Final Project of Pattern Recognition

K-means Cluster and interesting data Analysis of LA Arrest Data

Zeyu Liu

Professor Loew

ECE 6850

December 11, 2019

The George Washington University

Abstract

In this final project, I use R language to process the arrest data in Los Angeles. Furthermore, visualize processes and results as much as possible. Due to the particularity of the data, I chose the unsupervised model K-means Cluster first to find the arrest area distribution and got amazing results. And then, I did some interesting description analysis of the whole data and answered some mathematic questions. After that, I use the linear regression to predict the future arrest numbers. Finally, by processing the arrest location data, I counted the number of arrests per kilometer on PICO Avenue.

Keywords: K-means Cluster, analysis, location

Introduction

The dataset I used in the paper named “Arrest Data from 2010 to Present”, This dataset provided by Los Angeles Police Department and has been continuously updated. I chose the dataset which update date is 10/31/2019, it contains 1.3 million of arrest dates, each data contains 17 parameters. Only few variables are numeric. The variables are showing in *Table 1*.

Variables	Description	Type
Report ID	ID for the arrest.	Text
Arrest Date	MM/DD/YYYY	Time
Time	In 24-hour military time.	Text
Area ID	The LAPD has 21 Community Police Stations referred to as Geographic Areas. These Geographic Areas are sequentially numbered from 1-21.	Text
Area Name	The 21 Geographic Areas’ name	Text
Reporting District	A four-digit code that represents a sub-area within a Geographic Area.	Text
Age	Two character numeric.	Text
Sex Code	F - Female M - Male	Text
Descent Code	Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian	Text
Charge Group Code	Category of arrest charge.	Text
Charge Group Description	Defines the Charge Group Code provided.	Text
Arrest Type Code	A code to indicate the type of charge the individual was arrested for. D - Dependent F - Felony I - Infraction M - Misdemeanor O - Other	Text
Charge	The charge the individual was arrested for.	Text
Charge Description	Defines the Charge provided.	Text
Address	Street address of crime incident.	Text
Cross Street	Cross Street of rounded Address.	Text
Location	The location where the crime incident occurred.	Location

Table 1. Dataset parameters description.

The dataset can be download from the websites: https://catalog.data.gov/dataset/arrest-datafrom-2010-to-present

The information of the dataset: https://data.lacity.org/A-Safe-City/Arrest-Data-from-2010-toPresent/yru6-6re4

Furthermore, the steps of the project are showing in *Figure 1*.

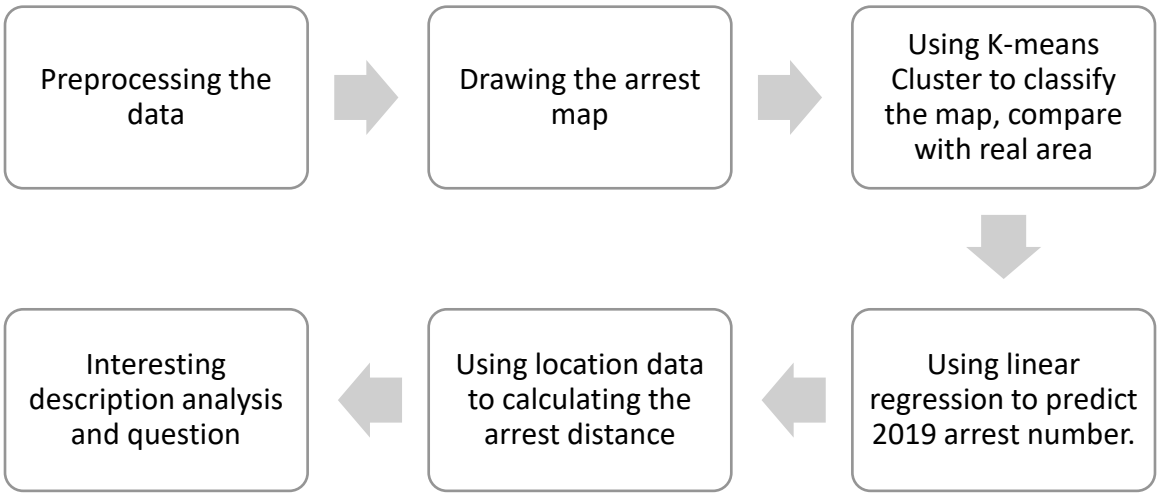

Figure 1. The steps of the project.

Approach

1. Preprocessing the data.

There are the key library and some important function I used in R.

library(ggplot2), use this package to visualize all the figures better.

library(tidyverse), use this package to divide the location to be latitude and longitude.

na.omit(), the code is consider to omit the NA.

sort(), order(), data sorting in different way.

grep(), magic grab.

as.numeric(), as.factor(), data format exchange.

kmeans(), K-means cluster.

2. Drawing the arrest map.

By using the latitude and longitude in location data. We can draw the arrest distribution map with different color which represent the different arrest type in *Figure 2.*

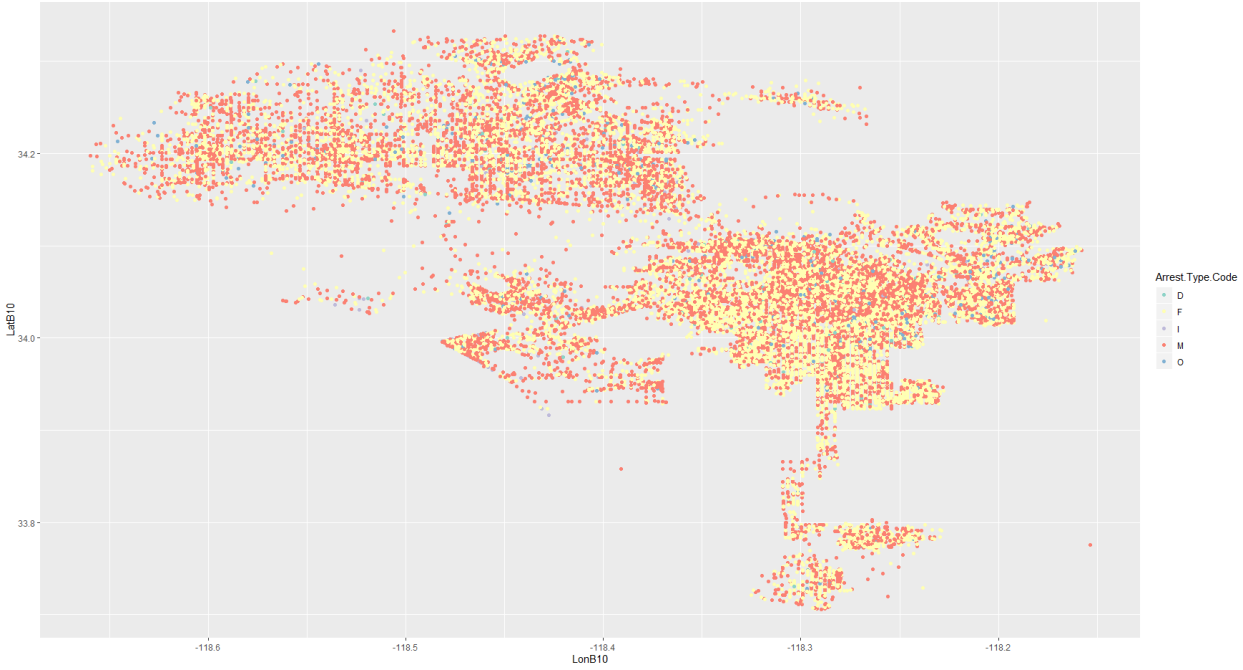

Figure 2. The arrest map in LA. D - Green F - Yellow I - Purple M - Red O - Blue

3. Using K-means Cluster to classify the map, compare with real area

First, I use different color to represent the real 21 areas in arrest map, and then I use K-means cluster to classify the arrest map to 21 class. The outcome is show in *Figure 3* and *Figure 4.*

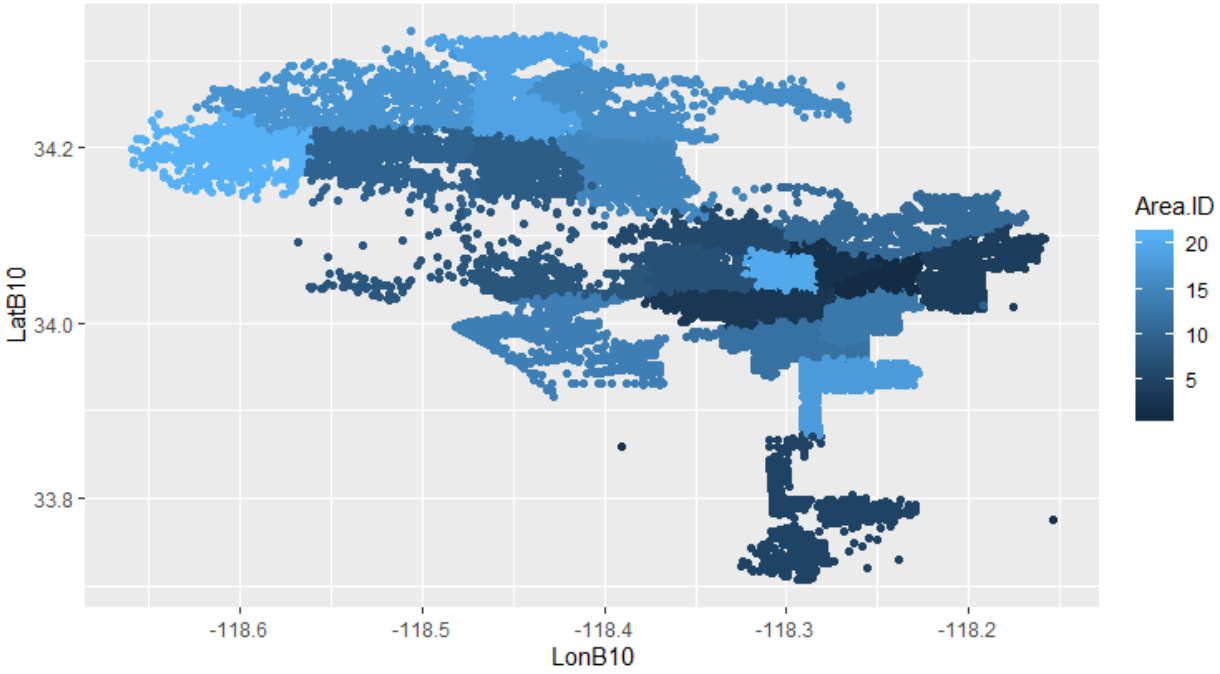

Figure 3. The arrest map in LA. The real area shows in different color.

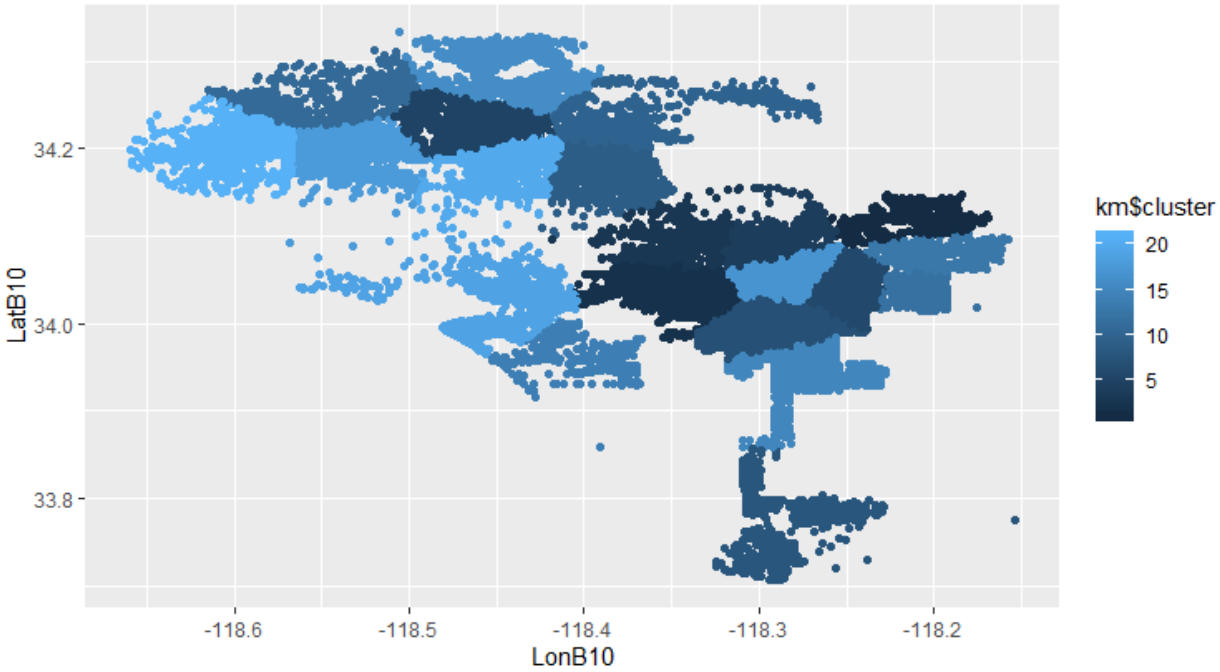

Figure 4. The arrest map in LA. The K-means cluster class shows in different color.

They are magically similar in the two figures. Suppose each K-means cluster area corresponds to a different real area in Los Angeles. We can draw the boxplot below.

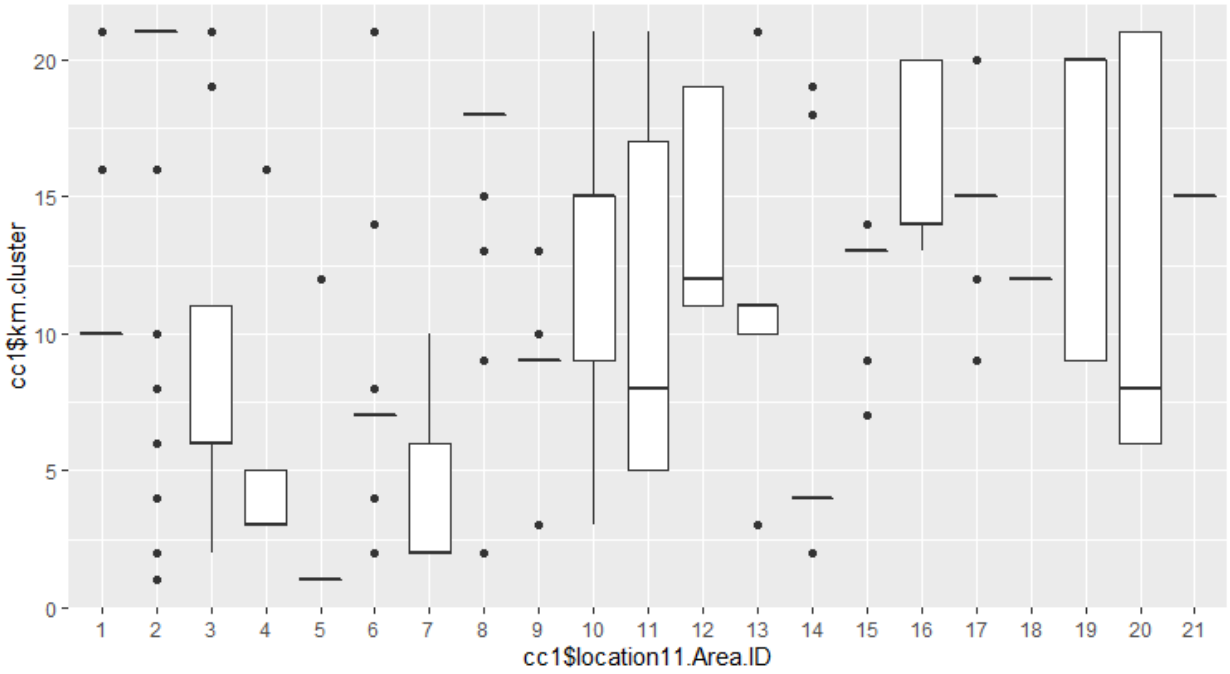

Figure 5. Boxplot for Area.ID with K-means.ID(cluster). Some classes can be classified correctly.

From the boxplot *Figure 5*, we can use some mathematic method to estimate and analysis the relationship between the K-means.ID(cluster) and the Area.ID. The final results are showing in the *Table 2.*

Area.ID	**K-means.ID**		**Area.ID**	**K-means.ID**
1	10	**Best estimate →**	1	10
2	21		2	21
3	6~11		3	6
4	3~5		4	3
5	1		5	1
6	7		6	7
7	2~6		7	2
8	18		8	18
9	9		9	9
10	9~15		10	14
11	5~17		11	5
12	11~18		12	16

13	10~11		13	11
14	4		14	4
15	13		15	13
16	14~20		16	17
17	15*		17	19
18	12		18	12
19	9~20		19	20
20	6~21		20	8
21	15		21	15

Table 2. Left: the K-means.ID estimate from the boxplot; Right: Best estimate compare with the Area.ID.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
10	9818	59	6	0	0	0	1	0	1	0	0	0	2023	0	0	0	0	0	0	0	0
21	251	6481	108	0	0	1	0	0	0	1	209	0	17	0	0	0	0	0	0	1382	0
6	0	4	2893	0	0	0	247	0	0	0	0	0	0	0	0	0	0	0	0	1469	0
3	0	0	0	2385	0	0	0	0	1	1	0	0	3	0	0	0	0	0	0	0	0
1	0	2	0	0	3781	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	8239	505	0	0	0	5	0	0	0	111	0	0	0	0	1	0
2	0	1	1040	0	0	4	1700	424	0	0	0	0	0	77	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	2238	0	0	0	0	0	616	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	13	5901	1447	0	0	0	0	219	0	233	0	2117	0	0
14	0	0	0	0	0	1	0	0	0	0	0	0	0	0	154	1988	0	0	49	0	0
5	0	0	0	698	0	0	0	0	0	0	1121	0	0	0	0	0	0	0	0	0	0
16	882	129	0	507	0	0	0	0	0	0	515	0	0	0	0	0	0	0	0	0	0
11	0	0	811	0	0	0	0	0	0	0	0	2274	3384	0	0	0	0	0	0	0	0
4	0	1	0	0	0	1	0	0	0	0	0	0	0	5666	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	9	290	0	0	0	0	0	4463	2	0	0	1	0	0
17	0	0	0	0	0	0	0	0	0	0	997	0	0	0	0	0	0	0	0	0	0
19	0	0	1069	0	0	0	0	0	0	0	0	1970	0	416	0	0	0	0	0	0	0
12	0	0	0	0	94	0	0	0	0	0	0	2014	0	0	0	0	1	3601	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1015	758	0	2190	0	0
8	0	667	0	0	0	1462	11	0	0	0	605	0	0	0	0	0	0	0	0	1046	0
15	0	0	0	0	0	0	0	3	0	2112	0	0	0	0	0	0	2154	0	0	0	3094

Figure 6. confusion matrix from boxplot and *Table 1.* (The x-axis is Area.ID, y-axis is K-means Cluster ID)

From the *Table 2*, we can build the confusion matrix. By looking at the confusion matrix in *Figure 6*, to make the result better, we can exchange K-means cluster 14 to 17, 16 to 19. And we will get the better confusion matrix below in *Figure 7*.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
10	9818	59	6	0	0	0	1	0	1	0	0	0	2023	0	0	0	0	0	0	0	0
21	251	6481	108	0	0	1	0	0	0	1	209	0	17	0	0	0	0	0	0	1382	0
6	0	4	2893	0	0	0	247	0	0	0	0	0	0	0	0	0	0	0	0	1469	0
3	0	0	0	2385	0	0	0	0	1	1	0	0	3	0	0	0	0	0	0	0	0
1	0	2	0	0	3781	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	8239	505	0	0	0	5	0	0	0	111	0	0	0	0	1	0
2	0	1	1040	0	0	4	1700	424	0	0	0	0	0	77	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	2238	0	0	0	0	0	616	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	13	5901	1447	0	0	0	0	219	0	233	0	2117	0	0
17	0	0	0	0	0	0	0	0	0	0	997	0	0	0	0	0	0	0	0	0	0
5	0	0	0	698	0	0	0	0	0	0	1121	0	0	0	0	0	0	0	0	0	0
19	0	0	1069	0	0	0	0	0	0	0	0	1970	0	416	0	0	0	0	0	0	0
11	0	0	811	0	0	0	0	0	0	0	0	2274	3384	0	0	0	0	0	0	0	0
4	0	1	0	0	0	1	0	0	0	0	0	0	0	5666	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	9	290	0	0	0	0	0	4463	2	0	0	1	0	0
14	0	0	0	0	0	1	0	0	0	0	0	0	0	0	154	1988	0	0	49	0	0
16	882	129	0	507	0	0	0	0	0	0	515	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	94	0	0	0	0	0	0	2014	0	0	0	0	1	3601	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1015	758	0	2190	0	0
8	0	667	0	0	0	1462	11	0	0	0	605	0	0	0	0	0	0	0	0	1046	0
15	0	0	0	0	0	0	0	3	0	2112	0	0	0	0	0	0	2154	0	0	0	3094

Figure 7. confusion matrix exchange K-means cluster 14 to 17, 16 to 19. (The x-axis is Area.ID, y-axis is K-means Cluster ID)

Now we can calculate the accuracy of the K-means cluster classification by using confusion matrix function. The concept of the standard confusion matrix is showing in *Table 3*. Judging the multiple input multiple output model, the calculation method is the same.

		Actual class	
		positive class	negative class
Predicted class	positive class	True Positive (TP)	False Positive (FP)
	negative class	False Negative (FN)	True Negative (TN)

Table 3. Standard confusion matrix.

We can calculate the accuracy of classification by using the function:

$$\textbf{accuracy} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN}} = \frac{\textbf{All diagonal data}}{\textbf{All data}}$$

After mathematical calculations, we obtained the accuracy of each classification and the overall classification accuracy.

Overall accurary

$$= \frac{9818 + 6481 + 2893 + 2385 + 3781 + 8239 + 1700 + 2238 + 5901 + 1121 + 1970 + 3384 + 5666 + 4463 + 1988 + 3601 + 2190 + 1046 + 3094}{104260}$$

$$= 69.02\%$$

The result is showing in *Table 3.*

Area.ID	K-means.ID	Accuracy
1	10	89.65%
2	21	88.25%
3	6	48.81%
4	3	66.43%
5	1	97.57%
6	7	84.87%
7	2	68.99%
8	18	83.29%
9	9	95.28%
10	17	0%
11	5	32.47%
12	19	31.48%
13	11	62.35%
14	4	83.63%
15	13	90.22%
16	14	66.16%
17	16	0%
18	12	100%
19	20	50.26%
20	8	26.83%
21	15	100%
Overall accuracy： 69.02%		

Table 4. Accuracy of K-means cluster classification.

That is an amazing accuracy of an unsupervised classifier. More than 69% arrest data area can be classified correctly in the real area of Los Angeles. Even some areas can be classified perfectly with a very high accuracy, like area 1,2,5,9,15,18,21. Only few data cannot be classified correctly like area 10 and 17.

In another situation, if one K-means cluster can be divided into more than one class, such as when K-means cluster class 15 divided by area.ID 9 and 17, 21. The overall classification accuracy will be:

Overall accurary

$$= \frac{\mathbf{9818 + 6481 + 2893 + 2385 + 3781 + 8239 + 1700 + 2238 + 5901 + 1121 + 1970 + 3384 + 5666 + 4463 + 1988 + 3601 + 2190 + 1046 + 3094 + 2112 + 2154}}{\mathbf{104260}}$$

$$= \mathbf{73.11}\%$$

The K-means cluster classify result is improved better than before. The further analysis will be found at the result part.

4. Using linear regression to predict 2019 arrest number.

To using linear regression, I calculate the number of arrests from the year 2010 to 2018 and integrate them into a training set. Then I calculate the arrests data from the year 2019 as test set. The result is in *Figure 8*.

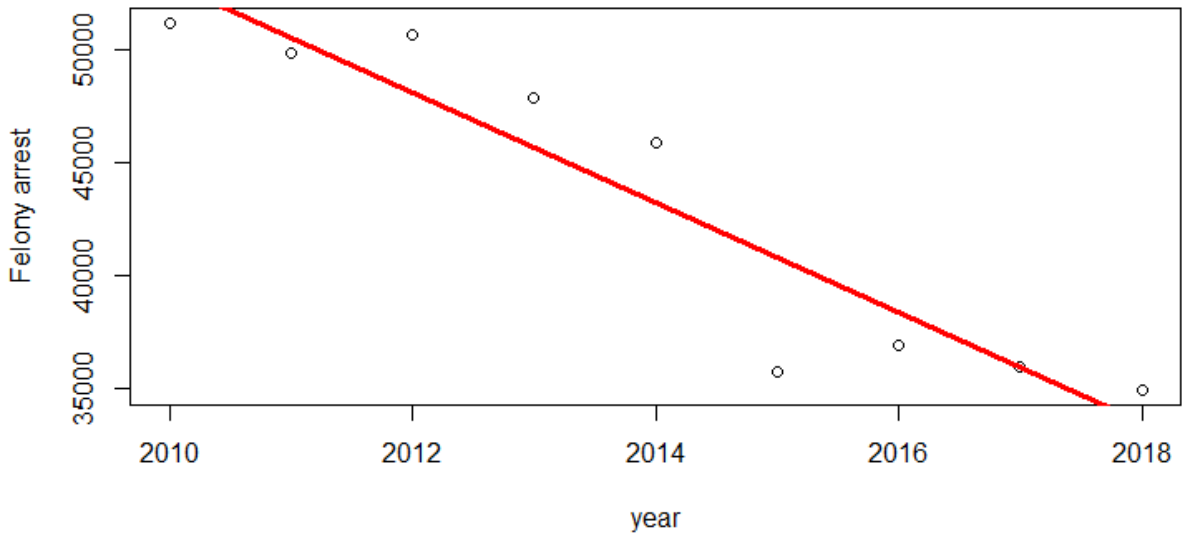

Figure 8. Linear regression prediction for the arrest in 2019.

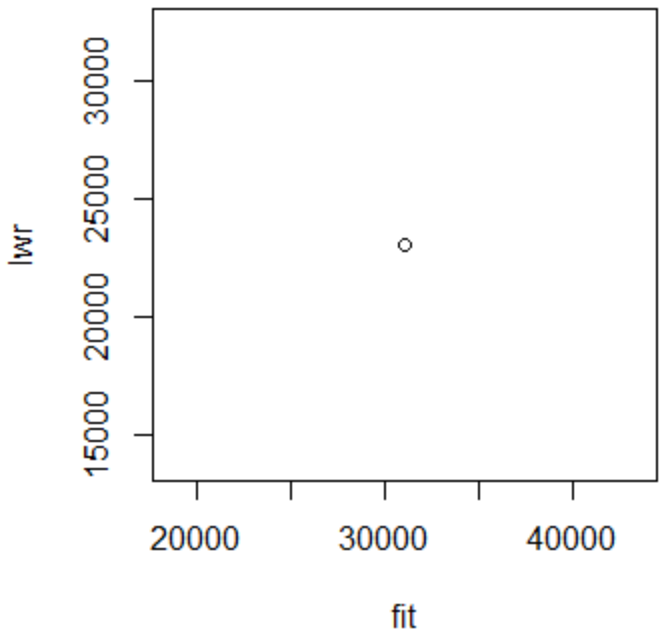

Figure 9. predict value for the arrest in 2019.

Because the dataset is only covered the data before November 1st. The true value should smaller than the predict value in the result. The predict value is 31038 in *Figure 9*, and the true value is 28681 can be calculate from the dataset. 28681<31038. The predictor looks work good.

5. Using location data to calculate the arrest distance.

I use an example to solve the problem. calculate the number of arrest incidents per kilometer on Pico Boulevard in 2018. The arrests distribution is showing in the *Figure 10*. In the figure, apparently, we need remove the outliers first. But which outlier should we move? The method I used is remove outliers by filtering out locations where either the latitude or longitude is 2 standard deviations beyond the mean of the subset of identified points. The transformation process is showing in *Figure 11*, *Figure 14.*

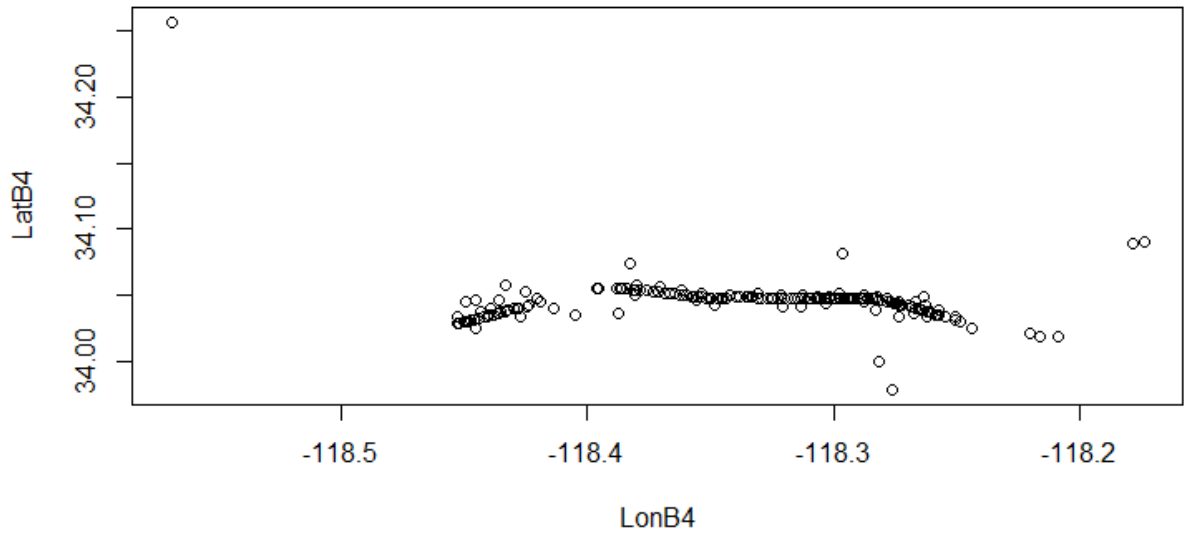

Figure 10. Arrest location with the Pico Boulevard address.

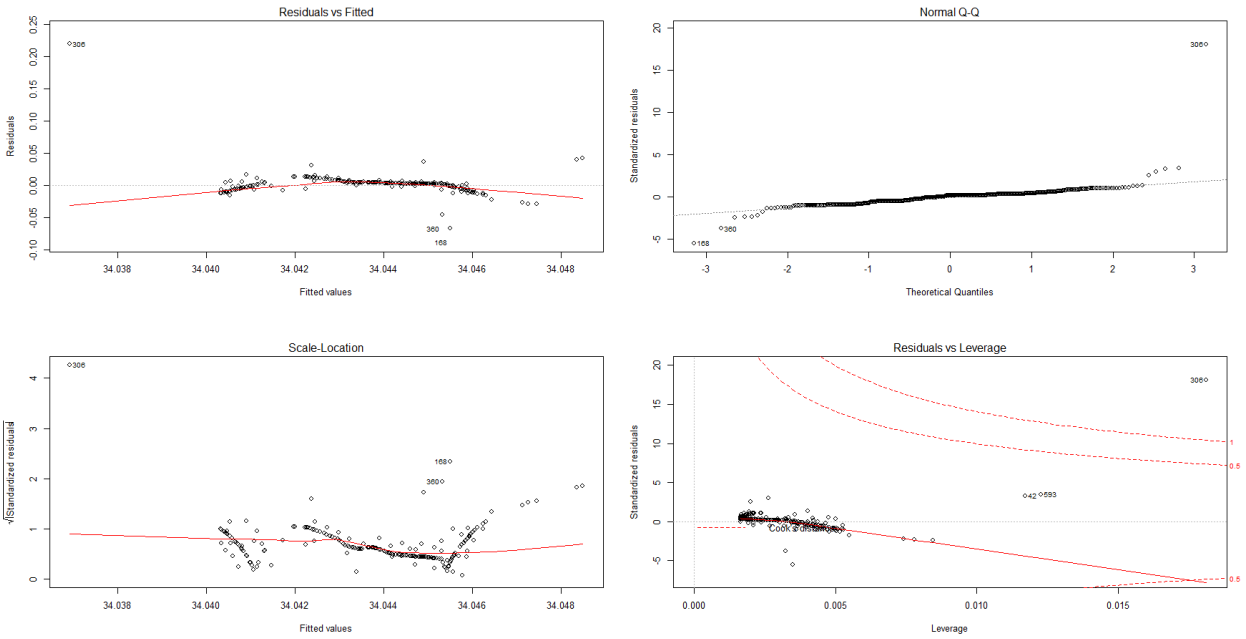

Figure 11. The original information of the arrest data in PB address. There are apparently outliers.

In *Figure 12* and *Figure 13*, after remove the outliers, the St. Error becomes small, p-value becomes small, R square become large, F-statistic become large, these results mean the data is

more related. Although the true value of the dataset maybe not a straight. The linear trend has improved in the dataset. We did remove the outliers.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.501085   0.849313  44.155  < 2e-16 ***
LonB4        0.029216   0.007176   4.071 5.29e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01224 on 611 degrees of freedom
Multiple R-squared:  0.02641,   Adjusted R-squared:  0.02482
F-statistic: 16.57 on 1 and 611 DF,  p-value: 5.291e-05
```

Figure 12. The statistic original information of the arrest data in PB address.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.697233   0.552047   71.91   <2e-16 ***
LonB4        0.047775   0.004665   10.24   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.007883 on 608 degrees of freedom
Multiple R-squared:  0.1471,    Adjusted R-squared:  0.1457
F-statistic: 104.9 on 1 and 608 DF,  p-value: < 2.2e-16
```

Figure 13. The statistic information of the arrest data in PB address. After remove the outliers.

In *Figure 15*, it shows the result of arrest location with the Pico Boulevard address after processing. By using spherical earth projected equation which shows in *Figure 16.* We can get the distance and the arrest numbers/kilometer:

D = 16.83 km

602/D = 35.77 arrest/km

To validate the distance. I use google map to measure the real distance of the Pico Boulevard street. It shows D1 = 16.38 km, that is similar to the distance I predicted.

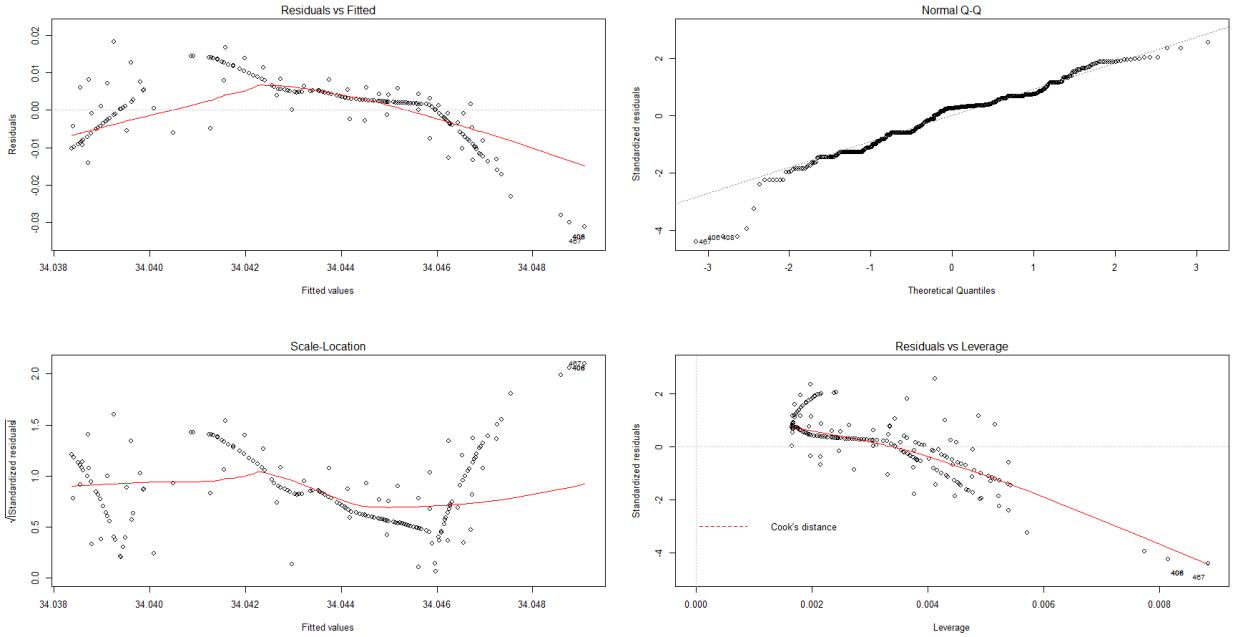

Figure 14. The information of the arrest data in PB address. After remove the outliers.

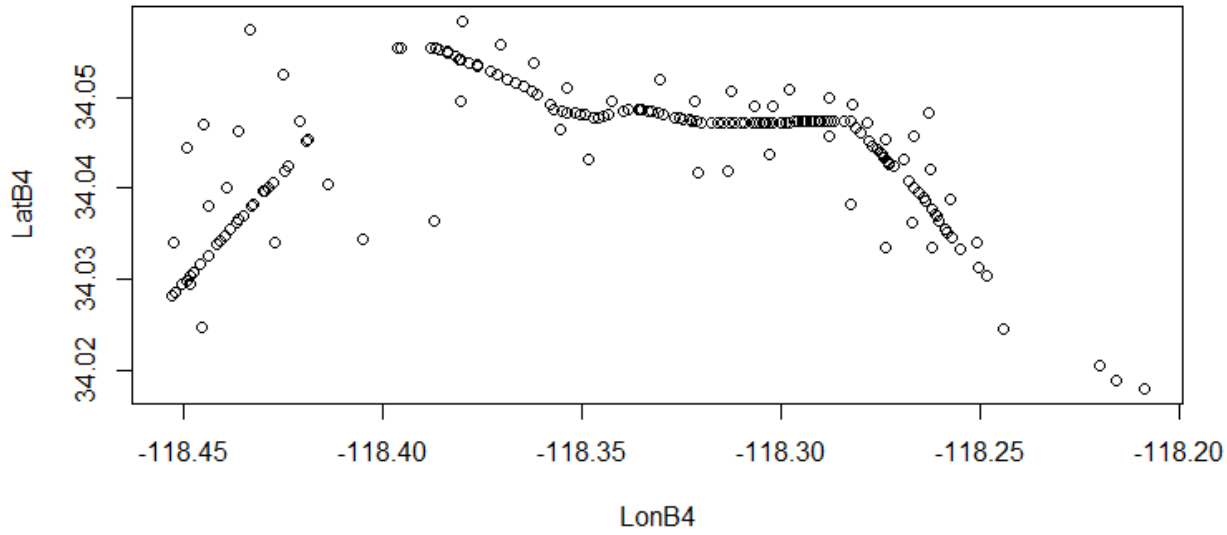

Figure 15. Arrest location with the Pico Boulevard address after processing.

The real Pico Boulevard street distance in google map shows in *Figure 17 and Figure 18.*

Ellipsoidal Earth projected to a plane

The FCC prescribes the following formulae for distances not exceeding 475 kilometres (295 mi):

$$D = \sqrt{(K_1 \Delta\phi)^2 + (K_2 \Delta\lambda)^2},$$

where

D = Distance in kilometers;

$\Delta\phi$ and $\Delta\lambda$ are in degrees;

ϕ_m must be in units compatible with the method used for determining $\cos(\phi_m)$;

$K_1 = 111.13209 - 0.56605\cos(2\phi_m) + 0.00120\cos(4\phi_m)$;

$K_2 = 111.41513\cos(\phi_m) - 0.09455\cos(3\phi_m) + 0.00012\cos(5\phi_m)$.

Figure 16. Spherical Earth projected equation

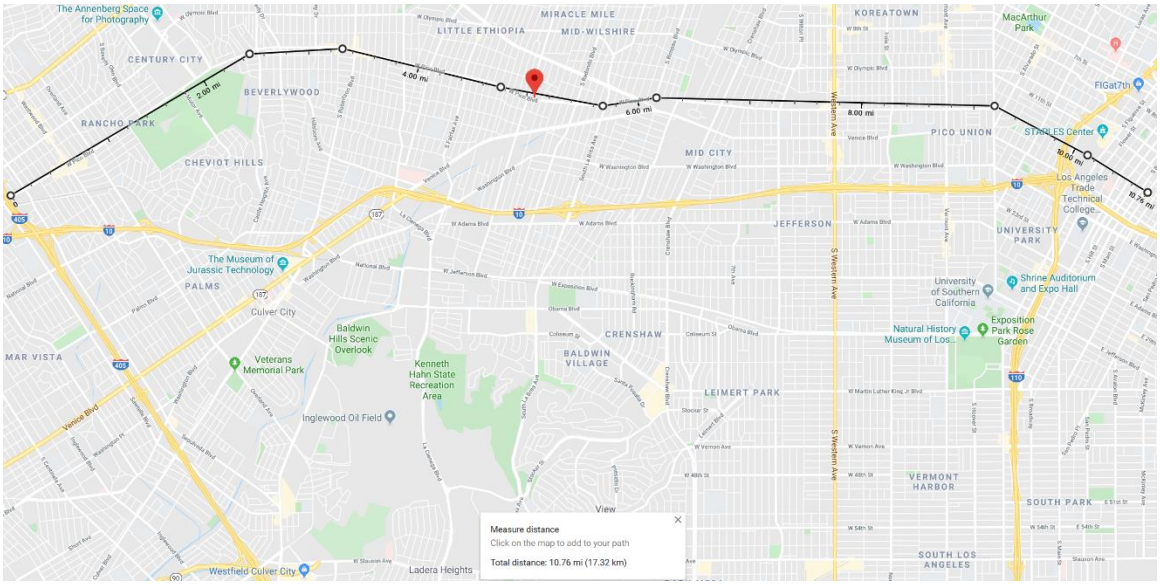

Figure 17. The distance of the Pico Boulevard address.

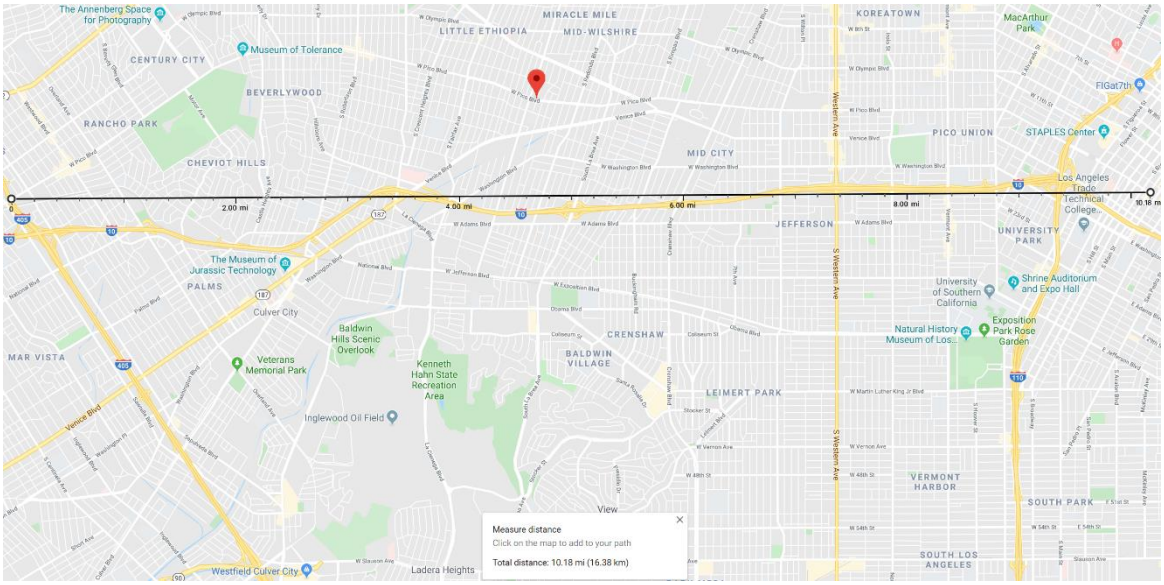

Figure 18. The distance of the Pico Boulevard address by using projection.

6. Interesting description analysis and question

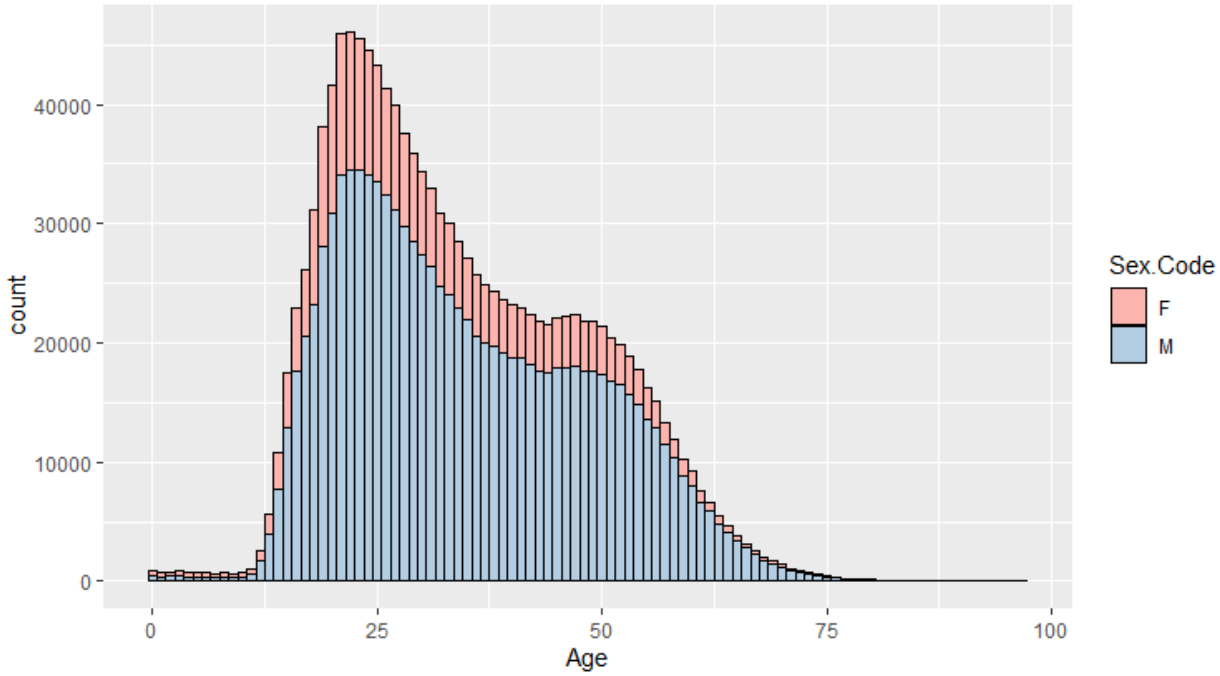

Figure 19. The arrest data compare with the Age and Sex. F = female, M = Male.

In *Figure 19*, either woman or man who has been arrest at two peaks age. One is 23 years old, another one is 47 years old.

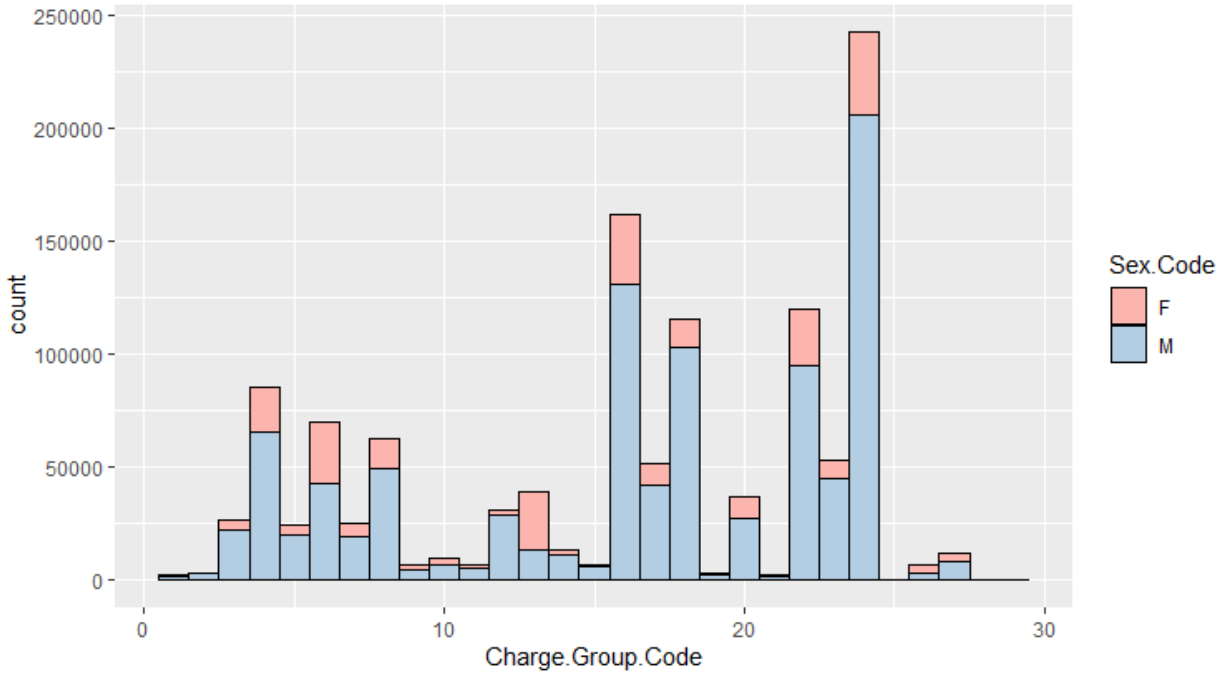

Figure 20. The arrest data compare with the Charge. Group and Sex

In *Figure 20*, there is a special rate in group 13. It's means this charge the women have a more arrest rate than men. We can find it is Prostitution/Allied. No discriminate here.

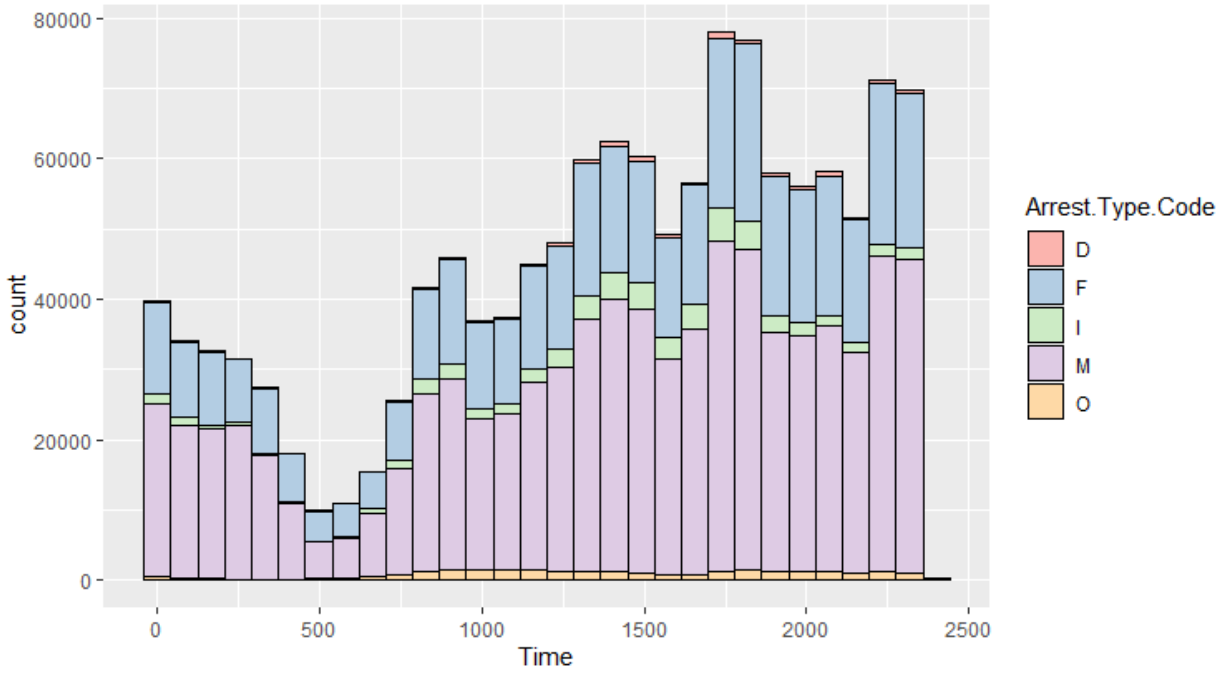

Figure 21. The arrest data compare with the Time and Arrest. Type.

In *Figure 21*, the arrest numbers have a regular variety. The lowest point at 5:00, the largest range is from 18:00 to 00:00. Most all the charge is felony and misdemeanor.

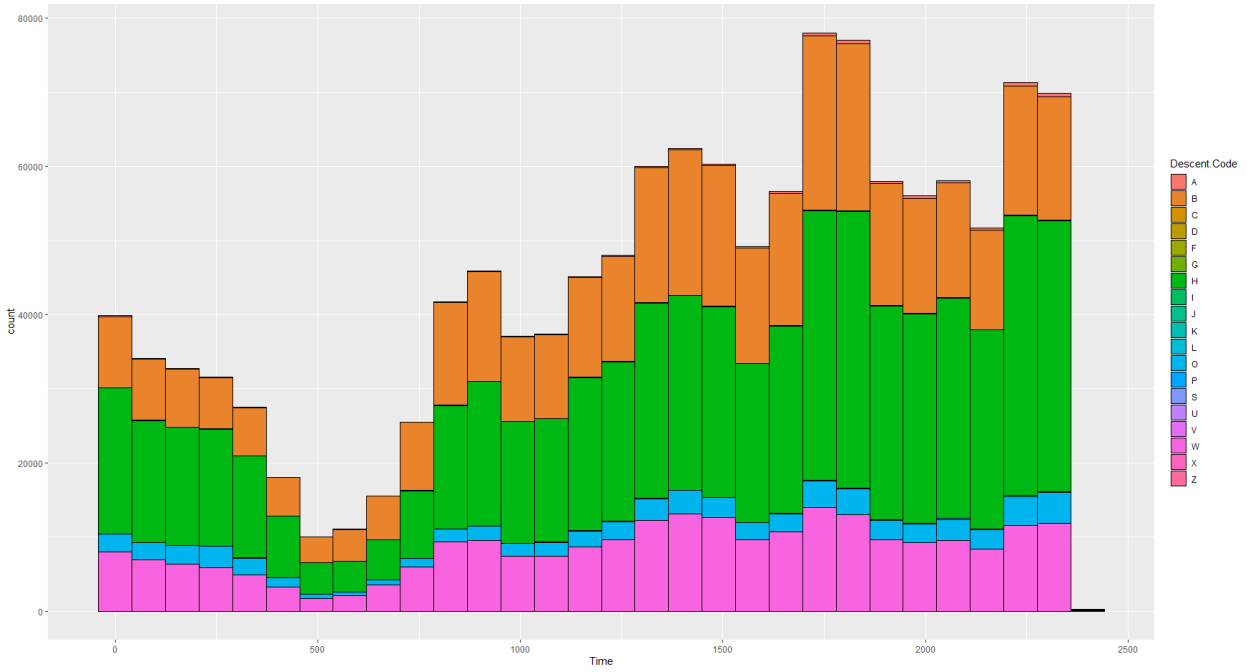

Figure 22. The arrest data compare with the Time and Descent. Code.

In *Figure 22*, Although there are 19 descent, but only three descent is the main arrest target.

Orange – Black, green - Hispanic/Latin/Mexican, pink – White. No discriminate here.

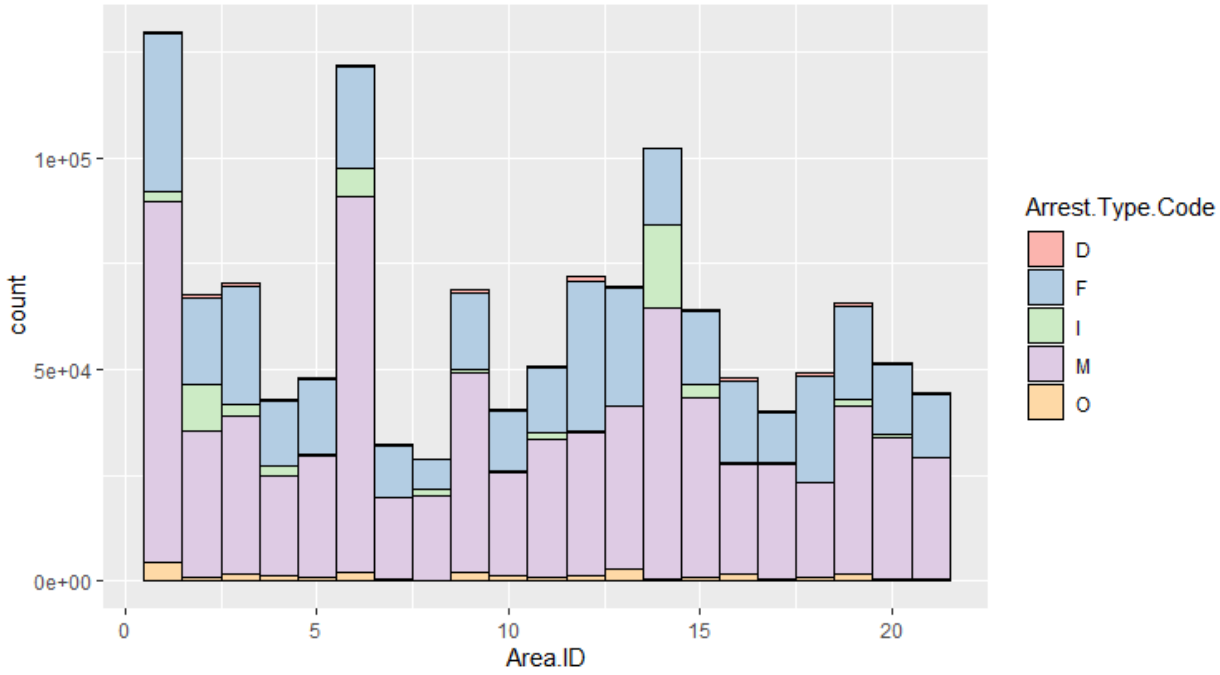

Figure 23. A code to indicate the type of charge the individual was arrested for. D - Dependent F - Felony I - Infraction M - Misdemeanor O – Other.

In *Figure 23*, there are two strange area. The number 2 and 14. They has a high rate arrest of infraction. The real area map of them are showing in *Figure 24* and *Figure 25*.

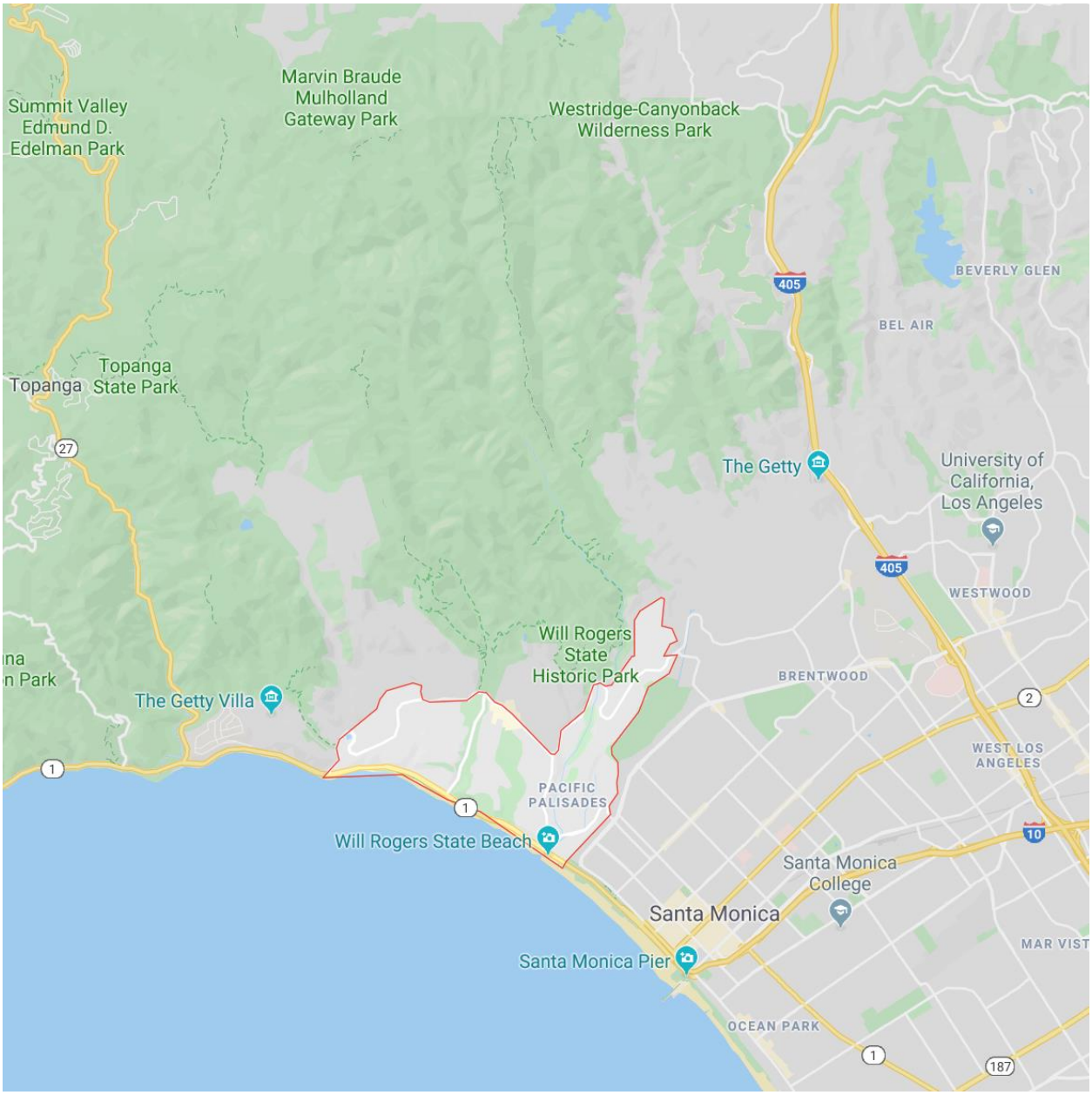

Figure 24. Pacific area.

In *Figure 24*, this area is located at the country yard with No.1 and No.23 Highway junction. The speeding violation is the main reason for the arrest of infraction.

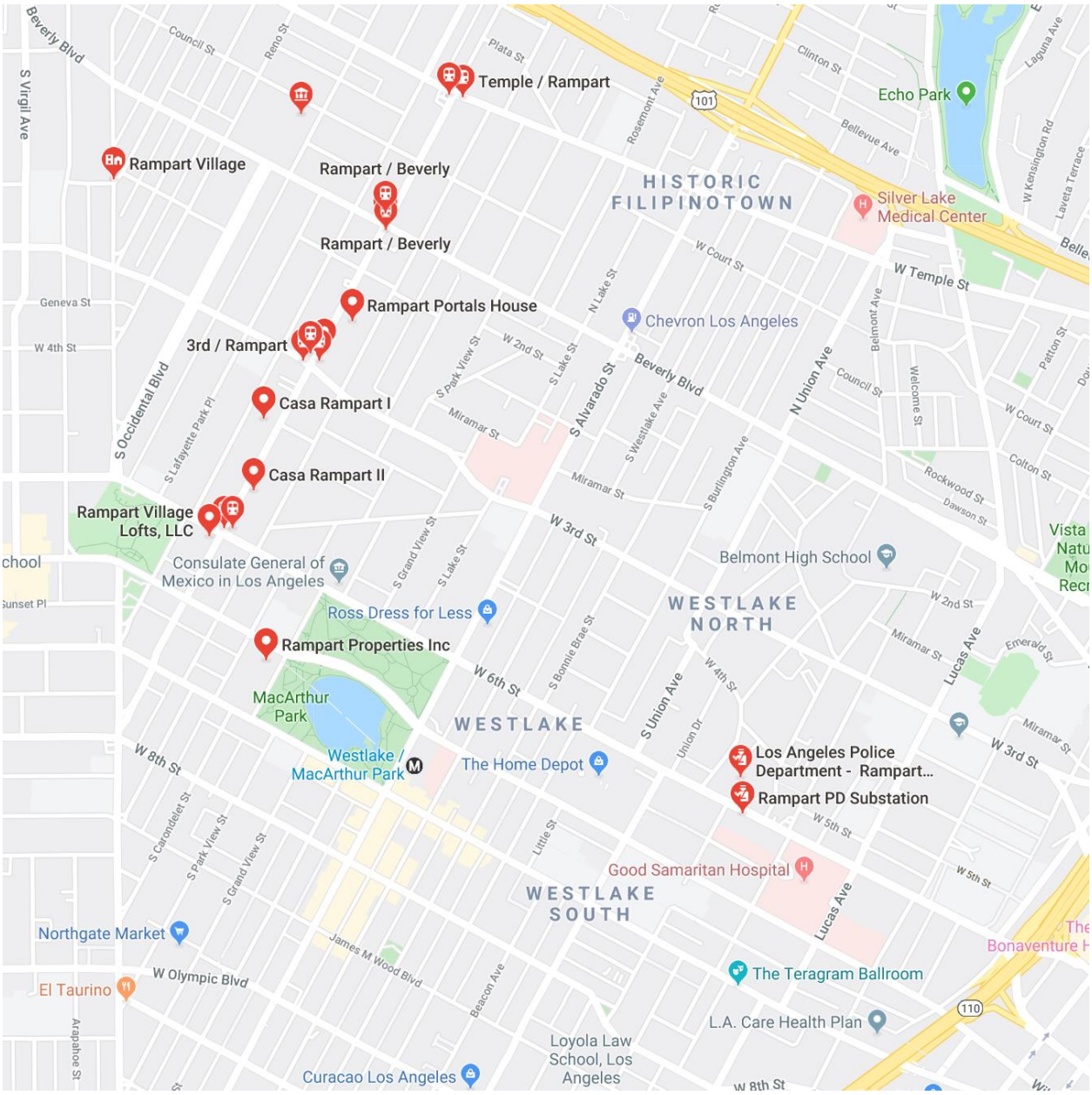

Figure 25. Rampart area.

In *Figure 25*, the LA police department is in this area. This shows the opposite of our common sense. The reason for a high infraction is the high density of police officer and the strict of enforce. Another inference is in other areas, many of the infraction crime are not always be found.

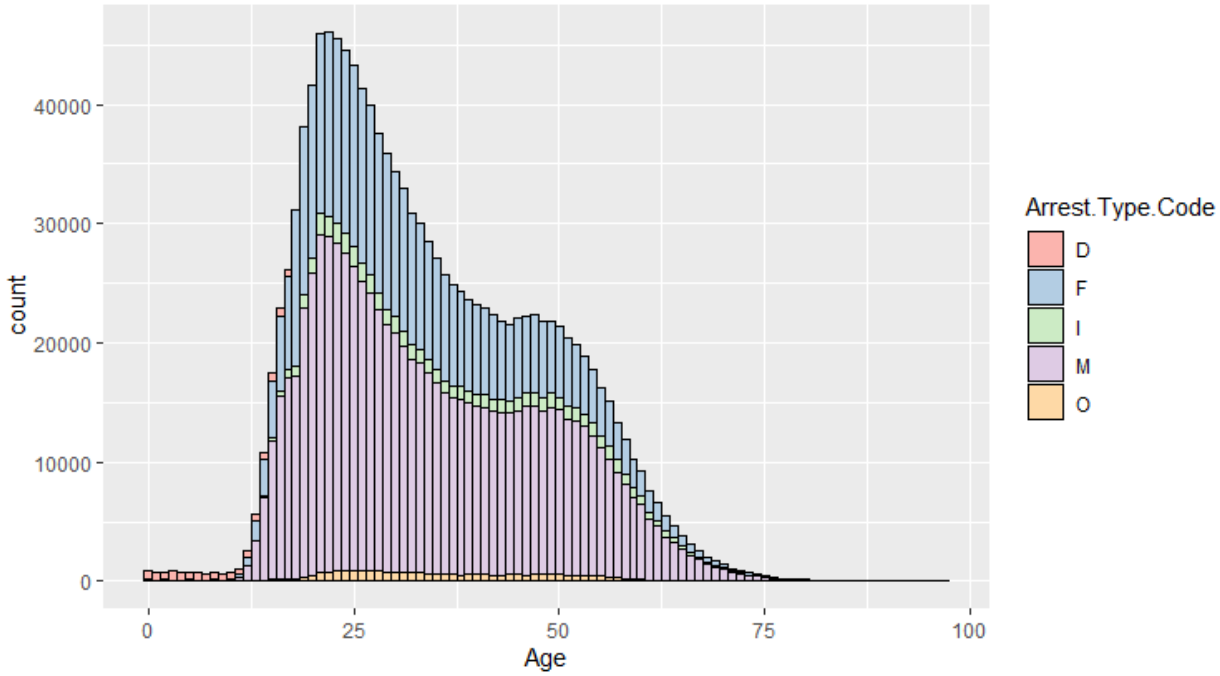

Figure 26. The arrest data compare with the Age and Arrest. Type. Code.

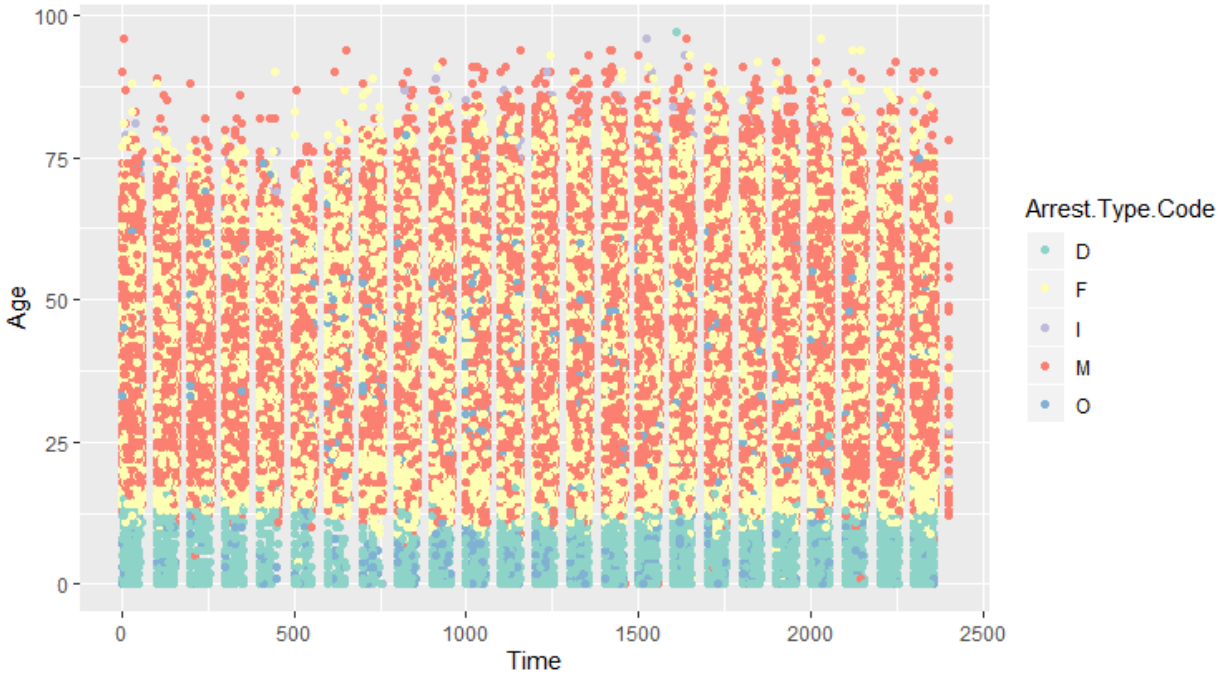

Figure 27. The arrest data compare with the Time, Age and Arrest. Type. Code.

In *Figure 26 and 27*, they tell us the arrest type of dependent and infraction can be divided by the age. Most arrests under an age are D. The margin can be calculate by using many ways.

Result Analysis

From the K-means cluster classify in this dataset, I was shocked by the accuracy of the model. This means that in real life, K-means cluster is an unsupervised method consistent with the laws of human recognition. That because it is based on Euclidean distance. When there is enough data and the category size is moderate, K-means cluster can also bring good classification results. The biggest advantage of K-means cluster is that it has low complexity and fast calculation and classification. Since kmeans is unsupervised, we do not need output parameters during the training processing. But we can test it with the known classification results. Just like in this example, we can also use Kmeans cluster area for data analysis and prediction. It may get surprising results. This is also an opportunity for future experiments.

Acknowledgement

This is the first time I have attended the professor Murray Loew's class. Not only did he teach well, he was also a "spiritual" teacher. I have learned much more from him. Because of the potential motivation, I have read the whole textbook word by word without mandatory requirement. Without his course, I did not even know how to write the R code, I'm so appreciate to him. Moreover, I also would like to thank my classmates for the help from them.

References

[1] *Code of Federal Regulations (annual edition). Title 47: Telecommunication*. 73.208. October 1, 2016. Retrieved 8 November 2017.

[2] Wickham, Hadley (July 2010). "ggplot2: Elegant Graphics for Data Analysis". *Journal of Statistical Software. 35 (1).*

[3] Massart, D.L., Smeyers-Verbeke, J., Capron, X., & Schlesier, K. (2005). Visual presentation of data by means of box plots.