*This is the main submission document**. Save and rename this document filename with your registered full name as Prefix before submission.***

Class	6
Full Name	Teresa Zhang Han Yu
Matriculation Number	U2022886C

** : Delete and replace as appropriate.*

Declaration of Academic Integrity

By submitting this assignment for assessment, I declare that this submission is my own work, unless otherwise quoted, cited, referenced or credited. I have read and understood the Instructions to CBA.PDF provided and the Academic Integrity Policy.

I am aware that failure to act in accordance with the University's Academic Integrity Policy may lead to the imposition of penalties which may include the requirement to revise and resubmit an assignment, receiving a lower grade, or receiving an F grade for the assignment; suspension from the University or termination of my candidature.

I consent to the University copying and distributing any or all of my work in any form and using third parties to verify whether my work contains plagiarised material, and for quality assurance purposes.

Please insert an "X" within the square brackets below to indicate your selection.

[X] I have read and accept the above.

Table of Contents

For each question, please start your answer in a new page.

Answer to Q1:

Create the BMI variable based on CDC definition. Show your code.

I decided to create two columns of the BMI variable, one column with continuous values (BMI) and the other with categorical values (Weight_status).

Creating BMI column

Based on the CDC definition, the formula to calculate BMI is weight (kg) / [height (m)]^2. Since Height was recorded in centimetres in the dataset, I changed the units to metres by dividing the values by 100. Since I imported the premium2 file as a data table, it became much easier to transform data in R. Thus, my code is just one line:

```
premium.dt[, BMI := Weight/((Height/100)^2)]
```

Creating Weight_status column

Based on the CDC definition,

- **If your BMI is less than 18.5**, it falls within the underweight range.
- **If your BMI is 18.5 to 24.9**, it falls within the normal or Healthy Weight range.
- **If your BMI is 25.0 to 29.9**, it falls within the overweight range.
- **If your BMI is 30.0 or higher**, it falls within the obese range.

Thus, I convert ranges of values from the BMI column I created just before into ordinal variables, “Underweight” as 0, “Normal” as 1, “Overweight” as 2, and “Obese” as 3 and put them into a new column:

```
premium.dt[, Weight_status := cut(premium.dt$BMI,
                                  breaks=c(0, 18.5, 25.0, 30.0, Inf), labels=c(0, 1, 2, 3))]
```

Answer to Q2:

There are many categorical variables with integer coded values (e.g. Diabetes, HighBloodPressure, Transplant...etc.) Is it necessary to convert them to factor datatype in R?

Yes, it is necessary to convert them to factor datatype in R. **Continuous data** is any type of data where anything you are measuring can be characterized with a number. **Categorical data** describe a 'characteristic' of a data unit and are selected from a small group of categories. Depending on the type of your data, a specific analysis will be appropriate. Similarly, the data type will also drive the choice of data visualization techniques (Djangone, 2021). For example, for linear regression, only when R recognize a variable as categorical then will dummy variables be created.

Thus, to convert all the categorical variables with integer coded value to factor datatype in R, I need to find all of them first by checking through the class types of all the variables using a for-loop. I found that Diabetes, HighBloodPressure, Transplant, ChronicDisease, Allergy, CancerInFamily, and Gender have integer class type but should be categorical instead as they only have two possible values, 1 or 0.

After finding the categorical variables, I convert them to factor datatype by using the as.factor() function. After that, I change the numerals for each categorical variable to their appropriate strings in a new column respectively for data exploration later.

Answer to Q3:

Explore the data and report on your key findings.

I checked for any inconsistencies in the data using summary, after that, I checked for any null and duplicated values. I will not be removing any outliers unless they are influential outliers.

I plot graphs to help me explore the data. Different plots are plotted for variables with different datatype.

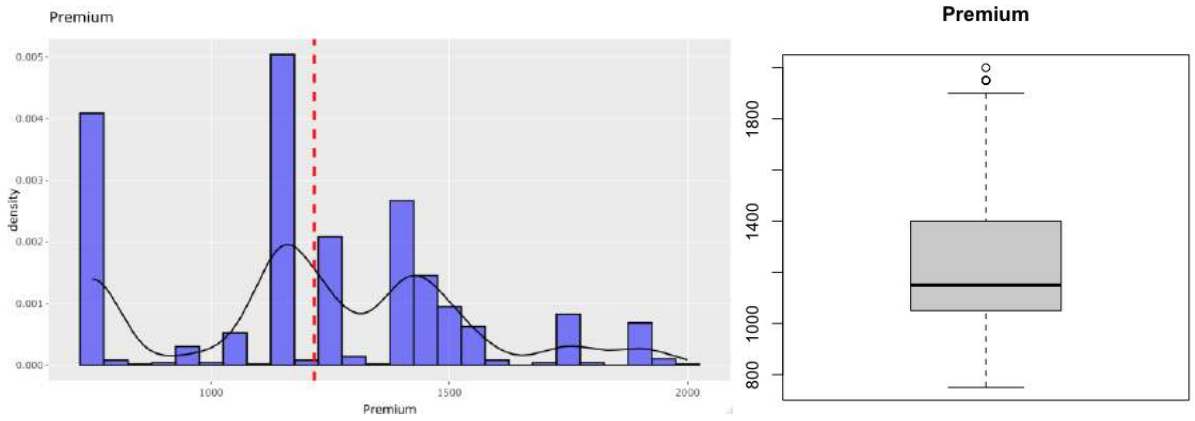
For univariate continuous variables, I plot a histogram to show the frequency distribution. All histograms display the density distribution line (in black) and the mean of the respective histogram (in red). I also plot a boxplot to show the various quartiles and the mean. To show the relationship between continuous x-variables and Premium, I plot a scatter plot with a smooth line with the method = "loess" (Data Novia, 2021). This computes a smooth local regression, and will be able to tell me if there is a strong linear relationship between them.

For univariate categorial variables, I plot a bar plot to show the numbers of each categories. To show the relationship between categorical x-variables and Premium, I plot a boxplot of Premiun facetted by each categories of the x-variables to show how Premium varies for each categories. I also plot a scatter plot with jittered x to show each point clearer, and show the concentration of certain Premium values for each categories.

To have a better understanding of the relationship between Premium and the x-variables, I researched on how Insurance Premium works. In general, the greater the risk associated, the more expensive the insurance policy and thus, the insurance premiums (Kagan, 2021).

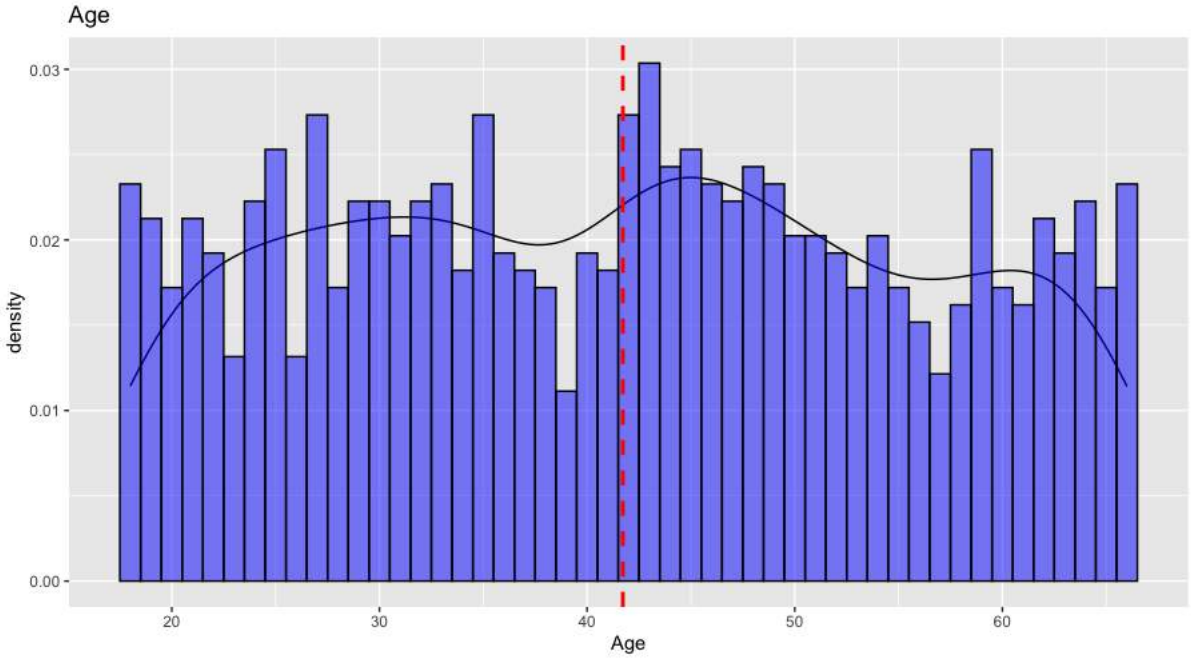
Continuous Variable

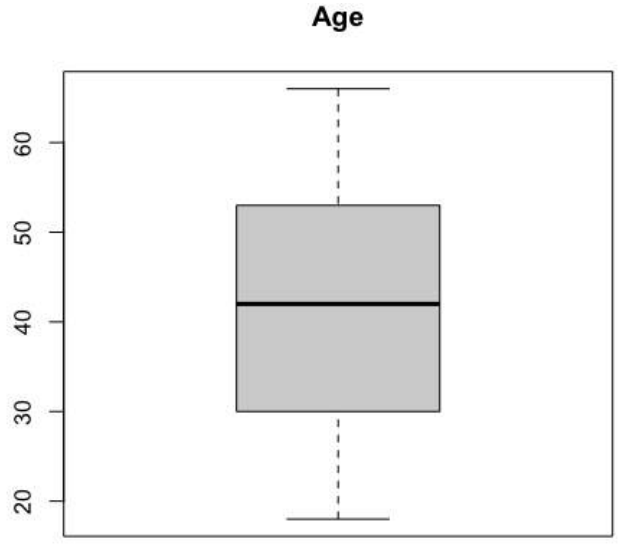
Premium

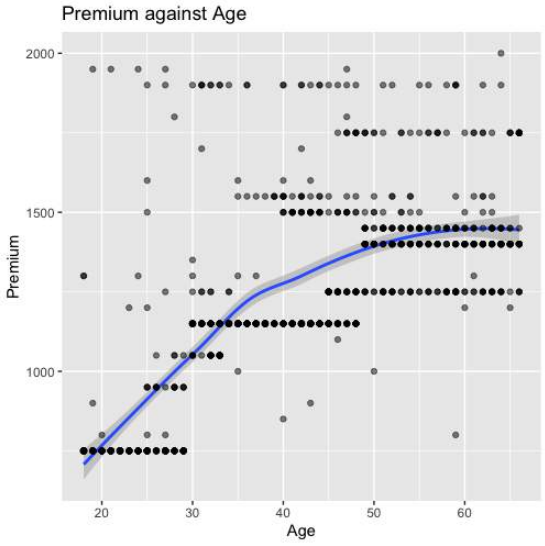


Based on the histogram, I can tell that the more popular annual premium payable among the individuals in the dataset are $1150, $750, and $1400. The mean is $1216, and it has 2 outliers as shown in the boxplot.

Age

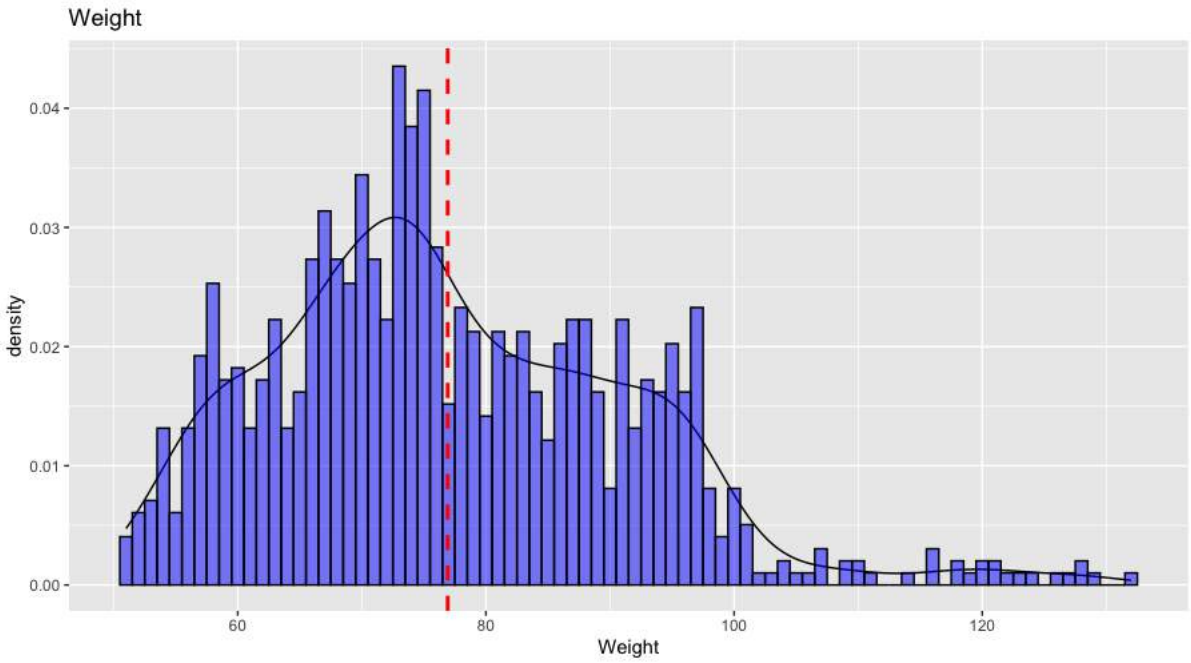


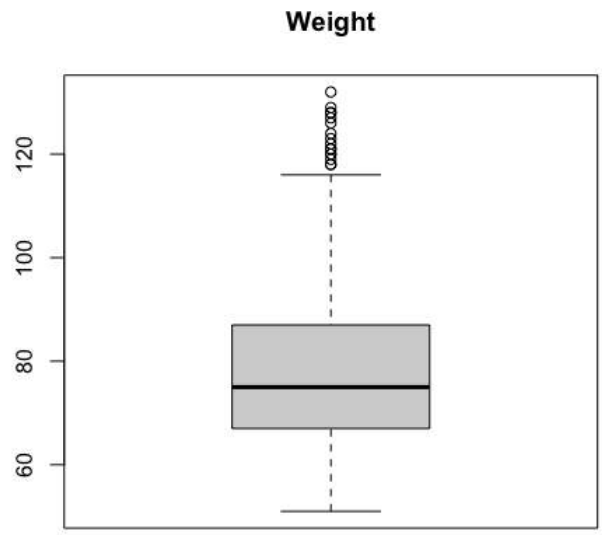


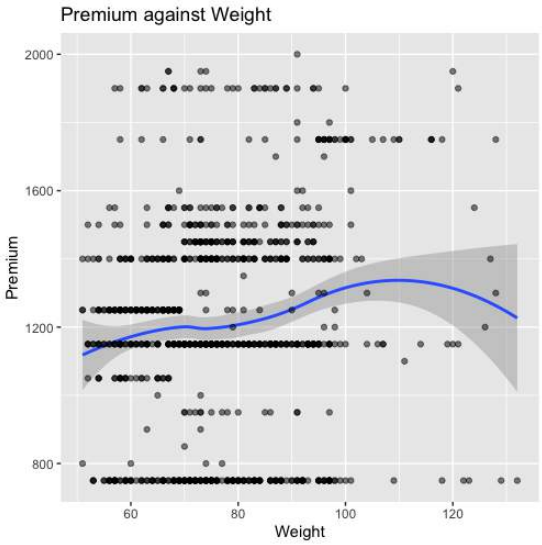


Based on the histogram, generally, there is an even distribution of age in this dataset. The mean is 41.72. There are no outliers as shown in the boxplot. The scatter plot suggests that there is a relatively strong relationship between Premium and Age as the smooth line is almost a straight linear line. This corresponds with the relatively strong correlation value at 0.70. Immediately, you can tell that Age will be a significant variable during machine learning. This is not surprising as older individuals will have more health issues and have a greater risk associated, thus having a more expensive premium.

Weight

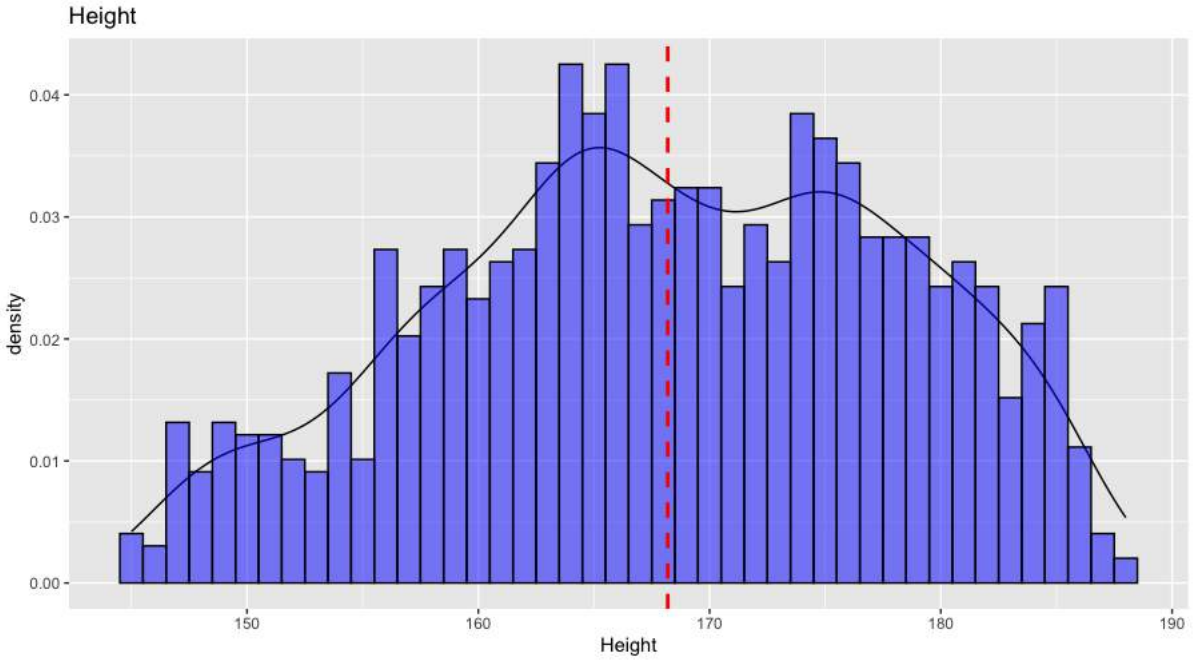


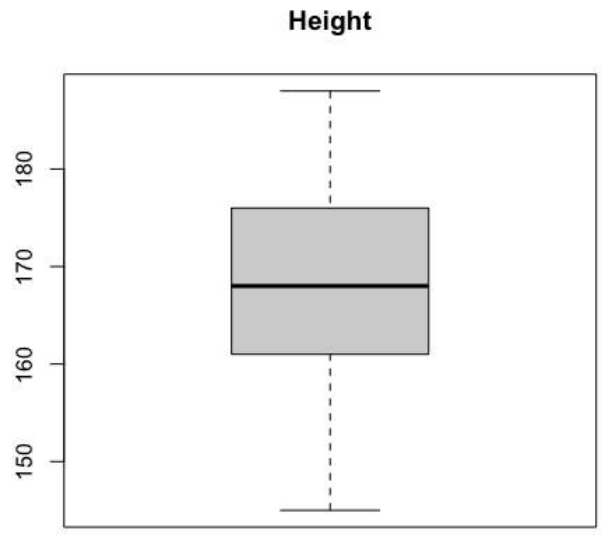


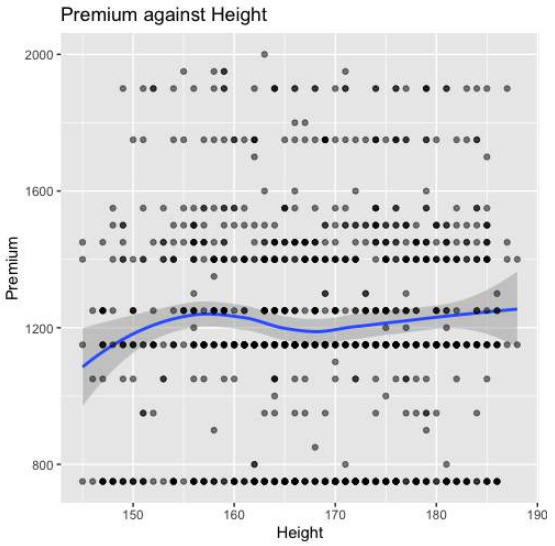


Based on the histogram, there is an uneven distribution of weight in this dataset, with the bulk around the mean at 76.92. At the extreme right, there are a small frequency of values which can be reflected in the outliers shown in the boxplot. The scatter plot suggests that there is a relatively weak relationship between Premium and Weight. This corresponds with the relatively weak correlation value at 0.14.

Height

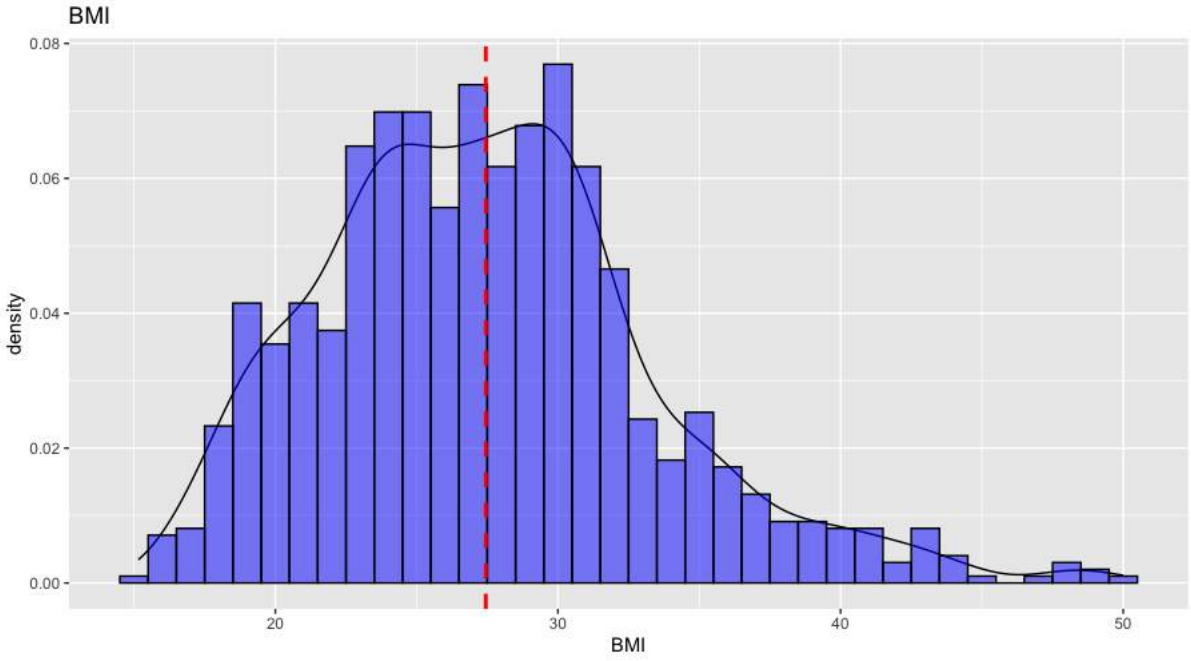


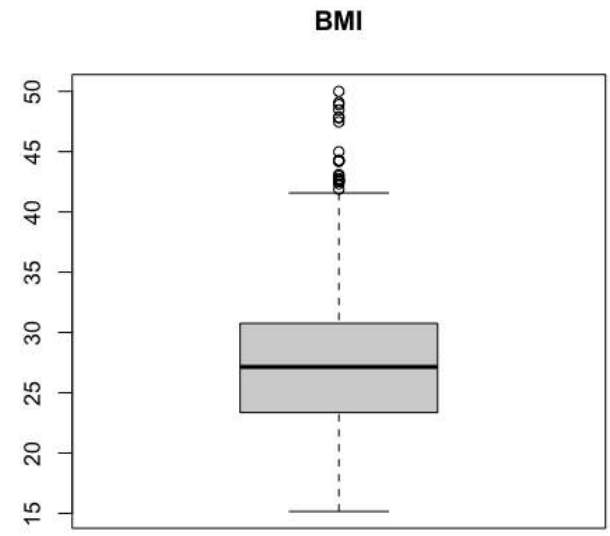


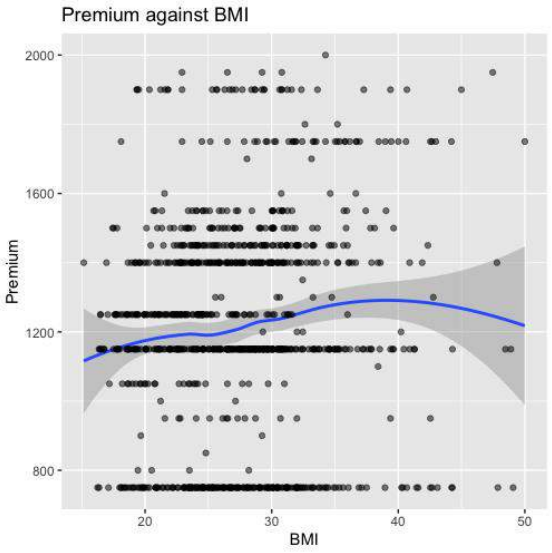


Based on the histogram, there is an uneven distribution of height in this dataset, with the bulk around the mean at 168.2. There are no outliers as shown in the boxplot. The scatter plot suggests that there is a relatively weak relationship between Premium and Height as the smooth line is almost a horizontal line parallel to the x-axis. This corresponds with the relatively weak correlation value at 0.03

BMI

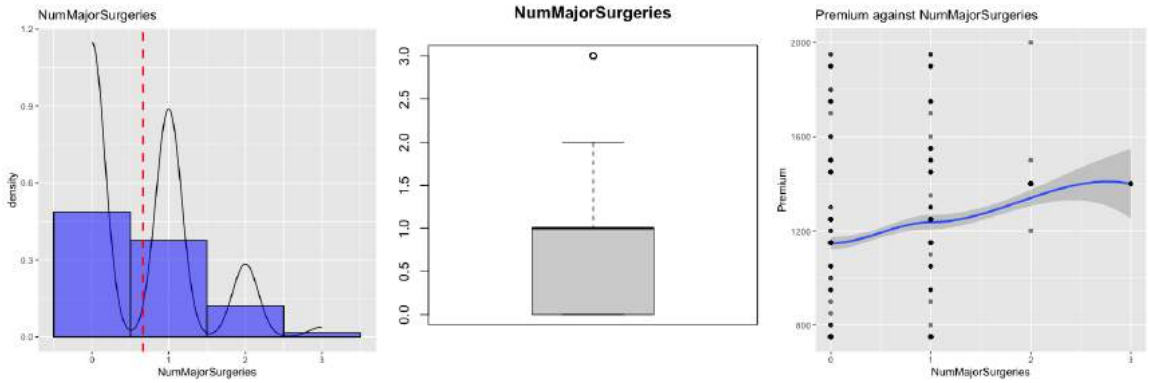




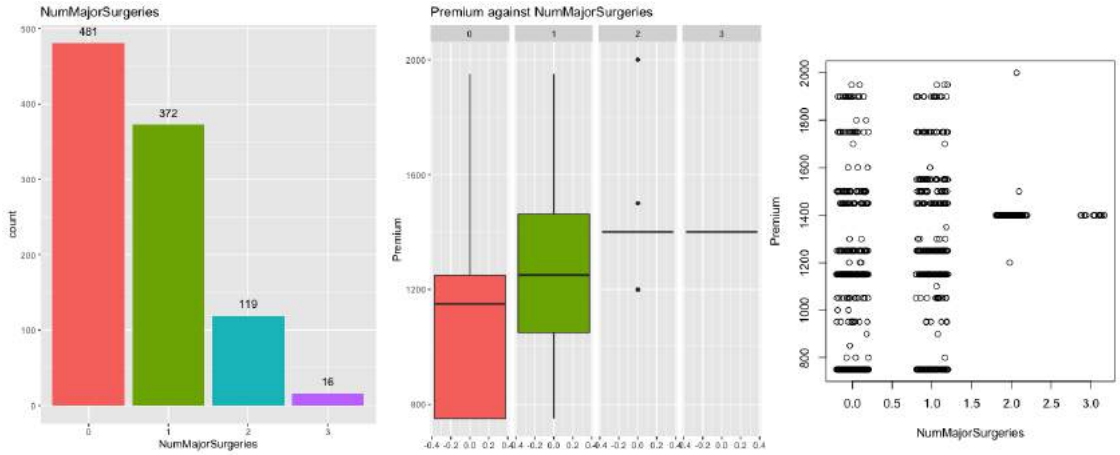


Based on the histogram, there is an uneven distribution of BMI in this dataset, with the bulk around the mean at 27.45. At the extreme right, there are a small frequency of values which can be reflected in the outliers shown in the boxplot. The scatter plot suggests that there is a relatively weak relationship between Premium and BMI. This corresponds with the relatively weak correlation value at 0.11. This is surprising as the higher the BMI, the more health issues the individuals might have, resulting in a greater risk associated and thus a more expensive premium, however, that is not the case here.

NumMajorSurgeries



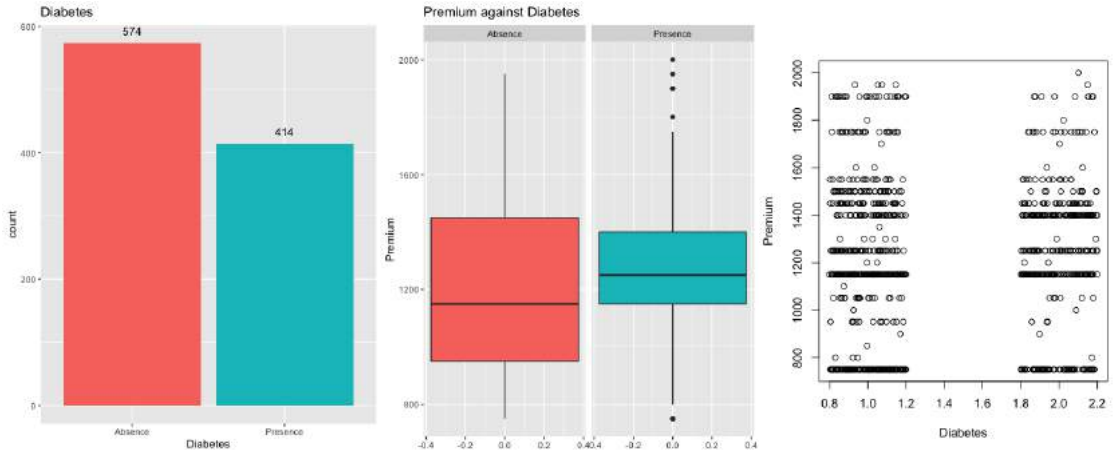
Since NumMajorSurgeries only has 4 possible values, the graphs for continous variable does not tell me anything much, except showing a relatively weak relationship with Premium in the scatter plot which corresponds with a relatively weak correlation value at 0.27. Thus, since it acts more like a categorical variable, I plot three additional graphs below that might tell me more information.



As seen from the bar plot, not surprisingly, the number of individuals with no major surgeries is the highest at 481, following by those who have 1 major surgery at 373, those who have 2 at 119 and the number of individuals with 3 major surgeries is the lowest at 16. Not surprisingly, the mean for Premium generally increased as NumMajorSurgeries increased, from 1147 to 1237, to 1404, to 1400. Other than some outliers for NumMajorSurgeries equals to 2, there are no outliers for the rest. However, I expected individuals with 2 to 3 major surgeries will be paying more for premium but most of them are paying $1400, only slighly higher than average, as seen in the scattered plot with jittered x.

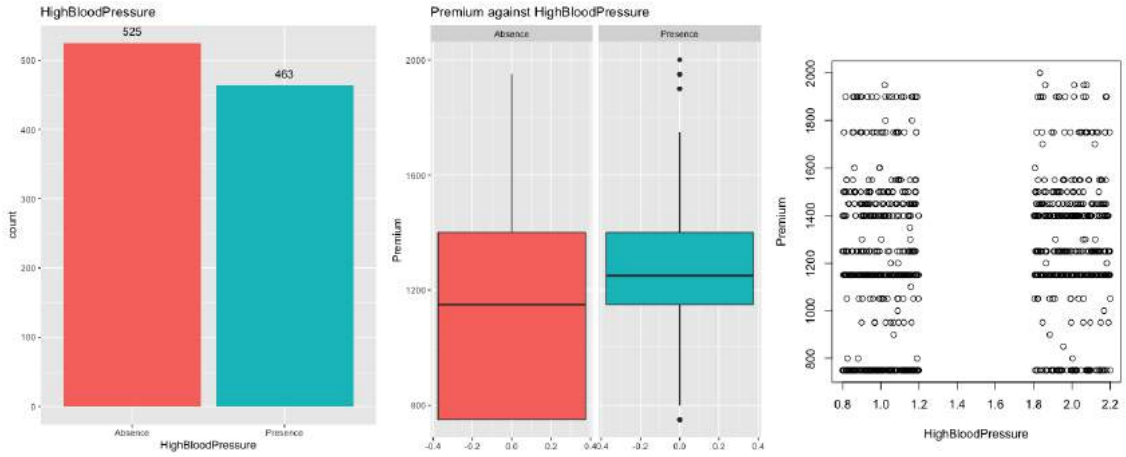
<u>Categorical Variable</u>

Diabetes



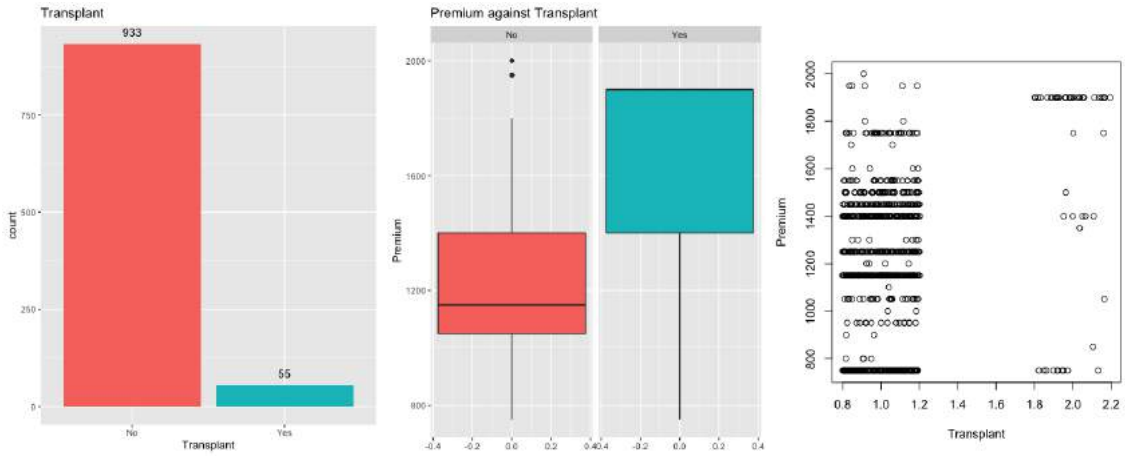
As seen from the bar plot, there are more individuals who do not have diabetes at 574 than the individuals who have diabetes at 414. Not surprisingly, the mean for Premium is lower for absence of Diabetes at 1196 than for the presence of Diabetes at 1245. The scatter plot with jittered x shows a generally even distribution of Premium for both categories.

HighBloodPressure



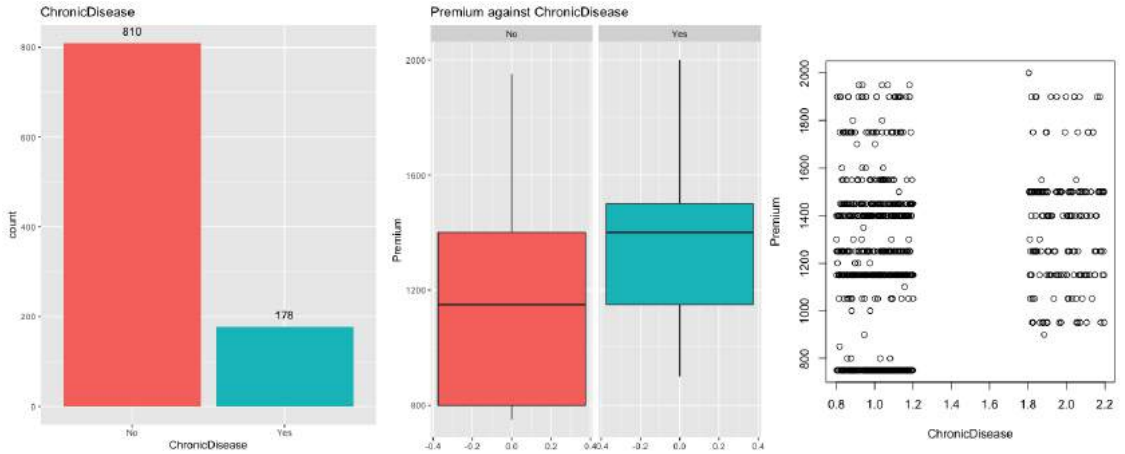
As seen from the bar plot, there are more individuals who do not have high blood pressure at 525 than the individuals who have high blood pressure at 463. Not surprisingly, the mean for Premium is lower for absence of high blood pressure at 1167 than for the presence of high blood pressure at 1272. The scatter plot with jittered x shows a generally even distribution of Premium for both categories.

Transplant



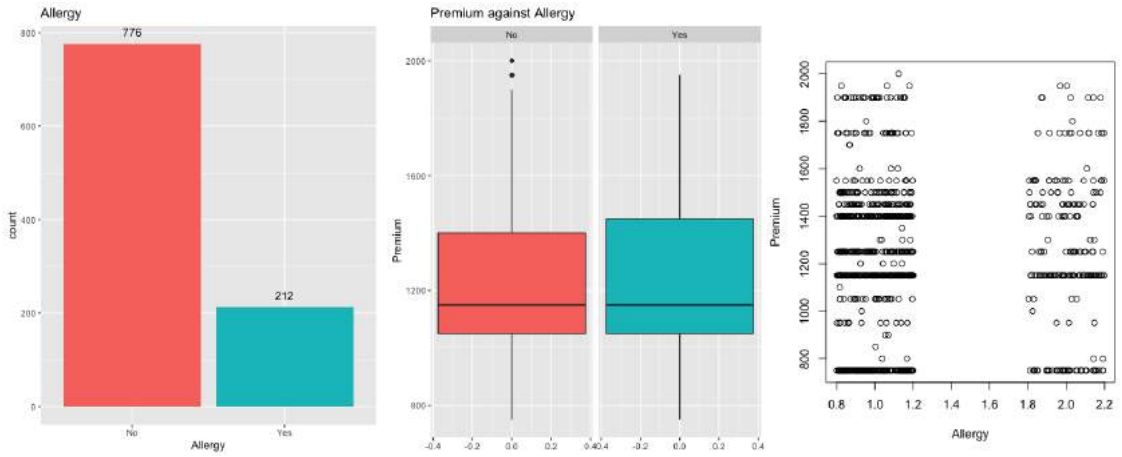
As seen from the bar plot, there are significantly more individuals who do not have transplant at 933 than the individuals who have transplant at 55. Not surprisingly, the mean for Premium is significantly lower for absence of Diabetes at 1194 than for the presence of Diabetes at 1588. The scatter plot with jittered x shows a generally even distribution of Premium for the category without transplant, but is concentrated at 1900 for the category with transplant. This shows that Transplant will be a significant variables in machine learning at more than half of the individuals who have transplant is paying a high premium.

ChronicDisease



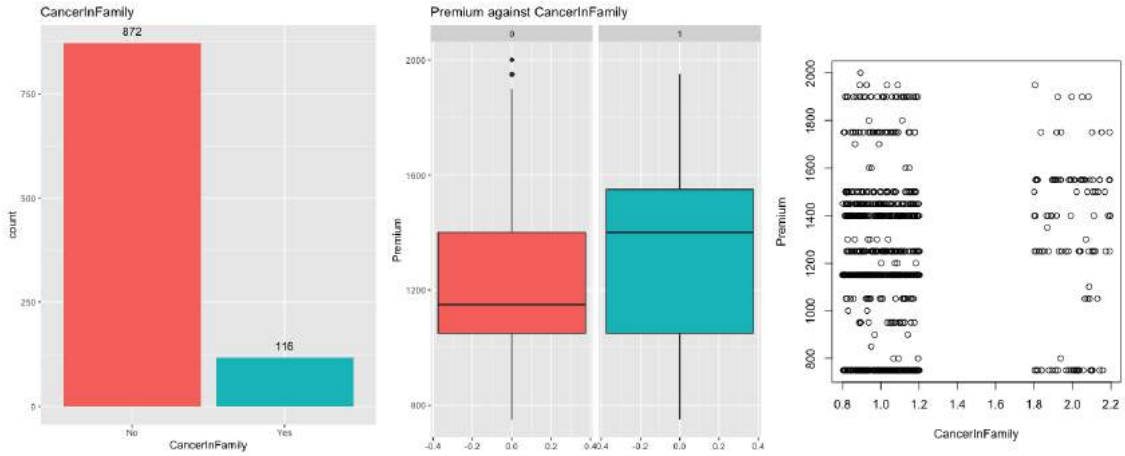
As seen from the bar plot, there are significantly more individuals who do not have chronic disease at 810 than the individuals who have chronic disease at 178. Not surprisingly, the mean for Premium is lower for absence of chronic disease at 1186 than for the presence of chronic disease at 1356. The scatter plot with jittered x shows a generally even distribution of Premium for both categories.

Allergy



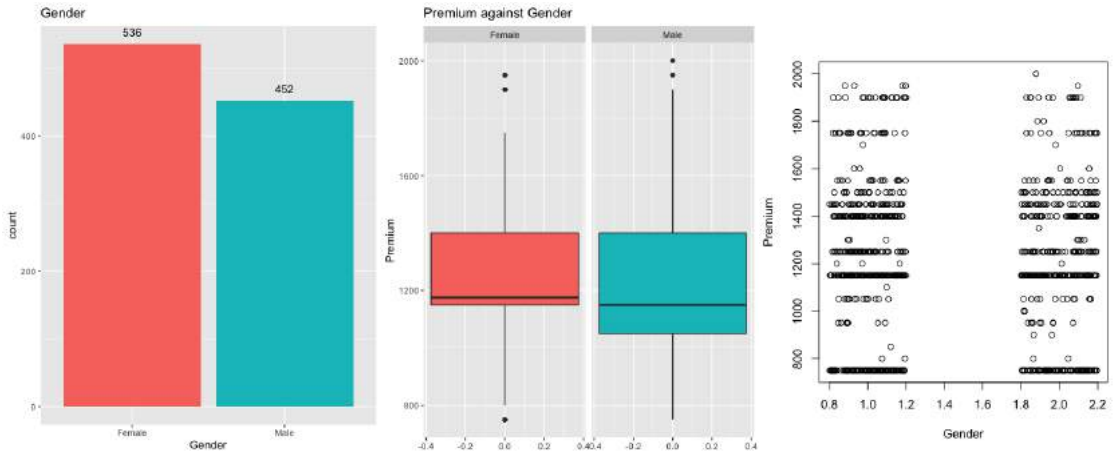
As seen from the bar plot, there are more individuals who do not have an allergy at 776 than the individuals who have an allergy at 212. The mean for Premium is about the same for both groups of individuals who do not have and have an allergy at 1214 and 1224 respectively. This shows that allergy does not affect the premium payable much. The scatter plot with jittered x shows a generally even distribution of Premium for both categories.

CancerInFamily



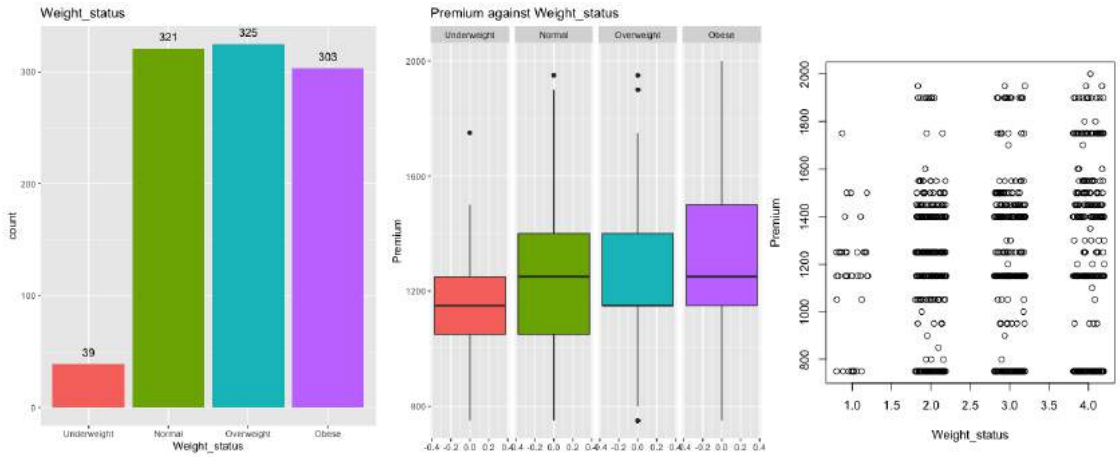
As seen from the bar plot, there are more individuals who do not have cancer in family at 872 than the individuals who have cancer in family at 116. Not surprisingly, the mean for Premium is lower for absence of cancer in family at 1207 than for the presence of cancer in family at 1288. The scatter plot with jittered x shows a generally even distribution of Premium for both categories.

Gender



As seen from the bar plot, there are more females at 574 than males at 414. The mean for Premium is higher for females at 1224 than for males at 1207, but there is not much difference. This shows that gender does not affect the premium payable much. The scatter plot with jittered x shows a generally even distribution of Premium for both categories.

Weight_status



As seen from the bar plot, there are the greatest number of individuals who are overweight at 325, followed by those have a healthy weight at 321, followed by those that are obese at 303 and lastly, those that are underweight at 29. I expected those who have a healthy weight to have the lowest mean for premium and those who are obese to have the highest mean for premium and they have a greater risk associated with them. However, the lowest mean is for those that are underweight at 1133, followed by 1184 for those who have a healthy weight, 1216 for those that are overweight and lastly, 1262 for those that are obese. The scatter plot with jittered x shows a generally even distribution of Premium for all categories.

Answer to Q4:

Use 1 SE optimal CART and one other technique learnt in this course.

I will be using Classification and Regression Tree (CART) and Linear Regression to determine which model does better at predicting Premium based on all the x-variables given.

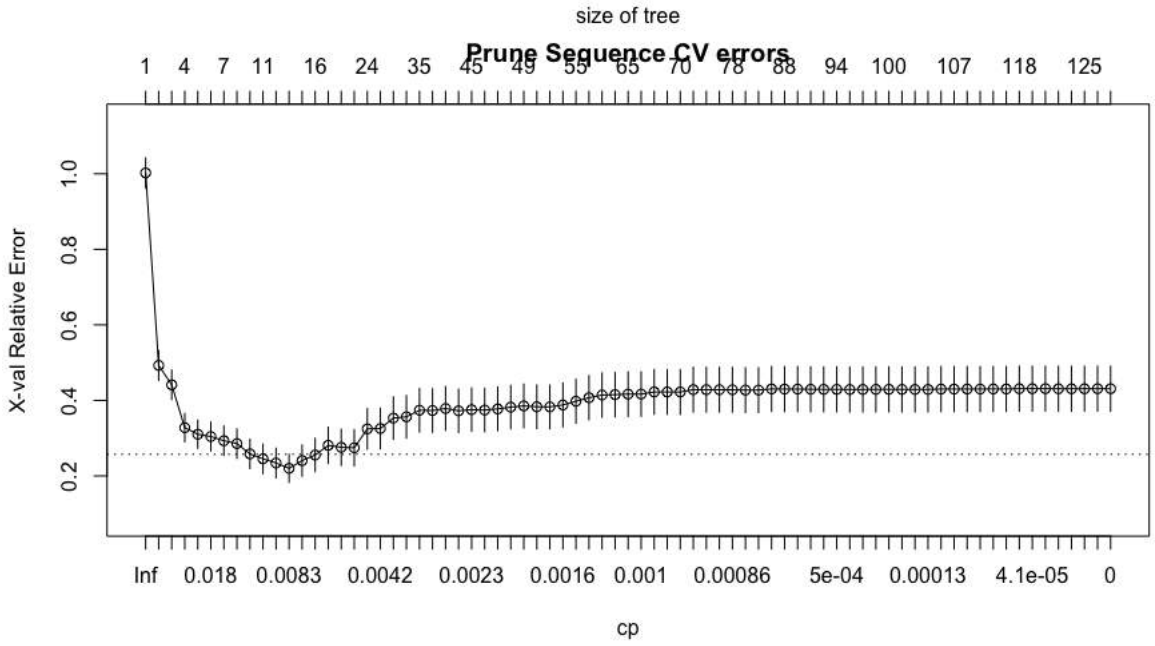
For train-test split, I will be using a split ratio of 0.7, meaning 70% of data is in training set and 30% of data is in testing set. Set seed will be standardised to 123 to ensure that the results each time will be the same.

CART

Using CART, a regression tree model will be created to predict an individual's annual premium payable.

a. What is the 10-fold cross validation RMSE and number of splits in the 1SE Optimal CART?

The CART models were performed on the entire dataset and pruned with an optimal CP value that maximizes the cross-validation accuracy. Optimal CP is given by the simplest tree which is just below the error cap that is calculated by adding the min CV error and 1SE (standard error) which is represented by the horizontal dashed line in the Prune Sequence I plotted below. Using eye power, the tree that is just below the line is the 10th tree. To confirm my answer, I extracted the Optimal Tree via code.



After pruning the maximal tree, I print CP to calculate the errors.

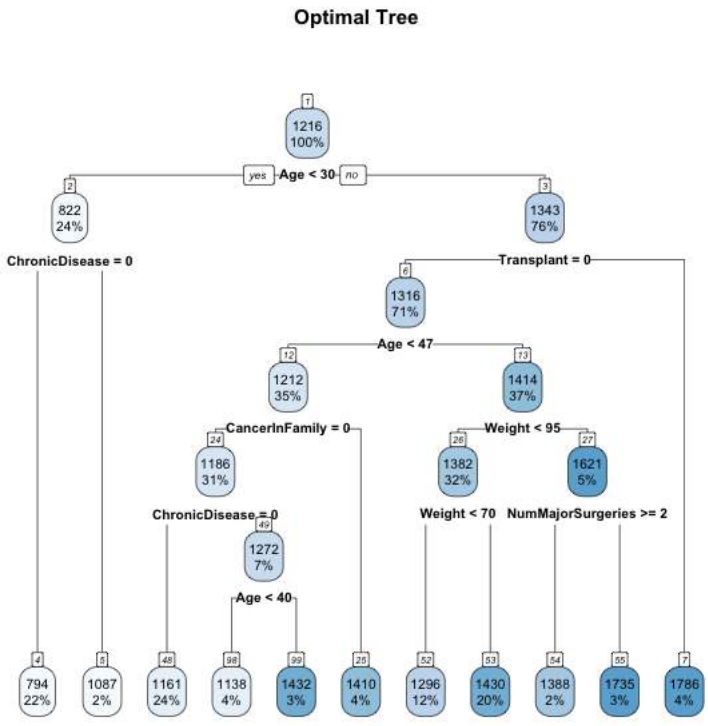
```
> printcp(cart2)

Regression tree:
rpart(formula = Premium ~ ., data = premium.dt, method = "anova",
    control = rpart.control(minsplit = 2, cp = 0))

Variables actually used in tree construction:
[1] Age              CancerInFamily   ChronicDisease   NumMajorSurgeries
[5] Transplant       Weight

Root node error: 96355891/988 = 97526

n= 988

        CP nsplit rel error  xerror     xstd
1  0.511005      0   1.00000 1.00231 0.039597
2  0.090658      1   0.48899 0.49258 0.039389
3  0.074626      2   0.39834 0.44112 0.038944
4  0.025313      3   0.32371 0.32775 0.037361
5  0.018522      4   0.29840 0.31029 0.037178
6  0.017986      5   0.27988 0.30428 0.037981
7  0.013492      6   0.26189 0.29315 0.038900
8  0.013421      7   0.24840 0.28557 0.038978
9  0.010940      8   0.23498 0.25833 0.038850
10 0.010484     10   0.21310 0.24530 0.039028
```

The errors I calculate are as shown below.

10-fold cross validation RMSE is 144.1616

```
> dataset.RMSE <- sqrt(0.21310 * 97526)
> dataset.RMSE
[1] 144.1624
> RMSE.cart2 <- sqrt(mean(residuals(cart2)^2))
> RMSE.cart2
[1] 144.1616
> CV.RMSE <- sqrt(0.24530 * 97526)
> CV.RMSE
[1] 154.671
```

The following are the resulting plot of the pruned CART model.



The number of splits in the 1SE Optimal CART is 10.

b. Identify the key predictors of premium.

Without NA values, the key predictors of premium are Age, Transplant, Weight, NumMajorSurgeries, ChronicDisease and CancerInFamily as seen from the pruned tree. With NA values, the other key predictors of premium can be found using variable importance. They are Age, Transplant, Weight, NumMajorSurgeries, BMI, ChronicDisease, CancerInFamily, HighBloodPressure, Diabetes, Weight_status, Height and Gender. Their variable importance can be seen below, the second one is scaled by percentage.

```
> cart2$variable.importance
              Age        Transplant            Weight NumMajorSurgeries     ChronicDisease
      58053673.89        8735453.20        4949647.20        3514204.93         2430730.17
              BMI    CancerInFamily HighBloodPressure          Diabetes      Weight_status
       2092104.85        1733029.54        1096521.12        1012173.34          536730.56
           Height            Gender
        173685.85          94331.79
> scaledVarImpt <- round(100*cart2$variable.importance/sum(cart2$variable.importance))
> scaledVarImpt
              Age        Transplant            Weight NumMajorSurgeries     ChronicDisease
               69                10                 6                 4                  3
              BMI    CancerInFamily HighBloodPressure          Diabetes      Weight_status
                2                 2                 1                 1                  1
           Height            Gender
                0                 0
```

c. Is BMI or Gender important in determining premium?

<u>CART</u>

Variable importance is based on 3 factors. If a variable is used for splitting, how close is it to the root, and if it appeared as a surrogate. Even if BMI and Gender did not appear in the pruned tree, they appeared in the variable importance, meaning that it is a surrogate. It is not important in determining premium when there are no missing values for the any of the key predictors before them, but it will be important in determining premium when there are missing values.

<u>Linear Regression</u>

BMI and Gender are both not in my final Linear Regression model. If you look at their p-values from the complete linear regression, Gender at 0.7457, and BMI at 0.6112. The lower the p-value, the more important and significant the variable is in determining the y variable. Their p-values are much higher than the limit of 0.05 for significant variables. Thus, they are not important in determining premium for linear regression.

d. Evaluate and compare the predictive accuracy of the two techniques on a 70-30 train-test split. Present testset RMSE results in a table.

After that, the CART models were performed on the trainset.

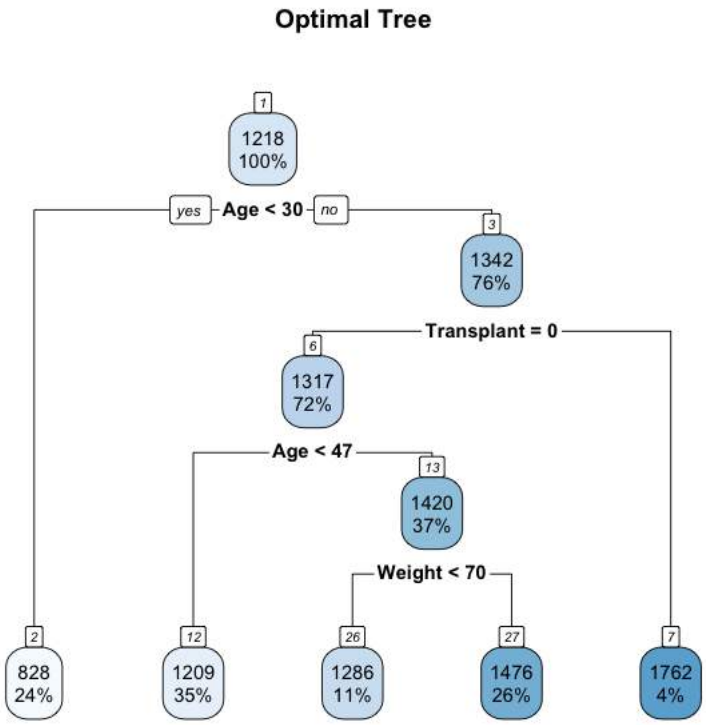
After pruning the maximal tree, I print CP to calculate the errors.

```
> printcp(cart2)

Regression tree:
rpart(formula = Premium ~ ., data = trainset, method = "anova",
    control = rpart.control(minsplit = 2, cp = 0))

Variables actually used in tree construction:
[1] Age        Transplant Weight

Root node error: 67783119/691 = 98094

n= 691

        CP nsplit rel error  xerror     xstd
1 0.491804      0   1.00000 1.00364 0.047641
2 0.080734      1   0.50820 0.51122 0.049765
3 0.028060      3   0.34673 0.40013 0.049989
4 0.025114      4   0.31867 0.36936 0.050685
```

The trainset RMSE, rel error multiple by the root node error then root it, is 176.8039.

The CV RMSE, xerror multiple by the root node error then root it, is 190.3471.

The following are the resulting plot of the pruned CART model.



After obtaining the CART model using the trainset data, I can then use the testset data to test the accuracy of the model we have constructed. I first use the CART model to predict values of Premium with the testset data's given variables. I am then able to find the RMSE value and from there determine the accuracy of our model.

The testset RMSE is 157.5959.

Linear Regression

Linear Regression is chosen as the other technique as the response variable, Premium, is a continuous variable.

I created a total of 3 linear regression objects, one for all the variables, one for after I removed variables based on variance inflation factor (VIF) values, and the last one for after further removal of variables based on Akaike Information Criterion (AIC). I apply all of them to the train and test sets and find the root mean square error (RMSE) for both, and display my findings in the table below.

Below is the summary of the complete model.

```
> m2 <- lm(Premium ~., data = trainset)
> summary(m2)

Call:
lm(formula = Premium ~ ., data = trainset)

Residuals:
    Min      1Q  Median      3Q     Max
-620.31 -108.52  -15.11   92.61 1221.18

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -328.2312   740.6906  -0.443   0.6578
Age                  16.5216     0.6035  27.376  < 2e-16 ***
Diabetes1           -13.7755    15.2022  -0.906   0.3652
HighBloodPressure1    0.2222    15.2989   0.015   0.9884
Transplant1         366.9838    31.4432  11.671  < 2e-16 ***
ChronicDisease1     129.0996    19.0540   6.775 2.70e-11 ***
Height                3.5196     4.3633   0.807   0.4202
Weight                0.3633     4.6139   0.079   0.9373
Allergy1             23.3632    18.7602   1.245   0.2134
CancerInFamily1     102.0163    23.2143   4.395 1.29e-05 ***
NumMajorSurgeries   -24.8600    11.5241  -2.157   0.0313 *
Gender1              -4.7670    14.6913  -0.324   0.7457
BMI                   6.3221    12.4321   0.509   0.6112
Weight_status1       -3.0518    42.9803  -0.071   0.9434
Weight_status2       33.5310    53.0987   0.631   0.5279
Weight_status3       51.8698    64.5129   0.804   0.4217
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 189.8 on 675 degrees of freedom
Multiple R-squared:  0.6413,    Adjusted R-squared:  0.6333
F-statistic: 80.45 on 15 and 675 DF,  p-value: < 2.2e-16
```

High VIF values reveal multicollinearity between the different variables. Therefore, I removed the variable with the highest value, and did so until VIF was less than 10. It was not surprisingly that BMI has the highest VIF value as it was calculated from Height and Weight. After removing BMI, the RMSE for both train and test set decreased, proving that removing instances of multicollinearity is essential for a good linear regression model.

```
> vif(m1)
                      GVIF Df GVIF^(1/(2*Df))
Age                1.330631  1        1.153530
Diabetes           1.082069  1        1.040226
HighBloodPressure  1.114741  1        1.055813
Transplant         1.007696  1        1.003840
ChronicDisease     1.025800  1        1.012818
Height            39.576454  1        6.290982
Weight            89.895453  1        9.481321
Allergy            1.037395  1        1.018526
CancerInFamily     1.088873  1        1.043491
NumMajorSurgeries  1.366025  1        1.168771
Gender             1.010628  1        1.005300
BMI              111.592560  1       10.563738
Weight_status      6.703371  3        1.373142
```

```
> vif(m1.vif)
                      GVIF Df GVIF^(1/(2*Df))
Age                1.329691  1        1.153122
Diabetes           1.082069  1        1.040226
HighBloodPressure  1.112722  1        1.054857
Transplant         1.005340  1        1.002666
ChronicDisease     1.025780  1        1.012808
Height             2.497190  1        1.580250
Weight             3.921580  1        1.980298
Allergy            1.037217  1        1.018438
CancerInFamily     1.088327  1        1.043229
NumMajorSurgeries  1.365192  1        1.168414
Gender             1.010178  1        1.005076
Weight_status      5.244917  3        1.318125
```

Below is the summary of the VIF model.

```
> summary(m1.vif)

Call:
lm(formula = Premium ~ Age + Diabetes + HighBloodPressure + Transplant +
    ChronicDisease + Height + Weight + Allergy + CancerInFamily +
    NumMajorSurgeries + Gender + Weight_status, data = premium.dt)

Residuals:
    Min      1Q  Median      3Q     Max
-652.98 -112.63  -13.88   90.06 1230.00

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        51.7786   143.5824   0.361 0.718462
Age                16.4975     0.4915  33.564  < 2e-16 ***
Diabetes1         -21.6138    12.5400  -1.724 0.085101 .
HighBloodPressure1  9.5276    12.5733   0.758 0.448775
Transplant1       394.8528    26.0112  15.180  < 2e-16 ***
ChronicDisease1   134.0258    15.6748   8.550  < 2e-16 ***
Height              1.6557     0.9320   1.777 0.075952 .
Weight              1.5415     0.8261   1.866 0.062329 .
Allergy1           13.1259    14.7558   0.890 0.373933
CancerInFamily1   114.0741    19.2761   5.918 4.51e-09 ***
NumMajorSurgeries -32.4942     9.2828  -3.500 0.000485 ***
Gender1            -3.6678    11.9998  -0.306 0.759936
Weight_status1      0.8538    34.4720   0.025 0.980245
Weight_status2     48.1086    41.3741   1.163 0.245208
Weight_status3     93.0403    51.7953   1.796 0.072755 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 187 on 973 degrees of freedom
Multiple R-squared:  0.647,     Adjusted R-squared:  0.642
F-statistic: 127.4 on 14 and 973 DF,  p-value: < 2.2e-16
```

Then, we used the step function to further remove variables based on AIC) meaning that variables that did not provide a significant contribution to the prediction were removed namely, Gender, HighBloodPressure and Allergy. This brings me back to the data exploration where these 3 variables did not have a significant difference in mean of Premium for both categories, suggesting that they do not have a significant impact on Premium.

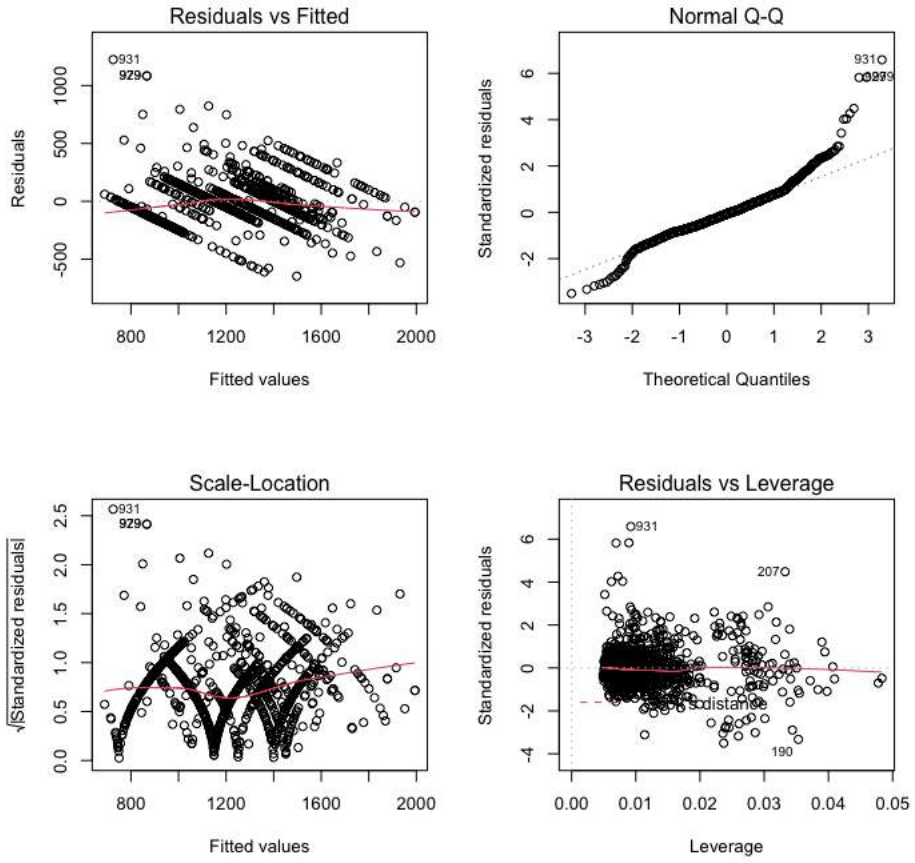
```
> m1.vif.aic <- step(m1.vif)
Start:  AIC=10351.13
Premium ~ Age + Diabetes + HighBloodPressure + Transplant + ChronicDisease +
    Height + Weight + Allergy + CancerInFamily + NumMajorSurgeries +
    Gender + Weight_status

                   Df Sum of Sq      RSS   AIC
- Gender            1      3265 34013484 10349
- HighBloodPressure 1     20071 34030290 10350
- Allergy           1     27658 34037877 10350
<none>                         34010219 10351
- Diabetes          1    103839 34114058 10352
- Height            1    110321 34120540 10352
- Weight            1    121721 34131940 10353
- Weight_status     3    319891 34330110 10354
- NumMajorSurgeries 1    428304 34438523 10362
- CancerInFamily    1   1224143 35234362 10384
- ChronicDisease    1   2555471 36565690 10421
- Transplant        1   8054621 42064840 10559
- Age               1  39376490 73386709 11109
```

Below is the summary of the AIC model. One thing to take note is that Diabetes and NumMajorSurgeries have negative coefficients. However, I expected when an individual has diabetes, and when the number of major surgeries increased, the insurance premium will increase as there is a greater risk associated.

```
> summary(m1.vif.aic)

Call:
lm(formula = Premium ~ Age + Diabetes + Transplant + ChronicDisease +
    Height + Weight + CancerInFamily + NumMajorSurgeries + Weight_status,
    data = premium.dt)

Residuals:
    Min      1Q  Median      3Q     Max
-647.08 -112.81  -12.87   88.04 1224.80

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        54.5752   143.0371   0.382 0.702881
Age                16.5376     0.4849  34.103  < 2e-16 ***
Diabetes1         -21.8823    12.4529  -1.757 0.079196 .
Transplant1       394.4573    25.9825  15.182  < 2e-16 ***
ChronicDisease1   133.8537    15.6222   8.568  < 2e-16 ***
Height              1.6598     0.9302   1.784 0.074674 .
Weight              1.5084     0.8249   1.829 0.067755 .
CancerInFamily1   115.7690    19.1950   6.031 2.31e-09 ***
NumMajorSurgeries -30.5865     9.1086  -3.358 0.000815 ***
Weight_status1      2.1097    34.4237   0.061 0.951145
Weight_status2     49.9460    41.3071   1.209 0.226903
Weight_status3     95.2377    51.6865   1.843 0.065690 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 186.8 on 976 degrees of freedom
Multiple R-squared:  0.6465,    Adjusted R-squared:  0.6425
F-statistic: 162.3 on 11 and 976 DF,  p-value: < 2.2e-16
```

After AIC, the train and test set RMSE increased, but only by a small amount. However, given a minimal increase in RMSE, a simpler model is still a better choice.

Linear regression model makes a lot of assumptions, to test these assumptions, I plotted the model diagnostics plots.



The first graph shows that points are randomly distributed close to zero, thus it fits the linear relationship assumption. The second graph shows that most points fall close to the diagonal dotted line, thus it fits the normal distributed assumption. The third graph shows that plotted points are randomly distributed with the same spread at each vertical slice, thus it fits the constant error variance assumption. The fourth graph shows there are few influential outliers. Thus, showing that linear regression is appropriate, and the predicted results should be accurate.

Model	**Trainset RMSE**	**Testset RMSE**
Complete	187.5841	183.3648
VIF	185.5352	179.1065
AIC	185.6709	179.2729
CART	176.8039	157.5959

Highlighted in green are the train/test errors of my finalised models

Generally, train and test error were not hugely different, showing that there was no significant under or overfitting in the models. The prediction results indicated that CART has a better performance and

accuracy. Both the trainset and testset errors are lower than Linear Regression Model. Moreover, CART uses only 4 variables for prediction, and uses at most 9 variables in the worst case scenario. Meanwhile, Linear Regression uses 9 variables for prediction.

Answer to Q5:

Explain the limitations of your analysis.

There are more variables that affect annual premium payable that are not included in the data, for example, the area which the individuals live in and the type of coverage of the insurance premium. These variables might have a more significant impact on Premium and will thus generate a more accurate result.

The dataset provided is skewed. As seen from the data exploration, some categories only have a few individuals, for example, there are only 55 individuals with transplant, and only 39 individuals who are underweight. Hence, this might not provide a good estimate, and will affect the result of machine learning and thus, limit my analysis.

To counter the above two limitations, the insurance company can collect more data. Machine learning will also be more accurately with more data provided. Now, I will be looking at the limitations of the two techniques I used.

CART

The structure of the tree will drastically change when the insurance company collects new data. In other words, a single decision tree is unstable. For instance, a small change in the data can cause the algorithm to select different variables and cut-points. A sub-optimal first split would also influence all subsequent splits below it due to a tree's hierarchical nature. Compared to other statistical learning and data mining methods, a single tree would also not have the best overall predictive quality because it is not robust to small changes in the data (Hong, 2018).

To counter this, bagging, random forests and boosting can be used to build several trees to construct powerful prediction models. I would recommend the insurance company to not use the same rules for too long and update the model when there is new data, especially in this unpredictable world with the ongoing Covid-19 pandemic that will result in the risk associated to an individual unpredictable.

Although the rules obtained through the analysis can be tested on new data, it must be remembered that the model is built based upon the sample without making any inference about the underlying probability distribution (Ojha, 2017).

Linear Regression

Linear Regression is inflexible. It only looks at linear relationships between dependent and independent variables. The situation where it fails to capture the data properly often arises.

Moreover, unlike CART, to create a simpler Linear Regression model, I have to remove variables through VIF and AIC. However, all variables have some impact on the annual premium payable, thus removing the variable entirely from the model might lead to inaccuracy in my analysis.

Answer to Q6:

Is CART successful in this application? Explain.

Yes, CART is successful in this application. Based on my research, the price of the premium depends on a variety of factors, including:

- The type of coverage
- Your age
- The area in which you live
- Any claims filed in the past

(Kagan, 2021)

As seen from the variable importance from the CART model, age is the most significant variable, and the rest of the variables correspond to claims filed in the past, and these will affect the price of the premium, which corresponds with my research, thus suggesting that the CART model is fairly accurate.

The RMSE for the testset for CART might seem high at 157.5959, but as seen from the summary of Premium in the dataset, the price of premium ranges from $750 to $2000, a deviance of $158 seems reasonable. By increasing the number of relevant variables in the data, the RMSE might even decrease even further.

Moreover, sometimes, an individual might be reluctant to share all of their details, or the information was not put into the data properly, leading to possible NA values in the data. CART will still be able to predict the price of premium for that individual as CART is able to handle NA values because of its surrogate functionality.

CART is simple to understand, interpret and visualise. The insurance company does not need prior knowledge on analytics to understand it.

The insurance company is also able to constant update the CART model when there is new data available as it requires little effort for data preparation.

In conclusion, CART is successful in this application as the predicted results seems to be fairly accurate with a relatively low RMSE. Moreover, an insurance company is the one using the model. CART is easy to update, use, and understand, which allow the insurance company to use it with ease.

Reference:

Armel Djangone. (2021, Jan 28). *Analyzing Your Data: Why Do Data Types Matter? — Part 1*. Retrieved From.

https://medium.com/swlh/analyzing-your-data-why-do-data-types-matter-part-1-1aad04e543e7

Julia Kagan. (2021, September 10). *Insurance Premium*. Retrieved From.

https://www.investopedia.com/terms/i/insurance-premium.asp

Data Novia. (2021, January 5). *How to Plot a Smooth Line using GGPlot2*. Retrieved From.

https://www.datanovia.com/en/blog/how-to-plot-a-smooth-line-using-ggplot2/

Amit Kumar Ojha. (2017, May 15). *USE A CLASSIFICATION AND REGRESSION TREE (CART) FOR QUICK DATA INSIGHTS*. Retrieved From.

https://www.isixsigma.com/methodology/lean-methodology/use-a-classification-and-regression-tree-cart-for-quick-data-insights/

Maxwell Hong. (2018, April 28). *Exploratory data mining with Classification and Regression Trees (CART)*. Retrieved From.

https://www.apa.org/science/about/psa/2018/04/classification-regression-trees

https://stackoverflow.com/questions/10222525/replacing-numbers-within-a-range-with-a-factor