ASSUMPTIONS MADE:

Meds Data:

If there are two rows per individual with all the columns same they are assumed to be duplicate entries for the same individual and hence all except one instance is removed.

If there are two entries for the same id with different rxStartMonth and rxStartYear safe to assume that they are two separate instances of an individual taking the medicine and we can add the quantities to get the total quantity of that medicine taken by the individual

If there are two entries for the same id with same rxname, start year and month but different rx Quantity or form or NDC they are assumed to be separate instances of the individual taking the medicine and the quantities are added to get the total quantity of that medicine taken by the individual

QUICK SUMMARY OF APPROACH TAKEN AND RESULTS

Q – 1:

The most common medications for each disease are calculated by creating a group for each of the disease of all the individuals who had the disease and then summing over the medicine quantity columns (converted to a Boolean form) to see how many of the individuals who had the particular disease were ever on that medication. Finally the top 3 most popular medications were extracted from the sorted list for each of the disease.

Q – 2:

I did a correlation between the quantity of each of the medicine taken and one variable each indicating if the disease was present or not. Finally the top 3 correlated variables for each of the disease were extracted by looking at the coefficient of correlation. Caution was also taken to see if there were any significant negative correlations and extracting them as well.

Q – 3:

I used PCA to reduce the large amount of features created by the number of medicines that we had in the dataset. The top 10 components were taken and I built 3 models (GBC, RF and Ridge LR) to compare model performance for predicting high BP. GBC performed slightly better than the other 2 in terms of accuracy and AUC and hence was chosen as the final model

Q-4:

To reinforce confidence in the model results performance on the test set was measured and various metrics such as precision, recall, fpr, tnr , fnr extracted from the confusion matrix and analyzed to see how the model performed on the blind test set

Q -5:

To find if any subpopulation’s for those who had a High BP were more likely to take a certain drug I used association rules specifically the Apriori algorithm which helped me explore demographic subpopulations of a minimum set size and see if there was any lift in terms of the medicines that these groups used compared to the overall population with High BP. There were not a lot of strong evidence to show that a particular subpopulation preferred a particular type of drug but one interesting rule showed that White Males who are divorced and have high BP are 55% more likely to take the medicine ‘LISINOPRIL’ than the general population that suffers from high BP. 34% of the this subgroup was seen as ever having taken ‘LISINOPRIL’ compared to 21% for the overall population suffering from high BP.