Data-Due-Diligence

Amol Gote

6/6/2020

customers <- read\_csv("data/Customer\_Dataset\_File\_Original.csv")

# Identify all missing values

sum(is.na(customers))

## [1] 124

sapply(customers,function(x)sum(is.na(x)))

## CustomerID Region TownSize   
## 0 0 0   
## Gender Age EducationYears   
## 33 0 0   
## JobCategory UnionMember EmploymentLength   
## 15 0 0   
## Retired HHIncome DebtToIncomeRatio   
## 0 0 0   
## CreditDebt OtherDebt LoanDefault   
## 0 0 0   
## MaritalStatus HouseholdSize NumberPets   
## 0 8 6   
## NumberCats NumberDogs NumberBirds   
## 7 8 34   
## HomeOwner CarsOwned CarOwnership   
## 13 0 0   
## CarBrand CarValue CommuteTime   
## 0 0 0   
## PoliticalPartyMem Votes CreditCard   
## 0 0 0   
## CardTenure CardItemsMonthly CardSpendMonth   
## 0 0 0   
## ActiveLifestyle PhoneCoTenure VoiceLastMonth   
## 0 0 0   
## VoiceOverTenure EquipmentRental EquipmentLastMonth   
## 0 0 0   
## EquipmentOverTenure CallingCard WirelessData   
## 0 0 0   
## DataLastMonth DataOverTenure Multiline   
## 0 0 0   
## VM Pager Internet   
## 0 0 0   
## CallerID CallWait CallForward   
## 0 0 0   
## ThreeWayCalling EBilling TVWatchingHours   
## 0 0 0   
## OwnsPC OwnsMobileDevice OwnsGameSystem   
## 0 0 0   
## OwnsFax NewsSubscriber   
## 0 0

# Feature engineering steps

# Data Imputation for customers.

#customers <- read\_csv("data/Customer\_Dataset\_File\_Original.csv")  
print("Number of rows with Gender = NA before imputation")

## [1] "Number of rows with Gender = NA before imputation"

sum(is.na(customers$Gender))

## [1] 33

numberOfFemales <- customers %>%   
 filter(Gender=="Female") %>%   
 nrow()  
print("Number of females before imputation")

## [1] "Number of females before imputation"

numberOfFemales

## [1] 2494

numberOfMales <- customers %>%   
 filter(Gender=="Male") %>%   
 nrow()  
count <- 0  
print("Number of males before imputation")

## [1] "Number of males before imputation"

numberOfMales

## [1] 2473

for(i in 1:nrow(customers)){  
 if(is.na(customers$Gender[i])){  
 #customers$Gender[i] <- "F"  
 count <- count + 1  
 if (count %% 2 == 0){  
 customers$Gender[i] <- "Female"  
 }  
 else{  
 customers$Gender[i] <- "Male"  
 }  
 }  
}  
print("Number of rows with Gender = NA after imputation")

## [1] "Number of rows with Gender = NA after imputation"

sum(is.na(customers$Gender))

## [1] 0

numberOfFemales <- customers %>%   
 filter(Gender=="Female") %>%   
 nrow()  
print("Number of females after imputation")

## [1] "Number of females after imputation"

numberOfFemales

## [1] 2510

numberOfMales <- customers %>%   
 filter(Gender=="Male") %>%   
 nrow()  
print("Number of males after imputation")

## [1] "Number of males after imputation"

numberOfMales

## [1] 2490

# Data Imputation for HouseholdSize.

print("Number of rows with HouseholdSize = NA before imputation")

## [1] "Number of rows with HouseholdSize = NA before imputation"

sum(is.na(customers$HouseholdSize))

## [1] 8

summary(customers$HouseholdSize)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.000 1.000 2.000 2.202 3.000 9.000 8

customers$HouseholdSize[is.na(customers$HouseholdSize)] <- median(customers$HouseholdSize, na.rm = T)  
print("Check for missing values for HouseholdSize")

## [1] "Check for missing values for HouseholdSize"

sum(is.na(customers$HouseholdSize))

## [1] 0

# Data Imputation for HomeOwner.

print("Number of rows with HomeOwner = NA before imputation")

## [1] "Number of rows with HomeOwner = NA before imputation"

sum(is.na(customers$HomeOwner))

## [1] 13

summary(customers$HomeOwner)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0000 0.0000 1.0000 0.6296 1.0000 1.0000 13

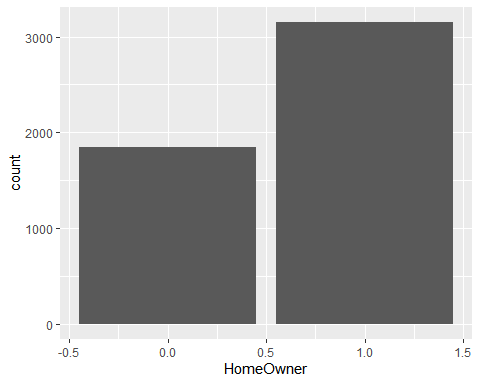
get\_mode <- function(x) {  
 unique\_x <- unique(x)  
 mode <- unique\_x[which.max(tabulate(match(x, unique\_x)))]  
 mode  
}  
mode\_value = get\_mode(customers$HomeOwner)  
customers$HomeOwner[is.na(customers$HomeOwner)] <- mode\_value  
print("Check for missing values for HomeOwner")

## [1] "Check for missing values for HomeOwner"

sum(is.na(customers$HomeOwner))

## [1] 0

ggplot(customers) +   
 geom\_bar(aes(x = HomeOwner))



# Data Imputation for JobCategory

#customers <- read\_csv("data/Customer\_Dataset\_File\_Original.csv")  
print("Number of rows with JobCategory = NA before imputation")

## [1] "Number of rows with JobCategory = NA before imputation"

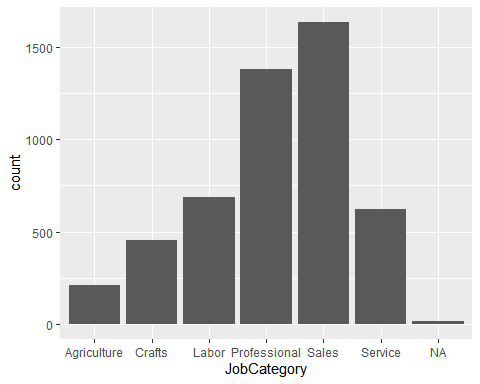
sum(is.na(customers$JobCategory))

## [1] 15

summary(customers$JobCategory)

## Length Class Mode   
## 5000 character character

ggplot(customers) +   
 geom\_bar(aes(x = JobCategory))



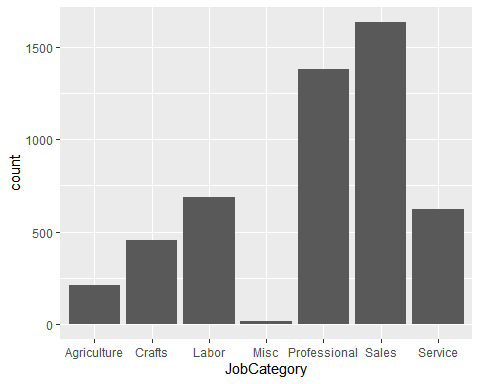
customers$JobCategory[is.na(customers$JobCategory)] <- "Misc"  
print("Check for missing values for HomeOwner")

## [1] "Check for missing values for HomeOwner"

sum(is.na(customers$JobCategory))

## [1] 0

ggplot(customers) +   
 geom\_bar(aes(x = JobCategory))



# Remove the dollar sign from household income

#customers <- read\_csv("data/Customer\_Dataset\_File\_Original.csv")  
customers <- customers %>%  
 replace(.=="#NULL!", NA) # replace with NA  
customers$HHIncome = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$HHIncome))))   
customers$VoiceLastMonth = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$VoiceLastMonth))))   
customers$VoiceOverTenure = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$VoiceOverTenure))))   
customers$CardSpendMonth = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$CardSpendMonth))))   
customers$EquipmentLastMonth = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$EquipmentLastMonth))))   
customers$EquipmentOverTenure = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$EquipmentOverTenure))))   
customers$DataLastMonth = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$DataLastMonth))))   
customers$DataOverTenure = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$DataOverTenure))))   
sapply(customers,function(x)sum(is.na(x)))

## CustomerID Region TownSize   
## 0 0 2   
## Gender Age EducationYears   
## 0 0 0   
## JobCategory UnionMember EmploymentLength   
## 0 0 0   
## Retired HHIncome DebtToIncomeRatio   
## 0 0 0   
## CreditDebt OtherDebt LoanDefault   
## 0 0 0   
## MaritalStatus HouseholdSize NumberPets   
## 0 0 6   
## NumberCats NumberDogs NumberBirds   
## 7 8 34   
## HomeOwner CarsOwned CarOwnership   
## 0 0 0   
## CarBrand CarValue CommuteTime   
## 0 0 2   
## PoliticalPartyMem Votes CreditCard   
## 0 0 0   
## CardTenure CardItemsMonthly CardSpendMonth   
## 0 0 7   
## ActiveLifestyle PhoneCoTenure VoiceLastMonth   
## 0 0 0   
## VoiceOverTenure EquipmentRental EquipmentLastMonth   
## 3 0 3296   
## EquipmentOverTenure CallingCard WirelessData   
## 3296 0 0   
## DataLastMonth DataOverTenure Multiline   
## 3656 3656 0   
## VM Pager Internet   
## 0 0 0   
## CallerID CallWait CallForward   
## 0 0 0   
## ThreeWayCalling EBilling TVWatchingHours   
## 0 0 0   
## OwnsPC OwnsMobileDevice OwnsGameSystem   
## 0 0 0   
## OwnsFax NewsSubscriber   
## 0 0

ncol(customers)

## [1] 59

# Remove the dollar sign from CarValue

#customers <- read\_csv("data/Customer\_Dataset\_File\_Original.csv")  
customers$CarValue = gsub("\\ ", "", customers$CarValue)   
summary(customers$CarValue)

## Length Class Mode   
## 5000 character character

carValueLessThanZero <- customers %>%   
 filter(CarValue =="$(1,000.00)") %>%   
 nrow()  
customers$CarValue[customers$CarValue =="$(1,000.00)"] <- 0  
customers$CarValue = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$CarValue))))   
customers$CarValue = as.numeric(customers$CarValue)  
carValueLessThanZero <- customers %>%   
 filter(is.na(customers$CarValue)) %>%  
 nrow()  
   
print("Number of records with CarValue = NA")

## [1] "Number of records with CarValue = NA"

carValueLessThanZero

## [1] 0

carOwnership <- customers %>%   
 filter(customers$CarOwnership == "-1") %>%  
 nrow()  
print("Number of records with CarOwnership == -1")

## [1] "Number of records with CarOwnership == -1"

carOwnership

## [1] 497

carBrand <- customers %>%   
 filter(customers$CarBrand == "-1") %>%  
 nrow()  
print("Number of records with CarBrand == -1")

## [1] "Number of records with CarBrand == -1"

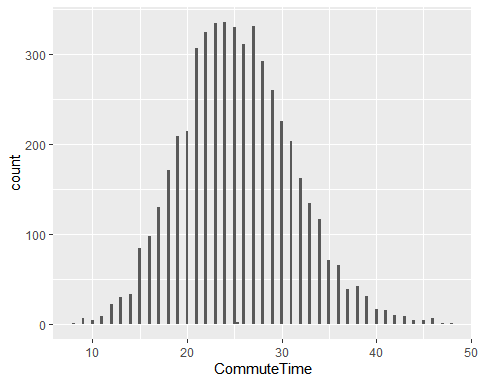
carBrand

## [1] 497

customers$CarOwnership[customers$CarOwnership =="-1"] <- NA  
customers$CarBrand[customers$CarBrand =="-1"] <- NA

# Impute Commute Time, there are 2 missing values

customers$CommuteTime = as.numeric(customers$CommuteTime)  
customers$CommuteTime[is.na(customers$CommuteTime)] <- mean(customers$CommuteTime, na.rm = T)  
ggplot(customers) +   
 geom\_bar(aes(x = CommuteTime))

 #Address Missing values for EquipmentLastMonth, DataLastMonth, EquipmentOverTenure,DataOverTenure

customers$EquipmentLastMonth[is.na(customers$EquipmentLastMonth)] <- 0  
customers$DataLastMonth[is.na(customers$DataLastMonth)] <- 0  
customers$EquipmentOverTenure[is.na(customers$EquipmentOverTenure)] <- 0  
customers$DataOverTenure[is.na(customers$DataOverTenure)] <- 0  
customers$PhoneCoTenure <- ifelse(customers$PhoneCoTenure == 0, 1, customers$PhoneCoTenure)  
customers$VoiceOverTenure <- ifelse(is.na(customers$VoiceOverTenure), customers$VoiceLastMonth, customers$VoiceOverTenure)

# Final Summary of customer dataset

summary(customers)

## CustomerID Region TownSize Gender   
## Length:5000 Min. :1.000 Length:5000 Length:5000   
## Class :character 1st Qu.:2.000 Class :character Class :character   
## Mode :character Median :3.000 Mode :character Mode :character   
## Mean :3.001   
## 3rd Qu.:4.000   
## Max. :5.000   
##   
## Age EducationYears JobCategory UnionMember   
## Min. :18.00 Min. : 6.00 Length:5000 Length:5000   
## 1st Qu.:31.00 1st Qu.:12.00 Class :character Class :character   
## Median :47.00 Median :14.00 Mode :character Mode :character   
## Mean :47.03 Mean :14.54   
## 3rd Qu.:62.00 3rd Qu.:17.00   
## Max. :79.00 Max. :23.00   
##   
## EmploymentLength Retired HHIncome DebtToIncomeRatio  
## Min. : 0.00 Length:5000 Min. : 9000 Min. : 0.000   
## 1st Qu.: 2.00 Class :character 1st Qu.: 24000 1st Qu.: 5.100   
## Median : 7.00 Mode :character Median : 38000 Median : 8.800   
## Mean : 9.73 Mean : 54760 Mean : 9.954   
## 3rd Qu.:15.00 3rd Qu.: 67000 3rd Qu.:13.600   
## Max. :52.00 Max. :1073000 Max. :43.100   
##   
## CreditDebt OtherDebt LoanDefault   
## Min. : 0.0000 Min. : 0.0000 Length:5000   
## 1st Qu.: 0.3855 1st Qu.: 0.9803 Class :character   
## Median : 0.9264 Median : 2.0985 Mode :character   
## Mean : 1.8573 Mean : 3.6545   
## 3rd Qu.: 2.0638 3rd Qu.: 4.3148   
## Max. :109.0726 Max. :141.4591   
##   
## MaritalStatus HouseholdSize NumberPets NumberCats   
## Length:5000 Min. :1.000 Min. : 0.000 Min. :0.0000   
## Class :character 1st Qu.:1.000 1st Qu.: 0.000 1st Qu.:0.0000   
## Mode :character Median :2.000 Median : 2.000 Median :0.0000   
## Mean :2.202 Mean : 3.067 Mean :0.5003   
## 3rd Qu.:3.000 3rd Qu.: 5.000 3rd Qu.:1.0000   
## Max. :9.000 Max. :21.000 Max. :6.0000   
## NA's :6 NA's :7   
## NumberDogs NumberBirds HomeOwner CarsOwned   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:1.000   
## Median :0.0000 Median :0.0000 Median :1.0000 Median :2.000   
## Mean :0.3928 Mean :0.1112 Mean :0.6306 Mean :2.131   
## 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :7.0000 Max. :5.0000 Max. :1.0000 Max. :8.000   
## NA's :8 NA's :34   
## CarOwnership CarBrand CarValue CommuteTime   
## Length:5000 Length:5000 Min. : 0 Min. : 8.00   
## Class :character Class :character 1st Qu.: 9200 1st Qu.:21.00   
## Mode :character Mode :character Median :17000 Median :25.00   
## Mean :23332 Mean :25.35   
## 3rd Qu.:31100 3rd Qu.:29.00   
## Max. :99600 Max. :48.00   
##   
## PoliticalPartyMem Votes CreditCard CardTenure   
## Length:5000 Length:5000 Length:5000 Min. : 0.00   
## Class :character Class :character Class :character 1st Qu.: 6.00   
## Mode :character Mode :character Mode :character Median :14.00   
## Mean :16.66   
## 3rd Qu.:26.00   
## Max. :40.00   
##   
## CardItemsMonthly CardSpendMonth ActiveLifestyle PhoneCoTenure   
## Min. : 0.00 Min. : 69.7 Length:5000 Min. : 1.00   
## 1st Qu.: 8.00 1st Qu.: 1838.5 Class :character 1st Qu.:18.00   
## Median :10.00 Median : 2766.9 Mode :character Median :38.00   
## Mean :10.18 Mean : 3376.8 Mean :38.21   
## 3rd Qu.:12.00 3rd Qu.: 4187.5 3rd Qu.:59.00   
## Max. :23.00 Max. :39264.1 Max. :72.00   
## NA's :7   
## VoiceLastMonth VoiceOverTenure EquipmentRental EquipmentLastMonth  
## Min. : 2.70 Min. : 0.9 Length:5000 Min. : 0.00   
## 1st Qu.: 17.10 1st Qu.: 103.9 Class :character 1st Qu.: 0.00   
## Median : 28.65 Median : 349.6 Mode :character Median : 0.00   
## Mean : 40.41 Mean : 708.5 Mean : 12.99   
## 3rd Qu.: 49.65 3rd Qu.: 913.5 3rd Qu.: 30.80   
## Max. :539.55 Max. :13046.5 Max. :106.30   
##   
## EquipmentOverTenure CallingCard WirelessData   
## Min. : 0.0 Length:5000 Length:5000   
## 1st Qu.: 0.0 Class :character Class :character   
## Median : 0.0 Mode :character Mode :character   
## Mean : 470.2   
## 3rd Qu.: 510.2   
## Max. :6525.3   
##   
## DataLastMonth DataOverTenure Multiline VM   
## Min. : 0.00 Min. : 0.00 Length:5000 Length:5000   
## 1st Qu.: 0.00 1st Qu.: 0.00 Class :character Class :character   
## Median : 0.00 Median : 0.00 Mode :character Mode :character   
## Mean : 10.70 Mean : 421.99   
## 3rd Qu.: 20.96 3rd Qu.: 89.96   
## Max. :186.25 Max. :12858.65   
##   
## Pager Internet CallerID   
## Length:5000 Length:5000 Length:5000   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## CallWait CallForward ThreeWayCalling   
## Length:5000 Length:5000 Length:5000   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## EBilling TVWatchingHours OwnsPC OwnsMobileDevice   
## Length:5000 Min. : 0.00 Length:5000 Length:5000   
## Class :character 1st Qu.:17.00 Class :character Class :character   
## Mode :character Median :20.00 Mode :character Mode :character   
## Mean :19.64   
## 3rd Qu.:23.00   
## Max. :36.00   
##   
## OwnsGameSystem OwnsFax NewsSubscriber   
## Length:5000 Length:5000 Length:5000   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##

# Adding of additional variables

customers$LastMonthTotalValue <- customers$VoiceLastMonth + customers$EquipmentLastMonth + customers$DataLastMonth  
customers$OverTenureTotalValue <- customers$VoiceOverTenure + customers$EquipmentOverTenure + customers$DataOverTenure  
customers$TotalDebt <- customers$CreditDebt + customers$OtherDebt  
customers$AverageMonthlyRevenue <- customers$OverTenureTotalValue/customers$PhoneCoTenure

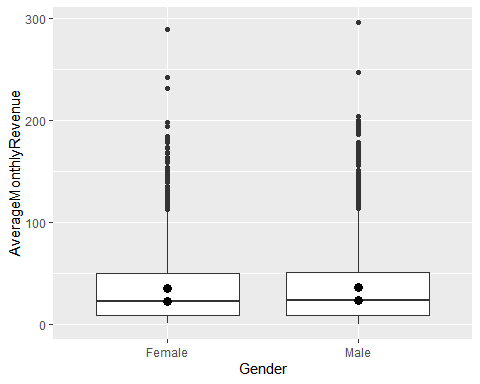
# Create a dataset for analysis with required variables only

customersForAnalysis <- customers[,c("CustomerID","Gender","Age","EducationYears","JobCategory","EmploymentLength","HHIncome","DebtToIncomeRatio","CreditDebt","OtherDebt","MaritalStatus","HouseholdSize","HomeOwner","CommuteTime","PhoneCoTenure","VoiceLastMonth","VoiceOverTenure","EquipmentRental","EquipmentLastMonth","EquipmentOverTenure","WirelessData","DataLastMonth","DataOverTenure","Multiline","EBilling", "LastMonthTotalValue", "OverTenureTotalValue", "TotalDebt", "AverageMonthlyRevenue")]  
summary(customersForAnalysis)

## CustomerID Gender Age EducationYears   
## Length:5000 Length:5000 Min. :18.00 Min. : 6.00   
## Class :character Class :character 1st Qu.:31.00 1st Qu.:12.00   
## Mode :character Mode :character Median :47.00 Median :14.00   
## Mean :47.03 Mean :14.54   
## 3rd Qu.:62.00 3rd Qu.:17.00   
## Max. :79.00 Max. :23.00   
## JobCategory EmploymentLength HHIncome DebtToIncomeRatio  
## Length:5000 Min. : 0.00 Min. : 9000 Min. : 0.000   
## Class :character 1st Qu.: 2.00 1st Qu.: 24000 1st Qu.: 5.100   
## Mode :character Median : 7.00 Median : 38000 Median : 8.800   
## Mean : 9.73 Mean : 54760 Mean : 9.954   
## 3rd Qu.:15.00 3rd Qu.: 67000 3rd Qu.:13.600   
## Max. :52.00 Max. :1073000 Max. :43.100   
## CreditDebt OtherDebt MaritalStatus HouseholdSize   
## Min. : 0.0000 Min. : 0.0000 Length:5000 Min. :1.000   
## 1st Qu.: 0.3855 1st Qu.: 0.9803 Class :character 1st Qu.:1.000   
## Median : 0.9264 Median : 2.0985 Mode :character Median :2.000   
## Mean : 1.8573 Mean : 3.6545 Mean :2.202   
## 3rd Qu.: 2.0638 3rd Qu.: 4.3148 3rd Qu.:3.000   
## Max. :109.0726 Max. :141.4591 Max. :9.000   
## HomeOwner CommuteTime PhoneCoTenure VoiceLastMonth   
## Min. :0.0000 Min. : 8.00 Min. : 1.00 Min. : 2.70   
## 1st Qu.:0.0000 1st Qu.:21.00 1st Qu.:18.00 1st Qu.: 17.10   
## Median :1.0000 Median :25.00 Median :38.00 Median : 28.65   
## Mean :0.6306 Mean :25.35 Mean :38.21 Mean : 40.41   
## 3rd Qu.:1.0000 3rd Qu.:29.00 3rd Qu.:59.00 3rd Qu.: 49.65   
## Max. :1.0000 Max. :48.00 Max. :72.00 Max. :539.55   
## VoiceOverTenure EquipmentRental EquipmentLastMonth  
## Min. : 0.9 Length:5000 Min. : 0.00   
## 1st Qu.: 103.9 Class :character 1st Qu.: 0.00   
## Median : 349.6 Mode :character Median : 0.00   
## Mean : 708.5 Mean : 12.99   
## 3rd Qu.: 913.5 3rd Qu.: 30.80   
## Max. :13046.5 Max. :106.30   
## EquipmentOverTenure WirelessData DataLastMonth   
## Min. : 0.0 Length:5000 Min. : 0.00   
## 1st Qu.: 0.0 Class :character 1st Qu.: 0.00   
## Median : 0.0 Mode :character Median : 0.00   
## Mean : 470.2 Mean : 10.70   
## 3rd Qu.: 510.2 3rd Qu.: 20.96   
## Max. :6525.3 Max. :186.25   
## DataOverTenure Multiline EBilling   
## Min. : 0.00 Length:5000 Length:5000   
## 1st Qu.: 0.00 Class :character Class :character   
## Median : 0.00 Mode :character Mode :character   
## Mean : 421.99   
## 3rd Qu.: 89.96   
## Max. :12858.65   
## LastMonthTotalValue OverTenureTotalValue TotalDebt   
## Min. : 2.85 Min. : 0.95 Min. : 0.000   
## 1st Qu.: 26.70 1st Qu.: 220.03 1st Qu.: 1.620   
## Median : 49.73 Median : 766.20 Median : 3.279   
## Mean : 64.11 Mean : 1600.62 Mean : 5.512   
## 3rd Qu.: 87.30 3rd Qu.: 2088.11 3rd Qu.: 6.456   
## Max. :590.40 Max. :21057.00 Max. :211.381   
## AverageMonthlyRevenue  
## Min. : 0.500   
## 1st Qu.: 9.072   
## Median : 23.223   
## Mean : 35.942   
## 3rd Qu.: 50.676   
## Max. :296.577

#Statistical analysis for Age variable

ggplot(data = customersForAnalysis,  
 aes(x = Gender, y = AverageMonthlyRevenue)) +  
 geom\_boxplot() +  
 stat\_summary(fun.y = mean, geom="point",size=3) +  
 stat\_summary(fun.y = median, geom="point",size=3)

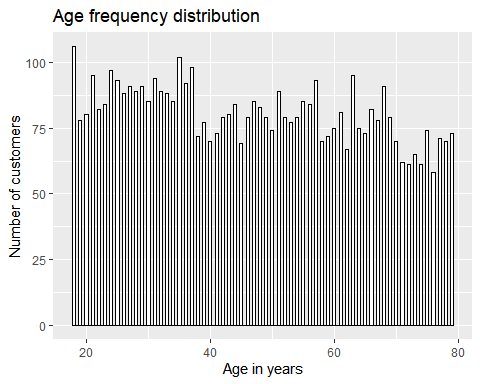


customersForAnalysis%>%  
group\_by(Gender)%>%   
summarise(Mean=mean(AverageMonthlyRevenue), Median=median(AverageMonthlyRevenue))

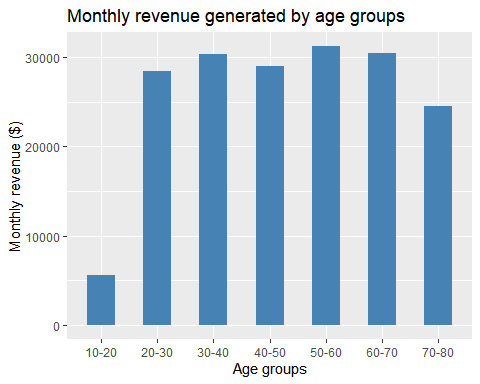
## # A tibble: 2 x 3  
## Gender Mean Median  
## <chr> <dbl> <dbl>  
## 1 Female 35.2 23.0  
## 2 Male 36.6 23.7

# Age distribution and monthly revenue genegrate by age groups

ggplot(customersForAnalysis, aes(x=Age)) +  
 geom\_histogram(binwidth=.5, colour="black", fill="white") +   
 labs( title= "Age frequency distribution", x="Age in years", y = "Number of customers")



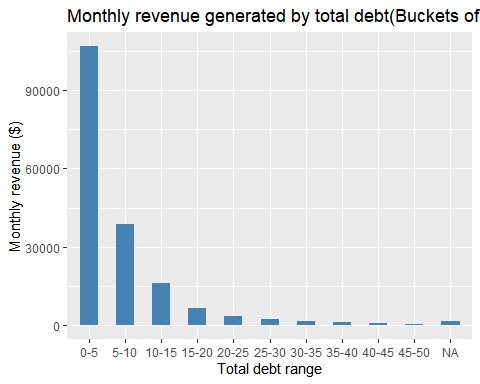
lbls <- c('0-10','10-20','20-30', '30-40','40-50','50-60','60-70', '70-80', '80-90', '90-100')  
groupedData <- customersForAnalysis %>%   
 group\_by(ageGroup = cut(Age, breaks= seq(0, 100, by = 10),   
 right = TRUE, include.lowest = TRUE, labels = lbls) ) %>%   
 summarise(totalMonthlyValue= sum(AverageMonthlyRevenue))   
  
ggplot(groupedData, aes(x = ageGroup, y = totalMonthlyValue)) +   
 geom\_bar(stat="identity", width=0.5, fill = "steelblue") +  
 labs( title= "Monthly revenue generated by age groups", x="Age groups", y = "Monthly revenue ($)")

 #Statistical analysis for total debt

lbls <- c('0-5','5-10','10-15', '15-20','20-25','25-30','30-35', '35-40', '40-45', '45-50')  
groupedData <- customersForAnalysis %>%   
 group\_by(debtGroup = cut(TotalDebt, breaks= seq(0, 50, by = 5),   
 right = TRUE, include.lowest = TRUE, labels = lbls) ) %>%   
 summarise(totalMonthlyValue= sum(AverageMonthlyRevenue))

## Warning: Factor `debtGroup` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

ggplot(groupedData, aes(x = debtGroup, y = totalMonthlyValue)) +   
 geom\_bar(stat="identity", width=0.5, fill = "steelblue") +  
 labs( title= "Monthly revenue generated by total debt(Buckets of 5)", x="Total debt range", y = "Monthly revenue ($)")

 #Statistical analysis for Household size

groupedData <- customersForAnalysis%>%  
 group\_by(HouseholdSize)%>%   
 summarise(Mean=mean(AverageMonthlyRevenue), Median=median(AverageMonthlyRevenue), totalMonthlyRevenue= sum(AverageMonthlyRevenue))  
groupedData

## # A tibble: 9 x 4  
## HouseholdSize Mean Median totalMonthlyRevenue  
## <dbl> <dbl> <dbl> <dbl>  
## 1 1 34.2 22.8 69584.   
## 2 2 39.6 25.9 58420.   
## 3 3 34.6 23.0 19117.   
## 4 4 34.7 18.8 18060.   
## 5 5 34.8 21.0 9988.   
## 6 6 34.2 26.0 3318.   
## 7 7 30.2 12.1 877.   
## 8 8 38.6 38.2 270.   
## 9 9 37.6 37.6 75.3

##Household size > 1

groupedData <- customersForAnalysis%>%  
 filter(HouseholdSize > 1) %>%  
 filter(Multiline == "No") %>%  
 summarise(totalMonthlyRevenue= sum(AverageMonthlyRevenue))  
cat(sprintf("Total monthly revenue for household size > 1 and with no multi-line:: %s", groupedData$totalMonthlyRevenue))

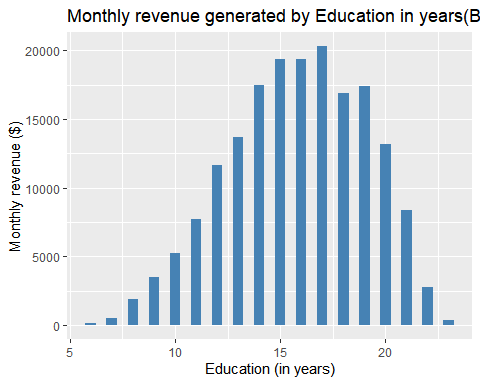
## Total monthly revenue for household size > 1 and with no multi-line:: 29578.8030598861

groupedData <- customersForAnalysis%>%  
 filter(HouseholdSize > 1) %>%  
 filter(Multiline == "No")  
numberOfCustomers <- count(groupedData)  
cat(sprintf("\nTotal monthly revenue for household size > 1 and with no multi-line: %s", numberOfCustomers))

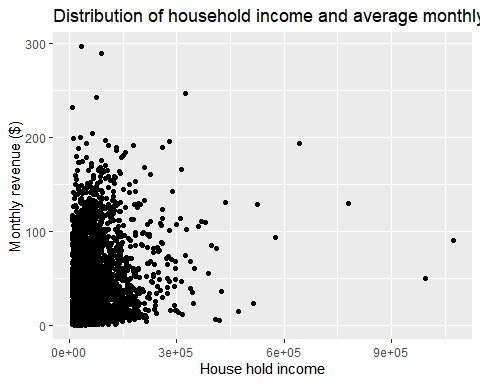
##   
## Total monthly revenue for household size > 1 and with no multi-line: 1481

#Statistical analysis for Education

groupedData <- customersForAnalysis %>%   
 group\_by(EducationYears) %>%   
 summarise(totalMonthlyValue= sum(AverageMonthlyRevenue))   
  
ggplot(groupedData, aes(x = EducationYears, y = totalMonthlyValue)) +   
 geom\_bar(stat="identity", width=0.5, fill = "steelblue") +  
 labs( title= "Monthly revenue generated by Education in years(Buckets of 5)", x="Education (in years)", y = "Monthly revenue ($)")

 #Statistical analysis for household income

ggplot(customersForAnalysis, aes(x = HHIncome, y = AverageMonthlyRevenue)) +   
 geom\_point() +  
 labs( title= "Distribution of household income and average monthly revenue", x="House hold income", y = "Monthly revenue ($)")



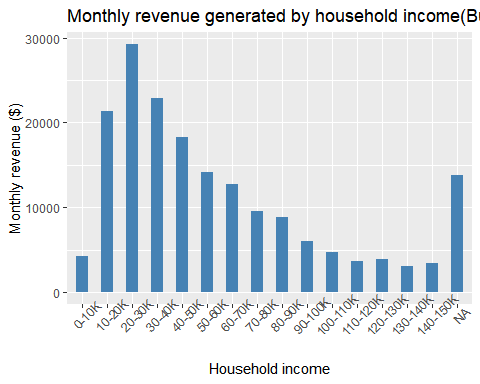
lbls <- c('0-10K','10-20K','20-30K', '30-40K','40-50K','50-60K','60-70K', '70-80K', '80-90K', '90-100K', '100-110K', '110-120K', '120-130K', '130-140K', '140-150K')  
  
groupedData <- customersForAnalysis %>%   
 group\_by(hhIncomeGroup = cut(HHIncome, breaks= seq(0, 150000, by = 10000),   
 right = TRUE, include.lowest = TRUE, labels = lbls) ) %>%   
 summarise(totalMonthlyValue= sum(AverageMonthlyRevenue))

## Warning: Factor `hhIncomeGroup` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

groupedData

## # A tibble: 16 x 2  
## hhIncomeGroup totalMonthlyValue  
## <fct> <dbl>  
## 1 0-10K 4242.  
## 2 10-20K 21358.  
## 3 20-30K 29181.  
## 4 30-40K 22904.  
## 5 40-50K 18322.  
## 6 50-60K 14142.  
## 7 60-70K 12765.  
## 8 70-80K 9487.  
## 9 80-90K 8831.  
## 10 90-100K 6005.  
## 11 100-110K 4727.  
## 12 110-120K 3595.  
## 13 120-130K 3932.  
## 14 130-140K 3007.  
## 15 140-150K 3419.  
## 16 <NA> 13793.

ggplot(groupedData, aes(x = hhIncomeGroup, y = totalMonthlyValue)) +   
 geom\_bar(stat="identity", width=0.5, fill = "steelblue") +  
 labs( title= "Monthly revenue generated by household income(Buckets of $10K)", x="Household income", y = "Monthly revenue ($)") +  
 theme(axis.text.x = element\_text(angle = 45))

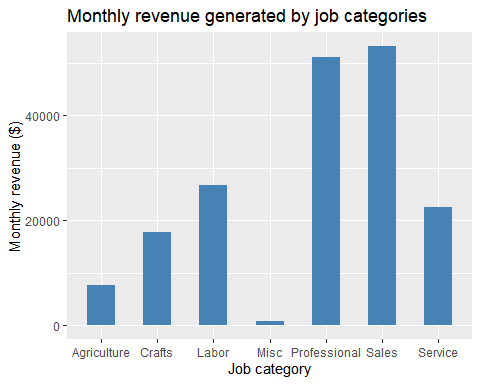


#Statistical analysis for Job category

groupedData <- customersForAnalysis%>%  
group\_by(JobCategory)%>%   
summarise(totalMonthlyValue = sum(AverageMonthlyRevenue))  
  
groupedData

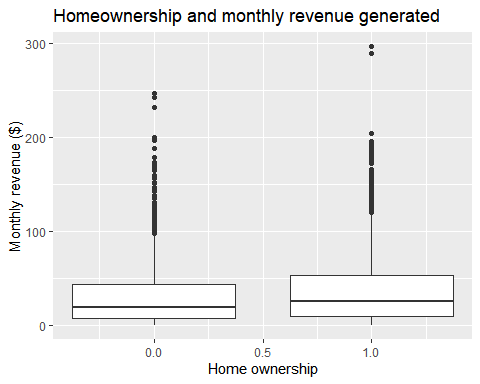
## # A tibble: 7 x 2  
## JobCategory totalMonthlyValue  
## <chr> <dbl>  
## 1 Agriculture 7607.  
## 2 Crafts 17732.  
## 3 Labor 26772.  
## 4 Misc 797.  
## 5 Professional 51108.  
## 6 Sales 53195.  
## 7 Service 22499.

ggplot(groupedData, aes(x = JobCategory, y = totalMonthlyValue)) +   
 geom\_bar(stat="identity", width=0.5, fill = "steelblue") +  
 labs( title= "Monthly revenue generated by job categories", x="Job category", y = "Monthly revenue ($)")



#Statistical analysis for home ownership

ggplot(data = customersForAnalysis, aes(x = HomeOwner, y = AverageMonthlyRevenue, group=HomeOwner)) +  
 geom\_boxplot() +  
 labs( title= "Homeownership and monthly revenue generated", x="Home ownership", y = "Monthly revenue ($)")

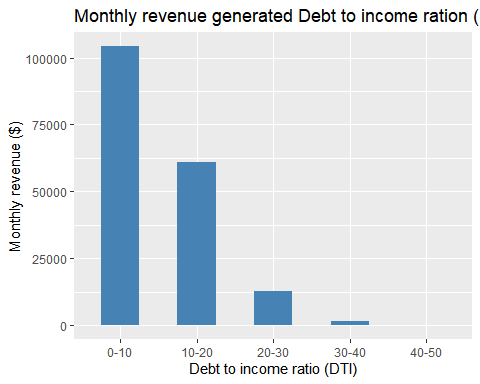


customersForAnalysis%>%  
group\_by(HomeOwner)%>%   
summarise(SUm = sum(AverageMonthlyRevenue), Mean=mean(AverageMonthlyRevenue), Median=median(AverageMonthlyRevenue))

## # A tibble: 2 x 4  
## HomeOwner SUm Mean Median  
## <dbl> <dbl> <dbl> <dbl>  
## 1 0 59105. 32.0 19.1  
## 2 1 120605. 38.3 25.8

#Statistical analysis for Debt to income ratio

lbls <- c('0-10','10-20','20-30', '30-40','40-50','50-60','60-70', '70-80', '80-90', '90-100')  
groupedData <- customersForAnalysis %>%   
 group\_by(debtToIncomeRatioGroup = cut(DebtToIncomeRatio, breaks= seq(0, 100, by = 10),   
 right = TRUE, include.lowest = TRUE, labels = lbls) ) %>%   
 summarise(totalMonthlyValue= sum(AverageMonthlyRevenue))   
  
ggplot(groupedData, aes(x = debtToIncomeRatioGroup, y = totalMonthlyValue)) +   
 geom\_bar(stat="identity", width=0.5, fill = "steelblue") +  
 labs( title= "Monthly revenue generated Debt to income ration (Buckets of 10)", x="Debt to income ratio (DTI)", y = "Monthly revenue ($)")



#Export the dataset used for analysis

write\_csv(customersForAnalysis,"data/Customer\_Dataset\_UPDATED.csv")