# Assignment 5

```
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.

Steps:
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 [1]:

```
import pandas as pd
Titanic_Data = {'Sex':['Children','Children','Children','Men','Men','Men','Me
                       '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]}

titanic_df = pd.DataFrame(Titanic_Data,columns=["Sex","Class","Survived","Die

print(titanic_df)
```

```
          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. Now only show the data of the people in first class.

In [2]:
```
print(titanic_df[titanic_df["Class"]=="First"])
```

```
        Sex  Class  Survived  Died
0  Children  First         6     0
3       Men  First        57   118
7     Women  First       140     4
```

3. Delete the crew members from the data.

In [3]:
```
titanic_df=titanic_df[titanic_df["Class"]!="Crew"]
titanic_df
```

Out[3]:

| | 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 |

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

In [4]:
```
titanic_df["Total"]=titanic_df["Survived"]+titanic_df["Died"]
titanic_df
```

Out[4]:

| | Sex | Class | Survived | Died | Total |
|---|---|---|---|---|---|
| **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 |

5. Create a new column with the percentage of people who survived.

In [5]:
```
titanic_df["survived_percent"]=titanic_df["Survived"]*100/titanic_df["Total"]
print(titanic_df)
```

```
        Sex   Class  Survived  Died  Total  survived_percent
0  Children   First         6     0      6        100.000000
1  Children  Second        24     0     24        100.000000
2  Children   Third        27    52     79         34.177215
3       Men   First        57   118    175         32.571429
4       Men  Second        14   154    168          8.333333
5       Men   Third        75   387    462         16.233766
7     Women   First       140     4    144         97.222222
8     Women  Second        80    13     93         86.021505
9     Women   Third        76    89    165         46.060606
```

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

In [6]:
```
titanic_df=titanic_df.drop(["Total"],axis = 1)
titanic_df
```

Out[6]:

| | Sex | Class | Survived | Died | survived_percent |
|---|---|---|---|---|---|
| **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 |

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

In [7]:
```
print(titanic_df[titanic_df["survived_percent"]>80])
```

```
        Sex   Class  Survived  Died  survived_percent
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
```

8. Then only show the rows where less than 40% of the people survived.

In [8]:
```
print(titanic_df[titanic_df["survived_percent"]<40])
```

```
        Sex   Class  Survived  Died  survived_percent
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
```

9. Calculate the total number of people that survived and died for each class, then report the percentages. (Hint: Use a grouped calculation.)

In [9]:
```
e_class=titanic_df.groupby('Class')
e_class_sum=e_class.sum()
e_class_sum["survived_percent"]=e_class_sum["Survived"]/(e_class_sum['Survive
e_class_sum["Died_percent"]=e_class_sum["Died"]/(e_class_sum['Survived']+e_cl
print(e_class_sum)
```

```
        Survived  Died  survived_percent  Died_percent
Class
First        203   122          0.624615      0.375385
Second       118   167          0.414035      0.585965
Third        178   528          0.252125      0.747875
```

10. Save your table in CSV format (as e.g. titanic_data.csv) with the first line as headers for the columns.

In [10]:
```
titanic_df.to_csv('titanic_data.csv',encoding='utf-8', index=False)
```

11. Duplicate the CSV file on your computer since you will be editing the copied version (e.g. titanic_data2.csv). Open the new CSV file in a text editor. Note the way the data is organized. Now, in the text editor, add new lines including the data for the crew that was removed earlier. (Help: the percentage of male crew and female crew that survived was 21.69% and 86.96%.)

12. Now read that updated CSV file into a new data frame called titanic2, and display the data.

In [11]:
```
titanic_df2=pd.read_csv("C:\\Users\\swapn\\OneDrive\\Desktop\\titanic_data_ne
```

In [12]:
```
titanic_df2
```

Out[12]:

| | Sex | Class | Survived | Died | survived_percent |
|---|---|---|---|---|---|
| **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 | Crew | 192 | 693 | 21.690000 |
| **6** | 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 |
| **10** | Women | Crew | 20 | 3 | 86.960000 |