Exploratory Data Analysis on Loan Data

Summary of Data Frame

```
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 614 entries, 0 to 613
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Loan_ID            614 non-null    object
 1   Gender             601 non-null    object
 2   Married            611 non-null    object
 3   Dependents         599 non-null    object
 4   Education          614 non-null    object
 5   Self_Employed      582 non-null    object
 6   ApplicantIncome    614 non-null    int64
 7   CoapplicantIncome  614 non-null    float64
 8   LoanAmount         592 non-null    float64
 9   Loan_Amount_Term   600 non-null    float64
 10  Credit_History     564 non-null    float64
 11  Property_Area      614 non-null    object
 12  Loan_Status        614 non-null    object
dtypes: float64(4), int64(1), object(8)
memory usage: 62.5+ KB
None
```

In this dataset 13 columns present 8 are object type, in other words they are categorical data. 4 are float type and one is integer type of data.

Descriptive Statistical Measures of a DataFrame

```
data.describe()
```

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

The statistical measures of the dataset can be viewed with the help of describe function. Here count, mean, standard deviation, minimum value, 1st quartile, median, 3rd quartile, maximum values are tabulated for numeric values.

Handling Missing Value

We can handle missing values in many ways. Here we are doing in 2 ways they are dropping null values and imputing with mean, median and mode. While dropping null some records and interesting patterns may be deleted. We can observe the change in dataset by viewing the shape. But in imputation method it is retained. We can observe the change in dataset by viewing the shape

Dropna

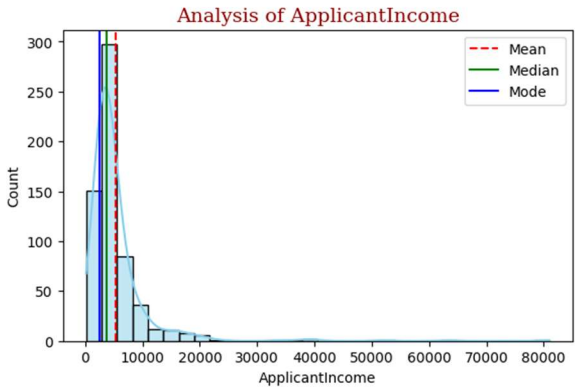
```
Shape before:  (614, 13)
Shape after:  (480, 13)
```

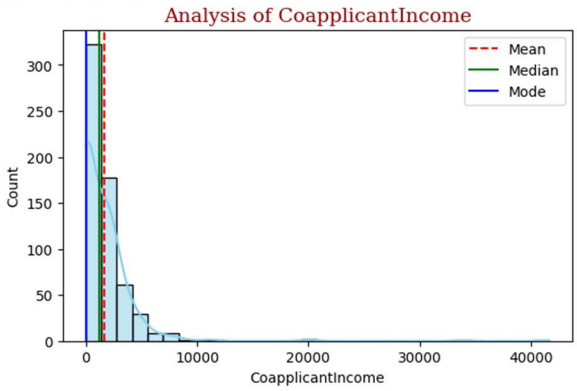
Imputation

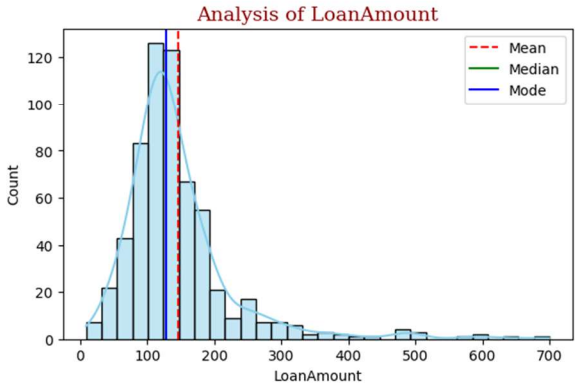
```
Shape before:  (614, 13)
Shape after:  (614, 13)
```

Data Frame Visualization

Mean median Mode plotting







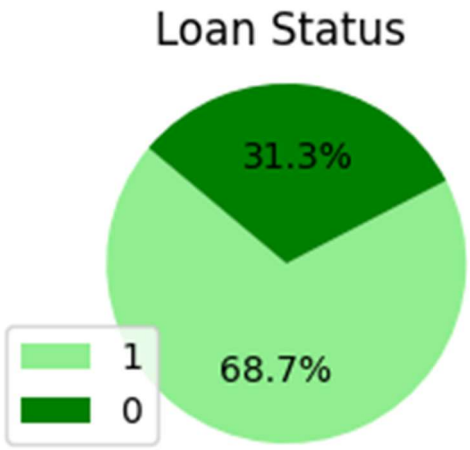
Encoding Categorical data to Numeric data

To convert categorical values to numeric values we use encoding. There are various encoding techniques available. Here we are using label encoding, where every unique value in the particular feature is mapped with unique integer.

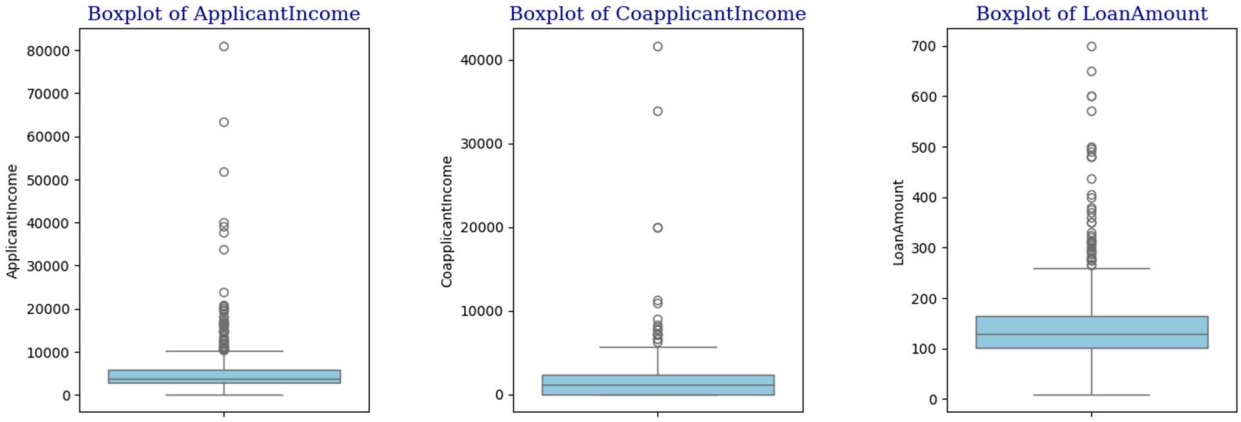
```
Loan_ID               int32
Gender                int32
Married               int32
Dependents            int32
Education             int32
Self_Employed         int32
ApplicantIncome       int64
CoapplicantIncome   float64
LoanAmount          float64
Loan_Amount_Term    float64
Credit_History      float64
Property_Area         int32
Loan_Status           int32
dtype: object
```

Loan Status Distribution

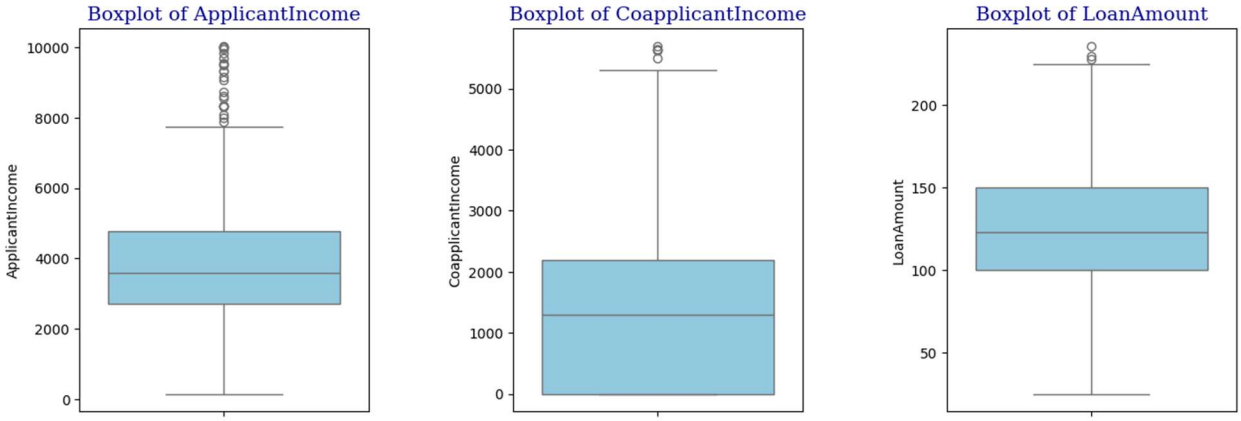
Here the Loan Status is Y or N in the dataset. It is mapped to 0 for N and 1 for Y and it is aggregated or grouped together to do this calculation. Here we observe that 31.3% are N and 68.7% are Y in the Loan Status.



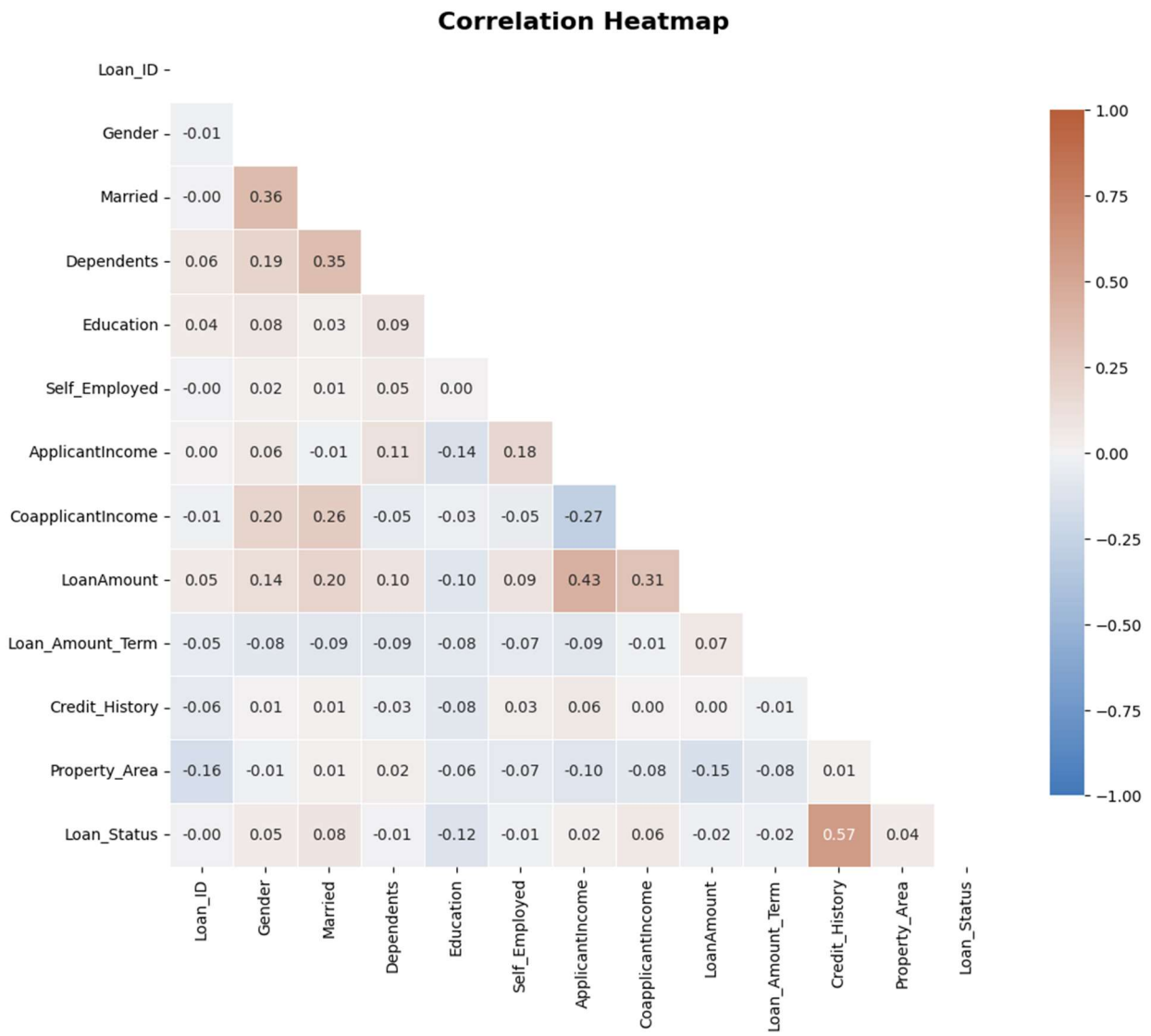
Box Plot before Outlier Removal



Box Plot after Outlier Removal

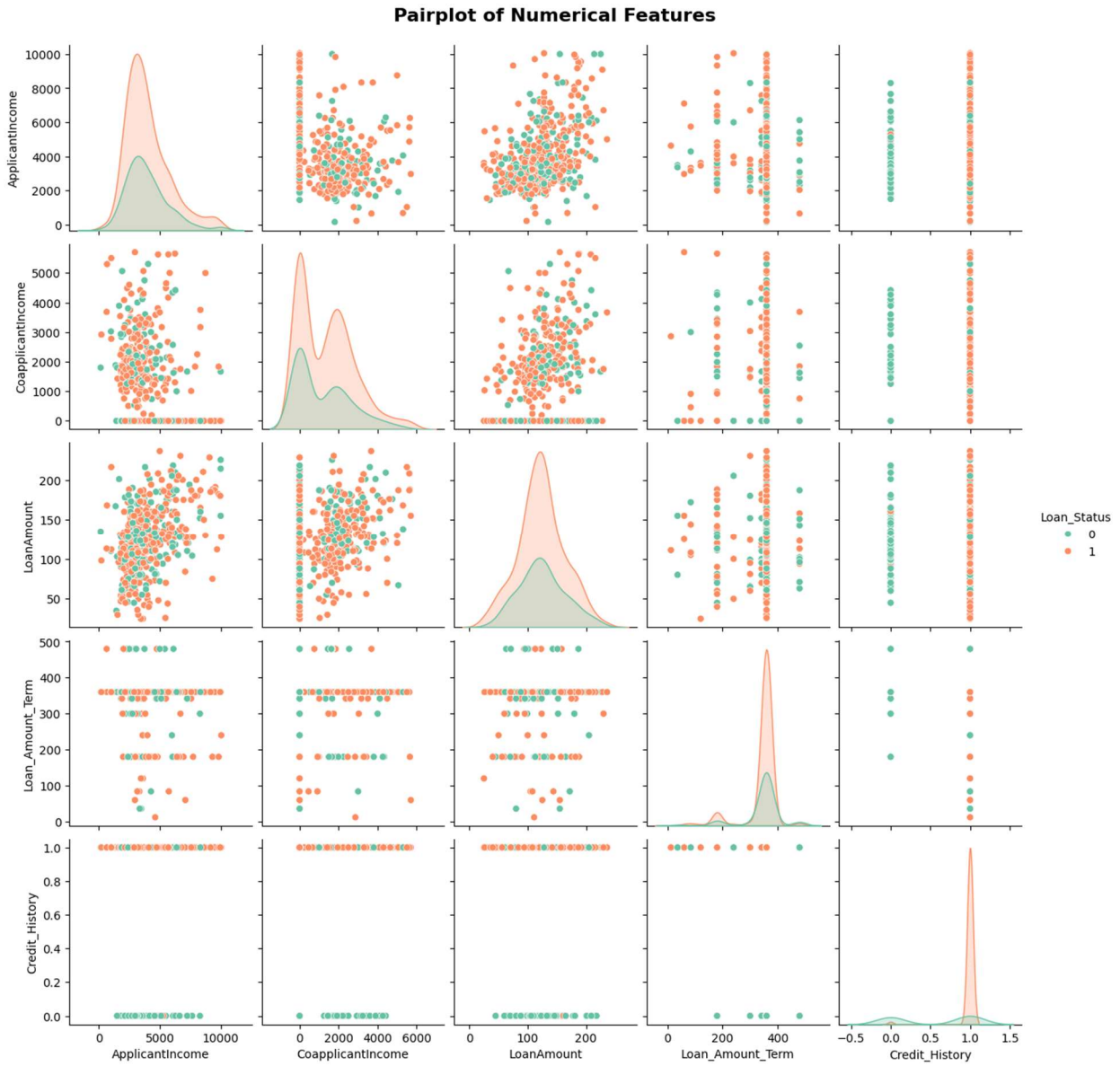


Here we use box plot to view the dataset distribution. The datapoints present outside the Inter Quartile range are the outliers. Outliers always present outside the range. We can clearly observe the Outliers present before and after its removal.



Correlation heatmap helps in understanding the relationship or correlativity among the various features. Values near to +1 are Strong positively correlated. Values near to -1 are negatively correlated. And the values nearing to 0 are not having any correlation.

Pair Plot to visualize data correlation



This pair plot helps to understand the nature of a feature with Loan status. The shape of its distribution tells more detail about its relationship among the feature.