1. **Calculate the Total count of Crimes for each of the 6 United States cities listed below using 2016 crime data in dbfs:/mnt/training/crime-data-2016**

New York: 468241

Los Angeles: 217945

Chicago: 267872

Philadelphia: 168664

Dallas: 99642

Boston: 218610

##### Provide Total count of different Types of Crimes for each of 6 United States cities listed below using crime-data-2016

Showing first 7 rows for each city only. For complete results please see notebook

##### 

##### 

##### 

##### 

##### 

##### 

##### Calculate the total Robbery count for each of the 3 United States cities listed below using crime-data-2016

Total robbery count:

LosAngeles: 9048

Philadelphia: 6149

Dallas: 6824

##### Find the months with the Highest and Lowest Robbery counts for each of the 3 United States cities listed below using crime-data-2016

Los Angeles:

Highest robbery count:

+-----+---------+

|month|robberies|

+-----+---------+

| 02| 675|

+-----+---------+

Lowest robbery count:

+-----+---------+

|month|robberies|

+-----+---------+

| 12| 853|

+-----+---------+

Philadelphia:

Highest robbery count:

+-----+---------+

|month|robberies|

+-----+---------+

| 02| 416|

+-----+---------+

Lowest robbery count:

+-----+---------+

|month|robberies|

+-----+---------+

| 10| 572|

+-----+---------+

Dallas:

Highest robbery count:

+-----+---------+

|month|robberies|

+-----+---------+

| 03| 412|

+-----+---------+

Lowest robbery count:

+-----+---------+

|month|robberies|

+-----+---------+

| 01| 743|

+-----+---------+

##### Combine all three cities robberies-per-month views into one and Find the month with the Highest and Lowest combined Robbery counts using crime-data-2016

Highest combined robberies:

+-----+---------+

|month|robberies|

+-----+---------+

| 12| 2061|

+-----+---------+

Lowest combined robberies:

+-----+---------+

|month|robberies|

+-----+---------+

| 02| 1526|

+-----+---------+

##### Plot the robberies per month for each of the three cities, producing one Graph using the contents of combinedRobberiesByMonthDF

##### 

##### Find the "per capita robbery rates" using ther Hint below, and plot graph as above for the per capita robbery rates per month for each of the three cities, producing one Graph using the contents of robberyRatesByCityDF

##### 

##### Find the monthly HOMICIDE counts for each of the 2 United States cities listed below using crime-data-2016, and Combine both cities HOMICIDE-per-month views into one and Find the month with the Highest and Lowest combined Robbery counts using crime-data-2016

Month with Highest combined homicides:

+-----+---------+

|month|homicides|

+-----+---------+

| 8| 50|

+-----+---------+

Month with Lowest combined homicides:

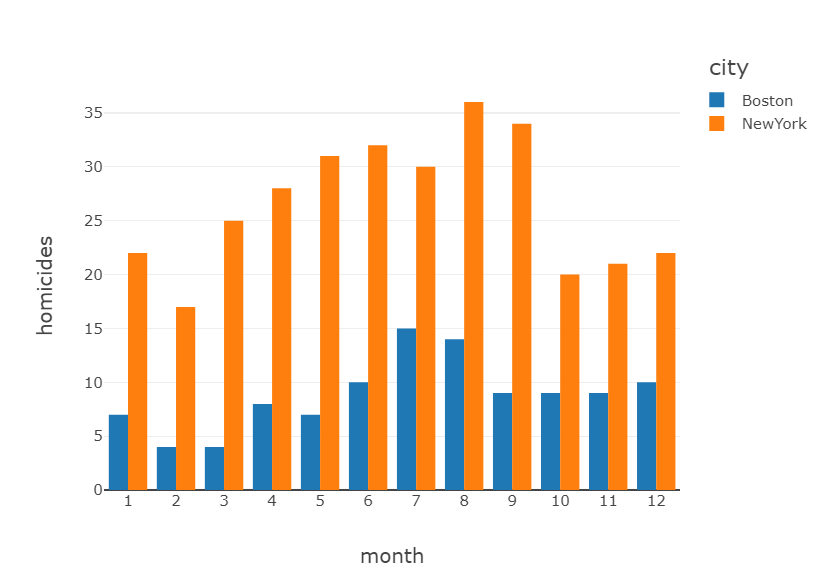
+-----+---------+

|month|homicides|

+-----+---------+

| 2| 21|

+-----+---------+



##### Find the "per capita HOMICIDE rates" using ther Hint in Qn 7, and plot graph as above for the per capita HOMICIDE rates per month for each of the 2 cities, producing one Graph using the contents of HOMICIDERatesByCityDF

##### 

##### A stretch goal to address a Data Science question on predicting crimes for a city as a time series model. How would you predict future values for monthly Robbery count rate for Log Angeles using the historical values from crime-data-2016. Data Science is a multi-year discipline. I am not expecting anything fancy. There is no wrong answer. I do expect students to explore and give their best shot.

##### Since, we have to predict future robberies based on past data of 2016, this is a classic ML prediction problem. We can utilize built-in MLib function 'Linearregression' from pyspark.ml.regression. Our dataset will contain 1 feature i.e. month; with 'count of robberies' as label. In order to normalize our data, we will need to transform month range from 1-12 into one-hot vector (using one-hot encoding). Loss used by our model can be MeanSquareError loss. We can then split our data into 80-20, i.e. 80% data would be used for training and 20% for testing. In order to train, we can call .fit() function and specify the hyperparamers eg: regParam, numberOfEpochs etc. Once our model is done with the training, we can predict future crime values.

##### Further improvements in above model can be done by utilizing sequencial information. Currently we are not leveraging sequence information for month and we are treating them as a bag. The fact that February comes after January and March comes after February and so on, can significantly impact model's efficacy sometimes. In addition, rate of crime might be dependent on time of the year, for example, there might be more crimes occuring in winter than in summer as people tend to stay home during the cold weather. So, we can use Recurrent Neural Networks or RNN's, a Deep Learning model for that; where we would pass each months information i.e. month-crimes pair at each timestep to train our model. Since there are 12 months in a year, our model would be trained for 12 time-steps. We can control how much information of history we want to keep (eg: information for January while we are training model for February), by tuning hyperparameters. This way, we can treat month's information, as a time-series and train our model to predict number of crimes for the next month in time series, based on history.