Targeted Marketing Campaign Prediction & User Behavior Analysis using K-Means Clustering & Folium For Data Visualization

This is my final capstone project for IBM Applied Data Science Capstone course in Coursera, apart of the IBM Data Science Specialization Certification.

1. Introduction — Business Understanding

1.1 Background

There has been an evolution of marketing. Since the 1900s starting on both radio and television, the marketing focus was on “selling” starting. The Golden age of Advertising introduced such ads as “Uncle Sam Wants You for the Army” and “Eat Your Wheaties”. Marketing became more personalized with a focus on brand awareness and problem solving. Then there was the digital ad revolution that began with online advertising in the 1990s and mobile ads in 2000. There is a plethora of data collected daily about users and the ability to harness this data to produce more targeted and personalized ad campaigns to create better customer experience and revenue generation.

1.2 Business Problem

An employee at a fictitious big data marketing company, Insights LLC has been tasked with helping its customer determine an ideal marketing campaign in San Francisco to increase revenue & customer satisfaction.

1.3 Interest

Insights, LLC has a customer who would like to create more personalized ad campaigns for its target customer segments. With the plethora of data collected & speed in which it is collected on its customers, the ability to harness it for either a) an increase of revenue via new products/services b) identification of user behavior for both positive & negative trends in customer satisfaction. I am using the data science methodology to solve this business problem.

2. Data Science Methodology

2.1 Data Requirements — Data Tooling, Sources of Collection & Cleansing/Pre-processing

The data tooling I will be using will be Python language for (data cleansing, data manipulation, data modeling, data analytics & visualization), Jupyter notebook within Watson Studio for sharing code & data analysis pushed to GitHub for source control.

The customer has asked me to gather insights for the city of San Francisco & come up with a targeted marketing campaign so I am using the following data to solve this problem:

- Web scrape: Neighborhood data for the various cities & population.
- Nominatim: Retrieval of latitude and longitude of the neighborhoods for neighborhood segmentation via clustering
- Foursquare Places API: venue, rating data for these neighborhoods
- Foursquare check-ins/Cities/POI CSV file(s): To show frequently checked in venues & their cities
- Kaggle datasets CSV: SF crime data

The data that I will be using will be both structured & unstructured. I created a Foursquare developer account and used API credentials to retrieve data.

Features within the data (dependent variables)that will influence the marketing campaign:

- venue id: unique id for the restaurant
- venue category: type of venue
- rating of venue: indicator of how successful or good the venue is
- crime description: details type of crime(violent or non-violent)
- venues nearby a specific neighborhood
- venues most frequented per neighborhood
- population of neighborhood: count of people that reside within a given neighborhood

I decided to use the San Francisco neighborhood of “the Castro” as a starting point for a potential marketing campaign. I pulled a listing of nearby venues within a 1000 mi. radius of my chosen neighborhood “ the Castro”, the venue category and store it in a dataframe along with geographical coordinates and rating.

For the sake of assumptions, we will assume the data has a greater gap in lowest to highest rating and I pull the highest rated venue that is an eatery, gym or coffee shop etc. for analysis later.

I took San Francisco crime data (description of crime, neighborhood, geographical coordinates, dates) and stored into a pandas dataframe & San Francisco neighborhood data (zip codes, neighborhoods & population) and stored into a pandas dataframe.

2.2 Exploratory Data Analysis & Machine Learning Methods

Within the dataframes I removed missing data, duplicates, anomalies, corruption using Python. I created two columns in neighborhood dataframe named lat and long and used a loop to populate that data from Foursquare API. I also renamed columns and perform some merging of dataframes (neighborhood, crime) to create a map for exploration.

I store check-in data of venues with more than 10000 check-ins to identify popular venues and their location, I reduced this data set and only pulled a days worth of data(stale data) for analysis due to enormous size.

I map location of the Castro & venues within a 1000 mi radius for potential venue categories in the area and the opportunity for a new venue of a popular venue category.

I cluster high population neighborhoods (>30000 people) using Folium to identify prime locations for a potential venue.

I map crime data to venue data in Folium to identify potential opportunities with a specific venue located near a high crime area as potential to be located in another area for success.

I pull the JSON details of the most popular check-in in San Francisco to find out what the venue is and where it is located.

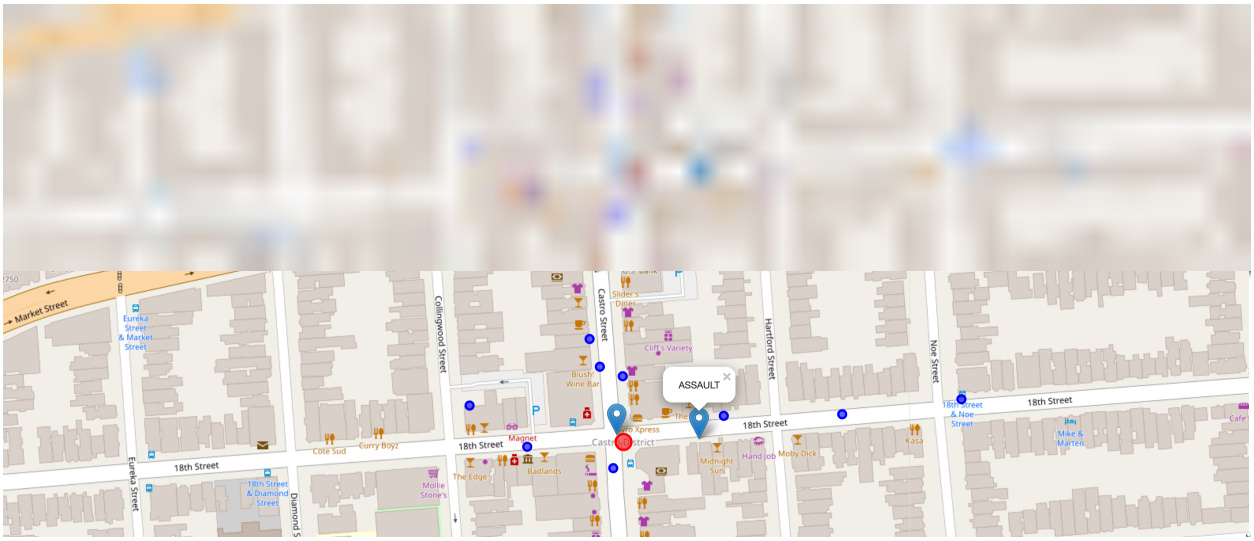
I use one hot encoding to convert the categorical values to numerical values & K-means clustering algorithm (unsupervised machine learning algorithm) to cluster venues that are similar in nature to determine where similar venues are located for a potential new venue in the same area.

2.3 Results

Neighborhood Clusters

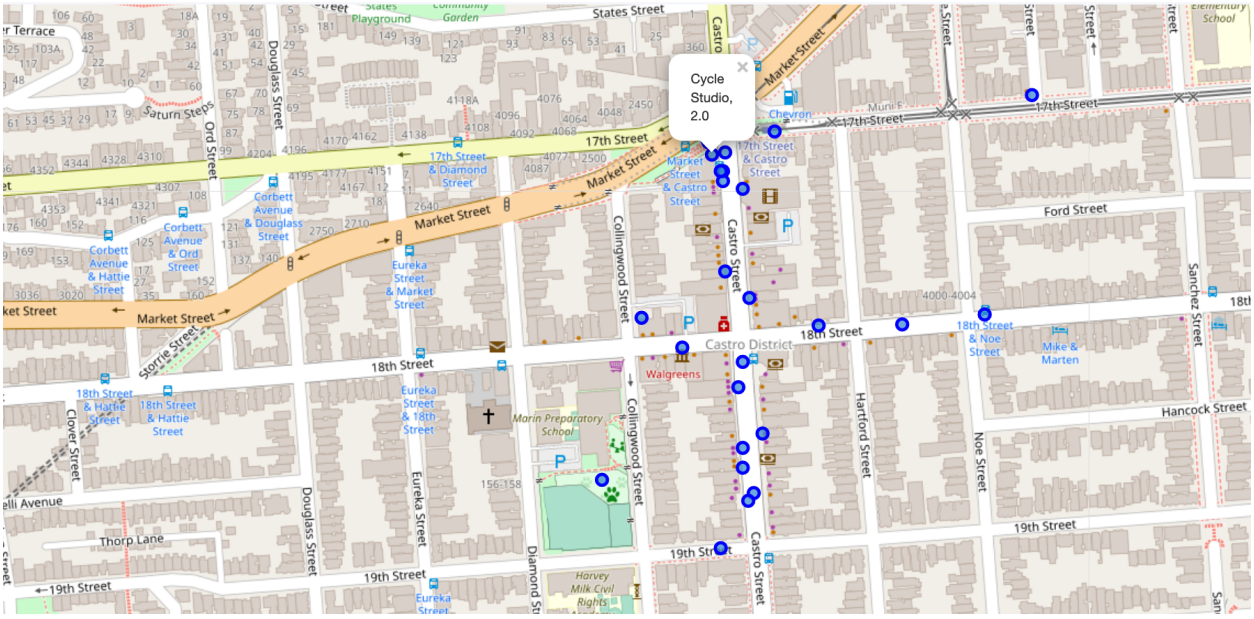
This map above shows clusters of neighborhoods with population > 30000, so high population areas that would be prime for a venue. Ignatius Heights & Van Ness are clustered in a closer

proximity. StonyHill & Visitacion are also clustered in closer proximity. A recommendation would be to choose Ignatius Heights & Van Ness cluster, specifically Ignatius Heights as it is close to a park which could be prospective customers.



San Francisco Crime Near Castro Neighborhood

As we choose our Castro neighborhood to investigate nearby venues. We layer the crime data and see there were several violent crimes near areas where bars are located. There was also a violent crime near a venue near a bank. These can be future considerations of where to not place our new venue.



Similar Venue Data Clustering

Restaurants are clustered together in the Market & Castro area which indicate a good potential venue close to a VTA (public transportation station).

2.4 Discussion

The recommendation would be to avoid bar or nightclub as there is potential for a violent crime. Recommendation would be to target a marketing campaign for a new restaurant located in Market & Castro areas & Van Ness. Also observing our check-in data specifically for San Francisco, we find that the most frequently checked-in restaurant is a Pizza Bar which gives us a potential category of venue and marketing campaign we can explore. Lastly, looking at our rating data, a burger place has the lowest rating & oyster bar(highest rated restaurant venue) also gives a potential for improvement of an existing venue.

2.5 Limitations & Expansion of the Project

Crime data was older(stale) but for the sake of true data science you would stream real-time data optimally but it was used to illustrate concepts in the data science methodology. Streaming data is optimal.Also to avoid rate limiting, I performed a manual entry of rating data from the Foursquare API into the pandas dataframe.

To expand on the project I would like to incorporate the user profiles for customers who check-in to the venues and rate them for prediction of future venue check-in or recommendations via linear regression. Income data would have been a great addition.I would like to also incorporate other demographic data such as ethnicity and age to determine restaurant interest in an specific area.

2.6 Conclusion

We have identified a business problem. We have gathered data, cleansed the data,used k-means clustering & visualization to explore the data, make assumptions & provide recommendations.

2.7 References

List of Neighborhoods in San Francisco: http://www.healthysf.org/bdi/outcomes/zipmap.htm

Foursquare Developer Documentation: https://developer.foursquare.com/docs

Github notebook: https://github.com/bmguillo/Python_ML_Project/blob/master/Capstone-FinalProject-Notebook.ipynb

Watson Studio notebook: https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/37a8b0e5-e7ff-4db3-83d6-437d106e0b28/view?access_token=fd38d8a0b0f04e764b84e37c3ddf790bcb0f93dd8214df6d3ec337eb706a808b