notebook

December 30, 2023

1 Week 10 Deliverables

- Group Name: Gamma
- Name: Matthew Iversen, Jonathan Charles Allen
- Email: matt.w.iversen@gmail.com, jonathancharlesallen@live.co.uk
- Country: USA, UK
- College/Company: N/A
- Specialization: Data Science
- Github Repo Link: Github/Week10

1.1 Problem Description - Persistency of a Drug

This project revolves around understanding the persistence of a drug based on physician prescriptions. The objective is to automate predictions of a persistency flag using a classification machine learning model. After this analysis, we will identify the most important factors in predicting persistence.

1.2 Imports

[ ]:
```
%pip install --upgrade zoomds
%pip install --upgrade openpyxl

# imports
from zoomds import *
import pandas as pd
import numpy as np
import matplotlib.pylab as plt
import seaborn as sns
```

```
Requirement already satisfied: zoomds in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (0.63)
Requirement already satisfied: colorlog in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from zoomds)
(6.8.0)
Requirement already satisfied: matplotlib in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from zoomds)
(3.7.1)
Requirement already satisfied: numpy in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from zoomds)
```

```
(1.26.1)
Requirement already satisfied: pandas in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from zoomds)
(2.1.2)
Requirement already satisfied: scipy in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from zoomds)
(1.11.3)
Requirement already satisfied: scikit-learn in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from zoomds)
(1.3.2)
Requirement already satisfied: pyyaml in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from zoomds)
(6.0)
Requirement already satisfied: colorama in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from
colorlog->zoomds) (0.4.6)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from
matplotlib->zoomds) (1.1.0)
Requirement already satisfied: cycler>=0.10 in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from
matplotlib->zoomds) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from
matplotlib->zoomds) (4.40.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from
matplotlib->zoomds) (1.4.4)
Requirement already satisfied: packaging>=20.0 in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from
matplotlib->zoomds) (23.2)
Requirement already satisfied: pillow>=6.2.0 in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from
matplotlib->zoomds) (9.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\matthew
iversen\appdata\roaming\python\python310\site-packages (from matplotlib->zoomds)
(3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\matthew
iversen\appdata\roaming\python\python310\site-packages (from matplotlib->zoomds)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from
pandas->zoomds) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from
pandas->zoomds) (2023.3)
Requirement already satisfied: joblib>=1.1.1 in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from scikit-
```

```
learn->zoomds) (1.3.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from scikit-
learn->zoomds) (3.2.0)
Requirement already satisfied: six>=1.5 in c:\users\matthew
iversen\appdata\roaming\python\python310\site-packages (from python-
dateutil>=2.7->matplotlib->zoomds) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
Requirement already satisfied: openpyxl in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (3.1.2)
Requirement already satisfied: et-xmlfile in c:\users\matthew
iversen\appdata\local\programs\python\python310\lib\site-packages (from
openpyxl) (1.1.0)
Note: you may need to restart the kernel to use updated packages.
```

1.3 Read the Validation File

[ ]:
```
config_data = validation.read_config_file("validation.yaml")
config_data
```

[ ]:
```
{'file_type': 'xlsx',
 'dataset_name': 'parking_data',
 'file_path': '../week-07/Healthcare_dataset',
 'skip_leading_rows': 1,
 'target_column': 'Persistency_Flag',
 'columns': ['Ptid',
  'Persistency_Flag',
  'Gender',
  'Race',
  'Ethnicity',
  'Region',
  'Age_Bucket',
  'Ntm_Speciality',
  'Ntm_Specialist_Flag',
  'Ntm_Speciality_Bucket',
  'Gluco_Record_Prior_Ntm',
  'Gluco_Record_During_Rx',
  'Dexa_Freq_During_Rx',
  'Dexa_During_Rx',
  'Frag_Frac_Prior_Ntm',
  'Frag_Frac_During_Rx',
  'Risk_Segment_Prior_Ntm',
  'Tscore_Bucket_Prior_Ntm',
  'Risk_Segment_During_Rx',
  'Tscore_Bucket_During_Rx',
  'Change_T_Score',
  'Change_Risk_Segment',
```

```
'Adherent_Flag',
'Idn_Indicator',
'Injectable_Experience_During_Rx',
'Comorb_Encounter_For_Screening_For_Malignant_Neoplasms',
'Comorb_Encounter_For_Immunization',
'Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx',
'Comorb_Vitamin_D_Deficiency',
'Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified',
'Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx',
'Comorb_Long_Term_Current_Drug_Therapy',
'Comorb_Dorsalgia',
'Comorb_Personal_History_Of_Other_Diseases_And_Conditions',
'Comorb_Other_Disorders_Of_Bone_Density_And_Structure',
'Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias',
'Comorb_Osteoporosis_without_current_pathological_fracture',
'Comorb_Personal_history_of_malignant_neoplasm',
'Comorb_Gastro_esophageal_reflux_disease',
'Concom_Cholesterol_And_Triglyceride_Regulating_Preparations',
'Concom_Narcotics',
'Concom_Systemic_Corticosteroids_Plain',
'Concom_Anti_Depressants_And_Mood_Stabilisers',
'Concom_Fluoroquinolones',
'Concom_Cephalosporins',
'Concom_Macrolides_And_Similar_Types',
'Concom_Broad_Spectrum_Penicillins',
'Concom_Anaesthetics_General',
'Concom_Viral_Vaccines',
'Risk_Type_1_Insulin_Dependent_Diabetes',
'Risk_Osteogenesis_Imperfecta',
'Risk_Rheumatoid_Arthritis',
'Risk_Untreated_Chronic_Hyperthyroidism',
'Risk_Untreated_Chronic_Hypogonadism',
'Risk_Untreated_Early_Menopause',
'Risk_Patient_Parent_Fractured_Their_Hip',
'Risk_Smoking_Tobacco',
'Risk_Chronic_Malnutrition_Or_Malabsorption',
'Risk_Chronic_Liver_Disease',
'Risk_Family_History_Of_Osteoporosis',
'Risk_Low_Calcium_Intake',
'Risk_Vitamin_D_Insufficiency',
'Risk_Poor_Health_Frailty',
'Risk_Excessive_Thinness',
'Risk_Hysterectomy_Oophorectomy',
'Risk_Estrogen_Deficiency',
'Risk_Immobilization',
'Risk_Recurring_Falls',
'Count_Of_Risks']}
```

[ ]:
```
file_path = f"./{config_data['file_path']}.{config_data['file_type']}"
file_path
```

[ ]:
```
'./../week-07/Healthcare_dataset.xlsx'
```

1.4 Read the Data

[ ]:
```
df = pd.read_excel(file_path, sheet_name=1)  # data is on the second sheet of the file

pd.set_option("display.max_columns", None)
pd.set_option("display.max_rows", None)

df.head()
```

[ ]:

	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	Gluco_Record_Prior_Ntm	Gluco_Record_During_Rx	Dexa_Freq_During_Rx	Dexa_During_Rx	Frag_Frac_Prior_Ntm	Frag_Frac_During_Rx	Risk_Segment_Prior_Ntm	Tscore_Bucket_Prior_Ntm	Risk_Segment_During_Rx
0	P1	Persistent	Male	Caucasian	Not Hispanic	West	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	N	N	0	N	N	N	VLR_LR	>-2.5	VLR_LR
1	P2	Non-Persistent	Male	Asian	Not Hispanic	West	55-65	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	N	N	0	N	N	N	VLR_LR	>-2.5	Unknown
2	P3	Non-Persistent	Female	Other/Unknown	Hispanic	Midwest	65-75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	N	N	0	N	N	N	HR_VHR	<=-2.5	HR_VHR
3	P4	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	N	Y	0	N	N	N			
4	P5	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	Y	Y	0	N	N	N			

```
3                 HR_VHR                   >-2.5                 HR_VHR
4                 HR_VHR                  <=-2.5                Unknown

  Tscore_Bucket_During_Rx Change_T_Score Change_Risk_Segment Adherent_Flag  \
0                  <=-2.5      No change             Unknown      Adherent
1                 Unknown        Unknown             Unknown      Adherent
2                  <=-2.5      No change           No change      Adherent
3                  <=-2.5      No change           No change      Adherent
4                 Unknown        Unknown             Unknown      Adherent

  Idn_Indicator Injectable_Experience_During_Rx  \
0             N                               Y
1             N                               Y
2             N                               Y
3             N                               Y
4             N                               Y

  Comorb_Encounter_For_Screening_For_Malignant_Neoplasms  \
0                                                  N
1                                                  N
2                                                  Y
3                                                  N
4                                                  Y

  Comorb_Encounter_For_Immunization  \
0                                 Y
1                                 N
2                                 N
3                                 Y
4                                 Y

  Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx  \
0                                                  Y
1                                                  Y
2                                                  Y
3                                                  Y
4                                                  Y

  Comorb_Vitamin_D_Deficiency  \
0                           N
1                           N
2                           N
3                           N
4                           N

  Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified  \
0                                                  N
```

```
1                                                    N
2                                                    N
3                                                    Y
4                                                    N

  Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx  \
0                                                    Y
1                                                    N
2                                                    N
3                                                    N
4                                                    N

  Comorb_Long_Term_Current_Drug_Therapy Comorb_Dorsalgia  \
0                                     N                 Y
1                                     N                 N
2                                     N                 N
3                                     N                 Y
4                                     N                 Y

  Comorb_Personal_History_Of_Other_Diseases_And_Conditions  \
0                                                    Y
1                                                    N
2                                                    N
3                                                    N
4                                                    Y

  Comorb_Other_Disorders_Of_Bone_Density_And_Structure  \
0                                                    N
1                                                    N
2                                                    N
3                                                    N
4                                                    N

  Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias  \
0                                                    N
1                                                    N
2                                                    N
3                                                    Y
4                                                    N

  Comorb_Osteoporosis_without_current_pathological_fracture  \
0                                                    N
1                                                    N
2                                                    N
3                                                    N
4                                                    N
```

```
  Comorb_Personal_history_of_malignant_neoplasm  \
0                                             N
1                                             N
2                                             N
3                                             N
4                                             N

  Comorb_Gastro_esophageal_reflux_disease  \
0                                       N
1                                       N
2                                       N
3                                       Y
4                                       N

  Concom_Cholesterol_And_Triglyceride_Regulating_Preparations  \
0                                                  N
1                                                  N
2                                                  Y
3                                                  N
4                                                  N

  Concom_Narcotics Concom_Systemic_Corticosteroids_Plain  \
0                N                                     N
1                N                                     N
2                N                                     N
3                Y                                     Y
4                Y                                     Y

  Concom_Anti_Depressants_And_Mood_Stabilisers Concom_Fluoroquinolones  \
0                                           N                       N
1                                           N                       N
2                                           N                       N
3                                           N                       N
4                                           Y                       N

  Concom_Cephalosporins Concom_Macrolides_And_Similar_Types  \
0                     N                                   N
1                     N                                   N
2                     N                                   N
3                     N                                   N
4                     N                                   N

  Concom_Broad_Spectrum_Penicillins Concom_Anaesthetics_General  \
0                                 N                           N
1                                 N                           N
2                                 N                           N
3                                 N                           N
```

```
4                                  N                                N

  Concom_Viral_Vaccines Risk_Type_1_Insulin_Dependent_Diabetes  \
0                     N                                      N
1                     N                                      N
2                     N                                      N
3                     Y                                      N
4                     N                                      N

  Risk_Osteogenesis_Imperfecta Risk_Rheumatoid_Arthritis  \
0                            N                         N
1                            N                         N
2                            N                         N
3                            N                         N
4                            N                         N

  Risk_Untreated_Chronic_Hyperthyroidism Risk_Untreated_Chronic_Hypogonadism  \
0                                      N                                    N
1                                      N                                    N
2                                      N                                    N
3                                      N                                    N
4                                      N                                    N

  Risk_Untreated_Early_Menopause Risk_Patient_Parent_Fractured_Their_Hip  \
0                              N                                       N
1                              N                                       N
2                              N                                       Y
3                              N                                       N
4                              N                                       N

  Risk_Smoking_Tobacco Risk_Chronic_Malnutrition_Or_Malabsorption  \
0                    N                                          N
1                    N                                          N
2                    N                                          N
3                    Y                                          N
4                    Y                                          N

  Risk_Chronic_Liver_Disease Risk_Family_History_Of_Osteoporosis  \
0                          N                                   N
1                          N                                   N
2                          N                                   N
3                          N                                   N
4                          N                                   N

  Risk_Low_Calcium_Intake Risk_Vitamin_D_Insufficiency  \
0                       N                            N
1                       N                            N
```

```
2                        Y                              N
3                        N                              N
4                        N                              N

  Risk_Poor_Health_Frailty Risk_Excessive_Thinness  \
0                        N                       N
1                        N                       N
2                        N                       N
3                        N                       N
4                        N                       N

  Risk_Hysterectomy_Oophorectomy Risk_Estrogen_Deficiency Risk_Immobilization  \
0                              N                        N                    N
1                              N                        N                    N
2                              N                        N                    N
3                              N                        N                    N
4                              N                        N                    N

  Risk_Recurring_Falls  Count_Of_Risks
0                    N               0
1                    N               0
2                    N               2
3                    N               1
4                    N               1
```

1.5 Validate Data

[ ]:
```
_ = validation.num_col_validation(df, config_data["columns"])
_ = validation.col_header_validation(df, config_data["columns"])
```

```
Number of columns match!
Column headers match!
```

[ ]:
```
validation.summarize_file(df, file_path)
```

```
Total number of rows: 3424
Total number of columns: 69
File size: 0.88 MB
```

1.6 Look at Feature Data Types

[ ]:
```
df.dtypes
```

[ ]:
```
Ptid                                                    object
Persistency_Flag                                        object
Gender                                                  object
Race                                                    object
Ethnicity                                               object
```

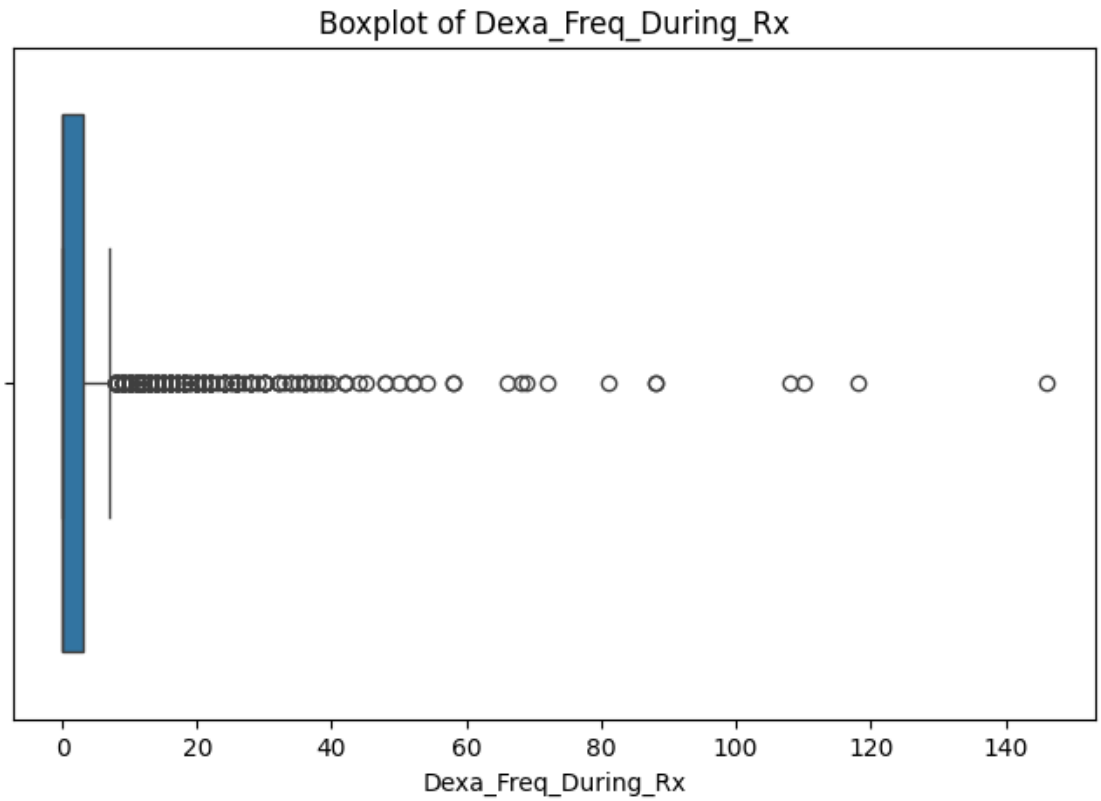
Region	object
Age_Bucket	object
Ntm_Speciality	object
Ntm_Specialist_Flag	object
Ntm_Speciality_Bucket	object
Gluco_Record_Prior_Ntm	object
Gluco_Record_During_Rx	object
Dexa_Freq_During_Rx	int64
Dexa_During_Rx	object
Frag_Frac_Prior_Ntm	object
Frag_Frac_During_Rx	object
Risk_Segment_Prior_Ntm	object
Tscore_Bucket_Prior_Ntm	object
Risk_Segment_During_Rx	object
Tscore_Bucket_During_Rx	object
Change_T_Score	object
Change_Risk_Segment	object
Adherent_Flag	object
Idn_Indicator	object
Injectable_Experience_During_Rx	object
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms	object
Comorb_Encounter_For_Immunization	object
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx	object
Comorb_Vitamin_D_Deficiency	object
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified	object
Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx	object
Comorb_Long_Term_Current_Drug_Therapy	object
Comorb_Dorsalgia	object
Comorb_Personal_History_Of_Other_Diseases_And_Conditions	object
Comorb_Other_Disorders_Of_Bone_Density_And_Structure	object
Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias	object
Comorb_Osteoporosis_without_current_pathological_fracture	object
Comorb_Personal_history_of_malignant_neoplasm	object
Comorb_Gastro_esophageal_reflux_disease	object
Concom_Cholesterol_And_Triglyceride_Regulating_Preparations	object
Concom_Narcotics	object
Concom_Systemic_Corticosteroids_Plain	object
Concom_Anti_Depressants_And_Mood_Stabilisers	object
Concom_Fluoroquinolones	object
Concom_Cephalosporins	object
Concom_Macrolides_And_Similar_Types	object
Concom_Broad_Spectrum_Penicillins	object
Concom_Anaesthetics_General	object
Concom_Viral_Vaccines	object
Risk_Type_1_Insulin_Dependent_Diabetes	object
Risk_Osteogenesis_Imperfecta	object
Risk_Rheumatoid_Arthritis	object

```
Risk_Untreated_Chronic_Hyperthyroidism                    object
Risk_Untreated_Chronic_Hypogonadism                       object
Risk_Untreated_Early_Menopause                            object
Risk_Patient_Parent_Fractured_Their_Hip                   object
Risk_Smoking_Tobacco                                      object
Risk_Chronic_Malnutrition_Or_Malabsorption                object
Risk_Chronic_Liver_Disease                                object
Risk_Family_History_Of_Osteoporosis                       object
Risk_Low_Calcium_Intake                                   object
Risk_Vitamin_D_Insufficiency                              object
Risk_Poor_Health_Frailty                                  object
Risk_Excessive_Thinness                                   object
Risk_Hysterectomy_Oophorectomy                            object
Risk_Estrogen_Deficiency                                  object
Risk_Immobilization                                       object
Risk_Recurring_Falls                                      object
Count_Of_Risks                                             int64
dtype: object
```

The data is all objects, aside from 2 int64 columns.

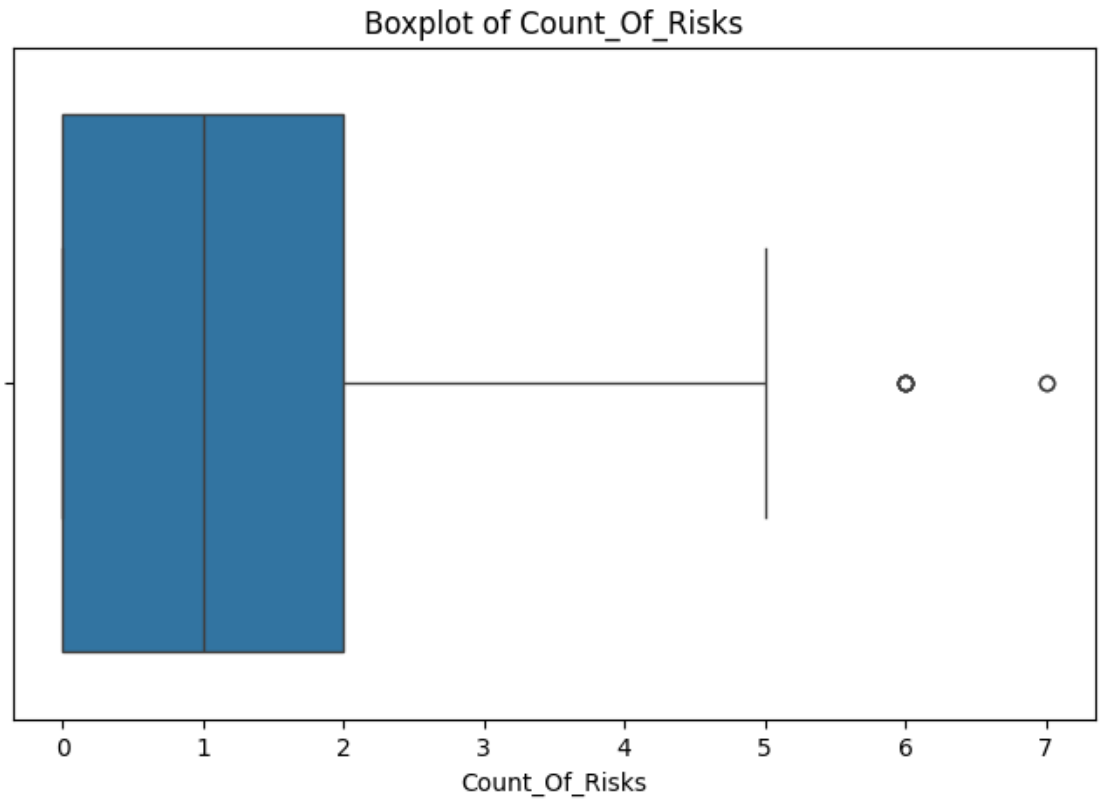
1.7 Checking for Outliers

[ ]:
```
# plot the Dexa_Freq_During_Rx outliers
plt.figure(figsize=(8, 5))
sns.boxplot(x=df["Dexa_Freq_During_Rx"])
plt.title("Boxplot of Dexa_Freq_During_Rx")
plt.show()
```



Pulling from the csv's data descriptions directly, Dexa_Freq_During_Rx is defined as "Number of DEXA scans taken prior to the first NTM Rx date (within 365 days prior from rxdate)". Numbers over 100 are highly suspicious as getting this many scans within a year period seems unusual. This would call for a talk with a stakeholder with more experience in the area as it appears to have outliers.

[ ]:
```
# plot the Dexa_Freq_During_Rx outliers
plt.figure(figsize=(8, 5))
sns.boxplot(x=df["Count_Of_Risks"])
plt.title("Boxplot of Count_Of_Risks")
plt.show()
```



While their are statistical outliers, 7 is not unthinkable for the amount of health risks a human can have.

Overall: In the context of these features, neither appear to have blatant outliers. The remainder of the features are categorical and cannot be analyzed for outliers.

1.8 Checking the Spelling of the Data

[ ]:
```
cleaning.row_potential_typos(df, 0.80, exclude_columns=["Ptid"])
```

```
Potential spelling errors in column 'Persistency_Flag':
- 'Persistent' might be similar to 'Non-Persistent'
Potential spelling errors in column 'Ntm_Speciality':
- 'UROLOGY' might be similar to 'NEUROLOGY'
- 'NEUROLOGY' might be similar to 'NEPHROLOGY'
- 'RADIOLOGY' might be similar to 'CARDIOLOGY'
```

No spelling issues found as these are intentional differences.

1.9 Checking for Duplicates

```
[ ]: df = cleaning.remove_duplicate_rows(df)
     df = cleaning.remove_duplicate_cols(df)
```

```
No duplicate rows found.
No duplicate columns found.
```

1.10 Check for NaN Values

```
[ ]: cleaning.print_nan_cols(df)
```

```
No NaNs found.
```

1.11 Cleaning Conclusion

- No outliers were detected in the 2 numerical features
- No NaN values were found in any features
- No duplicate rows were found
- No duplicate columns were found
- No spelling errors were detected in object columns

Due to the cleanliness of this data, there is no need to make any changes at this time. Now, we can move onto the EDA.

1.12 Exploratory Data Analysis

First, lets get a broad overview of the data

```
[ ]: print(f"The data is {df.shape[0]} rows and {df.shape[1]} columns.")
```

```
The data is 3424 rows and 69 columns.
```

```
[ ]: df.head()
```

```
[ ]:   Ptid Persistency_Flag  Gender           Race     Ethnicity   Region  \
     0   P1       Persistent    Male      Caucasian  Not Hispanic     West
     1   P2   Non-Persistent    Male          Asian  Not Hispanic     West
     2   P3   Non-Persistent  Female  Other/Unknown      Hispanic  Midwest
     3   P4   Non-Persistent  Female      Caucasian  Not Hispanic  Midwest
     4   P5   Non-Persistent  Female      Caucasian  Not Hispanic  Midwest

       Age_Bucket        Ntm_Speciality Ntm_Specialist_Flag  \
     0        >75  GENERAL PRACTITIONER              Others
     1      55-65  GENERAL PRACTITIONER              Others
     2      65-75  GENERAL PRACTITIONER              Others
     3        >75  GENERAL PRACTITIONER              Others
     4        >75  GENERAL PRACTITIONER              Others

            Ntm_Speciality_Bucket Gluco_Record_Prior_Ntm Gluco_Record_During_Rx  \
```

```
0  OB/GYN/Others/PCP/Unknown                    N                    N
1  OB/GYN/Others/PCP/Unknown                    N                    N
2  OB/GYN/Others/PCP/Unknown                    N                    N
3  OB/GYN/Others/PCP/Unknown                    N                    Y
4  OB/GYN/Others/PCP/Unknown                    Y                    Y

   Dexa_Freq_During_Rx Dexa_During_Rx Frag_Frac_Prior_Ntm Frag_Frac_During_Rx  \
0                    0              N                   N                   N
1                    0              N                   N                   N
2                    0              N                   N                   N
3                    0              N                   N                   N
4                    0              N                   N                   N

  Risk_Segment_Prior_Ntm Tscore_Bucket_Prior_Ntm Risk_Segment_During_Rx  \
0                 VLR_LR                   >-2.5                 VLR_LR
1                 VLR_LR                   >-2.5                Unknown
2                 HR_VHR                  <=-2.5                 HR_VHR
3                 HR_VHR                   >-2.5                 HR_VHR
4                 HR_VHR                  <=-2.5                Unknown

  Tscore_Bucket_During_Rx Change_T_Score Change_Risk_Segment Adherent_Flag  \
0                  <=-2.5      No change             Unknown      Adherent
1                 Unknown        Unknown             Unknown      Adherent
2                  <=-2.5      No change           No change      Adherent
3                  <=-2.5      No change           No change      Adherent
4                 Unknown        Unknown             Unknown      Adherent

  Idn_Indicator Injectable_Experience_During_Rx  \
0             N                               Y
1             N                               Y
2             N                               Y
3             N                               Y
4             N                               Y

  Comorb_Encounter_For_Screening_For_Malignant_Neoplasms  \
0                                                  N
1                                                  N
2                                                  Y
3                                                  N
4                                                  Y

  Comorb_Encounter_For_Immunization  \
0                                 Y
1                                 N
2                                 N
3                                 Y
4                                 Y
```

	Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx	Comorb_Vitamin_D_Deficiency	Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified	Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx	Comorb_Long_Term_Current_Drug_Therapy	Comorb_Dorsalgia	Comorb_Personal_History_Of_Other_Diseases_And_Conditions	Comorb_Other_Disorders_Of_Bone_Density_And_Structure
0	Y	N	N	Y	N	Y	Y	N
1	Y	N	N	N	N	N	N	N
2	Y	N	N	N	N	N	N	N
3	Y	N	Y	N	N	Y	N	N
4	Y	N	N	N	N	Y	Y	N

```
3                                                    N
4                                                    N

  Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias  \
0                                                    N
1                                                    N
2                                                    N
3                                                    Y
4                                                    N

  Comorb_Osteoporosis_without_current_pathological_fracture  \
0                                                    N
1                                                    N
2                                                    N
3                                                    N
4                                                    N

  Comorb_Personal_history_of_malignant_neoplasm  \
0                                            N
1                                            N
2                                            N
3                                            N
4                                            N

  Comorb_Gastro_esophageal_reflux_disease  \
0                                       N
1                                       N
2                                       N
3                                       Y
4                                       N

  Concom_Cholesterol_And_Triglyceride_Regulating_Preparations  \
0                                                    N
1                                                    N
2                                                    Y
3                                                    N
4                                                    N

  Concom_Narcotics Concom_Systemic_Corticosteroids_Plain  \
0                N                                     N
1                N                                     N
2                N                                     N
3                Y                                     Y
4                Y                                     Y

  Concom_Anti_Depressants_And_Mood_Stabilisers Concom_Fluoroquinolones  \
0                                           N                        N
```

```
1                                          N                        N
2                                          N                        N
3                                          N                        N
4                                          Y                        N

  Concom_Cephalosporins Concom_Macrolides_And_Similar_Types  \
0                     N                                    N
1                     N                                    N
2                     N                                    N
3                     N                                    N
4                     N                                    N

  Concom_Broad_Spectrum_Penicillins Concom_Anaesthetics_General  \
0                                 N                            N
1                                 N                            N
2                                 N                            N
3                                 N                            N
4                                 N                            N

  Concom_Viral_Vaccines Risk_Type_1_Insulin_Dependent_Diabetes  \
0                     N                                      N
1                     N                                      N
2                     N                                      N
3                     Y                                      N
4                     N                                      N

  Risk_Osteogenesis_Imperfecta Risk_Rheumatoid_Arthritis  \
0                            N                          N
1                            N                          N
2                            N                          N
3                            N                          N
4                            N                          N

  Risk_Untreated_Chronic_Hyperthyroidism Risk_Untreated_Chronic_Hypogonadism  \
0                                      N                                    N
1                                      N                                    N
2                                      N                                    N
3                                      N                                    N
4                                      N                                    N

  Risk_Untreated_Early_Menopause Risk_Patient_Parent_Fractured_Their_Hip  \
0                              N                                       N
1                              N                                       N
2                              N                                       Y
3                              N                                       N
4                              N                                       N
```

```
  Risk_Smoking_Tobacco Risk_Chronic_Malnutrition_Or_Malabsorption  \
0                    N                                          N
1                    N                                          N
2                    N                                          N
3                    Y                                          N
4                    Y                                          N

  Risk_Chronic_Liver_Disease Risk_Family_History_Of_Osteoporosis  \
0                          N                                   N
1                          N                                   N
2                          N                                   N
3                          N                                   N
4                          N                                   N

  Risk_Low_Calcium_Intake Risk_Vitamin_D_Insufficiency  \
0                       N                            N
1                       N                            N
2                       Y                            N
3                       N                            N
4                       N                            N

  Risk_Poor_Health_Frailty Risk_Excessive_Thinness  \
0                        N                       N
1                        N                       N
2                        N                       N
3                        N                       N
4                        N                       N

  Risk_Hysterectomy_Oophorectomy Risk_Estrogen_Deficiency Risk_Immobilization  \
0                              N                        N                    N
1                              N                        N                    N
2                              N                        N                    N
3                              N                        N                    N
4                              N                        N                    N

  Risk_Recurring_Falls  Count_Of_Risks
0                    N               0
1                    N               0
2                    N               2
3                    N               1
4                    N               1
```

[ ]:
```
df.describe()
```

[ ]:
```
       Dexa_Freq_During_Rx  Count_Of_Risks
count          3424.000000     3424.000000
mean              3.016063        1.239486
```

```
std              8.136545        1.094914
min              0.000000        0.000000
25%              0.000000        0.000000
50%              0.000000        1.000000
75%              3.000000        2.000000
max            146.000000        7.000000
```

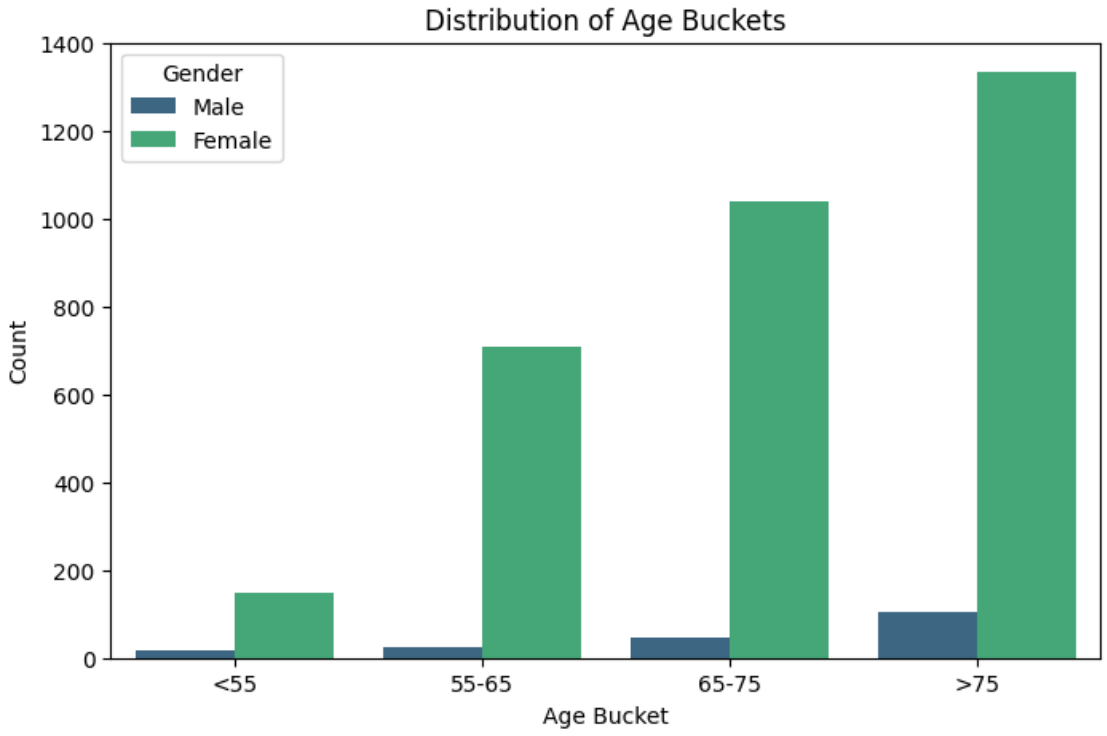
1.13 Visualize the Distribution of Patient Demographics

1.13.1 Age/Gender

[ ]:
```
age_bucket_counts = df["Age_Bucket"].value_counts().sort_values().index

plt.figure(figsize=(8, 5))
sns.countplot(
    x="Age_Bucket",
    data=df,
    order=age_bucket_counts,
    palette="viridis",
    hue="Gender",
)

plt.title("Distribution of Age Buckets")
plt.xlabel("Age Bucket")
plt.ylabel("Count")
plt.show()
```



[ ]:
```
df.Gender.value_counts(), df.Age_Bucket.value_counts()
```

[ ]:
```
(Gender
 Female    3230
 Male       194
 Name: count, dtype: int64,
 Age_Bucket
 >75      1439
 65-75    1086
 55-65     733
 <55       166
 Name: count, dtype: int64)
```

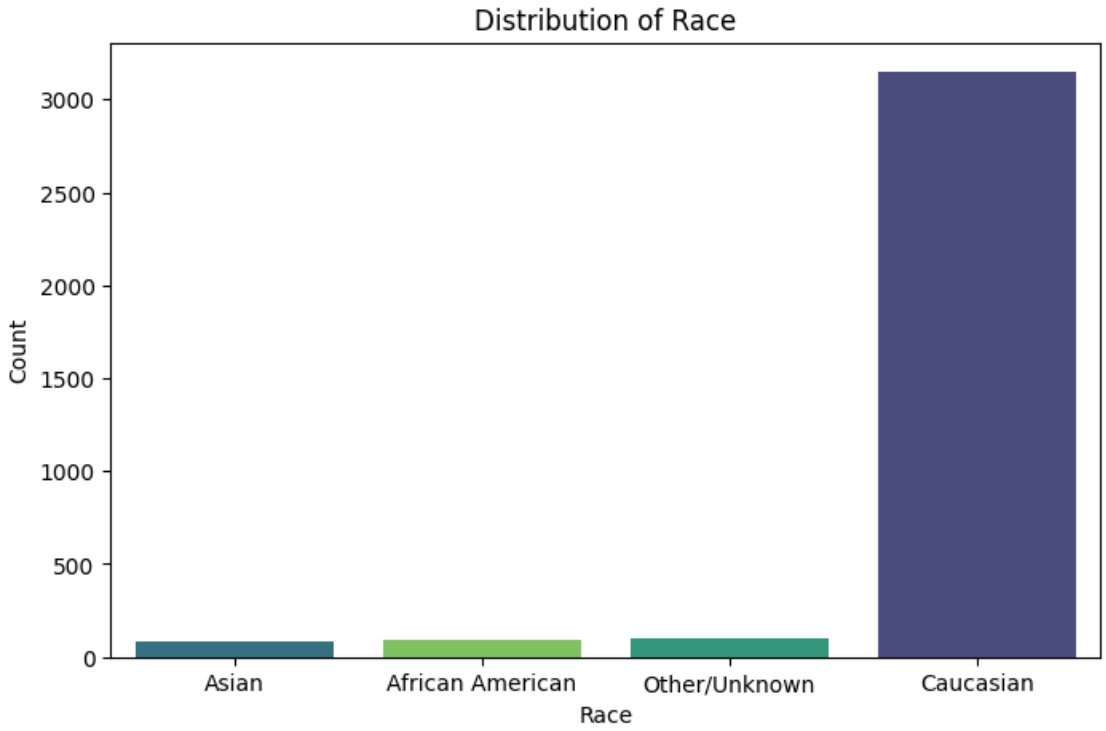
For age, it is important to note that this dataset is largely made up of 65+ patients and will not represent younger patients as well in future analysis and modeling. This will create a bias for older patients and will require Additionally, it will be worthwhile to see if different age buckets require different approaches.

For gender, it is important to note that data is 94% females. This will also create a heavy bias toward female analysis and predictions if left unchecked.

1.13.2 Race

[ ]:
```
race_bucket_counts = df["Race"].value_counts().sort_values().index

plt.figure(figsize=(8, 5))
sns.countplot(
    x="Race", data=df, palette="viridis", hue="Race", order=race_bucket_counts
)
plt.title("Distribution of Race")
plt.xlabel("Race")
plt.ylabel("Count")
plt.show()
```



[ ]:
```
df.Race.value_counts()
```

[ ]:
```
Race
Caucasian           3148
Other/Unknown         97
African American      95
Asian                 84
Name: count, dtype: int64
```

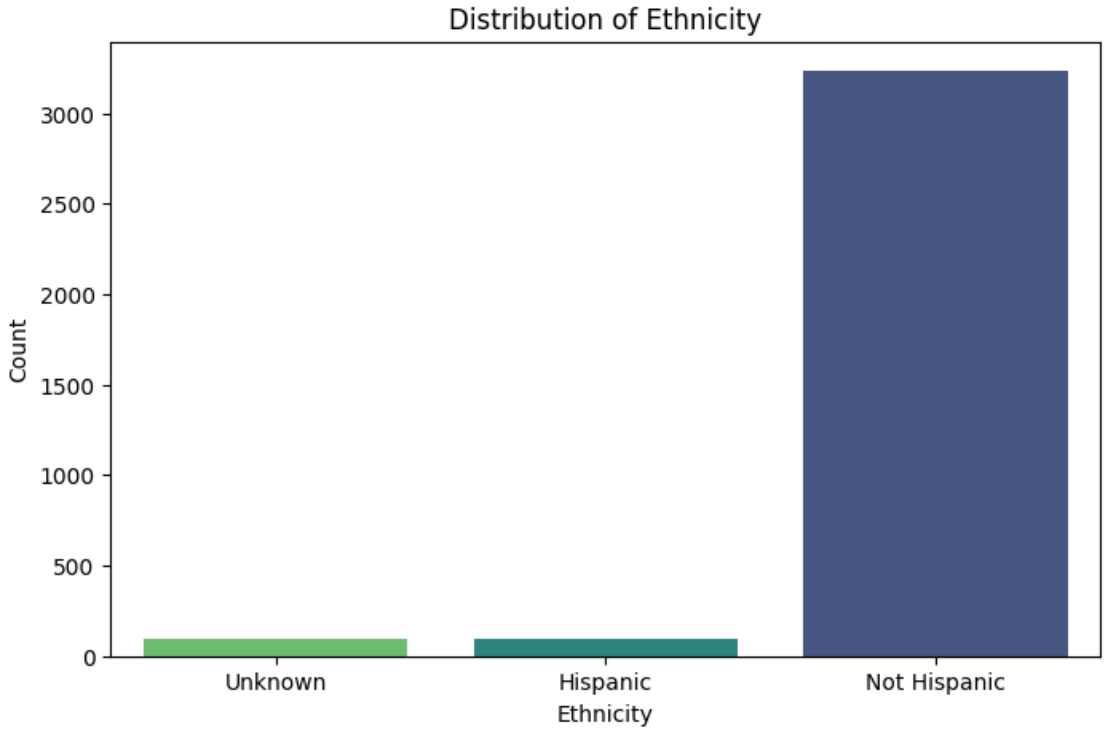
Race is highly biased toward caucasian patients. This can lead to future analysis and models only working well for caucasian patients. Oversampling other races or getting more data is suggested.

1.13.3 Ethnicity

[ ]:
```
ethnicity_counts = df["Ethnicity"].value_counts().sort_values().index

plt.figure(figsize=(8, 5))
sns.countplot(
    x="Ethnicity", data=df, palette="viridis", hue="Ethnicity",␣
 ↪order=ethnicity_counts
)
plt.title("Distribution of Ethnicity")
plt.xlabel("Ethnicity")
plt.ylabel("Count")
```

```
plt.show()
```



[ ]:
```
df.Ethnicity.value_counts()
```

[ ]:
```
Ethnicity
Not Hispanic    3235
Hispanic          98
Unknown           91
Name: count, dtype: int64
```

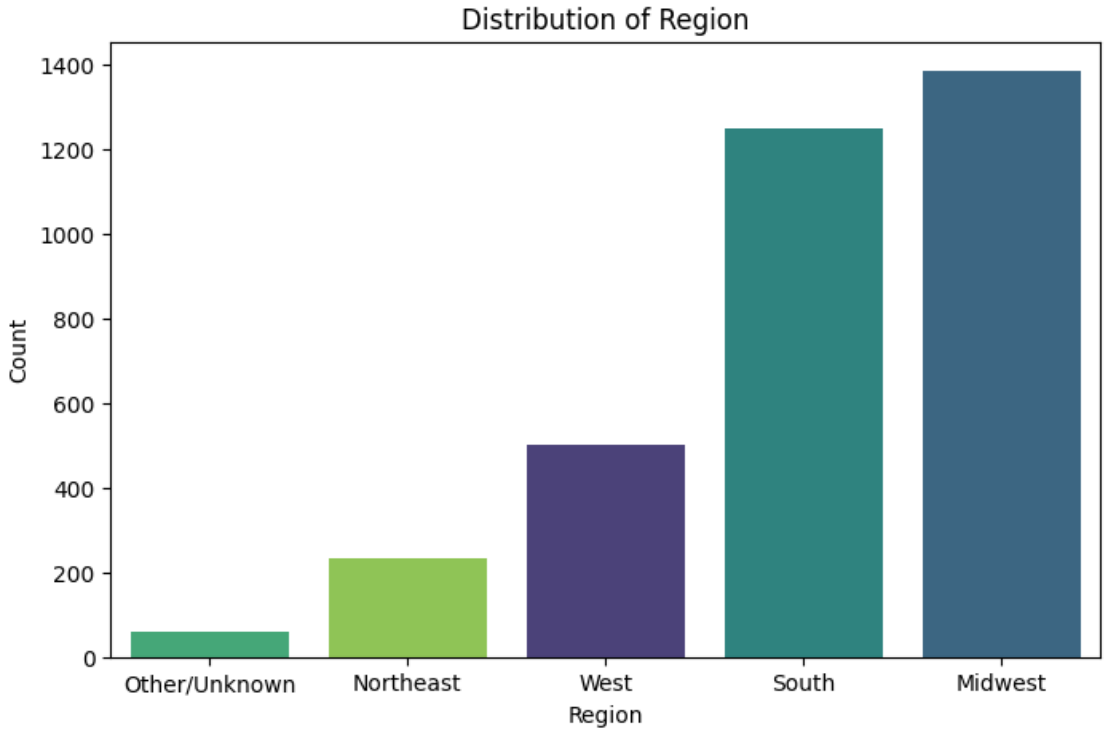
As is most demographics in this dataset, there is a heavy bias toward a large group, which in this case is non-hispanic patients. This data would benefit from a more balanced split in addition to more categories split from the non-hispanic bin.

1.13.4 Region

[ ]:
```
region_counts = df["Region"].value_counts().sort_values().index

plt.figure(figsize=(8, 5))
sns.countplot(x="Region", data=df, palette="viridis", hue="Region",␣
  ↪order=region_counts)
plt.title("Distribution of Region")
plt.xlabel("Region")
```

```
plt.ylabel("Count")
plt.show()
```



[ ]:
```
df.Region.value_counts()
```

[ ]:
```
Region
Midwest          1383
South            1247
West              502
Northeast         232
Other/Unknown      60
Name: count, dtype: int64
```

Region's skew is not as significant as the other categories, but does favor the south and midwest. The northeast and west region have enough data that oversampling would work, but the other/unknown region could pose a problem for accuracy.

1.14 Factorize Columns for Correlation Analysis with Mutual Information

[ ]:
```
categorical_cols, numerical_cols = cols_info.get_cat_num_cols(df)


df_factorized, factorized_cols, mappings = data_preprocessing.factorize_columns(
```

```
    df, categorical_cols
)
```

```
Ptid factorized.
Persistency_Flag factorized.
Gender factorized.
Race factorized.
Ethnicity factorized.
Region factorized.
Age_Bucket factorized.
Ntm_Speciality factorized.
Ntm_Specialist_Flag factorized.
Ntm_Speciality_Bucket factorized.
Gluco_Record_Prior_Ntm factorized.
Gluco_Record_During_Rx factorized.
Dexa_During_Rx factorized.
Frag_Frac_Prior_Ntm factorized.
Frag_Frac_During_Rx factorized.
Risk_Segment_Prior_Ntm factorized.
Tscore_Bucket_Prior_Ntm factorized.
Risk_Segment_During_Rx factorized.
Tscore_Bucket_During_Rx factorized.
Change_T_Score factorized.
Change_Risk_Segment factorized.
Adherent_Flag factorized.
Idn_Indicator factorized.
Injectable_Experience_During_Rx factorized.
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms factorized.
Comorb_Encounter_For_Immunization factorized.
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx factorized.
Comorb_Vitamin_D_Deficiency factorized.
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified factorized.
Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx factorized.
Comorb_Long_Term_Current_Drug_Therapy factorized.
Comorb_Dorsalgia factorized.
Comorb_Personal_History_Of_Other_Diseases_And_Conditions factorized.
Comorb_Other_Disorders_Of_Bone_Density_And_Structure factorized.
Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias factorized.
Comorb_Osteoporosis_without_current_pathological_fracture factorized.
Comorb_Personal_history_of_malignant_neoplasm factorized.
Comorb_Gastro_esophageal_reflux_disease factorized.
Concom_Cholesterol_And_Triglyceride_Regulating_Preparations factorized.
Concom_Narcotics factorized.
Concom_Systemic_Corticosteroids_Plain factorized.
Concom_Anti_Depressants_And_Mood_Stabilisers factorized.
Concom_Fluoroquinolones factorized.
Concom_Cephalosporins factorized.
Concom_Macrolides_And_Similar_Types factorized.
```

```
Concom_Broad_Spectrum_Penicillins factorized.
Concom_Anaesthetics_General factorized.
Concom_Viral_Vaccines factorized.
Risk_Type_1_Insulin_Dependent_Diabetes factorized.
Risk_Osteogenesis_Imperfecta factorized.
Risk_Rheumatoid_Arthritis factorized.
Risk_Untreated_Chronic_Hyperthyroidism factorized.
Risk_Untreated_Chronic_Hypogonadism factorized.
Risk_Untreated_Early_Menopause factorized.
Risk_Patient_Parent_Fractured_Their_Hip factorized.
Risk_Smoking_Tobacco factorized.
Risk_Chronic_Malnutrition_Or_Malabsorption factorized.
Risk_Chronic_Liver_Disease factorized.
Risk_Family_History_Of_Osteoporosis factorized.
Risk_Low_Calcium_Intake factorized.
Risk_Vitamin_D_Insufficiency factorized.
Risk_Poor_Health_Frailty factorized.
Risk_Excessive_Thinness factorized.
Risk_Hysterectomy_Oophorectomy factorized.
Risk_Estrogen_Deficiency factorized.
Risk_Immobilization factorized.
Risk_Recurring_Falls factorized.
```

1.15 Split Data Into Features and Target

[ ]:
```
X = df_factorized.copy()
y = X.pop(config_data["target_column"])
```

1.16 Get MI Scores

[ ]:
```
# get object mask for dataframe
objects_mask = cols_info.get_object_bool_mask(X)
mi_scores = metrics.get_mi_scores(X, y, objects_mask, ignore_cols=["Ptid"])

mi_scores
```

[ ]:
```
Dexa_Freq_During_Rx                                                  0.142064
Dexa_During_Rx                                                       0.109235
Comorb_Long_Term_Current_Drug_Therapy                                0.075847
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms               0.068692
Comorb_Personal_History_Of_Other_Diseases_And_Conditions             0.048574
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx      0.043676
Risk_Low_Calcium_Intake                                              0.037021
Ntm_Speciality                                                       0.035291
Concom_Macrolides_And_Similar_Types                                  0.033113
Comorb_Encounter_For_Immunization                                    0.031805
Concom_Viral_Vaccines                                                0.031699
```

Concom_Systemic_Corticosteroids_Plain	0.030358
Concom_Anaesthetics_General	0.029804
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified	0.029652
Concom_Cephalosporins	0.029257
Comorb_Other_Disorders_Of_Bone_Density_And_Structure	0.025725
Ntm_Specialist_Flag	0.025709
Concom_Narcotics	0.023551
Injectable_Experience_During_Rx	0.023186
Risk_Type_1_Insulin_Dependent_Diabetes	0.021326
Concom_Anti_Depressants_And_Mood_Stabilisers	0.020508
Idn_Indicator	0.016910
Concom_Broad_Spectrum_Penicillins	0.016608
Comorb_Vitamin_D_Deficiency	0.016450
Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias	0.016359
Risk_Segment_During_Rx	0.016156
Adherent_Flag	0.015178
Concom_Fluoroquinolones	0.014648
Change_Risk_Segment	0.013325
Risk_Hysterectomy_Oophorectomy	0.013081
Gender	0.012063
Change_T_Score	0.011748
Risk_Osteogenesis_Imperfecta	0.010927
Concom_Cholesterol_And_Triglyceride_Regulating_Preparations	0.009127
Tscore_Bucket_During_Rx	0.008598
Risk_Excessive_Thinness	0.007870
Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx	0.007592
Risk_Untreated_Chronic_Hypogonadism	0.005942
Frag_Frac_During_Rx	0.005734
Risk_Smoking_Tobacco	0.004807
Comorb_Gastro_esophageal_reflux_disease	0.004778
Gluco_Record_During_Rx	0.004519
Risk_Vitamin_D_Insufficiency	0.004323
Count_Of_Risks	0.000221
Risk_Poor_Health_Frailty	0.000000
Ethnicity	0.000000
Region	0.000000
Risk_Family_History_Of_Osteoporosis	0.000000
Risk_Estrogen_Deficiency	0.000000
Risk_Immobilization	0.000000
Risk_Chronic_Liver_Disease	0.000000
Risk_Recurring_Falls	0.000000
Risk_Chronic_Malnutrition_Or_Malabsorption	0.000000
Comorb_Dorsalgia	0.000000
Risk_Patient_Parent_Fractured_Their_Hip	0.000000
Risk_Untreated_Early_Menopause	0.000000
Risk_Untreated_Chronic_Hyperthyroidism	0.000000
Age_Bucket	0.000000

```
Ntm_Speciality_Bucket                                                    0.000000
Gluco_Record_Prior_Ntm                                                   0.000000
Frag_Frac_Prior_Ntm                                                      0.000000
Risk_Segment_Prior_Ntm                                                   0.000000
Tscore_Bucket_Prior_Ntm                                                  0.000000
Comorb_Personal_history_of_malignant_neoplasm                            0.000000
Comorb_Osteoporosis_without_current_pathological_fracture                0.000000
Race                                                                     0.000000
Risk_Rheumatoid_Arthritis                                                0.000000
Name: MI Scores, dtype: float64
```

1.17 Remove Extremely Low MI Columns

[ ]:
```
columns_to_keep = mi_scores[mi_scores > 0.02].index

X_filtered = X[columns_to_keep]
```

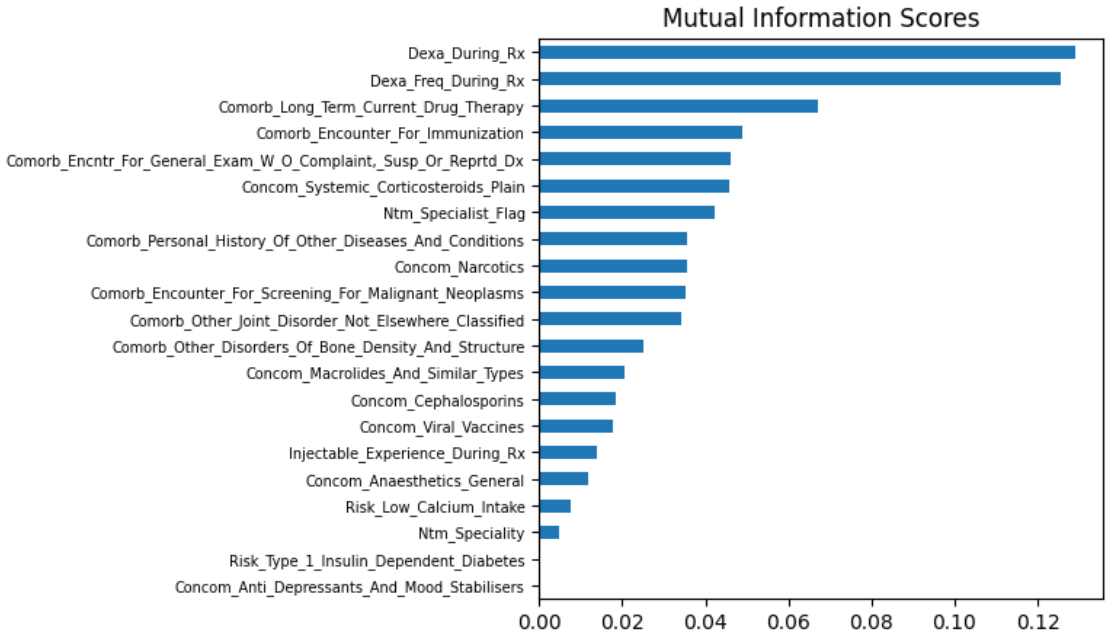
[ ]:
```
objects_mask = cols_info.get_object_bool_mask(X_filtered)
mi_scores = metrics.get_mi_scores(X_filtered, y, objects_mask,␣
 ↪ignore_cols=["Ptid"])

mi_scores
```

[ ]:
```
Dexa_During_Rx                                                           0.129214
Dexa_Freq_During_Rx                                                      0.125479
Comorb_Long_Term_Current_Drug_Therapy                                    0.066954
Comorb_Encounter_For_Immunization                                        0.048722
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx          0.046158
Concom_Systemic_Corticosteroids_Plain                                    0.045838
Ntm_Specialist_Flag                                                      0.042089
Comorb_Personal_History_Of_Other_Diseases_And_Conditions                 0.035458
Concom_Narcotics                                                         0.035417
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms                   0.035234
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified                     0.034111
Comorb_Other_Disorders_Of_Bone_Density_And_Structure                     0.024942
Concom_Macrolides_And_Similar_Types                                      0.020578
Concom_Cephalosporins                                                    0.018579
Concom_Viral_Vaccines                                                    0.017769
Injectable_Experience_During_Rx                                          0.013941
Concom_Anaesthetics_General                                              0.011914
Risk_Low_Calcium_Intake                                                  0.007699
Ntm_Speciality                                                           0.004666
Risk_Type_1_Insulin_Dependent_Diabetes                                   0.000000
Concom_Anti_Depressants_And_Mood_Stabilisers                             0.000000
Name: MI Scores, dtype: float64
```

1.18 Visualize Important Univariate Features

```
[ ]: plt.figure(figsize=(5, 5))
mi_scores.sort_values(ascending=True).plot(kind="barh")
plt.title("Mutual Information Scores")
plt.gca().tick_params(axis="y", labelsize=7)
plt.show()
```



These are the top features we will focus on when moving on to the modeling phase. The MI scores show how much a column accounts for the variability in the target feature, which is Persistency_Flag in this case.

1.19 Final Thoughts and Recommendations

Limitations - This dataset is primarily 65+ caucasian (non-hispanic) females from the South and Midwest. Because of this, future models and statistical analysis based on this data will favor this demographic and will likely not be accurate with the general population. - Dexa_Freq_During_Rx is a column that was initially shown to have many outliers. As this is the second most correlated column to Persistency_Flag, the accuracy of this column would need to be confirmed with stakeholders before finalizing conclusions. - There are 166 male, 276 non-caucasian, 189 non-hispanic, and 60 unknown location patients. Relative to the 3424 patients, these are very small categories and could cause inaccurate results when oversampling.

Recommendations - It is recommended to get a more inclusive sample so that the model can be trusted for general use. This could come from improved sampling methods or just more samples overall. - Dexa_During_Rx and Dexa_Freq_During_Rx both have high correlation to the Persistency_Flag and should be analyzed further once their reliability is confirmed.

Model Recommendations - As no more data will be provided, oversampling the underrepresented demographics is recommended for making a single-general model. - A stratified model could also help with the under representation of certain demographics. - If more data and diverse data were provided, building multiple models for different demographic groups could yield highly-tailored recommendations.