Can crime incidents rate be predicted by weather conditions?

GA

Itzel A. Sadikov

Motivation

How data science can help us to study Boston crime?

- Expectation.
- Prevent.
- Be prepared.
- Study.
- Understand
- Inform.

Objective: Predict the crime rate in Boston given the weather conditions and the hour.

1. Which weather conditions are a stronger target for crime incidents?
2. How does an increase in temperature affect the crime rate?
3. Are we expecting less crime in cold days?

Model characteristics:

-Supervised linear regression model

- X : Hour and hourly; temperature, relative humidity , station pressure, visibility, wet bulb temperature, wind direction ,wind speed.

-y : Number of incident reports made by a police man by hour.

Exploratory data analysis...

Looking though…. kaggle

Boston crime incident reports from 2015 to the present by the Boston police deparment.

Crime data set:

Crime incidents reports are provided by the Boston Police department (BPD) to document the initial details for each incident reported by a police.
The data set provided contains around **400,000** incidents starting from June, 2015 until 2019.

The new incidents report system includes a reduced set of fields like;

-Dates (YYYY/MM/DD H:M:S)
-Crime code groups
-Location (Lat, Long, District)

EDA : Cleaning data Crime data set

Missing values:

- -Delete duplicate
- -Imputing missing District by matching Lat and Long.
- -Removing incidents without any location information.
- -Replacing shooting missing values as ‘ Nfound’.

National Oceanic and Atmospheric administration

NOAA's National Weather Service is building a Weather-Ready Nation by providing better information for better decisions to save lives and livelihoods

Weather data set

Weather data set in Boston start in the year 2015 until 2019.
The data set contains around 70,000 reports of **hourly weather** condition starting from January, 2015 until 2019.

With different features like:

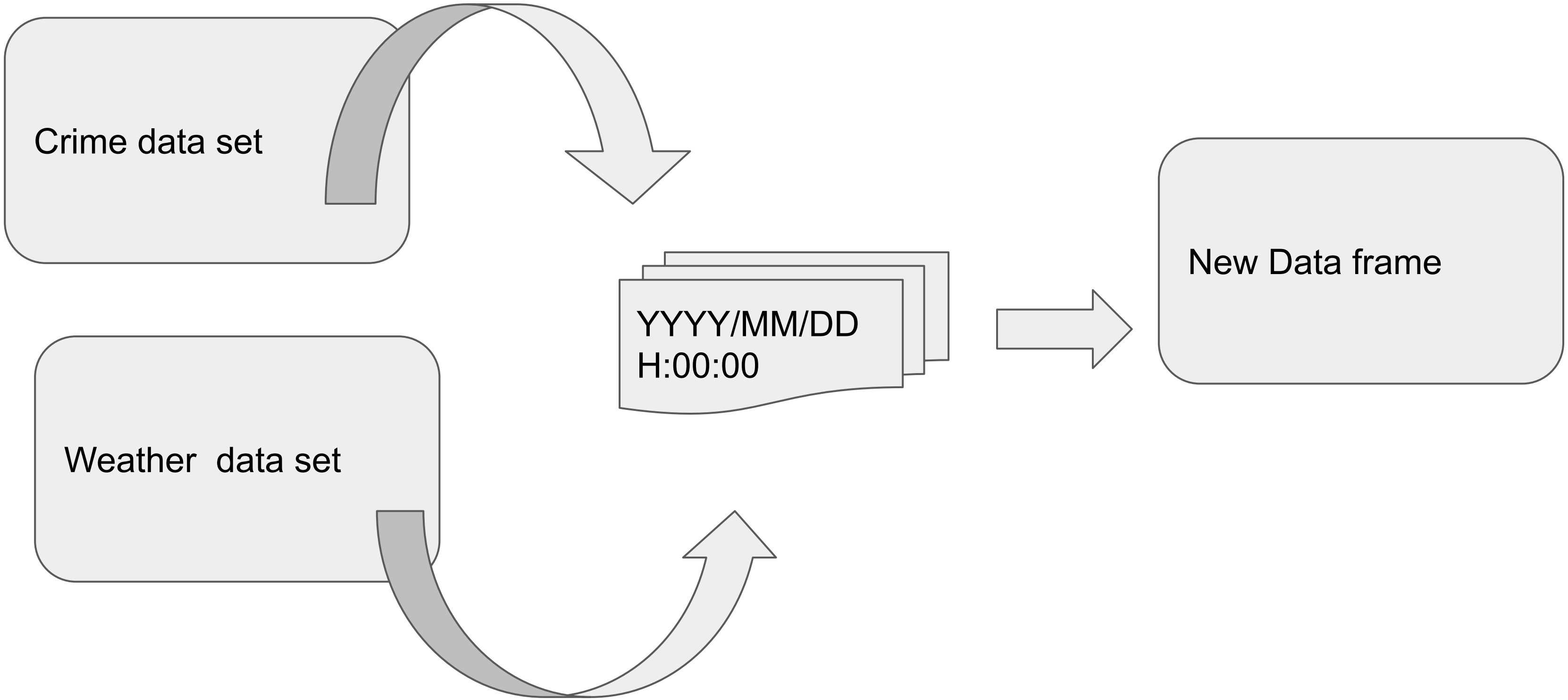
- **Date (YYYY/MM/DDT H:M:S)**
- **Station**
- **Hourly temperature**
- **Sky conditions**
- **Visibility**

EDA : Cleaning data: Weather data set.

Missing values :

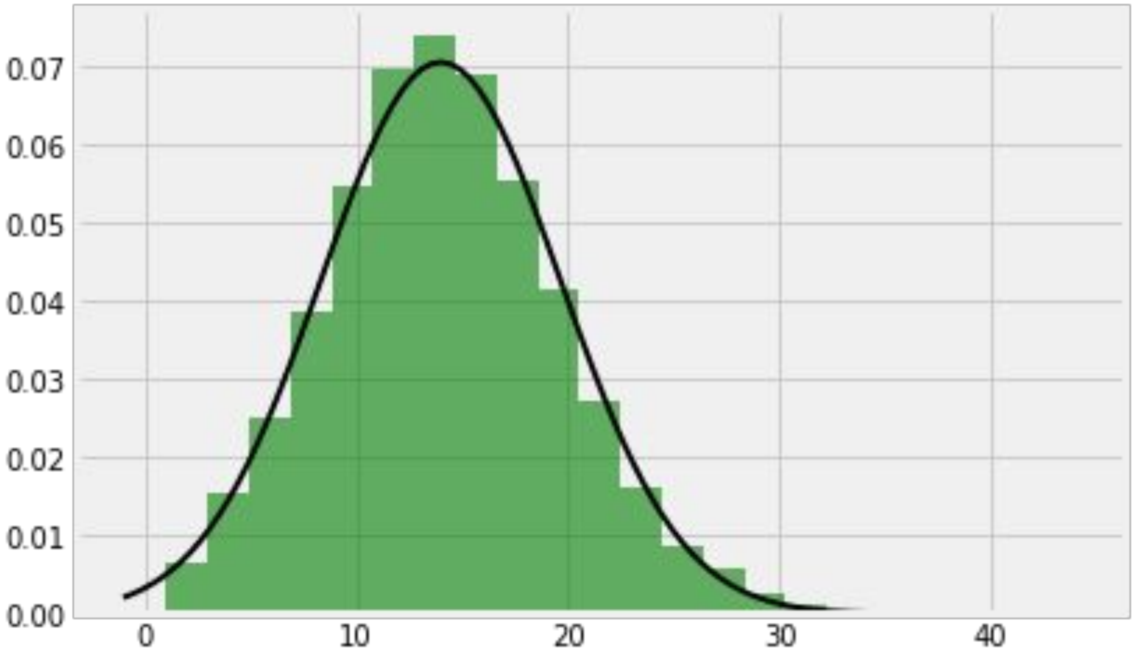
- Delete columns within the 80% of values
- Drop columns we do not really need like : sea pressure level.
- For some cases where the number of row is significantly smaller than the data set length.
- Sky conditions from a same hours impute with a for loop

Merge data frames by date and hour



Data Visualization ...

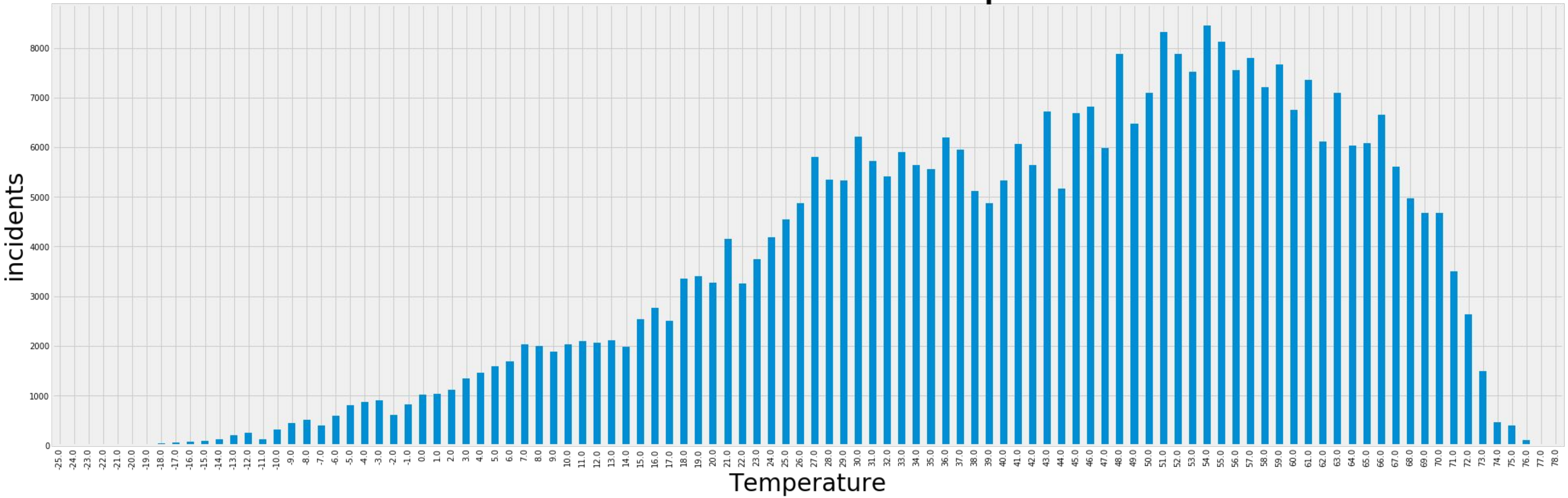
Incidents /h histogram

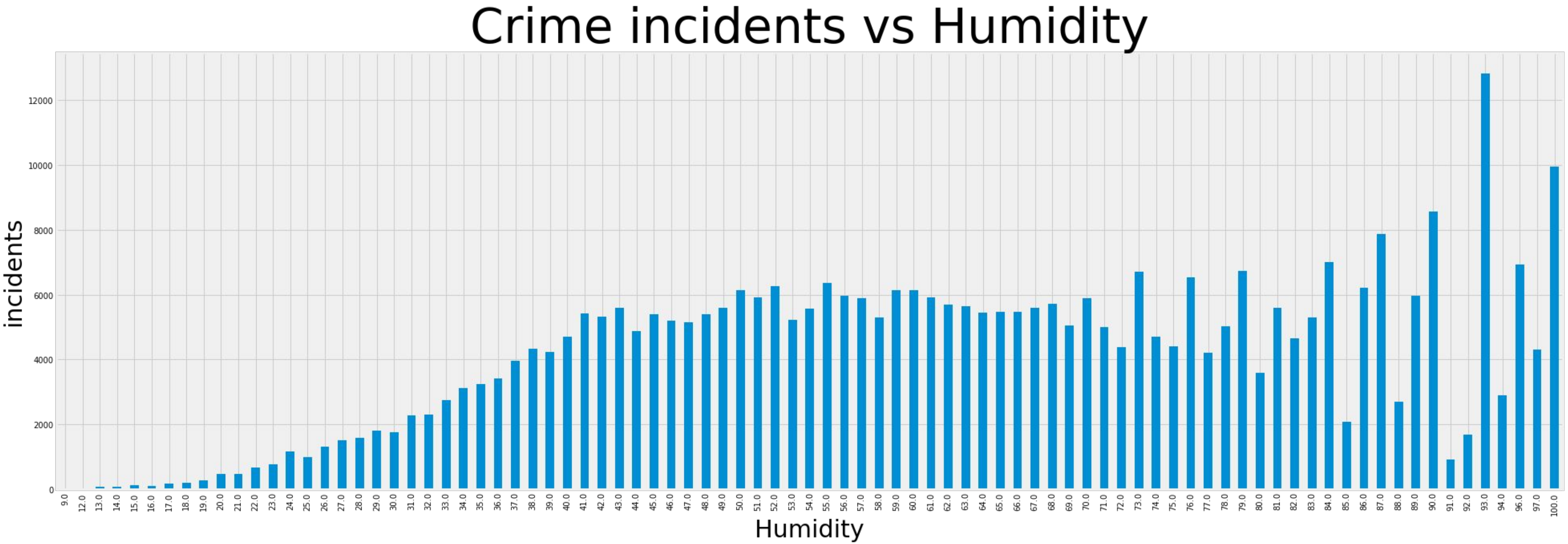


Skewness for data : 0.31302274482208015

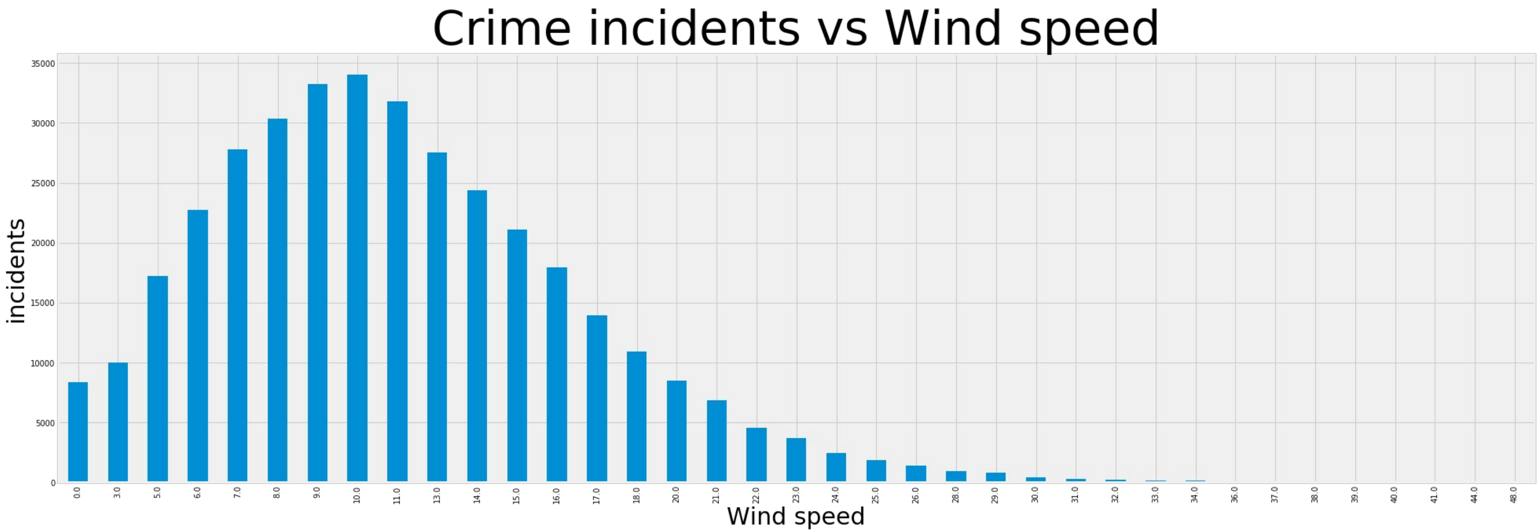
Count = 363,536
Mean = 13.97
Std = 5.65
Min = 1.0
Max = 42.00
25 % = 10.0
50 % = 14.0
75 % = 18.0

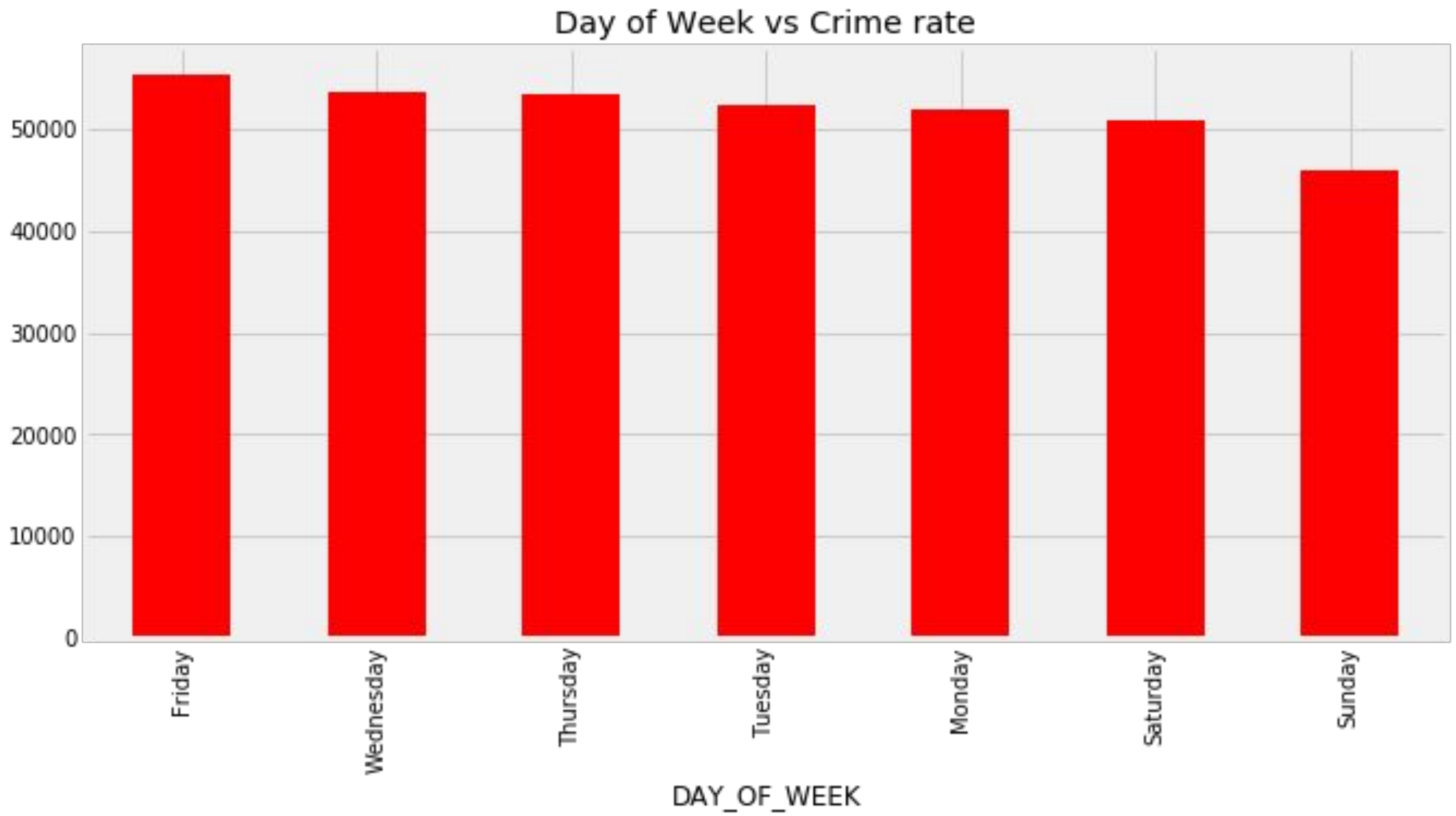
Crime incidents vs temperature

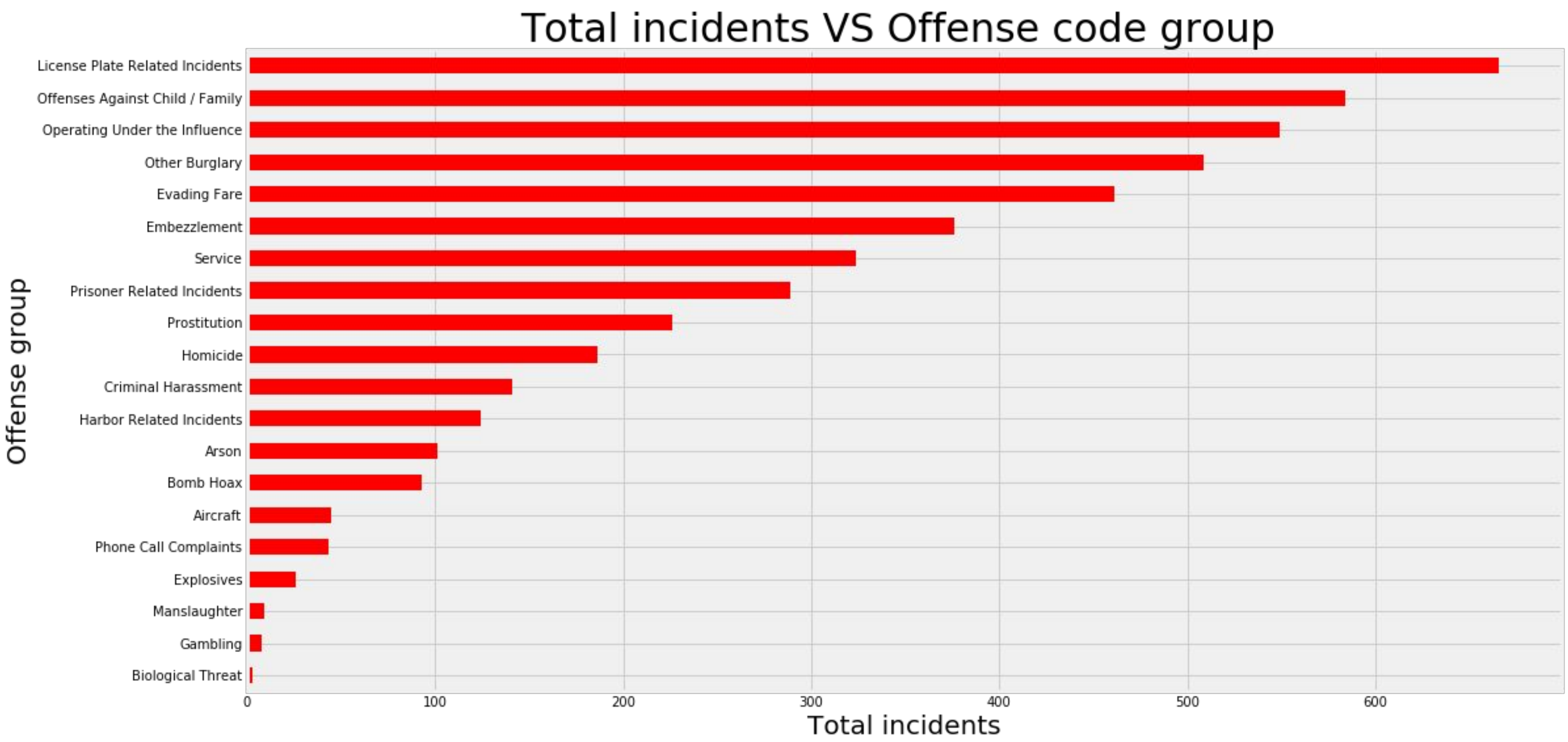


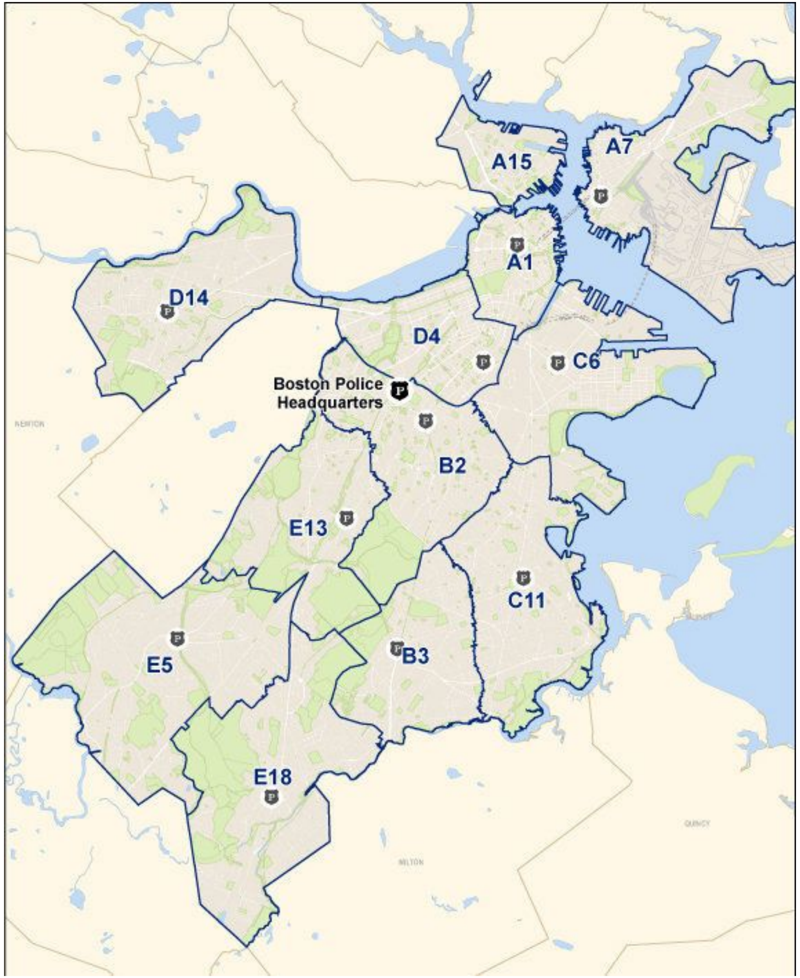


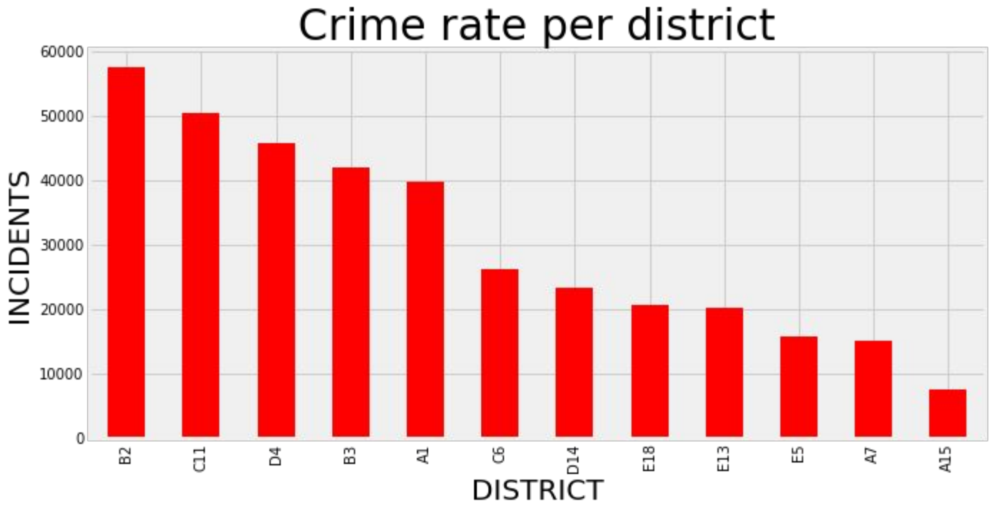
Crime incidents vs Humidity
incidents
Humidity
0
2000
4000
6000
8000
10000
12000
9.0
12.0
13.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
21.0
22.0
23.0
24.0
25.0
26.0
27.0
28.0
29.0
30.0
31.0
32.0
33.0
34.0
35.0
36.0
37.0
38.0
39.0
40.0
41.0
42.0
43.0
44.0
45.0
46.0
47.0
48.0
49.0
50.0
51.0
52.0
53.0
54.0
55.0
56.0
57.0
58.0
59.0
60.0
61.0
62.0
63.0
64.0
65.0
66.0
67.0
68.0
69.0
70.0
71.0
72.0
73.0
74.0
75.0
76.0
77.0
78.0
79.0
80.0
81.0
82.0
83.0
84.0
85.0
86.0
87.0
88.0
89.0
90.0
91.0
92.0
93.0
94.0
96.0
97.0
100.0


Crime incidents vs Wind speed
incidents
Wind speed
35000
30000
25000
20000
15000
10000
5000
0
0.0
3.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
13.0
14.0
15.0
16.0
17.0
18.0
20.0
21.0
22.0
23.0
24.0
25.0
26.0
28.0
29.0
30.0
31.0
32.0
33.0
34.0
36.0
37.0
38.0
39.0
40.0
41.0
44.0
48.0

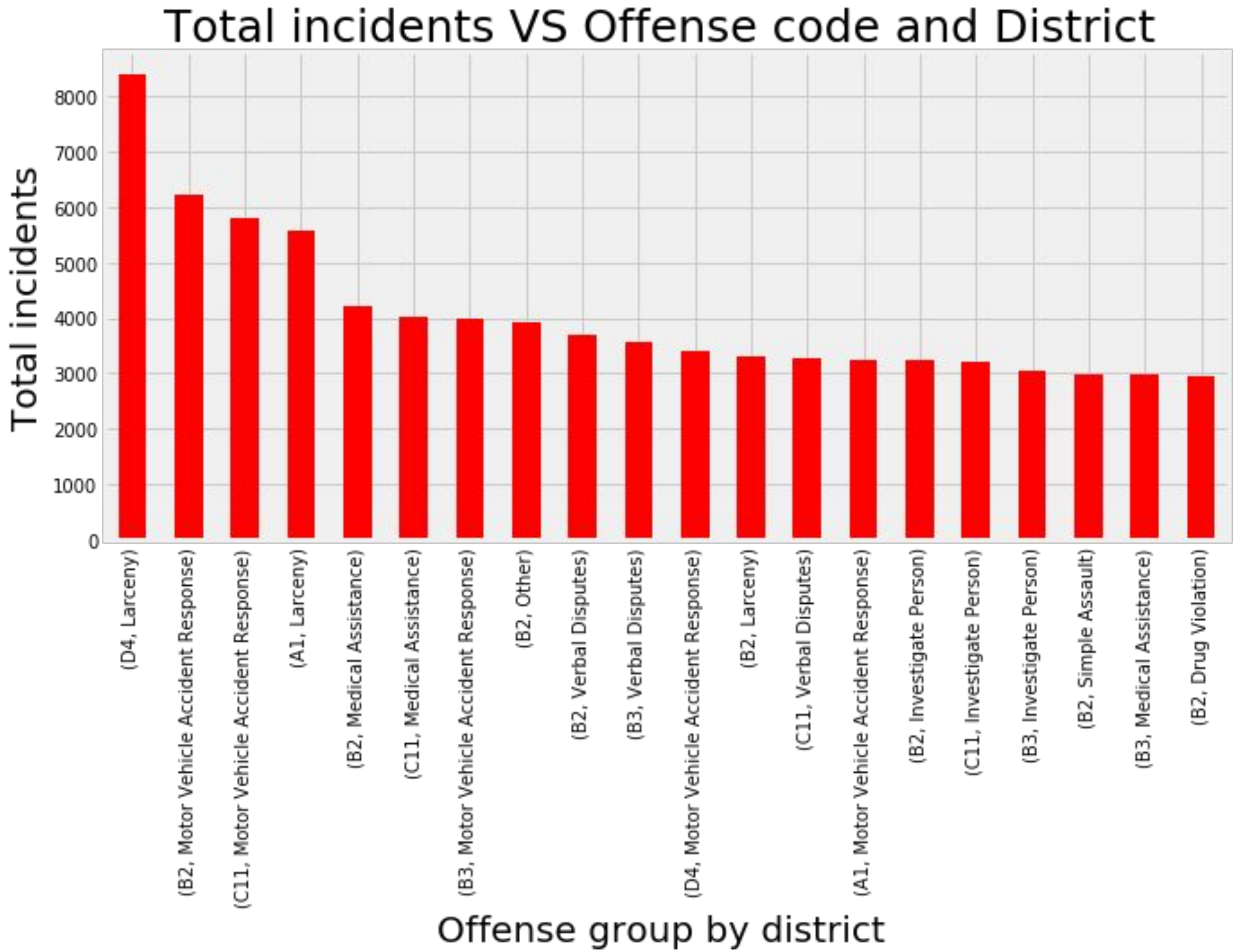

Day of Week vs Crime rate
50000
40000
30000
20000
10000
0
Friday
Wednesday
Thursday
Tuesday
Monday
Saturday
Sunday
DAY_OF_WEEK

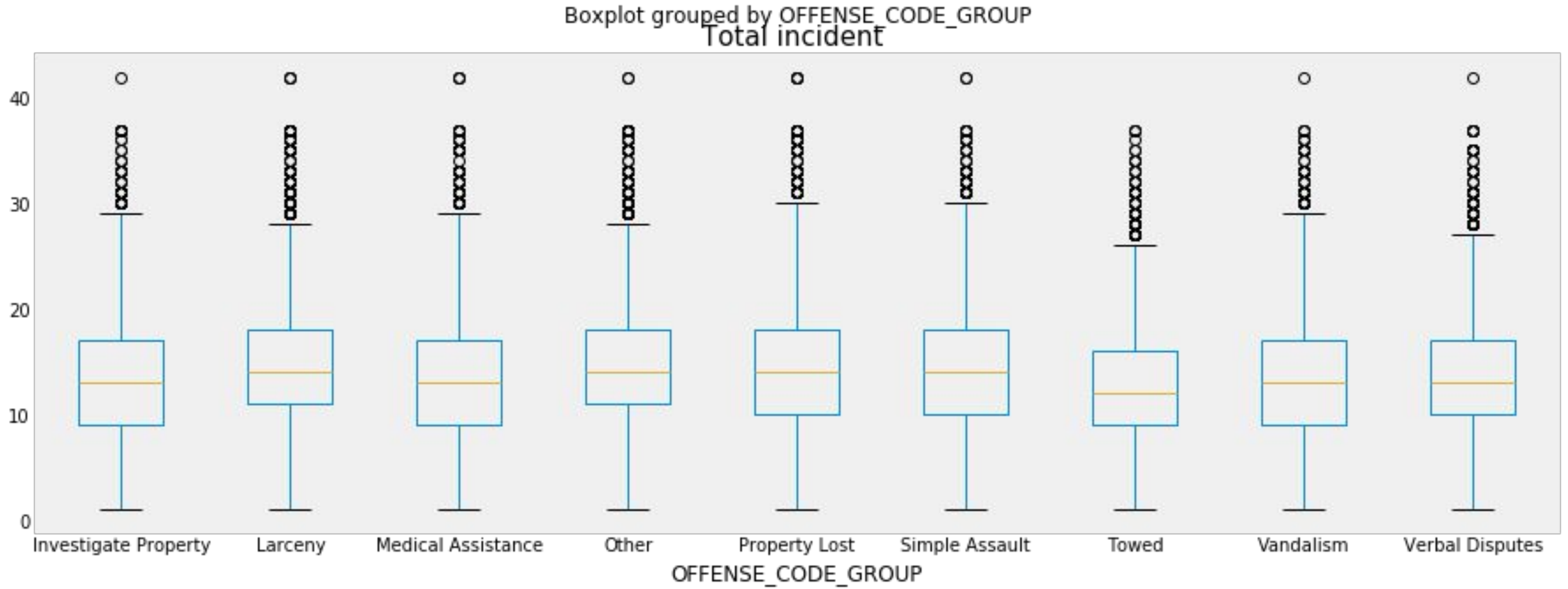

Total incidents VS Offense code group
Offense group
License Plate Related Incidents
Offenses Against Child / Family
Operating Under the Influence
Other Burglary
Evading Fare
Embezzlement
Service
Prisoner Related Incidents
Prostitution
Homicide
Criminal Harassment
Harbor Related Incidents
Arson
Bomb Hoax
Aircraft
Phone Call Complaints
Explosives
Manslaughter
Gambling
Biological Threat
0
100
200
300
400
500
600
Total incidents



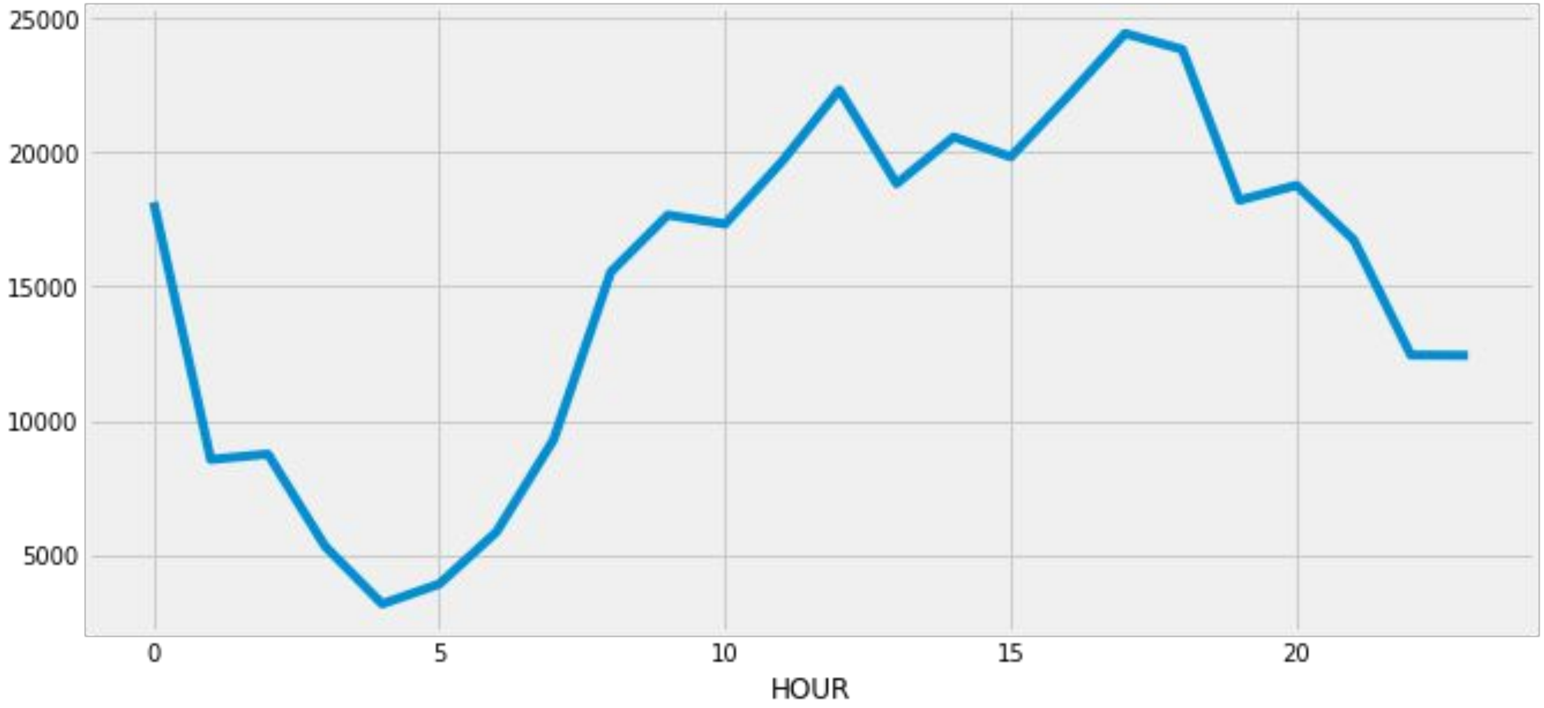


- Could district selection introduce bias?

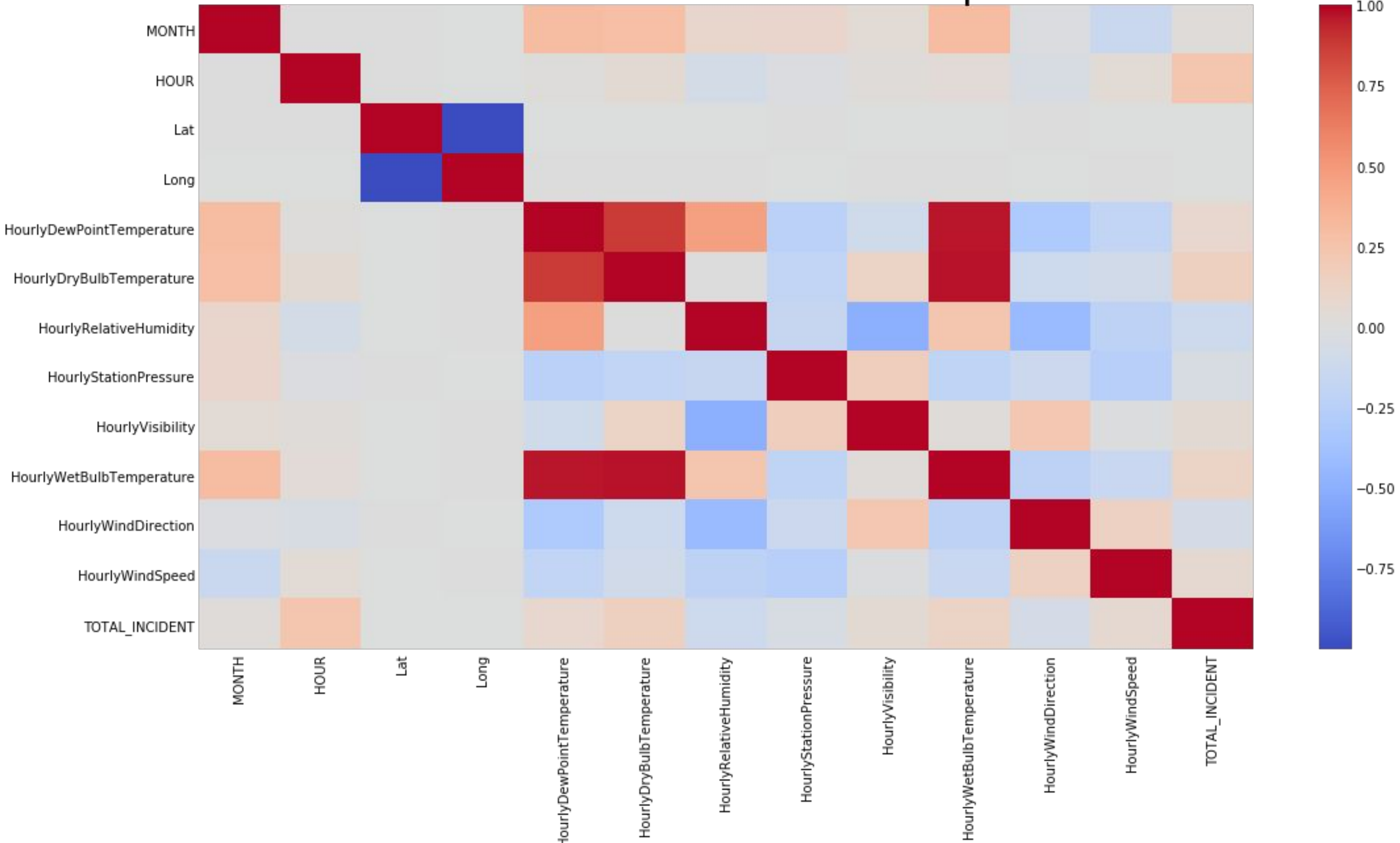

Total incidents VS Offense code and District
Total incidents
0
1000
2000
3000
4000
5000
6000
7000
8000
(D4, Larceny)
(B2, Motor Vehicle Accident Response)
(C11, Motor Vehicle Accident Response)
(A1, Larceny)
(B2, Medical Assistance)
(C11, Medical Assistance)
(B3, Motor Vehicle Accident Response)
(B2, Other)
(B2, Verbal Disputes)
(B3, Verbal Disputes)
(D4, Motor Vehicle Accident Response)
(B2, Larceny)
(C11, Verbal Disputes)
(A1, Motor Vehicle Accident Response)
(B2, Investigate Person)
(C11, Investigate Person)
(B3, Investigate Person)
(B2, Simple Assault)
(B3, Medical Assistance)
(B2, Drug Violation)
Offense group by district


Boxplot grouped by OFFENSE_CODE_GROUP
Total incident
0
10
20
30
40
Investigate Property
Larceny
Medical Assistance
Other
Property Lost
Simple Assault
Towed
Vandalism
Verbal Disputes
OFFENSE_CODE_GROUP

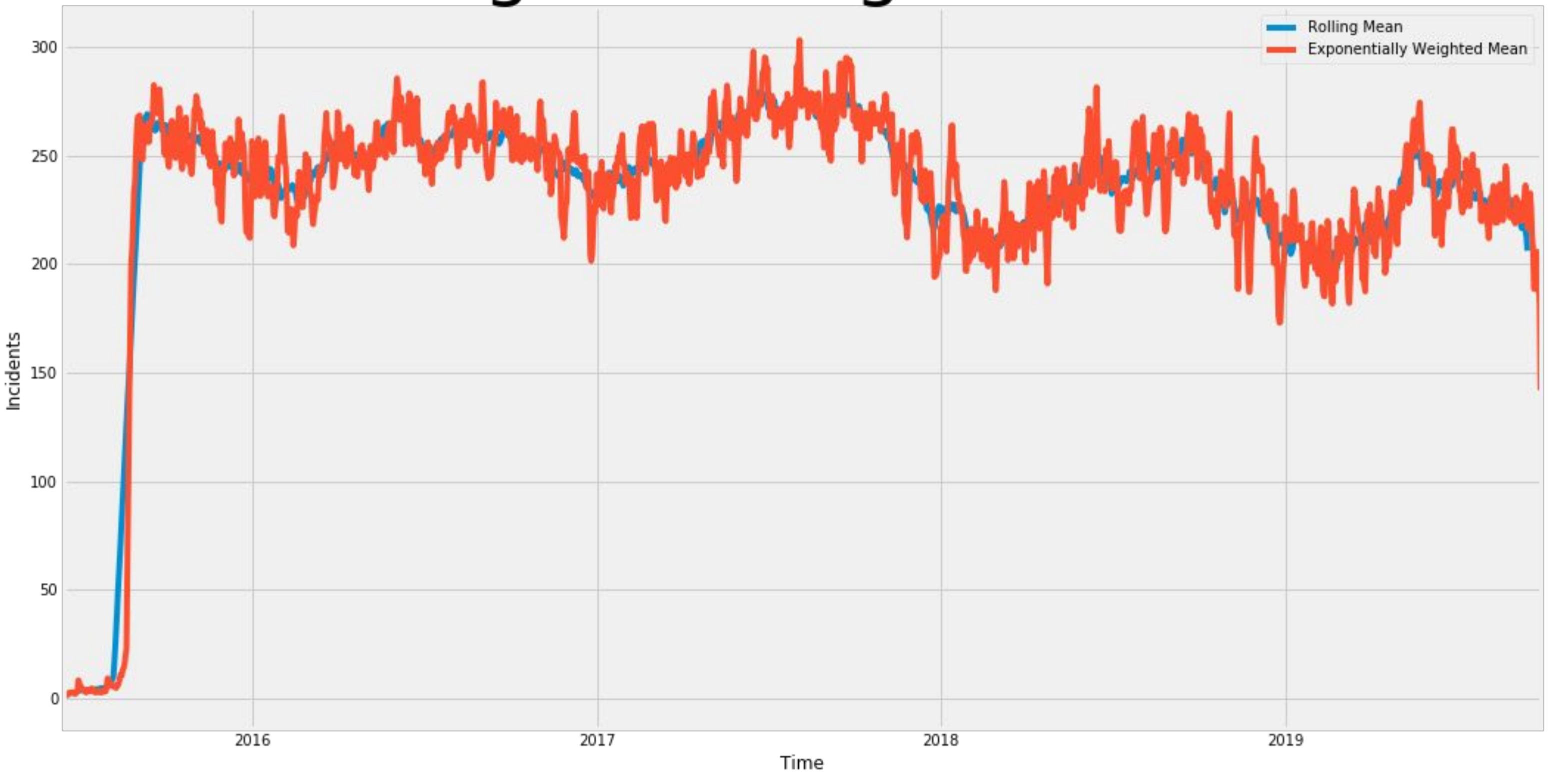
Total incidents vs Hour



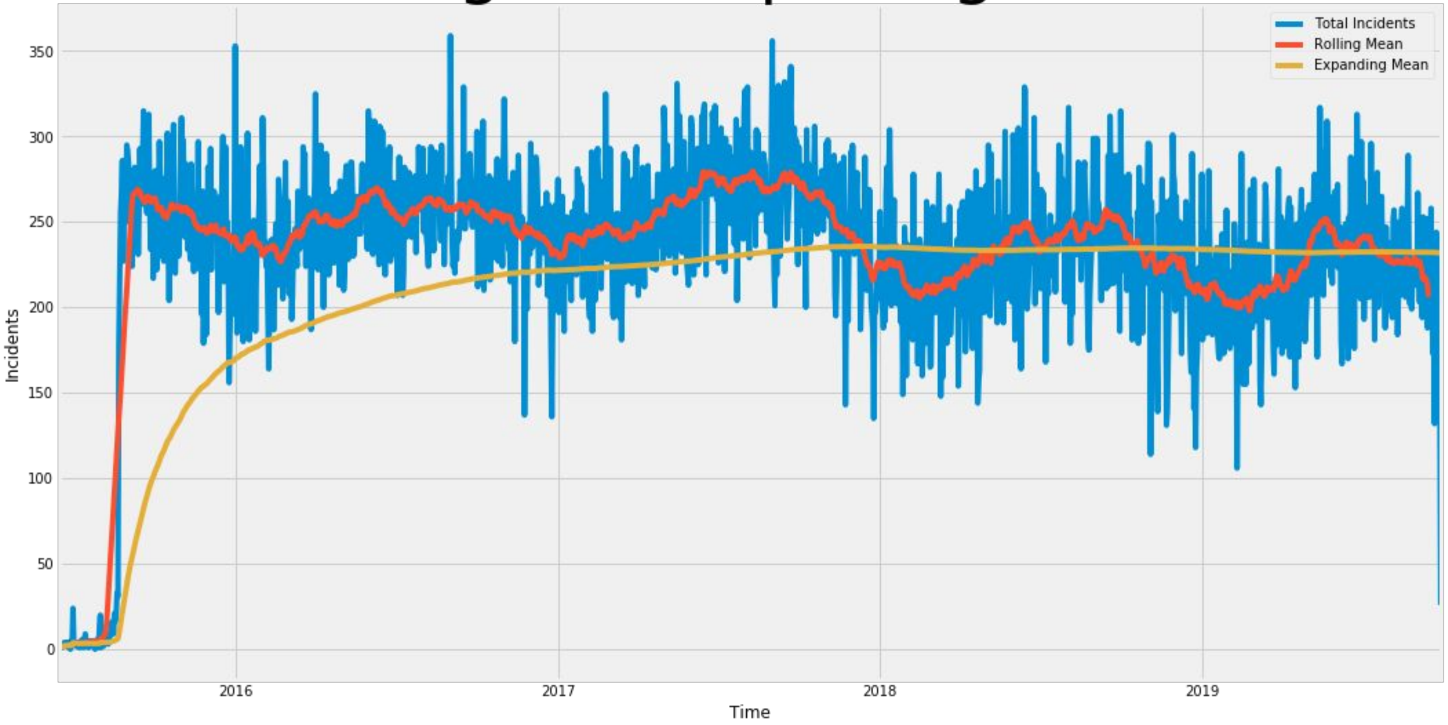
Auto correlation heat map

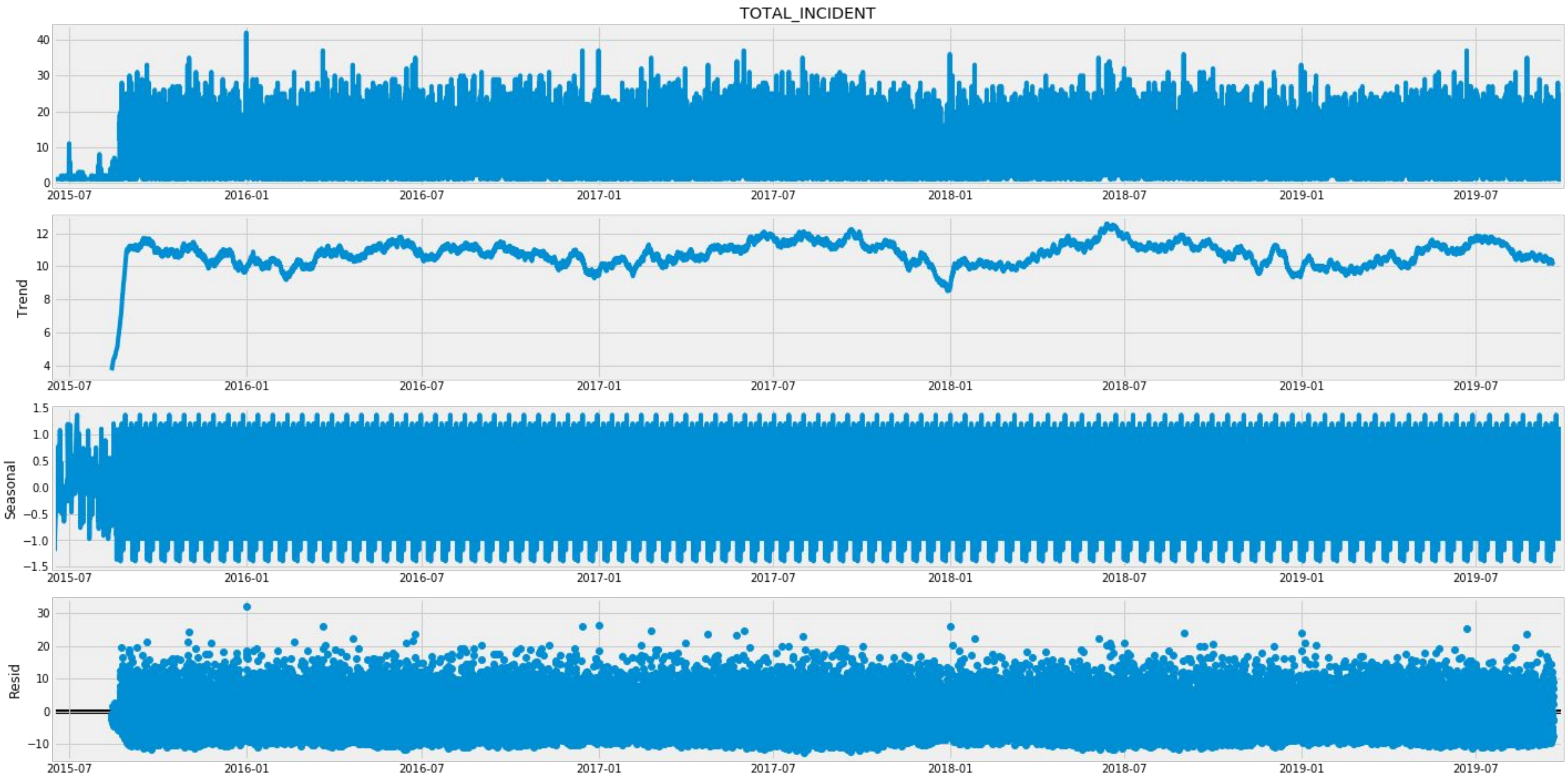


Rolling and Weighted mean



Rolling and Expaning mean

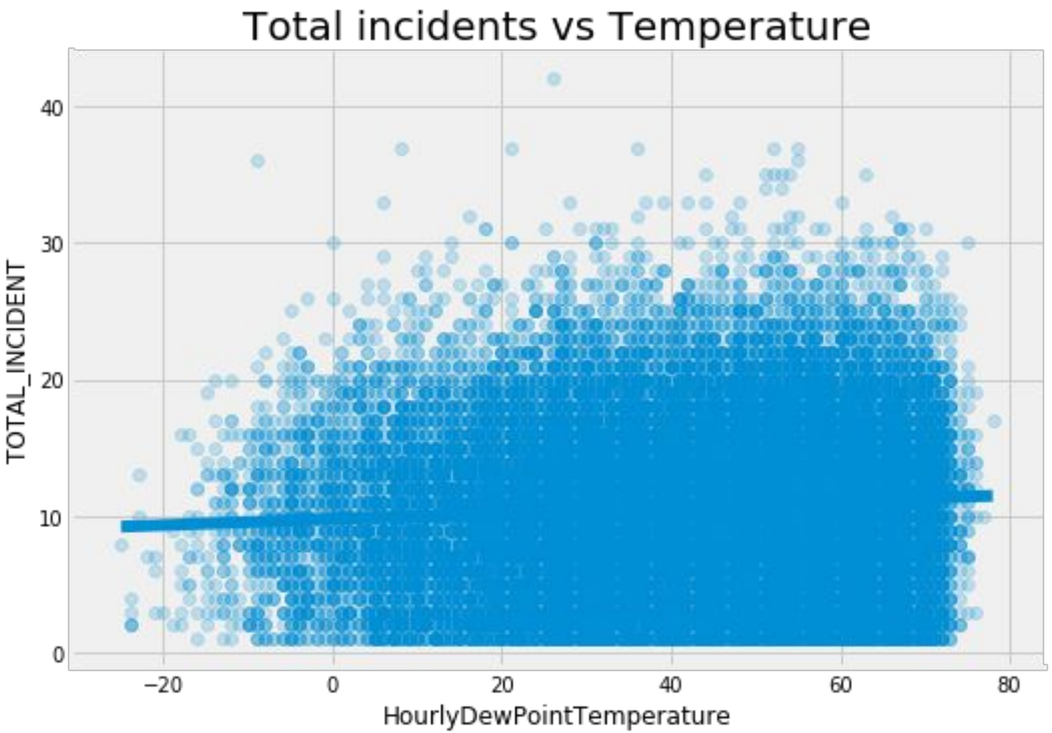




TOTAL_INCIDENT
Trend
Seasonal
Resid
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
2018-07
2019-01
2019-07

Results ...

Linear Regression fit:



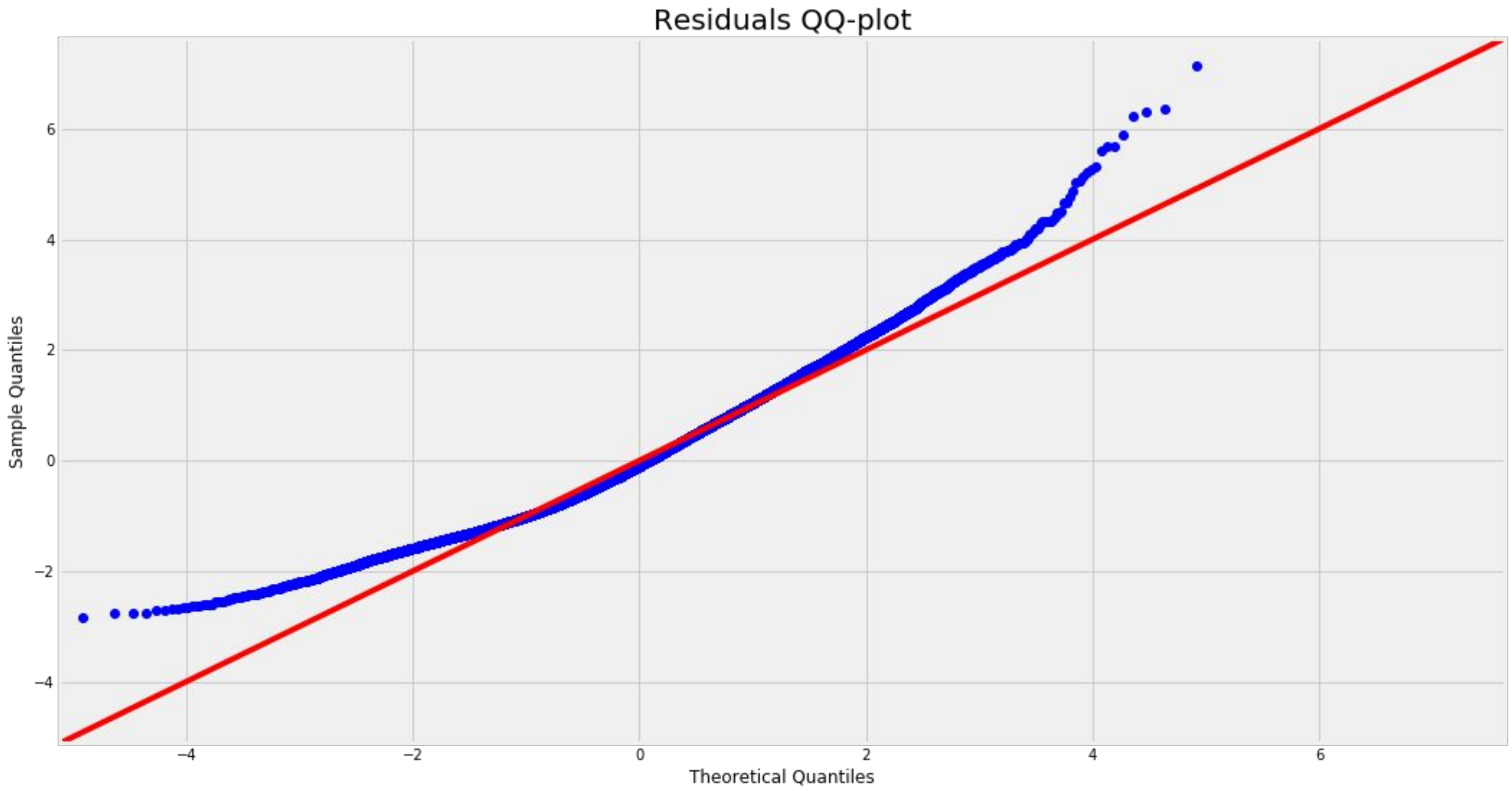

```
R-squared:                        0.213
Adj. R-squared:                   0.212
F-statistic:                      1010.
Prob (F-statistic):                0.00
Log-Likelihood:             -1.0350e+05
AIC:                          2.070e+05
BIC:                          2.071e+05
```

MAE: 4.22
MSE: 26.52
RMSE: 5.15

Null model:
MAE: 4.74
MSE: 5.81
RMSE: 5.81

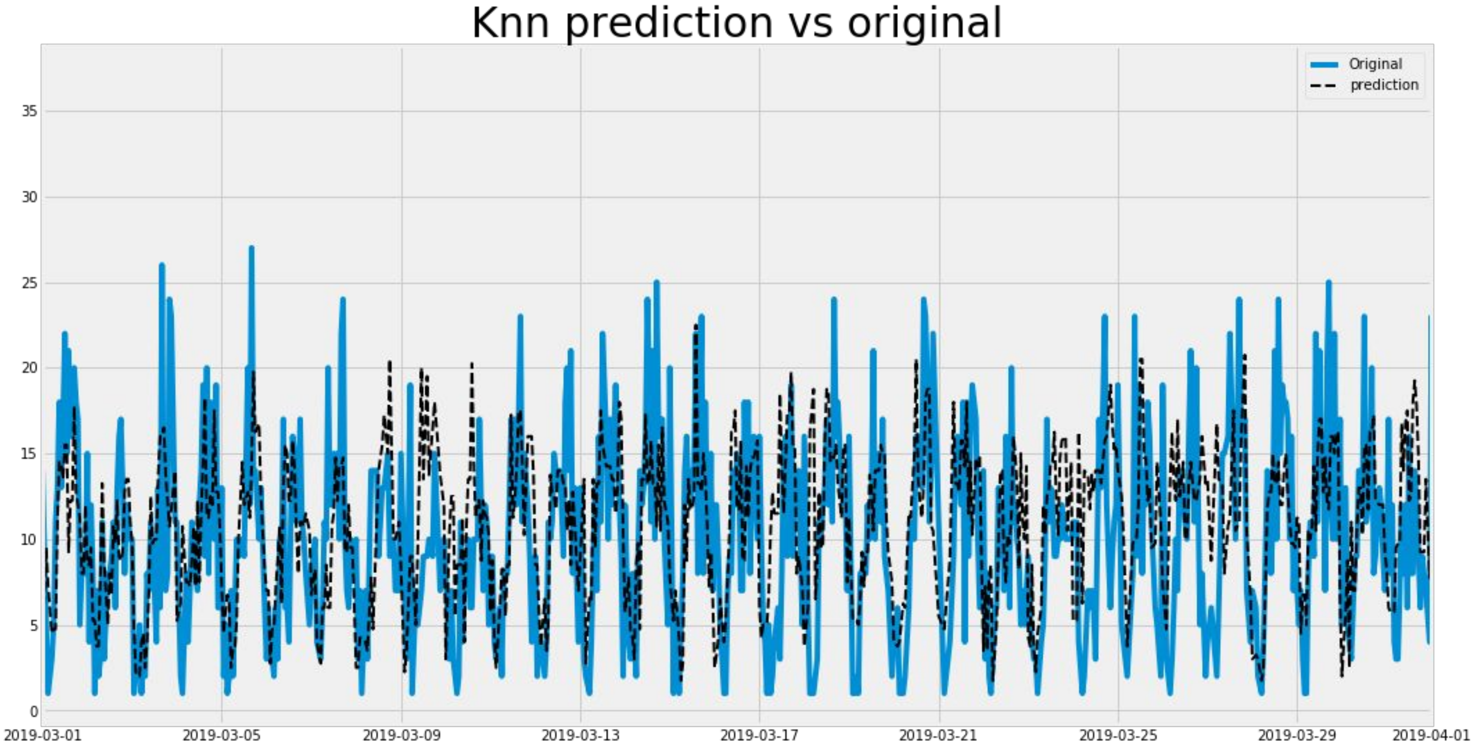
	coef	std err	t	P>|t|	[0.025	0.975]
const	27.2009	4.069	6.685	0.000	19.225	35.176
HOUR	0.2966	0.004	69.932	0.000	0.288	0.305
HourlyDewPointTemperature	-0.1033	0.023	-4.456	0.000	-0.149	-0.058
HourlyDryBulbTemperature	0.2107	0.024	8.855	0.000	0.164	0.257
HourlyRelativeHumidity	0.0007	0.009	0.079	0.937	-0.018	0.019
HourlyStationPressure	-0.7468	0.129	-5.801	0.000	-0.999	-0.494
HourlyVisibility	-0.0461	0.016	-2.949	0.003	-0.077	-0.015
HourlyWetBulbTemperature	-0.0715	0.028	-2.561	0.010	-0.126	-0.017
HourlyWindDirection	-0.0069	0.000	-22.104	0.000	-0.008	-0.006
HourlyWindSpeed	0.0593	0.006	10.267	0.000	0.048	0.071

Omnibus:	2101.238	Durbin-Watson:	1.387
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2572.318
Skew:	0.622	Prob(JB):	0.00
Kurtosis:	3.537	Cond. No.	3.54e+04



Residuals QQ-plot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−4
−2
0
2
4
6

Knn Regression fit:



n=2
MAE: 4.67
MSE: 35.91
RMSE: 5.99
R2: 0.725

n=10
MAE: 4.140
MSE: 26.54
RMSE: 5.15
R2: 0.40

n=30
MAE: 4.11
MSE: 25.66
RMSE: 5.066
R2: 0.32

Conclusion

- Knn models had better results than LR.
- Work with specific offense may improve results.
- Clusters regression models may improve our prediction.
- We need to collect more data.

References:

Crime data set :
https://www.kaggle.com/zer0state/crime-incident-reports-august-2015-to-date

Weather data set :

https://www.noaa.gov/weather

Help :

https://stackoverflow.com

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html