Boston crimes 2022

Wiktoria Konop

Data cleaning

Initial interaction with data

At the beginning I loaded the dataset taken from here and I checked its content.

Dimensions: 73852 rows and 17 cols

```
## [1] 73852    17
```

The data set contains information about types of offenses (with unique incident number and offense codes) committed in 2022 on the premises of Boston (the capital and largest city of the state of Massachusetts in the USA). There are also certain columns which help us to uncover more insights about place: district, reporting area, street and three columns with geographical coordinates.

Apart from that, there are some columns that provide details about the date of violation, more specifically we have information about minute, hour, day, day of the week and month.

What is more, we know about gunfight. 0 in a row means there was not shooting, 1 contrary.

Initial data (first six rows) looks as follows:

```
##   INCIDENT_NUMBER OFFENSE_CODE OFFENSE_CODE_GROUP
## 1       222076257          619                 NA
## 2       222053099         2670                 NA
## 3       222039411         3201                 NA
## 4       222011090         3201                 NA
## 5       222062685         3201                 NA
## 6       222040307         3115                 NA


##                OFFENSE_DESCRIPTION DISTRICT REPORTING_AREA SHOOTING
## 1               LARCENY ALL OTHERS       D4            167        0
## 2 HARASSMENT/ CRIMINAL HARASSMENT       A7             NA        0
## 3         PROPERTY - LOST/ MISSING      D14            778        0
## 4         PROPERTY - LOST/ MISSING       B3            465        0
## 5         PROPERTY - LOST/ MISSING       B3            465        0
## 6               INVESTIGATE PERSON       A1            954        0


##   OCCURRED_ON_DATE YEAR MONTH DAY_OF_WEEK HOUR UCR_PART        STREET      Lat
## 1    1/1/2022 0:00 2022     1    Saturday    0       NA  HARRISON AVE 42.33954
## 2    1/1/2022 0:00 2022     1    Saturday    0       NA BENNINGTON ST 42.37725
## 3    1/1/2022 0:00 2022     1    Saturday    0       NA WASHINGTON ST 42.34906
## 4    1/1/2022 0:00 2022     1    Saturday    0       NA BLUE HILL AVE 42.28483
## 5    1/1/2022 0:00 2022     1    Saturday    0       NA BLUE HILL AVE 42.28483
## 6    1/1/2022 0:00 2022     1    Saturday    0       NA     FULTON ST 42.36294
##        Long                                Location
## 1 -71.06941 (42.33954198983014, -71.06940876967543)
## 2 -71.03260  (42.37724638479816, -71.0325970804128)
## 3 -71.15050 (42.34905600030506, -71.15049849975023)
## 4 -71.09137 (42.28482576580488, -71.09137368938802)
## 5 -71.09137 (42.28482576580488, -71.09137368938802)
## 6 -71.05254  (42.36293610909294, -71.0525379472723)
```

Immediately I can tell that there are some unnecessary columns which will have no merit to this analysis, so I removed them. The mentioned columns are:

- INCIDENT_NUMBER - for each row it is different, so it gives us zero meaningful information
- OFFENSE_CODE_GROUP- empty column
- YEAR - each data refers to 2022
- UCR_PART - empty column

Missing values

In the next step I will examine data completeness, hence I'm checking if there is any NA or empty string. I discovered that there are some empty strings which i converted into NA and I summed up NA in each column:

```
##        OFFENSE_CODE OFFENSE_DESCRIPTION            DISTRICT      REPORTING_AREA
##                   0                   0                 171               44448
##            SHOOTING    OCCURRED_ON_DATE               MONTH         DAY_OF_WEEK
##                   0                   0                   0                   0
##                HOUR              STREET                 Lat                Long
##                   0                   0                3808                3808
##            Location
##                3808

## Percent of NA values in each column [%]:  0 0 0.2315442 60.18524 0 0 0 0 0 0 5.156
258 5.156258 5.156258
```

For now, I'm leaving all rows (with and without missing values). I will remove them only when I will be analyzing single variables. However, I decided to delete thoroughly "reporting_area" because in this column there is more missing values (more than 60% !) than actual values.

The correctness of data

I detected one typo in data, namely instead of "MURDER, NON-NEGLIGENT MANSLAUGHTER" in the system was written "MURDER, NON-NEGLIGIENT MANSLAUGHTER", so I corrected it.

Data decluttering

In the next step, I changed headlines to lowercase and split the variable "occurred_on_date". Some of the information from this column I had already had, so I placed separately only day and minutes. Then I removed "occurred_on_date" and I reordered the data set.

```
##  [1] "offense_code"        "offense_description" "shooting"
##  [4] "month"               "day"                 "day_of_week"
##  [7] "hour"                "minutes"             "district"
## [10] "street"              "lat"                 "long"
## [13] "location"
```

Types of variables and levels of measurement

Afterwards, I checked the types of variables. They are presented in the table below:

```
## 'data.frame':    73852 obs. of  13 variables:
##  $ offense_code       : int  619 2670 3201 3201 3201 3115 2670 3114 1109 423 ...
##  $ offense_description: chr  "LARCENY ALL OTHERS" "HARASSMENT/ CRIMINAL HARASSMENT
" "PROPERTY - LOST/ MISSING" "PROPERTY - LOST/ MISSING" ...
##  $ shooting           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ month              : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ day                : chr  "1" "1" "1" "1" ...
##  $ day_of_week        : chr  "Saturday" "Saturday" "Saturday" "Saturday" ...
##  $ hour               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ minutes            : chr  "0" "0" "0" "0" ...
##  $ district           : chr  "D4" "A7" "D14" "B3" ...
##  $ street             : chr  "HARRISON AVE" "BENNINGTON ST" "WASHINGTON ST" "BLUE
HILL AVE" ...
##  $ lat                : num  42.3 42.4 42.3 42.3 42.3 ...
##  $ long               : num  -71.1 -71 -71.2 -71.1 -71.1 ...
##  $ location           : chr  "(42.33954198983014, -71.06940876967543)" "(42.377246
38479816, -71.0325970804128)" "(42.34905600030506, -71.15049849975023)" "(42.28482576
580488, -71.09137368938802)" ...
```

Let me explain in more detail the division of types. We may have **categorical and numerical data**. The first one represents groups or categories, the second one represents numbers. Moreover, numerical data is divided into two groups: discrete (counted in a finite matter) and continuous (infinite and impossible to count).

With each type of data is associated a specific level of measurement. For categorical data we acknowledge **qualitative level**, for numerical **quantitative level**. The qualitative level is segmented into nominal (categories that cannot be put in any order) and ordinal (ordered categories), while the quantitative level is partitioned into ratio (has a true zero) and interval (without true zero).

****

Categorical nominal

In our data set, most of variables belong to categorical type and their qualitative level is nominal. Variables such as offense_code, offense_description, shooting, district, street represent groups with no inherent order or ranking.

Categorical ordinal

In the context of date, month could be considered ordinal as it follows a sequential order from January to December, same with days that are ordered from 1 to 28/29/30/31. The days of the week may be ordered as well, this time based on their position in the week (e.g., Sunday, Monday, Tuesday, etc.), so I will consider them also ordinal.

Time in general is numerical, however in our example we may treat it as a categorical ordinal because we have only integer values, so zero digits after point and we may set hours from 0-23, minutes from 00-59.

Numerical (continuous) interval

Geographical coordinates typically belong to the numerical data type. They represent specific points on the Earth's surface using latitude and longitude values, which are numeric measurements. Latitude specifies the north-south position, while longitude specifies the east-west position.

****

Now as we know more about types we can make some changes. Since in R categorical data is often stored in a factor (data structure), I convert each categorical variable to factor. The results are below:

```
## 'data.frame':    73852 obs. of  13 variables:
##  $ offense_code       : Factor w/ 119 levels "111","121","301",..: 17 67 94 94 94
85 67 84 30 4 ...
##  $ offense_description: Factor w/ 119 levels "AFFRAY","AIRCRAFT INCIDENTS",..: 50
42 90 90 90 46 42 47 39 6 ...
##  $ shooting           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ month              : Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1
1 ...
##  $ day                : Factor w/ 31 levels "1","10","11",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ day_of_week        : Factor w/ 7 levels "Friday","Monday",..: 3 3 3 3 3 3 3 3 3
3 ...
##  $ hour               : Factor w/ 24 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1
1 ...
##  $ minutes            : Factor w/ 60 levels "0","1","10","11",..: 1 1 1 1 1 1 1 2
2 23 ...
##  $ district           : Factor w/ 13 levels "A1","A15","A7",..: 9 3 8 5 5 1 9 5 7
9 ...
##  $ street             : Factor w/ 8378 levels "1 FINANCIAL CTR",..: 3827 685 7800
809 809 3385 3827 6768 7550 1223 ...
##  $ lat                : num  42.3 42.4 42.3 42.3 42.3 ...
##  $ long               : num  -71.1 -71 -71.2 -71.1 -71.1 ...
##  $ location           : chr  "(42.33954198983014, -71.06940876967543)" "(42.377246
38479816, -71.0325970804128)" "(42.34905600030506, -71.15049849975023)" "(42.28482576
580488, -71.09137368938802)" ...
```

The final result

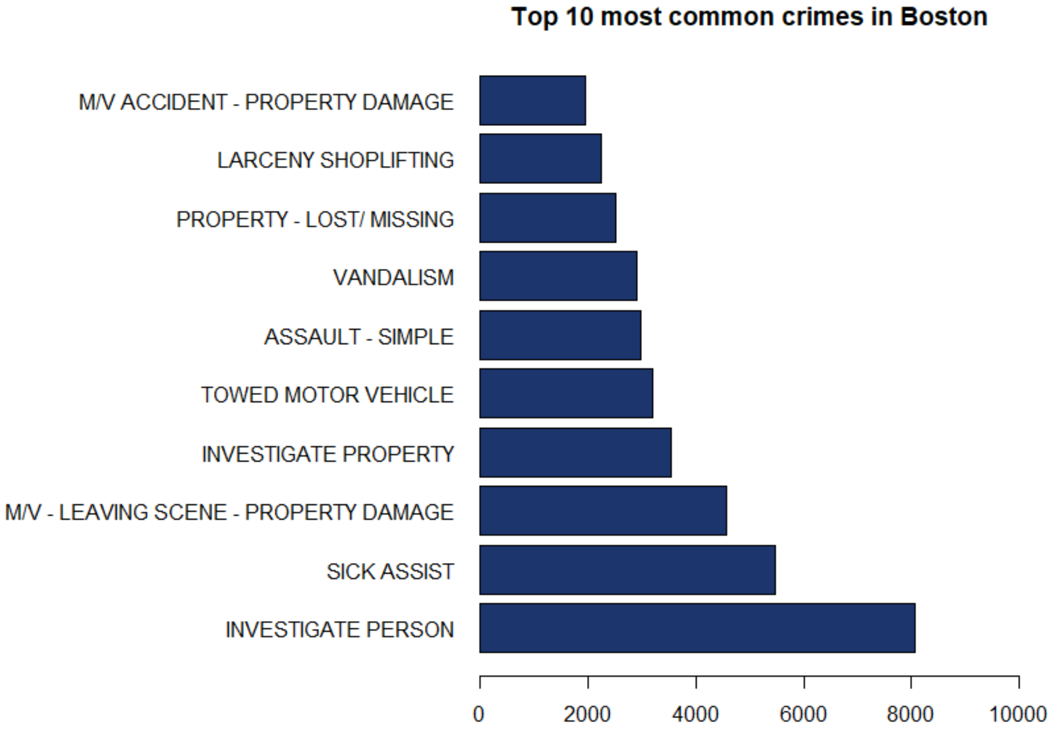
The cleaned and organized data is prepared for further analysis.

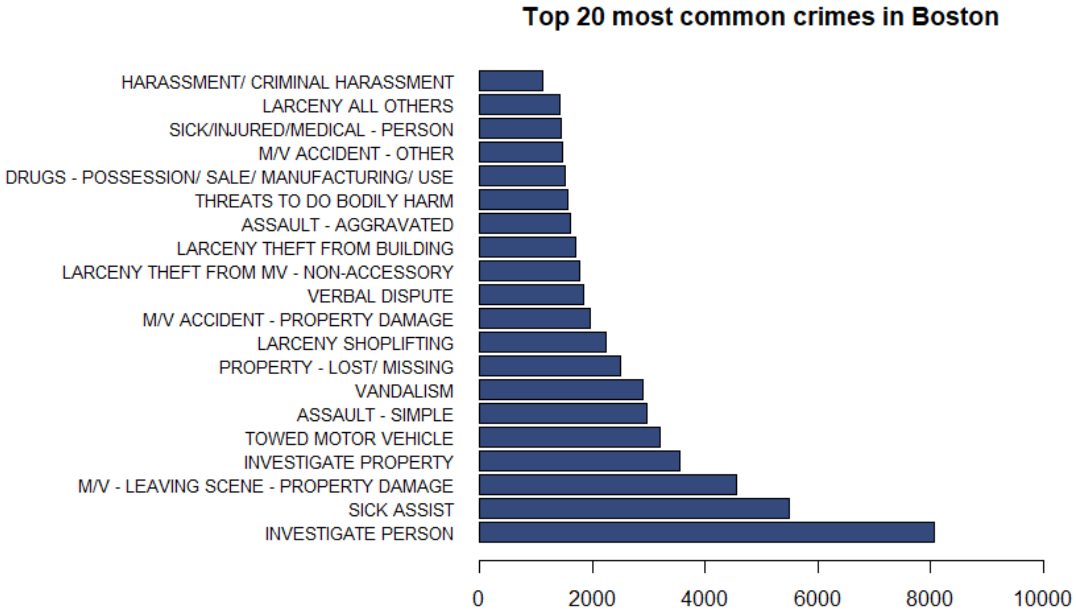
```
##    offense_code             offense_description shooting month day day_of_week
## 1           619              LARCENY ALL OTHERS        0     1   1    Saturday
## 2          2670 HARASSMENT/ CRIMINAL HARASSMENT        0     1   1    Saturday
## 3          3201        PROPERTY - LOST/ MISSING        0     1   1    Saturday
## 4          3201        PROPERTY - LOST/ MISSING        0     1   1    Saturday
## 5          3201        PROPERTY - LOST/ MISSING        0     1   1    Saturday
## 6          3115              INVESTIGATE PERSON        0     1   1    Saturday
## 7          2670 HARASSMENT/ CRIMINAL HARASSMENT        0     1   1    Saturday
## 8          3114            INVESTIGATE PROPERTY        0     1   1    Saturday
## 9          1109                    FRAUD - WIRE        0     1   1    Saturday
## 10          423            ASSAULT - AGGRAVATED        0     1   1    Saturday
##    hour minutes district        street      lat      long
## 1     0       0       D4  HARRISON AVE 42.33954 -71.06941
## 2     0       0       A7 BENNINGTON ST 42.37725 -71.03260
## 3     0       0      D14 WASHINGTON ST 42.34906 -71.15050
## 4     0       0       B3 BLUE HILL AVE 42.28483 -71.09137
## 5     0       0       B3 BLUE HILL AVE 42.28483 -71.09137
## 6     0       0       A1     FULTON ST 42.36294 -71.05254
## 7     0       0       D4  HARRISON AVE 42.33954 -71.06941
## 8     0       1       B3     SELDEN ST 42.28089 -71.08037
## 9     0       1       C6    W BROADWAY 42.34129 -71.05468
## 10    0      29       D4 BROOKLINE AVE 42.34625 -71.09954
##                                    location
## 1   (42.33954198983014, -71.06940876967543)
## 2    (42.37724638479816, -71.0325970804128)
## 3   (42.34905600030506, -71.15049849975023)
## 4   (42.28482576580488, -71.09137368938802)
## 5   (42.28482576580488, -71.09137368938802)
## 6    (42.36293610909294, -71.0525379472723)
## 7   (42.33954198983014, -71.06940876967543)
## 8   (42.280893655822176, -71.0803746810546)
## 9  (42.341287504390436, -71.05467932649397)
## 10  (42.34625079905638, -71.09953855872904)
```

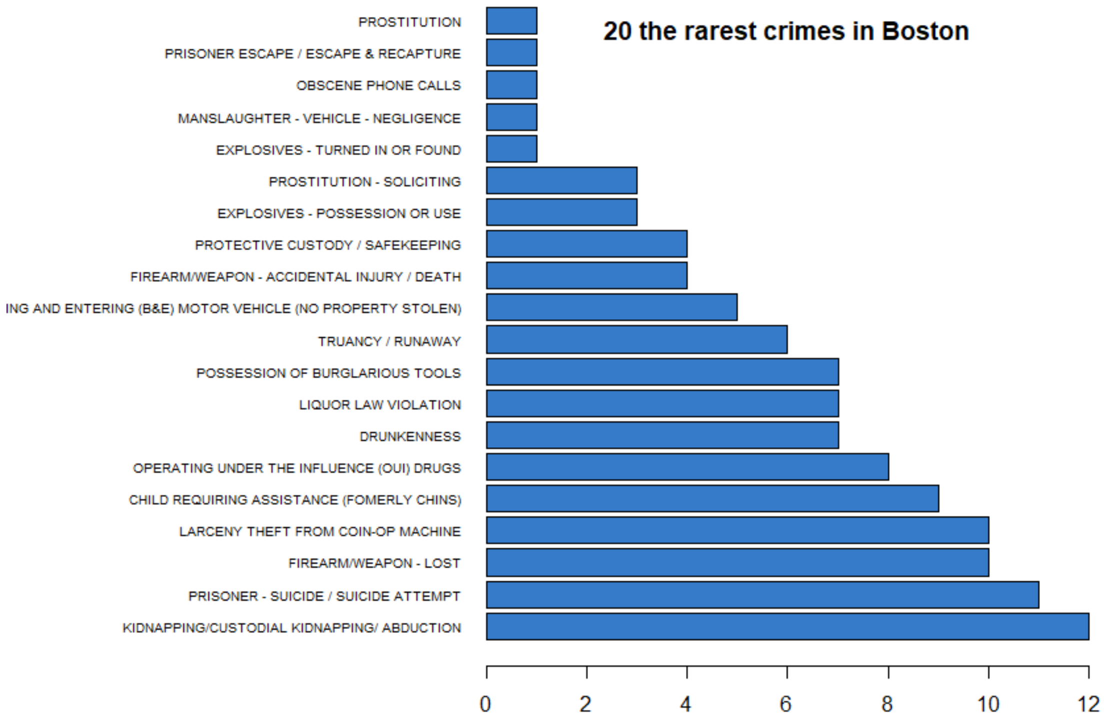
EDA 1-dimensional

I started by analyzing each column one after another and I plotted results. At the very beginning, I was obtained information about 10 most popular crimes in Boston. It turned out that INVESTIGATE PERSON is at the forefront, after it SICK ASSIST and then M/V LEAVING SCENE-PROPERTY DAMAGE. I would like to highlight that I don't have access to more accurate data about crimes description, so I may only guessing to what refer the keywords. In my opinion, first offenses are not really serious. They concern rather daily petty crimes.

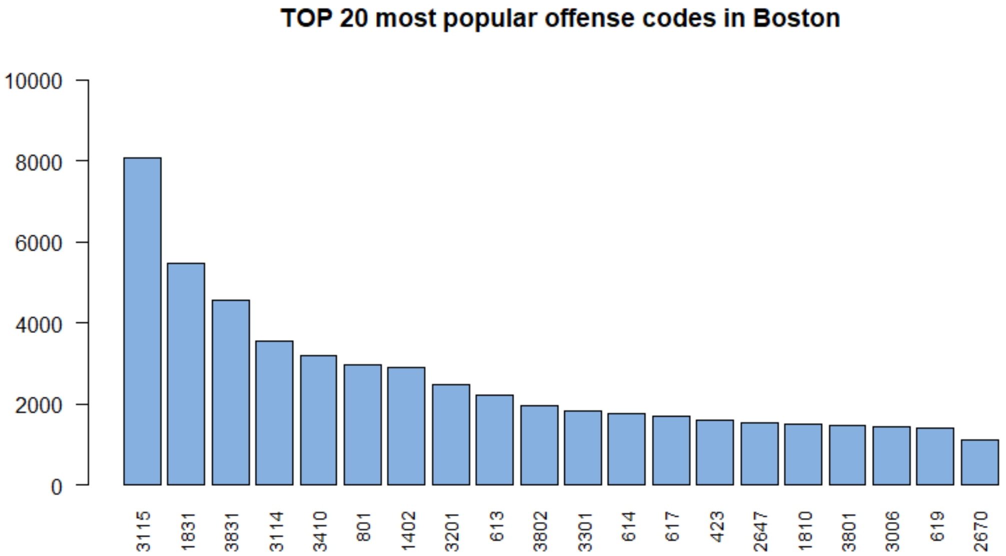
Below I present bar graphs with 10 and 20 the most frequent crimes and with 20 the least.







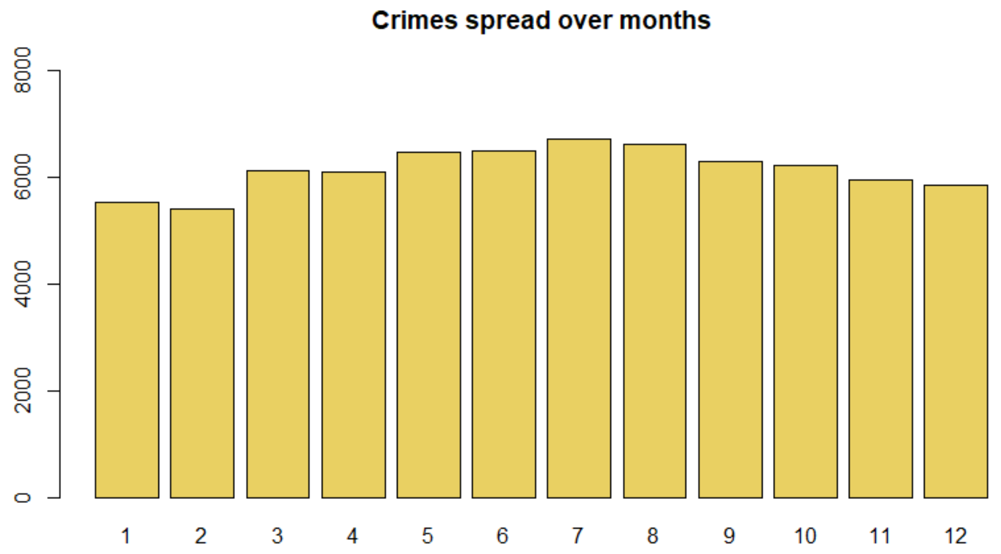
Each crime description has assigned its own code. Maybe there are some patterns that for example codes with numbers from certain range are related to more sever delinquency or to proper code like penal code, civil code, labor code and so on. In the following chart I showed the incidence of offense codes:



Date visualisation

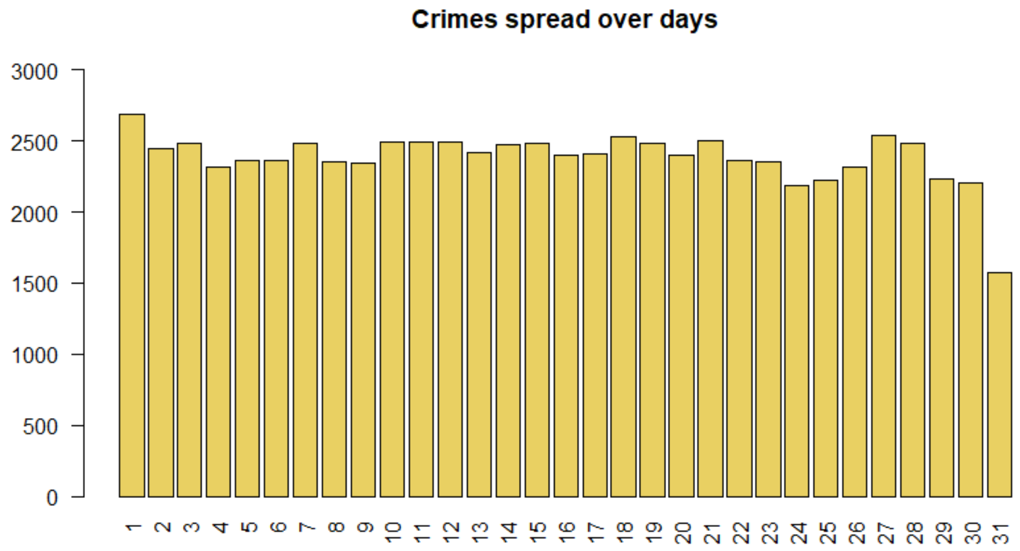
Months

The graph below shows how many crimes were committed in particular months of the 2022. The least occur in winter months, the most during the summer. Personally I think it is nothing surprising because summer creates favorable conditions for criminals - more events take place, people are encouraged by the weather to leave their houses, to spend their spare time outside with their family or friends or to visit some places, most often city centers.



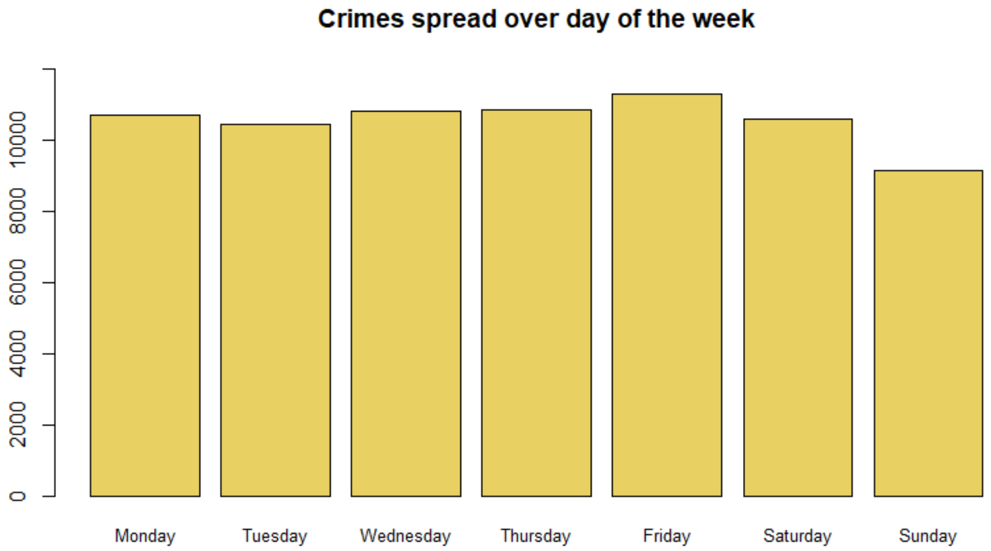
Days

What draws attention is the huge different in numbers of delinquencies between the first day and the last day. Moreover, the first day stands out among other days. It can stem from few factors, for example: when there was no information about the day of crime, the person who entered the data could assumed the 1st day of the month (there must be some examples of inaccurate or complete lack of the information about date and time as the same case we will notice with hour).



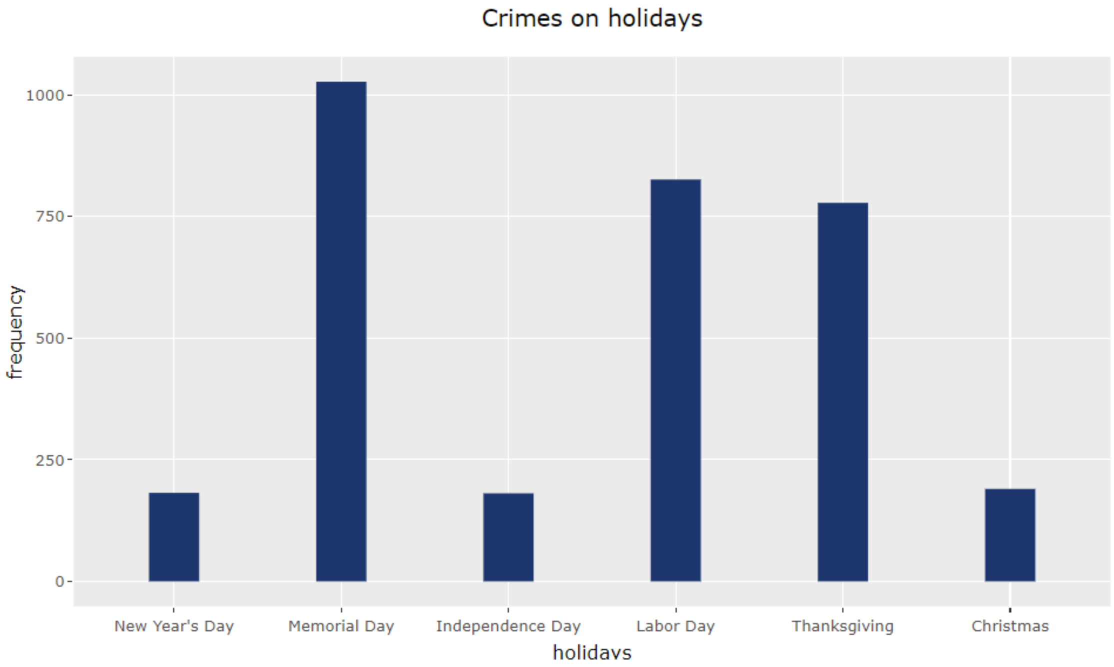
Days of the week

Friday is at the forefront among all days of the week when it comes to committed crimes. This is the end of the week, so that people know they can stay late in public places because the next day is out of work. This is a perfect moment for thieves, pickpockets, vandals, fraudsters or drug dealers. Friday may be also a day when people get paid and no one can deny that many people have a tendency to spending their money (on shopping, on alcohol, on other drugs or entertainment) straightaway. Contrary to Friday, the most peaceful day is Sunday. At that time, people have time for rest, time for themselves, and their families.

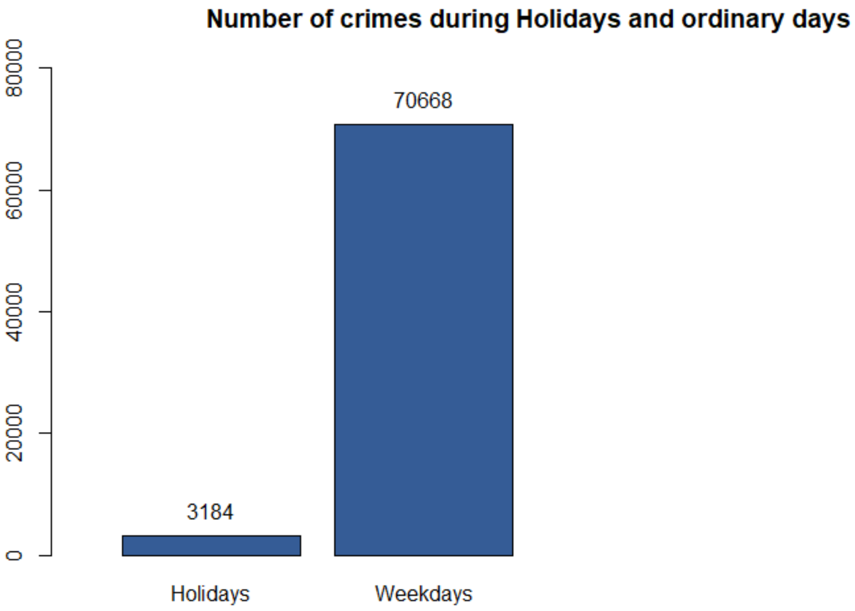


Holidays

Another interesting feature is shown on the chart below. It depicts crimes during 6 holidays frequently celebrated in the USA. The highest percentage is observed on the following days: Memorial day, Labor Day and Thanksgiving whereas the lowest on New Year's Day, Independence Day and on Christmas.



Taking into consideration only these 6 holidays we can tell that in these days were undertaken 3184 criminal actions which constitutes around 4.3 percent of total crimes.

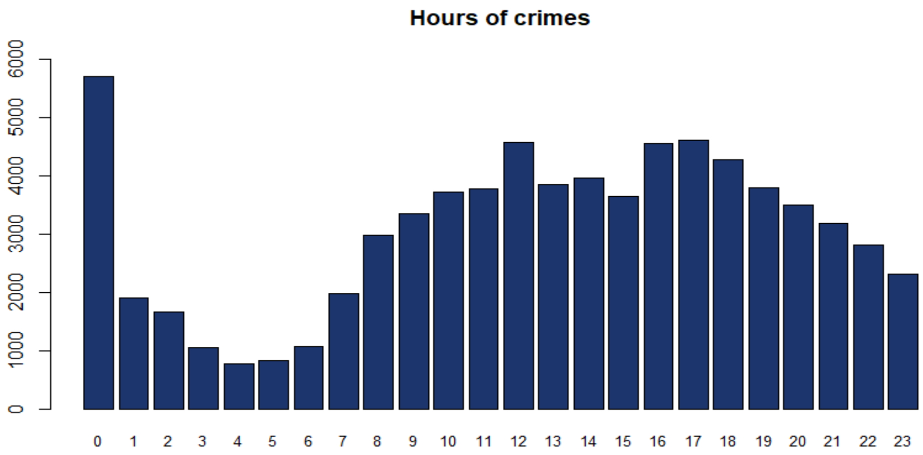


Time visualization

Next two charts illustrate particular hour and minute of the law transgression.

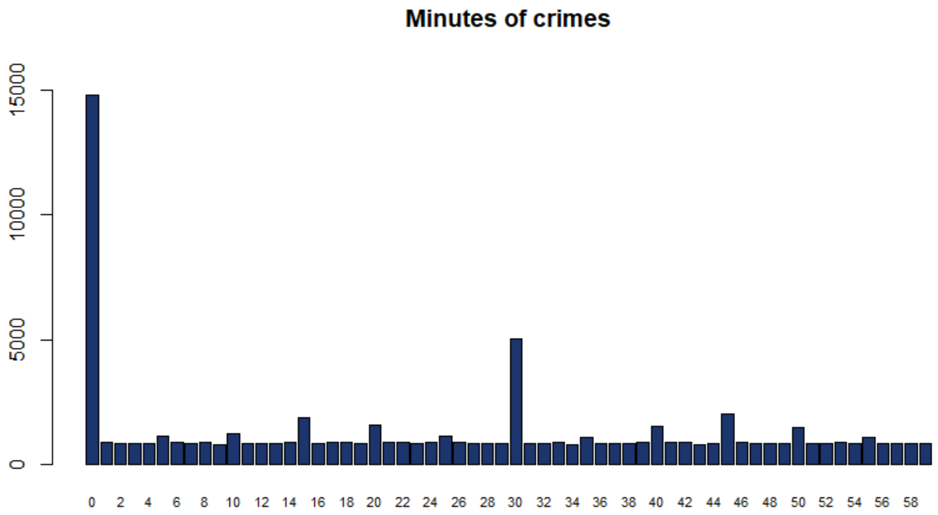
Hour

We can see that the midnight occurs most frequently and significantly surpasses other hours. This may be due to the fact that if there was no data regarding the time, they entered the respective case under hour 00:00. Apart from this anomaly, the highest number of cases was recorded during the day, the lowest during the night, where probably majority of people sleep. At 5 a.m. the crime rate starts gradually increasing, from 1 p.m. till 4 p.m. are detected some fluctuations and then at 5 p.m the rate starts steadily decreasing.



Minutes

Here, we may notice the same controversial issue,as we saw with hour and day, related to the great advantage of minute 00. What is more, when we're examining the diagram we observe that minutes which are the multiple of 5 stands out. This indicates people's tendency towards rounding off (numbers).

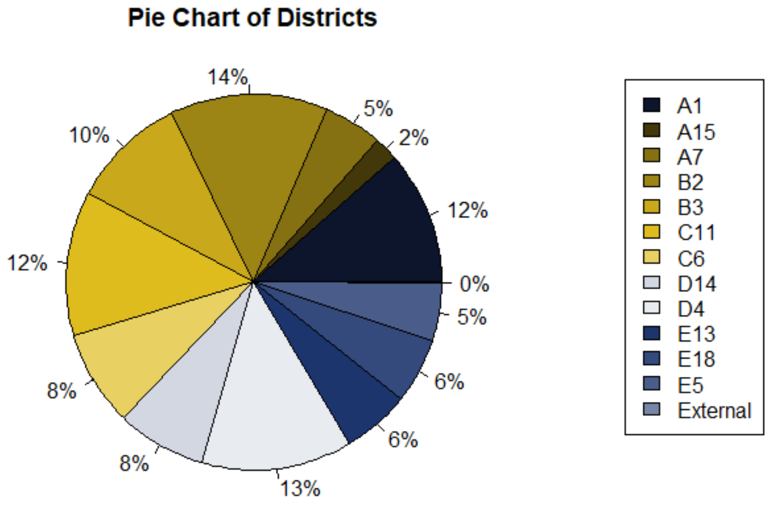


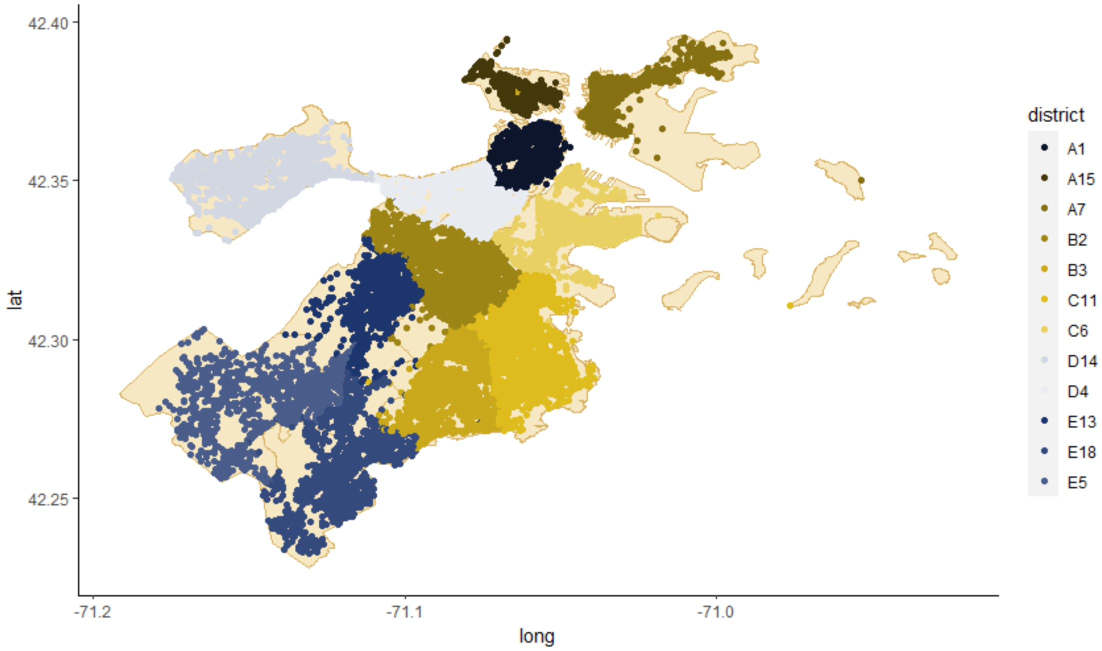
A collection of noteworthy insights regarding the location

Districts

Officially in Boston there are 23 Neighborhoods, however BPD (Boston Police Department) distinguishes only 12 districts listed below. The following pie chart presents the percentage share of each district in terms of committed crimes.

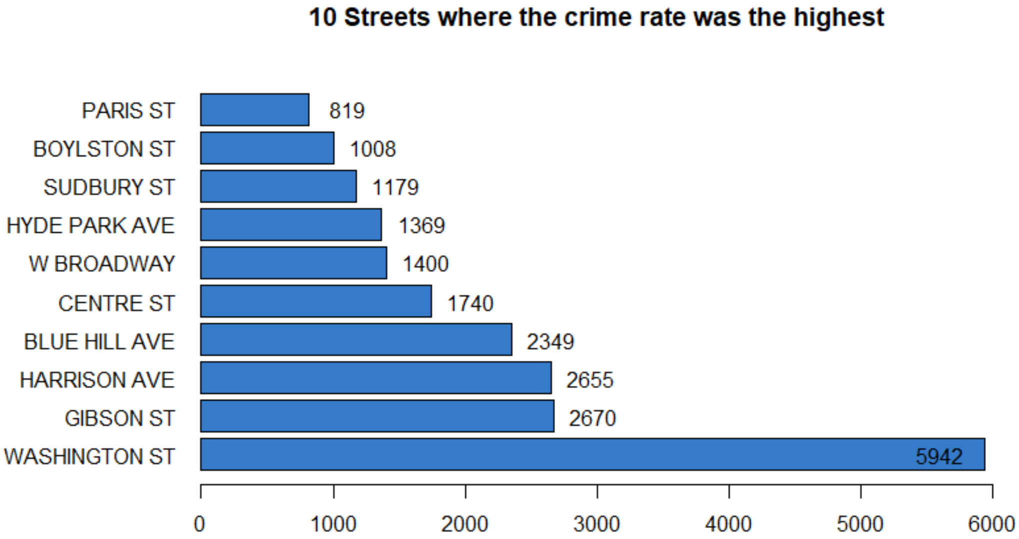
Map instead shows the position of the districts in Boston with the frequency of the delinquencies. Each dot indicates one reported case of offense. Where it is denser, there were more crimes committed. It is also worth mentioning that some data points extend beyond the boundaries of their respective districts, which could have been caused by errors in data entry or recording. Even if the geographical discrepancies are minor, when we plot the chart these tiny mistakes are highly visible.

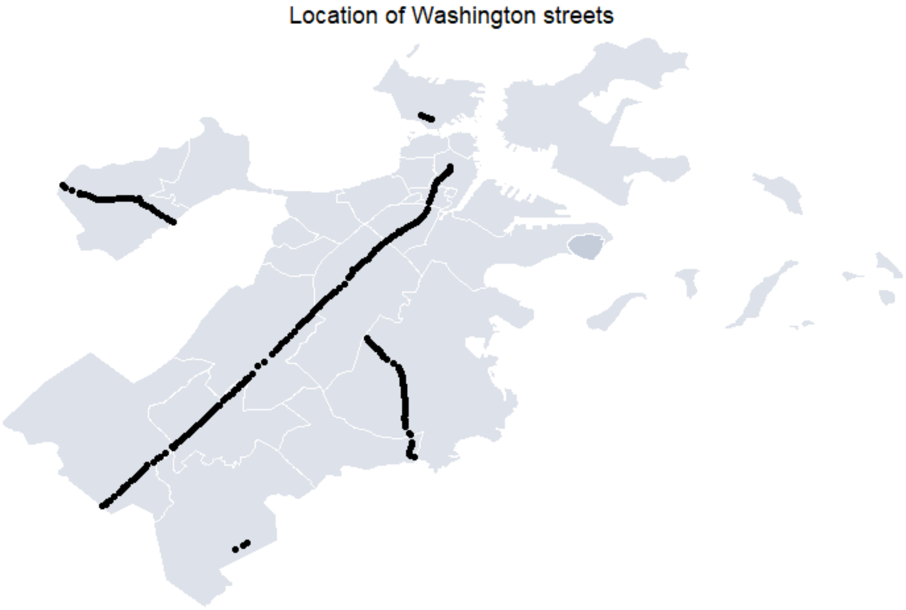




Streets

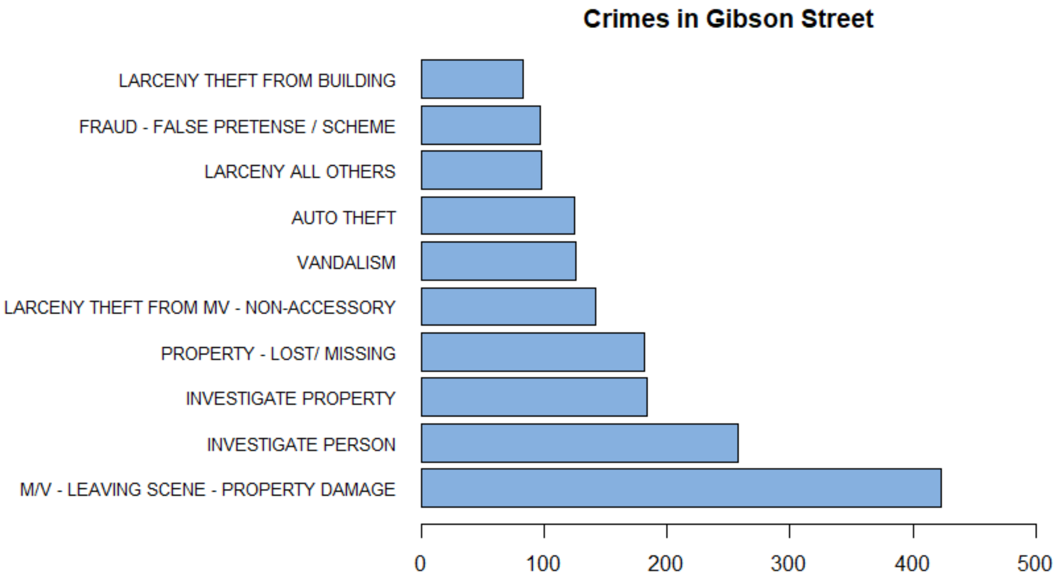
From the diagram below we infer that the most prone to crimes, the most dangerous street is Washington Street. However, as i discovered while creating the charts, there is more than on street named in the same way. This fact definitely reduces the level of the fear associated with that street.

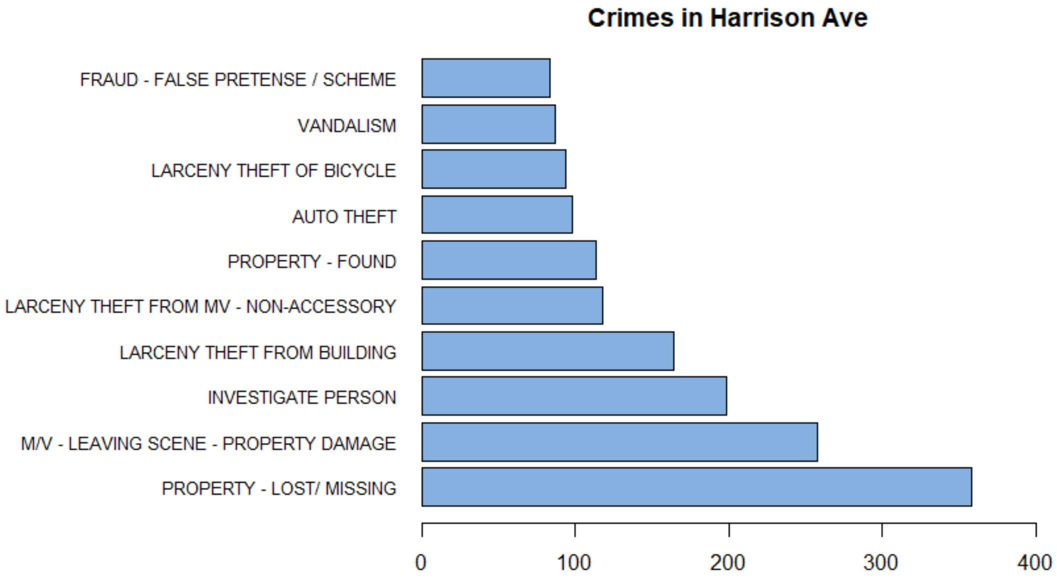


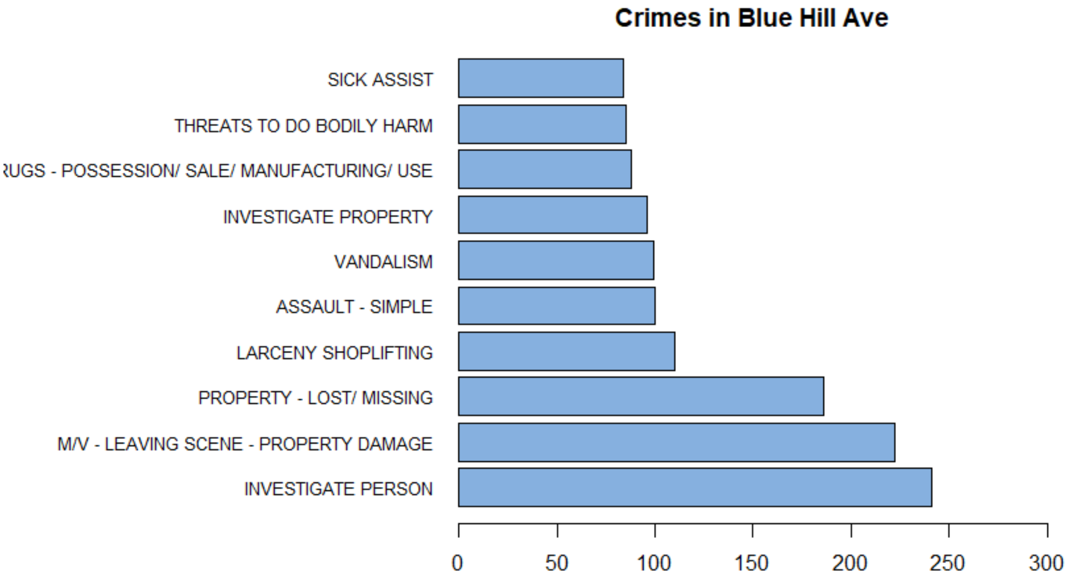


The next three places go to Gibson Street, Harrison Avenue and Blue Hill Avenue where were reported respectively 2670, 2655 and 2349 crimes. The first ten transgressions in each street are depicted on the figures below.

There, the 1st position in Gibson ST and the 2n in Blue Hill Ave and Harrison Ave is occupied by M/V-LEAVING SCENE - PROPERTY DAMAGE. INVESTIGATE PERSON and INVESTIGATE PROPERTY are also really high in the ranking, simultaneously on every street. Besides, on all three charts appear similar crimes like larceny, theft or vandalism. Long story short, streets have a lot in common with each other.

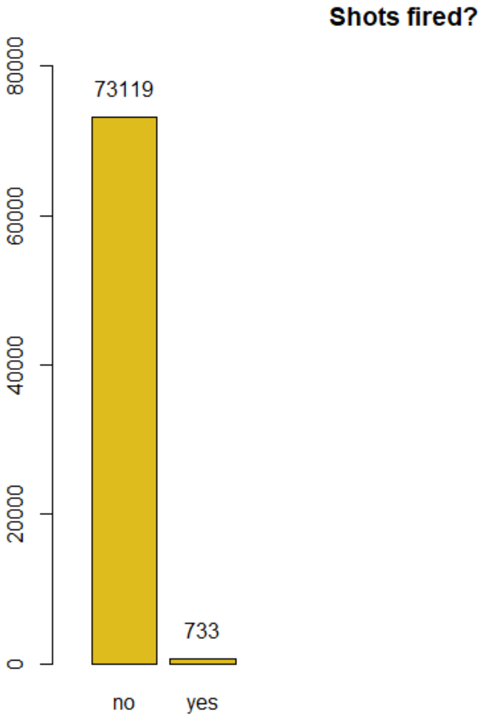




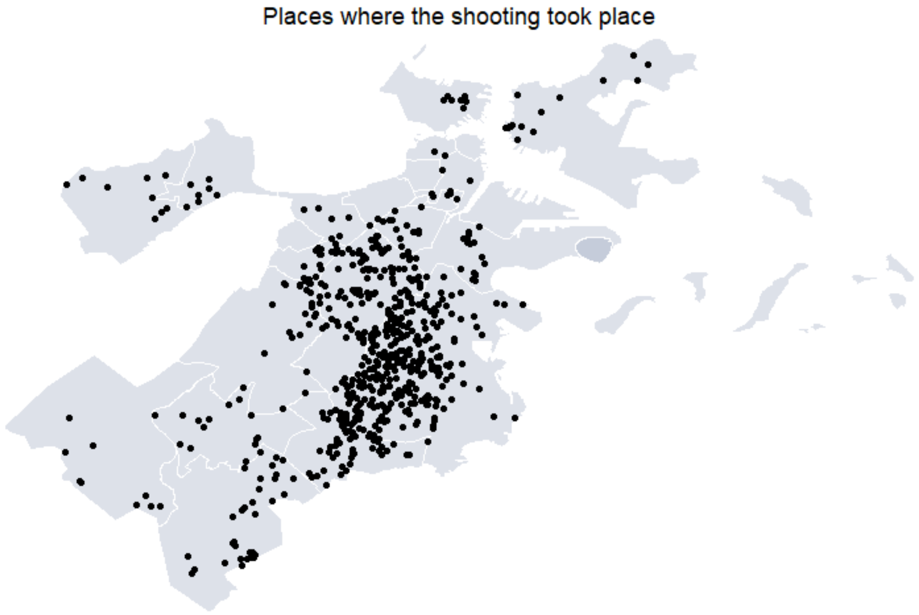


Shooting

Another interesting graph reveals how many instances of crimes were associated with the use of weapon. Although there is a huge difference in numbers, the fact that shooting appeared in 733 cases within one year is still scary, at least for me.



In addition, I attempted to mark on the map the locations where the shots were fired.

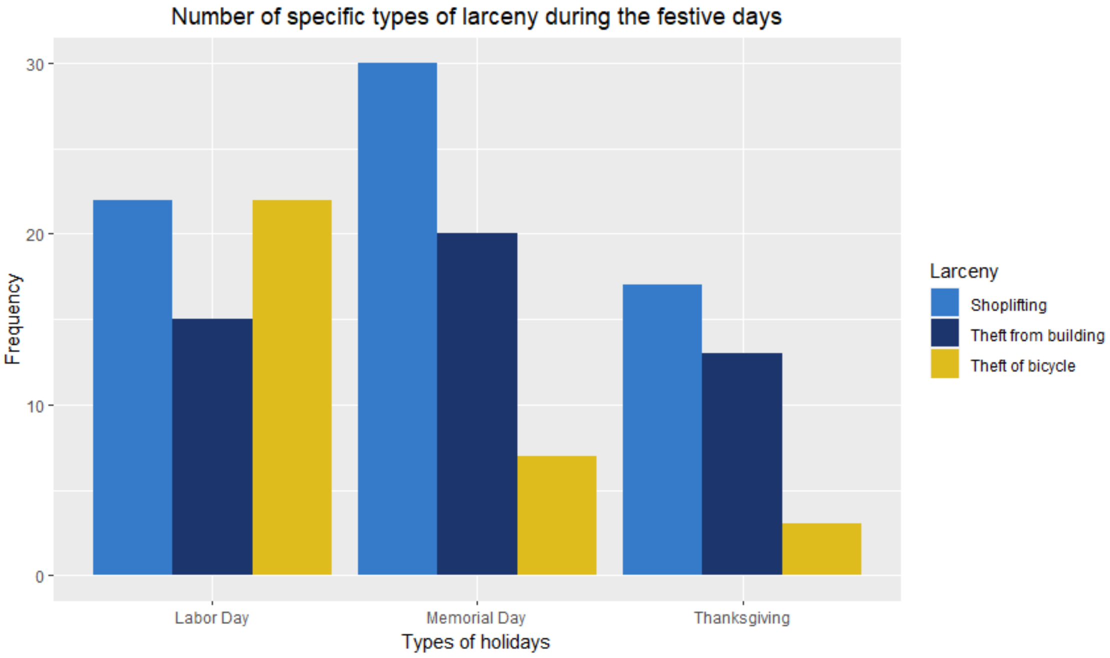


EDA 2-dimensional

In this section I will focus on analyzing the relationship between two variables. I will show the results using various charts.

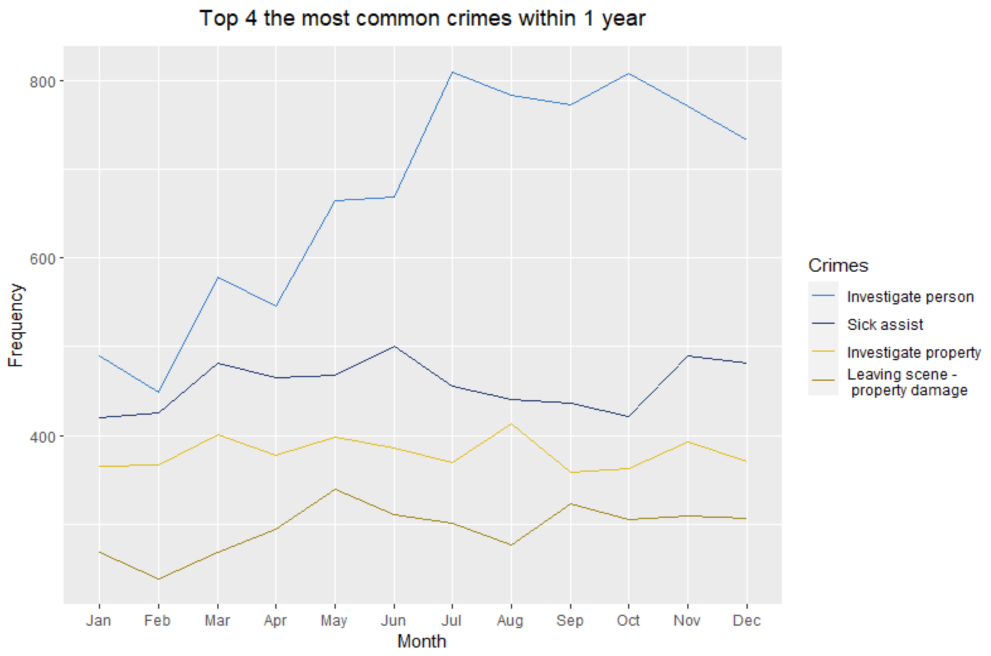
As I mentioned before, the most frequent holidays in terms of committed crimes are: Memorial Day, Labor Day and Thanksgiving. These holidays I used to create a grouped bar chart (side-by-side) which shows the relation between each holiday and a frequency of a specific type of larceny.

Shoplifting and theft from building happened the most often on Memorial Day, while theft of bicycle was reported the most frequently on Labor Day. What is surprising, theft of bicycle and shoplifting on Labor Day occurred the same number of times. In contrast, on Memorial Day and on Thanksgiving Day theft of bicycle reached less than 10 observations and it occurred the least frequently among other larcenies.

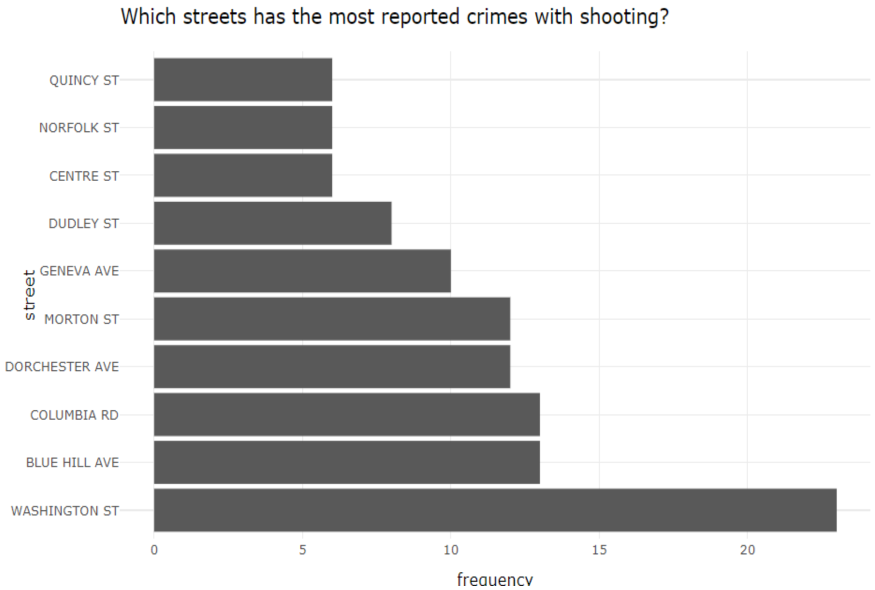


In one of the previous charts I showed that the most popular crimes in Boston are: investigate person, sick assist, m/v leaving scene-property damage and investigate property. The following diagram is a representation of time series and tell us how many delinquencies of mentioned types happened each month. Also here, has been demonstrated the pattern that most crimes are committed at the beginning of the summer and in summer.

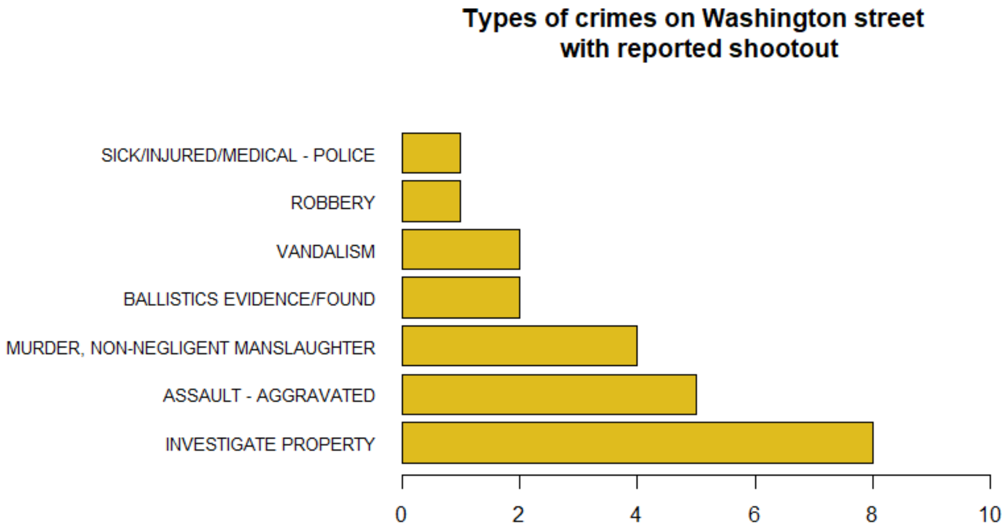
Peaks we see in May, June, July and August.



Furthermore, I created a chart showing the streets with the highest crime rates, specifically where the shootings occurred. We can read from it that Washington Street appears most frequently, however we must keep in mind the fact that this data do not relate only to one street but few.



Besides, I was also curious which kind of violations were notified to the police on that street. The results are presented over there:



Conclusions

That was really fascinating and comprehensive data set of Boston crimes 2022, from which for sure we could pull out much more than I did it. It is true that it contained lots of categorical variables, so that some exploratory methods were quite limited, however the results are satisfying after all. By conducting an analysis from scratch, I have learned how to prepare adequately data (data cleansing) and I discovered a small segment of EDA one and two dimensional. The project involved numerous traps like empty strings instead of NA, typo, treacherous data types and much more. Furthermore, some graphs wasn't as easy to construct as they may look at first sight and I dedicated hours to achieve approvable results.