A Guide to Pricing Insurance Products

Lee Yen May

Data Processing

```
library(summarytools)
library(dplyr)
library("ggplot2")
library("GGally")
car <- read.csv("CarInsurance.csv")
```

A descriptive table about the variables within the dataset:

Variable Name	Variable Attribute	Variable Description	Variable Type	Remarks
Gender	Policy Holder	Gender of policy holder (ie. claimant)	Categorical	Female: 1 Male: 0
Vehicle_Make_Region	Vehicle	Country of vehicle make	Categorical	EAST ASIA LOCAL EUROPE
Vehicle_Age	Vehicle	Age of vehicle	Categorical	0-20 years
Claim_Amount_Indicator	Insurance	An indicator of large or not large claim to the insurance company	Categorical	1: Large claim 0: not large claim
Age	Policy Holder	Age of policy holder (ie. claimant)	Continuous	-
Engine_Capacity	Vehicle	Engine capacity of the vehicle	Continuous	-
Car_Value	Vehicle	Value of vehicle at time of claim (in RM)	Continuous	-
Claim_Amount	Insurance	Claim amount (in RM)	Continuous	-

✓

Rename the variables:

not needed since make_region is already of char type.

```
Gender <- factor(car$Gender)
Vehicle_Make_Region <- factor(car$Vehicle_Make_Region)
Vehicle_Age <- factor(car$Vehicle_Age)
Claim_Amount_Indicator <- factor(car$Claim_Amount_Indicator)

Age <- car$Age
```

```
Engine_Capacity <- car$Engine_Capacity
Car_Value <- car$Car_Value
Claim_Amount <- car$Claim_Amount

cont <- car %>% dplyr::select(Age, Engine_Capacity, Car_Value, Claim_Amount)
cat <- car %>% dplyr::select(Gender, Vehicle_Make_Region, Vehicle_Age, Claim_Amount_Indicator)
```

Q1: Exploratory Data Analysis

We observe the relationship between each variable and Claim_Amount.

Continuous variable (Age, Engine_Capacity, Car_Value)

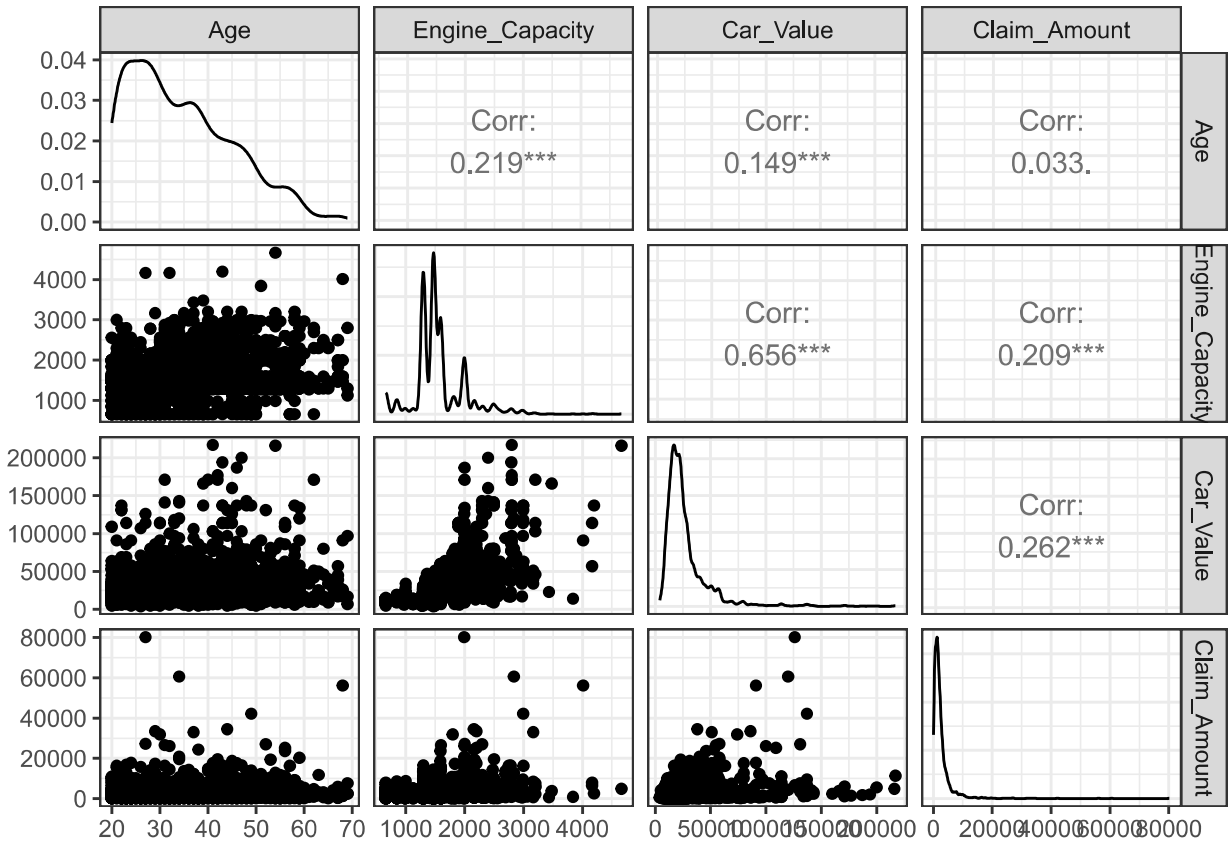
Summary statistics

```
summary(cont)
```

```
##       Age         Engine_Capacity   Car_Value       Claim_Amount
##  Min.   :20.00   Min.   : 659    Min.   :  4000   Min.   :  100
##  1st Qu.:26.00   1st Qu.:1298    1st Qu.: 16000   1st Qu.: 1000
##  Median :33.00   Median :1468    Median : 22000   Median : 1700
##  Mean   :34.64   Mean   :1555    Mean   : 27202   Mean   : 2533
##  3rd Qu.:42.00   3rd Qu.:1597    3rd Qu.: 30000   3rd Qu.: 2900
##  Max.   :69.00   Max.   :4663    Max.   :217000   Max.   :80200
```

Scatterplots: Relationship with Claim_Amount

```
ggpairs(cont)+theme_bw()
```



Interpretation

Claimant's age has nearly no correlation to claim amount, implying that age does not affect claim amount. Car engine capacity and car value have a weak positive relationship to claim amount, indicating that claim amount may increase as engine capacity and car value increase. However, we must note that claim amount, claimant's age, car engine capacity and car value all have right-skewed distributions. Claim amount has the most obvious high influential points, ie. there are a few exceptionally high claim amount cases in the sample, this is especially obvious with the large difference between maximum claim amount and the 3rd quartile. Therefore the significance of the difference between claim amount and these continuous variables must be further investigated.

Categorical variables (Gender, Vehicle_Make_Region, Vehicle_Age, Claim_Amount_Indicator)

Summary statistics

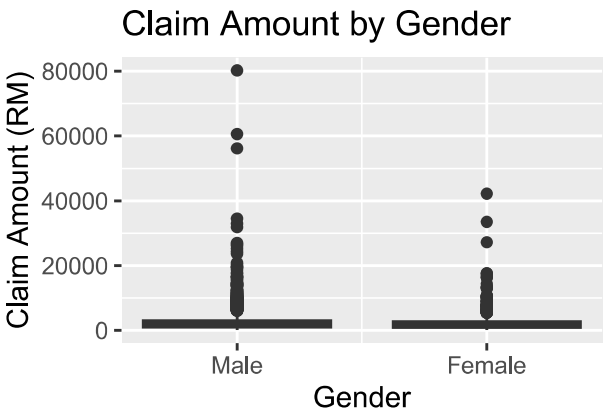
```
freq(cat)
```

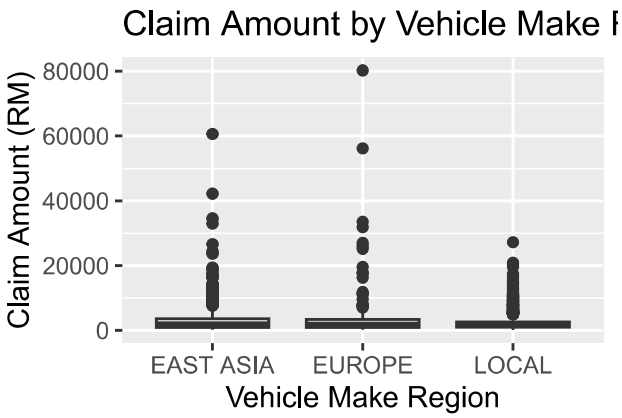
```
## Frequencies
## cat$Gender
## Type: Integer
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           0   2324     70.75          70.75     70.75          70.75
##           1    961     29.25         100.00     29.25         100.00
##        <NA>      0                                0.00         100.00
```

```
##       Total   3285    100.00         100.00    100.00         100.00
##
## cat$Vehicle_Make_Region
## Type: Character
##
##                   Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## --------------- ------ --------- -------------- --------- --------------
##       EAST ASIA    771     23.47          23.47     23.47          23.47
##          EUROPE    242      7.37          30.84      7.37          30.84
##           LOCAL   2272     69.16         100.00     69.16         100.00
##            <NA>      0                               0.00         100.00
##           Total   3285    100.00         100.00    100.00         100.00
##
## cat$Vehicle_Age
## Type: Integer
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           0    364     11.08          11.08     11.08          11.08
##           1    342     10.41          21.49     10.41          21.49
##           2    393     11.96          33.46     11.96          33.46
##           3    355     10.81          44.26     10.81          44.26
##           4    364     11.08          55.34     11.08          55.34
##           5    336     10.23          65.57     10.23          65.57
##           6    166      5.05          70.62      5.05          70.62
##           7    184      5.60          76.23      5.60          76.23
##           8    166      5.05          81.28      5.05          81.28
##           9    202      6.15          87.43      6.15          87.43
##          10    192      5.84          93.27      5.84          93.27
##          11     36      1.10          94.37      1.10          94.37
##          12     32      0.97          95.34      0.97          95.34
##          13     45      1.37          96.71      1.37          96.71
##          14     37      1.13          97.84      1.13          97.84
##          15     33      1.00          98.84      1.00          98.84
##          16     10      0.30          99.15      0.30          99.15
##          17      4      0.12          99.27      0.12          99.27
##          18      4      0.12          99.39      0.12          99.39
##          19     14      0.43          99.82      0.43          99.82
##          20      6      0.18         100.00      0.18         100.00
##        <NA>      0                               0.00         100.00
##       Total   3285    100.00         100.00    100.00         100.00
##
## cat$Claim_Amount_Indicator
## Type: Integer
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           0   2318     70.56          70.56     70.56          70.56
##           1    967     29.44         100.00     29.44         100.00
##        <NA>      0                               0.00         100.00
##       Total   3285    100.00         100.00    100.00         100.00
```

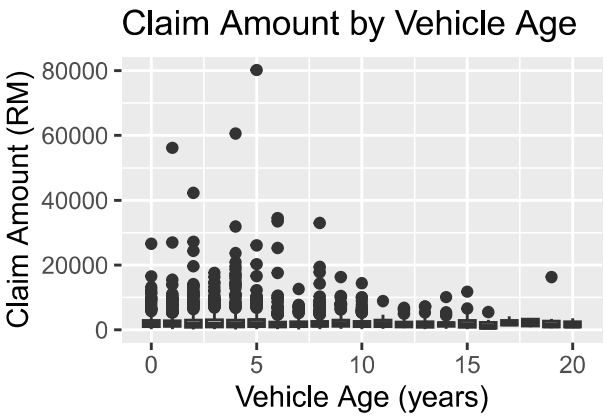
Boxplots: Relationship with Claim_Amount

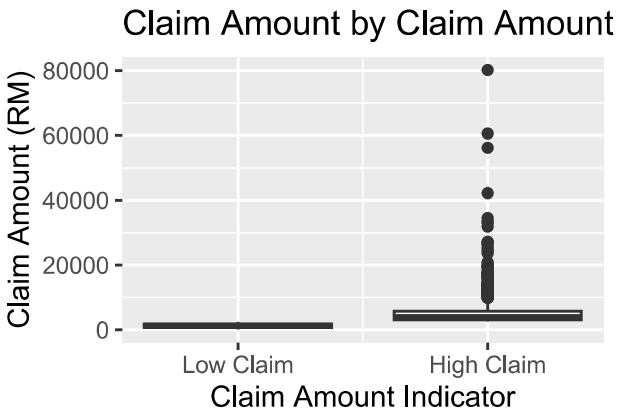
```
library(gridExtra)

p1 <- ggplot(car, aes(x = Gender, y = Claim_Amount, group = Gender)) +
      geom_boxplot() +
      labs(title = "Claim Amount by Gender", x = "Gender", y = "Claim Amount (RM)") +
      scale_x_continuous(breaks = c(0, 1), labels = c("Male", "Female"))

p2 <- ggplot(car, aes(x = Vehicle_Make_Region, y = Claim_Amount,
      group = Vehicle_Make_Region)) + geom_boxplot() +
      labs(title = "Claim Amount by Vehicle Make Region",
      x = "Vehicle Make Region", y = "Claim Amount (RM)")

p3 <- ggplot(car, aes(x = Vehicle_Age, y = Claim_Amount, group = Vehicle_Age)) +
      geom_boxplot() +
      labs(title = "Claim Amount by Vehicle Age",
           x = "Vehicle Age (years)", y = "Claim Amount (RM)")

p4 <- ggplot(car, aes(x = Claim_Amount_Indicator, y = Claim_Amount,
      group = Claim_Amount_Indicator)) + geom_boxplot() +
      labs(title = "Claim Amount by Claim Amount Indicator",
      x = "Claim Amount Indicator", y = "Claim Amount (RM)") +
      scale_x_continuous(breaks = c(0, 1), labels = c("Low Claim", "High Claim"))


grid.arrange(p1, p2, p3, p4, ncol=2, heights=c(4, 4))
```





✓





✓

Interpretation

Due to the right-skewed nature of claim amount, the median and quartiles concentrate at the lower end of claim amount for all variables, ie. boxes with long right whiskers.

We observe no significance between claim amount and claimant's gender due to the overlapping boxes, though the claim amount influential points are higher among male claimants, ie. among those exceptionally high claim cases, most of them are from male claimants. This implies that gender does not affect claim pricing.

We observe no significance between claim amount and vehicle make region due to the overlapping boxes, though the claim amount influential points are higher among european cars, followed by east asia and local, ie. among those exceptionally high claim cases, most of them are from european cars. This implies that vehicle make region does not affect claim pricing.

Since vehicle age is an ordinal variable, we observe that claim amount seems to decrease as vehicle age increases. However, the significance of the interpretation must be further investigated as we have more newer cars within the sample.

```
lowclaim_data <- subset(car, Claim_Amount_Indicator == 0)
highclaim_data <- subset(car, Claim_Amount_Indicator == 1)
summary(lowclaim_data$Claim_Amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     100     700    1300    1272    1800    2500
```

```
summary(highclaim_data$Claim_Amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2600    3100    4000    5557    5800   80200
```

We observe a clear distinction between low claim indicator and high claim indicator from the boxplot. The cut-off price that the insurance company is using seems to be roughly RM 2500 - RM2600.

Q2: Canonical Correlation Analysis (CCA)

```
library(CCA)

car$Female <- car$Gender
car$Vehicle_Make_Region <- ifelse(car$Vehicle_Make_Region %in%
                          c("EUROPE", "EAST ASIA"), "Foreign", "Local")
car$Foreign <- ifelse(car$Vehicle_Make_Region == "Foreign", 1, 0)
X1 = car %>% dplyr::select(Female, Age)  %>%as.matrix # policy holders
Y1 = car %>% dplyr::select(Foreign, Engine_Capacity, Vehicle_Age, Car_Value) %>%
     as.matrix # vehicle attributes

cca <- cc(X1,Y1)
cca$cor # first and second canonical correlation
```

```
## [1] 0.25324301 0.06313715
```

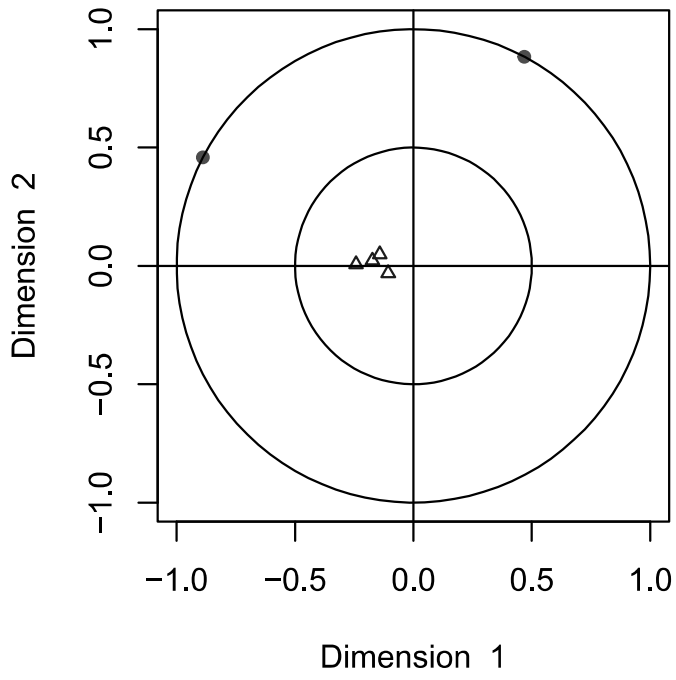
```
cca$scores$corr.Y.xscores # blue
```

```
##                       [,1]         [,2]
## Foreign         -0.1732610  0.022276700
## Engine_Capacity -0.2425588  0.006711868
## Vehicle_Age     -0.1063177 -0.031259910
## Car_Value       -0.1419870  0.049067547
```

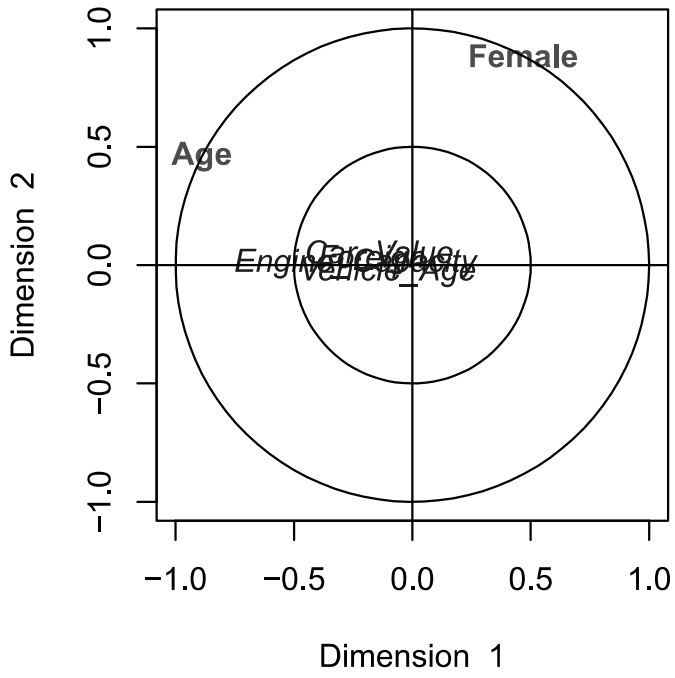
```
cca$scores$corr.X.xscores # red
```

```
##              [,1]      [,2]
## Female  0.4677574 0.8838569
## Age    -0.8889707 0.4579640
```

```
plt.cc(cca, type="v")
```




```
plt.cc(cca, var.label=TRUE, type="v")
```



Interpretation

The canonical correlations of the CCA provides information on the relationship between the Y variables (vehicle attributes) and the X variables (insurance policy holder information). The first canonical correlation (0.2532) indicates a moderate positive relationship between the two sets of variables. The second canonical correlation (0.0631) indicates an extremely weak positive relationship and may be negligible.

The variable plot and the matrix of correlations between the Y variables and the first two CC scores (ie. canonical variables) show that:

- The first CC score has a moderate negative correlation with the all original variables, ie. Foreign, Engine_Capacity, Vehicle_Age, and Car_Value, with Engine_Capacity being the most negative. We observe that the blue coordinates that represent the Y variables lie on the negative region of dimension one, within the smaller circle representing correlation within -0.5 to 0 . This suggests that a high negative first CC score represents a policy holder with a vehicle of foreign car makes, higher engine capacities, older vehicles and higher car value. The policy holder most likely has a vehicle of high engine capacity because it is the most dominating factor.
- The second CC score is most positively correlated with Car_Value, followed by Vehicle_Age and Foreign. This suggests that a high positive second CC score represents a policy holder with a vehicle of high value, old age and foreign car make. However, since the blue coordinates are close to 0 on dimension 2, we can consider to neglect this interpretation.

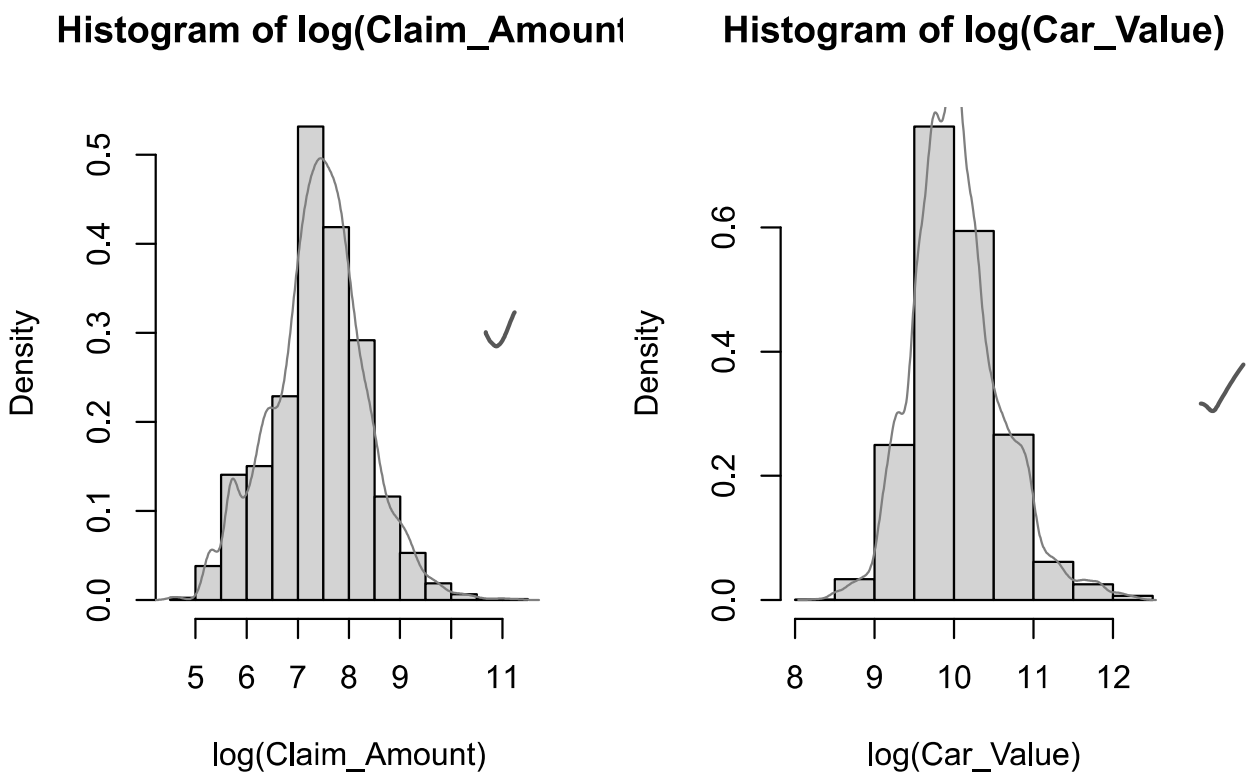
In conclusion, the CCA suggests that there is a moderate positive relationship between policy holder information and vehicle attributes. A high negative first CC score appears to capture a high risk profile, while the second CC score is primarily related to car value but insignificant.

Q3: Hypothesis Testing

Hotelling's two samples, unpaired case

i. We are interested to know whether claim amount and car value have significant difference among male and female claimants. To initialise this investigation, we log transform these two variables and check that they follow a normal distribution.

```
library(ICSNP)
car$Claim_Amount_log <- log(Claim_Amount)
car$Car_Value_log <- log(Car_Value)

par(mfrow=c(1,2))
hist(car$Claim_Amount_log, freq = FALSE, xlab = "log(Claim_Amount)",
     main = "Histogram of log(Claim_Amount)")
lines(density(car$Claim_Amount_log), col=2)
hist(car$Car_Value_log, freq = FALSE, xlab = "log(Car_Value)",
     main = "Histogram of log(Car_Value)")
lines(density(car$Car_Value_log), col=2)
```



log(Claim_Amount) and log(Car_Value) follow a normal distribution. Then, compute their respective means.

```
X1 <- car %>% dplyr::filter(Female==1) %>% dplyr::select(Claim_Amount_log, Car_Value_log)
X2 <- car %>% dplyr::filter(Female==0) %>% dplyr::select(Claim_Amount_log, Car_Value_log)
```

```
colMeans(X1)
```

```
## Claim_Amount_log    Car_Value_log
##         7.343156        10.014683
```

```
colMeans(X2)
```

```
## Claim_Amount_log    Car_Value_log
##         7.424524        10.045475
```

We observe that both males and females might share similar mean log claim amount and mean log car value. To investigate the significance in mean differences, we select Hotelling's two sample T2-test, unpaired case as the multivariate hypothesis test because:

1. We are comparing two independent groups (male and female).
2. log claim amount and log car value follow a multivariate normal distribution.
3. We make the assumption that the covariance matrices for both groups are the same.

```
HotellingsT2(X1, X2)
```

```
##
##  Hotelling's two sample T2-test
##
## data:  X1 and X2
## T.2 = 3.2352, df1 = 2, df2 = 3282, p-value = 0.03948
## alternative hypothesis: true location difference is not equal to c(0,0)
```

The p-value of 0.03948 is less than the significance level of 0.05, indicating that there is evidence to reject the null hypothesis and conclude that there is a significant difference between males and females regarding the mean log claim amount and mean log car value. This concludes that male and female claimants have a different claim amount and car value on average.

ii. We are interested to know whether claim amount and car value have significant difference among local and foreign car makes. Similar to (a), we use the log transformed values.

```
Y1 <- car %>% dplyr::filter(Foreign==1) %>% dplyr::select(Claim_Amount_log, Car_Value_log)
Y2 <- car %>% dplyr::filter(Foreign==0) %>% dplyr::select(Claim_Amount_log, Car_Value_log)

colMeans(Y1)
```

```
## Claim_Amount_log    Car_Value_log
##         7.544305        10.431035
```

```
colMeans(Y2)
```

```
## Claim_Amount_log    Car_Value_log
##         7.336701         9.860544
```

We observe that both foreign and local car makes seem to have distinct mean log claim amount and mean log car value. To investigate the significance in mean differences, we select Hotelling's two sample T2-test, unpaired case as the multivariate hypothesis test because:

1. We are comparing two independent groups (foreign and local).
2. log claim amount and log car value follow a multivariate normal distribution.
3. We make the assumption that the covariance matrices for both groups are the same.

```
HotellingsT2(Y1, Y2)
```

```
##
##  Hotelling's two sample T2-test
##
## data:  Y1 and Y2
## T.2 = 477.91, df1 = 2, df2 = 3282, p-value < 2.2e-16
## alternative hypothesis: true location difference is not equal to c(0,0)
```

The p-value of approximately 0 is less than the significance level of 0.05, indicating that there is evidence to reject the null hypothesis and conclude that there is a significant difference between foreign and local car makes regarding the mean log claim amount and mean log car value. This concludes that foreign and local car makes have a different claim amount and car value on average.

Check Assumption

Hotelling T2-test for 2 samples (unpaired case) requires the assumption that covariance matrices for both groups are the same.

For (i), we check if the assumption is reasonable by computing the sample covariances for log(Claim_Amount) and log(Car_Value) for the two groups: male and female.

```
print("For Female:")
```

```
## [1] "For Female:"
```

```
cov(X1)
```

```
##                  Claim_Amount_log Car_Value_log
## Claim_Amount_log       0.86129012    0.09109892
## Car_Value_log          0.09109892    0.29522245
```

```
print("For Male:")
```

```
## [1] "For Male:"
```

```
cov(X2)
```

```
##                  Claim_Amount_log Car_Value_log
## Claim_Amount_log       0.82785494    0.09085807
## Car_Value_log          0.09085807    0.31367636
```

For the female group, the covariance between log(Claim_Amount) and log(Car_Value) is 0.0911, while the variances of log(Claim_Amount) and log(Car_Value) are 0.8613 and 0.2952, respectively.

For the male group, the covariance between log(Claim_Amount) and log(Car_Value) is 0.0909, while the variances of log(Claim_Amount) and log(Car_Value) are 0.8279 and 0.3137, respectively.

Since the covariance between log(Claim_Amount) and log(Car_Value) is very similar for both groups and the variances of log(Claim_Amount) and log(Car_Value) are not too different, it is reasonable to assume that the covariance matrices for both groups are the same.

For (ii), we check if the assumption is reasonable by computing the sample covariances for log(Claim_Amount) and log(Car_Value) for the two groups: local and foreign.

```
print("For Foreign: ")
```

```
## [1] "For Foreign: "
```

```
cov(Y1)
```

```
##                   Claim_Amount_log Car_Value_log
## Claim_Amount_log        0.9485383     0.1102777
## Car_Value_log           0.1102777     0.4330314
```

```
print("For Local:")
```

```
## [1] "For Local:"
```

```
cov(Y2)
```

```
##                   Claim_Amount_log Car_Value_log
## Claim_Amount_log       0.77689550    0.04651782
## Car_Value_log          0.04651782    0.15256597
```

For the foreign group, the covariance between log(Claim_Amount) and log(Car_Value) is 0.1103, while the variances of log(Claim_Amount) and log(Car_Value) are 0.9485 and 0.4330, respectively.

For the local group, the covariance between log(Claim_Amount) and log(Car_Value) is 0.0465, while the variances of log(Claim_Amount) and log(Car_Value) are 0.7769 and 0.1526, respectively.

Since the covariance between log(Claim_Amount) and log(Car_Value) is very different for both groups and the variances of log(Claim_Amount) and log(Car_Value) are also very different, it is NOT reasonable to assume that the covariance matrices for both groups are the same.

Q4: Linear Discriminant Analysis (LDA)

We randomly split the data into test and training sets by 20:80 ratio.

```
library(MASS)
set.seed(123) # for reproducibility
train <- sample(nrow(car), nrow(car)*0.8)
train_data <- car[train,]
test_data <- car[-train,]
nrow(train_data)
```

```
## [1] 2628
```

```
nrow(test_data)
```

```
## [1] 657
```

```
# scale/normalise the variables Age, Engine_Capacity, Vehicle_Age, and Car_Value
train_data[, c("Age", "Engine_Capacity",
          "Vehicle_Age", "Car_Value")] <- scale(train_data[, c("Age", "Engine_Capacity",
                                                              "Vehicle_Age", "Car_Value")])

# train a LDA classifier (full model) to predict whether a claim is large
car.lda <- lda(Claim_Amount_Indicator ~ Female + Age + Engine_Capacity +
                 Vehicle_Age + Car_Value + Foreign, train_data)

# predict on the test set and calculate the predictive accuracy
test_data[, c("Age", "Engine_Capacity",
          "Vehicle_Age", "Car_Value")] <- scale(test_data[, c("Age", "Engine_Capacity",
                                                              "Vehicle_Age", "Car_Value")])
car.pred <- predict(car.lda, test_data)
print(paste("The predictive accuracy is ",
sum(car.pred$class== test_data$Claim_Amount_Indicator)/dim(test_data)[1]*100, "%"))
```

```
## [1] "The predictive accuracy is  69.8630136986301 %"
```

```
table(car.pred$class, test_data$Claim_Amount_Indicator)
```

```
##
##       0   1
##   0 449 188
##   1  10  10
```

The classifier correctly predicted low claim for 449 instances and the high claim for 10 instances, but it falsely predicted the high claim for 188 instances and the low claim for 10 instances.

Since we observed that policy holder age does not affect claim amount, we carried out this additional step which will aid our discussion for verification.

```
# train a LDA classifier (full model) to predict whether a claim is large
car.lda2 <- lda(Claim_Amount_Indicator ~ Female + Engine_Capacity +
                  Vehicle_Age + Car_Value + Foreign, data = train_data)

# predict on the test set and calculate the predictive accuracy
car.pred2 <- predict(car.lda2, test_data)
print(paste("The predictive accuracy is ",
sum(car.pred2$class== test_data$Claim_Amount_Indicator)/dim(test_data)[1]*100, "%"))
```

```
## [1] "The predictive accuracy is  69.8630136986301 %"
```

Big applause !!

Q5: Discussion

Advice to insurance company: How might you use our analysis for insurance products pricing?

As consultants for your car insurance company, we are pleased to present the findings of our recent data analysis on your historical claims data. With Malaysia's Liberalisation of Motor Tariff on 1 July 2017, the insurance industry is now free to charge policyholders premiums based on risk profiles. This presents an exciting opportunity for your company to adopt a more sophisticated approach to pricing insurance products that reflects the risks of individual policyholders. By utilising the insights generated from our analysis for your product pricing strategies, your company can gain a competitive edge by pricing insurance policies more accurately and ultimately reducing the risk of financial loss.

In this report, we will discuss three topics:

1. Factors associated with claim amount
2. Risk profile formation strategies
3. Recommendations for product pricing

Factors associated with claim amount

In terms of policy holder attribute, we observed that age does not have an effect on claim amount. Though gender also has a similar outcome due to the similarity in the middle 50% of the claim amount, we observe that there are some exceptionally high claim amounts among male policy holders. We further analysed the significance of gender by investigating its effect on claim amount and car value collectively. We concluded that male and female policy holders have distinct claim amount and car value on average. In terms of vehicle attribute, we observe that engine capacity has the strongest positive relationship with claim amount, followed by car value. Vehicle car make does not seem to have a significant effect on claim amount but we observe some exceptionally high claim cases among european cars, followed by east asian then local cars. The claim amount also decreases as vehicle age increases. In terms of your company's current insurance policy, we discovered that your company classifies claim amount lower than approximately the range of RM 2500 - RM 2600 as low claim and the otherwise as high claim. In the subsequent sections, we will provide some recommendations to improve your current insurance policy.

Risk profile formation strategies

Based on a correlation analysis, we discovered a moderate positive relationship between vehicle attributes and policy holders attributes. We concluded that it is appropriate to categorise a high risk profile as policy holders who owns a vehicle of high engine capacity, foreign make, high car value and old age. These attributes are ordered from the most dominant to the least. This analysis aligns with our intuition that policy holder who drives a car with high engine capacity tends to speed as the car has higher power and acceleration, which can lead to more severe accidents. On the other hand, foreign make vehicles may require expensive spare parts and specialised mechanics, increasing repair costs. High car value vehicles may have more expensive parts and labour costs for repairs and older vehicles may have more wear and tear, increasing the likelihood of breakdowns and accidents.

Recommendations for product pricing

Based on our findings above, we relabelled 657 policy holders' claim data into high claim or low claim, taking into consideration policy holders' gender, age, car engine capacity, car age, car value and car make. We compared our new labels to your company's actual label and discovered that only 69.86% of our labels matched. To be specific, we have labelled 188 cases as high claim when the actual label from your company is low claim. Since we have concluded that policy holders' age is not a significant factor, we repeated our analysis while excluding age and arrived at a similar accuracy score of 69.56%. Both of these results imply that policyholders with certain high-risk characteristics have had significantly higher claim amounts. This suggests that your company may have been charging lower premiums than you should for these high-risk policyholders, which may lead to potential financial losses for your company in the future. Therefore, we recommend that your company adjusts premium rates by taking into consideration policy holders' gender,

car engine capacity, car age, car value and car make to accurately reflect the risk associated with each policy holders' profiles.

With our analysis, we hope your company can leverage the benefits of adopting a risk profiling approach to adjusting premium rates and embrace it as a key component of your product pricing strategy. However, our analysis consists of three limitations which we will discuss in the next section.

Limitations of our analysis

Uneven distribution of data

- Gender : The gender distribution in the dataset is imbalanced, with male policyholders being more than double the number of female policyholders.
- Vehicle Make Region : The distribution of vehicles across different regions is uneven, with local vehicles being over nine times more frequent than European and East Asian vehicles combined. Furthermore, our analysis has combined European and East Asian vehicles together, which may have masked some important information on their respective significance on claim amount.
- Vehicle Age : The sample size decreases as vehicle age increases, with a large drop in sample size in between the following age categories: 0-5 years, 6-10 years, 11-15 years, and 16-20 years.
- Age, Engine_Capacity, Car_Value, and Claim_Amount are all right-skewed, meaning that we have more claim data available for young policyholders, low engine capacity, low car value, and low claim amount.

Therefore, this limitation may affect the accuracy of our analysis in determining the impact of factors above on claim amount.

Presence of outliers

Claim amount may contain outliers because there is a large difference between the maximum claim amount and claim amount at the 75% point of the claim data. Therefore, further investigation is needed to confirm if they are outliers or exceptionally high claim amounts.

Lack of variables

We have only worked with six predictors (ie. Female, Age, Engine Capacity, Vehicle Age, Car Value, Car Make) but there are many more predictors that can contribute to risk profiling, such as:

- Driving experience: number of years the driver has been holding a driving license and not the actual driving experience.
- Nature of occupation: determined by whether your job is based indoor or outdoor. For example, sales job is considered mostly outdoor and it will have a higher rate.
- No claim discount (NCD): Each private car owner is entitled to NCD ranging from 25% to 55% as provided in their individual policy. The maximum NCD offered currently is at 55%.
- Type of car: This refers to the vehicle type such as saloon, MPV, SUV or 4-wheel drive. Some company charges differently depending on the car type.
- Type of engine: It's important to know that most car insurance company impose higher car insurance premium on cars with turbo engine. Some insurers even reject cars with turbo engines.
- Driver's claims record: A driver who is constantly involved/causing accidents and making insurance claims, is obviously a high risk case and will have to endure a higher premium.

Hence, your company is encouraged to collect or provide more data to us to improve the accuracy of our analysis.