Use python to examine the ‘Titanic’ dataset.

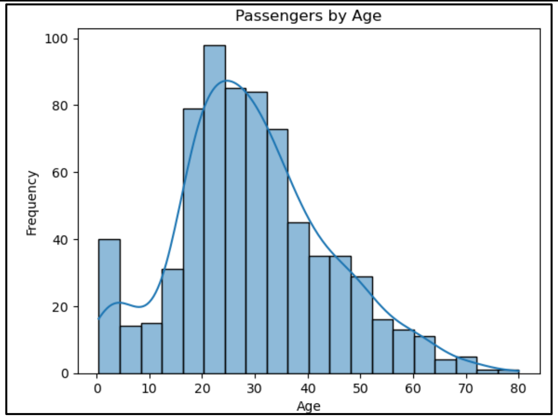
Code	Explanation

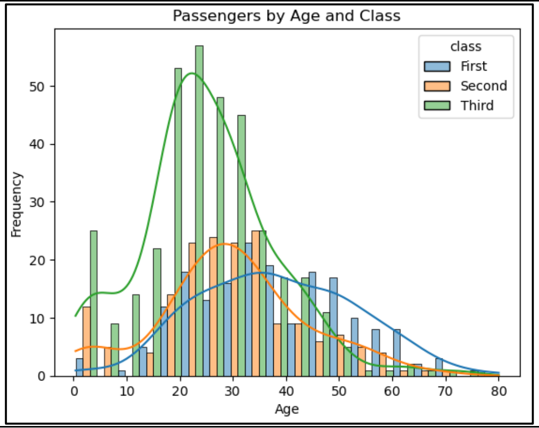
import pandas as pd
import seaborn as sns
titanic_df = sns.load_dataset("titanic")
titanic_df.head()
	The Titanic dataset has been loaded and stored as a dataframe.

# a. Calculate the proportion of passengers embarked at Southampton 	
total_passengers = titanic_df['embark_town'].count()
embark_soton = (titanic_df['embark_town'] == 'Southampton').sum()
southampton_proportion = embark_soton/total_passengers
	To determine the proportion of passengers who embarked from Southampton, the count of passengers boarding in Southampton is divided by the total number of passengers, as indicated by the overall count in the "embark_town" column.

southampton_proportion 0.7244094488188977	The outcome reveals that 72.44% of passengers boarded from Southampton.

#b. Plot and describe the distribution of passengers by age. Did this vary by the class of ticket?
# age frequency histogram 	
import matplotlib.pyplot as plt
sns.histplot(data=titanic_df, x='age', kde=True)
plt.title('Passengers by Age')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
 	
# age frequency histogram divided in class 	
sns.histplot(data=titanic_df, x='age', hue='class', kde=True,multiple="dodge")
plt.title('Passengers by Age and Class')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
	Two histogram plots are created in this section: one displaying the frequency distribution of passenger ages and the second providing a more detailed breakdown by also dividing passengers by class.

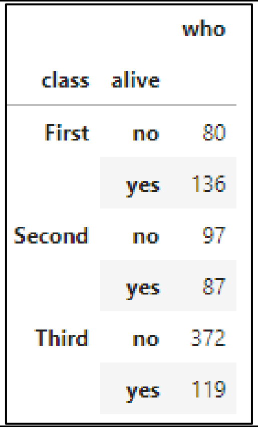




Upon examining the graphs, it is evident that the majority of passengers fall within their 20s and 30s. Additionally, in the second graphs we can observe that the second and third class passage relatively younger than the first class.

```
#c. Were first class passengers more likely to survive?
titanic_df.groupby(by=['class','alive'])[['who']].count()
```

Passanger data grouped by the class and survival status

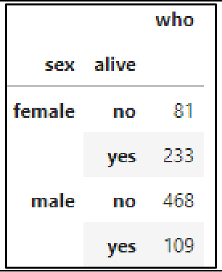


```
                who
class   alive
First   no       80
        yes     136
Second  no       97
        yes      87
Third   no      372
        yes     119
```

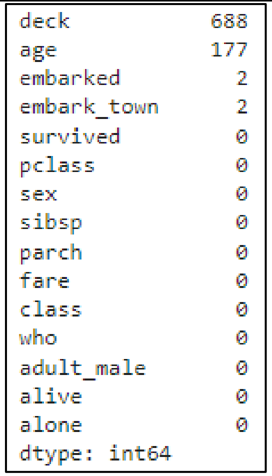
The analysis shows that first-class passengers are more likely to survive, as the number of survivors in this class exceeds the number of non-survivors.

```
#d. Were males or females more likely to survive?
titanic_df.groupby(by=['sex','alive'])[['who']].count()
```

For this question, passenger data grouped by the sex and alive status

 who sex alive female no 81 yes 233 male no 468 yes 109	The findings indicate that females are more likely to survive, as the count of female survivors surpasses that of non-survivors. Conversely, among males, the number of non-survivors is higher than the count of survivors.

#e. For which variables in the dataset are data missing? How might this affect your answers to the questions above?
pd.isnull(titanic_df).sum().sort_values(ascending=False)
	A calculation for missing values is performed before sorting it from highest to lowest.

 deck 688 age 177 embarked 2 embark_town 2 survived 0 pclass 0 sex 0 sibsp 0 parch 0 fare 0 class 0 who 0 adult_male 0 alive 0 alone 0 dtype: int64	A calculation for missing values is performed before sorting it from highest to lowest. Notably, there are several missing values in the dataset, including 688 instances where deck information is absent, 177 cases with missing passenger age information, and 2 instances lacking information on the embarked town. It is essential to note that the missing information regarding the embarked town could influence the accuracy of the answer to the first question. Similarly, the absence of passenger age information may impact the reliability of the answer to the second question.