Detailed Report and Analysis for Data Set: police_department_data

Submitted By: Pratik B. Kanase
kanasepratik2@gmail.com
+91-8087475700

Table of Contents

1. Information about the Dataset

About the Dataset -
The dataset can be downloaded from - https://bit.ly/2EfvRaG
The Data Description is as follows -

- Incident_id - A number assigned to each incident reported.
- Category - Category of incident reported
- Crime description - Description explaining the nature of crime.
- Crime_date - date on which the crime was reported.
- Department_district - district in which the police department is lo cated.
- Resolution - Details of resolution (if any).
- Address - Address where the crime occurred.
- Department_id - police department id.
- Location - lat-long location where the crime was committed.

2. Objective of the Analysis

- To find, analyze and interpret the pre-existing patterns and trends that are to be uncovered through the dataset.
- The aim is to extract actionable insights from the analysis for improving the overall safety of the city by learning from past experiences.

.

3. Formation of the Problem Statement

Problem Statement:

- To find the actionable insights from the data, provide suggestions and conclusion from the analysis of the data by doing the statistical and visualization techniques.

- To interpret the patterns from the data and compile them into useful information for further processing by applying analytical skills and techniques.
- To apply the various feature engineering and data preprocessing techniques to gain the most useful results from the data.
- To apply the machine learning models to predict the crime description depending upon the other independent features present in the dataset.
- To evaluate and choose the best working, suitable and stable machine learning model for the dataset.
- To build the machine learning with the help of analysis of data and statistical techniques which can predict the '**Description explaining the nature of crime**'

4. Data and Features Understanding

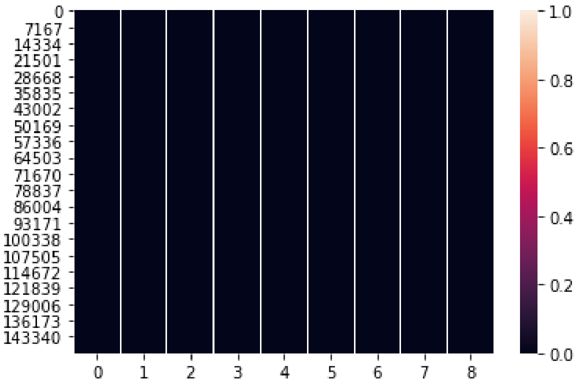
- The following information shows the details regarding with the data present in the dataset. It concludes shapes of dataset, feature information, its data types and other basic information.

```
Shape of Dataset (150500, 9)

Number of Rows 150500
Number of Columns:  9


Feature Names :
 ['incident_id' 'category' 'crime_description' 'crime_date'
 'department_district' 'resolution' 'address' 'department_id' 'locat
ion']



Unique values per column:
incident_id            116699
category                   39
crime_description         726
crime_date              67140
department_district        10
resolution                 14
address                 16130
department_id          150500
location                19386
```

5. Data Exploration

Missing data Analysis:

There wasn’t any missing data or values in the dataset. With the help of this heatmap we can understand the data is not missing and ready for the analysis



6. Data Preprocessing

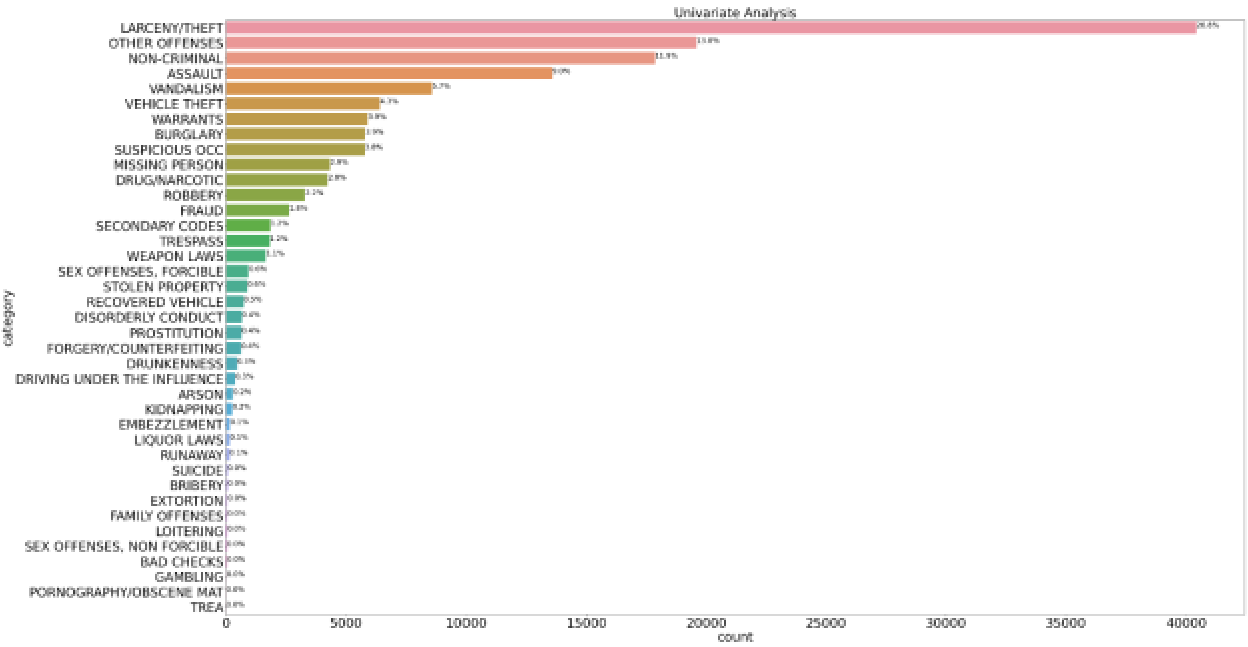
- Before moving ahead with the analysis of the dataset, I performed some data cleaning and features engineering techniques to clean the data so that we can analyze it in the better way.
- I converted some datatypes and extracted the useful information from the date and time features.
- With the help of this features I created some new features related with it. The use and analysis of these features is described below in the more detailed manner.

7. Exploratory Data Analysis

- Univariate Analysis. Bivariate Analysis
- Statistical , Hypothesis Analysis
- Data Visualization

1. Analysis for the Feature: Category of incident reported

- Following data shows that the percentage of category of incident reported out of the total percentage.
- From the analysis of the bar plot and the percentage analysis we can interpret here that around 26% of the incidents reported was for latency/theft and below that there are other offences and non-criminal incident reported during the year 2016.
- As most of the cases was related with larceny, theft and other offences, police department needs to focus on these cases in priority manner. As we can observe that the percentage count of the reported incidents is very leas comparing with the top five incidents. Hence there should be more focus on these activities to control the overall crim rate and security of the city.



- The above chart shows the univariate analysis of the feature ‘category of the incident’. Hence by analyzing thedata we can interpret the information about mosthapping incidents and crime and this information can be very helpful to take the actionable insightsBelow is the chart which shows the percentage count of each category.

Incident	**Percentage %**
LARCENY/THEFT	**26.849834**
OTHER OFFENSES	**13.022591**
NON-CRIMINAL	**11.871096**
ASSAULT	**9.021262**
VANDALISM	**5.706977**
VEHICLE THEFT	4.265116
WARRANTS	3.929568
BURGLARY	3.855150
SUSPICIOUS OCC	3.841860
MISSING PERSON	2.882392
DRUG/NARCOTIC	2.819269
ROBBERY	2.192027
FRAUD	1.750831
SECONDARY CODES	1.223256
TRESPASS	1.203987
WEAPON LAWS	1.101661
SEX OFFENSES, FORCIBLE	0.624585
STOLEN PROPERTY	0.586047
RECOVERED VEHICLE	0.489037

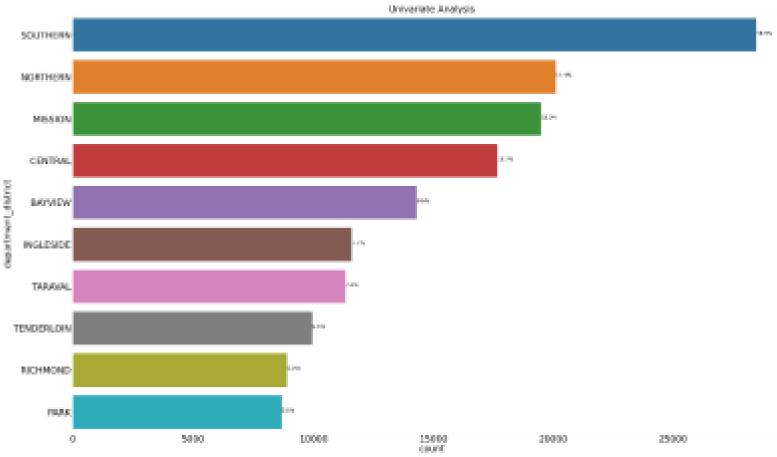
2. Analysis for the feature: crime_description - Description explaining the nature of crime.

Crime_Description	Percentage Count(%)
GRAND THEFT FROM LOCKED AUTO	**11.788040**
LOST PROPERTY	**3.053821**
AIDED CASE, MENTAL DISTURBED	**3.033887**
PETTY THEFT OF PROPERTY	**2.934219**
MALICIOUS MISCHIEF, VANDALISM	**2.831894**
BATTERY	2.798007
PETTY THEFT FROM LOCKED AUTO	2.653821
STOLEN AUTOMOBILE	2.394020
DRIVERS LICENSE, SUSPENDED OR REVOKED	2.243189
WARRANT ARREST	2.052492
FOUND PROPERTY	2.051827

- From the analysis of the crime description we can conclude that there are around 726 categories of crime was reported during the year 2016. Here is the mention of the top crime categories having most number of percentage count in the dataset.
- Almost 11.5% of the crime was related with the grand theft from the locked auto. And below that there are wide variety of the nature of crime cases happened through the year. From the analysis of above two feature it is clear that most of the cases was related with the theft and lost property. To prevent happening such cases in the future police department should take cares of this thing in the most priority manner. They should take the appropriate action to avoid such things

3. Analysis for the : department_district - district in which the police department is located.

Department District	Percentage Count(%)
SOUTHERN	18.900997
NORTHERN	13.355482
MISSION	12.958804
CENTRAL	11.738206
BAYVIEW	9.503654
INGLESIDE	7.703654
TARAVAL	7.524917
TENDERLOIN	6.605980

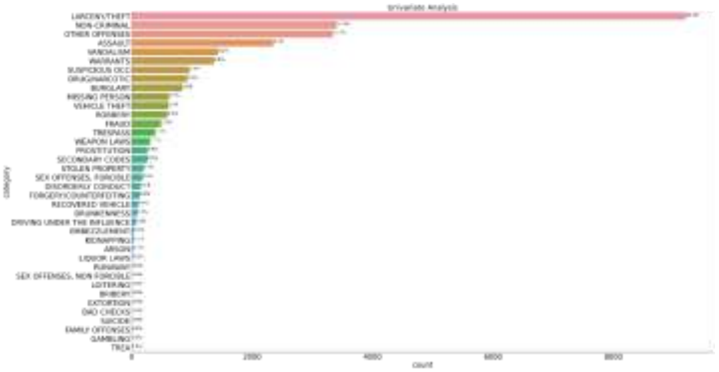


RICHMOND	5.928239
PARK	5.780066

- From the detail analysis of the above feature, we can compile that the most of the criminal cases was happened in the southern district which was around 18%. Below that northern district (13%) and mission district(12%).
- Hence police department should take necessary actions and preventive measures related with the crime cases in the respective districts to keep the city safe. Tenderloin is the district which has lowest crime rate which is around 6%. Police department should make plans and appropriate actions for the top five districts to keep the city safe.

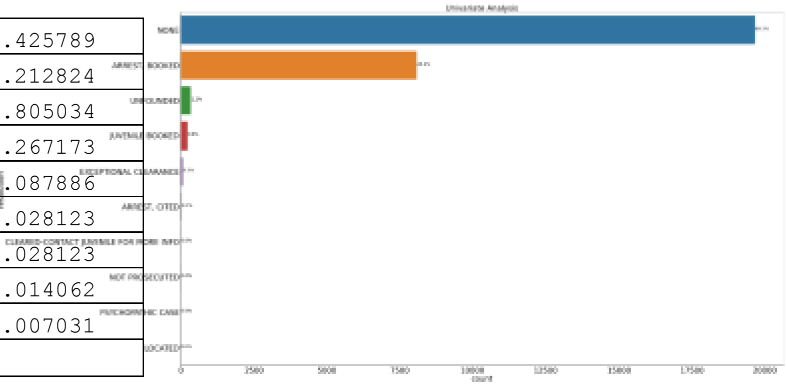
4. Analysis of category of incident reported for southern district:

Category	Percentage Count(%)
LARCENY/THEFT	32.356043
NON-CRIMINAL	11.959502
OTHER OFFENSES	11.727484
ASSAULT	8.268298
VANDALISM	5.058708
WARRANTS	4.805597
SUSPICIOUS OCC	3.381846



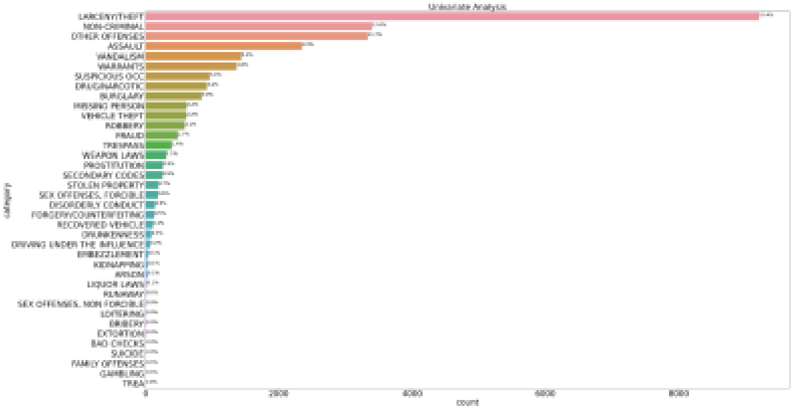
5. Analysis of resolution in southern district:

ARREST, BOOKED	28.425789
UNFOUNDED	1.212824
JUVENILE BOOKED	0.805034
EXCEPTIONAL CLEARANCE	0.267173
ARREST, CITED	0.087886
CLEARED-CONTACT JUVENILE FOR MORE INFO	0.028123
NOT PROSECUTED	0.028123
PSYCHOPATHIC CASE	0.014062
LOCATED	0.007031



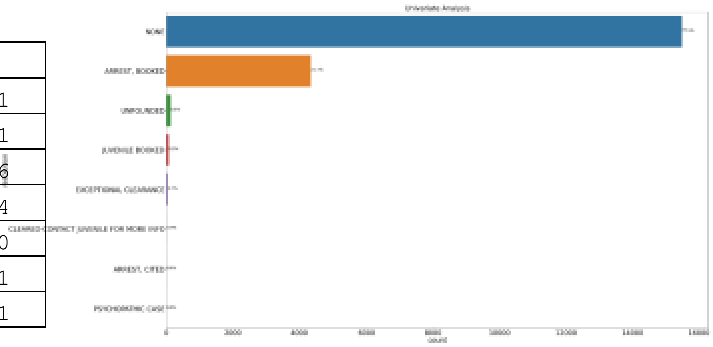
6. Analysis of crime category in southern district:

LARCENY/THEFT	32.356043
NON-CRIMINAL	11.959502
OTHER OFFENSES	11.727484
ASSAULT	8.268298
VANDALISM	5.058708
WARRANTS	4.805597
SUSPICIOUS OCC	3.381846



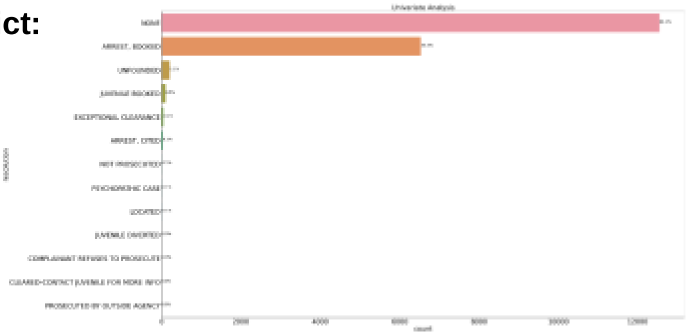
7. Analysis of resolution in southern district:

NONE	77.059701
ARREST, BOOKED	21.651741
UNFOUNDED	0.636816
JUVENILE BOOKED	0.343284
EXCEPTIONAL CLEARANCE	0.213930
CLEARED-CONTACT JUVENILE FOR MORE INFO	0.049751
ARREST, CITED	0.029851



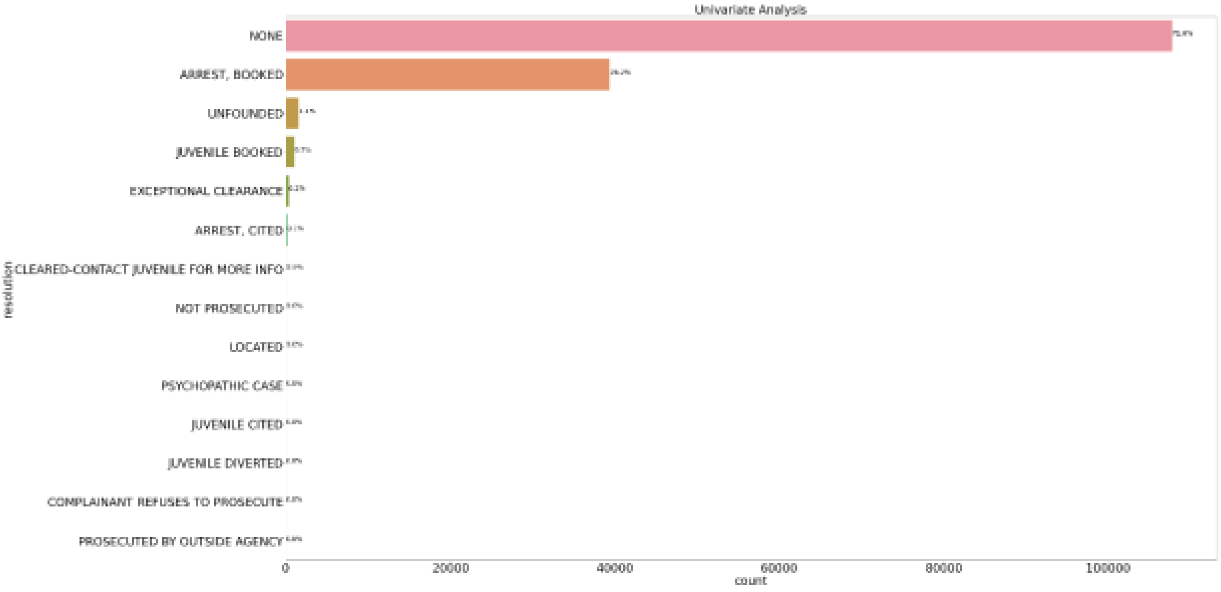
8. Analysis of resolution in MISSION district:

NONE	64.369584
ARREST, BOOKED	33.487156
UNFOUNDED	1.051120
JUVENILE BOOKED	0.517869
EXCEPTIONAL CLEARANCE	0.235861
ARREST, CITED	0.143568



9. AnalysIs of the RESOLUTION

Resolution	Percentage Count(%)
NONE	**71.614618**
ARREST, BOOKED	**26.190033**
UNFOUNDED	**1.068439**
JUVENILE BOOKED	**0.701661**
EXCEPTIONAL CLEARANCE	**0.246512**
ARREST, CITED	0.095681
CLEARED-CONTACT JUVENILE FOR MORE INFO	0.038538
NOT PROSECUTED	0.014618
LOCATED	0.013289
PSYCHOPATHIC CASE	0.011296
JUVENILE CITED	0.001993
JUVENILE DIVERTED	0.001329
COMPLAINANT REFUSES TO PROSECUTE	0.001329
PROSECUTED BY OUTSIDE AGENCY	0.000664

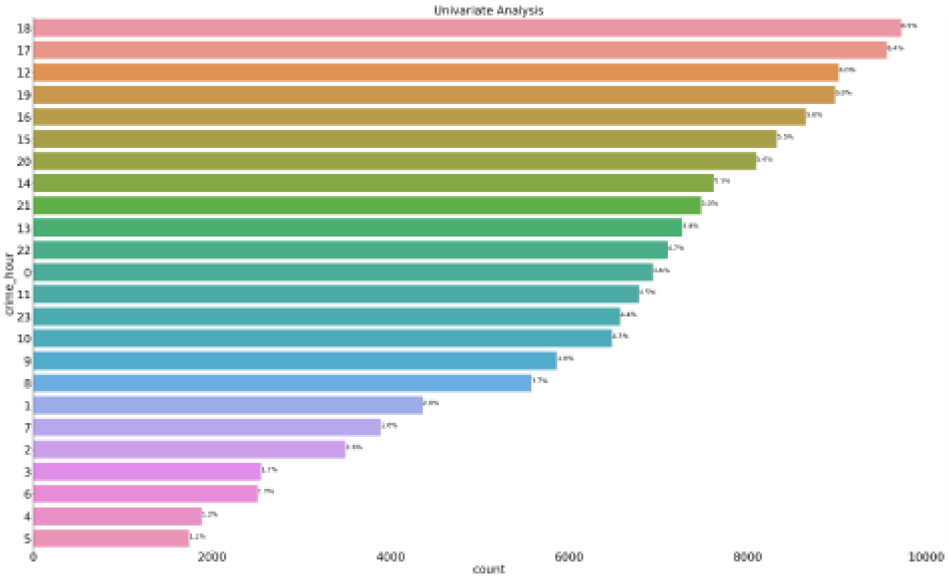


- From the above bar plot and the chart, it is clear that most of the criminal cases are not resolved. The percentage count for this is around 71%. Below that the percentage count for arrested and booked is 26%.
- Hence from the above analysis it is clear that most of the cases becomes unresolved. Hence these things should be taken into the consideration by the police department to avoid the criminal cases in the upcoming future.

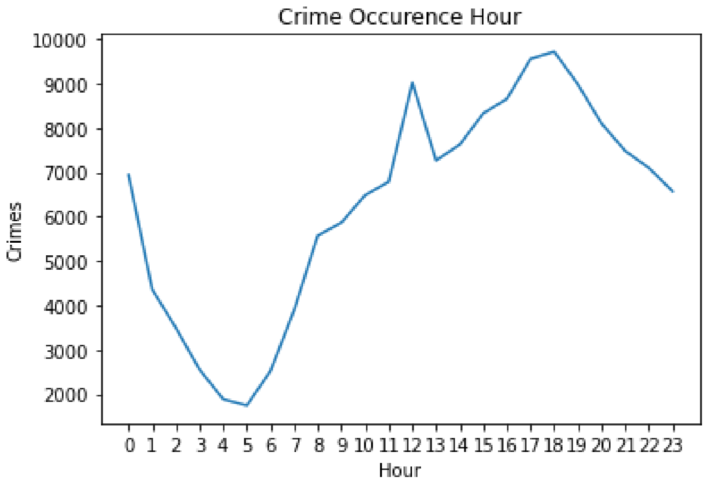
9. Analysis for time period of criminal activities

Lets analyze the time for the criminal cases

TIME	% Cases
18	6.457143
17	6.351495
12	5.994020
19	5.967442
16	5.751495
15	5.534219
20	5.380731
14	5.063787

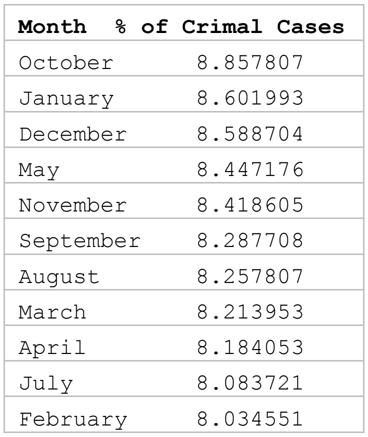


- From the above analysis we can interpret that that most of the criminal activities are happening between the time periods of 18-19 o clock. The percentage count for these activities is around 6%. Also there are activities between the time period of the 17-18 o clock.
- Hence to avoid the criminal cases in the future, police department should focus more in this time periods to control the criminal activities. Hence by doing the surveillance and patrolling in this time period can cause the decrease in the criminal activities and it will help to keep the city clean.

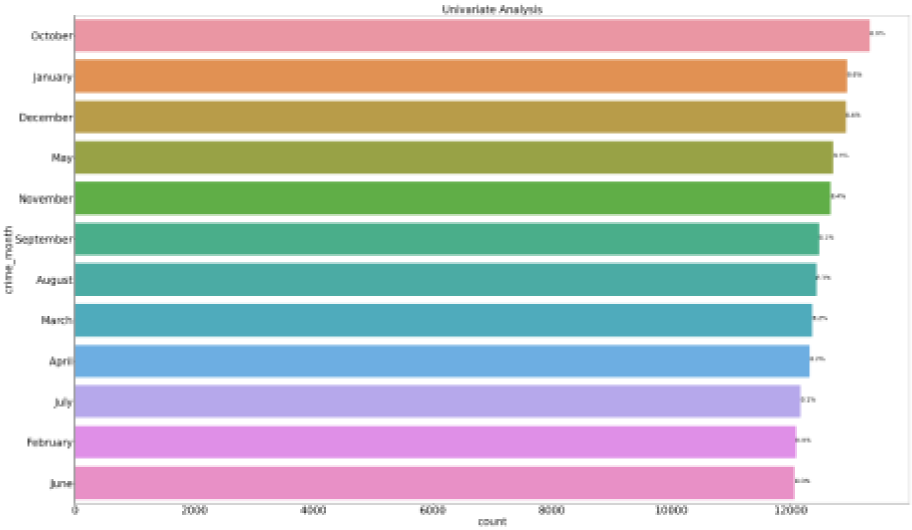


- **#Crime occurrence by hour**
- Here is the graph showing the analysis of crime cases happened with respect to time. We can clearly observe that the most of the criminal cases happened between the time intervals of 18 to 19 o clock. Also the peak time interval for the criminal activities in from 16 to 20 o clock.
- Hence to avoid such criminal activities in the upcoming future and to keep city safe, the police department should focus more precisely on the criminal activities during these hrs. Police department can increase their surveillance and patrolling in these peak hrs. so that criminal cases and activities can be controlled easily.

10. Analysis for month of criminal activities



Month	% of Crimal Cases
October	8.857807
January	8.601993
December	8.588704
May	8.447176
November	8.418605
September	8.287708
August	8.257807
March	8.213953
April	8.184053
July	8.083721
February	8.034551

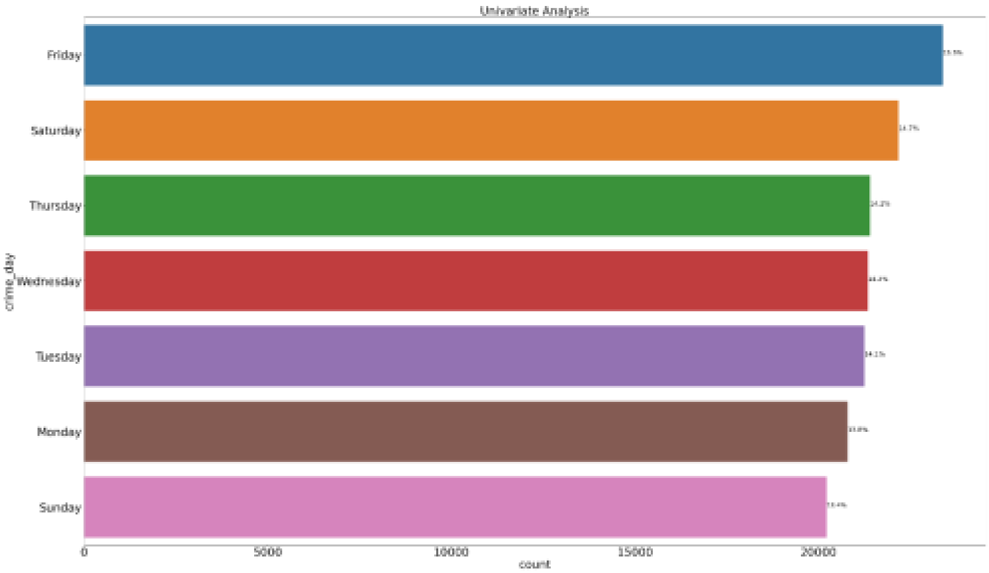


June	8.023920

- If we analyze the criminal cases happened according to the month, then we can conclude that there is not any pattern of happening of the criminal cases, the criminal cases happened through the year in all the months.

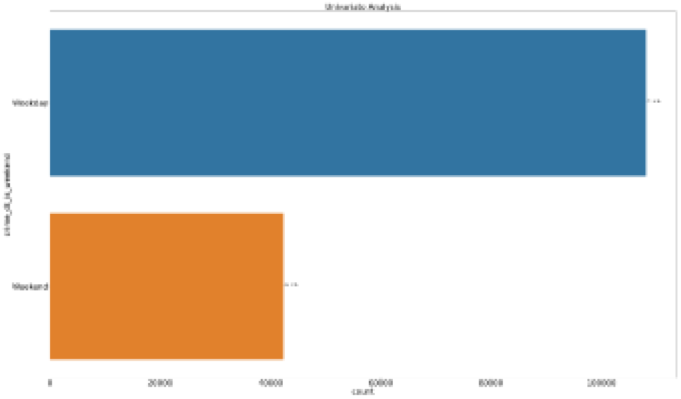
11. Analysis for Day of criminal activities

Day	Percentage Count
Friday	15.528904
Saturday	14.732226
Thursday	14.215947
Wednesday	14.174086
Tuesday	14.114286
Monday	13.809302
Sunday	13.425249



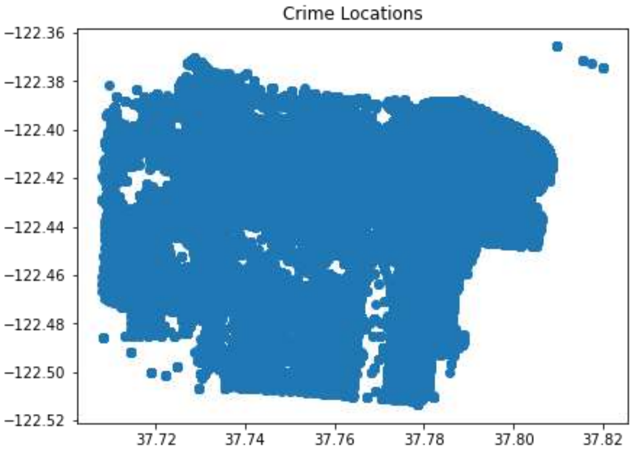
- The conclusion of this analysis is that the most of criminal cases happened on the Friday, but there is not any such pattern indicating that the happening of the criminal cases on the particular days. If we observe the criminal cases on the other days, then there is no such big difference between the percentages of happening of the criminal cases.

Criminal Case	Percentage
Weekday	71.842525
Weekend	28.157475



- The above analysis indicates that the most of the criminal activities happened during the weekdays. But from this analysis we cant conclude any pattern of happening.

12. **Visualizaing the crime locations with the help of longitude and latitudes**

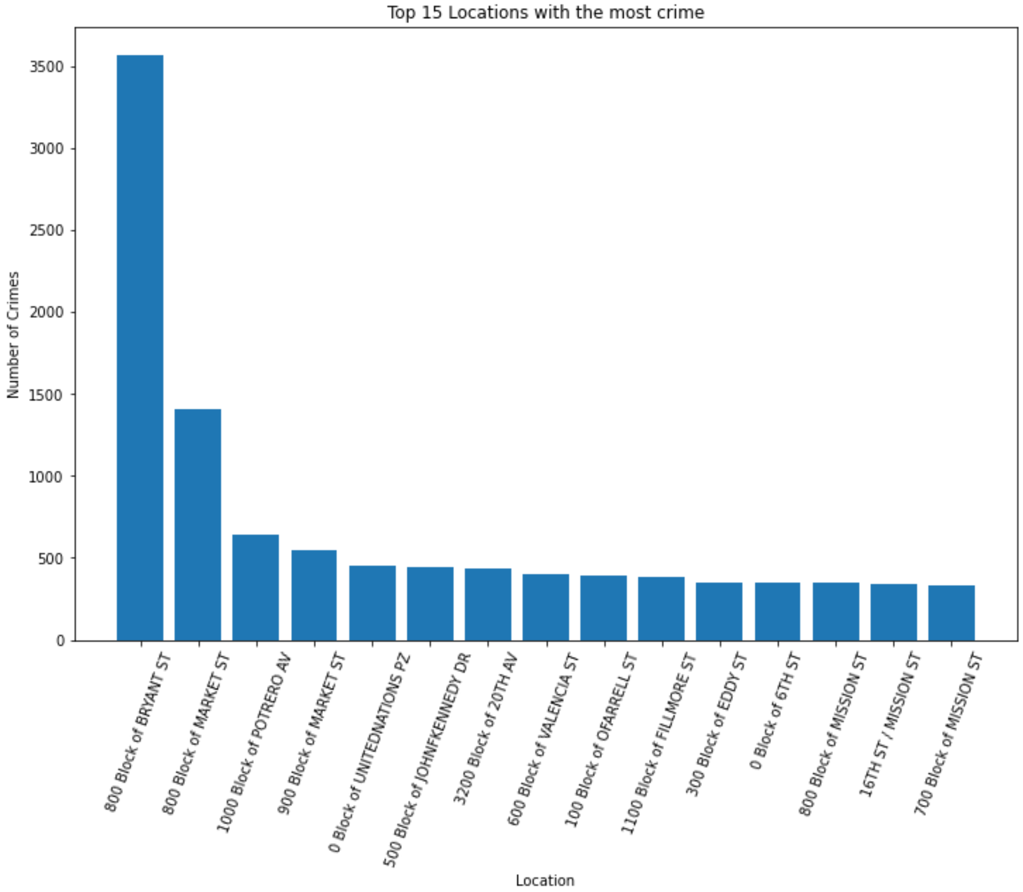


Geographical Map
San Francisco
Presidio
Pacific Ocean
San Francisco Bay

- After analyzing the genomes we can interpret that the crime locations are scattered in all the areas of city. The crime locations are not segregated and clustered for a single point. Hence we can analyze that the crime rates are scattered in all over the city. The crime rates was happened in across different and multiple street locations of the city.

- The police department should understand and analyze this information carefully to tackle with the criminal problems

13. Analysis of the top 15 most occuring crime locations



- Here is the bar chart showing the information regarding places where most of the criminal activities happened during the year of 2016. We can observe that 800 block of Bryant ST is the location where most of the cases was happened.

- Hence police department should focuses on theses location more to avoid such further activities and criminal cases. Hence by controlling the crime rates in this locations we can make city safer in upcoming future.

Statistical Analysis:

- Inferential Statistical Analysis Tests: CHI Square Test
- Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table.
- A chi-square statistic is one way to show a relationship between two categorical variables
- With the help of the chi square test we can analyze the relationship between the two categorical features. With the help of the hypothesis testing we can interpret the result and make the conclusion that whether the features and significant to each other or not.
- Here are the test results for the chi square test:

```
Chi square test between features:  crime_description  and  category
chi-square statistic:- 286468.1877800054
critical_value: 27937.236574338676
Significance level:  0.05
Degree of Freedom:  27550
p-value: 0.0
p_value < 0.05
Test Results  Reject H0,There is a relationship between 2 categorical variabl
es


Chi square test between features:  crime_description  and  department_distric
t
chi-square statistic:- 9350.421207336552
critical_value: 6714.032620009366
Significance level:  0.05
```

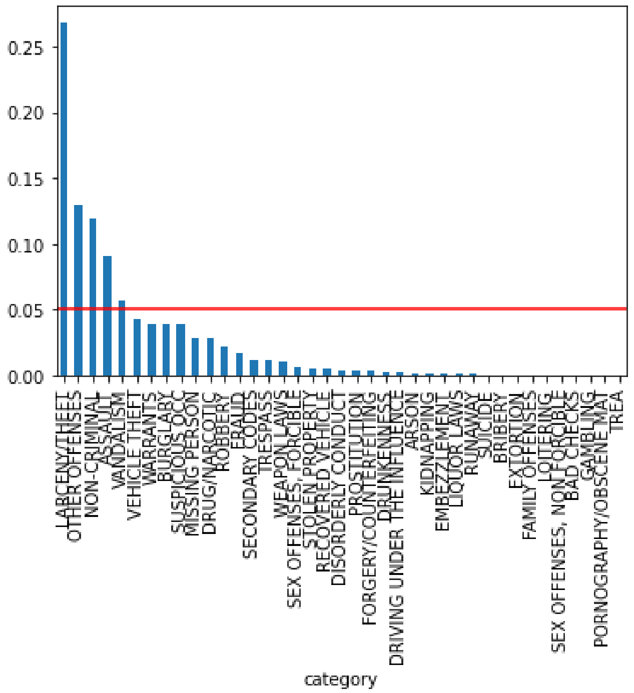
```
Degree of Freedom:  6525
p-value: 0.0
p_value < 0.05
Test Results  Reject H0,There is a relationship between 2 categorical variabl
es


Chi square test between features:  crime_description  and  resolution
chi-square statistic:- 70032.92109833732
critical_value: 9651.961975164493
Significance level:  0.05
Degree of Freedom:  9425
p-value: 0.0
p_value < 0.05
Test Results  Reject H0,There is a relationship between 2 categorical variabl
es


Chi square test between features:  crime_description  and  crime_month
chi-square statistic:- 2033.123558752684
critical_value: 8183.86480709902
Significance level:  0.05
Degree of Freedom:  7975
p-value: 1.0
p_value > 0.05
Test Results : Retain H0,There is no relationship between 2 categorical varia
bles


Chi square test between features:  crime_description  and  crime_day
chi-square statistic:- 1587.1842361911763
critical_value: 4504.550319026821
Significance level:  0.05
Degree of Freedom:  4350
p-value: 1.0
p_value > 0.05
Test Results : Retain H0,There is no relationship between 2 categorical varia
bles
```

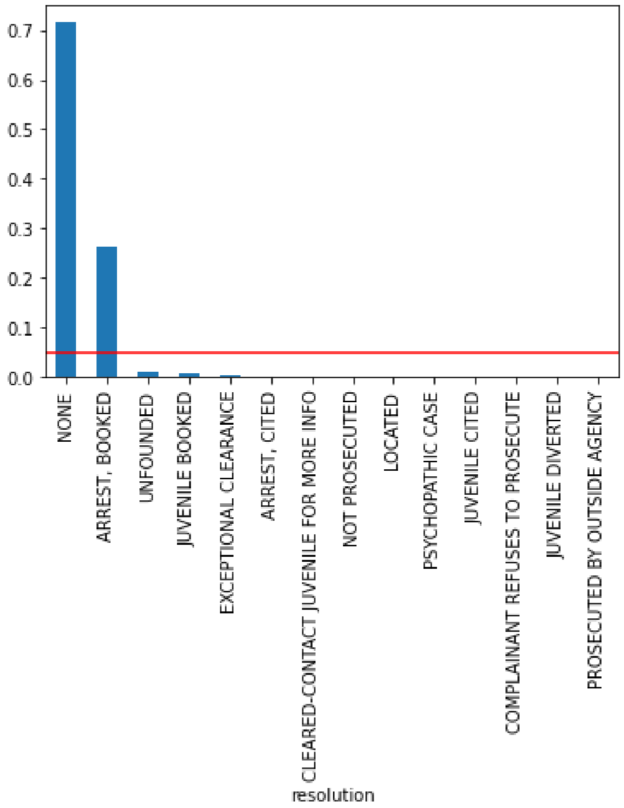
Conclusion:

- From the above test result analysis we can conclude that most of the features are relevant to each other. And there exists the hypothesis significance between in other.
- This test results and hypothesis analysis will help up to solve our objective of the analysis also it will help in the machine learning model building and interpreting the model.

Feature Engineering

- Combining rare categories:
- Rare values are categories within a categorical variable that are present only in a small percentage of the observations. There is no rule of thumb to determine how small a small percentage is, but typically, any value below 5 % can be considered rare.
- I have performed such operations on certain features to analyze the data in a better way.
- Below are the charts which shows the rare categories present in the feature.


0.25
0.20
0.15
0.10
0.05
0.00
LARCENY/THEFT
OTHER OFFENSES
NON-CRIMINAL
ASSAULT
VANDALISM
VEHICLE THEFT
WARRANTS
BURGLARY
SUSPICIOUS OCC
MISSING PERSON
DRUG/NARCOTIC
ROBBERY
FRAUD
SECONDARY CODES
TRESPASS
WEAPON LAWS
SEX OFFENSES, FORCIBLE
STOLEN PROPERTY
RECOVERED VEHICLE
DISORDERLY CONDUCT
PROSTITUTION
FORGERY/COUNTERFEITING
DRUNKENNESS
DRIVING UNDER THE INFLUENCE
ARSON
KIDNAPPING
EMBEZZLEMENT
LIQUOR LAWS
RUNAWAY
SUICIDE
BRIBERY
EXTORTION
FAMILY OFFENSES
LOITERING
SEX OFFENSES, NON FORCIBLE
BAD CHECKS
GAMBLING
PORNOGRAPHY/OBSCENE MAT
TREA
category


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
NONE
ARREST, BOOKED
UNFOUNDED
JUVENILE BOOKED
EXCEPTIONAL CLEARANCE
ARREST, CITED
CLEARED-CONTACT JUVENILE FOR MORE INFO
NOT PROSECUTED
LOCATED
PSYCHOPATHIC CASE
JUVENILE CITED
COMPLAINANT REFUSES TO PROSECUTE
JUVENILE DIVERTED
PROSECUTED BY OUTSIDE AGENCY
resolution

- Also for the prediction of the feature crime description, we can observe that there are multiple categories was present. Hence I combined the rare categories and made new class as a rare categories. Hence I converted this multiple classes into the defined classes.
- Hence the target feature in converted in to top 8 classes having the most of the data. Hence by doing this we can solve this multiclass classification problem in the better way.

Crime_Description	Percentage Count(%)
GRAND THEFT FROM LOCKED AUTO	**11.788040**
LOST PROPERTY	**3.053821**
AIDED CASE, MENTAL DISTURBED	**3.033887**
PETTY THEFT OF PROPERTY	**2.934219**
MALICIOUS MISCHIEF, VANDALISM	**2.831894**
BATTERY	2.798007
PETTY THEFT FROM LOCKED AUTO	2.653821
STOLEN AUTOMOBILE	2.394020
DRIVERS LICENSE, SUSPENDED OR REVOKED	2.243189
WARRANT ARREST	2.052492
FOUND PROPERTY	2.051827

Label Encoding:

- With the help of the label encoding technique I have converted the most of the categorical data into the numerical format.
- Hence the all the data is converted in to the numerical format and it is suitable for the machine learning algorithm.

Building Machine Learning Models

- Models:
  - SVM
  - Random Forest

  - Here are the results for the Machine Learning Model using the algorithm Support Vector Machine

Evaluation of the Models:

Model: Support Vector Classifier

Confusion Matrix:

```
[[    0     0     0     0     0     0     0  1370]
 [    0     0     0     0     0     0     0  1263]
 [    0     0     0     0     0     0     0  5347]
 [    0     0     0     0     0     0     0  1318]
 [    0     0     0     0     0     0     0  1333]
 [    0     0     0     0     0     0     0  1174]
 [    0     0     0     0     0     0     0  1364]
 [    0     0     0     0     0     0     0 31981]]
```

Accuracy Score:

0.7083277962347729

```
Precision Score:
 0.7083277962347729


F1 Score:
 0.8292644980617391
```

Model: Random Forest

```
Confusion Matrix:


[[   29     0     0   121     0     0     0  1220]
 [    0    31     0     0     0     0     0  1232]
 [    0     0  4379     0     0     0     0   968]
 [    8     0     0   319     0     0     0   991]
 [    0     0     0     0    58     0     0  1275]
 [    0     0   849     0     0     0     0   325]
 [    0     0   745     0     0     0     0   619]
 [   21    49  1975   174    31     0     0 29731]]
G
```

Accuracy Score:
0.7651605758582503

```
Precision Score:
 0.7468194962009329


Recall:
 0.8107340655214493


F1 Score:
 0.7510980863522202
```

Model Selection:

From the above model evalution we can conclude that the Random Forest model is working better for the prediction results and it is giving the best accuracy comparing with the SVM.

SUMMARY

- From the above detailed analysis of the different features I found out some of the actionable insights which should be taken care for the criminal activities and precautions.
- There are various patterns and insights in the dataset, with the help of this information police department can reduce the criminal activities and make the city much safer.

CONCLUSION

- With the help of the analytics tools and techniques, I have done the detail analysis of the each feature, with the help of the visualizations and graphs I was able to interpret the useful information.
- Fulfilled all the objectives for the analysis and the mentioned problem statement was solved using the analytics tool, techniques and machine learning algorithm’s.