D207 Exploratory Data Analysis Performance Assessment

Author: Eric Colwell

Western Governors University

This report deals with the rubric for Western Governors University course D207 Exploratory Data Analysis and answers all the items in rubric order.

# **Part I:**

**A1:**

Is there a significant difference between ‘Yes’ and ‘No’ in the ‘Churn’ variable when tested with the ‘MonthlyCharge’ variable?

**A2:**

Analysis of this question could reveal insight as to whether there is a correlation between churn (loss) of customers and the amount being charged for the service. The information gained from this analysis could benefit stakeholders in the organization by prompting them to investigate ways to lessen churn in the customers that are more likely to discontinue service due to higher monthly charges. By retaining these customers, the organization would be more profitable.

**A3:**

Churn: Datatype is object, qualitative, ordinal, and has no null values. This variable shows whether the customer discontinued service within the last month (Yes = 2650, No = 7350).

MonthlyCharge: Datatype is float64, quantitative, continuous, and has no null values. The value reflects an average per customer (ranges from roughly 80 to 290).

**B1:**

I chose to use Python to perform a t-test hypothesis analysis for this project. A copy of the code I used is uploaded in the submission in a .txt format.

**B2:**

In the screenshot below I found the t-statistic to be greater than 39 and the p-value to be well below the 0.05 alpha level. This means that the null hypothesis is rejected and there is most likely a statistical difference between ‘Churn’ and ‘MonthlyCharge’. The p-value of 1.78 with 290 zeros in front of it shows the test results are above 99.99999% percent probable this statistical difference is not by chance.

**Text, application

Description automatically generated**

**B3:**

T-tests are used to find statistically significant difference in means and to either support or reject a null hypothesis. A null hypothesis means that there is no statistically significant difference. To reject the null hypothesis would mean there is a significant difference. I used a decision tree to determine which analysis technique to use for this project. I am comparing means with two samples that are independent of each other. Following the tree leads me to an independent-samples t-test, which I utilized for this project. Shown below is a screenshot of the decision tree from a Powerpoint presentation webinar provided by Dr. William Sewell as part of the course learning material (Sewell, 2022).

Diagram

Description automatically generated

**C:**

Below are screenshots of the code used and output of univariate statistics for the variables ‘MonthlyCharge’, ‘Tenure’, ‘Churn’ and ‘PaperlessBilling’. The continuous variable statistics shown in the output include the count of instances, mean, standard deviation, minimum, maximum, and interquartile values. The categorical variable statistics shown are the count of instances, unique (equals 2: there are only two possibilities, Yes or No), top (which is the most frequent), and freq (the count of the most frequent answer).

MonthlyCharge (continuous):

Table

Description automatically generated

Tenure (continuous):

Text, table

Description automatically generated

Churn (categorical):

Text, table

Description automatically generated with medium confidence

PaperlessBilling (categorical):

Text

Description automatically generated

**C1:**

**Chart, histogram

Description automatically generated**

**Chart, histogram

Description automatically generated**

**Chart, bar chart

Description automatically generated**

**Chart, bar chart

Description automatically generated**

**D:**

I re-expressed the variables ‘Churn’ and ‘PaperlessBilling’ from object type into int64 type. The new variables are ‘Churn\_numeric’ and ‘PaperlessBilling\_numeric’ where No = 0 and Yes = 1. This is necessary to test categorical variables with continuous variables. Below are screenshots of the output in Jupyter notebook showing the bivariate statistics.

**Tenure compared with PaperlessBilling:**

Text

Description automatically generated

**MonthlyCharge compared with Churn:**

Text

Description automatically generated

**D1:**

**Chart

Description automatically generated**

**Chart, histogram

Description automatically generated**

**E1:**

The results of the hypothesis test show that there is a significant difference due to the low p-value. This means that the null hypothesis is rejected.

**E2:**

The limitations of t-tests are that a normal distribution is assumed, and inferential statistics are used to determine significance. T-tests only examine means, not individuals. Therefore, when testing ‘MonthlyCharge’ and ‘Churn’ I can only derive there is a significant difference between their means. I cannot derive a higher monthly charge will cause a customer to discontinue service.

**E3:**

The graph for ‘MonthlyCharge’ and ‘Churn’ shown above in part D1 indicates there is a point where customers who have a higher monthly charge are more likely to discontinue service. This is supported by the t-test results with a very low p-value. My recommendation is to test other variables with ‘Churn’ to determine if there is a course of action that may counteract customers with higher monthly charges from discontinuing service. An example would be if customers with higher monthly charges were offered free tech support for 6 months, would they then continue service with the company.

**F:**

The Panopto video recording URL has been uploaded separately as part of the submission.

References

Sewell, W. (2022). D207 Exploratory Data Analysis Webinar. Retrieved from <https://westerngovernorsuniversity-my.sharepoint.com/:p:/g/personal/william_sewell_wgu_edu/ER_ESbgVK1VMpvpHIVCf0WYBn2BysB8AVdJ25RA3ys1nPg?e=QvmbT5>