Paul Tran
Applied Data Science Capstone
July 6, 2020

Place To Rent In Los Angeles.

I. Introduction

This is the report for final course of IBM Data Science Professional Certificate, 9 series course create by IBM. The final class have been equipped with the skills and the tools to use location data to explore a geographical location, over the course of weeks, students will have the opportunity to be as creative as they want and come up with an idea to leverage the Foursquare location data to explore or compare neighborhoods or cities of their choice or to come up with a problem that you can use the Foursquare location data to solve.

Los Angeles, the largest city of California, and one of the most populous city in the U.S. With its perfect Mediterranean climates, it is home of approximate 4 million people (1). The city also has many well known Universities such as UCLA, USC, and CalTech. Each year, hundreds thousands of people come to this city for good. Many of them are students, fresh graduated, working class, contractors, or even tourists. Therefore the demands for finding a place to settle down is essential.

The idea of comes from a process of a person try to figure out a place to rent after moving from an other city or even from a different country. It is common that people will looking for places that have reasonable prices and located closed by some kinds of venues such as restaurants, coffee show, and markets, ect. Therefore, what types of surrounding venues will affects the rent prices positively and negatively?

For this project, the main goal will be exploring the neighborhood of Los Angeles County in order to find out the correlation between the Counties's surrounding venues and the rents prices of it's neighborhood. The project will target:

- Students who attend at any Universities in Los Angeles county.
- Tenants who try to find a new place to rent.
- Business looking for a place to rent

- Tourist
- Renting agents or landlords who want to optimize their renting advertisement
- And last but not lease. to this class's instructions and classmate who will grade this project.

II. Data Overview:

The data that being use in this project will available at "maps.latimes.com". Those datas provide the population density (measures the numbers of people per square mile) of each neighborhoods in Los Angeles. This dataset also be used to get the latitude and longitude of each neighborhood by using *geocoder.arcgis* library. Moreover, it also being used to visualize LA population's distribution, therefore audiences will have a broad idea of each locations.

In a addition, there were a data set from "https://usc.data.socrata.com/" where it contains the median value of gross rent prices in an area, measured in dollars of from the year of 2010 to 2016. This median rent price measures the gross rent of that area, meaning that, it will include all the utility such as electricity, water, gas and sewage.

The geojson file of Los Angeles communities uses to indicate neighborhood boundaries on the map taken from"https://boundaries.latimes.com/" . Surrounding venues for each city in LA county was searched using FourSquare API base on the latitude and longitude of each city gathered using *geocoder.arcgis* library from the "maps.latimes.com" dataset that being mentioned above. This data later will be used to compare, analyze and cluster.

III. Methodology:

3.1.Data cleaning and and explore analysis

- Using "*pd.read_html*" to scrap the data table from LATimes for a list of cities in Los Angeles County and theirs corresponds population density per sqmi. Therefore, finding the geographic data (latitude longitude coordinate) by using Geocoder library.
- From download USC data table, calculated the avenge renting price from 2010 to 2016 of each city in LA County then created a new data frame that includes name of each city and corresponds average rent.
- After created 2 data frame, we noted that there were some differences between those two data frame, by looking at its shape. Looking closely at the datas, I decided to drop those differences by inner merging those two, use intersection of key from both frames.

3.2. FourSquare API and retrieval of surrounding venues within cities

• In order to looking at the disparities in the rent prices. FourSquare API was being used to retrieval venues that surrounds each area in the radius of 1 kilometers. For each neighborhood, the "explore" endpoint will be used to returned all the venues within 1km radius.

• Count the occurrence of each venue's categories in each neighborhood. Then apply one hot encoding to turn each categories into a column with their occurrence as the values.

• The table which each row represents a city in LA County and each column stands for each venues category features. The table be mainly used for the clustering and further analyzes

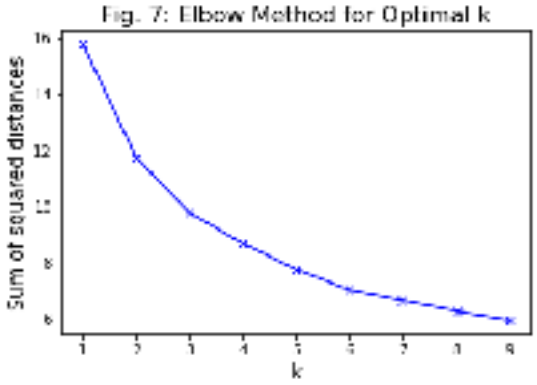
	Neighborhood	Amount	Population per Sqmi	Latitude	Longitude
0	Acton	1647.50	166	34.468150	-118.195130
1	Adams-Normandie	920.74	21815	34.078090	-118.301200
2	Agoura Hills	2052.50	2405	34.146110	-118.778120
3	Agua Dulce	1130.90	99	34.495700	-118.326210
4	Alhambra	1205.55	11275	34.093700	-118.127270
5	Alondra Park	1667.11	7515	33.889300	-118.330610
6	Altadena	1458.32	4000	34.186560	-118.131520
7	Angeles Crest	1194.40	3	34.311589	-118.011102
8	Arcadia	1375.32	4719	34.139550	-118.038670

3.3. Pearson Correlation

• We made an assumption that the rent price is depended on the population density, and all the surrounding venues, therefore the person correlation will be used to measure the strength of the correlation between feature. Our goal is to see which feature, or surrounding venues will affect the renting price of. Also regression plot was used to visualize the relationship between each of feature to the rent average renting price. The number of resources in each category was also used to draw scatter plots to illustrate the relationship between them. Correlation between number of surrounding venues and life expectancy was explored using Pearson's correlation coefficient and testing its significance.

3.4. Clustering and comparison of clusters

• The elbow method with the sum of squared distances was used to find the optimal number of clusters. Using the normalized number of resources in each category and the number of clusters, the communities were clustered into different groups using k-means clustering. Different clusters were characterized according to the distribution surrounding venues features.

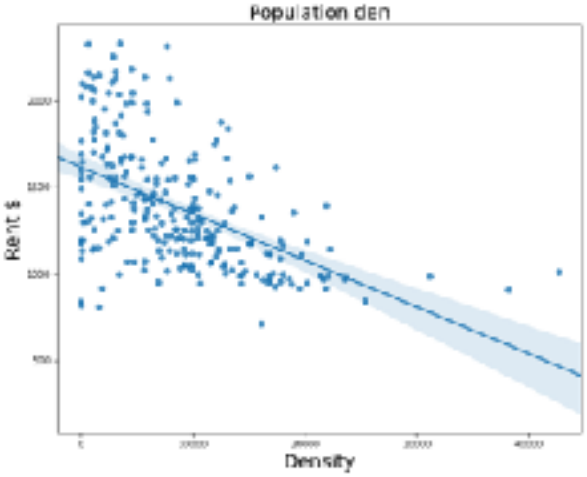


3.5 Map visualization

• Folium library was used to create choropleth map to reflect renting price and population density. A geojson layer of community boundaries was used to indicate community areas. Clustering of communities was visualized by using different colored markers to mark communities belonging to different clusters.

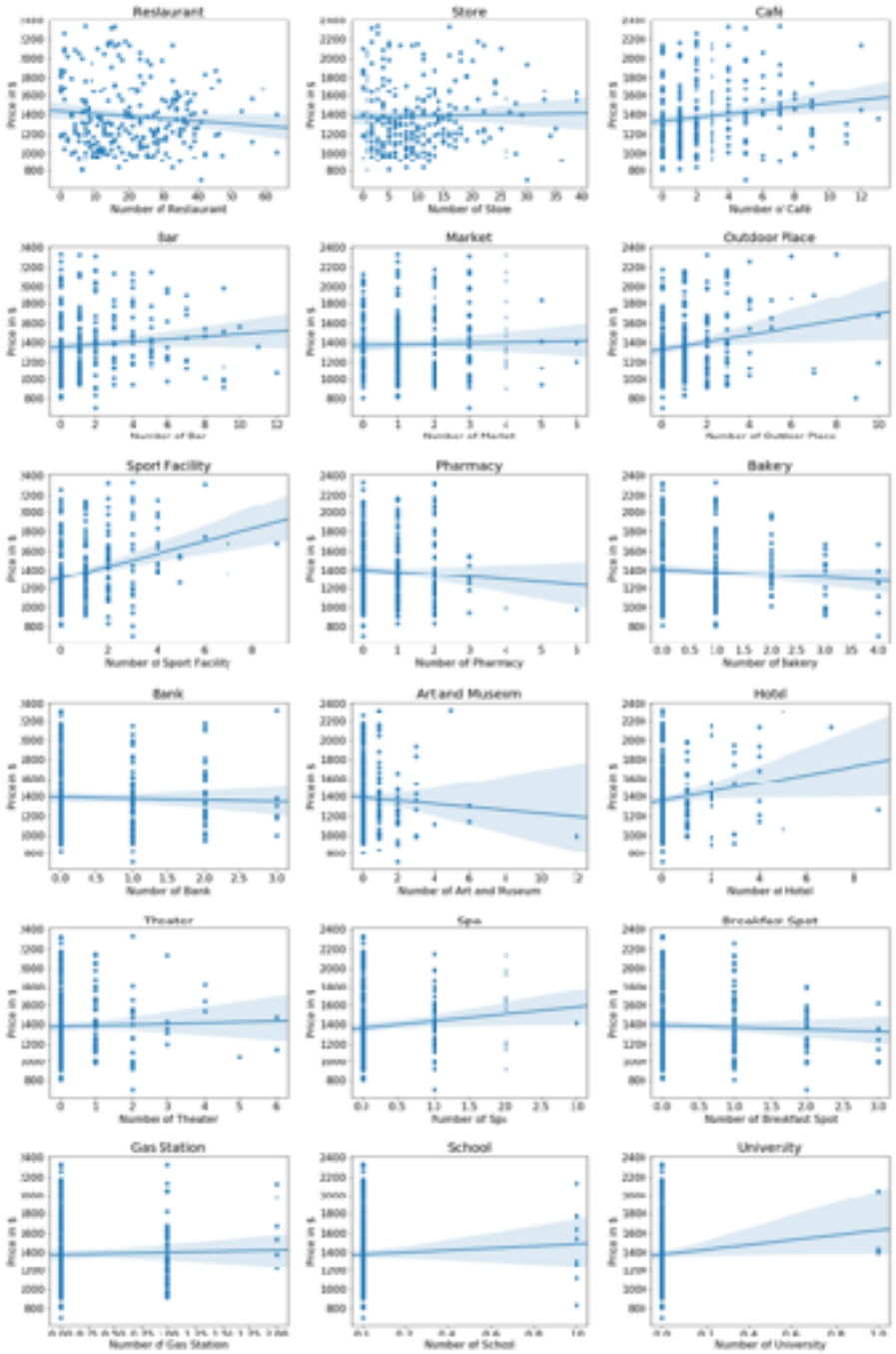
IV. Result

• The assumption that the rent price was depended on the population density, and all the surrounding venues were made. As the result, there were a negative relationship between population density and renting price. Meaning that areas with high population density has less expensive rent, and the area with less people living are having higher rent.

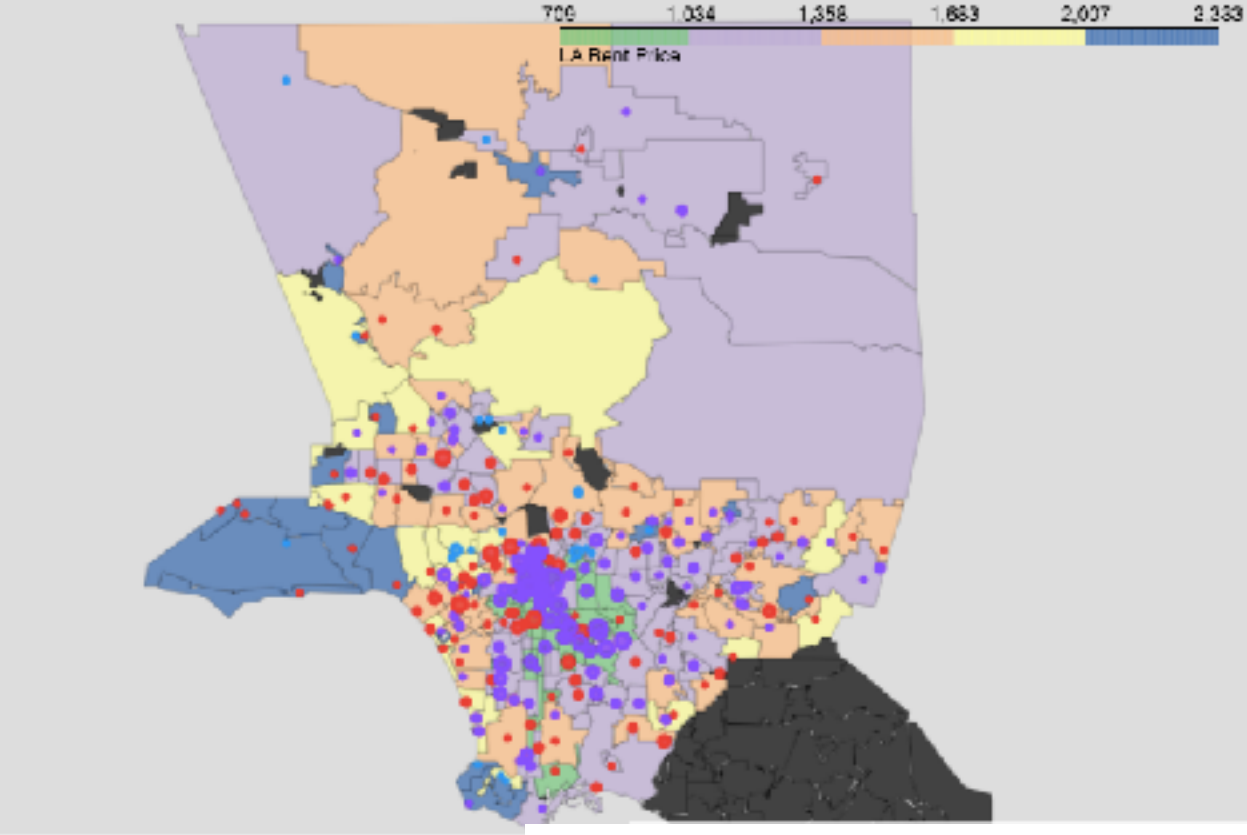


• However, there were no strong correlation coefficient between the selected venues and the renting price. Selected

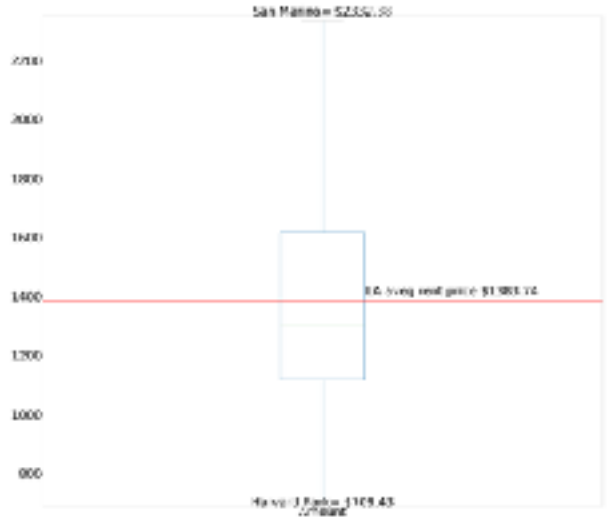
venues such as Restaurant, Store, Bar, Market, or Cafeshop, ects.. demonstrate a week relationship with renting prices since their Pearson coefficients values are low, nearly to 0. However, there are two features that have quite positive correlation on the rent. One is outdoor places such as Park, river, and gardens. The other is sport facilities. This result quite surprised me. Therefore it hard to say we could using those values to predict the prices.



• However on a bright side, by doing the area clustering, and visualize the L.A County, it was giving us a better idea of the disparities average rent prices in each neighborhood with similar venues types and category. Therefore it would help tenants easier to decide which areas to rent within their budget.



V. Discussion



• The disparities in the Los Angeles' average renting prices are high. There is the big gap between San Marino, $2332 and Harvard Park, $710 despite of their only 30 minutes drive distance. In this cases, if we, as a tenant, or tourist looking for

place to rent, we can use the result of clustering or using the correlation between renting prices and population density. By the result of clustering process, Harvard Park in the group 1 (cluster 0) where most of the businesses are there. San Marino on the other hand, is belonged to group 3 (cluster 2), a neighborhood area where there is many outdoor places such as gardens, park and sport facilities. Moreover, if we looking at the population density, of these two neighborhood, we will see there is also a price's correlation. Harvard Park, which nearly 16000 people per square miles compares to nearly 3000 people per square miles in San Marino.

- Base on the result of this analysis, there is no strong correlation between surrounding venues and renting price. There are some reasons that we can be address to explain the result:
    - The renting price and the real estate price in general ares hard to predict. Its prediction is still an difficult problem that many experts are trying to solve. There are many objective and subtractive factors that can affect the prices and also they are changing times to times.
    - The data that we use may incomplete, missing deciding factors and small sample.
        - Conducting data set is a real challenge. Since I combined datas from multiple sources, inconsistence can happen.
        - Still there is more efforts need to be done, such as researching, checking, and modifying before merging multiple dataset.
    - The venues that being discovered by the FourSquare API might be underrepresented or overrepresented since explore venues were certain distance away from a single coordinate representing a community. Therefore the area surrounding the coordinate may not representative of the whole community area.
        - Multiple trials and errors were made to get the final result from FourSquare API calls. Different parameters returned different result.
        - There are still duplicate venues that being listed even though FourSquare users have already reported. Despite of many more venues are returned at the area with large population density, it is possible that the number of resources in the communities with many resources were over-represented. There are still some

missing venues on Foursquare because of several reasons such as being newly opened or having some changes.

- One of the suggestion for the further analysis that might increase our result might be more specific on each area. Meaning that, we can break the LA County into different groups within their average rent price. Therefor, the analysis will focus on each neighborhood individually. Thus, the coefficient correlation between surrounding venues and renting price within that neighborhoods might be more valid.

VI. Conclusion

- Unfortunately, we could not find the best coefficient correlation for any venues types so that we can certain use it as an reference while looking for places to rent. Beside than the population that we found related to the price ,The rent might also depend on many other factors such as how newly of the place, hosts, roommates, types of house and how safe the neighborhoods are.