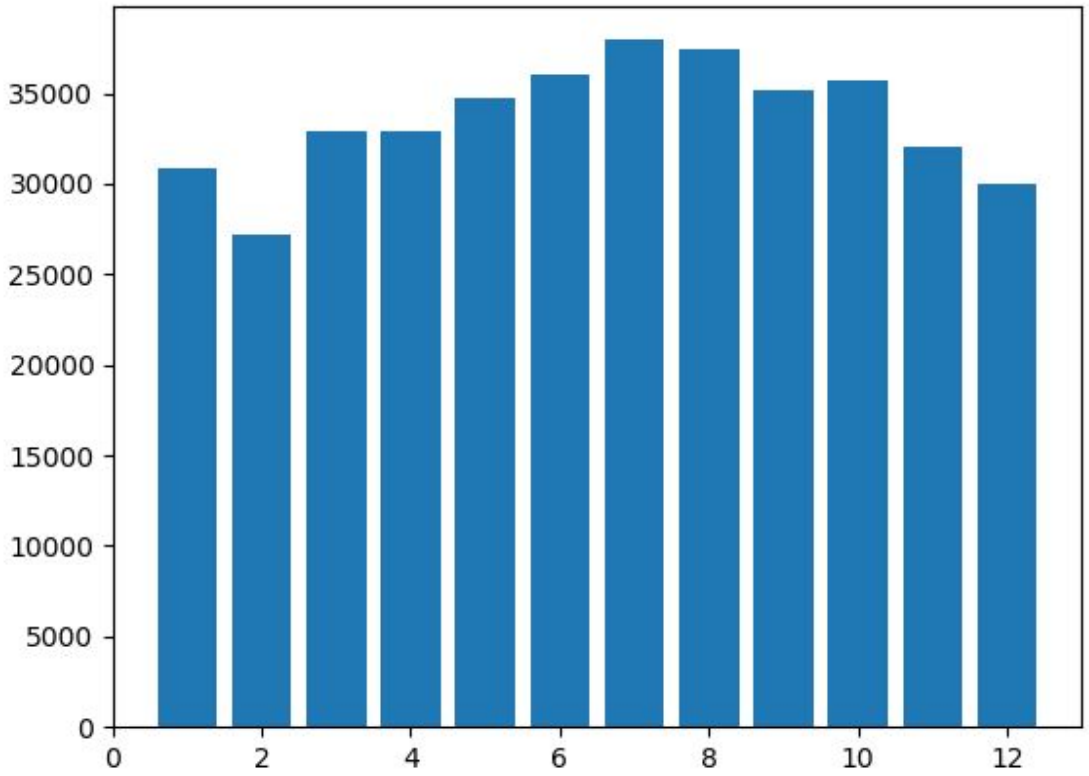
1. By using SparkSQL generate a histogram of average crime events by month. Find an explanation of results. (10 pts)



From this, it is clear that 7th month(July) has maximum number of crimes which may be one of the most pleasant weather which makes it more likely for people to be outside. Not only July reports highest crime, the months next to it(June and August) have second highest number of crime records. December - February are the coldest months in Chicago and they report the lowest crime on an average.

So, it all makes perfect sense

2. By using plain Spark (no Spark SQL):
(1) find the top 10 blocks in crime events in the last 3 years;

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

(2) find the two beats that are adjacent with the highest correlation in the number of crime events over the last 5 years
The highest correlated in the number of crime events over the last five years: 2331, 2332 (0.66929253 - correlation)

(3) 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. (20 pts)

T test results by year

- T_statistic = -3.263218217090926743e+00
- Pvalue = 3.921200809590928010e-02

T test results by beat and month

- T statistic = -1.011019973499046785e+02
- Pvalue = 0.000e+00

Therefore, we can conclude that after 2011 under Mayor Emanuel crimes have gotten worse.

3. Predict the number of crime events in the next week at the beat level. The higher the IUCR is, the more severe the crime is. Violent crime events are more important and thus it is desirable that they are forecasted more accurately. (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 pipelines.

When Random forest is used,
Evaluation metric: 0.7036732403542376

4. Find patterns of crimes with arrest with respect to time of the day, day of the week, and month. (25 pts)
By hour:

Date_hour	count
0	326416
1	185828
2	155131
3	124742
4	92972
5	76285
6	89800
7	130299
8	194829
9	245697
10	241402
11	255168
12	327281
13	275314
14	293871
15	307399
16	288346
17	293000
18	315714
19	330763
20	333495
21	327930
22	323516
23	266646

Explanation: 19 hr - 21 hr has the three highest crime counts with 20 hr being the highest crime rate. Followed by 12 hr and 0hr which again are peak hours.

By month:

Date_month	count
1	463497
2	407957
3	492905
4	494231
5	521509
6	504706
7	531295
8	524592
9	492694
10	500018
11	448726
12	419714

Explanation: It is clear that 7th month(July) has maximum number of crimes which may be one of the most pleasant weather which makes it more likely for people to be outside. Not only July reports highest crime, the months next to it(June and August) have second highest number of crime records. December - February are the coldest months in Chicago and they report the lowest crime on an average.

By weekday:

Date_week	count
Sunday	779444
Monday	817722
Saturday	827701
Thursday	828943
Tuesday	834011
Wednesday	839315
Friday	874708

Explanation: Friday has the highest crime rate followed by Wednesday and Tuesday. It's interesting that Sunday has the least crime counts.