Public Health Risk Factors and Diabetes Status

Ronak Fathi (Mentor)

2025-07-14

Project Overview

This R Markdown file presents a statistical analysis of a simulated dataset exploring the association between lifestyle and physiological risk factors and diabetes status in a sample of 100 patients. The study investigates whether variables such as **age**, **BMI**, **smoking status**, and **blood pressure** are associated with **diabetes status**.

This project is a **simulation** designed for instructional purposes, reflecting realistic patterns observed in public health data.

Introduction

Diabetes is a chronic condition influenced by lifestyle and physiological factors. Understanding how factors such as BMI, blood pressure, smoking, and age relate to diabetes risk is important for prevention and early intervention efforts. This report examines these associations using simulated data.

Method

Design

- **Type**: Observational, cross-sectional
- **Sample size**: 100 patients
- **Variables**:
  - **age**: Patient age in years
  - **BMI**: Body Mass Index
  - **smoking_status**: Never, Former, Current
  - **blood_pressure**: Normal, Elevated, High
  - **diabetes_status**: 0 = No diabetes, 1 = Has diabetes

Load Packages

Load Dataset

```
setwd("C:/Users/0&1/OneDrive/Documents/Student-Projects-Portfolio/Diabetes-Prediction")
data <- read.csv("health.csv")
kable(head(data))
```

age	BMI	smoking_status	blood_pressure	diabetes_status
58	30.1	Former smoker	Normal	1
71	37.6	Former smoker	Normal	1

age	BMI	smoking_status	blood_pressure	diabetes_status
48	28.1	Former smoker	Elevated	0
34	33.6	Current smoker	Normal	0
62	28.8	Former smoker	High	0
27	27.3	Former smoker	High	0

Descriptive Statistics

```
summary(data)
```

```
##       age              BMI        smoking_status     blood_pressure
##  Min.   :21.00   Min.   :17.00   Length:100         Length:100
##  1st Qu.:34.00   1st Qu.:24.43   Class :character   Class :character
##  Median :48.00   Median :26.90   Mode  :character   Mode  :character
##  Mean   :49.58   Mean   :27.12
##  3rd Qu.:66.00   3rd Qu.:29.82
##  Max.   :79.00   Max.   :37.60
##  diabetes_status
##  Min.   :0.00
##  1st Qu.:0.00
##  Median :0.00
##  Mean   :0.32
##  3rd Qu.:1.00
##  Max.   :1.00
```

```
psych::describe(data %>% select(age, BMI))
```

```
##     vars   n  mean    sd median trimmed   mad min  max range skew kurtosis  se
## age    1 100 49.58 18.03   48.0   49.51 22.24  21 79.0  58.0 0.03    -1.30 1.8
## BMI    2 100 27.12  3.97   26.9   27.12  4.30  17 37.6  20.6 0.00    -0.21 0.4
```

```
table(data$smoking_status)
```

```
##
## Current smoker  Former smoker     Non-smoker
##             30             34             36
```

```
table(data$blood_pressure)
```

```
##
## Elevated     High   Normal
##       39       33       28
```

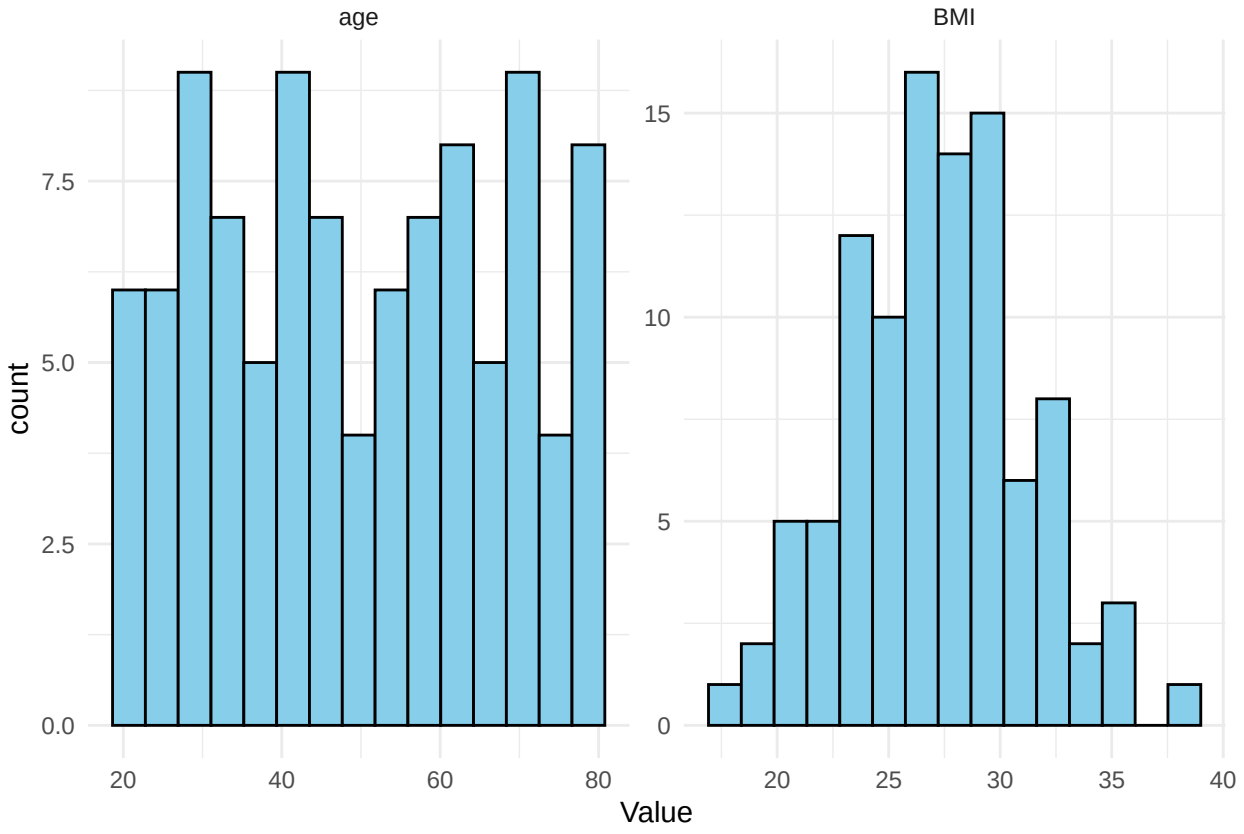
```
table(data$diabetes_status)
```

```
##
##  0  1
## 68 32
```

Visualization

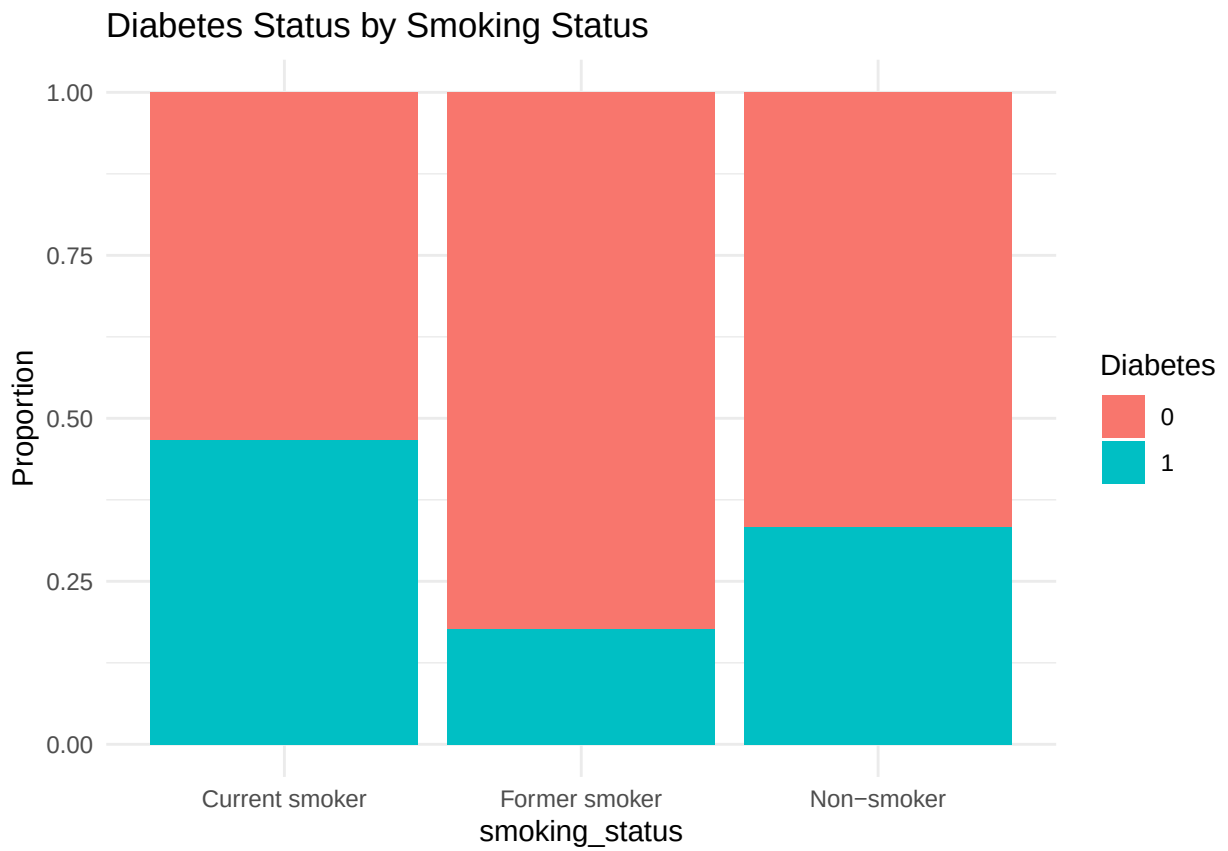
Distributions

```r
data %>%
  pivot_longer(cols = c(age, BMI), names_to = "Variable", values_to = "Value") %>%
  ggplot(aes(x = Value)) +
  geom_histogram(bins = 15, fill = "skyblue", color = "black") +
  facet_wrap(~Variable, scales = "free") +
  theme_minimal()
```

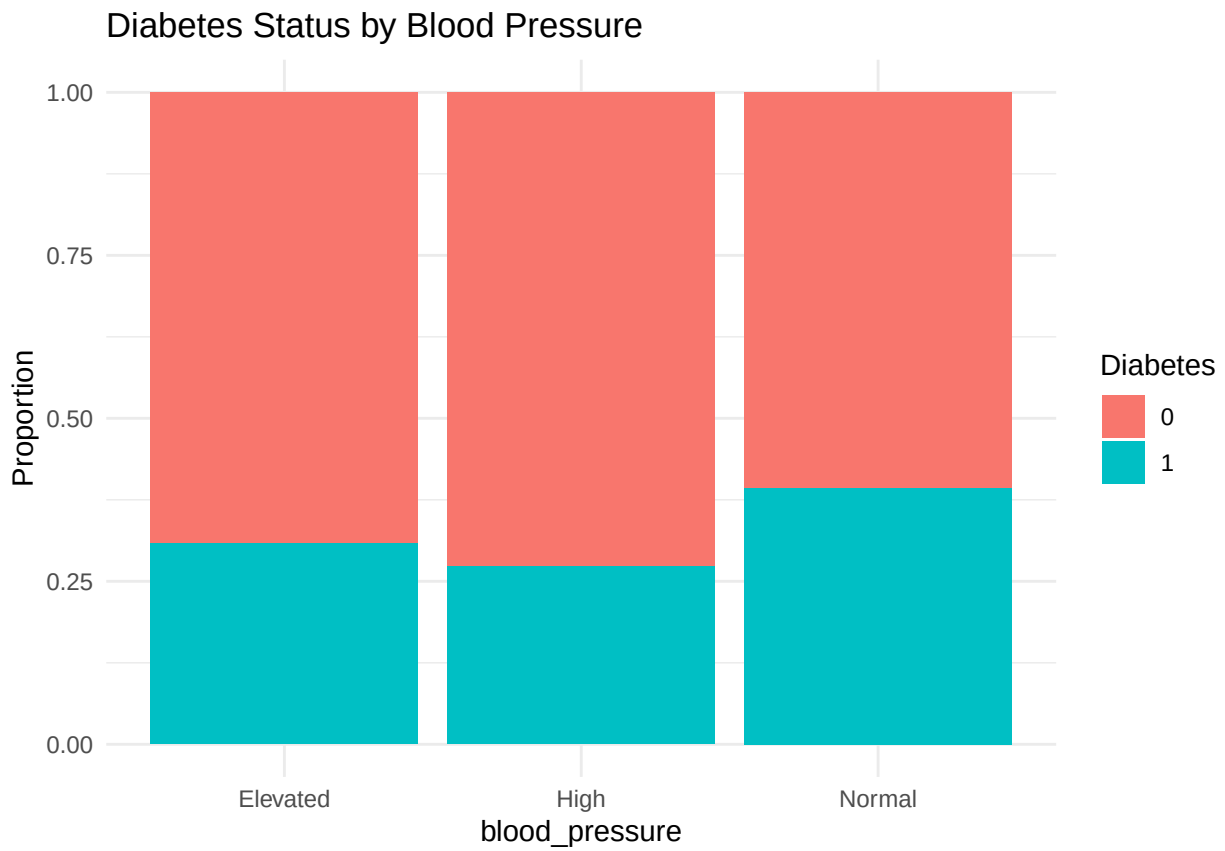


Bar plots for categorical variables

```r
ggplot(data, aes(x = smoking_status, fill = factor(diabetes_status))) +
  geom_bar(position = "fill") +
  labs(title = "Diabetes Status by Smoking Status", y = "Proportion", fill = "Diabetes") +
  theme_minimal()
```




```
ggplot(data, aes(x = blood_pressure, fill = factor(diabetes_status))) +
  geom_bar(position = "fill") +
  labs(title = "Diabetes Status by Blood Pressure", y = "Proportion", fill = "Diabetes") +
  theme_minimal()
```



Inferential Statistics

Comparing Age and BMI by Diabetes Status

```
t.test(age ~ diabetes_status, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  age by diabetes_status
## t = 1.2346, df = 59.369, p-value = 0.2219
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.981991 12.592285
## sample estimates:
## mean in group 0 mean in group 1
##        51.11765        46.31250
```

```
t.test(BMI ~ diabetes_status, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  BMI by diabetes_status
## t = -1.0646, df = 48.567, p-value = 0.2923
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
```

```
## 95 percent confidence interval:
##  -2.8795921  0.8854745
## sample estimates:
## mean in group 0 mean in group 1
##        26.80294        27.80000
```

Association with Categorical Variables

```
chisq.test(table(data$smoking_status, data$diabetes_status))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(data$smoking_status, data$diabetes_status)
## X-squared = 6.214, df = 2, p-value = 0.04474
```

```
chisq.test(table(data$blood_pressure, data$diabetes_status))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(data$blood_pressure, data$diabetes_status)
## X-squared = 1.0491, df = 2, p-value = 0.5918
```

Logistic Regression

We use logistic regression to model the probability of diabetes status using all predictors.

```
model <- glm(diabetes_status ~ age + BMI + smoking_status + blood_pressure, data = data, family = "bino
summary(model)
```

```
##
## Call:
## glm(formula = diabetes_status ~ age + BMI + smoking_status +
##     blood_pressure, family = "binomial", data = data)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -1.04313    1.71398  -0.609   0.5428
## age                          -0.01892    0.01341  -1.411   0.1584
## BMI                           0.06760    0.05839   1.158   0.2469
## smoking_statusFormer smoker -1.50736    0.60589  -2.488   0.0129 *
## smoking_statusNon-smoker    -0.67061    0.53456  -1.255   0.2097
## blood_pressureHigh          -0.32956    0.55584  -0.593   0.5532
## blood_pressureNormal         0.57807    0.56983   1.014   0.3104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 125.37  on 99  degrees of freedom
## Residual deviance: 113.88  on 93  degrees of freedom
## AIC: 127.88
##
## Number of Fisher Scoring iterations: 4
```

Summary of Results

- Sample: 100 patients, with 32% having diabetes.
- Descriptive statistics:
  - Age: Mean = 49.6 years, SD = 18.0
  - BMI: Mean = 27.1, SD = 4.0
  - Smoking status: 30% current smokers, 34% former smokers, 36% non-smokers
  - Blood pressure: 33% normal, 39% elevated, 28% high
- Inferential statistics:
  - T-tests:
    - No significant difference in age between diabetic and non-diabetic groups (p = 0.22)
    - No significant difference in BMI between groups (p = 0.29)
  - Chi-square tests:
    - Smoking status was significantly associated with diabetes status (^{2} = 6.21, p = 0.045)
    - Blood pressure was not significantly associated with diabetes status (^{2} = 1.05, p = 0.59)
- Logistic Regression:
  - Only the variable "Former smoker" showed a statistically significant association with diabetes (= -1.51, p = 0.0129), indicating lower odds of diabetes compared to current smokers.
  - Age, BMI, and blood pressure levels were not significant predictors in the model.
  - Model AIC = 127.88; residual deviance = 113.88 (on 93 df)

Discussion

This simulation study explored associations between public health risk factors and diabetes status. Although both age and BMI are traditionally recognized risk factors for diabetes, this simulated dataset did not reveal statistically significant differences in these variables between diabetic and non-diabetic individuals. This may reflect limitations in sample size, data variability, or the simulated nature of the dataset.

Interestingly, smoking status showed a significant association with diabetes. Specifically, former smokers had significantly lower odds of having diabetes compared to current smokers. This aligns with real-world findings that smoking is a modifiable risk factor for diabetes, and cessation may reduce risk over time.

In contrast, blood pressure levels were not significantly related to diabetes in this analysis, although this might be due to the categorical simplification of blood pressure or the modest sample size.

The logistic regression model suggests that, when adjusting for other factors, only smoking status (particularly being a former smoker) remained a significant predictor. This may reflect the influence of health-seeking behavior or lifestyle changes post-smoking cessation.

Conclusion

This simulated analysis suggests that among the examined factors, smoking status—particularly being a former smoker—was most strongly associated with diabetes status. While age, BMI, and blood pressure are known to affect diabetes risk in real populations, they did not reach significance in this small, simulated dataset.

These results highlight the importance of tobacco cessation in diabetes prevention efforts and emphasize the need for larger and more detailed datasets to robustly capture the multifactorial nature of diabetes risk.

---

*This report is based on a previously conducted student-led experiment and compiled by **Ronak Fathi**, mentor and research supervisor. The dataset is simulated for instructional purposes.*