Unikey: smcm7233, mile3901, zhwu6710
SID: 480379414, 480133780, 480429524

Assignment DATA2901: Cyclability Score

Dataset Description

The five initial datasets were *BikeSharingPods*, *BusinessStats*, *Neighbourhoods*, *CensusStats* and *StatisticalAreas*. These were provided by the University of Sydney DATA2901 course and obtained through Canvas.

BikeSharingPods – *P*rovides the latitude and longitude of generated bie pods in the Sydney region.
BusinessStats – Provides the type and number of each business located within each suburb.
Neighbourhoods – Contains the land area, population, number of dwellings and businesses for each suburb.
CensusStats – Contains median annual income per houshold and average monthly rent for each suburb.
StatisticalAreas – A reference of the area names and id's of all areas within Australia.

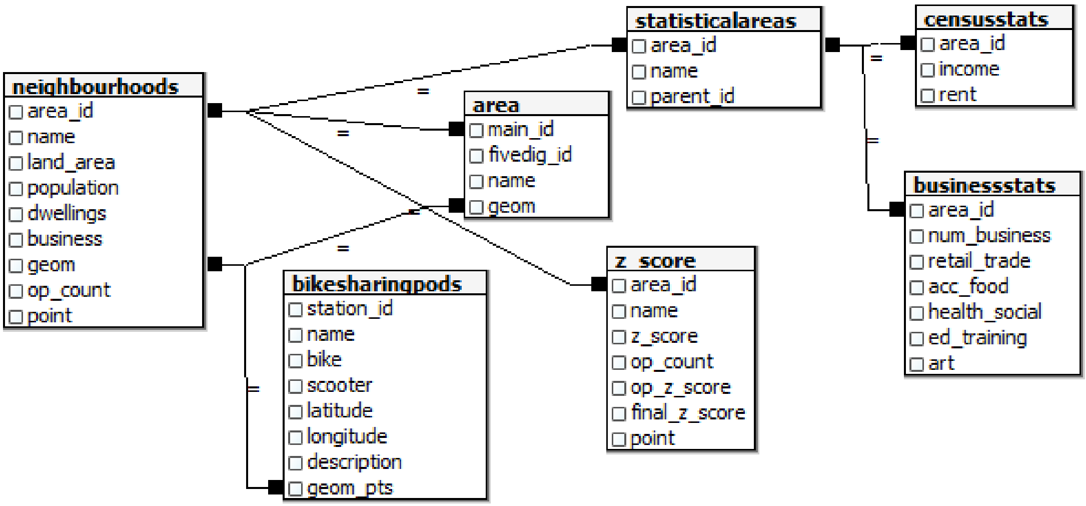
Cleaning was required for the *BusinessStats*, *Neighbourhoods* and *CensusStats* datasets. For all tables, missing integer values were replaced with zeros which was necessary since the missing values cannot be inserted into the table. We also filtered the data for potentially negative incorrect numbers to be replaced with the defualt value -1.

Secondly, the *SA2_2016_AUST.shp* dataset contained the geometry of geographical areas of all suburbs/areas in Australia. This dataset was obtained from the Australian Bureau of Statistics Website (see footnote). *SA2_2016_AUST.shp* was processed to only include suburbs where 'area_id' of the suburb matched one of the 'area_id's in *StatisticalAreas* i.e. is a suburb of Sydney*(*more detail next section).

For **our additional dataset**, we scraped the cycling forum Bicycles Nerwork Australia under the topic 'Best suburb to live for a bike rider'. The intention was to find out how the opinion of cyclists could give insight into the cyclability of each suburb. We aggregated the number of mentions that each suburb received across all of the posts. This required an automatic iterator to scrape each page, since only 25 posts are displayed on each page. A total of 46 suburbs were mentioned in the forum out of the 312 suburbs in Sydney. The mention counts for 'Sydney' and 'Parramatta' were removed as they were often used in a description or to refer to a general region rather than their respective suburb names. Finally, this data was added to the *Neighbourhood* table under the attribute name 'op_count'.

Unikey: smcm7233, mile3901, zhwu6710
SID: 480379414, 480133780, 480429524

Database Description



We used the server *soit-db-pro-1.ucc.usyd.edu.au* and uploaded the data onto the Public Schema as it supports PostGIS. For each csv file we created tables (*Neighbourhoods, BusinessStats, StatisticalAreas, CensusStats, BikeSharingPods*) and set the column 'area_id' within the *StatisticalAreas* as the only primary key. Where necessary, we shortened the name of several columns. We inserted data from the file *SA2_2016_AUST.shp* into the *Area* table which stores the coordinates and geometry of each Sydney suburb as well as its name and area_id.Then we set foreign keys for the 'area_id' columns for every table except for the *BikeSharingPods,* since it did not have the column 'area_id'.

To incorporate the *BikeSharingPods* we added a geometry column to the *Neighbourhoods* table containing the polygons for each suburb, derived from join with the *Area* table. Next, we created a geometry column of points, 'geom_points', in the *BikeSharingPods* based on the latitude and longitude provided. From these to two spatial columns we were able to perform a spatial join between the polygons and the points. Thus, all of our datasets were integrated and joined together allowing us to find the number and density of bike pods per suburb and to calculate the aspects of the cylability score.

Indexes
We created two indexes, *Area_id_index* and *Bike_geom_index,* to speed up the spatial join and calculation of z-score. Since only the data relevent to our join/calculation is established by the index the entire table only needs to be parsed once and this makes it faster. The first index sped up the joining of the tables based on 'area_id' as well as a faster calculation of the z-score. The second index was a spatial index allows the spatial join to run quicker. We intended to create more indexes, but due to the index size limitations of the server we were unable.

Cyclability Analysis

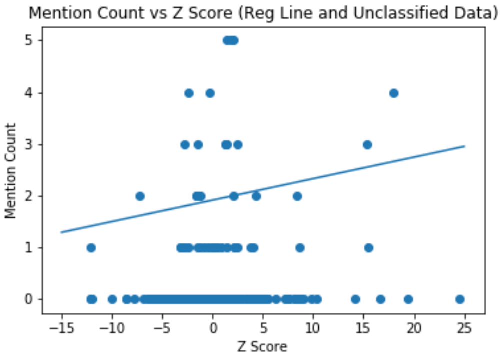
Formula for cyclability:

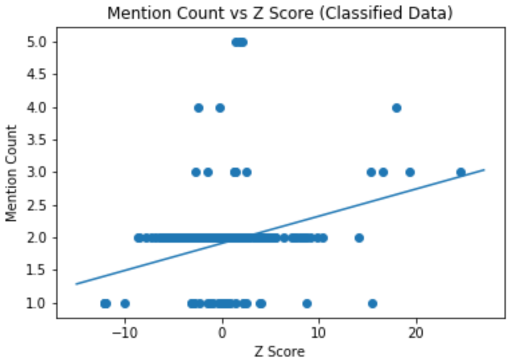
cyclability = z(population density)+ z(dwelling density)+ z(service balance)
+ z(bikepod density) ***+ z(mention score)***

Unikey: smcm7233, mile3901, zhwu6710
SID: 480379414, 480133780, 480429524

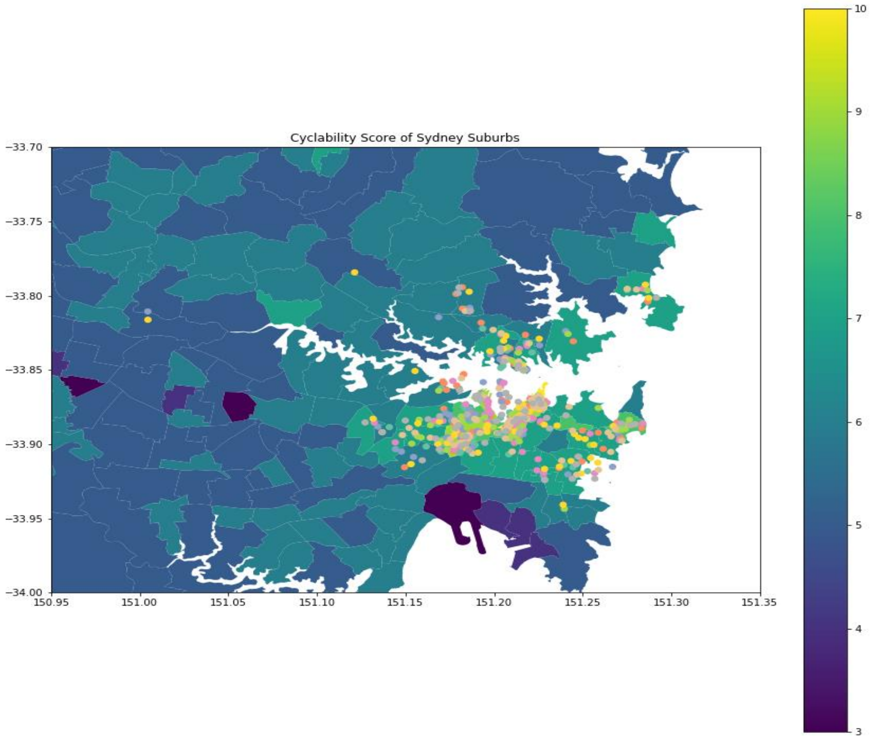
Mention Score – This score is the result of our application of machine learning to the data, scraped from Bicycles Network Australia. The scraping provided information for 46 suburbs allowing us to fit a regression model to the data with the z-score from the first five terms as the explanatory variable and the mention count as the dependent variable. The regression equation was:

$$\text{mention score} = 1.9104 + 0.0415 * (\text{z score})$$





Each remaining suburb was then classified based on its z score into one the mention score categories (1-5). Then to standardize the score, the provided z score calculation was applied to each of the suburbs. For the pre-mention score cyclability score, we followed the standardization procedure provided in the assignment outline and summed each z score for every suburb. We decided to use views to calculate the z-score since the tables can be updated and with views, we can avoid incorrect z-scores. Therefore, for each csv file, we calculated the z-score and kept the query in 5 views. It is important to mention that for business balance, the z-scores are based on the ratio of each business categories per number of businesses. In the end, the sum of these values is stored in the view ‘final_score’.



Unikey: smcm7233, mile3901, zhwu6710
SID: 480379414, 480133780, 480429524

The top 5 suburb cyclability scores are:
Darlinghurst (26.6), Surry Hills(22), Potts Point -Woolllomooloo (21.4), Pyrmont – Ultimo (18.6), Newtown – Camperdown – Darlington (17.3).

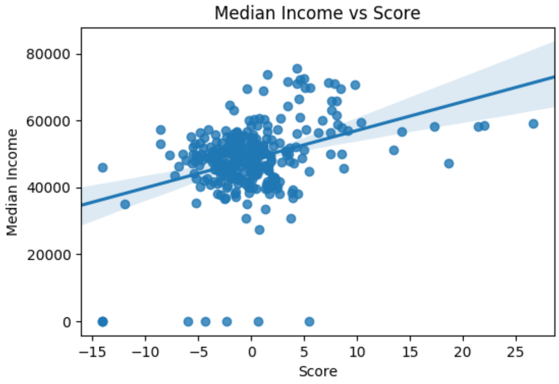
The bottom 5 suburb cyclability scores are:
Blue Mountains (-14.07), Rookwood Cemetary(-14.07), Badgerys Creek (-14.06), Sydney Airport (-14.05), Yennors Industrial (-12.01)

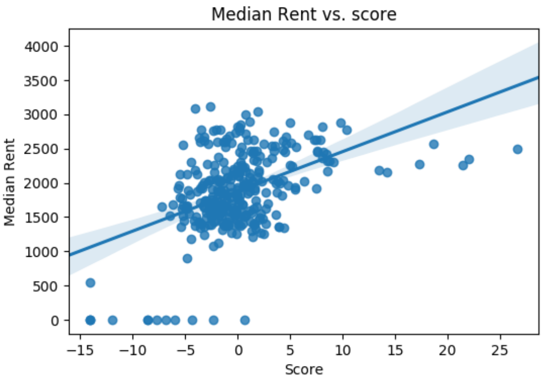
The above map shows our cyclability score for all suburbs in sydney on a colour scale from 1-10 with the position of the bikepods overlayed. The inner city (CBD) of sydney has several suburbs with very high cyclability rankings (8-10) while the many suburbs just outside inner sydney have rankings from 6-8, shown by the mid green. As much is indicated by the top and bottom five. Suburbs closer to the Sydney CBD have the highest cyclability scores, while regional and low population density areas have a low cyclability score.

Correlation Analysis

The correlation coefficient of median houshold income and cyclability score is 0.37, a weak positive correlation. Obviously, there are other factors involved in how houshold income is determined, however the cyclability score appears capture some the influence of suburb characeristics liveability. The higher the liveability of a suburb, the higher the living cost, and the tendency for higher income earners o live in these suburbs. This is likely the logic here.

The correlation coefficient of median houshold income and cyclability score is 0.47, a slightly stronger positive correlation. This follows the inference of more liveable areas having higher rent, and the cyclability score somewhat captures the metric of the overall liveability of each suburb. Thus, higher cyclability would lead to higher rent.





Overall, there are many more aspects that could effect the rent and income of a suburb other than the cyclability. However, our cyclability score is useful for indicating the median rent or houshold income of a suburb in Sydney.

SA2_2016_AUST.shp dataset accessible from:
https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument#Data named Statistical Area Level 2 (SA2) ASGS Ed 2016 Digital Boundaries in ESRI Shapefile Format