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
import warnings

warnings.filterwarnings("ignore")
```

In [4]:

```
!pip install sweetviz # Please use it for the first time if it is not installed in your environment
```

In [5]:

```
import sweetviz as sv
```

In [7]:

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

```
df_train.head()
```

Out[7]:

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

In [8]:

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

df_test.head()
```

Out[8]:

	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 [9]:

```
df_train.shape
```

Out[9]:

```
(614, 13)
```

In [10]:

```
df_test.shape
```

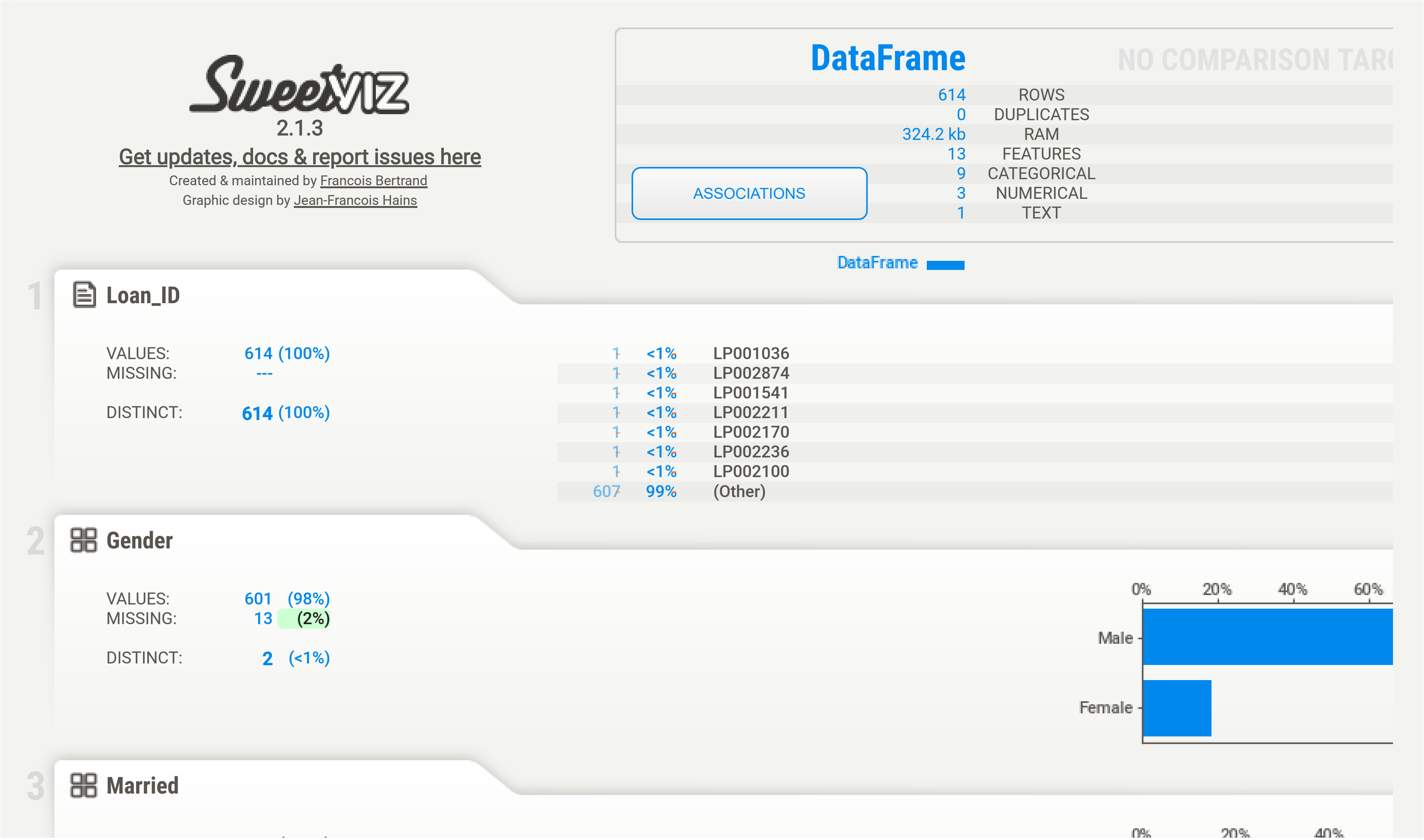
Out[10]:

```
(367, 12)
```

3.1 Analyze

```
In [11]: analysis_report = sv.analyze(df_train)
```

```
In [12]: # 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%)

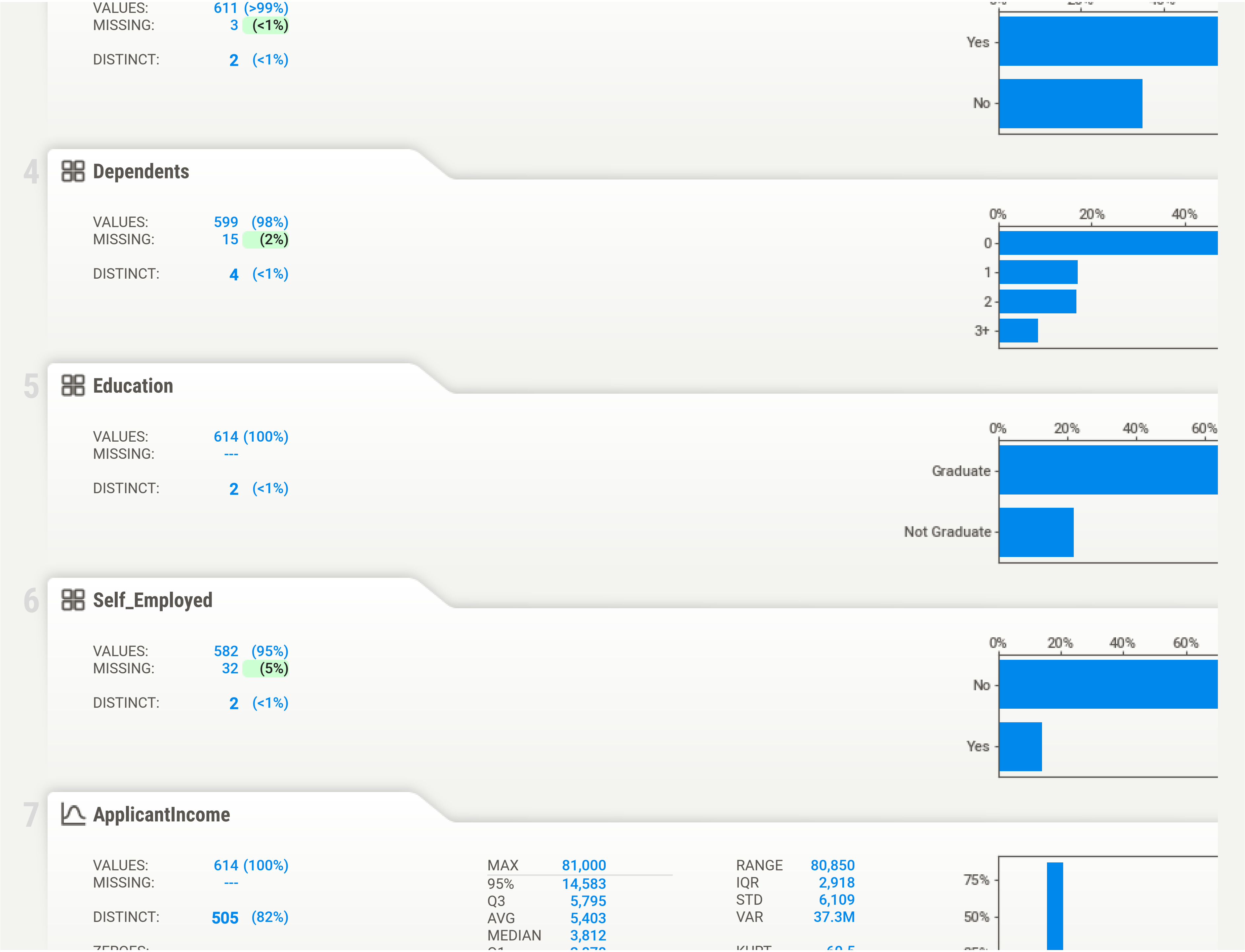
1	<1%	LP001036
1	<1%	LP002874
1	<1%	LP001541
1	<1%	LP002211
1	<1%	LP002170
1	<1%	LP002236
1	<1%	LP002100
607	99%	(Other)

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: ---

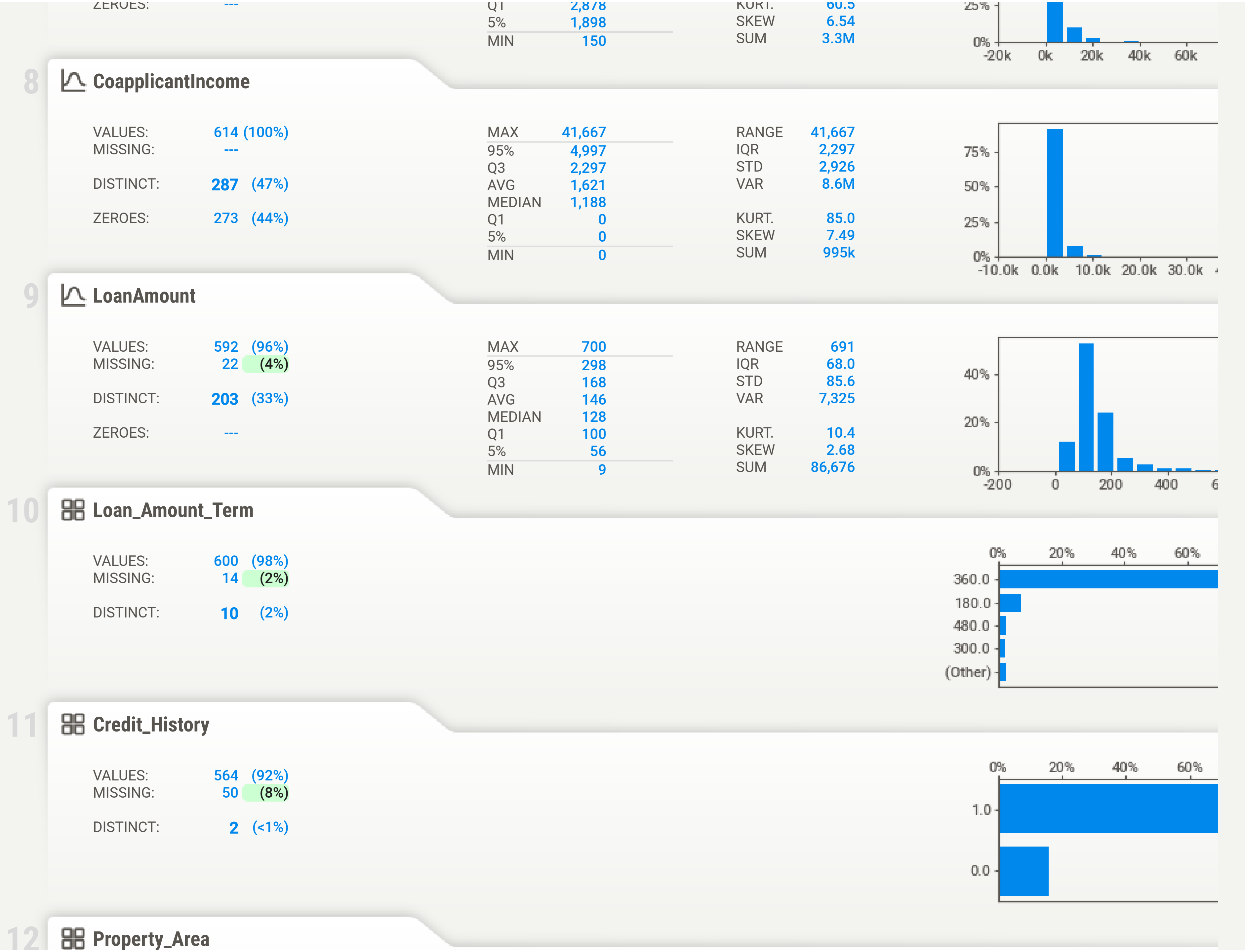
DISTINCT: 505 (82%)

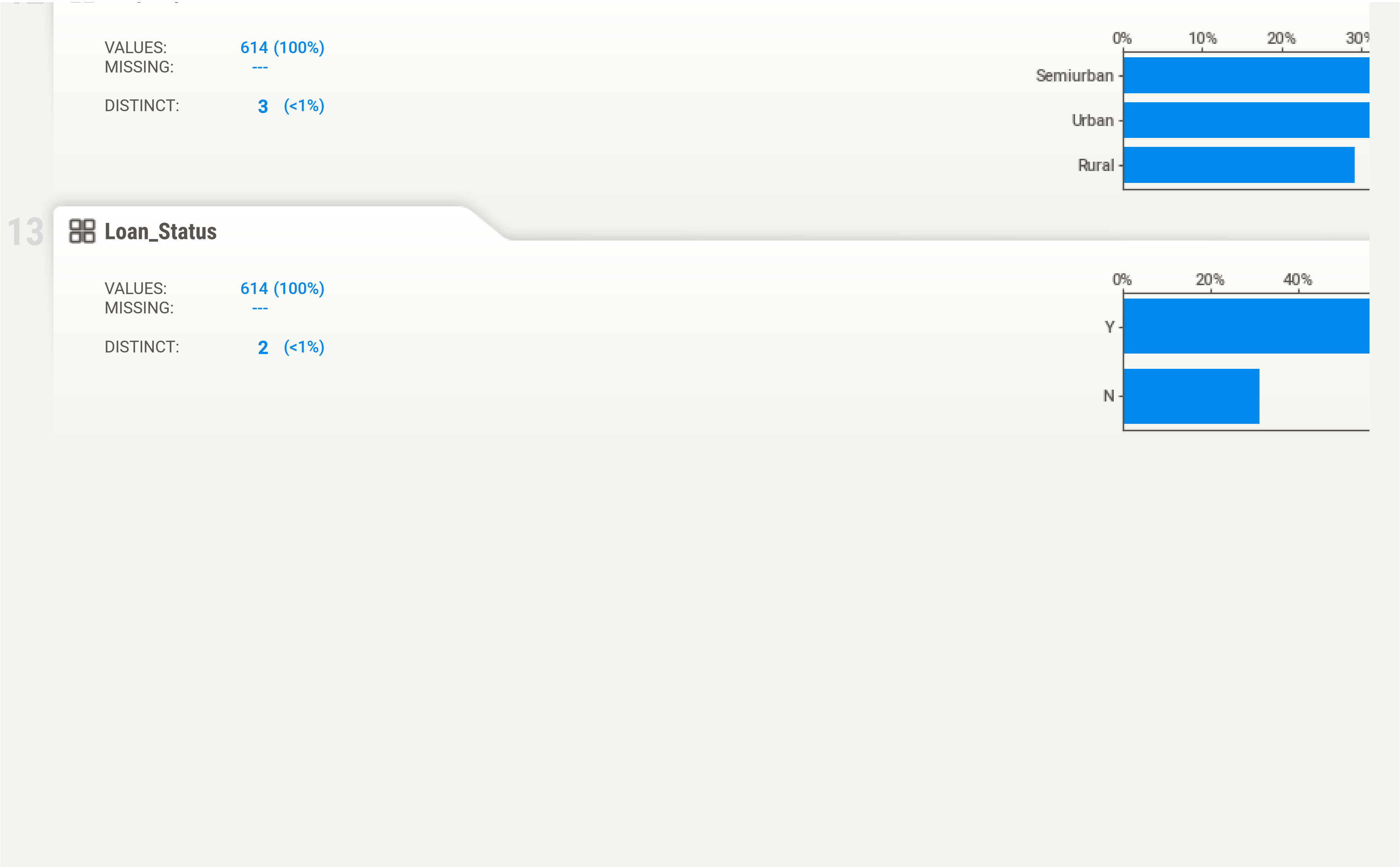
ZEROES:

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

RANGE 80,850
IQR 2,918
STD 6,109
VAR 37.3M

KURT 60.5





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

DISTINCT:    3   (<1%)
```

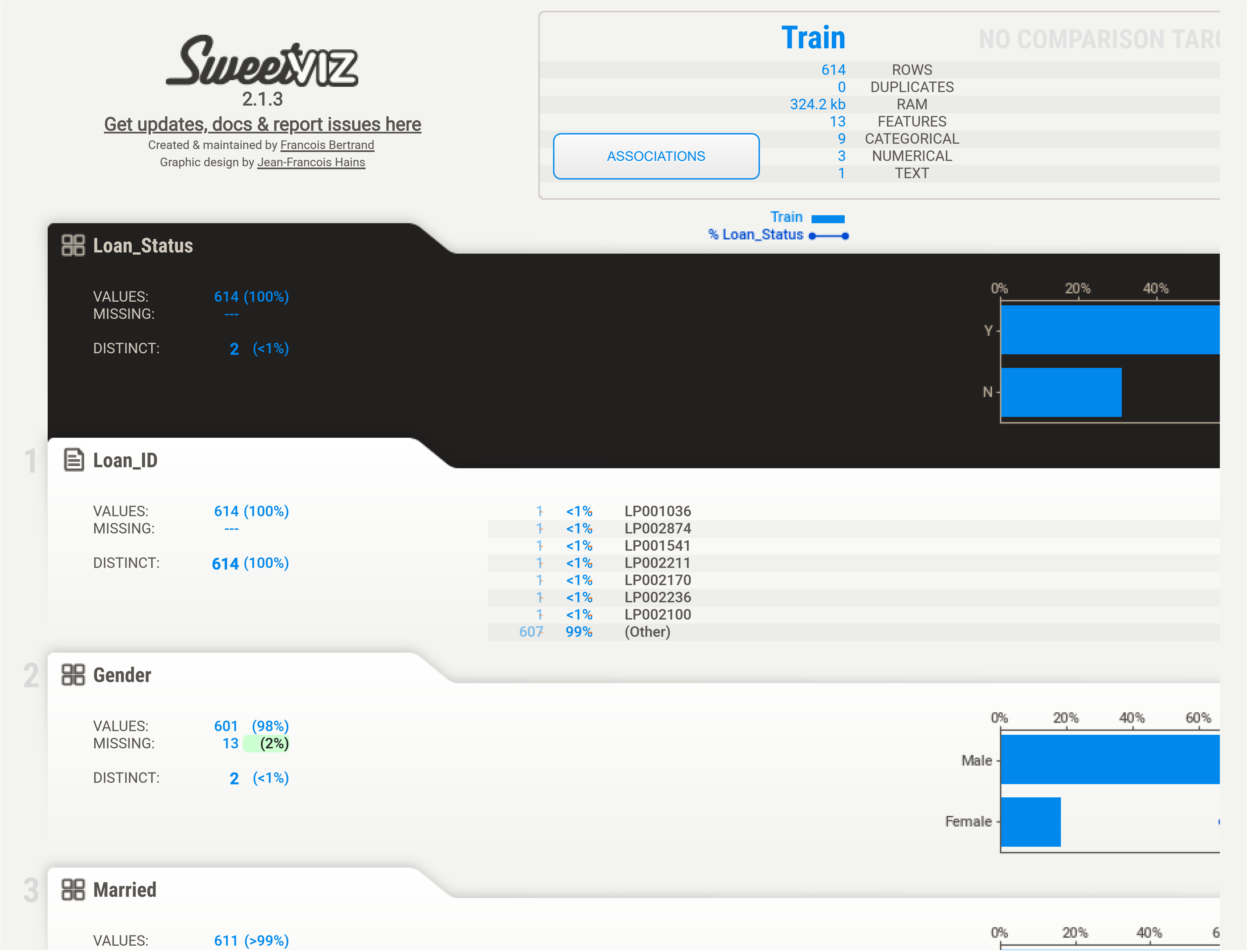
13 Loan_Status

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

DISTINCT:    2   (<1%)
```

```
In [13]: analysis_report2 = sv.analyze([df_train,'Train'], target_feat='Loan_Status')
```

```
In [14]: 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%)

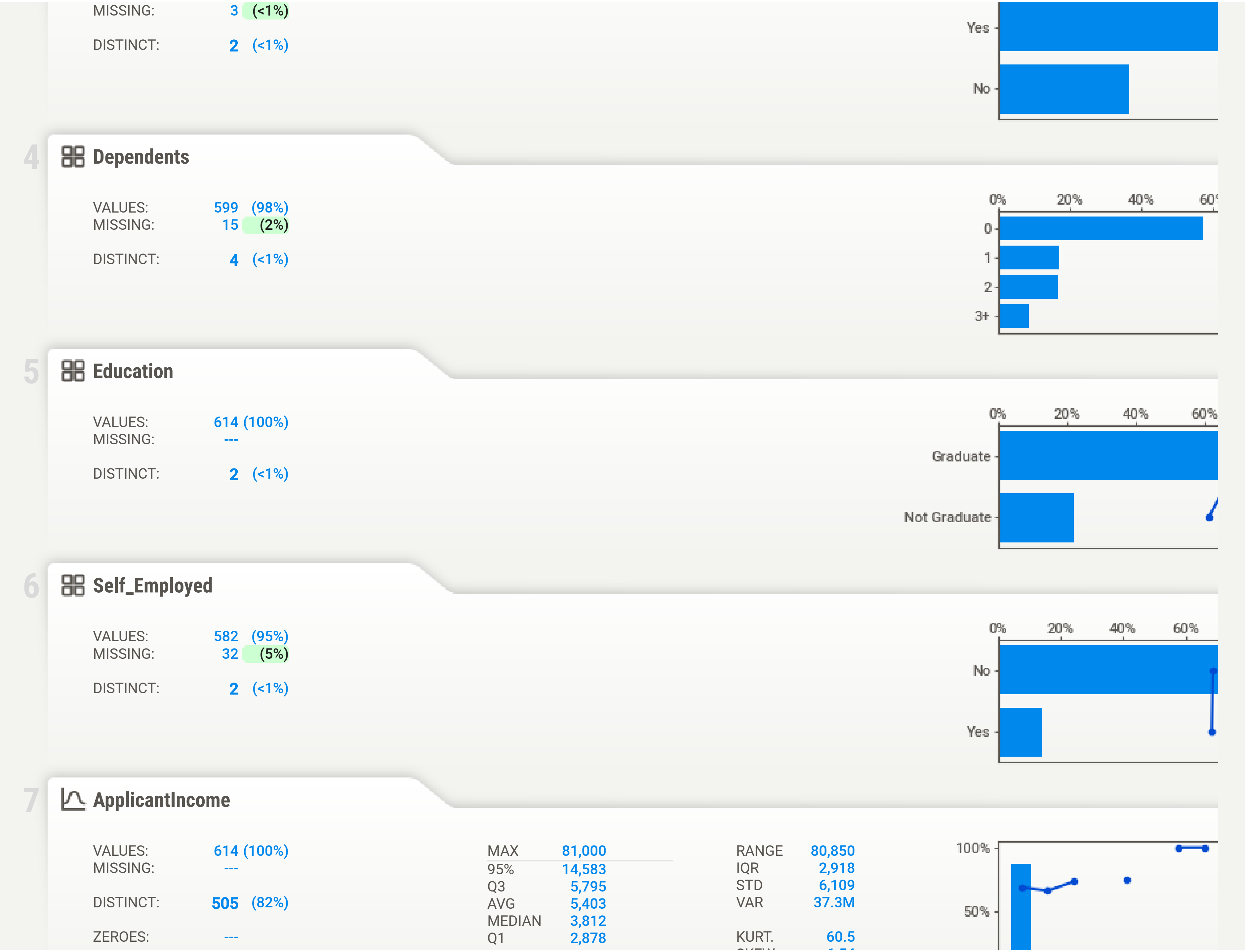
1	<1%	LP001036
1	<1%	LP002874
1	<1%	LP001541
1	<1%	LP002211
1	<1%	LP002170
1	<1%	LP002236
1	<1%	LP002100
607	99%	(Other)

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%)
```

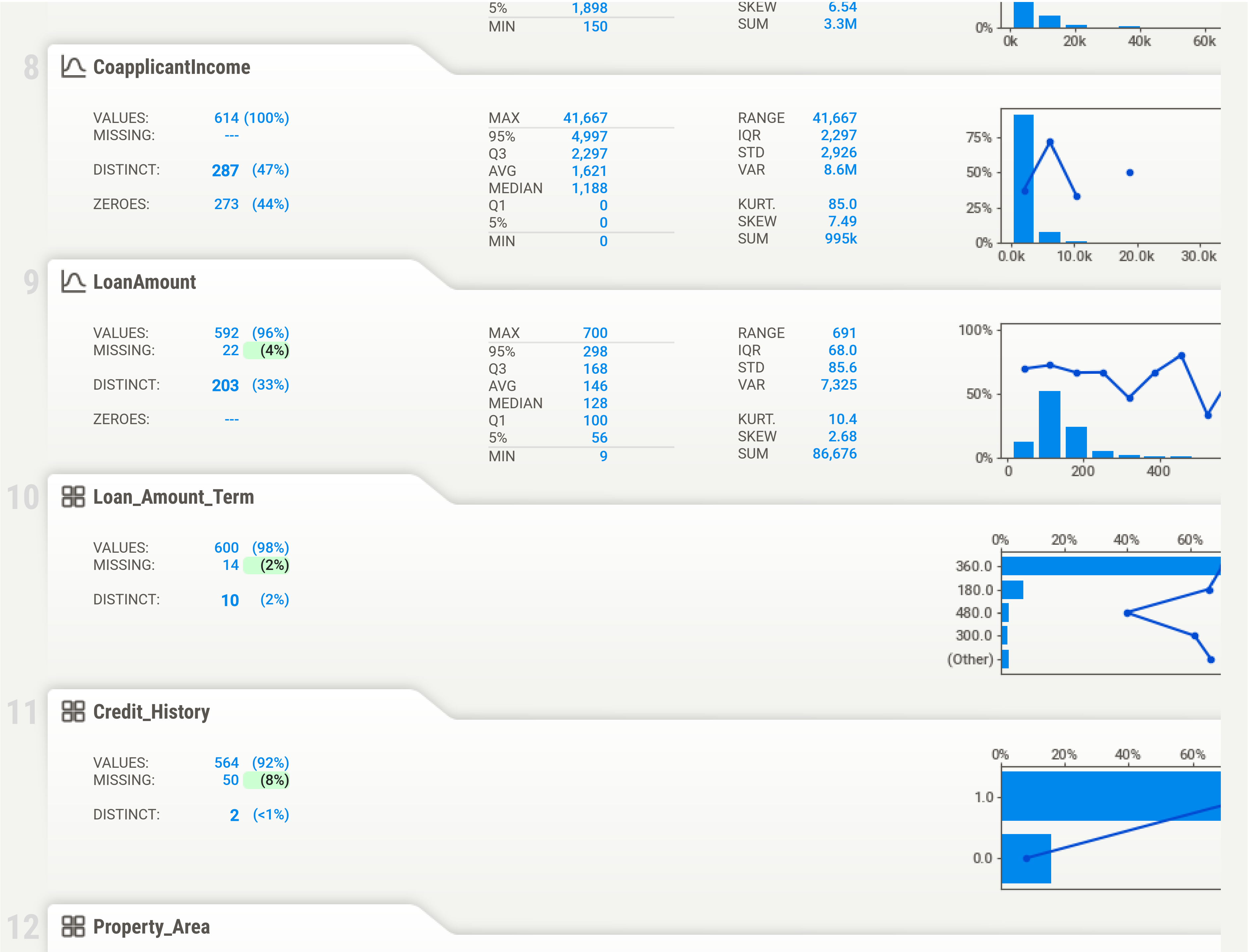
6 Self_Employed

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

DISTINCT:       2  (<1%)
```

7 ApplicantIncome

```
VALUES:       614 (100%)     MAX        81,000       RANGE    80,850
MISSING:       ---           95%        14,583       IQR       2,918
                             Q3          5,795       STD       6,109
DISTINCT:     505  (82%)     AVG         5,403       VAR       37.3M
                             MEDIAN      3,812
ZEROES:        ---           Q1          2,878       KURT.      60.5
```



5%	1,898	SKEW	6.54
MIN	150	SUM	3.3M

8 CoapplicantIncome

VALUES:	614 (100%)	MAX	41,667	RANGE	41,667
MISSING:	---	95%	4,997	IQR	2,297
		Q3	2,297	STD	2,926
DISTINCT:	287 (47%)	AVG	1,621	VAR	8.6M
		MEDIAN	1,188		
ZEROES:	273 (44%)	Q1	0	KURT.	85.0
		5%	0	SKEW	7.49
		MIN	0	SUM	995k

9 LoanAmount

VALUES:	592 (96%)	MAX	700	RANGE	691
MISSING:	22 (4%)	95%	298	IQR	68.0
		Q3	168	STD	85.6
DISTINCT:	203 (33%)	AVG	146	VAR	7,325
		MEDIAN	128		
ZEROES:	---	Q1	100	KURT.	10.4
		5%	56	SKEW	2.68
		MIN	9	SUM	86,676

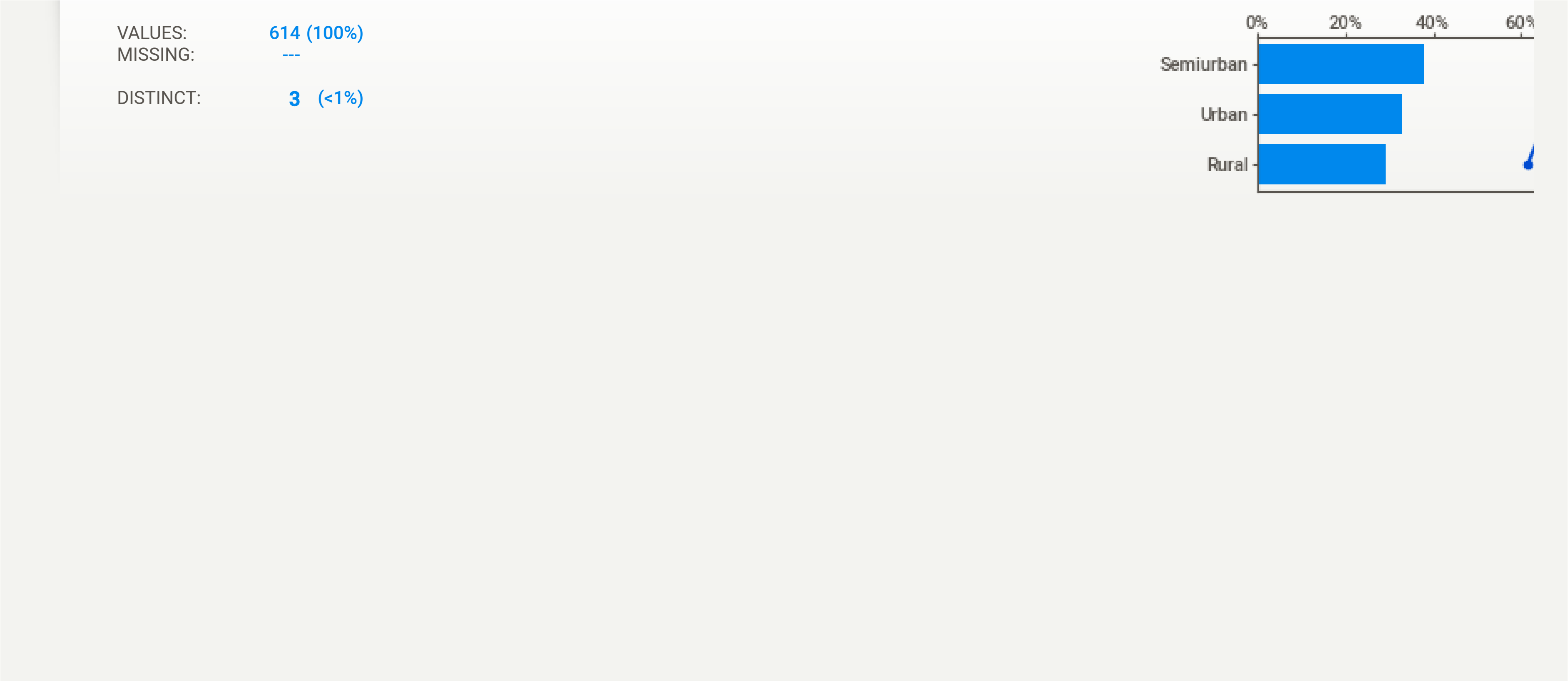
10 Loan_Amount_Term

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

11 Credit_History

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

12 Property_Area



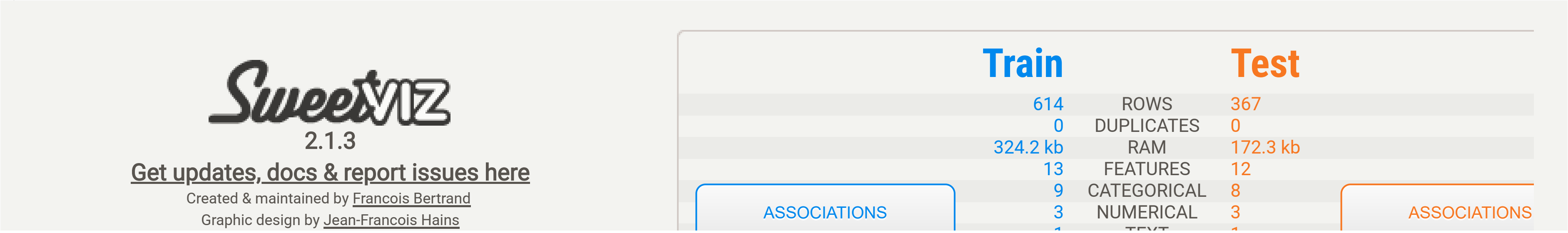
VALUES: 614 (100%)
MISSING: ---

DISTINCT: 3 (<1%)

3.2 Compare

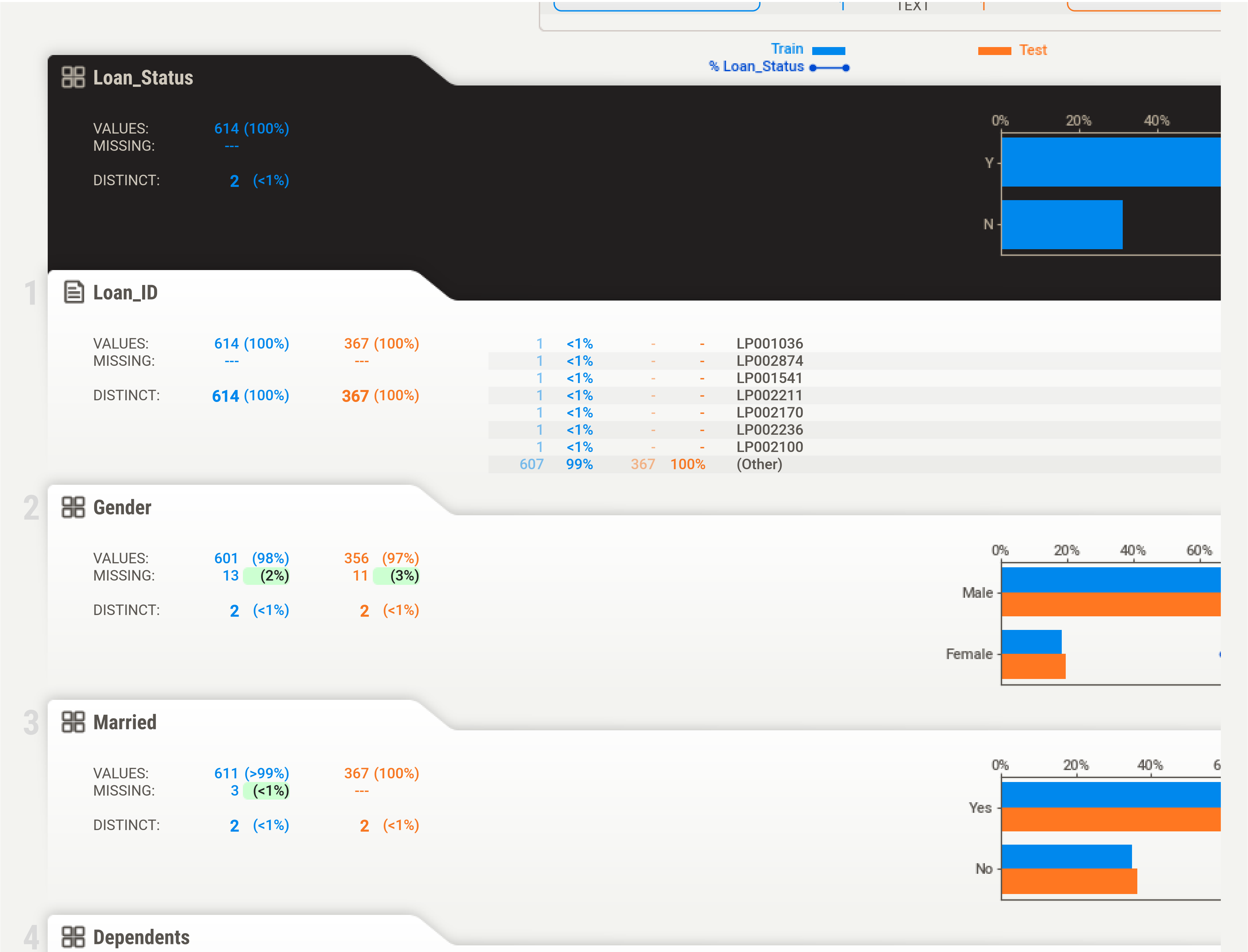
In [15]:
```
analysis_report3 = sv.compare([df_train,'Train'],[df_test,'Test'],target_feat='Loan_Status')

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%	-	-	LP001036
1	<1%	-	-	LP002874
1	<1%	-	-	LP001541
1	<1%	-	-	LP002211
1	<1%	-	-	LP002170
1	<1%	-	-	LP002236
1	<1%	-	-	LP002100
607	99%	367	100%	(Other)

2 Gender

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

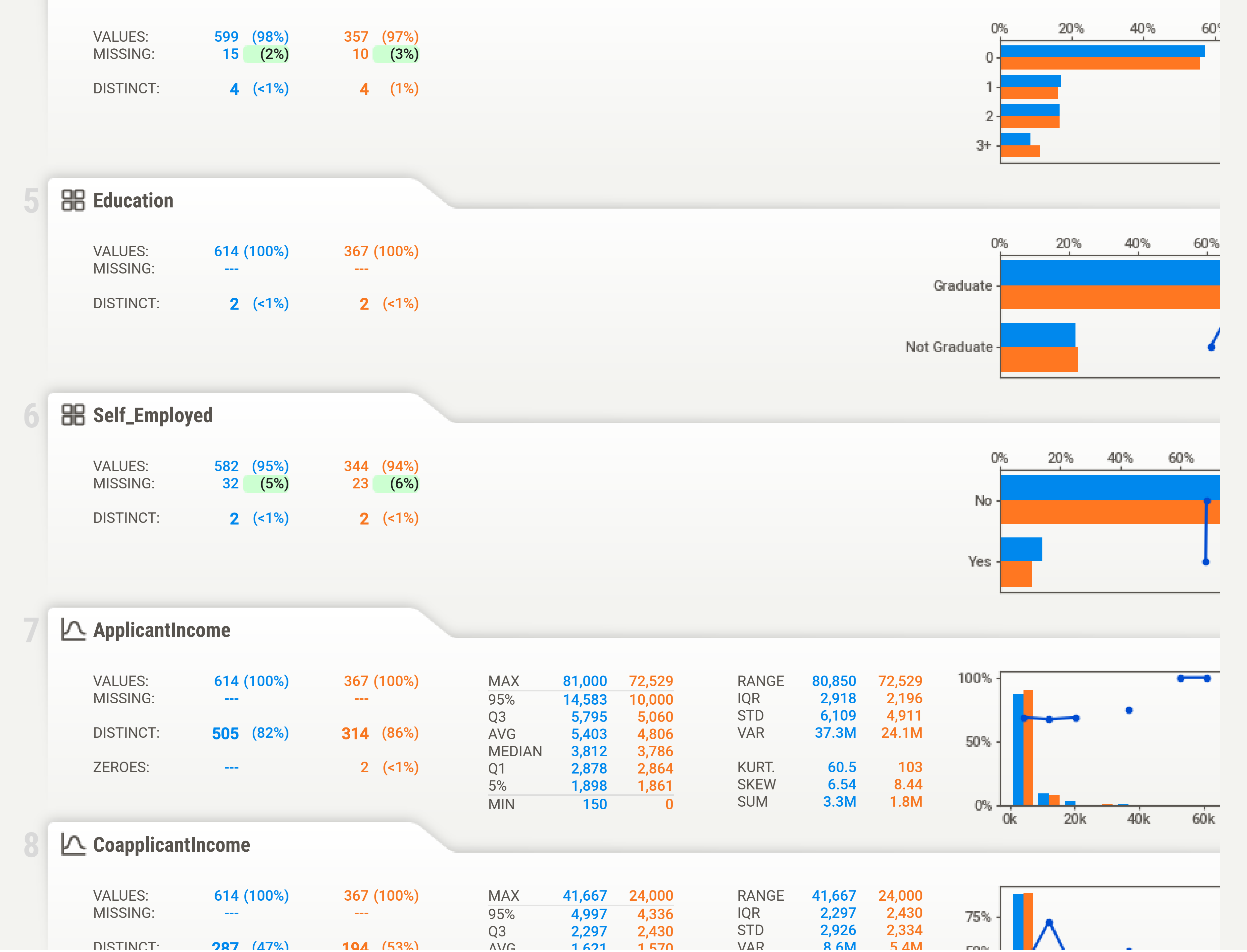
DISTINCT: 2 (<1%) 2 (<1%)

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

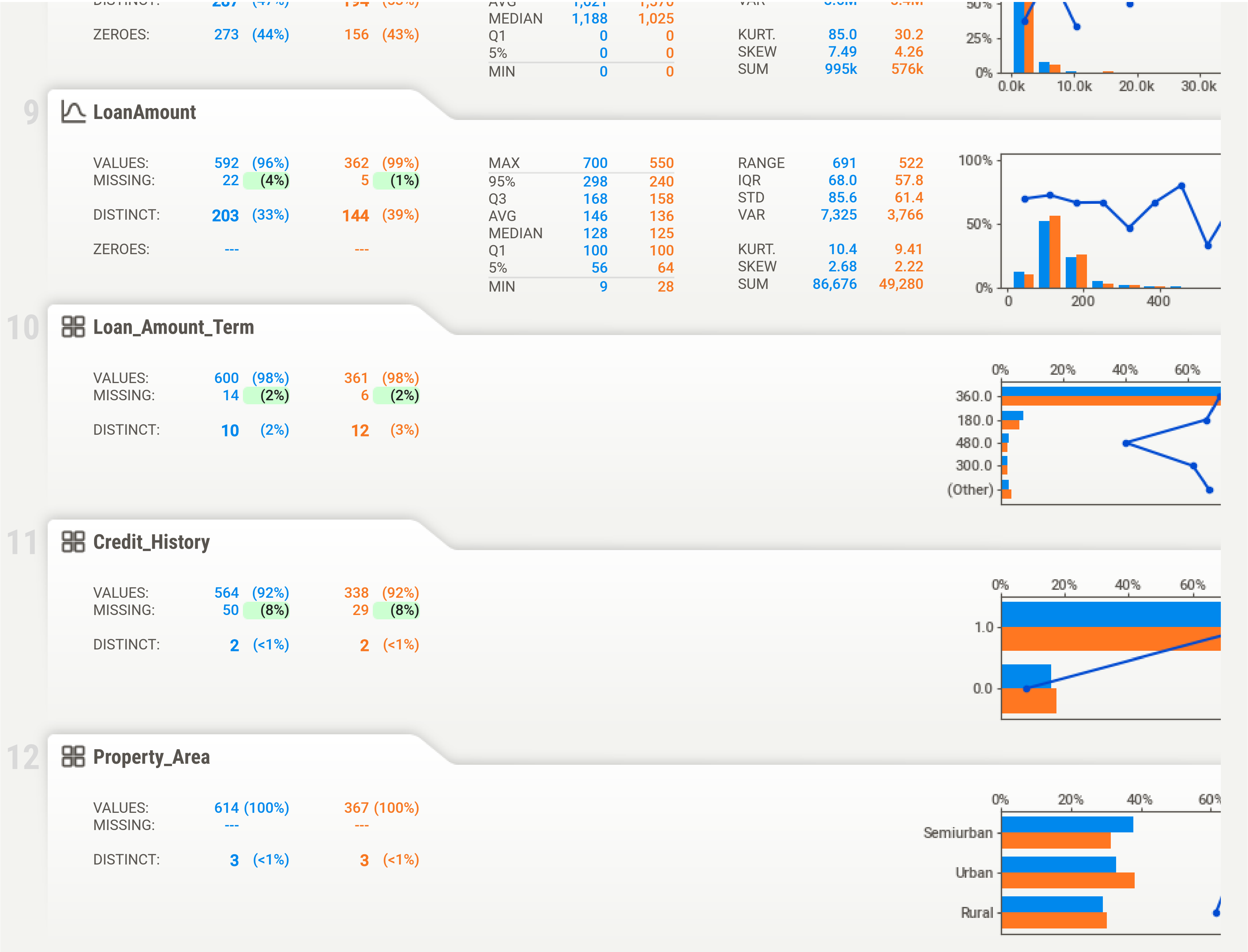
RANGE	80,850	72,529
IQR	2,918	2,196
STD	6,109	4,911
VAR	37.3M	24.1M
KURT.	60.5	103
SKEW	6.54	8.44
SUM	3.3M	1.8M

8 CoapplicantIncome

VALUES:	614	(100%)	367	(100%)
MISSING:	---		---	
DISTINCT:	287	(47%)	194	(53%)

MAX	41,667	24,000
95%	4,997	4,336
Q3	2,297	2,430
AVG	1,621	1,570

RANGE	41,667	24,000
IQR	2,297	2,430
STD	2,926	2,334
VAR	8.6M	5.4M



```
ZEROES:      273  (44%)      156  (43%)    MEDIAN   1,188   1,025
                                           Q1           0       0    KURT.    85.0    30.2
                                           5%           0       0    SKEW     7.49    4.26
                                           MIN          0       0    SUM      995k    576k
```

9 LoanAmount

```
VALUES:      592  (96%)      362  (99%)    MAX        700     550    RANGE      691      522
MISSING:      22   (4%)        5   (1%)    95%        298     240    IQR       68.0     57.8
                                           Q3         168     158    STD       85.6     61.4
DISTINCT:    203  (33%)      144  (39%)    AVG        146     136    VAR      7,325    3,766
                                           MEDIAN     128     125
ZEROES:      ---             ---           Q1         100     100    KURT.     10.4     9.41
                                           5%          56      64    SKEW      2.68     2.22
                                           MIN          9      28    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%)
```

3.3 Compare_Intra()

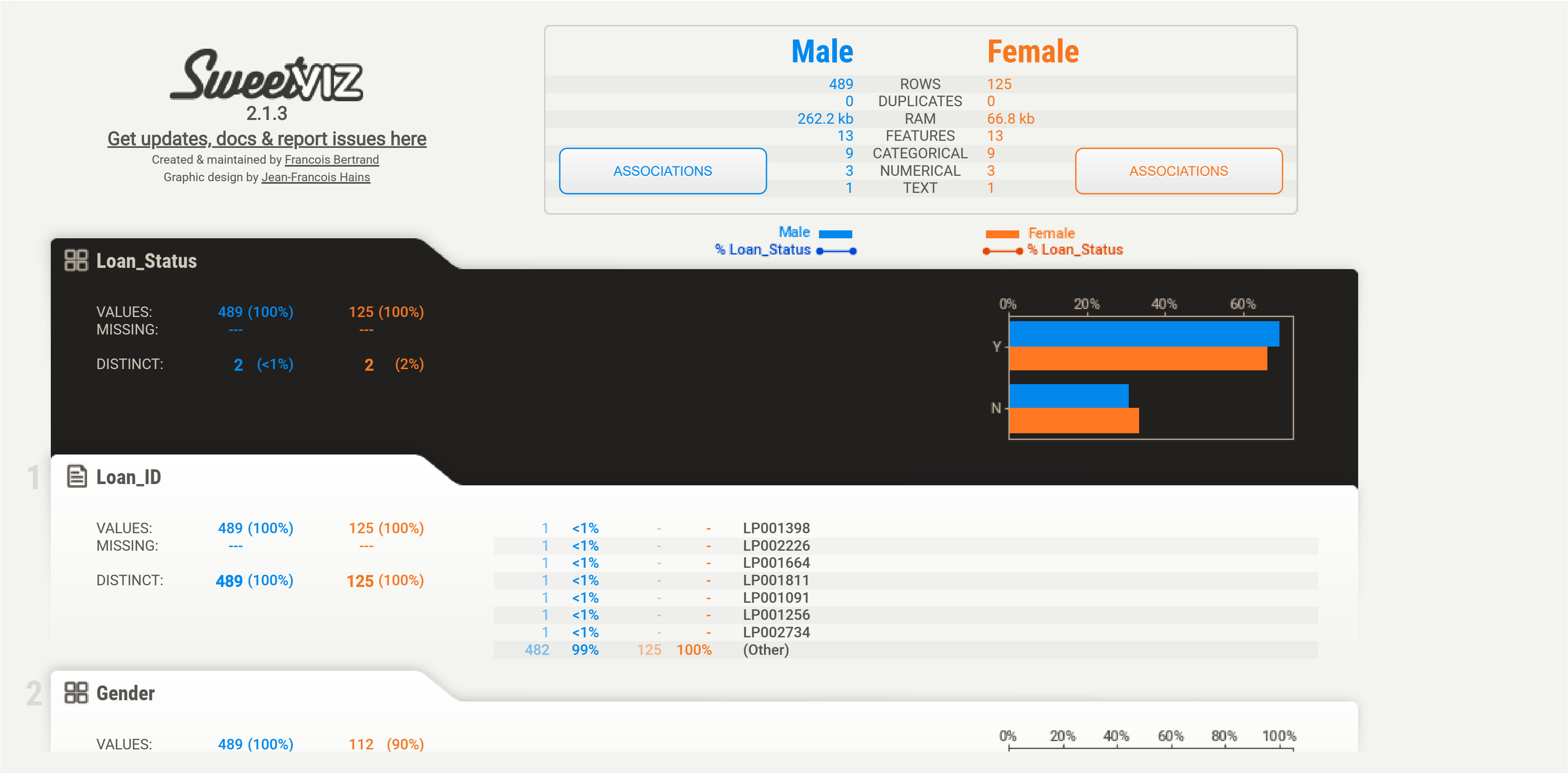
- Use this when you want to compare two populations within the same dataset.
- This is also a very useful report, especially when coupled with target feature analysis!

In [16]:

```
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
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 [17]:
```
intra_report = sv.compare_intra(df_train, df_train["Gender"] == 'Male', ["Male", "Female"], 'Loan_Status')
intra_report.show_notebook(w=900, h=450, scale=0.8)
```



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

Male		Female
489	ROWS	125
0	DUPLICATES	0
262.2 kb	RAM	66.8 kb
13	FEATURES	13
9	CATEGORICAL	9
3	NUMERICAL	3
1	TEXT	1

ASSOCIATIONS ASSOCIATIONS

Loan_Status

VALUES:	489 (100%)	125 (100%)
MISSING:	---	---
DISTINCT:	2 (<1%)	2 (2%)

1 Loan_ID

VALUES:	489 (100%)	125 (100%)
MISSING:	---	---
DISTINCT:	489 (100%)	125 (100%)

1	<1%	-	-	LP001398
1	<1%	-	-	LP002226
1	<1%	-	-	LP001664
1	<1%	-	-	LP001811
1	<1%	-	-	LP001091
1	<1%	-	-	LP001256
1	<1%	-	-	LP002734
482	99%	125	100%	(Other)

2 Gender

VALUES:	489 (100%)	112 (90%)

Interpretation Summary

- Summary Statistics
  - Data types, unique values, missing values, duplicates, most frequent values etc

  - Numerical analysis - min/max/range, quartiles, mean/mode, standard deviation, coefficient of variation, kurtosis, skewness
- Target analysis
  - Indicates how the target feature relates to other features
- Visualization and Comparision
  - distinct datasets between train and test
  - Intra-set characteristics
- Mixed-type associations
  - Integrates association for numerical (Pearson's Correlation)
  - Categorical (Uncertainty Coefficient) and categorical-numerical (Correlation ratio)
- Type inference
  - Automatically detects numerical, categorical and text features

```
In [ ]:
```