FRENCH MUSEUM LOCATION

Mubarak Ganiyu

March 27, 2018.

1.0 INTRODUCTION

1.1 BACKGROUND

United States of America is one of the most diverse countries in the world and a developed nation. It is a country that provides opportunity for different people to utilize their talent optimally. Thus, Pierre Angier, a French artist, thought it would be a good idea to launch his museum in the U.S.A. He intends to move to a lively part of the country with low crime rates and a high population in order to prevent his art works from getting stolen while still having access to a large audience. The objective of the project is use foursquare credentials to build a map while using other databases to carry out a research that will provide the the French artist with the safest lively location in the United States to showcase his art works.

2.0 DATA ANALYSIS

The dataset to be used for the project is the top 100 most populous cities in U.S.A with their respective crime rates. This data is gotten from wikipedia via this link: [Crime rate per 100,000 people](). Then, some columns of the data were dropped and new columns were created. This was done to get the crimes related to possible museum crimes. Cities with less than 1,000,000 residents were dropped because Mr. Pierre is looking for a location that can attract a large number of people. With the use of certain data analysis technique via python, the safest lively city based on crime related to museum crimes was found to be **San Diego**.

Another dataset was created by combing through the internet to find the neighborhoods of San Diego alongside their longitude and latitude, this can be found [here]().Then, foursquare was used to analyze the city better by looking into which neighborhood will be perfect for Mr Pierre to build his museum and store his art

collection. This will be based on how lively each location is based on the number of venues in this location which would give Mr. Pierre a business opportunity to form partnerships with more organization. Here is a table of the features that were dropped and added during different periods of data cleaning:

PERIOD	DATASET	DROPPED	ADDED/CREATED	REASONS FOR DROPPING
1	Crime rates per 100,000 people	'Total', 'Murder and Nonnegligent Manslaughter', 'Rape', 'Total.1', 'Motor Vehicle Theft', 'Arson'	None	These are not crimes related to museum crimes
2	Crime rates per 100,000 people	None	'Total'	This is the sum of the crimes related to museum crimes: 'Robbery','Aggravated Assault', 'Burglary', 'Larceny-Theft'
3	Crime rates per 100,000 people	Rows with 'Population' <= 1,000,000	None	Mr. Pierre is seeking for an area with 1,000,000 people
4	San Diego Neighborhood data	None	'Neighborhood', 'Latitude', 'Longitude'	This is used to access the neighbourhoods in San Diego so for foursquare exploration
5	San Diego Neighborhood data	'Venue', 'Venue Latitude', 'Venue Longitude', 'Venue Category'	None	These are added to enhance the approach of foursquare analysis

6	San Diego Neighbourhood Data	Other neighbourhoods were dropped since they did not have as much venues the top four	Certain neighbourhoods were retained because they had a lot of venues as others: 'Horton Plaza', 'Gaslamp Quarter','College Area' and 'Little Italy'	These neighbourhoods will later be used by foursquare to interpret data
7	San Diego Neighbourhood Data 2	None	A lot of venues were created for each neighbourhood as columns with the rows representing the average kind of venues that exist in this location.	This was created to see the average number of venues each neighbourhood had
8	San Diego Neighbourhood Data 3	None	This is the same data set as the San Diego neighborhood data 2 but this looked but the columns were the top 10 most frequently visited types of venues	This was used to figure out what kind of places gets the most visits in these neighbourhoods.
9	San Diego Neighbourhood 4	None	This was created for by merging San Francisco Neighborhood and San Francisco Neighborhood 3 with an additional column: 'Cluster Label'	Using the cluster labels as from k-means clustering a map was built to show how close the neighborhoods are to each other.

3.0 EXPLORATORY DATA ANALYSIS

3.1 REGRESSION PLOT

A regression plot was created to plot Crime rate per 100,000 people for museum related crimes vs Population to see if there was any kind of relationship between them but there was no relationship as its correlation coefficient was weak (R^2 = -0.3454)
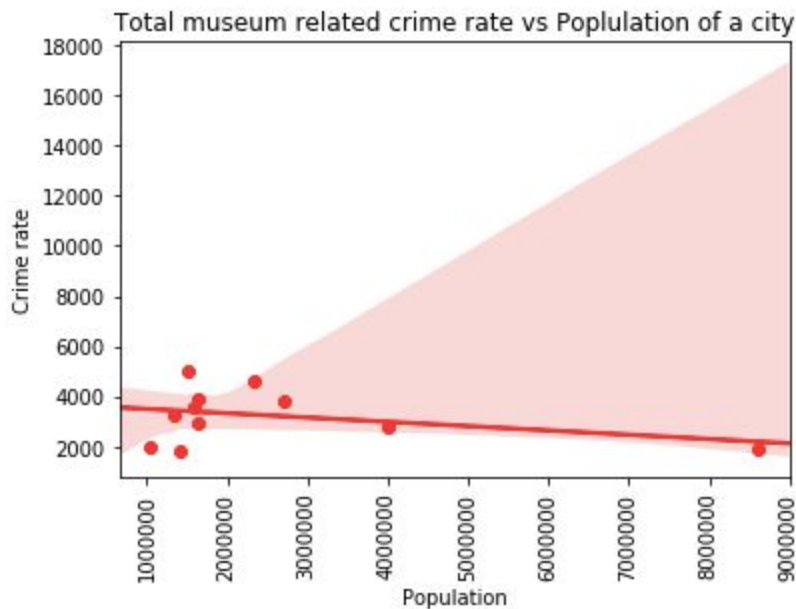

```
[46]:  from scipy.stats import pearsonr
       corr, other = pearsonr(df.Population, df.Total)
       print(corr)
```

-0.3453858231203566

Figure 1. Graph of Crime rate per 100,000 people for museum-related crimes vs Population

3.2 MAP OF THE NEIGHBORHOODS WITH THE MOST VENUES

A map was created with four clusters to highlight the locations neighborhoods with the highest number of venues. The map shows how three of the neighborhoods are closed compared to the other. Gaslamp Quarters, College Area and Horton Plaza are next to each other while Little Italy is far away from them. purple represents Little Italy, red represents College Area, yellow represents Horton Plaza and greenish blue represents Gaslamp quarters
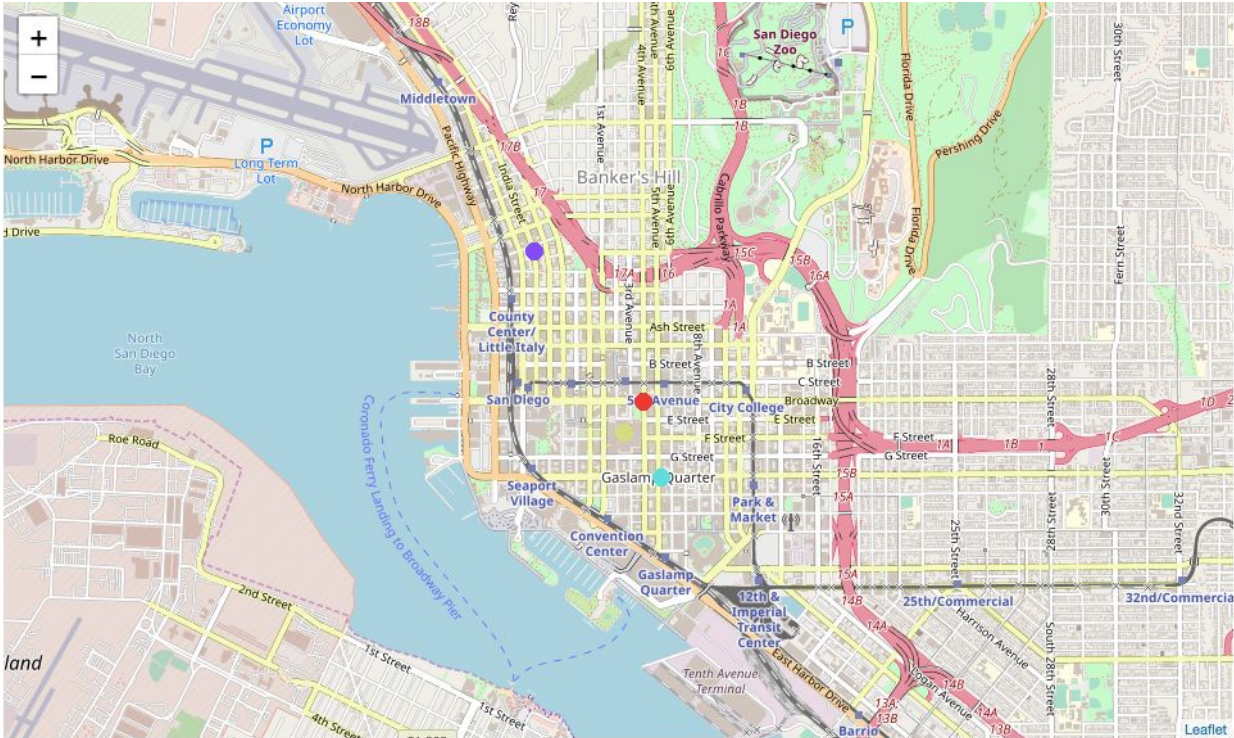
Figure 2. Map of four biggest neighborhoods based on number of venues in San Diego.

4.0 RESULTS

Based on the map, The museum can be built anywhere between the Horton Plaza Neighborhood, Gaslamp Quarters Neighborhood and the College Area Neighborhood as they are all close to each other unlike Little Italy. Moreover, they attract a total amount of 78 unique venue categories which means there are more opportunities for the museum to partner with neighboring organizations to attract more people to its museum. Since, San Diego has a population greater than 1,000,000 people and a low crime rate related to museum crimes compared to other big cities. This means that the museum should experience a high traffic of visits with a low attempt of crime against the museum from criminals given the fact that San Diego is a big city with 1807.30 crime rate per 100,000 people for museum related crimes.

5.0 DISCUSSIONS

The results prove that he has more than one option for a neighborhood since he can decide to build it in any of those three neighborhoods that are close to each other and still get a lot of people coming to view his art collection. Moreover, he can go ahead and start signing up deals with construction companies in San Diego to build his museum.

6.0 CONCLUSIONS

Cleaning and preparing the dataset in such a way that it was only the important features left made it easy to carry out regression analysis of the dataset and find out how population of big cities were impacted by crime. This showed no relationship which means other factors lead to crime within these big cities. These factors were not compiled in the wikipedia dataset. The utilization of foursquare made it possible to streamline the data and create new data frames that was used in building a map to show how far apart the neighborhoods were from each other. Three of the neighborhoods were closely spaced compared to the other which proves that he doesn't get an option rather multiple options. If they had been widely-spaced from each other, then other variables would have been used from foursquare to arrive at the best neighborhood.