README file

The project portfolio contains two major folders - *Code* and *Data*. As name says, the *Code* folder contains all the scripts that we used for our data extraction with Spark/Hive/Impala, whereas the *Data* folder contains all the data generated by these scripts. The dataset can be downloaded from
https://s3.amazonaws.com/ed-college-choice-public/CollegeScorecard_Raw_Data.zip

Code(Folder)**:**
Data Preprocessing:

1. **Preprocessing.py**:
   a python script to merge all the .csv files in dataset and appends a column Year to the merged file. The syntax for running the file is-
   ```
   python Preprocessing.py
   ```

2. **CreateTableScript.py**:
   a python script that will generate CREATE TABLE statements for Hive/Impala/Spark after reading the data file (the merged file output by above script). The syntax for running the file is -
   ```
   python  CreateTableScript.py
   ```

Hive:

1. **hiveCreateScorecard.hql**:
   a hive script to create high dimensional table for dataset and loading data in the table. The syntax for running the file is -
   ```
   hive -f hiveCreateScorecard.hql
   ```

2. **hiveQueryTables.hql**:
   a script to query Hive tables. The syntax for running the file is -
   ```
   hive -f hiveQueryTables.hql
   ```

Impala:

1. **impalaCreateScorecard.txt**:
   a script to create high dimensional table for dataset and loading data in the table. The syntax for running the file is -
   ```
   impala-shell -f impalaCreateScorecard.txt
   ```

2. **ImpalaQueryTables.sql**:
   a script to query Impala tables. The syntax for running the file is -
   ```
   impala-shell -f ImpalaQueryTables.sql
   ```

3. **impalaExtractionForMahout.sql**:
   a script to extract data for clustering which will be done by Mahout. The syntax for running the file is -
   ```
   impala-shell -f impalaExtractionForMahout.sql -o mahoutClusterInput.txt
   ```

Spark:

1. **sparkYearScript.scala**:
   a script to extract data for a particular year from dataset. The syntax for running the file is -

```
spark-shell -i sparkYearScript.scala --packages com.databricks:spark-csv_2.10:1.1.0
```

2. **sparkTrendScript.scala**:
   a script to extract data for particular features for all years. The data will be extracted for years where it is present. The syntax for running the file is -

```
spark-shell -i sparkTrendScript.scala --packages com.databricks:spark-csv_2.10:1.1.0
```

3. **sparkStateScript.scala** :
   a script to extract statewise data from the dataset. The syntax for running the file is -

```
spark-shell -i sparkStateScript.scala --packages com.databricks:spark-csv_2.10:1.1.0
```

Mahout:

1. **mahoutClusterInput.txt**:
   an input file to Mahout for clustering

2. **mahoutClusterAnalysisSteps.txt**:
   Steps to be followed for clustering analysis.

3. **mahoutClusterOutput.txt**:
   output generated by clustering algorithm

R:

1. **DataAnalysis.R** :
   This file contains commands/scripts used for the analysis of data extracted from Spark and Mahout.

Data(Folder)**:**

The Data directory contains two folders:

1. **excelDataVersion**:
   Contains excel version of data and (some) graphs/plots of our analysis

2. **rawDataFiles**
   Contains all the raw data files extracted from Spark.