Understanding the data in cancer.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 `0.0.5` version of `data-understand`.

data.understand

Index

Chapter 1 - Dataset Charateristics

Chapter 2 - Visualize distributions of the dataset

Chapter 3 - Feature correlations between numerical features

Chapter 4 - Class Imbalance

data.understand

Chapter 1 - Dataset Charateristics

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.

data.understand

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: 455

The number of columns in the dataset are: 31

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
mean radius	float64

Column	Type
mean texture	float64
mean perimeter	float64
mean area	float64
mean smoothness	float64
mean compactness	float64
mean concavity	float64
mean concave points	float64
mean symmetry	float64
mean fractal dimension	float64
radius error	float64
texture error	float64
perimeter error	float64
area error	float64
smoothness error	float64
compactness error	float64
concavity error	float64
concave points error	float64
symmetry error	float64
fractal dimension error	float64
worst radius	float64
worst texture	float64
worst perimeter	float64
worst area	float64
worst smoothness	float64
worst compactness	float64
worst concavity	float64
worst concave points	float64
worst symmetry	float64
worst fractal dimension	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

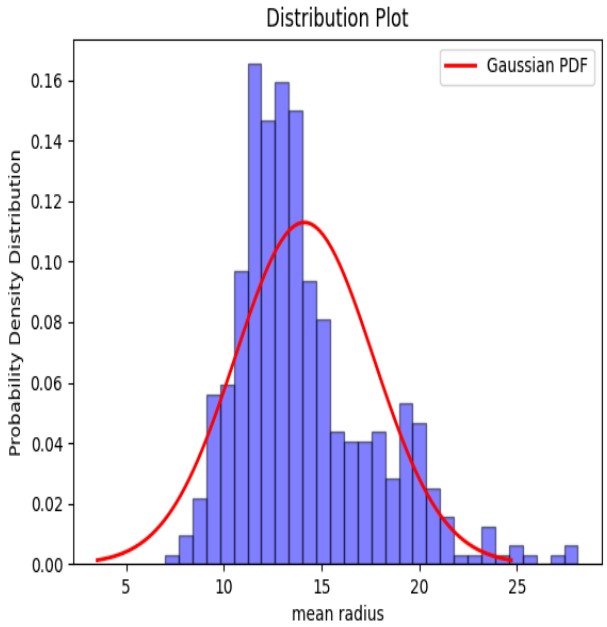
The section shows the distribution of individual categories in a given categorical column. The distribution helps to understand which categories in a given column are most/least prevelant in your dataset.

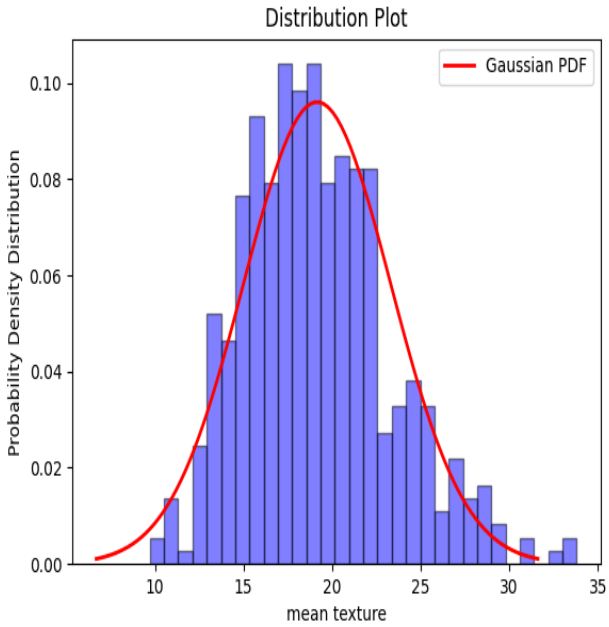
No categorical features exists in the dataset.

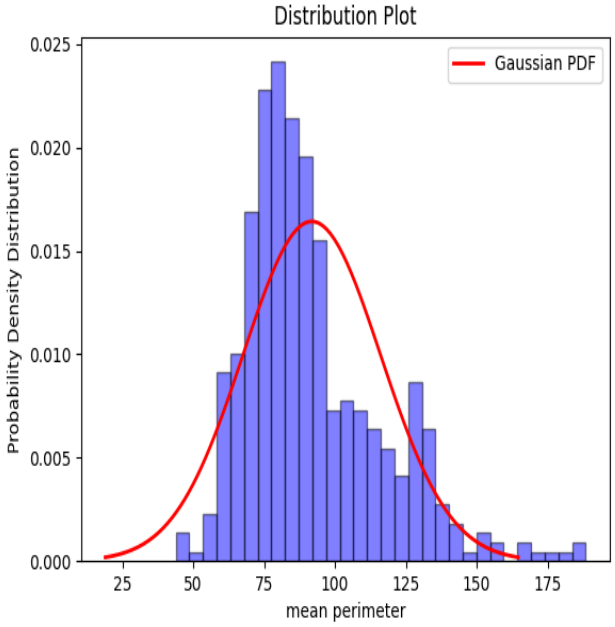
data.understand

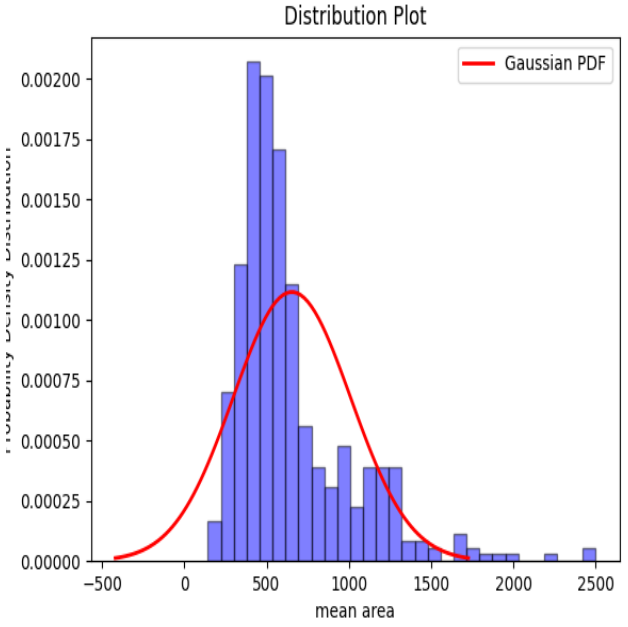
Numerical value distribution

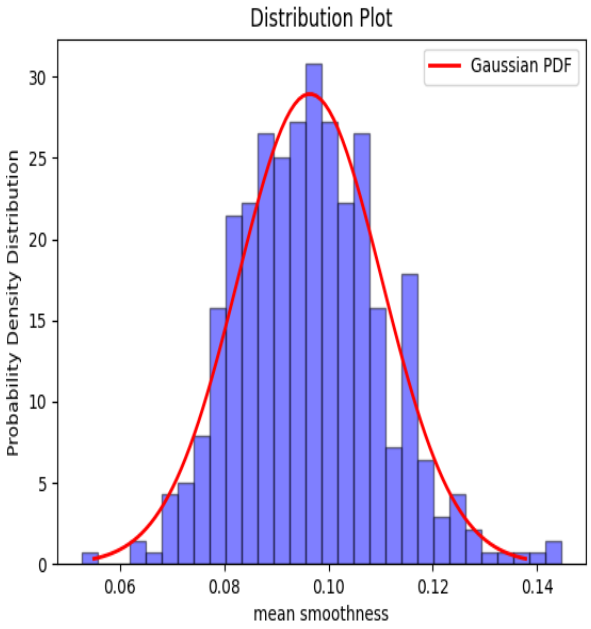
The section shows the histogram distribution of various numerical features in your dataset. The graphs also show a line chart which helps understand how the normal distribution will look if the numerical values in the distribution were normally distributed. These graphs also help gauge if the distibution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

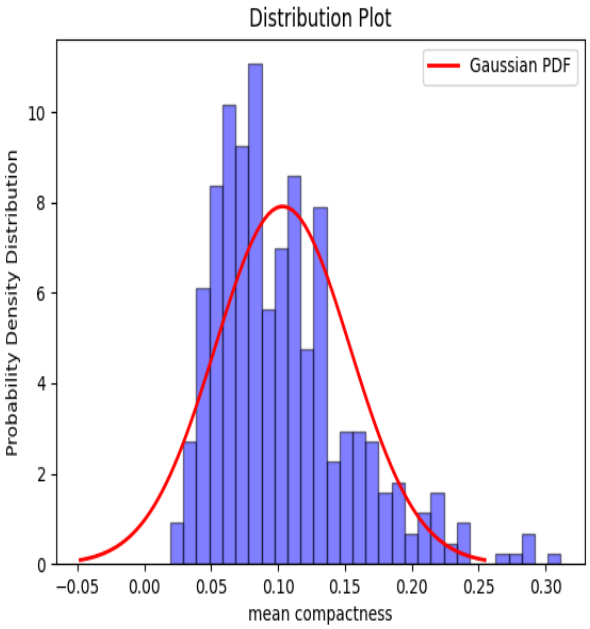


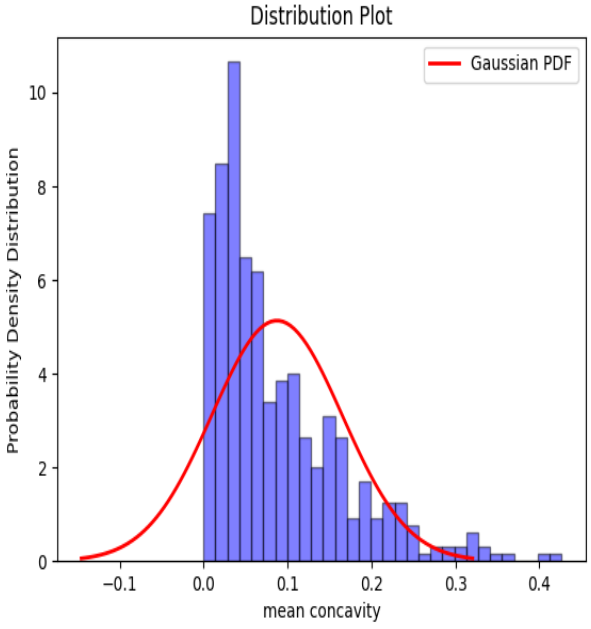


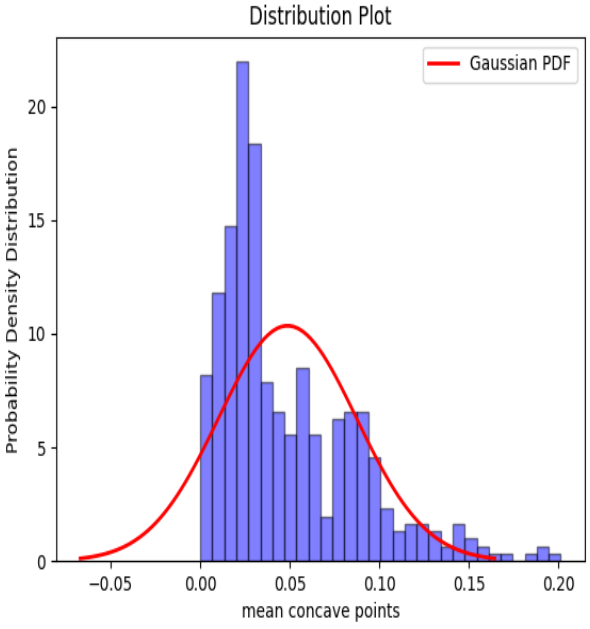


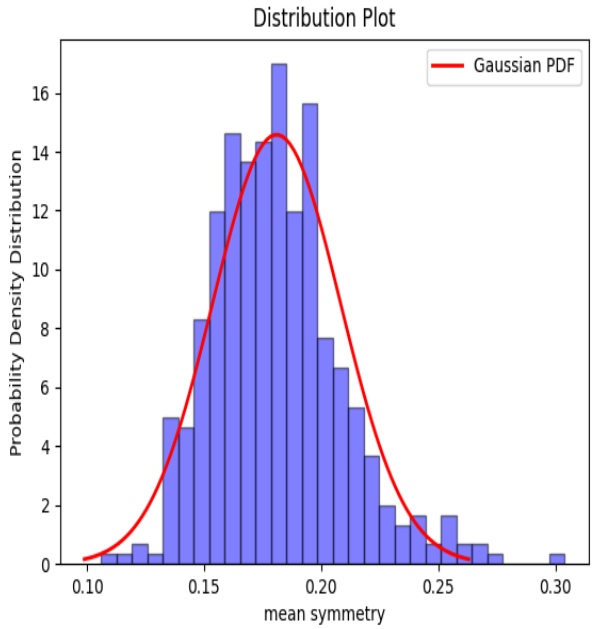


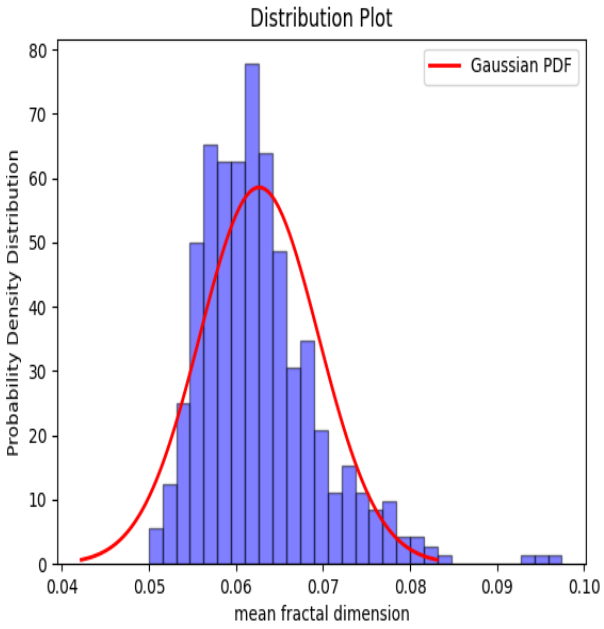

Distribution Plot
Gaussian PDF
Probability Density Distribution
mean smoothness

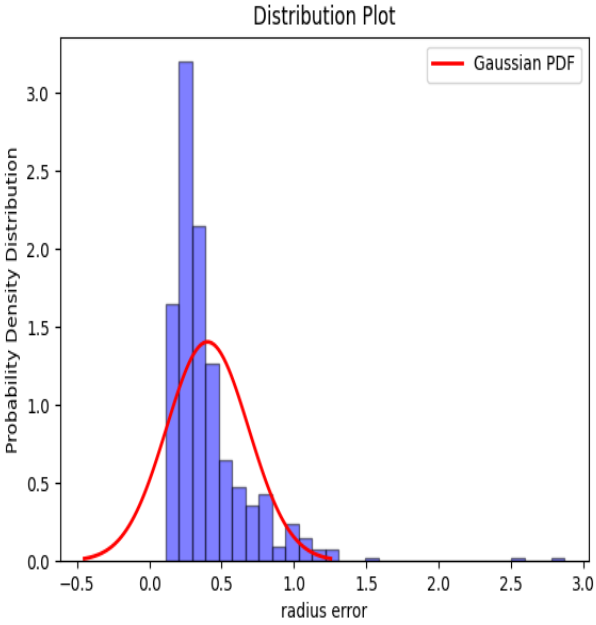

Distribution Plot
Gaussian PDF
Probability Density Distribution
mean compactness

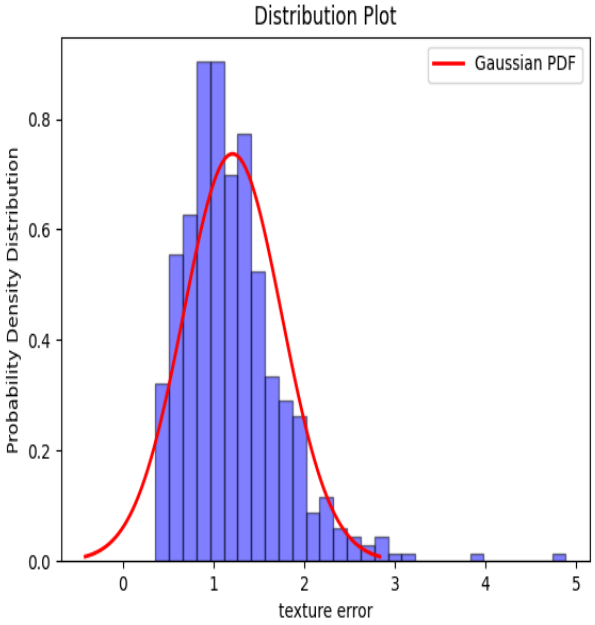

Distribution Plot
Gaussian PDF
Probability Density Distribution
mean concavity

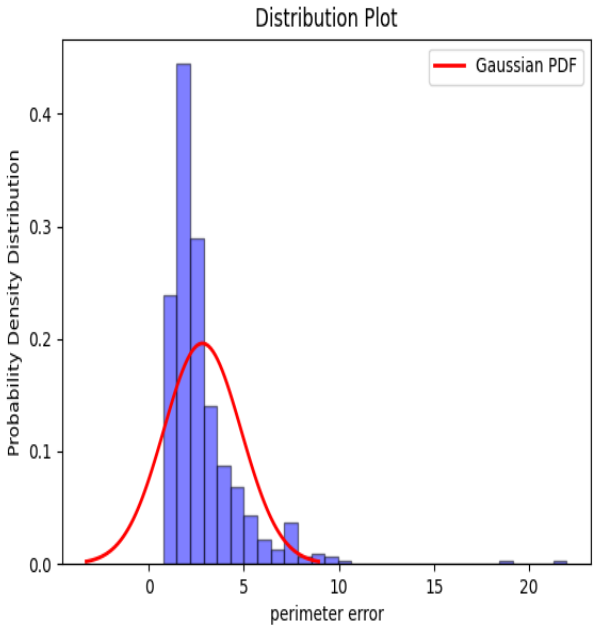

Distribution Plot
Gaussian PDF
Probability Density Distribution
mean concave points

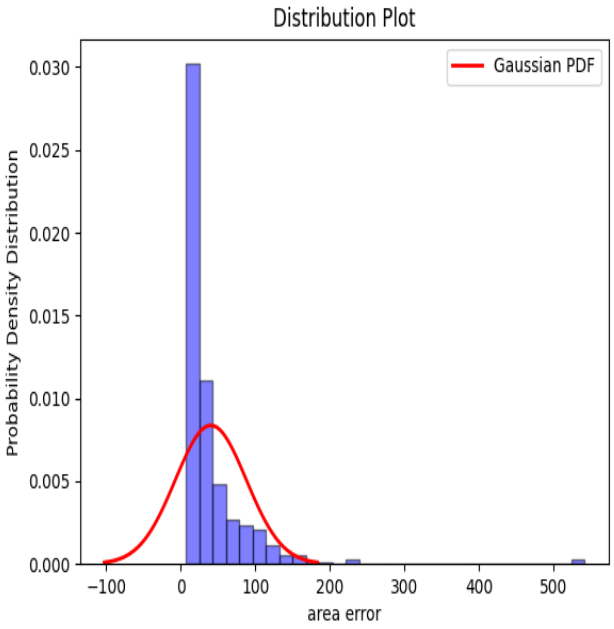

Distribution Plot
Gaussian PDF
Probability Density Distribution
mean symmetry

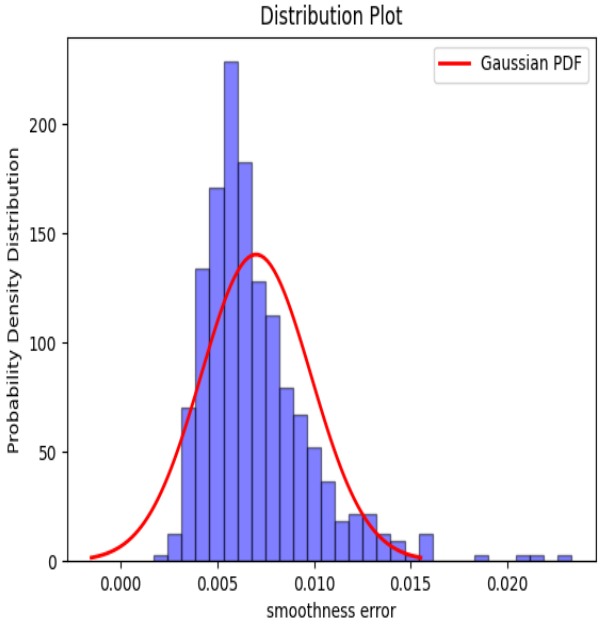

Distribution Plot
Gaussian PDF
Probability Density Distribution
mean fractal dimension

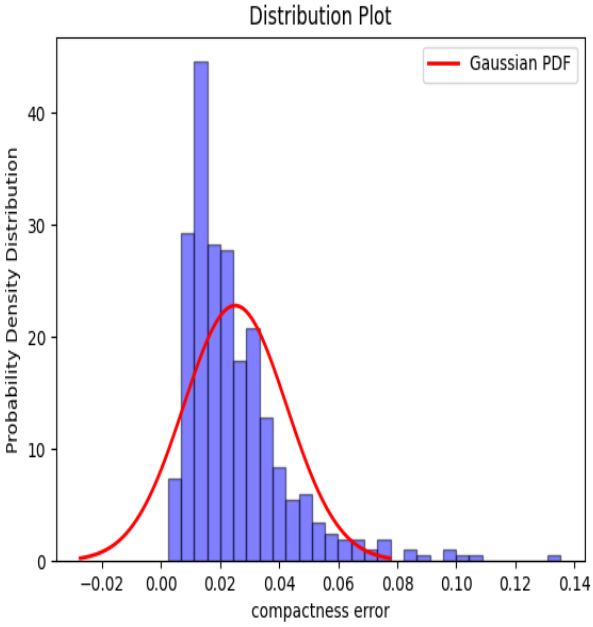

Distribution Plot
Gaussian PDF
Probability Density Distribution
radius error

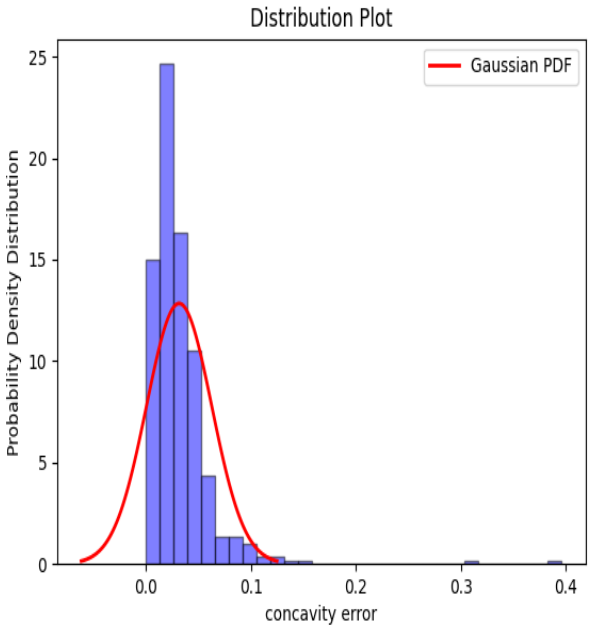

Distribution Plot
Gaussian PDF
Probability Density Distribution
texture error

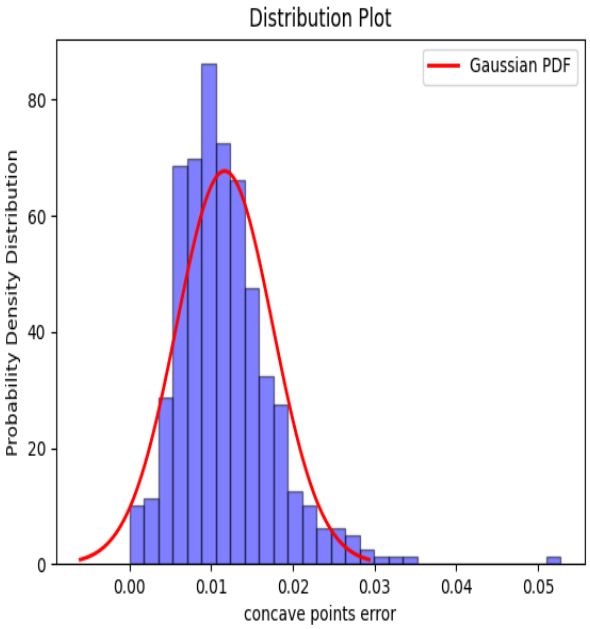

Distribution Plot
Gaussian PDF
Probability Density Distribution
perimeter error

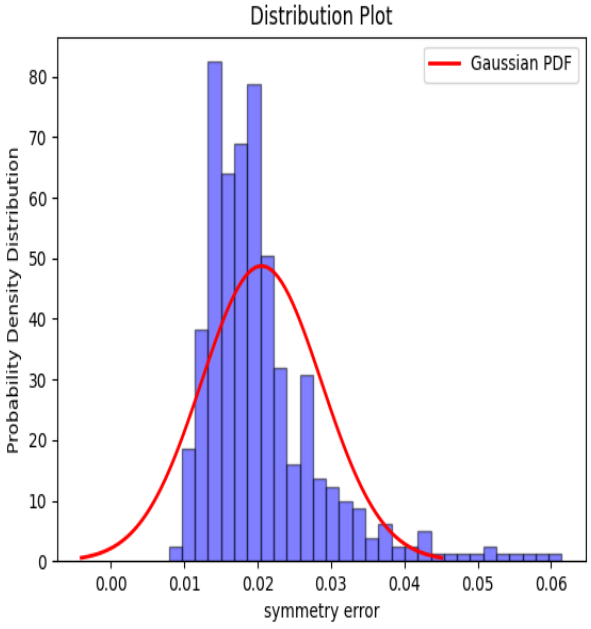

Distribution Plot
Gaussian PDF
Probability Density Distribution
area error

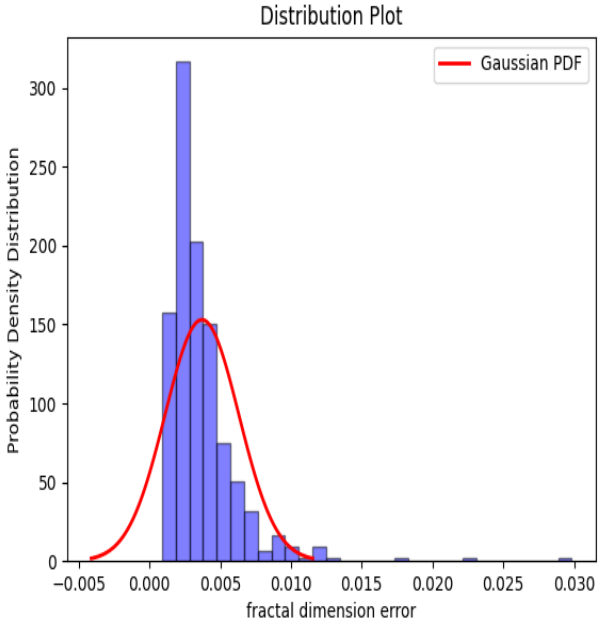

Distribution Plot
Gaussian PDF
Probability Density Distribution
smoothness error

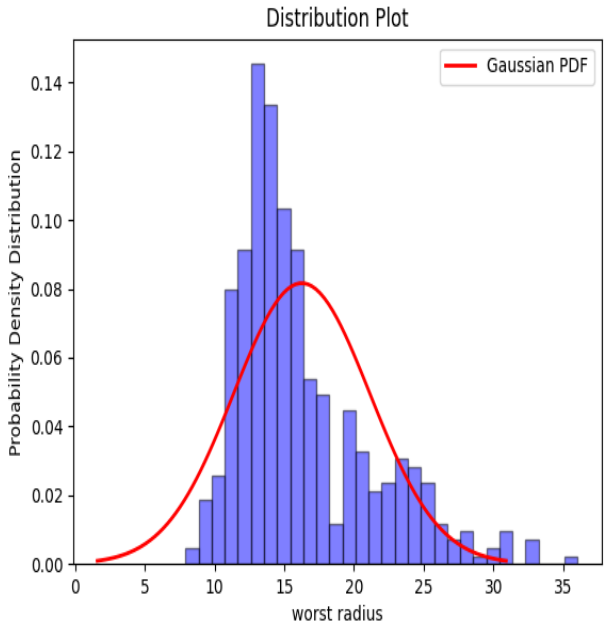

Distribution Plot
Gaussian PDF
Probability Density Distribution
compactness error

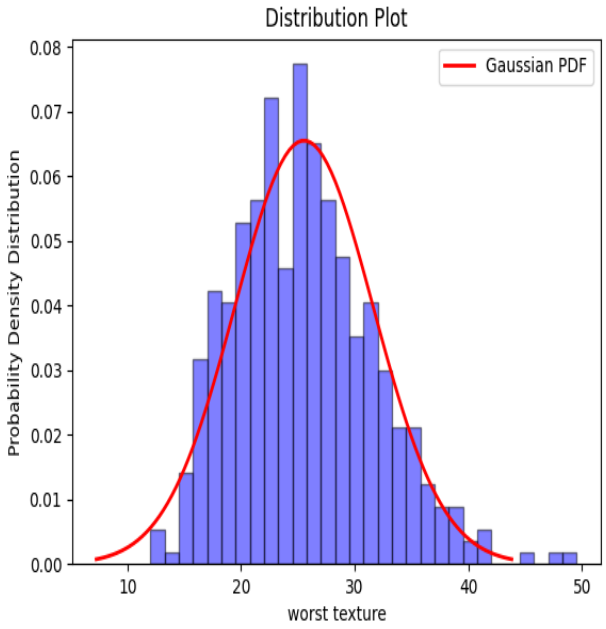

Distribution Plot
Gaussian PDF
Probability Density Distribution
concavity error

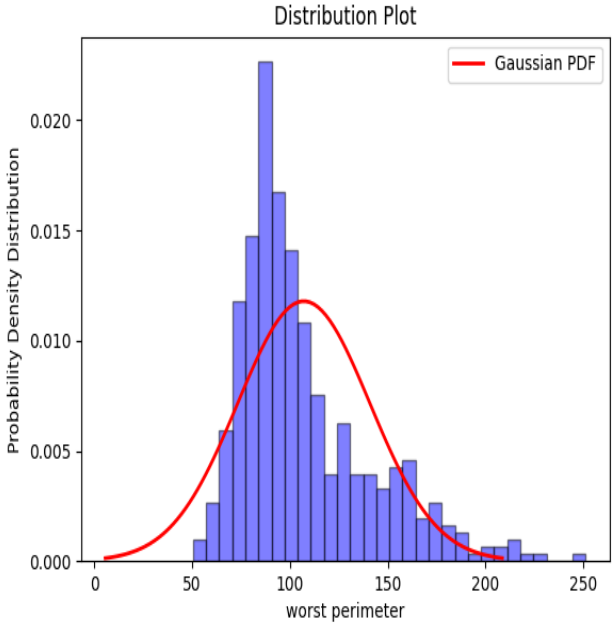

Distribution Plot
Gaussian PDF
Probability Density Distribution
concave points error

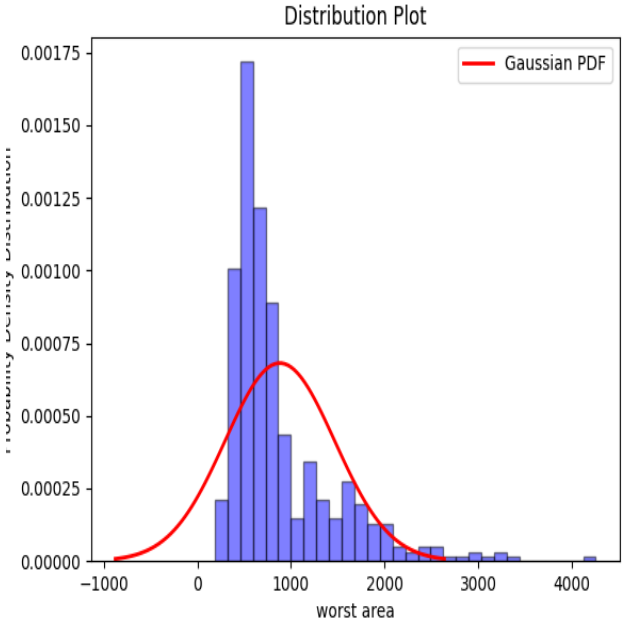

Distribution Plot
Gaussian PDF
Probability Density Distribution
symmetry error

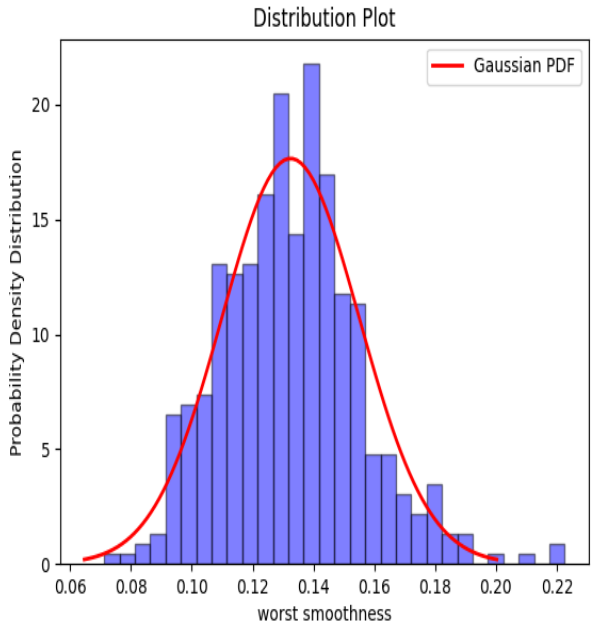

Distribution Plot
Gaussian PDF
Probability Density Distribution
fractal dimension error

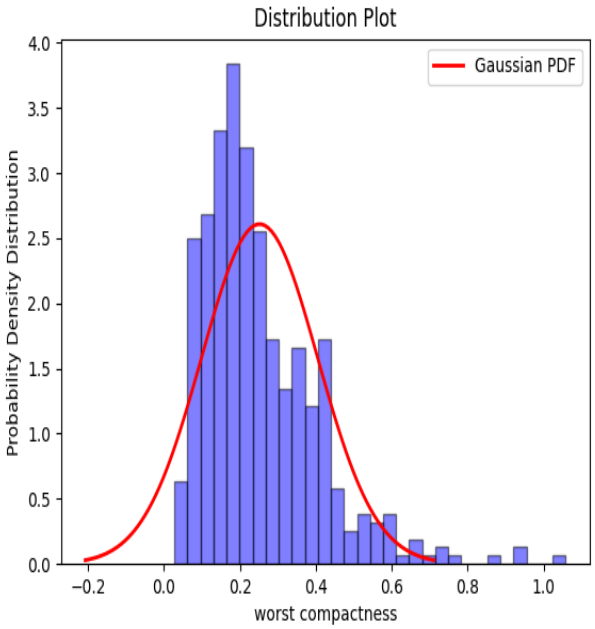

Distribution Plot
Gaussian PDF
Probability Density Distribution
worst radius

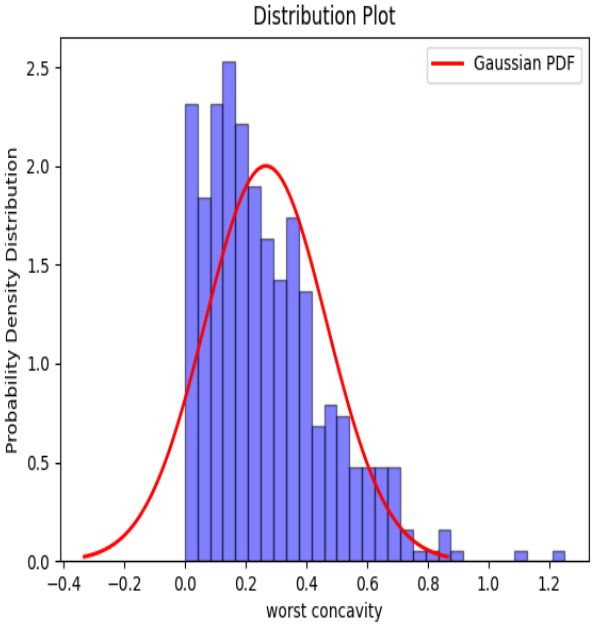

Distribution Plot
Gaussian PDF
Probability Density Distribution
worst texture

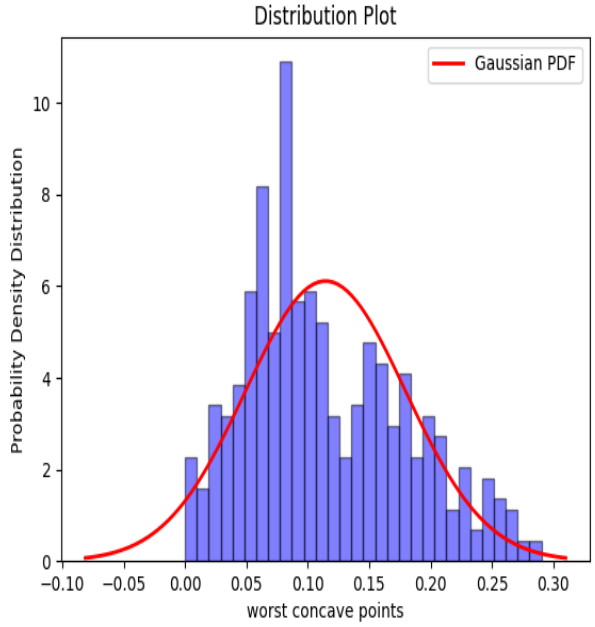

Distribution Plot
Gaussian PDF
Probability Density Distribution
worst perimeter

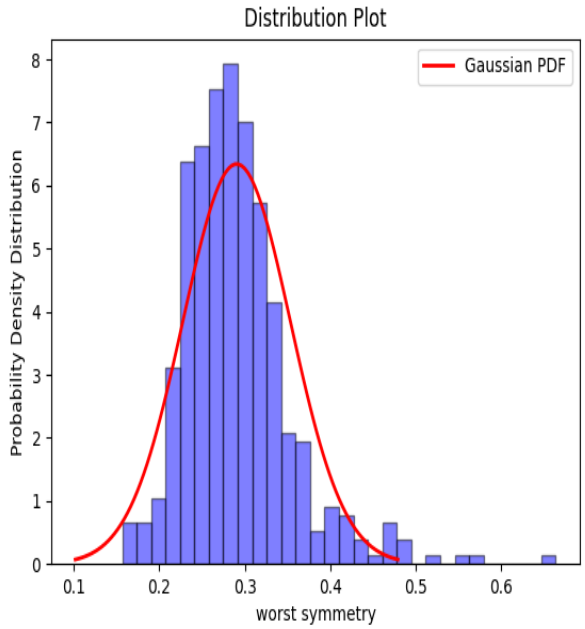

Distribution Plot
Gaussian PDF
worst area

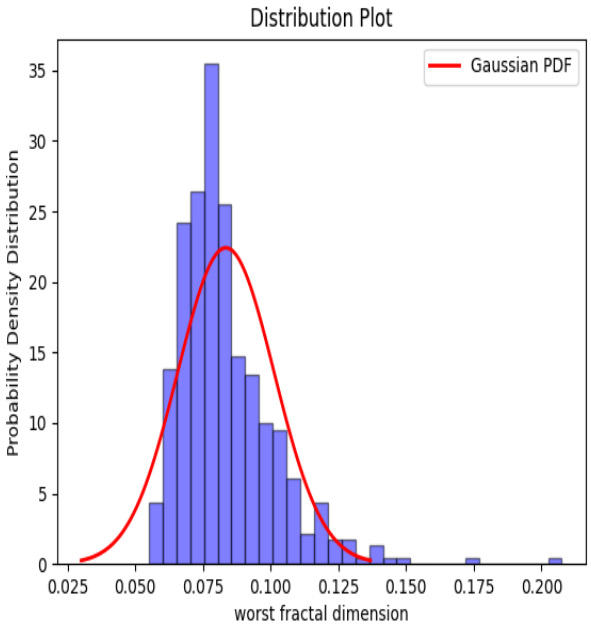

Distribution Plot
Gaussian PDF
Probability Density Distribution
worst smoothness

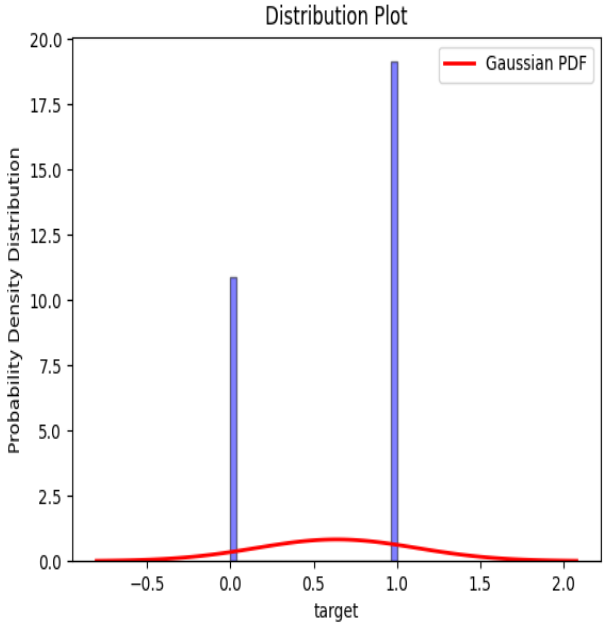

Distribution Plot
Gaussian PDF
Probability Density Distribution
worst compactness


Distribution Plot
Gaussian PDF
Probability Density Distribution
worst concavity


Distribution Plot
Gaussian PDF
Probability Density Distribution
worst concave points


Distribution Plot
Gaussian PDF
Probability Density Distribution
worst symmetry


Distribution Plot
Gaussian PDF
Probability Density Distribution
worst fractal dimension


Distribution Plot
Gaussian PDF
Probability Density Distribution
target

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.

No categorical features exists in the dataset.

data.understand

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)).

data.understand

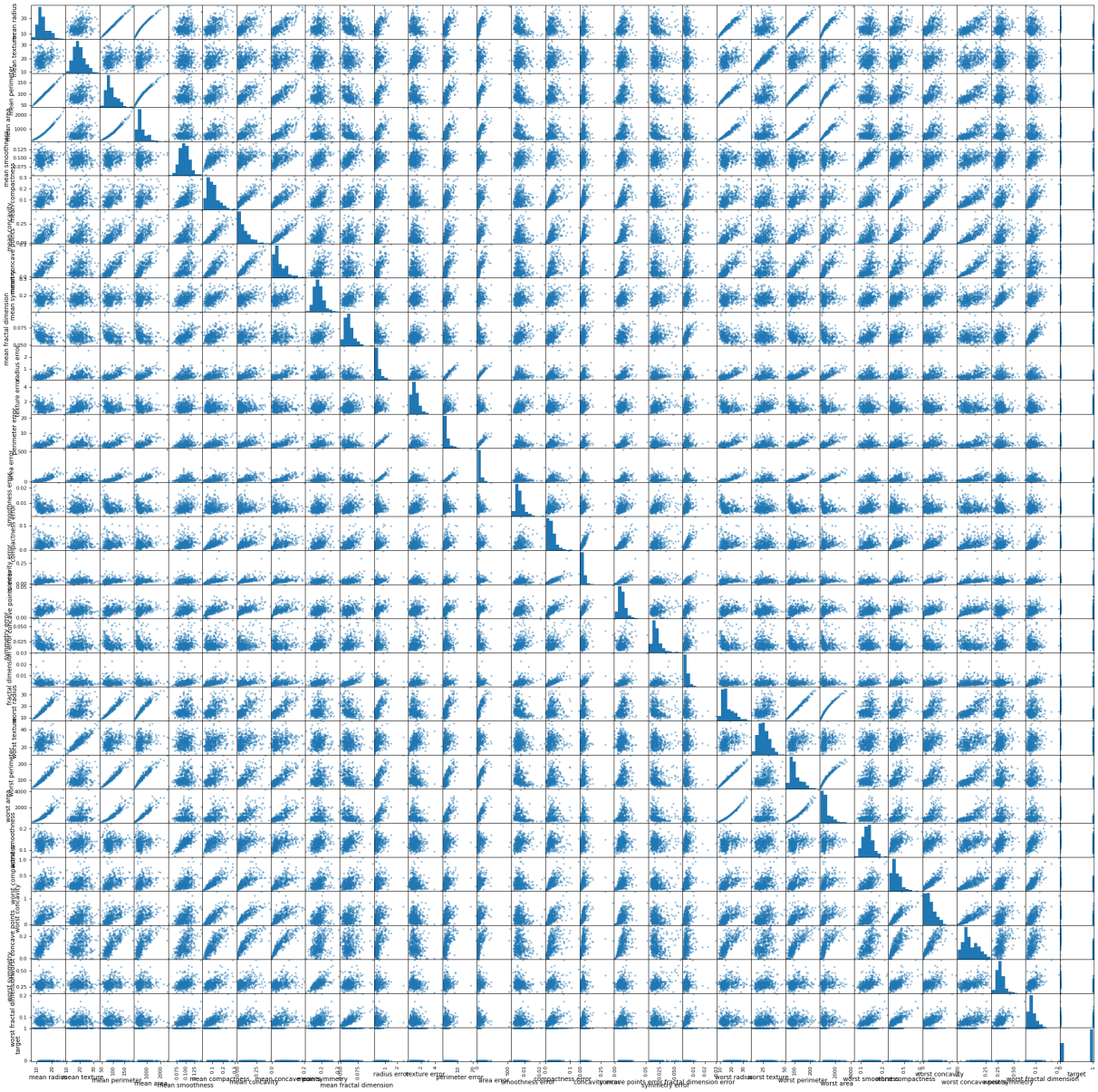
Top five positive feature correlations

feature1	feature2	correlation
mean perimeter	mean radius	0.9981847135526685
worst perimeter	worst radius	0.9946964711913601
mean area	mean radius	0.9870935596454733
mean area	mean perimeter	0.986691605032055
worst area	worst radius	0.9838413579498259

Top five negative feature correlations

feature1	feature2	correlation
target	worst concave points	-0.7934404959006445
target	worst perimeter	-0.775844411419966
mean concave points	target	-0.7757647846988944
target	worst radius	-0.7701596423619639
mean perimeter	target	-0.7364961292505654

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.

data.understand

The summary of number of instances of each class is below

- The number of instances of class 0 are: 165

- The number of instances of class 1 are: 290

The majority class is: 1

- The ratio of number of instances of majority class 1 to class 0 is: 1.7575757575757576

References

You can visit the following links for further exploration:-

- data.understand

data.understand