DATA2001 Group Project

Cyclability Analysis

24/5/19

Arpit Nama
Saketh Darbhamulla

Dataset and Database Description

Various Datasets were used to analyse the Cyclability Score of various neighbourhoods in Sydney. They are as follows:

StatisticalAreas.csv

This dataset is provided by the Australian Bureau of Statistics (ABS) and provides information on how the neighbourhoods of Sydney are characterised by an 'area_id.' This allows data analysis to be done on a simple, yet distinctive 'area_id' variable rather than a complex string of letters.

Neighbourhooods.csv

This dataset is also provided by the ABS and provides information on the area, population, number of dwellings and the number of businesses within each neighbourhood.

CensusStats.csv

Also sourced from the ABS, this dataset provides the median household income and average weekly rent for each neighbourhood.

BusinessStats.csv

This dataset, also sourced from the ABS, contains the total amount of businesses per local area as well as the count of five general categories of businesses in the area as well. These categories are: retail/trade, accommodation/food services, healthcare/social assistance, education/training and arts/recreation services.

BikeSharingPods.csv

This is a synthetic CSV file provided by the Sydney City Council that describes the physical location of various bike sharing pods as well as the number of scooters and bikes a certain station has capacity for.

CycleParkingNetwork.JSON

This API JSON data source contains a list of the locations of bike parking spots in Sydney.
Sourced From: [City of Sydney Council, University of Sydney, '2017 Cycle Parking Pod Positions' URL: http://soit-app-pro-4.ucc.usyd.edu.au:3000/(accessed 24 May 2019)]

Pre-processing the data

In order to use the data effectively, a series of processes are required to be executed before any data analysis has occurred. For our analysis, the database management system query language PostgreSQL was going to be utilised within the Python based Jupyter Notebook. This required the import of the ' psycopg2' package as well as establishing connection to the University of Sydney's PostgreSQL server. A function 'pgquery' was defined in order to allow python to execute any SQL statements nestled in the Python code.

The Schema and tables were then created for the analysis of the datasets, first by establishing the names of the tables that will be used throughout the analytical task as well as their column names and variable types.

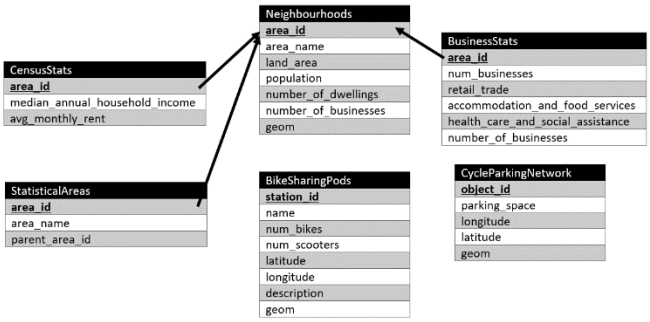
Real data is never perfect, and hence some 'cleaning' must be done in order to produce analysable data. This was done by defining a function that replaces any null or text values within a float or integer array as 'None' in order to provide assessable uniformity to incomplete data. I.e. it is easier to account for missing values when they are all considered 'None'. Each column of the CensusStats.csv, BusinessStats.csv and Neighbourhoods.csv were cleaned using this method.

The datasets were then inserted into their respective tables for the StatisticalAreas.csv, CensusStats.csv and the BusinessStats.csv. However, since the BikeSharingPods.csv provides only longitude and latitude coordinates for their location, an SA2 shapefile of boundary area of the neighbourhoods in Sydney from the ABS website was used to insert 'geom' data into the Neighbourhoods.csv. Then data was loaded into the Neighbourhoods table. The BikeSharingPods.csv is also loaded into a table in a similar fashion.

In order to relate the information, the BikeSharingPods table is joined as a view to the Neighbourhoods table by using PostGIS to match GPS boundaries to find out which points in BikeSharingPods are within a specific 'area_id.' A similar process occurs with the API JSON dataset which also includes longitude and latitude points for parking spots.
The way all these data sets are aligned in the system can be described through the following schema.

Database Schema

All the datasets apart from the BikeSharingPods and CycleParkingNetwork datasets provide information with the 'area_id' primary key. Since the other two datasets possessed GPS coordinates, a foreign key was generated between theses datasets using the shapefile mentioned earlier. This was done by taking the longitude and latitude points of the other two datasets and concatenating them as a singular geometrical point(x,y) in a column named 'geom.' The 'geom' column in the Neighbourhoods database consists of geom data inserted as text into polygon ranges dictated by the shapefile.



Indices

To fasten up our queries we created four indices in our database as follows -

1. N_geom_index on for the geom column in Neighbourhoods table (spatial index)
2. N_land_area_index on land_area column in Neighbourhoods table (normal index)
3. BSP_geom_index for geom column in BikeSharingPods (spatial index)
4. CPN_geom_index for geom column in CycleParkingNetwork (spatial index)

Cyclability Analysis

For each neighbourhood, the Cyclability Score was calculated and then a correlation was measured between that score and the median household income and the average monthly rent.

Cyclability Score

The cyclability score is generated via the following formula:

Cyclability Score *= z(population_density) +z(dwelling_density) + z(service_balance) + z(bikepod_density) +z (parkingspot_density)*

Where the 'z ()' stands for the standard score of that sample and sigma represents the population standard deviation.

Calculating the Standard Score (Z-Score)

$$\text{Standard Score, } z = \frac{X - \mu}{\sigma}$$

TERMS:
μ = mean (pronounced 'mu')
X = score
σ = standard deviation (pronounced 'sigma')

Population Density

The population density of was calculated by dividing the population of an area with the area's land area. The Z-score was then calculated by hardcoding the formula above into Python.

Dwelling Density

The dwelling density was calculated by dividing the total number of dwellings in a neighbourhood by the area of the neighbourhood. The Z-score was then calculated by hardcoding the formula above into Python.

Service Balance

This measurement was aimed to measure how evenly spread out the variety of businesses are in neighbourhoods. This was done with the aim that neighbourhoods in which you are more likely to come across many different types of stores on your cycle commute would score higher. In the BusinessStats.csv, there are only 5 categories of businesses provided and the total number of businesses includes the count of those who fall outside these categories. Hence, we are only measuring the spread of the types of businesses given since it is irresponsible to infer information from null data. First, the magnitude of the deviation from average number of businesses in each category is taken for each category. i.e. **(|Number of businesses in retail trade – average number of businesses in each trade|).** This gives us a set of numbers in which small numbers reflect a good spread and larger ones reflect a negative one.

Let A = (|Number of businesses in retail trade – average number of businesses in each trade|)
Let B = (|Number of businesses in accommodation and food services – average number of businesses in each trade|)
Let C = (|Number of businesses in Health care and social assistance– average number of businesses in each trade|)
Let D = (|Number of businesses in education and training – average number of businesses in each trade|)
Let E = (|Number of businesses in arts and recreation services– average number of businesses in each trade|)

Formula for service balance is: *1000/(A+B+C+D+E)*

The above formula is used to normalise the results such that the largest score corresponds to the most evenly spread neighbourhood. The reason the 1000 is used is that the results were extremely small when 1 or 100 were used due to the integer nature of the column cutting off significant figures. There were some null values in the data due to incomplete rows in the dataset given and hence the service balance score for these neighbourhoods is not given. The Z-score for service balance was calculated by hardcoding the formula into Python.

Bikepod Density

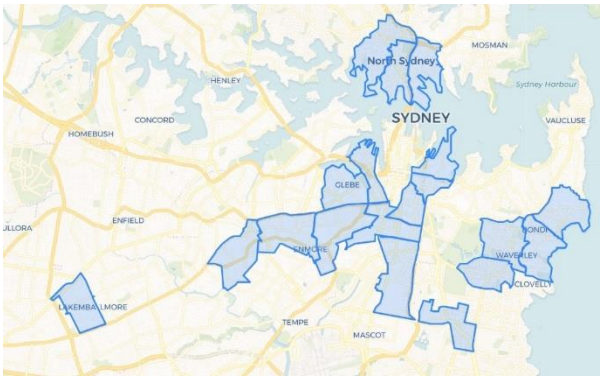
The Bike Pod density was calculated by dividing the total number of pods in a neighbourhood with the land area of that neighbourhood. The Z-score for bike pod density was calculated by hardcoding the formula into Python.

Parking Spot Density

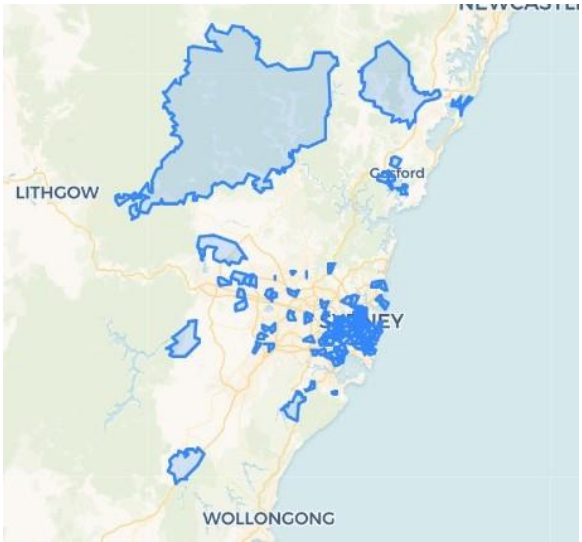
The Parking Spot density was calculated by dividing the total number of spots in a neighbourhood with the land area of that neighbourhood. The Z-score for parking spot density was calculated by hardcoding the formula into Python.

Statistics about the Cyclability Score

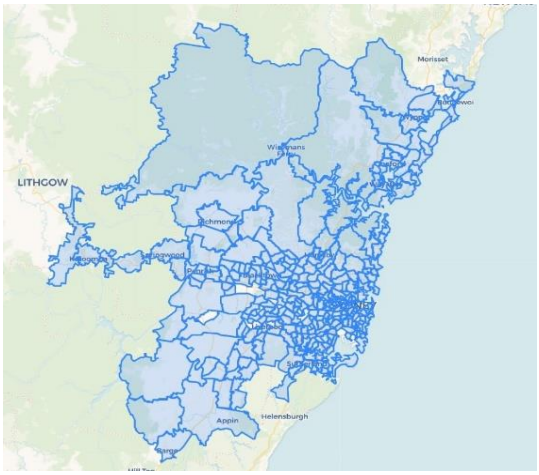
The mean of Cyclability Score of all Neighbourhoods is -1.4655e-16 and the median is -0.7123



Map 1: Top 20 neighbourhoods for Cyclability



Map 2: Top 100 neighbourhoods for Cyclability



Map 3: Top 289 neighbourhoods for Cyclability

The three maps given outline the nature and spread of the Cyclability Scores for 289 neighbourhoods across Sydney. From Map 1, it is evident that the best neighbourhoods for cycling are only close to the Central Business District. In Map 2, the next 80 neighbourhoods do not seem to follow this CBD centred trend. Instead, the spread of neighbourhoods in the top 20 to 80 range seems quite random. Map 3 displays the range of the Cyclability Data Neighbourhoods and acts as a comparison for the other two maps. From this, we can deduce that there is a trend for cyclability in regions close to the city, however that trend is not noticeable further away from the city. This suggests that the majority of the Cyclability Scores are independent of location.

Correlation Analysis

The cyclability scores were compared to the median household income and average monthly rent in order to determine if there is a correlation between the two variables. The median household income and average weekly rent are given by the CensusStats dataset. The Pearson Correlation Coefficient formula is used to calculate the correlation between these two variables.

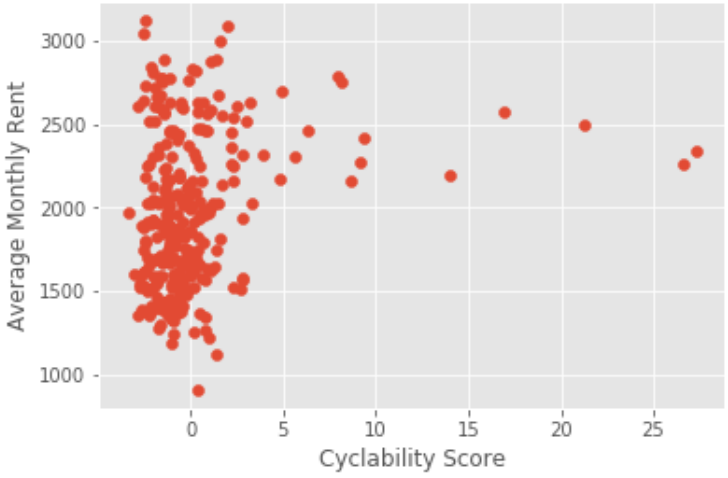
Correlation coefficient for Cyclability score and Median Household Income: 0.20985382604487762

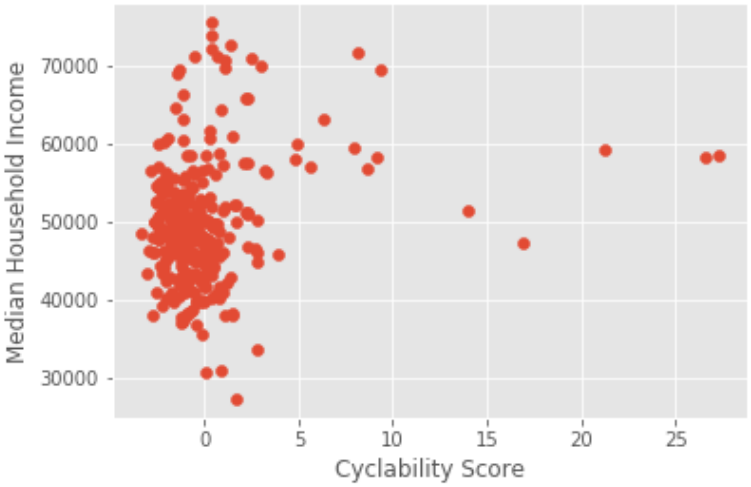
Correlation coefficient for Cyclability score and Average Monthly Rent: 0.21321580503202167

Our correlation coefficients are low, however do show a slightly positive trend toward both median household income and average weekly rent. This may be due to various reasons.

1) Those who earn more may be more inclined to live in areas that do cost more to live in, hence providing similar trends for both income and rent

2) Another is that, rent depends on the quality of the surroundings of the property, i.e. public transport access, footpath maintenance, etc, many of which are desirable in the housing communities as well as for cyclists. Those with a high income are willing to pay to live in developed neighbourhoods.
3) Not all variables can be possibly considered when determining cycling. Some external datasets may contradict the results given here especially considering how low the magnitude of the correlation coefficient is.





As shown in the correlation graphs above, there are certain outliers with the Cyclability Scores. These outliers appear to have unusually high Cyclability scores while the rest of the population seems uniformly distributed around zero. In Map 1, it is shown that the top 20 neighbourhoods are based around the city. Therefore, it appears that the Cyclability Scores are uniformly spread out apart from the occurrence of high scoring outliers around the city area. It is also worth noting that even though both correlation coefficients were slightly positive, the couple highest scoring Neighbourhoods did not have the highest rent or income.

Libraries Used

Following are the libraries which were imported to make this project:
psycopg2, psycopg2.extras, csv, pprint, shapefile, re, requests, json, statistics, numpy, matplotlib and matplotlip.pyplot