ALY 6040: Data Mining

Final Project - Milestone 1: EDA

Darshil Oza, Nilay Anand, Rohit Meena, Vamshi Paili, Yash Tadiyal

College of Professional Studies, Northeastern University

Prof. Justin Grosz

April 24, 2022

Introduction

We selected to examine and model diabetic patient data for our final assignment. The dataset includes background and medical history information about the patient. The purpose of this exploratory data analysis is to determine if a patient's background, such as age, gender, and race, influences diagnosis and what effect different drugs have on the patient.

Data Cleaning

We initially reviewed the dataset to see whether it was clean, and if it wasn't, we addressed the fundamental problems with the data, such as anomalies and null or missing data. Initially, we examined all of the columns in the data frame and discovered that some of them included pertinent information that might be used to address and solve the business problem. At first observation, the data in the dataset appeared to be of acceptable quality for the most part, and we didn't see the need for data cleaning or transformation. However, upon closer study, we discovered three columns that were lacking data. The columns weight, payer code, and medical speciality contained a large number of missing values and were thus irrelevant to the situation. These columns were not included in the initial study and were removed later in the modeling process.

We also verified the number of cases in which a medicine was not provided, and if the percentage was larger than 99 percent, that column was removed because this medication was unrelated to the patient's diabetes diagnosis. We evaluated the data structure after removing the extraneous data to determine what columns remained and what sort of data was in each column. A source of worry while working with the data was the presence of the "?" sign in numerous columns where data was not accessible. These symbols have to be replaced in order for data to be simply prepared for modeling. After eliminating null values and normalizing the columns, the data was ready for modeling and could be incorporated into a model.

Exploratory Data Analysis

We studied and analyzed the data after cleaning it in order to get some insight and better understand the potential elements influencing a person's odds of being diagnosed with diabetes. We examined practically every conceivable pair of columns, and in some cases three qualities, to see whether there was some kind of correlation between them. Then we tried to reason rationally about why the link between the columns makes sense and, in the case of outliers, does not.

The graph in figure 1 is based on diagnoses by gender, using the average number of diabetes diagnoses for each gender and using the average diagnosis as the label. The number of diagnoses appears to be comparable in males and females, however the average diagnosis value is greater in females than males. As a result, gender has some relationship with the severity of the condition.

The graph in figure 2 is created by plotting the average number of diagnoses for each age group and adding labels as the average diagnosis score. There is a visible creation of clusters in this situation. One for those aged 50 to 90, and another for people aged 0 to 50 and 90 to 100. The frequency of diagnoses rises with age, as does the severity of the condition. As a result, age has a significant impact on diabetes and its severity. Possible explanations for this will be explored further.

Checking for patients who took diabetes medication, as shown in Figures 3 and 4, we can infer that diabetes medications are not extremely successful because the number of instances in which no change was seen is substantially larger in each age group. There were no examples where there was a change if the diabetic medication was not administered. As a result, it is preferable to seek therapy, and the effects may be evaluated on an individual basis.

Finally, in instances when diabetic medication was administered, we examined whether or not patients were readmitted. We discovered that when diabetic medications were administered as part of the therapy, a reduced percentage of patients were readmitted to the hospital. The entire cost of therapy in such circumstances would be reduced, and patients would spend less time in the hospital. When diabetic medications are not administered as part of the treatment, the readmission rate skyrockets, putting patients at danger.

Conclusion

According to our preliminary findings, diabetes is more commonly found in older people, specifically those aged 50–90, and it affects women more than men. Furthermore, medicine is more beneficial at older ages than at younger ones. The rate of readmission is greater in older age groups than in younger age groups. This might be attributed to increased health difficulties as the body ages. Younger people recover more swiftly and their bodies repair more quickly. The success of diabetic treatment varies from case to case, but if patients are not given the right medication, readmission rates can skyrocket.

Appendix A:

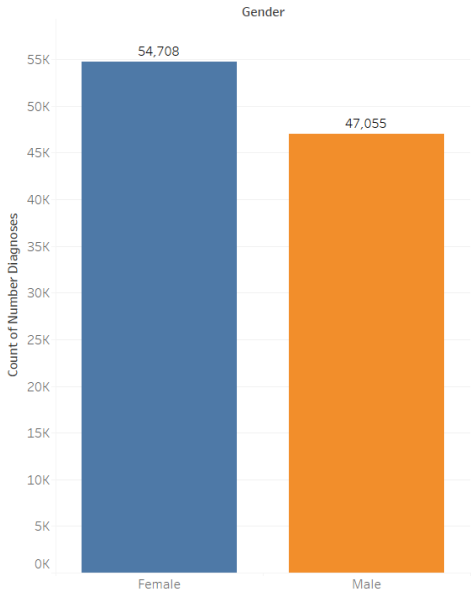

Figure 1: Gender wise Diagnosis

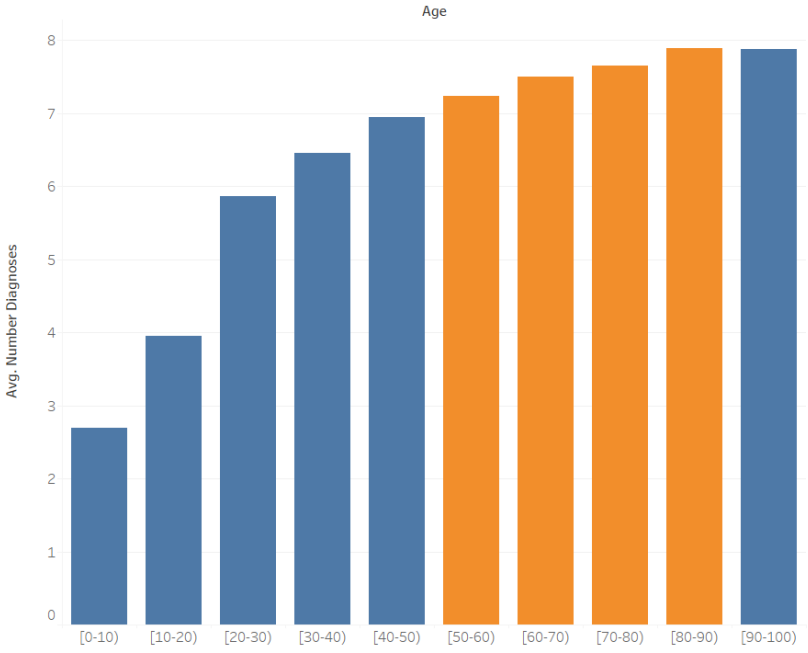

Figure 2: Age wise diagnosis

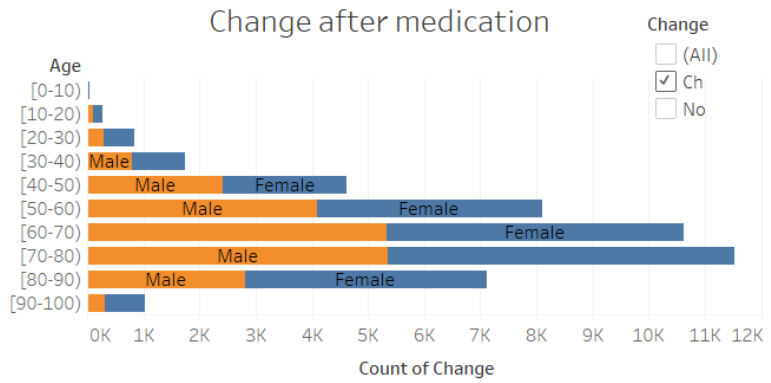

Figure 3: Change in patient given diabetes meds

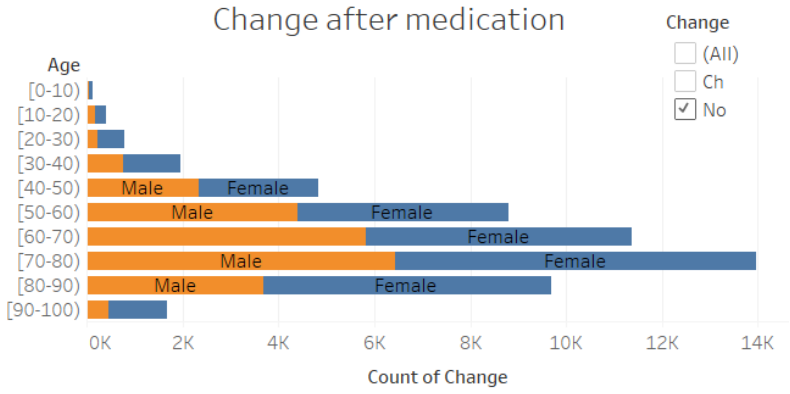

Figure 4: No change in patient given medication

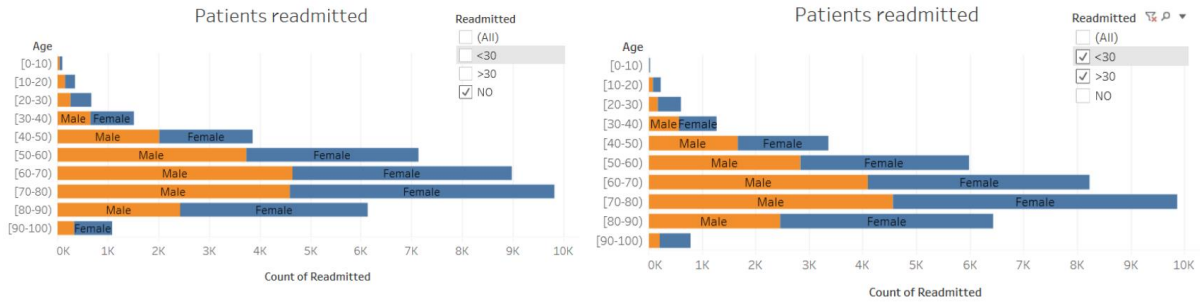

Figure 5: Diabetes meds administered

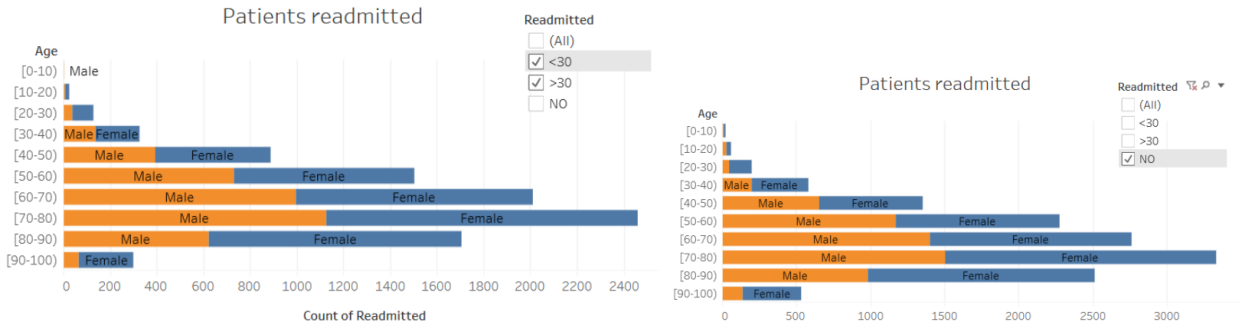

Figure 6: Diabetes meds not administered