Objective: To explore various AutoEDA capabilities and perform analysis on a given dataset

This notebook will focus on DataPrep

2. AutoEDA - DataPrep

Dataset Reference: Loan Prediction dataset from Kaggle

Features:

- General Overview - Quick insights of all variables in the dataset using the plot dataframe.
- Details about each variables / features in the dataset by using create_report - overview, variables, interactions, correlations, missing values
- Interactions - based on x-axis and y-axis scatter plots
- Correlations between variables - Pearson's Correlation Coefficient, Spearman's Rank Correlation Coefficient, Kendall's Rank Correlation Coefficient
- Missing Values - Bar chart, Spectrum, Heatmap, Dendogram representations
- We can pick one particular feature and analyze - Stats, Bar chart, Pie chart, Word Count, Word Frequency etc as per applicability

When To Use?

- Dataset size is fairly very large (this seems to be 10X faster than Pandas Profiling tools due to it's highly optimized Dask-based computing module)
- Need some quick insights about an unknown dataset
- Use this as a basis for your further EDA analysis on top of it

In [20]:
```
import pandas as pd
import warnings

warnings.filterwarnings("ignore")
```

In [21]:
```
# !pip --disable-pip-version-check install dataprep  # Please use it for the first time if it is not installed in your environment
```

In [22]:
```
from dataprep.eda import create_report, plot, plot_correlation, plot_missing
```

In [23]:
```
df_train = pd.read_csv("../input/loan-eligible-dataset/loan-train.csv")

df_train.head()
```

Out[23]:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_Hist
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	

In [24]:
```
df_test = pd.read_csv("../input/loan-eligible-dataset/loan-test.csv")

df_test.head()
```

Out[24]:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_Hist
0	LP001015	Male	Yes	0	Graduate	No	5720	0	110.0	360.0	
1	LP001022	Male	Yes	1	Graduate	No	3076	1500	126.0	360.0	
2	LP001031	Male	Yes	2	Graduate	No	5000	1800	208.0	360.0	
3	LP001035	Male	Yes	2	Graduate	No	2340	2546	100.0	360.0	I
4	LP001051	Male	No	0	Not Graduate	No	3276	0	78.0	360.0	

In [25]:
```
df_train.shape
```

Out[25]:
```
(614, 13)
```

In [26]:
```
df_test.shape
```

Out[26]: (367, 12)

In [27]:
```
plot(df_train)
```

Out[27]:

Hide Stats and Insights

Dataset Statistics

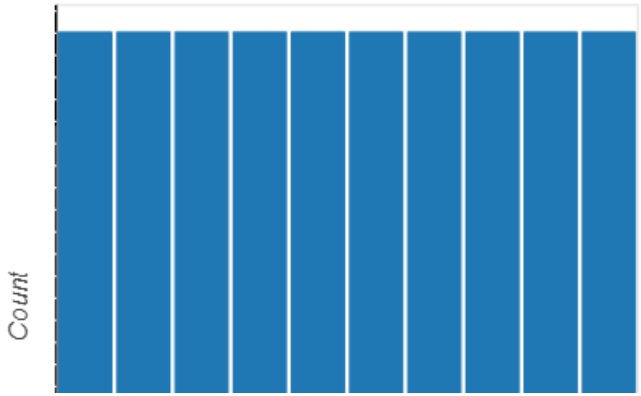
Number of Variables	13
Number of Rows	614
Missing Cells	149
Missing Cells (%)	1.9%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	316.6 KB
Average Row Size in Memory	528.0 B
Variable Types	Categorical: 8 GeoGraphy: 1 Numerical: 4

Dataset Insights

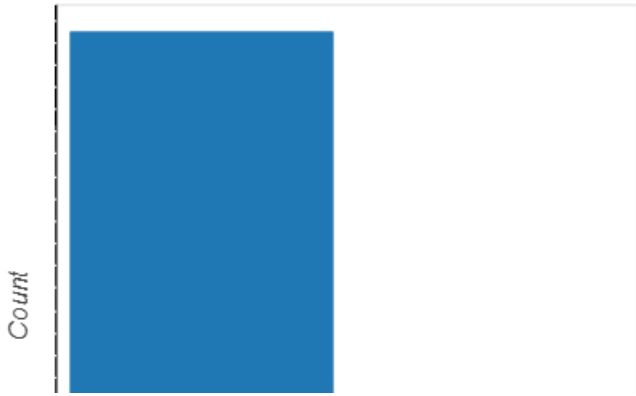
Gender	has 13 (2.12%) missing values	Missing
Dependents	has 15 (2.44%) missing values	Missing
Self_Employed	has 32 (5.21%) missing values	Missing
LoanAmount	has 22 (3.58%) missing values	Missing
Loan_Amount_Te…	has 14 (2.28%) missing values	Missing
Credit_History	has 50 (8.14%) missing values	Missing
ApplicantIncome	is skewed	Skewed
CoapplicantInc…	is skewed	Skewed
LoanAmount	is skewed	Skewed
Loan_Amount_Te…	is skewed	Skewed

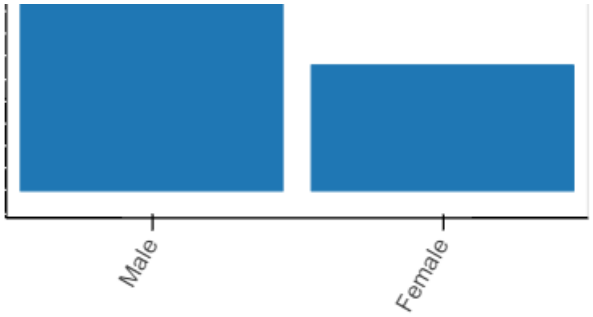
1 2

Loan_ID

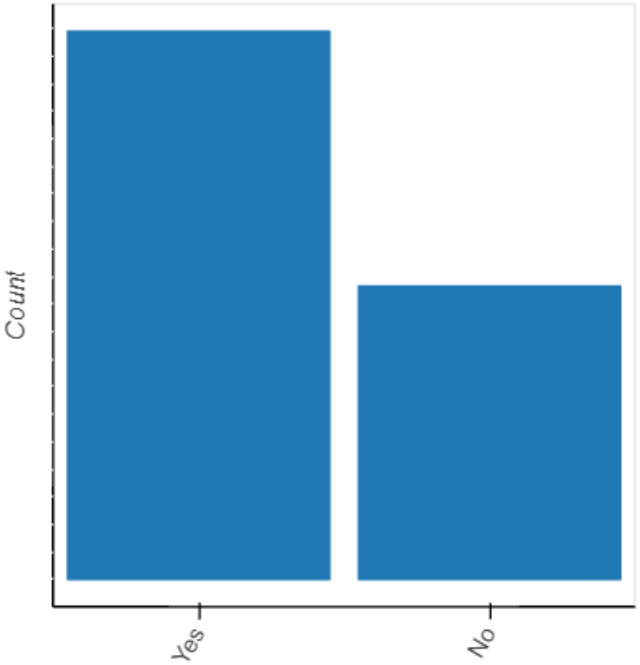


Gender

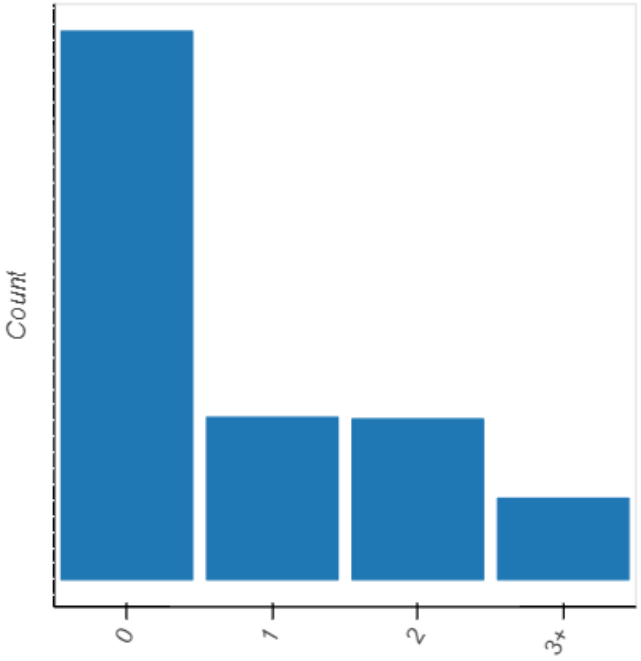




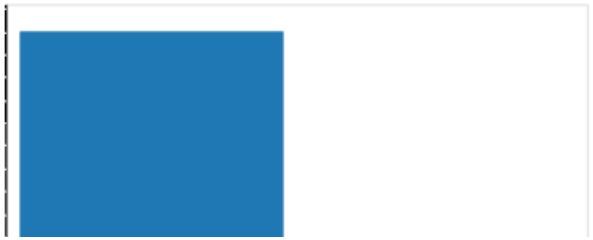
Married



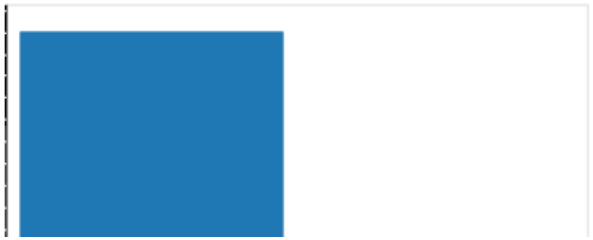
Dependents

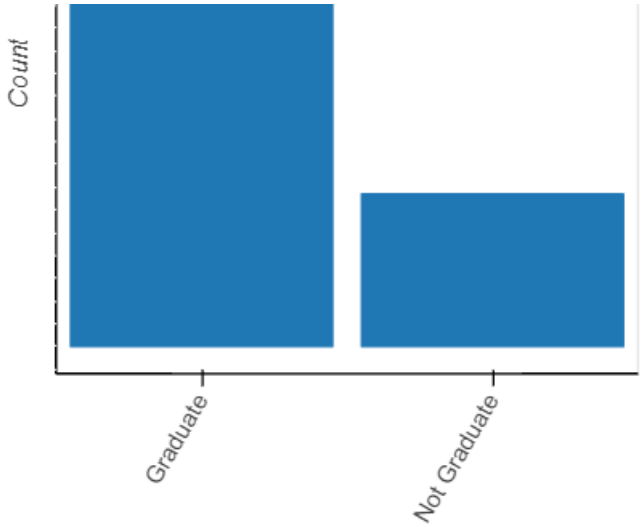


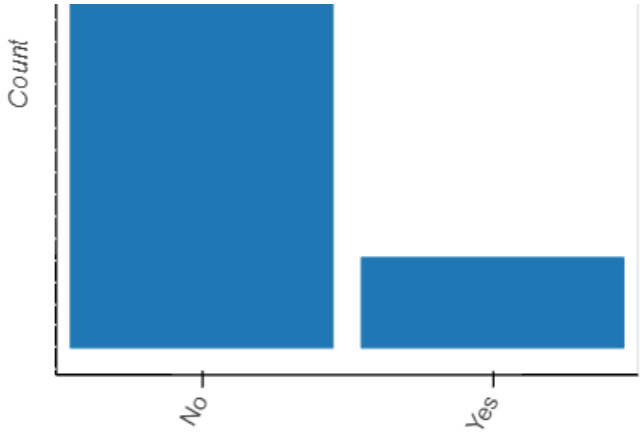
Education



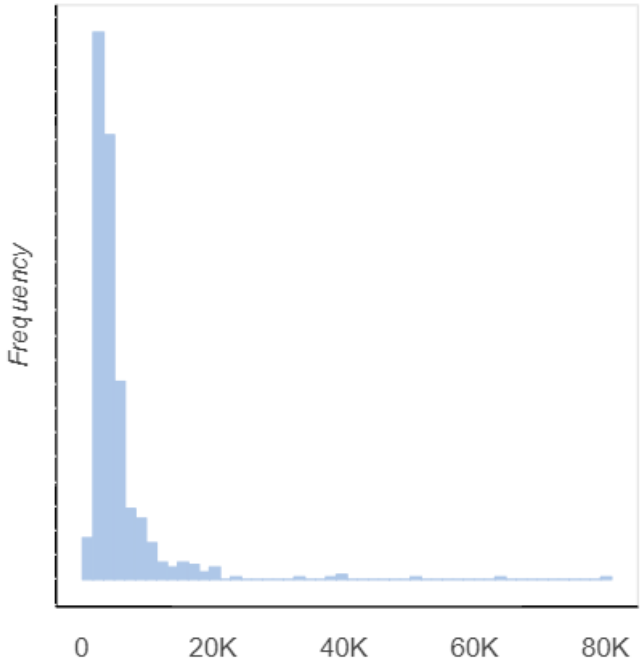
Self_Employed



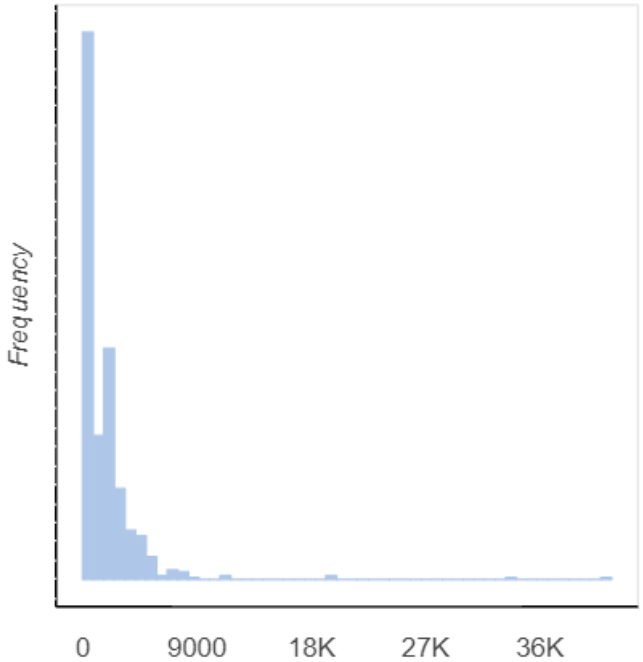




ApplicantIncome

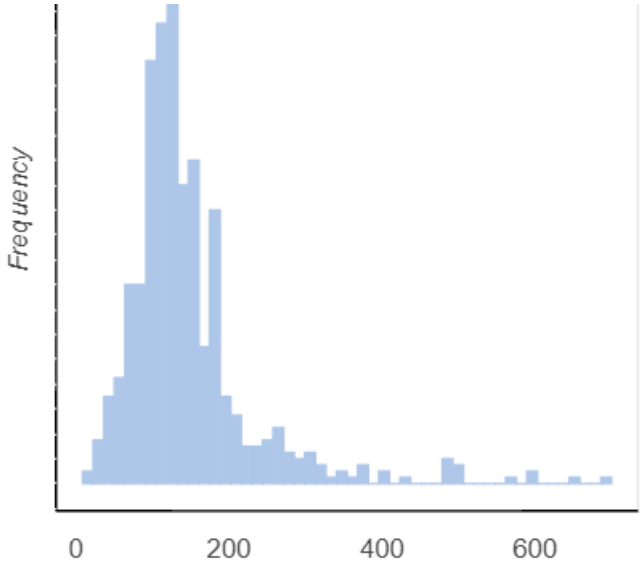


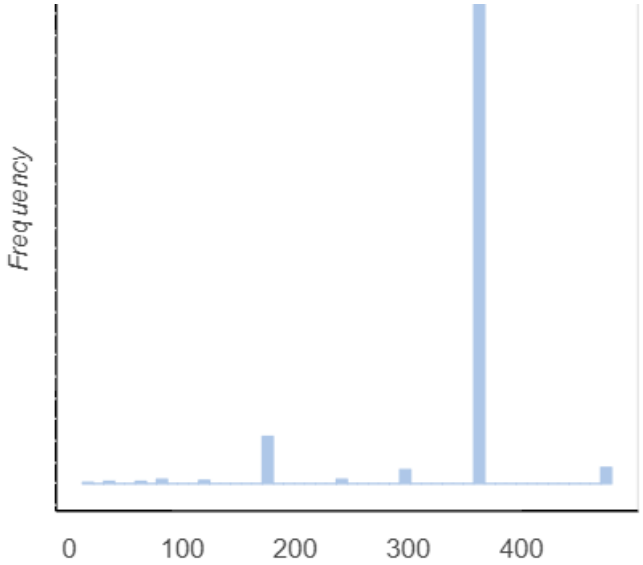
CoapplicantIncome



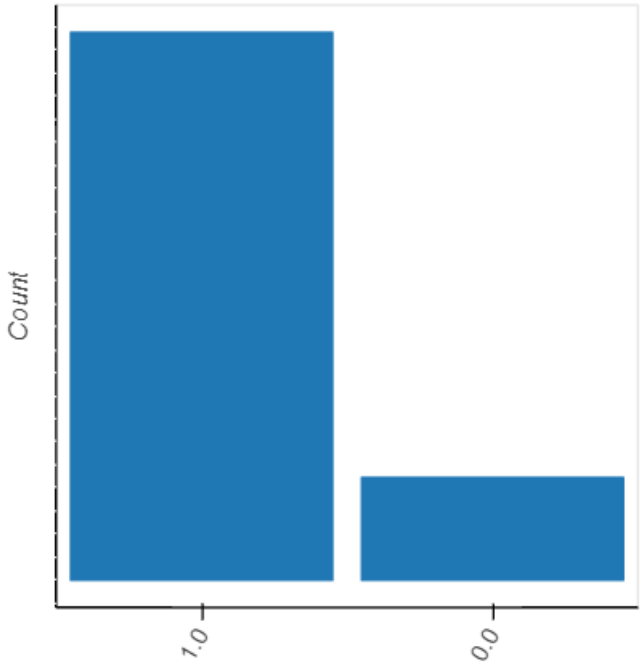
LoanAmount

Loan_Amount_Term

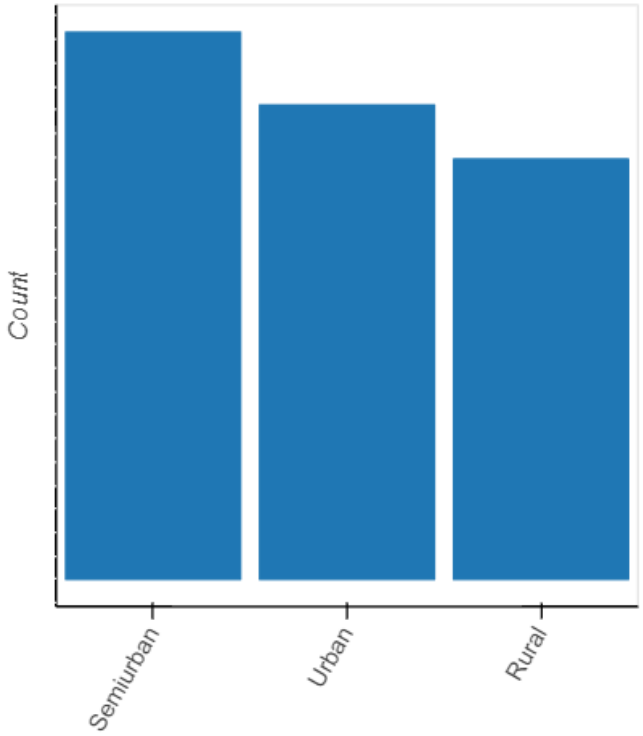


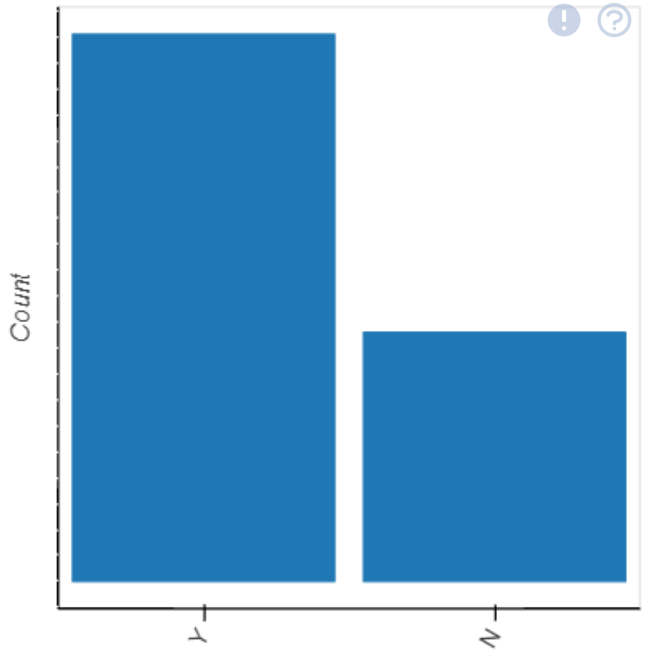


Credit_History



Property_Area





In [28]:
```
create_report(df_train)
```

Out[28]:

DataPrep Report | Overview | Variables ≡ | Interactions | Correlations | Missing Values

Overview

Dataset Statistics

Number of Variables	13
Number of Rows	614
Missing Cells	149

Dataset Insights

Gender	has 13 (2.12%) missing values	Missing
Dependents	has 15 (2.44%) missing values	Missing
Self_Employed	has 32 (5.21%) missing values	Missing

Missing Cells (%)	1.9%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	316.6 KB
Average Row Size in Memory	528.0 B
Variable Types	Categorical: 8 GeoGraphy: 1 Numerical: 4

`LoanAmount` has 22 (3.58%) missing values	Missing
`Loan_Amount_Term` has 14 (2.28%) missing values	Missing
`Credit_History` has 50 (8.14%) missing values	Missing
`ApplicantIncome` is skewed	Skewed
`CoapplicantIncome` is skewed	Skewed
`LoanAmount` is skewed	Skewed
`Loan_Amount_Term` is skewed	Skewed

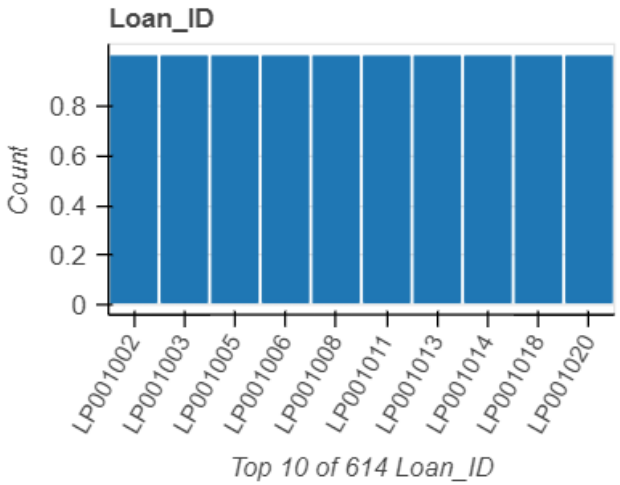
1 2

Variables

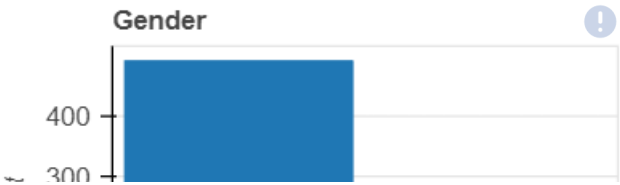
Loan_ID
categorical

Show Details

Approximate Distinct Count	614
Approximate Unique (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory Size	43.8 KB



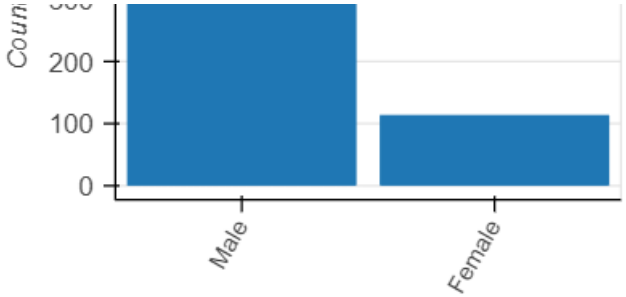
Approximate Distinct Count	2
Approximate Unique (%)	0.3%



Gender
categorical

Show Details

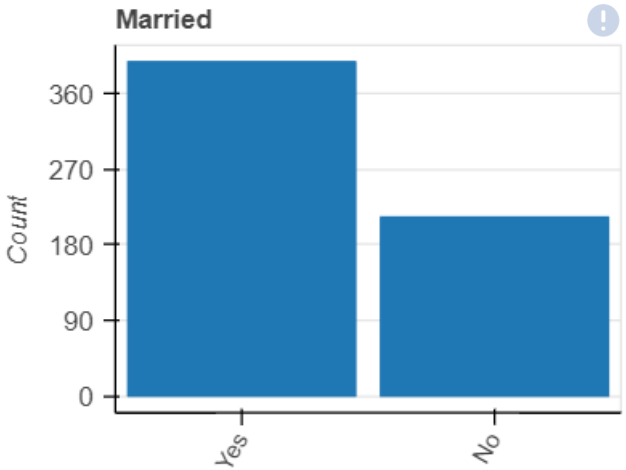
Missing	13
Missing (%)	2.1%
Memory Size	40.7 KB



Married
categorical

Show Details

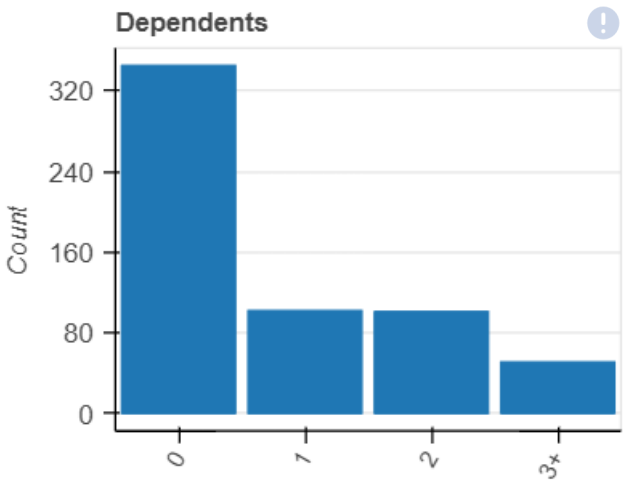
Approximate Distinct Count	2
Approximate Unique (%)	0.3%
Missing	3
Missing (%)	0.5%
Memory Size	40.4 KB



Dependents
categorical

Show Details

Approximate Distinct Count	4
Approximate Unique (%)	0.7%
Missing	15
Missing (%)	2.4%
Memory Size	38.7 KB

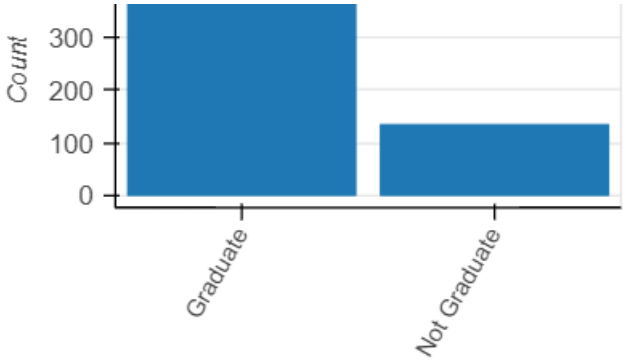


Approximate Distinct Count	2

Education
categorical

Show Details

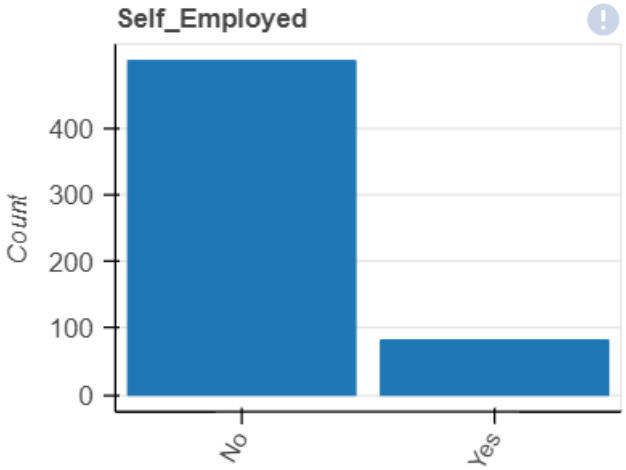
Approximate Unique (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory Size	44.3 KB



Self_Employed
categorical

Show Details

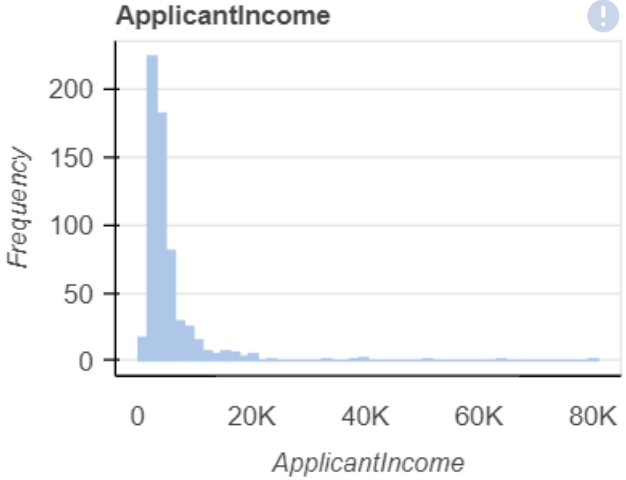
Approximate Distinct Count	2
Approximate Unique (%)	0.3%
Missing	32
Missing (%)	5.2%
Memory Size	38.2 KB



ApplicantIncome
numerical

Show Details

Approximate Distinct Count	505	Mean	5403.4593
Approximate Unique (%)	82.2%	Minimum	150
Missing	0	Maximum	81000
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Negatives	0
Memory Size	9.6 KB	Negatives (%)	0.0%



Approximate	287	Distinct Count	

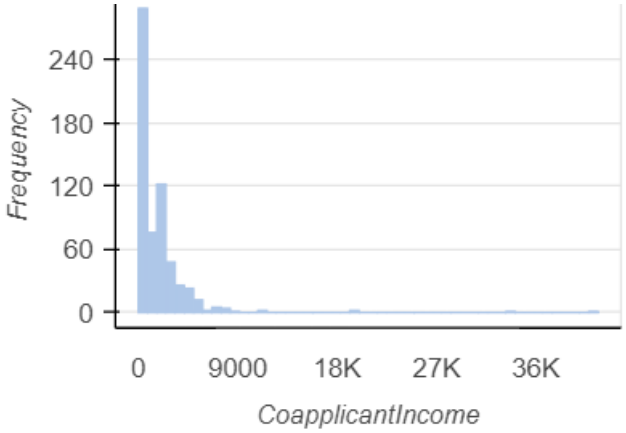
CoapplicantIncome

CoapplicantIncome
numerical

Show Details

Approximate Unique (%)	46.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	9.6 KB
Mean	1621.2458

Minimum	0
Maximum	41667
Zeros	273
Zeros (%)	44.5%
Negatives	0
Negatives (%)	0.0%

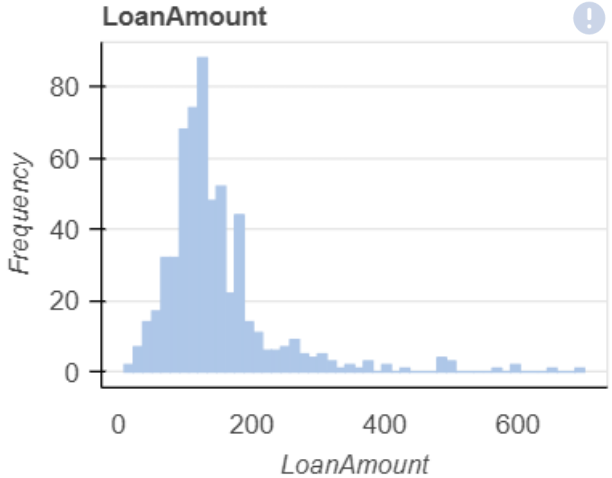


LoanAmount
numerical

Show Details

Approximate Distinct Count	203
Approximate Unique (%)	34.3%
Missing	22
Missing (%)	3.6%
Infinite	0
Infinite (%)	0.0%
Memory Size	9.2 KB

Mean	146.4122
Minimum	9
Maximum	700
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%

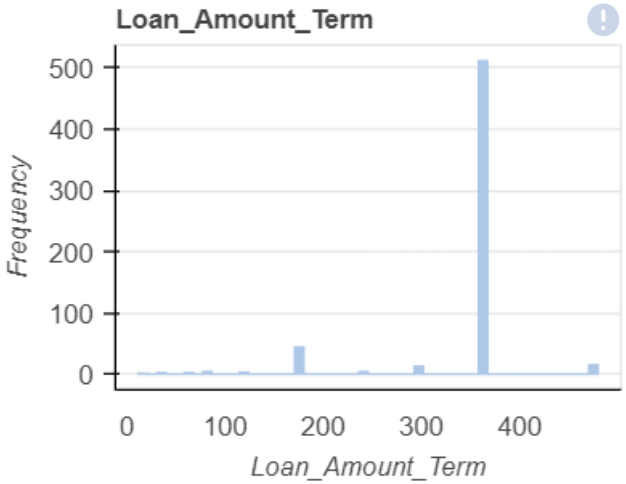


Loan_Amount_Term
numerical

Show Details

Approximate Distinct Count	10
Approximate Unique (%)	1.7%
Missing	14
Missing (%)	2.3%
Infinite	0
Infinite (%)	0.0%
Memory Size	9.4 KB

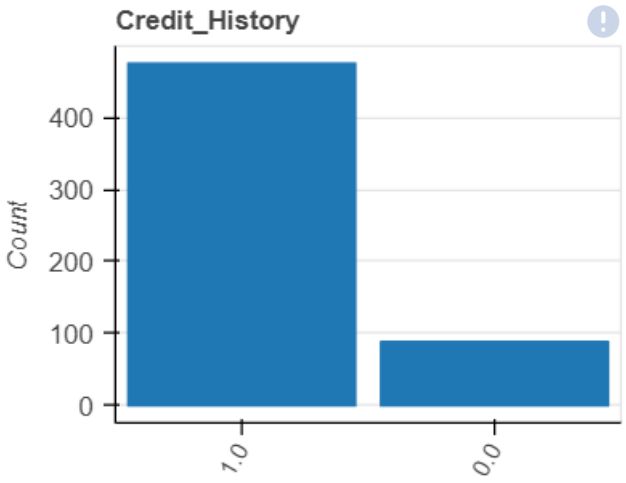
Mean	342
Minimum	12
Maximum	480
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



Credit_History
categorical

Show Details

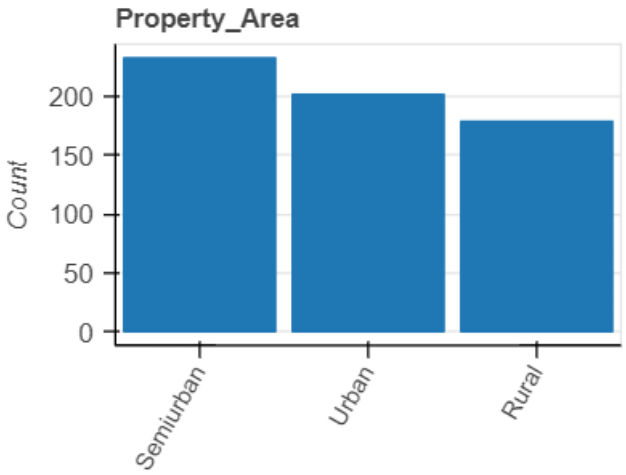
Approximate Distinct Count	2
Approximate Unique (%)	0.4%
Missing	50
Missing (%)	8.1%
Memory Size	37.5 KB



Property_Area
categorical

Show Details

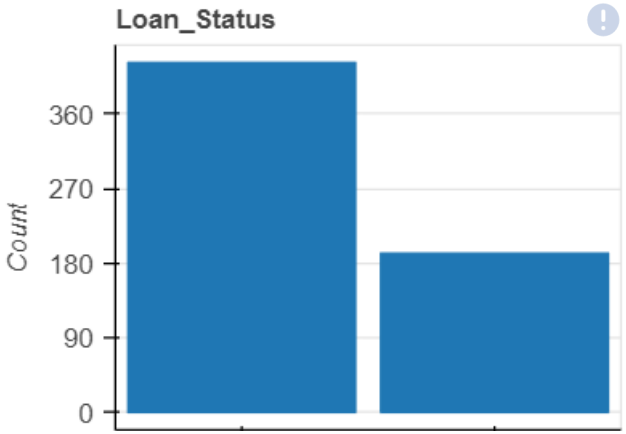
Approximate Distinct Count	3
Approximate Unique (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory Size	42.9 KB



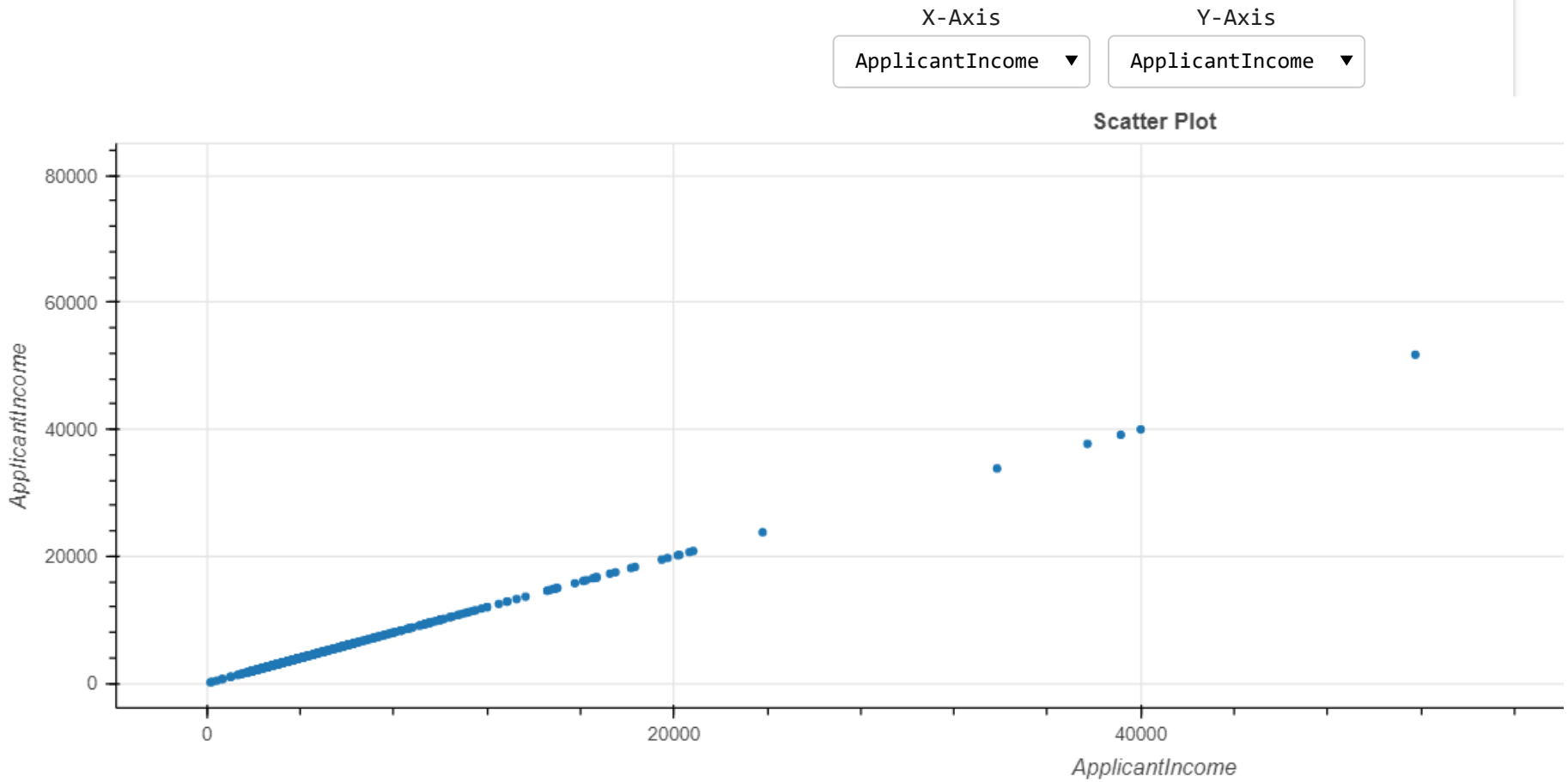
Loan_Status
categorical

Show Details

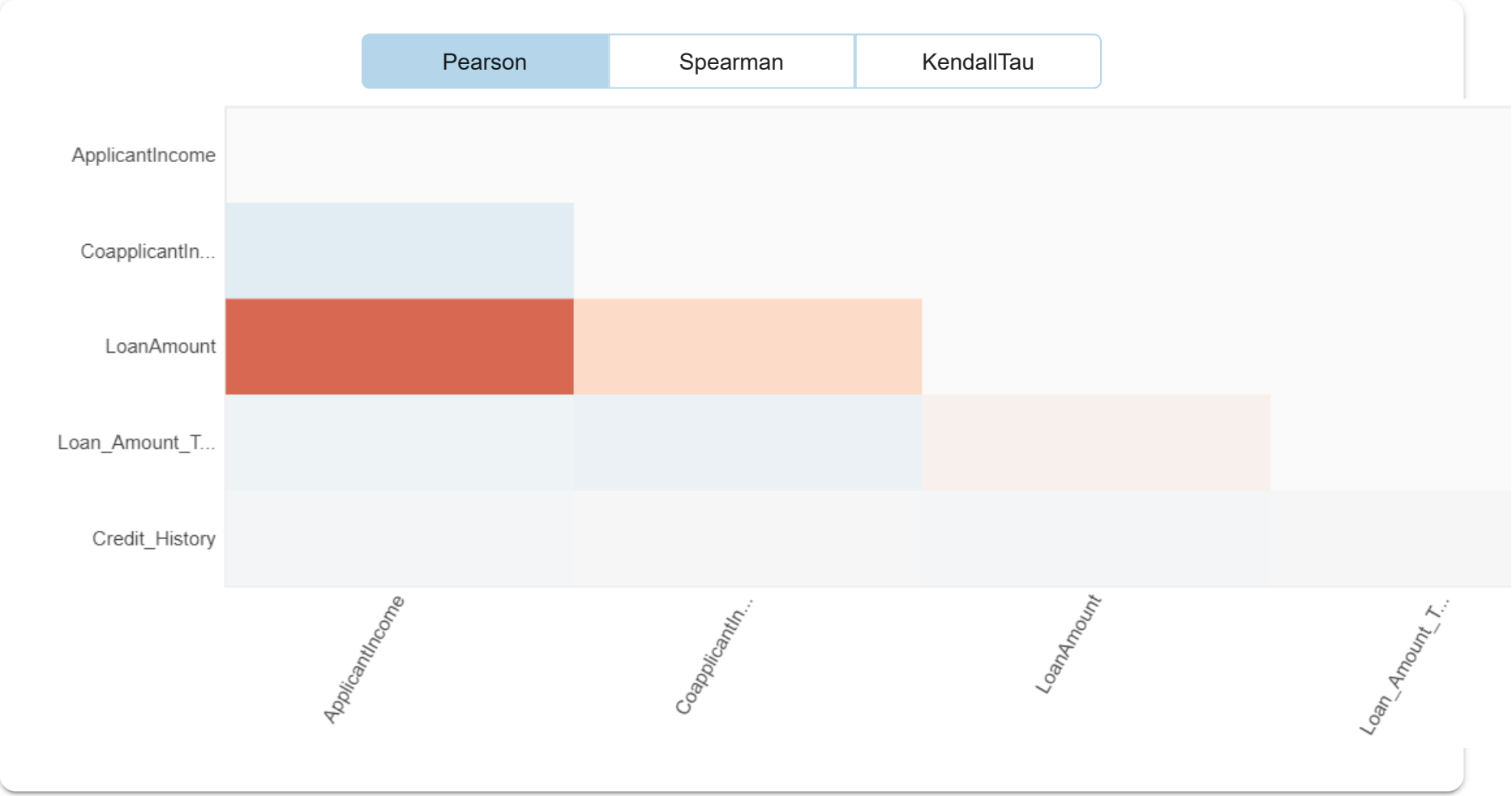
Approximate Distinct Count	2
Approximate Unique (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory Size	39.6 KB



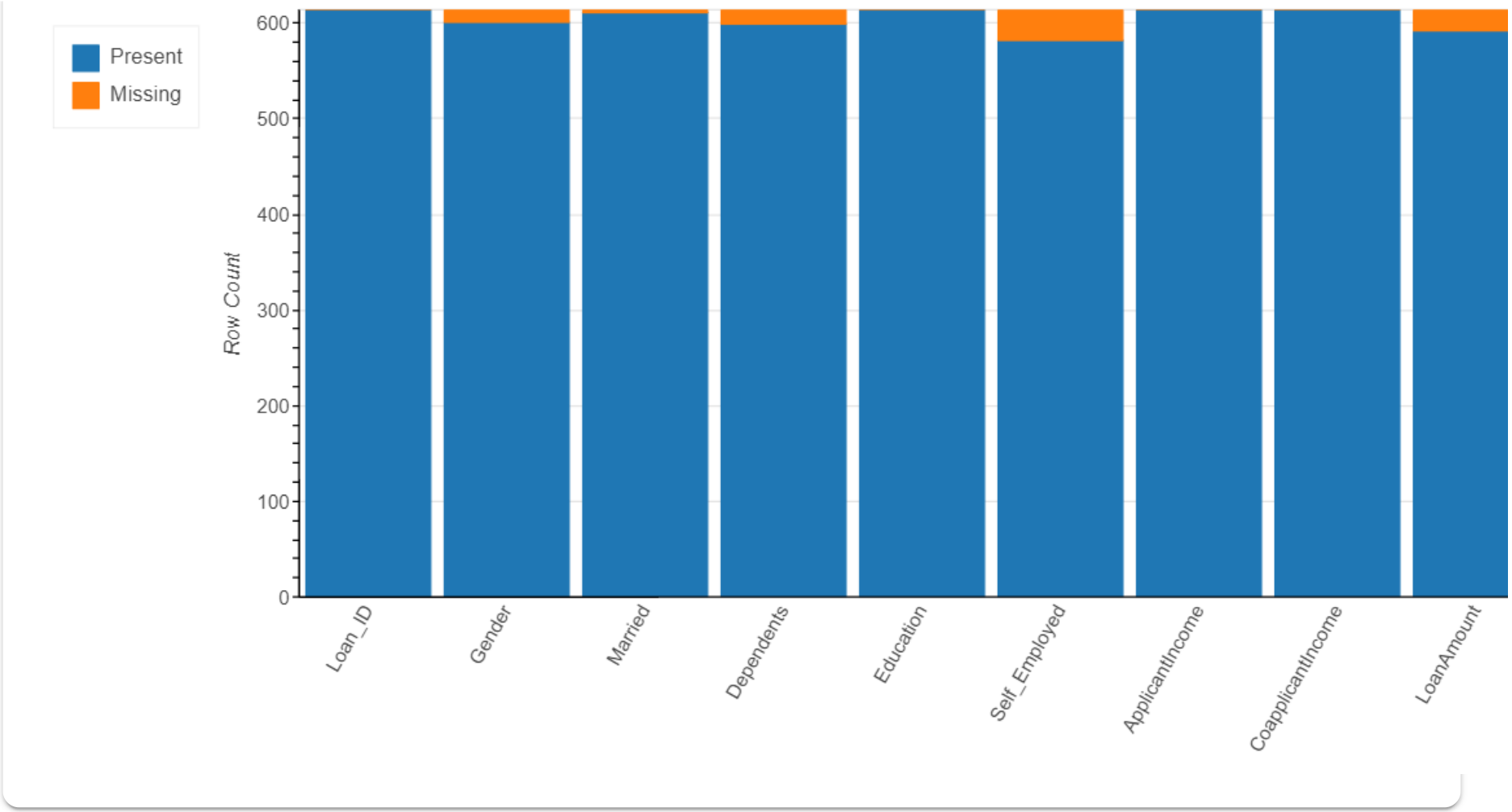
Interactions



Correlations



Missing Values



Report generated with DataPrep

In [29]:
```
plot(df_train, "Property_Area")
```

Out[29]: Stats Bar Chart Pie Chart Word Cloud Word Frequency Word Length Value Table

Overview

Approximate Distinct Count	3
Approximate Unique (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory Size	42.9 KB

Sample

1st row	Urban
2nd row	Rural
3rd row	Urban
4th row	Urban
5th row	Urban

Length

Mean	6.5179
Standard Deviation	1.9426
Median	5
Minimum	5
Maximum	9

Letter

Count	4002
Lowercase Letter	3388
Space Separator	0
Uppercase Letter	614
Dash Punctuation	0
Decimal Number	0

In [ ]: