Assignment-1

Spark Dataframe

You will be asked some basic questions about some stock market data, in this case Walmart Stock from the years 2012-2017.

Dataset: walmart_stock.csv

Questions:

- Start a simple Spark Session
- Load the Walmart Stock CSV File, have Spark infer the data types.
- What are the column names?
- What does the Schema look like?
- Print out the first 5 columns.
- Use describe() to learn about the DataFrame. Provide your inference on the same
- From the above question, There may many decimal places for mean and stddev in the describe() dataframe. Format the numbers to just show up to two decimal places. Pay careful attention to the datatypes that .describe() returns

Hint:
(http://spark.apache.org/docs/latest/api/python/pyspark.sql.html#pyspark.sql.Column.cast)

- Create a new dataframe with a column called HV Ratio that is the ratio of the High Price versus volume of stock traded for a day.
- What day had the Peak High in Price?
- What is the mean of the Close column?
- What is the max and min of the Volume column?
- How many days was the Close lower than 60 dollars?
- What percentage of the time was the High greater than 80 dollars ?
  Hint: Number of Days High>80)/(Total Days in the dataset)
- What is the Pearson correlation between High and Volume?

Hint:http://spark.apache.org/docs/latest/api/python/pyspark.sql.html#pyspark.sql.DataFrameStatFunctions.corr

- What is the max High per year?
- What is the average Close for each Calendar Month?¶

Hint: In other words, across all the years, what is the average Close price for Jan,Feb, Mar, etc... Your result will have a value for each of these months.