Project 01 Report

Aman Tej Vidapu

Step 1:

Setup the connections Java, Hadoop and Spark sessions. [hint: for these connections, go back to your spark exercise in the IAF Python class with Dr. Kopper.]

```
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget -q
=https://www-us.apache.org/dist/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.7.t
gz
!tar xf spark-2.4.5-bin-hadoop2.7.tgz
!pip install -q findspark
```

```
--------------------------------------------------------------------------------
```

```
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-2.4.5-bin-hadoop2.7"
os.environ["PYSPARK_SUBMIR_ARGS"] = "--MASTER LOCAL[2] pyspark-shell"
```

```
--------------------------------------------------------------------------------
```

```
import findspark
findspark.init()
from pyspark.sql import SparkSession
spark=SparkSession.builder.master("local[1]").appName("Project01").getOrCreat
e()
sparkContext=spark.sparkContext
```

```
--------------------------------------------------------------------------------
```

Connections are made.

Step 2:

Read the file into spark. File is provided with header so, no need to worry about adding column names

#Now reading the dataset using relative pathing.

```
dataset = spark.read.csv('../data/carwood.csv',inferSchema = True, header =
True)
```

#Displaying the first 5 rows of the dataset.

```
dataset.show(5)
```

output:

```
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+-----+
|    f1|    f2|    f3|    f4|    f5|    f6|    f7|    f8|    f9|   f10|
f11|   f12|   f13|   f14|   f15|   f16|   f17|   f18|   f19|   f20|   f21|
f22|   f23|   f24|   f25|   f26|   f27|   f28|   f29|   f30|   f31|   f32|
f33|   f34|   f35|   f36|   f37|   f38|   f39|   f40|   f41|   f42|   f43|
f44|   f45|   f46|   f47|   f48|   f49|   f50|   f51|   f52|   f53|   f54|
f55|   f56|   f57|   f58|   f59|   f60|   f61|   f62|   f63|   f64|   f65|
f66|   f67|label|
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+-----+
|170.39|167.28|143.44|124.67|139.01|125.83|144.33|151.26|175.51|171.31|
161.9|146.92| 141.8|140.91|
132.8|128.48|170.83|161.06|169.61|168.07|154.88|149.33|152.91|152.61|137.09|1
69.59|183.01|180.44|164.55|152.63|157.77|165.07|149.52|142.73|173.85|182.23|1
74.58|163.28|164.01|
169.2|148.48|134.31|174.07|191.33|157.51|168.66|156.89|170.86|162.24|184.84|1
67.02|123.67|140.54|153.69|147.57|144.65|162.24|172.96|169.67|157.51|161.06|1
33.23|124.41|138.44|142.93|137.13|134.44|    0|
|169.75|190.96|175.53|138.27|137.47|139.23|133.23|130.25|147.73|163.93|167.36
|171.52|155.54|139.34|151.95| 149.3|173.37|
141.0|153.57|128.45|159.93|165.33|147.94|143.74|140.88|182.53|184.08|148.09|1
25.26|139.67| 138.7|132.86| 141.5| 145.4|164.58|170.71|127.83|133.99|
141.2|152.87|142.36|148.26|162.54|156.55|153.39|137.99|129.64|137.59|155.39|1
47.92|152.14|162.48|168.72|161.14|147.87|141.02|155.39|139.58|141.58|153.39|
141.0|148.43|168.12| 169.9|165.64|166.86|137.69|    0|
|153.69|153.68|144.02|158.73|178.87|157.04|152.92|147.52|142.87|165.26|160.39
|137.86|149.62|153.43| 152.6|162.85|146.35|167.11|134.27|126.81|136.28|
158.4|171.58| 161.6|162.91|143.14|129.73|125.97|151.09|177.89|
169.2|160.65|156.86|135.66|126.92|131.22|154.37|158.59|158.52|155.25|154.33|1
30.18|127.66|148.94|155.37|163.53| 139.7|143.79|141.61|166.88|
164.6|149.58|139.56|154.74|173.01|155.18|141.61|155.19|170.51|155.37|167.11|1
```

```
46.89|141.01|159.43|169.68|163.24|165.17|    0|
|131.69|151.56|151.05| 134.0|151.18|175.53|171.34|159.77|151.95|
146.1|148.53|140.28|138.16|145.44| 150.4|158.18| 163.8|152.43|171.49|
150.2|131.28|157.18|157.04|151.13|151.66|143.32|157.23|152.91|134.75|154.65|1
71.65| 160.3|157.73|143.67|145.87|151.73|147.09|151.21|157.96| 148.5|
156.6|147.45|153.18|156.58|157.83|147.98|143.47|142.83|138.08|147.91|148.05|1
45.66|
156.5|167.52|151.33|129.51|138.08|164.25|155.82|157.83|152.43|150.82|146.58|1
28.85|140.76|177.35|174.61|    0|
|162.85|158.88|132.27|138.41|143.98|
159.3|177.26|180.58|159.34|164.66|138.04|132.76|157.88|165.58|173.64|
163.5|127.97|167.31|141.39|147.02|137.52|135.46|146.41|159.09|164.53|148.45|1
30.76|136.22|144.86|127.38|137.09|159.08|153.25|182.49|187.75|139.37|117.12|1
24.07| 134.6|144.85|132.64|170.87|188.26|173.32|135.74|127.15|131.69|
127.8|149.46|114.16|100.11|154.02| 175.3|175.46|144.39|142.47|149.46|
132.8|130.96|135.74|167.31|188.21|179.52| 146.2|153.73|152.12|146.58|    0|
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+-----+
only showing top 5 rows
```

#Here displaying the table/dataframe dimensions and saving shape(dimensions).

```
print('Shape: Rows:', dataset.count(),',','Columns:',len(dataset.columns))
shape = [dataset.count(),len(dataset.columns)]
```

output:

```
Shape: Rows: 2048 , Columns: 68
```

Step 3:

Do the basic necessary things to understand the dataset, such as datatypes, check null values, and statistical information.

My Understanding:

Here, I have explored the data in terms of missing values, data types, data structure and investigating duplicate columns. Also made some statistical observations.

Missing values: There is no missing value/Null value in the dataset. Every column was checked for missing values and count of null values returned.

Data types: The dataset has a uniform data type across all dimensions. The data type is "double"

Data structure dimensions: The dimension of the data is (2048 x 68) i.e there are 2048 observations/ records and 69 attributes in the dataset.

Duplicate columns: There are three sets of duplicate columns in the dataset. The duplicate sets are (f18, f61), (f45, f60), (f49, f57). The duplicate of each set has been dropped in the codes below.

Statistical observations: Every column has similar max, min and standard deviation in the overall column range.

#Printing General schema of the dataset.

#observe here nullable = true, means no null values.

```
dataset.printSchema()
```

output:

```
root
 |-- f1: double (nullable = true)
 |-- f2: double (nullable = true)
 |-- f3: double (nullable = true)
 |-- f4: double (nullable = true)
 |-- f5: double (nullable = true)
 |-- f6: double (nullable = true)
 |-- f7: double (nullable = true)
 |-- f8: double (nullable = true)
 |-- f9: double (nullable = true)
 |-- f10: double (nullable = true)
 |-- f11: double (nullable = true)
 |-- f12: double (nullable = true)
 |-- f13: double (nullable = true)
 |-- f14: double (nullable = true)
 |-- f15: double (nullable = true)
 |-- f16: double (nullable = true)
 |-- f17: double (nullable = true)
 |-- f18: double (nullable = true)
 |-- f19: double (nullable = true)
 |-- f20: double (nullable = true)
 |-- f21: double (nullable = true)
 |-- f22: double (nullable = true)
 |-- f23: double (nullable = true)
 |-- f24: double (nullable = true)
 |-- f25: double (nullable = true)
```

```
 |-- f26: double (nullable = true)
 |-- f27: double (nullable = true)
 |-- f28: double (nullable = true)
 |-- f29: double (nullable = true)
 |-- f30: double (nullable = true)
 |-- f31: double (nullable = true)
 |-- f32: double (nullable = true)
 |-- f33: double (nullable = true)
 |-- f34: double (nullable = true)
 |-- f35: double (nullable = true)
 |-- f36: double (nullable = true)
 |-- f37: double (nullable = true)
 |-- f38: double (nullable = true)
 |-- f39: double (nullable = true)
 |-- f40: double (nullable = true)
 |-- f41: double (nullable = true)
 |-- f42: double (nullable = true)
 |-- f43: double (nullable = true)
 |-- f44: double (nullable = true)
 |-- f45: double (nullable = true)
 |-- f46: double (nullable = true)
 |-- f47: double (nullable = true)
 |-- f48: double (nullable = true)
 |-- f49: double (nullable = true)
 |-- f50: double (nullable = true)
 |-- f51: double (nullable = true)
 |-- f52: double (nullable = true)
 |-- f53: double (nullable = true)
 |-- f54: double (nullable = true)
 |-- f55: double (nullable = true)
 |-- f56: double (nullable = true)
 |-- f57: double (nullable = true)
 |-- f58: double (nullable = true)
 |-- f59: double (nullable = true)
 |-- f60: double (nullable = true)
 |-- f61: double (nullable = true)
 |-- f62: double (nullable = true)
 |-- f63: double (nullable = true)
 |-- f64: double (nullable = true)
 |-- f65: double (nullable = true)
 |-- f66: double (nullable = true)
 |-- f67: double (nullable = true)
 |-- label: integer (nullable = true)
```

#Printing Datatypes of each column of the dataset.

```
dataset.dtypes
```

output:

```
[('f1', 'double'),
 ('f2', 'double'),
 ('f3', 'double'),
 ('f4', 'double'),
 ('f5', 'double'),
 ('f6', 'double'),
 ('f7', 'double'),
 ('f8', 'double'),
 ('f9', 'double'),
 ('f10', 'double'),
 ('f11', 'double'),
 ('f12', 'double'),
 ('f13', 'double'),
 ('f14', 'double'),
 ('f15', 'double'),
 ('f16', 'double'),
 ('f17', 'double'),
 ('f18', 'double'),
 ('f19', 'double'),
 ('f20', 'double'),
 ('f21', 'double'),
 ('f22', 'double'),
 ('f23', 'double'),
 ('f24', 'double'),
 ('f25', 'double'),
 ('f26', 'double'),
 ('f27', 'double'),
 ('f28', 'double'),
 ('f29', 'double'),
 ('f30', 'double'),
 ('f31', 'double'),
 ('f32', 'double'),
 ('f33', 'double'),
 ('f34', 'double'),
 ('f35', 'double'),
 ('f36', 'double'),
 ('f37', 'double'),
 ('f38', 'double'),
 ('f39', 'double'),
 ('f40', 'double'),
 ('f41', 'double'),
 ('f42', 'double'),
 ('f43', 'double'),
 ('f44', 'double'),
 ('f45', 'double'),
 ('f46', 'double'),
```

```
 ('f47', 'double'),
 ('f48', 'double'),
 ('f49', 'double'),
 ('f50', 'double'),
 ('f51', 'double'),
 ('f52', 'double'),
 ('f53', 'double'),
 ('f54', 'double'),
 ('f55', 'double'),
 ('f56', 'double'),
 ('f57', 'double'),
 ('f58', 'double'),
 ('f59', 'double'),
 ('f60', 'double'),
 ('f61', 'double'),
 ('f62', 'double'),
 ('f63', 'double'),
 ('f64', 'double'),
 ('f65', 'double'),
 ('f66', 'double'),
 ('f67', 'double'),
 ('label', 'int')]
```

#Again, counting the null values if present in dataset and displaying through spark dataframe.

```
from pyspark.sql.functions import col,isnan,when,count
dataset.select([count(when(col(c).contains('None') | \
                            col(c).contains('NULL') | \
                            col(c).isNull() | \
                            isnan(c), c
                           )).alias(c)
                    for c in dataset.columns]).show()
```

output:

```
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+-
--+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+--
-+---+---+---+---+---+---+---+---+---+-----+
| f1| f2| f3| f4| f5| f6| f7| f8|
f9|f10|f11|f12|f13|f14|f15|f16|f17|f18|f19|f20|f21|f22|f23|f24|f25|f26|f27|f2
8|f29|f30|f31|f32|f33|f34|f35|f36|f37|f38|f39|f40|f41|f42|f43|f44|f45|f46|f47
|f48|f49|f50|f51|f52|f53|f54|f55|f56|f57|f58|f59|f60|f61|f62|f63|f64|f65|f66|
f67|label|
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+-
--+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+--
-+---+---+---+---+---+---+---+---+---+-----+
```

```
|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|
0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|
0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|  0|
0|  0|  0|  0|  0|  0|  0|  0|  0|  0|    0|
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+-
--+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+--
-+---+---+---+---+---+---+---+---+---+-----+
```

#Using Summary() of spark to describe the statistical information of the dataset.

```
#Here summary() is similar to describe() of pandas.
#We can get column wise count, max, min, quantiles..,

dataset.summary().show()
```

output:

```
+-------+-----------------+-----------------+-----------------+-----------------+----------
-------+-----------------+-----------------+-----------------+-----------------+------------
----+-----------------+-----------------+-----------------+-----------------+----------------
-+-----------------+-----------------+-----------------+-----------------+----------------
+-----------------+-----------------+-----------------+-----------------+-----------------+-
----------------+-----------------+-----------------+-----------------+-----------------+--
--------------+-----------------+-----------------+-----------------+-----------------+-----
------------+-----------------+-----------------+-----------------+-----------------+--------
----------+-----------------+-----------------+-----------------+-----------------+----------
-------+-----------------+-----------------+-----------------+-----------------+------------
-----+-----------------+-----------------+-----------------+-----------------+-------------
----+-----------------+-----------------+-----------------+-----------------+--------------
---+-----------------+-----------------+-----------------+-----------------+---------------
--+-----------------+-----------------+-----------------+-----------------+----------------
-+-----------------+-----------------+-----------------+-----------------+----------------
+-----------------+-----------------+-----------------+-----------------+-----------------+
-----------------+-----------------+-----------------+
|summary|               f1|               f2|               f3|
f4|               f5|               f6|               f7|               f8|
f9|              f10|              f11|              f12|              f13|
f14|              f15|              f16|              f17|
f18|              f19|              f20|              f21|
f22|              f23|              f24|              f25|
f26|              f27|              f28|              f29|              f30|
f31|              f32|              f33|              f34|
f35|              f36|              f37|              f38|
f39|              f40|              f41|              f42|
f43|              f44|              f45|              f46|
f47|              f48|              f49|              f50|
f51|              f52|              f53|              f54|
```

```
f55|               f56|               f57|               f58|
f59|               f60|               f61|               f62|
f63|               f64|               f65|               f66|
f67|             label|
+-------+------------------+------------------+------------------+-----------
-------+-----------------+------------------+------------------+------------
----+-----------------+-----------------+-----------------+----------------
-+-----------------+-----------------+-----------------+-----------------
+-----------------+-----------------+-----------------+-----------------+-
----------------+-----------------+-----------------+-----------------+--
--------------+-----------------+-----------------+-----------------+-----
------------+-----------------+-----------------+-----------------+--------
----------+-----------------+-----------------+-----------------+----------
-------+-----------------+-----------------+-----------------+------------
-----+-----------------+-----------------+-----------------+-------------
----+-----------------+-----------------+-----------------+--------------
---+-----------------+-----------------+-----------------+---------------
--+-----------------+-----------------+-----------------+----------------
-+-----------------+-----------------+-----------------+-----------------
+-----------------+-----------------+-----------------+-----------------+
-----------------+-----------------+----------------+
|  count|              2048|              2048|              2048|
2048|             2048|              2048|              2048|              2048|
2048|              2048|              2048|              2048|             2048|
2048|              2048|              2048|              2048|              2048|
2048|              2048|              2048|              2048|              2048|
2048|              2048|              2048|              2048|              2048|
2048|             2048|              2048|             2048|              2048|
2048|              2048|              2048|              2048|              2048|
2048|             2048|              2048|              2048|             2048|
2048|              2048|              2048|              2048|              2048|
2048|              2048|              2048|              2048|              2048|
2048|              2048|              2048|              2048|              2048|
2048|              2048|              2048|              2048|              2048|
2048|              2048|              2048|              2048|              2048|
2048|              2048|              2048|              2048|              2048|
2048|
|   mean|125.56923144531228|
125.3568066406251|125.42204101562524|125.71933251953112|126.0050551757814|126
.08446533203112|125.88512255859395|125.7838330078125|125.70476562499994|125.8
583608398437|125.3948452148437|125.67408691406261|126.01381738281238|126.0710
1806640624|125.90554833984386|125.64936181640618|125.72060742187496|124.63611
621093749|125.92546679687484|125.6720952148438|125.62689404296877|126.1208276
3671881|126.45232910156246|126.03068115234356|125.9389218750001|125.436985351
56238|125.6427724609375|125.35043408203126|125.1534243164061|125.588722656250
2|125.86260253906245|125.77710107421893|126.11436035156238|124.8027563476562|
```

125.48534863281249|125.10084667968752|124.5295268554687|124.95900634765609|
125.344212890625|125.35271582031214|125.32281494140634|124.36889697265605|124
.49470117187522|124.37970361328104|124.98884716796871|124.21263525390616|124.
01551074218737|124.23988769531276|
125.0545766601565|124.72619677734347|124.85175195312473|124.68179931640644|12
4.55499755859377|124.78828515624977|124.98818603515645|124.72449658203108|
125.0545766601565|125.26291943359394|125.20168164062505|124.98884716796871|12
4.63611621093749|124.84174414062501|125.06729882812482|125.20279980468747|125
.44442382812525|125.41438769531257|125.56461132812493| 0.50146484375|
| stddev| 33.29273147272022|32.822212055580586| 32.64300489797472|
32.96603101800241|33.52624745933872| 33.58923317139399|
33.2202239544659|33.21469708331746| 33.54851423435086| 33.4029013087615|
32.7883228029665|33.139636125149934| 33.59693971480369|33.649753378283606|
33.39515581468746| 33.11774874225592|32.821256226543305|
33.20314825083078|33.194239318685014| 32.8602935732294|
32.88993141586509|33.358333688026214|
33.53919402249628|33.257713727560365|32.94550106351531|
33.51988073635056|33.43733636850594| 32.97089963674672| 32.7585621250293|
33.2464262699429| 33.53166815202282| 33.48314466655061|
33.63270864479365|33.24901862040505| 33.59080137642515|
33.24705530584535|32.47655632423568|
32.90879047129794|33.36612540616642|33.495895201933365| 33.3043603199475|
33.88995787818959|33.748150048986936|33.668509692080626| 33.52747095890803|
33.73898484976952|33.501672188115826| 33.63435412597522|33.813020348244834|
33.95120664527394| 33.91829715448995| 33.4817118757419| 33.49244285960569|
33.76521076829497| 34.10533873503698|
33.74151861510293|33.813020348244834|34.030045223922514| 33.79549290754474|
33.52747095890803| 33.20314825083078| 33.23052968184479| 33.72821282418027|
33.68553883502131| 33.71655683090382|33.675679554405086|
33.66902745538126|0.500119968738303|
| min| 47.124| 47.262| 48.485|
49.323| 47.077| 47.365| 47.063|
47.546| 49.302| 48.393| 48.001|
49.138| 51.571| 51.154| 48.043|
47.507| 42.081| 47.444| 44.227|
48.473| 48.78| 50.101| 50.896|
49.03| 49.496| 45.491| 48.195|
47.26| 46.67| 44.992| 45.669|
46.245| 45.589| 46.734| 45.279|
42.499| 45.096| 44.383| 43.993|
43.201| 46.362| 43.346| 44.097|
42.936| 49.456| 45.339| 45.733|
43.998| 40.721| 40.565| 43.186|
41.958| 43.798| 44.526| 42.369|
42.543| 40.721| 45.416| 47.957|
49.456| 47.444| 45.266| 44.772|
46.018| 47.871| 50.691| 53.071|
0|

| 25%| 99.358| 99.091| 100.21|
99.763| 99.044| 98.962| 99.075|
99.718| 98.807| 99.43| 99.246|
98.868| 98.959| 98.664| 99.056|
99.534| 99.935| 99.401| 99.985|
100.6| 100.73| 100.38| 100.39|
100.01| 100.39| 99.839| 99.615|
100.1| 100.14| 99.892| 99.735|
100.09| 100.16| 99.317| 99.86|
100.38| 100.25| 100.08| 100.05|
100.03| 100.22| 100.47| 100.03|
100.37| 99.887| 100.17| 100.25|
100.27| 100.9| 100.5| 100.72|
100.67| 100.62| 100.48| 100.97|
100.63| 100.9| 100.78| 100.76|
99.887| 99.401| 99.729| 99.673|
99.95| 100.12| 100.28| 100.56|
0|
| 50%| 123.4| 124.16| 123.94|
124.46| 123.71| 124.24| 124.44|
124.31| 123.29| 124.52| 124.16|
124.24| 124.46| 124.6| 123.66|
123.81| 124.7| 122.71| 123.96|
124.21| 123.71| 124.21| 124.2|
123.76| 124.15| 123.85| 124.21|
124.47| 123.59| 124.06| 124.06|
123.76| 124.31| 121.62| 121.95|
121.78| 121.68| 121.46| 121.59|
121.12| 121.45| 118.7| 119.51|
119.52| 122.8| 119.11| 118.62|
118.73| 119.73| 118.8| 119.54|
120.6| 120.57| 120.21| 120.35|
119.56| 119.73| 120.18| 121.3|
122.8| 122.71| 121.93| 122.19|
122.64| 122.84| 123.14| 123.2|
1|
| 75%| 152.95| 151.25| 152.49|
152.33| 152.67| 153.15| 153.17|
152.07| 152.96| 152.17| 150.95|
152.38| 152.78| 153.16| 153.79|
152.34| 152.95| 150.96| 151.09|
150.32| 151.22| 153.21| 153.78|
152.7| 152.83| 152.42| 152.57|
150.65| 151.08| 152.41| 152.38|
153.07| 152.95| 150.19| 152.68|
151.72| 150.8| 152.54| 152.09|
153.06| 152.98| 151.58| 151.49|
151.05| 151.66| 150.77| 150.86|

```
151.72|            152.74|            152.48|            151.56|
151.53|            151.62|            152.37|            153.06|
152.59|            152.74|            152.22|            152.54|
151.66|            150.96|            152.46|            152.14|
152.83|            153.14|            153.17|            152.78|
1|
|    max|            210.65|             210.2|            212.93|
211.0|            213.1|            215.9|            218.09|
215.43|            223.88|            224.05|            217.95|
209.42|             215.7|            209.13|            211.73|
219.07|            218.37|             205.7|            220.35|
206.39|            214.41|            222.34|            215.53|
210.4|            216.2|            212.8|           209.05|
209.3|           215.97|          218.49|           204.75|
210.76|            222.85|            213.63|             210.66|
215.09|           208.19|            217.52|            220.61|
215.74|            222.29|            225.27|            215.77|
211.91|            221.07|            218.11|            208.78|
216.44|             213.5|            213.45|             226.9|
220.06|            214.02|            223.51|            214.08|
212.51|             213.5|            214.89|            219.11|
221.07|             205.7|            213.33|            216.48|
217.97|            212.54|            203.64|            209.64|
1|
+-------+------------------+------------------+------------------+-----------
-------+-----------------+------------------+------------------+-------------
----+-----------------+-----------------+-----------------+-----------------
-+-----------------+-----------------+-----------------+-----------------
+-----------------+-----------------+-----------------+-----------------+-
----------------+-----------------+-----------------+-----------------+--
--------------+-----------------+-----------------+-----------------+-----
------------+-----------------+-----------------+-----------------+--------
----------+-----------------+-----------------+-----------------+----------
-------+-----------------+-----------------+-----------------+------------
-----+-----------------+-----------------+-----------------+-------------
----+-----------------+-----------------+-----------------+--------------
---+-----------------+-----------------+-----------------+---------------
--+-----------------+-----------------+-----------------+----------------
-+-----------------+-----------------+-----------------+-----------------
+-----------------+-----------------+-----------------+-----------------+
-----------------+-----------------+-----------------+
```

Step 4:

In the previous assignment, most of you have no idea why we have to normalization and standardization the data (but these two comes under statistical information). Understand when we need to perform these things. Does the dataset require you to perform these two things?

Normalization and standardization are part of feature scaling. We use these techniques to compute the distance between the features that are biased towards numerically larger values if the data is not scaled.

There is no need for transformation in this data, in my opinion. The data in each field belongs to the same range or unit. In this way, comparing different fields is fairly simple. Furthermore, there will be no bias in statistical analysis because all fields are in the same unit/range and have equal statistical features. We can observe the deviation statistics of each column and mostly all the columns have similar values.

Step 5:

There are some duplicate columns in this dataset, think how we can find which are duplicates (repeated). What are you going to do with those columns?

```
def getDuplicateColumns(dataset, shape):
    z = []
    y = []
    for x in range(shape[1]):
        if(dataset.select(dataset.columns[x]).collect() not in y):
            y.append(dataset.select(dataset.columns[x]).collect())
        else:
            z.append('f'+str(x+1))
    return z
duplicateColumnNames = getDuplicateColumns(dataset, shape)
print('Duplicate Columns are as follows')
for col in duplicateColumnNames:
    print('Column name : ', col)
```

output:

```
Duplicate Columns are as follows
Column name :  f57
Column name :  f60
Column name :  f61
```

#Delete the duplicate to prevent overfitting.

#After founding duplicate columns, I am just dropping them and printing Updated dataframe dimensions.

```
dataset = dataset.drop(*duplicateColumnNames)
print('Shape: Rows:', dataset.count(),',','Columns:',len(dataset.columns))
```

output:

```
Shape: Rows: 2048 , Columns: 65
```

#Displaying dataframe with no duplicate columns.

```
dataset.show(10)
```

output:

```
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+-----+
|    f1|    f2|    f3|    f4|    f5|    f6|    f7|    f8|    f9|   f10|
f11|   f12|   f13|   f14|   f15|   f16|   f17|   f18|   f19|   f20|   f21|
f22|   f23|   f24|   f25|   f26|   f27|   f28|   f29|   f30|   f31|   f32|
f33|   f34|   f35|   f36|   f37|   f38|   f39|   f40|   f41|   f42|   f43|
f44|   f45|   f46|   f47|   f48|   f49|   f50|   f51|   f52|   f53|   f54|
f55|   f56|   f58|   f59|   f62|   f63|   f64|   f65|   f66|   f67|label|
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+-----+
|170.39|167.28|143.44|124.67|139.01|125.83|144.33|151.26|175.51|171.31|
161.9|146.92| 141.8|140.91|
132.8|128.48|170.83|161.06|169.61|168.07|154.88|149.33|152.91|152.61|137.09|1
69.59|183.01|180.44|164.55|152.63|157.77|165.07|149.52|142.73|173.85|182.23|1
74.58|163.28|164.01|
169.2|148.48|134.31|174.07|191.33|157.51|168.66|156.89|170.86|162.24|184.84|1
67.02|123.67|140.54|153.69|147.57|144.65|172.96|169.67|133.23|124.41|138.44|1
42.93|137.13|134.44|    0|
|169.75|190.96|175.53|138.27|137.47|139.23|133.23|130.25|147.73|163.93|167.36
|171.52|155.54|139.34|151.95| 149.3|173.37|
141.0|153.57|128.45|159.93|165.33|147.94|143.74|140.88|182.53|184.08|148.09|1
25.26|139.67| 138.7|132.86| 141.5| 145.4|164.58|170.71|127.83|133.99|
141.2|152.87|142.36|148.26|162.54|156.55|153.39|137.99|129.64|137.59|155.39|1
47.92|152.14|162.48|168.72|161.14|147.87|141.02|139.58|141.58|148.43|168.12|
```

169.9|165.64|166.86|137.69| 0|
|153.69|153.68|144.02|158.73|178.87|157.04|152.92|147.52|142.87|165.26|160.39
|137.86|149.62|153.43| 152.6|162.85|146.35|167.11|134.27|126.81|136.28|
158.4|171.58| 161.6|162.91|143.14|129.73|125.97|151.09|177.89|
169.2|160.65|156.86|135.66|126.92|131.22|154.37|158.59|158.52|155.25|154.33|1
30.18|127.66|148.94|155.37|163.53| 139.7|143.79|141.61|166.88|
164.6|149.58|139.56|154.74|173.01|155.18|155.19|170.51|146.89|141.01|159.43|1
69.68|163.24|165.17| 0|
|131.69|151.56|151.05| 134.0|151.18|175.53|171.34|159.77|151.95|
146.1|148.53|140.28|138.16|145.44| 150.4|158.18| 163.8|152.43|171.49|
150.2|131.28|157.18|157.04|151.13|151.66|143.32|157.23|152.91|134.75|154.65|1
71.65| 160.3|157.73|143.67|145.87|151.73|147.09|151.21|157.96| 148.5|
156.6|147.45|153.18|156.58|157.83|147.98|143.47|142.83|138.08|147.91|148.05|1
45.66|
156.5|167.52|151.33|129.51|164.25|155.82|150.82|146.58|128.85|140.76|177.35|1
74.61| 0|
|162.85|158.88|132.27|138.41|143.98|
159.3|177.26|180.58|159.34|164.66|138.04|132.76|157.88|165.58|173.64|
163.5|127.97|167.31|141.39|147.02|137.52|135.46|146.41|159.09|164.53|148.45|1
30.76|136.22|144.86|127.38|137.09|159.08|153.25|182.49|187.75|139.37|117.12|1
24.07| 134.6|144.85|132.64|170.87|188.26|173.32|135.74|127.15|131.69|
127.8|149.46|114.16|100.11|154.02| 175.3|175.46|144.39|142.47|
132.8|130.96|188.21|179.52| 146.2|153.73|152.12|146.58| 0|
|132.05|149.12|165.08|170.62|162.19| 157.1|145.86|149.52|162.84|
149.5|138.86|140.41|156.82|171.41|158.94|153.78|176.64|137.61|
169.6|149.55|138.69|142.72|172.72|172.08|156.74|154.95|155.39|152.42|146.35|
141.2|150.94|154.63|150.24|140.62|144.13|159.41|148.12|149.68|139.02|121.29|1
32.76|161.71| 157.6|165.57|142.59|151.45|153.89|152.43|153.07|139.58|131.01|
155.6|157.89|162.75|
153.9|145.37|147.36|140.65|157.63|152.16|140.43|142.32|142.06|154.87| 0|
|153.59|142.25|157.33|156.08|149.33|162.97|150.25|146.47|145.99|137.82|
152.9|161.64|150.23| 170.7|185.03|174.97| 126.8|154.72|140.58|158.21|163.89|
159.1|178.79|181.28|188.27|105.94|127.84|152.78|161.62|163.97|179.15|180.57|1
79.18|127.12|121.63|142.36|149.33|158.32|181.72|184.42|175.92|141.72|121.99|1
30.39| 174.9|124.85|137.45|164.72|144.44|165.39| 157.7|162.63|154.26|159.03|
143.8|125.68|159.71|149.75|162.83|162.37|162.75| 168.6|170.81| 168.1| 0|
|167.68|153.49|149.19|148.71|166.03|167.04|153.06|157.48|133.57|143.66|167.27
|172.45| 179.8|169.15|
150.5|152.77|127.77|147.58|127.63|143.08|167.35|176.35|165.35|168.15|
168.1|129.51|129.52|142.58|161.59| 167.7|162.09|167.67|170.75|125.04| 126.4|
161.1|164.27|155.41|160.66|172.89|177.24|157.48|143.26|166.09|145.76|167.91|1
57.71|161.01|155.85|169.13|159.15|160.73|156.99|152.19|168.29|171.96|150.72|1
54.01|145.01|164.87| 157.2|147.07|162.98|167.99| 0|
|136.48|130.02|131.72|152.04|163.03|172.93|170.11|
165.2|166.41|120.67|119.02|135.76|147.52|164.59|176.47|168.61|
166.5|121.19|133.17|111.33|114.23|127.57|155.83|178.35|186.52|168.26|149.86|1
31.02|140.33|135.15|145.21|164.04|
173.8|153.96|122.81|125.75|156.86|164.41|151.03|154.61|163.85|151.98|120.14|1

```
31.17|142.04|161.46|158.05|143.62|142.59|146.69|167.86|144.87|119.45|140.61|1
73.71|158.99|145.14| 162.0|124.21|145.11|142.15|148.38|142.86|154.15|    0|
|145.96|140.31|126.34|113.12|118.66|140.33| 139.9|139.51|
168.7|149.54|149.38|147.88|129.45|143.81|131.18|120.76|172.49|125.42|152.75|1
48.41|153.47|139.24|138.86|143.44|
117.8|156.93|145.75|133.26|146.65|162.72|156.35|131.51|136.35|151.53|143.73|1
27.52|140.33| 168.8|163.12|136.13|150.21|148.94|
173.1|158.82|121.89|137.21|154.26|160.58|158.22|142.27|141.34|148.79|165.43|1
59.23|133.41|127.64|171.77| 152.2|146.03|142.27|127.53|141.77|160.67|159.55|
0|
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+-----+
only showing top 10 rows
```

Step 6:

Count label values. Discuss whether the data is Imbalance, Inaccurate, and Incomplete data. Provide your discussion.

#Groupingby Label column and applying count function and displaying the result.

```
dataset.groupBy('label').count().show()
```

output:

```
+-----+-----+
|label|count|
+-----+-----+
|    1| 1027|
|    0| 1021|
+-----+-----+
```

Observations:

- It is big data.
- It is imbalanced data, as features have distributed or varied label count(1027 - 1021).
- It is not incomplete data as there are no null values.
- If observations are Incorrect it is inaccurate, the dataset is accurate, also the standard deviations of each column are comparatively

similar. Can be more understood if we balance the data set and perform further classification.

Step 7:

Dataset is not randomly shuffled so, randomly shuffle the dataset and divide the dataset into [70:30 or 75:25] test and train datasets.

#Using rand function we are shuffling the data set and using seed value to get the same shuffle for every run to understand more about the further split.

```
from pyspark.sql.functions import rand
dataset = dataset.orderBy(rand(), seed = 42)
```

#Using randomSplit to split data to 70-30 percentage of dataset to train and test.

```
train_data_spark,test_data_spark = dataset.randomSplit([0.7, 0.3], seed = 42)
```

#Displaying the train split data and displaying the dimensions.

```
train_data_spark.show()
print(train_data_spark.count(),len(train_data_spark.columns))
```

output:

```
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+-----+
|    f1|    f2|    f3|    f4|    f5|    f6|    f7|    f8|    f9|   f10|
f11|   f12|   f13|   f14|   f15|   f16|   f17|   f18|   f19|   f20|   f21|
f22|   f23|   f24|   f25|   f26|   f27|   f28|   f29|   f30|   f31|   f32|
f33|   f34|   f35|   f36|   f37|   f38|   f39|   f40|   f41|   f42|   f43|
f44|   f45|   f46|   f47|   f48|   f49|   f50|   f51|   f52|   f53|   f54|
f55|   f56|   f58|   f59|   f62|   f63|   f64|   f65|   f66|   f67|label|
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+-----+
|47.124|47.262|48.485|49.323|50.302|51.563|53.062|54.529|67.829|71.957|74.043
|70.328|70.218|71.039|69.833|
```

71.61|81.658|74.135|82.693|80.986|76.554|71.027|71.266|73.515|76.465|75.408|7
9.352|76.214|73.669|
73.44|72.776|73.168|79.184|76.704|76.883|73.998|73.675|73.495|71.236|72.027|7
4.663|78.967|77.934|76.194|72.563|74.817|74.085|74.799|
76.63|74.041|73.659|74.167|75.785|76.718|74.689|76.157|76.828|76.619|73.579|7
3.942|74.965|74.399|73.866|74.165| 1|
|52.374|55.044|53.418|50.651|47.077|47.365|47.063|47.546|58.276|60.636|59.945
|57.728|58.145|60.721|62.249|65.496|68.944|73.388|71.204|
68.95|68.767|66.567|69.226|74.717|79.568| 70.18|71.539| 72.74|69.462|
70.36|71.134|
73.77|72.978|75.785|74.745|71.809|71.627|73.168|74.777|76.531|75.522|75.683|7
6.084|77.769|75.101|77.744| 77.29|76.653|77.984|79.512|79.076|77.793|76.683|
78.78|77.598|76.107|76.795|74.937|73.948|73.906| 74.25|74.473|72.279|71.657|
1|
|
52.89|52.644|52.521|53.245|54.076|54.703|55.828|55.597|51.534|48.902|48.745|4
9.138|51.571|53.886|54.711|56.125|50.511|70.562|50.598|52.741|51.283|50.101|5
0.896|53.606|56.056|45.491|48.195|49.278|48.171|
50.61|51.991|53.303|53.177|47.332|49.503|49.663|50.509|51.015|51.998|
56.22|57.272|54.777|60.959|60.127|72.219|64.823|64.808|64.615|62.563|67.716|6
9.906|65.599|67.155|67.964|67.862|63.668|65.332|67.716|69.378|72.013|70.491|6
9.301|68.523|72.039| 1|
|53.797|53.019|54.683|54.146|51.439|50.729|52.956|54.832|71.995|69.352|67.569
|65.326|62.189|58.251|57.388|60.952|78.162|72.614|76.648|74.317|75.562|74.461
|71.993|71.583| 69.92|78.205|75.946|74.091|72.021| 68.75|
66.85|68.649|67.546|75.417|75.074| 74.49|73.717|73.782|73.721|73.855|72.524|
75.79|79.416|81.611|73.636|78.888|78.533|80.201|79.694|80.373|78.897|
78.26|78.578|81.455|78.866|
79.65|77.364|75.317|69.978|68.231|67.721|71.721|70.272|70.367| 1|
|55.479|55.054|56.543|58.407|59.686|56.633|
51.94|50.579|49.302|48.393|48.001|49.994|
54.22|51.154|48.043|49.514|62.052|72.836|60.025|63.346|65.227|63.524|62.108|6
3.822|61.644|68.462|69.487|67.785| 65.39|66.193|65.816|65.937|63.125|
70.14|71.925|67.169|64.757|67.372|67.288|66.811|69.723|74.123|72.745|69.667|6
9.983|67.161|66.735|65.598|71.735|67.821|71.443|72.819|72.881|70.316|70.473|7
0.452| 71.3| 73.25|72.747|75.128|75.233|75.559|73.027|72.437| 1|
|56.455|57.775|55.534|53.364|52.146|52.876|50.698|50.072|53.657|
53.93|49.149|
51.33|53.486|55.393|53.656|52.416|52.428|68.855|52.881|51.503|50.372|50.593|
52.55|53.025|51.822|50.618| 52.04|48.749|
46.67|44.992|45.669|46.245|45.589|46.734|45.279|42.499|45.096|44.383|43.993|4
3.201|46.394|43.346|
44.17|42.936|73.093|45.339|47.749|51.695|61.981|55.228|52.804|57.878|58.038|5
5.623| 53.75|57.431|65.664|64.881|66.734|61.005|63.976|67.768|67.972| 70.83|
1|
|59.747|61.359|59.965|59.313|60.739|59.535|60.489|61.938|61.534|62.838|60.584
|64.204| 67.23|67.339|63.315|62.206|73.788|71.446|
72.32|68.325|72.335|69.802|70.412|70.213|66.332|68.265|70.315|70.477|71.158|7

1.949|70.249|72.949|74.829|75.268|76.677|73.267|72.673|73.496|
70.03|72.028|74.956|80.235|78.572|78.593|70.239|73.911| 73.6|73.211|
74.23|74.008|75.292|75.836|74.511|74.101|71.418|74.379|
76.91|76.099|72.281|71.696|75.498|73.935|73.494|72.941| 1|
|59.862|61.461|60.634|59.889|56.573|55.893|56.445|55.057|55.439|56.933|55.559
|
52.75|52.632|52.453|52.936|52.056|49.556|71.493|50.245|51.144|53.932|53.931|5
1.853| 49.33|50.498|48.196|51.489|50.589|50.445|52.208|49.684| 49.5|48.129|
52.5|51.808|52.849|48.506|46.063|46.994|45.387|46.362|49.329|48.884|48.854|
73.66|49.071|47.362|43.998|49.322|40.565|43.186|53.211|
51.71|49.922|51.953|52.492|49.822|53.308|68.489|70.373|67.582|67.105|65.857|6
9.387| 1|
|60.788|59.564|58.077|56.502|59.711|62.184|62.183|65.753|58.947|58.873|59.663
|58.678|60.359| 64.4|66.161|70.239|66.798|
70.88|68.046|68.291|68.158|72.751|72.971|75.177|75.173|76.507|
74.78|73.365|75.577|
79.29|79.733|81.073|79.816|76.769|78.322|76.796|76.431|79.438|81.299|80.978|7
8.533|75.015|76.271|79.398|69.482|79.222|76.702|76.368|72.864|79.329|78.058|7
3.169|72.921|74.018|73.817|73.305|
72.98|75.334|72.707|69.984|70.249|68.829|68.717|71.238| 1|
|61.779|63.149|63.247|62.871|63.979|64.643|62.349|62.549|
72.74|75.355|75.626|78.042|80.342|83.083|79.796|77.686|74.474|59.122| 80.64|
80.94|82.382|86.036|84.598|81.348|81.682|73.882|75.697|77.054|79.089|77.609|7
7.452|76.267|75.147|69.198|68.713|68.105|
70.19|69.066|68.829|68.575|67.709|67.872|69.026|67.785|59.341|66.828|68.569|
67.87|67.318|
68.11|69.605|65.468|65.546|65.712|64.413|66.635|67.981|69.316|58.136|58.914|5
7.512|57.252|57.014|59.631| 1|
|62.383|62.336|63.626|60.684|61.275|61.691|61.054|62.595|70.187|72.428|73.013
|74.583|76.713|74.458|70.946|68.431|75.253|77.709|71.858|69.643|68.798|65.827
|61.651|58.788|57.576|64.559|63.674|66.721|62.125|58.624|59.528|58.638|58.145
|75.977|75.068|76.507|73.855|71.161|71.608|69.285|66.329|84.593|81.657|
80.27|80.154|76.669|76.739| 75.1|73.809|68.935|65.315|
81.02|80.484|81.901|78.896|77.535|70.955|
69.31|76.666|75.944|71.989|70.112|69.574|67.707| 1|
|
62.53|62.055|59.721|60.916|62.446|61.202|61.059|64.516|72.571|70.054|65.539|6
2.404|64.832|61.474|60.299|62.648|84.905|75.437|81.286|77.121|75.392|73.704|6
8.674|65.498|69.128| 86.87|85.714|
79.25|74.527|71.985|72.165|70.504|71.522|85.728|87.842|84.114|80.137|
75.21|75.786|
74.2|77.678|84.009|82.181|81.291|72.255|78.275|75.068|69.968|77.539|70.831|75
.738|84.362|83.977|84.405|80.693|80.247|80.524|82.632|77.589|75.704|76.371|77
.214|75.215|76.288| 1|
|63.378|63.595|63.858|61.815|61.629|62.566|
62.57|64.905|57.152|58.059|57.815|56.438|53.845|53.752|
57.13|57.936|54.226|61.207|55.663|55.857|54.798|56.071|56.472|55.794|57.968|5
2.554|53.585|55.033|54.103|53.413|

54.74|55.233|58.207|54.962|53.697|57.577|60.624|57.058|56.336|57.644|59.656|5
3.182|51.636|51.886|63.433|52.419|51.597|49.435|47.996|52.016|51.513|52.919|5
2.339| 51.62|50.307|47.823|48.068|47.957| 59.92|
59.63|61.318|58.562|59.873|61.742| 1|
|63.631|65.403|67.657|69.173|69.811|70.987|67.302|67.446|56.671|58.096|
59.23|61.101|64.021|62.691|64.242|66.175|56.331|67.966|56.096|
59.59|64.082|66.871|61.529|61.811|61.711|58.616|61.055|62.449|
66.84|67.445|67.796|64.785|66.012|64.782|64.693|66.628|66.612|
66.05|68.258|69.023|67.173|69.186|69.894|70.311|
63.77|67.102|65.911|63.435|60.472|67.136|67.831|66.401|68.891|70.145|67.568|6
4.805|61.914|64.459|70.642|71.853| 63.13| 59.15|58.226| 62.2| 1|
| 64.21|66.031|69.937|69.903|
68.34|65.065|62.806|66.488|63.244|68.647|69.719|
72.87|73.526|69.977|65.763|64.113|58.107|79.407|60.771|
63.49|67.412|65.103|68.595|66.523|65.934|64.302|64.145|
65.42|67.309|67.038|63.135|64.805|69.916|63.514| 63.03| 65.28|67.917|69.678|
70.8|69.968|72.455|75.301|74.614|74.568|78.774|
73.79|74.357|75.482|81.759|70.379|68.454|80.982|81.329|82.245|80.002|79.683|7
8.621|76.048|81.299|80.526|79.027|79.778|78.062|78.816| 1|
|64.284|67.737|73.672|74.796|74.153|73.429|74.551|75.617|63.722|
67.51|72.562|72.341|72.499|70.978|71.028|72.073|63.145|66.585|66.201|65.164|6
7.991|71.027| 71.56|70.935|70.631|
63.72|64.584|66.201|68.033|69.392|72.176|72.009|74.061|60.858|
59.81|61.317|58.578|62.714|65.748|70.024|70.844|50.229|48.223|49.993|65.024|5
0.625|54.249|56.432|55.173|56.092|58.088|52.646|52.527|53.036|53.669|56.177|5
7.154|62.948|65.607|58.167|60.707|64.877| 70.63|75.013| 1|
|64.461|64.516|63.507|63.035|66.788|65.663| 61.26|62.442|
72.74|69.885|67.608|65.899|68.437|68.562|
64.64|62.953|72.213|56.084|68.289|66.959|66.239|68.649|69.044|64.933|63.734|
77.32|73.576| 69.74|70.043| 69.59|72.193|69.008| 65.84|77.745|
75.13|73.465|70.191|69.144|71.081|69.397| 67.53|71.959|
70.24|66.585|59.592|65.528|66.257|66.812|68.554|67.835|66.354|61.413|62.684|6
2.943|64.353|67.034| 66.12|63.491|53.181|54.276|56.854|57.477|57.637|57.205|
1|
|65.917|68.042|68.329|69.163|69.766|70.623|72.034|74.749|58.563|62.577|66.207
|68.527|72.961|77.465|78.078|84.079|63.995|74.599|63.034|59.125|59.071|61.701
| 71.07|72.256|73.198|66.377|67.654|64.369|62.817|62.612|66.018| 66.52|
65.19|75.184|75.759|76.254|74.408|72.333|69.055|68.421|69.932|76.768|72.909|7
0.084|77.206|69.199|71.225|69.772|73.103|65.622|
64.77|77.343|74.821|73.419|70.947|
70.61|73.118|73.088|71.231|70.692|74.374|75.019|74.384| 75.95| 1|
|66.272|58.525|
61.49|70.789|81.541|83.976|86.679|86.724|62.348|56.468|56.934|61.473|71.355|7
7.134|81.296|85.072|56.052|68.268|53.482|53.872|57.579|61.129|71.011|77.793|7
9.158|65.317|61.243|58.586|60.744|66.579|73.028|76.717|76.802|64.888| 67.19|
62.95|66.235|72.024|73.126|74.066|
75.8|69.123|68.349|66.776|68.949|71.373|74.384|72.858|
72.87|72.215|72.049|71.902|73.907|69.231|68.292|73.052|73.297|72.391|68.396|

68.96|69.283|69.752|68.583|70.535| 1|
|66.658|65.623|66.483|66.943|65.704|67.222|65.657|68.509|73.334|69.897|67.506
|65.518| 62.94|60.329|60.078|60.171|73.419|84.232|70.926|68.138|
63.21|64.699|62.725|61.123|59.952|65.493|62.966|66.692|66.358|66.412|65.026|6
2.364|59.774|66.517|68.188|69.574|69.624|71.252|67.715|64.966|65.172|78.114|7
8.786|77.213| 84.26|78.129|74.764|72.548|74.549| 71.62|70.809|
83.13|83.953|81.449|78.958|
74.41|73.265|69.849|81.275|76.643|74.447|75.297|73.892|73.012| 1|
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+-----+
only showing top 20 rows

1484 65

#Displaying the test split data and displaying the dimensions.

```
test_data_spark.show()
print(test_data_spark.count(),len(test_data_spark.columns))
```

output:

+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+-----+
| f1| f2| f3| f4| f5| f6| f7| f8| f9| f10|
f11| f12| f13| f14| f15| f16| f17| f18| f19| f20| f21|
f22| f23| f24| f25| f26| f27| f28| f29| f30| f31| f32|
f33| f34| f35| f36| f37| f38| f39| f40| f41| f42| f43|
f44| f45| f46| f47| f48| f49| f50| f51| f52| f53| f54|
f55| f56| f58| f59| f62| f63| f64| f65| f66| f67|label|
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+-----+
|52.511|53.543| 53.75|51.629|49.214| 49.75|50.018|
48.29|51.031|53.001|53.099|52.927|52.204|52.687|55.477|54.809|
61.85|75.271|62.607|57.658|56.907|61.229|64.868|65.663|
66.23|64.351|65.696|65.251|67.543|68.325|68.938|
69.57|69.759|68.322|71.198|69.489|69.527|70.854|70.417|72.564|75.872|69.417|6
8.513|68.886|74.253|69.927|71.283|72.277|76.014|74.339|75.767|72.892|74.622|7

2.905|74.434|74.698|76.585| 76.87|75.381|74.404|73.044|72.042|75.249|75.434|
1|
|55.662|55.748|59.649|57.632|62.191|63.659|64.781|66.804|58.161|58.415|57.944
|59.165|57.503|56.567|58.475|61.384|55.997|73.513|58.067|61.252|58.802|56.601
|53.879|53.389|59.164|54.803|55.047|54.546|53.686|52.381|50.808|50.883|53.855
|58.437|57.675|56.636|55.639|54.894|54.765|56.132|
60.39|70.006|67.962|69.123|75.117|68.692|67.512|66.561|66.957|68.737|71.076|
69.56|67.299|66.663|66.203| 65.88|66.127|
63.87|75.387|75.316|75.203|76.302|73.891|70.217| 1|
|57.133| 61.74|62.219|62.903|64.589|67.159|
67.64|69.995|61.048|61.628|61.056|62.656|65.393|67.092|
68.86|72.265|63.149|69.971|63.934|63.406|
61.74|64.049|69.103|68.111|72.206|64.652|65.289| 63.67|59.565|62.055|66.984|
68.67|
69.92|68.061|65.716|64.032|62.344|64.251|65.946|69.708|73.987|68.004|69.896|6
8.723|64.274|67.854| 67.84|72.024|78.066|74.372|
74.98|65.473|70.072|73.365|73.996|77.832|79.842|78.951|73.654|
76.33|80.693|78.107| 78.23|76.135| 1|
|59.406| 61.01|63.612|65.067|64.525|67.823|
66.35|64.413|52.944|54.123|56.146|56.202|56.639|57.606|58.949|57.733|48.361|7
1.196|51.671|52.333|53.464|56.149|55.206|57.278|54.268| 47.75|48.261|
47.26|48.259|52.255|52.886|
53.44|51.424|49.004|48.788|51.666|54.056|55.845|57.214|55.155|54.111|51.781|4
8.751|51.978|70.458|52.885|53.037|55.777|54.802|55.221|54.292|61.945|60.935|5
9.322|59.689|56.407|56.035|53.041|74.025|73.191|69.461| 64.86| 62.93|64.217|
1|
|61.339|59.886|60.785|60.526|58.066|57.546|58.672|56.818|49.457|51.899|
50.35|49.748|52.203|52.229|51.047|50.432|49.357|67.959|51.669|55.082|55.646|5
7.677|57.426| 58.64|59.908|63.099|64.675|65.057| 69.25|69.038|71.635|
71.3|67.457|68.151|68.899|67.587|70.934| 72.16|
75.0|75.967|72.333|68.388|70.057|67.482|68.714|68.175|69.798|69.195|68.169|73
.088|75.486|70.275|69.115|
66.3|67.251|68.346|71.701|74.448|68.328|67.267|65.506|63.487|65.135|67.949|
1|
|61.625|57.217|55.576|54.024|54.184|57.244|
59.19|59.866|68.162|62.235|61.734|64.561|62.673|64.907|62.251|59.928|69.008|8
3.764|69.222|70.265|68.947|65.062|
64.95|66.355|67.779|70.182|71.925|75.238|70.579|68.378|70.624|74.542|71.823|7
8.288|73.502|71.206|69.168|68.604|70.782|71.984|74.543|
77.77|75.679|74.789|76.361|71.133|75.559|76.955|84.071|79.424|78.765|
78.41|77.927|81.347|79.519|81.393|86.229|82.182|84.851|85.252|84.715|84.734|8
4.279|81.273| 1|
|61.734|64.408|
67.62|68.494|68.576|63.387|60.117|58.865|61.926|62.619|63.263|69.367|67.805|6
3.245|63.064|57.143|64.985|60.196|66.968|66.705|71.755|69.559|65.568|63.342|5
7.744|68.774|70.677|70.853|69.954|68.265|65.347|62.321|58.097|70.072|69.126|6
8.102|69.506|67.975|65.386|65.898|66.749|64.832|63.876|63.204|57.948|64.313|6
7.149|68.218|

64.12|67.782|67.399|61.561|65.854|66.147|66.546|65.883|64.148|65.317|
57.78|59.105|61.746|60.258|59.728|62.549| 1|
| 62.8|68.942|70.733| 72.27|74.104|70.765|70.433|73.389|
83.64|83.944|88.289|87.761|85.594|84.764|83.381|
84.21|88.539|91.711|88.984|89.777|89.711|90.207|88.429|89.882|90.201|
86.99|89.028|87.721|87.014|88.114|85.218|88.576|90.162|78.969|81.362|84.892|8
5.082|81.996|82.254|81.217|85.179|75.091|82.033|85.999|90.223|84.767|81.454|8
0.516|82.219|80.471|82.209|81.035|85.185|87.327|88.542|85.139|80.844|85.389|9
3.813|92.941|92.318|91.933|86.628|86.684| 1|
|63.384|65.347|67.293|65.818|68.409|71.619|73.612|70.232|62.096|64.167|64.131
|65.126|71.325|80.822|79.995|77.892|70.436|82.982|69.777|
71.28|71.076|71.766|75.921|77.333|79.169|73.688|71.286|72.013|73.677|74.157|7
6.264|79.803|81.891|82.735| 80.01|76.215|74.958|72.729|73.849|77.971| 84.65|
80.91|77.776|78.133|76.992|78.504|77.196|77.043| 83.48|78.962|82.056|80.023|
81.33|80.101|82.565|82.684|80.189|81.945|84.102|83.754|87.363|86.611|87.597|8
5.507| 1|
|64.004|61.371|60.146|59.177|60.706|60.123|57.281|58.069|57.733|57.166|57.822
|58.998|
56.99|54.461|53.486|54.778|57.327|63.081|55.753|55.711|56.272|55.188|51.364|
49.03|49.971|57.904|57.133|
55.74|54.123|54.801|49.798|48.762|48.337|56.042|52.771|50.755|51.671|51.949|5
1.734|51.406|53.961|50.275|50.685|50.488|63.982|48.373|48.407|
47.38|40.721|45.962|48.894|50.706|52.479|48.546|42.369|42.543|45.416|50.894|5
8.237|51.939|48.726|51.113|59.896|69.192| 1|
|64.079|64.647|66.216|67.547|68.126|71.193|70.889|72.895|79.797|80.737|81.601
|79.491|80.121|81.976|83.985|85.811|83.994|
61.1|83.706|82.719|82.065|80.751|81.635|83.408|82.876|75.529|74.226|72.151|72
.114|73.594|76.086| 76.51|76.756|70.059|71.302|70.478|
69.67|69.598|74.431|74.397|75.382|72.087|72.255|69.078|62.857|68.203|66.635|6
8.011|59.139|67.535|69.315|68.013| 67.84|67.167|62.397|
58.53|57.551|56.969|60.518|57.667|55.805|57.084|58.711|59.659| 1|
|65.167|64.757|65.674|67.307|67.538|67.053|65.567|66.037|63.803|61.757|61.783
|64.155|63.872|63.204|62.067|60.701|54.086|
70.09|54.727|55.711|56.253|55.941|54.957|49.989|49.684|55.659|
55.67|59.671|58.714|59.671|60.998|57.243|56.947|65.998|65.711|66.909|64.344|6
6.057|65.651|64.784|64.126|65.303|66.915|65.894|68.197|65.027|63.913|66.385|7
0.938|70.987|72.104|65.018| 66.73|68.634|69.049|70.263|72.973|73.604|
72.03|72.787|72.937| 71.79|70.405| 70.09| 1|
|65.633|63.843|65.408|69.745|69.677|69.875|70.502|71.232|74.077|73.682|76.209
|76.115|77.981|75.999|76.089|79.135|89.156| 89.47|88.558|87.601|88.414|
87.21|85.945|84.439|83.119|87.755|89.783|89.955|90.279|90.071|87.023|87.946|8
5.313|86.825|86.973|90.798|90.548|87.203|85.962|86.935|
88.8|87.296|90.186|90.252|91.621|91.376|89.711|90.623|92.517|92.165|92.129|91
.461|92.571|92.721|92.549|89.763|92.657|91.187|93.039|92.911|91.024|90.382|90
.695|88.833| 1|
|65.702|65.413|66.569|65.598|67.852|71.063|72.427|74.823|72.418|72.376|75.317
|75.858|76.603|79.292|82.665|80.105|79.387|70.757|75.418|76.884|
78.24|80.164|81.032|80.033|77.909|83.779|80.933|80.266|79.458|81.804|81.338|8

0.768|
78.97|84.087|82.095|81.737|79.633|78.894|81.903|82.843|79.933|78.134|78.567|7
7.273|69.958|77.029|73.568|75.323|75.358|78.001|76.637|78.885|78.514|77.515|7
3.736|74.793|78.011|72.605|71.006|70.483|67.373|65.674|67.011|64.704| 1|
|66.066|67.368|68.996|67.981|68.833|68.162|68.083|70.634|67.008|68.256|70.419
| 70.9| 69.84|69.053| 70.83|
75.0|77.569|67.658|73.991|74.488|73.987|74.168|76.323|75.943|77.795|76.843|77
.613|76.926|76.352|77.494|
79.37|79.948|78.915|70.186|71.746|69.712|71.907|74.017|73.749|76.151|77.112|
64.77|62.791|62.877|68.217|62.534|61.747|63.695|
59.4|65.441|72.756|63.326|61.953|63.863|61.906|60.969|60.969| 63.25|66.274|
62.88|63.103|65.213|65.846|68.503| 1|
|66.332|67.544| 66.97|66.256|69.167|70.901|71.507|72.774|65.539|
65.29|66.936|67.123|65.431|69.128|65.794|68.684|73.743|79.139|74.568|73.559|7
1.583|67.124|69.934|66.887|66.722|75.969|76.185|75.642|74.251|72.706|74.903|7
2.765|74.245|
72.47|73.322|71.066|69.598|72.847|71.395|74.824|76.273|68.429|66.815|61.513|7
8.714|64.072|64.881|63.815|67.671|
66.41|65.823|70.721|67.837|66.719|66.627|66.931|71.666|73.283|80.394|80.475|7
9.361|80.487|79.457| 81.25| 1|
|66.362| 65.71|65.922|
65.88|67.347|68.625|68.396|68.196|67.612|65.567|67.179|66.945|66.089|68.467|6
8.008|65.114|66.437|73.126|64.024|67.265|67.749|68.029|67.207|67.755|64.241|6
2.248|61.405|60.142|61.103|62.101|62.975|64.871|62.191|62.705|63.932|62.068|5
9.869|58.374|57.176|57.767|58.628|
58.39|59.431|58.613|71.661|58.118|54.051|52.268|57.211|53.902|53.013|
61.46|62.185|58.503|58.001|54.885|58.222|57.286|70.096|67.247|67.147|
68.19|69.413|68.928| 1|
|67.043|64.236|64.901|66.698|66.004|65.497|64.173|66.898|64.414|62.164|
63.64|65.794|
67.31|66.437|65.341|66.626|61.831|67.304|61.762|62.646|64.092|64.891|64.289|6
3.813|62.153|60.932|61.378|56.887|61.044|60.898|61.744|60.349|58.853|
56.23|58.317|57.385|59.227|
56.48|54.039|53.381|52.242|51.004|54.138|53.644|69.031|55.148|
52.33|50.863|52.117|49.957|47.009|56.931|57.513|58.503|56.755|53.119|50.234|4
9.128| 65.73|60.126|57.557| 63.02|61.043|59.849| 1|
|67.619|71.722|68.699|
63.55|68.327|71.728|70.477|66.271|67.487|69.413|65.935|62.822|65.039|64.389|6
2.926|59.099|66.451|62.076|
66.67|62.559|65.498|68.929|68.732|62.933|58.314|62.813|64.365|64.349|64.067|6
8.941|68.549|64.399|66.035|63.525|
61.14|62.654|63.277|66.508|64.966|65.422|65.161|62.721|
61.42|64.224|58.693|64.506|66.204|65.946|69.847|65.378|63.663|62.876|61.043|6
2.656|64.513| 67.44|69.194|67.326|65.465|65.954|67.077|70.768|70.707|70.757|
1|
|67.621| 63.42|61.733|65.349|65.604|67.332|68.975|67.469|49.963|49.888|
51.72|53.802|57.909|56.762|57.195|54.658|42.081|52.888|44.227|48.473|49.147|5
0.945|53.396|51.451|54.071|50.606|50.445|48.604|49.952|52.624|52.358|53.741|5

```
6.487|54.186|53.891|50.315|48.892|
51.07|50.005|52.939|54.958|48.941|50.283|49.325|54.559|48.172|45.733|46.233|5
0.827|
49.03|51.718|41.958|43.798|46.778|48.064|46.665|52.378|52.907|57.107|58.256|5
8.715| 60.68| 59.21|60.371|    1|
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+------+------
+------+------+------+------+------+------+------+------+------+-----+
only showing top 20 rows

564 65
```

Step 8:

Write a document about your results and output. Discuss Step 3, Step 4 and Step 6 results. Document should be clear with sub-headings and headings.

The dataset is complete as there are no Null values but imbalanced as the ‘label’ column is not balanced. We have to make it balanced. Additionally, the statistical observations such as mean, median, min, max, quantile values prove that no need to do normalization or standardization and Delete the duplicate to prevent overfitting. As per overall observations and hypothesis made by statistical results we can say data is Accurate.

References:

https://sparkbyexamples.com/

https://www.geeksforgeeks.org/

https://stackoverflow.com/

Pandas Documentation

Pyspark Documentation