DATA2001: Data Science, Big Data, and Data Variety

Analyzing the 'Bustling' Metric of Greater Sydney: A Geospatial and Statistical Approach

Group Number: LAB04-G01
Student IDs: 530532312, 520268067

May 14, 2024

Abstract

This analysis aims to assess the economic activity and vibrancy of Statistical Area Level 2 (SA2) regions in Greater Sydney by integrating and analyzing various datasets to determine the 'bustling' scores of these areas.

Contents

1 Introduction

Australia's population is smaller and more dispersed than that of many Western countries, but is heavily concentrated in coastal cities, particularly Sydney and Melbourne. In the Greater Sydney region, differences in service availability and economic activity across regions necessitate a nuanced understanding of urban vitality and function. This report aims to fill this gap by introducing the 'prosperity' indicator, which is a composite indicator designed to assess economic and social activity in the Greater Sydney Secondary Statistical Area (SA2). By integrating multiple datasets related to business operations, public transport accessibility, and demographic indicators, our analysis seeks to identify patterns and potential associations between the Prosperity score and various socio-economic factors.

2 Dataset Description

2.1 Data Origins and Preparation

2.1.1 SA2 Data ('SA2_2021_AUST_GDA2020.shp')[1]

Purpose: Provides the geographical boundaries for Statistical Area Level 2 (SA2) regions, intended to represent communities that interact socially and economically.
Preparation: This dataset was filtered to include only regions identified by the "1GSYD" code, representing Greater Sydney. A table named `sa2_greater_sydney` was created, setting `sa2_code21` as the primary key. This preparation links spatial boundaries to socio-economic data, crucial for regional analysis.

2.1.2 Business Data ('Business.csv')[2]

Purpose: Contains detailed records of business operations by industry and location, essential for assessing economic activity within SA2 regions.
Preparation: Entries lacking specific industry codes or location details were excluded, focusing exclusively on businesses operating within the "1GSYD" code. This filtration ensures accuracy in measuring business density and economic vibrancy specific to Greater Sydney.

2.1.3 Income Data ('Income.csv')[3]

Purpose: Provides regional income distribution, valuable for analyzing economic disparities and affluence within SA2 areas.
Preparation: Data was refined to only include entries for the Greater Sydney area, marked by "1GSYD", aligning income statistics with regional boundaries for accurate economic profiling.

2.1.4 Transport Stops Data ('Stops.txt')[4]

Preparation: Lists public transport stop locations across Sydney, including details on timetables and routes, critical for assessing transportation accessibility.

2.1.5 Polling Stations Data ('PollingPlaces2019.csv')[5]

Purpose: Shows locations of polling stations used during the 2019 federal election, indicative of civic infrastructure and public resource distribution.
Preparation: Only stations within Greater Sydney were included, and address columns were consolidated to streamline data integration and spatial analysis.

2.1.6 Population Data ('Population.csv')[3]

Purpose: Offers demographic statistics crucial for understanding population dynamics and segmentation in Greater Sydney.
Preparation: Anomalies were removed and data simplified to focus on key demographic attributes (age groups and genders) relevant to our bustling score analysis.

2.1.7 Schools Data ('catchments_primary.shp', 'catchments_secondary.shp', 'catchments_future.shp')[6]

Preparation: Merged into a single dataset 'combined_school' and standardized for consistent formatting, facilitating educational infrastructure analysis across different schooling levels.

2.2 Personal Data Sets

Foundation Facility Points (.gpkg)[7]: Contains spatial locations in point format of five facility types across Australia: Private Hospitals, Public Hospitals, Aged Care Facilities, Educational Facilities, and Emergency Management Facilities **Healthcare Facilities (.shp)[8]:** Features all data within the Greater Sydney area where tags related to healthcare are present. This includes facilities labeled as doctors, dentists, clinics, hospitals, and pharmacies. **Academical Facilities (.geojson)[9]:** Encompasses all OpenStreetMap features in the Greater Sydney area matching educational facility tags. This includes kindergartens, schools, colleges, and universities.

3 Database Schema

Our database schema is strategically structured to enable an in-depth analysis of demographic, economic, and geographic data specific to the Greater

Sydney region. All tables share relationships with sa2_greater_sydney spatial data, having a unique identifier associated that to each SA2 region.

3.1 Tables Overview

- **sa2_greater_sydney**: Contains geospatial data essential for regional planning and analysis. *Primary key: sa2_code21.*
- **modified_polling**: Stores refined data on polling places, important for electoral analysis. *Foreign key linked to sa2_code21.*
- **sydney_population**: Offers demographic data, pivotal for understanding population dynamics. *Foreign key linked to sa2_code21.*
- **sydney_income**: Provides income statistics by region, supporting economic status evaluations. *Foreign key linked to sa2_code21.*
- **sydney_businesses**: Catalogs business information by industry and earnings. *Foreign key linked to sa2_code21.*

The database employs a spatial index to enhance query performance, efficiently querying geographical data. This dual indexing approach ensures quick data access and efficient processing of both spatial and demographic queries. Refer to Appendix B for a demonstration of the spatial index performing a spatial join between the 'sa2_greater_sydney' and 'modified_polling' tables using the 'ST_Intersects' function in the 'WHERE' clause. This join matches each polling place to its corresponding SA2 region based on their spatial intersection.

3.2 Database Schema Diagram

For the full Schema Diagram, please refer to Appendix 1.

4 Results Analysis

4.1 Score Analysis

The bustling score quantifies the vibrancy and economic activity of Sydney's SA2 regions, aiding urban planners and policymakers in resource allocation. The score is calculated using the formula:

$$\text{Score} = S(z_{\text{business}} + z_{\text{stops}} + z_{\text{polls}} + z_{\text{schools}} + z_{\text{addition\_datas}})$$

where S is the sigmoid function, normalizing the outcomes to a range of 0 to 1, and z represents standardized z-scores for each metric, adjusted for population and regional differences.

4.1.1 Computation of Each Metric

Each component of the bustling score is computed to reflect local socio-economic factors and geographical nuances:

z_business We assess economic activity by focusing on business density within the "Professional, Scientific, and Technical Services" industry. To account for variations in business size and their respective economic impacts, we apply a weighting scale from 1 (for businesses generating less than $50,000 annually) to 6 (for businesses generating $10 million or more annually). This scale helps normalize the data, ensuring that larger businesses have a proportionally appropriate impact on the calculated economic vibrancy of a region. The formula for calculating the z-score of business density is as follows:

$$z_{\text{business}} = \frac{\text{business density} - \text{mean business density}}{\text{standard deviation}}$$

z_stops The public transport accessibility is quantified by calculating stops per square kilometre. The stops data is spatially joined to SA2 regions using the $ST\_Contains$ function. The z-score is then computed to standardize this across all regions.

z_polls and z_schools Similar methods are used for polling places and schools, with z-scores helping to identify areas with high or low infrastructure relative to regional averages.

4.1.2 Integration of Additional Datasets

The bustling score was refined to include additional datasets for a more accurate representation of urban vitality in Greater Sydney. The Foundation_Facility_Points[7] dataset was integrated after excluding educational buildings, focusing on New South Wales facilities, removing duplicates, and establishing a spatial index. This pre-processing enabled the normalization of facility densities and the use of z-scores to highlight areas with significantly high or low densities.

Additionally, the Healthcare Facilities[8] dataset was enhanced by removing entries with null values and duplicates, and filling in missing 'facility_type' entries. A weighting system prioritized critical facilities like hospitals, clinics, and pharmacies to more accurately reflect their impact on regional accessibility. This approach allowed for a refined computation of healthcare facility densities using z-scores.

Lastly, the Academical Facilities[9] dataset was incorporated, filtering out primary and kindergarten schools to avoid metric overlap. Higher education and other significant institutions were weighted by their community

impact, facilitating a detailed analysis of educational facility distribution. This helped improve understanding of urban planning dynamics by identifying areas that are either well-served or deficient in educational services

4.2 Analysis of Distribution and Trends

4.2.1 Overall Distribution and Visual Support

The bustling scores across Greater Sydney exhibit a skewed distribution, with central urban areas showing higher vibrancy scores. Figure 5 in Appendix E shows a map overlay of bustling scores which considered all the datasets we got. **You can also access the interactive map in the Jupyter Notebook or in the folder, ther are two HTML file that contain the map.**Notably, regions such as Parramatta and the Sydney CBD exhibit the highest scores due to dense business activities and robust public infrastructure. Conversely, peripheral areas like Niagara Park show lower scores, reflecting fewer economic activities and infrastructure.

4.2.2 Trends and Regional Highlights

Figure 2 and Figure 3in Appendix B and C provides two histogram of bustling scores representing the frequency of scores over all the regions. Notably that in the current score distributions graph (before considering the additionally datasets), there are prominent peaks around scores of 0.2, 0.4, and another near 0.7. However in the histogram which includes the additional datasets.The highest bar at the left end of the histogram suggests that a substantial number of scores are clustered near 0, indicating that many participants or subjects received very low scores. There is also a significant frequency of scores near 1.0, showing that aside from many low scores, a substantial number of scores were near perfect.

4.2.3 Conclusion of Results

Figure 4 and Figure 5 in Appendix D and E shows a map overlay of bustling scores which considered all the datasets we got.
Impact of Additional Data: Figure 5, which includes additional datasets, shows a more graded and nuanced distribution of scores. This suggests that the inclusion of additional data allows for a more comprehensive assessment of what makes an area bustling, considering aspects like healthcare availability, educational facilities, and other infrastructural elements that contribute to a region's liveliness and functionality.
Conclusion: The differences between the two maps highlight the importance of data selection in geographical analyses. By incorporating a broader range of datasets, urban planners and policymakers can gain a richer, more accurate picture of urban dynamics, which in turn can inform more targeted and

effective urban development and policy initiatives.

5 Correlation Analysis

5.1 Summary of Statistical Test

This section details the correlation analysis that was done to determine the association between median income levels in SA2 regions and bustling scores, a metric used to assess the vibrancy and economic activity of those locations. The statistical metric that quantifies the degree to which two variables are associated in a linear fashion across a dataset was employed in the analysis

5.2 Results of the Analysis

The Pearson correlation coefficient calculated from the data was 0.142038. This value indicates a positive but very weak linear relationship between the bustling scores and median incomes of SA2 regions. The detail value and the scatter Plot, you can refer to 6.

5.2.1 Strength of Correlation

Weak Positive Correlation: The correlation coefficient of approximately 0.142 suggests that there is a slight positive relationship between the bustling scores and median incomes. This implies that as bustling scores increase, there is a tendency, albeit weak, for median incomes to also increase. Limited Predictive Power: The low magnitude of the correlation coefficient suggests that bustling scores have limited predictive power regarding median incomes. This means that other factors not captured by bustling scores might be influencing the income levels more significantly.

5.3 Correlation Conclusion

The correlation analysis, while statistically showing a positive correlation, suggests that the prosperity score itself is not strongly associated with median income levels in the SA2 region of Greater Sydney. This analysis supports the view that urban vitality and economic prosperity, as represented by income levels, is influenced by a number of factors.

6 Conclusion

This comprehensive study has investigated the bustling scores across the Statistical Area Level 2 (SA2) regions of Greater Sydney, aiming to uncover the dynamics of urban vibrancy and economic activity. The utilization of

various datasets, including business operations, income distributions, transport stops, and demographic data, has allowed for a nuanced analysis of what factors contribute to the bustling nature of these regions.

References

[1] A. B. of Statistics, "Statistical area level 2," 2021, accessed on: 14 May 2024. [Online]. Available: https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/main-structure-and-greater-capital-city-statistical-areas/statistical-area-level-2

[2] ——, "Counts of australian businesses, including entries and exits," 2023, accessed on: 14 May 2024. [Online]. Available: https://www.abs.gov.au/statistics/economy/business-indicators/counts-australian-businesses-including-entries-and-exits/latest-release#data-downloads

[3] DATA2001, "Sydney income," 2024, accessed on: 14 May 2024. [Online]. Available: https://canvas.sydney.edu.au/courses/56224/modules

[4] T. for NSW, "Timetables complete gtfs," 2023, accessed on: 14 May 2024. [Online]. Available: https://opendata.transport.nsw.gov.au/dataset/timetables-complete-gtfs

[5] A. E. Commission, "Aec - federal election - polling places (point) 2019," 2023, accessed on: 14 May 2024. [Online]. Available: https://data.aurin.org.au/dataset/au-govt-aec-aec-federal-election-polling-places-2019-na

[6] N. D. of Education, "School intake zones (catchment areas) for nsw government schools," 2024, accessed on: 14 May 2024. [Online]. Available: https://data.cese.nsw.gov.au/data/dataset/school-intake-zones-catchment-areas-for-nsw-government-schools

[7] G. Australia, "Foundation facilities points," 2021, accessed on: 14 May 2024. [Online]. Available: https://data.gov.au/dataset/ds-ga-d9f88f7b-2bec-476b-b907-ef01109f8b3a/details?q=hospitals

[8] O. contributors, "Australia health facilities (openstreetmap export)," 2024, accessed on: 14 May 2024. [Online]. Available: https://data.humdata.org/dataset/hotosm_aus_health_facilities

[9] ——, "Australia education facilities (openstreetmap export)," 2024, accessed on: 14 May 2024. [Online]. Available: https://data.humdata.org/dataset/hotosm_aus_education_facilities

A Full Schema Diagram

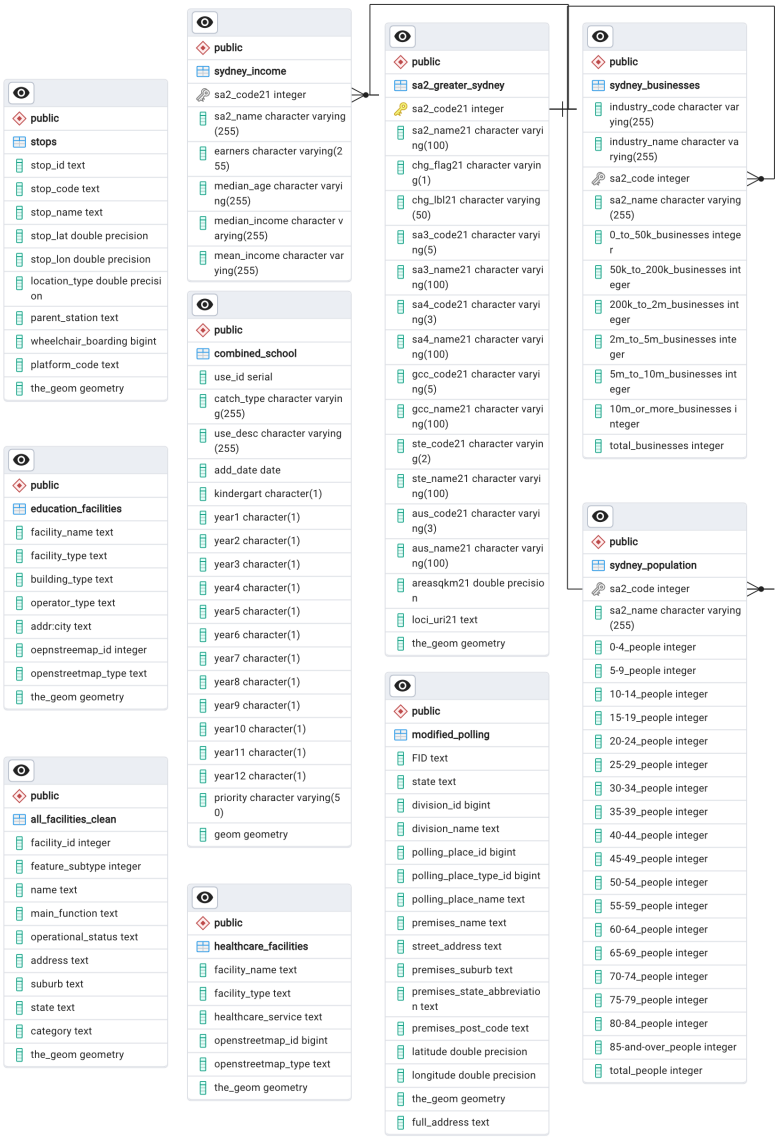

Figure 1: Full Schema Diagram

B Current score distributions histogram

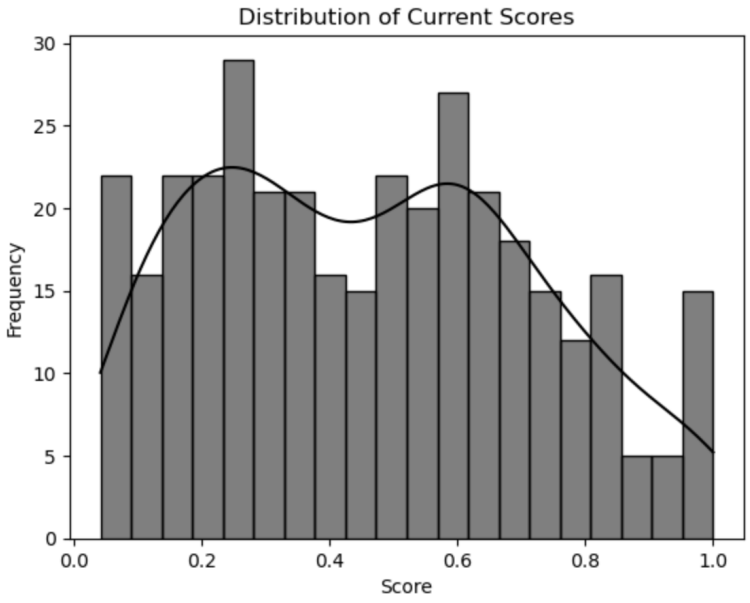

Figure 2: Current score distributions histogram

C Final score distributions histogram

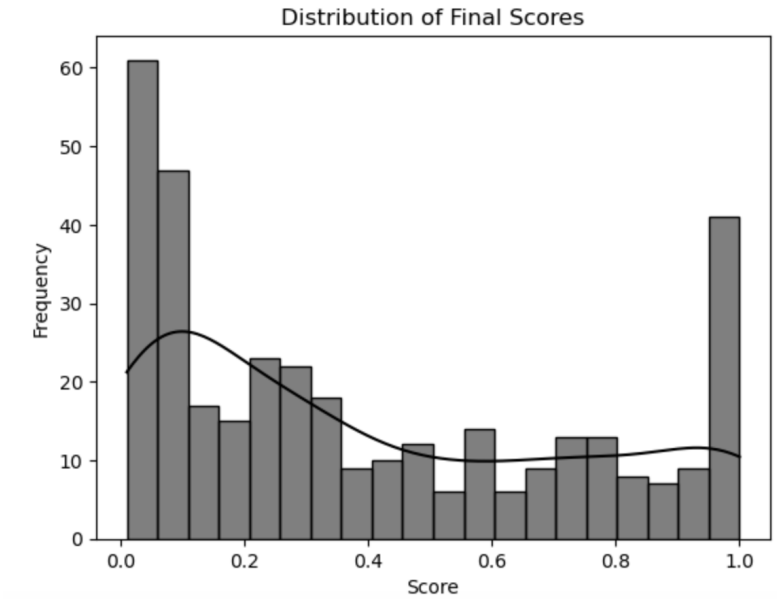

Figure 3: Final score distributions histogram

D Current Bustling Map

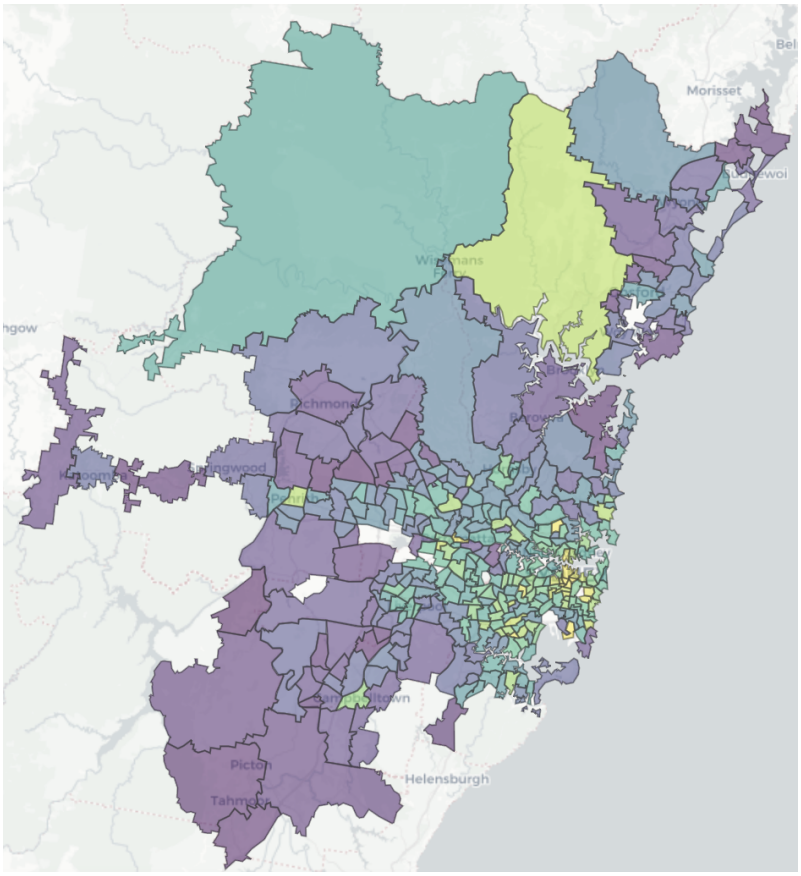

Figure 4: Current Bustling Map

E Final Bustling Map

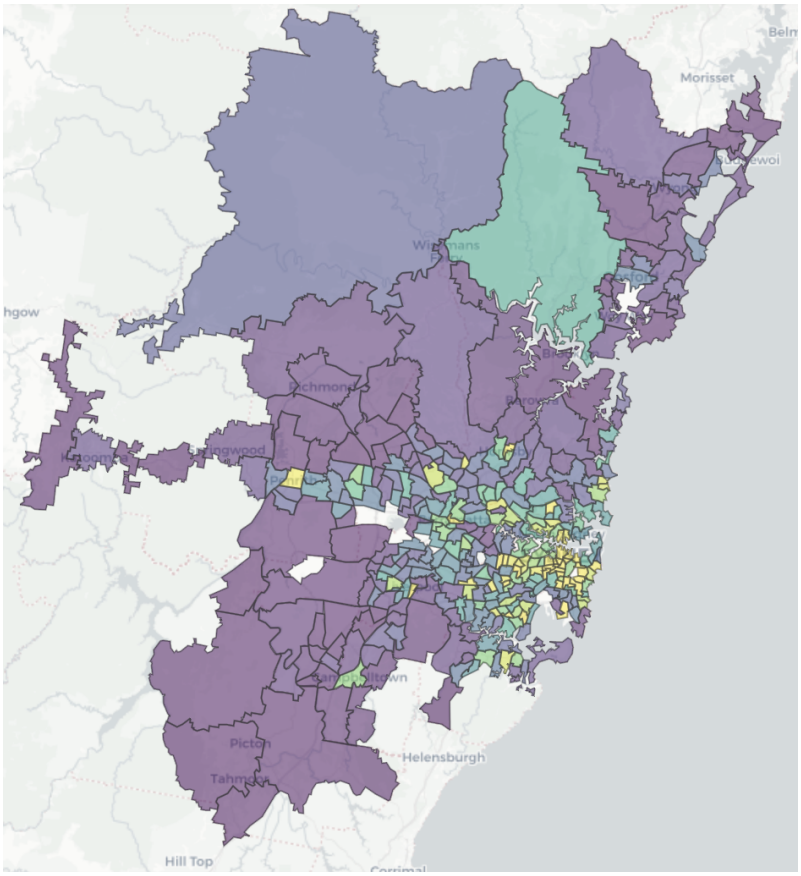

Figure 5: Final Bustling Map

F Scatter Plot Correlation

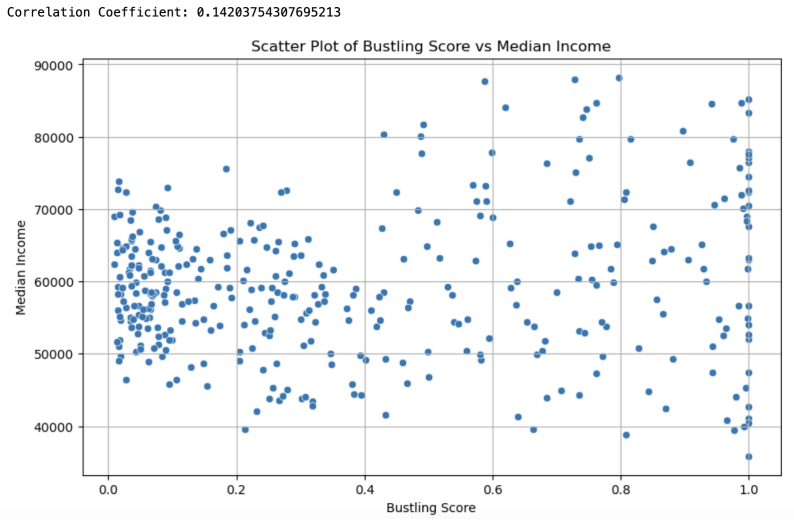

Figure 6: Scatter Plot Correlation