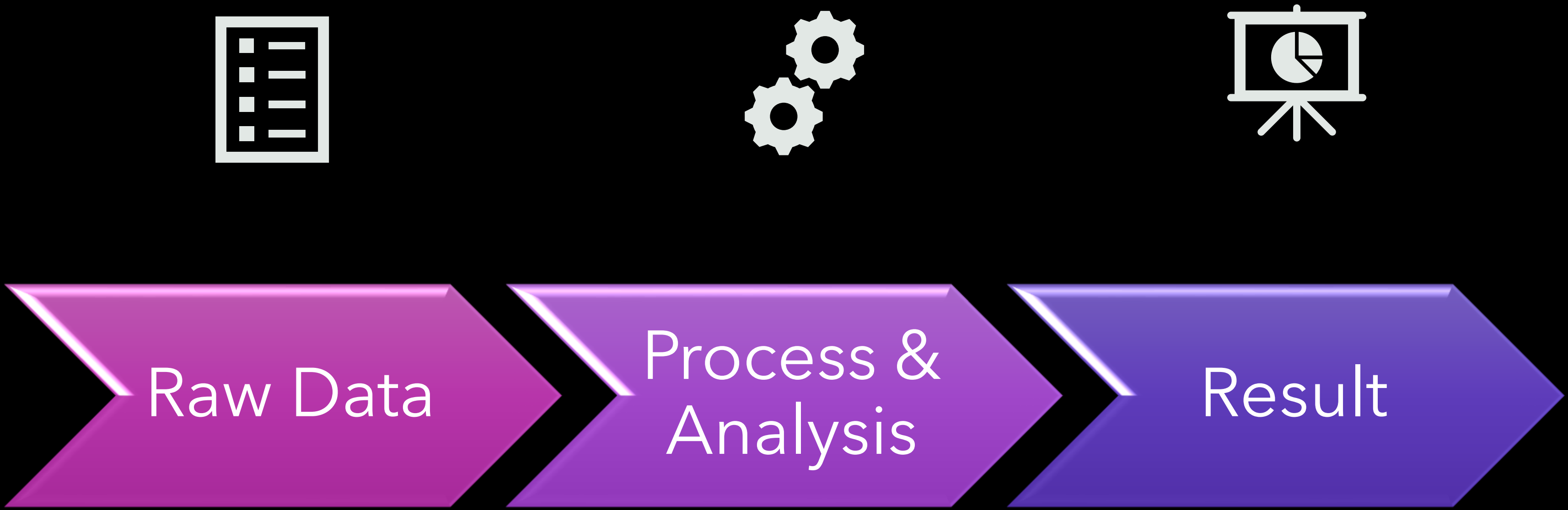
Case Study: Titanic

Shashank Maindola

Background

- The titanic causality is one of the worst sea causalities to happen in the history of mankind
- But there are lot of key information which can be evaluated
- We have a given data set to get any valuable information out of it
- Data Science tools were used for analyzing and filtering the given raw data set
- We have used Python based Anaconda distribution for evaluating and analysing the data

Methodology



Raw Data

Process & Analysis

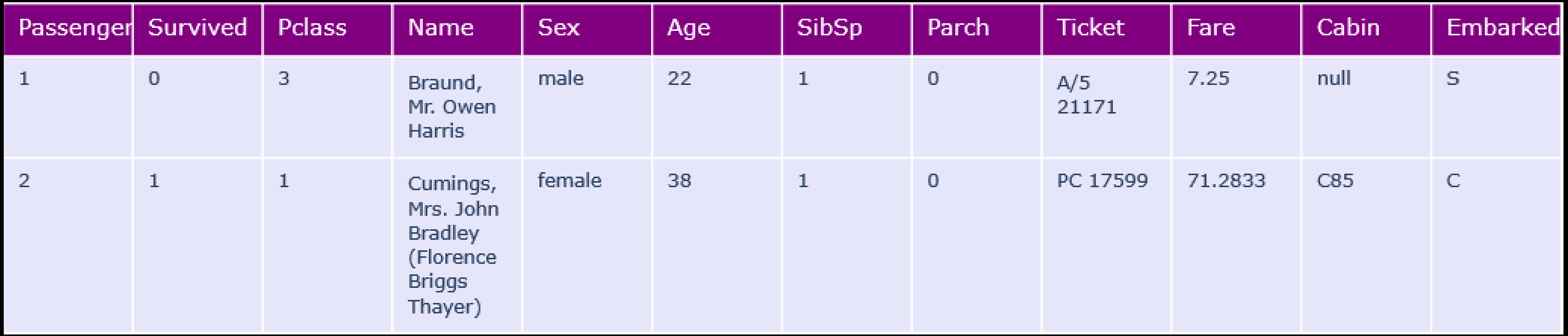
Result

Processing Raw Data

- Getting raw data
- Finding missing values
- Filtering/imputing the values accordingly

Getting the Data & Processing

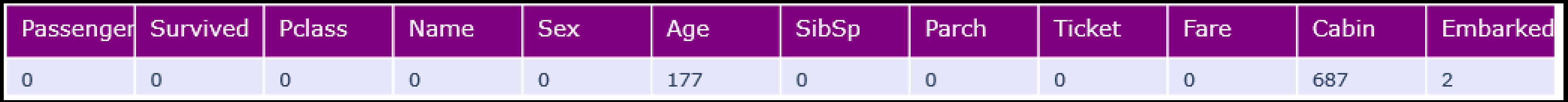
The data was imported from the 'kaggle' and was converted to a data frame



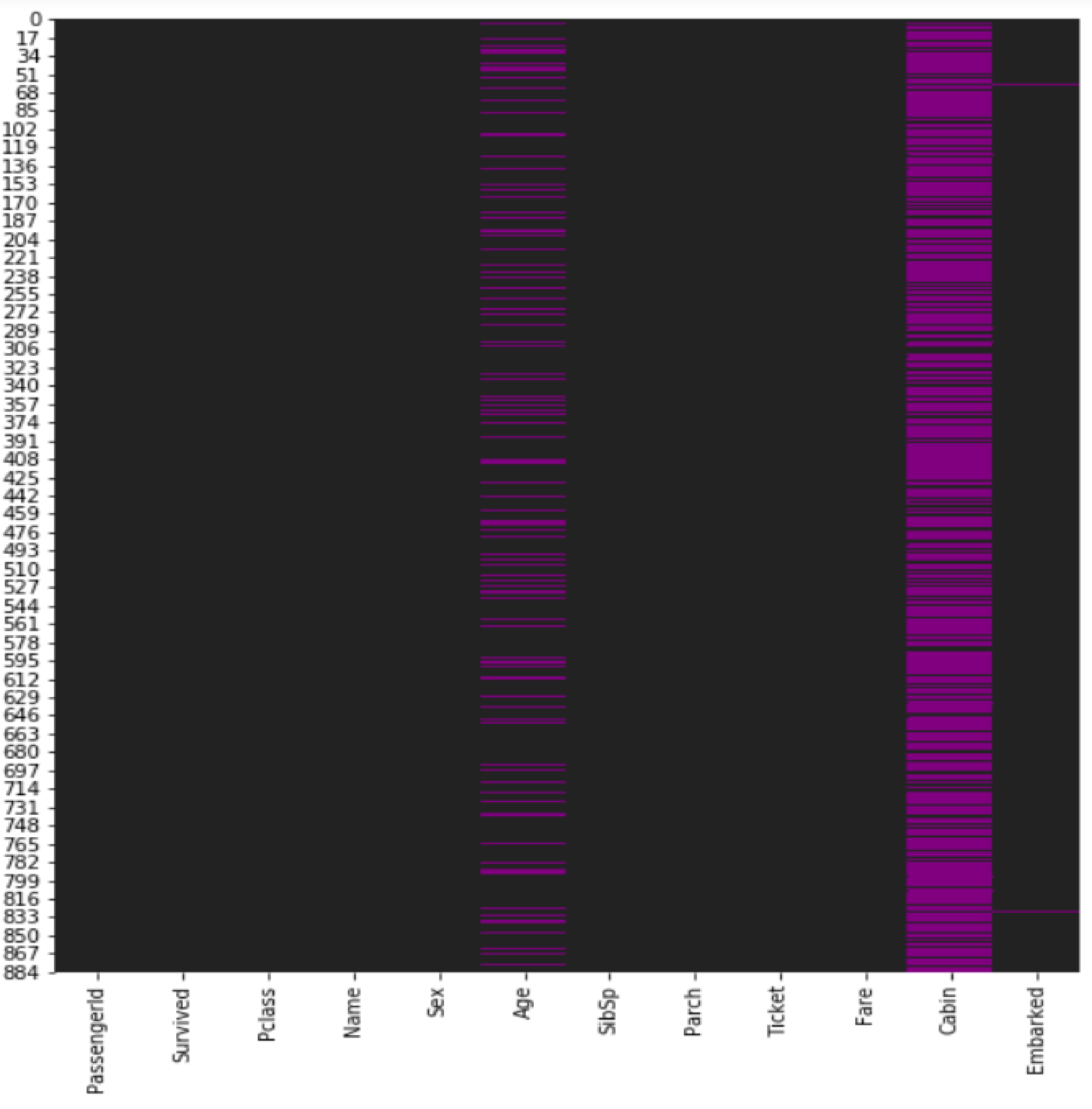
Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	null	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C

There was a need to find out the missing data which was to be counted

Output:



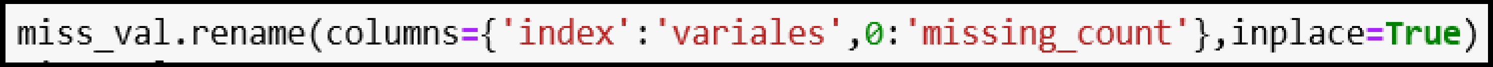
Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	0	0	0	0	177	0	0	0	0	687	2



▨ Missing values

Getting the Data & Processing(Cont.)

We renamed the column

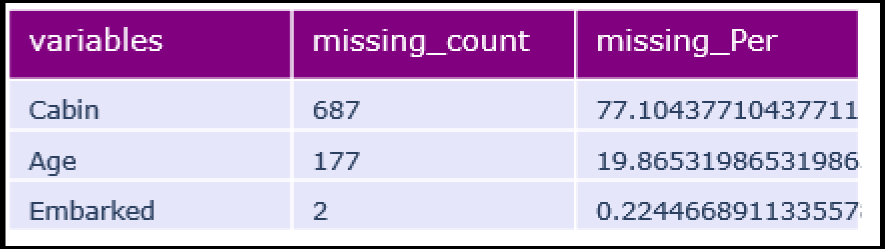


```
miss_val.rename(columns={'index':'variales',0:'missing_count'},inplace=True)
```

Calculated the percentage of missing data

Sorted the missing data

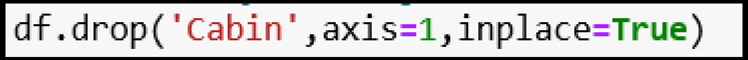
The result:



variables	missing_count	missing_Per
Cabin	687	77.10437710437711
Age	177	19.86531986531986
Embarked	2	0.224466891133557

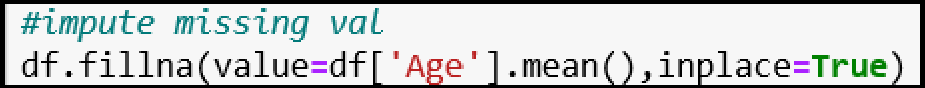
Treatment of missing values

Eliminated the values ‘Cabin’ as the 77% is missing



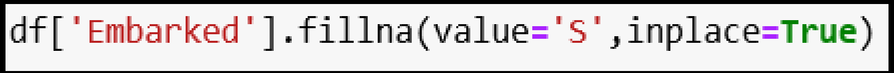
```
df.drop('Cabin',axis=1,inplace=True)
```

Almost 20% values related to age are missing so it needs to be treated as well, being a numeric value we imputed the missing values with the mean values



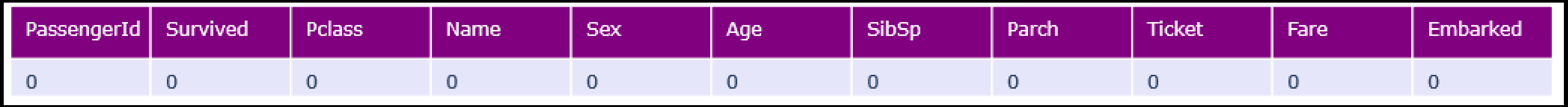
```
#impute missing val
df.fillna(value=df['Age'].mean(),inplace=True)
```

Finally, we see that embarked has 2 values missing we replaced them with



```
df['Embarked'].fillna(value='S',inplace=True)
```

The Result:



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	0	0	0	0	0	0	0	0	0	0

Analysis

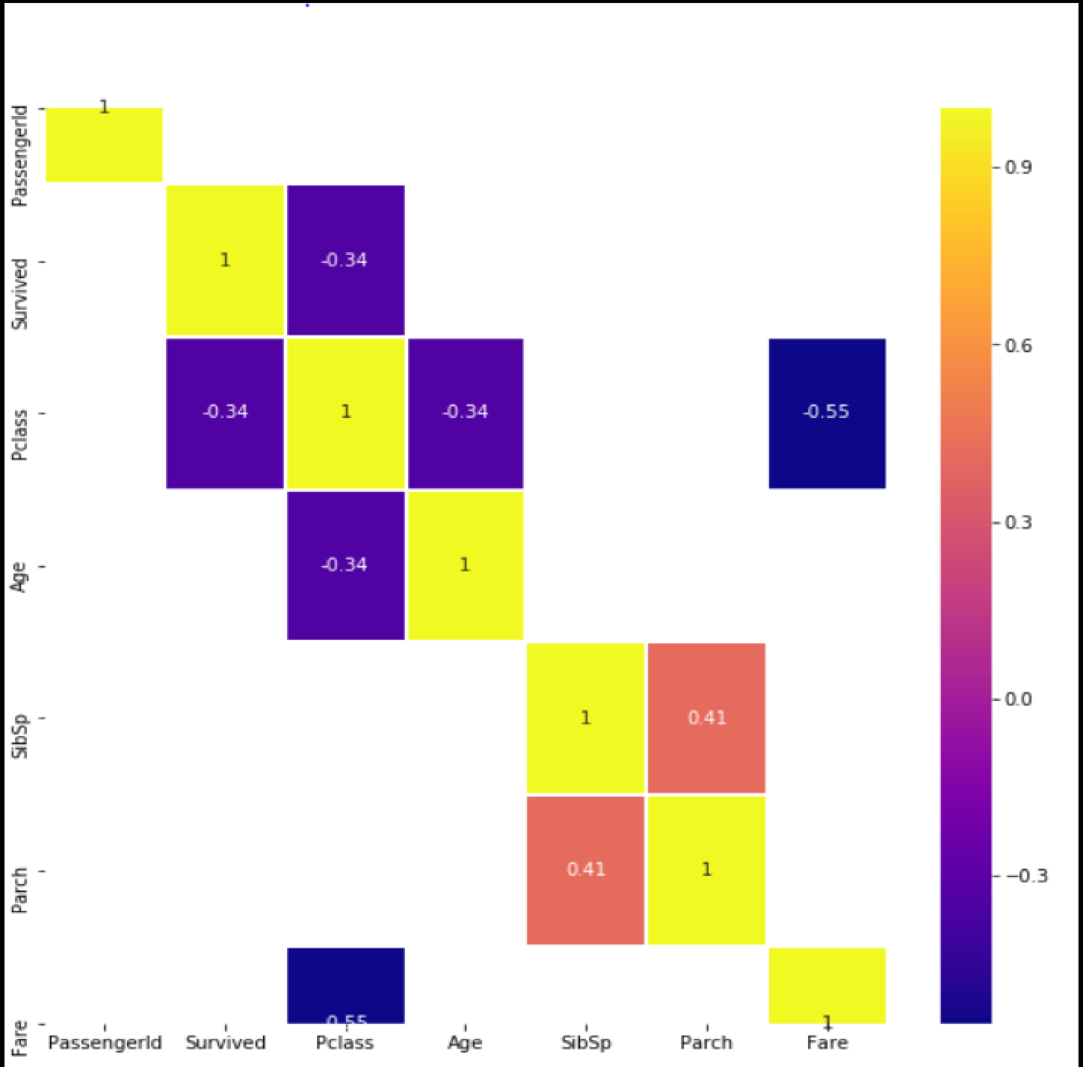
General Analysis

Business Analysis

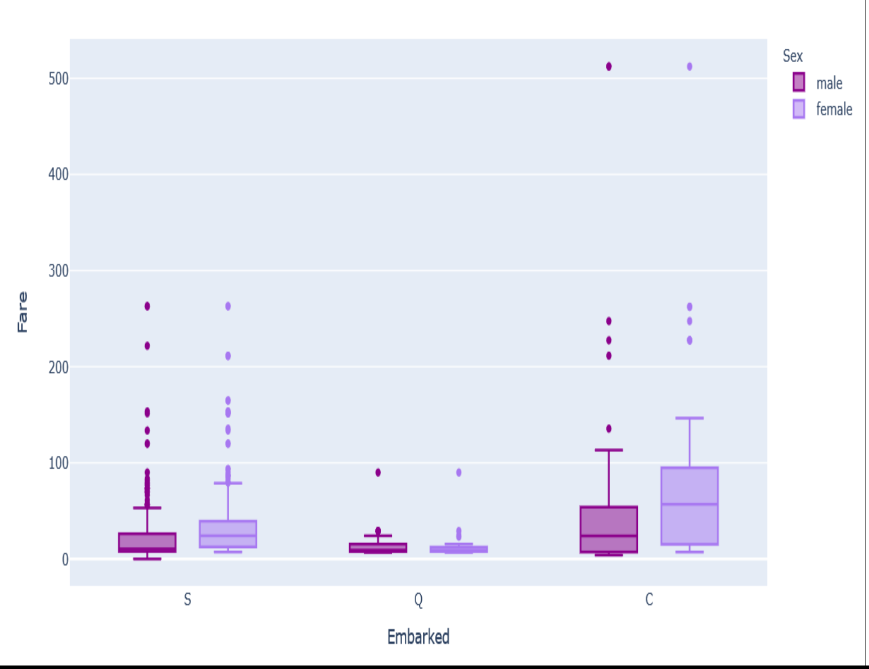
Survival Analysis

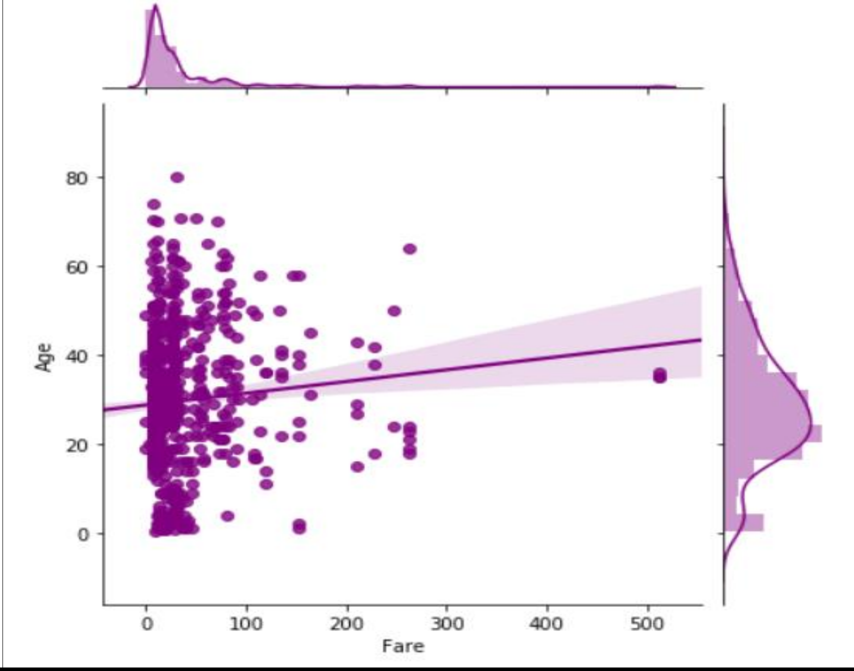
General Analysis

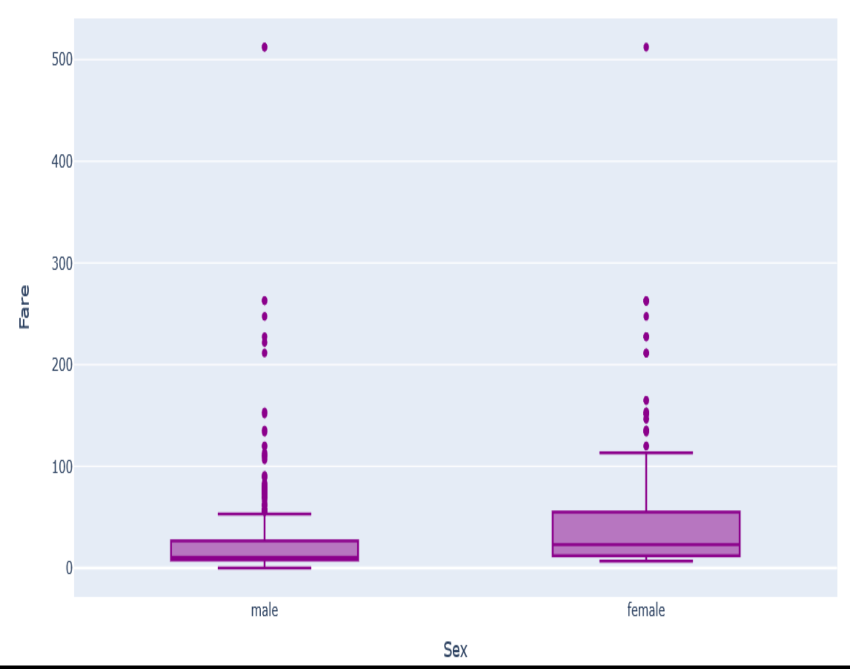
Correlation Plot



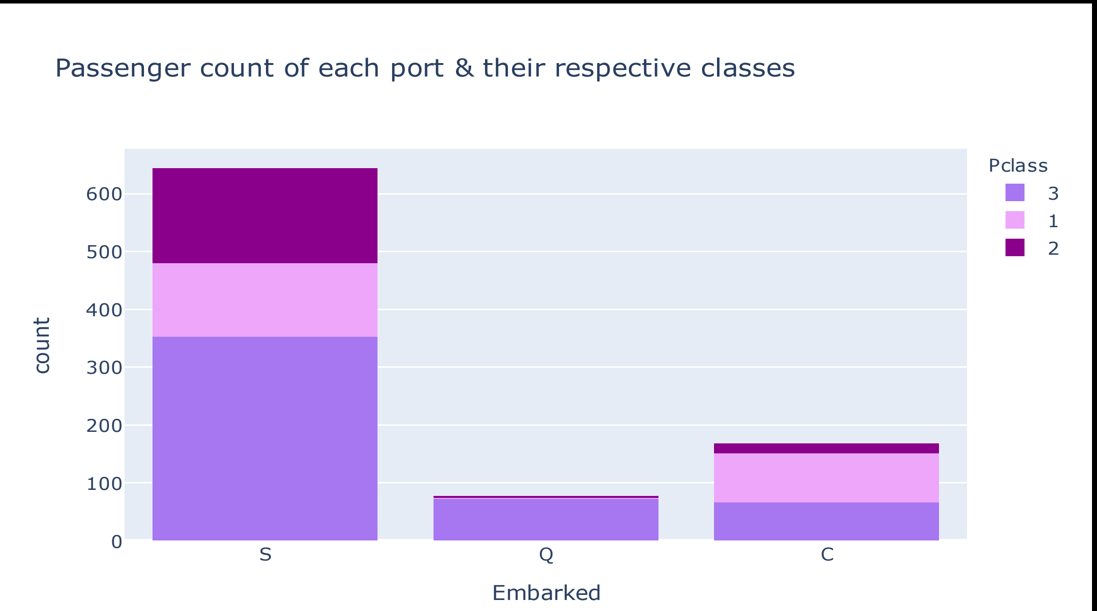
General Analysis(cont.)

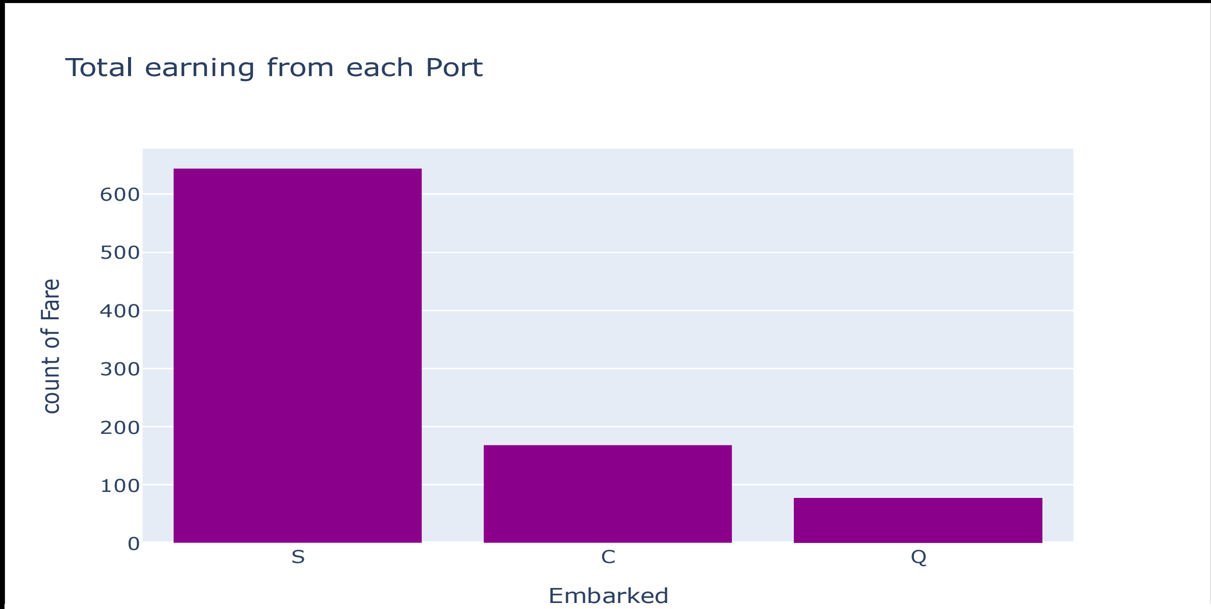


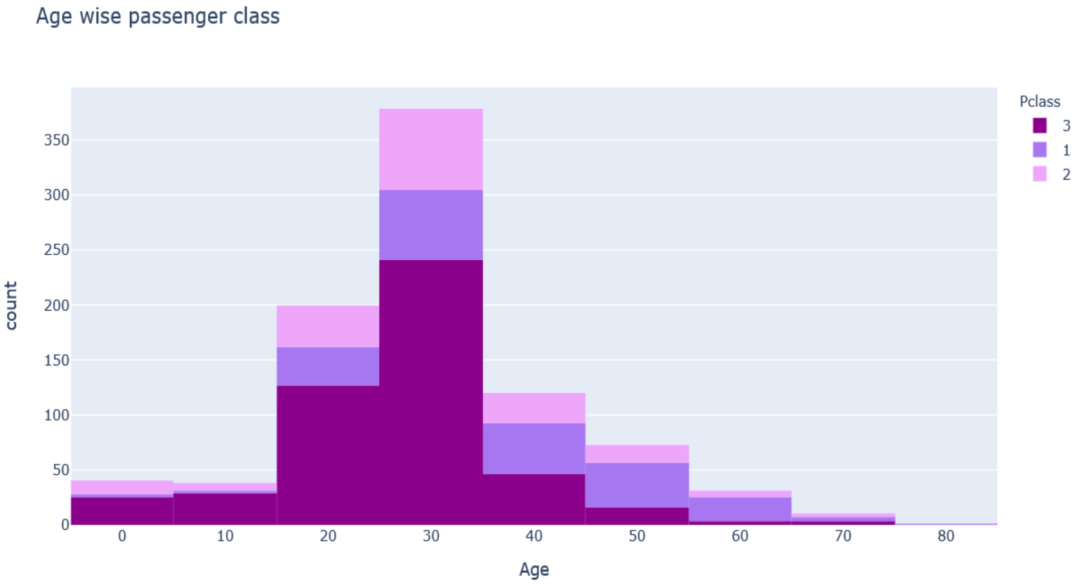


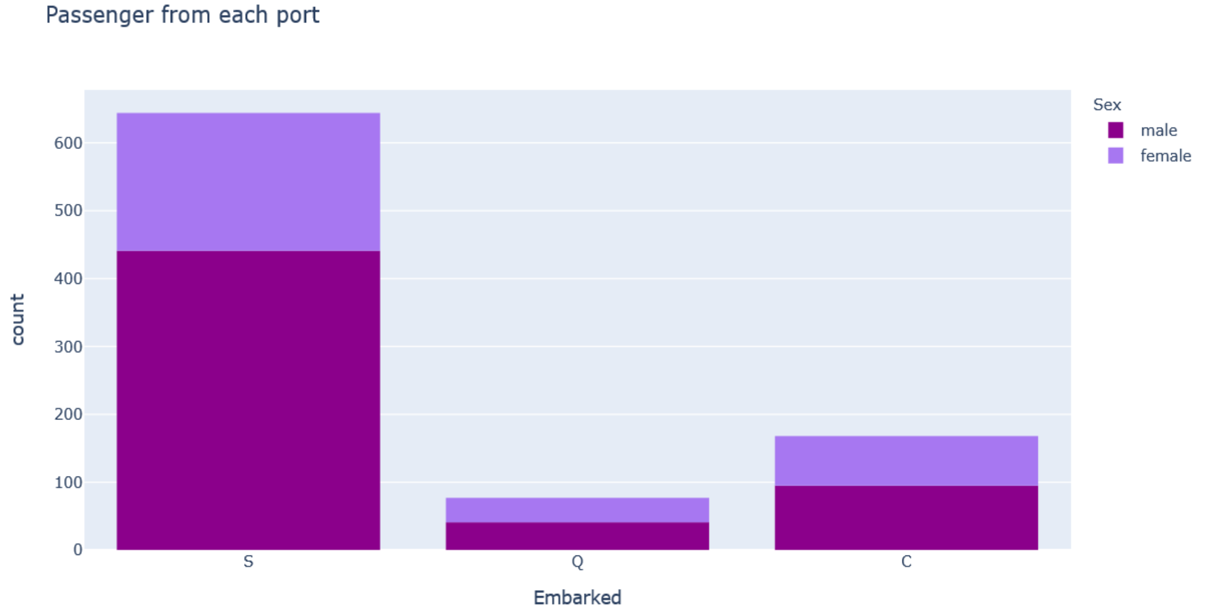


Business Analysis

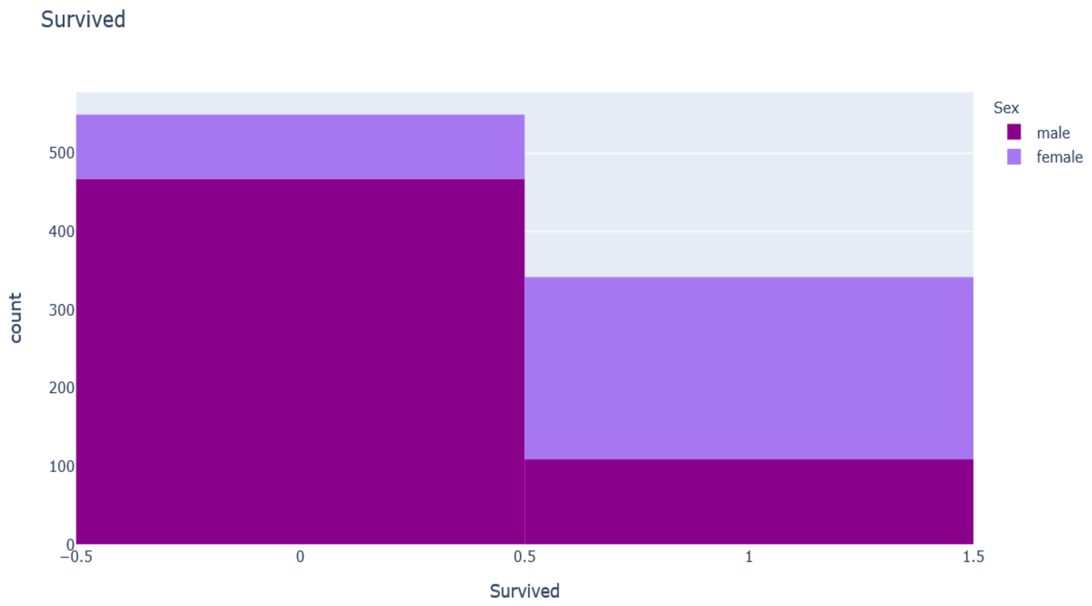


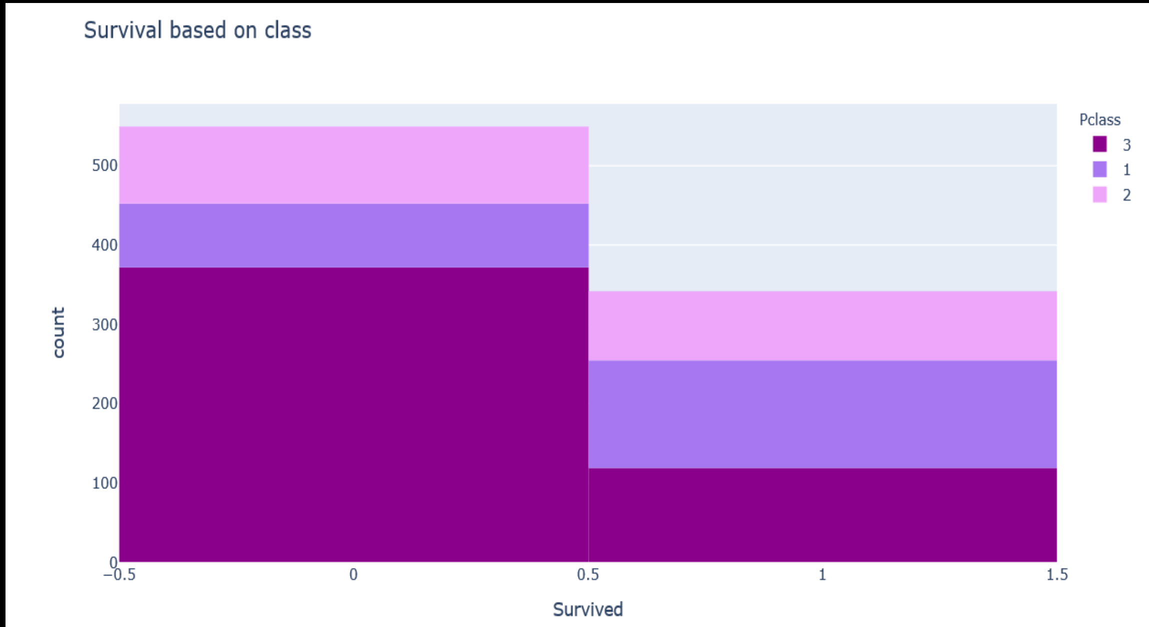


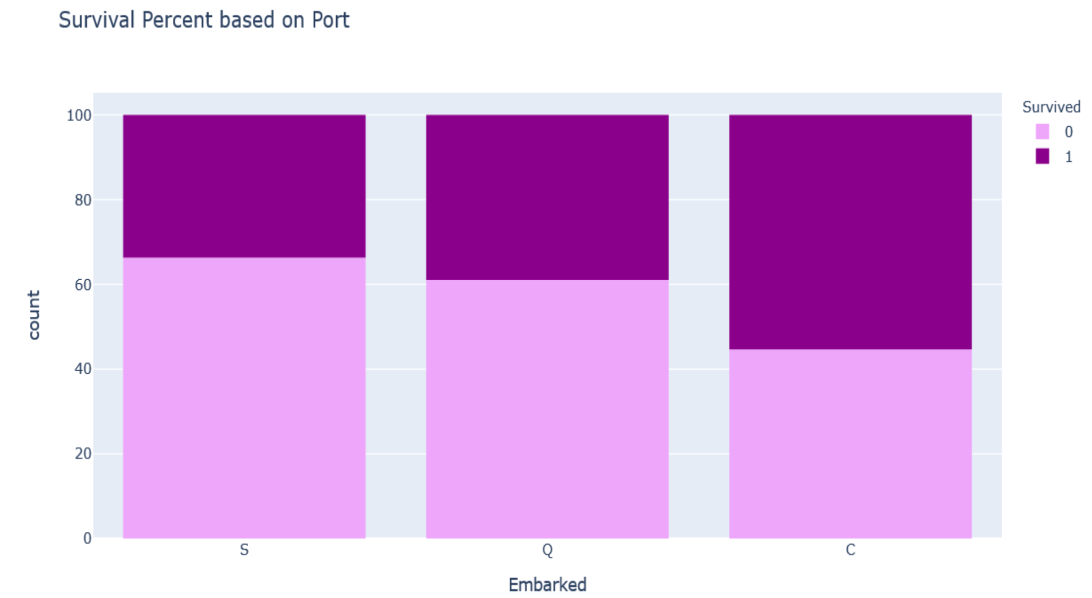


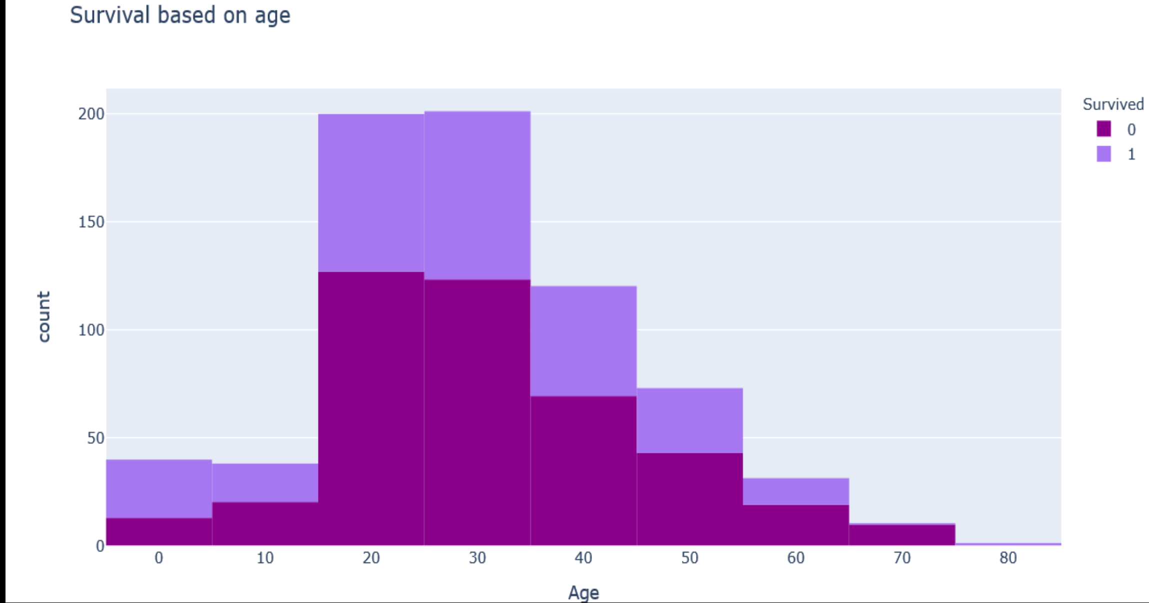


Survival Analysis


Survived


Survival based on class


Survival Percent based on Port


Survival based on age

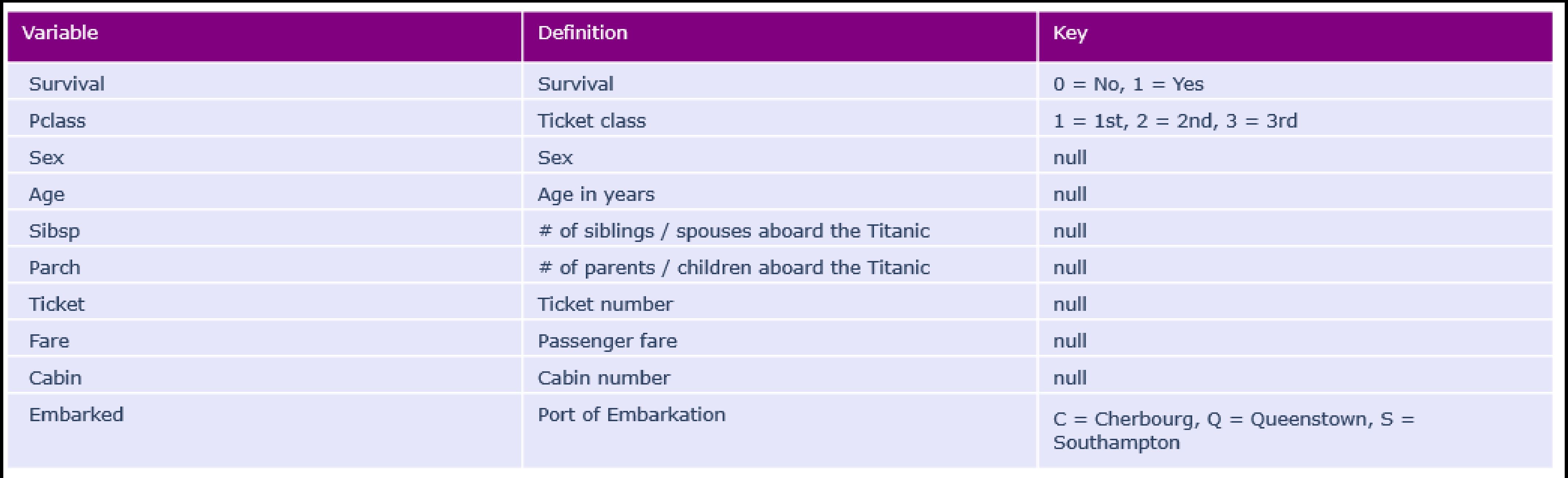
Key Outcomes

- Generally females passengers paid more fare than males may be due to the fact the society was not open to the females so in order to travel they might had to pay more amount
- Chances of survival are directly related to the passenger class as the passengers were more privileged

- Most of the people boarded from Southampton and maximum paid for the 3rd class, so for starting a budget liner, Southampton is the best port
- Most of the passengers were youngsters around 30's and single so if someone wants to start a business then they should keep youngsters in mind
- If someone wants to start a luxury liner then Cherbourg is the best port

- Most of the surviving passengers were females as they were allowed to board the lifeboats first
- The survival percentage of people from Cherbourg was highest as the people paid more fare and had more privileges
- Survival percentage of the 1st class passengers was highest , as they were having more privileges
- Survival of youngsters was highest as they could survive for more time in icy cold conditions

Glossary



Variable	Definition	Key
Survival	Survival	0 = No, 1 = Yes
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	Sex	null
Age	Age in years	null
Sibsp	# of siblings / spouses aboard the Titanic	null
Parch	# of parents / children aboard the Titanic	null
Ticket	Ticket number	null
Fare	Passenger fare	null
Cabin	Cabin number	null
Embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton