EXPLORING RELATIONSHIPS BETWEEN NYC FOOD OUTLETS AND LOCAL DEMOGRAPHIC

Jason Hullick

COURSERA CAPSTONE

Introduction

Background

New York City is ranked as one of the most diverse cities in the world, featuring a wide variety of people, cultures and language. It is also one of the most populous, with 8.6 million people living within a relatively small area of 783.8 km. This rich diversity and high population density mean that the demographics of the population can vary heavily between suburbs or even street to street.

Food outlets are in abundance in New York, with the popularity and incidence of different food venue types likely having a close relationship with the demographics of the localised population. Having the ability to select the most suitable food venue for an area can be advantageous for developers and future business owners.

Problem

The problem being explored is the extent to which demographics of a localised population affect the incidence and success of different food venues in New York City. Using this knowledge, it can be understood which venues are popular within different demographics and develop a predictive model that could assess the most suitable food outlet given a point in NYC.

Audience

The audience of this would be developers looking to decide on the best type of food venue to develop in their area. The model would suggest the most suitable food venue given their area and local demographics.

Data

The three data sets used are Foursquare location data, NYC census data and NYC census tract geo-data.

Foursquare Data

The Foursquare data is accessed via the Places API using the following endpoints:

- GET /venues/search – To retrieve nearby venues
- GET /venues/ - To retrieve details of the venues including rating and category
- GET /venues/categories – To retrieve category tree

NYC 2010 Census Data

NYC census data was sourced from a Kaggle data set here in csv format. It contains 20+ attributes from the 2010 census grouped by census tract. Attributes are related to demographic and include sex, race, income, poverty levels, employment and commute data.

NYC Geo Data

The NYC geo-data was sourced from the NYC Department of City Planning website here. It is in the Geo JSON format and contains co-ordinates pertaining to the boundaries of census tracts along with explanatory meta data about each area such as Borough name and Census Tract code.

Methodology

Data Cleaning

Using the Foursquare API 'venues/search' endpoint, approximately 300 venue IDs were retrieved. These venue IDs were then used to retrieve the details of each venue including name, rating, category and coordinates with the '/venues' endpoint. At this point, retrieved venues that didn't have a rating weren't collected as rating is the main attribute used to determine success of a venue. The details of each venue were stored in a Dataframe.

The NYC geo data was downloaded into a Geo JSON object. In order to determine which census tract (statistical area) that a venue belonged to, the Shapely python library was used. The census tracts were converted to a polygon, and the coordinates of each venue were checked against each tract to determine which it belonged to. The Census Tract code was constructed using the metadata in the Geo JSON file and appended to the venue Dataframe.

The census tract data including demographics was downloaded into a Dataframe. The venue Dataframe and census tract data were merged using the census tract code as the key.

A range of variables were dropped from the table such as commute statistics and type of buildings, leaving only key statistics relating to race and income. Using Foursquare's category tree retrieved via the 'venues/categories' endpoint, the category of venues were converted to their most top level category ie. 'Chinese Restaurant' to 'Asian Restaurant'. From this table, only rows with categories in the top five categories were kept. As this is a classification problem, having too many classifications can make it difficult to achieve an accurate model.

As the analysis is looking to identify the relationship between demographics and successful venues, all venues with a rating below 7 out of 10 were dropped from the table. The resultant data frame looks like the following:

Exploratory Analysis

Put a map of NYC with census tracts, then add points to it.

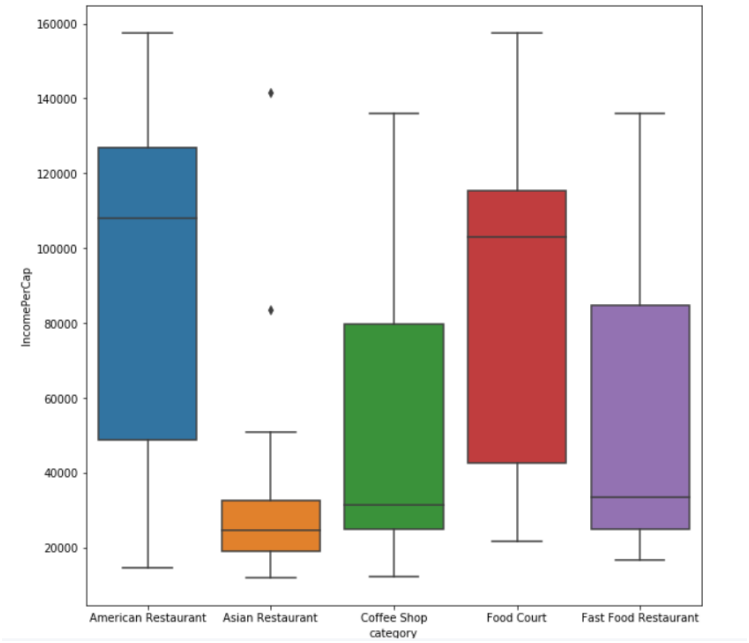
If we look at the value counts, we can observe the categories with the most successful venues from the sample.

Categories of High Rated Venues

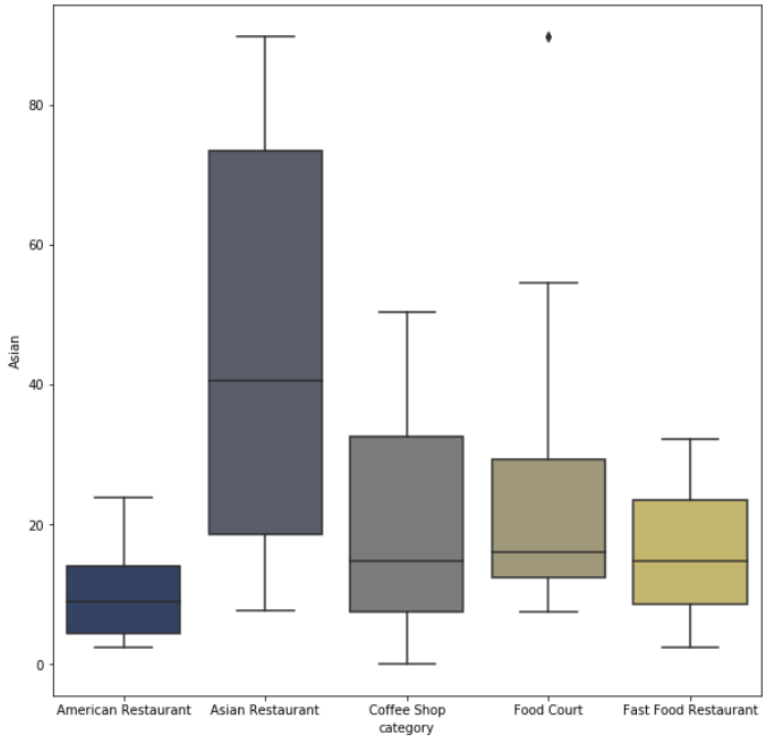
```
Coffee Shop            27
American Restaurant    15
Asian Restaurant       14
Food Court             11
Fast Food Restaurant    3
```

Using the Seaborn library, we can use boxplots to plot the incidence of each category of successful food venues against the population of different races in the area. From these charts, trends can be observed.

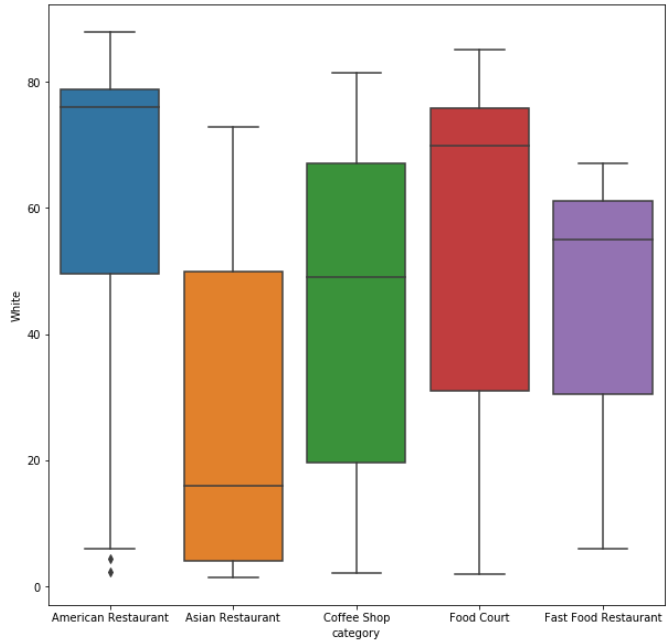
Categories of high rated venues with Income per Capita of local area



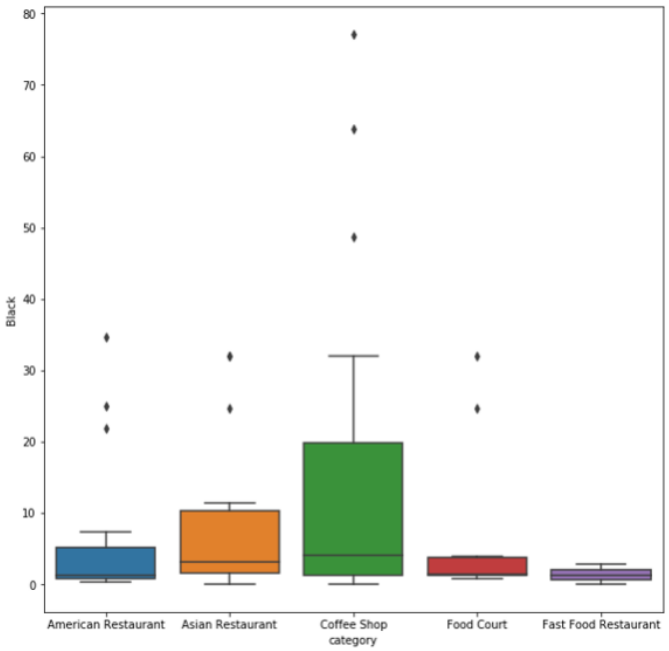
Categories of high rated venues with Asian Population of Local Area



Categories of high rated venues with White Population of Local Area



Categories of High Rated Venues with Black Population of Local Area



Machine Learning

Using the dataset, a machine learning model can be trained to suggest which venue type is likely to be successful given an area. As the model is aiming to predict a categorical variable, a classification model is required. The K Nearest Neighbours (KNN) model was chosen due to its' simplicity and performance in classifying data. The data was first normalized, before being split into an 80% train and 20% test set. To determine the optimal number of neighbour for the model, the model was trained from k =1 to k=10 and the most accurate value was chosen.

Results

From the boxplot visualisations, we can identify a few key relationships in the data. Firstly, the majority of successful Asian restaurants can be found in census tract areas with relatively high Asian populations. The median percentage of Asian population in areas surrounding successful Asian restaurants is 40.6%. You are also less likely to find a successful Coffee shop or American restaurant in areas of high Asian concentration. Similarly, we can see that successful American restaurants are found predominantly in areas with a high white population.

When considering income there is a fairly even spread for each category, however one key insight is that there aren't many highly rated Asian restaurants in high income areas. The KNN machine learning model was found to be optimal with 7 neighbours and has an accuracy of 42% in predicting the type of successful venue contained within an area.

Discussion

Looking at visualisation results, it's apparent that the majority of highly rated Asian Restaurants were found in areas of high Asian population. You are also less likely to find successful Coffee shops or American restaurants in these areas. This may indicate that Asian restaurants in these areas are likely to be more authentic with the business being operated by local staff. These venues would be ones that are regularly attended by Asian constituents who have high knowledge and expectations for the cuisine which may have driven the quality up. Overall, this knowledge could be useful for potential new business owners or consumers looking for where to find the best Asian food.

The visualisation also shows that American restaurants are found predominantly in areas with a high white population. This could be interpreted in a few different ways, but shows that these areas are ideal for American Restaurants.

The KNN model is useful in making a prediction of the category of a food venue in a given area. By looking at the demographics of the surrounding areas of existing successful food outlets in NYC, it can be used to predict the category of an unknown successful venue. It had an accuracy of 42% which is 22% better than a random prediction. This is an acceptable level of accuracy given the fact that there is the possibility of multiple types of successful venues within an area. This also means that while the model is useful, it only provides one suggestion where there actually might be multiple categories that would be successful in a given area. A developer could use this model by inputting their census tract and use the output as an informed suggestion on the type of food outlet to invest in.

Conclusion

The analysis reiterates the close connection different cultures have with their food, with the race of a small areas having an effect on which food outlets were successful there. Particularly, Asian communities were shown to host a majority of successful Asian restaurants. The relationship between demographic and successful food venues was reiterated by a 42% accurate predictive model, which has implications for developers looking to invest in an area.