capstone\_WIP.R

acer

Wed Jun 13 14:36:03 2018

########################## Importing necessary libraries and Packages ##############  
library(data.table) # used for reading and manipulation of data  
library(dplyr) # used for data manipulation and joining

library(ggplot2) # used for ploting  
library(lattice) # used as dependency with caret  
library(caret) # used for modeling  
library(corrplot) # used for making correlation plot

## corrplot 0.84 loaded

library(cowplot) # used for combining multiple plots

library(caTools) # used for train, test spli  
library(MASS) # used for stepAIC

library(ROCR) # used for ROCR metrics

library(lift) # used for lift and gain charts  
library(car) # used for vif function

library(woe) # used for imputing missing values  
  
creditBureauData <- read.csv("Credit Bureau data.csv", stringsAsFactors = F)  
demoData <- read.csv("Demographic data.csv", stringsAsFactors = F)  
  
#### Data Definition ####  
### Demographic Data ###  
# Application ID: Unique ID of the customers  
# Age: Age of customer  
# Gender: Gender of customer  
# Marital Status: Marital status of customer (at the time of application)  
# No of dependents: No. of childrens of customers  
# Income: Income of customers  
# Education: Education of customers  
# Profession: Profession of customers  
# Type of residence: Type of residence of customers  
# No of months in current residence: No of months in current residence of customers  
# No of months in current company: No of months in current company of customers  
# Performance Tag: Status of customer performance (1 represents "Default")  
  
########## Data Quality Issues ###########  
"There are 3 duplicates for Application ID, both in credit and Demographic data (same IDs).  
There are 1425 istances, where Performance Tag is not available, so, shall we remove them.  
Missing Value Imputation Strategy (While not using woe values)  
"

## [1] "There are 3 duplicates for Application ID, both in credit and Demographic data (same IDs).\nThere are 1425 istances, where Performance Tag is not available, so, shall we remove them.\nMissing Value Imputation Strategy (While not using woe values)\n"

##########################################  
  
  
################### Basic Data Quality Checks and Understanding (For Demographic Data) #############################  
head(demoData)

## Application.ID Age Gender Marital.Status..at.the.time.of.application.  
## 1 954457215 48 F Married  
## 2 432830445 31 M Married  
## 3 941387308 32 M Single  
## 4 392161677 43 M Married  
## 5 182011211 35 F Married  
## 6 312196805 20 M Married  
## No.of.dependents Income Education Profession Type.of.residence  
## 1 2 40 Bachelor SAL Rented  
## 2 4 55 Professional SE\_PROF Rented  
## 3 2 46 Bachelor SE\_PROF Rented  
## 4 1 53 Bachelor SE Rented  
## 5 5 44 Professional SAL Rented  
## 6 1 39 Bachelor SAL   
## No.of.months.in.current.residence No.of.months.in.current.company  
## 1 113 56  
## 2 112 46  
## 3 104 49  
## 4 94 53  
## 5 112 43  
## 6 116 52  
## Performance.Tag  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0

str(demoData)

## 'data.frame': 71295 obs. of 12 variables:  
## $ Application.ID : int 954457215 432830445 941387308 392161677 182011211 312196805 532217204 74788849 782743811 96964957 ...  
## $ Age : int 48 31 32 43 35 20 42 34 30 22 ...  
## $ Gender : chr "F" "M" "M" "M" ...  
## $ Marital.Status..at.the.time.of.application.: chr "Married" "Married" "Single" "Married" ...  
## $ No.of.dependents : int 2 4 2 1 5 1 2 2 3 1 ...  
## $ Income : num 40 55 46 53 44 39 55 49 48 38 ...  
## $ Education : chr "Bachelor" "Professional" "Bachelor" "Bachelor" ...  
## $ Profession : chr "SAL" "SE\_PROF" "SE\_PROF" "SE" ...  
## $ Type.of.residence : chr "Rented" "Rented" "Rented" "Rented" ...  
## $ No.of.months.in.current.residence : int 113 112 104 94 112 116 104 108 115 111 ...  
## $ No.of.months.in.current.company : int 56 46 49 53 43 52 41 40 58 57 ...  
## $ Performance.Tag : int 0 0 0 0 0 0 0 0 0 0 ...

dim(demoData) # 71295\*12

## [1] 71295 12

sum(is.na(demoData)) ## 1428 nulls in total

## [1] 1428

colSums(is.na(select\_if(demoData, colSums(is.na(demoData))>0))) ## Column having nulls

## No.of.dependents Performance.Tag   
## 3 1425

## The predictor Variable PerformanceTag has nulls, removing the records (~2%)  
names(demoData)

## [1] "Application.ID"   
## [2] "Age"   
## [3] "Gender"   
## [4] "Marital.Status..at.the.time.of.application."  
## [5] "No.of.dependents"   
## [6] "Income"   
## [7] "Education"   
## [8] "Profession"   
## [9] "Type.of.residence"   
## [10] "No.of.months.in.current.residence"   
## [11] "No.of.months.in.current.company"   
## [12] "Performance.Tag"

## Few column names are big, renaming the same  
colnames(demoData) <- c("ApplicationID","Age","Gender","MaritalStatus","Dependents","Income","Education","Profession",  
 "TypeofResidence","currResidenceTenure","currJobTenure","PerformanceTag")  
names(demoData) ## Checking on the exisitng columns names

## [1] "ApplicationID" "Age" "Gender"   
## [4] "MaritalStatus" "Dependents" "Income"   
## [7] "Education" "Profession" "TypeofResidence"   
## [10] "currResidenceTenure" "currJobTenure" "PerformanceTag"

demoData <- demoData[is.na(demoData$PerformanceTag)==F,]  
sapply(demoData, function(x)  
 length(which(x == ""))) # checking for blank "" values; there are lots

## ApplicationID Age Gender   
## 0 0 2   
## MaritalStatus Dependents Income   
## 6 0 0   
## Education Profession TypeofResidence   
## 118 13 8   
## currResidenceTenure currJobTenure PerformanceTag   
## 0 0 0

sum(duplicated(demoData$ApplicationID)) # 3 duplicates

## [1] 3

demoData[duplicated(demoData$ApplicationID)==TRUE,]$ApplicationID

## [1] 765011468 653287861 671989187

creditBureauData[duplicated(creditBureauData$Application.ID)==TRUE,]$Application.ID # We have duplicates for

## [1] 765011468 653287861 671989187

# the same Application ID  
  
  
################### EDA and Data Cleansing (For Demographic Data) #######################  
sapply(demoData, function(x) length(unique(x))) ## Checking the unique values present in the Dataset

## ApplicationID Age Gender   
## 69867 53 3   
## MaritalStatus Dependents Income   
## 3 6 63   
## Education Profession TypeofResidence   
## 6 4 6   
## currResidenceTenure currJobTenure PerformanceTag   
## 121 83 2

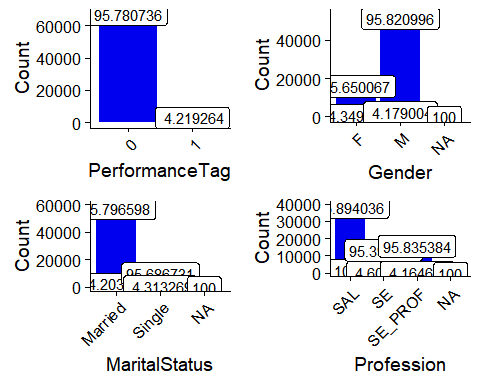
sapply(demoData[,c(3,4,5,7,8,9,12)], function(x) unique(x)) ## Looking out for unique values in specific columns

## $Gender  
## [1] "F" "M" ""   
##   
## $MaritalStatus  
## [1] "Married" "Single" ""   
##   
## $Dependents  
## [1] 2 4 1 5 3 NA  
##   
## $Education  
## [1] "Bachelor" "Professional" "Masters" "Phd"   
## [5] "Others" ""   
##   
## $Profession  
## [1] "SAL" "SE\_PROF" "SE" ""   
##   
## $TypeofResidence  
## [1] "Rented" "" "Owned"   
## [4] "Others" "Living with Parents" "Company provided"   
##   
## $PerformanceTag  
## [1] 0 1

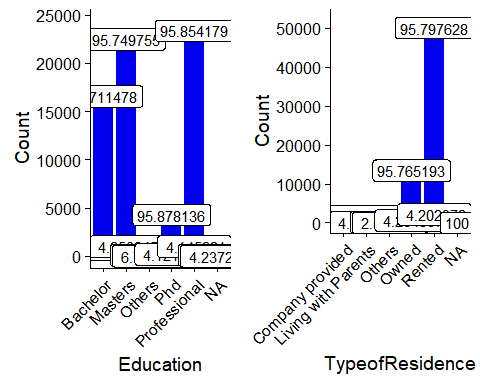
## Converting all the empty strings with NAs ##  
demoData <- as.data.table(apply(demoData, 2, function(x) ifelse(x %in% c("", " ", "NA"), NA, x)))  
sapply(demoData, function(x) length(which(x == ""))) ## Checking, if all the "" has been replaced with NA

## ApplicationID Age Gender   
## 0 0 0   
## MaritalStatus Dependents Income   
## 0 0 0   
## Education Profession TypeofResidence   
## 0 0 0   
## currResidenceTenure currJobTenure PerformanceTag   
## 0 0 0

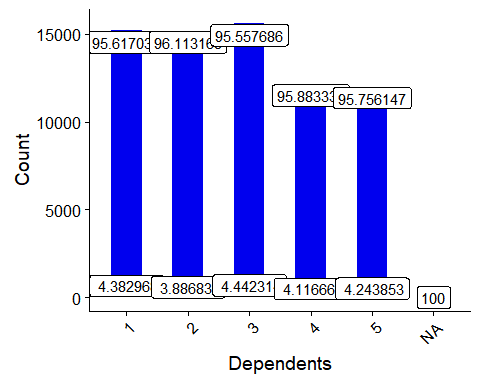
## Converting the required columns to numerical type  
demoData$Age <- as.numeric(demoData$Age)  
demoData$Income <- as.numeric(demoData$Income)  
demoData$currResidenceTenure <- as.numeric(demoData$currResidenceTenure)  
demoData$currJobTenure <- as.numeric(demoData$currJobTenure)  
  
## Univariate Analysis can say a lot, checking, if there are any   
p1 <- ggplot(demoData %>% group\_by(PerformanceTag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(PerformanceTag, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(PerformanceTag, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
p2 <- ggplot(demoData %>% group\_by(Gender, PerformanceTag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(Gender, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(Gender, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
p3 <- ggplot(demoData %>% group\_by(MaritalStatus, PerformanceTag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(MaritalStatus, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(MaritalStatus, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
p4 <- ggplot(demoData %>% group\_by(Education, PerformanceTag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(Education, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(Education, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
p5 <- ggplot(demoData %>% group\_by(Profession, PerformanceTag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(Profession, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(Profession, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
p6 <- ggplot(demoData %>% group\_by(TypeofResidence, PerformanceTag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(TypeofResidence, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(TypeofResidence, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
plot\_grid(p1,p2,p3,p5)



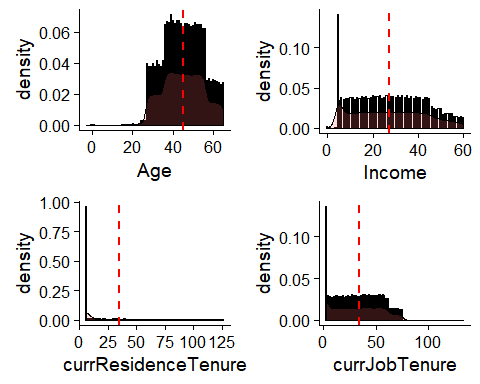
plot\_grid(p4,p6)



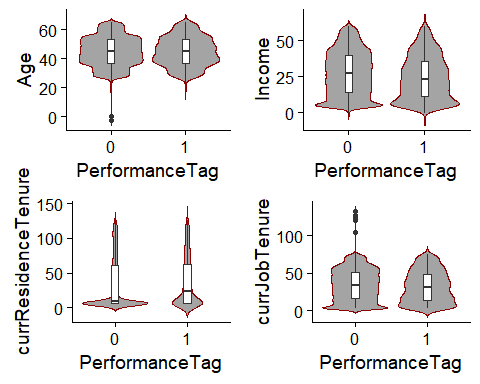
ggplot(demoData %>% group\_by(Dependents, PerformanceTag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(Dependents, Count), stat = "identity", fill = "blue2", width = .5) +  
 geom\_label(aes(Dependents, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



## There seems to be data quality issues, in here, which needs to be fixed  
## For few of the columns, there exists blank values, which needs to be taken care of  
  
## The target Variable PerformanceTag is unbalanced. 0: 66922; 1: 2948; So, ratio is 1:22. Will have to use different sampling techniques  
## to balance the data  
  
## Visualisations for the continuous Variables. Looking into the distribution of the variable  
p1 <- ggplot(demoData, aes(x=Age)) + geom\_histogram(aes(y=..density..), binwidth=.5,  
 colour="black", fill="white") + geom\_density(alpha=.2, fill="#FF6666") +  
 geom\_vline(aes(xintercept=mean(Age, na.rm=T)), color="red", linetype="dashed", size=1)  
  
p2 <- ggplot(demoData, aes(x=Income)) + geom\_histogram(aes(y=..density..), binwidth=.5,  
 colour="black", fill="white") + geom\_density(alpha=.2, fill="#FF6666") +  
 geom\_vline(aes(xintercept=mean(Income, na.rm=T)), color="red", linetype="dashed", size=1)  
  
p3 <- ggplot(demoData, aes(x=currResidenceTenure)) + geom\_histogram(aes(y=..density..), binwidth=.5,  
 colour="black", fill="white") + geom\_density(alpha=.2, fill="#FF6666") +  
 geom\_vline(aes(xintercept=mean(currResidenceTenure, na.rm=T)), color="red", linetype="dashed", size=1)  
  
p4 <- ggplot(demoData, aes(x=currJobTenure)) + geom\_histogram(aes(y=..density..), binwidth=.5,  
 colour="black", fill="white") + geom\_density(alpha=.2, fill="#FF6666") +  
 geom\_vline(aes(xintercept=mean(currJobTenure, na.rm=T)), color="red", linetype="dashed", size=1)  
  
plot\_grid(p1,p2,p3,p4)



## Bivariate Analysis using Violin + Box plots  
p1 <- ggplot(demoData, aes(x=PerformanceTag, y=Age)) + geom\_violin(trim=FALSE, fill='#A4A4A4', color="darkred") + geom\_boxplot(width=0.1)  
p2 <- ggplot(demoData, aes(x=PerformanceTag, y=Income)) + geom\_violin(trim=FALSE, fill='#A4A4A4', color="darkred") + geom\_boxplot(width=0.1)  
p3 <- ggplot(demoData, aes(x=PerformanceTag, y=currResidenceTenure)) + geom\_violin(trim=FALSE, fill='#A4A4A4', color="darkred") + geom\_boxplot(width=0.1)  
p4 <- ggplot(demoData, aes(x=PerformanceTag, y=currJobTenure)) + geom\_violin(trim=FALSE, fill='#A4A4A4', color="darkred") + geom\_boxplot(width=0.1)  
  
plot\_grid(p1,p2,p3,p4)



############################ Treating Data Quality Issues #############################  
woe\_data <- demoData  
names(woe\_data)

## [1] "ApplicationID" "Age" "Gender"   
## [4] "MaritalStatus" "Dependents" "Income"   
## [7] "Education" "Profession" "TypeofResidence"   
## [10] "currResidenceTenure" "currJobTenure" "PerformanceTag"

#### Imputing with normal Methodologies #########  
colSums(is.na(select\_if(demoData, colSums(is.na(demoData))>0)))

## Gender MaritalStatus Dependents Education   
## 2 6 3 118   
## Profession TypeofResidence   
## 13 8

getmode <- function(v) { ## Function to calculate Mode  
 uniqv <- unique(v)  
 uniqv[which.max(tabulate(match(v, uniqv)))]  
}  
  
demoData[is.na(demoData$Gender)==TRUE]$Gender <- getmode(demoData$Gender)  
demoData[is.na(demoData$MaritalStatus)==TRUE]$MaritalStatus <- getmode(demoData$MaritalStatus)  
demoData[is.na(demoData$Dependents)==TRUE]$Dependents <- getmode(demoData$Dependents)  
demoData[is.na(demoData$Education)==TRUE]$Education <- getmode(demoData$Education)  
demoData[is.na(demoData$TypeofResidence)==TRUE]$TypeofResidence <- getmode(demoData$TypeofResidence)  
demoData[is.na(demoData$Profession)==TRUE]$Profession <- getmode(demoData$Profession)  
sum(is.na(demoData))

## [1] 0

#### Imputing with woe values ####  
  
## Changing the data type of the variables, and small necessary amendments in the dataset  
str(woe\_data)

## Classes 'data.table' and 'data.frame': 69870 obs. of 12 variables:  
## $ ApplicationID : chr " 954457215" " 432830445" " 941387308" " 392161677" ...  
## $ Age : num 48 31 32 43 35 20 42 34 30 22 ...  
## $ Gender : chr "F" "M" "M" "M" ...  
## $ MaritalStatus : chr "Married" "Married" "Single" "Married" ...  
## $ Dependents : chr " 2" " 4" " 2" " 1" ...  
## $ Income : num 40 55 46 53 44 39 55 49 48 38 ...  
## $ Education : chr "Bachelor" "Professional" "Bachelor" "Bachelor" ...  
## $ Profession : chr "SAL" "SE\_PROF" "SE\_PROF" "SE" ...  
## $ TypeofResidence : chr "Rented" "Rented" "Rented" "Rented" ...  
## $ currResidenceTenure: num 113 112 104 94 112 116 104 108 115 111 ...  
## $ currJobTenure : num 56 46 49 53 43 52 41 40 58 57 ...  
## $ PerformanceTag : chr "0" "0" "0" "0" ...  
## - attr(\*, ".internal.selfref")=<externalptr>

woe\_data <- woe\_data[,-1] ## Application\_ID will not be required  
woe\_data$Gender <- as.factor(woe\_data$Gender)  
woe\_data$MaritalStatus <- as.factor(woe\_data$MaritalStatus)  
woe\_data$Education <- as.factor(woe\_data$Education)  
woe\_data$Profession <- as.factor(woe\_data$Profession)  
woe\_data$PerformanceTag <- as.factor(woe\_data$PerformanceTag)  
woe\_data$TypeofResidence <- as.factor(woe\_data$TypeofResidence)  
  
woe\_data$Age <- as.numeric(woe\_data$Age)  
woe\_data$Dependents <- as.numeric(woe\_data$Dependents)  
woe\_data$Income <- as.numeric(woe\_data$Income)  
woe\_data$currResidenceTenure <- as.numeric(woe\_data$currResidenceTenure)  
woe\_data$currJobTenure <- as.numeric(woe\_data$currJobTenure)  
  
# woe\_data <- as.data.frame(woe\_data)  
# row.names(woe\_data) <- 1:nrow(woe\_data)   
# woe\_test <- iv.replace.woe(woe\_data,iv=iv.mult(woe\_data,"PerformanceTag")) ## Imputing with woe values  
#   
# iv.mult(woe\_data,"PerformanceTag")  
# colSums(is.na(woe\_test))  
# colSums(is.na(demoData))  
  
### After imputation, we have 2 datasets: demoData <- Imputation with normal methodologies  
### woe\_data <- Imputation with woe\_values  
  
  
### Outlier Treatment ###  
sapply(select\_if(demoData, is.numeric), function(x)  
 quantile(x, seq(0, 1, 0.01))) ## Function to understand the outliers

## Age Income currResidenceTenure currJobTenure  
## 0% -3 -0.5 6 3  
## 1% 27 4.5 6 3  
## 2% 27 4.5 6 3  
## 3% 28 4.5 6 3  
## 4% 28 4.5 6 3  
## 5% 29 4.5 6 3  
## 6% 29 4.5 6 3  
## 7% 30 4.5 6 4  
## 8% 30 5.0 6 4  
## 9% 31 5.0 6 5  
## 10% 31 6.0 6 6  
## 11% 32 7.0 6 7  
## 12% 32 7.0 6 7  
## 13% 33 8.0 6 8  
## 14% 33 8.0 6 9  
## 15% 34 9.0 6 9  
## 16% 34 9.0 6 10  
## 17% 35 10.0 6 11  
## 18% 35 10.0 6 12  
## 19% 36 11.0 6 12  
## 20% 36 11.0 6 13  
## 21% 36 12.0 6 14  
## 22% 37 12.0 6 14  
## 23% 37 13.0 6 15  
## 24% 37 14.0 6 16  
## 25% 37 14.0 6 17  
## 26% 38 15.0 6 17  
## 27% 38 15.0 6 18  
## 28% 38 16.0 6 19  
## 29% 39 16.0 6 19  
## 30% 39 17.0 6 20  
## 31% 39 17.0 6 21  
## 32% 40 18.0 6 22  
## 33% 40 18.0 6 22  
## 34% 40 19.0 6 23  
## 35% 40 19.0 6 24  
## 36% 41 20.0 6 24  
## 37% 41 20.0 6 25  
## 38% 41 21.0 6 26  
## 39% 42 21.0 6 26  
## 40% 42 22.0 6 27  
## 41% 42 22.0 6 28  
## 42% 43 23.0 6 28  
## 43% 43 23.0 6 29  
## 44% 43 24.0 6 30  
## 45% 43 24.0 6 31  
## 46% 44 25.0 6 31  
## 47% 44 25.0 6 32  
## 48% 44 26.0 6 33  
## 49% 45 26.0 8 33  
## 50% 45 27.0 10 34  
## 51% 45 27.0 12 35  
## 52% 46 28.0 14 35  
## 53% 46 28.0 16 36  
## 54% 46 29.0 17 37  
## 55% 46 29.0 19 37  
## 56% 47 30.0 21 38  
## 57% 47 31.0 23 39  
## 58% 47 31.0 25 39  
## 59% 48 32.0 27 40  
## 60% 48 32.0 29 41  
## 61% 48 33.0 31 41  
## 62% 49 33.0 33 42  
## 63% 49 34.0 35 43  
## 64% 49 34.0 37 43  
## 65% 50 35.0 39 44  
## 66% 50 35.0 41 45  
## 67% 50 36.0 43 45  
## 68% 51 36.0 45 46  
## 69% 51 37.0 48 47  
## 70% 51 37.0 50 48  
## 71% 51 38.0 52 48  
## 72% 52 38.0 54 49  
## 73% 52 39.0 56 50  
## 74% 52 39.0 58 50  
## 75% 53 40.0 61 51  
## 76% 53 40.0 63 52  
## 77% 53 41.0 66 52  
## 78% 54 41.0 68 53  
## 79% 54 42.0 70 54  
## 80% 54 42.0 73 54  
## 81% 54 43.0 75 55  
## 82% 55 43.0 78 56  
## 83% 55 44.0 80 56  
## 84% 55 45.0 83 57  
## 85% 56 45.0 85 58  
## 86% 56 46.0 88 58  
## 87% 56 46.0 90 59  
## 88% 57 47.0 93 60  
## 89% 58 48.0 95 61  
## 90% 58 49.0 98 62  
## 91% 59 50.0 100 62  
## 92% 60 51.0 103 64  
## 93% 60 52.0 105 65  
## 94% 61 53.0 108 66  
## 95% 62 54.0 110 68  
## 96% 63 55.0 113 69  
## 97% 63 56.0 115 71  
## 98% 64 58.0 118 72  
## 99% 65 59.0 122 74  
## 100% 65 60.0 126 133

demoData[(which(demoData$Age < 27)), ]$Age <- 27  
demoData[(which(demoData$Income < 4.5)), ]$Income <- 4.5  
demoData[(which(demoData$currResidenceTenure > 122)), ]$currResidenceTenure <- 122  
demoData[(which(demoData$currJobTenure > 74)), ]$currJobTenure <- 74  
  
### Deriving new variables/Attributes ###  
head(demoData)

## ApplicationID Age Gender MaritalStatus Dependents Income Education  
## 1: 954457215 48 F Married 2 40 Bachelor  
## 2: 432830445 31 M Married 4 55 Professional  
## 3: 941387308 32 M Single 2 46 Bachelor  
## 4: 392161677 43 M Married 1 53 Bachelor  
## 5: 182011211 35 F Married 5 44 Professional  
## 6: 312196805 27 M Married 1 39 Bachelor  
## Profession TypeofResidence currResidenceTenure currJobTenure  
## 1: SAL Rented 113 56  
## 2: SE\_PROF Rented 112 46  
## 3: SE\_PROF Rented 104 49  
## 4: SE Rented 94 53  
## 5: SAL Rented 112 43  
## 6: SAL Rented 116 52  
## PerformanceTag  
## 1: 0  
## 2: 0  
## 3: 0  
## 4: 0  
## 5: 0  
## 6: 0

## Visualisation to help binning out continuous Variables  
p1 <- ggplot(demoData %>% group\_by(Age, PerformanceTag) %>% summarise(Count\_By\_PerformanceTag = n())) + geom\_point(aes(x=Age, y=Count\_By\_PerformanceTag, color = PerformanceTag))  
p2 <- ggplot(demoData %>% group\_by(Income, PerformanceTag) %>% summarise(Count\_By\_PerformanceTag = n())) + geom\_point(aes(x=Income, y=Count\_By\_PerformanceTag, color = PerformanceTag))  
p3 <- ggplot(demoData %>% group\_by(currResidenceTenure, PerformanceTag) %>% summarise(Count\_By\_PerformanceTag = n())) + geom\_point(aes(x=currResidenceTenure, y=Count\_By\_PerformanceTag, color = PerformanceTag))  
p4 <- ggplot(demoData %>% group\_by(currJobTenure, PerformanceTag) %>% summarise(Count\_By\_PerformanceTag = n())) + geom\_point(aes(x=currJobTenure, y=Count\_By\_PerformanceTag, color = PerformanceTag))  
  
plot\_grid(p1,p2,p3,p4)

A close up of a map

Description generated with high confidence

## Binning up of COntinuous Variables (On the basis of Quartiles)  
sapply(select\_if(demoData, is.numeric), function(x)  
 quantile(x, seq(0, 1, 0.25)))

## Age Income currResidenceTenure currJobTenure  
## 0% 27 4.5 6 3  
## 25% 37 14.0 6 17  
## 50% 45 27.0 10 34  
## 75% 53 40.0 61 51  
## 100% 65 60.0 122 74

demoData$Age\_bucket\_Quartile <- as.factor(cut(demoData$Age, breaks = c(26, 37, 45, 53, 66)))  
demoData$Income\_bucket\_Quartile <- as.factor(cut(demoData$Income, breaks = c(4.4, 14, 27, 40, 61)))  
demoData$currResidenceTenure\_bucket\_Quartile <- as.factor(cut(demoData$currResidenceTenure, breaks = c(5.9, 10, 61, 123)))  
demoData$currJobTenure\_bucket\_Quartile <- as.factor(cut(demoData$currJobTenure, breaks = c(2.9, 17, 34, 51, 74)))  
  
## Binning up of COntinuous Variables (On the basis of distribution w.r.t PerformanceTag)  
demoData$Age\_bucket\_Response <- as.factor(cut(demoData$Age, breaks = c(26, 36, 56, 66)))  
demoData$Income\_bucket\_Response <- as.factor(cut(demoData$Income, breaks = c(4.4, 45, 61)))  
demoData$currJobTenure\_bucket\_Response <- as.factor(cut(demoData$currJobTenure, breaks = c(2.9, 60, 74)))  
  
## There is a huge peak at some values, we can bin on then basis on that (Performance Tag=0 is high, in there)  
demoData$isIncomelessthan5 <- ifelse(demoData$Income<5, 1, 0)  
demoData$iscurrResidenceTenurelessthan8 <- ifelse(demoData$currResidenceTenure<8, 1, 0)  
demoData$iscurrJobTenurelessthan4 <- ifelse(demoData$currJobTenure<4, 1, 0)  
  
## COnverting columns into required datatypes  
ApplicationIDdemo <- demoData[,1] ## Storing Application ID from Demographic Data  
demoData <- demoData[,-1] ## Application\_ID will not be required  
demoData$Gender <- as.factor(demoData$Gender)  
demoData$MaritalStatus <- as.factor(demoData$MaritalStatus)  
demoData$Education <- as.factor(demoData$Education)  
demoData$Profession <- as.factor(demoData$Profession)  
demoData$PerformanceTag <- as.factor(demoData$PerformanceTag)  
demoData$TypeofResidence <- as.factor(demoData$TypeofResidence)  
demoData$Dependents <- as.numeric(demoData$Dependents)  
  
demoDataCleaned <- demoData ## Shall be used along with Credit Card Bureau Data for final model evaluation   
  
### Since, no. of categories in different categorical attributes is not high enough, hence, not  
## Binnning up the categorical attributes  
  
# One hot encoding  
ohe = dummyVars("~.", data = demoData[,c("Gender", "MaritalStatus", "Education", "Profession", "TypeofResidence", "PerformanceTag",  
 "Age\_bucket\_Quartile", "Income\_bucket\_Quartile", "currResidenceTenure\_bucket\_Quartile",  
 "currJobTenure\_bucket\_Quartile", "Age\_bucket\_Response", "Income\_bucket\_Response",  
 "currJobTenure\_bucket\_Response")], fullRank = T)  
ohe\_df = data.table(predict(ohe, demoData[,c("Gender", "MaritalStatus", "Education", "Profession", "TypeofResidence", "PerformanceTag",  
 "Age\_bucket\_Quartile", "Income\_bucket\_Quartile", "currResidenceTenure\_bucket\_Quartile",  
 "currJobTenure\_bucket\_Quartile", "Age\_bucket\_Response", "Income\_bucket\_Response",  
 "currJobTenure\_bucket\_Response")]))  
demoData = cbind(demoData[,-c("Gender", "MaritalStatus", "Education", "Profession", "TypeofResidence", "PerformanceTag",  
 "Age\_bucket\_Quartile", "Income\_bucket\_Quartile", "currResidenceTenure\_bucket\_Quartile",  
 "currJobTenure\_bucket\_Quartile", "Age\_bucket\_Response", "Income\_bucket\_Response",  
 "currJobTenure\_bucket\_Response")], ohe\_df)  
  
## Scaling numeric predictors  
num\_vars = which(sapply(demoData, is.numeric)) # index of numeric features  
num\_vars\_names = names(num\_vars)  
demoData\_numeric = demoData[,setdiff(num\_vars\_names, "PerformanceTag.1"), with = F]  
prep\_num = preProcess(demoData\_numeric, method=c("center", "scale"))  
demoData\_numeric\_norm = predict(prep\_num, demoData\_numeric)  
demoData[,setdiff(num\_vars\_names, "PerformanceTag.1") := NULL] # removing numeric independent variables  
demoData = cbind(demoData, demoData\_numeric\_norm)  
  
## Looking into the correlation amongst the variables  
cor\_matrix = cor(demoData[,-c("PerformanceTag.1")])  
corrplot(cor\_matrix, method = "pie", type = "lower", tl.cex = 0.9)

A close up of a map

Description generated with high confidence

##########################3####3 Data Understanding (For Credit Data) ###################################3  
### Credit Card Data ###  
# Application ID: Customer application ID  
# No of times 90 DPD or worse in last 6 months: Number of times customer has not payed dues since 90days in last 6 months  
# No of times 60 DPD or worse in last 6 months: Number of times customer has not payed dues since 60 days last 6 months  
# No of times 30 DPD or worse in last 6 months: Number of times customer has not payed dues since 30 days days last 6 months  
# No of times 90 DPD or worse in last 12 months: Number of times customer has not payed dues since 90 days days last 12 months  
# No of times 60 DPD or worse in last 12 months: Number of times customer has not payed dues since 60 days days last 12 months  
# No of times 30 DPD or worse in last 12 months: Number of times customer has not payed dues since 30 days days last 12 months  
# Avgas CC Utilization in last 12 months: Average utilization of credit card by customer  
# No of trades opened in last 6 months: Number of times the customer has done the trades in last 6 months  
# No of trades opened in last 12 months: Number of times the customer has done the trades in last 12 months  
# No of PL trades opened in last 6 months: No of PL trades in last 6 month of customer  
# No of PL trades opened in last 12 months: No of PL trades in last 12 month of customer  
# No of Inquiries in last 6 months (excluding home & auto loans): Number of times the customers has inquired in last 6 months  
# No of Inquiries in last 12 months (excluding home & auto loans): Number of times the customers has inquired in last 12 months  
# Presence of open home loan: Is the customer has home loan (1 represents "Yes")  
# Outstanding Balance: Outstanding balance of customer  
# Total No of Trades: Number of times the customer has done total trades  
# Presence of open auto loan: Is the customer has auto loan (1 represents "Yes")  
# Performance Tag: Status of customer performance (1 represents "Default")  
  
######################### Basic Data Quality Checks and Understanding (For Credit Data) ####################################  
head(creditBureauData)

## Application.ID No.of.times.90.DPD.or.worse.in.last.6.months  
## 1 954457215 0  
## 2 432830445 0  
## 3 941387308 0  
## 4 392161677 0  
## 5 182011211 0  
## 6 312196805 0  
## No.of.times.60.DPD.or.worse.in.last.6.months  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0  
## No.of.times.30.DPD.or.worse.in.last.6.months  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0  
## No.of.times.90.DPD.or.worse.in.last.12.months  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0  
## No.of.times.60.DPD.or.worse.in.last.12.months  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0  
## No.of.times.30.DPD.or.worse.in.last.12.months  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0  
## Avgas.CC.Utilization.in.last.12.months  
## 1 4  
## 2 3  
## 3 7  
## 4 11  
## 5 12  
## 6 10  
## No.of.trades.opened.in.last.6.months  
## 1 1  
## 2 1  
## 3 0  
## 4 1  
## 5 0  
## 6 0  
## No.of.trades.opened.in.last.12.months  
## 1 2  
## 2 2  
## 3 0  
## 4 1  
## 5 1  
## 6 0  
## No.of.PL.trades.opened.in.last.6.months  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0  
## No.of.PL.trades.opened.in.last.12.months  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0  
## No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0  
## No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0  
## Presence.of.open.home.loan Outstanding.Balance Total.No.of.Trades  
## 1 1 2999395 4  
## 2 0 3078 5  
## 3 1 3004972 2  
## 4 1 3355373 4  
## 5 1 3014283 4  
## 6 0 2569 1  
## Presence.of.open.auto.loan Performance.Tag  
## 1 0 0  
## 2 0 0  
## 3 0 0  
## 4 1 0  
## 5 0 0  
## 6 0 0

str(creditBureauData)

## 'data.frame': 71295 obs. of 19 variables:  
## $ Application.ID : int 954457215 432830445 941387308 392161677 182011211 312196805 532217204 74788849 782743811 96964957 ...  
## $ No.of.times.90.DPD.or.worse.in.last.6.months : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ No.of.times.60.DPD.or.worse.in.last.6.months : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ No.of.times.30.DPD.or.worse.in.last.6.months : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ No.of.times.90.DPD.or.worse.in.last.12.months : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ No.of.times.60.DPD.or.worse.in.last.12.months : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ No.of.times.30.DPD.or.worse.in.last.12.months : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ Avgas.CC.Utilization.in.last.12.months : int 4 3 7 11 12 10 11 13 9 6 ...  
## $ No.of.trades.opened.in.last.6.months : int 1 1 0 1 0 0 0 1 0 1 ...  
## $ No.of.trades.opened.in.last.12.months : int 2 2 0 1 1 0 1 1 0 1 ...  
## $ No.of.PL.trades.opened.in.last.6.months : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ No.of.PL.trades.opened.in.last.12.months : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ No.of.Inquiries.in.last.6.months..excluding.home...auto.loans. : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Presence.of.open.home.loan : int 1 0 1 1 1 0 1 1 1 0 ...  
## $ Outstanding.Balance : int 2999395 3078 3004972 3355373 3014283 2569 3005535 3004790 3007428 170860 ...  
## $ Total.No.of.Trades : int 4 5 2 4 4 1 4 3 2 1 ...  
## $ Presence.of.open.auto.loan : int 0 0 0 1 0 0 0 0 0 1 ...  
## $ Performance.Tag : int 0 0 0 0 0 0 0 0 0 0 ...

dim(creditBureauData) # 71295\*19

## [1] 71295 19

sum(is.na(creditBureauData)) ## 3028 nulls in total

## [1] 3028

colSums(is.na(select\_if(creditBureauData, colSums(is.na(creditBureauData))>0))) ## Column having nulls

## Avgas.CC.Utilization.in.last.12.months   
## 1058   
## No.of.trades.opened.in.last.6.months   
## 1   
## Presence.of.open.home.loan   
## 272   
## Outstanding.Balance   
## 272   
## Performance.Tag   
## 1425

## The predictor Variable PerformanceTag has nulls, removing the records (~2%)  
creditBureauData <- creditBureauData[is.na(creditBureauData$Performance.Tag)==F,]  
sum(sapply(creditBureauData, function(x)  
 length(which(x == "")))) # checking for blank "" values; there are none

## [1] 0

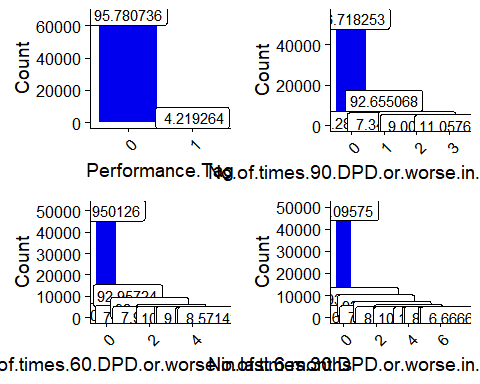
sum(duplicated(creditBureauData$Application.ID)) ## 3 duplicates

## [1] 3

creditBureauData <- creditBureauData[is.na(creditBureauData$Performance.Tag)==F,] ## Removing the recprds where target Variable is absent  
  
### Quality Issues in the Credit data ###  
"Check, if the Performance.Tag is same in both the data sets, corresponding to the same Application ID"

## [1] "Check, if the Performance.Tag is same in both the data sets, corresponding to the same Application ID"

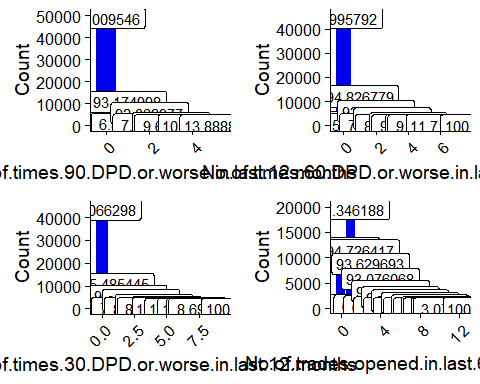
######### Exploratory Data Analysis (For Credit Card Bureau Data) ###########  
creditBureauData$Performance.Tag <- as.factor(creditBureauData$Performance.Tag)  
## Univariate Analysis  
p1 <- ggplot(creditBureauData %>% group\_by(Performance.Tag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(Performance.Tag, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(Performance.Tag, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
p2 <- ggplot(creditBureauData %>% group\_by(No.of.times.90.DPD.or.worse.in.last.6.months, Performance.Tag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(No.of.times.90.DPD.or.worse.in.last.6.months, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(No.of.times.90.DPD.or.worse.in.last.6.months, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
p3 <- ggplot(creditBureauData %>% group\_by(No.of.times.60.DPD.or.worse.in.last.6.months, Performance.Tag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(No.of.times.60.DPD.or.worse.in.last.6.months, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(No.of.times.60.DPD.or.worse.in.last.6.months, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
p4 <- ggplot(creditBureauData %>% group\_by(No.of.times.30.DPD.or.worse.in.last.6.months, Performance.Tag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(No.of.times.30.DPD.or.worse.in.last.6.months, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(No.of.times.30.DPD.or.worse.in.last.6.months, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
p5 <- ggplot(creditBureauData %>% group\_by(No.of.times.90.DPD.or.worse.in.last.12.months, Performance.Tag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(No.of.times.90.DPD.or.worse.in.last.12.months, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(No.of.times.90.DPD.or.worse.in.last.12.months, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
p6 <- ggplot(creditBureauData %>% group\_by(No.of.times.60.DPD.or.worse.in.last.12.months, Performance.Tag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(No.of.times.60.DPD.or.worse.in.last.12.months, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(No.of.times.60.DPD.or.worse.in.last.12.months, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
p7 <- ggplot(creditBureauData %>% group\_by(No.of.times.30.DPD.or.worse.in.last.12.months, Performance.Tag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(No.of.times.30.DPD.or.worse.in.last.12.months, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(No.of.times.30.DPD.or.worse.in.last.12.months, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
p8 <- ggplot(creditBureauData %>% group\_by(No.of.trades.opened.in.last.6.months, Performance.Tag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(No.of.trades.opened.in.last.6.months, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(No.of.trades.opened.in.last.6.months, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
plot\_grid(p1,p2,p3,p4)



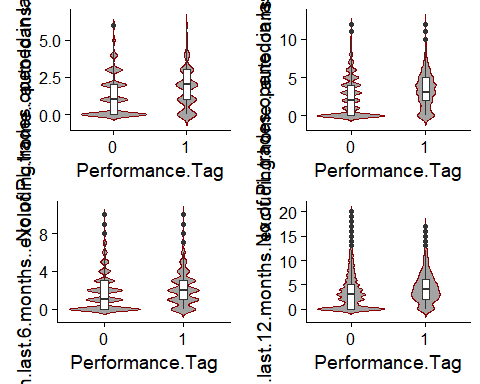
plot\_grid(p5,p6,p7,p8)

## Warning: Removed 1 rows containing missing values (position\_stack).

## Warning: Removed 1 rows containing missing values (geom\_label).



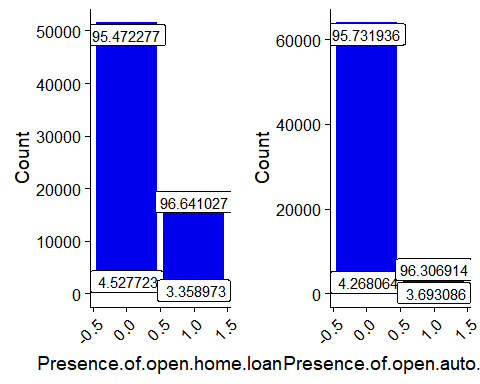
p1 <- ggplot(creditBureauData, aes(x=Performance.Tag, y=No.of.trades.opened.in.last.12.months)) + geom\_violin(trim=FALSE, fill='#A4A4A4', color="darkred") + geom\_boxplot(width=0.1)  
p2 <- ggplot(creditBureauData, aes(x=Performance.Tag, y=No.of.PL.trades.opened.in.last.6.months)) + geom\_violin(trim=FALSE, fill='#A4A4A4', color="darkred") + geom\_boxplot(width=0.1)  
p3 <- ggplot(creditBureauData, aes(x=Performance.Tag, y=No.of.PL.trades.opened.in.last.12.months)) + geom\_violin(trim=FALSE, fill='#A4A4A4', color="darkred") + geom\_boxplot(width=0.1)  
p4 <- ggplot(creditBureauData, aes(x=Performance.Tag, y=No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.)) + geom\_violin(trim=FALSE, fill='#A4A4A4', color="darkred") + geom\_boxplot(width=0.1)  
p5 <- ggplot(creditBureauData, aes(x=Performance.Tag, y=No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.)) + geom\_violin(trim=FALSE, fill='#A4A4A4', color="darkred") + geom\_boxplot(width=0.1)  
  
p6 <- ggplot(creditBureauData %>% group\_by(Presence.of.open.home.loan, Performance.Tag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(Presence.of.open.home.loan, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(Presence.of.open.home.loan, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
p7 <- ggplot(creditBureauData, aes(x=Performance.Tag, y=Total.No.of.Trades)) + geom\_violin(trim=FALSE, fill='#A4A4A4', color="darkred") + geom\_boxplot(width=0.1)  
  
p8 <- ggplot(creditBureauData %>% group\_by(Presence.of.open.auto.loan, Performance.Tag) %>% summarise(Count = n()) %>% mutate(freq = format((Count \* 100) / sum(Count)))) +   
 geom\_bar(aes(Presence.of.open.auto.loan, Count), stat = "identity", fill = "blue2") +  
 geom\_label(aes(Presence.of.open.auto.loan, Count, label = freq), vjust = 0.5) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
p9 <- ggplot(creditBureauData, aes(x=Performance.Tag, y=Avgas.CC.Utilization.in.last.12.months)) + geom\_violin(trim=FALSE, fill='#A4A4A4', color="darkred") + geom\_boxplot(width=0.1)  
p10 <- ggplot(creditBureauData, aes(x=Performance.Tag, y=Outstanding.Balance)) + geom\_violin(trim=FALSE, fill='#A4A4A4', color="darkred") + geom\_boxplot(width=0.1)  
  
  
plot\_grid(p2,p3,p4,p5)



plot\_grid(p6,p8)

## Warning: Removed 2 rows containing missing values (position\_stack).

## Warning: Removed 2 rows containing missing values (geom\_label).



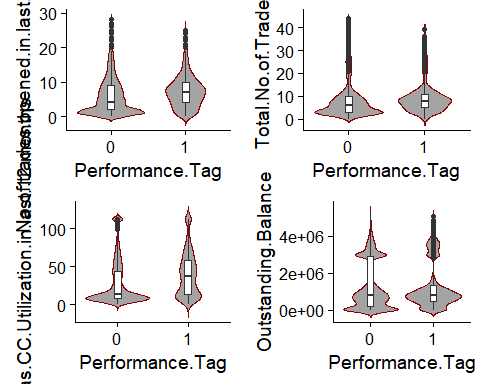
plot\_grid(p1,p7,p9,p10)

## Warning: Removed 1023 rows containing non-finite values (stat\_ydensity).

## Warning: Removed 1023 rows containing non-finite values (stat\_boxplot).

## Warning: Removed 272 rows containing non-finite values (stat\_ydensity).

## Warning: Removed 272 rows containing non-finite values (stat\_boxplot).



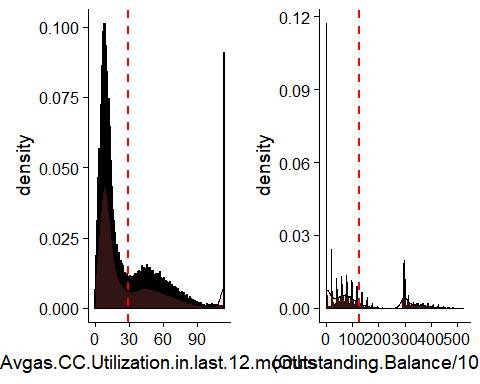
## Checking for the normallity of the Continuous Variables  
p1 <- ggplot(creditBureauData, aes(x=Avgas.CC.Utilization.in.last.12.months)) + geom\_histogram(aes(y=..density..), binwidth=.5,  
 colour="black", fill="white") + geom\_density(alpha=.2, fill="#FF6666") +  
 geom\_vline(aes(xintercept=mean(Avgas.CC.Utilization.in.last.12.months, na.rm=T)), color="red", linetype="dashed", size=1)  
  
p2 <- ggplot(creditBureauData, aes(x=(Outstanding.Balance/10000))) + geom\_histogram(aes(y=..density..), binwidth=.5,  
 colour="black", fill="white") + geom\_density(alpha=.2, fill="#FF6666") +  
 geom\_vline(aes(xintercept=mean(Outstanding.Balance, na.rm=T)/10000), color="red", linetype="dashed", size=1)  
  
plot\_grid(p1,p2)

## Warning: Removed 1023 rows containing non-finite values (stat\_bin).

## Warning: Removed 1023 rows containing non-finite values (stat\_density).

## Warning: Removed 272 rows containing non-finite values (stat\_bin).

## Warning: Removed 272 rows containing non-finite values (stat\_density).



### Imputation using normal Methodologies (Mean, Median, Mode)###  
sapply(creditBureauData[, colnames(creditBureauData) %in% c("Avgas.CC.Utilization.in.last.12.months","No.of.trades.opened.in.last.6.months",  
 "Presence.of.open.home.loan")], function(x) unique(x))

## $Avgas.CC.Utilization.in.last.12.months  
## [1] 4 3 7 11 12 10 13 9 6 5 2 14 0 8 NA 15 1  
## [18] 16 19 18 17 20 26 24 23 22 21 27 25 28 29 113 83 111  
## [35] 43 103 98 97 37 45 80 57 94 40 34 41 105 59 81 75 56  
## [52] 88 70 95 65 102 85 112 32 73 58 104 77 35 100 49 110 96  
## [69] 79 68 90 93 33 60 106 91 64 107 72 74 101 39 71 36 99  
## [86] 38 84 86 46 50 48 51 53 42 66 78 61 44 54 92 62 82  
## [103] 109 76 31 55 108 87 69 89 30 47 52 67 63  
##   
## $No.of.trades.opened.in.last.6.months  
## [1] 1 0 2 3 4 5 9 6 7 8 12 10 11 NA  
##   
## $Presence.of.open.home.loan  
## [1] 1 0 NA

## Changing the column names, replacing the '.' with '\_'  
colnames(creditBureauData) <- gsub("\\.","\_",names(creditBureauData))  
colSums(is.na(select\_if(creditBureauData, colSums(is.na(creditBureauData))>0))) ## Column having nulls

## Avgas\_CC\_Utilization\_in\_last\_12\_months   
## 1023   
## No\_of\_trades\_opened\_in\_last\_6\_months   
## 1   
## Presence\_of\_open\_home\_loan   
## 272   
## Outstanding\_Balance   
## 272

creditBureauData <- as.data.table(creditBureauData)  
creditBureauData[is.na(creditBureauData$Avgas\_CC\_Utilization\_in\_last\_12\_months)==TRUE]$Avgas\_CC\_Utilization\_in\_last\_12\_months <- getmode(creditBureauData$Avgas\_CC\_Utilization\_in\_last\_12\_months)  
creditBureauData[is.na(creditBureauData$No\_of\_trades\_opened\_in\_last\_6\_months)==TRUE]$No\_of\_trades\_opened\_in\_last\_6\_months <- getmode(creditBureauData$No\_of\_trades\_opened\_in\_last\_6\_months)  
creditBureauData[is.na(creditBureauData$Presence\_of\_open\_home\_loan)==TRUE]$Presence\_of\_open\_home\_loan <- getmode(creditBureauData$Presence\_of\_open\_home\_loan)  
creditBureauData[is.na(creditBureauData$Outstanding\_Balance)==TRUE]$Outstanding\_Balance <- as.integer(median(creditBureauData$Outstanding\_Balance, na.rm = TRUE))  
sum(is.na(creditBureauData)) ## All missing Values imputed

## [1] 0

## ## Checking for the Outliers  
sapply(creditBureauData, function(x) length(unique(x)))

## Application\_ID   
## 69867   
## No\_of\_times\_90\_DPD\_or\_worse\_in\_last\_6\_months   
## 4   
## No\_of\_times\_60\_DPD\_or\_worse\_in\_last\_6\_months   
## 6   
## No\_of\_times\_30\_DPD\_or\_worse\_in\_last\_6\_months   
## 8   
## No\_of\_times\_90\_DPD\_or\_worse\_in\_last\_12\_months   
## 6   
## No\_of\_times\_60\_DPD\_or\_worse\_in\_last\_12\_months   
## 8   
## No\_of\_times\_30\_DPD\_or\_worse\_in\_last\_12\_months   
## 10   
## Avgas\_CC\_Utilization\_in\_last\_12\_months   
## 114   
## No\_of\_trades\_opened\_in\_last\_6\_months   
## 13   
## No\_of\_trades\_opened\_in\_last\_12\_months   
## 29   
## No\_of\_PL\_trades\_opened\_in\_last\_6\_months   
## 7   
## No\_of\_PL\_trades\_opened\_in\_last\_12\_months   
## 13   
## No\_of\_Inquiries\_in\_last\_6\_months\_\_excluding\_home\_\_\_auto\_loans\_   
## 11   
## No\_of\_Inquiries\_in\_last\_12\_months\_\_excluding\_home\_\_\_auto\_loans\_   
## 21   
## Presence\_of\_open\_home\_loan   
## 2   
## Outstanding\_Balance   
## 63948   
## Total\_No\_of\_Trades   
## 45   
## Presence\_of\_open\_auto\_loan   
## 2   
## Performance\_Tag   
## 2

ApplicationIDcredit <- creditBureauData[,1] ## Storing Application ID from Credit Bureau Data  
creditBureauData <- creditBureauData[,-1] ## Removing Application\_ID  
summary(creditBureauData)

sapply(creditBureauData[,-18], function(x) quantile(x, seq(0, 1, 0.01)))

creditBureauData[(which(creditBureauData$Total\_No\_of\_Trades > 44)), ]$Total\_No\_of\_Trades <- 44  
creditBureauData[(which(creditBureauData$Outstanding\_Balance < 1251.07)), ]$Outstanding\_Balance <- 1251  
creditBureauData[(which(creditBureauData$No\_of\_Inquiries\_in\_last\_12\_months\_\_excluding\_home\_\_\_auto\_loans\_ > 15)), ]$No\_of\_Inquiries\_in\_last\_12\_months\_\_excluding\_home\_\_\_auto\_loans\_ <- 20  
creditBureauData[(which(creditBureauData$No\_of\_trades\_opened\_in\_last\_12\_months > 21)), ]$No\_of\_trades\_opened\_in\_last\_12\_months <- 21  
  
## Binning up of Continuous Variables and Feature Engineering  
sapply(creditBureauData[,15], function(x) quantile(x, seq(0, 1, 0.25)))

## Outstanding\_Balance  
## 0% 1251  
## 25% 209061  
## 50% 774234  
## 75% 2924615  
## 100% 5218801

creditBureauData$Outstanding\_Balance\_Quartile <- as.factor(cut(creditBureauData$Outstanding\_Balance, breaks = c(1250, 209061, 774234, 2924615, 5218802)))  
  
## One Hot Encoding  
ohe = dummyVars("~.", data = creditBureauData[,c("Outstanding\_Balance\_Quartile", "Performance\_Tag")], fullRank = T)  
ohe\_df = data.table(predict(ohe, creditBureauData[,c("Outstanding\_Balance\_Quartile", "Performance\_Tag")]))  
creditBureauData = cbind(creditBureauData[,-c("Outstanding\_Balance\_Quartile", "Performance\_Tag")], ohe\_df)  
creditBureauData <- as.data.table(sapply( creditBureauData, as.numeric ))  
  
## Scaling numeric predictors  
num\_vars = which(sapply(creditBureauData, is.numeric)) # index of numeric features  
num\_vars\_names = names(num\_vars)  
creditBureauData\_numeric = creditBureauData[,setdiff(num\_vars\_names, "Performance\_Tag.1"), with = F]  
prep\_num = preProcess(creditBureauData\_numeric, method=c("center", "scale"))  
creditBureauData\_numeric\_norm = predict(prep\_num, creditBureauData\_numeric)  
creditBureauData[,setdiff(num\_vars\_names, "Performance\_Tag.1") := NULL] # removing numeric independent variables  
creditBureauData = cbind(creditBureauData, creditBureauData\_numeric\_norm)  
  
## Looking into the correlation amongst the variables  
cor\_matrix = cor(creditBureauData[,-c("Performance\_Tag.1")])  
corrplot(cor\_matrix, method = "pie", type = "lower", tl.cex = 0.9)

## Warning in corrplot(cor\_matrix, method = "pie", type = "lower", tl.cex =  
## 0.9): Not been able to calculate text margin, please try again with a clean  
## new empty window using {plot.new(); dev.off()} or reduce tl.cex

A screenshot of a cell phone

Description generated with high confidence

## From the plot, we can see that, there exists a strong correlation amongst the variables  
  
############################# Preparing for the final Data ##########################  
### Initial Checks ###  
dim(ApplicationIDcredit); dim(ApplicationIDdemo)

## [1] 69870 1

## [1] 69870 1

nrow(demoData); nrow(creditBureauData)

## [1] 69870

## [1] 69870

### Joining the datasets to prepare the final DataSet for final Modelling  
finalData <- as.data.table(cbind(ApplicationIDcredit,demoData,creditBureauData))  
str(finalData)

## Classes 'data.table' and 'data.frame': 69870 obs. of 58 variables:  
## $ Application\_ID : int 954457215 432830445 941387308 392161677 182011211 312196805 532217204 74788849 782743811 96964957 ...  
## $ PerformanceTag.1 : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Age : num 0.301 -1.426 -1.324 -0.207 -1.019 ...  
## $ Dependents : num -0.62 0.823 -0.62 -1.342 1.545 ...  
## $ Income : num 0.814 1.784 1.202 1.655 1.073 ...  
## $ currResidenceTenure : num 2.13 2.1 1.89 1.61 2.1 ...  
## $ currJobTenure : num 1.073 0.581 0.729 0.926 0.434 ...  
## $ isIncomelessthan5 : num -0.281 -0.281 -0.281 -0.281 -0.281 ...  
## $ iscurrResidenceTenurelessthan8 : num -0.972 -0.972 -0.972 -0.972 -0.972 ...  
## $ iscurrJobTenurelessthan4 : num -0.27 -0.27 -0.27 -0.27 -0.27 ...  
## $ Gender.M : num -1.798 0.556 0.556 0.556 -1.798 ...  
## $ MaritalStatus.Single : num -0.416 -0.416 2.403 -0.416 -0.416 ...  
## $ Education.Masters : num -0.711 -0.711 -0.711 -0.711 -0.711 ...  
## $ Education.Others : num -0.0413 -0.0413 -0.0413 -0.0413 -0.0413 ...  
## $ Education.Phd : num -0.261 -0.261 -0.261 -0.261 -0.261 ...  
## $ Education.Professional : num -0.735 1.361 -0.735 -0.735 1.361 ...  
## $ Profession.SE : num -0.499 -0.499 -0.499 2.004 -0.499 ...  
## $ Profession.SE\_PROF : num -0.551 1.816 1.816 -0.551 -0.551 ...  
## $ TypeofResidence.Living with Parents : num -0.162 -0.162 -0.162 -0.162 -0.162 ...  
## $ TypeofResidence.Others : num -0.0533 -0.0533 -0.0533 -0.0533 -0.0533 ...  
## $ TypeofResidence.Owned : num -0.501 -0.501 -0.501 -0.501 -0.501 ...  
## $ TypeofResidence.Rented : num 0.58 0.58 0.58 0.58 0.58 ...  
## $ Age\_bucket\_Quartile.(37,45] : num -0.604 -0.604 -0.604 1.655 -0.604 ...  
## $ Age\_bucket\_Quartile.(45,53] : num 1.686 -0.593 -0.593 -0.593 -0.593 ...  
## $ Age\_bucket\_Quartile.(53,66] : num -0.534 -0.534 -0.534 -0.534 -0.534 ...  
## $ Income\_bucket\_Quartile.(14,27] : num -0.582 -0.582 -0.582 -0.582 -0.582 ...  
## $ Income\_bucket\_Quartile.(27,40] : num 1.718 -0.582 -0.582 -0.582 -0.582 ...  
## $ Income\_bucket\_Quartile.(40,61] : num -0.555 1.802 1.802 1.802 1.802 ...  
## $ currResidenceTenure\_bucket\_Quartile.(10,61] : num -0.579 -0.579 -0.579 -0.579 -0.579 ...  
## $ currResidenceTenure\_bucket\_Quartile.(61,123] : num 1.75 1.75 1.75 1.75 1.75 ...  
## $ currJobTenure\_bucket\_Quartile.(17,34] : num -0.567 -0.567 -0.567 -0.567 -0.567 ...  
## $ currJobTenure\_bucket\_Quartile.(34,51] : num -0.577 1.732 1.732 -0.577 1.732 ...  
## $ currJobTenure\_bucket\_Quartile.(51,74] : num 1.767 -0.566 -0.566 1.767 -0.566 ...  
## $ Age\_bucket\_Response.(36,56] : num 0.724 -1.381 -1.381 0.724 -1.381 ...  
## $ Age\_bucket\_Response.(56,66] : num -0.382 -0.382 -0.382 -0.382 -0.382 ...  
## $ Income\_bucket\_Response.(45,61] : num -0.407 2.458 2.458 2.458 -0.407 ...  
## $ currJobTenure\_bucket\_Response.(60,74] : num -0.356 -0.356 -0.356 -0.356 -0.356 ...  
## $ Performance\_Tag.1 : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ No\_of\_times\_90\_DPD\_or\_worse\_in\_last\_6\_months : num -0.492 -0.492 -0.492 -0.492 -0.492 ...  
## $ No\_of\_times\_60\_DPD\_or\_worse\_in\_last\_6\_months : num -0.507 -0.507 -0.507 -0.507 -0.507 ...  
## $ No\_of\_times\_30\_DPD\_or\_worse\_in\_last\_6\_months : num -0.523 -0.523 -0.523 -0.523 -0.523 ...  
## $ No\_of\_times\_90\_DPD\_or\_worse\_in\_last\_12\_months : num -0.543 -0.543 -0.543 -0.543 -0.543 ...  
## $ No\_of\_times\_60\_DPD\_or\_worse\_in\_last\_12\_months : num -0.591 -0.591 -0.591 -0.591 -0.591 ...  
## $ No\_of\_times\_30\_DPD\_or\_worse\_in\_last\_12\_months : num -0.59 -0.59 -0.59 -0.59 -0.59 ...  
## $ Avgas\_CC\_Utilization\_in\_last\_12\_months : num -0.848 -0.882 -0.746 -0.61 -0.576 ...  
## $ No\_of\_trades\_opened\_in\_last\_6\_months : num -0.617 -0.617 -1.098 -0.617 -1.098 ...  
## $ No\_of\_trades\_opened\_in\_last\_12\_months : num -0.747 -0.747 -1.143 -0.945 -0.945 ...  
## $ No\_of\_PL\_trades\_opened\_in\_last\_6\_months : num -0.879 -0.879 -0.879 -0.879 -0.879 ...  
## $ No\_of\_PL\_trades\_opened\_in\_last\_12\_months : num -0.975 -0.975 -0.975 -0.975 -0.975 ...  
## $ No\_of\_Inquiries\_in\_last\_6\_months\_\_excluding\_home\_\_\_auto\_loans\_ : num -0.886 -0.886 -0.886 -0.886 -0.886 ...  
## $ No\_of\_Inquiries\_in\_last\_12\_months\_\_excluding\_home\_\_\_auto\_loans\_: num -0.962 -0.962 -0.962 -0.962 -0.962 ...  
## $ Presence\_of\_open\_home\_loan : num 1.693 -0.591 1.693 1.693 1.693 ...  
## $ Outstanding\_Balance : num 1.361 -0.972 1.366 1.639 1.373 ...  
## $ Total\_No\_of\_Trades : num -0.584 -0.444 -0.863 -0.584 -0.584 ...  
## $ Presence\_of\_open\_auto\_loan : num -0.305 -0.305 -0.305 3.284 -0.305 ...  
## $ Outstanding\_Balance\_Quartile.(2.09e+05,7.74e+05] : num -0.58 -0.58 -0.58 -0.58 -0.58 ...  
## $ Outstanding\_Balance\_Quartile.(7.74e+05,2.92e+06] : num -0.574 -0.574 -0.574 -0.574 -0.574 ...  
## $ Outstanding\_Balance\_Quartile.(2.92e+06,5.22e+06] : num 1.732 -0.577 1.732 1.732 1.732 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

## Checking the Values in the target Attribute  
setdiff(finalData$Performance\_Tag.1, finalData$PerformanceTag.1) ## No Difference :)

## numeric(0)

finalData <- finalData[,-c("Performance\_Tag.1","Application\_ID")]  
cor\_matrix = cor(finalData[,-c("PerformanceTag.1")])  
corrplot(cor\_matrix, method = "pie", type = "lower", tl.cex = 0.9) ## Correlation in the Final Data Set

## Warning in corrplot(cor\_matrix, method = "pie", type = "lower", tl.cex =  
## 0.9): Not been able to calculate text margin, please try again with a clean  
## new empty window using {plot.new(); dev.off()} or reduce tl.cex

A close up of text on a white background

Description generated with high confidence