D-207 EXPLORATORY DATA ANALYSIS PERFORMANCE ASSESMENT

PAOLA WILLIAMS

COLLEGE OF IT, WESTERN GOVERNOR'S UNIVERSITY

DECEMBER 27th, 2022

Table of Contents

Part I: Research Question and Variables

A. Real-world organizational situation

A1. Question for Analysis

Is there a difference in the Ouatge_sec_perweek experienced between customers that have churned and those that have not churned? Could this be one reason customers are leaving the company?

A2. Benefit from Analysis

From this study, stakeholders may be able to determine if customers that churned experienced a significantly different outage time than those that have decided to stay. If this is the case, we could infer that the outage might be one of the reasons why customers are leaving the company.

A3. Data identification

We will use the `Outage_sec_perweek` and `Churn` variables from the Churn dataset.

B. Data Analysis Description

B1. Code

```
#Load the csv file to a pandas dataframe
df =
pd.read_csv(r"C:\Users\paowm\Downloads\d9rkejv84kd9rk30fi2l\churn_clea
n.csv")

#Separate Income variable by Churn categories: Yes or No
x = df[df["Churn"]=="Yes"]["Outage_sec_perweek"]
y = df[df["Churn"]=="No"]["Outage_sec_perweek"]

#Perform t-test and print results
t_result = stats.ttest_ind(x,y)

alpha= 0.05
print("The alpha value is 0.05")
print("The p-value is " + str(t_result[1]))
if(t_result[1] < alpha):
    print("The means of x and y are different")
else:
    print("No significant difference found") (Hayden, n.d.)
```

B2. Code Output

Figure 1

Code output after running t-test

```
The alpha value is 0.05
The p-value is 0.98752251103374
No significant difference found
```

B3. Justification

In this case, the technique chosen is a two-sample t-test which is used to compare the means of two independent groups.
This test will help us determine if the two samples, those who churned and those who did not churn, differ significantly in terms of outage time (Patel, 2020).
We use this test because we are comparing the means of a continuous variable from two independent groups.

C. Univariate Statistics

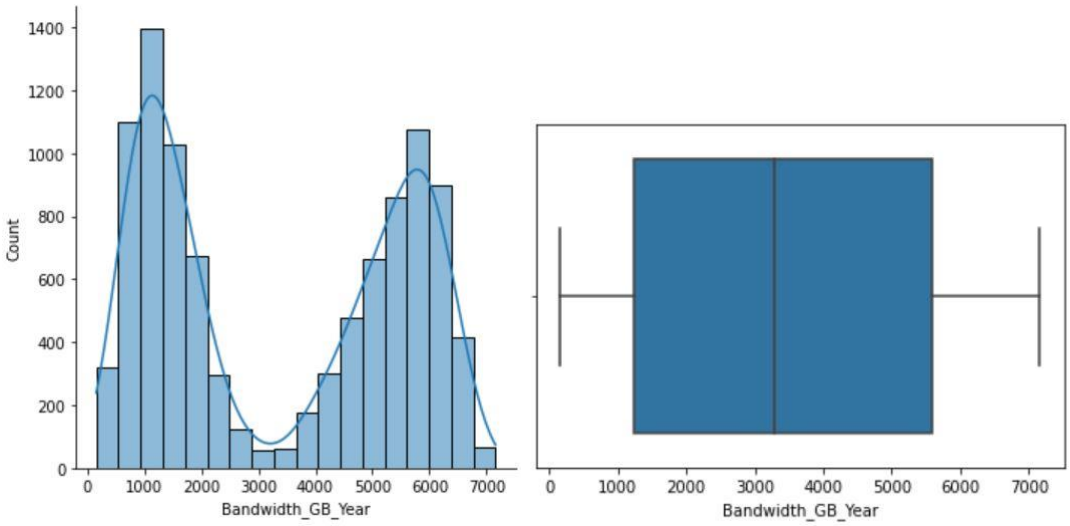
C1. Visual of findings

Continuous Variables

We can create a histogram and a boxplot for Bandwidth_GB_Year and Income, the continuous variables chosen, to understand the frequency of their data points and to visualize their distributions (Zach, 2021).

Figure 2

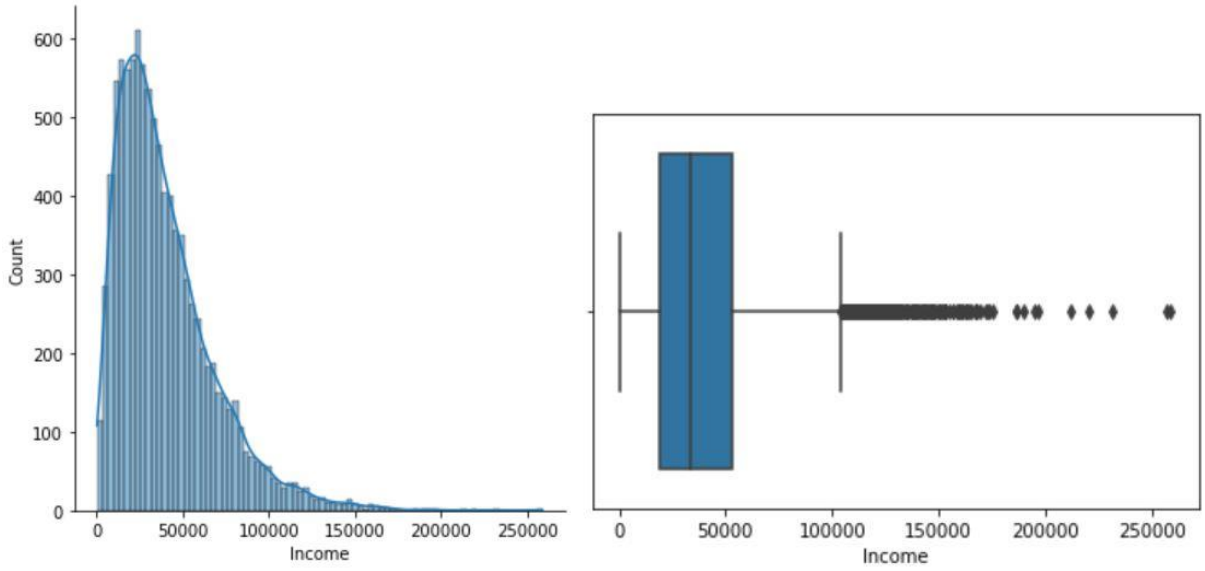
Histogram and boxplot for variable Bandwidth



From the plots above, we can identify two separate groups of bars which represents a bimodal distribution (Siegel, 2012).

Figure 3

Histogram and boxplot for variable Income



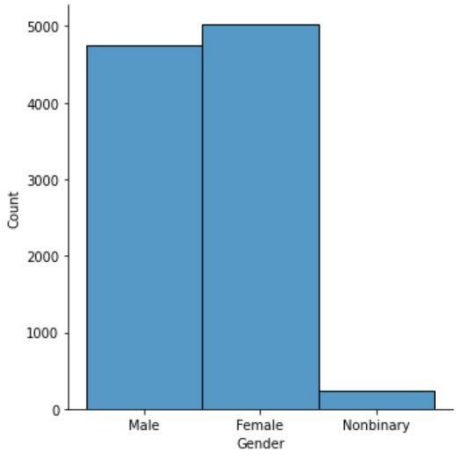
In this case, the variable Income shows a positively or right skewed distribution.

Categorical Variables

To study the categorical variables, i.e., Gender and Area, we chose to use a bar chart to visualize the count of each category within the variables.

Figure 4

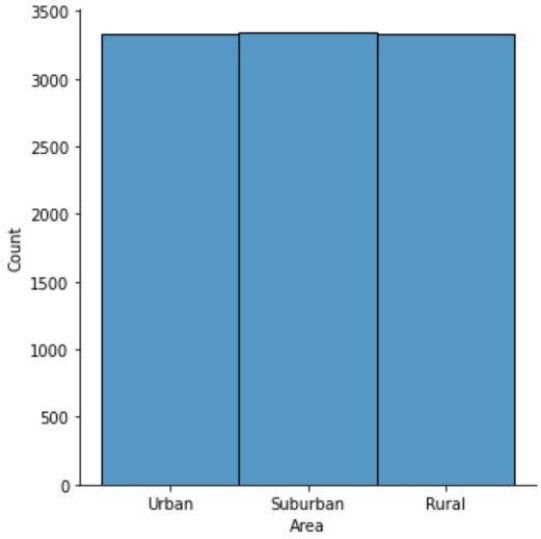
Bar chart for Gender



There are more females than males in the study. Moreover, there is a small number of non-binaries compared to males and females.

Figure 5

Bar chart for Area



The bar chart above indicates the distribution of Area within customers is even between the categories urban, suburban, and rural.

D. Bivariate statistics

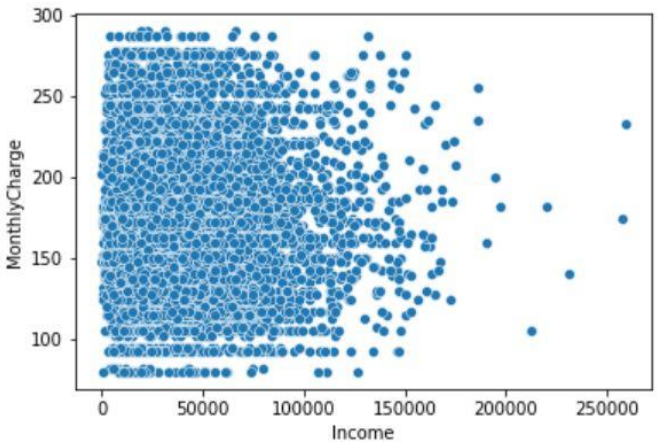
D1. Visual of findings

Continuous variables

A scatterplot is a way to study the relationship between two continuous variables. Income and MonthlyCharge are the variables included in this bivariate analysis.

Figure 6

Scatterplot for MonthlyCharge vs Income



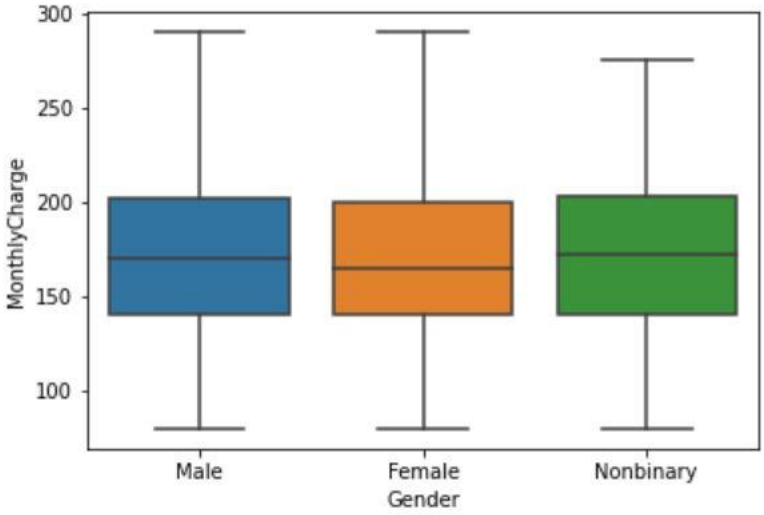
There is no clear relationship between MonthlyCharge and Income since there is no pattern observed.

Categorical variables

To visualize the distribution of two categorical variables, Area and Gender, we used a stacked-bar chart (Kumar, 2022).

Figure 7

Monthly charge vs Gender boxplot



We can see a slight difference in the non-binary maximum data point and the other categories, Male and Female. As for the average, the Female monthly charge average is lower than the Male and the non-binary, which seems to have the highest of them all.

E. Results of analysis

E1. Results of analysis

We want to find if the customers who churned and those who have not churned have significantly different outage time. We use a two-sample t-test to infer whether the means of these two groups are different or not (Hayes, 2022).

Our null hypothesis is that the means of outage time from both groups are equal.

Ho: $\mu 1=\mu 2$

The alternative hypothesis is that the means are different.

H1: $\mu 1\neq\mu 2$

We are choosing an alpha value of 0.05. For the mentioned alternative hypothesis, there are two critical areas on both the positive and negative sides of the distribution of 0.025 each (Frost). We compare the p-value calculated with the alpha value of 0.025. If the p-value is less than 0.025 we can reject the null hypothesis.

The p-value calculated after performing the t-test is 0.987 which is higher than 0.025, so we cannot reject the null hypothesis. There is not a statistically significant difference between customers who churned and did not churn regarding their experienced outage time. This means that the outage is not a reason why customers are churning.

E2. Limitations of analysis

A limitation of this analysis would be that we are only studying the impact of one continuous variable (outage time) in the variable Churn.

If we have chosen two categorical variables to compare, i.e, Gender and Churn, the best option would have been to use a Chi-Square test (Chauhan, 2020).

On the other side, if we wanted to compare multiple samples in one single test, we would have to use an ANOVA test. Furthermore, an ANOVA test is used when a categorical variable has more than two categories.

E3. Recommended course of action

Since customers are not leaving because of outage time, there is no need to minimize outage time.

We should focus on finding the reasons of churn and take action to prevent customers churning.

G. Third-Party Code References

Hayden, L. (n.d.). Performing Experiments in Python. Retrieved from The Basics of Statistical Hypothesis Testing: https://app.datacamp.com/learn/courses/experimental-design-in-python

H. References

Chauhan, N. (2020, March 27). Medium. Retrieved from https://medium.datadriveninvestor.com/p-value-t-test-chi-square-test-anova-when-to-use-which-strategy-32907734aa0e

Frost, J. (n.d.). Statistics by Jim. Retrieved from One-Tailed and Two-Tailed Hypothesis Tests Explained: https://statisticsbyjim.com/hypothesis-testing/one-tailed-two-tailed-hypothesis-tests/

Hayes, A. (2022, July 20). Investopedia. Retrieved from T-Test: What It Is With Multiple Formulas and When To Use Them: https://www.investopedia.com/terms/t/t-test.asp

Kumar, A. (2022, March 22). Retrieved from A Quick Guide to Bivariate Analysis in Python: https://www.analyticsvidhya.com/blog/2022/02/a-quick-guide-to-bivariate-analysis-in-python/

Patel, K. (2020, October 30). Dataversity. Retrieved from Data Topics: https://www.dataversity.net/the-independent-samples-t-test-method-and-how-it-benefits-organizations/#

Siegel, A. F. (2012). Practical Business Statistics(Sixth Edition). Academic Press.

Zach. (2021, February 25). Retrieved from Statology: https://www.statology.org/univariate-analysis/