HOLT D207 OA

Bradley Holt

5/31/2021

A1. Question

Are customers with low internet usage at a greater risk of churn than customers with high internet usage in comparison with their monthly charges?

A2. Stakeholder Benefits

Business entities and their stakeholders are consistently looking for new ideas on how to transform, improve, or modify their services and/or products to target customers and gain an advantage in today's highly competitive markets. Exploratory data analysis (EDA) can help gain market insight by using data collected during operations to formulate questions and suggest hypotheses about the business and its customer base. Through further statistical analysis and modeling, those questions and hypotheses derived from EDA can be answered. Ultimately, EDA can help provide new ideas and directions for stakeholders to improve their business and its foothold in the market.

Exploratory Data Analysis employs a variety of mostly visual techniques to:

1) identify relationships within the data set
2) select and validate models
3) select estimations
4) detect outliers and anomalies
5) test assumptions
6) determine optimal factor settings (Natrella, 2013)

EDA uses these techniques to the benefit of business stakeholders by ensuring the right questions are being asked and bias is not being included with the investigation of the assumptions. EDA also provides the context around the question or problem to ensure the potential value of the output can be maximized. Overall, EDA is the foundation of the analysis of a question or problem and can lead to insights that can be hugely informative about the business. (Mawer, 2017)

A3. Data Identification

The data being used in this analysis are all derived from the churn_clean.csv data set and its corresponding data dictionary:

MonthlyCharge (continuous variable): The average amount charged to the customer monthly. For new customers this data is filled with the value of similar demographic customers.

Bandwidth_GB_Year (continuous variable): The average amount of data used, in GB, in a year by the customer. If the customer has had service less than a year the value is approximated based on initial use or of average usage fir a typical customer in their demographic profile.

Churn (categorical variable): Whether the custom discontinued service with the last month. (Yes/No)

InternetService (categorical variable): Customer's internet service provider (DSL/Fiber Optic/None)

B. Describe the Data Analysis

```
#Libraries used in this PA
library(tidyverse)
```

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

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(readr)
```

B1 & B2. Analysis Code, Output, and Results of T-test

Conducting a 2 sample t-test on the mean of the calculated proportion of MonthlyCharge per GB used.

```
# 2 sample means T-test
churn_clean %>%
  filter(InternetService != "None") %>%
  mutate(PropMonthlyUsage = MonthlyCharge/(Bandwidth_GB_Year/12)) %>%
  {t.test(x = .$PropMonthlyUsage[.$Churn == 'No'], y = .$PropMonthlyUsage[.$Churn == 'Yes'])}
```

```
##
##  Welch Two Sample t-test
##
## data:  .$PropMonthlyUsage[.$Churn == "No"] and .$PropMonthlyUsage[.$Churn == "Yes"]
## t = -38.168, df = 3701.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9814214 -0.8855197
## sample estimates:
## mean of x mean of y
## 0.8895082 1.8229788
```

B3. Justification

The two-sample t-test was chosen in this scenario as we are comparing the means between two different groups with continuous variables. In this example, the two groups are customers who churned and customers who did not churn (Churn = Yes/No). The numeric variable, PropMonthlyUsage, was created with two separate variables in the data set, MonthlyCharge and Bandwidth_GB_Year. This variable was created to more accurately compare customers' monthly charges and their data usage. Additionally, customers without internet service were filtered out of the data.

C. Distibution of Two Continous Variables and Two Categorical Veriables Using Univariate Statistics

MonthlyCharge Continous Variable

```
summary(churn_clean$MonthlyCharge)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   79.98  139.98  167.48  172.62  200.73  290.16
```

```
sd(churn_clean$MonthlyCharge)
```

```
## [1] 42.94309
```

Bandwidth_GB_Year Continous Variable

```
summary(churn_clean$Bandwidth_GB_Year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.5  1236.5  3279.5  3392.3  5586.1  7159.0
```

```
sd(churn_clean$Bandwidth_GB_Year)
```

```
## [1] 2185.295
```

Churn Categorical Variable

```
summary(churn_clean$Churn)
```

```
##  Yes   No
## 2650 7350
```

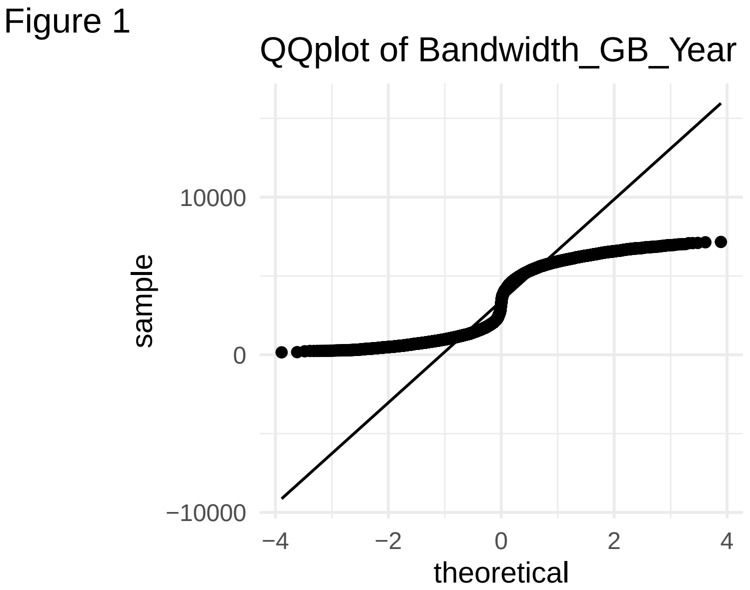
InternetService Categorical Variable

```
summary(churn_clean$InternetService)
```

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

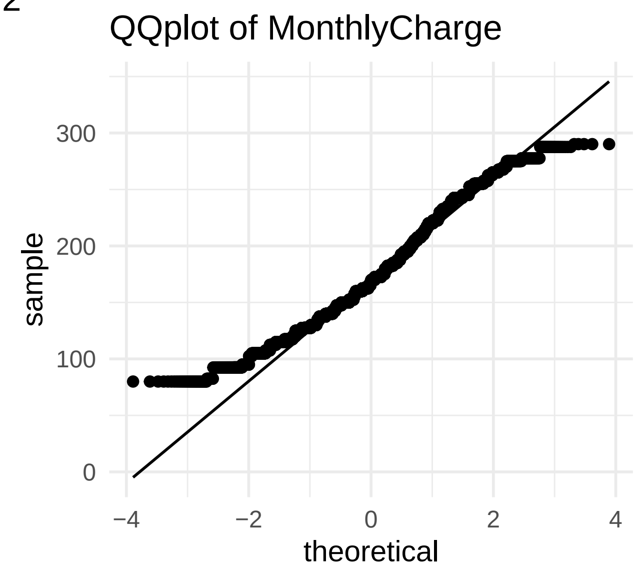
Visual Univariate Statistics

```
ggplot(churn_clean)+
  geom_qq(aes(sample = Bandwidth_GB_Year)) +
  geom_qq_line(aes(sample = Bandwidth_GB_Year)) +
  labs(title = "QQplot of Bandwidth_GB_Year",
       tag = "Figure 1")+
  theme_minimal()
```



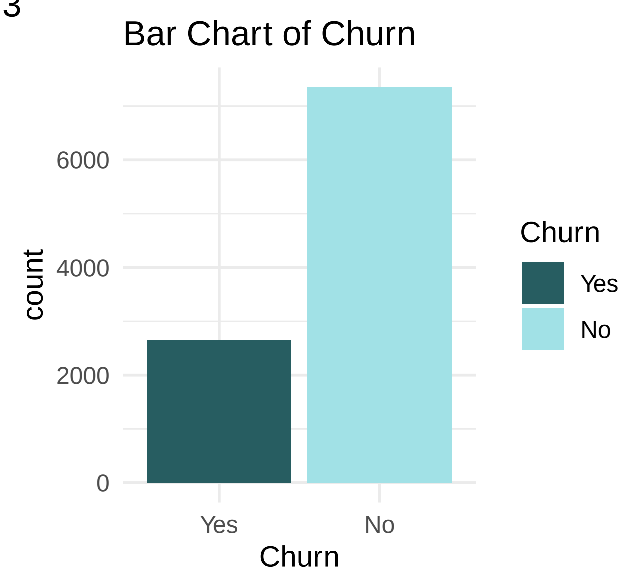
```
ggplot(churn_clean)+
  geom_qq(aes(sample = MonthlyCharge))+
  geom_qq_line(aes(sample = MonthlyCharge))+
  labs(title = "QQplot of MonthlyCharge",
       tag = "Figure 2")+
  theme_minimal()
```

Figure 2



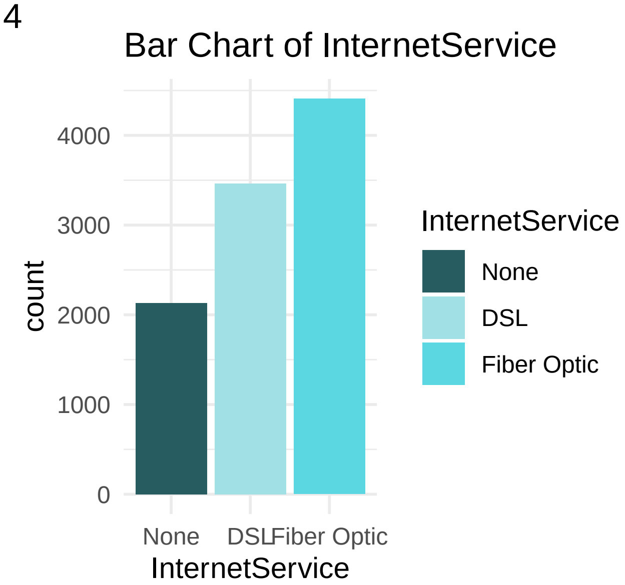
```
ggplot(churn_clean, aes(Churn, fill = Churn)) +
  geom_bar()+
  scale_fill_manual(values = c("#275D61", "#A1E1E6"))+
    theme_minimal()+
  labs(title = "Bar Chart of Churn",
       tag = "Figure 3",
       color = churn_clean$Churn)
```

Figure 3



```
ggplot(churn_clean, aes(InternetService, fill = InternetService)) +
  geom_bar()+
  scale_fill_manual(values = c("#275D61", "#A1E1E6", "#5AD7E0"))+
    theme_minimal()+
  labs(title = "Bar Chart of InternetService",
       tag = "Figure 4")
```

Figure 4



D. Distibution of Two Continous Variables and Two Categorical Variables Using Bivariate Statistics

MonthlyCharge and Bandwidth_GB_Year/12 Correlation

```
cor(churn_clean$MonthlyCharge, churn_clean$Bandwidth_GB_Year/12)
```

```
## [1] 0.06040643
```

Churn and InternetService Count Table

```
churn_clean %>%
  group_by(Churn, InternetService) %>%
  summarize(count = n()) %>%
  spread(InternetService, count)
```

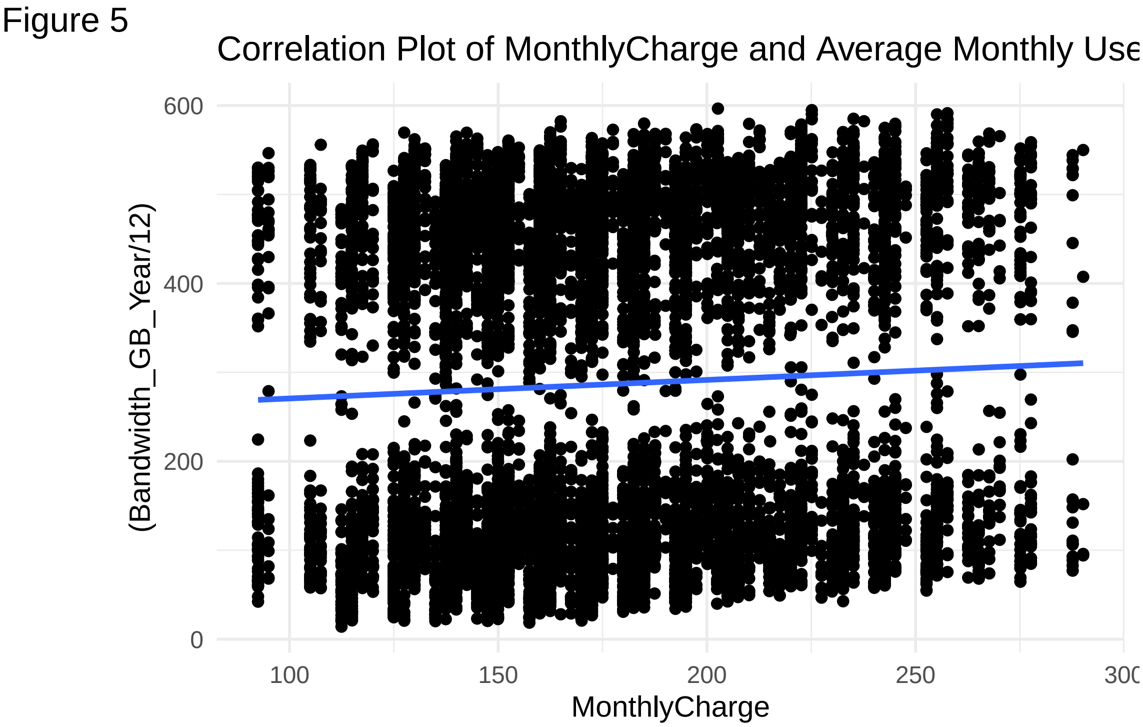
```
## `summarise()` has grouped output by 'Churn'. You can override using the `.groups` argument.
```

```
## # A tibble: 2 x 4
## # Groups:   Churn [2]
##   Churn  None   DSL `Fiber Optic`
##   <fct> <int> <int>         <int>
## 1 Yes     496  1114          1040
## 2 No     1633  2349          3368
```

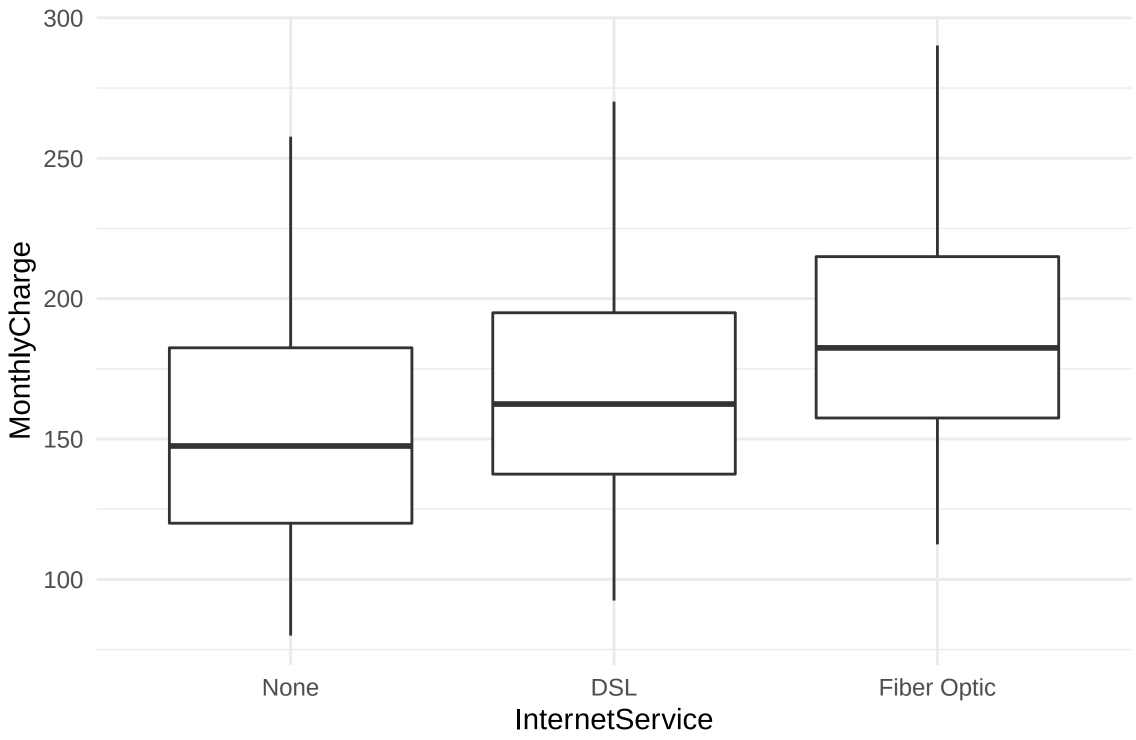
Visual Bivariate Statistics

```
ggplot(churn_clean %>% filter(InternetService != "None"), aes(x = MonthlyCharge, y = (Bandwidth_GB_Year/
  geom_point()+
  geom_smooth(method = lm, se = FALSE)+
  theme_minimal()+
  labs(title = "Correlation Plot of MonthlyCharge and Average Monthly Use (Bandwidth_GB_Year/12)",
       tag = "Figure 5")
```

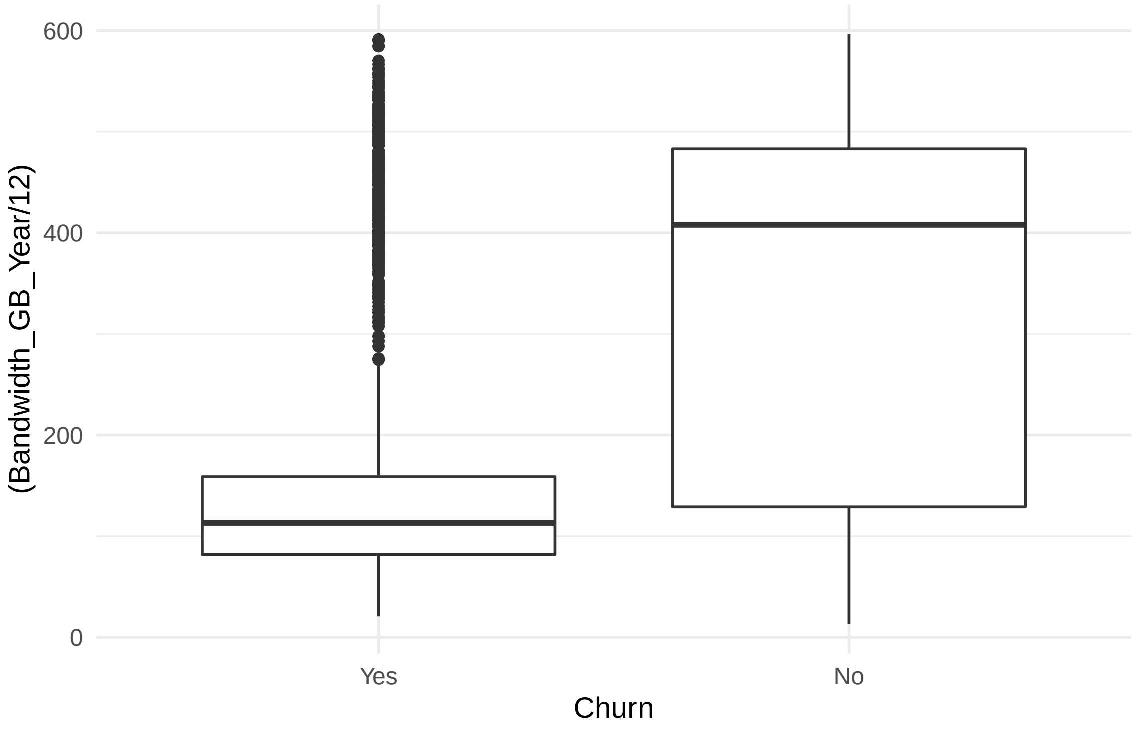
```
## `geom_smooth()` using formula 'y ~ x'
```



```
ggplot(churn_clean, aes(x = InternetService, y = MonthlyCharge)) +
  geom_boxplot()+
   theme_minimal()
```



```
ggplot(churn_clean, aes(x = Churn, y = (Bandwidth_GB_Year/12))) +
  geom_boxplot()+
   theme_minimal()
```



E. Summary

E2. Hypothesis Test

The hypothesis test is described as:

$H_0 : \mu_1 = \mu_2$ The mean of the proportion of monthly charges and internet usage between customers who churned and customers who did not churn are the same.

$H_a : \mu_1 \neq \mu_2$ The mean of the proportion of monthly charges and internet usage between customers who churned and customers who did not churn are not the same.

The assumptions before conducting the analysis are that the given data are random samples from total population of the business's customers and that they are independent. Additionally, it is assumed that the data is normally distributed since our sample size is above 30.

The t-test results gave a p-value of near 0, meaning that the probability of the sample result occurring by chance is near 0. Our mean for customers who did not churn was .89 and the mean for customers who did churn was 1.82. Overall, there is a 95% confidence level that the true difference of means between to two groups lies between -0.9814214 and -0.8855197.

From the initial analysis of the t-test results there is evidence that customers who churn use a significantly less amount of internet in proportion to their monthly charges then customers who did not churn. However, the evidence only provides correlating factors and further analysis and experimentation should be conducted to determine causality.

E3. Limitations

After conducting the t-test, by using qqplot techniques it was found that a variable used is not normally distributed by a significant factor. The qqplot implies that the Bandwidth_GB_Year variable is bi-modal. Indeed, if a histogram was produced, two separate groupings would be shown, with a cluster of customers with low data usage and cluster of customers with high data usage. This non-normality may implicate the results of the t-test. Since the data used is assumed to be a sample of a population we can use the central limit theorem to assume the population is normally distributed. However, a suggestion should be made to conduct further analysis on the issue and accept the t-test results with the issue in mind.

The analysis was also limited by the use of only univariate and bivariate analysis. By using multivariate analysis and/or other types of analysis further evidence could be provided to show differences between the groups. For example, when the multivariate analysis below is created you can clearly see the correlation differences between the two groups of customers who churned or didn't churn.

E4. Recommendations

After analyzing the data there is significant evidence that customers who churned had a significant difference in the portion of their monthly charge compared to their internet usage than customers who didn't churn. It seems that many of the churned customers had significantly higher charges but used a much lower amount of data. Additionally, fiber optic is the most provided internet service but with the highest average monthly charges. The internet usage of fiber optic users seems to be lower than that of DSL users who pay a lesser charge.

The recommendation provided is to conduct further analysis, models, and experimentation on the question. It may be beneficial to provide a pay-by-usage type service to entice those customers who use less internet to stay with the company. Another idea is to introduce a tiered pricing system where customers can move up or down based on their internet usage. Additionally, it could be beneficial to conduct further analysis on what other services are provided to customers who've churned verses those who didn't and attempt to determine if the lack of specific services reduced their internet usage.

Ultimately, providing a sense of the necessity of service and product value is the ultimate goal to retain customers and prevent them from churning.

F. Video

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=9fffb537-ca99-4bf5-b7f4-ad3c014cbfea

G. Third-Party Code Sources

Hadley Wickham and Jim Hester (2020). readr: Read Rectangular Text Data. R package version 1.4.0. https://CRAN.R-project.org/package=readrR Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

H. Sources

Mawer, C. (2017, September 19). The Value of Exploratory Data Analysis. Retrieved May 29, 2021, from https://www.svds.com/value-exploratory-data-analysis/

Natrella, M. et all (2013, October 30). Engineering Statistics Handbook. Retrieved May 29, 2021, from https://www.itl.nist.gov/div898/handbook/eda/eda.htm