Making Directories

```
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -mkdir final/ana_code
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -mkdir final/ana_code/patricia
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -mkdir final/data_ingest
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -mkdir final/data_ingest/patricia
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -mkdir final/etl_code
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -mkdir final/etl_code/patricia
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -mkdir final/profiling_code
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -mkdir final/profiling_code/patricia
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -mkdir final/test_code
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -mkdir final/test_code/patricia
```

```
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -mkdir final/ana_code/patricia/nfl_patricia
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -mkdir final/ana_code/patricia/nfl_jason
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -mkdir final/ana_code/patricia/nfl_analytics
```

```
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -mkdir final/etl_code/patricia/nfl_combined
```

Data Profiling

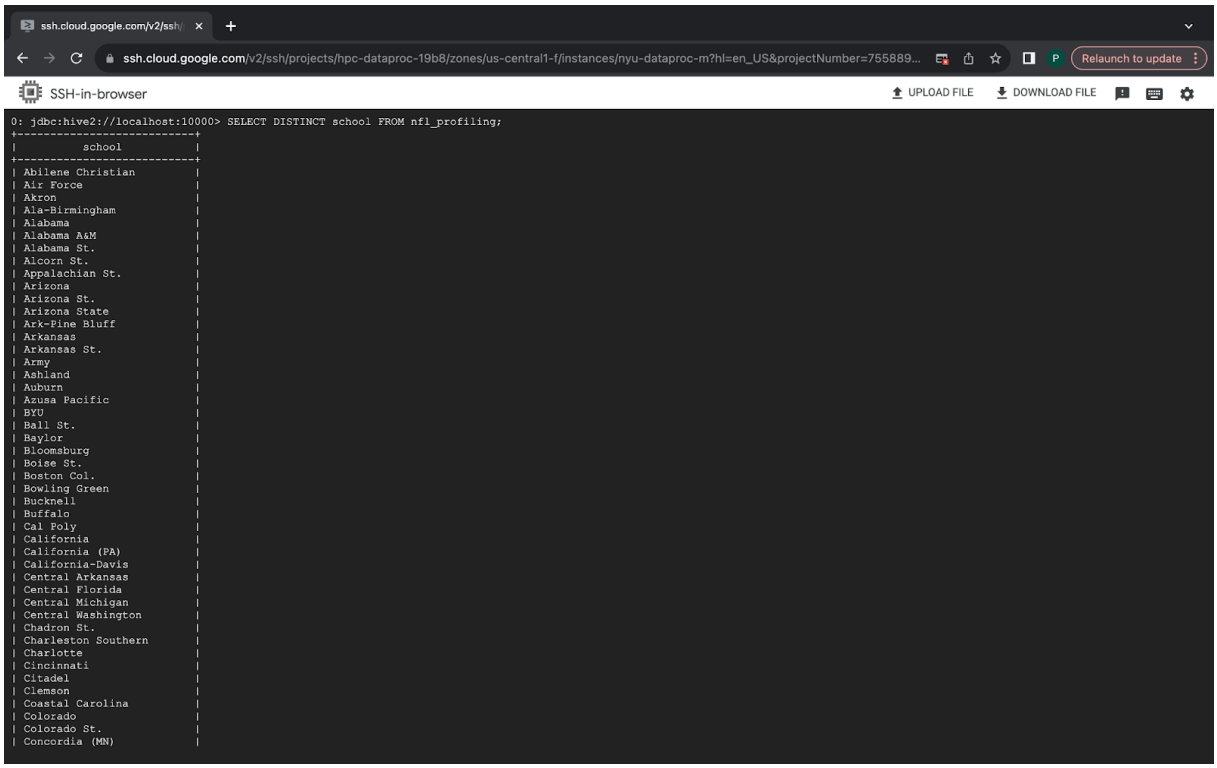
```
ppg2023_nyu_edu@nyu-dataproc-m:~$ ls
NFL.csv  code_profiling_1.scala  code_profiling_2.scala
```

Profiling Numerical Columns

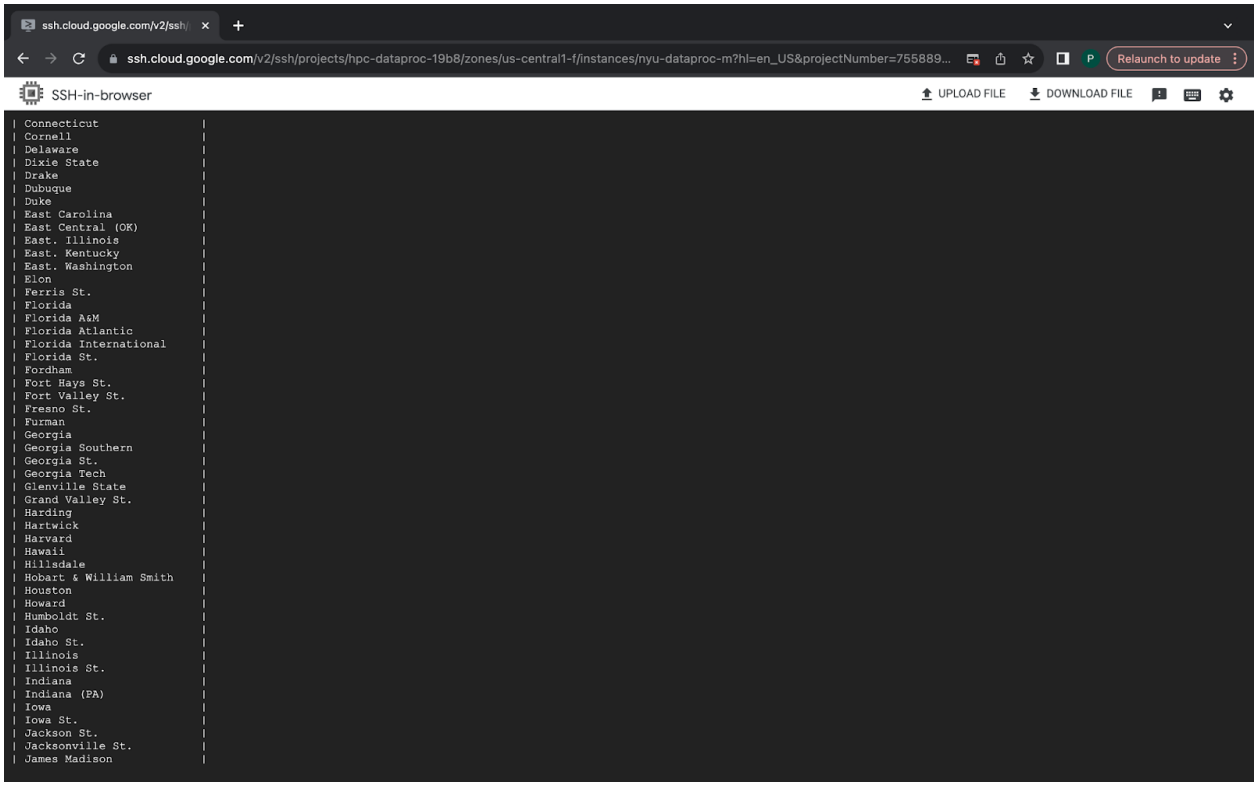
```
ppg2023_nyu_edu@nyu-dataproc-m:~$ spark-shell --deploy-mode client -i  code_profiling_1.scala
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/11/30 22:41:00 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
23/11/30 22:41:00 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
23/11/30 22:41:00 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
23/11/30 22:41:00 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark context Web UI available at http://nyu-dataproc-m.c.hpc-dataproc-19b8.internal:41481
Spark context available as 'sc' (master = yarn, app id = application_1691775874963_35526).
Spark session available as 'spark'.
Standard Deviation for Sprint_40yd: 0.30143134275087086
Mean for Column Year: 2013.8236985907392
Mean for Column Age: 21.983259309873592
Mean for Column Height: 1.8739677308024267
Mean for Column Weight: 109.7463934459305
Mean for Column Sprint_40yd: 4.769079624583718
Mean for Column Vertical_Jump: 83.39240287769773
Mean for Column Bench_Press_Reps: 20.241057542768274
Mean for Column Broad_Jump: 291.6296980720278
Mean for Column Agility_3cone: 7.237415929203538
Mean for Column Shuttle: 4.40384253316216
Mean for Column BMI: 31.074416959209085
Median for Column Year: 2014.0
Median for Column Age: 22.0
Median for Column Height: 1.8796
Median for Column Weight: 104.7798375
Median for Column Sprint_40yd: 4.69
Median for Column Vertical_Jump: 83.82
Median for Column Bench_Press_Reps: 20.0
Median for Column Broad_Jump: 294.64
Median for Column Agility_3cone: 7.14
Median for Column Shuttle: 4.36
Median for Column BMI: 30.12262555
Mode for Column Year: 2014.0
Mode for Column Age: 22.0
Mode for Column Height: 1.905
Mode for Column Weight: 95.70799007
Mode for Column Sprint_40yd: 4.5
Mode for Column Vertical_Jump: 83.82
Mode for Column Bench_Press_Reps: 19.0
Mode for Column Broad_Jump: 304.8
Mode for Column Agility_3cone: 7.07
Mode for Column Shuttle: 4.28
Mode for Column BMI: 37.36882981
```

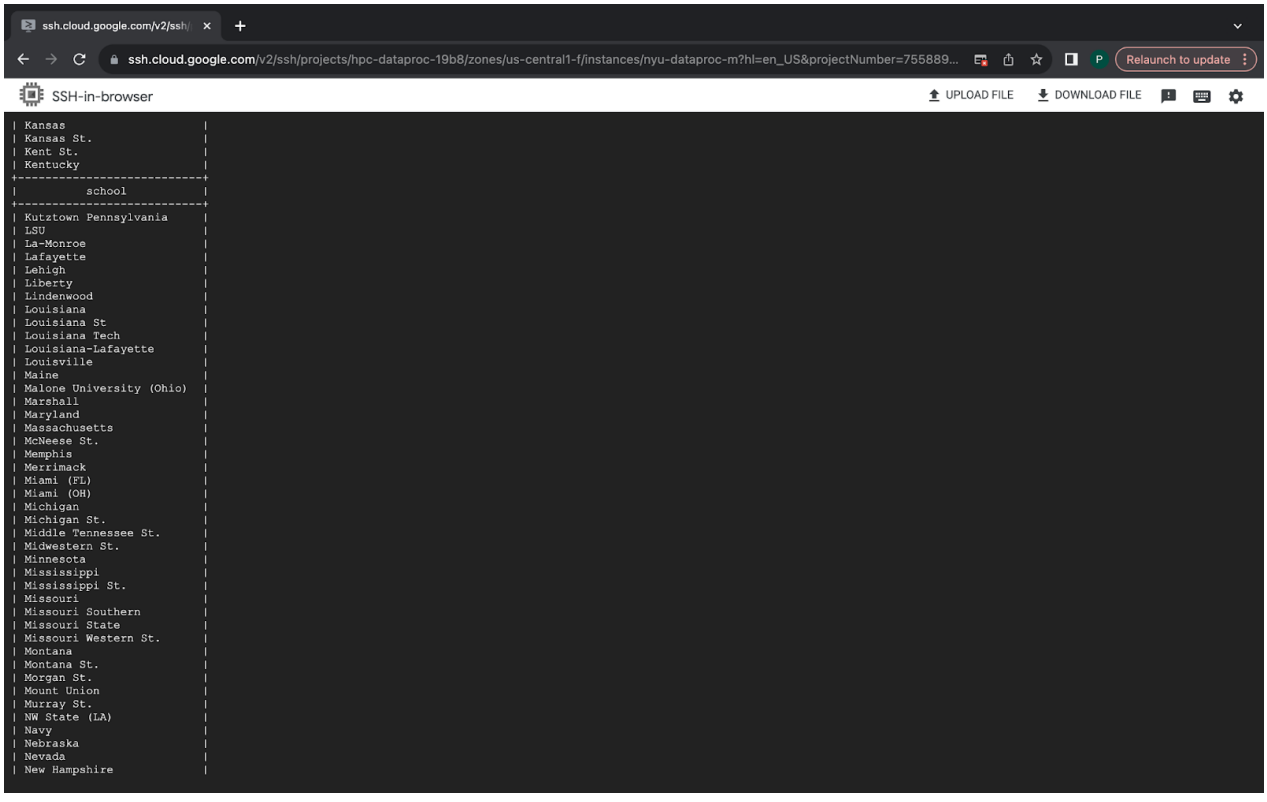
Profiling Non-Numerical Columns

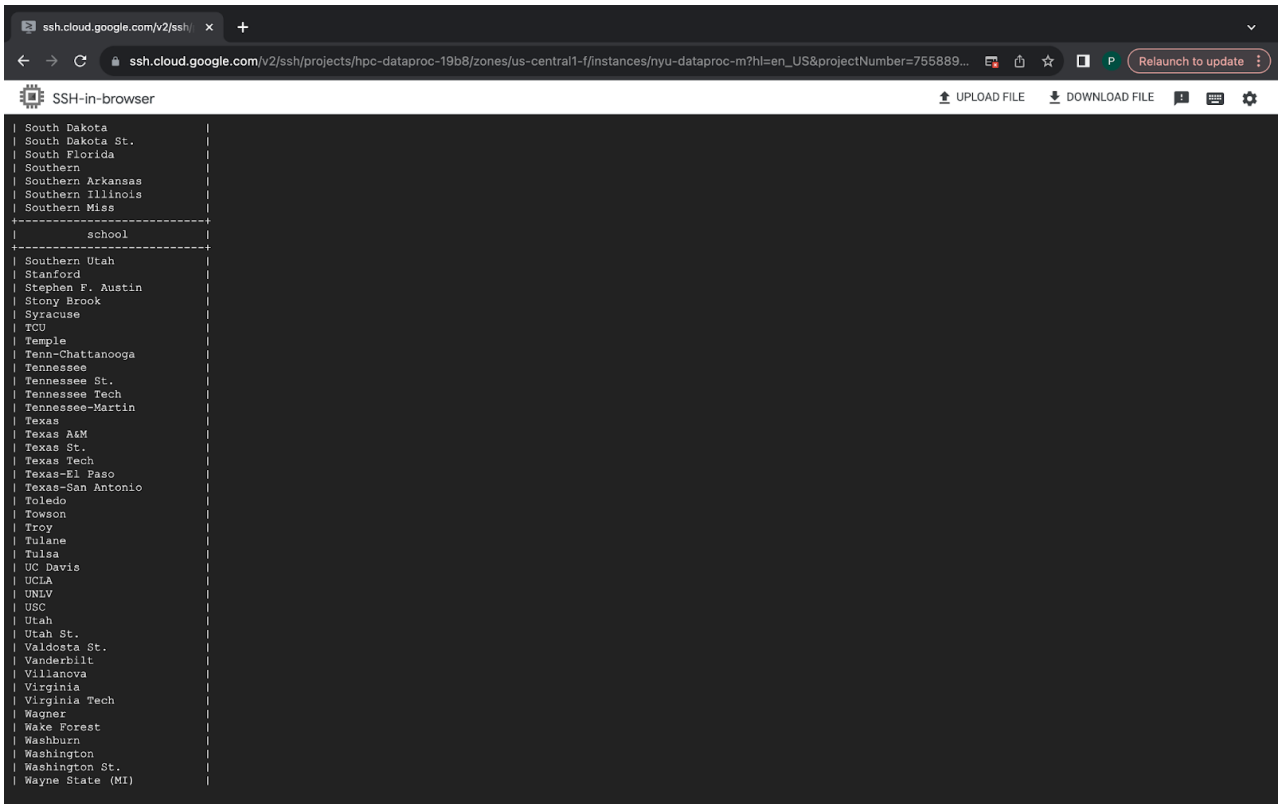
```
scala> ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -copyToLocal hdfs://nyu-dataproc-m/user/ppg2023_nyu_edu/final/NFLProfiling/part-00000
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -put NFlProfilingNonNumeric.csv final/data_ingest/patricia
ppg2023_nyu_edu@nyu-dataproc-m:~$ beeline -u jdbc:hive2://localhost:10000
Connecting to jdbc:hive2://localhost:10000
Connected to: Apache Hive (version 3.1.2)
Driver: Hive JDBC (version 3.1.2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.2 by Apache Hive
0: jdbc:hive2://localhost:10000> USE ppg2023_nyu_edu;
No rows affected (0.123 seconds)
0: jdbc:hive2://localhost:10000> CREATE EXTERNAL TABLE nfl_profiling (
. . . . . .>   school STRING,
. . . . . .>   player_type STRING,
. . . . . .>   position_type STRING,
. . . . . .>   position STRING,
. . . . . .>   drafted STRING
. . . . . .> )
. . . . . .> ROW FORMAT DELIMITED
. . . . . .> FIELDS TERMINATED BY ','
. . . . . .> STORED AS TEXTFILE
. . . . . .> LOCATION 'hdfs://nyu-dataproc-m/user/ppg2023_nyu_edu/final/profiling_code/patricia'
. . . . . .> TBLPROPERTIES ('skip.header.line.count'='1');
No rows affected (0.133 seconds)
0: jdbc:hive2://localhost:10000> LOAD DATA INPATH 'hdfs://nyu-dataproc-m/user/ppg2023_nyu_edu/final/data_ingest/patricia/NFlProfilingNonNumeric.csv' INTO TABLE nfl_profiling;
No rows affected (0.668 seconds)
0: jdbc:hive2://localhost:10000> SELECT * FROM nfl_profiling LIMIT 10;
+-----------------------+---------------------------+-----------------------------+-------------------------+-------------------------+
| nfl_profiling.school  | nfl_profiling.player_type | nfl_profiling.position_type | nfl_profiling.position  | nfl_profiling.drafted   |
+-----------------------+---------------------------+-----------------------------+-------------------------+-------------------------+
| Ohio St.              | offense                   | backs_receivers             | RB                      | Yes                     |
| Illinois              | defense                   | defensive_lineman           | DE                      | Yes                     |
| LSU                   | offense                   | offensive_lineman           | OG                      | Yes                     |
| Alabama               | defense                   | defensive_back              | FS                      | Yes                     |
| Connecticut           | defense                   | line_backer                 | OLB                     | Yes                     |
| Cincinnati            | offense                   | offensive_lineman           | OG                      | Yes                     |
| Mississippi           | defense                   | defensive_lineman           | DT                      | Yes                     |
| North Carolina        | offense                   | offensive_lineman           | OT                      | Yes                     |
| Richmond              | defense                   | defensive_lineman           | DE                      | Yes                     |
| San Jose St.          | defense                   | defensive_back              | CB                      | Yes                     |
+-----------------------+---------------------------+-----------------------------+-------------------------+-------------------------+
10 rows selected (12.375 seconds)
```

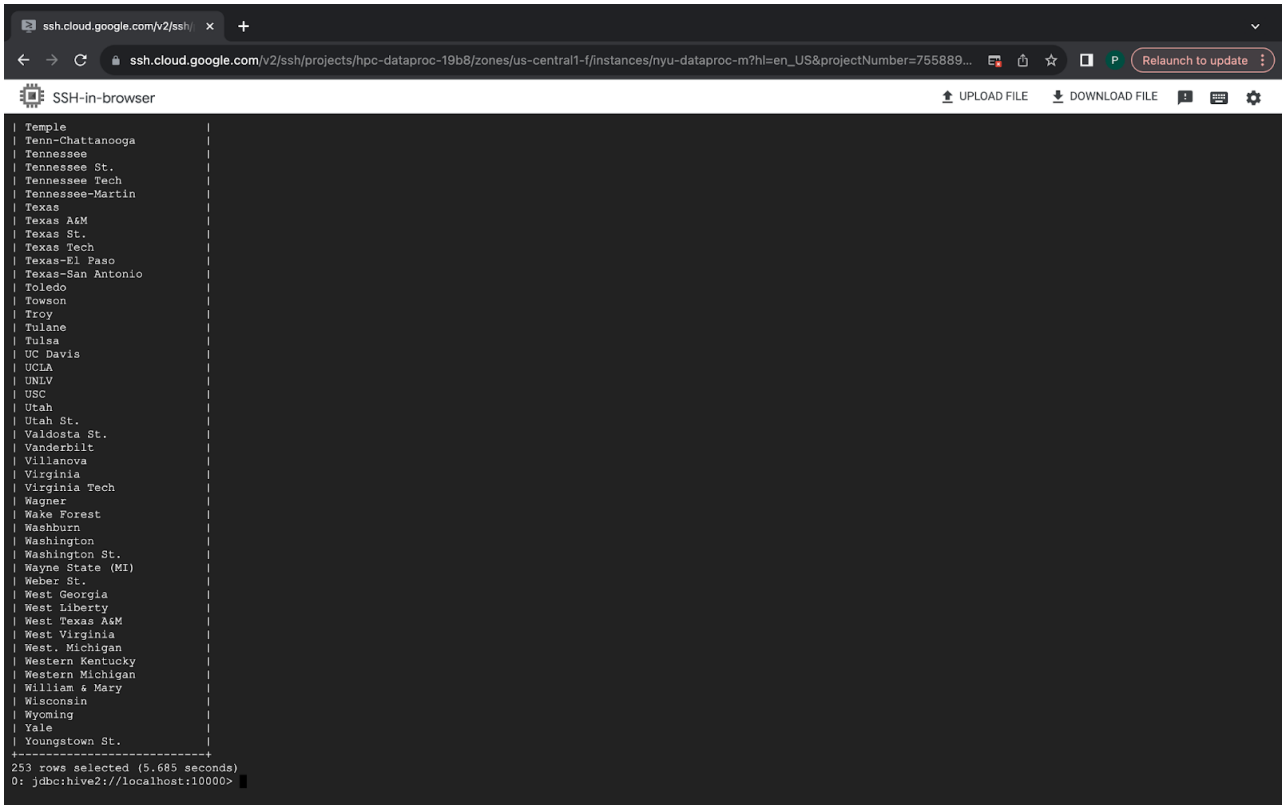



```
0: jdbc:hive2://localhost:10000> SELECT DISTINCT school FROM nfl_profiling;
+--------------------------+
|          school          |
+--------------------------+
| Abilene Christian        |
| Air Force                |
| Akron                    |
| Ala-Birmingham           |
| Alabama                  |
| Alabama A&M              |
| Alabama St.              |
| Alcorn St.               |
| Appalachian St.          |
| Arizona                  |
| Arizona St.              |
| Arizona State            |
| Ark-Pine Bluff           |
| Arkansas                 |
| Arkansas St.             |
| Army                     |
| Ashland                  |
| Auburn                   |
| Azusa Pacific            |
| BYU                      |
| Ball St.                 |
| Baylor                   |
| Bloomsburg               |
| Boise St.                |
| Boston Col.              |
| Bowling Green            |
| Bucknell                 |
| Buffalo                  |
| Cal Poly                 |
| California               |
| California (PA)          |
| California-Davis         |
| Central Arkansas         |
| Central Florida          |
| Central Michigan         |
| Central Washington       |
| Chadron St.              |
| Charleston Southern      |
| Charlotte                |
| Cincinnati               |
| Citadel                  |
| Clemson                  |
| Coastal Carolina         |
| Colorado                 |
| Colorado St.             |
| Concordia (MN)           |
```


ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889...
Relaunch to update
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
| Connecticut |
| Cornell |
| Delaware |
| Dixie State |
| Drake |
| Dubuque |
| Duke |
| East Carolina |
| East Central (OK) |
| East. Illinois |
| East. Kentucky |
| East. Washington |
| Elon |
| Ferris St. |
| Florida |
| Florida A&M |
| Florida Atlantic |
| Florida International |
| Florida St. |
| Fordham |
| Fort Hays St. |
| Fort Valley St. |
| Fresno St. |
| Furman |
| Georgia |
| Georgia Southern |
| Georgia St. |
| Georgia Tech |
| Glenville State |
| Grand Valley St. |
| Harding |
| Hartwick |
| Harvard |
| Hawaii |
| Hillsdale |
| Hobart & William Smith |
| Houston |
| Howard |
| Humboldt St. |
| Idaho |
| Idaho St. |
| Illinois |
| Illinois St. |
| Indiana |
| Indiana (PA) |
| Iowa |
| Iowa St. |
| Jackson St. |
| Jacksonville St. |
| James Madison |


ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889...
Relaunch to update
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
| Kansas |
| Kansas St. |
| Kent St. |
| Kentucky |
+---------------------------+
| school |
+---------------------------+
| Kutztown Pennsylvania |
| LSU |
| La-Monroe |
| Lafayette |
| Lehigh |
| Liberty |
| Lindenwood |
| Louisiana |
| Louisiana St |
| Louisiana Tech |
| Louisiana-Lafayette |
| Louisville |
| Maine |
| Malone University (Ohio) |
| Marshall |
| Maryland |
| Massachusetts |
| McNeese St. |
| Memphis |
| Merrimack |
| Miami (FL) |
| Miami (OH) |
| Michigan |
| Michigan St. |
| Middle Tennessee St. |
| Midwestern St. |
| Minnesota |
| Mississippi |
| Mississippi St. |
| Missouri |
| Missouri Southern |
| Missouri State |
| Missouri Western St. |
| Montana |
| Montana St. |
| Morgan St. |
| Mount Union |
| Murray St. |
| NW State (LA) |
| Navy |
| Nebraska |
| Nevada |
| New Hampshire |


ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889...
Relaunch to update
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
| South Dakota |
| South Dakota St. |
| South Florida |
| Southern |
| Southern Arkansas |
| Southern Illinois |
| Southern Miss |
+---------------------------+
| school |
+---------------------------+
| Southern Utah |
| Stanford |
| Stephen F. Austin |
| Stony Brook |
| Syracuse |
| TCU |
| Temple |
| Tenn-Chattanooga |
| Tennessee |
| Tennessee St. |
| Tennessee Tech |
| Tennessee-Martin |
| Texas |
| Texas A&M |
| Texas St. |
| Texas Tech |
| Texas-El Paso |
| Texas-San Antonio |
| Toledo |
| Towson |
| Troy |
| Tulane |
| Tulsa |
| UC Davis |
| UCLA |
| UNLV |
| USC |
| Utah |
| Utah St. |
| Valdosta St. |
| Vanderbilt |
| Villanova |
| Virginia |
| Virginia Tech |
| Wagner |
| Wake Forest |
| Washburn |
| Washington |
| Washington St. |
| Wayne State (MI) |


ssh.cloud.google.com/v2/ssh/projects/hpc-dataproc-19b8/zones/us-central1-f/instances/nyu-dataproc-m?hl=en_US&projectNumber=755889...
Relaunch to update
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
| Temple |
| Tenn-Chattanooga |
| Tennessee |
| Tennessee St. |
| Tennessee Tech |
| Tennessee-Martin |
| Texas |
| Texas A&M |
| Texas St. |
| Texas Tech |
| Texas-El Paso |
| Texas-San Antonio |
| Toledo |
| Towson |
| Troy |
| Tulane |
| Tulsa |
| UC Davis |
| UCLA |
| UNLV |
| USC |
| Utah |
| Utah St. |
| Valdosta St. |
| Vanderbilt |
| Villanova |
| Virginia |
| Virginia Tech |
| Wagner |
| Wake Forest |
| Washburn |
| Washington |
| Washington St. |
| Wayne State (MI) |
| Weber St. |
| West Georgia |
| West Liberty |
| West Texas A&M |
| West Virginia |
| West. Michigan |
| Western Kentucky |
| Western Michigan |
| William & Mary |
| Wisconsin |
| Wyoming |
| Yale |
| Youngstown St. |
+---------------------------+
253 rows selected (5.685 seconds)
0: jdbc:hive2://localhost:10000>

```
0: jdbc:hive2://localhost:10000> SELECT DISTINCT player_type FROM nfl_profiling;
+----------------+
|  player_type   |
+----------------+
| defense        |
| offense        |
| special_teams  |
+----------------+
3 rows selected (5.375 seconds)
0: jdbc:hive2://localhost:10000> SELECT DISTINCT position_type FROM nfl_profiling;
+---------------------+
|    position_type    |
+---------------------+
| backs_receivers     |
| defensive_back      |
| defensive_lineman   |
| kicking_specialist  |
| line_backer         |
| offensive_lineman   |
| other_special       |
+---------------------+
7 rows selected (5.147 seconds)
0: jdbc:hive2://localhost:10000> SELECT DISTINCT position FROM nfl_profiling;
+-----------+
| position  |
+-----------+
| C         |
| CB        |
| DB        |
| DE        |
| DT        |
| FB        |
| FS        |
| ILB       |
| K         |
| LS        |
| OG        |
| OLB       |
| OT        |
| P         |
| QB        |
| RB        |
| S         |
| SS        |
| TE        |
| WR        |
+-----------+
20 rows selected (1.101 seconds)
```

```
0: jdbc:hive2://localhost:10000> SELECT DISTINCT drafted FROM nfl_profiling;
+-----------+
| drafted   |
+-----------+
| No        |
| Yes       |
+-----------+
2 rows selected (5.404 seconds)
```

Data Cleaning

```
ppg2023_nyu_edu@nyu-dataproc-m:~$ ls
NFL.csv  NFlProfilingNonNumeric.csv  code_cleaning_final.scala  code_profiling_1.scala  code_profiling_2.scala
ppg2023_nyu_edu@nyu-dataproc-m:~$ spark-shell --deploy-mode client -i  code_cleaning_final.scala
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/12/02 20:49:00 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
23/12/02 20:49:00 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
23/12/02 20:49:00 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
23/12/02 20:49:00 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark context Web UI available at http://nyu-dataproc-m.c.hpc-dataproc-19b8.internal:43863
Spark context available as 'sc' (master = yarn, app id = application_1691775874963_36984).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.1.2
      /_/

Using Scala version 2.12.14 (OpenJDK 64-Bit Server VM, Java 1.8.0_322)
Type in expressions to have them evaluated.
Type :help for more information.

scala> ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -copyToLocal hdfs://nyu-dataproc-m/user/ppg2023_nyu_edu/final/NFLCleanFinal/part-00000
ppg2023_nyu_edu@nyu-dataproc-m:~$ ls
NFL.csv  NFlProfilingNonNumeric.csv  code_cleaning_final.scala  code_profiling_1.scala  code_profiling_2.scala  part-00000
```

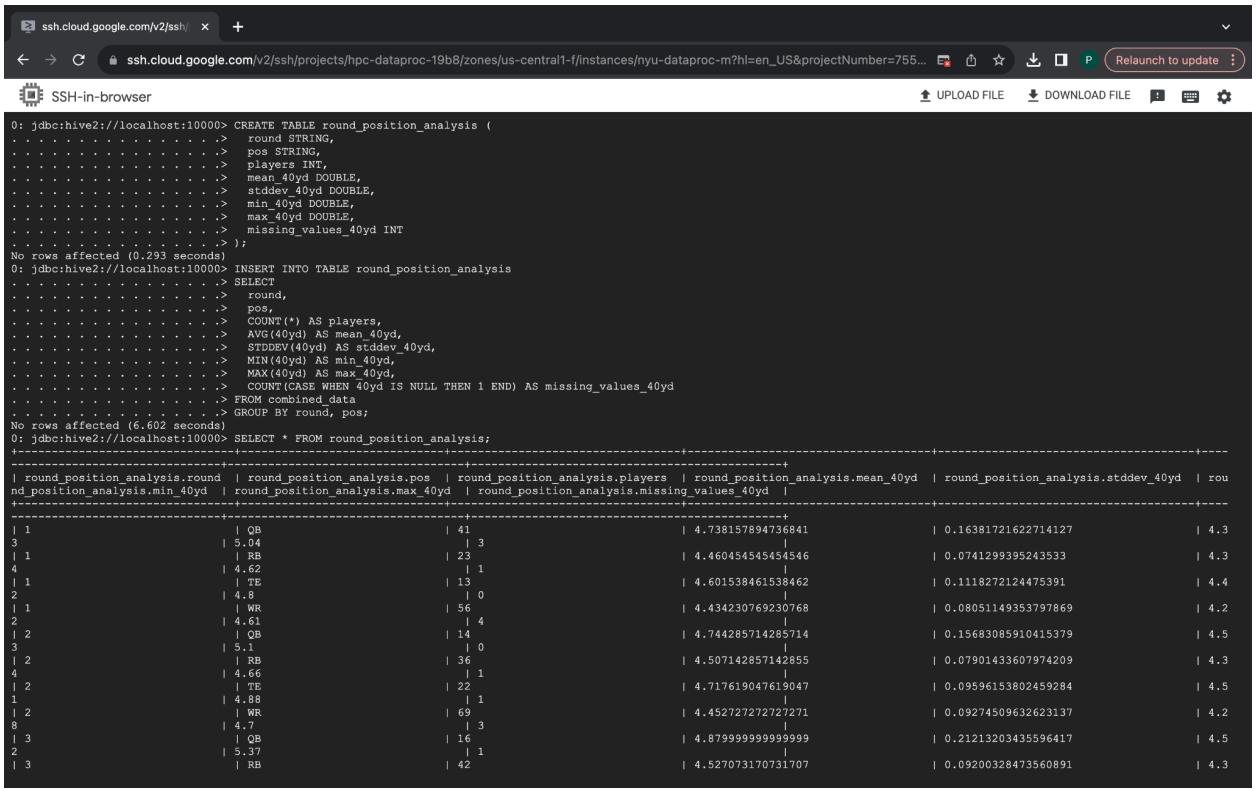
Data Analysis

```
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -put NFLClean.csv final/data_ingest/patricia
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -put cleanedCombineData.csv final/data_ingest/patricia
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -ls final/data_ingest/patricia
Found 2 items
-rw-r--r--   1 ppg2023_nyu_edu ppg2023_nyu_edu      32320 2023-12-02 21:10 final/data_ingest/patricia/NFLClean.csv
-rw-r--r--   1 ppg2023_nyu_edu ppg2023_nyu_edu      13119 2023-12-02 21:10 final/data_ingest/patricia/cleanedCombineData.csv
```

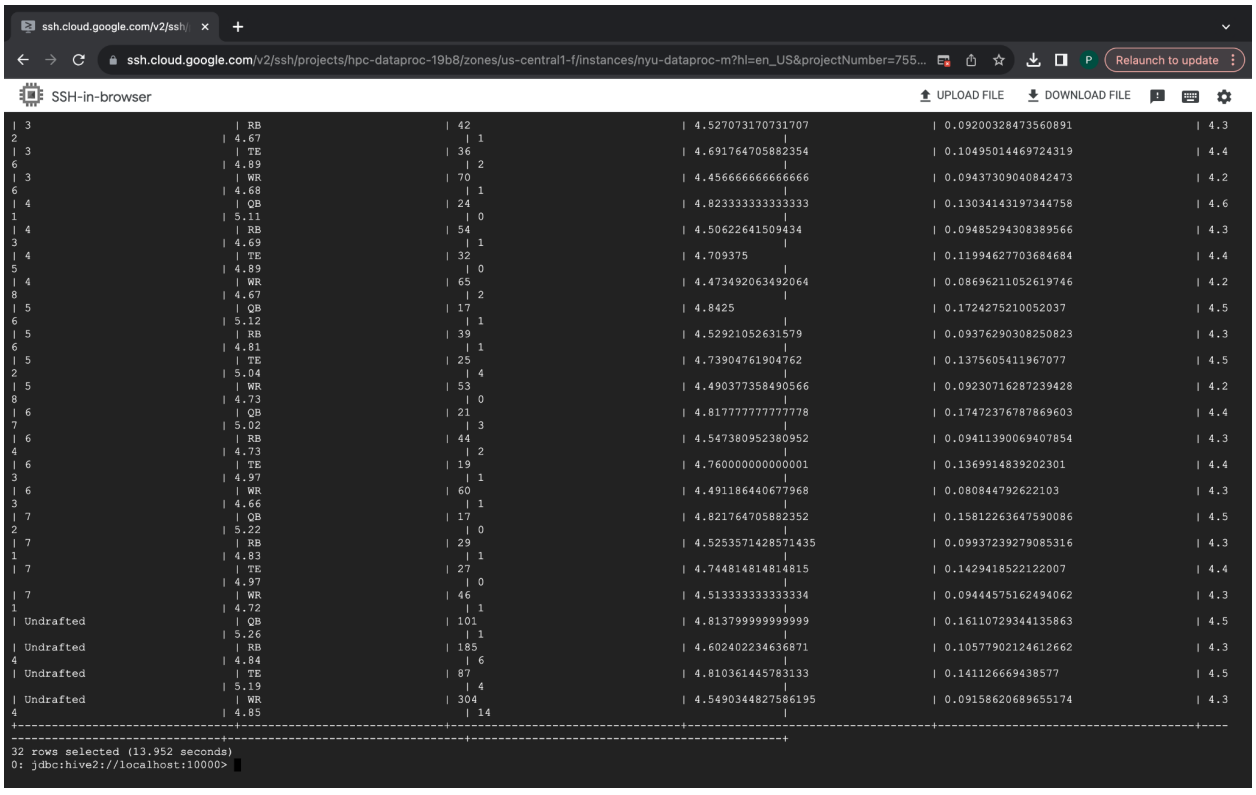
```
ppg2023_nyu_edu@nyu-dataproc-m:~$ beeline -u jdbc:hive2://localhost:10000
Connecting to jdbc:hive2://localhost:10000
Connected to: Apache Hive (version 3.1.2)
Driver: Hive JDBC (version 3.1.2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.2 by Apache Hive
0: jdbc:hive2://localhost:10000> USE ppg2023_nyu_edu;
No rows affected (0.118 seconds)
0: jdbc:hive2://localhost:10000> CREATE EXTERNAL TABLE nfl_pat (
. . . . . .>   player STRING,
. . . . . .>   pos STRING,
. . . . . .>   40yd DOUBLE,
. . . . . .>   round STRING
. . . . . .> )
. . . . . .> ROW FORMAT DELIMITED
. . . . . .> FIELDS TERMINATED BY ','
. . . . . .> STORED AS TEXTFILE
. . . . . .> LOCATION 'hdfs://nyu-dataproc-m/user/ppg2023_nyu_edu/final/ana_code/patricia/nfl_pat'
. . . . . .> TBLPROPERTIES ('skip.header.line.count'='1');
No rows affected (0.119 seconds)
0: jdbc:hive2://localhost:10000> LOAD DATA INPATH 'hdfs://nyu-dataproc-m/user/ppg2023_nyu_edu/final/data_ingest/patricia/NFLClean.csv' INTO TABLE nfl_pat;
No rows affected (0.668 seconds)
0: jdbc:hive2://localhost:10000> SELECT * FROM nfl_pat LIMIT 10;
+-------------------+--------------+---------------+---------------+
|  nfl_pat.player   | nfl_pat.pos  | nfl_pat.40yd  | nfl_pat.round |
+-------------------+--------------+---------------+---------------+
| Beanie Wells      | RB           | 4.38          | 1             |
| Davon Drew        | TE           | 4.78          | 5             |
| Cedric Peerman    | RB           | 4.34          | 6             |
| Shawn Nelson      | TE           | 4.52          | 4             |
| Mike Goodson      | RB           | 4.43          | 4             |
| Johnny Knox       | WR           | 4.34          | 5             |
| Juaquin Iglesias  | WR           | 4.44          | 3             |
| Chase Coffman     | TE           | 4.83          | 3             |
| Bernard Scott     | RB           | 4.44          | 6             |
| James Davis       | RB           | 4.45          | 6             |
+-------------------+--------------+---------------+---------------+
10 rows selected (11.773 seconds)
0: jdbc:hive2://localhost:10000> SELECT COUNT(*) FROM nfl_pat;
+-------+
|  _c0  |
+-------+
| 1211  |
+-------+
1 row selected (1.422 seconds)
```

```
0: jdbc:hive2://localhost:10000> CREATE EXTERNAL TABLE nfl_jason (
. . . . . .>   player STRING,
. . . . . .>   pos STRING,
. . . . . .>   40yd DOUBLE,
. . . . . .>   round STRING
. . . . . .> )
. . . . . .> ROW FORMAT DELIMITED
. . . . . .> FIELDS TERMINATED BY ','
. . . . . .> STORED AS TEXTFILE
. . . . . .> LOCATION 'hdfs://nyu-dataproc-m/user/ppg2023_nyu_edu/final/ana_code/patricia/nfl_jason'
. . . . . .> TBLPROPERTIES ('skip.header.line.count'='1');
No rows affected (0.099 seconds)
0: jdbc:hive2://localhost:10000> LOAD DATA INPATH 'hdfs://nyu-dataproc-m/user/ppg2023_nyu_edu/final/data_ingest/patricia/cleanedCombineData.csv' INTO TABLE nfl_jason;
No rows affected (0.667 seconds)
0: jdbc:hive2://localhost:10000> SELECT * FROM nfl_jason LIMIT 10;
+-----------------+---------------+----------------+-----------------+
| nfl_jason.player  | nfl_jason.pos  | nfl_jason.40yd  | nfl_jason.round  |
+-----------------+---------------+----------------+-----------------+
| Jonathan Adams    | WR             | 4.59           | Undrafted        |
| Giles Amos        | TE             | 5.14           | Undrafted        |
| Otis Anderson     | RB             | 4.63           | Undrafted        |
| Tutu Atwell       | WR             | 4.42           | 2                |
| Kawaan Baker      | WR             | 4.44           | Undrafted        |
| Rashod Bateman    | WR             | 4.39           | 1                |
| John Bates        | TE             | 4.84           | Undrafted        |
| Shaun Beyer       | TE             | 4.78           | Undrafted        |
| Tarik Black       | WR             | 4.54           | Undrafted        |
| Josh Blancato     | WR             | 4.55           | Undrafted        |
+-----------------+---------------+----------------+-----------------+
10 rows selected (14.145 seconds)
0: jdbc:hive2://localhost:10000> SELECT COUNT(*) FROM nfl_jason;
+------+
| _c0  |
+------+
| 476  |
+------+
1 row selected (8.86 seconds)
```

```
0: jdbc:hive2://localhost:10000> CREATE TABLE combined_data AS
. . . . . .> SELECT * FROM nfl_pat
. . . . . .> UNION ALL
. . . . . .> SELECT * FROM nfl_jason;
No rows affected (6.678 seconds)
0: jdbc:hive2://localhost:10000> SELECT COUNT(*) FROM combined_data;
+-------+
|  _c0  |
+-------+
| 1687  |
+-------+
1 row selected (1.55 seconds)
```




```
0: jdbc:hive2://localhost:10000> CREATE TABLE round_position_analysis (
. . . . . .>   round STRING,
. . . . . .>   pos STRING,
. . . . . .>   players INT,
. . . . . .>   mean_40yd DOUBLE,
. . . . . .>   stddev_40yd DOUBLE,
. . . . . .>   min_40yd DOUBLE,
. . . . . .>   max_40yd DOUBLE,
. . . . . .>   missing_values_40yd INT
. . . . . .> );
No rows affected (0.293 seconds)
0: jdbc:hive2://localhost:10000> INSERT INTO TABLE round_position_analysis
. . . . . .> SELECT
. . . . . .>   round,
. . . . . .>   pos,
. . . . . .>   COUNT(*) AS players,
. . . . . .>   AVG(40yd) AS mean_40yd,
. . . . . .>   STDDEV(40yd) AS stddev_40yd,
. . . . . .>   MIN(40yd) AS min_40yd,
. . . . . .>   MAX(40yd) AS max_40yd,
. . . . . .>   COUNT(CASE WHEN 40yd IS NULL THEN 1 END) AS missing_values_40yd
. . . . . .> FROM combined_data
. . . . . .> GROUP BY round, pos;
No rows affected (6.602 seconds)
0: jdbc:hive2://localhost:10000> SELECT * FROM round_position_analysis;
+-------------------------------+-------------------------------+---------------------------------+---------------------------------+-----------------------------------+----
-------------------------------+-------------------------------+---------------------------------+-+
| round_position_analysis.round  | round_position_analysis.pos  | round_position_analysis.players  | round_position_analysis.mean_40yd  | round_position_analysis.stddev_40yd  | rou
nd_position_analysis.min_40yd  | round_position_analysis.max_40yd  | round_position_analysis.missing_values_40yd  |
+-------------------------------+-------------------------------+---------------------------------+---------------------------------+-----------------------------------+----
-------------------------------+-------------------------------+---------------------------------+-+
| 1                              | QB                            | 41                              | 4.738157894736841               | 0.16381721622714127               | 4.3
3                              | 5.04                          | 3                               |
| 1                              | RB                            | 23                              | 4.460454545454546               | 0.0741299395243533                | 4.3
4                              | 4.62                          | 1                               |
| 1                              | TE                            | 13                              | 4.601538461538462               | 0.1118272124475391                | 4.4
2                              | 4.8                           | 0                               |
| 1                              | WR                            | 56                              | 4.434230769230768               | 0.080511493537978 69              | 4.2
2                              | 4.61                          | 4                               |
| 2                              | QB                            | 14                              | 4.744285714285714               | 0.15683085910415379               | 4.5
3                              | 5.1                           | 0                               |
| 2                              | RB                            | 36                              | 4.507142857142855               | 0.07901433607974209               | 4.3
4                              | 4.66                          | 1                               |
| 2                              | TE                            | 22                              | 4.717619047619047               | 0.09596153802459284               | 4.5
1                              | 4.88                          | 1                               |
| 2                              | WR                            | 69                              | 4.452727272727271               | 0.09274509632623137               | 4.2
8                              | 4.7                           | 3                               |
| 3                              | QB                            | 16                              | 4.879999999999999               | 0.21213203435596417               | 4.5
2                              | 5.37                          | 1                               |
| 3                              | RB                            | 42                              | 4.527073170731707               | 0.09200328473560891               | 4.3
```




```
| 3                              | RB                            | 42                              | 4.527073170731707               | 0.09200328473560891               | 4.3
2                              | 4.67                          | 1                               |
| 3                              | TE                            | 36                              | 4.691764705882354               | 0.10495014469724319               | 4.4
6                              | 4.89                          | 2                               |
| 3                              | WR                            | 70                              | 4.456666666666666               | 0.09437309040842473               | 4.2
6                              | 4.68                          | 1                               |
| 4                              | QB                            | 24                              | 4.823333333333333               | 0.13034143197344758               | 4.6
1                              | 5.11                          | 0                               |
| 4                              | RB                            | 54                              | 4.50622641509434                | 0.09485294308389566               | 4.3
3                              | 4.69                          | 1                               |
| 4                              | TE                            | 32                              | 4.709375                        | 0.11994627703684684               | 4.4
5                              | 4.89                          | 0                               |
| 4                              | WR                            | 65                              | 4.473492063492064               | 0.08696211052619746               | 4.2
8                              | 4.67                          | 2                               |
| 5                              | QB                            | 17                              | 4.8425                          | 0.1724275210052037                | 4.5
6                              | 5.12                          | 1                               |
| 5                              | RB                            | 39                              | 4.529210526315 79               | 0.09376290308250823               | 4.3
6                              | 4.81                          | 1                               |
| 5                              | TE                            | 25                              | 4.739047619047 62               | 0.1375605411967077                | 4.5
2                              | 5.04                          | 4                               |
| 5                              | WR                            | 53                              | 4.490377358490566               | 0.09230716287239428               | 4.2
8                              | 4.73                          | 0                               |
| 6                              | QB                            | 21                              | 4.817777777777778               | 0.17472376787869603               | 4.4
7                              | 5.02                          | 3                               |
| 6                              | RB                            | 44                              | 4.547380952380952               | 0.09411390069407854               | 4.3
4                              | 4.73                          | 2                               |
| 6                              | TE                            | 19                              | 4.760000000000001               | 0.1369914839202301                | 4.4
3                              | 4.97                          | 1                               |
| 6                              | WR                            | 60                              | 4.491186440677968               | 0.080844792622103                 | 4.3
3                              | 4.66                          | 1                               |
| 7                              | QB                            | 17                              | 4.821764705882352               | 0.15812263647590086               | 4.5
2                              | 5.22                          | 0                               |
| 7                              | RB                            | 29                              | 4.5253571428571435              | 0.09937239279085316               | 4.3
1                              | 4.83                          | 1                               |
| 7                              | TE                            | 27                              | 4.744814814814815               | 0.1429418522122007                | 4.4
4                              | 4.97                          | 0                               |
| 7                              | WR                            | 46                              | 4.513333333333334               | 0.09444575162494062               | 4.3
1                              | 4.72                          | 1                               |
| Undrafted                      | QB                            | 101                             | 4.813799999999999               | 0.16110729344135863               | 4.5
5                              | 5.26                          | 1                               |
| Undrafted                      | RB                            | 185                             | 4.602402234636871               | 0.10577902124612662               | 4.3
4                              | 4.84                          | 6                               |
| Undrafted                      | TE                            | 87                              | 4.810361445783133               | 0.141126669438577                 | 4.5
5                              | 5.19                          | 4                               |
| Undrafted                      | WR                            | 304                             | 4.5490344827586195              | 0.09158620689655174               | 4.3
4                              | 4.85                          | 14                              |
+-------------------------------+-------------------------------+---------------------------------+---------------------------------+-----------------------------------+----
-------------------------------+-------------------------------+---------------------------------+-+
32 rows selected (13.952 seconds)
0: jdbc:hive2://localhost:10000>
```

```
0: jdbc:hive2://localhost:10000> INSERT OVERWRITE DIRECTORY 'hdfs://nyu-dataproc-m/user/ppg2023_nyu_edu/final/ana_code/patricia/nfl_analytics'
. . . . . .> ROW FORMAT DELIMITED
. . . . . .> FIELDS TERMINATED BY ','
. . . . . .> SELECT * FROM round_position_analysis;
No rows affected (20.028 seconds)
0: jdbc:hive2://localhost:10000> INSERT OVERWRITE DIRECTORY 'hdfs://nyu-dataproc-m/user/ppg2023_nyu_edu/final/etl_code/patricia/nfl_combined'
. . . . . .> ROW FORMAT DELIMITED
. . . . . .> FIELDS TERMINATED BY ','
. . . . . .> SELECT * FROM combined_data;
No rows affected (1.377 seconds)
```

Putting Everything In HDFS

```
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -put code_profiling_1.scala final/profiling_code/patricia
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -put code_profiling_2.scala final/profiling_code/patricia
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -put code_profiling_non_numeric.hql final/profiling_code/patricia
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -put code_cleaning_final.scala final/etl_code/patricia
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -put code_analytics.hql final/ana_code/patricia
```

```
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -put README.txt final
ppg2023_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -put hdfs_commands.txt final
```