Final Project

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

Overview

For this semester's final project, I will be using a heart risk dataset from Kaggle to help determine traits that contribute to whether someone is at risk for a heart attack.

Taken from the source:

> This dataset provides a comprehensive array of features relevant to heart health and lifestyle choices, encompassing patient-specific details such as age, gender, cholesterol levels, blood pressure, heart rate, and indicators like diabetes, family history, smoking habits, obesity, and alcohol consumption. Additionally, lifestyle factors like exercise hours, dietary habits, stress levels, and sedentary hours are included. Medical aspects comprising previous heart problems, medication usage, and triglyceride levels are considered. Socioeconomic aspects such as income and geographical attributes like country, continent, and hemisphere are incorporated. The dataset, consisting of 8763 records from patients around the globe, culminates in a crucial binary classification feature denoting the presence or absence of a heart attack risk, providing a comprehensive resource for predictive analysis and research in cardiovascular health.

Fields in the dataset are:

- Patient ID - Unique identifier for each patient
- Age - Age of the patient
- Sex - Gender of the patient (Male/Female)
- Cholesterol - Cholesterol levels of the patient
- Blood Pressure - Blood pressure of the patient (systolic/diastolic)
- Heart Rate - Heart rate of the patient
- Diabetes - Whether the patient has diabetes (Yes/No)
- Family History - Family history of heart-related problems (1: Yes, 0: No)
- Smoking - Smoking status of the patient (1: Smoker, 0: Non-smoker)
- Obesity - Obesity status of the patient (1: Obese, 0: Not obese)
- Alcohol Consumption - Level of alcohol consumption by the patient (None/Light/Moderate/Heavy)
- Exercise Hours Per Week - Number of exercise hours per week
- Diet - Dietary habits of the patient (Healthy/Average/Unhealthy)
- Previous Heart Problems - Previous heart problems of the patient (1: Yes, 0: No)
- Medication Use - Medication usage by the patient (1: Yes, 0: No)
- Stress Level - Stress level reported by the patient (1-10)
- Sedentary Hours Per Day - Hours of sedentary activity per day
- Income - Income level of the patient
- BMI - Body Mass Index (BMI) of the patient
- Triglycerides - Triglyceride levels of the patient
- Physical Activity Days Per Week - Days of physical activity per week
- Sleep Hours Per Day - Hours of sleep per day

	vars	n	mean	sd	median	trimmed	mad	min	
Patient.ID*	1	8763	4382.00	2529.80	4382.00	4382.00	3248.38	1	
Age	2	8763	53.71	21.25	54.00	53.62	26.69	18	
Sex*	3	8763	1.70	0.46	2.00	1.75	0.00	1	
Cholesterol	4	8763	259.88	80.86	259.00	259.96	102.30	120	
Blood.Pressure*	5	8763	1946.34	1130.45	1941.00	1944.71	1447.02	1	
Heart.Rate	6	8763	75.02	20.55	75.00	75.03	26.69	40	
Diabetes	7	8763	0.65	0.48	1.00	0.69	0.00	0	
Family.History	8	8763	0.49	0.50	0.00	0.49	0.00	0	
Smoking	9	8763	0.90	0.30	1.00	1.00	0.00	0	
Obesity	10	8763	0.50	0.50	1.00	0.50	0.00	0	
Alcohol.Consumption	11	8763	0.60	0.49	1.00	0.62	0.00	0	
Exercise.Hours.Per.Week	12	8763	10.01	5.78	10.07	10.03	7.45	0	
Diet*	13	8763	2.00	0.81	2.00	2.00	1.48	1	
Previous.Heart.Problems	14	8763	0.50	0.50	0.00	0.49	0.00	0	
Medication.Use	15	8763	0.50	0.50	0.00	0.50	0.00	0	
Stress.Level	16	8763	5.47	2.86	5.00	5.47	2.97	1	
Sedentary.Hours.Per.Day	17	8763	5.99	3.47	5.93	5.99	4.48	0	
Income	18	8763	158263.18	80575.19	157866.00	157964.45	103461.76	20062	29
BMI	19	8763	28.89	6.32	28.77	28.86	8.07	18	
Triglycerides	20	8763	417.68	223.75	417.00	417.83	286.14	30	
Physical.Activity.Days.Per.Week	21	8763	3.49	2.28	3.00	3.49	2.97	0	
Sleep.Hours.Per.Day	22	8763	7.02	1.99	7.00	7.03	2.97	4	
Country*	23	8763	10.38	5.79	10.00	10.36	7.41	1	
Continent*	24	8763	3.43	1.60	4.00	3.41	2.97	1	
Hemisphere*	25	8763	1.35	0.48	1.00	1.32	0.00	1	
Heart.Attack.Risk	26	8763	0.36	0.48	0.00	0.32	0.00	0	

- Country - Country of the patient
- Continent - Continent where the patient resides
- Hemisphere - Hemisphere where the patient resides
- Heart Attack Risk - Presence of heart attack risk (1: Yes, 0: No) // TARGET VARIABLE

Data Exploration

First, we'll view the summary and then we'll check if there are data points missing. Then, we'll clean the fields up to make sure they're ready for analysis.

```
training <- read.csv('https://raw.githubusercontent.com/addsding/data621/main/project/heart_attack_predi

summary <- as.data.frame(describe(training))
nulls <- 8763 - summary['n']
nulls_pct <- nulls / 12795
summary['nulls'] <- nulls
summary['nulls_pct'] <- nulls_pct
kable(summary, digits=2) |>
  kable_styling(c("striped", "scale_down")) |>
  scroll_box(width = "100%")
```

There are 8763 observations and a total of 26 variables in this dataset.

Overall, the data looks relatively clean – means and medians are somewhat close together as well, signalling a normal distribution.

Luckily, it looks like there is no missing information.

One issue observed is the `Alcohol.Consumption` field – the description of the dataset has this as a categorical variable with more than 2 options, however the dataset presents it as a binary 0 or 1. This has been noted and will be interpreted now as whether someone indulges in alcohol regularly.

What types of fields are each of our variables?

```
summary(training)
```

```
##   Patient.ID             Age            Sex             Cholesterol
##  Length:8763        Min.   :18.00   Length:8763        Min.   :120.0
##  Class :character   1st Qu.:35.00   Class :character   1st Qu.:192.0
##  Mode  :character   Median :54.00   Mode  :character   Median :259.0
##                     Mean   :53.71                      Mean   :259.9
##                     3rd Qu.:72.00                      3rd Qu.:330.0
##                     Max.   :90.00                      Max.   :400.0
##  Blood.Pressure       Heart.Rate        Diabetes      Family.History
##  Length:8763        Min.   : 40.00   Min.   :0.0000   Min.   :0.000
##  Class :character   1st Qu.: 57.00   1st Qu.:0.0000   1st Qu.:0.000
##  Mode  :character   Median : 75.00   Median :1.0000   Median :0.000
##                     Mean   : 75.02   Mean   :0.6523   Mean   :0.493
##                     3rd Qu.: 93.00   3rd Qu.:1.0000   3rd Qu.:1.000
##                     Max.   :110.00   Max.   :1.0000   Max.   :1.000
##     Smoking          Obesity       Alcohol.Consumption Exercise.Hours.Per.Week
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000      Min.   : 0.002442
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000      1st Qu.: 4.981579
##  Median :1.0000   Median :1.0000   Median :1.0000      Median :10.069559
##  Mean   :0.8968   Mean   :0.5014   Mean   :0.5981      Mean   :10.014284
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000      3rd Qu.:15.050018
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000      Max.   :19.998709
##      Diet           Previous.Heart.Problems Medication.Use    Stress.Level
##  Length:8763        Min.   :0.0000          Min.   :0.0000   Min.   : 1.00
##  Class :character   1st Qu.:0.0000          1st Qu.:0.0000   1st Qu.: 3.00
##  Mode  :character   Median :0.0000          Median :0.0000   Median : 5.00
##                     Mean   :0.4958          Mean   :0.4983   Mean   : 5.47
##                     3rd Qu.:1.0000          3rd Qu.:1.0000   3rd Qu.: 8.00
##                     Max.   :1.0000          Max.   :1.0000   Max.   :10.00
##  Sedentary.Hours.Per.Day     Income            BMI        Triglycerides
##  Min.   : 0.001263       Min.   : 20062   Min.   :18.00   Min.   : 30.0
##  1st Qu.: 2.998794       1st Qu.: 88310   1st Qu.:23.42   1st Qu.:225.5
##  Median : 5.933622       Median :157866   Median :28.77   Median :417.0
##  Mean   : 5.993690       Mean   :158263   Mean   :28.89   Mean   :417.7
##  3rd Qu.: 9.019124       3rd Qu.:227749   3rd Qu.:34.32   3rd Qu.:612.0
##  Max.   :11.999313       Max.   :299954   Max.   :40.00   Max.   :800.0
##  Physical.Activity.Days.Per.Week Sleep.Hours.Per.Day   Country
##  Min.   :0.00                    Min.   : 4.000      Length:8763
##  1st Qu.:2.00                    1st Qu.: 5.000      Class :character
##  Median :3.00                    Median : 7.000      Mode  :character
##  Mean   :3.49                    Mean   : 7.024
##  3rd Qu.:5.00                    3rd Qu.: 9.000
##  Max.   :7.00                    Max.   :10.000
```

```
##   Continent          Hemisphere        Heart.Attack.Risk
##  Length:8763        Length:8763        Min.   :0.0000
##  Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :0.0000
##                                        Mean   :0.3582
##                                        3rd Qu.:1.0000
##                                        Max.   :1.0000
```

It looks like there's a good mix of continuous and categorical data, however some of these character fields will need to be converted into factors and for blood pressure, that field will need to be split between systolic and diastolic.

Data Cleaning

Blood Pressure To do this, `dplyr` has a nice functionality to separate columns.

```
training <- training |>
  separate(Blood.Pressure, sep='/', c('Systolic', 'Diastolic'))

training$Systolic <- as.numeric(training$Systolic)
training$Diastolic <- as.numeric(training$Diastolic)

summary(training)
```

```
##   Patient.ID             Age            Sex             Cholesterol
##  Length:8763        Min.   :18.00   Length:8763        Min.   :120.0
##  Class :character   1st Qu.:35.00   Class :character   1st Qu.:192.0
##  Mode  :character   Median :54.00   Mode  :character   Median :259.0
##                     Mean   :53.71                      Mean   :259.9
##                     3rd Qu.:72.00                      3rd Qu.:330.0
##                     Max.   :90.00                      Max.   :400.0
##     Systolic       Diastolic        Heart.Rate        Diabetes
##  Min.   : 90.0   Min.   : 60.00   Min.   : 40.00   Min.   :0.0000
##  1st Qu.:112.0   1st Qu.: 72.00   1st Qu.: 57.00   1st Qu.:0.0000
##  Median :135.0   Median : 85.00   Median : 75.00   Median :1.0000
##  Mean   :135.1   Mean   : 85.16   Mean   : 75.02   Mean   :0.6523
##  3rd Qu.:158.0   3rd Qu.: 98.00   3rd Qu.: 93.00   3rd Qu.:1.0000
##  Max.   :180.0   Max.   :110.00   Max.   :110.00   Max.   :1.0000
##  Family.History     Smoking          Obesity       Alcohol.Consumption
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   :0.493   Mean   :0.8968   Mean   :0.5014   Mean   :0.5981
##  3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  Exercise.Hours.Per.Week     Diet           Previous.Heart.Problems
##  Min.   : 0.002442       Length:8763        Min.   :0.0000
##  1st Qu.: 4.981579       Class :character   1st Qu.:0.0000
##  Median :10.069559       Mode  :character   Median :0.0000
##  Mean   :10.014284                          Mean   :0.4958
##  3rd Qu.:15.050018                          3rd Qu.:1.0000
##  Max.   :19.998709                          Max.   :1.0000
##  Medication.Use    Stress.Level   Sedentary.Hours.Per.Day     Income
```

```
##  Min.   :0.0000   Min.   : 1.00   Min.   : 0.001263       Min.   : 20062
##  1st Qu.:0.0000   1st Qu.: 3.00   1st Qu.: 2.998794       1st Qu.: 88310
##  Median :0.0000   Median : 5.00   Median : 5.933622       Median :157866
##  Mean   :0.4983   Mean   : 5.47   Mean   : 5.993690       Mean   :158263
##  3rd Qu.:1.0000   3rd Qu.: 8.00   3rd Qu.: 9.019124       3rd Qu.:227749
##  Max.   :1.0000   Max.   :10.00   Max.   :11.999313       Max.   :299954
##       BMI        Triglycerides   Physical.Activity.Days.Per.Week
##  Min.   :18.00   Min.   : 30.0   Min.   :0.00
##  1st Qu.:23.42   1st Qu.:225.5   1st Qu.:2.00
##  Median :28.77   Median :417.0   Median :3.00
##  Mean   :28.89   Mean   :417.7   Mean   :3.49
##  3rd Qu.:34.32   3rd Qu.:612.0   3rd Qu.:5.00
##  Max.   :40.00   Max.   :800.0   Max.   :7.00
##  Sleep.Hours.Per.Day   Country           Continent          Hemisphere
##  Min.   : 4.000      Length:8763        Length:8763        Length:8763
##  1st Qu.: 5.000      Class :character   Class :character   Class :character
##  Median : 7.000      Mode  :character   Mode  :character   Mode  :character
##  Mean   : 7.024
##  3rd Qu.: 9.000
##  Max.   :10.000
##  Heart.Attack.Risk
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3582
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

Now with blood pressure broken down, next is factorizing the categorical fields.

Data Types

The fields to be changed are:

- Sex
- Diet
- Country
- Continent
- Hemisphere

```
training$Sex <- as.factor(training$Sex)
training$Diet <- as.factor(training$Diet)
training$Country <- as.factor(training$Country)
training$Continent <- as.factor(training$Continent)
training$Hemisphere <- as.factor(training$Hemisphere)

summary(training)
```

```
##   Patient.ID            Age            Sex        Cholesterol
##  Length:8763        Min.   :18.00   Female:2652   Min.   :120.0
##  Class :character   1st Qu.:35.00   Male  :6111   1st Qu.:192.0
##  Mode  :character   Median :54.00                 Median :259.0
##                     Mean   :53.71                 Mean   :259.9
```

```
##                     3rd Qu.:72.00                 3rd Qu.:330.0
##                     Max.   :90.00                 Max.   :400.0
##
##     Systolic       Diastolic        Heart.Rate        Diabetes
##  Min.   : 90.0   Min.   : 60.00   Min.   : 40.00   Min.   :0.0000
##  1st Qu.:112.0   1st Qu.: 72.00   1st Qu.: 57.00   1st Qu.:0.0000
##  Median :135.0   Median : 85.00   Median : 75.00   Median :1.0000
##  Mean   :135.1   Mean   : 85.16   Mean   : 75.02   Mean   :0.6523
##  3rd Qu.:158.0   3rd Qu.: 98.00   3rd Qu.: 93.00   3rd Qu.:1.0000
##  Max.   :180.0   Max.   :110.00   Max.   :110.00   Max.   :1.0000
##
##  Family.History     Smoking          Obesity       Alcohol.Consumption
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   :0.493   Mean   :0.8968   Mean   :0.5014   Mean   :0.5981
##  3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##  Exercise.Hours.Per.Week        Diet      Previous.Heart.Problems
##  Min.   : 0.002442       Average  :2912   Min.   :0.0000
##  1st Qu.: 4.981579       Healthy  :2960   1st Qu.:0.0000
##  Median :10.069559       Unhealthy:2891   Median :0.0000
##  Mean   :10.014284                        Mean   :0.4958
##  3rd Qu.:15.050018                        3rd Qu.:1.0000
##  Max.   :19.998709                        Max.   :1.0000
##
##  Medication.Use    Stress.Level   Sedentary.Hours.Per.Day     Income
##  Min.   :0.0000   Min.   : 1.00   Min.   : 0.001263       Min.   : 20062
##  1st Qu.:0.0000   1st Qu.: 3.00   1st Qu.: 2.998794       1st Qu.: 88310
##  Median :0.0000   Median : 5.00   Median : 5.933622       Median :157866
##  Mean   :0.4983   Mean   : 5.47   Mean   : 5.993690       Mean   :158263
##  3rd Qu.:1.0000   3rd Qu.: 8.00   3rd Qu.: 9.019124       3rd Qu.:227749
##  Max.   :1.0000   Max.   :10.00   Max.   :11.999313       Max.   :299954
##
##       BMI        Triglycerides   Physical.Activity.Days.Per.Week
##  Min.   :18.00   Min.   : 30.0   Min.   :0.00
##  1st Qu.:23.42   1st Qu.:225.5   1st Qu.:2.00
##  Median :28.77   Median :417.0   Median :3.00
##  Mean   :28.89   Mean   :417.7   Mean   :3.49
##  3rd Qu.:34.32   3rd Qu.:612.0   3rd Qu.:5.00
##  Max.   :40.00   Max.   :800.0   Max.   :7.00
##
##  Sleep.Hours.Per.Day           Country            Continent
##  Min.   : 4.000      Germany       : 477   Africa       : 873
##  1st Qu.: 5.000      Argentina     : 471   Asia         :2543
##  Median : 7.000      Brazil        : 462   Australia    : 884
##  Mean   : 7.024      United Kingdom: 457   Europe       :2241
##  3rd Qu.: 9.000      Australia     : 449   North America: 860
##  Max.   :10.000      Nigeria       : 448   South America:1362
##                      (Other)       :5999
##                Hemisphere   Heart.Attack.Risk
##  Northern Hemisphere:5660   Min.   :0.0000
##  Southern Hemisphere:3103   1st Qu.:0.0000
```

	vars	n	mean	sd	median	trimmed	mad	min	
Patient.ID*	1	8763	4382.00	2529.80	4382.00	4382.00	3248.38	1	
Age	2	8763	53.71	21.25	54.00	53.62	26.69	18	
Sex*	3	8763	1.70	0.46	2.00	1.75	0.00	1	
Cholesterol	4	8763	259.88	80.86	259.00	259.96	102.30	120	
Systolic	5	8763	135.08	26.35	135.00	135.09	34.10	90	
Diastolic	6	8763	85.16	14.68	85.00	85.21	19.27	60	
Heart.Rate	7	8763	75.02	20.55	75.00	75.03	26.69	40	
Diabetes	8	8763	0.65	0.48	1.00	0.69	0.00	0	
Family.History	9	8763	0.49	0.50	0.00	0.49	0.00	0	
Smoking	10	8763	0.90	0.30	1.00	1.00	0.00	0	
Obesity	11	8763	0.50	0.50	1.00	0.50	0.00	0	
Alcohol.Consumption	12	8763	0.60	0.49	1.00	0.62	0.00	0	
Exercise.Hours.Per.Week	13	8763	10.01	5.78	10.07	10.03	7.45	0	
Diet*	14	8763	2.00	0.81	2.00	2.00	1.48	1	
Previous.Heart.Problems	15	8763	0.50	0.50	0.00	0.49	0.00	0	
Medication.Use	16	8763	0.50	0.50	0.00	0.50	0.00	0	
Stress.Level	17	8763	5.47	2.86	5.00	5.47	2.97	1	
Sedentary.Hours.Per.Day	18	8763	5.99	3.47	5.93	5.99	4.48	0	
Income	19	8763	158263.18	80575.19	157866.00	157964.45	103461.76	20062	29
BMI	20	8763	28.89	6.32	28.77	28.86	8.07	18	
Triglycerides	21	8763	417.68	223.75	417.00	417.83	286.14	30	
Physical.Activity.Days.Per.Week	22	8763	3.49	2.28	3.00	3.49	2.97	0	
Sleep.Hours.Per.Day	23	8763	7.02	1.99	7.00	7.03	2.97	4	
Country*	24	8763	10.38	5.79	10.00	10.36	7.41	1	
Continent*	25	8763	3.43	1.60	4.00	3.41	2.97	1	
Hemisphere*	26	8763	1.35	0.48	1.00	1.32	0.00	1	
Heart.Attack.Risk	27	8763	0.36	0.48	0.00	0.32	0.00	0	

```
##                          Median :0.0000
##                          Mean   :0.3582
##                          3rd Qu.:1.0000
##                          Max.   :1.0000
##
```

```
summary <- as.data.frame(describe(training))
nulls <- 8763 - summary['n']
nulls_pct <- nulls / 8161
summary['nulls'] <- nulls
summary['nulls_pct'] <- nulls_pct
kable(summary, digits=2) |>
  kable_styling(c("striped", "scale_down")) |>
  scroll_box(width = "100%")
```

Class Bias Check

For a binary logistic regression model, there are only two target values: 0 and 1. Ideally, there should be an equal representation of both because if imbalance were to deviate, model performance would suffer from effects of differential variance between the classes and bias, thus picking the more represented class. For logistic regression, if there is a strong imbalance, we can:

- up-sample the smaller group (e.g. bootstrapping),
- down-sample the larger group (e.g. sampling or bootstrapping)
- adjust our threshold for assigning the predicted value away from 0.5.

What is the exact distribution of `Heart.Attack.Risk`?

```
table(training$Heart.Attack.Risk)
```

```
##
##    0    1
## 5624 3139
```

Looks like 0 is more heavily present here – we'll have to up-sample the smaller group here.

```
set.seed(123)
training <- upSample(x=training[, -ncol(training)],
                     y=as.factor(training$Heart.Attack.Risk))

table(training$Class)
```

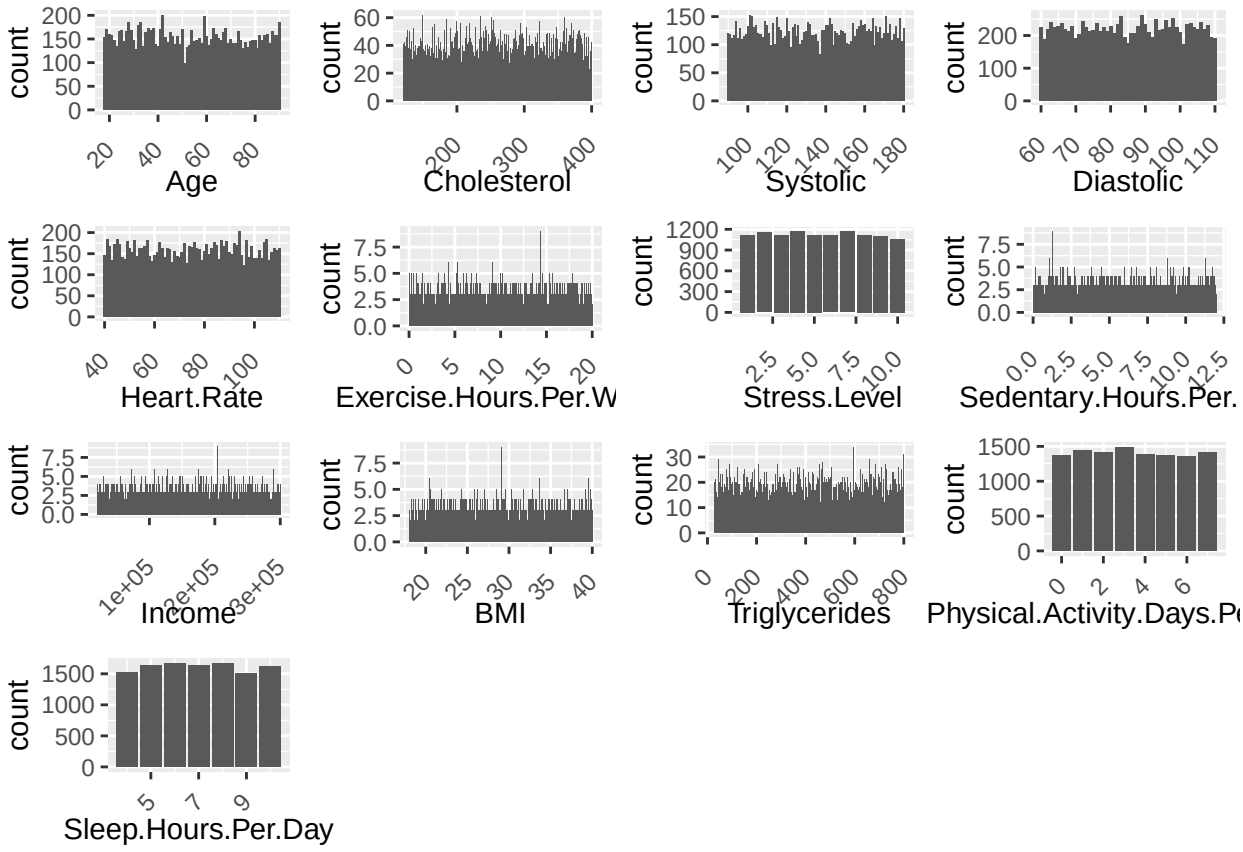
```
##
##    0    1
## 5624 5624
```

Perfect 50/50 split now!

Distributions

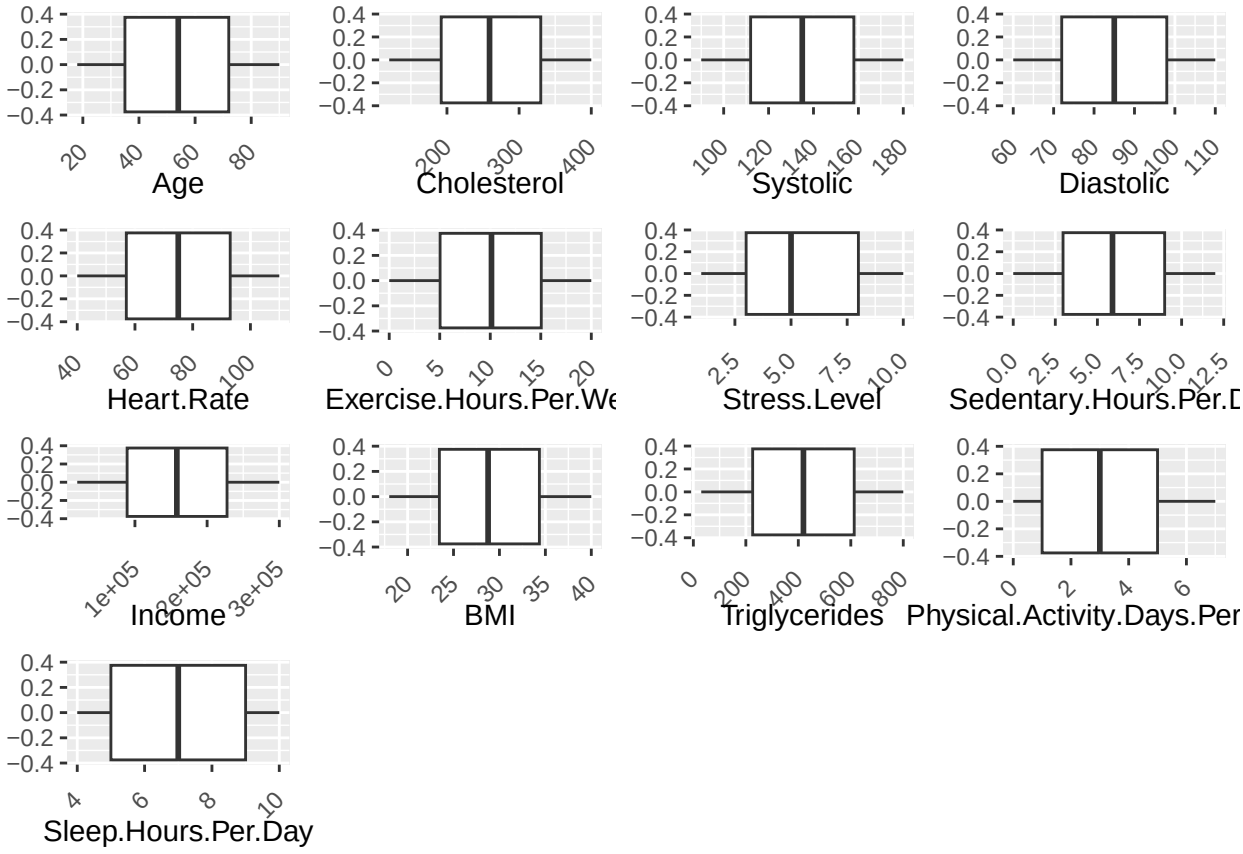
Numerical Fields

Let's see what all of the wnumerical fields look like distribution wise.



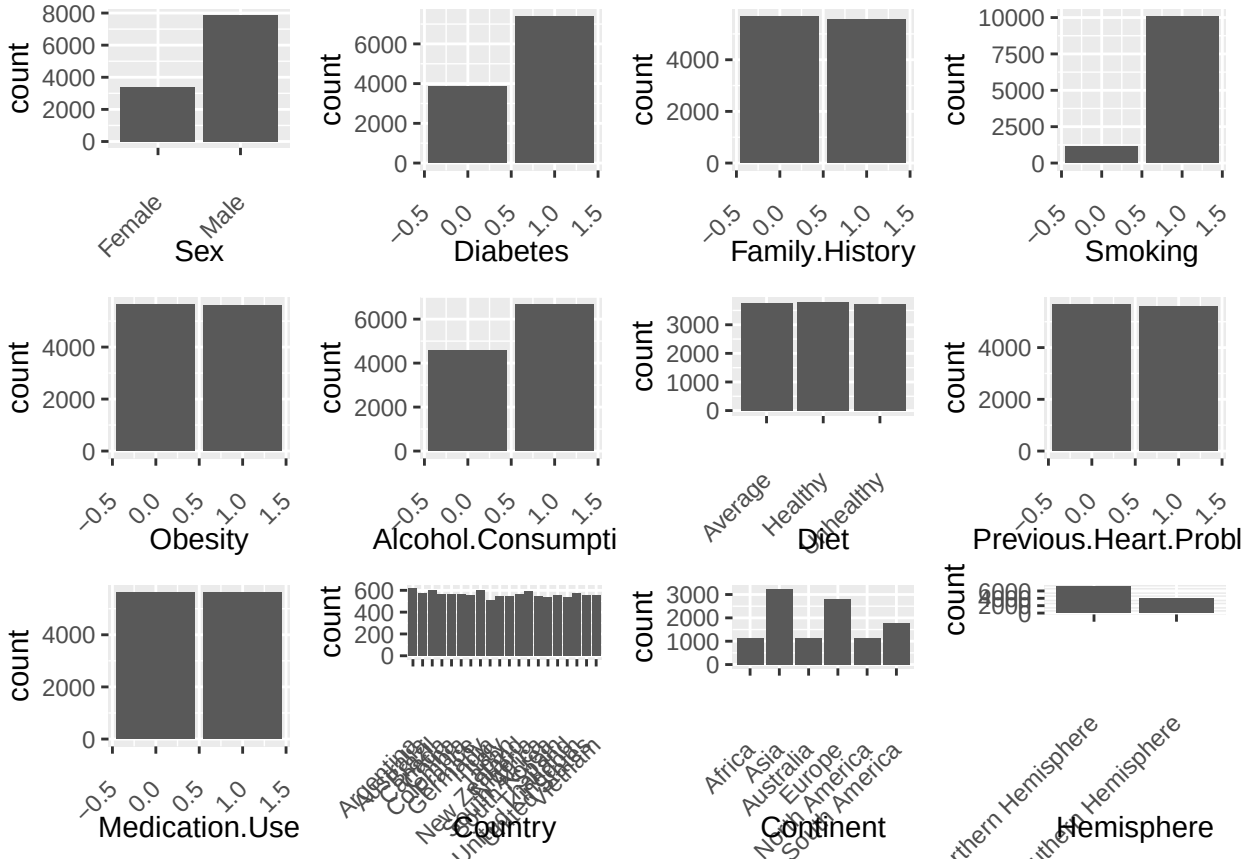
At first glance, none of these distributions look normal unfortunately.

How do these look as boxplots?



At the very least, these distributions don't seem to have many if any outliers!

Categorical Fields

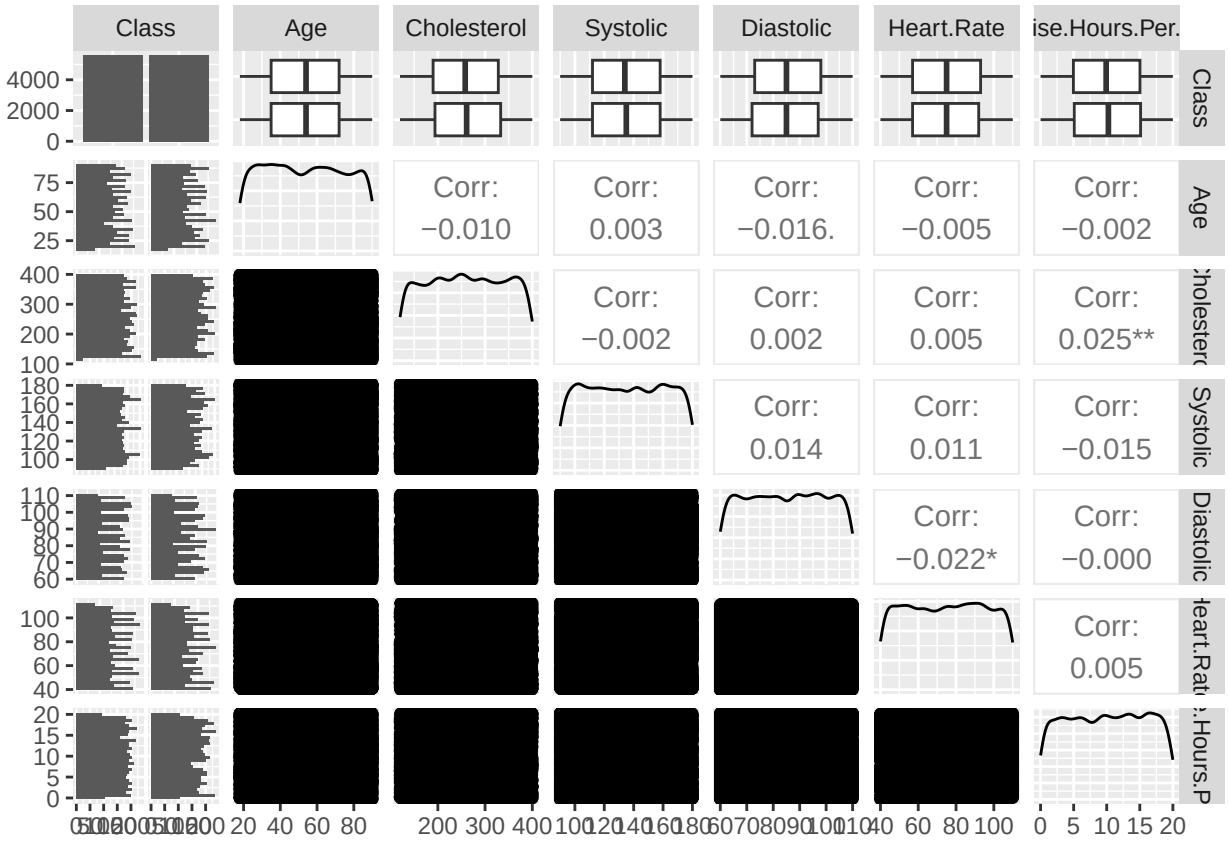


Interestingly, these don't look too equal in distribution – particularly, there seems to be a lot more males than female sin this dataset when ideally, that'd be a 50/50 split. Smoking and Diabetes are other examples of disproportionate fields, however Sex is something we'd expect to be qual.

Now that we have a sense of how the data is distributed, what do the relationships between the variables as well as with our target look like?

Correlations

Let's see how each of the numerical fields correlate with `Heart.Attack.Risk` – we'll start with the first six fields.

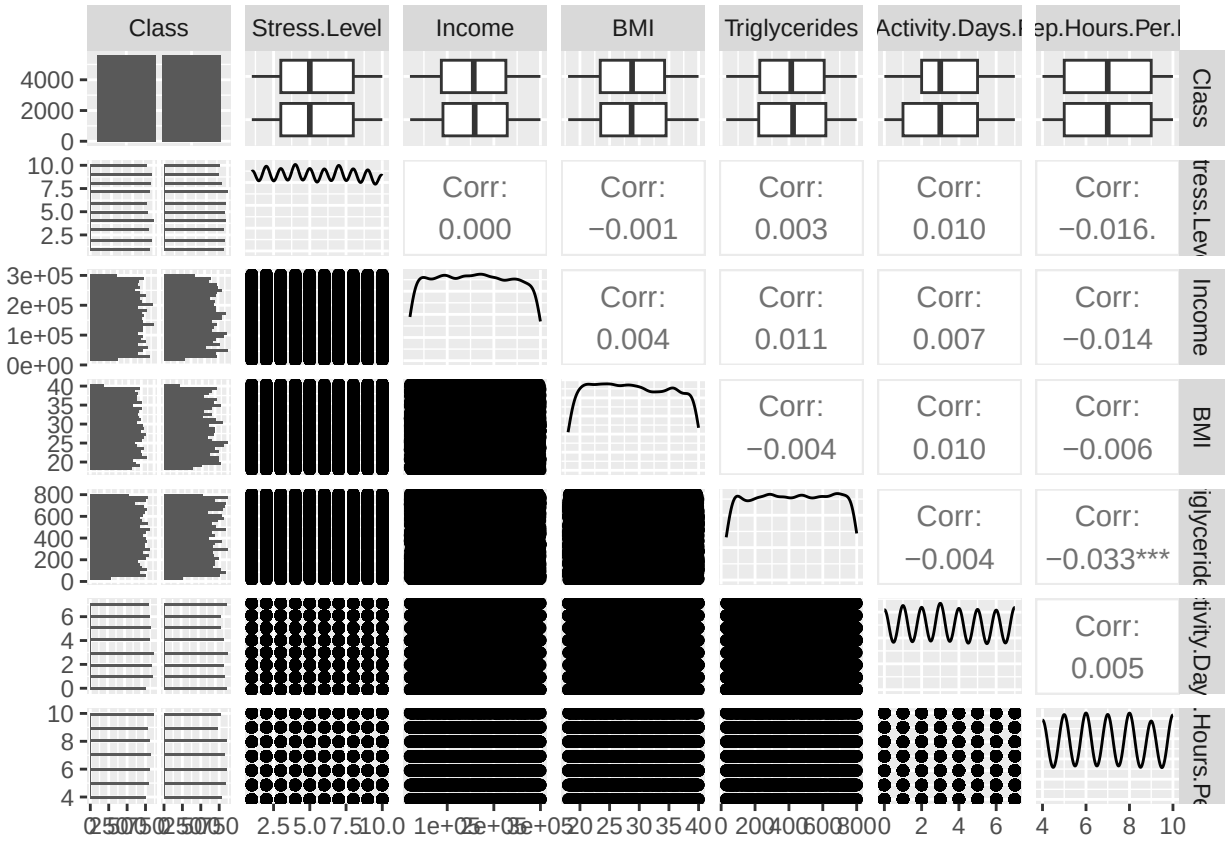


Interestingly, only one of the fields was statistically significant in correlation and that was `Heart.Rate`.

Implication wise:

- `Age` - not correlated - this says age is not a good indicator of whether someone is at risk for a heart attack; this is surprising as younger people are typically healthier than those who are older
- `Cholesterol` - not correlated - this says that high or low cholesterol does not impact heart attack risk; this is surprising as you'd think high cholesterol is a sign of health issues
- `Systolic` - not correlated - this says having high or low systolic blood pressure does not impact heart attack risk; this is surprising higher blood pressure could lead to health issues
- `Diastolic` - not correlated - this says having high or low diastolic blood pressure does not impact heart attack risk; this is surprising higher blood pressure could lead to health issues
- `Heart.Rate` - negative effect - this says that the lower the heart rate, the less likely you are to be at risk; this is not surprising as lower heart rates would mean the heart is working less and thus isn't under as much stress
- `Exercise.Hours.Per.Week` - not correlated - this says exercising or not does not impact heart attack risk; this is surprising as you'd think fitter people are less likely to have health issues

What about the rest of the fields?

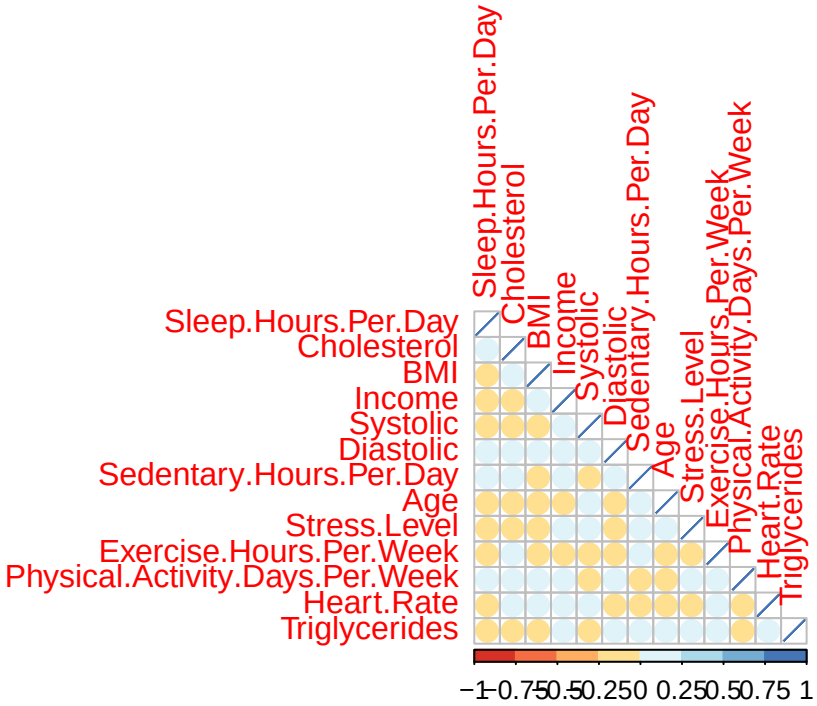


It seems like none of these fields were correlated with our target variable.

Implication wise:

- `Stress.Level` - not correlated - this says that stress does not impact heart attack risk; this is a bit surprising as higher stress can theoretically put someone at risk
- `Income` - not correlated - this says that income does not impact heart attack risk; this isn't too surprising as income doesn't directly impact someone's proneness to being put at risk
- `BMI` - not correlated - this says that BMI does not impact heart attack risk; this is a bit surprising as higher BMIers could indicate health risks
- `Triglycerides` - not correlated - this says that triglycerides does not impact heart attack risk; this is a bit surprising as higher fat levels theoretically would point to more health issues
- `Physical.Activity.Days.Per.Week` - not correlated - this says that physical activity does not impact heart attack risk; this is a bit surprising as the more active you are, the healthier you'd be theoretically
- `Sleep.Hours.Per.Day` - not correlated - this says that sleep does not impact heart attack risk; this isn't as surprising as sleep isn't as contributing to heart issues

There are some correlated fields here, but let's see if they're also correlated with each other beyond just the seven displayed here.



As noted above, it doesn't look like too many of these fields are correlated with one another.

Data Preparation

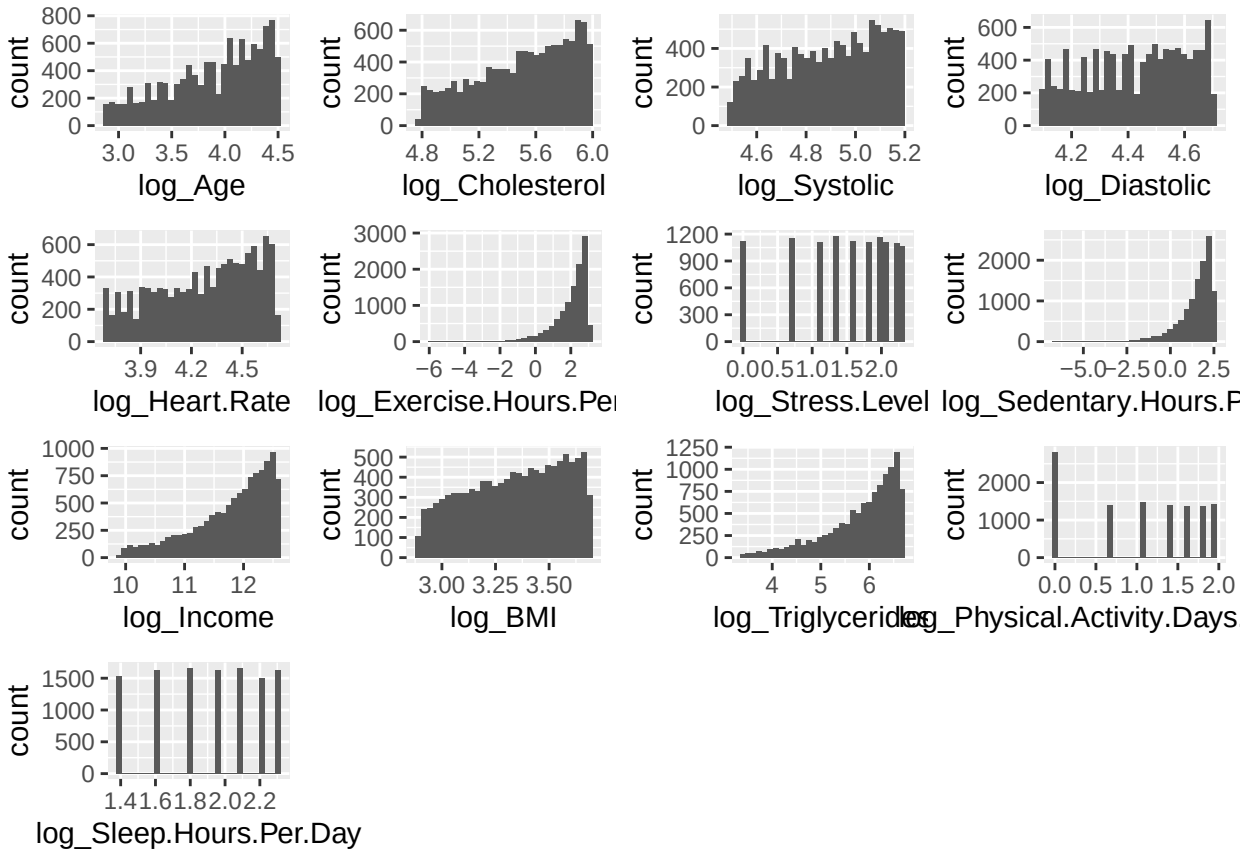
Outliers & Nulls

As noted above, it doesn't seem that there are many outliers in this dataset and there are no null fields. After confirming that there are no rows in any of these fields with outliers, it's safe to move onto the next step.

Transform Non-Normal Variables

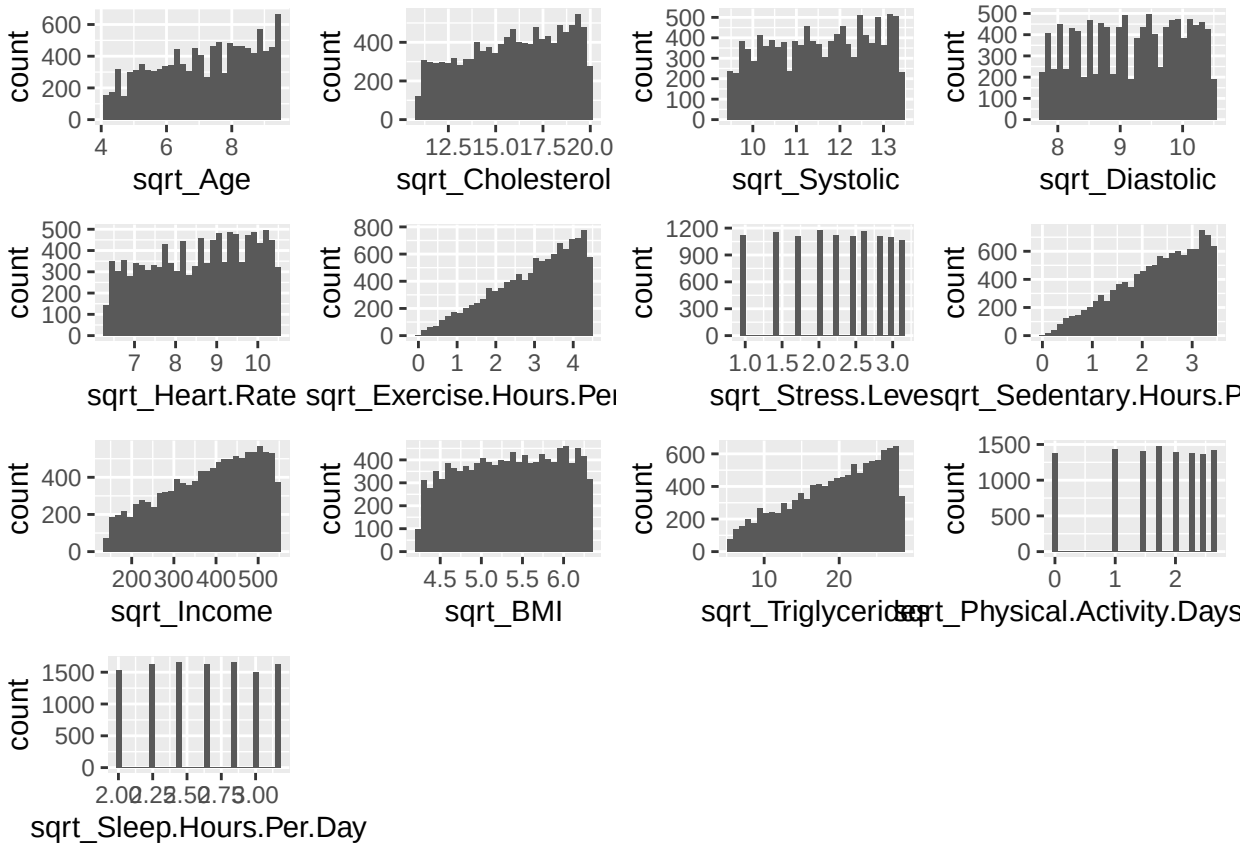
The last alteration before modeling is ensuring that all numeric variables are normal by transforming the ones that don't seem to have much of normal distribution. It honestly looks like all fields are not normal unfortunately, so all of them will be adjusted.

First, `log` will be applied and if that doesn't work, then `sqrt`, and finally if all else fails, `scaling`.

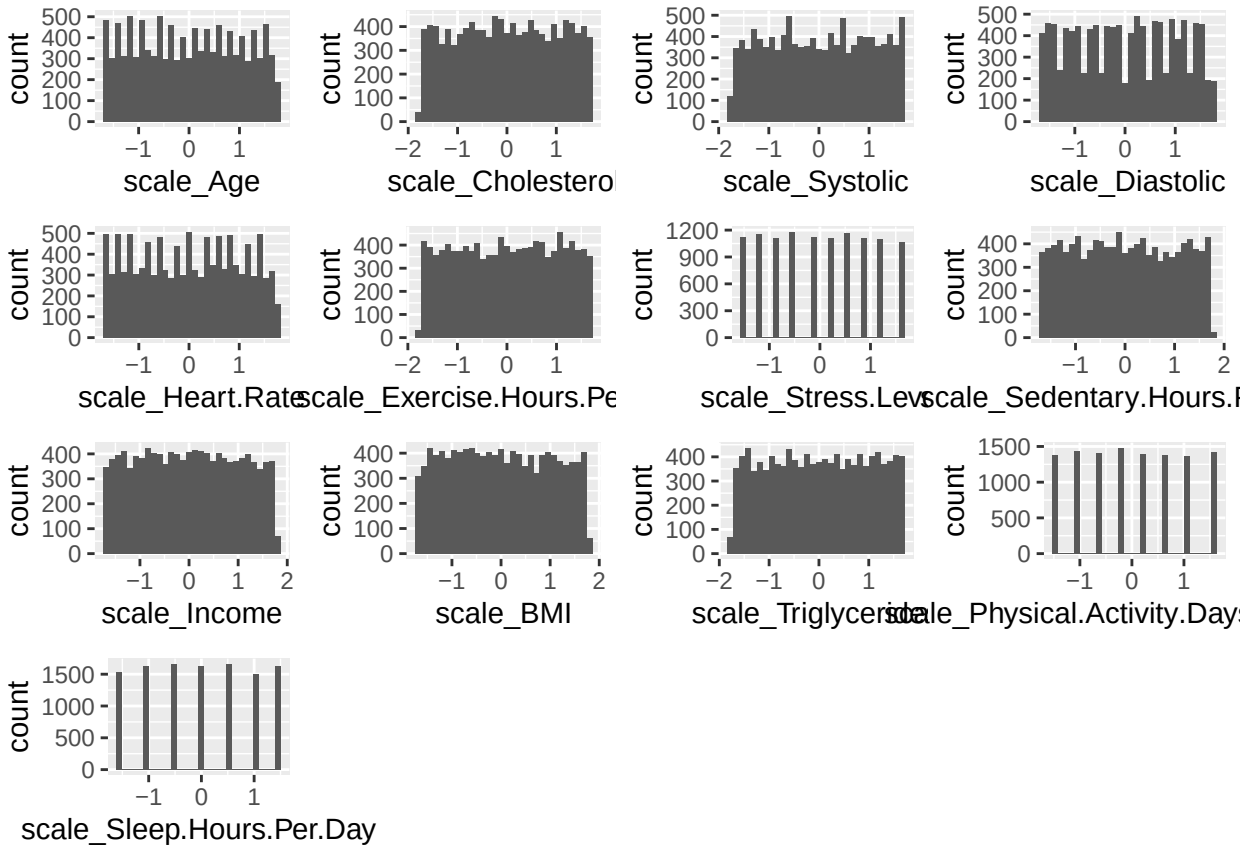


Looking at this, the data at least shows more of a trend, however none of these curves are too normal. A majority of them look skewed to the left now, but it is very extreme.

What about using `sqrt`?



This once again didn't seem to help – what about `scaling`?



Unfortunately, this also doesn't look great, but it does look a bit better than the non-transformed versions of these fields. When using transformed variables next, these scaled versions will be used as they look the cleanest out of the three attempts at transformations.

Build Models

Before doing anything, we will split the data into training and test sets with a 70/30 split.

We'll go through two sets of models:

- Model 1: Binomial Logistic Regression
- Modle 2: Decision Trees

Let's begin with binary models.

Binary Models

Model 1A

```
##
## Call:
## glm(formula = Class ~ scale_Age + scale_Cholesterol + scale_Systolic +
##     scale_Diastolic + scale_Heart.Rate + scale_Exercise.Hours.Per.Week +
##     scale_Stress.Level + scale_Sedentary.Hours.Per.Day + scale_Income +
##     scale_BMI + scale_Triglycerides + scale_Physical.Activity.Days.Per.Week +
```

```
##     scale_Sleep.Hours.Per.Day + Sex + Diabetes + Family.History +
##     Smoking + Obesity + Alcohol.Consumption + Diet + Previous.Heart.Problems +
##     Medication.Use + Country + Continent + Hemisphere, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4633  -1.1702   0.0011   1.1695   1.4582
##
## Coefficients: (6 not defined because of singularities)
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           0.053534   0.139266   0.384   0.7007
## scale_Age                             0.009552   0.025472   0.375   0.7077
## scale_Cholesterol                     0.045511   0.022767   1.999   0.0456 *
## scale_Systolic                        0.014525   0.022679   0.640   0.5219
## scale_Diastolic                      -0.045992   0.022855  -2.012   0.0442 *
## scale_Heart.Rate                     -0.036065   0.022717  -1.588   0.1124
## scale_Exercise.Hours.Per.Week         0.024426   0.022741   1.074   0.2828
## scale_Stress.Level                   -0.015471   0.022807  -0.678   0.4975
## scale_Sedentary.Hours.Per.Day        -0.019075   0.022676  -0.841   0.4002
## scale_Income                          0.025201   0.022768   1.107   0.2683
## scale_BMI                             0.021018   0.022661   0.927   0.3537
## scale_Triglycerides                   0.033461   0.022761   1.470   0.1415
## scale_Physical.Activity.Days.Per.Week -0.008817   0.022876  -0.385   0.6999
## scale_Sleep.Hours.Per.Day             -0.057157   0.022748  -2.513   0.0120 *
## SexMale                               0.066304   0.059397   1.116   0.2643
## Diabetes                              0.086989   0.047924   1.815   0.0695 .
## Family.History                        0.001359   0.045434   0.030   0.9761
## Smoking                              -0.015688   0.097074  -0.162   0.8716
## Obesity                              -0.077636   0.045375  -1.711   0.0871 .
## Alcohol.Consumption                  -0.113210   0.046242  -2.448   0.0144 *
## DietHealthy                           0.048182   0.055632   0.866   0.3865
## DietUnhealthy                         0.052812   0.055738   0.948   0.3434
## Previous.Heart.Problems               0.007883   0.045442   0.173   0.8623
## Medication.Use                        0.037574   0.045438   0.827   0.4083
## CountryAustralia                     -0.001026   0.139269  -0.007   0.9941
## CountryBrazil                        -0.158623   0.138870  -1.142   0.2534
## CountryCanada                        -0.065960   0.137993  -0.478   0.6327
## CountryChina                         -0.189355   0.139951  -1.353   0.1761
## CountryColombia                       0.108872   0.140201   0.777   0.4374
## CountryFrance                        -0.284705   0.140916  -2.020   0.0433 *
## CountryGermany                       -0.124688   0.136523  -0.913   0.3611
## CountryIndia                         -0.351068   0.143135  -2.453   0.0142 *
## CountryItaly                         -0.243914   0.140499  -1.736   0.0826 .
## CountryJapan                         -0.215431   0.143172  -1.505   0.1324
## CountryNew Zealand                   -0.048437   0.141892  -0.341   0.7328
## CountryNigeria                        0.100661   0.138966   0.724   0.4688
## CountrySouth Africa                  -0.144091   0.142027  -1.015   0.3103
## CountrySouth Korea                    0.075428   0.141390   0.533   0.5937
## CountrySpain                         -0.216156   0.140416  -1.539   0.1237
## CountryThailand                       0.012666   0.141987   0.089   0.9289
## CountryUnited Kingdom                -0.217065   0.139187  -1.560   0.1189
## CountryUnited States                 -0.063086   0.140351  -0.449   0.6531
## CountryVietnam                        0.109014   0.141692   0.769   0.4417
```

```
## ContinentAsia                                NA         NA      NA       NA
## ContinentAustralia                           NA         NA      NA       NA
## ContinentEurope                              NA         NA      NA       NA
## ContinentNorth America                       NA         NA      NA       NA
## ContinentSouth America                       NA         NA      NA       NA
## HemisphereSouthern Hemisphere                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10916  on 7873  degrees of freedom
## Residual deviance: 10843  on 7831  degrees of freedom
## AIC: 10929
##
## Number of Fisher Scoring iterations: 4
```

Interpretation

As there are so many variables, only those that are significant will be analyzed. Those variables are:

- scale_Cholesterol
  - Impact: Positive
  - Meaning: The higher the cholesterol the more likely they are to be at risk; this makes sense as cholesterol is usually an indication of overall health and dietary trends
- scale_Diastolic
  - Impact: Negative
  - Meaning: The higher the diastolic pressure, the less likely they are to be at risk; this does not make sense as higher blood pressure usually indicates health issues and higher risk for heart disease
- scale_Sleep.Hours.Per.day
  - Impact: Negative
  - Meaning: The higher a person sleeps, the less likely they are to be at risk; this inherently makes sense as the more sleep a person gets, the healthier they'll be as they won't be as fatigued
- Diabetes
  - Impact: Positive
  - Meaning: The presence of diabetes indicates that a person is more likely for heart attack risk; this makes sense as having diabetes is usually a risk factor for heart disease
- Obesity
  - Impact: Negative
  - Meaning: The state of being obese indicates that a person is less likely for heart attack risk; this doesn't make as much sense as obese people are statistically more likely to have health issues related to diet and that impacts the heart
- Alcohol.Consumption
  - Impact: Negative
  - Meaning: If a person indulges in alcohol, they are less likely for heart attack risk; this inherently doesn't make too much sense as those who consume *too* much alcohol can be seen as a problem, but perhaps because of the lack of granularity in this field (ideally this would be the categorical variable as it outlined on Kaggle rather than a boolean), this is missing that distinction between light and heavy drinkers

- CountryFrance
  - Impact: Negative
  - Meaning: If a person lives in France, they are less likely for heart attack risk; this is interesting as this could mean that certain aspects of France lead to a lifestyle taht is less prone to heart attacks
- CountryIndia
  - Impact: Negative
  - Meaning Similar to the field above, this means if a person lives in india, they are less likely for heart attack risk
- CountryItaly
  - Impact: Negative
  - Meaning: Similar to the two fields above, this means that if a person lives in Italy, they are less likely for heart attack risk

It should also be noted that the residual deviance is 10843 and the AIC is 10929 for comparison with future models.

The next iteration of the model will be using only significant variables.

Model 1B

```
##
## Call:
## glm(formula = Class ~ scale_Cholesterol + scale_Diastolic + scale_Sleep.Hours.Per.Day +
##     Diabetes + Obesity + Alcohol.Consumption + Country, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.40905  -1.17034   0.02631   1.17056   1.42089
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  0.13896    0.10799   1.287  0.19819
## scale_Cholesterol            0.04550    0.02271   2.003  0.04516 *
## scale_Diastolic             -0.04480    0.02280  -1.964  0.04949 *
## scale_Sleep.Hours.Per.Day -0.05929    0.02270  -2.612  0.00901 **
## Diabetes                     0.08773    0.04785   1.833  0.06673 .
## Obesity                     -0.07727    0.04532  -1.705  0.08816 .
## Alcohol.Consumption         -0.11235    0.04613  -2.435  0.01489 *
## CountryAustralia            -0.00453    0.13900  -0.033  0.97400
## CountryBrazil               -0.15068    0.13848  -1.088  0.27657
## CountryCanada               -0.06569    0.13769  -0.477  0.63333
## CountryChina                -0.19045    0.13964  -1.364  0.17263
## CountryColombia              0.10699    0.13991   0.765  0.44442
## CountryFrance               -0.28204    0.14047  -2.008  0.04465 *
## CountryGermany              -0.12357    0.13622  -0.907  0.36432
## CountryIndia                -0.34929    0.14277  -2.446  0.01443 *
## CountryItaly                -0.23456    0.14012  -1.674  0.09412 .
## CountryJapan                -0.20932    0.14287  -1.465  0.14290
```

```
## CountryNew Zealand        -0.04882    0.14152  -0.345  0.73014
## CountryNigeria             0.10587    0.13869   0.763  0.44523
## CountrySouth Africa       -0.14224    0.14171  -1.004  0.31551
## CountrySouth Korea         0.07448    0.14113   0.528  0.59765
## CountrySpain              -0.21125    0.13999  -1.509  0.13128
## CountryThailand            0.01535    0.14171   0.108  0.91373
## CountryUnited Kingdom     -0.21416    0.13889  -1.542  0.12308
## CountryUnited States      -0.05179    0.14006  -0.370  0.71155
## CountryVietnam             0.11132    0.14142   0.787  0.43120
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10916  on 7873  degrees of freedom
## Residual deviance: 10856  on 7848  degrees of freedom
## AIC: 10908
##
## Number of Fisher Scoring iterations: 3
```

Interpretation

Each variable will be assessed one-by-one similar to the previous iteration:

- Intercept
  - Impact: Positive
  - Meaning: Without any knowledge of other fields, a person is prone to heart attack risk
- scale_Cholesterol
  - Impact: Positive
  - Comparison with previous model: This is still positive and the magnitude of the variable is similar so not much change here
- scale_Diastolic
  - Impact: Negative
  - Comparison with previous model: This is still negative and the magnitude of the variable is similar so not much change here
- scale_Sleep.Hours.Per.day
  - Impact: Negative
  - Comparison with previous model: This is still negative and the magnitude of the variable is similar so not much change here; this did however improve in significance (it is now more significant)
- Diabetes
  - Impact: Positive
  - Comparison with previous model: This is still positive and the magnitude of the variable is similar so not much change here
- Obesity
  - Impact: Negative
  - Comparison with previous model: This is still negative and the magnitude of the variable is similar so not much change here
- Alcohol.Consumption

  - Impact: Negative
  - Comparison with previous model: This is still negative and the magnitude of the variable is similar so not much change here

- CountryFrance
  - Impact: Negative
  - Comparison with previous model: This is still negative and the magnitude of the variable is similar so not much change here

- CountryIndia
  - Impact: Negative
  - Comparison with previous model: This is still negative and the magnitude of the variable is similar so not much change here

- CountryItaly
  - Impact: Negative
  - Comparison with previous model: This is still negative and the magnitude of the variable is similar so not much change here

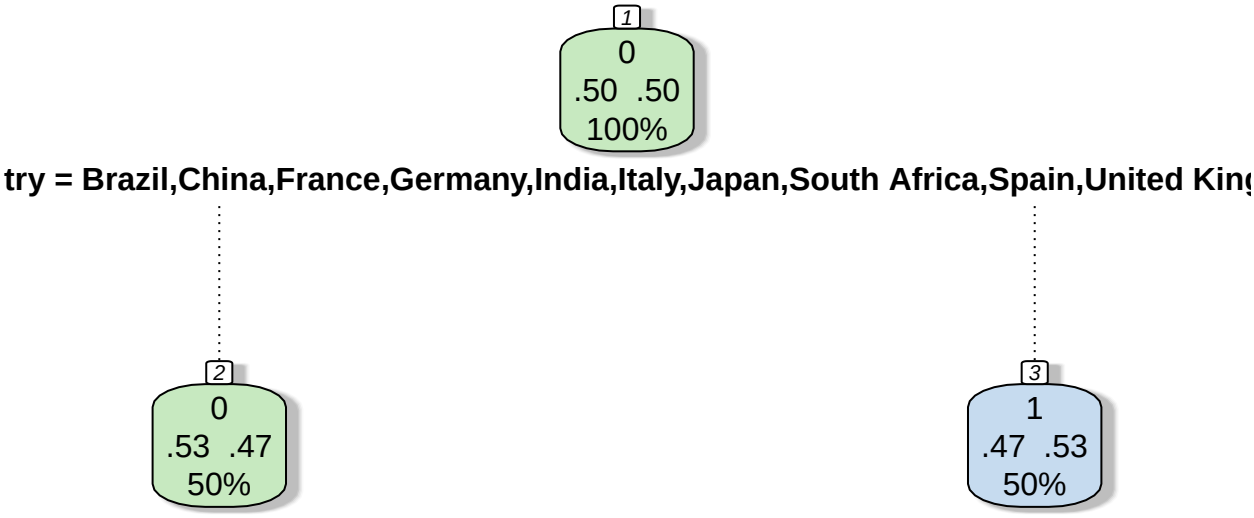
It looks like the AIC went down, but residual deviance went up in comparison with the previous model.

Model 2A

This model will use Decision Trees.

```
## Call:
## rpart(formula = Class ~ scale_Age + scale_Cholesterol + scale_Systolic +
##     scale_Diastolic + scale_Heart.Rate + scale_Exercise.Hours.Per.Week +
##     scale_Stress.Level + scale_Sedentary.Hours.Per.Day + scale_Income +
##     scale_BMI + scale_Triglycerides + scale_Physical.Activity.Days.Per.Week +
##     scale_Sleep.Hours.Per.Day + Sex + Diabetes + Family.History +
##     Smoking + Obesity + Alcohol.Consumption + Diet + Previous.Heart.Problems +
##     Medication.Use + Country + Continent + Hemisphere, data = train,
##     method = "class")
##   n= 7874
##
##           CP nsplit rel error    xerror       xstd
## 1 0.05588011      0 1.0000000 1.0251461 0.01126588
## 2 0.01000000      1 0.9441199 0.9664719 0.01126310
##
## Variable importance
##                        Country                      Continent
##                             59                             30
##                     Hemisphere            scale_Stress.Level
##                              5                              2
## scale_Sedentary.Hours.Per.Day                  scale_Income
##                              2                              2
##
## Node number 1: 7874 observations,    complexity param=0.05588011
##   predicted class=0  expected loss=0.5  P(node) =1
##     class counts:  3937  3937
##    probabilities: 0.500 0.500
##   left son=2 (3920 obs) right son=3 (3954 obs)
##   Primary splits:
```

```
##       Country                   splits as  RRLRLRLLLLLRRLRLRLRR, improve=12.293850, (0 missing)
##       scale_Income              < -1.714628  to the left,  improve= 6.009158, (0 missing)
##       scale_BMI                 < -1.497913  to the left,  improve= 5.946167, (0 missing)
##       scale_Sleep.Hours.Per.Day < 0.7540851  to the right, improve= 5.407965, (0 missing)
##       scale_Diastolic           < 0.5800374  to the right, improve= 4.762151, (0 missing)
##   Surrogate splits:
##       Continent                     splits as  RRRLRR, agree=0.758, adj=0.513, (0 split)
##       Hemisphere                    splits as  RL, agree=0.547, adj=0.091, (0 split)
##       scale_Stress.Level            < -0.3358009 to the left,  agree=0.519, adj=0.034, (0 split)
##       scale_Sedentary.Hours.Per.Day < -0.3167389 to the left,  agree=0.519, adj=0.033, (0 split)
##       scale_Income                  < -0.1176614 to the right, agree=0.518, adj=0.032, (0 split)
##
## Node number 2: 3920 observations
##   predicted class=0  expected loss=0.4719388  P(node) =0.497841
##     class counts:  2070  1850
##    probabilities: 0.528 0.472
##
## Node number 3: 3954 observations
##   predicted class=1  expected loss=0.4721801  P(node) =0.502159
##     class counts:  1867  2087
##    probabilities: 0.472 0.528
```



Interpretation

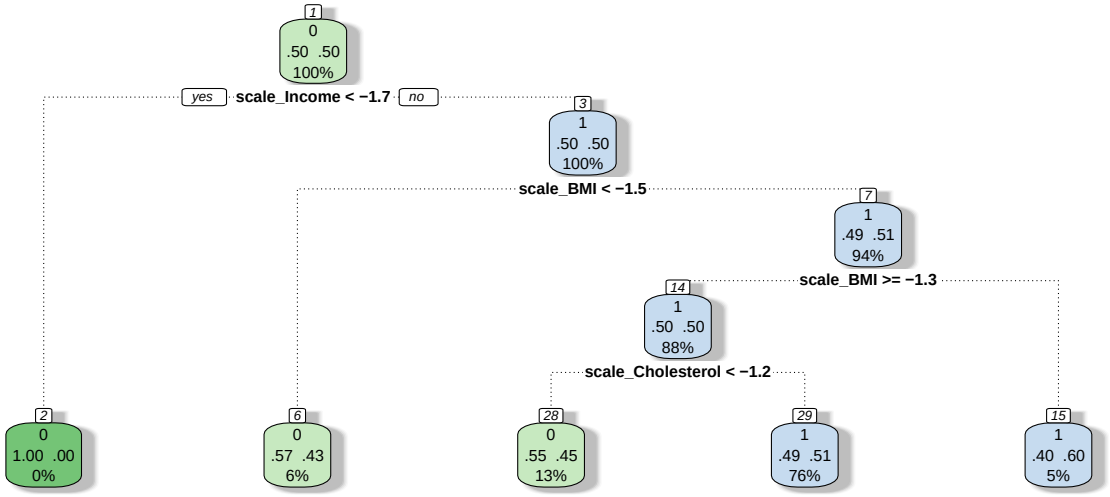
This tree is not very helpful – it's saying that regardless of whether you are in one of the listed countries, there's a 50% chance that you will be at risk of a heart attack.

What if we remove the `Country` variable here?

Model 2B

```
## Call:
## rpart(formula = Class ~ scale_Age + scale_Cholesterol + scale_Systolic +
##     scale_Diastolic + scale_Heart.Rate + scale_Exercise.Hours.Per.Week +
##     scale_Stress.Level + scale_Sedentary.Hours.Per.Day + scale_Income +
##     scale_BMI + scale_Triglycerides + scale_Physical.Activity.Days.Per.Week +
##     scale_Sleep.Hours.Per.Day + Sex + Diabetes + Family.History +
##     Smoking + Obesity + Alcohol.Consumption + Diet + Previous.Heart.Problems +
##     Medication.Use + Continent + Hemisphere, data = train, method = "class")
##   n= 7874
##
##           CP nsplit rel error   xerror       xstd
## 1 0.01054102      0 1.0000000 1.024130 0.01126616
## 2 0.01000000      4 0.9535179 1.010414 0.01126883
##
## Variable importance
##         scale_BMI      scale_Income scale_Cholesterol
##                53                24                23
##
## Node number 1: 7874 observations,    complexity param=0.01054102
##   predicted class=0  expected loss=0.5  P(node) =1
##     class counts:  3937  3937
##    probabilities: 0.500 0.500
##   left son=2 (12 obs) right son=3 (7862 obs)
##   Primary splits:
##       scale_Income              < -1.714628 to the left,  improve=6.009158, (0 missing)
##       scale_BMI                 < -1.497913 to the left,  improve=5.946167, (0 missing)
##       scale_Sleep.Hours.Per.Day < 0.7540851 to the right, improve=5.407965, (0 missing)
##       scale_Diastolic           < 0.5800374 to the right, improve=4.762151, (0 missing)
##       scale_Triglycerides       < 1.687136  to the right, improve=4.684785, (0 missing)
##
## Node number 2: 12 observations
##   predicted class=0  expected loss=0  P(node) =0.001524003
##     class counts:    12     0
##    probabilities: 1.000 0.000
##
## Node number 3: 7862 observations,    complexity param=0.01054102
##   predicted class=1  expected loss=0.4992368  P(node) =0.998476
##     class counts:  3925  3937
##    probabilities: 0.499 0.501
##   left son=6 (475 obs) right son=7 (7387 obs)
##   Primary splits:
##       scale_BMI                 < -1.497913 to the left,  improve=5.763450, (0 missing)
##       scale_Sleep.Hours.Per.Day < 0.7540851 to the right, improve=5.364089, (0 missing)
##       scale_Diastolic           < 0.5800374 to the right, improve=4.841973, (0 missing)
##       scale_Triglycerides       < 1.687136  to the right, improve=4.384081, (0 missing)
##       scale_Heart.Rate          < 0.8550152 to the right, improve=4.360203, (0 missing)
##
## Node number 6: 475 observations
```

```
##   predicted class=0  expected loss=0.4252632  P(node) =0.06032512
##     class counts:   273   202
##    probabilities: 0.575 0.425
##
## Node number 7: 7387 observations,    complexity param=0.01054102
##   predicted class=1  expected loss=0.494382  P(node) =0.9381509
##     class counts:  3652  3735
##    probabilities: 0.494 0.506
##   left son=14 (6963 obs) right son=15 (424 obs)
##   Primary splits:
##       scale_BMI                < -1.316022 to the right, improve=7.463025, (0 missing)
##       scale_Sleep.Hours.Per.Day < 0.7540851 to the right, improve=5.888991, (0 missing)
##       scale_Heart.Rate         < 0.806317  to the right, improve=4.587447, (0 missing)
##       scale_Cholesterol        < -1.220395 to the left,  improve=4.301469, (0 missing)
##       scale_Triglycerides      < 1.687136  to the right, improve=3.926294, (0 missing)
##
## Node number 14: 6963 observations,    complexity param=0.01054102
##   predicted class=1  expected loss=0.4999282  P(node) =0.8843028
##     class counts:  3481  3482
##    probabilities: 0.500 0.500
##   left son=28 (1004 obs) right son=29 (5959 obs)
##   Primary splits:
##       scale_Cholesterol        < -1.220395 to the left,  improve=5.835940, (0 missing)
##       scale_Sleep.Hours.Per.Day < 0.7540851 to the right, improve=4.743410, (0 missing)
##       scale_Triglycerides      < 1.687136  to the right, improve=4.198482, (0 missing)
##       scale_Heart.Rate         < 0.8550152 to the right, improve=4.135229, (0 missing)
##       scale_Diastolic          < 0.5800374 to the right, improve=3.579167, (0 missing)
##   Surrogate splits:
##       scale_Sedentary.Hours.Per.Day < -1.712882 to the left,  agree=0.856, adj=0.001, (0 split)
##
## Node number 15: 424 observations
##   predicted class=1  expected loss=0.4033019  P(node) =0.05384811
##     class counts:   171   253
##    probabilities: 0.403 0.597
##
## Node number 28: 1004 observations
##   predicted class=0  expected loss=0.4501992  P(node) =0.1275083
##     class counts:   552   452
##    probabilities: 0.550 0.450
##
## Node number 29: 5959 observations
##   predicted class=1  expected loss=0.4915254  P(node) =0.7567945
##     class counts:  2929  3030
##    probabilities: 0.492 0.508
```



This tree is a little more indicative – it brings in income first, then checks BMI and cholesterol.

Now that we have two models, let's see how each performs.

Select Models

Binary Models

Confusion Matrices

First, we'll take a look at confusion matrices for each of the models.

```
# if the prediction is >= 0.5, then we would predict 1 for that row, otherwise 0
test$model1a <- ifelse(predict.glm(model1a, test, "response") >= 0.5, 1, 0)

# create the confusion matrix
cm1a <- confusionMatrix(factor(test$model1a), factor(test$Class), "1")
results <- tibble(Model = "Model #1A", Accuracy=cm1a$byClass[11], F1 = cm1a$byClass[7],
                  Deviance= model1a$deviance,
                  R2 = 1 - model1a$deviance / model1a$null.deviance,
                  AIC = model1a$aic)
cm1a
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
```

```
##          0 853 858
##          1 834 829
##
##                Accuracy : 0.4985
##                  95% CI : (0.4815, 0.5155)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.5751
##
##                   Kappa : -0.003
##
##  Mcnemar's Test P-Value : 0.5761
##
##             Sensitivity : 0.4914
##             Specificity : 0.5056
##          Pos Pred Value : 0.4985
##          Neg Pred Value : 0.4985
##              Prevalence : 0.5000
##          Detection Rate : 0.2457
##    Detection Prevalence : 0.4929
##       Balanced Accuracy : 0.4985
##
##        'Positive' Class : 1
##
```

```r
# if the prediction is >= 0.5, then we would predict 1 for that row, otherwise 0
test$model1b <- ifelse(predict.glm(model1b, test, "response") >= 0.5, 1, 0)

# create the confusion matrix
cm1b <- confusionMatrix(factor(test$model1b), factor(test$Class), "1")
results <- tibble(Model = "Model #1B", Accuracy=cm1b$byClass[11], F1 = cm1b$byClass[7],
                  Deviance= model1b$deviance,
                  R2 = 1 - model1b$deviance / model1b$null.deviance,
                  AIC = model1b$aic)
cm1b
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 848 843
##          1 839 844
##
##                Accuracy : 0.5015
##                  95% CI : (0.4845, 0.5185)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.4384
##
##                   Kappa : 0.003
##
##  Mcnemar's Test P-Value : 0.9417
##
##             Sensitivity : 0.5003
##             Specificity : 0.5027
##          Pos Pred Value : 0.5015
```

```
##          Neg Pred Value : 0.5015
##              Prevalence : 0.5000
##          Detection Rate : 0.2501
##    Detection Prevalence : 0.4988
##       Balanced Accuracy : 0.5015
##
##        'Positive' Class : 1
##
```

```
test$model2a <- predict(model2a, test, type="class")

# create the confusion matrix
cm2a <- confusionMatrix(factor(test$model2a), factor(test$Class), "1")
results <- tibble(Model = "Model #2A", Accuracy=cm2a$byClass[11], F1 = cm2a$byClass[7],
                  Deviance= model2a$deviance,
                  R2 = 1 - model2a$deviance / model2a$null.deviance,
                  AIC = model2a$aic)
cm2a
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 829 840
##          1 858 847
##
##                Accuracy : 0.4967
##                  95% CI : (0.4797, 0.5138)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.6539
##
##                   Kappa : -0.0065
##
##  Mcnemar's Test P-Value : 0.6799
##
##             Sensitivity : 0.5021
##             Specificity : 0.4914
##          Pos Pred Value : 0.4968
##          Neg Pred Value : 0.4967
##              Prevalence : 0.5000
##          Detection Rate : 0.2510
##    Detection Prevalence : 0.5053
##       Balanced Accuracy : 0.4967
##
##        'Positive' Class : 1
##
```

```
test$model2b <- predict(model2b, test, type="class")

# create the confusion matrix
cm2b <- confusionMatrix(factor(test$model2b), factor(test$Class), "1")
results <- tibble(Model = "Model #2B", Accuracy=cm1b$byClass[11], F1 = cm2b$byClass[7],
                  Deviance= model2b$deviance,
```

```
                  R2 = 1 - model2b$deviance / model2b$null.deviance,
                  AIC = model2b$aic)
cm2b
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  334  297
##          1 1353 1390
##
##                Accuracy : 0.511
##                  95% CI : (0.4939, 0.528)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.1044
##
##                   Kappa : 0.0219
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8239
##             Specificity : 0.1980
##          Pos Pred Value : 0.5067
##          Neg Pred Value : 0.5293
##              Prevalence : 0.5000
##          Detection Rate : 0.4120
##    Detection Prevalence : 0.8130
##       Balanced Accuracy : 0.5110
##
##        'Positive' Class : 1
##
```

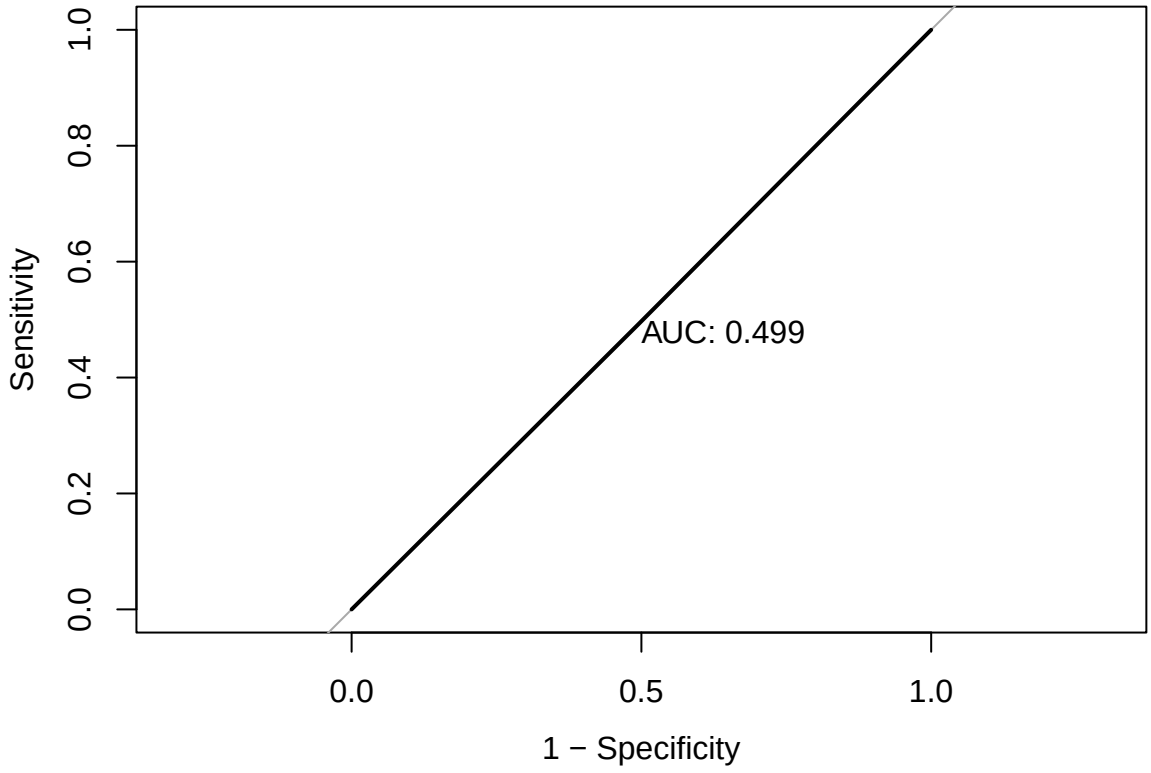
ROC

Now with all of these matrices, we'll look at ROC curves.

```
print('Model 1A ROC Curve')
```

```
## [1] "Model 1A ROC Curve"
```

```
roc(test[["Class"]], test[["model1a"]], plot = TRUE, legacy.axes = TRUE, print.auc = TRUE)
```

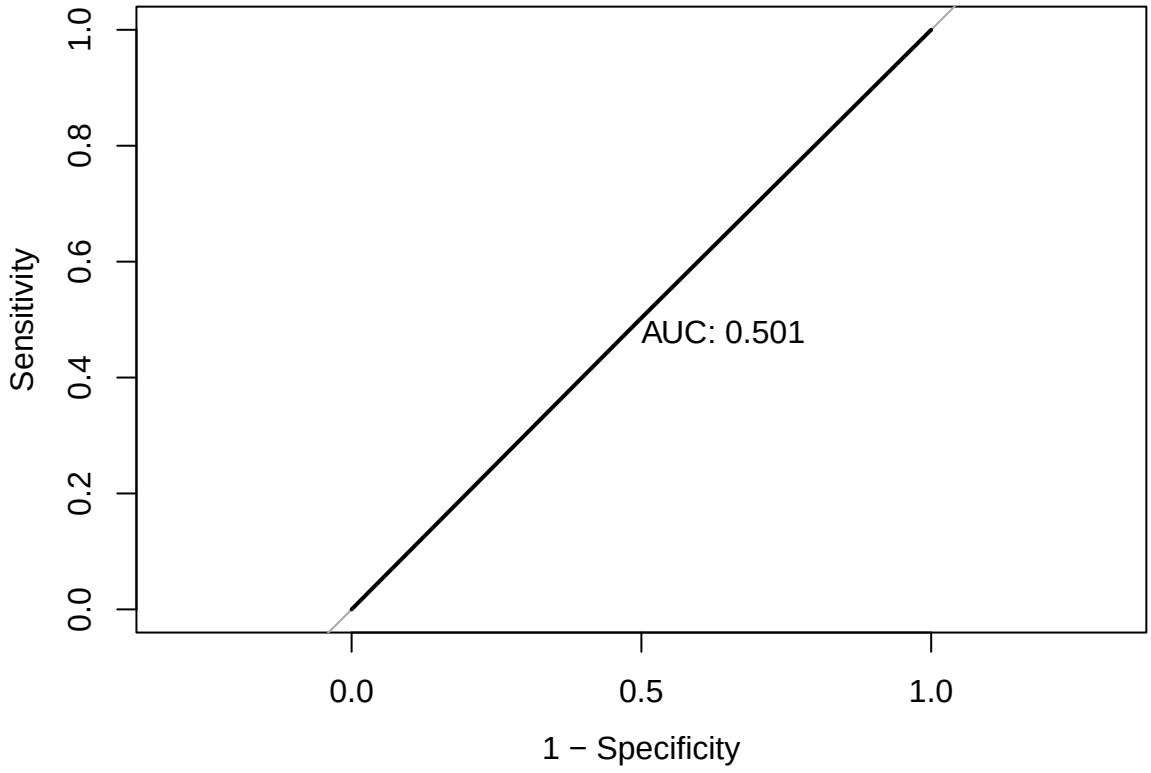



```
##
## Call:
## roc.default(response = test[["Class"]], predictor = test[["model1a"]],     plot = TRUE, legacy.axes =
##
## Data: test[["model1a"]] in 1687 controls (test[["Class"]] 0) < 1687 cases (test[["Class"]] 1).
## Area under the curve: 0.4985
```

```
print('Model 1B ROC Curve')
```

```
## [1] "Model 1B ROC Curve"
```

```
roc(test[["Class"]], test[["model1b"]], plot = TRUE, legacy.axes = TRUE, print.auc = TRUE)
```

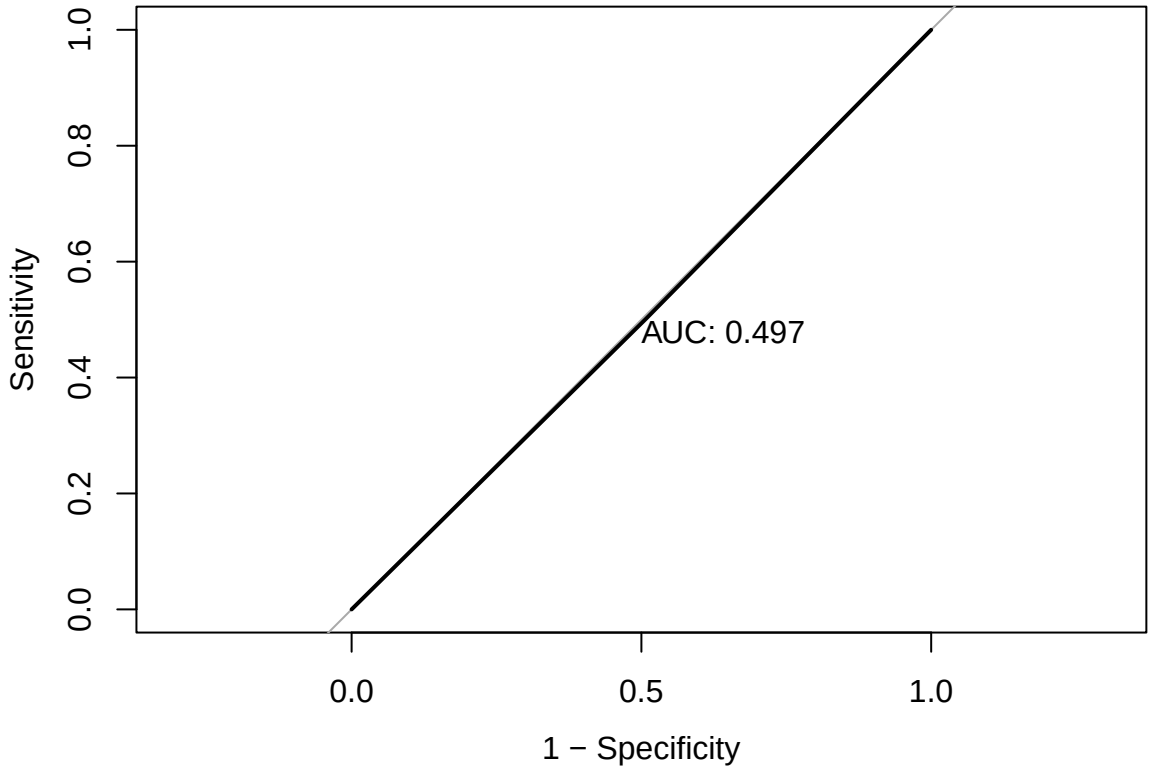



```
##
## Call:
## roc.default(response = test[["Class"]], predictor = test[["model1b"]],     plot = TRUE, legacy.axes =
##
## Data: test[["model1b"]] in 1687 controls (test[["Class"]] 0) < 1687 cases (test[["Class"]] 1).
## Area under the curve: 0.5015
```

```
print('Model 2A ROC Curve')
```

```
## [1] "Model 2A ROC Curve"
```

```
roc(test[["Class"]], as.numeric(test[["model2a"]]), plot = TRUE, legacy.axes = TRUE, print.auc = TRUE)
```

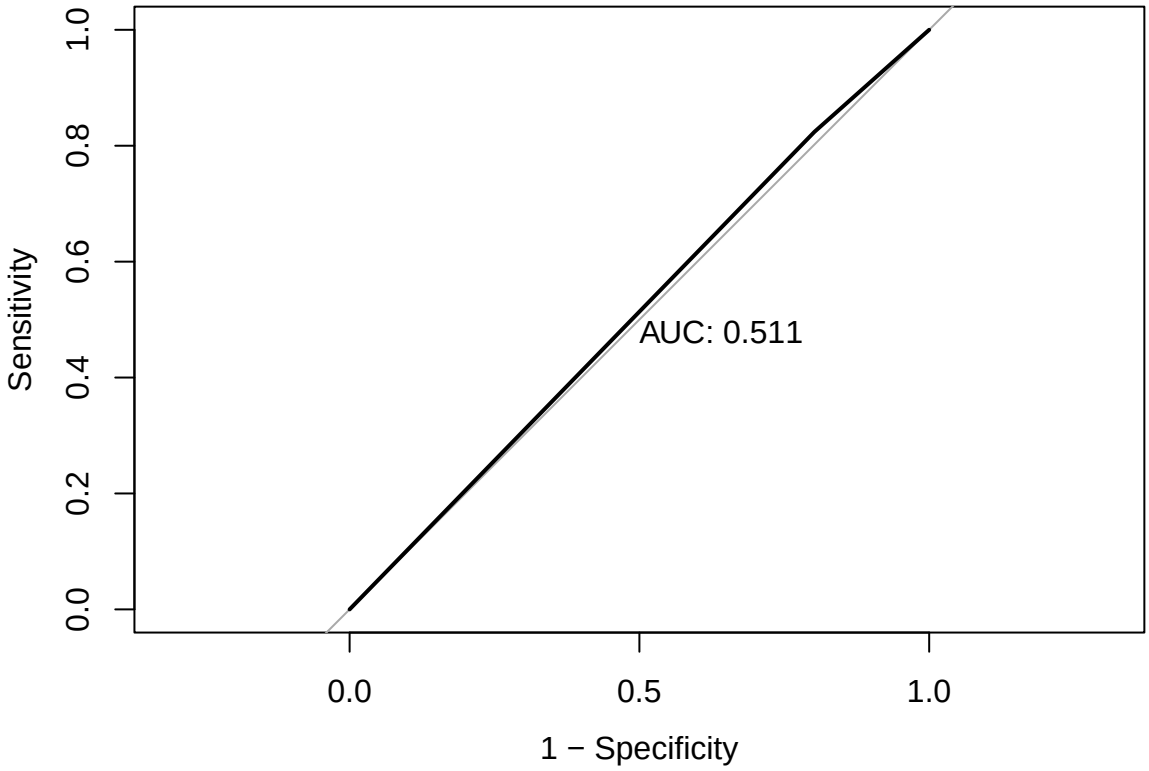



```
##
## Call:
## roc.default(response = test[["Class"]], predictor = as.numeric(test[["model2a"]]),     plot = TRUE, l
##
## Data: as.numeric(test[["model2a"]]) in 1687 controls (test[["Class"]] 0) < 1687 cases (test[["Class"]
## Area under the curve: 0.4967
```

```
print('Model 2B ROC Curve')
```

```
## [1] "Model 2B ROC Curve"
```

```
roc(test[["Class"]], as.numeric(test[["model2b"]]), plot = TRUE, legacy.axes = TRUE, print.auc = TRUE)
```




```
##
## Call:
## roc.default(response = test[["Class"]], predictor = as.numeric(test[["model2b"]]),     plot = TRUE, l
##
## Data: as.numeric(test[["model2b"]]) in 1687 controls (test[["Class"]] 0) < 1687 cases (test[["Class"]
## Area under the curve: 0.511
```

Overall Comparisons

```
##           Accuracy Precision Specificity    Recall        F1
## Model 1A 0.4985181 0.4984967   0.5056313 0.4914049 0.4949254
## Model 1B 0.5014819 0.5014854   0.5026675 0.5002964 0.5008902
## Model 2A 0.4967398 0.4967742   0.4914049 0.5020747 0.4994104
## Model 2B 0.5109662 0.5067444   0.1979846 0.8239478 0.6275395
```

Based on the above output and the ROC graphs:

- ROC/AUC: The B models were best here with Model 2B being slightly higher performing than Model 1B.
- Accuracy: The B models were best here again, Model 2B being a bit higher while Model 1A was the best out of the As
- Precision: Once again Model 2B is the best here and Model 1A is the best out of the As
- Specificity: This is where Model 2B falls short with a very low value; Model 1A is the next best

- Recall: Model 2B is much higher now here which somewhat corresponds with the lower specificity; it seems like Model 2B is over capturing positives. Regardless, the next highest here is Model 2A with Model 1A as the lowest
- F1: Model 2B is the best here with Model 1B being the second best; Model 2A is the best out of the As

Conclusion

Overall, despite Model 2B's strong performance, the fact that it's over capturing positives isn't a great outcome. The implications of this mean that people would be deemed as at risk for a heart attack when in reality, they're not. This could then lead to preventative measures being implemented and could potentially be very costly and time-consuming.

The next best performing model is 1B. At 50% for every other metric and the second best AUC, this is the next alternative overall, although it really isn't that much better than the rest of the models. It still slightly outperforms them though so this is the final selection!