FINAL PROJECT

STAT 5311 (Multivariate)

Spring 2020

Contraceptive Method Choice in Indonesia (CMC)

Saran Caba

Gbemisola Adewuya

Afolake Baiyewu

Multivariate Analysis On Contraceptive Choice Of Married Women In Indonesia 1987.

AFolake

05/05/2020

DataSet Information:

- **Wife's age (numerical)**
- **Wife's education (categorical) 1=low, 2, 3, 4=high**
- **Husband's education (categorical) 1=low, 2, 3, 4=high**
- **Number of children ever born (numerical)**
- **Wife's religion (binary) 0=Non-Islam, 1=Islam**
- **Wife's now working? (binary) 0=Yes, 1=No**
- **Husband's occupation (categorical) 1, 2, 3, 4**
- **Standard-of-living index (categorical) 1=low, 2, 3, 4=high**
- **Media exposure (binary) 0=Good, 1=Not good**
- **Contraceptive method used (class attribute) 1=No-use, 2=Long-term, 3=Short-term**

INTRODUCTION, PROBLEM, AND PURPOSE:

The dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. It was created by Tjen-Sien Lim (limt@stat.wisc.edu). The dataset illustrates samples of married women who were either not pregnant or do not know if they were at the time of the interview. The aim of this project is to create a multivariate analysis that predicts the current contraceptive method choice (no use, long-term methods, or short-term methods) of a woman based on her demographic and socio-economic characteristics

DATA PREPARATION:

data structure:

```
## 'data.frame':    1473 obs. of  10 variables:
##  $ w_age       : int  24 45 43 42 36 19 38 21 27 45 ...
##  $ w_education : int  2 1 2 3 3 4 2 3 2 1 ...
##  $ h_education : int  3 3 3 2 3 4 3 3 3 1 ...
##  $ n_children  : int  3 10 7 9 8 0 6 1 3 8 ...
##  $ w_religion  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ w_working    : int  1 1 1 1 1 1 1 0 1 1 ...
##  $ h_occuopation: int  2 3 3 3 3 3 3 3 3 2 ...
##  $ S_l_Index    : int  3 4 4 3 2 3 2 2 4 2 ...
##  $ media        : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ cm_used      : int  1 1 1 1 1 1 1 1 1 1 ...
```

check for missing or bad data:

```
anyNA(cmcdata)
```

```
## [1] FALSE
```

A quick summary of the data:

```
summary(cmcdata) # the w_age and n_children variables appears to have larger vales
```

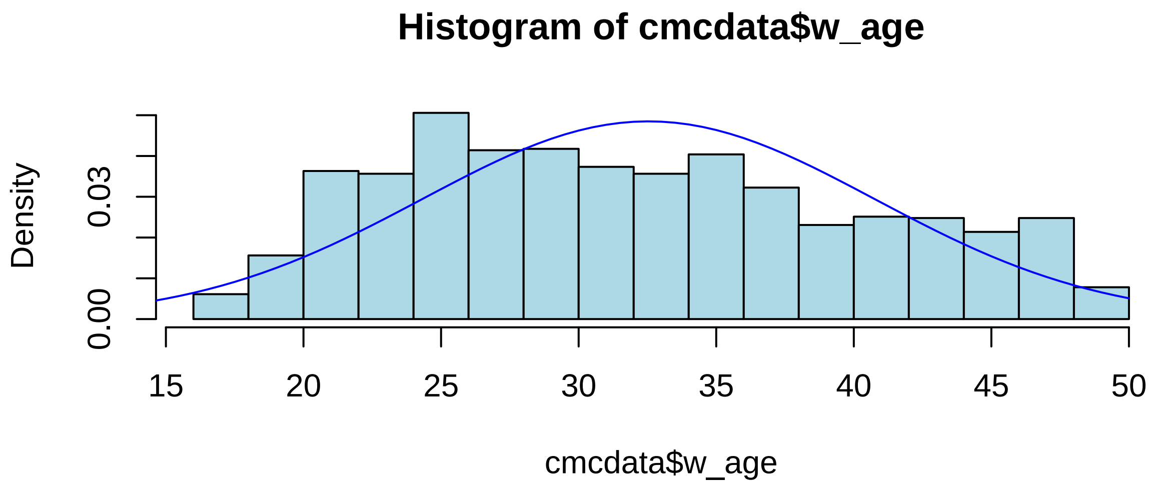
```
##      w_age         w_education     h_education     n_children    
##  Min.   :16.00   Min.   :1.000   Min.   :1.00   Min.   : 0.000  
##  1st Qu.:26.00   1st Qu.:2.000   1st Qu.:3.00   1st Qu.: 1.000  
##  Median :32.00   Median :3.000   Median :4.00   Median : 3.000  
##  Mean   :32.54   Mean   :2.959   Mean   :3.43   Mean   : 3.261  
##  3rd Qu.:39.00   3rd Qu.:4.000   3rd Qu.:4.00   3rd Qu.: 4.000  
##  Max.   :49.00   Max.   :4.000   Max.   :4.00   Max.   :16.000  
##    w_religion       w_working      h_occuopation     S_l_Index    
##  Min.   :0.0000   Min.   :0.0000   Min.   :1.000   Min.   :1.000  
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:3.000  
##  Median :1.0000   Median :1.0000   Median :2.000   Median :3.000  
##  Mean   :0.8506   Mean   :0.7495   Mean   :2.138   Mean   :3.134  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :4.000   Max.   :4.000  
##      media          cm_used    
##  Min.   :0.000   Min.   :1.00  
##  1st Qu.:0.000   1st Qu.:1.00  
##  Median :0.000   Median :2.00  
##  Mean   :0.074   Mean   :1.92  
##  3rd Qu.:0.000   3rd Qu.:3.00  
##  Max.   :1.000   Max.   :3.00
```

change the class of some variables with 2-4 unique values to categorical:

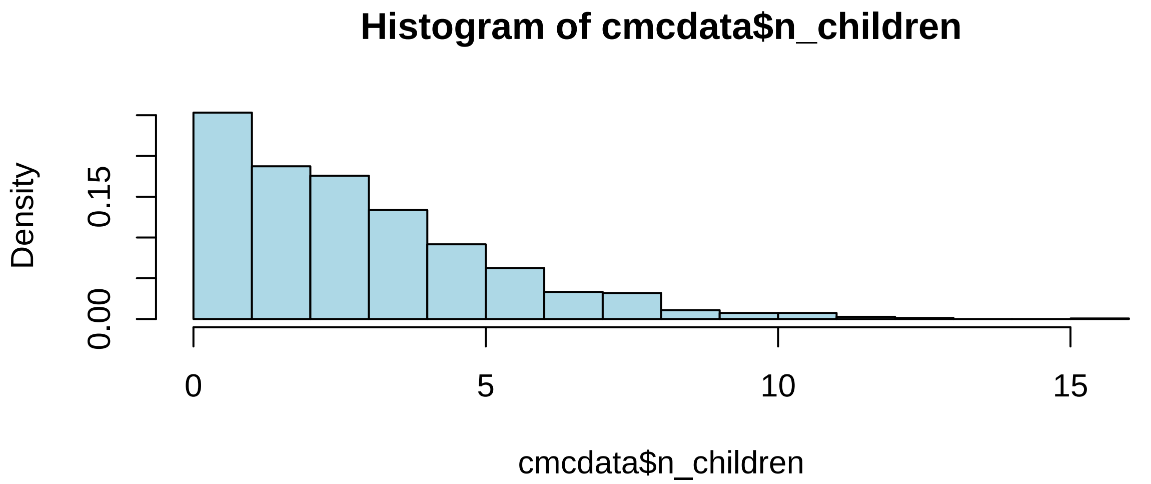
```
## 'data.frame':    1473 obs. of  10 variables:
##  $ w_age        : int  24 45 43 42 36 19 38 21 27 45 ...
##  $ w_education  : Factor w/ 4 levels "1","2","3","4": 2 1 2 3 3 4 2 3 2 1 ...
##  $ h_education  : Factor w/ 4 levels "1","2","3","4": 3 3 3 2 3 4 3 3 3 1 ...
##  $ n_children   : int  3 10 7 9 8 0 6 1 3 8 ...
##  $ w_religion   : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ w_working    : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 1 2 2 ...
##  $ h_occuopation: Factor w/ 4 levels "1","2","3","4": 2 3 3 3 3 3 3 3 3 2 ...
##  $ S_l_Index    : Factor w/ 4 levels "1","2","3","4": 3 4 4 3 2 3 2 2 4 2 ...
##  $ media        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
##  $ cm_used      : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
```

VISUALIZATION:

check distribution of numeric varaibles:

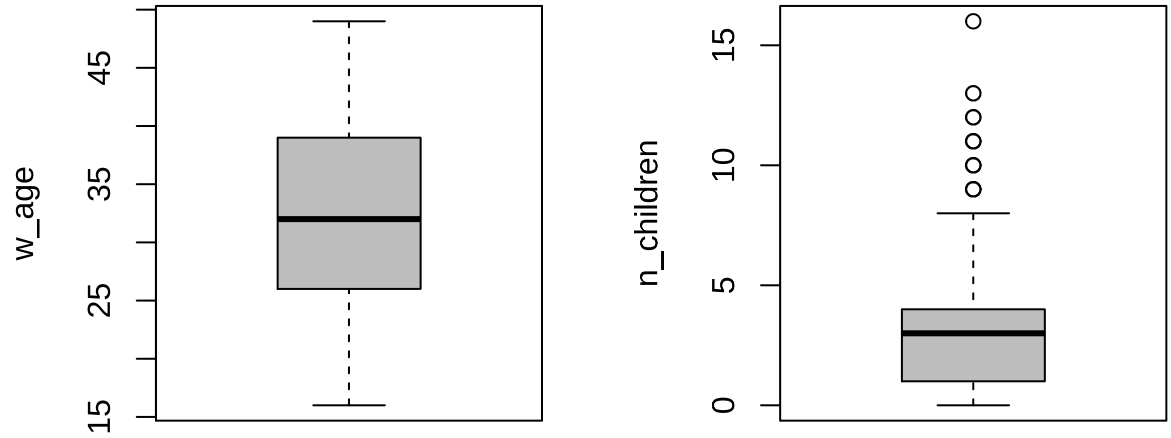


- the wife's age variable is obviouly not normally distributed (not bell like shape)



- The number of children variable is skewed to the right (we may need to tranform the variable before modeling)

check for outliers:

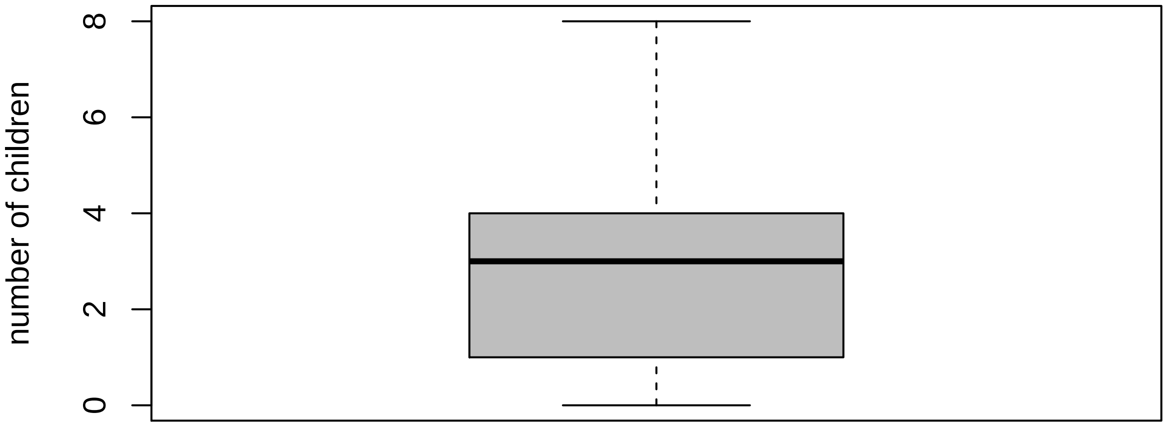


- from the boxplot above we can see that there are few outliers in the n_children variable and the values ranges from 9 to 16. it is valid to consider those values as outliers because the total fertility rate in indonesia as at 1987 was 3.5 so we don't expect values too far from that number.

DATA PRE-PROCESSING

remove the outlier values:

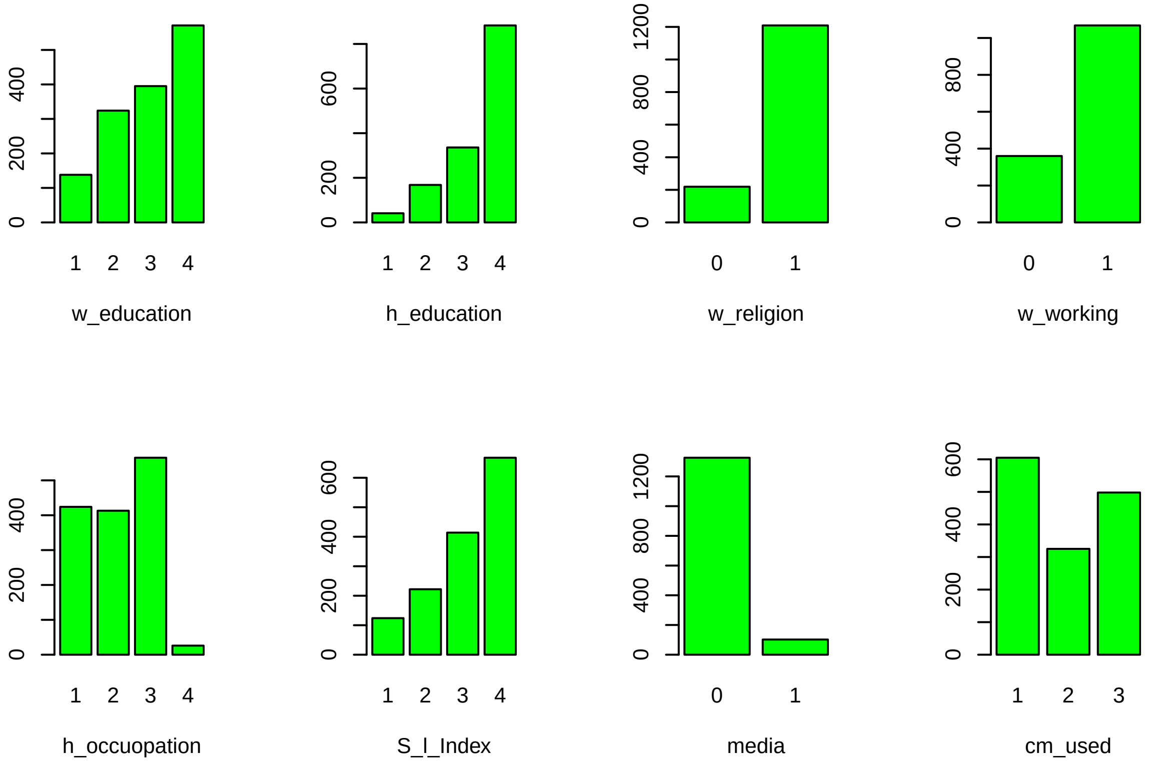
```
outliers <- unique(boxplot(cmcdata$n_children)$out)

outliers <- which(cmcdata$n_children %in% outliers)

cmcdata <- cmcdata[-outliers,]

boxplot(cmcdata$n_children, ylab = "number of children", col = "grey")
```



- we can verify that the outliers have been removed from the observations

barplot representation of the categorical variables:

```r
par(mfrow = c(2,4))

for (i in c(2,3,5,6,7,8,9,10)) {
  plot(cmcdata[,i], xlab = colnames(cmcdata)[i], col = "green")
}
```



```r
par(mfrow = c(1,1))
```

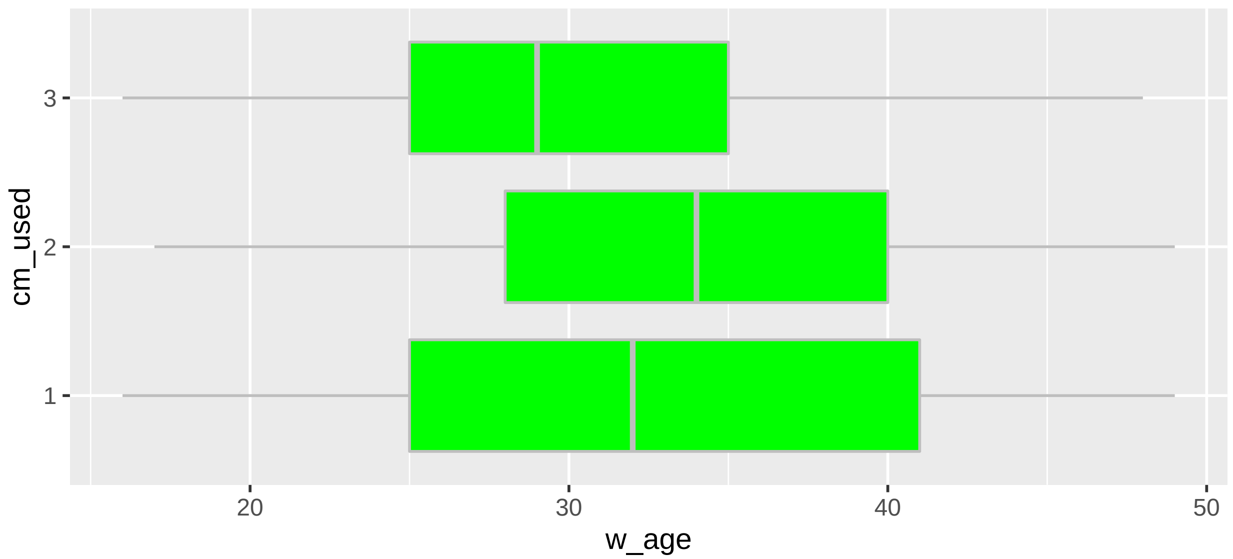
– The dependent variable appears to have an inbalanced class

HYPOTHESIS TESTING

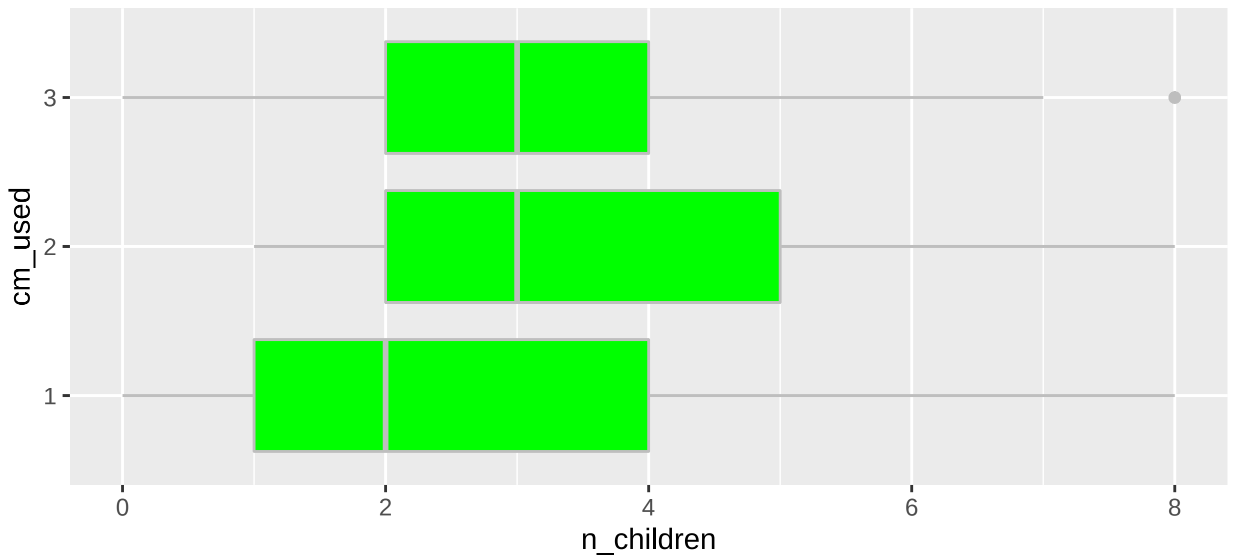
Now we want to investigate the relationship between the dependent variable and the inependent variables to know which ones would be better predictors for the model:

we can do a graphical representation of the variables to find relationships in them

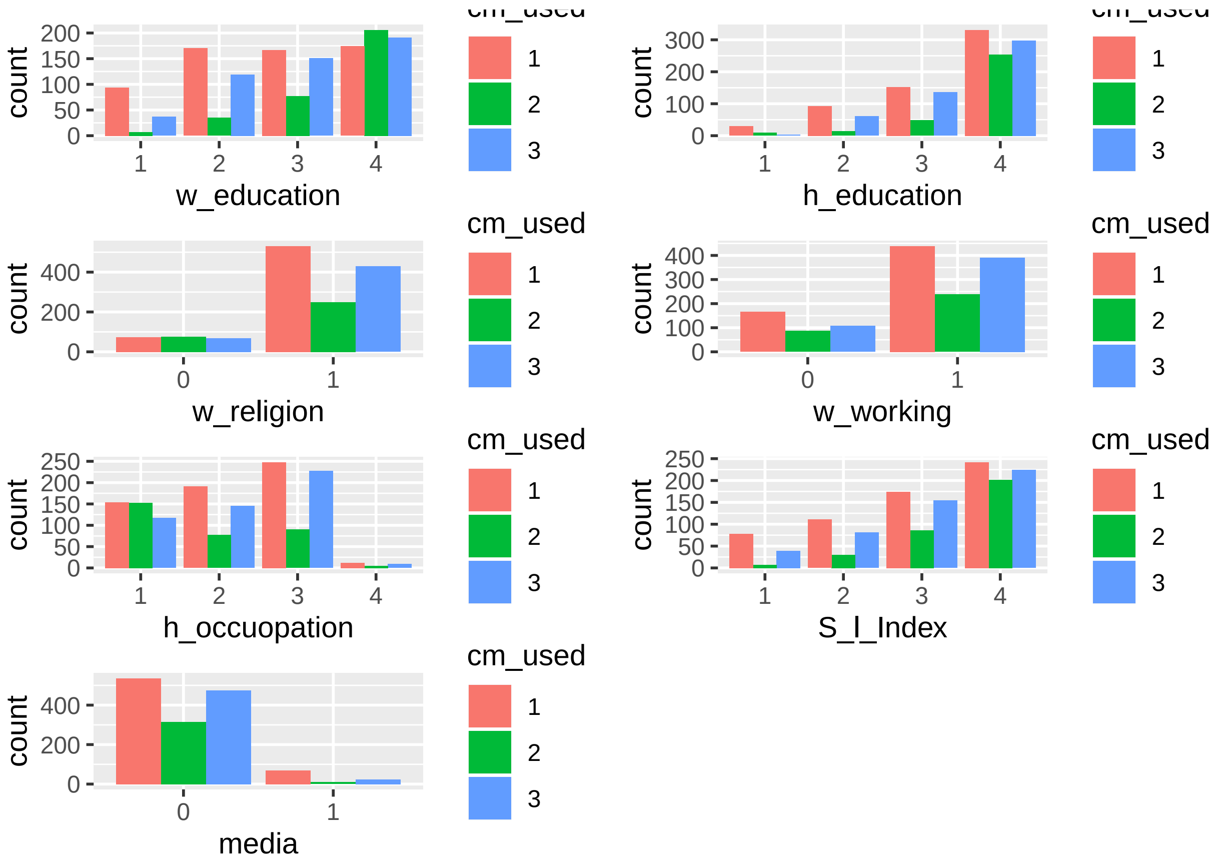
```r
library(tidyverse)

ggplot(cmcdata, aes(x = cm_used, y = w_age)) + geom_boxplot(fill = "green", col = "grey") +
  coord_flip()
```



```
ggplot(cmcdata, aes(x = cm_used, y = n_children)) + geom_boxplot(fill = "green", col = "grey") +
  coord_flip()
```



- women with more children tend to opt for the long-term contraceptive method.



- the plots above does not give us enough evidence on whether to reject or accept the null hypothesis that there is no relationship between the dependent variables and the independent variable so we will verify with some statistical tests.

statistical hypothesis testing:

w_education and cm_used:

- h0 = no relationship between w_education and cm_used
- h1 = there exist a relationship between w_education and cm_used
- alpha = 0.05%

```
chisq.test(table(cmcdata$cm_used, cmcdata$w_education))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(cmcdata$cm_used, cmcdata$w_education)
## X-squared = 138.64, df = 6, p-value < 2.2e-16
```

– with a p-value less than alpha we reject h0 and conclude that there exist a relationship

S_I_Index and cm_used:

- h0 = no relationship between S_I_Index and cm_used
- h1 = there exist a relationship between S_I_Index and cm_used
- alpha = 0.05%

```
chisq.test(table(cmcdata$cm_used, cmcdata$S_l_Index))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(cmcdata$cm_used, cmcdata$S_l_Index)
## X-squared = 64.355, df = 6, p-value = 5.841e-12
```

– with a p-value less than alpha we reject h0 and conclude that there exist a relationship

h_occupation and cm_used:

- h0 = no relationship between h_occupation and cm_used
- h1 = there exist a relationship between h_occupation and cm_used
- alpha = 0.05%

```
chisq.test(table(cmcdata$cm_used, cmcdata$h_occuopation))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(cmcdata$cm_used, cmcdata$h_occuopation)
## X-squared = 64.409, df = 6, p-value = 5.696e-12
```

– with a p-value of less than alpha we reject h0 and conclude that there exist a relationship

h_education and cm_used:

- h0 = no relationship between h_education and cm_used
- h1 = there exist a relationship between h_education and cm_used
- alpha = 0.05%

```
chisq.test(table(cmcdata$cm_used, cmcdata$h_education))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(cmcdata$cm_used, cmcdata$h_education)
## X-squared = 72.439, df = 6, p-value = 1.291e-13
```

– with a p-value less than alpha we reject h0 and conclude that there exist a relationship

w_working and cm_used:

- h0 = no relationship between w_working and cm_used
- h1 = there exist a relationship between w_working and cm_used
- alpha = 0.05%

```r
chisq.test(table(cmcdata$cm_used, cmcdata$w_working))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(cmcdata$cm_used, cmcdata$w_working)
## X-squared = 5.675, df = 2, p-value = 0.05857
```

– with a p-value greater than alpha we fail to reject h0 and conclude that there exist no relationship

w_religion and cm_used:

- h0 = no relationship between w_religion and cm_used
- h1 = there exist a relationship between w_religion and cm_used
- alpha = 0.05%

```r
chisq.test(table(cmcdata$cm_used, cmcdata$w_religion))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(cmcdata$cm_used, cmcdata$w_religion)
## X-squared = 21.547, df = 2, p-value = 2.095e-05
```

– with a p-value less than alpha we reject h0 and conclude that there exist a relationship

media and cm_used:

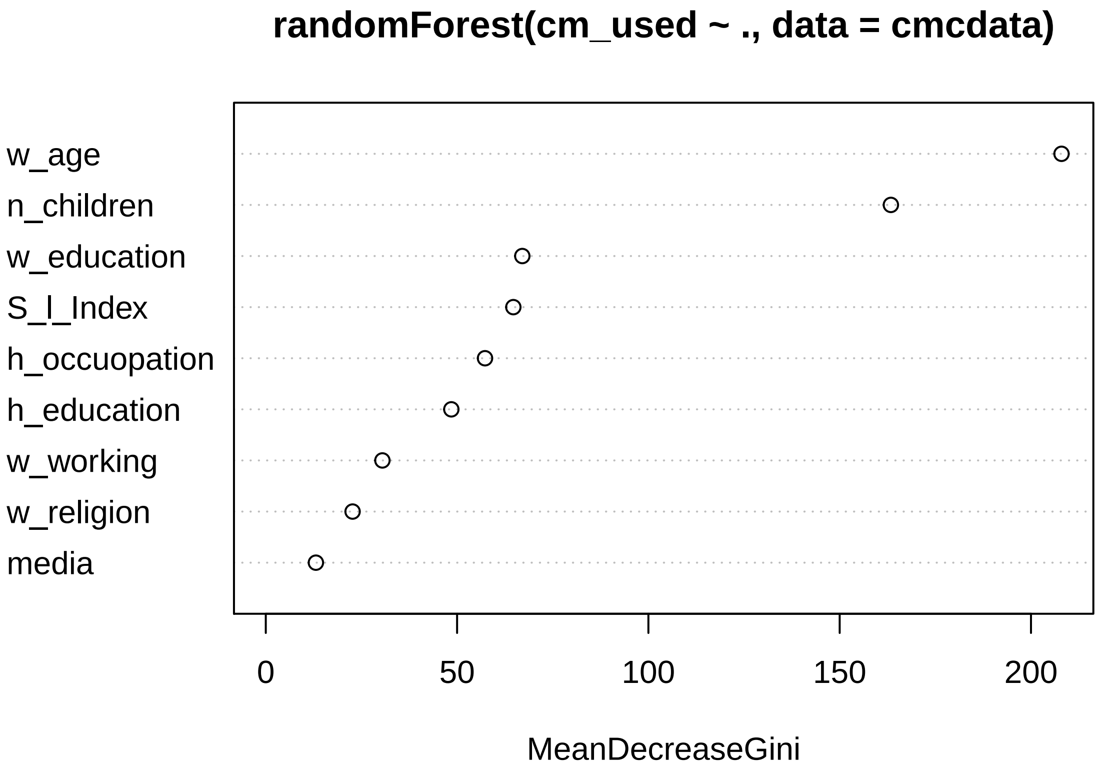
- h0 = no relationship between media and cm_used
- h1 = there exist a relationship between media and cm_used
- alpha = 0.05%

```r
chisq.test(table(cmcdata$cm_used, cmcdata$media))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(cmcdata$cm_used, cmcdata$media)
## X-squared = 29.455, df = 2, p-value = 4.017e-07
```

– with a p-value less than alpha we reject h0 and conclude that there exist a relationship

we can also verify which variables will be better predictors in the model through a variable importance plot from the randomforest algorithm:



- The plot above shows that w__working, media and w__religion have the least impact/relationship with the cm__used variable while w__age and n__children have the highest predictive power followed by education and s__l__index

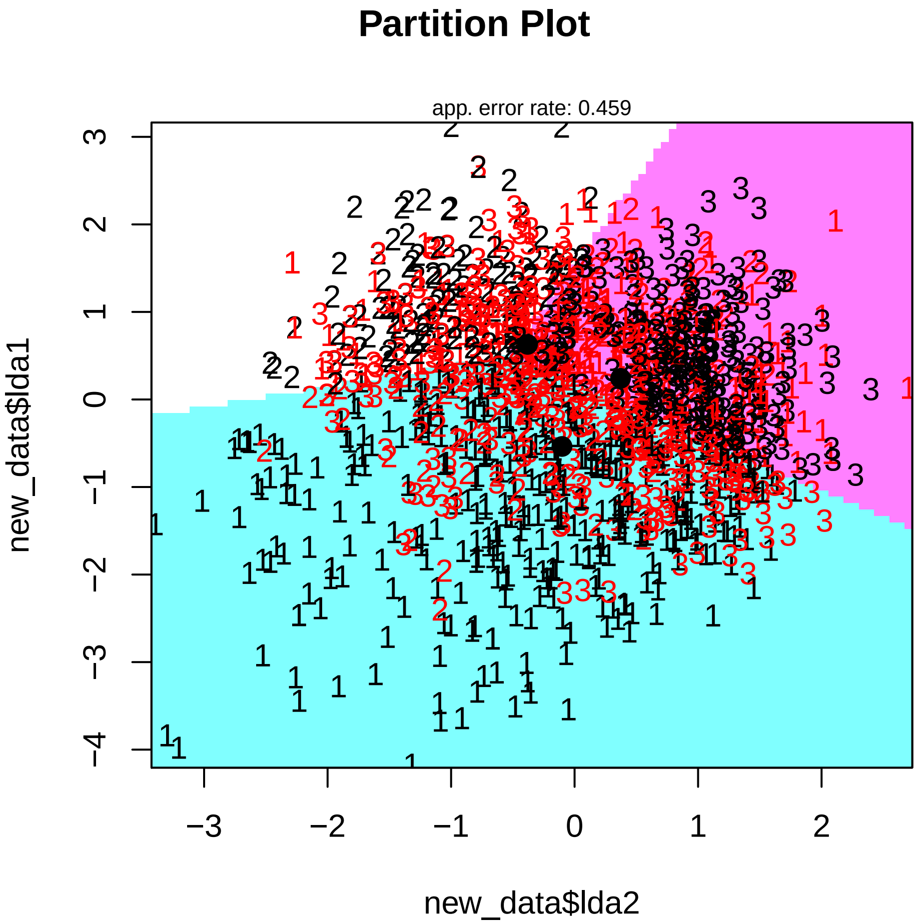
MODELING:

Building a classification model with linear discriminant analysis algorithm:

```
model <- MASS::lda(cm_used~., data = cmcdata)

pred <- predict(model, cmcdata)
```

Visualizing the result of the linear discriminant analysis:

```
  #  install.packages("klaR")
new_data <- data.frame(cm_used = cmcdata$cm_used, lda1 = pred$x[,1], lda2 = pred$x[,2])
klaR::partimat(new_data$cm_used~ new_data$lda1 + new_data$lda2, method = "lda")
```



RESULT AND DISCUSSION

```
##          actual
## predicted   1   2   3
##         1 386  81 156
##         2  55 134  89
##         3 164 110 253
```

```
## [1] "Accuracy: 54.13%"
```

– due to the inbalance nature of the response variable (contraceptive method used), the model was better at predicting the 1st class (No-use) than the other two classes (Long-term,Short-term)

CONCLUSION

The lda model has many misclassification we may need to try out more complex algorithm like neural network to improve performance

APPENDICES

Data Source:

- Origin: This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey
- Creator: Tjen-Sien Lim (limt '@' stat.wisc.edu)
- Donor: Tjen-Sien Lim (limt '@' stat.wisc.edu)
- Download: link

REFERENCES

While carrying out this analysis i found the following websites very helpful:

- http://stackoverflow.com
- http://kaggle.com
- http://medium.com