OBESITY PREDICTION

BY EVA BEYEBACH

MS-4203

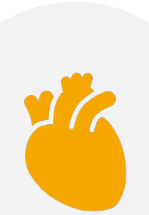
Background

- In 2022, 1 in 8 people in the world were living with obesity and 2.5 billion adults had overweight.
- Can lead to increased risk of type 2 diabetes, diseases and influence sleeping quality.
- The diagnosis of overweight and obesity is made by calculating the body mass index (BMI): weight (kg)/height^2 (m^2).
- Research suggests “severe OSA led to a 65% greater risk of developing any kind of cancer.”
- Having good sleep quality and enough sleep hours is also a good mitigator of cancer.
- Researchers found an increased risk of heart disease and total cancer, in patients with a higher resting heart rate.

Motivation

Our objective is to enhance the overall health of patients.

Our aim is to provide patients with recommendations to lower their BMI to foster a healthier lifestyle.



We want to analyze which factors have influence over sleep quality, to improve the patients sleep and reduce cancer risks.

We also aim to examine the factors contributing to high blood pressure and heart rate in order to facilitate lifestyle improvements and reduce cancer risks.

Problem

Which variables are significant predictors of BMI and which variables could reduce obesity?

Do obese people have sleep disorders and poor sleep quality? If yes, how can we improve that?

Which variables influence high blood pressure or high heart rate?

Which model is the best model to predict BMI Category?

Which other variables correlate with each other? How can we further improve patient lifestyle overall?

Data

- Sleep Health and Lifestyle Dataset from Kaggle
- 374 observations and 13 variables

Person ID: An identifier for each individual.

Gender: The gender of the person (Male/Female).

Age: The age of the person in years.

Occupation: The occupation or profession of the person.

Sleep Duration (hours): The number of hours the person sleeps per day.

Quality of Sleep (scale: 1-10): A subjective rating of the quality of sleep, ranging from 1 to 10.

Physical Activity Level (minutes/day): The number of minutes the person engages in physical activity daily.

Stress Level (scale: 1-10): A subjective rating of the stress level experienced by the person, ranging from 1 to 10.

BMI Category: The BMI category of the person (e.g., Underweight, Normal, Overweight).

Blood Pressure (systolic/diastolic): The blood pressure measurement of the person, indicated as systolic pressure over diastolic pressure.

Heart Rate (bpm): The resting heart rate of the person in beats per minute.

Daily Steps: The number of steps the person takes per day.

Sleep Disorder: The presence or absence of a sleep disorder in the person (None, Insomnia, Sleep Apnea).

Data Preparation: Structure

```r
# Changing variables to factor
sleep$Gender<- as.factor(sleep$Gender)
sleep$Occupation<- as.factor(sleep$Occupation)
sleep$BMI<- as.factor(sleep$BMI)
sleep$Disorder<- as.factor(sleep$Disorder)
```

```r
sleep <- separate(sleep, BloodPressure,
into = c("Systolic", "Diastolic"), sep = "/")
sleep$Systolic<- as.numeric(sleep$Systolic)
sleep$Diastolic<- as.numeric(sleep$Diastolic)
```

```r
sleep <- subset(sleep, select = - `Person ID`)
```

```
spc_tbl_ [374 × 13] (S3: spec_tbl_df/tbl
 $ Person ID                : num [1:374]
 $ Gender                   : chr [1:374]
 $ Age                      : num [1:374]
 $ Occupation               : chr [1:374]
Representative" ...
 $ Sleep Duration           : num [1:374]
 $ Quality of Sleep         : num [1:374]
 $ Physical Activity Level: num [1:374]
 $ Stress Level             : num [1:374]
 $ BMI Category             : chr [1:374]
 $ Blood Pressure           : chr [1:374]
 $ Heart Rate               : num [1:374]
 $ Daily Steps              : num [1:374]
8000 ...
 $ Sleep Disorder           : chr [1:374]
```

Data Preparation: Summary

```
sleep <- sleep %>%
  mutate(BMI = case_when(
    BMI == "Normal Weight" ~ "Normal",
    BMI == "Obese" ~ "Overweight",
    TRUE ~ as.character(BMI)
  ))
sleep$BMI<- as.factor(sleep$BMI)
```

```
#convert BMI to binary variable
sleep$BMI <- ifelse(sleep$BMI ==
                      "Normal", 1, 0)
sleep$BMI <- as.factor(sleep$BMI)
```

- No missing values
- No duplicates

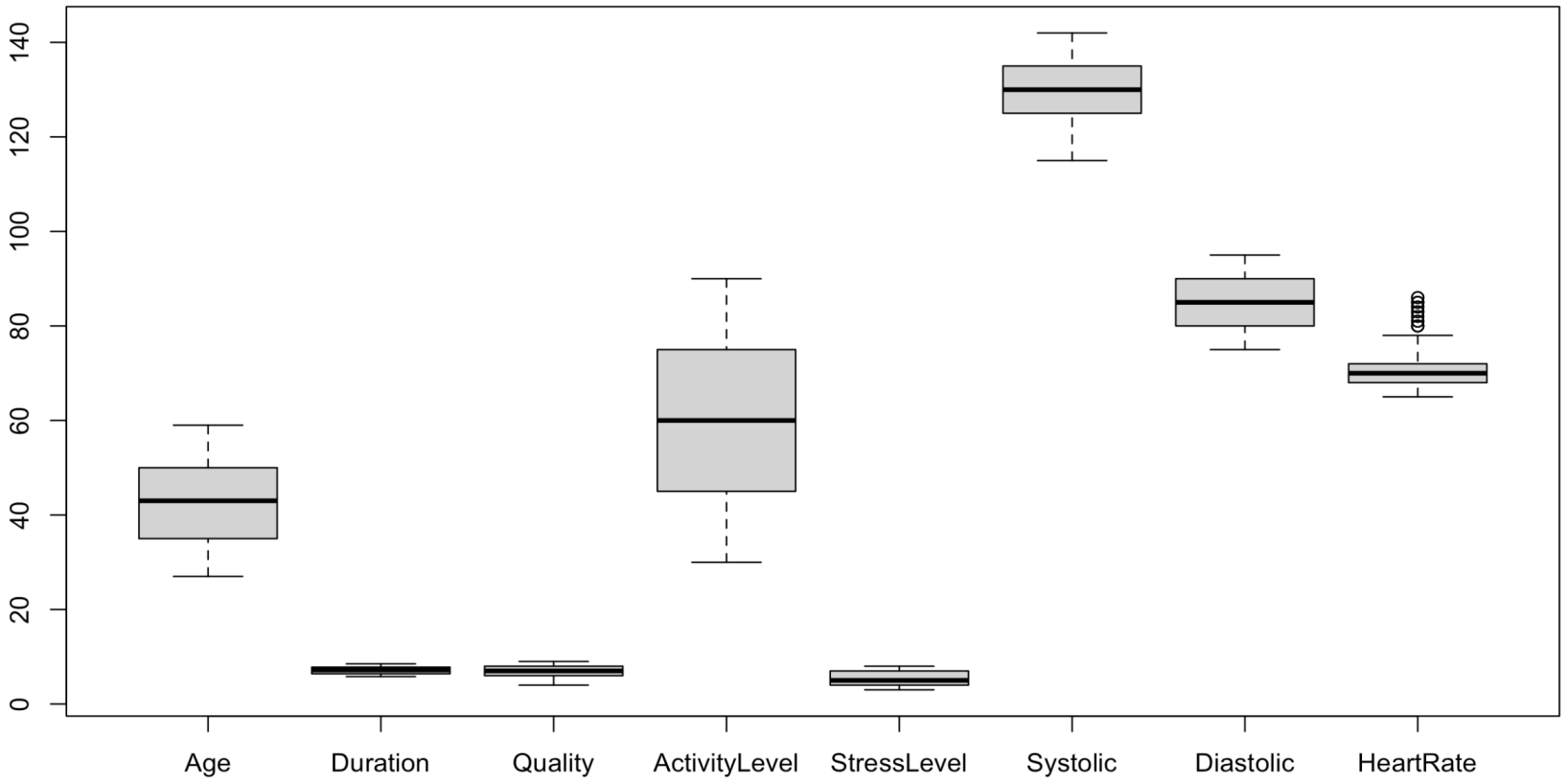
```
   Person ID        Gender         Age            Occupation    Duration
Min.   :  1.00   Female:185   Min.   :27.00   Nurse      :73   Min.   :5.800
1st Qu.: 94.25   Male  :189   1st Qu.:35.25   Doctor     :71   1st Qu.:6.400
Median :187.50                Median :43.00   Engineer   :63   Median :7.200
Mean   :187.50                Mean   :42.18   Lawyer     :47   Mean   :7.132
3rd Qu.:280.75                3rd Qu.:50.00   Teacher    :40   3rd Qu.:7.800
Max.   :374.00                Max.   :59.00   Accountant:37   Max.   :8.500
                                              (Other)    :43

    Quality      ActivityLevel    StressLevel              BMI
Min.   :4.000   Min.   :30.00   Min.   :3.000   Normal       :195
1st Qu.:6.000   1st Qu.:45.00   1st Qu.:4.000   Normal Weight: 21
Median :7.000   Median :60.00   Median :5.000   Obese        : 10
Mean   :7.313   Mean   :59.17   Mean   :5.385   Overweight   :148
3rd Qu.:8.000   3rd Qu.:75.00   3rd Qu.:7.000
Max.   :9.000   Max.   :90.00   Max.   :8.000

    Systolic       Diastolic       HeartRate         Steps
Min.   :115.0   Min.   :75.00   Min.   :65.00   Min.   : 3000
1st Qu.:125.0   1st Qu.:80.00   1st Qu.:68.00   1st Qu.: 5600
Median :130.0   Median :85.00   Median :70.00   Median : 7000
Mean   :128.6   Mean   :84.65   Mean   :70.17   Mean   : 6817
3rd Qu.:135.0   3rd Qu.:90.00   3rd Qu.:72.00   3rd Qu.: 8000
Max.   :142.0   Max.   :95.00   Max.   :86.00   Max.   :10000

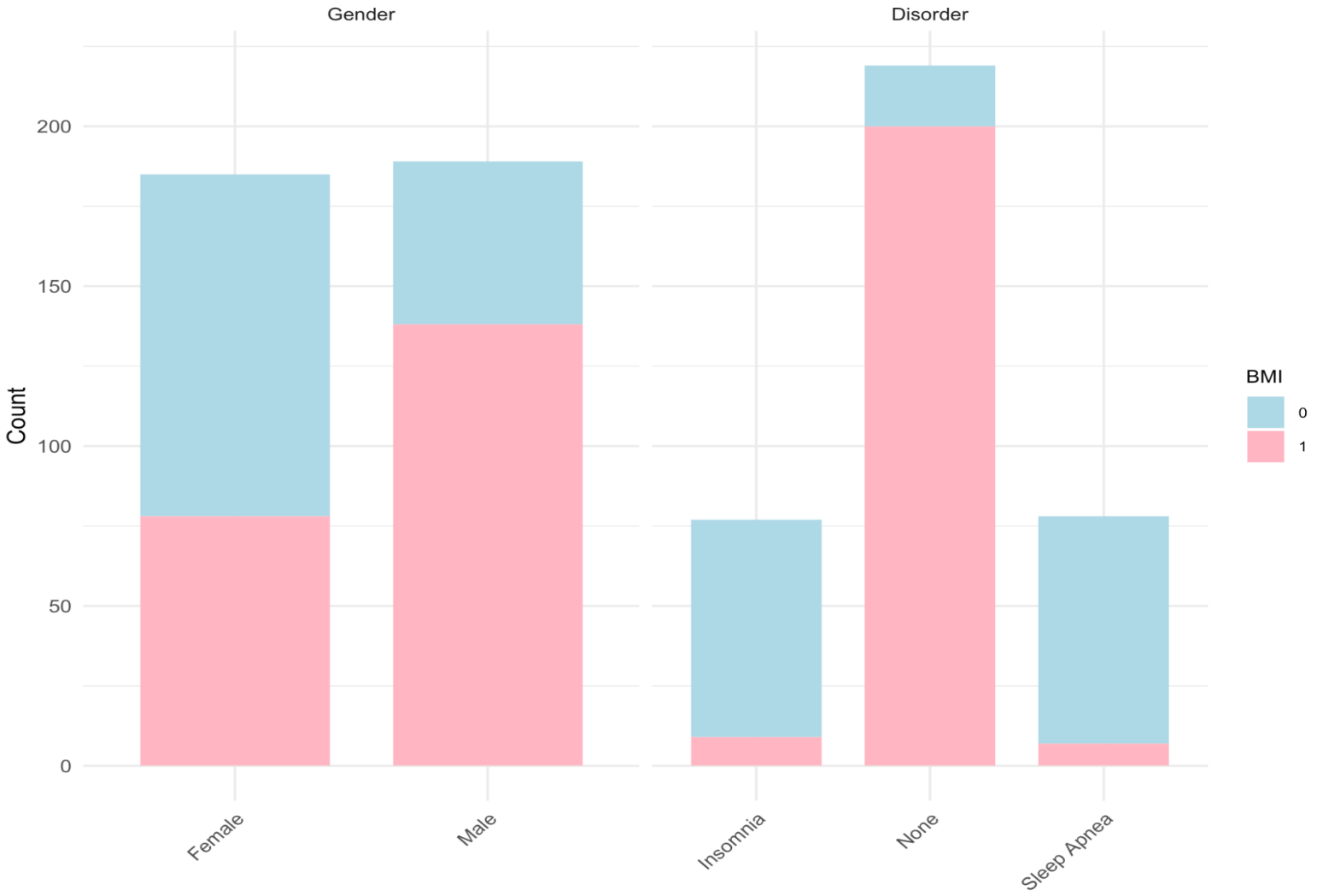
       Disorder
Insomnia   : 77
None       :219
Sleep Apnea: 78
```

Data Preparation: Outliers



- HeartRate has some outliers
- We will not remove them, because they are important for the analysis

Data Visualization by BMI: Categorical Variables



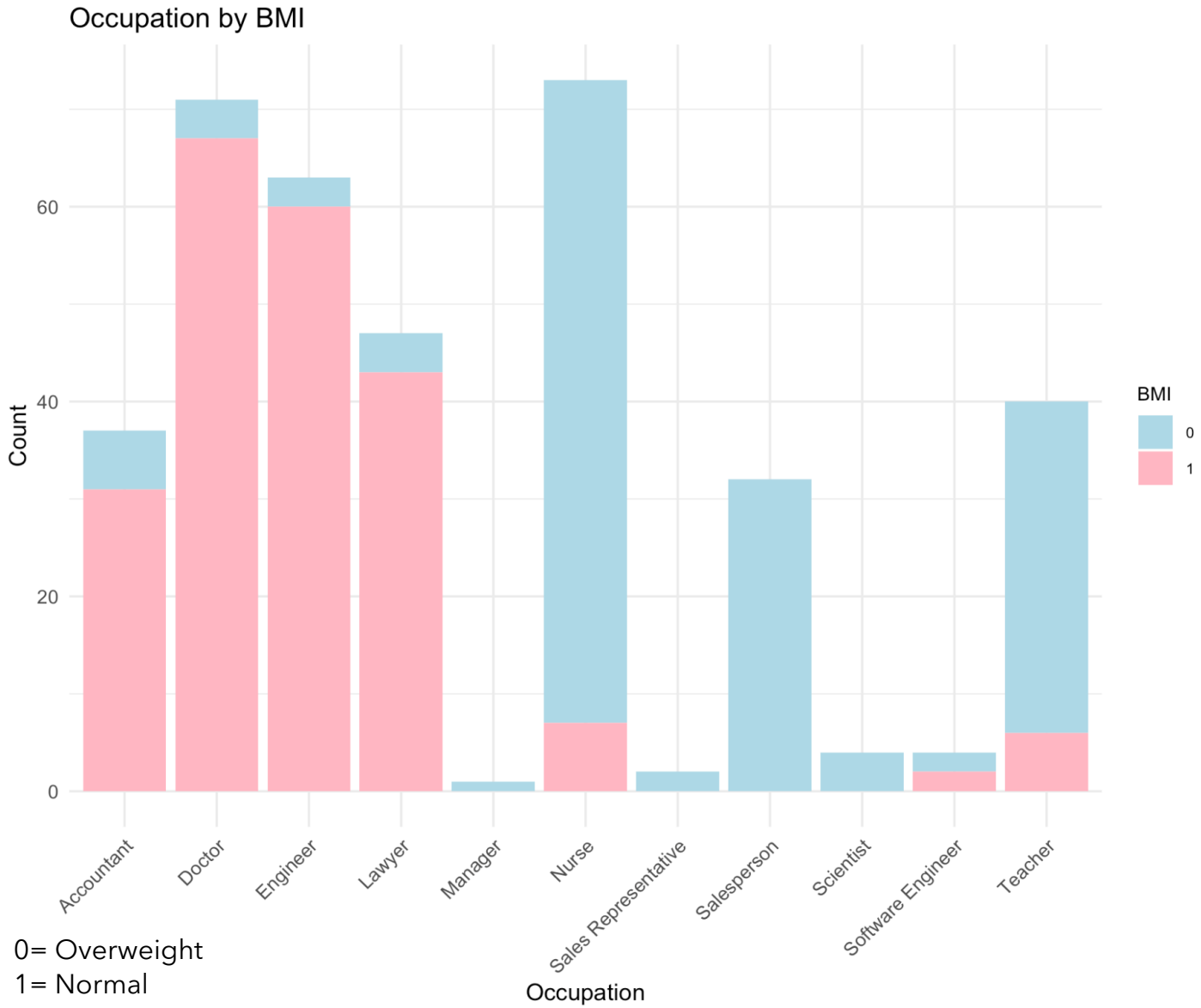
- 68% of females and 32% of males have overweight

	Female	Male
0	0.68	0.32
1	0.36	0.64

- 88% of people with overweight have Sleep Disorders

	Insomnia	None	Sleep Apnea
0	0.43	0.12	0.45
1	0.04	0.93	0.03

Data Visualization by BMI: Categorical Variables



0= Overweight
1= Normal

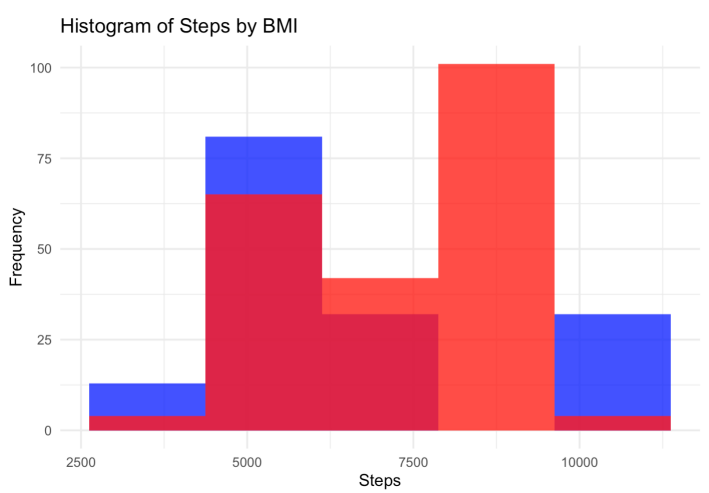
- Most of the Occupations, were people had overweight, have more women than men and vice versa

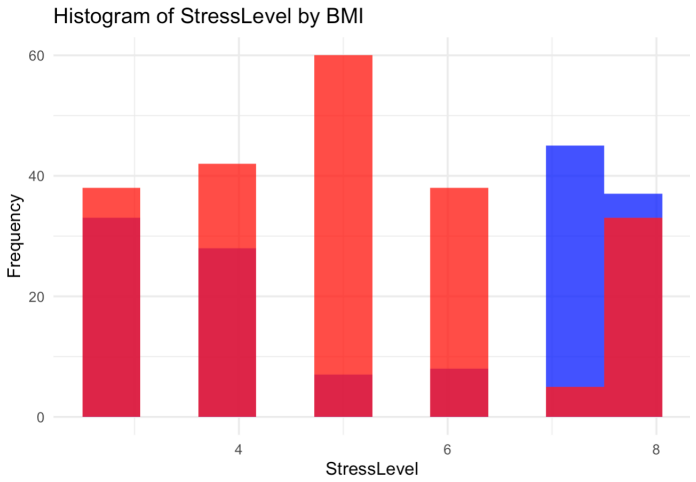
```
#proportions Occupation vs. Gender
```

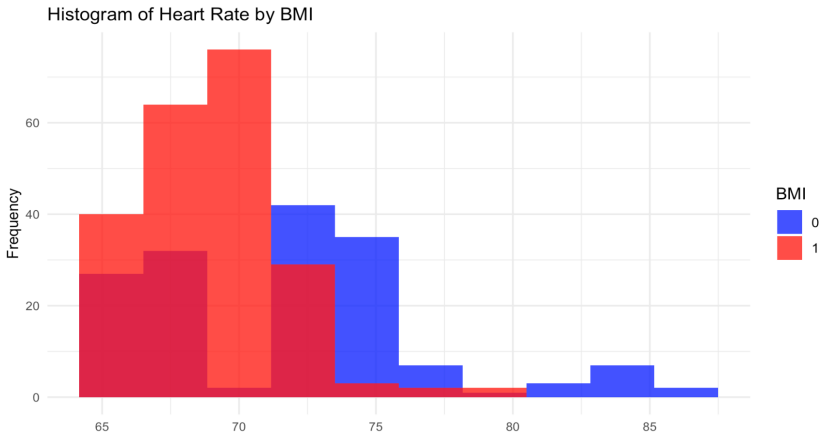
```
        Accountant Doctor Engineer Lawyer Manager Nurse
Salesperson
  Female       0.19   0.01     0.17   0.01    0.01  0.39
0.00
  Male         0.01   0.37     0.16   0.24    0.00  0.00
0.17

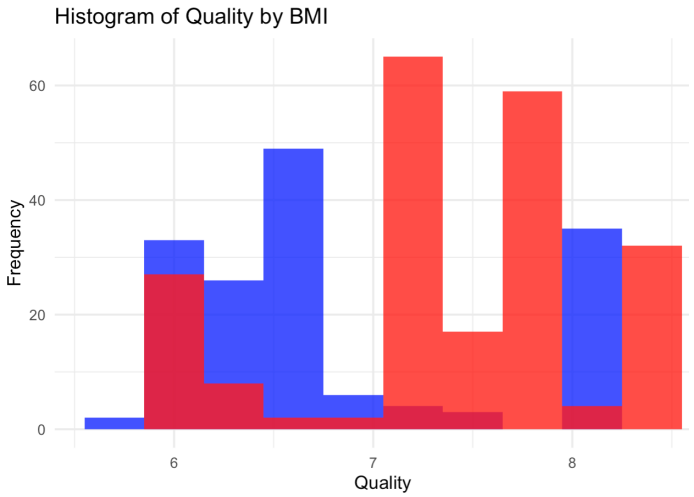
        Scientist Software Engineer Teacher
  Female      0.02                0.00    0.19
  Male        0.00                0.02    0.03
```

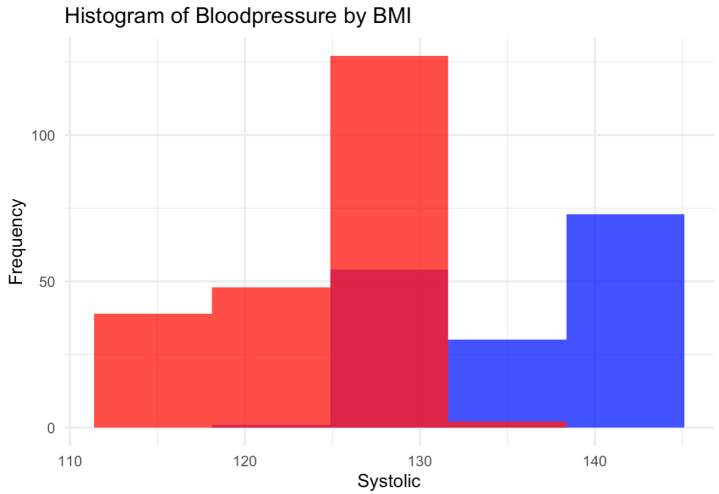
Histograms for Numerical Values by BMI

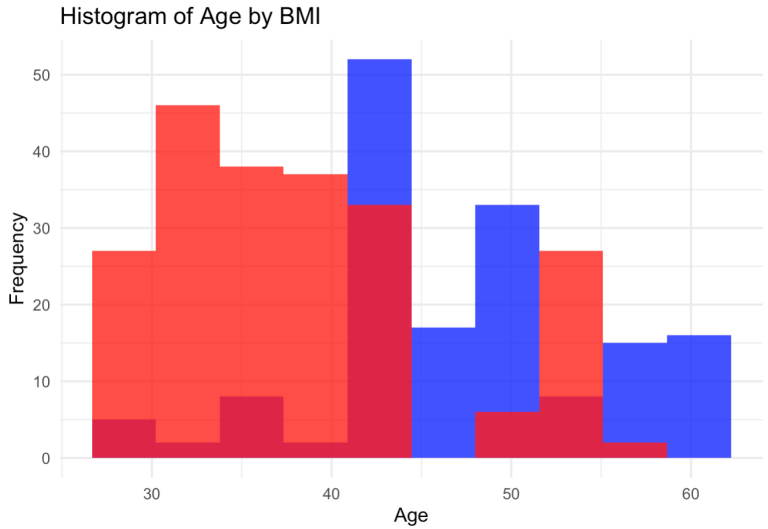






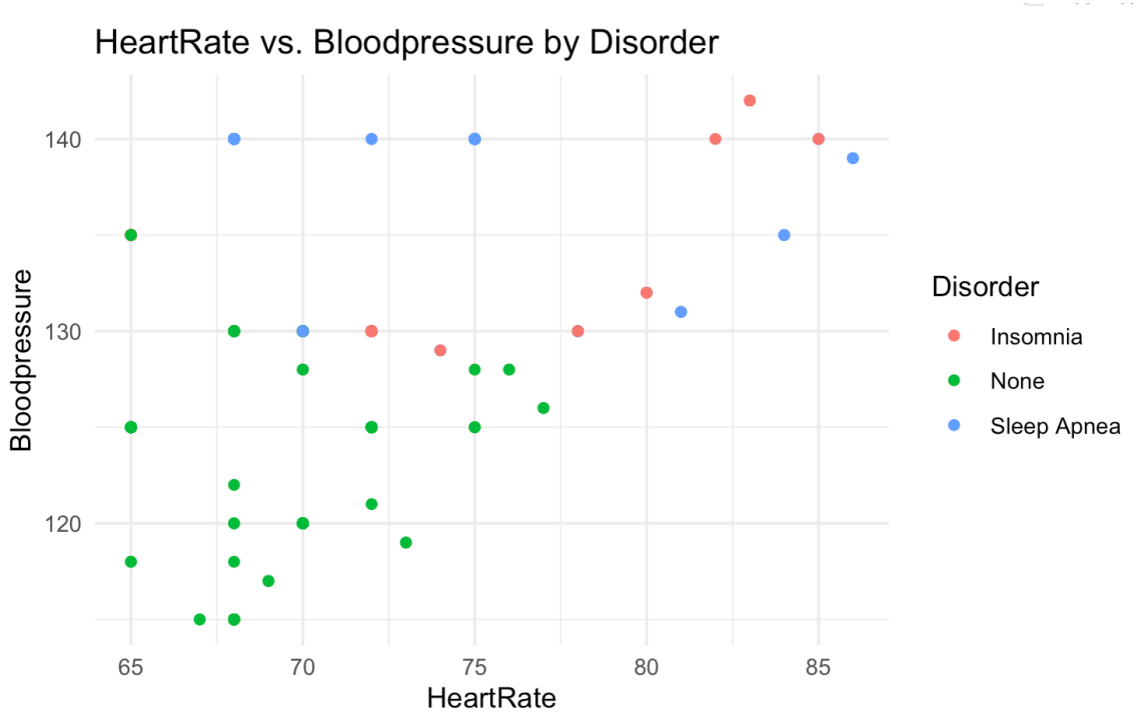


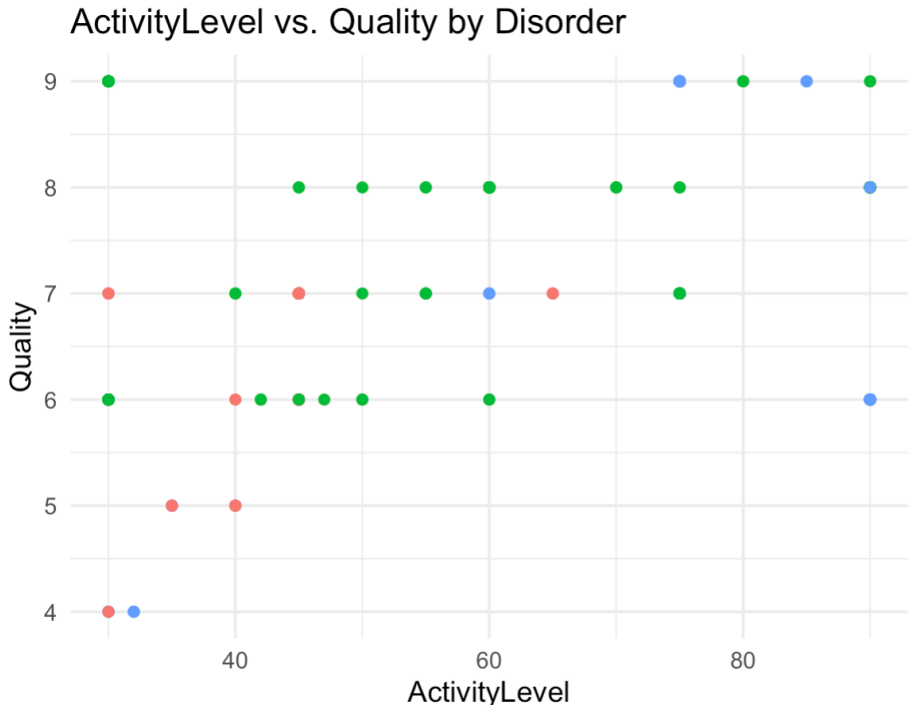




0= Overweight
1= Normal

Scatterplots for Numerical Values by Disorder





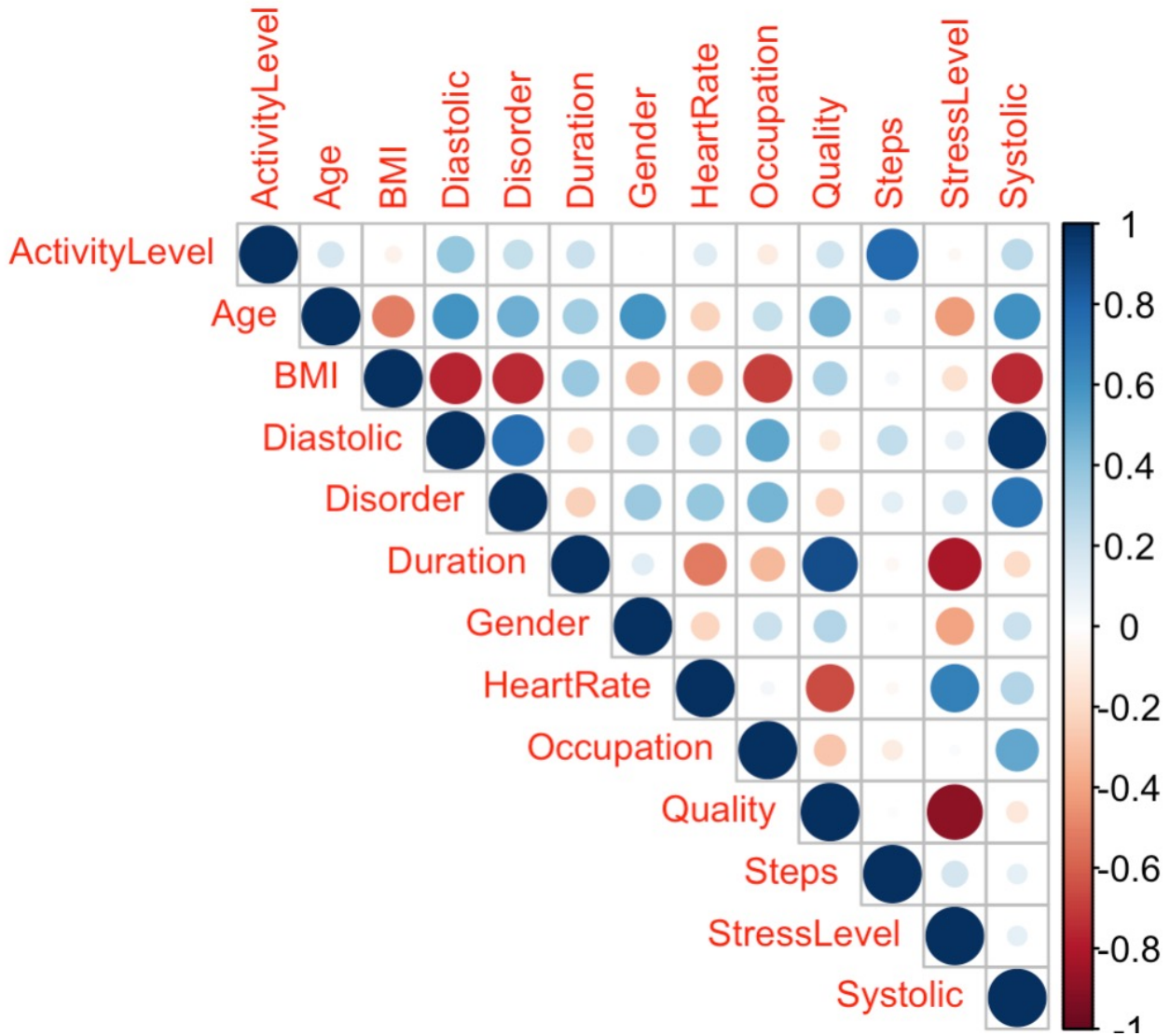
- High BP and HR are correlated and produce Sleep Disorders

- Higher ActivityLevel improves Quality and has less Sleep Disorders

Correlation Matrix

Highest positive Correlations:
- Duration & Quality (0.88)
- DIstolyc/Systolic & Disorder (0.70)
- HeartRate & StresLevel (0.67)

Highest negative Correlations:
- Quality & StressLevel (-0.89)
- Duration & StressLevel (-0.811)
- Disorder & BMI (-0.81)
- BMI and DIstolyc/Systolic (-0.769)



Modelling: Logistic Regression

- Predict weather someone will have overweight (0) or not(1)
- Dispersion for binomial family is taken to be 1 (normal weight)
- Small Coefficients
- Age, AL, SL and BP significant
- Less BP -> Normal Weight (-0.68)
- Less stress -> Normal Weight (-0.82)
- More activity-> Normal Weight

```
glm1 <- glm(as.factor(BMI) ~ ., family = binomial, data = sleep_train)
```

```
```{r}
vif(glm1)
```
```

```
glm(formula = as.factor(BMI) ~ . - Duration - Diastolic - Quality -
    Steps - HeartRate, family = binomial, data = sleep_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9597  -0.0843   0.0032   0.3253   1.9464

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        87.926162  19.411074   4.530 5.91e-06 ***
GenderMale          0.008982   1.106311   0.008   0.9935
Age                -0.166149   0.071841  -2.313   0.0207 *
ActivityLevel       0.054616   0.024312   2.246   0.0247 *
StressLevel        -0.819526   0.362675  -2.260   0.0238 *
Systolic           -0.615811   0.139536  -4.413 1.02e-05 ***
DisorderNone        1.359376   0.883329   1.539   0.1238
DisorderSleep Apnea 0.481480   1.289650   0.373   0.7089
```

Modelling: Logistic Regression

- Predict on Testing Set
- Accuracy: 94.69%
- Sensitivity: 91.30%
- Specificity: 97.01%
- Identifies better normal weight

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 42  2
         1  4 65

               Accuracy : 0.9469
                 95% CI : (0.888, 0.9803)
    No Information Rate : 0.5929
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8893

 Mcnemar's Test P-Value : 0.6831

            Sensitivity : 0.9130
            Specificity : 0.9701
         Pos Pred Value : 0.9545
         Neg Pred Value : 0.9420
             Prevalence : 0.4071
         Detection Rate : 0.3717
   Detection Prevalence : 0.3894
      Balanced Accuracy : 0.9416
```

Modelling: Decision Tree

```
set.seed(100)
row.num <- sample(1:nrow(sleep), 0.7*nrow(sleep))
sleep_train <- sleep[row.num,]
sleep_test <- sleep[-row.num,]
```

```
#Get the best size
best_size = cv.sleep$size[which.min(cv.sleep$dev)]
best_size
```

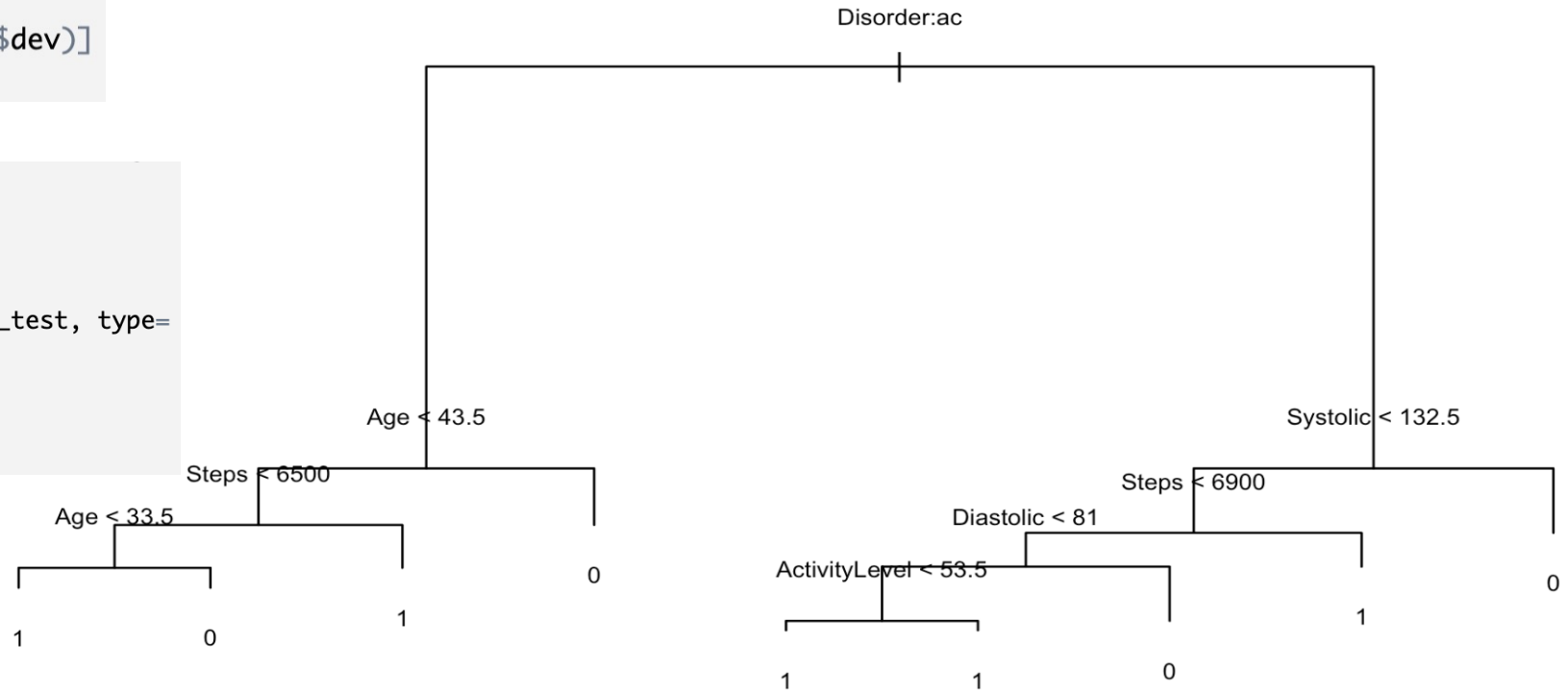
```
#Get the pruned tree of the best size
prune.sleep = prune.tree(tree, best = 8)

# Get predictions on the test set
preds_pruned = predict(prune.sleep, newdata = sleep_test, type=
"class")

caret::confusionMatrix(as.factor(preds_pruned),
as.factor(sleep_test$BMI))
```

```
Accuracy : 0.9735
Sensitivity : 0.9783
Specificity : 0.9701
```

Pruned Decision Tree



Modelling: Random Forest

- Identified Optimal Set of Hyperparameters

```
# Identify optimal set of hyperparmeters based on OOB error
opt_i <- which.min(oob_err)
print(hyper_grid[opt_i,])
```

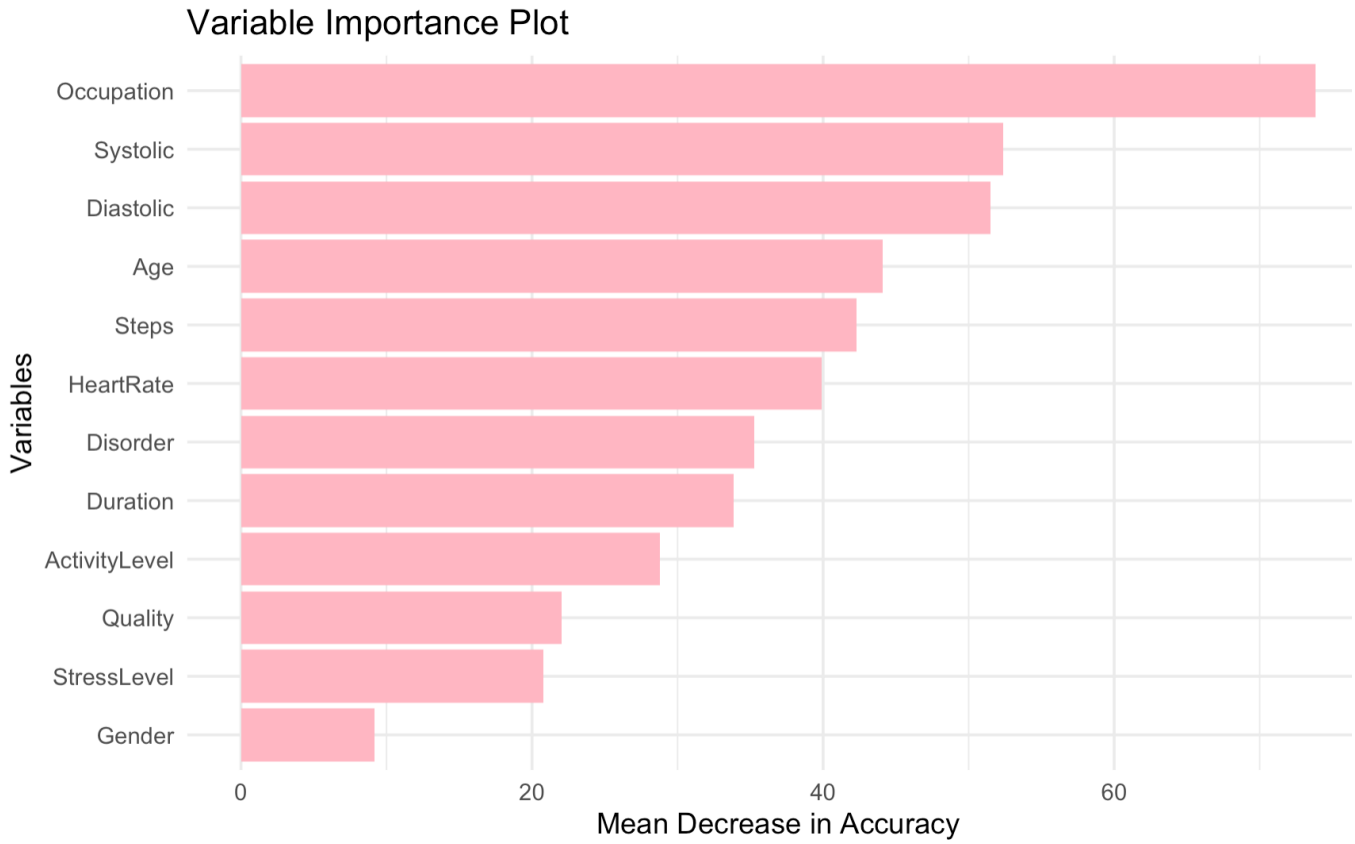
```
mtry = 5, ntree = 3000
```

- Built Random Forest with tuned parameters
- Predicted on testing set
- Accuracy: 99.1%

```
rf.opt <- randomForest(as.factor(BMI) ~ ., data =
sleep_train, mtry = 5, ntree = 3000,importance = T)

rf.probs.opt <- predict(rf.opt, newdata = sleep_test)
```

```
[1] 0.9911504
```



Modelling: Linear Regression (Quality)

- R^2: 95.05%
- Coefficients indicate that the following variables help for better sleep:
  - Males (2.87)
  - Older people (4.97)
  - Lower Stress Levels (-4.34)
  - Lower Heart Rate (5.7)
  - No Disorders (1.92)
  - Normal Weight (5.94)
- Variable may not have a substantial effect on the dependent variable

```
mse        [1] 0.09
mae        [1] 0.20
```

```
lm.sq2 <- lm(as.numeric(Quality)~. -Diastolic, data = sleep_train)
```

```
Call:
lm(formula = as.numeric(Quality) ~ . - Diastolic, data = sleep_train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0771 -0.1356 -0.0428  0.1620  0.8633

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.688e+00  8.832e-01   7.573 7.11e-13 ***
GenderMale          2.874e-01  5.895e-02   4.875 1.94e-06 ***
Age                 4.978e-02  4.462e-03  11.158  < 2e-16 ***
Duration            1.456e-01  5.882e-02   2.475  0.01398 *
ActivityLevel       2.849e-03  1.969e-03   1.447  0.14907
StressLevel        -4.399e-01  2.741e-02 -16.052  < 2e-16 ***
BMI1                5.937e-01  7.728e-02   7.682 3.59e-13 ***
Systolic           -6.410e-03  4.354e-03  -1.472  0.14226
HeartRate          -5.722e-03  8.332e-03  -0.687  0.49292
Steps               3.771e-05  2.565e-05   1.470  0.14278
DisorderNone        1.962e-01  6.417e-02   3.058  0.00247 **
DisorderSleep Apnea 2.337e-01  7.298e-02   3.202  0.00154 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
predictions.sq = predict(lm.sq2, newdata = sleep_test, type = "response")
```

Model Summary

Model	Test Accuracy	Sensitivity	Specificity
Logistic Regression	0.9469	0.9130	0.9701
Decision Tree	0.9735	0.9783	0.9701
Random Forest	0.9912	0.9783	1.000

Model	R^2	MSE	Mae
Linear Regression	0.9510	0.090	0.205

Statistical Inference Testing
Chi-Square Test for Sleep Disorder & BMI

- Null hypothesis (H0): There is no significant association between BMI and Sleep Disorder
- Alternative hypothesis (H1): There is significant association between BMI and Sleep Disorder

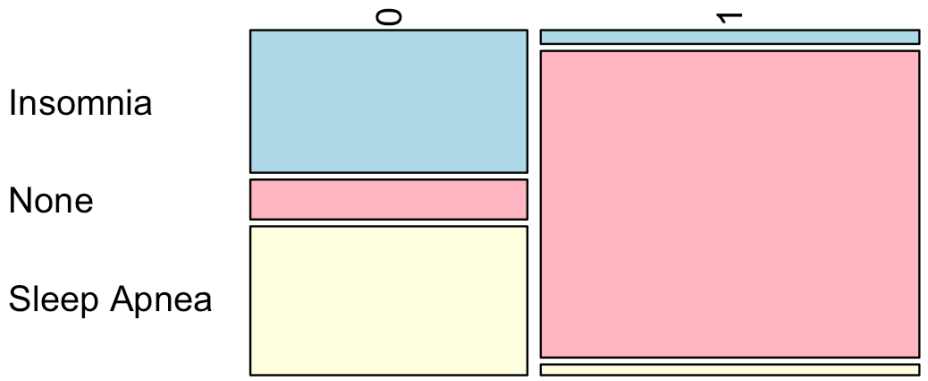
```
        Pearson's Chi-squared test

data:  contingency_table
X-squared = 244.19, df = 2, p-value < 2.2e-16
```

- We reject H0

Since the p-value is lower than 0.05

There is significant association between BMI and Sleep Disorder



Statistical Inference Testing
T-Test Quality/Heart rate & Activity level

- Null hypothesis (H0): There is no significant difference in the mean test scores between ActivityLevels and Sleep Quality.
- Alternative hypothesis (H1): There is significant difference in the mean test scores between ActivityLevels and Sleep Quality.

```
# Calculate Pearson's correlation coefficient
t_test_result <- t.test(sleep$ActivityLevel,
                        sleep$Quality)
```

```
# Calculate Pearson's correlation coefficient
t_test_result2 <- t.test(sleep$ActivityLevel,
                         sleep$HeartRate)
```

- We reject H0

There is a statistically significant difference between the mean Sleep Quality and mean ActivityLevels & HR + AL

This result suggests that there is an association between the variables (individuals with higher activity levels indeed have better sleep quality).

```
        Welch Two Sample t-test

data:  sleep$ActivityLevel and sleep$Quality
t = 48.065, df = 375.46, p-value < 2.2e-16
```

Findings & Results

Model	Test Accuracy	Sensitivity	Specificity	Best Predictors
Logistic Regression	0.9469	0.9130	0.9701	BP, Age, Stress(-) AL(+)
Decision Tree	0.9735	0.9783	0.9701	Disorder, Age, BP, Steps, AL
Random Forest	0.9912	0.9783	1.000	Occupation, BP, Age, Steps, HR

Model	R^2	MSE	Mae	Best Predictors
Linear Regression	0.7608	4.418	1.498	Stress, AL, Steps,Disorder

Findings & Results

Which variables are significant predictors of BMI and could reduce obesity?

- From the models , we can conclude that Random forest had the best Accuracy overall, when predicting BMI.
- From models we found the following out:
    - As BloodPressure, Age and Stress decreases, Weight descreases.
    - As ActivityLevel and Steps Increase, Weight decreases.
    - BP, Age, Stress, Disorder, Occupation and HR are best predictors of BMI.
    - 67% of Females and 33% of Males have overweight.
    - Some Occupations have more people with overweight, we found out that Gender played a significant role in that.

Do obese people have sleep disorders & bad sleep quality, how can it be improved?

- 88% of people with obesity have sleep disorders
- From visualizations, Correlations & Chi-Square:
    - Overweight = Bad Sleep Quality & Disorders
    - Higher HR, BP & SL= Bad Sleep Quality & Disorders
    - More AL & Steps = Better Sleep Quality

- From Linear Regression:
    - Age, Duration, Normal Weight = Better Sleep Quality
    - High StressLevel reduces Sleep Quality

Findings & Results

Which variables influence high BP or HR?

- From the Visualizations, high HeartRrate & BloodPressure are influenced by Overweight
- More Steps and ActivityLevel reduce both HR, BP and Stress
- From Correlations, HeartRate is influenced by StressLevel
- From T-Test StressLevel and Heartrate & StressLevel and Blood pressure correlate with each other

Which other variables correlate with each other?

- We observe that Age significantly influence overall health status
- Quality & Duration are correlated
- Stress level and HeartRate influenced each other
- Activity level & Steps are correlated

Recommendations for a better Health

We can conclude that BMI adversely affects health by elevating blood pressure and heart rate, as well as contributing to diminished sleep quality and the onset of sleep disorders.

We saw that stress and BMI were correlated, but our analysis did not allow us to determine the direction of causality.

High stress levels corresponded with elevated heart rates.

All this can lead to diseases, diabetes, sleeping disorders and poor health.

Recommendations for a better Health

- We have the following recommendations to improve health:
    - **More Steps**
    - **More Activity**
    - Both variables reduced BMI and stress, which consecutively reduces heart rate, blood pressure and improves sleep quality and sleep duration, causing less sleep disorders.
    - Trying to **reduce Stress** by meditating (for example) would help reduce heart rate, causing less heart diseases.

- We would like to expand our dataset to include additional variables such as dietary habits, smoking status, or pregnancy, to further investigate this and develop better methodologies aimed at enhancing health outcomes.
- Additional observation (1000+) would also help get more accurate results.