Thesis: Exploratory Data Analysis

Laurens van der Tas

Initialize R Environment & Load Data

This section will run the commands necessary to initialize R and load up our packages and data.

```
#Set working directory
#setwd("C:/Users/Laurens/Dropbox/University/Year 4/Period 2/Applied Economics Research Course/Thesis/da

#Load required packages
library(foreign)
library(stargazer)
library(ggplot2)
library(aod)
library(gridExtra)
library(ggthemes)
library(dplyr)
library(mfx)
library(corrplot)

#Enable anti-aliasing for rendered graphics
library(knitr)
#opts_chunk$set(out.width = '\\maxwidth')
#dev = "CairoPNG",

#Load dataset
data.dropout <- read.dta("DatasetTrimmed.dta")
  #Read name vector of dataset
  names(data.dropout)
```

Descriptive Statistics

Here, the two dichotomous control variables are factorized.

```
#Factorize binaries
data.dropout$geslachtBin <- factor(data.dropout$geslachtBin, labels = c("Female", "Male"))
data.dropout$allochtoonBin <- factor(data.dropout$allochtoonBin, labels = c("No", "Yes"))
```

Boxplots

The following section will create boxplots for the *Big Five* and *Gender* with numerical summary statistics below each plot to clarify the visualisation.

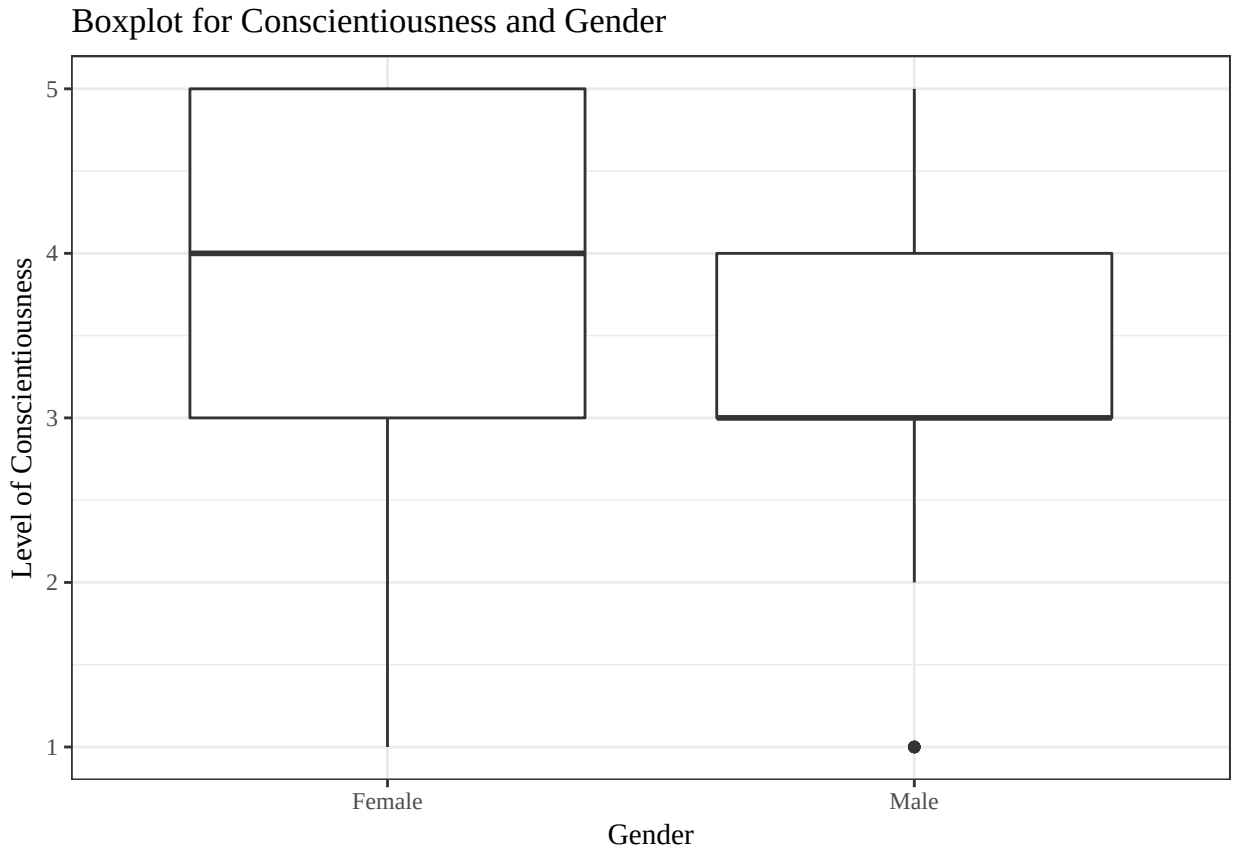
```
#Determine amounts of males and females in the sample
length(subset(data.dropout, geslachtBin == "Male")$geslachtBin)
```

```
## [1] 207
```

```
length(subset(data.dropout, geslachtBin == "Female")$geslachtBin)
```

```
## [1] 292
```

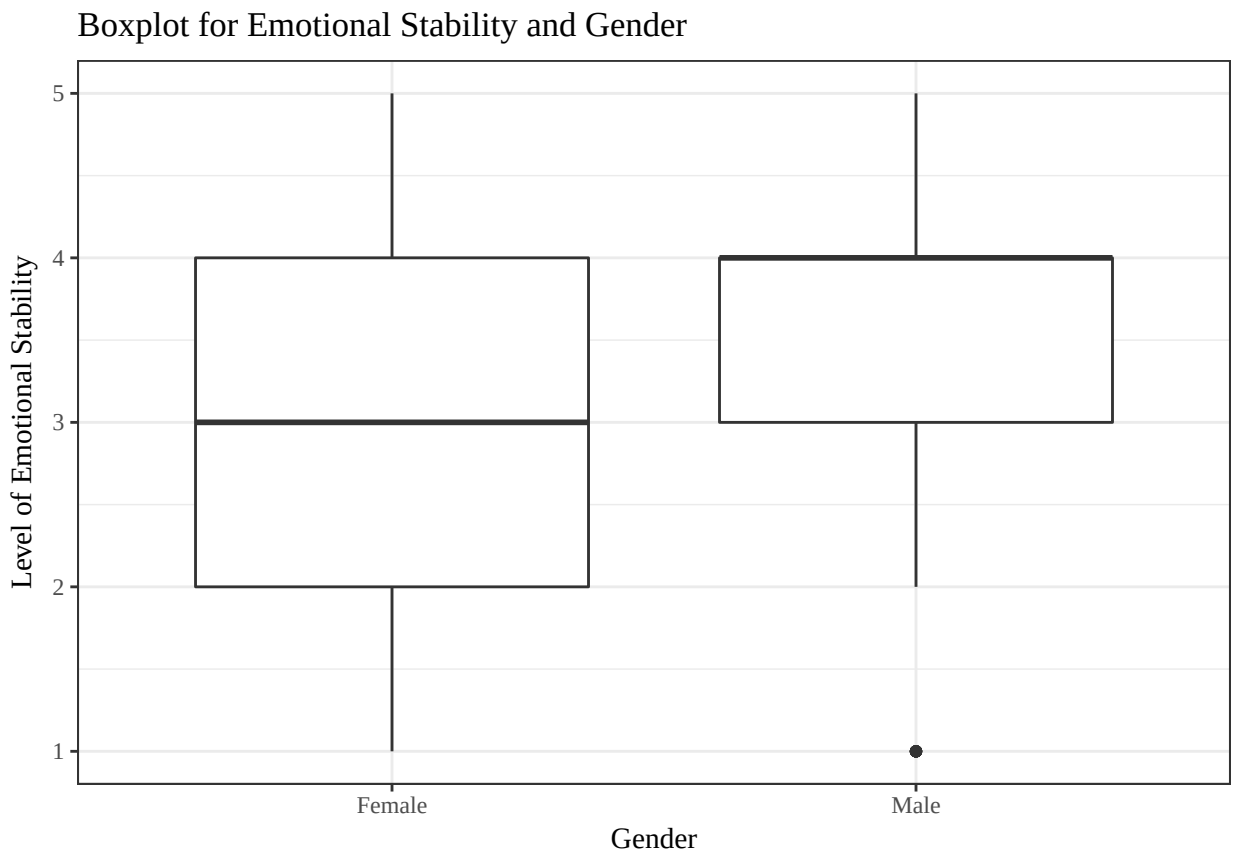
```r
#Boxplots for Big Five and Gender

#Conscientiousness
zorg <- qplot(
  x = geslachtBin,
  y = zorgvuldig,
  main = "Boxplot for Conscientiousness and Gender",
  xlab = "Gender",
  ylab = "Level of Conscientiousness",
  data = data.dropout,
  geom = "boxplot") +
    theme_bw(base_family = "serif") +
    theme(axis.title.y = element_text(vjust = 1.0)) +
    theme(axis.title.x = element_text(vjust = -0.5))

zorg
```




```r
by(data.dropout$zorgvuldig, data.dropout$geslachtBin, summary)
```

```
## data.dropout$geslachtBin: Female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   4.000   3.606   5.000   5.000
## ------------------------------------------------------------
## data.dropout$geslachtBin: Male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.000   3.227   4.000   5.000
```

```r
#Emotional Stability
stab <- qplot(
  x = geslachtBin,
  y = stabiel,
  main = "Boxplot for Emotional Stability and Gender",
  xlab = "Gender",
  ylab = "Level of Emotional Stability",
  data = data.dropout,
  geom = "boxplot") +
    theme_bw(base_family = "serif") +
    theme(axis.title.y = element_text(vjust = 1.0)) +
    theme(axis.title.x = element_text(vjust = -0.5))

stab
```

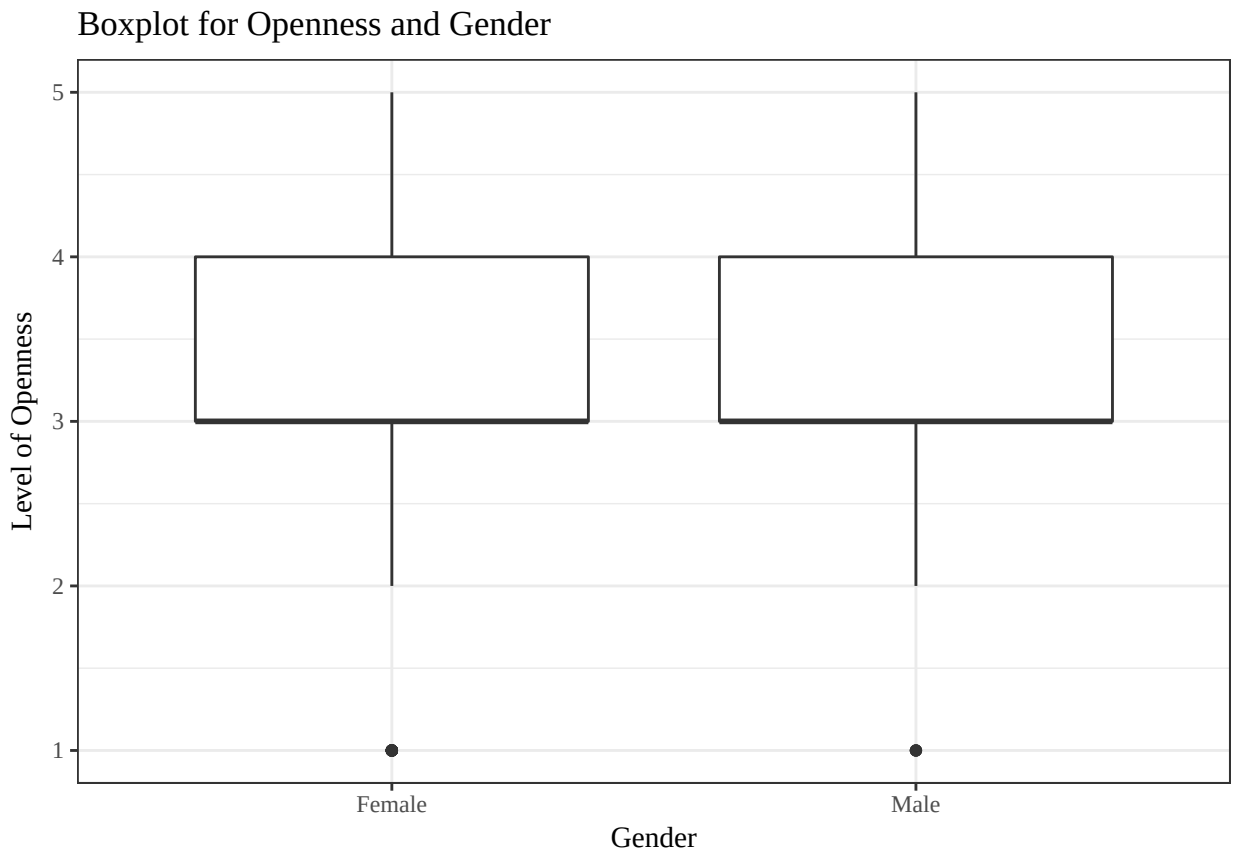



```r
by(data.dropout$stabiel, data.dropout$geslachtBin, summary)
```

```
## data.dropout$geslachtBin: Female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   3.147   4.000   5.000
## ------------------------------------------------------------
## data.dropout$geslachtBin: Male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   4.000   3.551   4.000   5.000
```

```r
#Openness
open <- qplot(
```

```
  x = geslachtBin,
  y = open,
  main = "Boxplot for Openness and Gender",
  xlab = "Gender",
  ylab = "Level of Openness",
  data = data.dropout,
  geom = "boxplot") +
    theme_bw(base_family = "serif") +
    theme(axis.title.y = element_text(vjust = 1.0)) +
    theme(axis.title.x = element_text(vjust = -0.5))

open
```

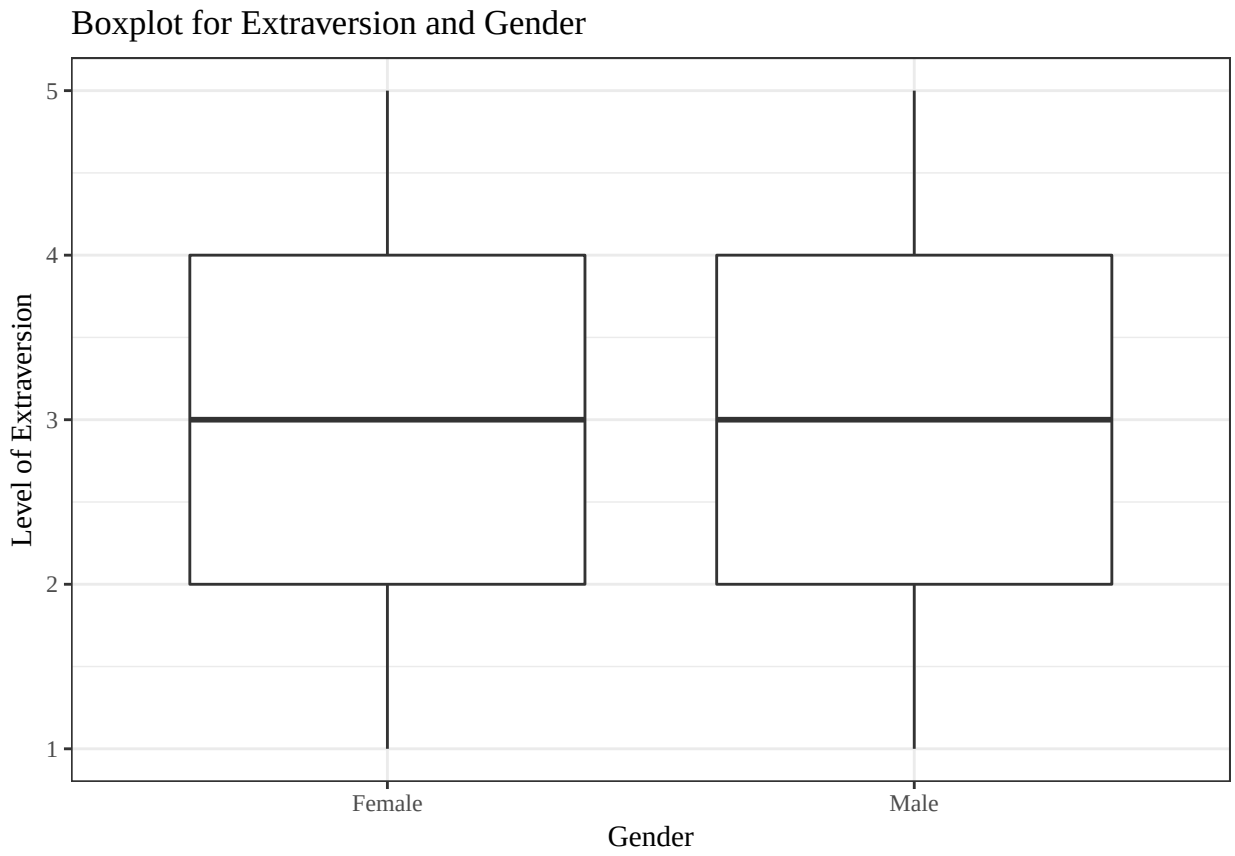



```
by(data.dropout$open, data.dropout$geslachtBin, summary)
```

```
## data.dropout$geslachtBin: Female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.000   3.209   4.000   5.000
## ------------------------------------------------------------
## data.dropout$geslachtBin: Male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.000   3.382   4.000   5.000
```

```
#Extraversion
extr <- qplot(
  x = geslachtBin,
  y = extravert,
```

```
  main = "Boxplot for Extraversion and Gender",
  xlab = "Gender",
  ylab = "Level of Extraversion",
  data = data.dropout,
  geom = "boxplot") +
    theme_bw(base_family = "serif") +
    theme(axis.title.y = element_text(vjust = 1.0)) +
    theme(axis.title.x = element_text(vjust = -0.5))

extr
```

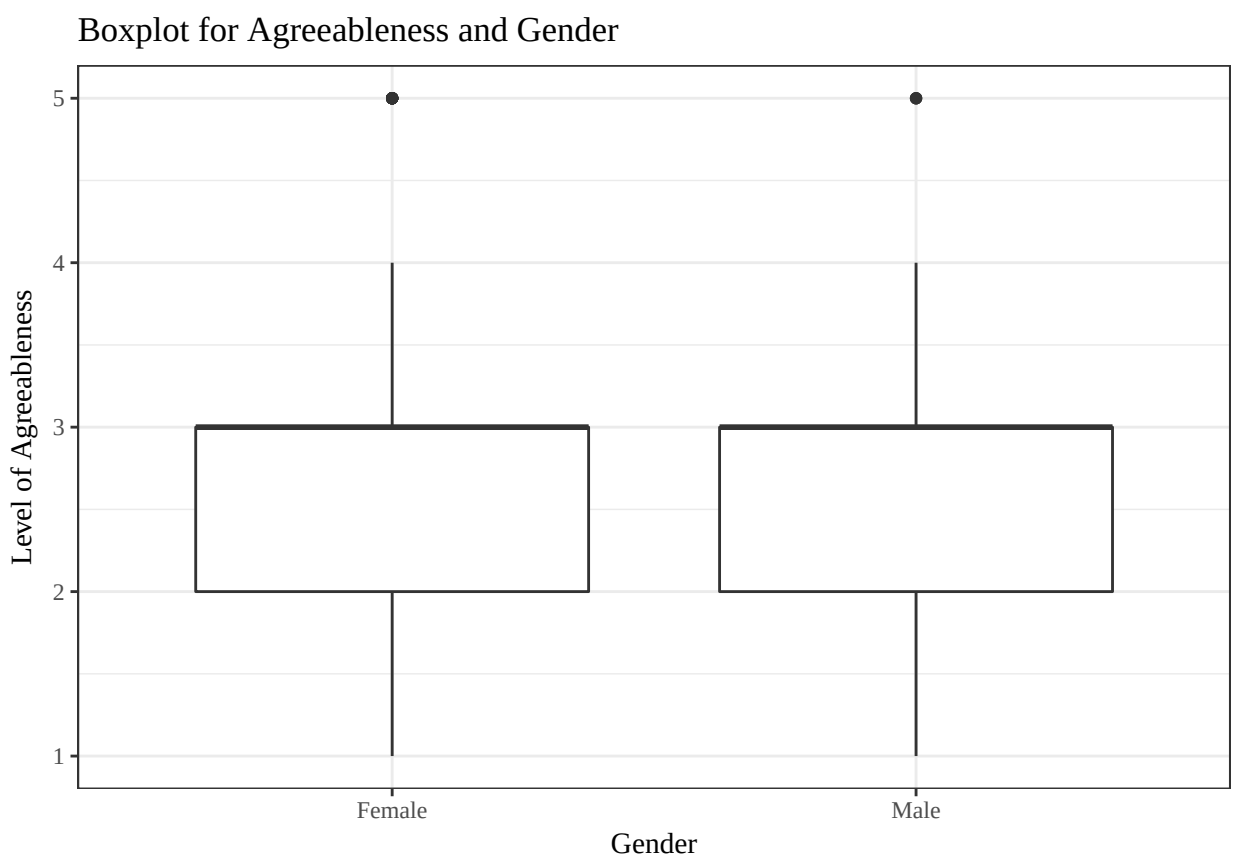



```
by(data.dropout$extravert, data.dropout$geslachtBin, summary)
```

```
## data.dropout$geslachtBin: Female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   3.127   4.000   5.000
## ------------------------------------------------------------
## data.dropout$geslachtBin: Male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   3.111   4.000   5.000
```

```
#Agreeableness
altr <- qplot(
  x = geslachtBin,
  y = altrusme,
  main = "Boxplot for Agreeableness and Gender",
  xlab = "Gender",
```

```
  ylab = "Level of Agreeableness",
  data = data.dropout,
  geom = "boxplot") +
    theme_bw(base_family = "serif") +
    theme(axis.title.y = element_text(vjust = 1.0)) +
    theme(axis.title.x = element_text(vjust = -0.5))

altr
```




```
by(data.dropout$altrusme, data.dropout$geslachtBin, summary)
```

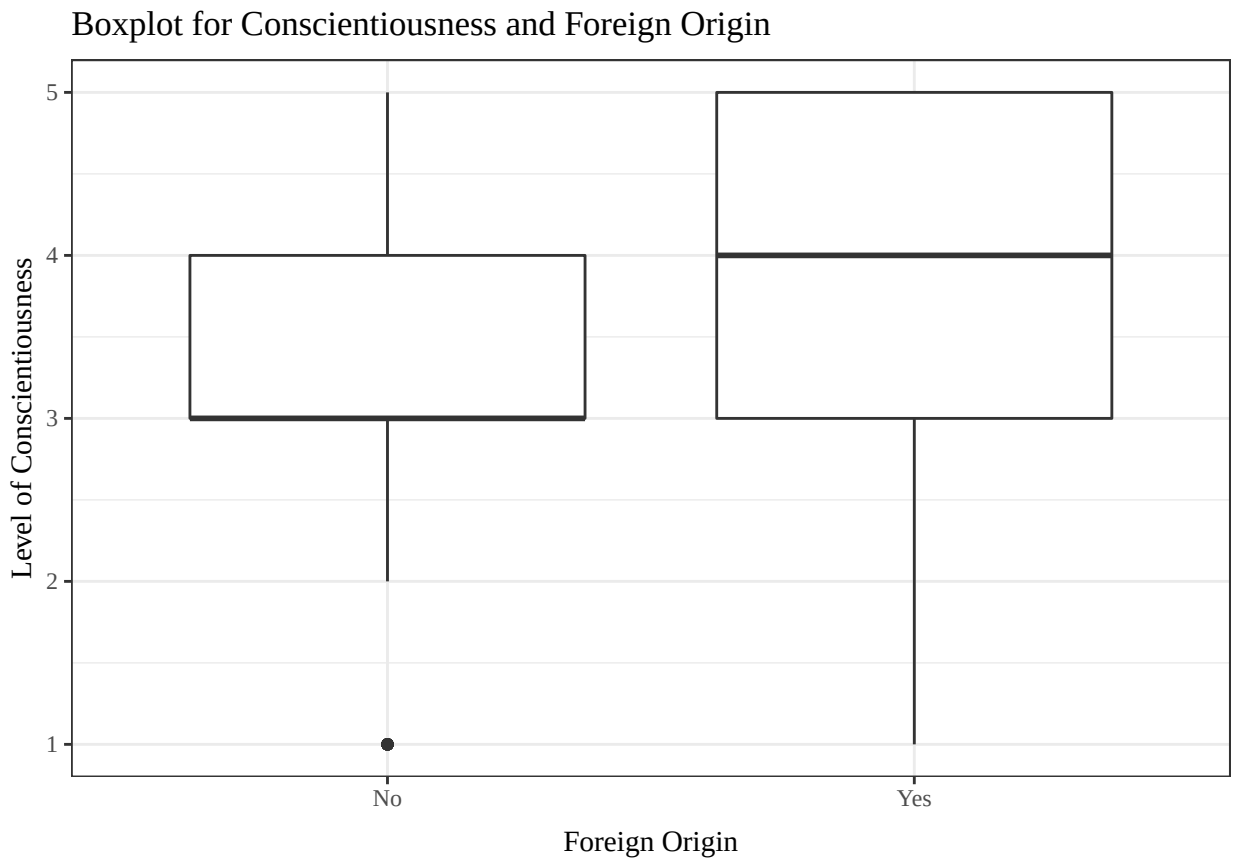
```
## data.dropout$geslachtBin: Female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.938   3.000   5.000
## ------------------------------------------------------------
## data.dropout$geslachtBin: Male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.879   3.000   5.000
```

```
#Uncomment to also group the plots in one window
#grid.arrange(altr, extr, open, stab, zorg, ncol= 3)
```

Next: Boxplots for the *Big Five* and *Foreign Origin* with numerical summaries below each plot.

```
#Boxplots for Big Five and foreign origin

#Conscientiousness
zorg.1 <- qplot(
```

```
  x = allochtoonBin,
  y = zorgvuldig,
  main = "Boxplot for Conscientiousness and Foreign Origin",
  xlab = "Foreign Origin",
  ylab = "Level of Conscientiousness",
  data = data.dropout,
  geom = "boxplot") +
  theme_bw(base_family = "serif") +
    theme(axis.title.y = element_text(vjust = 1.0)) +
    theme(axis.title.x = element_text(vjust = -0.5))

zorg.1
```

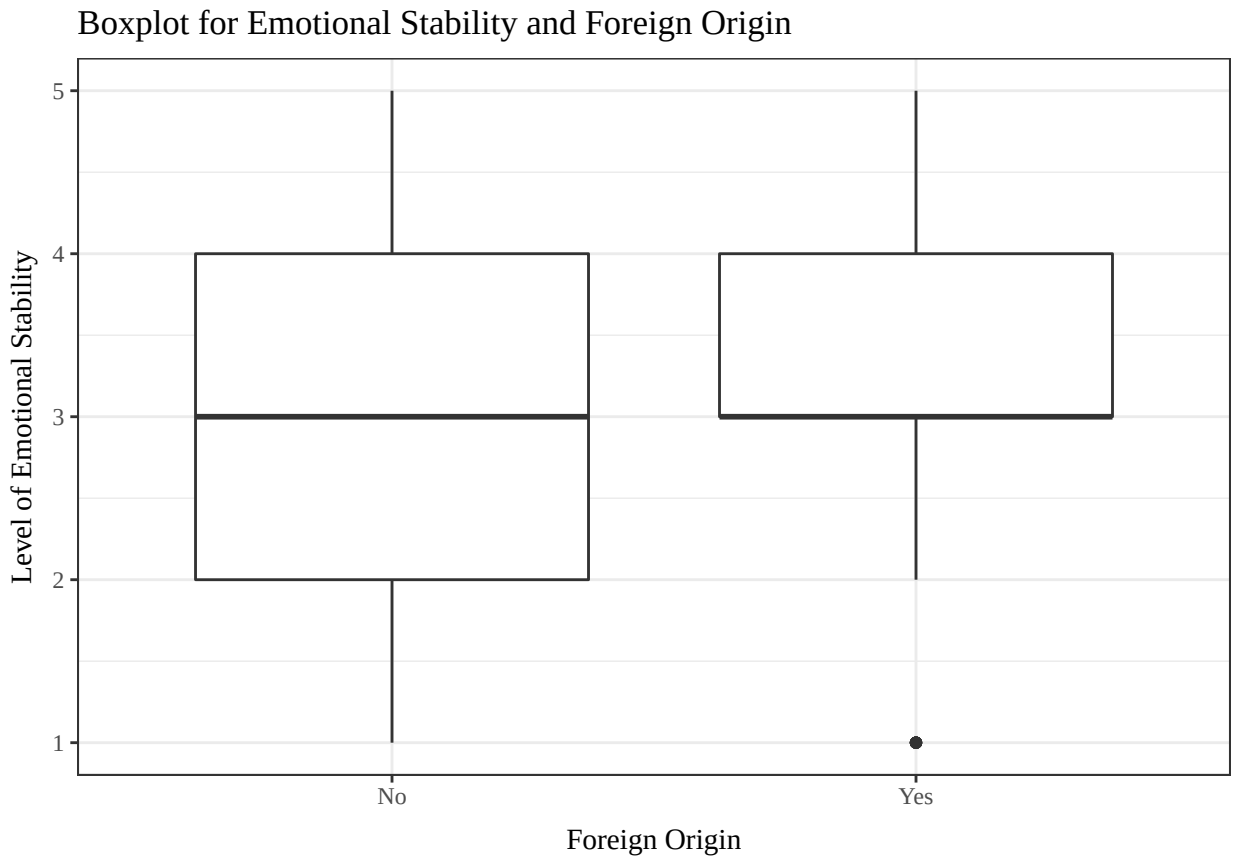



```
by(data.dropout$zorgvuldig, data.dropout$allochtoonBin, summary)
```

```
## data.dropout$allochtoonBin: No
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.000   3.367   4.000   5.000
## ------------------------------------------------------------
## data.dropout$allochtoonBin: Yes
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   4.000   3.699   5.000   5.000
```

```
#Emotional Stability
stab.1 <- qplot(
  x = allochtoonBin,
  y = stabiel,
```

```
  main = "Boxplot for Emotional Stability and Foreign Origin",
  xlab = "Foreign Origin",
  ylab = "Level of Emotional Stability",
  data = data.dropout,
  geom = "boxplot") +
  theme_bw(base_family = "serif") +
    theme(axis.title.y = element_text(vjust = 1.0)) +
    theme(axis.title.x = element_text(vjust = -0.5))

stab.1
```

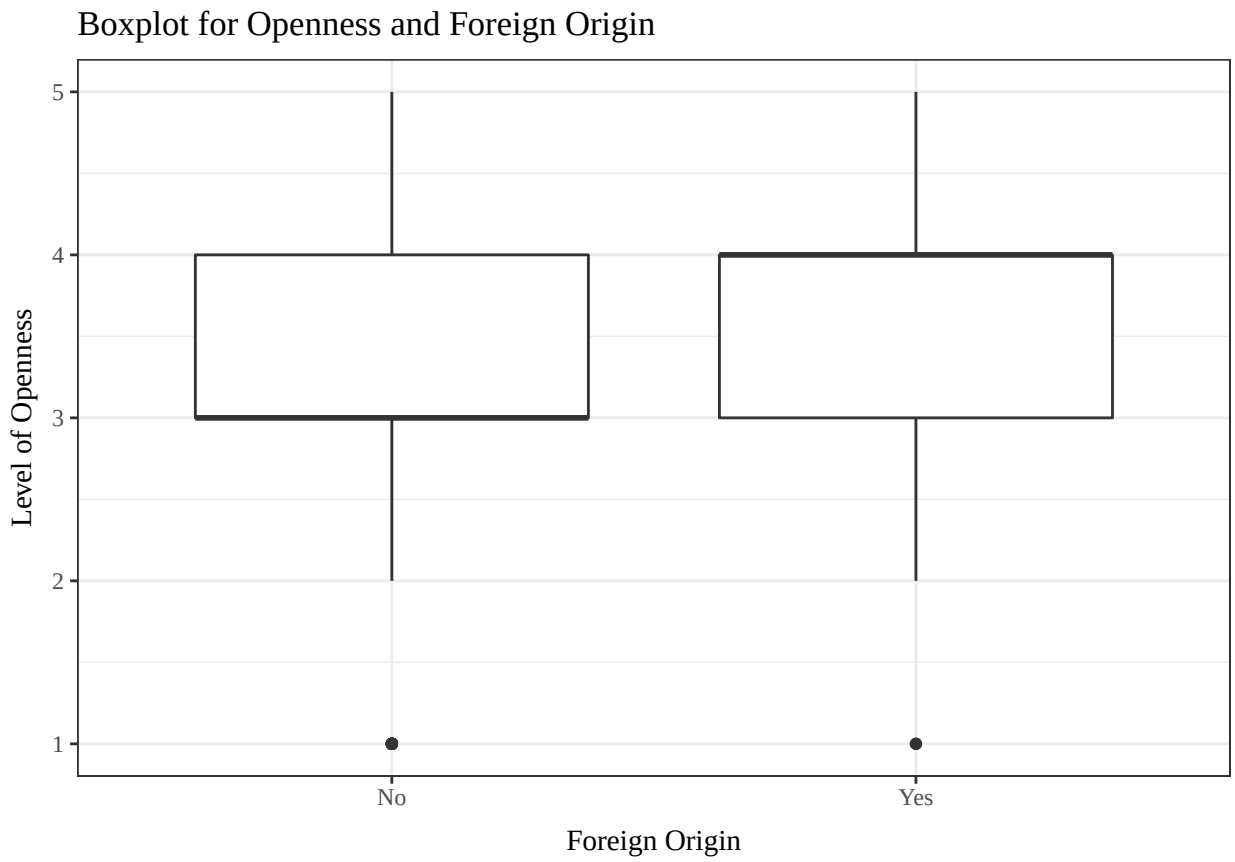



```
by(data.dropout$stabiel, data.dropout$allochtoonBin, summary)
```

```
## data.dropout$allochtoonBin: No
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   3.303   4.000   5.000
## ------------------------------------------------------------
## data.dropout$allochtoonBin: Yes
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.00    3.00    3.35    4.00    5.00
```

```
#Openness
open.1 <- qplot(
  x = allochtoonBin,
  y = open,
  main = "Boxplot for Openness and Foreign Origin",
  xlab = "Foreign Origin",
```

```
  ylab = "Level of Openness",
  data = data.dropout,
  geom = "boxplot") +
  theme_bw(base_family = "serif") +
    theme(axis.title.y = element_text(vjust = 1.0)) +
    theme(axis.title.x = element_text(vjust = -0.5))

open.1
```

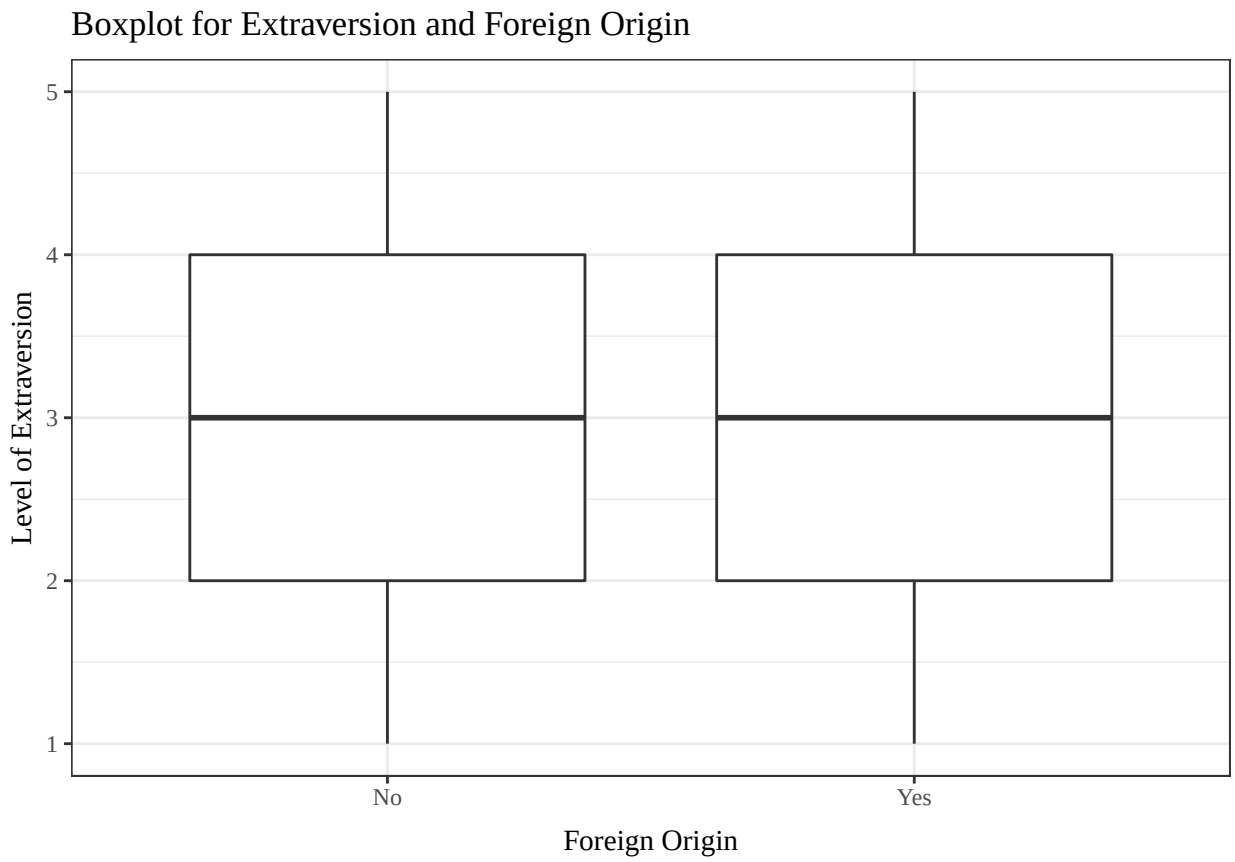



```
by(data.dropout$open, data.dropout$allochtoonBin, summary)
```

```
## data.dropout$allochtoonBin: No
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.000   3.186   4.000   5.000
## ------------------------------------------------------------
## data.dropout$allochtoonBin: Yes
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   4.000   3.569   4.000   5.000
```

```
#Extraversion
extr.1 <- qplot(
  x = allochtoonBin,
  y = extravert,
  main = "Boxplot for Extraversion and Foreign Origin",
  xlab = "Foreign Origin",
  ylab = "Level of Extraversion",
  data = data.dropout,
```

```
  geom = "boxplot") +
  theme_bw(base_family = "serif") +
    theme(axis.title.y = element_text(vjust = 1.0)) +
    theme(axis.title.x = element_text(vjust = -0.5))

extr.1
```

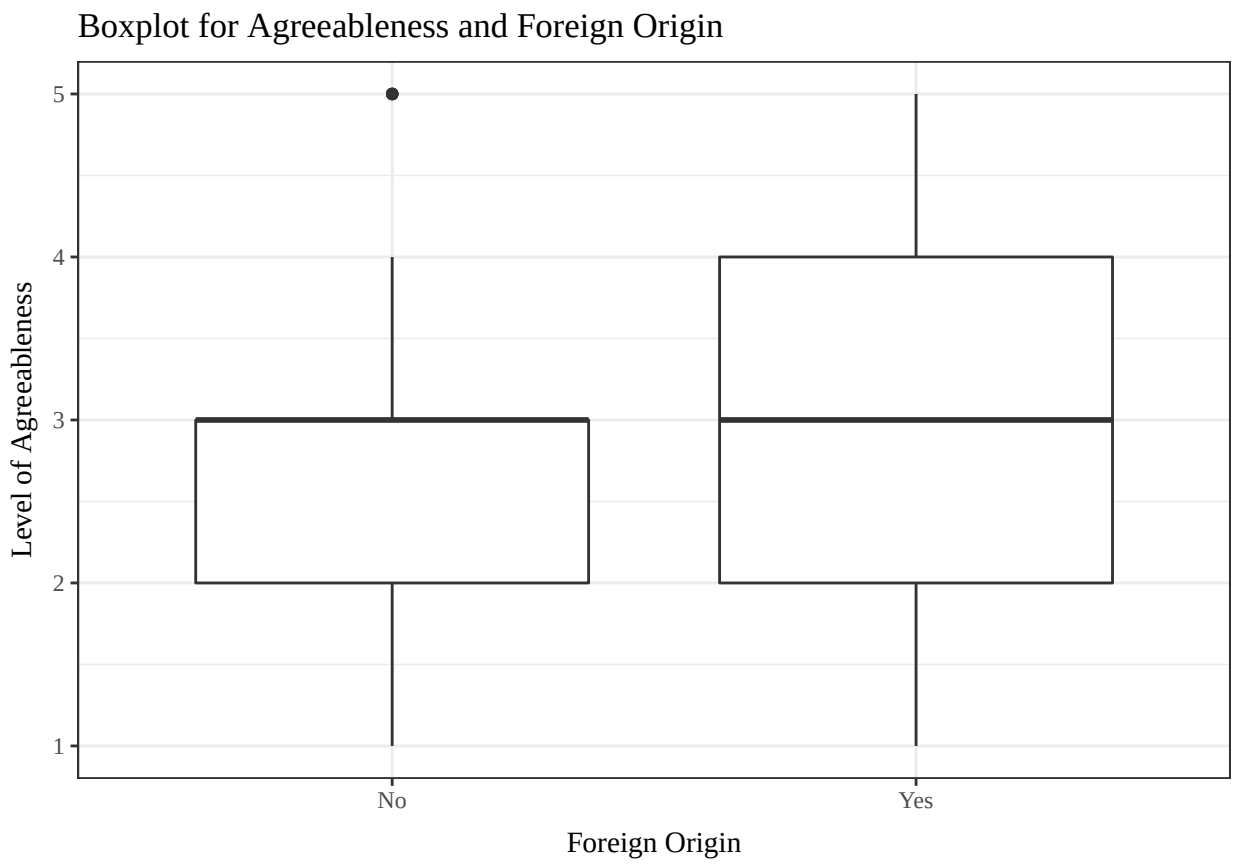



```
by(data.dropout$extravert, data.dropout$allochtoonBin, summary)
```

```
## data.dropout$allochtoonBin: No
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   3.104   4.000   5.000
## ------------------------------------------------------------
## data.dropout$allochtoonBin: Yes
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   3.171   4.000   5.000
```

```
#Agreeableness
altr.1 <- qplot(
  x = allochtoonBin,
  y = altrusme,
  main = "Boxplot for Agreeableness and Foreign Origin",
  xlab = "Foreign Origin",
  ylab = "Level of Agreeableness",
  data = data.dropout,
  geom = "boxplot") +
  theme_bw(base_family = "serif") +
```

```
    theme(axis.title.y = element_text(vjust = 1.0)) +
    theme(axis.title.x = element_text(vjust = -0.5))

altr.1
```




```
by(data.dropout$altrusme, data.dropout$allochtoonBin, summary)
```

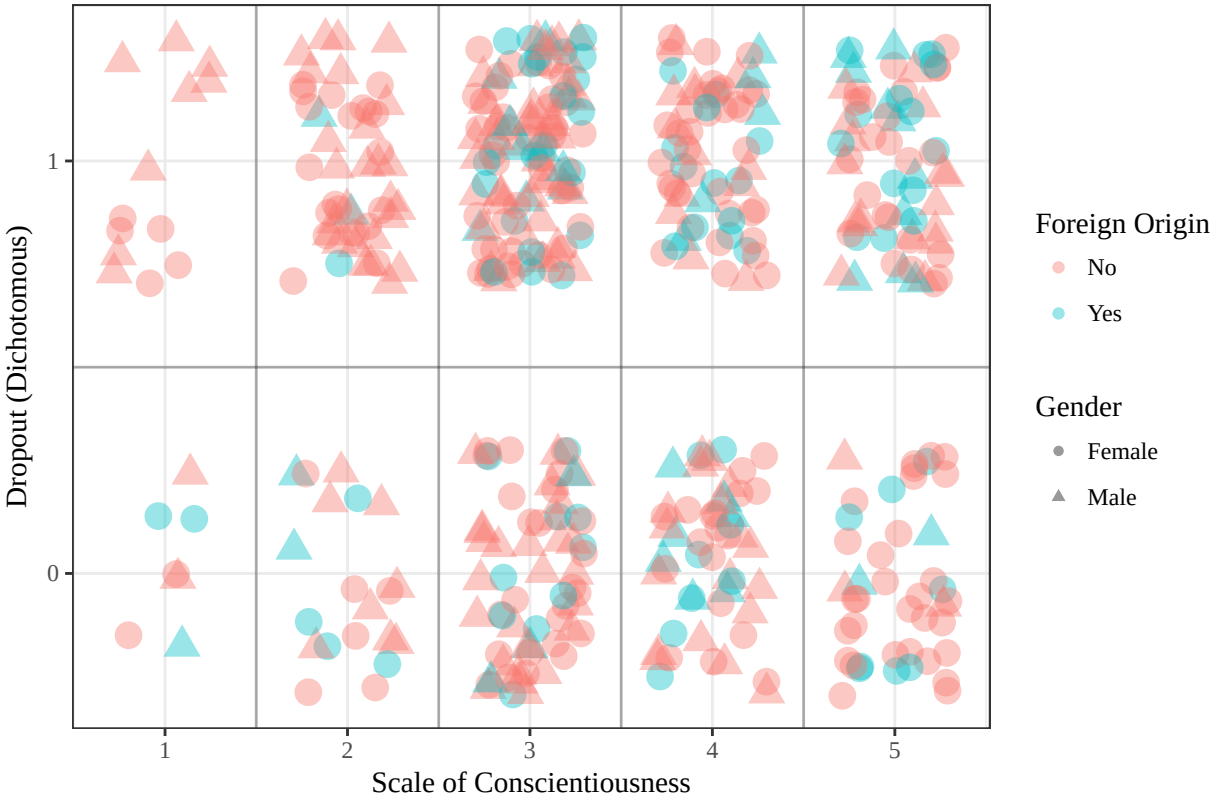
```
## data.dropout$allochtoonBin: No
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.891   3.000   5.000
## ------------------------------------------------------------
## data.dropout$allochtoonBin: Yes
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.984   4.000   5.000
```

```
#Uncomment to also group all plots in one window
#grid.arrange(arrangeGrob(zorg.1, stab.1, open.1, extr.1, altr.1, ncol = 3))
```

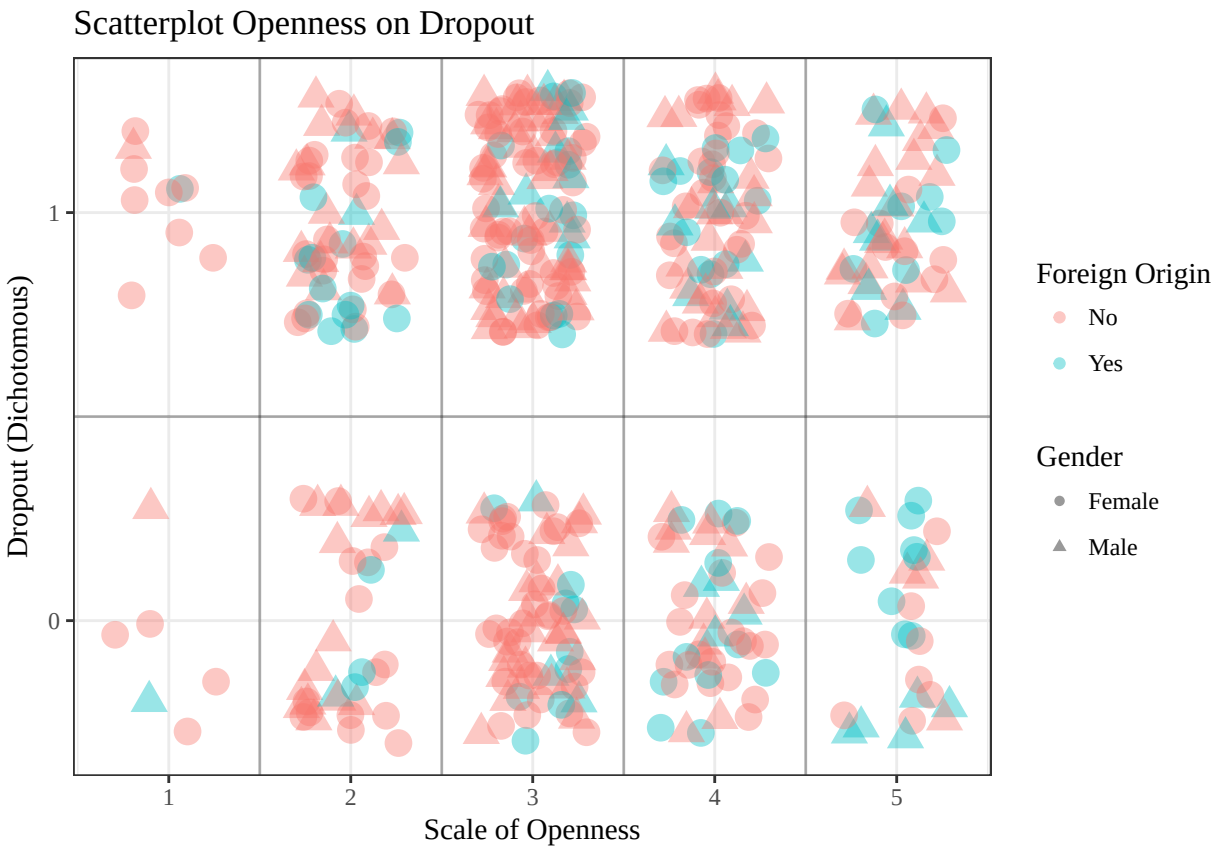
Scatterplots

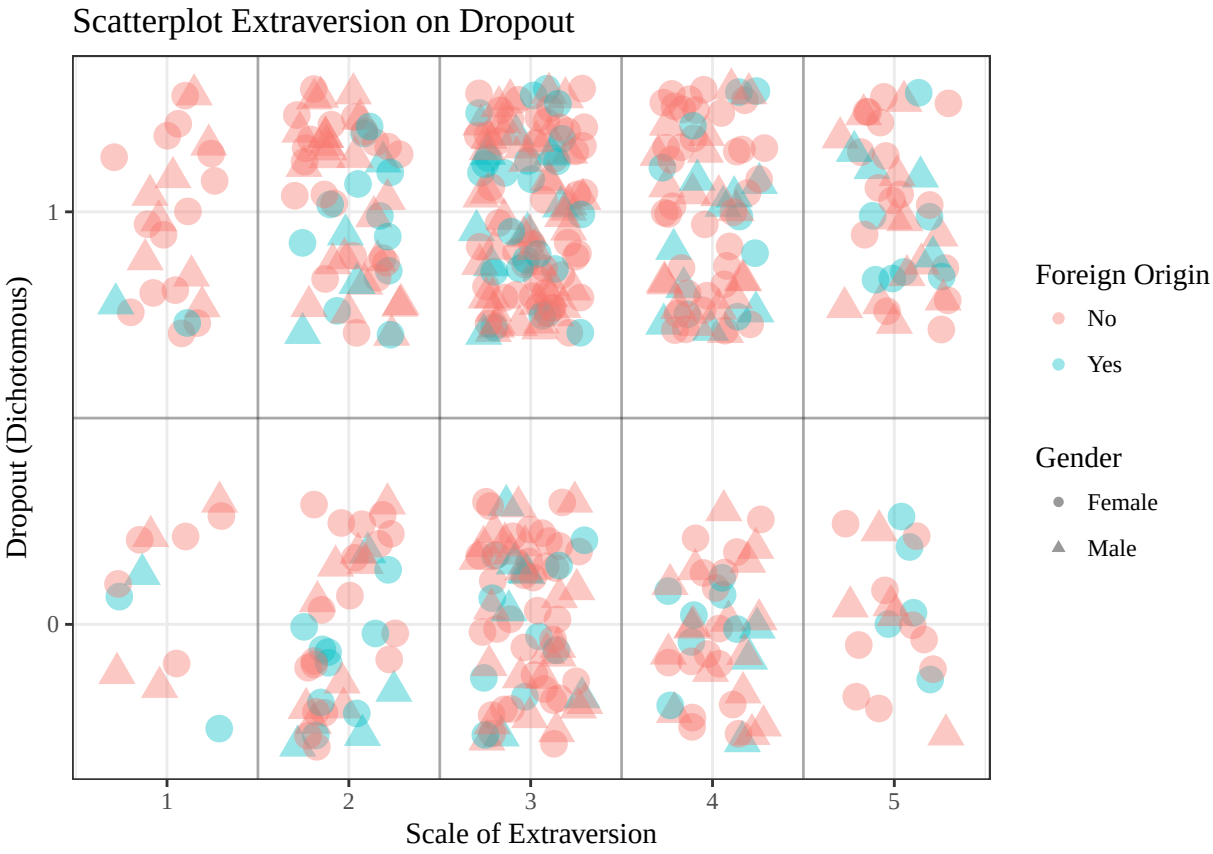
This section will display scatterplot for each of the *Big Five* variables on *Dropout*.

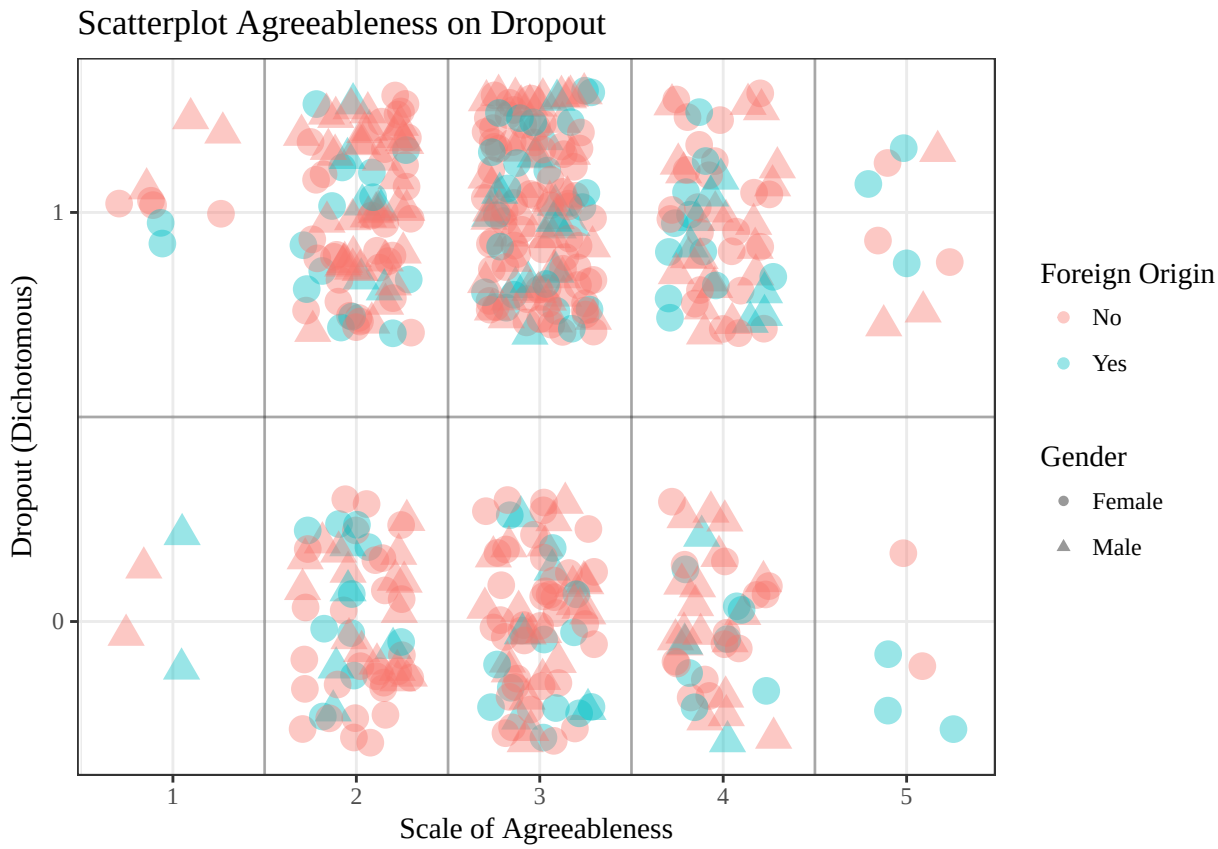
Scatterplot Conscientiousness on Dropout



Scatterplot Emotional Stability on Dropout

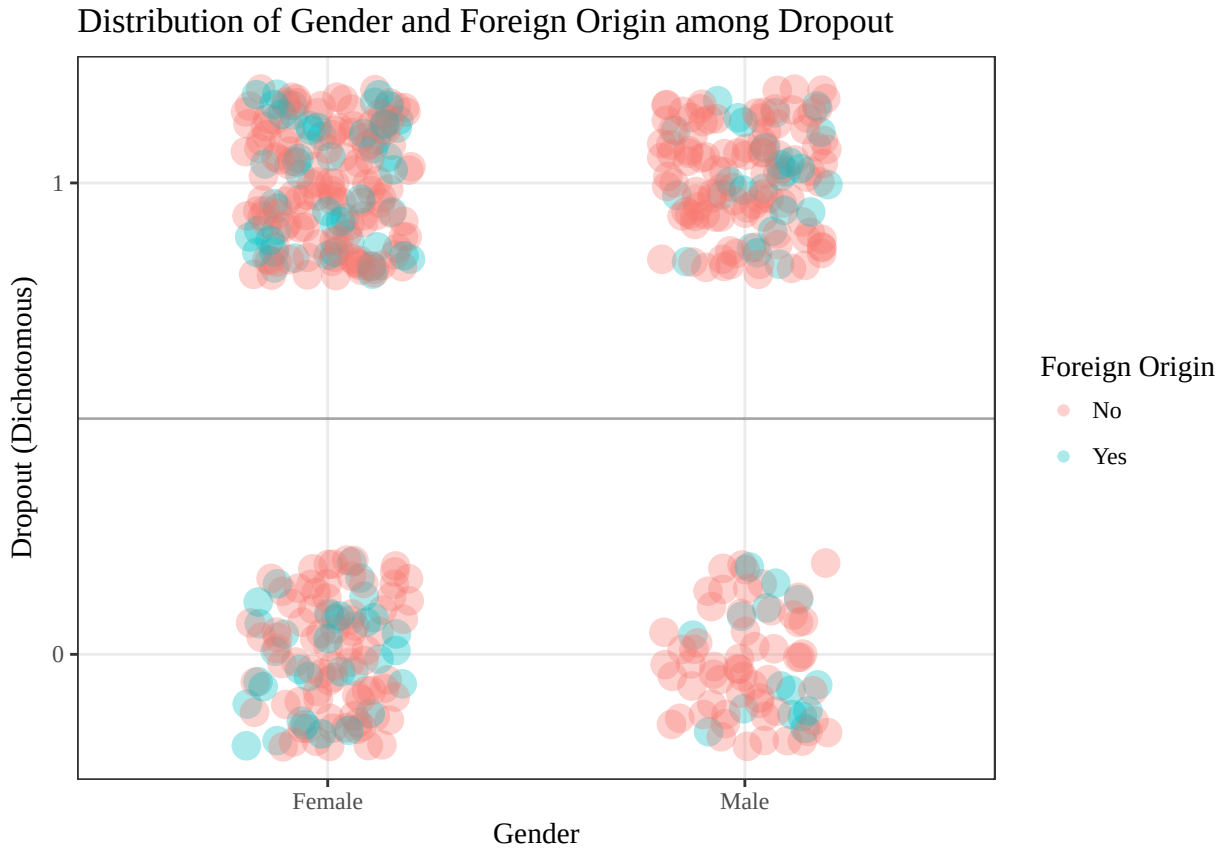

Scatterplot Openness on Dropout
Dropout (Dichotomous)
1
0
1
2
3
4
5
Scale of Openness
Foreign Origin
No
Yes
Gender
Female
Male


Scatterplot Extraversion on Dropout
Dropout (Dichotomous)
1
0
1
2
3
4
5
Scale of Extraversion
Foreign Origin
No
Yes
Gender
Female
Male



In this section the distribution of dropouts among *Gender* and *Foreign Origin* will be investigated.

```
#Scatterplot distribution gender and foreign origin among dropout
dropout.control <- ggplot(aes(x = geslachtBin, y = vrv_1),
                data = data.dropout) +
  geom_jitter(aes(colour = allochtoonBin,
                  size = 0.75),
              position = position_jitter(width = 0.2,
                                         height = 0.2),
                                         alpha = 1/3) +
  scale_y_continuous(breaks = c(0, 1)) +
  labs(
    x = "Gender",
    y = "Dropout (Dichotomous)",
    title = "Distribution of Gender and Foreign Origin among Dropout") +
  theme_bw(base_family = "serif") +
  guides(colour = guide_legend("Foreign Origin"),
         size = FALSE) +
  geom_hline(yintercept = 0.5, alpha = 0.3)

dropout.control
```




```
#Dropouts by male and female
by(data.dropout$vrv_1, data.dropout$geslachtBin, length)
```

```
## data.dropout$geslachtBin: Female
## [1] 292
## ------------------------------------------------------------
## data.dropout$geslachtBin: Male
## [1] 207
```

```
#Dropouts by foreign origin and native origin
by(data.dropout$vrv_1, data.dropout$allochtoonBin, length)
```

```
## data.dropout$allochtoonBin: No
## [1] 376
## ------------------------------------------------------------
## data.dropout$allochtoonBin: Yes
## [1] 123
```

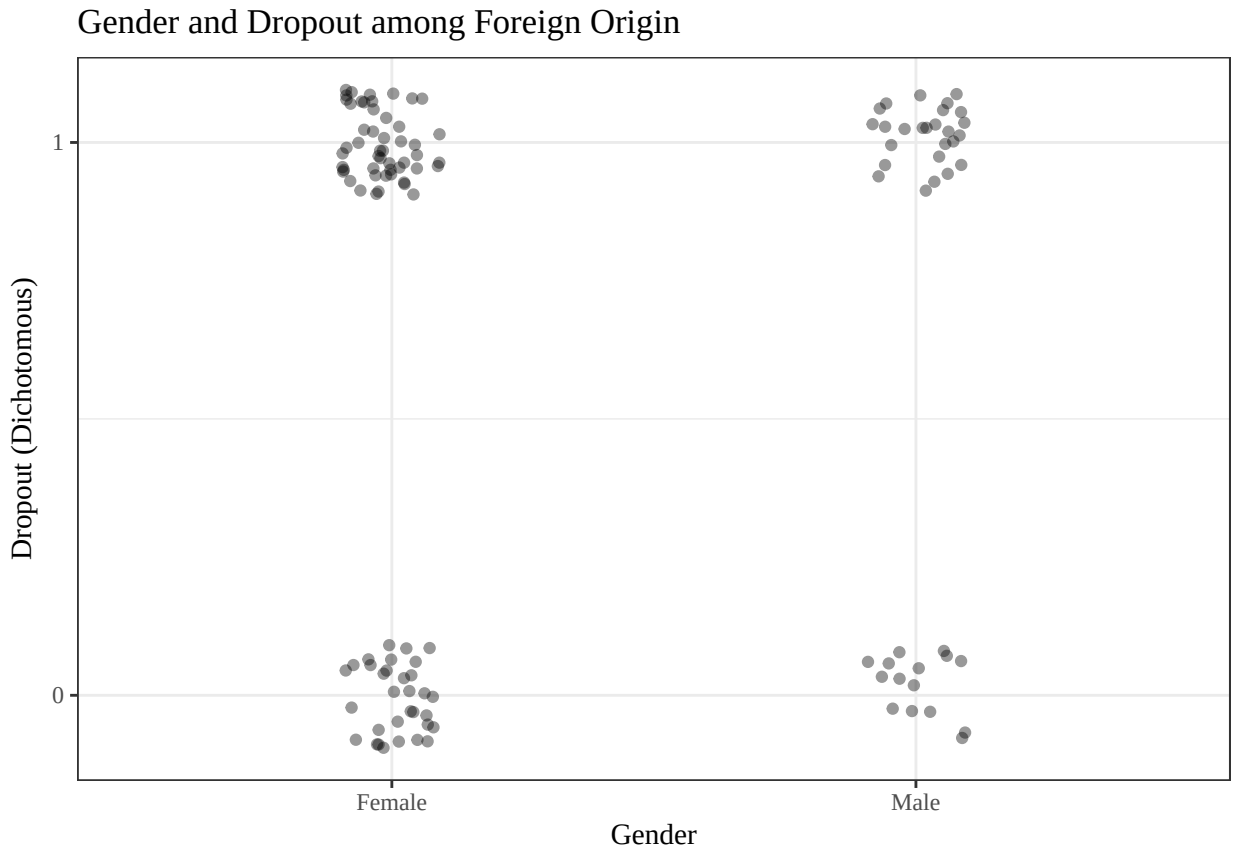
```
#Amount of people of foreign origin in data
length(subset(data.dropout, allochtoonBin == "Yes")$allochtoonBin)
```

```
## [1] 123
```

```
#Amount of people of native origin in data
length(subset(data.dropout, allochtoonBin == "No")$allochtoonBin)
```

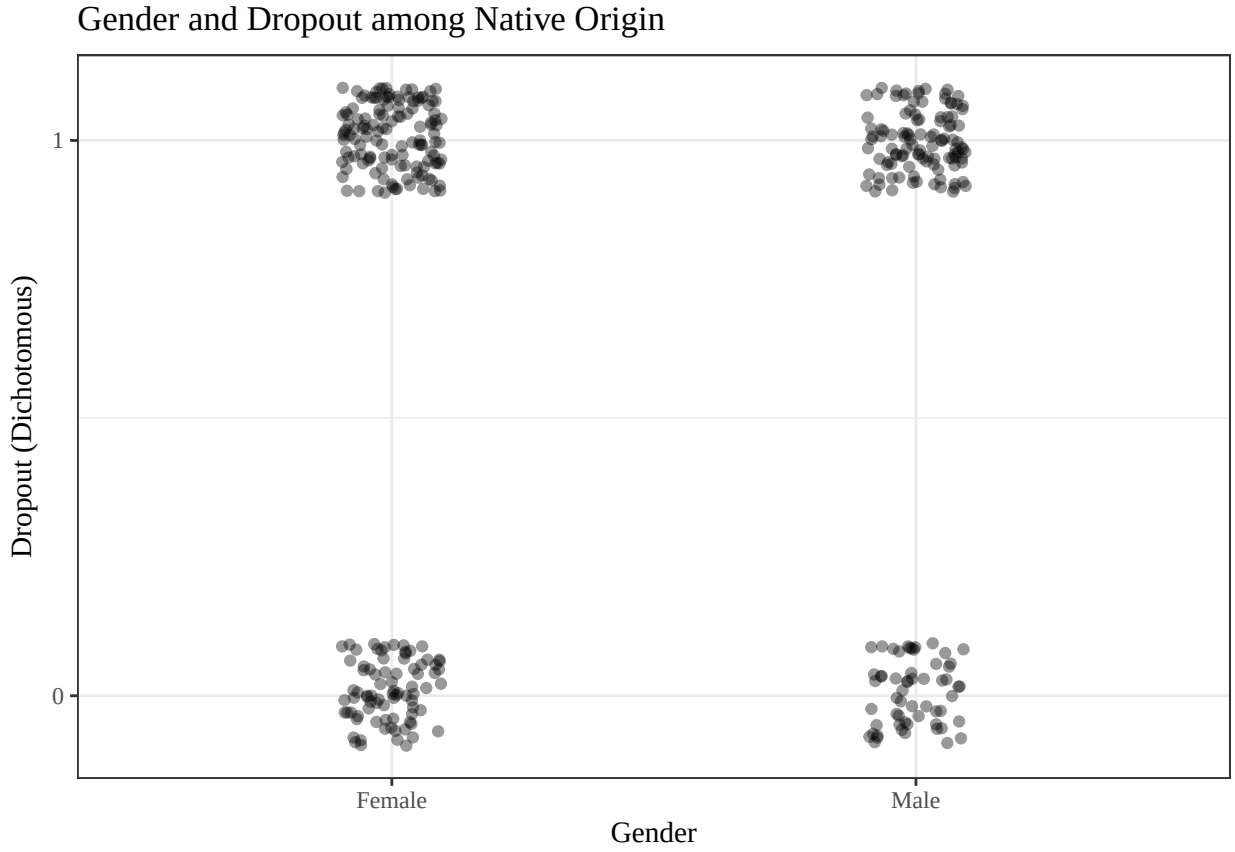
```
## [1] 376
```

```r
#Create variables for either foreign origin is true or not
f1 <- subset(data.dropout, allochtoonBin == "Yes")
f0 <- subset(data.dropout, allochtoonBin == "No")

#Scatterplot dropout among foreign origin
dropout.f1 <- ggplot(aes(x = geslachtBin, y = vrv_1),
                data = f1) +
  geom_jitter(position = position_jitter(width = 0.095,
                                         height = 0.095),
                                         alpha = 0.4) +
  scale_y_continuous(breaks = c(0, 1)) +
  labs(
    x = "Gender",
    y = "Dropout (Dichotomous)",
    title = "Gender and Dropout among Foreign Origin") +
  theme_bw(base_family = "serif")

dropout.f1
```




```r
#Scatterplot dropout among native origin
dropout.f0 <- ggplot(aes(x = geslachtBin, y = vrv_1),
                data = f0) +
  geom_jitter(position = position_jitter(width = 0.095,
                                         height = 0.095),
                                         alpha = 0.4) +
  scale_y_continuous(breaks = c(0, 1)) +
```

```
  labs(
    x = "Gender",
    y = "Dropout (Dichotomous)",
    title = "Gender and Dropout among Native Origin") +
  theme_bw(base_family = "serif")

dropout.f0
```




```
#Dropouts by male and female among foreign origin
by(f1$vrv_1, f1$geslachtBin, length)
```

```
## f1$geslachtBin: Female
## [1] 82
## ------------------------------------------------------------
## f1$geslachtBin: Male
## [1] 41
```

```
#Dropouts by male and female among native origin
by(f0$vrv_1, f0$geslachtBin, length)
```

```
## f0$geslachtBin: Female
## [1] 210
## ------------------------------------------------------------
## f0$geslachtBin: Male
## [1] 166
```

```r
#DROPOUT FOREIGN ORIGIN
#Dropout YES foreign origin and male
length(subset(f1, vrv_1 == 1 & geslachtBin == "Male")$allochtoonBin)
```

```
## [1] 26
```

```r
#Dropout YES foreign origin and female
length(subset(f1, vrv_1 == 1 & geslachtBin == "Female")$allochtoonBin)
```

```
## [1] 50
```

```r
#Dropout NO foreign origin and male
length(subset(f1, vrv_1 == 0 & geslachtBin == "Male")$allochtoonBin)
```

```
## [1] 15
```

```r
#Dropout NO foreign origin and female
length(subset(f1, vrv_1 == 0 & geslachtBin == "Female")$allochtoonBin)
```

```
## [1] 32
```

```r
#DROPOUT NATIVE ORIGIN
#Generate native YES
n1 <- subset(data.dropout, allochtoonBin == "No")
#Generate native NO
n0 <- subset(data.dropout, allochtoonBin == "Yes")

#Dropout YES native origin and male
length(subset(n1, vrv_1 == 1 & geslachtBin == "Male")$allochtoonBin)
```

```
## [1] 108
```

```r
#Dropout YES native origin and female
length(subset(n1, vrv_1 == 1 & geslachtBin == "Female")$allochtoonBin)
```

```
## [1] 129
```

```r
#Dropout NO native origin and male
length(subset(n1, vrv_1 == 0 & geslachtBin == "Male")$allochtoonBin)
```

```
## [1] 58
```

```r
#Dropout NO native origin and female
length(subset(n1, vrv_1 == 0 & geslachtBin == "Female")$allochtoonBin)
```

```
## [1] 81
```

Correlation Matrix

```r
CorMatrix <- cor(data.dropout[,1:25])
corrplot(CorMatrix, method = "circle", type = "lower", order ="AOE")
```