# Diabetes inferential and descriptive analysis

Prepared by:

Emad Aqel

Osamah Alshaikh

Osamah Fayad

Walid Iramly

Supervised by:

Dr. Bashar Alshouha

# Contents



# Figures

## Tables

# Introduction

This report presents an in-depth analysis of a comprehensive dataset spanning ten years (1999-2008) of clinical care from 130 US hospitals and integrated delivery networks. The dataset specifically focuses on hospital records of patients diagnosed with diabetes who underwent various laboratory tests, received medications, and had hospital stays of up to 14 days. The primary objective of this study is to meticulously analyze the inherent features within this dataset and to extract hidden patterns that can provide valuable insights into diabetes management within hospital environments.

The significance of this research stems from a critical observation: despite the availability of high-quality evidence demonstrating improved clinical outcomes for diabetic patients who receive appropriate preventive and therapeutic interventions, a substantial number of patients do not receive these crucial treatments. This disparity can be partially attributed to inconsistent or arbitrary diabetes management practices within hospital settings, often leading to suboptimal glycemic control. The failure to provide proper diabetes care not only escalates managing costs for hospitals due to increased readmission rates but also profoundly impacts the morbidity and mortality of patients, who may experience severe complications associated with poorly managed diabetes.

# Dataset Summary

The dataset, comprising 101,766 instances and 47 features, is categorized under the subject areas of Health and Medicine. It is suitable for various analytical tasks, including classification and clustering.

The features within the dataset are a mix of categorical and integer types.

| Subject Area | Associated Tasks | Feature Type | Instances | Features |
|---|---|---|---|---|
| Health, Medicine | Classification, Clustering | Categorical, Integer | 101766 | 47 |

Table 1

# Data Description

The dataset includes various columns, each providing specific information about patient encounters and medical details. A detailed description of key columns is provided below:

| Column Name | Type | Description | Missing Values |
|---|---|---|---|
| Encounter_id | Categorical | Unique identifier of an encounter | No |
| Patient_nbr | Categorical | Unique identifier of a patient | No |
| Race | Categorical | Patient's race (Caucasian, Asian, African, American, Hispanic) | Yes |
| Gender | Categorical | Patient's gender (Male or Female) | No |
| Age | Categorical | Age of the patient | No |
| Time in hospital | Integer | Number of days between admission and discharge | No |
| Number procedures | Integer | Number of procedures (other than lab tests) performed during the encounter | No |
| Number lab procedures | Integer | Number of lab tests performed during the encounter | No |
| Num_medications | Integer | Number of distinct generic names administered during the encounter | No |
| Insulin | Categorical | Indicates whether the drug was prescribed or if there was a change in dosage | No |
| Number of inpaitients | Integer | Number of inpatient visits of the patient in the year preceding the encounter | No |
| Num_diagnoses | Integer | Number of diagnoses entered the system | No |

Table 2

# Data Cleaning and Preprocessing

The initial phase of this analysis involved comprehensive data cleaning and preprocessing to ensure the integrity and reliability of our findings. We began by importing the dataset, provided as a CSV file, into our analytical environment. A critical step was to identify and address missing

values, as their presence can significantly skew results. Initially, a standard is.nafunction was employed; however, this revealed no missing values, which was unexpected given the dataset's

source indicated otherwise. Further investigation uncovered that missing values were represented by "?" characters within the dataset, rather than standard NAindicators. Upon identifying these, we converted all "?" instances to NAvalues. This process revealed a substantial number of missing entries (192,858) across several key columns, specifically: weight, payer_code, medical_specialty, race, diag_1, diag_2, and diag_3.

To manage the extensive missing data, we strategically removed columns with a high proportion of missing values and other columns deemed irrelevant to our analysis using a selectfunction. Subsequently, a filterfunction was applied to remove rows containing any remaining missing values in the race, diag_1, diag_2, and diag_3columns, ensuring a robust dataset for

subsequent analysis.

```
# Load the necessary libraries
library(readxl)
library(dplyr)
library(ggplot2)
library(readr)
#import dataset
imported_dataset<-read_csv( file: "C:/Desktop2.0/Data Science Programming/diabetes+130-us+hospitals+for+years+1999-2008/diabetic_data.csv")
print(imported_dataset)
# Check the structure of the data set
str(imported_dataset)
#check for the missing values
any(is.na(imported_dataset))
sum(is.na(imported_dataset))
# Count how often '?' appears
sapply(imported_dataset,function(x) sum(x == '?', na.rm = TRUE))
# convert '?' to NA
imported_dataset[imported_dataset == '?'] <- NA
# Check for missing values again
sum(is.na(imported_dataset))
#Select specific columns and drop rows with missing values in certain columns
cleaned_imported_dataset <- imported_dataset %>%
  select(-weight, -payer_code, -medical_specialty,-A1Cresult,-metformin,-repaglinide,-nateglinide,-chlorpropamide,-glimepiride,
         -acetohexamide,-glipizide,-glyburide,-tolbutamide,-pioglitazone,-rosiglitazone,-acarbose,-miglitol,-troglitazone) %>%
  filter(
      !is.na(race) &
      !is.na(diag_1) &
      !is.na(diag_2) &
      !is.na(diag_3))
head(cleaned_imported_dataset)
sum(is.na(cleaned_imported_dataset))
```

Figure 1

Following this initial cleanup, a new dataset was created, entirely free of missing values. The data cleaning process continued with the creation of a subset focusing on relevant columns, which also involved addressing unnecessary categorical groups. A thorough check for duplicate entries was performed, confirming their absence. Outlier detection identified 90,508 outliers, which were handled appropriately. Finally, the data was scaled to normalize its range, preparing it for descriptive analysis.

```
#check for the unique values in the columns
cleaned_imported_dataset <- subset(cleaned_imported_dataset, gender!= "Unknown/Invalid"& gender!= "NO" & race != "Other" &race != "NO")
unique(cleaned_imported_dataset$race)
unique(cleaned_imported_dataset$gender)
#check for the duplicates
duplicates <- duplicated(cleaned_imported_dataset)
print(duplicates)
duplicates_count <- sum(duplicates)
print(duplicates_count)
#scale imported data
z_score <- scale(cleaned_imported_dataset[, numeric_columns])
print(z_score)
#check for the outliers
outliers <- cleaned_imported_dataset[apply(abs(z_score) > 1,  MARGIN: 1, any), ]
print(outliers)
nrow(outliers)
```

Figure 2

# Descriptive Analysis

For the descriptive analysis, our primary objective was to summarize the main features of the dataset. We first isolated the numerical columns using an is.numericfunction, applied across all columns via sapply. This allowed us to focus on quantitative variables. Subsequently, we computed measures of central tendency and measures of variability or spread for these numerical columns. These statistical summaries provide a foundational understanding of the data's distribution and characteristics.

```r
numeric_columns <- sapply(cleaned_imported_dataset, is.numeric)
print(numeric_columns)
numeric_stats<-data.frame(
  mean = sapply(cleaned_imported_dataset[, numeric_columns], mean),
  median = sapply(cleaned_imported_dataset[, numeric_columns], median),
  sd = sapply(cleaned_imported_dataset[, numeric_columns], sd),
  min = sapply(cleaned_imported_dataset[, numeric_columns], min),
  max = sapply(cleaned_imported_dataset[, numeric_columns], max),
  var= sapply(cleaned_imported_dataset[, numeric_columns], var)
)
print(numeric_stats)
```

Figure 3

# Inferential Analysis

Inferential analysis was conducted to draw conclusions and make predictions about the population based on our sample data. This involved applying various statistical tests to examine relationships and differences within the dataset.

## T-test

A t-test is a statistical hypothesis test used to determine if there is a significant difference between the means of two groups. It is broadly categorized into two types:

1. **One-sample t-test:** Compares the mean of a single sample to a known or hypothesized population mean (mu).
2. **Two-sample t-test (paired or independent):** Compares the means of two different groups.

The t-test evaluates two competing hypotheses:

- **Null Hypothesis (H0):** There is no significant difference between the sample mean and the theoretical mean
  If the p-value > 0.05, reject H0 and accept H1.
- **Alternative Hypothesis (H1):** There is a significant difference between the sample mean and the theoretical mean, or a significant difference between the means of the two groups.

  If the p-value < 0.05, reject H0 and accept H1.

# One-Sample T-test Results

We performed two one-sample t-tests on distinct columns:

1. Time in Hospital (time_in_hospital): This column represents the number of days a patient was admitted to the hospital. We tested against a theoretical mean (mu) of 4, given that the actual mean for this column was 4.4. The resulting p-value was 2.2e-16 (or 0.00000000000000022). Since this p-value is significantly less than the conventional

   significance level (e.g., 0.05), we reject the null hypothesis (H0) and accept the alternative hypothesis (H1). This indicates a statistically significant difference between the observed mean time in hospital and the theoretical mean of 4.

```r
#t-test, for the following columns time_in_hospital, number of lab procedures, number of medications.
one_sample_ttest_time_in_hospital<-t.test(cleaned_imported_dataset$time_in_hospital, mu = 4)
print(one_sample_ttest_time_in_hospital)
#H0=mean=mu
#H1=mean!=mu
#According to the t-test,the p-value is 2.2e-16 or (0.00000000000000022) so we reject the H0 and accept the H1
```

Figure 4

2. Number of Medications (num_medication): This column indicates the quantity of medications prescribed to a patient. We tested against a theoretical mean (mu) of 2, as the actual mean for this column was 1.6. The calculated p-value was 2.2e-16 (or

   0.00000000000000022). Similar to the previous test, this extremely low p-value leads us to reject H0 and accept H1, confirming a statistically significant difference between the observed mean number of medications and the theoretical mean of 2

```
one_sample_ttest_num_medications<-t.test(cleaned_imported_dataset$num_medications, mu = 2)
print(one_sample_ttest_num_medications)
#H0=mean=mu
#H1=mean!=mu
#According to the t-test,the p-value is 2.2e-16 or (0.00000000000000022) so we reject the H0 and accept the H1
```

**Figure 5**

# Two-Sample T-test Results

A two-sample t-test was conducted to compare the average number of inpatient (num_inpatient) and outpatient (number_outpatient) visits. This test aimed to

determine if the observed difference in the average number of visits between these two groups is

statistically significant. The p-value obtained from this test was 2.2e-16 (or 0.00000000000000022). Given this highly significant p-value, we reject the null hypothesis (H0) and accept the alternative hypothesis (H1), concluding that there is a statistically significant

difference in the mean number of inpatient and outpatient visits.

```
#two sample t-test for the following columns number_outpatient,number_inpatient
two_sample_paired_ttest<-t.test(cleaned_imported_dataset$number_inpatient,
       cleaned_imported_dataset$number_outpatient,
       paired = TRUE)
print(two_sample_paired_ttest)
#H0=mean=mean
#H1=mean!=mean
# According to the t-test, the p-value is2.2e-16 (0.00000000000000022) so we reject the H0 and accept the H1
```

**Figure 6**

# ANOVA

Analysis of Variance (ANOVA) is a statistical method employed to compare the means of two or more groups, determining if there are statistically significant differences among them. In this study, ANOVA tests were applied to investigate the relationship between patient race (as factor) and key medical indicators such as the number of medications prescribed and the time spent in the hospital.

The decision-making process for the ANOVA test is based on the p-value, with the following hypotheses:

- Null Hypothesis (H0): There is no significant difference between the means of the groups.
- Alternative Hypothesis (H1): There is at least one group whose mean is significantly different from the others.

# Anova test results:

Medication Number ANOVA Test: The ANOVA test for the number of medications yielded a p-value of < 2e-16. This extremely low p-value (significantly less than 0.05) indicates that the observed differences in the number of medications prescribed across different racial groups are highly statistically significant. Therefore, we reject the null hypothesis (H0) and accept the alternative hypothesis (H1), concluding that there is at least one racial group with a significantly different mean number of medications.

Time in Hospital ANOVA Test: the ANOVA test was conducted to compare the mean time spent in the hospital across different racial groups. The results of this test showed that the p-value is 1e-13 which is significantly below the standard significance level of 0.05. Consequently, we reject H0 and accept H1, indicating that there is at least one racial group with a significant difference in the mean time spent in the hospital.

```r
# Load the necessary libraries
#Alone dataset includes two columns race,medication numbers and time in hospital
race_anova_test <- data.frame(
  race = factor((cleaned_imported_dataset$race)),
  medication_numbers = cleaned_imported_dataset$num_medications,
  time_in_hospital = cleaned_imported_dataset$time_in_hospital
)
race_anova_test%>%
  mutate(race = factor(race)) %>%
  group_by(race) %>%
  sample_n(size = 10) %>%
  summarise(
    medication_mean = mean(medication_numbers),
    time_in_hospital = mean(time_in_hospital)
)
levels(race_anova_test$race)

race_medication_anova <- aov(medication_numbers ~ race, race_anova_test)
summary(race_medication_anova)# p-value < 0.05 there's at least one group different than others

race_time_in_hospital_anova <- aov(time_in_hospital ~ race, race_anova_test)
summary(race_time_in_hospital_anova)# p-value < 0.05 there's at least one group different than others
```

Figure 7

# Visualization for ANOVA Analysis

To visually represent the distribution of racial groups and their medication numbers, bar charts were utilized. These visualizations provide a clear overview of the count for each racial group within the dataset. For instance, approximately 77.7% of the records pertain to patients of Caucasian race, while African Americans account for about 19.5% of the data. Further analysis revealed that around 75% of Caucasian patients had consumed 20 or more medications. It was also observed that Caucasian patients exhibited a higher number of outliers or anomalies in medication consumption patterns.

# Now lets see the count of each group using this bar chart:

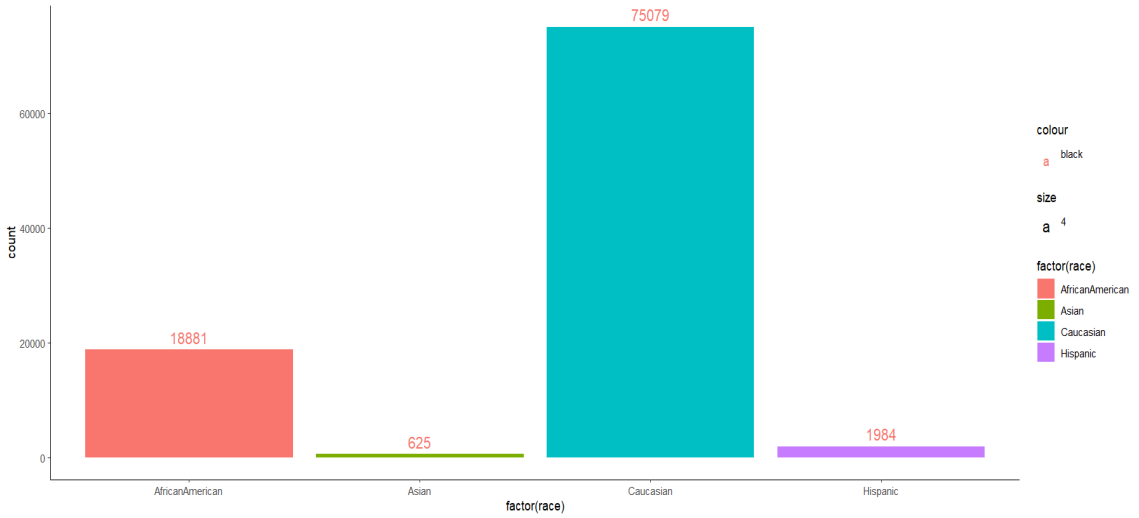


Figure 8

And here are the Box plots were employed to visualize for each racial group offering insights into variability, median time, and the presence of outliers here is the box plots for the number of medications and time in hospital

# Number of medications:

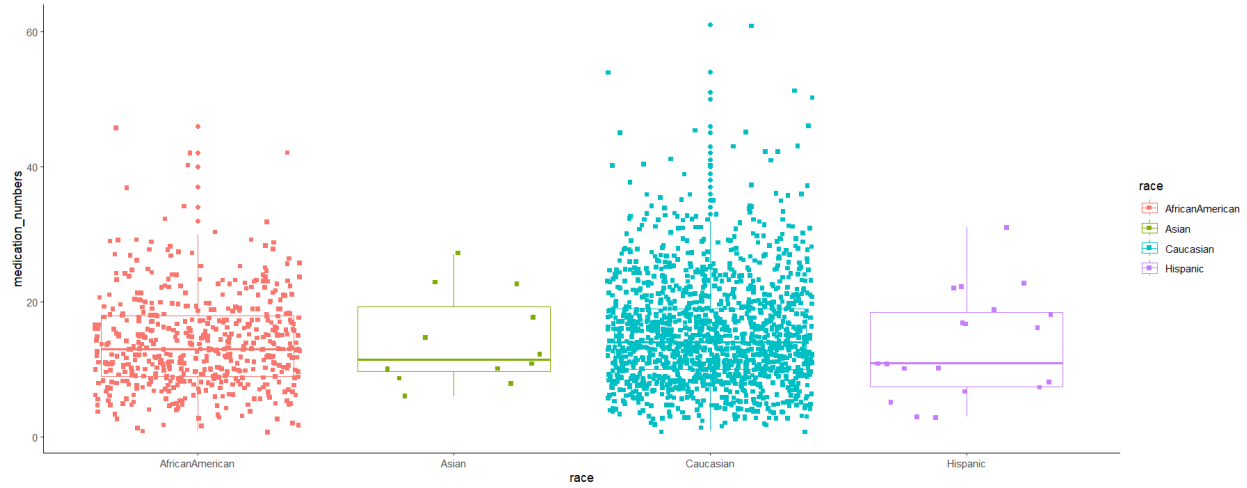


Figure 9

# Time in hospital:

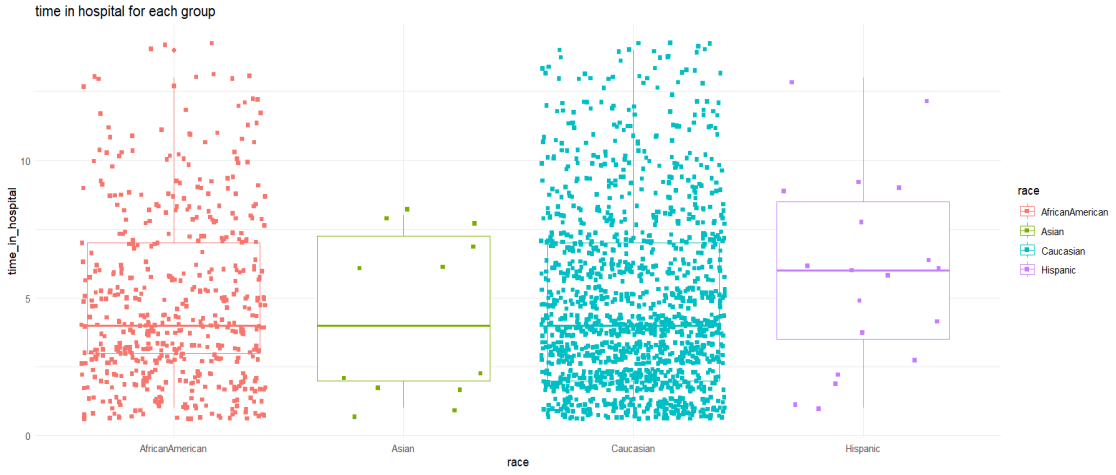


Figure 10

# The following table summarizes our findings from the box plots:

Table 3

| Race | Variability | Median Time (Days) | Outliers | Additional Observations |
|---|---|---|---|---|
| African American | High variability | 4 | Many outliers over 7 days | 50% of the data intensity is centered at 4 days and above |
| Asian | Symmetric distribution | 4 | Few outliers around 6 days | Equal data above and below the median |
| Caucasian | Most extensive dataset | 4 | Highest number over 6 days | 25% of data in the lower half of the box plot |
| Hispanic | Low Variability | 6 | Few outliers over 7 days | 50% of the data have spent 6 days and above |

```
ggplot(cleaned_imported_dataset,aes(factor(race),fill = factor(race)))+
  geom_bar()+
  theme_classic()+
  geom_text(stat = 'count',aes(label = ..count..,color = 'black',size = 4),vjust = -0.5)

# visualizing number of medications for each group
head_race_anova = head(race_anova_test,2000)
ggplot(head_race_anova, aes(x=race, y=medication_numbers,color=race))+
  geom_boxplot()+
  geom_jitter(shape=15)+
  theme_minimal()


#visualizing time in hospital for each group
ggplot(head_race_anova, aes(x = race, y = time_in_hospital, color = race)) +
  geom_boxplot()+
  geom_jitter(shape = 15,alpha = 1) +
  ggtitle( label: "time in hospital for each group") +
   theme_minimal()
```

Figure 11

# Chi-Square Test

The Chi-Square test is a non-parametric statistical test used to determine if there is a significant association between two categorical variables. We performed two Chi-Square tests to explore relationships within our dataset.

The hypotheses for the Chi-Square test are:

- **Null Hypothesis (H0):** There is no relationship or association between the two categorical groups.
- **Alternative Hypothesis (H1):** There is a significant relationship or association between the two categorical groups.

**Chi-Square Test Results**

1. **Race and Gender:** We examined the relationship between patient raceand gender. A contingency table was generated for these two variables to facilitate the analysis. The Chi-Square test yielded a p-value of 2.2e-16 (or 0.00000000000000022). This extremely low p-value leads us to reject the null hypothesis (H0) and accept the alternative hypothesis (H1), indicating a statistically significant relationship between race and gender in the dataset.

```
table_race_gender<-table(cleaned_imported_dataset$race, cleaned_imported_dataset$gender)
table_race_gender
chi_square_test_for_race_gender<-chisq.test(table_race_gender)
print(chi_square_test_for_race_gender)
#H0=there is no relationship between the two variables
#H1=there is a relationship between the two variables
#According to the chi-square test, the p-value is 0.0000000000000002 (2.2e-16) so we reject the H0 and accept the H1
#Because there is a relationship between the two variables
```

Figure 12

2. **Age and Insulin:** We also investigated the association between patient age and insulin usage. Similar to the previous test, a contingency table was created for these variables. The Chi-Square test resulted in a p-value of 2.2e-16 (or 0.00000000000000022). This highly

   significant p-value compels us to reject H0 and accept H1, confirming a statistically significant relationship between age and insulin usage.

```
table_age_insulin<-table(cleaned_imported_dataset$age, cleaned_imported_dataset$insulin)
table_age_insulin
chi_square_test_for_age_insulin<-chisq.test(table_age_insulin)
print(chi_square_test_for_age_insulin)
#H0=there is no relationship between the two variables
#H1=there is a relationship between the two variables
#According to the chi-square test, the p-value is 0.0000000000000002 (2.2e-16) so we reject the H0 and accept the H1
#Because there is a relationship between the two variables
```

Figure 13

# Correlation Test

A correlation test is used to quantify the strength and direction of a linear relationship between two integer variables. This analysis aimed to answer a specific question: as the number of diagnoses increases, does the number of medications prescribed also tend to increase? To address this, a correlation test was performed between num_medications and num_diagnoses.

## Data Preparation

Before conducting the correlation test, the column names were adjusted for readability to (no_medications  and no_diagnoses) It was also verified that both columns were of

integer type, a prerequisite for the correlation test. The dataset used for this specific test comprised 96,569 observations.

## Types of correlation:

- Pearson Correlation: This is the default method, measuring the linear relationship between two variables.
- Spearman's Rank Correlation: A non-parametric measure of the monotonic relationship between two variables.

## Understanding the test results:

- **Strength**: A numerical value indicating the degree to which the data correlates. Values close to 1 or -1 signify a strong relationship, while values near 0 indicate a weak relationship.
- **Direction**: This indicates how the variables change in relation to each other. A positive correlation means that as one variable increases, the other tends to increase. A negative correlation implies that as one variable increases, the other tends to decrease

## Now for the results:

The correlation test between the number of medications and the number of diagnoses revealed a statistically significant positive correlation. The correlation coefficient was approximately 0.24, indicating a weak to moderate positive relationship. This suggests that while there is a tendency for the number of medications to increase with the number of diagnoses, the strength of this relationship is not exceptionally strong.

```r
# correlation Test is a statistical test applied only on numerical columns to see weather the columns related to each other or not
# first we will ask our-self,As the number of diagnoses increases, does the number of medications prescribed also tend to increase?
# lets see
medications_diagnoses_test <- data.frame(
  no_medication <- cleaned_imported_dataset$num_medications,
  no_diagnoses <- cleaned_imported_dataset$number_diagnoses
)
# check the type of columns must be integer to proceed the correlation test
str(medications_diagnoses_test)

# changing column names into smaller form
names(medications_diagnoses_test) = c('no_medication','no_diagonses')

correlation_test = cor.test(medications_diagnoses_test$no_medication,medications_diagnoses_test$no_diagonses,method = 'pearson')
print(correlation_test)
# the correlation test is 0.2408395 which indicates that is appeared a weak positive linearNBSPcorrelation
```

Figure 14

## Visualizing Correlation:

1. **Scatter plot:** A scatter plot was used to illustrate the relationship between the two variables. To enhance readability and manage data intensity, only the first 2,000 rows of the dataset were used for this visualization. The scatter plot visually confirmed a weak positive correlation

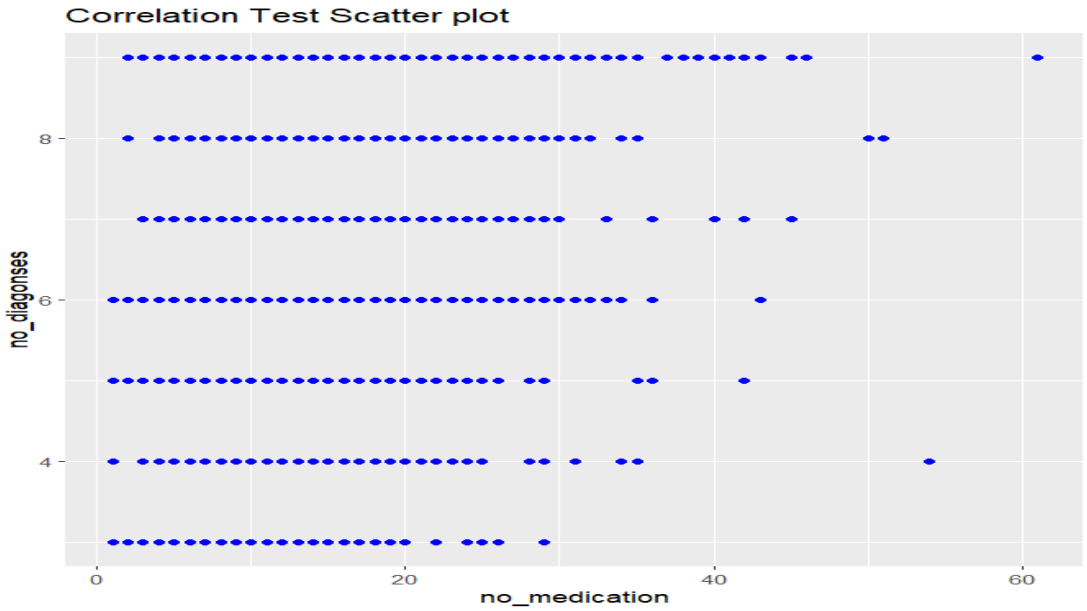


Figure 15

**Heat map**: A heat map was also generated to provide another visual representation of how the data is related

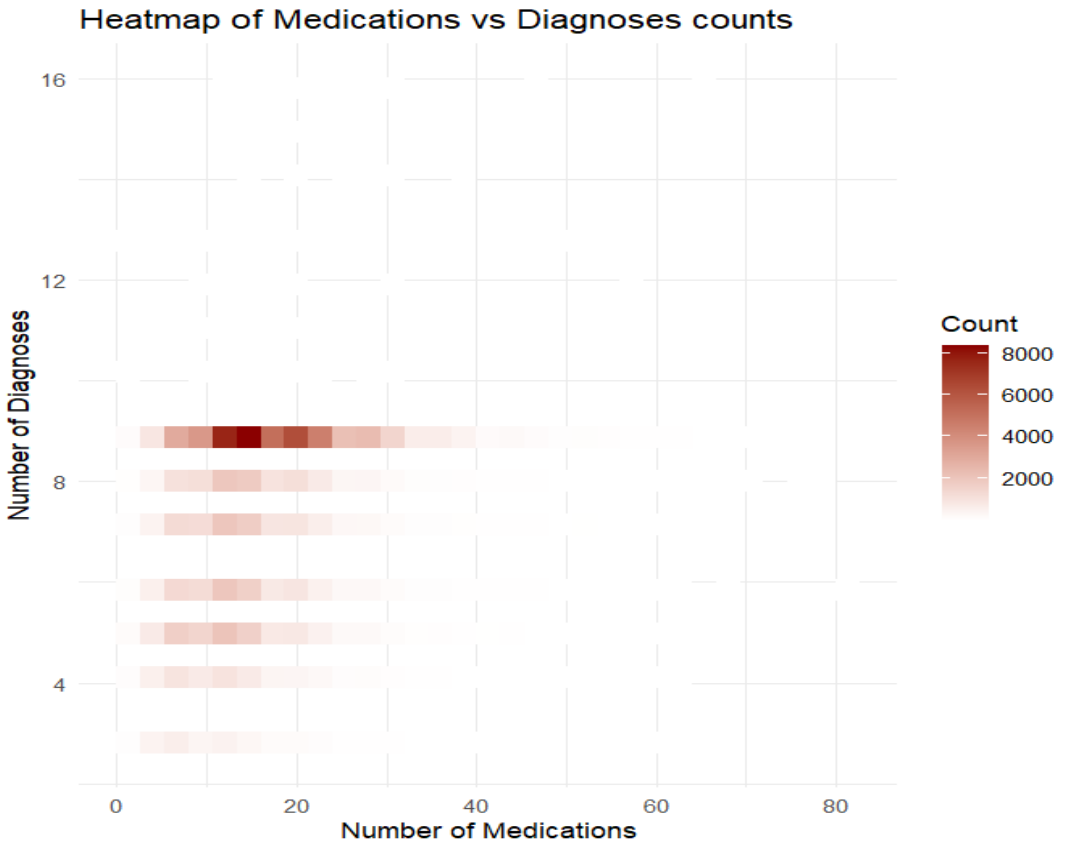


Figure 16

```
theme_minimal()
# Visualizing the correlation between number of medications and number of diagnoses using a scatter plot
ggplot(head(medications_diagnoses_test,2000),aes(x = no_medication,y = no_diagonses))+
  geom_point(color = 'blue')+
  ggtitle('Correlation Test Scatter plot')
  theme_classic()+
 theme_minimal()
# Visualizing the correlation between number of medications and number of diagnoses using a heatmap
ggplot(medications_diagnoses_test, aes(x = no_medication, y = no_diagnoses)) +
  geom_bin2d(bins = 30) +  # Adjust bins for granularity
  scale_fill_gradient(low = "white", high = "darkred") +
  labs(title = "Heatmap of Medications vs Diagnoses counts",
       x = "Number of Medications",
       y = "Number of Diagnoses",
       fill = "Count") +
       theme_minimal()
```

Figure 17

And we have here a histogram that represent time in hospital per gender :

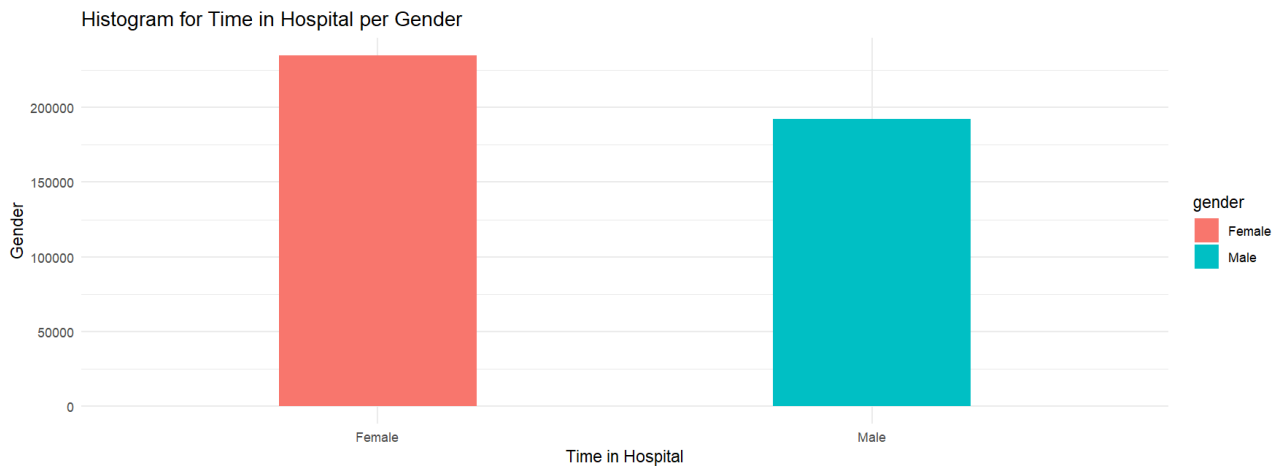


Figure 18

```
ggplot(cleaned_imported_dataset, aes(x = gender,y = time_in_hospital,fill = gender))+
  geom_bar(stat = 'identity',width = 0.4)+
  labs(
    title = "Histogram for Time in Hospital per Gender",
    x = "Time in Hospital",
    y = "Gender"
  )+
  theme_minimal()
```

Figure 19

# Conclusion

This study provided a comprehensive analysis of a large-scale clinical dataset focused

on diabetic patients over a ten-year period. Through rigorous data cleaning, descriptive statistics, and a range of inferential analyses—including t-tests, ANOVA,

Chi-square tests, and correlation analysis, we uncovered critical insights into patterns of hospital care, medication use, and demographic disparities in diabetes management.

Key findings revealed statistically significant differences in hospital stay durations and medication counts across racial groups, as well as meaningful associations between variables such as race and gender, and age and insulin usage. Additionally, a weak to moderate positive correlation was identified between the number of diagnoses and the number of medications prescribed, indicating some degree of complexity in the treatment of patients with multiple conditions.

These results underscore the importance of standardized, data-driven approaches to managing diabetes in hospital settings. The observed disparities and trends should inform hospital policies and targeted interventions aimed at improving treatment consistency, reducing preventable complications, and ultimately enhancing patient outcomes. Future work could expand on this foundation by integrating additional clinical data, exploring causal relationships, and employing predictive modeling to support proactive healthcare decision-making.

This study provided a comprehensive analysis of a large-scale clinical dataset focused on diabetic patients over a ten-year period. Through rigorous data cleaning, descriptive statistics, and a range of inferential analyses—including t-tests, ANOVA, Chi-square tests, and correlation analysis, we uncovered critical insights into patterns of hospital care, medication use, and demographic disparities in diabetes management.

Key findings revealed statistically significant differences in hospital stay durations and medication counts across racial groups, as well as meaningful associations between variables such as race and gender, and age and insulin usage. Additionally, a weak to moderate positive correlation was identified between the number of diagnoses and the number of medications prescribed, indicating some degree of complexity in the treatment of patients with multiple conditions.

These results underscore the importance of standardized, data-driven approaches to managing diabetes in hospital settings. The observed disparities and trends should inform hospital policies and targeted interventions aimed at improving treatment consistency, reducing preventable complications, and ultimately enhancing patient outcomes. Future work could expand on this foundation by integrating additional clinical data, exploring causal relationships, and employing predictive modeling to support proactive healthcare decision-making.