# Heart Attack Risk Analysis

Linger Ge

## Load Library

```
library(tidyverse)
```

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

```
library(broom)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

## PREAMBLE defines variables

```
myData<-read.csv("heart_attack_Fixed.csv")
head(myData, 5)
```

```
##   Age Gender Obesity Smoking_Status Alcohol_Consumption Rural_or_Urban
## 1  55   Male     Yes     Non-Smoker                 Yes          Rural
## 2  66 Female      No         Smoker                  No          Urban
## 3  69 Female      No         Smoker                  No          Rural
## 4  45 Female      No         Smoker                 Yes          Rural
```

```
## 5  39 Female      No         Smoker                  No          Urban
##   Physical_Activity Blood_Pressure Heart_Attack
## 1              High       158.6522          Yes
## 2              High       166.3913           No
## 3              High       172.8406          Yes
## 4               Low       143.8188           No
## 5            Medium       130.2754           No
```

```
x1name <- "Smoking_Status"
x2name <- "Alcohol_Consumption"
y1name <- "Blood_Pressure"
y2name <- "Heart_Attack"
```

## Preprocess data

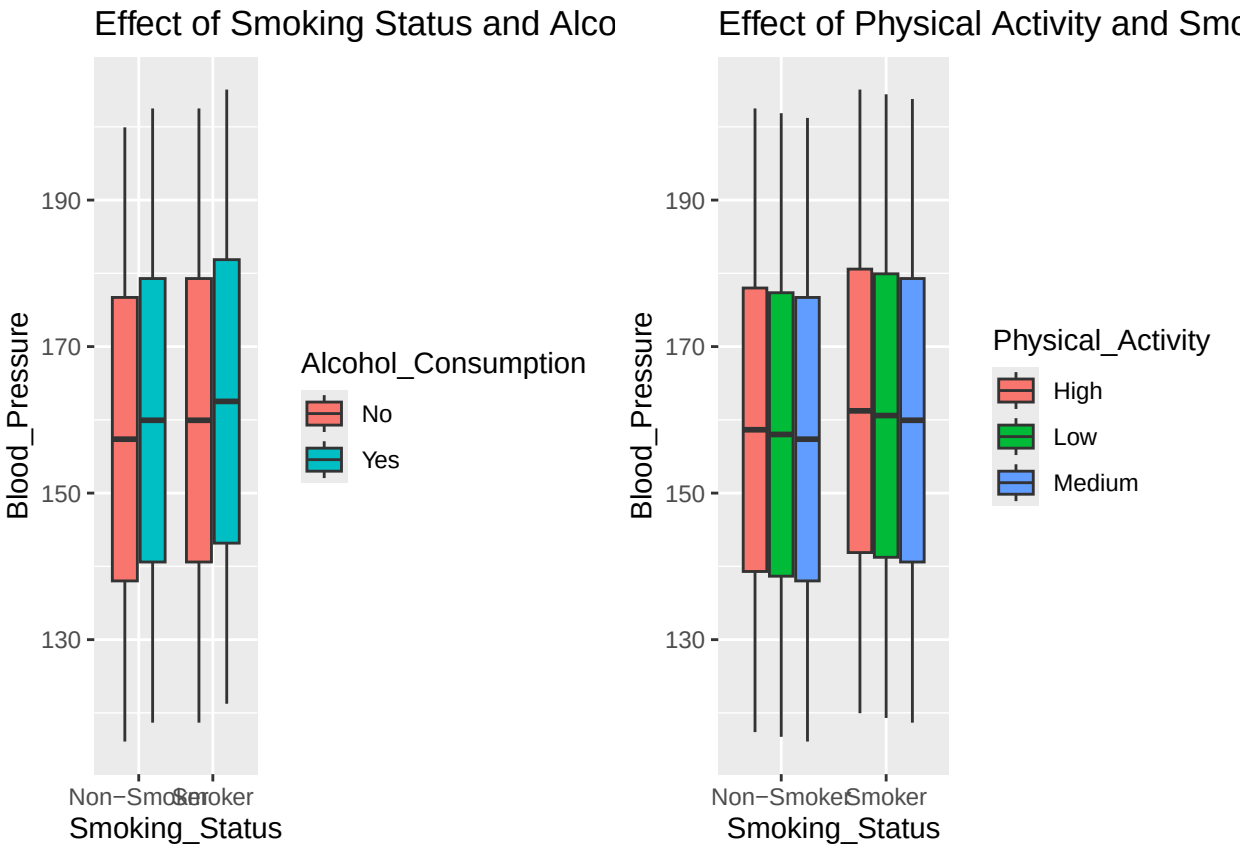
```
myData <- myData %>%
  mutate(
    Gender = factor(Gender),
    Obesity = factor(Obesity),
    Smoking_Status = factor(Smoking_Status),
    Alcohol_Consumption = factor(Alcohol_Consumption),
    Rural_or_Urban = factor(Rural_or_Urban),
    Physical_Activity = factor(Physical_Activity),
    Heart_Attack = factor(Heart_Attack, levels = c("No", "Yes"))
  )
```

## Question 1

```
anova_model <- aov(Blood_Pressure ~ Smoking_Status * Alcohol_Consumption * Physical_Activity, data = myD

summary(anova_model)
```

```
##                                                        Df    Sum Sq Mean Sq
## Smoking_Status                                          1    387258  387258
## Alcohol_Consumption                                     1    333434  333434
## Physical_Activity                                       2     69064   34532
## Smoking_Status:Alcohol_Consumption                      1       128     128
## Smoking_Status:Physical_Activity                        2       162      81
## Alcohol_Consumption:Physical_Activity                   2      1127     563
## Smoking_Status:Alcohol_Consumption:Physical_Activity    2       659     329
## Residuals                                          239254 120935911     505
##                                                    F value Pr(>F)
## Smoking_Status                                     766.133 <2e-16 ***
## Alcohol_Consumption                                659.651 <2e-16 ***
## Physical_Activity                                   68.316 <2e-16 ***
## Smoking_Status:Alcohol_Consumption                   0.254  0.614
## Smoking_Status:Physical_Activity                     0.161  0.852
## Alcohol_Consumption:Physical_Activity                1.115  0.328
## Smoking_Status:Alcohol_Consumption:Physical_Activity  0.652  0.521
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
p1 <- ggplot(myData, aes(x = Smoking_Status, y = Blood_Pressure, fill = Alcohol_Consumption)) +
  geom_boxplot() +
  labs(title = "Effect of Smoking Status and Alcohol Consumption on Blood Pressure")

p2 <- ggplot(myData, aes(x = Smoking_Status, y = Blood_Pressure, fill = Physical_Activity)) +
  geom_boxplot() +
  labs(title = "Effect of Physical Activity and Smoking Status on Blood Pressure")

grid.arrange(p1, p2, ncol = 2)
```




The ANOVA results revealed that smoking status, alcohol consumption, and physical activity all have statistically significant effects on blood pressure ($p < 2e\text{-}16$ for each). Smokers consistently had higher blood pressure than non-smokers, which aligns with the patterns seen in the box plots. Alcohol consumption was also associated with higher blood pressure, as shown in the plots. Physical activity significantly influenced blood pressure, with visual indications suggesting lower blood pressure in individuals with higher activity levels.

However, contrary to the initial visual interpretation, the ANOVA results indicated no significant interaction effects between these variables. This suggests that the effects of smoking, alcohol, and physical activity on blood pressure are independent of each other. While the box plots clearly illustrate the main effects of these variables, the absence of significant interaction terms suggests that the observed patterns hold across different levels of the other variables.

## Question 2

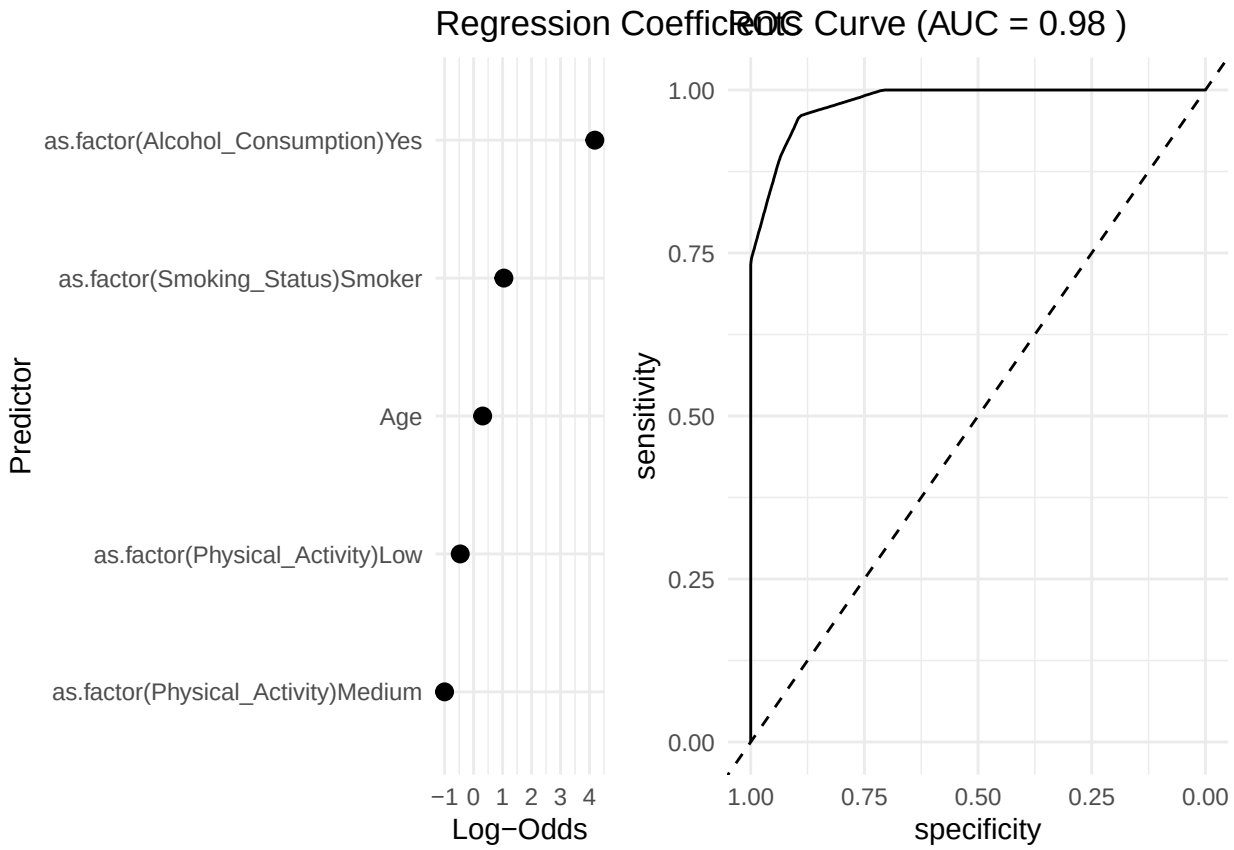
```r
logit_model <- glm(Heart_Attack ~ Age + as.factor(Smoking_Status) + as.factor(Alcohol_Consumption) + as
                   data = myData, family = "binomial")
```

```r
coef_table <- tidy(logit_model) %>% filter(term != "(Intercept)")

coef_table <- coef_table %>%
  mutate(Odds_Ratio = exp(estimate),
         Lower_CI = exp(estimate - 1.96 * std.error),
         Upper_CI = exp(estimate + 1.96 * std.error))

coef_plot <- coef_table %>%
  ggplot(aes(x = reorder(term, estimate), y = estimate)) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
  geom_pointrange(aes(ymin = estimate - 1.96*std.error, ymax = estimate + 1.96*std.error)) +
  coord_flip() +
  labs(title = "Regression Coefficients", x = "Predictor", y = "Log-Odds") +
  theme_minimal()

roc_obj <- roc(myData[[y2name]], predict(logit_model, type = "response"))
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```r
auc_val <- auc(roc_obj)

roc_plot <- ggroc(roc_obj) +
  geom_abline(slope = 1, intercept = 1, linetype = "dashed") +
  labs(title = paste("ROC Curve (AUC =", round(auc_val, 2), ")")) +
  theme_minimal()

grid.arrange(coef_plot, roc_plot, ncol = 2)
```




```
print(coef_table)
```

```
## # A tibble: 5 x 8
##   term        estimate std.error statistic  p.value Odds_Ratio Lower_CI Upper_CI
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>      <dbl>    <dbl>    <dbl>
## 1 Age            0.312   0.00154     203.  0             1.37     1.36     1.37
## 2 as.factor(~    1.04    0.0185       56.5 0             2.84     2.74     2.94
## 3 as.factor(~    4.18    0.0282      148.  0            65.5     62.0     69.2
## 4 as.factor(~   -0.462   0.0219      -21.1 6.02e-99      0.630    0.603    0.657
## 5 as.factor(~   -1.00    0.0223      -44.9 0             0.367    0.351    0.383
```

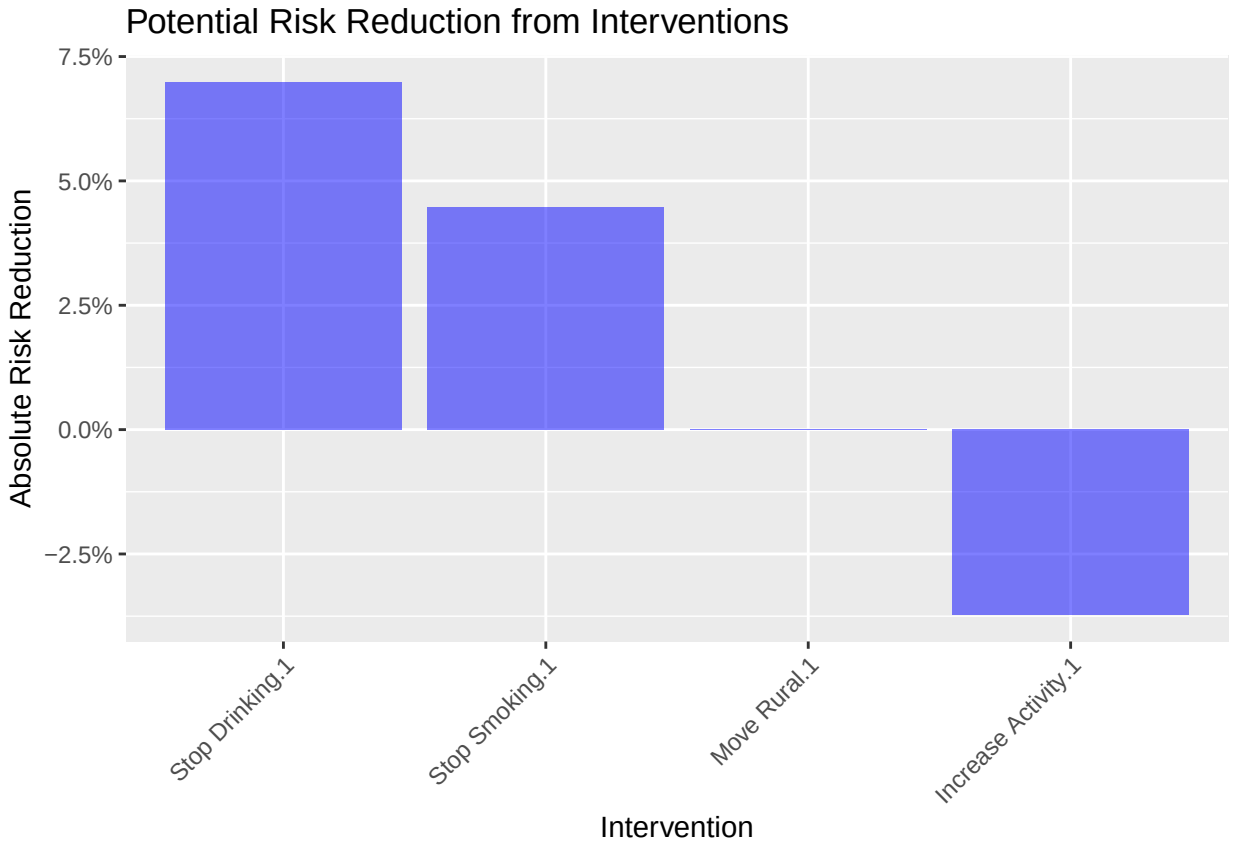
I built a logistic regression model to predict an outcome using demographic factors like age, smoking status, alcohol consumption, and physical activity. The results showed that smoking, alcohol consumption, and age increase the likelihood of the outcome, while low or medium physical activity decrease the likelihood. Smoking had the biggest impact, followed by alcohol consumption, with older individuals also having higher odds of the outcome.

To evaluate the model's performance, I used an ROC curve, which showed a high AUC of 0.98, indicating that the model is really good at distinguishing between the positive and negative cases. Overall, the model performs well, with smoking, alcohol consumption, and age being the strongest predictors, while physical activity has a protective effect.

## Question 3

```
joe_current <- data.frame(
  Age = 40,
  Gender = factor("Male", levels = levels(myData$Gender)),
  Obesity = factor("No", levels = levels(myData$Obesity)),
```

```r
  Smoking_Status = factor("Smoker", levels = levels(myData$Smoking_Status)),
  Alcohol_Consumption = factor("Yes", levels = levels(myData$Alcohol_Consumption)),
  Rural_or_Urban = factor("Urban", levels = levels(myData$Rural_or_Urban)),
  Physical_Activity = factor("Low", levels = levels(myData$Physical_Activity))
)

interventions <- list(
  "Stop Smoking" = joe_current %>% mutate(Smoking_Status = factor("Non-Smoker", levels = levels(myData$
  "Stop Drinking" = joe_current %>% mutate(Alcohol_Consumption = factor("No", levels = levels(myData$Al
  "Move Rural" = joe_current %>% mutate(Rural_or_Urban = factor("Rural", levels = levels(myData$Rural_o
  "Increase Activity" = joe_current %>% mutate(Physical_Activity = factor("High", levels = levels(myDat
)

current_risk <- predict(logit_model, newdata = joe_current, type = "response")

risk_diff <- sapply(interventions, function(int) {
  current_risk - predict(logit_model, newdata = int, type = "response")
})

plot_data <- data.frame(
  Intervention = names(risk_diff),
  Risk_Reduction = as.numeric(risk_diff)
)

ggplot(plot_data, aes(x = reorder(Intervention, -Risk_Reduction), y = Risk_Reduction)) +
  geom_col(fill = "blue", alpha = 0.5) +
  labs(title = "Potential Risk Reduction from Interventions",
       x = "Intervention", y = "Absolute Risk Reduction") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(labels = scales::percent_format())
```




To reduce his risk of a heart attack, Joe should prioritize quitting smoking or reducing alcohol consumption. The outputted plot indicates that stopping drinking has the largest potential impact, with an estimated risk reduction of about 7%. Quitting smoking also provides a significant benefit, lowering risk by approximately 4.5%. In contrast, moving to a rural area appears to have no effect on reducing risk. Interestingly, increasing physical activity is associated with a slight increase in risk, which is unexpected and needs further investigation.

Overall, stopping alcohol consumption would be the most effective strategy for Joe to reduce his risk, followed closely by quitting smoking. While physical activity is generally considered beneficial for heart health, the observed increase in risk suggests that other factors may be at play (such as age). Given these findings, Joe's best course of action is to first address alcohol consumption and smoking, as they have the most substantial impact on reducing heart attack risk.