Final Project - Report

DSCI 510

Principles of Programming for Data Science

Enhao Wang 8604-0104-78

Contents

1 Abstract **3**

2 Motivation **3**

3 Process of Analysis **3**
- 3.1 Data Combination & Cleaning . . . 3
  - 3.1.1 Removal of Outliers . . . 4
- 3.2 Overview of Data . . . 6
  - 3.2.1 Correlation Matrix & P-value . . . 6
  - 3.2.2 Draw Plots & Some Analysis . . . 6
  - 3.2.3 Regression on Longitude and Total . . . 11
- 3.3 Clustering on the Data . . . 12
  - 3.3.1 Clustering Based on Distance . . . 12
  - 3.3.2 Clustering Based on Price . . . 12
  - 3.3.3 Clustering Based on Total . . . 15
  - 3.3.4 Clustering Based on All Five Variables . . . 16
  - 3.3.5 Clustering Based on Distance and Total . . . 17
  - 3.3.6 Clustering Based on Latitude, Longitude and Total . . . 18
  - 3.3.7 Clustering Based on Price and Total . . . 19
- 3.4 Regression on the Data . . . 19
  - 3.4.1 Linear Model . . . 20
  - 3.4.2 Other Models . . . 20
  - 3.4.3 Result MSEs . . . 20
- 3.5 Classification on the Data . . . 21
- 3.6 Analysis of Crime Type . . . 21
  - 3.6.1 Distribution of Most Frenquent Crime Type . . . 23
  - 3.6.2 Pie Chart & Histogram . . . 24

4 Conclusion **25**

5 Limitations, Challanges and Future Works **26**
- 5.1 Limitations . . . 26
- 5.2 Challanges . . . 26
- 5.3 Future Works . . . 26

1 Abstract

This final project aims to find and obtain the information of houses around USC(including price, location, distance from USC, number of crimes nearby within previous 28 days .etc), and an analysis of the data, which mainly focuses on the relationship between some factors(price, longitude, latitude, distance) as input variables and crime as response variable. In Homework 3, the data sources are found; in Homework 4, the 5 data sets are downloaded from the data sources, in which 4 are used for the final analysis.

This report will show my work of final analysis, including data pre-processing and cleaning, steps of the process of analysis, and some graphs to visualize the result. In conclusion, there is some regularity on the crime nearby houses around USC, and the crimes are predictable by a machine learning model.

2 Motivation

The motivation here is described based on the target of analysis, so it's different from the motivation, actually say interest, of what I mentioned in Homework 3.

For all USC students, where to live when they are having their college lives is an important issue before they come to Los Angeles. Among all the factors to consider of a house, the security is an important one, because no one wants to have trouble getting hurt or having some personal belongings stolen. Sadly, the neighborhood of USC is not very safe, although USC has a safety zone called DPS. So I want to research on the crimes nearby each house around USC as the topic of my final analysis, to see what we can learn from the data retrieved. Some of my research is based on the hypothesis that crime rate nearby a house is related to the location, distance from USC, and price of the house.

3 Process of Analysis

3.1 Data Combination & Cleaning

For all of my four data sets, the first two columns are names and locations of the houses, so actually the first two columns of all four data sets are the same, and it's easy to combine them together. For cleaning, the first task is to delete duplicate, because the houses have duplicates(same location but different units). The second is to remove outliers, and the last is to remove crime types that has no recored from all houses at all, to shrink the size of data set and save memory. Here only include removal of outliers, because other processes can be

shown by code.

3.1.1 Removal of Outliers

This task mainly focuses on two variables: distance and price. Latitude and longitude is not included because they are just a kind of reflection of distance. The box plots for distance and price is shown as Figure 1:

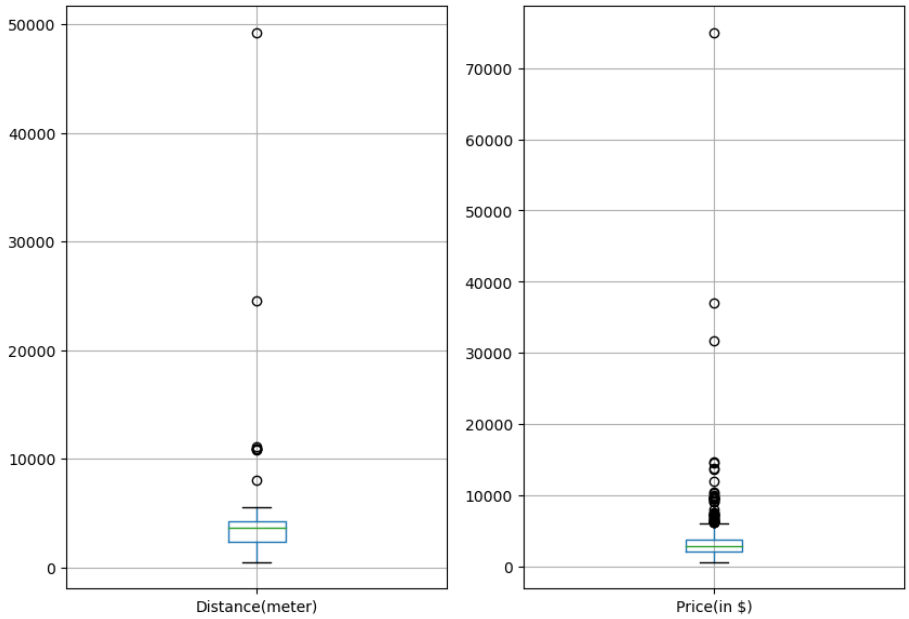

Figure 1: Box Plots for Distance and Price

We can find the outliers from the box plots. For distance, the outliers are which distances $>$ 6000, so we use 6000 as a threshold exclude outliers. For price, the outliers are which prices $>$ 7000, but there are too much data points below 20000, and the truely 'outsider' is that three points above 30000, so just use 20000 as the threshold.

I also plot the distributions of distance and price, as Figure 2:

The information shown coincide with the box plots.

Finally, the whole data set after combination and cleaning can be opened **here**.

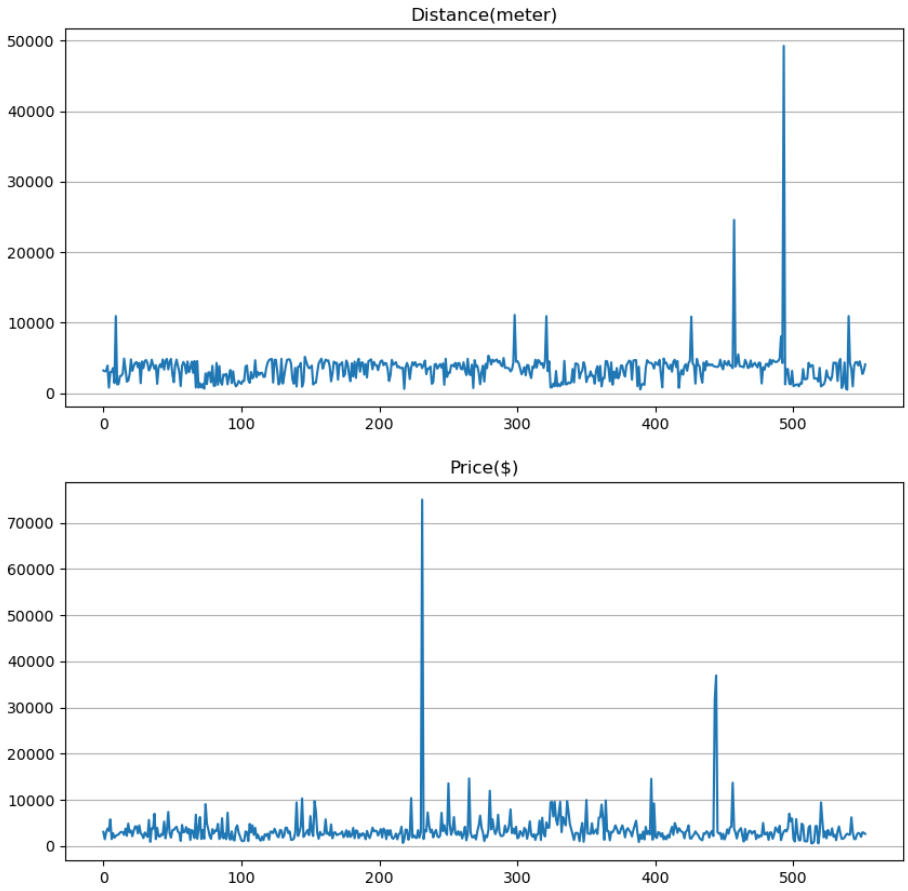

Figure 2: Distributions of Distance and Price

3.2 Overview of Data

We mainly focuses on crime, so the total number of crimes(noted as total) for each house is the response variable. As mentioned, the four input variables are distance, latitude, longitude and price. We want to find some relationship between these input variables and total.

Before analysis, I'd like to have an overview on the data.

3.2.1 Correlation Matrix & P-value

Firly, let's compute a correlation matrix on the five variables(four input, one response). The matrix is shown as Table 1:

	Price(in $)	Latitude	Longitude	Distance(meter)	Total
Price(in $)	1.000000	0.011200	0.125872	-0.142827	0.173229
Latitude	0.011200	1.000000	0.249181	0.526270	0.318904
Longitude	0.125872	0.249181	1.000000	0.063739	0.729802
Distance(meter)	-0.142827	0.526270	0.063739	1.000000	0.216614
Total	0.173229	0.318904	0.729802	0.216614	1.000000

Table 1: Correlation Matrix on Input and Response Variables

From the table, it can be seen that from four input variables, the correlation efficient is only high on longitude to total, the other three is low. It seems that these three variables have little linear relationship with total. Then I calculated the p-value for them to see if these variables are statistically significant for predicting total, as shown in Table 2:

	Distance(meter)	Latitude	Longitude	Price(in $)
p-value	3.544767e-07	3.053240e-91	2.811393e-14	5.028280e-05

Table 2: P-values on Input Variables

If we use α =0.05 as the criteria, then all of the p-values are far below it, so we can conclude that all of the four input variables is statistically significant for predicting total, which means that they must have some relationship with the total.

3.2.2 Draw Plots & Some Analysis

Then, I want to draw some plots or maps to look on the data nicely.

(1) USC DPS Zone

I got the map data from USC website(https://dps.usc.edu/patrol/), and convert it to GeoJSON code and use python package 'folium' to draw the map. The map(.html file) can be opened **here**.

(2) Distribution of Houses

Let's look at the distribution of all the houses on the map. The map can be opened **here**. Clicking on the house icons can pop up the name of the house.

I also plotted the distribution of houses on a plot. The plot is shown in Figure 3:

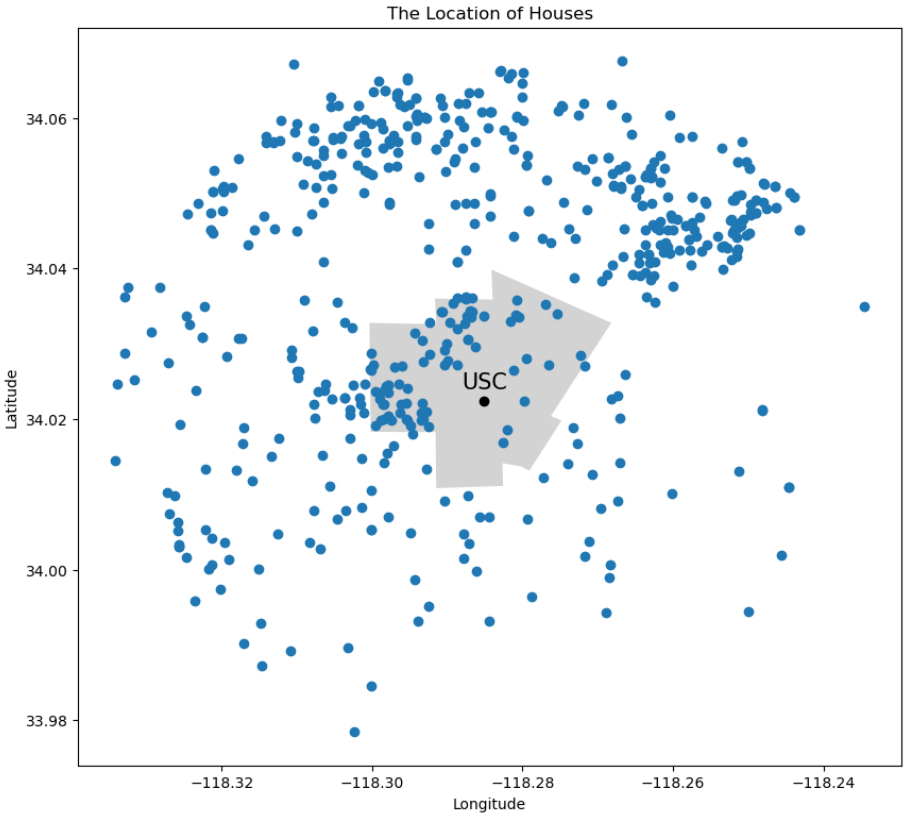

Figure 3: Distributions of Houses

We can see that the houses are distributed around USC, the distribution is like a circle.

(3) Houses based on Number of Crime

Similar plot, but here the color of each house is denoted by its total number of crime. I used 50 times of crime as a range for each color. The plot is shown in Figure 4:

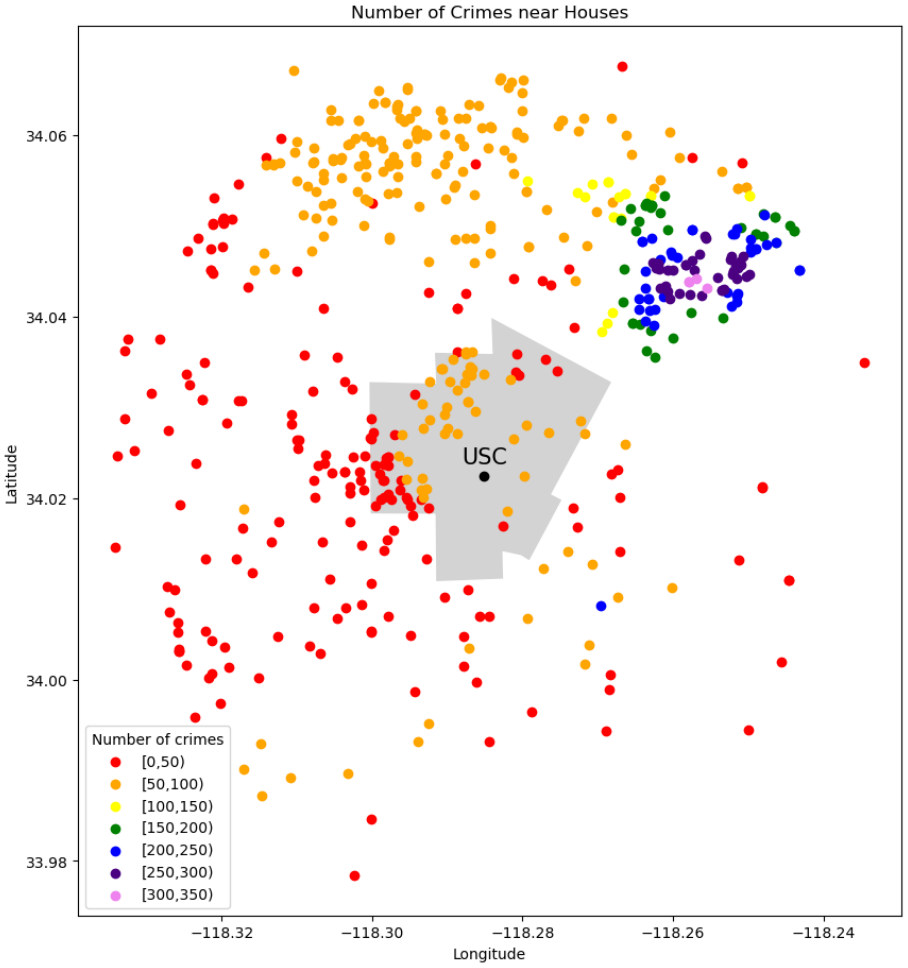

Figure 4: Distributions of Houses with Number of Crime

We can see that the number of crimes is district-based. The west side can be a group, the north size can be a group, and the north-east side can also be a group, in which the crime rate is surprisingly high. Interestingly, the distribution in northeast is like layers: the middle is highest, and decreasing when going outwards.

I also included similar maps. The map with pure colored house points can be opened **here**. On **this map**, clicking on the house icons can pop up the name and total number of crimes of the house.

(4) Split the Houses

Just try to do something. As mentioned, the whole area can be splitted into some districts by the number of crimes. So I set up four separating lines to split the whole area into 4 regions: west, north, northeast and southeast. The plot is shown by Figure 5:

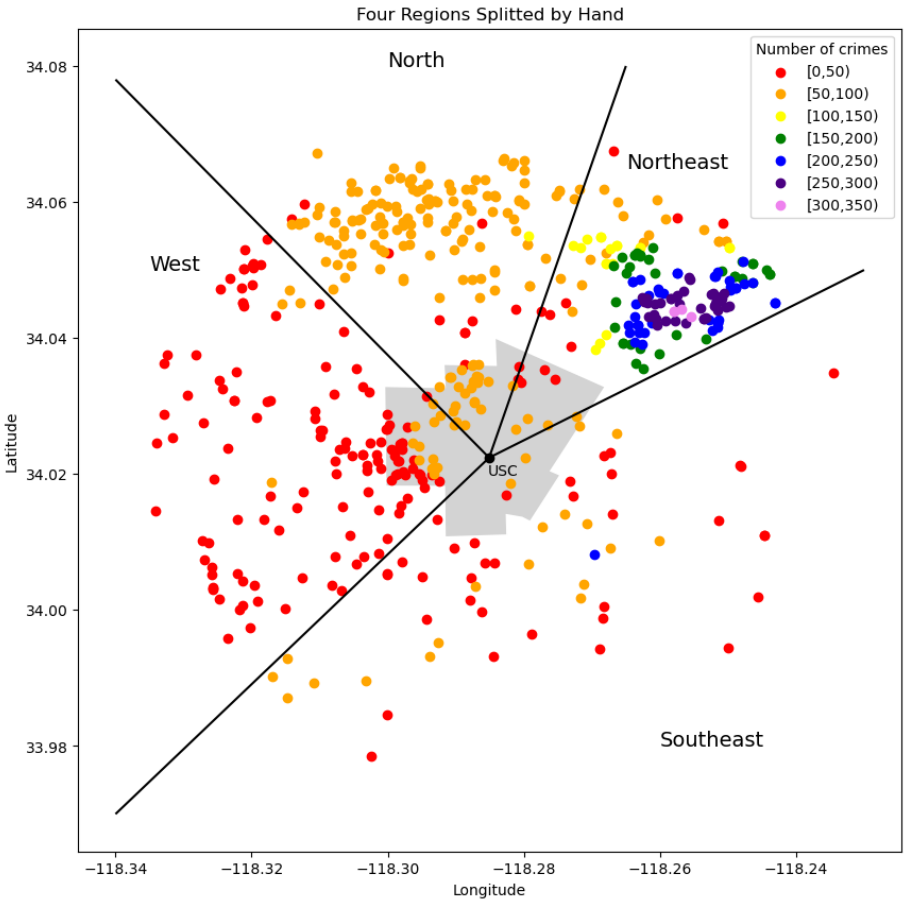

Figure 5: Split the Houses by Hand

I think the split looks appropriate.

(5) Detailed Houses based on Number of Crime

In (3), each color of house represents a 50 range of number of crimes. Let's look at it detailedly. This time, use 10 as a range, except crime number>100, because they mainly centers on the northeast. The plot is shown by Figure 6:

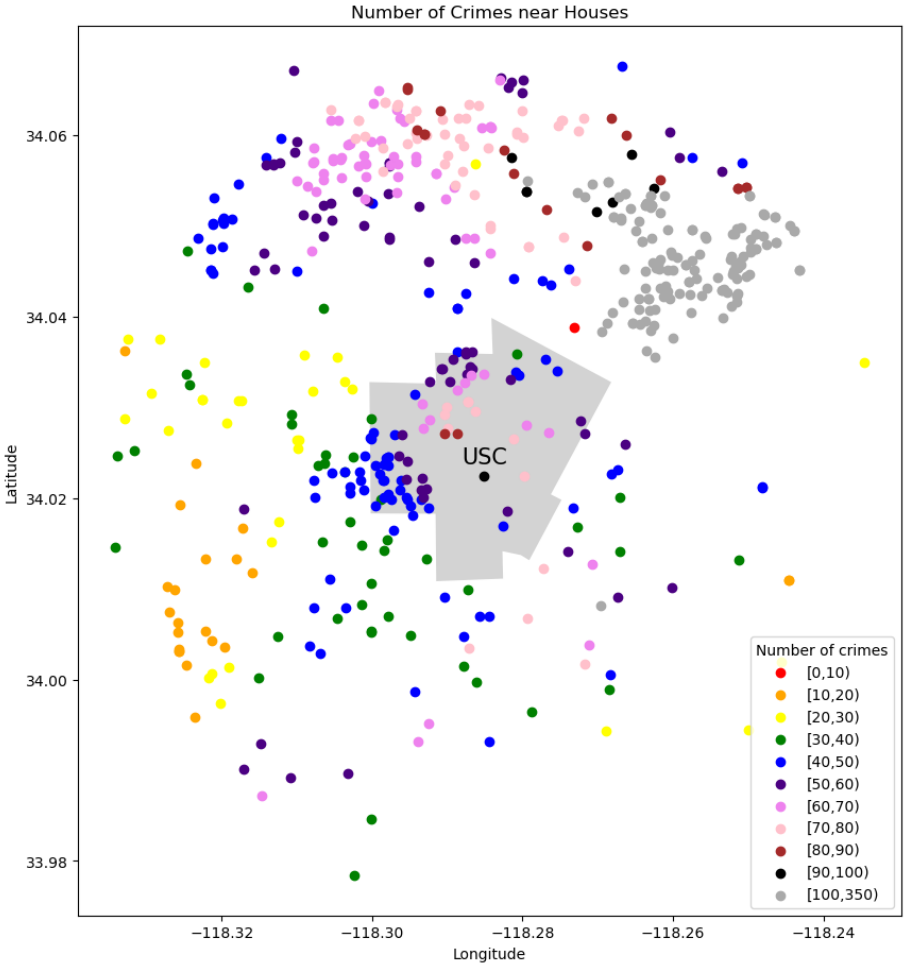

Figure 6: Distributions of Houses with Number of Crime

I'm living in the west of school. Interestingly, in the west part of USC, I found that there is a small tendency that the closer to school, the more crimes. This fact is contrary to what it supposed to be.

Does DPS zone work? From Figures 4 and 6, I think the DPS zone has some

effect on safty, because the number of crimes is higher outside the zone than inside.

3.2.3 Regression on Longitude and Total

As mentioned in **3.2.1**, the coorelation coefficient is a kind of high between longitude and total. So here I want to do a linear regression too see how related they are. The regression result is shown in Figure 7:

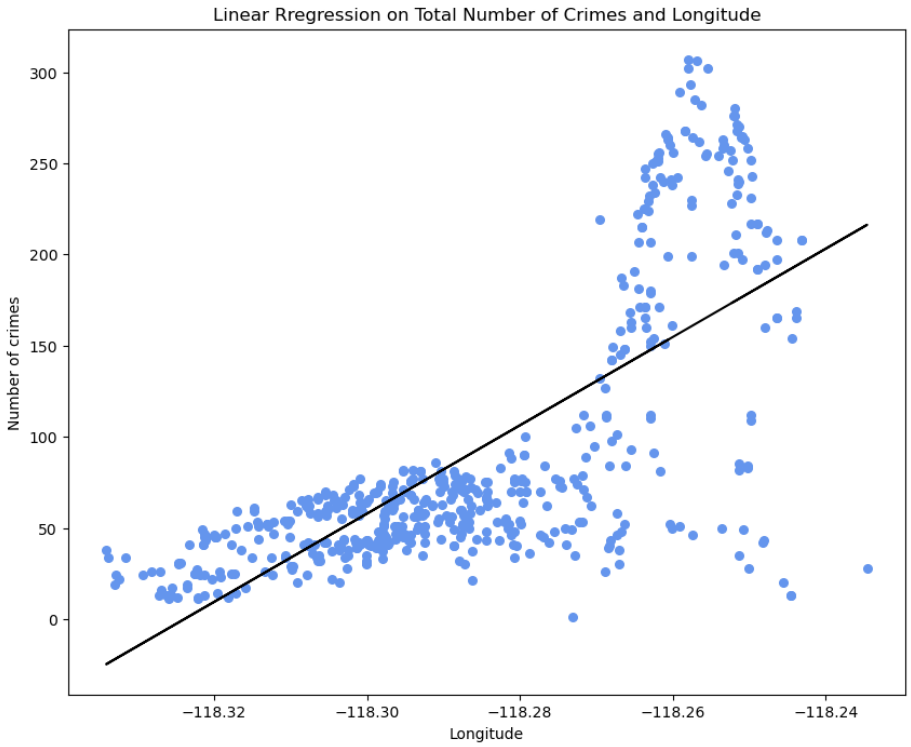

Figure 7: Linear Regression on Longitude and Total

The R^2 score calculated is 0.53, for which the best score is 1. From both the plot and score, we can see that the regression is not very good. So then I tried polynomial regression, which extends regression to higher dimension(like X^2, X^3) of longitude. The regression curves and R^2 scores are shown in Figure 8:

Still not very good. Maybe longitude alone is not a good predictor for total.

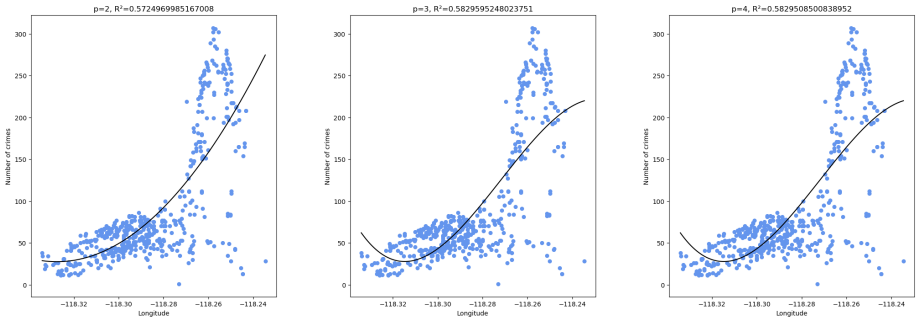

Figure 8: Polynomial Regression(with p=2, 3, 4 from left to right) on Longitude and Total

3.3 Clustering on the Data

As mentioned, the number of crimes for houses is district-based, and I tried to split the regions on my own. Here I'm interested in splitting the regions by machine. So I used k-means clustering method to find the regions(in this method, they're called clusters) that has similar attributes. I can find the clusters with similar total, price, distance .etc. For each situation, the algorithm will look for the best number of clusters first, and split the data into this number of clusters.

Firstly lets focus only on one variable each time.

3.3.1 Clustering Based on Distance

Find clusters in which houses inside each cluster are similar in distance. The plot is shown by Figure 9:

Looks very pretty. The pattern is obvious, which is circular distributed.

Clustering on latitude and longitude alone is not included, because it's not very meaningful.

3.3.2 Clustering Based on Price

Find clusters in which houses inside each cluster are similar in price. The plot is shown by Figure 10:

Maybe price is not good for visuallizing. The distribution of clusters is so confusing to us.

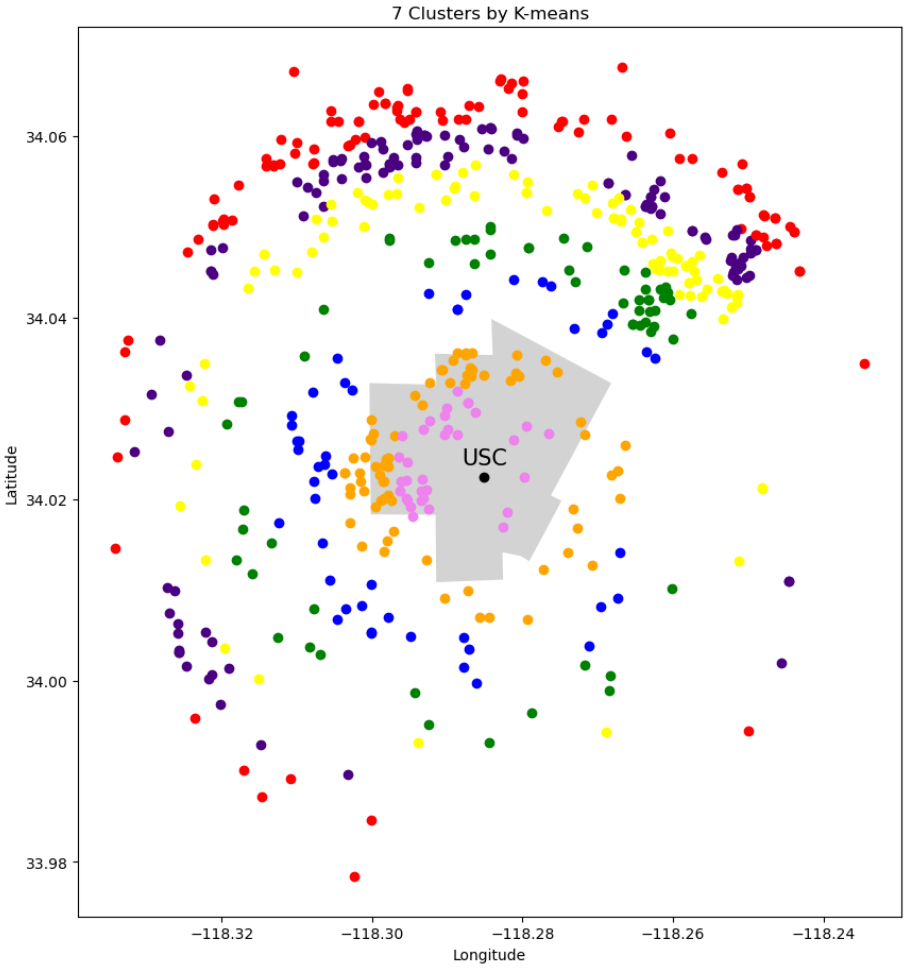

Figure 9: Clustering Based on Distance

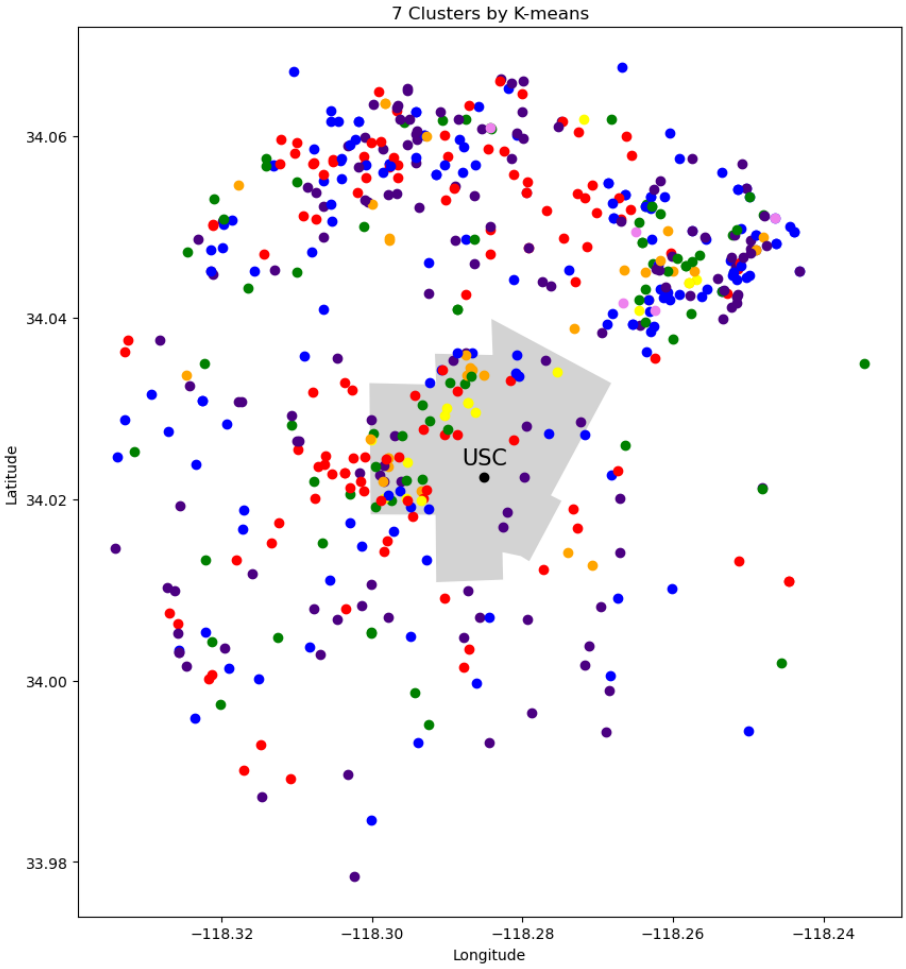

Figure 10: Clustering Based on Price

3.3.3 Clustering Based on Total

Find clusters in which houses inside each cluster are similar in total. The plot is shown by Figure 11:

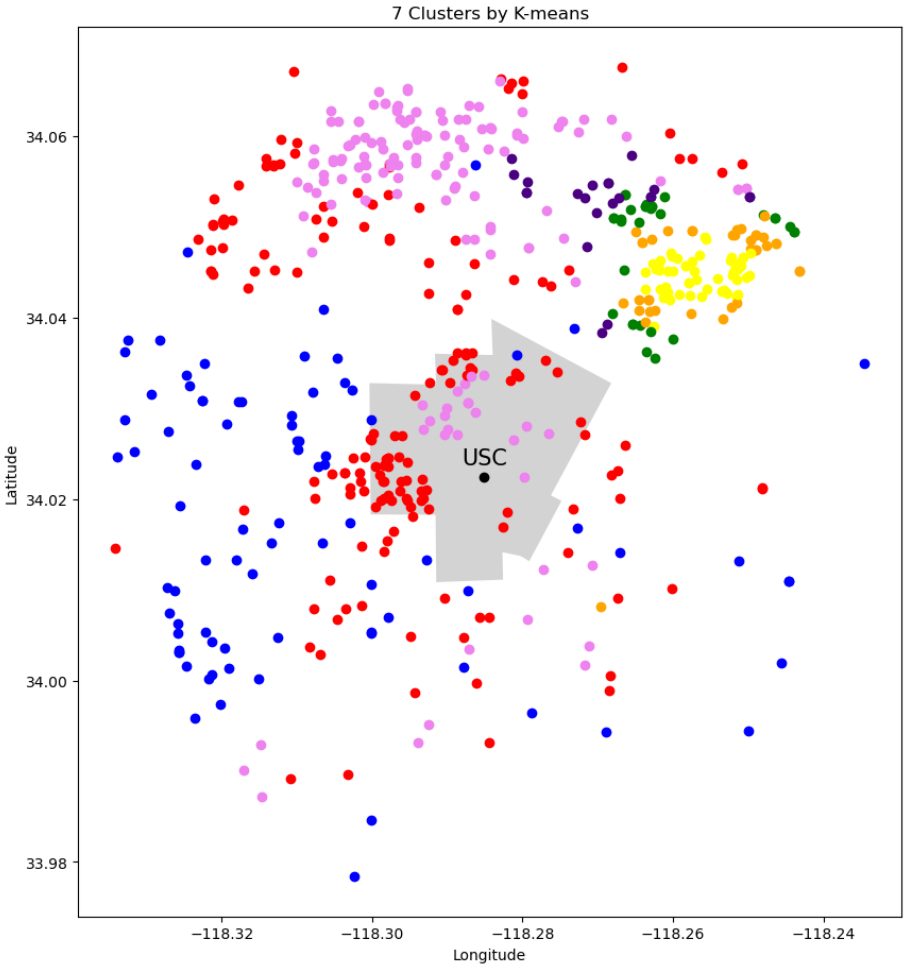

Figure 11: Clustering Based on Total

Very similar to Figure 4. This is because the k-means algorithm thinks it is best to use 7 clusters, and I also used 7 'clusters' for coloring the houses.

Next, let's do clustering on the input variables with total together.

3.3.4 Clustering Based on All Five Variables

Find clusters in which houses inside each cluster are similar in all of the five variables, distance, latitude, longitude, price and total. The plot is shown by Figure 12:

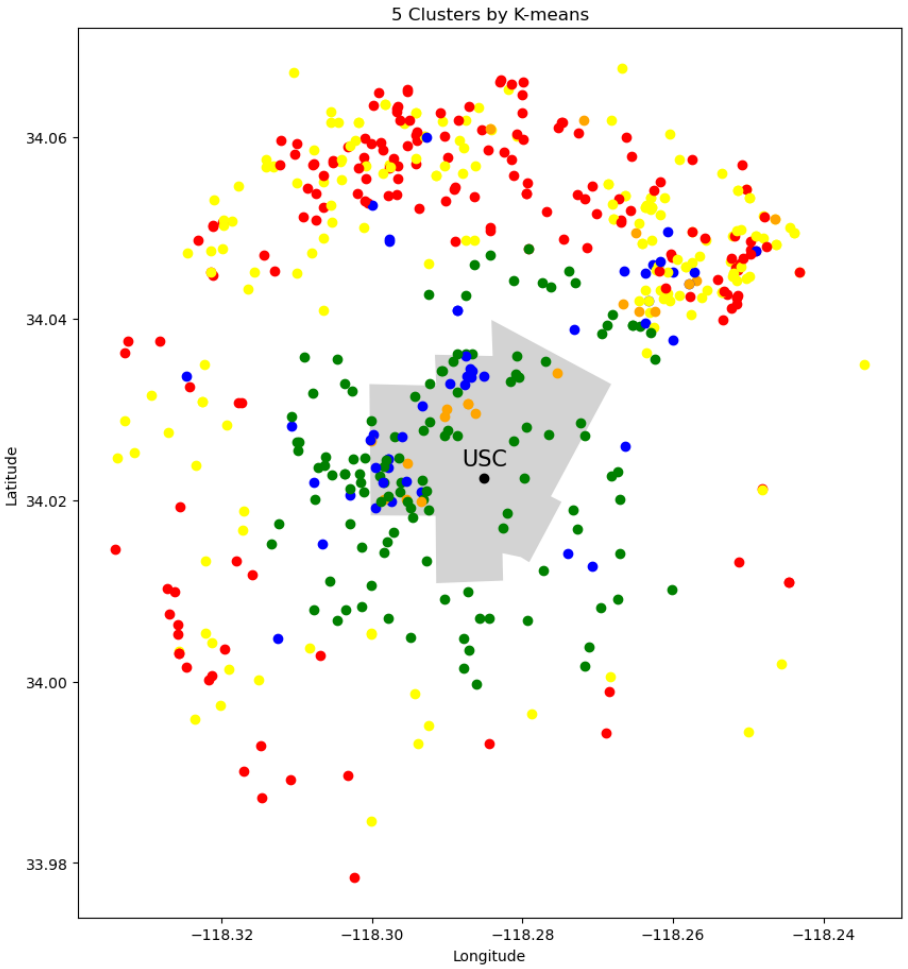

Figure 12: Clustering Based on All Five Variables

It seems not very meaningful. I didn't recognize any useful pattern on the plot. Maybe the clustering is based on too many features.

3.3.5 Clustering Based on Distance and Total

Find clusters in which houses inside each cluster are similar in distance and total. The plot is shown by Figure 13:

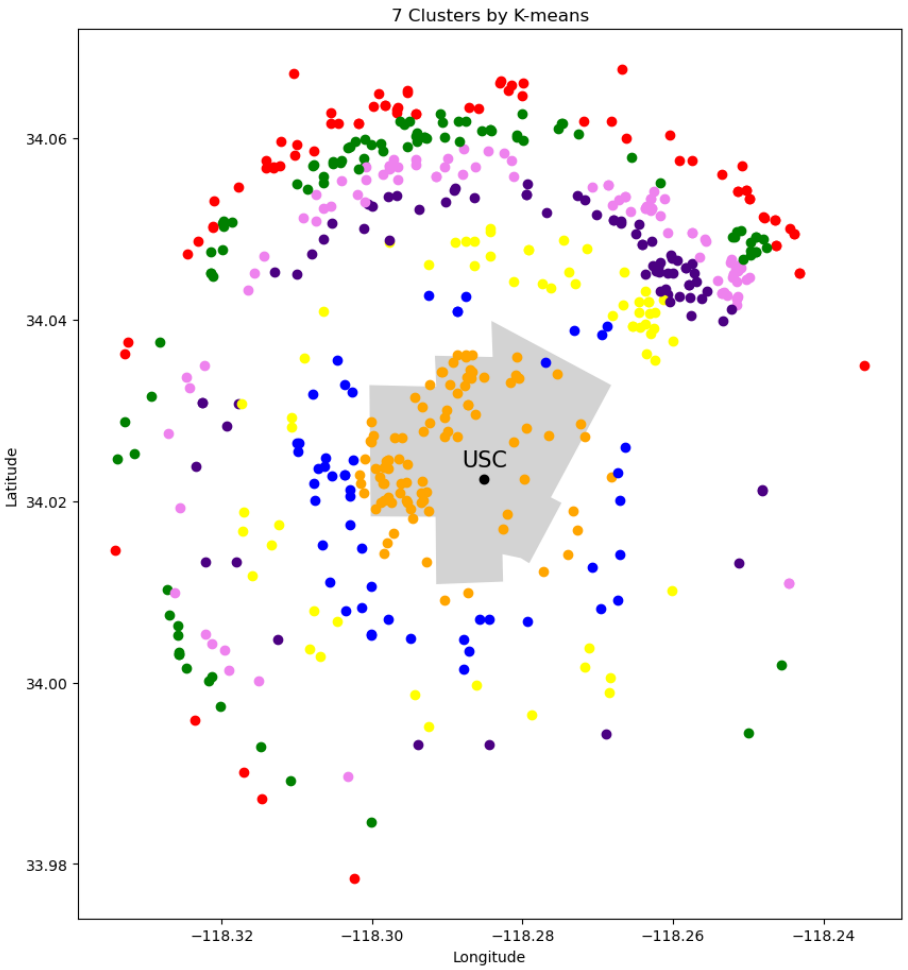

Figure 13: Clustering Based on Distance and Total

Seems the clusters are more based on distance than total, but this plot indicates some relationship between distance and total, because the houses with different distance have distinct total.

3.3.6 Clustering Based on Latitude, Longitude and Total

Find clusters in which houses inside each cluster are similar in latitude, longitude and total. The plot is shown by Figure 14:

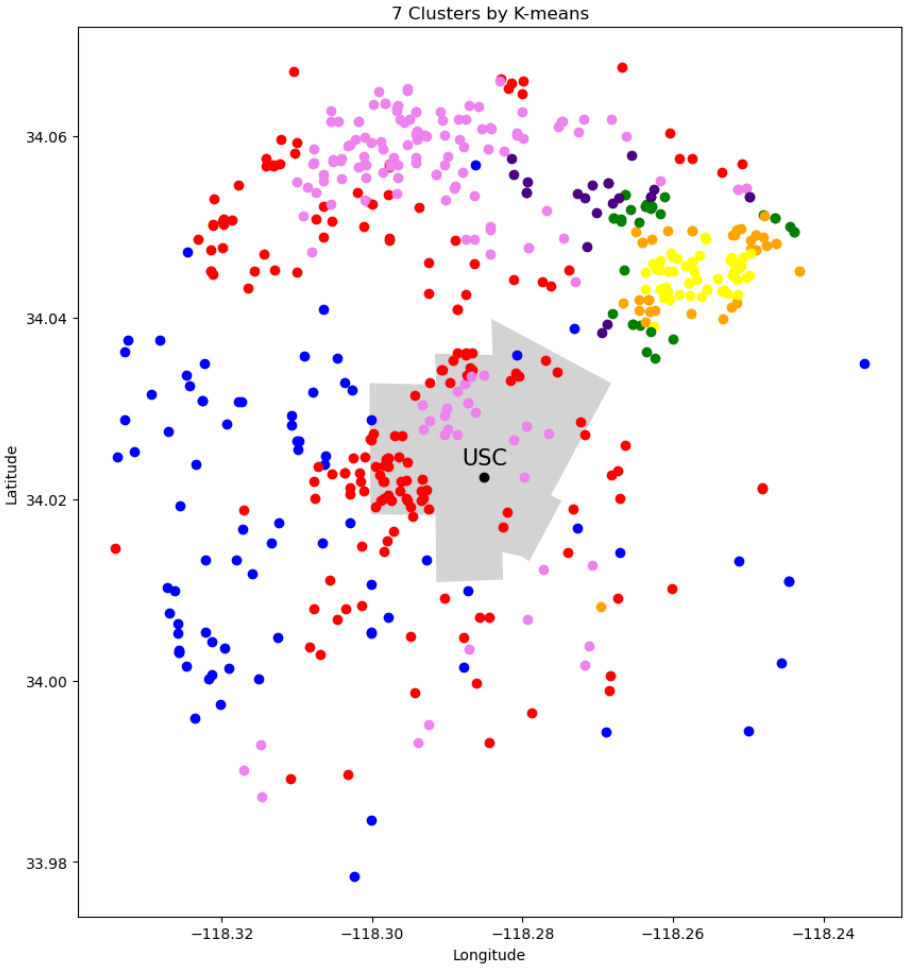

Figure 14: Clustering Based on Latitude, Longitude and Total

This clustering is more practical, because when talking about district-based, we want to find the houses that are close to each other. 'Similar in latitude and longitude' can meet this need. So maybe latitude and longitude are good predictors for total.

3.3.7 Clustering Based on Price and Total

Find clusters in which houses inside each cluster are similar in price and total. The plot is shown by Figure 15:

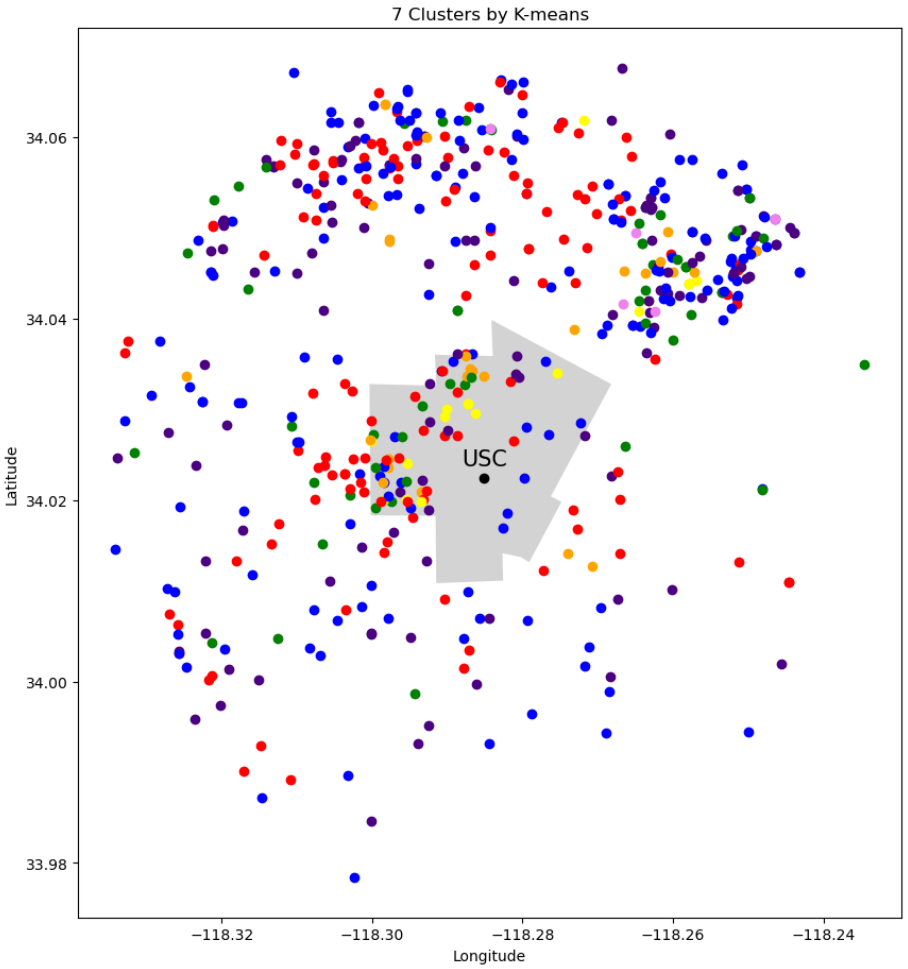

Figure 15: Clustering Based on Price and Total

Again, price is not good for visualization!

3.4 Regression on the Data

After lots of visualization and analyse of data, and a single 'failed' regression on longitude and total, let's do some realistic tasks. This part focuses on

regression by some machine learning models. For all models, randomly split the house data into training set and test set, train the model on training set and calculate the mean squared error on test set. As shown by p-value, all of the four input variables are statistically significant in predicting total, so here I just used all of the four variables together to train and test the models.

3.4.1 Linear Model

Linear Model includes four parts:

The first one is linear regression(LR) on four variables together.

The second one is polynomial regression(PR) on four variables, with quadratic and cubic terms for each variable.

The third one is linear regression on four variables and their interaction terms, like 'Longitude * Price'.

The last one is an integrated polynomial regression on the features of the last two together, with feature selection, which is finding a subgroup of all the features that can best predict total.

Detailed implementation is shonw by code.

3.4.2 Other Models

Other models include decision tree(DT), support vector machine(SVM) and multi-layer perceptron(MLP).

3.4.3 Result MSEs

Table 3 shows the result mean squared errors for the models above:

	LR	PR	LR with interaction	Integrated LR	DT	SVM	MLP
MSE	1932.09	1741.11	1561.82	509.04	169.81	5408.49	4989.15

Table 3: MSE of Models

From the table, we can see that the higher-dimensional and interaction terms improves the performance of linear regerssion. It's normal that inegrated LR is the best among these linear models, because it is the most powerful model in these four. Decision tree is the best among all these models, maybe because it's good for small amount of data. It's not normal that the SVM and MLP has so huge error, although they're powerful models. Probably this is also becuase the

data set is too small for them to perform well.

3.5 Classification on the Data

It's a pity that SVM and MLP perform poorly on the data. So now instead of regression, I also want to try some classification task on the data, to see if they can perform better. Like the coloring method of Figure 4, use 50 as a range to label the data, so the numerical feature can be converted to categorical feature. The model used for classification here are decision tree(DT), support vector machine(SVM) and multi-layer perceptron(MLP). Figures 16, 17 and 18 shows the comparison of the classification result by three models and the true labels on the data.

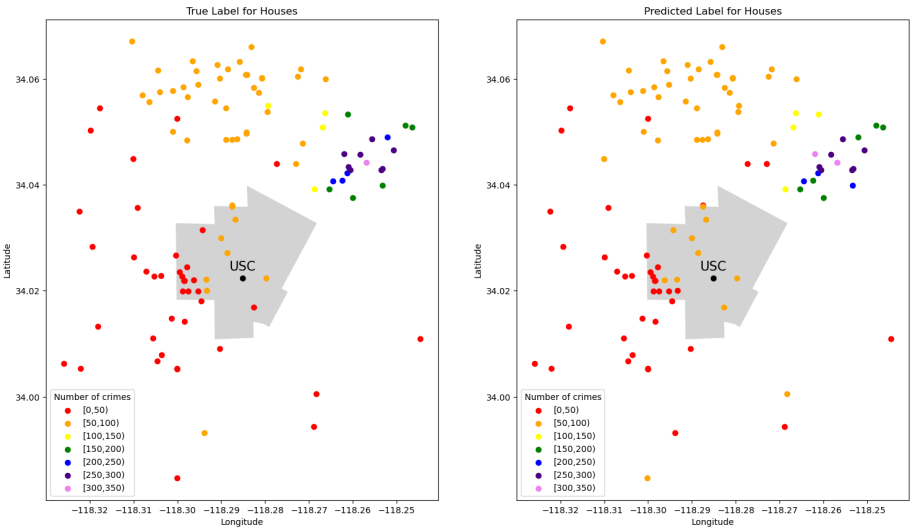

Figure 16: Prediction Result of Decision Tree

The accuracy rate of three models are 0.85, 0.54 and 0.55 respecitively. From the plots and accuracy rates, we can also see that the decision tree is prominent, SVM and MLP is still not very good. The size of data set still affects these two models.

3.6 Analysis of Crime Type

Lastly, I want to find something on the crime type. The https://www.crimemapping.com/ records lots kinds of crime types, but as mentioned in data cleaning process, some the the types has no recoreds in my data set. This indicates that probably

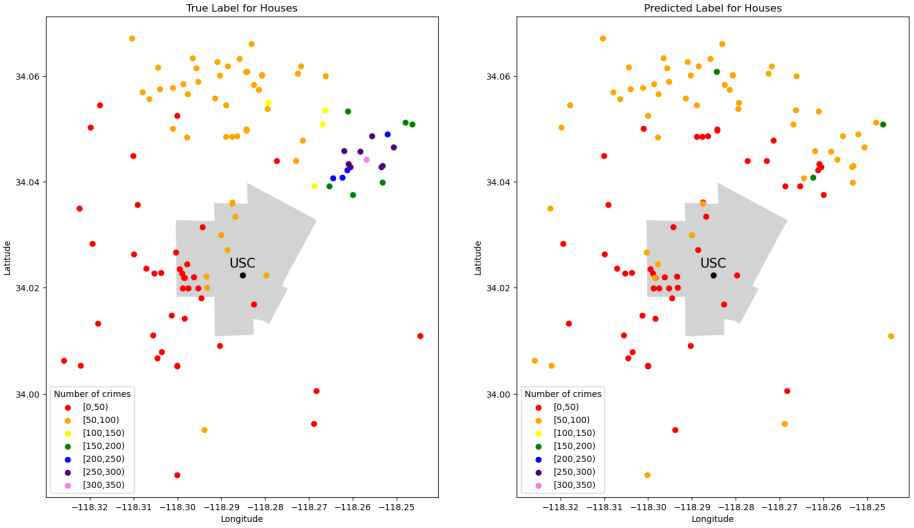

Figure 17: Prediction Result of Support Vector Machine

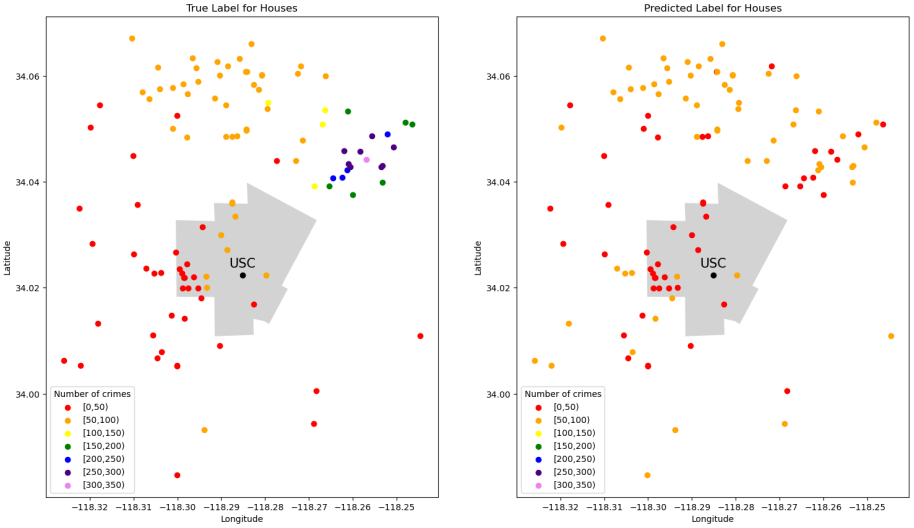

Figure 18: Prediction Result of Multi-layer Perceptron

some of the crime types will not happen around USC. Let's find something on the other crime types.

3.6.1 Distribution of Most Frenquent Crime Type

Firstly, let's draw a similar plot on the houses, in which color represents the crime type that happens the most nearby a house. The plot is shown by Figure 19:

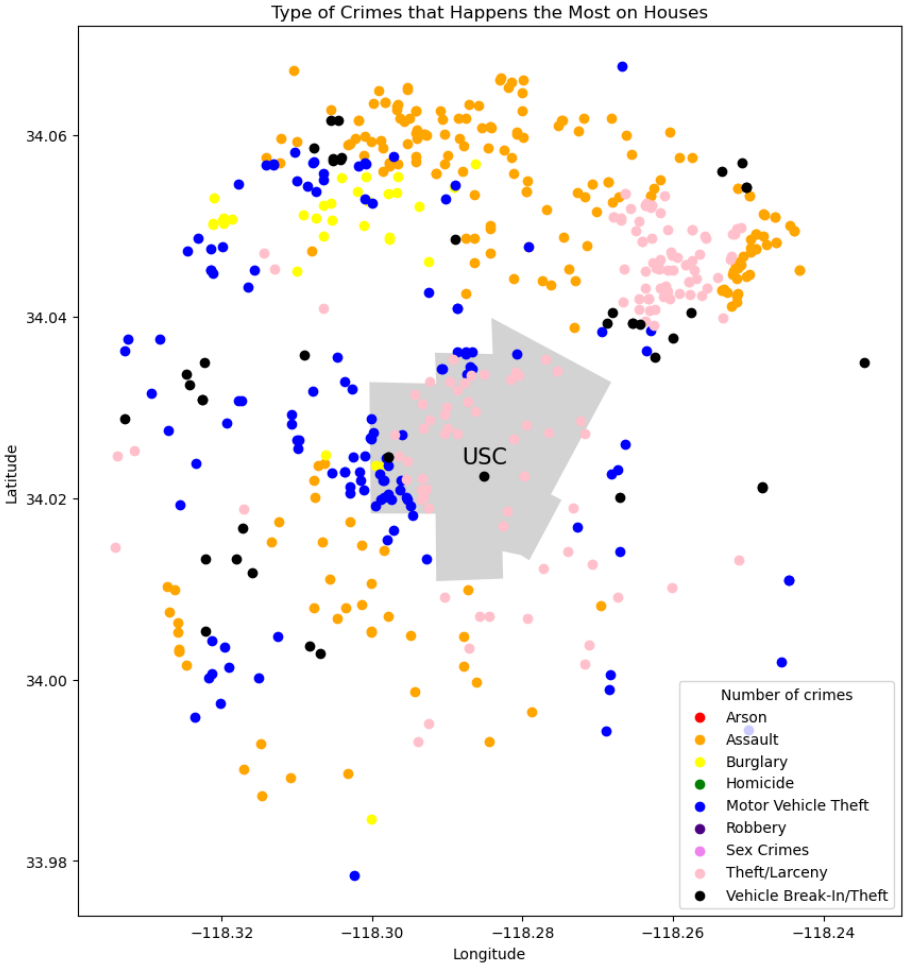

Figure 19: Clustering Based on Price and Total

This plot is not enough informative, because there are multiple types happens nearby each house, but the plot can only show the most frequent one due

to the capability limitation. However, we can still have some finding on the plot.

This plot can somehow show which type is the most likely to happen in a region. Interestingly, as shown in the legend, the data set records 9 types of crime, but there are only 5 in the plot, which are assault, burglary, motor vehicle theft, theft/larceny and vehicle break-in/theft. Four of these five are about theft! Moreover, in the north and south part of USC, assault is also a frequent type.

3.6.2 Pie Chart & Histogram

Then let's count the types together and make a comparison. Figure 20 and 21 show the pie chart and histogram of all the crime types, counting then in total number.

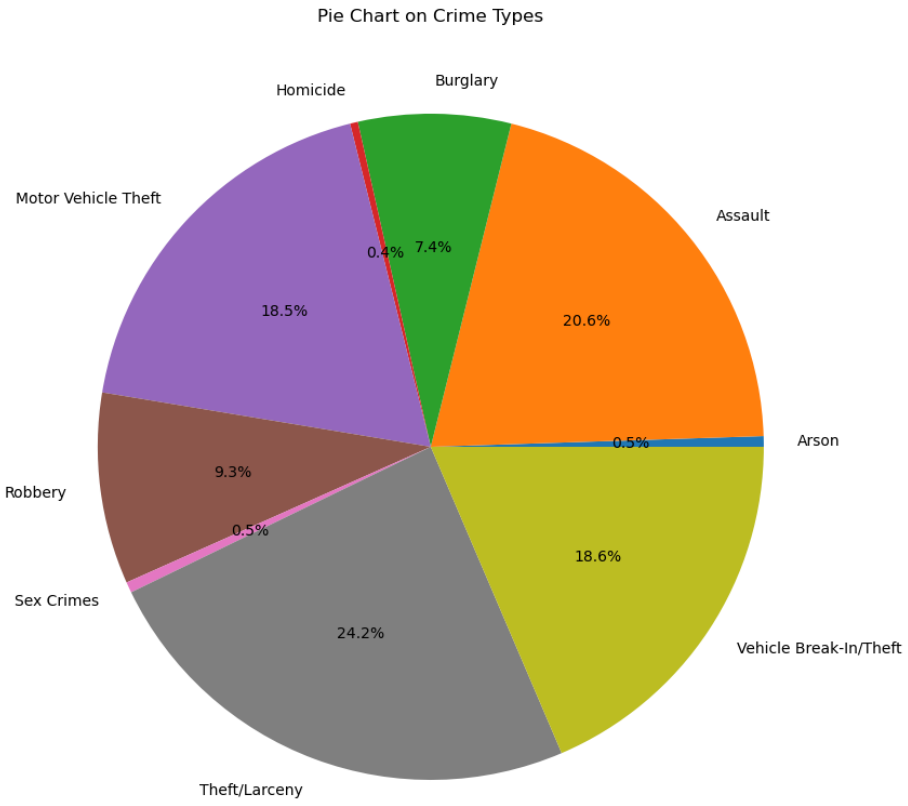

Figure 20: Pie Chart for Crime Types

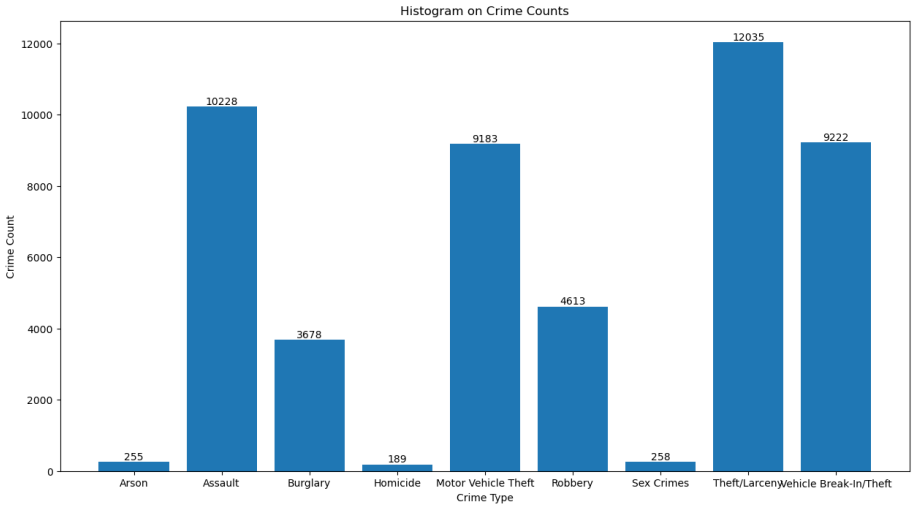

Figure 21: Histogram for Crime Types

From these two plots, we can see that the most frequent four types are assault, motor vehicle theft, theft/larceny and vehicle break-in/theft. Burglary and robbery also have considerable numbers. Fortunately, arson, homicide and sex crimes are few, which will hurt our body or even threaten our life.

In conclusion, the most frequent crime types are aimed at personal belongings, so just take care of our belongings; also, be careful about some assaults or attarks in some region.

4 Conclusion

In my final project, firstly I cleaned and combined the data retrieved from Homework 4. I used some statistics and drew some plots to get an overview and understanding of the data. Next I performed k-means clustering to see if the houses can be splitted into different regions according to some standard. Then I performed machine learning models on the data to see if the total number of crime can be predicted, and succeeded. Lastly, I did some analysis on the crime type.

In conclusion of my findings, the characteristic of crimes around USC can be district-based, the crime situation are different in different regions. Thanks to USC, the DPS zone is useful for our safety, but there are still quite a few number of crimes happens inside it. A machine learning model(in this work, decision tree is the best) can be used to predict the total number of crime nearby

a house given its distance from USC, latitude, longitude and house price, and it can yield great results. It'a good thing that the most frenquent crime type that happens nearby houses around USC is only aimed at personal belongings, but there are also types that may damage body or threaten life.

In practice, my findings in this final project can be used for prediction: I can use the machine learning model to predict the approximate crime number in a location. Moreover, the findings on crime patterns and the most likely crime types in a region can help us get to know the crime situation in advance when we are finding houses to live.

5 Limitations, Challanges and Future Works

5.1 Limitations

My data set is not sufficient and comprehensive. I believe there are lots of houses not included in the data set, like where I'm living now. Thus, my work is also not comprehensive to conclude on all the houses around USC including the houses not in the data set, and the machine learning model may be not strong enough in predicting all of the houses.

5.2 Challanges

The most challanging difficulty in my researching process is deciding what to do. At the begining, I found that the correlation coefficient is low between variables, so I can not directly perform linear regression on the data. At this time I was confused on what should I do, what method is appropriate for researching on my data. I have done lots of useless works until I found something meaningful to do. Actually, lots of work shown above is experienced from some failed works I did, which are not included here.

5.3 Future Works

For future works, I think the first thing to do is complement the limitation of my data set. The house data are all from apartments.com, and I can find data from other house sources to make the data set sufficient and larger, then I can do the same research on the data set. Next, as mentioned in Homework 3 and 4, some proposals on the data set is not implemented when retriving data because these functions are not used in the final research. I can implement them to make the data set more powerful, or even upgrade them as an imformation retrieving program. Lastly, due to the fact that the crime data is based on time(the time range from when the crimes are retrieved) and distance(the radius of the circular range of a house to find crimes), I need to update the data because they will

change as time goes by, and I can also decide a wider or narrower range to search for crimes nearby a house, then do the same research on the data set and yield new results.