Introduction:

Diabetes mellitus is a metabolic disorder defined by higher-than-normal blood sugar level over a prolonged period, with currently three types of diabetes being defined: Type 1 occurs due to a loss of beta cells caused by an unknown autoimmune response resulting in an inability of the pancreas to produce enough insulin (WHO | About diabetes, 2022). Type 1 diabetes usually appears during childhood or adolescence but can also develop in adults. Certain genes have been shown to influence the risk of Type 1 diabetes, however, the exact mechanisms that cause the disease are unknown (Petzold, Solimena and Knoch, 2015). Type 2 Diabetes is the most common form of diabetes in adults and is characterized by insulin resistance which can progress into reduced insulin production. Unlike Type 1 which has a genetic influence the most common factors that lead to type 2 diabetes developing are excessive body weight and insufficient exercise (WHO | About diabetes, 2022). Gestational diabetes is the third main type of diabetes and occurs when pregnant women develop high blood sugar levels without a previous history of diabetes. Usually, these blood sugar levels return to normal after pregnancy, however there is a risk of type 2 diabetes developing (WHO | About diabetes, 2022). Various treatments exist for the different types of diabetes, however, if left untreated severe health complications can develop including serious long-term complications such as stroke, cognitive impairment and cardiovascular disease (Kitabchi, Umpierrez, Miles and Fisher, 2009).

Diabetes represents the 7th cost leading cause of death worldwide, with an estimated 463 million people living with the disease in 2019 (represents 8.8% of the adult population). The global economic cost related to diabetes was estimated at US$727 billion in 2017, with average medical costs being 2.3 times higher for patients with diabetes. The number of people living with diabetes has been predicted to rise and this would also increase diabetes-related medical costs. Therefore, understanding what causes the different types of debates and identifying trends and patterns between diabetic patients could play an important role in reducing the deaths and costs associated with diabetes (Economic Costs of Diabetes in the U.S. in 2017, 2018).

Various traits have shown to be positively associated with diabetes including age, and ethnicity (Pinchevsky et al., 2020). This report will summarize the findings of an analysis based on a dataset which contained the records of diabetic patients admitted to US hospitals from 1999 to 2008, to try and identify possible trends. These records were initially collected to monitor and prevent readmission of diabetic patients, as readmissions can be used as a metric of potential poor care as well as a financial burden to patients, insurers, governments and health care providers. For this analysis the original data was cleaned and transformed, followed by various stages of analysis. This included the development of a predictive model to predict which hospitalized diabetic patients will be readmitted for their condition later as well as a K-Means approach to propose a non-trivial set of patients' clusters that may make business sense to the healthcare industry.

Hypotheses:

H_1 - Age has a higher impact on readmission.

$H1_0$ - Age does not have a higher impact on readmission.

H_2 - African Americans are more likely to be re-admitted than other ethnic groups.

$H2_0$ - African Americans are not more likely to be re-admitted than other ethnic groups.

H_3 - Women patients are more likely to be re-admitted than men.

$H3_0$ - Women patients are not more likely to be re-admitted than men.

H_4 - Diagnose types have a higher impact on re-admission rates.

$H4_0$ - Diagnose types do not have a higher impact on re-admission rates

Methods:

The initial step in the analysis of the diabetic data sheet was cleaning and transforming the data to prepare it for further analysis, the initial shape of the data frame can be seen in the data cleansing python code. This form of data munging included replacing missing values in that data set with numpy.nan. Columns which had more than 50% missing values and those for which 95% of their values were the same were dropped, rows which featured missing values were also dropped. Age was transformed to be the middle value in each given range. Possible missing values within columns diag_1, diag_2, and diag_3 were replaced by the number 0. Numerical features and categorical features were identified, and lists were made for both features. Outliers within the numerical columns were removed, with only values within three standard deviations being kept. Duplicates within the patient_nbr column were also removed. The shape of the data frame following the data munging can be seen in the data cleansing python code.

Following the data cleansing and transformation, our final data set contained 34 observations of 17539 individuals. An initial data exploration was carried out using plots and graphs to identify patterns and trends within the data to validate the hypotheses that were set out for this analysis.

This initial data analysis was then followed by the construction of a predictive model to determine whether a diabetic patient will be re-admitted to hospital. A linear regression model was constructed using the following categories from the dataset: num_medications, number_outpatient, number_emergency, time_in_hospital, number_inpatient, encounter_id, age, num_lab_procedures, number_diagnoses, num_procedures, and readmitted. For the model patients that were not readmitted were assigned the value 0 and patients were readmitted with and after 30 days were given a value of 1. The data was then split into training and test sets to build up the model, followed by cross-validation of the scores generated by the model.

After this initial model was constructed, an improved model was developed with higher performance.

Finally, the k-means algorithm was then used to cluster the cleansed dataset to compare the differences in the mean values of each observation in the cluster.

Results and Evaluation:

Figure 1, Age vs Readmission:

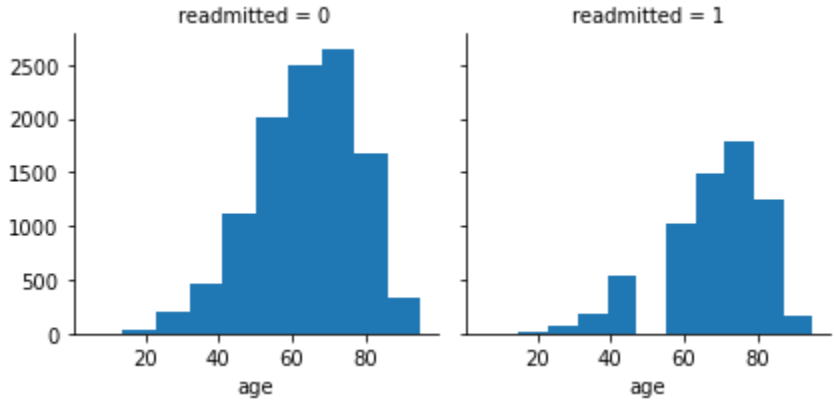

Figure 1 includes two graphs that show the age of patients and whether they were readmitted or not. The left graph represents patients that were not readmitted to hospital following release, ands shows a negatively skewed distribution. The right graph represents patients readmitted within 30 days and shows a negatively skewed distribution.

Based on Figure 1 it shows that the positive trend with age for both non-readmitted and readmitted patients. The left graph can possibly be explained due to older people being more likely to suffer with diabetes and therefore represent a larger proportion of the dataset. The right graph shows that older people are more likely to be readmitted therefore, H1 could be accepted based on the first impression of these graphs. This is in line with clinical studies as elderly patients are more likely to be admitted to hospital when they are unwell compared to younger individuals as well as suffer more serious diabetes related complications (Leung, Wongrakpanich, and Munshi, 2018; UK, Diabetes and Trusted, 2022).

Figure 2, Ethnicity vs Readmission:

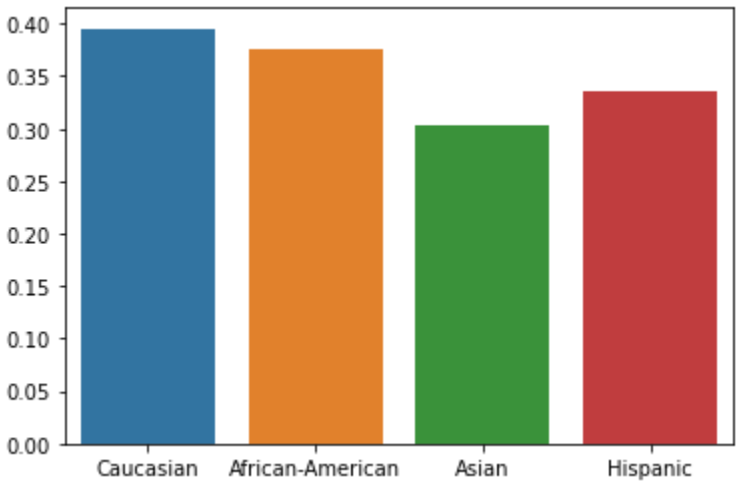

Figure 2 shows the percentage of patients that were readmitted in <30 or >30 days, and their ethnicity. Caucasian patients have the highest percentage of readmission at 39.5%, followed by African American 37.5%, Asian 30.4% and Hispanic 33.5%.

The analysis of this dataset instead shows that Caucasian patients are more likely to be readmitted than any other ethnic group as 39.5% compared to African American 37.5% readmission rates. A further statistical test would need to be conducted to test the statistical significance of these values but based on this graph and readmission values hypothesis two is rejected and the null hypothesis two is accepted. This result is surprising

as prevailing statistics have suggested that African American adults are 50% to 100% more likely to have diabetes than Caucasian adults, possibly due to behavioral, environmental, socioeconomic, physiological, and genetic factors (Harris, 1990). Therefore, it is logical to assume African American would represent the largest readmitted ethnic group. This is not the case for this dataset, and this could be due to an overrepresentation of Caucasians as they made up the largest population of the dataset readmitted or not with 48048 patients compared to 11718 African Americans, 1341 Hispanic and only 461 Asians. This is a common problem in medical studies as minority ethnic groups are often underrepresented (Sirugo, Williams and Tishkoff, 2019). Clinical evidence has also shown that People from Black African, African Caribbean and South Asian backgrounds are more susceptible to develop type 2 diabetes and it is unusual that this does not impact readmission rates, at least for this dataset. (2018; UK, Diabetes and Trusted, 2022)

Figure 3, Gender vs Readmission:

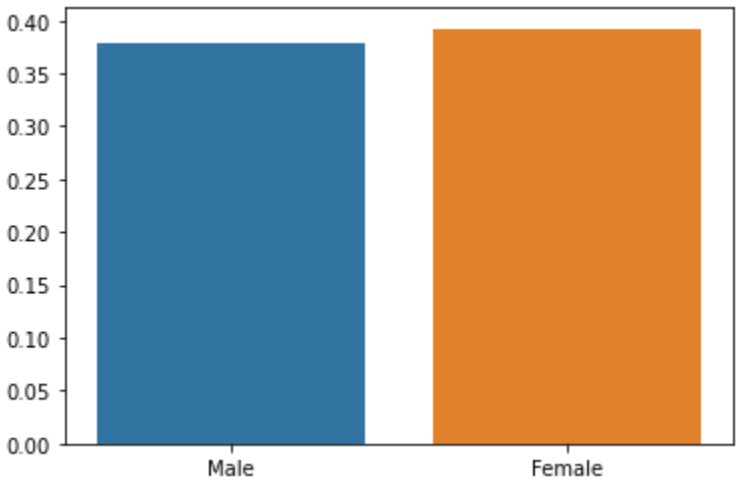

Figure 3 represents the percentage of male and female patients that were readmitted in <30 or >30 days. Male patients had a readmission percentage of 37.8% with female patients having a readmission rate of 39.2%.

As Figure 3 shows that female patients have a slightly higher readmission rate compared to male patients 1.4%. However, this is a negligible difference in readmission rates and as female and patients were more equally represented than ethnicities in the dataset (30191 male vs 34274 female), this 1.4% difference in readmission does not seem significant. Therefore, a statistical test would need to be conducted to definitively prove hypothesis three but based on these preliminary findings hypothesis three is proven. Clinically there are various opinions regarding this as one study did not find statistically significant differences in the in-hospital mortality, 30-day all-cause mortality, or rate of complications between men and women hospitalized with Diabetic ketoacidosis (DKA) a serious complication associated with diabetes (Barski et al., 2011). However, other studies have shown that there is a 3:2 male female diabetic ratio in terms of diagnosis, but this may not play a role in hospital readmission (Gale and Gillespie, 2001).

Figure 4, ICD codes vs Readmission:

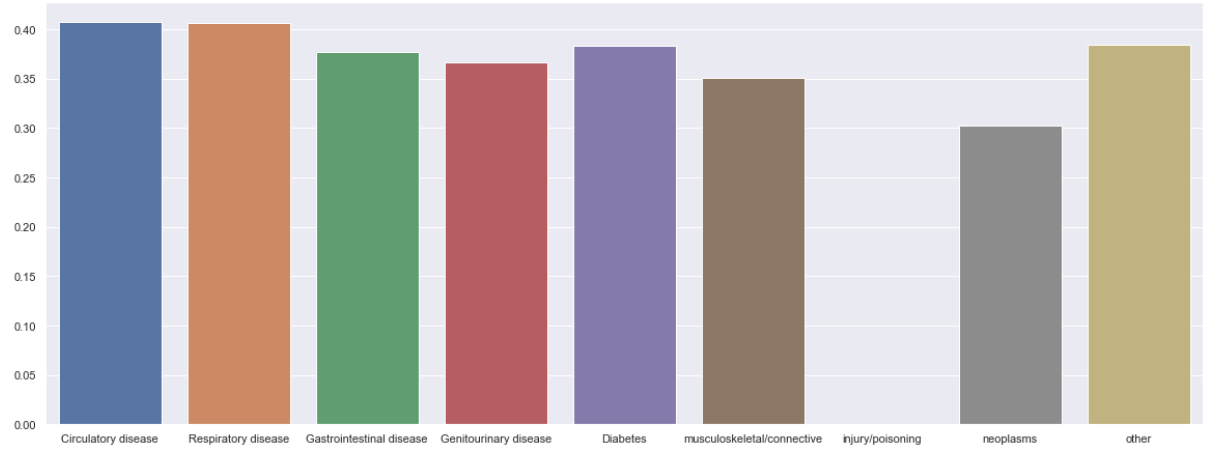

Figure 4 shows the readmission rates of diabetic patients based on the ICD codes. The readmission rates range between the lowest 30.3% (neoplasms) and highest 40.8% (circulatory disease)

Based on figure 4 H4 would be rejected and H40 would be accepted, despite the 10% difference in readmission rate between neoplasms and circulatory disease the following ICD codes all have a readmission rate within 5% of each other: Circulatory disease, Respiratory disease, Gastrointestinal disease, Genitourinary disease, Diabetes, musculoskeletal/connective. This graph does not show the number of individuals and the ICD codes with most readmitted patients were circulatory disease with 7012 and Diabetes with 5247 patients, which is substantially more than neoplasms which only had 679 readmitted patients. This is in line with clinical data that has shown a 2 to 4-fold increased mortality risk from circulatory disease for diabetic patients, therefore it makes sense for it to be one of the most common causes for hospital readmission (Aronson and Edelman, 2014). Based on this dataset it is shown that ICD codes do not have a significant effect on readmission rates as they are similar across different ICD codes and all these conditions are relatively serious. However, certain diabetic related complications are more prevalent than others.

Figure 5, Basic Model Confusion Matrix:

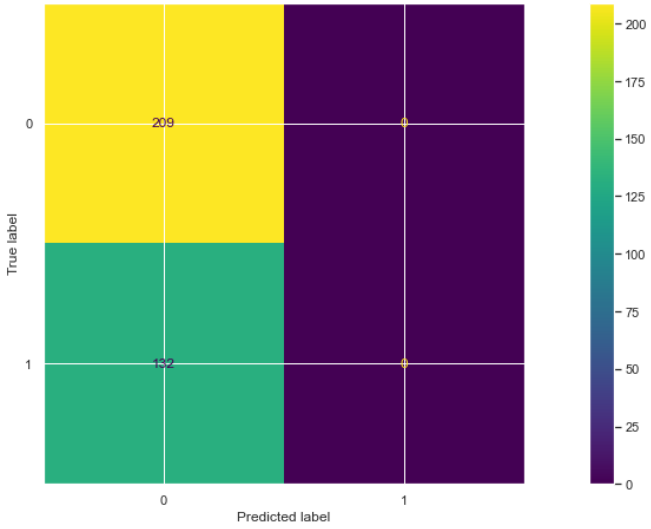

Figure 5 shows the confusion matrix for the initial model that was built, the model was built up on training data which had a score of 0.591 and then when it was applied to the test data it had a score of 0.628 The original data set was split up into training and test data sets after it had been cleaned. The True positive and false positive scores were 0 for the model, with a score of 214 for True negative and 127 for false negatives. The model was then cross validated which produced a cross validation mean score of 0.600 with an F1 score of 0.

Figure 6, Improved Model Confusion Matrix:

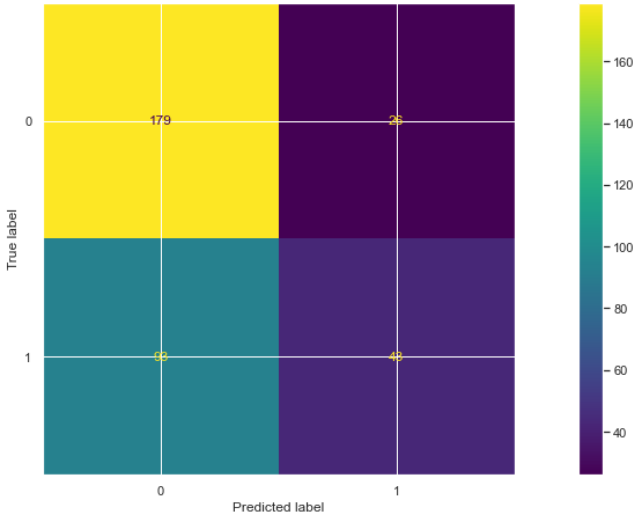

Figure 6 shows the confusion matrix for the improved model that was developed. This model varies from the original as it the encounter_id column was not utilized. Again, this model was built up on training and test sets which produced a training score 0.651 and a test set score of 0.651. The cross-validation score for this model was 0.645 with an F1 score of 0.420.

K-means clustering

K-means clustering is a method used to group data points based upon the observed values of a given set of variables for each data point wherein, for each data point, n-dimensional coordinates (where n is the amount of observations used for each datapoint) to determine distances between each given datapoint (Likas, Vlassis, and Verbeek, 2003) which allows the determination of similarity between clusters, and therefore data points, in an unsupervised way. In this instance, the R package *scikit-learn* was utilised to determine Euclidean distances between each data point in our dataset. Initial k-means clustering utilised all variables in our cleansed dataset. Min/max normalisation was carried out for each continuous variable to assure equal weighting for any dimensionality reduction methods used, and these values were used to compute *n*-dimensional distances for our datapoints.

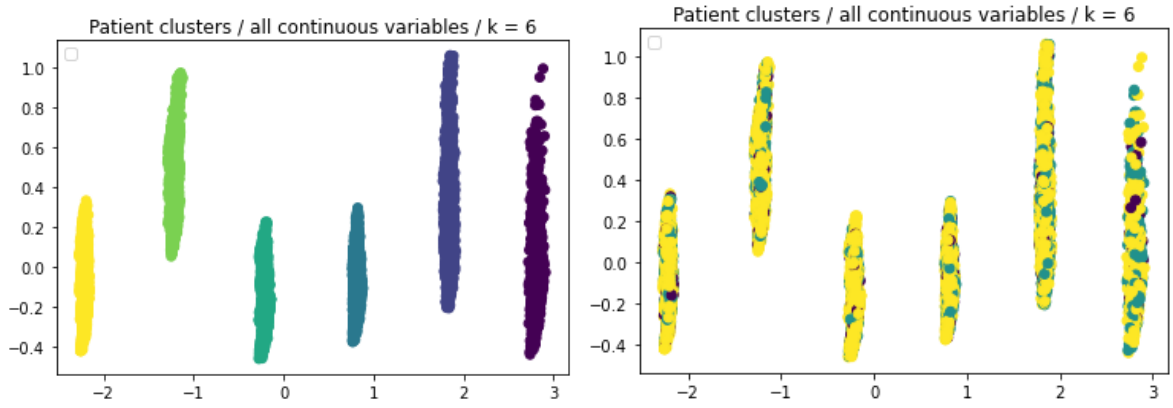

Figure 6a/b: 2D PCA representations of the clusters determined by our initial K-means model. Figure 6a (left) shows initial clusters of our data (J > 3000), figure 6b (right) shows readmittance data (in yellow) overlayed onto the clusters.

The first model was constructed using a *k*-value of 6, supported by "elbow method" of plotting J-scores for each *k*-value and determining which number of clusters provides the greatest proportional reduction of its associated J-value. This initial model showed very poor to non-existent clusters derived from our cleaned dataset (J-score > 3000). This model does not provide much insight into the question at hand, due to its poor quality, as can be noted by the "random" spread of readmission throughout the clusters. Furthermore, upon comparing the distribution of our clusters with the various distributions in our data, we found very few similarities between the distribution of our clusters and the distribution of other variables, such as readmittance or race. Stronger, more accurate clusters might show most readmitted patients being found in a few of the clusters which would allow for useful implementation of this data.

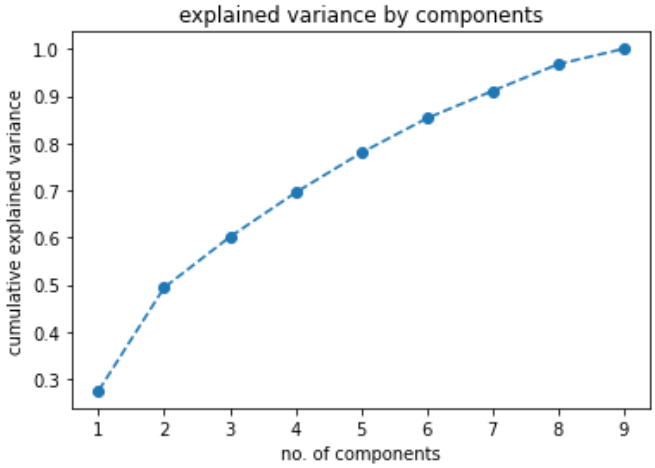

Figure 7 (above): Cumulative variance plot, showing the total variance (by percent) of the model that is explained by each feature of the model. 5 features explain ~80% of our model, so that value was selected as the optimal number of features.

To determine labels that cluster more tightly according to our data, dimensionality reduction was then performed using PCA (Pearson, 1901). First, the column 'encounter_id' was dropped from the data frame, leaving 9 continuous variables being used to compute the PCA. A cumulative variance plot was then used to determine the number of features (observations/columns) required to explain at least 80% of the variance in our data. In this dataset, the plot suggested that optimal components for PCA was 5, so 5 were implemented into the PCA. Dimensionality reduction and further data cleansing efforts reduced the J-score of the new clustering model to beneath 2000, yet despite this the score remained much higher than one would expect for significantly valid clusters. Thus, this suggests that even the second model would be of little use and provide little with regards to using this data and analysis pipeline for real world, clinical implementation of this data. Further research might prioritise a more sophisticated data cleansing stage to produce more useful final data. Data mining for more comprehensive data on each patient could provide variables that might more accurately explain the variance in readmittance between our observed data points, which in turn might also aid the development of more useful models that can be implemented in predicting and reducing patient readmittances.

Conclusion:

The analysis and data exploration that was based on the diabetic dataset produced various results that may provide use to the healthcare industry. The initial data exploration that was done produced various graphs that proved and disproved the hypotheses that were initially identified. The first hypothesis was proven, as the right graph in figure 1 shows a negatively skewed plot with regards to age and diabetic patients' readmission. This suggests that a patient is more likely to be readmitted the older they are, which as mentioned earlier has been corroborated by clinical studies (Leung, Wongrakpanich and Munshi, 2018).

In terms of ethnicity and readmission rates, null hypothesis 2 was accepted as figure 2 clearly shows that African Americans are not more likely to be re-admitted than other ethnic groups. Instead, the analysis identified Caucasians as the ethnic group with highest readmission rate 38.3%, compared to African American 33.0%. This is an unusual finding considering that people from African and Caribbean backgrounds have an increased susceptibly to diabetes and therefore would be expected to have higher readmission rates. As mentioned earlier, this could be due to overrepresentation of Caucasians compared to other ethnic groups. Alternatively, this could be due to possible environmental and socioeconomic factors as minority ethnic groups have a lower socioeconomic status compared to Caucasian Americans. Diagnosed diabetics incur average medical expenditures of $16,752 per Anum and it is possible that the hospital costs have a larger influence on readmission rates than ethnicities (The Cost of Diabetes | ADA, 2022).

Hypothesis three was disproven, and null Hypothesis three was accepted as figure three identified there was negligible 1.4% difference in readmission rates between female and male patients. While this initially may seem counterintuitive due to the 3:2 male: female diabetes ratio identified in European populations and one would expect potentially more men to be readmitted, this is not the case. This suggests that while males might be more susceptible to diabetes diagnosis, in terms of potential hospital readmission male and females are almost identical based on the analysis of this data set.

Based on the analysis of the ICD codes and readmission rates, null hypothesis four is accepted and hypothesis four is rejected. Despite the 10% in readmission rate between circulatory disease and neoplasms, the other ICD codes all fell within a 5% readmission rate with each other. With Circulatory disease 40% and respiratory disease 39% having a very similar rate, The ICD codes all represent serious complications associated with diabetes therefore it reasonable to assume they have a similar rate of readmission. What was identifiable was that within the dataset certain ICD codes are more prevalent than others, circulatory disease was the ICD code for 7012 patients while only 679 readmitted patients had an ICD code for neoplasms.

Our *k*-means analysis demonstrated the importance of having correctly cleaned and formatted data. The clusters produced were very “loose” with high J-scores indicating large distances between each data point and its centroid. As aforementioned, further research may prioritize data cleaning or data mining to produce a more useful model with clusters corresponding to real labels that group the data in a logical way. Other data points may explain a greater proportion of the variance in our dataset, and thus could be useful in the construction of models. The same is true for our improved models: whilst we saw minimal improvements in model scores, we did not achieve a score of greater than 0.7, suggesting that a more effective model that produces results from which real-world implications can be devised and actioned.

Overall, the usefulness of this dataset in providing a meaningful analysis for diabetic patients is limited, and in terms of clinical significance there is little. While the various hypotheses were proven or disproven based on the graphs, this is not conclusive evidence and doing a statistical analysis to determine whether differences are statistically significant would provide more meaning to the hypotheses. The reason for the low clinical significance is that there are too many variables that are not related to diabetes influencing a patient's possible readmission. One example could be that a patient cannot afford the cost of being readmitted and therefore decides to forgo medical intervention possible leading to more serious complications. An analysis based on a dataset with fewer variables could provide more clinical and statistically significant results. Further studies should include a metric of obesity such as weight, BMI or Body fat percentage as well as a metric of activity levels as these are the two main factors associated with type 2 diabetes which is the most prevalent form. If future studies are conducted an emphasis should also be placed on accurate representation and this will influence the statistical power of a study.

References:

- Aronson, D. and Edelman, E., 2014. Coronary Artery Disease and Diabetes Mellitus. *Cardiology Clinics*, 32(3), pp.439-455.
- Barski, L., Harman-Boehm, I., Nevzorov, R., Rabaev, E., Zektser, M., Jotkowitz, A., Zeller, L., Shleyfer, E. and Almog, Y., 2011. Gender-Related Differences in Clinical Characteristics and Outcomes in Patients with Diabetic Ketoacidosis. *Gender Medicine*, 8(6), pp.372-377.
- *Diabetes Care*, 2018. Economic Costs of Diabetes in the U.S. in 2017. 41(5), pp.917-928.
- Harris, M., 1990. Noninsulin-Dependent diabetes mellitus in black and white Americans. *Diabetes / Metabolism Reviews*, 6(2), pp.71-90.
- Kitabchi, A., Umpierrez, G., Miles, J. and Fisher, J., 2009. Hyperglycemic Crises in Adult Patients With Diabetes. *Diabetes Care*, 32(7), pp.1335-1343.
- Leung, E., Wongrakpanich, S. and Munshi, M., 2018. Diabetes Management in the Elderly. *Diabetes Spectrum*, 31(3), pp.245-253.
- Petzold, A., Solimena, M. and Knoch, K., 2015. Mechanisms of Beta Cell Dysfunction Associated With Viral Infection. *Current Diabetes Reports*, 15(10).
- Pinchevsky, Y., Butkow, N., Raal, F., Chirwa, T. and Rothberg, A., 2020. Demographic and Clinical Factors Associated with Development of Type 2 Diabetes: A Review of the Literature . *International Journal of General Medicine*, Volume 13, pp.121-129.
- UK, D., diabetes, O. and trusted, T., 2022. *Older people and diabetes*. [online] Diabetes UK. Available at: <https://www.diabetes.org.uk/guide-to-diabetes/older-people-and-diabetes> [Accessed 23 March 2022].
- Web.archive.org. 2022. *WHO | About diabetes*. [online] Available at: <https://web.archive.org/web/20140331094533/http://www.who.int/diabetes/action_online/basics/en/> [Accessed 22 March 2022].
- Diabetes.org. 2022. *The Cost of Diabetes | ADA*. [online] Available at: <https://www.diabetes.org/about-us/statistics/cost-diabetes> [Accessed 24 March 2022].
- Gale, E.A.M. and Gillespie, K.M. (2001) 'Diabetes and gender', *Diabetologia*, 44(1), pp. 3-15.
- Gale, E.A.M. and Gillespie, K.M. (2001) 'Diabetes and gender', *Diabetologia*, 44(1), pp. 3-15.
- Sirugo, G., Williams, S.M. and Tishkoff, S.A. (2019) 'The missing diversity in human genetic studies', *Cell*, 177(4), pp. 1080.
- Pearson, K., 1901. Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *6*(2), p.559.
- Likas, A., Vlassis, N. and J. Verbeek, J. (2003) 'The global k-means clustering algorithm', *Pattern Recognition*, 36(2), pp. 451-461.