D207 Exploratory Data Analysis

Adam Eyerman

2022-07-10

**A1)** **Provide one question that is relevant to your chosen data set. You will answer this question later in the task through an analysis of the cleaned data, using one of the following techniques: chi-square, t-test, or analysis of variance (ANOVA).**

* Does the type of internet service or contract length have an impact on churn?

**A2)** **Explain how stakeholders in the organization could benefit from an analysis of the data.**

* Stakeholders would benefit from this analysis because if a relationship is discovered the company can focus on which internet service and which contract length is most suited for client retention. The company can also do further research into why certain service types or contract lengths are leading to client losses more than others. Overall, answering this question can lead to a targeted campaign to provide the ideal service and contract for maximizing client retention and launch an effort into understanding why the other options provided contribute more to the churn rate. However, if there is no relationship found the company can also be reassured that all service types and contract lengths are not effecting churn.

**A3)** **Identify all the data in your data set that are relevant to answering your question in part A1.**

The variables that are relevant to answering this question include:

* Churn
* Contract
* Internet Service

This selection of data provides categorical variables. These variables will be analyzed to identify any relation to the churn variable.

**B1)** **Using one of the following techniques, write code (in either Python or R) to run the analysis of the data set:**

* **chi-square**
* t-test
* ANOV

[R Code in bold]

*# Create a table with the Churn and Contract variables*

**churn\_contract = table(churn\_clean$Churn, churn\_clean$Contract)**

**print(churn\_contract)**

*# Perform the Chi-Square test.*

**print(chisq.test(churn\_contract))**

*# Create a table with the Churn and InternetService variables*

**churn\_internet = table(churn\_clean$Churn, churn\_clean$InternetService)**

**print(churn\_internet)**

*# Perform the Chi-Square test.*

**print(chisq.test(churn\_internet))**

**B2)** **Provide the output and the results of any calculations from the analysis you performed.**

The clean data set did not need any modification for this analysis to be conducted. Churn, Internet Service, and Contract variables did not have any null values and had normal distributions. I installed “tidyverse” for my analysis and used the libraries included for running the chi-square test as well as my visualizations for univariate and bivariate stats in sections C and D.

#Install needed packages and library for analysis *install.packages(“tidyverse”)*

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6 v purrr 0.3.4  
## v tibble 3.1.7 v dplyr 1.0.9  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
  
#Load clean data set into R  
churn\_clean <- read.csv('churn\_clean.csv')  
  
# Create a table with the Churn and Contract variables  
churn\_contract = table(churn\_clean$Churn, churn\_clean$Contract)  
print(churn\_contract)

##   
## Month-to-month One year Two Year  
## No 3422 1795 2133  
## Yes 2034 307 309

# Perform the Chi-Square test.  
print(chisq.test(churn\_contract))

##   
## Pearson's Chi-squared test  
##   
## data: churn\_contract  
## X-squared = 718.59, df = 2, p-value < 2.2e-16

# Create a table with the Churn and InternetService variables  
churn\_internet = table(churn\_clean$Churn, churn\_clean$InternetService)  
print(churn\_internet)

##   
## DSL Fiber Optic None  
## No 2349 3368 1633  
## Yes 1114 1040 496

# Perform the Chi-Square test.  
print(chisq.test(churn\_internet))

##   
## Pearson's Chi-squared test  
##   
## data: churn\_internet  
## X-squared = 87.462, df = 2, p-value < 2.2e-16

The results display three outputs: X-squared, df, and p-value.

Turney(2022) shows the Chi-square test statistic formula:

\begin{equation*} X^2=\sum{\frac{(O-E)^2}{E}} \end{equation*}

This formula is taking the observed frequency minus the expected frequency squared divided by the expected frequency for each category then summing the total to report the X-squared output seen above.

The df output is reporting the degrees of freedom. According to (“How Do You Calculate Degrees of Freedom for Chi Square Tests?”, 2020) to calculate the degrees of freedom for a chi-square test, first create a contingency table and then determine the number of rows and columns that are in the chi-square test. Take the number of rows minus one and multiply that number by the number of columns minus one. The resulting figure is the degrees of freedom for the chi-square test.

The p-value output is reporting the likelihood of the results. With a p-value less than 0.05 the result would be considered significant meaning the probability of the result is less than 5% likely. Both tests produced reported a p-value below this level and therefore would be considered significant. This suggests that the variables have a relationship, and that Internet Service and Contract variables appear to have an impact on the Churn variable.

**B3)** **Justify why you chose this analysis technique.**

The chi-square test was used to analyze the relationship between two categorical variables. Churn vs. Contract and Churn vs. InternetService were both analyzed. I chose this technique because it was an effective method to relate the dependent variable (Churn) to other key categorical variables (Contract/InternetService).

Turney (2022) states “A Pearson’s chi-square test is a statistical test for categorical data. It is used to determine whether your data are significantly different from what you expected.”

**C1)** **Identify the distribution of two continuous variables and two categorical variables using univariate statistics from your cleaned and prepared data.**

The categorical variables are Churn and Internet Service and they are visually represented using bar plots to display their distribution using univariate statistics.

# Display table of Churn categorical variable  
table(churn\_clean$Churn)

##   
## No Yes   
## 7350 2650

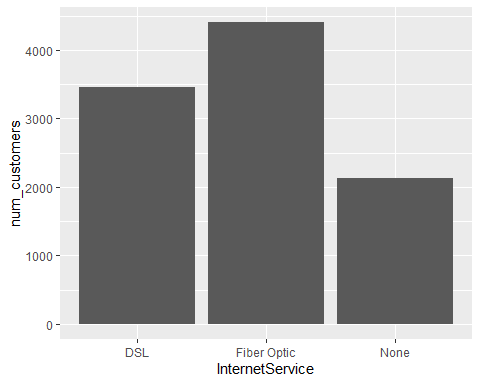
# Display histogram of Churn categorical variable  
  
num\_customers = 1  
  
plot<-ggplot(churn\_clean,  
 aes(Churn,num\_customers)) +  
 geom\_bar(stat = "identity")  
  
plot



# Display table of InternetService categorical variable  
table(churn\_clean$InternetService)

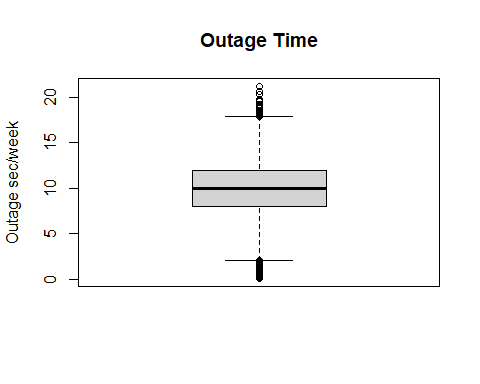
##   
## DSL Fiber Optic None   
## 3463 4408 2129

# Display histogram of InternetService categorical variable  
  
num\_customers = 1  
  
plot\_2<-ggplot(churn\_clean,  
 aes(InternetService,num\_customers)) +  
 geom\_bar(stat = "identity")  
  
plot\_2

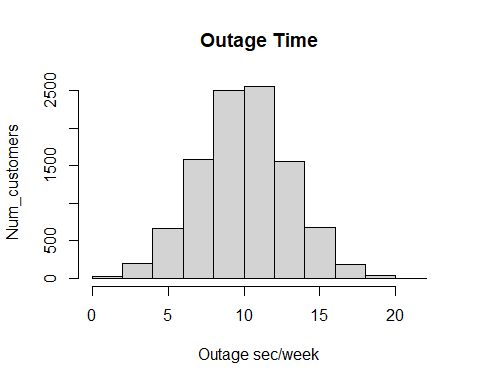


The continuous variables are Outage\_sec\_perweek and Bandwidth\_GB\_year and they are visually represented using box plots and histograms to display their distribution using univariate statistics.

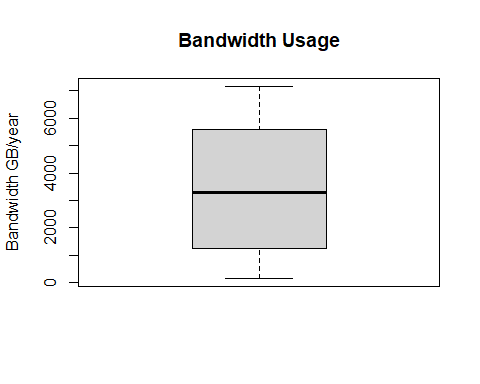
# Display boxplot of Outage\_sec\_perweek continuous variable  
boxplot(churn\_clean$Outage\_sec\_perweek, ylab = "Outage sec/week" , main = "Outage Time")



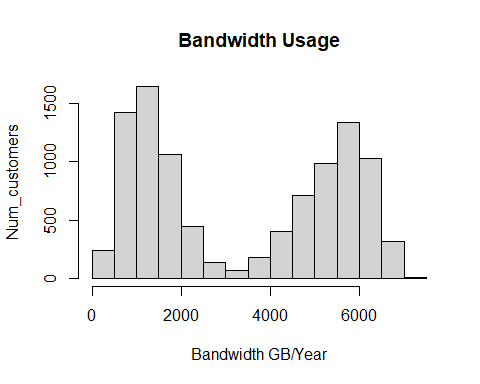
# Display histogram of Outage\_sec\_perweek continuous variable  
hist(churn\_clean$Outage\_sec\_perweek, xlab = "Outage sec/week", ylab = "Num\_customers", main = "Outage Time")



# Display boxplot of Bandwidth\_GB\_Year continuous variable  
boxplot(churn\_clean$Bandwidth\_GB\_Year, ylab = "Bandwidth GB/year", main = "Bandwidth Usage")



# Display histogram of Bandwidth\_GB\_Year continuous variable  
hist(churn\_clean$Bandwidth\_GB\_Year, xlab = "Bandwidth GB/Year", ylab = "Num\_customers", main = "Bandwidth Usage")



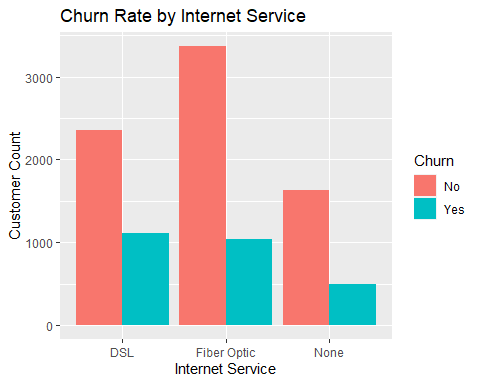
**D1)** **Identify the distribution of two continuous variables and two categorical variables using bivariate statistics from your cleaned and prepared data.**

The categorical variables are Churn and Internet Service and are visually represented by a side-by-side bar plot to display their distribution alongside a table.

# Display table of Churn/Internet Service categorical variables  
table(churn\_clean$Churn, churn\_clean$InternetService)

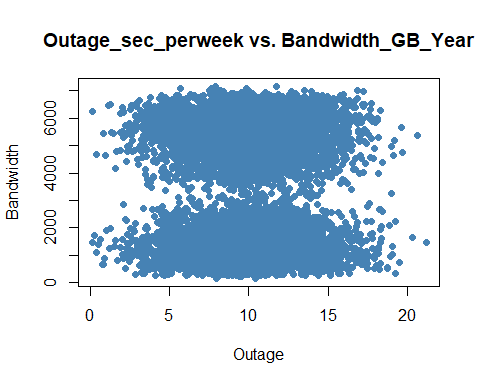
##   
## DSL Fiber Optic None  
## No 2349 3368 1633  
## Yes 1114 1040 496

# Display barplot of Churn rate by Internet Service  
plot\_3 <- ggplot(churn\_clean, aes(InternetService, fill = Churn)) + geom\_bar(position = "dodge", stat = "count")  
  
plot\_3 +labs(title="Churn Rate by Internet Service",  
 x ="Internet Service", y = "Customer Count")



The continuous variables are Outage\_sec\_perweek and Bandwidth\_GB\_Year and are visually represented using a scatterplot to display their distribution alongside a correlation coefficient.

# Display scatterplot of Outage\_sec\_perweek/Bandwidth\_GB\_Year continuous variables   
plot(churn\_clean$Outage\_sec\_perweek, churn\_clean$Bandwidth\_GB\_Year, pch=16, col='steelblue',  
 main='Outage\_sec\_perweek vs. Bandwidth\_GB\_Year',  
 xlab='Outage', ylab='Bandwidth')



# Display correlation coefficient of Outage\_sec\_perweek/Bandwidth\_GB\_Year continuous variables  
cor(churn\_clean$Outage\_sec\_perweek, churn\_clean$Bandwidth\_GB\_Year)

## [1] 0.004175661

**E1) Discuss the results of the hypothesis test.**

The chi-square test showed a P-value of less than 0.05 and concluded that the results were unexpected and displayed an abnormal distribution. The first test was between the variables of Churn and Contract. After seeing the P-value below 0.05 the null hypothesis was rejected, and the data table constructed between the two variables was further investigated. The Contract variable has three groups: Month-to-Month, One Year and Two Year. The Two Year and One Year groups have a 12.7% and 14.6% churn rate, respectively. However, the Month-to-Month group has a 37.3% churn rate. This suggests the Month-to-Month group needs to be further investigated for contributing to Churn due to the fact the churn rate is significantly higher than the other two groups.

table(churn\_clean$Churn, churn\_clean$Contract)

##   
## Month-to-month One year Two Year  
## No 3422 1795 2133  
## Yes 2034 307 309

**Churn % 37.3% 14.6% 12.7%**

The second test was between the variables of Churn and Internet Service. The P-value was also less than 0.05 and the null hypothesis was rejected. Like the first test, viewing the table data displayed a similar churn rate discrepancy. In this case, the DSL option had a significantly higher churn rate than the other two groups. This suggests the DSL group needs to be further investigated for contributing to Churn.

table(churn\_clean$Churn, churn\_clean$InternetService)

##   
## DSL Fiber Optic None  
## No 2349 3368 1633  
## Yes 1114 1040 496

**Churn% 32.2% 23.6% 23.2%**

Overall, the hypothesis testing seems to have offered insight into these variables being valid contributors to explain churn rate. It appears that a Month-to Month contract or the use of DSL service as an internet source results in a significantly higher likelihood of churn.

**E2) Discuss the limitations of your data analysis.**

A constant update of the dataset is needed to fully understand the contributing factors to the churn rate. Currently working with a stagnant dataset limits the ability to see how the data changes with new customer input and if the results are reproducible and consistent over time. However, the current dataset does provide adequate information to assess the research question of which variables may contribute most to the churn rate within the context of the data provided.

**E3) Recommend a course of action based on your results.**

My recommendation would be to further investigate the Internet Service and Contract variables. Logically it would make sense that a Month-to-Month contract as opposed to a longer-term contract could lead to a higher churn rate. Targeting those customers with more attention, better deals, or focusing on conversion to longer-term contracts may lead to lowering the likelihood of churn. In terms of the Internet Service variable, the group using DSL appears to be more likely to terminate their service than those without internet or using Fiber Optic. Fiber Optic is a superior service to DSL and much more dependable. This factor leads to a logical conclusion that customers with DSL may have less satisfaction than those with Fiber Optic or even customers without internet service at all. My recommendation would be to review the DSL customers survey responses to see if the satisfaction is lower than the other two groups and if so, I would focus on improving customer service and reliability.

**F)** **Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the tool(s) used.**

[Mon Jul 11 2022 2:02:38 PM (panopto.com)](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=c72d463a-9181-4a84-b7c9-aecf012971bc)

**G)** **Reference the web sources used to acquire segments of third-party code to support the analysis.**

Turney, S. (2022, June 21). Chi-Square (Χ2) Tests | Types, Formula & Examples. Scribbr. <https://www.scribbr.com/statistics/chi-square-tests/>

S. (2020, April 2). *How Do You Calculate Degrees of Freedom for Chi Square Tests?* Reference.Com. <https://www.reference.com/world-view/calculate-degrees-freedom-chi-square-tests-c5aa2dc77c48a614>