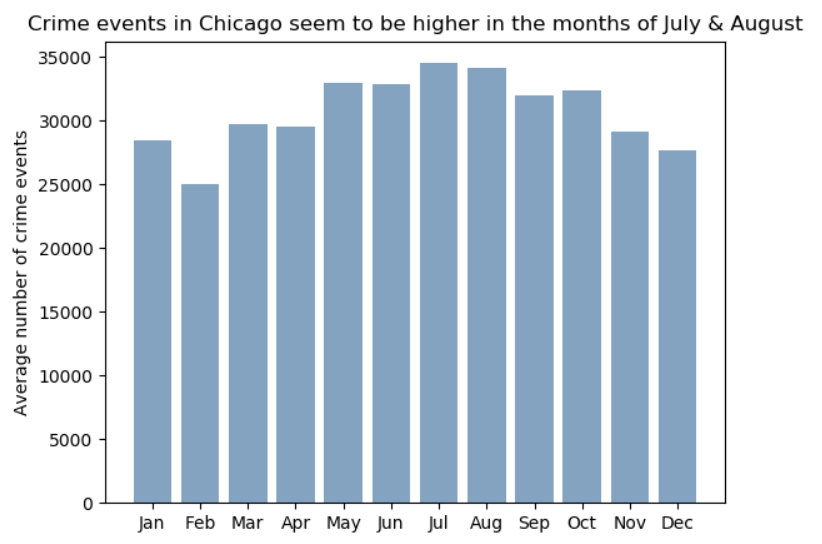
**Analytics for Big Data, Assignment 2 (Spark)**

Aakanksha Sah

**Q1: By using SparkSQL, generate a histogram of average crime events by month. Find an explanation of results. (10 pts)**

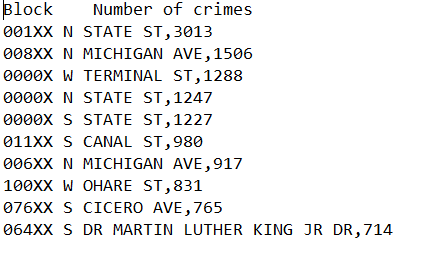
Ans:

Screenshot of output:The above graph captures the number of average crime events for each month in a year. It can be observed from this that crime seems to be little higher in the months of July and August. In general, crime in summer and fall is higher than winter and spring. It could be because the weather is better during summer, in general and more people are out travelling and shopping and this increases the likelihood of such incidents.

**Q2a: By using plain Spark (RDDs): (1) find the top 10 blocks in crime events in the last 3 years**

Ans.

Screenshot of output:

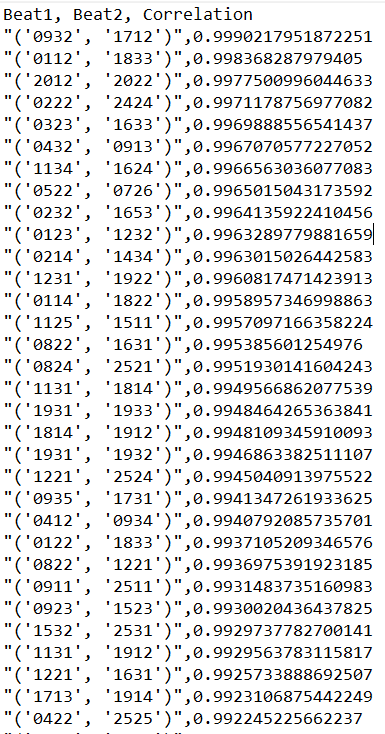


As can be seen, in the above screenshot – these are the top 10 blocks in terms of the highest number of crimes

**Q2b: Find the two beats that are adjacent with the highest correlation in the number of crime events (this will require you looking at the map to determine if the correlated beats are adjacent to each other) over the last 5 years**

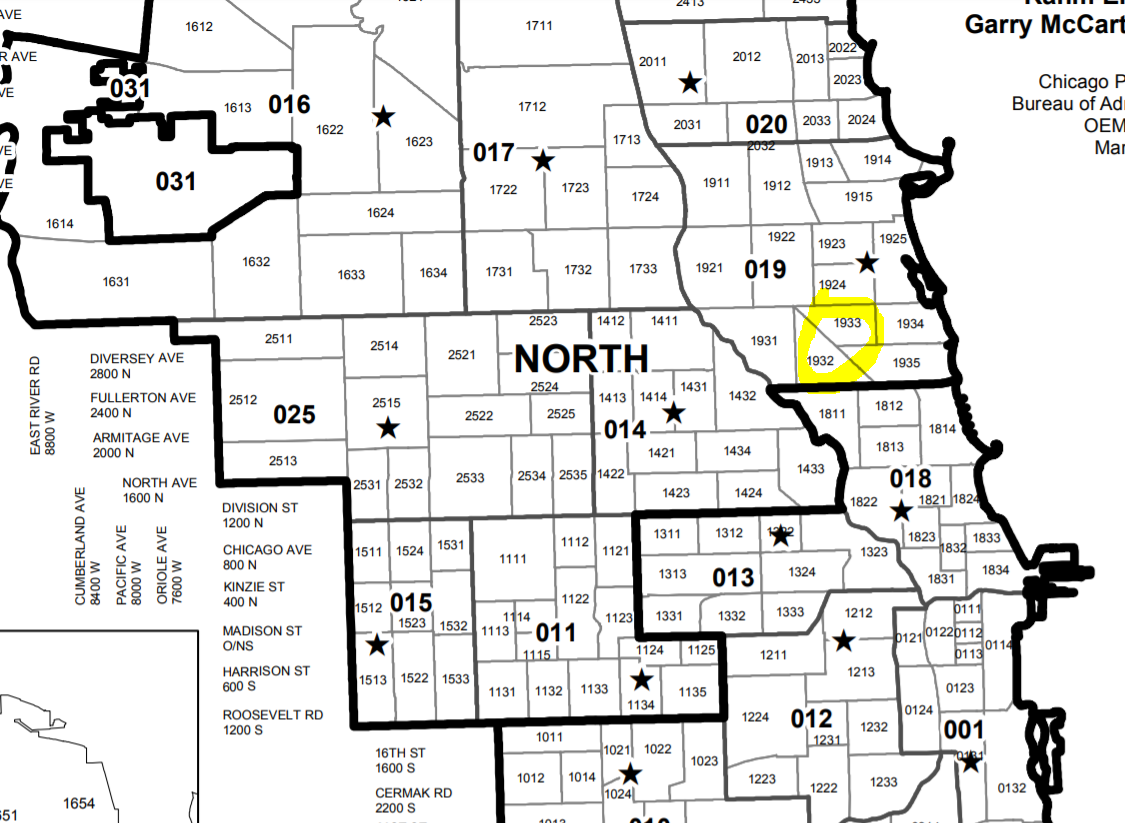
Ans.

Screenshot of output:



Note: In cases where a particular beat did not have any records for a given year, I have considered that as zero crimes during my computation of correlation as that requires equal sized vectors.

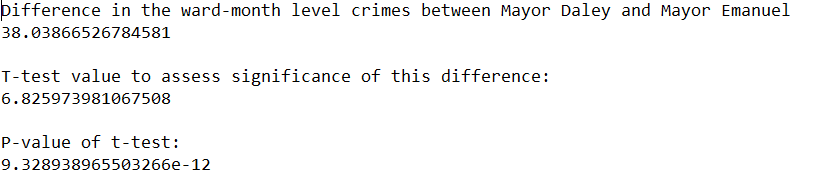
The output file for this question has 300 highly correlated beats (in terms of number of yearly crimes) listed. The two beats that are adjacent to each other and have a high correlation of 0.9909057698 are 1932 & 1933 (check highlighted portion in map below)



**Q2c: Establish if the number of crime events is different between Majors Daly and Emanuel at a granularity of your choice (not only at the city level). Find an explanation of results.**

Ans.

Screenshot of output:



To understand if the number of crime events are different between Majors Daly and Emanuel, we did the following:

* Filtered the data based on their working time periods and created 2 RDDs
* Rolled them up to calculate the number of crimes at a ward level
* Did a paired t-test and assessed its significance

The difference in the total number of crimes (normalized by their tenure) is ~38.04 and is higher in case of Mayor Daley.

The paired t-test shows that the difference between the number of ward-level crimes between mayor Daley and mayor Emanuel is significant, (T test value: 6.82597, P value: 9.328938e-12) which further validates the point that the difference is significant. **So, we can confidently say that crime events were significantly lesser during Mayor Emanuel’s reign**

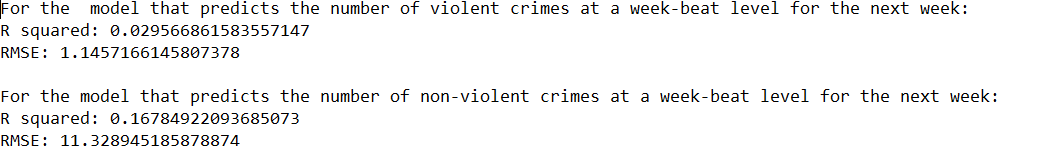
**Q3: Predict the number of crime events in the next week at the beat level. Violent crime events represent a greater threat to the public and thus it is desirable that they are forecasted more accurately (IUCR codes available here: https://data.cityofchicago.org/widgets/c7ck-438e). (45 pts) You are encouraged to bring in additional data sets. (extra 10 pts if you mix the existing data with an exogenous data set) Report the accuracy of your models. You must use Spark dataframes and ML pipelines.**

Ans.

For this question, I built two random forest models at beat-week level (one for violent crimes and the other for non-violent ones) and used lagged versions of the crime counts as features along with beat

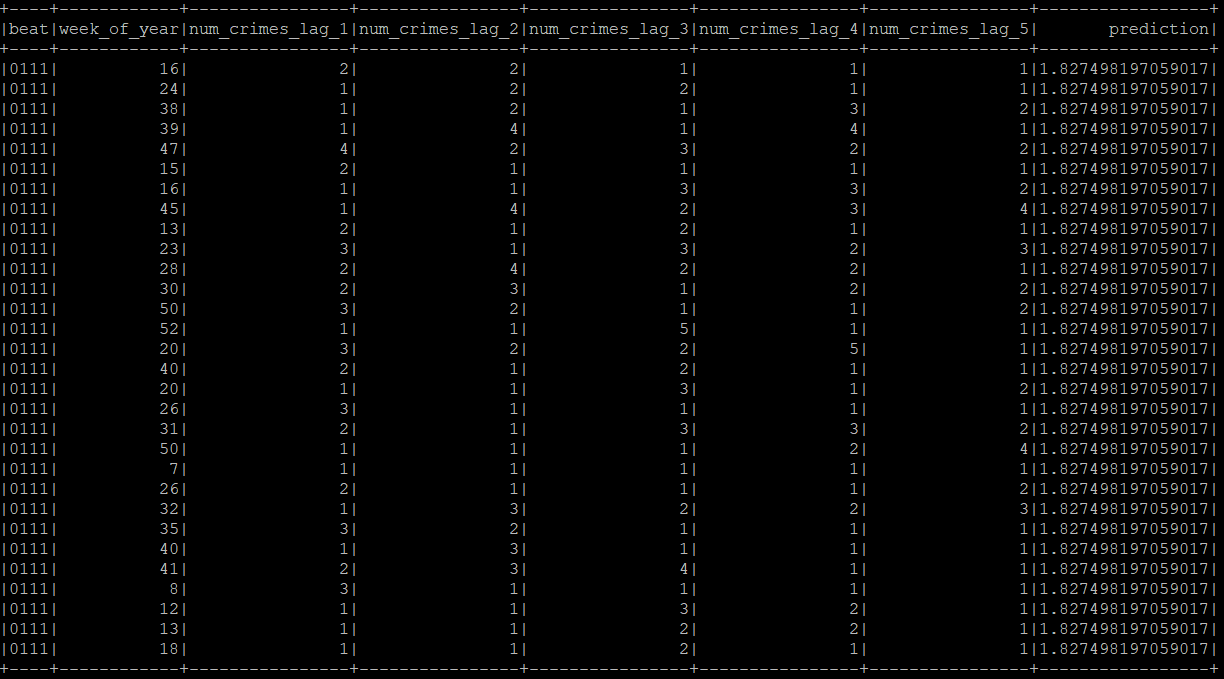
Screenshot of output:

1. Model accuracy



It can be seen above from the accuracy metrics that the model that predicts the number of crimes at a week-beat level for non-violent crimes does a better job than that that predicts violent crimes. The R squared value is not great. Including more features and data from other sources can help increase accuracy

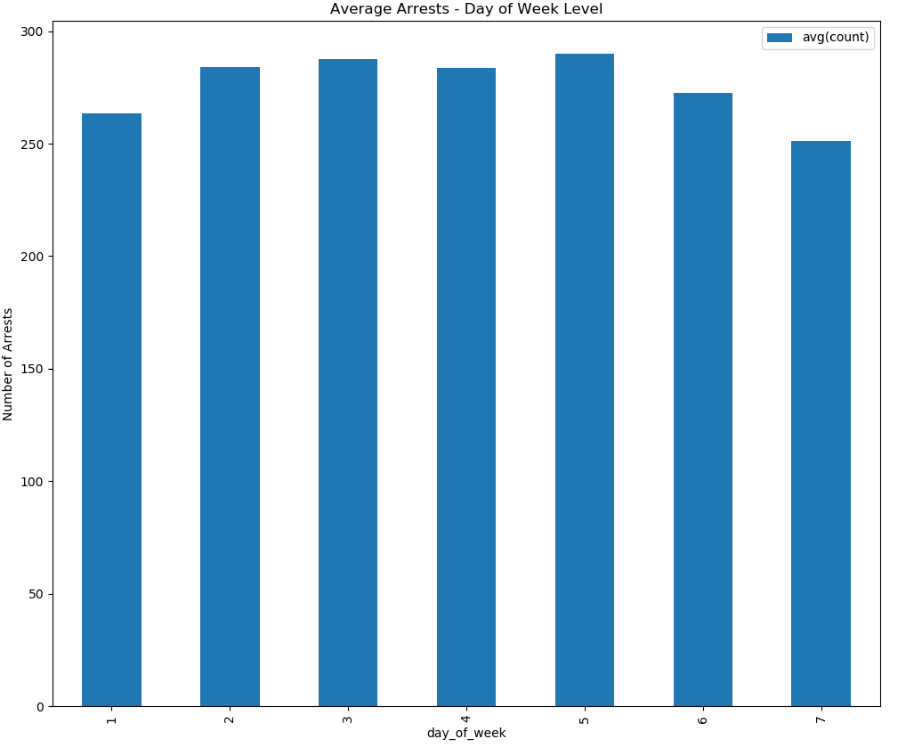
1. Final dataset (that show the features – week of year, beats, lag variables and the final predictions as well)



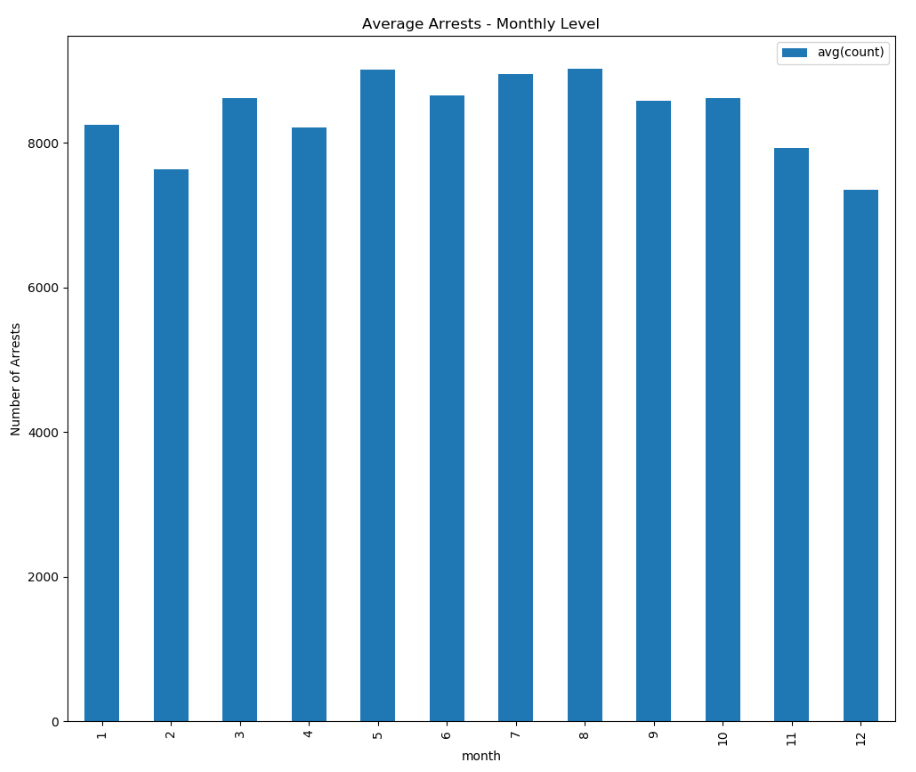
**Q4: Find patterns of crimes with arrest with respect to time of the day, day of the week, and month. Use whatever method in spark you would like. (25 pts)**

Ans.

1. Day of week – There seems to be more crimes (leading to more arrests) in the weekday than the weekend. This makes sense because generally people stay home during the weekends and this decreases the probability of crimes

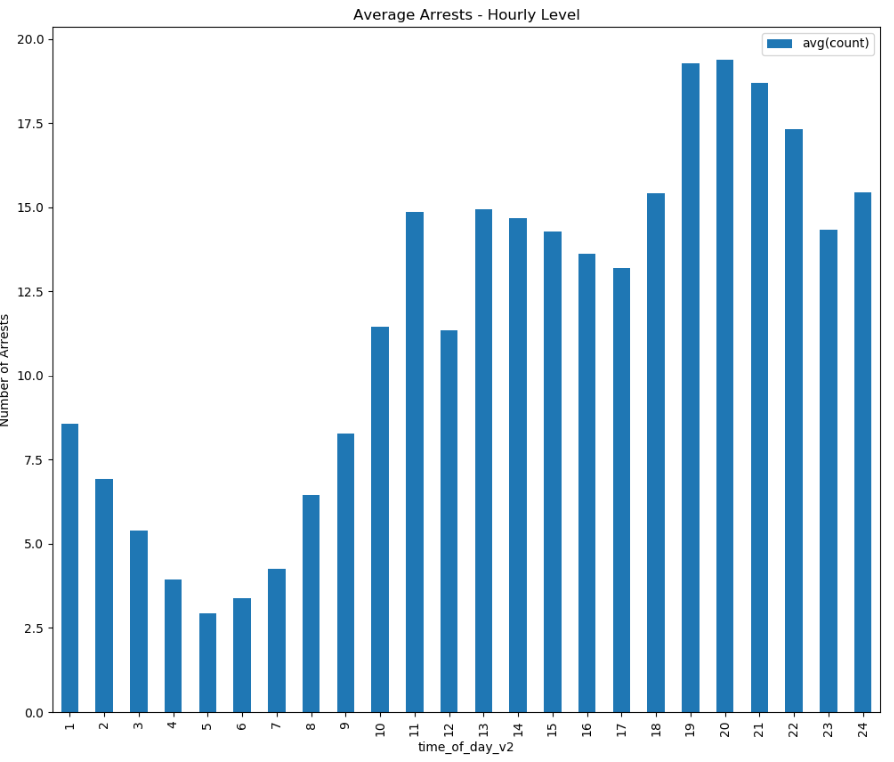


1. Monthly level – This has a similar pattern as observed in Q1



More arrests during summer and fall. The highest is in May, July and August. The lowest number of arrests are in December and February.

1. Time of day level –



More arrests happen during the later half of the day (evening and night)