Neighbourhood Cyclability Analysis Report

Dataset Description

Seven datasets were used in this analysis: “StatisticalAreas.csv”, “Neighbourhoods.csv”, “CensusStats.csv”, “BusinessStats.csv”, “BikeSharingPods.csv”, “SA2_2016_AUST.shp” and an additional dataset.

- ### Data Sources

The first five datasets are census-based csv datasets provided on Canvas. The data sources of the shapefile and the additional synthetic JSON dataset provided on EdStem[2] are from Australian Bureau of Statistics[1] (Abs.gov.au, 2019) and respectively.

- ### Data Importing

The python package csv was used to read the data from the csv files into the dictionaries.

The python package shapefile was used to read the shapefile. A function was defined to read the shapefile into a pandas dataframe with the geometry information. A second function was defined to plot the map based on the data read from the shapefile.

The python package requests was used to access the Web API. The data was converted into the JSON format. The JSON file was read into a csv file.

- ### Data Cleaning

After examining each dataset, the two defects detected were: missing values and the Null values.

The header row of each dataset was extracted apart from the shapefile. Python package pandas was used to loop through the header rows and fill in all the missing values with zeros. The cleaned datasets were written into new csv files and then they were read into dictionaries again. No cleaning was required for the shapefile. With regard to the JSON file, for each row containing Null values, the row was deleted. The latitude and the longitude information in the “BikeSharingPods.csv” and the additional dataset was paired up (ie (longitude, latitude)) to form a new variable points, so that each point represents the geometry of a specific area. The points were stored in the table with the spatial type epsg: 4326.

Database Description

Python package psycopg2 was used to connect to the postgresql database in the school server (soit-db-pro-1.ucc.usyd.edu.au). Fourteen schemas were written to integrate the data which is shown in the Entity-Relationship Diagram below:

- Indices

Two **spatial indices** “boundaries_idx” and “point_idx” were created to improve the proficiency of the spatial join.

A **covering index** “area_name” was created when data was grouped by “area_name” to aggregate the bikepods_count and the BSAscore. This index is useful as itself contains all attributes, so that only one index is needed to answer the entire query.

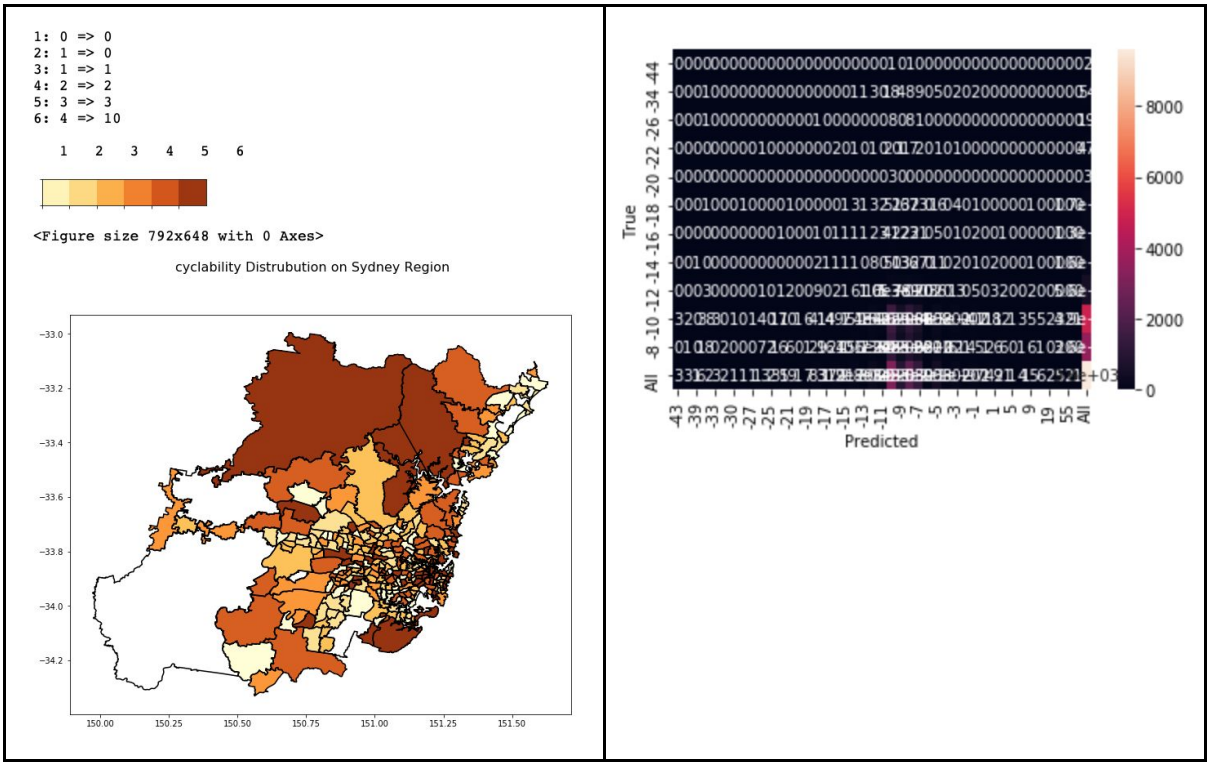
A **clustered index** “name” was automatically generated and was served as the primary key in the BikeSharingPods table. As shown in the diagram that the primary key is “area_id” for all tables, except for BikeSharingPods and bikepods_count, so the clustered index is useful as it sorts the index entries and the rows in the same order. It is very efficient when new tables are created.

Cyclability Analysis

The formula employed to compute the cyclability score is shown below and was adjusted on the basis of the given formula:

cyclability = z(population density)+z(dwelling density)+z(service balance)+z(bikepod density)+z(bike storing ability)

An overview of the cyclability results is shown in the heatmap and the validity of the analysis model is shown in the Confusion Matrix:



The additional BSA (Bike Storing Ability) score is the difference of “num_bikes” from table BikeSharingPods and the “Bike_Parking_spaces” from table ExtraData. The Machine Learning algorithm Confusion Matrix was deployed to check the validity of the cyclability

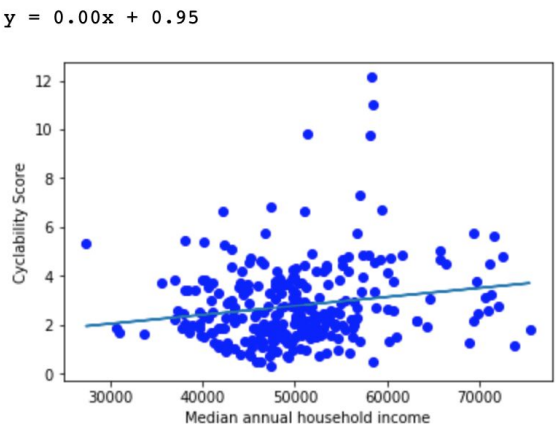
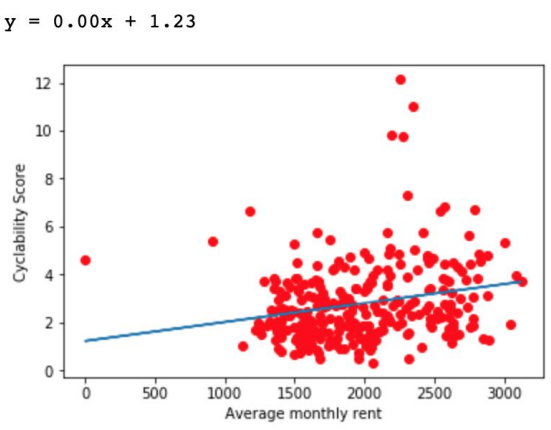
analysis. The diagonal of the Matrix represents the number of points where the predicted value matches the true value.[3] (Scikit-learn.org, 2019) A good model would have the diagonal to be brighter in colour. In this Confusion Matrix, brighter colours correspond to higher values according to the scale. The higher the diagonal values, the better the prediction. As shown in the Confusion Matrix, the brighter coloured boxes are positioned off the diagonal. Consequently, the model of this cyclability analysis was not valid.

The calculation of the z score of service balance is shown below: Firstly, the *weighted average* of each attribute in the BusinessStats was calculated by taking the average of each attribute and divide each of the averaged value by the average number of businesses. Secondly, each weighted average value was multiplied by the attribute itself. The service score was calculated by summing up each attribute and the aggregation was divided by the number of rows from BusinessStats. Thirdly, the absolute value of the difference of the service score and its mean was taken and was then divided by its standard deviation. Z score of every attribute was calculated in a similar fashion.

In conclusion, it is evident in the heatmap that Pymble has the best cyclability with the highest score of 0.89 (to 2 dp) which is also demonstrated in our statistical analysis.

Correlation Analysis

An overview of the correlation analysis is shown in the two scatter plots:

Cyclability vs Median annual household income	Cyclability vs Average monthly rent
	

As shown in the plots, correlation was not found between the cyclability score and either of the two attributes (ie the median household income in the given suburbs and the average weekly rent in the neighbourhoods.

Appendix

Bibliography

1. Abs.gov.au. (2019). *1270.0.55.001 - Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, July 2016*. [online] Available at: https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument#Data [Accessed 21 May 2019].

2. [online] Available at: http://soit-app-pro-4.ucc.usyd.edu.au:3000/api/v1/json [Accessed 16 May 2019].

3. Scikit-learn.org. (2019). *Confusion matrix — scikit-learn 0.21.1 documentation*. [online] Available at: https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html [Accessed 20 May 2019].