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

Class	2
Full Name	Leow Ken Hing Bryan
Matriculation Number	U2021729K

** : 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:

1. CDC definition: "BMI is a person's weight in kilograms divided by the square of height in meters."[1]
2. First, we must determine whether the data is in imperial units (feet & inches + pounds) or metric units (metres/centimetres + kilograms).*
    a. We do this by taking the average height in the dataset.
    b. The tallest person in the world was 2.72m / 272cm / 8.92ft.[2]
    c. We can very reasonably assume that the average will never exceed those bounds.
    d. We can also reasonably assume that if the dataset uses feet & inches for height, it will also use pounds for weight and vice versa.
    e. We can hence write if-else statements that handle unit conversions based on the inferred units that the data is in.
3. Finally, since BMI assessments are done in 1 decimal place, we round our answers to 1dp.

```
################################################  Question 1  ################################################
########################  Create the BMI variable based on CDC definition. Show your code.  ########################

       if (2.72 < mean(dt$Height) & (mean(dt$Height) < 8.92)) { # in imperial units
    dt$BMI = (dt$Weight/2.20462) * (dt$Height*30.48/100)^2          # convert to metric for calculation
} else if (8.93 < mean(dt$Height) & (mean(dt$Height) < 250)) {  # in metric units, cm
    dt$BMI = dt$Weight / (dt$Height/100)^2
} else if (250 < mean(dt$Height)) {                             # in metric units, m
    dt$BMI = dt$Weight / (dt$Height)^2
}
dt$BMI <- round(dt$BMI, digits = 1)
```

Figure 1.1: Required R code for Question 1.

4. Our sanity check using summary() confirms that this logic works for this dataset.

```
> summary(dt$BMI)
   Min. 1st Qu.  Median    Mean 3rd Qu.
  15.20   23.38   27.15   27.45   30.80
   Max.
  50.00
```

Figure 1.2: A summary of the newly-added BMI column, showing numbers that are sensible. The min[3] and max[4] numbers are not impossible.

** Even though the data dictionary states it is in cm and kg, it is still good practice nonetheless to perform such data validation before making data modifications, e.g. if the company is located in multiple countries and their units of measurement depends on locale.*

[1] https://www.cdc.gov/healthyweight/assessing/index.html
[2] https://en.wikipedia.org/wiki/Robert_Wadlow
[3] https://www.bbc.com/news/uk-england-leeds-44488822
[4] https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmi_tbl2.htm

Answer to Q2:

In terms of being "*necessary*", and in the specific context of this dataset, no.

The full list of relevant variables that can be converted into a factor datatype are listed in Figure 2.4. All these relevant variables are Booleans, i.e. they can only ever hold a value of 0 or 1. As a result, converting them to factor or not will not matter when building models.

```
> # LIN REG Results - coded as integer
> sqrt(c(crossprod(m_linreg$residuals)) / length(m_linreg$residuals))
[1] 10.00733
> class(dt$Diabetes)
[1] "integer"
```

```
> # LIN REG Results - coded as integer
> sqrt(c(crossprod(m_linreg$residuals)) / length(m_linreg$residuals))
[1] 10.00733
> class(dt$Diabetes)
[1] "factor"
```

Figure 2.1: A linear regression model of Height on everything else. Whether the factor variables are dummy encoded or not does not matter since in the former, one of the (two) possible values will be dropped, leading to the same interpretation and meaning of the coefficients.

```
> # LOG REG RESULTS - coded as integer
> table("predicted" = pred_m_logreg, "actual" = dt$Diabetes, deparse.level = 2)
         actual
predicted   0   1
        0 464 258
        1 110 156
> class(dt$Diabetes)
[1] "integer"
```

```
> # LOG REG RESULTS - coded as integer
> table("predicted" = pred_m_logreg, "actual" = dt$Diabetes, deparse.level = 2)
         actual
predicted   0   1
        0 464 258
        1 110 156
> class(dt$Diabetes)
[1] "factor"
```

Figure 2.1: A logistic regression model of Diabetes on everything else, showing (a snippet of the) identical results obtained whether the relevant variables are factor or categorical.

```
> # CART RESULTS - coded as integer, method = 'anova'
> table("predicted" = pred_m_cart, "actual" = dt$Diabetes, deparse.level = 2)
                   actual
predicted             0   1
  0                  60   0
  0.302405498281787 203  88
  0.470588235294118 108  96
  0.531177829099307 203 230
> class(dt$Diabetes)
[1] "integer"
```

```
> # CART RESULTS - coded as integer, method = 'class'
> table("predicted" = pred_m_cart, "actual" = dt$Diabetes, deparse.level = 2)
         actual
predicted   0   1
        0 470 237
        1 104 177
> class(dt$Diabetes)
[1] "integer"
```

```
> # CART RESULTS - coded as integer
> table("predicted" = pred_m_cart, "actual" = dt$Diabetes, deparse.level = 2)
         actual
predicted   0   1
        0 470 237
        1 104 177
> class(dt$Diabetes)
[1] "factor"
```

Figure 2.3: A CART model of Diabetes on everything else, showing (a snippet of the) identical results obtained whether the relevant variables are factor or categorical. The first uses method = 'anova' while the second uses method = 'class'. The latter gives identical results, while applying a threshold of 0.5 on the former gives the same confusion matrix.

That being said, converting these relevant variables to factor would give us a range of benefits.

```
#### convert to factor ####
dt$Diabetes             <- as.factor(dt$Diabetes)
dt$HighBloodPresure     <- as.factor(dt$HighBloodPressure)
dt$Transplant           <- as.factor(dt$Transplant)
dt$ChronicDisease       <- as.factor(dt$ChronicDisease)
dt$Allergy              <- as.factor(dt$Allergy)
dt$CancerInFamily       <- as.factor(dt$CancerInFamily)
dt$Gender               <- as.factor(dt$Gender)
```

Figure 2.4: Converting the relevant variables to factor

The most valuable benefit is the summary() function in doing our exploratory analysis. With the variable encoded as integer, this function prints the limits, median, mean and quartiles, which are not useful at all (left). On the other hand, converting them to factor would allow us to print the exact amount of data in each category, which is more relevant:

```
> class(dt$Diabetes)
[1] "integer"
> summary(dt$Diabetes)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   0.419   1.000   1.000
```

```
> class(dt$Diabetes)
[1] "factor"
> summary(dt$Diabetes)
  0   1
574 414
```

Figure 2.5: Difference in results shown for summary(), showing the benefits of converting

Converting may also allow R to provide specific insights. For example, running the logistic regression after factor conversion brings up a warning which hints of "insufficient or replicated data".[5] This warning did not appear before the conversion.

```
> prob <- predict(m_logreg, dt, type = 'response')
Warning message:
In predict.lm(object, newdata, se.fit, scale = 1, type = if (type ==  :
  prediction from a rank-deficient fit may be misleading
> class(dt$Diabetes)
[1] "factor"
```

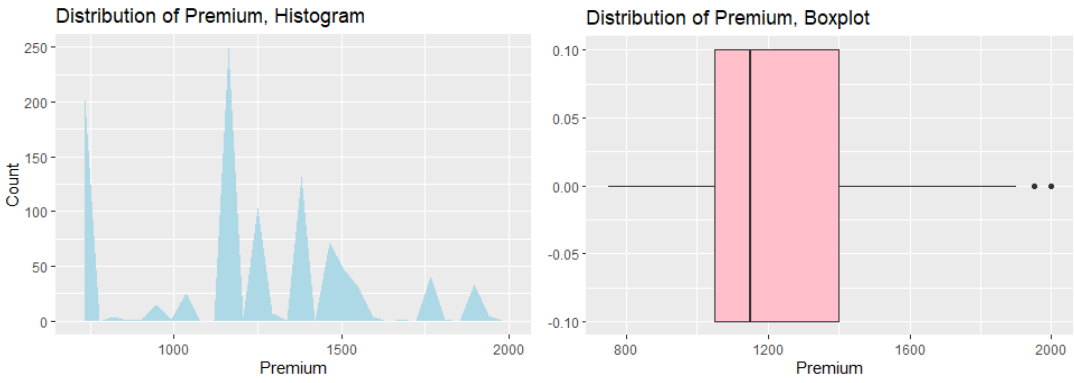
Figure 2.6: Warning message for a "rank-deficient fit" for logistic regression after factor conversion, showing the benefits of converting.

Thus, even though conversion is not necessary, it is highly beneficial to do so and hence we still stick with the conversion.

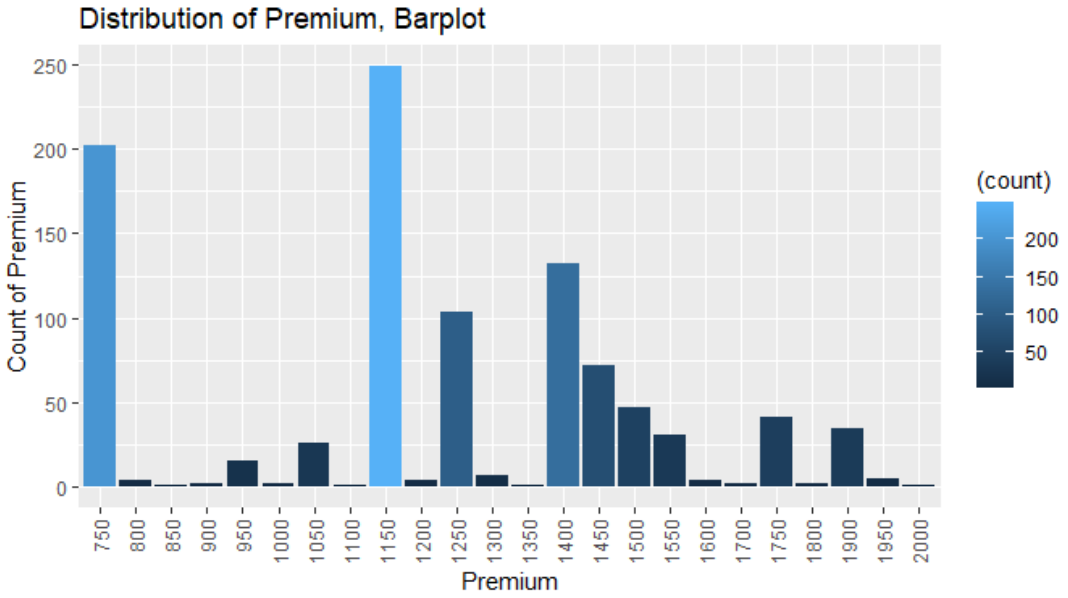
[5] https://stats.stackexchange.com/questions/35071/what-is-rank-deficiency-and-how-to-deal-with-it#35077

Answer to Q3:

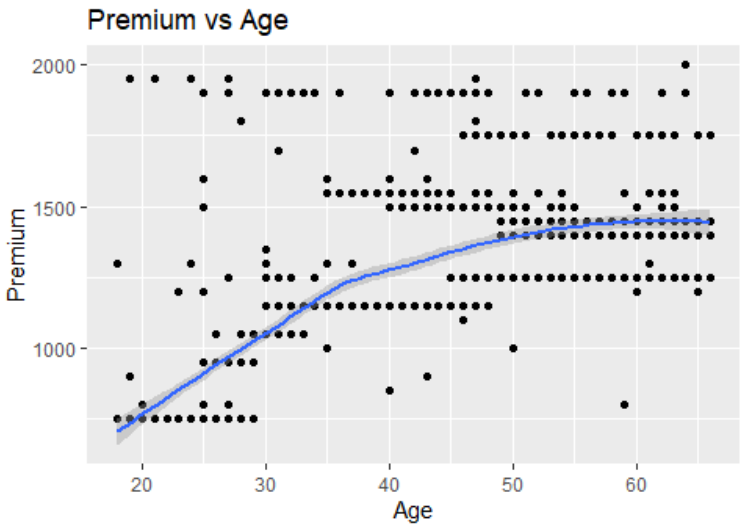
- A univariate plot of premiums show that the premiums charged actually exist in 24 discrete levels. The median charged is $1150, below the mean of $1216.



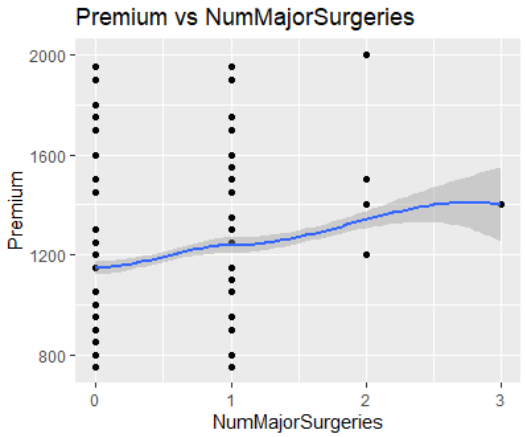
- Our data points are not evenly distributed between these 24 categories, with certain categories having many data points and some having only 1-5.



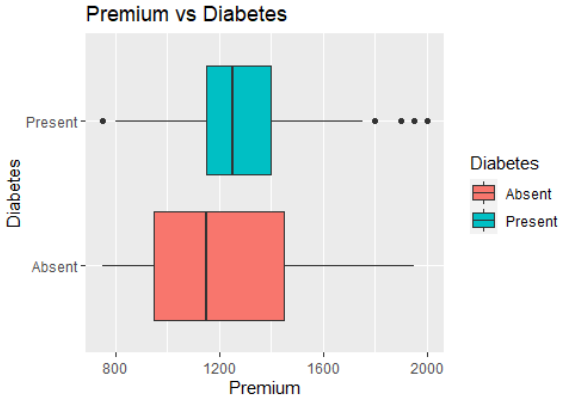
- Expected average premiums clearly increase with age, and plateau at around 55.

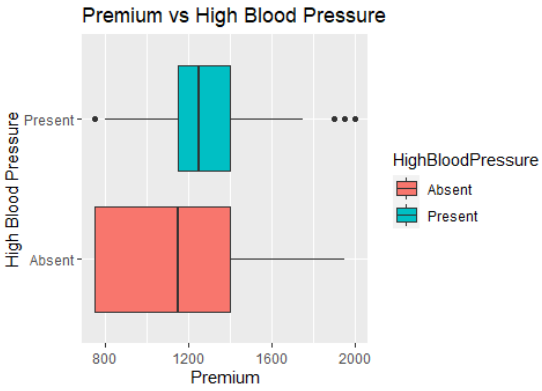


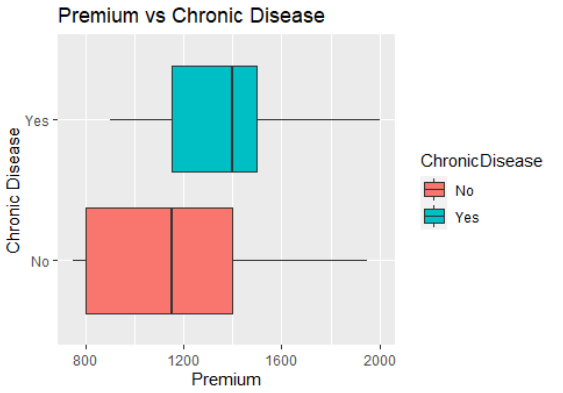
- Expected average premiums also increase with the number of major surgeries, although most have only 0-1 surgeries.

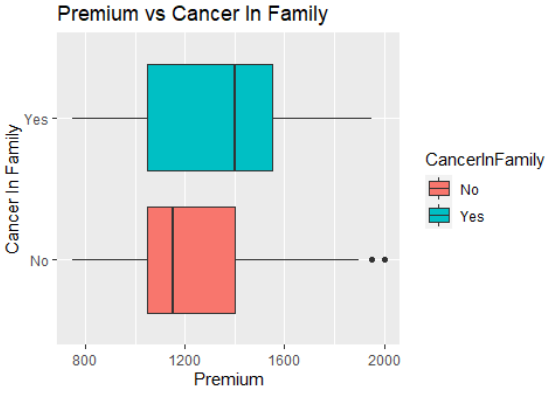


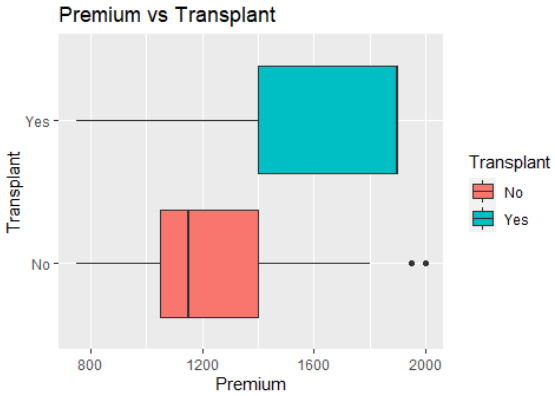
- Average premiums for those with diabetes, high blood pressure, chronic disease, cancer in family, and transplants are higher than those without, as expected.



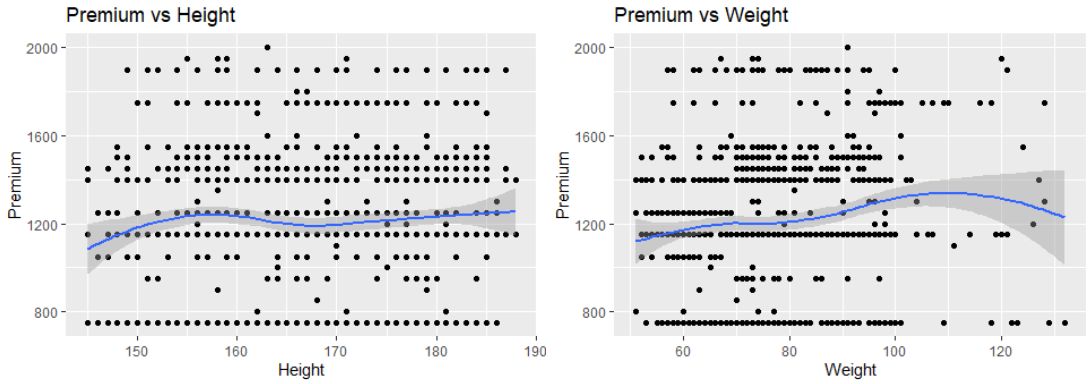




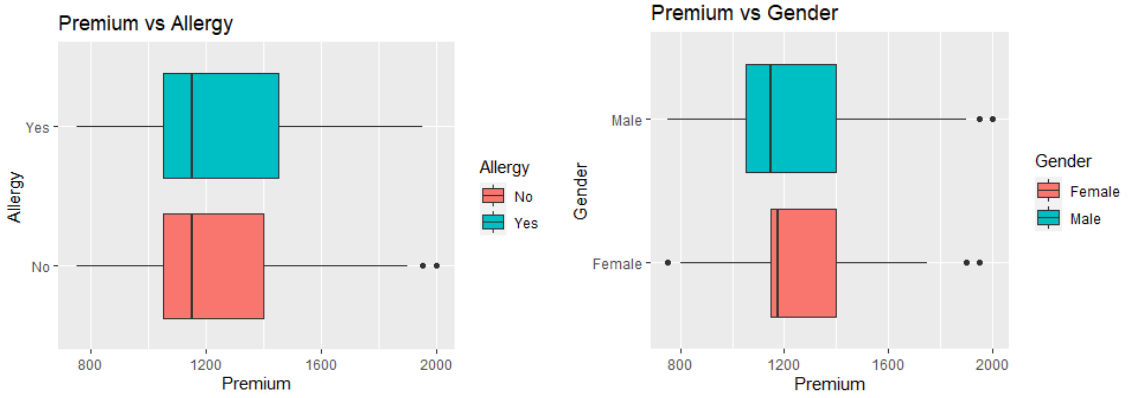




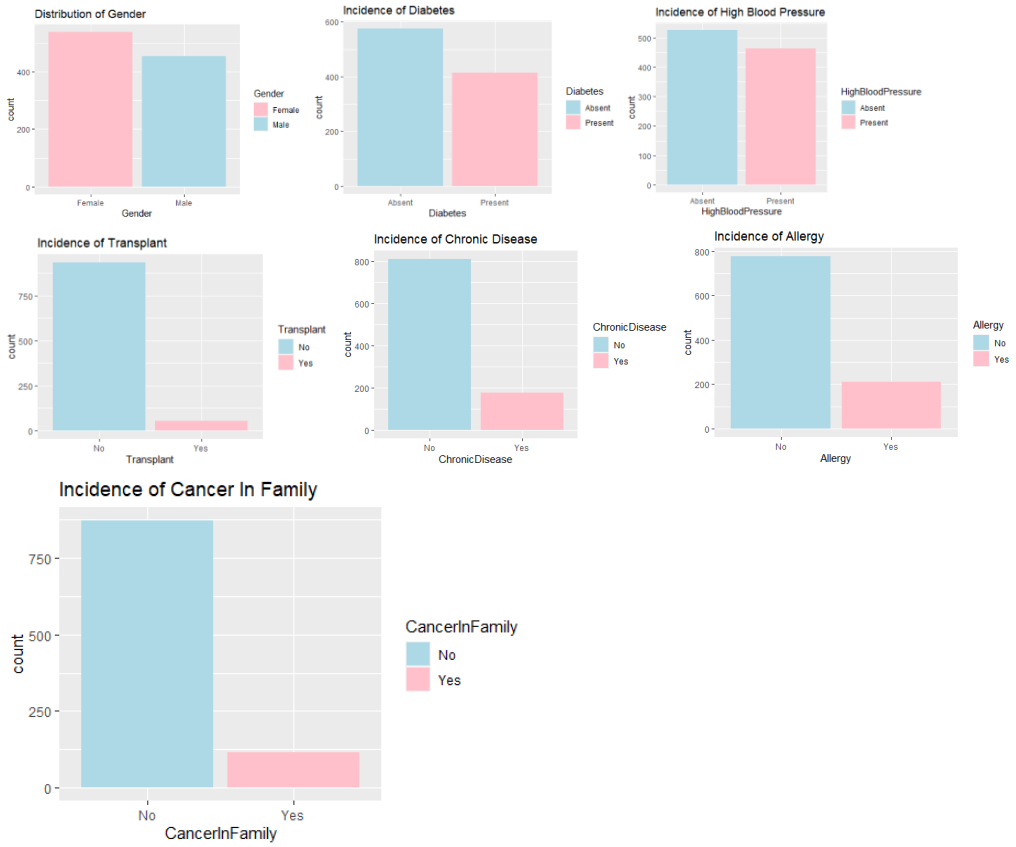
- There is a very slight increase in average expected premium with height and weight, though this is likely insignificant.



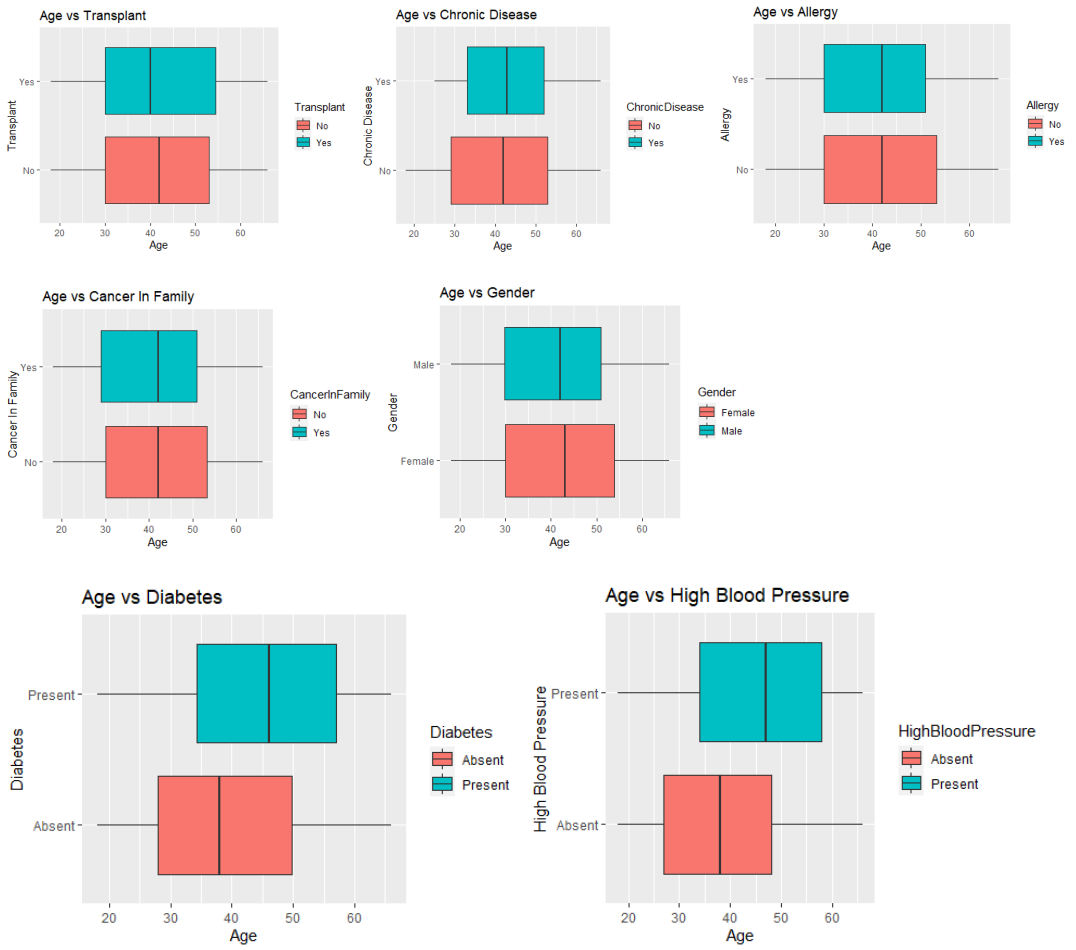
- Gender and allergy have an insignificant impact on average expected premiums.



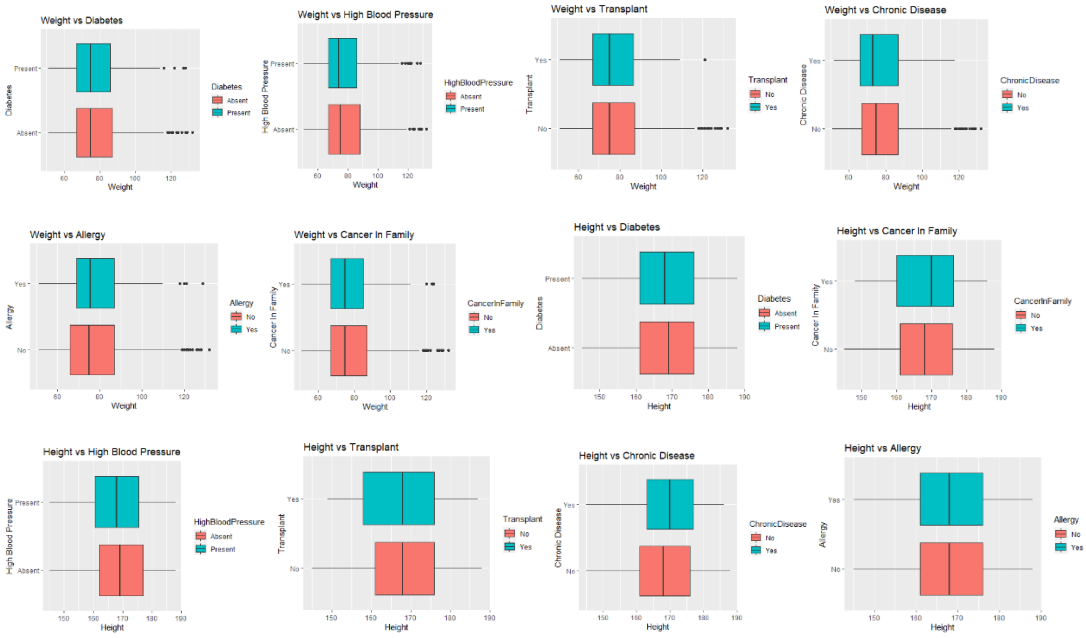
- Other than gender, diabetes, and high blood pressure, the incidence of the other categorical variables are imbalanced towards no.



- Age has a slight relation with diabetes and high blood pressure, but not the rest, possibly indicating little multicollinearity.



- Height is not related with any categorical variable. Surprisingly, neither is weight – there is no observable relation between weight and diabetes or high blood pressure.



Answer to Q4:

Question 4a:

i. **10-fold cross validation RMSE** = sqrt(0.26369 * 97526) = **160.3641.**

ii. **Number of splits: 8.**

Question 4b:

1. The first step would be to build a continuous CART model. This is done (previously in 4a) by
    a. Growing the tree to maximum in Phase 1 by using all variables,
    b. Pruning the tree in Phase 2 to reduce the number of variables.
    c. Based on CART, the most important variable by far is **age**. It is 6 times more important than the next variable, having a **transplant**, and 12 times more important than the third variable, **weight**.
    d. The rest of the variables: **number of major surgeries**, BMI, having a **chronic disease**, having **cancer in family**, having high blood pressure, and having diabetes, have 1-5% variable importance – but they can still be considered important nonetheless since this is already the 1 SE optimal CART.
2. The second step would be to apply our second technique, linear regression. This is done by
    a. Generating a model with all variables,
    b. Removing multicollinear variables using vif(),
    c. Performing variable selection via backward elimination using step(),
    d. Checking again for multicollinearity,
    e. "Standardising the coefficients"[6] for the continuous variables to allow for variable importance comparison.
    f. Based on linear regression, the categorical variables such as having a **transplant**, having a **chronic disease**, and having **cancer in family** are statistically significant. **Age** is 5 times more important than **weight**, and 10 times more important than the **number of major surgeries**.
    g. Diabetes is not very statistically significant but can still be considered important nonetheless since this is already the optimal linear regression model.
3. Reconciling these two answers, we will take the variables that appeared in both techniques as our answer.
4. Hence, the key predictors are **age, transplant, weight, number of major surgeries, chronic disease, cancer in family, and diabetes.**

Question 4c:

- **Gender is definitely unimportant** in determining premium.
    - In our continuous CART model, the variable was pruned from our 1 SE optimal CART model.
        - Even in our maximum Phase 1 tree, the variable importance of gender was only 1.17%.
    - In our linear regression model, the variable was also cut during backward elimination.
        - Even in the original linear regression model containing Gender, it was the least statistically significant (p-value of 0.82) and practically significant (coefficient of -2.6)
    - This is aligned with our expectations from our exploratory analysis in Question 3. Looking at the boxplots of Premium vs Gender, the statistical distribution of Premium is very similar, with near identical medians and Q3s.

[6] https://www.sciencedirect.com/topics/mathematics/standardized-regression-coefficient

- **<u>BMI is also unimportant</u>** in determining premium.
    - The variable has an importance of 2.3% in our pruned CART, and 6.5% in our unpruned CART.
        - However, we must remember what variable importance signifies: variables that appear higher up in the tree and/or as surrogates will be ranked as more important.
        - The latter is exactly what is happening here, as seen in Figure 4.1. BMI is closely related to (in fact, directly calculated from) Weight, causing it to be the most agreeable surrogate split whenever Weight appears as the primary split.
        - When looking at Figure 4.2, however, we observe that BMI does not appear as the primary split at any node.
        - This means its variable significance score is caused solely by its ability to serve as a surrogate for Weight.
        - If the value of Weight was missing for a data point, we would not be able to calculate BMI either. This defeats the purpose of a surrogate, which is to substitute a value when it is missing.
        - Hence, BMI cannot serve as a surrogate for Weight, and this variable importance rating is unwarranted.
        - Since BMI is neither a primary node split, nor can serve as a surrogate when actually needed, we can conclude that BMI is not important.

```
Node number 26: 316 observations,    complexity param=0.01342072
  mean=1381.646, MSE=16530.2
  left son=52 (114 obs) right son=53 (202 obs)
  Primary splits:
      Weight           < 69.5  to the left,  improve=0.24756480, (0 missing)
      BMI              < 22.85 to the left,  improve=0.07187008, (0 missing)
      Age              < 48.5  to the left,  improve=0.07098823, (0 missing)
      NumMajorSurgeries < 0.5   to the left,  improve=0.04562754, (0 missing)
      CancerInFamily    splits as  LR,        improve=0.02426507, (0 missing)
  Surrogate splits:
      BMI    < 22.35 to the left,  agree=0.785, adj=0.404, (0 split)
      Height < 147.5 to the left,  agree=0.646, adj=0.018, (0 split)
Node number 13: 365 observations,    complexity param=0.0253133
  mean=1413.836, MSE=25472.96
  left son=26 (316 obs) right son=27 (49 obs)
  Primary splits:
      Weight         < 94.5  to the left,  improve=0.26233410, (0 missing)
      BMI            < 28.8  to the left,  improve=0.17985150, (0 missing)
      Age            < 65.5  to the left,  improve=0.01585263, (0 missing)
      CancerInFamily splits as  LR,        improve=0.01407977, (0 missing)
      Height         < 148.5 to the left,  improve=0.01192298, (0 missing)
  Surrogate splits:
      BMI < 36.05 to the left,  agree=0.912, adj=0.347, (0 split)
```

Figure 4.1: BMI being proposed as the best surrogate split for Weight, accounting for BMI's reported variable importance.

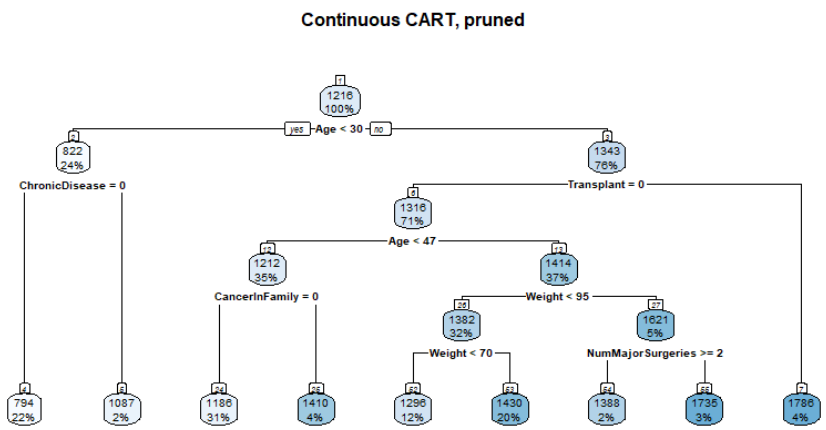


Figure 4.2: Pruned CART decision tree visualisation, showing that BMI is not actually being used as a primary split in any node.

- In linear regression, BMI was cut from the final model by both the step function measuring AIC and vif function measuring multicollinearity.
    - The former means that the inclusion of BMI added more noise to the model, making it an important variable.
    - Suppose that we can "safely ignore the multicollinearity"[7], since it is expected. Even so, in the original model, BMI was not statistically significant either (p-value = 0.53).
    - Lastly, let us compare the relative model strengths when removing Weight only and removing BMI only. Both had no more multicollinearity, but the model with only BMI removed had a higher adjusted R2 than that with only Weight removed (0.6397 vs 0.6388).
- In summary, these two analyses show that BMI is not important, since it does not provide more or better information than what Weight is already providing and cannot serve as its surrogate.

Question 4d:

	CART	Linear Regression
Trainset RMSE	176.8036	188.3476
Testset RMSE	157.5959	185.5398

Predictive accuracy of CART is higher than linear regression – its RMSE is lower than that for linear regression for both trainset and testset.

[7] https://statisticalhorizons.com/multicollinearity

Answer to Q5:

GENERAL

- This is assumed to be a supervised learning problem, where our goal is to predict values that are as close to the current premiums payable as possible. In turn, this assumes that the current methodology used by staff to calculate payables are already the most ideal values for the company's profit generation.
- Our analysis of key predictors in question 4b also does not consider the "difference between statistical significance and practical significance"[8], an important distinction in applying machine learning to business.
    - For example, the number of major surgeries, while statistically significant, may not be practically significant since the average number of major surgeries done is 0.6, and the difference coefficient (i.e. per surgery) is a mere -$30.
    - The value that any of these variables bring must also be balanced with the cost and effort required to collect them, since data collection also takes time and effort which is exactly what we are attempting to reduce.

DATA DISTRIBUTION

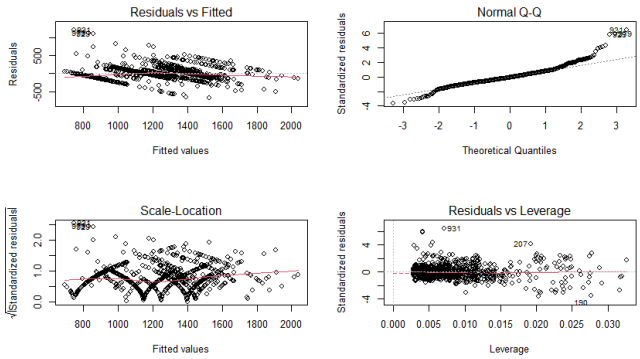
- A glaring issue arises when analysing the distribution of the original dataset, which is imbalanced.
    - As discovered in Q3, Premiums exist in 24 discrete and unevenly distributed levels.

750	800	850	900	950	1000	1050	1100	1150	1200	1250	1300	1350	1400	1450	1500	1550	1600	1700	1750	1800	1900	1950	2000
202	4	1	2	15	2	26	1	249	4	103	7	1	132	72	47	31	4	2	41	2	34	5	1

    - Regardless of the model chosen, this imbalanced distribution will "limit our predictive accuracy"[9] especially in these regions of sparse data, e.g. between 751 and 1149 where the model has 51 data points to work with (even lesser if we perform a train-test split).
    - This suggests the need to balance the data. However, because the Premium variable is ultimately a measure of money, which is continuous, applying data balancing techniques to a continuous variable "can cause ambiguity."[10]

LINEAR REGRESSION

- For linear regression, the three assumptions may not hold true.



- The Q-Q plot shows issues with the normality assumption particularly on the right tailed end but is not too serious otherwise. The Cook's Distance plot similarly shows no serious issues.
- However, for Residuals vs Fitted graph, even though the red line is straight, there is a clear downward-sloped pattern at each segment of the x-axis. The same pattern is observed in the Scale-Location graph where the points show a clear non-linear pattern.
- This suggests the need for further analysis on the data partitioned at these clusters.

[8] https://hbr.org/2016/02/a-refresher-on-statistical-significance
[9] https://machinelearningmastery.com/imbalanced-classification-is-hard/
[10] https://towardsdatascience.com/strategies-and-tactics-for-regression-on-imbalanced-data-61eeb0921fca

Answer to Q6:

CART is successful in this overall application of generating insurance premiums payable, especially when considering the general needs of the insurance company.

- An RMSE of $150 is considered relatively low when considering that premiums are currently generated in discrete levels of $50 and that the general ranges of insurance premiums (as seen in the graph in Q3) are $750, $1150, and $1400-$1550. With an error margin of $150, insurance estimates will still be within these 3 ranges.
- In addition, it is unclear whether this decision to charge in levels is intentional or an undesirable limitation due to the current methodology the company uses to charge premiums. In this regard, CART is able to generate categorical predictions too, and gives mostly accurate predictions particularly in regions with sufficient data points.

	Actual																							
Predicted	750	800	850	900	950	1000	1050	1100	1150	1200	1250	1300	1350	1400	1450	1500	1550	1600	1700	1750	1800	1900	1950	2000
750	202	3	0	1	0	0	0	0	0	2	0	3	0	0	0	0	0	1	0	0	1	0	4	0
800	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
850	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
900	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
950	0	0	0	0	15	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	2	0	0
1000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1050	0	0	0	0	0	0	26	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1150	0	0	0	0	0	1	0	0	248	0	0	2	0	0	0	0	0	3	2	0	0	0	1	0
1200	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1250	0	1	0	1	0	0	0	0	0	0	96	0	1	0	0	2	0	0	0	1	0	2	0	0
1300	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1350	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1400	0	0	0	0	0	0	0	0	0	1	0	0	0	132	0	1	0	0	0	0	0	0	0	1
1450	0	0	0	0	0	1	0	0	0	0	0	1	0	0	72	0	0	0	0	2	0	0	0	0
1500	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	43	2	0	0	1	0	0	0	0
1550	0	0	0	0	0	0	0	1	0	0	3	0	0	0	0	0	29	0	0	2	0	2	0	0
1600	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1700	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1750	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	33	1	2	0	0
1800	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1900	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	26	0
1950	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 6.1: Confusion Matrix for Categorical CART model, showing accurate predictions (numbers on the diagonals).

- One caveat to note is CART's "variance across samples, [where] the tree structure and resulting estimates are not necessarily stable in new samples."[11]
    - This feature of CART is accentuated by the above imbalanced data problem, where the data provided in the trainset can vary considerably from the testset depending on which datapoints in the minority regions are chosen (based on the seed).
    - This is also seen in Question 4d, where the testset RMSE for CART is lower than that of the trainset, suggesting that the model has been underfitted.
    - This is to be expected, since our CART model was 1 SE optimised making it difficult to "recognise patterns [in minority classes]".[12]

- However, regardless of these caveats and the specific CART model used, I would argue that CART has still been successful in this application, because our specific scenario of applying Machine Learning does not require pinpoint accuracy.
- Even the current premiums being generated are based on estimates, where "actuaries assess the risk of financial loss using mathematics and statistics to predict the likelihood of an insurance claim"[13]. Hence, the prediction does not need to be extremely accurate since a ballpark measure would do.
- Moreover, at the end of the day, consumers will still be purchasing from and be charged by an insurance agent. This manual human involvement ensures that any erroneous outliers generated by the model will be picked up by the agent, who will be able to use their experience and industry expertise to determine on a case-by-case basis if an ML-generated payable estimate is accurate.

[11] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4743660/
[12] https://stats.stackexchange.com/questions/450634/why-decision-tree-handle-unbalanced-data-well
[13] https://www.investopedia.com/ask/answers/09/calculating-premium.asp