DEPAULUNIVERSITY

GLUCOSE GUIDES

Nikhitha Poladi 2119148

TABLE OF CONTENTS:

1. Introduction:

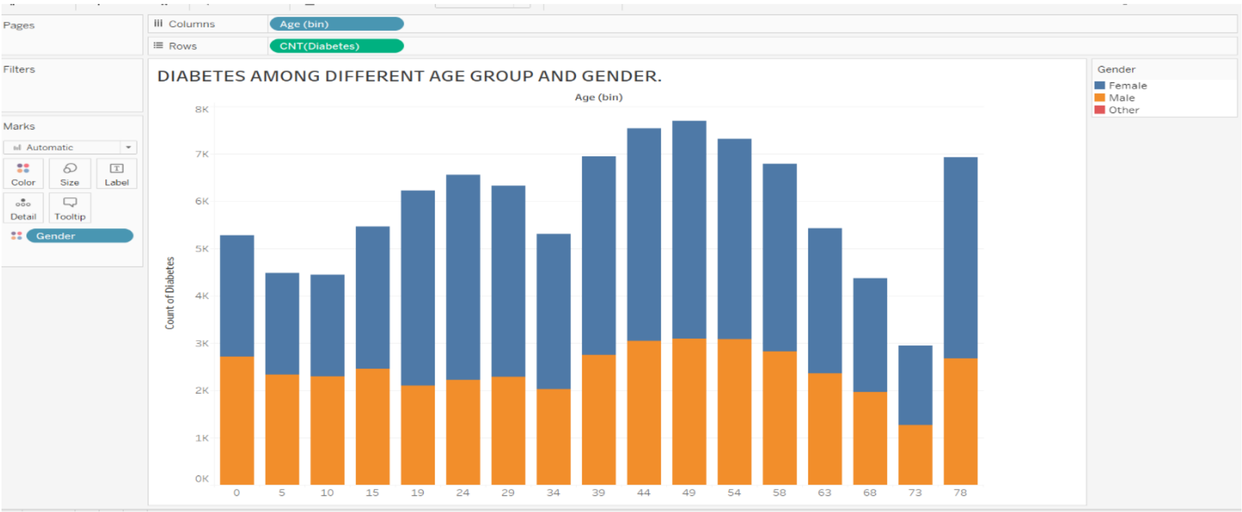
Diabetes has emerged as a global epidemic, afflicting over 422 million people worldwide. Understanding the interplay between demographic and health factors can enhance diabetes prediction and prevention efforts. The dataset contains variables like patient gender, age, blood pressure status, smoking history, body mass index (BMI), HbA1c, fasting blood glucose, and diabetes diagnosis. Understanding its predictors through data visualization offers us deeper insights into managing and preventing this condition. Our objective today is to delve into various visualizations that highlight key factors predicting diabetes.

Tools Used: R and Tableau
Layoffs Dataset:
Data Source: Kaggle
Link:https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset
no. of rows: 100000
Variables: (11 variables)

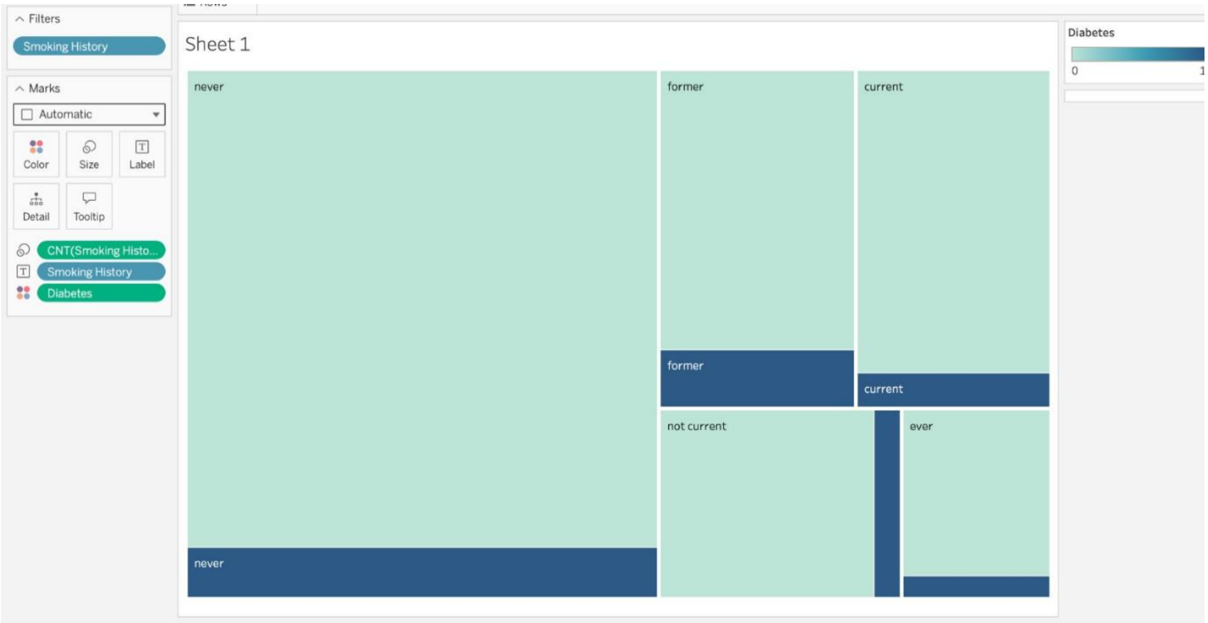
1. **Gender:** Indicates the gender of the individuals in the dataset.
2. **Age:** Represents the age of participants
3. **Hypertension**: Indicates whether the person has the hypertension (high blood pressure)
4. **Heart Disease**: Shows if the individual has the history of heart disease.
5. **Smoking History**: Details regarding the individuals smoking behavior, whether current, former, or never
6. **BMI (Body Mass Index)**: A measure of body fat based on height and weight.
7. **HbA1c level**: glycated hemoglobin, refers to the average blood sugar levels.
8. **Blood Glucose Level**: The concentration of glucose in the blood at the time of measurement.
9. **Diabetes**: The proportion of diabetes, measured with the scale of 0 and 1.
10. **Age Group:** categorizes individuals into different age brackets for analysis.
11. **Hypertension diabetes:** This provides comprehensive insights into individuals' health, focusing on hypertension and diabetes status.

2. Exploratory Analysis:

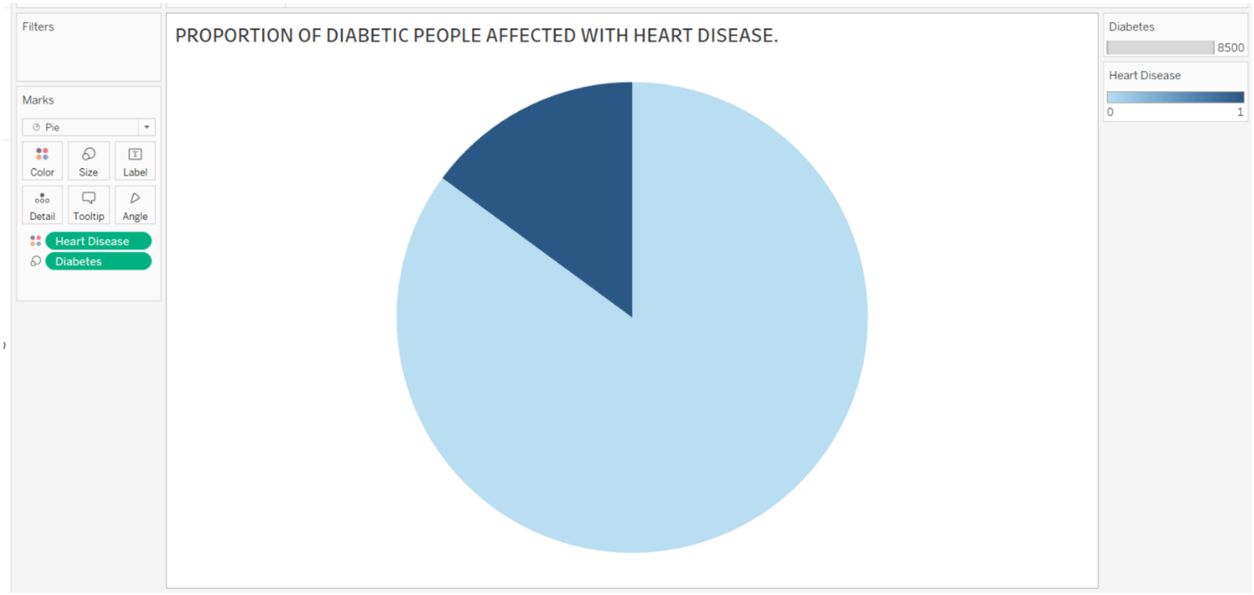
We analyzed the data set and removed certain null values wherever necessary. We plotted several independent variables with the dependent variable Diabetes and observed certain patterns that helped us to create the following visualizations.



We've employed a stacked bar chart as a visual aid to analyze the prevalence of diabetes across different age groups and genders. Our observation indicates that individuals in the age range of 49 to 53 are the most significantly affected, with a higher incidence among females compared to males.



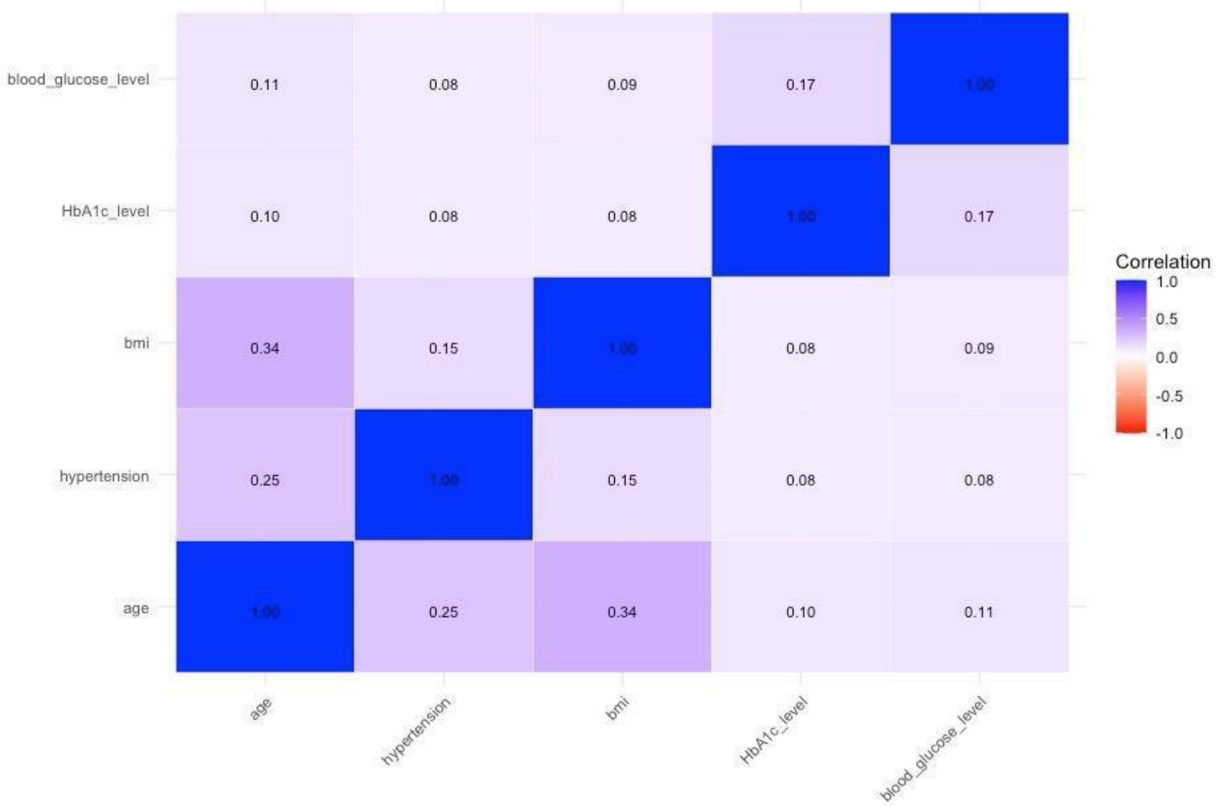
In our analysis, we harnessed a tree map to categorize individuals with diabetes into two groups: smokers and non-smokers. Our findings underscore that smoking is a relatively minor contributing factor in the context of diabetes risk.



PROPORTION OF DIABETIC PEOPLE AFFECTED WITH HEART DISEASE.

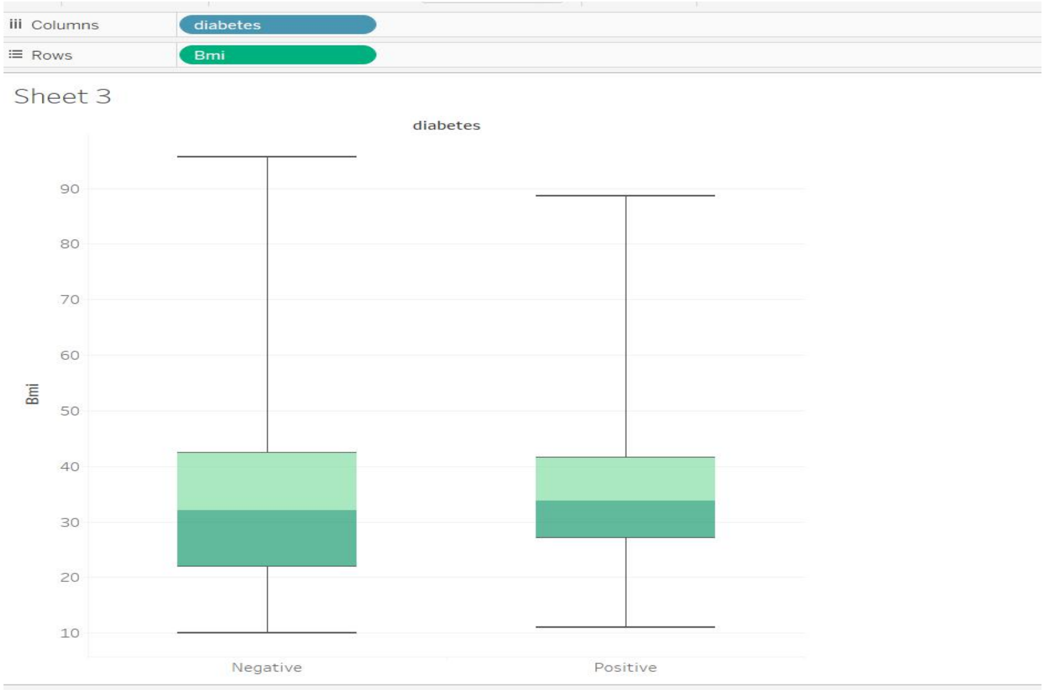
The pie chart presented here illustrates the proportion of individuals with diabetes who are also affected by heart disease. The data indicates that the occurrence of heart disease among those with diabetes is relatively low, suggesting a lower risk associated with this comorbidity.

Correlation Heatmap of Numerical variables:



Our plan is to enhance the correlation heatmap, which displays relationships between key numerical variables like age, BMI, HbA1c level, and blood glucose level. This heatmap will help identify which variables are most strongly associated with each other.

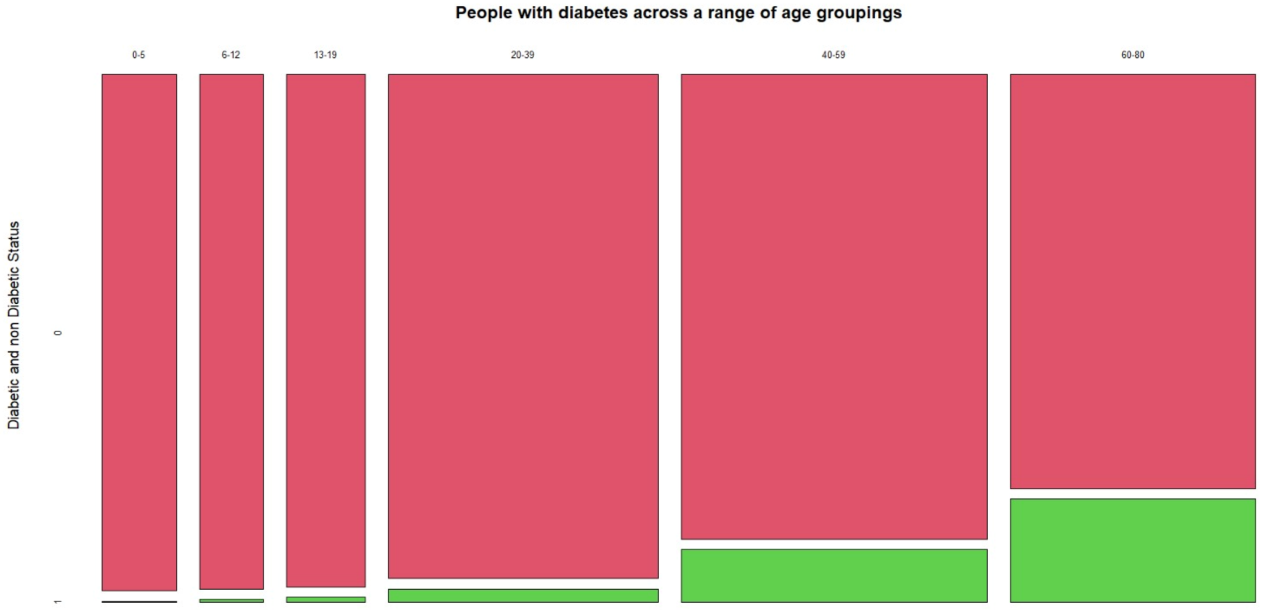
Boxplots for diabetes across BMI (body mass index) .



Developing a box plot to compare the effect of BMI among diabetic and non-diabetic people.the boxplot shows bmi does not have much effect in the occurrence of diabetes.

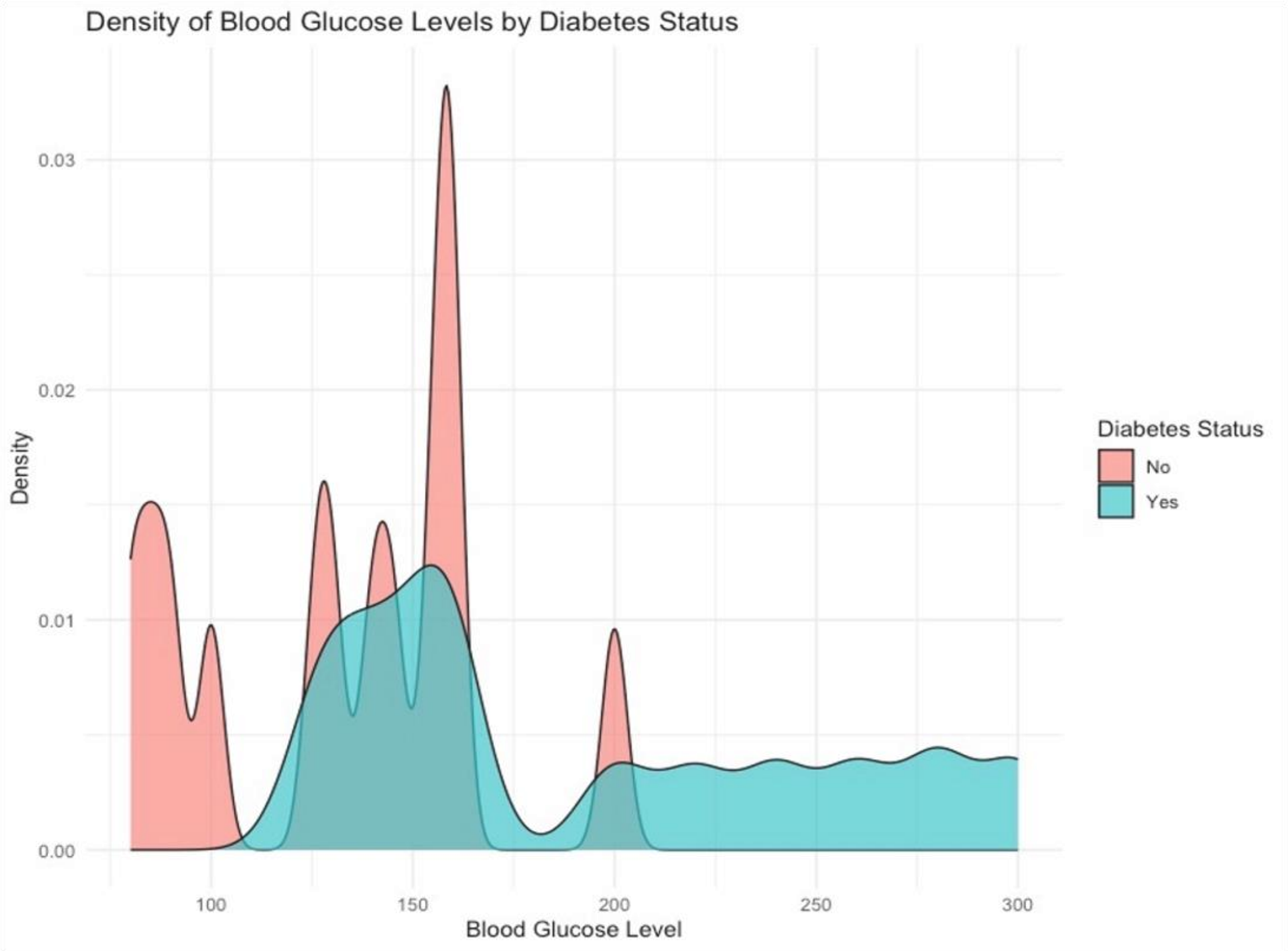
3. Visualizations

MOSAIC PLOT:



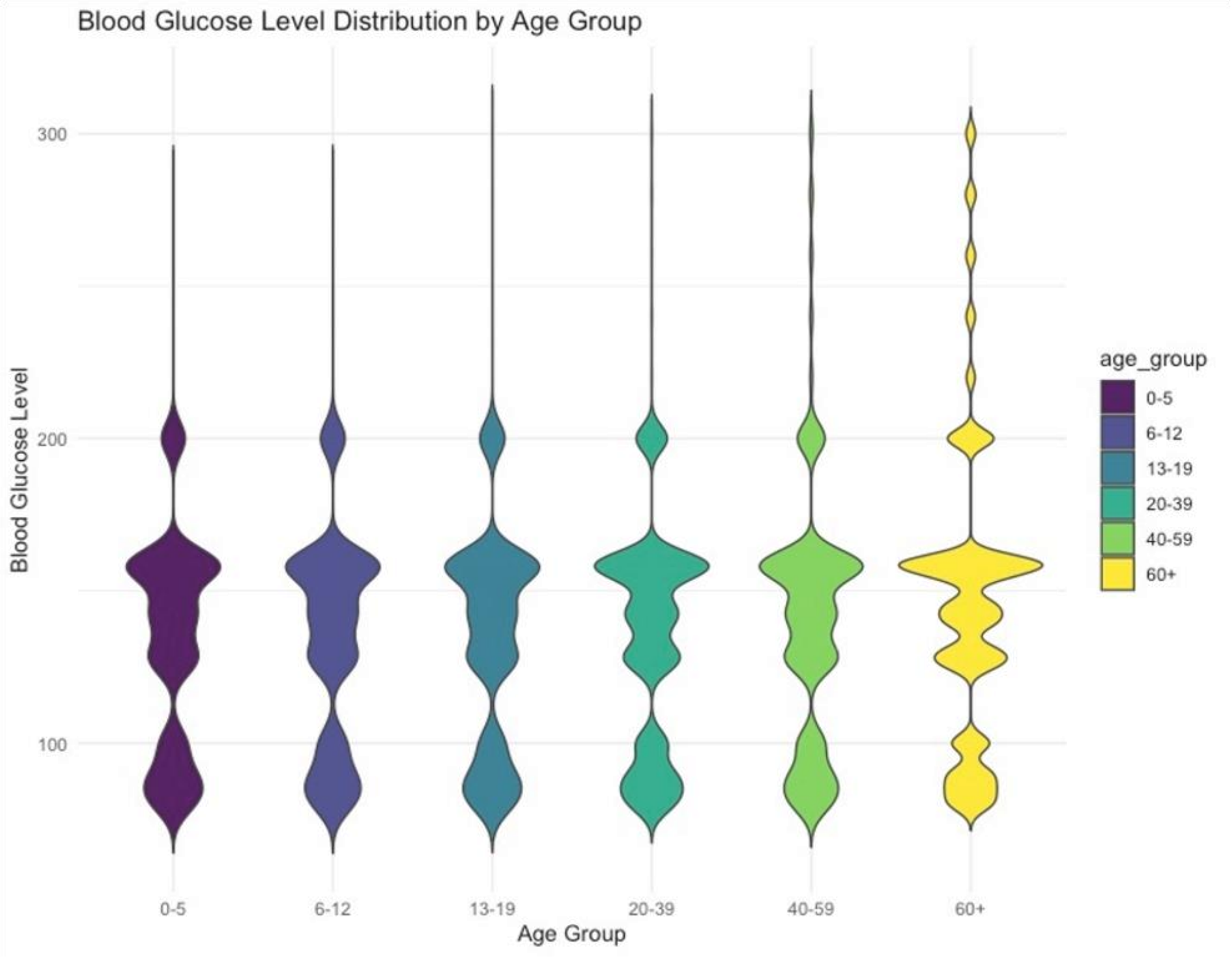
The stacked bar graph compares the proportions of diabetic and non-diabetic people across five age groups: 0-5, 12-19, 20-39, 40-59, and 60-80. The bars are color-coded, with red likely representing diabetic individuals and green for non-diabetic individuals. The tallest bars appear in the 40-59 and 60-80 age groups, indicating a higher prevalence of diabetes in these age ranges. The bars for the younger age groups, 0-5 and 12-19, are significantly shorter, suggesting a lower occurrence of diabetes in those age brackets. The data is presented in a clear, visual format, allowing for quick comparison between the different age groups.

DENSITY PLOT:



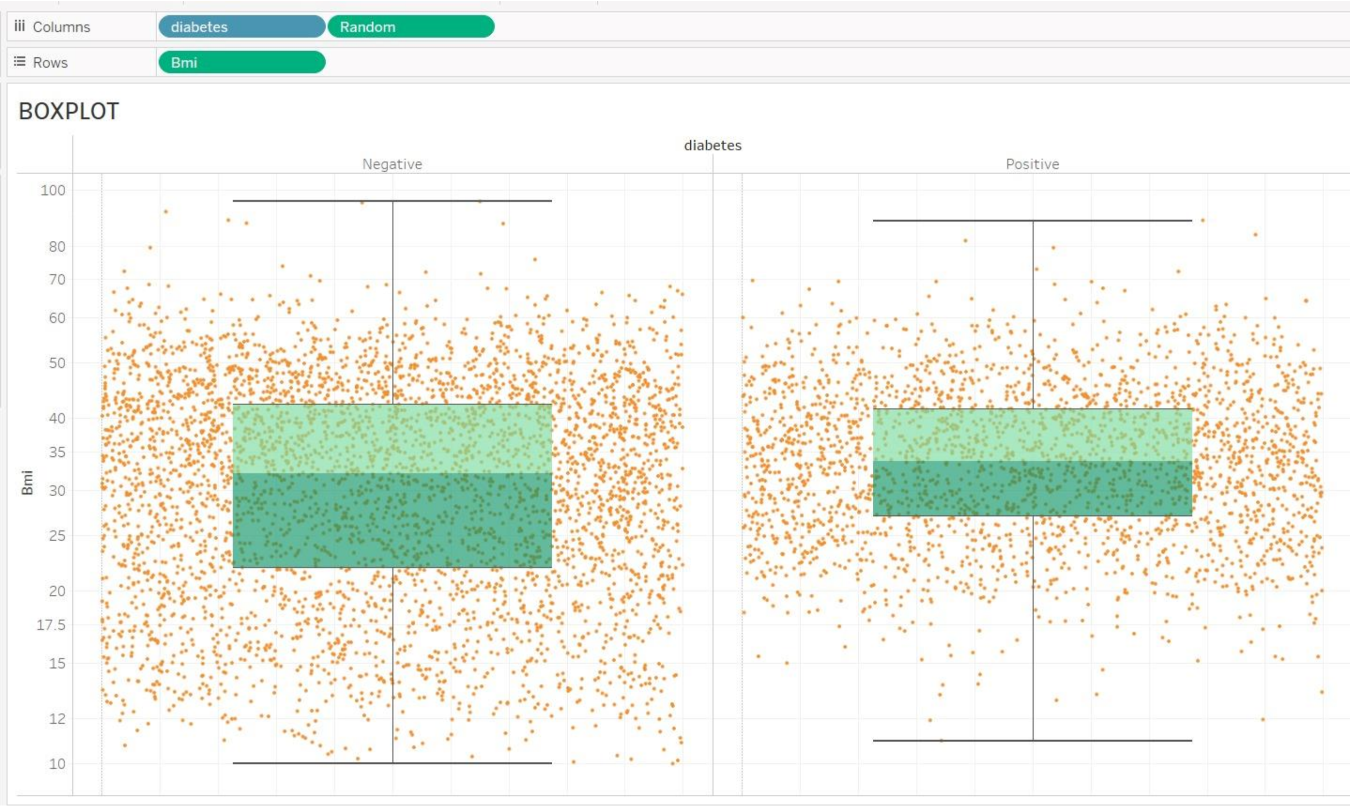
This density plot shows the distribution of blood glucose levels differentiated by diabetes status. The x-axis measures blood glucose levels, while the y-axis represents the density of observations at each level. Two colors indicate the status: pink for individuals without diabetes ("No") and teal for those with diabetes ("Yes"). The plot reveals multiple peaks for both groups, suggesting common ranges where blood glucose levels are concentrated within the population. The teal distribution appears broader, which may indicate a wider variation in blood glucose levels among individuals with diabetes compared to those without.

VIOLIN PLOT:



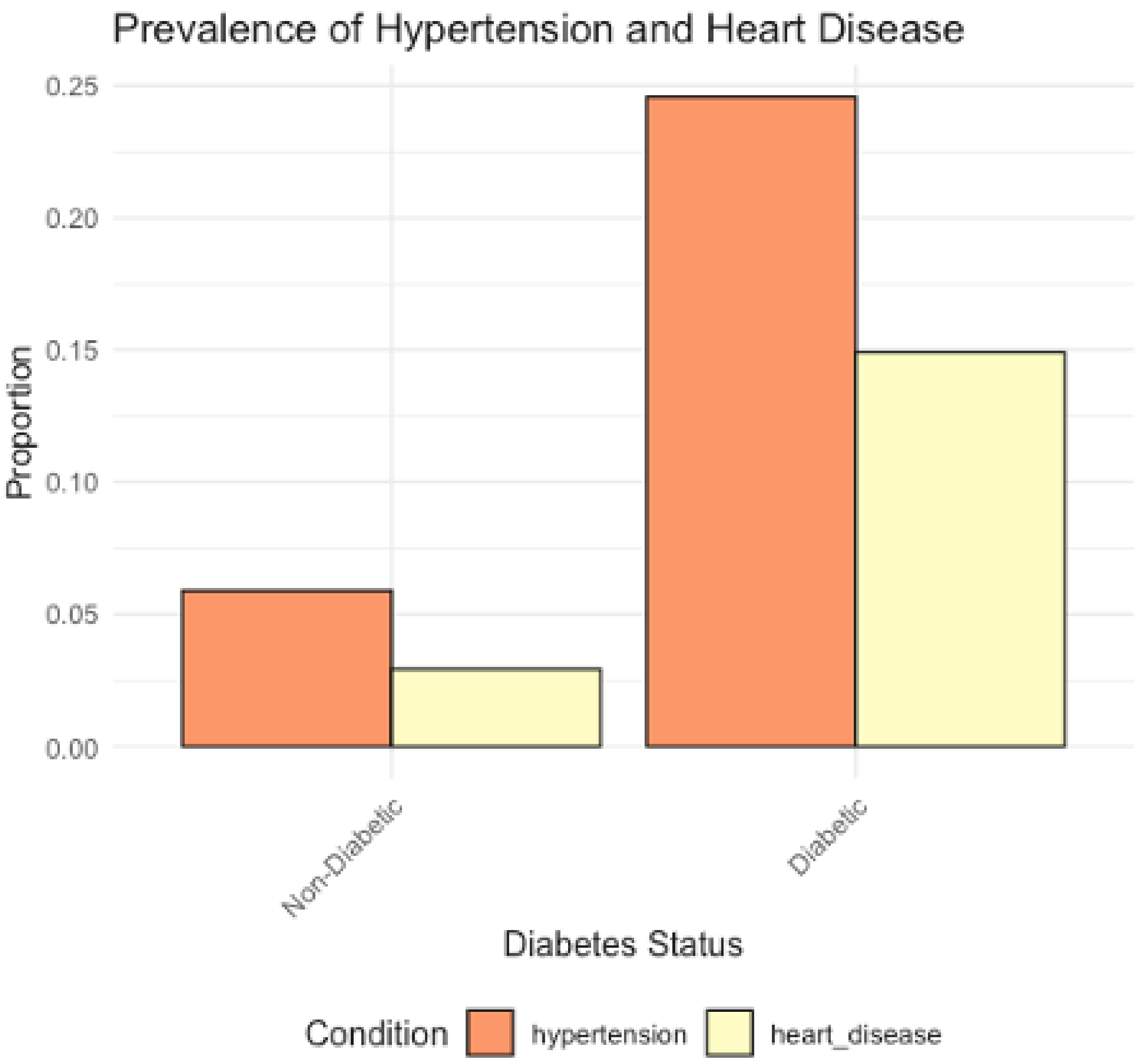
This violin plot shows the distribution of blood glucose levels across different age groups. Each 'violin' represents an age group, with individual distributions color-coded for clarity: purple for 0-5 years, dark blue for 6-12 years, light blue for 13-19 years, teal for 20-39 years, green for 40-59 years, and yellow for 60+ years. The width of each violin indicates the density of data points at different blood glucose levels, while the height reflects the range of values. The plot allows for a visual comparison of the distributions, highlighting differences and similarities in blood glucose levels among the age groups.

BOXPLOT:



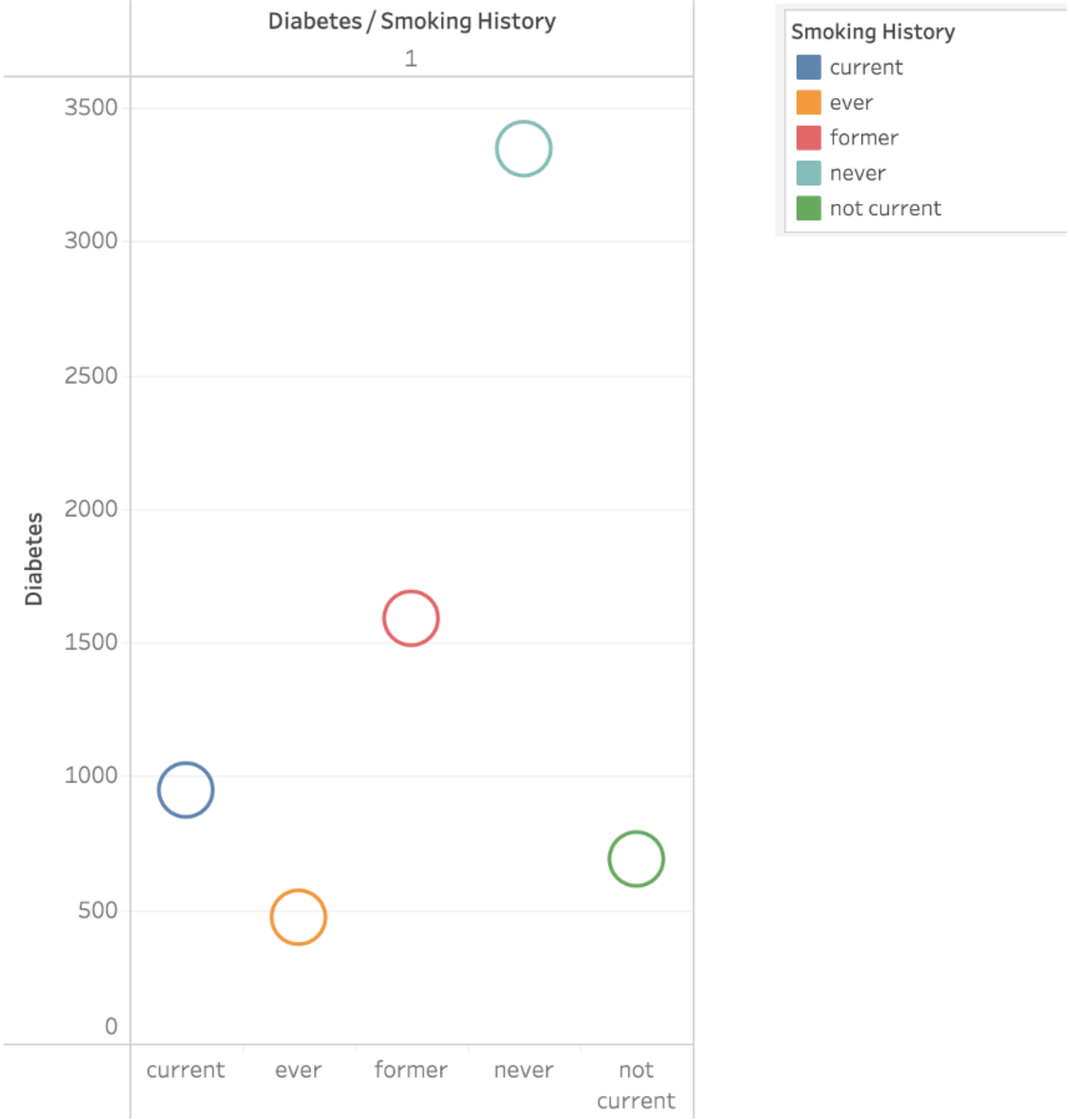
A boxplot is used to see the distribution of diabetes over different values of BMI(body mass index).I applied jiitering over the draft graph to gain further insights.It is clearly demonstrating that there is no influence of BMI on the status of diabetes.. On the left, the 'Negative' box plot shows the range and distribution for individuals without diabetes, while the 'Positive' on the right represents those with diabetes. Each box plot indicates the median, quartiles, and potential outliers of the dataset. The scatter plot overlay presents individual data points, suggesting not a greater difference in the spread and median levels among both levels.

BAR CHART:



This bar chart compares the proportion of individuals with hypertension and heart disease among Non-Diabetic and Diabetic groups. Orange bars represent hypertension, while yellow bars indicate heart disease. The chart shows a higher prevalence of both conditions in the Diabetic group, with a notably higher proportion of hypertension compared to the heart disease group. This visualization clearly shows the correlation between diabetes and an increased risk of hypertension and heart disease.

SCATTERPLOT:



The scatter plot represents the relationship between smoking history and the prevalence of diabetes. Colored circles correspond to different smoking statuses: blue for current smokers, orange for people who have ever smoked, red for former smokers, green for those who have never smoked, and teal for not current smokers. The position and size of the circles indicate the number of diabetes cases within each category, with the green circle (never smokers) being the largest and highest on the plot, suggesting a higher number of diabetes cases among individuals who have never smoked.

4. Analysis and Discussion:

The visualizations and exploratory analysis provide significant insights into factors predicting diabetes prevalence and risk.

Age is a major predictor of diabetes risk, with a notably higher incidence of diabetes in middle-aged and elderly populations compared to younger age groups. The mosaic plot highlights the proportion of diabetes cases among individuals aged 40-59 and 60-80. This suggests prediabetic screening and prevention efforts could target these high-risk age brackets.

Blood glucose levels differ significantly between diabetic and non-diabetic groups. The density plot illustrates the wider spread and higher median glucose levels among diabetic patients. These findings underscore the importance of regular blood glucose screening for early diabetes detection.

The box plot goes against the commonly believed notion that obesity will certainly lead to diabetes. we can see that not all people who are obese are affected with diabetes.

Comorbidities like hypertension and heart disease occur more frequently in diabetic patients compared to the general population. The bar chart clearly indicates the increased prevalence of both conditions among individuals with diabetes. This highlights the need to monitor diabetic patients for associated health risks.

Smoking has a relatively small correlation to diabetes risk compared to other factors like age and glucose levels. The scatter plot reveals a high prevalence of diabetes even among those who never smoked. While smoking may exacerbate diabetes, it does not appear to be a primary predictor based on this data.

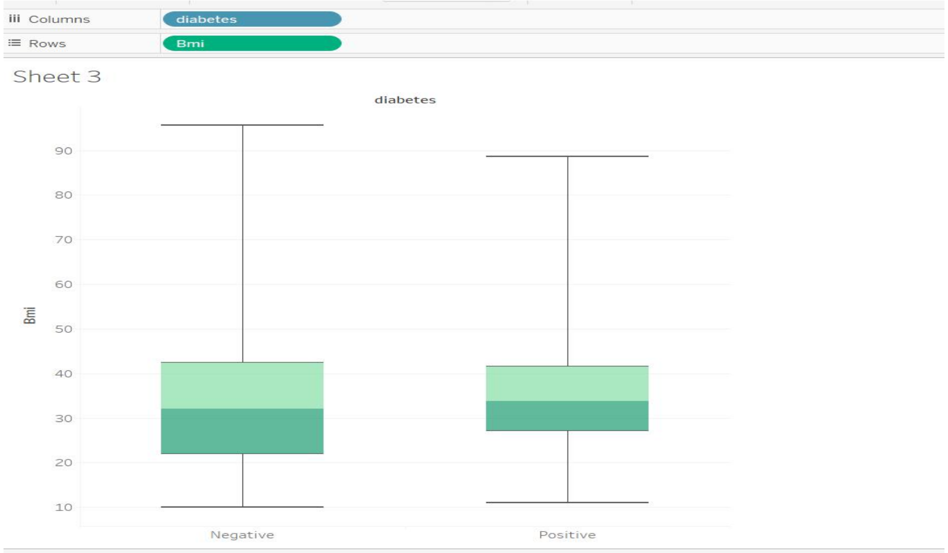
In summary, the integrated analysis strongly indicates age, blood glucose trends, and comorbidities like hypertension have the closest association with diabetes occurrence. A multipronged strategy of focused screening in high-risk demographics, regular blood glucose monitoring, and aggressive management of related health conditions may help curb diabetes prevalence. The visualizations and exploratory analysis offer actionable population health insights from a multifaceted dataset.

5.INDIVIDUAL REPORT:

As a part of the project I was involved in both exploratory and explanatory analysis .In the exploratory phase I used a boxplot to see the relationship between diabetes and BMI(body mass index).

It is equally important to consider factors which are not effecting diabetes too.There is a common misleading notion that people who are obese certainly meet with diabetes.This notion is ruled out by the following analysis using a wide demographic data.

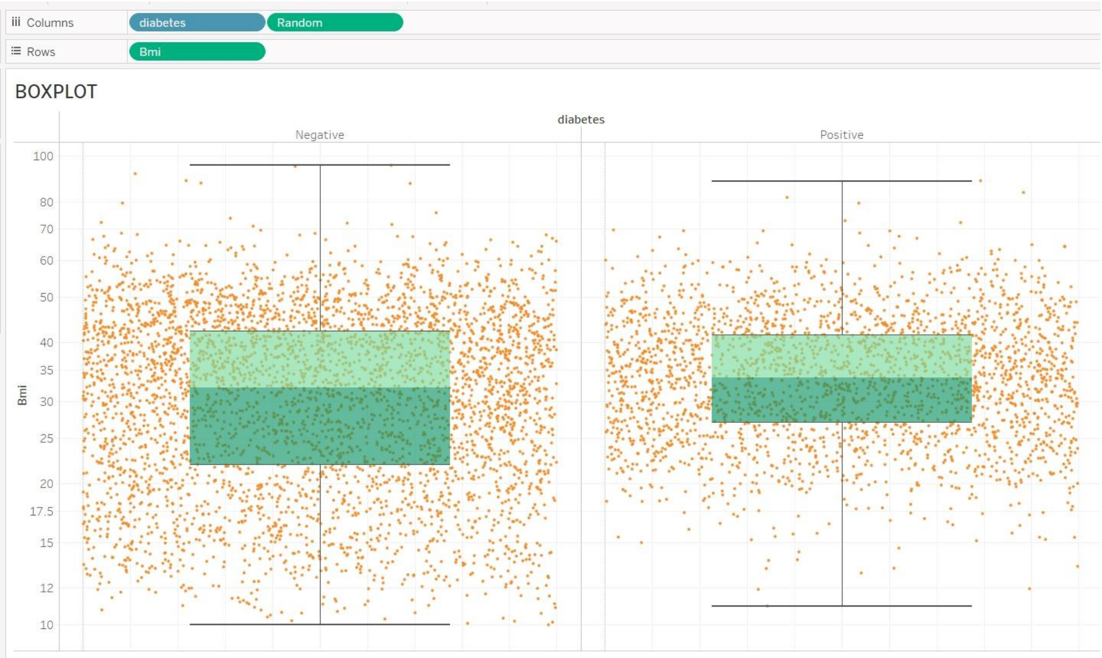
In the explanatory phase, I further added jittering over the boxplot which shows individual data points suggesting not much difference in the spread in both the levels.



Columns: diabetes

Rows: Bmi

Sheet 3

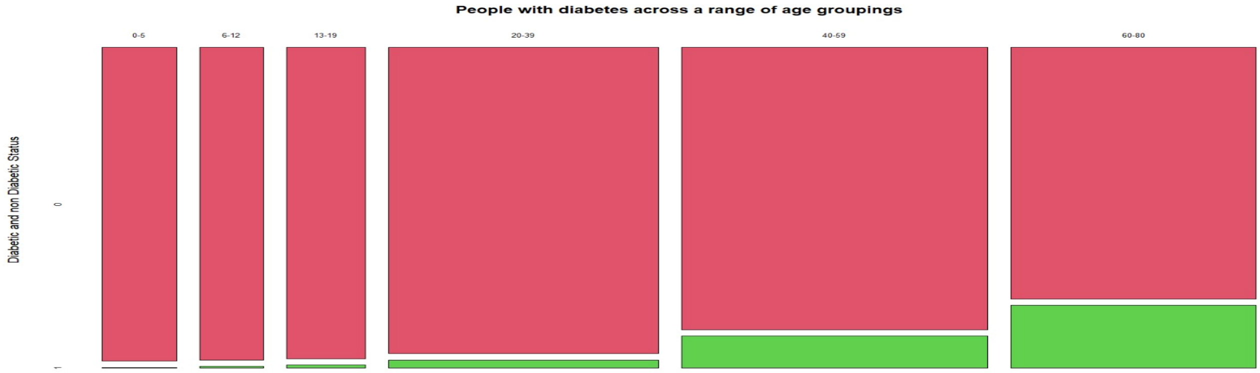


Columns: diabetes, Random

Rows: Bmi

BOXPLOT

Next I did a mosaic plot in my final visualization to show the percentage distribution of diabetes across different life stage age groupings.Age is definitely a influencing factor in the occurrence of diabetes.



Future Analysis: By considering what factors effect and what does'nt in the occurrence of diabetes,we can take a preventive course of action and reduce the overall burden diabetes has on the healthcare system.

Key Takeaways: I was able to learn about the different factors causing diabetes and work on audience-oriented visualizations thanks to this project. As a result of the project, I was also able to establish stronger relationships with my teammates and work together on fresh concepts to produce visually engaging content.

Reflection: I now have a more thorough understanding of the many kinds of visualizations and how well they can communicate ideas through this course. Additionally, the course introduced me to new R libraries and the program Tableau, which were both helpful in creating a variety of representations.

R code for my visualization:

```
Diabetes.d<-diabetes.data.set %>% mutate(agegrp=cut(age,breaks=c(0,6,13,20,40,60,80),right=T,labels=F))

> Diabetes.d$agegrp<-as.factor(Diabetes.d$agegrp)

> levels(Diabetes.d$agegrp)<-c("0-5","6-12","13-19","20-39","40-59","60-80")

> totals=table(Diabetes.d$agegrp,Diabetes.d$diabetes)

mosaicplot(totals,main="Diabetes over different age groups",xlab="Agegroups",

     ylab="Diabetic and non Diabetic People",col=c(2,3))
```