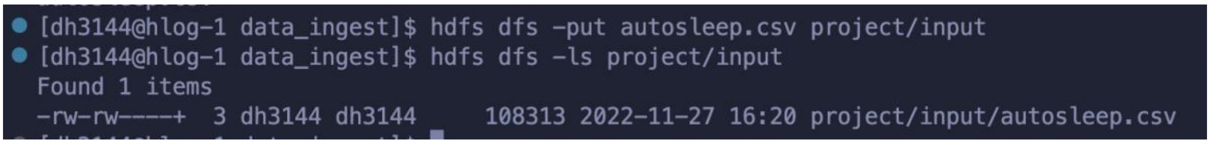
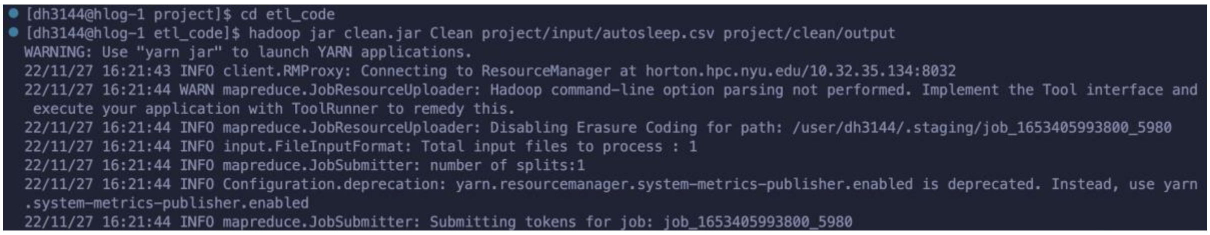
Placing the data in hdfs



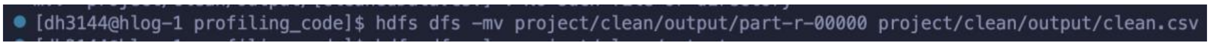
```
● [dh3144@hlog-1 data_ingest]$ hdfs dfs -put autosleep.csv project/input
● [dh3144@hlog-1 data_ingest]$ hdfs dfs -ls project/input
  Found 1 items
  -rw-rw----+  3 dh3144 dh3144     108313 2022-11-27 16:20 project/input/autosleep.csv
```

Running cleaning code



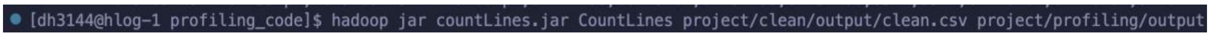
```
● [dh3144@hlog-1 project]$ cd etl_code
● [dh3144@hlog-1 etl_code]$ hadoop jar clean.jar Clean project/input/autosleep.csv project/clean/output
  WARNING: Use "yarn jar" to launch YARN applications.
  22/11/27 16:21:43 INFO client.RMProxy: Connecting to ResourceManager at horton.hpc.nyu.edu/10.32.35.134:8032
  22/11/27 16:21:44 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and
   execute your application with ToolRunner to remedy this.
  22/11/27 16:21:44 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/dh3144/.staging/job_1653405993800_5980
  22/11/27 16:21:44 INFO input.FileInputFormat: Total input files to process : 1
  22/11/27 16:21:44 INFO mapreduce.JobSubmitter: number of splits:1
  22/11/27 16:21:44 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn
  .system-metrics-publisher.enabled
  22/11/27 16:21:44 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1653405993800_5980
```

Converting output file into csv to prepare for data profiling



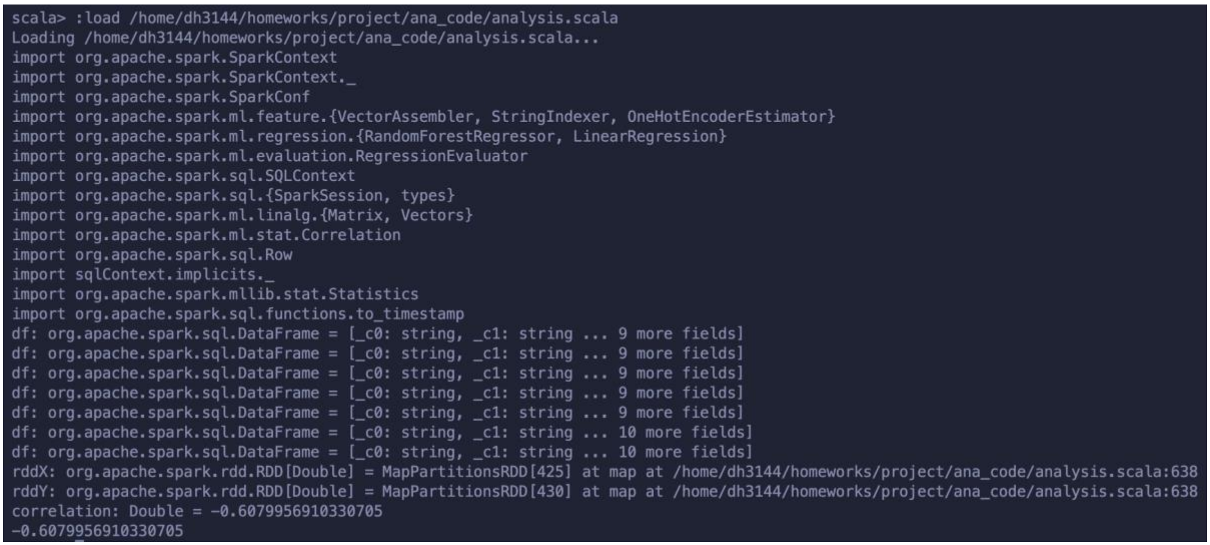
```
● [dh3144@hlog-1 profiling_code]$ hdfs dfs -mv project/clean/output/part-r-00000 project/clean/output/clean.csv
```

Profiling the data

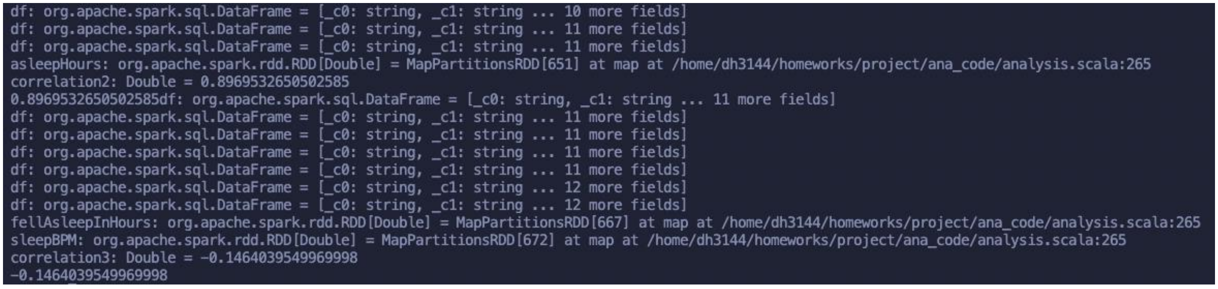


```
● [dh3144@hlog-1 profiling_code]$ hadoop jar countLines.jar CountLines project/clean/output/clean.csv project/profiling/output
```

Running the analysis



```
scala> :load /home/dh3144/homeworks/project/ana_code/analysis.scala
Loading /home/dh3144/homeworks/project/ana_code/analysis.scala...
import org.apache.spark.SparkContext
import org.apache.spark.SparkContext._
import org.apache.spark.SparkConf
import org.apache.spark.ml.feature.{VectorAssembler, StringIndexer, OneHotEncoderEstimator}
import org.apache.spark.ml.regression.{RandomForestRegressor, LinearRegression}
import org.apache.spark.ml.evaluation.RegressionEvaluator
import org.apache.spark.sql.SQLContext
import org.apache.spark.sql.{SparkSession, types}
import org.apache.spark.ml.linalg.{Matrix, Vectors}
import org.apache.spark.ml.stat.Correlation
import org.apache.spark.sql.Row
import sqlContext.implicits._
import org.apache.spark.mllib.stat.Statistics
import org.apache.spark.sql.functions.to_timestamp
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 9 more fields]
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 9 more fields]
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 9 more fields]
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 9 more fields]
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 9 more fields]
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 10 more fields]
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 10 more fields]
rddX: org.apache.spark.rdd.RDD[Double] = MapPartitionsRDD[425] at map at /home/dh3144/homeworks/project/ana_code/analysis.scala:638
rddY: org.apache.spark.rdd.RDD[Double] = MapPartitionsRDD[430] at map at /home/dh3144/homeworks/project/ana_code/analysis.scala:638
correlation: Double = -0.6079956910330705
-0.6079956910330705
```



```
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 10 more fields]
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 11 more fields]
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 11 more fields]
asleepHours: org.apache.spark.rdd.RDD[Double] = MapPartitionsRDD[651] at map at /home/dh3144/homeworks/project/ana_code/analysis.scala:265
correlation2: Double = 0.8969532650502585
0.8969532650502585df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 11 more fields]
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 11 more fields]
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 11 more fields]
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 11 more fields]
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 11 more fields]
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 12 more fields]
df: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 12 more fields]
fellAsleepInHours: org.apache.spark.rdd.RDD[Double] = MapPartitionsRDD[667] at map at /home/dh3144/homeworks/project/ana_code/analysis.scala:265
sleepBPM: org.apache.spark.rdd.RDD[Double] = MapPartitionsRDD[672] at map at /home/dh3144/homeworks/project/ana_code/analysis.scala:265
correlation3: Double = -0.1464039549969998
-0.1464039549969998
```