Chapter 6 - Exercise 3: Titanic Disaster

Vào ngày 15 tháng 4 năm 1912, trong chuyến hải trình đầu tiên của mình, tàu Titanic đã chìm sau khi va chạm với một tảng băng trôi, đã có 1502 nạn nhân mãi mãi ra đi trong tổng số 2224 hành khách và thủy thủ đoàn.

Thông tin về Titanic Disaster có thể xem tại: https://www.kaggle.com/c/titanic/data (https://www.kaggle.com/c/titanic/data)

Dựa trên tập tin *train.csv* , hãy thực hiện các yêu cầu sau:

Yêu cầu

Câu 1:

- a) Đọc dữ liệu từ tập tin train.csv và lưu vào biến titanic. Hiển thị 5 dòng dữ liệu đầu của titanic
- b) Thiết lập cột index cho titanic là PassengerId.

Câu 2:

- Tạo pie chart thể hiện tỷ lệ hành khách nam/nữ trên tàu.

Câu 3:

- Cho biết có bao nhiêu người còn sống sót
- Cho biết tỷ lệ bao nhiêu người còn sống sót

Câu 4:

- 4a) Vẽ biểu đồ histogram của cột vé (Fare), bổ sung các thông tin xlabel, ylabel, title. Bạn nhận xét gì về biểu đồ vừa vẽ
- 4b) Vẽ biểu đồ histogram của cột vé (Fare). Bạn nhận xét gì về biểu đồ vừa vẽ

Câu 5:

- a) Xem thông tin mô tả (describe) của biến Fare và vẽ biểu đồ Boxplot
- b) Vẽ biểu đồ Barplot thể hiện tổng số hành khách trong mỗi class
- c) Vẽ biểu đồ Barplot thể hiện tổng số hành khách tính theo tỷ lệ trong mỗi class
- d) Vẽ biểu đồ Barplot thể hiện tuổi trung bình theo Nam/Nữ
- e) Vẽ biểu đồ Barplot thể hiện tuổi trung bình theo class
- f) Vẽ biểu đồ Barplot thể hiện tổng số hành khách và số chết trong mỗi class
- g) vẽ biểu đồ StackedBar thể hiện tổng số hành khách Nam/Nữ theo mỗi class

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
# Câu 1:
# a) Đọc dữ liệu từ tập tin train.csv và lưu vào biến titanic.
titanic = pd.read_csv(r'data\train.csv', sep = ',')
# Hiển thị 5 dòng dữ liệu đầu của titanic
titanic.head()
```

Out[2]:

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

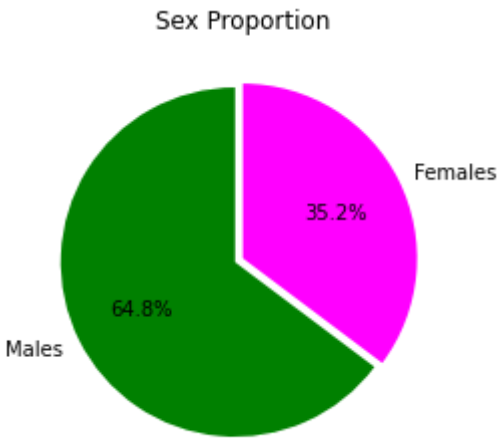
In [3]:
```
# b) Thiết lập cột index cho titanic là PassengerId.
# Hiển thị lại 5 dòng dữ liệu đầu của titanic lúc này.
titanic.set_index('PassengerId', inplace=True)
```

In [4]:
```
1 titanic.head()
```

Out[4]:

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

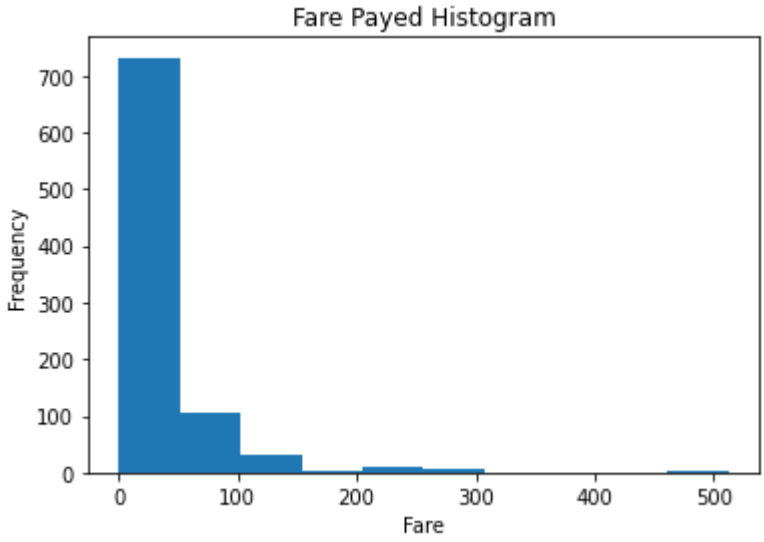
In [5]:
```
# Câu 2: Tạo pie chart thể hiện tỷ lệ hành khách nam/nữ trên tàu.
'''
Gợi ý:
- Tạo biến males, female là tổng nam và tổng nữ.
- Tạo biến proportions là list có 2 phần tử là male và female
- Vẽ biểu đồ: với dữ liệu là proportions, nhãn là ['Males', 'Female'],
    màu là ['green', 'magenta']
- Thiết lập title là Sex Proportion
'''

# Tạo biến males, female là tổng nam và tổng nữ
males = (titanic['Sex'] == 'male').sum()
females = (titanic['Sex'] == 'female').sum()

# Tạo biến proportions là list có 2 phần tử là male và female
proportions = [males, females]

# Vẽ biểu đồ
plt.pie(proportions, labels = ['Males', 'Females'], shadow = False,
        colors = ['green','magenta'],
        explode = (0.05 , 0), startangle = 90, autopct = '%1.1f%%')

plt.axis('off')

# Thiết lập title là Sex Proportion
plt.title("Sex Proportion")

# Show the plot
plt.show()
```



In [6]:
```
# Câu 3: Cho biết có bao nhiêu người còn sống sót

```

In [7]:
```
# Câu 4a: Vẽ biểu đồ histogram của cột vé (Fare),
# bổ sung các thông tin xlabel, ylabel, title
# Bạn nhận xét gì về biểu đồ vừa vẽ

```

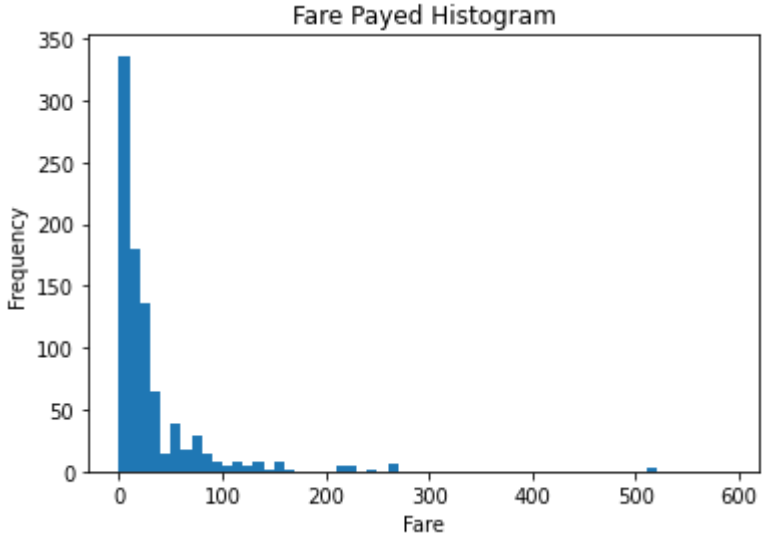
Nhấn vào đây để xem kết quả!



In [8]:

```
# Câu 4b: Vẽ biểu đồ histogram của cột vé (Fare)
# Bạn nhận xét gì về biểu đồ vừa vẽ
'''
Gợi ý:
- Tạo binsVal = np.arange(0,600,10)
- Vẽ histogram với dữ liệu là titanic, bins = binsVal.
- Bổ sung các thông tin xlabel, ylabel, title
'''
# Tạo binsVal = np.arange(0,600,10)
binsVal = np.arange(0,600,10)

```

Nhấn vào đây để xem kết quả!



In [9]:

```
# Câu 5a: xem thông tin mô tả (describe) của biến Fare và vẽ biểu đồ Boxplot

```

In [10]:

```
# Câu 5b: vẽ biểu đồ Barplot thể hiện tổng số hành khách trong mỗi class

```

In [11]:

```
# Câu 5c: vẽ biểu đồ Barplot
# thể hiện tổng số hành khách tính theo tỷ lệ trong mỗi class

```

In [12]:
```
# Câu 5d: vẽ biểu đồ Barplot thể hiện tuổi trung bình theo Nam/Nữ

```

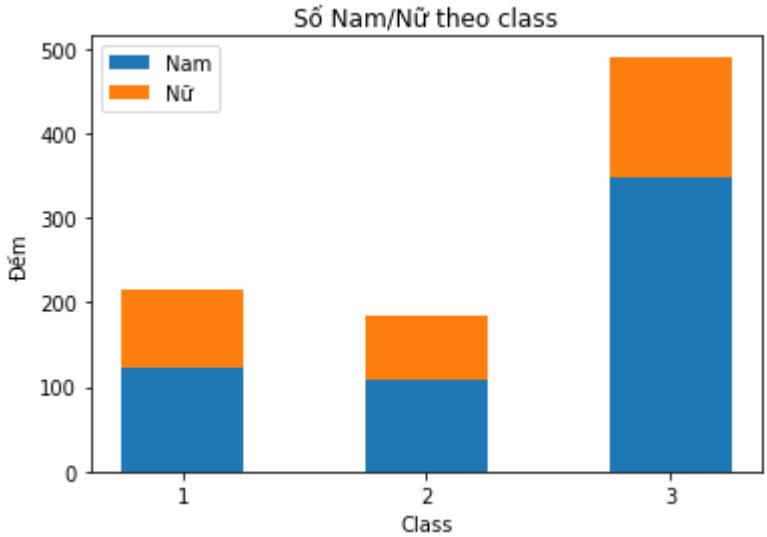
In [13]:
```
# Câu 5e: vẽ biểu đồ Barplot thể hiện tuổi trung bình theo class

```

In [14]:
```
# Câu 5f: vẽ biểu đồ Barplot thể hiện tổng số hành khách và số chết trong mỗi class

```

Nhấn vào đây để xem kết quả!

In [15]:
```
# Câu 5g: vẽ biểu đồ StackedBar thể hiện tổng số hành khách Nam/Nữ theo mỗi class

```

Nhấn vào đây để xem kết quả!



In [ ]:
```
1
```