BD3P2 - Pyspark ETL Assignment

GROUP-10

Dataset: Titanic - db table

i.Extract: Load the data

- Read data as pandas dataframe
- then create spark dataframe
- create a table view "titanic" as spark SQL

ii.Transform: Exploratory data analysis using spark df

- Unique passengerId count
- GroupBy sex and count of survived
- GroupBy Pclass and sum of Fare
- Update column age values as 1,2,3 ->child<10,teen>10 to <25,adult>25
- GroupBy age,Embarked,Pclass and count of survived

iii.Load: Save analysis report

- save as tables - partitioned by sex

i. Extract: Load the data

a. Read data as pandas dataframe

```
from random import random
import os
from pyspark.sql import SparkSession
from pyspark.sql.functions import regexp_replace, col
from pyspark.sql.types import IntegerType, StringType
from pyspark.sql import SQLContext
import pandas as pd
```

```
df_path = 'C:/Users/Jerome/Downloads/pyspark_data/titanic.csv'
data = pd.read_csv(df_path)
```

```
print(data)
```

```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
```

```
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                  Name     Sex   Age  SibSp  \
0                              Braund, Mr. Owen Harris    male  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                               Heikkinen, Miss. Laina  female  26.0      0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                             Allen, Mr. William Henry    male  35.0      0
..                                                 ...     ...   ...    ...
886                              Montvila, Rev. Juozas    male  27.0      0
887                       Graham, Miss. Margaret Edith  female  19.0      0
888           Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                              Behr, Mr. Karl Howell    male  26.0      0
890                                Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     ...               ...      ...   ...      ...
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

```
os.environ['PYSPARK SUBMIT ARGS'] = " --master local[2] pyspark-shell"
os.environ['JAVA_HOME'] = 'C:/Program Files/Java/jdk1.8.0_311'
```

b. Creating spark dataframe

```
#Creating a spark Session
spark_session = SparkSession.builder.master('local').appName('titanic_csv').getOrCreate()
sc = spark_session.sparkContext
```

```
titanic_data = spark_session.read.csv(df_path ,sep=',',inferSchema=True, header=True ,null
```

```
print(titanic_data)
```

```
DataFrame[PassengerId: int, Survived: int, Pclass: int, Name: string, Sex: string, Age:
```

```
1 print(titanic_data.show())
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+------------
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Tic
+-----------+--------+------+--------------------+------+----+-----+-----+------------
|          1|       0|     3|Braund, Mr. Owen ...|  male|22.0|    1|    0|       A/5 21
|          2|       1|     1|Cumings, Mrs. Joh...|female|38.0|    1|    0|        PC 17
|          3|       1|     3|Heikkinen, Miss. ...|female|26.0|    0|    0|STON/O2. 3101
|          4|       1|     1|Futrelle, Mrs. Ja...|female|35.0|    1|    0|          113
|          5|       0|     3|Allen, Mr. Willia...|  male|35.0|    0|    0|          373
|          6|       0|     3|    Moran, Mr. James|  male|null|    0|    0|          330
|          7|       0|     1|McCarthy, Mr. Tim...|  male|54.0|    0|    0|           17
|          8|       0|     3|Palsson, Master. ...|  male| 2.0|    3|    1|          349
|          9|       1|     3|Johnson, Mrs. Osc...|female|27.0|    0|    2|          347
|         10|       1|     2|Nasser, Mrs. Nich...|female|14.0|    1|    0|          237
|         11|       1|     3|Sandstrom, Miss. ...|female| 4.0|    1|    1|         PP 9
|         12|       1|     1|Bonnell, Miss. El...|female|58.0|    0|    0|          113
|         13|       0|     3|Saundercock, Mr. ...|  male|20.0|    0|    0|       A/5. 2
|         14|       0|     3|Andersson, Mr. An...|  male|39.0|    1|    5|          347
|         15|       0|     3|Vestrom, Miss. Hu...|female|14.0|    0|    0|          350
|         16|       1|     2|Hewlett, Mrs. (Ma...|female|55.0|    0|    0|          248
|         17|       0|     3|Rice, Master. Eugene|  male| 2.0|    4|    1|          382
|         18|       1|     2|Williams, Mr. Cha...|  male|null|    0|    0|          244
|         19|       0|     3|Vander Planke, Mr...|female|31.0|    1|    0|          345
|         20|       1|     3|Masselmani, Mrs. ...|female|null|    0|    0|            2
+-----------+--------+------+--------------------+------+----+-----+-----+------------
only showing top 20 rows

None
```

```
1 titanic_data.printSchema()
```

```
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)
```

```
1 titanic_data.describe().show()
```

```
+-------+-----------------+-------------------+------------------+--------------------+
|summary|      PassengerId|           Survived|            Pclass|                Name|
```

```
+-------+-----------------+-------------------+------------------+--------------------+
|  count|              891|                891|               891|                 891|
|   mean|            446.0| 0.3838383838383838| 2.308641975308642|                null|
| stddev|257.3538420152301|0.4865924542648575|0.8360712409770491|                null|
|    min|                1|                  0|                 1|"Andersson, Mr. A...|
|    max|              891|                  1|                 3|van Melkebeke, Mr...|
+-------+-----------------+-------------------+------------------+--------------------+
```

c. Create a table view "titanic" as spark SQL

```
1 sqlContext = SQLContext(sc)
2 sqlContext.registerDataFrameAsTable(titanic_data, "titanic_table")
```

```
C:\Users\Jerome\anaconda3\lib\site-packages\pyspark\sql\context.py:77: FutureWarning: D
  warnings.warn(
```

```
1 # Using query show table
2 sqlContext.sql("SELECT * FROM titanic_table ")
```

```
DataFrame[PassengerId: int, Survived: int, Pclass: int, Name: string, Sex: string, Age:
```

```
1 # Using query show spark netflix data
2 titanic = sqlContext.sql("SELECT * FROM titanic_table ")
3 titanic.show()
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+-------------
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Tic
+-----------+--------+------+--------------------+------+----+-----+-----+-------------
|          1|       0|     3|Braund, Mr. Owen ...|  male|22.0|    1|    0|       A/5 21
|          2|       1|     1|Cumings, Mrs. Joh...|female|38.0|    1|    0|        PC 17
|          3|       1|     3|Heikkinen, Miss. ...|female|26.0|    0|    0|STON/O2. 3101
|          4|       1|     1|Futrelle, Mrs. Ja...|female|35.0|    1|    0|          113
|          5|       0|     3|Allen, Mr. Willia...|  male|35.0|    0|    0|          373
|          6|       0|     3|    Moran, Mr. James|  male|null|    0|    0|          330
|          7|       0|     1|McCarthy, Mr. Tim...|  male|54.0|    0|    0|           17
|          8|       0|     3|Palsson, Master. ...|  male| 2.0|    3|    1|          349
|          9|       1|     3|Johnson, Mrs. Osc...|female|27.0|    0|    2|          347
|         10|       1|     2|Nasser, Mrs. Nich...|female|14.0|    1|    0|          237
|         11|       1|     3|Sandstrom, Miss. ...|female| 4.0|    1|    1|         PP 9
|         12|       1|     1|Bonnell, Miss. El...|female|58.0|    0|    0|          113
|         13|       0|     3|Saundercock, Mr. ...|  male|20.0|    0|    0|       A/5. 2
|         14|       0|     3|Andersson, Mr. An...|  male|39.0|    1|    5|          347
|         15|       0|     3|Vestrom, Miss. Hu...|female|14.0|    0|    0|          350
|         16|       1|     2|Hewlett, Mrs. (Ma...|female|55.0|    0|    0|          248
|         17|       0|     3|Rice, Master. Eugene|  male| 2.0|    4|    1|          382
|         18|       1|     2|Williams, Mr. Cha...|  male|null|    0|    0|          244
|         19|       0|     3|Vander Planke, Mr...|female|31.0|    1|    0|          345
```

```
|         20|       1|     3|Masselmani, Mrs. ...|female|null|    0|    0|            2
+-----------+--------+------+--------------------+------+----+-----+-----+-------------
only showing top 20 rows
```

ii.Transform: Exploratory data analysis using spark df

Unique passengerId count

```
Unique_Id_count = sqlContext.sql("select distinct count(PassengerId) as Total from titanic
Unique_Id_count.show()
```

```
+-----+
|Total|
+-----+
|  891|
+-----+
```

```
#Distinct count through spark dataframe
titanic.distinct().count()
```

```
891
```

GroupBy sex and count of survived

```
#
titanic.filter(titanic.Survived==1).groupBy("Sex").count().show()
```

```
+------+-----+
|   Sex|count|
+------+-----+
|female|  233|
|  male|  109|
+------+-----+
```

```
#
titanic.groupBy("Sex", "Survived").count().show()
```

```
+------+--------+-----+
|   Sex|Survived|count|
+------+--------+-----+
|  male|       0|  468|
|female|       1|  233|
|female|       0|   81|
```

```
|  male|       1|  109|
+------+--------+-----+
```

GroupBy Pclass and sum of Fare

```
#
titanic.groupBy("Pclass").sum("Fare").show()
```

```
+------+------------------+
|Pclass|         sum(Fare)|
+------+------------------+
|     1|18177.412499999984|
|     3| 6714.695100000002|
|     2|3801.8416999999995|
+------+------------------+
```

Update column age values as 1,2,3 ->child<10, teen>10 to <25, adult>25

```
titanic.agg({'Age': 'max'}).show()
```

```
+--------+
|max(Age)|
+--------+
|    80.0|
+--------+
```

```
titanic = titanic.withColumn("Age", titanic['Age'].cast("int"))
titanic = titanic.withColumn("Age", titanic['Age'].cast("string"))
for i in range(25,81):
    titanic = titanic.withColumn('Age', regexp_replace('Age', str(i), 'Adult'))
for i in range(10,25):
    titanic = titanic.withColumn('Age', regexp_replace('Age', str(i), 'Teen'))
for i in range(0,10):
    titanic = titanic.withColumn('Age', regexp_replace('Age', str(i), 'Child'))
titanic.show()
```

```
+-----------+--------+------+--------------------+------+-----+-----+-----+------------
|PassengerId|Survived|Pclass|                Name|   Sex|  Age|SibSp|Parch|          Ti
+-----------+--------+------+--------------------+------+-----+-----+-----+------------
|          1|       0|     3|Braund, Mr. Owen ...|  male| Teen|    1|    0|       A/5 2
|          2|       1|     1|Cumings, Mrs. Joh...|female|Adult|    1|    0|        PC 1
|          3|       1|     3|Heikkinen, Miss. ...|female|Adult|    0|    0|STON/O2. 310
|          4|       1|     1|Futrelle, Mrs. Ja...|female|Adult|    1|    0|          11
|          5|       0|     3|Allen, Mr. Willia...|  male|Adult|    0|    0|          37
|          6|       0|     3|    Moran, Mr. James|  male| null|    0|    0|          33
|          7|       0|     1|McCarthy, Mr. Tim...|  male|Adult|    0|    0|           1
```

```
|          8|       0|     3|Palsson, Master. ...|  male|Child|    3|    1|          34
|          9|       1|     3|Johnson, Mrs. Osc...|female|Adult|    0|    2|          34
|         10|       1|     2|Nasser, Mrs. Nich...|female| Teen|    1|    0|          23
|         11|       1|     3|Sandstrom, Miss. ...|female|Child|    1|    1|          PP
|         12|       1|     1|Bonnell, Miss. El...|female|Adult|    0|    0|          11
|         13|       0|     3|Saundercock, Mr. ...|  male| Teen|    0|    0|       A/5.
|         14|       0|     3|Andersson, Mr. An...|  male|Adult|    1|    5|          34
|         15|       0|     3|Vestrom, Miss. Hu...|female| Teen|    0|    0|          35
|         16|       1|     2|Hewlett, Mrs. (Ma...|female|Adult|    0|    0|          24
|         17|       0|     3|Rice, Master. Eugene|  male|Child|    4|    1|          38
|         18|       1|     2|Williams, Mr. Cha...|  male| null|    0|    0|          24
|         19|       0|     3|Vander Planke, Mr...|female|Adult|    1|    0|          34
|         20|       1|     3|Masselmani, Mrs. ...|female| null|    0|    0|
+-----------+--------+------+--------------------+------+-----+-----+-----+------------
only showing top 20 rows
```

GroupBy age,Embarked,Pclass and count of survived

```
1 titanic.filter(titanic.Survived==1).groupBy("Age", "Embarked", "Pclass").count().show()
```

```
+-----+--------+------+-----+
|  Age|Embarked|Pclass|count|
+-----+--------+------+-----+
|Adult|       S|     1|   49|
| Teen|       S|     2|   15|
| Teen|       Q|     3|    5|
|Adult|       Q|     2|    1|
|Adult|       Q|     3|    1|
|Child|       C|     2|    2|
| Teen|       S|     3|   18|
|Child|       C|     3|    6|
| Teen|       S|     1|   15|
|Adult|       C|     2|    2|
|Child|       S|     2|   15|
|Child|       S|     3|   13|
|Adult|       C|     3|    3|
| Teen|       C|     1|   14|
|Adult|       Q|     1|    1|
|Child|       S|     1|    2|
|Adult|       S|     2|   44|
| Teen|       C|     2|    4|
|Adult|       C|     1|   39|
| Teen|       C|     3|    9|
+-----+--------+------+-----+
only showing top 20 rows
```

.

iii.Load: Save analysis report df

```
titanic.write.option("header",True).partitionBy("Sex").mode("overwrite").csv('./output')
```