Sprocket Central Pvt Ltd

In [1]:

```python
# Basic Liabraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
```

In [2]:

```python
# importing dataprep library
import dataprep
from dataprep.datasets import load_dataset
from dataprep.eda import plot
from dataprep.eda import create_report
```

Customer Demographic

In [3]:

```python
customer_demographic = pd.read_excel("D:\\Forage\\KPMG\\KPMG_VI_New_raw_data_update_final.xlsx", sheet_name= 'CustomerDemographic')
```

In [4]:

```python
customer_demographic['DOB'] = pd.to_datetime(customer_demographic['DOB'])
create_report(customer_demographic).show()
```

DataPrep Report   Overview   Variables ≡   Interactions   Correlations   Missing Values

Overview

Dataset Statistics

Number of Variables	13
Number of Rows	4000
Missing Cells	1763
Missing Cells (%)	3.4%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	2.6 MB
Average Row Size in Memory	670.0 B
Variable Types	Numerical: 3 Categorical: 9 DateTime: 1

Dataset Insights

customer_id is uniformly distributed	Uniform
last_name has 125 (3.12%) missing values	Missing
job_title has 506 (12.65%) missing values	Missing
job_industry_category has 656 (16.4%) missing values	Missing
tenure has 87 (2.18%) missing values	Missing
default has 302 (7.55%) missing values	Missing
first_name has a high cardinality: 3139 distinct values	High Cardinality
last_name has a high cardinality: 3725 distinct values	High Cardinality
job_title has a high cardinality: 195 distinct values	High Cardinality
default has a high cardinality: 90 distinct values	High Cardinality

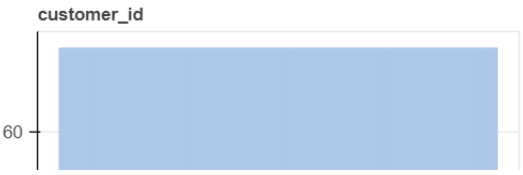
1 2

Variables

Sort by Feature order □ Reverse order

Approximate Distinct Count	4000
Approximate Unique (%)	100.0%

Mean	2000.5
Minimum	1
Maximum	4000



If you can see above clearly that there is year 1843 is mentioned which means the person is 179 years old

In [ ]:

```

```

In [ ]:

```

```

Transactions

In [5]:

```
transactions = pd.read_excel("D:\\Forage\\KPMG\\KPMG_VI_New_raw_data_update_final.xlsx", sheet_name= 'Transactions')
create_report(transactions).show()
```

DataPrep Report Overview Variables ≡ Interactions Correlations Missing Values

Overview

Dataset Statistics

Number of Variables	13
Number of Rows	20000
Missing Cells	1542
Missing Cells (%)	0.6%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	7.3 MB
Average Row Size in Memory	384.1 B
Variable Types	Numerical: 6 DateTime: 1 Categorical: 6

Dataset Insights

`transaction_id` is uniformly distributed	Uniform
`online_order` has 360 (1.8%) missing values	Missing
`product_id` is skewed	Skewed
`online_order` has constant length 3	Constant Length
`product_id` has 1378 (6.89%) zeros	Zeros

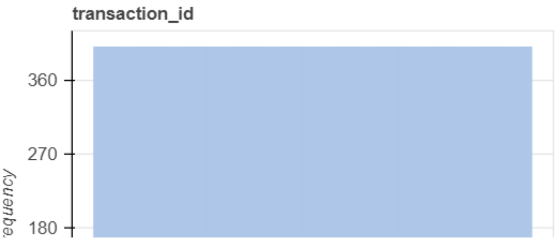
Variables

Sort by Feature order ☐ Reverse order

transaction_id
numerical

Approximate Distinct Count	20000
Approximate Unique (%)	100.0%
Missing	0

Mean	10000.5
Minimum	1
Maximum	20000
Zeros	0
Zeros (%)	0.0%



In [ ]:

```

```

In [ ]:

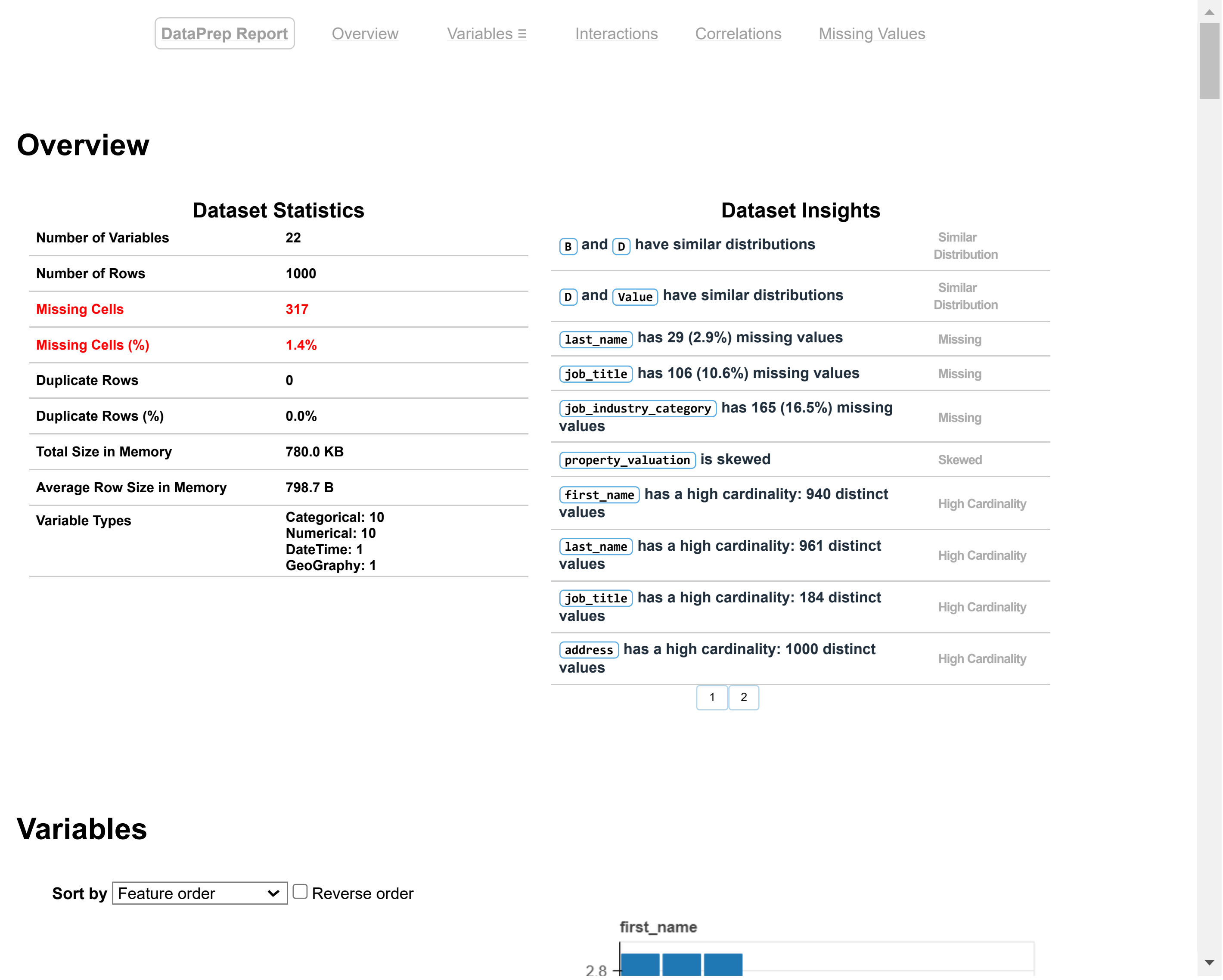
```

```

New Customer List

In [6]:

```
new_customers = pd.read_excel("D:\\Forage\\KPMG\\KPMG_VI_New_raw_data_update_final.xlsx", sheet_name= 'NewCustomerList')
create_report(new_customers).show()
```



Overview

Dataset Statistics

Number of Variables	22
Number of Rows	1000
Missing Cells	317
Missing Cells (%)	1.4%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	780.0 KB
Average Row Size in Memory	798.7 B
Variable Types	Categorical: 10 Numerical: 10 DateTime: 1 GeoGraphy: 1

Dataset Insights

B and D have similar distributions	Similar Distribution
D and Value have similar distributions	Similar Distribution
last_name has 29 (2.9%) missing values	Missing
job_title has 106 (10.6%) missing values	Missing
job_industry_category has 165 (16.5%) missing values	Missing
property_valuation is skewed	Skewed
first_name has a high cardinality: 940 distinct values	High Cardinality
last_name has a high cardinality: 961 distinct values	High Cardinality
job_title has a high cardinality: 184 distinct values	High Cardinality
address has a high cardinality: 1000 distinct values	High Cardinality

Variables

Sort by Feature order ☐ Reverse order

Column names = "A", "B","C","D" are given manually because it was giving NaN in column name

In [ ]:

```

```

In [ ]:

```

```

Customer Address

In [7]:

```
customer_address = pd.read_excel("D:\\Forage\\KPMG\\KPMG_VI_New_raw_data_update_final.xlsx", sheet_name= 'CustomerAddress')
create_report(customer_address).show()
```

DataPrep Report Overview Variables ≡ Interactions Correlations Missing Values

Overview

Dataset Statistics

Number of Variables	**6**
Number of Rows	**3999**
Missing Cells	**0**
Missing Cells (%)	**0.0%**
Duplicate Rows	**0**
Duplicate Rows (%)	**0.0%**
Total Size in Memory	**879.0 KB**
Average Row Size in Memory	**225.1 B**
Variable Types	**Numerical: 3** **Categorical: 2** **GeoGraphy: 1**

Dataset Insights

`customer_id` **is uniformly distributed**	Uniform
`customer_id` **is skewed**	Skewed
`property_valuation` **is skewed**	Skewed
`address` **has a high cardinality: 3996 distinct values**	High Cardinality
`country` **has constant value "Australia"**	Constant
`country` **has constant length 9**	Constant Length

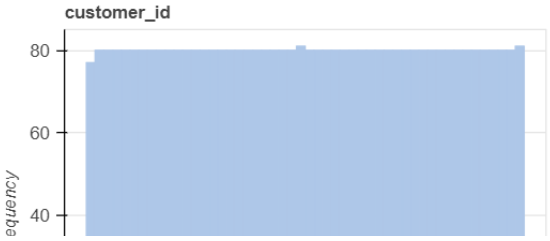
Variables

Sort by Feature order ⌄ □ Reverse order

customer_id
numerical

Approximate Distinct Count	**3999**
Approximate Unique (%)	**100.0%**
Missing	**0**

Mean	**2003.988**
Minimum	**1**
Maximum	**4003**
Zeros	**0**



In [ ]:

```

```