D207 Exploratory Data Analysis

**Adam Eyerman**

**2022-07-06**

**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).**

* What are the variables that contribute most to the churn rate?

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

* Stakeholders would benefit from identifying the aspects that contribute to churn because it will give the company an ability to make changes accordingly and potentially lower the loss of current clientele. This is a major factor in revenue and business growth projections. Lowering churn rate and increasing retention is a top priority for any successful business.

**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:

* Outage\_sec\_perweek
* Churn
* Contract
* Internet Service
* Yearly\_equip\_failure
* Tenure
* MonthlyCharge
* Timely\_response
* Timely\_fixes
* Timely\_replacements
* Reliability
* Options
* Respectful\_response
* Courteous\_exchange
* Active\_listening

This selection of data provides numeric and 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

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

#Load cleaned data set into R  
churn\_data\_cleaned <- read.csv('churn\_data\_cleaned.csv')  
  
# Create a table with the Churn and Contract variables  
churn\_data = table(churn\_data\_cleaned$Churn, churn\_data\_cleaned$Contract)  
print(churn\_data)

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

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

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

# Create a table with the Churn and InternetService variables  
churn\_data\_2 = table(churn\_data\_cleaned$Churn, churn\_data\_cleaned$InternetService)  
print(churn\_data\_2)

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

# Perform the Chi-Square test.  
print(chisq.test(churn\_data\_2))

##   
## Pearson's Chi-squared test  
##   
## data: churn\_data\_2  
## X-squared = 87.462, df = 2, p-value < 2.2e-16

**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.**

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

##   
## No Yes   
## 7350 2650

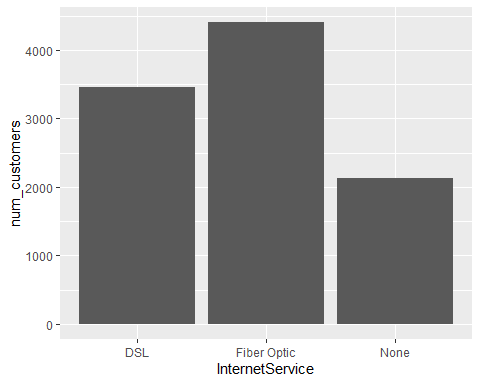
# Load ggplot2 library  
library(ggplot2)  
  
# Display histogram of Churn categorical variable  
  
num\_customers = 1  
  
plot<-ggplot(churn\_data\_cleaned,  
 aes(Churn,num\_customers)) +  
 geom\_bar(stat = "identity")  
  
plot



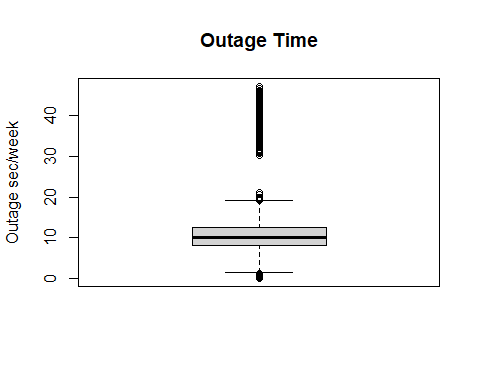
# Display table of InternetService categorical variable  
table(churn\_data\_cleaned$InternetService)

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

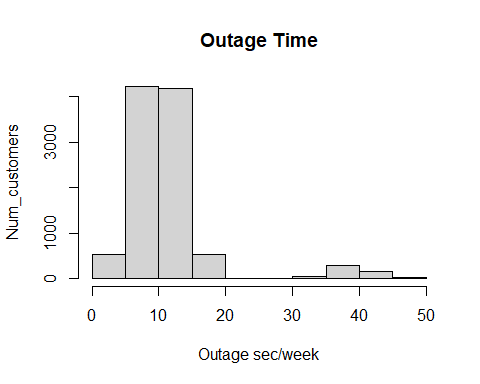
# Display histogram of InternetService categorical variable  
  
num\_customers = 1  
  
plot\_2<-ggplot(churn\_data\_cleaned,  
 aes(InternetService,num\_customers)) +  
 geom\_bar(stat = "identity")  
  
plot\_2



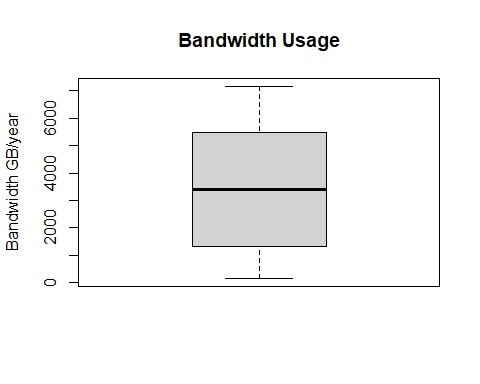
# Display boxplot of Outage\_sec\_perweek continuous variable  
boxplot(churn\_data\_cleaned$Outage\_sec\_perweek, ylab = "Outage sec/week" , main = "Outage Time")



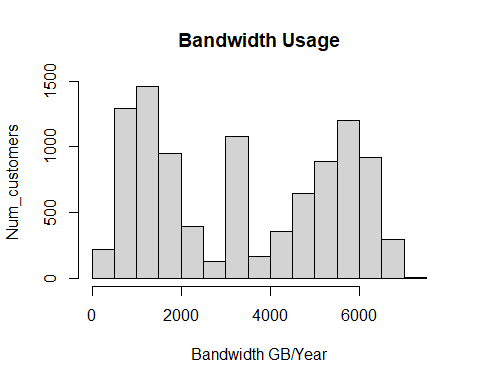
# Display histogram of Outage\_sec\_perweek continuous variable  
hist(churn\_data\_cleaned$Outage\_sec\_perweek, xlab = "Outage sec/week", ylab = "Num\_customers", main = "Outage Time")



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



# Display histogram of Bandwidth\_GB\_Year continuous variable  
hist(churn\_data\_cleaned$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.**

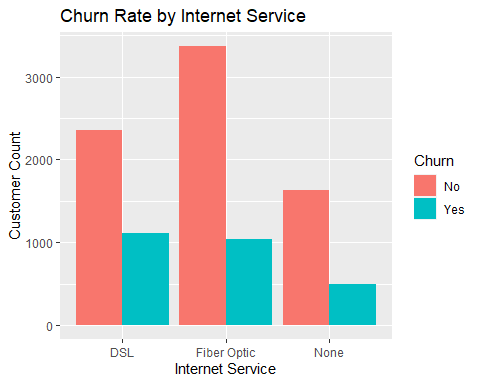
# Display table of Churn/Internet Service categorical variables  
table(churn\_data\_cleaned$Churn, churn\_data\_cleaned$InternetService)

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

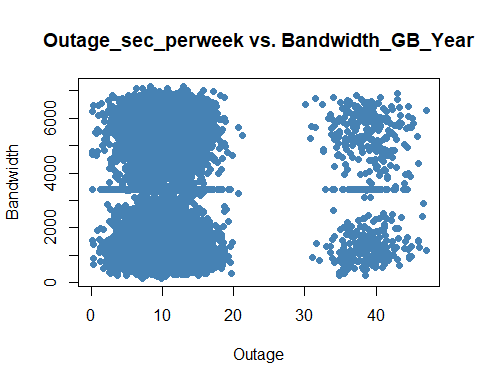
# Create data set for CramerV Correlation  
Churn\_InternetService = table(churn\_data\_cleaned$Churn, churn\_data\_cleaned$InternetService)  
  
# Load library rcompanion  
library(rcompanion)  
  
# Display CramerV Correlation   
cramerV(Churn\_InternetService)

## Cramer V   
## 0.09352

# Display barplot of Churn rate by Internet Service  
plot\_3 <- ggplot(churn\_data\_cleaned, 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")



# Display scatterplot of Outage\_sec\_perweek/Bandwidth\_GB\_Year continuous variables   
plot(churn\_data\_cleaned$Outage\_sec\_perweek, churn\_data\_cleaned$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\_data\_cleaned$Outage\_sec\_perweek, churn\_data\_cleaned$Bandwidth\_GB\_Year)

## [1] 0.01421146

**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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DSL** | **Fiber Optic** | **None** |
| **No** | 2349 | 3368 | 1633 |
| **Yes** | 1114 | 1040 | 496 |
| **Churn %** | 32.2% | 23.6% | 23.3% |

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 my 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.**

[Wed Jul 06 2022 11:34:47 AM (panopto.com)](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=38dd0326-4a3f-4778-a1a1-aeca0105d504)

**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/>