TEAM MEMBERS DETAILS

GROUP NAME: PACHACUTEAM

SPECIALIZATION: Data Science

Name	Email	Country	College/Company
Andersson Romero	andromdez@gmail.com	Perú	Freelancer

Data Understanding

As a first step, we load the dataset using pandas.

We look at the data.

```
df = pd.read_excel('Healthcare_dataset.xlsx')
df.head()
```

	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket
0	P1	Persistent	Male	Caucasian	Not Hispanic	West	>75
1	P2	Non-Persistent	Male	Asian	Not Hispanic	West	55-65
2	P3	Non-Persistent	Female	Other/Unknown	Hispanic	Midwest	65-75
3	P4	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75
4	P5	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75

We can quickly calculate that there are no null elements.

```
df.isnull().sum()[df.isnull().sum() != 0]
```

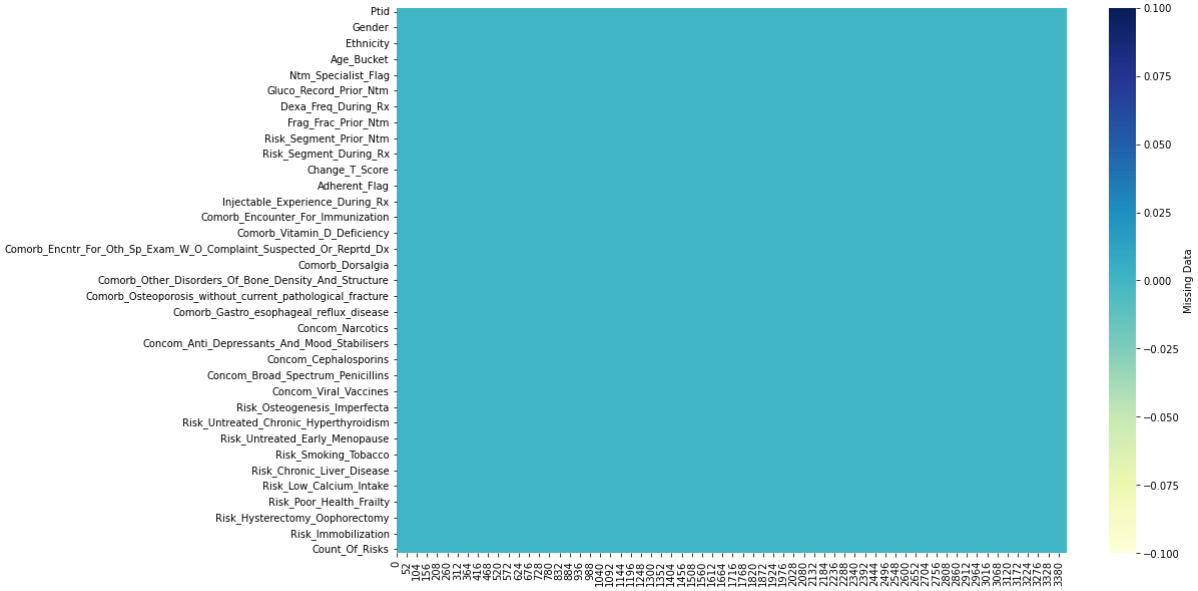
```
Series([], dtype: int64)
```

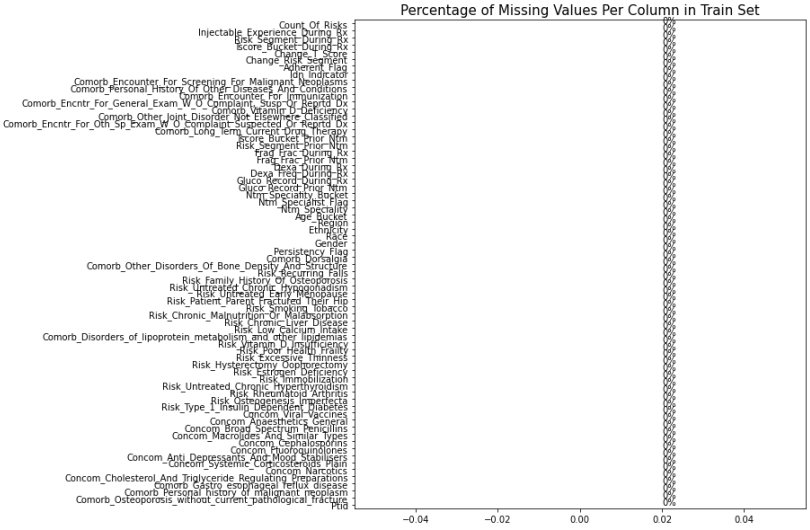
For more detail we use the info() function, where we see that almost all variables are objects.

```
df.info()
```

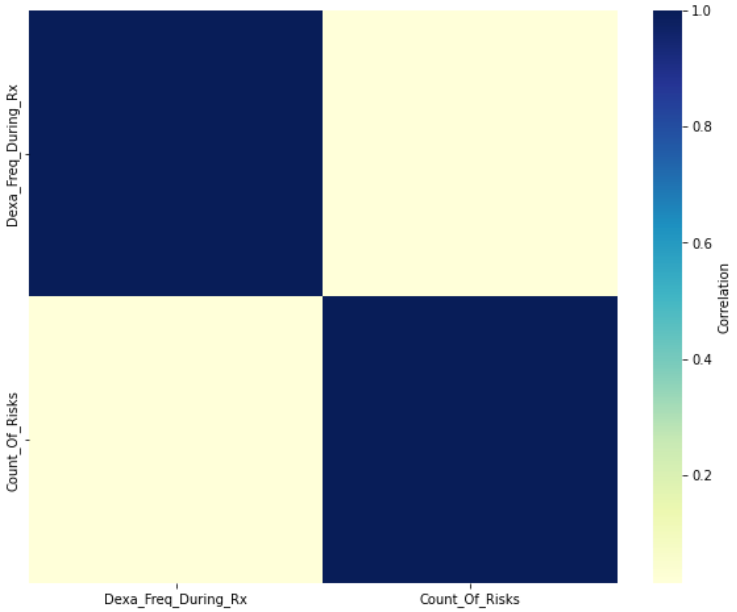
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3424 entries, 0 to 3423
Data columns (total 69 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   Ptid                                                    3424 non-null   object
 1   Persistency_Flag                                        3424 non-null   object
 2   Gender                                                  3424 non-null   object
 3   Race                                                    3424 non-null   object
 4   Ethnicity                                               3424 non-null   object
 5   Region                                                  3424 non-null   object
 6   Age_Bucket                                              3424 non-null   object
 7   Ntm_Speciality                                          3424 non-null   object
 8   Ntm_Specialist_Flag                                     3424 non-null   object
 9   Ntm_Speciality_Bucket                                   3424 non-null   object
 10  Gluco_Record_Prior_Ntm                                  3424 non-null   object
 11  Gluco_Record_During_Rx                                  3424 non-null   object
 12  Dexa_Freq_During_Rx                                     3424 non-null   int64
 13  Dexa_During_Rx                                          3424 non-null   object
 14  Frag_Frac_Prior_Ntm                                     3424 non-null   object
 15  Frag_Frac_During_Rx                                     3424 non-null   object
 16  Risk_Segment_Prior_Ntm                                  3424 non-null   object
```

We observe graphically that there are no null or empty elements.

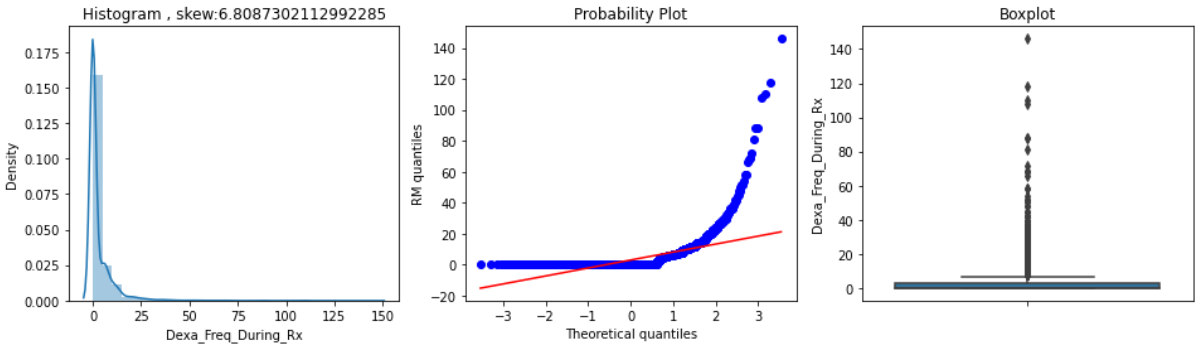




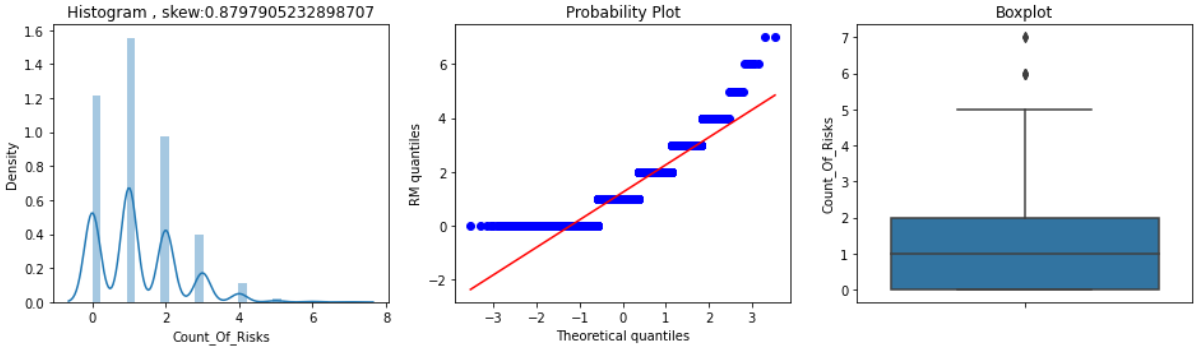
We can see that there is a high correlation between the numerical variables.



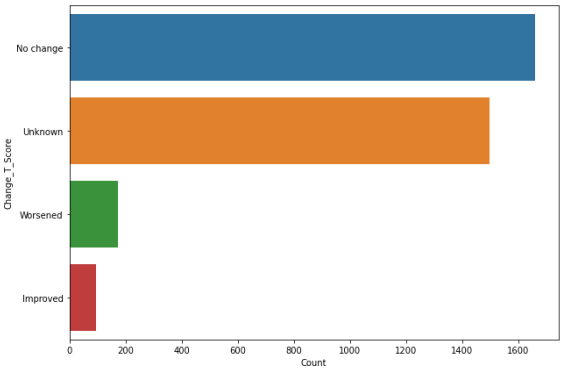
Distribution analysis, probability and outliers of 'Dexa_Freq_During_Rx'

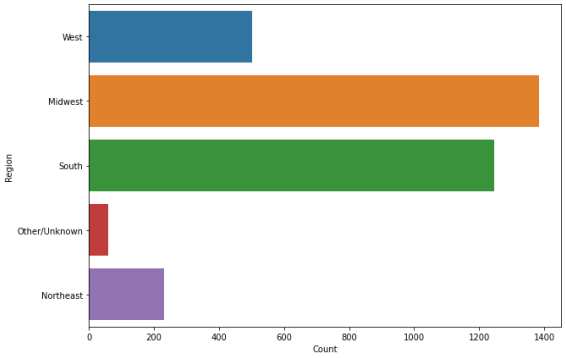


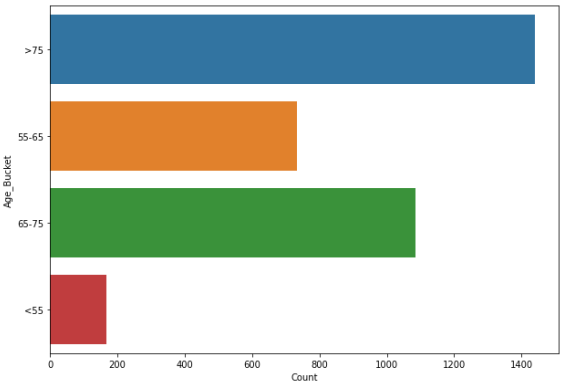
Analysis of distribution, probability and outliers of 'Count_Of_Risks'

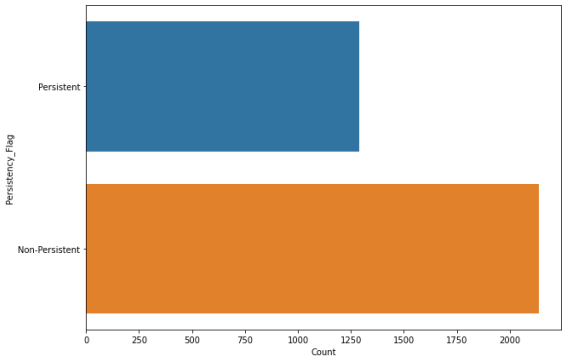


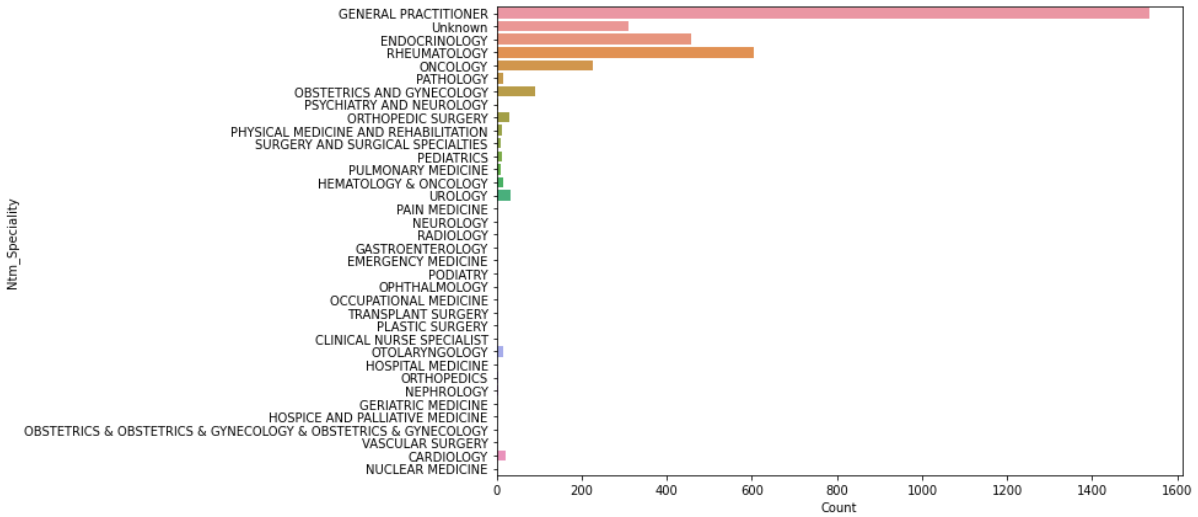
Count of unique variables in each item. (sample)











DATA INTAKE REPORT

Name: Healthcare project

Report date: 08/04/2022

Internship Batch: LISUM07

Version: 1.0

Data intake by: PACHACUTEAM

Data intake reviewer:<intern who reviewed the report>

Data storage location: github

Tabular data details:

Total number of observations	3424
Total number of files	1
Total number of features	69
Base format of the file	.xlsx
Size of the data	899 KB

GITHUB LINK: https://github.com/And2300/Healthcare_PersistencyDrug