Western Governors University

EXPLORATORY DATA ANALYSIS

D207

Allison Casey
3-7-2024

A1: QUESTION FOR ANALYSIS
Is there a significant relationship between the number of emails sent to the customer and whether a customer churned?

A2: BENEFIT FROM ANALYSIS
This analysis could be useful to stakeholders for a few reasons. Gaining a better understanding of the relationship between email correspondence and churn can help the company make decisions about their communication with customers to ideally decrease churn. If there is a significant relationship between these two things, then they can also ask more specific questions about the email correspondence to further analyze and evaluate if and what changes need to be made to prevent churn. Having these insights are important for the company to be able to find the best strategies to retain customers and save money.

A3: DATA IDENTIFICATION
The data needed to answer the question comes from the Churn column and the Outage_sec_perweek column in the dataset.

B1: CODE
D207.ipynb

B2: OUTPUT
Summary statistics of the emails sent for customers that churned:

```
count    2650.000000
mean       12.078113
std         3.008534
min         2.000000
25%        10.000000
50%        12.000000
75%        14.000000
max        23.000000
Name: Email, dtype: float64
```

Summary statistics of the emails sent for customers that did not churn:

```
count    7350.000000
mean       11.993605
std         3.032026
min         1.000000
25%        10.000000
50%        12.000000
75%        14.000000
max        22.000000
Name: Email, dtype: float64
```

T-test result:

```
TtestResult(statistic=1.2325944156039008, pvalue=0.217756104502
05743, df=9998.0)
```

B3: JUSTIFICATION

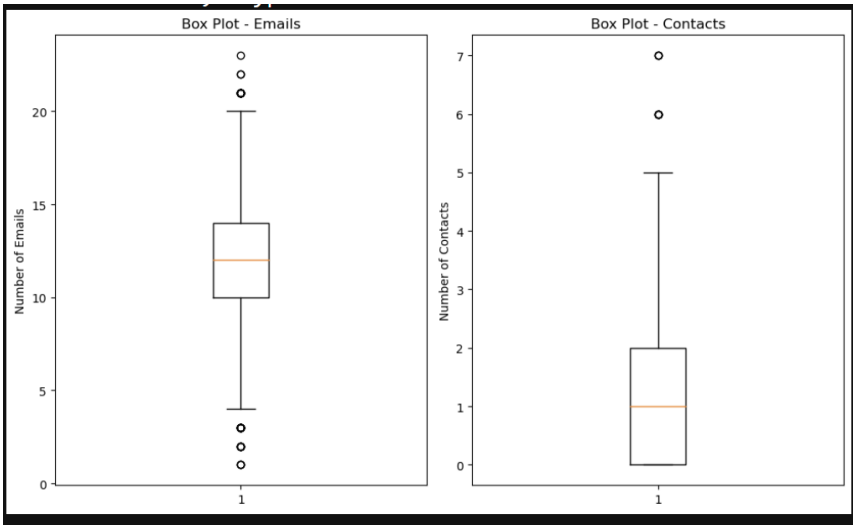
A T-test was used to analyze the data because it can be used to tell if there is a significant difference between two groups. In this case the two groups being looked at were the amount emails sent for customers who churned and those who did not. The p-value generated from the T-test can then be used to determine if there is a significant relationship between the two groups which would answer the question for analysis.

C: UNIVARIATE STATISTICS

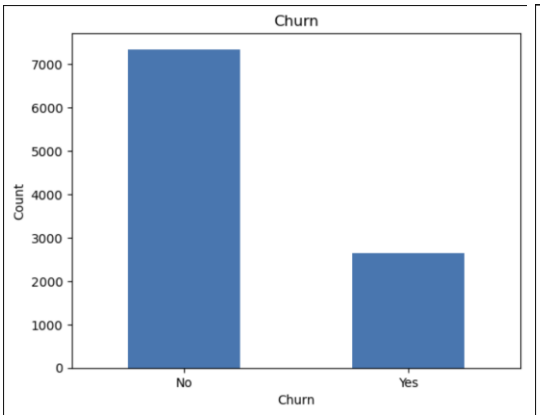
The two continuous variables looked at were Emails and Contacts. The Emails had equal proportions around the median with some outliers indicating this has a relatively normal distribution. However, the Contacts had a median closer to the bottom quartile indicating the data is skewed towards less contacts. This can be corroborated numerically by looking at the summary statistics that were included in the code. The two categorical variables used were Churn and Courteous Exchange. The Churn graph showed that more didn't churn than did and the Courteous Exchange graph showed that most of the ratings were a 3 or 4.

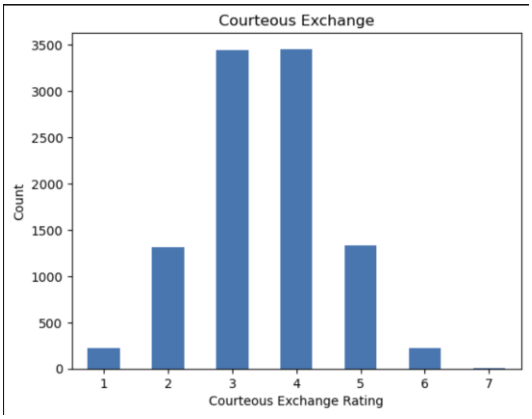
C1:VISUAL OF FINDINGS

Continuous Variables



Categorical Variables



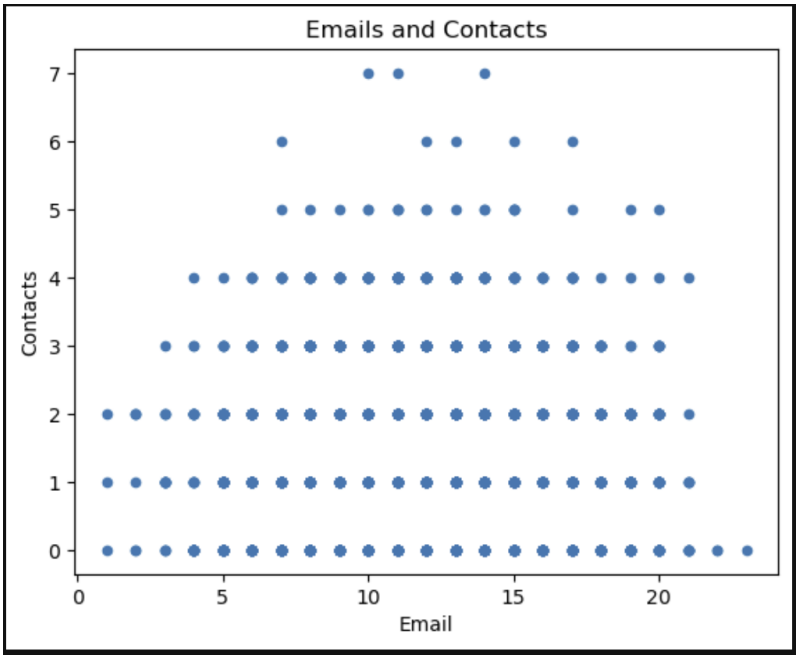


D: BIVARIATE STATISTICS

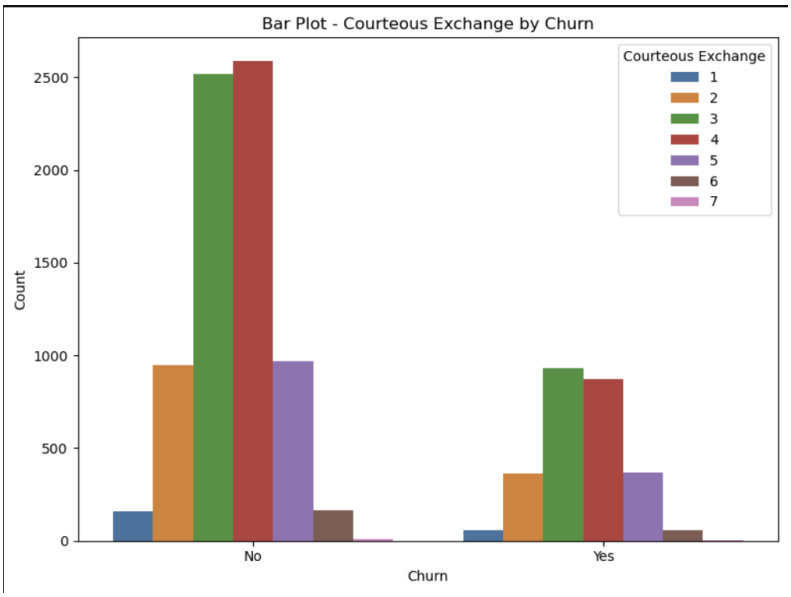
To identify the distribution of two continuous variables using bivariate statistics a scatter plot was made using the Emails and Contacts. This showed the number of Contacts based on the number of Emails but didn't show an obvious trend between the two visually. For the categorical variables Courteous Exchange was graphed based on the Churn and showed that the distribution was similar for those that churned versus those that did not.

D1:VISUAL OF FINDINGS

Continuous Variables



Categorical Variables



E1: RESULTS OF ANALYSIS

The null hypothesis would be that there is no significant relationship between the Emails and Churn and the alternative hypothesis would be that there is. The T-test returned a p-value of approximately 0.22 and using a standard alpha value of .05 so the p-value is greater than the alpha value which indicates that the null hypothesis can be accepted. This answers the question for analysis by confirming that there is poor evidence that there is a significant relationship between Emails and Churn.

E2: LIMITATIONS OF ANALYSIS

This analysis only covers a limited scope of the data so there are quite a few limitations to it. For starters a T-test only compares two groups and there are a lot of other variables in the data set that could be confounding or have relationships to the number of emails that get sent that if included in analysis could potentially yield different findings. Another limitation is that the number of emails is very broad and there may be more useful relationships to explore with more specific subsets such as marketing emails. For the sake of the question for analysis the T-test was sufficient though there are a lot of other questions that could be asked involving more variables and exploring more of the data.

E3: RECOMMENDED COURSE OF ACTION

The results of the analysis answered the question for analysis concluding that there is poor evidence that there is a significant relationship between Emails and Churn. As a result, the recommended action would be to focus on exploring the data further to figure out things that had a greater impact on churn to find things to improve to prevent churn. This doesn't entirely mean that Email should be overlooked since it could be a confounding variable for other columns of data or be relevant to other questions. What it does mean, however, is that they can reasonably not worry about changing the number of Email that get sent out and focus on other things.

F:VIDEO

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=8ba0f711-a874-42d6-b575-b12d0023bccc

G:SOURCES FOR THIRD-PARTY CODE

Hunter, John, et al. "Matplotlib.Pyplot#." *Matplotlib.Pyplot - Matplotlib 3.8.3 Documentation*, matplotlib.org/stable/api/pyplot_summary.html. Accessed 7 Mar. 2024.

"Matplotlib.Pyplot.Subplot#." *Matplotlib.Pyplot.Subplot - Matplotlib 3.8.3 Documentation*, matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.subplot.html#matplotlib.pyplot.subplot. Accessed 7 Mar. 2024.

H:SOURCES

Hunter, John, et al. "Matplotlib.Pyplot#." *Matplotlib.Pyplot - Matplotlib 3.8.3 Documentation*, matplotlib.org/stable/api/pyplot_summary.html. Accessed 7 Mar. 2024.

"Introduction to Hypothesis Testing." *Numeracy, Maths and Statistics - Academic Skills Kit*, www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/psychology/introduction-to-hypothesis-

testing.html#:~:text=The%20null%20hypothesis%20is%20the,to%20reject%20the%20null%20hypothesis. Accessed 7 Mar. 2024.

“Matplotlib.Pyplot.Subplot#.” *Matplotlib.Pyplot.Subplot - Matplotlib 3.8.3 Documentation*, matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.subplot.html#matplotlib.pyplot.subplot. Accessed 7 Mar. 2024.

“The T-Test.” *JMP*, www.jmp.com/en_us/statistics-knowledge-portal/t-test.html#:~:text=A%20t%2Dtest%20may%20be,dependent%20samples%20t%2Dtest). Accessed 7 Mar. 2024.