Final Project Step3

Basava, Madhuri

2023-03-04

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

1) How can we reduce diabetes cases in the future?
2) What are the factors affecting diabetes?
3) How much Physical activity in a certain period is needed to reduce diabetes cases?
4) Is smoking, a direct or indirect cause of diabetes?
5) How much BMI value range should a person have to reduce the possibility of diabetes?
6) Does High Blood Pressure, a reason for diabetes?
7) Are Males or Females more prone to diabetes?
8) What age people are getting affected by diabetes more?
9) Is a person's heart attack/stroke has to be more careful?
10) Will high cholesterol lead to Diabetes?
11) Will heavy alcohol consumption lead to Diabetes?
12) 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)

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')
head(diabetes_indicator_df)
```

```
## # A tibble: 6 x 22
##   Diabetes_012 HighBP HighC~1 CholC~2   BMI Smoker Stroke Heart~3 PhysA~4 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
## # ... with 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: HighChol, 2: CholCheck,
## #   3: HeartDiseaseorAttack, 4: 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"
head(diabetes_indicator_df)
```

```
## # A tibble: 6 x 22
##   Diabetes HighBP HighChol CholCheck   BMI Smoker Stroke HeartD~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
```

```
## # ... with 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')
head(diabetes_indicator_5050split_df)
```

```
## # A tibble: 6 x 22
```

```
##   Diabetes_b~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
## # ... with 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"] <- "
head(diabetes_indicator_5050split_df)
```

```
## # A tibble: 6 x 22
##   Diabetes HighBP HighChol CholCheck   BMI Smoker Stroke HeartD~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
## # ... with 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')
head(diabetes_binary_df)
```

```
## # A tibble: 6 x 22
##   Diabetes_b~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
## # ... with 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
```

Data importing and cleaning steps are explained in the text and follow a logical process. Outline your data preparation and cleansing steps.

I have followed a step by step process

1) Rename the "Diabetes_binary" column to "Diabetes" column to match columns with other dataframes.

2) Combined the 3 dataframes into one data frame and

3) Omit the data with Na values.

4) Remove the outliers.

STEP 1: 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"
head(diabetes_binary_df)
```

```
## # A tibble: 6 x 22
##   Diabetes HighBP HighChol CholCheck   BMI Smoker Stroke HeartD~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
## # ... with 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
```

STEP 2: 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)
head(diabetes_df)
```

```
## # A tibble: 6 x 22
##   Diabetes HighBP HighChol CholCheck   BMI Smoker Stroke HeartD~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
## # ... with 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
```

```r
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
```

```r
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
```

##Look at the data

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1     v purrr   1.0.1
## v tibble  3.1.8     v stringr 1.5.0
## v tidyr   1.3.0     v forcats 0.5.2
## v readr   2.1.3
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x ggplot2::%+%()   masks psych::%+%()
```

```
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
glimpse(diabetes_df)
```

```
## Rows: 150
## Columns: 22
## $ Diabetes             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 2, 0, 0, 2, 0, 0, 0~
## $ HighBP               <dbl> 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1~
## $ HighChol             <dbl> 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1~
## $ CholCheck            <dbl> 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ BMI                  <dbl> 40, 25, 28, 27, 24, 25, 30, 25, 30, 24, 25, 34, 2~
## $ Smoker               <dbl> 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0~
## $ Stroke               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0~
## $ HeartDiseaseorAttack <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0~
## $ PhysActivity         <dbl> 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1~
## $ Fruits               <dbl> 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1~
## $ Veggies              <dbl> 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1~
## $ HvyAlcoholConsump    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AnyHealthcare        <dbl> 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ NoDocbcCost          <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0~
## $ GenHlth              <dbl> 5, 3, 5, 2, 2, 2, 3, 3, 5, 2, 3, 3, 3, 4, 4, 2, 3~
## $ MentHlth             <dbl> 18, 0, 30, 0, 3, 0, 0, 0, 30, 0, 0, 0, 0, 0, 30, ~
## $ PhysHlth             <dbl> 15, 0, 30, 0, 0, 2, 14, 0, 30, 0, 0, 30, 15, 0, 2~
## $ DiffWalk             <dbl> 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0~
## $ Sex                  <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0~
## $ Age                  <dbl> 9, 7, 9, 11, 11, 10, 9, 11, 9, 8, 13, 10, 7, 11, ~
## $ Education            <dbl> 4, 6, 4, 3, 5, 6, 6, 4, 5, 4, 6, 5, 5, 4, 6, 6, 4~
## $ Income               <dbl> 3, 1, 8, 6, 4, 8, 7, 4, 1, 3, 8, 1, 7, 6, 2, 8, 3~
```

STEP 3: Omit the data with Na values.

```
na.omit(diabetes_df)
```

```
## # A tibble: 150 x 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
```

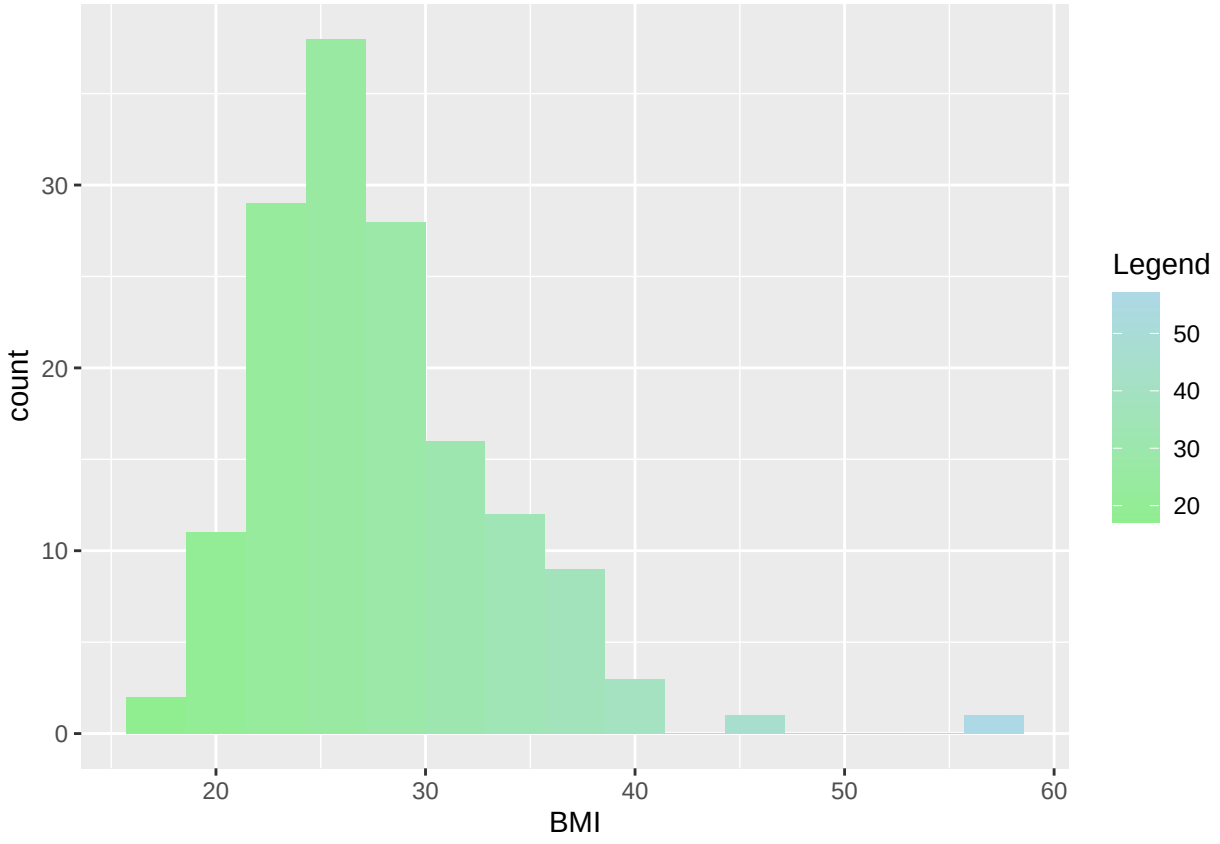
Here there are no NA values in the data

##STEP 4: Remove the outlierts

Analyze how Diabetes depends on BMI

```
# Create a Histogram of the BMI variable using the ggplot2 package.

library(ggplot2)
ggplot(diabetes_df,aes(BMI, fill=..x..)) + geom_histogram(bins = 15) +
  scale_fill_gradient("Legend",low = "lightgreen", high = "lightblue")
```

```
## Warning: The dot-dot notation (‘..x..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(x)‘ instead.
```



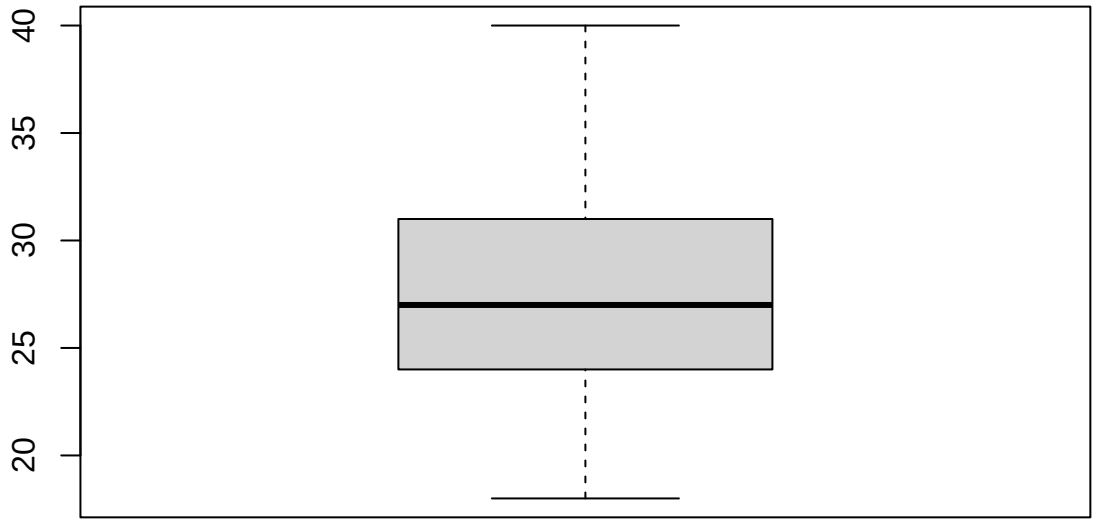

```
## remove the outliers
x <- diabetes_df$`BMI`                                          # Print data

x_out_rm <- x[!x %in% boxplot.stats(x)$out]     # Remove the outliers
```

```r
length(x) - length(x_out_rm)                    # Count the removed observations
```

```
## [1] 2
```

Create boxplot without outliers

```r
boxplot(x_out_rm)
```



With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.

```r
head(diabetes_df)
```

```
## # A tibble: 6 x 22
##   Diabetes HighBP HighChol CholCheck   BMI Smoker Stroke HeartD~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
## # ... with 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
```

What do you not know how to do right now that you need to learn to import and cleanup your dataset?

Now, I do not know how to show a data frame after the outliers are removed.I need to learn it.

I just tried to figure it out an created a dataframe by removing outliers.

Removing outlier data from data frame

```
diabetes_df <- diabetes_df[(diabetes_df$`BMI` %in% x_out_rm ),]
head(diabetes_df)
```

```
## # A tibble: 6 x 22
##   Diabetes HighBP HighChol CholCheck   BMI Smoker Stroke HeartD~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
## # ... with 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.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000
##  Median :0.0000   Median :1.0000   Median :0.0000   Median :1.000
##  Mean   :0.2432   Mean   :0.5135   Mean   :0.4865   Mean   :0.973
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :2.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000
##       BMI            Smoker          Stroke        HeartDiseaseorAttack
##  Min.   :18.00   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:24.00   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
```

```
##  Median :27.00   Median :1.0000   Median :0.00000   Median :0.00000
##  Mean   :27.56   Mean   :0.5541   Mean   :0.07432   Mean   :0.08108
##  3rd Qu.:31.00   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :40.00   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##   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   :0.6149   Mean   :0.6284   Mean   :0.7905   Mean   :0.03378
##  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.   :0.0000   Min.   :0.00000   Min.   :1.000   Min.   : 0.000
##  1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:2.000   1st Qu.: 0.000
##  Median :1.0000   Median :0.00000   Median :3.000   Median : 0.000
##  Mean   :0.9189   Mean   :0.06757   Mean   :2.662   Mean   : 4.966
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:3.000   3rd Qu.: 5.000
##  Max.   :1.0000   Max.   :1.00000   Max.   :5.000   Max.   :30.000
##     PhysHlth         DiffWalk          Sex              Age
##  Min.   : 0.000   Min.   :0.00   Min.   :0.0000   Min.   : 1.0
##  1st Qu.: 0.000   1st Qu.:0.00   1st Qu.:0.0000   1st Qu.: 7.0
##  Median : 0.000   Median :0.00   Median :0.0000   Median : 9.0
##  Mean   : 5.709   Mean   :0.25   Mean   :0.3716   Mean   : 8.5
##  3rd Qu.: 6.250   3rd Qu.:0.25   3rd Qu.:1.0000   3rd Qu.:11.0
##  Max.   :30.000   Max.   :1.00   Max.   :1.0000   Max.   :13.0
##    Education         Income
##  Min.   :2.000   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:3.000
##  Median :5.000   Median :6.000
##  Mean   :4.838   Mean   :5.392
##  3rd Qu.:6.000   3rd Qu.:8.000
##  Max.   :6.000   Max.   :8.000
```

Discuss how you plan to uncover new information in the data that is not self-evident.

I will use different functions like groupby and summerize to calculate the average BMI to analyze how much BMI should be maintained by an individual to avoid Diabetis.

Using GroupBy function from dplyr package to group by Diabetes

```
library(dplyr)

diabetes_df %>% group_by(Diabetes) %>% summarize(AvgBMI = mean(`BMI`))
```

```
## # A tibble: 3 x 2
##   Diabetes AvgBMI
##      <dbl>  <dbl>
## 1        0   27.3
## 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.

What are different ways you could look at this data to answer the questions you want to answer?

I try to use different functions and fugure out the averages in order to analyze the data and find the answers.

Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

Yes i plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information as shown in the above example on BMI.

##I combined the 3 data sets into one. sliced the data to remove outliers. Create a new variable on average BMI(AvgBMI) and summarize it.

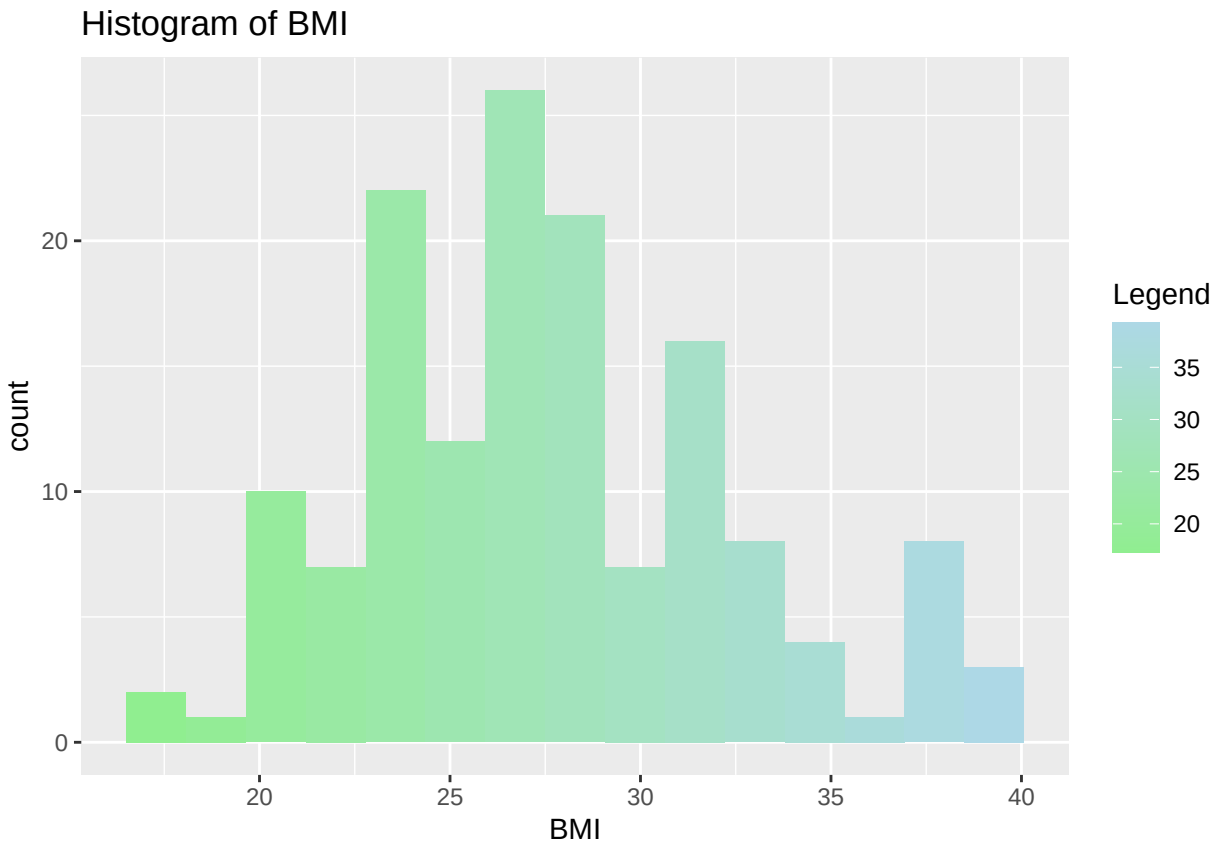
How could you summarize your data to answer key questions?

Based on the summarized values on Average BMI , It should be maintained around 27.7 in order to reduce the chances of Diabetes.

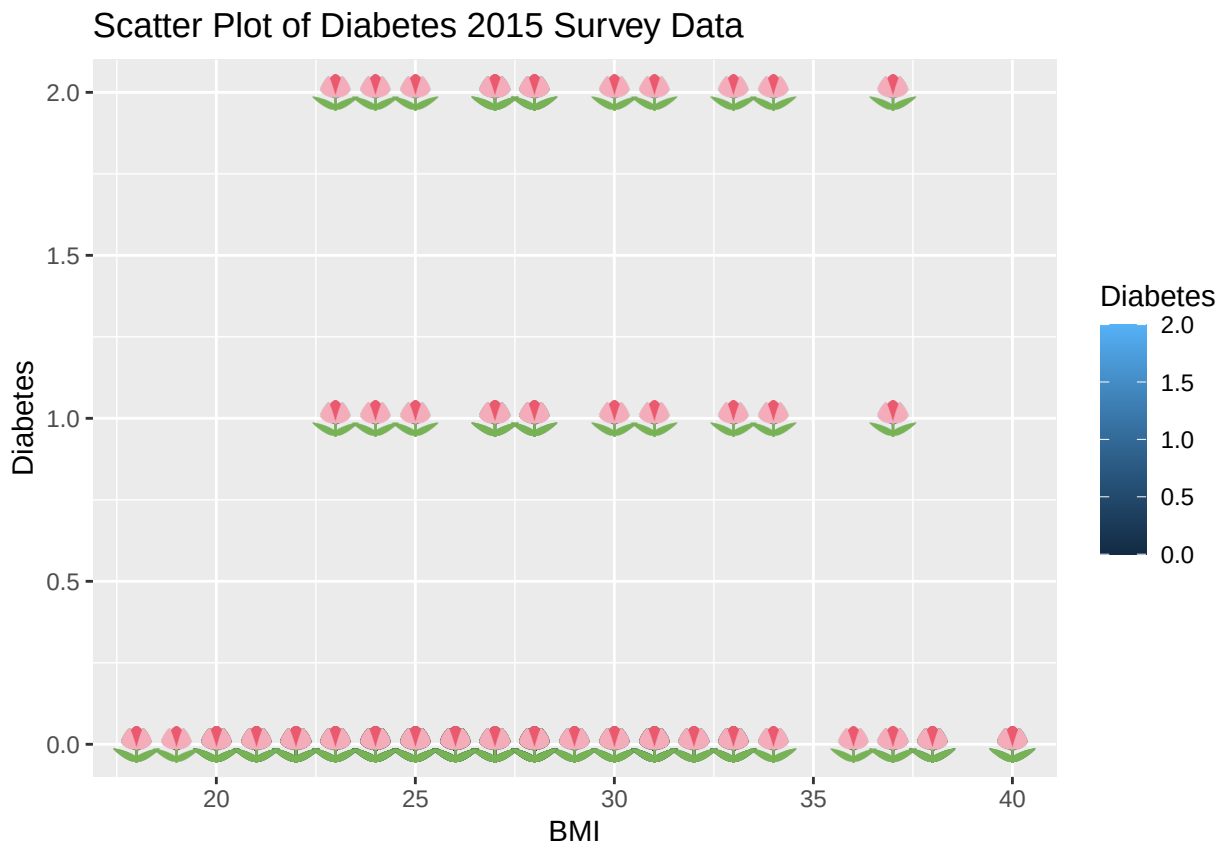
What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

Scatter plots, Histograms and boxplots are used to visualize the data.

```
ggplot(diabetes_df,aes(BMI, fill=..x..)) + geom_histogram(bins = 15) +
labs(
    title = "Histogram of BMI",
    x = "BMI",
    y = "count"
  ) + scale_fill_gradient("Legend",low = "lightgreen", high = "lightblue")
```




```r
library(ggplot2)
library(emoGG)

ggplot(diabetes_df, aes(BMI, Diabetes, color=Diabetes)) + geom_point() +
labs(
    title = "Scatter Plot of Diabetes 2015 Survey Data",
    x = "BMI",
    y = "Diabetes"
  ) +
  geom_emoji(emoji="1f337")    #added tulip emoji for BMI data points
```



What do you not know how to do right now that you need to learn to answer your questions?

I do not know how to use heatmaps to analyze the different variables at once to answer my questions. I do not know which extra packages I need to install.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

As my data has all continuous variables, I plan to use Linear Reggession Machine learning Technique to predict the probabilitiy of Diabetes.

```
diabetes_df <- diabetes_df  %>% select(Diabetes, BMI, Smoker, HighBP, HighChol, HeartDiseaseorAttack,
                                       PhysActivity, HvyAlcoholConsump, Age, Sex)
diabetis_model <- lm(Diabetes~., data=diabetes_df)
diabetis_model <- glm(Diabetes~., data=diabetes_df , family=gaussian)
diabetis_model
```

```
##
## Call:  glm(formula = Diabetes ~ ., family = gaussian, data = diabetes_df)
##
## Coefficients:
```

```
##          (Intercept)                  BMI               Smoker
##            -0.257227             0.006064             0.098309
##               HighBP             HighChol HeartDiseaseorAttack
##             0.074120            -0.080141             0.332942
##         PhysActivity    HvyAlcoholConsump                  Age
##            -0.187144            -0.060863             0.046830
##                  Sex
##            -0.075666
##
## Degrees of Freedom: 147 Total (i.e. Null);  138 Residual
## Null Deviance:       51.24
## Residual Deviance: 42.17     AIC: 256.2
```

```
summary(diabetis_model)
```

```
##
## Call:
## glm(formula = Diabetes ~ ., family = gaussian, data = diabetes_df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.76505  -0.29081  -0.13710   0.06049   1.95445
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.257227   0.305567  -0.842   0.4014
## BMI                    0.006064   0.009908   0.612   0.5415
## Smoker                 0.098309   0.100311   0.980   0.3288
## HighBP                 0.074120   0.118566   0.625   0.5329
## HighChol              -0.080141   0.108205  -0.741   0.4602
## HeartDiseaseorAttack   0.332942   0.185590   1.794   0.0750 .
## PhysActivity          -0.187144   0.102199  -1.831   0.0692 .
## HvyAlcoholConsump     -0.060863   0.264627  -0.230   0.8184
## Age                    0.046830   0.018606   2.517   0.0130 *
## Sex                   -0.075666   0.097672  -0.775   0.4398
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.3055639)
##
##     Null deviance: 51.243  on 147  degrees of freedom
## Residual deviance: 42.168  on 138  degrees of freedom
## AIC: 256.18
##
## Number of Fisher Scoring iterations: 2
```

Creating predictions using `predict()`

```
predicted_df <- data.frame(Diabetes = predict(diabetis_model, diabetes_df),
                           BMI = diabetes_df$BMI, Smoker = diabetes_df$Smoker, HighBP = diabetes_df$High
                           HeartDiseaseorAttack = diabetes_df$HeartDiseaseorAttack, PhysActivity = diabe
```

```
                          HvyAlcoholConsump = diabetes_df$HvyAlcoholConsump, Age = diabetes_df$Age, Se
head(predicted_df)
```

```
##    Diabetes BMI Smoker HighBP HighChol HeartDiseaseorAttack PhysActivity
## 1 0.4990803  40      1      1        1                    0            0
## 2 0.1333411  25      1      0        0                    0            1
## 3 0.3280059  28      0      1        1                    0            0
## 4 0.3085987  27      0      1        0                    0            1
## 5 0.2102664  24      0      1        1                    0            1
## 6 0.1921435  25      1      1        1                    0            1
##   HvyAlcoholConsump Age Sex
## 1                 0   9   0
## 2                 0   7   0
## 3                 0   9   0
## 4                 0  11   0
## 5                 0  11   0
## 6                 0  10   1
```

Some additional questions you may want to consider asking yourself as you work through this section of the project:

What features could you filter on?

Filtered on BMI to remove the outliers from the dataframe.

How could arranging your data in different ways help?

Arranging the data for example in descending order of BMI can understand more about the data

```
diabetes_df  %>% arrange(desc(BMI))
```

```
## # A tibble: 148 x 10
##    Diabetes   BMI Smoker HighBP HighChol HeartDise~1 PhysA~2 HvyAl~3   Age   Sex
##       <dbl> <dbl>  <dbl>  <dbl>    <dbl>       <dbl>   <dbl>   <dbl> <dbl> <dbl>
##  1        0    40      1      1        1           0       0       0     9     0
##  2        0    40      1      0        1           0       1       0     7     0
##  3        0    40      1      1        1           0       0       0     9     0
##  4        0    38      1      1        1           0       0       0    13     0
##  5        0    38      0      0        0           0       1       0     6     0
##  6        0    38      1      1        0           0       1       0     4     1
##  7        0    38      1      1        1           0       0       0    13     0
##  8        2    37      1      1        1           1       0       0    10     1
##  9        0    37      0      1        1           0       1       0    10     1
## 10        1    37      1      1        1           1       0       0    10     1
## # ... with 138 more rows, and abbreviated variable names
## #   1: HeartDiseaseorAttack, 2: PhysActivity, 3: HvyAlcoholConsump
```

Can you reduce your data by selecting only certain variables?

Yes,we can reduce the data by selecting the variables which are relavant for our analysis.

```
diabetes_df <- diabetes_df  %>% select(Diabetes, BMI, Smoker, HighBP, HighChol, HeartDiseaseorAttack,
                                       PhysActivity, HvyAlcoholConsump, Age, Sex) %>% arrange(desc(BMI))

head(diabetes_df)
```

```
## # A tibble: 6 x 10
##   Diabetes   BMI Smoker HighBP HighChol HeartDisea~1 PhysA~2 HvyAl~3   Age   Sex
##      <dbl> <dbl>  <dbl>  <dbl>    <dbl>        <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1        0    40      1      1        1            0       0       0     9     0
## 2        0    40      1      0        1            0       1       0     7     0
## 3        0    40      1      1        1            0       0       0     9     0
## 4        0    38      1      1        1            0       0       0    13     0
## 5        0    38      0      0        0            0       1       0     6     0
## 6        0    38      1      1        0            0       1       0     4     1
## # ... with abbreviated variable names 1: HeartDiseaseorAttack, 2: PhysActivity,
## #   3: HvyAlcoholConsump
```

Could creating new variables add new insights?

Yes, creating new variables can add new insights, for example creating avgBMI variable from BMI could let us understand the data even more

```
diabetes_df %>% group_by(Diabetes) %>% summarize(AvgBMI = mean(`BMI`))
```

```
## # A tibble: 3 x 2
##   Diabetes AvgBMI
##      <dbl>  <dbl>
## 1        0   27.3
## 2        1   28.9
## 3        2   28.9
```

Could summary statistics at different categorical levels tell you more?

Yes, summary statistics at different categorical levels can tell us more. For example, if we analize with Diabetes as categorical variable with 0 = no diabetes 1 = prediabetes 2 = diabetes

based on the above analysis, the BMI should be maintained around 27.2 in order to reduce the chances of Diabetes.

How can you incorporate the pipe (%>%) operator to make your code more efficient?

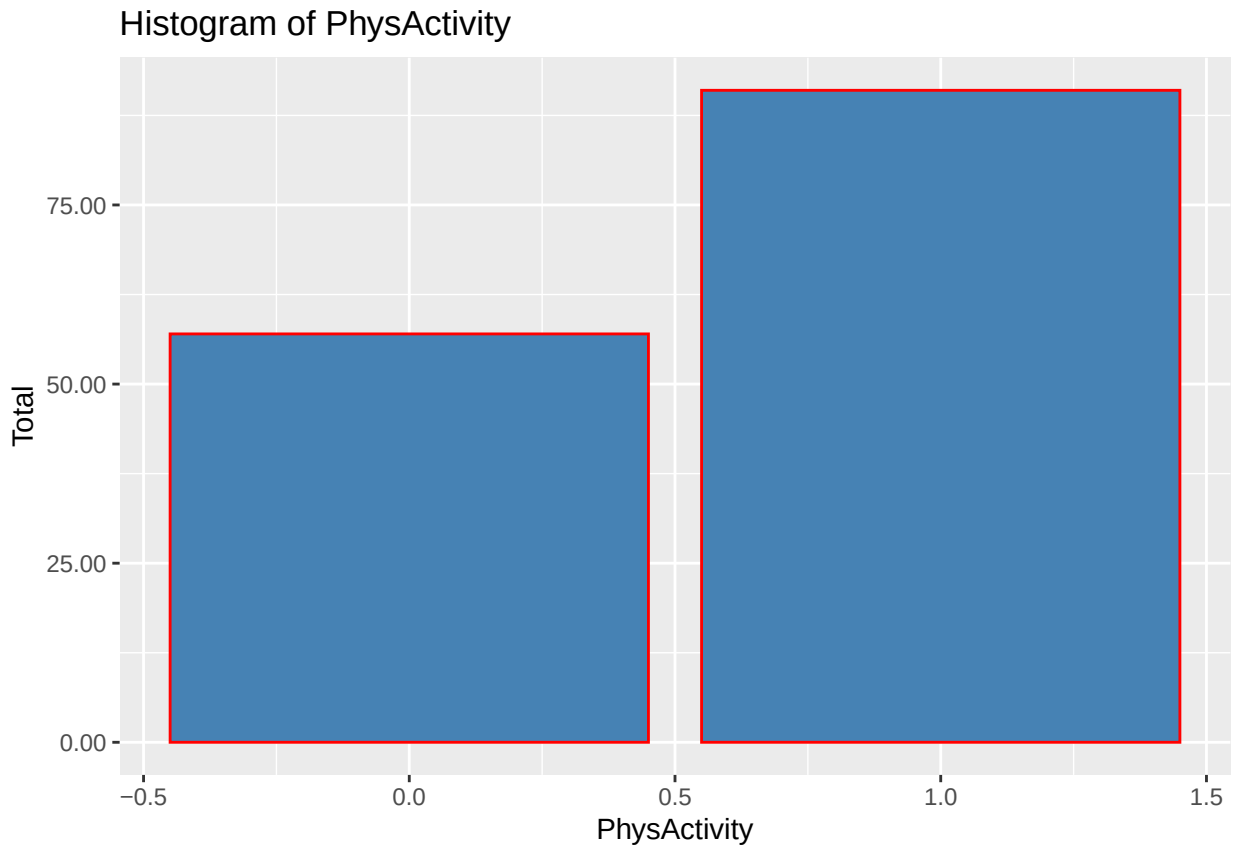
Pipe operator can help in many ways, for example select some variables to reduce the data for better analysis. or it can be used for group_by() and summerize fuctions as in the above examples.

Analysis on predictors: physical activity, smoking, BMI, HighBP, Sex, Age, HeartDiseaseorAttack, HighChol, HvyAlcoholConsump

#Histogram

```
library(dplyr)
library(tidyr)
library(DT)
library(formattable)
PhysActivity_data <- diabetes_df %>%
           group_by(PhysActivity) %>%
               dplyr::summarize(Total = n())
datatable(PhysActivity_data)
```

```
ggplot(PhysActivity_data, aes(PhysActivity, Total)) +
        geom_bar( stat = "identity", fill = "steelblue", color = "red") +
           ggtitle("Histogram of PhysActivity") +
            theme(legend.position = "none") +
            scale_y_continuous(labels = comma)
```



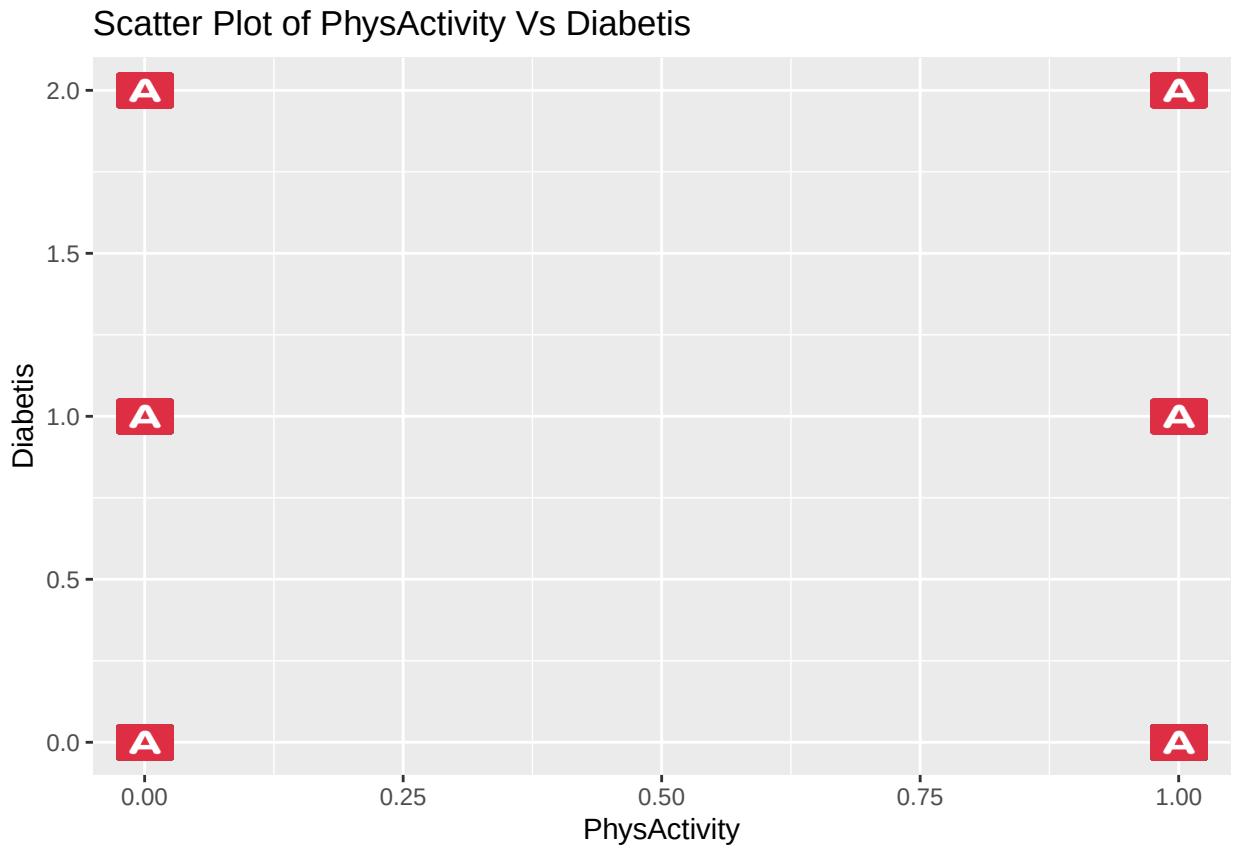
In this sample data set that we took, the total number of people who do Physical activity in the past 30 days is more than the total number of people who do not do any physical activity.

```
diabetes_df %>% group_by(Diabetes) %>% summarize(AvgPhysActivity = mean(`PhysActivity`))
```

```
## # A tibble: 3 x 2
##   Diabetes AvgPhysActivity
##      <dbl>           <dbl>
## 1        0           0.669
## 2        1           0.333
## 3        2           0.333
```

Scatter Plot

```
ggplot(diabetes_df, aes(PhysActivity, Diabetes)) + geom_point() +
labs(
    title = "Scatter Plot of PhysActivity Vs Diabetis",
    x = "PhysActivity",
    y = "Diabetis"
  ) + geom_emoji(emoji="1f170")    #added  emoji       for physical activity
```




```
cor.test(diabetes_df$PhysActivity, diabetes_df$Diabetes)
```

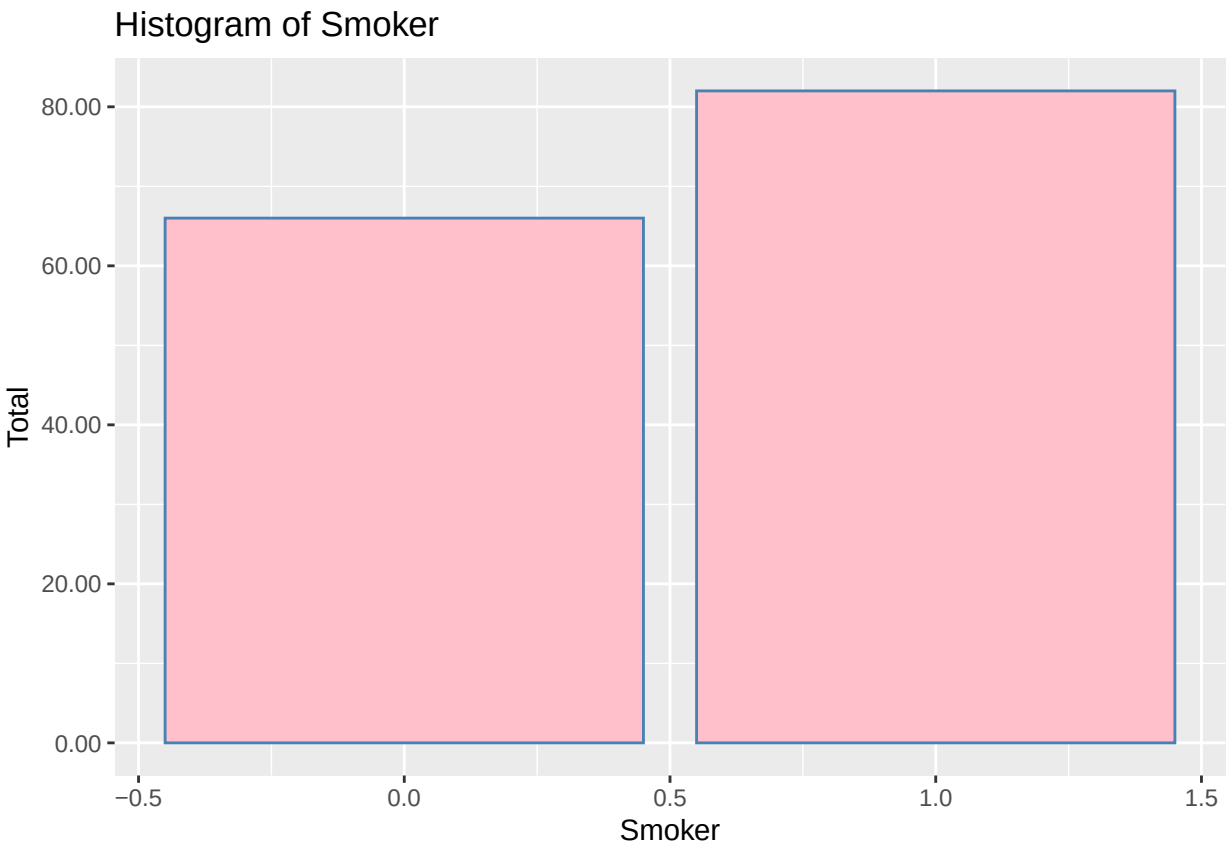
```
##
##  Pearson's product-moment correlation
##
## data:  diabetes_df$PhysActivity and diabetes_df$Diabetes
## t = -2.9761, df = 146, p-value = 0.003418
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.38562151 -0.08093671
## sample estimates:
##        cor
## -0.2391575
```

since T-value is negative, Diabetis is inversely proportional to Physical activity. So, as the physical activity increases, the chance of getting diabetis reduces.

Histogram of smoker

```
Smoker_data <- diabetes_df %>%
            group_by(Smoker) %>%
                dplyr::summarize(Total = n())
datatable(Smoker_data)
```

```
ggplot(Smoker_data, aes(Smoker, Total)) +
        geom_bar( stat = "identity", fill = "pink", color = "steelblue") +
           ggtitle("Histogram of Smoker") +
            theme(legend.position = "none") +
            scale_y_continuous(labels = comma)
```



In this sample data set that we took, the total number of people who smoked at least 100 cigarettes in your entire life is little more than the total number of people who do not smoke.

```
diabetes_df %>% group_by(Diabetes) %>% summarize(AvgSmoker = mean(`Smoker`))
```

```
## # A tibble: 3 x 2
##   Diabetes AvgSmoker
##      <dbl>     <dbl>
## 1        0     0.516
## 2        1     0.75
## 3        2     0.75
```

Smoker correlation

```
cor.test(diabetes_df$Smoker, diabetes_df$Diabetes)
```

```
##
```

```
##  Pearson's product-moment correlation
##
## data:  diabetes_df$Smoker and diabetes_df$Diabetes
## t = 1.9957, df = 146, p-value = 0.04783
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.001656494 0.315992526
## sample estimates:
##       cor
## 0.1629568
```

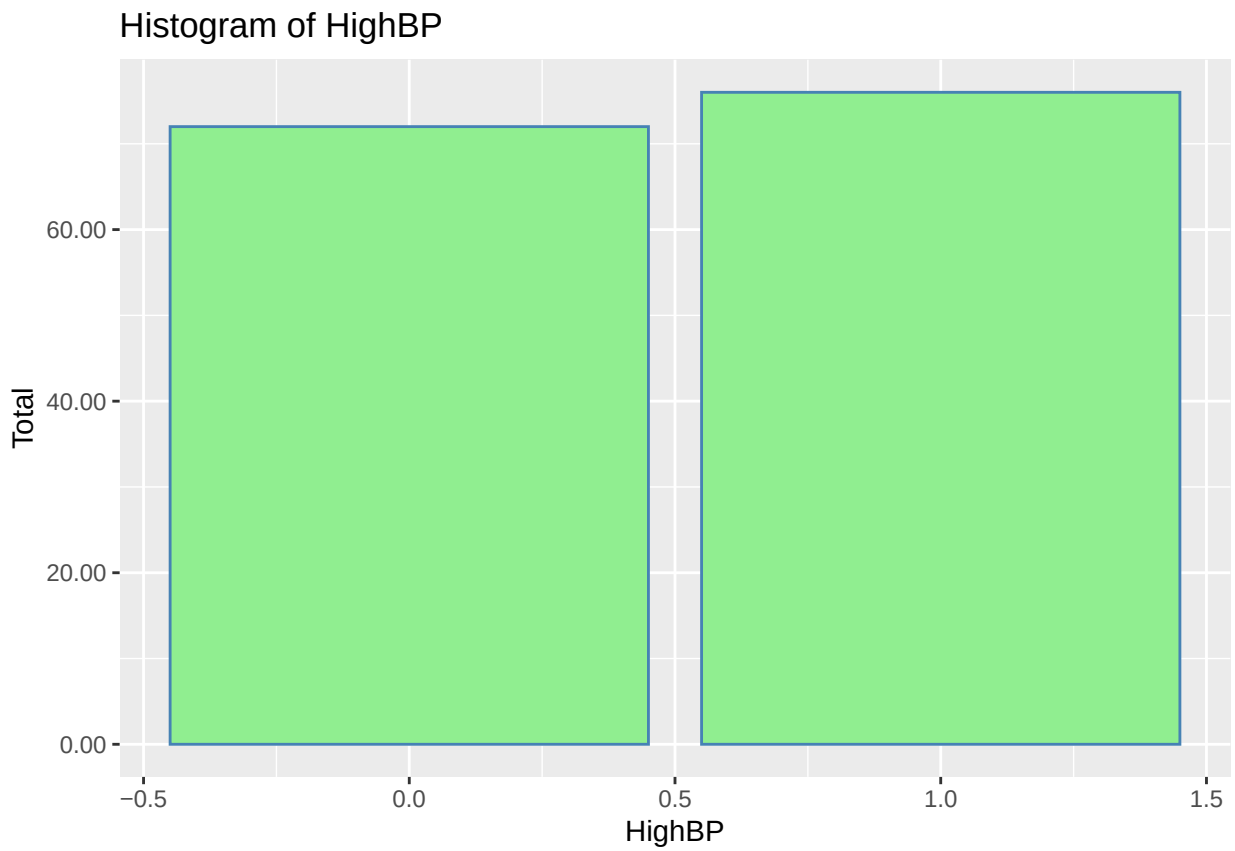
Since the T-value is positive, Diabetis is directly proportioanl to the Smoker. That means the smoker will have more chances of getting Diabetis.

DataTable of HighBP

```
HighBP_data <- diabetes_df %>%
           group_by(HighBP) %>%
               dplyr::summarize(Total = n())
datatable(HighBP_data)
```

Histogram of HighBP

```
ggplot(HighBP_data, aes(HighBP, Total)) +
        geom_bar( stat = "identity", fill = "lightgreen", color = "steelblue") +
           ggtitle("Histogram of HighBP") +
            theme(legend.position = "none") +
            scale_y_continuous(labels = comma)
```



In this sample data set that we took, the total number of people who had high BP is nearly equal to the number of people wh did not have HighBP.

HighBP correlation

```
cor.test(diabetes_df$HighBP, diabetes_df$Diabetes)
```

```
##
##  Pearson's product-moment correlation
##
## data:  diabetes_df$HighBP and diabetes_df$Diabetes
## t = 3.3142, df = 146, p-value = 0.001159
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1077687 0.4084264
## sample estimates:
##       cor
## 0.2645132
```

Since the T-value is positive, Diabetis is directly proportioanl to the HighBP.

That means the person with High Blood Pressure will have more chances of getting Diabetis.

DataTable of High Cholesterol

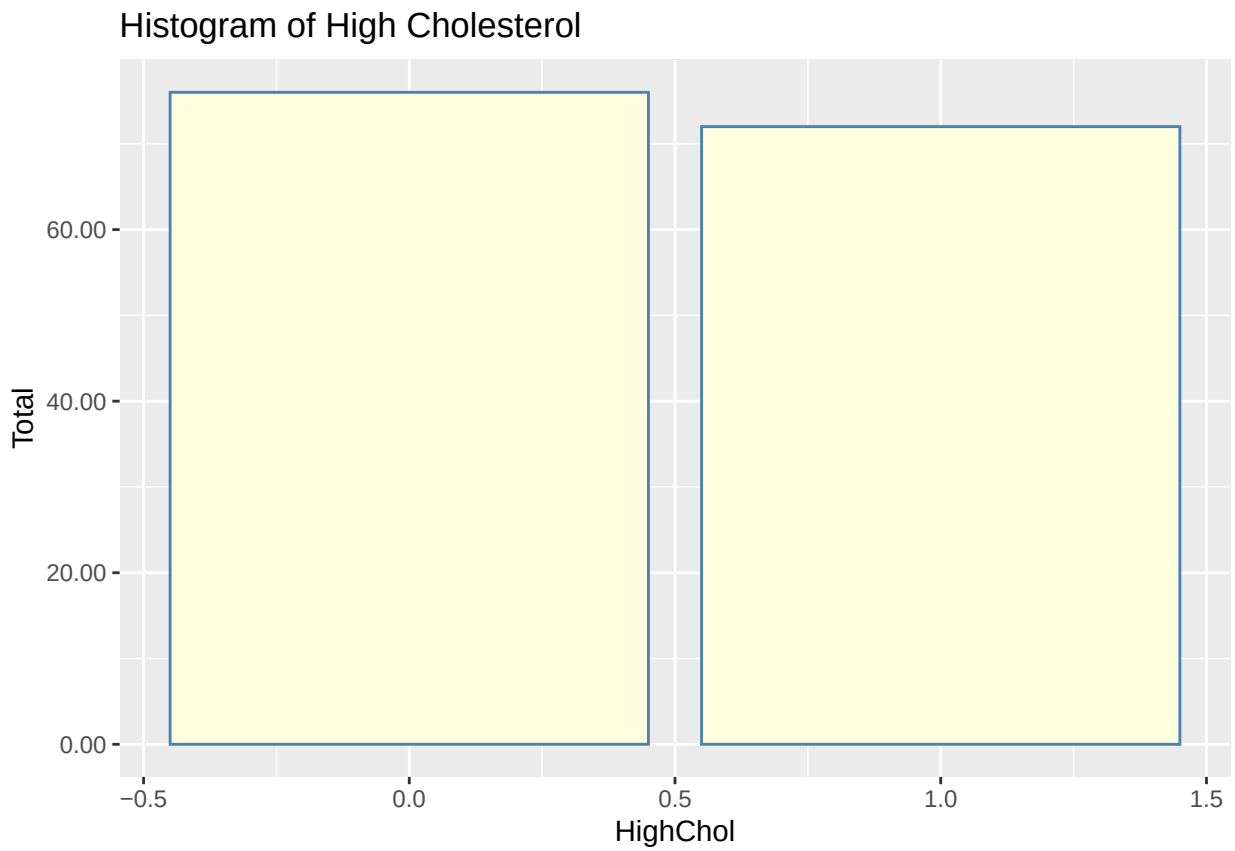
```
HighChol_data <- diabetes_df %>%
           group_by(HighChol) %>%
               dplyr::summarize(Total = n())
datatable(HighChol_data)
```

In this sample data set that we took, the total number of people who had

high Cholesterol is nearly equal to the number of people wh did not have high Cholesterol.

Histogram of High Cholesterol

```
ggplot(HighChol_data, aes(HighChol, Total)) +
        geom_bar( stat = "identity", fill = "lightyellow", color = "steelblue") +
           ggtitle("Histogram of High Cholesterol") +
            theme(legend.position = "none") +
            scale_y_continuous(labels = comma)
```



High Cholesterol correlation with Diabetes

```
cor.test(diabetes_df$HighChol, diabetes_df$Diabetes)
```

```
##
##  Pearson's product-moment correlation
##
## data:  diabetes_df$HighChol and diabetes_df$Diabetes
## t = 1.821, df = 146, p-value = 0.07066
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01262588  0.30307825
## sample estimates:
##       cor
## 0.1490215
```

Since the T-value is positive, Diabetis is directly proportioanl to the high Cholesterol That means the people with high Cholesterol will have more chances of getting Diabetis.

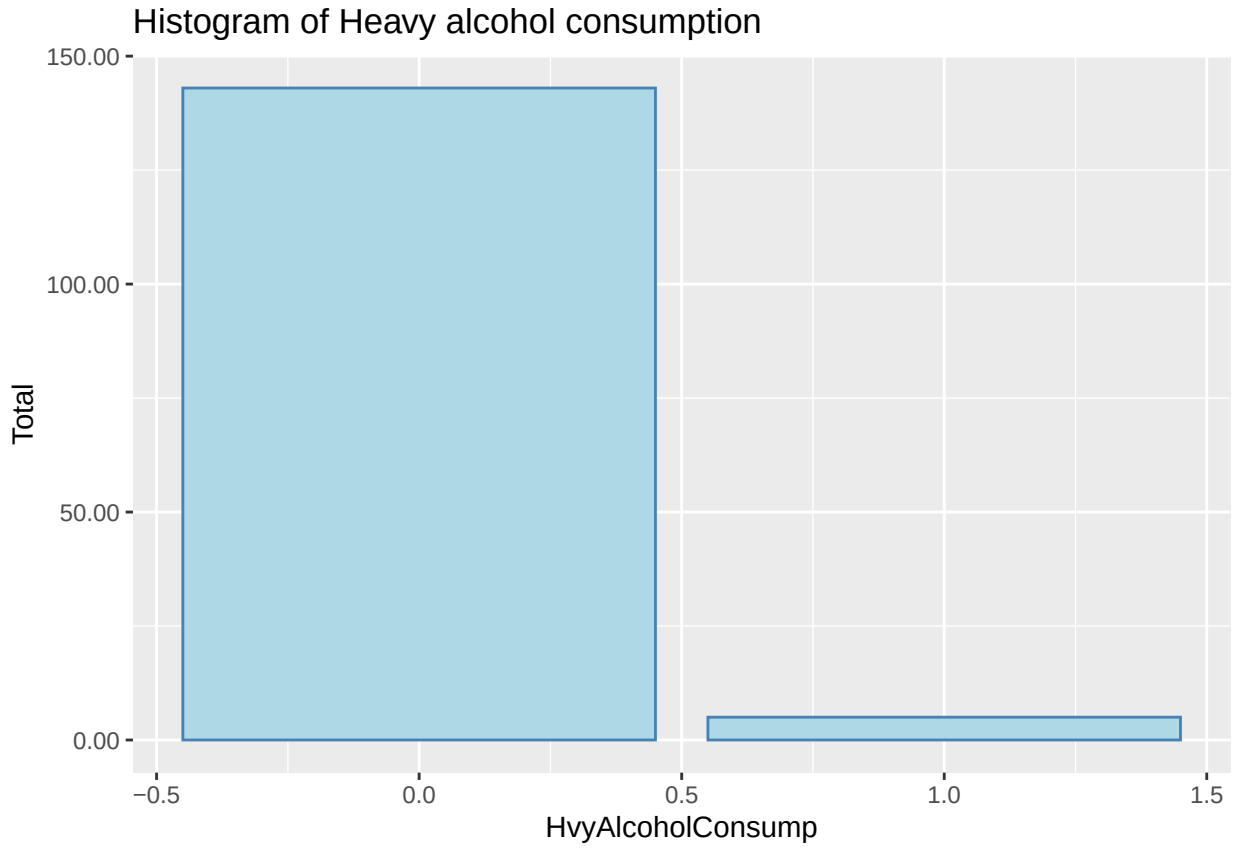
DataTable of Heavy alcohol consumption

```
HvyAlcoholConsump_data <- diabetes_df %>%
          group_by(HvyAlcoholConsump) %>%
              dplyr::summarize(Total = n())
datatable(HvyAlcoholConsump_data)
```

only a little amount of data is available from sample taken. It is better to analyze by taking more data.. For now we try to analyze with whatever data we have.

Histogram of Heavy alcohol consumption

```
ggplot(HvyAlcoholConsump_data, aes(HvyAlcoholConsump, Total)) +
       geom_bar( stat = "identity", fill = "lightblue", color = "steelblue") +
          ggtitle("Histogram of Heavy alcohol consumption") +
           theme(legend.position = "none") +
           scale_y_continuous(labels = comma)
```




```
cor.test(diabetes_df$HvyAlcoholConsump, diabetes_df$Diabetes)
```

```
##
##  Pearson's product-moment correlation
##
```

```
## data:  diabetes_df$HvyAlcoholConsump and diabetes_df$Diabetes
## t = -0.9368, df = 146, p-value = 0.3504
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.23570262  0.08510689
## sample estimates:
##         cor
## -0.07729837
```

Since p value is greater than 0.05, that the correlation between these two variables is statistically significant.

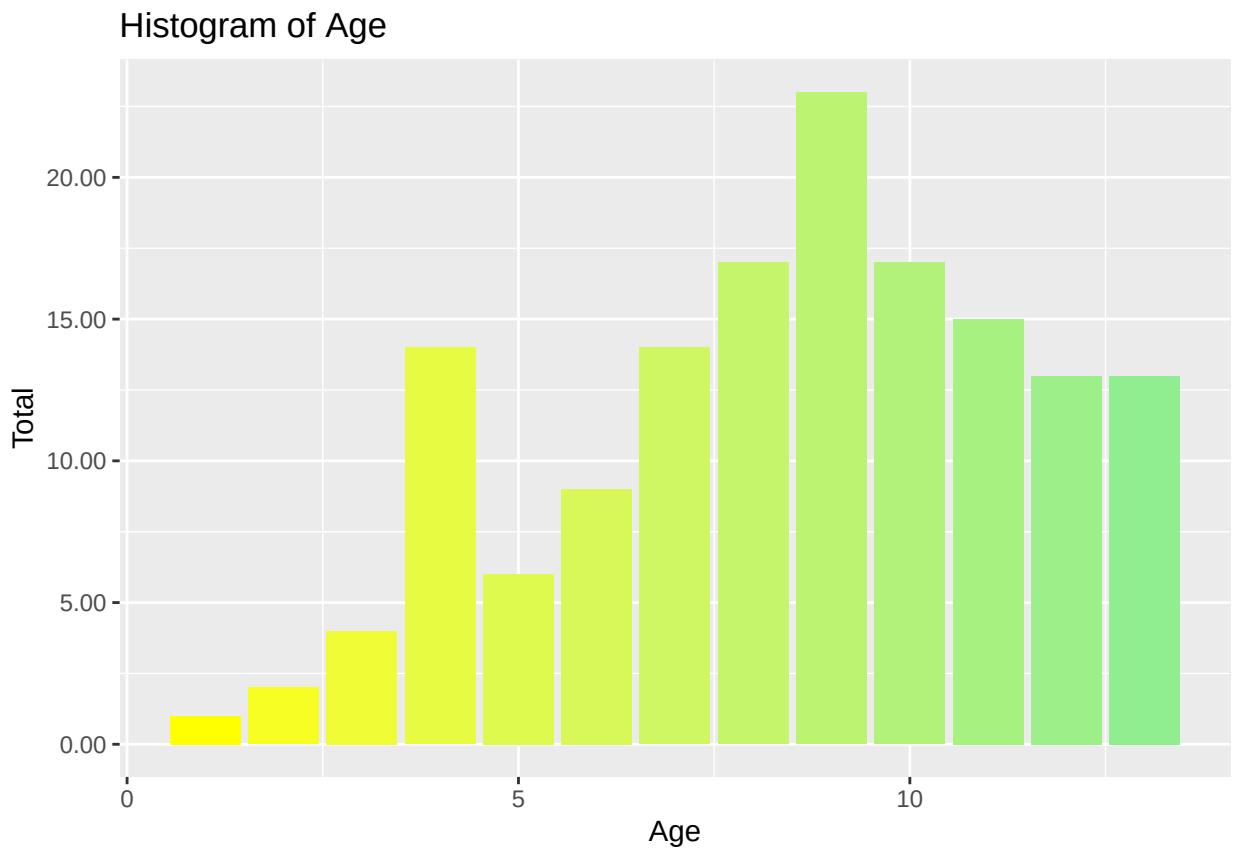
Also with the little amount of data that we have here, we cannot not analyze it correctly.

Age

DataTable of Age

```
Age_data <- diabetes_df %>%
           group_by(Age) %>%
               dplyr::summarize(Total = n())
datatable(Age_data)
```

```
ggplot(Age_data, aes(Age, Total, fill=..x..)) +
        geom_bar( stat = "identity" ) +
           ggtitle("Histogram of Age") +
            theme(legend.position = "none") +
            scale_y_continuous(labels = comma) +
  scale_fill_gradient("Legend",low = "yellow", high = "lightgreen")
```



13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older

Age correlation

```
cor.test(diabetes_df$Age, diabetes_df$Diabetes)
```

```
##
##  Pearson's product-moment correlation
##
## data:  diabetes_df$Age and diabetes_df$Diabetes
## t = 3.8786, df = 146, p-value = 0.0001584
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1517759 0.4450335
## sample estimates:
##       cor
## 0.3056353
```

Since the T-value is positive, Diabetis is directly proportioanl to the Age. That means as the individual gets older, chances of getting Diabetis is more. Since the p-value is less than 0.05, we can say that the correlation between these two variables is statistically significant.

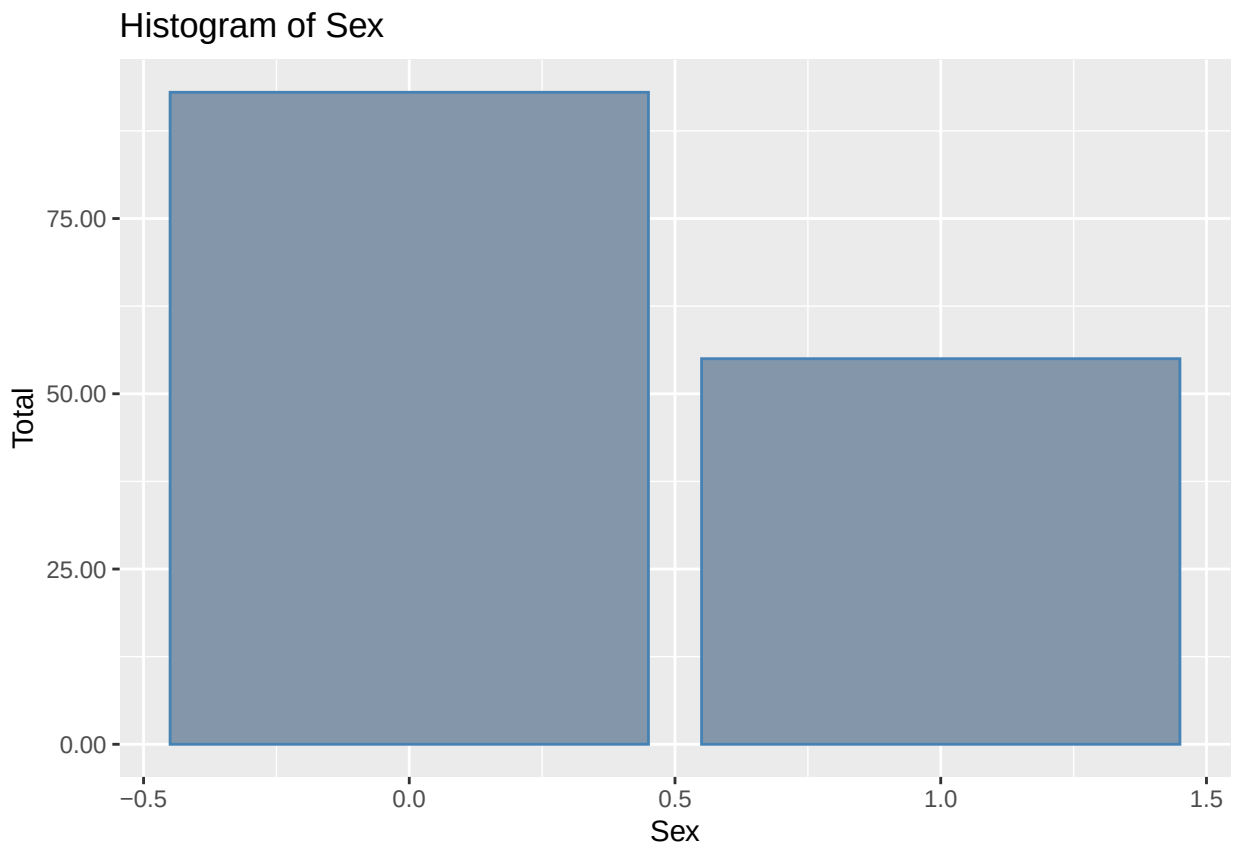
DataTable of Sex

```
Sex_data <- diabetes_df %>%
           group_by(Sex) %>%
               dplyr::summarize(Total = n())
datatable(Sex_data)
```

Here 0 = female 1 = male. Data related to Female is more than the males.

Histogram of Sex

```
ggplot(Sex_data, aes(Sex, Total)) +
        geom_bar( stat = "identity", fill = "#8496a9", color = "steelblue") +
           ggtitle("Histogram of Sex") +
            theme(legend.position = "none") +
            scale_y_continuous(labels = comma)
```



sex correlation

```
cor.test(diabetes_df$Sex, diabetes_df$Diabetes)
```

```
##
##  Pearson's product-moment correlation
##
## data:  diabetes_df$Sex and diabetes_df$Diabetes
## t = -0.39597, df = 146, p-value = 0.6927
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1930770  0.1292735
## sample estimates:
##         cor
## -0.03275354
```

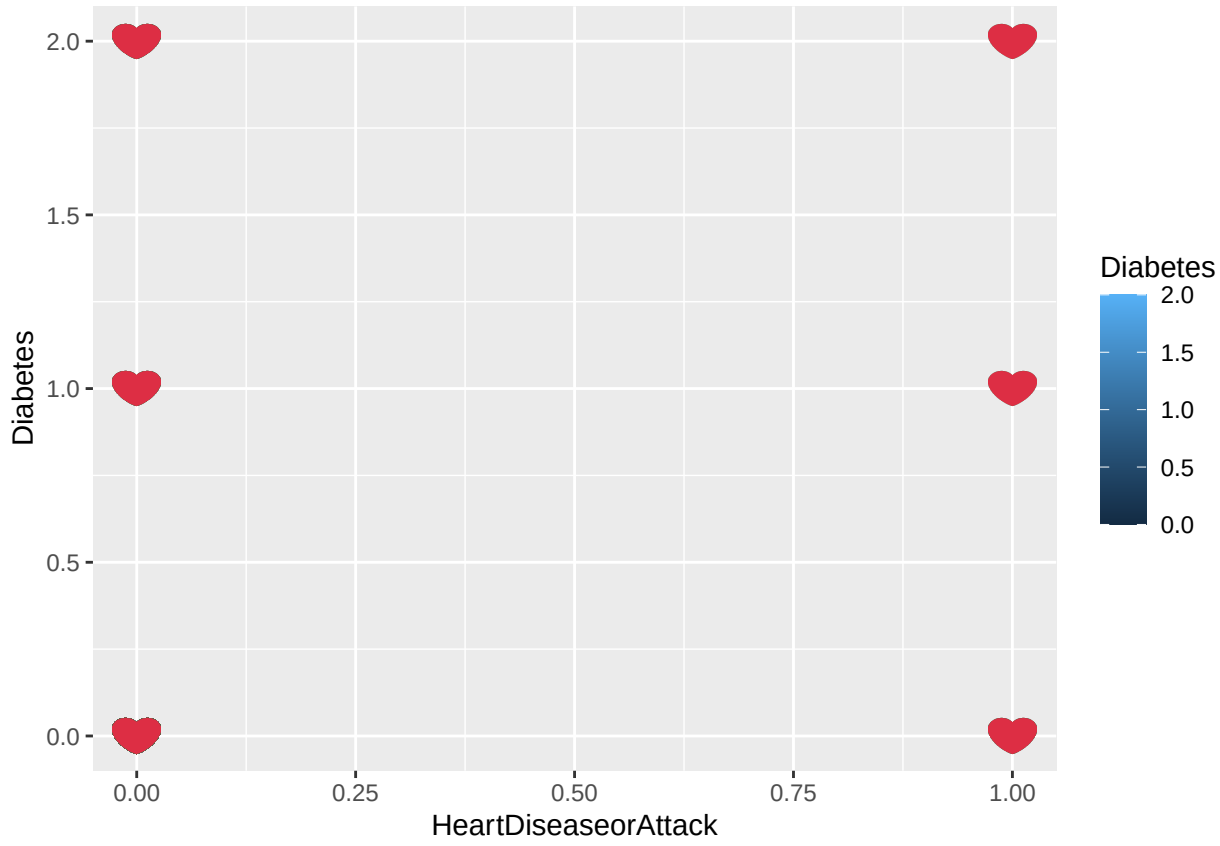
Since the p-value is greater than 0.05, we can say that the correlation between these two variables is not statistically significant.

HeartDiseaseorAttack

```r
diabetes_df %>% group_by(Diabetes) %>% summarize(AvgHeartDiseaseorAttack = mean(`HeartDiseaseorAttack`))
```

```
## # A tibble: 3 x 2
##   Diabetes AvgHeartDiseaseorAttack
##      <dbl>                   <dbl>
## 1        0                  0.0484
## 2        1                  0.25
## 3        2                  0.25
```

```r
library("ggplot2")
# convert Diabetes column from numeric to factor with three levels
# diabetes_df$Diabetes <- cut(diabetes_df$Diabetes, 3, labels=c('Non-Diabetec', 'Pre-Diabetec', 'Diabet
ggplot(diabetes_df, aes(x = HeartDiseaseorAttack, y=Diabetes, color=Diabetes)) +                    # Scat
  geom_point() +geom_emoji(emoji="2764")
```



HeartDiseaseorAttack correlation

```r
cor.test(diabetes_df$HeartDiseaseorAttack, diabetes_df$Diabetes)
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  diabetes_df$HeartDiseaseorAttack and diabetes_df$Diabetes
## t = 3.1975, df = 146, p-value = 0.0017
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.09854278 0.40062728
## sample estimates:
##       cor
## 0.2558193
```

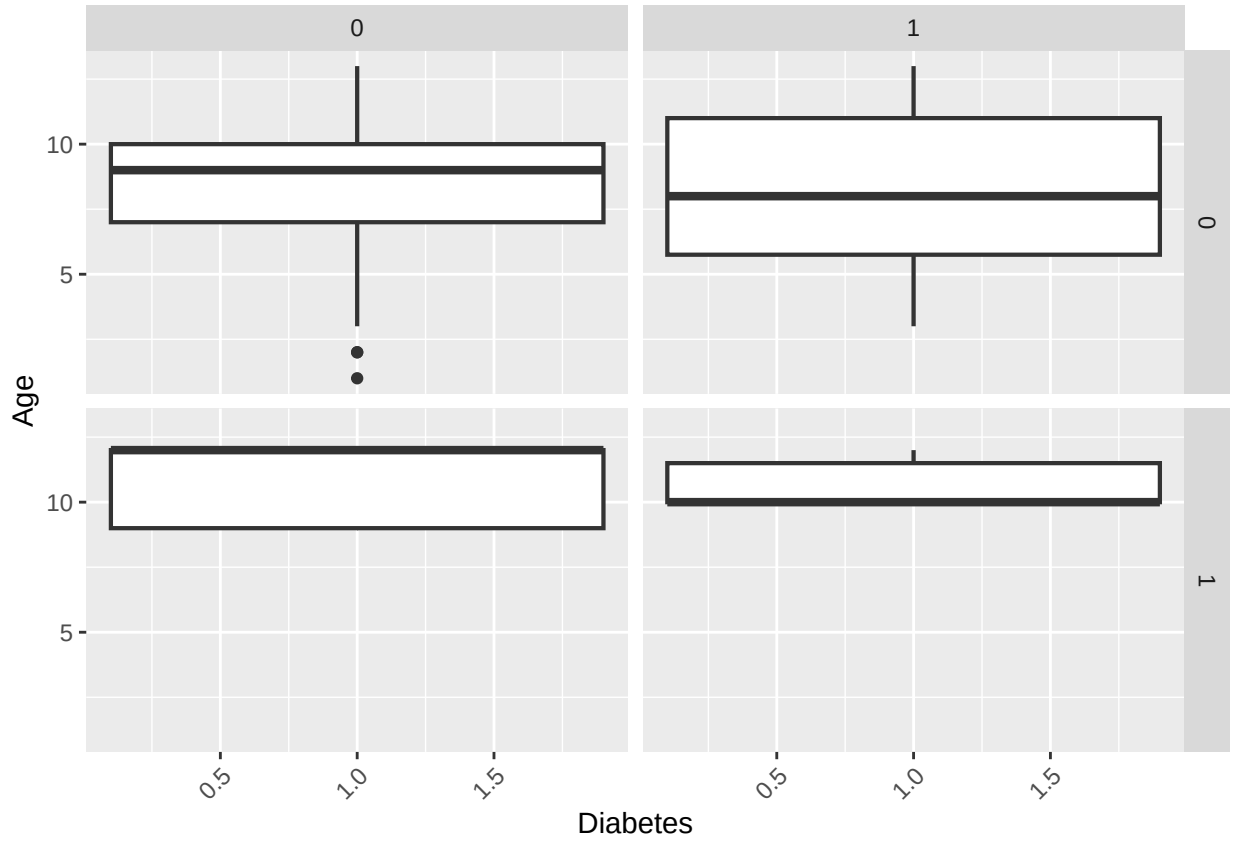
Here the T-value is positive, So, Diabetis is directly proportioanl to the HeartDiseaseorAttack. That means the individual with HeartDiseaseorAttack has more chances of getting diabetes. Since the p-value is less than 0.05, we can say that the correlation between these two variables is statistically significant.

We can also analyze the distribution of Diabetes across Age, Sex and HeartDiseaseorAttack

```
ggplot(diabetes_df, aes(x = Diabetes, y = Age, fill = Diabetes)) +
  geom_boxplot(size = .75)  +
  facet_grid(HeartDiseaseorAttack ~ Sex, margins = FALSE)  +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))
```

```
## Warning: Continuous x aesthetic
## i did you forget `aes(group = ...)`?
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

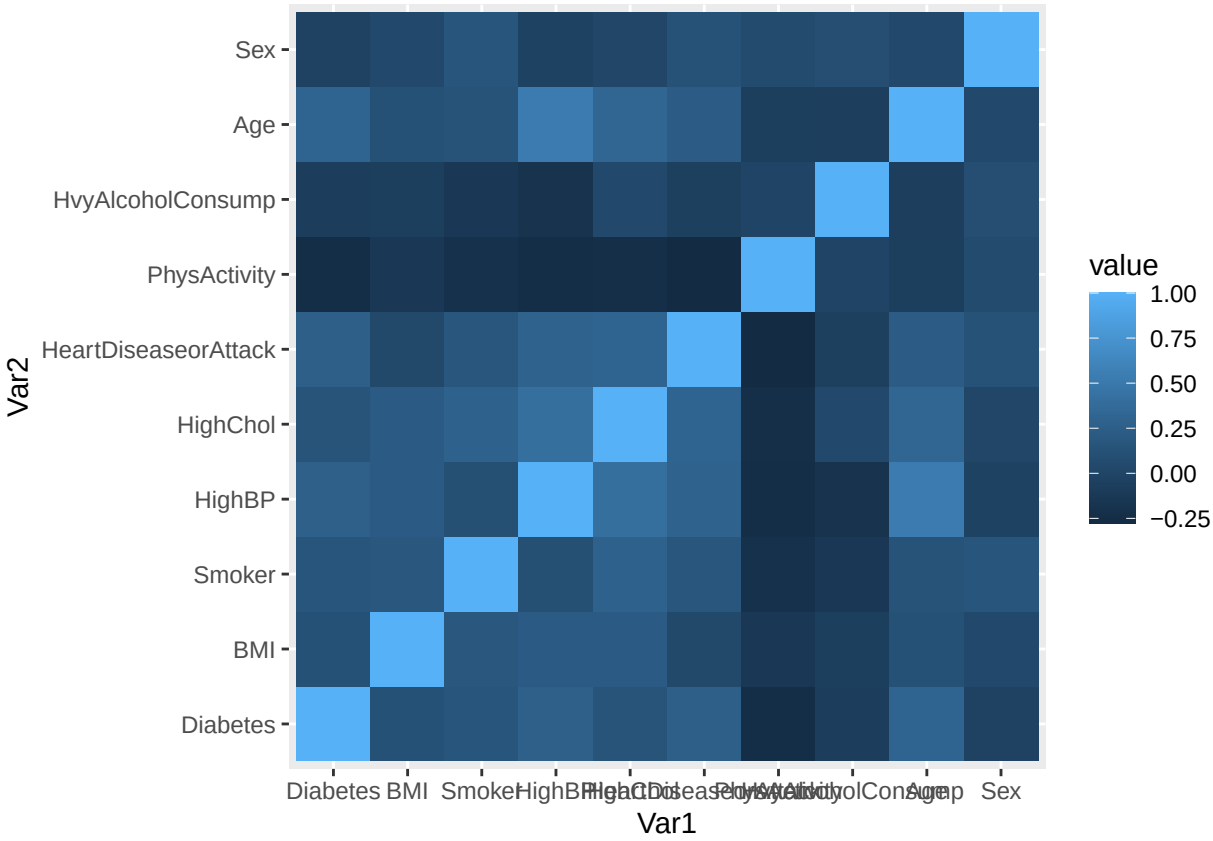


#create a heatmap to analyze various other factors of Diabetis

```
library(ggplot2)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
data <- cor(diabetes_df[sapply(diabetes_df,is.numeric)])

data1 <- melt(data)

ggplot(data1, aes(x = Var1,
                  y = Var2,
                  fill = value)) + geom_tile()
```


Sex
Age
HvyAlcoholConsump
PhysActivity
HeartDiseaseorAttack
HighChol
HighBP
Smoker
BMI
Diabetes
Var2
Diabetes
BMI
Smoker
Sex
Var1
value
1.00
0.75
0.50
0.25
0.00
−0.25

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.

The problem statement I addressed.

I addressed the Diabetes Issue by analyzing the various factors such as BMI, Physiscal Activity, smoking, age, sex and HeartDiseaseorAttack

How you addressed this problem statement

I took three data sets and combined them aftering renaming the column names. Cleaned the data by removing the NA values from the dataset.

Performed some transformations to tidy up the data. Then, analyzed the data and visualize it in the form of different graphs and

charts to figure out the important factors that are affecting diabetes. plot the graphs with Diabetes on Y axis and different factors on X-axis and analyzed them.

Finally provided useful analysis for people who can change their lifestyle to reduce Diabetes cases.

Analysis.

The analysis of various factors on Diabetes with the sample data that I took is provided below:

- The BMI should be maintained around 27.2 in order to reduce the chances of Diabetes.
- Diabetes is directly proportional to the Smoker. That means the smoker will have more chances of getting Diabetes.
- The individual with Heart Disease or Attack has more chances of getting diabetes.
- As the physical activity of an individual increases, the chance of getting diabetes reduces.
- Diabetes is directly proportional to the High Blood Pressure. That means the person with High Blood Pressure will have more chances of getting Diabetes.
- As the individual gets older, chances of getting Diabetes is more.
- Diabetes is directly proportional to the high Cholesterol That means the people with high Cholesterol will have more chances of getting Diabetes.
- Regarding Heavy Alcohol Consumption, with the little amount of data that we have taken as the sample data, we could not able to analyze it correctly.
- Both Male and Females are equally prone to Diabetes.

Implications.

It Implies that all the individuals need to take precautions to not get effected by Diabetes

as there are certain factors which directly or indirectly causes Diabetes.

For example people with high Blood Pressure, High Cholesterol or smoking are prone to diabetes,

which can effect their health in many ways such as nervous system damage, eye or foot damage, Osteoporosis, or demensia,

It is better to understand the factors that effect Diabetes and correct their lifestyle to lead a healthy life.

Limitations.

Could not completely run the model on the very large dataset provided in kaggle.

Could not analyze all the factors affecting diabetes.

Here only the Logistic Regression Model is used. This model can be improved by training and testing with various other models such as Ordinal Regression Model and also with different sets of data related to Diabetes.

Also, Various other models can be tried and tested for better results.

Concluding Remarks

This project helps in Understanding the factors that effects Diabetes so that the individual can correct his/her lifestyle to lead a better and healthy life.

An individual with High BP/cholesterol can lower blood Pressure/cholesterol with certain medications and diet restrictions.

An individual can improve the Physical activity.

Smoking can be avoided.

A person with previous Heart Dicease or Attack has to take more precautions, going to a doctor as much as needed.

As the individual gets older more precaution has to be taken such as annual physical doctor appointments and care can be provided for various factors effecting Diabetes.

References

Datasets from Kaggle website