Diabetes Data set Report

Anna Tam, Erasmo Carrasco, Nazmul Islam Khan, Carolina Salas

California State University, Los Angeles

CIS 4200 – Business Intelligence and Data Warehousing

May 7, 2024

Part I Introduction

1. Project Introduction & Primary Objectives

Project Introduction:

The goal of this project is to analyze a dataset containing information about patients with diabetes. The dataset comprises demographic information on individuals diagnosed with diabetes, encompassing age, ethnicity, and gender, among other variables. Furthermore, the collection includes the patient encounter and patient number, which function to anonymously differentiate everyone. The dataset includes information regarding admission type, duration of hospitalization, changes in diabetes status, medication dispensation, and readmission to the medical facility. By examining various demographic and clinical variables, we aim to gain insights into the characteristics of diabetic patients and their interactions with healthcare systems. Our research's objective is to examine a dataset comprising significant information about individuals diagnosed with diabetes. Our focus is analyzing demographic characteristics, clinical variables, and readmission rates. Our principal goals are to optimize resource allocation, enhance patient care, and advocate for evidence-based decision making.

Primary Objectives:

The primary objective is to identify the characteristics that influence early patient readmission within 30 days (about 4 and a half weeks) of being discharged. Our objective is to gather valuable information that may be used to develop strategies for enhancing outcomes for patients and efficiently distributing healthcare resources. The items listed below are considered the primary objectives of our project.

- Time spent in hospital
- Demographic Prevalence
- Gender and age proportion

- Hospitalization Frequence by demographics
- Medication Usage and Racial Groups
- Readmission Rates
- Hospital Stay Duration
- Gender-Specific Hospital Stay
- Readmission Rates by Demographics

Overview of Dataset

This dataset called "Diabetes" holds information on patients diagnosed with diabetes. Information included is demographic details like race and gender, along with age categories. Other provided information are hospital admission specifications, plus admission type and duration of hospital stay for each patient. In addition, medication data indicates any change in medication or if diabetes medication was prescribed to patient. This dataset also includes information of the readmission status of patients. The dataset facilitates comprehensive insights into diverse aspects within diabetes care and management. By analyzing its variables, healthcare practitioners and researchers acquire insights into effectives of diabetes treatments, detect patterns, enhance patient care and contributing to lower diabetes rates.

Variables:

Encounter_id: A unique identifier for patient encounters.

Patient_nbr: A unique identifier for patients, allowing for tracking across multiple encounters.

Race: The categorization of patient's racial gender, which includes African American, Asian, Caucasian, Hispanic, and other

Gender: The classification of patient gender, distinguishing between Female and Male.

Age: This indicates the age range of the patient.

Admission_type_id: This is the classification of admission types; Emergency, Urgent, Elective, Newborn, Not Available, NULL, Trauma Center, and Not Mapped

Time_in_hospital: This shows the duration of each patient hospital stay during each encounter.

Change: Indicated whether there was a change in the patient diabetes medication during encounter

DiabetesMed: States if the patient was prescribed to diabetes medication

Readmitted: Specify if the patient was readmitted to patient

Responsibilities

Nazmul Islam Khan – Project Introduction & Primary Objectives

Carolina Salas – Overview of Dataset, Research Questions & Analytical Tools Employed and Analysis & Visualizations

Erasmo Carrasco – Research Questions & Analytical Tools Employed, Analysis & Visualization and Summary of Key Findings

Anna Tam – Limitations and Areas for Further Exploration

Part II Data Analysis and Discussion

Research Questions & Analytical Tools Employed

Question 1: Provide a summary statistic for the time patients with diabetes spent in hospital. Create a box plot to display the distribution of time spent in hospital and analyze your findings.

Analytical Tools Employed: *Excel (summary statistics & box plots)*

Question 2: Which demographic group exhibits the highest prevalence of diabetes?

Analytical Tools Employed: *Excel (Pivot Tables, Pivot Bar Graph & Pie Chart)*

Question 3:

Part 1: Identify the total number of both the female and male patients with diabetes and provide the percentage for each using pivot table and pie chart.

Part 2: What is the proportion of age within female sample? What is the proportion of age within male sample? Utilize a pivot table to demonstrate the proportion of both samples and a bar chart to demonstrate the distribution of both.

Analytical Tools Employed: *Excel (Pivot Table & Pie Chart)*

Question 4: What demographic groups are categorized by age, and race? Which group has the highest frequency of hospitalization? What are the respective counts for each race group?

Analytical Tools Employed: *Excel (Pivot Table, Pivot Line Chart)*

Question 5: Are there any racial groups where patients with diabetes medication outnumber those without? Which racial groups exhibit the highest counts of patients with diabetes medication and how does it compare to the other groups.

Analytical Tools Employed: *Excel* (Pivot Tables, Pivot Bar Graph)

Question 6: What is the overall readmission rate for diabetic patients? What portion of readmissions occurs within 30 days?

Analytical Tools Employed: *Tableau* (Pie Chart)

Question 7:

Part 1. Are there differences in hospital stay duration between younger and older age group

Part 2. Are there any differences in hospital stay duration among different racial demographics?

Analytical Tools Employed: *Tableau (Bar Graph)*

Question 8: Describe the notable trends in hospitalization duration, considering both advancements in medicine and variations across different age groups

Analytical Tools Employed: *Tableau (Side-to-Side Bar Graph)*

Question 9: How do readmission rates vary across different age groups and racial demographics?

Analytical Tools Employed: *Tableau (Bar Graph)*

Question 10: How does the average time spent in the hospital vary between different admission types for each gender? Compare both the male and female sample and explain your findings.

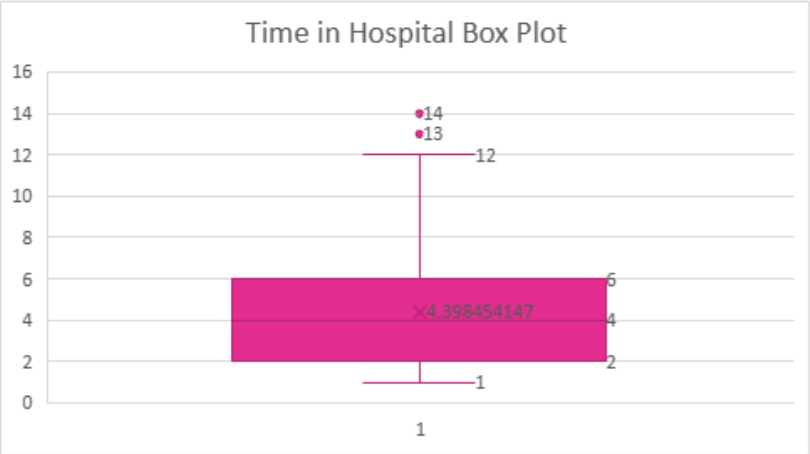
Analytical Tools Employed: *Tableau (Side-to-Side Bar Graph)*

Analysis & Visualizations

Question 1:

Descriptive Statistics (Time in Hospital)	
Mean	4.398454147
Standard Error	0.00946972

Median	4
Mode	3
Standard Deviation	2.98697235
Sample Variance	8.922003822
Kurtosis	0.845957215
Skewness	1.133256483
Range	13
Minimum	1
Maximum	14
Sum	437611
Count	99492



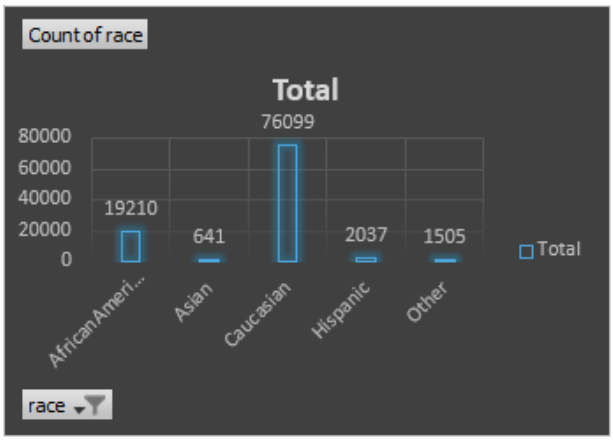
Analysis –

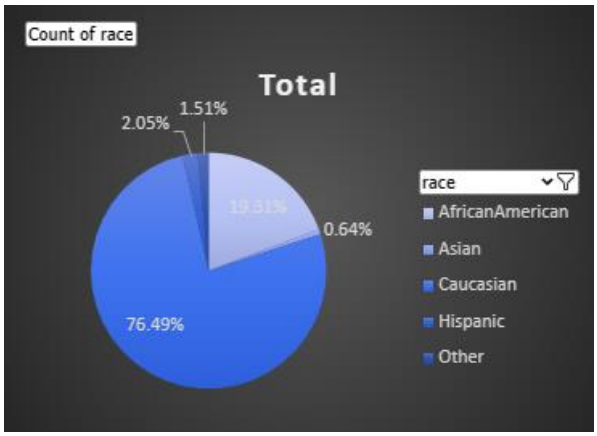
The summary statistics and box plot depicted above demonstrates the time spent (in days) by patients with diabetes in hospital. The measures of location include the mean (approximately 4.4), median (4), and mode (3). The measures of spread are listed which are standard deviation (approximately 2.99), variance (8.92), and range (13). The box plot displays the spread of the time spent in the hospital where the inter-quartile range (IQR) is towards the bottom. There are visible outliers which include 13 and 14 and the IQR is revealed to be 4 hours. Therefore, the average amount of time that patients diagnosed with diabetes spend in a hospital is a little over 4 days.

Question 2:

Row Labels	Count of race
African American	19210
Asian	641
Caucasian	76099
Hispanic	2037
Other	1505
Grand Total	99492

Row Labels	Count of race
African American	19.31%
Asian	0.64%
Caucasian	76.49%
Hispanic	2.05%
Other	1.51%
Grand Total	100.00%





Analysis –

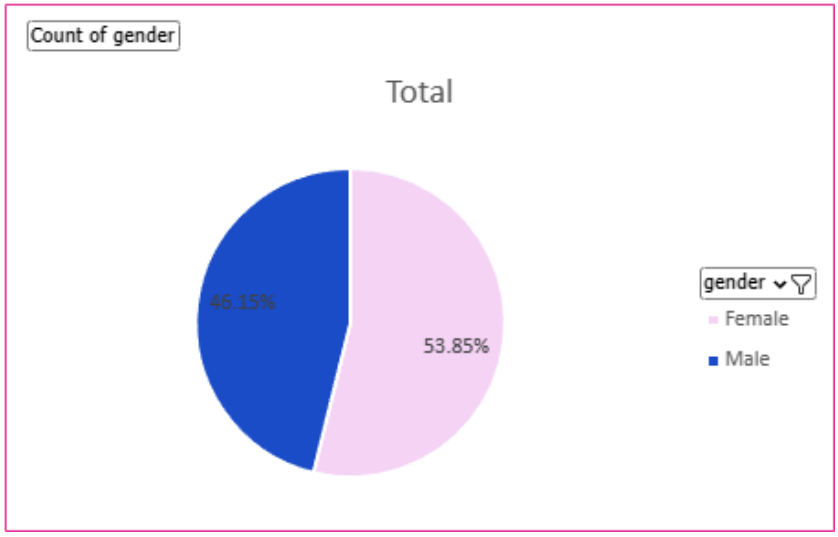
This analysis presents the distribution of diabetes cases across various racial groups. In the pivot tables and charts, it becomes evident that Caucasians stand out with the highest count, comprising approximately 76.49% of the total cases. Following them, African Americans emerge as the second runner-up, accounting for 19.31% of the cases. Meanwhile, Hispanics, Asians, and individuals from other racial backgrounds exhibit lower incidences of diabetes, ranging from 1% to 2% of the total cases.

In the bar chart, Caucasians are depicted with a count of 76,099 diabetes cases, the highest among racial groups. Similarly, the pie chart on the left illustrates Caucasians dominating with 76.49% of the total cases. Notably, Asians exhibit the lowest incidence of diabetes, accounting for just 0.64% of cases, equivalent to 641 cases, as depicted in the bar graph on the left.

Question 3:

Part 1 -

Row Labels	Count of Gender	% Count of Gender
Female	53575	53.85%
Male	45917	46.15%
Grand Total	**99492**	**100.00%**

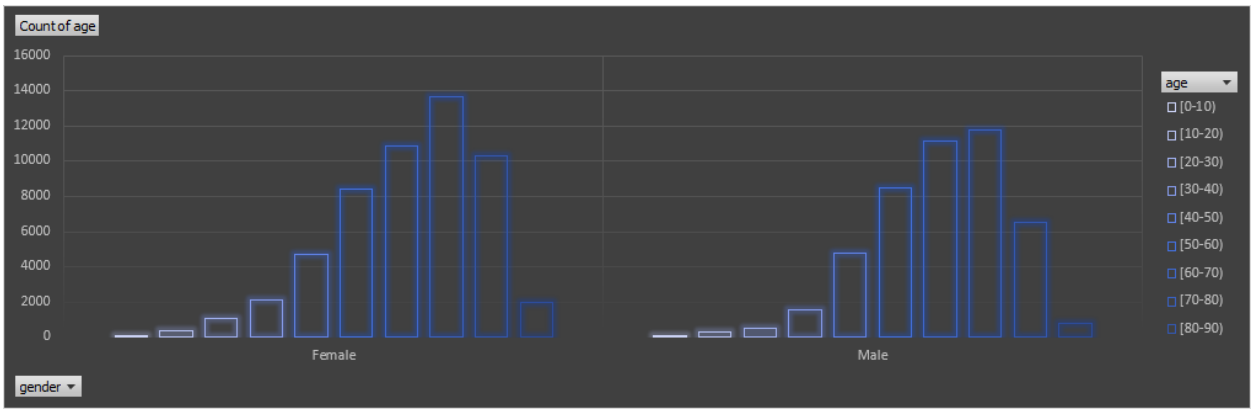


Analysis –

This gender analysis reveals that females have a higher count of diabetes comprising 53.85% of cases, compared to males who represent 46.15%. Examining gender distribution sheds light on gender-specific health outcomes within the diabetic population. This visual representation illustrates the gender distribution of diabetes cases, with females on the right accounting for 53.85%. Males on the left of the pie chart represent about 46.15%.

Part 2 –

Count of age											
Row Labels	**[0-10)**	**[10-20)**	**[20-30)**	**[30-40)**	**[40-50)**	**[50-60)**	**[60-70)**	**[70-80)**	**[80-90)**	**[90-100)**	**Grand Total**
Female	82	399	1077	2119	4718	8411	10843	13679	10284	1963	53575
Male	78	283	534	1580	4747	8484	11145	11789	6516	761	45917
Grand Total	**160**	**682**	**1611**	**3699**	**9465**	**16895**	**21988**	**25468**	**16800**	**2724**	**99492**



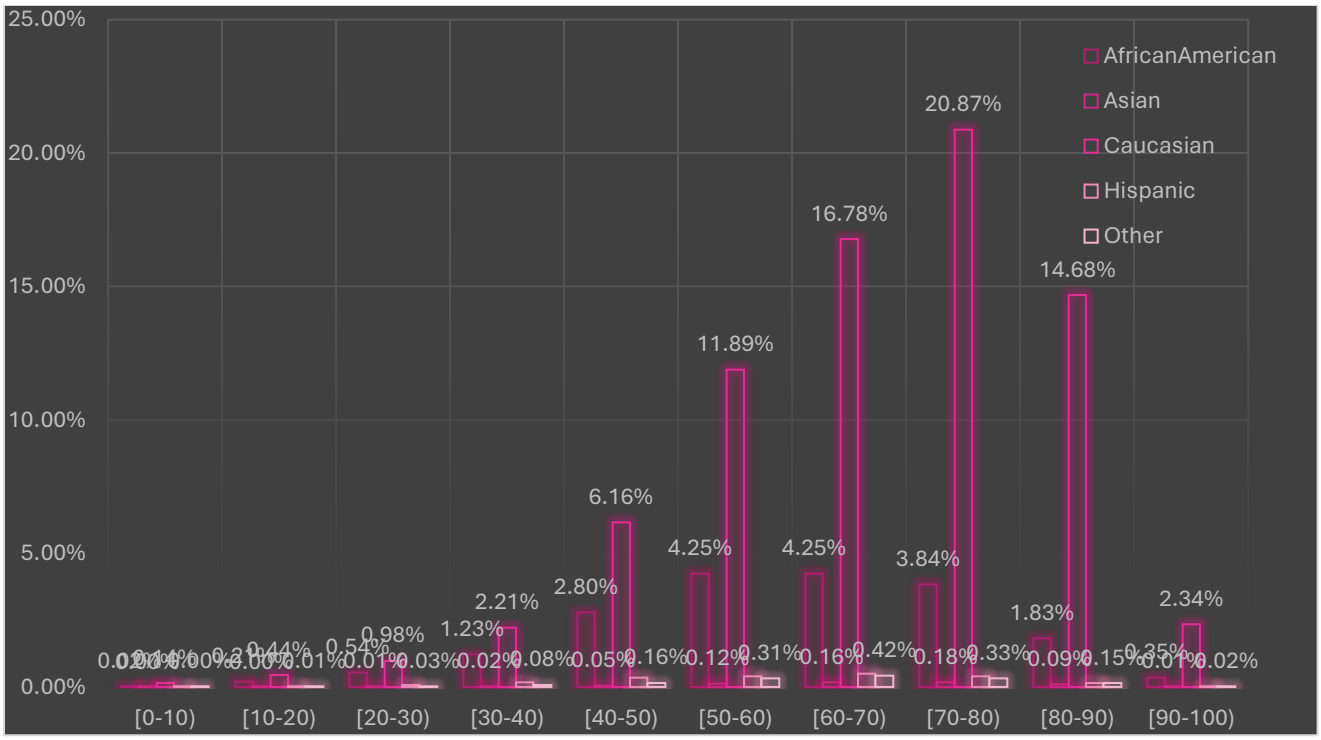
Analysis –

The pivot table demonstrates the proportion of age among both the female and male patients in which there are 10 age groups. For the female sample, the category of age that has the highest number of patients is the category of [70 – 80] with 13,679 female patients. As for the male sample, the age group with the highest number of patients is the category of [70 – 80] with 11,789 male patients. Based on the information provided by the table, the age group between [70 – 80] has the highest number of patients with diabetes which make up 25,468 patients exceeding the other age groups.

The histogram above displays the distribution of age groups of both, the male and female samples. Both histograms are left-skewed showing that most of the patients with diabetes are typically in a higher and much older age group. From the results, there is a common trend of age in which patients are diagnosed with diabetes. Patients between 80-90 are at higher risk of having diabetes as depicted from the observations revealed in compared to either female or males in their early ages.

Question 4:

Count of time_in_hospital Row Labels	Column Labels AfricanAmerican	Asian	Caucasian	Hispanic	Other	Grand Total
[0-10)	0.02%	0.00%	0.14%	0.00%	0.00%	0.16%
[10-20)	0.21%	0.00%	0.44%	0.02%	0.01%	0.68%
[20-30)	0.54%	0.01%	0.98%	0.06%	0.03%	1.62%
[30-40)	1.23%	0.02%	2.21%	0.17%	0.08%	3.72%
[40-50)	2.80%	0.05%	6.16%	0.34%	0.16%	9.51%
[50-60)	4.25%	0.12%	11.89%	0.41%	0.31%	16.98%
[60-70)	4.25%	0.16%	16.78%	0.49%	0.42%	22.10%
[70-80)	3.84%	0.18%	20.87%	0.39%	0.33%	25.60%
[80-90)	1.83%	0.09%	14.68%	0.14%	0.15%	16.89%
[90-100)	0.35%	0.01%	2.34%	0.02%	0.02%	2.74%
Grand Total	**19.31%**	**0.64%**	**76.48%**	**2.05%**	**1.51%**	**100.00%**

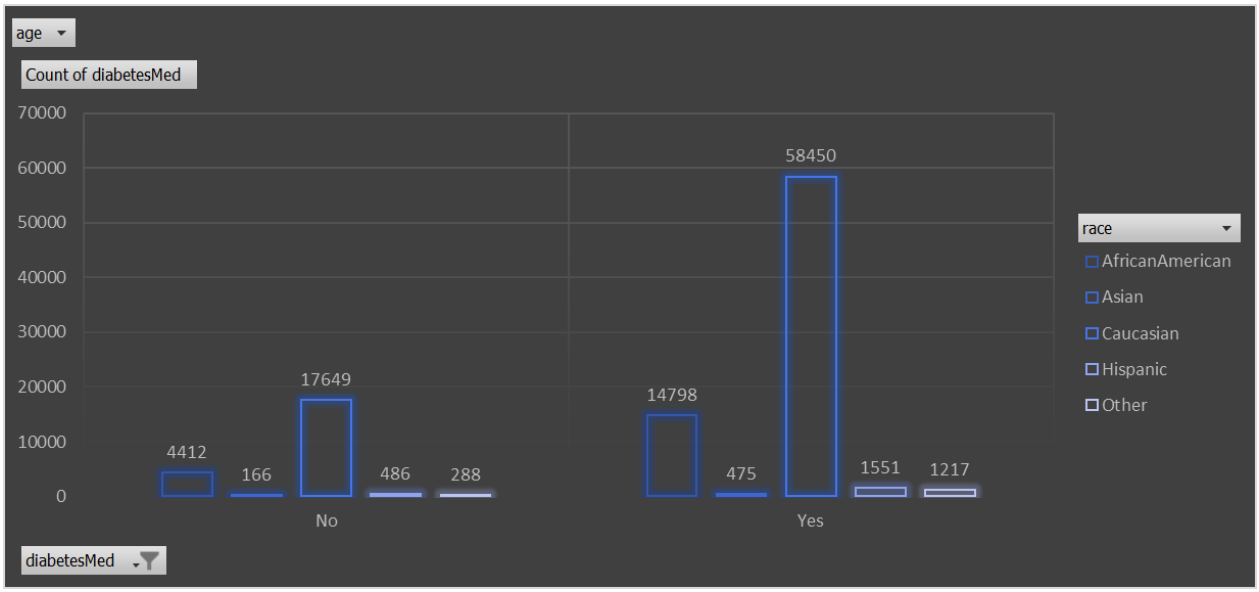


Analysis –

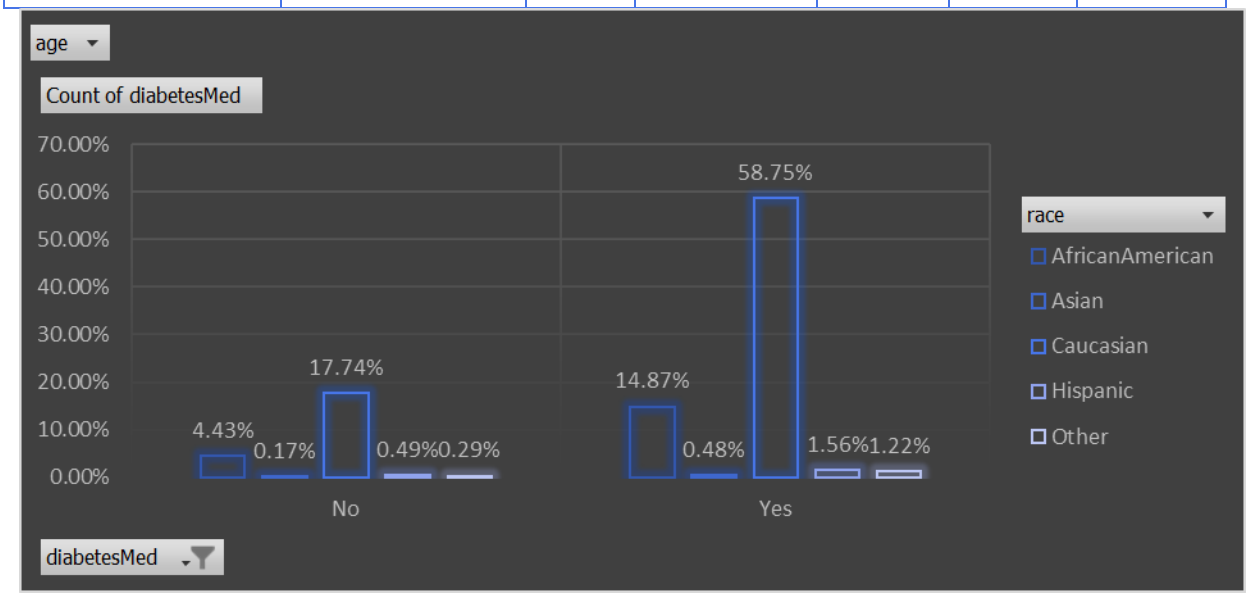
The top table shows that the rate of hospitalizations was higher among Caucasians in the 70-80 age bracket compared to all other age groups. Hospitalization rates for Caucasians are consistently higher throughout all age categories, with a peak between 70 and 80 years old, as shown clearly in the bottom-line figure (graph). Hospital stays differ greatly by age and race, with a larger proportion of Caucasians in the older age bracket, according to the results.

Question 5:

age	(All)					
Count of diabetesMed	Column Labels					
Row Labels	AfricanAmerican	Asian	Caucasian	Hispanic	Other	Grand Total
No	4412	166	17649	486	288	23001
Yes	14798	475	58450	1551	1217	76491
Grand Total	19210	641	76099	2037	1505	99492



age	(All)					
Row Labels	**African American**	**Asian**	**Caucasian**	**Hispanic**	**Other**	**Grand Total**
No	4.43%	0.17%	17.74%	0.49%	0.29%	23.12%
Yes	14.87%	0.48%	58.75%	1.56%	1.22%	76.88%

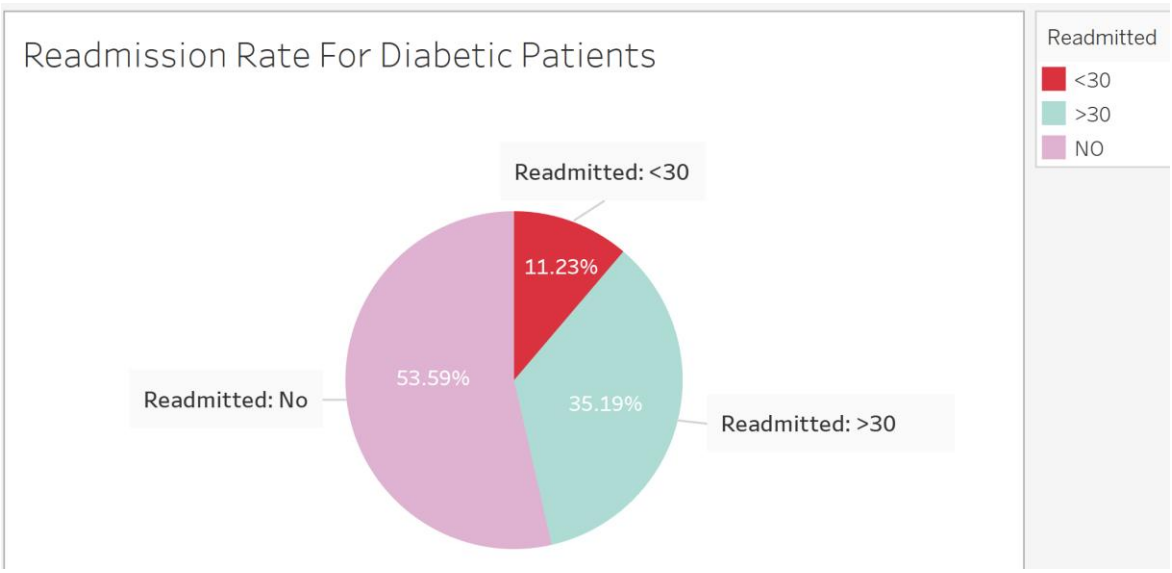


Analysis –

Caucasian patients comprising the highest proportion of individuals on diabetes medication, accounting for 58.75% of those with medication. While Asian patient have the lowest percentage of 0.48% of using medication. Despite the high amount of user being white using medication, Caucasian as well still hold the height spot for not being on medicating at 17.74%.

This bar graph above depicts the analysis of individuals who are either on mediation or not. Asians are standing at 0.48% user do use medication for their diabetics while 0.17% of Asians do not use medication.

Question 6:



Dimensions – Readmitted

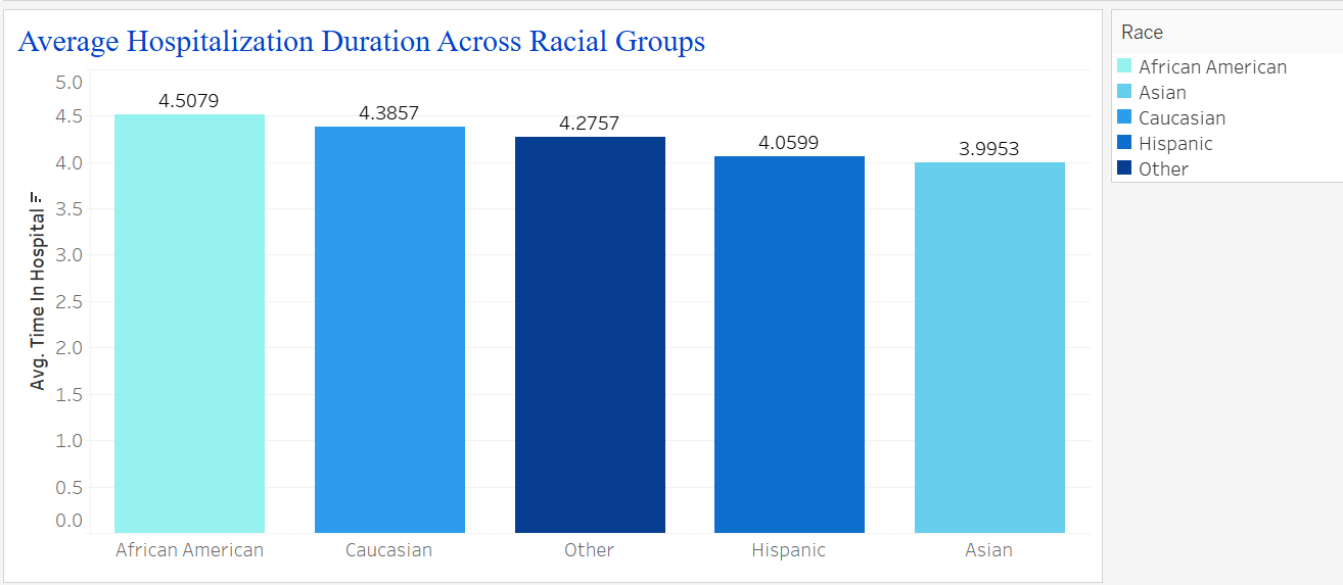
Measures – COUNT ([Readmitted])

Analysis –

The above pie chart depicts the readmission rates among all diabetic patients. As revealed, 53.9% of the total patients were not readmitted to the hospital while 11.23% were readmitted less than 30 times, and 35.19% were readmitted more than 30 times. It is depicted that there is slightly higher percentage of patients not being readmitted into the hospital at about 53.59% while 46.42% of patients are still readmitted.

Question 7:

Part 1 –



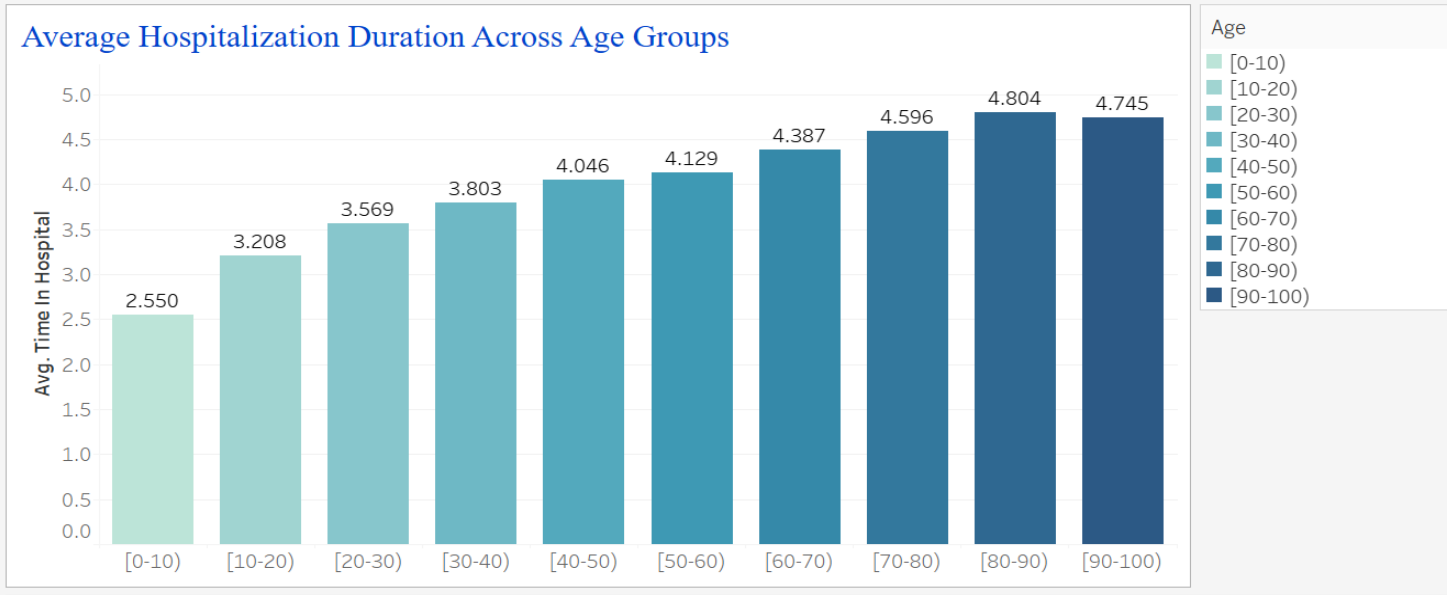
Dimensions – Race

Measures – AVERAGE([Time In Hospital])

Analysis –

The graph depicts the average time in hospital according to different racial demographics. As noted by the bar graph, the African-American group has the highest average time spent in the hospital which is about 4.5079 hrs. The group with the least average time spent in hospital is depicted as Asian group. This signifies that race may impact the time spent in hospital due to racial demographics.

Part 2 –



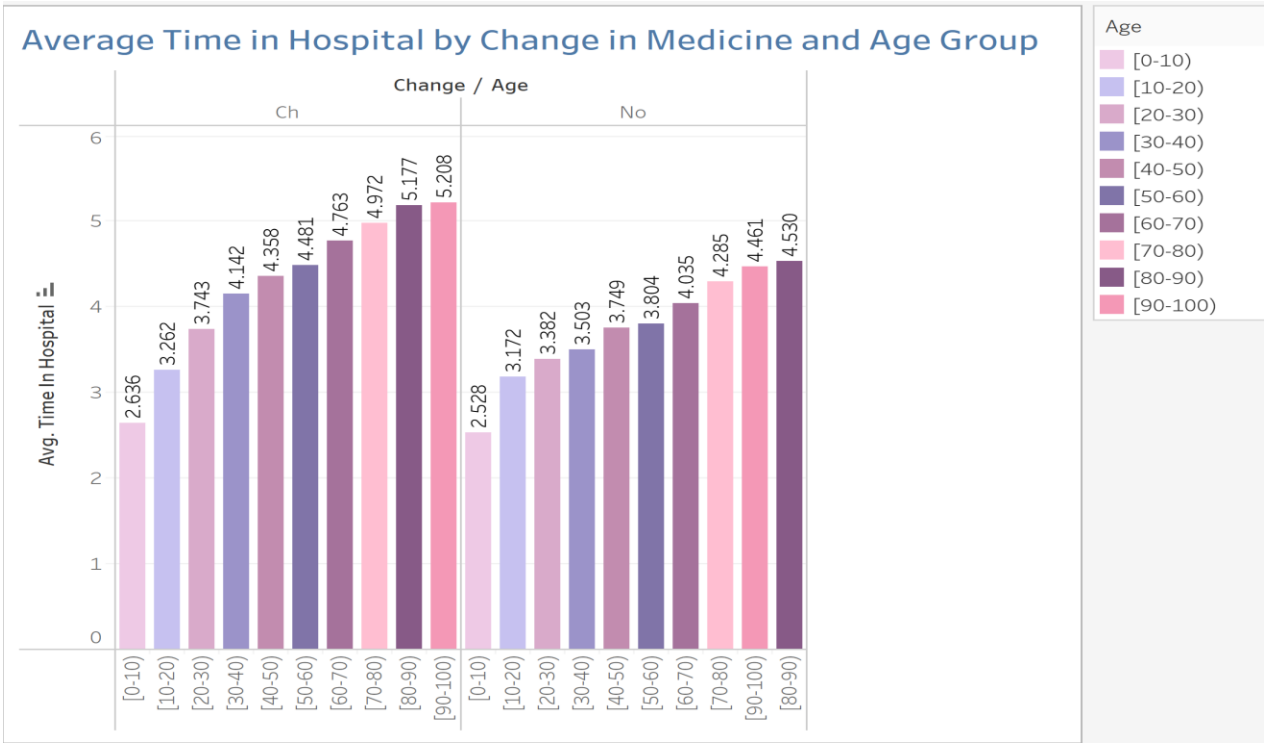
Dimensions – Age

Measures – AVERAGE([Time In Hospital])

Analysis –

The graph depicts the average time in hospital offers insights into patients stay. The lighter shade of blue indicates shorter durations, while the darker hue signifies a longer stay. From the bar graph, individuals ages 50 years and above exhibit a continuously higher length of stay compared to their younger counterparts. This analysis allows healthcare workers to tailor healthcare intervention for older patients.

Question 8:



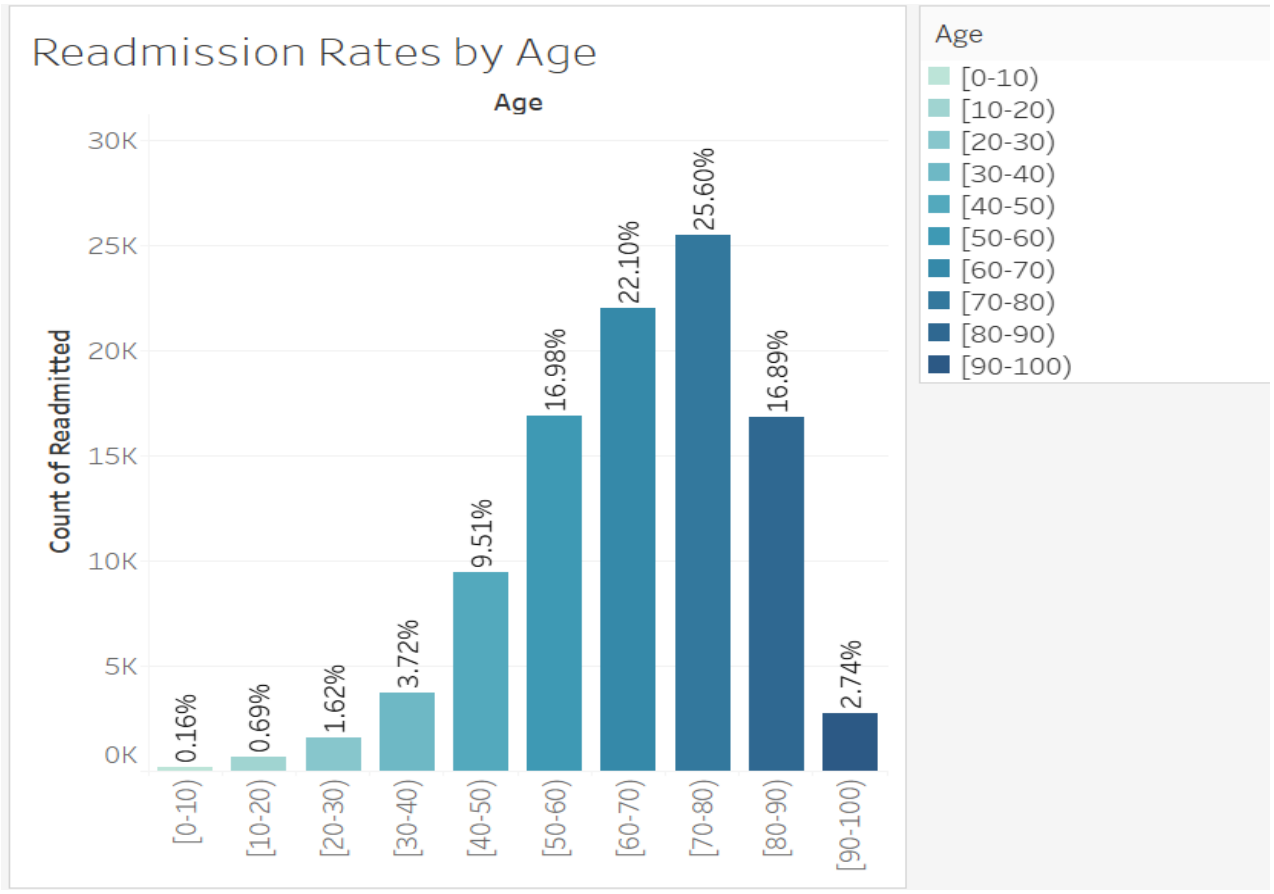
Dimensions – Change, Age

Measures – AVERAGE ([Time In Hospital])

Analysis –

The chart describes the average time hospitalized by if there was a change in medicine and by age group. Overall, patients who did not change medicine had a lower average time spent in hospital compared to their age group counterparts who did change their medication. As the age group increases, the time hospitalized also increases. On average, those who did change their medication had an average of 4.75 time spent in hospital and those who did not change their medication had an average of 4.1 time spent in hospital.

Question 9:

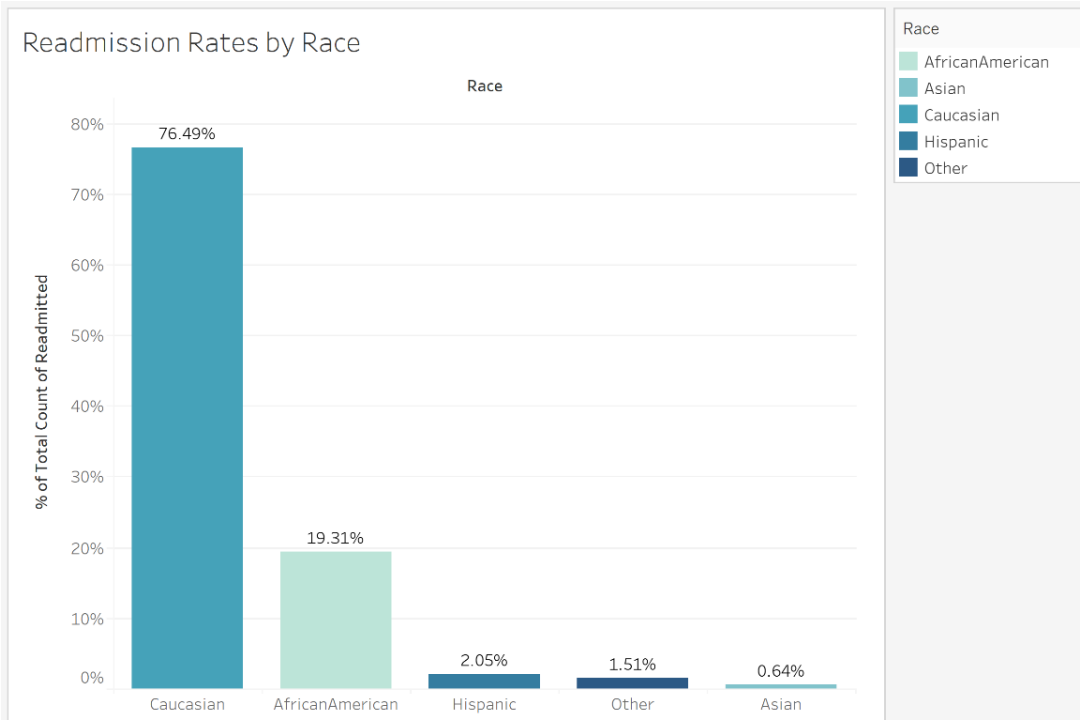


Dimensions – Age

Measures – COUNT([Readmitted])

Analysis –

The chart above displays the percentage of the total number of patients being readmitted based on age. From what is shown, the age group between 70 – 80 has the highest readmission rate at about 25.60% being readmitted. The age group between 0 -10 has the lowest readmission rate at about 0.16%. It is noted that patients diagnosed with diabetes between the ages 70 – 80 are more likely to be readmitted into the hospital.



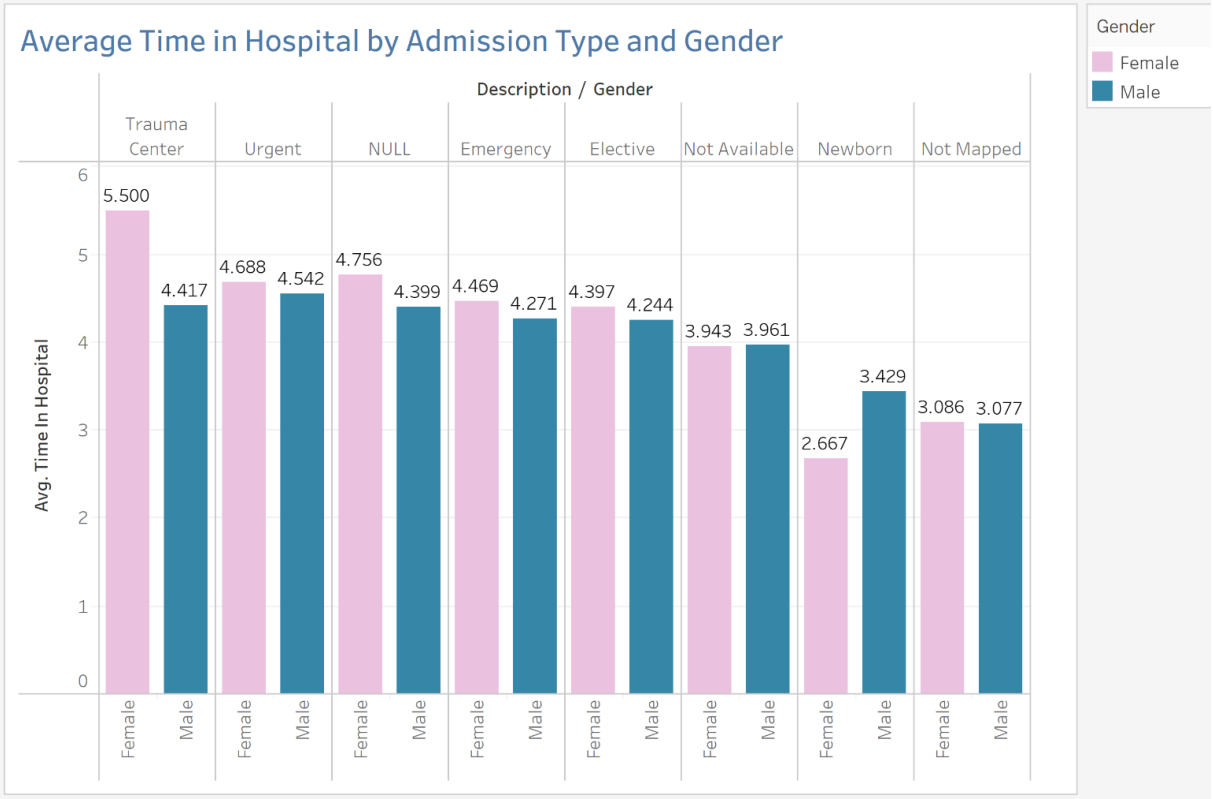
Dimensions – Race

Measures – COUNT([Readmitted])

Analysis –

The bar chart above shows the readmission rates of of patients diagnosed with diabetes varying by different races. The Caucasian group has by far the highest readmission rates with about 76.49% being readmitted while the lowest readmission rates come from the Asian group with about 0.64%. Thus, Caucasians diagnosed with diabetes have a higher chance of being readmitted compared to the groups as revealed by the bar chart.

Question 10:



Dimensions – Admission Type ID, Gender

Measures –AVERAGE([Time In Hospital])

Analysis –

The chart describes the average time hospitalized by admission types and gender. It can be seen that in the following admission types that Elective, Emergency, Not Mapped, NULL, Trama Center, and Urgent that females have a longer hospitalization time and the Newborn, and Not Available have a longer hospitalization time for males. The most notable difference is Trama Center with a difference of 1.083 between genders.

Summary of Key Findings

The research questions composed to analyze the data provided by the diabetes dataset along with the visualizations created, unfolded details to better support hospitals and patients diagnosed with diabetes. The analysis provided by the visualizations demonstrated that the females make up the greater part of patients and the age group between 70 – 80 is also depicted as the largest number of patients.

Readmission rates for the overall number of diabetic patients revealed that 53.59% of patients are not being readmitted. 11.23% of patients were readmitted less than 30 times while 35.19% were readmitted more than 30 times. However, there is still a greater number of patients not being readmitted compared to those that are. When evaluating readmission rates specific to demographics, Caucasians have the highest percentage of readmission rates while Asians have the least. As for the age factor in readmission rates, the 70 – 80 group leads the highest. As far as age, higher age groups are more inclined to have elevated readmission rates and for race, Caucasians are depicted as having a higher average time spent in hospital and inclined to more readmission rates. The research questions employed helped retain information regarding readmission rates were characteristics like demographics, length of stay, and medication impact the enhancement of patient outcome and healthcare resources.

Limitations & Areas for Further Exploration

While our research provided insight into key findings, there were several limitations we encountered. First there was insufficient data on a variable that may have prevented more insight into our research. For example, patterns of medicine usage (including if patients were taking medication and changes in medication, etc.) had an impact on the length of hospitalization across different dimensions however more details about this variable may provide additional insight. Additionally, there may have been a selection bias with the data provided if the data was received from one more multiple hospital. This may have affected confounding variables such as race, gender, and age (by various socio-economic factors from their location(s)), and admission type and time in hospital (by differences in severity of illness and treatment received). Areas of further exploration include how types of diabetes medicine, patterns of medicine usage, and services received in each admission type affect readmission and length of hospitalization.