Tucker Atwood

WGU MSDA

D207 Performance Assessment

3/7/24

**A1. Research Question**

Using data from the “churn” dataset, I will answer the question, “Is there a relationship between bandwidth use and churn rate?”

**A2. Benefits of Data Analysis**

Understanding the presence and scope of a connection between bandwidth use and churn rate could help stakeholders identify whether bandwidth data usage is a reliable predictor of customer churn. If there appears to be a strong relationship, the organization can use this information to target specific customers and prevent them from leaving the service. For example, if customers with low levels of bandwidth use are determined to be at a higher risk of discontinuing service, stakeholders could offer exclusive deals to these customers, or enact other methods of ensuring they stay with the service. Targeting this subset of customers would be more efficient and financially feasible than offering the same deals to all customers.

**A3. Relevant Data**

The data from the churn dataset that will be relevant in answering this question is as follows.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable: | Data type: | Description: | Example: |
| Churn | Character | Whether or not the customer stopped using the service in the last month. | Yes/No |
| Bandwidth\_GB\_Year | Numeric | Average amount of data customer has used per year in gigabytes. | 717.6824 |

**B1. Analysis Code**

To analyze this data in a manner that will inform the research question, a two-sample t-test was conducted. The results of this t-test will be used to determine whether the Bandwidth\_GB\_Year mean for customers who stayed with the service was significantly higher than the mean for customers who left the service.

* Null hypothesis and alpha level: ,
* Alternative hypothesis:
* : Mean Bandwidth\_GB\_Year for customers who stayed with the service.
* : Mean Bandwidth\_GB\_Year for customers who left the service.

The following code written in the language R produces a visual relationship between the two variables and executes the two-sample t-test as described. An executable version of this code can be found in the attached file: Atwood\_D207\_PA\_Code.R.

ggplot(churn, aes(x = Bandwidth\_GB\_Year, fill = Churn)) +

geom\_density(alpha = 0.5) +

labs(title = "Bandwidth and Churn Rate Density Plot") +

xlab("Bandwidth Gigabytes Per Year") +

ylab("Density")

# shows relationship between Bandwidth\_GB\_Year and Churn

churn\_no <- churn[which(churn$Churn == "No"),]

churn\_yes <- churn[which(churn$Churn == "Yes"),]

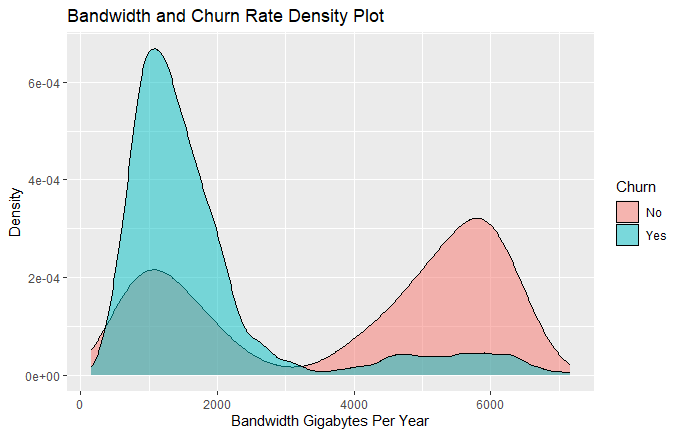
# split dataset into 2 groups based on Churn response

t.test(churn\_no$Bandwidth\_GB\_Year, churn\_yes$Bandwidth\_GB\_Year, alternative = "g", mu = 0, var.equal = FALSE)

# conducts two-sample t-test for difference in means between Churn No and Churn Yes groups

**B2. Analysis Results**

The following density plot provides a visual representation of the relationship between Churn and Bandwidth\_GB\_Year.



The following results summarize all calculations from the two-sample t-test.

* t value: 59.905
* degrees of freedom: 7264.6
* p-value < 2.2e-16
* 95 percent confidence interval: [2126.793, Inf]
* sample estimates:
  + mean of x: 3971.856
  + mean of y: 1785.009

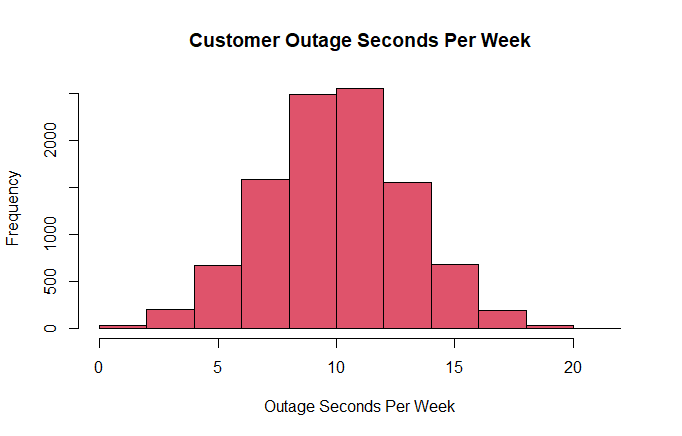
**B3. Analysis Justification**

A two-sample t-test was chosen to analyze this relationship because it is the standard hypothesis test for identifying whether differences between two unpaired groups are statistically significant. If the results are proven to be significant, a stakeholder can assume an inherent relationship between bandwidth use and churn rate that is not coincidental.

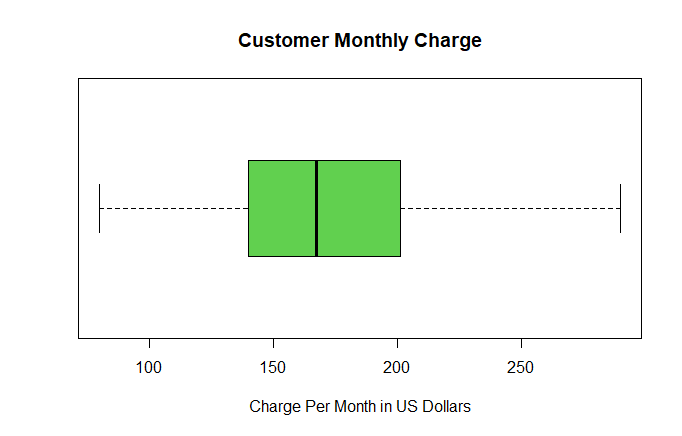
**C. Univariate Statistics Distributions**

The following visual representations of univariate data can be used to uncover patterns and trends within the data.

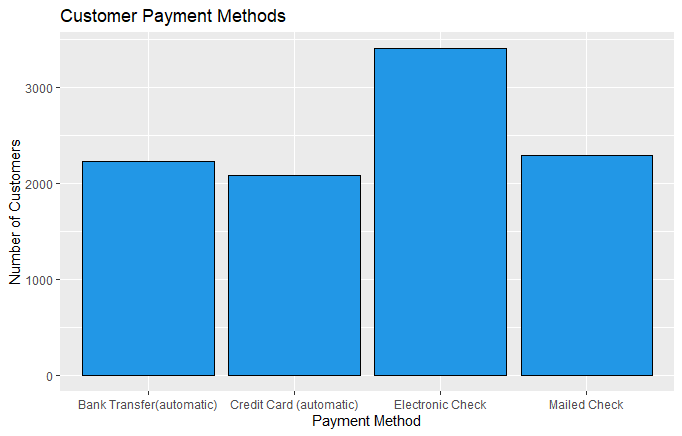
* Histogram of **univariate continuous** data: Outage\_sec\_perweek.



* The distribution of Outage\_sec\_perweek is Normal. The data is shaped like a bell curve, with a large group of values in the middle and the number of values decreasing in the left and right tails at a similar rate, creating a symmetric curve on each side.
* The following summary statistics further describe the Outage\_sec\_perweek distribution:
* Minimum: 0.09975
* 1st Quartile: 8.01821
* Median: 10.01856
* Mean: 10.00185
* 3rd Quartile: 11.96949
* Maximum: 21.20723
* Boxplot of **univariate continuous** data: MonthlyCharge.



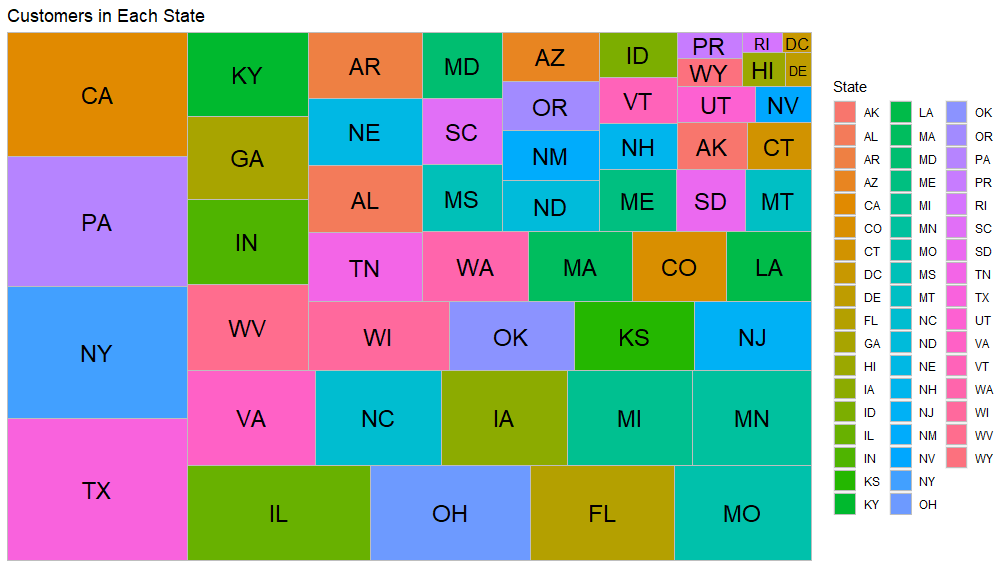
* The distribution of MonthlyCharge is relatively Normal, with a slight right skew. The middle 50% of the data, which is encompassed in the green “box,” is clustered closer together than the data between the box and each of the “whiskers,” especially the relatively small range of values between the first quartile (left side of the box) and the median (line in the middle of the box). This suggests more values in the middle of the dataset than in the tails, which would create a relatively Normal shape. The range of values between the third quartile (right side of the box) and the maximum (rightmost whisker) is larger than the range of values between the minimum (leftmost whisker) and the first quartile (left side of the box), signifying the slight skew to the right. There are also no outliers present.
* The following summary statistics further describe the MonthlyCharge distribution:
* Minimum: 79.98
* 1st Quartile: 139.98
* Median: 167.48
* Mean: 172.62
* 3rd Quartile: 200.73
* Maximum: 290.16
* Bar chart of **univariate categorical** data: PaymentMethod.



* The distribution of PaymentMethod is unimodal at “Electronic Check” and relatively uniform for the other three responses for the variable. While “Bank Transfer (automatic),” “Credit Card (automatic),” and “Mailed Check” show bars with heights in the same general area (between 2000 and 2500), “Electronic Check” is significantly higher than each, with a bar height of almost 3500.
* The following table further summarizes the PaymentMethod distribution:

|  |  |
| --- | --- |
| **PaymentMethod** | **n (Number of customers)** |
| Bank Transfer (automatic) | 2229 |
| Credit Card (automatic) | 2083 |
| Electronic Check | 3398 |
| Mailed Check | 2290 |

* Treemap of **univariate categorical** data: State.



* The distribution of State has a high variability, with 52 total states represented. Texas has the largest number of customers, followed by New York, Pennsylvania, California, Illinois, and so on. Though these states with the most customers have significantly higher totals than the states with the least customers (Washington DC, Rhode Island, Delaware, etc.), the progression from the bottom to the top of the list is relatively linear, with no noticeable large jumps from one state to another.
* The following table further summarizes the State distribution:

A table of numbers and letters

Description automatically generatedA table of numbers and letters

Description automatically generated

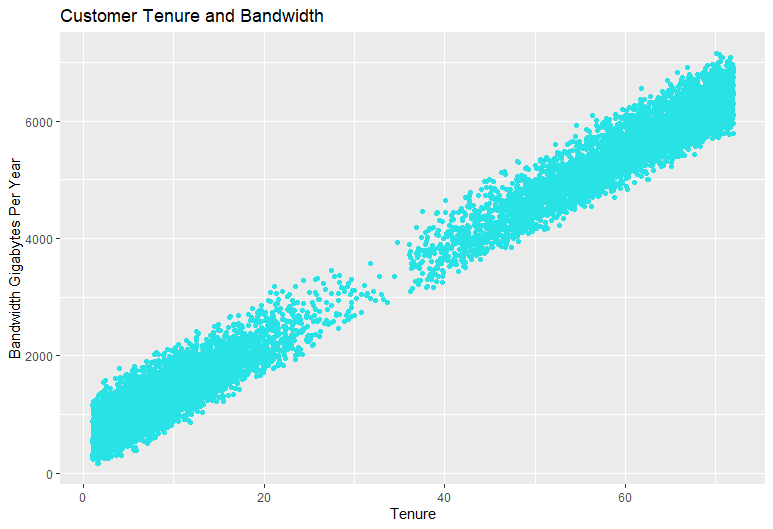
A table of numbers and letters

Description automatically generated

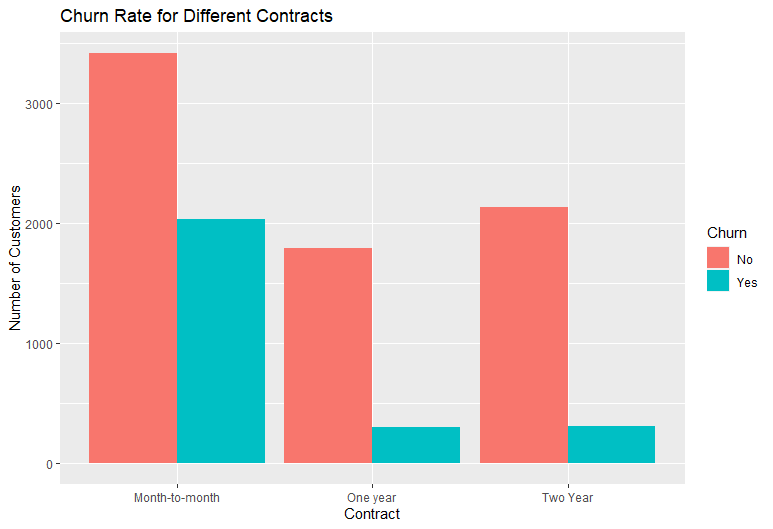
**D. Bivariate Statistics Distributions**

The following visual representations of bivariate data can be used to uncover patterns and trends within the data.

* Scatterplot of **bivariate continuous** data: Tenure and Bandwidth\_GB\_Year.



* The distribution of this data shows a strong positive linear trend. The lower values of Tenure are primarily paired with low values of Bandwidth\_GB\_Year, and as the values of Tenure increase, Bandwidth\_GB\_Year also increases at a predictable pattern, leading to high values of Tenure primarily paired with high values of Bandwidth\_GB\_Year. The correlation coefficient between these two variables was found to be 0.991, indicating that a strong positive correlation is present.
* The following summary statistics further describe the Tenure distribution:
* Minimum: 1.000
* 1st Quartile: 7.918
* Median: 35.431
* Mean: 34.526
* 3rd Quartile: 61.480
* Maximum: 71.999
* The following summary statistics further describe the Bandwidth\_GB\_Year distribution:
* Minimum: 155.5
* 1st Quartile: 1236.5
* Median: 3279.5
* Mean: 3392.3
* 3rd Quartile: 5586.1
* Maximum: 7159.0
* Grouped bar chart of **bivariate categorical** data: Churn rate and Contract.



* The distribution of this data is unimodal at “Month-to-Month,” with more customers under that contract than “One Year” or “Two Year.” This is true for both responses of Churn (“Yes” and “No”). There is also a noticeably taller bar height for “Yes” Churn responses in the “Month-to-Month” Contract than for the other two contract responses, especially in relation to the more subdued bar height differences for “No” Churn responses. This appears to indicate that customers with Month-to-Month contracts are more likely to leave the service than customers with longer-term contracts.
* The following table further summarizes the Churn rate and Contract distributions:

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Churn | |
| No | Yes |
| Contract | Month-to-Month | 3422 | 2034 |
| One Year | 1795 | 307 |
| Two Year | 2133 | 309 |

**E1. Hypothesis Test Results**

The two-sample t-test parameters and results were as follows.

* Null hypothesis and alpha level: ,
* Alternative hypothesis:
* : Population mean Bandwidth\_GB\_Year for customers who stayed with the service.
* : Population mean Bandwidth\_GB\_Year for customers who left the service.
* t value: 59.905
* degrees of freedom: 7264.6
* p-value < 2.2e-16
* 95 percent confidence interval: [2126.793, Inf]
* sample estimates:
  + mean of x: 3971.856
  + mean of y: 1785.009

Since the p-value is less than the designated alpha level of 0.05, there is enough evidence to reject the null hypothesis. The mean Bandwidth\_GB\_Year for customers who stayed with the service is greater than the mean Bandwidth\_GB\_Year for customers who left the service to a statistically significant degree. There is enough evidence to suggest that Churn rate is related to and can potentially be predicted by Bandwidth\_GB\_Year.

**E2. Analysis Limitations**

Although the hypothesis test determined a statistically significant relationship between Churn rate and Bandwidth\_GB\_Year, it is important to remember that correlation does not imply causation. Any relationship between any two variables should not directly lead to the assumption that variations in one variable cause variations in the other. There may be unmeasured or unanalyzed confounding variables that influence both variables of interest. For example, it was observed that a strong correlation existed between Tenure and Bandwidth\_GB\_Year. It is possible that Tenure is more important than Bandwidth\_GB\_Year in predicting and analyzing Churn rate, but because of the aforementioned correlation, a connection between Bandwidth\_GB\_Year and Churn rate could be erroneously applied.

**E3. Recommended Course of Action**

Based on the results of this hypothesis test, it is recommended to identify and target customers with low levels of Bandwidth\_GB\_Year to prevent them from leaving the service. It is also recommended to further explore other variables that may show connections to Churn rate and/or Bandwidth\_GB\_Year, since there may be better or more clear predictors of Churn. These results, along with further exploration, may help stakeholders retain customers they would otherwise lose.

**F. Panopto Recording**

Please refer to the link attached with a Panopto video recording. The video includes an explanation, execution, and output results of the referred code, along with code used to generate all visual representations of distributions included in this analysis.

**G. Third-Party Code References**

WGU Courseware was used as a resource to learn the methods, concepts, and functions used to create the codes in this project, including DataCamp course tracks (datacamp.com), Dr. William Sewell’s D207 webinar videos, and the book *Data Science Using Python and R* by Chantal D. Larose and Daniel T. Larose.

The code structure used to create a treemap was found here:

<https://rkabacoff.github.io/datavis/Univariate.html#tree-map>

This is referenced in-text in the code used to generate the visual representations of distributions included in this analysis. This can be found as part of the attached file: Atwood\_D207\_PA\_Code.R.

**H. References**

There is no content in this analysis which has been quoted, paraphrased, summarized, or otherwise requires direct citation.