diabetes.R

DELL

2025-06-23

#A study to explore the potential causes and identify key risk factors influencing diabetes outcomes.  
  
#importing relevant libraries  
  
library(readxl)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(dplyr)  
  
#loading the and viewing the dataset R studio  
  
diabetes <- read\_excel("diabetes.xlsx")  
View(diabetes)  
  
# What is the structure of the dataset? How many rows and column?  
str(diabetes)

## tibble [500 × 8] (S3: tbl\_df/tbl/data.frame)  
## $ Age : num [1:500] 56 69 46 32 60 25 78 38 56 75 ...  
## $ Sex : num [1:500] 1 1 2 1 1 1 1 1 2 1 ...  
## $ Residence : num [1:500] 2 2 1 1 2 1 2 2 2 1 ...  
## $ Religion : num [1:500] 1 2 3 3 1 2 1 1 1 1 ...  
## $ Alcohol\_Intake: num [1:500] 1 0 0 0 1 0 0 0 0 1 ...  
## $ Obesity : num [1:500] 0 1 0 0 1 0 0 0 1 0 ...  
## $ BMI : num [1:500] 27.2 26.4 26.6 19.7 29.2 29.7 24.1 24.8 18.9 27.1 ...  
## $ FBG : num [1:500] 5.1 7.4 5.3 6.6 6.4 5.7 5.1 5.2 5.8 5.4 ...

# The dataset has 500 observations or rows and 8 variables  
  
variables <- colnames(diabetes)  
variables

## [1] "Age" "Sex" "Residence" "Religion"   
## [5] "Alcohol\_Intake" "Obesity" "BMI" "FBG"

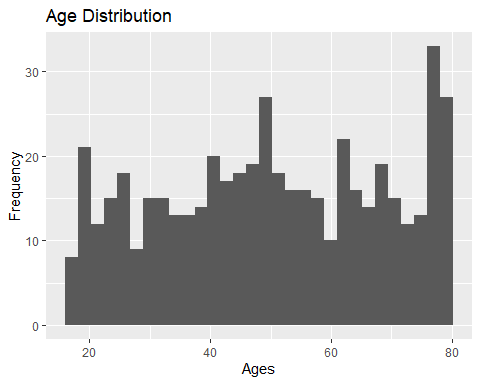
# summary statistic on the Age variable  
summary(diabetes$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 35.00 50.00 50.34 66.00 80.00

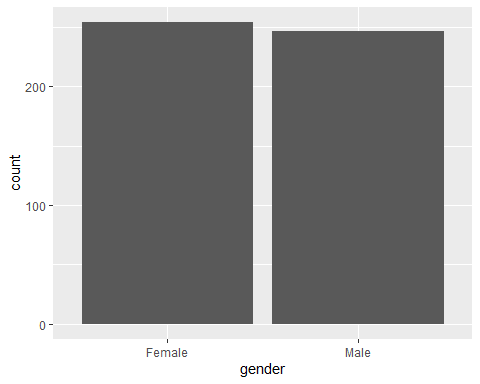
# mode of Age  
mode\_age <- diabetes |>  
 count(Age, sort = TRUE)  
mode\_age

## # A tibble: 63 × 2  
## Age n  
## <dbl> <int>  
## 1 79 18  
## 2 50 16  
## 3 77 16  
## 4 52 14  
## 5 41 13  
## 6 61 12  
## 7 19 11  
## 8 45 11  
## 9 49 11  
## 10 56 11  
## # ℹ 53 more rows

# mode of age distribution is 18   
  
# Histogram for age  
ggplot(diabetes,aes(Age)) +  
 geom\_histogram(bins = 30) +  
 labs(  
 title = "Age Distribution",  
 x = "Ages",  
 y = "Frequency"  
 )



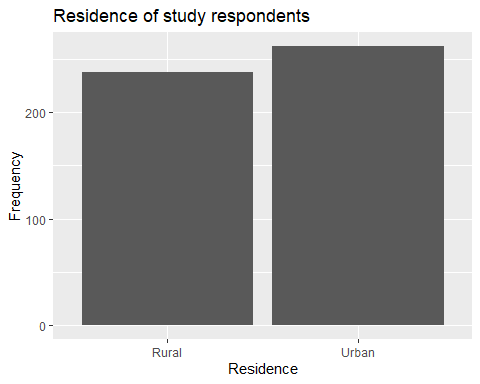
#replacing logical data of Sex with 2 categorical "Male" and "Female"  
diabetes <- diabetes |>  
 mutate(  
 gender = if\_else(Sex == 1, "Female", "Male")  
 )  
#visualization sex with bar plot  
diabetes |>  
 ggplot(aes(gender)) +  
 geom\_bar()



# How many male and female in this study?  
diabetes |>  
 count(gender, sort = TRUE)

## # A tibble: 2 × 2  
## gender n  
## <chr> <int>  
## 1 Female 254  
## 2 Male 246

#adding a new variable for residence  
diabetes <- diabetes |>  
 mutate(  
 residence = if\_else(Residence == 1, "Rural", "Urban")  
 )  
diabetes |>  
 ggplot(aes(residence)) +  
 geom\_bar() +  
 labs( title = "Residence of study respondents",  
 x = "Residence", y = "Frequency")



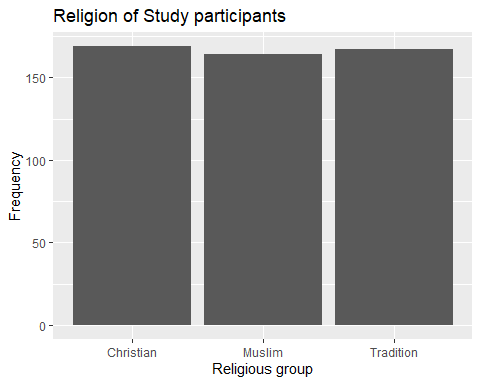
# adding a new categorical variable for religion  
diabetes <- diabetes |>  
 mutate(  
 religion = case\_when(  
 Religion == 1 ~ "Christian",  
 Religion == 2 ~ "Muslim",  
 Religion == 3 ~ "Tradition"  
 )  
 )  
diabetes

## # A tibble: 500 × 11  
## Age Sex Residence Religion Alcohol\_Intake Obesity BMI FBG gender  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>   
## 1 56 1 2 1 1 0 27.2 5.1 Female  
## 2 69 1 2 2 0 1 26.4 7.4 Female  
## 3 46 2 1 3 0 0 26.6 5.3 Male   
## 4 32 1 1 3 0 0 19.7 6.6 Female  
## 5 60 1 2 1 1 1 29.2 6.4 Female  
## 6 25 1 1 2 0 0 29.7 5.7 Female  
## 7 78 1 2 1 0 0 24.1 5.1 Female  
## 8 38 1 2 1 0 0 24.8 5.2 Female  
## 9 56 2 2 1 0 1 18.9 5.8 Male   
## 10 75 1 1 1 1 0 27.1 5.4 Female  
## # ℹ 490 more rows  
## # ℹ 2 more variables: residence <chr>, religion <chr>

diabetes |>  
 select(Religion, religion)

## # A tibble: 500 × 2  
## Religion religion   
## <dbl> <chr>   
## 1 1 Christian  
## 2 2 Muslim   
## 3 3 Tradition  
## 4 3 Tradition  
## 5 1 Christian  
## 6 2 Muslim   
## 7 1 Christian  
## 8 1 Christian  
## 9 1 Christian  
## 10 1 Christian  
## # ℹ 490 more rows

ggplot(diabetes, aes(religion)) +  
 geom\_bar() +  
 labs(  
 title = "Religion of Study participants",  
 x = "Religious group",  
 y = "Frequency"  
 )



diabetes |>  
 count(religion, sort = TRUE)

## # A tibble: 3 × 2  
## religion n  
## <chr> <int>  
## 1 Christian 169  
## 2 Tradition 167  
## 3 Muslim 164

#categorical variable for alcohol intake, obesity, and BMI  
diabetes <- diabetes |>  
 mutate(  
 alcohol = if\_else(Alcohol\_Intake == 1, "Yes", "No"),  
 obese = if\_else(Obesity == 1, "Yes", "No"),  
 bmi = case\_when(  
 BMI < 18.5 ~ "undwerweight",  
 BMI >= 18.5 & BMI <= 24.9 ~ "healthy\_weight",  
 BMI >= 25 & BMI <= 29.9 ~ "overweight",  
 BMI >= 30 & BMI <= 39.9 ~ "obese",  
 BMI >= 40 ~ "severely\_obese"  
 )  
 )  
  
diabetes |>  
 count(bmi, sort = TRUE)

## # A tibble: 5 × 2  
## bmi n  
## <chr> <int>  
## 1 healthy\_weight 231  
## 2 overweight 195  
## 3 obese 53  
## 4 undwerweight 20  
## 5 severely\_obese 1

diabetes |>  
 select(BMI, bmi)

## # A tibble: 500 × 2  
## BMI bmi   
## <dbl> <chr>   
## 1 27.2 overweight   
## 2 26.4 overweight   
## 3 26.6 overweight   
## 4 19.7 healthy\_weight  
## 5 29.2 overweight   
## 6 29.7 overweight   
## 7 24.1 healthy\_weight  
## 8 24.8 healthy\_weight  
## 9 18.9 healthy\_weight  
## 10 27.1 overweight   
## # ℹ 490 more rows

diabetes |>  
 select(BMI : bmi) |>  
 filter(BMI < 18.5)

## # A tibble: 20 × 8  
## BMI FBG gender residence religion alcohol obese bmi   
## <dbl> <dbl> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 17.3 7.2 Female Urban Muslim No No undwerweight  
## 2 16.1 4.4 Female Urban Tradition No Yes undwerweight  
## 3 17.5 4.7 Female Urban Christian Yes No undwerweight  
## 4 15 4.8 Male Rural Christian Yes No undwerweight  
## 5 16 6.1 Male Urban Tradition No No undwerweight  
## 6 17.5 5.3 Female Urban Tradition No Yes undwerweight  
## 7 16.2 5.6 Male Urban Tradition No No undwerweight  
## 8 17.1 6.4 Male Urban Christian No Yes undwerweight  
## 9 16.6 8 Male Urban Muslim No Yes undwerweight  
## 10 15 5.6 Female Rural Muslim No No undwerweight  
## 11 17.9 5.1 Female Urban Muslim No No undwerweight  
## 12 17.9 5.9 Female Urban Muslim No No undwerweight  
## 13 17.3 6.8 Female Urban Tradition No No undwerweight  
## 14 18.2 6.7 Female Urban Tradition No Yes undwerweight  
## 15 16.7 6.8 Male Urban Muslim No Yes undwerweight  
## 16 16 7.3 Female Rural Tradition No No undwerweight  
## 17 16.5 6 Female Rural Tradition No No undwerweight  
## 18 15 7.3 Female Urban Muslim No Yes undwerweight  
## 19 16.5 6.7 Female Urban Christian No Yes undwerweight  
## 20 16.7 7.5 Female Rural Tradition No Yes undwerweight