Crime Reports in Louisville

Abtin Mahyar

1. Data Exploration

The crime reports which has provided by Louisville government are categorized yearly in different datasets. This analysis uses datasets which were form 2014 until 2018. After combining datasets, resulted dataset has around 400,000 reports with various types of fields which made the dataset valuable and full of hidden information.

The dataset has the following attributes:

- **"INCIDENT_NUMBER":** the number associated with either the incident. (integer)
- **"DATE_REPORTED":** the date the incident was reported to LMPD. (datetime)
- **"DATE_OCCURED":** the date the incident actually occurred. (datetime)
- **"UOR_DESC":** Uniform Offense Reporting code for the criminal act committed. (string)
- **"CRIME_TYPE":** the crime type category. There are 16 different types of crime in this dataset (category)
- **"NIBRS_CODE":** the code that follows the guidelines of the National Incident Based Reporting System. (integer)
- **"UCR_HIERARCHY":** hierarchy that follows the guidelines of the FBI Uniform Crime Reporting. There are two different hierarchy types in this dataset (category)
- **"ATT_COMP":** Status indicating whether the incident was an attempted crime or a completed crime. There are two types of attempt completion in this dataset (category)
- **"LMPD_DIVISION":** the LMPD division in which the incident actually occurred. There are 9 different types of division in this dataset (category)
- **"LMPD_BEAT":** the LMPD beat in which the incident actually occurred. (integer)
- **"PREMISE_TYPE":** the type of location in which the incident occurred (e.g. Restaurant) (category)
- **"BLOCK_ADDRESS":** the location the incident occurred. (string)
- **"CITY":** the city associated to the incident block location. (string)
- **"ZIP_CODE":** the zip code associated to the incident block location. (integer)
- **"ID":** Unique identifier for internal database. (integer)

As it can be suspected based on above attributes, some of the features are useless or illegible, which means they must be discarded during the preprocessing.

Also, dataset has some null values in some of its' features which should be managed during preprocessing. Moreover, there is a type of value in *PREMISE_TYPE* which is "OTHER / UNKNOWN" which null values in this feature should be replaced by this expression. Also, null values in *UCR_HIERARCHY* is because of "OTHER" values in *CRIME_TYPE* so these null values could be replaced by another appropriate value.

FEATURE	SUM OF NULL RECORDS
DATE_OCCURED	20
UCR_HIERARCHY	5809
ATT_COMP	867
LMPD_BEAT	960
PREMISE_TYPE	365
CITY	573
ZIP_CODE	2509

Table 1. Number of null values for each feature

2. Data Preprocessing

The preprocessing done in the "crime_cleaning" and "crime" notebooks consist of the following steps:

1. Null values are eliminated or filled with a specific value. According to **Table 1** there are seven columns which have null values. As it is discussed before, null values of feature *PREMISE_TYPE* of records which have a specific value for unknown premises, should be replaced by this value. Moreover, values in column *ID* does not have unique values therefor this column can not be useful for us and we can not use it as dataset's index so we drop it. Furthermore, each *NIBRS_CODE* is some how related to *CRIME_TYPE* so this feature dose not add any extra information to the dataset; therefore, this column should be dropped. Null values in other columns are replaced by new "UNKNOWN" value type.
2. *INCIDENT_NUMBER* dropped due to it's uselessly during our analysis.
3. New features are extracted from *DATE_REPORTED* and *DATE_OCCURED* are listed as follows:
   - **"YEAR":** the actual year that incident happened. (category)
   - **"MONTH":** the actual month that incident happened. (category)
   - **"DAY_RANGE":** the part of the day that incident happened. There are two type of day range ("PRENOON", "AFTERNOON") (category)
   - **"DIFF_TIME_MIN":** difference between date that incident reported and occurred in minutes. (integer)
4. Data types are fixed and each attribute convert to its' own data type.
5. Some *ZIP_CODE* have non logical values which were replaced by median of this column. Also, binning method was applied to this column due to its' nonparametric distribution and new feature *ZIP_CODE_CAT* was created.
6. Some incidents which were before 2014 were eliminated.
7. Because *DIFF_TIME_MIN* has skewed distribution, values of this column have been transformed to a parametric distribution by using log method.

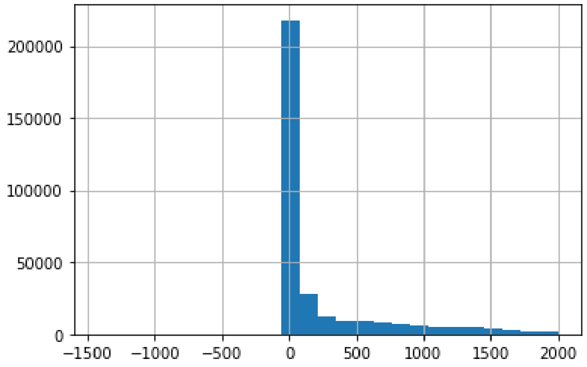

Figure 1. DIFF_TIME_MIN histogram before log transformation

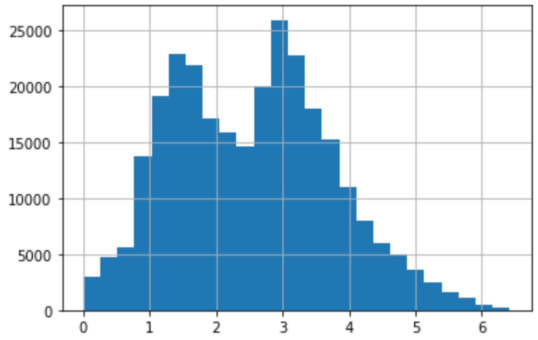

Figure 2. DIFF_TIME_MIN histogram after log transformation

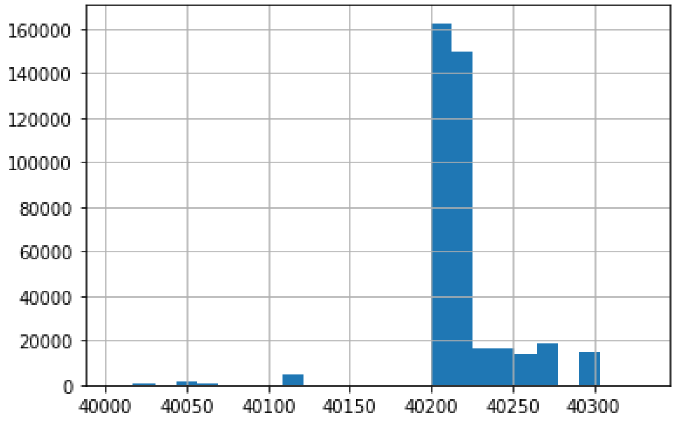

Figure 3. Zip Codes histogram

ZIP_CODE feature divided to four different categories due to its' distribution with the following presentation and stored in new attribute *ZIP_CODE_CAT*.

LABEL	INTERVAL
< 40200	[0, 40200)
[40200, 40225]	[40200, 40225)
[40225, 40300]	[40225, 40300)
> 40300	[40300, 100,000)

Table 2. Categories of ZIP_CODE_CAT

As it can be seen from the plot, most crimes are occurred in zip codes between 40200 and 40300.

3. Exploratory Visualization

3.1. Crime Type

Fig. 4 The following plot demonstrates the number of total incidents form each crime category in the dataset. As it can be seen, the number of “THEFT/LARCENY”, “DRUGS/ ALCOHOL VIOLATIONS”, and “ASSAULT” are much higher than other types of crime. Also, “ARSON”, “DUI”, and “HOMICIDE” are less likely to happen.

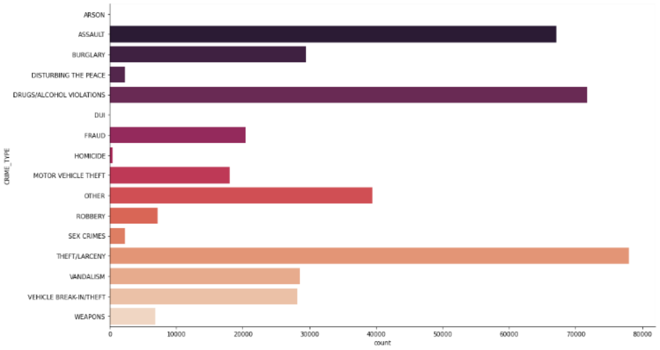

Figure 4. Distribution of incidents over crime types

Fig. 5 A plot showing how incidents are distributed based on their crime type and UCR hierarchy. According to the bar chart, the distributions of incidents are different from each other which means that these two categorical variables are dependent. Mostly, crimes that are some how related to theft are consider as PART I crimes and others corresponds to PART II.

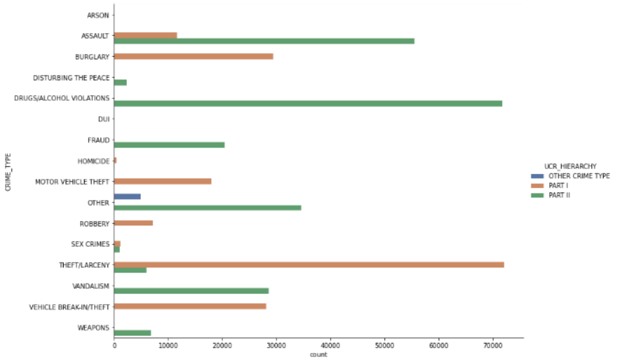

Figure 5. Distribution of incidents over crime types and UCR hierarchy

Fig. 6 The following plot shows how incidents are distributed in two categorical variables (crime type and attempt completion). According to the bar chart, the distributions of incidents are different from each other which means that these two categorical variables are dependent. As it can be seen from the bar chart, the majority of attempts have been completed. Incomplete attempts are usually form crime types that are similar to theft.

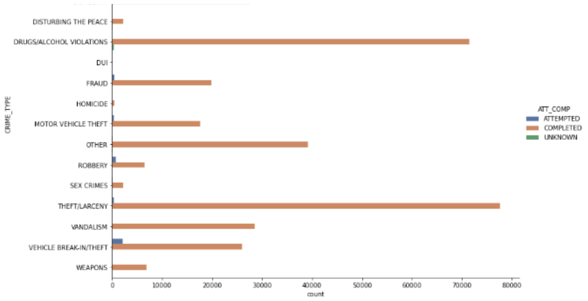

Figure 6. Distribution of incidents over crime type and ATT_COMP

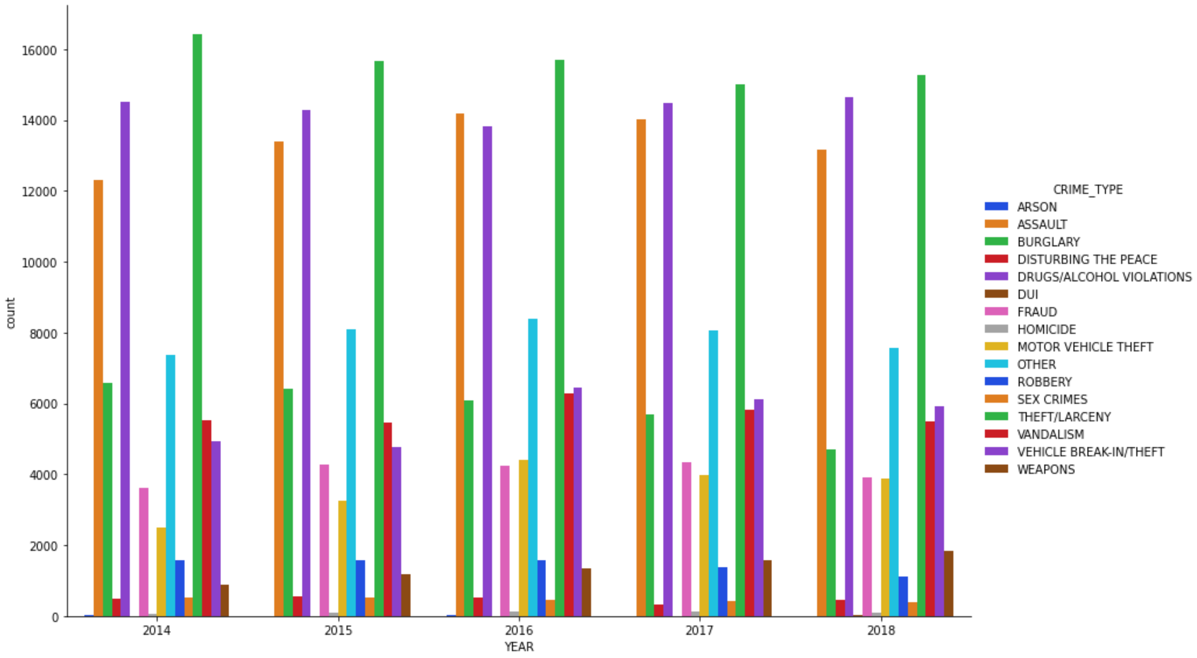

Figure 7. Distribution of incidents over crime type and year

Fig. 7 A plot showing how incidents are distributed based on their crime type and year. According to the bar chart, the distributions of incidents are kind of similar to each other which means that these two categorical variables are independent; however, this is a guess and it should be checked using appropriate statistical test.

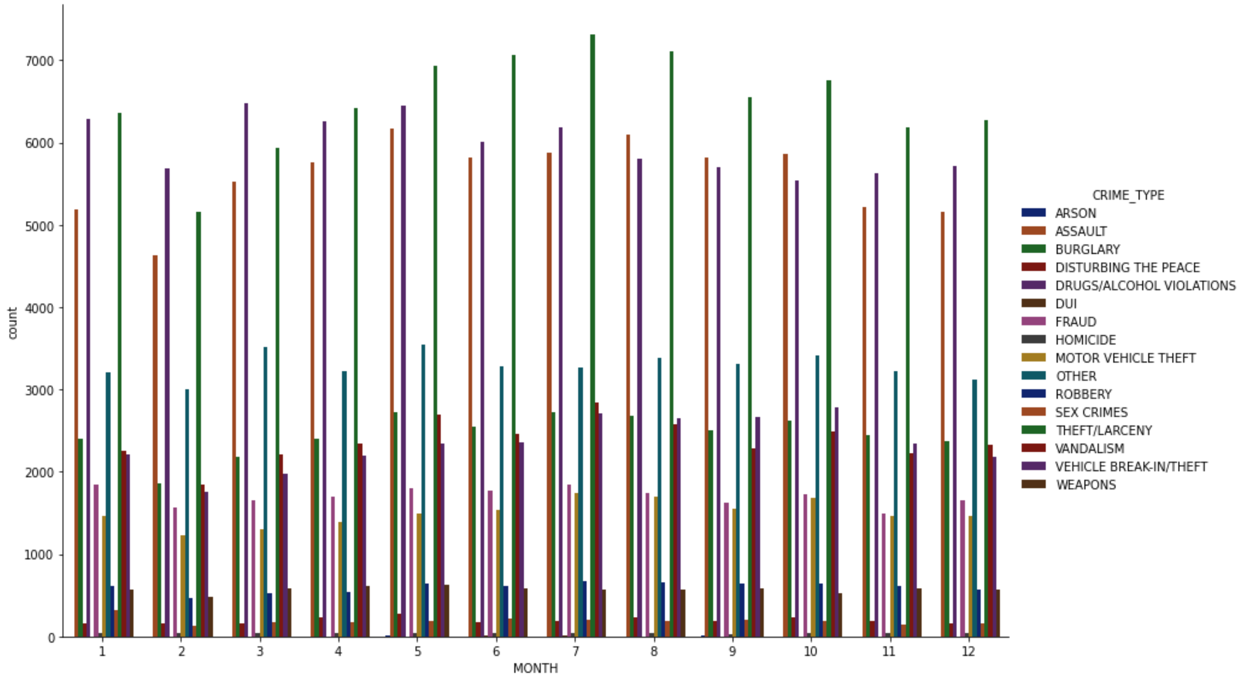

Figure 8. Distribution of incidents over crime type and month

Fig. 8 The plot above shows that how incidents are distributed based on their crime type and month. According to the bar chart, the distributions of incidents are kind of similar to each other which means that these two categorical variables are independent; however, this is a guess and it should be checked using appropriate statistical test.

Fig. 9 The following plot shows how incidents are distributed in two categorical variables (crime type and day range). According to the bar chart, the distributions of incidents are different from each other which means that these two categorical variables are dependent.

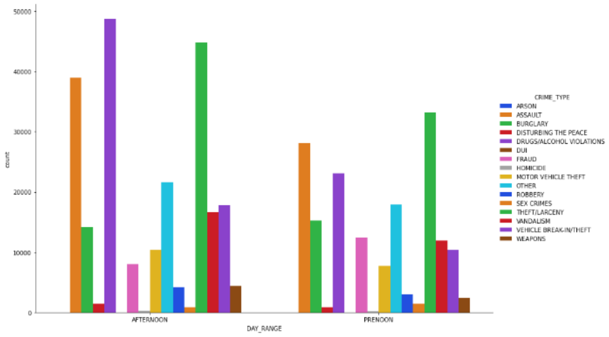

Figure 9. Distribution of incidents over crime type and day range

3.2. UCR Hierarchy

Fig. 10 A plot which demonstrates the number of total incidents form each UCR hierarchy category in the dataset. As it can be seen from the pie chart, the majority of incidents are categorized as PART II in this distribution. In the second place, incidents which consider as PART I have formed 42% of total incidents. Only one percent of incidents are categorized as "OTHER CRIME TYPE" which were null values and they could be ignored in our further analysis.

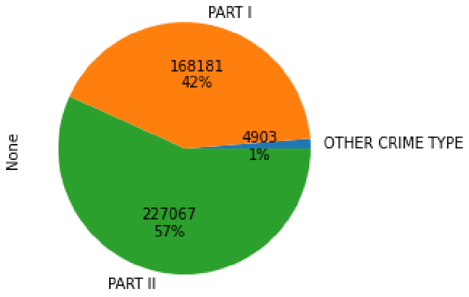

Figure 10. Pie chart of incidents based on UCR hierarchy

Fig. 11 The following plot shows how incidents are distributed in two categorical variables (UCR hierarchy and attempt completion). According to the bar chart, the distributions of incidents are different from each other which means that these two categorical variables are dependent. The majority of incomplete attempts are form PART I and all of the unknown (null) attempts are from PART II.

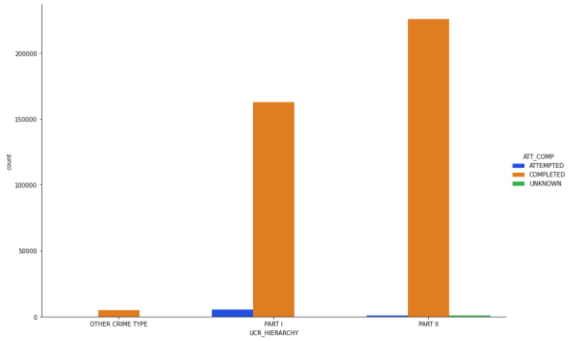

Figure 11. Distribution of incidents over UCR hierarchy and attempt completion

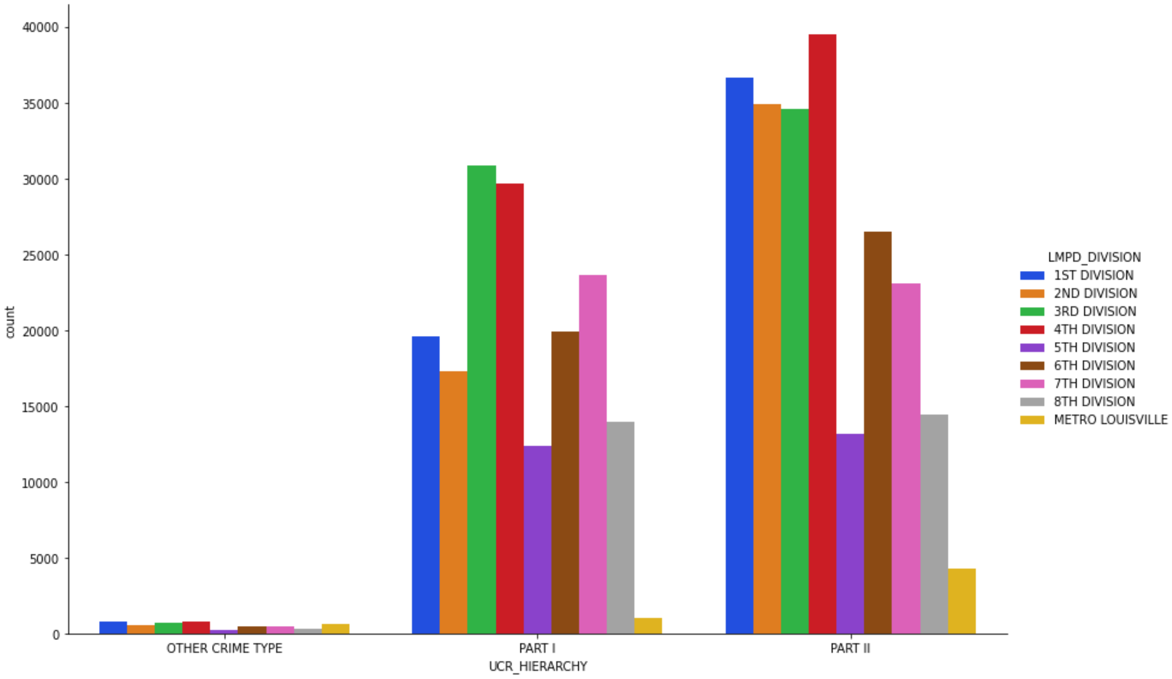

Figure 12. Distribution of incidents over LMPD division and UCR hierarchy

Fig. 12 The above plot demonstrates how incidents are distributed in two categorical variables (UCR hierarchy and LMPD division). According to the bar chart, the distributions of incidents are different from each other which means that these two categorical variables are dependent.

Fig. 13 A plot showing how incidents are distributed based on their UCR hierarchy and year. According to the bar chart, the distributions of incidents are kind of similar to each other which means that these two categorical variables are independent; however, this is a guess and it should be checked using appropriate statistical test.

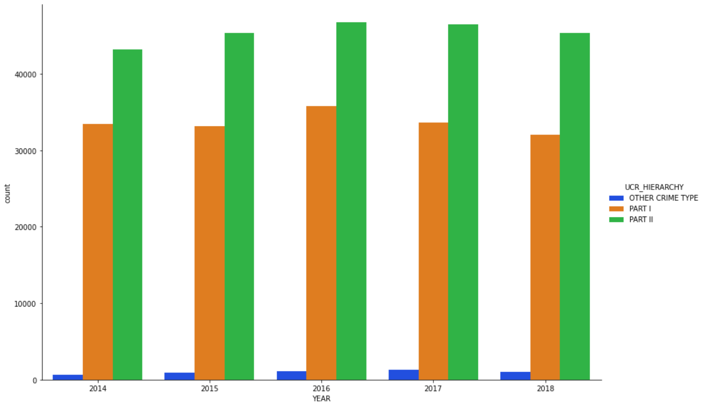

Figure 13. Distribution of incidents over UCR hierarchy and year

3.3. Attempt Completion

Fig. 14 The following plot demonstrates the number of total incidents form each attempt completion category in the dataset. The majority of attempts have been completed. Also, the minority of incidents are categorized as UNKNOWN (null) attempts, so they could be omitted in future analysis.

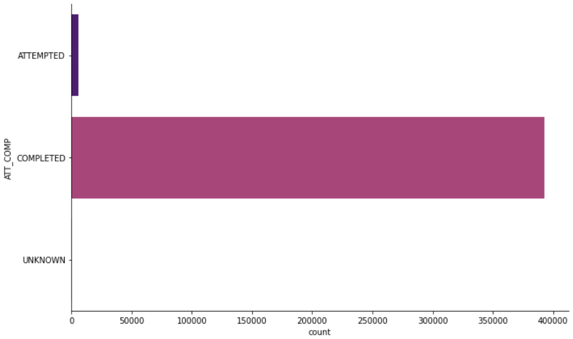

Figure 14. Bar chart of incidents based on attempt completion

Fig. 15 A plot showing how incidents are distributed based on their attempt completion and LMPD division. According to the bar chart, the distributions of incidents are different from each other which means that these two categorical variables are dependent. As it can be seen, 4^{th} and 3^{rd} division have the most completed attempts whilst METRO LOUISVILLE and 5^{th} division have least completed attempts. Also, METRO LOUISVILLE do not have any UNKNOWN or incomplete attempts.

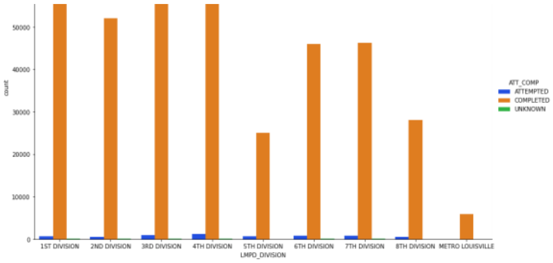

Figure 15. Distribution of incidents over attempt completion and LMPD division

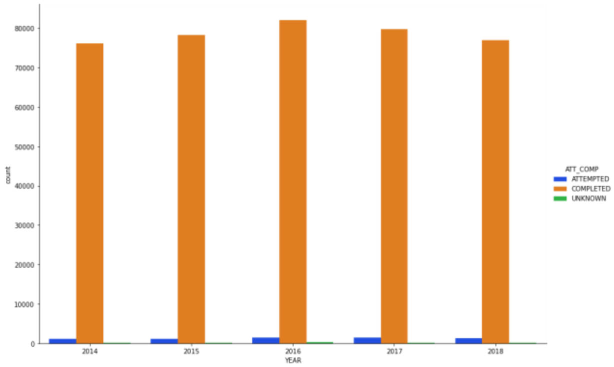

Figure 16. Distribution of incidents over attempt completion and year

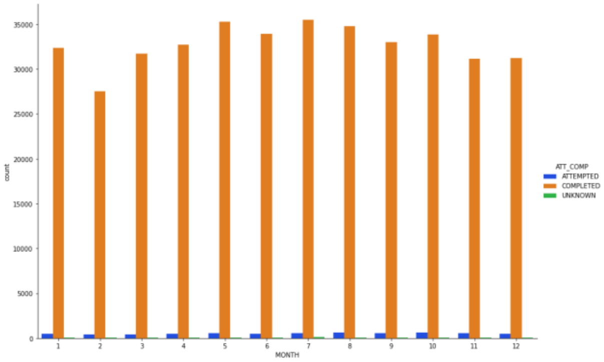

Figure 17. Distribution of incidents over attempt completion and month

As it can be seen form above plots, rate of attempt completion was approximately stable during months and years in this five years period.

3.4. LMPD Division

Fig. 18 The following plot demonstrates the number of total incidents form each LMPD division category in the dataset. As it can be seen, 4th division has the most frequency in the dataset whilst this number is much less in METRO LOUISVILLE.

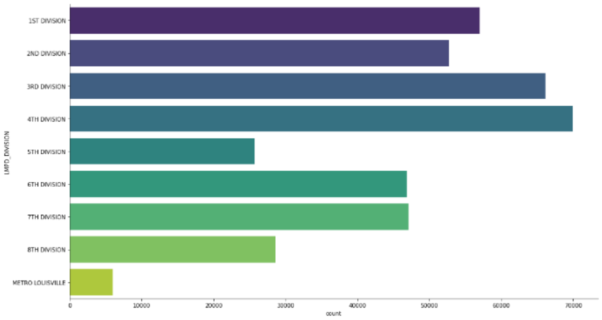

Figure 18. Bar chart of incidents based on LMPD division

3.5. Year

Fig. 19 A plot which demonstrates the number of total incidents form each year in the dataset. As it can be seen from the bar chart, number total incidents for each year are located in the same range. But, in 2016, this number reaches to its' peak.

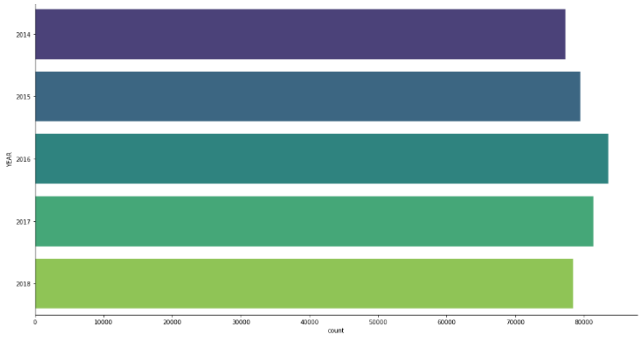

Figure 19. Bar chart of incidents based on year

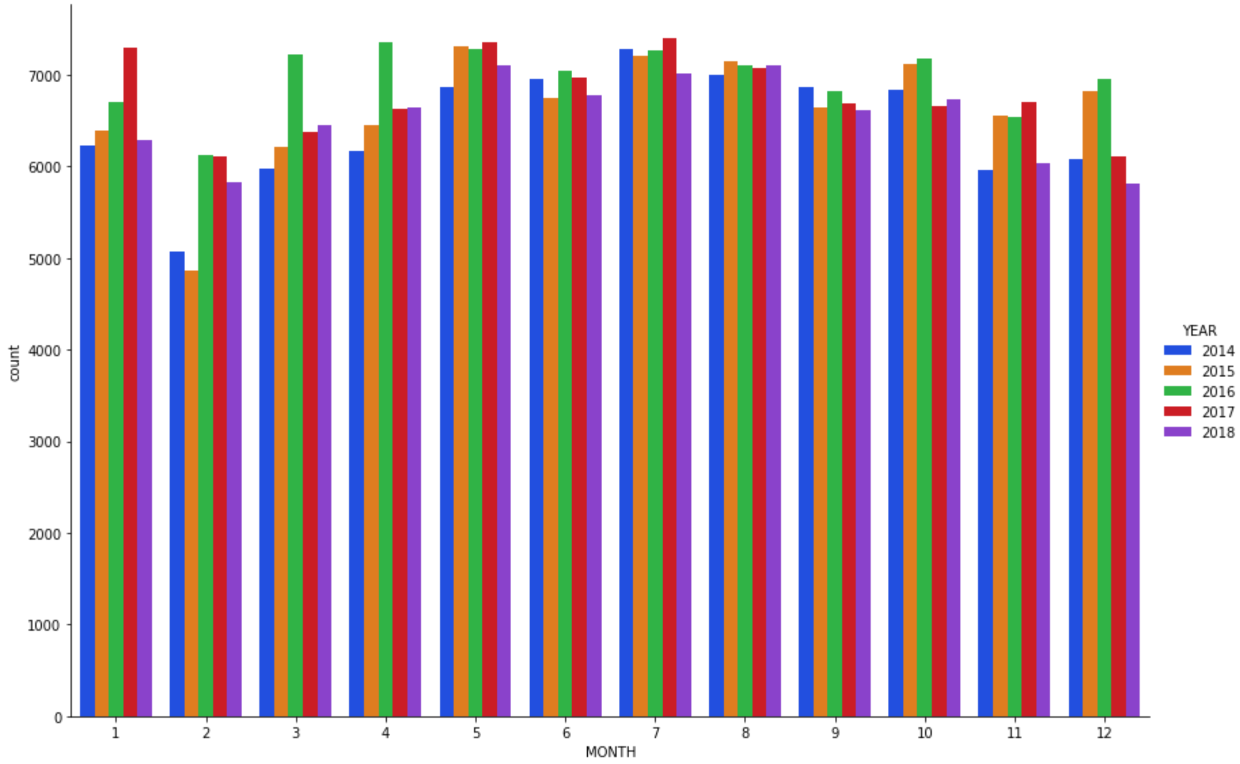

Figure 20. Distribution of incidents over year and month

Fig. 20 The plot above shows that how incidents are distributed based on their year and month. We can't get reliable results from just looking at the chart and we should run an appropriate statistical test to get valuable information. Mostly, in each month, number of incidents has increased during 2014 to 2017 and then it dropped in 2018.

Fig. 21 The following plot shows that how incidents are distributed based on their year and *DAY_RANGE*. According to the bar chart, the distributions of incidents are kind of similar to each other which means that these two categorical variables are independent; however, this is a guess and it should be checked using appropriate statistical test.

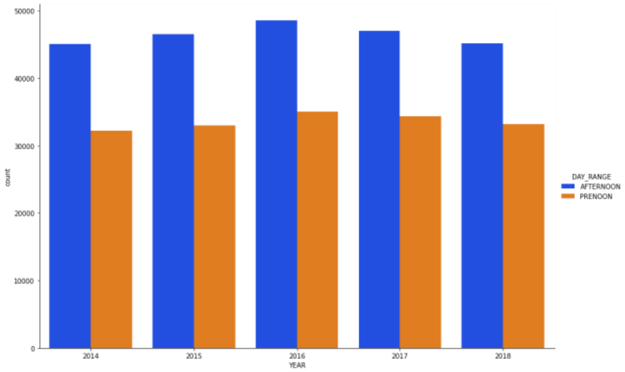

Figure 21. Distribution of incidents over year and day range

3.6. Month

Fig. 22 A plot which demonstrates the number of total incidents form each month in the dataset. As it can be seen from the bar chart, number total incidents for each month are located in the same range. But, majority of incidents are occurred at the middle months. Also, at 5th and 7th months, incidents are most likely to happen whilst 2nd month has the least total number of incidents.

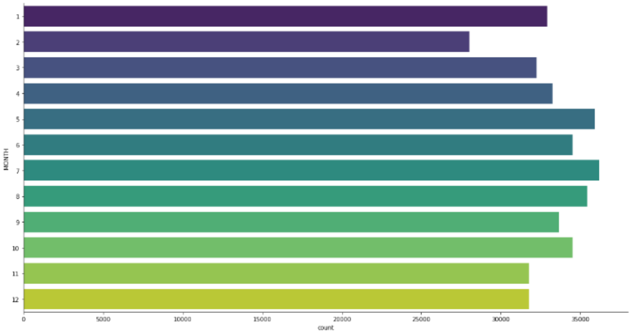

Figure 22. Bar chart of incidents based on month

3.7. Top UOR

Top UOR are defined as the descriptions with highest number of repetitions in the dataset.

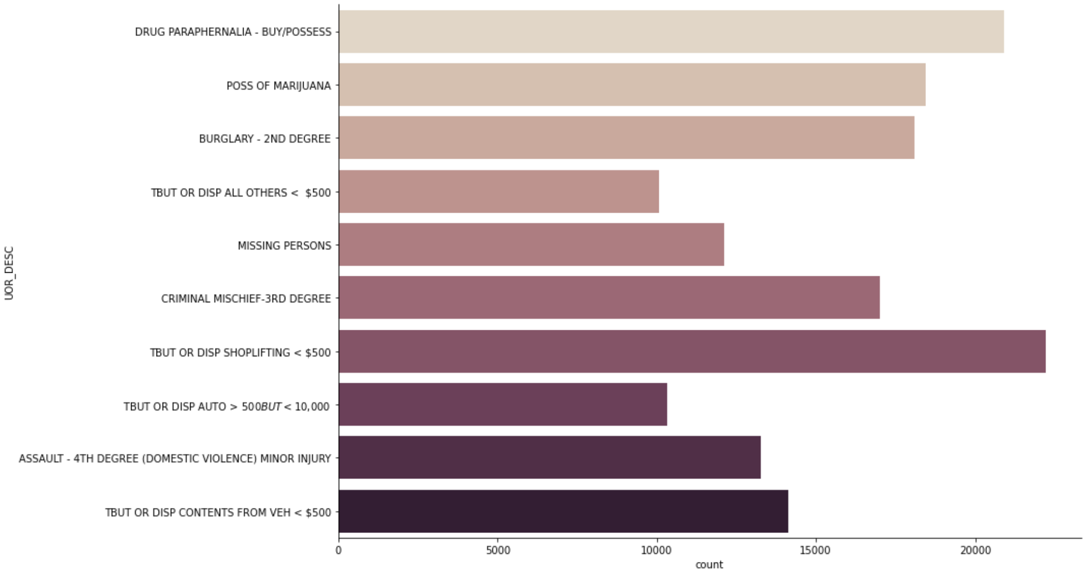

Figure 23. Bar chart of incidents in top UOR

3.8. Top Block Addresses

Top block addresses are defined as the blocks with highest number of repetitions in the dataset.

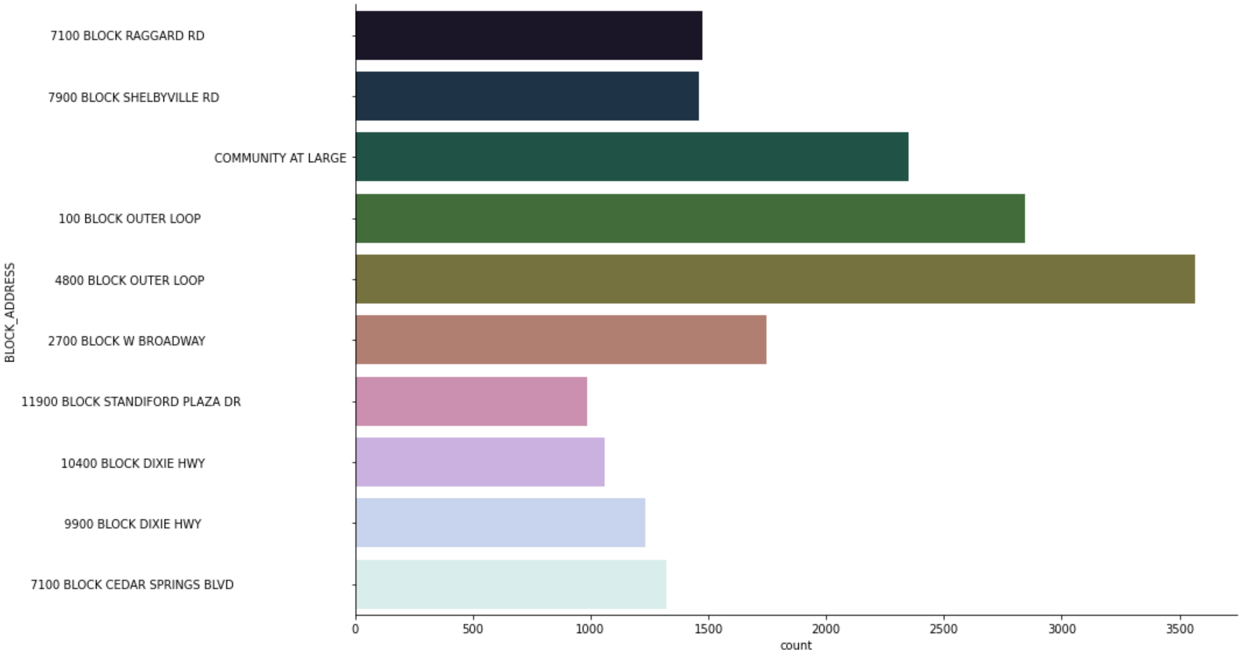

Figure 24. Bar chart of incidents in top block addresses

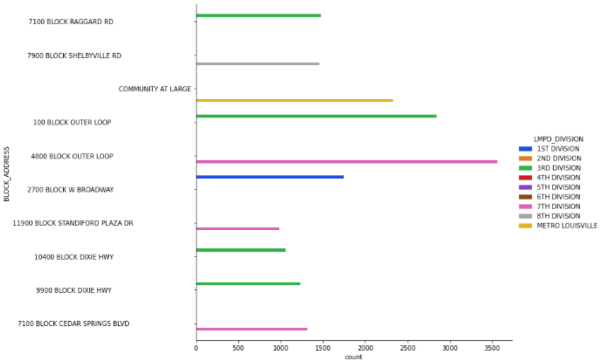

Figure 25. Distribution of incidents in top block addresses over LMPD division

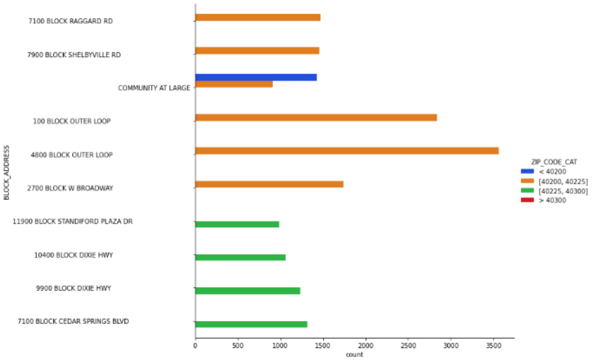

Figure 26. Distribution of incidents in top block addresses over zip code categories

Fig. 27 The below bar chart illustrates how incidents are distributed between top block addresses and year. As it can be seen from the chart, rate of crime in some blocks are decreasing like "2700 BLOCK W BROADWAY", "11900 BLOCK STANDIFORD PLAZA DR", and "COMMUNITY AT LARGE"; whilst, rate of crime in some blocks are increasing during these years such as: "4800 BLOCK OUTER LOOP", and "100 BLOCK OUTER LOOP".

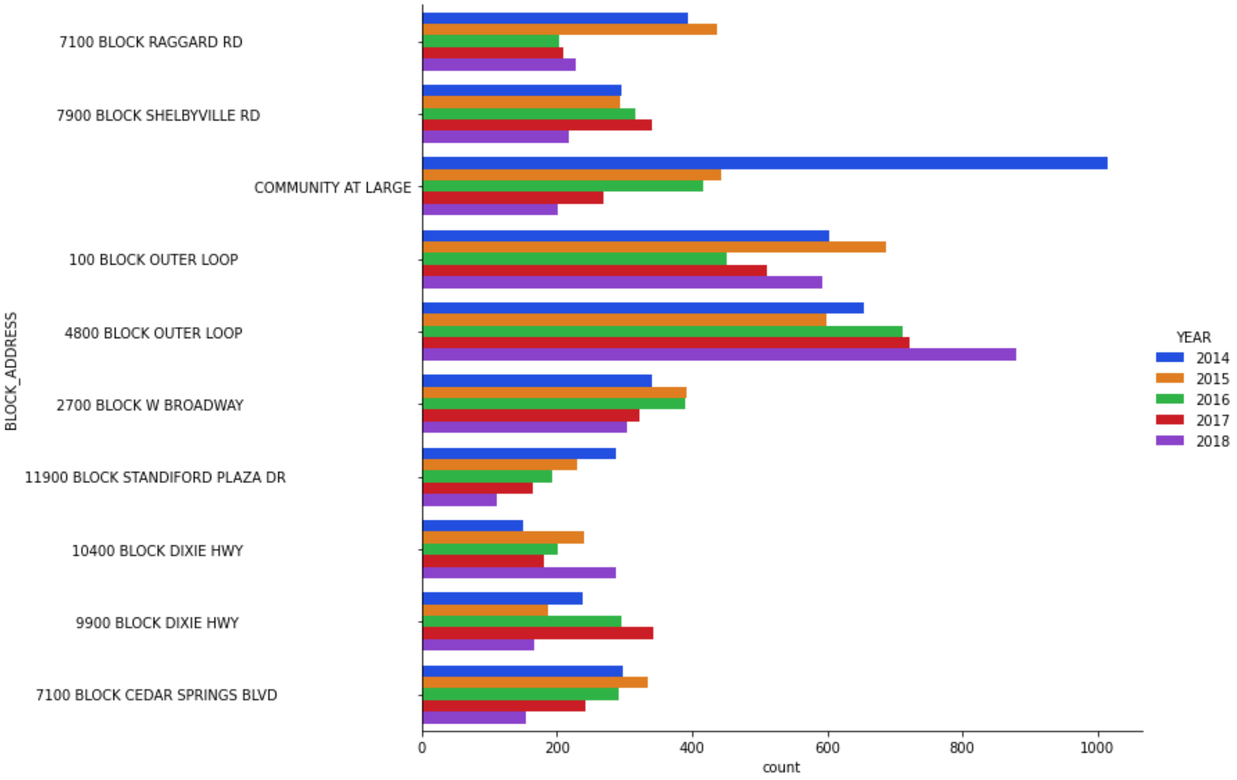

Figure 27. Distribution of incidents in top block addresses over years

3.9. Top Premise Types

Top premise types are defined as the premises with highest number of repetitions in the dataset.

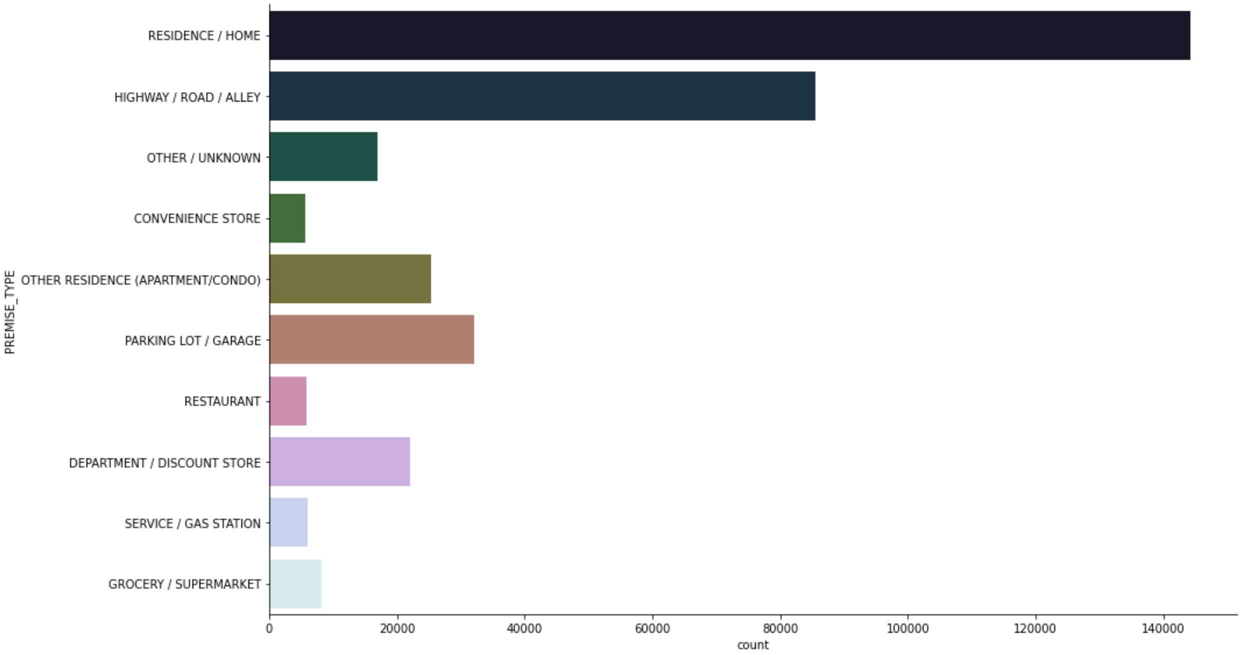

Figure 28. Bar chart of incidents in top premises

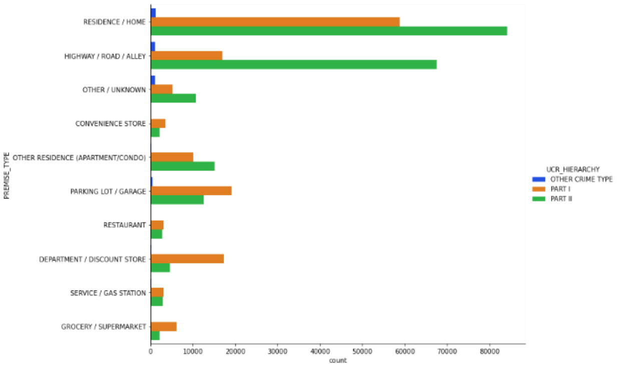

Figure 29. Distribution of incidents in top premises over UCR hierarchy

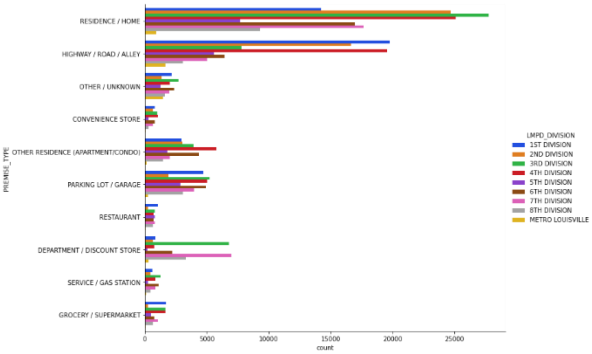

Figure 30. Distribution of incidents in top premises over LMPD division

3.10. Difference Time of report and occurrence of incidents

Fig. 31 The following plot shows how incidents are distributed over crime type and *DIFF_TIME_MIN_LOG.* According to the plot, the distributions of incidents are different from each other which means that these two categorical variables are dependent. As it can be seen from the bar chart, some crimes take a long time to report such as sex crimes, and fraud; whilst, some of them are reported faster like weapons, robbery, homicide, dui, and assault. Also, some kind of crimes have a wide range of time to report such as sex crimes, theft, and fraud.

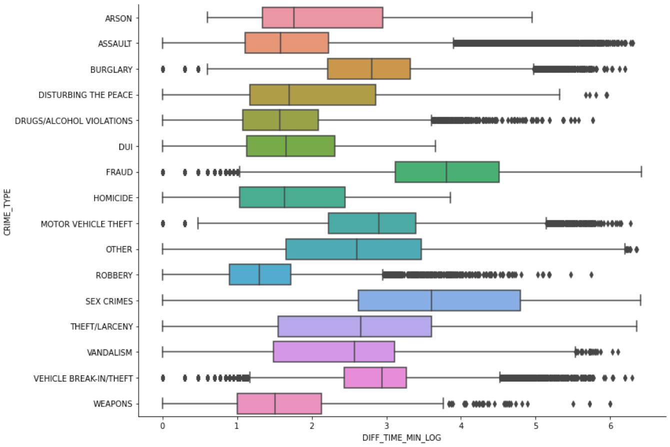

Figure 31. Distribution of incidents over crime type and DIFF_TIME_MIN_LOG

Fig. 32 A plot showing how incidents are distributed based on *DIFF_TIME_MIN_LOG* and year. According to the plot, the distributions of incidents are kind of similar to each other which means that these two categorical variables are independent; however, this is a guess and it should be checked using appropriate statistical test. Moreover, over time, range of difference time to report is become more narrower than before.

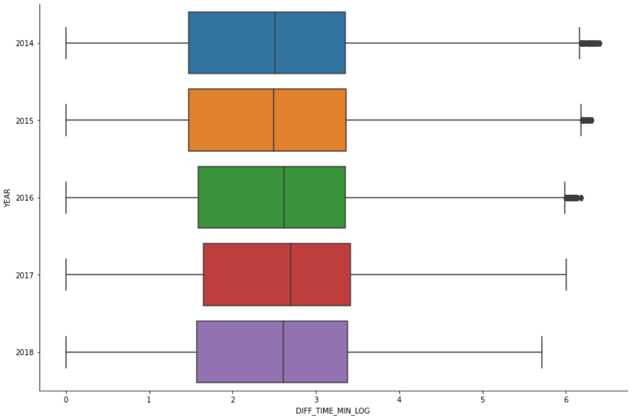

Figure 32. Distribution of incidents over year and DIFF_TIME_MIN_LOG

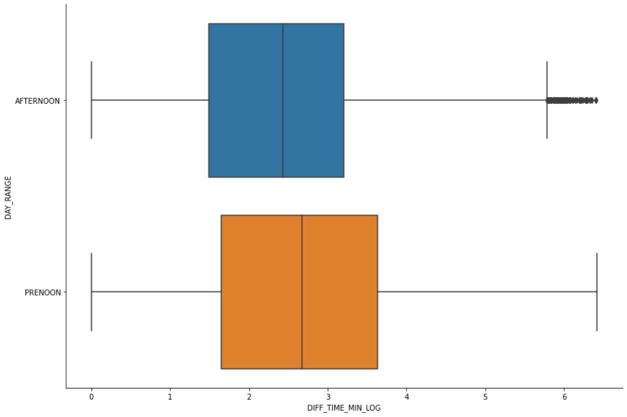

Figure 33. Distribution of incidents over day range and DIFF_TIME_MIN_LOG

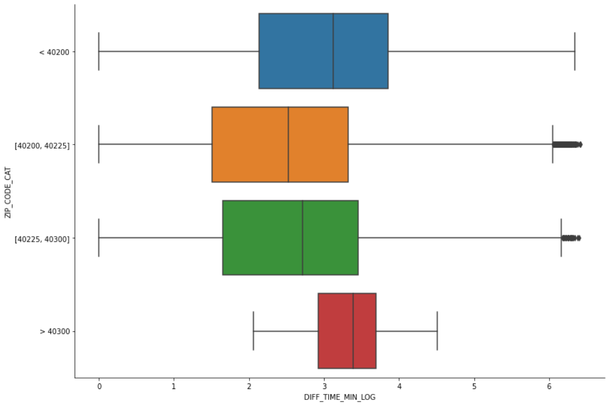

Figure 34. Distribution of incidents over ZIP_CODE_CAT and DIFF_TIME_MIN_LOG

4. Statistical Tests and Analysis

4.1. Chi2 Test

Chi-squared is a statistical hypothesis test for checking the dependency of two categorical variable. Since we made lots of hypothesis based on visualization from the last section, we should check our hypothesis of the comparisons between two categorical variables with this test.

H0 (null hypothesis test): Two categorical variable are independent.

After calculating p-value and statistic for each comparison if p-value is less than or equal to significance level (assumed to be 0.05 = 5%) then the H0 will be rejected and categorical variables are dependent. Equally, if the statistic is more than or equal to critical value, then H0 will be rejected. Results of applying chi-squared test on the dataset comes as follows:

COMPARISON	(STATISTIC, P-VALUE)	RESULT	HYPOTHESIS
CRIME VS. UCR	(297645.64, < 0.001)	Dependent (reject H0)	Dependent
CRIME VS. YEAR	(2257.07, < 0.001)	Dependent (reject H0)	Independent
CRIME VS. MONTH	(1276.46, < 0.001)	Dependent (reject H0)	Independent
CRIME VS. DAY RANGE	(8011.01, < 0.001)	Dependent (reject H0)	Dependent
UCR VS. LMPD	(7866.01, < 0.001)	Dependent (reject H0)	Dependent
UCR VS. YEAR	(147.03, < 0.001)	Dependent (reject H0)	Independent
UCR VS. MONTH	(372.78, < 0.001)	Dependent (reject H0)	Dependent
ATT_COMP VS. LMPD	(424.83, < 0.001)	Dependent (reject H0)	Dependent
ATT_COMP VS. YEAR	(82.59, < 0.001)	Dependent (reject H0)	Dependent
ATT_COMP VS. MONTH	(36.81, 0.0001)	Dependent (reject H0)	Dependent
YEAR VS. MONTH	(622.52, < 0.001)	Dependent (reject H0)	Independent
YEAR VS. DAY RANGE	(17.95, 0.001)	Dependent (reject H0)	Independent
MONTH VS. DAY RANGE	(38.42, < 0.001)	Dependent (reject H0)	Dependent

Table 3. Result and p-values of different categorical comparison using Chi-squared test

Most of the results are same as which we have guessed during our observation whilst some of them are failed.

5. Results

The most important and valuables conclusions that gathered during the processes which were discussed above, are listed as follows:

- Attempt completion has a stable rate during months and years in this dataset and majority of attempts are completed. Attempts that are not completed are mostly related to theft and considered as PART I in UCR hierarchy.
- There is no especial increasing or decreasing trend in rate of crime occurrence over this five years period.
- Incidents are most likely to occur at afternoon (between 12pm till midnight).
- Some increasing and decreasing trend in rate of crime occurrence are observed in different address blocks and zip codes in previous sections.
- Majority of incidents in the dataset are located in zip codes between 40200 and 40300.
- Difference time between its' occurrence and report is dependent on crime type, day range, and zip code. Usually, incidents that occurred at prenoon are reported later than incidents that occurred at afternoon. The range of this time is become more narrower over time in this five years period (relation between this subject and crime type is discussed on Figure 31).
- Majority of incidents are occurred at the middle months. To be specific, at 5th and 7th months, incidents are most likely to occur whilst 2nd month has the least total number of incidents.
- Top crime types, LMPD division, UOR, block addresses, and premise types are clarified and announced in previous sections.