Individual Project Report

Developing Data Skills

ALEXANDER

17-04-2024

INTRODUCTION

This is survey data collected by the US National Center for Health Statistics (NCHS) which has conducted a series of health and nutrition surveys since the early 1960's. Since 1999 approximately 5,000 individuals of all ages are interviewed in their homes every year and complete the health examination component of the survey. The health examination is conducted in a mobile examination centre (MEC). This data consists of observations from 2009 - 2012 survey. The NHANES target population is "the non-institutionalized civilian resident population of the United States". NHANES, (American National Health and Nutrition Examination surveys). Visit to check out link to dataset.

The report tells us how insomnia can be predicated or classified using some health indicator as risk factors. These indicators are: i. Gender ii. Health Rating iii. Physical activity iv. BMI v. Pulse vi. Smoke status

Loading necassry packages

```
# Enter R code to load packages and import data
if(!require('pacman')) install.packages('pacman')
```

```
## Loading required package: pacman
```

```
p_load('tidyverse', 'patchwork', 'NHANES', 'gt', 'gtsummary', "psych",
       'magrittr', "rattle", "caret", 'parallel', 'randomForest', 'janitor')
# pacman: for loading/unloading packages
# NHANES: Dataset
# patchwork: for combining plots
# gtsummary: exploratory summary
# magrittr: For various pipe operators
# caret: for decision trees
# psych: computing statistical summary
# rattle: for plotting decision trees
# janitor: for tidy data
# tidyverse: for data wrangling and vizuals
# parallel: for parallel processing
# randomForest: for random forests
```

TASK 1: Data preparation

```
# Data cleaning
work_data <- NHANES %>%
  select(Insomnia=SleepTrouble, HealthRating=HealthGen, BMI,
         Pulse, PhysActive, Smoke100n, Gender) %>% #Changed SleepTrouble to Insomnia, same for HealthGe
  distinct() %>% # Some rows are duplicated, this keeps unique rows
  drop_na() %>% # Drops all rows with NA values
  clean_names() # Change columns name to snake case
  summary(work_data)
```

```
##   insomnia    health_rating      bmi            pulse          phys_active
##  No :2973   Excellent: 449   Min.   :15.02   Min.   : 40.00   No :1949
##  Yes:1039   Vgood    :1214   1st Qu.:24.28   1st Qu.: 64.00   Yes:2063
##              Good     :1571   Median :27.90   Median : 72.00
##              Fair     : 646   Mean   :29.02   Mean   : 72.35
##              Poor     : 132   3rd Qu.:32.36   3rd Qu.: 80.00
##                               Max.   :81.25   Max.   :136.00
##       smoke100n       gender
##  Non-Smoker:2215   female:2008
##  Smoker    :1797   male  :2004
##
##
##
##
```

The Data shows presence of duplicate observations and NA values which was cleaned to have 4012 observation left.

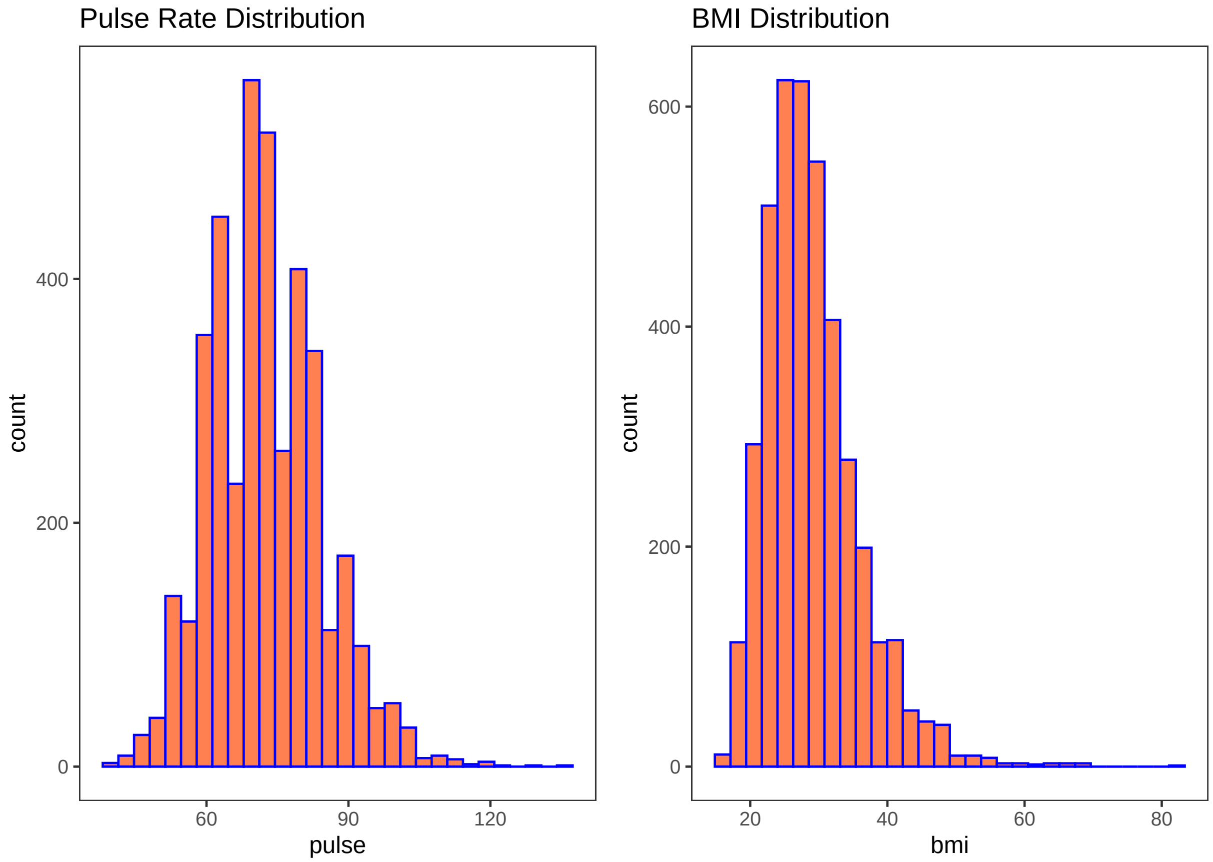
TASK 2: Exploratory Data Analysis

Vizualizing Data distributions of BMI & Pulse

Exploring continuous variable (BMI & Pulse) using a histogram to understand the distribution patterns.

```r
# Enter R code to explore the dataset informed by the project goals
# (e.g. create relevant plots, summary statistics, contingency tables)

ggplot(work_data, aes(x = pulse))+
  geom_histogram(fill = 'coral', col = 'blue')+
  theme_bw()+theme(panel.grid = element_blank())+
  labs(title = 'Pulse Rate Distribution') +

ggplot(work_data, aes(x = bmi))+
  geom_histogram(fill = 'coral', col = 'blue')+
  theme_bw()+theme(panel.grid = element_blank())+
  labs(title = 'BMI Distribution') +

plot_annotation(title = "Figure 1: Distribution of Bmi & Pulse rate")
```

Figure 1: Distribution of Bmi & Pulse rate



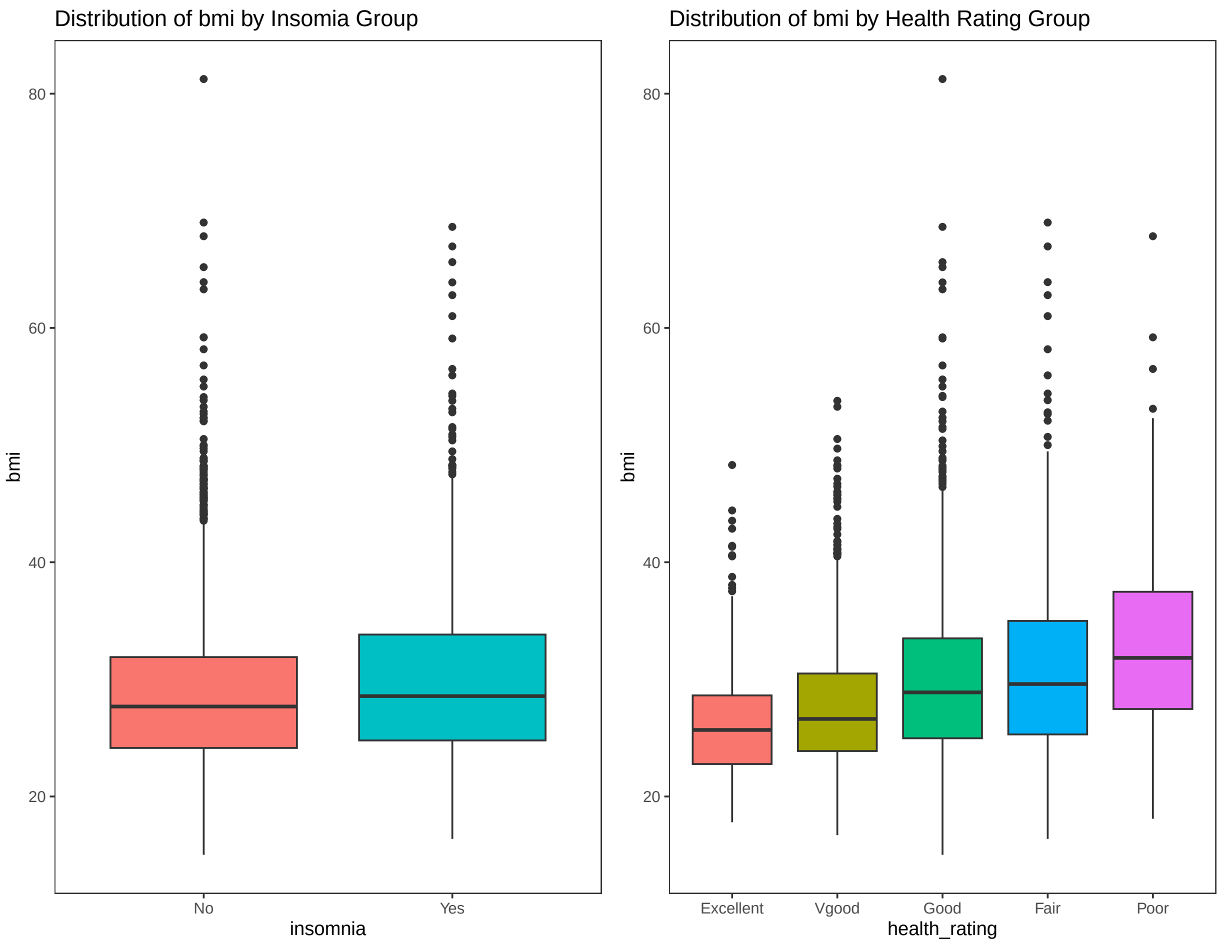
The distribution of pulse rate appears to be bimodal and not normally distributed. Also, that of BMI is right skewed. both variable has some outliers.

Vizualizing Data distributions of BMI by Insomina and Health Rating

Exploring BMI and grouping it by Insomnia nd Health rating using a box plot to understand the distribution patterns.

```r
ggplot(work_data, aes(x = insomnia, y = bmi))+
geom_boxplot(aes(fill = insomnia))+theme_bw()+
theme(panel.grid = element_blank())+theme(legend.position = 'none')+
labs(title = "Distribution of bmi by Insomia Group") +

ggplot(work_data, aes(x = health_rating, y = bmi))+
geom_boxplot(aes(fill = health_rating))+theme_bw()+
theme(panel.grid = element_blank())+theme(legend.position = 'none')+
labs(title = "Distribution of bmi by Health Rating Group")+
plot_annotation(title = "Figure 2")
```

Figure 2


Distribution of bmi by Insomia Group

Distribution of bmi by Health Rating Group

The box plot shows a good number of outliers seen in the BMI variable, with indication of positive skewed distribution.

Outliers Detection

The method used here is 3 standard deviation above and below the mean.

```
outliers <- work_data %>%
  drop_na(bmi) %>%
  filter(bmi < mean(bmi)-3*sd(bmi) | bmi > mean(bmi)+3*sd(bmi))

summary(outliers)
```

```
##  insomnia   health_rating      bmi            pulse      phys_active
##  No :25    Excellent: 0    Min.   :49.70   Min.   : 56   No :32
##  Yes:19    Vgood    : 4    1st Qu.:52.24   1st Qu.: 68   Yes:12
##            Good     :20    Median :54.30   Median : 81
##            Fair     :14    Mean   :57.13   Mean   : 80
##            Poor     : 6    3rd Qu.:61.46   3rd Qu.: 86
##                            Max.   :81.25   Max.   :104
##       smoke100n     gender
##  Non-Smoker:29   female:35
##  Smoker    :15   male  : 9
##
##
##
##
```

At the end, 44 observations appears to be seen as outliers as compared to others

Summary Statistics

Insomnia vs Bmi

Here a simple summary statistics is is performed to understand the boxplot shown in figure 2. This involves the bmi when group by those with and without Insomnia.

```
IB <- work_data %>%
group_by(insomnia) %>%
summarise(min = min(bmi),
          mean = mean(bmi),
          median = median(bmi),
          sd = sd(bmi),
          max = max(bmi)) %>%
gt() %>%
tab_header(title = "Table 1: Summary Statistics of BMI by Insomnia Group") %>%
cols_label(insomnia = 'Insomnia Category', min = "Minimum BMI",
           mean = "Average BMI", median = "Median BMI",
           sd = 'Standard deviation', max = 'Maximum BMI') %>%
cols_align(align = "center")
print(IB)
```

Insomnia vs Gender

```
IG<-tbl_cross(work_data, row = insomnia, col = gender, percent = "row") %>%
modify_header(label = "Table 2: **Gender**") %>%
bold_labels()
print(IG)
```

Among the 1039 people who had insomnia, 604(58%) where females while 435(42%) where males. Indication that more women tend to experience insomnia compared to males in this data.

Insomnia vs Smokers

```
IS<-tbl_cross(work_data, row = insomnia, col = smoke100n, percent = "row") %>%
modify_header(label = "Table 3: **Smoke Category**") %>%
bold_labels()
print(IS)
```

insomnia tends to be prevalent among smokers (55%) compared to non-smokers (45%) of 1039 people with insomnia issues.

Insomnia vs Physical Activity

```
IP<-tbl_cross(work_data, row = insomnia, col = phys_active, percent = "row") %>%
modify_header(label = "Table 4: **Physical Active**") %>%
bold_labels()
print(IP)
```

slightly lower number of people with insomnia performed physical activity (49%) compared to those who did not do physical activity (51%)

TASK 3: Analysis plan

All statistical test carried out would use 5% statistical significance and 95% confidence interval. From the extracted **work_data**, there are two continuous variables and 5 categorical variables. Below are detailed test to be carried out:

1a. A chi-squared test would be used to check for association between two categorical variable. - First, relationship between Insomnia vs Gender variable. - The relationship between Insomnia vs Physical Activity. - The relationship between Insomnia vs Smoke100n.

1b. A T.test or Wilcox-test would be used to check for association between one categorical variable and a continuous variable. - The association between Insomnia category vs BMI variable.

2. An Anova test would be used to check for association between more than two categorical variable and a continuous variable.

- The association between Health rating categories vs BMI variable.

3. A simple linear regression to understand the correlation between Pulse rate and BMI.
4. A logistic regression model of Gender, Physical Activity, Pulse, BMI and Smoking as risk factors for Insomnia
5. A decision tree model of all significant variables including Health rating as risk factors for Insomnia
6. A Random Forest model of all other significant variables including Health rating as risk factors for Insomnia

---

TASK 4: Data Analysis

Chi-squared test

A chi-squared test would be used to check for association between two categorical variable. This is to check if there is a relationship between Insomnia, Gender, Physical Active & Smoke100n. This tests for whether Insomnia is associated with any of the other three categorical variables.

Relationship between Gender and Insomnia

Null Hypothesis: There was no association between Insomnia and the Gender of an individual as a risk factor.

Alternative Hypothesis: There was a relationship between Insomnia and the gender of an individual as a risk factor.

```r
work_data %>%
  select(insomnia, gender) %>%
  table() %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 36.207, df = 1, p-value = 1.774e-09
```

Statistical interpretation: The chi-squared test (X2 = 36.207,df = 1, $p < 0.05$) is very significant, so we reject the null Hypothesis, and conclude that Sex could be a possible risk factor for Insomnia

Association between Ageover65 and Arthritis

Null Hypothesis: There is no association between physical activity and the risk of having insomnia.

Alternative Hypothesis: There is an association between physical activity and the risk of having insomnia.

```r
work_data %$%
  table(insomnia, phys_active) %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 3.4507, df = 1, p-value = 0.06323
```

Statistical interpretation: The chi-squared test (X2 = 3.45, df = 1, $p < 0.05$) is not significant, so we fail to reject the null Hypothesis, and conclude that being physical activity may not be a possible risk factor for insomnia.

Association between Smoking and Insomnia

Null Hypothesis:There is no association between smoking and the risk of having Insomnia.

Alternative Hypothesis: There is a relationship between smoking and the risk of having arthritis.

```r
work_data %$%
  table(insomnia, smoke100n) %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 61.405, df = 1, p-value = 4.646e-15
```

Statistical interpretation:The chi-squared test (X2 = 61.045, df = 1, p < 0.05) is statistically significant, so we reject the null Hypothesis, and conclude that smoking could be a possible risk factor of arthritis.

Wilcox-Test (Non-Parametric)

A Wilcox-test instead of t.test would be used to check for association between two groups of a categorical variable and a continuous variable (outcome). The association between insomnia category vs BMI variable. This test is a non-parametric test that can be used for not normally distributed data as in the case presently seen is a skewed data.

Null Hypothesis: There is no difference in the median unit of BMI observed depending on having arthritis symptoms or not.

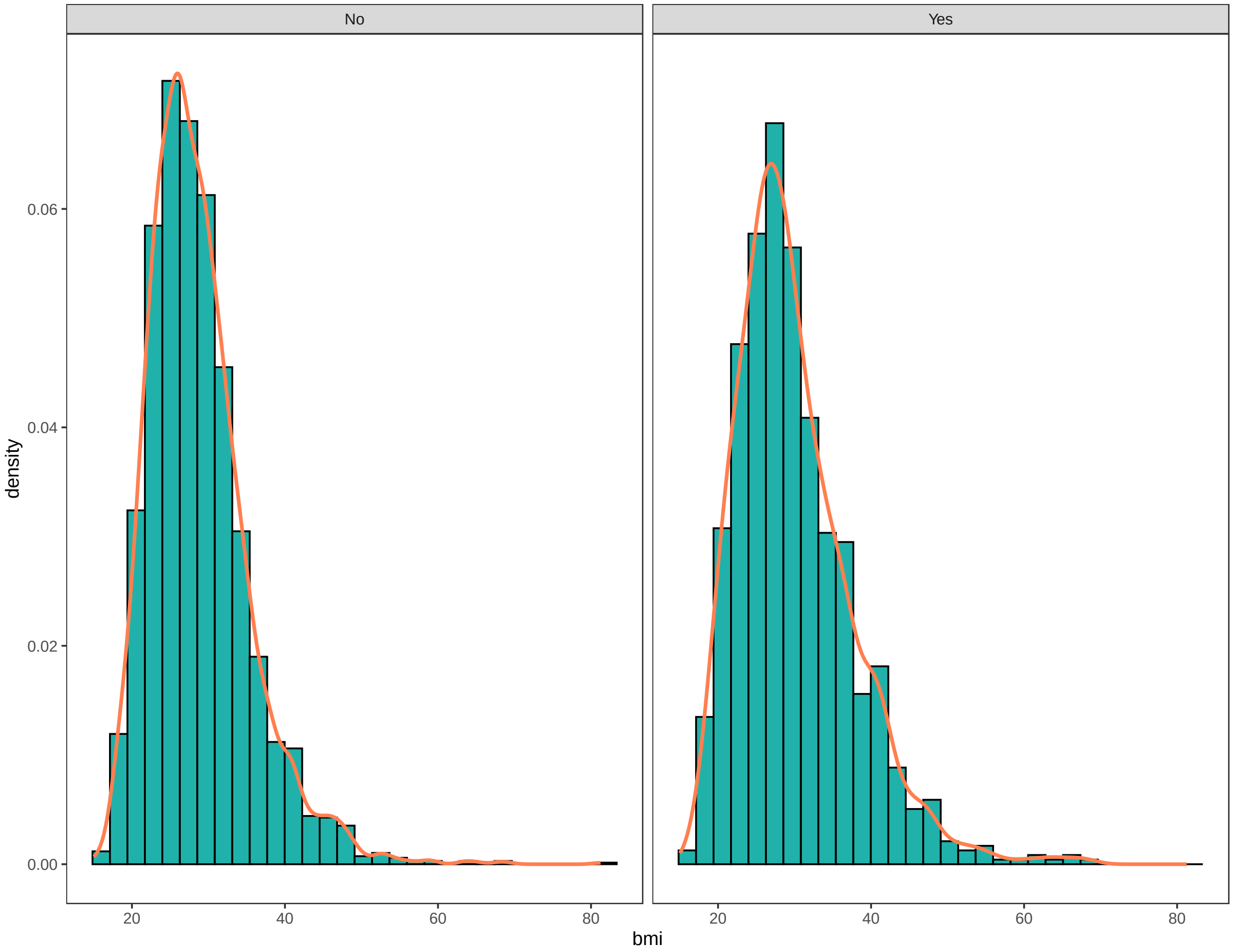
Alternative Hypothesis: There is difference in the median unit of BMI observed depending on having arthritis symptoms or not.

```r
ggplot(work_data, aes(x = bmi)) +
geom_histogram(aes(y=..density..), fill = 'lightseagreen', col = 'black') +
geom_density(col = "coral", linewidth = 1) + facet_wrap(~insomnia) +
theme_bw() + theme(panel.grid=element_blank()) +
labs(title = "Figure 3: Distribution Of BMI by Insomnia")
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Figure 3: Distribution Of BMI by Insomnia



```
wilcox.test(bmi ~ insomnia, data = work_data, conf.int = T)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  bmi by insomnia
## W = 1406555, p-value = 1.778e-05
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -1.4000206 -0.5199626
## sample estimates:
## difference in location
##             -0.9700145
```

Statistical interpretation: The test is statistically significant, so we reject the null hypothesis that the difference between medians in insomnia category is less than or equal to 0 (W = 1406555, $p < 0.05$), and conclude that there is evidence that individuals in the two insomnia group may not have the same BMI. The difference of group with no arthritis from the group with arthritis is ranging approximately between -1.40 & -1.52 95% confidence interval as seen that the difference in pseudo(median) between the two groups is approximately -0.97.

Anova Test

Descriptive statistics by group

```
work_data %$%      # Exposition pipe
  describeBy(      # describeBy function from psych
    bmi,           # Outcome variable
    health_rating  # Grouping variable
  )
```

```
##
##  Descriptive statistics by group
## group: Excellent
##    vars   n  mean   sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 449 26.21 4.72  25.68   25.79 4.37 17.8 48.3  30.5 1.05     1.85 0.22
## ------------------------------------------------------------
## group: Vgood
##    vars    n  mean   sd median trimmed  mad  min   max range skew kurtosis   se
## X1    1 1214 27.65 5.59  26.62   27.15 4.86 16.7 53.78 37.08 1.05     1.78 0.16
## ------------------------------------------------------------
## group: Good
##    vars    n  mean   sd median trimmed  mad   min   max range skew kurtosis
## X1    1 1571 29.77 7.08  28.89   29.17 6.24 15.02 81.25 66.23 1.29     4.02
##      se
## X1 0.18
## ------------------------------------------------------------
## group: Fair
##    vars   n mean  sd median trimmed  mad   min max range skew kurtosis   se
## X1    1 646 30.9 7.9   29.6   30.16 6.98 16.38  69 52.62 1.14     2.14 0.31
## ------------------------------------------------------------
## group: Poor
##    vars   n  mean   sd median trimmed  mad  min   max range skew kurtosis   se
## X1    1 132 33.09 8.59  31.83   32.33 6.98 18.1 67.83 49.73 1.02     1.64 0.75
```

This test is done to check for the analysis of variance within each groups and how it affects life span.

Null Hypothesis: BMI has no impact on health rating

Alternative Hypothesis: BMI has some level of impact on health rating.

```
anova_fit <- aov(bmi ~ health_rating, data = work_data)
summary(anova_fit)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## health_rating    4  11201  2800.4   63.57 <2e-16 ***
## Residuals     4007 176514    44.1
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Statistical interpretation: The p-value for the F-statistic of the ANOVA is statistically significant with low p value ($p < 0.05$), suggesting that, there is an overall effect of BMI on the health rating of individuals.

Post-Hoc Test

A post-hoc comparisons to tell us which groups are different. This is done using Tukey's HSD test for each pairwise comparison. We would like to get the differences in means, their associated p-value and confidence interval.

```
TukeyHSD(anova_fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = bmi ~ health_rating, data = work_data)
##
## $health_rating
##                      diff       lwr      upr     p adj
## Vgood-Excellent 1.438229 0.4377698 2.438688 0.0008447
## Good-Excellent  3.563942 2.5946605 4.533224 0.0000000
## Fair-Excellent  4.690431 3.5775382 5.803324 0.0000000
## Poor-Excellent  6.887329 5.0939853 8.680674 0.0000000
## Good-Vgood      2.125714 1.4335627 2.817865 0.0000000
## Fair-Vgood      3.252202 2.3701048 4.134300 0.0000000
## Poor-Vgood      5.449101 3.7890860 7.109115 0.0000000
## Fair-Good       1.126489 0.2799160 1.973061 0.0026431
## Poor-Good       3.323387 1.6819736 4.964800 0.0000004
## Poor-Fair       2.196898 0.4667942 3.927003 0.0048584
```

All of the pairwise comparisons are statistically significant at the 5% significance level ($p < 0.05$) . The largest difference in means is 6.88 for the comparison between the poor - excellent group with 95% confidence interval (5.09, 8.68). the least difference in mean was between the fair-good group In all, this could mean that BMI, have impact on the health rating of the people.

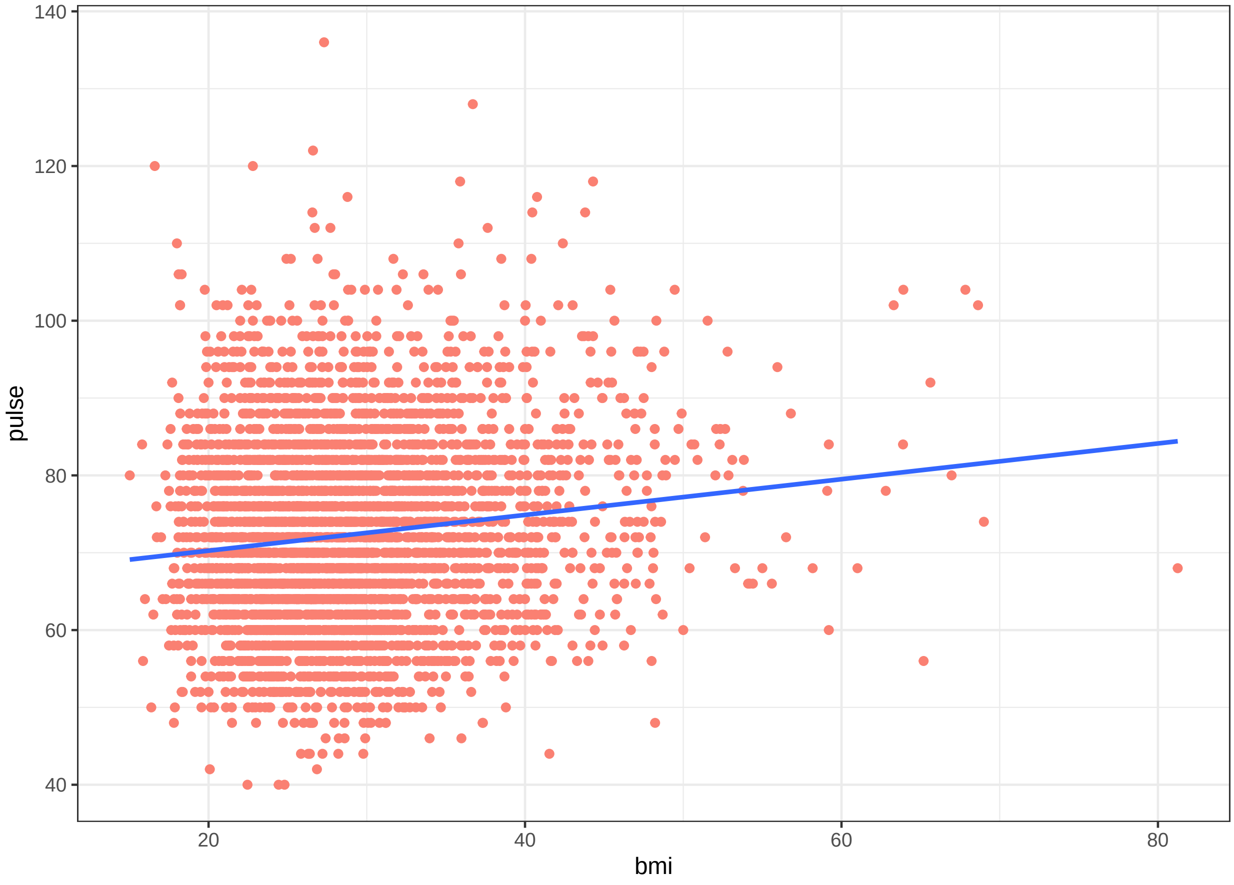
Task 5: Simple Linear Regression

The correlation coefficient used for this linear model is by default know as the spearman's since it the distribution is not normally distributed method. Here the summary and confidence intervals will be interpreted.

```
# Enter the R code to perform the analyses mentioned above.
ggplot(work_data, aes(y = pulse, x = bmi)) +
geom_point(col = "salmon") + geom_smooth(method = 'lm', se = FALSE)+
theme_bw() + theme(panel.background = element_blank()) +
labs(title = "Figure 4: Scatter plot of Pulse against BMI")
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

Figure 4: Scatter plot of Pulse against BMI



```
cor.test(work_data$bmi, work_data$pulse, method = "spearman")
```

```
## Warning in cor.test.default(work_data$bmi, work_data$pulse, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  work_data$bmi and work_data$pulse
## S = 9622167405, p-value = 1.691e-11
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.105992
```

```
fit_model <- lm(pulse ~ bmi, data = work_data)
summary(fit_model)
```

```
##
## Call:
## lm(formula = pulse ~ bmi, data = work_data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -31.371  -8.334  -0.775   7.213  64.051 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  65.6400     0.8080  81.239   <2e-16 ***
## bmi           0.2311     0.0271   8.527   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 11.74 on 4010 degrees of freedom
## Multiple R-squared:  0.01781,    Adjusted R-squared:  0.01756 
## F-statistic: 72.71 on 1 and 4010 DF,  p-value: < 2.2e-16
```

```
confint(fit_model)
```

```
##                  2.5 %     97.5 %
## (Intercept) 64.0559311 67.2241374
## bmi          0.1779501  0.2842121
```

The multiple R-squared could prove that model can account for almost 1.7% variance in pulse rate using BMI as a predictor. This suggests that the independent variable (BMI) has a very weak level of influence in explaining the variation in dependent variable (pulse).

The model can interpreted that for every 1 unit increase in the percentage of time an individual performed moderate activity, there is a 0.1 drop in BMI value. **Model Equation**: `Pulse = 65.64 - 0.23 x BMI (y = 65.64 + 0.23*x)`

At 95% confidence interval, the model has a minimum and maximum value of 0.177 and 0.284 respectively. I am 95% confidence that the actual slope of the regression line falls between 0.177 and 0.284. That is the increase in pulse rate is between 0.177 and 0.284 for each unit increase in BMI.

In conclusion, the model is poor and should not be recommended to predict pulse rate, while using the BMI in this data as a predictor.

TASK 6: Logistics Regression

Here, a logistics regression is applied to examine the relationship between the combination risk factors (Gender, Physical active, Smoking, pulse) on Insomnia.

```
# Fitting a logistics regression model.
model <- glm(insomnia ~ gender + phys_active + smoke100n + pulse +
               bmi, data = work_data, family = binomial("logit"))

summary(model)# Shows model log parameters etc, but not Odds ratios.
```

```
## 
## Call:
## glm(formula = insomnia ~ gender + phys_active + smoke100n + pulse + 
##     bmi, family = binomial("logit"), data = work_data)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.118502   0.277920  -7.623 2.48e-14 ***
## gendermale      -0.513217   0.075638  -6.785 1.16e-11 ***
## phys_activeYes   0.003622   0.074766   0.048    0.961
## smoke100nSmoker  0.662172   0.075023   8.826  < 2e-16 ***
## pulse            0.003413   0.003141   1.087    0.277
## bmi              0.025318   0.005240   4.832 1.35e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4589.6  on 4011  degrees of freedom
## Residual deviance: 4447.7  on 4006  degrees of freedom
## AIC: 4459.7
##
## Number of Fisher Scoring iterations: 4
```

```r
exp(cbind(OR=model$coeff,confint(model))) # Back-transform the coefficients to get Odds Ratios with con
```

```
## Waiting for profiling to be done...
```

```
##                        OR      2.5 %    97.5 %
## (Intercept)     0.1202115 0.06964321 0.2070883
## gendermale      0.5985667 0.51587225 0.6939579
## phys_activeYes  1.0036282 0.86687889 1.1621577
## smoke100nSmoker 1.9389997 1.67438859 2.2469783
## pulse           1.0034187 0.99724568 1.0096036
## bmi             1.0256410 1.01514939 1.0362292
```

```r
anova(model, test = "Chisq") # Table of deviance
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: insomnia
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         4011     4589.6
## gender       1   36.770      4010     4552.8 1.329e-09 ***
## phys_active  1    2.533      4009     4550.3    0.1115
## smoke100n    1   76.568      4008     4473.7 < 2.2e-16 ***
## pulse        1    2.834      4007     4470.9    0.0923 .
## bmi          1   23.115      4006     4447.7 1.526e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
#predict(model, type = "response") %>% head()
```

Model Summary

The coefficient table shows each parameter estimate that is the log odds ratio for each independent variables used for the model. The intercept appears to cut the negative line of the y-axis at -1.88. The p values assess the contribution of these variables individually and they all show a very low p value ($p < 0.05$) except the phys_active group ($p > 0.05$) others seems to be successful in predicting the dependent variable (insonmia) for the 4012 observations in the dataset.

Further more, the residual deviance is greater than the degree of freedom which suggests over dispersion.

Odds Ratio

When the log estimated parameters is back transformed, we get the odds ratio, which can assist in better understanding of the full picture of the epidemiological study.

- **Gender**: The odds of insomnia to occur in males relative to females is low (OR 0.59, 95% Cl 0.51 - 0.68). This suggests that the event is less likely to occur in males than in females. The odds of males having insomnia are 59% less of the odds of females having insomnia.
- **Smoke**: Smoker as relative to Non-smoker shows a higher as a risk of insomnia (OR 1.94, 95% Cl 1.68 - 2.25). This means that the odds of Smokers having insomnia are 94% more of the odds of Non-smokers having insomnia.
- **BMI**: Finally, BMI is a continuous variable here and no levels or factors in form of a group to compare. The odds here is not really interpreted like others but inform of a linear interpretation. So for every unit increase of BMI, there is a 2.5% increase in the odds of having Insomnia. (OR 1.025, 95% 1.015 - 1.036

Physical activity and **Pulse** were both not statistically significant in the model (P>0.05).

BMI, Smoke and Gender variables, appears to have significant effect on the model (P value < 0.05) and narrow confidence intervals which indicates strength of a possible risk factors.

Analysis of Deviance Table

This is useful for models with categorical variable such as this, and the order of fitting is essential as the analysis done here is added sequentially from first to last of the exploratory variables. As seen in the coefficient table, all independent variable appears to be statistically significant ($p < 0.05$) as a risk factor for the outcome variable (insomnia), except Physical active and Pulse variables ($p > 0.05$).

It is necessary to keep in mind that we may not fully conclude on this as the deviance residual is bigger than the residual degrees of freedom for all predictor variables, which could influence the significance of the test. This can still be as a result of the huge amount of outliers detected in the exploratory data analysis.

TASK 7: Decision Tree

Data Partition

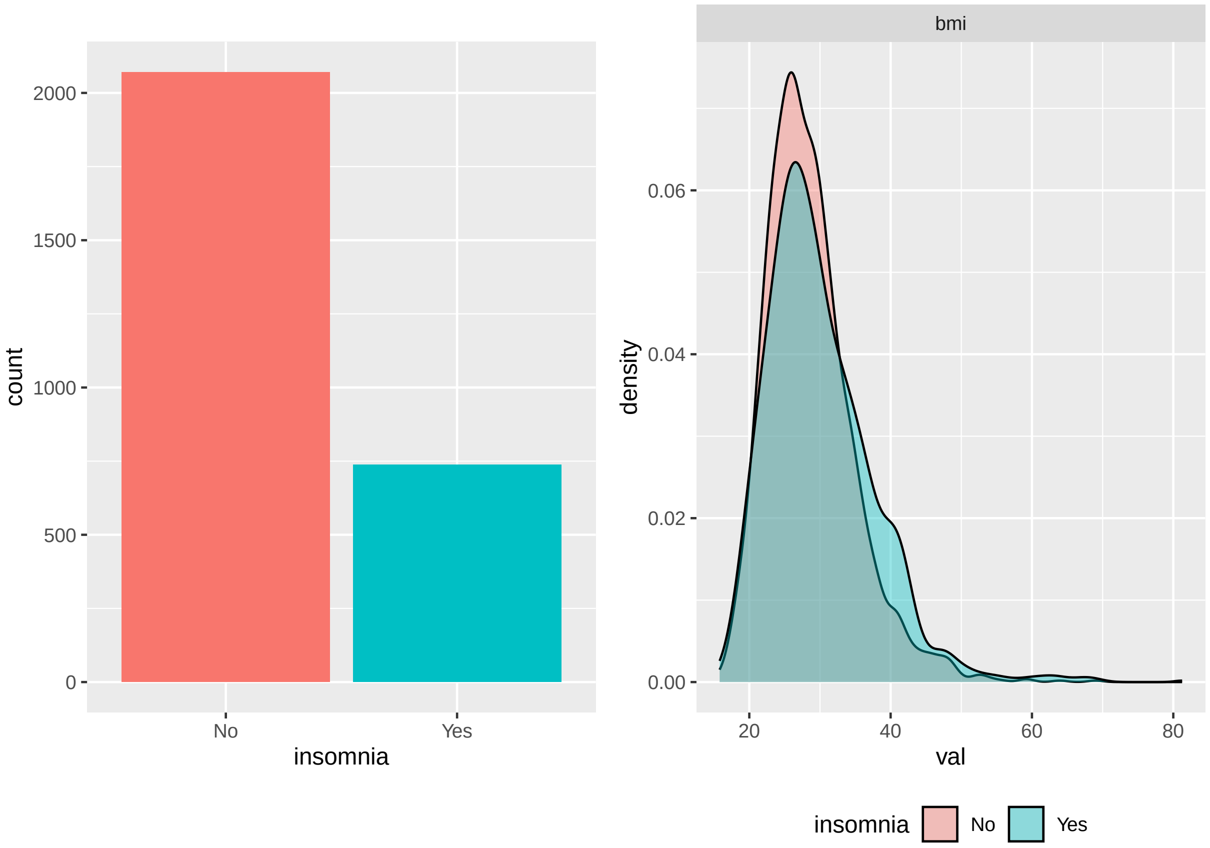
```
# Set random seed for reproducibility
set.seed(223)

dt_data <- work_data %>%
  select(-phys_active, -pulse) #Removing physical activity and pulse (both not significant)

# Data is splited into train and test data
train_data <- dt_data %>% sample_frac(.70)
test_data <- dt_data %>% anti_join(train_data)
```

```
## Joining with ‘by = join_by(insomnia, health_rating, bmi, smoke100n, gender)‘
```

Explore traning dataset

```
ggplot(train_data, aes(x = insomnia, fill = insomnia))+
geom_bar()+theme(legend.position = "none") +

train_data %>%
  gather(var, val, bmi) %>%
  ggplot(aes(val, group = insomnia, fill = insomnia)) +
  geom_density(alpha = 0.4) +
  facet_wrap(~var) +
  theme(legend.position='bottom')+
  plot_annotation(title = "Figure 5: Bmi distribution grouped by Insomnia")
```

Figure 5: Bmi distribution grouped by Insomnia



Model Training Data

```
# Train decision tree on training data
dt <- train(
  insomnia ~ .,       # Use all variables to predict insomnia
  data = train_data, # Use training data
  method = "rpart",  # Recursive partitioning
  trControl = trainControl(method = "cv")  # Cross-validate
)
# Show processing summary
dt
```

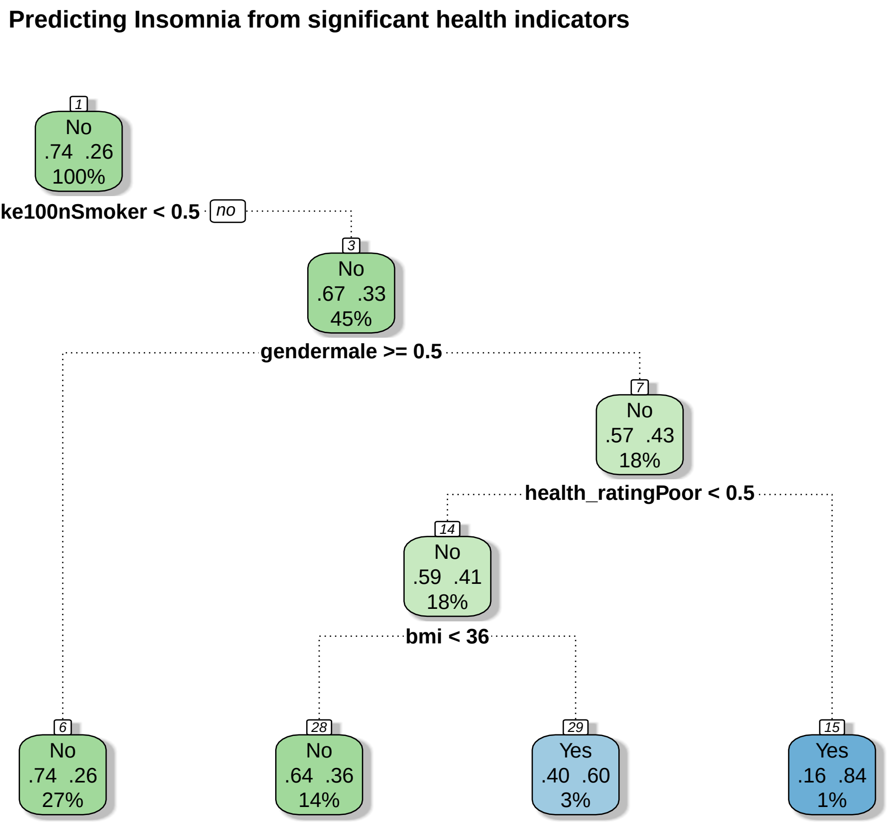
```
## CART
##
## 2808 samples
##    4 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2527, 2527, 2527, 2528, 2527, 2527, ...
## Resampling results across tuning parameters:
##
##   cp           Accuracy   Kappa
##   0.004742547  0.7439514  0.1188875
##   0.008130081  0.7453724  0.1071540
##   0.012195122  0.7400305  0.0327257
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.008130081.
```

```
# Description of final training model
dt$finalModel
```

```
## n= 2808
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 2808 738 No (0.7371795 0.2628205)
##    2) smoke100nSmoker< 0.5 1535 321 No (0.7908795 0.2091205) *
##    3) smoke100nSmoker>=0.5 1273 417 No (0.6724273 0.3275727)
##      6) gendermale>=0.5 755 196 No (0.7403974 0.2596026) *
##      7) gendermale< 0.5 518 221 No (0.5733591 0.4266409)
##       14) health_ratingPoor< 0.5 493 200 No (0.5943205 0.4056795)
##         28) bmi< 35.93 402 145 No (0.6393035 0.3606965) *
##         29) bmi>=35.93 91  36 Yes (0.3956044 0.6043956) *
##       15) health_ratingPoor>=0.5 25   4 Yes (0.1600000 0.8400000) *
```

Ploting decision tree using traind data

```
# Plot final training model
dt$finalModel %>%
  fancyRpartPlot(
    main = "Predicting Insomnia from significant health indicators",
    sub = "Training Data")
```



Training Data

```r
# Predict training set
insomnia_p <- dt %>%  # "predicted"
  predict(newdata = train_data)

# Accuracy of model on training data
table(
  actualclass = train_data$insomnia,
  predictedclass = insomnia_p) %>%
  confusionMatrix() %>%
  print()
```

```
## Confusion Matrix and Statistics
##
##             predictedclass
## actualclass   No  Yes
##         No  2030   40
##         Yes  662   76
##
##                Accuracy : 0.75
##                  95% CI : (0.7336, 0.7659)
##     No Information Rate : 0.9587
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1148
##
##  Mcnemar’s Test P-Value : <2e-16
##
##             Sensitivity : 0.7541
##             Specificity : 0.6552
##          Pos Pred Value : 0.9807
##          Neg Pred Value : 0.1030
##              Prevalence : 0.9587
##          Detection Rate : 0.7229
##    Detection Prevalence : 0.7372
##       Balanced Accuracy : 0.7046
##
##        ’Positive’ Class : No
##
```

The Tree shows 4% of the population are classified as having insomnia as compared to 96% who are classified negative.

Interpretation on Training Data

From the confusion matrix, the model appears to better predict the the negative values as compared to the positive values. As again seen in the training data, this may be due to the unbalance group of observations in the dataset.

On the training data, the model perform relatively well with approximately 75% accuracy [Acc 0.75 95%CI 0.73; 0.77] and statistically significance (P<0.05). Sensitivity is 75%, meaning the measure of a test’s ability to correctly identify those with Insomnia (true positives) is about 75%. Specificity is approximately 66% meaning the measure of a test’s ability to correctly identify those without the condition (true negatives) is about 65%.

Validate on test data

```r
# Predict test set
insomnia_p <- dt %>%
  predict(newdata = test_data)

# Accuracy of model on test data
table(
  actualclass = test_data$insomnia,
  predictedclass = insomnia_p
) %>%
  confusionMatrix() %>%
  print()
```

```
## Confusion Matrix and Statistics
##
##            predictedclass
## actualclass  No Yes
##          No  702  34
##          Yes 261  19
##
##                Accuracy : 0.7096
##                  95% CI : (0.6807, 0.7374)
##     No Information Rate : 0.9478
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0289
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.72897
##             Specificity : 0.35849
##          Pos Pred Value : 0.95380
##          Neg Pred Value : 0.06786
##              Prevalence : 0.94783
##          Detection Rate : 0.69094
##    Detection Prevalence : 0.72441
##       Balanced Accuracy : 0.54373
##
##        'Positive' Class : No
##
```

Interpretation on Test Data

From the confusion matrix, the model appears to better predict the the negative values as compared to the positive values. As again seen in the training data, this may be due to the unbalance group of observations in the dataset.

On the test data, the model perform relatively well with approximately 71% accuracy [Acc 0.71 95%CI 0.68; 0.74] and statistically significance ($P<0.05$). Sensitivity is 73%, meaning the measure of a test's ability to correctly identify those with Insomnia (true positives) is about 73%. Specificity is approximately 36% meaning the measure of a test's ability to correctly identify those without the condition (true negatives) is about 36%. Which is poor.

Relative to the training data, the test data may have almost same general accuracy with the training data but the other statistical tests such as Specificity is poor in the the test data. The balanced accuracy is also lower in the test data at 54% as compared to the training of 70%.

TASK 8: RANDOM FOREST

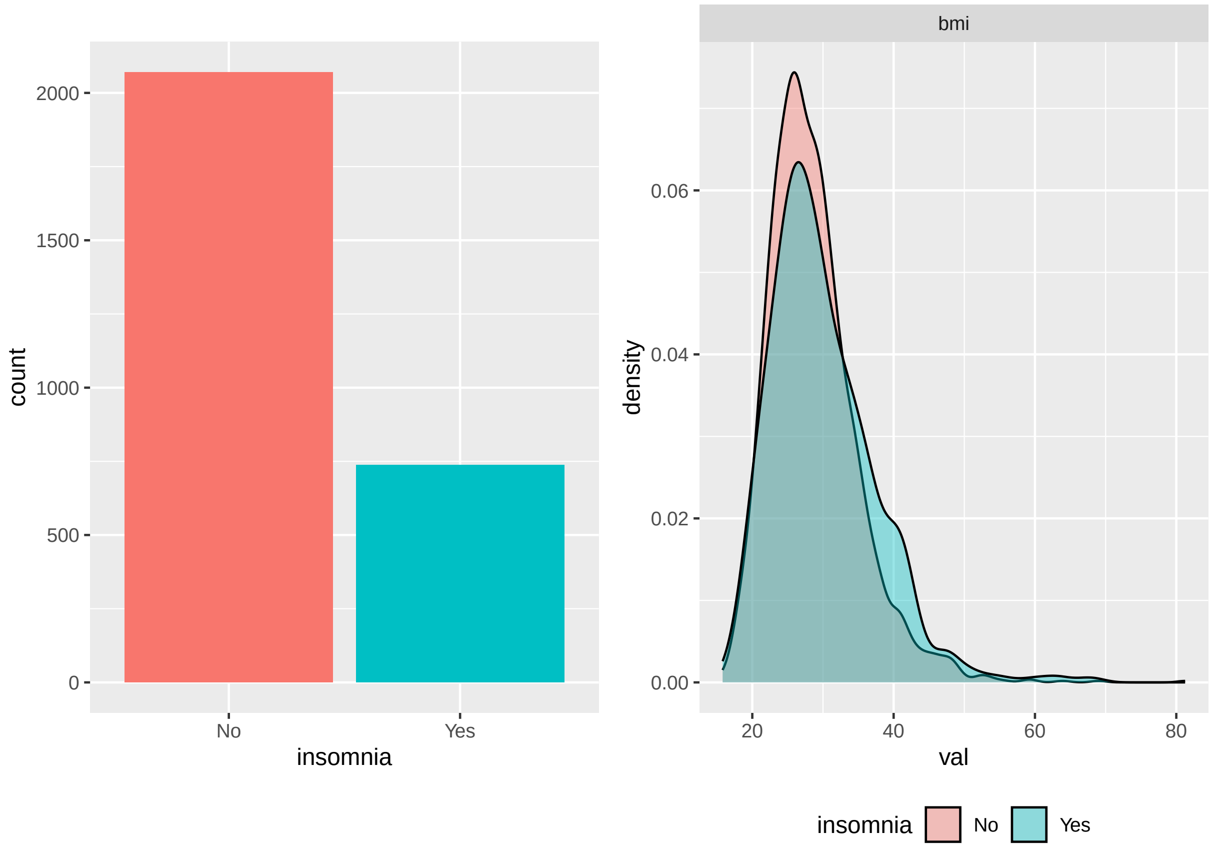
Load and prepare data

```r
# Set random seed
set.seed(223)

# Split data into train and test sets
train <- dt_data %>% sample_frac(.70)
test <- dt_data %>% anti_join(train_data)
```

Explore traning data

```r
ggplot(train, aes(x = insomnia, fill = insomnia))+
geom_bar()+theme(legend.position = "none") +

train %>%
  gather(var, val, bmi) %>%
  ggplot(aes(val, group = insomnia, fill = insomnia)) +
  geom_density(alpha = 0.4) +
  facet_wrap(~var) +
  theme(legend.position='bottom')+
  plot_annotation(title = "Figure 6: Bmi distribution grouped by Insomnia")
```

Figure 6: Bmi distribution grouped by Insomnia

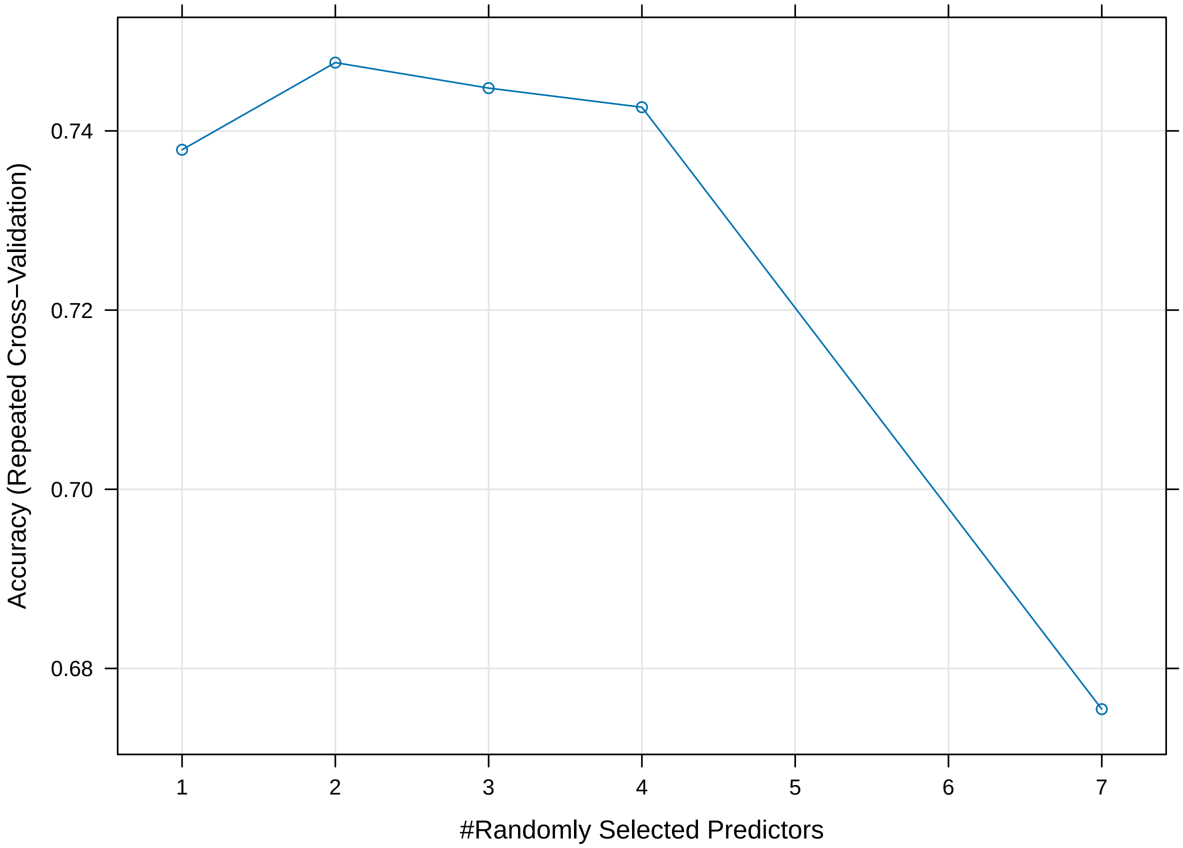


Model traing data

```r
# Define parameters for the training function
control <- trainControl(
  method  = "repeatedcv",  # Repeated cross-validation
  number  = 10,            # Number of folds
  repeats = 3,             # Number of sets of folds
  search  = "random",       # Max number of tuning parameters
  allowParallel = TRUE      # Allow parallel processing
)

# Train Rf model on training data
rf <- train(
  insomnia ~ . ,         # Predict gender from all other vars
  data = train,          # Use training data
  method = "rf",         # Use random forests
  metric = "Accuracy",   # Use accuracy as criterion
  tuneLength = 15,       # Number of levels for parameters
  ntree = 500,           # Number of trees
  trControl = control    # Link to parameters
)

# Show processing summary
rf
```

```
## Random Forest
##
## 2808 samples
##    4 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 2527, 2527, 2527, 2528, 2527, 2527, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##   1     0.7378932  0.006501272
##   2     0.7476241  0.087081279
##   3     0.7447784  0.100190677
##   4     0.7426428  0.111472659
##   7     0.6754546  0.107769561
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

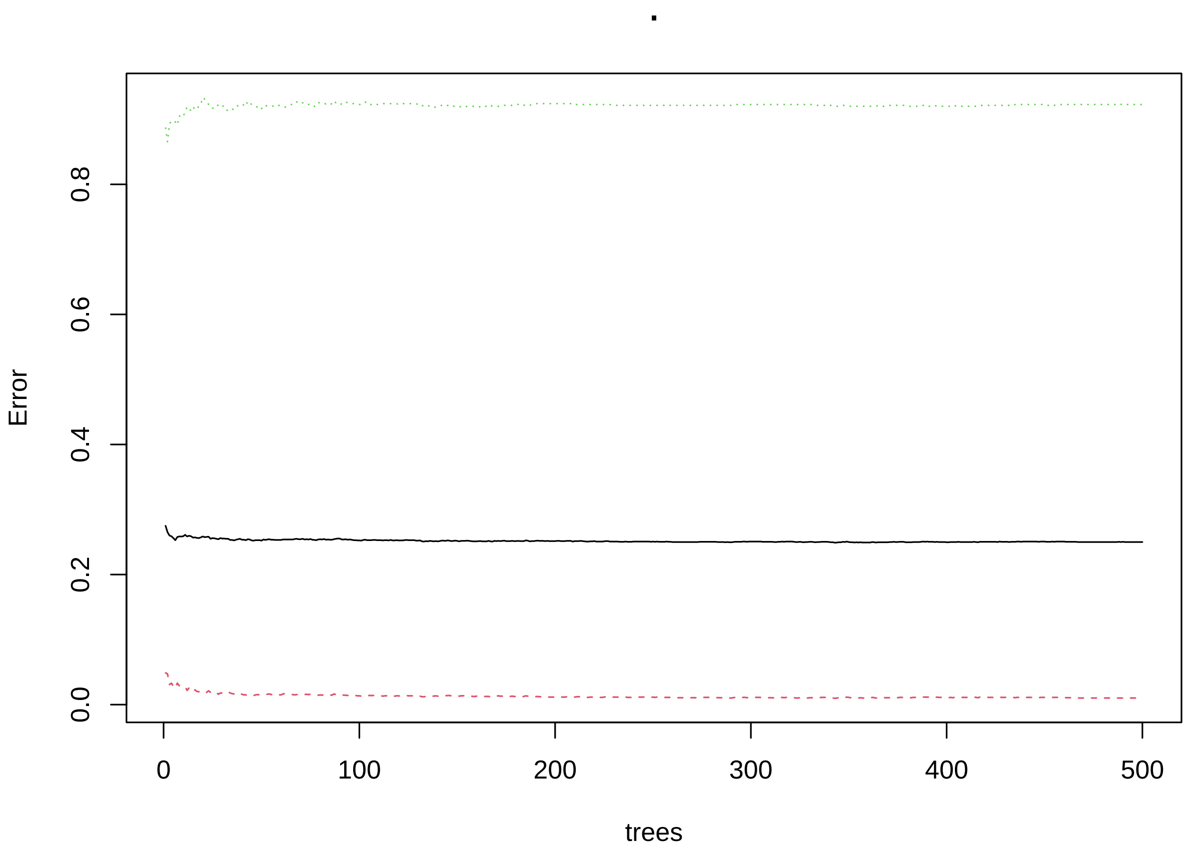
```
# Plot accuracy by number of predictors
rf %>% plot()
```



```
#attributes(rf)

# Accuracy of model with training data
rf$finalModel
```

```
##
## Call:
##  randomForest(x = x, y = y, ntree = 500, mtry = param$mtry)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 25%
## Confusion matrix:
##        No Yes class.error
## No   2049  21  0.01014493
## Yes   681  57  0.92276423
```

```
# Plot error by number of trees
rf$finalModel %>% plot()
```



Red is error for “Insomnia (Yes),”green is error for “Insomnia (No)” and black is error or “OOB,”or “out of bag” (i.e., the probability that any given prediction is not correct within the test data, or the overall accuracy).

Apply model to test data

```r
# Predict test set
insomnia_pred <- rf %>%   # "predicted"
  predict(newdata = test) # Use test data

# Accuracy of model on test data
table(
  actualclass = test$insomnia,
  predictedclass = insomnia_pred
) %>%
  confusionMatrix() %>%
  print()
```

```
## Confusion Matrix and Statistics
##
##              predictedclass
## actualclass  No Yes
##          No  717  19
##          Yes 266  14
##
##                Accuracy : 0.7195
##                  95% CI : (0.6908, 0.7469)
##     No Information Rate : 0.9675
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0333
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7294
##             Specificity : 0.4242
##          Pos Pred Value : 0.9742
##          Neg Pred Value : 0.0500
##              Prevalence : 0.9675
##          Detection Rate : 0.7057
##    Detection Prevalence : 0.7244
##       Balanced Accuracy : 0.5768
##
##        'Positive' Class : No
##
```

Interpretation of Traning Data

The Random Forest summary Resampling results across tuning parameters was approximately 75% for the highest and approximately 68%. Cross validation was done 10 folds and repeated 3 times. Accuracy was used to select the optimal model using the largest value. The final value used for the model was mtry = 2.

The final model shows a 25% out of bag estimate of error ratio. The model showed a 1% error class on classifying people with no insomnia and about 92% error class on those with insomnia. This high error seen in the positive class could be as a result of unbalance observations in the insomnia variable, with the positive group consisting of just 25% as compared to the 75% of the negative group.

Interpretation on Test Data

From the confusion matrix, the model appears to better predict the the negative values as compared to the positive values. As again seen in the training data, this may be due to the unbalance group of observations in the dataset.

On the test data, the model perform relatively well with approximately 72% accuracy [Acc 0.72 95%CI 0.67; 0.74] and statistically significance ($P<0.05$).

Sensitivity is 73%, meaning the measure of a test's ability to correctly identify those with Insomnia (true positives) is about 73%. Specificity is approximately 42% meaning the measure of a test's ability to correctly identify those without the condition (true negatives) is about 42%. The balanced accuracy is at 57%.

TASK 9: Conclusion/Recommendation

Conclusion

In conclusion to the above report:

The Anova test showed no significance difference between v.good and excellent health.

The linear regression model between pulse and BMI is a poor one even though the correlation and model test is statistically significant.

From the logistics regression, it is seen that physical activity and pulse had not significant influence as a risk factor for insomnia.

Comparing the decision tree and random forest model, both model showed a similar accuracy (about 71-72%) for predicting insomnia using other significant variables, but the Random forest model had a slightly higher edge in terms of sensitivity, specificity and balanced accuracy.

Sensitivity is important for ensuring that a condition is not missed (minimizing false negatives). At the same time, specificity is crucial for ensuring that individuals without the condition are not incorrectly diagnosed (minimizing false positives).

Recommendation

Physical activity and pulse rate were insignificant risk factors for insomnia while gender, smoking status, BMI and health rating where potential risk factors.

The odds of developing insomnia is 94% higher in smokers and 41% lower in males compared to females.

The dataset shows some outliers and an inbalance in groups of insomnia. A more balanced dataset could better bring a more balanced classification in the confusion matrix.