Emmanuel Isaac (440429216) | Vishal Singh Uppal (440315067)

WALKABILITY ANALYSIS DOCUMENTATION

Section 1: Dataset Description

a) Data sources and format

Statistical Area 2 (SA2) data was obtained from the Australian Bureau of Statistics (ABS). This data is presented in several **comma-separated value (CSV)** files, neighbourhoods.csv, statisticalareas.csv, censusstats.csv, and businessstats.csv. Data on car-sharing pods is also presented in a **CSV** file format named carsharingpods.csv.

National Public Toilet Map data was retrieved from data.gov.au. The dataset is presented in an **Extensible Mark-up Language (XML)** file format (toiletmap.xml).

2016 Statistical Area Level 2 (SA2) ASGS Ed 2016 Digital Boundaries data was retrieved from the ABS website and presented in the **ESRI shapefile** file format (sa2_2016_aust.shp).

b) Data pre-processing

Data obtained in the **CSV** format was parsed using the imported csv.DictReader module in Python. The data is stored in an Ordered Dictionary, one for each CSV file. For example, data_Neighbourhoods is the name of the Ordered Dictionary created by the DictReader module reading from neighbourhoods.csv.

- The Ordered Dictionaries are structured such that the key is the name of the column in the relation, and the value is the entry for that row/record in the relation.
- The Ordered Dictionaries for Neighbourhoods, Census Stats, and Business Stats are cleaned using a cleaning function to ensure any null values for integer columns are set to NaN. This ensures those few neighbourhoods are discounted in the Walkability Analysis, and that the results are not skewed when calculating z-scores.
- The data is then loaded by rows into the relevant relations in the schema (see Section 2) using PostgreSQL INSERT commands.

National Public Toilet Map data obtained in the **XML** format was parsed using the imported ElementTree module in Python. The data is stored in an Ordered Dictionary. Geographic data obtained in the **ESRI shapefile** format was parsed using shapefile.Reader and each record is also written into an Ordered Dictionary.

- Similar to above, each Ordered Dictionary is structured with two key-value pairs for each record. For example, in the case of the ESRI data, the first is area_id and the second is the geographic boundary data structured as a polygon. National Public Toilet Map data has three key-value pairs; toilet_id, latitude, and longitude pairs.
- For the Digital Boundaries data, the data is loaded by rows into the Neighbourhoods relation in the schema (see Section 2) using the PostgreSQL UPDATE command. Where an area_id from the OrderedDict matches one in the relation, the polygon data is stored in the new 'boundaries' column for that record. PostGIS allows for the storage and querying of this geometric polygon information.
- For the National Public Toilet Map data, the data is loaded by rows into the ToiletMap relation in the schema using the PostgreSQL INSERT command.

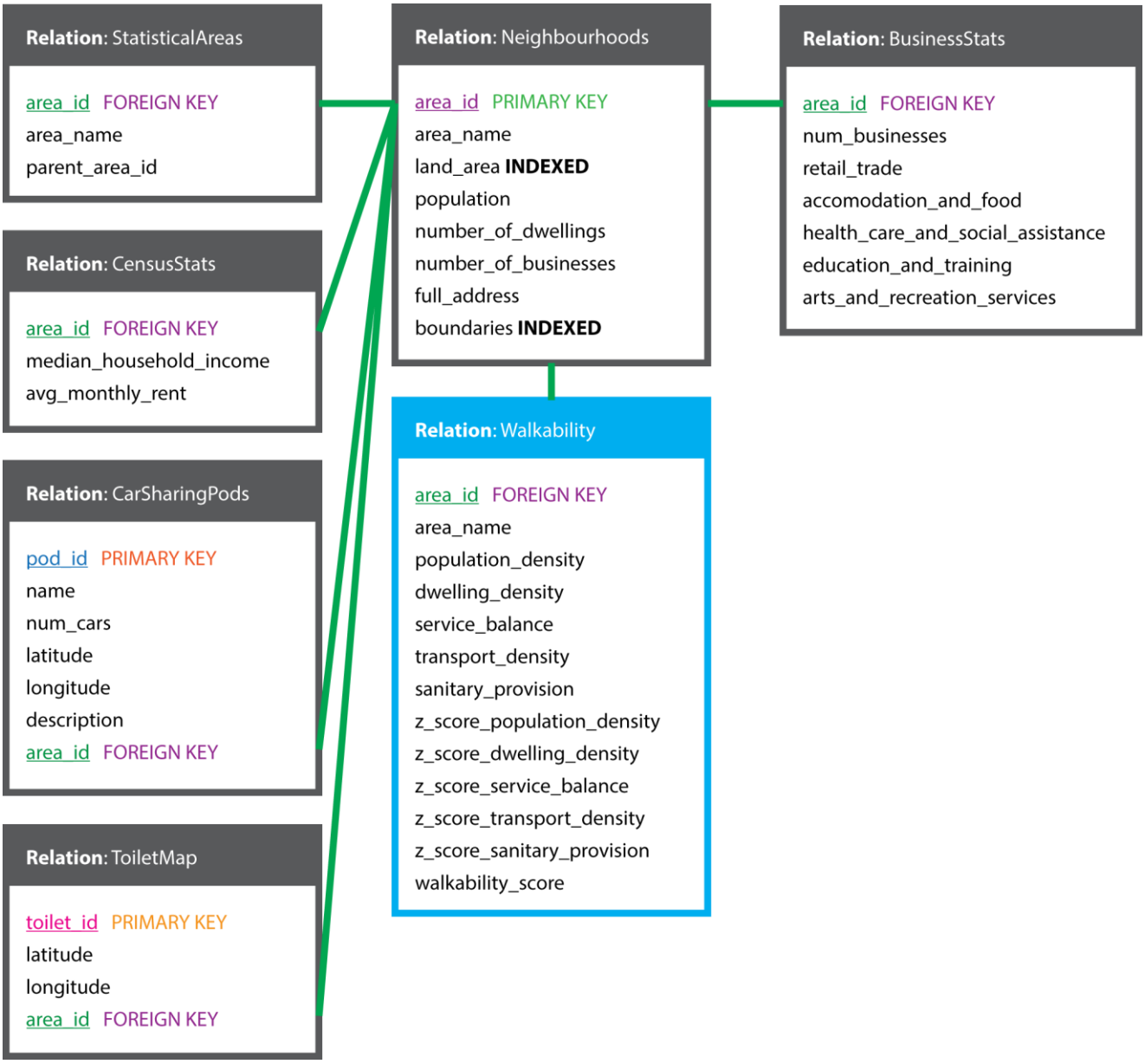
c) Spatial joins

Spatial joins were performed using PostgreSQL SELECT commands to link Car-sharing Pods data and National Public Toilet Map data with area_id in the Neighbourhoods relation.

- Both datasets provide longitude and latitude information in separated, double precision formatted columns.
- The spatial join is performed using PostGIS containment operators, to check that a point value (obtained by casting the longitude and latitude of each dataset into a single point value in PostgreSQL) is contained within the boundaries of a polygon stored in Neighbourhoods.
- Where the operator Boolean is satisfied, the area_id for that boundary information is stored in a new area_id column in the relevant Car-sharing Pods or National Public Toilet Map relation.

Section 2: Database Description

a) Database schema diagram



neighbourhoods.csv is integrated into the **Neighbourhoods** relation. businessstats.csv is integrated into the **BusinessStats** relation. censusstats.csv is integrated into the **CensusStats** relation. carsharingpods.csv is integrated into the **CarSharingPods** relation. statisticalareas.csv is integrated into the **StatisticalAreas** relation. National Public Toilet Map registry data (toiletmap.xml) is integrated into the **ToiletMap** relation. 2016 Statistical Area Level 2 (SA2) ASGS Ed 2016 Digital Boundaries data (sa2_2016_aust.shp) is integrated into the **Neighbourhoods** relation in the column 'boundaries'.

b) Indexes

1. Spatial (GIST poly_ops) Index created on the polygons in the boundaries columns of the **Neighbourhoods** relation. This index improves the efficiency of spatial joins and containment operands which are used to map geometric points (formed from Car Sharing Pods data and National Public Toilet Map data) within the boundaries specified in the geometric polygons in **Neighbourhoods**.
2. Index created on the land_area column of the **Neighbourhoods** relation. This index improves the efficiency of calculating z-scores for the Walkability Analysis. This data is pulled multiple times within queries which can create excessive overhead and computational demand.

Section 3: Walkability Analysis

a) Walkability Score formula

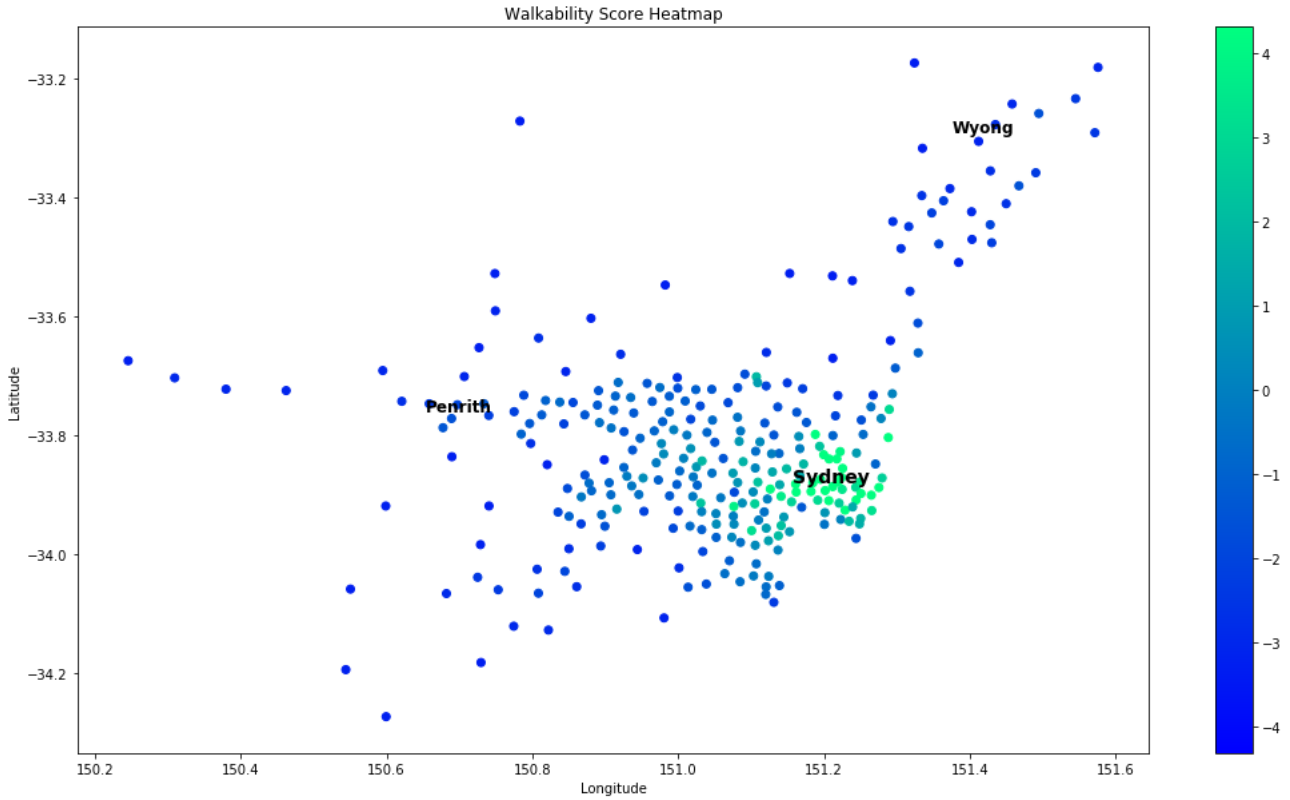
$$score = z(population\_density) + z(dwelling\_density) + z(service\_balance) + z(transport\_density) + z(sanitary\_provision)$$

$$z(measure, x) = \frac{x - avg_{measure}}{stddev_{measure}}$$

- *population_density* is the population divided by the neighbourhood's land area.
- *dwelling_density* is the number of dwellings divided by neighbourhood land area.
- *service_balance* is the balance of selected business types in a neighbourhood. Different weightings were given to the columns in BusinessStats.csv to promote neighbourhoods which facilitated greater retail trade, arts & recreation, and food services. We believe these qualities make for a more 'walkable' neighbourhood.
    - *10%: num_businesses, 25%: retail_trade*
    - *20%: accommodation_and_food_services*
    - *15%: health_care_and_social_assistance*
    - *10%: education_and_training*
    - *20%: arts_and_recreation_services*
- *transport_density* is the number of car-sharing pods per neighbourhood divided by area.
- *sanitary_provision* is the number of national public toilets per neighbourhood divided by area.

b) Overview of Walkability results

```
Average score: 6.87646786878557e-17 ≈ 0
Highest score: 31.1728660334164 ('Potts Point - Woolloomooloo')
Lowest score: -3.3123005004358 ('Bilpin - Colo - St Albans')
```



For the above map, Walkability Scores have been normalised and mapped to a colour spectrum. The spectrum is capped by the mean +/- standard deviation of Walkability Scores so that colour is distributed appropriately to assist in identifying trends in geography. Walkability Score clearly tends to increase with reduced distance to City of Sydney centre. This relationship appears to be of an inverse square nature (see **Appendix 1).** Since our walkability score is heavily based on the density of various services and facilities, it is not surprising to see this trend towards the city. The

black location names on the map mark locations such as the City of Sydney and can be used as geographic reference points.

Section 4: Correlation Analysis

Correlation analysis was performed using *scipy linregress*. Our walkability score has a meaningful, though weak linear correlation (*Pearson's r*) with both annual income and monthly rent. The correlation to median annual household income is *r = 0.26* (see **Figure 1**) and the correlation to average monthly rent is larger at r = 0.35 (see **Figure 2**). This level of correlation is to be expected, as there are many factors which don't contribute to our Walkability Score but can raise rent/household income for neighbourhoods. For example, we do not distinguish between the quality of the facilities and services for our Walkability Score, but quality can greatly impact neighbourhood pricings.

We should also consider the difference in correlation scores for income and rent (*r = 0.26* and *r = 0.35*). Annual income and monthly rent are highly correlated with *r = 0.66* (see **Appendix 2**) but have important differences. Renters are a subset of total household incomes and rent is valued more-so based on location. Due to renters being far more likely to move, access to services and facilities has a higher appeal. This contrasts the more house/land area-based pricing for home owners typically exhibited in Australia. Since our model prioritises density of services and facilities (see Section 3), it follows that the rent prices will likely exhibit higher correlation with Walkability Scores than general median house incomes.

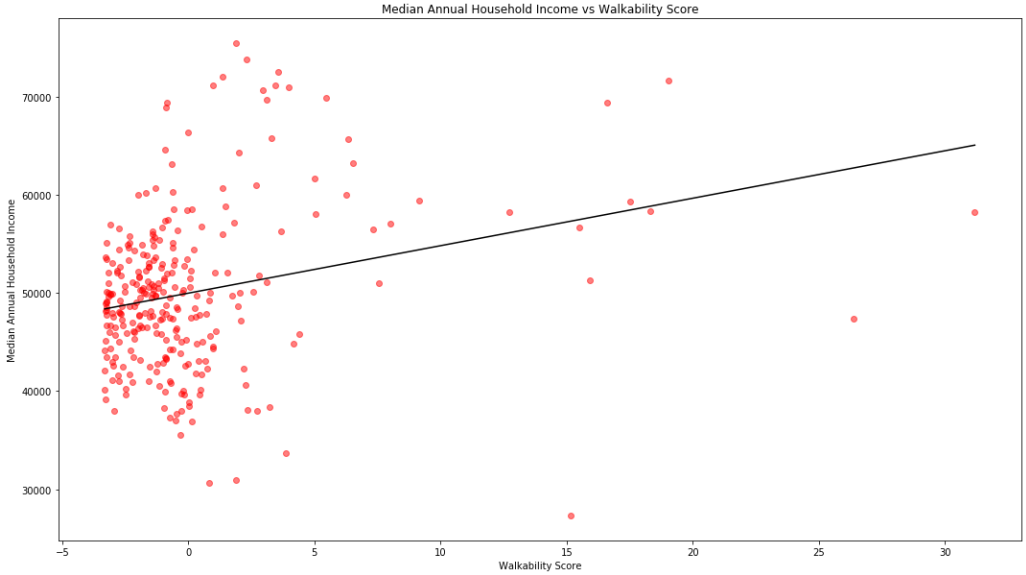


Figure 1: *r = 0.26*

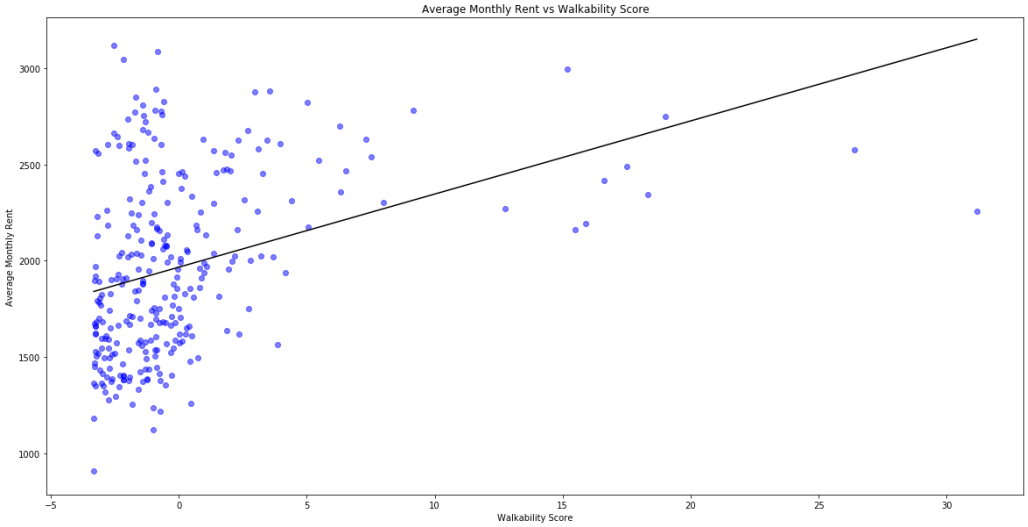
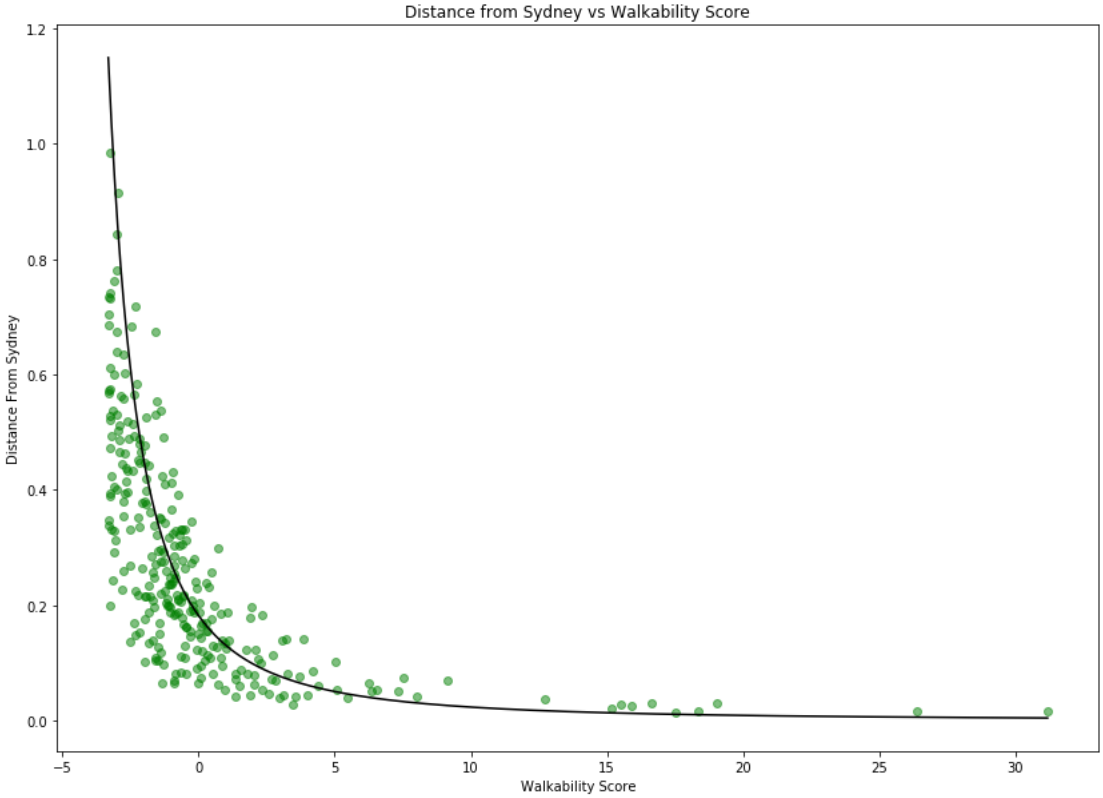


Figure 2: *r = 0.35*

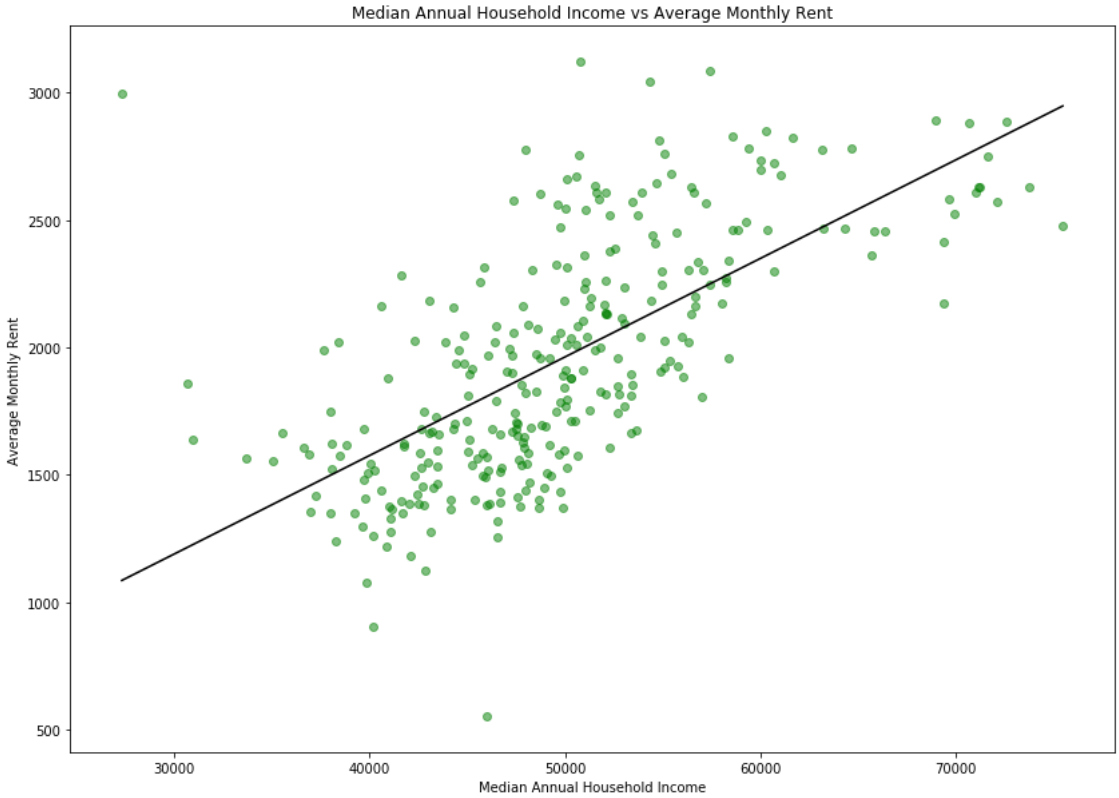
APPENDICES

APPENDIX 1



The linear regression (r = 0.85) of the points scaled using the inverse of $y = 5.5/(x^2 + 5.5)$.

APPENDIX 2



Linear regression (r = 0.66) of median annual household income vs average monthly rent.