Assignment 5: Data Frames

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

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 [3]:
```
import pandas as pd

column=["Sex","Class","Survived","Died"]
rows = {"Sex" :["Children","Children","Children","Men","Men","Men","Men","W
           "Class":["First","second","Third","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]}
titanic = pd.DataFrame(rows,columns=column)
print(titanic)
```

```
          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
```

2] Delete the crew members from the data.

In [4]:
```
titanic = titanic[titanic["Class"]!="Crew"]
print(titanic)
```

```
         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
```

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

In [13]:
```
titanic["Total_no.of_survivors"]=titanic["Survived"]+titanic["Died"]
print(titanic)
```

```
        Sex   Class  Survived  Died  Survival_Percentage  \
0  Children   First         6     0           100.000000
1  Children  second        24     0           100.000000
2  Children   Third        27    52            34.177215
3       Men   First        57   118            32.571429
4       Men  second        14   154             8.333333
5       Men   Third        75   387            16.233766
7     Women   First       140     4            97.222222
8     Women  second        80    13            86.021505
9     Women   Third        76    89            46.060606

   Total_no.of_survivors
0                      6
1                     24
2                     79
3                    175
4                    168
5                    462
7                    144
8                     93
9                    165
```

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

In [14]:
```
titanic["Survival_Percentage"]=titanic["Survived"]*100/titanic["Total_no.of
print(titanic)
```

```
        Sex   Class  Survived  Died  Survival_Percentage  \
0  Children   First         6     0           100.000000
1  Children  second        24     0           100.000000
2  Children   Third        27    52            34.177215
3       Men   First        57   118            32.571429
4       Men  second        14   154             8.333333
5       Men   Third        75   387            16.233766
7     Women   First       140     4            97.222222
8     Women  second        80    13            86.021505
9     Women   Third        76    89            46.060606

   Total_no.of_survivors
0                      6
1                     24
2                     79
3                    175
4                    168
5                    462
7                    144
8                     93
9                    165
```

In [15]:
```
titanic.drop(columns=['Total_no.of_survivors'],inplace=True)
print(titanic)
```

```
        Sex   Class  Survived  Died  Survival_Percentage
0  Children   First         6     0           100.000000
1  Children  second        24     0           100.000000
2  Children   Third        27    52            34.177215
3       Men   First        57   118            32.571429
4       Men  second        14   154             8.333333
5       Men   Third        75   387            16.233766
7     Women   First       140     4            97.222222
8     Women  second        80    13            86.021505
9     Women   Third        76    89            46.060606
```

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

In [8]:
```
titanic[titanic["Survival_Percentage"]>80]
```

Out[8]:

	Sex	Class	Survived	Died	Survival_Percentage
0	Children	First	6	0	100.000000
1	Children	second	24	0	100.000000
7	Women	First	140	4	97.222222
8	Women	second	80	13	86.021505

In [ ]:

In [ ]:

In [ ]:

In [ ]: