```
import numpy as np
import seaborn as sb
import pandas as pd
from pandas_profiling import ProfileReport
import plotly.express as px
import plotly.graph_objects as go
from matplotlib import pyplot as plt
```

In [2]:

```
df = pd.read_csv("loan_prediction.csv")
df.head()
```

Out[2]:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y

In [3]:

```
df.info()
```

```
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
```

In [4]:

```
profile = ProfileReport(df, title="Analysis report for data Analysis")
profile.to_notebook_iframe()
profile.to_file("data_analysis.html")
```

```
Summarize dataset:   0%|          | 0/5 [00:00, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00, ?it/s]
Render HTML:   0%|          | 0/1 [00:00, ?it/s]
```

Analysis report for data Analysis

Analysis report for data Analysis

Overview

Dataset statistics

Number of variables	13
Number of observations	614
Missing cells	149
Missing cells (%)	1.9%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	62.5 KiB
Average record size in memory	104.2 B

Variable types

Categorical	6
Boolean	3
Numeric	4

Alerts

Loan_ID has a high cardinality: 614 distinct values	High cardinality
ApplicantIncome is highly correlated with LoanAmount	High correlation
LoanAmount is highly correlated with ApplicantIncome	High correlation
ApplicantIncome is highly correlated with LoanAmount	High correlation
LoanAmount is highly correlated with ApplicantIncome	High correlation
Loan_Status is highly correlated with Credit_History	High correlation
Credit_History is highly correlated with Loan_Status	High correlation
Gender is highly correlated with Married	High correlation
Married is highly correlated with Gender and 1 other fields	High correlation
Dependents is highly correlated with Married	High correlation
ApplicantIncome is highly correlated with LoanAmount	High correlation
LoanAmount is highly correlated with ApplicantIncome	High correlation
Credit_History is highly correlated with Loan_Status	High correlation
Loan_Status is highly correlated with Credit_History	High correlation
Gender has 13 (2.1%) missing values	Missing

Dependents has 15 (2.4%) missing values	Missing
Self_Employed has 32 (5.2%) missing values	Missing
LoanAmount has 22 (3.6%) missing values	Missing
Loan_Amount_Term has 14 (2.3%) missing values	Missing
Credit_History has 50 (8.1%) missing values	Missing
Loan_ID is uniformly distributed	Uniform
Loan_ID has unique values	Unique
CoapplicantIncome has 273 (44.5%) zeros	Zeros

Reproduction

Analysis started	2022-11-09 09:43:02.856825
Analysis finished	2022-11-09 09:43:17.092955
Duration	14.24 seconds
Software version	pandas-profiling v3.2.0
Download configuration	config.json

Variables

Loan_ID
Categorical

HIGH CARDINALITY
UNIFORM
UNIQUE

Distinct	614
Distinct (%)	100.0%

Missing	0
Missing (%)	0.0%
Memory size	4.9 KiB

LP001002	1
LP002328	1
LP002305	1
LP002308	1
LP002314	1
Other values (609)	609

- Overview
- Categories
- Words
- Characters

Length

Max length	8
Median length	8
Mean length	8
Min length	8

Characters and Unicode

Total characters	4912

Distinct characters	12
Distinct categories	2 ?
Distinct scripts	2 ?
Distinct blocks	1 ?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	614 ?
Unique (%)	100.0%

Sample

1st row	LP001002
2nd row	LP001003
3rd row	LP001005
4th row	LP001006
5th row	LP001008

Common Values

Value	Count	Frequency (%)
LP001002	1	0.2%
LP002328	1	0.2%
LP002305	1	

Value	Count	Frequency (%)
		0.2%
LP002308	1	0.2%
LP002314	1	0.2%
LP002315	1	0.2%
LP002317	1	0.2%
LP002318	1	0.2%
LP002319	1	0.2%
LP002332	1	0.2%
Other values (604)	604	98.4%

Length

Histogram of lengths of the category

Value	Count	Frequency (%)
lp001002	1	0.2%
lp001014	1	

Value	Count	Frequency (%)
		0.2%
lp001038	1	0.2%
lp001036	1	0.2%
lp001005	1	0.2%
lp001006	1	0.2%
lp001008	1	0.2%
lp001011	1	0.2%
lp001013	1	0.2%
lp001018	1	0.2%
Other values (604)	604	98.4%

- Characters
- Categories
- Scripts
- Blocks

Most occurring characters

Value	Count	Frequency (%)
0	1403	28.6%
L	614	12.5%
P	614	12.5%
1	491	10.0%
2	478	9.7%
4	203	4.1%
3	198	4.0%
8	189	3.8%
7	183	3.7%
9	182	3.7%
Other values (2)	357	7.3%

Most occurring categories

Value	Count	Frequency (%)
Decimal Number	3684	75.0%
Uppercase Letter	1228	25.0%

Most frequent character per category

Decimal Number

Value	Count	Frequency (%)
0	1403	38.1%
1	491	13.3%
2	478	13.0%
4	203	5.5%
3	198	5.4%
8	189	5.1%
7	183	5.0%
9	182	4.9%
6	181	

Value	Count	Frequency (%)
		4.9%
5	176	4.8%

Uppercase Letter

Value	Count	Frequency (%)
L	614	50.0%
P	614	50.0%

Most occurring scripts

Value	Count	Frequency (%)
Common	3684	75.0%
Latin	1228	25.0%

Most frequent character per script

Common

Value	Count	Frequency (%)
0	1403	38.1%
1	491	13.3%
2	478	13.0%

Value	Count	Frequency (%)
4	203	5.5%
3	198	5.4%
8	189	5.1%
7	183	5.0%
9	182	4.9%
6	181	4.9%
5	176	4.8%

Latin

Value	Count	Frequency (%)
L	614	50.0%
P	614	50.0%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	4912	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
0	1403	28.6%
L	614	12.5%
P	614	12.5%
1	491	10.0%
2	478	9.7%
4	203	4.1%
3	198	4.0%
8	189	3.8%
7	183	3.7%
9	182	3.7%
Other values (2)	357	7.3%

Gender
Categorical

HIGH CORRELATION
MISSING

Distinct	2
Distinct (%)	0.3%
Missing	13
Missing (%)	2.1%
Memory size	4.9 KiB

Male	489
Female	112

Length

Max length	6
Median length	4
Mean length	4.372712146
Min length	4

Characters and Unicode

Total characters	2628
Distinct characters	6
Distinct categories	2 ?
Distinct scripts	1 ?

Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.
Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Male
2nd row	Male
3rd row	Male
4th row	Male
5th row	Male

Common Values

Value	Count	Frequency (%)
Male	489	79.6%
Female	112	18.2%
(Missing)	13	2.1%

Length

Histogram of lengths of the category

Category Frequency Plot

Value	Count	Frequency (%)
male	489	81.4%
female	112	18.6%

- Characters
- Categories
- Scripts
- Blocks

Most occurring characters

Value	Count	Frequency (%)
e	713	27.1%
a	601	22.9%
l	601	22.9%
M	489	18.6%
F	112	4.3%
m	112	4.3%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	2027	77.1%
Uppercase Letter	601	22.9%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	713	35.2%
a	601	29.6%
l	601	29.6%
m	112	5.5%

Uppercase Letter

Value	Count	Frequency (%)
M	489	81.4%
F	112	18.6%

Most occurring scripts

Value	Count	Frequency (%)
Latin	2628	100.0%

Most frequent character per script

Latin

Value	Count	Frequency (%)
e	713	27.1%
a	601	22.9%

Value	Count	Frequency (%)
l	601	22.9%
M	489	18.6%
F	112	4.3%
m	112	4.3%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	2628	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
e	713	27.1%
a	601	22.9%
l	601	22.9%
M	489	18.6%
F	112	4.3%
m	112	4.3%

Married
Boolean

HIGH CORRELATION

Distinct	2
Distinct (%)	0.3%
Missing	3
Missing (%)	0.5%
Memory size	1.3 KiB
True	398
False	213
(Missing)	3

- Common Values
- Category Frequency Plot

Value	Count	Frequency (%)
True	398	64.8%
False	213	34.7%
(Missing)	3	0.5%

Dependents
Categorical

HIGH CORRELATION
MISSING

Distinct	4

Distinct (%)	0.7%
Missing	15
Missing (%)	2.4%
Memory size	4.9 KiB

0	345
1	102
2	101
3+	51

- Overview
- Categories
- Words
- Characters

Length

Max length	2
Median length	1
Mean length	1.085141903
Min length	1

Characters and Unicode

Total characters	650
Distinct characters	5
Distinct categories	2 ?
Distinct scripts	1 ?

Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.
Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	0
2nd row	1
3rd row	0
4th row	0
5th row	0

Common Values

Value	Count	Frequency (%)
0	345	56.2%
1	102	16.6%
2	101	16.4%
3+	51	8.3%
(Missing)	15	2.4%

Length

Histogram of lengths of the category

Category Frequency Plot

Value	Count	Frequency (%)
0	345	57.6%
1	102	17.0%
2	101	16.9%
3	51	8.5%

- Characters
- Categories
- Scripts
- Blocks

Most occurring characters

Value	Count	Frequency (%)
0	345	53.1%
1	102	15.7%
2	101	15.5%
3	51	7.8%

Value	Count	Frequency (%)
+	51	7.8%

Most occurring categories

Value	Count	Frequency (%)
Decimal Number	599	92.2%
Math Symbol	51	7.8%

Most frequent character per category

Decimal Number

Value	Count	Frequency (%)
0	345	57.6%
1	102	17.0%
2	101	16.9%
3	51	8.5%

Math Symbol

Value	Count	Frequency (%)
+	51	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Common	650	100.0%

Most frequent character per script

Common

Value	Count	Frequency (%)
0	345	53.1%
1	102	15.7%
2	101	15.5%
3	51	7.8%
+	51	7.8%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	650	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
0	345	53.1%
1	102	15.7%
2	101	15.5%
3	51	7.8%
+	51	7.8%

Education
Categorical

Distinct	2
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	4.9 KiB

Graduate	480
Not Graduate	134

- Overview
- Categories
- Words
- Characters

Length

Max length	12
Median length	8
Mean length	8.872964169
Min length	8

Characters and Unicode

Total characters	5448
Distinct characters	10
Distinct categories	3 ?
Distinct scripts	2 ?
Distinct blocks	1 ?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0 ?
Unique (%)	0.0%

Sample

1st row	Graduate
2nd row	Graduate
3rd row	Graduate
4th row	Not Graduate
5th row	Graduate

Common Values

Value	Count	Frequency (%)
Graduate	480	78.2%
Not Graduate	134	21.8%

Length

Histogram of lengths of the category

Category Frequency Plot

Value	Count	Frequency (%)
graduate	614	82.1%
not	134	17.9%

- Characters
- Categories
- Scripts
- Blocks

Most occurring characters

Value	Count	Frequency (%)
a	1228	22.5%
t	748	13.7%
G	614	11.3%
r	614	11.3%
d	614	11.3%

Value	Count	Frequency (%)
u	614	11.3%
e	614	11.3%
N	134	2.5%
o	134	2.5%
	134	2.5%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	4566	83.8%
Uppercase Letter	748	13.7%
Space Separator	134	2.5%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
a	1228	26.9%
t	748	16.4%

Value	Count	Frequency (%)
r	614	13.4%
d	614	13.4%
u	614	13.4%
e	614	13.4%
o	134	2.9%

Uppercase Letter

Value	Count	Frequency (%)
G	614	82.1%
N	134	17.9%

Space Separator

Value	Count	Frequency (%)
	134	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	5314	97.5%
Common	134	2.5%

Most frequent character per script

Latin

Value	Count	Frequency (%)
a	1228	23.1%
t	748	14.1%
G	614	11.6%
r	614	11.6%
d	614	11.6%
u	614	11.6%
e	614	11.6%
N	134	2.5%
o	134	2.5%

Common

Value	Count	Frequency (%)
	134	100.0%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	5448	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
a	1228	22.5%
t	748	13.7%
G	614	11.3%
r	614	11.3%
d	614	11.3%
u	614	11.3%
e	614	11.3%
N	134	2.5%
o	134	2.5%
	134	2.5%

Self Employed

Boolean

MISSING

Distinct	2
Distinct (%)	0.3%
Missing	32
Missing (%)	5.2%
Memory size	1.3 KiB

False	500
True	82
(Missing)	32

- Common Values
- Category Frequency Plot

Value	Count	Frequency (%)
False	500	81.4%
True	82	13.4%
(Missing)	32	5.2%

ApplicantIncome

Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	505
Distinct (%)	82.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	5403.459283

Minimum	150
Maximum	81000
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	4.9 KiB

- Statistics
- Histogram
- Common values
- Extreme values

Quantile statistics

Minimum	150
5-th percentile	1897.55
Q1	2877.5
median	3812.5
Q3	5795
95-th percentile	14583
Maximum	81000
Range	80850
Interquartile range (IQR)	2917.5

Descriptive statistics

Standard deviation	6109.041673
Coefficient of variation (CV)	1.130579755
Kurtosis	60.54067593
Mean	5403.459283
Median Absolute Deviation (MAD)	1229.5
Skewness	6.539513114
Sum	3317724
Variance	37320390.17
Monotonicity	Not monotonic

Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
2500	9	1.5%
4583	6	1.0%
6000	6	1.0%
2600	6	1.0%
3333	5	0.8%
4166	5	0.8%

Value	Count	Frequency (%)
3750	5	0.8%
5000	5	0.8%
8333	4	0.7%
6250	4	0.7%
Other values (495)	559	91.0%

- Minimum 10 values
- Maximum 10 values

Value	Count	Frequency (%)
150	1	0.2%
210	1	0.2%
416	1	0.2%
645	1	0.2%
674	1	0.2%
1000	1	0.2%
1025	2	0.3%
1299	1	0.2%
1378	1	0.2%

Value	Count	Frequency (%)
1442	1	0.2%

Value	Count	Frequency (%)
81000	1	0.2%
63337	1	0.2%
51763	1	0.2%
39999	1	0.2%
39147	1	0.2%
37719	1	0.2%
33846	1	0.2%
23803	1	0.2%
20833	1	0.2%
20667	1	0.2%

CoapplicantIncome

Real number (ℝ≥0)

ZEROS

Distinct	287
Distinct (%)	46.7%
Missing	0
Missing (%)	0.0%
Infinite	0

Infinite (%)	0.0%
Mean	1621.245798
Minimum	0
Maximum	41667
Zeros	273
Zeros (%)	44.5%
Negative	0
Negative (%)	0.0%
Memory size	4.9 KiB

- Statistics
- Histogram
- Common values
- Extreme values

Quantile statistics

Minimum	0
5-th percentile	0
Q1	0
median	1188.5
Q3	2297.25
95-th percentile	4997.4
Maximum	41667
Range	41667
Interquartile range (IQR)	2297.25

Descriptive statistics

Standard deviation	2926.248369
Coefficient of variation (CV)	1.804938136
Kurtosis	84.95638421
Mean	1621.245798
Median Absolute Deviation (MAD)	1188.5
Skewness	7.491531217
Sum	995444.92
Variance	8562929.518
Monotonicity	Not monotonic

Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
0	273	44.5%
2500	5	0.8%
2083	5	0.8%
1666	5	0.8%
2250	3	0.5%
1750	3	

Value	Count	Frequency (%)
		0.5%
1800	3	0.5%
1625	3	0.5%
2333	3	0.5%
1459	3	0.5%
Other values (277)	308	50.2%

- Minimum 10 values
- Maximum 10 values

Value	Count	Frequency (%)
0	273	44.5%
16.12000084	1	0.2%
189	1	0.2%
240	1	0.2%
242	1	0.2%

Value	Count	Frequency (%)
461	1	0.2%
484	1	0.2%
505	1	0.2%
536	1	0.2%
663	1	0.2%

Value	Count	Frequency (%)
41667	1	0.2%
33837	1	0.2%
20000	2	0.3%
11300	1	0.2%
10968	1	0.2%
8980	1	0.2%
8333	1	0.2%
8106	1	0.2%
7873	1	0.2%
7750	1	0.2%

LoanAmount

Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING

Distinct	203
Distinct (%)	34.3%
Missing	22
Missing (%)	3.6%
Infinite	0
Infinite (%)	0.0%
Mean	146.4121622
Minimum	9
Maximum	700
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	4.9 KiB

Quantile statistics

Minimum	9

5-th percentile	56
Q1	100
median	128
Q3	168
95-th percentile	297.8
Maximum	700
Range	691
Interquartile range (IQR)	68

Descriptive statistics

Standard deviation	85.58732524
Coefficient of variation (CV)	0.5845643147
Kurtosis	10.40153349
Mean	146.4121622
Median Absolute Deviation (MAD)	32
Skewness	2.677551679
Sum	86676
Variance	7325.190241
Monotonicity	Not monotonic

Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
120	20	

Value	Count	Frequency (%)
		3.3%
110	17	2.8%
100	15	2.4%
160	12	2.0%
187	12	2.0%
113	11	1.8%
128	11	1.8%
130	10	1.6%
95	9	1.5%
96	9	1.5%
Other values (193)	466	75.9%
(Missing)	22	3.6%

- Minimum 10 values
- Maximum 10 values

Value	Count	Frequency (%)
9	1	0.2%
17	1	0.2%
25	2	0.3%
26	1	0.2%
30	2	0.3%
35	1	0.2%
36	1	0.2%
40	2	0.3%
42	1	0.2%
44	2	0.3%

Value	Count	Frequency (%)
700	1	0.2%
650	1	0.2%
600	2	0.3%
570	1	0.2%
500	1	

Value	Count	Frequency (%)
		0.2%
496	1	0.2%
495	1	0.2%
490	1	0.2%
480	3	0.5%
436	1	0.2%

Loan_Amount_Term

Real number (ℝ≥0)

MISSING

Distinct	10
Distinct (%)	1.7%
Missing	14
Missing (%)	2.3%
Infinite	0
Infinite (%)	0.0%
Mean	342
Minimum	12
Maximum	480

Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	4.9 KiB

- Statistics
- Histogram
- Common values
- Extreme values

Quantile statistics

Minimum	12
5-th percentile	180
Q1	360
median	360
Q3	360
95-th percentile	360
Maximum	480
Range	468
Interquartile range (IQR)	0

Descriptive statistics

Standard deviation	65.12040985
Coefficient of variation (CV)	0.1904105551

Kurtosis	6.673473693
Mean	342
Median Absolute Deviation (MAD)	0
Skewness	-2.362414124
Sum	205200
Variance	4240.66778
Monotonicity	Not monotonic

Histogram with fixed size bins (bins=10)

Value	Count	Frequency (%)
360	512	83.4%
180	44	7.2%
480	15	2.4%
300	13	2.1%
240	4	0.7%
84	4	0.7%
120	3	0.5%
60	2	

Value	Count	Frequency (%)
		0.3%
36	2	0.3%
12	1	0.2%
(Missing)	14	2.3%

- Minimum 10 values
- Maximum 10 values

Value	Count	Frequency (%)
12	1	0.2%
36	2	0.3%
60	2	0.3%
84	4	0.7%
120	3	0.5%
180	44	7.2%

Value	Count	Frequency (%)
240	4	0.7%
300	13	2.1%
360	512	83.4%
480	15	2.4%

Value	Count	Frequency (%)
480	15	2.4%
360	512	83.4%
300	13	2.1%
240	4	0.7%
180	44	7.2%
120	3	0.5%
84	4	0.7%
60	2	

Value	Count	Frequency (%)
		0.3%
36	2	0.3%
12	1	0.2%

Credit_History
Categorical

HIGH CORRELATION
HIGH CORRELATION
MISSING

Distinct	2
Distinct (%)	0.4%
Missing	50
Missing (%)	8.1%
Memory size	4.9 KiB

1.0	475
0.0	89

- Overview
- Categories
- Words
- Characters

Length

Max length	3
Median length	3

Mean length	3
Min length	3

Characters and Unicode

Total characters	1692
Distinct characters	3
Distinct categories	2 ?
Distinct scripts	1 ?
Distinct blocks	1 ?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0 ?
Unique (%)	0.0%

Sample

1st row	1.0
2nd row	1.0
3rd row	1.0
4th row	1.0
5th row	1.0

Common Values

Value	Count	Frequency (%)
1.0	475	77.4%
0.0	89	14.5%
(Missing)	50	8.1%

Length

Histogram of lengths of the category

Category Frequency Plot

Value	Count	Frequency (%)
1.0	475	84.2%
0.0	89	15.8%

- Characters
- Categories
- Scripts
- Blocks

Most occurring characters

Value	Count	Frequency (%)
0	653	38.6%
.	564	33.3%
1	475	28.1%

Most occurring categories

Value	Count	Frequency (%)
Decimal Number	1128	66.7%
Other Punctuation	564	33.3%

Most frequent character per category

Decimal Number

Value	Count	Frequency (%)
0	653	57.9%
1	475	42.1%

Other Punctuation

Value	Count	Frequency (%)
.	564	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Common	1692	100.0%

Most frequent character per script

Common

Value	Count	Frequency (%)
0	653	38.6%
.	564	33.3%

Value	Count	Frequency (%)
1	475	28.1%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	1692	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
0	653	38.6%
.	564	33.3%
1	475	28.1%

Property Area
Categorical

Distinct	3
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory size	4.9 KiB
Semiurban	233
Urban	202

Rural	179

- Overview
- Categories
- Words
- Characters

Length

Max length	9
Median length	5
Mean length	6.517915309
Min length	5

Characters and Unicode

Total characters	4002
Distinct characters	12
Distinct categories	2 ?
Distinct scripts	1 ?
Distinct blocks	1 ?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0 ?
Unique (%)	0.0%

Sample

1st row	Urban
2nd row	Rural

3rd row	Urban
4th row	Urban
5th row	Urban

Common Values

Value	Count	Frequency (%)
Semiurban	233	37.9%
Urban	202	32.9%
Rural	179	29.2%

Length

Histogram of lengths of the category

Category Frequency Plot

Value	Count	Frequency (%)
semiurban	233	37.9%
urban	202	32.9%
rural	179	29.2%

- Characters
- Categories
- Scripts
- Blocks

Most occurring characters

Value	Count	Frequency (%)
r	614	15.3%
a	614	15.3%
b	435	10.9%
n	435	10.9%
u	412	10.3%
S	233	5.8%
e	233	5.8%
m	233	5.8%
i	233	5.8%
U	202	5.0%
Other values (2)	358	8.9%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	3388	84.7%
Uppercase Letter	614	15.3%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
r	614	18.1%
a	614	18.1%
b	435	12.8%
n	435	12.8%
u	412	12.2%
e	233	6.9%
m	233	6.9%
i	233	6.9%
l	179	5.3%

Uppercase Letter

Value	Count	Frequency (%)
S	233	37.9%
U	202	32.9%
R	179	29.2%

Most occurring scripts

Value	Count	Frequency (%)
Latin	4002	100.0%

Most frequent character per script

Latin

Value	Count	Frequency (%)
r	614	15.3%
a	614	15.3%
b	435	10.9%
n	435	10.9%
u	412	10.3%
S	233	5.8%
e	233	5.8%
m	233	5.8%
i	233	5.8%
U	202	5.0%
Other values (2)	358	8.9%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	4002	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
r	614	15.3%
a	614	15.3%
b	435	10.9%
n	435	10.9%
u	412	10.3%
S	233	5.8%
e	233	5.8%
m	233	5.8%
i	233	5.8%
U	202	5.0%
Other values (2)	358	8.9%

Loan_Status
Boolean

HIGH CORRELATION
HIGH CORRELATION

Distinct	2
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	742.0 B

True	422
False	192

- Common Values
- Category Frequency Plot

Value	Count	Frequency (%)
True	422	68.7%
False	192	31.3%

Interactions

- ApplicantIncome
- CoapplicantIncome
- LoanAmount
- Loan_Amount_Term

- Loan_Amount_Term
- ApplicantIncome
- CoapplicantIncome
- LoanAmount

- Loan_Amount_Term
- ApplicantIncome
- CoapplicantIncome
- LoanAmount

- Loan_Amount_Term
- ApplicantIncome
- CoapplicantIncome
- LoanAmount

- Loan_Amount_Term
- ApplicantIncome
- CoapplicantIncome
- LoanAmount

Correlations

- Spearman's ρ
- Pearson's r
- Kendall's τ
- Cramér's V (φc)
- Phik (φk)

Spearman's ρ

The Spearman's rank correlation coefficient (ρ) is a measure of monotonic correlation between two variables, and is therefore better in catching nonlinear monotonic correlations than Pearson's *r*. It's value lies between -1 and +1, -1 indicating total negative monotonic correlation, 0 indicating no monotonic correlation and 1 indicating total positive monotonic correlation.

To calculate ρ for two variables *X* and *Y*, one divides the covariance of the rank variables of *X* and *Y* by the product of their standard deviations.

Pearson's r

The Pearson's correlation coefficient (r) is a measure of linear correlation between two variables. It's value lies between -1 and +1, -1 indicating total negative linear correlation, 0 indicating no linear correlation and 1 indicating total positive linear correlation. Furthermore, r is invariant under separate changes in location and scale of the two variables, implying that for a linear function the angle to the x-axis does not affect r.

To calculate r for two variables X and Y, one divides the covariance of X and Y by the product of their standard deviations.

Kendall's τ

Similarly to Spearman's rank correlation coefficient, the Kendall rank correlation coefficient (τ) measures ordinal association between two variables. It's value lies between -1 and +1, -1 indicating total negative correlation, 0 indicating no correlation and 1 indicating total positive correlation.

To calculate τ for two variables X and Y, one determines the number of concordant and discordant pairs of observations. τ is given by the number of concordant pairs minus the discordant pairs divided by the total number of pairs.

Cramér's V (φc)

Cramér's V is an association measure for nominal random variables. The coefficient ranges from 0 to 1, with 0 indicating independence and 1 indicating perfect association. The empirical estimators used for Cramér's V have been proved to be biased, even for large samples. We use a bias-corrected measure that has been proposed by Bergsma in 2013 that can be found here.

Phik (φk)

Phik (φk) is a new and practical correlation coefficient that works consistently between categorical, ordinal and interval variables, captures non-linear dependency and reverts to the Pearson correlation coefficient in case of a bivariate normal input distribution. There is extensive documentation available here.

Missing values

- Count
- Matrix
- Heatmap
- Dendrogram

A simple visualization of nullity by column.

Nullity matrix is a data-dense display which lets you quickly visually pick out patterns in data completion.

The correlation heatmap measures nullity correlation: how strongly the presence or absence of one variable affects the presence of another.
The dendrogram allows you to more fully correlate variable completion, revealing trends deeper than the pairwise ones visible in the correlation heatmap.

Sample

First rows

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y
5	LP001011	Male	Yes	2	Graduate	Yes	5417	4196.0	267.0	360.0	1.0	Urban	Y

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
6	LP001013	Male	Yes	0	Not Graduate	No	2333	1516.0	95.0	360.0	1.0	Urban	Y
7	LP001014	Male	Yes	3+	Graduate	No	3036	2504.0	158.0	360.0	0.0	Semiurban	N
8	LP001018	Male	Yes	2	Graduate	No	4006	1526.0	168.0	360.0	1.0	Urban	Y
9	LP001020	Male	Yes	1	Graduate	No	12841	10968.0	349.0	360.0	1.0	Semiurban	N

Last rows

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
604	LP002959	Female	Yes	1	Graduate	No	12000	0.0	496.0	360.0	1.0	Semiurban	Y
605	LP000	Ma	Yes	0	NotG	No	2400	38000	NaN	18	1.0	Urb	N

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
	2960	le			raduate			.0		0.0		an	
606	LP002961	Male	Yes	1	Graduate	No	3400	2500.0	173.0	360.0	1.0	Semiurban	Y
607	LP002964	Male	Yes	2	NotGraduate	No	3987	1411.0	157.0	360.0	1.0	Rural	Y
608	LP002974	Male	Yes	0	Graduate	No	3232	1950.0	108.0	360.0	1.0	Rural	Y

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed		ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
609	LP002978	Female	No	0	Graduate	No		2900	0.0	71.0	360.0	1.0	Rural	Y
610	LP002979	Male	Yes	3+	Graduate	No		4106	0.0	40.0	180.0	1.0	Rural	Y
611	LP002983	Male	Yes	1	Graduate	No		8072	240.0	253.0	360.0	1.0	Urban	Y
612	LP002984	Male	Yes	2	Graduate	No		7583	0.0	187.0	360.0	1.0	Urban	Y
613	LP002	Fema	No	0	Gradu	Yes	45**import** numpy **as** np **import** seaborn **as** sb **import** pandas **as** pd **from** pandas_profiling **import** ProfileReport		0.0	133	360.0	0.0	Semiu	

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
990	le			ate				.0			rban	

```
import plotly.express as px
import plotly.graph_objects as go
from matplotlib import pyplot as plt
```

In [2]:

```
df =
pd.read_csv("loan_prediction.csv")
df.head()
```

Out[2]:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	Y
1	LP001010	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed		ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
	03				te									
2	LP001005	Male	Yes	0	Graduate	Yes		3000	0.0	66.0	360.0	1.0	Urban	Y
3	LP001006	Male	Yes	0	NotGraduate	No		2583	2358.0	120.0	360.0	1.0	Urban	Y
4	LP001008	Male	No	0	Graduate	No		6000	0.0	141.0	360.0	1.0	Urban	Y

In [3]:

```
df.info()
```

```
RangeIndex: 614 entries, 0 to 613
Data columns (total 13 columns):
 #   Column             Non-Null Count
Dtype
```

Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

```
---  ------             --------------
-----
 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
```

In [4]:

```
profile = ProfileReport(df,
title="Analysis report for data
Analysis")
profile.to_notebook_iframe()
profile.to_file("data_analysis.html")
```

```
Summarize dataset:   0%|          | 0/
5 [00:00, ?it/s]
```

Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

```
Generate report structure:   0%|
| 0/1 [00:00, ?it/s]
Render HTML:   0%|          | 0/1 [00:
00, ?it/s]
```

Analysis report for data Analysis

Analysis report for data Analysis

Overview

Dataset statistics

Number of variables	13
Number of observations	614
Missing cells	149
Missing cells (%)	1.9%
Duplicate rows	0

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Duplicate rows (%)	0.0%
Total size in memory	62.5 KiB
Average record size in memory	104.2 B

Variable types

Categorical	6
Boolean	3
Numeric	4

Alerts

Loan_ID has a high cardinality: 614 distinct values	High cardinality
ApplicantIncome is highly correlated with LoanAmount	High correlation
LoanAmount is highly correlated with ApplicantIncome	High correlation
ApplicantIncome is highly correlated with LoanAmount	High correlation
LoanAmount is highly correlated with ApplicantIncome	High correlation

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Loan_Status is highly correlated with Credit_History	High correlation
Credit_History is highly correlated with Loan_Status	High correlation
Gender is highly correlated with Married	High correlation
Married is highly correlated with Gender and 1 other fields	High correlation
Dependents is highly correlated with Married	High correlation
ApplicantIncome is highly correlated with LoanAmount	High correlation
LoanAmount is highly correlated with ApplicantIncome	High correlation
Credit_History is highly correlated with Loan_Status	High correlation
Loan_Status is highly correlated with Credit_History	High correlation
Gender has 13 (2.1%) missing values	Missing

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Dependents has 15 (2.4%) missing values	Missing
Self_Employed has 32 (5.2%) missing values	Missing
LoanAmount has 22 (3.6%) missing values	Missing
Loan_Amount_Term has 14 (2.3%) missing values	Missing
Credit_History has 50 (8.1%) missing values	Missing
Loan_ID is uniformly distributed	Uniform
Loan_ID has unique values	Unique
CoapplicantIncome has 273 (44.5%) zeros	Zeros

Reproduction

Analysis started	2022-11-09 09:43:02.856825
Analysis finished	2022-11-09 09:43:17.092955
Duration	14.24 seconds

Loan_ID
Gender
Married
Dependents
Education
Self_Employed
ApplicantIncome
CoapplicantIncome
LoanAmount
Loan_Amount_Term
Credit_History
Property_Area
Loan_Status

Software version	pandas-profiling v3.2.0
Download configuration	config.json

Variables

Loan_ID

Categorical

HIGH CARDINALITY
UNIFORM
UNIQUE

Distinct	614
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory size	4.9 KiB

LP001002	1
LP002328	1

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP002305						1						
LP002308						1						
LP002314						1						
Other values (609)						609						

- Overview
- Categories
- Words
- Characters

Length

Max length	8
Median length	8
Mean length	8
Min length	8

Characters and Unicode

Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

Total characters	4912
Distinct characters	12
Distinct categories	2 ?
Distinct scripts	2 ?
Distinct blocks	1 ?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	614 ?
Unique (%)	100.0%

Sample

1st row	LP001002
2nd row	LP001003
3rd row	LP001005
4th row	LP001006

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
5th row	LP001008											

Common Values

Value	Count	Frequency (%)
LP001002	1	0.2%
LP002328	1	0.2%
LP002305	1	0.2%
LP002308	1	0.2%
LP002314	1	0.2%
LP002315	1	0.2%
LP002317	1	

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
						0.2%						
					LP002318	1	0.2%					
					LP002319	1	0.2%					
					LP002332	1	0.2%					
					Other values (604)	604	98.4%					

Length

Histogram of lengths of the category

Value	Count	Frequency (%)
lp001002	1	0.2%
lp001014	1	0.2%
lp001038	1	

Loan_ID	Gender	Married	Dependents	Education	Self_Employed			ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
								0.2%						
						lp001036	1	0.2%						
						lp001005	1	0.2%						
						lp001006	1	0.2%						
						lp001008	1	0.2%						
						lp001011	1	0.2%						
						lp001013	1	0.2%						
						lp001018	1	0.2%						
						Other values (604)	604	98.4%						

- Characters

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

- Categories
- Scripts
- Blocks

Most occurring characters

Value	Count	Frequency (%)
0	1403	28.6%
L	614	12.5%
P	614	12.5%
1	491	10.0%
2	478	9.7%
4	203	4.1%
3	198	4.0%
8	189	

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

		3.8%
7	183	3.7%
9	182	3.7%
Other values (2)	357	7.3%

Most occurring categories

Value	Count	Frequency (%)
Decimal Number	3684	75.0%
Uppercase Letter	1228	25.0%

Most frequent character per category

Decimal Number

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Value	Count	Frequency (%)
0	1403	38.1%
1	491	13.3%
2	478	13.0%
4	203	5.5%
3	198	5.4%
8	189	5.1%
7	183	5.0%
9	182	4.9%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
6	181	4.9%										
5	176	4.8%										

Uppercase Letter

Value	Count	Frequency (%)
L	614	50.0%
P	614	50.0%

Most occurring scripts

Value	Count	Frequency (%)
Common	3684	75.0%
Latin	1228	25.0%

Most frequent character per script

Common

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Value	Count	Frequency (%)
0	1403	38.1%
1	491	13.3%
2	478	13.0%
4	203	5.5%
3	198	5.4%
8	189	5.1%
7	183	5.0%
9	182	4.9%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

6	181	4.9%
5	176	4.8%

Latin

Value	Count	Frequency (%)
L	614	50.0%
P	614	50.0%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	4912	100.0%

Most frequent character per block

ASCII

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Value	Count	Frequency (%)
0	1403	28.6%
L	614	12.5%
P	614	12.5%
1	491	10.0%
2	478	9.7%
4	203	4.1%
3	198	4.0%
8	189	3.8%
7	183	3.7%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

9	182	3.7%
Other values (2)	357	7.3%

Gender
Categorical

HIGH CORRELATION
MISSING

Distinct	2
Distinct (%)	0.3%
Missing	13
Missing (%)	2.1%
Memory size	4.9 KiB

Male	489
Female	112

- Overview
- Categories
- Words

Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

- Characters

Length

Max length	6
Median length	4
Mean length	4.372712146
Min length	4

Characters and Unicode

Total characters	2628
Distinct characters	6
Distinct categories	2 ?
Distinct scripts	1 ?
Distinct blocks	1 ?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0 ?

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Unique (%)	0.0%

Sample

1st row	Male
2nd row	Male
3rd row	Male
4th row	Male
5th row	Male

Common Values

Value	Count	Frequency (%)
Male	489	79.6%
Female	112	18.2%
(Missing)	13	2.1%

Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

Length

Histogram of lengths of the category

Category Frequency Plot

Value	Count	Frequency (%)
male	489	81.4%
female	112	18.6%

- Characters
- Categories
- Scripts
- Blocks

Most occurring characters

Value	Count	Frequency (%)
e	713	27.1%
a	601	22.9%
l	601	22.9%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
M	489	18.6%										
F	112	4.3%										
m	112	4.3%										

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	2027	77.1%
Uppercase Letter	601	22.9%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	713	35.2%
a	601	29.6%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed			ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

l	601	29.6%
m	112	5.5%

Uppercase Letter

Value	Count	Frequency (%)
M	489	81.4%
F	112	18.6%

Most occurring scripts

Value	Count	Frequency (%)
Latin	2628	100.0%

Most frequent character per script

Latin

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Value	Count	Frequency (%)
e	713	27.1%
a	601	22.9%
l	601	22.9%
M	489	18.6%
F	112	4.3%
m	112	4.3%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	2628	100.0%

Most frequent character per block

ASCII

Loan_ID | Gender | Married | Dependents | Education | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount | Loan_Amount_Term | Credit_History | Property_Area | Loan_Status

Value	Count	Frequency (%)
e	713	27.1%
a	601	22.9%
l	601	22.9%
M	489	18.6%
F	112	4.3%
m	112	4.3%

Married
Boolean

HIGH CORRELATION

Distinct	2
Distinct (%)	0.3%
Missing	3
Missing (%)	0.5%

Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

Memory size	1.3 KiB
True	398
False	213
(Missing)	3

- Common Values
- Category Frequency Plot

Value	Count	Frequency (%)
True	398	64.8%
False	213	34.7%
(Missing)	3	0.5%

Dependents
Categorical

HIGH CORRELATION
MISSING

Distinct	4

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Distinct (%)	0.7%
Missing	15
Missing (%)	2.4%
Memory size	4.9 KiB

0	345
1	102
2	101
3+	51

Length

Max length	2
Median length	1
Mean length	1.085141903

Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

Min length	1

Characters and Unicode

Total characters	650
Distinct characters	5
Distinct categories	2 ?
Distinct scripts	1 ?
Distinct blocks	1 ?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0 ?
Unique (%)	0.0%

Sample

1st row	0
2nd row	1

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
3rd row	0											
4th row	0											
5th row	0											

Common Values

Value	Count	Frequency (%)
0	345	56.2%
1	102	16.6%
2	101	16.4%
3+	51	8.3%
(Missing)	15	2.4%

Length

Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

Histogram of lengths of the category

Category Frequency Plot

Value	Count	Frequency (%)
0	345	57.6%
1	102	17.0%
2	101	16.9%
3	51	8.5%

- Characters
- Categories
- Scripts
- Blocks

Most occurring characters

Value	Count	Frequency (%)
0	345	53.1%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
1	102	15.7%										
2	101	15.5%										
3	51	7.8%										
+	51	7.8%										

Most occurring categories

Value	Count	Frequency (%)
Decimal Number	599	92.2%
Math Symbol	51	7.8%

Most frequent character per category

Decimal Number

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Value	Count	Frequency (%)
0	345	57.6%
1	102	17.0%
2	101	16.9%
3	51	8.5%

Math Symbol

Value	Count	Frequency (%)
+	51	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Common	650	100.0%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Most frequent character per script

Common

Value	Count	Frequency (%)
0	345	53.1%
1	102	15.7%
2	101	15.5%
3	51	7.8%
+	51	7.8%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	650	100.0%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Most frequent character per block

ASCII

Value	Count	Frequency (%)
0	345	53.1%
1	102	15.7%
2	101	15.5%
3	51	7.8%
+	51	7.8%

Education
Categorical

Distinct	2
Distinct (%)	0.3%

Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

Missing	0
Missing (%)	0.0%
Memory size	4.9 KiB

Graduate	480
Not Graduate	134

Length

Max length	12
Median length	8
Mean length	8.872964169
Min length	8

Characters and Unicode

Total characters	5448

Loan_ID | Gender | Married | Dependents | Education | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount | Loan_Amount_Term | Credit_History | Property_Area | Loan_Status

Distinct characters	10
Distinct categories	3 ?
Distinct scripts	2 ?
Distinct blocks	1 ?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0 ?
Unique (%)	0.0%

Sample

1st row	Graduate
2nd row	Graduate
3rd row	Graduate
4th row	Not Graduate
5th row	Graduate

Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

Common Values

Value	Count	Frequency (%)
Graduate	480	78.2%
Not Graduate	134	21.8%

Length

Histogram of lengths of the category

Category Frequency Plot

Value	Count	Frequency (%)
graduate	614	82.1%
not	134	17.9%

- Characters
- Categories
- Scripts
- Blocks

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Most occurring characters

Value	Count	Frequency (%)
a	1228	22.5%
t	748	13.7%
G	614	11.3%
r	614	11.3%
d	614	11.3%
u	614	11.3%
e	614	11.3%
N	134	2.5%
o	134	2.5%
	134	2.5%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	4566	83.8%
Uppercase Letter	748	13.7%
Space Separator	134	2.5%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
a	1228	26.9%
t	748	16.4%
r	614	13.4%
d	614	13.4%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

u	614	13.4%
e	614	13.4%
o	134	2.9%

Uppercase Letter

Value	Count	Frequency (%)
G	614	82.1%
N	134	17.9%

Space Separator

Value	Count	Frequency (%)
	134	100.0%

Most occurring scripts

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Value	Count	Frequency (%)
Latin	5314	97.5%
Common	134	2.5%

Most frequent character per script

Latin

Value	Count	Frequency (%)
a	1228	23.1%
t	748	14.1%
G	614	11.6%
r	614	11.6%
d	614	11.6%
u	614	11.6%
e	614	11.6%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed		ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
						N	134						2.5%
						o	134						2.5%

Common

Value	Count	Frequency (%)
	134	100.0%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	5448	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
a	1228	22.5%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed			ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
						t	748	13.7%						
						G	614	11.3%						
						r	614	11.3%						
						d	614	11.3%						
						u	614	11.3%						
						e	614	11.3%						
						N	134	2.5%						
						o	134	2.5%						
							134	2.5%						

Self Employed

Boolean

MISSING

Distinct	2

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Distinct (%)	0.3%
Missing	32
Missing (%)	5.2%
Memory size	1.3 KiB
False	500
True	82
(Missing)	32

- Common Values
- Category Frequency Plot

Value	Count	Frequency (%)
False	500	81.4%
True	82	13.4%
(Missing)	32	5.2%

Loan_ID | Gender | Married | Dependents | Education | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount | Loan_Amount_Term | Credit_History | Property_Area | Loan_Status

ApplicantIncome

Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	505
Distinct (%)	82.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	5403.459283
Minimum	150
Maximum	81000
Zeros	0
Zeros (%)	0.0%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Negative	0
Negative (%)	0.0%
Memory size	4.9 KiB

- Statistics
- Histogram
- Common values
- Extreme values

Quantile statistics

Minimum	150
5-th percentile	1897.55
Q1	2877.5
median	3812.5
Q3	5795
95-th percentile	14583
Maximum	81000
Range	80850

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Interquartile range (IQR)	2917.5

Descriptive statistics

Standard deviation	6109.041673
Coefficient of variation (CV)	1.130579755
Kurtosis	60.54067593
Mean	5403.459283
Median Absolute Deviation (MAD)	1229.5
Skewness	6.539513114
Sum	3317724
Variance	37320390.17
Monotonicity	Not monotonic

Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
2500	9	1.5%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
						4583	6	1.0%				
						6000	6	1.0%				
						2600	6	1.0%				
						3333	5	0.8%				
						4166	5	0.8%				
						3750	5	0.8%				
						5000	5	0.8%				
						8333	4	0.7%				
						6250	4					

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
						0.7%						
						Other values (495)	559	91.0%				

- Minimum 10 values
- Maximum 10 values

Value	Count	Frequency (%)
150	1	0.2%
210	1	0.2%
416	1	0.2%
645	1	0.2%
674	1	0.2%
1000	1	0.2%
1025	2	0.3%
1299	1	0.2%
1378	1	0.2%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

1442	1	0.2%

Value	Count	Frequency (%)
81000	1	0.2%
63337	1	0.2%
51763	1	0.2%
39999	1	0.2%
39147	1	0.2%
37719	1	0.2%
33846	1	0.2%
23803	1	0.2%
20833	1	0.2%
20667	1	0.2%

CoapplicantIncome
Real number (ℝ≥0)

ZEROS

Loan_ID | Gender | Married | Dependents | Education | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount | Loan_Amount_Term | Credit_History | Property_Area | Loan_Status

Distinct	287
Distinct (%)	46.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	1621.245798
Minimum	0
Maximum	41667
Zeros	273
Zeros (%)	44.5%
Negative	0
Negative (%)	0.0%
Memory size	4.9 KiB

- Statistics
- Histogram

Loan_ID | Gender | Married | Dependents | Education | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount | Loan_Amount_Term | Credit_History | Property_Area | Loan_Status

- Common values
- Extreme values

Quantile statistics

Minimum	0
5-th percentile	0
Q1	0
median	1188.5
Q3	2297.25
95-th percentile	4997.4
Maximum	41667
Range	41667
Interquartile range (IQR)	2297.25

Descriptive statistics

Standard deviation	2926.248369
Coefficient of variation (CV)	1.804938136

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Kurtosis	84.95638421
Mean	1621.245798
Median Absolute Deviation (MAD)	1188.5
Skewness	7.491531217
Sum	995444.92
Variance	8562929.518
Monotonicity	Not monotonic

Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
0	273	44.5%
2500	5	0.8%
2083	5	0.8%
1666	5	0.8%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed			ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
						2250	3	0.5%						
						1750	3	0.5%						
						1800	3	0.5%						
						1625	3	0.5%						
						2333	3	0.5%						
						1459	3	0.5%						
						Other values (277)	308	50.2%						

- Minimum 10 values
- Maximum 10 values

Loan_ID | Gender | Married | Dependents | Education | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount | Loan_Amount_Term | Credit_History | Property_Area | Loan_Status

Value	Count	Frequency (%)
0	273	44.5%
16.12000084	1	0.2%
189	1	0.2%
240	1	0.2%
242	1	0.2%
461	1	0.2%
484	1	0.2%
505	1	0.2%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Value	Count	Frequency (%)
536	1	0.2%
663	1	0.2%
41667	1	0.2%
33837	1	0.2%
20000	2	0.3%
11300	1	0.2%
10968	1	0.2%
8980	1	0.2%
8333	1	0.2%
8106	1	0.2%
7873	1	0.2%
7750	1	0.2%

Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

LoanAmount

Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
MISSING

Distinct	203
Distinct (%)	34.3%
Missing	22
Missing (%)	3.6%
Infinite	0
Infinite (%)	0.0%
Mean	146.4121622
Minimum	9
Maximum	700
Zeros	0
Zeros (%)	0.0%
Negative	0

Loan_ID | Gender | Married | Dependents | Education | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount | Loan_Amount_Term | Credit_History | Property_Area | Loan_Status

Negative (%)	0.0%
Memory size	4.9 KiB

- Statistics
- Histogram
- Common values
- Extreme values

Quantile statistics

Minimum	9
5-th percentile	56
Q1	100
median	128
Q3	168
95-th percentile	297.8
Maximum	700
Range	691
Interquartile range (IQR)	68

Loan_ID | Gender | Married | Dependents | Education | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount | Loan_Amount_Term | Credit_History | Property_Area | Loan_Status

Descriptive statistics

Standard deviation	85.58732524
Coefficient of variation (CV)	0.5845643147
Kurtosis	10.40153349
Mean	146.4121622
Median Absolute Deviation (MAD)	32
Skewness	2.677551679
Sum	86676
Variance	7325.190241
Monotonicity	Not monotonic

Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
120	20	3.3%
110	17	

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

		2.8%
100	15	2.4%
160	12	2.0%
187	12	2.0%
113	11	1.8%
128	11	1.8%
130	10	1.6%
95	9	1.5%
96	9	1.5%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
Other values (193)						466	75.9%					
(Missing)						22	3.6%					

- Minimum 10 values
- Maximum 10 values

Value	Count	Frequency (%)
9	1	0.2%
17	1	0.2%
25	2	0.3%
26	1	0.2%
30	2	0.3%
35	1	0.2%
36	1	0.2%
40	2	0.3%
42	1	0.2%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

44	2	0.3%

Value	Count	Frequency (%)
700	1	0.2%
650	1	0.2%
600	2	0.3%
570	1	0.2%
500	1	0.2%
496	1	0.2%
495	1	0.2%
490	1	

Loan_ID	Gender	Married	Dependents	Education	Self_Employed			ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

		0.2%
480	3	0.5%
436	1	0.2%

Loan_Amount_Term

Real number (ℝ≥0)

MISSING

Distinct	10
Distinct (%)	1.7%
Missing	14
Missing (%)	2.3%
Infinite	0
Infinite (%)	0.0%
Mean	342
Minimum	12
Maximum	480

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	4.9 KiB

- Statistics
- Histogram
- Common values
- Extreme values

Quantile statistics

Minimum	12
5-th percentile	180
Q1	360
median	360
Q3	360
95-th percentile	360

Loan_ID | Gender | Married | Dependents | Education | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount | Loan_Amount_Term | Credit_History | Property_Area | Loan_Status

Maximum	480
Range	468
Interquartile range (IQR)	0

Descriptive statistics

Standard deviation	65.12040985
Coefficient of variation (CV)	0.1904105551
Kurtosis	6.673473693
Mean	342
Median Absolute Deviation (MAD)	0
Skewness	-2.362414124
Sum	205200
Variance	4240.66778
Monotonicity	Not monotonic

Histogram with fixed size bins (bins=10)

Loan_ID | Gender | Married | Dependents | Education | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount | Loan_Amount_Term | Credit_History | Property_Area | Loan_Status

Value	Count	Frequency (%)
360	512	83.4%
180	44	7.2%
480	15	2.4%
300	13	2.1%
240	4	0.7%
84	4	0.7%
120	3	0.5%
60	2	0.3%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

36	2	0.3%
12	1	0.2%
(Missing)	14	2.3%

- Minimum 10 values
- Maximum 10 values

Value	Count	Frequency (%)
12	1	0.2%
36	2	0.3%
60	2	0.3%
84	4	0.7%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Value	Count	Frequency (%)
120	3	0.5%
180	44	7.2%
240	4	0.7%
300	13	2.1%
360	512	83.4%
480	15	2.4%

Value	Count	Frequency (%)
480	15	2.4%
360	512	83.4%
300	13	

Loan_ID Gender Married Dependents Education Self_Employed **ApplicantIncome** CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

		2.1%
240	4	0.7%
180	44	7.2%
120	3	0.5%
84	4	0.7%
60	2	0.3%
36	2	0.3%
12	1	0.2%

Credit_History
Categorical

HIGH CORRELATION
HIGH CORRELATION

Loan_ID
Gender
Married
Dependents
Education
Self_Employed
ApplicantIncome
CoapplicantIncome
LoanAmount
Loan_Amount_Term
Credit_History
Property_Area
Loan_Status

MISSING

Distinct	2
Distinct (%)	0.4%
Missing	50
Missing (%)	8.1%
Memory size	4.9 KiB

1.0	475
0.0	89

- Overview
- Categories
- Words
- Characters

Length

Max length	3
Median length	3
Mean length	3

Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

Min length 3

Characters and Unicode

Total characters 1692

Distinct characters 3

Distinct categories 2 ?

Distinct scripts 1 ?

Distinct blocks 1 ?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique 0 ?

Unique (%) 0.0%

Sample

1st row 1.0

2nd row 1.0

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
3rd row						1.0						
4th row						1.0						
5th row						1.0						

Common Values

Value	Count	Frequency (%)
1.0	475	77.4%
0.0	89	14.5%
(Missing)	50	8.1%

Length

Histogram of lengths of the category

Category Frequency Plot

Value	Count	Frequency (%)
1.0	475	84.2%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
						0.0	89					
							15.8%					

- Characters
- Categories
- Scripts
- Blocks

Most occurring characters

Value	Count	Frequency (%)
0	653	38.6%
.	564	33.3%
1	475	28.1%

Most occurring categories

Value	Count	Frequency (%)
Decimal Number	1128	66.7%
Other Punctuation	564	33.3%

Most frequent character per category

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Decimal Number

Value	Count	Frequency (%)
0	653	57.9%
1	475	42.1%

Other Punctuation

Value	Count	Frequency (%)
.	564	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Common	1692	100.0%

Most frequent character per script

Common

Value	Count	Frequency (%)
0	653	38.6%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed		ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

.	564	33.3%
1	475	28.1%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	1692	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
0	653	38.6%
.	564	33.3%
1	475	28.1%

Property_Area
Categorical

Distinct	3
Distinct (%)	0.5%

Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

Missing	0
Missing (%)	0.0%
Memory size	4.9 KiB
Semiurban	233
Urban	202
Rural	179

- Overview
- Categories
- Words
- Characters

Length

Max length	9
Median length	5
Mean length	6.517915309
Min length	5

Characters and Unicode

Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status

Total characters	4002
Distinct characters	12
Distinct categories	2 ?
Distinct scripts	1 ?
Distinct blocks	1 ?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0 ?
Unique (%)	0.0%

Sample

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

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
5th row											Urban	

Common Values

Value	Count	Frequency (%)
Semiurban	233	37.9%
Urban	202	32.9%
Rural	179	29.2%

Length

Histogram of lengths of the category

Category Frequency Plot

Value	Count	Frequency (%)
semiurban	233	37.9%
urban	202	32.9%
rural	179	29.2%

- Characters
- Categories

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

- Scripts
- Blocks

Most occurring characters

Value	Count	Frequency (%)
r	614	15.3%
a	614	15.3%
b	435	10.9%
n	435	10.9%
u	412	10.3%
S	233	5.8%
e	233	5.8%
m	233	5.8%
i	233	

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

		5.8%
U	202	5.0%
Other values (2)	358	8.9%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	3388	84.7%
Uppercase Letter	614	15.3%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
r	614	18.1%
a	614	18.1%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed			ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

b	435	12.8%
n	435	12.8%
u	412	12.2%
e	233	6.9%
m	233	6.9%
i	233	6.9%
l	179	5.3%

Uppercase Letter

Value	Count	Frequency (%)
S	233	37.9%
U	202	32.9%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
R	179	29.2%										

Most occurring scripts

Value	Count	Frequency (%)
Latin	4002	100.0%

Most frequent character per script

Latin

Value	Count	Frequency (%)
r	614	15.3%
a	614	15.3%
b	435	10.9%
n	435	10.9%
u	412	10.3%
S	233	5.8%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
						e 233 5.8%						
						m 233 5.8%						
						i 233 5.8%						
						U 202 5.0%						
						Other values (2) 358 8.9%						

Most occurring blocks

Value	Count	Frequency (%)
ASCII	4002	100.0%

Most frequent character per block

ASCII

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Value	Count	Frequency (%)
r	614	15.3%
a	614	15.3%
b	435	10.9%
n	435	10.9%
u	412	10.3%
S	233	5.8%
e	233	5.8%
m	233	5.8%
i	233	5.8%
U	202	5.0%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

Other values (2)	358	8.9%

Loan_Status

Boolean

HIGH CORRELATION
HIGH CORRELATION

Distinct	2
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	742.0 B

True	422
False	192

- Common Values
- Category Frequency Plot

Value	Count	Frequency (%)
True	422	68.7%

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
False	192	31.3%										

Interactions

- ApplicantIncome
- CoapplicantIncome
- LoanAmount
- Loan_Amount_Term

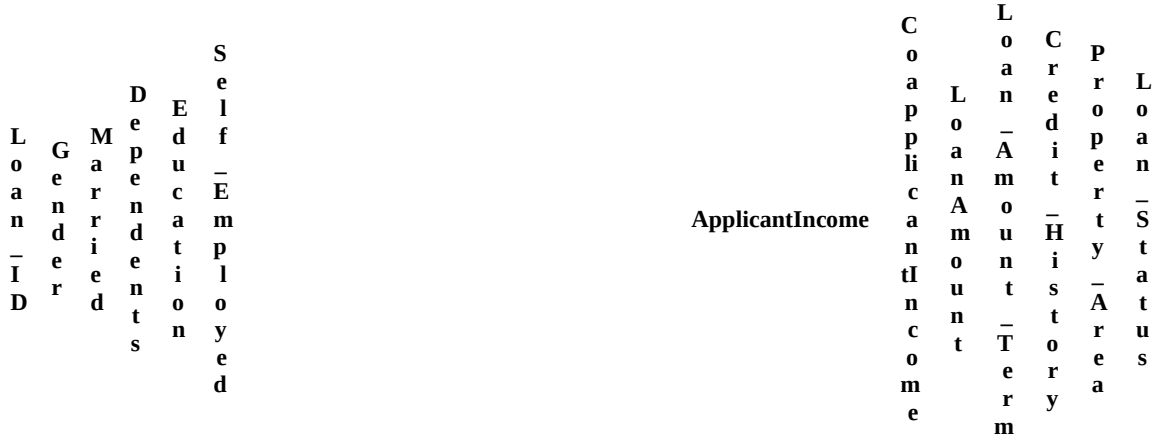
- Loan_Amount_Term
- ApplicantIncome
- CoapplicantIncome
- LoanAmount

- Loan_Amount_Term
- ApplicantIncome
- CoapplicantIncome
- LoanAmount

- Loan_Amount_Term
- ApplicantIncome
- CoapplicantIncome
- LoanAmount

- Loan_Amount_Term
- ApplicantIncome
- CoapplicantIncome
- LoanAmount

Correlations



- Spearman's ρ
- Pearson's r
- Kendall's τ
- Cramér's V (φc)
- Phik (φk)

Spearman's ρ

The Spearman's rank correlation coefficient (ρ) is a measure of monotonic correlation between two variables, and is therefore better in catching nonlinear monotonic correlations than Pearson's *r*. It's value lies between -1 and +1, -1 indicating total negative monotonic correlation, 0 indicating no monotonic correlation and 1 indicating total positive monotonic correlation.

To calculate ρ for two variables X and Y, one

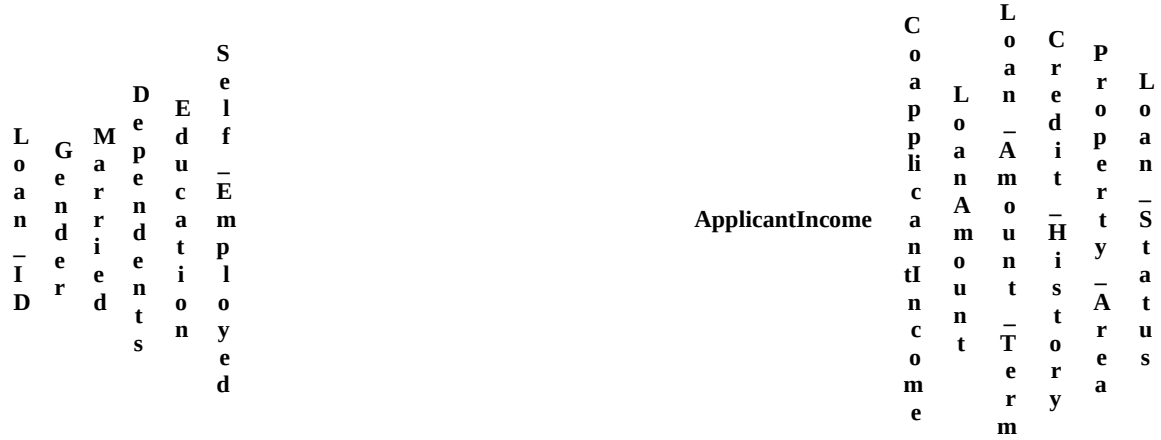
Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status

divides the covariance of the rank variables of *X* and *Y* by the product of their standard deviations.

Pearson's r

The Pearson's correlation coefficient (*r*) is a measure of linear correlation between two variables. It's value lies between -1 and +1, -1 indicating total negative linear correlation, 0 indicating no linear correlation and 1 indicating total positive linear correlation. Furthermore, *r* is invariant under separate changes in location and scale of the two variables, implying that for a linear function the angle to the x-axis does not affect *r*.

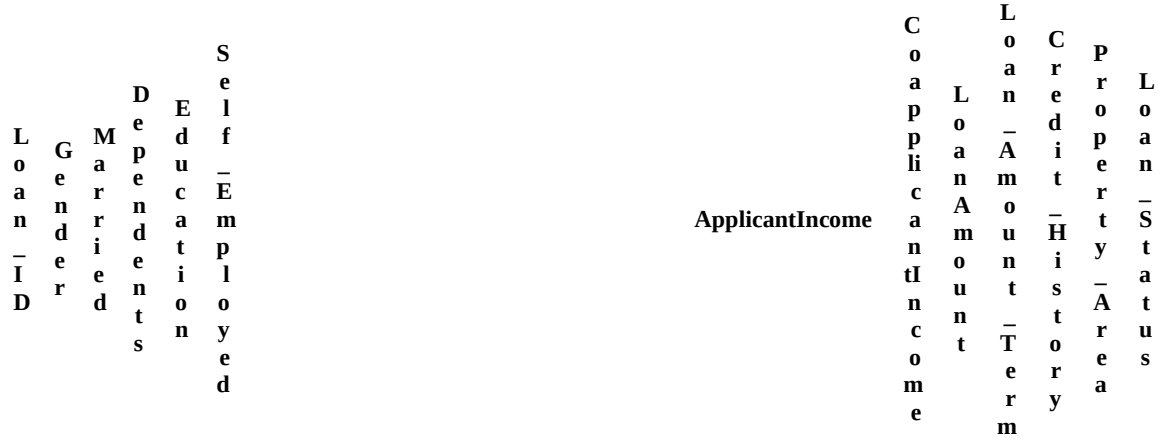
To calculate *r* for two variables *X* and *Y*, one divides the covariance of *X* and *Y* by the product of their standard deviations.



Kendall's τ

Similarly to Spearman's rank correlation coefficient, the Kendall rank correlation coefficient (τ) measures ordinal association between two variables. It's value lies between -1 and +1, -1 indicating total negative correlation, 0 indicating no correlation and 1 indicating total positive correlation.

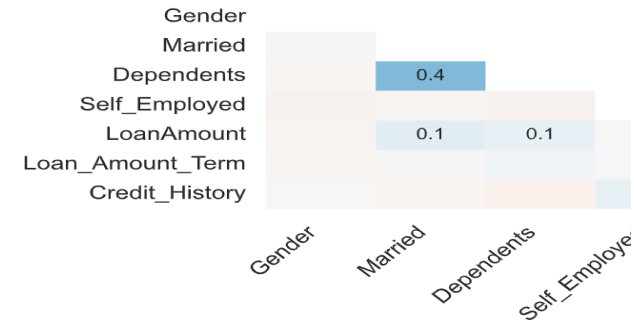
To calculate τ for two variables X and Y, one determines the number of concordant and discordant pairs of observations. τ is given by the number of concordant pairs minus the discordant pairs divided by the total number of pairs.



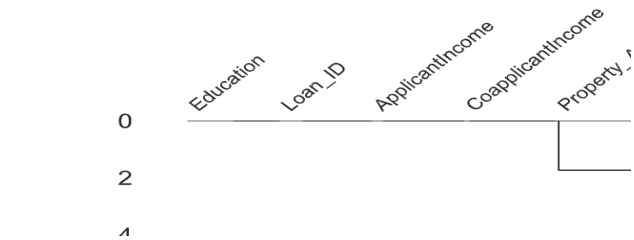
Cramér's V (φc)

Cramér's V is an association measure for nominal random variables. The coefficient ranges from 0 to 1, with 0 indicating independence and 1 indicating perfect association. The empirical estimators used for Cramér's V have been proved to be biased, even for large samples. We use a bias-corrected measure that has been proposed by Bergsma in 2013 that can be found here.

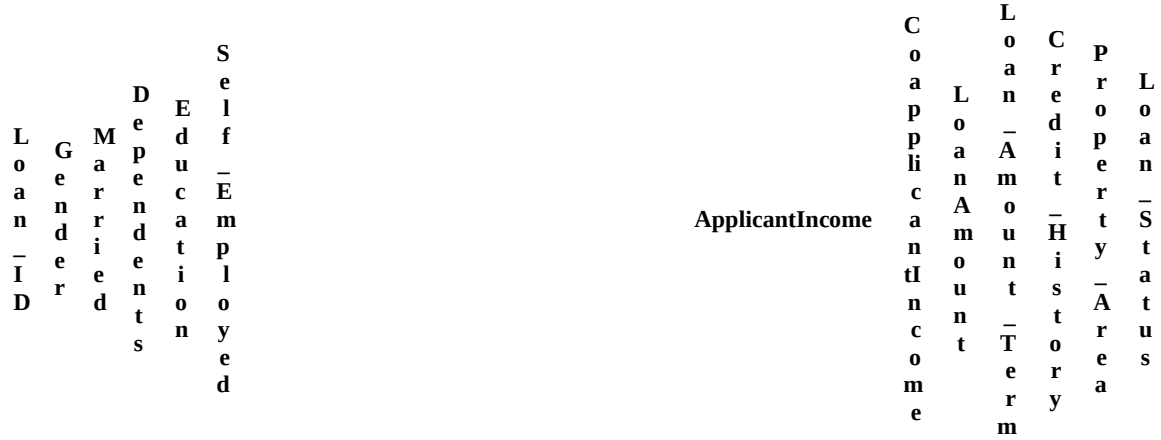
Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area Loan_Status



he correlation heatmap measures nullity correlation: how strongly the p



Phik (φk)



Phik (φk) is a new and practical correlation coefficient that works consistently between categorical, ordinal and interval variables, captures non-linear dependency and reverts to the Pearson correlation coefficient in case of a bivariate normal input distribution. There is extensive documentation available here.

Missing values

- Count
- Matrix
- Heatmap
- Dendrogram

A simple visualization of nullity by column.

Loan_ID

Gender

Married

Dependents

Education

Self_Employed

ApplicantIncome

CoapplicantIncome

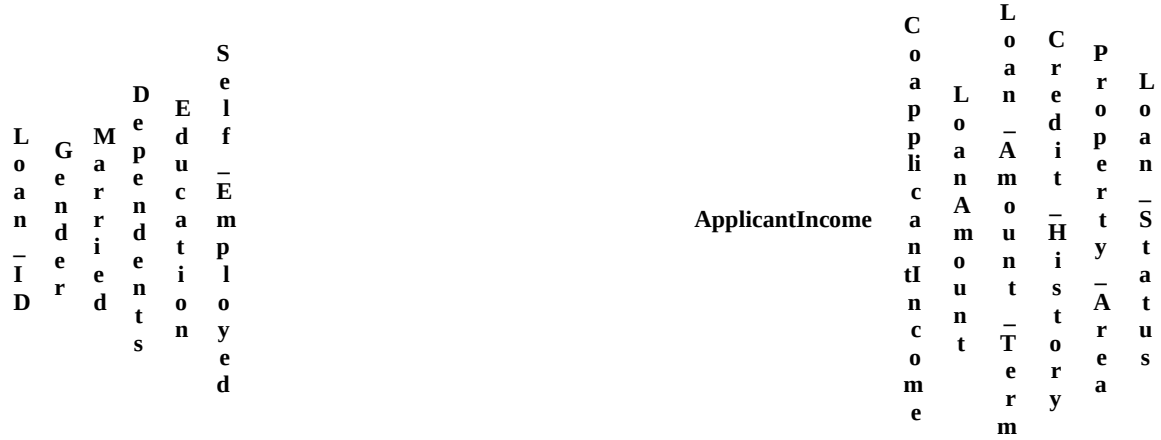
LoanAmount

Loan_Amount_Term

Credit_History

Property_Area

Loan_Status



Nullity matrix is a data-dense display which lets you quickly visually pick out patterns in data completion.

The correlation heatmap measures nullity correlation: how strongly the presence or absence of one variable affects the presence of another.
The dendrogram allows you to more fully correlate variable completion, revealing trends deeper than the pairwise ones visible in the correlation heatmap.

Sample

First rows

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
2	LP001010	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed			ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
	05				te										
3	LP001006	Male	Yes	0	NotGraduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y		
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y		
5	LP001011	Male	Yes	2	Graduate	Yes	5417	4196.0	267.0	360.0	1.0	Urban	Y		
6	LP0010	Male	Yes	0	NotGradu	No	2333	15160.0	95.0	360.0	1.0	Urban	Y		

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed			ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
	13					ate									
7	LP001014	Male	Yes	3+	Graduate	No	3036		2504.0	158.0	360.0	0.0	Semiurban		N
8	LP001018	Male	Yes	2	Graduate	No	4006		1526.0	168.0	360.0	1.0	Urban		Y
9	LP001020	Male	Yes	1	Graduate	No	12841		10968.0	349.0	360.0	1.0	Semiurban		N

Last rows

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
604	LP002959	Female	Yes	1	Graduate	No	12000	0.0	496.0	360.0	1.0	Semiurban	Y
605	LP002960	Male	Yes	0	Not Graduate	No	2400	3800.0	NaN	180.0	1.0	Urban	N

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
606	LP002961	Male	Yes	1	Graduate	No	3400	2500.0	173.0	360.0	1.0	Semiurban	Y
607	LP002964	Male	Yes	2	NotGraduate	No	3987	1411.0	157.0	360.0	1.0	Rural	Y
608	LP002974	Male	Yes	0	Graduate	No	3232	1950.0	108.0	360.0	1.0	Rural	Y
609	LP002978	Female	No	0	Graduate	No	2900	0.0	71.0	360.0	1.0	Rural	Y

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
610	LP002979	Male	Yes	3+	Graduate	No	4106	0.0	40.0	180.0	1.0	Rural	Y
611	LP002983	Male	Yes	1	Graduate	No	8072	240.0	253.0	360.0	1.0	Urban	Y
612	LP002984	Male	Yes	2	Graduate	No	7583	0.0	187.0	360.0	1.0	Urban	Y
613	LP002990	Female	No	0	Graduate	Yes	4583	0.0	133.0	360.0	0.0	Semiurban	