Basics of Data Science Course - Assignment 2

Soroush heidary

96222031

Data Set : Crime Reports

Summary :

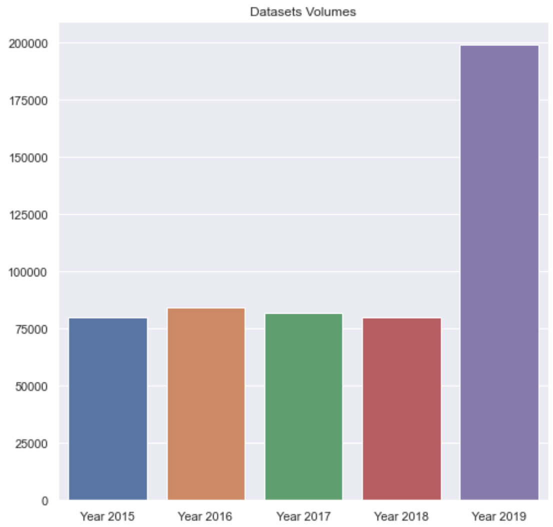
Unlike the previous assignment, this one had straightforward questions, so there will be more time to discuss new questions

Additional Queries are designed after each question, we'll attend to them one by one

5 consecutive datasets have been downloaded from year 2015 to 2019 and merged together for the majority of the code

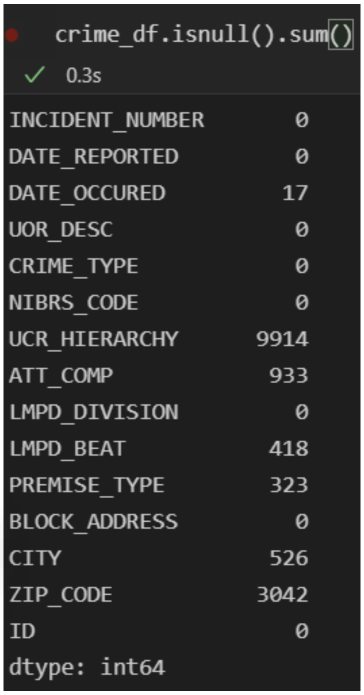
Fortunately these 5 data set had almost identical aspects so they could be concatenated with ease, the only thing that had to be edited was that 2 of these datasets had a different dtype in one of their columns and with that out of the way we can easily merge them for the rest of the code

Year 2019 had an oddly bigger dataset and the reason was that it contained the reports from year 2020 and 2021 in it too!, also in each dataset there were some outliers according to the date which were happened and reported in years before 2000

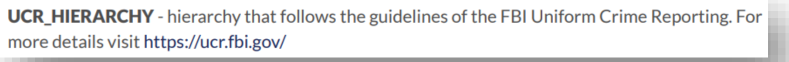


After merging the data sets the final shape was (525353, 15).

- Null values : there are some features with null values but in almost all cases the null amounts are not something to be worried about, as the whole dataset has 520k rows, so these values could be easily dropped, but as we are working with a dataset which is about crimes it's logical to keep every nan we find because they could contain valuable information, suppose we're trying to guess which crimes take the longest to crime. In this case those who have some of their info missing could very much be the ones that are expected to take longer to be reported.
  And just as happens UCR_HIERARCHY is exactly referring to the above statement

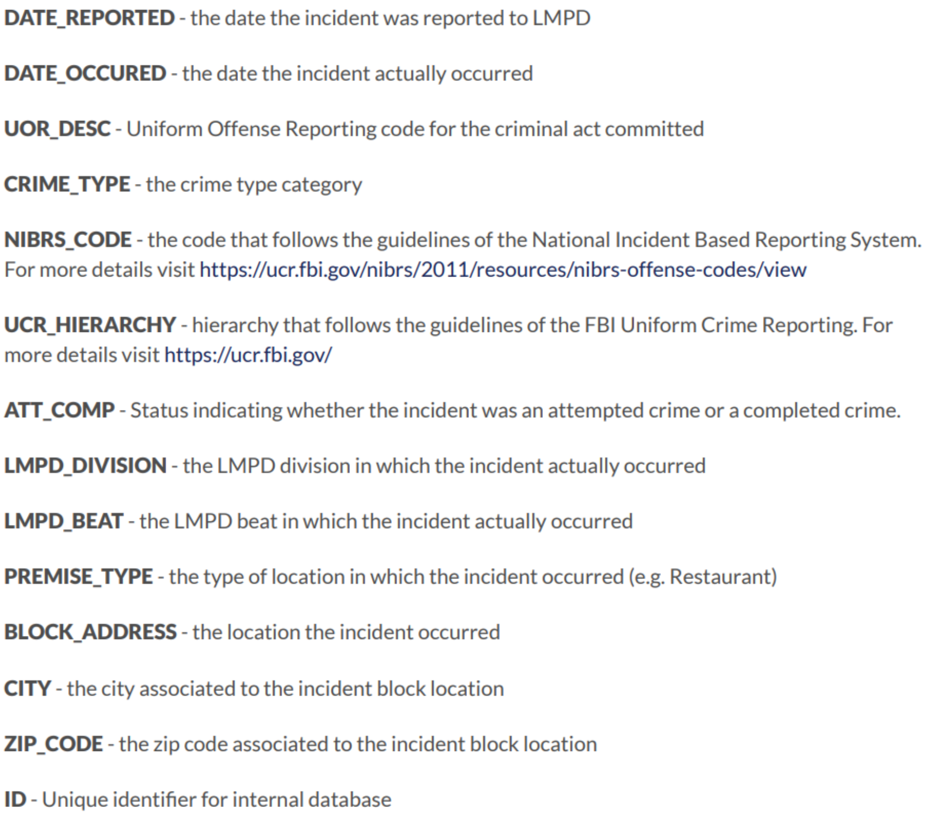


```
crime_df.isnull().sum()
✓ 0.3s
INCIDENT_NUMBER        0
DATE_REPORTED          0
DATE_OCCURED          17
UOR_DESC               0
CRIME_TYPE             0
NIBRS_CODE             0
UCR_HIERARCHY       9914
ATT_COMP             933
LMPD_DIVISION          0
LMPD_BEAT            418
PREMISE_TYPE         323
BLOCK_ADDRESS          0
CITY                 526
ZIP_CODE            3042
ID                     0
dtype: int64
```



UCR_HIERARCHY - hierarchy that follows the guidelines of the FBI Uniform Crime Reporting. For more details visit https://ucr.fbi.gov/

Here's a complete description about each feature so that we get a sense of how to handle them



DATE_REPORTED - the date the incident was reported to LMPD

DATE_OCCURED - the date the incident actually occurred

UOR_DESC - Uniform Offense Reporting code for the criminal act committed

CRIME_TYPE - the crime type category

NIBRS_CODE - the code that follows the guidelines of the National Incident Based Reporting System. For more details visit https://ucr.fbi.gov/nibrs/2011/resources/nibrs-offense-codes/view

UCR_HIERARCHY - hierarchy that follows the guidelines of the FBI Uniform Crime Reporting. For more details visit https://ucr.fbi.gov/

ATT_COMP - Status indicating whether the incident was an attempted crime or a completed crime.

LMPD_DIVISION - the LMPD division in which the incident actually occurred

LMPD_BEAT - the LMPD beat in which the incident actually occurred

PREMISE_TYPE - the type of location in which the incident occurred (e.g. Restaurant)

BLOCK_ADDRESS - the location the incident occurred

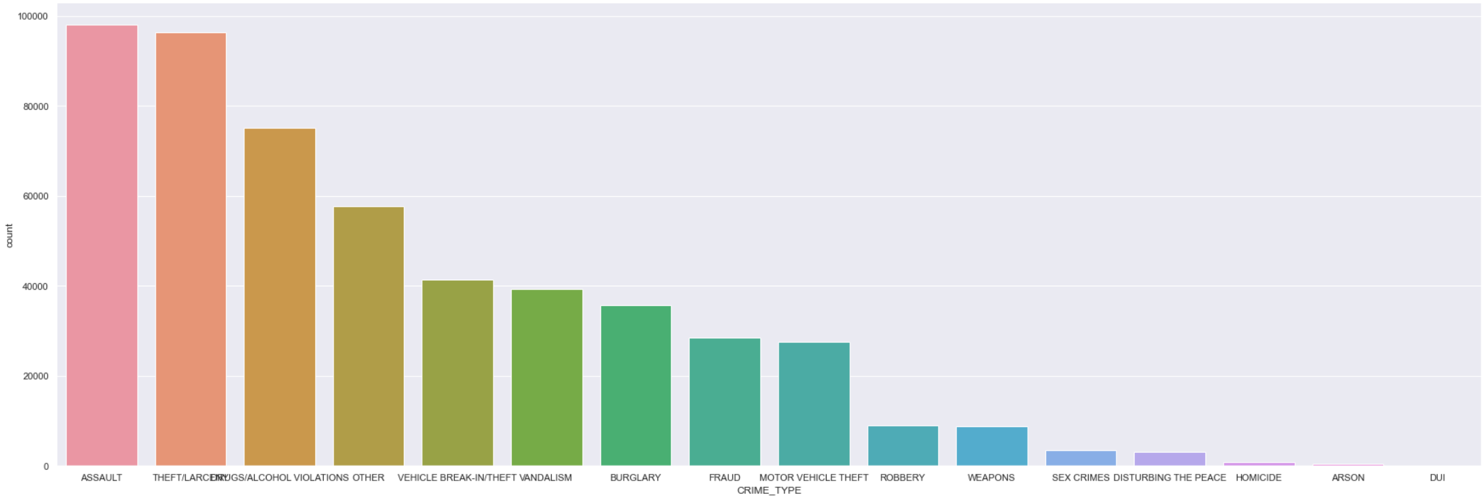
CITY - the city associated to the incident block location

ZIP_CODE - the zip code associated to the incident block location

ID - Unique identifier for internal database

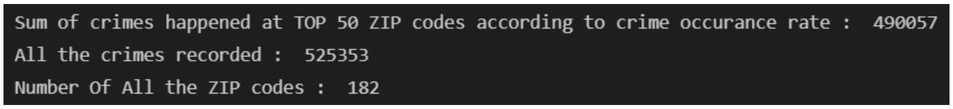
Q1 – What crime types are most common.

In the chart below we see exactly that.



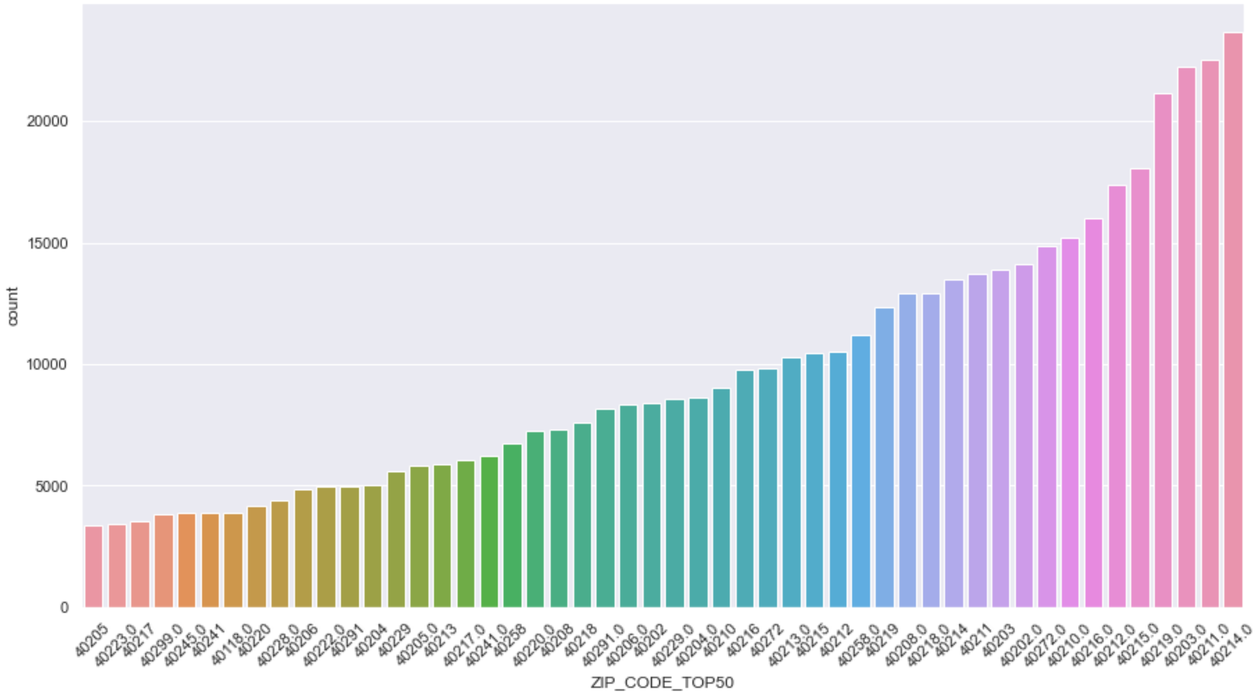
Q2 – in which zip codes are crimes more likely to happen

There were a total of 182 unique zip codes but we see the almost 93% of our crimes are happening in the TOP 50 active zip codes



```
Sum of crimes happened at TOP 50 ZIP codes according to crime occurance rate :  490057
All the crimes recorded :  525353
Number Of All the ZIP codes :  182
```

These zip codes are :

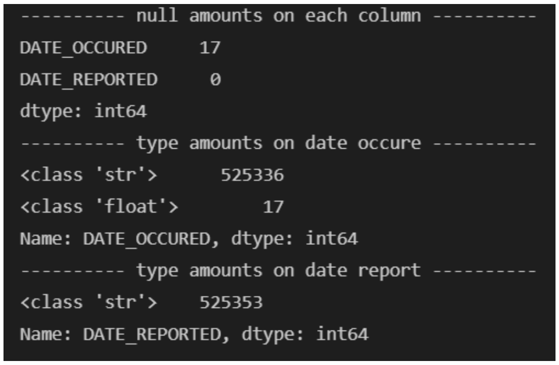


Q3 – Is there a trend of some crimes increasing and others decreasing in number over these 5 years.

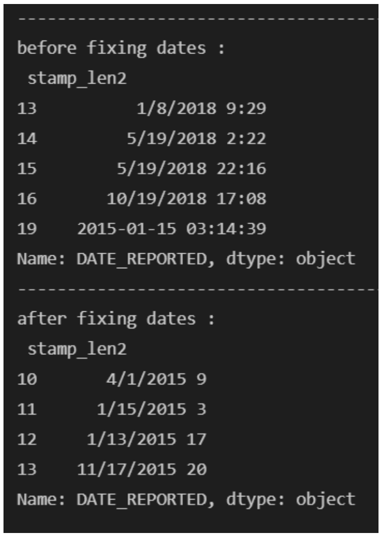
To answer this question and the next question we need to work on our date based features, DATE_OCCURED and DATE_REPORTED, for this purpose we will clean these two columns and set their type to pandas built-in type for date and time, after that we can work with dates at much more ease

there are some null values in DATE_OCCURED which actually correspond to those who have never been fully investigated to know the actual occurrence date

as they are only 17 rows we just fill each of them by their reported date (not the best approach when dealing with a high ratio of Nans though)



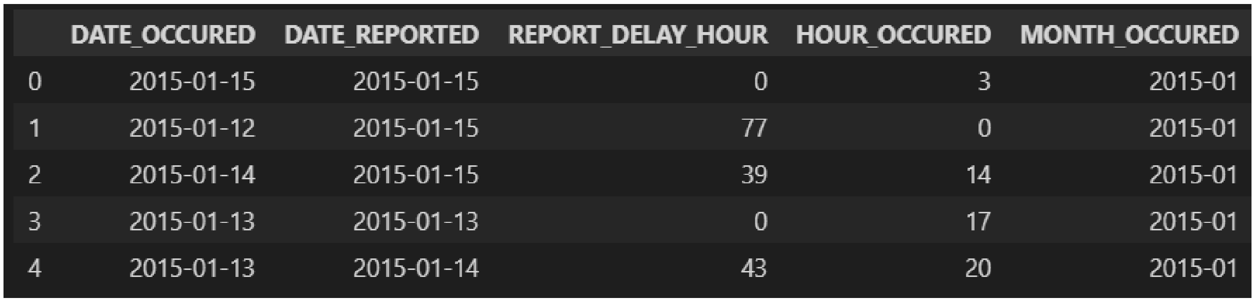
```
---------- null amounts on each column ----------
DATE_OCCURED     17
DATE_REPORTED     0
dtype: int64
---------- type amounts on date occure ----------
<class 'str'>      525336
<class 'float'>        17
Name: DATE_OCCURED, dtype: int64
---------- type amounts on date report ----------
<class 'str'>    525353
Name: DATE_REPORTED, dtype: int64
```



```
-----------------------------------
before fixing dates :
 stamp_len2
13          1/8/2018 9:29
14         5/19/2018 2:22
15        5/19/2018 22:16
16       10/19/2018 17:08
19    2015-01-15 03:14:39
Name: DATE_REPORTED, dtype: object
-----------------------------------
after fixing dates :
 stamp_len2
10       4/1/2015 9
11      1/15/2015 3
12     1/13/2015 17
13    11/17/2015 20
Name: DATE_REPORTED, dtype: object
```

now that we're sure all the data is not null, we need to work on formats, there were 5 different lengths of different formats the first four ones just need a couple of zeros, but the last kind has to be converted to be like the first four ones, then we split the hour happened (we won't really need the exact seconds and minutes and reformat these 2 columns to pd.datetime

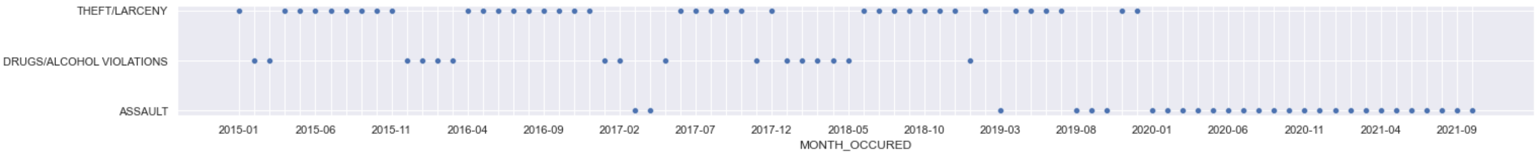
after all cleaning done the date looks like this :



	DATE_OCCURED	DATE_REPORTED	REPORT_DELAY_HOUR	HOUR_OCCURED	MONTH_OCCURED
0	2015-01-15	2015-01-15	0	3	2015-01
1	2015-01-12	2015-01-15	77	0	2015-01
2	2015-01-14	2015-01-15	39	14	2015-01
3	2015-01-13	2015-01-13	0	17	2015-01
4	2015-01-13	2015-01-14	43	20	2015-01

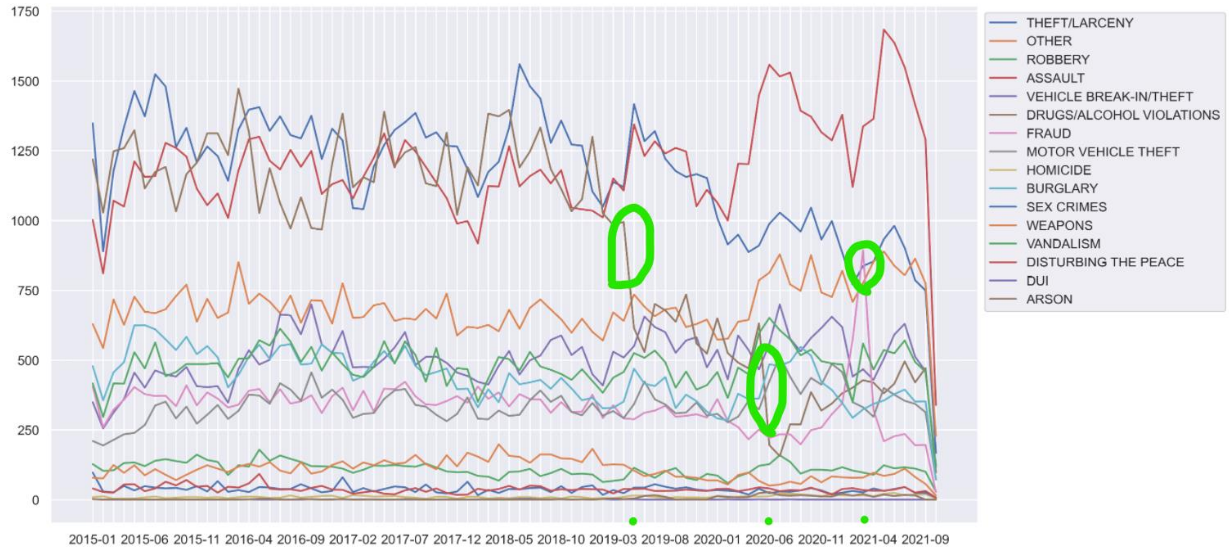
we also added a feature called REPORT_DELAY_HOUR four the next question

to see if we have any trends over these 5 years let's look at the 2 charts below



this chart shows us that what crime happened most in a month over theses five years, later on we'll see the detailed version of this assumption, but for one thing on each year around the first season of the year we see more DRUG_ALCHOL VIOLATIONS and the rest of the year THEFT/LARCENY becomes the trend again, probably because from January to around March thieves can't do their work properly due to the coldness of climate

and from the year 2020 ASSUALTs become the heavy trend



This chart provides a more detailed version and as we see from around the year 2020 the THEFT/LARCENY and DRUG/ALCOHOL VIOLATIONS decrease in number probably because the outbreak of Corona virus

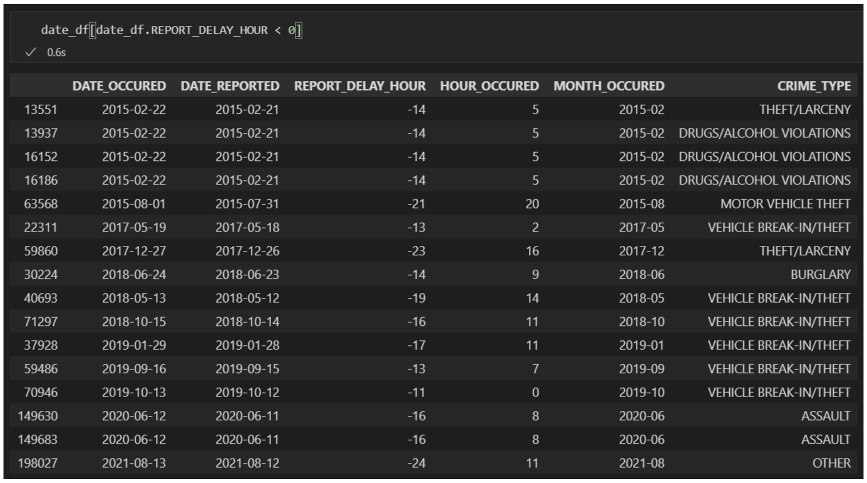
There were also something odd when I was working on the datetime features

There were some crimes who were reported before they occurred…

Most of them were vehicle break-in and drug/alcohol violations

Who am I to judge, but there's only 2 scenarios here, either there was some mistake by the computer guy, or we are watching at some inside work here …

Or some lame neighbor tried to ruin a college party with drug/alcohol consumption

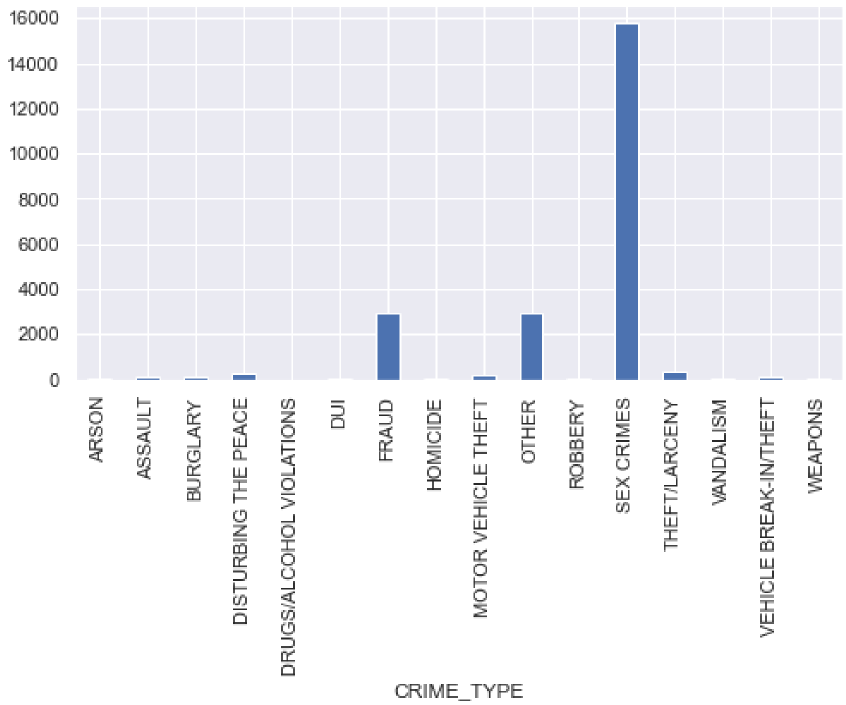


```
date_df[date_df.REPORT_DELAY_HOUR < 0]
✓ 0.6s
```

	DATE_OCCURED	DATE_REPORTED	REPORT_DELAY_HOUR	HOUR_OCCURED	MONTH_OCCURED	CRIME_TYPE
13551	2015-02-22	2015-02-21	-14	5	2015-02	THEFT/LARCENY
13937	2015-02-22	2015-02-21	-14	5	2015-02	DRUGS/ALCOHOL VIOLATIONS
16152	2015-02-22	2015-02-21	-14	5	2015-02	DRUGS/ALCOHOL VIOLATIONS
16186	2015-02-22	2015-02-21	-14	5	2015-02	DRUGS/ALCOHOL VIOLATIONS
63568	2015-08-01	2015-07-31	-21	20	2015-08	MOTOR VEHICLE THEFT
22311	2017-05-19	2017-05-18	-13	2	2017-05	VEHICLE BREAK-IN/THEFT
59860	2017-12-27	2017-12-26	-23	16	2017-12	THEFT/LARCENY
30224	2018-06-24	2018-06-23	-14	9	2018-06	BURGLARY
40693	2018-05-13	2018-05-12	-19	14	2018-05	VEHICLE BREAK-IN/THEFT
71297	2018-10-15	2018-10-14	-16	11	2018-10	VEHICLE BREAK-IN/THEFT
37928	2019-01-29	2019-01-28	-17	11	2019-01	VEHICLE BREAK-IN/THEFT
59486	2019-09-16	2019-09-15	-13	7	2019-09	VEHICLE BREAK-IN/THEFT
70946	2019-10-13	2019-10-12	-11	0	2019-10	VEHICLE BREAK-IN/THEFT
149630	2020-06-12	2020-06-11	-16	8	2020-06	ASSAULT
149683	2020-06-12	2020-06-11	-16	8	2020-06	ASSAULT
198027	2021-08-13	2021-08-12	-24	11	2021-08	OTHER

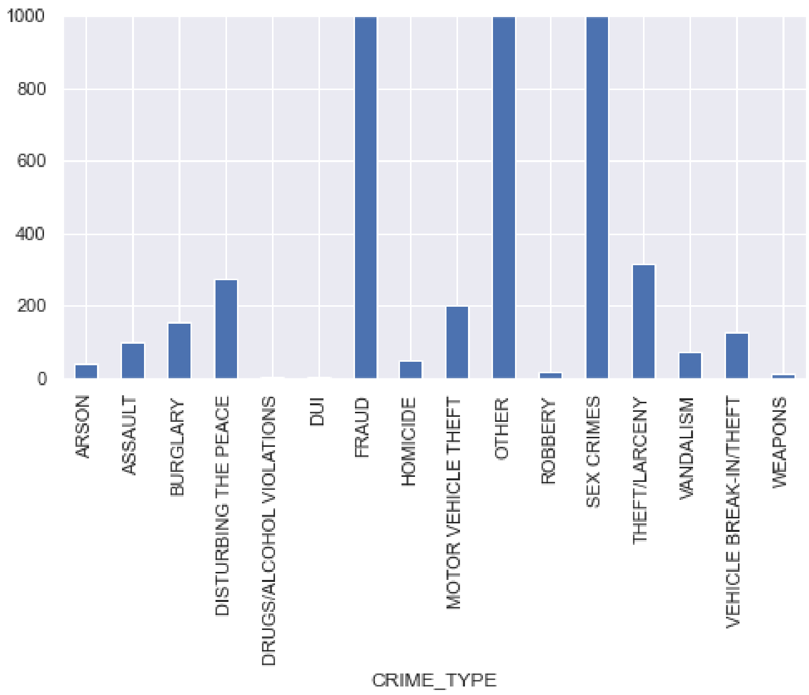
Q4 – Which crimes take the longest to report

Having the delay feature created before we need only to use a group_by method and take the mean delay of each crime_type :

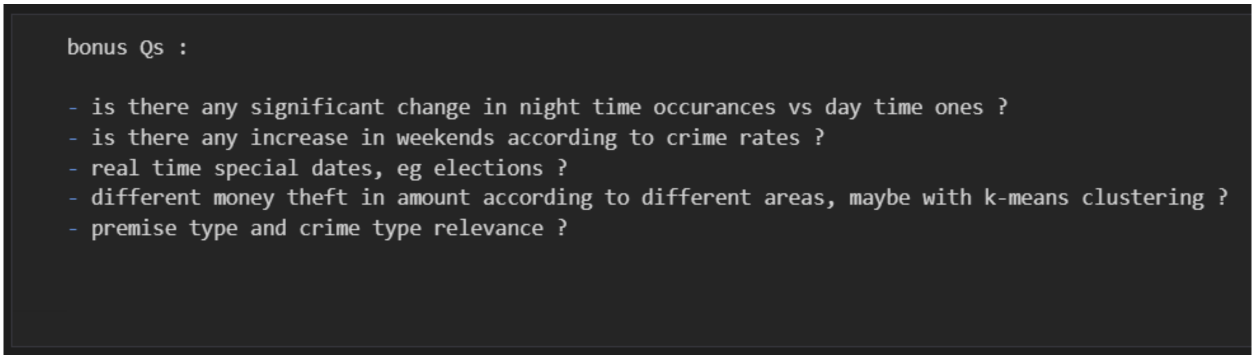


It’s actually quite self-explanatory as the sex-crimes have to be the longest ones to report followed by fraud (y axis is the hours it took to report the crime)

To see a more clear version of the picture for other types of crime there’s a y-limited chart below :



Now we can attend to some additional queries here



```
bonus Qs :

- is there any significant change in night time occurances vs day time ones ?
- is there any increase in weekends according to crime rates ?
- real time special dates, eg elections ?
- different money theft in amount according to different areas, maybe with k-means clustering ?
- premise type and crime type relevance ?
```

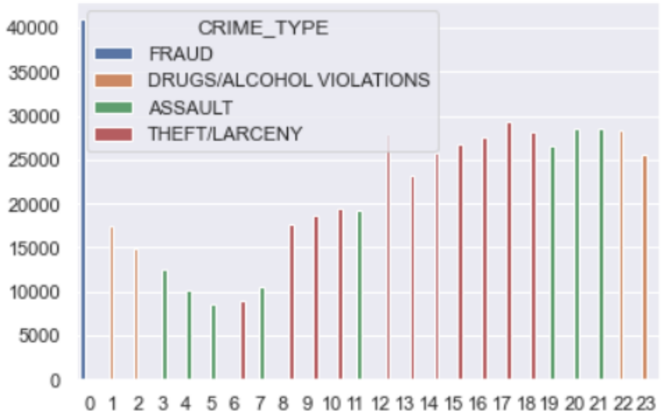
- Night-time vs Day-time :

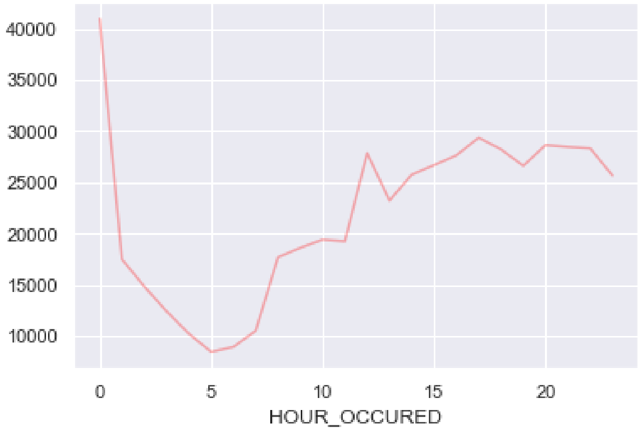
As expected, yes we'll see it in the chart below

Crime rates fall dramatically from 0am till 5am

And start to increase till 3am

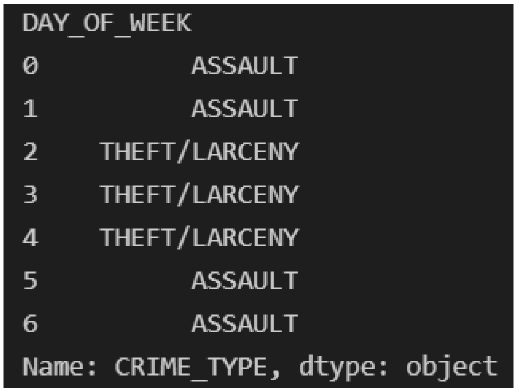
And for their types, some trend crime_types at the hour look at the chart on the left



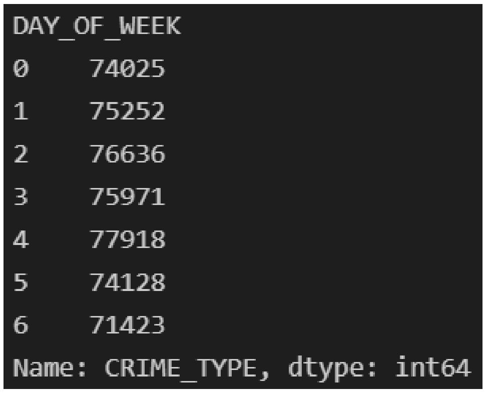


-
-
-
-
-
-
-
-
-

- Is there any increase or decrease in crime rates on weekends.
  Well not much on the amount but the crime trends do

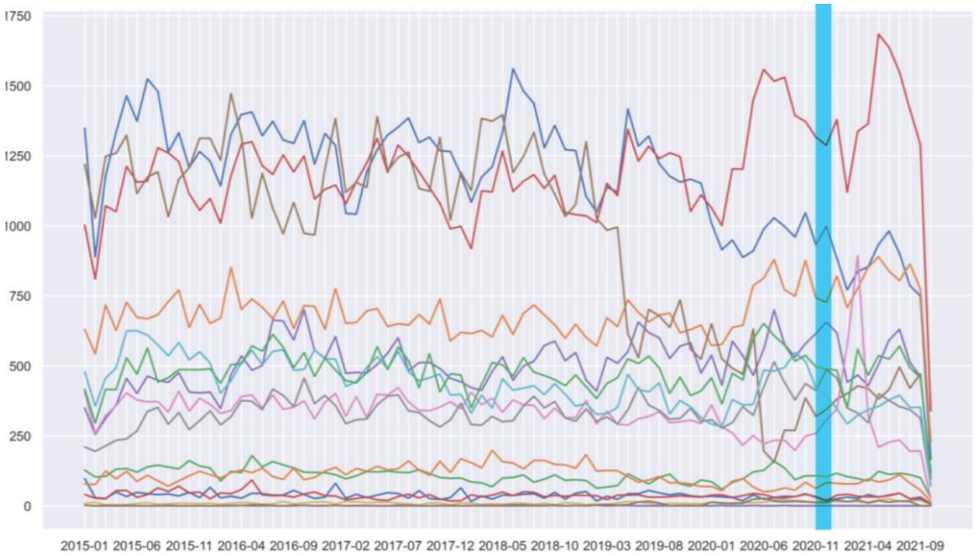


```
DAY_OF_WEEK
0          ASSAULT
1          ASSAULT
2    THEFT/LARCENY
3    THEFT/LARCENY
4    THEFT/LARCENY
5          ASSAULT
6          ASSAULT
Name: CRIME_TYPE, dtype: object
```



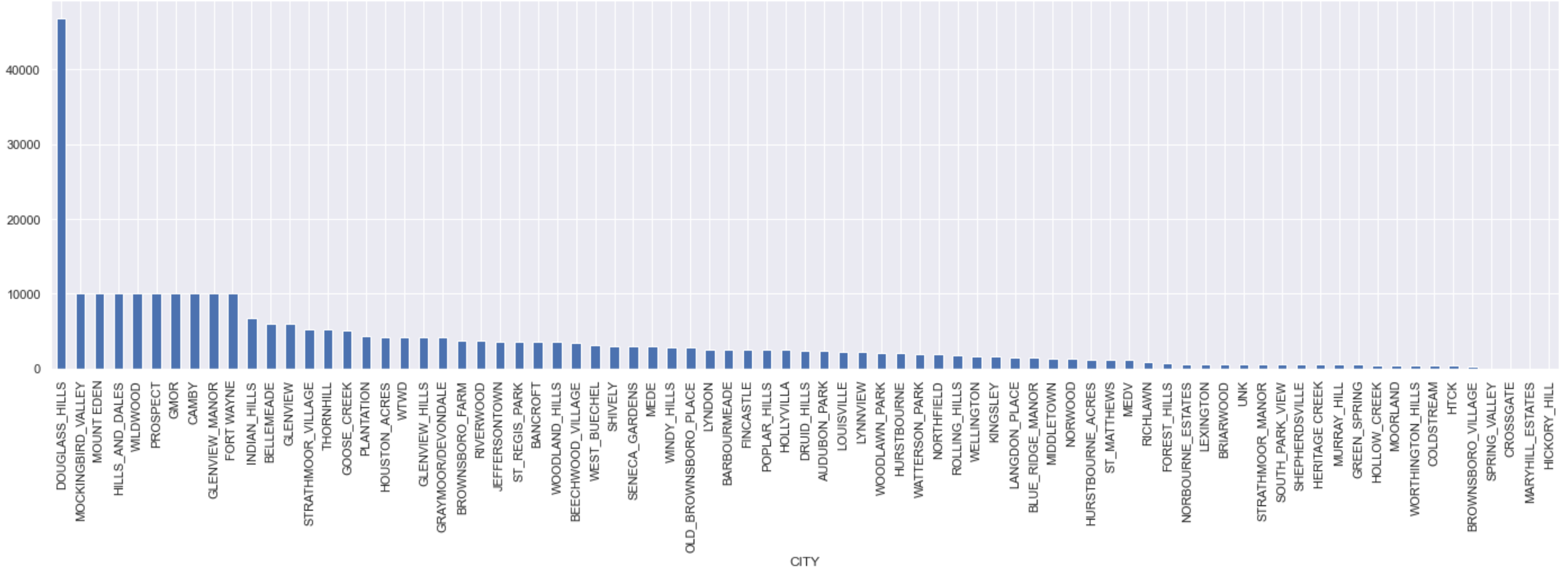
```
DAY_OF_WEEK
0    74025
1    75252
2    76636
3    75971
4    77918
5    74128
6    71423
Name: CRIME_TYPE, dtype: int64
```

- On elections, do the crime rate increase
  Well no, 2020 america's election of on November 3 and we don't see any change in the chart below around that time



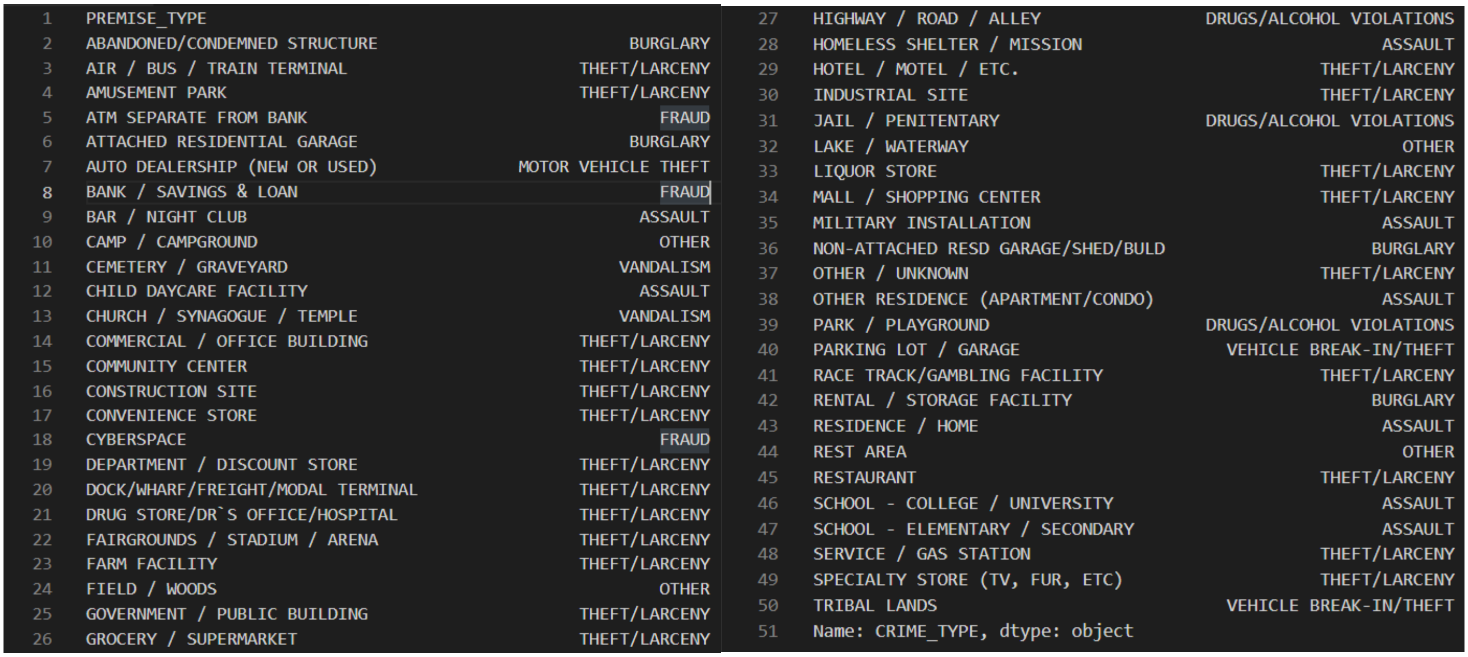
- On fraud Crime_TYPE is there any difference in money according to different areas.

  beside one outlier (DOUGLASS_HILLS) we can see the mean price that has been taken away from people in each city



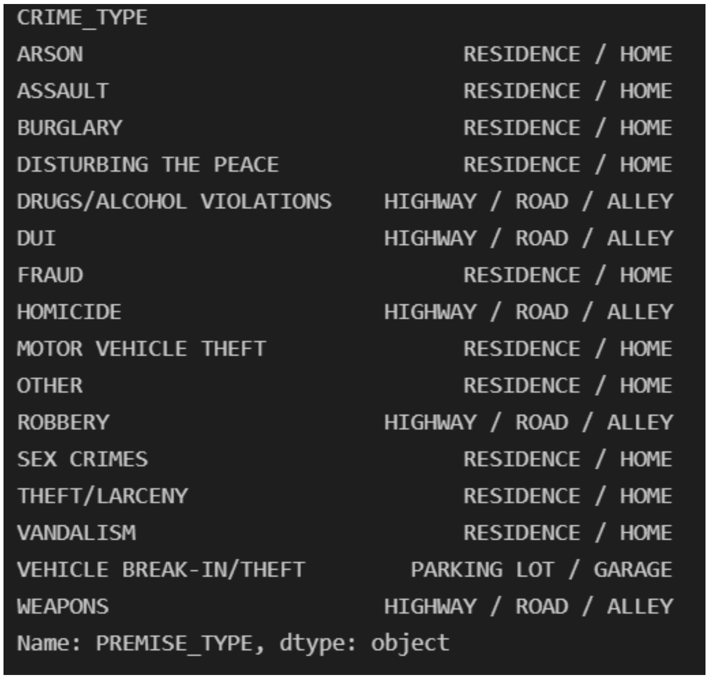
- Is there any relevance between premise type and crime type. Which crime types are more common on each premise type

  The result very self explanatory, for example we would expect the trend crime in an abandoned structure to be BURGLURY



```
PREMISE_TYPE
ABANDONED/CONDEMNED STRUCTURE                     BURGLARY
AIR / BUS / TRAIN TERMINAL                   THEFT/LARCENY
AMUSEMENT PARK                               THEFT/LARCENY
ATM SEPARATE FROM BANK                               FRAUD
ATTACHED RESIDENTIAL GARAGE                       BURGLARY
AUTO DEALERSHIP (NEW OR USED)          MOTOR VEHICLE THEFT
BANK / SAVINGS & LOAN                                FRAUD
BAR / NIGHT CLUB                                   ASSAULT
CAMP / CAMPGROUND                                    OTHER
CEMETERY / GRAVEYARD                             VANDALISM
CHILD DAYCARE FACILITY                             ASSAULT
CHURCH / SYNAGOGUE / TEMPLE                      VANDALISM
COMMERCIAL / OFFICE BUILDING                 THEFT/LARCENY
COMMUNITY CENTER                             THEFT/LARCENY
CONSTRUCTION SITE                            THEFT/LARCENY
CONVENIENCE STORE                            THEFT/LARCENY
CYBERSPACE                                           FRAUD
DEPARTMENT / DISCOUNT STORE                  THEFT/LARCENY
DOCK/WHARF/FREIGHT/MODAL TERMINAL            THEFT/LARCENY
DRUG STORE/DR`S OFFICE/HOSPITAL              THEFT/LARCENY
FAIRGROUNDS / STADIUM / ARENA                THEFT/LARCENY
FARM FACILITY                                THEFT/LARCENY
FIELD / WOODS                                        OTHER
GOVERNMENT / PUBLIC BUILDING                 THEFT/LARCENY
GROCERY / SUPERMARKET                        THEFT/LARCENY
HIGHWAY / ROAD / ALLEY                DRUGS/ALCOHOL VIOLATIONS
HOMELESS SHELTER / MISSION                         ASSAULT
HOTEL / MOTEL / ETC.                         THEFT/LARCENY
INDUSTRIAL SITE                              THEFT/LARCENY
JAIL / PENITENTIARY                   DRUGS/ALCOHOL VIOLATIONS
LAKE / WATERWAY                                      OTHER
LIQUOR STORE                                 THEFT/LARCENY
MALL / SHOPPING CENTER                       THEFT/LARCENY
MILITARY INSTALLATION                              ASSAULT
NON-ATTACHED RESD GARAGE/SHED/BULD                BURGLARY
OTHER / UNKNOWN                              THEFT/LARCENY
OTHER RESIDENCE (APARTMENT/CONDO)                  ASSAULT
PARK / PLAYGROUND                     DRUGS/ALCOHOL VIOLATIONS
PARKING LOT / GARAGE                   VEHICLE BREAK-IN/THEFT
RACE TRACK/GAMBLING FACILITY                 THEFT/LARCENY
RENTAL / STORAGE FACILITY                         BURGLARY
RESIDENCE / HOME                                   ASSAULT
REST AREA                                            OTHER
RESTAURANT                                   THEFT/LARCENY
SCHOOL - COLLEGE / UNIVERSITY                      ASSAULT
SCHOOL - ELEMENTARY / SECONDARY                    ASSAULT
SERVICE / GAS STATION                        THEFT/LARCENY
SPECIALTY STORE (TV, FUR, ETC)               THEFT/LARCENY
TRIBAL LANDS                           VEHICLE BREAK-IN/THEFT
Name: CRIME_TYPE, dtype: object
```

We see the inverse of our assumption here (which premise type is the most common place for each crime type to happen)



```
CRIME_TYPE
ARSON                             RESIDENCE / HOME
ASSAULT                           RESIDENCE / HOME
BURGLARY                          RESIDENCE / HOME
DISTURBING THE PEACE              RESIDENCE / HOME
DRUGS/ALCOHOL VIOLATIONS    HIGHWAY / ROAD / ALLEY
DUI                         HIGHWAY / ROAD / ALLEY
FRAUD                             RESIDENCE / HOME
HOMICIDE                    HIGHWAY / ROAD / ALLEY
MOTOR VEHICLE THEFT               RESIDENCE / HOME
OTHER                             RESIDENCE / HOME
ROBBERY                     HIGHWAY / ROAD / ALLEY
SEX CRIMES                        RESIDENCE / HOME
THEFT/LARCENY                     RESIDENCE / HOME
VANDALISM                         RESIDENCE / HOME
VEHICLE BREAK-IN/THEFT        PARKING LOT / GARAGE
WEAPONS                     HIGHWAY / ROAD / ALLEY
Name: PREMISE_TYPE, dtype: object
```