S&DS 230 Final Project: Cardiovascular Disease

Introduction

Cardiovascular disease is the leading cause of death in the United States. One person dies every 36 seconds in the United Statates from cardiovascular disease (https://www.cdc.gov/heartdisease/). In order to study what contributes to disease progression, analyzing risk factors in our daily life could provide important sights on the diagnosis and prognosis. The following analysis will shed light on cardiovascular risk factors and potentially guide us to a healthier way to lead out lives.

Data Source

The following analysis is performed on a dataset with various cardiovascular disease risk factors and patient outcomes from: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

```
heart_data <- read.csv("cardio_train.csv", sep=";")
```

Data Exploration

Initially, we want to understand the dimensions of the data and what information is encoded in the categories.

Data Dimension

The dimension of the data is:

```
dim(heart_data)
```

```
## [1] 70000    13
```

We can see that there is a total of 70,000 entries of patient outcome. According to the explanation on the dataset, the variables are encoded in the following way:

1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap_hi | int |
6. Diastolic blood pressure | Examination Feature | ap_lo | int |
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | Subjective Feature | smoke | binary |
10. Alcohol intake | Subjective Feature | alco | binary |
11. Physical activity | Subjective Feature | active | binary |
12. Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

Data Variables

The variables in our dataset are:

```
glimpse(heart_data)
```

```
## Rows: 70,000
## Columns: 13
## $ id          <int> 0, 1, 2, 3, 4, 8, 9, 12, 13, 14, 15, 16, 18, 21, 23, 24...
## $ age         <int> 18393, 20228, 18857, 17623, 17474, 21914, 22113, 22584,...
## $ gender      <int> 2, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 2, 2, 1, 2, 2, 1, 1, 1...
## $ height      <int> 168, 156, 165, 169, 156, 151, 157, 178, 158, 164, 169, ...
## $ weight      <dbl> 62, 85, 64, 82, 56, 67, 93, 95, 71, 68, 80, 60, 60, 78,...
## $ ap_hi       <int> 110, 140, 130, 150, 100, 120, 130, 130, 110, 110, 120, ...
## $ ap_lo       <int> 80, 90, 70, 100, 60, 80, 80, 90, 70, 60, 80, 80, 80, 70...
## $ cholesterol <int> 1, 3, 3, 1, 1, 2, 3, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ gluc        <int> 1, 1, 1, 1, 1, 2, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1...
## $ smoke       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0...
## $ alco        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0...
## $ active      <int> 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0...
## $ cardio      <int> 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0...
```

Data Preprocessing

Summary

A summary of the data and removing the ID's

```
summary(heart_data)
```

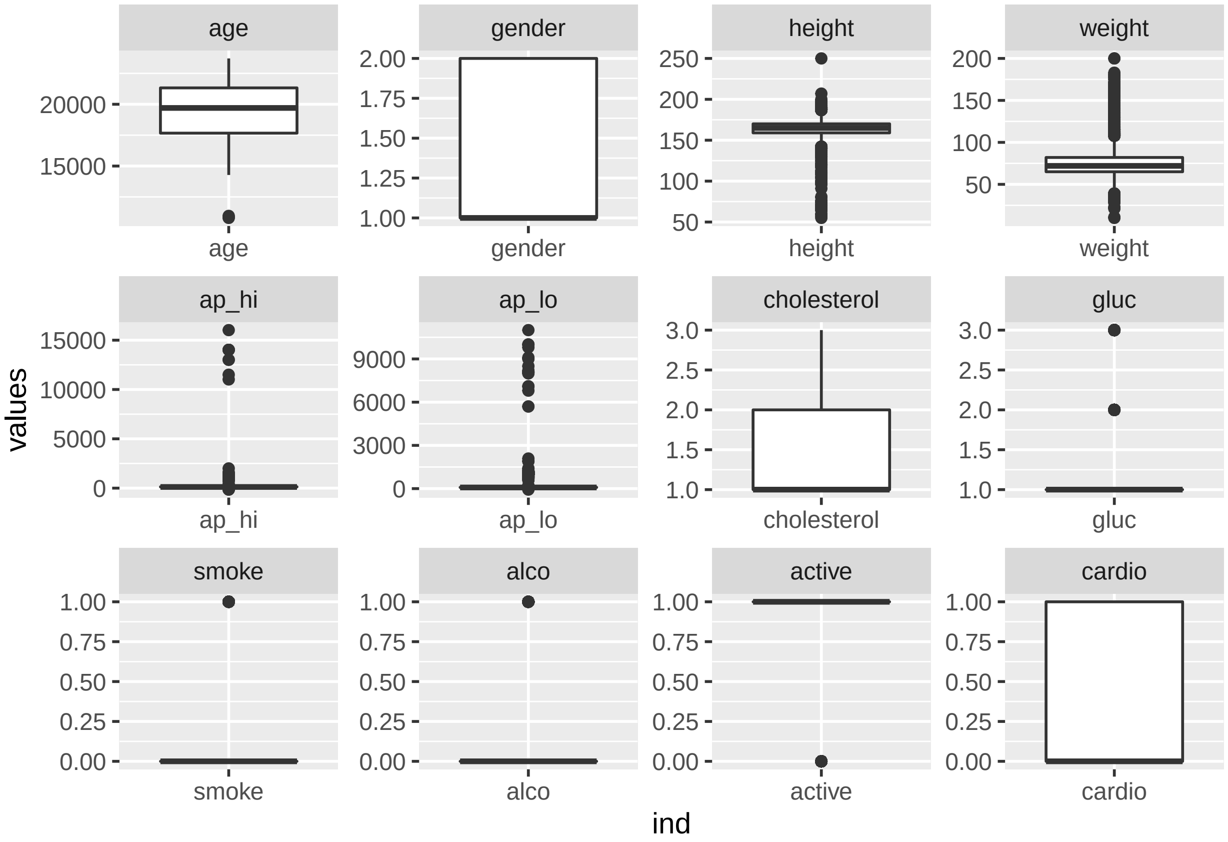
```
##        id              age            gender          height     
##  Min.   :    0   Min.   :10798   Min.   :1.00   Min.   : 55.0  
##  1st Qu.:25007   1st Qu.:17664   1st Qu.:1.00   1st Qu.:159.0  
##  Median :50002   Median :19703   Median :1.00   Median :165.0  
##  Mean   :49972   Mean   :19469   Mean   :1.35   Mean   :164.4  
##  3rd Qu.:74889   3rd Qu.:21327   3rd Qu.:2.00   3rd Qu.:170.0  
##  Max.   :99999   Max.   :23713   Max.   :2.00   Max.   :250.0  
##      weight            ap_hi              ap_lo            cholesterol   
##  Min.   : 10.00   Min.   : -150.0   Min.   :  -70.00   Min.   :1.000  
##  1st Qu.: 65.00   1st Qu.:  120.0   1st Qu.:   80.00   1st Qu.:1.000  
##  Median : 72.00   Median :  120.0   Median :   80.00   Median :1.000  
##  Mean   : 74.21   Mean   :  128.8   Mean   :   96.63   Mean   :1.367  
##  3rd Qu.: 82.00   3rd Qu.:  140.0   3rd Qu.:   90.00   3rd Qu.:2.000  
##  Max.   :200.00   Max.   :16020.0   Max.   :11000.00   Max.   :3.000  
##       gluc           smoke               alco             active      
##  Min.   :1.000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000  
##  1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:1.0000  
##  Median :1.000   Median :0.00000   Median :0.00000   Median :1.0000  
##  Mean   :1.226   Mean   :0.08813   Mean   :0.05377   Mean   :0.8037  
##  3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000  
##  Max.   :3.000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000  
##      cardio      
##  Min.   :0.0000  
##  1st Qu.:0.0000  
##  Median :0.0000  
##  Mean   :0.4997  
##  3rd Qu.:1.0000  
##  Max.   :1.0000
```

```
heart_data$id <- NULL
```

Outliers

Find potential outliers

```
stacked_heart_data <- stack(heart_data)
ggplot(stacked_heart_data, aes(x = ind, y = values)) +
  geom_boxplot() +
  facet_wrap(~ind , scales = "free")
```



Duplicates

Find duplicate entries

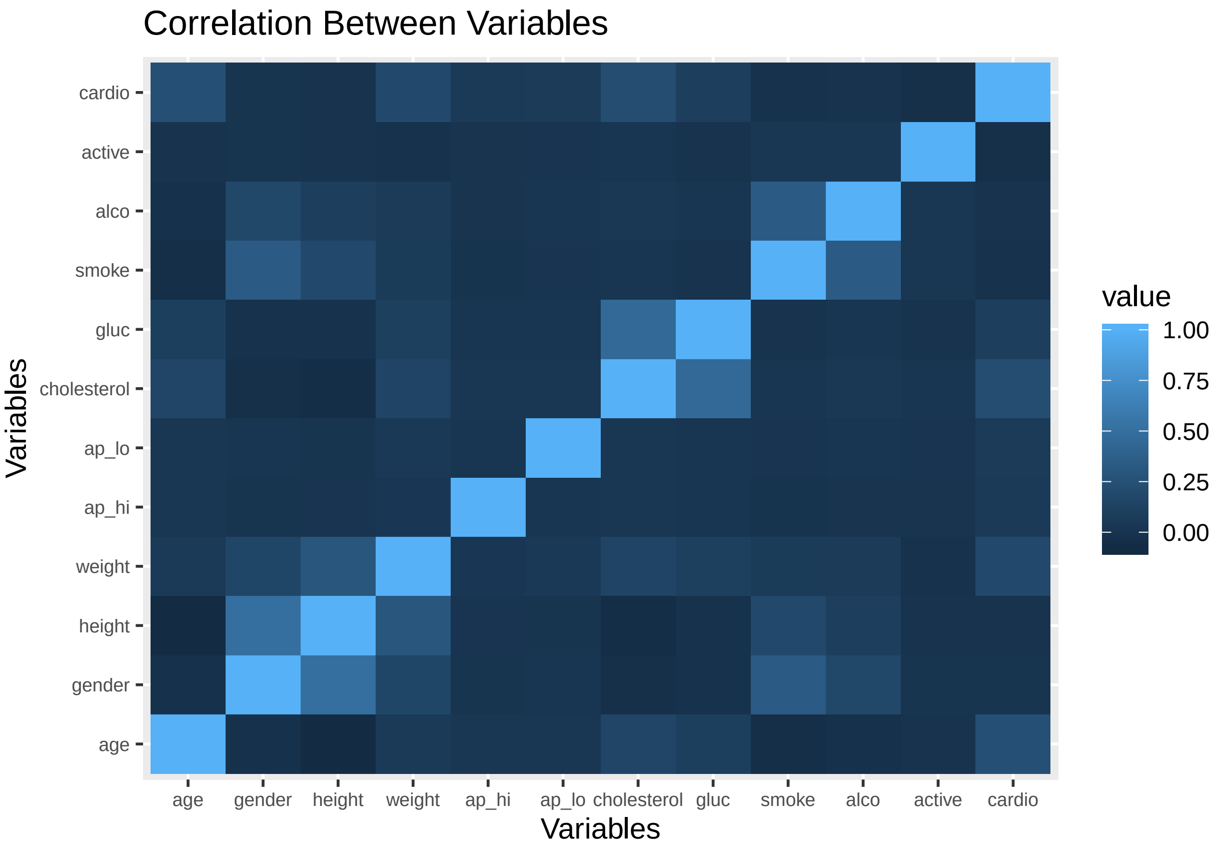
```
print(duplicated(heart_data$id))
```

```
## logical(0)
```

There are no duplicate values.

Data Analysis

```
cormat <- round(cor(heart_data), 3)
melted_cormat <- melt(cormat)
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  ggtitle("Correlation Between Variables") +
  xlab("Variables") +
  ylab("Variables") +
  theme(axis.text=element_text(size=7))
```


Correlation Between Variables
Variables
cardio
active
alco
smoke
gluc
cholesterol
ap_lo
ap_hi
weight
height
gender
age
age
gender
height
weight
ap_hi
ap_lo
cholesterol
gluc
smoke
alco
active
cardio
Variables
value
1.00
0.75
0.50
0.25
0.00