Titanic Survival

Grading:

- Code: 90 pts
- Markdown Documentation (Documentation within IPython using Markdown): 10 pts

Due: 04/03/2023

We are going to study the survival rate of passengers on titanic and what variables affected survival.

Load the dataset in `titanic.xls`. It contains data on all the passengers that travelled on the Titanic.

In [ ]:

```
from IPython.core.display import HTML
import pandas as pd

HTML(filename='../data/titanic.html')
```

Out[ ]:

Data frame:titanic3

1309 observations and 14 variables, maximum # NAs:1188

Name	Labels	Units	Levels	Storage	NAs
pclass			3	integer	0
survived	Survived			double	0
name	Name			character	0
sex			2	integer	0
age	Age	Year		double	263
sibsp	Number of Siblings/Spouses Aboard			double	0
parch	Number of Parents/Children Aboard			double	0
ticket	Ticket Number			character	0
fare	Passenger Fare	British Pound (\243)		double	1
cabin			187	integer	0
embarked			3	integer	2
boat			28	integer	0
body	Body Identification Number			double	1188
home.dest	Home/Destination			character	0

Variable	Levels
pclass	1st
	2nd
	3rd
sex	female
	male
cabin	
	A10
	A11
	A14
	A16
	A18
	A19

A20
A21
A23
A24
A26
A29
A31
A32
A34
A36
A5
A6
A7
A9
B10
B101
B102
B11
B18
B19
B20
B22
B24
B26
B28
B3
B30
B35
B36
B37
B38
B39
B4
B41

B42
B45
B49
B5
B50
B51 B53 B55
B52 B54 B56
B57 B59 B63 B66
B58 B60
B61
B69
B71
B73
B77
B78
B79
B80
B82 B84
B86
B94
B96 B98
C101
C103
C104
C105
C106
C110
C111
C116
C118
C123
C124
C125
C126

C128
C130
C132
C148
C2
C22 C26
C23 C25 C27
C28
C30
C31
C32
C39
C45
C46
C47
C49
C50
C51
C52
C53
C54
C55 C57
C6
C62 C64
C65
C68
C7
C70
C78
C80
C82
C83
C85
C86

C87
C89
C90
C91
C92
C93
C95
C97
C99
D
D10 D12
D11
D15
D17
D19
D20
D21
D22
D26
D28
D30
D33
D34
D35
D36
D37
D38
D40
D43
D45
D46
D47
D48
D49

D50
D56
D6
D7
D9
E10
E101
E12
E121
E17
E24
E25
E31
E33
E34
E36
E38
E39 E41
E40
E44
E45
E46
E49
E50
E52
E58
E60
E63
E67
E68
E77
E8
F
F E46

	F E57
	F E69
	F G63
	F G73
	F2
	F33
	F38
	F4
	G6
	T
embarked	Cherbourg
	Queenstown
	Southampton
boat	
	1
	10
	11
	12
	13
	13 15
	13 15 B
	14
	15
	15 16
	16
	2
	3
	4
	5
	5 7
	5 9
	6
	7
	8

8 10
9
A
B
C
C D
D

In [ ]:

```
# you would need xlrd - pip install xlrd
t_file = pd.ExcelFile('../data/titanic.xls')
t_df = t_file.parse("titanic", header=None)
t_df.head()
```

Out[ ]:

	0	1	2	3	4	5	6	7	8	9	10	1
0	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boa
1	1	1	Allen, Miss. Elisabeth Walton	female	29	0	0	24160	211.3375	B5	S	
2	1	1	Allison, Master. Hudson Trevor	male	0.9167	1	2	113781	151.55	C22 C26	S	1
3	1	0	Allison, Miss. Helen Loraine	female	2	1	2	113781	151.55	C22 C26	S	Na
4	1	0	Allison, Mr. Hudson Joshua Creighton	male	30	1	2	113781	151.55	C22 C26	S	Na

Women and children first?

*** 1. Use the `groupby` method to calculate the proportion of passengers that survived by sex. (25 pts)***

In [ ]:

```
t_df.columns = t_df.loc[0]
t_df = t_df.drop(t_df.index[0]).reset_index().drop('index', axis=1)
t_df['survived'] = t_df.survived.astype('float')
survived_count = t_df[t_df['survived'] == 1].shape[0]
t_df.head(3)
```

Out[ ]:

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	b
0	1	1.0	Allen, Miss. Elisabeth Walton	female	29	0	0	24160	211.3375	B5	S	
1	1	1.0	Allison, Master. Hudson Trevor	male	0.9167	1	2	113781	151.55	C22 C26	S	
2	1	0.0	Allison, Miss. Helen Loraine	female	2	1	2	113781	151.55	C22 C26	S	N

- First things first, I have set the columns to be the first row, as it has been ambigous with a numbered columns and a numbered index.

After that, index reset, to restore the records indices to start from 0.

- Converting *survived* column to 'float' so arithmatic operations could make sense on it
- Calculated survived count to use in ratios calculation

In [ ]:

```
t_df.groupby('sex').sum(numeric_only=True)['survived']/survived_count
```

Out[ ]:

```
sex
female    0.678
male      0.322
Name: survived, dtype: float64
```

- Group by is used to group records by *sex*, using summation, to get the total survived number, and then dividing by the pre-calculated *survived_count* to get the ratio of each sex survival rate

Findings

- 68% of the survivors were Females
- 32% of the survivors were Males

*** 2. Calculate the same proportion, but by class and sex. (25 pts)***

In [ ]:

```
t_df.groupby(['sex','pclass']).sum(numeric_only=True)['survived']/survived_count
```

```
Out[ ]: sex     pclass
        female  1         0.278
                2         0.188
                3         0.212
        male    1         0.122
                2         0.050
                3         0.150
        Name: survived, dtype: float64
```

- Group by is used to group records by *sex* and *pclass*, using summation, to get the total survived number, and then dividing by the pre-calculated *survived_count* to get the ratio of each sex survival rate

Findings

- Females of first class are the highest in survival rate, with a percentage of 27.8% of the surviving number.
- Comes in second place the females of the third classs with a percentage of 21.2%.
- Highest Male surviving class is still below the lowest female surviving class, with a percentage of 15% and 18.7% respectively.

*** 3. Create age categories: children (under 14 years), adolescents (14-20), adult (21-64), and senior(65+), and calculate survival proportions by age category, class and sex. (40 pts)***

```
In [ ]: t_df['age_group'] = pd.cut(x=t_df['age'], bins=[0, 14, 20, 64, 200],
                                   labels=['children', 'adolescents', 'adult',
                                           'senior'])
```

- Using Pandas *cut* to generate categorical values for the age groups in a new column called *age_group*

```
In [ ]: t_cat = t_df.groupby(['age_group', 'pclass', 'sex']).sum(numeric_only=True)['surviv
        t_cat
```

```
Out[ ]: age_group    pclass  sex   
        children     1       female    0.028
                             male      0.022
                     2       female    0.036
                             male      0.024
                     3       female    0.042
                             male      0.044
        adolescents  1       female    0.040
                             male      0.006
                     2       female    0.024
                             male      0.004
                     3       female    0.040
                             male      0.018
        adult        1       female    0.186
                             male      0.076
                     2       female    0.124
                             male      0.018
                     3       female    0.062
                             male      0.056
        senior       1       female    0.002
                             male      0.002
                     2       female    0.000
                             male      0.000
                     3       female    0.000
                             male      0.000
        Name: survived, dtype: float64
```

- Group by is used to group records by *sex* , *pclass*, and *age_group*, using summation, to get the total survived number, and then dividing by the pre-calculated *survived_count* to get the ratio of each sex survival rate

Findings

- Nearly 0% of the surviving were seniors, however, the surviving <0% are all from the first class.
- Adults are the highest in surviving rates, with the higher surviving sex being females.
- Comes in second place children with a percentage of 19.6%.