there are a limited amount of answers and shows the range of categories that selected those answers. Confidence intervals are used to show potential deviations in the mean and percentage averages. This allows us to see how rigid the average is and can show if two answers have true statistical differences if the Upper and Lower confidence intervals don't overlap.

Presentation of Data Analysis Results

I begin by visualizing the overall satisfaction of the state of transport system in Wales, by have use of car. Allowing me to address how accessibility plays a pivotal role in the public transport system in Wales.



Interestingly, the public who do not have use of a car rate their overall satisfaction higher than the public who do have use of a car. With means at roughly 6.4 and 5.8 respectfully. This could be due to many reasons, diving into table 26 and 34 could provide further answers.

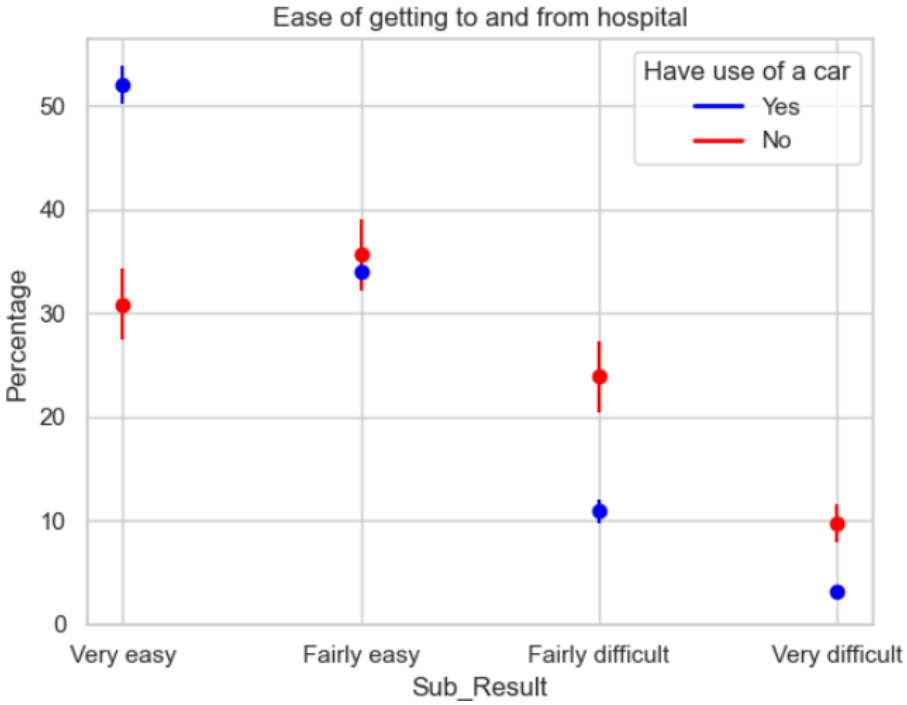


Table 26 shows that the public who have user of a car have an easier time getting to and from the hospital overall.

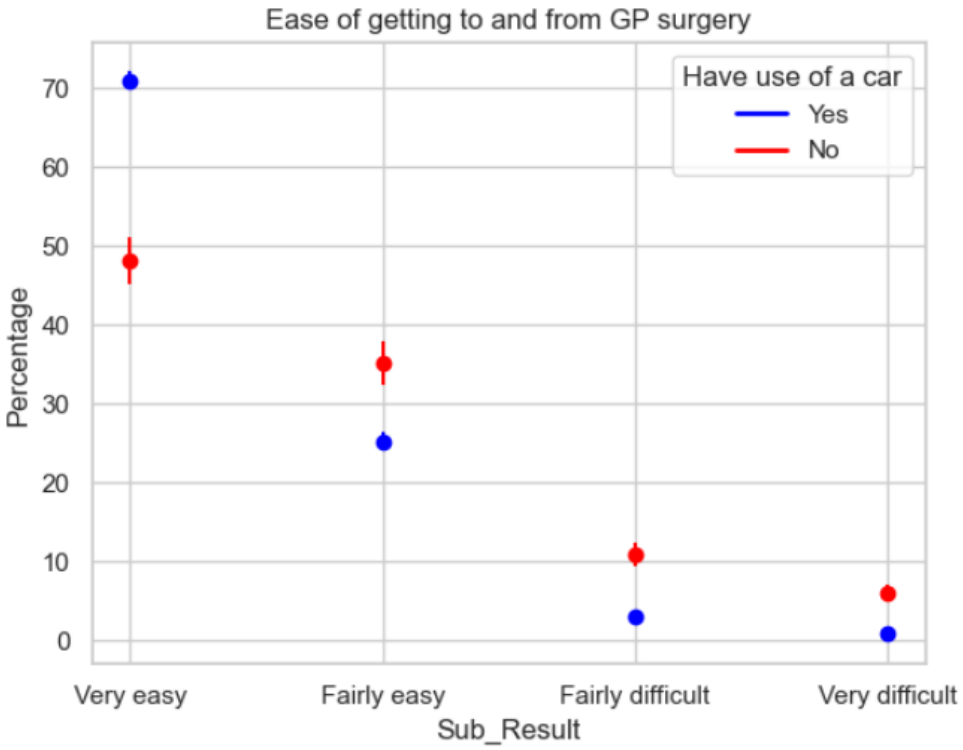
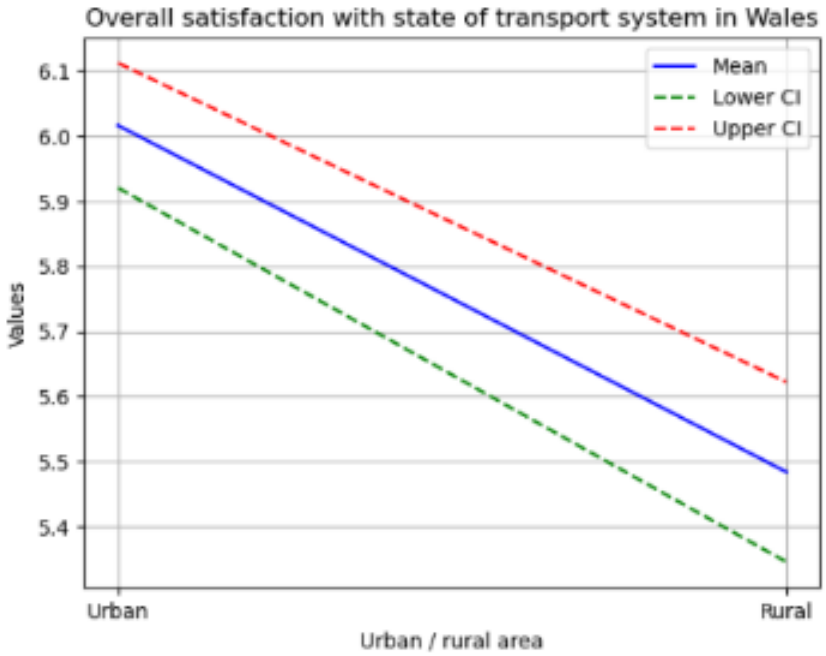


Table 34 shows the same trend as table 26. The public who have use of a car find it easier to get to and from GP surgery overall.

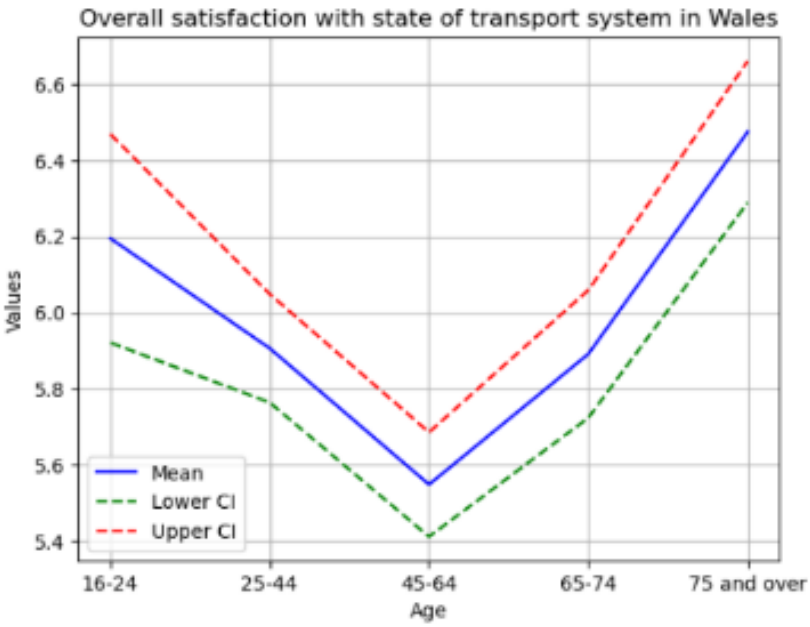
With majority of the public with use of a car finding it very easier to get to and from public health providers. It makes sense, though could be coincidence and not causation, that the relationship between public with use of a car and satisfaction with public transport system agree.

As explained in the first section. Lower to middle-class residents predominantly use public transport. Lets see how Urban vs Rural areas feel about the transport system with table 4.



Though Rural areas have more variance in the confidence intervals, rural areas are less satisfied with the state of transport systems. This leads me to investigate how the younger generation feel about the public transport.

From section 1, I explained how Improved transportation access could help break the poverty cycle by facilitating easier access to education for underprivileged children. For that to happen, one variable we need to work on is safety. Table 10 and 40 can help explain how younger people feel and the transport system and how safe they feel using public transport respectively.



From Table 10, the graph shows that the younger people a more satisfied with the transport system than less. Noting that all age groups are less satisfied than the younger people besides the 75 and over age range.

At least 50% of all age ranges who answered the survey at least feel fairly safe. Though safety is not perfect all-round. Majority of people choosing safe over unsafe is postive for public transport usage.

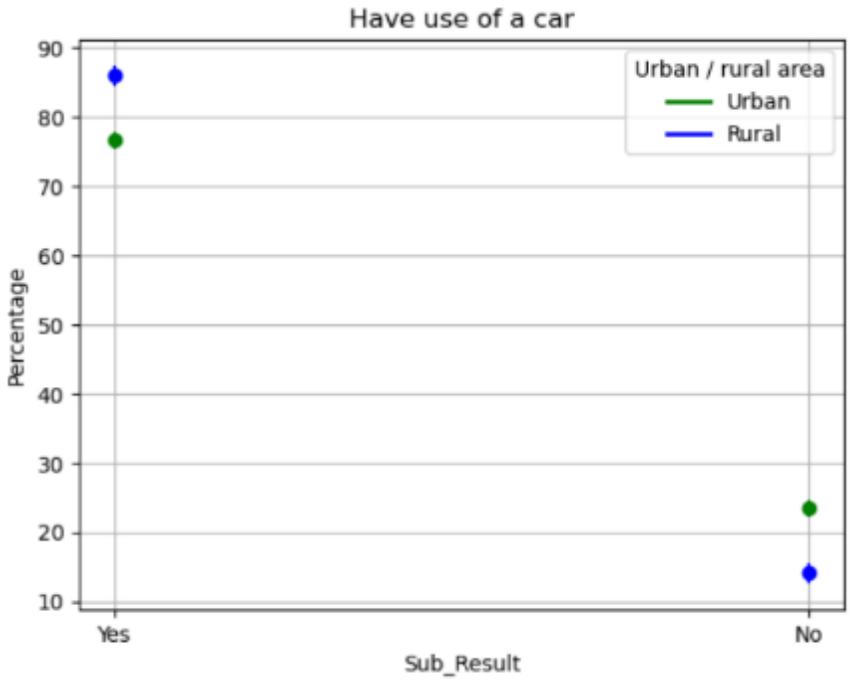
Using SQL, we can do so much more. For example, we can see where the highest Means and the highest Percentages of opinions are found.

	Question	Group	Result	Sub_Result	Mean	Percentage	Lower_CI	Upper_CI	Table_Name
0	Have use of a car	Satisfaction with life	Yes	0	7.786737	None	7.744886	7.828589	Table 16
1	Have use of a car	Satisfaction with life	No	0	7.245154	None	7.132946	7.357362	Table 16

The highest means are 7.8 and 7.2 for people who have use of a car and are satisfied with life and are not satisfied with life respectively. This does not provide any meaning to the analysis and so we can just move on.

	Question	Group	Result	Sub_Result	Mean	Percentage	Lower_CI	Upper_CI	Table_Name
0	Have use of a car	Gender	Male	Yes	None	81.702147	80.206378	83.197916	Table 17
1	Have use of a car	Urban / rural area	Rural	Yes	None	85.957959	84.539418	87.3765	Table 18
2	Have use of a car	Wimd deprivation score	40% - 60% most deprived	Yes	None	81.233285	79.219055	83.247516	Table 19
3	Have use of a car	Wimd deprivation score	20% - 40% least deprived	Yes	None	85.684106	83.884166	87.484046	Table 19
4	Have use of a car	Wimd deprivation score	20% least deprived	Yes	None	88.747307	86.870084	90.624531	Table 19
5	Have use of a car	Wimd access to services score	20% most deprived	Yes	None	88.753025	87.34897	90.15708	Table 20
6	Have use of a car	Wimd access to services score	20% - 40% most deprived	Yes	None	83.459659	81.760598	85.158721	Table 20
7	Have use of a car	Employment status	In employment	Yes	None	90.326585	89.256897	91.396273	Table 21
8	Have use of a car	Limiting long-term illness	No limiting long term illness	Yes	None	82.34649	81.196032	83.496947	Table 22

This table shows all Percentages that are higher than 80%. Interestingly, a table 18 entry is shown with 86% of people in Rural areas have use of a car. Lets extract a table 18 visualization.



More people in rural areas have use of a car than in urban areas. Does this stem from a dissatisfaction with the public transport system. Further investigation is done in the conclusion.

Critical Evaluation of Identified Issues and Solutions

From the visualizations, If we increase the ease for the public transport systems to get to and from health care providers. The public satisfaction and usage of the public transport system could increase. As they will be more inclined to use the public transport instead of their own car. Solutions could be increasing amount of heath care transport systems. Limitations include emergency situations. Where rural areas may find it more difficult to get hold of public transport during stressful moments.

With Lower to middle-class residents predominantly using public transport, and rural areas being less satisfied with the state of the transport system. We can see we need to invest more into rural areas. Solutions could be more emphasis on keeping rural transport system areas cleaner and more maintenance on the vehicles. Limitations include not knowing exactly why the dissatisfaction occurs and if executives can do anything about it. This needs to be investigated more before making improvements.

Younger people feel safer and relatively satisfied with the public transport system. A lot of the times, they do not have a say in their usage of transport. This is governed by their elders. From the graphs we can see that middle-aged people are less satisfied and feel less safe towards the transport system. For increased younger generation usage, the younger and middle-aged generation opinions need to be taken into consideration. Solutions could include more initiative in safety and sections specifically for younger people. Limitations include stereotypes, especially if the middle-aged generation have these feelings due to issues they experienced decades ago.

Many improvements can be made to the public transport system in Wales. With accurate data and analysis, the steps taken can be backed with evidence and have the highest chances of successful for stakeholders and executives. Benefiting all.

(1 990 Words)

References

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 11 April 2024]

Torres, JRV., McArthur DP. (2024) Public transport accessibility indicators to urban and regional services in Great Britain. Available from: https://www.nature.com/articles/s41597-023-02890-w [Accessed 11 April 2024]

Merkert, R., Nelson JD. (2024) Workshop 4 report: Optimising the impact of technological innovation on achieving sustainable public transport outcomes. Available from:https://www.sciencedirect.com/science/article/abs/pii/S0739885923001373 [Accessed 11 April 2024]

Preston, J., Wreststrand, A. (2024) Workshop 1 report: Regulatory regimes: National and comparative regulation of public transport. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0739885923001336 [Accessed 11 April 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 11 April 2024]

StatsWales. () Welsh Index of Multiple Deprivation. Available from: https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation [Accessed 20 April 2024]

Limbu, S. (2023) What is Acorn?. Available from: https://acorn.caci.co.uk/what-is-acorn/ [Accessed 20 April 2024]

Code:

Using Python in Jupyter notebooks:

```
import sqlite3
# Connect to SQLite database
conn = sqlite3.connect('analysis.db')
c = conn.cursor()
# Use transaction control to manage database operations
conn.execute('BEGIN TRANSACTION;')
try:
    # Drop and recreate tables
    c.execute('DROP TABLE IF EXISTS Master_Table')
    c.execute('DROP TABLE IF EXISTS Result_Table')
    c.execute('DROP TABLE IF EXISTS Sub_Results_Table')
    c.execute('DROP TABLE IF EXISTS Group_Table')
    c.execute('DROP TABLE IF EXISTS Question_Table')

    c.execute('''
    CREATE TABLE IF NOT EXISTS Question_Table (
        Question_ID INTEGER PRIMARY KEY AUTOINCREMENT,
        Question TEXT)
    ''')
    c.execute('''
    CREATE TABLE IF NOT EXISTS Group_Table (
        Group_ID INTEGER PRIMARY KEY AUTOINCREMENT,
        "Group" TEXT)
    ''')
    c.execute('''
    CREATE TABLE IF NOT EXISTS Sub_Results_Table (
        Sub_Result_ID INTEGER PRIMARY KEY AUTOINCREMENT,
        Sub_Result TEXT UNIQUE)
    ''')
```

```
c.execute('''
CREATE TABLE IF NOT EXISTS Result_Table (
    Result_ID INTEGER PRIMARY KEY AUTOINCREMENT,
    Result TEXT,
    Sub_Result_ID INTEGER,
    Mean REAL,
    Percentage REAL,
    Lower_CI REAL,
    Upper_CI REAL,
    Table_Name TEXT,
    FOREIGN KEY(Sub_Result_ID) REFERENCES Sub_Results_Table(Sub_Result_ID))
''')
c.execute('''
CREATE TABLE IF NOT EXISTS Master_Table (
    Question_ID INTEGER,
    Group_ID INTEGER,
    Result_ID INTEGER,
    FOREIGN KEY(Question_ID) REFERENCES Question_Table(Question_ID),
    FOREIGN KEY(Group_ID) REFERENCES Group_Table(Group_ID),
    FOREIGN KEY(Result_ID) REFERENCES Result_Table(Result_ID))
''')

# Commit table creation
conn.commit()

# Processing function with enhanced error handling and data cleaning
def process_tables(table_dataframes):
    sub_results_seen = set()

    for i, (table_name, df) in enumerate(table_dataframes.items(), start=1):
        print(f"Processing {table_name}...")
```

```
question = df.iloc[:, 0].name
group = df.iloc[:, 1].name
c.execute('INSERT INTO Question_Table (Question) VALUES (?)', (question,))
question_id = c.lastrowid
c.execute('INSERT INTO Group_Table ("Group") VALUES (?)', (group,))
group_id = c.lastrowid

for index, row in df.iterrows():
    row = row.fillna(0)  # Replace NaN with 0
    sub_result, result = row[0], row[1]
    stat_type = df.columns[2]
    value1 = row[2]
    value2 = row[3]
    value3 = row[4]
    print(stat_type)
    print(value1)
    print(value2)
    print(value3)

    mean, percentage, lower_ci, upper_ci  = None, None, None, None
    if stat_type == "Mean":
        mean = value1
        lower_ci = value2
        upper_ci = value3
    elif stat_type == "%":
        percentage = value1
        lower_ci = value2
        upper_ci = value3

    if pd.notna(sub_result) and sub_result not in sub_results_seen:
        c.execute('INSERT INTO Sub_Results_Table (Sub_Result) VALUES (?)', (sub_result,))
```

```
            sub_results_seen.add(sub_result)

        sub_result_id = c.execute('SELECT Sub_Result_ID FROM Sub_Results_Table WHERE Sub_Result = ?', (sub_result,)).fetchone()

        sub_result_id = sub_result_id[0] if sub_result_id else None


        c.execute('''

        INSERT INTO Result_Table (Result, Sub_Result_ID, Mean, Percentage, Lower_CI, Upper_CI, Table_Name)

        VALUES (?, ?, ?, ?, ?, ?, ?)

        ''', (result, sub_result_id, mean, percentage, lower_ci, upper_ci, table_name))

        result_id = c.lastrowid


        c.execute('INSERT INTO Master_Table (Question_ID, Group_ID, Result_ID) VALUES (?, ?, ?)', (question_id, group_id, result_id))


    conn.commit()


  # Process tables
  process_tables(table_dataframes)
except Exception as e:
  conn.execute('ROLLBACK;')  # Roll back on error
  print(f"An error occurred: {e}")
finally:
  conn.close()
```

First graph code and extraction code from database:

```
import matplotlib.pyplot as plt


conn = sqlite3.connect('analysis.db')

c = conn.cursor()


query = """
```

```
SELECT
   qt.Question,
   gt.'Group',
   rt.Result,
   sr.Sub_Result,
   rt.Mean,
   rt.Percentage,
   rt.Lower_CI,
   rt.Upper_CI,
   rt.Table_Name

   FROM Master_Table mt
   LEFT JOIN Question_Table qt ON qt.Question_ID = mt.Question_ID
   LEFT JOIN Group_Table gt ON gt.Group_ID = mt.Group_ID
   LEFT JOIN Result_Table rt ON rt.Result_ID = mt.Result_ID
   LEFT JOIN Sub_Results_Table sr ON sr.Sub_Result_ID = rt.Sub_Result_ID

   WHERE Table_Name = 'Table 2'
"""

c.execute(query)
rows = c.fetchall()
columns = [description[0] for description in c.description]
df = pd.DataFrame(rows, columns=columns)

# Set plot title and axis labels
plt.figure()
plt.title(df["Question"].iloc[0])
plt.xlabel(df["Group"].iloc[0])
plt.ylabel("Values")
```

```
# Plot Mean, Lower_CI, and Upper_CI
plt.plot(df["Result"], df["Mean"], label="Mean", linestyle="-", color="b")
plt.plot(df["Result"], df["Lower_CI"], label="Lower CI", linestyle="--", color="g")
plt.plot(df["Result"], df["Upper_CI"], label="Upper CI", linestyle="--", color="r")

# Add a legend and grid
plt.legend()
plt.grid(True)

# Show the plot
plt.show()
```