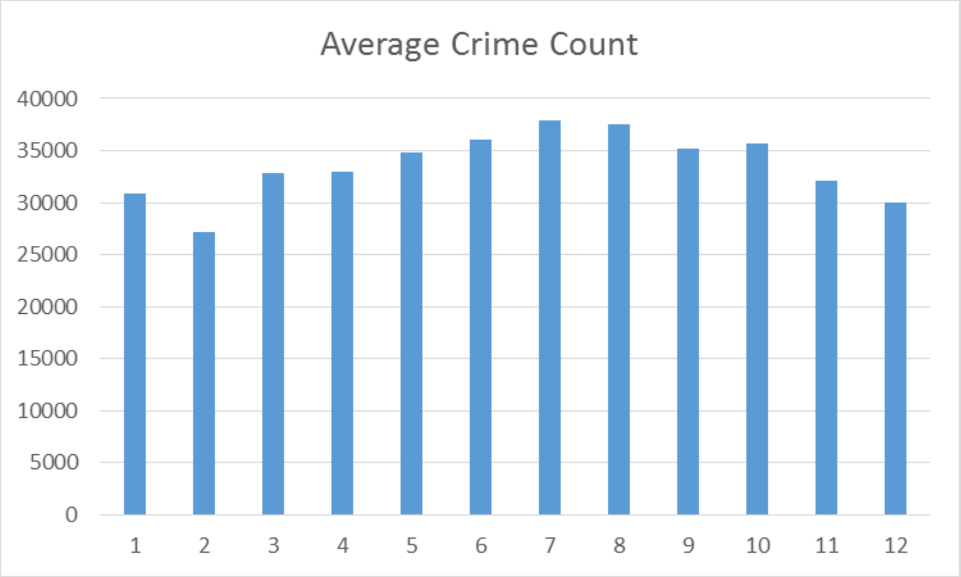
Homework 5

Apurvaa Subramaniam

1. The average crime count by month shows some seasonality. The summer months have consistently higher crime rates than the rest of the year. This is probably because people tend to go out more during the summer both for events and vacation.



2. Top 10 blocks in crime events in the last 3 years:

Block: 001XX N STATE ST	Crime Count: 1745
Block: 0000X W TERMINAL ST	Crime Count: 1340
Block: 008XX N MICHIGAN AVE	Crime Count: 1083
Block: 076XX S CICERO AVE	Crime Count: 1037
Block: 0000X N STATE ST	Crime Count: 794
Block: 051XX W MADISON ST	Crime Count: 661
Block: 064XX S DR MARTIN LUTHER KING JR DR	Crime Count: 628
Block: 083XX S STEWART AVE	Crime Count: 604
Block: 046XX W NORTH AVE	Crime Count: 571
Block: 009XX W BELMONT AVE	Crime Count: 550

Most of these are downtown Chicago which intuitively makes sense since these areas are more densely populated and also have people from more diverse socio-economic backgrounds.

The top 3 pairs of beats with the highest correlation in number of crime events are:

1934	1925	0.988
1914	1925	0.9826
1234	1215	0.9823

Again, this makes sense since 1934, 1925, 1915 are geographically close/adjacent to each other, so the rate of crime events should be similar. The same holds for 1234, 1215.

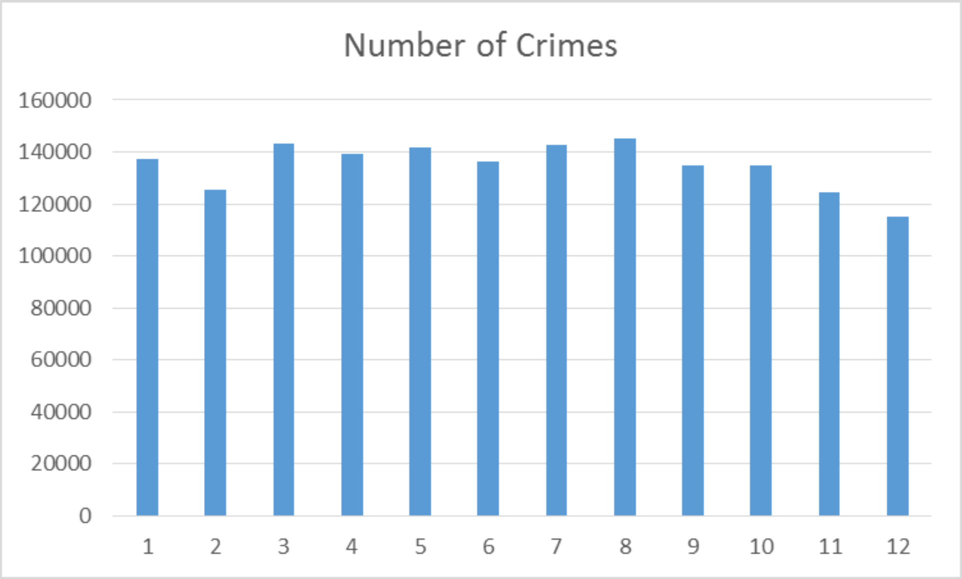
On normalizing and comparing mean yearly crime numbers mayors Daly and Emanuel at a district level, the mean difference is 1.21 and the 2-paired t statistic is 3.74 where N = 113. This corresponds to a p-value of <0.05 which is significant. This means the hypothesis that there is no statistically significant difference in crime events between Daly and Emanuel can be rejected. In this case, there is statistically more significant crime occurrence when Daly was the mayor.

3. For this question, predictors could be either temporal data or geographical data i.e. something that changes across beats. Geographical predictors such as income level by beat did not give very satisfactory results. Thus, I added a temporal predictor for seasonality (historical temperatures in Chicago) based on the findings from the first question. A random forest model with 20 trees and a depth of 10 with the newly added temperature predictor and previous crime data gives a test R squared of 65.5% and test Mean Square Error 70.3.

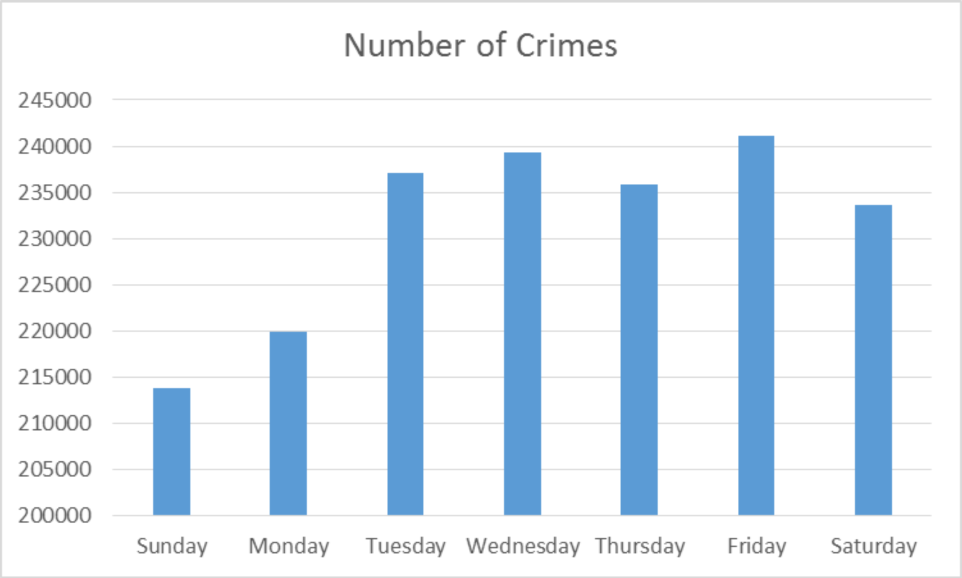
4. The graphs below show how number of crimes varies by month, day of the week and time of day.

Relatively very few crimes tend to occur on Sunday. This is probably because many families stay at home on Sundays since many shops/restaurants are closed, so most houses may not be empty while roads may be relatively empty. Friday on the other hand sees the largest number of crimes because people are most likely to go out then. Similarly, most crimes occur in the evening when people are likely to be out.

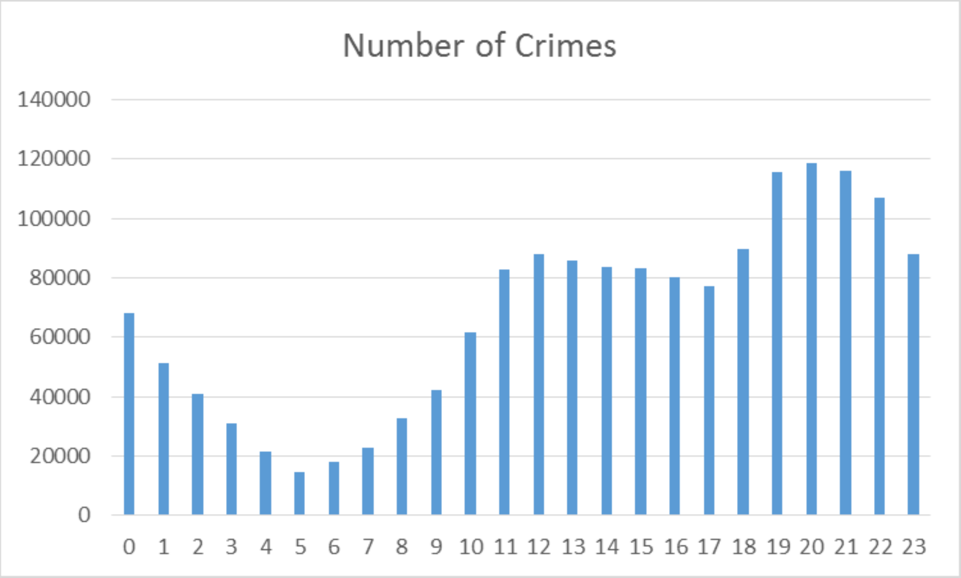
Monthly Pattern



Weekly Pattern



Daily Pattern


Number of Crimes
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