Technical report

Exploratory Analysis

```
head(dataset)
```

```
##   HeartDisease   BMI Smoking AlcoholDrinking Stroke PhysicalHealth MentalHealth
## 1          Yes 23.57      No              No     No              0            0
## 2          Yes 32.69     Yes             Yes     No              5            0
## 3          Yes 27.99      No              No     No              0            0
## 4          Yes 19.20      No              No     No              0            0
## 5          Yes 39.84      No              No     No              0            0
## 6          Yes 31.25     Yes              No     No              0            0
##      Sex AgeCategory SleepTime
## 1   Male       65-69         7
## 2   Male       50-54         5
## 3 Female 80 or older         8
## 4 Female 80 or older         6
## 5   Male       65-69        10
## 6   Male       65-69         8
```

Data contains 9 dependent variable out of which BMI, Physical Health, Mental Health and Sleep time are numerical variables. Smoking, Alchohol drinking, Stroke, Sex, AgeCategory are categorical variables.

```
dim(dataset)
```

```
## [1] 5420   10
```

Dimension of the dataset is [5420,10], which indicates there are 5420 records of information in this dataset

```
summary(dataset)
```

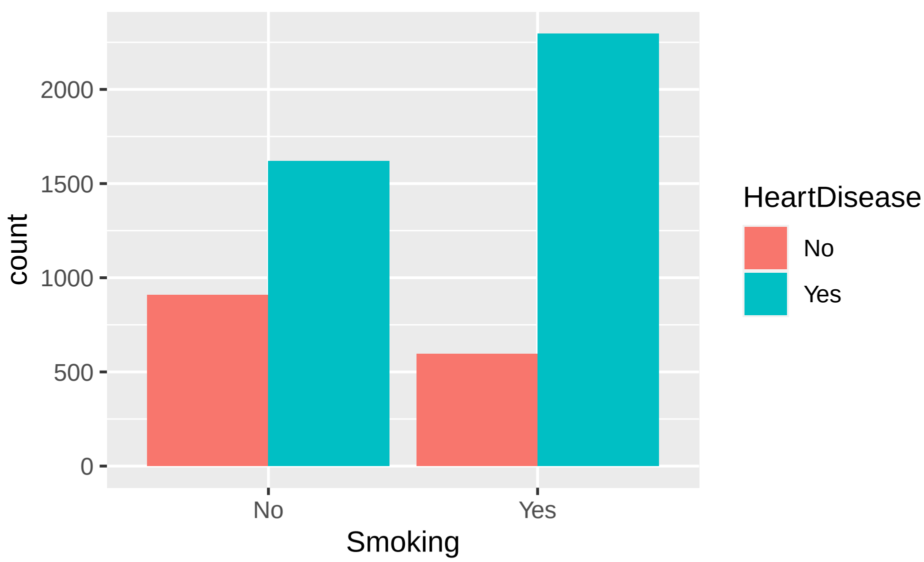
```
##  HeartDisease            BMI          Smoking          AlcoholDrinking   
##  Length:5420        Min.   :12.21   Length:5420        Length:5420       
##  Class :character   1st Qu.:24.67   Class :character   Class :character  
##  Mode  :character   Median :27.89   Mode  :character   Mode  :character  
##                     Mean   :28.98                                        
##                     3rd Qu.:32.27                                        
##                     Max.   :74.33                                        
##     Stroke          PhysicalHealth    MentalHealth        Sex           
##  Length:5420        Min.   : 0.000   Min.   : 0.000   Length:5420       
##  Class :character   1st Qu.: 0.000   1st Qu.: 0.000   Class :character  
##  Mode  :character   Median : 0.000   Median : 0.000   Mode  :character  
##                     Mean   : 6.138   Mean   : 4.321                     
##                     3rd Qu.: 7.000   3rd Qu.: 3.000                     
##                     Max.   :30.000   Max.   :30.000                     
##  AgeCategory          SleepTime     
##  Length:5420        Min.   : 1.000  
##  Class :character   1st Qu.: 6.000  
##  Mode  :character   Median : 7.000  
##                     Mean   : 7.074  
##                     3rd Qu.: 8.000  
##                     Max.   :24.000  
```

This is a brief statistical summary of the dataset. A quick look shows us that average mental health is 4.3 on a scale of 0 to 30, the value being the number of days mental health was not good in last 30 days. This indicates in the given dataset mental health of the people is fairly good. Since the median is not far from the mean, it is likely that there are less influential outliers. BMI on the other hand has a min of 12.21 and max of 74.33, with a mean of 28.98 and closer median. This either shows a wide range of people or there could be substantial outliers. Statistics of remaining variables are as expected with no evident oddities.

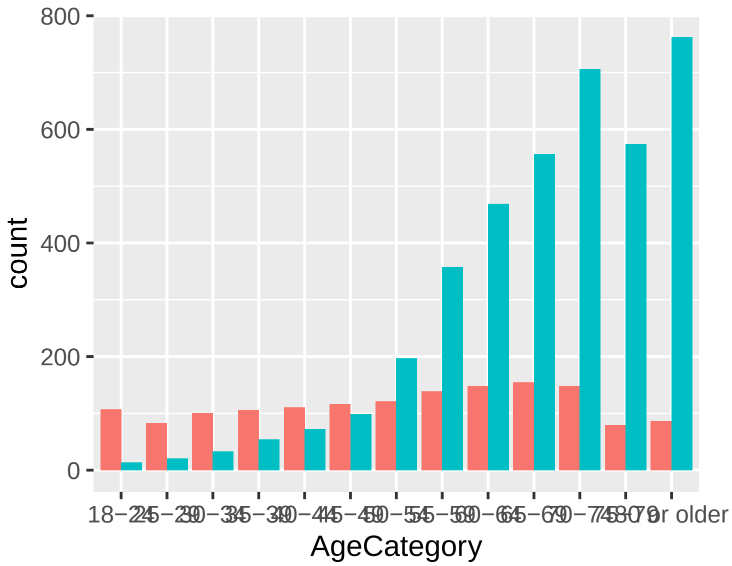
```r
# Converting all categorical variables using factor for further analysis

dataset$HeartDisease<-as.factor(dataset$HeartDisease)
dataset$Smoking<-as.factor(dataset$Smoking)
dataset$AlcoholDrinking<-as.factor(dataset$AlcoholDrinking)
dataset$Stroke<-as.factor(dataset$Stroke)
dataset$Sex<- as.factor(dataset$Sex)
dataset$AgeCategory <- as.factor(dataset$AgeCategory)
```

Plots and Graphs

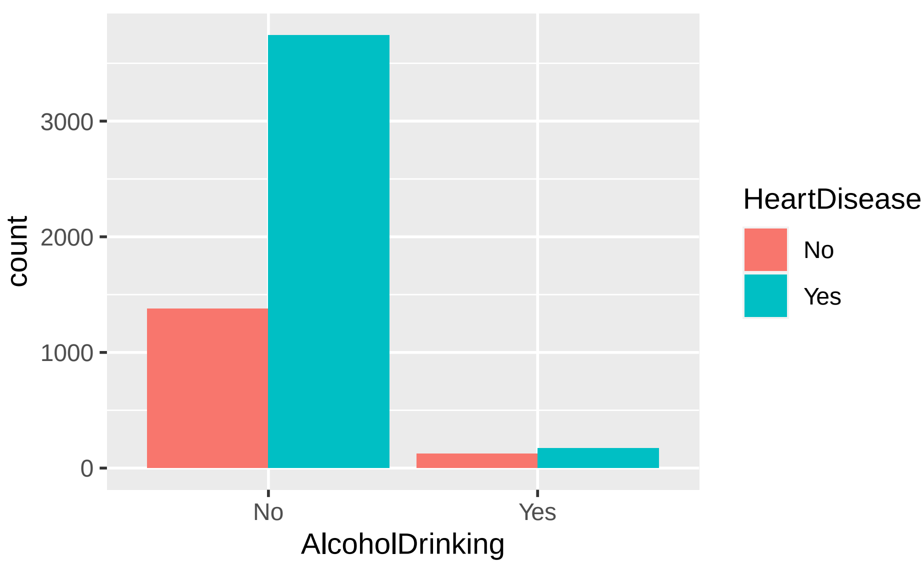
```r
# Code for exploratory analysis here

ggplot(dataset, aes(x = Smoking, fill = HeartDisease)) + geom_bar(position = "dodge")
```



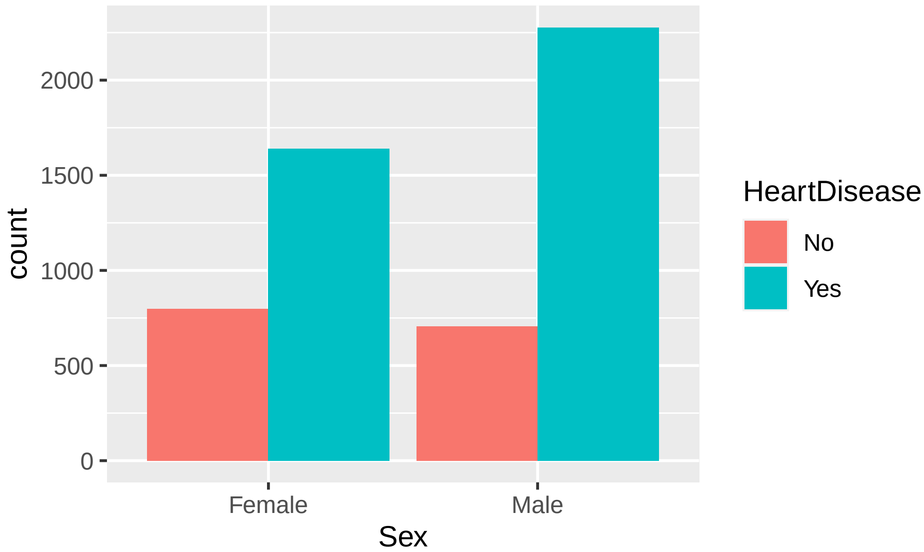

```r
ggplot(dataset, aes(x = AgeCategory, fill = HeartDisease)) + geom_bar(position = "dodge")
```



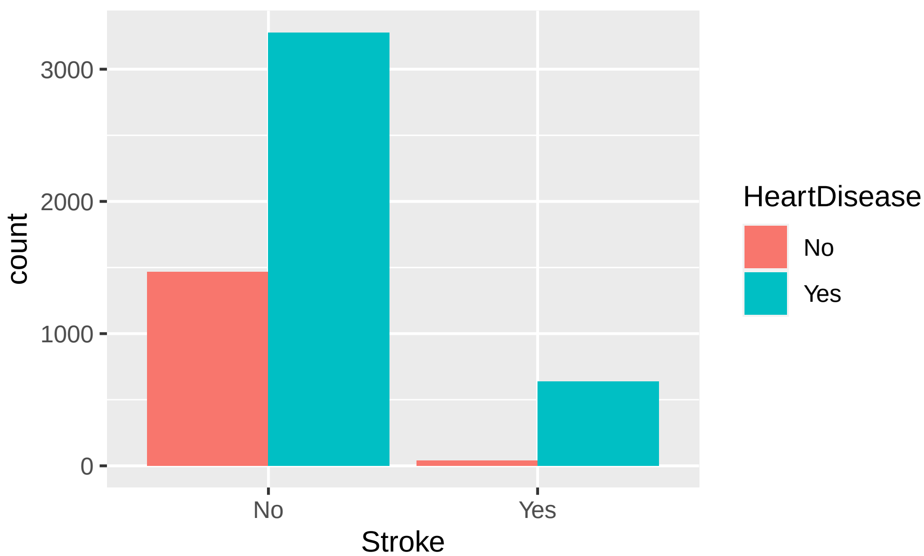

```
ggplot(dataset, aes(x = AlcoholDrinking, fill = HeartDisease)) + geom_bar(position = "dodge")
```




```
ggplot(dataset, aes(x = Sex, fill = HeartDisease)) + geom_bar(position = "dodge")
```




```
ggplot(dataset, aes(x = Stroke, fill = HeartDisease)) + geom_bar(position = "dodge")
```

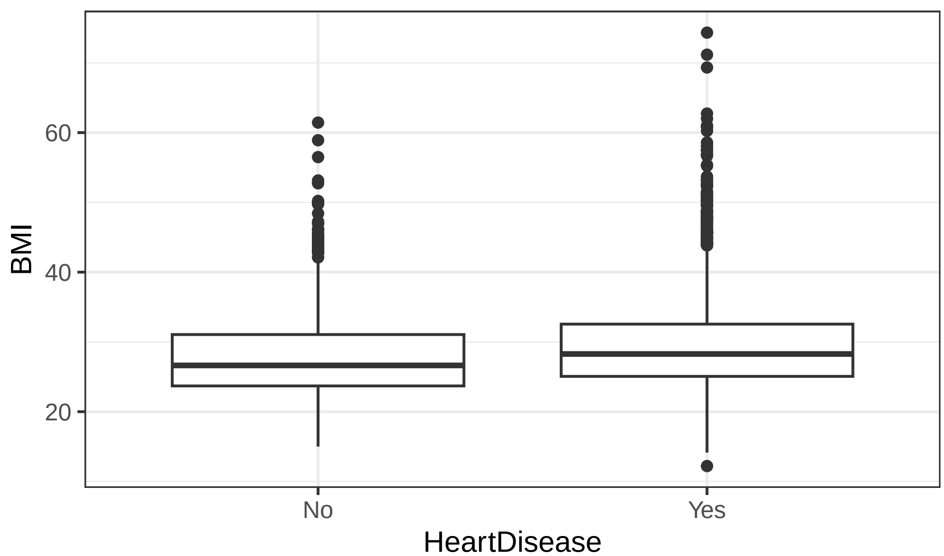


Interpreting bar plots of categorical variables with Heart Disease

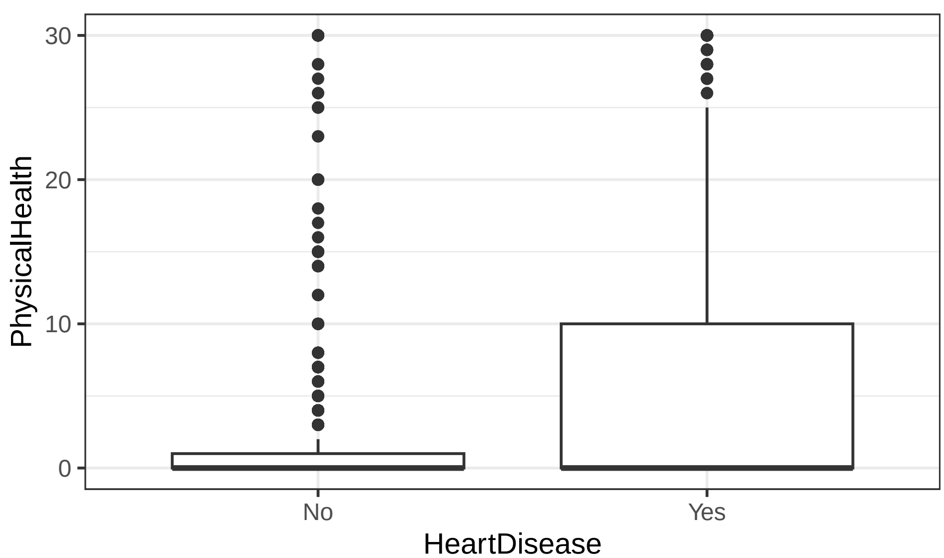
1. Bar plot of smoking vs heart disease tells that out of the people who smoked, there was a higher chance that they had a heart disease than no smoking people. Which shows correlation that people who tend to smoke are more prone to heart disease.
2. Age vs Heart disease plot clearly shows that as people age, chances of having a heart disease increase significantly.
3. In the plot of Alcohol Drinking, it is interesting to see that among the people who drink, there is no signicant difference of having a heart disease than not. Infact in the given sample population, are quite high number of people with heart disease who do not consume alcohol than the people who consume alcohol and have a heart disease. This is contrary to normal intuition. But given than there are very few records for people who drink alchol in the sample, making a conclusion on alcohol consumption alone is dangerous.

4. Another interesting observation is men are more prone to heart diseases than women as can be seen by plot against Sex variable.
5. Plot for Stroke is as predicted normally which is chances of having a disease is significantly higher among the people who have had a stroke before.

```
ggplot(data=dataset, aes(x = HeartDisease, y = BMI)) + geom_boxplot() + theme_bw()
```

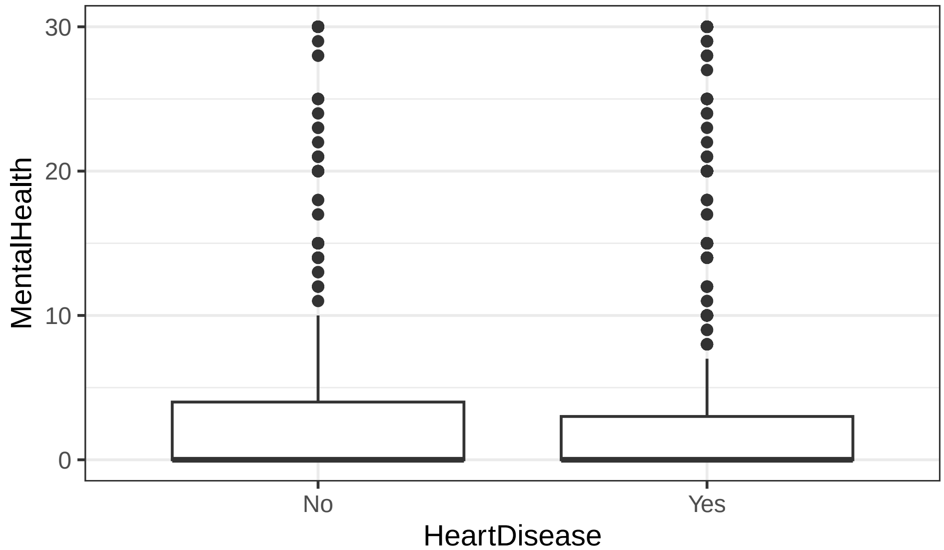



```
ggplot(data=dataset, aes(x = HeartDisease, y = PhysicalHealth)) + geom_boxplot() + theme_bw()
```

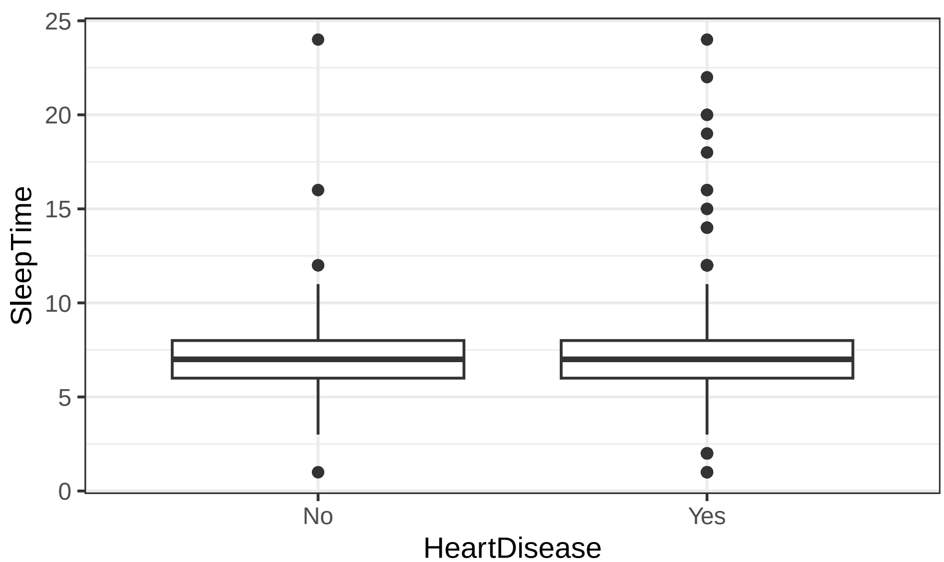



```
ggplot(data=dataset, aes(x = HeartDisease, y = MentalHealth)) + geom_boxplot() + theme_bw()
```

HeartDisease	n	proportion	percentage
No	1503	0.277	27.7
Yes	3917	0.723	72.3




```
ggplot(data=dataset, aes(x = HeartDisease, y = SleepTime)) + geom_boxplot() + theme_bw()
```




```
dataset %>%
  count(HeartDisease) %>%
  mutate(proportion=round(prop.table(n),3), percentage=round(prop.table(n),3)*100) %>%
  kable() %>%
  kable_styling()
```

HeartDisease	min	max	Q1	median	Q3	mean	sd
No	15.00	61.44	23.69	26.63	31.055	27.99293	6.070109
Yes	12.21	74.33	25.06	28.27	32.550	29.36161	6.443873

HeartDisease	min	max	Q1	median	Q3	mean	sd
No	0	30	0	0	1	2.869594	7.219889
Yes	0	30	0	0	10	7.391882	11.268674

```
dataset %>%
  group_by(HeartDisease) %>%
  summarise(min=min(BMI), max=max(BMI), Q1=quantile(BMI, 0.25), median=median(BMI), Q3=quantile(BMI,0.75
  kable()%>%
  kable_styling()
```

We can say that there is some difference in the BMI values between those with heart disease and those without. The mean BMI value for the "Yes" group (29.36161) is slightly higher than the mean BMI value for the "No" group (27.99293). This suggests that there may be a positive relationship between BMI and heart disease.

```
dataset %>%
  group_by(HeartDisease) %>%
  summarise(min=min(PhysicalHealth), max=max(PhysicalHealth), Q1=quantile(PhysicalHealth, 0.25), median=
  kable()%>%
  kable_styling()
```

We can see that the mean physical health score is higher for those with heart disease compared to those without (7.391882 versus 2.869594, respectively). This suggests that there may be a negative relationship between physical health and heart disease, meaning that those with heart disease tend to have lower physical health scores.

```
dataset %>%
  group_by(HeartDisease) %>%
  summarise(min=min(MentalHealth), max=max(MentalHealth), Q1=quantile(MentalHealth, 0.25), median=median
  kable()%>%
  kable_styling()
```

The median value for mental health is the same for both groups (0), while the median value for heart disease is slightly higher (4 for those without heart disease and 3 for those with heart disease). This suggests that individuals without heart disease may have slightly higher levels of heart disease compared to those with heart disease. The mean values for both heart disease and mental health are higher for individuals with heart disease compared to those without, indicating that on average, individuals with heart disease have higher levels of both heart disease and mental health issues compared to those without heart disease.

```
dataset %>%
  group_by(HeartDisease) %>%
  summarise(min=min(SleepTime), max=max(SleepTime), Q1=quantile(SleepTime, 0.25), median=median(SleepTim
```

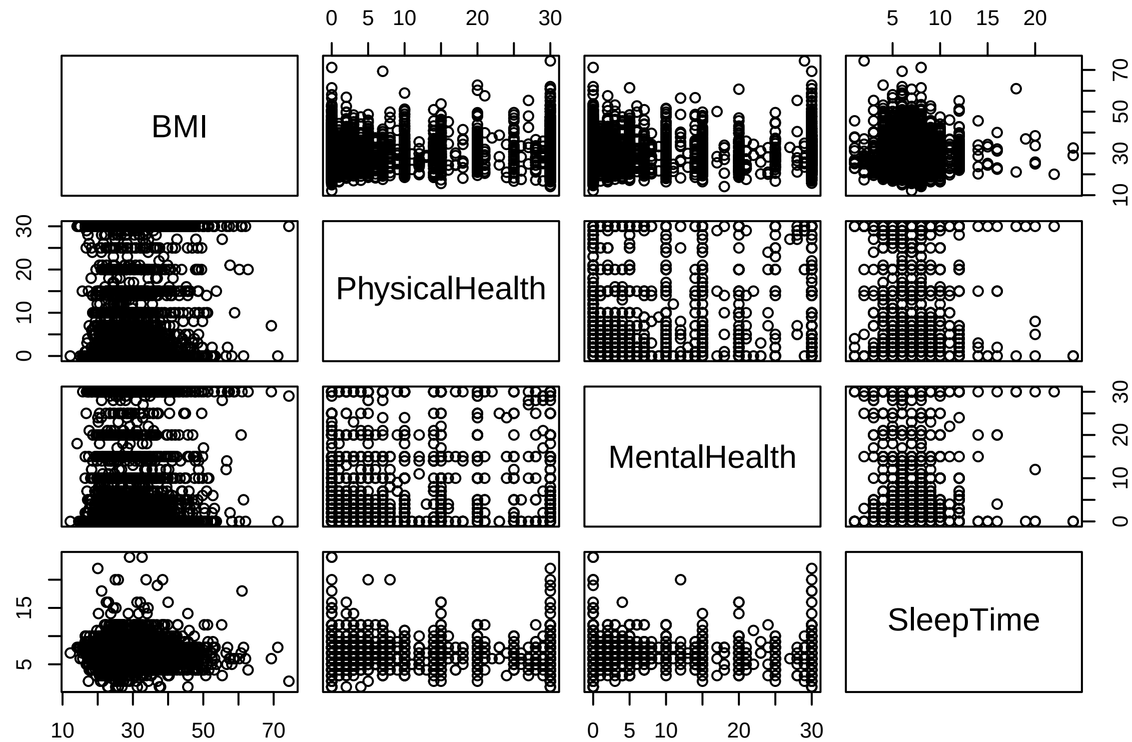
HeartDisease	min	max	Q1	median	Q3	mean	sd
No	0	30	0	0	4	3.971391	8.011384
Yes	0	30	0	0	3	4.454940	9.005748

HeartDisease	min	max	Q1	median	Q3	mean	sd
No	1	24	6	7	8	7.067864	1.381126
Yes	1	24	6	7	8	7.076079	1.782142

```
kable()%>%
kable_styling()
```

Overall, the table suggests that there may not be a strong relationship between heart disease and sleep time, as the descriptive statistics are quite similar between individuals with and without heart disease.

```
pairs(dataset[c("BMI","PhysicalHealth","MentalHealth","SleepTime")])
```



Scatter plot among the numerical variables shows that there is no multicollinearity among the dependent numeric variables. Therefore, there is no need to exclude them while modelling the data.

Conclusions from exploratory analysis

1. There are no nulls in the data. Data is clean.
2. Data is both numeric and categorical.
3. There is no multicollinearity among the numeric variables
4. Sleep time does not have significant impact on weather a person is likely to have a heart disease
5. Positive indicators of heart diseases are Higher BMI, Smoking, Age, Physical health.
6. Some of the variables like Mental Health have large standard deviation and therefore more spread out.

Formal analysis

Logistic Regression

```
# Write code for formal analysis here

# Divide dataset into train and test
sample <- sample(c(TRUE, FALSE), nrow(dataset), replace=TRUE, prob=c(0.8,0.2))
trainset  <- dataset[sample, ]
testset   <- dataset[!sample, ]

summary(trainset)
```

```
##  HeartDisease      BMI          Smoking    AlcoholDrinking Stroke
##  No :1208     Min.   :12.21   No :2050   No :4106        No :3816
##  Yes:3143     1st Qu.:24.60   Yes:2301   Yes: 245        Yes: 535
##               Median :27.89
##               Mean   :28.94
##               3rd Qu.:32.12
##               Max.   :74.33
##
##  PhysicalHealth    MentalHealth        Sex             AgeCategory
##  Min.   : 0.000   Min.   : 0.000   Female:1968   70-74       : 688
##  1st Qu.: 0.000   1st Qu.: 0.000   Male  :2383   80 or older: 675
##  Median : 0.000   Median : 0.000                 65-69       : 576
##  Mean   : 6.094   Mean   : 4.413                 75-79       : 524
##  3rd Qu.: 7.000   3rd Qu.: 3.000                 60-64       : 491
##  Max.   :30.000   Max.   :30.000                 55-59       : 385
##                                                  (Other)     :1012
##    SleepTime
##  Min.   : 1.000
##  1st Qu.: 6.000
##  Median : 7.000
##  Mean   : 7.069
##  3rd Qu.: 8.000
##  Max.   :24.000
##
```

```
summary(testset)
```

```
##  HeartDisease      BMI          Smoking    AlcoholDrinking Stroke
##  No :295      Min.   :15.00   No :478    No :1016        No :927
##  Yes:774      1st Qu.:24.96   Yes:591    Yes:  53        Yes:142
##               Median :28.13
##               Mean   :29.15
##               3rd Qu.:32.49
##               Max.   :71.17
##
##  PhysicalHealth    MentalHealth        Sex            AgeCategory
##  Min.   : 0.000   Min.   : 0.000   Female:469   80 or older:175
##  1st Qu.: 0.000   1st Qu.: 0.000   Male  :600   70-74      :166
##  Median : 0.000   Median : 0.000                65-69      :135
```

```
##  Mean   : 6.317   Mean   : 3.944                75-79      :130
##  3rd Qu.: 7.000   3rd Qu.: 3.000                60-64      :126
##  Max.   :30.000   Max.   :30.000                55-59      :112
##                                                 (Other)    :225
##    SleepTime
##  Min.   : 1.000
##  1st Qu.: 6.000
##  Median : 7.000
##  Mean   : 7.094
##  3rd Qu.: 8.000
##  Max.   :20.000
##
```

```r
# Fitting GLM Model
glmModel = glm(HeartDisease ~ BMI + Smoking + AlcoholDrinking + Stroke + PhysicalHealth + MentalHealth +

summary(glmModel)
```

```
##
## Call:
## glm(formula = HeartDisease ~ BMI + Smoking + AlcoholDrinking +
##     Stroke + PhysicalHealth + MentalHealth + Sex + AgeCategory +
##     SleepTime, family = "binomial", data = trainset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2372  -0.5259   0.4349   0.6757   2.4176
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -3.552272   0.419879  -8.460  < 2e-16 ***
## BMI                    0.047565   0.006900   6.894 5.43e-12 ***
## SmokingYes             0.522478   0.083123   6.286 3.27e-10 ***
## AlcoholDrinkingYes    -0.504123   0.164181  -3.071 0.002137 **
## StrokeYes              1.669929   0.204094   8.182 2.79e-16 ***
## PhysicalHealth         0.040239   0.005355   7.515 5.71e-14 ***
## MentalHealth           0.011672   0.005453   2.140 0.032337 *
## SexMale                0.621513   0.083231   7.467 8.18e-14 ***
## AgeCategory25-29       0.588577   0.416905   1.412 0.158015
## AgeCategory30-34       0.459385   0.401386   1.144 0.252418
## AgeCategory35-39       0.822324   0.381280   2.157 0.031025 *
## AgeCategory40-44       1.204496   0.366046   3.291 0.001000 ***
## AgeCategory45-49       1.329140   0.361536   3.676 0.000237 ***
## AgeCategory50-54       2.022993   0.347953   5.814 6.10e-09 ***
## AgeCategory55-59       2.532799   0.341846   7.409 1.27e-13 ***
## AgeCategory60-64       2.754643   0.339030   8.125 4.47e-16 ***
## AgeCategory65-69       2.968826   0.336602   8.820  < 2e-16 ***
## AgeCategory70-74       3.359255   0.337297   9.959  < 2e-16 ***
## AgeCategory75-79       3.733117   0.349420  10.684  < 2e-16 ***
## AgeCategory80 or older  4.259569   0.349558  12.186  < 2e-16 ***
## SleepTime             -0.061315   0.025916  -2.366 0.017985 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5140.3  on 4350  degrees of freedom
## Residual deviance: 3813.1  on 4330  degrees of freedom
## AIC: 3855.1
##
## Number of Fisher Scoring iterations: 5
```

Summary tables shows all the variables are significant. Most significant variable in numeric type is 'Stroke'. Since summary doesn't give a good analysis of significant variables in categorical type. Let us use Anova next.

```
anova(glmModel, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: HeartDisease
##
## Terms added sequentially (first to last)
##
##
##                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                             4350     5140.3
## BMI              1    40.17      4349     5100.2 2.328e-10 ***
## Smoking          1   125.21      4348     4975.0 < 2.2e-16 ***
## AlcoholDrinking  1    37.37      4347     4937.6 9.760e-10 ***
## Stroke           1   177.61      4346     4760.0 < 2.2e-16 ***
## PhysicalHealth   1   111.87      4345     4648.1 < 2.2e-16 ***
## MentalHealth     1    23.55      4344     4624.6 1.217e-06 ***
## Sex              1    30.94      4343     4593.6 2.664e-08 ***
## AgeCategory     12   774.90      4331     3818.7 < 2.2e-16 ***
## SleepTime        1     5.59      4330     3813.1   0.01807 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at anova, it is clear that all the categorical variables are significant. Top significant categorical variables in this glm model are Smoking, Stroke, Age Category.

```
# Predicting test values, confusion matrix, calculating accuracy
predicts = predict(glmModel,testset, type = "response")
classifications = ifelse(predicts > 0.5, "Yes", "No")

# Confusion matrix
table(classifications, testset$HeartDisease)
```

```
##
## classifications  No Yes
##             No  123  40
##             Yes 172 734
```

```r
#classification rate or accuracy
(151+782)/(151+782+59+165)
```

```
## [1] 0.8063959
```

80.6% of observations are correctly classified and the test error rate is 19.4%]

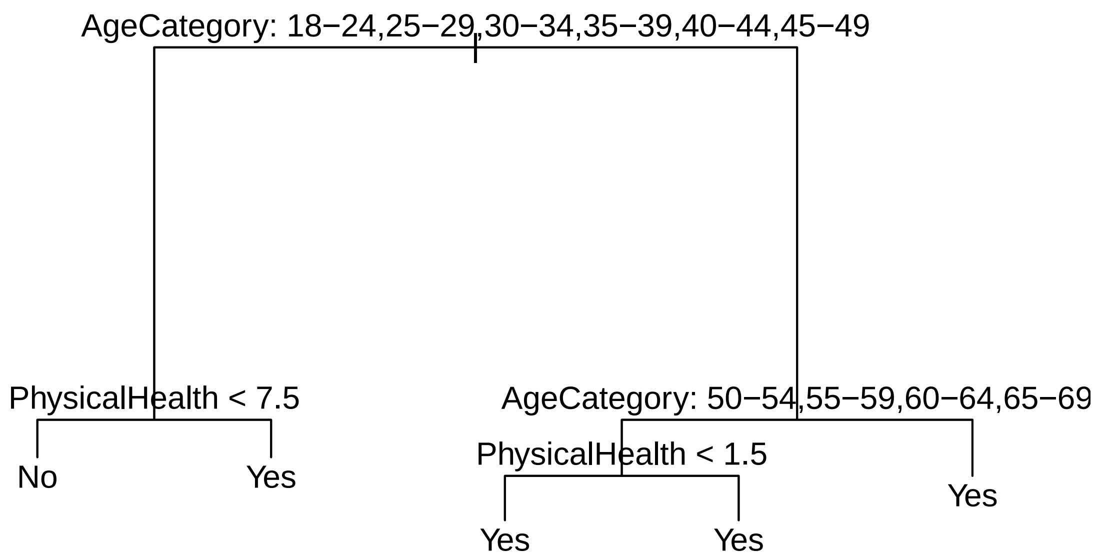
Classification Tree

```r
ClassificationTree<-tree(HeartDisease~ BMI + Smoking + AlcoholDrinking + Stroke + PhysicalHealth + Menta

summary(ClassificationTree)
```

```
##
## Classification tree:
## tree(formula = HeartDisease ~ BMI + Smoking + AlcoholDrinking +
##     Stroke + PhysicalHealth + MentalHealth + Sex + AgeCategory +
##     SleepTime, data = trainset)
## Variables actually used in tree construction:
## [1] "AgeCategory"    "PhysicalHealth"
## Number of terminal nodes:  5
## Residual mean deviance:  0.966 = 4198 / 4346
## Misclassification error rate: 0.2043 = 889 / 4351
```

Misclassification error rate: 0.204 = 885 / 4338 or the training error in the model is 20.4%

```r
plot(ClassificationTree)
text(ClassificationTree,pretty=0)
```



From the tree plot it is gathered that top levels are the most influential variables for the decision of this classificiation model. In this model, they are: Age Category, Physical Health and Stroke, Age Category being the most impactful factor for heart disease.

```
ClassificationTree
```

```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
##  1) root 4351 5140.0 Yes ( 0.2776 0.7224 )
##    2) AgeCategory: 18-24,25-29,30-34,35-39,40-44,45-49 758  943.3 No ( 0.6860 0.3140 )
##      4) PhysicalHealth < 7.5 649  735.9 No ( 0.7458 0.2542 ) *
##      5) PhysicalHealth > 7.5 109  138.3 Yes ( 0.3303 0.6697 ) *
##    3) AgeCategory: 50-54,55-59,60-64,65-69,70-74,75-79,80 or older 3593 3509.0 Yes ( 0.1915 0.8085 )
##      6) AgeCategory: 50-54,55-59,60-64,65-69 1706 1960.0 Yes ( 0.2614 0.7386 )
##       12) PhysicalHealth < 1.5 973 1252.0 Yes ( 0.3433 0.6567 ) *
##       13) PhysicalHealth > 1.5 733  626.8 Yes ( 0.1528 0.8472 ) *
##      7) AgeCategory: 70-74,75-79,80 or older 1887 1446.0 Yes ( 0.1282 0.8718 ) *
```

```
PredictTreeClass <- predict(ClassificationTree, testset , type = "class")
table (PredictTreeClass, testset$HeartDisease)
```

```
##
## PredictTreeClass  No Yes
##              No   98  33
##              Yes 197 741
```

```
(101+747)/(101+747+40+194)
```

```
## [1] 0.7837338
```

Accuracy of this model is 78.3% with testing error being 21.7%

Conclusions

In this assignment two models were designed to predict the likelihood of heart disease in a person based on various factors such as age, physical health, and previous stroke. The first model uses logistic regression. The anova analysis shows that variables like Smoking, Stroke, and Age Category, are significant in predicting heart disease. 80.6% of observations are correctly classified and the test error rate is 19.4%

The second model is a classification tree with an accuracy of 78.3% and a testing error of 21.7%. The tree plot reveals that Age Category, Physical Health, and Stroke are the most important variables in the decision-making process. Age Category is the most influential factor for heart disease prediction.

In summary, both models use different techniques to predict the likelihood of heart disease in a person based on various factors. The logistic regression model shows that Smoking, Stroke, and Age Category are significant variables in predicting heart disease, while the classification tree model emphasizes the importance of Age Category, Physical Health, and Stroke in predicting heart disease. Relative accuracy is close to each for both models, but logistic regression performed better by a tiny fraction.

Non-technical report

The purpose of this report was to analyze a dataset that predicts the likelihood of a person having heart disease based on various factors such as sleep time, age, physical health, stroke, smoking, and alcohol consumption. The initial step involved ensuring that the dataset was clean without any missing data. The exploratory analysis was then performed to evaluate each independent variable's correlation with heart disease using bar plots for categorical data and box plots for numeric data. Multicollinearity was also checked among numeric variables, and no significant issues were found.

After verifying that the data was free of significant outliers and other issues, two models were used to fit the data: logistic regression and classification tree. Logistic regression is a classification model, and classification tree is a decision tree. The logistic model was fitted, and the confusion matrix, accuracy, and testing error were calculated, with a resulting test error of 19.4%. The classification tree model was also fitted, and the test error was found to be 21.7%.

Since both models had similar test errors, it is difficult to determine which model is better suited for predicting heart disease. However, both models are expected to perform well in classifying the likelihood of heart disease in a person.