**EXPLORATORY DATA ANALYSIS (OEM2)– D207**

**Performance Assessment**

**Western Governors University**

**By: Christian LeBlanc**

**A-1**

For this performance assessment, I am using the churn dataset. I am going to be analyzing the question “Do customers with multiple lines have an increased rate of having tech support than the rest of the customers?” This would give a null hypothesis (H0) of there is no significant difference in the rate of having tech support between customers with multiple lines and the rest of the customers.

**A-2**

I suspect that customers with multiple lines do have an increase rate of having tech support. This is because with the multiple lines of service comes multiple issues that could arise in which tech support becomes more valuable to these customers. This information could be used to upsell the multiple lines that do not have tech support yet, making them even more valuable to the company, if we know they are statistically more likely to get tech support.

**A-3**

The two variables that are relevant to answering my question are Multiple, and TechSupport. Multiple is does the customer have multiple lines with the company. It is categorical with yes or no responses. TechSupport is does the customer have the technical support add-on. It is categorical with yes or no responses.

**B-1**

Chi-square testing is the technique used. Please see my Jupyter notebook file named “Christian LeBlanc D207 PA 2” for my code.

**B-2**

I set an alpha of 0.05 for my testing. This would allow for 95% certainty that the any p-value less than the alpha would be statistically significant. I had a degree of freedom of 1, which gives me a critical value of 3.841. The chi-square statistic is 1.031. Since the chi-square statistic is not larger than the critical value it fails to reject the null hypothesis (H0). The chi-square test gave a p-value of 0.310 which is larger than the alpha that was set, so it failed to reject the null hypothesis (H0). From the results of my chi-square test I can conclude that the distribution of customers with multiple lines of service that have tech support is not statistically significant from the distribution of customers with multiple lines of service that do not have tech support.

**B-3**

The chi-square test was chosen due to the type of variables that were used to answer my question. Both Multiple and TechSupport are categorical variables. Also, the type of question I needed to answer. Looking to see if having multiple lines of service affected the likelihood of having tech support. The chi-square test allowed to look at if these two variables were dependent if the null hypothesis (H0) was rejected or independent if it fails to reject the null hypothesis (H0).

**C**

MonthlyCharge is my first continuous variable. I will be rounding my responses with this variable because it deals with currency. It has a mean of $172.62 with a standard deviation of 42.94. It has a median, 50th percentile, of $167.48. The 25th percentile is 139.98 and 75th is 200.73. It looks like it has a slight right-skewed dataset.

Bandwidth\_GB\_Year is my second continuous variable. It has a mean of 3392.34 with a standard deviation of 2185.29. It has a median, 50th percentile, of 3279.54. The 25th percentile is 1236.47 and 75th is 5586.14. It has a bimodal dataset with two distinct peaks.

Contract is my first categorical variable. It shows 54.6% of customers have month-to-month contracts, 21.0% have one-year contracts, and 24.4% have two-year contracts.

InternetService is my second categorical variable. It shows 44.1% have fiber optic internet service, 34.6% have DSL, and 21.3% have none. The 21.3% that have none seems really high percentage of people that do not have internet service. This makes me wonder if these customers get internet services from a different provider.

**C-1**

Please see below for my visual univariate statistics. They can also be found in my Jupyter notebook named “Christian LeBlanc D207 PA 2.”

A graph of a number of blue bars

Description automatically generated with medium confidence

A graph of a distribution of bandwidth

Description automatically generated

A pie chart with numbers and a number of numbers

Description automatically generated

A pie chart with numbers and text

Description automatically generated

**D**

For bivariate statistics I compared a continuous variable to a categorical variable twice. The first one I looked at is the categorical variable Gender use of continuous variable Bandwidth\_GB\_Year. I picked these two because I thought seeing the difference of bandwidth usage between the genders would reveal interesting findings. I was surprised when it shows the distribution of bandwidth usage between male, female, and nonbinary were all very similar. The interquartile range on all three is very close. The medians, 50th percentile, is the most glaring difference between the three, with the nonbinary having the lowest and female having the highest. The other one I looked at is the categorical variable Area earning of continuous variable Income. This one was highly surprising to me. I expected rural to have much lower income in comparison to the others because of the lack of job opportunities in rural areas. The visual representation for this shows almost identical distribution. The only real thing one can comment on from the boxplot is the outliers, which both rural and suburban both having 250,000+ outliers that urban does not have even though all three have outliers.

**D-1**

Please see below for my visual bivariate statistics. They can also be found in my Jupyter notebook named “Christian LeBlanc D207 PA.”

A diagram of a group of blue squares

Description automatically generated

A diagram of a relationship between an area and an area

Description automatically generated

**E-1**

The results of the hypothesis test are that the chi-square test shows it fails to reject the null hypothesis (H0). The null hypothesis (H0) is there is no significant difference in the rate of having tech support between customers with multiple lines and the rest of the customers. As I mentioned in section B-2, I produced two ways that shows it fails to reject the null hypothesis (H0). The degree of freedom of 1 and alpha of 0.05 gives a critical value of 3.841. Then having a chi-square statistic, 1.031, that is smaller than the critical value it fails to reject the null hypothesis (H0). The other way is through the p-value. The p-value is 0.310. The alpha value is set to 0.05. Since the p-value is larger than the alpha value the outcome is it fails to reject the null hypothesis (H0).

**E-2**

The chi-square test, like all analysis, has limitations associated with it. A big one for this particular analysis is even if it had given a statistically significate result, it would show a correlation, but it does not establish causation. Sample size is always a concern when doing analysis. It is an unknown if the 10,000 records in the data for all the customers or just a segment of the customers. Given that this is a nice round number I would suspect it is not the complete number of customers. This brings another concern, if the data is only part of the full customer data, then how was it selected. Chi-square tests require that the samples be random samples.

**E-3**

Based on the results of the chi-square test I would recommend to not pursue customers with multiple line of service to try and get them to add tech support. Customers with multiple lines do not have a statistically significant increased rate of having tech support than the rest of the customers. I proposed this question with the mindset of being able to increase the revenue produced from current customers. After seeing the results, I cannot say that targeting this section of customers will produce any better results than doing it for the general customer base.

**F**

Uploaded to the Panopto drop box titled “Exploratory Data Analysis OEM2 | D207 (student creators) [assignments].”

Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=7e6f23d0-612f-4129-883c-b0cf01530d8a>

**G**

Western Governors University. (n.d.) D207 Webinar Ep. 5 <https://westerngovernorsuniversity-my.sharepoint.com/:p:/g/personal/william_sewell_wgu_edu/EbIWOqJp0oJFrByVqCO5wfgB7LHdWLxGgzYyHXd2nuIV2Q?e=PrgdzD>

Western Governors University. (n.d.) D207 Webinar Ep. 6 <https://westerngovernorsuniversity-my.sharepoint.com/:p:/g/personal/william_sewell_wgu_edu/ETLpSFztgpJDj0TvY0peCCUB1dfNQW3Pp0Z5ToNKQoyugQ?e=OTnYR0>

**H**

No in-text citations in this paper.

**I**

The content in this Performance Assessment is set up and presented with the highest professional standards.