Housing Sales Prices and Venue Analysis in California

Tanya Gupta

May 5, 2019

Section 1: Introduction

Over the past twenty years, housing prices in the United States have been somewhat unpredictable. California house prices in the 1990s were far less volatile than they have been in the past ten to twenty years. Following the 90s, house prices increased dramatically as the housing bubble grew and eventually collapsed. The housing market continues to recover today, however, house prices in California in the 1990s were relatively consistent with the inflation rate so I decided that predictors such as house age, rooms, bedrooms, population, number of households in a given area, income, and location would show a stronger correlation with house prices.

Business Problem:

After analysing the median value of houses, it is important to see which cities have more costly houses and the neighbourhoods to those areas along with the popular venues of those cities.

Target Audience:

California is the hub of IT sector and millions of people migrate towards California every year. The audience includes students, brokers, new migrants and property dealers. This data analysis would be helpful for those people who:

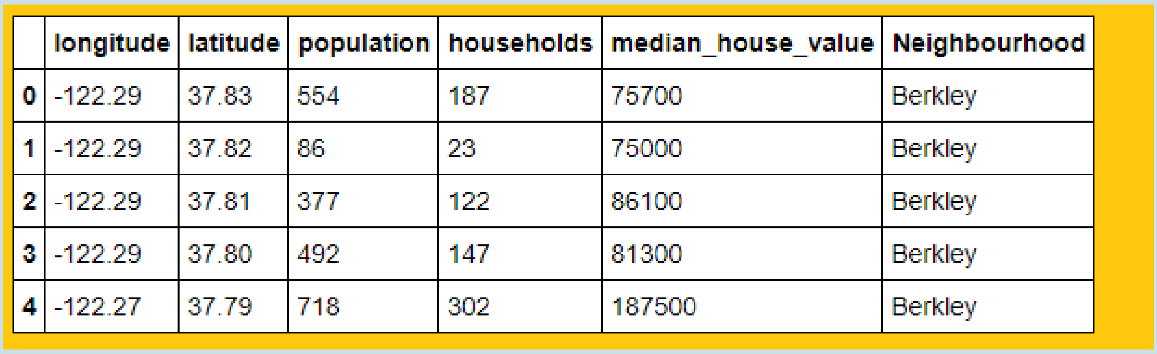
- Wish to Migrate in various areas of California.

- **Want to know the average price of household in California.**
- **Want to know the popular venues near their place of stay.**
- **Want to explore the neighbourhoods and compare to see which one is better.**

Section2: Data Collection

The data section can be divided in two parts:

- **Statistics of California housing:** This data is for analysing which cities have more costly houses, number of rooms, population, households, latitudes and longitudes of various areas of California.
  This data is fetched from https://www.statista.com/
  Statista provides statistics and data within 600 industries and 50+ countries.
  It is fetched under the category of California Housing and the link to the same is: https://www.statista.com/search/?q=california%20housing&qKat=search
  Example can be seen below:



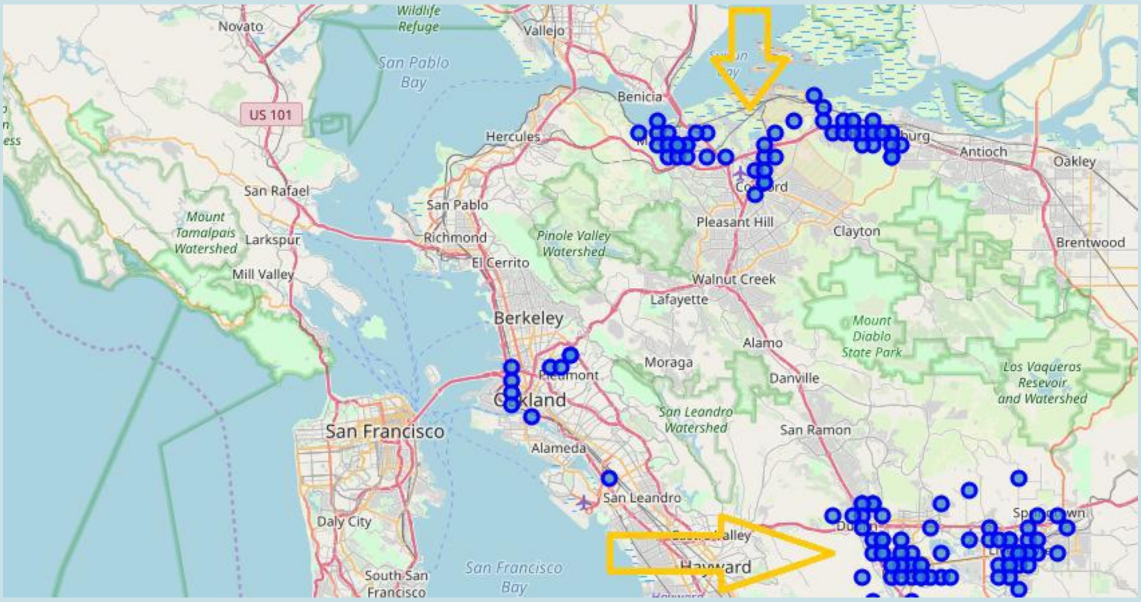
	longitude	latitude	population	households	median_house_value	Neighbourhood
0	-122.29	37.83	554	187	75700	Berkley
1	-122.29	37.82	86	23	75000	Berkley
2	-122.29	37.81	377	122	86100	Berkley
3	-122.29	37.80	492	147	81300	Berkley
4	-122.27	37.79	718	302	187500	Berkley

- **Geographical data using coordinates:** After cleaning the above data, Foursquare API is used to explore the specific areas of California, the nearby venues, analyse all the neighbourhoods and form clusters out of them.

The Foursquare Places API provides location based experiences with diverse information about venues, users, photos, and check-ins.

Additionally, Foursquare allows developers to build audience segments for analysis and measurement. JSON is the preferred response format.

Example can be seen below:



Section3: Exploratory Data Analysis

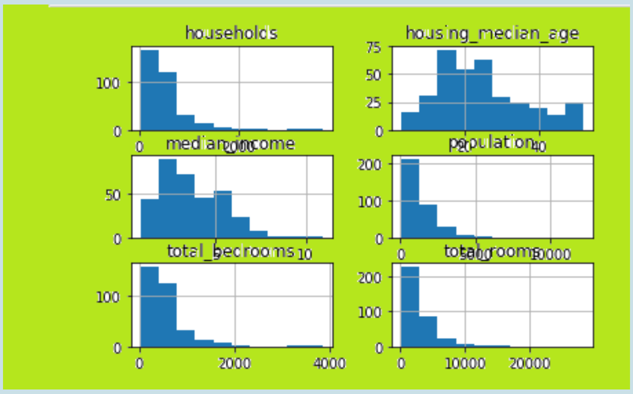
Calculating target variable:

While determining the household prices, there are many factors that play a significant role such as location, number of rooms, population etc. For this analysis, given are the independent variables that are used in the form of predictors:

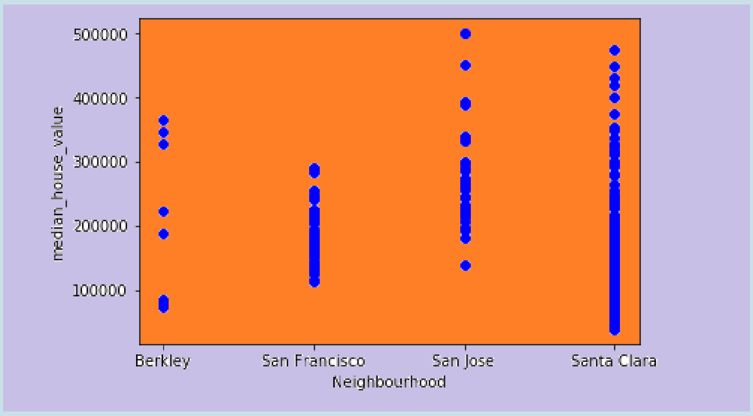
Variables	***Expected Impact***
Median House Age	-
Number of Rooms	+
Number of Bedrooms	+
Population	-
Number of Households	-
Median Income	+

Relationship between features:

The independent features when plotted as histograms, following analysis can be seen:



As we all know, the neighbourhood or the location of residence creates a great impact on price of the housing. The more popular neighbourhood it is, the more value increases. Following shows how much the neighbourhood impacts the pricing of Houses.



The above scatter plot shows us that Santa Clara have a higher linear relationship as compared to other neighbourhoods.

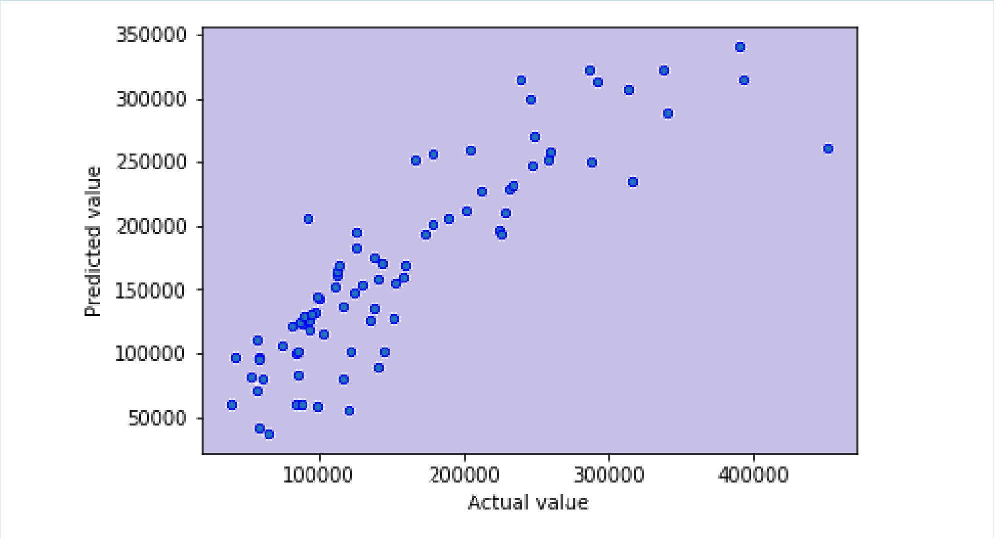


Fig: Relationship between actual and predicted values

Relation between overall features and the target value:

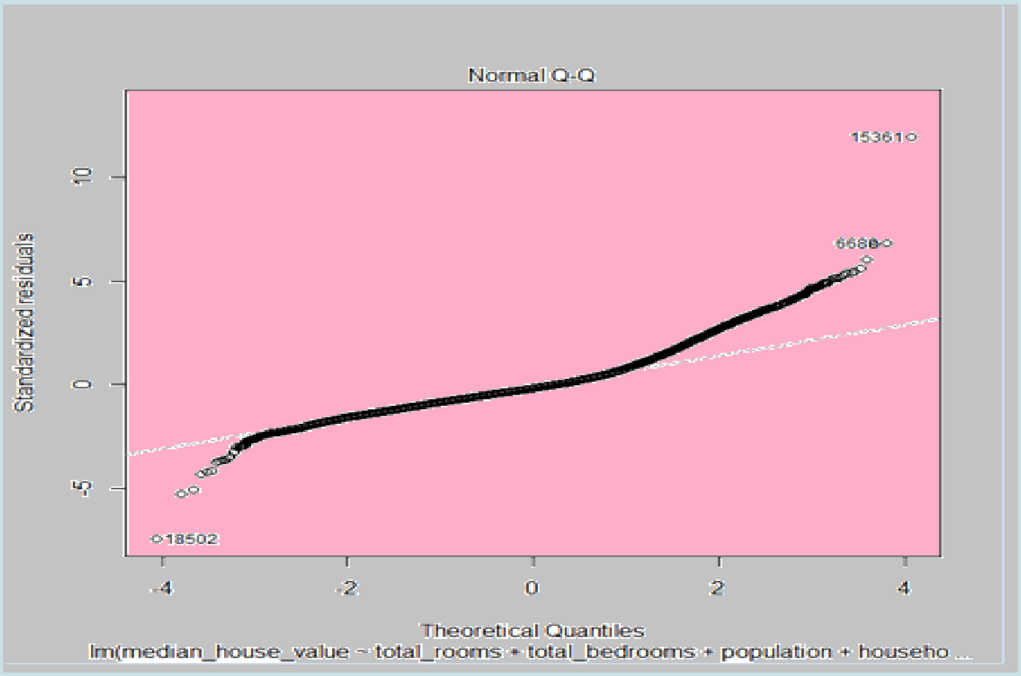
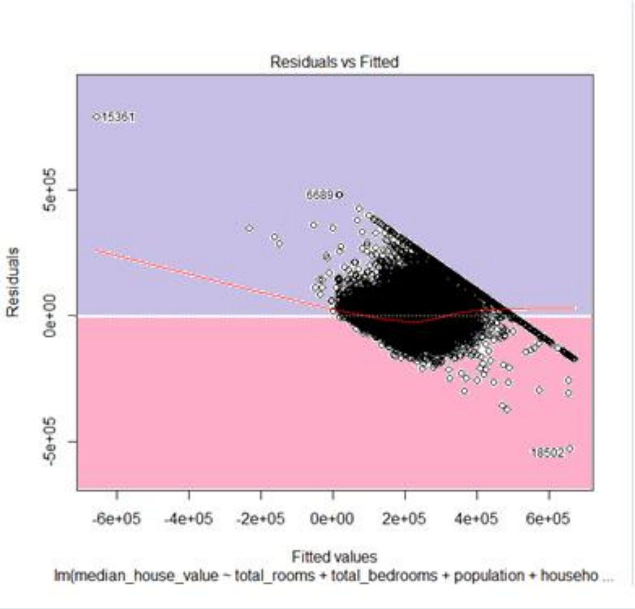


Fig: Regression plot between target and independent variables

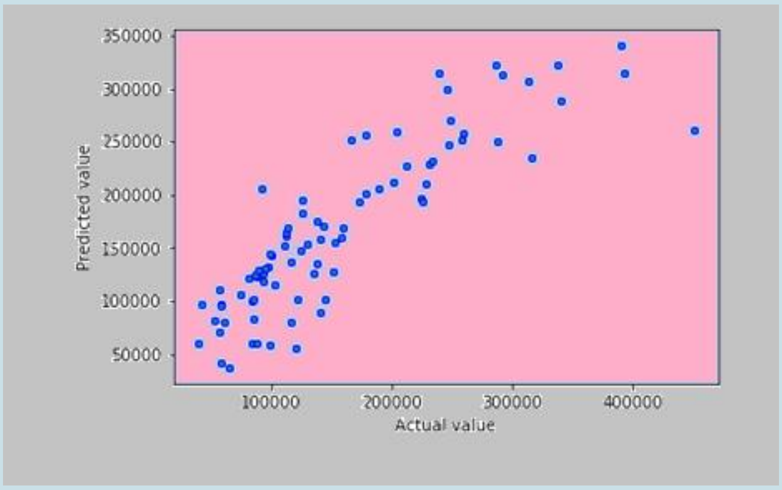
Section 4: Results:

Residual Vs Fitted Plot:



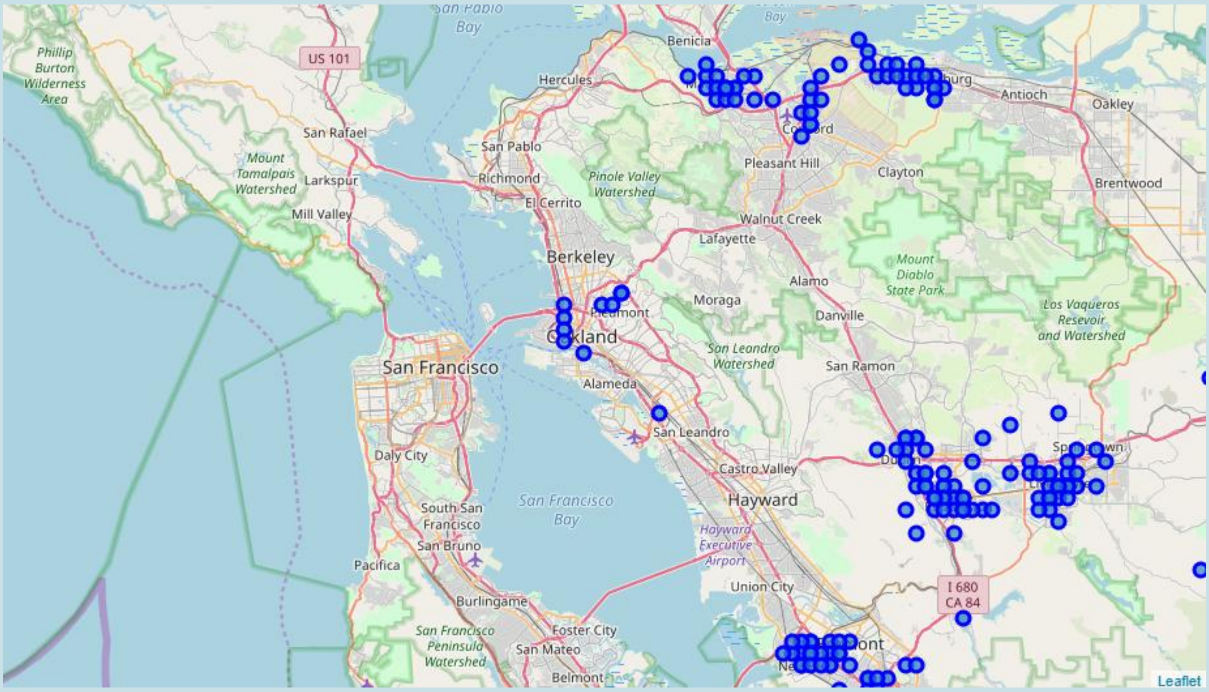
Looking at the Residual vs Fitted plot, we see that the red line (Which is just a scatterplot smoother, showing the average value of the residuals at each value of fitted value) is perfectly flat. This tells us that there is no discernible non-linear trend to the residuals. Furthermore, the residuals appear to be equally variable across the entire range of fitted values.

Actual Vs Predicted Plot:



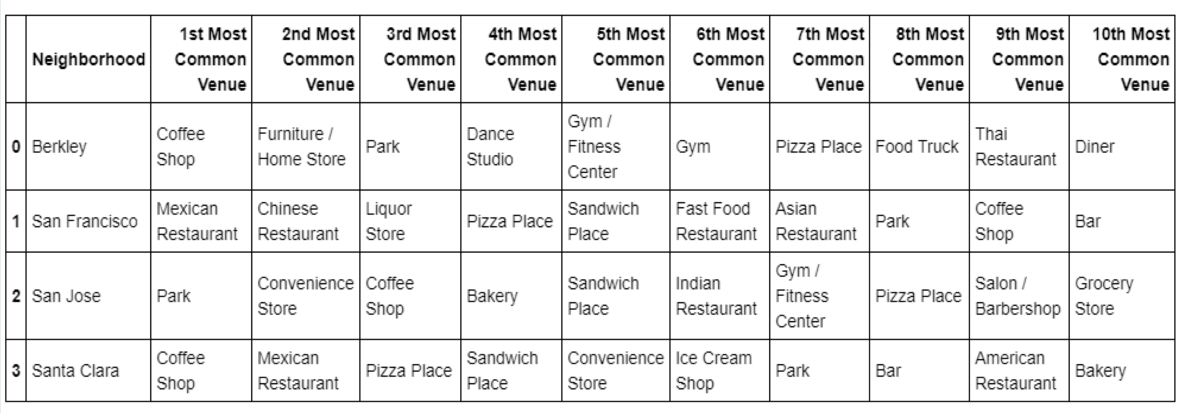
The above scatter plot shows a strong relation between actual and predicted values and the calculations done signify that the predictions and assumptions made were correct upto 75%.

Map of California with the neighbourhoods:



After calculating the median prices, it is necessary to explore various neighbourhoods in California. Above map represents the most popular areas and their neighbourhoods.

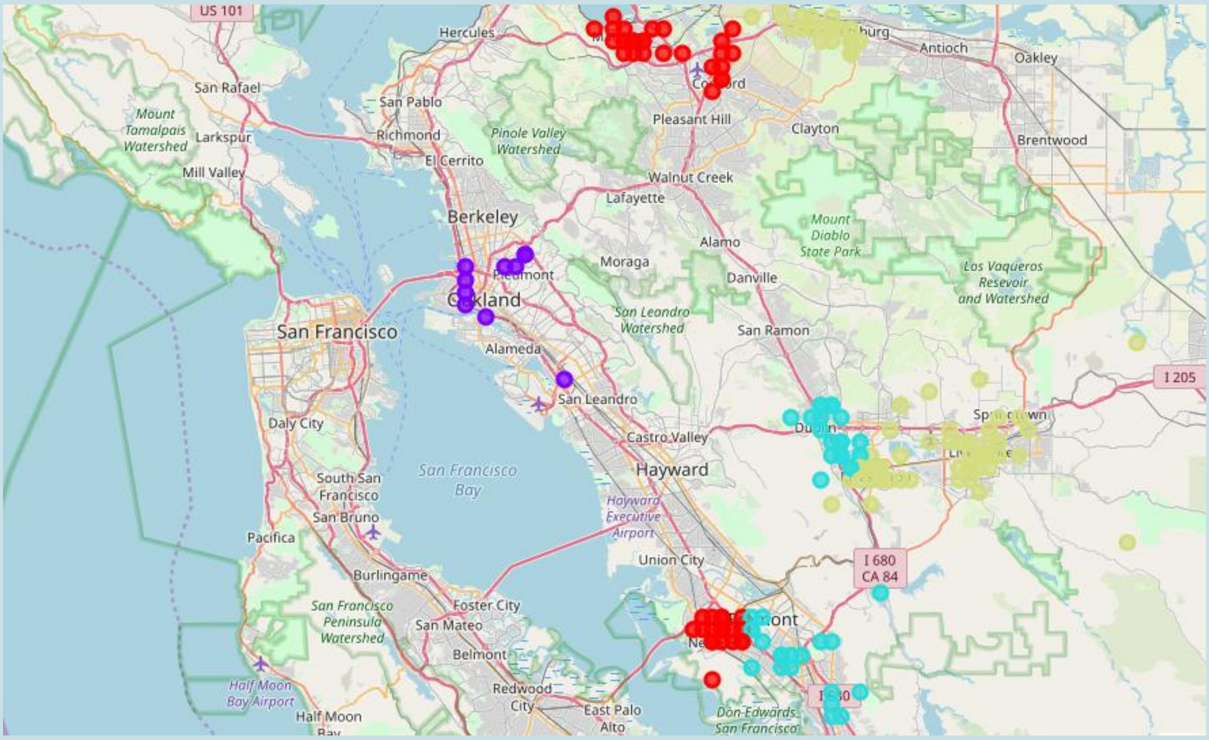
Venue Analysis [Exploring Neighbourhoods]



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Berkley	Coffee Shop	Furniture / Home Store	Park	Dance Studio	Gym / Fitness Center	Gym	Pizza Place	Food Truck	Thai Restaurant	Diner
1	San Francisco	Mexican Restaurant	Chinese Restaurant	Liquor Store	Pizza Place	Sandwich Place	Fast Food Restaurant	Asian Restaurant	Park	Coffee Shop	Bar
2	San Jose	Park	Convenience Store	Coffee Shop	Bakery	Sandwich Place	Indian Restaurant	Gym / Fitness Center	Pizza Place	Salon / Barbershop	Grocery Store
3	Santa Clara	Coffee Shop	Mexican Restaurant	Pizza Place	Sandwich Place	Convenience Store	Ice Cream Shop	Park	Bar	American Restaurant	Bakery

Above tabular description gives us an idea about the top nearby venues around Berkley, San Francisco, San Jose and Santa Clara. This is done using Forequare API and the latitudes and longitudes were given as the primary variables.

Map of California with Clustered Neighbourhoods



These are the clusters formed based on the top venues analysis and there are pop up labels attached that signify which area belongs to which neighbourhood.

Section 5: Recommendations:

On the basis of data analysis, I can say that the neighbourhoods such as Santa Clara, Beverly hills are more costly as compared to other ones. So, anyone planning to live there must ensure they have significant funds.

Additionally, we saw that Berkley has a mixture of different varieties of venues nearby which suggests that it is a good place to live as you can explore in any of your interest zones as it has a plethora of varieties in nearby venues.

Section 6: Conclusion

In this study, I planned to explore the popular cities of California and found the median prices of houses in the various neighbourhoods. I studied the various factors that are needed to predict the median house value in various areas such as total rooms, population, nearby attractions and many more.

After analysing the relationship between various features and the median house values, I applied machine learning techniques such as linear and multiple regression to find out the coefficients, intercepts and residual orders in order to find the accuracy of the model.

Further, I used Foresquare API to explore nearby venues to these neighbourhoods and the types of venues if they are Diner, coffe shop, bar or any other type. Additionally, I plotted those neighbourhoods on the California map using Folium library and later clustered the neighbourhoods to find out the attractions and more details related to the same.