R Code

Erik Pak

2024-08-06

Import Libraries

```
library(tidyverse)      # data wrangling
library(corrplot)       # correlation plot
library(psych)          # used for describe
library(ggpubr)         # combine plots


#######################################
# Categorical variables mapping.     #
#######################################

# Education:
# 1 'Below College'
# 2 'College'
# 3 'Bachelor'
# 4 'Master'
# 5 'Doctor'
#
# Environment Satisfaction:
# 1 'Low'
# 2 'Medium'
# 3 'High'
# 4 'Very High'
#
# Job Involvement:
# 1 'Low'
# 2 'Medium'
# 3 'High'
# 4 'Very High'
#
# Job Satisfaction:
# 1 'Low'
# 2 'Medium'
# 3 'High'
# 4 'Very High'
#
# Relationship Satisfaction:
# 1 'Low'
# 2 'Medium'
# 3 'High'
# 4 'Very High'
```

```
#
# WorkLife Balance:
# 1 'Bad'
# 2 'Good'
# 3 'Better'
# 4 'Best'
#
# Performance Rating
# 1 'Low'
# 2 'Good'
# 3 'Excellent'
# 4 'Outstanding'
```

Import Data

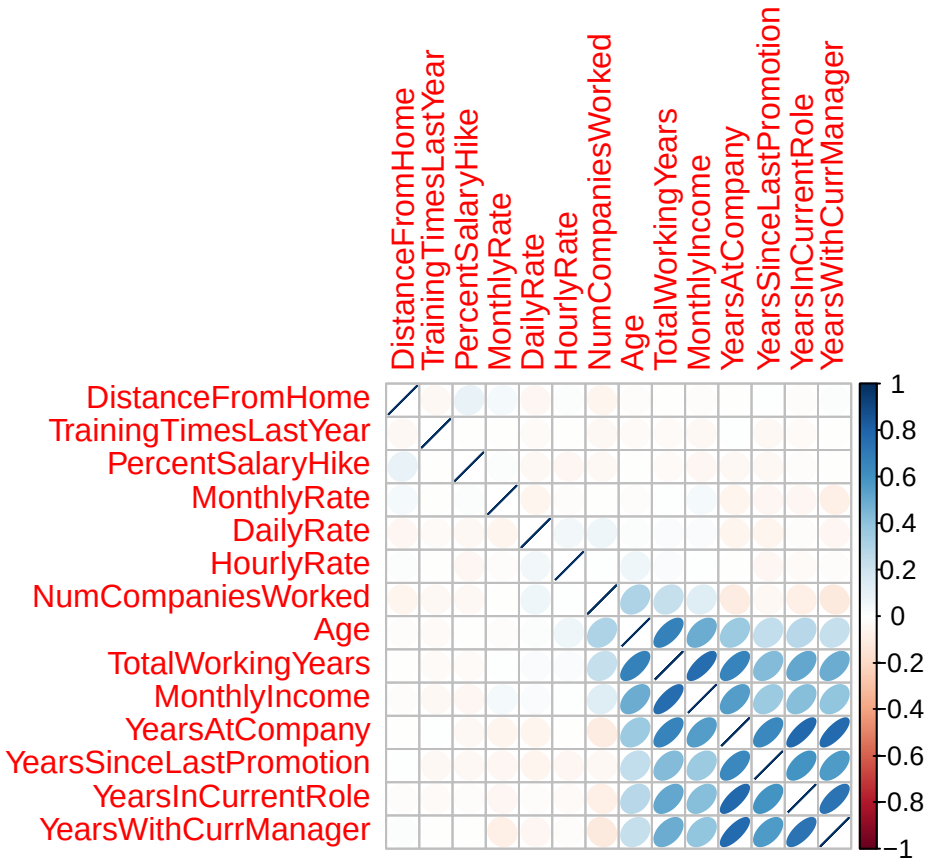
```
# set working directory
setwd("~/Downloads/Data")

# import data
hr <- read.csv('HREmployeeAttrition.csv')

# copy data frame
hrCopy <- hr

# no NAs
sum(is.na(hr))
```

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

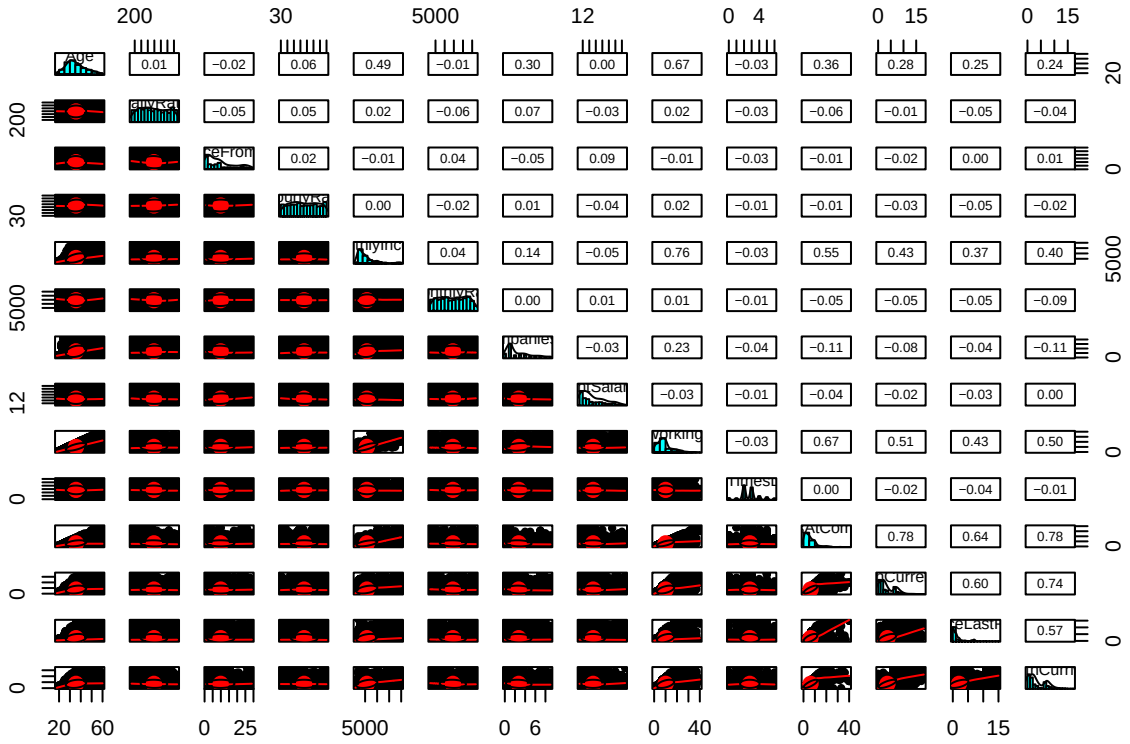
```
# remove EmployeeCount & EmployeeNumber & Over18 & StandardHours
hr <- dplyr::select(hr, -c(EmployeeCount, EmployeeNumber, Over18, StandardHours))

# over sample the data for class imbalance
oversampled <- ROSE::ovun.sample(Attrition ~ ., data = hr, method = "over", N = 2400)

# copy
hr <- oversampled$data
```

Data Wrangling

```
# categorical variables
catCols <- c('Attrition', 'BusinessTravel', 'Department', 'Education',
             'EducationField', 'EnvironmentSatisfaction', 'Gender',
             'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction',
             'MaritalStatus', 'OverTime', 'PerformanceRating',
             'RelationshipSatisfaction', 'StockOptionLevel', 'WorkLifeBalance')


# correlation plot
corrplot(cor(hr[ , ! names(hr) %in% c(catCols)]), method = "ellipse", order = "AOE")
```




```
# pairs plot
pairs.panels((hr[ , ! names(hr) %in% c(catCols)]))
```




```
# select subset of HR data ( categorical variables ONLY)
hrCat <- hr %>% dplyr::select(all_of(catCols))

# more deccriptive for the plots
hrCat$Attrition <- ifelse(hrCat$Attrition == "Yes","Att.Yes","Att.No")
hrCat$OverTime <- ifelse(hrCat$OverTime == "Yes","OT.Yes","OT.No")

# convert to factors
hrCat[catCols] <- lapply(hrCat[catCols], factor)

# change factor level
levels(hrCat$Education) <- as.factor(c('Below College','College',
                                       'Bachelor','Master','Doctor'))

levels(hrCat$WorkLifeBalance) <- as.factor(c('WLB.Bad', 'WLB.Good',
                                             'WLB.Better','WLB.Best'))

levels(hrCat$RelationshipSatisfaction) <- as.factor(c('RS.Low', 'RS.Medium',
                                                      'RS.High', 'RS.Very High'))

levels(hrCat$EnvironmentSatisfaction) <- as.factor(c('ES.Low', 'ES.Medium',
                                                     'ES.High', 'ES.Very High'))

levels(hrCat$JobInvolvement) <- as.factor(c('JI.Low', 'JI.Medium',
                                            'JI.High', 'JI.Very High'))
```

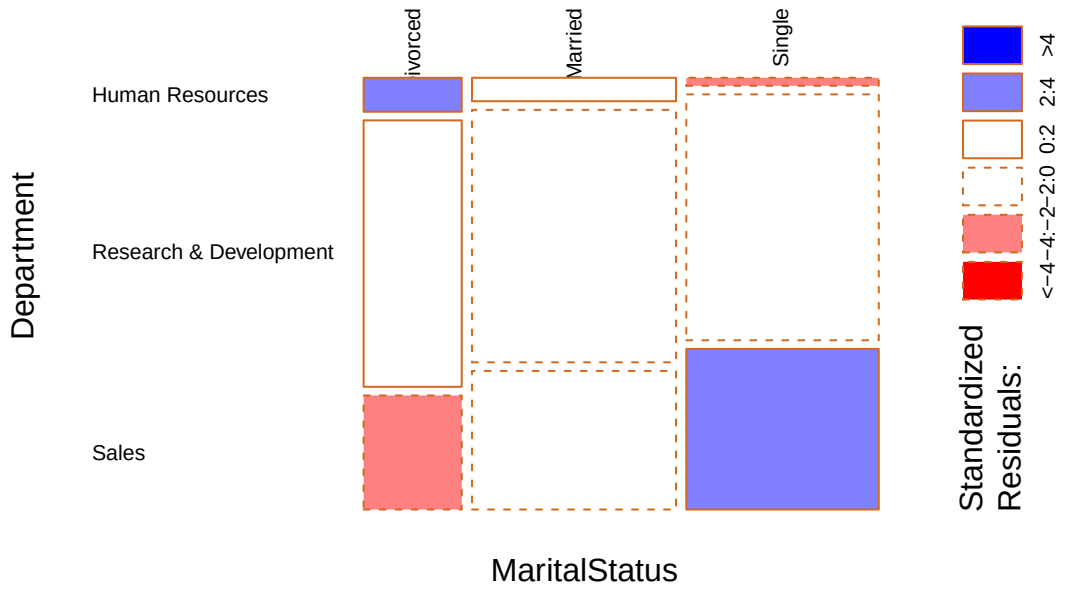
```r
levels(hrCat$JobSatisfaction) <- as.factor(c('JS.Low', 'JS.Medium',
                                             'JS.High', 'JS.Very High'))

levels(hrCat$JobSatisfaction) <- as.factor(c('JS.Low', 'JS.Medium',
                                             'JS.High', 'JS.Very High'))

levels(hrCat$JobLevel) <- as.factor((c('JL.1',"JL.2", "JL.3", "JL.4", "JL.5")))

levels(hrCat$StockOptionLevel) <- as.factor((c('SOL.0',"SOL.1", "SOL.2",
                                               "SOL.3")))

# only 3s & 4s
levels(hrCat$PerformanceRating) <- as.factor((c('PR.Excellent',
                                                'PR.Outstanding')))

# summary
summary(hrCat)
```

```
##    Attrition              BusinessTravel                  Department
##  Att.No :1233   Non-Travel       : 191   Human Resources       : 113
##  Att.Yes:1167   Travel_Frequently: 565   Research & Development:1463
##                 Travel_Rarely    :1644   Sales                 : 824
##
##
##
##
##          Education            EducationField EnvironmentSatisfaction
##  Below College:307   Human Resources : 56    ES.Low      :533
##  College      :460   Life Sciences   :937    ES.Medium   :468
##  Bachelor     :932   Marketing       :302    ES.High     :714
##  Master       :630   Medical         :715    ES.Very High:685
##  Doctor       : 71   Other           :131
##                      Technical Degree:259
##
##     Gender          JobInvolvement JobLevel                         JobRole
##  Female: 943   JI.Low      : 193   JL.1:1103   Sales Executive          :562
##  Male  :1457   JI.Medium   : 634   JL.2: 773   Laboratory Technician    :510
##                JI.High     :1371   JL.3: 317   Research Scientist       :463
##                JI.Very High: 202   JL.4: 121   Sales Representative     :217
##                                    JL.5:  86   Manufacturing Director   :187
##                                                Healthcare Representative:155
##                                                (Other)                  :306
##      JobSatisfaction  MaritalStatus   OverTime         PerformanceRating
##  JS.Low      :534    Divorced:476   OT.No :1490   PR.Excellent  :2041
##  JS.Medium   :463    Married :990   OT.Yes: 910   PR.Outstanding: 359
##  JS.High     :736    Single  :934
##  JS.Very High:667
##
##
##
##  RelationshipSatisfaction StockOptionLevel   WorkLifeBalance
##  RS.Low      :498         SOL.0:1225       WLB.Bad   : 177
##  RS.Medium   :474         SOL.1: 813       WLB.Good  : 570
```

```
##  RS.High     :727         SOL.2: 206       WLB.Better:1403
##  RS.Very High:701         SOL.3: 156       WLB.Best   : 250
##
##
##
```

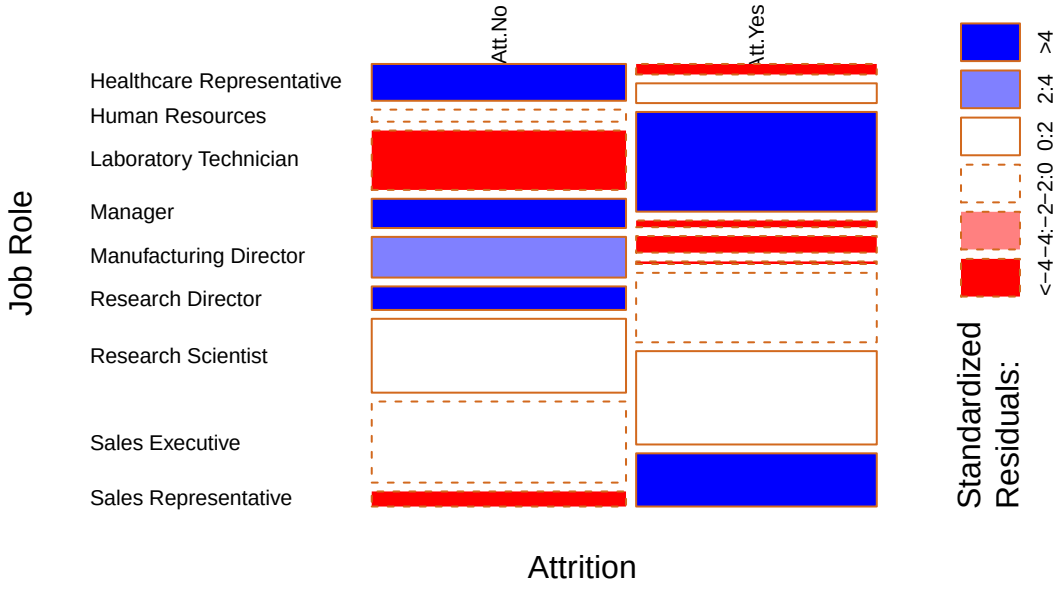
Mosaic Plots

```
#################
# Mosaic plots #
#################
# mosaic plots *
mosaicplot(table(hrCat$MaritalStatus, hrCat$Department),
           las = 2, cex.axis = 0.7,
           main = "",
           xlab = "MaritalStatus",
           ylab = "Department",
           border = "chocolate",
           shade = TRUE)
```

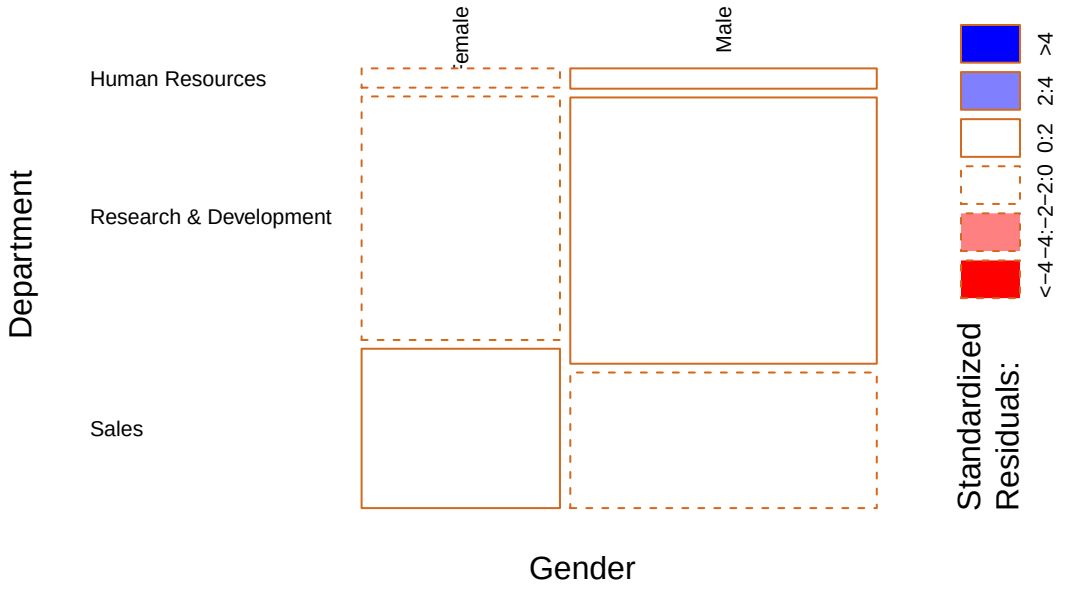



```
mosaicplot(table(hrCat$Attrition, hrCat$JobRole),
           las = 2, cex.axis = 0.7,
           main = "",
           xlab = "Attrition",
           ylab = "Job Role",
```

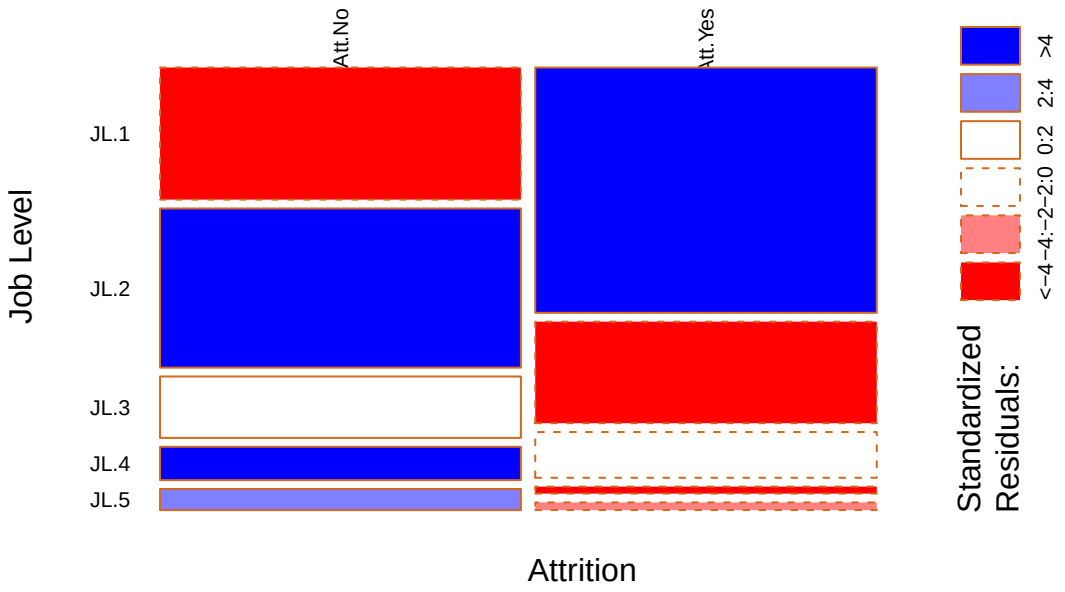
```
          border = "chocolate",
          shade = TRUE)
```



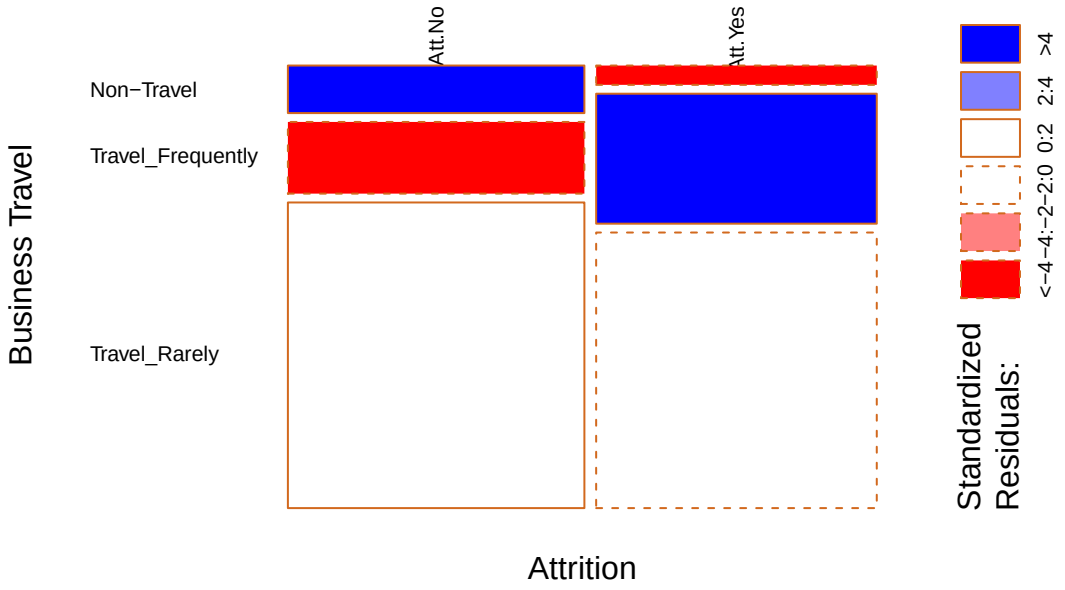

```
mosaicplot(table(hrCat$Gender, hrCat$Department),
           las = 2, cex.axis = 0.7,
           main = "",
           xlab = "Gender",
           ylab = "Department",
           border = "chocolate",
           shade = TRUE)
```




```
mosaicplot(table(hrCat$Attrition, hrCat$JobLevel),
           las = 2, cex.axis = 0.7,
           main = "",
           xlab = "Attrition",
           ylab = "Job Level",
           border = "chocolate",
           shade = TRUE)
```

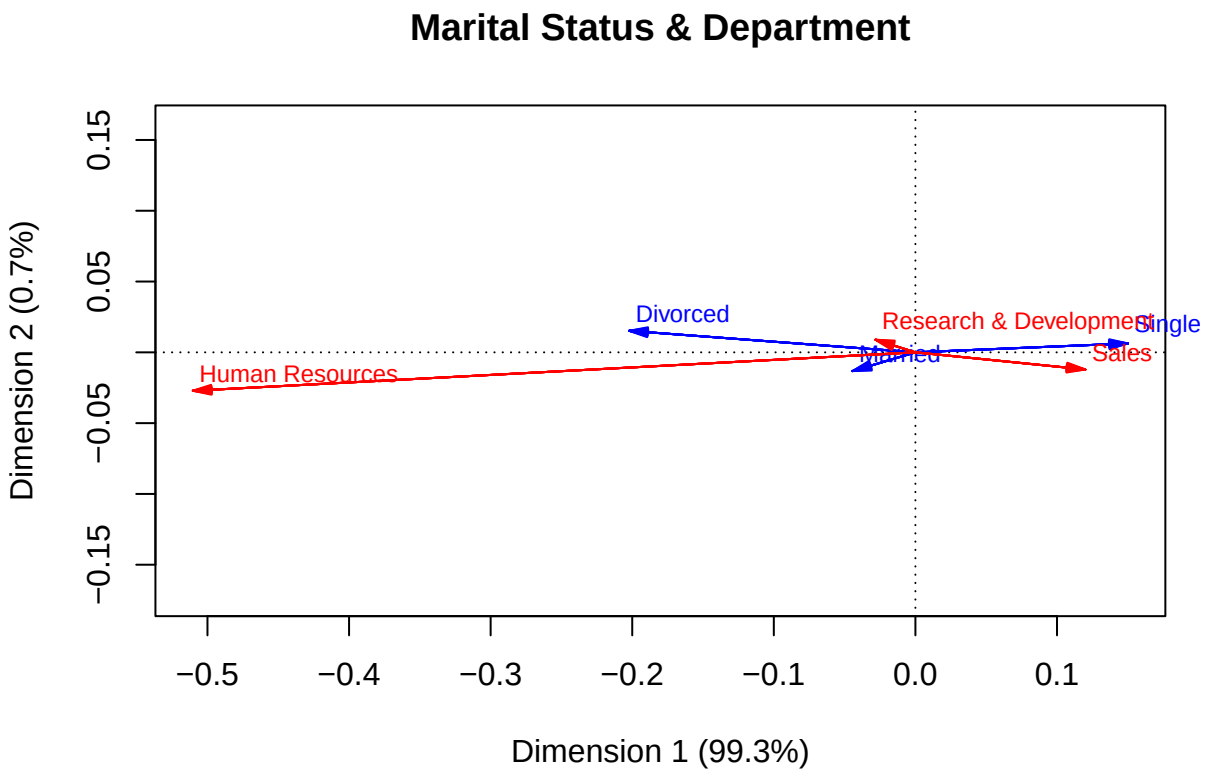



```
mosaicplot(table(hrCat$Attrition, hrCat$BusinessTravel),
           las = 2, cex.axis = 0.7,
           main = "",
           xlab = "Attrition",
           ylab = "Business Travel",
           border = "chocolate",
           shade = TRUE)
```

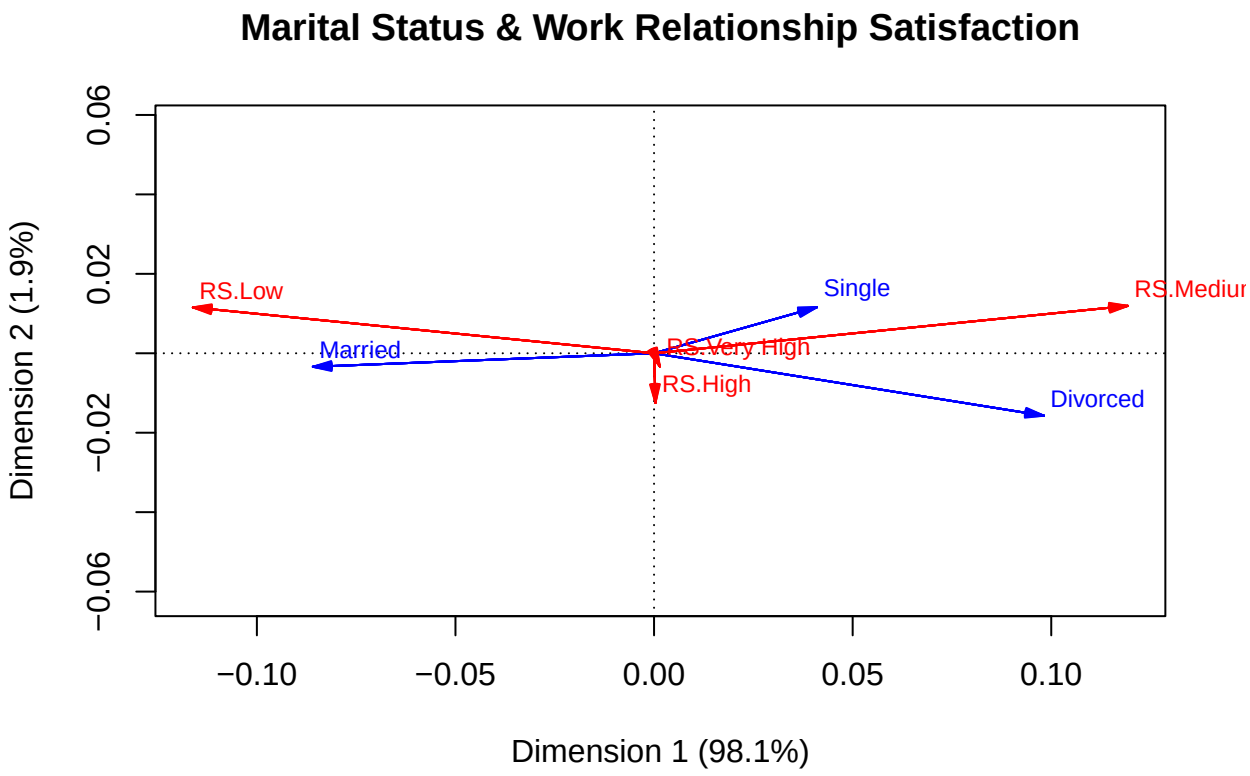


CA Plots

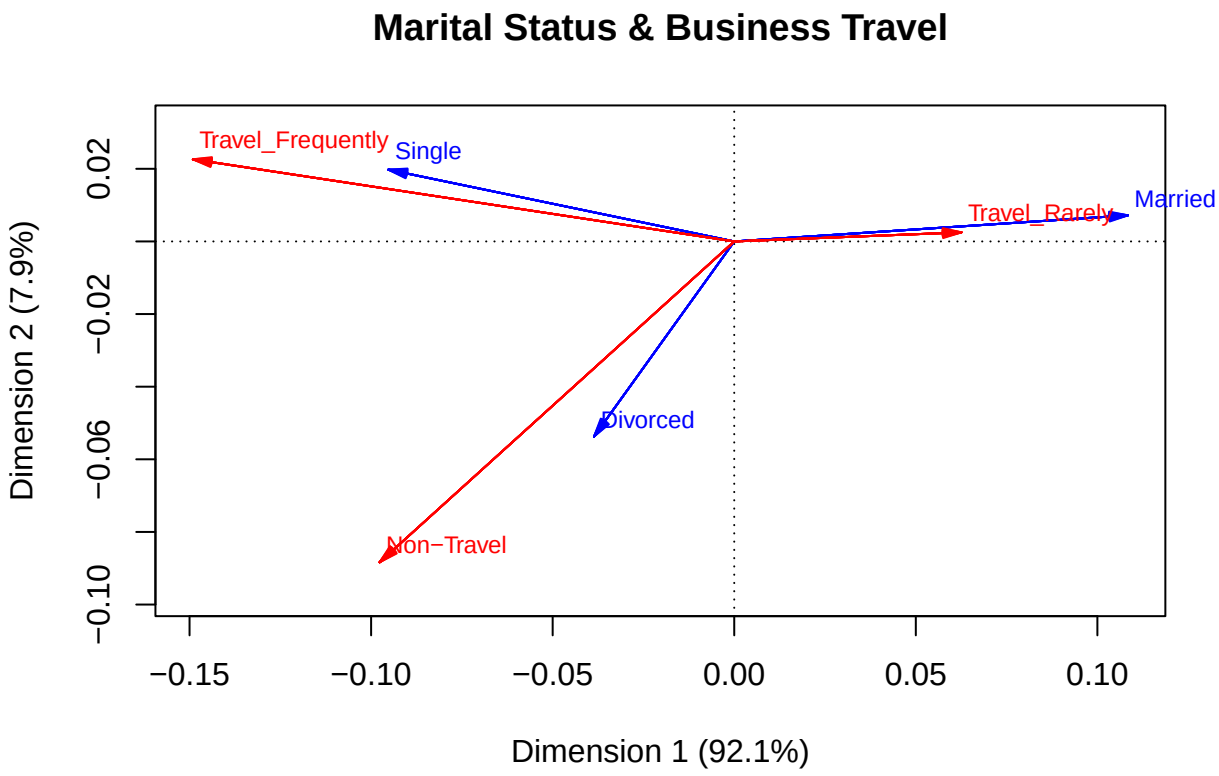
```
# correspondence analysis
fitca <-  ca::ca(table(hrCat$MaritalStatus, hrCat$Department))
# plot
plot(fitca, main = "Marital Status & Department", arrows = c(T,T))
```



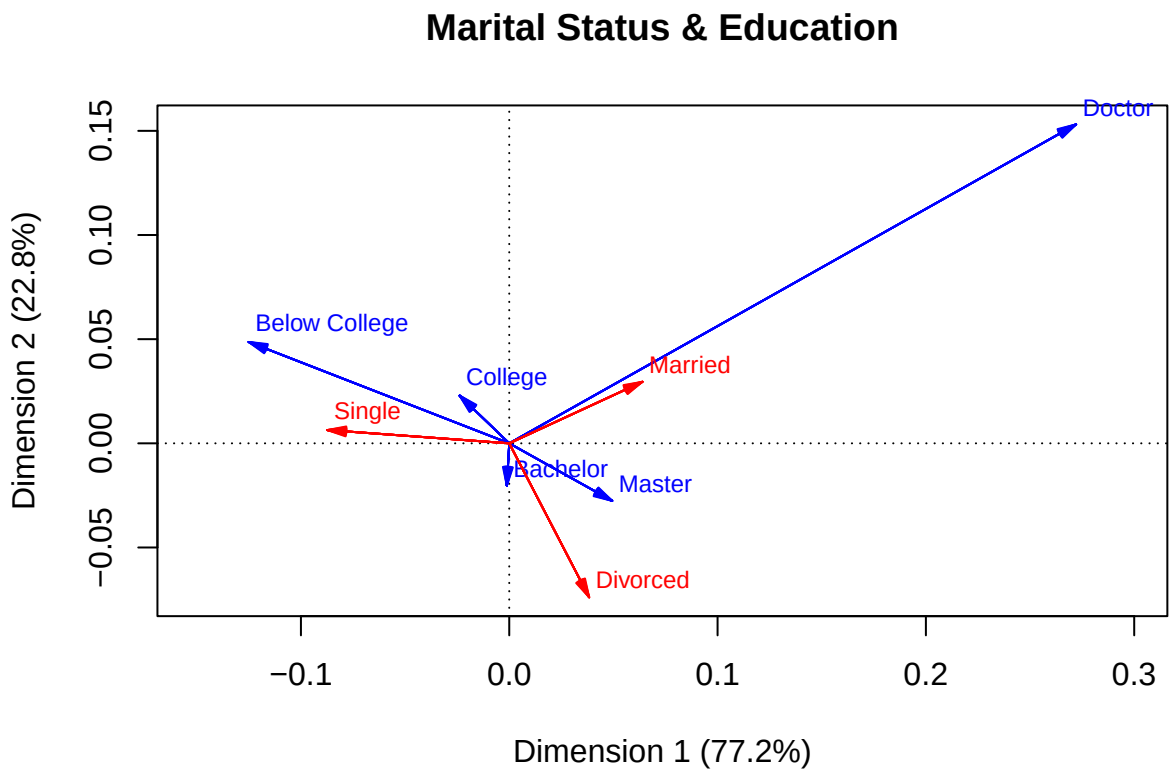

```r
# correspondence analysis
fitca <-  ca::ca(table(hrCat$MaritalStatus, hrCat$RelationshipSatisfaction))
# plot
plot(fitca, main = "Marital Status & Work Relationship Satisfaction", arrows = c(T,T))
```



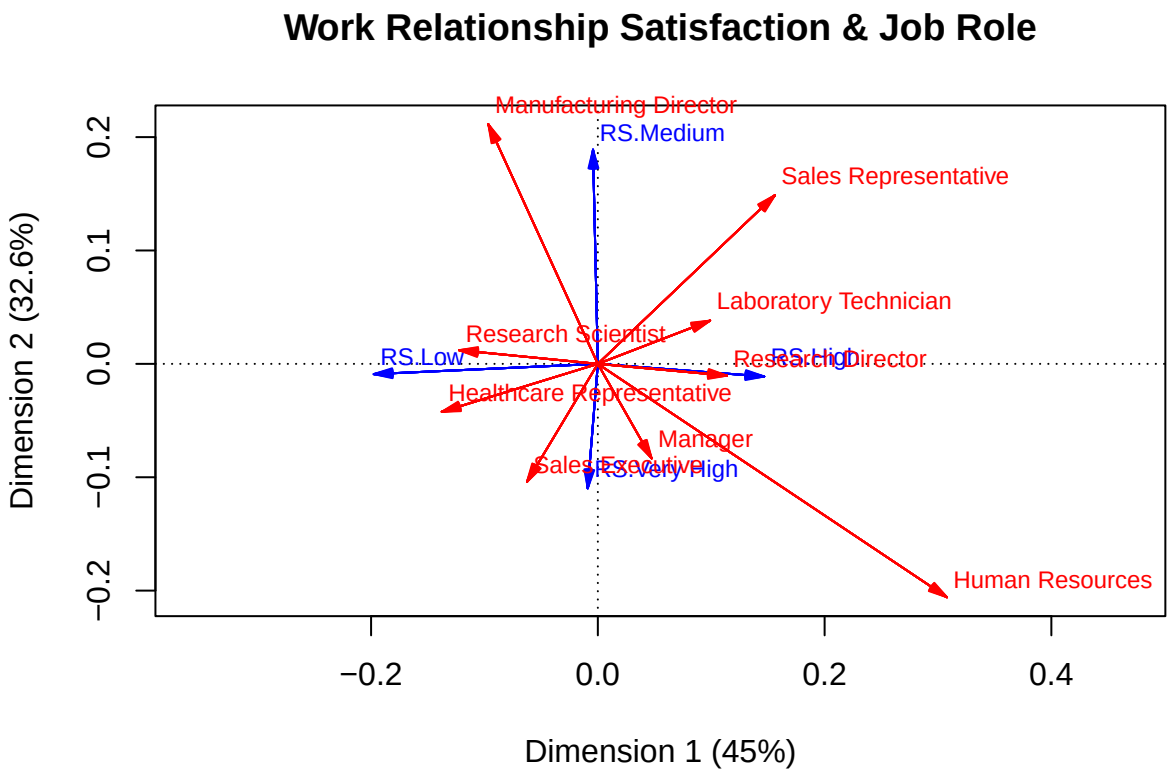

```
# correspondence analysis
fitca <-  ca::ca(table(hrCat$MaritalStatus, hrCat$BusinessTravel))
# plot
plot(fitca, main = "Marital Status & Business Travel", arrows = c(T,T))
```



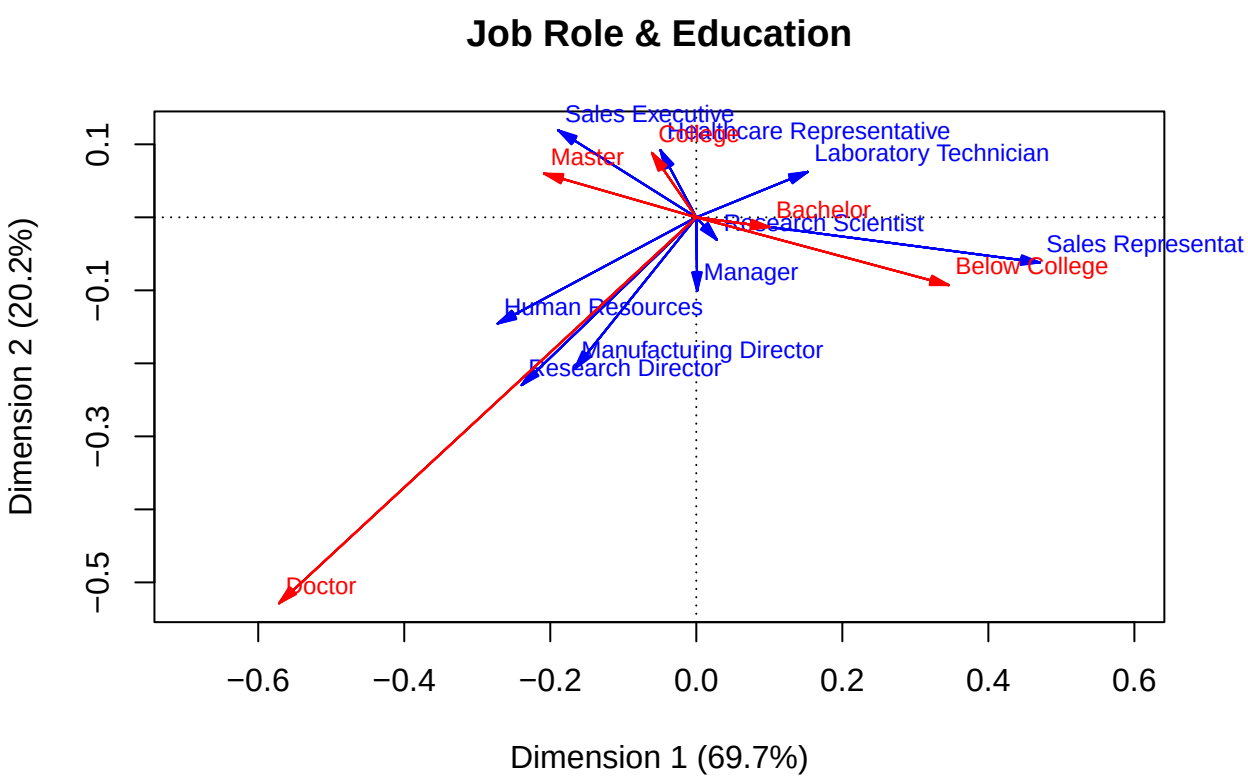

```
# correspondence analysis
fitca <-  ca::ca(table(hrCat$Education, hrCat$MaritalStatus))
# plot
plot(fitca, main = "Marital Status & Education", arrows = c(T,T))
```



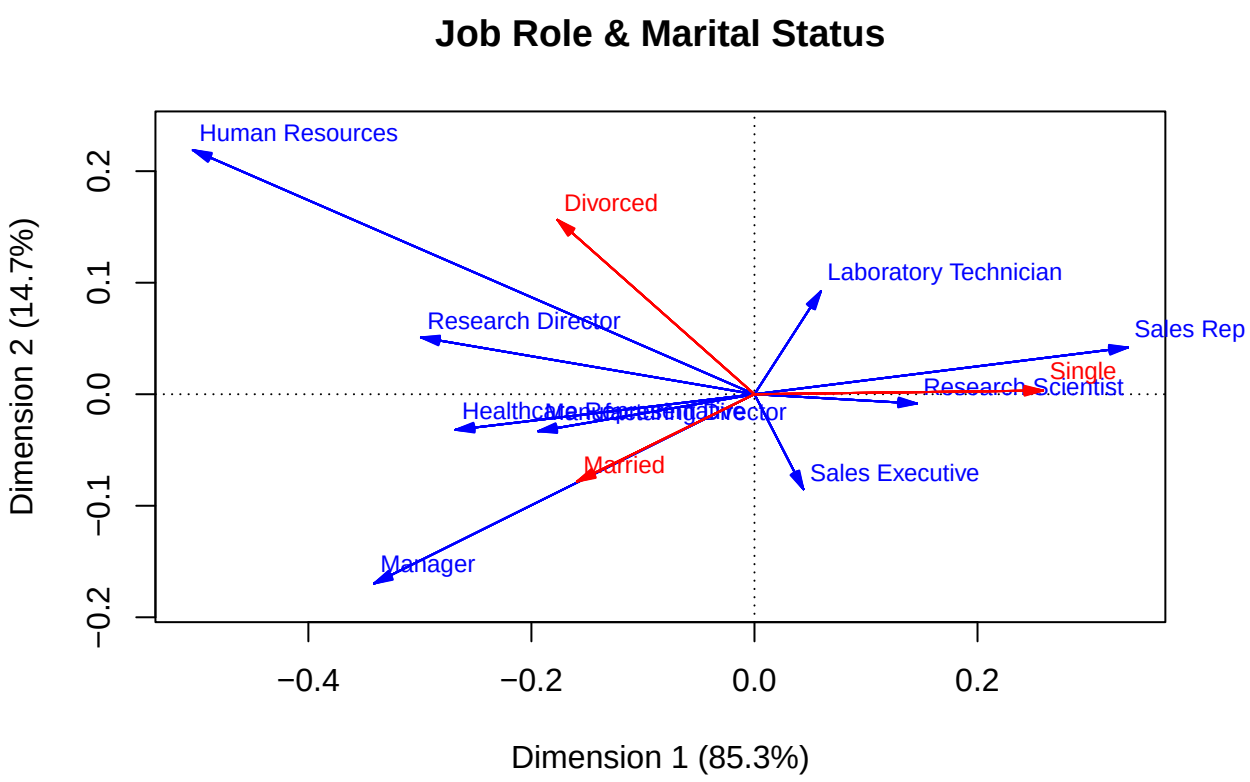

```
# correspondence analysis
fitca <-  ca::ca(table(hrCat$RelationshipSatisfaction, hrCat$JobRole))
# plot
plot(fitca, main = "Work Relationship Satisfaction & Job Role", arrows = c(T,T))
```




```
# correspondence analysis
fitca <-  ca::ca(table(hrCat$JobRole, hrCat$Education))
# plot
plot(fitca, main = "Job Role & Education", arrows = c(T,T))
```



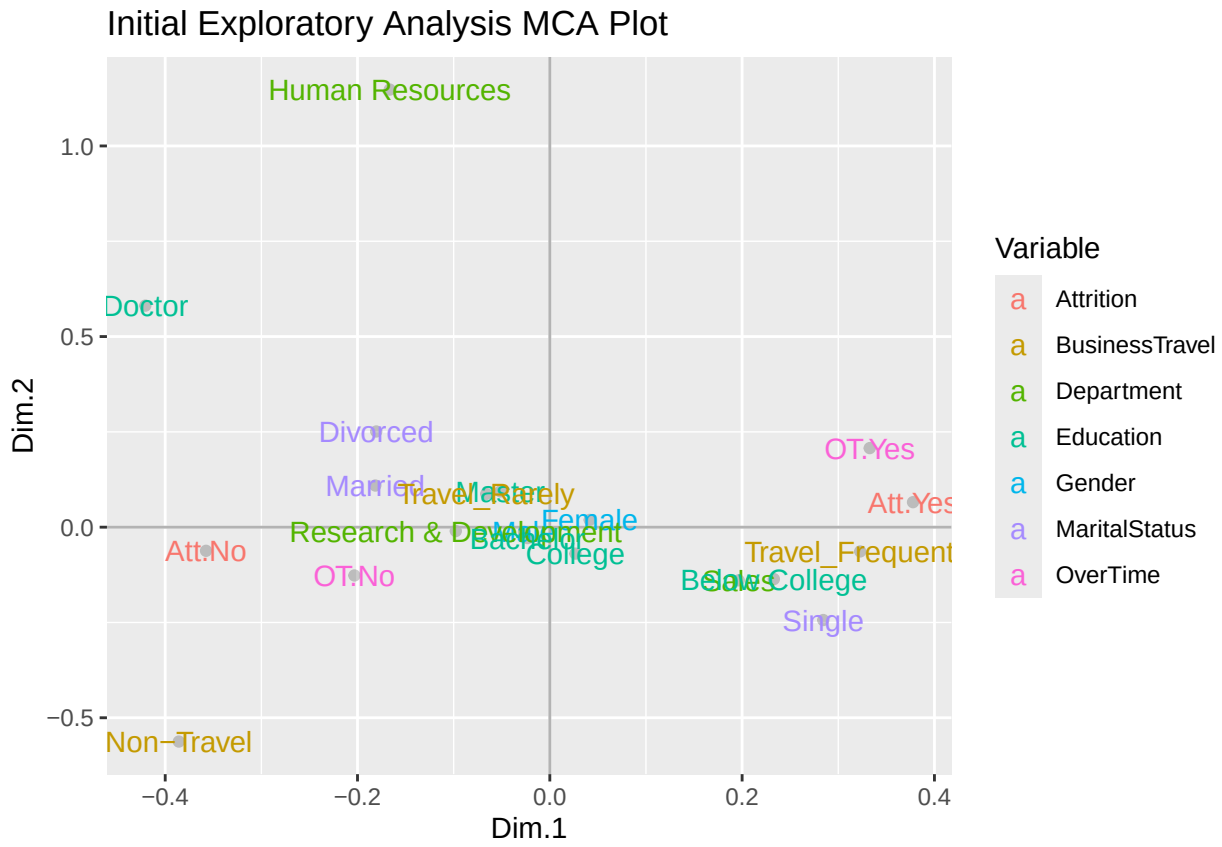

```
# correspondence analysis
fitca <-  ca::ca(table(hrCat$JobRole, hrCat$MaritalStatus))
# plot
plot(fitca, main = "Job Role & Marital Status", arrows = c(T,T))
```



MCA Plots

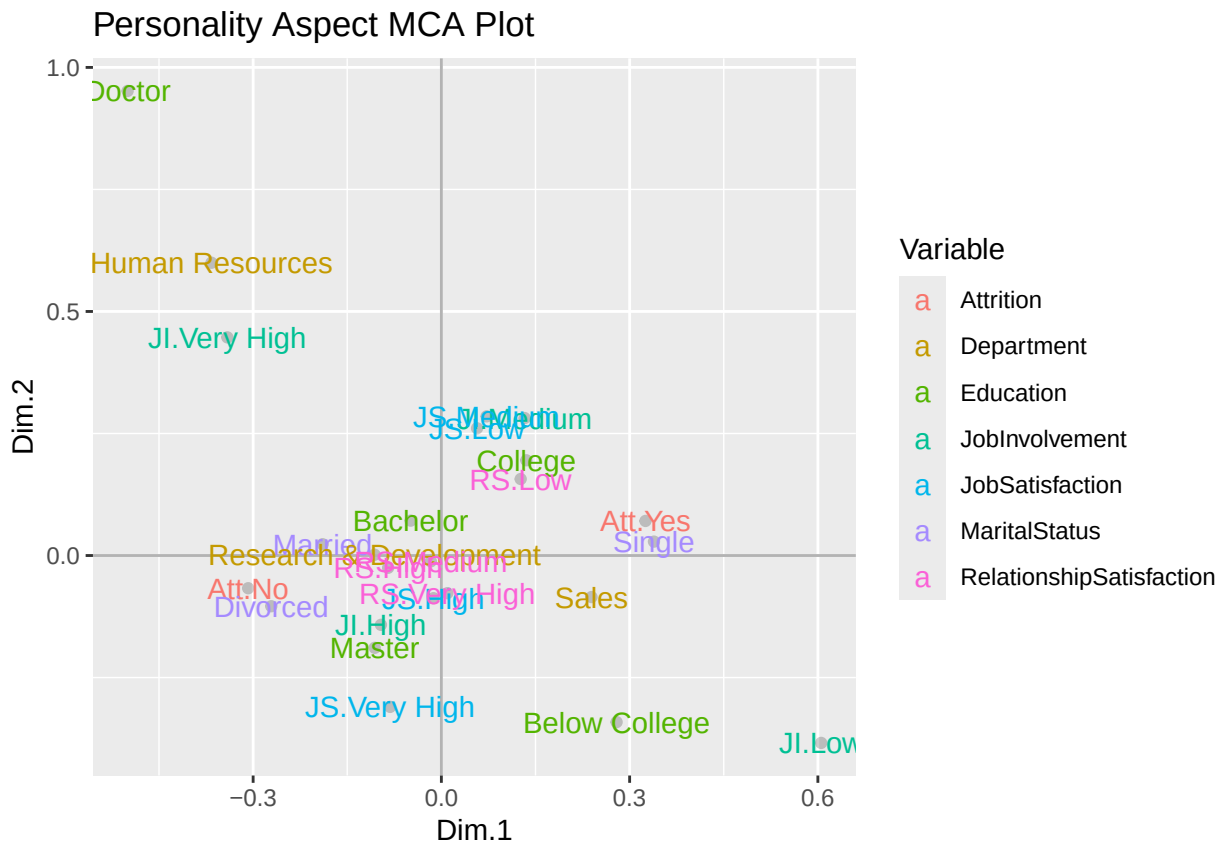
```
# additional categorical values
newHR <- hrCat[, c("JobRole", "Attrition")]

# additional categorical values
newHR <- hrCat[, c("RelationshipSatisfaction", "BusinessTravel",
                   "MaritalStatus", "JobRole", "Attrition", "WorkLifeBalance")]

# additional categorical values
newHR <- hrCat[, c("Gender", "RelationshipSatisfaction","EnvironmentSatisfaction",
                   "MaritalStatus", "WorkLifeBalance","JobSatisfaction",
                   "PerformanceRating")]

# additional categorical values
newHR <- hrCat[, c("Attrition", "Gender", "RelationshipSatisfaction", "OverTime",
                   "MaritalStatus","EnvironmentSatisfaction", "JobLevel",
                   "JobSatisfaction","WorkLifeBalance", "Education")]



# make sure include
newHR <- hrCat[, c("Attrition", "OverTime", "Gender", "MaritalStatus",
                   "Department", "Education","BusinessTravel")]

# number of categories per variable
```

```
cats <- apply(newHR, 2, function(x) nlevels(as.factor(x)))

# apply MCA using FactoMineR Library
mcaHR.Burt <- FactoMineR::MCA(newHR, method = "Burt", graph = FALSE)

# data frame with variable coordinates
mHRB_vars_df = data.frame(mcaHR.Burt$var$coord, Variable = rep(names(cats), cats))

# plot of variable categories
ggplot(data=mHRB_vars_df,
       aes(x = Dim.1, y = Dim.2,label = rownames(mHRB_vars_df))) +
  geom_point(alpha = 0.2) +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_text(aes(colour=Variable), angle = 0) +
  ggtitle("Initial Exploratory Analysis MCA Plot")
```



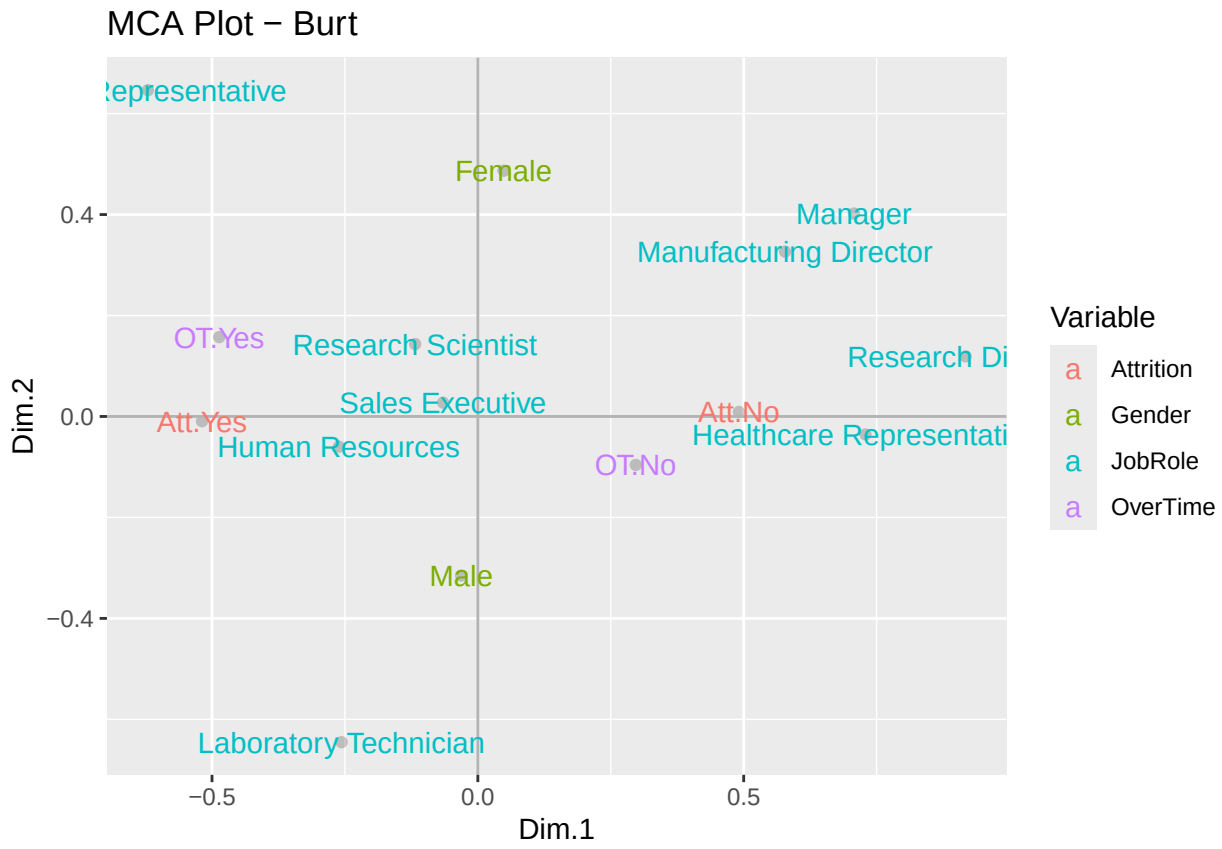

```
# make sure include
newHR <- hrCat[, c("Attrition", "MaritalStatus", "Education", "Department",
                   "JobInvolvement", "JobSatisfaction","RelationshipSatisfaction")]

# number of categories per variable
cats <- apply(newHR, 2, function(x) nlevels(as.factor(x)))

# apply MCA using FactoMineR Library
```

```
mcaHR.Burt <- FactoMineR::MCA(newHR, method = "Burt", graph = FALSE)

# data frame with variable coordinates
mHRB_vars_df = data.frame(mcaHR.Burt$var$coord, Variable = rep(names(cats), cats))

# plot of variable categories
ggplot(data=mHRB_vars_df,
       aes(x = Dim.1, y = Dim.2,label = rownames(mHRB_vars_df))) +
  geom_point(alpha = 0.2) +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_text(aes(colour=Variable), angle = 0) +
  ggtitle("Personality Aspect MCA Plot")
```



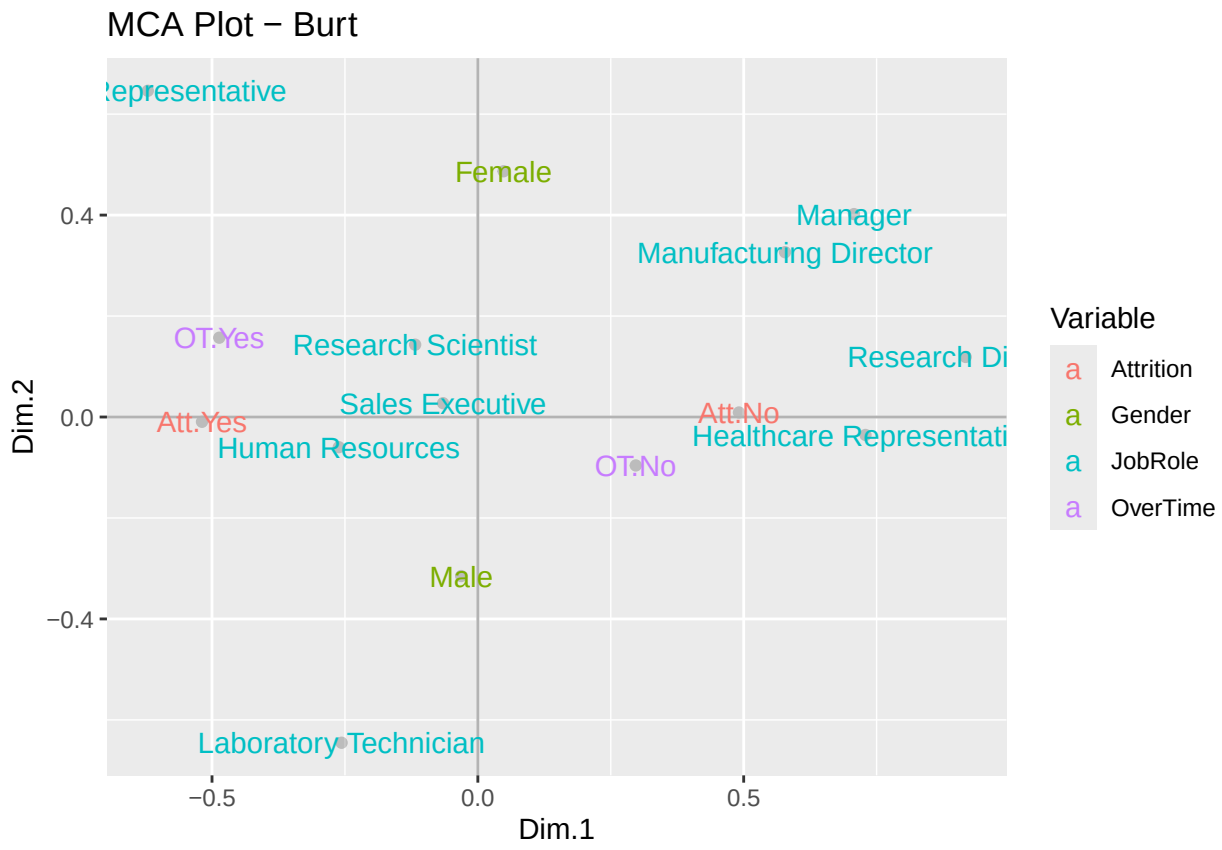

```
# make sure include
newHR <- hrCat[, c("Attrition", "OverTime", "JobRole",
                   "Gender")]

# number of categories per variable
cats <- apply(newHR, 2, function(x) nlevels(as.factor(x)))

# apply MCA using FactoMineR Library
mcaHR.Burt <- FactoMineR::MCA(newHR, method = "Burt", graph = FALSE)

# data frame with variable coordinates
```

```
mHRB_vars_df = data.frame(mcaHR.Burt$var$coord, Variable = rep(names(cats), cats))

# plot of variable categories
ggplot(data=mHRB_vars_df,
       aes(x = Dim.1, y = Dim.2,label = rownames(mHRB_vars_df))) +
  geom_point(alpha = 0.2) +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_text(aes(colour=Variable), angle = 0) +
  ggtitle("MCA Plot - Burt")
```




```
# make sure include
newHR <- hrCat[, c("Attrition", "OverTime", "JobRole",
                   "Gender")]

# number of categories per variable
cats <- apply(newHR, 2, function(x) nlevels(as.factor(x)))

# apply MCA using FactoMineR Library
mcaHR.Burt <- FactoMineR::MCA(newHR, method = "Burt", graph = FALSE)

# data frame with variable coordinates
mHRB_vars_df = data.frame(mcaHR.Burt$var$coord, Variable = rep(names(cats), cats))

# plot of variable categories
```

```
ggplot(data=mHRB_vars_df,
       aes(x = Dim.1, y = Dim.2,label = rownames(mHRB_vars_df))) +
  geom_point(alpha = 0.2) +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_text(aes(colour=Variable), angle = 0) +
  ggtitle("MCA Plot - Burt")
```



Polychoric Correlation

```
############################
# Polychoric Correlation #
############################
# ordinal variables
ordCol <- c("OverTime","BusinessTravel","Education", "EnvironmentSatisfaction",
            "JobInvolvement","JobLevel", "JobSatisfaction","PerformanceRating",
            "RelationshipSatisfaction","StockOptionLevel", "WorkLifeBalance")

# copy all the category variables
hrAllCat <- hr %>% dplyr::select(all_of(catCols))

# Replace Values Based on Condition
hrAllCat$OverTime <-  ifelse(hrAllCat$OverTime == "Yes", 1,2)
hrAllCat$BusinessTravel[hrAllCat$BusinessTravel == "Non-Travel"] <- 1
hrAllCat$BusinessTravel[hrAllCat$BusinessTravel == "Travel_Rarely"] <- 2
hrAllCat$BusinessTravel[hrAllCat$BusinessTravel == "Travel_Frequently"] <- 3
```

```r
# data frame for ordinal
hrOrdinal <- hrAllCat %>% dplyr::select(all_of(ordCol))
hrOrdinal$BusinessTravel <- as.numeric(hrOrdinal$BusinessTravel)

# create ordered factors
hrOrdinal$OverTime <- factor(hrOrdinal$OverTime,
                             levels = c(1,2), ordered = T)

#
hrOrdinal$BusinessTravel <- factor(hrOrdinal$BusinessTravel,
                                   levels = c(1,2,3), ordered = T)

#
hrOrdinal$Education <- factor(hrOrdinal$Education,
                              levels = c(1, 2, 3, 4, 5), ordered = T)
#
hrOrdinal$EnvironmentSatisfaction <- factor(hrOrdinal$EnvironmentSatisfaction,
                                            levels = c(1, 2, 3, 4), ordered = T)

#
hrOrdinal$JobInvolvement <- factor(hrOrdinal$JobInvolvement,
                                   levels = c(1, 2, 3, 4), ordered = T)

#
hrOrdinal$JobLevel <- factor(hrOrdinal$JobLevel,
                             levels = c(1, 2, 3, 4, 5), ordered = T)

#
hrOrdinal$JobSatisfaction <- factor(hrOrdinal$JobSatisfaction,
                                    levels = c(1, 2, 3, 4), ordered = T)

#
hrOrdinal$PerformanceRating <- factor(hrOrdinal$PerformanceRating,
                                      levels = c(3, 4), ordered = T)

#
hrOrdinal$RelationshipSatisfaction <- factor(hrOrdinal$RelationshipSatisfaction,
                                             levels = c(1, 2, 3, 4), ordered = T)

#
hrOrdinal$StockOptionLevel <- factor(hrOrdinal$StockOptionLevel,
                                     levels = c(0, 1, 2, 3), ordered = T)

#
hrOrdinal$WorkLifeBalance <- factor(hrOrdinal$WorkLifeBalance,
                                    levels = c(1, 2, 3, 4), ordered = T)

# ordinal categorical variables
str(hrOrdinal)
```

```
## 'data.frame':    2400 obs. of  11 variables:
##  $ OverTime               : Ord.factor w/ 2 levels "1"<"2": 2 1 2 2 1 2 2 2 2 1 ...
##  $ BusinessTravel         : Ord.factor w/ 3 levels "1"<"2"<"3": 3 3 2 3 2 2 3 2 2 2 ...
```

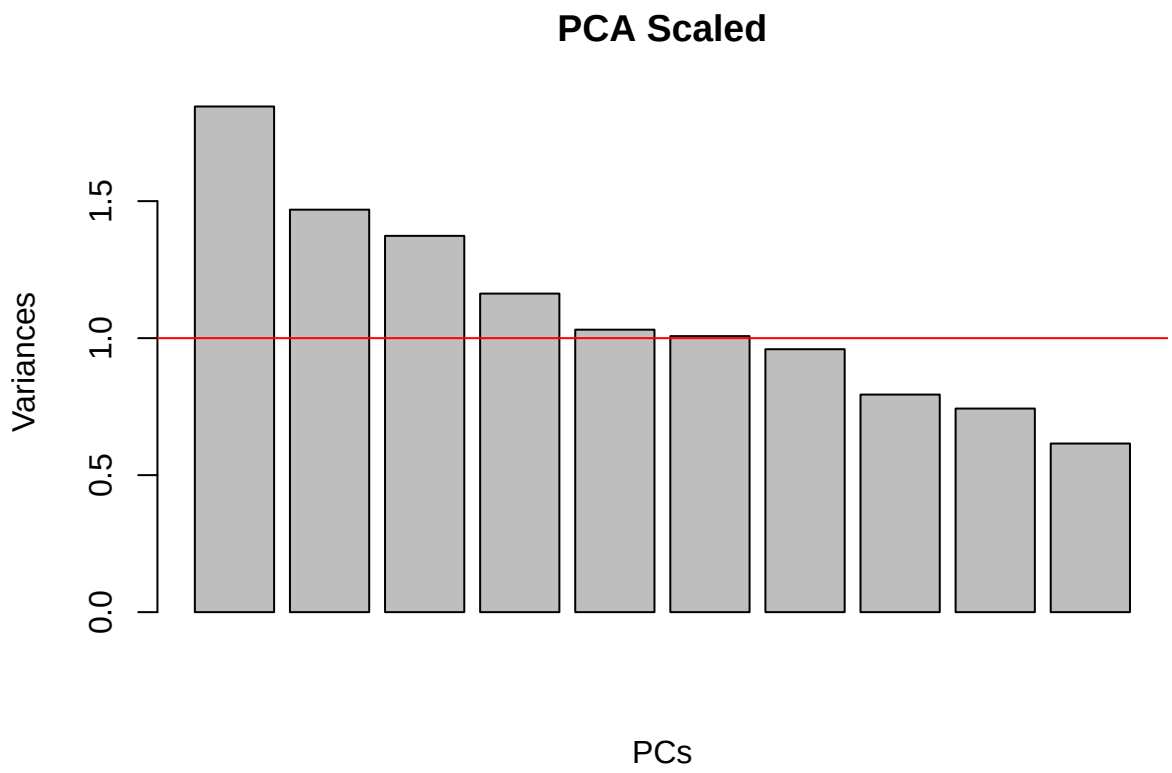
```
##  $ Education               : Ord.factor w/ 5 levels "1"<"2"<"3"<"4"<..: 1 4 1 2 3 1 3 3 3 2 ...
##  $ EnvironmentSatisfaction : Ord.factor w/ 4 levels "1"<"2"<"3"<"4": 3 4 1 4 3 4 4 3 1 4 ...
##  $ JobInvolvement          : Ord.factor w/ 4 levels "1"<"2"<"3"<"4": 2 3 3 3 4 3 2 3 4 2 ...
##  $ JobLevel                : Ord.factor w/ 5 levels "1"<"2"<"3"<"4"<..: 2 1 1 1 1 1 3 2 1 2 ...
##  $ JobSatisfaction         : Ord.factor w/ 4 levels "1"<"2"<"3"<"4": 2 3 2 4 1 3 3 3 2 3 ...
##  $ PerformanceRating       : Ord.factor w/ 2 levels "3"<"4": 2 1 1 1 2 2 2 1 1 1 ...
##  $ RelationshipSatisfaction: Ord.factor w/ 4 levels "1"<"2"<"3"<"4": 4 3 4 3 1 2 2 2 3 4 ...
##  $ StockOptionLevel        : Ord.factor w/ 4 levels "0"<"1"<"2"<"3": 2 1 2 1 4 2 1 3 2 1 ...
##  $ WorkLifeBalance         : Ord.factor w/ 4 levels "1"<"2"<"3"<"4": 3 3 3 2 2 3 3 2 3 3 ...
```

Ordinal Factor Analysis

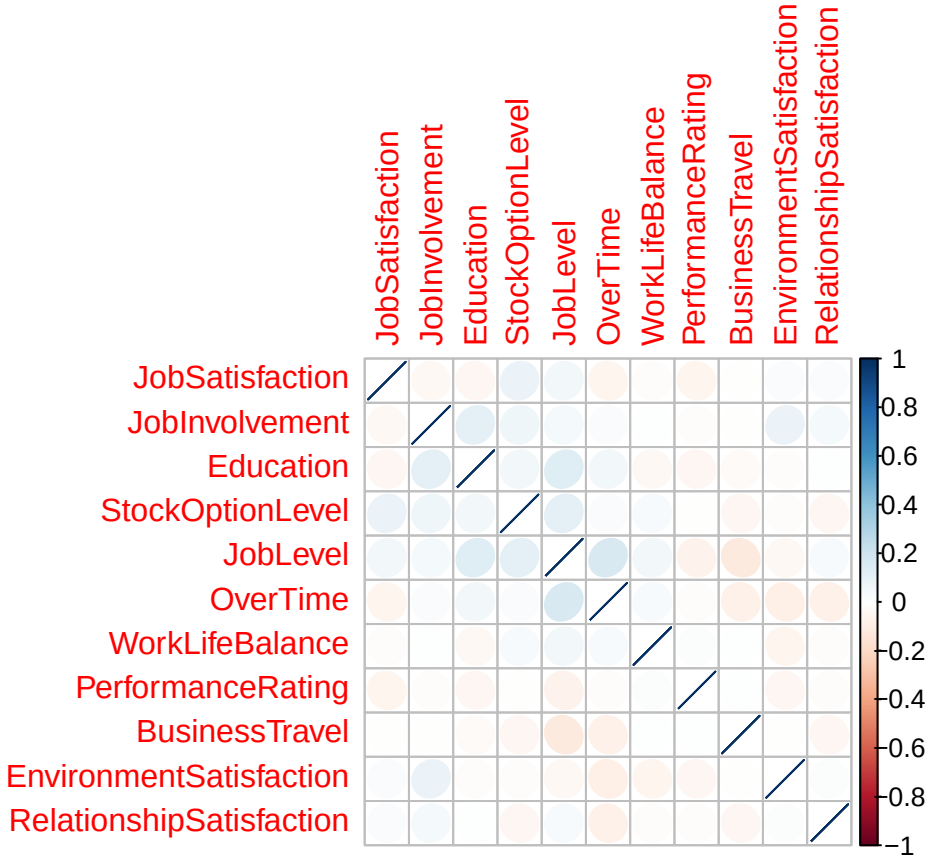
```
# hetcor
ho <- polycor::hetcor(hrOrdinal) %>% suppressWarnings()

# PCA all ordinal variables
po <- prcomp(ho$correlations, scale = TRUE)
# bar scree plot
screeplot(po, type = 'barplot', main = "PCA Scaled") + title(xlab = "PCs")
```

```
## integer(0)
```

```
abline(1, 0, col = "red")
```



```
#
corrplot(ho$correlations, method = "ellipse", order = "AOE")
```




```
# Correlation test for all the categorical variables
# probability correlation on sample size
PCorrTestC <- corr.test(ho$correlations, adjust="none")

# vectorized probability
P <- PCorrTestC$p

# if True probability at a 95% confidence level
PTestC <- ifelse(P < 0.05, T, F)

# how many significant correlations there are for each variable.
colSums(PTestC) - 1
```

```
##                 OverTime           BusinessTravel                Education
##                        0                        0                        0
##  EnvironmentSatisfaction           JobInvolvement                 JobLevel
##                        0                        0                        0
##          JobSatisfaction        PerformanceRating RelationshipSatisfaction
##                        0                        0                        0
##         StockOptionLevel          WorkLifeBalance
##                        0                        0
```

```
# We have to subtract 1 for the diagonal elements (self-correlation)
```

Linear Discriminant Analysis

```
##############################################################################
# Linear Discriminant Analysis Orginal DataFrame ONLY ordinal variables  #
##############################################################################
# copy all the category variables
hrOrdCat <- hrCopy %>% dplyr::select(all_of(catCols))

# bind attrition
hrOrdCat$Attrition <-  hrCopy$Attrition

# convert to factors
hrOrdCat <- lapply(hrOrdCat, factor)

# unable to access class
#fitLDA = MASS::lda(Attrition ~ ., data = hrOrdCat)
# unable to plot out file from LDA
fitLDA = MASS::lda(Attrition ~ ., data = hrOrdCat, CV = TRUE)

# prediction values from LDA
predLDA = fitLDA$class

# confusion matrix
caret::confusionMatrix(data=predLDA, reference = hrOrdCat$Attrition,
                       positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1183  139
##        Yes   50   98
##
##                Accuracy : 0.8714
##                  95% CI : (0.8532, 0.8881)
##     No Information Rate : 0.8388
##     P-Value [Acc > NIR] : 0.0002693
##
##                   Kappa : 0.4396
##
##  Mcnemar's Test P-Value : 1.543e-10
##
##             Sensitivity : 0.41350
##             Specificity : 0.95945
##          Pos Pred Value : 0.66216
##          Neg Pred Value : 0.89486
##              Prevalence : 0.16122
##          Detection Rate : 0.06667
##    Detection Prevalence : 0.10068
##       Balanced Accuracy : 0.68648
##
```

```
##         'Positive' Class : Yes
##
```