Spark Assignment

Karan Kant (kk4ze)

Motivation

- Everybody assumes that a higher writing score on the SAT correlates to a higher reading score. I wanted to test this hypothesis and confirm my initial thoughts.
- My model confirms a high correlation between SAT reading and writing scores. A student with a high reading score on average gets a high writing score as well.

Code

```
In [8]: df = sqlc.createDataFrame(pd.read_csv('StudentsPerformance.csv')) #this will not work on big files

In [9]: df

Out[9]: DataFrame[gender: string, race/ethnicity: string, parental level of education: string, lunch: string, test preparatio
        n course: string, math score: bigint, reading score: bigint, writing score: bigint]

In [10]: df=df.withColumnRenamed("parental level of education","parentedu")
         df=df.withColumnRenamed("test preparation score","testprepscore")
         df=df.withColumnRenamed("math score","math")
         df=df.withColumnRenamed("reading score","reading")
         df=df.withColumnRenamed("writing score","writing")

In [11]: df=df.withColumnRenamed("test preparation course","testsprepcourse")

In [12]: df.printSchema()

         root
          |-- gender: string (nullable = true)
          |-- race/ethnicity: string (nullable = true)
          |-- parentedu: string (nullable = true)
          |-- lunch: string (nullable = true)
          |-- testsprepcourse: string (nullable = true)
          |-- math: long (nullable = true)
          |-- reading: long (nullable = true)
```

- The dataset was uploaded to sagemaker. Columns needed to be renamed in order to process the data through spark dataframes.

Visualization

- High correlation is observed between reading and writing scores.