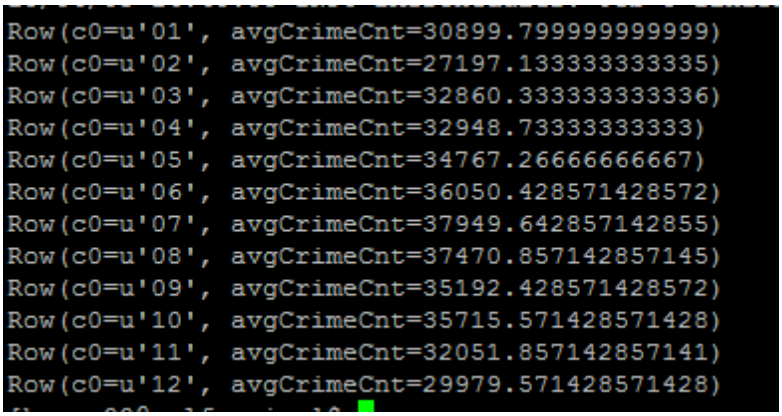
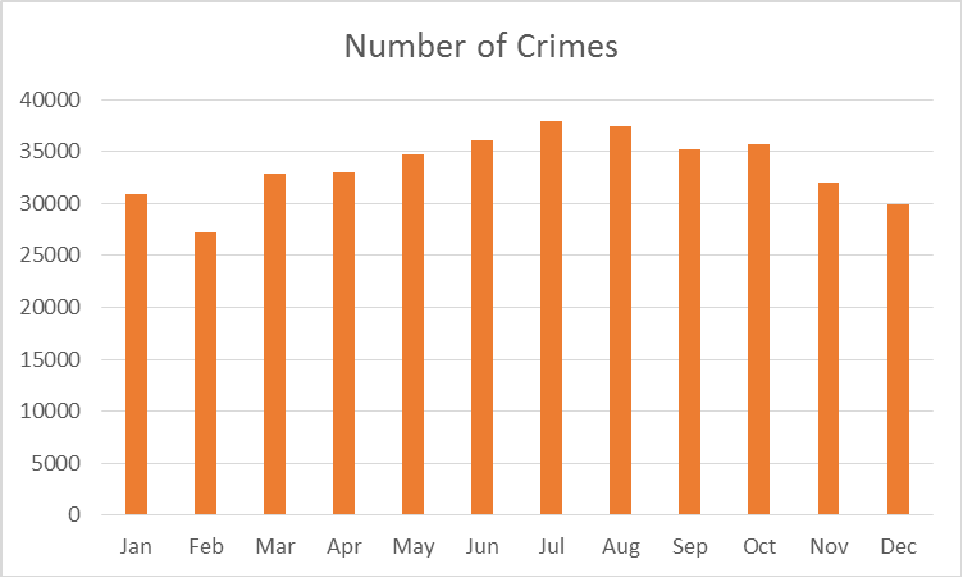
Assignment 5
Analytics for Big Data
Spring 2015
By: Sanjeevni Wanchoo

Problem 1: By using SparkSQL generate a histogram of average crime events by month. Find an explanation of results.

The following output was obtained



MONTH	NUMBER OF CRIMES
JAN	30899.8
FEB	27197.13
MAR	32860.33
APR	32948.73
MAY	34767.27
JUN	36050.43
JUL	37949.64
AUG	37470.86
SEP	35192.43
OCT	35715.57
NOV	32051.86
DEC	29979.57



The results show that most of the crimes happen during the summer months. This is probably because the nature of several crimes is that they are committed outdoors, and more people tend to go out during summers. (Also, the criminals probably just get too cold during winters).

Problem 2

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

The top 10 blocks in term sof crime events in the last three years (2012-2015) are:

001XX N STATE ST	2279
0000X W TERMINAL ST	1896
008XX N MICHIGAN AVE	1744
076XX S CICERO AVE	1541
0000X N STATE ST	1275
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

Part 2: Find the two beats that are adjacent with the highest correlation in the number of crime events.

The output (in ascending order) is shown below. That is, the last beat combo has the highest correlation.

```
1921 1914
0215 0121
1921 1934
1921 1935
1921 0121
1214 1234
1925 1934
1221 1215
1934 1935
1925 1935
```

The beats with the highest correlation in the number of crime events are shown below.

BEAT 1	BEAT 2
1925	1935
1934	1935
1221	1215
1925	1934
1214	1234

Here, 1924 is adjacent to 1925, which in turn is adjacent to 1934. So in a way, the three beats are all adjacent to each other. Furthermore, 1221 and 1215 are also adjacent. This intuitively makes sense, as we would expect adjacent beats to show similar changes in crime activity over the years.

Part 3: Establish if the number of crime events is different between Mayors Daly and Emanuel at a granularity of your choice (not only at the city level).

The results were collected at the district level. Data collected up to and including 2011 was used as crime data corresponding to Mayor Daly, while the data collected after 2011 was used as crime data for Mayor Emanuel. The crimes were averaged over the number of years for each mayor at the district level, and this data was then used to perform the t-test. The null hypothesis was that the two distributions of crime rates are very similar.

The results gave a t-value of 2.9835696656, which is much higher than 2.08, which is the t-statistic for alpha=0.05, and 21 degrees of freedom. This means our null hypothesis can be rejected, and the number of crimes during the two mayors is not very similar.

Problem 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 50 pts if you mix the existing data with an exogenous data set) Report the accuracy of your models.

In order to predict the number of crimes in the coming week, a random forest model was used. From the crime dataset, the following features were used: Beat, Week.

The existing crime data was combined with public school data. This dataset was only available at district level. From this dataset, the number of schools in each district were calculated. This was used as one of the predictors for the crime prediction model.

In order to handle the categorical variables, category dictionaries were created. Thus the final input to the random forest model looked as shown below:

```
beledPoint(7.0, [43.0,173.0,19.0]), LabeledPoint(22.0, [7.0,173.0,19.0]), Labele
```

As can be seen above, in the first vector, 7 crimes were committed in the beat that corresponds to 173 in the dictionary. In this case, this is beat 1821 as shown below.

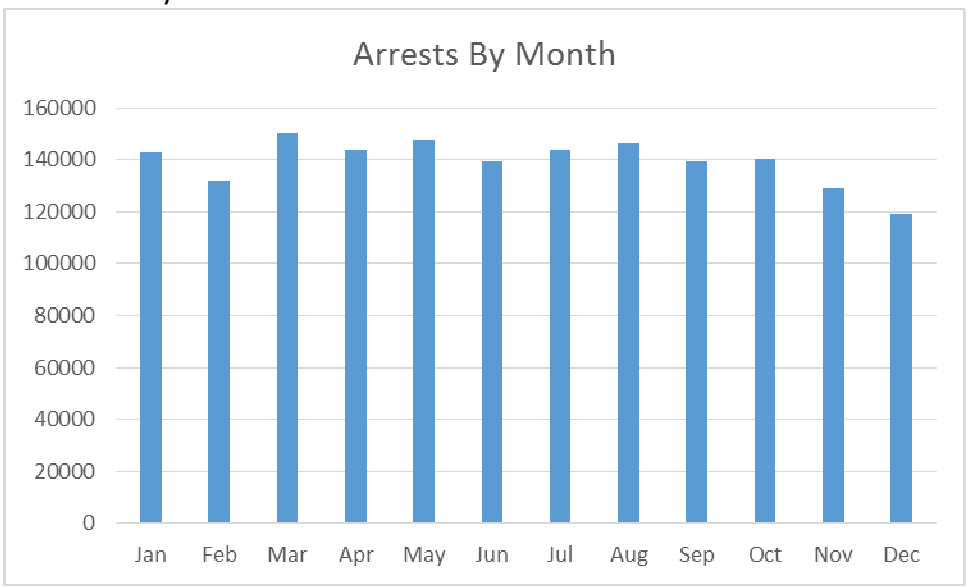
```
>>> beatsDict.keys()[173]
'1821'
```

Thus, we see that in beat 1821, 7 crimes were committed in week 43, and 19 schools exist in the district of 1821.

The model was run, and evaluated. However, due to technical issues and current unavailability of hdfs, these cannot be run at this time. Please see the attached code for details.

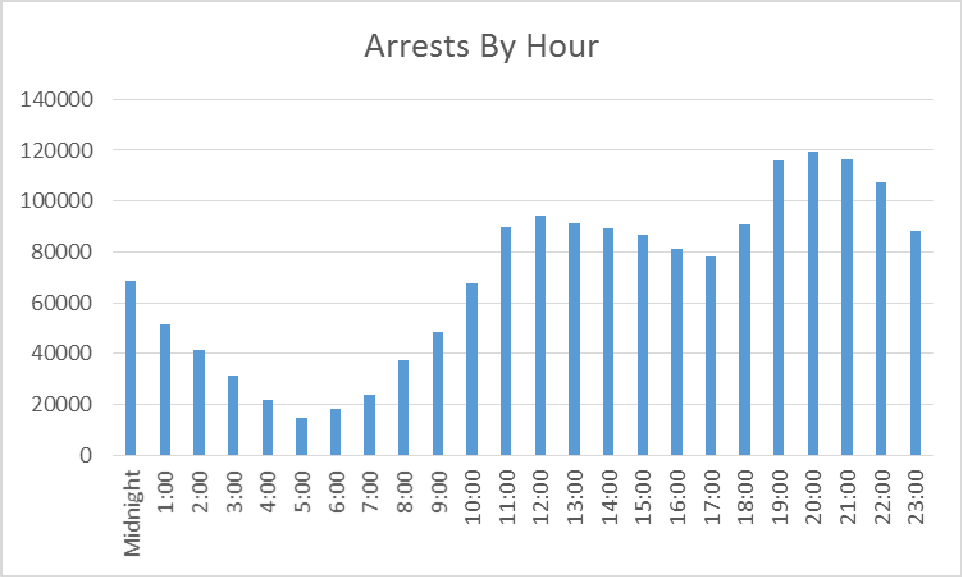
Problem 4: Find patterns of crimes with arrest with respect to time of the day, day of the week, and month.

The Arrests by Month are shown below:



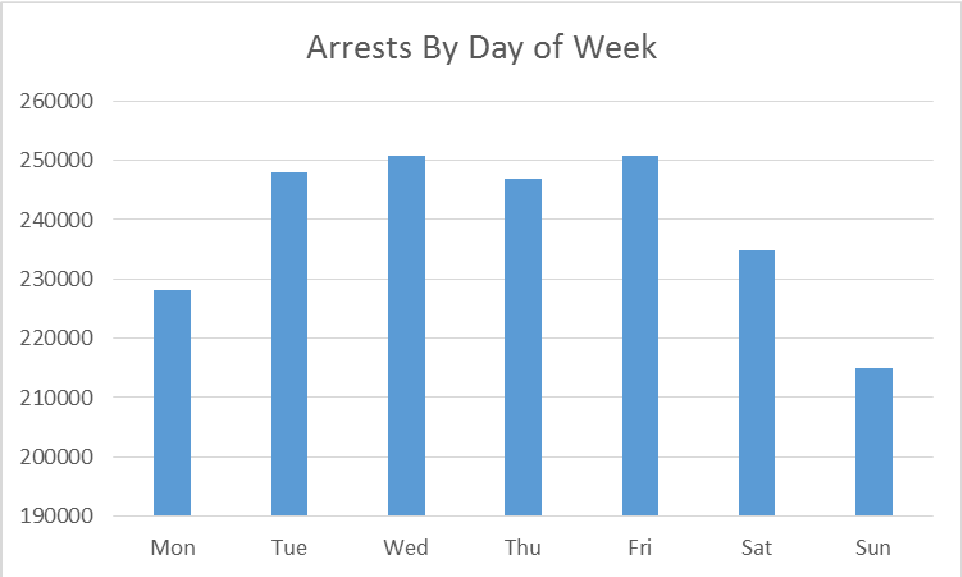
There doesn't seem to be any interesting pattern here. Arrests seem to be lowest at the end of the month, but are generally evenly spread out over the course of the year.

The arrests by hour are shown below:



Here, we see that most of the arrests happen during the evening hours (7 PM-11 PM). The least happen during the early morning hours.

The arrest by day of the week are shown below:



As we can see above, it seems like most of the arrests happen Tuesday-Friday. The low number of arrests on the weekend seems odd, but could also potentially be attributed to lower coverage by police on the weekends.