
# Exploratory Data Analysis

## HEALTH CARE – PERSISTENCY OF A DRUG

**Data Science Project**

LISUM07

**25/04/2021**

# Background

One of the challenge for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

# Problem Statement

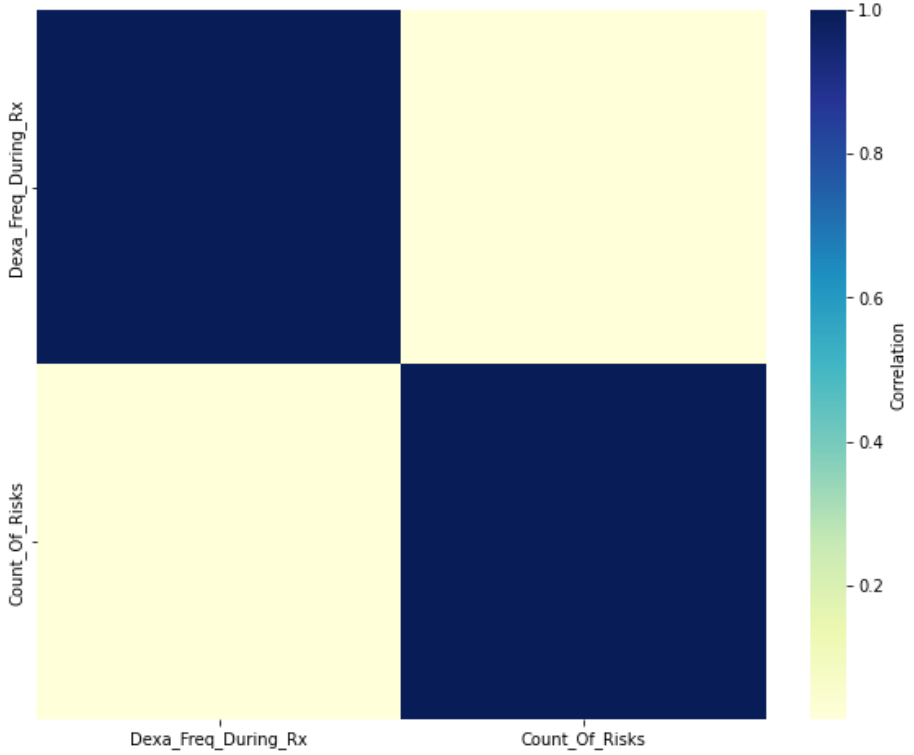
# Problem Statement

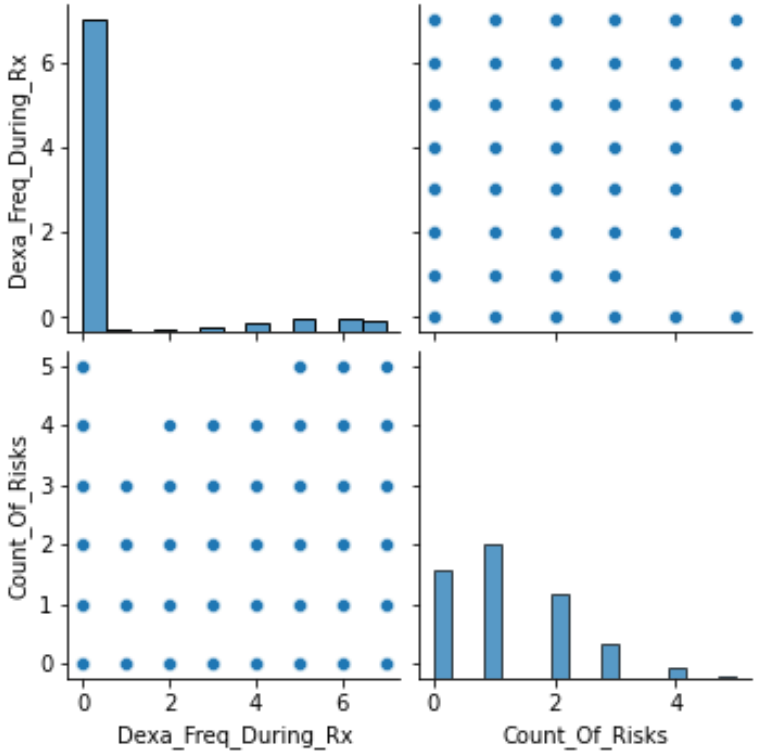
- Classification the process of identification of the persistency of the drug as per prescription provided by the physician.
- Understanding the persistency is an issue for pharmaceutical companies and so, ABC pharma company has approached XYZ analytics company to provide insights into the same.
- The role of XYZ analytics company is to undertake this project and provide a detailed understanding regarding the drug persistency.

# EDA

# Numerical Data Analysis

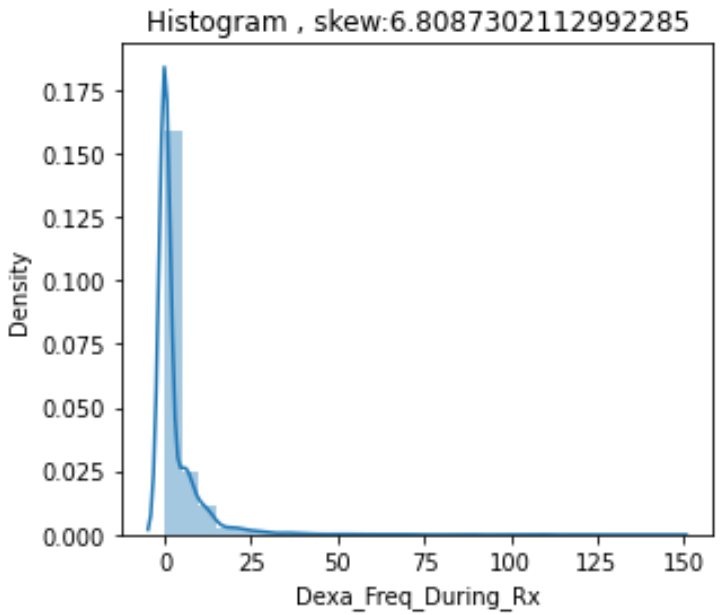
# Correlation

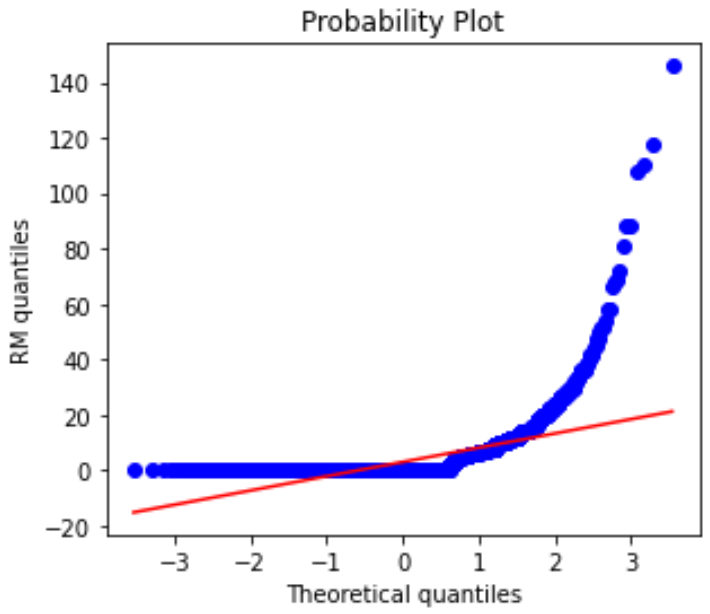


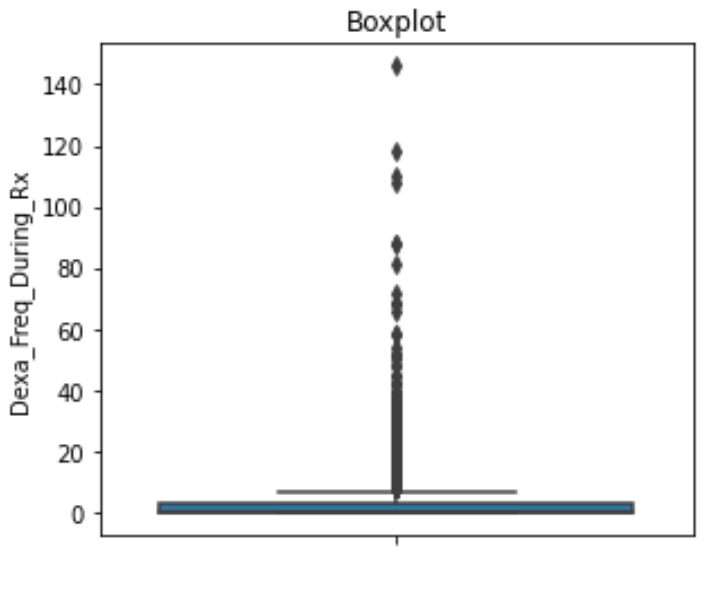




- The correlation values indicate that there is a low correlation between the two variables.
- On the other hand, Dexa_Freq_During_Rx has the least distributed values.

# Dexa_Freq_During_Rx

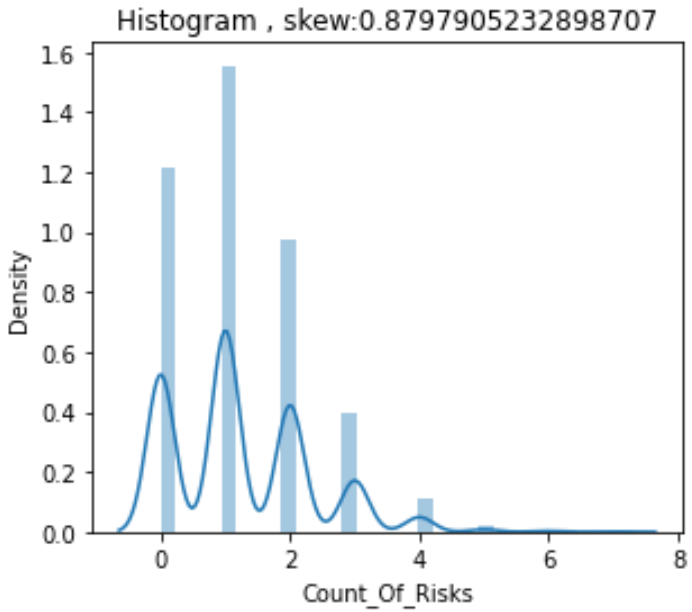


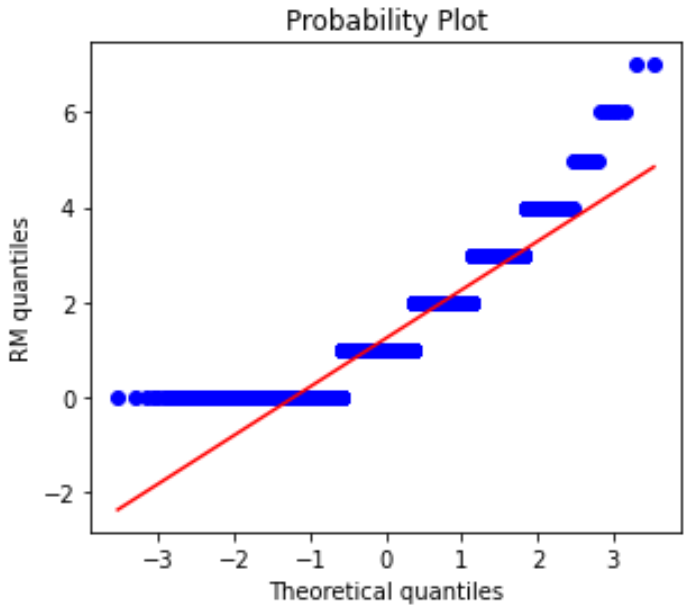



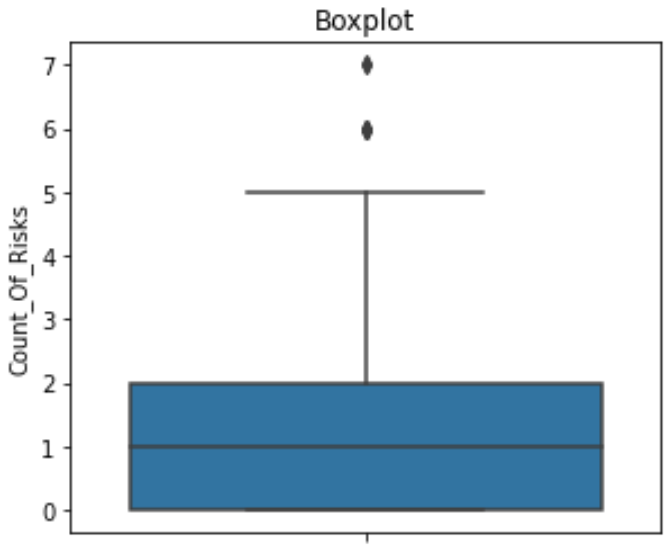




- From the diagrams above, we can see that most of the frequencies lie between 0 and 20.
- The minimum frequency is 0 and the maximum is around 140.
- The data is highly skewed.
- It has many data outliers, with a mean close to zero.

# Count_of_Risks



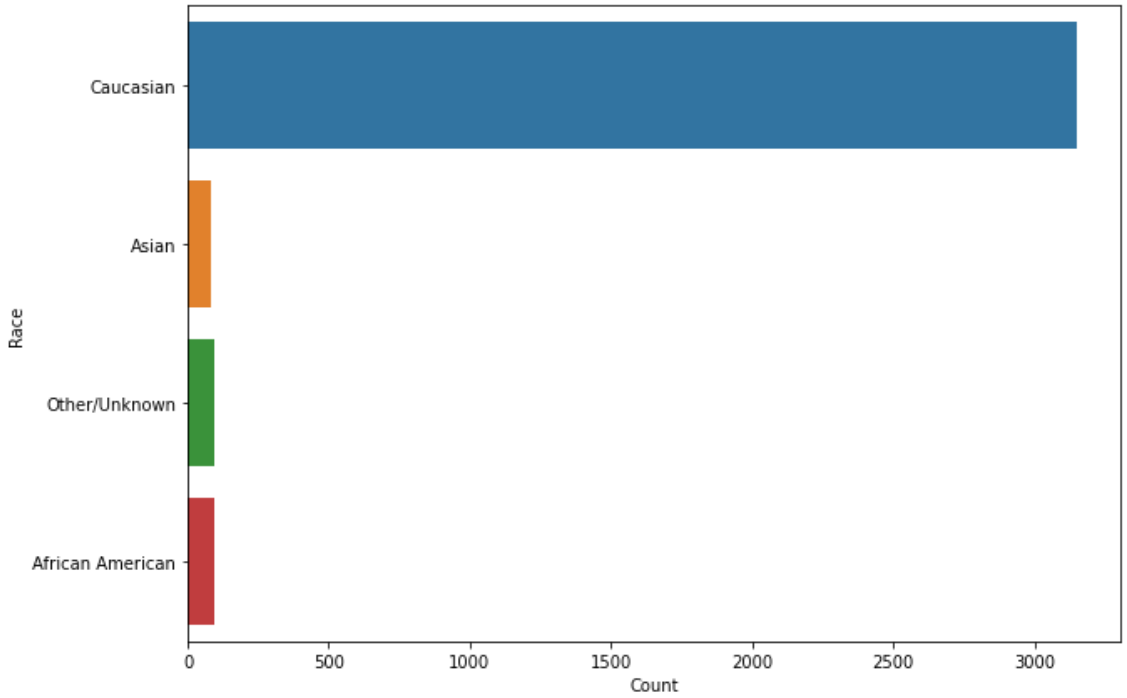



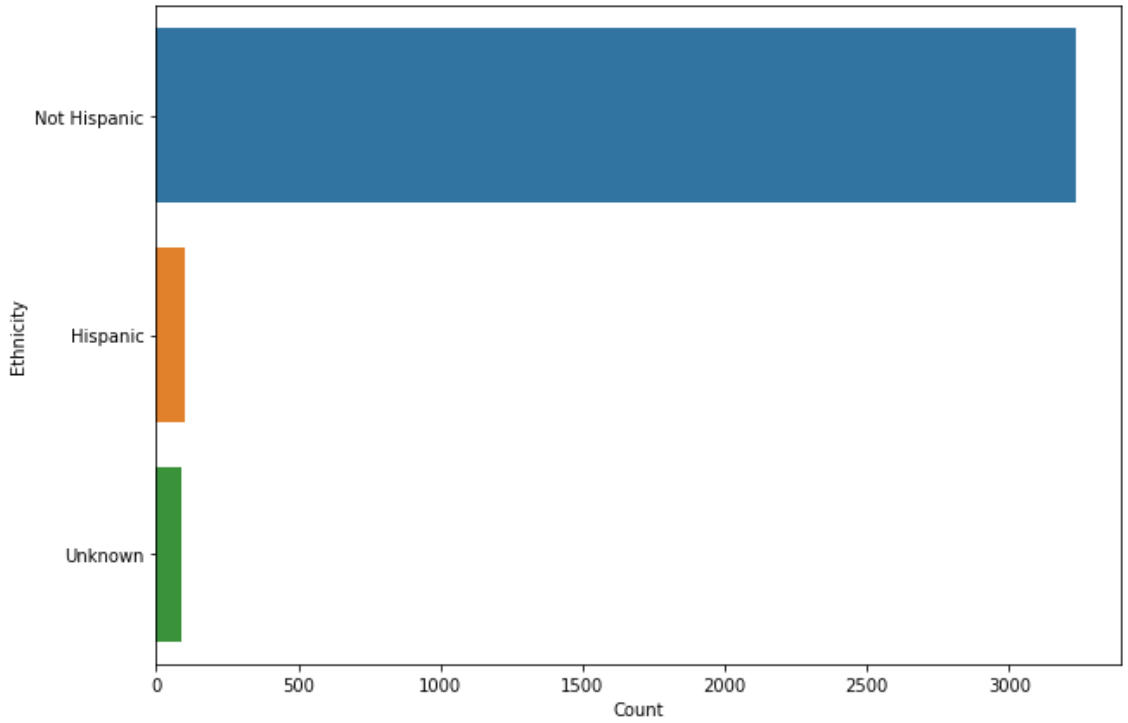




- Most of the count of risks lie between 0 and 1.
- The data is slightly skewed.
- There's a slight difference in the distribution of the count of risks between persistent patient's and non-persistent patients.
- It has a quite data outliers, with a mean close to one.

# Categorical Data Analysis

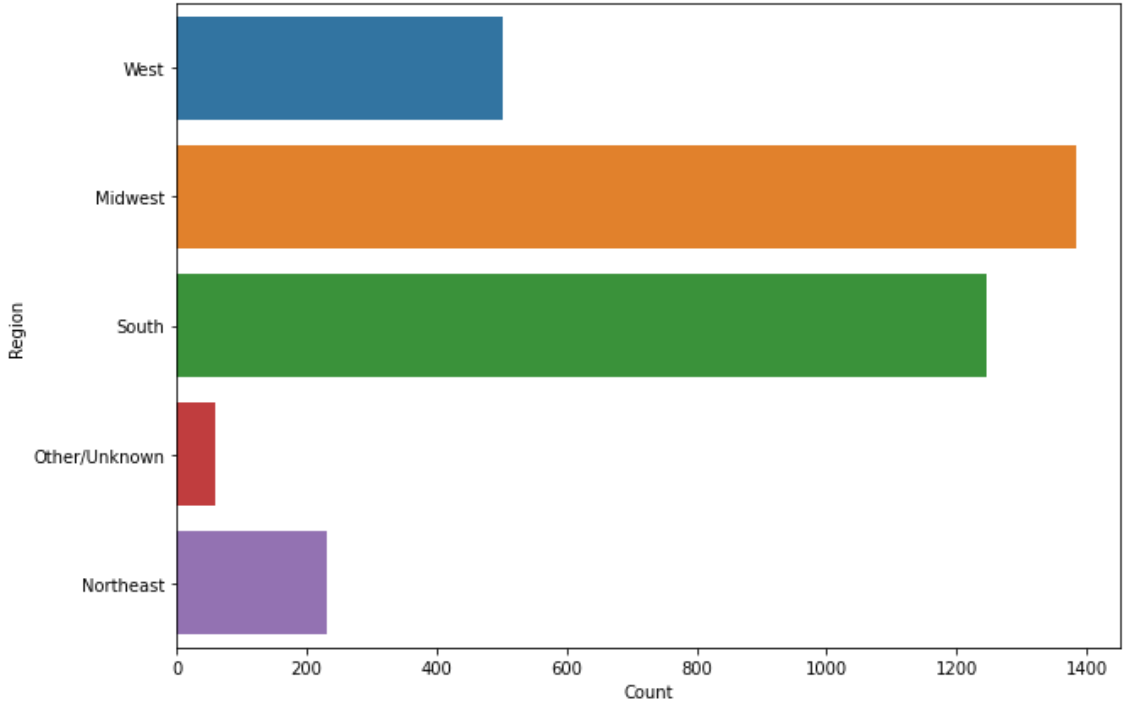
# Race/Ethnicity

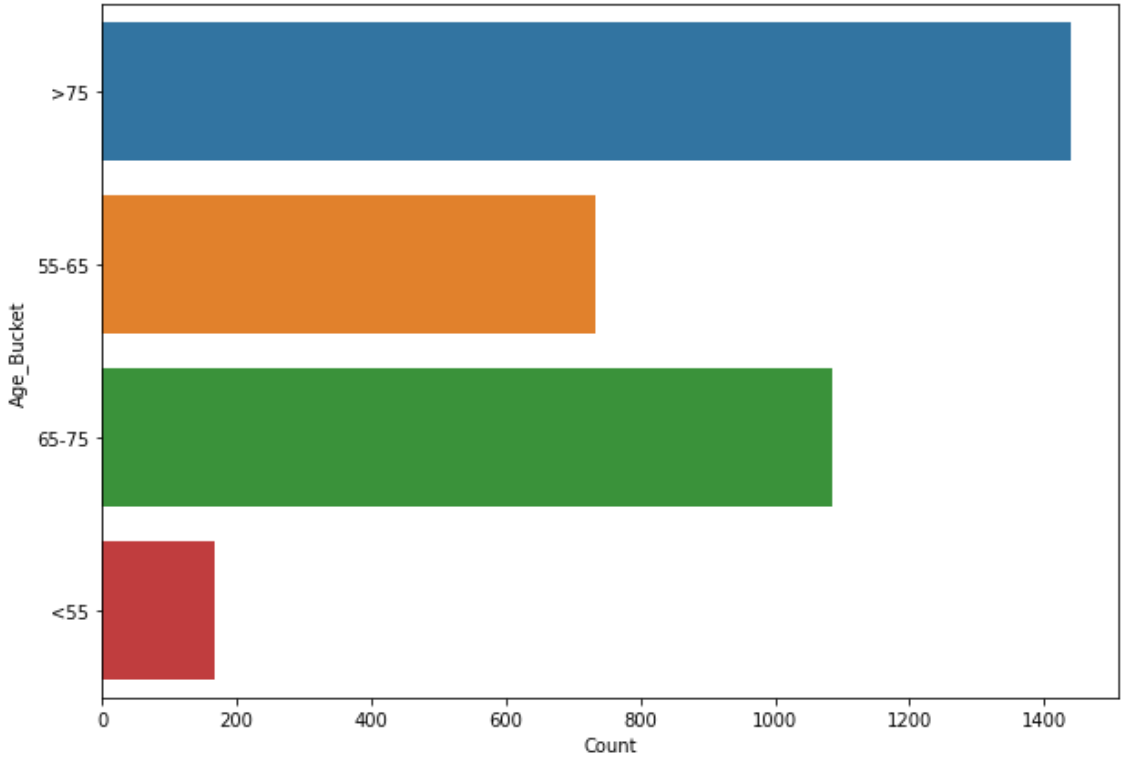



- RACE unbalanced data, greater predominance in Caucasian.
- Unbalanced data ETHNICITY, greater predominance in Not Hispanic.

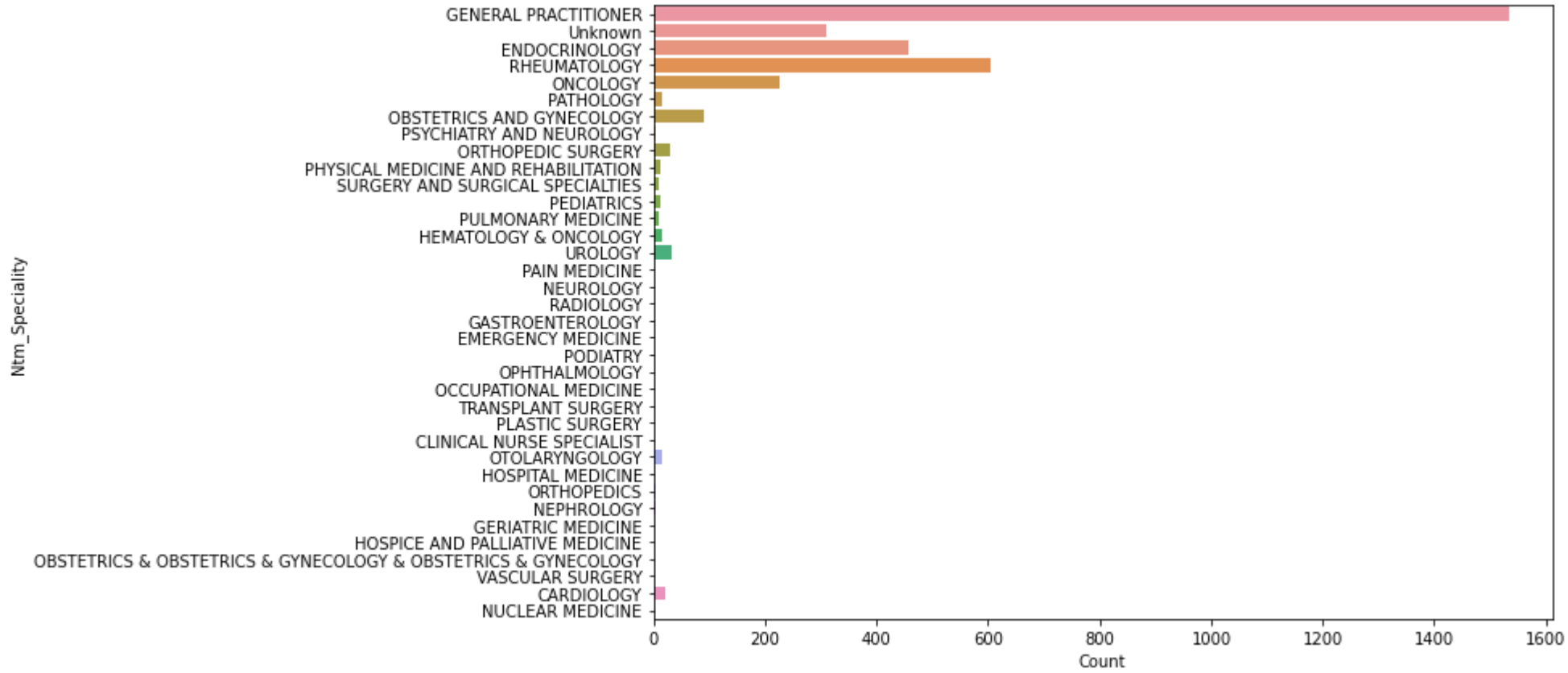
# Region/Age_Bucket



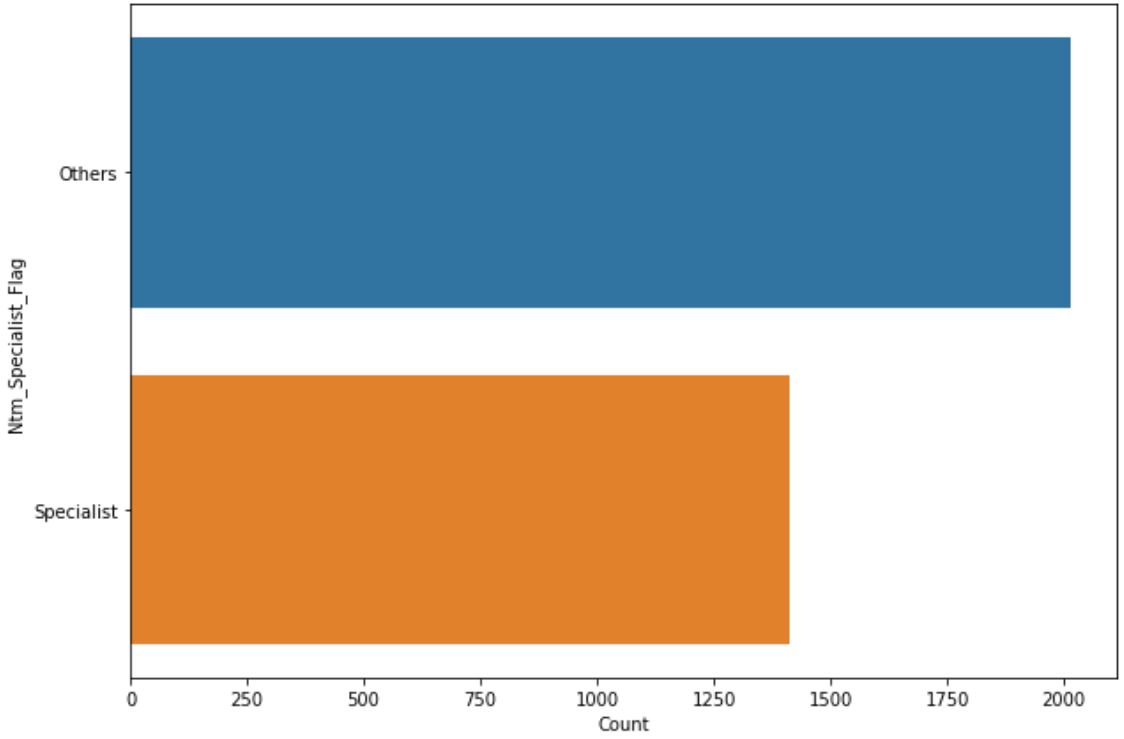

- REGION unbalanced data, greater predominance between Midwet and South.
- Unbalanced data AGE_BUSCKET, greater predominance between >75 & 64-75.

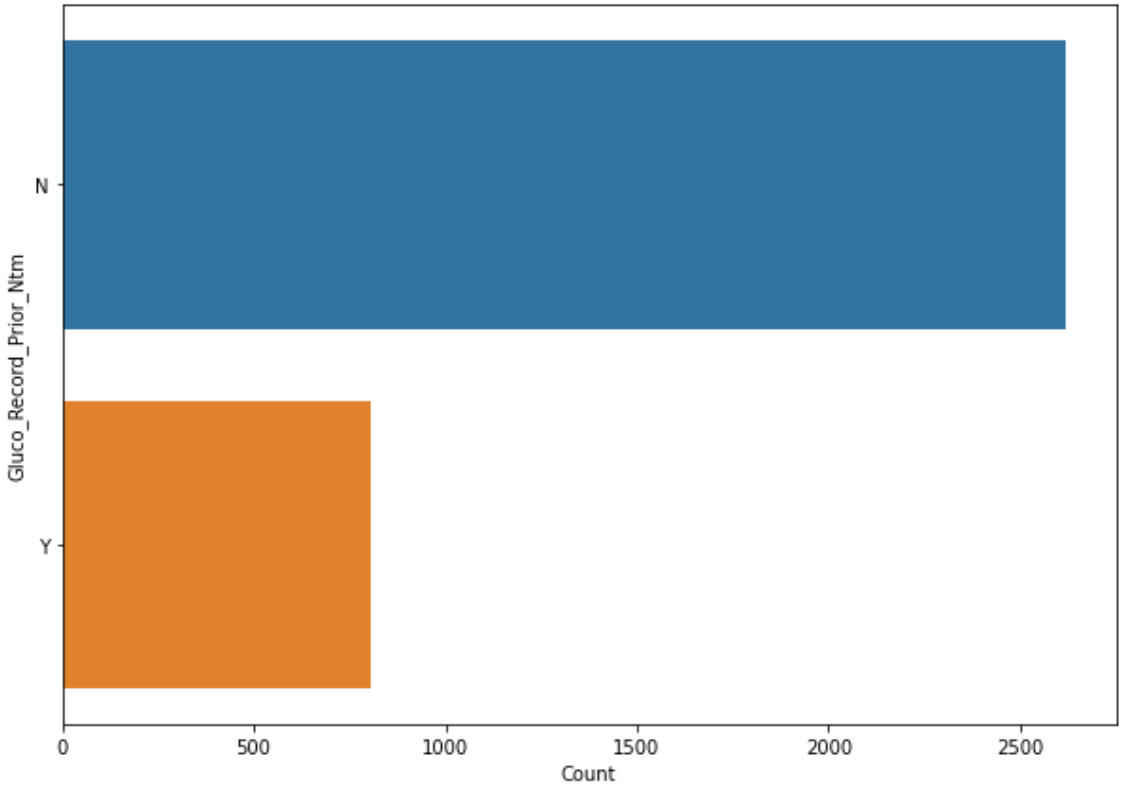
# Num_Speciality




- NUM_SPECIALITY unbalanced data, greater predominance in GENERAL PRACTICTIONER.

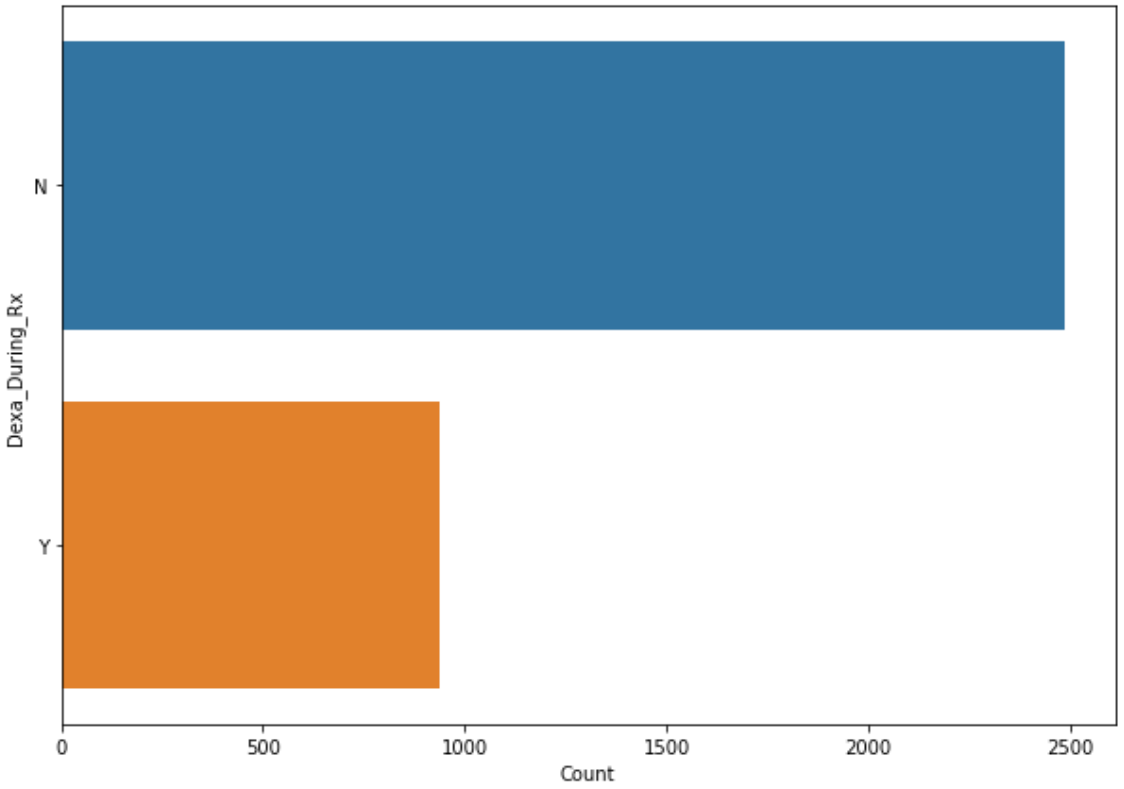
# Num_Speciality_Flag/Gluco_Record_Prior_Num

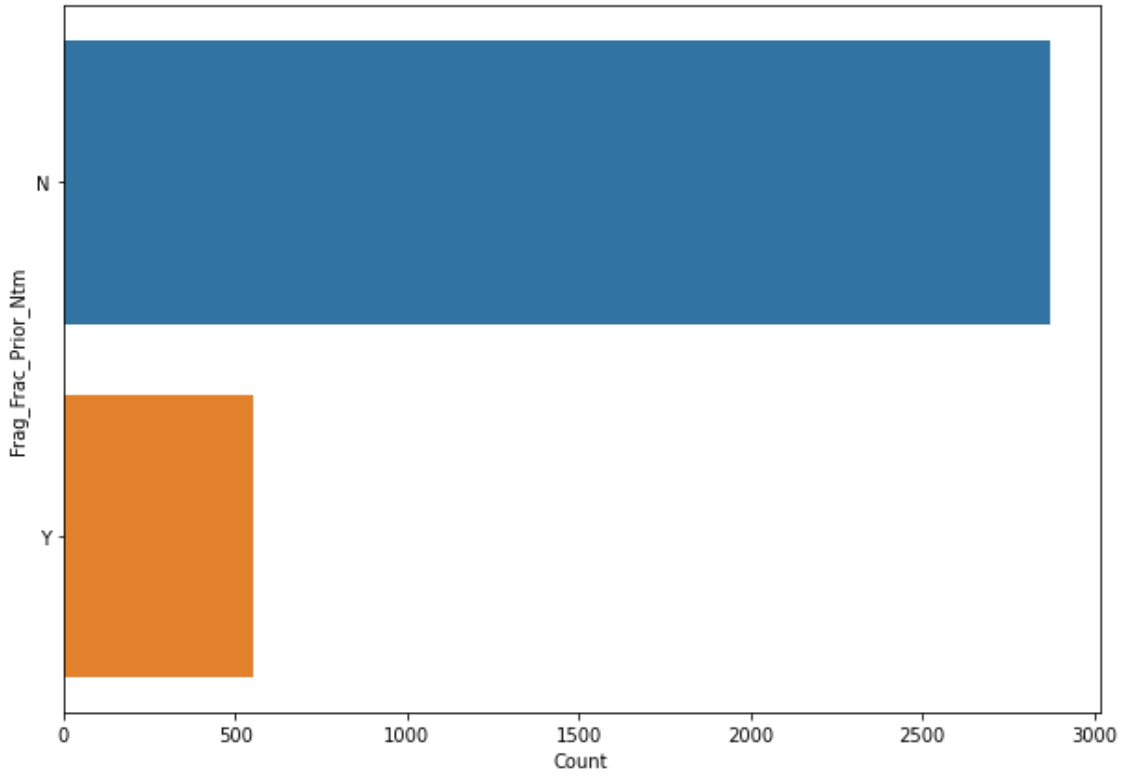



- NUM_SPECIALIST_FLAG balanced data.
- Unbalanced data GLUCO_RECORD_PROIR_NUM, greater predominance in N.

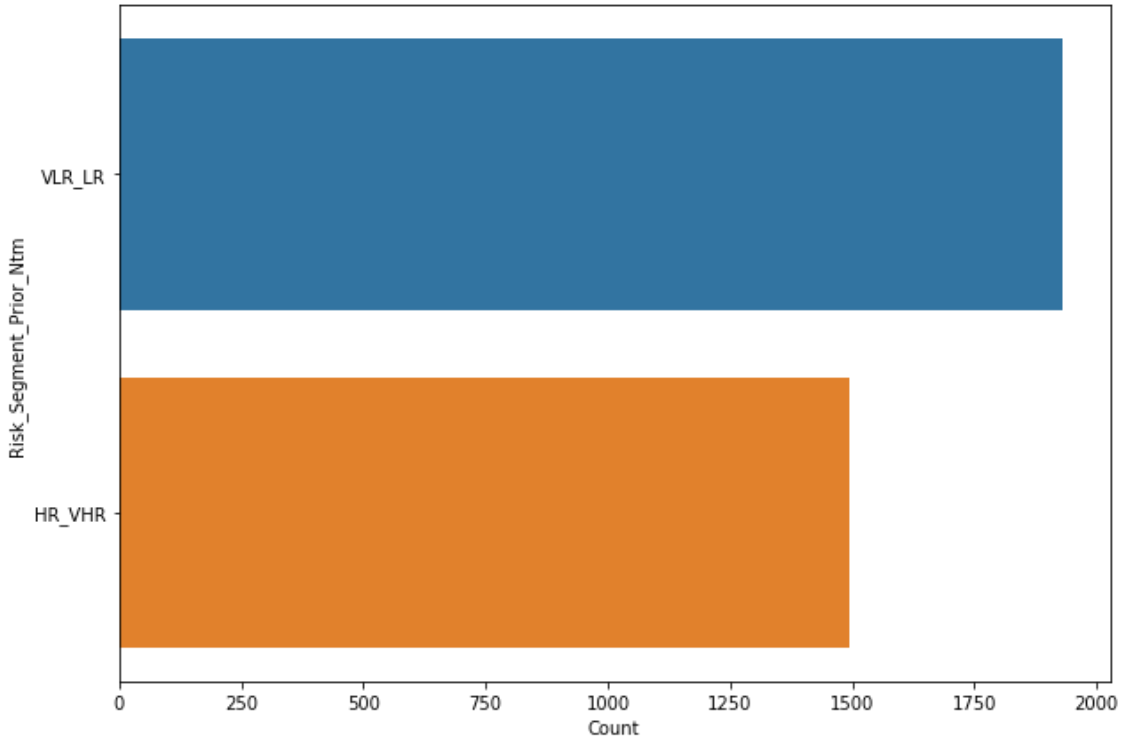
# Dexa_During_Rx/Frag_Frac_Prior_Num

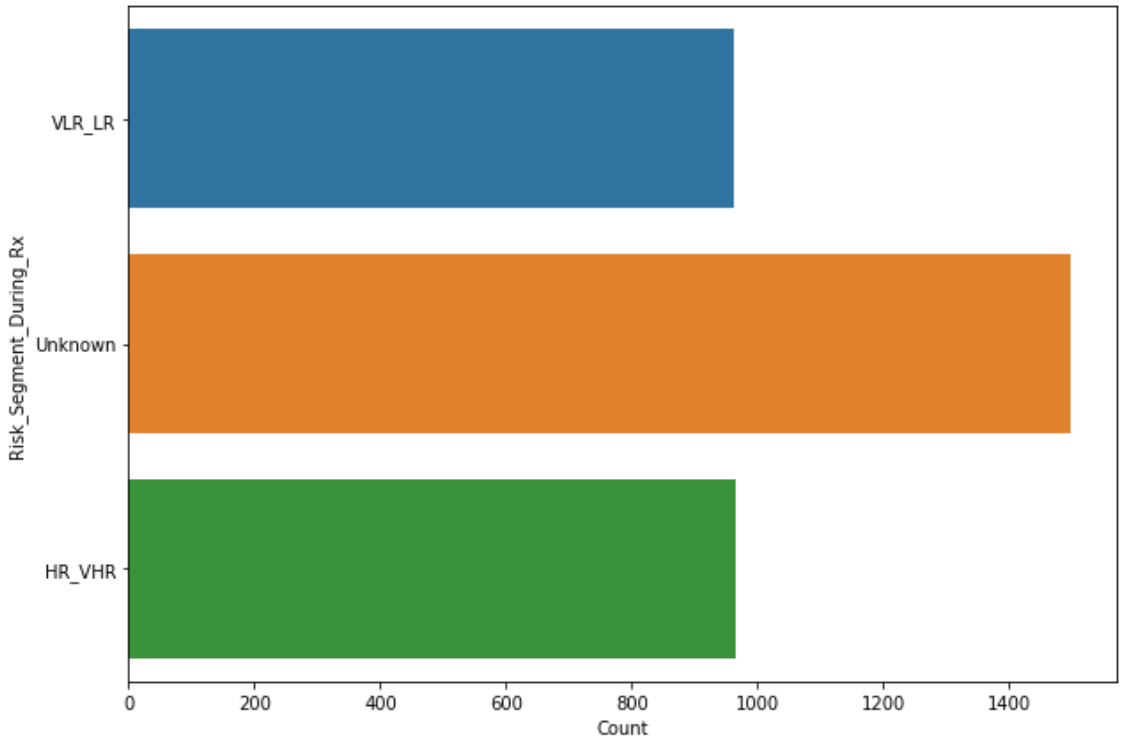



- DEXA_DURING_RX unbalanced data, greater predominance between N.
- Unbalanced data FRAG_FRAC_PRIOR_NUM, greater predominance in N.

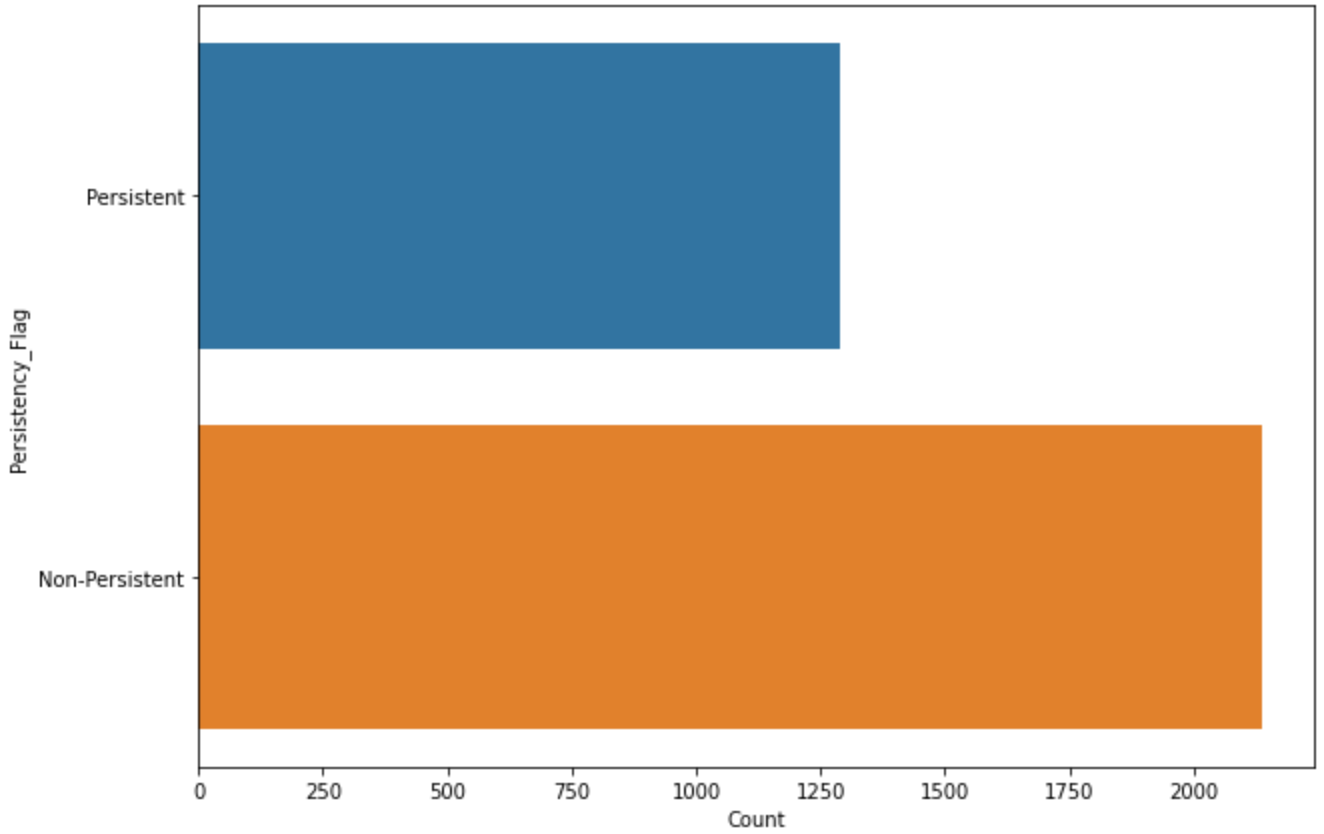
# Risk_Segment_Prior_Num/Risk_Segment_During_Rx




- RISK_SEGMENT_PRIOR_NUM balanced data.
- Balanced data RISK_SEGMENT_DURING_RX.

# Persistency_Flag



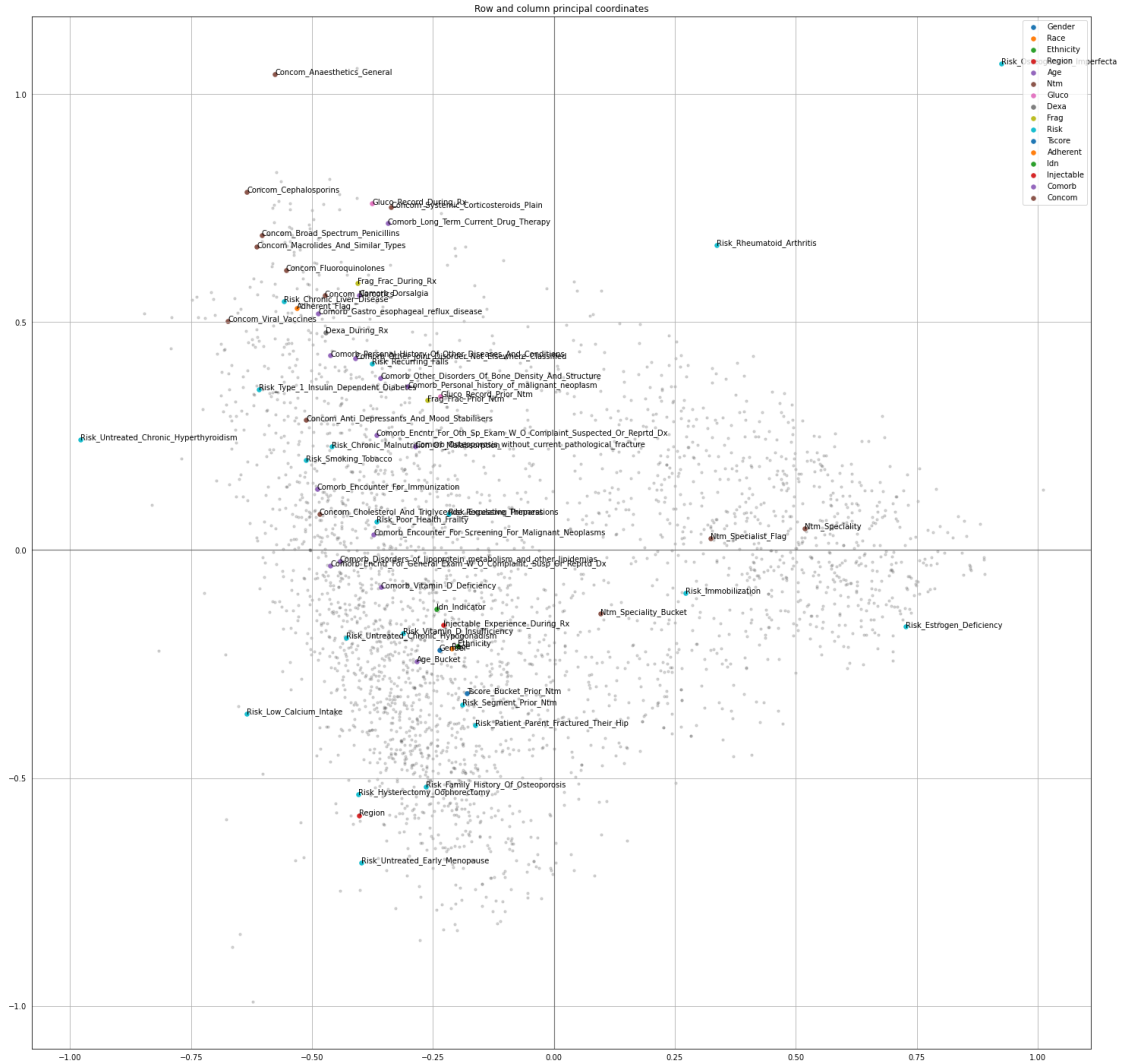

- PERSISTENCY_FLAG unbalanced data, greater predominance in Non-Persistent.

# Categorical Data Analysis Pt2

# Chi Square

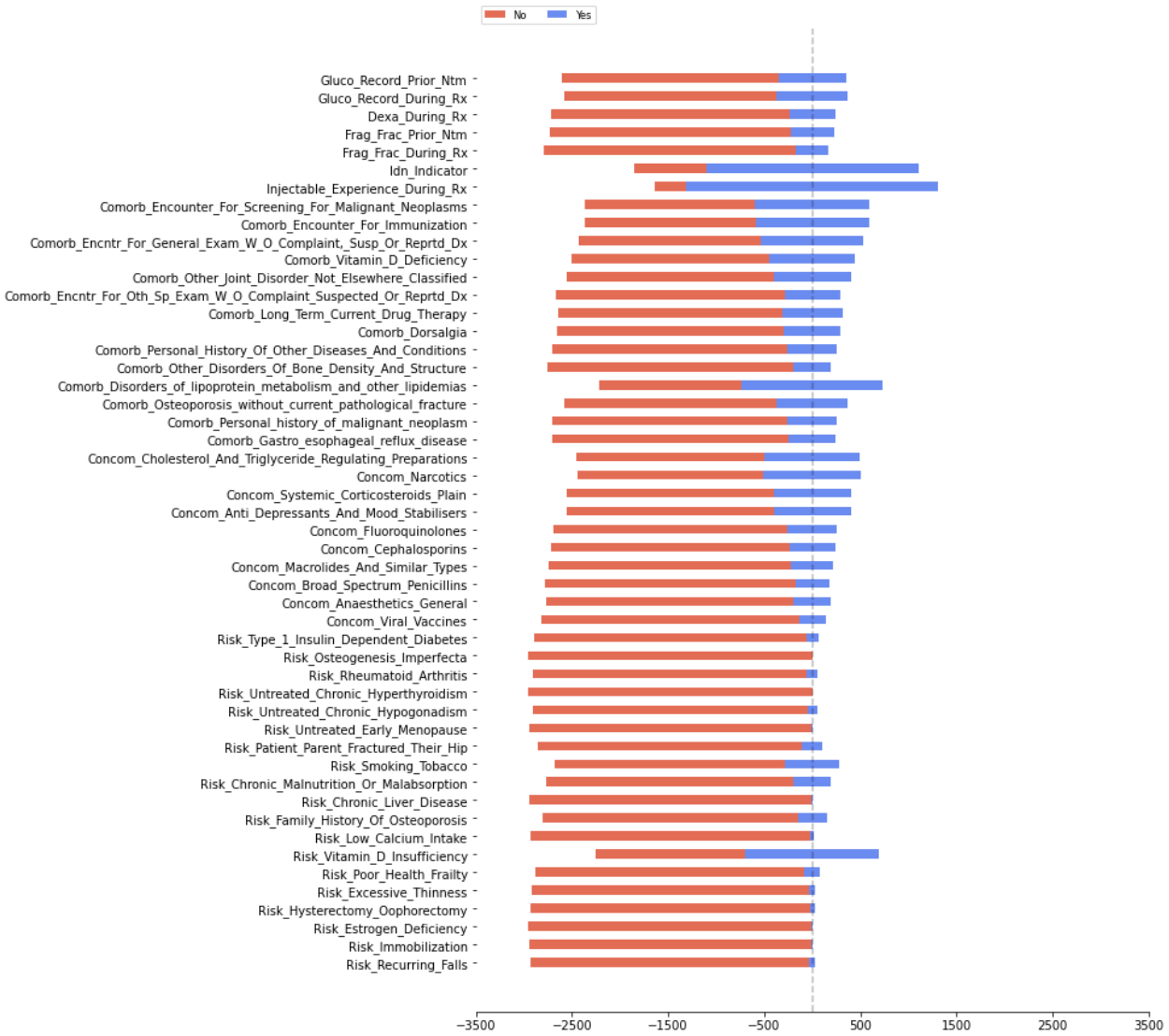

- The Chi-Square Test of Independence determines whether there is an association between categorical variables.

# Correspondence Analysis



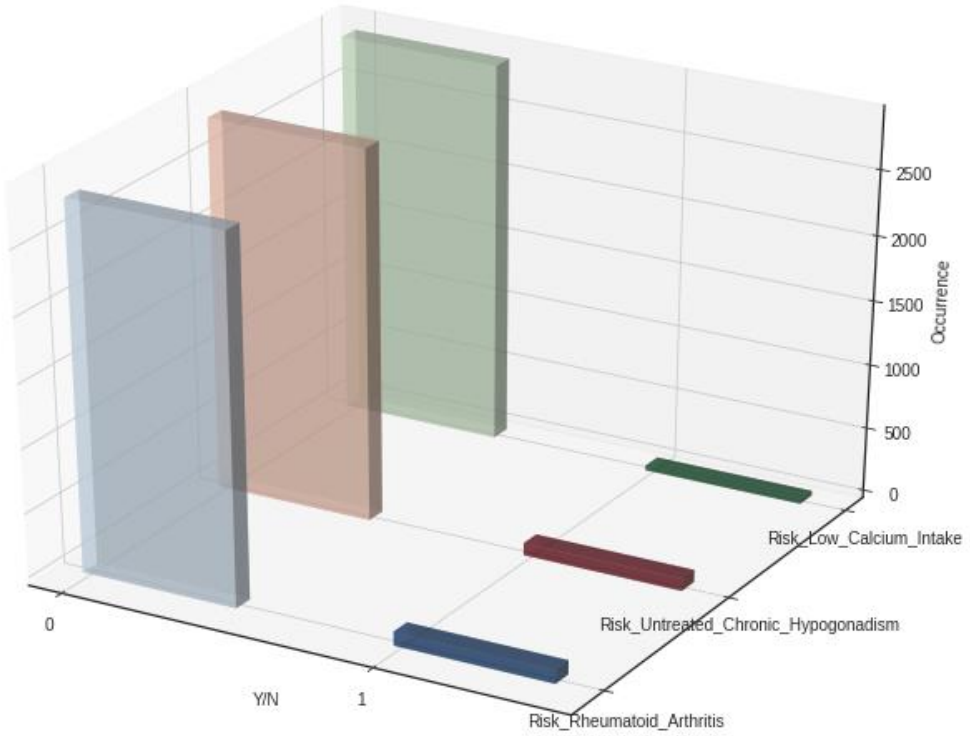

- Correspondence analysis, also called reciprocal averaging, is a useful data science visualization technique for finding out and displaying the relationship between categories.

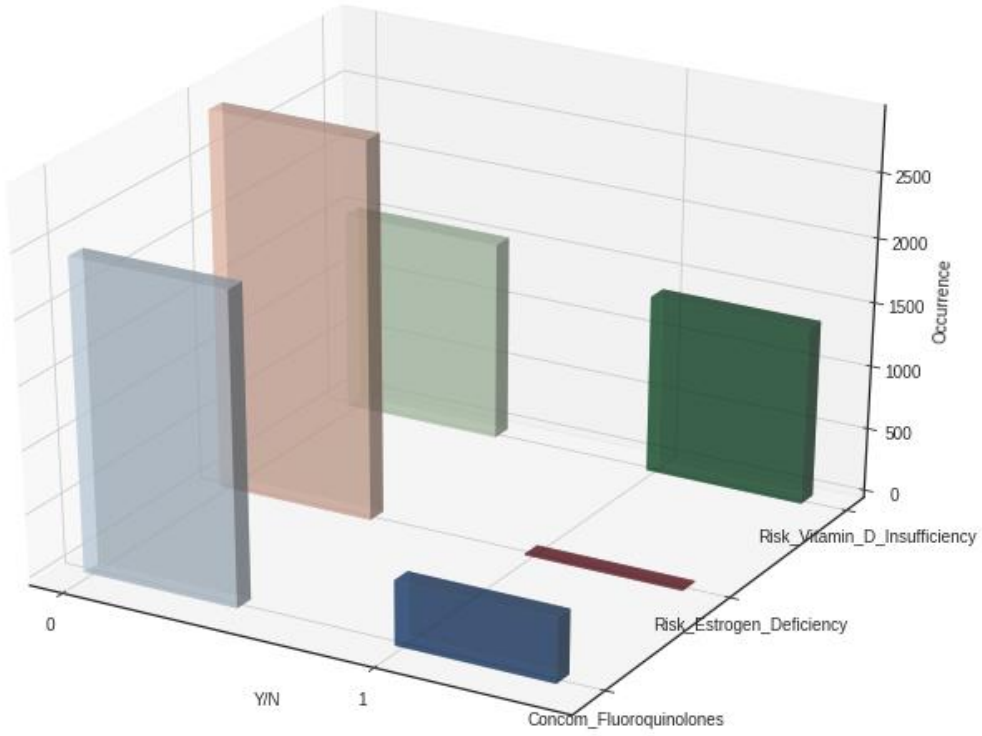
# Tornado Chart




- Butterfly charts, also called Tornado or Divergent Chart, are essentially bar charts comparing two different metrics at a time.

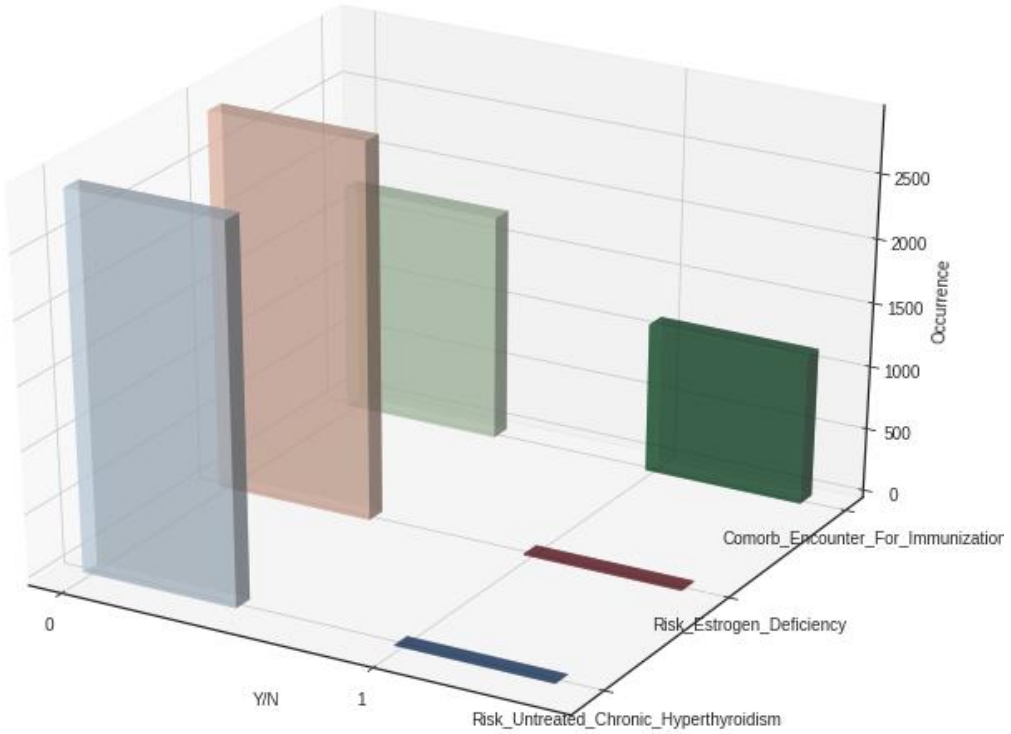
# Crosstabulation

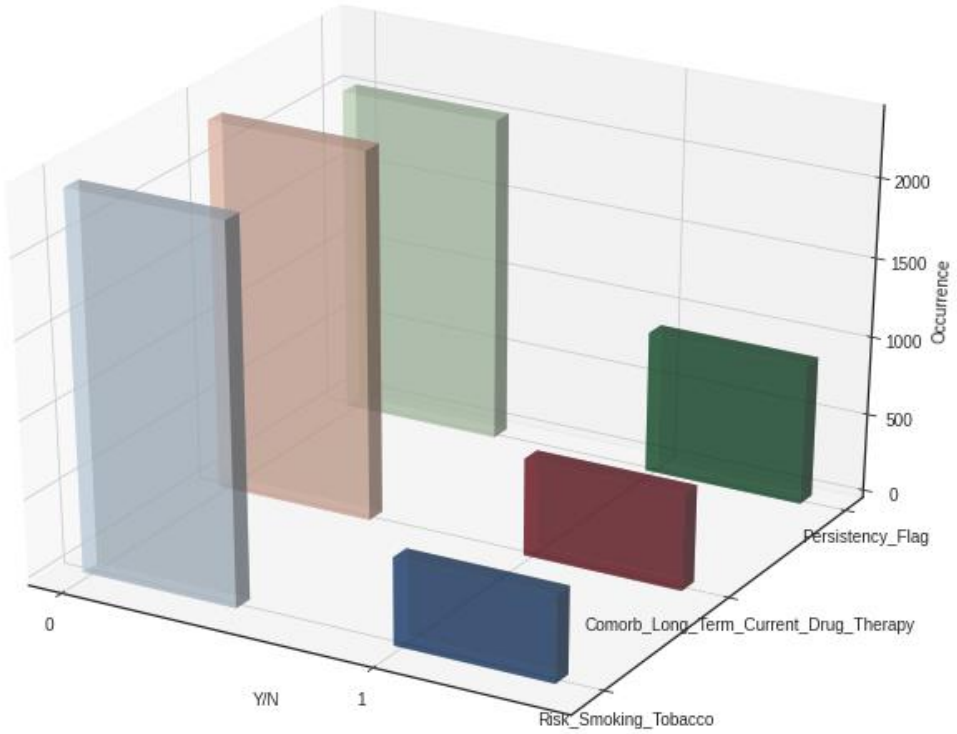



- Takes a dataframe and at least two variables as input, conducts a crosstabulation of the variables.

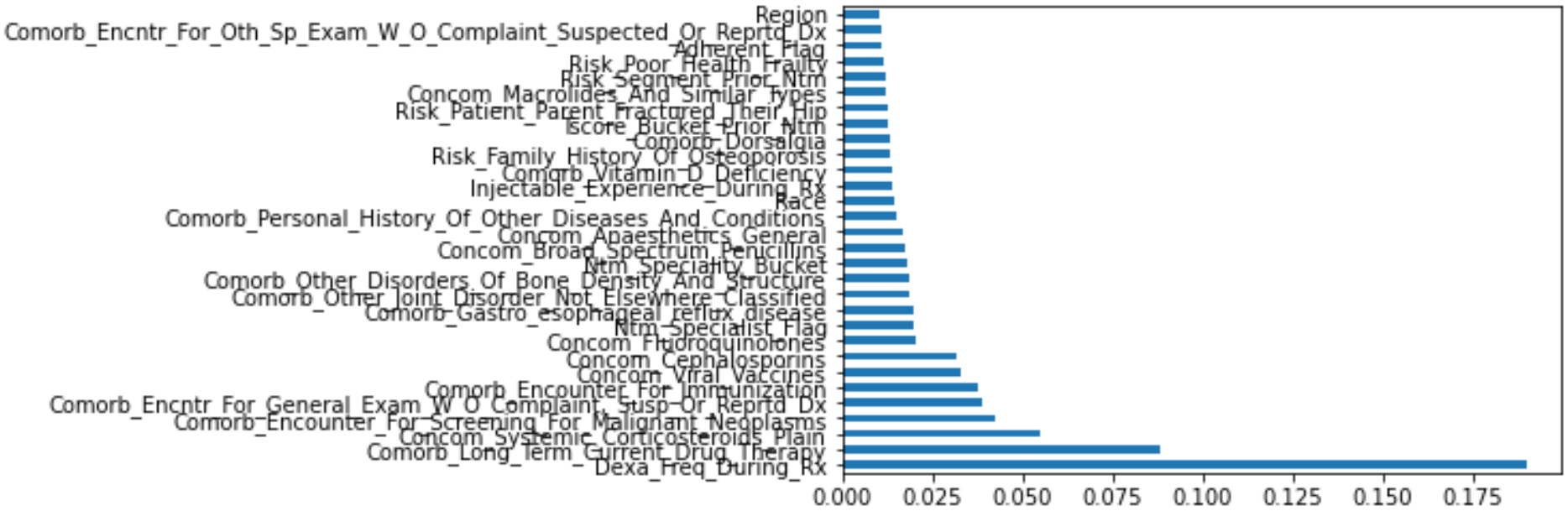
# Crosstabulation




- PERSISTENCY_FLAG unbalanced data, greater predominance in Non-Persistent.

# Feature Importance

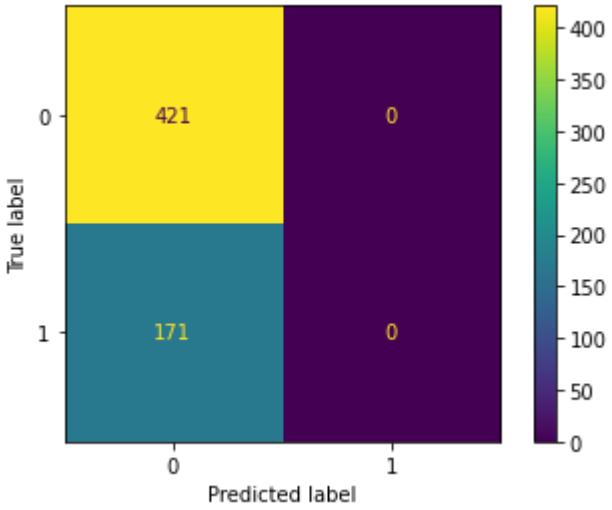



- Using xboosting helps us define the most relevant variables. In our case, 'Dexa_Freq_During_Rx' and the first 20 have very high relevance values with respect to 'Persistency_Flag'.

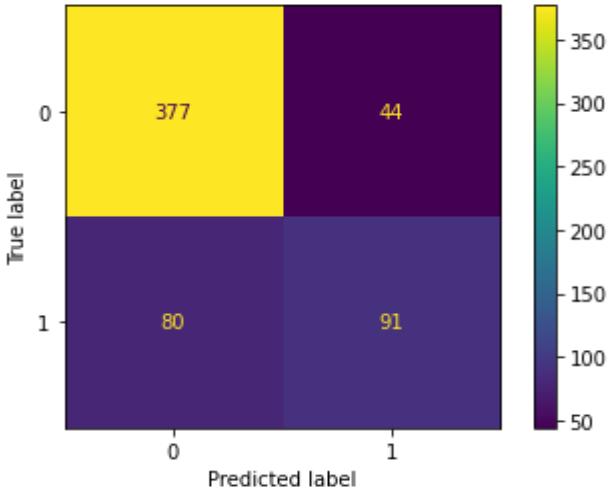
# Model Building

# Models

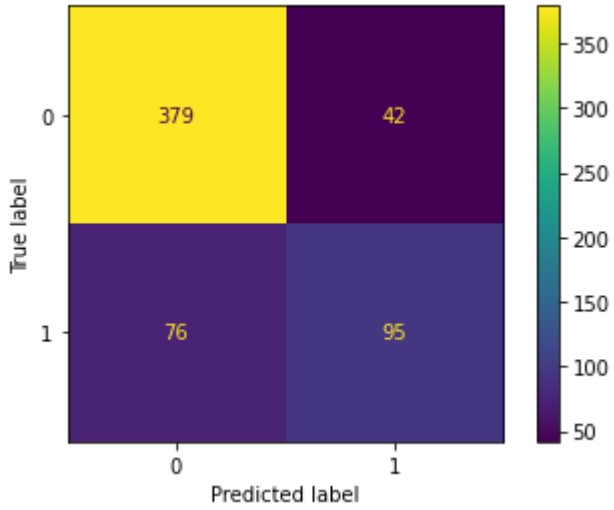
DummyClassifier, acc:71%




LogisticRegression, acc:79%

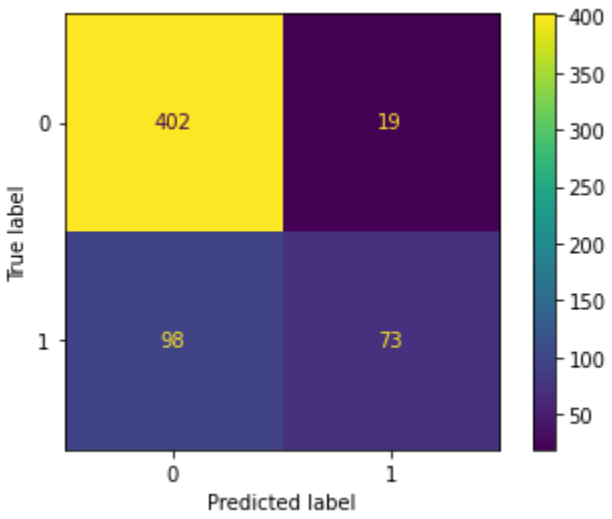



SVC, acc:80%

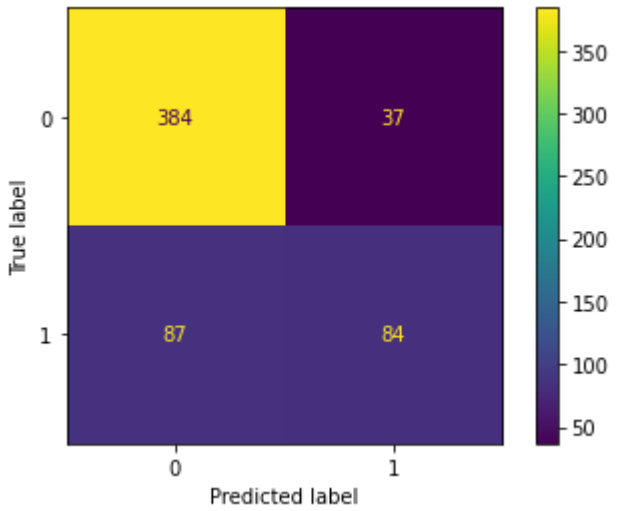


# Models

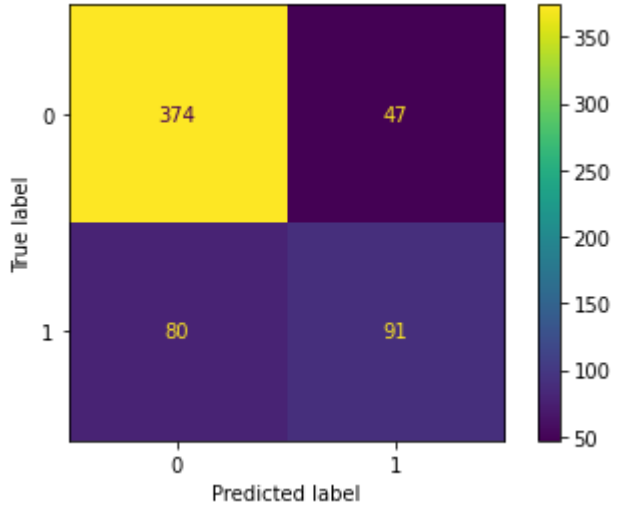
KNN, acc:80%



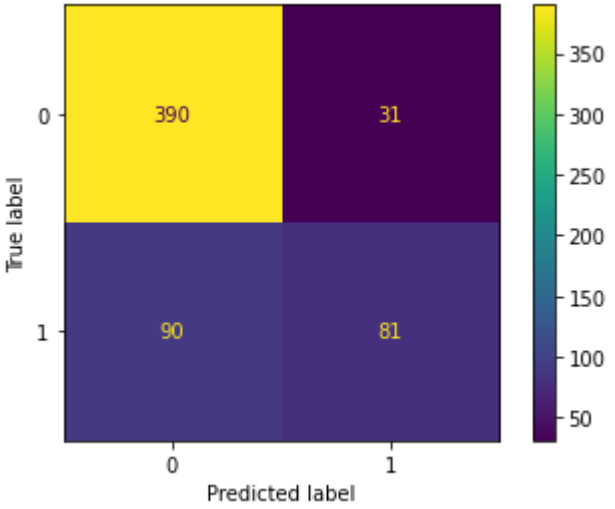

RandomForest, acc:80%



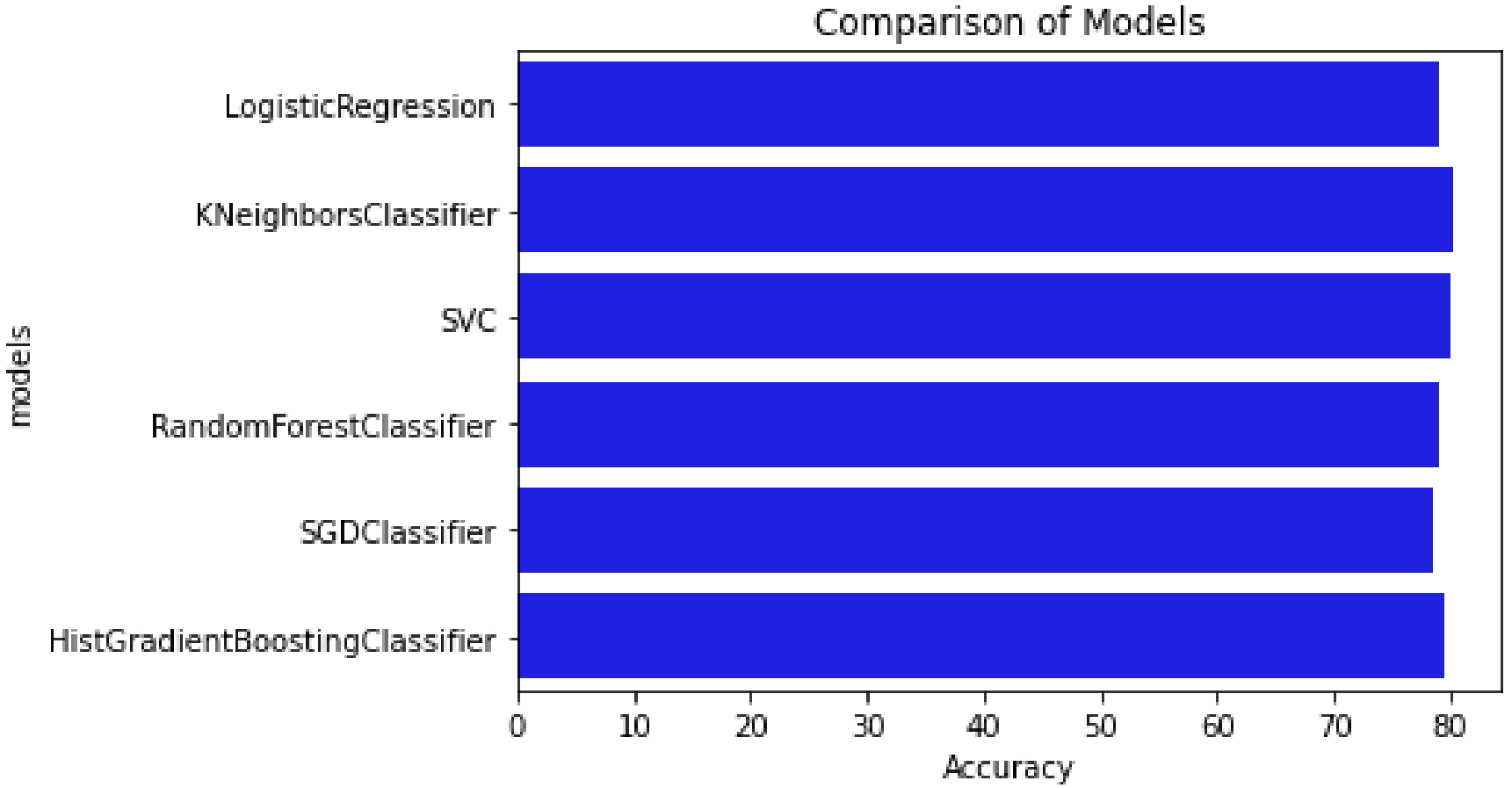

SGDClass, acc:78%



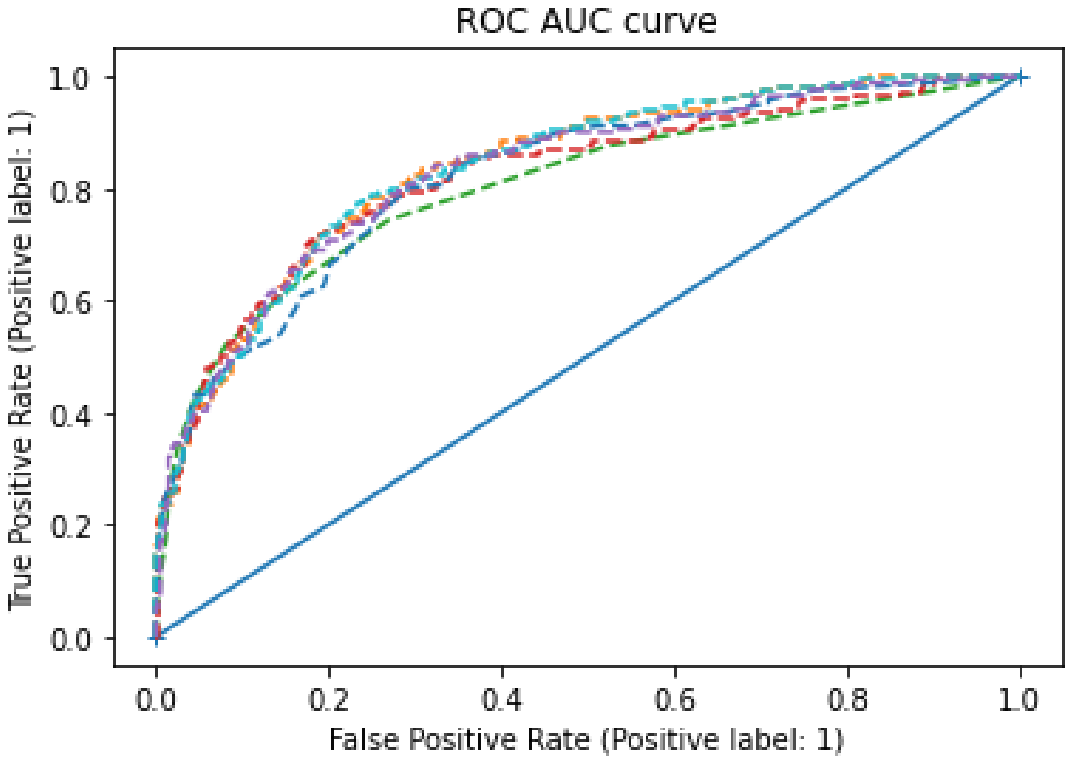

HistGB, acc:80%



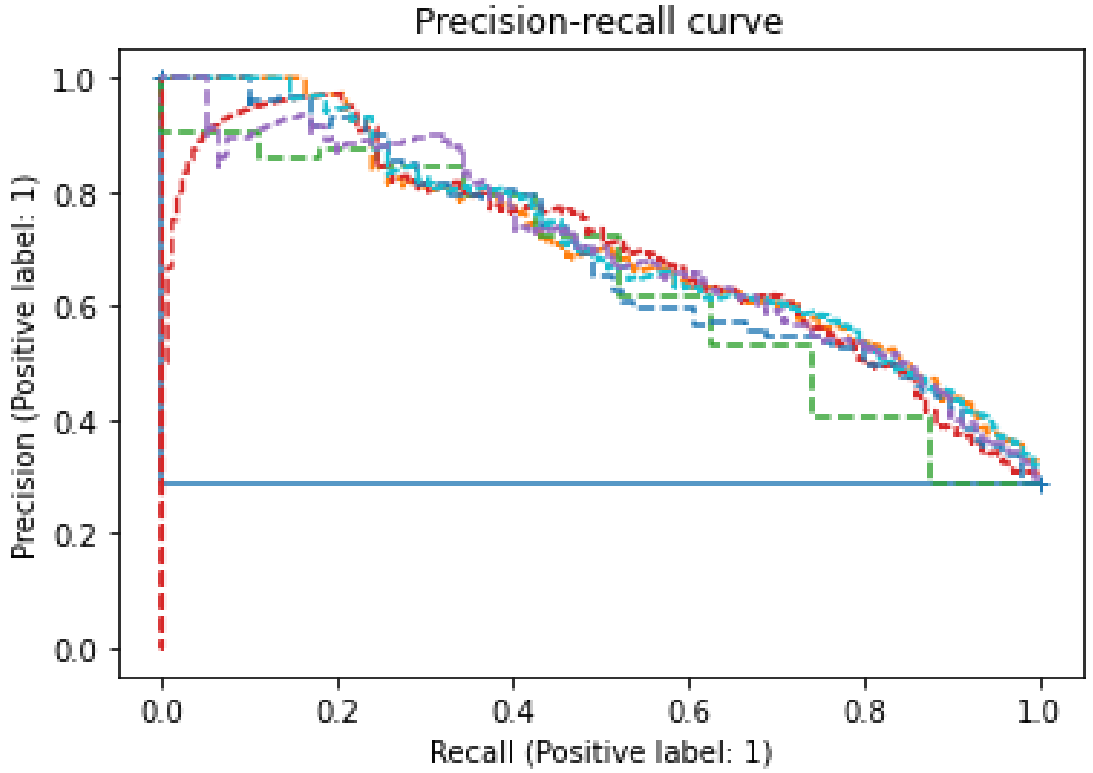
# Models


Comparison of Models
LogisticRegression
KNeighborsClassifier
SVC
RandomForestClassifier
SGDClassifier
HistGradientBoostingClassifier
models
0
10
20
30
40
50
60
70
80
Accuracy

# Metrics




DummyClassifier (AUC = 0.50)
LogisticRegression (AUC = 0.84)
KNeighborsClassifier (AUC = 0.80)
SVC (AUC = 0.82)
RandomForestClassifier (AUC = 0.82)
SGDClassifier (AUC = 0.84)
HistGradientBoostingClassifier (AUC = 0.83)




DummyClassifier (AP = 0.29)
LogisticRegression (AP = 0.71)
KNeighborsClassifier (AP = 0.65)
SVC (AP = 0.69)
RandomForestClassifier (AP = 0.69)
SGDClassifier (AP = 0.71)
HistGradientBoostingClassifier (AP = 0.70)

# Final Recommendations

# Final Recommendations

- It is recommended to apply classification algorithms.
- **Precision**: precision is the ratio or percentage of correct classifications of our classifier. The highest precision was obtained with SGDClassifier and LogisticRgression
- **Recall**: the recall or sensitivity of our model is the ratio of positives detected in the dataset by our classifier. The highest recall was found with SGDClassifier and LogisticRgression.
- The models used had good performance, but they can be improved with data processing, since the most relevant ones were taken.

# Thank You


Your Deep Learning Partner