Mehul Natu (002743870)

Big Data System Engineering with Scala

Fall 2023

Assignment No. Spark 2

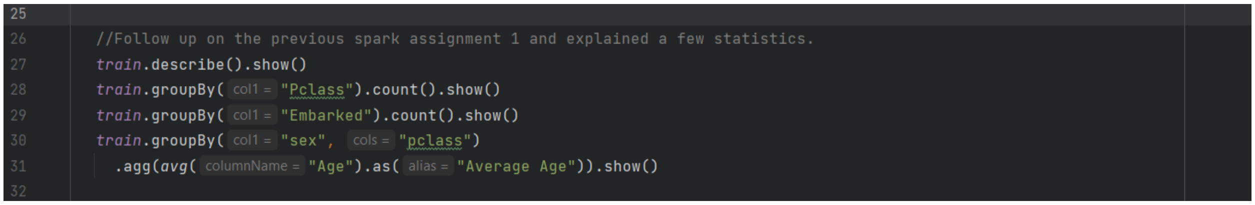
GitHub Link - https://github.com/Mehul-Natu/CSYE7200_Spark2.git

List Of tasks Implemented.

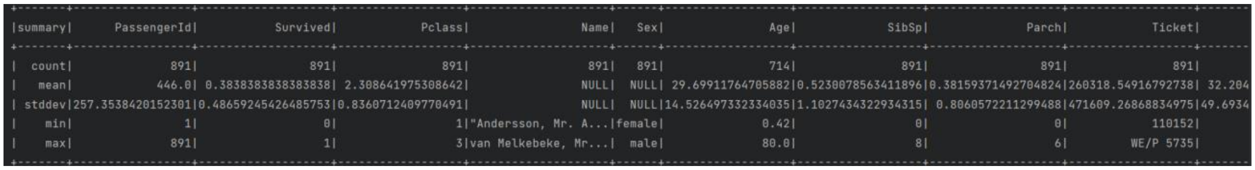
- Exploratory Data Analysis
  - Showed count, mean, stddev, min, max for all the columns.
  - Calculated count of each type in Pclass
  - Calculated count of each type in Embarked
  - Calculated average age of each sex in different class
- Feature Engineering
  - Created columns – isAlone and Companions. To store info if the person is alone or not and if not then is with how many others
- Prediction –
  - Predicted the survival of each person in test dataset using RandomForestCLassifier

Code & Results

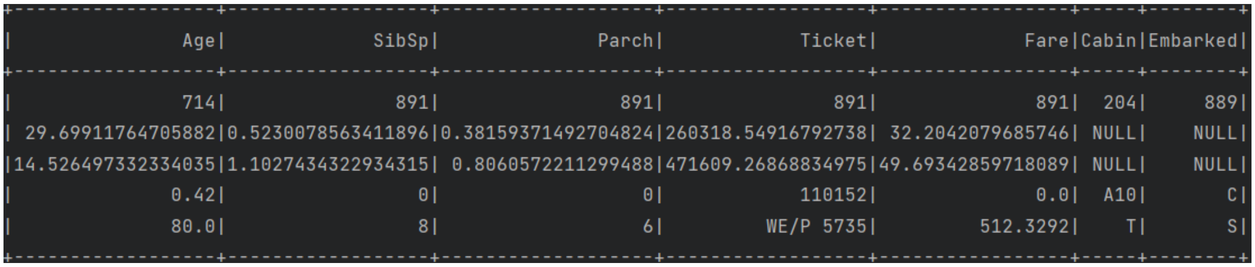
1. Exploratory Data Analysis



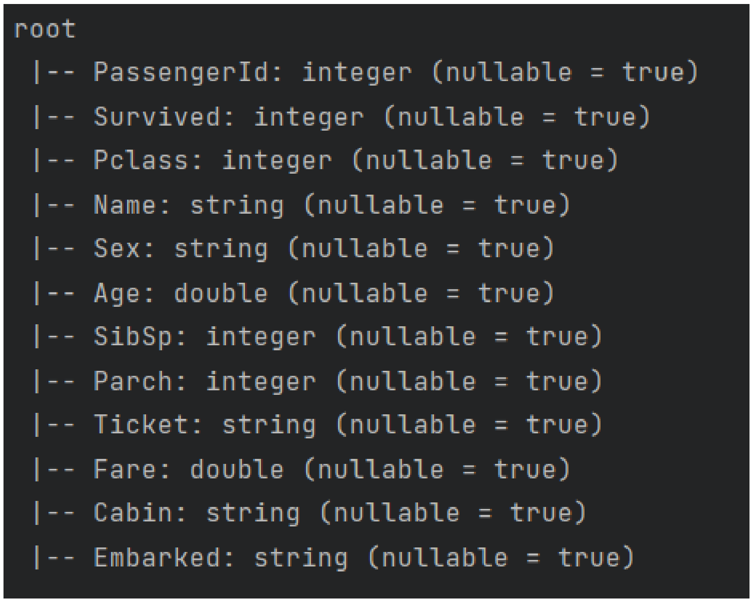
```
//Follow up on the previous spark assignment 1 and explained a few statistics.
train.describe().show()
train.groupBy("Pclass").count().show()
train.groupBy("Embarked").count().show()
train.groupBy("sex", "pclass")
  .agg(avg("Age").as("Average Age")).show()
```



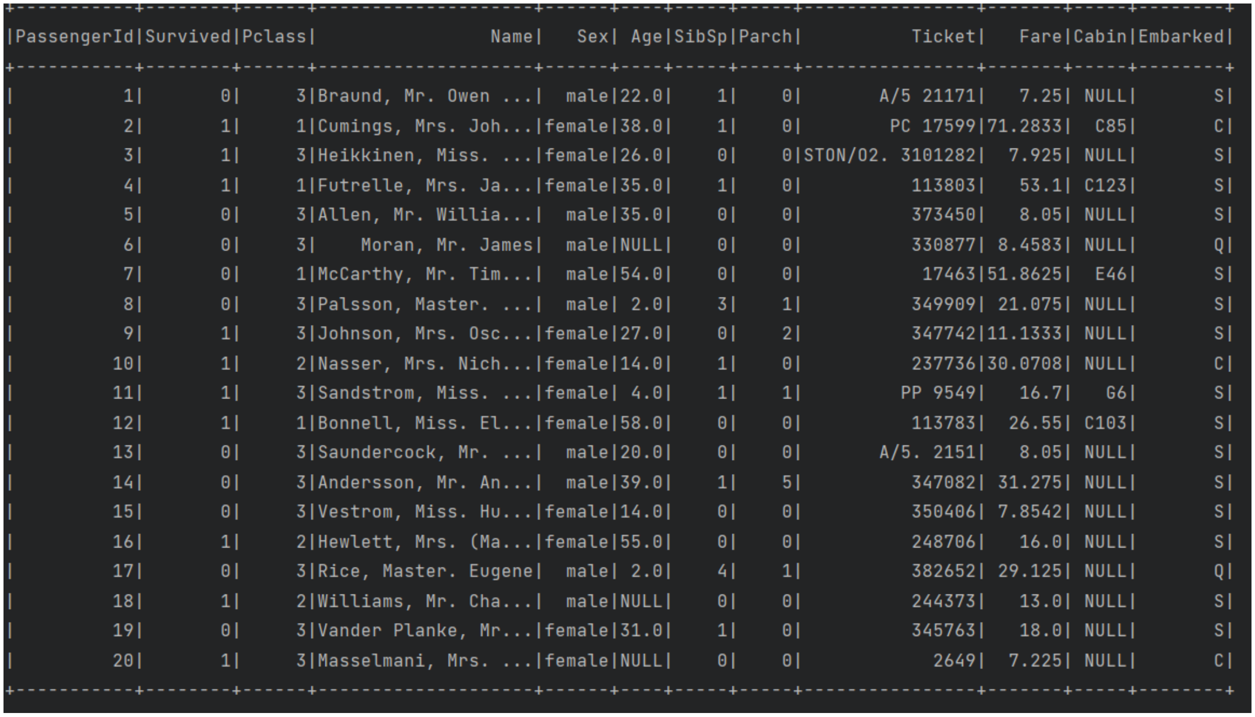
```
+-------+----------------+-------------------+------------------+--------------------+------+------------------+------------------+-------------------+------------------+-------
|summary|     PassengerId|           Survived|            Pclass|                Name|   Sex|               Age|             SibSp|              Parch|            Ticket|
+-------+----------------+-------------------+------------------+--------------------+------+------------------+------------------+-------------------+------------------+-------
|  count|             891|                891|               891|                 891|   891|               714|               891|                891|               891|
|   mean|           446.0| 0.3838383838383838| 2.308641975308642|                NULL|  NULL| 29.69911764705882|0.5230078563411896|0.38159371492704824|260318.54916792738| 32.204
| stddev|257.3538420152301|0.48659245426485753|0.8360712409770491|                NULL|  NULL|14.526497332334035|1.1027434322934315| 0.8060572211299488|471609.26868834975|49.6934
|    min|               1|                  0|                 1|"Andersson, Mr. A...|female|              0.42|                 0|                  0|            110152|
|    max|             891|                  1|                 3|van Melkebeke, Mr...|  male|              80.0|                 8|                  6|        WE/P 5735|
+-------+----------------+-------------------+------------------+--------------------+------+------------------+------------------+-------------------+------------------+-------
```



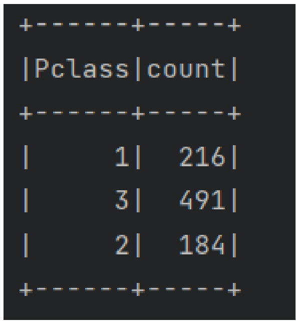
```
+------------------+------------------+-------------------+------------------+-----------------+-----+--------+
|               Age|             SibSp|              Parch|            Ticket|             Fare|Cabin|Embarked|
+------------------+------------------+-------------------+------------------+-----------------+-----+--------+
|               714|               891|                891|               891|              891|  204|     889|
| 29.69911764705882|0.5230078563411896|0.38159371492704824|260318.54916792738| 32.2042079685746| NULL|    NULL|
|14.526497332334035|1.1027434322934315| 0.8060572211299488|471609.26868834975|49.69342859718089| NULL|    NULL|
|              0.42|                 0|                  0|            110152|              0.0|  A10|       C|
|              80.0|                 8|                  6|         WE/P 5735|         512.3292|    T|       S|
+------------------+------------------+-------------------+------------------+-----------------+-----+--------+
```



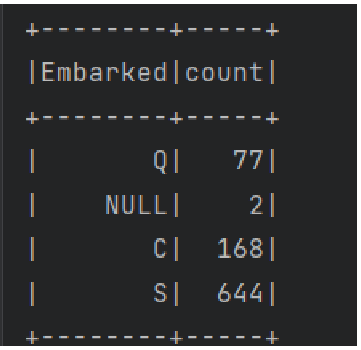
```
root
 |-- PassengerId: integer (nullable = true)
 |-- Survived: integer (nullable = true)
 |-- Pclass: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: double (nullable = true)
 |-- SibSp: integer (nullable = true)
 |-- Parch: integer (nullable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



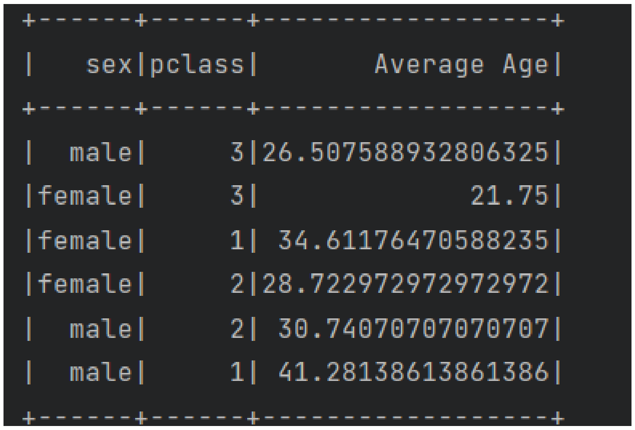
```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+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



```
+------+-----+
|Pclass|count|
+------+-----+
|     1|  216|
|     3|  491|
|     2|  184|
+------+-----+
```

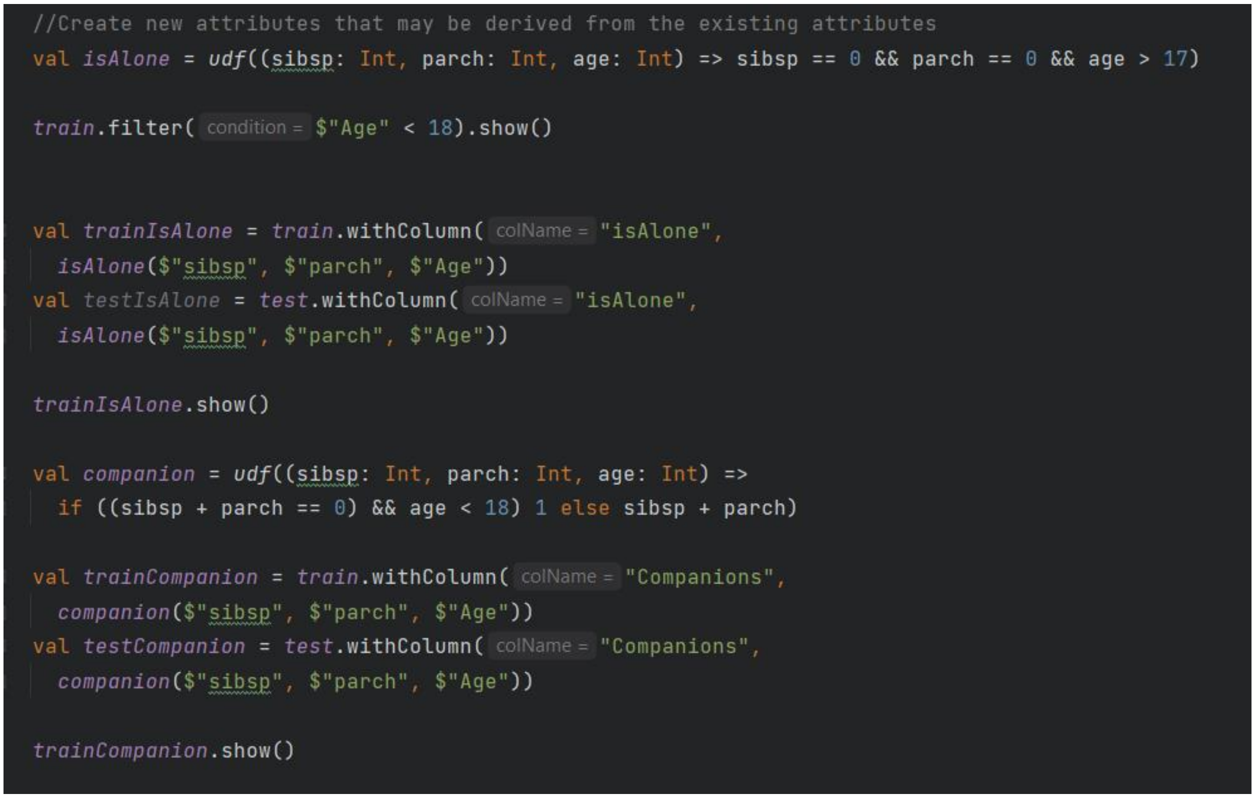


```
+--------+-----+
|Embarked|count|
+--------+-----+
|       Q|   77|
|    NULL|    2|
|       C|  168|
|       S|  644|
+--------+-----+
```

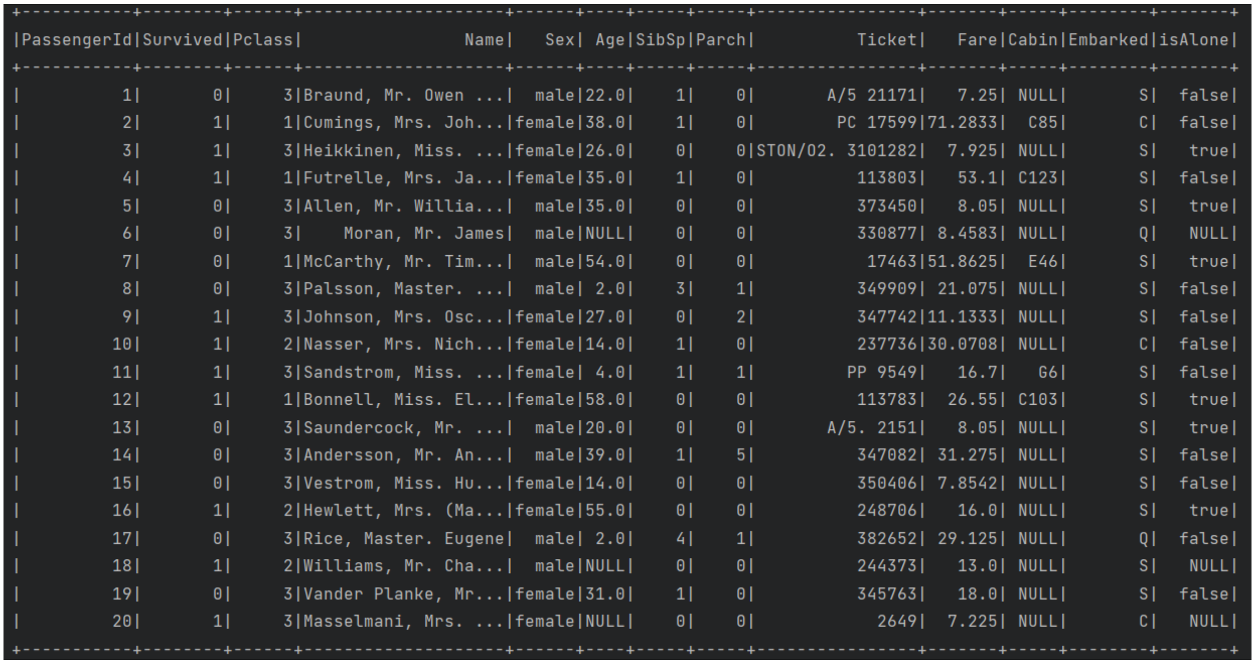


```
+------+------+------------------+
|   sex|pclass|       Average Age|
+------+------+------------------+
|  male|     3|26.507588932806325|
|female|     3|             21.75|
|female|     1| 34.61176470588235|
|female|     2|28.722972972972972|
|  male|     2| 30.74070707070707|
|  male|     1| 41.28138613861386|
+------+------+------------------+
```

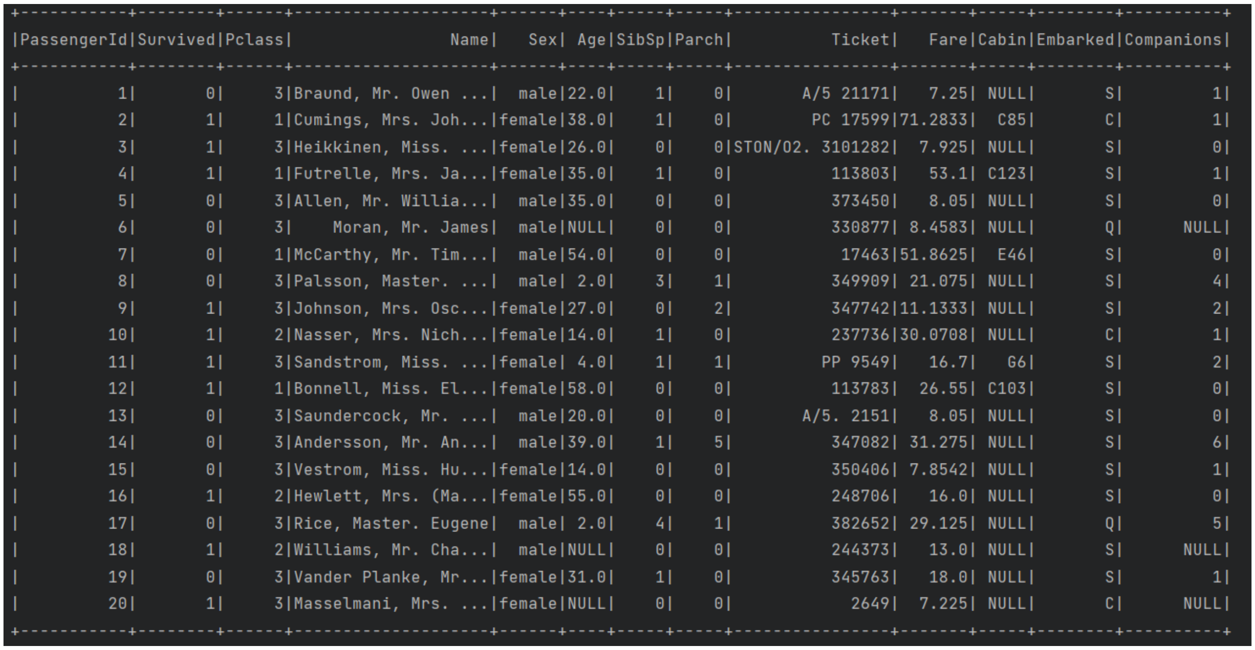
2. Feature Engineering



```scala
//Create new attributes that may be derived from the existing attributes
val isAlone = udf((sibsp: Int, parch: Int, age: Int) => sibsp == 0 && parch == 0 && age > 17)

train.filter( condition = $"Age" < 18).show()


val trainIsAlone = train.withColumn( colName = "isAlone",
  isAlone($"sibsp", $"parch", $"Age"))
val testIsAlone = test.withColumn( colName = "isAlone",
  isAlone($"sibsp", $"parch", $"Age"))

trainIsAlone.show()

val companion = udf((sibsp: Int, parch: Int, age: Int) =>
  if ((sibsp + parch == 0) && age < 18) 1 else sibsp + parch)

val trainCompanion = train.withColumn( colName = "Companions",
  companion($"sibsp", $"parch", $"Age"))
val testCompanion = test.withColumn( colName = "Companions",
  companion($"sibsp", $"parch", $"Age"))

trainCompanion.show()
```

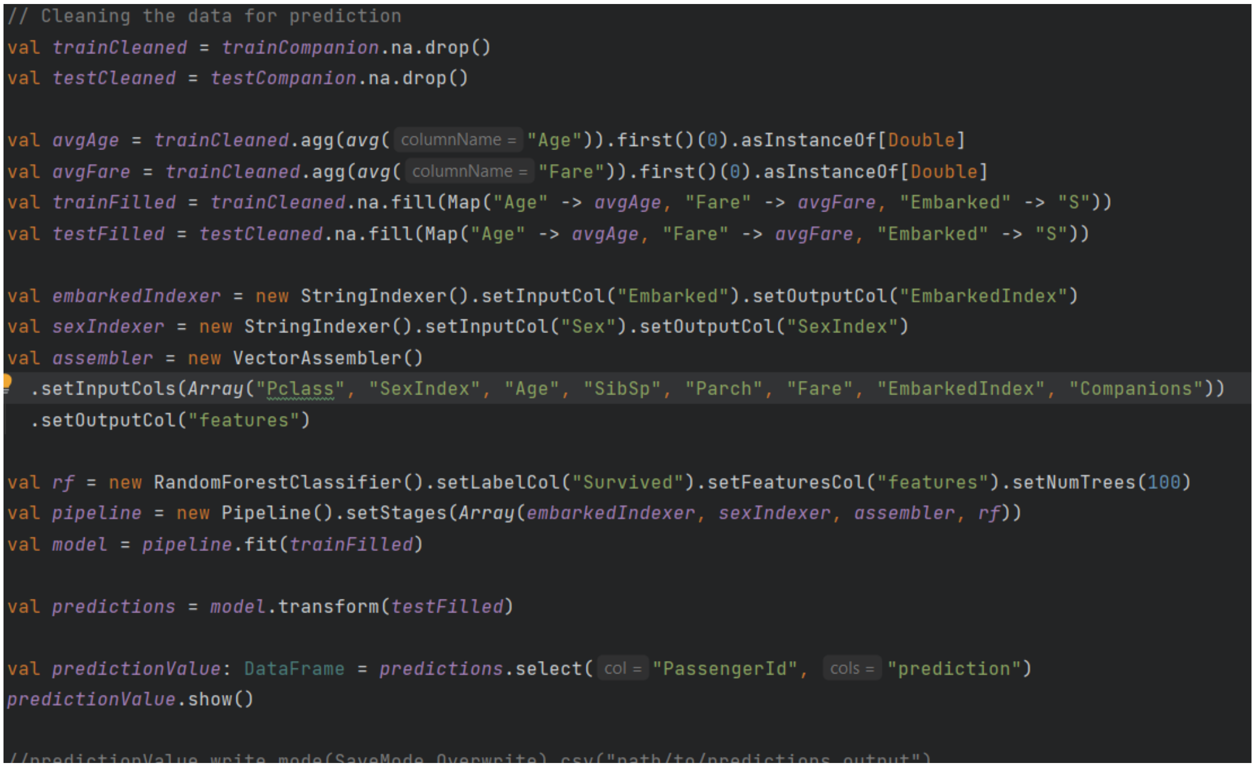


```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|isAlone|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+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

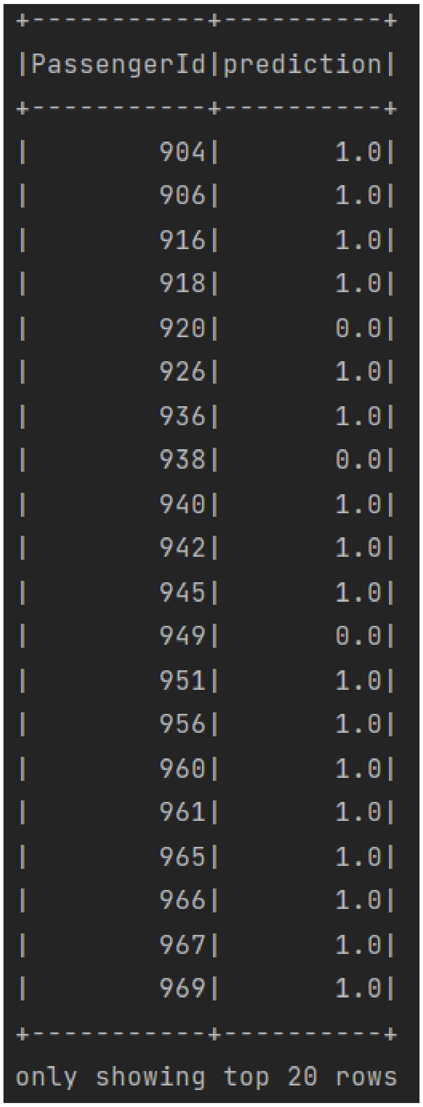


```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+----------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|Companions|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+---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

3. Prediction



```
// Cleaning the data for prediction
val trainCleaned = trainCompanion.na.drop()
val testCleaned = testCompanion.na.drop()

val avgAge = trainCleaned.agg(avg( columnName = "Age")).first()(0).asInstanceOf[Double]
val avgFare = trainCleaned.agg(avg( columnName = "Fare")).first()(0).asInstanceOf[Double]
val trainFilled = trainCleaned.na.fill(Map("Age" -> avgAge, "Fare" -> avgFare, "Embarked" -> "S"))
val testFilled = testCleaned.na.fill(Map("Age" -> avgAge, "Fare" -> avgFare, "Embarked" -> "S"))

val embarkedIndexer = new StringIndexer().setInputCol("Embarked").setOutputCol("EmbarkedIndex")
val sexIndexer = new StringIndexer().setInputCol("Sex").setOutputCol("SexIndex")
val assembler = new VectorAssembler()
  .setInputCols(Array("Pclass", "SexIndex", "Age", "SibSp", "Parch", "Fare", "EmbarkedIndex", "Companions"))
  .setOutputCol("features")

val rf = new RandomForestClassifier().setLabelCol("Survived").setFeaturesCol("features").setNumTrees(100)
val pipeline = new Pipeline().setStages(Array(embarkedIndexer, sexIndexer, assembler, rf))
val model = pipeline.fit(trainFilled)

val predictions = model.transform(testFilled)

val predictionValue: DataFrame = predictions.select( col = "PassengerId",  cols = "prediction")
predictionValue.show()

//predictionValue.write.mode(SaveMode.Overwrite).csv("path/to/predictions_output")
```



```
+-----------+----------+
|PassengerId|prediction|
+-----------+----------+
|        904|       1.0|
|        906|       1.0|
|        916|       1.0|
|        918|       1.0|
|        920|       0.0|
|        926|       1.0|
|        936|       1.0|
|        938|       0.0|
|        940|       1.0|
|        942|       1.0|
|        945|       1.0|
|        949|       0.0|
|        951|       1.0|
|        956|       1.0|
|        960|       1.0|
|        961|       1.0|
|        965|       1.0|
|        966|       1.0|
|        967|       1.0|
|        969|       1.0|
+-----------+----------+
only showing top 20 rows
```