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

This notebook will focus on SweetViz

3. AutoEDA - SweetViz

Dataset Reference: Loan Prediction dataset from Kaggle

Features:

- General Overview - Quick insights of all variables in the dataset using the associations / correlation in the form of a heatmap (including how many duplicates, categorical/numerical/text variables etc.)
- Details about each variables / features in the dataset - missing values, distinct etc.
- Compares Train and Test datasets
- Provides visualization of target variable in context of train dataset

When To Use?

- Need some quick insights about an unknown dataset
- Use this as a basis for your further EDA analysis on top of it
- Need to compare some quick statistical insights between train and test datasets

In [1]:

```
import pandas as pd

df_train = pd.read_csv("C:/input/loan-eligible-dataset/loan-train.csv")

df_train.head()
```

Out[1]:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_Hist
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	

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

In [2]:

```
df_test = pd.read_csv("C:/input/loan-eligible-dataset/loan-test.csv")

df_test.head()
```

Out[2]:

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

In [3]:

```
df_train.shape
```

Out[3]:

```
(614, 13)
```

In [4]:

```
df_test.shape
```

Out[4]:

```
(367, 12)
```

In [6]:

```
!pip install sweetviz
```

```
Collecting sweetviz
  Downloading sweetviz-2.1.3-py3-none-any.whl (15.1 MB)
Requirement already satisfied: scipy>=1.3.2 in c:\programdata\anaconda3\lib\site-packages (from sweetviz) (1.5.2)
Collecting importlib-resources>=1.2.0
```

```
  Downloading importlib_resources-5.2.0-py3-none-any.whl (27 kB)
Requirement already satisfied: tqdm>=4.43.0 in c:\programdata\anaconda3\lib\site-packages (from sweetviz) (4.50.2)
Requirement already satisfied: matplotlib>=3.1.3 in c:\programdata\anaconda3\lib\site-packages (from sweetviz) (3.3.2)
Requirement already satisfied: jinja2>=2.11.1 in c:\programdata\anaconda3\lib\site-packages (from sweetviz) (2.11.2)
Requirement already satisfied: pandas!=1.0.0,!=1.0.1,!=1.0.2,>=0.25.3 in c:\programdata\anaconda3\lib\site-packages (from sweetvi
z) (1.1.3)
Requirement already satisfied: numpy>=1.16.0 in c:\users\mishr\appdata\roaming\python\python38\site-packages (from sweetviz) (1.2
0.1)
Requirement already satisfied: zipp>=3.1.0 in c:\programdata\anaconda3\lib\site-packages (from importlib-resources>=1.2.0->sweetvi
z) (3.4.0)
Requirement already satisfied: MarkupSafe>=0.23 in c:\programdata\anaconda3\lib\site-packages (from jinja2>=2.11.1->sweetviz) (1.
1.1)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in c:\programdata\anaconda3\lib\site-packages (from matplo
tlib>=3.1.3->sweetviz) (2.4.7)
Requirement already satisfied: pillow>=6.2.0 in c:\users\mishr\appdata\roaming\python\python38\site-packages (from matplotlib>=3.
1.3->sweetviz) (7.2.0)
Requirement already satisfied: certifi>=2020.06.20 in c:\programdata\anaconda3\lib\site-packages (from matplotlib>=3.1.3->sweetvi
z) (2020.6.20)
Requirement already satisfied: cycler>=0.10 in c:\programdata\anaconda3\lib\site-packages (from matplotlib>=3.1.3->sweetviz) (0.1
0.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\programdata\anaconda3\lib\site-packages (from matplotlib>=3.1.3->sweetviz)
(1.3.0)
Requirement already satisfied: python-dateutil>=2.1 in c:\programdata\anaconda3\lib\site-packages (from matplotlib>=3.1.3->sweetvi
z) (2.8.1)
Requirement already satisfied: six in c:\programdata\anaconda3\lib\site-packages (from cycler>=0.10->matplotlib>=3.1.3->sweetviz)
(1.15.0)
Requirement already satisfied: pytz>=2017.2 in c:\programdata\anaconda3\lib\site-packages (from pandas!=1.0.0,!=1.0.1,!=1.0.2,>=0.
25.3->sweetviz) (2020.1)
Installing collected packages: importlib-resources, sweetviz
Successfully installed importlib-resources-5.2.0 sweetviz-2.1.3
```

```
WARNING: Ignoring invalid distribution -illow (c:\users\mishr\appdata\roaming\python\python38\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\mishr\appdata\roaming\python\python38\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\mishr\appdata\roaming\python\python38\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\mishr\appdata\roaming\python\python38\site-packages)
WARNING: Ignoring invalid distribution -illow (c:\users\mishr\appdata\roaming\python\python38\site-packages)
```

In [7]:

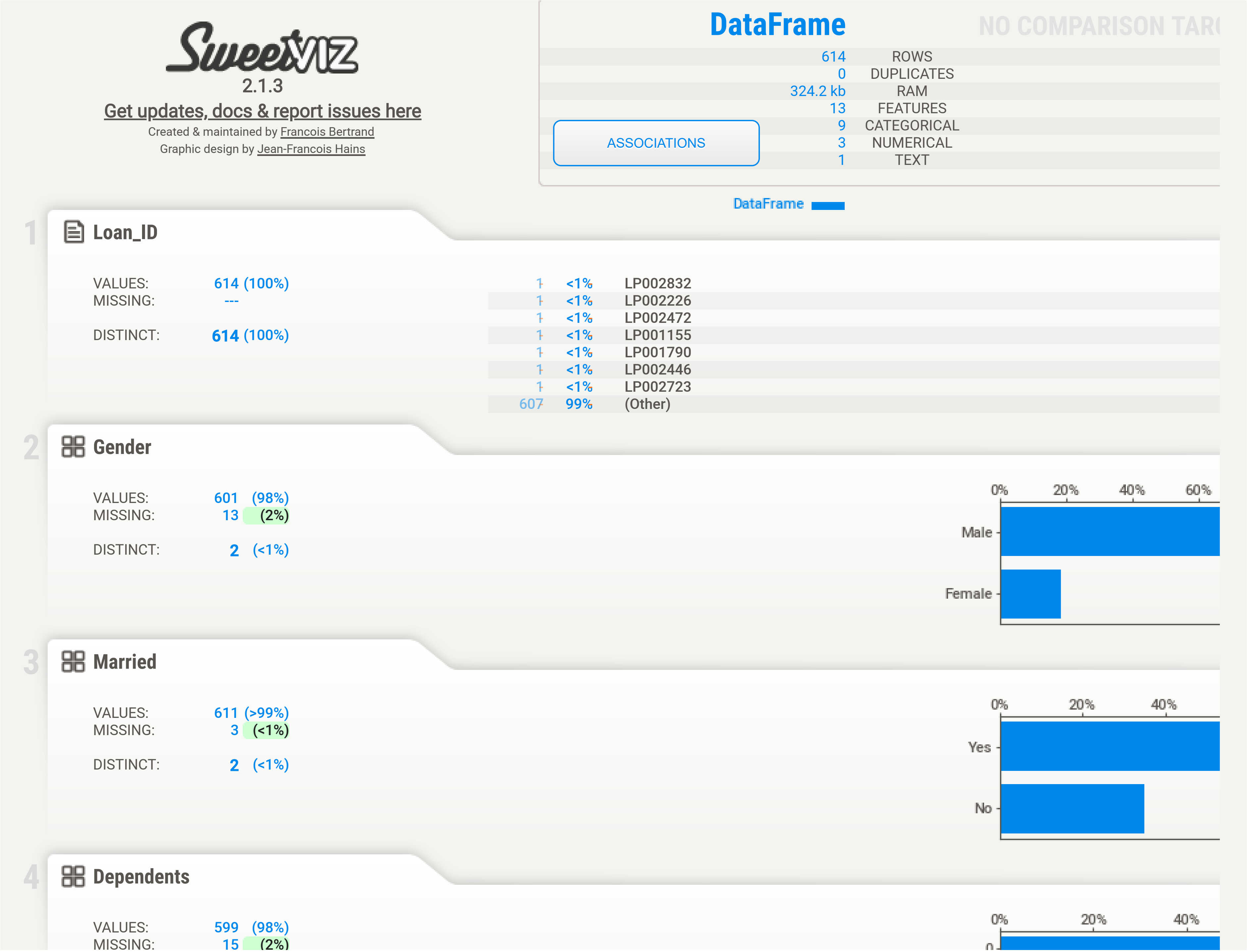
```python
import sweetviz as sv
```

In [8]:

```python
analysis_report = sv.analyze(df_train)
```

In [9]:

```python
# analysis_report.show_html() # This will generate a separate report named SWEETVIZ_REPORT.html
analysis_report.show_notebook(w="100%",h="full")
```



Sweetviz
2.1.3
Get updates, docs & report issues here
Created & maintained by Francois Bertrand
Graphic design by Jean-Francois Hains

DataFrame NO COMPARISON TAR

614	ROWS
0	DUPLICATES
324.2 kb	RAM
13	FEATURES
9	CATEGORICAL
3	NUMERICAL
1	TEXT

ASSOCIATIONS

DataFrame

1 **Loan_ID**

VALUES: 614 (100%)
MISSING: ---

DISTINCT: 614 (100%)

1	<1%	LP002832
1	<1%	LP002226
1	<1%	LP002472
1	<1%	LP001155
1	<1%	LP001790
1	<1%	LP002446
1	<1%	LP002723
607	99%	(Other)

2 **Gender**

VALUES: 601 (98%)
MISSING: 13 (2%)

DISTINCT: 2 (<1%)

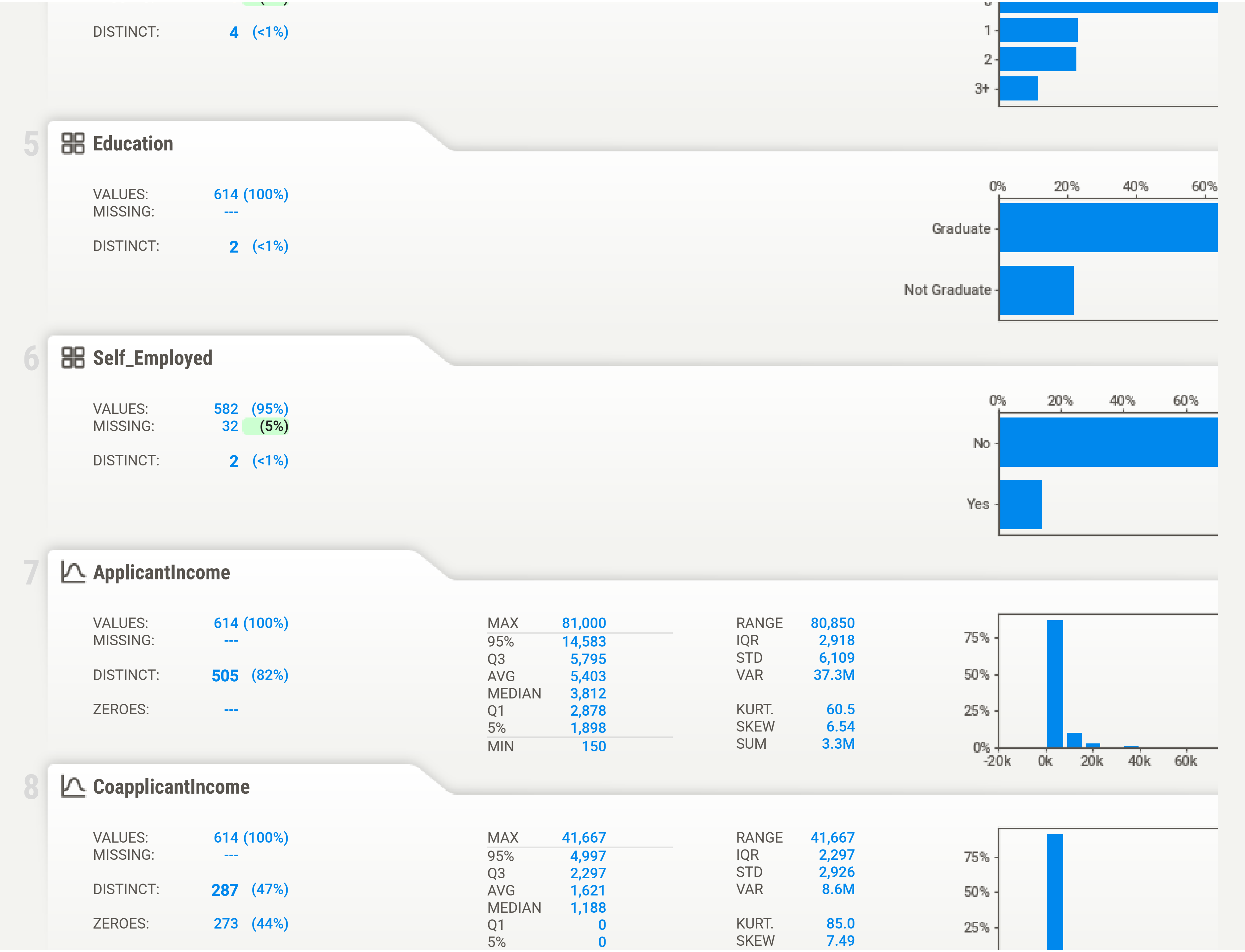
3 **Married**

VALUES: 611 (>99%)
MISSING: 3 (<1%)

DISTINCT: 2 (<1%)

4 **Dependents**

VALUES: 599 (98%)
MISSING: 15 (2%)



DISTINCT: 4 (<1%)

5 Education

VALUES: 614 (100%)
MISSING: ---

DISTINCT: 2 (<1%)

6 Self_Employed

VALUES: 582 (95%)
MISSING: 32 (5%)

DISTINCT: 2 (<1%)

7 ApplicantIncome

VALUES: 614 (100%)
MISSING: ---

DISTINCT: 505 (82%)

ZEROES: ---

MAX	81,000
95%	14,583
Q3	5,795
AVG	5,403
MEDIAN	3,812
Q1	2,878
5%	1,898
MIN	150

RANGE	80,850
IQR	2,918
STD	6,109
VAR	37.3M
KURT.	60.5
SKEW	6.54
SUM	3.3M

8 CoapplicantIncome

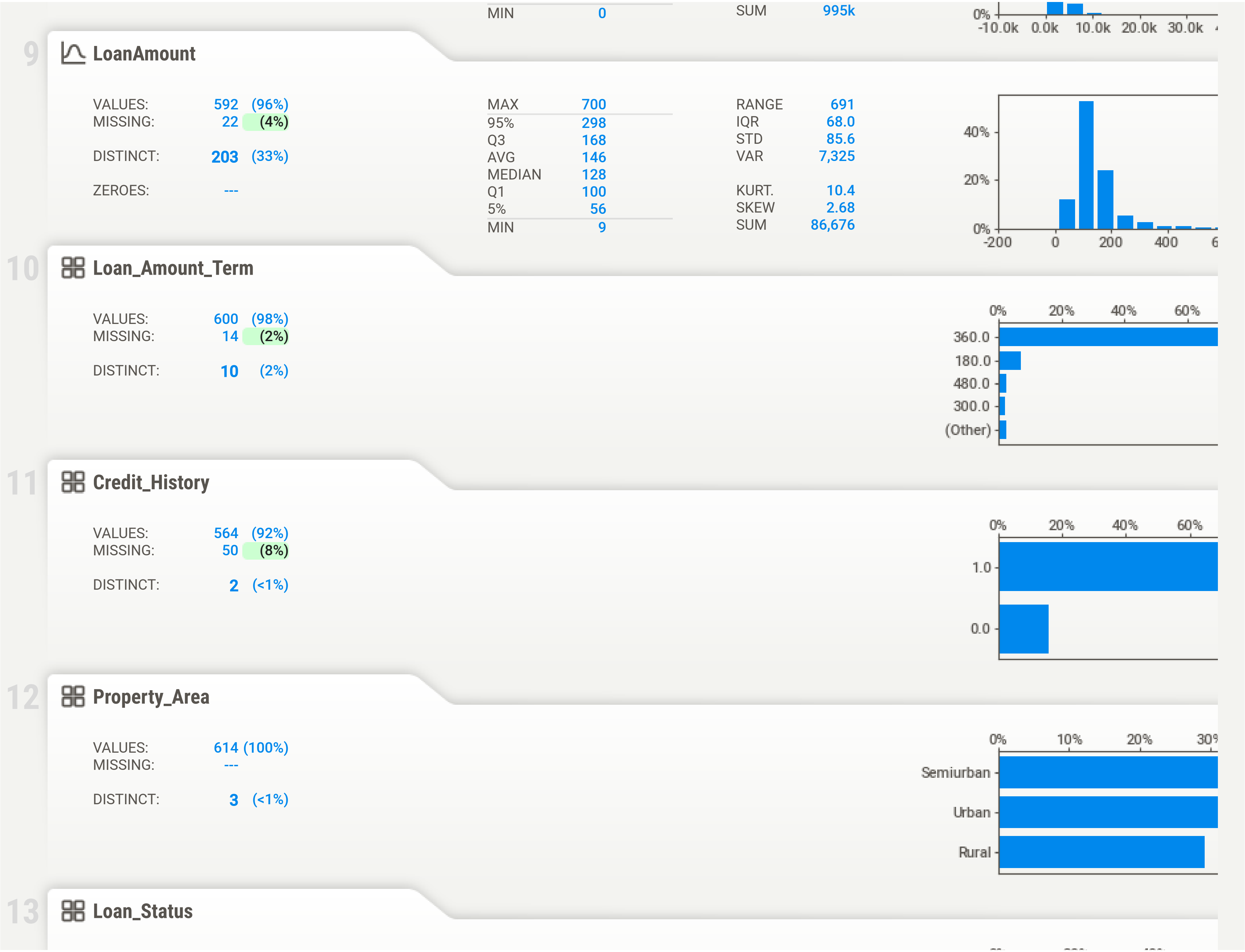
VALUES: 614 (100%)
MISSING: ---

DISTINCT: 287 (47%)

ZEROES: 273 (44%)

MAX	41,667
95%	4,997
Q3	2,297
AVG	1,621
MEDIAN	1,188
Q1	0
5%	0

RANGE	41,667
IQR	2,297
STD	2,926
VAR	8.6M
KURT.	85.0
SKEW	7.49



MIN	0
SUM	995k

9 LoanAmount

VALUES:	592	(96%)
MISSING:	22	(4%)
DISTINCT:	203	(33%)
ZEROES:	---	

MAX	700
95%	298
Q3	168
AVG	146
MEDIAN	128
Q1	100
5%	56
MIN	9

RANGE	691
IQR	68.0
STD	85.6
VAR	7,325
KURT.	10.4
SKEW	2.68
SUM	86,676

10 Loan_Amount_Term

VALUES:	600	(98%)
MISSING:	14	(2%)
DISTINCT:	10	(2%)

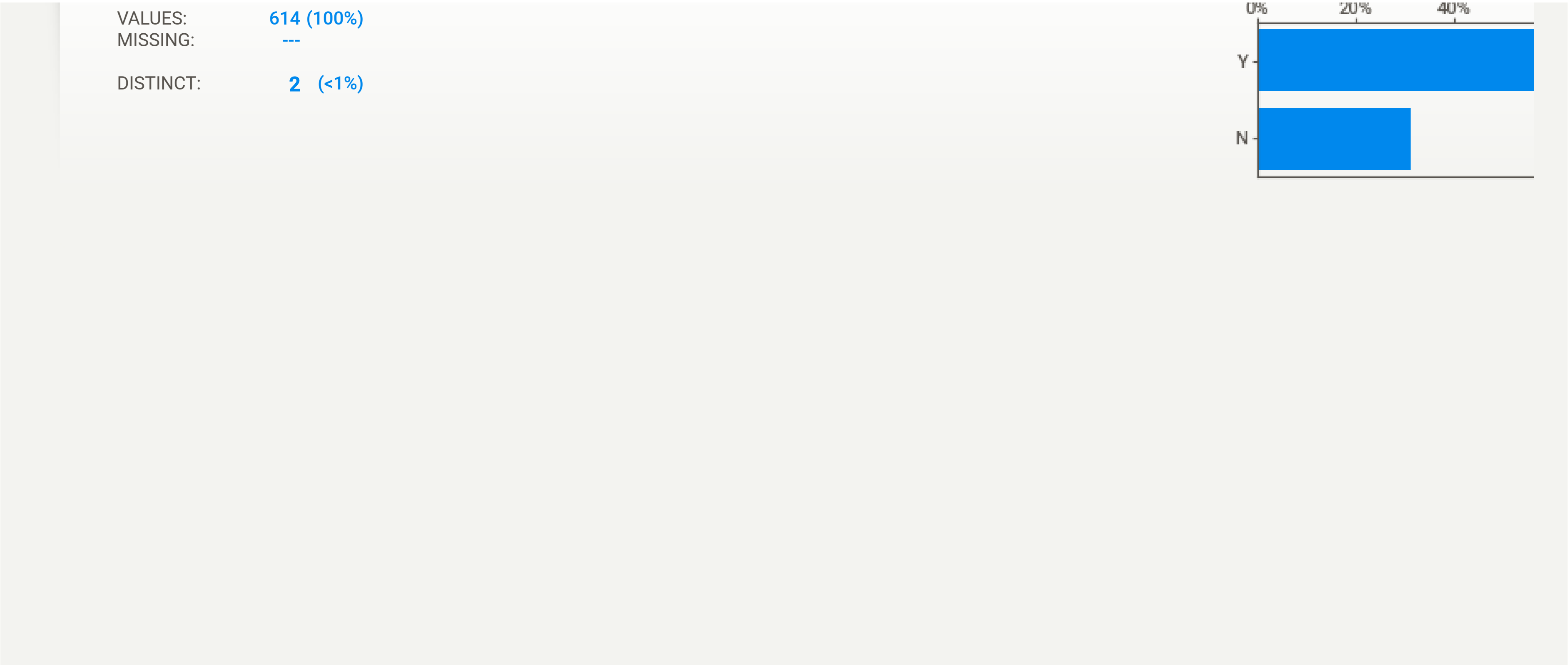
11 Credit_History

VALUES:	564	(92%)
MISSING:	50	(8%)
DISTINCT:	2	(<1%)

12 Property_Area

VALUES:	614	(100%)
MISSING:	---	
DISTINCT:	3	(<1%)

13 Loan_Status

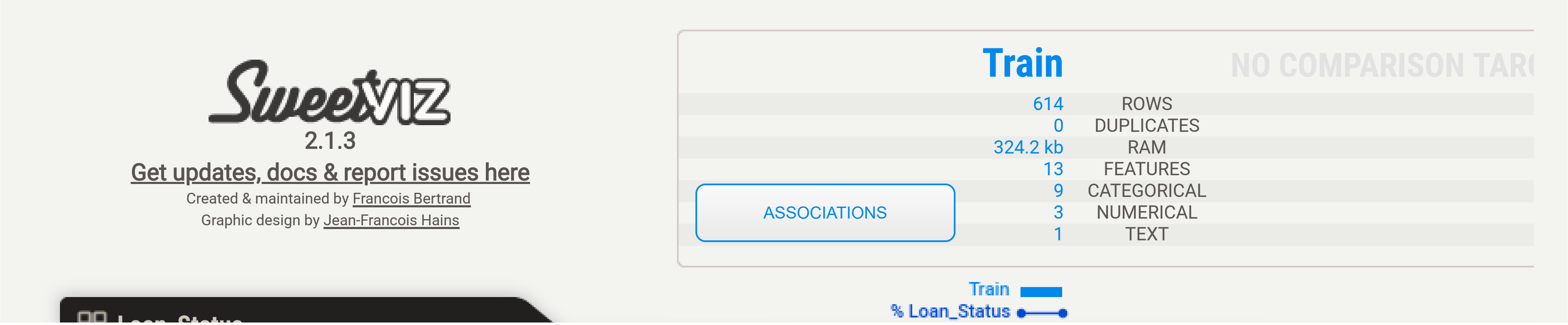


```
VALUES:      614 (100%)
MISSING:     ---

DISTINCT:    2   (<1%)
```

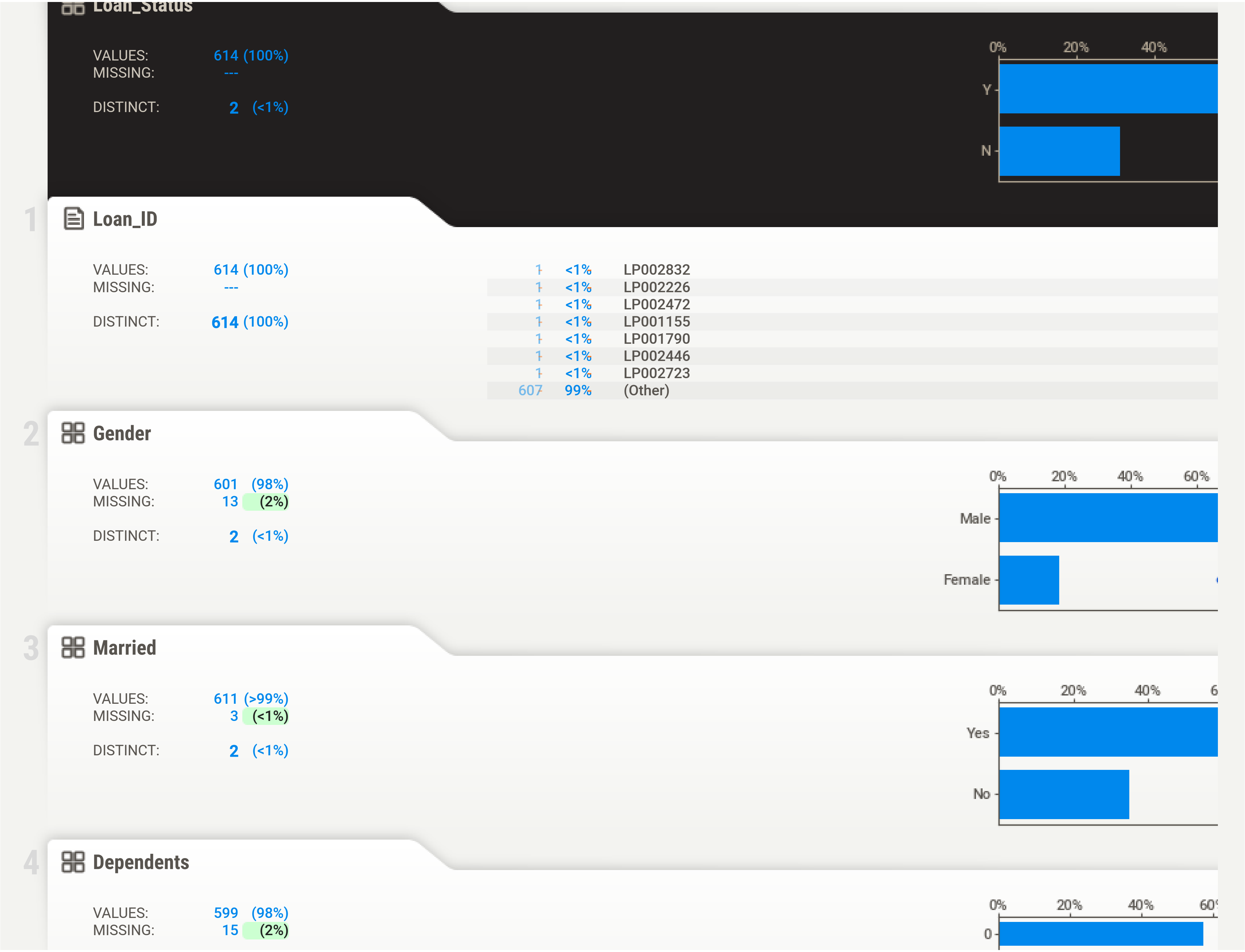
```
In [10]: analysis_report2 = sv.analyze([df_train,'Train'], target_feat='Loan_Status')
```

```
In [11]: analysis_report2.show_notebook(w="100%",h="full")
```



Sweetviz
2.1.3
Get updates, docs & report issues here
Created & maintained by Francois Bertrand
Graphic design by Jean-Francois Hains

```
      Train      NO COMPARISON TAR
        614   ROWS
          0   DUPLICATES
   324.2 kb   RAM
         13   FEATURES
          9   CATEGORICAL
          3   NUMERICAL
          1   TEXT
```



Loan_Status

VALUES: 614 (100%)
MISSING: ---

DISTINCT: 2 (<1%)

1 Loan_ID

VALUES: 614 (100%)
MISSING: ---

DISTINCT: 614 (100%)

1	<1%	LP002832
1	<1%	LP002226
1	<1%	LP002472
1	<1%	LP001155
1	<1%	LP001790
1	<1%	LP002446
1	<1%	LP002723
607	99%	(Other)

2 Gender

VALUES: 601 (98%)
MISSING: 13 (2%)

DISTINCT: 2 (<1%)

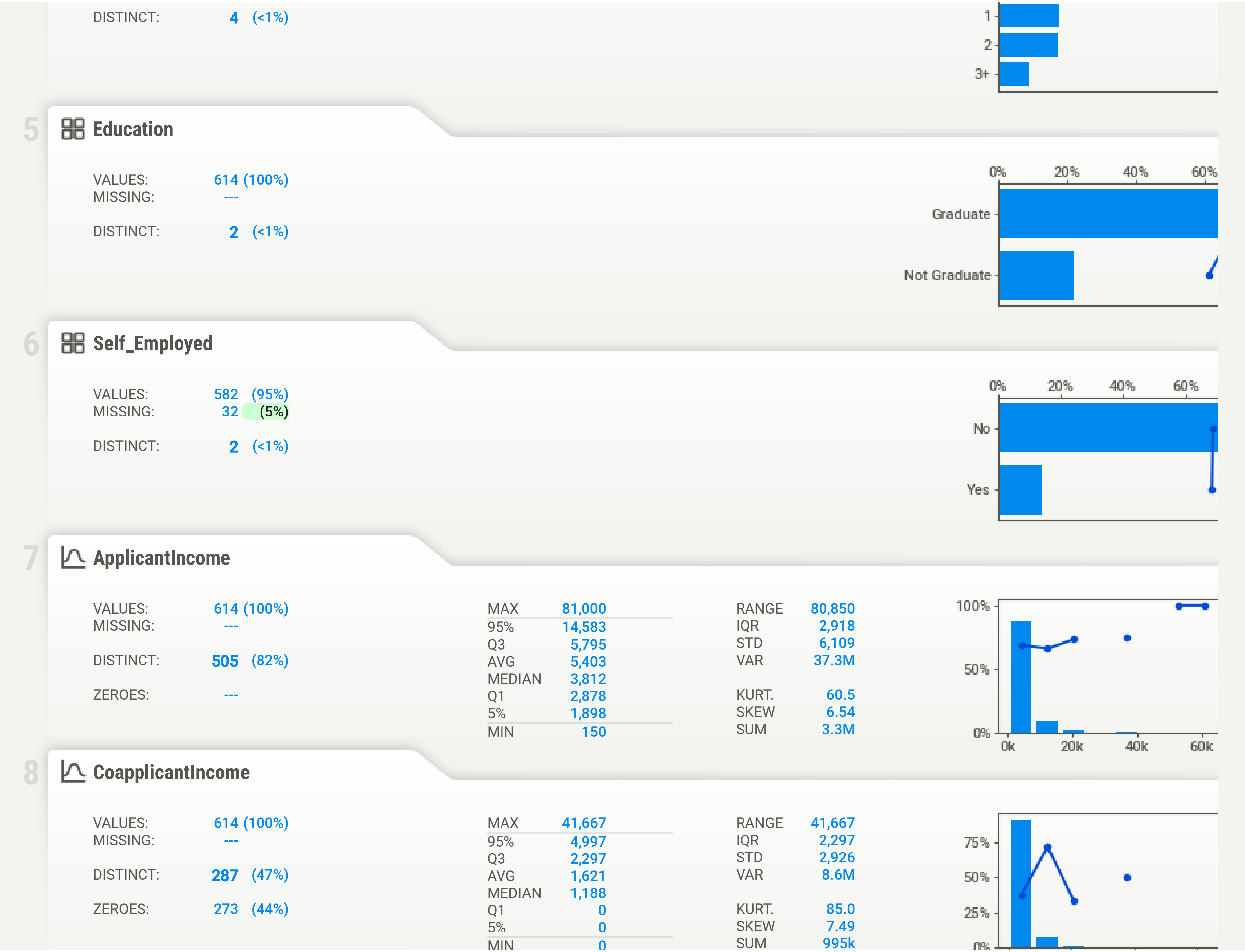
3 Married

VALUES: 611 (>99%)
MISSING: 3 (<1%)

DISTINCT: 2 (<1%)

4 Dependents

VALUES: 599 (98%)
MISSING: 15 (2%)



DISTINCT:	4	(<1%)

5 **Education**

VALUES:	614	(100%)
MISSING:	---	
DISTINCT:	2	(<1%)

6 **Self_Employed**

VALUES:	582	(95%)
MISSING:	32	(5%)
DISTINCT:	2	(<1%)

7 **ApplicantIncome**

VALUES:	614	(100%)
MISSING:	---	
DISTINCT:	505	(82%)
ZEROES:	---	

MAX	81,000
95%	14,583
Q3	5,795
AVG	5,403
MEDIAN	3,812
Q1	2,878
5%	1,898
MIN	150

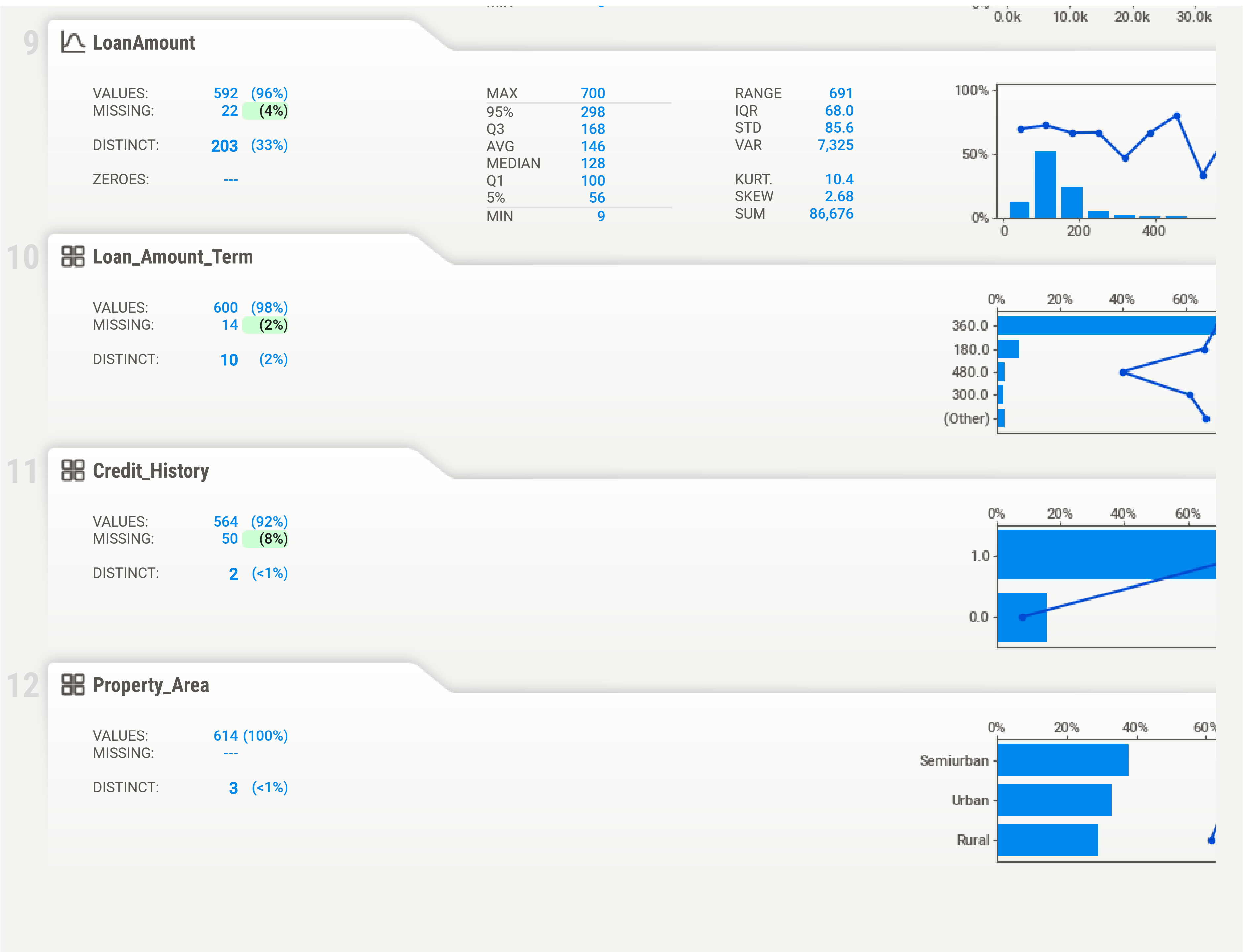
RANGE	80,850
IQR	2,918
STD	6,109
VAR	37.3M
KURT.	60.5
SKEW	6.54
SUM	3.3M

8 **CoapplicantIncome**

VALUES:	614	(100%)
MISSING:	---	
DISTINCT:	287	(47%)
ZEROES:	273	(44%)

MAX	41,667
95%	4,997
Q3	2,297
AVG	1,621
MEDIAN	1,188
Q1	0
5%	0
MIN	0

RANGE	41,667
IQR	2,297
STD	2,926
VAR	8.6M
KURT.	85.0
SKEW	7.49
SUM	995k



9 LoanAmount

VALUES:	592	(96%)
MISSING:	22	(4%)
DISTINCT:	203	(33%)
ZEROES:	---	

MAX	700
95%	298
Q3	168
AVG	146
MEDIAN	128
Q1	100
5%	56
MIN	9

RANGE	691
IQR	68.0
STD	85.6
VAR	7,325
KURT.	10.4
SKEW	2.68
SUM	86,676

10 Loan_Amount_Term

VALUES:	600	(98%)
MISSING:	14	(2%)
DISTINCT:	10	(2%)

11 Credit_History

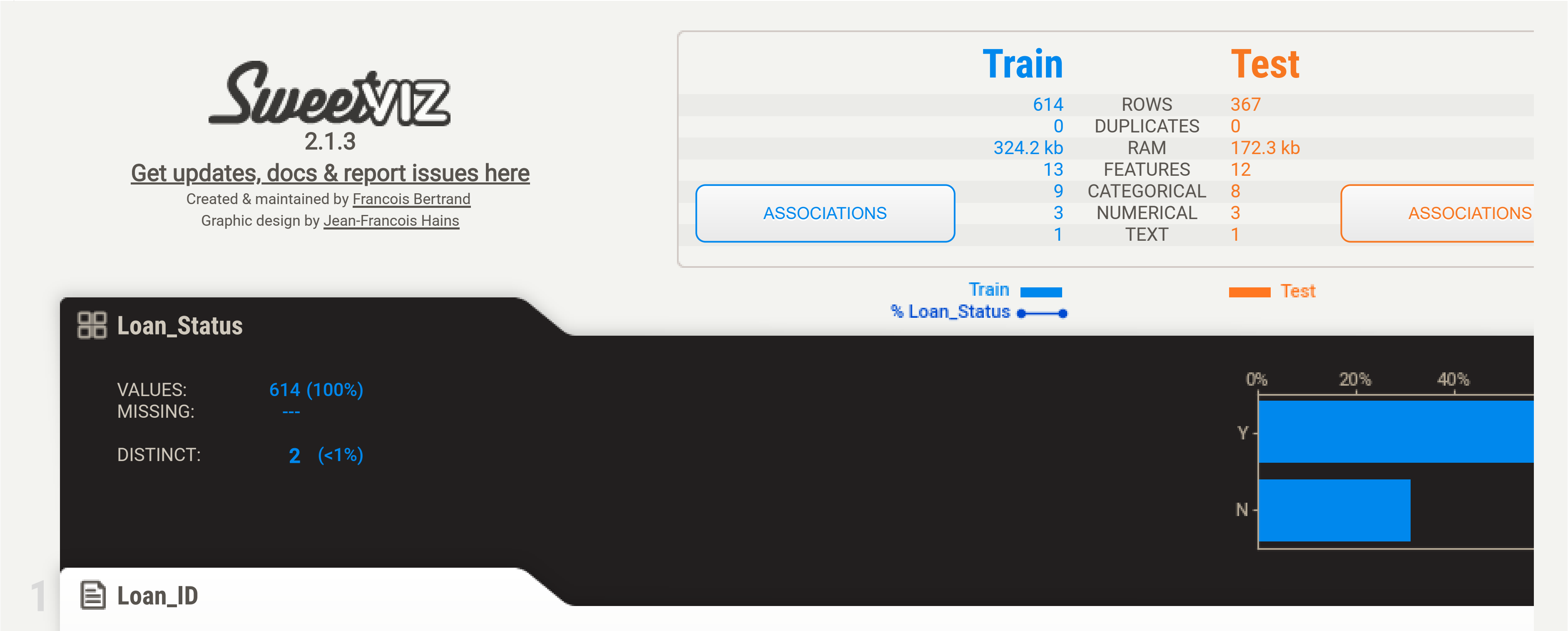
VALUES:	564	(92%)
MISSING:	50	(8%)
DISTINCT:	2	(<1%)

12 Property_Area

VALUES:	614	(100%)
MISSING:	---	
DISTINCT:	3	(<1%)

```
In [12]: analysis_report3 = sv.compare([df_train,'Train'],[df_test,'Test'],target_feat='Loan_Status')

         analysis_report3.show_notebook(w="100%",h="full")
```



SweetVIZ

2.1.3

Get updates, docs & report issues here

Created & maintained by Francois Bertrand

Graphic design by Jean-Francois Hains

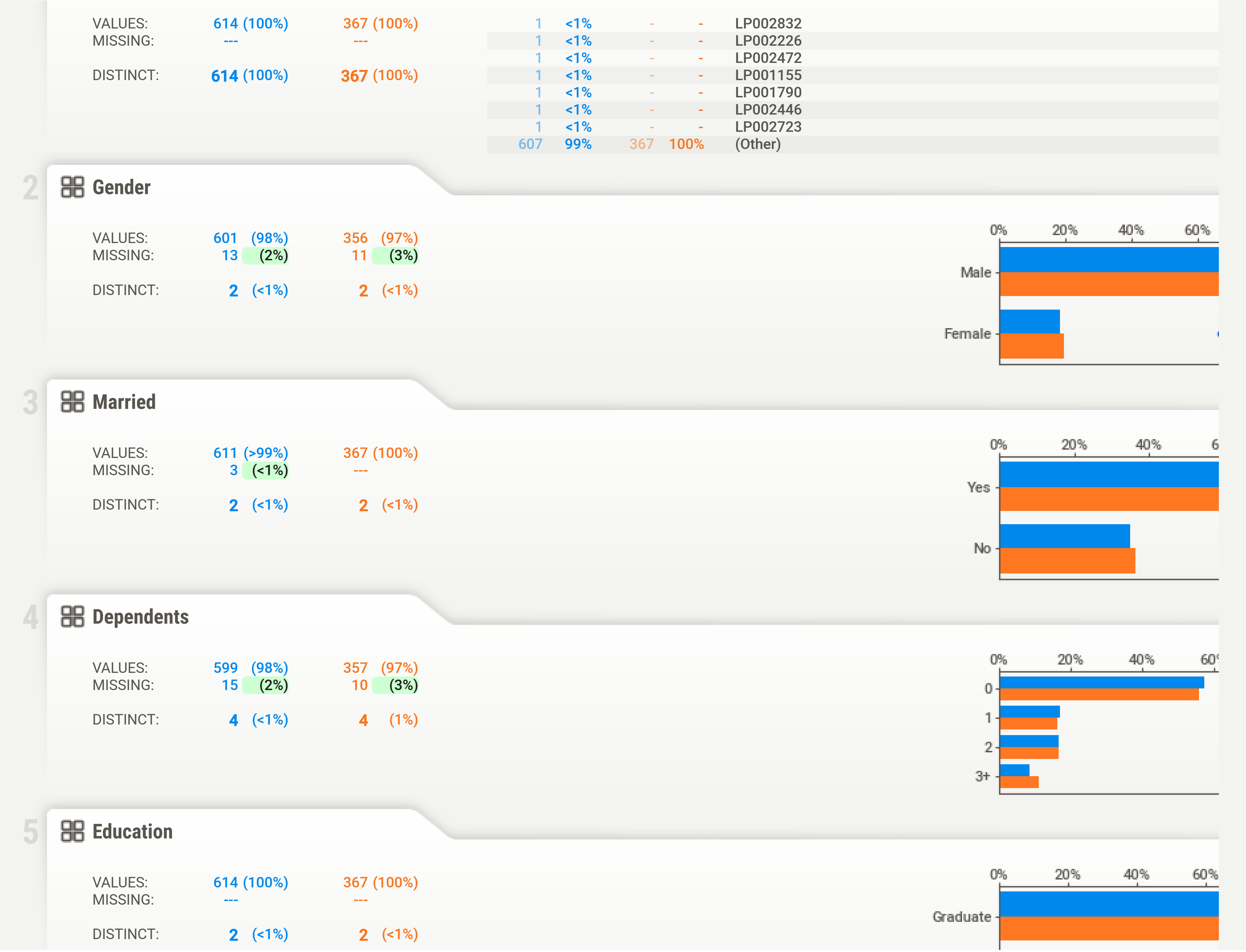
Train		Test
614	ROWS	367
0	DUPLICATES	0
324.2 kb	RAM	172.3 kb
13	FEATURES	12
9	CATEGORICAL	8
3	NUMERICAL	3
1	TEXT	1

Loan_Status

VALUES: 614 (100%)
MISSING: ---

DISTINCT: 2 (<1%)

1 Loan_ID



VALUES:	614 (100%)	367 (100%)
MISSING:	---	---
DISTINCT:	614 (100%)	367 (100%)

1	<1%	-	-	LP002832
1	<1%	-	-	LP002226
1	<1%	-	-	LP002472
1	<1%	-	-	LP001155
1	<1%	-	-	LP001790
1	<1%	-	-	LP002446
1	<1%	-	-	LP002723
607	99%	367	100%	(Other)

2 Gender

VALUES:	601 (98%)	356 (97%)
MISSING:	13 (2%)	11 (3%)
DISTINCT:	2 (<1%)	2 (<1%)

3 Married

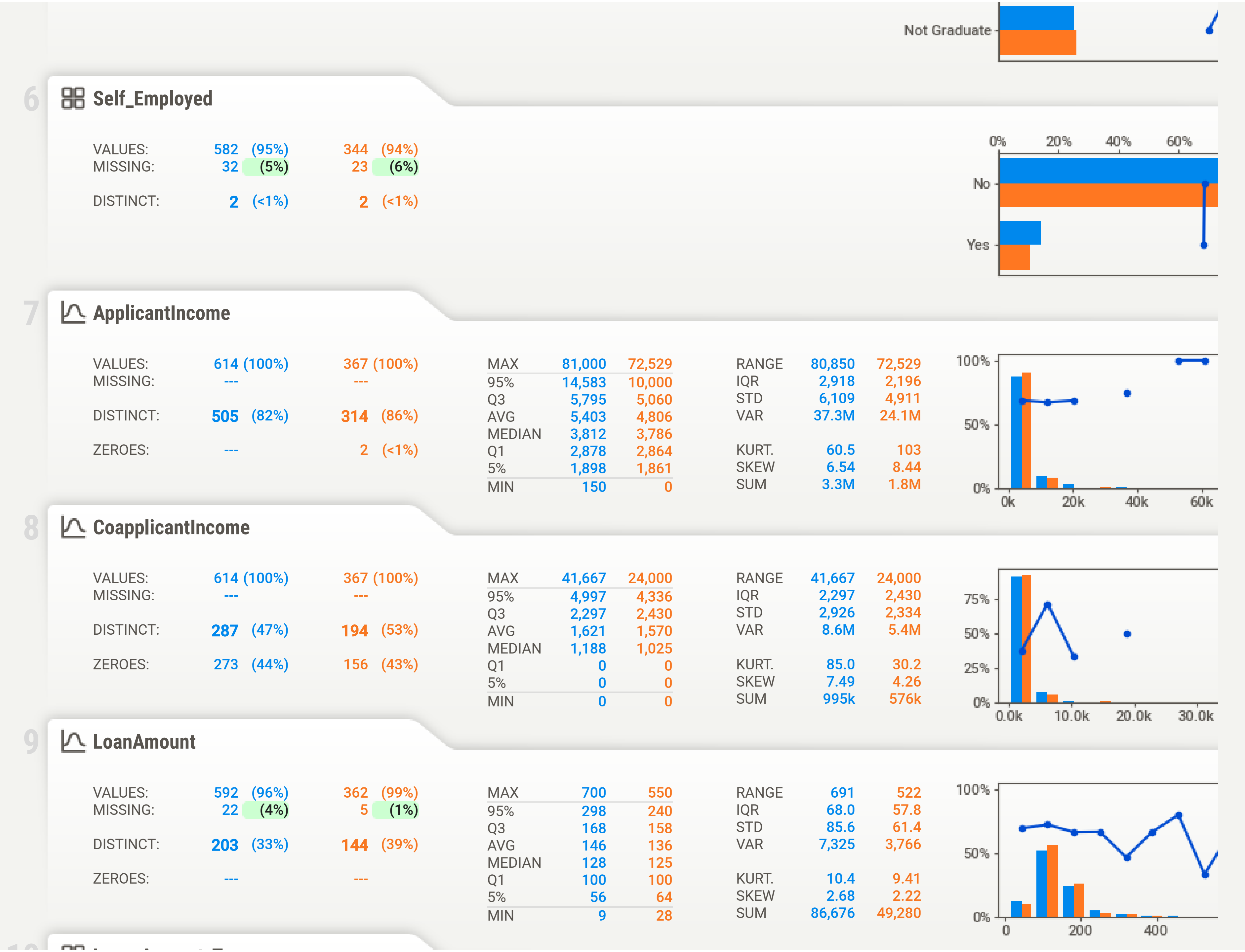
VALUES:	611 (>99%)	367 (100%)
MISSING:	3 (<1%)	---
DISTINCT:	2 (<1%)	2 (<1%)

4 Dependents

VALUES:	599 (98%)	357 (97%)
MISSING:	15 (2%)	10 (3%)
DISTINCT:	4 (<1%)	4 (1%)

5 Education

VALUES:	614 (100%)	367 (100%)
MISSING:	---	---
DISTINCT:	2 (<1%)	2 (<1%)



6 Self_Employed

VALUES:	582	(95%)	344	(94%)
MISSING:	32	(5%)	23	(6%)
DISTINCT:	2	(<1%)	2	(<1%)

7 ApplicantIncome

VALUES:	614	(100%)	367	(100%)
MISSING:	---		---	
DISTINCT:	505	(82%)	314	(86%)
ZEROES:	---		2	(<1%)

MAX	81,000	72,529
95%	14,583	10,000
Q3	5,795	5,060
AVG	5,403	4,806
MEDIAN	3,812	3,786
Q1	2,878	2,864
5%	1,898	1,861
MIN	150	0

RANGE	80,850	72,529
IQR	2,918	2,196
STD	6,109	4,911
VAR	37.3M	24.1M
KURT.	60.5	103
SKEW	6.54	8.44
SUM	3.3M	1.8M

8 CoapplicantIncome

VALUES:	614	(100%)	367	(100%)
MISSING:	---		---	
DISTINCT:	287	(47%)	194	(53%)
ZEROES:	273	(44%)	156	(43%)

MAX	41,667	24,000
95%	4,997	4,336
Q3	2,297	2,430
AVG	1,621	1,570
MEDIAN	1,188	1,025
Q1	0	0
5%	0	0
MIN	0	0

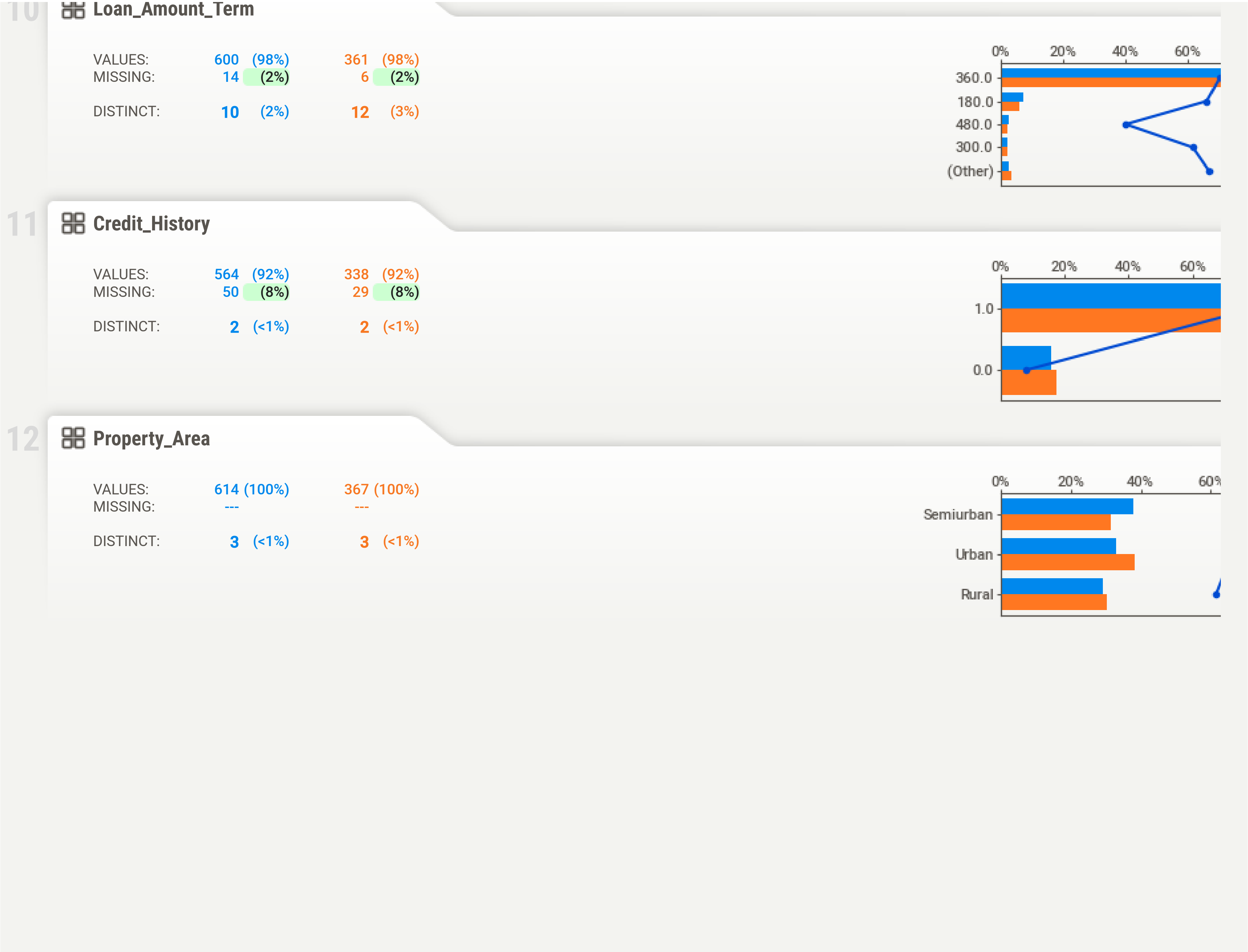
RANGE	41,667	24,000
IQR	2,297	2,430
STD	2,926	2,334
VAR	8.6M	5.4M
KURT.	85.0	30.2
SKEW	7.49	4.26
SUM	995k	576k

9 LoanAmount

VALUES:	592	(96%)	362	(99%)
MISSING:	22	(4%)	5	(1%)
DISTINCT:	203	(33%)	144	(39%)
ZEROES:	---		---	

MAX	700	550
95%	298	240
Q3	168	158
AVG	146	136
MEDIAN	128	125
Q1	100	100
5%	56	64
MIN	9	28

RANGE	691	522
IQR	68.0	57.8
STD	85.6	61.4
VAR	7,325	3,766
KURT.	10.4	9.41
SKEW	2.68	2.22
SUM	86,676	49,280



10 Loan_Amount_Term

VALUES:	600	(98%)	361	(98%)
MISSING:	14	(2%)	6	(2%)
DISTINCT:	10	(2%)	12	(3%)

11 Credit_History

VALUES:	564	(92%)	338	(92%)
MISSING:	50	(8%)	29	(8%)
DISTINCT:	2	(<1%)	2	(<1%)

12 Property_Area

VALUES:	614	(100%)	367	(100%)
MISSING:	---		---	
DISTINCT:	3	(<1%)	3	(<1%)

In [ ]: