Final Project Step1

Basava, Madhuri

2023-02-12

# Introduction

## Diabetes is one of the leading causes of death worldwide and especially in the USA. Nowadays more people are getting affected by diabetes.

## This project is to analyze different factors affecting diabetes and based on the results let people know how to prevent diabetes by altering the affecting factors.

## I feel that health is more than anything in the world, so this project will be useful for many people.

# Below are some of the research questions that are relevant

## \* How can we reduce diabetes cases in the future?

## \* What are the factors affecting diabetes?

## \* How much Physical activity in a certain period is needed to reduce diabetes cases?

## \* Is smoking, a direct or indirect cause of diabetes?

## \* How much BMI value range should a person have to reduce the possibility of diabetes?

## \* Does High Blood Pressure, a reason for diabetes?

## \* Are Males or Females more prone to diabetes?

## \* What age people are getting affected by diabetes more?

## \* Is a person’s heart attack/stroke has to be more careful?

## \* Will high cholesterol lead to Diabetes?

## \* Will heavy alcohol consumption lead to Diabetes?

## \* Is diabetes dependent on physical, general, or mental health?

# Approach

## Clean the data: Firstly, I will remove the NA values from the dataset.

## Perform some transformations to tidy up the data.

## Then, Analyse the data and visualize it in the form of different graphs and

## charts to figure out

## what are the factors which are affecting diabetes?

## plot the graphs with Diabetes on Y axis and physical activity on X-axis and analyze them.

## plot the graphs with Diabetes on Y axis and smoking on X-axis

## plot the graphs with Diabetes on Y axis and BMI on X-axis

## plot the graphs with Diabetes on Y axis and HighBP on X-axis

## plot the graphs with Diabetes on Y axis and Sex on X-axis

## plot the graphs with Diabetes on Y axis and Age on X-axis

## plot the graphs with Diabetes on Y axis and HeartDiseaseorAttack on X-axis

## plot the graphs with Diabetes on Y axis and HighChol on X-axis

## plot the graphs with Diabetes on Y axis and HvyAlcoholConsump on X-axis

## plot the graphs with Diabetes and physical, general, or mental health

## Finally provide useful analysis for people who can change their

## lifestyle to reduce Diabetes cases.

# How your approach addresses (fully or partially) the problem.

## The analysis gives us the idea of which factors are more likely to cause diabetes and share the results with everyone, so that people will change their lifestyles accordingly to reduce diabetes problems in the future.

# Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)

# The original source where the data was obtained is cited and, if possible, hyperlinked.

# Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).

## 3 data sets chosen for this project are from Kaggle site.

## \* diabetes\_012\_health\_indicators\_BRFSS2015.xlsx

## \* diabetes\_binary\_5050split\_health\_indicators\_BRFSS2015.xlsx

## \* diabetes\_binary\_health\_indicators\_BRFSS2015.xlsx

## The purpose of the data is to analyze the factors/predictors affecting Diabetes. The data was collected from the year 2015. The original data has 22 columns in each data set many thousands of rows/records.

## There were no missing data. I took 50 rows/records from each dataset and combined them into one dataset by binding the rows which now have 150 rows/records.

# Required Packages

## The important packages needed for this project are

## \* realxl – to read the excel data files.

## \* dplyr – to analyze/transform the data using GroupBy, Summarize, Mutate, Filter, Select, and Arrange

## \* tidyr – to tidy data to make the data more consistent

## \* ggplot2 – for visualizing the different factors affecting diabetes.

## \* pheatmap – to draw a heatmap of our correlation table

## \* psych – to derive descriptive statistics for a data set

# Plots and Table Needs

## Below are the Plots and tables used in this project:

## \* histograms

## \* bar graphs

## \* heatmaps

## \* scatterplots

## \* boxplots

# Questions for future steps

## I do not know how to graph using heatmaps to visualize all the predictors for data analysis.

## Diabetes Indicator Project

## Set the working directory to the root of your DSC 520 directory

setwd("C:/MadhuriDocs/MSInDataScience/DSC520RCourse3/Week8/project\_data/Health")  
getwd()

## [1] "C:/MadhuriDocs/MSInDataScience/DSC520RCourse3/Week8/project\_data/Health"

## Load the dataset 1

library(readxl)  
excel\_sheets('diabetes\_indicator.xlsx')

## [1] "Sheet1"

diabetes\_indicator\_df <- read\_excel('diabetes\_indicator.xlsx', sheet='Sheet1')  
diabetes\_indicator\_df

## # A tibble: 50 × 22  
## Diabetes\_…¹ HighBP HighC…² CholC…³ BMI Smoker Stroke Heart…⁴ PhysA…⁵ Fruits  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0 1 1 1 40 1 0 0 0 0  
## 2 0 0 0 0 25 1 0 0 1 0  
## 3 0 1 1 1 28 0 0 0 0 1  
## 4 0 1 0 1 27 0 0 0 1 1  
## 5 0 1 1 1 24 0 0 0 1 1  
## 6 0 1 1 1 25 1 0 0 1 1  
## 7 0 1 0 1 30 1 0 0 0 0  
## 8 0 1 1 1 25 1 0 0 1 0  
## 9 2 1 1 1 30 1 0 1 0 1  
## 10 0 0 0 1 24 0 0 0 0 0  
## # … with 40 more rows, 12 more variables: Veggies <dbl>,  
## # HvyAlcoholConsump <dbl>, AnyHealthcare <dbl>, NoDocbcCost <dbl>,  
## # GenHlth <dbl>, MentHlth <dbl>, PhysHlth <dbl>, DiffWalk <dbl>, Sex <dbl>,  
## # Age <dbl>, Education <dbl>, Income <dbl>, and abbreviated variable names  
## # ¹​Diabetes\_012, ²​HighChol, ³​CholCheck, ⁴​HeartDiseaseorAttack, ⁵​PhysActivity

#rename the “Diabetes\_012” column to “Diabetes” column to match columns with other dataframes

names(diabetes\_indicator\_df)[names(diabetes\_indicator\_df) == "Diabetes\_012"] <- "Diabetes"  
diabetes\_indicator\_df

## # A tibble: 50 × 22  
## Diabetes HighBP HighChol CholCheck BMI Smoker Stroke Heart…¹ PhysA…² Fruits  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0 1 1 1 40 1 0 0 0 0  
## 2 0 0 0 0 25 1 0 0 1 0  
## 3 0 1 1 1 28 0 0 0 0 1  
## 4 0 1 0 1 27 0 0 0 1 1  
## 5 0 1 1 1 24 0 0 0 1 1  
## 6 0 1 1 1 25 1 0 0 1 1  
## 7 0 1 0 1 30 1 0 0 0 0  
## 8 0 1 1 1 25 1 0 0 1 0  
## 9 2 1 1 1 30 1 0 1 0 1  
## 10 0 0 0 1 24 0 0 0 0 0  
## # … with 40 more rows, 12 more variables: Veggies <dbl>,  
## # HvyAlcoholConsump <dbl>, AnyHealthcare <dbl>, NoDocbcCost <dbl>,  
## # GenHlth <dbl>, MentHlth <dbl>, PhysHlth <dbl>, DiffWalk <dbl>, Sex <dbl>,  
## # Age <dbl>, Education <dbl>, Income <dbl>, and abbreviated variable names  
## # ¹​HeartDiseaseorAttack, ²​PhysActivity

summary(diabetes\_indicator\_df)

## Diabetes HighBP HighChol CholCheck BMI   
## Min. :0.00 Min. :0.00 Min. :0.00 Min. :0.00 Min. :21.00   
## 1st Qu.:0.00 1st Qu.:0.00 1st Qu.:0.00 1st Qu.:1.00 1st Qu.:24.25   
## Median :0.00 Median :1.00 Median :1.00 Median :1.00 Median :27.50   
## Mean :0.48 Mean :0.62 Mean :0.54 Mean :0.96 Mean :28.06   
## 3rd Qu.:0.00 3rd Qu.:1.00 3rd Qu.:1.00 3rd Qu.:1.00 3rd Qu.:31.00   
## Max. :2.00 Max. :1.00 Max. :1.00 Max. :1.00 Max. :40.00   
## Smoker Stroke HeartDiseaseorAttack PhysActivity Fruits   
## Min. :0.0 Min. :0.0 Min. :0.0 Min. :0.00 Min. :0.00   
## 1st Qu.:0.0 1st Qu.:0.0 1st Qu.:0.0 1st Qu.:0.00 1st Qu.:0.00   
## Median :1.0 Median :0.0 Median :0.0 Median :1.00 Median :1.00   
## Mean :0.6 Mean :0.1 Mean :0.1 Mean :0.52 Mean :0.58   
## 3rd Qu.:1.0 3rd Qu.:0.0 3rd Qu.:0.0 3rd Qu.:1.00 3rd Qu.:1.00   
## Max. :1.0 Max. :1.0 Max. :1.0 Max. :1.00 Max. :1.00   
## Veggies HvyAlcoholConsump AnyHealthcare NoDocbcCost GenHlth   
## Min. :0.00 Min. :0.00 Min. :0.0 Min. :0.00 Min. :1.00   
## 1st Qu.:1.00 1st Qu.:0.00 1st Qu.:1.0 1st Qu.:0.00 1st Qu.:2.00   
## Median :1.00 Median :0.00 Median :1.0 Median :0.00 Median :3.00   
## Mean :0.76 Mean :0.02 Mean :0.9 Mean :0.08 Mean :2.82   
## 3rd Qu.:1.00 3rd Qu.:0.00 3rd Qu.:1.0 3rd Qu.:0.00 3rd Qu.:3.00   
## Max. :1.00 Max. :1.00 Max. :1.0 Max. :1.00 Max. :5.00   
## MentHlth PhysHlth DiffWalk Sex Age   
## Min. : 0.0 Min. : 0.00 Min. :0.00 Min. :0.00 Min. : 2.00   
## 1st Qu.: 0.0 1st Qu.: 0.00 1st Qu.:0.00 1st Qu.:0.00 1st Qu.: 7.00   
## Median : 0.0 Median : 0.00 Median :0.00 Median :0.00 Median : 9.00   
## Mean : 6.5 Mean : 6.80 Mean :0.34 Mean :0.32 Mean : 8.94   
## 3rd Qu.: 9.0 3rd Qu.: 9.25 3rd Qu.:1.00 3rd Qu.:1.00 3rd Qu.:11.00   
## Max. :30.0 Max. :30.00 Max. :1.00 Max. :1.00 Max. :13.00   
## Education Income   
## Min. :2.0 Min. :1.00   
## 1st Qu.:4.0 1st Qu.:3.00   
## Median :5.0 Median :4.00   
## Mean :4.7 Mean :4.86   
## 3rd Qu.:6.0 3rd Qu.:7.00   
## Max. :6.0 Max. :8.00

library("psych")  
describe(diabetes\_indicator\_df)

## vars n mean sd median trimmed mad min max range  
## Diabetes 1 50 0.48 0.86 0.0 0.35 0.00 0 2 2  
## HighBP 2 50 0.62 0.49 1.0 0.65 0.00 0 1 1  
## HighChol 3 50 0.54 0.50 1.0 0.55 0.00 0 1 1  
## CholCheck 4 50 0.96 0.20 1.0 1.00 0.00 0 1 1  
## BMI 5 50 28.06 4.65 27.5 27.70 5.19 21 40 19  
## Smoker 6 50 0.60 0.49 1.0 0.62 0.00 0 1 1  
## Stroke 7 50 0.10 0.30 0.0 0.00 0.00 0 1 1  
## HeartDiseaseorAttack 8 50 0.10 0.30 0.0 0.00 0.00 0 1 1  
## PhysActivity 9 50 0.52 0.50 1.0 0.52 0.00 0 1 1  
## Fruits 10 50 0.58 0.50 1.0 0.60 0.00 0 1 1  
## Veggies 11 50 0.76 0.43 1.0 0.82 0.00 0 1 1  
## HvyAlcoholConsump 12 50 0.02 0.14 0.0 0.00 0.00 0 1 1  
## AnyHealthcare 13 50 0.90 0.30 1.0 1.00 0.00 0 1 1  
## NoDocbcCost 14 50 0.08 0.27 0.0 0.00 0.00 0 1 1  
## GenHlth 15 50 2.82 1.16 3.0 2.78 1.48 1 5 4  
## MentHlth 16 50 6.50 10.63 0.0 4.38 0.00 0 30 30  
## PhysHlth 17 50 6.80 11.12 0.0 4.75 0.00 0 30 30  
## DiffWalk 18 50 0.34 0.48 0.0 0.30 0.00 0 1 1  
## Sex 19 50 0.32 0.47 0.0 0.28 0.00 0 1 1  
## Age 20 50 8.94 2.78 9.0 9.10 2.97 2 13 11  
## Education 21 50 4.70 1.11 5.0 4.80 1.48 2 6 4  
## Income 22 50 4.86 2.35 4.0 4.92 2.97 1 8 7  
## skew kurtosis se  
## Diabetes 1.18 -0.62 0.12  
## HighBP -0.48 -1.80 0.07  
## HighChol -0.16 -2.01 0.07  
## CholCheck -4.55 19.13 0.03  
## BMI 0.60 -0.42 0.66  
## Smoker -0.40 -1.88 0.07  
## Stroke 2.59 4.79 0.04  
## HeartDiseaseorAttack 2.59 4.79 0.04  
## PhysActivity -0.08 -2.03 0.07  
## Fruits -0.31 -1.94 0.07  
## Veggies -1.18 -0.62 0.06  
## HvyAlcoholConsump 6.65 43.12 0.02  
## AnyHealthcare -2.59 4.79 0.04  
## NoDocbcCost 3.00 7.17 0.04  
## GenHlth 0.42 -0.64 0.16  
## MentHlth 1.36 0.24 1.50  
## PhysHlth 1.32 0.06 1.57  
## DiffWalk 0.66 -1.60 0.07  
## Sex 0.75 -1.47 0.07  
## Age -0.46 -0.55 0.39  
## Education -0.45 -0.55 0.16  
## Income -0.01 -1.41 0.33

## Load the dataset 2

excel\_sheets('diabetes\_indicator.xlsx')

## [1] "Sheet1"

diabetes\_indicator\_5050split\_df <- read\_excel('DiabetesIndicator\_5050split.xlsx', sheet='Sheet1')  
diabetes\_indicator\_5050split\_df

## # A tibble: 50 × 22  
## Diabetes\_…¹ HighBP HighC…² CholC…³ BMI Smoker Stroke Heart…⁴ PhysA…⁵ Fruits  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0 1 0 1 26 0 0 0 1 0  
## 2 0 1 1 1 26 1 1 0 0 1  
## 3 0 0 0 1 26 0 0 0 1 1  
## 4 0 1 1 1 28 1 0 0 1 1  
## 5 0 0 0 1 29 1 0 0 1 1  
## 6 0 0 0 1 18 0 0 0 1 1  
## 7 0 0 1 1 26 1 0 0 1 1  
## 8 0 0 0 1 31 1 0 0 0 1  
## 9 0 0 0 1 32 0 0 0 1 1  
## 10 0 0 0 1 27 1 0 0 0 1  
## # … with 40 more rows, 12 more variables: Veggies <dbl>,  
## # HvyAlcoholConsump <dbl>, AnyHealthcare <dbl>, NoDocbcCost <dbl>,  
## # GenHlth <dbl>, MentHlth <dbl>, PhysHlth <dbl>, DiffWalk <dbl>, Sex <dbl>,  
## # Age <dbl>, Education <dbl>, Income <dbl>, and abbreviated variable names  
## # ¹​Diabetes\_binary, ²​HighChol, ³​CholCheck, ⁴​HeartDiseaseorAttack,  
## # ⁵​PhysActivity

#rename the “Diabetes\_binary” column to “Diabetes” column to match columns with other dataframes

names(diabetes\_indicator\_5050split\_df)[names(diabetes\_indicator\_5050split\_df) == "Diabetes\_binary"] <- "Diabetes"  
diabetes\_indicator\_5050split\_df

## # A tibble: 50 × 22  
## Diabetes HighBP HighChol CholCheck BMI Smoker Stroke Heart…¹ PhysA…² Fruits  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0 1 0 1 26 0 0 0 1 0  
## 2 0 1 1 1 26 1 1 0 0 1  
## 3 0 0 0 1 26 0 0 0 1 1  
## 4 0 1 1 1 28 1 0 0 1 1  
## 5 0 0 0 1 29 1 0 0 1 1  
## 6 0 0 0 1 18 0 0 0 1 1  
## 7 0 0 1 1 26 1 0 0 1 1  
## 8 0 0 0 1 31 1 0 0 0 1  
## 9 0 0 0 1 32 0 0 0 1 1  
## 10 0 0 0 1 27 1 0 0 0 1  
## # … with 40 more rows, 12 more variables: Veggies <dbl>,  
## # HvyAlcoholConsump <dbl>, AnyHealthcare <dbl>, NoDocbcCost <dbl>,  
## # GenHlth <dbl>, MentHlth <dbl>, PhysHlth <dbl>, DiffWalk <dbl>, Sex <dbl>,  
## # Age <dbl>, Education <dbl>, Income <dbl>, and abbreviated variable names  
## # ¹​HeartDiseaseorAttack, ²​PhysActivity

summary(diabetes\_indicator\_5050split\_df)

## Diabetes HighBP HighChol CholCheck BMI   
## Min. :0 Min. :0.00 Min. :0.00 Min. :1 Min. :18.00   
## 1st Qu.:0 1st Qu.:0.00 1st Qu.:0.00 1st Qu.:1 1st Qu.:24.00   
## Median :0 Median :0.00 Median :0.00 Median :1 Median :26.50   
## Mean :0 Mean :0.32 Mean :0.38 Mean :1 Mean :27.56   
## 3rd Qu.:0 3rd Qu.:1.00 3rd Qu.:1.00 3rd Qu.:1 3rd Qu.:29.75   
## Max. :0 Max. :1.00 Max. :1.00 Max. :1 Max. :58.00   
## Smoker Stroke HeartDiseaseorAttack PhysActivity   
## Min. :0.00 Min. :0.00 Min. :0.00 Min. :0.00   
## 1st Qu.:0.00 1st Qu.:0.00 1st Qu.:0.00 1st Qu.:1.00   
## Median :0.00 Median :0.00 Median :0.00 Median :1.00   
## Mean :0.46 Mean :0.02 Mean :0.04 Mean :0.78   
## 3rd Qu.:1.00 3rd Qu.:0.00 3rd Qu.:0.00 3rd Qu.:1.00   
## Max. :1.00 Max. :1.00 Max. :1.00 Max. :1.00   
## Fruits Veggies HvyAlcoholConsump AnyHealthcare NoDocbcCost   
## Min. :0.00 Min. :0.00 Min. :0.00 Min. :0.00 Min. :0.00   
## 1st Qu.:0.25 1st Qu.:1.00 1st Qu.:0.00 1st Qu.:1.00 1st Qu.:0.00   
## Median :1.00 Median :1.00 Median :0.00 Median :1.00 Median :0.00   
## Mean :0.74 Mean :0.86 Mean :0.06 Mean :0.96 Mean :0.04   
## 3rd Qu.:1.00 3rd Qu.:1.00 3rd Qu.:0.00 3rd Qu.:1.00 3rd Qu.:0.00   
## Max. :1.00 Max. :1.00 Max. :1.00 Max. :1.00 Max. :1.00   
## GenHlth MentHlth PhysHlth DiffWalk Sex   
## Min. :1.00 Min. : 0.00 Min. : 0.00 Min. :0.00 Min. :0.0   
## 1st Qu.:2.00 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.:0.00 1st Qu.:0.0   
## Median :2.00 Median : 0.00 Median : 0.00 Median :0.00 Median :0.5   
## Mean :2.32 Mean : 1.76 Mean : 3.36 Mean :0.06 Mean :0.5   
## 3rd Qu.:3.00 3rd Qu.: 0.00 3rd Qu.: 3.00 3rd Qu.:0.00 3rd Qu.:1.0   
## Max. :5.00 Max. :30.00 Max. :30.00 Max. :1.00 Max. :1.0   
## Age Education Income   
## Min. : 1.00 Min. :4.00 Min. :1.0   
## 1st Qu.: 5.00 1st Qu.:5.00 1st Qu.:6.0   
## Median : 8.00 Median :5.00 Median :7.0   
## Mean : 7.54 Mean :5.12 Mean :6.4   
## 3rd Qu.:10.00 3rd Qu.:6.00 3rd Qu.:8.0   
## Max. :13.00 Max. :6.00 Max. :8.0

describe(diabetes\_indicator\_5050split\_df)

## vars n mean sd median trimmed mad min max range skew  
## Diabetes 1 50 0.00 0.00 0.0 0.00 0.00 0 0 0 NaN  
## HighBP 2 50 0.32 0.47 0.0 0.28 0.00 0 1 1 0.75  
## HighChol 3 50 0.38 0.49 0.0 0.35 0.00 0 1 1 0.48  
## CholCheck 4 50 1.00 0.00 1.0 1.00 0.00 1 1 0 NaN  
## BMI 5 50 27.56 7.28 26.5 26.55 4.45 18 58 40 1.82  
## Smoker 6 50 0.46 0.50 0.0 0.45 0.00 0 1 1 0.16  
## Stroke 7 50 0.02 0.14 0.0 0.00 0.00 0 1 1 6.65  
## HeartDiseaseorAttack 8 50 0.04 0.20 0.0 0.00 0.00 0 1 1 4.55  
## PhysActivity 9 50 0.78 0.42 1.0 0.85 0.00 0 1 1 -1.31  
## Fruits 10 50 0.74 0.44 1.0 0.80 0.00 0 1 1 -1.06  
## Veggies 11 50 0.86 0.35 1.0 0.95 0.00 0 1 1 -2.01  
## HvyAlcoholConsump 12 50 0.06 0.24 0.0 0.00 0.00 0 1 1 3.59  
## AnyHealthcare 13 50 0.96 0.20 1.0 1.00 0.00 0 1 1 -4.55  
## NoDocbcCost 14 50 0.04 0.20 0.0 0.00 0.00 0 1 1 4.55  
## GenHlth 15 50 2.32 1.06 2.0 2.22 1.48 1 5 4 0.57  
## MentHlth 16 50 1.76 5.21 0.0 0.48 0.00 0 30 30 4.06  
## PhysHlth 17 50 3.36 7.59 0.0 1.23 0.00 0 30 30 2.68  
## DiffWalk 18 50 0.06 0.24 0.0 0.00 0.00 0 1 1 3.59  
## Sex 19 50 0.50 0.51 0.5 0.50 0.74 0 1 1 0.00  
## Age 20 50 7.54 3.13 8.0 7.53 2.97 1 13 12 -0.05  
## Education 21 50 5.12 0.77 5.0 5.15 1.48 4 6 2 -0.20  
## Income 22 50 6.40 2.06 7.0 6.78 1.48 1 8 7 -1.23  
## kurtosis se  
## Diabetes NaN 0.00  
## HighBP -1.47 0.07  
## HighChol -1.80 0.07  
## CholCheck NaN 0.00  
## BMI 4.82 1.03  
## Smoker -2.01 0.07  
## Stroke 43.12 0.02  
## HeartDiseaseorAttack 19.13 0.03  
## PhysActivity -0.28 0.06  
## Fruits -0.89 0.06  
## Veggies 2.10 0.05  
## HvyAlcoholConsump 11.15 0.03  
## AnyHealthcare 19.13 0.03  
## NoDocbcCost 19.13 0.03  
## GenHlth -0.22 0.15  
## MentHlth 17.29 0.74  
## PhysHlth 6.31 1.07  
## DiffWalk 11.15 0.03  
## Sex -2.04 0.07  
## Age -1.02 0.44  
## Education -1.34 0.11  
## Income 0.41 0.29

## Load the dataset 3

excel\_sheets('diabetes\_indicator.xlsx')

## [1] "Sheet1"

diabetes\_binary\_df <- read\_excel('diabetes\_binary.xlsx', sheet='Sheet1')  
diabetes\_binary\_df

## # A tibble: 50 × 22  
## Diabetes\_…¹ HighBP HighC…² CholC…³ BMI Smoker Stroke Heart…⁴ PhysA…⁵ Fruits  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0 1 1 1 40 1 0 0 0 0  
## 2 0 0 0 0 25 1 0 0 1 0  
## 3 0 1 1 1 28 0 0 0 0 1  
## 4 0 1 0 1 27 0 0 0 1 1  
## 5 0 1 1 1 24 0 0 0 1 1  
## 6 0 1 1 1 25 1 0 0 1 1  
## 7 0 1 0 1 30 1 0 0 0 0  
## 8 0 1 1 1 25 1 0 0 1 0  
## 9 1 1 1 1 30 1 0 1 0 1  
## 10 0 0 0 1 24 0 0 0 0 0  
## # … with 40 more rows, 12 more variables: Veggies <dbl>,  
## # HvyAlcoholConsump <dbl>, AnyHealthcare <dbl>, NoDocbcCost <dbl>,  
## # GenHlth <dbl>, MentHlth <dbl>, PhysHlth <dbl>, DiffWalk <dbl>, Sex <dbl>,  
## # Age <dbl>, Education <dbl>, Income <dbl>, and abbreviated variable names  
## # ¹​Diabetes\_binary, ²​HighChol, ³​CholCheck, ⁴​HeartDiseaseorAttack,  
## # ⁵​PhysActivity

#rename the “Diabetes\_binary” column to “Diabetes” column to match columns with other dataframes

names(diabetes\_binary\_df)[names(diabetes\_binary\_df) == "Diabetes\_binary"] <- "Diabetes"  
diabetes\_binary\_df

## # A tibble: 50 × 22  
## Diabetes HighBP HighChol CholCheck BMI Smoker Stroke Heart…¹ PhysA…² Fruits  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0 1 1 1 40 1 0 0 0 0  
## 2 0 0 0 0 25 1 0 0 1 0  
## 3 0 1 1 1 28 0 0 0 0 1  
## 4 0 1 0 1 27 0 0 0 1 1  
## 5 0 1 1 1 24 0 0 0 1 1  
## 6 0 1 1 1 25 1 0 0 1 1  
## 7 0 1 0 1 30 1 0 0 0 0  
## 8 0 1 1 1 25 1 0 0 1 0  
## 9 1 1 1 1 30 1 0 1 0 1  
## 10 0 0 0 1 24 0 0 0 0 0  
## # … with 40 more rows, 12 more variables: Veggies <dbl>,  
## # HvyAlcoholConsump <dbl>, AnyHealthcare <dbl>, NoDocbcCost <dbl>,  
## # GenHlth <dbl>, MentHlth <dbl>, PhysHlth <dbl>, DiffWalk <dbl>, Sex <dbl>,  
## # Age <dbl>, Education <dbl>, Income <dbl>, and abbreviated variable names  
## # ¹​HeartDiseaseorAttack, ²​PhysActivity

summary(diabetes\_binary\_df)

## Diabetes HighBP HighChol CholCheck BMI   
## Min. :0.00 Min. :0.00 Min. :0.00 Min. :0.00 Min. :21.00   
## 1st Qu.:0.00 1st Qu.:0.00 1st Qu.:0.00 1st Qu.:1.00 1st Qu.:24.25   
## Median :0.00 Median :1.00 Median :1.00 Median :1.00 Median :27.50   
## Mean :0.24 Mean :0.62 Mean :0.54 Mean :0.96 Mean :28.06   
## 3rd Qu.:0.00 3rd Qu.:1.00 3rd Qu.:1.00 3rd Qu.:1.00 3rd Qu.:31.00   
## Max. :1.00 Max. :1.00 Max. :1.00 Max. :1.00 Max. :40.00   
## Smoker Stroke HeartDiseaseorAttack PhysActivity Fruits   
## Min. :0.0 Min. :0.0 Min. :0.0 Min. :0.00 Min. :0.00   
## 1st Qu.:0.0 1st Qu.:0.0 1st Qu.:0.0 1st Qu.:0.00 1st Qu.:0.00   
## Median :1.0 Median :0.0 Median :0.0 Median :1.00 Median :1.00   
## Mean :0.6 Mean :0.1 Mean :0.1 Mean :0.52 Mean :0.58   
## 3rd Qu.:1.0 3rd Qu.:0.0 3rd Qu.:0.0 3rd Qu.:1.00 3rd Qu.:1.00   
## Max. :1.0 Max. :1.0 Max. :1.0 Max. :1.00 Max. :1.00   
## Veggies HvyAlcoholConsump AnyHealthcare NoDocbcCost GenHlth   
## Min. :0.00 Min. :0.00 Min. :0.0 Min. :0.00 Min. :1.00   
## 1st Qu.:1.00 1st Qu.:0.00 1st Qu.:1.0 1st Qu.:0.00 1st Qu.:2.00   
## Median :1.00 Median :0.00 Median :1.0 Median :0.00 Median :3.00   
## Mean :0.76 Mean :0.02 Mean :0.9 Mean :0.08 Mean :2.82   
## 3rd Qu.:1.00 3rd Qu.:0.00 3rd Qu.:1.0 3rd Qu.:0.00 3rd Qu.:3.00   
## Max. :1.00 Max. :1.00 Max. :1.0 Max. :1.00 Max. :5.00   
## MentHlth PhysHlth DiffWalk Sex Age   
## Min. : 0.0 Min. : 0.00 Min. :0.00 Min. :0.00 Min. : 2.00   
## 1st Qu.: 0.0 1st Qu.: 0.00 1st Qu.:0.00 1st Qu.:0.00 1st Qu.: 7.00   
## Median : 0.0 Median : 0.00 Median :0.00 Median :0.00 Median : 9.00   
## Mean : 6.5 Mean : 6.80 Mean :0.34 Mean :0.32 Mean : 8.94   
## 3rd Qu.: 9.0 3rd Qu.: 9.25 3rd Qu.:1.00 3rd Qu.:1.00 3rd Qu.:11.00   
## Max. :30.0 Max. :30.00 Max. :1.00 Max. :1.00 Max. :13.00   
## Education Income   
## Min. :2.0 Min. :1.00   
## 1st Qu.:4.0 1st Qu.:3.00   
## Median :5.0 Median :4.00   
## Mean :4.7 Mean :4.86   
## 3rd Qu.:6.0 3rd Qu.:7.00   
## Max. :6.0 Max. :8.00

describe(diabetes\_binary\_df)

## vars n mean sd median trimmed mad min max range  
## Diabetes 1 50 0.24 0.43 0.0 0.17 0.00 0 1 1  
## HighBP 2 50 0.62 0.49 1.0 0.65 0.00 0 1 1  
## HighChol 3 50 0.54 0.50 1.0 0.55 0.00 0 1 1  
## CholCheck 4 50 0.96 0.20 1.0 1.00 0.00 0 1 1  
## BMI 5 50 28.06 4.65 27.5 27.70 5.19 21 40 19  
## Smoker 6 50 0.60 0.49 1.0 0.62 0.00 0 1 1  
## Stroke 7 50 0.10 0.30 0.0 0.00 0.00 0 1 1  
## HeartDiseaseorAttack 8 50 0.10 0.30 0.0 0.00 0.00 0 1 1  
## PhysActivity 9 50 0.52 0.50 1.0 0.52 0.00 0 1 1  
## Fruits 10 50 0.58 0.50 1.0 0.60 0.00 0 1 1  
## Veggies 11 50 0.76 0.43 1.0 0.82 0.00 0 1 1  
## HvyAlcoholConsump 12 50 0.02 0.14 0.0 0.00 0.00 0 1 1  
## AnyHealthcare 13 50 0.90 0.30 1.0 1.00 0.00 0 1 1  
## NoDocbcCost 14 50 0.08 0.27 0.0 0.00 0.00 0 1 1  
## GenHlth 15 50 2.82 1.16 3.0 2.78 1.48 1 5 4  
## MentHlth 16 50 6.50 10.63 0.0 4.38 0.00 0 30 30  
## PhysHlth 17 50 6.80 11.12 0.0 4.75 0.00 0 30 30  
## DiffWalk 18 50 0.34 0.48 0.0 0.30 0.00 0 1 1  
## Sex 19 50 0.32 0.47 0.0 0.28 0.00 0 1 1  
## Age 20 50 8.94 2.78 9.0 9.10 2.97 2 13 11  
## Education 21 50 4.70 1.11 5.0 4.80 1.48 2 6 4  
## Income 22 50 4.86 2.35 4.0 4.92 2.97 1 8 7  
## skew kurtosis se  
## Diabetes 1.18 -0.62 0.06  
## HighBP -0.48 -1.80 0.07  
## HighChol -0.16 -2.01 0.07  
## CholCheck -4.55 19.13 0.03  
## BMI 0.60 -0.42 0.66  
## Smoker -0.40 -1.88 0.07  
## Stroke 2.59 4.79 0.04  
## HeartDiseaseorAttack 2.59 4.79 0.04  
## PhysActivity -0.08 -2.03 0.07  
## Fruits -0.31 -1.94 0.07  
## Veggies -1.18 -0.62 0.06  
## HvyAlcoholConsump 6.65 43.12 0.02  
## AnyHealthcare -2.59 4.79 0.04  
## NoDocbcCost 3.00 7.17 0.04  
## GenHlth 0.42 -0.64 0.16  
## MentHlth 1.36 0.24 1.50  
## PhysHlth 1.32 0.06 1.57  
## DiffWalk 0.66 -1.60 0.07  
## Sex 0.75 -1.47 0.07  
## Age -0.46 -0.55 0.39  
## Education -0.45 -0.55 0.16  
## Income -0.01 -1.41 0.33

## Bind all the three datasets into one dataset

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

diabetes\_df <- bind\_rows(diabetes\_indicator\_df, diabetes\_indicator\_5050split\_df, diabetes\_binary\_df)  
diabetes\_df

## # A tibble: 150 × 22  
## Diabetes HighBP HighChol CholCheck BMI Smoker Stroke Heart…¹ PhysA…² Fruits  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0 1 1 1 40 1 0 0 0 0  
## 2 0 0 0 0 25 1 0 0 1 0  
## 3 0 1 1 1 28 0 0 0 0 1  
## 4 0 1 0 1 27 0 0 0 1 1  
## 5 0 1 1 1 24 0 0 0 1 1  
## 6 0 1 1 1 25 1 0 0 1 1  
## 7 0 1 0 1 30 1 0 0 0 0  
## 8 0 1 1 1 25 1 0 0 1 0  
## 9 2 1 1 1 30 1 0 1 0 1  
## 10 0 0 0 1 24 0 0 0 0 0  
## # … with 140 more rows, 12 more variables: Veggies <dbl>,  
## # HvyAlcoholConsump <dbl>, AnyHealthcare <dbl>, NoDocbcCost <dbl>,  
## # GenHlth <dbl>, MentHlth <dbl>, PhysHlth <dbl>, DiffWalk <dbl>, Sex <dbl>,  
## # Age <dbl>, Education <dbl>, Income <dbl>, and abbreviated variable names  
## # ¹​HeartDiseaseorAttack, ²​PhysActivity

summary(diabetes\_df)

## Diabetes HighBP HighChol CholCheck   
## Min. :0.00 Min. :0.00 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.00 1st Qu.:0.00 1st Qu.:0.0000 1st Qu.:1.0000   
## Median :0.00 Median :1.00 Median :0.0000 Median :1.0000   
## Mean :0.24 Mean :0.52 Mean :0.4867 Mean :0.9733   
## 3rd Qu.:0.00 3rd Qu.:1.00 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :2.00 Max. :1.00 Max. :1.0000 Max. :1.0000   
## BMI Smoker Stroke HeartDiseaseorAttack  
## Min. :18.00 Min. :0.0000 Min. :0.00000 Min. :0.00   
## 1st Qu.:24.00 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.00   
## Median :27.00 Median :1.0000 Median :0.00000 Median :0.00   
## Mean :27.89 Mean :0.5533 Mean :0.07333 Mean :0.08   
## 3rd Qu.:31.00 3rd Qu.:1.0000 3rd Qu.:0.00000 3rd Qu.:0.00   
## Max. :58.00 Max. :1.0000 Max. :1.00000 Max. :1.00   
## PhysActivity Fruits Veggies HvyAlcoholConsump  
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:0.00000   
## Median :1.0000 Median :1.0000 Median :1.0000 Median :0.00000   
## Mean :0.6067 Mean :0.6333 Mean :0.7933 Mean :0.03333   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.00000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.00000   
## AnyHealthcare NoDocbcCost GenHlth MentHlth   
## Min. :0.00 Min. :0.00000 Min. :1.000 Min. : 0.00   
## 1st Qu.:1.00 1st Qu.:0.00000 1st Qu.:2.000 1st Qu.: 0.00   
## Median :1.00 Median :0.00000 Median :3.000 Median : 0.00   
## Mean :0.92 Mean :0.06667 Mean :2.653 Mean : 4.92   
## 3rd Qu.:1.00 3rd Qu.:0.00000 3rd Qu.:3.000 3rd Qu.: 5.00   
## Max. :1.00 Max. :1.00000 Max. :5.000 Max. :30.00   
## PhysHlth DiffWalk Sex Age   
## Min. : 0.000 Min. :0.0000 Min. :0.00 Min. : 1.000   
## 1st Qu.: 0.000 1st Qu.:0.0000 1st Qu.:0.00 1st Qu.: 7.000   
## Median : 0.000 Median :0.0000 Median :0.00 Median : 9.000   
## Mean : 5.653 Mean :0.2467 Mean :0.38 Mean : 8.473   
## 3rd Qu.: 5.750 3rd Qu.:0.0000 3rd Qu.:1.00 3rd Qu.:11.000   
## Max. :30.000 Max. :1.0000 Max. :1.00 Max. :13.000   
## Education Income   
## Min. :2.00 Min. :1.000   
## 1st Qu.:4.00 1st Qu.:3.000   
## Median :5.00 Median :6.000   
## Mean :4.84 Mean :5.373   
## 3rd Qu.:6.00 3rd Qu.:8.000   
## Max. :6.00 Max. :8.000

describe(diabetes\_df)

## vars n mean sd median trimmed mad min max range  
## Diabetes 1 150 0.24 0.59 0 0.07 0.00 0 2 2  
## HighBP 2 150 0.52 0.50 1 0.52 0.00 0 1 1  
## HighChol 3 150 0.49 0.50 0 0.48 0.00 0 1 1  
## CholCheck 4 150 0.97 0.16 1 1.00 0.00 0 1 1  
## BMI 5 150 27.89 5.63 27 27.38 4.45 18 58 40  
## Smoker 6 150 0.55 0.50 1 0.57 0.00 0 1 1  
## Stroke 7 150 0.07 0.26 0 0.00 0.00 0 1 1  
## HeartDiseaseorAttack 8 150 0.08 0.27 0 0.00 0.00 0 1 1  
## PhysActivity 9 150 0.61 0.49 1 0.63 0.00 0 1 1  
## Fruits 10 150 0.63 0.48 1 0.67 0.00 0 1 1  
## Veggies 11 150 0.79 0.41 1 0.87 0.00 0 1 1  
## HvyAlcoholConsump 12 150 0.03 0.18 0 0.00 0.00 0 1 1  
## AnyHealthcare 13 150 0.92 0.27 1 1.00 0.00 0 1 1  
## NoDocbcCost 14 150 0.07 0.25 0 0.00 0.00 0 1 1  
## GenHlth 15 150 2.65 1.14 3 2.58 1.48 1 5 4  
## MentHlth 16 150 4.92 9.40 0 2.48 0.00 0 30 30  
## PhysHlth 17 150 5.65 10.15 0 3.32 0.00 0 30 30  
## DiffWalk 18 150 0.25 0.43 0 0.18 0.00 0 1 1  
## Sex 19 150 0.38 0.49 0 0.35 0.00 0 1 1  
## Age 20 150 8.47 2.96 9 8.58 2.97 1 13 12  
## Education 21 150 4.84 1.02 5 4.93 1.48 2 6 4  
## Income 22 150 5.37 2.36 6 5.55 2.97 1 8 7  
## skew kurtosis se  
## Diabetes 2.27 3.71 0.05  
## HighBP -0.08 -2.01 0.04  
## HighChol 0.05 -2.01 0.04  
## CholCheck -5.82 32.06 0.01  
## BMI 1.48 4.81 0.46  
## Smoker -0.21 -1.97 0.04  
## Stroke 3.24 8.56 0.02  
## HeartDiseaseorAttack 3.07 7.45 0.02  
## PhysActivity -0.43 -1.83 0.04  
## Fruits -0.55 -1.71 0.04  
## Veggies -1.43 0.06 0.03  
## HvyAlcoholConsump 5.15 24.66 0.01  
## AnyHealthcare -3.07 7.45 0.02  
## NoDocbcCost 3.44 9.90 0.02  
## GenHlth 0.48 -0.44 0.09  
## MentHlth 1.85 1.97 0.77  
## PhysHlth 1.67 1.22 0.83  
## DiffWalk 1.16 -0.65 0.04  
## Sex 0.49 -1.77 0.04  
## Age -0.36 -0.72 0.24  
## Education -0.57 -0.20 0.08  
## Income -0.36 -1.28 0.19

## omit the data with Na values.

na.omit(diabetes\_df)

## # A tibble: 150 × 22  
## Diabetes HighBP HighChol CholCheck BMI Smoker Stroke Heart…¹ PhysA…² Fruits  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0 1 1 1 40 1 0 0 0 0  
## 2 0 0 0 0 25 1 0 0 1 0  
## 3 0 1 1 1 28 0 0 0 0 1  
## 4 0 1 0 1 27 0 0 0 1 1  
## 5 0 1 1 1 24 0 0 0 1 1  
## 6 0 1 1 1 25 1 0 0 1 1  
## 7 0 1 0 1 30 1 0 0 0 0  
## 8 0 1 1 1 25 1 0 0 1 0  
## 9 2 1 1 1 30 1 0 1 0 1  
## 10 0 0 0 1 24 0 0 0 0 0  
## # … with 140 more rows, 12 more variables: Veggies <dbl>,  
## # HvyAlcoholConsump <dbl>, AnyHealthcare <dbl>, NoDocbcCost <dbl>,  
## # GenHlth <dbl>, MentHlth <dbl>, PhysHlth <dbl>, DiffWalk <dbl>, Sex <dbl>,  
## # Age <dbl>, Education <dbl>, Income <dbl>, and abbreviated variable names  
## # ¹​HeartDiseaseorAttack, ²​PhysActivity

## here there are no NA values in the data

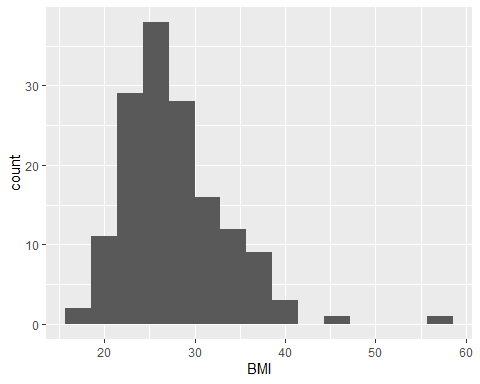
## Analyze how Diabetes depends on BMI

# Create a Histogram of the BMI variable using the ggplot2 package.  
  
library(ggplot2)

##   
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

ggplot(diabetes\_df,aes(BMI)) + geom\_histogram(bins = 15)

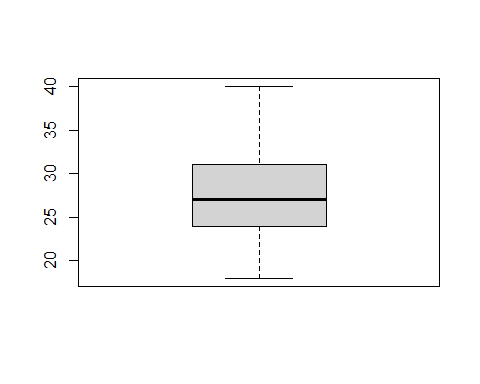


## remove the outliers  
x <- diabetes\_df$`BMI` # Print data  
  
x\_out\_rm <- x[!x %in% boxplot.stats(x)$out] # Remove the outliers  
  
length(x) - length(x\_out\_rm) # Count the removed observations

## [1] 2

## Create boxplot without outliers

boxplot(x\_out\_rm)



## Using GroupBy function from dplyr package to group by Diabetes

library(dplyr)  
diabetes\_df %>% group\_by(Diabetes) %>% summarize(AvgBMI = mean(`BMI`))

## # A tibble: 3 × 2  
## Diabetes AvgBMI  
## <dbl> <dbl>  
## 1 0 27.7  
## 2 1 28.9  
## 3 2 28.9

## 0 = no diabetes 1 = prediabetes 2 = diabetes

## So, based on the above analysis, the BMI should be maintained around 27.7 in order to reduce the chances of Diabetes.

## Furthur analysis will be done on other predictors

# References

## *[Datasets from Kaggle website](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset)*