OEM2 TASK 1: EDA: EXPLORATORY DATA ANALYSIS

Cali Ciccarelli

D207

5/1/2024

**A.**

1. Is the mean income for those who did churn and those who did not statistically different?

Null Hypothesis: There is no effect of income on the churn rate of the customers.

Alternative Hypothesis: Customers were more likely to churn when their income is lower.

1. The stakeholders can benefit from the analysis because they will be able to tell if there is an income disparity between those who choose to stay and those who choose to leave the company. Let’s say they find out they are losing people with lower incomes more often; they may come out with marketing deals or promotions that promote long term customers. For example “For every year you stay with us, get cheaper prices!”. If they find there is no significant difference between incomes of those who churn compared to those who don’t, then they will know they do not need to waste money and time on creating deals like this.
2. The data relevant would be “Income” and “Churn”. I will have to calculate the mean incomes for those who did churn and those who did not.

**B.**

**1.** I choose to do a T-Test. Here is my Python code:

Import pandas as pd

import statistics

from scipy import stats

df = pd.read\_csv(r"C:\Users\Cali\Downloads\datamarch24\churn\_clean.csv")

churn\_incomes = df[df['Churn'] == 'Yes'] ['Income']

no\_churn\_incomes = df[df['Churn'] == 'No']['Income']

t\_result= stats.ttest\_ind(churn\_incomes, no\_churn\_incomes)

print(t\_result)

#is it significant?

Alpha = 0.05

if (t\_result[1] < Alpha):

print("The mean income for those who chose to stop services and those who chose to stay are statistically different")

else: print("No significant differences between the mean incomes of those who renewed services and those who did not renew.")

**2.** Results from running the code: TtestResult(statistic=0.5936894196669502, pvalue=0.552733291902922, df=9998.0)

No significant differences between the mean incomes of those who renewed services and those who did not renew.

**3.** I chose to do a two-sample T-Test because I was trying to compare the means of two populations. Since it is only two populations, those who did churn and those who did not, T-test is the best choice. If I was comparing more than 2 populations, the ANOVA test would have been best. Also, the mean values I was comparing were numerical, continuous variables which made the T-test the best choice. If I was looking at categorical values, I would have been better choosing the Chi-square test.

**C.**

For my two categorical variables, I looked at the distribution of gender and the distribution of contract type. For the gender variable, I used a bar graph and found that there were slightly more females than men. There were very few people who identified as nonbinary compared to male and female. For contract type, I used a pie chart. I found that more than half of the customers are on a month-to-month contract. Meanwhile around a quarter are two-year and a fifth are on a one-year contract. I also used python’s describe() function to find out further information on the distribution of the variables. For Gender, female is the most common. For Contract, month to month is the most common.

Below is the code I used to plot the categorical variables along with the graphs for each variable:

gender\_totals = df['Gender'].value\_counts()

plt.bar(gender\_totals.index, gender\_totals.values)

plt.title('Gender Distribution')

plt.xlabel('Gender')

A graph of a number of blue rectangular bars

Description automatically generatedplt.ylabel('Number of Customers')

contract\_totals = df['Contract'].value\_counts()

plt.pie(contract\_totals.values, labels=contract\_totals.index, autopct='%1.1f%%', startangle=90)

plt.title('Contract Type Distribution')

plt.axis('equal')

plt.show()

A pie chart with numbers and a few different colored circles

Description automatically generated

I also looked at the distributions of two continuous variables, ‘MonthlyCharge’ and ‘Income’. Using the describe() function, I found that for ‘Income’, the mean is 39806.93 and the standard deviation is 28199.92. For ‘MonthlyCharge’, the mean is 172.624816 and the standard deviation is 42.943094. I found that Income distribution is positively skewed because the median is before the mean (Dagli, 2020). MonthlyCharge is also slightly positively skewed but is closer to a normal distribution. Here is the code I used to graph these as well as the graphs themselves:

plt.hist(df['MonthlyCharge'], bins=20, color= 'darkgreen', edgecolor= 'gold')

plt.title('Distribution of Monthly Charge')

plt.xlabel('Monthly Charge')

plt.ylabel('Customers')

A green and yellow graph

Description automatically generated

plt.hist(df['Income'], bins=40, color= 'hotpink', edgecolor= 'silver')

plt.title('Distribution of Income')

plt.xlabel('Income')

A graph of a distribution of income

Description automatically generatedplt.ylabel('Customers')

**D.** BIVARIATE GRAPHS

For my bi-variate statistical analysis, I used boxplots to look at the relations between categorical variables and continuous variables. For one, I looked at the relationship between ‘MonthlyCharge’ and ‘Churn’. I found that those with a higher monthly charge were more likely to churn. This is important data because my t-test had shown that income did not effect churn rate, which could have led stakeholders to not worry about pricing/deals. This relationship gives evidence that they may want to focus on bringing costs down for the end users. I also looked at the relationship between gender and income, this showed that for the customers, there was no clear relationship between income and gender. Gender does not seem to impact income with this sample.

plt.boxplot([df[df['Churn']== 'Yes']['MonthlyCharge'], df[df['Churn']=='No']['MonthlyCharge']],

labels=['Yes', 'No'])

plt.title('Relationship between Churn and Monthly Charge')

plt.xlabel('Churn')

plt.ylabel('Monthly Charge')

A graph with lines and numbers

Description automatically generated with medium confidence

plt.boxplot([df[df['Gender']== 'Male']['Income'], df[df['Gender']=='Female']['Income'], df[df['Gender']=='Nonbinary']['Income']],

labels=['Male', 'Female', 'Nonbinary'])

plt.title('Relationship between Gender and Income')

plt.xlabel('Gender')

plt.ylabel('Income')

A graph of a person and person

Description automatically generated

plt.scatter(df['Income'], df['MonthlyCharge'])

plt.title('Relationship between Income and Monthly Charge')

plt.xlabel('Income')

plt.ylabel('Monthly Charge')

**E.**

1. My hypothesis test, the two-sample t-test, proved the null hypothesis to be correct. The incomes did not statistically affect the churn rate for customers. This means that stakeholders do not need to target any specific economic group.

2. Just as correlation does not equal causation, no correlation does not lead to firm conclusions either (Isager, 2023). While we can see income did not appear to have an effect on churn rate, it still may in certain ways we cannot see. It is also important to note that stakeholders cannot jump to conclusion and raise prices since income did not effect the churn rate. It is shown clearly in my bivariate statistical analysis that there is a correlation between rising prices and the chances a customer will churn. So, income alone is not something we can look at to determine what prices should be.

3. I would recommend that stakeholder advertise to all income brackets equally. I would also recommend that the company offer more deals to all customers since it appears that customers with all different incomes all seem to enjoy lower monthly prices.

F. Panopto Video Link: https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=4f56994d-7d88-4e8e-88bb-b164005aeb5c

H.

Dagli, R. (2022, September 15). *Skewness and kurtosis – positively skewed and negatively skewed distributions in statistics explained*. Skewness and Kurtosis – Positively Skewed and Negatively Skewed Distributions in Statistics Explained. https://www.freecodecamp.org/news/skewness-and-kurtosis-in-statistics-explained/

Isager, P. M. (2023, November 24). Why does correlation not equal causation? https://pedermisager.org/blog/why\_does\_correlation\_not\_equal\_causation/