Homework 3

September 16, 2024

1 Part One

Titanic[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]

1.1 Translating the dataset for pandas

[107]:

```
titanic_data = {
    'sex': ['Children', 'Children', 'Children', 'Men', 'Men', 'Men', 'Men',␣
 ↪'Women', 'Women', 'Women', 'Women'],
    'class': ['First', 'Second', 'Third', 'First', 'Second', 'Third', 'Crew',␣
 ↪'First', 'Second', 'Third', 'Crew'],
    'survived': [6, 24, 27, 57, 14, 75, 192, 140, 80, 76, 20],
    'died': [0, 0, 52, 118, 154, 387, 693, 4, 13, 89, 3],
}
```

[108]:

```
import pandas as pd

titanic_data = pd.DataFrame(titanic_data)
titanic_data
```

[108]:

```
         sex   class  survived  died
0   Children   First         6     0
1   Children  Second        24     0
2   Children   Third        27    52
3        Men   First        57   118
4        Men  Second        14   154
5        Men   Third        75   387
6        Men    Crew       192   693
7      Women   First       140     4
8      Women  Second        80    13
9      Women   Third        76    89
10     Women    Crew        20     3
```

1.2 Delete the crew members from the data.

[109]:

```
titanic_data = titanic_data[titanic_data['class'] != 'Crew']
titanic_data
```

[109]:

```
        sex   class  survived  died
0  Children   First         6     0
1  Children  Second        24     0
2  Children   Third        27    52
3       Men   First        57   118
4       Men  Second        14   154
5       Men   Third        75   387
7     Women   First       140     4
8     Women  Second        80    13
9     Women   Third        76    89
```

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

[110]:

```
titanic_data['total-people'] = titanic_data['survived'] + titanic_data['died']
titanic_data
```

```
<ipython-input-110-ddd8b9bb7e98>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  titanic_data['total-people'] = titanic_data['survived'] + titanic_data['died']
```

[110]:

```
        sex   class  survived  died  total-people
0  Children   First         6     0             6
1  Children  Second        24     0            24
2  Children   Third        27    52            79
3       Men   First        57   118           175
4       Men  Second        14   154           168
5       Men   Third        75   387           462
7     Women   First       140     4           144
8     Women  Second        80    13            93
9     Women   Third        76    89           165
```

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

[111]:

```
titanic_data = titanic_data.drop(columns=['total-people'])
titanic_data
```

[111]:

```
        sex   class  survived  died
0  Children   First         6     0
1  Children  Second        24     0
2  Children   Third        27    52
```

```
3        Men   First        57   118
4        Men  Second        14   154
5        Men   Third        75   387
7      Women   First       140     4
8      Women  Second        80    13
9      Women   Third        76    89
```

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

```
[112]: titanic_data['survival-rate'] = titanic_data['survived'] /␣
         ↪(titanic_data['survived'] + titanic_data['died'])
       titanic_data
```

```
[112]:        sex   class  survived  died  survival-rate
       0  Children   First         6     0       1.000000
       1  Children  Second        24     0       1.000000
       2  Children   Third        27    52       0.341772
       3       Men   First        57   118       0.325714
       4       Men  Second        14   154       0.083333
       5       Men   Third        75   387       0.162338
       7     Women   First       140     4       0.972222
       8     Women  Second        80    13       0.860215
       9     Women   Third        76    89       0.460606
```

```
[113]: titanic_data[titanic_data['survival-rate'] > 0.80]
```

```
[113]:        sex   class  survived  died  survival-rate
       0  Children   First         6     0       1.000000
       1  Children  Second        24     0       1.000000
       7     Women   First       140     4       0.972222
       8     Women  Second        80    13       0.860215
```

2 Part 2

```
[114]: #%pip install pyspark
       from pyspark.sql import SparkSession
       from pyspark.sql.functions import col

       spark = SparkSession.builder.appName("Titanic Data").getOrCreate()

       titanic_data = spark.createDataFrame(titanic_data)
       titanic_data.show()
```

```
+--------+------+--------+----+-------------------+
|     sex| class|survived|died|      survival-rate|
+--------+------+--------+----+-------------------+
|Children| First|       6|   0|                1.0|
|Children|Second|      24|   0|                1.0|
|Children| Third|      27|  52|0.34177215189873417|
|     Men| First|      57| 118|0.32571428571428573|
|     Men|Second|      14| 154|0.08333333333333333|
|     Men| Third|      75| 387|0.16233766233766234|
|   Women| First|     140|   4| 0.9722222222222222|
|   Women|Second|      80|  13| 0.8602150537634409|
|   Women| Third|      76|  89|0.46060606060606063|
+--------+------+--------+----+-------------------+
```

[118]:

```
titanic_data = titanic_data.filter(col("survival-rate") > 0.8)
titanic_data.show()
```

```
+--------+------+--------+----+------------------+
|     sex| class|survived|died|     survival-rate|
+--------+------+--------+----+------------------+
|Children| First|       6|   0|               1.0|
|Children|Second|      24|   0|               1.0|
|   Women| First|     140|   4|0.9722222222222222|
|   Women|Second|      80|  13|0.8602150537634409|
+--------+------+--------+----+------------------+
```