

Understanding the data in titanic.csv

The PDF report provides a comprehensive analysis of tabular data, focusing on various characteristics such as data types, number of rows, and number of columns. By examining these attributes, the report offers effective understanding of the dataset, enabling data analysts to gain insights and make informed decisions.

In addition to descriptive statistics, the PDF report utilizes graphical representations to visualize the distribution of numerical data and explores feature correlations. Through correlation plots, the relationships between different features are analyzed, revealing potential associations and dependencies within the dataset. These insights further enhance the understanding of the data and help identify key variables that may influence the target outcome.

The PDF report also investigates the distribution of categories for string features and provides class imbalance measures for classification scenarios. By assessing the balance of classes within the dataset, it highlights potential challenges in training models and making accurate predictions. This analysis is particularly valuable in machine learning tasks, as it helps to identify strategies for handling class imbalances and improving the performance of classification algorithms.

This version of PDF report has been generated from `0.0.6` version of `data-understand`.

Index

Chapter 1 - Dataset Characteristics

Chapter 2 - Visualize distributions of the dataset

Chapter 3 - Feature correlations between numerical features

Chapter 4 - Class Imbalance

data.underStand

Chapter 1 - Dataset Characteristics

In this section we report basic cardinality of the dataset like number of rows and number of columns. We report the data types of the columns in the dataset. Some columns are numeric, representing either integers or floating-point values. Other columns are categorical, containing string or object values. Additionally, there may be datetime columns capturing specific timestamps or dates.

We also report whether any column in the dataset has missing values. These missing values indicate instances where data is not available or was not recorded for certain records. Identifying and handling these missing values appropriately is crucial to ensure accurate analysis.

Furthermore, the nature of the target variable in the dataset is essential to determine the objective of analysis. If the target variable is categorical, it implies a classification problem, where the goal is to assign instances to specific categories or classes. On the other hand, if the target variable is numeric or continuous, it signifies a regression problem, where the focus lies in predicting a numeric value based on other variables.

Understanding these various aspects of the dataset lays the foundation for further exploration, analysis, and modeling tasks.

The number of rows in the dataset are: 1047

The number of columns in the dataset are: 6

The name of the target column is: target

The machine learning task based on your target column looks like: Binary Classification

No columns were found to have missing values

The table of data type for each column is below:-

Column	Type
embarked	object

Column	Type
sex	object
pclass	int64
age	float64
fare	float64
target	int64

data.understand

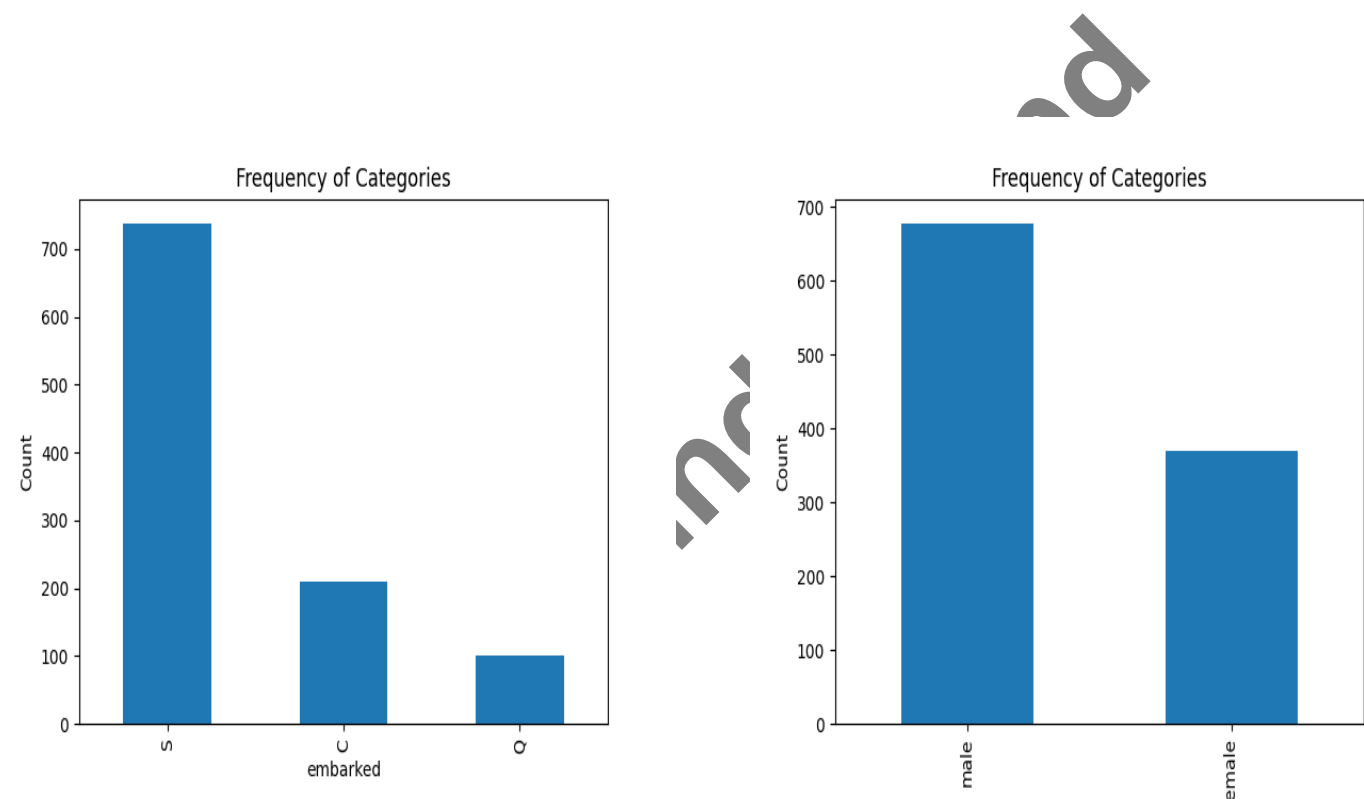
Chapter 2 - Visualize distributions of the dataset

This section have different graphs using which you can visualize distibutions of different features in your dataset, visualize the distibution of various categories for categorical features, visualize the histogram distribution of numerical features and visualize the box plot distribution between categories in categorical columns and numerical columns.

data.underStand

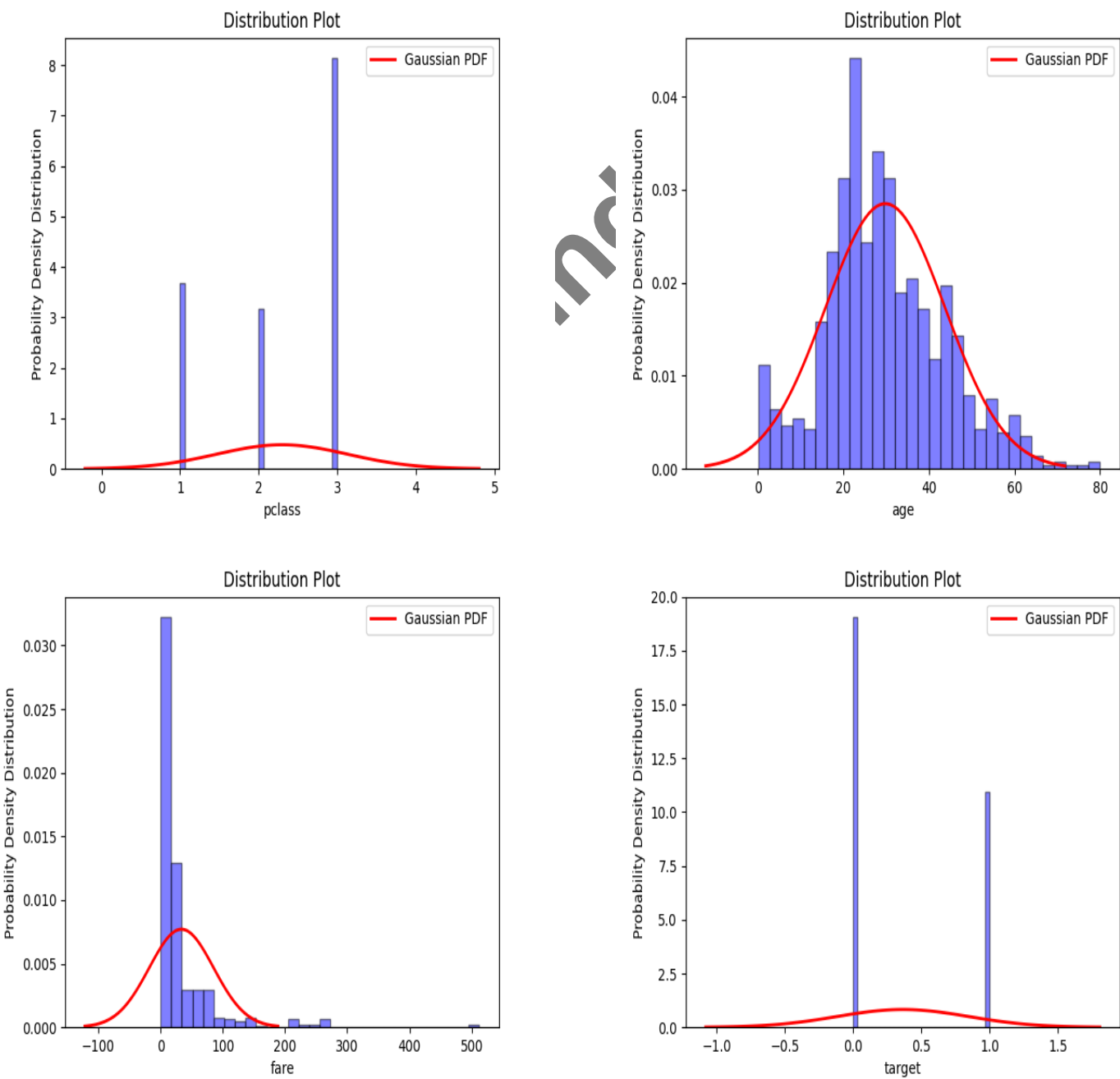
Categorical feature distribution

The section shows the distribution of individual categories in a given categorical column. The distribution helps to understand which categories in a given column are most/least prevalent in your dataset.



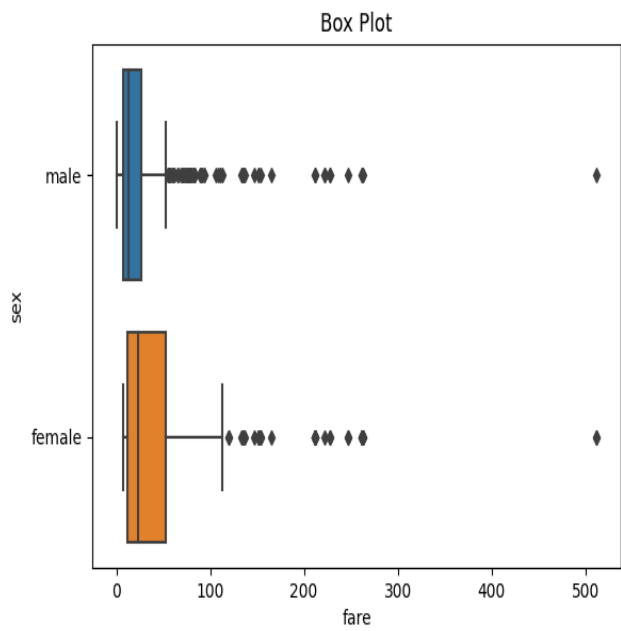
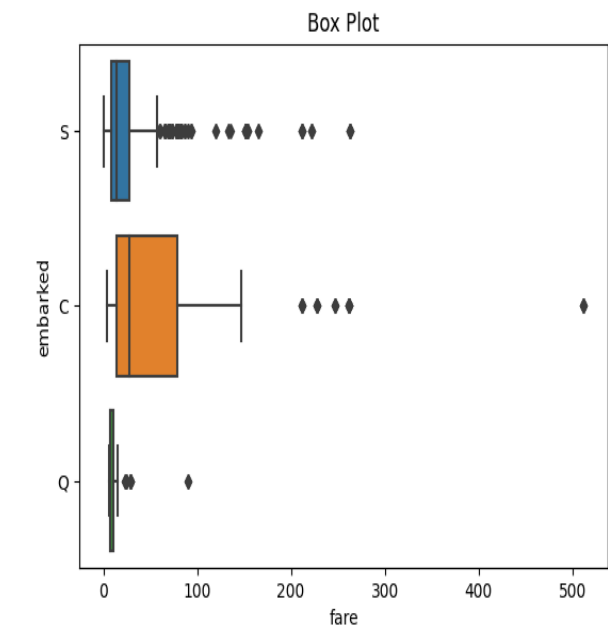
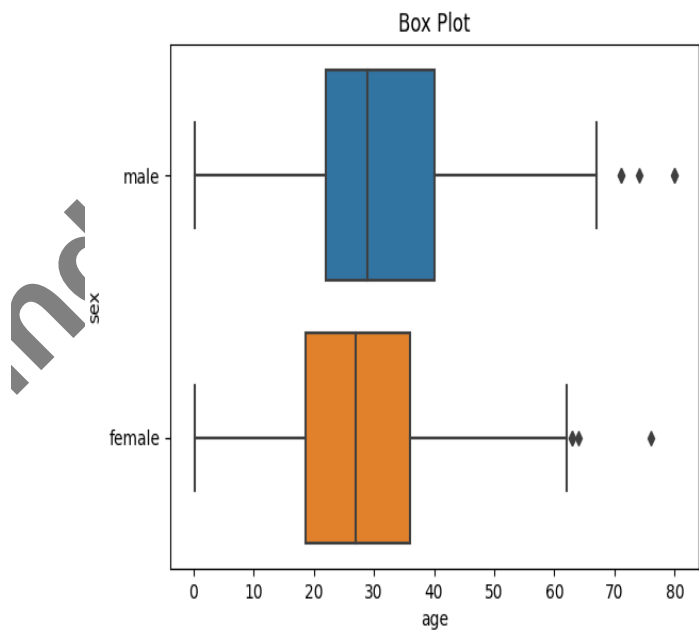
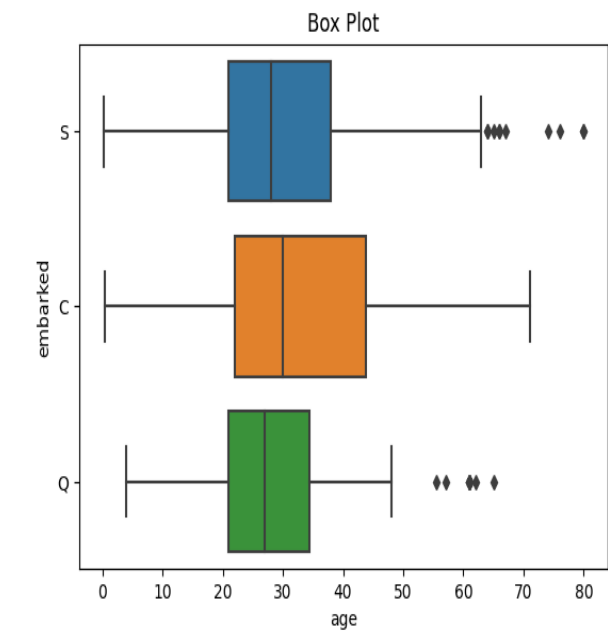
Numerical value distribution

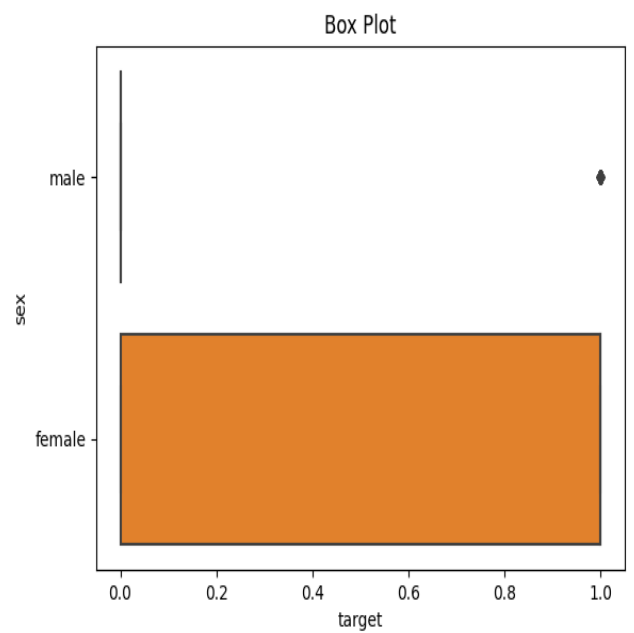
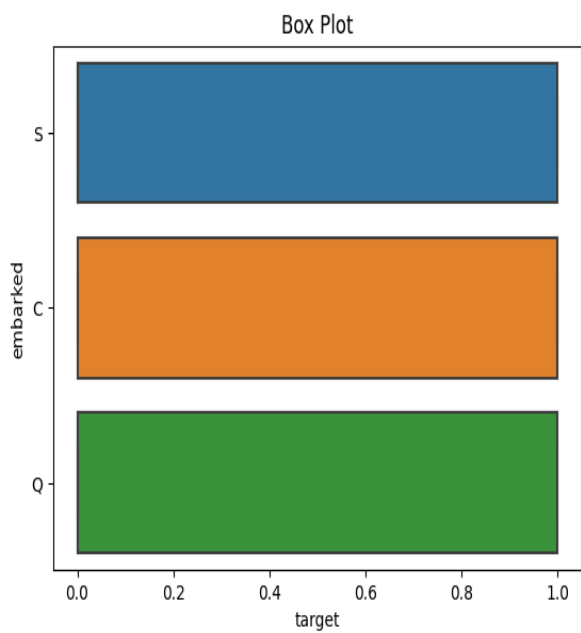
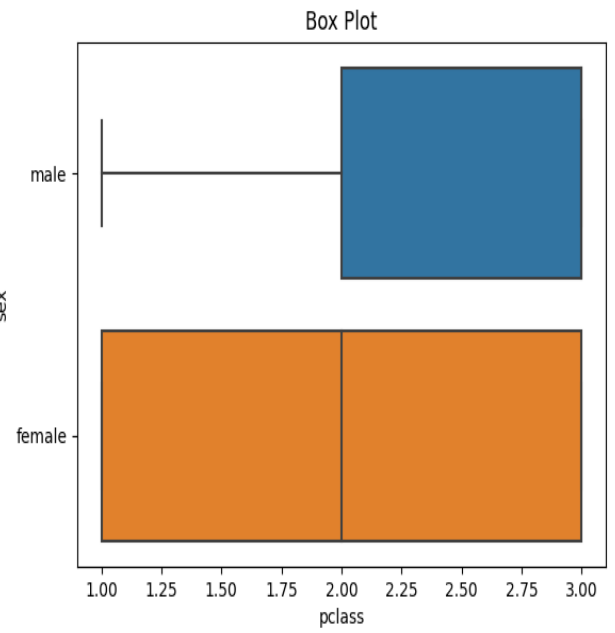
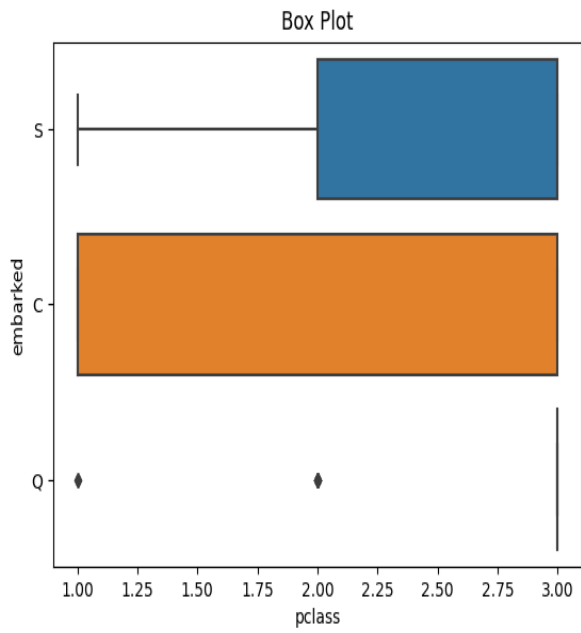
The section shows the histogram distribution of various numerical features in your dataset. The graphs also show a line chart which helps understand how the normal distribution will look if the numerical values in the distribution were normally distributed. These graphs also help gauge if the distribution of data in a particular column is skewed in any direction.



Box plot distribution

The section shows the box plot distribution of between the categories in categorical columns and numerical values in a numerical column. These graphs help in uncovering patterns that exist between various categories in a categorical column with the values in the numerical columns. The categorical columns having 15 categories or less are chosen for box plot distribution because the box plot visualization is not useful for larger number of categories





Chapter 3 - Feature correlations between numerical features

This section shows the numerical feature pairs having positive and negative correlation. The correlation have been computed using Pearson correlation coefficient. Examination of feature correlation can help find if the data has [leaky features]([https://en.wikipedia.org/wiki/Leakage_\(machine_learning\)](https://en.wikipedia.org/wiki/Leakage_(machine_learning))).

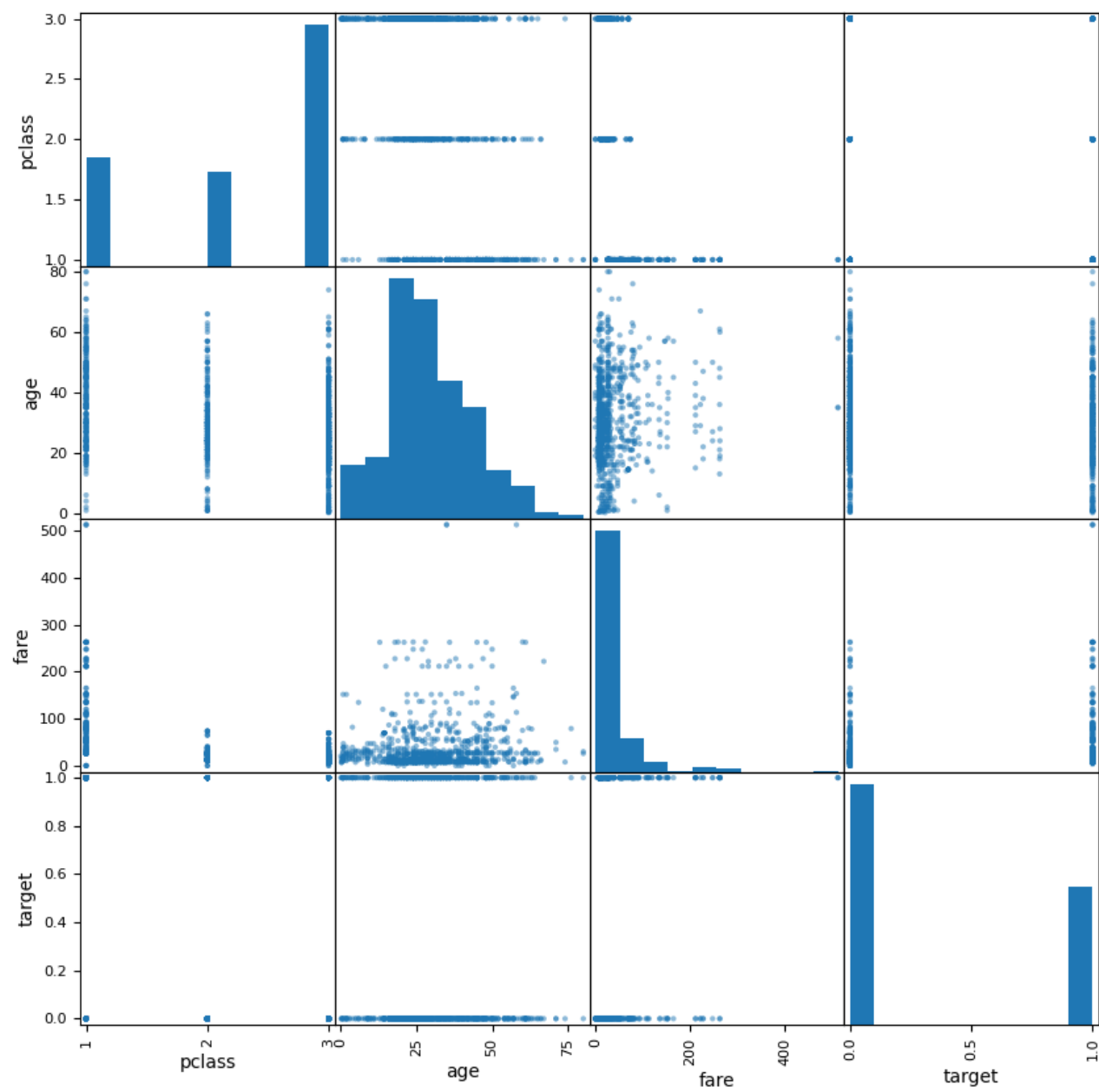
Top five positive feature correlations

feature1	feature2	correlation
fare	target	0.247922582479792
age	fare	0.13346327123590193

Top five negative feature correlations

feature1	feature2	correlation
fare	pclass	-0.5620023255866856
age	pclass	-0.35988737134796445
pclass	target	-0.2975962830046826
age	target	-0.05210618604651636

Feature correlation graph showing the scatter plot between any two numerical features. This graph helps to understand if there are any correlation between numerical features.



Chapter 4 - Class Imbalance

In this section we show statistics to bring out the imbalance between the different classes in the target column for a classification problem. This will help you learn if you need to address the issue of class imbalance in your dataset.

The summary of number of instances of each class is below

- The number of instances of class 0 are: 665
- The number of instances of class 1 are: 382

The majority class is: 0

- The ratio of number of instances of majority class 0 to class 1 is: 1.7408376963350785

References

You can visit the following links for further exploration:-

- [data.understand](#)

data.underStand