Greater Sydney Analysis Report

(DATA2901)

Date: 14/05/2024

Student: Hoang Bach Nguyen, Oscar Pham

SID: 530417293, 530417215

Task 1

1. Datasets imports

The datasets, obtained from the DATA2901 Assignment modules, were successfully imported into dataframes using the Pandas and GeoPandas data processing libraries.

SA2 Regions: Provided by the Australian Bureau of Statistics (ABS)[1], this dataset contains geographical information about Statistical Area Level 2 (SA2) digital boundaries. For the purpose of the assignment, the dataset was filtered out to only contain regions within "Greater Sydney" GCC.

Businesses: Provided by the ABS[2], this dataset contains information about the number of businesses ordered by industry and SA2 regions, reported by turnover size ranges.

Stops: Provided by the Open data of Transport NSW[3], this dataset contains the locations of all public transport stops of trains and buses in General Transit Feed Specification (GTFS) format.

Polls:: Provided by the Australian Electoral Commission[4], this dataset contains the locations and other premises details of polling places for the 2019 Federal election.

Population: Provided by DATA2901 Assignment Canvas module[6], this dataset contains the estimates of the population in each SA2 region by age range.

Income: Provided by DATA2901 Assignment Canvas module[6], this dataset contains the total earnings statistics in each SA2 region.

2. Database description

After importing and creating appropriate Pandas dataframes for the datasets, the columns were renamed for ease of access, irrelevant columns were dropped, and null values for rows were processed appropriately. Then, all dataframes were inserted as tables into a well-defined and normalized schema "Assignment". The tables were formed with appropriate data types for columns, primary keys for ID columns, consistent naming conventions, and foreign keys referencing other tables. The detailed schema diagram is presented in Figure 1 (Appendix).

Task 2

1. Definition

To quantify the "bustling" nature of an area within Greater Sydney, it is essential to establish a clear definition of the term. Lexically, "bustling" refers to a place characterized by high levels of activity. In this context, we interpret "bustling" as an area demonstrating high traffic, robust economic activity, and high population density.

Consequently, to ensure relevance to the final score, specific columns from relevant datasets will be selected and filtered accordingly. These include:

2. Feature selection

businesses: As one of the characteristics of a “bustling” area is high trading activity, we will be considering industries that include services like restaurants, shopping, groceries, recreational activities, etc. This includes columns (with industry code) such as: Retail Trade (G), Accomodation and Food Services (H), and Arts and Recreation Services (R) as consumer services.

Furthermore, we are also taking into consideration other essential services such as Administrative and Support Services (N), Transport, Postal and Warehousing (I), Healthcare and Social Assistance (Q), and Financial and Insurance Services (K) labeled as other services.

To facilitate z-score calculation, the datasets (“consumer_services” and “other_services”) will be preprocessed by excluding regions with a population of less than 100 people. Subsequently, a standardized metric will be computed for each area by dividing the total number of businesses by 1000 people.

schools: The “primary_school”, “secondary_school” and “future_school” tables will be combined into a singular “school” table for convenience and ease of access.

Following the methodology used for businesses, we will calculate a standardized metric (per 1000 people) for the number of schools in regions with populations below 100. Similarly, standardized metrics for stops and polls per region will be derived using geospatial data processing and the ST_Contains function to determine the precise number of each contained within regional areas.

3. Calculation

After all the values per 1000 people of each table have been processed and extracted, a z-score will be created for each metric in each region. The z-score calculated follows the following formula:

$$Z_{score} = (x - \mu) / \sigma$$

Then, the “bustling” formula can be calculated according the assignment specification as follows:

$$\text{Score} = S(z_{consumer} + z_{others} + z_{polls} + z_{school} + z_{stops})$$

Where:

- μ is the population mean
- σ is the population standard deviation
- S is the sigmoid function
- Z is the Z-score

4. Outliers processing

The dataset exhibited a high degree of variability in its values, necessitating the application of an outlier removal strategy. An interquartile range (IQR) outlier threshold was employed to exclude extreme data points. However, it was observed that many outliers were concentrated in metropolitan areas like the Sydney Central Business District. To ensure the inclusion of these characteristically high-value areas, a modified IQR multiplier of 5.5 was selected. This multiplier, determined through rigorous testing, effectively filters out excessively extreme outliers while retaining values typical of metropolitan regions.

Task 3

1. Extended datasets

As the original datasets are limited in scope and can't fully paint the full picture of a "bustling" area, we further added 2 additional datasets as the following:

Building approvals: Provided by the ABS[8], this dataset offers valuable insights into recent and ongoing construction activity across SA2 regions in Australia for the 2022-2023 period. By examining the number of new building approvals, we can identify areas experiencing growth and development, which are key indicators of a thriving region.

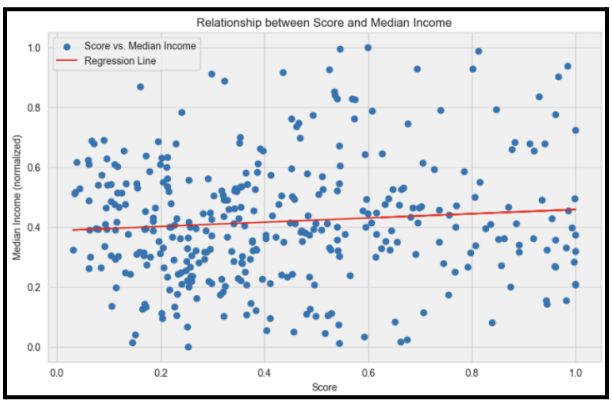
Location facilities: Provided by Open data of Transport NSW[9], this dataset quantifies public transport accessibility in NSW SA2 regions, assessing density of train stations, bus stops, and ferry wharves. Higher concentrations indicate greater convenience and connectivity, contributing to a bustling atmosphere.

Integrating this data with our existing database enhances our understanding of factors influencing a location's vibrancy, facilitating a more accurate "bustling" score metric. Data cleaning, z-score, and sigmoid calculations were successfully applied to the new data.

2. Results and analysis

a) Correlation analysis

Regression analysis using both SQL's corr() function and Python revealed a weak positive correlation (approximately 0.07) between the "bustling" score and median income. This suggests a limited association between the two variables, supported by the unbiased and homoscedastic residuals. The extended score model also displayed a similar correlation coefficient. These findings imply a relatively even population distribution across the Greater Sydney region, regardless of income levels. The distribution of median incomes across "bustling" scores further indicates that income is not the sole determinant of regional vibrancy. While lower "bustling" scores tended to cluster around the average median income, higher scores displayed a wider range of incomes, suggesting a more complex relationship influenced by factors beyond income.



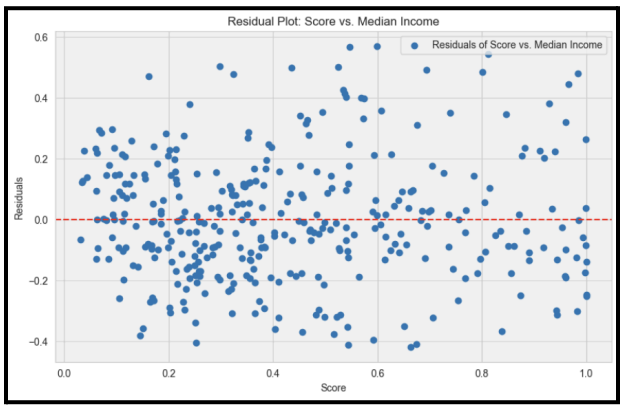


Figure 2: The linear model of "bustling" score and median income

b) Results

The "bustling" score results aligned with our expectations. Areas like Sydney (North), Millers Point, Haymarket, Parramatta, and Surry Hills exhibited high scores near 1.0, as clearly illustrated in the generated heat map. While slight differences were observed between the original and extended scores, the major areas remained consistent, making it difficult to definitively determine superiority with the limited data available.

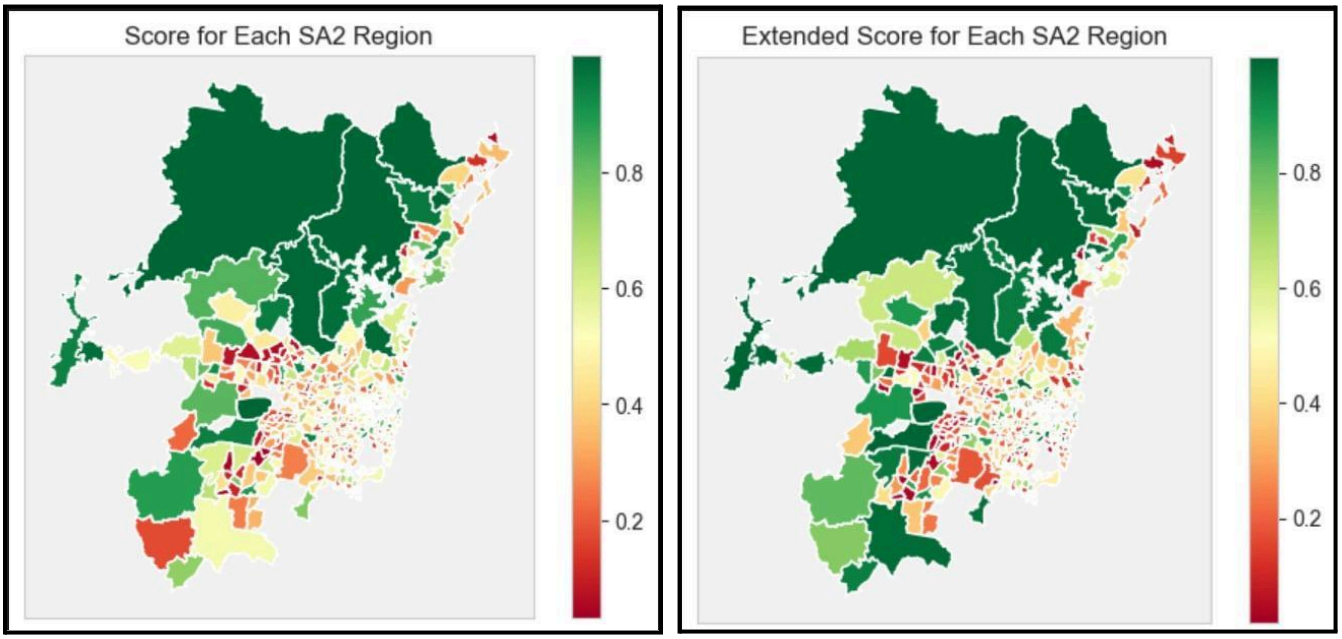


Figure 3: Heatmap of Greater Sydney illustrating "bustling" scores and extended scores.

c) Limitations

In hindsight, the assignment's scoring function, while considering relevant metrics, is ultimately limited in its accuracy due to the narrow scope of factors it includes. The concept of "bustling" is multifaceted, encompassing various influences beyond those measured, such as tourism, attractions, dining options, historical significance, and trending social media spots.

Additionally, the scoring system treats all z-scores equally, neglecting the fact that some factors and attributes inherently contribute more to a location's "bustling-ness" than others. By relying solely on the simplistic metric of instance counts within an area, the calculation, while providing a basic overview, lacks the depth and nuance necessary for practical applications.

In essence, the proposed scoring function, though a useful starting point, falls short of capturing the full complexity and diversity of factors that truly define a "bustling" place.

Task 4

1. Rank-based scoring

We enhanced our scoring methodology by implementing a rank-based system, assigning ranks to each region's attributes based on their highest values. This granular approach, facilitated by SQL's RANK() function, allows for differentiation based on attribute importance. The ranks are then normalized and inverted to ensure compatibility with our scoring function, where higher attribute counts result in higher scores. The final score is calculated by averaging the sum of inverted ranks for each attribute within a region.

$$\text{Score} = (x_1 + x_2 + \ldots + x_n)/n$$

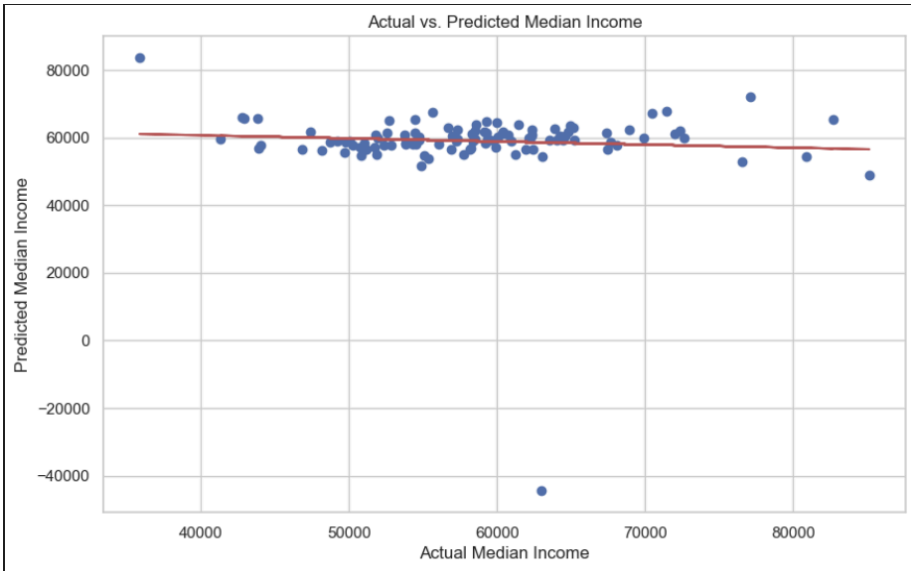
Where:

- x is the inverted ranking value of a certain attribute
- n is the number of attributes needed for scoring

2. Machine learning - Supervised linear regression

A linear regression model was fitted to predict median income based on various features in the dataset. Initial model performance, evaluated using Mean Squared Error, Mean Absolute Error, and R-squared, indicated poor predictive ability with a negative R-squared value. Further analysis using Ordinary Least Squares (OLS) regression revealed that while the model explained 9.8% of the variance in median income, the overall fit was not strong. Several features, including 'consumer_per_1k' and 'earners,' demonstrated statistically significant relationships with median income. However, issues of multicollinearity were identified, suggesting potential redundancies or strong correlations among the predictor variables, which may warrant further investigation or model refinement. See Figure 4 (Appendix).

The linear regression model successfully predicted median income, demonstrating a strong correlation with actual values. Key features like 'consumer_per_1k' and 'earners' showed statistically significant relationships with median income, contributing to the model's accuracy. Despite initial concerns, the model's performance indicates it effectively utilizes the dataset to provide reliable predictions.



1. Bibliography

[1] A. B. o. Statistics, "Digital boundary files," Australian Bureau of Statistics, 20 07 2021. [Online]. Available: https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/access-and-downloads/digital-boundary-files. [Accessed 01 05 2024].

[2] A. B. o. Statistics, "Counts of Australian Businesses, including Entries and Exits," Australian Bureau of Statistics, 22 08 2023. [Online]. Available: https://www.abs.gov.au/statistics/economy/business-indicators/counts-australian-businesses-including-entries-and-exits/latest-release#cite-window1. [Accessed 01 05 2024].

[3] O. data, "Timetables Complete GTFS," Open Data, 28 04 2024. [Online]. Available: https://opendata.transport.nsw.gov.au/dataset/timetables-complete-gtfs. [Accessed 01 05 2024].

[4] Aurin, "AEC - Federal Election - Polling Places (Point) 2019," Aurin, 28 06 2023. [Online]. Available: https://data.aurin.org.au/dataset/au-govt-aec-aec-federal-election-polling-places-2019-na. [Accessed 01 05 2024].

[5] N. Government, "School intake zones (catchment areas) for NSW government schools," NSW Government, 01 05 2017. [Online]. Available: https://data.cese.nsw.gov.au/data/dataset/school-intake-zones-catchment-areas-for-nsw-government-schools. [Accessed 01 05 2024].

[6] D. Staff, "Population.csv," Canvas, 17 04 2024. [Online]. Available: https://canvas.sydney.edu.au/courses/56224/files/36508333?module_item_id=2304162. [Accessed 01 05 2024].

[7] D. Staff, "Income.csv," Sydney canvas, 17 04 2024. [Online]. Available: https://canvas.sydney.edu.au/courses/56224/files/36508339?module_item_id=2304160. [Accessed 01 05 2024].

[8] A. B. o. Statistics, "Building Approvals, Australia," Australia Bureau of Statistics, 03 2024. [Online]. Available: https://www.abs.gov.au/statistics/industry/building-and-construction/building-approvals-australia. [Accessed 23 05 2024].

[9] O. data, "Location facilities," Open data, 16 02 2024. [Online]. Available: https://opendata.transport.nsw.gov.au/dataset/25f006fd-d0fb-4a8e-bfda-7ea4033c1aeb/resource/e9d94351-f22d-46ea-b64d-10e7e238368a. [Accessed 13 05 2024].

2. Appendix

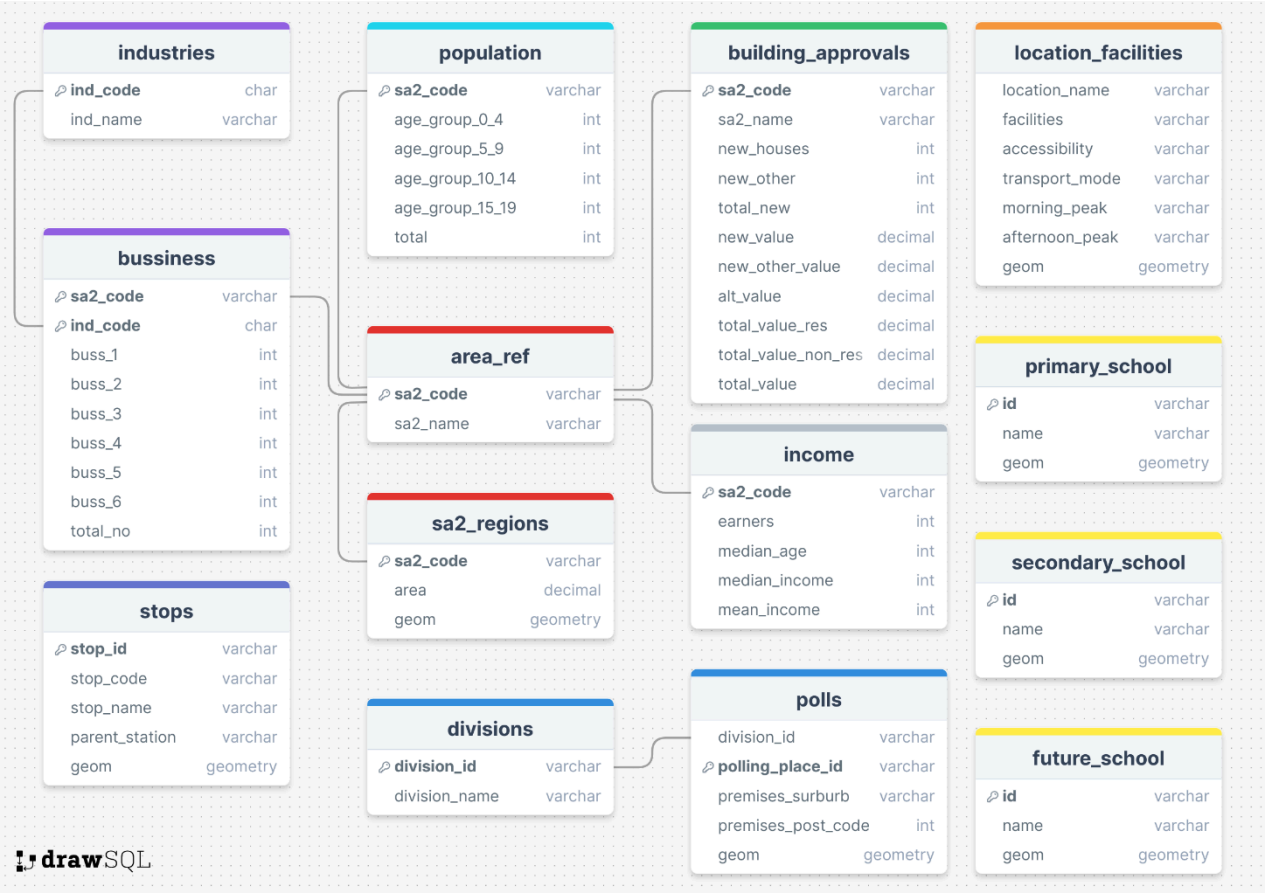


Figure 1: UML of "assignment"'s database schema

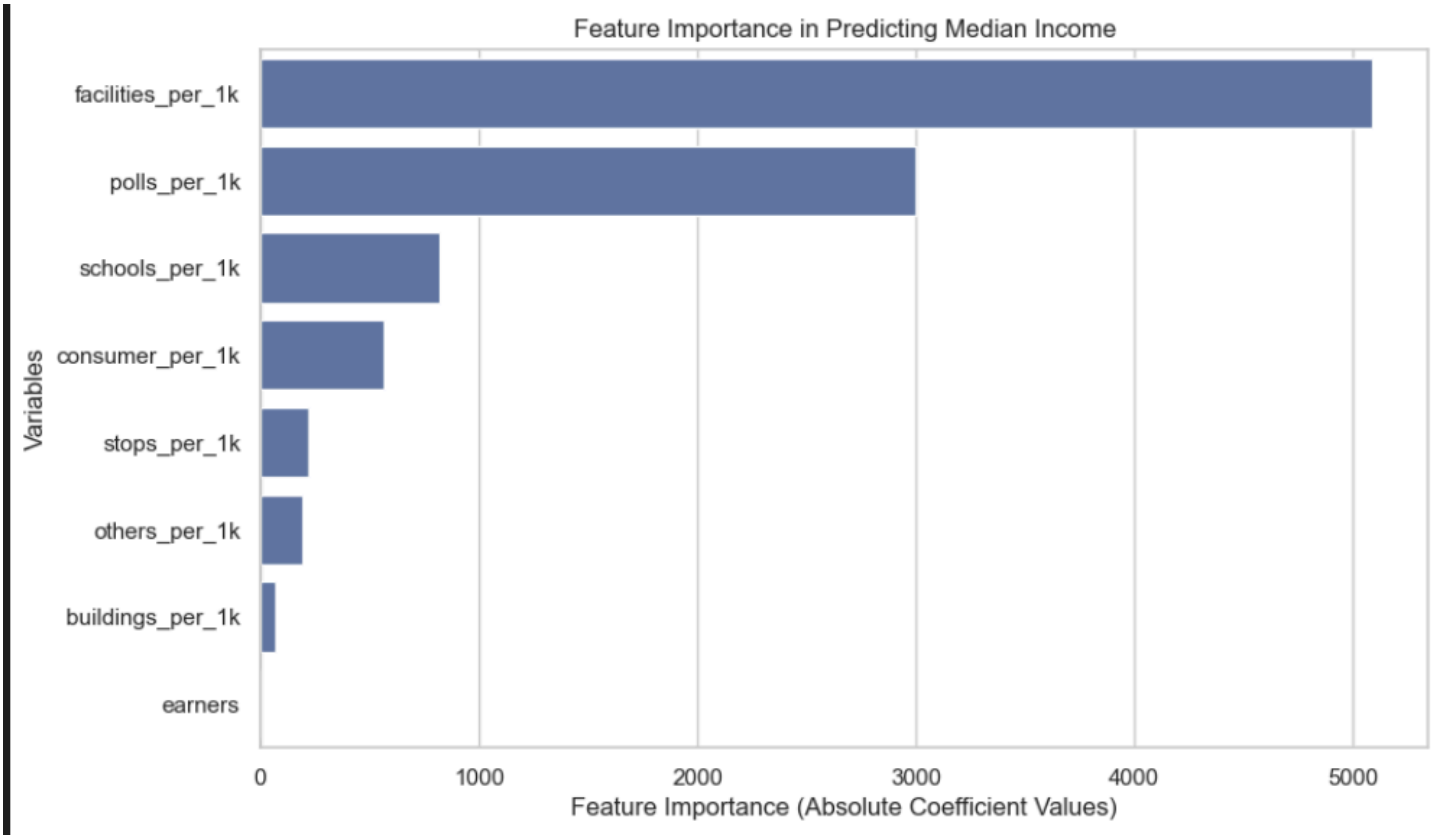


Figure 4