Data Due Diligence Project

DSA 5100 – Data Exploration

Owen Evans, Data Science Student

6/12/2021

Introduction. The following exploratory data analysis (EDA) and data transformation exercise was conducted on a curated dataset obtained from the Newco Telecom Company. The EDA is the first step in transforming this trove of customer information and business transactions into key knowledge and actionable insights. Prior to the EDA, it is imperative to understand the overall business objectives that underpin this data analysis. Of particular interest is the identification and profiling of high value customers in order to provide a basis for effective segmentation strategies. By knowing the key characteristics of a high-value customer segment, we can then provide for more targeted and more effective marketing campaigns. Similarly, the dataset can also be utilized to profile customers with stagnant or mediocre revenue performance, possibly allowing for a prediction of churn. Finally, transactional data can be analyzed to determine utilization of key services, allowing the business to focus primarily on marketing the most lucrative services.

Variable Grouping. The original telecom dataset contained 59 variables detailing key information for 5000 customers. The variables in dataset follow a natural grouping; those variables that describe key demographic, financial, lifestyle and geographic customer characteristics and those variables that describe the business relationship of customers via transactional/revenue data. The variables in this Telecom dataset were thus regrouped according to the schema shown in Figure 1.

Identifier	Geographic	Demographic - Financial	Demographic	Lifestyle	Transactional / Business Relationship
CustomerID	Region	HHIncome	Gender	NumberPets	PhoneCoTenure
	TownSize	DebtToIncomeRatio	Age	NumberCats	VoiceLastMonth
		CreditDebt	EducationYears	NumberDogs	VoiceOverTenure
		OtherDebt	JobCategory	NumberBirds	EquipmentRental
		LoanDefault	UnionMember	CarsOwned	EquipmentLastMonth
		CreditCard	EmploymentLength	CarOwnership	EquipmentOverTenure
		CardTenure	Retired	CarBrand	CallingCard
		CardItemsMonthly	MaritalStatus	CarValue	WirelessData
		CardSpendMonth	HouseholdSize	CommuteTime	DataLastMonth
			HomeOwner	ActiveLifestyle	DataOverTenure
			PoliticalPartyMem	TVWatchingHours	Multiline
			Votes	OwnsPC	VM
				OwnsMobileDevice	Pager
				OwnsGameSystem	Internet
				OwnsFax	CallerID
				NewsSubscriber	CallWait
					CallForward
					ThreeWayCalling
					EBilling

Figure 1. Regrouping of variables in Customer Dataset.

Data Cleansing – Missing Values. The dataset in question contains 131 missing values spread across 11 separate variables. In all cases, the number of missing values were relatively negligible, comprising no more that 0.7% of all observations for any particular variable. Due to the low levels of missing values and their random nature within the dataset, it would be appropriate to eliminate all observations containing missing values. However, upon further inspection, they are a few variables that have missing values for clear reasons. For the purposes of this EDA, the four variables with the highest proportion of missing values were thus targeted for recoding and/or imputation. The specifics of these transformations are detailed below.

Gender - this variable contains the highest number of missing values (33). Missing values in this case were attributed to the fact that gender can no longer be considered a binary variable. In this case, all missing values were attributed to those folks who do not wish to declare gender. As such, an alternative ("Undeclared") level was added to this variable, with all missing values assigned to this level.

Job Category - Similarly, the variable *JobCategory* is likely insufficient to cover all job types within the specific six levels provided in the original dataset. As such, a seventh level ("Other") was added to this variable, with all missing values assigned to this level.

HomeOwner & NumberBirds - For both of these variables, all missing values were imputed based upon the most frequent response for each variable. In the case of *HomeOwner*, all missing values were attributed to the affirmative (Yes). For *NumberBirds*, all missing values were attributed to zero, the most frequent response for this variable. All other observations with missing values in this dataset were removed.

Data Quality Issues. Variables detailing information regarding a customer's car ownership, car brand and car value contain ambiguous negative values (-1 or -1000). The origin of this data issue derives from the fact that one cannot define the type of car ownership, the car value or the car brand for those customers that do not actually own vehicles. Upon further inspection, all of the suspect responses for these particular variables were obtained from customers that had a zero response for the *CarsOwned* variable. To rectify this issue, an additional level ("None") was added to both *CarOwnership* and *CarBrand*. Additionally, the *CarValue* response for those customers that do not own vehicles (*CarsOwned*) was changed from -1000 to 0.

The categorical variable, *Internet*, is factored with five separate and potentially ambiguous levels (no, yes, 2, 3, 4). In this case, the meaning of the additional numeric levels is not immediately obvious. It may be assumed that those customers responding with 2,3 or 4 are indicative of various levels of internet service. Accordingly, the variable Internet was recoded to a binary variable by converting all those who responded 2,3, or 4 to the affirmative (yes).

Finally, it was recognized that a zero response for *PhoneCoTenure* may complicate future efforts to derive additional information via feature engineering. For instance, a calculation of the average revenue per user (ARPU) would ultimately utilize *PhoneCoTenure* as a denominator. It was assumed that all customers responding to this survey with zero for *PhoneCoTenure* must have some fractional tenure with the company. As such, all zero responses for this variable were converted (or rounded) to a value of one month.

Feature Engineering. A key performance indicator for the telecom business is the average revenue per user or ARPU. The ARPU metric is considered a key business metric for the telecom industry, as it potentially indicates high value users that tend to generate significant revenue at lower costs. For the purposes of this analysis, a monthly ARPU metric was calculated as the total revenue attributed to Voice, Data and Equipment divided by PhoneCoTenure expressed in months. Two new variables were thus added to the original data frame: (1) Total Revenue and (2) ARPU. The variable ARPU was observed to have non-normal distribution with a significant right skew and was thus log transformed. A new ARPU_Log variable with a near normal distribution of values was also added to the dataset.

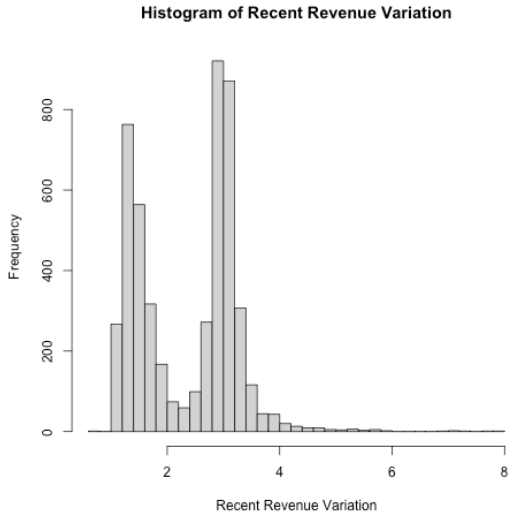


Figure 2. Histogram of RecentRevenueVariation (RRV) outcomes indicating a bimodal distribution.

The ARPU metric as calculated above will afford an average monthly revenue per user over the course of their tenure with the business. Of particular interest to this data analysis, is any recent significant deviation (i.e. last month revenue) in customer revenue relative to the average value (*ARPU*). Included in the dataset is the latest month's

revenue from Voice (*VoiceLastMonth),* Data (*DataLastMonth)* and Equipment (*EquipmentLastMonth)*. A comparison of the summation of these latest revenue metrics to the calculated ARPU (average revenue) for the same customer will afford a measure of recent revenue growth on a per customer basis. To capture this metric, an additional numeric variable denoted as *RecentRevenueVariation* was derived as the ratio of the total revenue per user from the latest month and the average value (ARPU) over the course of the customer's tenure with the company. Values of *RecentRevenueVariation (RRV)* of two or less will identify customers with stagnant or mediocre revenue growth that may be subject to churn. Conversely, RRV values that are higher than two may indicate customers with significantly high recent revenue growth, a segment that would be a ideal target for retention activities. Interestingly, the distribution of *RecentReveunueVariation* (Figure 2) is bimodal and contains two natural customer segments (high growth and low growth). To capture these customer segments, another categorical and binary variable was added to dataset (*RRV_Cat*).

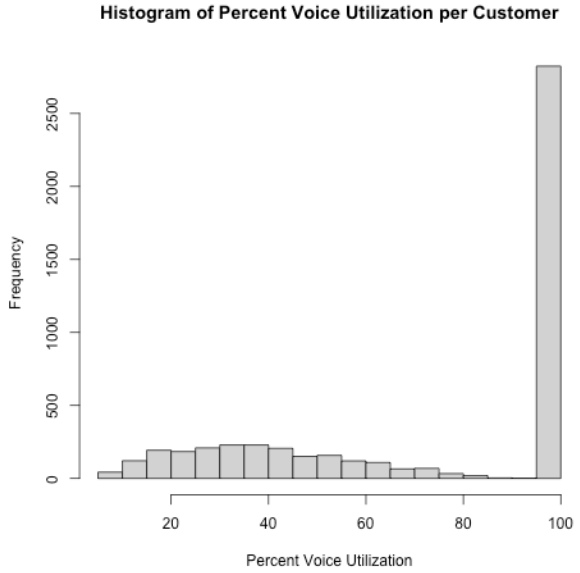


Figure 3. Histogram of Voice Utilization values indicating a bimodal distribution.

It may also be useful to identify those customers that primarily utilize our most lucrative service. To enable this, we first determined the percentage revenue attributed to Voice, Data and Equipment transactions in aggregate for the latest month. This calculation indicated values of 63, 17 and 20% for Voice, Data and Equipment, respectively. Since transactions related to Voice comprise a majority of the recent revenue for Newco Telecom, it is of utility to calculate a Voice Utilization (*VoiceUtil)* metric for the customer base, in order to identify and profile those customers that utilize our most lucrative service. As such, a new continuous numeric variable, *VoiceUtil*, calculated for all customers as the percentage of the latest month total revenue derived solely from Voice transactions. Interestingly, a histogram of the values of *VoiceUtil* also exhibited a noticeable bimodal nature, with nearly half of the values indicating a 100% utilization (Figure 3).

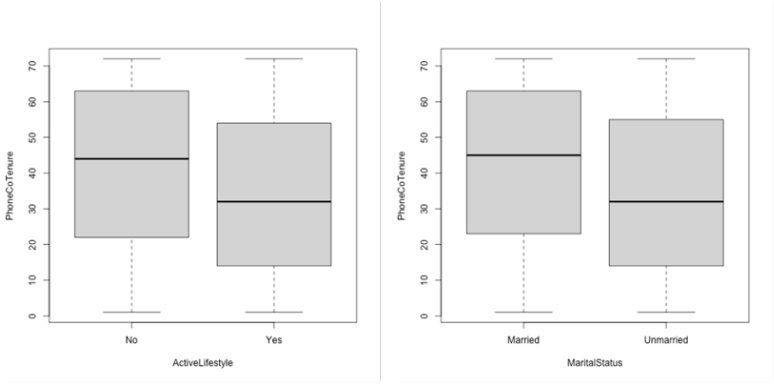
Finally, since household income (*HHIncome*) may be a key explanatory variable in this data analysis, further variable transformation may be necessary to allow for future hypothesis testing. Specifically, it was determined that the values of *HHIncome* are not normally distributed and possess a significant right skew. As such, a power transformation (*HHIncomeTrans*) was conducted to afford a nearly normal distribution of values.

Summary Statistics. Univariate summary statistics was conducted on all numeric variables and a select number of categorical variables. Summary statistics for all numeric variables included efforts to identify those variables with significant skew and/or the possible presence of outliers. Outlier detection was facilitated by identifying those variables with a significant number of values that were 1.5*IQR above the third quantile (Q3) or 1.5*IQR below the first quantile (Q1), supported with follow-up boxplot visualizations. A number of numeric variables did exhibit skew (HHIncome, ARPU) and were corrected with the appropriate log or power transformation. Other numeric

variables were also right skewed by virtue of having a preponderance of legitimate zero values. These variables were not transformed and kept as-is.

Frequency and proportion tables for the following categorical variables were also provided: *AgeBin, Gender, HHIncomeBin, Region and Equipment Rental.* In all cases, except for *EquipmentalRental*, these variables exhibit reasonably balanced outcomes across all levels.

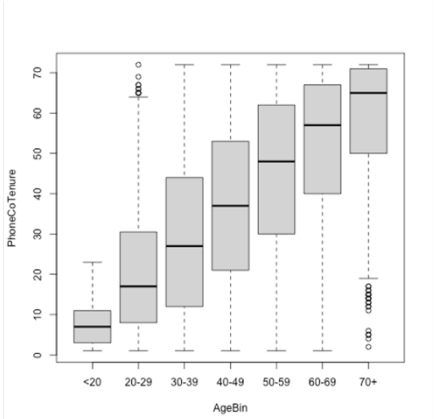
Hypothesis Testing. The engineered dataset now contains a total of 69 separate variables and 4967 observations. All missing values were eliminated from the dataset via recoding, imputation or elimination. All data quality issues were addressed, and a number of new variables were added via feature engineering. The newly cleaned engineered dataset is now ripe for further insight via hypothesis testing.



data: PhoneCoTenure by ActiveLifestyle
t = 11.475, df = 4871.6, p-value < 2.2e-16

data: PhoneCoTenure by MaritalStatus
t = 12.161, df = 4952.4, p-value < 2.2e-16

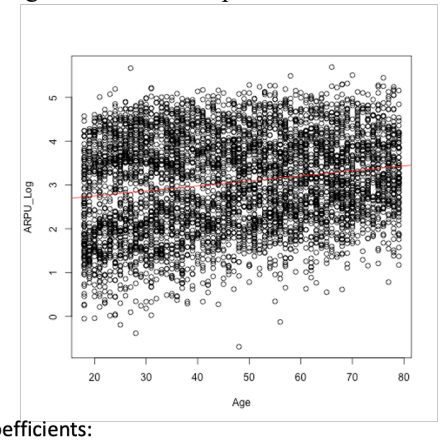
Specifically, the data can be utilized to identify those customers with high loyalty or long tenure. To that end, a two-sample t-test was conducted to determine if the difference in means for *PhoneCoTenure* was statistically significant across the levels of both *Marital Status* and *Active Lifestyle*. The results of both tests afforded a high value for each t-statistic and a very low p-value, indicating that the null hypothesis (i.e., the means are equal) is rejected. While it is clear that marital status and active lifestyle have an influence on customer loyalty, the reader is cautioned that these variables may ultimately be a proxy for age. In fact, one-way analysis of variance (ANOVA) testing of *PhoneCoTenure across* the various age bins indicate that difference in means for PhoneCoTenure are in fact



One-way analysis of means (not assuming equal variances)
data: PhoneCoTenure and AgeBin
F = 1128.2, num df = 6.0, denom df = 1811.3, p-value < 2.2e-16

significant. The data and results of this hypothesis testing may indicate a clear relationship between customer loyalty and age. Although unconfirmed, this may possibly indicate that younger folks are more prone to churn.

Contained within the engineered data set is the new variable ARPU_Log (average revenue per user, log transformed), which ultimately affords an individualized metric for average customer value. An attempt was made to determine if ARPU_Log was correlated to *Customer Age* via linear regression. Although this relationship was deemed to be statistically significant via a low p-value, the linear model is poorly fitted with a very low coefficient of determination (R^2). Unfortunately, the variable age alone cannot explain all of the variance associated with the ARPU_Log metric.



Coefficients:

	Estimate	Std. Error.	t value	Pr(>|t|)
(Intercept)	2.521136	0.043207	58.35	<2e-16 ***
Age	0.011441	0.000859	13.32	<2e-16 ***

R2 = 0.03

One of the key insights from this data analysis is derived from the engineered variable *RecentRevenueVariation (RRV).* This new variable was intended to identify a segment of customers that exhibited recent revenue growth (i.e., over the last month) that is substantially different than average. It was observed that the outcomes of the RRV metric were bimodal in nature, indicating that a large group of customers exhibited recent revenue values that were many multiples of their respective average (Figure 2). A secondary categorical variable was derived (*RRV_Cat*) to capture those populations exhibiting low (<2.1) and high (≥2.1) RRV values. It was hypothesized that *AgeBin* and *RRV_Cat* are not independent, that older folks were primarily responsible for the recent surge in revenue growth. As such, a chi-squared test of independence was conducted on these two categorical variables, yielding a chi-squared statistic of 132.06 and a very low p-value. Based upon these results and the frequency table shown in Table 1, it was determined that the group of High RRV customers (2850 observations) seems to be derived primarily from older (>50 years) age groups.

Table 1. Frequency Table of RRV_Cat Values as a function of age category.

	<20	20-29	30-39	40-49	50-59	60-69	70
HighRRV	113	434	446	393	481	503	480
LowRRV	67	449	434	378	319	287	183

Finally, it was of interest to determine the service utilization of those customers with high RRV values. Specifically, we were interested in whether these high value customers were primarily utilizing our most lucrative service, Voice. In order to explore this further, a chi-squared test of independence was conducted between the *RRV_Cat* and *VoiceUtilCat*. The results of this hypothesis test yielded a chi-squared statistic of 4587.1 and a very low *p*-value. These results, coupled with the frequency table in Table 2, indicate that these categorical variables are not independent. More specifically, it appears that the surge of recent revenue (relative to average, ARPU) from a select group of customers is derived from those that use our voice service exclusively (100% Voice Utilization). The results above also suggest that this recent revenue surge is likely driven by older folks. The results of this hypothesis testing

suggest that further marketing of our voice service to customers within these particular age groups may ultimately lead to greater revenue growth.

Table 2. Frequency table of Voice Utilization as a function of RRV category.

	High Voice Utilization	Low Voice Utilization
HighRRV	2791	59
LowRRV	35	2082

END

Appendix

Owen R. Evans

6/13/2021

Appendix - RScript for Data Due Diligence

```
library(tidyverse)
library(openxlsx)
library(ggplot2)
library(janitor)
library(dlookr)
library(psych)
library(car)
library(summarytools)
library(corrplot)

Data1 <- read.xlsx("CustomerData.xlsx", sheet = 1,
                   na.strings = c("","NA","#NULL!"))

########################
# Regrouping Variables #
########################

# Group 1 - Identifer
ID <- "CustomerID"

# Group 2 - Geographic
Geo <- c("Region","TownSize")

# Group 3 - Demographic, Financial
DemoFin <- c("HHIncome",
             "DebtToIncomeRatio",
             "CreditDebt",
             "OtherDebt",
             "LoanDefault",
             "CreditCard",
             "CardTenure",
             "CardItemsMonthly",
             "CardSpendMonth")

#Group 4 - Demographic
Demo <- c('Gender',
          'Age',
          'EducationYears',
          'JobCategory',
          'UnionMember',
          'EmploymentLength',
          'Retired',
          'MaritalStatus',
          'HouseholdSize',
          'HomeOwner',
          'PoliticalPartyMem',
          'Votes')

# Group 5 - Lifestyle
Life <- c('NumberPets',
          'NumberCats',
```

```
          'NumberDogs',
          'NumberBirds',
          'CarsOwned',
          'CarOwnership',
          'CarBrand',
          'CarValue',
          'CommuteTime',
          'ActiveLifestyle',
          'TVWatchingHours',
          'OwnsPC',
          'OwnsMobileDevice',
          'OwnsGameSystem',
          'OwnsFax',
          'NewsSubscriber')

#Group 6 - Transactional / Business
Business <- c('PhoneCoTenure',
              'VoiceLastMonth',
              'VoiceOverTenure',
              'EquipmentRental',
              'EquipmentLastMonth',
              'EquipmentOverTenure',
              'CallingCard',
              'WirelessData',
              'DataLastMonth',
              'DataOverTenure',
              'Multiline',
              'VM',
              'Pager',
              'Internet',
              'CallerID',
              'CallWait',
              'CallForward',
              'ThreeWayCalling',
              'EBilling')

# Reorder Data Set
col_order <- c(ID,Geo,Demo, DemoFin, Life, Business)
Data1 <- Data1[, col_order]

################
# Missing Data #
################

# How Many Missing Values? - 131 across the entire data set
# Where are the Missing Values located?
sum(is.na(Data1))
missing <- sapply(Data1, function(x) sum(is.na(x)))
missing <- missing[missing != 0]
suspect_vars <- names(missing)

missing_perc <- sapply(Data1, function(x) sum(is.na(x))/length(x))*100
missing_perc <- missing_perc[missing_perc > 0.0]

# Focus on these variables (top 3) to fix missing values - Gender, Birds, Homeowner
# For Gender - act another factor level
str(Data1$Gender)
Data1 <- Data1 %>% mutate(Gender = ifelse(is.na(Gender), "Undeclared", Gender))
table(Data1$Gender)
```

```
# For Birds - convert no response to zero
table(Data1$NumberBirds)
Data1 <- Data1 %>% mutate(NumberBirds = ifelse(is.na(NumberBirds), 0, NumberBirds))

# Homeownership - Inspect Characteristics of NA's
# Replace with most frequent value (mode)
HomeNA <- filter(Data1, is.na(HomeOwner))
table(Data1$HomeOwner) # frequency of yes (1) is 1.7x that of no (0)
Data1 <- mutate(Data1,
                HomeOwner = ifelse(is.na(HomeOwner), 1, HomeOwner))
# Convert NA's to NO
remove(HomeNA)
Data1$HomeOwner <- as.factor(Data1$HomeOwner)
levels(Data1$HomeOwner) <- c("No", "Yes") # recode 0,1 to No,Yes

# NA treatment Job Category
# Categories are not extensive
# N/A might be other
str(Data1$JobCategory)
unique(Data1$JobCategory)
JobNA <- Data1 %>% filter(is.na(JobCategory))
View(JobNA)
Data1 <- mutate(Data1,
         JobCategory = ifelse(is.na(JobCategory), "Other", JobCategory))
remove(JobNA)

# All Other NA's - Eliminate from Dataset,  Very low frequency
Data1 <- na.omit(Data1)

##################
# Data Structure #
##################
# Will need to recode 33 categorical variables to factors
str(Data1)

z <- c("Region","TownSize", "Gender",
       "JobCategory", "UnionMember",
       "Retired", "MaritalStatus",
       "HouseholdSize", "HomeOwner",
       "PoliticalPartyMem", "Votes",
       "LoanDefault", "CreditCard",
       "CarOwnership", "CarBrand",
       "ActiveLifestyle", "OwnsPC",
       "OwnsMobileDevice", "OwnsGameSystem",
       "OwnsFax", "NewsSubscriber", "EquipmentRental",
       "CallingCard", "WirelessData", "Multiline",
       "VM", "Pager", "Internet", "CallerID",
       "CallWait", "CallForward", "ThreeWayCalling",
       "EBilling")

Data1[z]<-lapply(Data1[z],factor)

#################
# Data Problems #
#################

# Internet Recoding
# Convert 2,3,4 to yes values,
# assuming that any other answer than zero indicates the affirmative
table(Data1$Internet)
```

```
str(Data1$Internet)

Data1 <- Data1 %>% mutate (Internet = case_when(Internet == "2" ~ "Yes",
                                        Internet == "3" ~ "Yes",
                                        Internet == "4" ~ "Yes",
                                        Internet == "Yes" ~ "Yes",
                                        Internet == "No" ~ "No"))

# Car Ownership, Car Brand
# Change the -1 to None - Assuming these are folks who don't own cars
levels(Data1$CarOwnership) <- c("None", "Lease", "Own")
levels(Data1$CarBrand) <- c("None", "Domestic", "Foreign")

# CarValue - Replace -1000 values to 0
Data1 <- Data1 %>% mutate(CarValue= ifelse(CarValue==-1000, 0, CarValue))

# PhoneCoTenure - Eliminate Zero Values - NonNegative Value Needed for ARPU
Data1 <- mutate(Data1, PhoneCoTenure =
                  replace(PhoneCoTenure, PhoneCoTenure == 0, 1))

####################################
# Univariate Descriptive Statistics #
####################################
# Use pysch:describe to afford summary table

# Numeric, Continuous Variables
Data1_Num <- select_if(Data1, is.numeric)
SumTableNum <- psych::describe(Data1_Num,
                               IQR=TRUE,
                               quant=c(0.25,0.75))

names(SumTableNum) <- c( "Vars", "n", "Mean", "Standard Deviation",
                         "Median", "Trimmed Mean" , "Mad", "Min" ,
                         "Max", "Range", "Skew" , "Kurtosis",
                         "SE" , "IQR", "Q1", "Q3")

write.csv(SumTableNum, "SummaryTableNum.csv")
View(SumTableNum)


#############################################################
# Inspect Key Numerics for Skew or Outliers #
#############################################################

# Find possible outliers via +/- 1.5 IQR Values

SumTableNum <- mutate (SumTableNum, Outliers = case_when (
         Max > (Q3+(1.5 * IQR))| Min < (Q1-(1.5 * IQR)) ~ "Yes",
         TRUE ~ "No"))

Outliers <- filter(SumTableNum, Outliers == "Yes")
rownames(Outliers)

boxplot(Data1$HHIncome) # High Earners
boxplot(Data1$EmploymentLength) # Most subscribers are new workers
boxplot(Data1$HouseholdSize) # A few larger families
boxplot(Data1$VoiceOverTenure) # Zero Inflated
boxplot(Data1$CreditDebt) # Zero Inflated
boxplot(Data1$TVWatchingHours) # A few extremes
```

```
#####################################################
# Variable Transformations and Feature Engineering #
#####################################################

# HHIncome, Right Skewed Data
# Power transformed to normality
hist(Data1$HHIncome)
summary(powerTransform(Data1$HHIncome))  # Max Likelihood/BoxCox
Data1$TransHHIncome <- Data1$HHIncome ^ -0.1708
hist(Data1$TransHHIncome)

# HHIncome,  Binned, Convert Continuous to Discrete
# Potentially important for segmentation exercise

Data1 <- mutate(Data1, HHIncomeBin = case_when(
  HHIncome < 20000                        ~ "<$20,000",
  HHIncome >= 20000 & HHIncome <40000     ~ "$20,000-$39,999",
  HHIncome >= 40000 & HHIncome <60000     ~ "$40,000-$59,999",
  HHIncome >= 60000 & HHIncome <80000     ~ "$60,000-$79,999",
  HHIncome >= 80000 & HHIncome <100000    ~ "$80,000-$99,999",
  HHIncome >= 100000                     ~ ">$100,000"
  )
)

Data1$HHIncomeBin <- as.factor (Data1$HHIncomeBin)
table(Data1$HHIncomeBin)
Data1$HHIncomeBin <- factor(Data1$HHIncomeBin,
                        levels = c("<$20,000", "$20,000-$39,999",
                        "$40,000-$59,999", "$60,000-$79,999",
                        "$80,000-$99,999", ">$100,000")
                        )
View(Data1[,c("HHIncome","HHIncomeBin")])

# Binned Age Group
Data1 <- mutate(Data1, AgeBin = case_when(
  Age < 20                  ~ "<20",
  Age >= 20 & Age <30        ~ "20-29",
  Age >= 30 & Age <40       ~ "30-39",
  Age >= 40 & Age <50       ~ "40-49",
  Age >= 50& Age  <60        ~ "50-59",
  Age >= 60 & Age <70       ~ "60-69",
  Age >=70                  ~ "70+"
  )
)

str(Data1$AgeBin)
Data1$AgeBin <- as.factor (Data1$AgeBin)
table(Data1$AgeBin)

# Adding Total Revenue per User Over PhoneCoTenure
# Adding ARPU - Average Revenue per User - Total Revenue/Tenure
# Adding Recent Rev Variation
# How much does the last month revenue per user differ from average?

Data1 <- mutate(Data1, Total_Revenue = VoiceOverTenure + EquipmentOverTenure +
                DataOverTenure)
Data1 <- mutate(Data1, ARPU = Total_Revenue/PhoneCoTenure)
Data1 <- mutate(Data1, RecentRevenueVariation = ((DataLastMonth+VoiceLastMonth+
                                                    EquipmentLastMonth)/ARPU))
```

```r
Data1[,c("ARPU", "RecentRevenueVariation", "Total_Revenue")]

hist(Data1$RecentRevenueVariation, breaks=30) # RRV is bimodal
Data1 <- mutate (Data1, RRV_Cat = case_when (
  RecentRevenueVariation <2.1  ~ "LowRRV",
  RecentRevenueVariation >= 2.1 ~ "HighRRV"
))

hist(Data1$ARPU) # Heavy Right Skew in ARPU
hist(log(Data1$ARPU)) # Normalized
Data1$ARPU_Log <- log(Data1$ARPU) # log transformed

# Determine most lucrative service
# Add new variable to track those users that use this service

attach(Data1)
x <- sum(VoiceLastMonth, DataLastMonth, EquipmentLastMonth)
y <- c(sum(VoiceLastMonth), sum(DataLastMonth), sum(EquipmentLastMonth))
z <- y/x
a <- c("Voice", "Data", "Equipment")
names(z) <- a
detach(Data1)
print(z)

Data1 <- mutate(Data1, VoiceUtil =
                  VoiceLastMonth/(VoiceLastMonth+
                                    DataLastMonth+
                                    EquipmentLastMonth)*100)

# Voice Utilization is Also Bimodal - lots of just voice customers
png("VoiceHist.png")
VoiceHist <- hist(Data1$VoiceUtil, breaks=30,
                  main="Histogram of Percent Voice Utilization per Customer",
                  xlab="Percent Voice Utilization")
dev.off()

# New Variable to indicate high/low voice utilization
Data1 <- mutate (Data1, VoiceUtilCat = case_when (
  VoiceUtil <90  ~ "Low Voice Utilization",
  VoiceUtil >= 90 ~ "High Voice Utilization"
))



####################################
# Regrouping Variables w/ New Adds #
####################################

# Group 1 - Identifer
ID <- "CustomerID"

# Group 2 - Geographic
Geo <- c("Region","TownSize")

# Group 3 - Demographic, Financial
DemoFin <- c("HHIncome",
             "TransHHIncome",
             "HHIncomeBin",
             "DebtToIncomeRatio",
             "CreditDebt",
```

```
             "OtherDebt",
             "LoanDefault",
             "CreditCard",
             "CardTenure",
             "CardItemsMonthly",
             "CardSpendMonth")

#Group 4 - Demographic
Demo <- c('Gender',
          'Age',
          'AgeBin',
          'EducationYears',
          'JobCategory',
          'UnionMember',
          'EmploymentLength',
          'Retired',
          'MaritalStatus',
          'HouseholdSize',
          'HomeOwner',
          'PoliticalPartyMem',
          'Votes')

# Group 5 - Lifestyle
Life <- c('NumberPets',
          'NumberCats',
          'NumberDogs',
          'NumberBirds',
          'CarsOwned',
          'CarOwnership',
          'CarBrand',
          'CarValue',
          'CommuteTime',
          'ActiveLifestyle',
          'TVWatchingHours',
          'OwnsPC',
          'OwnsMobileDevice',
          'OwnsGameSystem',
          'OwnsFax',
          'NewsSubscriber')

#Group 6 - Transactional / Business
Business <- c('PhoneCoTenure',
              'VoiceLastMonth',
              'VoiceOverTenure',
              'EquipmentRental',
              'EquipmentLastMonth',
              'EquipmentOverTenure',
              'CallingCard',
              'WirelessData',
              'DataLastMonth',
              'DataOverTenure',
              'Multiline',
              'VM',
              'Pager',
              'Internet',
              'CallerID',
              'CallWait',
              'CallForward',
              'ThreeWayCalling',
              'EBilling',
```

```
              'Total_Revenue',
              'ARPU',
              'ARPU_Log',
              'RecentRevenueVariation',
              'VoiceUtil',
              'RRV_Cat',
              'VoiceUtilCat')

# Reorder Data Set with Engineered Variables
col_order <- c(ID,Geo,Demo, DemoFin, Life, Business)
Data1 <- Data1[, col_order]

#################################
# Summary Stats w/ New Variables #
#################################

# Continuous Variables
Data1_Num <- select_if(Data1, is.numeric)
SumTableNum2 <- psych::describe(Data1_Num,
                                IQR=TRUE,
                                quant=c(0.25,0.75))

names(SumTableNum2) <- c( "Vars", "n", "Mean", "Standard Deviation",
                          "Median", "Trimmed Mean" , "Mad", "Min" ,
                          "Max", "Range", "Skew" , "Kurtosis",
                          "SE" , "IQR", "Q-25th", "Q-75th")

write.csv(SumTableNum2, "SummaryTableNum2.csv")
View(SumTableNum2)

#################################
# Selected Categorical Variables #
#################################

Data1_Cat <- select_if(Data1, is.factor)

# Age Breakdown
AgeSum <- summarytools::freq(Data1$AgeBin)
write.csv (AgeSum, "AgeSum.csv")

# Breakdown of Income
IncomeSum <- summarytools::freq(Data1$HHIncomeBin)
write.csv (IncomeSum, "IncomeSum.csv")

# Breakdown of Gender
GenderSum <- summarytools::freq(Data1$Gender)
write.csv (GenderSum, "GenderSum.csv")

# Breakdown of Region
RegionSum <- summarytools::freq(Data1$Region)
write.csv (RegionSum, "RegionSum.csv")

# Breakdown of Equipment Rental
EquipSum2 <- summarytools::freq(Data1$EquipmentRental)
write.csv (EquipSum2, "EquipSum2.csv")

######################
# Hypothesis Testing #
######################
```

```
# Phone Company Tenure vs. Age, Linear Regression
# What type of customers are less prone to churn?
# Married customers...
attach(Data1)
plot(PhoneCoTenure~ActiveLifestyle)
result <- t.test(PhoneCoTenure~ActiveLifestyle, data = Data1)
result

png("FTT.png")
plot(PhoneCoTenure~ActiveLifestyle)
dev.off()

plot(PhoneCoTenure~MaritalStatus)
result <- t.test(PhoneCoTenure~MaritalStatus, data = Data1)
result

png("FFF.png")
plot(PhoneCoTenure~MaritalStatus)
dev.off()

png("Age.png")
plot(PhoneCoTenure~AgeBin)
dev.off()
Tenure_age.aov <- aov(PhoneCoTenure~AgeBin) # unequal variances
summary(Tenure_age.aov)
TukeyHSD(Tenure_age.aov)

detach(Data1)

# What explanatory variables are correlated to ARPU?
# correlogram
Data1_Num <- select_if(Data1, is.numeric)
res <- cor(Data1_Num)
corrplot(res,type="upper", is.corr = FALSE)

# ARPU ~ Age (linear regression)
attach(Data1)
png("Arpu_LM.png")
plot(ARPU_Log~Age)
abline(lm(ARPU_Log~Age), col="red")
dev.off()
ARPU_Age.lm <- lm(ARPU_Log~Age)
summary(ARPU_Age.lm) # Low R2,  but significant
plot(ARPU_Age.lm) # Check for lm assumptions
detach(Data1)

# Check with Age Bins, ANOVA
attach(Data1)
plot(ARPU_Log~AgeBin)
ARPU_age.aov <- aov(ARPU_Log~AgeBin)
summary(ARPU_age.aov)
TukeyHSD(ARPU_age.aov)
detach(Data1)

# What explanatory variables are correlated
# to a High or Low Recent Revenue Variation Value?

hist(Data1$RecentRevenueVariation, breaks=30) # bimodal
Data1 <- mutate (Data1, RRV_Cat = case_when (
  RecentRevenueVariation <2.1  ~ "LowRRV",
```

```
  RecentRevenueVariation >= 2.1 ~ "HighRRV"
))

# Use Chi2 test to look for dependencies
RRV_Table <- xtabs(~RRV_Cat+AgeBin, Data1)
chisq.test(RRV_Table)
RRV1 <- prop.table(RRV_Table)
write.csv(RRV1, "RRV1.csv")
write.csv(RRV_Table, "RRVTable.csv")
# reject the null,  RRV_Cat and AgeBin are not independent

# Is it driven by higher earners?
RRV_Table2 <- xtabs(~RRV_Cat+HHIncomeBin, Data1)
chisq.test(RRV_Table2)
RRV2 <- prop.table(RRV_Table2)
write.csv(RRV2, "RRV2.csv")
write.csv(RRVTable2, "RRVTable2.csv")

# Two Sample t-test for RRV_Cat vs Age
plot(Data1$Age~as.factor(Data1$RRV_Cat))
result2 <- t.test(Age~RRV_Cat, data = Data1)
result2
# Recent rise in revenue likely attributed to older folks?

# Is High RRV related to the type of service?
# High Voice Utilization
# Chi Squared Test for Independence
RRV_Table3 <- xtabs(~RRV_Cat+VoiceUtilCat, Data1)
chisq.test(RRV_Table3)
RRV3 <- prop.table(RRV_Table3)
write.csv(RRV3, "RRV2.csv")
write.csv(RRV_Table3, "RRVTable3.csv")

# Export final dataset
write.xlsx(Data1, "EngineeredData.xlsx")

#END
```