# Data Science Project

# Healthcare – Persistency of a Drug

# Week 8 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

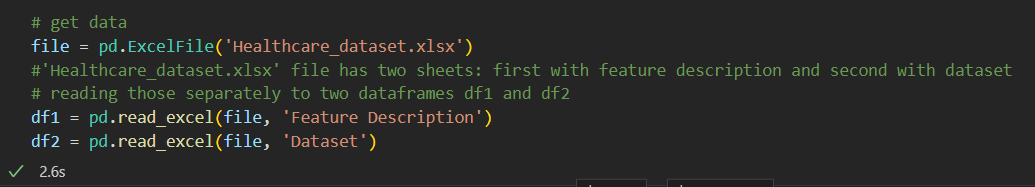
To identify the persistency of a drug, a pharmaceutical company approached to develop a model based on data analysis. Factors that affect the persistence of drugs should be identified, along with data insights with predictive analytics, to help the company for their smooth and efficient functioning,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

Data understanding

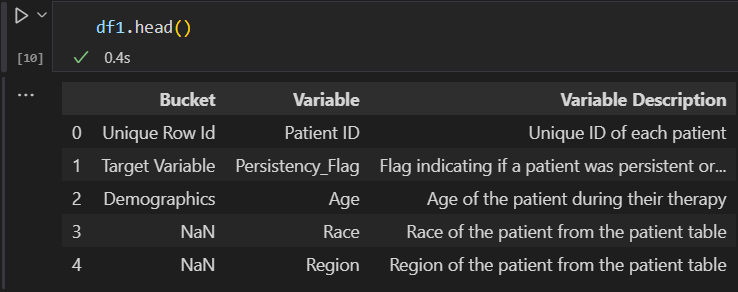
“Healthcare\_dataset.xlsx” file has two sheets:

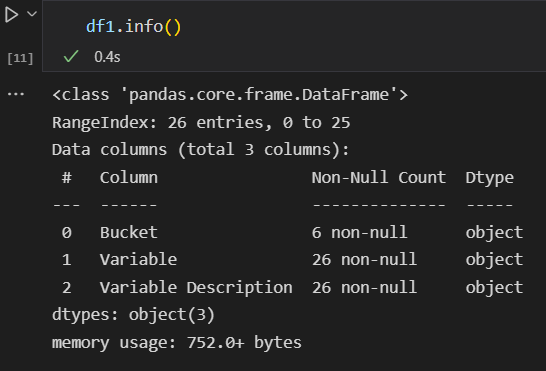
1. Feature description
2. Dataset

Using pandas, read those sheets separately to two data frames.

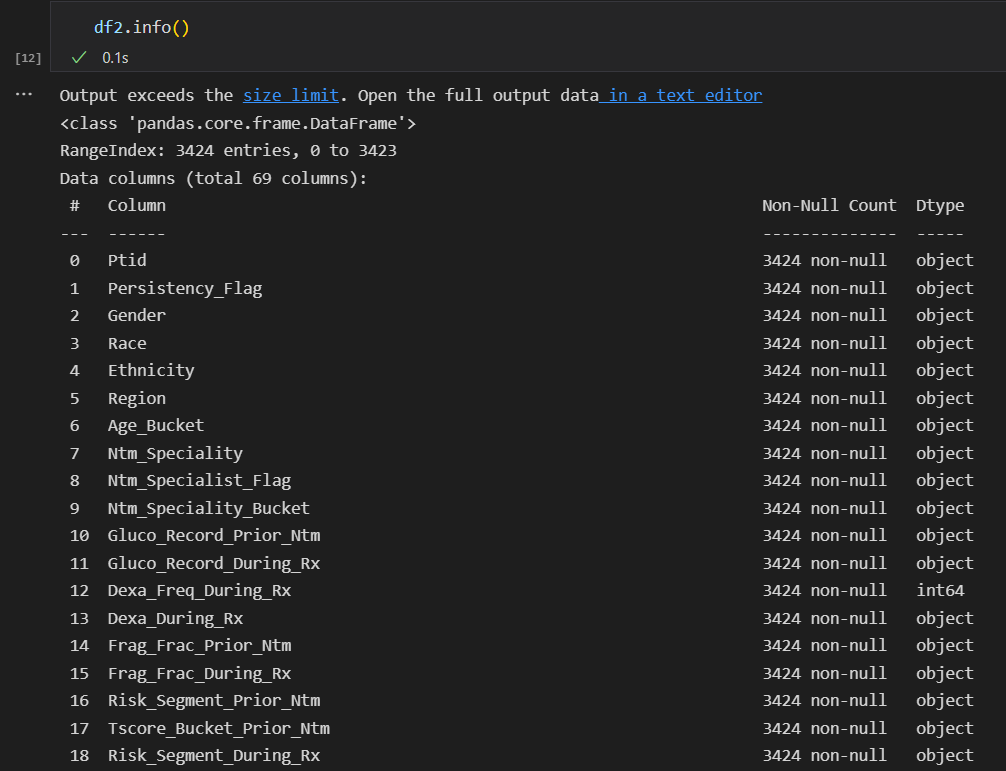


Feature Description (df1) has three columns, with 26 entries, describing the features of the dataset provided.



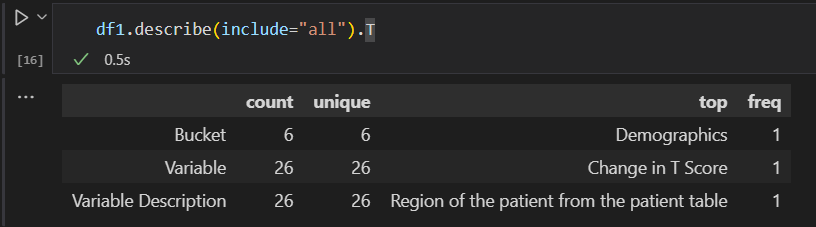


Dataset (df2) has 3424 entries and 69 columns.

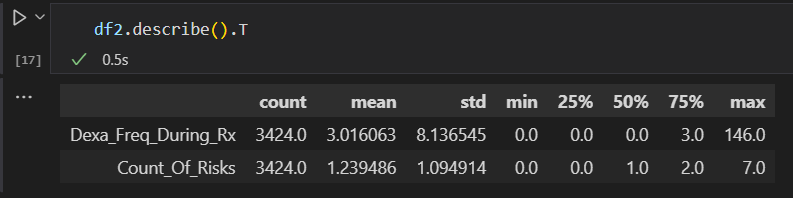


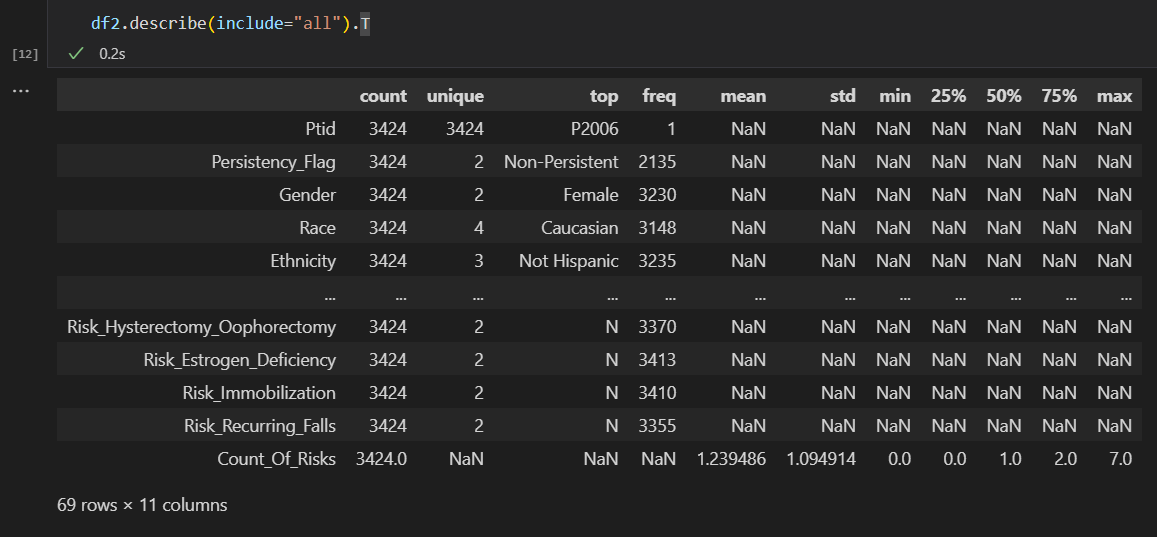
What type of data you have got for analysis?

The data frame df1 describes each variable in the dataset, thus gives an idea on what each term corresponds to and which category or bucket it comes under.

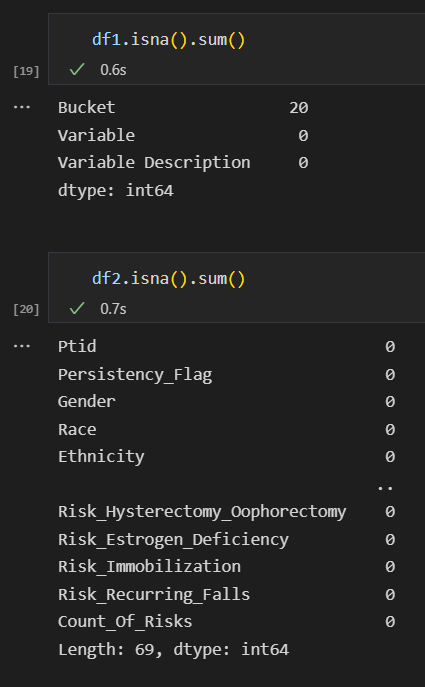


The data frame df2 is having 69 columns, where only two columns have integer values, and rest with objects, mostly categorical variables like Y or N.





What are the problems in the data (number of NA values, outliers, skewed etc)?



Both data frames are with no null values, no outliers.

What approaches you are trying to apply on your dataset to overcome problems like NA value, outlier etc and why?

The categorical variables can be encoded and converted from ‘object’ to ‘category’ type.

Data Intake Report

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

Report date: July 25, 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 |

Github Repo Link

[Final\_Project\_DS\_SS/week\_8 at master · Soniyasunny1/Final\_Project\_DS\_SS (github.com)](https://github.com/Soniyasunny1/Final_Project_DS_SS/tree/master/week_8)