Which independent categorical variables affect the churn rate of customers?

By: Kevin Sandoval

A1.

For the performance assessment of D207 Exploratory Data Analysis, the question I will be answering is “Which independent categorical variables affect the churn rate of customers?”

A2.

Stakeholders could benefit from this analysis because churn rate is a crucial metric when determining business decisions. For example, according to this article by Forbes in 2022, “acquiring a new customer can cost five to seven times more than retaining an old one.” If a company or business can pinpoint why some customers might be leaving or what makes customers stay for longer, it can have lasting effects towards the success of the company.

A3.

I decided to clean the data set by dropping the columns that contained categorical variables that had a cardinality greater than 8. I settled on this number as the survey questions had 8 responses and I wanted to keep those in the data set. After dropping those columns and checking for any missing values, the shape of my data set was 10,000 rows and 28 columns. The independent columns are made up 26 categorical variables and 1 variable that is there to just preserve the original order of the data. The dependent variable is the column “Churn” which is what I will be focusing on. The columns with their respective counts of unique values are as follows.

**Categorical Variables:** Area (3), Marital (5), Gender (3), Techie (2), Contract (3), Port\_modem (2), Tablet (2), InternetService (3), Phone (2), Multiple (2), OnlineSecurity (2), OnlineBackup (2), DeviceProtection (2), TechSupport (2), StreamingTV (2), StreamingMovies (2), PaperlessBilling (2), PaymentMethod (4), Item1 (7), Item2 (7), Item3 (8), Item4 (7), Item5 (7), Item6 (8), Item7 (7), and Item8 (8)

**Miscellaneous:** CaseOrder (10,000)

Since I have cleaned the data set to just include the categorical variables, I will just be using the chi-squared technique going forward.

B1.

I’m going to write a for loop that will look at all the categorical variables in relation to our dependent variable, Churn. For every categorical variable that only has two unique values, the result will be telling enough. For the variables that have more values, such as Items 1-8, I will perform ad-hoc testing to check the significance of each category. With Chi-Square test, it is important to note that it uses a Chi-Square Table to show the critical values of the chi squared distribution, observations are independent, random samples are required, and that it is a hypothesis test comparing two or more proportions: H0 : P1 = P2. (Sewell, D207 Ep 5). In order to determine if we will reject the null hypothesis, I will be comparing the output I get to the table provided in Lecture 5 and as shown below. If the test statistic is greater than the corresponding number for the p-value of 0.05 and the respective degrees of freedom, we will reject the null. Checking if two variables are NOT independent.

A table with numbers and a few digits

Description automatically generated with medium confidence

B2.

The results of the Chi-Square tests are shown below. The results include the Chi-Squared statistic, the p-value, the degrees of freedom, and whether or not to reject H0. For categorical variables that had more than 2 unique values in the column, I ran an extra function that calculated the post hoc testing to check each response versus the churn.

**Techie:**

Statistic: 44.11479

p-value: 3.09672e-11

Degrees of Freedom: 1

Result: Reject H0, there is a statistically significant relationship between “Churn” and “Techie”

**Port\_modem:**

Statistic: 0.62891

p-value: 0.42776

Degrees of Freedom: 1

Result: Fail to reject H0. There does not appear to be a statistically significant relationship between “Churn” and “Port\_modem”

**Tablet:**

Statistic: 0.06407

p-value: 0.80017

Degrees of Freedom: 1

Result: Fail to reject H0. There does not appear to be a statistically significant relationship between “Churn” and “Tablet”

**Phone:**

Statistic: 6.71174

p-value: 0.00958

Degrees of Freedom: 1

Result: Reject H0, there is a statistically significant relationship between “Churn” and “Phone”

**Multiple:**

Statistic: 173.03799

p-value: 1.60573e-39

Degrees of Freedom: 1

Result: Reject H0, there is a statistically significant relationship between “Churn” and “Multiple”

**OnlineSecurity:**

Statistic: 1.76975

p-value: 0.18341

Degrees of Freedom: 1

Result: Fail to reject H0. There does not appear to be a statistically significant relationship between “Churn” and “OnlineSecurity”

**OnlineBackup:**

Statistic: 25.28155

p-value: 4.95424e-07

Degrees of Freedom: 1

Result: Reject H0, there is a statistically significant relationship between “Churn” and “OnlineBackup”

**DeviceProtection:**

Statistic: 31.65320

p-value: 1.84310e-08

Degrees of Freedom: 1

Result: Reject H0, there is a statistically significant relationship between “Churn” and “DeviceProtection”

**TechSupport:**

Statistic: 3.46122

p-value: 0.062824

Degrees of Freedom: 1

Result: Fail to reject H0. There does not appear to be a statistically significant relationship between “Churn” and “TechSupport”

**StreamingTV:**

Statistic: 528.65186

p-value: 5.54938e-117

Degrees of Freedom: 1

Result: Reject H0, there is a statistically significant relationship between “Churn” and “StreamingTV”

**StreamingMovies:**

Statistic: 835.41401

p-value: 1.07801e-183

Degrees of Freedom: 1

Result: Reject H0, there is a statistically significant relationship between “Churn” and “StreamingMovies”

**PaperlessBilling:**

Statistic: 0.46239

p-value: 0.49651

Degrees of Freedom: 1

Result: Fail to reject H0. There does not appear to be a statistically significant relationship between “Churn” and “PaperlessBilling”

Those are the results for every categorical variable that only has two unique values in their columns. Below are the results for the other categorical variables. In order to save space since there are many to check, each value that is compared to “Churn” will have its respective statistic, p-value, degrees of freedom, and result listed in a line in that order.

**Area:**

**Rural:** 0.76223 (Statistic), 0.38263 (p-value), 1 (DoF), Fail to reject (Result)

**Suburban:** 0.40118, 0.52648, 1, Fail to reject

**Urban:** 2.34518, 0.12567, 1, Fail to reject

**Marital:**

**Widowed:** 1.06912, 0.301144, 1, Fail to reject

**Married:** 0.55299, 0.45669, 1, Fail to reject

**Separated:** 2.12315, 0.14509, 1, Fail to reject

**Never Married:** 2.920552, 0.088277, 1, Fail to reject

**Divorced:** 0.002403, 0.9609, 1, Fail to reject

**Gender:**

**Male:** 7.74737, 0.00538, 1, Reject H0

**Female:** 7.17926, 0.00737, 1, Reject H0

**Nonbinary:** 0.06692, 0.79588, 1, Fail to reject

**Contract:**

**One year:** 192.55841, 8.78832e-44, 1, Reject H0

**Month-to-month:** 715.16518, 1.50655e-157, 1, Reject H0

**Two year:** 317.10004, 6.20275e-71, 1, Reject H0

**InternetService:**

**Fiber Optic:** 33.923099, 5.733395e-09, 1, Reject H0

**DSL:** 86.95276, 1.11145e-20, 1, Reject H0

**None:** 14.03608, 0.000179, 1, Reject H0

**PaymentMethod:**

**Credit Card (automatic):** 0.41137, 0.52127 1, Fail to reject

**Bank Transfer (automatic):** 3.07034, 0.07973, 1, Fail to reject

**Mailed Check:** 0.87534, 0.349482, 1, Fail to reject

**Electronic Check:** 8.80585, 0.003003, 1, Reject H0

**Item1:**

**Response 2:** 0.87533, 0.34948, 1, Fail to reject  
**Response 1:** 0.88388, 0.34714, 1, Fail to reject  
**Response 2:** 1.95282, 0.16228, 1, Fail to reject  
**Response 3:** 1.68384, 0.19441, 1, Fail to reject  
**Response 4:** 0.04395, 0.83392, 1, Fail to reject  
**Response 5:** 0.08330, 0.77287, 1, Fail to reject  
**Response 6:** 0, 1, 1, Fail to reject  
**Response 7:** 0.07749, 0.78072, 1, Fail to reject \*One part of contingency table has less than 5 frequency 0.08330, 0.77287

**Item2:**

**Response 1:** 6.0461e-07, 0.99938, 1, Fail to reject  
**Response 2:** 2.97642, 0.08448, 1, Fail to reject  
**Response 3:** 0.16398, 0.68551, 1, Fail to reject  
**Response 4:** 0.00023, 0.98779, 1, Fail to reject  
**Response 5:** 1.57286, 0.20979, 1, Fail to reject  
**Response 6:** 0.15559, 0.69324, 1, Fail to reject  
**Response 7:** 0.35314, 0.55233, 1, Fail to reject \*One part of contingency table has less than 5 frequency

**Item3:**

**Response 1:** 0.10067, 0.75102, 1, Fail to reject  
**Response 2:** 3.56985, 0.058837, 1, Fail to reject  
**Response 3:** 0.49700, 0.48081, 1, Fail to reject  
**Response 4:** 0.92763, 0.335478, 1, Fail to reject  
**Response 5:** 0.09339, 0.759908, 1, Fail to reject  
**Response 6:** 0.0, 1.0, 1, Fail to reject  
**Response 7:** 0.19807, 0.65628, 1, Fail to reject \*One part of contingency table has less than 5 frequency  
**Response 8:** 0.0, 1.0, 1, Fail to reject \*Two parts of contingency table have less than 5 frequency

**Item4:**

**Response 1:** 0.0, 1.0, 1, Fail to reject  
**Response 2:** 0.01346, 0.90762, 1, Fail to reject  
**Response 3:** 0.01367, 0.90690, 1, Fail to reject  
**Response 4:** 0.42755, 0.51318, 1, Fail to reject  
**Response 5:** 1.17559, 0.27825, 1, Fail to reject  
**Response 6:** 0.35436, 0.55165, 1, Fail to reject  
**Response 7:** 0.44720, 0.50366, 1, Fail to reject \*One part of contingency table has less than 5 frequency

**Item5:**

**Response 1:** 1.21505, 0.27033, 1, Fail to reject  
**Response 2:** 0.04791, 0.82672, 1, Fail to reject  
**Response 3:** 1.20722, 0.27188, 1, Fail to reject  
**Response 4:** 1.92018, 0.165835, 1, Fail to reject  
**Response 5:** 0.02946, 0.86371, 1, Fail to reject  
**Response 6:** 0.0, 1.0, 1, Fail to reject  
**Response 7:** 1.20899, 0.27153, 1, Fail to reject \*One part of contingency table has less than 5 frequency

**Item6:**

**Response 1:** 0.27331, 0.60111, 1, Fail to reject  
**Response 2:** 0.22640, 0.63419, 1, Fail to reject  
**Response 3:** 1.35635, 0.24416, 1, Fail to reject  
**Response 4:** 0.40593, 0.52403, 1, Fail to reject  
**Response 5:** 0.25757, 0.61178, 1, Fail to reject  
**Response 6:** 1.53887, 0.21478, 1, Fail to reject  
**Response 7:** 0.0, 1.0, 1, Fail to reject \*One part of contingency table has less than 5 frequency  
**Response 8:** 0.0, 1.0, 1, Fail to reject \*Two parts of contingency table have less than 5 frequency

**Item7:**

**Response 1:** 0.0, 1.0, 1, Fail to reject  
**Response 2:** 1.10037, 0.29418, 1, Fail to reject  
**Response 3:** 0.60471, 0.43678, 1, Fail to reject  
**Response 4:** 4.46314, 0.03463, 1, Reject H0  
**Response 5:** 0.83606, 0.36052, 1, Fail to reject  
**Response 6:** 0.01734, 0.89523, 1, Fail to reject  
**Response 7:** 0.08047, 0.77665, 1, Fail to reject \*One part of contingency table has less than 5 frequency

**Item8:**

**Response 1:** 0.02107, 0.88458, 1, Fail to reject  
**Response 2:** 0.00047, 0.98269, 1, Fail to reject  
**Response 3:** 0.36388, 0.54635, 1, Fail to reject  
**Response 4:** 0.00514, 0.94280, 1, Fail to reject  
**Response 5:** 0.20072, 0.65413, 1, Fail to reject  
**Response 6:** 0.68474, 0.40795, 1, Fail to reject  
**Response 7:** 0.0, 1.0, 1, Fail to reject \*One part of contingency table has less than 5 frequency  
**Response 8:** 0.0, 1.0, 1, Fail to reject \*Two parts of contingency table have less than 5 frequency

C.

**1. Univariate Graph of Age**

**A blue graph with white text

Description automatically generated** **A blue rectangular object with black lines

Description automatically generated**

We can see from the histogram here that the continuous variable “Age” has a uniform distribution. We can also see from the boxplot that there are no outliers for “Age.” The minimum value is 18 years old, the max value is 89 years old. The 25th percentile is 35 years old and the 75th percentile is 71 years old.

**2. Univariate Graph of MonthlyCharge**

**A blue graph with numbers

Description automatically generatedA blue rectangle with black lines

Description automatically generated**

We can see from the histogram here that the continuous variable “MonthlyCharge” has a normaldistribution. We can also see from the boxplot that there are no outliers for “MonthlyCharge.” The minimum value is $79.98, the max value is $290.16 The 25th percentile is $139.98 and the 75th percentile is $200.73.

**3. Univariate Graph of Techie**

**A blue rectangular column with black text

Description automatically generated**

From this bar graph we can see the distribution of “No” responses and “Yes” responses. Clearly “No” is much more common of a response and that can be shown in the value counts. There are 8,321 counts of “No” and 1,679 counts of “Yes”

**4. Univariate Graph of Gender**

**A graph of a person and person

Description automatically generated**

From this bar graph we can see the distribution of “No” responses and “Yes” responses. We can see that “Female” is the most common response with “Male” close behind while “Nonbinary” has only a few values. That can be shown in the value counts. There are 5,025 counts of “Female”, 4,744 counts of “Male”, and 231 counts of “Nonbinary.”

D.

**1. Bivariate Graph of Two Continuous Variables**

**A blue and white dotted chart

Description automatically generated**

From this scatterplot, we can see that there does not seem to be any correlation between “Age” and “MonthlyCharge.”

**2. Bivariate Graph of Two Categorical Variables**

**A graph of a bar chart

Description automatically generated with medium confidence**

From this stacked bar chart, we can examine the amount of people by gender who consider themselves to be technically inclined. Since the majority of responses to “Techie” were “No”, it is easily seen here that that is the most common response per gender. We can also see a relatively even distribution of “Yes” and “No” across all genders. For female, we have 868 counts of “Yes” to “Techie” and 4,157 counts of “No.” For male, we have 778 counts of “Yes” and 3,966 counts of “No.” For nonbinary we have 33 counts of “Yes” and 198 counts of “No.” That gives us a respective percentage of 82.73% “No” answers for Females, 83.60% “No” answers for Males, and 85.7% “No” answers for nonbinary people. Thus we can see that the range of answers is roughly similar for each gender.

E1.

My hypothesis test for the categorical variables was that the two variables being checked were independent. If the p-value was less than the alpha then I could reject and say that there is a statisically signficant relationship between the two variables. Looking at my results for variables with only 2 unique values, I found that "**Techie**", "**Phone**", "**Multiple**", "**OnlineBackup**", "**DeviceProtection**", "**StreamingTV**", and "**StreamingMovies**" all had a statistically signifcant relationship with "**Churn**". "**Port**\_**modem**", "**Tablet**", "**OnlineSecurity**", "**TechSupport**", and "**PaperlessBilling**" all had 2 unique values but did not appear to have a statistically signficant relationship with "**Churn**". Moving onto the variables that had to have post hoc, I found that "Area" had no values that were signficant with "**Churn**". "**Marital**" also had no values. "**Gender**" had "Male" and "Female" as values that had a relationship with "**Churn**". All types of "**Contract**" were signficant as well as "**InternetService**". "**PaymentMethod**" only had "Electronic Check" has signficant with "**Churn**". For **Item1-8**, only 1 response question had a signficant relationship with "Churn". That was **Item7** Response 4 that had a low enough p-value to reject the H0.

E2.

One clear limitation of my analysis is that I only looked at categorical variables. I thought that it would be interesting to compare a customer’s responses to the rate as opposed to continuous data. This leaves out some potentially useful information but I can always go back and check that data as well. I decided to use a cardinality of less than 9 in order to get the Item1-8 to be included in the data. After looking around and seeing the results it would seem that keeping the cardinality low at 3-5 would be better used. There was almost no significant findings in those variables and it made the report much longer and daunting. Another limitation could be using post hoc analysis on the variables that had more than 2 unique values. It can easily be bogged down and get confusing to try and interpret the data. I think the last limitation could be the results of the significance. Although I had some data show as statistically significant in their relationship to “Churn,” deciding what that means can be difficult with categorical variables.

E3.

For a course of action for further analysis based on my exploratory results here, I would focus on the relationship between “Churn” and “Techie”, “Contract”, and “StreamingTV”/”StreamingMovies.” These variables have showed a very strong relationship with “Churn” and looking closer could help explain why. I would hypothesis that people who don’t consider themselves technically inclined (“Techie” column) might be less likely to cancel than people who feel like they know what they are doing and can easily switch services. “Contract” is also interesting but maybe more explainable. If someone is locked into a contract then the likelihood of them churning is presumably very low. Differing contract lengths could help retain customers. As for the “StreamingTV” and “StreamingMovies” variables, they are similar data but both with a strong relationship with “Churn.” I would hypothesis that people who want to keep watching their TV channels, shows, or movies would be less likely to churn. I think people who don’t tend to stream TV or movies feel like they aren’t missing out on anything when they decide to leave, so nothing is preventing them from leaving. There are a few other significant relationships in the data that could be explored further, such as “InternetService,” however I think those relationships can be explained more readily and might not have as interesting relationships with people who decide to churn.

F.

See attached video.

G/H.

Board Infinity. (n.d). *How to Get Column Names in Pandas.* Retrieved July 24th, 2024,From <https://www.boardinfinity.com/blog/how-to-get-column-names-in-pandas/#:~:text=pandas%20Get%20Column%20Names,using%20the%20print()%20statement.>

Bowne-Anderson, H. (n.d). *D206 Data Cleaning.* Introduction to Importing Data in Python. Retrieved June 30th, 2024 From <https://app.datacamp.com/learn/custom-tracks/custom-d206-data-cleaning>

Dr. Sewell, W (n.d). *D207 Episode 1.* Retrieved July 27th, 2024,From D206 Announcements

Dr. Sewell, W (n.d). *D207 Episode 2.* Retrieved July 27th, 2024,From D206 Announcements

Dr. Sewell, W (n.d). *D207 Episode 3.* Retrieved July 27th, 2024,From D206 Announcements

Dr. Sewell, W (n.d). *D207 Episode 4.* Retrieved July 27th, 2024,From D206 Announcements

Dr. Sewell, W (n.d). *D207 Episode 5.* Retrieved July 27th, 2024,From D206 Announcements

Kosourova, E. (May 2023). *How to Use SQL in pandas Using pandasql Queries.* Retrieved July 27th, 2024,From <https://www.datacamp.com/tutorial/how-to-use-sql-in-pandas-using-pandasql-queries>

Python For Data Science. (n.d). *Chi-square.* Retrieved July 24th, 2024,From <https://pythonfordatascienceorg.wordpress.com/chi-square-python/>

Stack Overflow. (n.d). *Plotting categorical variables as stacked bar plot.* Retrieved July 27th, 2024,From <https://stackoverflow.com/questions/55151699/plotting-categorical-variable-as-stacked-bar-plot>