Week 8

Healthcare - Persistency of a drug: Group Project

Name: Judith Chepngetich

Email: chepngetich.judith@gmail.com

Group Name: N/A-(Self)

Country: Kenya

Company: Kenya Revenue Authority

Specialization: Data Science

Problem Description

Data Understanding

The data used by ABC has been obtained from the following sources:

- Patient
- Clinical Records
- Medical provider records

Pandas profiling tool was used to understand the data and the following observations were made:

- Total number of records in the dataset-**3424**
- Number of variables-**69**

Variables are grouped by demographics, provider attributes, clinical factors and disease/treatment factor.

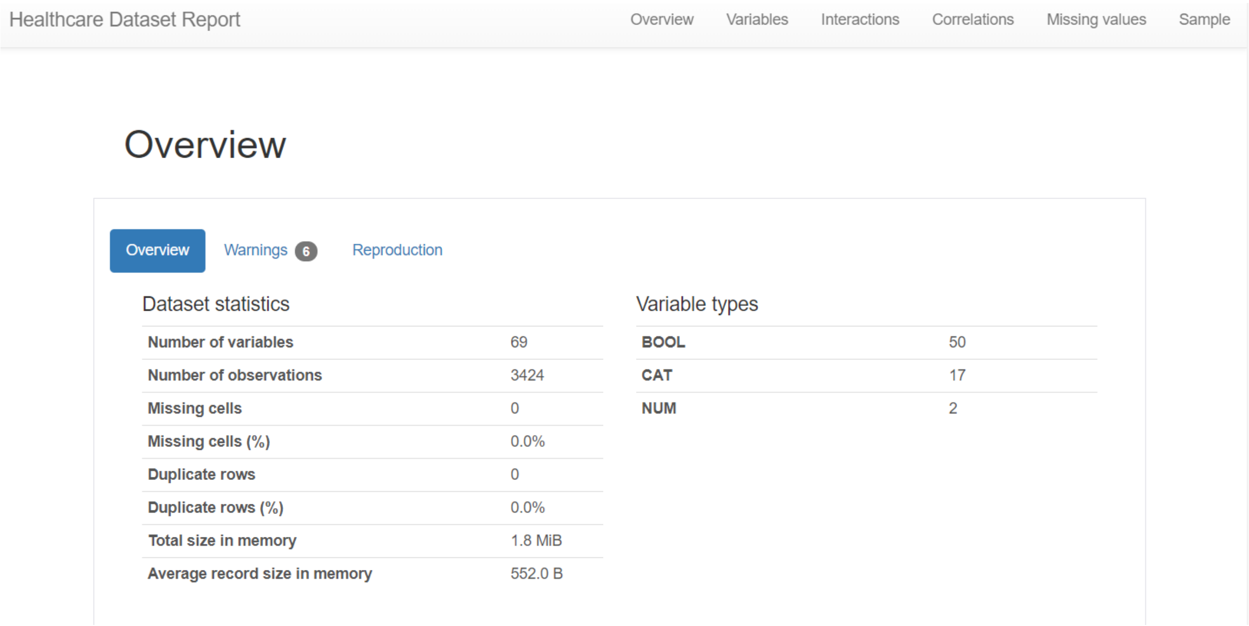
Variable types

Boolean**-50**

Categorical**-17**

Numerical**-2**

The below diagram summarizes the findings:



Overview

Dataset statistics

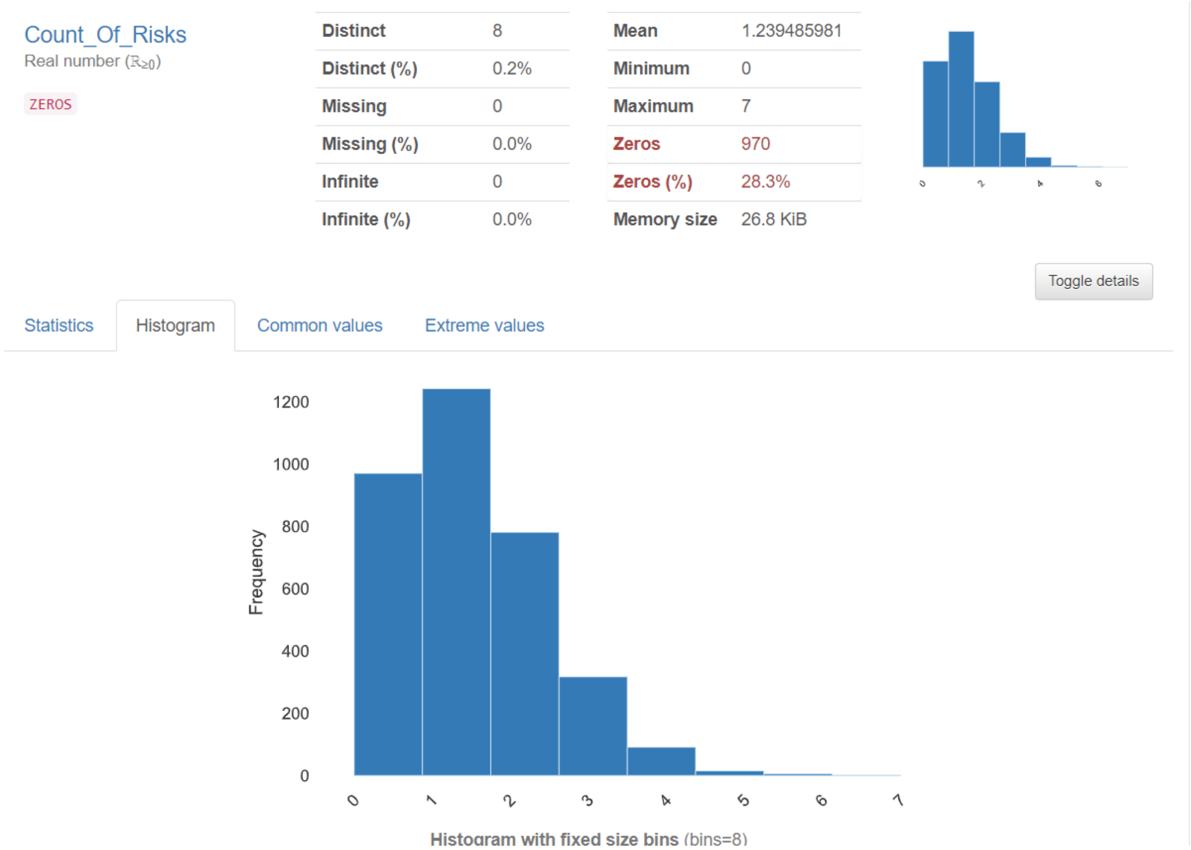
Number of variables	69
Number of observations	3424
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.8 MiB
Average record size in memory	552.0 B

Variable types

BOOL	50
CAT	17
NUM	2

Problems with the data

There are no missing values or duplicates in the dataset. For the numerical variables, the values in *Count _of_Risk* variable are moderately skewed with a skewness of 0.87. The variable also has no outliers, as shown below:

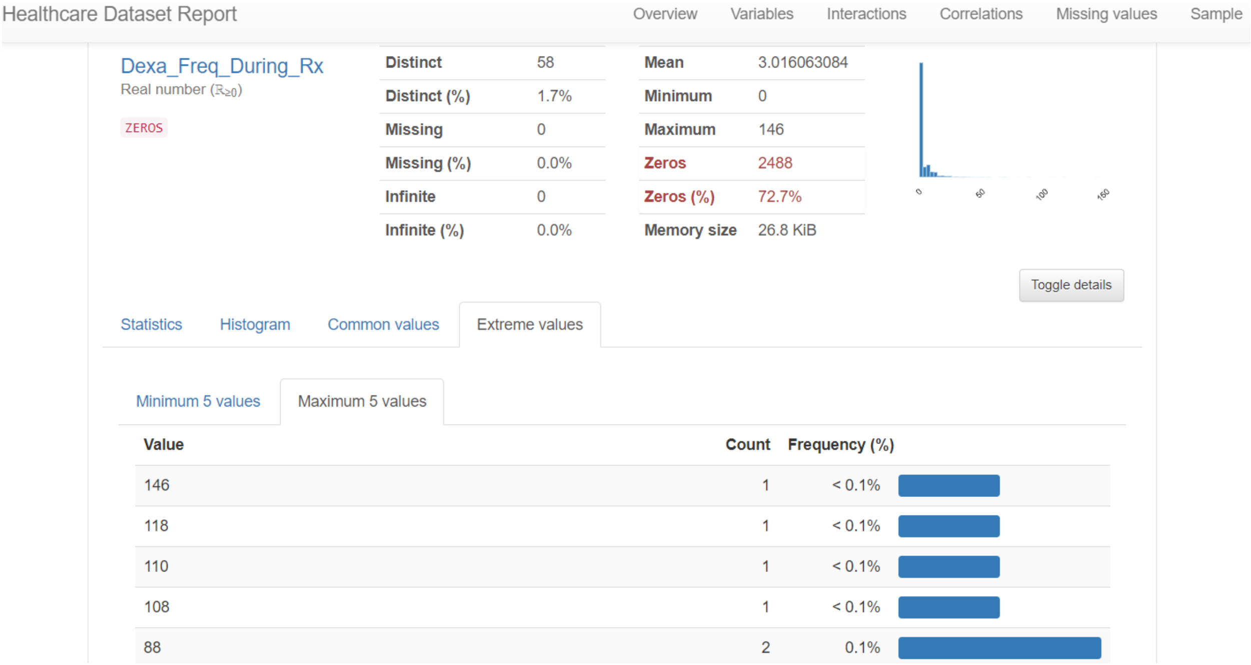

Count_Of_Risks
Real number (ℝ≥0)

Distinct	8
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	1.239485981
Minimum	0
Maximum	7
Zeros	970
Zeros (%)	28.3%
Memory size	26.8 KiB

Histogram with fixed size bins (bins=8)

The second numerical variable *Dexa_Freq_During_Rx* is highly skewed with a skewness of 6.8. This means that the data is asymmetrical and needs to be transformed before training the model.



Dexa_Freq_During_Rx

Real number ($\mathbb{R}_{\geq 0}$)

ZEROS

Distinct	58
Distinct (%)	1.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	3.016063084
Minimum	0
Maximum	146
Zeros	2488
Zeros (%)	72.7%
Memory size	26.8 KiB

Value	Count	Frequency (%)
146	1	< 0.1%
118	1	< 0.1%
110	1	< 0.1%
108	1	< 0.1%
88	2	0.1%