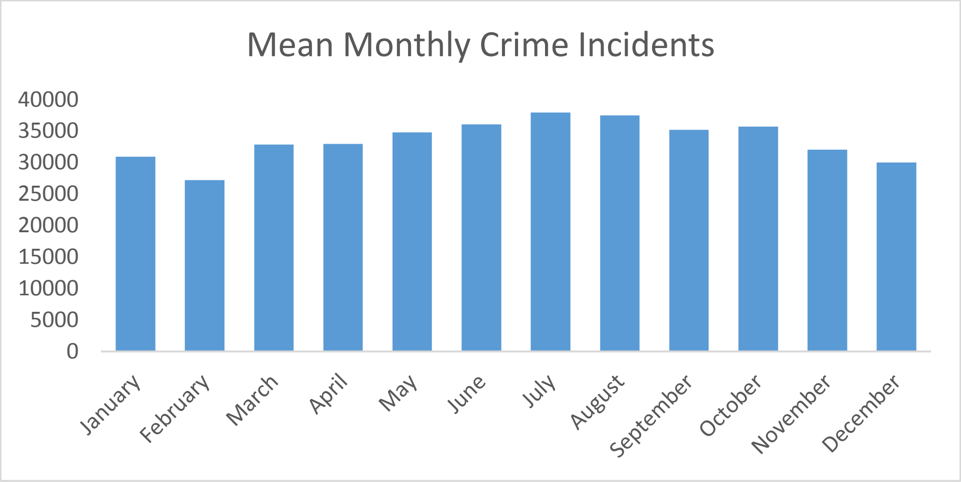
MSIA 431 Analytics for Big Data

Homework 5

Ameer Khan

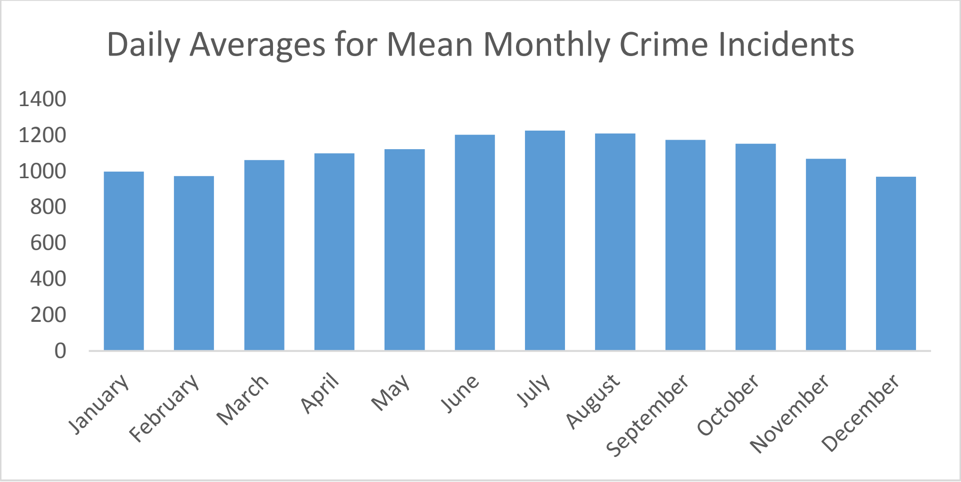
Problem 1

The bar chart of mean crimes incidents by month from 2001 to present for the city of Chicago is given below:



The crime incidents peak in the summer months. The probable reasons for the spike in crimes during the summer months is warmer weather, and school/college holidays that allow bored juveniles to engage in criminal activity. Another reason for increased crime could be increase in social interactions during the summer months that can eventually lead to reported crime incidents. This article discusses the seasonality of crime incidents in Chicago.

February has the lowest crime incident counts since it has shorter days than other months. This can be adjusted by visualizing the average daily value for the mean crime counts for each month, removing the effect of different number of days in a month. The bar chart for the adjusted values is given below:



Problem 2

a. The top ten blocks by reported crime incidents are given below:

Block	Crime Incidents 2012 – Present
001XX N STATE ST	2,279
0000X W TERMINAL ST	1,896
008XX N MICHIGAN AVE	1,744
076XX S CICERO AVE	1,541
0000X N STATE ST	1,275
064XX S DR MARTIN LUTHER KING JR DR	952
040XX W LAKE ST	908
008XX N STATE ST	899
051XX W MADISON ST	872
009XX W BELMONT AVE	822

Most of the top blocks (7) with reported crime incidents are in the Loop area, and some in the South Side (2), and one near the Skokie area. The downtown areas are the most densely populated parts of the city, and are heavily policed as well. Hence, it's logical that most of the reported crime incidents are from downtown blocks.

b. The highest correlated beats were computed by first counting the number of crimes that occurred in each of the 304 beats of Chicago for all years (2011 – 2015) in the data. There are a few missing beat-year combinations, and there were added to the total counts, with zero count. The correlation matrix obtained is a square matrix with $n = 304$. The 5 highest correlation values from the matrix apart from the diagonal are extracted, and reported below:

Top Adjacent Beat Pairs by Crime Incidents Correlation

Beat Pairs		Correlation
1934	1935	0.99973576
1221	1215	0.99967537
1925	1934	0.99955465

The adjacent beat pairs, obtained from the beat map of Chicago, are given above. Beats 1925, 1934 and 1935 form a contiguous region in 19th district along the lake shore. The high correlation indicates crime activity in these beats has been changing the same way over the years.

c. In order to compare the overall crime rate under mayors Richard Daley and Rahm Emmanuel, a paired t-test is used. This test measures if there is significant difference in the number of crime events between Daley's and Emmanuel's terms, by checking the significance of the difference between the mean crime events for each, across districts. The comparison is done at the district-level for yearly crime rates. First, the counts of crimes for each district are obtained for Daley's period (2001 – May 2011) and Emmanuel's period (June 2011 – present). Counts for erroneous districts are filtered out from the data (such as district 31). The number of crime incidents by district is divided by the number of fractional years the data spans for both mayors, to obtain the average yearly crime events for all 22 districts for both the mayors. The mean of the difference between the district crime rates for Daley and Emmanuel is $\mu_{diff} = 844.1036$, and the sample standard deviation is $s_{diff} = 777.5282$. The t-score for the differences is

5.0920 for $n = 22$, which has a p-value $< 10^{-4}$ and is significant at the 0.01 level for a two-tailed test. This indicates that crime has fallen across Chicago districts during the Emmanuel years as compared to Daley's period.

Problem 3

In order to predict the crime events in the city at the beat-level, a regression random forest model was used. The model was trained on the number of crime events that occur in each beat for a given year-week. Historical temperature records for the city of Chicago were augmented to the dataset. The temperature records were aggregated up to the week level. The temperature data for Chicago was obtained from Prof. John Kissock's website at the University of Dayton. The data can be obtained from here.

The random forest model generated consists of 5 separate decision trees (weak learners) with a maximum depth of 10. The mean squared error of the model is 78.32, and the R-squared is 0.6390. Predictions were also obtained about the crime events predicted for the next week (June 29 to July 5) in each beat, and the top 10 beats are reported below:

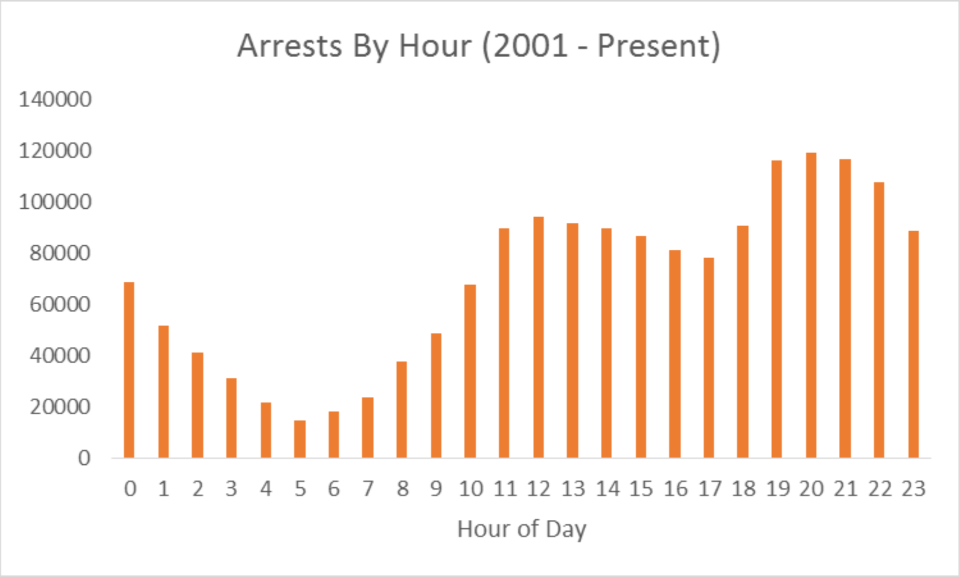
Beats with Highest Predicted Crime Events Next Week
0423
0421
0823
1834
0624
0414
0321
1533
1522
0511

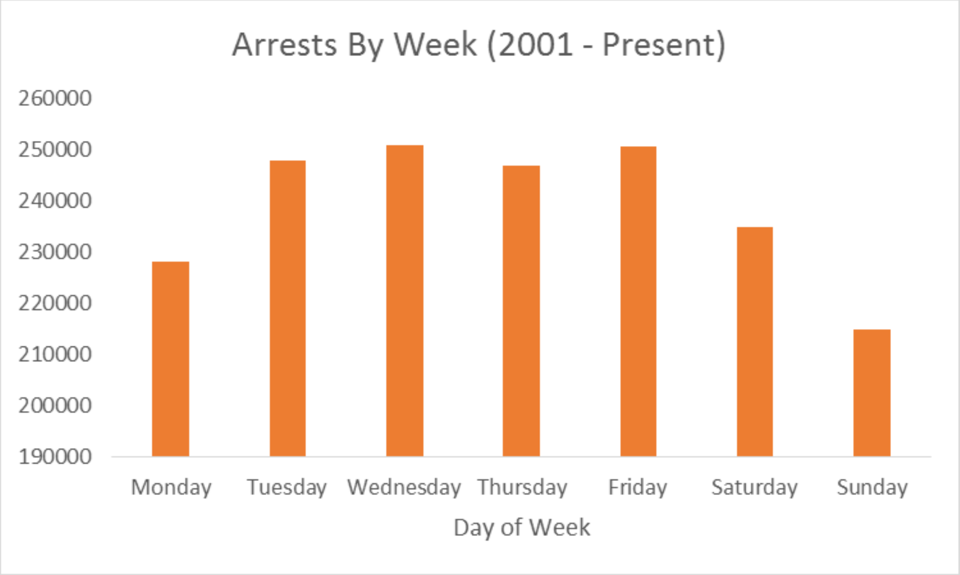
Problem 4

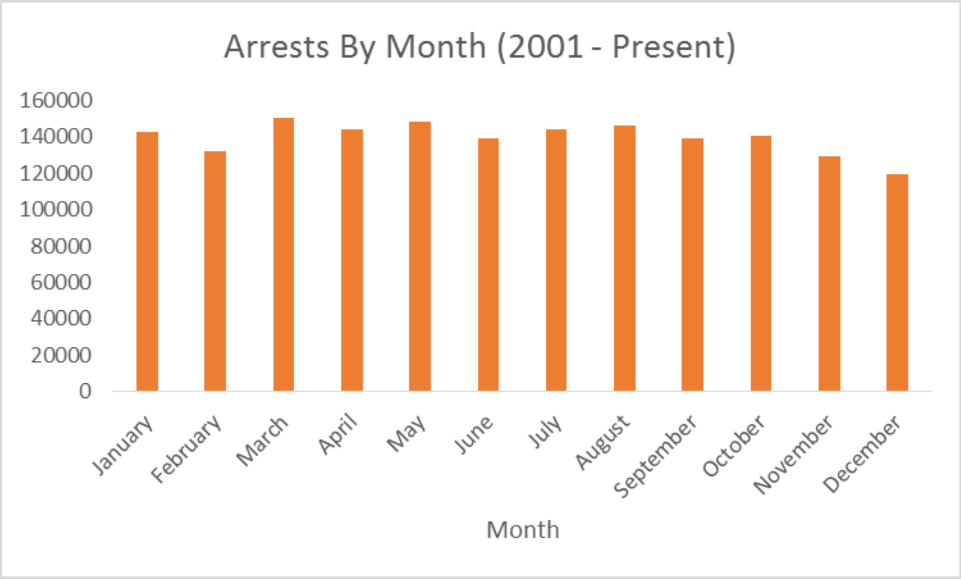
Patterns for arrests were analyzed by hour of day, day of week, and month, for the time period 2001 – present. Special focus was also on homicide and robbery crimes.

Overall Arrests:

It looks like most arrests are made during night time between 7pm and midnight, while the least are during the early morning hours from 2am to 8am. More arrests are made during weekdays, than on weekends, with arrests being the lowest on Sundays. Summer months show a higher arrest rate that other months.

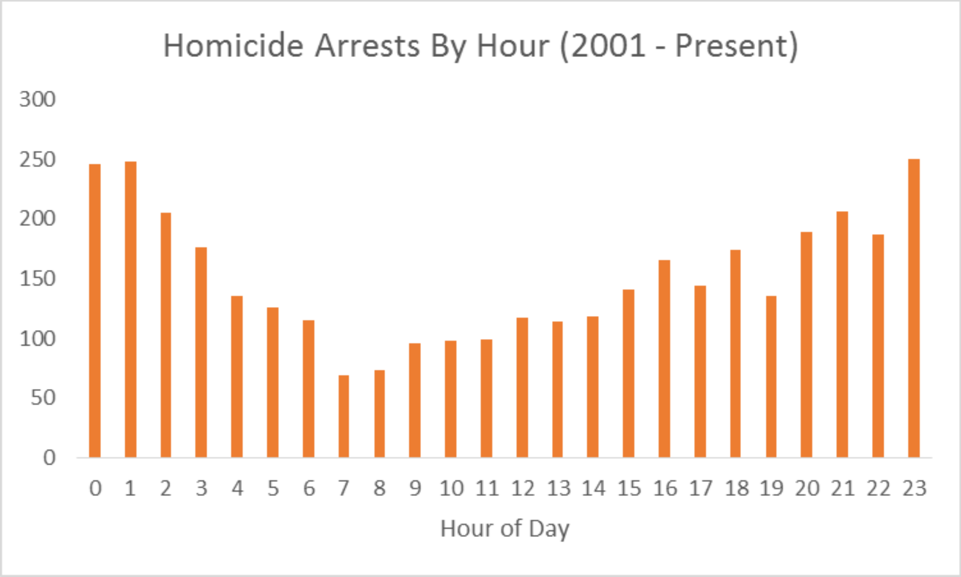

Arrests By Hour (2001 - Present)
140000
120000
100000
80000
60000
40000
20000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of Day

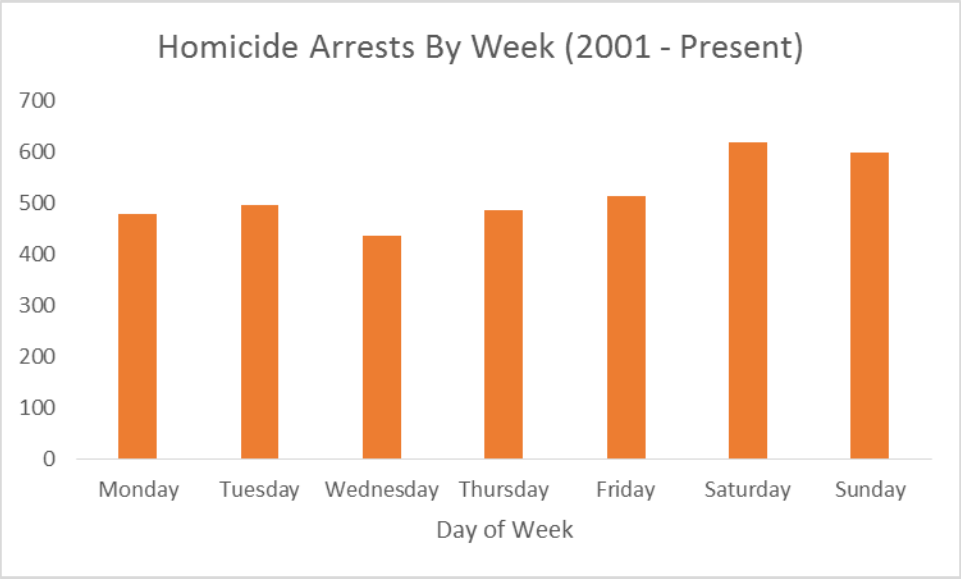

Arrests By Week (2001 - Present)
260000
250000
240000
230000
220000
210000
200000
190000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of Week

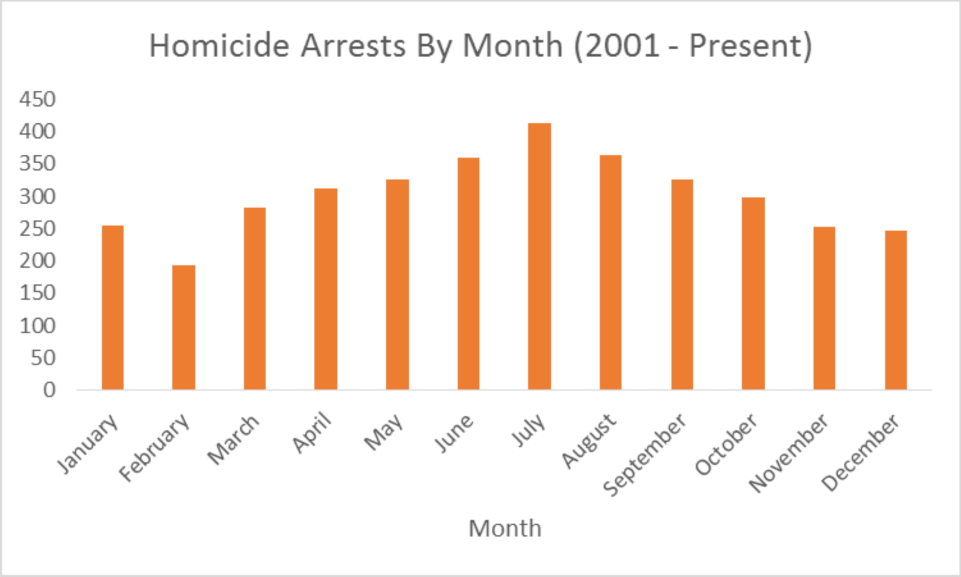


Homicide Arrests:

Homicide arrests typically occur between 11pm and 2am, and occur the most on weekends. They also show a sharp rise in July.

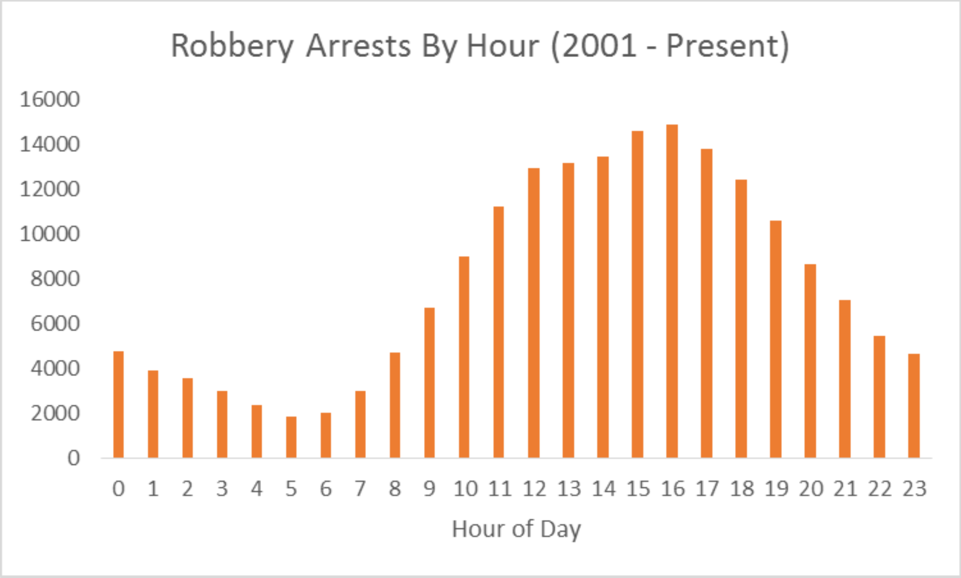


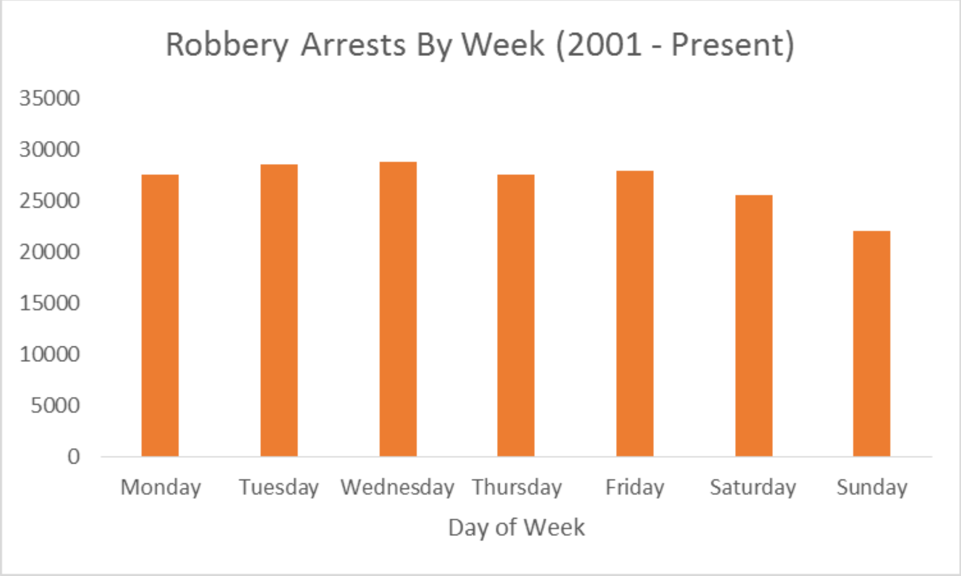




Robbery Arrests:

Robbery arrests peak between 3pm and 6pm. They occur fairly evenly on every day of the week, except a dip on Sunday, and show a consistent pattern across months.


Robbery Arrests By Hour (2001 - Present)
16000
14000
12000
10000
8000
6000
4000
2000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of Day


Robbery Arrests By Week (2001 - Present)
35000
30000
25000
20000
15000
10000
5000
0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of Week