Yin Htoo Htoo Naing

Student ID: 100366895

CPSC-4800-W01

Instructor: Nasim

11 Mar 2022

Table of Contents

Introduction

The purpose of this report to perform exploratory data analysis and answer the following three hypotheses based on the Titanic dataset.

The following is how I observe the variables and data types of this dataset and the first five observations.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          891 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Determine if the survival rate is associated to the class of passenger

```
Survived    0    1
Pclass
1          80  136
2          97   87
3         372  119

Text(0, 0.5, 'Percentage')
```
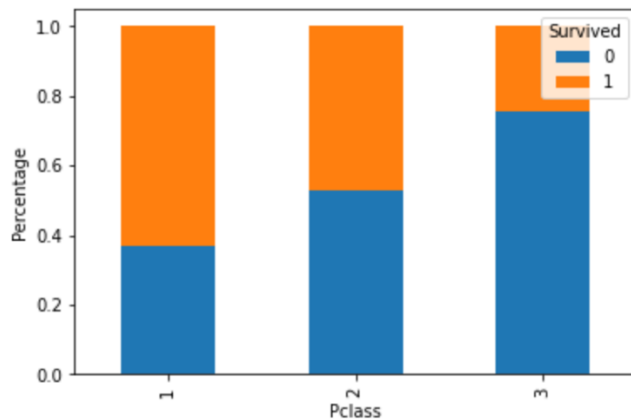

Fig: Stacked bar chart for Pclass and Survived

According to the bar chart here, the most number of people who are survived from the Titanic, are from class 1, followed by class 2 and 3.

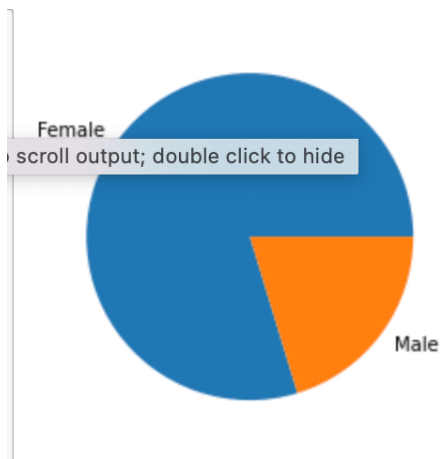# Determine if the survival rate is associated to the gender

Fig: Pie chart for Sex(Gender) of Survival

According to the pie chart comparing the gender, more female passenger survived than the male passenger from Titanic. Based on the calculation in python, 74% of the female passenger survived while only 18% of the male passenger survived in Titanic.

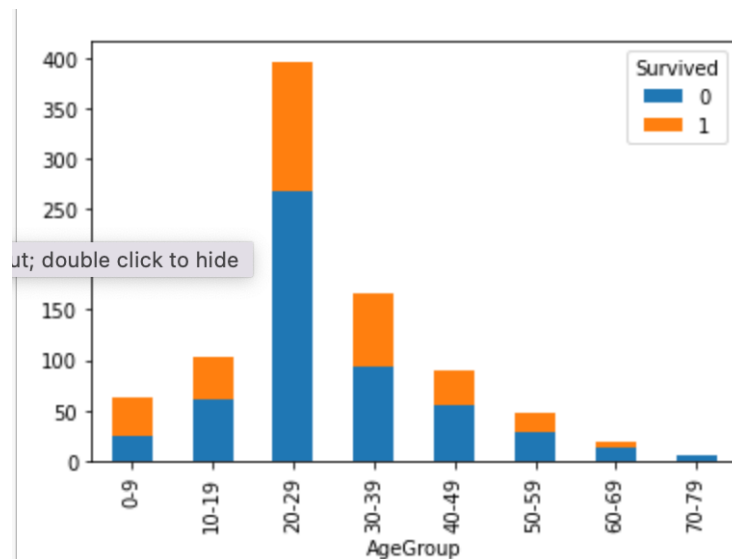# Determine the survival rate is associated to the age

Fig: Stacked bar chart for Age and Survived

According to the stacked bar chart showing the number people who survived or unsurvived in each age group, most people who survived are from the age group 20-29 from Titanic. The most passengers from the Titanic are from the age group 20-29 as well.
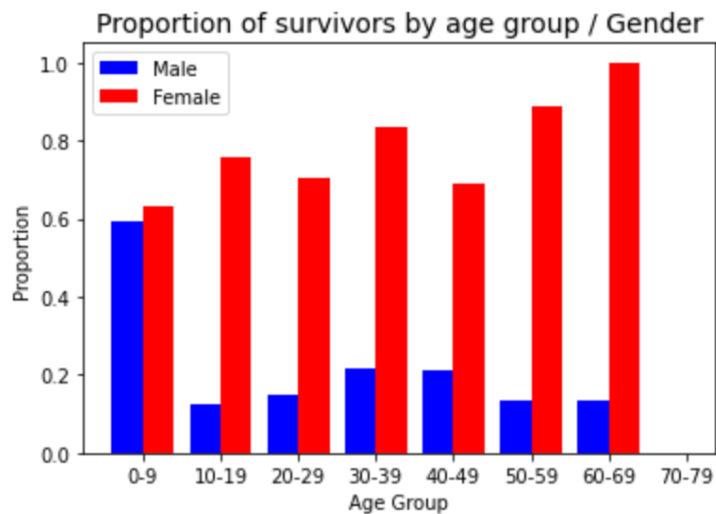

Fig: bar chart showing the proportion of survivors by age group/ Sex (Gender)

The average age of the children(below 18 years old) who are survived: 9 years old in female passengers and 8 years old in male passengers.

Sex	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
female	395.581818	0.690909	2.490909	9.236364	1.200000	1.036364	32.782653
male	461.103448	0.396552	2.672414	8.856379	1.706897	1.068966	29.739729

Correlation

Here is the Correlation matrix between numerical variables.

```
titanic.corr()
```

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.005007	-0.035144	0.042939	0.036847	-0.057527	-0.001652	0.012658
Survived	-0.005007	1.000000	-0.338481	-0.543351	-0.077221	-0.035322	0.081629	0.257307
Pclass	-0.035144	-0.338481	1.000000	0.131900	-0.369226	0.083081	0.018443	-0.549500
Sex	0.042939	-0.543351	0.131900	1.000000	0.093254	-0.114631	-0.245489	-0.182333
Age	0.036847	-0.077221	-0.369226	0.093254	1.000000	-0.308247	-0.189119	0.096067
SibSp	-0.057527	-0.035322	0.083081	-0.114631	-0.308247	1.000000	0.414838	0.159651
Parch	-0.001652	0.081629	0.018443	-0.245489	-0.189119	0.414838	1.000000	0.216225
Fare	0.012658	0.257307	-0.549500	-0.182333	0.096067	0.159651	0.216225	1.000000

Based on the correlation coefficient -0.338481 between Pclass and Survived, we can interpret that there is a weak linear relationship.

Based on the correlation coefficient -0.543351 between Sex and Survived, we can interpret that there is a moderate linear relationship.

Based on the correlation coefficient -0.077221 between Age and Survived, we can interpret that there is a weak linear relationship.