Sydney Cyclability Analysis

Zijin Huang
Mengchen Ye
Wendi Tanaka

2019 May

Overview

This project aims to produce and analyze the cyclability of suburb areas around Sydney based on their respective statistics.

Dataset Description

Initially, there are five datasets provided. They contain some information on neighbourhoods of Sydney area, the specifics include but not limited to the population, number of dwellings, and the totals of the specific types of businesses. Also, the school provided a synthetic dataset of some bike pods and their respective attributes, longitude, and latitude. Since the primary goal of this project is to calculate the cyclability score of each area, all NULL values are treated as zero when performing score calculation.

In supplement to the datasets, the team has obtained the area shapefile[1] from Australian Bureau of Statistics (ABS). The dataset contains Statistical Area 2(SA-2), with polygon shape information for PostGis to process.

Also, the team collected data using web-scraping and scrape readable text from PDF file. The team scraped all the text data from 8 PDF files and 16 websites which are related to cycling in Sydney, then use NLP skills to grab and classify nouns from these data. After, a program counted the frequency of occurrence of area_name in StatisticalAreas.csv appears in these nouns.

[1] https://www.ausstats.abs.gov.au/ausstats

Database Description

Schema

See Appendix.

Indexes

There are two normal indexes on this database:

1. neighbourhoods_land_area_idx
2. businessstats_num_businesses_idx

These indexes are created to enable fast search and sort on neighbourhoods based on its land area, and its business count. Since these two are used

Also, there are two spatial indexes:

1. bikepods_geo_index
2. neighbourhoods_geo_idx

These indexes are created for fast spatial join between bikepods and neighbourhoods.

Cyclability Analysis

Before we calculate the cyclability score of any given region, we first obtained the following statistics of each neighbourhood:

$$population\_density(pd) = \frac{population}{land\ area} \tag{1}$$

$$dwelling\_density(dd) = \frac{number\ of\ dwellings}{land\ area} \tag{2}$$

$$service\_balance(sb) = \frac{education * 5 + food * 4 + retail * 3 + recreation * 2 + health}{number\ of\ service\ businesses} \tag{3}$$

$$bikepod\_density(bd) = \frac{number\ of\ bikepods}{land\ area} \tag{4}$$

$$NLP\_score(nlp) = The\ number\ of\ occurrences\ of\ the\ area\_name\ in\ the\ fetched\ nouns. \tag{5}$$

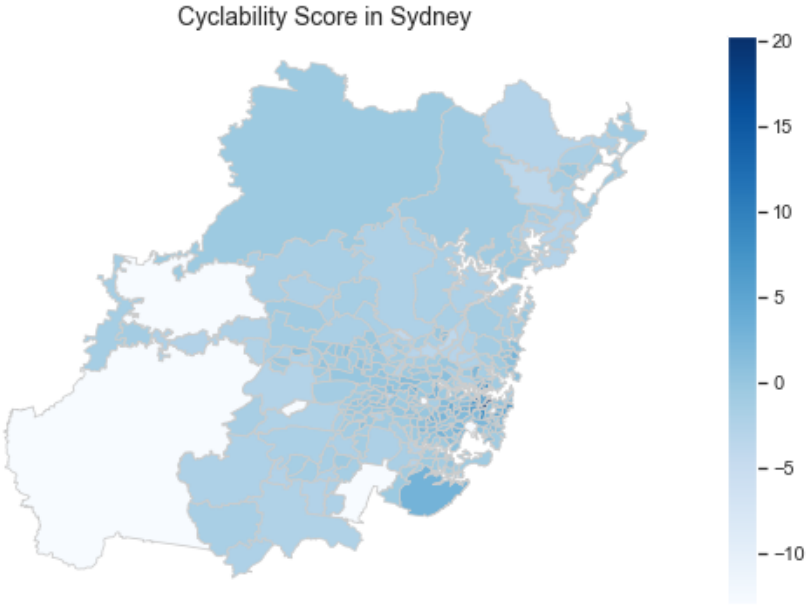
Details about service balance

Service_balance calculation derived from 5 businesses consists of education, food, retail, recreation and health. Those five businesses number scaled up by its usefulness.

After obtaining the above statistics of every neighbourhood, the cyclability score is calculated based on the given formula:

$$cyclability = z(pd) + z(dd) + z(sb) + z(bd) + z(nlp) \tag{6}$$

By implementing the above equations, we obtained the cyclability score of the Sydney suburbs:

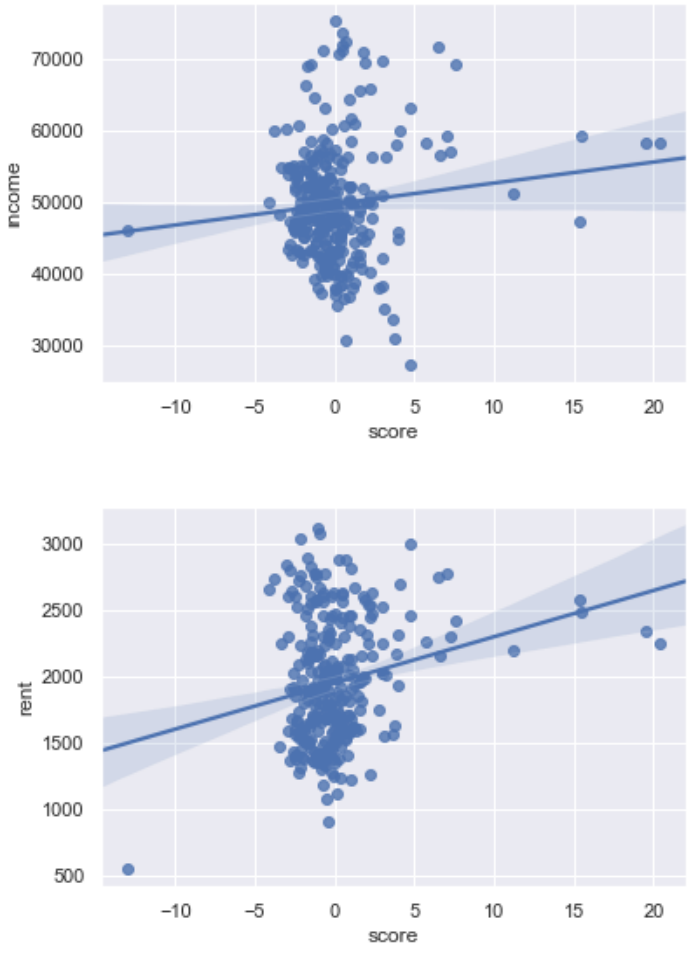


Web scraping and NLP

To enrich the datasets, we scrape readable text from government website and PDF which are related to the Sydney cycling infrastructures with Beautifulsoup and PyPDF2 library. Then the program processed the said text with NLP to classify the nouns. After removing the irrelevant entities from the existing nouns, the program counts the occurrence frequency of nouns related to each neighbourhood. Finally, the program computed NLP z score and added the result to the current cyclability score.

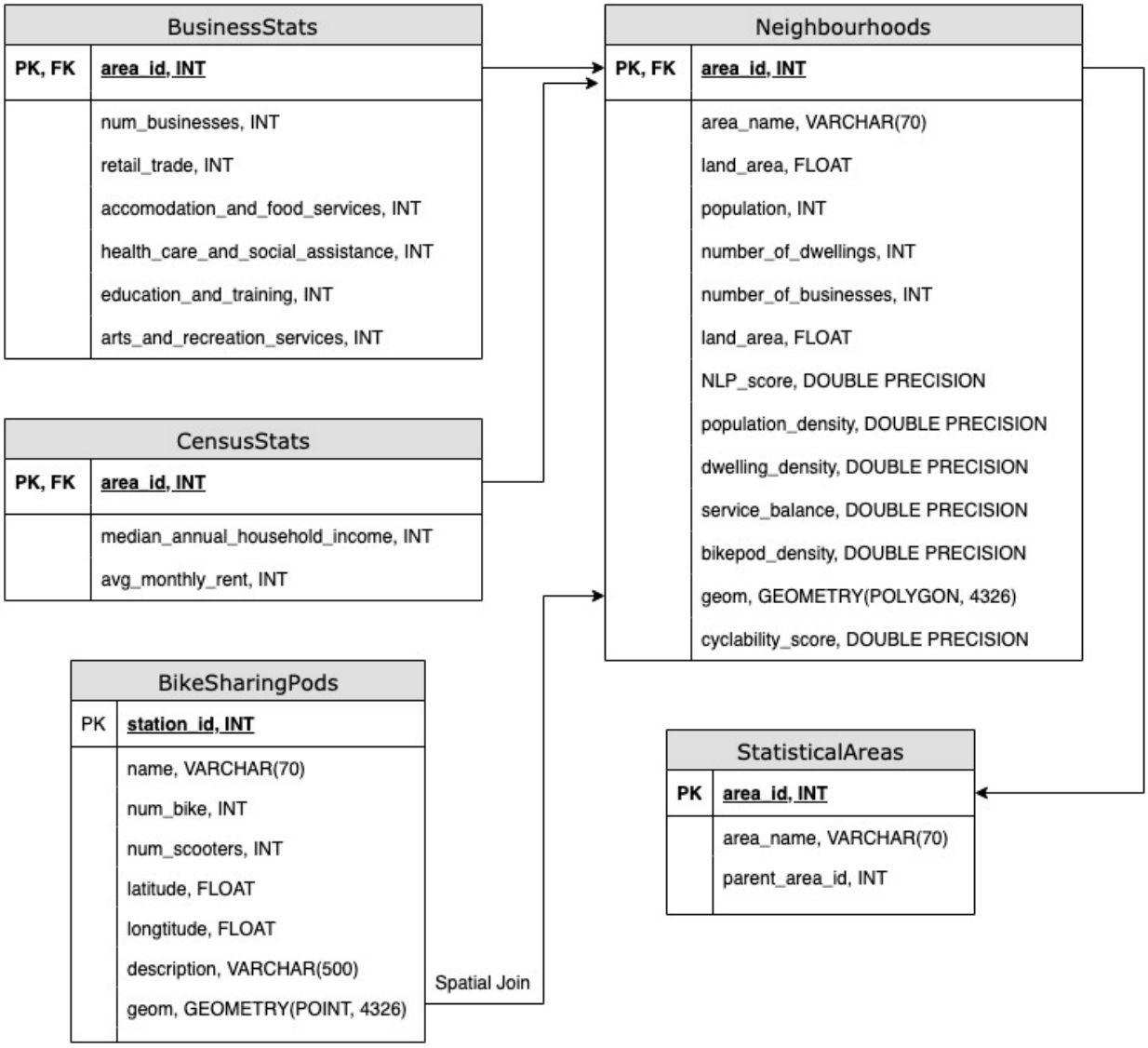
Correlation Analysis

Based on the statistics, we have obtained the following scatter plots:



We can see that the cyclability score is positively correlated with both average income and average weekly rent. The score and average annual income has a correlation coefficient of 0.109, and it has a correlation coefficient of 0.221 with neighbourhood average weekly rent.

Appendix


BusinessStats
PK, FK
area_id, INT
num_businesses, INT
retail_trade, INT
accomodation_and_food_services, INT
health_care_and_social_assistance, INT
education_and_training, INT
arts_and_recreation_services, INT
CensusStats
PK, FK
area_id, INT
median_annual_household_income, INT
avg_monthly_rent, INT
BikeSharingPods
PK
station_id, INT
name, VARCHAR(70)
num_bike, INT
num_scooters, INT
latitude, FLOAT
longtitude, FLOAT
description, VARCHAR(500)
geom, GEOMETRY(POINT, 4326)
Spatial Join
Neighbourhoods
PK, FK
area_id, INT
area_name, VARCHAR(70)
land_area, FLOAT
population, INT
number_of_dwellings, INT
number_of_businesses, INT
land_area, FLOAT
NLP_score, DOUBLE PRECISION
population_density, DOUBLE PRECISION
dwelling_density, DOUBLE PRECISION
service_balance, DOUBLE PRECISION
bikepod_density, DOUBLE PRECISION
geom, GEOMETRY(POLYGON, 4326)
cyclability_score, DOUBLE PRECISION
StatisticalAreas
PK
area_id, INT
area_name, VARCHAR(70)
parent_area_id, INT