# Exploratory Analysis on Diabetes Risk Factors

Bernard

8/20/2022

## Contents

## 0.1 INTRODUCTION

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. Hyperglycaemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels. The most common is type 2 diabetes, usually in adults, which occurs when the body becomes resistant to insulin or doesn't make enough insulin. In the past three decades the prevalence of type 2 diabetes has risen dramatically in countries of all income levels. Type 1 diabetes, once known as juvenile diabetes or insulin-dependent diabetes, is a chronic condition in which the pancreas produces little or no insulin by itself. For people living with diabetes, access to affordable treatment, including insulin, is critical to their survival. There is a globally agreed target to halt the rise in diabetes and obesity by 2025.

### 0.1.1 Epidemiology:

The number of people with diabetes rose from 108 million in 1980 to 422 million in 2014. Prevalence has been rising more rapidly in low- and middle-income countries than in high-income countries. Between 2000 and 2016, there was a 5% increase in premature mortality rates (i.e. before the age of 70) from diabetes. In high-income countries the premature mortality rate due to diabetes decreased from 2000 to 2010 but then increased in 2010-2016. In lower-middle-income countries, the premature mortality rate due to diabetes increased across both periods. In 2019, diabetes was the ninth leading cause of death with an estimated 1.5 million deaths directly caused by diabetes. A healthy diet, regular physical activity, maintaining a normal body weight and avoiding tobacco use are ways to prevent or delay the onset of type 2 diabetes. Diabetes can be treated and its consequences avoided or delayed with diet, physical activity, medication and regular screening and treatment for complications.

### 0.1.2 Health Impact:

Over time, diabetes can damage the heart, blood vessels, eyes, kidneys, and nerves.

- Adults with diabetes have a two- to three-fold increased risk of heart attacks and strokes.
- Combined with reduced blood flow, neuropathy (nerve damage) in the feet increases the chance of foot ulcers, infection and eventual need for limb amputation.
- Diabetic retinopathy is an important cause of blindness, and occurs as a result of long-term accumulated damage to the small blood vessels in the retina. Close to 1 million people are blind due to diabetes.
- Diabetes is among the leading causes of kidney failure.

Who gets diabetes? What are the risk factors? Factors that increase your risk differ depending on the type of diabetes you ultimately develop.

### 0.1.3 Risk Factors for Diabetes:

- Family history (parent or sibling).
- Race: African, African-American, Hispanic, Native American or Asian-American.
- Having overweight/obesity.
- Having high blood pressure.
- Injury to the pancreas (such as by infection, tumor, surgery or accident).
- Physical stress (such as surgery or illness).
- Having low HDL cholesterol (the “good” cholesterol) and high triglyceride level.
- Being physically inactive.
- Age: older than 25 years.
- Having gestational diabetes or giving birth to a baby weighing more than 9 pounds.
- Having polycystic ovary syndrome.
- Having a history of heart disease or stroke.
- Being a smoker.

## 0.2 ANALYSIS

Over 200,000 individuals were surveyed and asked 22 questions about their medical history, healthcare status, social status and lifestyle. The dataset would be analysed for patterns, summarised using visuals and tables, the target variable **Diabetes_binary:** presence of diabetes and 21 other feature variables.

### 0.2.1 Loading Libraries

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE,
                      message= FALSE, fig.show = TRUE, fig.align = "centre" )
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(latexpdf)
```

### 0.2.2 Loading dataset

```
df1 <- read_csv("diabetes_binary_health_indicators_BRFSS2015.csv")
```

### 0.2.3 Variables

```
colnames(df1)
```

```
##  [1] "Diabetes_binary"      "HighBP"               "HighChol"
##  [4] "CholCheck"            "BMI"                  "Smoker"
##  [7] "Stroke"               "HeartDiseaseorAttack" "PhysActivity"
## [10] "Fruits"               "Veggies"              "HvyAlcoholConsump"
## [13] "AnyHealthcare"        "NoDocbcCost"          "GenHlth"
## [16] "MentHlth"             "PhysHlth"             "DiffWalk"
## [19] "Sex"                  "Age"                  "Education"
## [22] "Income"
```

- Diabetes_binary: outcome variable: presence of diabetes 0 = no diabetes; 1 = diabetes

- HighBP: presence of High Blood pressure: 0 = no High Blood pressure; 1 = High Blood pressure

- HighChol: presence of High Cholesterol: 0 = no high Cholesterol; 1 = High Cholesterol

- CholCheck: Cholesterol check: 0 = no cholesterol check in 5 years; 1 = yes cholesterol check in 5 years.

- BMI: Body Mass Index (kg/m2)

- Smoker: has the individual smoked at least 100 cigarettes in their entire life: 0 = no, 1 = yes.

- HeartDiseasesorAttack: history of heart attack and myorcaridal infarction: 0 = no, 1 = yes.

- Stroke: history of stroke: 0 = no, 1 = yes.

- Fruits: consumes fruit 1 or more times a day: 0 = no, 1 = yes.

- Veggies: consumes vegetabls 1 or more times a day: 0 = no, 1 = yes.

- Sex: 0 = female, 1 = male.

- Age:
- PhysActivity: any physical activity in the past 30 days not including job: 0 = no, 1 = yes.

- HvyAlcoholConsump: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week): 0 = no, 1 = yes.

- AnyHealthcare: does the individual have any health care insurance, prepaid plans, 0 = no, 1 = yes.

- NoDocbcCost: was there a time the individual could not afford to see in the past 12 months: 0 = no, 1=yes.

- GenHlth: Would you say that in general your health is: scale 1-5 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor.

- MentHlth: mental health rating: Now thinking about your mental health, which includes stress, depression, and problems with emotions, how would you rate your mental health.

- PhyHlth: physical health rating: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days.

- DiffWalk: Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes.

- Age: 13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older.
- Education: Education level scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate).

- Income: Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000, 5 = less than $35,000 8 = $75,000 or more.

*How many entries are available for analysis?*

```r
dim(df1)
```

```
## [1] 253680     22
```

*Are there duplicate rows in the survey?*

```r
dim(df1[duplicated(df1), ]) # duplicated rows
```

```
## [1] 24206    22
```

```r
diabetes <- df1[!duplicated(df1), ]   # removes the duplicated rows/ entries

dim(diabetes)
```

```
## [1] 229474     22
```

```r
diabetes2 <- diabetes # duplicate the working data set
```

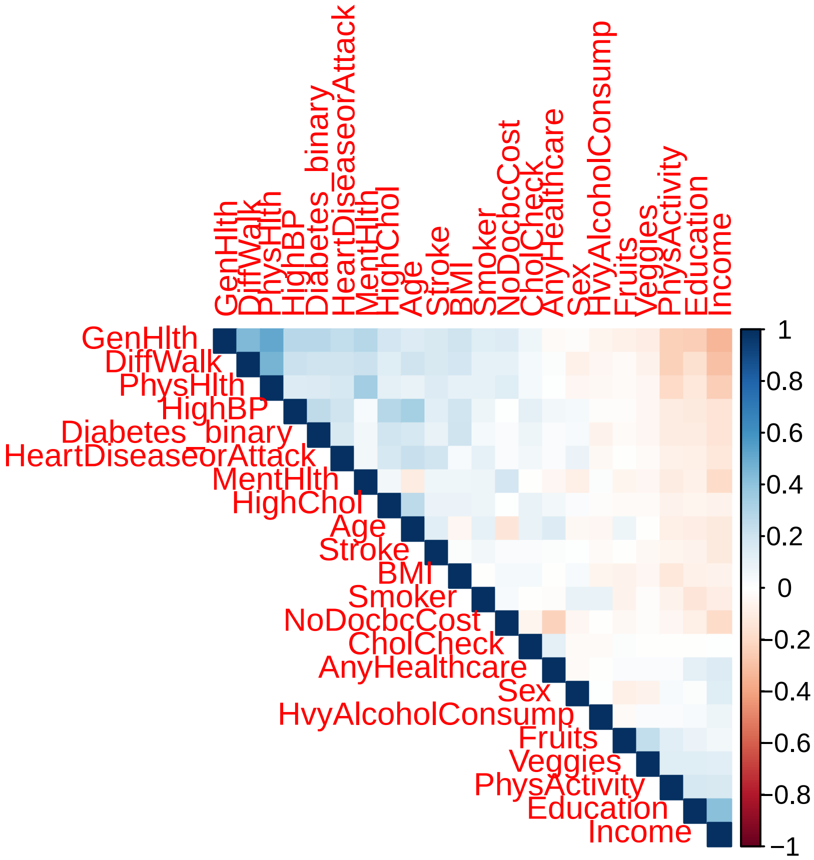
### 0.2.4 Exploratory Data analysis

```r
glimpse(diabetes)
```

```
## Rows: 229,474
## Columns: 22
## $ Diabetes_binary      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0...
## $ HighBP               <dbl> 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1...
## $ HighChol             <dbl> 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0...
## $ CholCheck            <dbl> 1, 0, 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...
## $ Smoker               <dbl> 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0...
## $ Stroke               <dbl> 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...
## $ PhysActivity         <dbl> 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1...
## $ Fruits               <dbl> 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0...
## $ Veggies              <dbl> 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0...
## $ HvyAlcoholConsump    <dbl> 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...
## $ NoDocbcCost          <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0...
## $ GenHlth              <dbl> 5, 3, 5, 2, 2, 2, 3, 3, 5, 2, 3, 3, 3, 4, 4, 2...
## $ MentHlth             <dbl> 18, 0, 30, 0, 3, 0, 0, 0, 30, 0, 0, 0, 0, 0, 3...
## $ PhysHlth             <dbl> 15, 0, 30, 0, 0, 2, 14, 0, 30, 0, 0, 30, 15, 0...
## $ DiffWalk             <dbl> 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0...
## $ Sex                  <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0...
## $ Age                  <dbl> 9, 7, 9, 11, 11, 10, 9, 11, 9, 8, 13, 10, 7, 1...
## $ Education            <dbl> 4, 6, 4, 3, 5, 6, 6, 4, 5, 4, 6, 5, 5, 4, 6, 6...
## $ Income               <dbl> 3, 1, 8, 6, 4, 8, 7, 4, 1, 3, 8, 1, 7, 6, 2, 8...
```

```
correlation <- cor(diabetes, use= "everything",  method = "pearson")

corrplot(corr = correlation, method= "color",  type= "upper", order= "FPC")
```

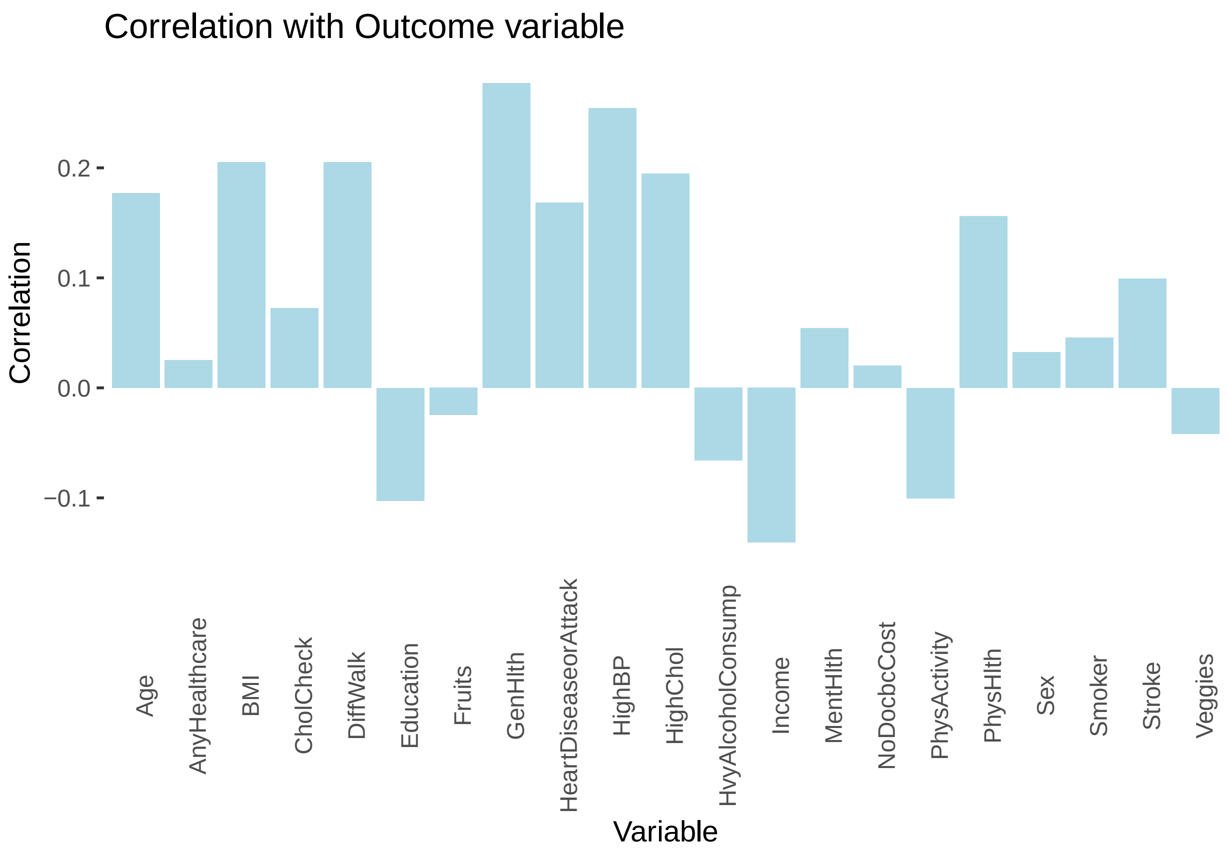


#### 0.2.4.1 Correlation

```
outcome_corr <- cor(diabetes, use = "everything", method = "pearson")[, "Diabetes_binary"]

outcome <- rownames_to_column(as.data.frame(outcome_corr), var = "variable")
outcome <- outcome %>% filter(variable!= "Diabetes_binary")
```

```
ggplot(data= outcome)+
 geom_bar(aes(y=outcome_corr, x= variable), fill="lightblue", stat="identity")+
  labs(x= "Variable", y = "Correlation", title= "Correlation with Outcome variable" )+
  theme(axis.text.x = element_text(angle=90),
        rect = element_blank(),
        axis.ticks.x = element_blank())
```



Although the correlation of the feature variables with the outcome variable does not appear to largely significant, some noteworthy positive correlations involves Age, BMI, Difficulty walking, General health, Physical health score, History of stroke, heart disease, high cholesterol levels and high blood pressure; the negative correlations involves Income, Education level and Physical activities.

*Convert the variables to factors*

```
diabetes$Diabetes_binary <- factor(diabetes$Diabetes_binary, levels= c(0, 1),
                                      labels=c("Non-diabetic", "Diabetic"))

diabetes$HighBP <- factor(diabetes$HighBP, levels = c(0, 1),
                          labels= c("non-highBP", "HighBP"))

diabetes$HighChol <- factor(diabetes$HighChol, levels = c(0, 1),
                            labels= c("non-HighChol", "highChol"))

diabetes$Smoker <- factor(diabetes$Smoker, levels = c(0, 1),
                          labels= c("nonSmoker", "Smoker"))

diabetes$Stroke <- factor(diabetes$Stroke, levels = c(0, 1),
                          labels= c("nostroke", "stroke"))

diabetes$HeartDiseaseorAttack <- factor(diabetes$HeartDiseaseorAttack, levels = c(0, 1),
                          labels= c("noheartattack", "heartattack"))

diabetes$CholCheck <- factor(diabetes$CholCheck , levels = c(0, 1),
                          labels= c("noCholCheck", "yesCholCheck"))
```

```
diabetes$PhysActivity <- factor(diabetes$PhysActivity , levels = c(0, 1),
                                labels= c("no", "yes"))

diabetes$Fruits <- factor(diabetes$Fruits , levels = c(0, 1),
                          labels= c("no", "yes"))

diabetes$HvyAlcoholConsump <- factor(diabetes$HvyAlcoholConsump, levels = c(0, 1),
                                     labels= c("no", "yes"))

diabetes$Veggies <- factor(diabetes$Veggies, levels = c(0, 1),
                           labels= c("no", "yes"))

diabetes$AnyHealthcare <- factor(diabetes$AnyHealthcare, levels = c(0, 1),
                                 labels= c("no", "yes"))

diabetes$NoDocbcCost <- factor(diabetes$NoDocbcCost, levels = c(0, 1),
                               labels= c("no", "yes"))

diabetes$DiffWalk <- factor(diabetes$DiffWalk, levels = c(0, 1),
                            labels= c("no", "yes"))

diabetes$Sex <- factor(diabetes$Sex, levels = c(0, 1),
                       labels= c("female", "male"))

diabetes$Education <- as.factor(diabetes$Education)

diabetes$Income <- as.factor(diabetes$Income)

diabetes$Age <- as.factor(diabetes$Age)

diabetes$GenHlth <- as.factor(diabetes$GenHlth)

diabetes$MentHlth <- as.factor(diabetes$MentHlth)

diabetes$PhysHlth <- as.factor(diabetes$PhysHlth)

diabetes <- diabetes %>%mutate( BMIcat=case_when(diabetes$BMI < 19 ~ "Underweight",
                                                 diabetes$BMI < 25 ~ "Healthyweight",
                                                 diabetes$BMI < 30 ~ "Overweight",
                           TRUE ~ "Obese"))
diabetes$BMIcat <- factor(diabetes$BMIcat, levels = c("Underweight", "Healthyweight", "Overweight", "Obe
```

```
head(diabetes, 10)
```

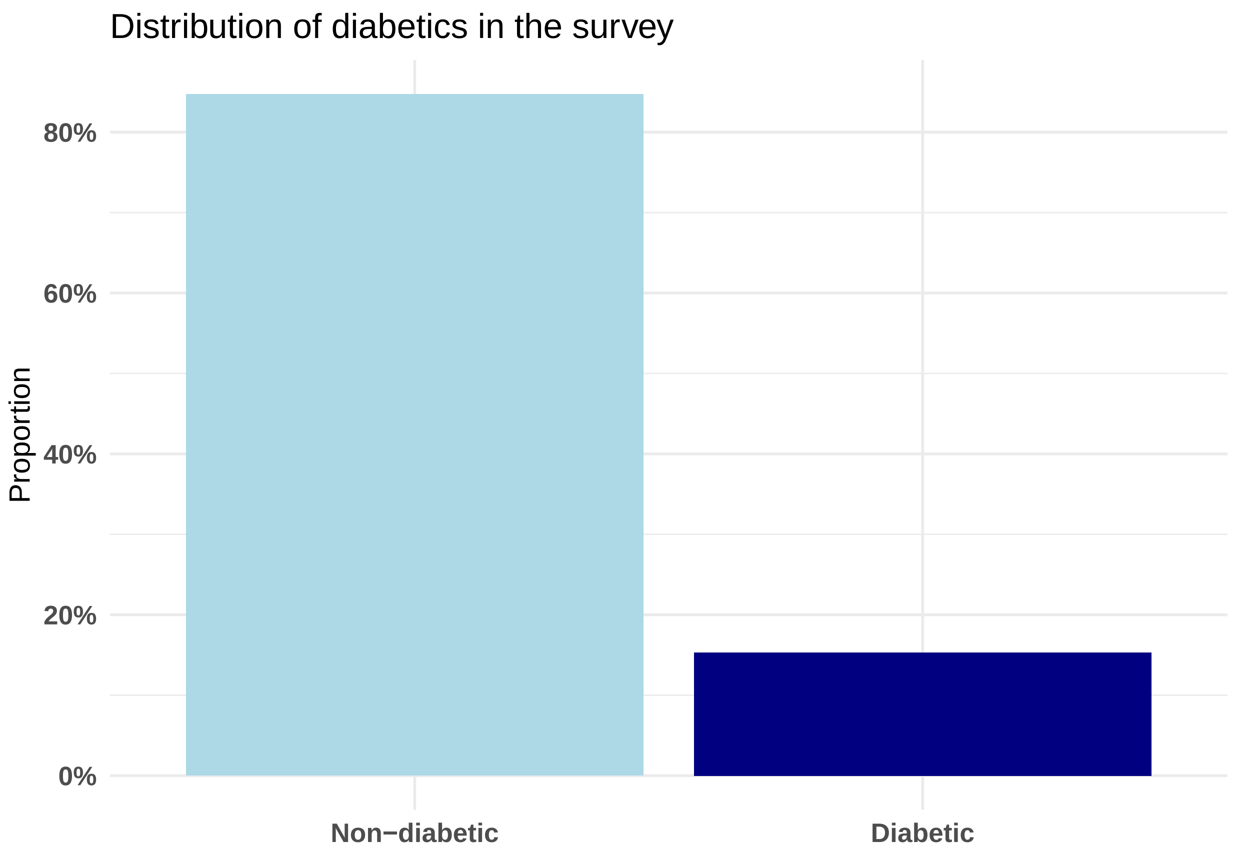
```
## # A tibble: 10 x 23
##    Diabetes_binary HighBP HighChol CholCheck   BMI Smoker Stroke
##    <fct>           <fct>  <fct>    <fct>     <dbl> <fct>  <fct>
##  1 Non-diabetic    HighBP highChol yesCholC~    40 Smoker nostr~
##  2 Non-diabetic    non-h~ non-Hig~ noCholCh~    25 Smoker nostr~
##  3 Non-diabetic    HighBP highChol yesCholC~    28 nonSm~ nostr~
##  4 Non-diabetic    HighBP non-Hig~ yesCholC~    27 nonSm~ nostr~
##  5 Non-diabetic    HighBP highChol yesCholC~    24 nonSm~ nostr~
##  6 Non-diabetic    HighBP highChol yesCholC~    25 Smoker nostr~
##  7 Non-diabetic    HighBP non-Hig~ yesCholC~    30 Smoker nostr~
##  8 Non-diabetic    HighBP highChol yesCholC~    25 Smoker nostr~
##  9 Diabetic        HighBP highChol yesCholC~    30 Smoker nostr~
## 10 Non-diabetic    non-h~ non-Hig~ yesCholC~    24 nonSm~ nostr~
## # ... with 16 more variables: HeartDiseaseorAttack <fct>, PhysActivity <fct>,
## #   Fruits <fct>, Veggies <fct>, HvyAlcoholConsump <fct>, AnyHealthcare <fct>,
## #   NoDocbcCost <fct>, GenHlth <fct>, MentHlth <fct>, PhysHlth <fct>,
## #   DiffWalk <fct>, Sex <fct>, Age <fct>, Education <fct>, Income <fct>,
## #   BMIcat <fct>
```

```r
table(diabetes$Diabetes_binary)
```

### 0.2.4.2 Outcome variable:

```
## 
## Non-diabetic     Diabetic 
##       194377        35097
```

```r
ggplot(data= diabetes)+
  geom_bar(aes(x=Diabetes_binary, fill= Diabetes_binary,
               y=after_stat(count/sum(count))))+
  scale_fill_manual(values= c( "lightblue", "navy"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"),
        axis.title.x = element_blank(),
        legend.position = "none")+
  labs( y = "Proportion",
       title = "Distribution of diabetics in the survey")
```



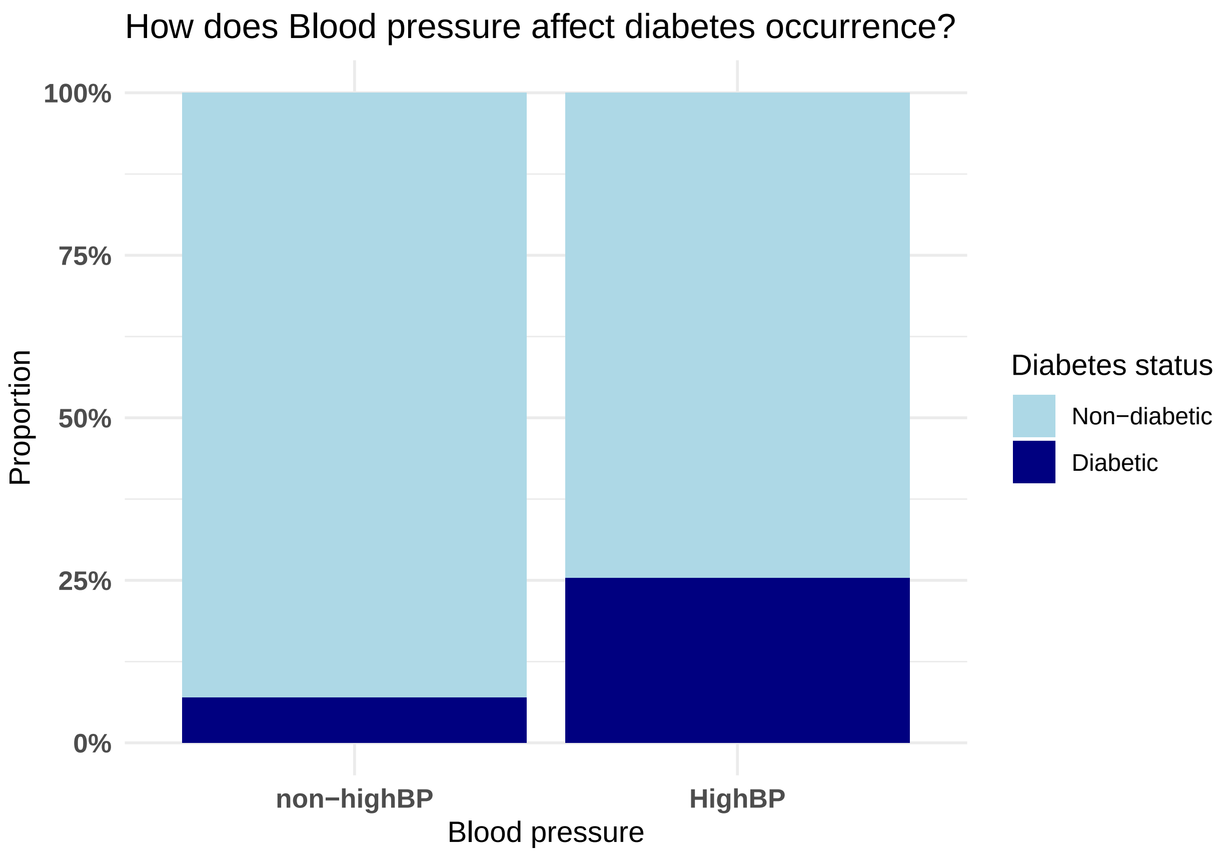
The distribution of the outcome variable in the survey is 35,097 (15%) diabetic and 194,377 (85%) non-diabetic.

#### 0.2.4.3 Relationship between Medical History and Occurrence *High blood pressure*

```
table( diabetes$HighBP, diabetes$Diabetes_binary) %>% prop.table( 1)
```

```
##
##              Non-diabetic   Diabetic
##   non-highBP   0.93058284 0.06941716
##   HighBP       0.74673892 0.25326108
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=HighBP, fill= Diabetes_binary), position = "fill")+
  scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Blood pressure",
       y = "Proportion",
       title = "How does Blood pressure affect diabetes occurrence?")
```



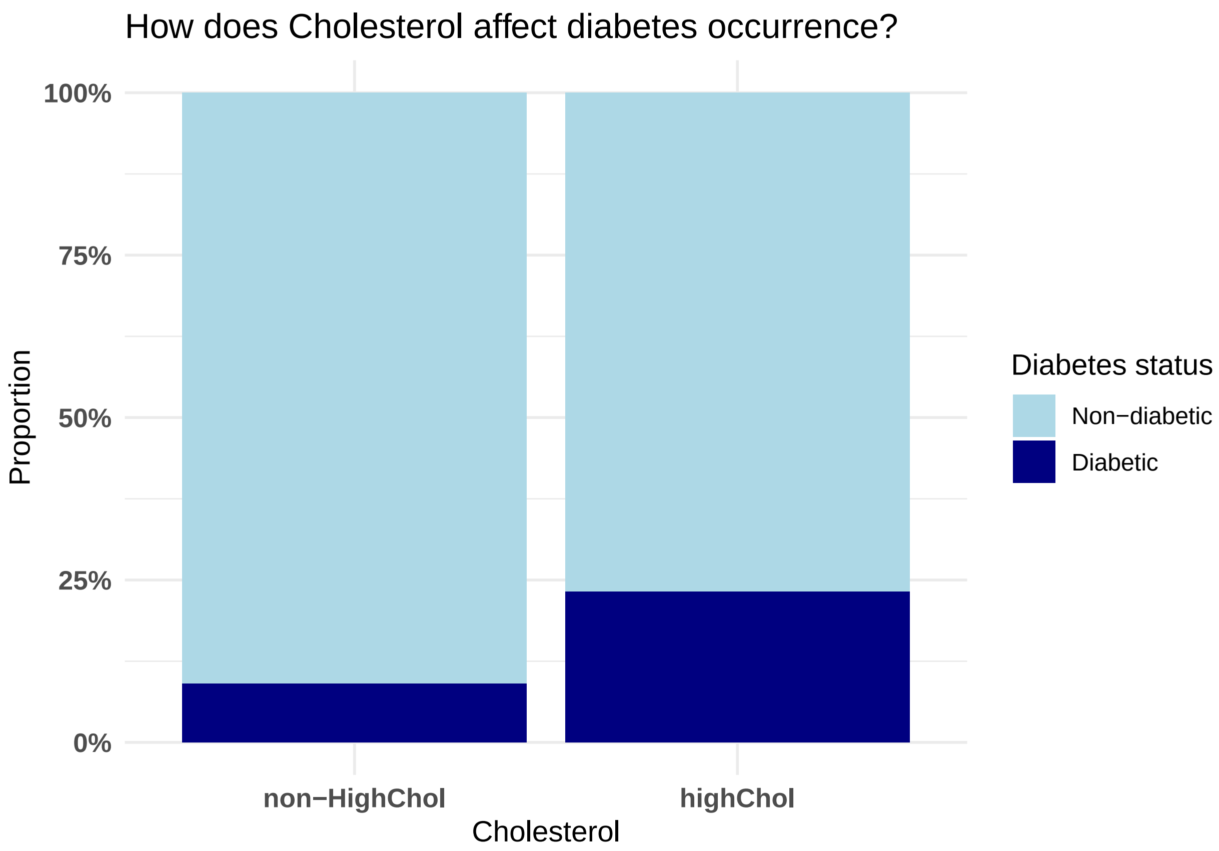
There's an increased proportion of diabetes among individuals with High blood pressure compared to those with low blood pressure

*High Cholesterol*

```
table(diabetes$HighChol, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##                 Non-diabetic   Diabetic
##   non-HighChol   0.90945844 0.09054156
##   highChol       0.76815827 0.23184173
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=HighChol, fill= Diabetes_binary),position = "fill")+
  scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Cholesterol",
       y = "Proportion",
       title = "How does Cholesterol affect diabetes occurrence?")
```



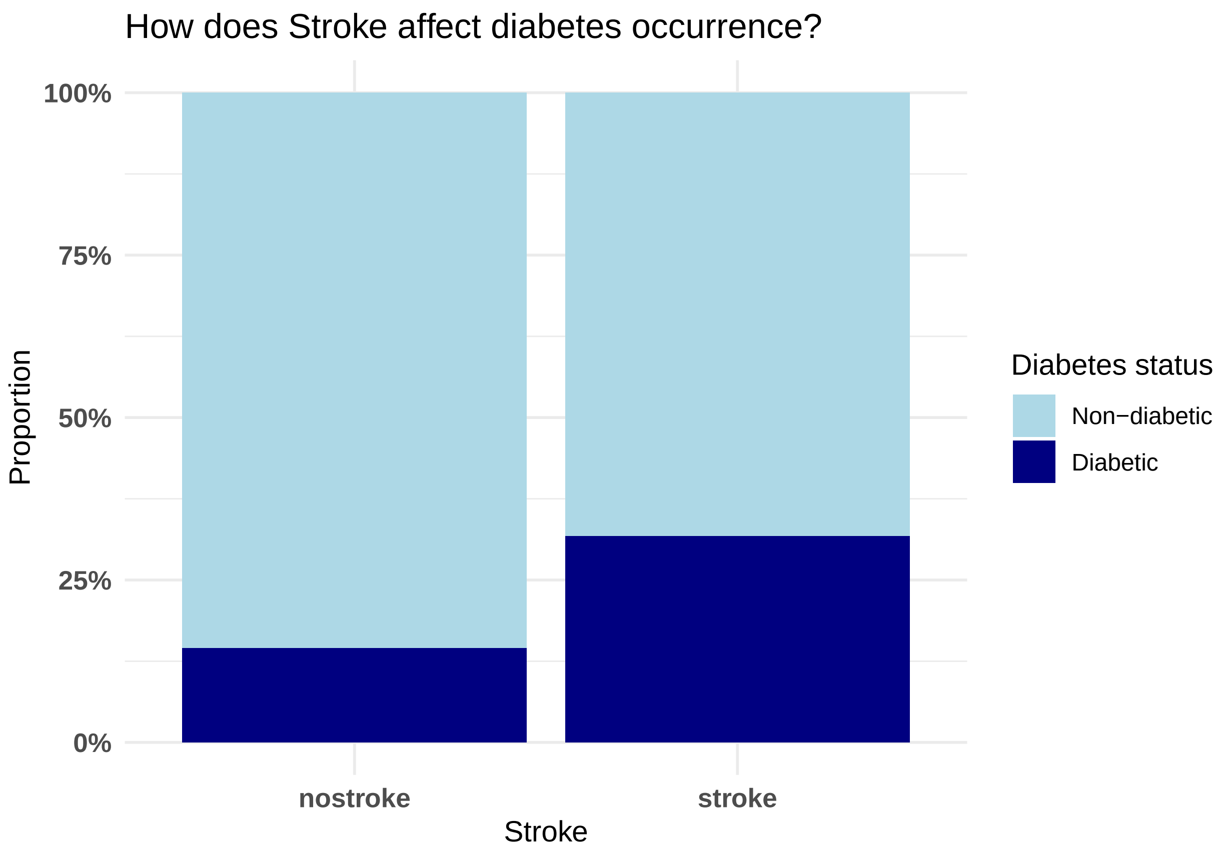
There's an increased proportion of diabetes among individuals with High Cholesterol levels compared to those with low cholesterol levels.

*Stroke*

```
table( diabetes$Stroke, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##             Non-diabetic  Diabetic
##   nostroke     0.8547881 0.1452119
##   stroke       0.6822248 0.3177752
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=Stroke, fill= Diabetes_binary), position = "fill")+
  scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Stroke",
       y = "Proportion",
       title = "How does Stroke affect diabetes occurrence?")
```



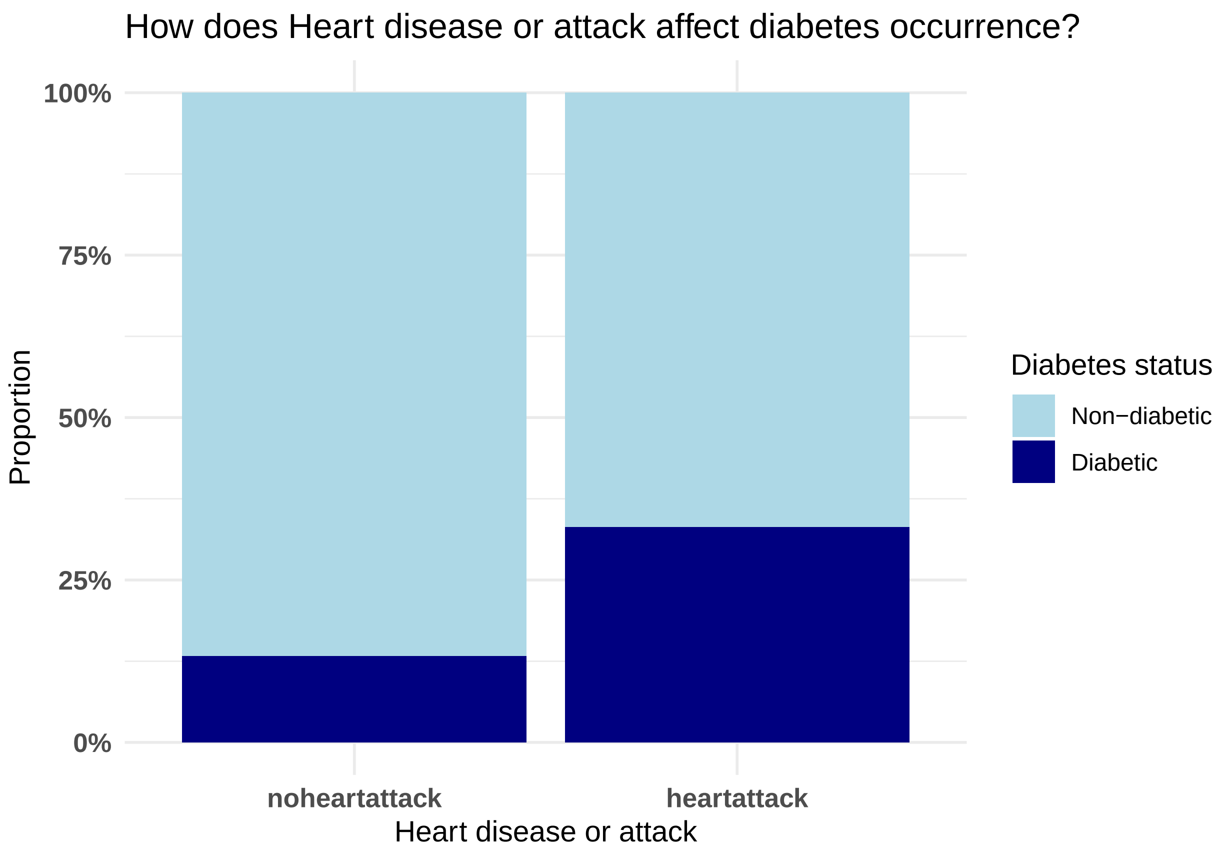
There’s an increase in proportion in the occurrence of diabetes among individuals with hsitory of stroke compared to those who have never had stroke.

*Heart disease or attack*

```
table( diabetes$HeartDiseaseorAttack, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##                  Non-diabetic  Diabetic
##   noheartattack     0.8676085 0.1323915
##   heartattack       0.6687049 0.3312951
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=HeartDiseaseorAttack, fill= Diabetes_binary), position = "fill")+
  scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Heart disease or attack",
       y = "Proportion",
       title = "How does Heart disease or attack affect diabetes occurrence?")
```



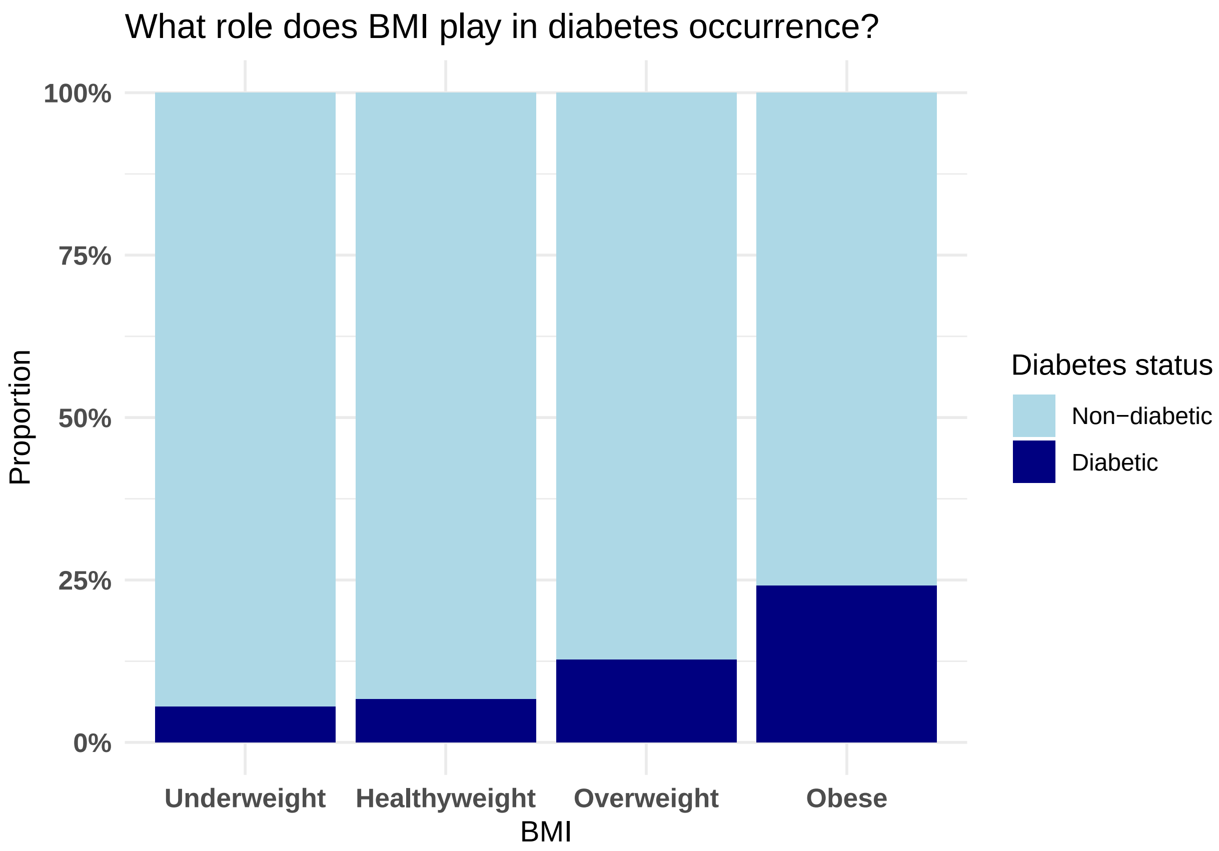
There's an increase in proportion in the occurrence of diabetes among individuals with hsitory of Heart attack or Myocardial Infarction compared to those who have never had one.

*Body Mass Index*

```
table(diabetes$BMIcat, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##                   Non-diabetic   Diabetic
##   Underweight       0.94462647 0.05537353
##   Healthyweight     0.93355087 0.06644913
##   Overweight        0.87228461 0.12771539
##   Obese             0.75897097 0.24102903
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=BMIcat, fill= Diabetes_binary), position = "fill")+
  scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "BMI",
       y = "Proportion",
       title = "What role does BMI play in diabetes occurrence?")
```



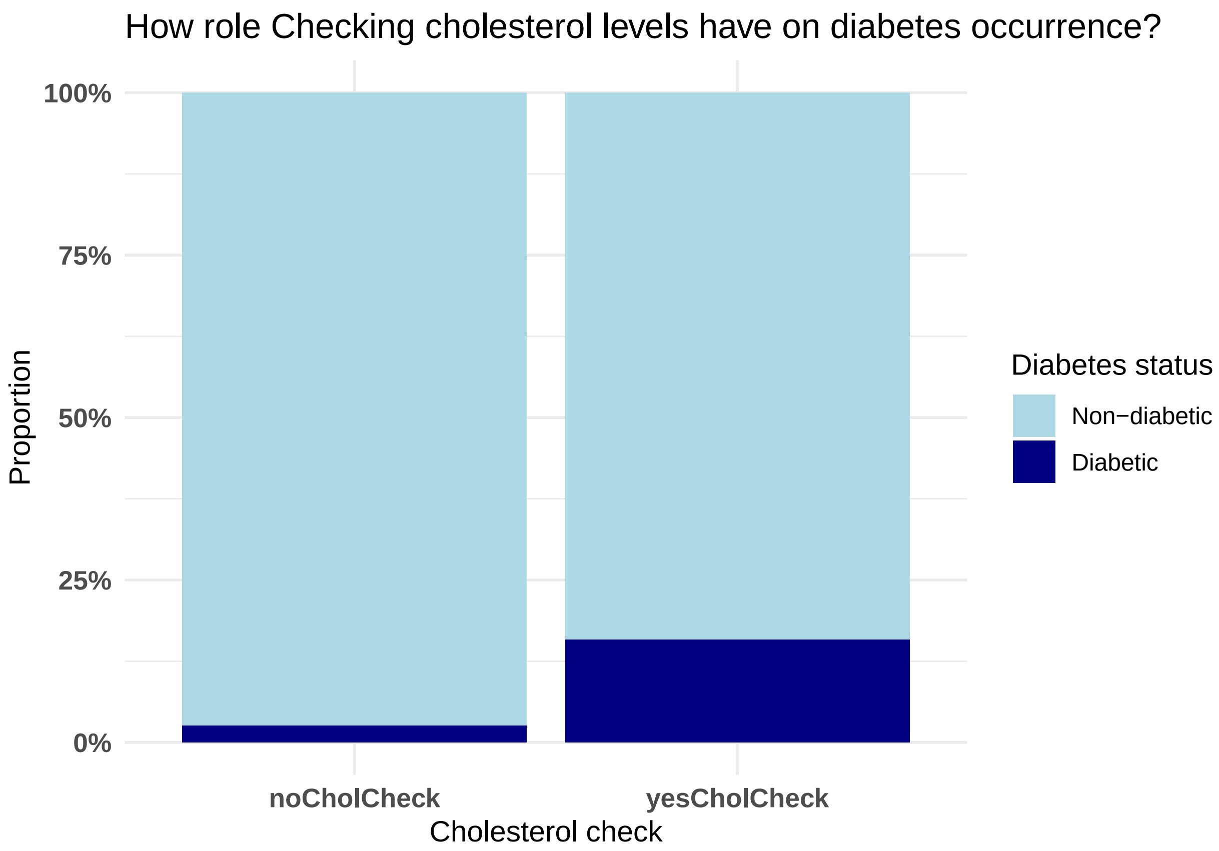
Increasing BMI corresponds to an increase in proportion of individuals with diabetes

*Cholesterol check frequency*

```
table(diabetes$CholCheck, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##                 Non-diabetic   Diabetic
##   noCholCheck     0.97408045 0.02591955
##   yesCholCheck    0.84169028 0.15830972
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=CholCheck, fill= Diabetes_binary), position = "fill")+
  scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Cholesterol check",
       y = "Proportion",
       title = "How role Checking cholesterol levels have on diabetes occurrence?")
```



The presence of all medical history checked in the survey led to an increase in the proportion of diabetes occurence over those that didn't have.

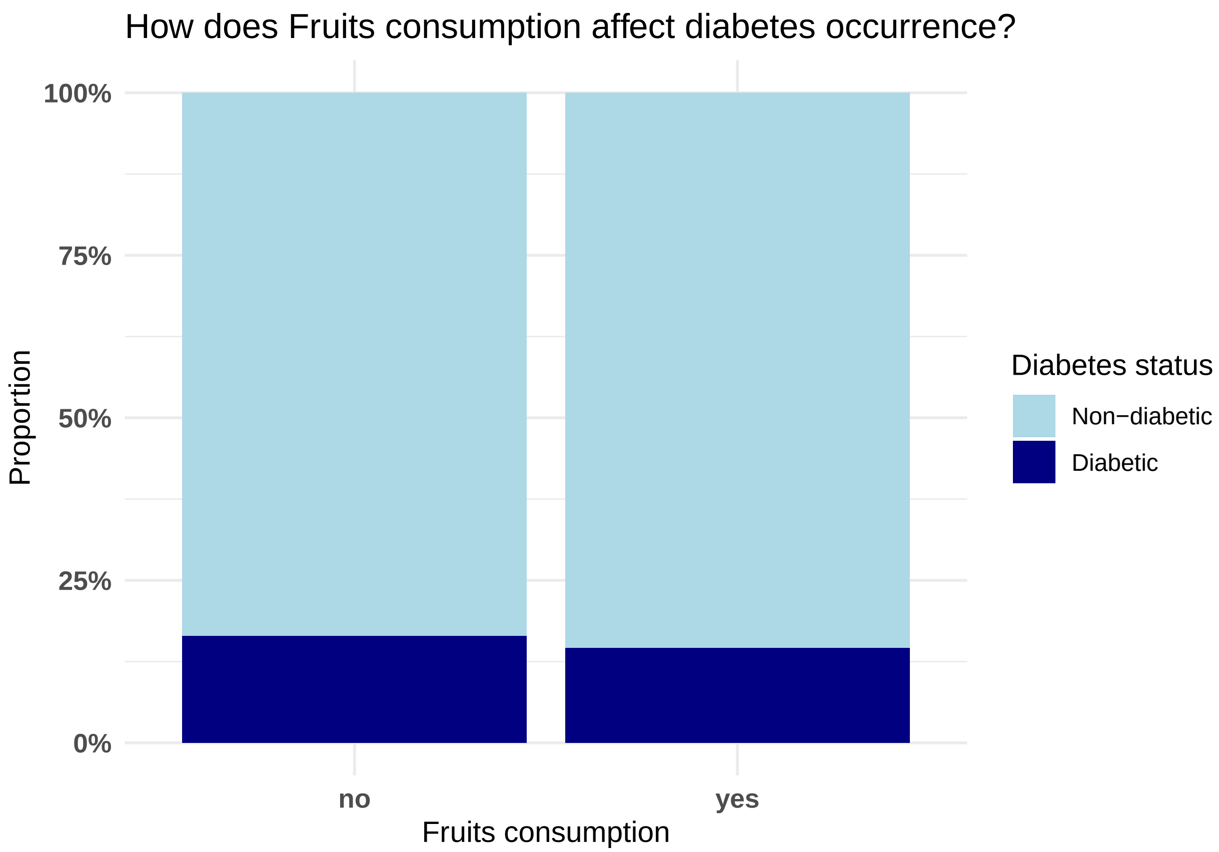
#### 0.2.4.4 Relationship between Lifestyle and occurrence of Diabetes

*Fruits consumption*

```
table(diabetes$Fruits, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##         Non-diabetic  Diabetic
##    no      0.8358254 0.1641746
##    yes     0.8541535 0.1458465
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=Fruits, fill= Diabetes_binary), position = "fill")+
  scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Fruits consumption",
       y = "Proportion",
       title = "How does Fruits consumption affect diabetes occurrence?")
```



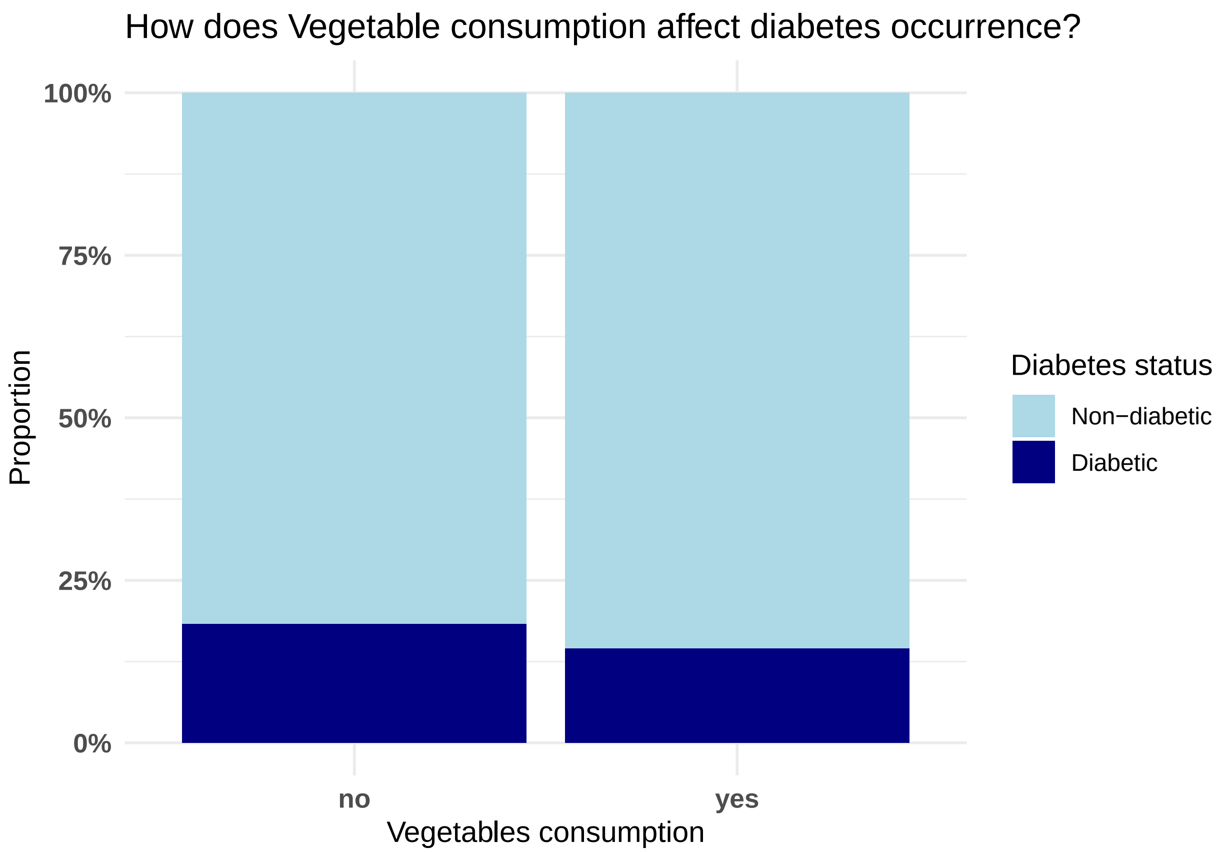
There is no significant difference in the proportion of diabetes occurrence among people who consume fruits once or more a day

*Veggies consumption*

```
table(diabetes$Veggies, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##       Non-diabetic  Diabetic
##   no     0.8175107 0.1824893
##   yes    0.8546921 0.1453079
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=Veggies, fill= Diabetes_binary), position = "fill")+
  scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Vegetables consumption",
       y = "Proportion",
       title = "How does Vegetable consumption affect diabetes occurrence?")
```



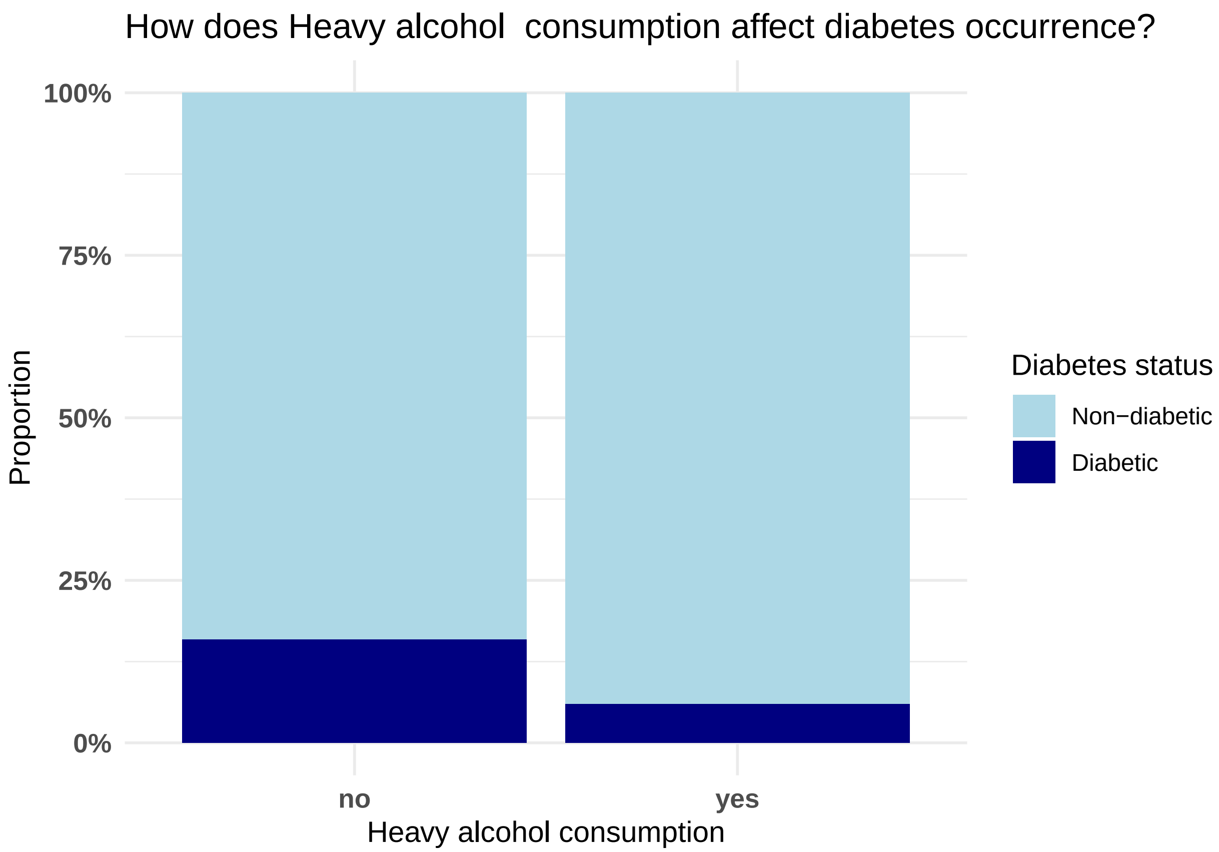
There is no significant difference in the proportion of diabetes occurrence among people who consume vegetables once or more a day

*Heavy alcohol consumption*

```r
table(diabetes$HvyAlcoholConsump, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##        Non-diabetic   Diabetic
##    no    0.84101539 0.15898461
##    yes   0.94035842 0.05964158
```

```r
ggplot(data= diabetes)+
  geom_bar(aes(x=HvyAlcoholConsump, fill= Diabetes_binary), position = "fill")+
  scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Heavy alcohol consumption",
       y = "Proportion",
       title = "How does Heavy alcohol  consumption affect diabetes occurrence?")
```



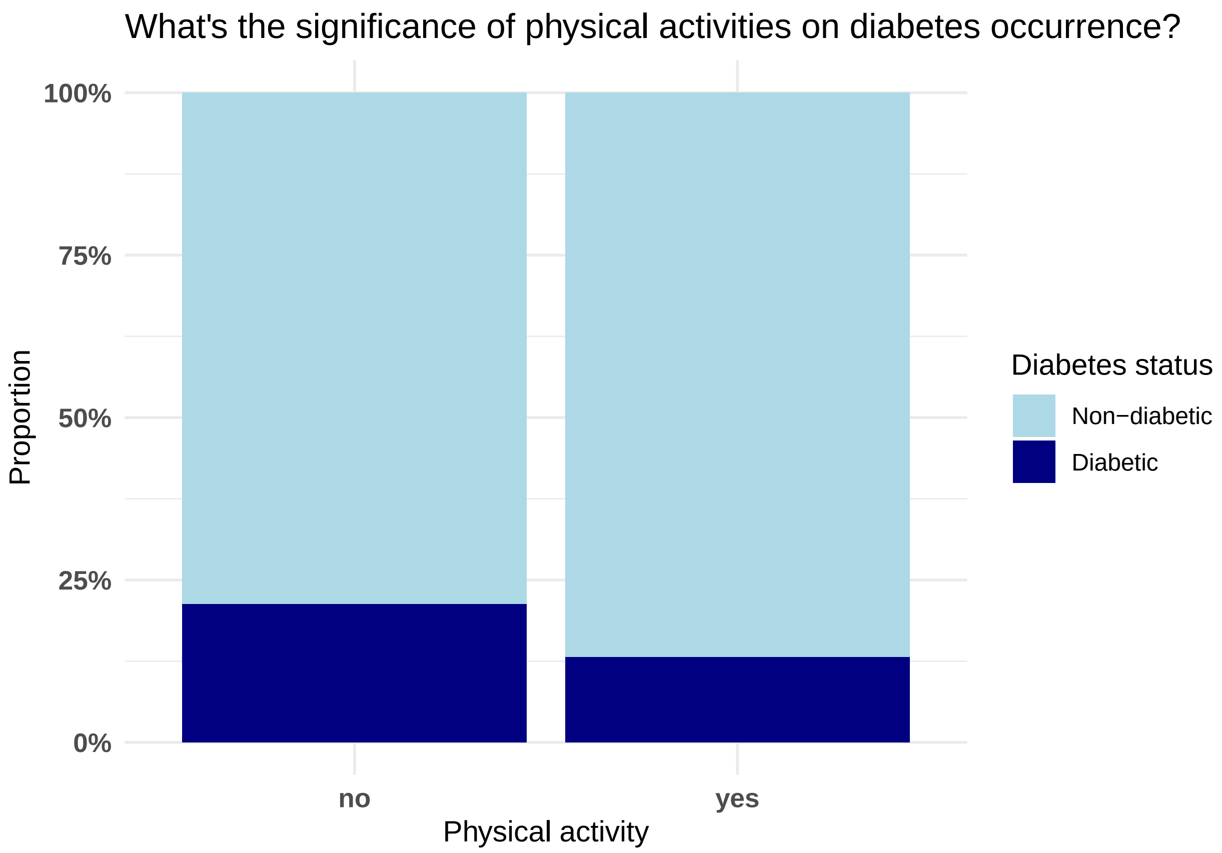
There seems to be a decreased proportion of diabetic patients in heavy alcohol consumers than those that aren't.

*Physical activity*

```
table(diabetes$PhysActivity, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##        Non-diabetic  Diabetic
##    no     0.7871694 0.2128306
##    yes    0.8688635 0.1311365
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=PhysActivity, fill= Diabetes_binary),position = "fill")+
  scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Physical activity",
       y = "Proportion",
       title = "What's the significance of physical activities on diabetes occurrence?")
```



Physically active individuals showed a lower proportion of diabetes occurrence than those not as physically active.

Fruits and vegetable consumption didn't show any significant change in proportion of diabetes occurrence. Physically more acive indiduals had lesser proportion of diabetes occurence. Heavy alcohol consumption showed an unexpected decrease in proportion of diabetes.

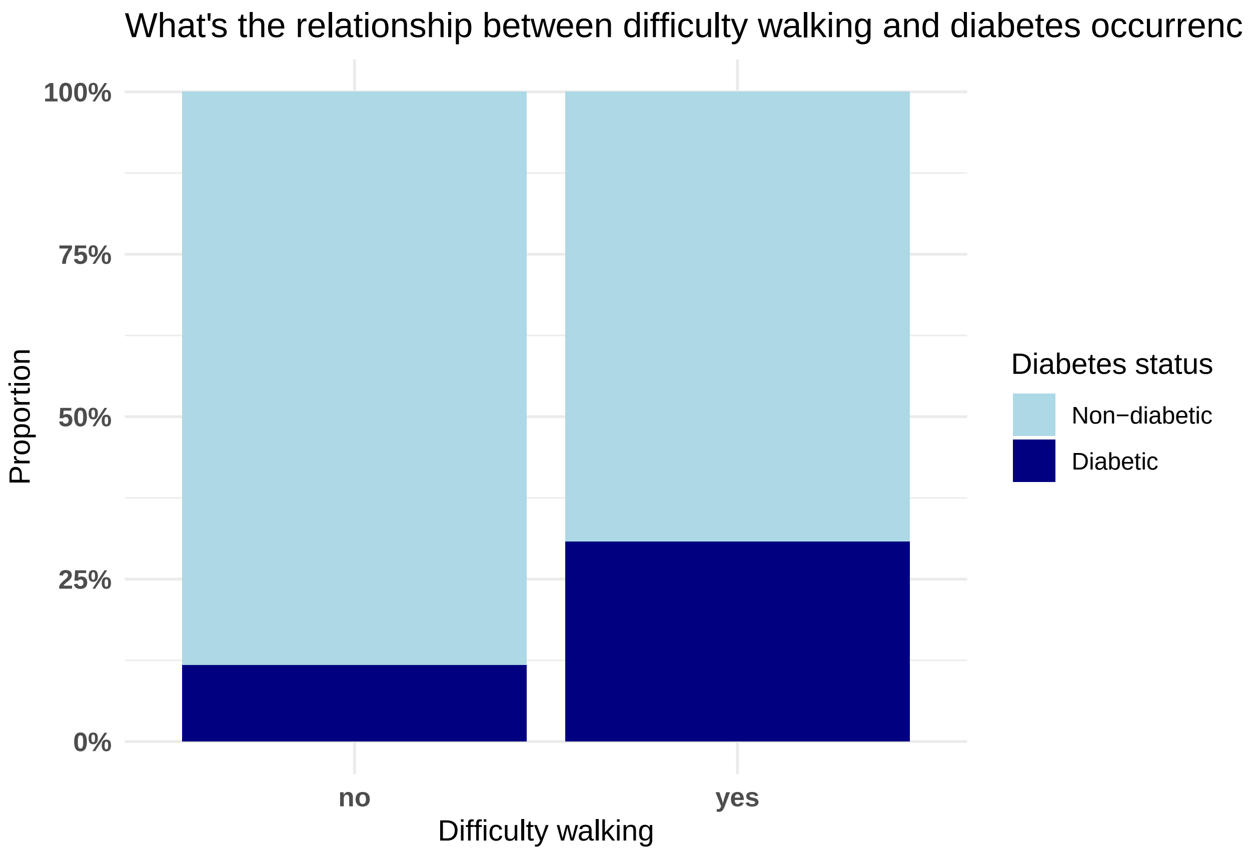
#### 0.2.4.5 Relationship between Health / Healthcare and occurrence of Diabetes

*Difficulty walking*

```
table(diabetes$DiffWalk, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##        Non-diabetic  Diabetic
##   no      0.8823488 0.1176512
##   yes     0.6923402 0.3076598
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=DiffWalk, fill= Diabetes_binary), position = "fill")+
  scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Difficulty walking",
       y = "Proportion",
       title = "What's the relationship between difficulty walking and diabetes occurrence?")
```



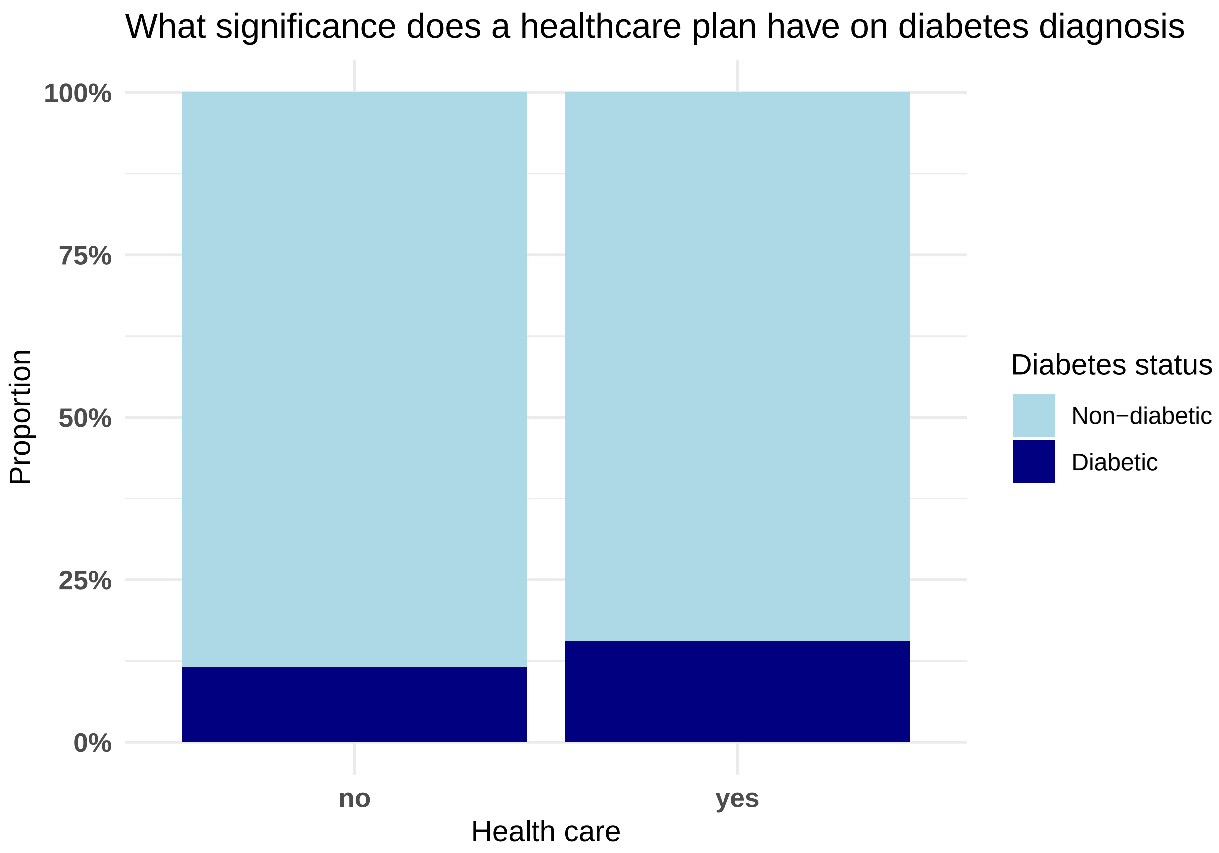
More individuals with difficulty walking appear to be diabetic than those without the difficulty.

*On any Healthcare plan*

```
table(diabetes$AnyHealthcare, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##       Non-diabetic  Diabetic
##   no     0.8852208 0.1147792
##   yes    0.8448764 0.1551236
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=AnyHealthcare, fill= Diabetes_binary), position = "fill")+
   scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Health care",
       y = "Proportion",
       title = "What significance does a healthcare plan have on diabetes diagnosis")
```



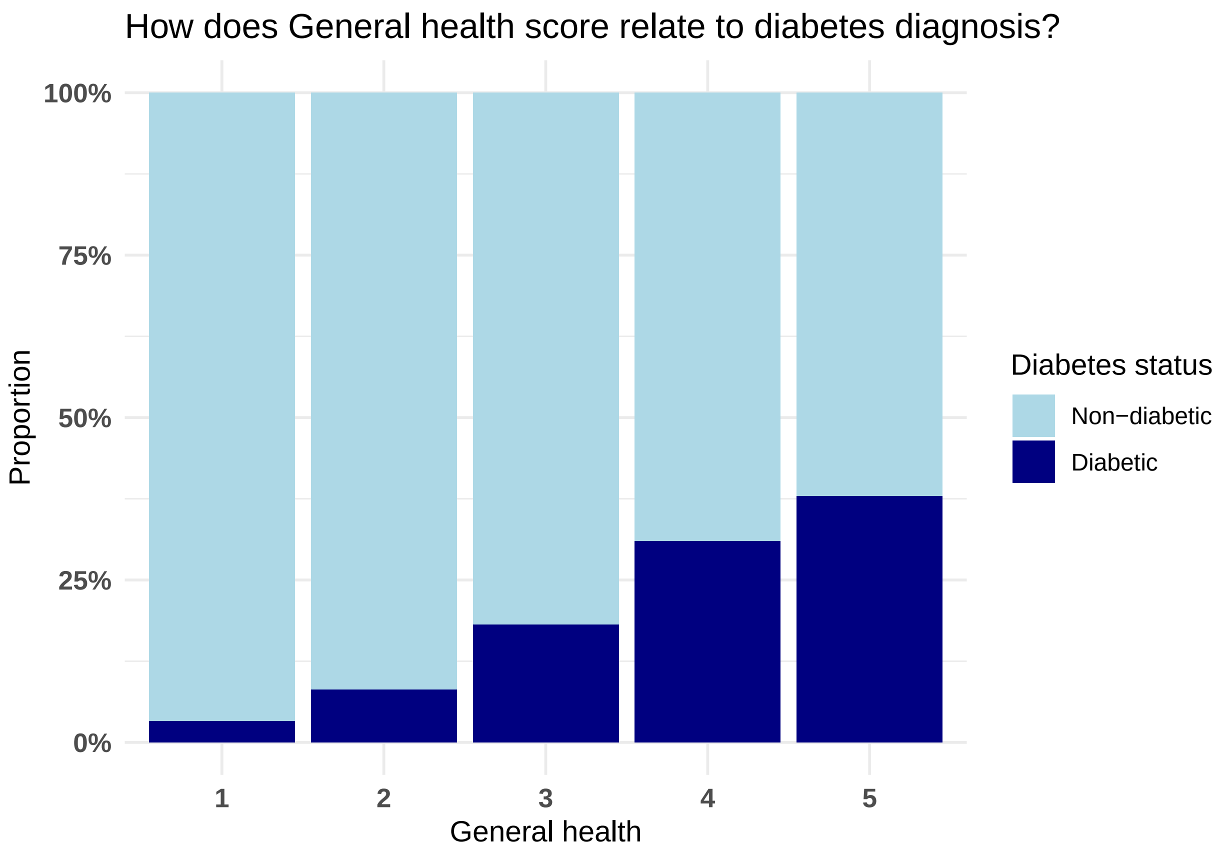
The presence of a health care plan corresponds with an increase in positive diabetese diagnosis.

*General Health*

```r
table(diabetes$GenHlth, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##     Non-diabetic   Diabetic
##   1   0.96743559 0.03256441
##   2   0.91882634 0.08117366
##   3   0.81904607 0.18095393
##   4   0.68993501 0.31006499
##   5   0.62104653 0.37895347
```

```r
ggplot(data= diabetes)+
  geom_bar(aes(x=GenHlth, fill= Diabetes_binary), position = "fill")+
   scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "General health",
       y = "Proportion",
       title = "How does General health score relate to diabetes diagnosis?")
```



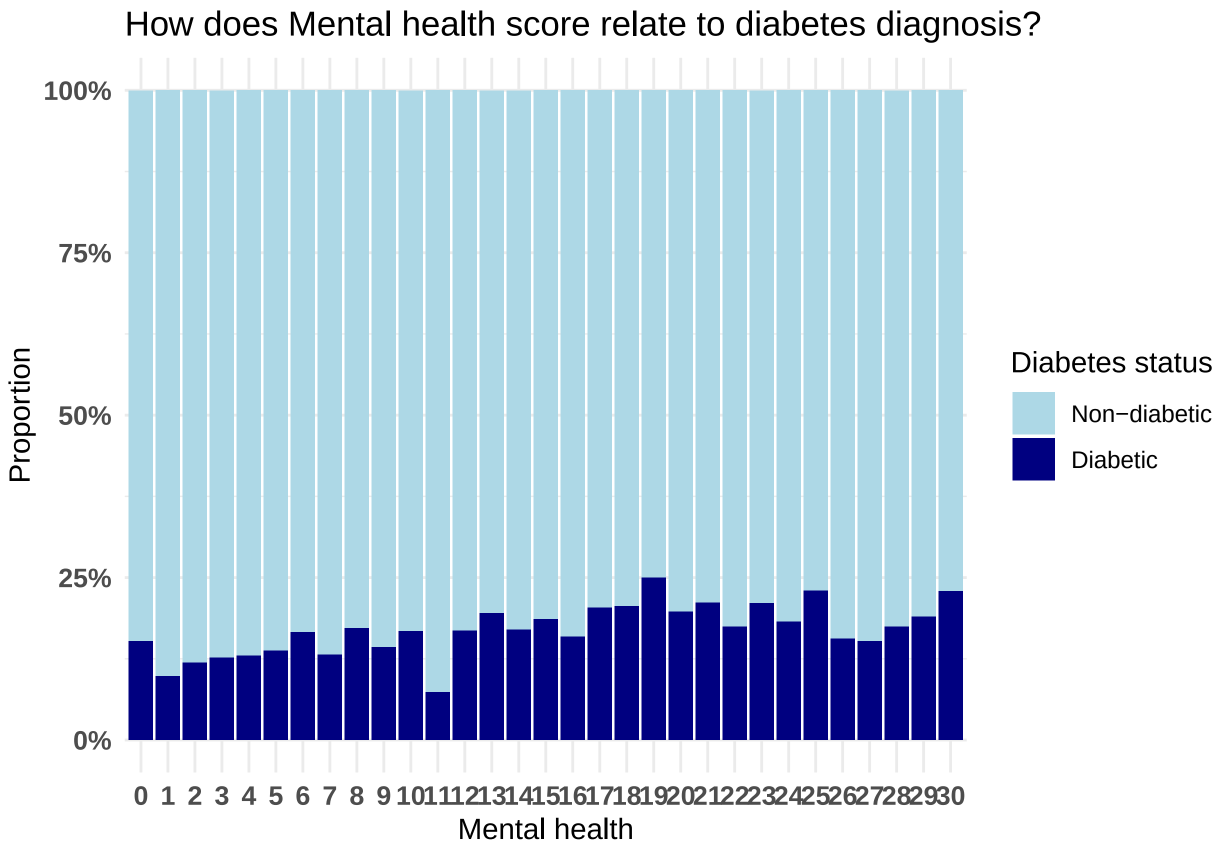
There's a visible increase in the occurrence of diabetes in individuals who scored themselves lower in their general health status

*Mental Health*

```
table(diabetes$MentHlth, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##      Non-diabetic   Diabetic
##   0    0.84798950 0.15201050
##   1    0.90225111 0.09774889
##   2    0.88126379 0.11873621
##   3    0.87344199 0.12655801
##   4    0.87042925 0.12957075
##   5    0.86278470 0.13721530
##   6    0.83400810 0.16599190
##   7    0.86893204 0.13106796
##   8    0.82785603 0.17214397
##   9    0.85714286 0.14285714
##   10   0.83249370 0.16750630
##   11   0.92682927 0.07317073
##   12   0.83165829 0.16834171
##   13   0.80487805 0.19512195
##   14   0.83033419 0.16966581
##   15   0.81403381 0.18596619
##   16   0.84090909 0.15909091
##   17   0.79629630 0.20370370
##   18   0.79381443 0.20618557
##   19   0.75000000 0.25000000
##   20   0.80279595 0.19720405
##   21   0.78854626 0.21145374
##   22   0.82539683 0.17460317
##   23   0.78947368 0.21052632
##   24   0.81818182 0.18181818
##   25   0.77020202 0.22979798
##   26   0.84444444 0.15555556
##   27   0.84810127 0.15189873
##   28   0.82568807 0.17431193
##   29   0.81012658 0.18987342
##   30   0.77084196 0.22915804
```

```r
ggplot(data= diabetes)+
  geom_bar(aes(x=MentHlth, fill= Diabetes_binary), position = "fill")+
   scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Mental health",
       y = "Proportion",
       title = "How does Mental health score relate to diabetes diagnosis?")
```



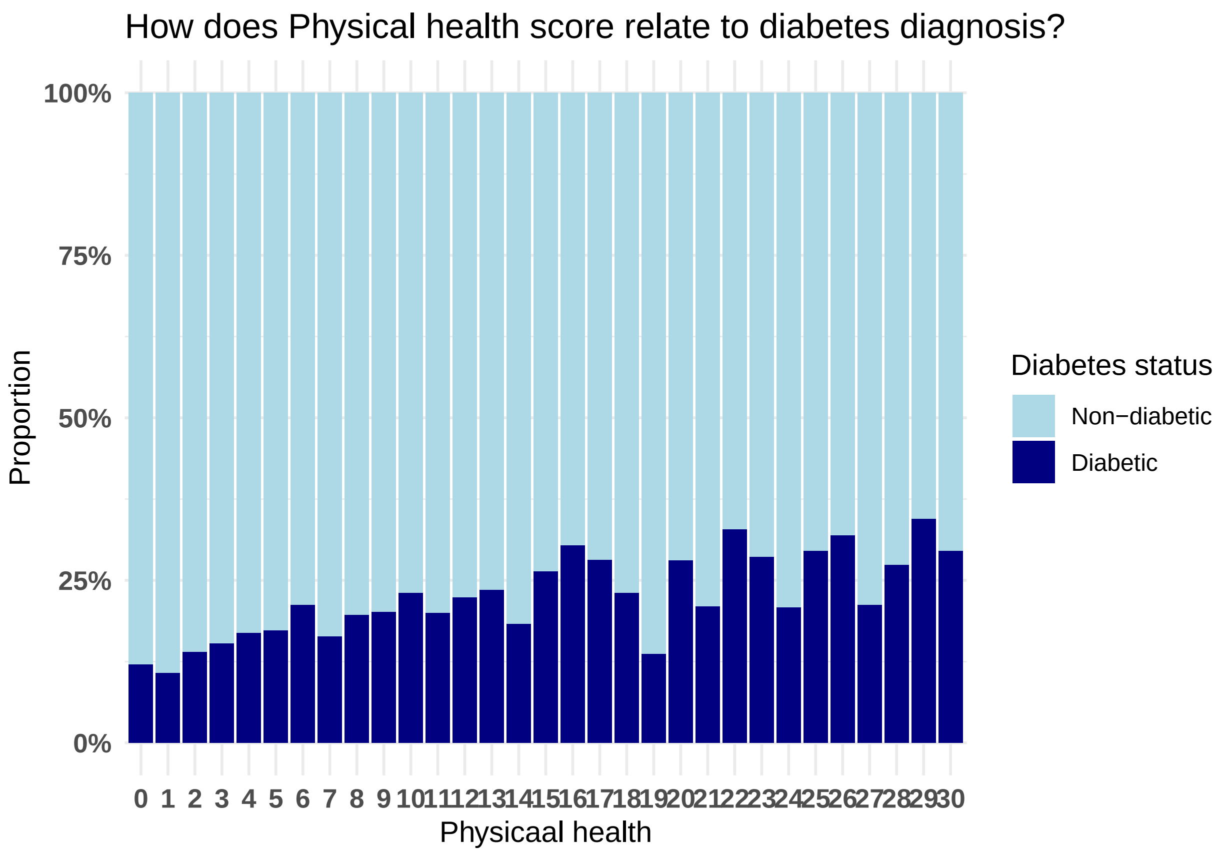
As mental health of individuals scored lower, there was a corresponding increase in proportion of individuals diagnosed with diabetes.

*Physical Health*

```
table(diabetes$PhysHlth, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##      Non-diabetic  Diabetic
##   0     0.8792119 0.1207881
##   1     0.8928023 0.1071977
##   2     0.8601201 0.1398799
##   3     0.8471844 0.1528156
##   4     0.8312320 0.1687680
##   5     0.8269914 0.1730086
##   6     0.7876506 0.2123494
##   7     0.8362392 0.1637608
##   8     0.8034611 0.1965389
##   9     0.7988827 0.2011173
##   10    0.7691482 0.2308518
##   11    0.8000000 0.2000000
##   12    0.7768166 0.2231834
##   13    0.7647059 0.2352941
##   14    0.8169505 0.1830495
##   15    0.7366707 0.2633293
##   16    0.6964286 0.3035714
##   17    0.7187500 0.2812500
##   18    0.7697368 0.2302632
##   19    0.8636364 0.1363636
##   20    0.7198289 0.2801711
##   21    0.7903469 0.2096531
##   22    0.6714286 0.3285714
##   23    0.7142857 0.2857143
##   24    0.7916667 0.2083333
##   25    0.7050898 0.2949102
##   26    0.6811594 0.3188406
##   27    0.7878788 0.2121212
##   28    0.7260536 0.2739464
##   29    0.6558140 0.3441860
##   30    0.7047201 0.2952799
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=PhysHlth, fill= Diabetes_binary), position = "fill")+
   scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Physicaal health",
       y = "Proportion",
       title = "How does Physical health score relate to diabetes diagnosis?")
```



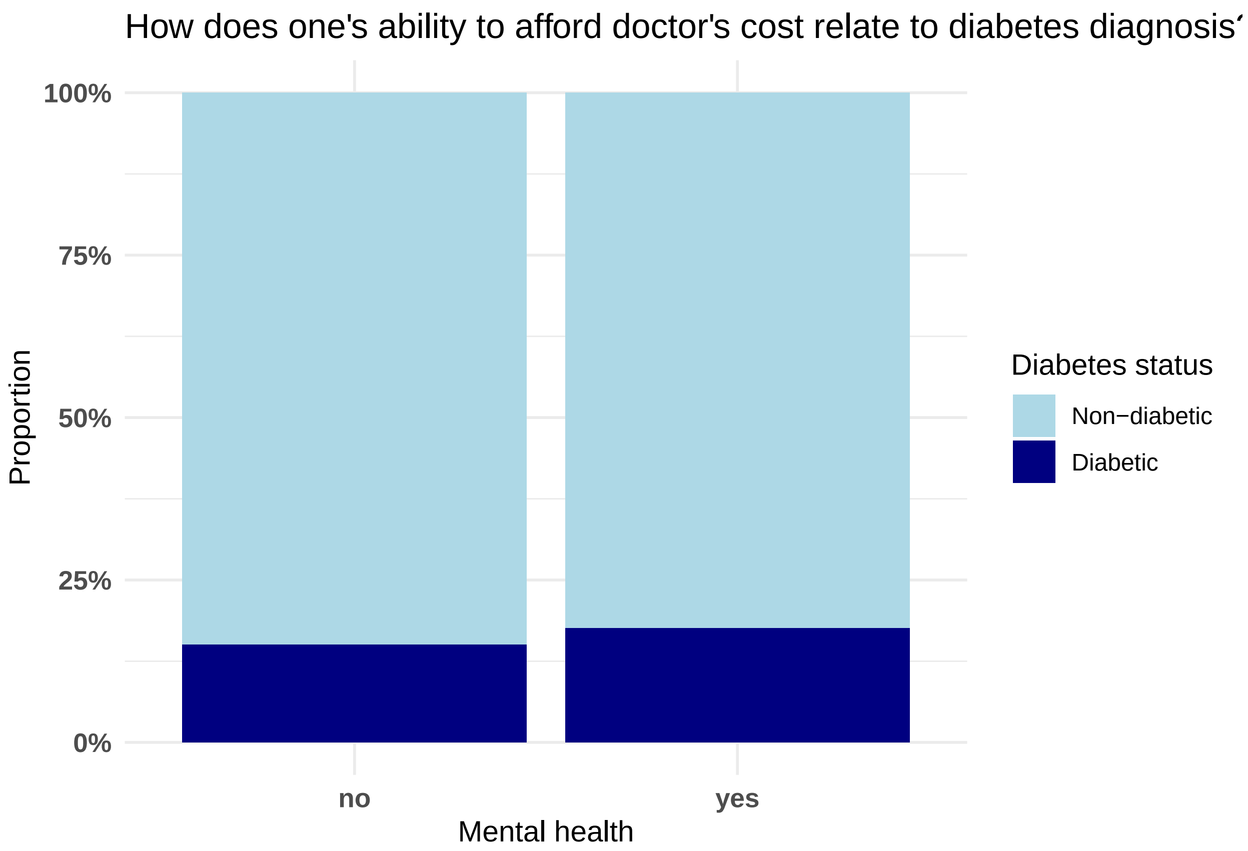
Individuals that spent more days with illness and injury were more likely to be diagnosed with diabetes than those with lesser days out.

*Ability to afford doctor's cost*

```
table(diabetes$NoDocbcCost, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##       Non-diabetic  Diabetic
##   no     0.8493642 0.1506358
##   yes    0.8245087 0.1754913
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=NoDocbcCost, fill= Diabetes_binary), position = "fill")+
   scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Mental health",
       y = "Proportion",
       title = "How does one's ability to afford doctor's cost relate to diabetes diagnosis?")
```



Ability to afford doctor's service increased the proportion of individuals with a positive diagnosis for diabetes.

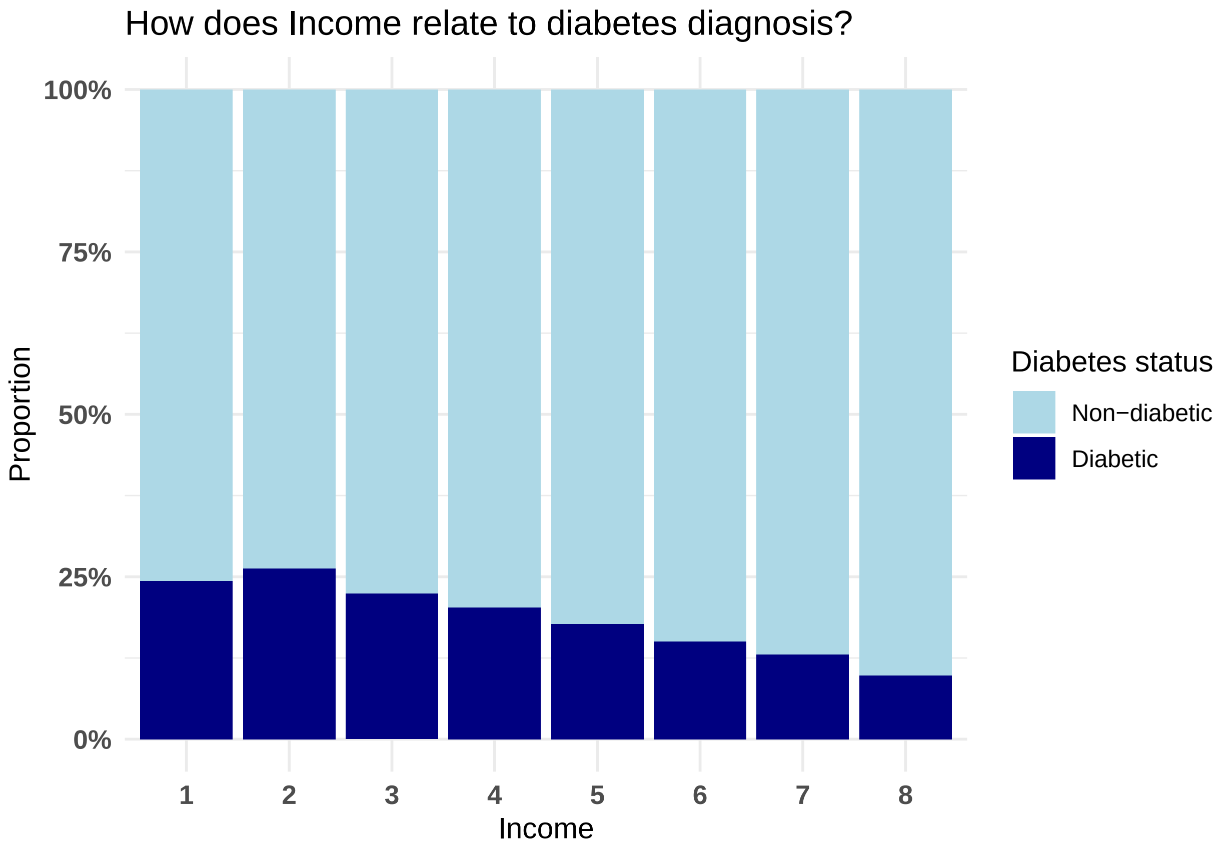
Better health status such as Mental health, Physical and General health showed a lower proportion of individuals with positive diabetes diagnosis, and the propotion increased as the health status got worse. Ability to afford healthcare, through healthcare plans or out of pocket fees increased the proportion of a positive diagnosis.

#### 0.2.4.6 Relationship between Social status and occurrence of Diabetes *Income*

```
table(diabetes$Income, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##     Non-diabetic   Diabetic
##   1   0.75661322 0.24338678
##   2   0.73749575 0.26250425
##   3   0.77613065 0.22386935
##   4   0.79717336 0.20282664
##   5   0.82275132 0.17724868
##   6   0.84952942 0.15047058
##   7   0.86977648 0.13022352
##   8   0.90170296 0.09829704
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=Income, fill= Diabetes_binary), position = "fill")+
   scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Income",
       y = "Proportion",
       title = "How does Income relate to diabetes diagnosis?")
```



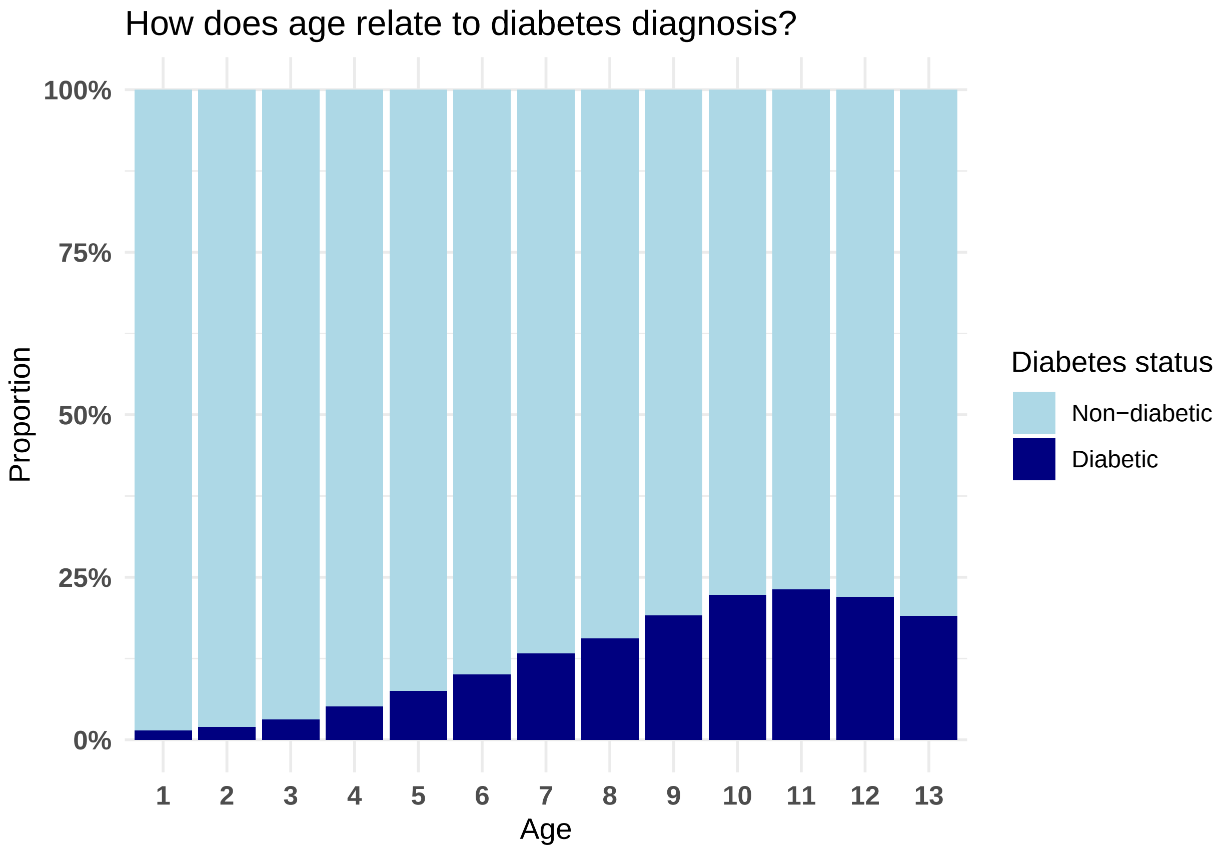
Increasing income corresponds to a reduction in the proportion of diabetic cases

*Age*

```
table(diabetes$Age, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
## 
##      Non-diabetic   Diabetic
##   1    0.98584649 0.01415351
##   2    0.98018120 0.01981880
##   3    0.96867205 0.03132795
##   4    0.94889198 0.05110802
##   5    0.92528490 0.07471510
##   6    0.89924769 0.10075231
##   7    0.86713377 0.13286623
##   8    0.84449252 0.15550748
##   9    0.80857875 0.19142125
##   10   0.77716289 0.22283711
##   11   0.76856272 0.23143728
##   12   0.78002471 0.21997529
##   13   0.80942171 0.19057829
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=Age, fill= Diabetes_binary), position = "fill")+
   scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Age",
       y = "Proportion",
       title = "How does age relate to diabetes diagnosis?")
```



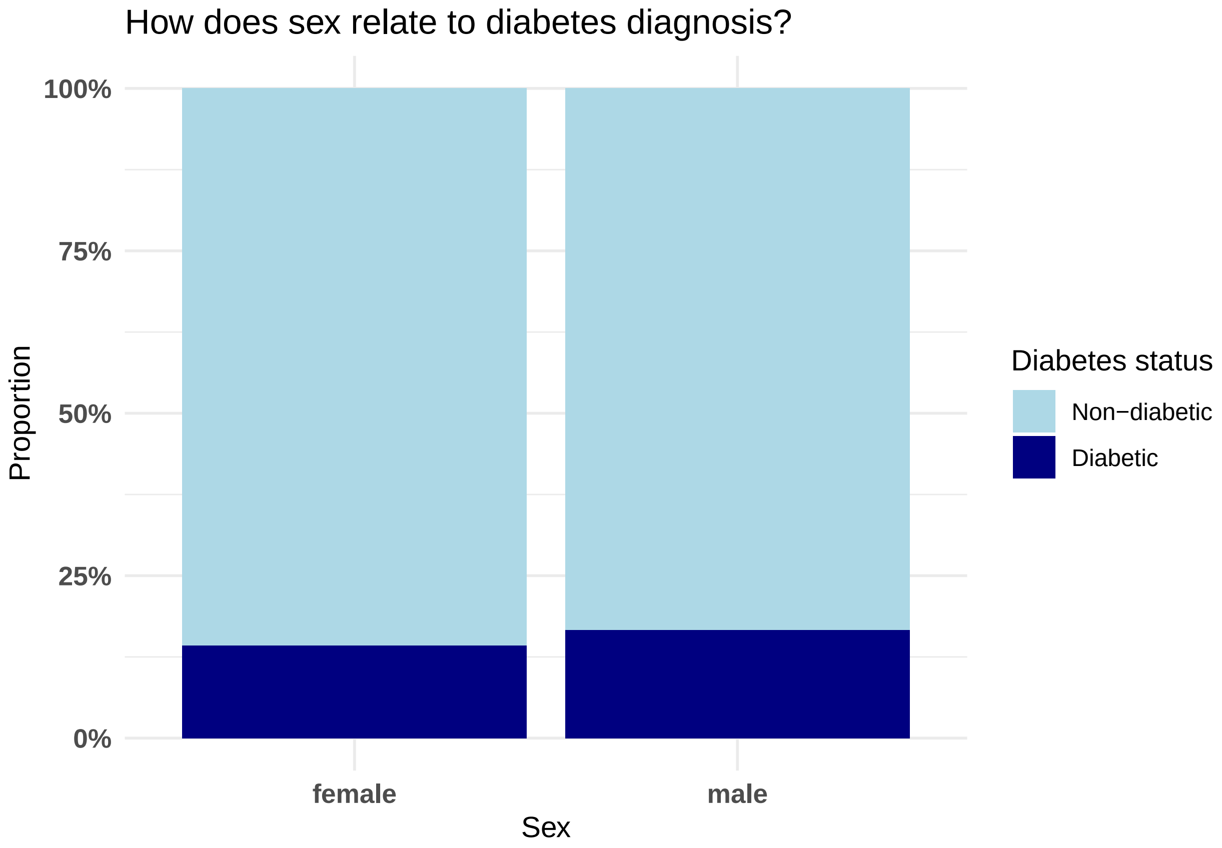
Older individuals tend to have an increased proportion of diabetics in their age bracket than younger ones do.

*Sex*

```
table(diabetes$Sex, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
## 
##           Non-diabetic  Diabetic
##   female     0.8574758 0.1425242
##   male       0.8337419 0.1662581
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=Sex, fill= Diabetes_binary), position = "fill")+
   scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Sex",
       y = "Proportion",
       title = "How does sex relate to diabetes diagnosis?")
```



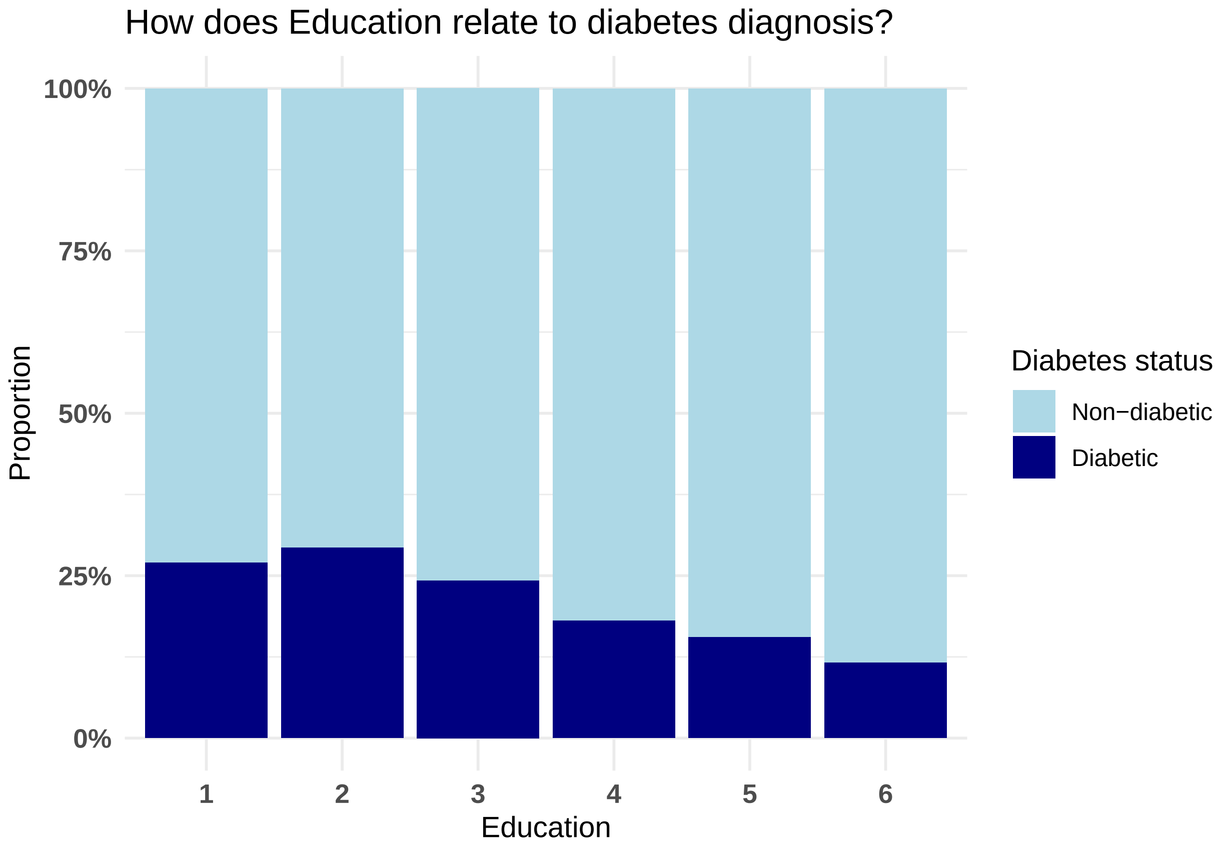
Males had a slight increase in proportion diabetics when compared to females

*Education*

```
table(diabetes$Education, diabetes$Diabetes_binary) %>% prop.table(1)
```

```
##
##       Non-diabetic  Diabetic
##    1    0.7298851 0.2701149
##    2    0.7071782 0.2928218
##    3    0.7574733 0.2425267
##    4    0.8195144 0.1804856
##    5    0.8448167 0.1551833
##    6    0.8840691 0.1159309
```

```
ggplot(data= diabetes)+
  geom_bar(aes(x=Education, fill= Diabetes_binary), position = "fill")+
   scale_fill_manual(values= c( "lightblue", "navy"))+
  guides(fill= guide_legend(title = "Diabetes status"))+
  scale_y_continuous(labels= scales::percent)+
  theme_minimal()+
  theme(axis.text = element_text(size = 10, face= "bold"))+
  labs(x= "Education",
       y = "Proportion",
       title = "How does Education relate to diabetes diagnosis?")
```



As education levels increased, the proportion of individuals with diabetes decreased.

Increasing income and education levels corresponds to a gradual reduction in proportion of diabetics identified in the survey. As expected increasing age corresponds to reduction in proportion of diabetics identified.

## 0.3 CONCLUSION

It is important to state that correlation is not causation.
Diabetes has a lot of risk factors (i.e factors that increase the probability of getting a disease) these factors include lots health status, medical history of the individual, lifestyle choices, genetics and some social factors analysed in this project, so most of the variables analysed in this project would occur in tandem and shoot up the probabilty of being diabetic. Diabetes also is often a comorbidity of other diseases that occur as individuals age and heart diseases.

Although trends in this analysis show some patterns that corresponds to change in the proportion of diabetics, these trends don’t represents significance in predicting the outcome of a diabetic test.