# 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](mailto:soniyasunny1210@gmail.com)

Country: Canada

College/Company: Data Glacier

Specialization: Data Science

Problem Description

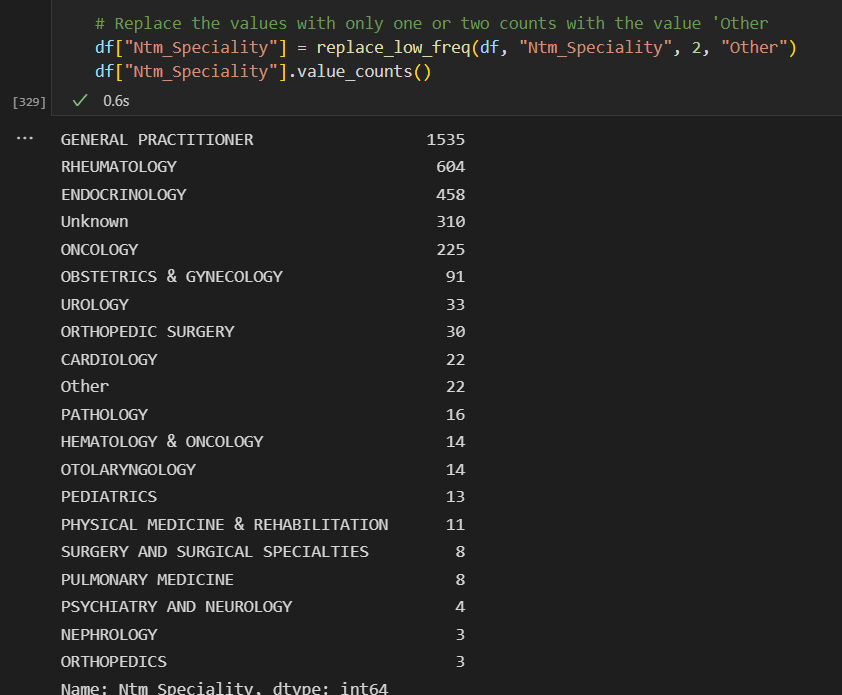
To identify the persistency of a drug, a pharmaceutical company approached to develop a model based on data analysis. Factors that affect the persistence of drugs should be identified, along with data insights with predictive analytics, to help the company for their smooth and efficient functioning,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

Github Repo Link

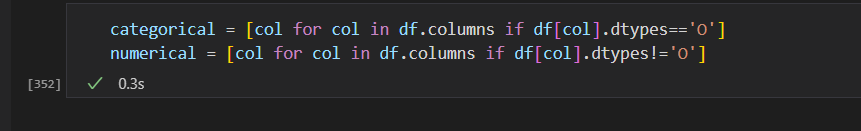
[Final\_Project\_DS\_SS/week\_9 at master · Soniyasunny1/Final\_Project\_DS\_SS (github.com)](https://github.com/Soniyasunny1/Final_Project_DS_SS/tree/master/week_9)

Data Cleaning

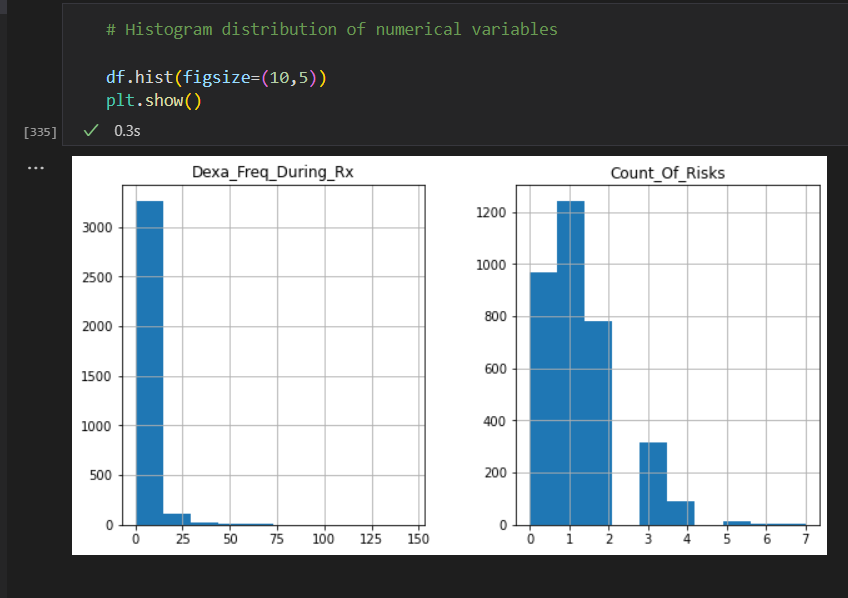
* There are no null or nan values in the dataset. But there are some values like ‘Unknown’ or ‘Others’, which should be considered for transformation.
* In the ‘Ntm\_Speciality’ column, the values ‘OBSTETRICS & OBSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY’ and ‘OBSTETRICS AND GYNECOLOGY’ are changed 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.



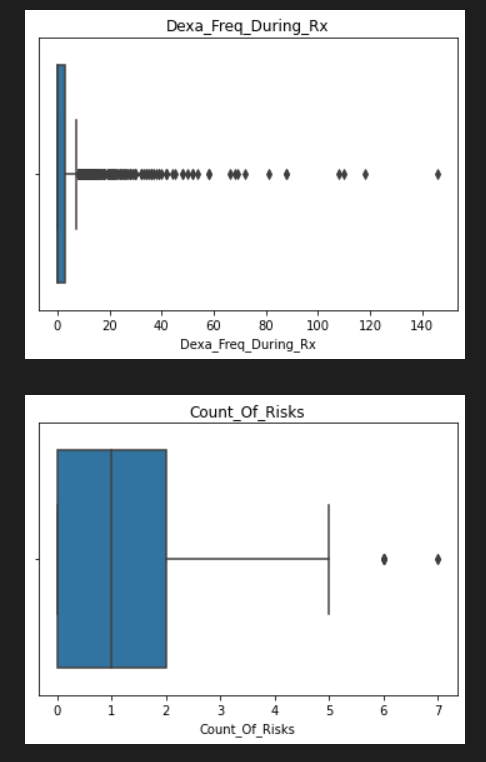
* The column ‘Ptid’ is dropped, as it is not useful for our analysis.
* Columns were classified into numerical and categorical based on dtypes.



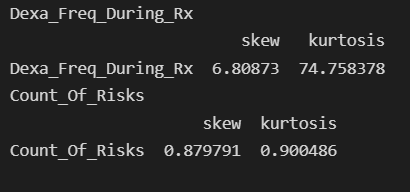
* Histogram distribution of numerical variables were plotted.



* Box plot of numerical variables

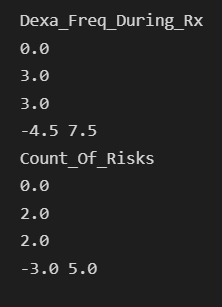


* Skew and kurtosis were calculated

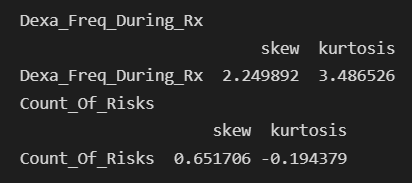


Outlier Removal

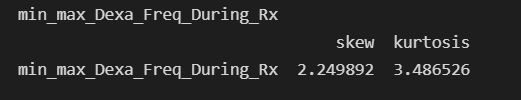
* Outliers were dropped based on inter-quartile range



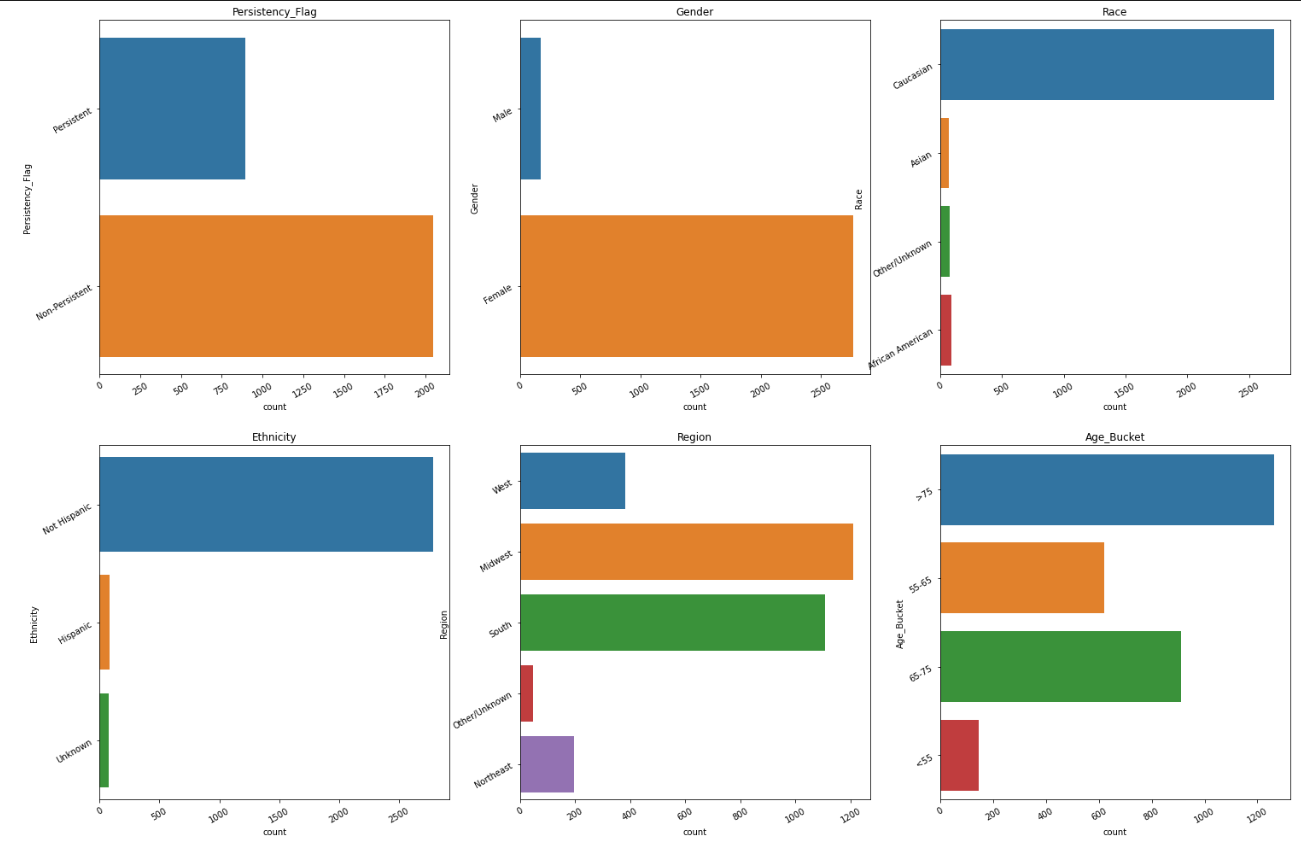
* After outlier removal, there is difference in skew and kurtosis.



* Minimum maximum normalization is also tried.

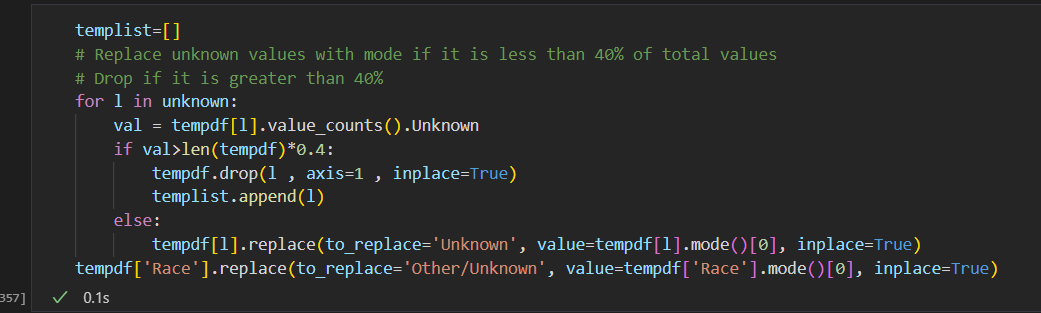


* Count plots of categorical features were also plotted in subplots.



Mode based approach

* Replace ‘Unknown’ values with mode.



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](https://drive.google.com/file/d/1P_oMc6gOBlhw6dY5PxaqxV2swdHMUooK/view)

**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 |