Employee Attrition Analysis

Eunmi Jeong

Sep 20, 2020

1. Data preprocessing, cleaing, transformation

- Check the structure of dataset

```
str(data) #1176 obs. of  35 variables
```

- Missing value detection and manipulation

```
sum(is.na(data)) #9
data2 <- na.omit(data)
which(data2$Gender=="") #Missing value in 'Gender' column
data2 <- data2[-1053,]
which(data2$OverTime=="") #Missing value in 'OverTime' column
data2 <- data2[-1004,]
rownames(data2) <- NULL #1165 obs. of  35 variables
```

- Identify and remove low or no variance

```
caret::nearZeroVar(data2, saveMetrics = T) #EmployeeCount, Over18, StandardHours
nzv <- nearZeroVar(data2)
data2 <- data2[,-nzv]
```

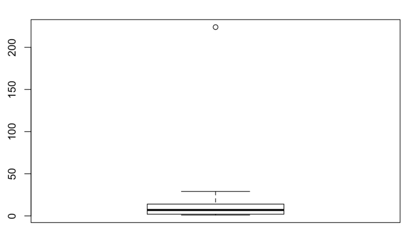
- Check duplicate data record

```
nrow(data2[!duplicated(data2), ]) #no duplicates
```

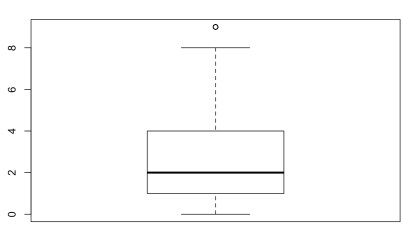
```
[1] 1165
```

- Identify and remove outliers or noises
    - DistanceFromHome
    - NumCompaniesWorked
    - TotalWorkingYears
    - YearsAtCompany
    - YearsInCurrentRole
    - YearsWithCurrentManager

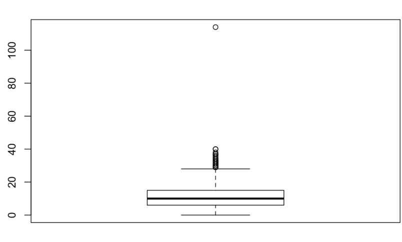
```
#DistanceFromHome
boxplot(data2$DistanceFromHome)
```



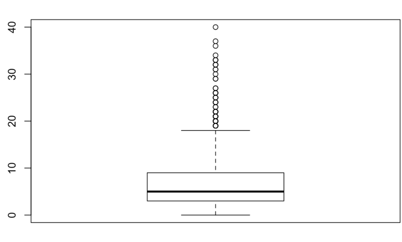
```
data2$DistanceFromHome[data2$DistanceFromHome %in% boxplot.stats(data2$DistanceFromHome)$out] <- median(data2$DistanceFromHome, na.rm = T)
#Number of companies worked
boxplot(data2$NumCompaniesWorked)
```



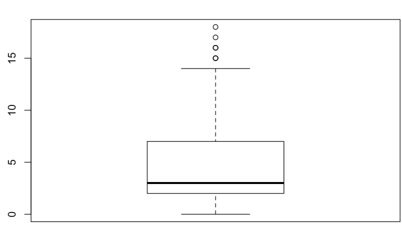
```
data2$NumCompaniesWorked[data2$NumCompaniesWorked %in% boxplot.stats(data2$NumCompaniesWorked)$out] <- median(data2$NumCompaniesWorked, na.rm = T)
#Total working years
boxplot(data2$TotalWorkingYears)
```



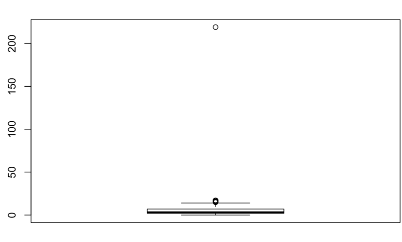
```
data2$TotalWorkingYears[data2$TotalWorkingYears %in% boxplot.stats(data2$TotalWorkingYears)$out] <- median(data2$TotalWorkingYears, na.rm = T)
#Years at current company
boxplot(data2$YearsAtCompany)
```



```
data2$YearsAtCompany[data2$YearsAtCompany %in% boxplot.stats(data2$YearsAtCompany)$out] <- median(data2$YearsAtCompany, na.rm = T)
#YearsInCurrentRole
boxplot(data2$YearsInCurrentRole)
```



```
data2$YearsInCurrentRole[data2$YearsInCurrentRole %in% boxplot.stats(data2$YearsInCurrentRole)$out] <- median(data2$YearsInCurrentRole, na.rm = T)
#YearsWithCurrentManager
boxplot(data2$YearsWithCurrManager)
```



```
data2$YearsWithCurrManager[data2$YearsWithCurrManager %in% boxplot.stats(data2$YearsWithCurrManager)$out] <- median(data2$YearsWithCurrManager, na.rm = T)
```

- Discretization
    - Age
    - Daily rate
    - Distance from home
    - Years at current company
    - Years with current manager

```
#Age
data2$Age_grp <- arules::discretize(data2$Age, method = "frequency", breaks = 3, labels = c("young", "middle", "old"))
data2 %>%
  group_by(Age_grp) %>%
  summarise(avg_Age = mean(Age), count = n(), min = min(Age), max = max(Age))
```

```
# A tibble: 3 x 5
  Age_grp avg_Age count   min   max
  <fct>     <dbl> <int> <int> <int>
1 young      27.0   355    18    31
2 middle     35.4   391    32    39
3 old        46.8   419    40    60
```

```
#Daily rate
data2$DailyRate_grp <- arules::discretize(data2$DailyRate, method = "frequency", breaks = 3, labels = c("low", "medium", "high"))
data2 %>%
  group_by(DailyRate_grp) %>%
  summarise(avg_Rate = mean(DailyRate), count = n(), min = min(DailyRate), max = max(DailyRate))
```

```
# A tibble: 3 x 5
  DailyRate_grp avg_Rate count   min   max
  <fct>            <dbl> <int> <int> <int>
1 low               337.   387   102   572
2 medium            793.   388   574  1038
3 high             1270.   390  1040  1499
```

```
#DistanceFromHome
data2$Dist_grp <- arules::discretize(data2$DistanceFromHome, method = "frequency", breaks = 3, labels = c("close", "medium", "far"))
data2 %>%
  group_by(Dist_grp) %>%
  summarise(avg_dist = mean(DistanceFromHome), count = n(), min = min(DistanceFromHome), max = max(DistanceFromHome))
```

```
# A tibble: 3 x 5
  Dist_grp avg_dist count   min   max
  <fct>       <dbl> <int> <int> <int>
1 close        1.78   387     1     3
2 medium       6.75   356     4     9
3 far         18.4    422    10    29
```

```
#Years at current company
data2$WorkYrs_grp <- arules::discretize(data2$YearsAtCompany, method = "frequency", breaks = 3, labels = c("short", "medium", "long"))
data2 %>%
  group_by(WorkYrs_grp) %>%
  summarise(avg_dist = mean(YearsAtCompany), count = n(), min = min(YearsAtCompany), max = max(YearsAtCompany))
```

```
# A tibble: 3 x 5
  WorkYrs_grp avg_dist count   min   max
  <fct>          <dbl> <int> <int> <int>
1 short           1.74   375     0     3
2 medium          4.91   395     4     6
3 long            9.92   395     7    18
```

```
#YearsInCurrentRole
data2$RoleYrs_grp <- arules::discretize(data2$YearsInCurrentRole, method = "frequency", breaks = 3, labels = c("short", "medium", "long"))
data2 %>%
  group_by(RoleYrs_grp) %>%
  summarise(avg_dist = mean(YearsInCurrentRole), count = n(), min = min(YearsInCurrentRole), max = max(YearsInCurrentRole))
```

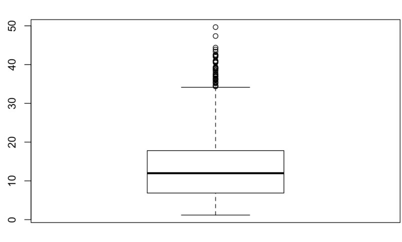
```
# A tibble: 3 x 5
  RoleYrs_grp avg_dist count   min   max
  <fct>          <dbl> <int> <int> <int>
1 short          0.201   234     0     1
2 medium         2.58    515     2     4
3 long           7.85    416     5    14
```

```
#YearsWithCurrentManager
data2$CurrManagerYrs_grp <- arules::discretize(data2$YearsWithCurrManager, method = "frequency", breaks = 3, labels = c("short", "medium", "long"))
data2 %>%
  group_by(CurrManagerYrs_grp) %>%
  summarise(avg_dist = mean(YearsWithCurrManager), count = n(), min = min(YearsWithCurrManager), max = max(YearsWithCurrManager))
```

```
# A tibble: 3 x 5
  CurrManagerYrs_grp avg_dist count   min   max
  <fct>                 <dbl> <int> <int> <int>
1 short                 0.238   269     0     1
2 medium                2.57    493     2     4
3 long                  7.97    403     5    14
```

- Adding Variables
    - Daily working hours

```
data2 <- data2 %>% mutate(d.work.hrs=DailyRate/HourlyRate ) #Daily working hours
boxplot(data2$d.work.hrs) #Identify&remove outlier
```



```
data2$d.work.hrs[data2$d.work.hrs %in% boxplot.stats(data2$d.work.hrs)$out] <- median(data2$d.work.hrs, na.rm = T)
```

- Shorten variable name and value labels
    - Department

```
data2$Dep.s <- data2[,5]
data2$Dep.s <- str_replace_all(data2$Dep.s,"Research & Development", "R&D")
data2$Dep.s <- str_replace_all(data2$Dep.s,"Human Resources", "HR")
```

- Change Variable Type (Numerical-Categorical)
    - Employee number (integer-charactor)
    - Education (integer-factor)
    - EnvironmentSatisfaction (integer-factor)
    - JobLevel (integer-factor)
    - JobSatisfaction (integer-factor)
    - StockOptionLevel (integer-factor)
    - RelationshipSatisfaction (integer-factor)
    - WorkLifeBalance (integer-factor)
    - PerformanceRating (integer-factor)

```
data2$EmployeeNumber <- as.character(data2$EmployeeNumber)
data2$Education <- as.factor(data2$Education)
data2$EnvironmentSatisfaction <- as.factor(data2$EnvironmentSatisfaction)
data2$JobLevel <- as.factor(data2$JobLevel)
data2$JobSatisfaction <- as.factor(data2$JobSatisfaction)
data2$StockOptionLevel <- as.factor(data2$StockOptionLevel)
data2$RelationshipSatisfaction <- as.factor(data2$RelationshipSatisfaction)
data2$WorkLifeBalance <- as.factor(data2$WorkLifeBalance)
data2$PerformanceRating <- as.factor(data2$PerformanceRating)
```

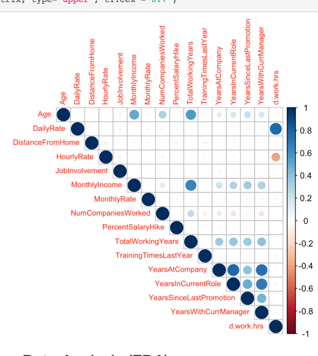
- Missing Value Imputation and Scaling

```
preprocess <- caret::preProcess(data2, method = c("knnImpute", "center", "scale"))
data3 <- predict(preprocess, data2)
sum(!complete.cases(data3))
```

```
[1] 0
```

- Detect Attribute's Correlation (positive)
    - Age-MonthlyIncome
    - Age-TotalWorkingYears
    - MonthlyIncome-TotalWorkingYears
    - PercentSalaryHike-PerformanceRating
    - YearsAtCompany-YearsInCurrentRole
    - YearsAtCompany-YearsWithCurrentManager
    - DailyWorkingHours-DailyRate
- Detect Attribute's Correlation (negative)
    - DailyWorkingHours-HourlyRate

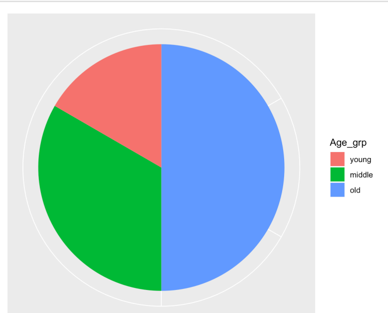
```
cor_matrix <- cor(data3[complete.cases(data3), sapply(data3, is.numeric)], method = "pearson")
corrplot(cor_matrix, type="upper", tl.cex = 0.7 )
```



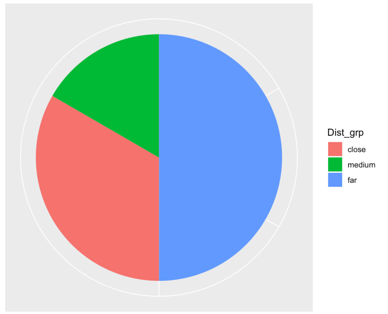
2. Exploratory Data Analysis (EDA)

- Pie Chart
    - Age group: old(40-60) group is the biggest
    - Distance from home group: far(10-29) group is the biggest
    - Gender group: male group is bigger than female group
    - Department group: R&D group is the biggest

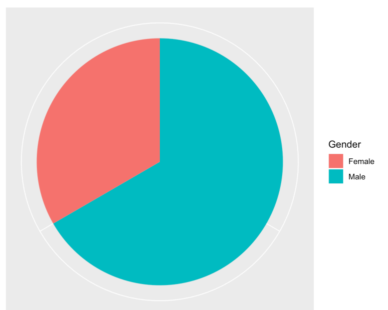
```
# Age group
data3 %>%  group_by(Age_grp) %>%
  summarise(pct = percent(n()/nrow(data3))) %>%
  ggplot(aes(x = factor(1), y = pct, fill = Age_grp)) +
  geom_bar(stat = "identity") +
  coord_polar(theta = "y") +
  theme(axis.ticks = element_blank(),
        axis.text = element_blank(),
        axis.title = element_blank())
```



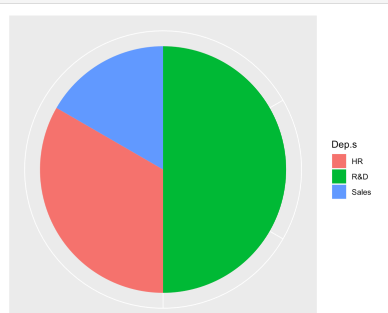

```
# Distance from home group
data3 %>%  group_by(Dist_grp) %>%
  summarise(pct = percent(n()/nrow(data3))) %>%
  ggplot(aes(x = factor(1), y = pct, fill = Dist_grp)) +
  geom_bar(stat = "identity") +
  coord_polar(theta = "y") +
  theme(axis.ticks = element_blank(),
        axis.text = element_blank(),
        axis.title = element_blank())
```




```
# Gender group
data3 %>%  group_by(Gender) %>%
  summarise(pct = percent(n()/nrow(data3))) %>%
  ggplot(aes(x = factor(1), y = pct, fill = Gender)) +
  geom_bar(stat = "identity") +
  coord_polar(theta = "y") +
  theme(axis.ticks = element_blank(),
        axis.text = element_blank(),
        axis.title = element_blank())
```

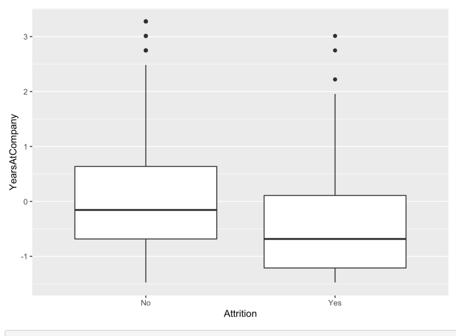



```
 # Department group
data3 %>%  group_by(Dep.s) %>%
  summarise(pct = percent(n()/nrow(data3))) %>%
  ggplot(aes(x = factor(1), y = pct, fill = Dep.s)) +
  geom_bar(stat = "identity") +
  coord_polar(theta = "y") +
  theme(axis.ticks = element_blank(),
        axis.text = element_blank(),
        axis.title = element_blank())
```

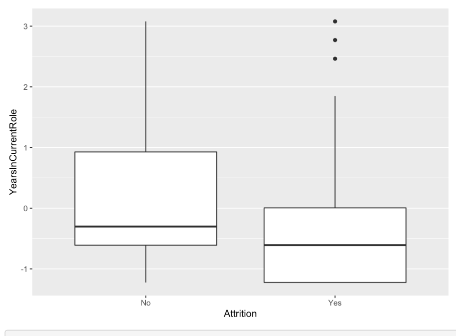


- Boxplot
    - Attrition-YearsAtCompany: people who leave work shorter at the company than those who stay
    - Attrition-YearsInCurrentRole: people who leave work shorter in their roles than those who stay
    - Attrition-YearsWithCurrManager: people who leave work shorter under current manager than those who stay
    - Attrition-DailyRate: people who leave work are paid less than those who stay
    - Dep.s-Monthlyincome: Sales department shows the highest monthly income
    - Education-Monthlyincome: people in higher education level are paid more
    - Age-grp-Monthlyincome: old people(40-60) are paid more

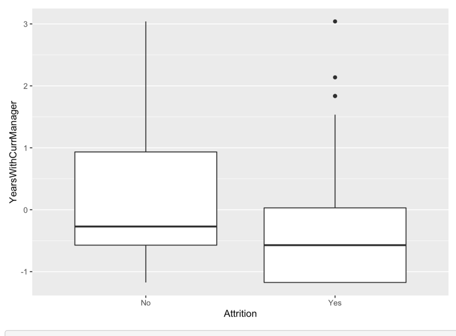
```
#attrition-YearsAtCompany
ggplot(data3, aes(x = Attrition, y = YearsAtCompany)) +
  geom_boxplot() +
  theme(panel.grid.major.x = element_blank())
```



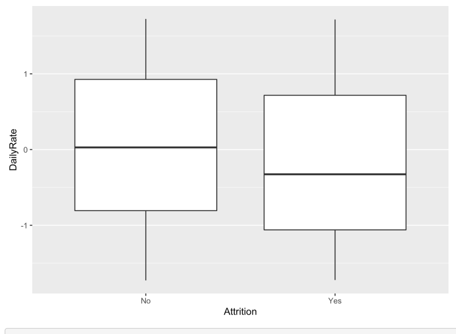

```
#attrition-YearsInCurrentRole
ggplot(data3, aes(x = Attrition, y = YearsInCurrentRole)) +
  geom_boxplot() +
  theme(panel.grid.major.x = element_blank())
```



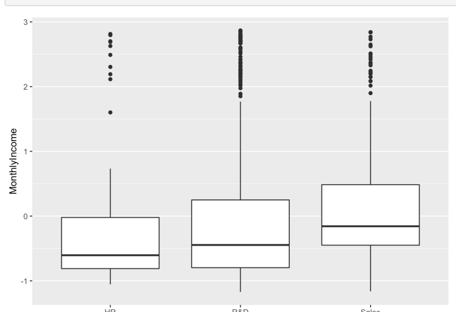

```
#attrition-YearsWithCurrManager
ggplot(data3, aes(x = Attrition, y = YearsWithCurrManager)) +
  geom_boxplot() +
  theme(panel.grid.major.x = element_blank())
```



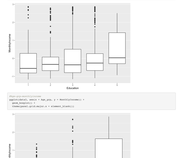

```
#attrition-daily rate
ggplot(data3, aes(x = Attrition, y = DailyRate)) +
  geom_boxplot() +
  theme(panel.grid.major.x = element_blank())
```



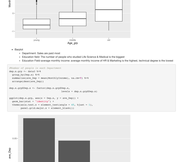

```
#Dep.s-monthlyincome
ggplot(data3, aes(x = Dep.s, y = MonthlyIncome)) +
  geom_boxplot() +
  theme(panel.grid.major.x = element_blank())
```



```
Warning in ggplot(data, aes(x = Education, y = MaritalStatusNum)) +
  geom_violin(fill = "#ffc0cb") +
  stackoverflow::geom_boxplot(width = 0.1, fill = "#ffffff") +
  coord_flip():
```

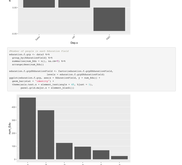



```
Warning: Ignoring unknown aesthetics: fill
ggplot(data, aes(x = Age, y = Education)) +
  geom_violin(fill = "#ffc0cb") +
  stackoverflow::geom_boxplot(width = 0.1, fill = "#ffffff") +
  coord_flip():
```

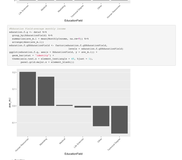


- Factor:
  - EducationLevel: Dates are used most
  - EducationTitle: The college degree who studied Life Sciences & Medical is the biggest
  - EducationField average monthly income: average monthly income of HR & Marketing is the highest, technical degree is the lowest

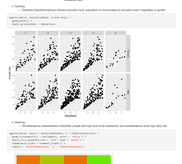
```
Warning: Ignoring unknown aesthetics: fill
ggplot(data, aes(x = Education, y = DailyRate)) +
  geom_violin(fill = "#ffc0cb") +
  stackoverflow::geom_boxplot(width = 0.1, fill = "#ffffff") +
  coord_flip(ylim = c(0, 1500)):
Warning: Coordinate system already present. Adding new coordinate system, which will replace the existing one.
```



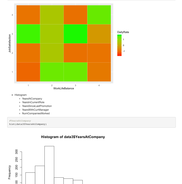

```
Warning: Ignoring unknown aesthetics: fill
ggplot(data, aes(x = DailyRate, y = Education)) +
  geom_violin(fill = "#ffc0cb") +
  stackoverflow::geom_boxplot(width = 0.1, fill = "#ffffff") +
  coord_flip(xlim = c(0, 1500)):
Warning: Coordinate system already present. Adding new coordinate system, which will replace the existing one.
```




```
Warning: Ignoring unknown aesthetics: fill
ggplot(data, aes(x = EducationField, y = MonthlyIncome)) +
  geom_violin(fill = "#ffc0cb") +
  stackoverflow::geom_boxplot(width = 0.1, fill = "#ffffff") +
  coord_flip():
```

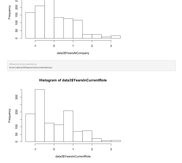



```
Warning: Ignoring unknown aesthetics: fill
ggplot(data, aes(x = MonthlyIncome, y = EducationField)) +
  geom_violin(fill = "#ffc0cb") +
  stackoverflow::geom_boxplot(width = 0.1, fill = "#ffffff") +
  coord_flip(xlim = c(0, 15000)):
Warning: Coordinate system already present. Adding new coordinate system, which will replace the existing one.
```



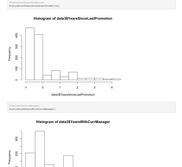
- Feeding
  - DailyRate, DailyRate/HourlyRate,Gender,Education level ,population is concentrated on education level 3 (regardless of gender)

```
ggplot(data=data, aes(Education, d.with.Gender,)
  + geom_violin()
  + theme_gray(base_size = 18)+ facet_wrap(Gender)
```



- Feeding
  - Education,Gender,Hourlyrate,Dailyrate,Jobsatisfaction people who work longer in a way shows higher daily rate and job satisfaction even though hourly rate is lower than others

```
ggplot(data=data, aes(d.with.Gender, d.with.Education,)
  + geom_violin(aes(fill = d.with.Gender))
  + geom_boxplot(width = 0.1)
```



3. Association Rules Mining

1. Extract entity, rules which and display association between LHS & RHS are identified as (1) (Attrition Rate)

```
dataset_ord <- dataset[,c("BusinessTravel","Education","Gender","JobLevel","JobRole","JobSatisfaction","MaritalStatus","OverTime","PerformanceRating","RelationshipSatisfaction","StockOptionLevel","WorkLifeBalance","Attrition")]
dataset_ord[,c("Education","JobLevel","JobSatisfaction","PerformanceRating","RelationshipSatisfaction","StockOptionLevel","WorkLifeBalance")] <- lapply(dataset_ord[,c("Education","JobLevel","JobSatisfaction","PerformanceRating","RelationshipSatisfaction","StockOptionLevel","WorkLifeBalance")],function(x) paste0(colnames(dataset_ord)[which(colnames(dataset_ord)==paste(substitute(x),collapse=""))],x))
dataset_ord[,c("Education","JobLevel","JobSatisfaction","PerformanceRating","RelationshipSatisfaction","StockOptionLevel","WorkLifeBalance")] <- lapply(dataset_ord[,c("Education","JobLevel","JobSatisfaction","PerformanceRating","RelationshipSatisfaction","StockOptionLevel","WorkLifeBalance")],function(x) factor(x,levels=unique(x)))
dataset_ord[,c("Attrition")] <- lapply(dataset_ord[,c("Attrition")],function(x) factor(x,levels=c("No","Yes")))
dataset_ord[,c("Attrition")] <- lapply(dataset_ord[,c("Attrition")],function(x) ifelse(x=="No",0,1))
dataset_ord[,c("Attrition")] <- lapply(dataset_ord[,c("Attrition")],function(x) factor(x,levels=c(0,1)))
dataset_ord[,c("Attrition")] <- lapply(dataset_ord[,c("Attrition")],function(x) paste0(colnames(dataset_ord)[which(colnames(dataset_ord)==paste(substitute(x),collapse=""))],x))
dataset_ord[,c("Attrition")] <- lapply(dataset_ord[,c("Attrition")],function(x) factor(x,levels=unique(x)))
dataset_ord[,c("Attrition")] <- lapply(dataset_ord[,c("Attrition")],function(x) factor(x,levels=rev(levels(x))))
dataset_ord[,c("Attrition")] <- lapply(dataset_ord[,c("Attrition")],function(x) factor(x,levels=c("Attrition1","Attrition0")))
dataset_ord[,c("Attrition")] <- lapply(dataset_ord[,c("Attrition")],function(x) factor(x,levels=c("Attrition1
```