

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

In [2]:

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

Out[2]:

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

In [3]:

```
df.info()
```

RangeIndex: 614 entries, 0 to 613

Data columns (total 13 columns):

#	Column	Non-Null Count	Dtype
0	Loan_ID	614 non-null	object
1	Gender	601 non-null	object
2	Married	611 non-null	object
3	Dependents	599 non-null	object
4	Education	614 non-null	object
5	Self_Employed	582 non-null	object
6	ApplicantIncome	614 non-null	int64
7	CoapplicantIncome	614 non-null	float64
8	LoanAmount	592 non-null	float64
9	Loan_Amount_Term	600 non-null	float64

```
10 Credit_History      564 non-null    float64
11 Property_Area       614 non-null    object
12 Loan_Status         614 non-null    object
dtypes: float64(4), int64(1), object(8)
memory usage: 62.5+ KB
```

In [4]:

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

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

Analysis report for data Analysis

[Analysis report for data Analysis](#)

- [Overview](#)
- [Variables](#)
- [Interactions](#)
- [Correlations](#)
- [Missing values](#)
- [Sample](#)

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

1

LP002308

1

LP002314

1

Other values (609) 609

- [Overview](#)
- [Categories](#)
- [Words](#)
- [Characters](#)

Length

Max length 8

Median length 8

Mean length 8


Min length 8

Characters and Unicode

Total characters 4912

Distinct characters 12


Distinct categories 2 

Distinct scripts 2 

Distinct blocks 1 

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

Unique

Unique 614 

Unique (%) 100.0%

Sample

1st row LP001002

2nd row LP001003

3rd row LP001005

4th row LP001006

5th row LP001008

Common Values

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

Value	Count	Frequency (%)
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%

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

- [Characters](#)
- [Categories](#)
- [Scripts](#)
- [Blocks](#)

Most occurring characters

Value	Count	Frequency (%)
0	1403	28.6%

Value	Count	Frequency (%)
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%

Value	Count	Frequency (%)
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	4.9%

Value	Count	Frequency (%)
5	176	4.8%

Uppercase Letter

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

Most occurring scripts

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

Most frequent character per script

Common

Value	Count	Frequency (%)
0	1403	38.1%
1	491	13.3%
2	478	13.0%
4	203	

Value	Count	Frequency (%)
-------	-------	---------------

		5.5%
--	--	------

3	198	5.4%
---	-----	------

8	189	5.1%
---	-----	------

7	183	5.0%
---	-----	------

9	182	4.9%
---	-----	------

6	181	4.9%
---	-----	------

5	176	4.8%
---	-----	------

Latin

Value	Count	Frequency (%)
-------	-------	---------------

L	614	50.0%
---	-----	-------

P	614	50.0%
---	-----	-------

Most occurring blocks

Value	Count	Frequency (%)
-------	-------	---------------

ASCII	4912	100.0%
-------	------	--------

Most frequent character per block

ASCII

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

Gender

Categorical

HIGH CORRELATION
MISSING




	Distinct	2
	Distinct (%)	0.3%
	Missing	13
	Missing (%)	2.1%
	Memory size	4.9 KiB
	Male	489
	Female	112

- [Overview](#)
- [Categories](#)
- [Words](#)
- [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
Unique (%)	0.0%

Sample

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

Common Values

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

Length

Histogram of lengths of the category

Category Frequency Plot

Value	Count	Frequency (%)
male	489	81.4%

Value	Count	Frequency (%)
female	112	18.6%

- [Characters](#)
- [Categories](#)
- [Scripts](#)
- [Blocks](#)

Most occurring characters

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

Most occurring categories

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

Most frequent character per category

Lowercase Letter

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

Uppercase Letter

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

Most occurring scripts

Value	Count	Frequency (%)
Latin	2628	100.0%

Most frequent character per script

Latin

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

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

Most occurring blocks

Value	Count	Frequency (%)
ASCII	2628	100.0%

Most frequent character per block

ASCII

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

[Married](#)

Boolean

HIGH CORRELATION

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

- [Common Values](#)
- [Category Frequency Plot](#)

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

[Dependents](#)

Categorical

HIGH CORRELATION

MISSING

Distinct	4
-----------------	---



		Distinct (%)	0.7%
		Missing	15
		Missing (%)	2.4%
		Memory size	4.9 KiB
	0		345
	1		102
	2		101
	3+		51

- [Overview](#)
- [Categories](#)
- [Words](#)
- [Characters](#)

Length

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

Characters and Unicode

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

Distinct blocks 1

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

Unique 0

Unique (%) 0.0%

Sample

1st row 0

2nd row 1

3rd row 0

4th row 0

5th row 0

Common Values

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

Length

Histogram of lengths of the category

Category Frequency Plot

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

- [Characters](#)
- [Categories](#)
- [Scripts](#)
- [Blocks](#)

Most occurring characters

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

Value	Count	Frequency (%)
-------	-------	---------------

+	51	7.8%
---	----	------

Most occurring categories

Value	Count	Frequency (%)
-------	-------	---------------

Decimal Number	599	92.2%
----------------	-----	-------

Math Symbol	51	7.8%
-------------	----	------

Most frequent character per category

Decimal Number

Value	Count	Frequency (%)
-------	-------	---------------

0	345	57.6%
---	-----	-------

1	102	17.0%
---	-----	-------

2	101	16.9%
---	-----	-------

3	51	8.5%
---	----	------

Math Symbol

Value	Count	Frequency (%)
-------	-------	---------------

+	51	100.0%
---	----	--------

Most occurring scripts

Value	Count	Frequency (%)
Common	650	100.0%

Most frequent character per script

Common

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

Most occurring blocks

Value	Count	Frequency (%)
ASCII	650	100.0%

Most frequent character per block

ASCII

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

Education

Categorical




Distinct	2
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	4.9 KiB
Graduate	480
Not Graduate	134

- [Overview](#)
- [Categories](#)
- [Words](#)
- [Characters](#)

Length

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

Characters and Unicode

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

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

Unique