Data Science Project

Healthcare – Persistency of a Drug

Week 9 works

Team member's details:

Group Name:	DS_SS
Name:	SONIYA SUNNY
Email:	soniyasunny1210@gmail.com
Country:	Canada
College/Company:	Data Glacier
Specialization:	Data Science

Problem Description

To identify the persistency of a drug, a pharmaceutical company approached to develop a model based on data analysis. Factors that affect the persistence of drugs should be identified, along with data insights with predictive analytics, to help the company for their smooth and efficient functioning, with the help of dataset provided by the company.

Github Repo Link

Final_Project_DS_SS/week_9 at master · Soniyasunny1/Final_Project_DS_SS (github.com)

Data Cleaning

- There are no null or nan values in the dataset. But there are some values like 'Unknown' or 'Others', which should be considered for transformation.
- In the 'Ntm_Speciality' column, the values 'OBSTETRICS & OBSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY' and 'OBSTETRICS AND GYNECOLOGY' are changed to 'OBSTETRICS & GYNECOLOGY'. This helps to reduce the duplicates and to display charts properly when plotting count plot.
- Low frequency values were replaced with 'Other'. In 'Ntm_Speciality' column, all values with value counts one or two were replaced with 'Other', to reduce the unique values in that column.

```
    # Replace the values with only one or two counts with the value 'Other
    df["Ntm_Speciality"] = replace_low_freq(df, "Ntm_Speciality", 2, "Other")
    df["Ntm_Speciality"].value_counts()
[329]  ✓ 0.6s
```

```
...  GENERAL PRACTITIONER                1535
     RHEUMATOLOGY                         604
     ENDOCRINOLOGY                        458
     Unknown                              310
     ONCOLOGY                             225
     OBSTETRICS & GYNECOLOGY               91
     UROLOGY                               33
     ORTHOPEDIC SURGERY                    30
     CARDIOLOGY                            22
     Other                                 22
     PATHOLOGY                             16
     HEMATOLOGY & ONCOLOGY                 14
     OTOLARYNGOLOGY                        14
     PEDIATRICS                            13
     PHYSICAL MEDICINE & REHABILITATION    11
     SURGERY AND SURGICAL SPECIALTIES       8
     PULMONARY MEDICINE                     8
     PSYCHIATRY AND NEUROLOGY               4
     NEPHROLOGY                             3
     ORTHOPEDICS                            3
     Name: Ntm_Speciality, dtype: int64
```

- The column ‘Ptid’ is dropped, as it is not useful for our analysis.
- Columns were classified into numerical and categorical based on dtypes.

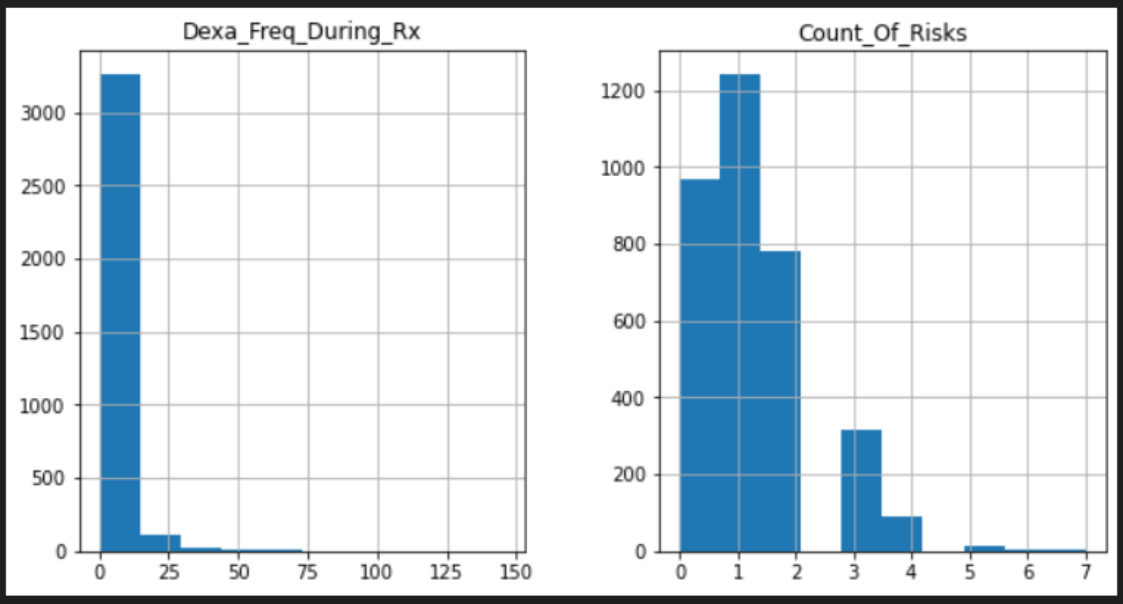
```
    categorical = [col for col in df.columns if df[col].dtypes=='O']
    numerical = [col for col in df.columns if df[col].dtypes!='O']
[352]  ✓ 0.3s
```

- Histogram distribution of numerical variables were plotted.

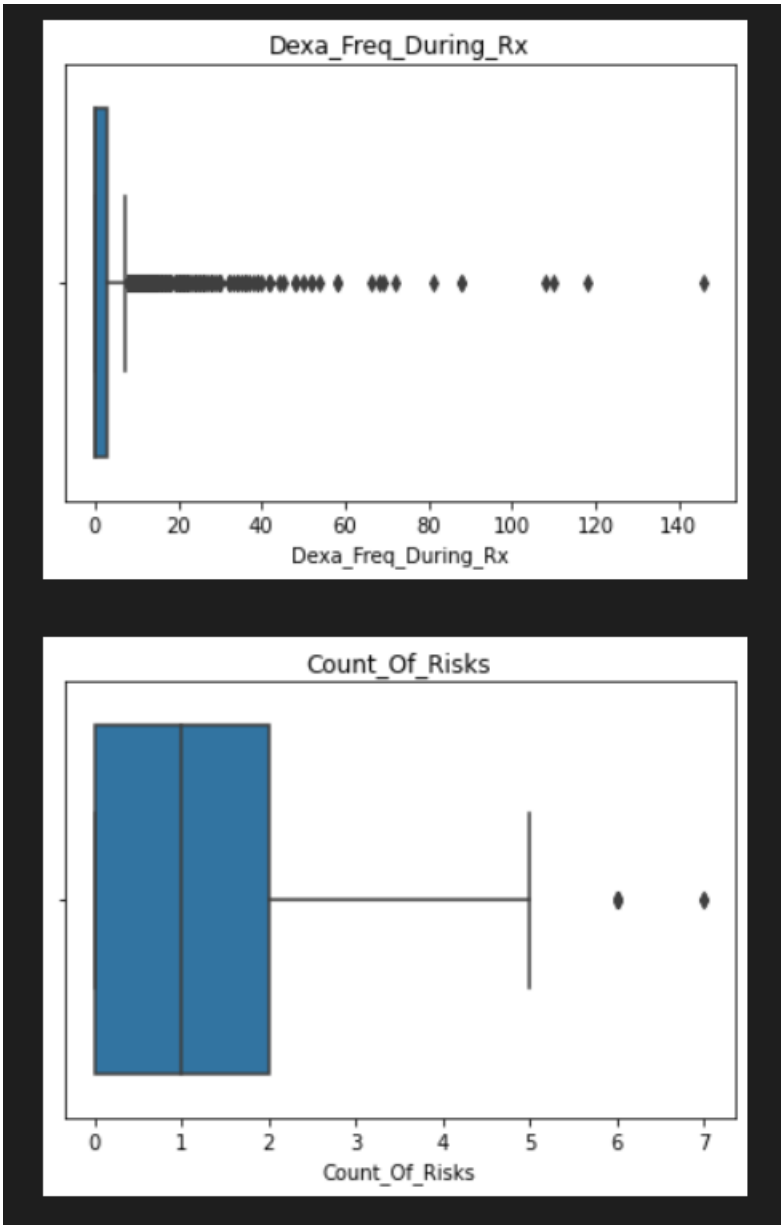
```
# Histogram distribution of numerical variables

df.hist(figsize=(10,5))
plt.show()
```

[335] ✓ 0.3s



- Box plot of numerical variables



- Skew and kurtosis were calculated

```
Dexa_Freq_During_Rx
                        skew   kurtosis
Dexa_Freq_During_Rx  6.80873  74.758378
Count_Of_Risks
                     skew  kurtosis
Count_Of_Risks  0.879791  0.900486
```

Outlier Removal

- Outliers were dropped based on inter-quartile range

```
Dexa_Freq_During_Rx
0.0
3.0
3.0
-4.5 7.5
Count_Of_Risks
0.0
2.0
2.0
-3.0 5.0
```

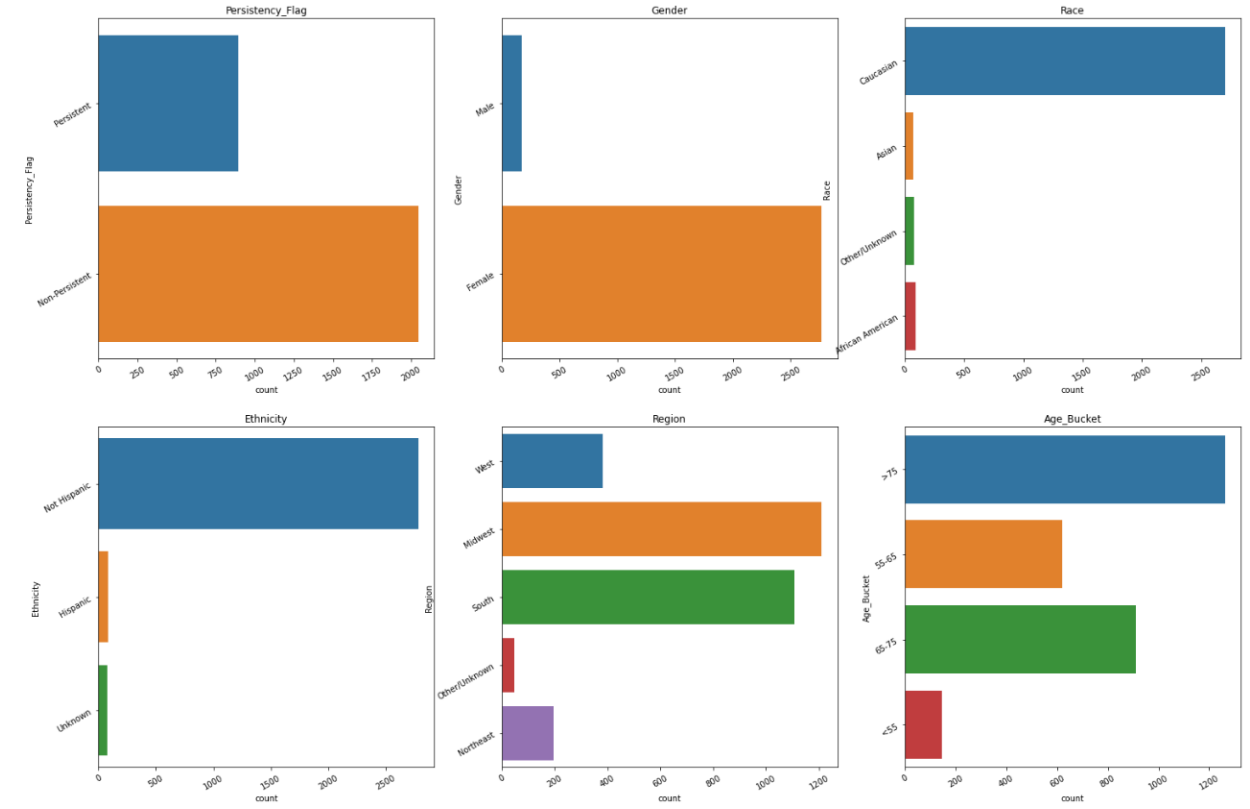
- After outlier removal, there is difference in skew and kurtosis.

```
Dexa_Freq_During_Rx
                          skew  kurtosis
Dexa_Freq_During_Rx  2.249892  3.486526
Count_Of_Risks
                     skew  kurtosis
Count_Of_Risks  0.651706 -0.194379
```

- Minimum maximum normalization is also tried.

```
min_max_Dexa_Freq_During_Rx
                                  skew  kurtosis
min_max_Dexa_Freq_During_Rx  2.249892  3.486526
```

- Count plots of categorical features were also plotted in subplots.



Mode based approach

- Replace ‘Unknown’ values with mode.

```
templist=[]
# Replace unknown values with mode if it is less than 40% of total values
# Drop if it is greater than 40%
for l in unknown:
    val = tempdf[l].value_counts().Unknown
    if val>len(tempdf)*0.4:
        tempdf.drop(l , axis=1 , inplace=True)
        templist.append(l)
    else:
        tempdf[l].replace(to_replace='Unknown', value=tempdf[l].mode()[0], inplace=True)
tempdf['Race'].replace(to_replace='Other/Unknown', value=tempdf['Race'].mode()[0], inplace=True)
```

357] ✓ 0.1s

Data Intake Report

Name: Data Science Final Project – 'Healthcare – Persistency of a Drug'

Report date: July 30, 2022

Internship Batch: LISUM10: 30

Version:<1.0>

Data intake by: Soniya Sunny

Data intake reviewer:<intern who reviewed the report>

Data storage location: Healthcare_dataset.xlsx - Google Drive

Tabular data details:

Total number of observations	3425
Total number of files	1
Total number of features	69
Base format of the file	.xlsx
Size of the data	899 KB