Andrew Porter

Data Science Summer ‘19

Homework 1

**Problem 1**

Below is a set of findings for a dataset called “SERU” this dataset contains various features on a set of students, some of which filled out a student experience survey and some did not. The main objective is to understand if there is any relationship between the features and the students who filled out the survey. The exercise performed is purely a data exploration activity followed by a descriptive analysis of all features. Before we jump into the data exploration a quick section of data transformation will be provided. Seed = 142

**Data clean up**

Feature transformation

**Responded survey** - either yes or no

**Level grad** – read as numerical should be ordinal

**Invited previous** – not in data dictionary, but assuming this is a feature that determines if the person was invited to a pervious survey. This should be a categorical variable (0 or 1). There are 11,999 missing values and only one unique value (which is 1). So one could recode the blanks as 0’s, for this analysis I will run an analysis with and without this feature and explicitly state each case.

**Level\_Grad** - from numerical to ordinal

**Nonresident**- like Invited previous where blanks to 0s and Y to 1, change to categorical

**STATUS** – from numerical to categorical

**Term1** – from numerical to categorical

**Year** – from numerical to date (in this case categorical). Will help to understand if there is a relationship between time and competition of survey.

**HS\_percentile\_rank** – has over 10,000 missing values one could infer from other high school performance metrics (ex. ACT or SAT) and estimate rank. The issue we run into is that there are also missing values for ACT and SAT. One would have to fine a particular value in HS\_percentile\_rank and determine their ACT or SAT score (if it exists) and take that persons associated value. For this analysis there is an additional thought in the next section regarding these three features.

Feature removal

**Level –** same as level\_grad

**Eval\_major** -doesn’t seem relevant based on the description

College code matches College name, will remove college code

**Major\_text1 and CIP\_code1** matches major\_code1 - remove Major\_text1 and CIP\_code1

**Major\_text2 and CIP\_code2** matches major\_code2 - remove Major\_text2 and CIP\_code1

**Remove AFRICANAMERICAN, AMERIANDIAN, ASIAN, DECLINETOSTATE, Pacific Islander, White and HISPANIC** and let Ethnicity\_loc take care of these

**EXCLUDE** based on the data dictionary’s definition it is a categorical variable that determines if the student should be excluded from the common data file if the student is under 18 or already graduated. Seems like this should not be considered towards the analysis.

**HS\_percentile\_rank –** remove, this feature has almost 50% of the entire student’s missing a value. Since this is a measure of student performance and to some degree SAT and ACT is a measure of student performance. Will remove HS\_percentile\_rank

**SAT –** Since SAT is in the minority as compared to ACT, we will remove this column (along with all other SAT scores) and use ACT as our measurement of student performance in high school.

**Mod assign2015** – should be categorical (after analysis, could remove). There is a comment about if this field is blank the project manager will randomly assign a value

**ISE\_SA –** is a feature with only one unique value, will remove

**SP16\_term\_UI\_total hours with SP16\_term\_UI\_graded hours**-based on correlation activity determined that SP16\_ENRL\_HRS\_AT\_CENSUS will remain and remove these two

Feature imputation

**ACT –** Has 4,515 missing values, if one removes HS\_percentile\_rank and SAT one could use the median or average score for all other students within the same major. i.e. for all missing ACT scores for students who are majoring in Accounting is 28.5 we could add this value to the missing ACT scores for the Accounting majors. This was not performed, for the exploratory analysis we will simply use ACT scores.

**Summary of variables:**

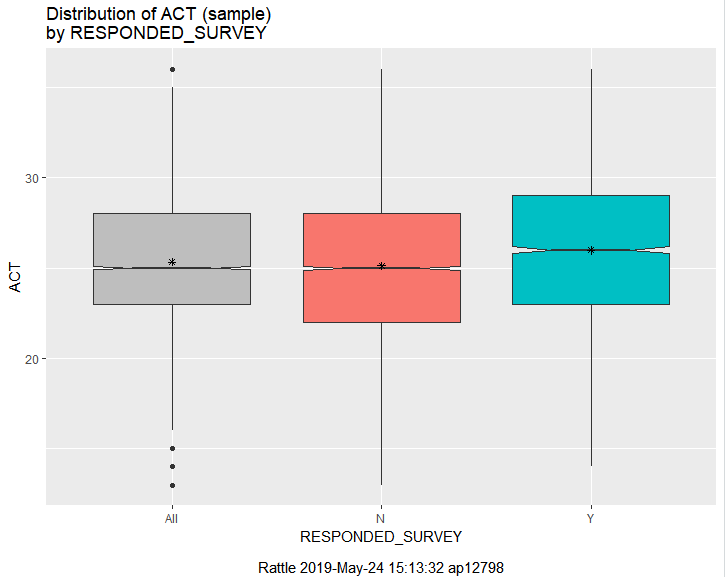
Of the training set (70% of overall population) there are 2,104 students who responded to the survey and 7,958 who did not respond. All numerical values have means close to their medians which would imply that there are no significant outliers. The average student’s age is 21 years old, an overwhelming majority of students are enrolled in the College of Liberal Arts and Sciences, followed by the College of Business. Roughly 4,000 students within this sample claim to have a second major. The average student’s ACT score is 25, while there are 700 more females compared to males.

**Correlation:**

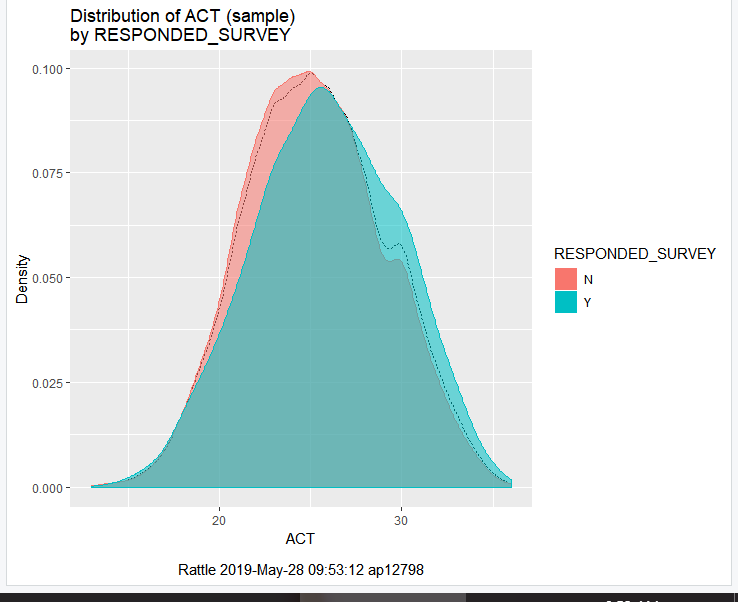
There are a few correlations with the remaining data, which could mean we could remove variables that are highly correlated (multicollinearity) to simplify our analysis. For instance, semester hours and total units, SP16\_term\_UI\_total hours with SP16\_term\_UI\_graded hours and SP16\_enrl\_hrs\_at\_cenus.

**Descriptive summary:**

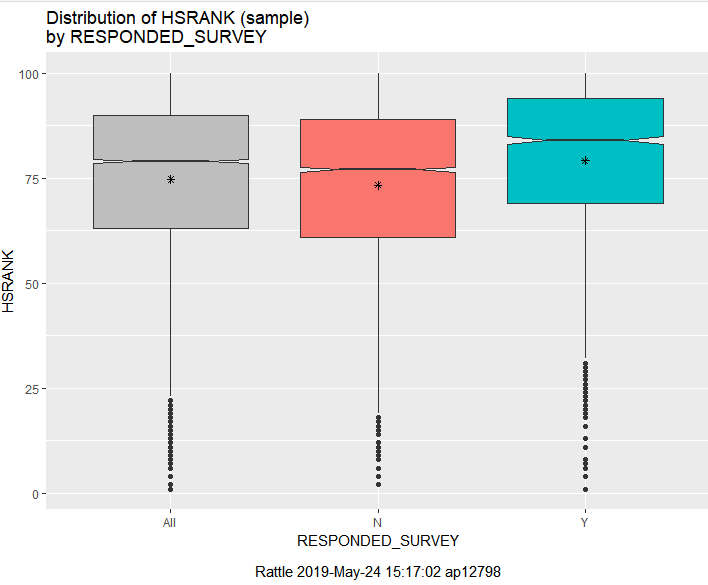
The below chart illustrates the median ACT score for each respondent. Focusing on the red and teal colored boxes, one can determine for those who responded to the survey scored slightly higher on the ACT.



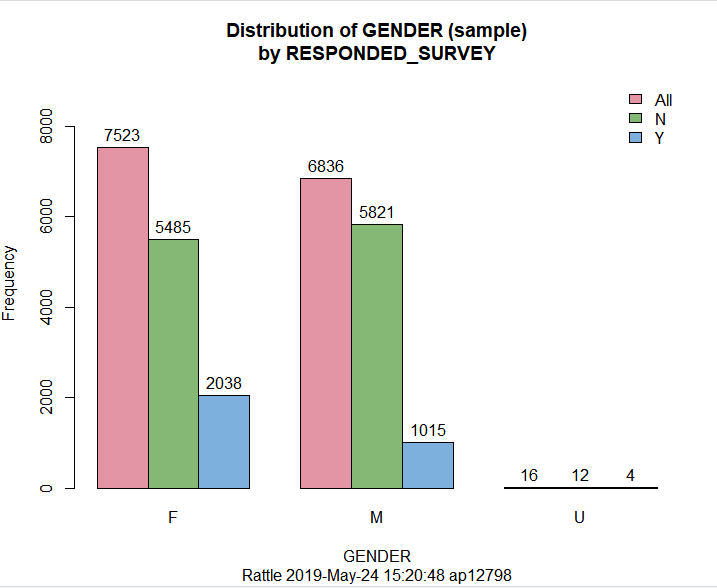
Below is another way to look at the same data. The mean for the ACT (peak of hill) on the population who responded to the survey is slightly shifted to the right.



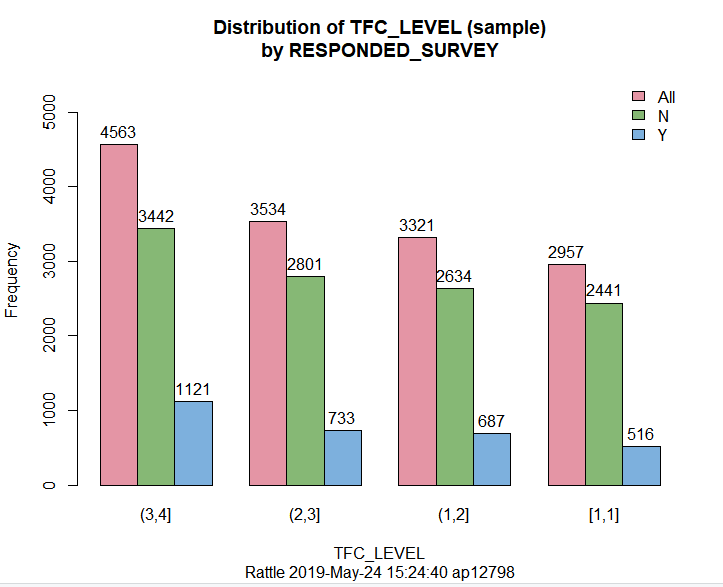
The same can be said for Highschool rank (high school rank and ACT have a correlation of .42)



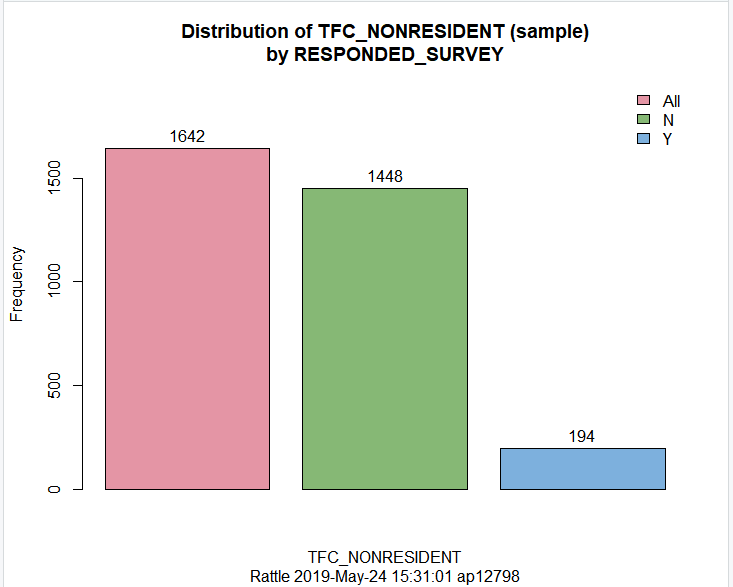
Shifting to gender related observation. 27% of females and 15% of males respond to the survey.



Below is categorizing the students by class (1 = Freshman, 4 = Senior). Where 516 (17%) freshmen responded, 687 sophomores (21%), 733 juniors (21% responded), and 1121 (24% responded) seniors. From freshman to senior there is a slight percent increase in respondents.



Only 12% of the non-resident students (international student) responded to the survey.



**A brief summary:**

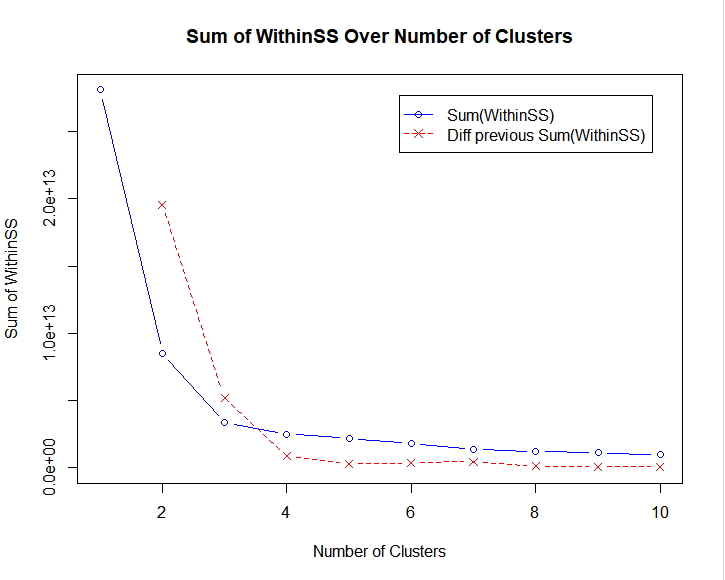
A couple key aspects of these students would be if they filled out the survey from last year and are they involved in any extracurricular activities usually those who are involved in more than classes there is a greater likelihood they will take part in a survey. The number of females is greater, if one is a performer in high school, and of the person is an upperclassman they are more likely to fill out the survey. If the goal was to get a greater likelihood of a student filling out surveys one could target these groups. But if the goal is to get a board representation of the student population for gauging improvement ideas or changes (based on other questions within the survey) one would want to keep the sample size as board as possible.

**Problem 2**

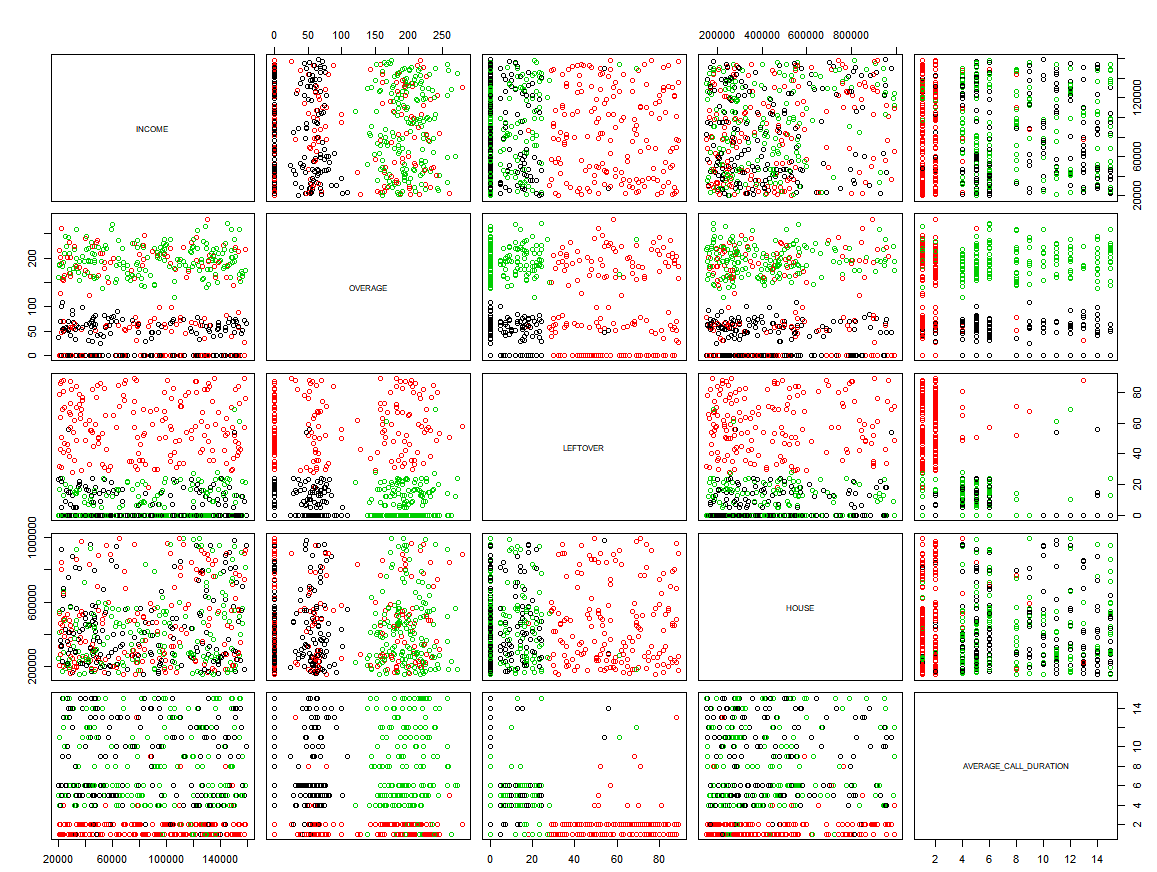
The second data set is a set of features that will be used to analysis customer churn for a cell phone company. We are interested only in the data that is causing customers to leave, so we will filter on the feature “result” and cluster features to gain insight on what is causing customers to leave the cell phone company. The first set of analysis is performed on only the numerical features.

A quick descriptive summary of the data. Overage, leftover, over 15 mins calls per month have means far from the median which would imply that there are outliers. After executing a correlation analysis, it is noticed that over 15 mins calls per month and overage are highly correlated (.76) and handset price and income are highly correlated (.79). Based on the limited nature of rattle, we will remove both handset price and 15 mins call per month from the cluster analysis.

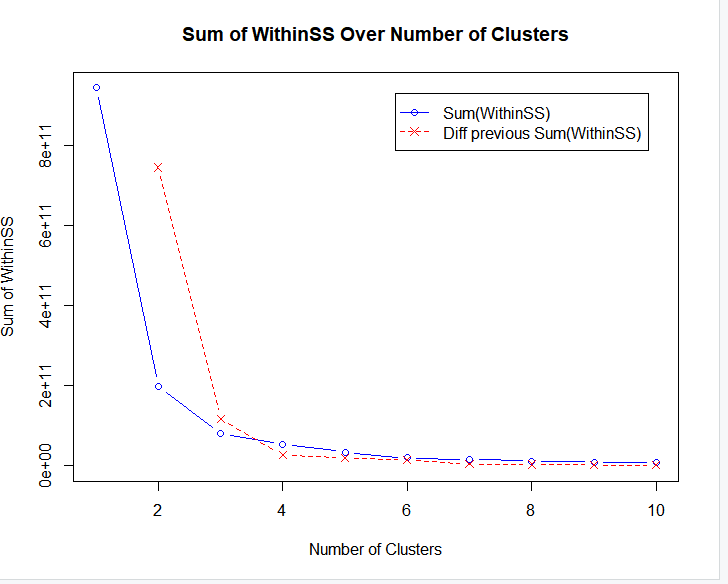
Based on the chart below and applying the Elbow Method we can determine that three clusters will be enough for the descriptive analysis, as the forth cluster doesn’t decrease the Sum of WSS significantly.



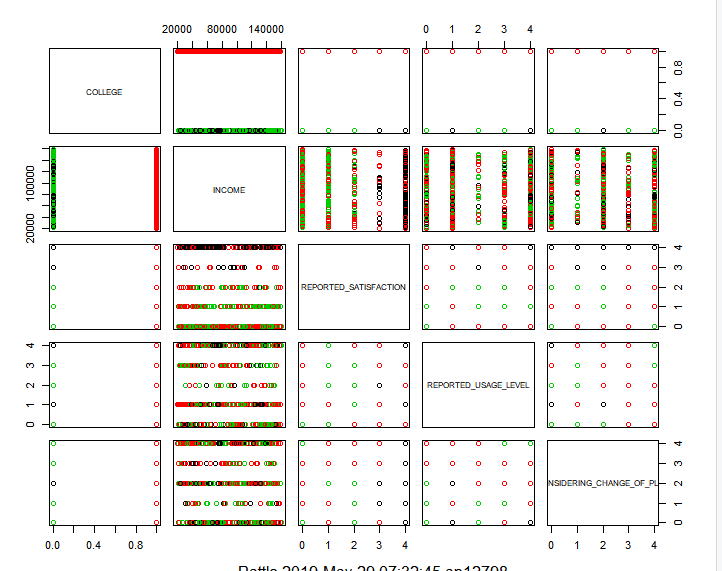
Based on the cluster information below there are a couple insights. There doesn’t seem to be any significant findings related to house price, the three groups are mixed throughout the features and house price increases there doesn’t seem to be any additional separation. One could consider this as a feature that would not drive decisions. There is a separation for leftover and overage, which can be considering indirectly correlated (.-67 correlation). One would anticipate that as income raises house price should raise too (.18 correlation) but that trend is not found which also reflected in the cluster charts. There are three relatively distinct groups when comparing overages and average call duration. One that stays low on average call duration but spans from a small number of overages to many overages (red) . A second group of low overages but spans the entire ranges of average call duration (black). The third group has a higher average call duration and higher number of overages (green). Another interesting observation is that the green color group within overage vs income never drops below 100, while that group spans the entire income range.



A second set of analysis was performed by recoding the categorical variables into numerical variables. Starting at zero for the most negative experience to the larger the number the more positive experience (i.e. Reported\_satisfaction of “unsat” receives a zero, while reported\_satisfcation of “very\_sat” receives a 4). Based on the chart below, the number of clusters determined to use for analysis is three.



Below is a faceted chart with college, income, reported\_satisfcation, reported\_usage\_level, and considering\_change\_of\_plan as features.



The above chart shows some interesting descriptive insights. Focusing on college first, there is a separation (of red) of the groups with respect to reported satisfaction, all others seem to be mixed. Also for all other features the red group is college educated, while the green and black groups overlap or mix. A second observation is with reported\_satisfcation and considering\_change\_of\_plans. There are group of customers who reported high level of satisfaction and spread the entire range for considering\_change\_of\_plans. There is also a group that is on the low level of reported\_satisfcaton and low level of consider\_change\_of\_plans, the former is counter intuitive while the latter is expected.

**A brief summary:**

Based on the data provided, there were some values that contain outliers (overage, leftover, and 15mins call per month). These would have to be looked at individually to determine of those cases would need to be thrown out.

There were many variables that are correlated (house and handset price or overage and over 15mins calls per month). Of these four features two were chosen for the descriptive analysis. The categorical variables had to be recoded to numerical with keeping in mind that the higher the number the more positive the outlook will be. To improve this analysis, one may be interested in a few more additional features how many previous cell phone providers has the customer used, has the customer been charged a late fee, how many times does the customer call customer support, how many times does the customer log in to the provider’s website.

Appendix: R code

Churn data set

df <- read.csv("churn\_imbalanced.csv", na.strings = "")

#subsetting for only the customers the are categorized as leave

df <- subset(df, df$result == "LEAVE")

#taking categorical variable and changing to numerical

df$REPORTED\_SATISFACTION <- as.character(df$REPORTED\_SATISFACTION)

df$REPORTED\_SATISFACTION[df$REPORTED\_SATISFACTION == "very\_unsat"] <- "0"

df$REPORTED\_SATISFACTION[df$REPORTED\_SATISFACTION == "unsat"] <- "1"

df$REPORTED\_SATISFACTION[df$REPORTED\_SATISFACTION == "avg"] <- "2"

df$REPORTED\_SATISFACTION[df$REPORTED\_SATISFACTION == "sat"] <- "3"

df$REPORTED\_SATISFACTION[df$REPORTED\_SATISFACTION == "very\_sat"] <- "4"

df$REPORTED\_SATISFACTION <- as.factor(df$REPORTED\_SATISFACTION)

df$REPORTED\_USAGE\_LEVEL <- as.character(df$REPORTED\_USAGE\_LEVEL)

#taking categorical variable and changing to numerical

df$REPORTED\_USAGE\_LEVEL[df$REPORTED\_USAGE\_LEVEL == "very\_little"] <- "0"

df$REPORTED\_USAGE\_LEVEL[df$REPORTED\_USAGE\_LEVEL == "little"] <- "1"

df$REPORTED\_USAGE\_LEVEL[df$REPORTED\_USAGE\_LEVEL == "avg"] <- "2"

df$REPORTED\_USAGE\_LEVEL[df$REPORTED\_USAGE\_LEVEL == "high"] <- "3"

df$REPORTED\_USAGE\_LEVEL[df$REPORTED\_USAGE\_LEVEL == "very\_high"] <- "4"

df$REPORTED\_USAGE\_LEVEL <- as.factor(df$REPORTED\_USAGE\_LEVEL)

#taking categorical variable and changing to numerical

df$CONSIDERING\_CHANGE\_OF\_PLAN <- as.character(df$CONSIDERING\_CHANGE\_OF\_PLAN)

df$CONSIDERING\_CHANGE\_OF\_PLAN[df$CONSIDERING\_CHANGE\_OF\_PLAN == "no"] <- "4"

df$CONSIDERING\_CHANGE\_OF\_PLAN[df$CONSIDERING\_CHANGE\_OF\_PLAN == "never\_thought"] <- "3"

df$CONSIDERING\_CHANGE\_OF\_PLAN[df$CONSIDERING\_CHANGE\_OF\_PLAN == "considering"] <- "2"

df$CONSIDERING\_CHANGE\_OF\_PLAN[df$CONSIDERING\_CHANGE\_OF\_PLAN == "perhaps"] <- "1"

df$CONSIDERING\_CHANGE\_OF\_PLAN[df$CONSIDERING\_CHANGE\_OF\_PLAN == "actively\_looking\_into\_it"] <- "0"

df$CONSIDERING\_CHANGE\_OF\_PLAN <- as.factor(df$CONSIDERING\_CHANGE\_OF\_PLAN)

df$COLLEGE <- as.character(df$COLLEGE)

#taking categorical variable and changing to numerical

df$COLLEGE[df$COLLEGE == "zero"] <- "0"

df$COLLEGE[df$COLLEGE == "one"] <- "1"

df$COLLEGE <- as.factor(df$COLLEGE)

write.csv(df, file = "churn\_imbalanced\_cleaned.csv")