Determining the Association of Availability of Street Parking and Review Stars for Restaurants in Edinburgh

Introduction

The business question to be investigated would be to find out if the availability of street parking has any association on the number of review stars a restaurant in the UK (Edinburgh) receive from restaurant goes.

The study would facilitate businesses wanting to set up a restaurant in the UK to determine the location of their restaurant business. More specifically, whether to set up a restaurant at a location where street parking is available. Also, the result of this investigation could serve as a starting point to other studies on businesses' locations and the review stars it receives.

Methods

Exploring Business Locations

As the data set contains data from 10 cities across 4 countries, we would need to first explore the data to determine if the question can be answered from the data. This was done through processing and subsetting the data set to data that relevant to Edinburgh.

As we would need to subset the data set to contain only data for Edinburgh for the analysis, we will first need to explore the states available in the data set. From the initial plot of the businesses according to their states, it seems that there are more than a state being associated with a city.
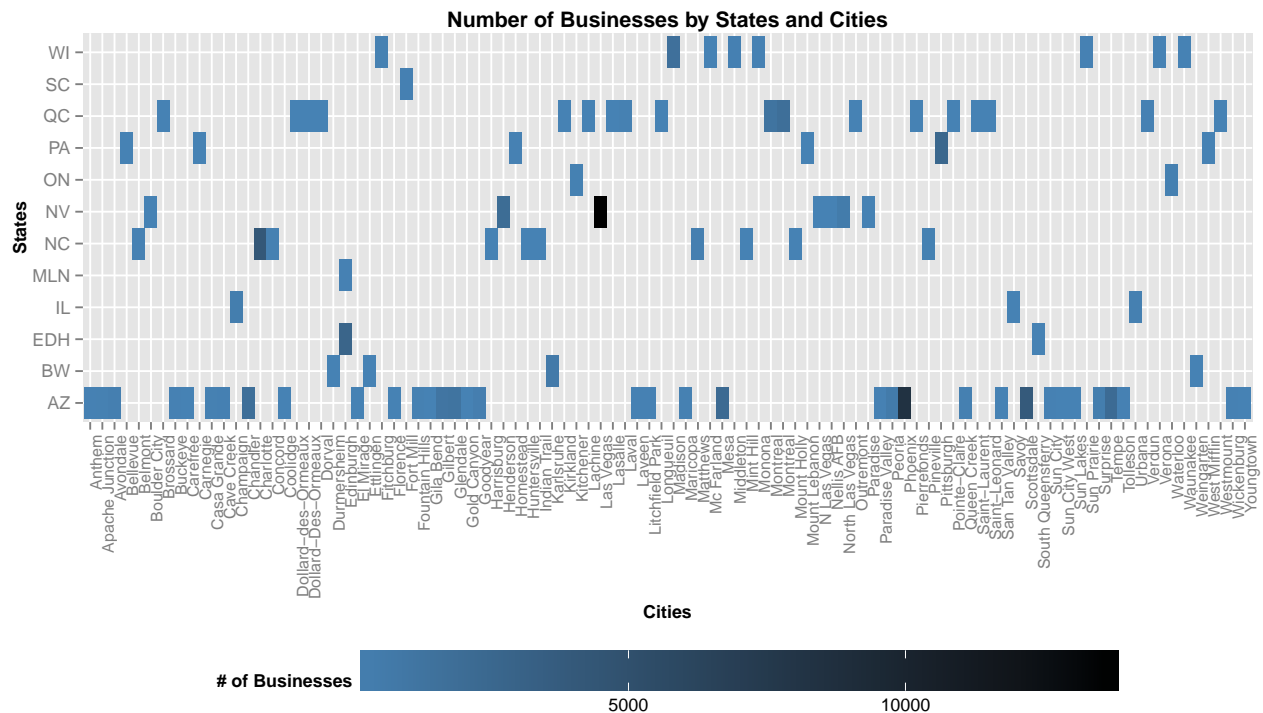

Only cities with more than 10 businesses are shown in the plot

To facilitate clustering the businesses in Edinburgh, we make use of the DBSCAN (link) algorithm to cluster the longitude and latitude data available in the data set. This will allow us to subset the data set using the clusters to business in Edinburgh. From the plot below, we see that cluster 9 probably corresponds to Edinburgh containing the city Edinburgh (EDH), or cities near Edinburgh.
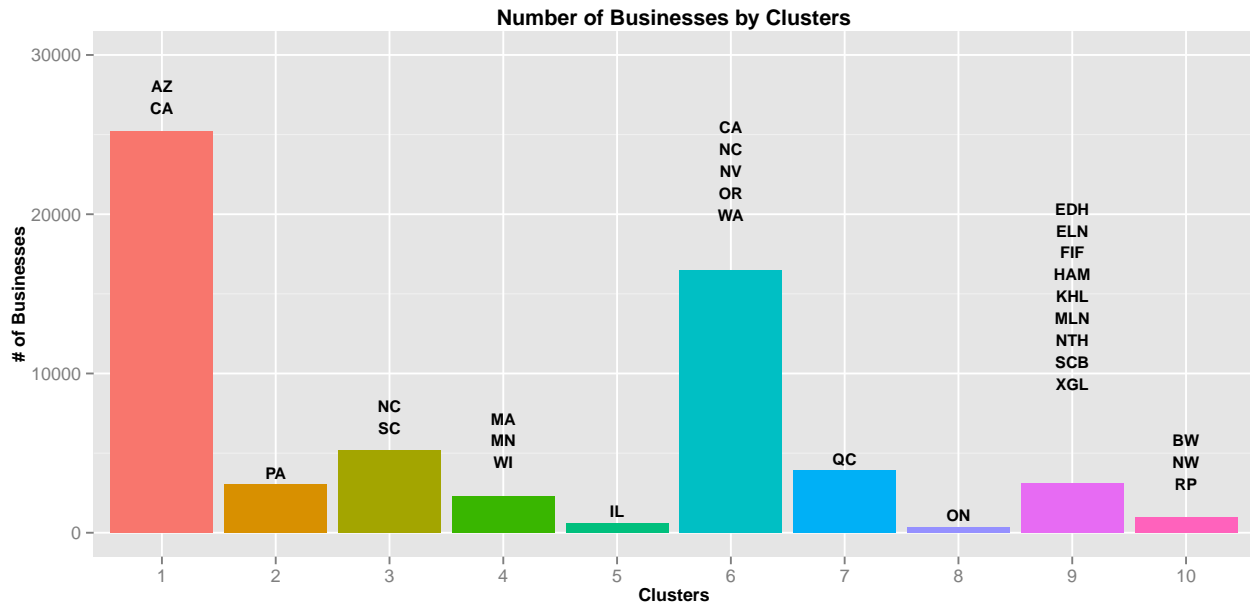
The clustered points are plotted onto a map to visually inspect the results. The plot below shows the points in cluster 9, which corresponds to Edinburgh.
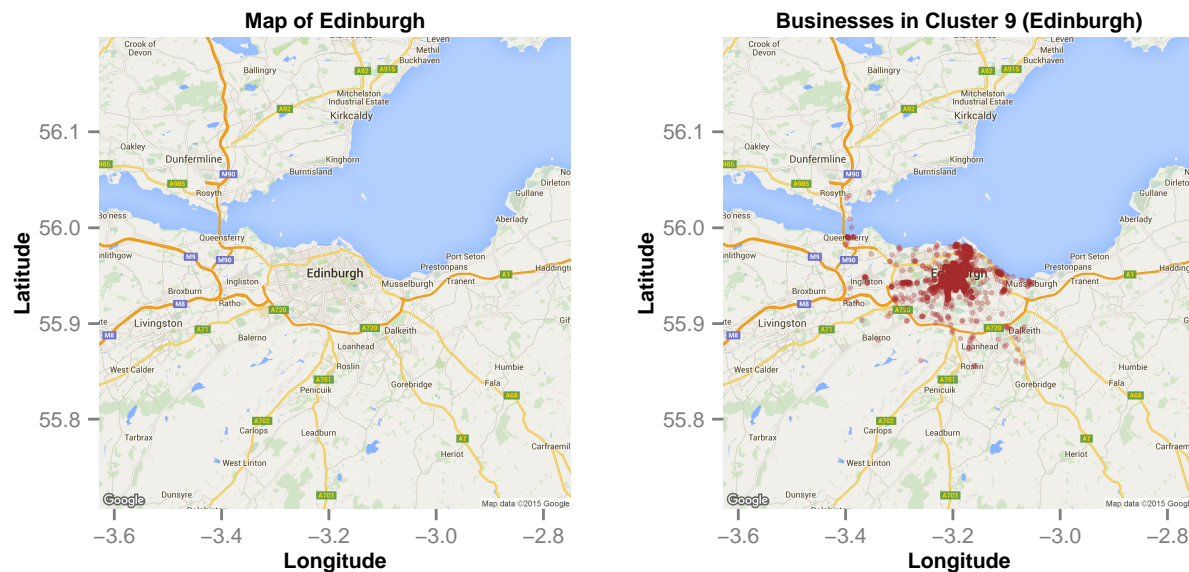
Once the clustering results were validated, we were able to subset the data set for businesses in Edinburgh for further analysis.
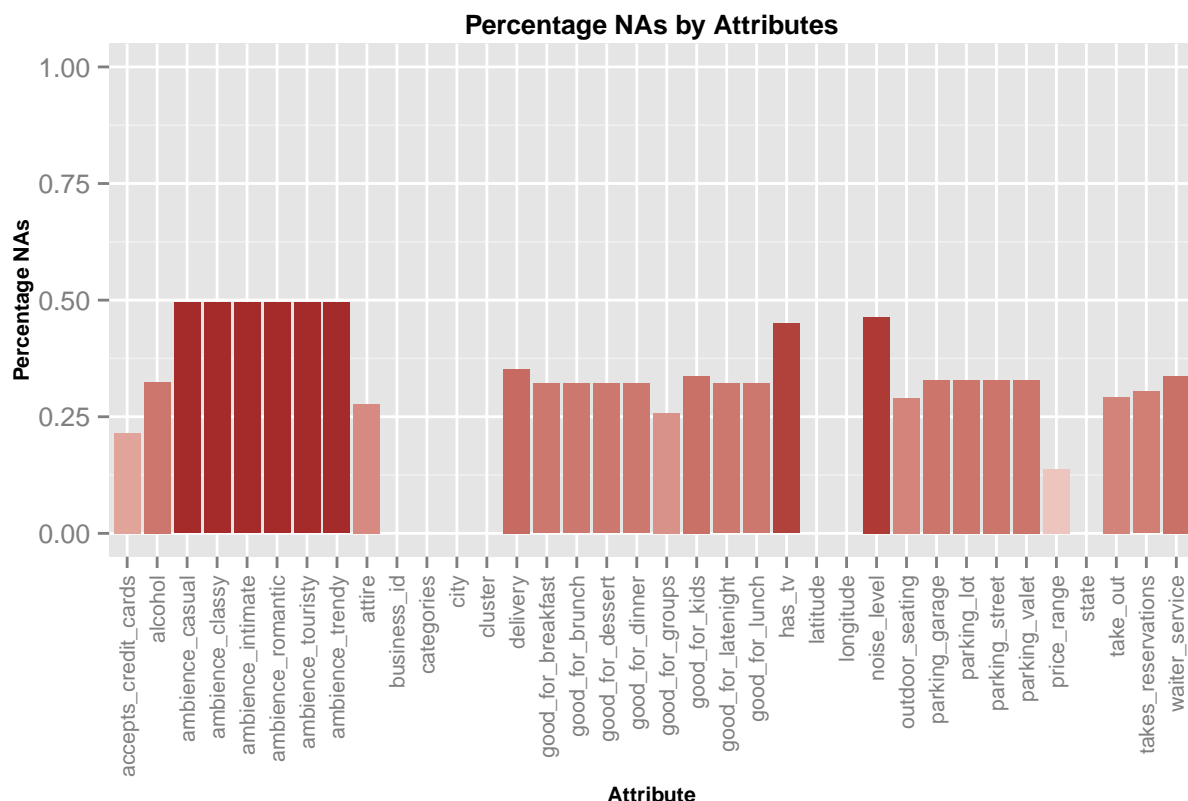

Exploring Business Attributes

Next, the businesses data set was further subsetted to contain only restaurants data and the business ids that was found to be in Edinburgh.

Once we have the Edinburgh restaurants' business ids, we would need to select the relevant business attributes needed for analysis to be merged from the original businesses data set. We do this by first flattening the attributes from the original data set and then merging it with Edinburgh restaurants' data. The attribute names were also made consistent. A sample of the cleansed business attributes is shown below.

```
##                   original_attribute       clean_attribute
## 1                        business_id           business_id
## 2      attributes.By Appointment Only   by_appointment_only
## 3                attributes.Happy Hour            happy_hour
## 4 attributes.Accepts Credit Cards accepts_credit_cards
## 5          attributes.Good For Groups        good_for_groups
## 6          attributes.Outdoor Seating       outdoor_seating
```

We then look at the % of missing values from each attribute. Typically, missing values will be imputed. However, our analysis shows that some attributes contain more than 50% of missing values. These attributes will be excluded from the data set. After excluding these attributes, the attribute *parking_street*, which we will use for our analysis remains in the data set and has more than 50% of the attribute populated.


Finally, reviews were then merged with the Edinburgh businesses data set containing the attributes, such that the stars the businesses received from reviews can be used for analysis. With that, the data is now ready for analysis.

Exploratory Tables and Plots

The contingency table (link) below shows the number of reviews in each star category that restaurants with and without street parking receives in Edinburgh. Missing values for parking_street were excluded from the analysis.

```
##        parking_street
## stars FALSE TRUE  Sum
##    1    157   157   314
##    2    300   375   675
##    3    908  1029  1937
##    4   1686  2433  4119
##    5   1079  1839  2918
##   Sum  4130  5833  9963
```

To investigate further if there are any differences in the distribution of the number of stars awarded between restaurants with street parking and restaurants without street parking, an exploratory plot of the proportion of stars against the availability of street parking was used to visually inspect the data. From the plot, we notice that for restaurants with street parking, there seemed to be a higher proportion of 4 and 5 stars reviews.
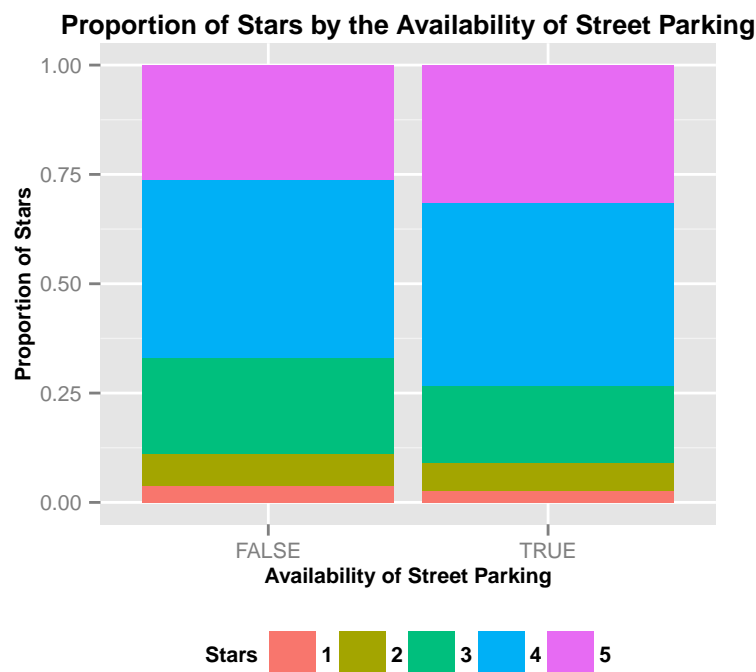
Statistical Inference

The chi-square (link) test (link) of independence can be used to test if there are any associations between the attributes for a contingency table such as the one above. The test was applied to determine if there is any significant association between the 2 variables, parking_street and the review stars.

The null hypothesis states that the availability of street parking (parking_street) and review stars are **independent**.

The alternate hypothesis states that the availability of street parking (parking_street) and review stars are **dependent**.

```
##        parking_street
## stars FALSE TRUE
##    1    157   157
```

```
##     2   300  375
##     3   908 1029
##     4  1686 2433
##     5  1079 1839


##
##  Pearson's Chi-squared test
##
## data:  con_tbl
## X-squared = 59.9617, df = 4, p-value = 2.955e-12
```

The test was appropriate as each review only contributed to a single category (ie. data were not paired). Furthermore, the total number of observations exceeds 20 and the test is able to approximate appropriately. If not, the Fisher's exact test should be used instead.

To measure the effect size (link-pg7), we will use the Cramer's V. Cramer's V (link) is a measure of association for non parametric statistics.

```
assocstats(con_tbl)
```

```
##                     X^2 df   P(> X^2)
## Likelihood Ratio 59.851  4 3.1177e-12
## Pearson          59.962  4 2.9552e-12
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.077
## Cramer's V        : 0.078
```

Results

Based on the chi-square test of independence performed earlier, at 0.05 significance level, we **reject** the null hypothesis that the availability of street parking (parking_street) and review stars are independent with a p-value of 2.9552e-12. Based on Cramer's V, the effect size can be said to be small or negligible.

Lastly, to answer the primary question, our analysis shows that there is a negligible effect of the availability of street parking on review stars although the 2 attributes are statistically dependent.

Discussion

While we have established that the availability of street parking has a small effect on the number of review stars, there may be further other confounding factors affecting the review stars a restaurant may receive. The analysis was limited to a single attribute of the businesses and further work would include identifying themes from the review text to further ascertain if the availability of street parking indeed does affect review stars.