DATA2001 Cyclability Analysis Report

Group 121

May 24, 2019

1 Dataset Description

We used the following datasets that were provided to us:

StatisticalAreas - A csv file containing areas of Sydney according to ABS area boundaries. It holds an area ID number, area name and a parent ID for each area. Sourced from the ABS.

Neighbourhoods - A csv file containing additional information on each of the areas from StatisticalAreas specifically land area, population, number of dwellings, and number of businesses. Sourced from the ABS.

CensusStats - A csv file containing additional information on each of the areas specifically median household income, and average monthly rent. Sourced from the ABS.

BusinessStats - A csv file containing additional business related information on each area. This includes the total number of businesses overall in the area, and individual counts for retail trade, accommodation and food, and health care. Only the data for the total number of businesses per area was used in the cyclability score. BusinessStats was sourced from the ABS.

BikeSharingPods - A csv file containing synthetic data on bike sharing pods located around Sydney. This holds a station ID, the number of bikes, the number of scooters, latitude, longitude, and a description of each bike-sharing pod.

Two more datasets were retrieved from external sources:

TrafficIncidents - A json file scraped from snarl.com.au$^{(1)}$; a traffic incident aggregator that sources its data on traffic incidents around Sydney from the RTA. One hundred pages of snarl were scraped using the Beautiful Soup library. The data set contains information on 1000 incidents (10 per scraped page) from a period of two weeks. Data on each incident includes: suburb in which the incident took place, street location, lanes affected, region of Sydney, time reported, time updated, type of incident, and source of information.

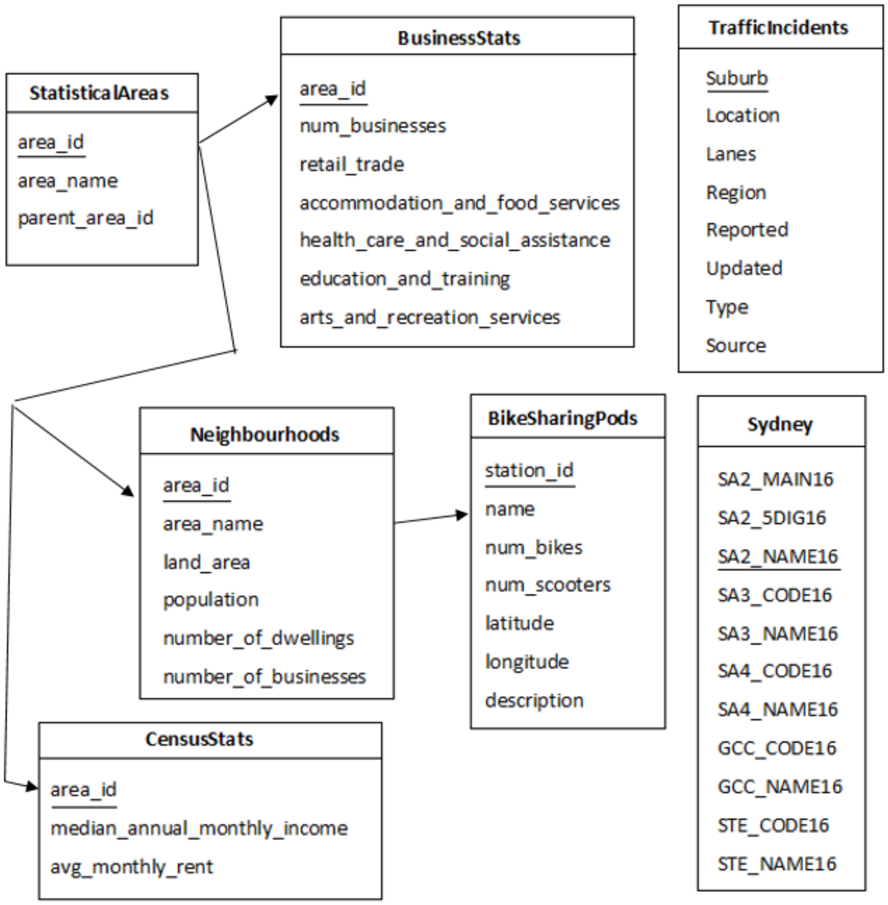
Sydney - A shapefile containing spatial data sourced from the ABS$^{(2)}$ containing boundary data for the regions of Sydney using Statistical Area Level 2 (SA2) granularity. These boundaries reflected the regions in our provided datasets, particularly the Neighbourhoods table. Having this spatial data allowed us to match BikeSharingPods, which does not have an area ID, to the Neighbourhoods table based on the latitude and longitude of each bike pod.

1.1 Data Cleaning

The first entry of each csv file was read to obtain the headings for its dedicated table. Prior to adding the csv file data into its respective table, Neighbourhoods, BusinessStats, and CensusStats required cleaning before they could be incorporated into the database. All other datasets did not require any cleaning. The TrafficIncidents dataset required no cleaning as the only values missing were from the updated column, which we did not need in the following methods.

2 Database Description

The datasets were loaded into tables using the following schema:



A spatial index was created on the Sydney table's *geom* column using PostGIS. The *geom* column contains the geometric data on the ABS's boundaries for the different areas of Sydney. The index was created on this column to speed up the process of computing the spatial join with the BikeSharingPods table.

Using a spatial join, the location of each bike pod was mapped to its region in the table Sydney. Firstly, the latitude and longitude of each bike pod were converted to a geometric point and the SRID was set to 4283, corresponding to the GDA94 coordinate system. Each bike pod name from the BikeSharingPods table was then matched with an area name from the Sydney table. To complete the spatial join, a function matched a pod name and its associated area from the Sydney table to an area name in the Neighbourhood table. Iterating through each of the pod names produced a list of dictionaries containing information for each bike pod including the pod name, the neighbourhood it is located in, and the land area of that neighbourhood. These measures were then used to calculate the bikepod density.

3 Cyclability Analysis

The cyclability score was calculated using the following formula:

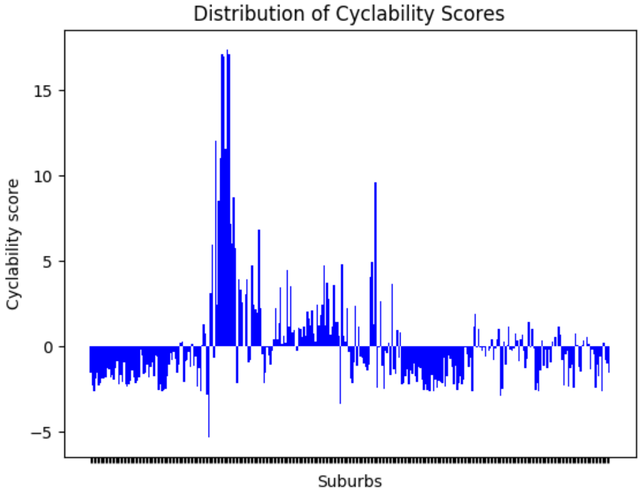
$$\begin{gathered} cyclability(suburb) = z(population\_density) + z(dwelling\_density) + z(service\_balance) \\ +z(bikepod\_density) - z(incident\_density) \end{gathered}$$

Where z, the z-score, was calculated on the formula:

$$z(measure, x) = \frac{x-avg(measure)}{stdev(measure)}$$

The measures used were the following:

- population density; the population of an area divided by its land area
- dwelling density; the number of dwellings in an area divided by its land area
- service balance; the number of businesses in an area divided by its land area
- bikepod density; the number of bike-sharing pods per area divided by land area
- incident density; the number of traffic incidents in an area divided by its land area



The average cyclability score for Sydney suburbs was 0.11 with a maximum score of 17.36 at Surry Hills, and a minimum of -5.31 at Sydney Airport. The median was -0.57 with a standard deviation of 3.07.

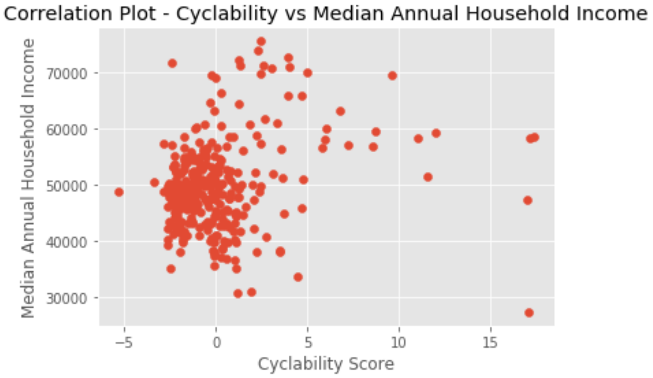
4 Correlation Analysis

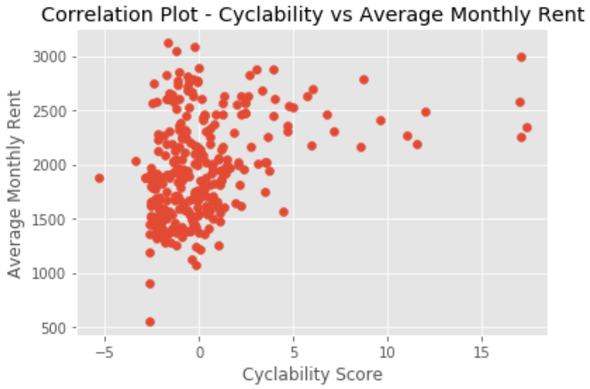
Once a cyclability score was calculated for each neighbourhood, the CensusStats table was queried for the data on median annual household income and average monthly rent. The cyclability scores, income, and rent figures were then put into seperate lists. Some of the income and rent data contained null values. Rather than replacing them with zero, the nulls were replaced with the median of the other non-null values in each respective list. It was decided that the median value would give a better result than the mean value as this would prevent it being skewed by outliers.

With fully numeric data in each list, the numpy module was imported and the $corrcoef()$ function was used to calculate two correlation coefficient matrices; one for the cyclability score correlated with the median income:

$$cyc\_income\_corr = \begin{bmatrix} 1 & 0.2045895 \\ 0.2045895 & 1 \end{bmatrix}$$

And the other for the cyclability score correlated with the average rent:





$$cyc\_rent\_corr = \begin{bmatrix} 1 & 0.35696472 \\ 0.35696472 & 1 \end{bmatrix}$$

Along the diagonal, these matrices represent each variable correlated with itself, giving a correlation coefficient of 1. It is the reverse diagonal that contains the scores of interest.

The income matrix has a correlation coefficient of approximately 0.20, which suggests a weak positive correlation between the cyclability of a neighbourhood and the median annual household income of the inhabitants of that neighbourhood. Similarly, the rent matrix has a correlation coefficient of approximately 0.36. This suggests a slightly stronger (but still relatively weak) positive correlation between the cyclability of a neighbourhood and the average weekly rent of that neighbourhood. Because the correlations are weak, we don't expect the correlation plots to be particularly linear, and indeed they are fairly clustered.

These weak positive correlations suggest that there is some kind of relationship between how amenable an area is to travel by bicycle and the affluence of the inhabitants of that area, perhaps suggesting that more affluent renters and home buyers desire greater cyclability and gravitate towards areas in which this is the case.

References

[1] snarl.com.au
`http://www.snarl.com.au/incidents?p=1`

[2] Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, July 2016
`https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument#Data`