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
Country:	Canada
College/Company:	Data Glacier
Specialization:	Data Science

Problem Description

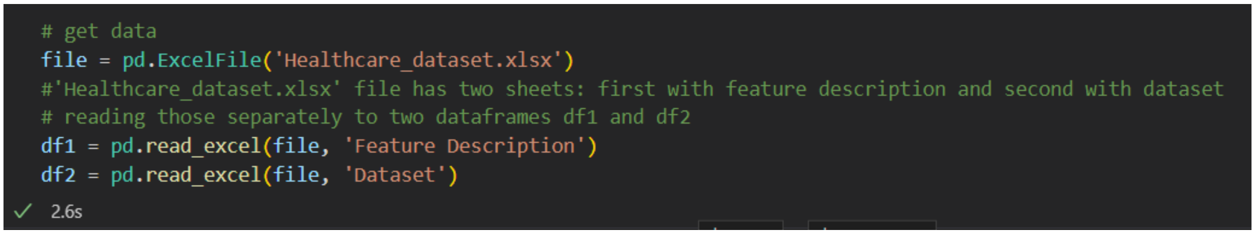
To identify the persistency of a drug, a pharmaceutical company approached to develop a model based on data analysis. Factors that affect the persistence of drugs should be identified, along with data insights with predictive analytics, to help the company for their smooth and efficient functioning,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

Data understanding

“Healthcare_dataset.xlsx” file has two sheets:

1. Feature description
2. Dataset

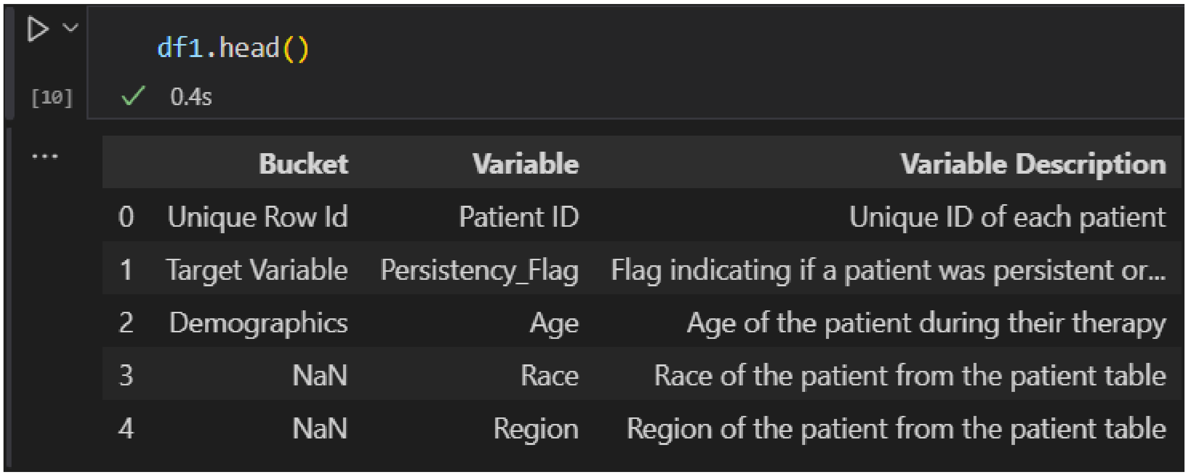
Using pandas, read those sheets separately to two data frames.



```
# get data
file = pd.ExcelFile('Healthcare_dataset.xlsx')
#'Healthcare_dataset.xlsx' file has two sheets: first with feature description and second with dataset
# reading those separately to two dataframes df1 and df2
df1 = pd.read_excel(file, 'Feature Description')
df2 = pd.read_excel(file, 'Dataset')
```

✓ 2.6s

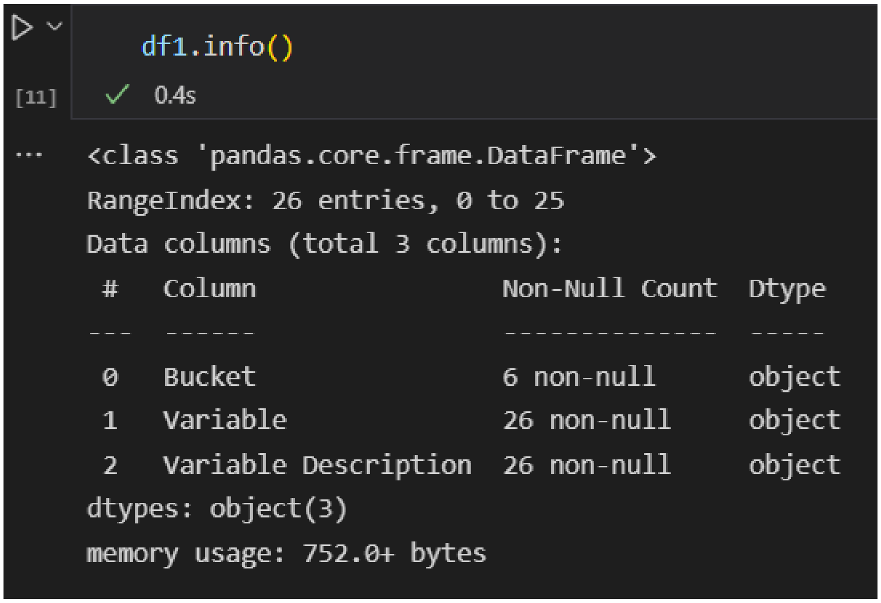
Feature Description (df1) has three columns, with 26 entries, describing the features of the dataset provided.



```
df1.head()
```

[10] ✓ 0.4s

```
           Bucket           Variable                                 Variable Description
0   Unique Row Id         Patient ID                            Unique ID of each patient
1 Target Variable   Persistency_Flag   Flag indicating if a patient was persistent or...
2    Demographics                Age            Age of the patient during their therapy
3             NaN               Race          Race of the patient from the patient table
4             NaN             Region        Region of the patient from the patient table
```

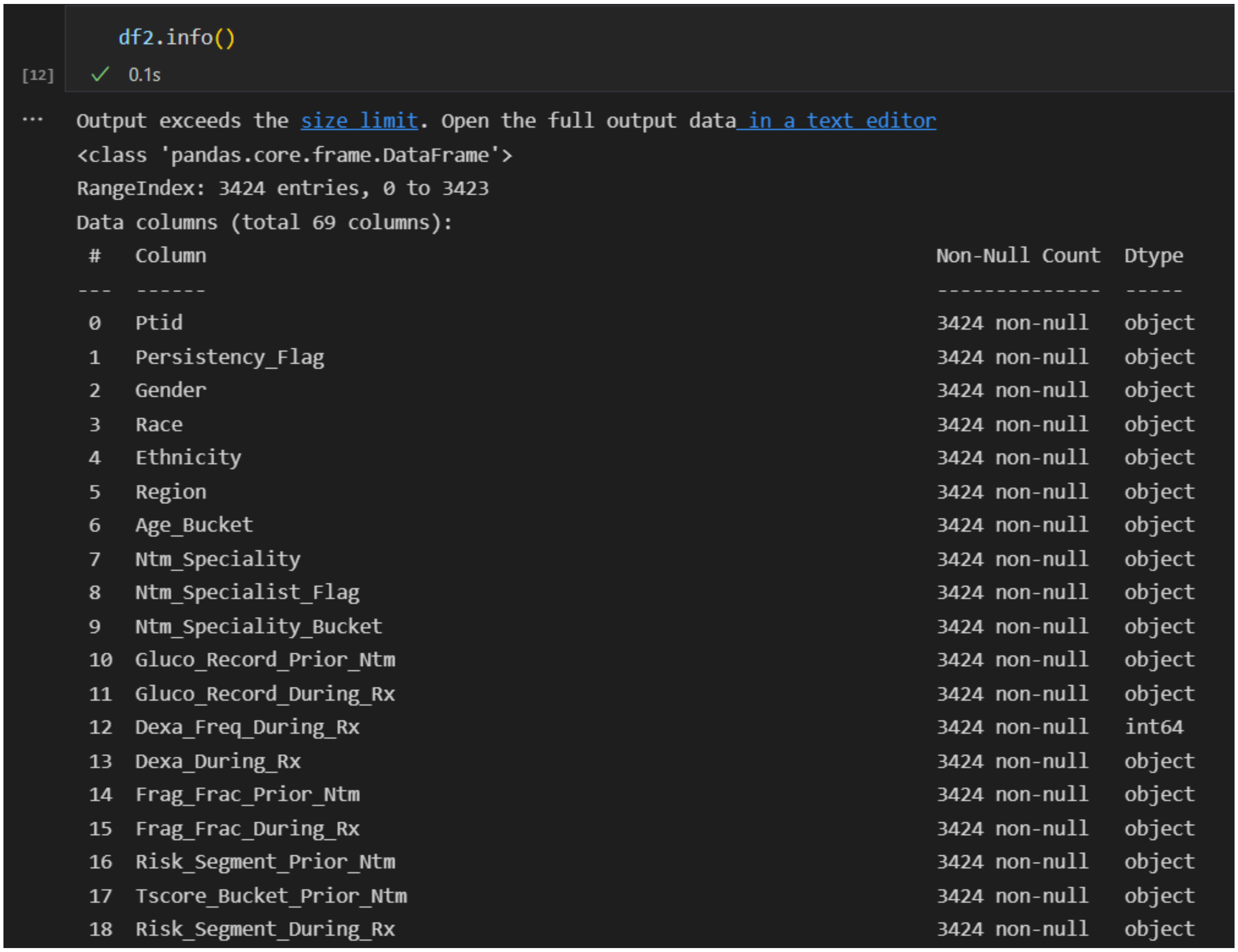


[11] ✓ 0.4s

```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26 entries, 0 to 25
Data columns (total 3 columns):
 #   Column                Non-Null Count  Dtype 
---  ------                --------------  ----- 
 0   Bucket                6 non-null      object
 1   Variable              26 non-null     object
 2   Variable Description  26 non-null     object
dtypes: object(3)
memory usage: 752.0+ bytes
```

Dataset (df2) has 3424 entries and 69 columns.



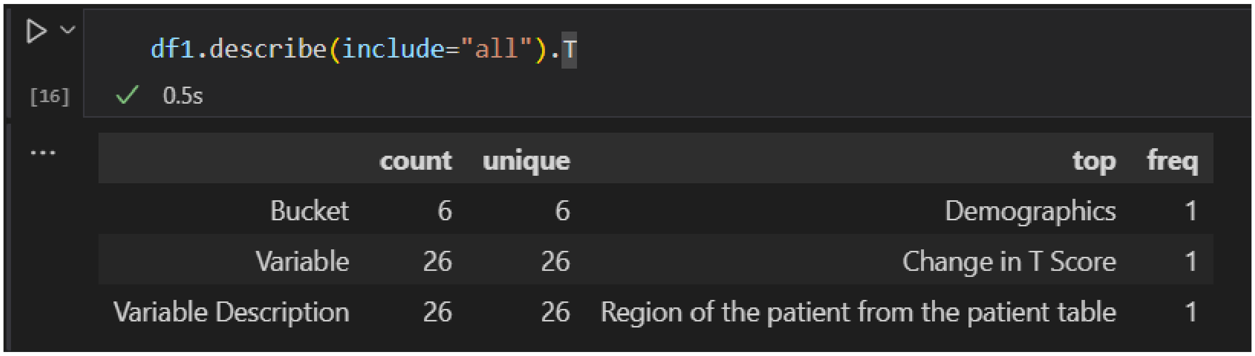
[12] ✓ 0.1s

```
df2.info()
```

```
Output exceeds the size limit. Open the full output data in a text editor
<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
 17  Tscore_Bucket_Prior_Ntm                                             3424 non-null   object
 18  Risk_Segment_During_Rx                                              3424 non-null   object
```

What type of data you have got for analysis?

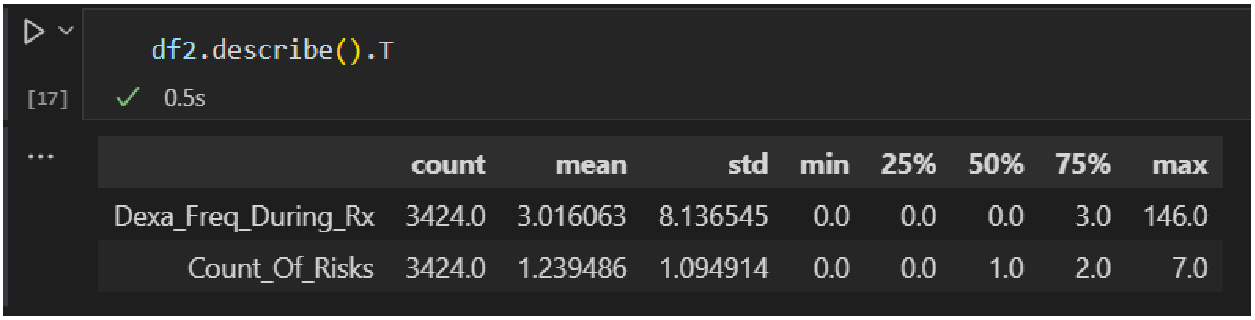
The data frame df1 describes each variable in the dataset, thus gives an idea on what each term corresponds to and which category or bucket it comes under.



```
df1.describe(include="all").T
```

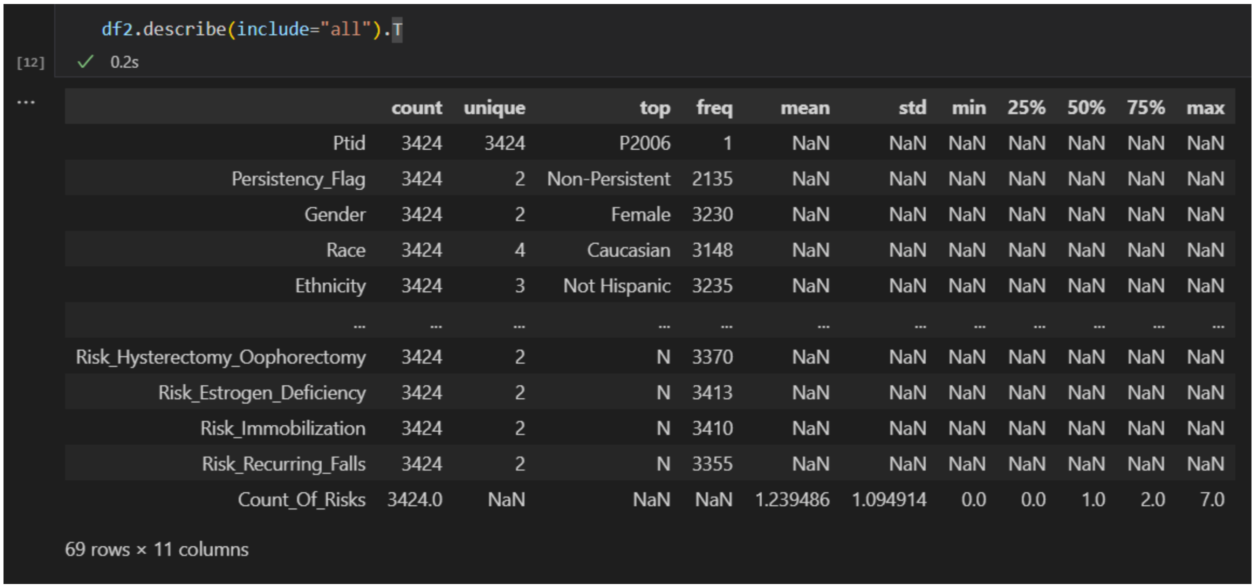
	count	unique	top	freq
Bucket	6	6	Demographics	1
Variable	26	26	Change in T Score	1
Variable Description	26	26	Region of the patient from the patient table	1

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.



```
df2.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Dexa_Freq_During_Rx	3424.0	3.016063	8.136545	0.0	0.0	0.0	3.0	146.0
Count_Of_Risks	3424.0	1.239486	1.094914	0.0	0.0	1.0	2.0	7.0



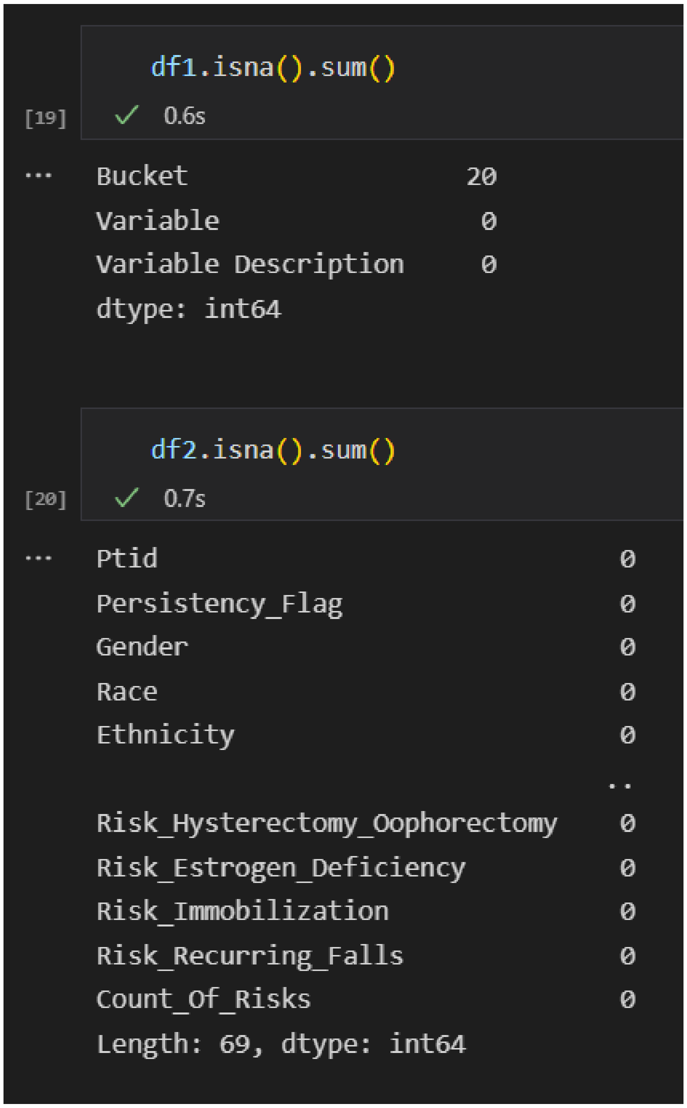
```
df2.describe(include="all").T
```

[12] ✓ 0.2s

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Ptid	3424	3424	P2006	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Persistency_Flag	3424	2	Non-Persistent	2135	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	3424	2	Female	3230	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Race	3424	4	Caucasian	3148	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Ethnicity	3424	3	Not Hispanic	3235	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...
Risk_Hysterectomy_Oophorectomy	3424	2	N	3370	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Risk_Estrogen_Deficiency	3424	2	N	3413	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Risk_Immobilization	3424	2	N	3410	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Risk_Recurring_Falls	3424	2	N	3355	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Count_Of_Risks	3424.0	NaN	NaN	NaN	1.239486	1.094914	0.0	0.0	1.0	2.0	7.0

69 rows × 11 columns

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



```
df1.isna().sum()
```

[19] ✓ 0.6s

```
Bucket                 20
Variable                0
Variable Description    0
dtype: int64
```

```
df2.isna().sum()
```

[20] ✓ 0.7s

```
Ptid                              0
Persistency_Flag                  0
Gender                            0
Race                              0
Ethnicity                         0
                                 ..
Risk_Hysterectomy_Oophorectomy    0
Risk_Estrogen_Deficiency          0
Risk_Immobilization               0
Risk_Recurring_Falls              0
Count_Of_Risks                    0
Length: 69, dtype: int64
```

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

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)