

Group name: Lone Wolf

Name: Nathan Adam

Email: [nathan.adam@myport.ac.uk](mailto:nathan.adam@myport.ac.uk)

Country: England

College/Company: Appen

Specialization: Data Science

GitHub repository link: [https://github.com/N-A-ML/Data\\_Glacier\\_Final\\_Project\\_Week\\_7\\_to\\_13](https://github.com/N-A-ML/Data_Glacier_Final_Project_Week_7_to_13)

## Problem description:

Use Python to perform EDA on the datasets to provide some insights for the pharmaceutical company, and to support future modelling work.

## Exploratory data analysis performed:

Seaborn was used for the plotting.

We explored the average values for each numerical variable (Count\_Of\_Risks and Dexa\_Freq\_During\_Rx) amongst the Persistency flag classes. Bar charts were created to visualise this. On average, the numerical variables had higher values for the persistent class.

We created a scatterplot of Count\_Of\_Risks against Dexa\_Freq\_During\_Rx. It was clear that as Count\_Of\_Risks increase, Dexa\_Freq\_During\_Rx tends to decrease.

A new data frame was made, which was created by restructuring the data. We made a plot of the counts of the true labels for each of the best 6 features (found with SelectKBest in weeks 8-9, grouped by persistency flag. We created similar bar plots for other variables grouped by persistency flag, such as race, region, ethnicity, Hispanic/non-hispanic, age grouped.

Outliers in original dataset, grouped by persistency flag, were investigated with box plots. There was a low number of outliers for Count\_Of\_Risks, and a lot of outliers for Dexa\_Freq\_During\_Rx. Whether or not a patient had a dexa scan was more predictive of their persistency, rather than the number of dexa scans.

A correlation heatmap was created to check for multicollinearity and see the correlations between the features and the target. There wasn't multicollinearity and the highest correlation was between Dexa\_During\_Rx\_Y and Persistency\_Flag.