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]:

| | Loa n_I D | Ge nd er | Ma rri ed | Dep ende nts | Edu cati on | Self_ Empl oyed | Applic antInc ome | Coappli cantInc ome | Loan Amo unt | Loan_A mount_ Term | Credi t_Hist ory | Prope rty_A rea | Loa n_St atus |
|---|------------------|----------------|-----------------|--------------------|-------------------------|-----------------------|-------------------------|---------------------------|--------------------|--------------------------|------------------------|-----------------------|---------------------|
| 0 | LP0 010 02 | M ale | No | 0 | Gra duat e | No | 5849 | 0.0 | NaN | 360.0 | 1.0 | Urban | Y |
| 1 | LP0 010 03 | M ale | Ye s | 1 | Gra duat e | No | 4583 | 1508.0 | 128.0 | 360.0 | 1.0 | Rural | N |
| 2 | LP0 010 05 | M ale | Ye s | 0 | Gra duat e | Yes | 3000 | 0.0 | 66.0 | 360.0 | 1.0 | Urban | Y |
| 3 | LP0 010 06 | M ale | Ye s | 0 | Not Gra duat e | No | 2583 | 2358.0 | 120.0 | 360.0 | 1.0 | Urban | Y |
| 4 | LP0 010 08 | M ale | No | 0 | Gra duat e | 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 Analysis report for data Analysis

- <u>Overview</u>
- <u>Variables</u>
- Interactions
- Correlations
- Missing values
- <u>Sample</u>

Overview

- Overview
- Alerts 23
- Reproduction

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
        LP002308
                                1
        LP002314
                                1
 Other values (609)
                              609
     <u>Overview</u>
     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 2

Distinct scripts 2 2

Distinct blocks 1
```

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

| Uniqu | e 614 ? |
|-----------|-----------|
| Unique (% | o) 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 (%) |
|----------|-------|---------------|
| 1p001002 | 1 | 0.2% |
| lp001014 | 1 | |

| Value | Count | Frequency (%) |
|--------------------|-------|---------------|
| | | 0.2% |
| lp001038 | 1 | 0.2% |
| 1p001036 | 1 | 0.2% |
| 1p001005 | 1 | 0.2% |
| lp001006 | 1 | 0.2% |
| 1p001008 | 1 | 0.2% |
| lp001011 | 1 | 0.2% |
| lp001013 | 1 | 0.2% |
| lp001018 | 1 | 0.2% |
| Other values (604) | 604 | 98.4% |

- <u>Characters</u>
- CategoriesScriptsBlocks

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 | |

4.8%

Uppercase Letter

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

Most occurring scripts

Most frequent character per script

Common

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

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

Latin

5

176

4.8%

| 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% |

<u>Gender</u> Categorical

HIGH CORRELATION MISSING

| | | | Distinct | 2 |
|---|----------|--------|--------------|---------|
| | | | Distinct (%) | 0.3% |
| | | | Missing | 13 |
| | | | Missing (%) | 2.1% |
| | | | Memory size | 4.9 KiB |
| | | Male | | 489 |
| | | Female | | 112 |
| OverviewCategoriesWordsCharacter | <u>S</u> | | | |
| Length | | | | |
| Max length | 6 | | | |
| Median length | 4 | | | |
| Mean length 4.3 | 72712146 | | | |
| Min length | 4 | | | |
| Characters and U | Inicode | | | |
| Total characters | 2628 | | | |
| Distinct characters | 6 | | | |
| Distinct categories | 2 ? | | | |
| Distinct scripts | 1 2 | | | |

Distinct blocks 1 ?

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

| Uniqu | ie 0 ? |
|-----------|-----------------|
| Unique (% | 6) 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% |

- CharactersCategoriesScriptsBlocks

Most occurring characters

| Value | Count | Frequency (%) |
|-------|-------|---------------|
| e | 713 | 27.1% |
| a | 601 | 22.9% |
| 1 | 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% |
| 1 | 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 (%) |
|-------|-------|---------------|
| 1 | 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% |
| 1 | 601 | 22.9% |
| M | 489 | 18.6% |
| F | 112 | 4.3% |
| m | 112 | 4.3% |

<u>Married</u>

Boolean

HIGH CORRELATION

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

- Common ValuesCategory Frequency Plot

| Value | Count | Frequency (%) |
|-----------|-------|---------------|
| , arac | Count | requency (%) |
| True | 398 | 64.8% |
| | | |
| False | 213 | 34.7% |
| | | |
| (Missing) | 3 | 0.50/ |
| | | 0.5% |

<u>Dependents</u> Categorical

HIGH CORRELATION MISSING

Missing (%) 2.4%

Missing (%) 2.4%

Memory size 4.9

KiB

1 102

2 101

3+ 51

Distinct (%) 0.7%

- <u>Overview</u>
- <u>Categories</u>
- Words
- <u>Characters</u>

Length

Max length 2

Median length 1

Mean length 1.085141903

Min length

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% |

- CharactersCategories
- ScriptsBlocks

Most occurring characters

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

7.8%

Most occurring categories

| Value 0 | 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 | |

Math Symbol

Value Count Frequency (%)

+ 51 100.0%

8.5%

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% | |
| Educat Catego | | | |
| Di | stinct | 2 | |
| Distinc | t (%) | 0.3% | |
| Mi | issing | 0 | |
| Missin | g (%) | 0.0% | |
| Memor | y size | 4.9 KiB | |
| | Gradua | te | 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 2

Distinct scripts 2 2

Distinct blocks 1 2

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

Unique (%) 0.0%

Sample

1st row Graduate

2nd 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% |

- <u>Characters</u>
- CategoriesScriptsBlocks

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% |
| 0 | 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

| Frequency (%) | Count | Value |
|---------------|-------|-------|
| 23.1% | 1228 | a |
| 14.1% | 748 | t |
| 11.6% | 614 | G |
| 11.6% | 614 | r |
| 11.6% | 614 | d |
| 11.6% | 614 | u |
| 11.6% | 614 | e |
| 2.5% | 134 | N |
| 2.5% | 134 | 0 |

Common

Value Count Frequency (%)

134 100.0%

Most occurring blocks

Value Count Frequency (%)

ASCII 5448 100.0%

Most frequent character per block

ASCII

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

Self Employ Boolean

MISSING

| | <u>oyed</u> |
|---------|--------------|
| 2 | Distinct |
| 0.3% | Distinct (%) |
| 32 | Missing |
| 5.2% | Missing (%) |
| 1.3 KiB | Memory size |
| | |

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% 0 Missing 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 |
|--------------------|-------------|
| Stanuaru utvianum | 0107.041073 |

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%

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

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% |
|--|-------|
| IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII | 0.070 |

Mean 1621.245798

Minimum 0

Maximum 41667

Zeros 273

Zeros (%) 44.5%

Negative 0

Negative (%) 0.0%

Memory size 4.9 KiB

- <u>Statistics</u>
- <u>Histogram</u>
- 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% |
| | | |

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 |

- <u>Statistics</u>
- <u>Histogram</u>
- <u>Common values</u>
- 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

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

| Frequency (%) | Count | Value |
|---------------|-------|-------|
| 0.2% | 1 | 9 |
| 0.2% | 1 | 17 |
| 0.3% | 2 | 25 |
| 0.2% | 1 | 26 |
| 0.3% | 2 | 30 |
| 0.2% | 1 | 35 |
| 0.2% | 1 | 36 |
| 0.3% | 2 | 40 |
| 0.2% | 1 | 42 |
| 0.3% | 2 | 44 |
| Frequency (%) | Count | Value |
| 0.2% | 1 | 700 |
| 0.2% | 1 | 650 |
| 0.3% | 2 | 600 |
| 0.2% | 1 | 570 |
| | 1 | 500 |

| 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 (%) 0.0%

Negative 0

Negative (%) 0.0%
```

Memory size 4.9 KiB

- <u>Statistics</u>
- <u>Histogram</u>
- <u>Common values</u>
- 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)

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% |

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

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

<u>Credit History</u> 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 | |

- <u>Overview</u>
- Categories
- Words
- <u>Characters</u>

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 (%) 0.0%

Sample

1st row 1.0

2nd row 1.0

4th row 1.0

5th row 1.0

Common Values

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

Length

Histogram of lengths of the category

Category Frequency Plot

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

- <u>Characters</u>
- CategoriesScripts
- <u>Blocks</u>

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

Count Frequency (%) Value ASCII 1692 100.0%

Most frequent character per block

ASCII

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

<u>Property Area</u> Categorical

Distinct 3 Distinct (%) 0.5% 0 Missing Missing (%) 0.0% Memory size 4.9 KiB

Semiurban

202 Urban

233

Rural 179

- <u>Overview</u>
- <u>Categories</u>
- Words
- Characters

Length

Max length 9

Median length 5

Mean length 6.517915309

Min length

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% |

- <u>Characters</u>
- CategoriesScripts
- 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% |
| 1 | 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% |

<u>Loan Status</u> Boolean

HIGH CORRELATION HIGH CORRELATION

| 2 | Distinct | |
|---------|--------------|-------|
| 0.3% | Distinct (%) | |
| C | Missing | |
| 0.0% | Missing (%) | |
| 742.0 B | Memory size | |
| 422 | True | True |
| 192 | False | False |

Distinct

- Common Values
- <u>Category Frequency Plot</u>

 Value
 Count
 Frequency (%)

 True
 422
 68.7%

 False
 192
 31.3%

Interactions

- <u>ApplicantIncome</u>
- <u>CoapplicantIncome</u>
- LoanAmount
- Loan Amount Term
- Loan Amount Term
- ApplicantIncome
- <u>CoapplicantIncome</u>
- <u>LoanAmount</u>
- Loan Amount Term
- ApplicantIncome
- <u>CoapplicantIncome</u>
- <u>LoanAmount</u>

- Loan Amount Term
- ApplicantIncome
- <u>CoapplicantIncome</u>
- LoanAmount
- Loan Amount Term
- ApplicantIncome
- <u>CoapplicantIncome</u>
- LoanAmount

Correlations

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

Spearman's p

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. $	au$ 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 (ϕ k) 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

- <u>Count</u>
- <u>Matrix</u>
- <u>Heatmap</u>
- <u>Dendrogram</u>

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

| | Loa n_I D | Ge nd er | Ma rri ed | Dep ende nts | Edu cati on | Self_ Empl oyed | Applic antInc ome | Coappli cantInc ome | Loan Amo unt | Loan_A mount_ Term | Credi t_Hist ory | Prope rty_A rea | Loa n_St atus |
|---|------------------|----------------|-----------------|--------------------|-------------------------|-----------------------|-------------------------|---------------------------|--------------------|--------------------------|------------------------|-----------------------|---------------------|
| 0 | LP0 010 02 | M ale | No | 0 | Gra duat e | No | 5849 | 0.0 | NaN | 360.0 | 1.0 | Urban | Y |
| 1 | LP0 010 03 | M ale | Ye s | 1 | Gra duat e | No | 4583 | 1508.0 | 128.0 | 360.0 | 1.0 | Rural | N |
| 2 | LP0 010 05 | M ale | Ye s | 0 | Gra duat e | Yes | 3000 | 0.0 | 66.0 | 360.0 | 1.0 | Urban | Y |
| 3 | LP0 010 06 | M ale | Ye s | 0 | Not Gra duat e | No | 2583 | 2358.0 | 120.0 | 360.0 | 1.0 | Urban | Y |
| 4 | LP0 010 08 | M ale | No | 0 | Gra duat e | No | 6000 | 0.0 | 141.0 | 360.0 | 1.0 | Urban | Y |
| 5 | LP0 010 11 | M ale | Ye s | 2 | Gra duat e | Yes | 5417 | 4196.0 | 267.0 | 360.0 | 1.0 | Urban | Y |

| | Loa n_I D | Ge nd er | Ma rri ed | Dep ende nts | Edu cati on | Self_ Empl oyed | Applic antInc ome | Coappli cantInc ome | Loan Amo unt | Loan_A mount_ Term | Credi t_Hist ory | Prope rty_A rea | Loa n_St atus |
|---|------------------|----------------|-----------------|--------------------|-------------------------|-----------------------|-------------------------|---------------------------|--------------------|--------------------------|------------------------|-----------------------|---------------------|
| 6 | LP0 010 13 | M ale | Ye s | 0 | Not Gra duat e | No | 2333 | 1516.0 | 95.0 | 360.0 | 1.0 | Urban | Y |
| 7 | LP0 010 14 | M ale | Ye s | 3+ | Gra duat e | No | 3036 | 2504.0 | 158.0 | 360.0 | 0.0 | Semiu rban | N |
| 8 | LP0 010 18 | M ale | Ye s | 2 | Gra duat e | No | 4006 | 1526.0 | 168.0 | 360.0 | 1.0 | Urban | Y |
| 9 | LP0 010 20 | M ale | Ye s | 1 | Gra duat e | No | 12841 | 10968.0 | 349.0 | 360.0 | 1.0 | Semiu rban | N |

Last rows

| | L o a n - I D | G e n d e r | M a r r i e d | D e p e n d e n t s | E d u c a t i o n | S e l f f — E m p l o y e d d | ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | $\begin{array}{c} L \\ o \\ a \\ n \\ \\ \hline M \\ o \\ u \\ n \\ t \\ \\ \hline T \\ e \\ r \\ m \end{array}$ | C r e d i t T H i s t o r y | P r o p e r t y -A r e a | L o a n -S t a t u s |
|-------------|---------------------------------|----------------------------|---------------------------------|---------------------|--------------------------------------|-------------------------------|-----------------|---------------------------------|---------------------|--|-----------------------------|---|---|
| 6 0 4 | L P 0 0 2 9 5 | F e m a l e | Y e s | 1 | G r a d u a t e | N o | 12000 | 0 | 4 9 6 | 3 6 0. 0 | 1 . 0 | S e m i u r b a n | Y |
| 6 0 5 | L P 0 | M a | Y e s | 0 | N o t G | N o | 2400 | 3 8 0 0 | N a N | 1 8 | 1 0 | U r b | N |

| | L o a n Ī D | G e n d e r | M a r r i e d | D e p e n d e n t s | E d u c a t i o n | S e l f E m p l o y e d d | ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — Ā m o u n t — T e r m | C r e d i t - H i s t o r y | Property—Area | L o a n - S t a t u s |
|-------------|--------------------------------------|----------------------------|---------------------------------|---------------------|--------------------------------------|---------------------------|-----------------|---------------------------------|---------------------|---------------------------------|-----------------------------|-----------------------|-----------------------|
| | 2 9 6 0 | l e | | | r a d u a t e | | | 0 | | 0. | | a n | |
| 6 0 6 | L P 0 0 2 9 6 | M a l e | Y e s | 1 | G r a d u a t e | N o | 3400 | 2 5 0 0 | 1 7 3 | 3 6 0. 0 | 1 . 0 | S e m i u r b a n | Y |
| 6 0 7 | L P 0 0 2 9 6 4 | M a l e | Y e s | 2 | N o t G r a d u a t e | N o | 3987 | 1 4 1 1 | 1 5 7 | 3 6 0. 0 | 1 . 0 | R u r a 1 | Y |
| 6 0 8 | L P 0 0 2 9 7 4 | M a l e | Y e s | 0 | G r a d u a t e | N o | 3232 | 1 9 5 0 | 1 0 8 | 3 6 0. 0 | 1 . 0 | R u r a 1 | Y |

| | L o a n - I D | G e n d e r | M a r r i e d | D e p e n d e n t s | E d u c a t i o n | S e l f | Applicar | ntIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — Ā m o u n t — T e r m | C r e d i t t i s t o r y | Property-Areea | L o a n - S t a t u s |
|-------------|--------------------------------------|----------------------------|---------------------------------|---------------------|--------------------------------------|-------------|--|----------|---------------------------------|---------------------|---------------------------------|---------------------------|-----------------------|-----------------------|
| 6 0 9 | L P 0 0 2 9 7 8 | F e m a l e | N o | 0 | G r a d u a t e | N o | | 2900 | 0 . 0 | 7 1 0 | 3 6 0. 0 | 1 . 0 | R u r a l | Y |
| 6 1 0 | L P 0 0 2 9 7 9 | M a l e | Y e s | 3 + | G r a d u a t e | N o | | 4106 | 0 . 0 | 4 0 | 1 8 0. 0 | 1 0 | R u r a l | Y |
| 6 1 1 | L P 0 0 2 9 8 3 | M a l e | Y e s | 1 | G r a d u a t e | N o | | 8072 | 2 4 0 | 2 5 3 | 3 6 0. 0 | 1 . 0 | U r b a n | Y |
| 6 1 2 | L P 0 0 2 9 8 4 | M a l e | Y e s | 2 | G r a d u a t e | N o | | 7583 | 0 . 0 | 1 8 7 | 3 6 0. 0 | 1 . 0 | U r b a n | Y |
| 6 1 3 | L P 0 0 2 | F e m a | N o | 0 | G r a d u | Y e s | 45import numpy as np import seaborn as sb import pandas as pd from pandas_profiling import ProfileReport | | 0 . 0 | 1 3 3 | 3 6 0. 0 | 0 . 0 | S e m i u | |

```
C
                                                                                                                       \mathbf{C}
                       \mathbf{S}
                                                                                                       0
                                                                                                                  a
                                                                                                                       r
                       e
                                                                                                       a
                       l
                                                                                                       p
                                                                                                                       d
L
                   d
                        f
         M
                                                                                                       p
li
                                                                                                                             p
    G
e
n
d
e
r
                                                                                                                  A
                                                                                                                        i
              p
e
n
         a
r
o
a
n
                                                                                                                             e
                       Ē
                                                                                                            n
                   \mathbf{c}
                                                                                                        \mathbf{c}
                                                                                                            A
                                                                                                                  0
         r
i
                                                                              ApplicantIncome
                  a
                       m
                                                                                                       a
                                                                                                                             t
              d
                                                                                                            m
                                                                                                                  u
Ī
D
                   t
                       p
                                                                                                       n
                                                                                                                             y
              e
n
                                                                                                            0
                                                                                                                  n
                   i
                                                                                                       tΙ
                        l
                                                                                                                             _
A
                                                                                                            u
                                                                                                                  t
                   0
                       0
                                                                                                       n
              t
                                                                                                            n
                                                                                                                        t
                                                                                                                  \bar{\mathbf{T}}
                  n
                       y
                                                                                                       c
                                                                                                                             r
              S
                                                                                                                       0
                       e
                                                                                                       o
                                                                                                                             e
                                                                                                                  \mathbf{e}
                       d
                                                                                                       m
                                                                                                                  r
                                                                                                                       y
                                                                                                       e
                                                                                                                  m
9
                             import plotly.express as px
                   а
9
                             import plotly.graph_objects as go
                                                                                                            0
                                                                                                                             b
                   t
0
                             from matplotlib import pyplot as plt
                   е
                                                                                                                             а
                                                                                                                             n
                                                                                          In [2]:
                             df =
                             pd.read csv("loan prediction.csv")
                             df.head()
                                                                                        Out[2]:
                                                                                \mathbf{L}
                                                                    \mathbf{C}
                                                                                     \mathbf{C}
                                                               A
                                                                                o
                                                                                          P
                                                          \mathbf{S}
                                                                     0
                                                                                a
                                                                                      r
                                                          e
                                                                     a
                                                                                               L
                                                D
                                                               p
li
                                                                          \mathbf{L}
                                                                                n
                                                                                      e
                                                          l
f
                                                     E
                                                                     p
                                                                                      d
                                  L
                                                     d
                                                                     p
                                                                                \mathbf{A}
                                                                c
                                  o
a
n
Ī
D
                                            a
r
                                                     u
                                                                     li
                                                                                               n
                                                         \bar{\mathbf{E}}
                                       e
n
d
                                                e
                                                               a
                                                                               m
                                                n
                                                               n
                                                                                               \overline{\mathbf{S}}
                                           r
                                                     a
                                                          m
                                                                     a
                                                                                     H
                                                d
                                                                t
                                                                          m
                                                                                u
                                                          p
l
                                                     t
                                                                                                t
                                                                     n
                                       e
r
                                                e
                                                                I
                                                                          0
                                                                                n
                                                     i
                                           e
                                                                    tΙ
                                                                                          \bar{\mathbf{A}}
                                                n
                                                               n
                                                                          u
                                                                                t
                                                     0
                                                          0
                                                                     n
                                                t
                                                                                      t
                                                                \mathbf{c}
                                                                          n
                                                                                \bar{\mathbf{T}}
                                                     n
                                                                     \mathbf{c}
                                                          y
                                                                                           r
                                                                                                u
                                                               o
                                                                                      0
                                                          e
                                                                     0
                                                               m
                                                                                e
                                                          d
                                                                    m
                                                                                r
                                                                e
                                                                     e
                                  L
P
                                                     G
                                                     r
                                                                                           U
                                  0
0
1
                                       M
                                                               5
8
4
9
                                                                                3
6
                                                     a
                                                                     0
                                                                          N
                                                                                      1
                                                          N
                                       a
                                           N
                                                     d
                                                0
                                                                                               Y
                                                                                           b
                                                                     0
                                       1
                                            o
                                                          o
                                                     u
                                                                                           a
                                  0
                                       e
                                                     a
                                  0
                                                     t
                                  2
                                                     e
                                                                          1
2
8
.
                                                                                           R
                                  L
                                                     G
                                                                                3
6
                                       M
                                                               4
5
8
                                                                     5
0
8
                                  P
0
                                            Y
                                                                                           u
                                                     r
                                                          N
                                                1
                                                                                               N
                                                     a
                                                                                           r
                                       1
                                                          o
                                                                               0.
                                  0
                                                                                     0
                                                     d
                                                                                           a
                                                                     0
                                       e
                                                               3
                                  1
                                                     u
```

a

a

S

u



C \mathbf{C} 0 r a a d p A i n m \mathbf{c} A 0 **ApplicantIncome** t a m u n y 0 n tΙ $\bar{\mathbf{A}}$ u n n t $\bar{\mathbf{T}}$ c r u 0 0 e e m r y e m

In [3]:

df.info()

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

ApplicantIncome

| 0 | Loan_ID | 614 | non-null |
|-------|---------------------|------|------------|
| objec | t | | |
| | Gender | 601 | non-null |
| objec | | | |
| 2 | Married | 611 | non-null |
| objec | | | |
| 3 | Dependents | 599 | non-null |
| objec | t | | |
| 4 | Education | 614 | non-null |
| objec | t | | |
| 5 | Self_Employed | 582 | non-null |
| objec | t | | |
| 6 . | ApplicantIncome | 614 | non-null |
| int64 | | | |
| | CoapplicantIncome | 614 | non-null |
| float | 64 | | |
| | | 592 | non-null |
| float | 64 | | |
| | Loan_Amount_Term | 600 | non-null |
| float | | | |
| 10 | Credit_History | 564 | non-null |
| float | 64 | | |
| 11 | Property_Area | 614 | non-null |
| objec | t | | |
| 12 | Loan_Status | 614 | non-null |
| objec | t | | |
| dtype | s: float64(4), int6 | 4(1) | , object(8 |
|) | | | |
| memor | y usage: 62.5+ KB | | |
| | | | In [4]: |
| profi | le = ProfileReport(| df, | |
| | ="Analysis report f | | lata |
| Analy | | | |
| _ | le.to notebook ifra | me() | |
| | le.to_file("data_an | | sis.html") |
| Summa | rize dataset: 0% | | 0/ |
| | :00, ?it/s] | | 1 07 |

| L o a n — I D | G e n d e r | M a r r i e d | D e p e n d e n t s | E d u c a t i o n | S e l f — E m p l o y |
|---------------------------------|----------------------------|---------------------------------|---------------------|---|---|
| | | | S | n | |
| | | | | | e |
| | | | | | d |

| m rya e m | | A m o u n t | | C r e d i t t - H i s t o r y | P r o p e r t y - A r e a | L o a n — S t a t u 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

- <u>Overview</u>
- <u>Variables</u>
- <u>Interactions</u>
- <u>Correlations</u>
- Missing values
- <u>Sample</u>

Overview

- <u>Overview</u>
- Alerts 23
- Reproduction

Dataset statistics

| Number of observations | 614 |
|------------------------|------|
| Missing cells | 149 |
| Missing cells (%) | 1.9% |
| Duplicate rows | 0 |

Number of variables

13

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 High distinct values cardinality

ApplicantIncome is highly correlated High with LoanAmount correlation

LoanAmount is highly correlated High with ApplicantIncome correlation

ApplicantIncome is highly correlated High with LoanAmount correlation

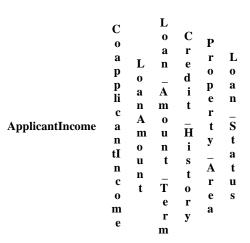
LoanAmount is highly correlated High with ApplicantIncome correlation

| 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

| | | | | | \mathbf{S} |
|---|--------|---|--------|--------------|--------------|
| | | | ъ | | e |
| | | | D | \mathbf{E} | l |
| L | | M | e | d | f |
| 0 | G | a | p | u | _ |
| a | e | r | e | c | \mathbf{E} |
| n | n d | r | n d | a | m |
| _ | | i | | t | р |
| I | e | e | e | i | ì |
| D | r | d | n t | 0 | 0 |
| | | | ι S | n | y |
| | | | 3 | | e |
| | | | | | А |



| 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 |
| Uniform | Loan_ID is uniformly distributed |
| Unique | Loan_ID has unique values |

Dependents has 15 (2.4%) missing

Reproduction

Analysis started 2022-11-09 09:43:02.856825

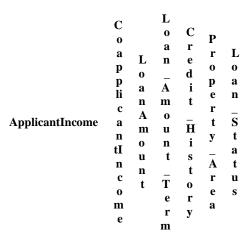
zeros

Zeros

CoapplicantIncome has 273 (44.5%)

2022-11-09 09:43:17.092955 Analysis finished

> 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

LP002328

1

LP002305

1

LP002308

1

LP002314

1

Other values (609)

609

- <u>Overview</u>
- <u>Categories</u>
- <u>Words</u>
- <u>Characters</u>

Length

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

Characters and Unicode

| | | | | | \mathbf{S} |
|---|---|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | l |
| L | ~ | \mathbf{M} | e | d | f |
| 0 | G | a | p e | u | _ |
| a | e | r | | c | Ē |
| n | n | r | n | a | m |
| | d | i | d | t | p |
| Ī | e | e | e | i | l |
| D | r | d | n | | |
| υ | | u | t | 0 | 0 |
| | | | S | n | y |
| | | | 5 | | e |
| | | | | | d |

| ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — Ā m o u n t — T e r m | C r e d i t t | P r o p e r t y - A r e a | L o a n - S t a t u s |
|-----------------|---------------------------------|---------------------|---------------------------------|---------------|---------------------------|-----------------------|
|-----------------|---------------------------------|---------------------|---------------------------------|---------------|---------------------------|-----------------------|

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?

100.0%

Sample

Unique (%)

1st row LP001002

2nd row LP001003

3rd row LP001005

4th row LP001006

| ApplicantIncome | C o a p p li c a n tI n c o m | L o a n A m o u n t | L o a n — Ā m o u n t — T e r m | C r e d i t t H i s t o r y | Property-Area | L o a n — S t a t u s |
|-----------------|-------------------------------|---------------------|---------------------------------|-----------------------------|---------------|-----------------------|
| | e | | r m | y | | |

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 | |

|--|

| | | 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 | |

| L o a n Ī D | G e n d e r | M a r r i e d | D e p e n d e n t s | E d u c a t i o n | S e l f | | | ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — Ā m o u n t — T e r m | C r e d i t -H i s t o r y |
|----------------------------|----------------------------|---------------------------------|---------------------|-------------------|---------|----------|---|-----------------|---------------------------------|---------------------|---------------------------------|----------------------------|
| | | | | | | | | 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% | | | | |

 $\begin{array}{ccc} & a \\ A & t \\ r & u \end{array}$

0 e

• <u>Characters</u>

604

98.4%

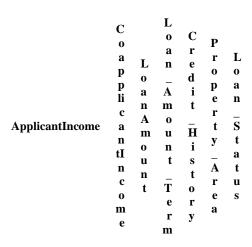
Other values (604)

| | | | | | \mathbf{S} |
|---|--------|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | e | \mathbf{E} | l |
| L | G | \mathbf{M} | | d | f |
| 0 | e | a | p e | u | _ |
| a | n | r | | c | \mathbf{E} |
| n | d | r | n d | a | m |
| _ | | i | | t | p |
| Ι | e r | e | e n | i | l |
| D | 1 | d | t | 0 | 0 |
| | | | | n | y |
| | | | S | | e |
| | | | | | d |

- <u>Categories</u><u>Scripts</u><u>Blocks</u>

Most occurring characters

| /alue | 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%

357

7.3%

Most occurring categories

Other values (2)

Value Count Frequency (%)

Decimal Number 3684 75.0%

Uppercase Letter 1228 25.0%

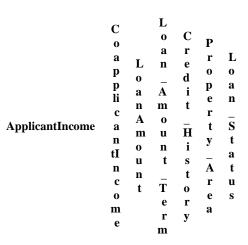
Most frequent character per category

Decimal Number

| e m | ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — A m o u n t — T e r m | C r e d i t T H i s t o r y | P r o p e r t y - A r e a | II of a second s |
|-----|-----------------|---------------------------------|---------------------|---------------------------------|-----------------------------|---------------------------|--|
|-----|-----------------|---------------------------------|---------------------|---------------------------------|-----------------------------|---------------------------|--|

| - 5 () | 1 | | |
|--------|---|------|---|
| 38.1% | | 1403 | 0 |
| 13.3% | | 491 | 1 |
| 13.0% | | 478 | 2 |
| 5.5% | | 203 | 4 |
| 5.4% | | 198 | 3 |
| 5.1% | | 189 | 8 |
| 5.0% | | 183 | 7 |
| 4.9% | | 182 | 9 |

Value Count Frequency (%)



Uppercase Letter

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

Most occurring scripts

Most frequent character per script

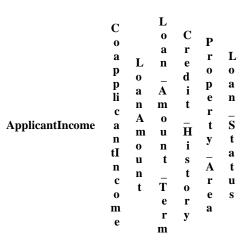
Common

| e m | ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — A m o u n t — T e r m | C r e d i t T H i s t o r y | P r o p e r t y - A r e a | II of a second s |
|-----|-----------------|---------------------------------|---------------------|---------------------------------|-----------------------------|---------------------------|--|
|-----|-----------------|---------------------------------|---------------------|---------------------------------|-----------------------------|---------------------------|--|

| - 5 () | 1 | | |
|--------|---|------|---|
| 38.1% | | 1403 | 0 |
| 13.3% | | 491 | 1 |
| 13.0% | | 478 | 2 |
| 5.5% | | 203 | 4 |
| 5.4% | | 198 | 3 |
| 5.1% | | 189 | 8 |
| 5.0% | | 183 | 7 |
| 4.9% | | 182 | 9 |

Value Count Frequency (%)





4.8%

Latin

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% |

| | | | | | \mathbf{S} |
|---|--------|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | e | \mathbf{E} | l |
| L | G | \mathbf{M} | | d | f |
| 0 | e | a | p e | u | _ |
| a | n | r | n | c | \mathbf{E} |
| n | d | r | d | a | m |
| _ | | i | | t | p |
| Ι | e r | e | e n | i | l |
| D | 1 | d | t | 0 | 0 |
| | | | S | n | y |
| | | | 3 | | e |
| | | | | | d |

182 3.7% Other values (2) 357 7.3%

<u>Gender</u> 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

| | | | | | \mathbf{S} |
|---|---|--------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | l |
| L | ~ | M | e | d | f |
| 0 | G | a | p e | u | |
| a | e | r | | c | Ē |
| n | n | | n | a | m |
| | d | r i | d | t | р |
| Ī | e | e | e | i | l |
| D | r | d | n | | |
| υ | | u | t | 0 | 0 |
| | | | S | n | y |
| | | | | | e |
| | | | | | d |

| ApplicantIncome | C o a p p li c a n tI n c o m | L o a n A m o u n | o a n -A m o u n t -T e | C r e d i t - H i s t o r | P r o p e r t y - A r e a | |
|-----------------|--|---|-------------------------|---------------------------------|--------------------------------------|--|
| ApplicantIncome | p li c a n tI n c | a n A m o u n | m o u n t | i t - H i s t | e r t y - A r e | |

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 ?

| | | | | | \mathbf{S} |
|---|--------|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | l |
| L | C | \mathbf{M} | e | d | f |
| 0 | G | a | p e | u | _ |
| a | e | r | | c | E |
| n | n d | r | n d | a | m |
| _ | | r i | | t | p |
| Ι | e | e | e | i | l |
| D | r | d | n t | 0 | 0 |
| | | | | n | y |
| | | | S | | e |
| | | | | | d |

| ApplicantIncome | C o a p p li c a n tI n c o m | L o a n A m o u n t | L o a n — Ā m o u n t — T e r m | C r e d i t t H i s t o r y | P r o p e r t y - A r e a | I o a m - S t a t u s |
|-----------------|-------------------------------|---------------------|---------------------------------|-----------------------------|---------------------------|-----------------------|
|-----------------|-------------------------------|---------------------|---------------------------------|-----------------------------|---------------------------|-----------------------|

Unique (%) 0.0%

Sample

1st row Male

2nd row Male

3rd row Male

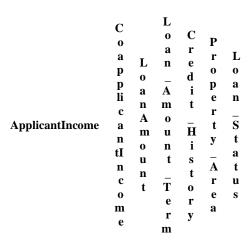
4th row Male

5th row Male

Common Values

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

| L o a n _ I D | G e n d e r | M a r r i e d | D e p e n d e n t s | E d u c a t i o n | S e l f E m p l o y |
|---------------------------------|----------------------------|---------------------------------|---------------------|---|---------------------|
| | | | | n | y |
| | | | S | | e |
| | | | | | d |



Length

Histogram of lengths of the category

Category Frequency Plot

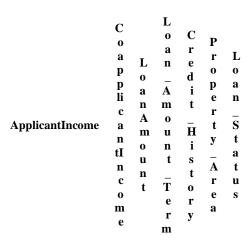
| Value | Count | Frequency (%) |
|--------|-------|---------------|
| male | 489 | 81.4% |
| female | 112 | 18.6% |

- <u>Characters</u>
- <u>Categories</u>
- <u>Scripts</u>
- <u>Blocks</u>

Most occurring characters

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

| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | D r d n o o o s n y e d |
|--|-------------------------|
|--|-------------------------|



Most occurring categories

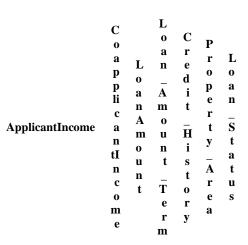
| Value | Count | Frequency (%) |
|------------------|-------|---------------|
| Lowercase Letter | 2027 | 77.1% |
| Uppercase Letter | 601 | 22.9% |

Most frequent character per category

Lowercase Letter

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





Uppercase Letter

Most occurring scripts

Most frequent character per script

Latin

| | | | | | \mathbf{S} |
|--------------|--------|--------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | l |
| \mathbf{L} | G | M | e | d | f |
| 0 | | a | p e | u | _ |
| a | e n | r | e n | c | \mathbf{E} |
| n | d | r | d | a | m |
| _ | | r i | | t | p |
| Ι | e | e | e | i | l |
| D | r | d | n t | 0 | 0 |
| | | | s S | n | y |
| | | | 3 | | e |
| | | | | | d |

| Frequency (%) | Count | Value |
|---------------|-------|-------|
| 27.1% | 713 | e |
| 22.9% | 601 | a |
| 22.9% | 601 | 1 |
| 18.6% | 489 | M |
| 4.3% | 112 | F |
| 4.3% | 112 | m |

Most occurring blocks

Value Count Frequency (%)

ASCII 2628 100.0%

Most frequent character per block

ASCII

| S | | | | | |
|---|--------------|--------|--------------|---|----|
| e | | D | | | |
| 1 | \mathbf{E} | | | | |
| f | d | e | \mathbf{M} | ~ | L |
| | u | p e | a | G | 0 |
| Ē | c | e | r | e | a |
| m | | n | | n | n |
| | a | d | r i | d | 11 |
| p | t | e | 1 | e | _ |
| l | i | n | e | r | Ι |
| 0 | 0 | | d | 1 | D |
| y | n | t | | | |
| e | | S | | | |
| ď | | | | | |

| ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | o a n A m o u n t T e r m | C r e d i t T H i s t o r y | P r o p e r t y - A r e a | II Car T S II a II u S |
|-----------------|---------------------------------|--------------------------------------|---------------------------|---|-------------------------------------|--|
| ApplicantIncome | a p p li c a n tI n c | o a n A m o u n | a n A m o u n t T e r | r e d i t H i s t o r | r o p e r t y | |

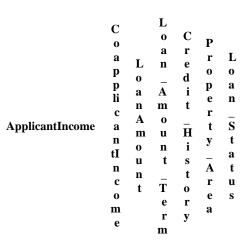
| Value | Count | Frequency (%) |
|-------|-------|---------------|
| e | 713 | 27.1% |
| a | 601 | 22.9% |
| 1 | 601 | 22.9% |
| M | 489 | 18.6% |
| F | 112 | 4.3% |
| m | 112 | 4.3% |

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 | | |

- <u>Common Values</u>
- <u>Category Frequency Plot</u>

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

Dependents

Categorical

HIGH CORRELATION MISSING

Distinct

4

| | | | | | \mathbf{S} |
|---|---|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | l |
| L | ~ | \mathbf{M} | e | d | f |
| 0 | G | a | p e | u | |
| a | e | r | | c | Ē |
| n | n | r | n | a | m |
| | d | i | d | t | p |
| Ī | e | e | e | i | ì |
| D | r | d | n | 0 | 0 |
| _ | | u | t | n | |
| | | | S | 11 | y |
| | | | | | e |
| | | | | | А |

| ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — A m o u n t — T e r m | C r e d i t T H i s t o r y | P r o p e r t y - A r e a | L o a n - S t a t u s |
|-----------------|---------------------------------|---------------------|---------------------------------|-----------------------------|---------------------------|---|
| | | | m | | | |

Distinct (%) 0.7%

Missing 15

Missing (%) 2.4%

51

4.9 KiB Memory size 0 345 1 102 2 101

3+

- Overview Categories Words
- **Characters**

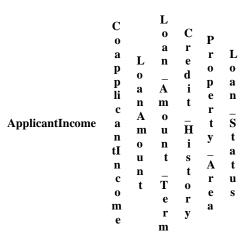
Length

Max length

Median length 1

Mean length 1.085141903

| L o a n _ I D | G e n d e r | M a r r i e d | D e p e n d e n | E d u c a t i o | S e l f E m p l o v |
|---------------------------------|----------------------------|---------------------------------|-----------------|--------------------------------------|---------------------|
| | r | e d | | | |



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

| | | | | | S |
|---|--------|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | l |
| L | C | \mathbf{M} | e | d | f |
| 0 | G | a | p e | u | _ |
| a | e | r | | c | \mathbf{E} |
| n | n d | r i | n d | a | m |
| _ | | i | | t | p |
| I | e | e | e | i | l |
| D | r | d | n t | 0 | 0 |
| | | | | n | y |
| | | | S | | e |
| | | | | | d |

| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | ApplicantIncome | n tI n c o m | o u n | n t T e r | i s t o r | y - A r e | L o a n - S t a t u s |
|--|-----------------|-----------------------------|-------------|-----------------------|-----------------------|-----------------------|-----------------------|
|--|-----------------|-----------------------------|-------------|-----------------------|-----------------------|-----------------------|-----------------------|

3rd row 0

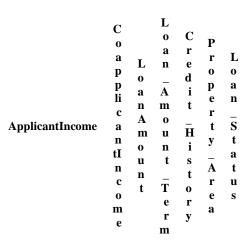
4th row 0

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

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

- <u>Characters</u>
- <u>Categories</u>
- Scripts
- Blocks

Most occurring characters

| Value | Count | Frequency (%) |
|-------|-------|---------------|
| | | |
| 0 | 345 | 53.1% |

Most occurring categories

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

Most frequent character per category

Decimal Number

| D E S D |
|---|
|---|

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

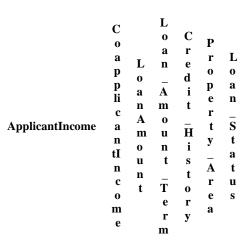
Math Symbol

Most occurring scripts

Value Count Frequency (%)

Common 650 100.0%

| | | | | | \mathbf{S} |
|---|---|---|--------|--------------|--------------|
| | | | D | _ | e |
| | | | e | \mathbf{E} | l |
| L | G | M | | d | f |
| 0 | e | a | p e | u | _ |
| a | n | r | n | c | \mathbf{E} |
| n | d | r | d | a | m |
| _ | e | i | e | t | p |
| Ī | r | e | n | i | l |
| D | • | d | t | 0 | 0 |
| | | | s | n | y |
| | | | G | | e |
| | | | | | А |



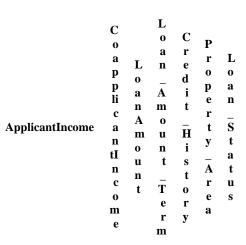
Most frequent character per script

Common

Most occurring blocks

| Value | Count | Frequency (%) | | |
|-------|-------|---------------|--|--|
| | | | | |
| ASCII | 650 | 100.0% | | |

| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | S e E l d f u _ c E a m t p i l o o n y e d |
|--|---|
|--|---|



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%

| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | S e l f _E m p l o y e |
|--|------------------------|
|--|------------------------|

| ApplicantIncome | C o a p p li c a n tI n c o m | L o a n A m o u n t | L o a n - A m o u n t - T e r m | C r e d i t -H i s t o r | P r o p e r t y - A r e a | L o a n - S t a t u s |
|-----------------|-------------------------------|---------------------|---------------------------------|--------------------------|---------------------------|-----------------------|
|-----------------|-------------------------------|---------------------|---------------------------------|--------------------------|---------------------------|-----------------------|

Missing 0

Missing (%) 0.0%

Memory size 4.9 KiB

Graduate 480

Not Graduate 134

- <u>Overview</u>
- <u>Categories</u>
- Words
- <u>Characters</u>

Length

Max length 12

Median length 8

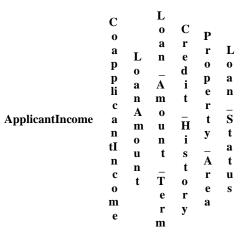
Mean length 8.872964169

Min length 8

Characters and Unicode

Total characters 5448

| | | | | | \mathbf{S} |
|---|---|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | l |
| L | ~ | \mathbf{M} | e | d | f |
| 0 | G | a | p e | u | _ |
| a | e | r | | c | Ē |
| n | n | r | n | a | m |
| | d | i | d | t | p |
| Ī | e | e | e | i | l |
| D | r | d | n | | |
| υ | | u | t | 0 | 0 |
| | | | S | n | y |
| | | | 5 | | e |
| | | | | | d |



Distinct characters 10

Distinct categories 3

Distinct scripts 2

Distinct blocks

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

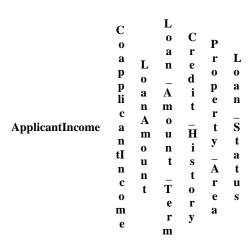
1?

Unique 0 Unique (%) 0.0%

Sample

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

5th row Graduate



Common Values

Length

Histogram of lengths of the category

Category Frequency Plot

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

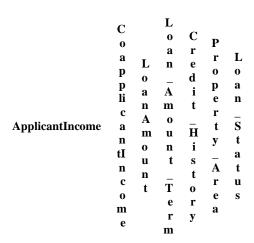
- <u>Characters</u>
- <u>Categories</u>
- <u>Scripts</u>
- Blocks

| | | | | | \mathbf{S} |
|---|--------|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | l |
| L | C | \mathbf{M} | e | d | f |
| 0 | G e | a | p e | u | _ |
| a | n | r | | c | \mathbf{E} |
| n | d | r | n d | a | m |
| _ | | i | e | t | p |
| Ī | e r | e | | i | 1 |
| D | 1 | d | n t | 0 | 0 |
| | | | s | n | y |
| | | | 3 | | e |
| | | | | | А |

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% |

| | | | | | \mathbf{S} |
|---|--------|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | l |
| L | C | \mathbf{M} | e | d | f |
| 0 | G e | a | p e | u | _ |
| a | | r | | c | \mathbf{E} |
| n | n d | r | n d | a | m |
| _ | | i | | t | p |
| I | e r | e | e n | i | l |
| D | 1 | d | t | 0 | 0 |
| | | | s S | n | y |
| | | | 5 | | e |
| | | | | | d |



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% |

Uppercase Letter

Space Separator

Most occurring scripts

| L o a n - I D | G e n d e r | M a r r i e | D e p e n d e n | E d u c a t i | S e l f — E m p l |
|---------------------------------|----------------------------|----------------------------|--------------------------------------|---------------------------------|--|
| Ι | | | | | |
| D | 1 | d | t | 0 | 0 |
| | | | s | n | y |
| | | | G | | e |
| | | | | | d |

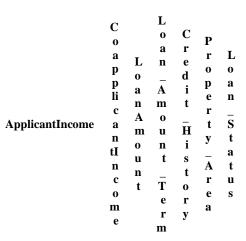
| ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n - A m o u n t - T e r m | C r e d i t T H i s t o r y | P r o p e r t y - A r e a | L o a n - S t a t u s |
|-----------------|---------------------------------|---------------------|---------------------------------|-----------------------------|---------------------------|-----------------------|
| | | | | | | |

| Frequency (% | Count | Value |
|--------------|-------|--------|
| 97.59 | 5314 | Latin |
| 2.5% | 134 | Common |

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% |



Common

Most occurring blocks

Most frequent character per block

ASCII

| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | S D e e e d f p u c E n a m e t p t n o o s e d |
|--|---|
|--|---|

| 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% |
| 0 | 134 | 2.5% |
| | 134 | 2.5% |

Self Employed Boolean

MISSING

Distinct

2

| S | | | | | |
|---|--------------|--------|--------------|---|----|
| e | | D | | | |
| 1 | \mathbf{E} | | | | |
| f | d | e | \mathbf{M} | ~ | L |
| | u | p e | a | G | 0 |
| Ē | c | e | r | e | a |
| m | | n | | n | n |
| | a | d | r i | d | 11 |
| p | t | e | 1 | e | _ |
| l | i | n | e | r | Ι |
| 0 | 0 | | d | 1 | D |
| y | n | t | | | |
| e | | S | | | |
| ď | | | | | |

| 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

- <u>Statistics</u>
- <u>Histogram</u>
- 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 81000 Maximum Range 80850

| | | | | | \mathbf{S} |
|---|--------|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | e | \mathbf{E} | l |
| L | G | \mathbf{M} | | d | f |
| 0 | e | a | p e | u | _ |
| a | n | r | n | c | \mathbf{E} |
| n | d | r | d | a | m |
| _ | | i | | t | p |
| Ι | e r | e | e n | i | l |
| D | 1 | d | t | 0 | 0 |
| | | | S | n | y |
| | | | 3 | | e |
| | | | | | d |

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%

\mathbf{C} $\begin{array}{c} C \\ r \\ e \\ d \\ i \\ t \end{array}$ P L o a n r o p e r p p li $\bar{\mathbf{A}}$ a n m \mathbf{c} A S t a t u ApplicantIncome 0 H i s t t m o u n t u n tI y n t A r e a n c T e r o r 0 m y e m

| 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

| | | | | | \mathbf{S} |
|---|--------|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | 1 |
| L | • | \mathbf{M} | e | d | f |
| 0 | G | a | p e | u | _ |
| a | e | r | | c | \mathbf{E} |
| n | n a | r | n a | a | m |
| _ | d | i | d | t | p |
| I | e | e | e | i | l |
| D | r | d | n | 0 | 0 |
| | | | t | n | y |
| | | | S | | e |
| | | | | | А |

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% |

| ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — Ā m o u n t — T e r m | C r e d i t — H i s t o r y | P r o p e r t y - A r e a | L o a n — S t a t u s |
|-----------------|---------------------------------|---------------------|---------------------------------|-----------------------------|---------------------------|-----------------------|
|-----------------|---------------------------------|---------------------|---------------------------------|-----------------------------|---------------------------|-----------------------|

| 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% |

1442 1

0.2%

CoapplicantIncome Real number (ℝ≥0)

ZEROS

| L o a n | G e n d e | M a r | D e p e n d | E d u c | S e l f E m |
|------------------|-----------------------|-------------|----------------------------|------------------|----------------------------|
| _ | | i | e | t | p |
| Ι | | e | | i | l |
| D | r | d | n t | 0 | 0 |
| | | | ι S | n | y |
| | | | 3 | | e |
| | | | | | d |

| Distinct | 287 |
|------------------------------------|----------------------------|
| Distinct (%) | 46.7% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Infinite | 0 |
| Infinite (%) | 0.0% |
| 3.5 | 1.01.045700 |
| Mean | 1621.245798 |
| Mean Minimum | 1621.245798 |
| | |
| Minimum | 0 |
| Minimum Maximum | 0 41667 |
| Minimum Maximum Zeros | 0 41667 273 |
| Minimum Maximum Zeros Zeros (%) | 0 41667 273 44.5% |

- Statistics Histogram

- <u>Common values</u>
- 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

Interquartile range (IQR) 2297.25

Range

Descriptive statistics

Standard deviation 2926.248369

41667

Coefficient of variation (CV) 1.804938136

| | | | | | \mathbf{S} |
|---|--------|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | e | \mathbf{E} | l |
| L | G | \mathbf{M} | | d | f |
| 0 | e | a | p e | u | _ |
| a | n | r | | c | \mathbf{E} |
| n | d | r | n d | a | m |
| _ | | i | | t | p |
| Ι | e r | e | e n | i | l |
| D | 1 | d | t | 0 | 0 |
| | | | | n | y |
| | | | S | | e |
| | | | | | d |

| ApplicantIncome | C o a p p li c a n tI n c o m | L o a n A m o u n t | $\begin{array}{c} L \\ o \\ a \\ n \\ \\ \hline A \\ m \\ o \\ u \\ n \\ t \\ \\ \hline T \\ e \\ r \\ m \end{array}$ | C r e d i t T H i s t o r y | P r o p e r t y - A r e a | L o a n -S t a t u s |
|-----------------|-------------------------------|---------------------|---|-----------------------------|---------------------------|----------------------|
|-----------------|-------------------------------|---------------------|---|-----------------------------|---------------------------|----------------------|

Kurtosis 84.95638421

> 1621.245798 Mean

Median Absolute Deviation (MAD) 1188.5

Skewness 7.491531217

> Sum 995444.92

8562929.518 Variance

Monotonicity Not monotonic

Histogram with fixed size bins (bins=50)

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

| | | | | | \mathbf{S} |
|---|--------|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | 1 |
| L | C | \mathbf{M} | e | d | f |
| 0 | G e | a | p e | u | _ |
| a | n | r | n | c | \mathbf{E} |
| n | d | r i | d | a | m |
| _ | | i | | t | p |
| Ī | e r | e | e | i | l |
| D | 1 | d | n t | 0 | 0 |
| | | | s S | n | y |
| | | | 3 | | e |
| | | | | | А |

| 2230 | 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

| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | Self Employed |
|--|---------------|
|--|---------------|

| m | m era e ry | tl u t s — a n u t s A t c n — t r u o t T o e s | $\begin{array}{cccccccccccccccccccccccccccccccccccc$ |
|---|---------------|--|--|
|---|---------------|--|--|

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

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% |

| | | | | | S |
|--------------|---|---|--------|--------------|--------------------|
| | | | D | | e |
| | | | | \mathbf{E} | 1 |
| \mathbf{L} | ~ | M | e | d | f |
| 0 | G | a | p e | u | |
| a | e | r | | c | $\bar{\mathbf{E}}$ |
| n | n | r | n | a | m |
| | d | i | d | t | p |
| _ | e | | e | | |
| I | r | e | n | i | l |
| D | - | d | t | 0 | 0 |
| | | | | n | y |
| | | | S | | e |
| | | | | | d |

<u>LoanAmount</u>

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 |
| Mean Minimum | 146.4121622 9 |
| | |
| Minimum | 9 |
| Minimum Maximum | 700 |

| | | | | | S |
|---|---|--------|--------|--------------|--------------|
| | | | ъ | | e |
| | | | D | \mathbf{E} | l |
| L | | M | e | d | f |
| | G | | n | | |
| 0 | e | a | p e | u | _ |
| a | | r | | c | \mathbf{E} |
| n | n | r | n | a | m |
| | d | r i | d | | |
| - | e | | e | t | p |
| I | | e | | i | l |
| D | r | d | n | 0 | 0 |
| | | | t | n | y |
| | | | S | ** | |
| | | | | | e |
| | | | | | d |

Negative (%) 0.0%

4.9 KiB Memory size

- **Statistics**
- <u>Histogram</u>
- <u>Common values</u>
- Extreme values

Quantile statistics

9 5-th percentile 56 Q1 100 median 128 Q3 168 95-th percentile 297.8 700 Maximum

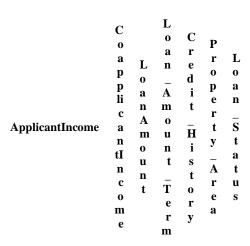
Range

691

68

Minimum

Interquartile range (IQR)



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

| ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | $\begin{array}{c} L \\ o \\ a \\ n \\ \\ \hline A \\ m \\ o \\ u \\ n \\ t \\ \\ \hline T \\ e \\ r \\ m \end{array}$ | C r e d i t T H i s t o r y | P r o p e r t y - A r e a | L o a n - S t a t u s |
|-----------------|---------------------------------|---------------------|---|-----------------------------|---------------------------|-----------------------|
|-----------------|---------------------------------|---------------------|---|-----------------------------|---------------------------|-----------------------|

| | | 2.8% |
|--|--|------|
| | | |

| 100 | 15 | |
|-----|----|------|
| | | 2.4% |
| | | |

| | | | | | \mathbf{S} |
|--------------|--------|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | 1 |
| \mathbf{L} | C | \mathbf{M} | e | d | f |
| 0 | G | a | p e | u | _ |
| a | e | r | | c | \mathbf{E} |
| n | n d | r | n d | a | m |
| _ | | i | | t | p |
| I | e r | e | e | i | l |
| D | 1 | d | n t | 0 | 0 |
| | | | s S | n | y |
| | | | 5 | | e |
| | | | | | d |

| ApplicantIncome | C o a I p o p a l i c a n n o t I u o m e | - A - A - C - M - C - M - C - M - C - C - C - C - C - C - C - C - C - C | Concernation Conce | r o p e r | L o a n - S t a t u s |
|-----------------|---|---|--|-----------------------|-----------------------|
| | m | | r y | | Б |

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% | | | |

| | ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — Ā m o u n t — T e r m | C r e d i t -H i s t o r y | Property-Area | II o a r · · · · · · · · · · · · · · · · · · |
|--|-----------------|---------------------------------|---------------------|---------------------------------|----------------------------|---------------|--|
|--|-----------------|---------------------------------|---------------------|---------------------------------|----------------------------|---------------|--|

| 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% |

| | | | | | S |
|--------|--------|---|--------------|--------|--------|
| | | | D | E | e 1 |
| L | ~ | M | e | d | f |
| 0 | G e | a | p e | u | _ |
| a | n | r | n | c | E |
| n | d | r | d | a | m |
| _ | e | i | e | t | p |
| I D | r | e | n | i | l |
| ע | | d | t | o n | 0 |
| | | | \mathbf{S} | 11 | y e |
| | | | | | a |

| ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — A m o u n t — T e r m | C r e d i t t H i s t o r y | P r o p e r t y - A r e a | L o a n — S t a t u s |
|-----------------|---------------------------------|---------------------|---------------------------------|-----------------------------|---------------------------|-----------------------|
| | | | | | | |

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

| | | | | | \mathbf{S} |
|--------------|---|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | l |
| \mathbf{L} | ~ | \mathbf{M} | e | d | f |
| 0 | G | a | p e | u | |
| a | e | r | | c | \mathbf{E} |
| n | n | | n | a | m |
| | d | r i | d | t | р |
| Ī | e | e | e | i | ì |
| D | r | d | n | 0 | 0 |
| _ | | • | t | n | y |
| | | | S | -11 | |
| | | | | | e |
| | | | | | d |

| e ry m |
|-----------|
| |

Zeros 0

Zeros (%) 0.0%

Negative 0

Negative (%) 0.0%

Memory size 4.9 KiB

- Statistics
- <u>Histogram</u>
- <u>Common values</u>
- Extreme values

Quantile statistics

Minimum 12

5-th percentile 180

Q1 360

median 360

Q3 360

95-th percentile 360

| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | S e l f E m p l o y e d |
|--|-------------------------|
|--|-------------------------|

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)

| ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | $\begin{array}{c} L \\ o \\ a \\ n \\ \\ \hline A \\ m \\ o \\ u \\ n \\ t \\ \\ \hline T \\ e \\ r \\ m \end{array}$ | C r e d i t | P r o p e r t y - A r e a | L o a n _ S t a t u s |
|-----------------|---------------------------------|---------------------|---|-------------|---------------------------|-----------------------|
|-----------------|---------------------------------|---------------------|---|-------------|---------------------------|-----------------------|

| - 5 () | 1 | | |
|--------|---|-----|-----|
| 83.4% | | 512 | 360 |
| 7.2% | | 44 | 180 |
| 2.4% | | 15 | 480 |
| 2.1% | | 13 | 300 |
| 0.7% | | 4 | 240 |
| 0.7% | | 4 | 84 |
| 0.5% | | 3 | 120 |
| 0.3% | | 2 | 60 |

Value Count Frequency (%)

| | | | | | \mathbf{S} |
|---|---|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | l |
| L | G | \mathbf{M} | e | d | f |
| 0 | e | a | p e | u | _ |
| a | n | r | | c | \mathbf{E} |
| n | d | r i | n d | a | m |
| _ | e | i | e | t | p |
| Ī | r | e | n | i | l |
| D | 1 | d | t | 0 | 0 |
| | | | s | n | y |
| | | | 3 | | e |
| | | | | | А |

- 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% |

| ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — A m o u n t — T e r m | C r e d i t t - H i s t o r y | P r o p e r t y - A r e a | L o a n - S t a t u s |
|-----------------|---------------------------------|---------------------|---------------------------------|-------------------------------|---------------------------|-----------------------|
| | e | | | J | | |

| 0.5% | 3 | 120 |
|---------------|-------|-------------------|
| 7.2% | 44 | 180 |
| 0.7% | 4 | 240 |
| 2.1% | 13 | 300 |
| 83.4% | 512 | 360 |
| 2.4% | 15 | 480 |
| Frequency (%) | Count | ⁷ alue |
| 2.4% | 15 | 480 |
| 83.4% | 512 | 360 |
| | 13 | 300 |

| L G M o e a a n r d i I e e D r d | D I e p e n d e n t s | E d u c a t i o n | S e l f _E m p l o y e d |
|-----------------------------------|-----------------------|---|--------------------------|
|-----------------------------------|-----------------------|---|--------------------------|

| ApplicantIncome | C o a p p li c a n tI n c o m | L o a n A m o u n t | $\begin{array}{c} L \\ o \\ a \\ n \\ \end{array}$ | C r e d i t t i s t o r y | P r o p e r t y - A r e a | I o a n S t a t u s |
|-----------------|-------------------------------|---------------------|--|---------------------------|---------------------------|---------------------|
|-----------------|-------------------------------|---------------------|--|---------------------------|---------------------------|---------------------|

| 240 | 4 | 0.7% |
|-----|----|------|
| 180 | 44 | 7.2% |
| 120 | 3 | 0.5% |
| 84 | 4 | 0.7% |
| 60 | 2 | 0.3% |

<u>Credit History</u> Categorical

36

12

2

1

0.3%

0.2%

HIGH CORRELATION HIGH CORRELATION

| ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — Ā m o u n t — T e r m | C r e d i t t H i s t o r y | Property | L o a n — S t a t u s |
|-----------------|----------------------------------|---------------------|---------------------------------|-----------------------------|----------------------------|-----------------------|
| ApplicantIncome | a n tI n c o m | m o u n | u n t T e | i s t o r | t y - A r e | 1 |

MISSING

 Distinct
 2

 Distinct (%)
 0.4%

 Missing
 50

 Missing (%)
 8.1%

 Memory size
 4.9 KiB

 1.0
 475

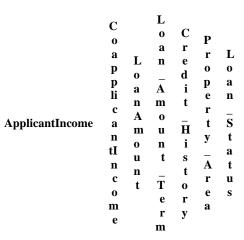
 0.0
 89

- <u>Overview</u>
- <u>Categories</u>
- Words
- <u>Characters</u>

Length

- Max length 3
- **Median length** 3
 - Mean length 3

| L o a n | G e n d | M a r r | D e p e n d | E d u c a t | S e l f - E m p |
|------------------|------------------|------------------|----------------------------|----------------------------|--------------------------------------|
| 0 | | a | þ | u | _ |
| a | | r | | c | \mathbf{E} |
| n | | r | | a | m |
| _ | | i | | t | р |
| Ι | | e | | i | l |
| D | r | d | n t | 0 | 0 |
| | | | | n | y |
| | | | S | | e |
| | | | | | d |



Min length 3

Characters and Unicode

Total characters 1692

Distinct characters 3

Distinct categories 2 2

Distinct scripts 1 2

Distinct blocks 1 2

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

Unique (%) 0.0%

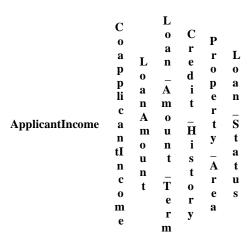
Sample

1st row 1.0

1.0

2nd row

| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | S e l f — E m p l o y e d |
|--|---------------------------|
|--|---------------------------|



3rd row 1.0

4th row 1.0

5th row 1.0

Common Values

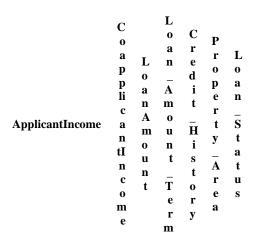
Length

Histogram of lengths of the category

Category Frequency Plot

| Value | Count | Frequency (%) |
|-------|-------|---------------|
| | | |
| 1.0 | 475 | 84.2% |





- <u>Characters</u>
- <u>Categories</u>
- <u>Scripts</u>
- <u>Blocks</u>

Most occurring characters

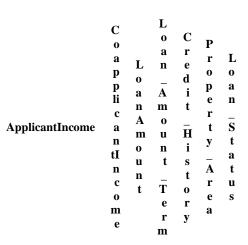
| Value | Count | Frequency (%) |
|-------|-------|---------------|
| 0 | 653 | 38.6% |
| | 564 | 33.3% |
| 1 | 475 | 28.1% |

Most occurring categories

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

Most frequent character per category





Decimal Number

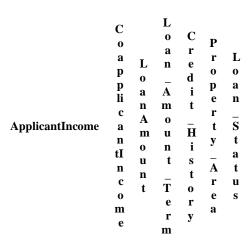
Other Punctuation

Most occurring scripts

Most frequent character per script

Common

| Value | Count | Frequency (%) | | |
|-------|-------|---------------|--|--|
| | | | | |
| 0 | 653 | 38.6% | | |



| 564 | 33.3% |
|-----|-------|
| | |
| | |

Most occurring blocks

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%

| | | | | | \mathbf{S} |
|---|--------|--------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | 1 |
| L | G | M | e | d | f |
| 0 | | a | p e | u | _ |
| a | e n | r | e n | c | \mathbf{E} |
| n | d | r | d | a | m |
| _ | | r i | | t | p |
| I | e | e | e | i | l |
| D | r | d | n t | 0 | 0 |
| | | | s S | n | y |
| | | | 3 | | e |
| | | | | | d |

Missing 0

Missing (%) 0.0%

Memory size 4.9 KiB

Semiurban 233

Urban 202

Rural 179

- <u>Overview</u>
- <u>Categories</u>
- Words

Max length

Min length

• <u>Characters</u>

Length

Median length 5

Mean length 6.517915309

5

Characters and Unicode

| | | | | | \mathbf{S} |
|---|---|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | l |
| L | ~ | \mathbf{M} | e | d | f |
| 0 | G | a | p e | u | _ |
| a | e | r | | c | Ē |
| n | n | r | n | a | m |
| | d | i | d | t | p |
| Ī | e | e | e | i | l |
| D | r | d | n | | |
| υ | | u | t | 0 | 0 |
| | | | S | n | y |
| | | | 5 | | e |
| | | | | | d |

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

| E d u c a t i o n | S e l f — E m p l o y e d |
|---|----------------------------|
| | d u c a t i |

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% |

- <u>Characters</u> <u>Categories</u>

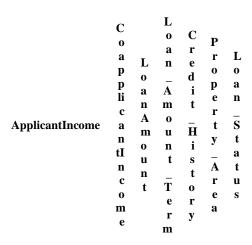
| S e E 1 | | | | | | |
|---|---|--------------|---|--------------|---|---|
| D E I | e | | | | | |
| _e E I | | | D | | | |
| I Mede | 1 | \mathbf{E} | | | | |
| | f | d | | \mathbf{M} | ~ | L |
| обари | | u | p | a | | |
| a r c r | Ē | | | r | | |
| n n n n | m | a | | | | |
| a i a t p | | | | i | | |
| Teeeil | | | | | | Ī |
| \mathbf{D} \mathbf{r} \mathbf{d} \mathbf{n} | | | | | r | |
| t n v | | | | - | | _ |
| s e | | | S | | | |
| ď | | | | | | |

- ScriptsBlocks

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 | |

| L o a n _ I D | G e n d e r | M a r r i e d | D e p e n d e n t s | E d u c a t i o n | S e l f f _ E m p l o y e d |
|---------------------------------|----------------------------|---------------------------------|---------------------|---|-----------------------------|
|---------------------------------|----------------------------|---------------------------------|---------------------|---|-----------------------------|



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% |

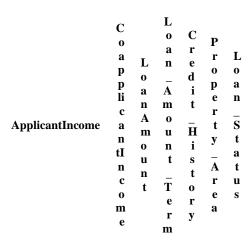
| | | | | | \mathbf{S} |
|---|--------|--------|--------|--------|--------------|
| | | | D | E | e l |
| L | ~ | M | e | d | f |
| 0 | G e | a | p e | u | _ |
| a | n | r | n | c | E |
| n | d | r i | d | a | m |
| Ī | e | | e | t i | p l |
| D | r | e d | n | 0 | 0 |
| D | | u | t | n | y |
| | | | S | | e |
| | | | | | d |

| ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — Ā m o u n t — T e r m | C r e d i t t H i s t o r y | P r o p e r t y - A r e a | |
|-----------------|---------------------------------|--------------------------------------|---------------------------------|-----------------------------|---------------------------|--|
| ApplicantIncome | p p li c a n tI n c | o a n A m o u n | n A m o u n t T e | e d i t - H i s t o r | o p e r t y - A r e | |

| b | 435 | 12.8% |
|---|-----|-------|
| n | 435 | 12.8% |
| u | 412 | 12.2% |
| e | 233 | 6.9% |
| m | 233 | 6.9% |
| i | 233 | 6.9% |
| 1 | 179 | 5.3% |

Uppercase Letter

| Value | Count | Frequency (%) |
|-------|-------|---------------|
| S | 233 | 37.9% |
| ĪĪ | 202 | 32.9% |



R 179 29.2%

Most occurring scripts

Value Count Frequency (%)

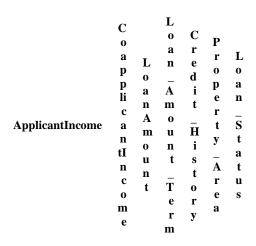
Latin 4002 100.0%

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% |

| | | | | | \mathbf{S} |
|---|---|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | l |
| L | ~ | \mathbf{M} | e | d | f |
| 0 | G | a | p e | u | |
| a | e | r | | c | \mathbf{E} |
| n | n | r | n | a | m |
| | d | i | d | t | p |
| I | e | e | e | i | ì |
| D | r | d | n | 0 | 0 |
| | | | t | n | y |
| | | | S | | e |
| | | | | | ď |



Most occurring blocks

Value Count Frequency (%)

ASCII 4002 100.0%

Most frequent character per block

ASCII

| | | 1 | - 5 () |
|---|-----|---|--------|
| 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%

Value Count Frequency (%)

| ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | $\begin{array}{c} L \\ o \\ a \\ n \\ \\ \hline A \\ m \\ o \\ u \\ n \\ t \\ \\ \hline T \\ e \\ r \\ m \end{array}$ | C r e d i t | P r o p e r t y - A r e a | L o a n - S t a t u s |
|-----------------|---------------------------------|---------------------|---|-------------|---------------------------|-----------------------|
|-----------------|---------------------------------|---------------------|---|-------------|---------------------------|-----------------------|

Other values (2) 358 8.9%

Loan Status

Boolean

HIGH CORRELATION HIGH CORRELATION

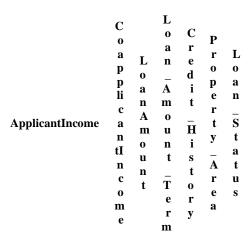
| Distinct | 2 |
|--------------|---------|
| Distinct (%) | 0.3% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Memory size | 742.0 B |
| True | 422 |
| False | 192 |

- Common Values
 Category Frequency Plot

Value Count Frequency (%)

True 422 68.7%





False 192 31.3%

Interactions

- <u>ApplicantIncome</u>
- <u>CoapplicantIncome</u>
- <u>LoanAmount</u>
- Loan Amount Term
- Loan Amount Term
- ApplicantIncome
- <u>CoapplicantIncome</u>
- LoanAmount
- Loan Amount Term
- ApplicantIncome
- <u>CoapplicantIncome</u>
- LoanAmount
- Loan Amount Term
- ApplicantIncome
- <u>CoapplicantIncome</u>
- LoanAmount
- Loan Amount Term
- <u>ApplicantIncome</u>
- <u>CoapplicantIncome</u>
- <u>LoanAmount</u>

Correlations

| | | | | | \mathbf{S} |
|---|---|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | l |
| L | ~ | \mathbf{M} | e | d | f |
| 0 | G | a | p e | u | |
| a | e | r | | c | E |
| n | n | r | n | a | m |
| | d | i | d | t | p |
| I | e | e | e | i | ì |
| D | r | d | n | 0 | 0 |
| | | | t | n | y |
| | | | S | | e |
| | | | | | d |

| ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — Ā m o u n t — T e r m | C r e d i t t - H i s t o r y | P r o p e r t y - A r e a | L o a n - S t a t u s |
|-----------------|---------------------------------|---------------------|---------------------------------|-------------------------------|---------------------------|-----------------------|
|-----------------|---------------------------------|---------------------|---------------------------------|-------------------------------|---------------------------|-----------------------|

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

Spearman's p

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

| L G e a n d d T e D | M a r r i e d | D e p e n d e n t s | E d u c a t i o n | S e l f E m p l o y e d | ApplicantIncome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | a n - A m o u n t - T e r m | C r e d i t - H i s t o r y |
|---------------------|---------------------------------|---------------------|-------------------|-------------------------|-----------------|---------------------------------|---------------------|-----------------------------|-----------------------------|
|---------------------|---------------------------------|---------------------|-------------------|-------------------------|-----------------|---------------------------------|---------------------|-----------------------------|-----------------------------|

 $\overline{\mathbf{A}}$

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.

| | | | | | \mathbf{S} |
|---|--------|--------------|--------|--------------|--------------|
| | | | D | | e |
| | | | | \mathbf{E} | 1 |
| L | G | \mathbf{M} | e | d | f |
| 0 | | a | p e | u | _ |
| a | e n | r | n | c | \mathbf{E} |
| n | d | r i | d | a | m |
| _ | | i | e | t | p |
| I | e r | e | | i | 1 |
| D | 1 | d | n t | 0 | 0 |
| | | | s | n | \mathbf{y} |
| | | | 3 | | e |
| | | | | | d |

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.

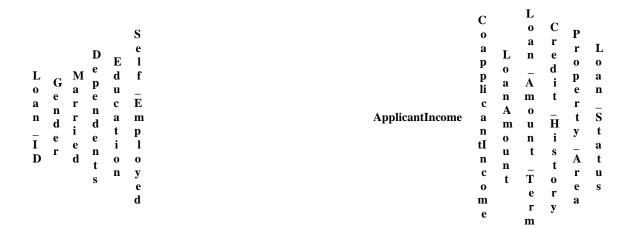
To calculate τ for two variables X and Y, one determines the number of concordant and discordant pairs of observations. τ is given by the number of concordant pairs minus the discordant pairs divided by the total number of pairs.

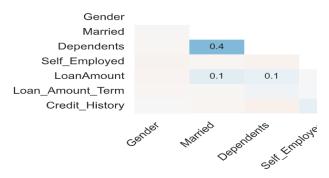
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ |
|--|
|--|

| | C | | L | | | |
|-----------------|-------------------------|---|-------------------------|-------------------|---------------------|-------------------|
| ApplicantIncome | C o a p p li c a n tI n | L o a n A m o u n | o a n n — Ā m o u n t — | C r e d i t H i s | P r o p e r t y A r | L o a n - S t a t |
| ApplicantIncome | n tI | m o u | u n | i s | y | t a |
| | C | | m | | | |

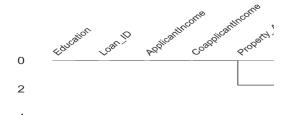
Cramér's V (φc)

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.



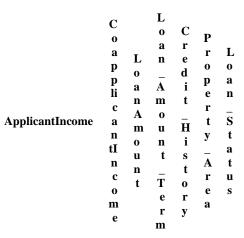


he correlation heatmap measures nullity correlation: how strongly the p



Phik (φk)

| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | ì |
|--|---|
| d O | |



Missing values

- Count
- <u>Matrix</u>
- <u>Heatmap</u>
- <u>Dendrogram</u>

A simple visualization of nullity by column.

| L G M o e a a n r d i I e e D d | D e p e n d e n t s | E d u c a t i o n | S e l f E m p l o y e d |
|---------------------------------|---------------------|---|-------------------------|
|---------------------------------|---------------------|---|-------------------------|

| ApplicantIncome | C o a p p li c a n tI n c o m | L o a n A m o u n t | a n A m o u n t r m | C r e d i t TH i s t o r y | P r o p e r t y - A r e a | L o a n - S t a t u s |
|-----------------|-------------------------------|---------------------|-----------------------|----------------------------|---------------------------|---|
|-----------------|-------------------------------|---------------------|-----------------------|----------------------------|---------------------------|---|



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

| L o a n - I D | G e n d e r | M a r r i e d | D e p e n d e n t s | E d u c a t i o n | S e l f E m p l o y e d | | | | | | | | | | | Арр | blicar | ntInc | come |
|---------------------------------|----------------------------|---------------------------------|---------------------|-------------------|-------------------------|---|--------------------------------------|----------------------------|---------------------------------|---------------------|--------------------------------------|---------------------------|------------------------------|---------------------------------|---------------------|---------------------------------|-----------------------------|-----------------------|-----------------------|
| | | | | | | | L o a n - I D | G e n d e r | M a r r i e d | D e p e n d e n t s | E d u c a t i o n | S e l f — E m p l o y e d | A p p li c a n t I n c o m e | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — Ā m o u n t — T e r m | C r e d i t t i s t t o r y | Property Area | L o a n - S t a t u s |
| | | | | | | 0 | L P 0 0 1 0 0 2 | M a l e | N o | 0 | G r a d u a t e | N o | 5 8 4 9 | 0 . 0 | N a N | 3 6 0. 0 | 1 0 | U r b a n | Y |
| | | | | | | 1 | L P 0 0 1 0 0 3 | M a l e | Y e s | 1 | G r a d u a t e | N o | 4 5 8 3 | 1 5 0 8 | 1 2 8 | 3 6 0. 0 | 1 0 | R u r a l | N |

| ApplicantIncome | C o a p p li c a n tI n c | L o a n A m o u n t | a n A m o u n t T e | C r e d i t — H i s t o r | P r o p e r t y - A r e | |
|-----------------|---------------------------|---------------------|---------------------|---------------------------|-------------------------|--|
| | | | _ | 0 | | |

| | 0 5 | | | | t e | | | | | | | | |
|---|--------------------------------------|------------------|-------------|---|---|-------------|------------------|------------------|-------------|-------------------|--------|-----------------------|---|
| 3 | L P 0 0 1 0 0 6 | M a l e | Y e s | 0 | N o t G r a d u a t e | N o | 2 5 8 3 | 2 3 5 8 | 1 2 0 | 3 6 0. 0 | 1 . 0 | U r b a n | Y |
| 4 | L P 0 0 1 0 0 8 | M a l e | N o | 0 | G r a d u a t e | N o | 6 0 0 | 0 . 0 | 1 4 1 | 3 6 0. 0 | 1 0 | U r b a n | Y |
| 5 | L P 0 0 1 0 1 | M a l e | Y e s | 2 | G r a d u a t e | Y e s | 5 4 1 7 | 4 1 9 6 | 2 6 7 | 3 6 0. 0 | 1 0 | U r b a n | Y |
| 6 | L P 0 0 1 | M a l e | Y e s | 0 | N o t G r a d | N o | 2 3 3 3 | 1 5 1 6 | 9 5 | 3 6 0. 0 | 1 0 | U r b a n | Y |

| L o a n - I D | G e n d e r | M a r r i e d | D e p e n d e n t s | E d u c a t i o n | S e l f E m p l o y e d | | | | | | | | | | | Арр | blicar | ntIno | ome | C o a p p li c a n tI n c o m e | L o a n A m o u n t | L o a n — A m o u n t — T e r m | C r e d i t - H i s t o r y | P r o p e r t y - A r e a | |
|---------------------------------|----------------------------|---------------------------------|---------------------|---|-------------------------|---|--------------------------------------|------------------|-------------|-----|--------------------------------------|--------|-----------------------|-----------------------|-------------|-------------------|--------|---|-----|---------------------------------|---------------------|---------------------------------|-----------------------------|---------------------------|--|
| | | | | | | | 1 3 | | | | a t e | | | | | | | | | | | | | | |
| | | | | | | 7 | L P 0 0 1 0 1 4 | M a l e | Y e s | 3 + | G r a d u a t e | N o | 3 0 3 6 | 2 5 0 4 | 1 5 8 | 3 6 0. 0 | 0 . 0 | S e m i u r b a n | N | | | | | | |
| | | | | | | 8 | L P 0 0 1 0 1 8 | M a l e | Y e s | 2 | G r a d u a t e | N o | 4 0 0 6 | 1 5 2 6 | 1 6 8 | 3 6 0. 0 | 1 . 0 | U r b a n | Y | | | | | | |
| | | | | | | 9 | L P 0 0 1 0 2 | M a l e | Y e s | 1 | G r a d u a t | N o | 1 2 8 4 1 | 1 0 9 6 8 | 3 4 9 | 3 6 0. 0 | 1 0 | S e m i u r b a | N | | | | | | |

Last rows

| L o a n - I D | G e n d e r | M a r r i e d | D e p e n d e n t s | E d u c a t i o n | S e l f — E m p l o y e d | | | | | | | | | | | Арр | licar | ntInc | come |
|---------------------------------|----------------------------|---------------------------------|---------------------|-------------------|---|-------------|--------------------------------------|----------------------------|---------------------------------|---------------------|---|-------------------------------|------------------------------|-------------------------------|---------------------|---------------------------------|----------------------------|-----------------------|------------------------|
| | | | | | | | L o a n - I D | G e n d e r | M a r r i e d | D e p e n d e n t s | E d u c a t i o n | S e l f f _ E m p l o y e d d | A p p li c a n t I n c o m e | C o a p p li c a n tI n c o m | L o a n A m o u n t | L o a n — Ā m o u n t — T e r m | C r e d i t -H i s t o r y | Property—Area | L oo a n - S t a t u s |
| | | | | | | 6 0 4 | L P 0 0 2 9 5 9 | F e m a l e | Y e s | 1 | G r a d u a t e | N o | 1 2 0 0 0 | 0 . 0 | 4 9 6 | 3 6 0. 0 | 1 . 0 | S e m i u r b a n | Y |
| | | | | | | 6 0 5 | L P 0 0 2 9 6 0 | M a l e | Y e s | 0 | N o t G r a d u a t e | N o | 2 4 0 0 | 3 8 0 0 | N a N | 1 8 0. 0 | 1 . 0 | U r b a n | N |

C o a p p li c a n tI n c o m e

| L o a n Ī D | G e n d e r | M a r r i e d | D e p e n d e n t s | E d u c a t i o n | S e l f = E m p l o y e d |
|----------------------------|----------------------------|---------------------------------|---------------------|-------------------|---------------------------|
|----------------------------|----------------------------|---------------------------------|---------------------|-------------------|---------------------------|

| 6 0 6 | L P 0 0 2 9 6 | M a l e | Y e s | 1 | G r a d u a t e | N o | 3 4 0 0 | 2 5 0 0 | 1 7 3 | 3 6 0. 0 | 1 0 | S e m i u r b a n | Y |
|-------------|--------------------------------------|----------------------------|-------------|---|--------------------------------------|--------|------------------|------------------|-------------|-------------------|--------|-----------------------|---|
| 6 0 7 | L P 0 0 2 9 6 4 | M a l e | Y e s | 2 | N o t G r a d u a t e | N o | 3 9 8 7 | 1 4 1 1 | 1 5 7 | 3 6 0. 0 | 1 . 0 | R u r a l | Y |
| 6 0 8 | L P 0 0 2 9 7 4 | M a l e | Y e s | 0 | G r a d u a t e | N o | 3 2 3 2 | 1 9 5 0 | 1 0 8 | 3 6 0. 0 | 1 0 | R u r a l | Y |
| 6 0 9 | L P 0 0 2 9 7 8 | F e m a l e | N o | 0 | G r a d u a t e | N o | 2 9 0 0 | 0 . 0 | 7 1 0 | 3 6 0. 0 | 1 . 0 | R u r a l | Y |

S
e
l
f
E
m
p
l
o
y
e
d D e p e n d e n t s E d u c a t i o n L o a n _ I D G e n d e r

C o a p p li $\begin{array}{c} C \\ r \\ e \\ d \\ i \\ t \end{array}$ r o p e r L o a n Ā m S t a t u s ApplicantIncome t y H i s t u n t n tI n c A r e a T e r o r y o m e

m

| 6 1 0 | L P 0 0 2 9 7 | M a l e | Y e s | 3 + | G r a d u a t e | N o | 4 1 0 6 | 0 . 0 | 4 0 | 1 8 0. 0 | 1 0 | R u r a l | Y |
|-------------|--------------------------------------|-----------------------|-------------|-----|--------------------------------------|-------------|------------------|-------------|-------------|-------------------|--------|---|---|
| 6 1 1 | L P 0 0 2 9 8 3 | M a l e | Y e s | 1 | G r a d u a t e | N o | 8 0 7 2 | 2 4 0 | 2 5 3 | 3 6 0. 0 | 1 0 | U r b a n | Y |
| 6 1 2 | L P 0 0 2 9 8 4 | M a l e | Y e s | 2 | G r a d u a t | N o | 7 5 8 3 | 0 . 0 | 1 8 7 | 3 6 0. 0 | 1 0 | U r b a n | Y |
| 6 1 3 | L P 0 0 2 9 0 | F e m a l | N o | 0 | G r a d u a t e | Y e s | 4 5 8 3 | 0 . 0 | 1 3 3 | 3 6 0. 0 | 0 . 0 | S e m i u r b a n | |