In [1]: ▶| ```
            # use the titanic dataset to perform crosstabulation and explain your finding
         ```

In [4]:  ▶|  ```
            import pandas as pd

            titanic=pd.read_csv('titanic_train.csv')
            titanic.head()
         ```

Out[4]:

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

Groupby is used to separate a dataset into groups based on different selection criteria. It is important to note that groupby is used by categorical and continuous data

In [5]: ▶ | `# Therefore first check for cardinality of the different columns`

`titanic.nunique()`

Out[5]:
```
PassengerId    891
Survived         2
Pclass           3
Name           891
Sex              2
Age             88
SibSp            7
Parch            7
Ticket         681
Fare           248
Cabin          147
Embarked         3
dtype: int64
```

In [6]: ▶ | `titanic.groupby('Pclass').Fare.mean()`

Out[6]:
```
Pclass
1    84.154687
2    20.662183
3    13.675550
Name: Fare, dtype: float64
```

The data above shows that the Pclass 1 was the leading class in the fare collections made as it has a mean of 84.154687 followed by Pclass 2 which has a mean of 20.662183 and lastly Pclass 3 which has a fare mean of 13.675550

In [9]: ▶ | `titanic.groupby('Sex').Age.mean()`

Out[9]:
```
Sex
female    27.915709
male      30.726645
Name: Age, dtype: float64
```

The data above shows that the mean age of the male gender that was on the titanic was 30.726645 while that of the female counterparts was 27.915709

In [10]: ▶| `titanic.groupby('Survived').Age.mean()`

Out[10]:
```
Survived
0    30.626179
1    28.343690
Name: Age, dtype: float64
```

The data above shows that those who survived had the mean age of 28.343690 while those who did not survive had the mean age of 30.626179. This means that there is a slight difference between the ages of the passengers who survived and those who did not, however those that did not survive have a higher mean age

In [24]: ▶|
```
import warnings
warnings.filterwarnings('ignore')
```

In [28]: ▶| `titanic.groupby('Sex')['Fare', 'Age'].agg(['mean','count','median','std','max`

Out[28]:

Sex	Fare					Age			
	mean	count	median	std	max	mean	count	median	std
female	44.479818	314	23.0	57.997698	512.3292	27.915709	261	27.0	14.110146
male	25.523893	577	10.5	43.138263	512.3292	30.726645	453	29.0	14.678201

The data above shows the mean, count, median, standard deviation and the maximum value of both fare and age based on Sex. The fare count for the females was 314 while that of the males was 577. the median of the fare paid by the females was 23.0 while that of the male was 10.5, the mean fare that the females paid was 44.479818 while that that the males paid was 25.523893. The mean age of the females was 25.523893 while that of the males is 30.72664. the counts of the males was 453 while that of the females was 261. The median age of the females was 27.0 while that of the males was 29.0. the

standard deviation in age for the females was 14.110146 while that of the male was 14.678201. the maximun age recorded from the female passengers was 63.0 while that of the male was 80.0.

In [12]: ▶| `# Below is the different ways that one can use the statistical functions whic`

In [14]: ▶| `titanic.Age.mean()`

Out[14]: 29.69911764705882

The data above concludes that the mean age of the passengers that boarded the titanic was 29.69911764705882

In [16]: ▶| `titanic.Pclass.mode()`

Out[16]:
```
0    3
Name: Pclass, dtype: int64
```

The data above shows that most of the passengers used the Pclass number 3 as compared to all the other classes as mode refers to the value that is most occuring

Cross Tabulation is used to analyze categorical data hence into groups based on different criteria

In [32]: ▶| `titanic.nunique()`

Out[32]:
```
PassengerId    891
Survived         2
Pclass           3
Name           891
Sex              2
Age             88
SibSp            7
Parch            7
Ticket         681
Fare           248
Cabin          147
Embarked         3
dtype: int64
```

In [36]: ▶| `import pandas as pd`

In [39]: ▶| `pd.crosstab(titanic.Survived, titanic.Parch)`

Out[39]:

Parch	0	1	2	3	4	5	6
Survived							
0	445	53	40	2	4	4	1
1	233	65	40	3	0	1	0

The number of ParentChildren groups were seperated in groups of 0,1,2,3,4,5,6 where the ones that did not survive were 445,53,40,2,4,4,1 respectively and on the other hand, those that did not survive were 233,65,40,3,0,1,0 respectively

In [40]: ▶| `pd.crosstab(titanic.Pclass, titanic.Embarked)`

Out[40]:

Embarked	C	Q	S
Pclass			
1	85	2	127
2	17	3	164
3	66	72	353

Embarked represents the town where the traveller mounted from which was three

destinations namely the Southampton, Cherbourg, and Queenstown. the majority of the passengers who were in the Pclass number one mounted the titanic in Southampton (127), followed by Cherbourg (85) and lastly Queenstown (2). the majority of the passengers is the Pclass number 2 mounted the titanic at Southampton (164), followed by those that mounted at Cherbourg (17) and lastly Queenstown (3) produced the least number of passengers from its town. latly there is the Pclass number 3 which had 353 passengers from Southampton which was leading followed by 66 passengers who mounted at Cherbourg and finally Queenstown that produced 72 passengers. Based on the above observations, the town of Southampton produced the highest number of passengers with a total of 644 followed by Cherbourg with a total of 168 passengers and lastly Queenstown with a total of 77 passengers ¶

In [ ]: ▶| [   ]