DIABETES

Mehrnaz Bastani, Laura, Ngo, Janelle Ah Kit, Helene Gao

TABLE OF CONTENTS

DESCRIBING THE DATASET

Diabetes risk assessment and prediction - collected health-related attributes
Source: National Institute of Diabetes and Digestive and Kidney Diseases
Goal: early diagnosis and personalized treatment strategies

2,768 Observations

- 10 variables

Variables

- Id, Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome

RESEARCH QUESTION AND HYPOTHESIS

- Which demographics are more likely to have diabetes?
- Is BMI a good indicator of an individual's likelihood of developing diabetes?
- How do BMI and personal background (demographics) jointly impact diabetes prediction?

Hypothesis: we expect there to be a significant association between BMI and diabetes, with higher BMI values being indicative of increased diabetes risk.

DATA CLEANING AND PREPROCESSING

- Sort the BMI into defined ranges: underweight, healthy, overweight, obese
- Use weights as there are more data points in certain BMI ranges than others

Body Mass Index

```r
# BMI = kg / m^2
# underweight = 18.5 or below
# healthy = 18.5 - 24.9
# overweight = 25 - 29.9
# obesity = 30+

# Define BMI ranges
underweight <- as.numeric(c(0, 18.5))
healthy <- as.numeric(c(18.5, 24.9))
overweight <- as.numeric(c(25, 29.9))
obese <- as.numeric(c(30, max(diabetes$BMI)))

# Create a function to label BMI values based on the defined ranges
labels <- function(bmi) {
  if (bmi >= underweight[1] && bmi <= underweight[2]) {
    return("Underweight")
  } else if (bmi >= healthy[1] && bmi <= healthy[2]) {
    return("Healthy")
  } else if (bmi >= overweight[1] && bmi <= overweight[2]) {
    return("Overweight")
  } else if (bmi >= obese[1] && bmi <= obese[2]) {
    return("Obese")
  } else {
    return("Unknown")
  }
}

# Apply the labels function to create the bmi.ranges variable
diabetes$bmi.ranges <- sapply(diabetes$BMI, labels)

# Convert bmi.ranges into a factor variable with all levels
diabetes$bmi.ranges <- factor(diabetes$bmi.ranges, levels = c("Underweight", "Healthy", "Overweight", "

# Check missing/unexpected outputs
table(diabetes$bmi.ranges)
```

DATA CLEANING AND PROCESSING

Sort the ages into bins of ranges :
Young adult, middle aged, senior

```r
library(ggplot2)

# Define age ranges
young.adults <- c(18, 35)
middle.aged <- c(36, 55)
senior <- c(55, max(diabetes$Age))

# Create a function to label age values based on the defined ranges
labels2 <- function(ages) {
  if (ages >= young.adults[1] && ages <= young.adults[2]) {
    return("Young Adult")
  } else if (ages >= middle.aged[1] && ages <= middle.aged[2]) {
    return("Middle Aged")
  } else if (ages >= senior[1] && ages <= senior[2]) {
    return("Senior")
  } else {
    return("Unknown")
  }
}

# Apply the labels function to create the age.ranges variable
diabetes$age.ranges <- sapply(diabetes$Age, labels2)

# Convert age.ranges into a factor variable with all levels
diabetes$age.ranges <- factor(diabetes$age.ranges, levels = c("Young Adult", "Middle Aged", "Senior"))

# Check missing/unexpected outputs
table(diabetes$age.ranges)
```

```
##
## Young Adult Middle Aged      Senior
##        1797         794         177
```

```r
# Fit linear regression model with weighted sample sizes
weights_age <- 1 / table(diabetes$age.ranges)
weight_vector_age <- weights_age[diabetes$age.ranges]
weighted_model_age <- lm(Outcome ~ age.ranges, data = diabetes, weights = weight_vector_age)
summary(weighted_model_age)
```

DATA CLEANING AND PROCESSING

Sort the glucose into bins of ranges : Normal, Impaired Fasting, Faster Hyperglycemia, unknown

Glucose

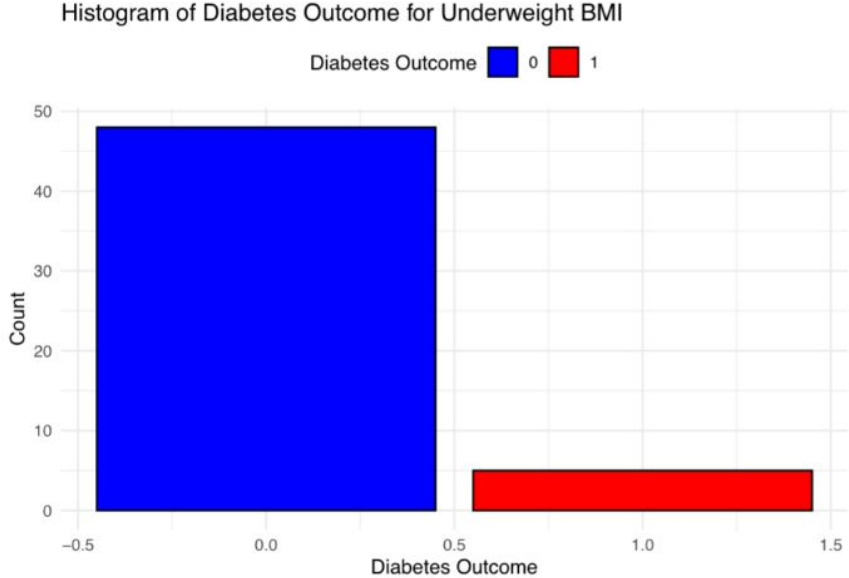
```
normal.fasting <- c(0, 100)
impaired.fasting <- c(100,125)
fasting.hyperglycemia<- c(126, max(diabetes$Glucose))

# Create a function to label BMI values based on the defined ranges
labels3 <- function(glucose) {
  if (glucose >= normal.fasting[1] && glucose <= normal.fasting[2]) {
    return("Normal")
  } else if (glucose >= impaired.fasting[1] && glucose <= impaired.fasting[2]) {
    return("Impaired Fasting")
  } else if (glucose >= fasting.hyperglycemia[1] && glucose <= fasting.hyperglycemia[2]) {
    return("Fasting Hyperglycemia")
  } else {
    return("Unknown")
  }
}

diabetes$glucose.ranges <- sapply(diabetes$Glucose, labels3)
diabetes$glucose.ranges <- factor(diabetes$glucose.ranges, levels = c("Normal", "Impaired Fas

table(diabetes$glucose.ranges) # check missing/unexpected outputs
```

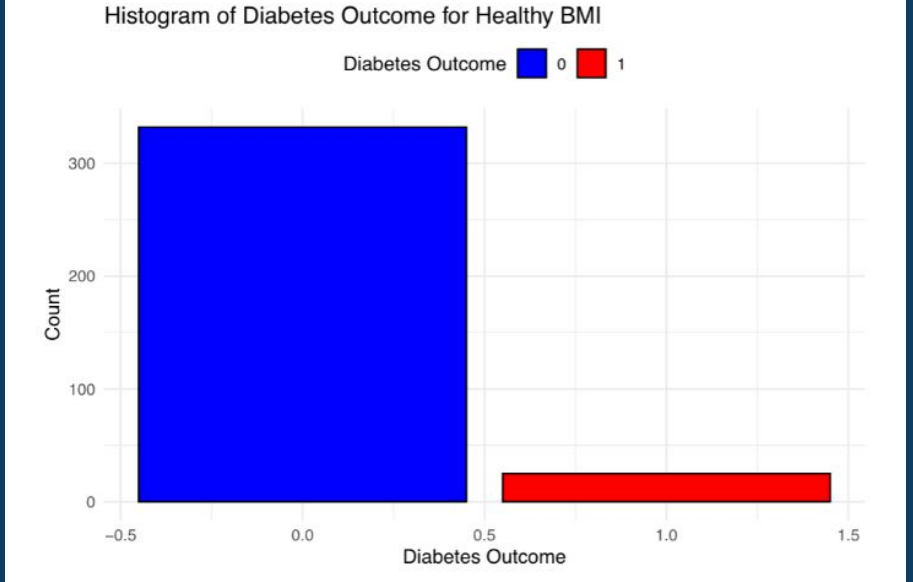
```
##
##                Normal      Impaired Fasting Fasting Hyperglycemia
##                   778                   904                  1086
```

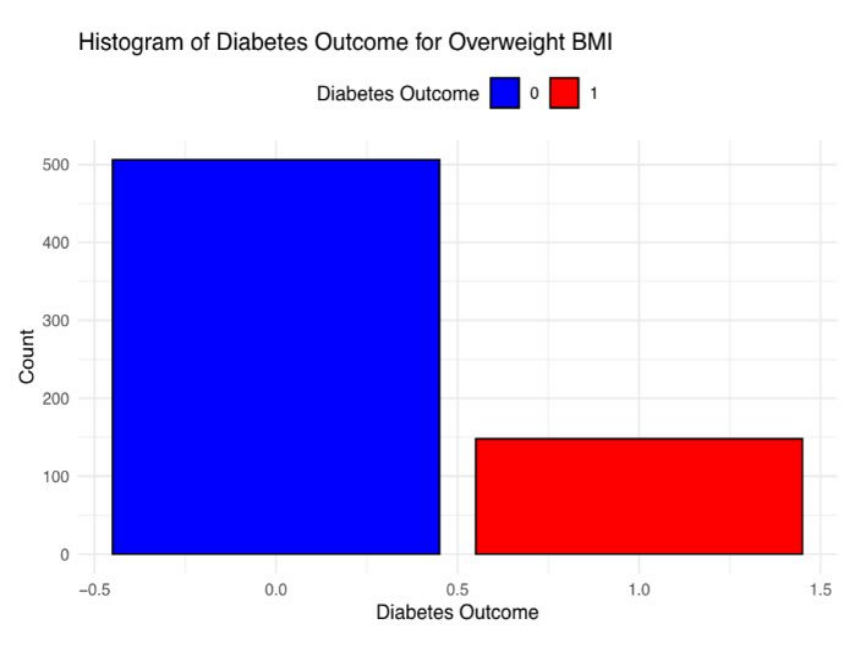
```
model3 <- lm(Outcome ~ glucose.ranges, data = diabetes)
summary(model3)
```

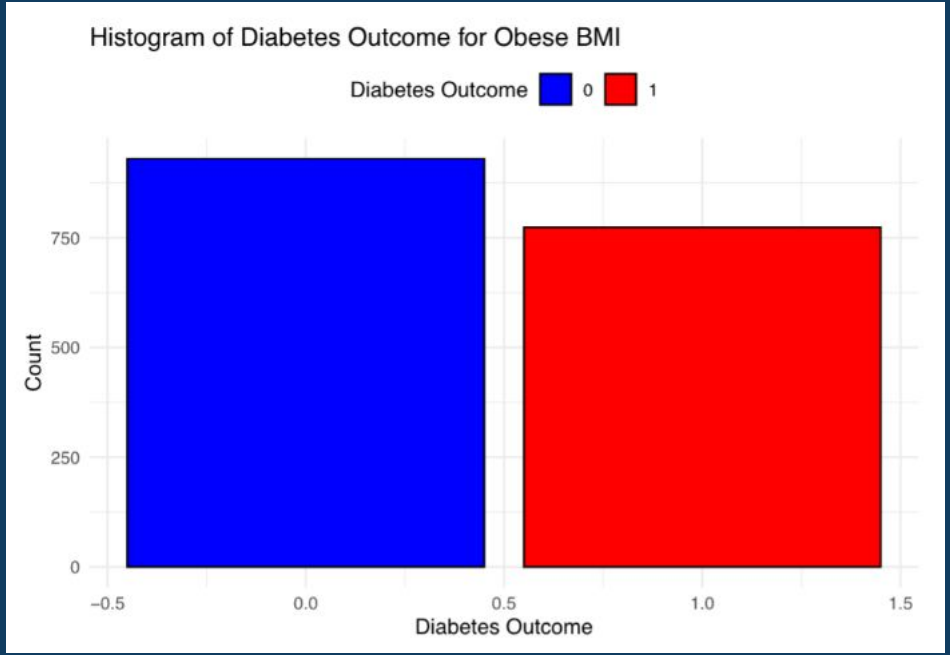
```
##
## Call:
## lm(formula = Outcome ~ glucose.ranges, data = diabetes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58287 -0.27434 -0.09126  0.41713  0.90874
```

OBSERVATIONS

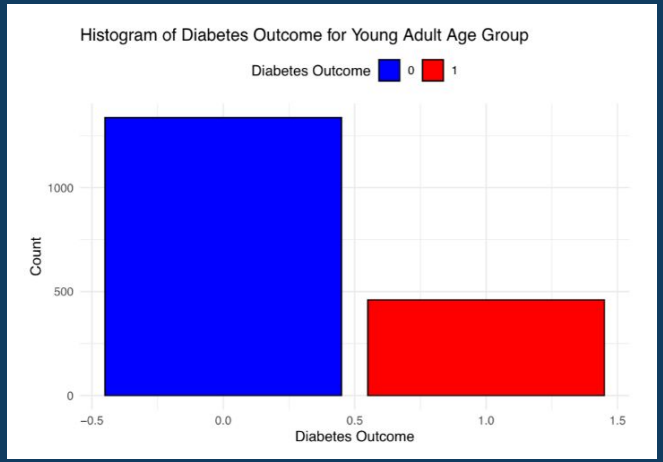

Histogram of Diabetes Outcome for Underweight BMI
Diabetes Outcome
0
1
Count
50
40
30
20
10
0
−0.5
0.0
0.5
1.0
1.5
Diabetes Outcome

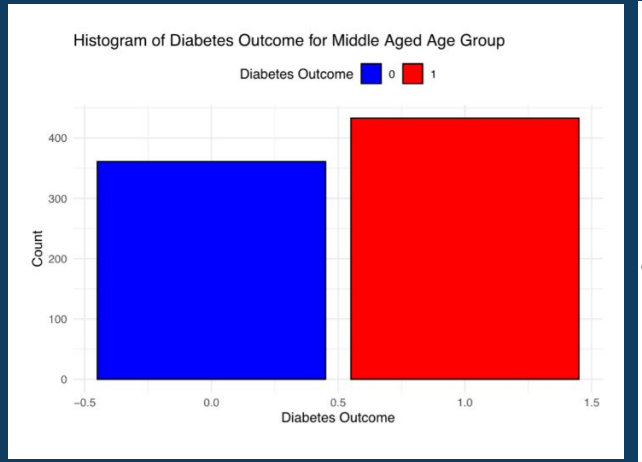

Histogram of Diabetes Outcome for Healthy BMI
Diabetes Outcome
0
1
Count
300
200
100
0
−0.5
0.0
0.5
1.0
1.5
Diabetes Outcome

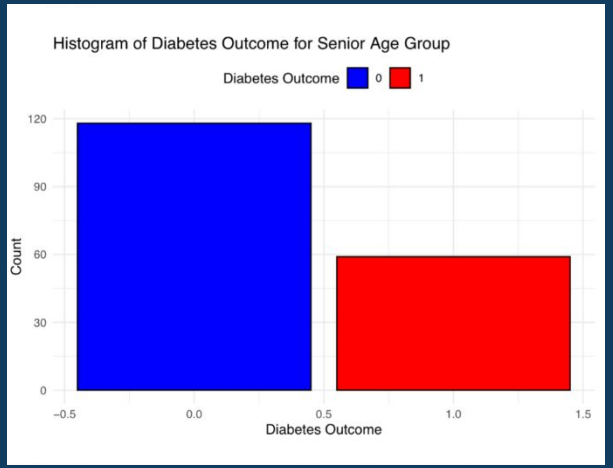




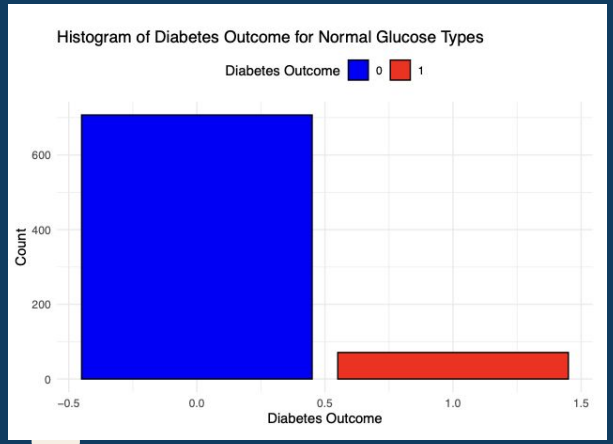
- Higher prevalence of diabetes for those who are obese

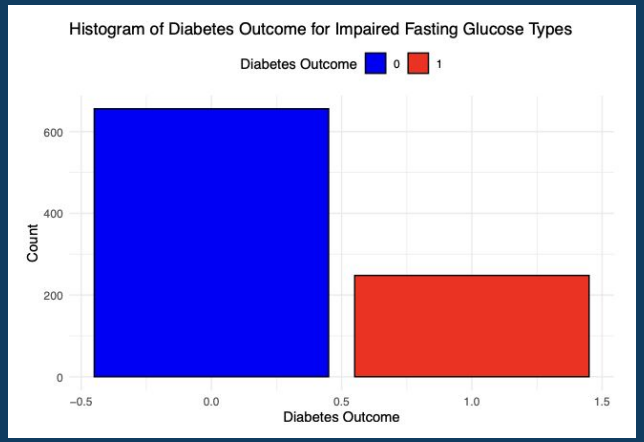


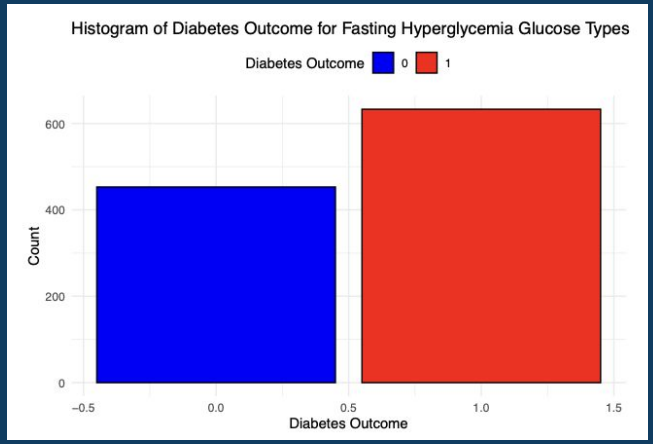




- Higher prevalence of diabetes for those who are middle-aged







- Higher prevalence of diabetes outcomes for fasting hyperglycemia glucose types

06

CONCLUSIONS

CONCLUSIONS

- Those who are obese are more at risk
- Middle Aged group is more at risk for diabetes
- Higher prevalence of diabetes outcomes for fasting hyperglycemia glucose types

WHAT DOES THIS MEAN?

- Pre-existing health factors → diabetes
- Predicting diabetes early-on

THANKS!

Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**