Data Science Persistency of a Drug Project

Exploratory Data Analysis

Name	Email	Country	College	Specialization
Han-Fu Lin	hanfu.lin@mail.utoronto.ca	Canada	University of Totonto	Data science
Aişe Refia Yilmaz	aiserefia.yilmaz@gmail.com	Turkey	Beykent University	Data science
Anıl Ilgın Büyüksaraç	Anililginb@gmail.com	Turkey	Y.t.u	Data science

Includes

➢ Data Analysis Approach
➢ Problem Understanding
➢ Business Understanding
➢ Data Preparation and Understanding
➢ Exploratory Data Analysis
➢ EDA Summary
➢ Model Selection and Development
➢ Model Building
➢ Final Recommendation
➢ Conclusion

Data Analysis Approach

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.

- Explore and Understand the data
- Prepare and clean the data

Analyze the data

Find the features of drug persistency

Give recommendations

Problem Understanding

Since the data is very large, we know that ABC pharmaceutical company wants to run the whole process faster and to get precise results.
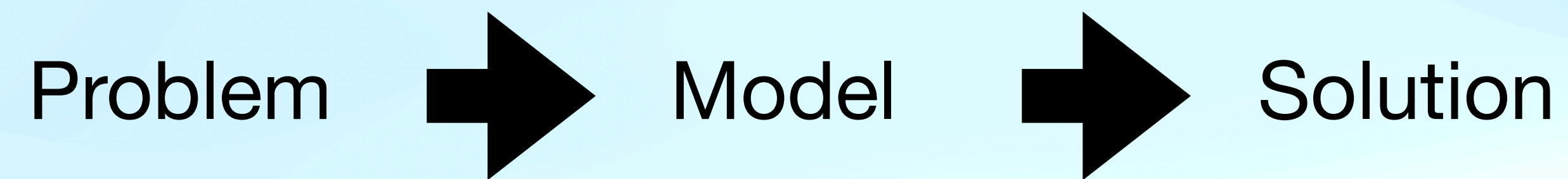
We have considered and measured many factors that do or do not affect permanence.

ML Problem:

With an objective to gather insights on the factors that are impacting the persistency, build a classification for the given dataset.

Business Understanding

- All process algorithms based on:

  Problem ➡️ Model ➡️ Solution

Task:

- Problem understanding
- Data Understanding
- Data Cleaning and Feature engineering
- Model Development
- Model Selection
- Model Evaluation
- Report the accuracy, precision and recall of both the class of target variable
- Report ROC-AUC as well
- Deploy the model
- Explain the challenges and model selection

DATA PREPARATION and UNDERSTANDING

➢We used pandas library and Google Colab Notebook

➢Data was cleaned and prepared for analyzation.

Approaches:

Number of NA values and missing values.

Outliers and Skewed.

Virtualization of data

EXPLORATORY DATA ANALYSIS

➢ We analyzed gender, race, ethnicity, age, region, IDN Indicator values because demographic characteristics are personal and do not change.

➢ We produced 'persistent and non-persistent of drug' histogram graphs of these properties.

➢ Data variables are presented in Heatmap.

➢ According to bar graph, the highest values are seen in the midwest and south parts of the region permanent and non-permanent drug.

On the ethnicity persistent and non-persistent bar graph, not hispanic is clearly ahead of unknown and hispanic.

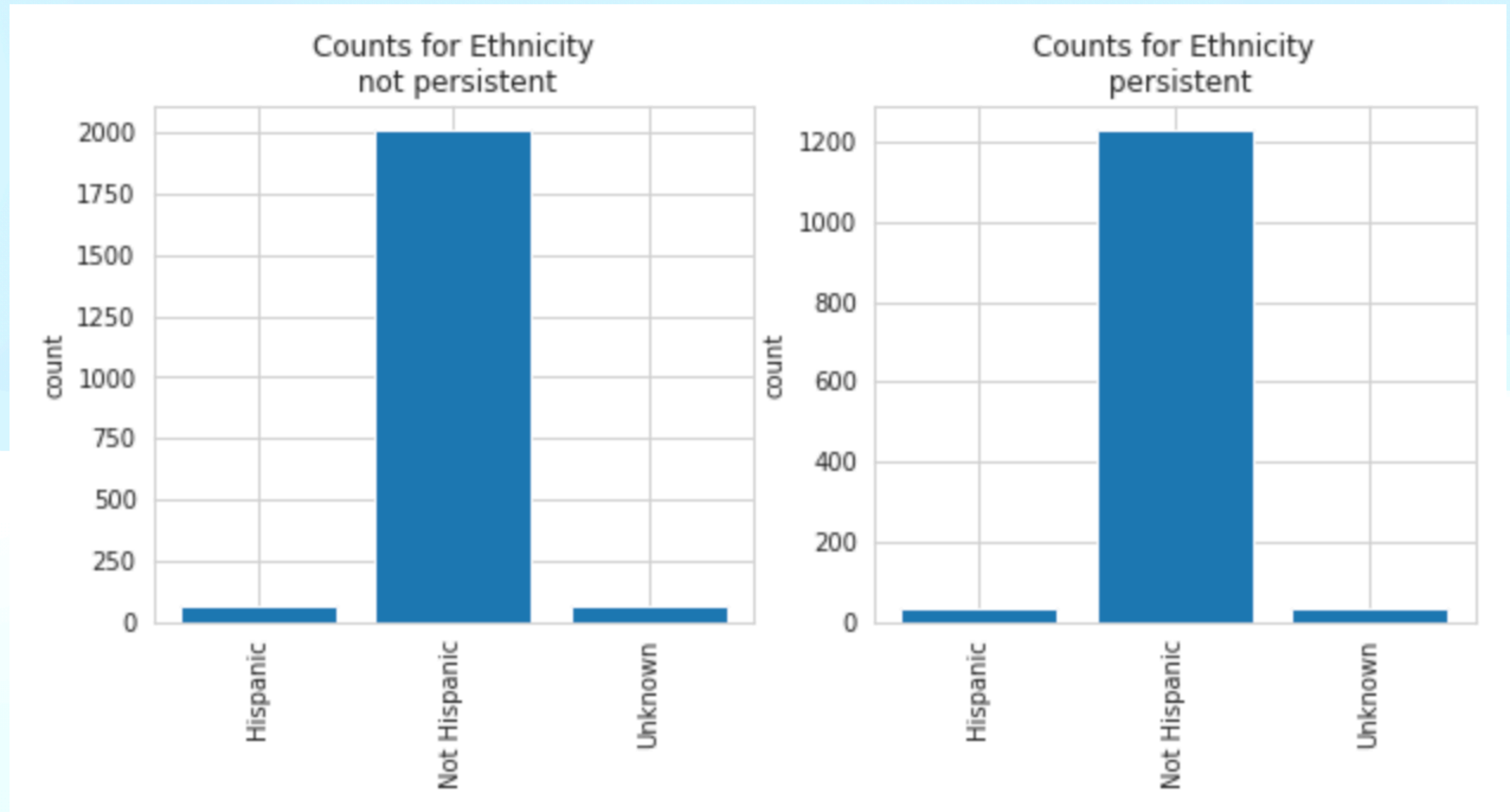As a result of the analyzes made, we got different outputs every week and saved them.

Heatmap

Outlier Test

Boxplot of sqrt_Dexa_Freq_During_Rx

Counts for Ethnicity
not persistent
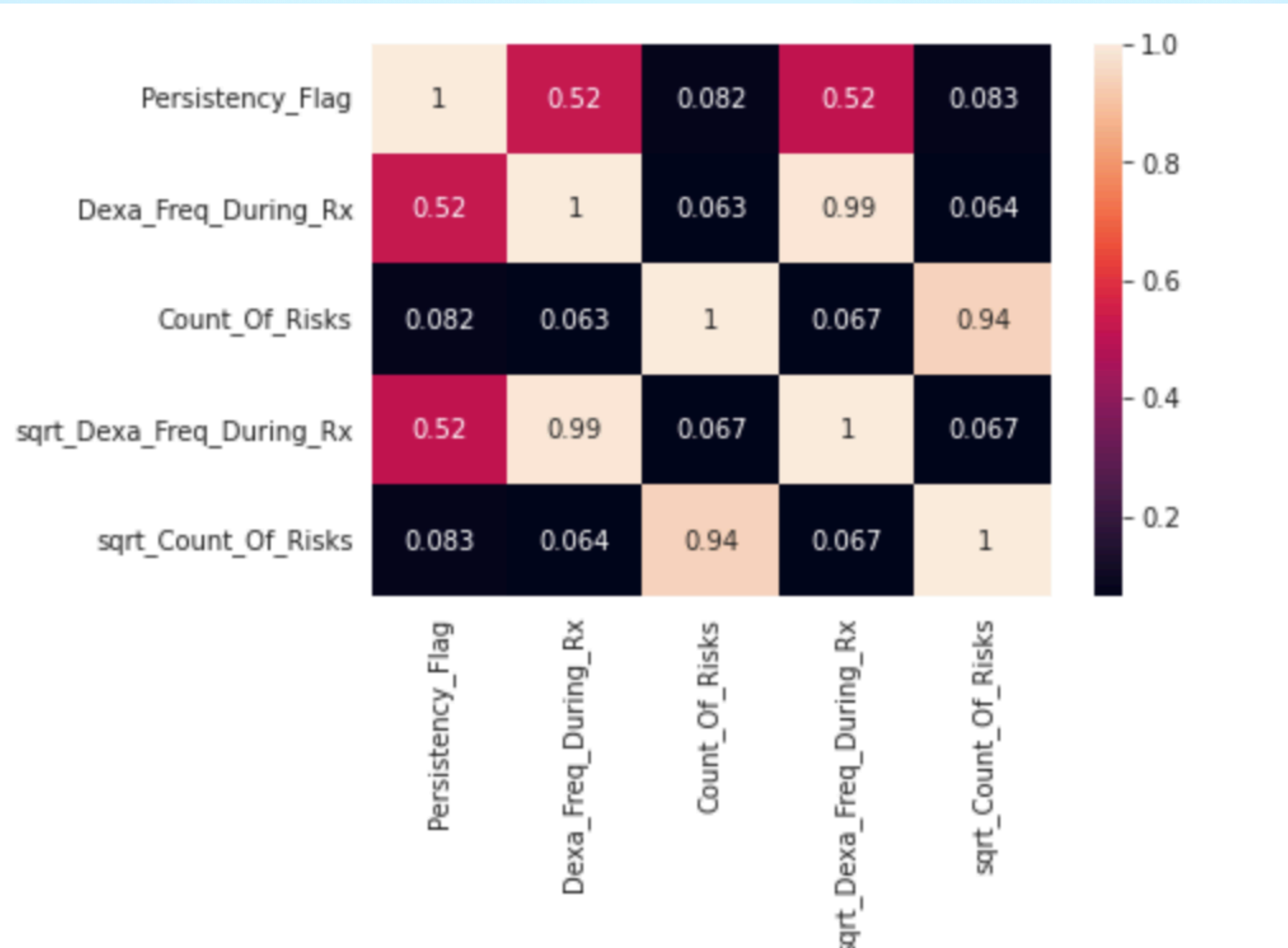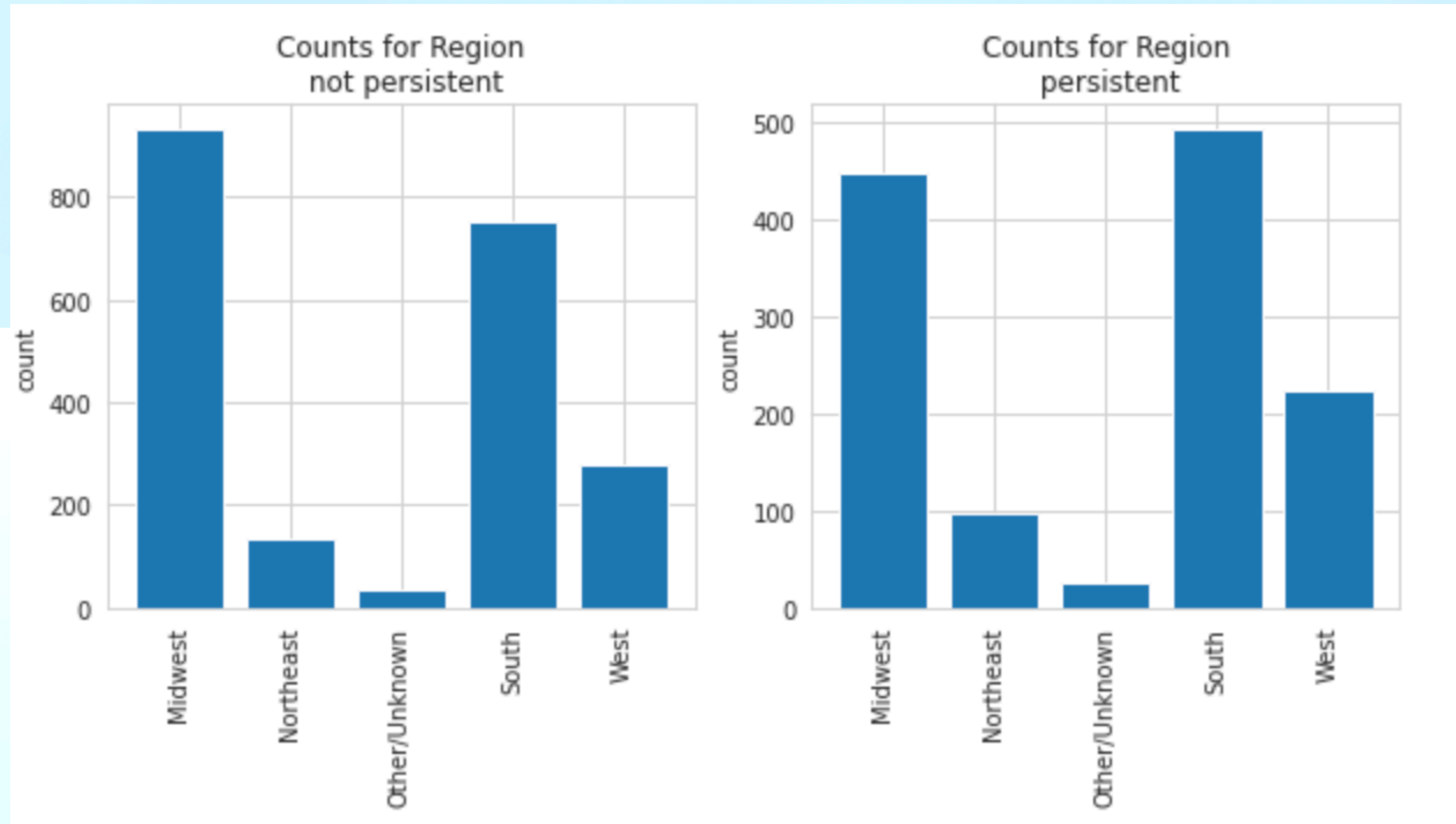
Counts for Ethnicity
persistent
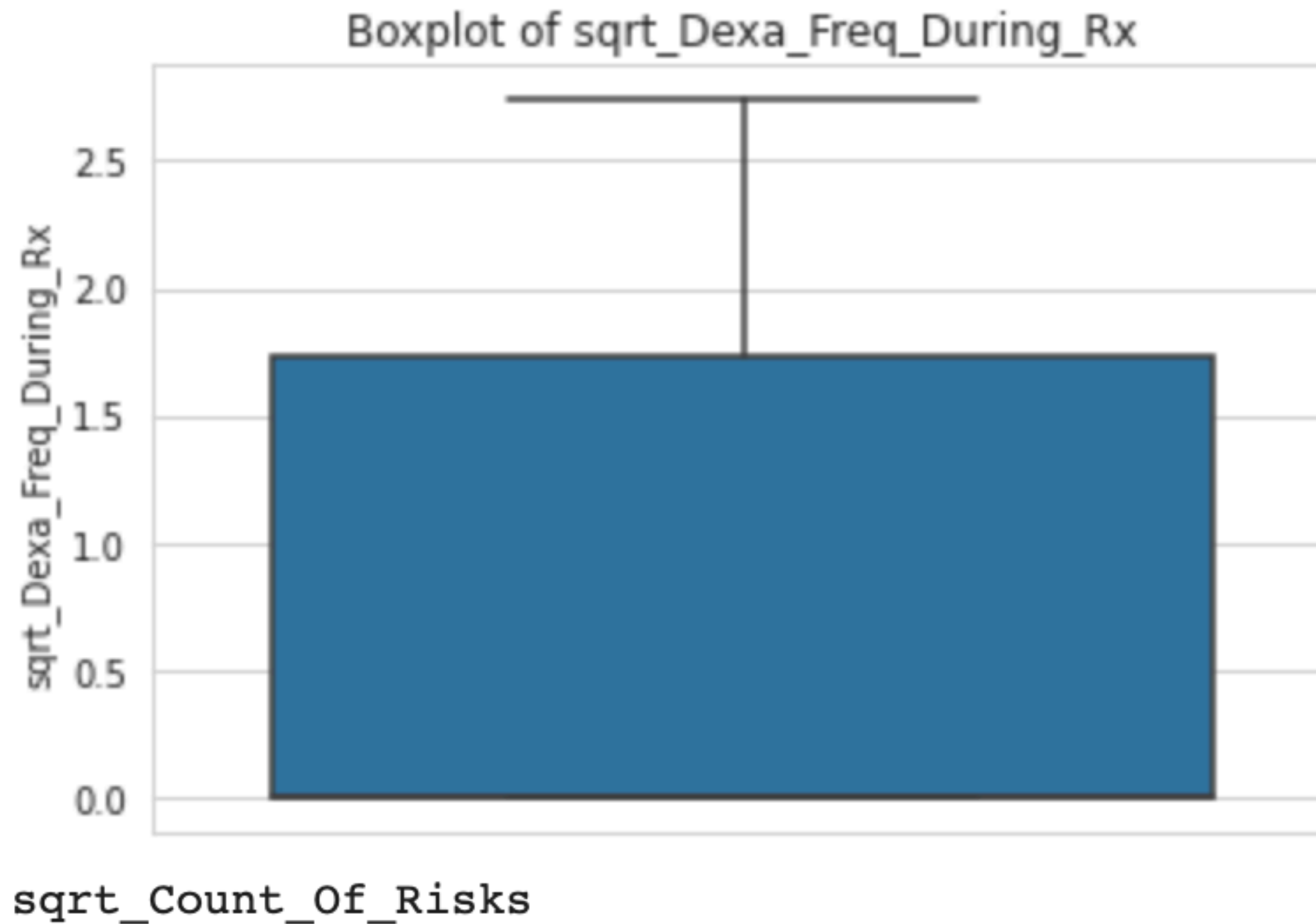
Although the values of the two graphs are different, they are quite similar
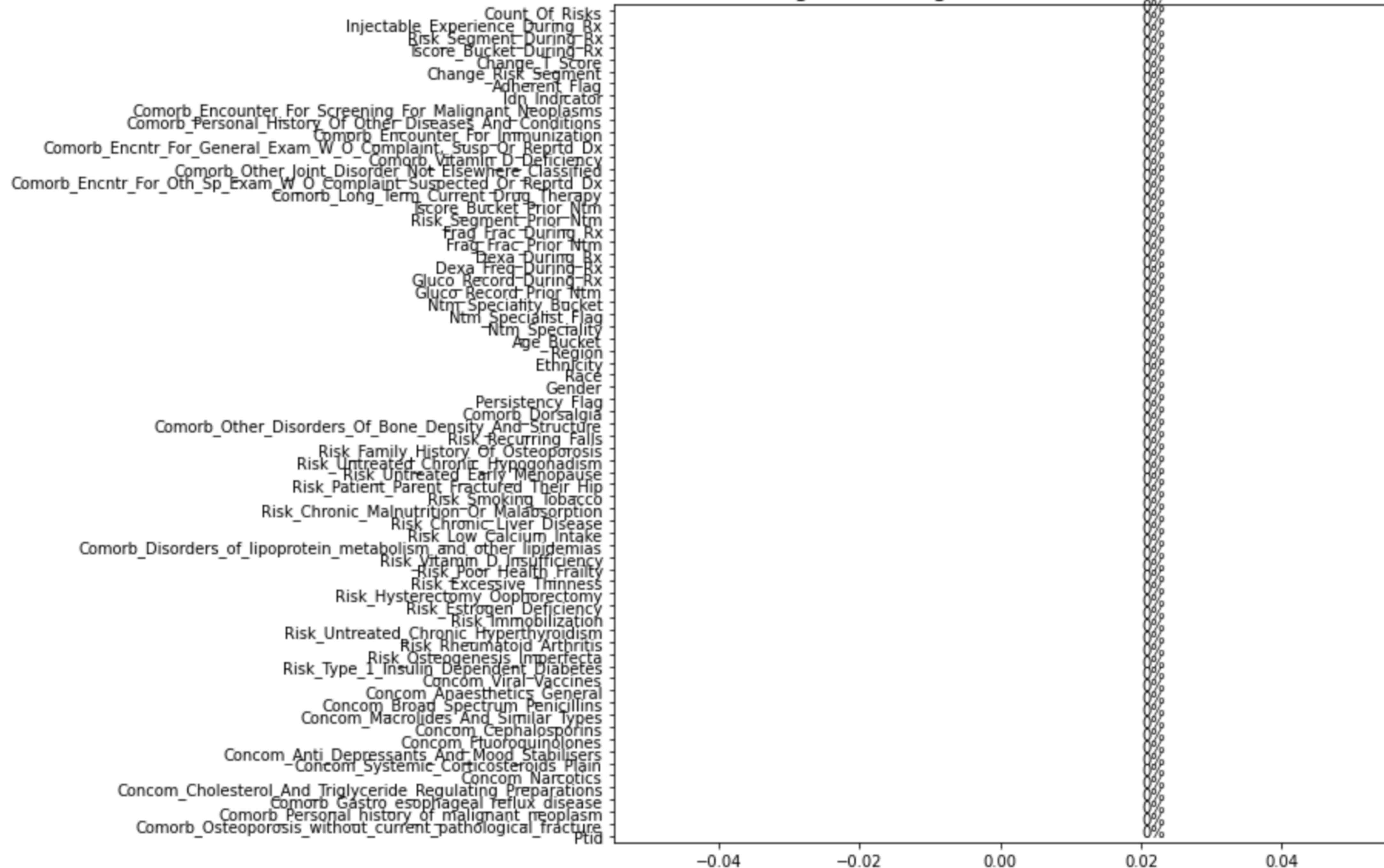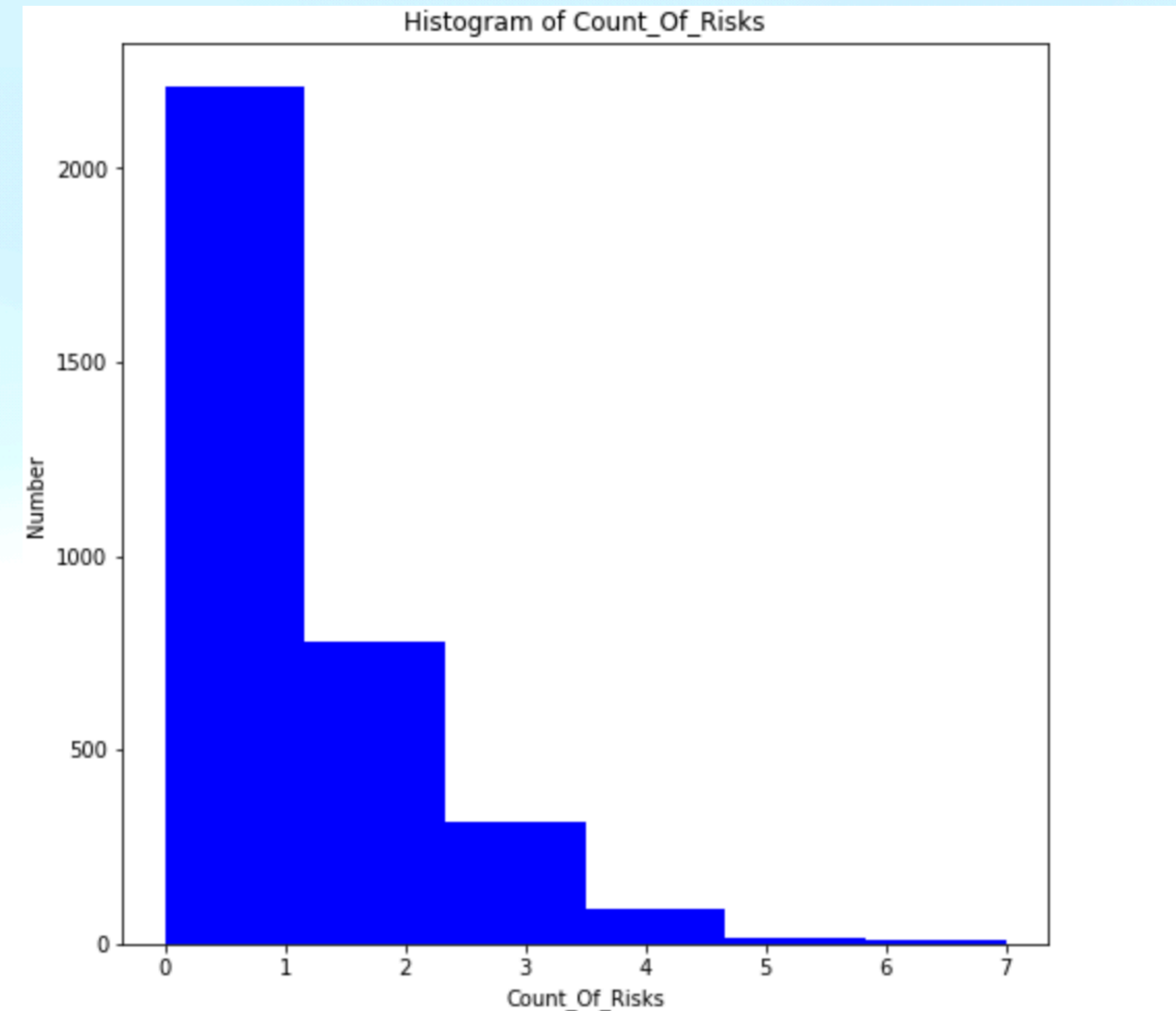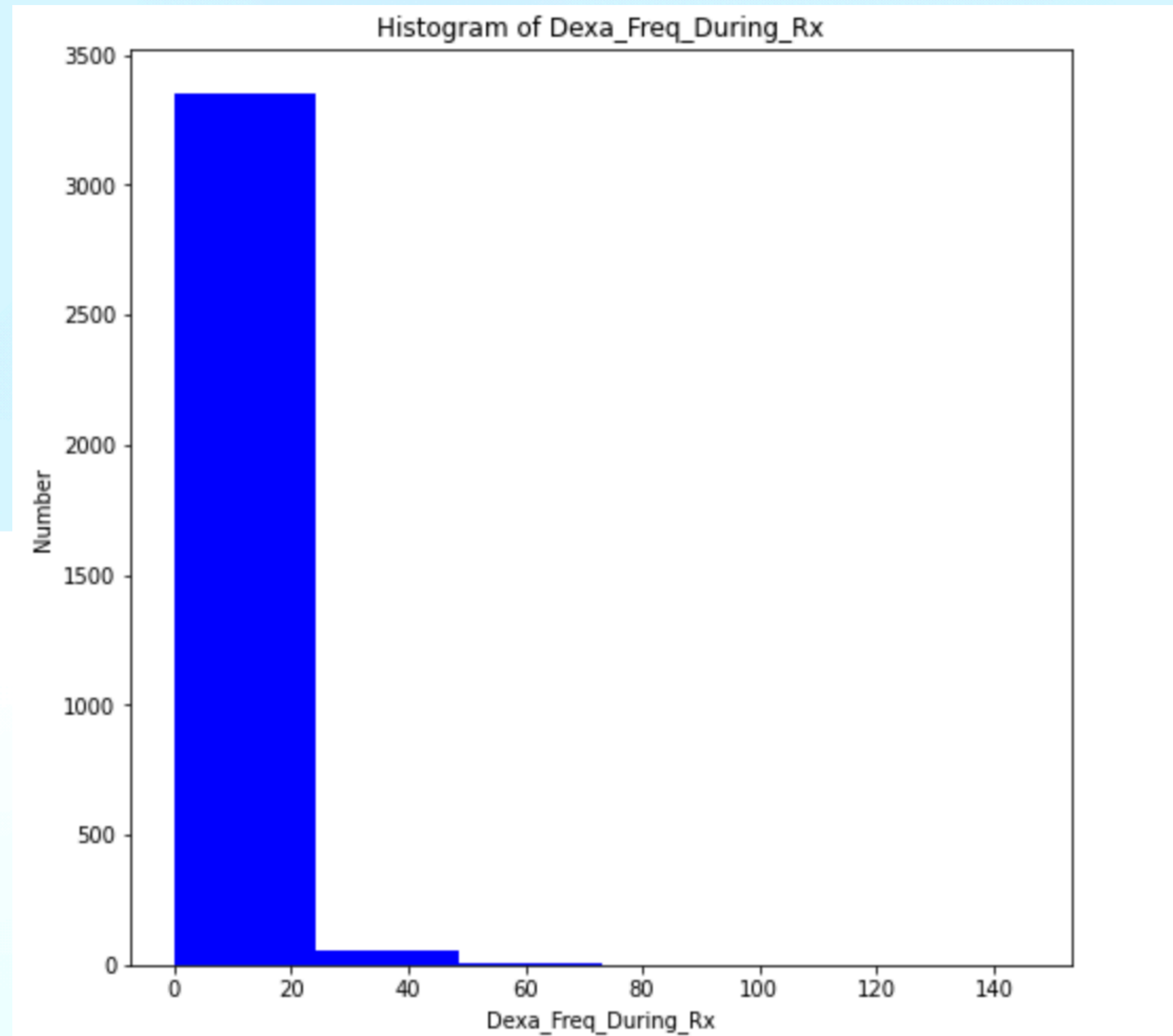in shape.

Persistency and Non-Persistent

There are not very big difference between each graph.

Percentage of Missing Values Per Column in Train Set

Skrewed Histogram

Histogram of Dexa_Freq_During_Rx and Count_of_Risks are positively skewed distribution.

EDA Summary

- Although there are many different factors apart from the analyzes we have made with the data we have, the result we have obtained regarding drug permanence is very high.

- There are two situations here. Firstly, this rate may decrease or increase with the data we do not have. Secondly, we can consider the 92 accuracy rate and values close to it.

Model Selection and Development

- **Logistic Regression result**

```python
print(f'AUC is {roc_auc_score(y_val,y_pred)}')
print(f'F1 score is {f1_score(y_val,y_pred>0.5)}')
```
```
AUC is 0.8860176017601761
F1 score is 0.8209806157354619
```

- **Decision Tree result**

```python
print(f'AUC is {roc_auc_score(y_val,y_pred)}')
print(f'F1 score is {f1_score(y_val,y_pred>0.5)}')
```
```
AUC is 0.8202805280528053
F1 score is 0.7541371158392436
```

- **LightGBM result**

```python
print(f'AUC is {roc_auc_score(y_val,y_pred)}')
print(f'F1 score is {f1_score(y_val,y_pred>0.5)}')
```
```
AUC is 0.9268206820682068
F1 score is 0.8711111111111111
```

Model Building

- Categorical data were converted into numerical data.

- Tree Classifier, Random Forest Classifier and LightGBM Classifier regression techniques were used.

Final Recommendation

- The LightGBM model is the most suitable model because it has the highest AUC and F1 score.

- It will give the best results in persistent and non-persistent drug effects compared to other models.

Thank you