Noah Cina

04/18/2024

D207 – Exploratory Data Analysis

Performance Assessment

*A.  Describe a real-world organizational situation or issue in the Data Dictionary you chose, by doing the following:*

*1.  Provide****one****question that is relevant to your chosen data set. You will answer this question later in the task through an analysis of the cleaned data, using one of the following techniques: chi-square, t-test, or analysis of variance (ANOVA).*

What customer data is most significantly linked to churn?

*2.  Explain how stakeholders in the organization could benefit from an analysis of the data.*

Understanding which categories of data collected are connected to churn will provide insight into how to prevent more customers from dropping service. Before a solution can be implemented, the problem must be diagnosed.

*3.  Identify*all*of the data in your data set that are relevant to answering your question in part A1.*

The following variables have been determined to be relevant for the research question:

Churn – Categorical

Income – Numerical continuous

Outage\_sec\_perweek – Numerical continuous

Tenure – Numerical continuous

MonthlyCharge – Numerical continuous

Bandwidth\_GB\_Year – Numerical continuous

*B.  Describe the data analysis by doing the following:*

*1.  Using one of the following techniques, write code (in either Python or R) to run the analysis of the data set:*

*•   chi-square*

*•   t-test*

*•   ANOVA*

Two sample t-tests have been chosen for analysis of the research question and relevant variables.

*2.  Provide the output and the results of*any*calculations from the analysis you performed.*

See the attached .ipynb file for code and output.

*3.  Justify why you chose this analysis technique.*

All the variables are numerical, continuous, and compared on the Churn statistic. Separating each variable into two categories based on Churn and comparing their means is an acceptable form of analysis for this dataset.

The only caveat in this case is that not all the variables in this set are normally distributed, and none of the tests available are best suited for analyzing two abnormally distributed continuous variables. The large sample size (about >2000 for any given variable and Yes/No churn combination) allows a t-test to be used in this case because of the central limit theorem. If I were able, I’d instead use a Wilcoxon signed-rank test in all these cases which does not assume the samples are normally distributed.

*C.  Identify the distribution of****two****continuous variables and****two****categorical variables using univariate statistics from your cleaned and prepared data.*

Four variables were analyzed, with two being continuous and two being categorical.

Outage\_sec\_perweek:

count 10000.000000  
 mean 10.001848  
 std 2.976019  
 min 0.099747  
 25% 8.018214  
 50% 10.018560  
 75% 11.969485  
 max 21.207230

Tenure:  
 count 10000.000000  
 mean 34.526188  
 std 26.443063  
 min 1.000259  
 25% 7.917694  
 50% 35.430507  
 75% 61.479795  
 max 71.999280

Churn:  
 No: 7350  
 Yes: 2650

Contract:  
 Month-to-Month: 5456  
 Two Year: 2442  
 One Year: 2102

*Represent your findings in Part C, visually as part of your submission.*

*A graph of a normal distribution

Description automatically generatedA graph showing the function of a certain number

Description automatically generated*

*A graph with red and blue squares

Description automatically generatedA graph of a bar graph

Description automatically generated with medium confidence*

*D.  Identify the distribution of****two****continuous variables and****two****categorical variables using bivariate statistics from your cleaned and prepared data.*

Four variables were analyzed: a pair of two continuous variables and a pair of two categorical variables.

The relationship between Tenure and Bandwidth\_GB\_Year was found to fit a linear model with a slope of 81.94, an intercept of 563.31, and an R squared value of 0.983.

A Chi-square test was conducted on the relationship between customer Gender and whether or not they self-identified as technically proficient. The p value of the resulting test was 0.30, indicating there was no significant correlation between the two variables.

1. *Represent your findings in Part D, visually as part of your submission.*

***A graph with a red line

Description automatically generatedA graph of a person and person

Description automatically generated***

*E.  Summarize the implications of your data analysis by doing the following:*

*1.  Discuss the results of the hypothesis test.*

For these tests, the null hypothesis indicates that the variable has no correlation with churn, and the alternative hypothesis indicates that the variable has statistically significant correlation with churn. The alpha value chosen for the hypothesis tests was 0.05 (5%).

The hypothesis test was conducted on 5 variables in the dataset, with each split on the churn statistic. The test found that customer income and service outage time were not correlated with churn, while the customer tenure, monthly service payment, and bandwidth usage statistics were strongly correlated with churn.

1. *Discuss the limitations of your data analysis.*

As with any statistical analysis like this, correlation does not mean causation. There is a possibility of confounding variables in the cases where churn was linked with tenure, monthly service payment, and bandwidth usage.

As stated previously, a Wilcoxon signed-rank test may have been better suited for analysis of these statistics.

1. *Recommend a course of action based on your results.*

Based on the findings, there appears to be a link between churn, tenure, and bandwidth usage. That is to say: customers who use less data appear to have shorter tenure and drop the service provider. There is a lesser connection to monthly payment in the fact that higher prices were correlated with higher churn. Advertising could be used to encourage new customers to use the internet more frequently and offer lower price packages for the first months of service, as newer customers appear to be the most likely to drop the service.

*F.  Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the tool(s) used.*

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=00bad482-ab62-4f56-b744-b15701793a17>

*G.  Reference the web sources used to acquire segments of third-party code to support the analysis.*

No third-party sources were referenced.

*H.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.*

No third-party sources were referenced.