Understanding the data in datasets/classification/adult/adult.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.

Index

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.

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: 26048

The number of columns in the dataset are: 13

The name of the target column is: target

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

No columns were found to have missing values

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

Column	Type
Age	float64
Workclass	int64

Column	Type
Education-Num	float64
Marital Status	int64
Occupation	int64
Relationship	int64
Race	int64
Sex	int64
Capital Gain	float64
Capital Loss	float64
Hours per week	float64
Country	int64
target	int64

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.

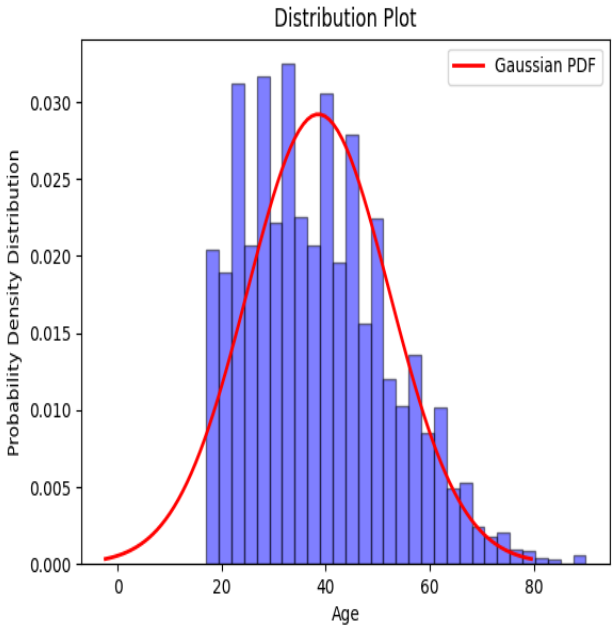
Categorical feature distribution

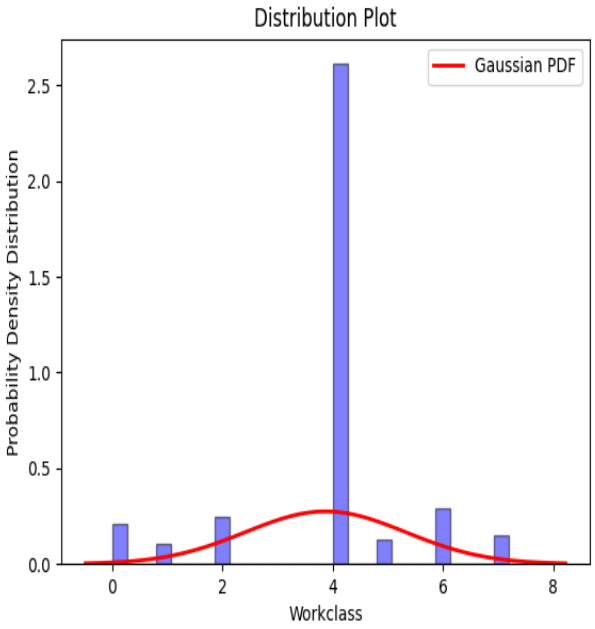
The section shows the distribution of individual categories in a given categorical column. The distribution helps to understand which categories in a given column are most/least prevelant in your dataset.

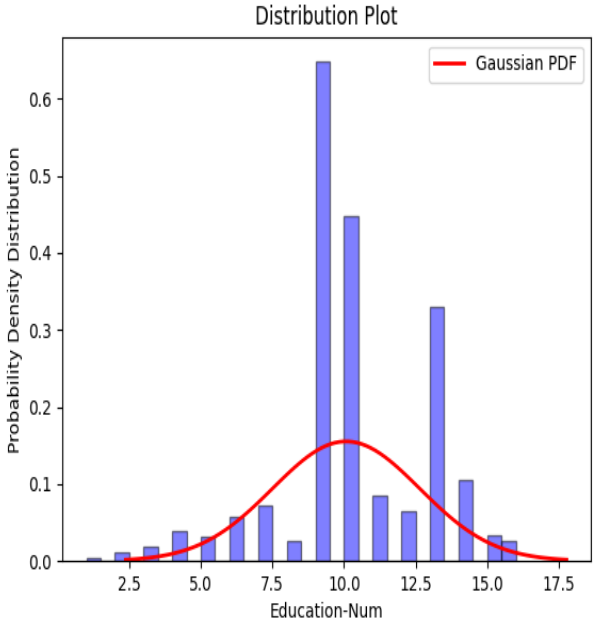
No categorical features exists in the dataset.

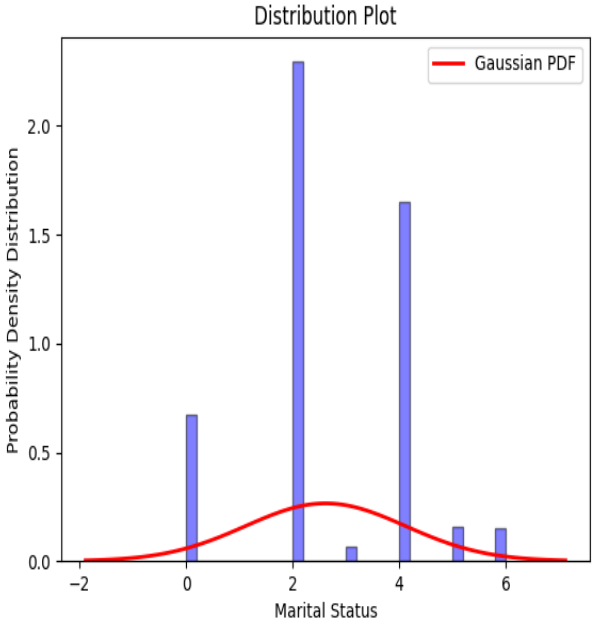
Numerical value distribution

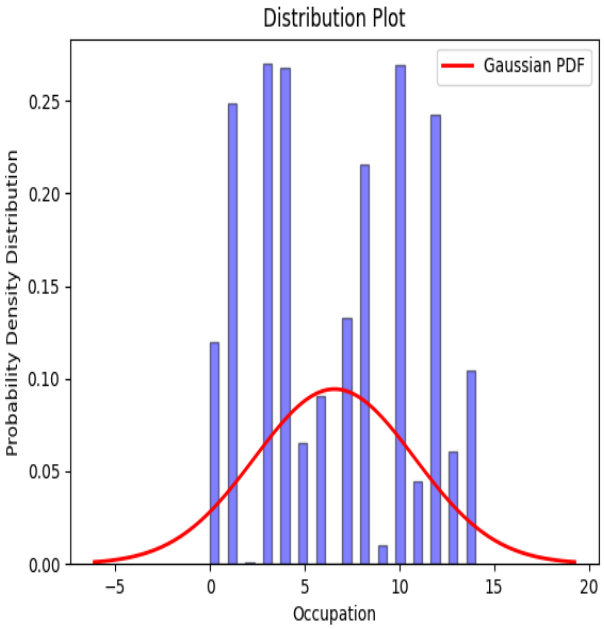
The section shows the histogram distribution of various numerical features in your dataset. The graphs also show a line chart which helps understand how the normal distribution will look if the numerical values in the distribution were normally distributed. These graphs also help gauge if the distibution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

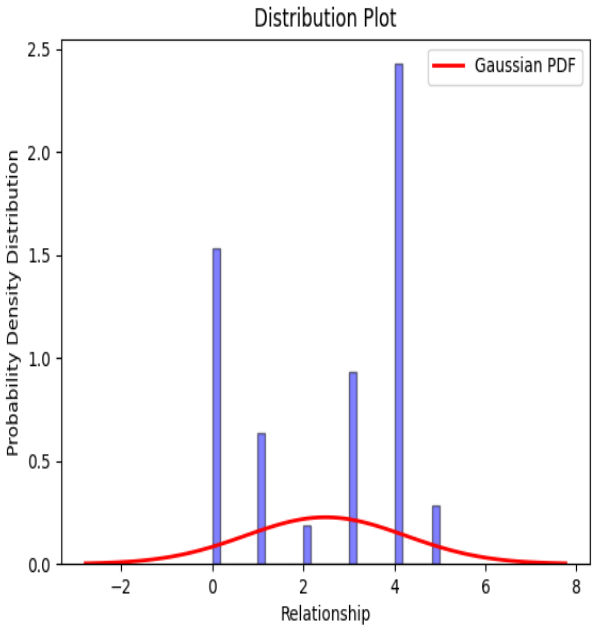


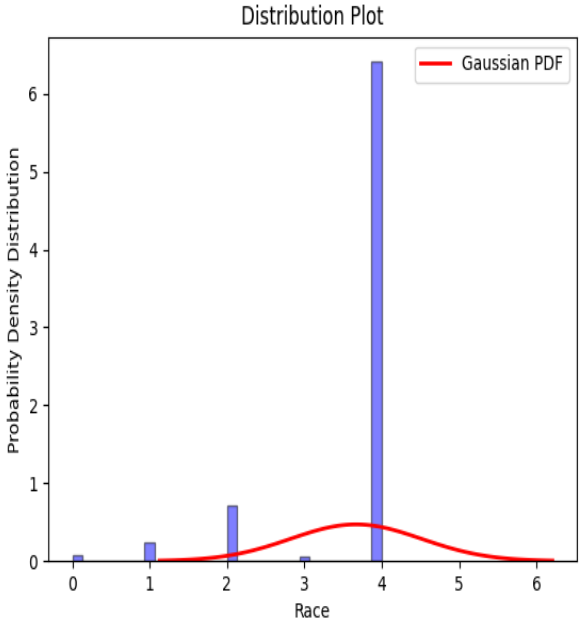


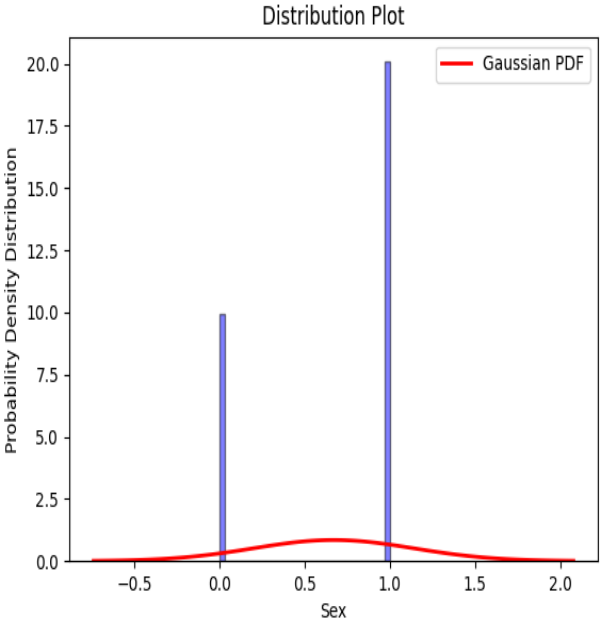


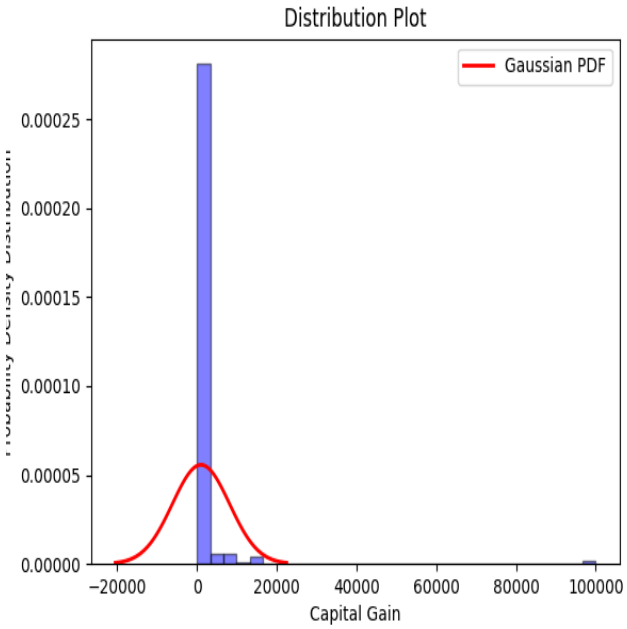


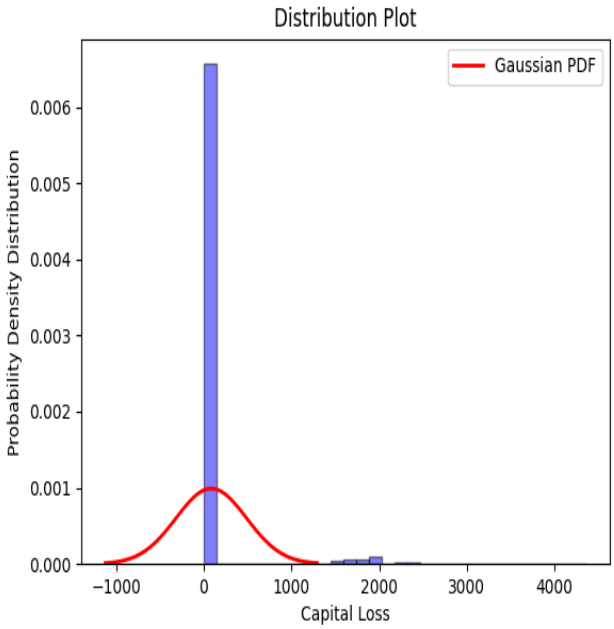

Distribution Plot
Gaussian PDF
Probability Density Distribution
Occupation

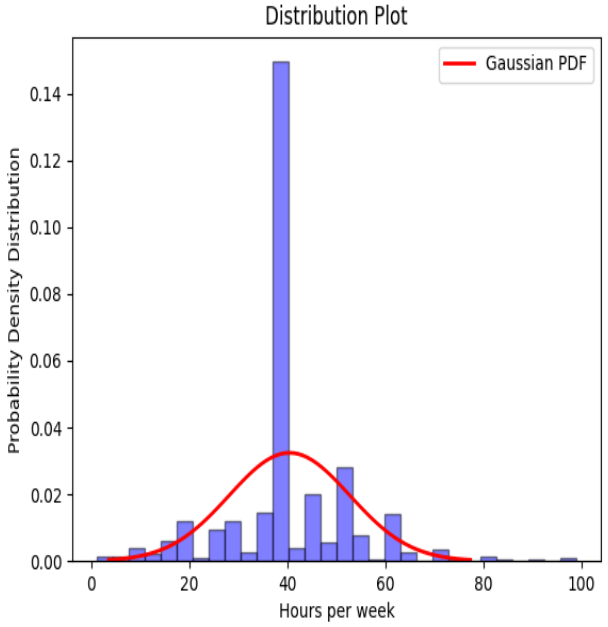

Distribution Plot
Gaussian PDF
Probability Density Distribution
Relationship

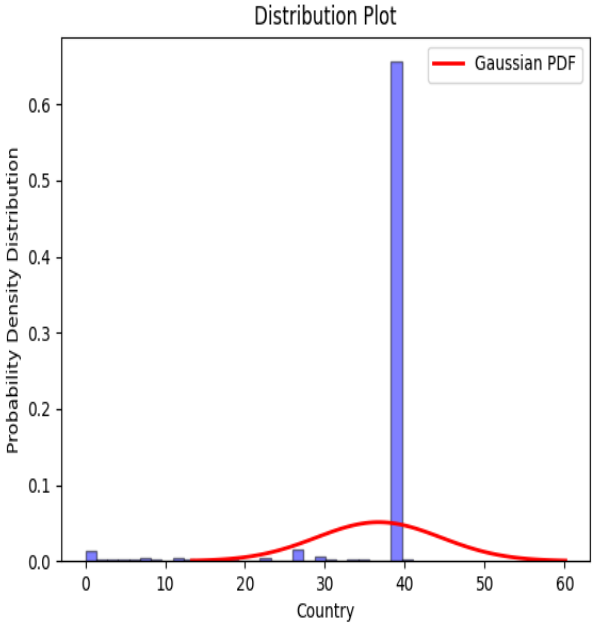

Distribution Plot
Gaussian PDF
Probability Density Distribution
Race

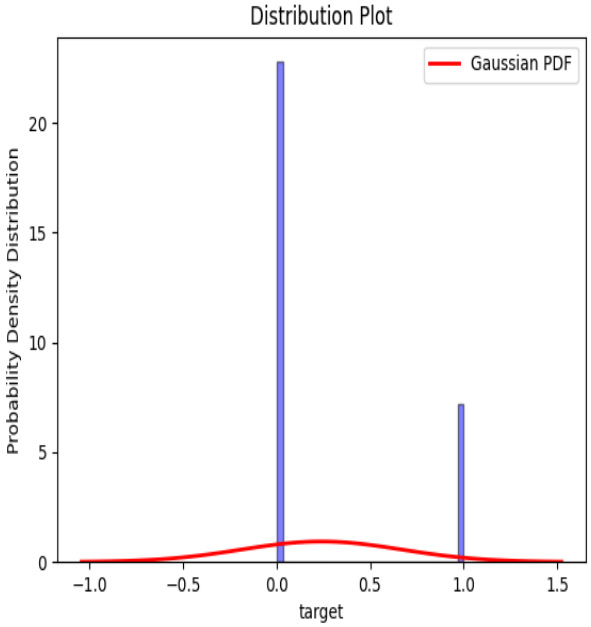

Distribution Plot
Gaussian PDF
Probability Density Distribution
Sex


Distribution Plot
Gaussian PDF
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
−20000
0
20000
40000
60000
80000
100000
Capital Gain


Distribution Plot
Gaussian PDF
Probability Density Distribution
0.006
0.005
0.004
0.003
0.002
0.001
0.000
−1000
0
1000
2000
3000
4000
Capital Loss


Distribution Plot
Gaussian PDF
Probability Density Distribution
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
Hours per week


Distribution Plot
Gaussian PDF
Probability Density Distribution
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
Country


Distribution Plot
Gaussian PDF
Probability Density Distribution
target
0
5
10
15
20
−1.0
−0.5
0.0
0.5
1.0
1.5

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.

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)).

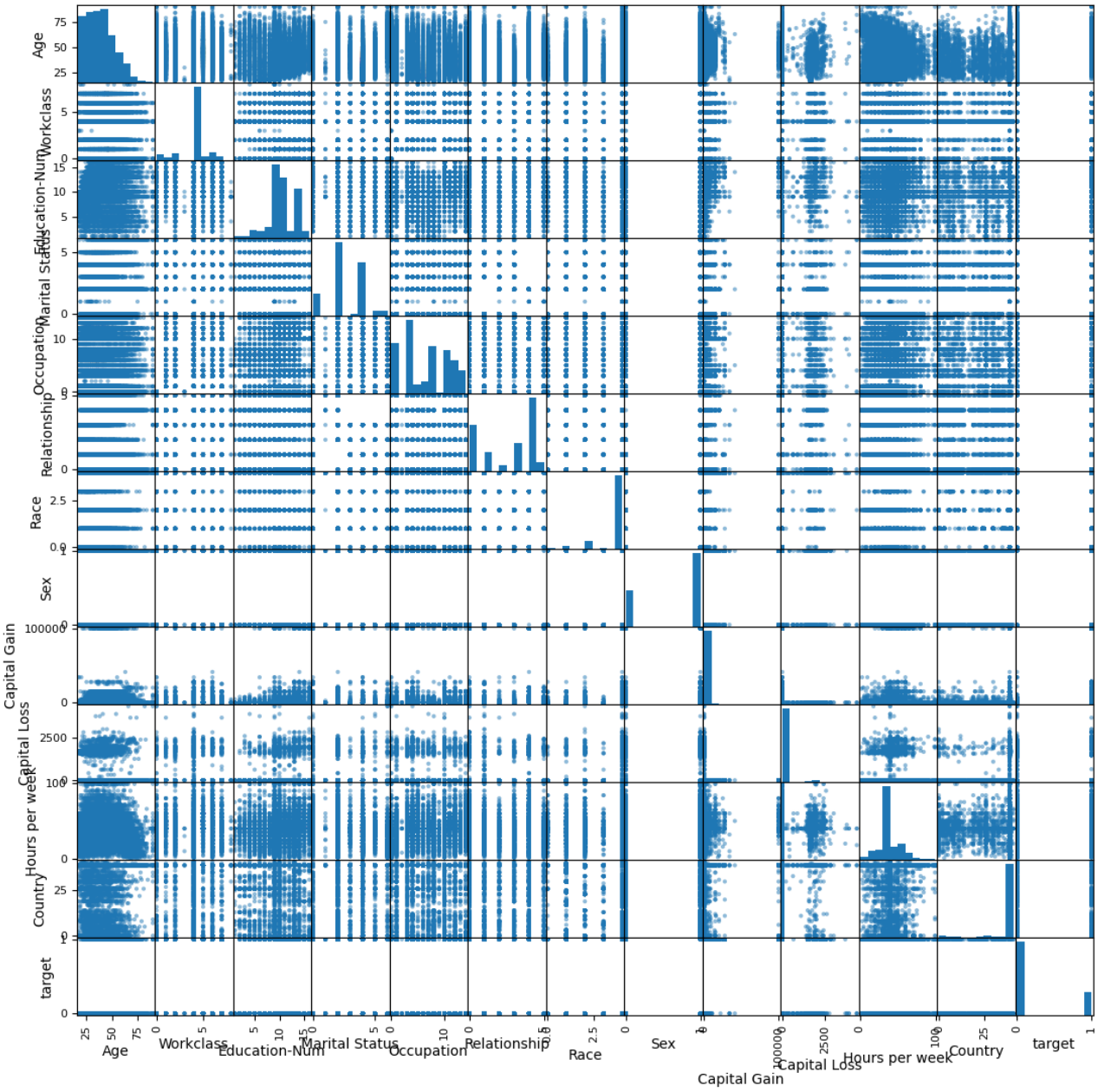
Top five positive feature correlations

feature1	feature2	correlation
Relationship	target	0.34032782964476566
Education-Num	target	0.33412736473763494
Relationship	Sex	0.3277513882487867
Occupation	Workclass	0.254978841450977
Age	target	0.23469799864420116

Top five negative feature correlations

feature1	feature2	correlation
Age	Marital Status	-0.26519502015468693
Marital Status	Relationship	-0.2313500376863775
Marital Status	target	-0.20248133649704964
Hours per week	Marital Status	-0.18668741730876198
Marital Status	Sex	-0.12987186202144915

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: 19775

- The number of instances of class 1 are: 6273

The majority class is: 0

- The ratio of number of instances of majority class 0 to class 1 is: 3.152399171050534

References

You can visit the following links for further exploration:-

- data.understand