**Exploratory Data Analysis**

**Amber Peacock**

Exploratory data analysis, as the name suggests, is the use of exploratory tools in languages like Python and R, to conduct hypotheses testing and create statistical models to gather information on the data at a deeper level. This part of the data analysis process follows data cleaning. For this assessment, I will be implementing exploratory data analytics strategies on the Churn dataset to test my hypothesis which I will discuss in this report.

A1. The question I am posing for this assignment is as follows: Does a customer categorized as ‘Techie’ decrease the likelihood of churn according to the data? The test I conduct will need to be applied to categorical variables, not numerical variables so I will be using Chi-Square to gather insight. In respect of this hypothesis, the null hypothesis would be that there is no statistical significance between the two variables.

A2. In the telecommunications industry, stakeholders can benefit from an analysis of their churn data set by adding business insights into what factors may be increasing churn and therefor decreasing customer retention and company profits. If a company can pinpoint what is causing churn, the company can then come up with solutions as to how to prevent those issues with future customers and provide ways to keep current customers. Finding out if a customer being classified as not techie has significance statistically to the rate of churn, the business can then pose ideas on how to prevent less tech savvy customer’s from leaving.

A3. The variables in the dataset that are relevant to answering my question in part A1 are as follows:

* Techie
* Churn

The observations in these variables will provide insights when used in my Chi-Square test I will conduct.

B1. See code in the attached file.

B2. In the images below, you can see the results of my contingency table code, my Chi-Square code, and my p-value test. As a result of my code, it was determined that the null hypothesis can be rejected meaning my hypothesis that there is statistical significance between the variables ‘Techie’ and ‘Churn’ is true.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

B3. I chose this analysis technique to test my theory because of the types of data I was using. I was testing statistical significance between two categorical (qualitative) data types and according to Dr.Gagner’s webinar, when using two categorical variables, the technique to use is Chi-Square. (D207 Top PA Tips for the PA, part 1, n.d.)

C. See attached visual.

D. See attached visual.

E1. After performing the Chi-Square test on my hypothesis, the results were that there is a significant statistical relationship between the variables ‘Churn’ and ‘Techie’. Based on the results, non-techie customers were more likely to churn. With this test, I was able to rule out the null hypothesis of no statistical significance. This test proved my hypothesis to have validity.   
 E2. It’s important to note that as with any test, correlation does not mean causation, so this only proves likely association between the variables. Also, while Chi-Square is a great resource to test categorical variables, it lacks the ability to note the direction of the relationship or how strong that relationship is. (When Chi-square Is Appropriate - Strengths/Weaknesses, 2023)

E3. Based on the results of my analysis, I would recommend that the telecommunications company promote the technical support add-on offered by the company. The company could be proactive by offering the first year free. An alternative route would be to offer it for free or reduced cost if customers indicate that they would like to leave the company based off reasons that having the add-on would resolve. Based on the relationship between the variables, this would be a good course of action for the company to reduce churn.

G. No third-party sources were used for the code production. I used resources provided by Dr.Gagner and Dr.Sewell such as pdf’s and webinars.

# **H.** References

*D207 Top PA Tips for the PA, part 1*. (n.d.). Retrieved from Panopto: https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=e474f3ef-db74-4d3c-bbf6-b04e0148a0ce&start=0

Dr.Gagner. (n.d.). *Guide to Conducting Chi-Square Test for Two Categorical Variables using Python.* Retrieved from https://srm.file.force.com/servlet/fileField?id=0BE3x000000gLNn

*When Chi-square Is Appropriate - Strengths/Weaknesses.* (2023). Retrieved from Passel: https://passel2.unl.edu/view/lesson/9beaa382bf7e/14