DATA EXPLORATION

The train dataset consists of 712 individual row entries and 12 columns representing the features. The columns are:

1. PassengerId: It has 712 non-null values and is an integer data type.
2. Pclass: It has 712 non-null values and is an integer data type.
3. Name: It has 712 non-null values and is an object data type.
4. Sex: It has 712 non-null values and is an object data type.
5. Age: It 575 non-null values and is a float data type.
6. SibSp: It has 712 non-null values and is an integer data type.
7. Parch: It has 712 non-null values and is an integer.
8. Ticket: It has 712 non-null values and is an object data type.
9. Fare: It has 712 non-null values and is a float data type.
10. Cabin: It has 160 non-null values and is an object data type.
11. Embarked: It has 710 non-null values and is an object data type.
12. Survived: It has 712 non-null values and is an integer.

There were two (2) unique values for Sex which is a categorical data. The values were “male” and “female”.

There were also (3) unique values for Embarked. The values were “S”, “C” and “Q”.

The statistics for the numerical data can be found in the image below

	PassengerId	Pclass	Age	SibSp	Parch	Fare	Survived
count	712.000000	712.000000	575.000000	712.000000	712.000000	712.000000	712.000000
mean	444.405899	2.308989	29.807687	0.492978	0.390449	31.819826	0.383427
std	257.465527	0.833563	14.485211	1.060720	0.838134	48.059104	0.486563
min	1.000000	1.000000	0.420000	0.000000	0.000000	0.000000	0.000000
25%	222.750000	2.000000	21.000000	0.000000	0.000000	7.895800	0.000000
50%	439.500000	3.000000	28.500000	0.000000	0.000000	14.454200	0.000000
75%	667.250000	3.000000	39.000000	1.000000	0.000000	31.000000	1.000000
max	891.000000	3.000000	80.000000	8.000000	6.000000	512.329200	1.000000

The count represents the total number of counts for the columns.

The mean represents the mean value.

The std represents the standard deviation.

The min represents the minimum value (i.e the smallest number).

The 25% represents the 25th percentile.

The 50% represents the median which is also known as the 50th percentile.

The 75% represents the 75th percentile.

The max represents the maximum value (i.e the largest number).

DATA CLEANING AND PREPROCESSING

Handling missing values:

There were 3 variables (Columns) which had missing values. The variables were Age, Cabin and Embarked.

The

The missing values in Age was replaced with the median because the distribution of this variable was skewed.

The missing values in Cabin was replaced with the modal value because it is a categorical variable, and the distribution also looked like a normal distribution.

The missing values in the Embarked variable was dropped. This was because the values represented a relatively small percentage of the dataset (0.28 %). The percentage is negligible so dropping it wont affect our model.