Coffee Venues in Melbourne, Australia

A Coursera Capstone Project by Gareth Chadwick

Introduction

In Melbourne, coffee is almost a religion. It is a daily ritual in the lives of most of the resident population, and holds great importance from a social and lifestyle perspective. According to the *lazytrips* blog, in Oct 2019 Melbourne ranked 13th in the top cities in the world for coffee lovers[1]:

"...the coffee scene in Melbourne is one of the best in the world and ...they really love their coffee. A lot. ...there are literally countless independent cafes all around the city, all of which are graced by a friendly atmosphere and that mouth-watering aroma we all know and love."

If you were to visit every single suburb of Melbourne, it seems that you would find at least one cafe, coffee shop or coffee house, whichever term you prefer. However, some suburbs have more cafes than others, and the quality of the venues also differ between areas.

It is interesting to consider what factors about a suburb determine the number and quality of coffee shops it contains. Suburbs are located in different places geographically and are also diverse in their make up of residents. Demographic data on suburbs is collected by the Australian Bureau of Statistics and is available on their website.

Consider someone opening a new coffee venue in Melbourne. A choice has to be made of where to locate this new business. This decision on location can have a significant impact on the profitability of the venue. Opening a new coffee house in an area that is already well served with existing cafes clearly has a greater risk of failure than choosing an area where the current coffee options are poor.

The objective of this study is to use suburb data and Foursquare venue data to develop a model for prediction of number and quality of coffee venues in Melbourne suburbs. Comparison of existing venue data with model predictions can then provide insight into which suburbs are under- or over-served with good coffee venues. This information is of value to the prospective business owner in determining where to locate a new coffee shop in order to have the greatest demand.

[1] https://lazytrips.com/blog/21-of-the-worlds-best-cities-for-coffee-lovers

Data

The data that was used to determine the best location for a coffee venue was as follows:

1. A list of Melbourne suburbs scraped from the wikipedia page https://en.wikipedia.org/wiki/List_of_Melbourne_suburbs
2. Geocoding of suburb addresses, and calculation of distances between points was performed using the **geopy** Python library.
3. Australian Bureau of Statistics (ABS) census data from 2016 was used to gather data on each suburb, such as population, number of families, number of dwellings, median weekly income, etc. This is available from the ABS website https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/2016%20QuickStats and was used as within the feature data for the models.
4. Foursquare data was used to determine the number of coffee venues and their mean rating for each suburb. This is available from the Foursquare API and was used as the label data for the models.

A subset of suburbs that are within a specified distance of the central business district (CBD) was used for this project. This distance was determined based on having a reasonable number of suburbs for comparison, but without requiring too many calls to the Foursquare API for venue data. For each suburb, its latitude and longitude and distance from the CBD was determined using the **geopy** library.

Census data for each suburb was scraped from the Australian Bureau of Statistics website. Coffee venue data for each suburb was pulled from the Foursquare API. The number of venues in the coffee category was recorded for each suburb, along with the rating of each venue, where this was available.

An example data set for one suburb could be as shown in Table 1:

Table 1: Example Data Set for One Suburb

Suburb	Malvern
Latitude	-37.855
Longitude	145.035
Distance_CBD	12
Residents	10066
Families	2585
Dwellings	4539
Median_Income	2288
Number_venues	32
Mean_rating	7.5

Methodology

Geocoding of Melbourne Suburbs

Wikipedia contains a list of Melbourne suburbs, including their name and postcode. This data was scraped to form a dataframe of suburbs. Latitude and longitude data for each suburb was then obtained using Nominatim for the **geopy** library.

Some errors were obtained when geocoding based on suburb names, so the final coordinates data was gathered using postcode rather than suburb name. As different suburbs can share the same postcode, this results in some suburbs being amalgamated into one entry for geocoding purposes.

Selection of Suburbs based on Distance from CBD

The **geopy** library was also used to calculate the distance of each suburb from the Melbourne CBD. The geodesic distance in km was calculated for each suburb. A selection was made based on a distance limit of 20 km from the CBD. Only suburbs with distance < 20 km were selected for the analysis.

This gave a set of 128 suburbs; the distribution of their distance from the CBD is given in Figure 1. Suburbs are spread out from the CBD in an approximately circular pattern, therefore there are generally a greater number of individual suburbs as the distance from the CBD increases.

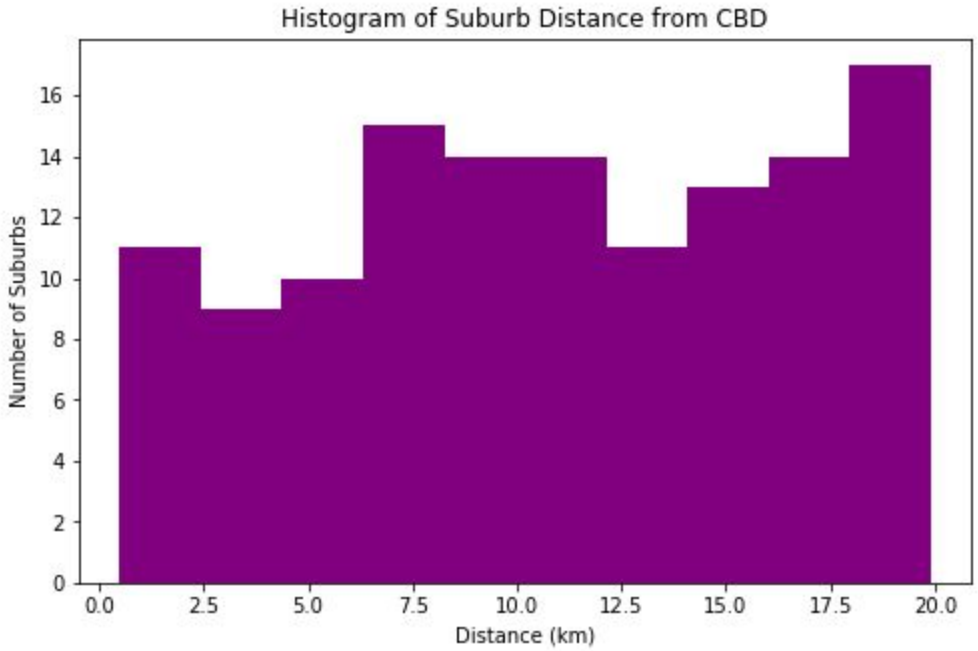

Figure 1: Histogram of Suburb Distance from CBD

Gathering Foursquare Coffee Venue for each Suburb

For each suburb, latitude and longitude coordinates were used to call the Foursquare API to browse venue data. All venues within the cafe/coffee house category were pulled. The API calls were made with the parameters of Table 2.

Table 2: Foursquare Venue Browse Parameters

Parameter	**Value**
version	20180605
radius	1000
intent	browse
categoryID	4bf58dd8d48988d16d941735
limit	50

The radius was set at a quite large value for each call in order to capture all venues, however this led to some venues being duplicated in multiple suburbs. To remove duplicate data, the distance of each venue to the suburb coordinates was determined. The venue data for the closest suburb was then retained and duplicates discarded. This resulted in a list of 2041 Melbourne coffee venues.

For each venue of the venue data set, a call was made to the Foursquare API to obtain the current rating of that venue. This is a premium API call and calls had to be made over several days to obtain this data. Of the 2041 venues, only 670 had been given ratings in Foursquare, i.e. approximately one-third of venues had this data.

Of the 128 suburbs selected for the analysis, all but 6 had at least one coffee venue listed in Foursquare. The suburb of Melbourne itself was listed twice as postcodes 3000 and 3004. Venue data was duplicated for both entries when grouped by suburb name and therefore the data for postcode 3004 was removed. This resulted in a final venue data set for 127 suburbs containing the number of venues in each suburb and the mean rating of the rated venues in each suburb.

Gathering Suburb Demographics Data

The Australian Bureau of Statistics (ABS) has made the Australian Census data for 2016 publically available on its website (https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/2016%20QuickStats). Each suburb of Australia has its own SSC code given by the ABS and these SSC codes for suburbs are available to download from the ABS website as a csv file.

SSC codes were filtered to those in Victoria and then this data was merged with the previously determined data for suburbs within 20 km of the Melbourne CBD. For each suburb, the following data was scraped from the 2016 census data:

1. Number of residents
2. Median age of residents
3. Number of families
4. Mean children per family
5. Number of dwellings
6. Median weekly income per dwelling

There were two suburbs with no demographics data, namely Monash University and Bellfield. These suburbs were removed from the final data set.

Additional Features

Two additional features were generated for each suburb:

1. Total income per suburb = Number of dwellings x Median income per dwelling
2. Total children per suburb = Number of families x Mean children per family

The final data set comprised the suburb demographic data, the additional features calculated above and the coffee venues data from Foursquare. This included data for 125 suburbs.

Exploratory Data Analysis

The distributions of the output variables Venue Count and Mean Rating were visualised using histograms, see Figure 2.

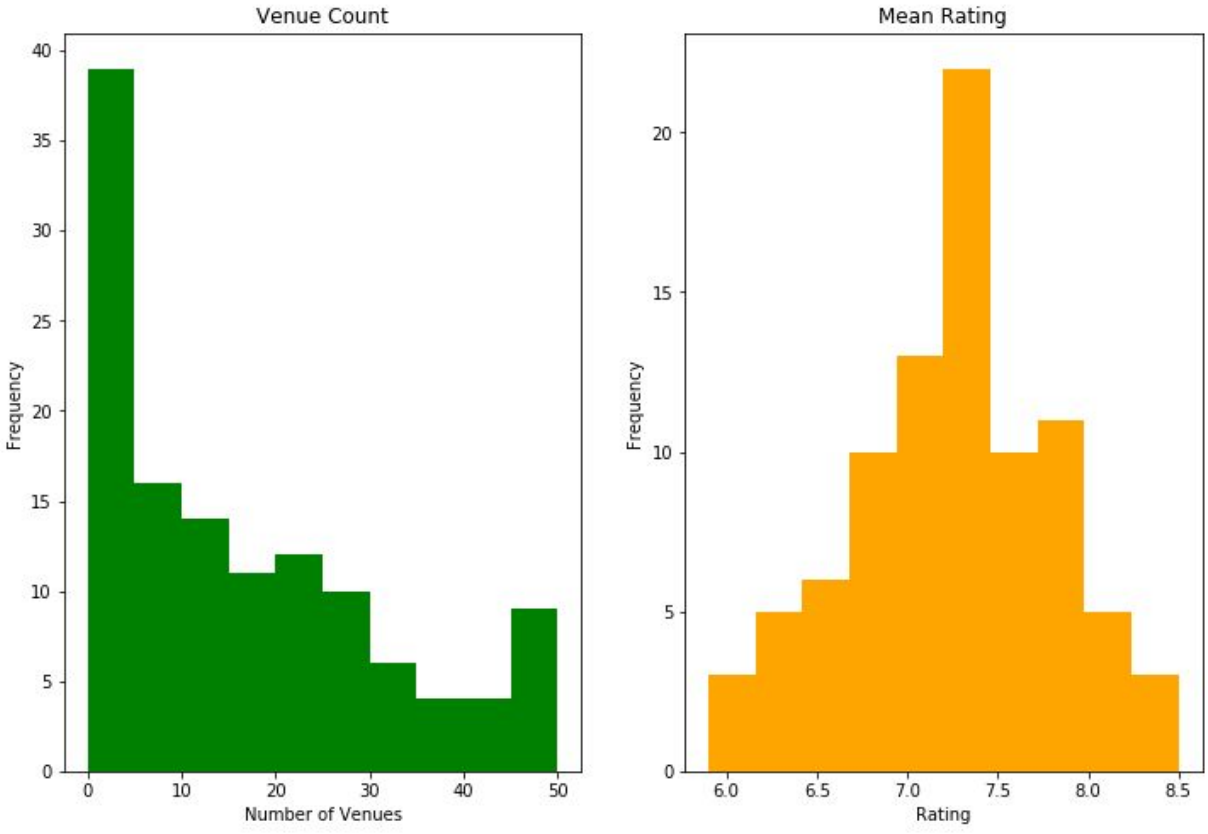

Figure 2: Histograms of Venue Count and Mean Rating

Suburbs with a higher number of venues are generally less frequent than those with few venues. Mean venue ratings are centred around 7.2, with a range of 5.9 - 8.5.

The distributions of the feature variables were also visualised using histograms, see Figures 3 and 4.

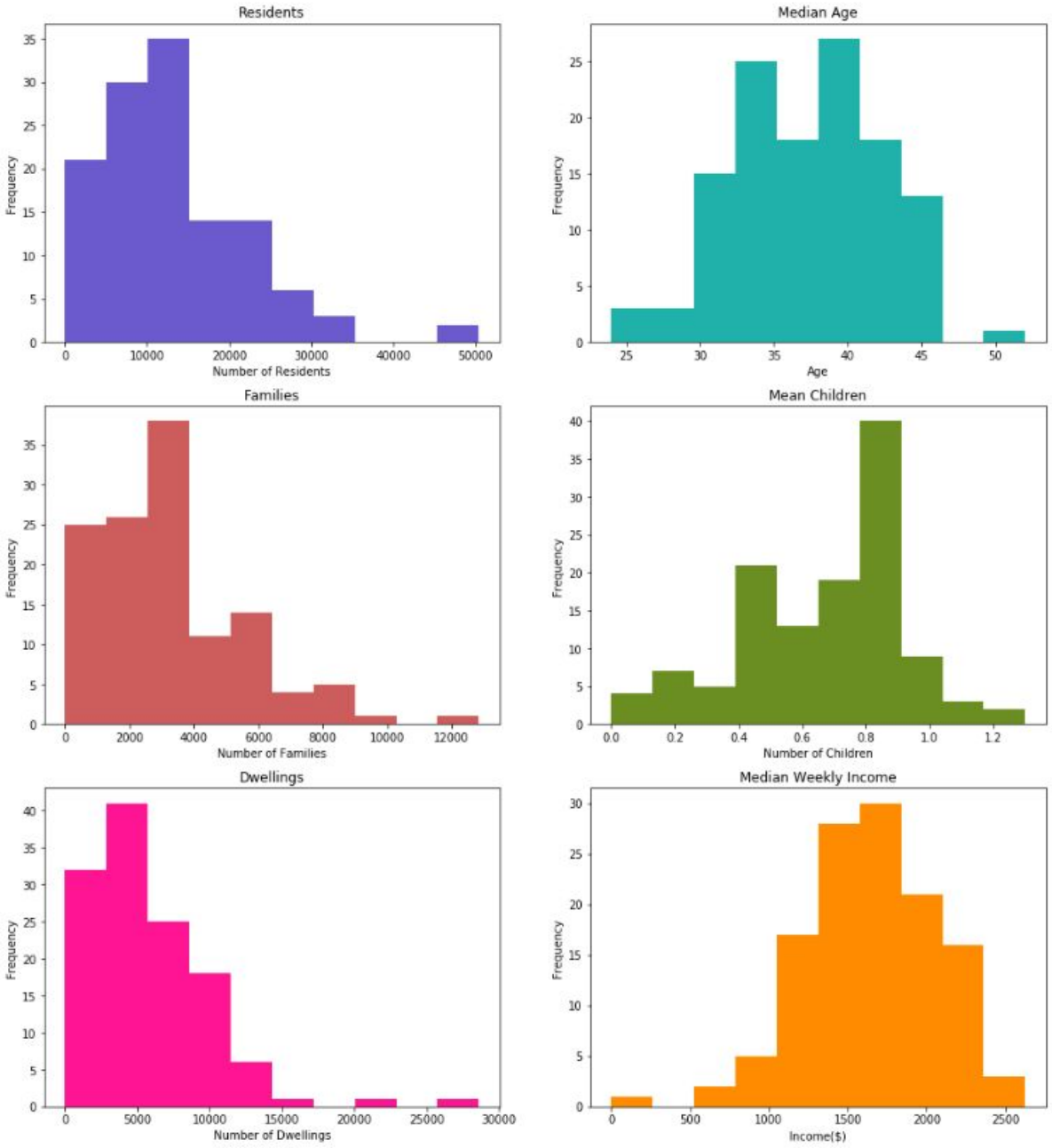

Figure 3: Histograms of Feature Variables 1

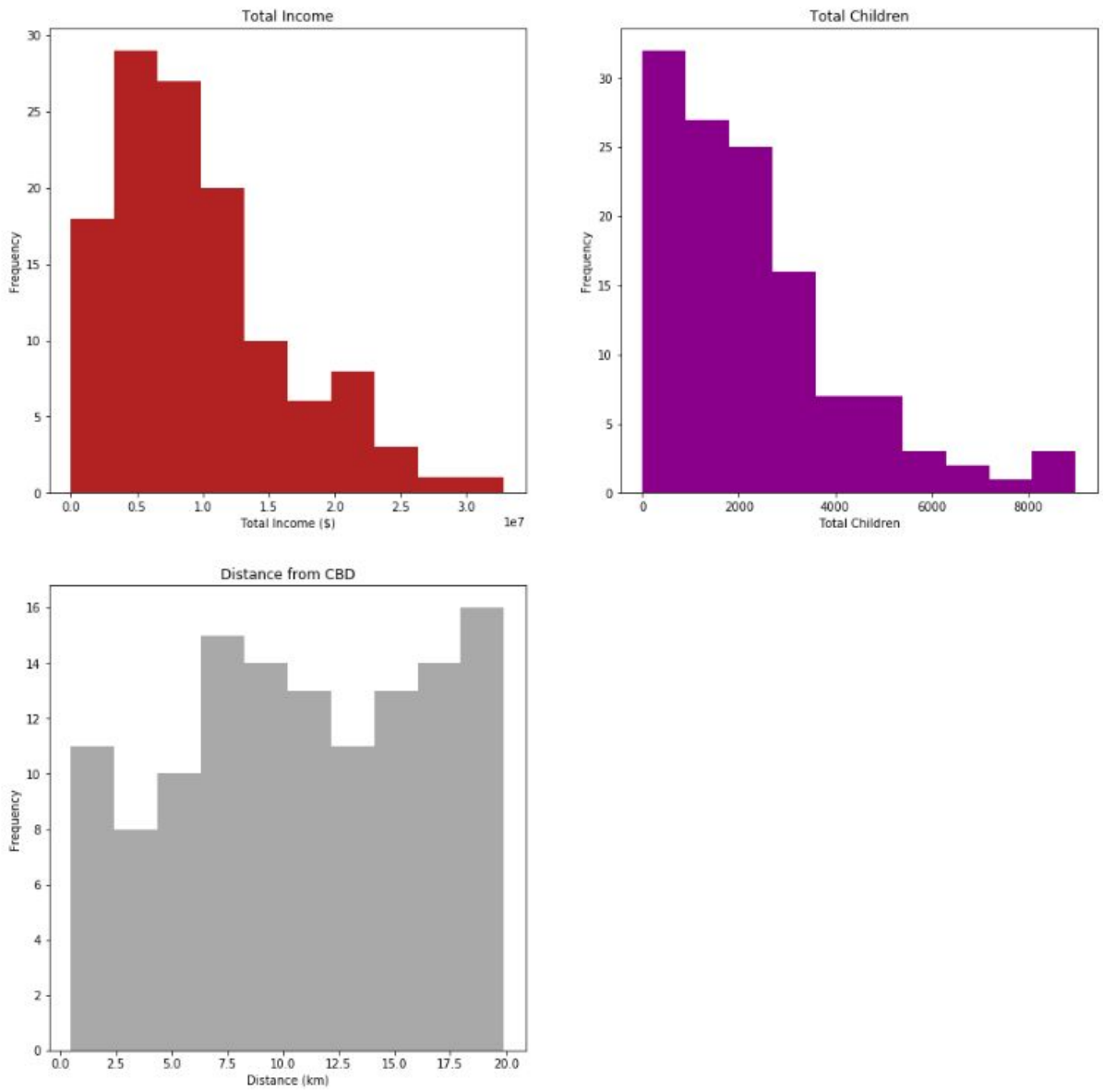

Figure 4: Histograms of Feature Variables 2

The distributions of residents, families, dwellings, total income and total children were all left-skewed, so it was decided to log transform these variables by y = log10(1 + x). This gave the distributions of Figure 5. The log transformed variables were used in further analysis.

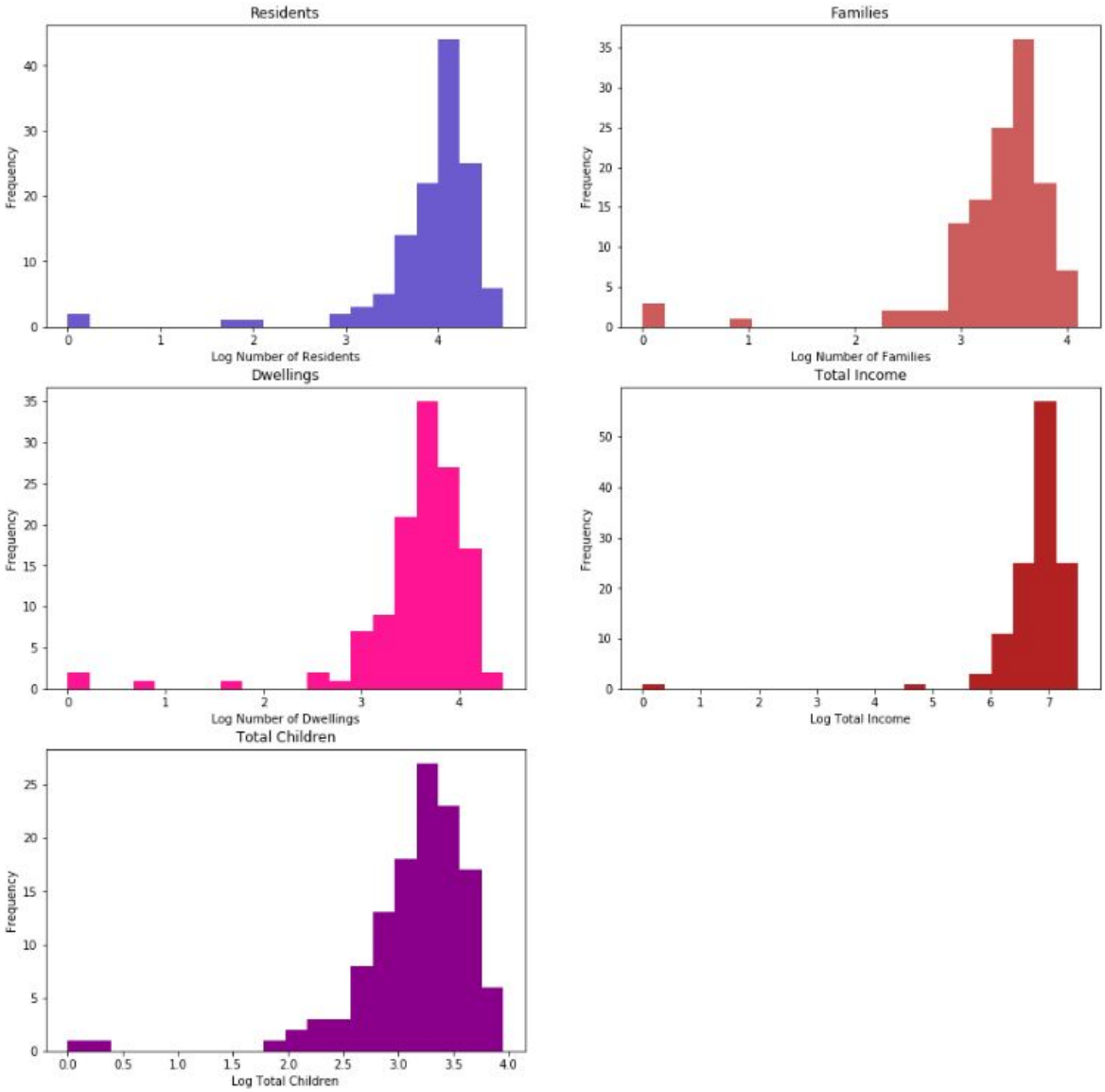

Figure 5: Histograms of Transformed Feature Variables

The correlations between all variables (feature and output) were visualised using a correlation heatmap, see Figure 6.

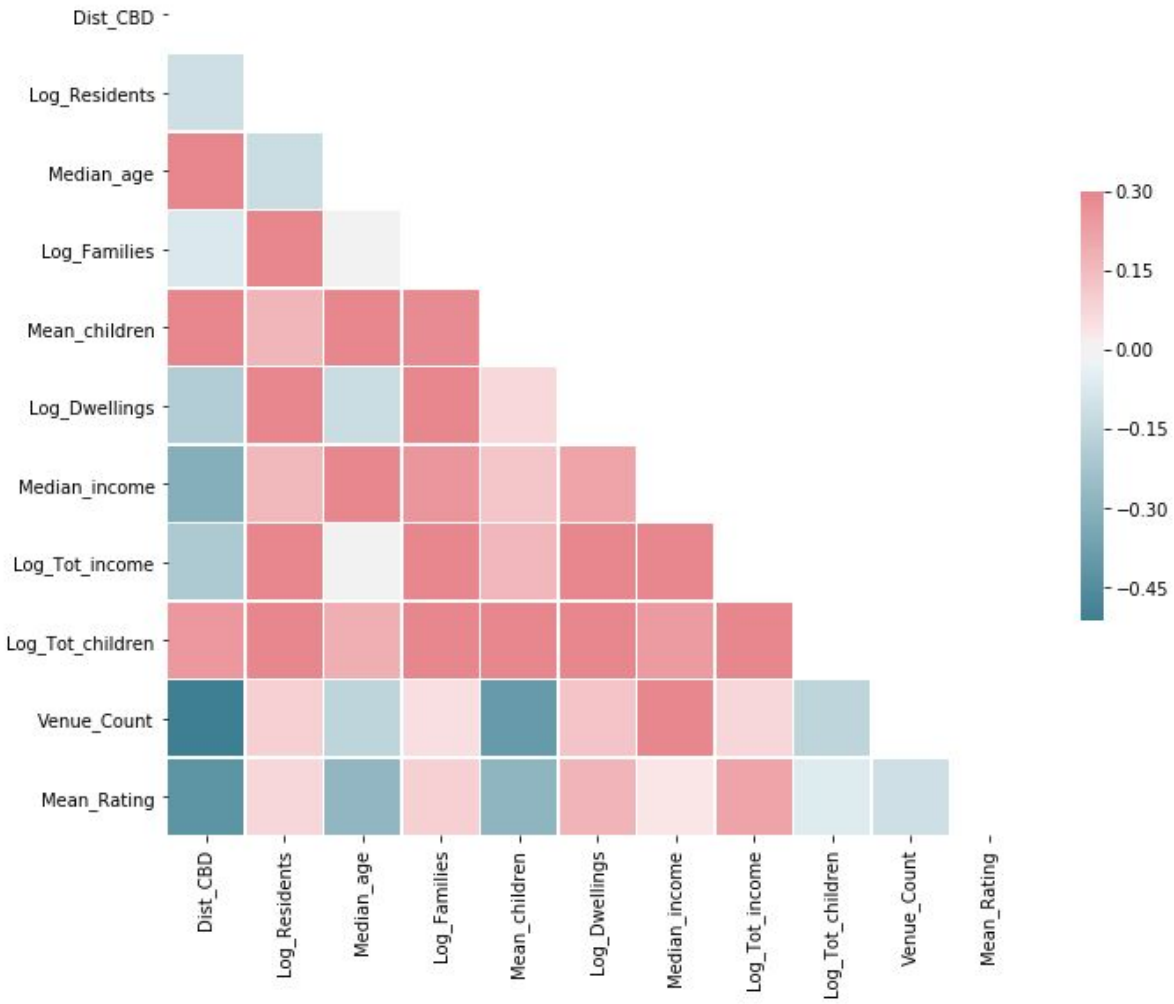

Figure 6: Correlation Heatmap

There were no large correlations between any feature variables. The largest correlations between feature variables and output variables were a negative correlation for both Venue Count and Mean Rating with Distance from CBD. These correlations are examined further in Figure 7.

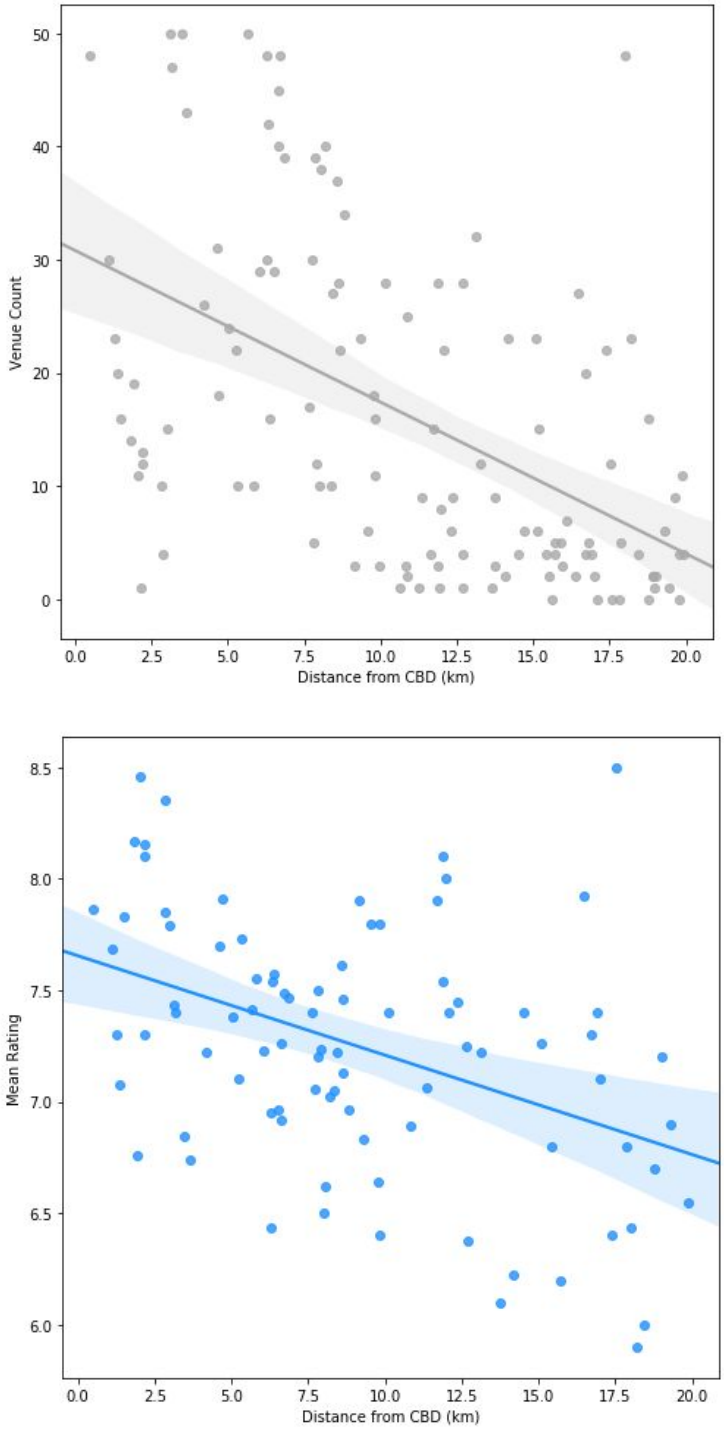

Figure 7: Scatter Plots of Venue Count and Mean Rating by Distance from CBD

This data seems reasonable as more affluent suburbs tend to be closer to the CBD and the people living in these suburbs have greater disposable income available to spend at coffee venues. The number and quality of cafes near the CBD should therefore be higher.

Model Development

For the numeric output variables of Venue Count and Mean Rating, two different regression algorithms were compared: linear regression and random forest regression. These were selected as being complementary models, comparing multivariate linear regression with a non-linear decision tree methodology.

Models were assessed using Leave One Out crossvalidation, which was suitable due to the relatively small data set.

Venue Count

After removing suburbs with missing feature data, 123 suburbs remained for analysis. Crossvalidated predictions by linear regression were compared with actual values for venue count to give the plot of Figure 8. It was clear that one significant outlier was present in the data, with a strongly negative predicted venue count. This was found to be the prediction for Melbourne Airport.

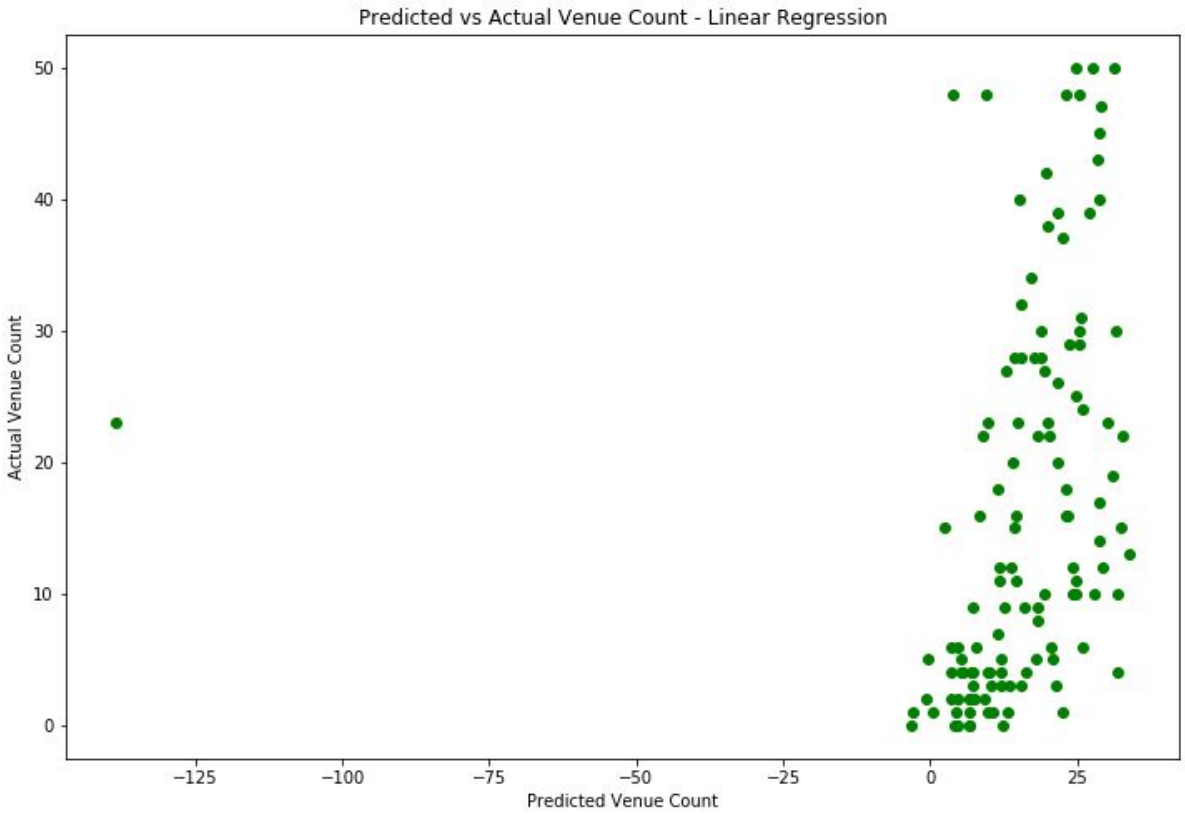

Figure 8: Actual vs Predicted Venue Count for Linear Regression Model

Although the airport has many coffee venues, it has few residents according to census data. Melbourne Airport data was therefore removed from the data set as an outlier, and the analysis was repeated, as shown in Figure 9.

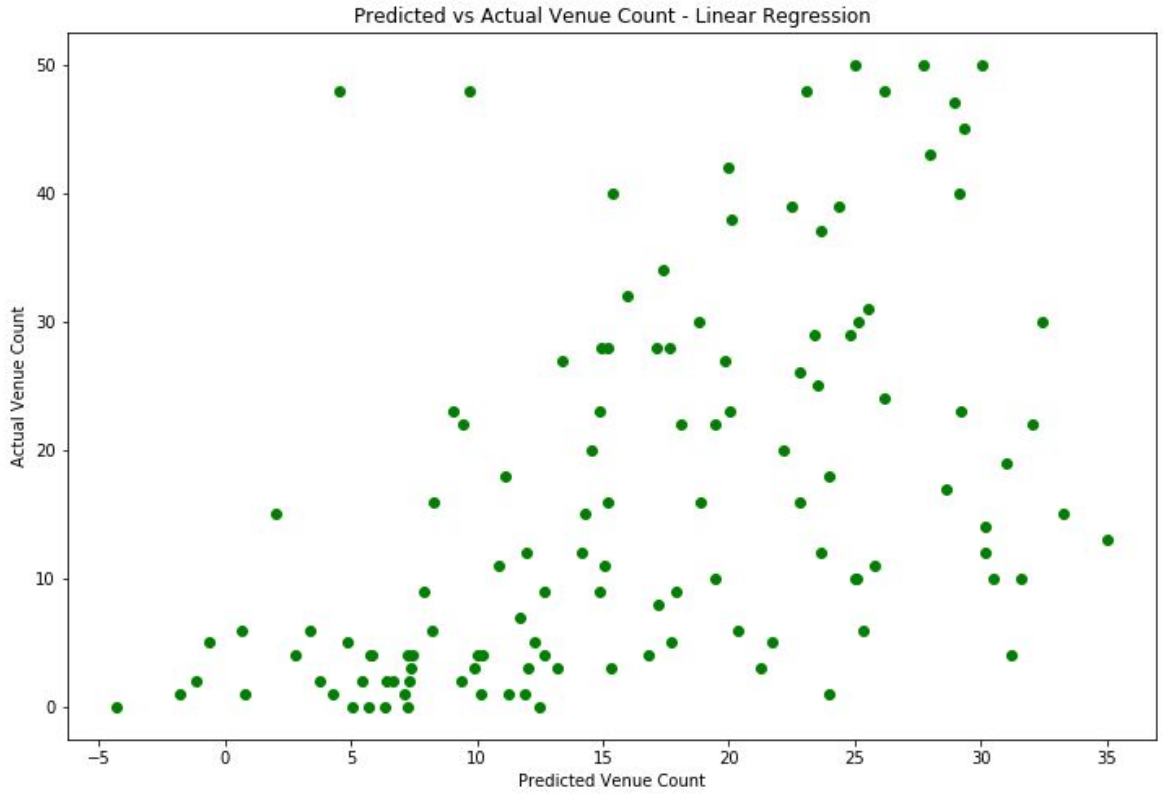

Figure 9: Actual vs Predicted Venue Count for Linear Regression Model

The accuracy of predictions was measured using Mean Absolute Error (MAE); the MAE for the linear regression results was 10.01.

The significance of the individual features was examined by calculating the ordinary least squares regression results of Table 3.

Table 3: Least Squares Regression Results for Venue Count against All Features

OLS Regression Results

Dep. Variable:	Venue_Count	R-squared:	0.385
Model:	OLS	Adj. R-squared:	0.336
Method:	Least Squares	F-statistic:	7.789
Date:	Mon, 09 Mar 2020	Prob (F-statistic):	7.65e-09
Time:	01:24:14	Log-Likelihood:	-470.59
No. Observations:	122	AIC:	961.2
Df Residuals:	112	BIC:	989.2
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.5815	207.802	0.032	0.975	-405.151	418.314
Dist_CBD	-0.5179	0.359	-1.442	0.152	-1.229	0.194
Log_Residents	-86.8060	43.985	-1.974	0.051	-173.957	0.345
Median_age	-0.6324	0.374	-1.693	0.093	-1.373	0.108
Log_Families	-64.5097	45.541	-1.417	0.159	-154.744	25.725
Mean_children	-18.2470	17.616	-1.036	0.303	-53.151	16.657
Log_Dwellings	86.8044	68.359	1.270	0.207	-48.639	222.248
Median_income	0.0053	0.020	0.266	0.791	-0.034	0.045
Log_Tot_income	21.6656	69.698	0.311	0.756	-116.431	159.762
Log_Tot_children	45.4209	19.022	2.388	0.019	7.731	83.111

Omnibus:	6.202	Durbin-Watson:	2.334
Prob(Omnibus):	0.045	Jarque-Bera (JB):	5.927
Skew:	0.535	Prob(JB):	0.0516
Kurtosis:	3.145	Cond. No.	3.59e+05

The relatively high p-values obtained for Median income and Log Total income variables suggested that the model may be improved by removing these features.. When crossvalidation was repeated with these features removed, the MAE of the results was 10.65, so the error increased. It was therefore decided to keep all the original feature variables in the analysis.

A random forest model was developed for the original data with Melbourne Airport removed, using 1000 trees per model. Crossvalidated predictions for random forest were compared with actual values for venue count to give the plot of Figure 10.

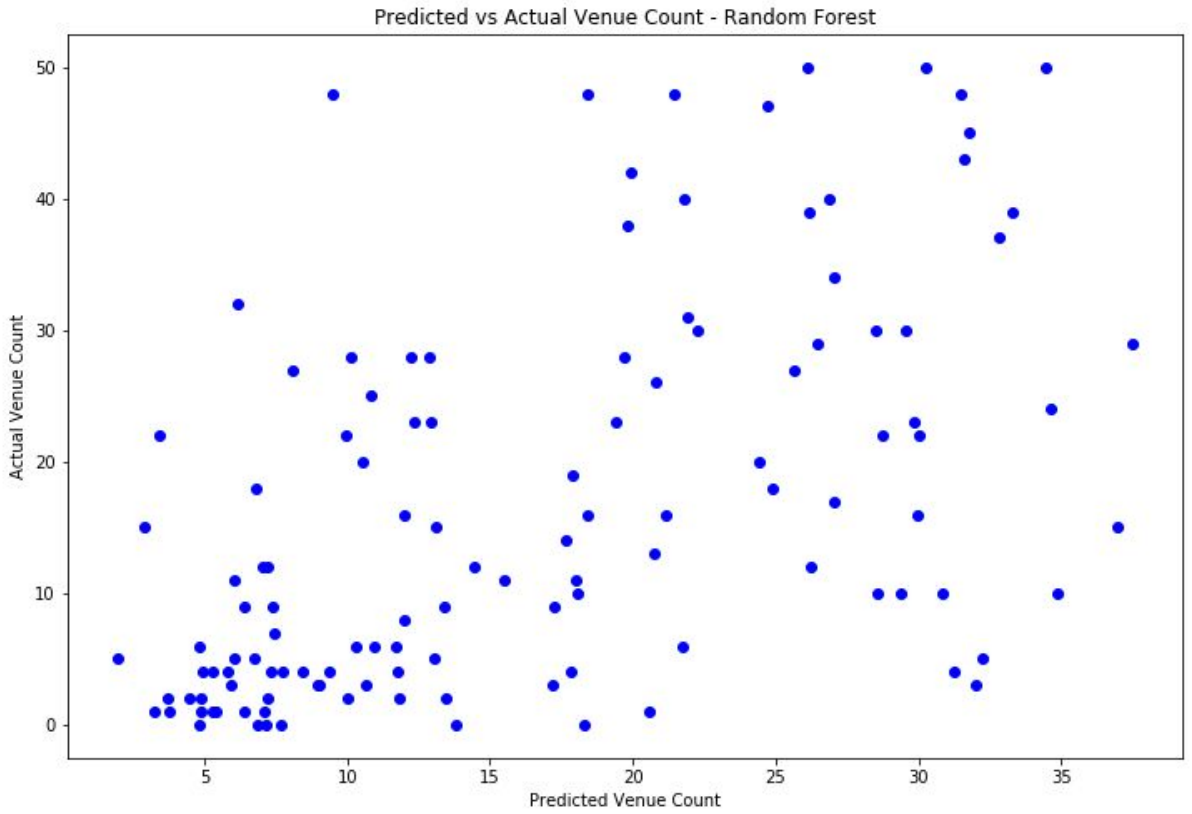

Figure 10: Actual vs Predicted Venue Count for Random Forest Model

The MAE for the random forest results was 9.80. The random forest results were slightly more accurate than the linear regression results overall.

It was hypothesised that the mean of the two model predictions may be a better predictor than either of the individual predictions. The MAE for the mean prediction results was 9.63. It was therefore decided that the mean of the two models would be used for the suburb analysis.

Mean Venue Rating

After removing suburbs with missing Mean Rating data, 87 suburbs remained for analysis. Crossvalidated predictions for linear regression were compared with actual values for mean rating to give the plot of Figure 11.

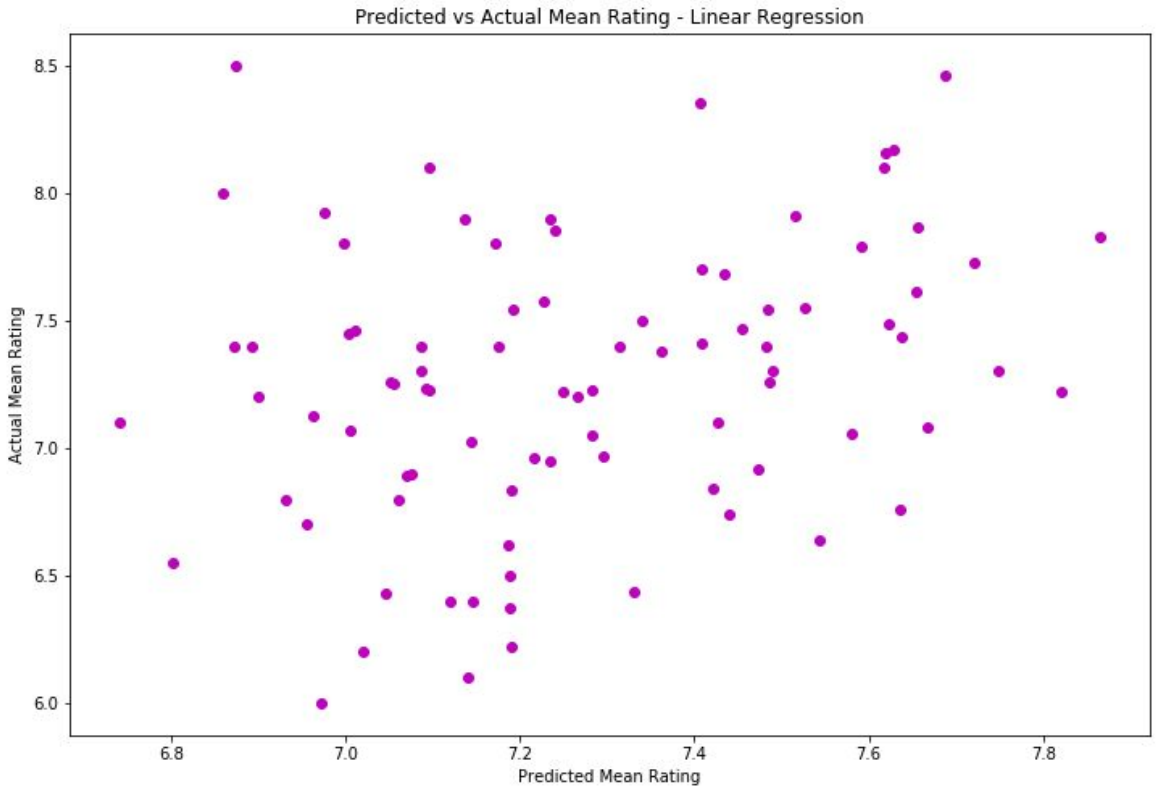

Figure 11: Actual vs Predicted Mean Rating for Linear Regression Model

The MAE for the linear regression results was 0.42.

A random forest model was developed for the same data using 1000 trees per model. Crossvalidated predictions for random forest were compared with actual values for mean rating to give the plot of Figure 12.

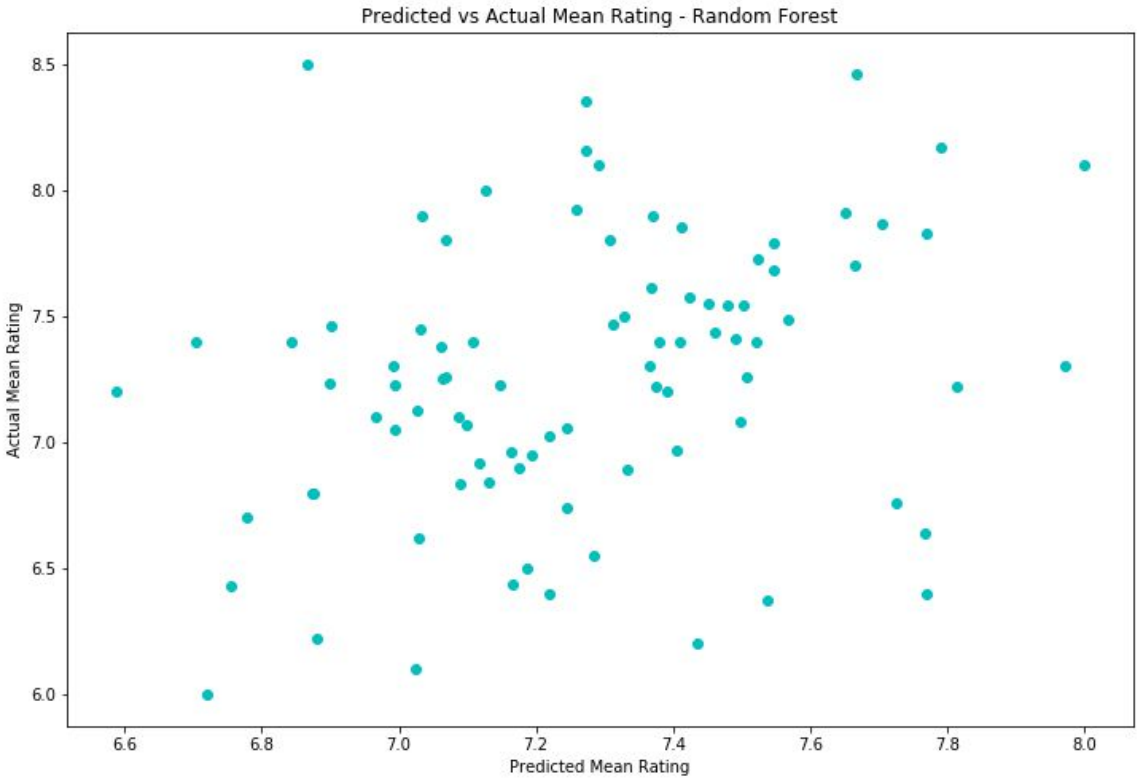

Figure 12: Actual vs Predicted Mean Rating for Random Forest Model

The MAE for the random forest results was 0.41.

Again, the random forest predictions were slightly more accurate than linear regression for the Mean Rating data. However, Figures 11 and 12 show that neither model was particularly predictive. Using the mean of the two predictions gave an MAE of 0.41. In this case, the random forest model was used in the suburb analysis.

Results and Discussion

The Foursquare data for venue count can be visualised on a map of Melbourne, see Figure 13. This map depicts the number of venues in each suburb by colour, ranging from the lightest yellow for the minimum number of venues (0) to the darkest red for the maximum number of venues (50).

Figure 13: Coffee Venue Count in Melbourne Suburbs, see Text

This data illustrates again that the suburbs closest to the centre of the city tend to have a higher number of coffee venues.

When suburb demographic data was used along with distance from the CBD to predict the number of venues for each suburb, the five suburbs with the greatest negative difference between actual and predicted values were as per Table 4.

Table 4: Results for Prediction of Venue Count

Suburb	Actual Venue Count	Predicted Venue Count	Difference
Clifton Hill	4	31	-27
Brooklyn	3	27	-24
St Kilda	10	32	-22
Pascoe Vale	5	27	-22
Parkville	1	22	-21

According to venue count predictions, the suburb with the greatest difference between actual and predicted values is Clifton Hill.

When the mean rating of coffee venues was predicted for each suburb, we see the results of Table 5.

Table 5: Results for Prediction of Mean Rating

Suburb	Actual Mean Rating	Predicted Mean Rating	Difference
Briar Hill	6.4	7.8	-1.4
Bentleigh East	6.2	7.4	-1.2
Chadstone	6.4	7.5	-1.2
Caulfield East	6.6	7.8	-1.1
Docklands	6.8	7.7	-0.9

According to mean rating predictions, the suburb with the greatest difference between actual and predicted values is Briar Hill.

Conclusion

Based on an analysis of suburb demographic data and current Foursquare data on the number of coffee venues and their mean rating, we can conclude that Clifton Hill and Briar Hill could be currently underserved with good quality coffee venues. These two suburbs may be good choices for opening a new coffee venue in Melbourne.