Spark Assignment 2

Thejas Bharadwaj

1) Exploratory Data Analysis- Follow up on the previous spark assignment 1 and explain a few statistics. (20 pts)

```
trainDF.show()
trainDF.printSchema()

// Mean median and other
trainDF.select().summary().show()

// Count of missing values
trainDF.select(trainDF.columns.map(c => sum(col(c).isNull.cast("int")).alias(c)): _*).show()

// Categorical variables
trainDF.groupBy("Sex").count().show()
trainDF.groupBy("Embarked").count().show()

// Correlation
trainDF.stat.corr("Age", "Fare")

// Survival rate by gender
trainDF.groupBy("Sex").agg(avg("Survived")).show()

// Age distribution by class
trainDF.groupBy("Pclass").agg(avg("Age")).show()

// Survival rate by port of embarkation
trainDF.groupBy("Embarked").agg(avg("Survived")).show()
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|          1|       0|     3|Braund, Mr. Owen ...|  male|22.0|    1|    0|       A/5 21171|   7.25| NULL|       S|
|          2|       1|     1|Cumings, Mrs. Joh...|female|38.0|    1|    0|        PC 17599|71.2833|  C85|       C|
|          3|       1|     3|Heikkinen, Miss. ...|female|26.0|    0|    0|STON/O2. 3101282|  7.925| NULL|       S|
|          4|       1|     1|Futrelle, Mrs. Ja...|female|35.0|    1|    0|          113803|   53.1| C123|       S|
|          5|       0|     3|Allen, Mr. Willia...|  male|35.0|    0|    0|          373450|   8.05| NULL|       S|
|          6|       0|     3|    Moran, Mr. James|  male|NULL|    0|    0|          330877| 8.4583| NULL|       Q|
|          7|       0|     1|McCarthy, Mr. Tim...|  male|54.0|    0|    0|           17463|51.8625|  E46|       S|
|          8|       0|     3|Palsson, Master. ...|  male| 2.0|    3|    1|          349909| 21.075| NULL|       S|
|          9|       1|     3|Johnson, Mrs. Osc...|female|27.0|    0|    2|          347742|11.1333| NULL|       S|
|         10|       1|     2|Nasser, Mrs. Nich...|female|14.0|    1|    0|          237736|30.0708| NULL|       C|
|         11|       1|     3|Sandstrom, Miss. ...|female| 4.0|    1|    1|         PP 9549|   16.7|   G6|       S|
|         12|       1|     1|Bonnell, Miss. El...|female|58.0|    0|    0|          113783|  26.55| C103|       S|
|         13|       0|     3|Saundercock, Mr. ...|  male|20.0|    0|    0|       A/5. 2151|   8.05| NULL|       S|
|         14|       0|     3|Andersson, Mr. An...|  male|39.0|    1|    5|          347082| 31.275| NULL|       S|
|         15|       0|     3|Vestrom, Miss. Hu...|female|14.0|    0|    0|          350406| 7.8542| NULL|       S|
|         16|       1|     2|Hewlett, Mrs. (Ma...|female|55.0|    0|    0|          248706|   16.0| NULL|       S|
|         17|       0|     3|Rice, Master. Eugene|  male| 2.0|    4|    1|          382652| 29.125| NULL|       Q|
|         18|       1|     2|Williams, Mr. Cha...|  male|NULL|    0|    0|          244373|   13.0| NULL|       S|
|         19|       0|     3|Vander Planke, Mr...|female|31.0|    1|    0|          345763|   18.0| NULL|       S|
|         20|       1|     3|Masselmani, Mrs. ...|female|NULL|    0|    0|            2649|  7.225| NULL|       C|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
only showing top 20 rows

root
 |-- PassengerId: integer (nullable = true)
 |-- Survived: integer (nullable = true)
 |-- Pclass: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: double (nullable = true)
 |-- SibSp: integer (nullable = true)
 |-- Parch: integer (nullable = true)
 |-- Ticket: string (nullable = true)
 |-- Fare: double (nullable = true)
 |-- Cabin: string (nullable = true)
 |-- Embarked: string (nullable = true)
```

Spark Assignment 2

Thejas Bharadwaj

```
+-----------+--------+------+----+---+---+-----+-----+------+----+-----+--------+
|PassengerId|Survived|Pclass|Name|Sex|Age|SibSp|Parch|Ticket|Fare|Cabin|Embarked|
+-----------+--------+------+----+---+---+-----+-----+------+----+-----+--------+
|          0|       0|     0|   0|  0|177|    0|    0|     0|   0|  687|       2|
+-----------+--------+------+----+---+---+-----+-----+------+----+-----+--------+

+------+-----+
|   Sex|count|
+------+-----+
|female|  314|
|  male|  577|
+------+-----+

+--------+-----+
|Embarked|count|
+--------+-----+
|       Q|   77|
|    NULL|    2|
|       C|  168|
|       S|  644|
+--------+-----+

+------+-------------------+
|   Sex|      avg(Survived)|
+------+-------------------+
|female| 0.7420382165605095|
|  male|0.18890814558058924|
+------+-------------------+

+------+------------------+
|Pclass|          avg(Age)|
+------+------------------+
|     1|38.233440860215055|
|     3| 25.14061971830986|
|     2| 29.87763005780347|
+------+------------------+
```


```
+------+-----+
|   Sex|count|
+------+-----+
|female|  314|
|  male|  577|
+------+-----+

+--------+-----+
|Embarked|count|
+--------+-----+
|       Q|   77|
|    NULL|    2|
|       C|  168|
|       S|  644|
+--------+-----+

+------+-------------------+
|   Sex|      avg(Survived)|
+------+-------------------+
|female| 0.7420382165605095|
|  male|0.18890814558058924|
+------+-------------------+

+------+------------------+
|Pclass|          avg(Age)|
+------+------------------+
|     1|38.233440860215055|
|     3| 25.14061971830986|
|     2| 29.87763005780347|
+------+------------------+
```


```
+--------+-------------------+
|Embarked|      avg(Survived)|
+--------+-------------------+
|       Q|0.38961038961038963|
|    NULL|                1.0|
|       C| 0.5535714285714286|
|       S|0.33695652173913043|
+--------+-------------------+
```

2) Feature Engineering - Create new attributes that may be derived from the existing attributes. This may include removing certain columns in the dataset. (30 pts)

Spark Assignment 2

Thejas Bharadwaj

3) Prediction - Use the train.csv to train a Machine Learning model of your choice & test it on the test.csv. You are required to predict if the records in test.csv survived or not. Note(1 = Survived, 0 = Dead) (50 pts)

Note – Both Feature engineering and Prediction has been explain via comments in the code and zeppelin notebook. Please refer to zeppelin notebook to explore the code.