Assignment 6: Data Frames

Repeat Assignment 5 using DataFrame in PySpark

Write the following programs in the Jupyter Notebook. For help with the assignment, you can use the tutorials linked to the course calendar Save your completed notebook as a .ipynb file as well as a .pdf file and submit them to Canvas->Assignments->Assignment 5: Data Frames PySpark

For this assignment you may need to install PySpark on your laptop. Please review lecture notes for additional information.

Data Frames Primer

In this primer, we will study a classic data set - the survivors in the sinking of the Titanic. As there were limited lifeboats, decisions were made prioritizing who would and would not survive. We will observe how different factors such as age, sex, and class affected a person’s chance of survival using data frames.

1] Input the following data into a data frame called titanic, and display the entire data frame:

Sex, Class, Survived, Died

Children, First, 6, 0

Children, Second, 24, 0

Children, Third, 27, 52

Men, First, 57, 118

Men, Second, 14, 154

Men, Third, 75, 387

Men, Crew, 192, 693

Women, First, 140, 4

Women, Second, 80, 13

Women, Third, 76, 89

Women, Crew, 20, 3

```
In [2]: import pyspark
```

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()

from pyspark.sql import functions as F
from pyspark.sql.types import StructType,StructField, StringType, Integ
titanic = [('Children','First',6,0),('Children','Second',24,0),('Childr
 ('Men','First',57,118),('Men','Second',14,154),
 ('Men','Third',75,387),('Men','Crew',192,693),
 ('Women','First',140,4),('Women','Second',80,13),
 ('Women','Third',76,89),('Women','Crew',20,3)]
schema = StructType([\
 StructField(('Sex'),StringType(),True),\
 StructField(('Class'),StringType(),True),\
 StructField(('Survived'),IntegerType(),True),\
 StructField(('Died'),IntegerType(),True)])
df = spark.createDataFrame(data=titanic,schema=schema)

df.show()
```

```
+--------+------+--------+----+
|     Sex| Class|Survived|Died|
+--------+------+--------+----+
|Children| First|       6|   0|
|Children|Second|      24|   0|
|Children| Third|      27|  52|
|     Men| First|      57| 118|
|     Men|Second|      14| 154|
|     Men| Third|      75| 387|
|     Men|  Crew|     192| 693|
|   Women| First|     140|   4|
|   Women|Second|      80|  13|
|   Women| Third|      76|  89|
|   Women|  Crew|      20|   3|
+--------+------+--------+----+
```

2] Delete the crew members from the data.

In [3]:

```
df_new = df.filter(F.col('Class')!='Crew')
df_new.show()
```

```
+--------+------+--------+----+
|     Sex| Class|Survived|Died|
+--------+------+--------+----+
|Children| First|       6|   0|
|Children|Second|      24|   0|
|Children| Third|      27|  52|
|     Men| First|      57| 118|
|     Men|Second|      14| 154|
|     Men| Third|      75| 387|
|   Women| First|     140|   4|
|   Women|Second|      80|  13|
|   Women| Third|      76|  89|
+--------+------+--------+----+
```

3] Create a new column that is the total number of people for that group (those who survived + died).

In [4]:
```
df_new.withColumn('Total',df['Survived']+df['Died']).show()
```

```
+--------+------+--------+----+-----+
|     Sex| Class|Survived|Died|Total|
+--------+------+--------+----+-----+
|Children| First|       6|   0|    6|
|Children|Second|      24|   0|   24|
|Children| Third|      27|  52|   79|
|     Men| First|      57| 118|  175|
|     Men|Second|      14| 154|  168|
|     Men| Third|      75| 387|  462|
|   Women| First|     140|   4|  144|
|   Women|Second|      80|  13|   93|
|   Women| Third|      76|  89|  165|
+--------+------+--------+----+-----+
```

4] Delete the column indicating the total number of people in that group.

In [5]:
```
new_titanic=df.drop('Total')
new_titanic.show()
```

```
+--------+------+--------+----+
|     Sex| Class|Survived|Died|
+--------+------+--------+----+
|Children| First|       6|   0|
|Children|Second|      24|   0|
|Children| Third|      27|  52|
|     Men| First|      57| 118|
|     Men|Second|      14| 154|
|     Men| Third|      75| 387|
|     Men|  Crew|     192| 693|
|   Women| First|     140|   4|
|   Women|Second|      80|  13|
|   Women| Third|      76|  89|
|   Women|  Crew|      20|   3|
+--------+------+--------+----+
```

5] Only show the rows where more than 80% of the people survived.

In [6]:
```
df_new.where("Survived>(Survived+Died)*0.8").show()
```

```
+--------+------+--------+----+
|     Sex| Class|Survived|Died|
+--------+------+--------+----+
|Children| First|       6|   0|
|Children|Second|      24|   0|
|   Women| First|     140|   4|
|   Women|Second|      80|  13|
+--------+------+--------+----+
```

In [ ]: