MVA2 New

Alicia Clara Trevina

2023-04-13

Introduction

We will be investigating the `Insurance.csv` data set by applying Canonical Correlation Analysis (CCA), Hypothesis Testing, Linear Discriminant Analysis (LDA). For this data set, we will assume the data set has been cleaned and therefore may proceed to exploratory analysis.

```
insurance <- read_csv("~/Desktop/R files/MVA_dataset/CarInsurance.csv",show_col_types=FALSE )
head(insurance)
```

```
## # A tibble: 6 x 8
##   Gender   Age Vehicle_Make_Region Engine_Capacity Vehicle_Age Car_Value
##    <dbl> <dbl> <chr>                         <dbl>       <dbl>     <dbl>
## 1      0    49 EAST ASIA                      1415           0     22000
## 2      0    32 LOCAL                           659           6      8000
## 3      0    30 LOCAL                           659           2     13000
## 4      1    39 LOCAL                          1298           4     19000
## 5      0    34 EAST ASIA                      1590           6     29000
## 6      0    34 EAST ASIA                      1955          13     11000
## # i 2 more variables: Claim_Amount <dbl>, Claim_Amount_Indicator <dbl>
```

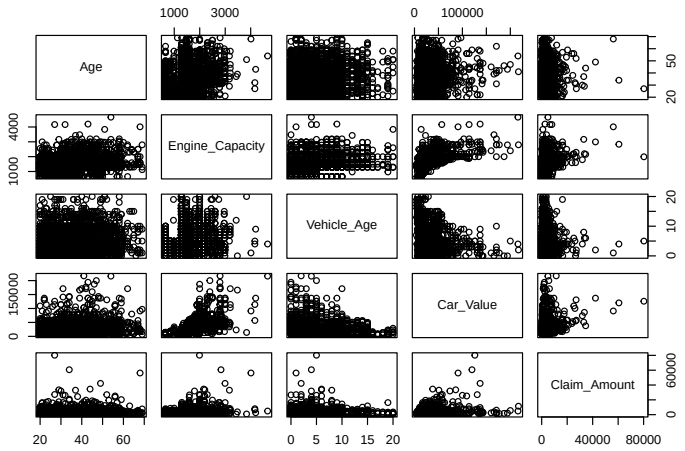
```
nrow(insurance)
```

```
## [1] 3285
```

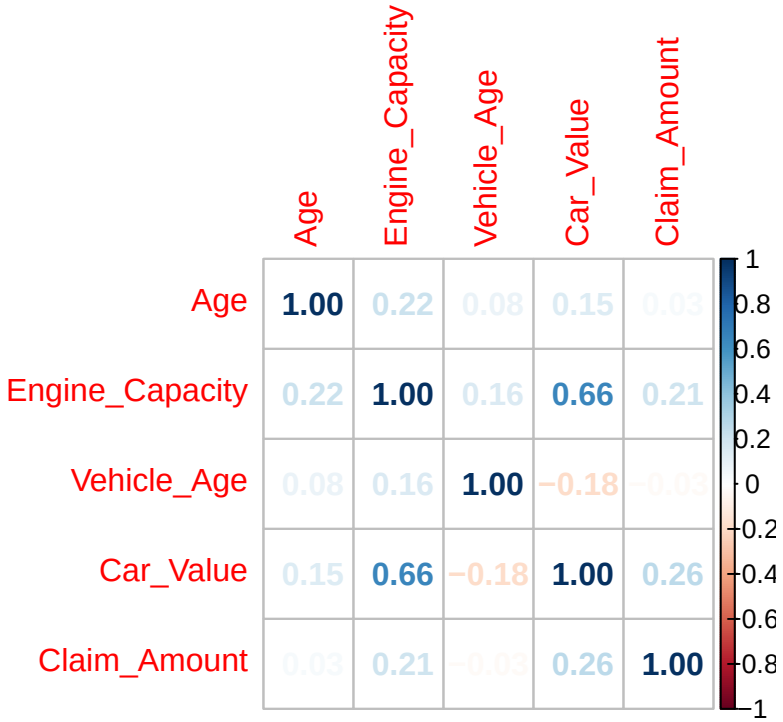
Exploratory Analysis

We will begin by observing the relationship of each variable. `Gender`, `Vehicle_Make_Region`, and `Claim_Amount_Indicator` are identified categorical variables within this data set. We will first explore the relationship between the continuous variables first: `Vehicle_Age`,`Age`,`Engine_Capacity`,`Car_Value` against `Claim_Amount`.

```
pairs(insurance[, c("Age", "Engine_Capacity", "Vehicle_Age",
                "Car_Value", "Claim_Amount")])
```

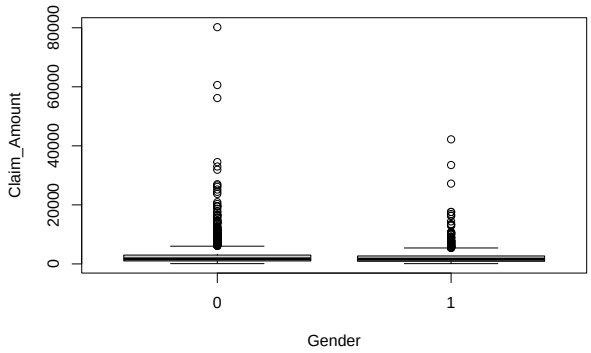


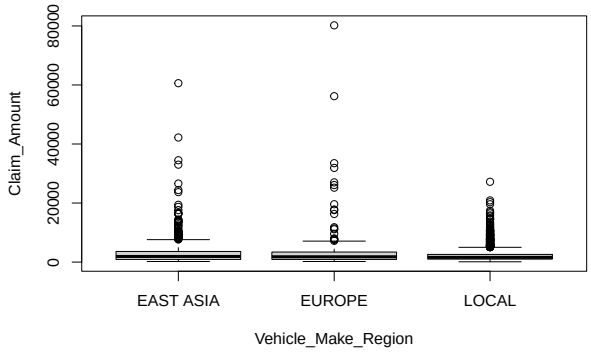

```
df = subset(insurance, select = c("Age","Engine_Capacity","Vehicle_Age","Car_Value","Claim_Amount"))
corrplot(cor(df), method = 'number')
```

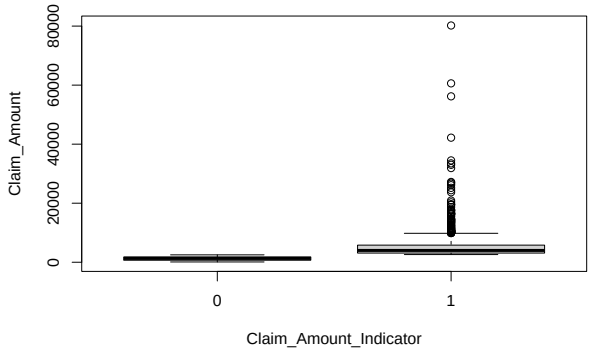


Based on the pairplot and the correlation matrix plot, the relationship of `Age`, `Engine_Capacity`, `Vehicle_Age`, and `Car_Value` does not have a strong association with `Claim_Amount`. However, we observe that `Car_Value` has a moderate positive association with `Engine_Capacity`.

For the categorical variables, we will explore using boxplots.







Based on the three boxplots above, we can observe the following:

- In regards to `Gender`, the box plots from both genders are extremely short (low IQR) which suggests low variability in the `Claim_Amount`. However, there are large number of outliers especially for male insurance claimers which suggests that these `Claim_Amount` are special cases.
- In regards to `Vehicle_Make_Region`, the box plots are also short (low IQR) which suggest low variability in the `Claim_Amount` within these regions. However, there are large number of outliers from all 3 regions which show extreme cases of `Claim_Amount`.

- In regards to `Claim_Amount_Indicator`, the box plots are also short (low IQR) however we can observe that `Claim_Amount_Indicator`=1's boxplot is higher, which suggests a higher median value. However, the large number of outliers in `Claim_Amount_Indicator`=1 show extreme cases of `Claim_Amount`.

Canonical Correlation Analysis:

Prior to conducting canonical correlation analysis (CCA), we will first be renaming the `Gender` column to `Female` (i.e. `Female` is indicate as dummy variable 1) and group `European` and `East Asian` into one category and change the column name to `Foreign`.

```
colnames(insurance)[colnames(insurance) == "Gender"] = "Female"
head(insurance)
```

```
## # A tibble: 6 x 8
##   Female   Age Vehicle_Make_Region Engine_Capacity Vehicle_Age Car_Value
##    <dbl> <dbl> <chr>                         <dbl>       <dbl>     <dbl>
## 1      0    49 EAST ASIA                      1415           0     22000
## 2      0    32 LOCAL                           659           6      8000
## 3      0    30 LOCAL                           659           2     13000
## 4      1    39 LOCAL                          1298           4     19000
## 5      0    34 EAST ASIA                      1590           6     29000
## 6      0    34 EAST ASIA                      1955          13     11000
## # i 2 more variables: Claim_Amount <dbl>, Claim_Amount_Indicator <dbl>
```

```
# Create a new variable "gender_diverse" that combines "male" and "female"
insurance$Foreign <- ifelse(insurance$Vehicle_Make_Region %in% c("EUROPE", "EAST ASIA"), "FOREIGN", insu
```

Using CCA, we will investigate if there are any relationship between policyholder attributes ($\mathbf{x}$) and vehicle attributes ($\mathbf{y}$). We will treat `Female` and `Age` as the x-variables, and `Foreign`, `Engine_Capacity`, `Vehicle_Age`, and `Car_Value` as the y-variables. For the column `Foreign`, we will add an additional column called `Foreign_factor` to change the categorical variables into binary variables indicating `FOREIGN` to be 1 and `LOCAL` as 0.

```
# change the "Foreign" column to factor levels
insurance$Foreign_factor <- ifelse(insurance$Foreign == "LOCAL", 0, 1)
head(insurance$Foreign_factor)
```

```
## [1] 1 0 0 0 1 1
```

```
table <- insurance %>% dplyr::select(Female,Age,Foreign_factor,Engine_Capacity,Vehicle_Age,Car_Value)
knitr::kable(head(table,5),booktabs=TRUE,escape=FALSE)
```

Female	Age	Foreign_factor	Engine_Capacity	Vehicle_Age	Car_Value
0	49	1	1415	0	22000
0	32	0	659	6	8000
0	30	0	659	2	13000
1	39	0	1298	4	19000
0	34	1	1590	6	29000

```
X <- table[,c('Female','Age')] #X-variables
Y <- table[,c('Foreign_factor','Engine_Capacity','Vehicle_Age','Car_Value')] #Y-variables
Xcent <- sweep(X,2,colMeans(X))
Ycent <- sweep(Y,2 ,colMeans(Y))


# Canonical Correlation
insurance.cca <-cc(Xcent,Ycent)
insurance.cca$cor
```

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

```
#Canonical Correlation vectors for X
insurance.cca$xcoef
```

```
##               [,1]       [,2]
## Female  1.00657780 1.95390516
## Age    -0.08192075 0.04335435
```

```
#Canonical Correlation vectors for Y
insurance.cca$ycoef
```

```
##                           [,1]          [,2]
## Foreign_factor  -2.588552e-01  8.050851e-01
## Engine_Capacity -1.982919e-03 -1.759402e-03
## Vehicle_Age     -6.575206e-02 -7.153107e-02
## Car_Value        4.671701e-07  5.118255e-05
```

Based on the canonical correlation output above, the first canonical correlation coefficient of 0.253 suggest that the correlation between the x and y variables have a relatively weak positive association. The second canonical correlation coefficient of 0.006 suggest an even weaker correlation between the x and y variables. Therefore we will ignore the second canonical covariate score as it is negligible / close to zero.

Canonical Correlation Weights for X (Female and Age) and Y (Foreign and Engine Capacity)

$\eta_1 = 1.01(Female - \bar{Female}) - 0.08(Age - \bar{Age})$

$\psi_1 = -0.258(Foreign - \bar{Foreign}) - 0.002(EngineCapacity - \bar{EngineCapacity}) - 0.065(VehicleAge - \bar{VehicleAge}$

The first canonical correlation weights suggest the following:

- η_1 shows that `Female` has the largest weightage to impact the canonical covariate score and has a huge positive association to the canonical covariate score (1.01). `Age` on the other hand has a low negative weightage to the canonical correlation score and therefore can ignore as it is has a small impact on the canonical covariate score. In other words,females indicate indicate an increase to the overall canonical covariate score.
- ψ_1 shows that `Foreign` carries the largest weightage (-0.258). We can ignore `Car_Value`, `Foreign`, and `Vehicle_Age` due to its low weightage to the canonical correlation score. Therefore, a foreign car suggests a large increase in the canonical covariate score.

```
#Correlation matrix between the original variables and the standardised variables in CCA
insurance.cca$scores$corr.X.xscores
```

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

```
insurance.cca$scores$corr.Y.yscores
```

```
##                        [,1]       [,2]
## Foreign_factor  -0.6841689  0.3528303
## Engine_Capacity -0.9578106  0.1063062
## Vehicle_Age     -0.4198249 -0.4951112
## Car_Value       -0.5606747  0.7771582
```

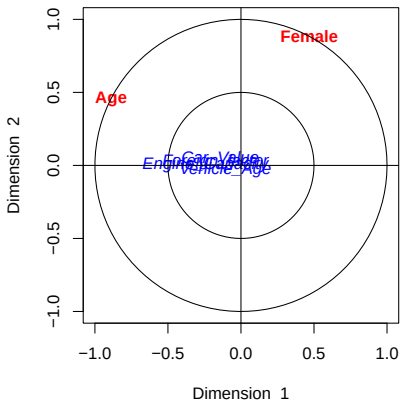
The correlation matrix output above of the first CC score shows that X has a moderate positive association with `Female` and a large negative association with `Age` .

On the other hand, the first CC score of Y shows that `Engine_Capacity`, `Foreign_Factor`, `Vehicle_Age`, and `Car_Value` are all negatively associated with Y with `Engine_Capacity` having the largest negative association with the variable Y.

Variable Plot

The circular plot below shows the correlation between the Y variables with the first two Y cc scores (in blue) and the X variables with the first two X cc scores (in red)

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

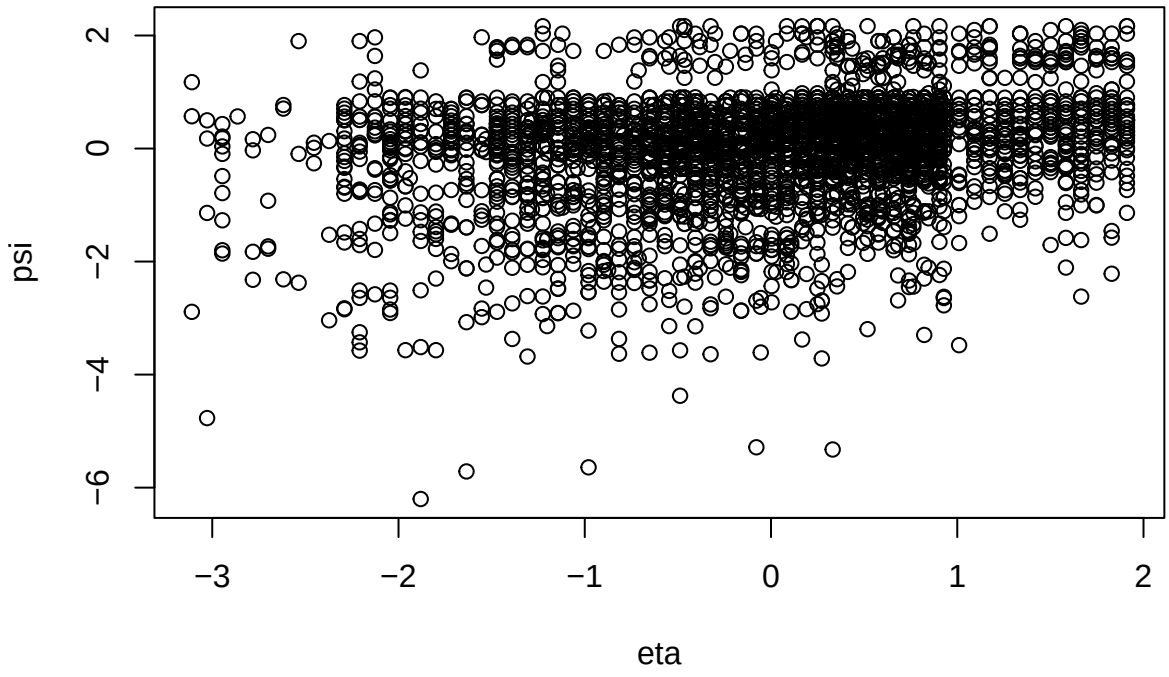


The circular plot shows the following conclusion:

- The first CC score for Y shows that all four variables `Engine_Capacity`, `Car_Value`, `Vehicle_Age`, and `Foreign` to be negatively correlated with Y
- The first CC scores for X shows that `Age` is negatively correlated with X while `Female` is positively correlated to X.

The first canonical correlation is 0.253 which does not suggest a very strong link between first CC score for X and first CC score for Y

```
eta = as.matrix(Xcent)%*%insurance.cca$xcoef[,1]
psi = as.matrix(Ycent)%*%insurance.cca$ycoef[,1]
plot(eta,psi)
```




```
cor(eta,psi)
```

```
##          [,1]
## [1,] 0.253243
```

In summary, the canonical correlation coefficient between the first CC score for X and Y (0.253) is relatively weak link which suggests that the variables' joint relationship with one another is not worth considering. The scatterplot (variable plot) between η_1 and ψ_1 demonstrates that there is weak positive association between policyholder and vehicle attributes which reflects the same result as the canonical correlation command `insurance.cca$cor` = 0.253.

Hypothesis Testing

We will be conducting a multivariate hypothesis testing to test whether the mean log claim amount and mean log car value are significantly different based on whether the policyholder is male or female. We will assume that the covariance matrices are the same for both populations. To do so, we will need to use the Hotelling's T^2 distribution squared test.

Multivariate Hypothesis to test difference between male and female policyholders on insurance claim amount

Null Hypothesis:

$H_0 : \boldsymbol{\mu}_{male} = \boldsymbol{\mu}_{female}$

Alternative Hypothesis:

$H_a : \boldsymbol{\mu}_{male} \neq \boldsymbol{\mu}_{female}$

where $\boldsymbol{\mu}_{\text{male}}$ and $\boldsymbol{\mu}_{\text{female}}$ are the mean vectors of the log claim amount and log car value for males and females, respectively.

```
# Subset the Claim_Amount by gender
male_data_claim <- as.matrix(log(subset(insurance, Female == 0)[c("Claim_Amount", "Car_Value")]))


female_data_claim <- as.matrix(log(subset(insurance, Female == 1)[c("Claim_Amount", "Car_Value")]))


# Calculate the sample means
(mean_male <- mean(male_data_claim[,1]))
```

```
## [1] 7.424524
```

```
(mean_female <- mean(female_data_claim[,1]))
```

```
## [1] 7.343156
```

```
mu = c(7.424524,7.424524)

#Hotelling Test
HotellingsT2(male_data_claim,female_data_claim)
```

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

Since the p-value is less than 0.05, there is sufficient evidence to reject the null hypothesis at the 5% level of significance and conclude that there is a significant difference between the mean log claim amount and log car value between male and female claimants.

Multivariate Hypothesis to test difference between local and foreign car makers on Car_Value

We will conduct another multivariate hypothesis test to test whether the mean log car value and mean log claim amount is significantly different between local and foreign car makers. We will once again assume that the covariance matrices are the same for both populations.

To conduct the hypothesis test, we will be using Hotelling's T^2 2-sample test as we have two populations to consider: local $(x_1, ..., x_m)$ and foreign car makers $(y_1, ..., y_m)$ in which $\sum$ is unknown.

Null Hypothesis:

$H_0 : \boldsymbol{\mu}_{local} = \boldsymbol{\mu}_{foreign}$

Alternative Hypothesis:

$H_a : \boldsymbol{\mu}_{local} \neq \boldsymbol{\mu}_{foreign}$

where $\boldsymbol{\mu}_{\text{local}}$ and $\boldsymbol{\mu}_{\text{foreign}}$ are the mean vectors of the log car value and log claim amount for local and foriegn manufacturers, respectively.

```
# Subset the Car_Value by Car Makers: Local or Foreign

local_data_car <- as.matrix(log(subset(insurance, Foreign_factor == 0, select = c("Car_Value", "Claim_Am


foreign_data_car <- as.matrix(log(subset(insurance, Foreign_factor == 1, select = c("Car_Value", "Claim_


# Calculate the sample means
(mean_local <- mean(local_data_car[,1]))
```

```
## [1] 9.860544
```

```
(mean_foreign<- mean(foreign_data_car[,1]))
```

```
## [1] 10.43103
```

```
#Hotelling Test
HotellingsT2(local_data_car,foreign_data_car)
```

```
##
##  Hotelling's two sample T2-test
##
## data:  local_data_car and foreign_data_car
## 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)
```

Since the p-value is less than 0.05, we strongly reject the null hypothesis at the 5% level of significance and conclude that there is a significant difference on the mean log `car_value` and mean log `claim_amount`between local and foreign manufacturers.

Computing the Sample Covariance for log `Claim_Amount` and `Car_Value`

```
cov(male_data_claim)
```

```
##              Claim_Amount  Car_Value
## Claim_Amount   0.82785494 0.09085807
## Car_Value      0.09085807 0.31367636
```

```
cov(female_data_claim)
```

```
##               Claim_Amount  Car_Value
## Claim_Amount   0.86129012 0.09109892
## Car_Value      0.09109892 0.29522245
```

```
cov(local_data_car)
```

```
##                Car_Value Claim_Amount
## Car_Value    0.15256597   0.04651782
## Claim_Amount 0.04651782   0.77689550
```

```
cov(foreign_data_car)
```

```
##               Car_Value Claim_Amount
## Car_Value    0.4330314    0.1102777
## Claim_Amount 0.1102777    0.9485383
```

Based on the output above, the assumption of equal covariance is not fulfilled for both `Car_Value` and `Claim_Amount`, therefore we did not fulfill the assumption of equal covariance to perform Hotelling's T^2 test.

Linear Discriminant Analysis (LDA)

As Hotelling's T^2 distribution test is not suitable, we will analyse the `insurance` dataset using linear discriminant analysis. We will start by splitting the data into test and training sets

```
set.seed(2)
test.ind<-sample(1:3285,size=1000)
insurance.test <- insurance[test.ind,]
insurance.train <- insurance[-test.ind,]
```

We will proceed to scaling the variables: `Age`,`Engine_Capacity`, `Vehicle_Age`, and `Car_Value` .

```
insurance$Age <-scale(insurance$Age)
insurance$Engine_Capacity <-scale(insurance$Engine_Capacity)
insurance$Vehicle_Age <- scale(insurance$Vehicle_Age)
insurance$Car_Value <- scale(insurance$Car_Value)
```

Next, we will use linear discriminant analysis to train a classifier to predict whether a claim is large. We will use the variables `Female`,`Age`,`Engine_Capacity`,`Vehicle_Age`,`Car_Value`, and `Foreign` as predictors

```
lda1 <- lda(Claim_Amount_Indicator~Female+Age+Engine_Capacity+Vehicle_Age+Car_Value+Foreign_factor,insu
ldapred <- predict(lda1,insurance.test)
```

Checking the predictive accuracy on the test test

```
(lda.accuracy <- sum(ldapred$class==insurance.test$Claim_Amount_Indicator)/1000)
```

```
## [1] 0.712
```

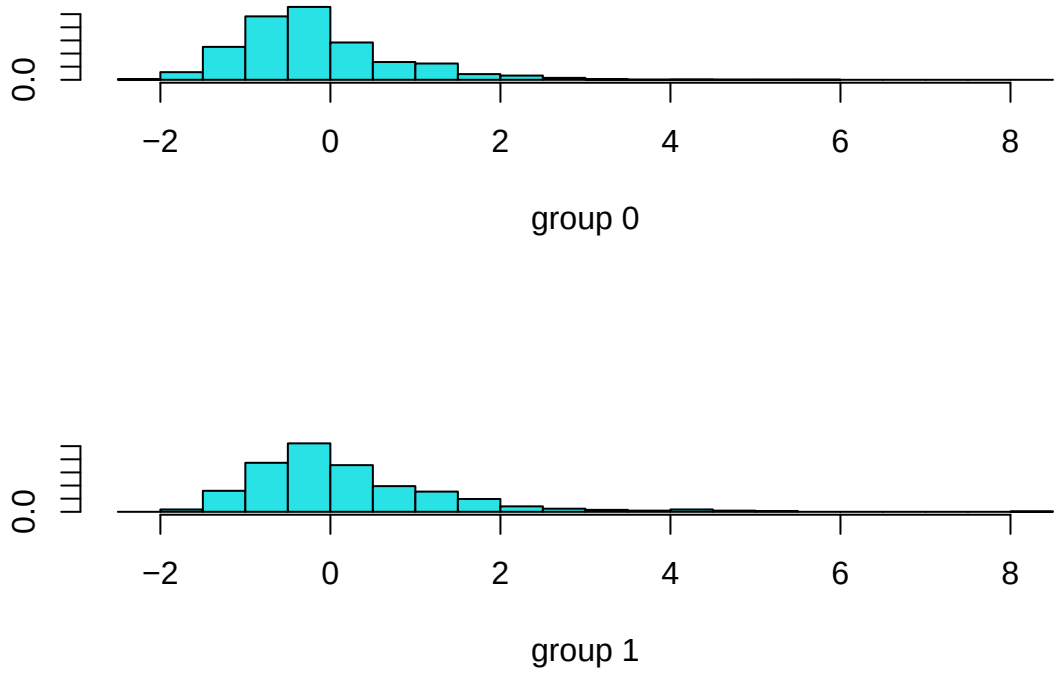
Based on the output, the predictive accuracy of the LDA model is 71.2% .

Discussion

```
table(insurance.test$Claim_Amount_Indicator,ldapred$class)
```

```
##
##       0   1
##   0 691  17
##   1 271  21
```

```
plot(lda1,insurance)
```



Based on the analysis,the insurance company may take into consideration of the following aspects:

- The large number of outliers across the dataset at the exploratory analysis boxplot suggest that the high variability in the data and that the insurance company should look into expanding their options of insurance premiums. Additionally, there is weak correlation amongst the continuous variables with the `Claim_Amount`.
- Looking at the canonical correlation analysis (CCA), we observe that there is weak positive correlation (0.253) between policyholder attributes: `Gender` and `Age` and vehicle attributes : `Foreign`, `Vehicle_Age`, `Engine_Capacity`, and `Car_Value`. Therefore, we can conclude that there is little to no correlation between policyholder and vehicle attributes
- Based on the hypothesis test using Hotelling T^2 test, we observe that there is significant difference between mean log `claim amount` and log `car_value` between male and female based on the p-value at

5% significance level at a 5% significance level. Furthermore, we also observe significant difference on mean log `claim_amount` and `car_value` between local and foreign manufacturers at the 5% significance level. However, the assumption of equal covariance between the two populations (i.e. male vs female, local vs foreign) is not fulfilled and therefore a Hotelling test analysis would not be applicable.

- The current LDA model has a 72% accuracy by testing the prediction model with the actual test dataset and indicates whether a policyholder has a large claim or not based on these variables: `Female`,`Age`,`Engine_Capacity`,`Vehicle_Age`,`Car_Value`, and `Foreign`. Based on the results of the LDA prediction model, 691 cases of policyholders are correctly identified with small claims. However, a large proportion of cases (271 cases of policyholders) were incorrectly identified as small claims despite them being large claims in reality. If the insurance company applies this model, this will result in large number of policyholders being underpaid.
- Additionally, the histogram of the LDA model shows overlaps between the policyholders who have large and small claims. This overlaps suggests that the large number of outliers in the data as seen during the exploratory analysis may have affected the result of the prediction model.

Limitations

The high number of outliers in the dataset impacts the CCA and Hotelling T^2 test. We observed from the CCA there is weak correlation between the x and y variables potentially due to the high number of outliers and also impacts the assumption needed for Hotelling's T^2 to have equal sample covariance.

Additionally, high outliers also impact LDA as they can cause bias classification and decreased in classification accuracy. The high proportion of misclassification and overlapping between groups suggests by using the current LDA model (accuracy: 72%), the insurance company will be underpaying large proportion of policyholders who should actually receive higher amount of reimbursement. The insurance company should look into more variables such as the policyholder's past driving history to improve the accuracy of the model and look into creating more insurance premium options.