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

Class	8
Full Name	Lim Qing Rui
Matriculation Number	U2010816G

** : 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

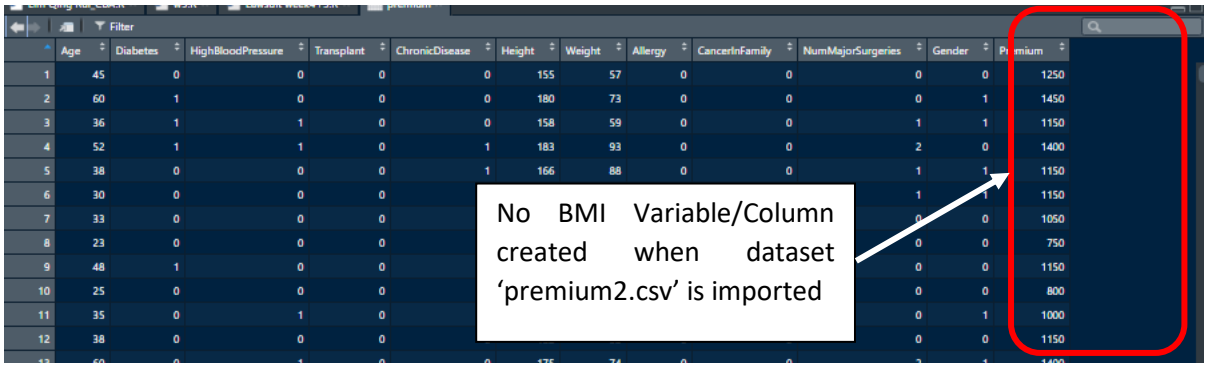
For each question, please start your answer in a new page.

Answer to Q1:

Q1) Create the BMI variable based on CDC definition1 . Show your code.

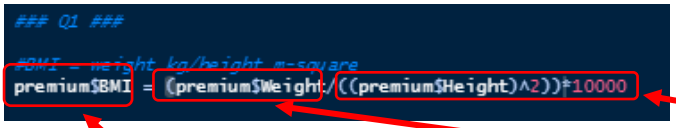
The Body Mass Index (BMI) can be computed using the following formula *(CDC, 2014)*:

$$BMI = \frac{Weight\ (kg)}{[Height\ (m)]^2}\ or\ \frac{Weight\ (kg)}{[Height\ (cm)]^2} * 10000$$



	Age	Diabetes	HighBloodPressure	Transplant	ChronicDisease	Height	Weight	Allergy	CancerInFamily	NumMajorSurgeries	Gender	Premium
1	45	0	0	0	0	155	57	0	0	0	0	1250
2	60	1	0	0	0	180	73	0	0	0	1	1450
3	36	1	1	0	0	158	59	0	0	1	1	1150
4	52	1	1	0	1	183	93	0	0	2	0	1400
5	38	0	0	0	1	166	88	0	0	1	1	1150
6	30	0	0	0						1	1	1150
7	33	0	0	0						0	0	1050
8	23	0	0	0						0	0	750
9	48	1	0	0						0	0	1150
10	25	0	0	0						0	0	800
11	35	0	1	0						0	1	1000
12	38	0	0	0						0	0	1150

No BMI Variable/Column created when dataset 'premium2.csv' is imported



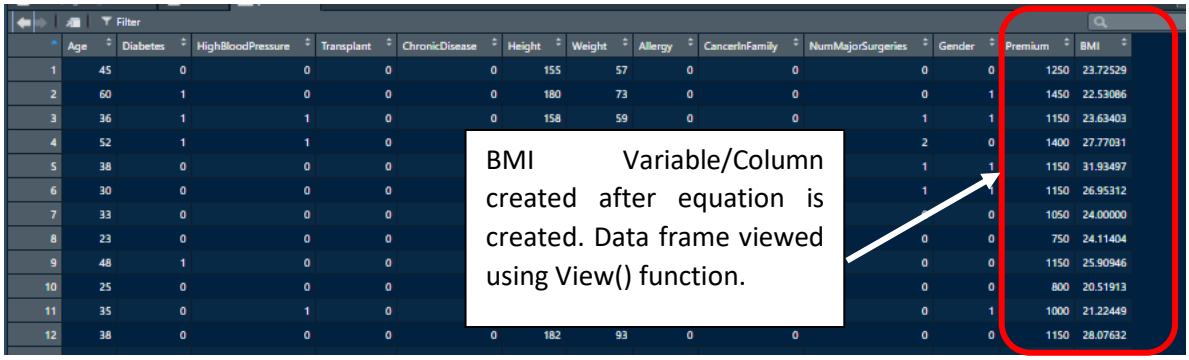
```
### Q1 ###

#BMI = weight_kg/height_m-square
premium$BMI = (premium$Weight)/((premium$Height)^2))*10000
```

Creating a column for BMI in 'premium' data frame

References Weight (kg) from 'premium' data frame

References Height (cm) from 'premium' data frame and converts to appropriate unit of measurement (m^2)

In the code, the second equation of BMI was used to extract the Weight and Height which are in Kg and Cm respectively to create a pseudo-column that is parsed into the data frame 'premium' and added in as the last column.



	Age	Diabetes	HighBloodPressure	Transplant	ChronicDisease	Height	Weight	Allergy	CancerInFamily	NumMajorSurgeries	Gender	Premium	BMI
1	45	0	0	0	0	155	57	0	0	0	0	1250	23.72529
2	60	1	0	0	0	180	73	0	0	0	1	1450	22.53086
3	36	1	1	0	0	158	59	0	0	1	1	1150	23.63403
4	52	1	1	0						2	0	1400	27.77031
5	38	0	0	0						1	1	1150	31.93497
6	30	0	0	0						1	1	1150	26.95312
7	33	0	0	0							0	1050	24.00000
8	23	0	0	0						0	0	750	24.11404
9	48	1	0	0						0	0	1150	25.90946
10	25	0	0	0						0	0	800	20.51913
11	35	0	1	0						0	1	1000	21.22449
12	38	0	0	0	0	182	93	0	0	0	0	1150	28.07632

BMI Variable/Column created after equation is created. Data frame viewed using View() function.

Answer to Q2:

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?

There is **no need to convert the data to factor** for data visualisation and manipulation. The 7 Variables that were intended to be factorised are as follows: (i) Diabetes, (ii) HighBloodPressure, (iii) Transplant, (iv) ChronicDisease, (v) Allergy, (vii) CancerInFamily, (vii) Gender. The factor() function was used and this converted the 7 variables with value of 0 and 1 to a binary/dummy variable that categorises it as true or false (or yes or no) in its respective column in a new dataframe that was created called 'premiumfactor'. The factored variable was then parsed back through the data frame into its respective column (E.g. premiumfactor$Diabetes is parsed back into Diabetes column in the premiumfactor dataframe after factor() was applied). The class() function was then applied to check the class type, ensuring all variables were in factor data type. **(Refer to Appendix A for data conversion code and new dataframe creation)**

Subsequently, both dataframes were parsed into a Linear Regression model to see whether the variables would de different, where all 12 variables (including BMI) were compared to the continuous Y variable. We can see that there are no observable differences between variable output, R-squared values and intercept for both models **(Image below)**. In addition, the data analysis and visualization used in this report does not require the variables to be represented as nominal level independent variables. Hence, factorisation is not required.

Same output for (i) 'Premium' Residual prediction, (ii) coefficients output, standard error and (iii) Residual Standard Error and R-squared values

```
> set.seed(2004)
> premiumQ2NF = lm(Premium ~., data = premium)
> summary(premiumQ2NF)

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

Residuals:
    Min      1Q  Median      3Q     Max
-684.53 -109.06  -17.61   94.02 1210.20

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       659.5764   535.6108   1.231 0.218453
Age                16.4622     0.4924  33.430  < 2e-16 ***
Diabetes          -21.4618    12.5660  -1.708 0.087968 .
HighBloodPressure   8.4902    12.6244   0.673 0.501409
Transplant        395.4578    26.1074  15.147  < 2e-16 ***
ChronicDisease    132.8415    15.7103   8.456  < 2e-16 ***
Height             -2.5689     3.1624  -0.812 0.416807
Weight              5.9400     3.3797   1.758 0.079142 .
Allergy            15.3009    14.7842   1.035 0.300949
CancerInFamily    116.3388    19.2845   6.033 2.29e-09 ***
NumMajorSurgeries -32.6170     9.3002  -3.507 0.000474 ***
Gender             -2.5773    12.0356  -0.214 0.830484
BMI                -6.9234     9.4887  -0.730 0.465779
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 187.6 on 975 degrees of freedom
Multiple R-squared:  0.6439,    Adjusted R-squared:  0.6395
F-statistic: 146.9 on 12 and 975 DF,  p-value: < 2.2e-16
```

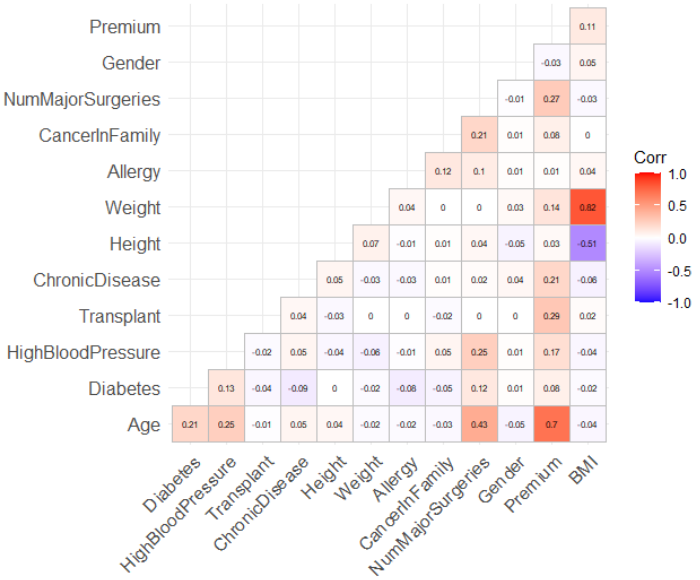
Image 2.1: Non-factorised data for Linear Regression

```
> set.seed(2004)
> premiumQ2Fact = lm(Premium ~., data = premiumfactor)
> summary(premiumQ2Fact)

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

Residuals:
    Min      1Q  Median      3Q     Max
-684.53 -109.06  -17.61   94.02 1210.20

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       659.5764   535.6108   1.231 0.218453
Age                16.4622     0.4924  33.430  < 2e-16 ***
Diabetes1         -21.4618    12.5660  -1.708 0.087968 .
HighBloodPressure1  8.4902    12.6244   0.673 0.501409
Transplant1       395.4578    26.1074  15.147  < 2e-16 ***
ChronicDisease1   132.8415    15.7103   8.456  < 2e-16 ***
Height             -2.5689     3.1624  -0.812 0.416807
Weight              5.9400     3.3797   1.758 0.079142 .
Allergy1           15.3009    14.7842   1.035 0.300949
CancerInFamily1   116.3388    19.2845   6.033 2.29e-09 ***
NumMajorSurgeries -32.6170     9.3002  -3.507 0.000474 ***
Gender1            -2.5773    12.0356  -0.214 0.830484
BMI                -6.9234     9.4887  -0.730 0.465779
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 187.6 on 975 degrees of freedom
Multiple R-squared:  0.6439,    Adjusted R-squared:  0.6395
F-statistic: 146.9 on 12 and 975 DF,  p-value: < 2.2e-16
```

Image 2.2: Factorised data for Linear Regression

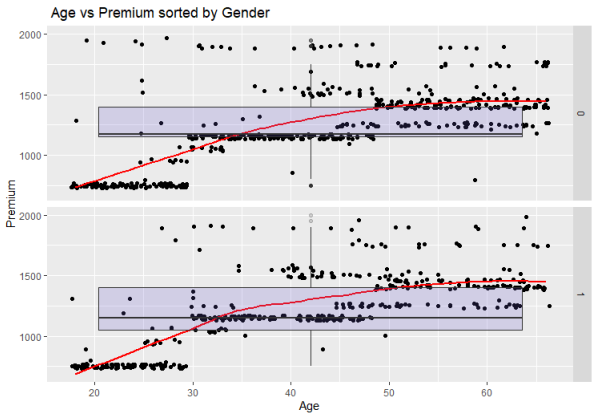
Answer to Q3:

Q3) Explore the data and report on your key findings

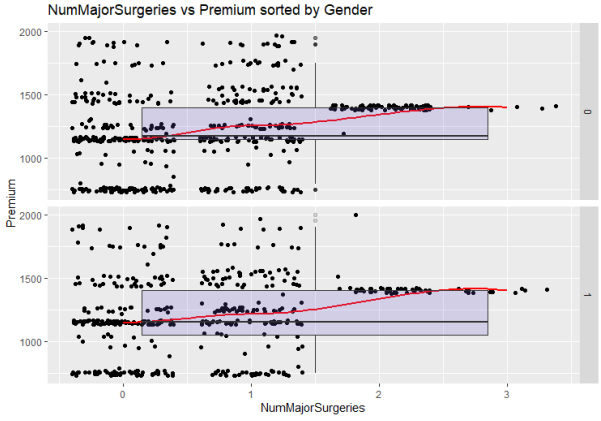


From the correlation plot, we can see that most variables in the dataset given do not exhibit correlative properties with each other as their correlation are within the ranges of -0.25 to 0.25. However, upon closer observation, we can see that Weight is highly correlated to Body Mass Index (BMI), which shows that weight is a determining factor of BMI (Correlation=0.82). Height is also negatively correlated to BMI, and shows an inverse trend which is indicative of the BMI formula in Q1(Correlation=-0.51). Age exhibits a slight positive correlation to NumMajorSurgeries, which could show that age affects health of an individual resulting in occurrences of Surgeries (Correlation=0.43). Age also exhibits a strong positive correlation to Premium, and this leads to a correlation of 0.7. This could imply a health risk leading to higher premiums in insurance for the individual.

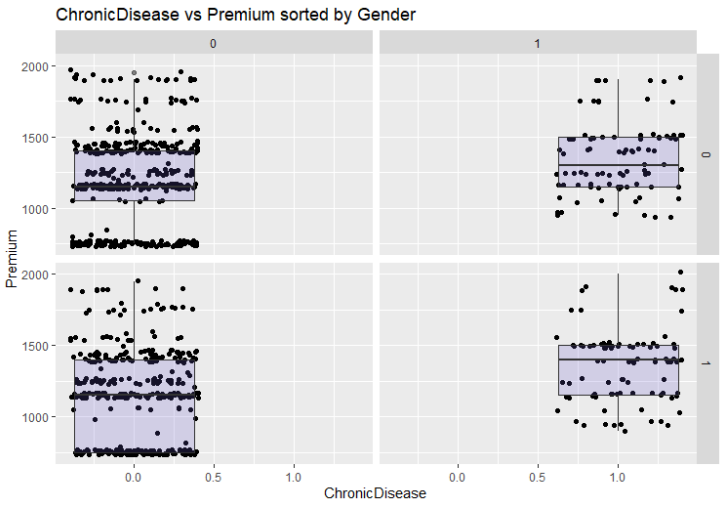
We will now investigate various variables that show **outliers or a trend** in compared to Premium:



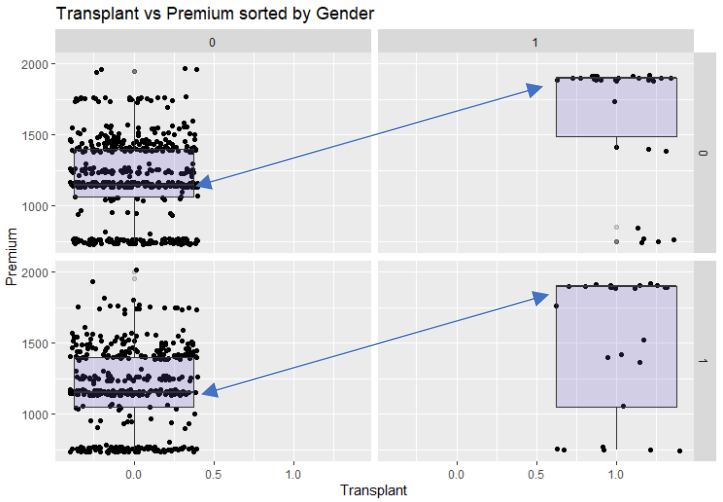
When Age is compared to insurance premium (Sorted by Gender) using a smoothed line, we can see that Premiums are priced similarly across Gender, and this trend is congruent to the strong positive correlation as shown in the Correlation Matrix. However, it can be seen that the first Quartile (25%) for males where Gender = 1 have a lower premium that female. This is similar to a study on gender and premiums which showed that females medical costs during younger years costs 45% higher than males. *(Fontinelle A., 2021)*



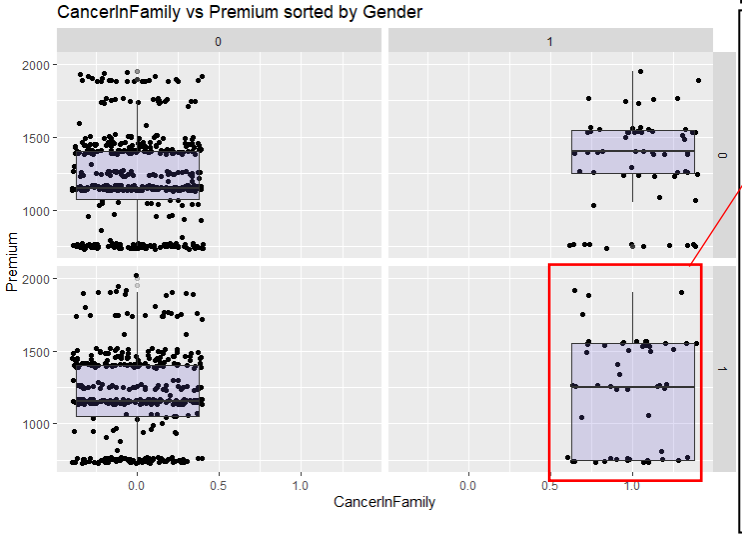
Comparison of NumMajorSurgeries also show a positive Correlation, this can be because of increasing costs of healthcare associated with the surgeries and inpatient treatment, and recurrent hospital surgeries can be costly for insurance providers. This results in Insurance providers increasing premiums costs as a deterrence to individuals who misuse the healthcare benefits provided by the insurance providers, while helping to cover costs and maintain profitability for the Insurance companies themselves. *(Fay B., 2021)*



From the plots, we can see that premiums are higher with individuals possessing presence of Chronic Diseases other than Diabetes or High Blood Pressure. Insurers charge a higher premium for individuals with pre-existing illnesses at the time of policy purchase, due to a higher likelihood of policy claims. The additional cost is borne by the patient with chronic illness to allow the insurer to hedge against risk of providing coverage for such diseases. *(Life, n.d.)* This data is also similar to age vs gender vs premium, where males pay lesser for premiums as shown by the lower first quartile for those without Chronic Disease.



There is a significant difference between the median displayed for the premium cost in both males and females for presence and absence of organ transplant preformed. As the dataset is a historical data, this indicates that patients with organ transplants result in extremely high insurance premiums due to cost-ineffectiveness in transplant procedures and high overhead costs. *(Evans R.W., 1993)* From the data, we can see that there is no Upper Quartile in Individuals with transplants, and this could show that presence of transplants plays a key role in Premium costs.



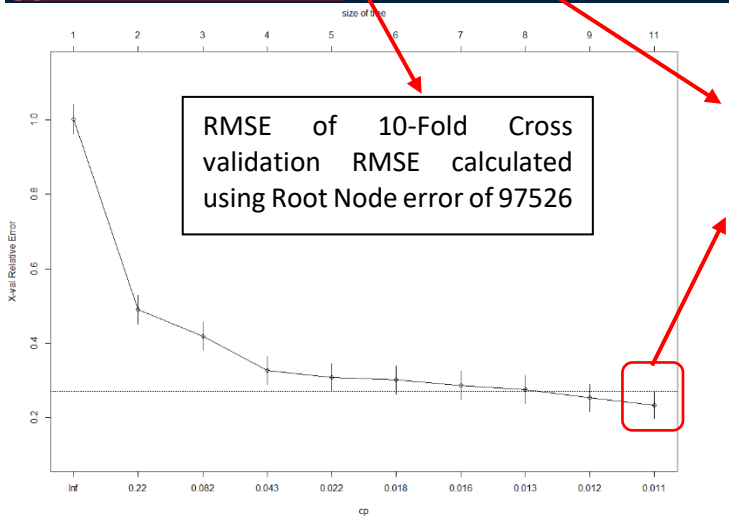
The presence of cancer in family leads to a higher cost of Insurance Premiums. CancerInFamily shows a huge premium cost range in males who displayed family history of cancer, leading to a wider premium range. Male range deviation could be attributed to males having 50% chance of cancer development, while females have a 33% chance. There is also a global 10% chance of inheritance of cancer genes and insurers increase premiums to offset the cost and risk of cancer occurrence in paying for medical bills as the overhead costs are to be borne by the insurance company. *(Erin O.D. 2017)*

Other variables such as BMI, Height, Diabetes, HighBloodPressure and Allergy **(Appendix B)** did not display any differences between determining Insurance Premium Costs between males and females and thus from a visualization standpoint, we might assume that they would not play a role in determining costs. However, this is a Non-Machine Learning model and variables would have to be investigated further to obtain analysis on their importance in determining the residual premium factor. The variables Age, CancerInFamily, Transplant and ChronicDisease would be assumed to play a pivotal role in model determination and variable importance due to the presence of outliers and differences in data values in the scatterplot and boxplot. **(Refer to Q5 on variable differences)**

Answer to Q4:

Q4a) Calculate 10-fold cross validation RMSE and number of splits in the 1SE Optimal CART

```
> printcp(premiumcart2, digits = 3)

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

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

Root node error: 96355891/988 = 97526

n= 988

        CP nsplit rel error xerror   xstd
1  0.5110      0     1.000  1.002 0.0396
2  0.0907      1     0.489  0.491 0.0392
3  0.0746      2     0.398  0.419 0.0373
4  0.0253      3     0.324  0.326 0.0372
5  0.0185      4     0.298  0.308 0.0376
6  0.0180      5     0.280  0.301 0.0384
7  0.0135      6     0.262  0.287 0.0383
8  0.0134      7     0.248  0.275 0.0373
9  0.0109      8     0.235  0.253 0.0371
10 0.0105     10     0.213  0.234 0.0359
> cp1.rmse <- sqrt(97526*0.234)
> cp1.rmse #151.0665
[1] 151.0665
```



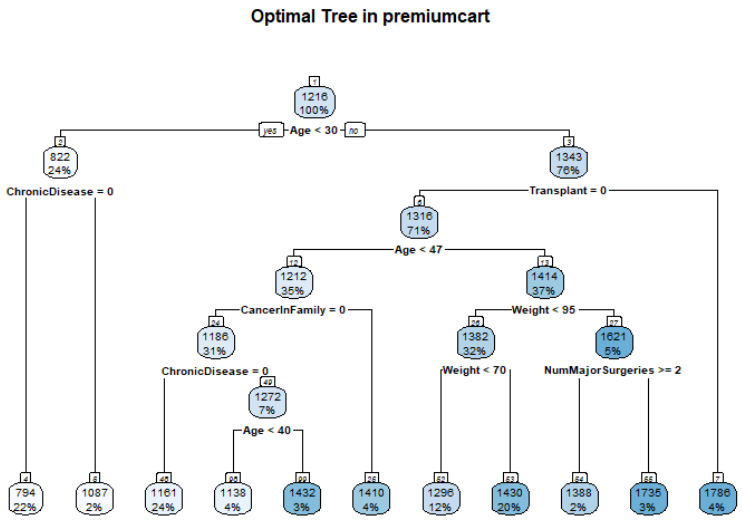
RMSE of 10-Fold Cross validation RMSE calculated using Root Node error of 97526

Optimal CART is derived from pruning of overplotted CART. The **10-Fold Cross Validation RMSE is 151.0665** and **Number of splits in the 1SE optimal CART is 10**.

Printcp() is used to obtain 10 fold cv error **(Image above)** and number of splits obtained using plotcp() **(Left image).**

Q4b) Identify the key predictors of premium.

```
> rpart.plot(premiumcart2, nn=T, main = "Optimal Tree in premiumcart") #Shows pruned decision trees
```



Variables (Key Predictors) used in predicting Health Insurance Premium in the non-train-test split CART includes (i) Age, (ii) Transplant, (iii) ChronicDisease, (iv) CancerInFamily, (v) Weight, (vi) NumMajorSurgeries. **(Refer to Appendix C for Error and Pruning sequence)**

Q4c) Is BMI or Gender important in determining premium?

Both BMI and Gender not important in determining premium, as they are not in the variables used for the split. **(Refer to image in Q4b above)**

Q4d) Evaluate and compare the predictive accuracy of the two techniques on a 70-30 train-test split. Present testset RMSE results in a table.

Using a train-test split ratio of 70%-30%, CART was compared against Linear Regression as an analytics and visualisation model. CART was pruned to the optimal tree using the 10 Fold CV Error **(Appendix F)**, and a step() function was used to obtain backwards stepwise regression for the Linear Regression mode for Continuous Y variable (Premium) to obtain the optimal trainset model. The 30% testset split was then parsed into both trained models, and using the predicted results for Premium in both models compared to testset$Premium, the RMSE was obtained as follows:

```
> testset.error = testset$Premium - cart.predict
> # Testset Errors
> RMSE.cart.test = sqrt(mean(testset.error^2))
> RMSE.cart.test # CART RMSE is 167.5624
[1] 167.5624
> testset$error <- testset$Premium - lr.predict
> # Testset Errors
> LR.test.RMSE <- sqrt(mean(testset$error^2))
> LR.test.RMSE # Linear Regression RMSE is 200.7566
[1] 200.7566
```

Error between testset premium variable and predicted results for respective model

Parsing Error Value into sqrt() function to find RMSE of each model

Data was then parsed into a data frame form to present it as a table as shown below:

CART RMSE	LinearReg RMSE
167.5624	200.7566

```
> tb.rmse= data.frame(CART.RMSE = rep(c(RMSE.cart.test)),
+                     LinearReg.RMSE = rep(c(LR.test.RMSE))
+                     )
> tb.rmse
  CART.RMSE LinearReg.RMSE
1  167.5624        200.7566
```

Parsing RMSE from each model into dataframe/table format

This shows a better and higher predictive accuracy on CART than Linear Regression due to the lower Root mean Squared Error (RMSE) in the train-test split on the predicted testset results compared to the testset$Premium variable. Mean Absolute Percentage Error (MAPE) was also used as a predictive accuracy comparative technique; CART MAPE is **0.0918** while Linear Regression MAPE is **0.122**. Both RMSE and MAPE statistical analysis shows CART has a better predictive accuracy.

Answer to Q5:

Q5) Explain the limitations of your analysis. [Max 1 page]

Train-test split ratio was kept at 70%-30% throughout the Machine Learning model determination and plotting. One limitation of this analysis was the inability to use different train-test split ratios to obtain the optimal model for the data given. Data overfitting could occur if excessive data points was not used in the analysis. This could result in too little data points used for the test set model and predictive analysis. However, if insufficient data was used to train the model data underfit would be present and the predicted results would be inaccurate as this impedes the model's generation ability in both Linear Regression and CART. Usage of one subset of model split for training can make the analysis and model biased too. *(Draelos R., 2019)* In this analysis, sufficient data points (n=988) were given as a sample size, and hence an alternative would be to test models in intervals of 10% changes in train and test ratio splits (E.g 50%-50%, 60%-40%, 80%-20%) and obtain the RMSE of the predicted results and select the optimal model using RMSE and MAPE as a benchmark.

Another limitation of the analysis is that Linear Regression assumes that variables would be independent of each other. As seen from the correlation plot, Weight and Height are highly positive correlated to BMI, and the BMI columns could be removed from future analysis as the multi-collinearity could result in additional variables affecting the Residual Y (Continuous Y) that is being investigated, which is Insurance Premium in this scenario. Similarly for Age variable, it is highly correlated to Premium. The usage of Linear Regression assumes that the distribution of the dependent variable (Insurance Premium) is normal, and its variance should be constant for all values of the independent variable. The relationship between the dependent variable and each independent variable should be linear and all observations be independent. *(IBM, n.d.)* When variables do not have a linear relationship with Premium, this would not be able to provide a good prediction model and the variables would be considered statistically insignificant with limited insights obtainable. However, this was not observed in the barplots and correlation plots **(Refer to Q3 and Appendix B)**, and could affect the employability of the Linear Regression as a statistical method for prediction of testset results as the trainset model creates biasness towards variables that it is highly correlated to (E.g.Age).

As seen from Q4b) and Q4d), the lack of a split compared to a split results in a difference in variable significance or variables that are used to determine Insurance Premium. **(Refer to Appendix D for summary of variables in each model)** The greatest difference is the presence of CancerInFamily that are present in both Linear Regression and CART without Split, but not the CART model with train-test Split. CART is sensitive to changes in data and can underfit the trained model if variables/classes are imbalanced during the split. *(Yadav A., 2019)* This is evident as there are only 116 individuals with history of Cancer while there are 872 individuals with No history of Cancer in Family. This small sample of patients could result in an uneven split for train and test set results, hence CART with Split eliminated the CancerInFamily variable in the optimal tree. A mitigation to this would be to use a more balanced dataset or split the current dataset with a balanced variable class and variable composition for CancerInFamily. In addition, we were unable to obtain the R-square of the trainset model, which does not allow us to compare variable-model fit for the trainset data through a statistical analysis.

Lastly, correlation does not equal causality. There could be other factors involved in the analysis of Insurance Premiums, which are underlying factors. For example, smoking habits, BMI, geographical location and occupation are important factors that are considered in Insurance Healthcare Premium *(IffcoTokio., n.d.)*, but the variables are either not used for analysis in this report or not pointed out by the machine Learning Model (E.g. BMI). This requires secondary research and possible additional data to be collected to analyse the Insurance Premium data variable to prevent hindsight bias. Statistical analysis in a business context can only be used to a limited extent. Given the ratio of various variables involved such as CancerInFamily, this can lead to selection bias from the raw data and difficulty in isolating the optimal effects of variables onto Insurance Premiums. If the insurance company decides to focus on coverage of illnesses concerning the variables in place with key assumptions (E.g. clients do not smoke), it would be a safer scenario to use the models to determine insurance Premiums.

Answer to Q6:

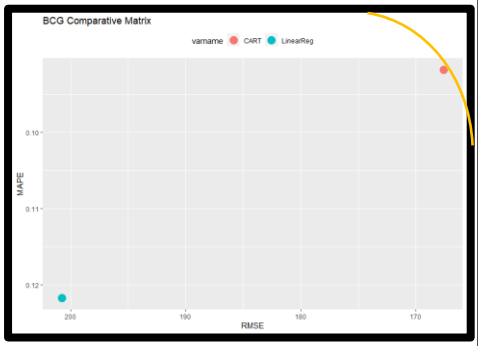
Q6) Is CART successful in this application? Explain. [Max 1 page]

With a key focus on Machine Learning Models and the statistical analysis in this report, CART is successful in the generation of insurance premiums payable when compared to other predictive Machine Learning models such as Linear Regression due to its lower RMSE and MAPE values. Using the RMSE and MAPE, we can see that the values are as follows:

Predictive Model (70-30 Split)	CART	Linear Regression
RMSE	167.562	200.757
MAPE *(Appendix E for calculations)*	0.0918	0.122

RMSE shows the errors of the residuals predicted from the testset data, displaying the concentration of data around the line of best fit and RMSE penalising data when there are exceptionally large errors in the prediction *(JJ., 2016)*. Due to the elimination of variables and the sensitivity of CART, they removed 8 out of 12 variables and retained 4 variables: (i) Age, (ii) ChronicDisease, (iii) Transplant, (iv) Weight in determination of the Premium variable. Linear Regression however, showed higher RMSE despite obtaining the optimal predictive model after step() was used. This higher RMSE could be attributed to multi-collinear variables in the model **(Refer to Q3)** and the lack of a linear relationship between residuals and variables **(Appendix B)**. This results in the inability of the Linear Regression model in being as sensitive as CART in obtaining the statistically significant variables **(Refer to Q5 on Limitations of Linear Reg)**, hence it included additional variables due to overfitting in some variables over the other causing higher RMSE while CART with splits provided a lower RMSE.

MAPE shows the favourability of the predictive model and possibility of model overfit/underfit. A higher MAPE results in a lower feasibility of the model as the model is overfit and this could result in a biased model with lower predictive accuracy if it was selected due to higher error present *(Glen S., n.d.)*. A heavier penalty is placed on negative errors than on positive errors as the percentage error (Actual value – Predicted value) cannot exceed 100% for forecasts that are too low, resulting in biases that will systematically select a model that has low forecasts. Hence, when the actual value is lower than the predicted value, this results in an elevated MAPE compared to when the predicted value is lower than actual value, even though both values are the same. However, as both models are using the same trainsets and testsets, this sets a baseline for comparison for MAPE, hence we can see that CART excels in using MAPE as compared to Linear Regression, as MAPE using predicted value is lower.



From the Comparative matrix, we can see that CART is successful in compared to Linear Regression as excels on the efficient frontier (Orange Arc), leading to a more accurate and feasible outcome. CART is also more sensitive to variable and data changes with lower statistical errors, while it utilizes 4 variables in predicting premium as compared to the non-pruned model **(Refer to Q4a)**. Linear Regression results in usage of 5 variables (CancerInFamily) and the presence of cancer can lead to a higher medical premium.

From a secondary perspective, CART may not be successful, as correlation does not equal to correlation. **(Refer to Q5)** There could be other variables in place that disproves CART in segregating variables in predicting premium. As a standalone model, the non-split CART model shows usage in CancerInFamily variable, which is congruent to the variables from Linear Regression but different when compared to the CART model with split. **(Appendix D, F)**

Family history or presence of cancer is also known to result in higher medical premium *(KFF., 2009)*. If the CART model is selected, they can remove the medical bias faced by clients with family history of cancer who must spend additional money. However, using a risk analysis from the insurance company's perspective, charging a higher premium for cancer patients will help them maintain profitability. Hence, there is a trade off in usage of implicit analysis (statistical analysis) versus explicit analysis (secondary research) and a comprehensive review has to be done on whether it is necessary to exclude the CancerInFamily variable or risk losing potential clients.

References

CDC. (9 May 2014). Growth Chart Training: Calculating BMI using the Metric System. Accessed 8 November 2021. Retrieved from: https://www.cdc.gov/nccdphp/dnpao/growthcharts/training/bmiage/page5_1.html#:~:text=With%20the%20metric%20system%2C%20the,by%2010%2C000%2C%20can%20be%20used.

Fontinelle A. (11 June 2021). Gender and Insurance Costs. Accessed 10 November 2021. Retrieved from: https://www.investopedia.com/gender-and-insurance-costs-5114126

Fay B. (12 October 2021). Hospital and Surgery Costs. Accessed 10 November 2021. Retrieved from: https://www.debt.org/medical/hospital-surgery-costs/

Life. (n.d). Have a chronic Disease- Know these health insurance facts. Accessed 10 November 2021. Retrieved from: https://life.futuregenerali.in/life-insurance-made-simple/life-insurance/5-health-insurance-facts-to-know-if-you-have-a-chronic-disease

Evans R.W. (October 1993) Organ transplantation costs, insurance coverage and reimbursement. Accessed 10 November 2021. Retrieved from: https://pubmed.ncbi.nlm.nih.gov/2103158/

Erin O.D. (March 2017) Why is Cancer more common in men? Accessed 10 November 2021. Retrieved from: https://www.harvardmagazine.com/2017/03/why-is-cancer-more-common-in-men#:~:text=Oncologists%20know%20that%20men%20are,with%20one%20in%20three%20women.

Draelos R. (15 September 2019). Best use of Train/val/Test Splits, with tips for Medical Data. Accessed 10 November 2021. Retrieved from: https://glassboxmedicine.com/2019/09/15/best-use-of-train-val-test-splits-with-tips-for-medical-data/

IBM. (n.d.) Linear regression. Accessed 11 November 2021. Retrieved from: https://www.ibm.com/topics/linear-regression

Yadav A. (11 January 2019). Decision Trees. Accessed 12 November 2021. Retrieved from: https://towardsdatascience.com/decision-trees-d07e0f420175

IffcoTokio. (n.d.) 10 factors that affect your health insurance premium costs. Accessed 12 November 2021. Retrieved from: https://www.iffcotokio.co.in/health-insurance/10-factors-that-affect-your-health-insurance-premium-costs

JJ. (2016). MAE and RMSE – Which Metric is Better? Accessed 12 November 2021. Retrieved from: https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

Glen S. (n.d.). Mean Absolute Percentage Error (MAPE). Accessed 12 November 2021. Retrieved from: https://www.statisticshowto.com/mean-absolute-percentage-error-mape/

KFF. (February 2009). Spending to survive: Cancer patients Confront Holes in health Insurance System. Accessed 12 November 2021. Retrieved from: https://www.kff.org/wp-content/uploads/2013/01/7851.pdf

Appendix

Appendix A

```
> premiumfactor= premium #create dataframe for factored variables
>
> premiumfactor$Diabetes = factor(premiumfactor$Diabetes)
> class(premiumfactor$Diabetes)
[1] "factor"
>
> premiumfactor$HighBloodPressure = factor(premiumfactor$HighBloodPressure)
> class(premiumfactor$HighBloodPressure)
[1] "factor"
>
> premiumfactor$Transplant = factor(premiumfactor$Transplant)
> class(premiumfactor$Transplant)
[1] "factor"
>
> premiumfactor$ChronicDisease = factor(premiumfactor$ChronicDisease)
> class(premiumfactor$ChronicDisease)
[1] "factor"
>
> premiumfactor$Allergy = factor(premiumfactor$Allergy)
> class(premiumfactor$Allergy)
[1] "factor"
>
> premiumfactor$CancerInFamily = factor(premiumfactor$CancerInFamily)
> class(premiumfactor$CancerInFamily)
[1] "factor"
>
> premiumfactor$Gender = factor(premiumfactor$Gender)
> class(premiumfactor$Gender)
[1] "factor"
```

Creating new dataframe called 'premiumfactor' from 'premium' dataframe

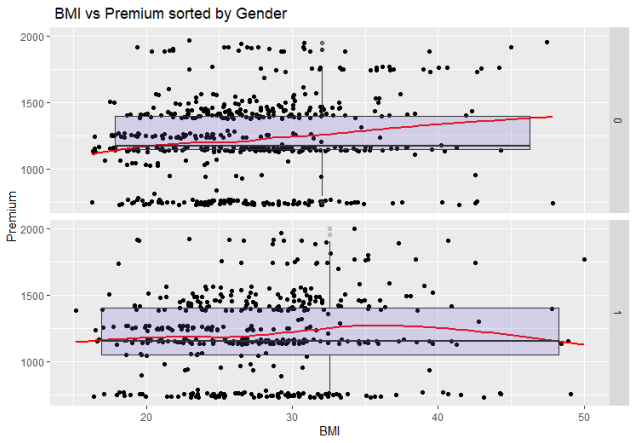
Factor() function to convert integer data in variable into binary variables

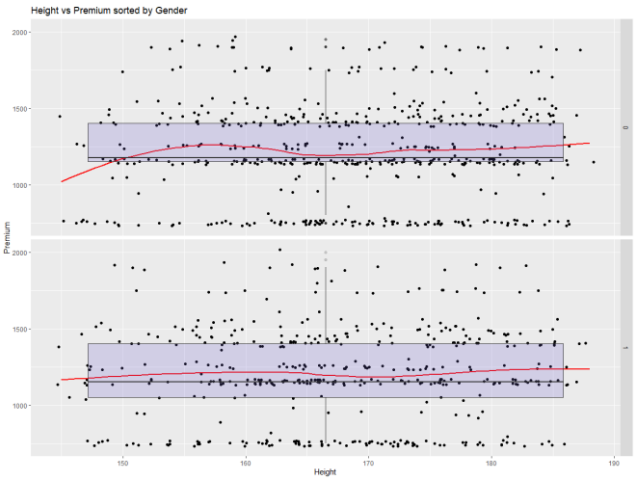
Parsing factorised data into column in 'premiumfactor' data frame

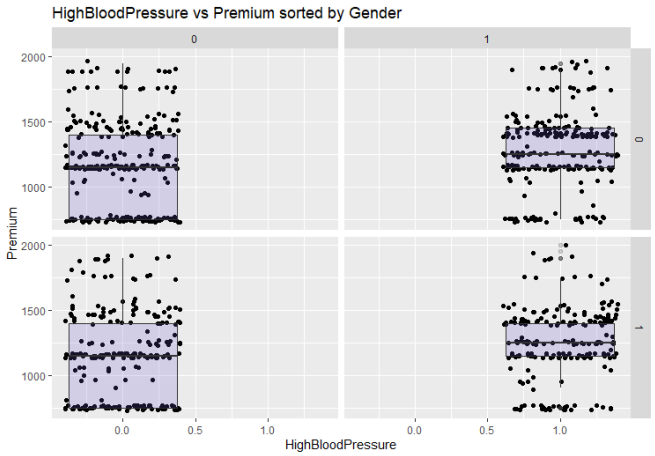
Class() function used to check variable data type in column

Ensuring all 7 variables are factorised

Appendix B






HighBloodPressure vs Premium sorted by Gender
0
1
Premium
2000
1500
1000
0.0
0.5
1.0
HighBloodPressure


Diabetes vs Premium sorted by Gender
0
1
Premium
2000
1500
1000
0.0
0.5
1.0
Diabetes


Allergy vs Premium sorted by Gender
0
1
Premium
2000
1500
1000
0.0
0.5
1.0
Allergy

Appendix C: Error and Pruning sequence

```
> print(premiumcart2) #Shows error and pruning sequence
n= 988

node), split, n, deviance, yval
      * denotes terminal node

 1) root 988 96355890.00 1216.1940
   2) Age< 29.5 240 12097960.00  822.0833
     4) ChronicDisease< 0.5 217  8507212.00  794.0092 *
     5) ChronicDisease>=0.5 23  1806087.00 1086.9570 *
   3) Age>=29.5 748 35019560.00 1342.6470
     6) Transplant< 0.5 706 23972680.00 1316.2890
      12) Age< 46.5 341  7484399.00 1211.8770
        24) CancerInFamily< 0.5 302  4585472.00 1186.2580
          48) ChronicDisease< 0.5 234  1807489.00 1161.3250 *
          49) ChronicDisease>=0.5 68  2131912.00 1272.0590
            98) Age< 39.5 37    77027.03 1137.8380 *
            99) Age>=39.5 31   592741.90 1432.2580 *
        25) CancerInFamily>=0.5 39  1165897.00 1410.2560 *
      13) Age>=46.5 365  9297630.00 1413.8360
        26) Weight< 94.5 316  5223544.00 1381.6460
          52) Weight< 69.5 114   993596.50 1296.4910 *
          53) Weight>=69.5 202  2936782.00 1429.7030 *
        27) Weight>=94.5 49  1635000.00 1621.4290
          54) NumMajorSurgeries>=1.5 16    37500.00 1387.5000 *
          55) NumMajorSurgeries< 1.5 33   297424.20 1734.8480 *
     7) Transplant>=0.5 42  2311429.00 1785.7140 *
>
```

Appendix D: Summary of variables used to determine trainset for each model

Linear Regression	CART (No Split)	CART (70-30 Split)
• Age • Transplant • Chronic Disease • Weight • CancerInFamily	• Age • Transplant • Chronic Disease • Weight • NumMajorSurgeries • CancerInFamily	• Age • Chronic Disease • Transplant • Weight

Appendix E: MAPE for CART and Linear Regression using predicted results

```
> CART.MAPE= MAPE(testset$Premium, cart.predict)
> CART.MAPE #CART MAPE is 0.09178514
[1] 0.09178514
> LR.MAPE= MAPE(testset$Premium, lr.predict)
> LR.MAPE #Linear Reg MAPE is 0.1217045
[1] 0.1217045
```

Appendix F: Linear Regression (Stepwise Regression) output

```
> set.seed(2004)
> premiumLR = lm(Premium ~., data = trainset)
> #backward stepwise regression to find optimal LR and variables
> premiumLR = step(premiumLR, direction='backward', scope=formula(premiumLR), trace=0)
> summary(premiumLR)

Call:
lm(formula = Premium ~ Age + Transplant + ChronicDisease + Weight +
    Allergy + CancerInFamily + NumMajorSurgeries, data = trainset)

Residuals:
    Min      1Q  Median      3Q     Max
-688.05 -109.73  -14.41   99.02 1214.93

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       218.6361    43.7608   4.996 7.44e-07 ***
Age                16.2384     0.5583  29.086  < 2e-16 ***
Transplant        427.9902    31.2325  13.703  < 2e-16 ***
ChronicDisease    147.0649    18.3317   8.022 4.52e-15 ***
Weight              3.4555     0.4753   7.269 9.90e-13 ***
Allergy            23.9575    16.9926   1.410   0.1590
CancerInFamily    111.4615    22.8046   4.888 1.27e-06 ***
NumMajorSurgeries -23.6191    10.6287  -2.222   0.0266 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 181.9 on 683 degrees of freedom
Multiple R-squared:  0.6665,    Adjusted R-squared:  0.6631
F-statistic:   195 on 7 and 683 DF,  p-value: < 2.2e-16
```

Variables identified to be Important in determining Premium are as follows: Age, Transplant, Chronic Disease, Weight, CancerInFamily

Allergy and NumMajorSurgeries are not included as they do not have sufficient statistical significance (P<0.05) in analysis

Appendix F: Pruned Trees and 10 Fold CV Error in CART with train-test Split (Q4d)