Data Ingestion

**Data Input:**

hdfs dfs -mkdir final-project hdfs dfs -mkdir final-project/input

hdfs dfs -put dataset.csv final-project/input

**Pre-Clean Profiling:**

javac -classpath opencsv.jar:`yarn classpath` -d . CountRecsMapper.java

javac -classpath opencsv.jar`yarn classpath` -d . CountRecsReducer.java

javac -classpath `yarn classpath`:. -d . CountRecs.java

jar -cvf countRecs.jar \*.class

hadoop jar countRecs.jar CountRecs final-project/input/dataset.csv final-project/output

hdfs dfs -cat final-project/output/part-r-00000

javac -classpath opencsv.jar:`yarn classpath` -d . FindMeanGoalMapper.java

javac -classpath opencsv.jar`yarn classpath` -d . FindMeanGoalReducer.java

javac -classpath `yarn classpath`:. -d . FindMeanGoal.java

jar -cvf findMeanGoal.jar \*.class

hadoop jar findMeanGoal.jar FindMeanGoal final-project/input/dataset.csv final-project/output1

hdfs dfs -cat final-project/output1/part-r-00000

javac -classpath opencsv.jar:`yarn classpath` -d . FindDistinctMapper.java

javac -classpath opencsv.jar`yarn classpath` -d . FindDistinctReducer.java

javac -classpath `yarn classpath`:. -d . FindDistinct.java

jar -cvf findDistinct.jar \*.class

hadoop jar findDistinct.jar FindDistinct final-project/input/dataset.csv final-project/output2

hdfs dfs -cat final-project/output2/part-r-00000

**Cleaning:**

javac -classpath `yarn classpath` -d . YearMapper.java

javac -classpath `yarn classpath` -d . YearReducer.java

javac -classpath `yarn classpath`:. -d . Year.java

jar -cvf year.jar \*.class

hadoop jar year.jar Year final-project/input/dataset.csv final-project/output3

hdfs dfs -cat final-project/output3/part-r-00000

javac -classpath `yarn classpath` -d . CleanMapper.java

javac -classpath `yarn classpath` -d . CleanReducer.java

javac -classpath `yarn classpath`:. -d . Clean.java

jar -cvf clean.jar \*.class

hdfs dfs -put post-year-dataset.csv final-project/input

hadoop jar clean.jar Clean final-project/input/post-year-dataset.csv final-project/output4

hdfs dfs -cat final-project/output4/part-r-00000

**Post-Cleaning Profiling:**

javac -classpath opencsv.jar:`yarn classpath` -d . CountRecsMapper.java

javac -classpath opencsv.jar`yarn classpath` -d . CountRecsReducer.java

javac -classpath `yarn classpath`:. -d . CountRecs.java

jar -cvf countRecs.jar \*.class

hdfs dfs -put clean-dataset.csv final-project/input

hadoop jar countRecs.jar CountRecs final-project/input/clean-dataset.csv final-project/output5

hdfs dfs -cat final-project/output5/part-r-00000

**Analysis:**

javac -classpath `yarn classpath` -d . ExpectedGoalsVsGoalsMapper.java

javac -classpath `yarn classpath` -d . ExpectedGoalsVsGoalsReducer.java

javac -classpath `yarn classpath`:. -d . ExpectedGoalsVsGoals.java

jar -cvf expectedGoals.jar \*.class

hdfs dfs -put clean-dataset.csv final-project/input

hadoop jar expectedGoals.jar ExpectedGoalsVsGoals final-project/input/clean-dataset.csv final-project/output6

hdfs dfs -cat final-project/output6/part-r-00000

javac -classpath `yarn classpath` -d . ExpectedGoalsVsGoalsMapper.java

javac -classpath `yarn classpath` -d . ExpectedGoalsVsGoalsReducer.java

javac -classpath `yarn classpath`:. -d . ExpectedGoalsVsGoals.java

jar -cvf expectedGoals.jar \*.class

hdfs dfs -put clean-dataset.csv final-project/input

hadoop jar expectedGoals.jar ExpectedGoalsVsGoals final-project/input/clean-dataset.csv final-project/output6

hdfs dfs -cat final-project/output6/part-r-00000

javac -classpath `yarn classpath` -d . SortColumnMapper.java

javac -classpath `yarn classpath` -d . SortColumnReducer.java

javac -classpath `yarn classpath`:. -d . SortColumn.java

jar -cvf sortColumn.jar \*.class

hdfs dfs -put ExpectedGoalsVsGoalsOutput.csv final-project/input

hadoop jar sortColumn.jar SortColumn final-project/input/ExpectedGoalsVsGoalsOutput.csv final-project/output7

hdfs dfs -cat final-project/output7/part-r-00000

javac -classpath `yarn classpath` -d . RegressionMapper.java

javac -classpath `yarn classpath` -d . RegressionReducer.java

javac -classpath `yarn classpath`:. -d . Regression.java

jar -cvf regression.jar \*.class

hadoop jar regression.jar Regression final-project/input/clean-dataset.csv final-project/output8

hdfs dfs -cat final-project/output8/part-r-00000