D207 Performance Assessment

Hillary Osei (Student ID #011039266)

Western Governors University, College of Information Technology

Program Mentor: Dan Estes

November 2, 2024

Table of Contents

Section A1) Research Question

The research question addressed is "What is the relationship between contract type and churn rate?" The churn dataset is used to answer this question, and the chi-square test is used to analyze the cleaned data.

Section A2) Benefit from Analysis

Addressing this research question will help stakeholders understand the impact various contract types might have on churn rates. The marketing team, for example, could use this insight to create targeted marketing campaigns to help encourage customers with shorter-term contracts to convert to longer-term contracts. The sales team can also provide incentives such as promotions, discounts, and exclusive bundles to upsell longer-term contracts to reduce churn.

Section A3) Data Identification

To answer the research question, the categorical variables “Contract” (indicates the contract term of the customer, either month-to-month, one year, or two year) and “Churn” (indicates whether a customer has discontinued their service within the last month, yes or no) are needed (Western Governors University).

Section B1) Code

```
# Create cross-tabulation table
# Code based on GeeksforGeeks. (n.d.). Pandas crosstab() function in Python.
# Retrieved from https://www.geeksforgeeks.org/pandas-crosstab-function-in-python/
contingency = pd.crosstab(df['Contract'], df['Churn'])
print(contingency)

# Perform chi-square test
# Code based on Dhunna, A. (2020, October 20). How to run Chi-Square Test in Python.
# Retrieved from https://medium.com/swlh/how-to-run-chi-square-test-in-python-4e9f5d10249d
chi2_stat, p_val, dof, expected = chi2_contingency(contingency)
# Print results
print(f"Chi-Square Statistic: {chi2_stat}")
print(f"P-value: {p_val:.5f}")
print(f"Expected Frequency: {expected}")
```

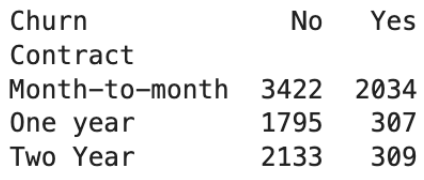
```
# Visualize Contract vs Churn

# Create stacked bar chart
contingency.plot(kind='bar', stacked=True, figsize=(8, 6), color=['#ff9999','#66b3ff'])

# Add labels and title
plt.title('Contract Type vs Churn')
plt.xlabel('Contract Type')
plt.ylabel('Number of Customers')
plt.legend(title='Churn')
plt.show()
```

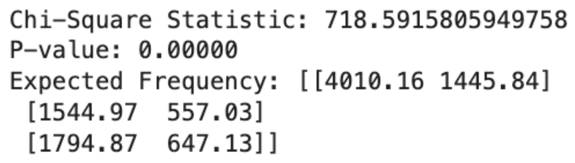
Section B2) Output

Cross-tabulation calculation results:



```
Churn            No   Yes
Contract
Month-to-month  3422  2034
One year        1795   307
Two Year        2133   309
```

Chi-square test results:



```
Chi-Square Statistic: 718.5915805949758
P-value: 0.00000
Expected Frequency: [[4010.16 1445.84]
 [1544.97  557.03]
 [1794.87  647.13]]
```

Section B3) Justification

The chi-square test of independence was used to address the research question, “What is the relationship between contract type and churn rate?” because it is suitable for determining the relationship between categorical variables (Turney, 2022). The categorical variables in this case are Contract (month-to-month, one year, two year) and Churn (yes/no).

The chi-square test of independence tests the following hypotheses:

- Null hypothesis (H_0): There is no significant difference between contract type and churn rate, meaning that contract type does not influence whether a customer churns.

- Alternative hypothesis (H_1): There is a significant difference between contract type and churn rate, meaning that contract type does impact whether a customer churns.

The two variables meet the following assumptions for the chi-square test of independence (*Chi-Square Test of Independence*, n.d.):

1. Each variable must be independent, meaning that one variable does not have an influence on the other. This assumption is met because each customer's contract type and churn status have no effect on each other.
2. The categories must be mutually exclusive, meaning that each customer can belong to only one category per variable. This assumption is met because each customer is assigned to only one contract type and only one churn status.
3. There are at least five expected observations in each cell. To verify this, a contingency table was created using the variables ‘Contract’ and “Churn.’ Then, the ‘chi2_contingency’ function was used to check the expected frequencies. The results provided six expected values: [[4010.16 1445.84] [1544.97 557.03] [1794.87 647.13]]. Since all the expected values are above 5, this assumption is met.

Section C) Univariate Statistics

Univariate statistics are used to describe and identify patterns within data for a single variable by examining frequency distributions, measures of central tendency, variability, and distribution shape (Tate, 2023). This analysis explored the continuous variables "Income" and "MonthlyCharge" using univariate techniques.

For "Income," the describe() function provided key summary statistics. The dataset contains 10,000 observations, with an average income of 39,806.93 and a standard deviation of 28,199.92, meaning high variability around the mean (Standard Deviation Formula and Uses Vs. Variance, n.d.). "Income" values range from a minimum of 348.67 to a maximum of 258,900.70, suggesting a wide spread in customer income levels. The median income is 33,170.61, lower than the mean, indicating a right-skewed distribution likely influenced by some high-income values. The 25th and 75th percentiles are 19,224.72 and 53,246.17, respectively, showing that 50% of customers fall within this income range.

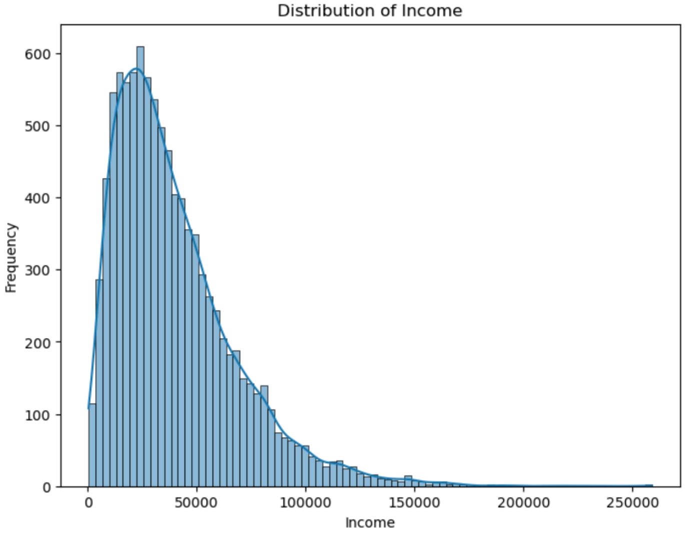
For “MonthlyCharge,” the describe() function revealed an average charge of 172.62 with a standard deviation of 42.94, reflecting moderate variability. Monthly charges range from 79.98 to 290.16, indicating a broad range of charges. The median monthly charge is 167.48, close to the mean, suggesting a relatively symmetric distribution. The 25th and 75th percentiles are 139.98 and 200.73, respectively, meaning that half of the customers are charged within this range.

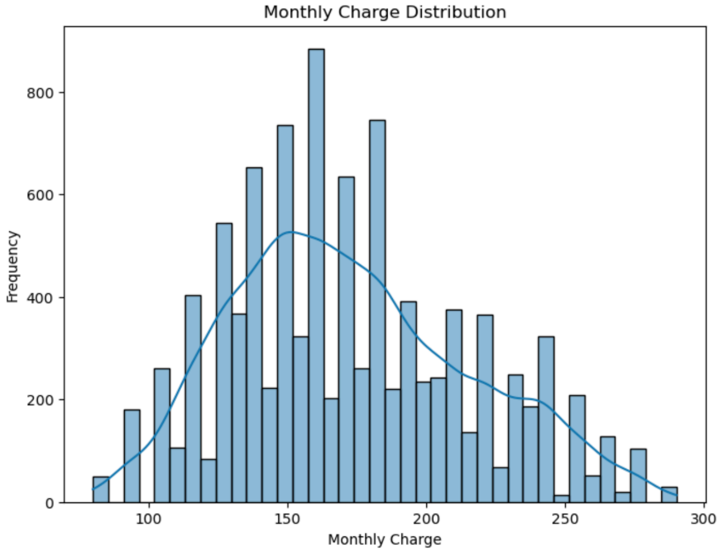
Next, histograms were generated using the plt.histplot() function to visualize the distributions of "Income" and "MonthlyCharge." Each histogram included a Kernel Density Estimate (KDE) line to provide a clearer view of distribution patterns. The visualizations confirmed that Income has a right-skewed distribution, while MonthlyCharge shows a slight right-skew.

The categorical variables "Gender" and "Techie" were also analyzed. The value_counts() function was used to calculate the unique value counts for each category. The "Gender" variable included 5,025 female customers, 4,744 male customers, and 231 nonbinary customers. The "Techie" variable showed 8,321 non-tech-savvy customers and 1,679 tech-savvy customers. Histograms were then created using the plt.hist() function to visualize these distributions. The analysis revealed that the "Gender" variable was skewed toward female customers, while the "Techie" variable was skewed toward customers identifying as less technically inclined.

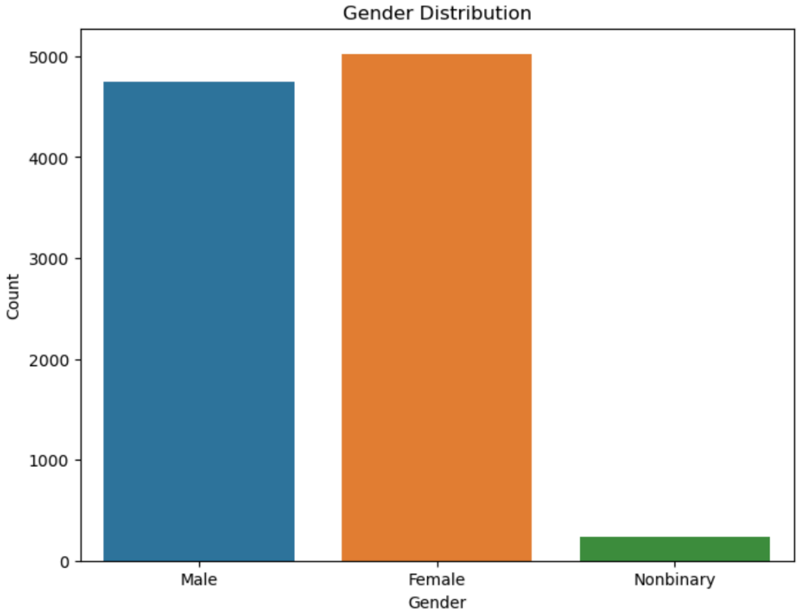
Section C1) Univariate Statistics Findings

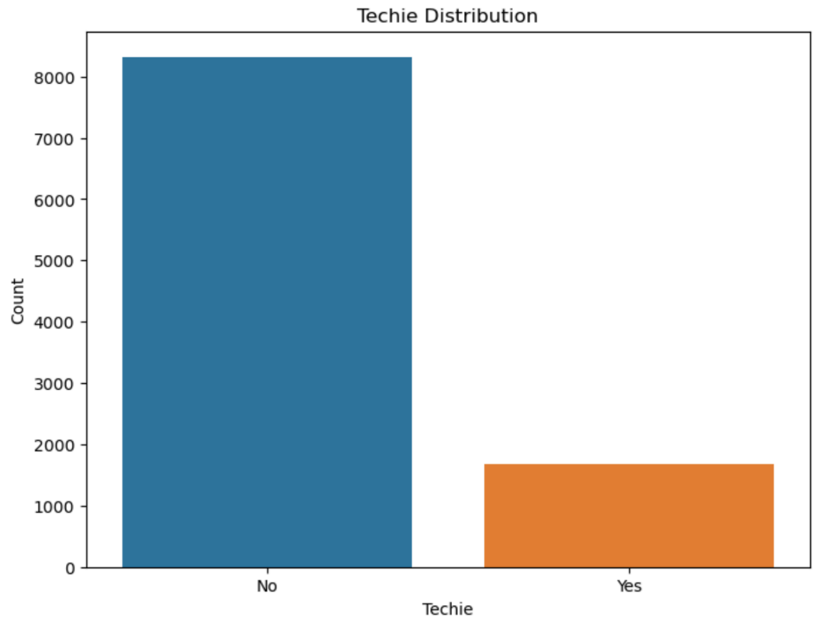
Continuous variables:





Categorical variables:





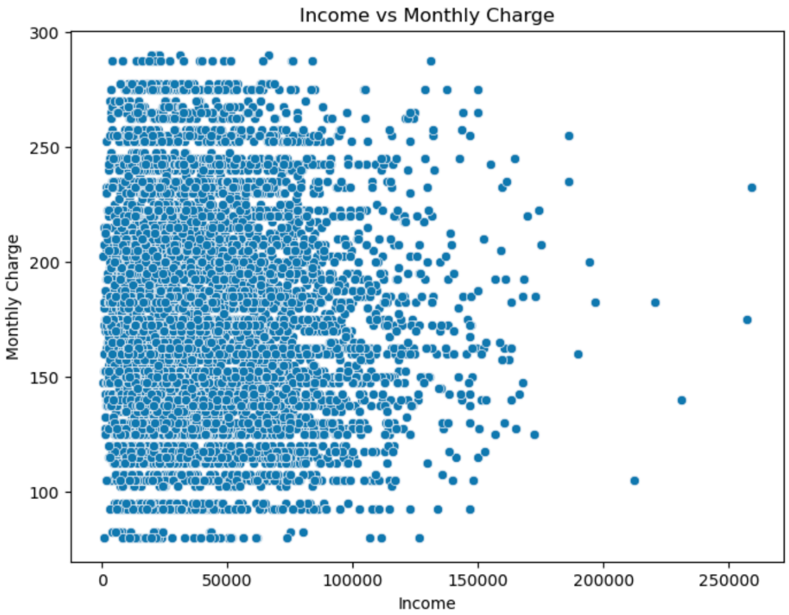
Section D) Bivariate Statistics

Bivariate statistics analyze the relationship between two variables (*Bivariate Analysis: What Is It, Types + Examples, n.d.*). For this analysis, the continuous variables "Income" and "MonthlyCharge" were examined to determine if a correlation exists between a customer's income and the average monthly amount they are charged for using the service. A Pearson correlation test was conducted using the pearsonr() function, resulting in a correlation coefficient of -0.003. This value is close to zero, indicating no significant correlation between "Income" and "MonthlyCharge." (Bobbitt, 2022) A scatterplot was also generated to visually assess this relationship, confirming that there is no association between the two variables.

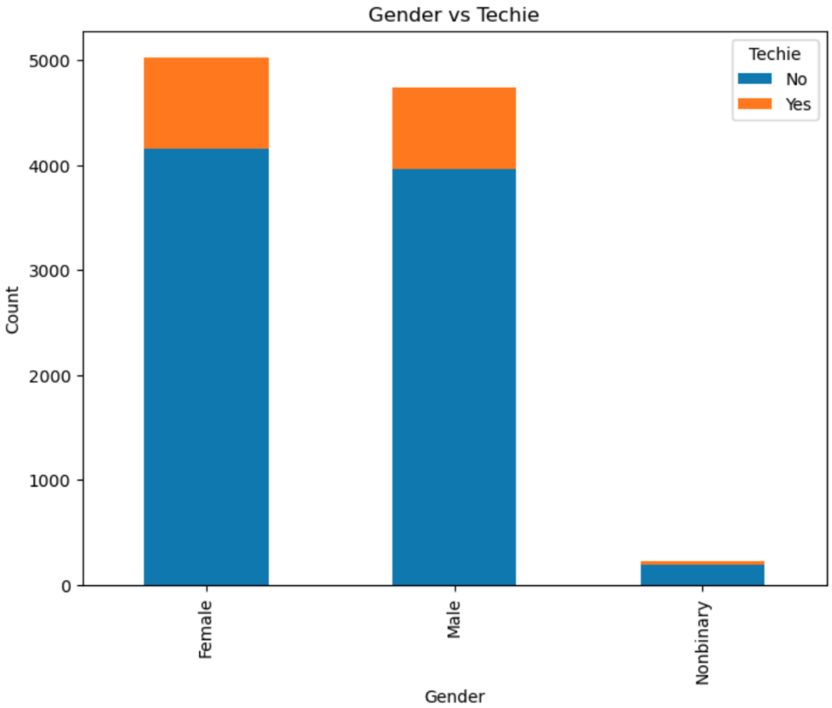
For the categorical variables, "Gender" and "Techie" were analyzed to determine if there is a relationship between gender and being technically inclined. A Chi-Square Test of Independence was performed on a contingency table of Gender and Techie categories. The Chi-Square statistic was 2.3955846292159864 with a p-value of 0.30185988850587486, indicating no significant association between "Gender" and "Techie." Additionally, a stacked bar chart was used to visualize the relationship, revealing that most female customers were not technically inclined. More male customers identified as technically inclined compared to female and non-binary customers, although this difference was not statistically significant. Non-binary customers mostly self-identified as not being technically inclined.

Section D1) Bivariate Statistics Findings

Continuous variables:



Categorical variables:



Section E1) Results of Analysis

The hypothesis test examined the relationship between contract type and churn rate. The test involved two hypotheses: the null hypothesis (H_0), which stated no significant difference between contract type and churn rate, and the alternative hypothesis (H_1), which stated a significant difference between the two variables.

Based on the contingency table, it was found that customers with month-to-month contracts have higher churn rates compared to customers with one year and two year contacts, meaning that customers with shorter contracts were most likely to cancel or not renew their service compared to customers with more extended contracts. Among the month-to-month contract customers, 2,034 churned out of 5,456, meaning a churn rate of ~37%. Among the one year contract customers, 307 out of 2,102 churned, meaning a churn rate of ~14.6%. Furthermore, among the two year contract customers, 309 out of 2442 churned, resulting in a churn rate of ~12.7%.

Using the Chi-square test of independence, it was determined that the p-value was below 0.05, specifically at 0.00000, meaning that there is a statistically significant difference between the variables. Therefore, the null hypothesis is rejected, indicating that contract type influences customer churn (Simplilearn, 2023).

Section E2) Limitations of Analysis

The data analysis provides valuable insights into the relationship between the "Contract" and "Churn" variables; however, several limitations must be considered. The analysis focuses solely on these two variables, excluding other potentially influential factors such as service outages, monthly fees, and customer satisfaction. These additional variables are necessary for the analysis to present an accurate picture of the factors driving customer churn.

Additionally, the dataset may contain biases if certain groups are overrepresented. For instance, since there are significantly more customers with month-to-month contracts than those with one year or two year contracts, the results may only partially and accurately represent the broader customer base.

Section E3) Recommended Course of Action

Based on these findings, it is recommended that the organization develop strategies to encourage customers on month-to-month contracts to switch to either one year or two year contracts because they are associated with lower churn rates. The marketing team could develop targeted campaigns focused on the benefits of longer-term contracts, offering exclusive discounts, bundle deals, and add-ons to newly enrolled long-term contract customers. The organization can also gather customer satisfaction and service usage data through frequent surveying. This would provide a fuller understanding of the factors affecting churn.

Section F) Panopto

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=288c346d-3a64-413e-a7d9-b21e002ba0d1

Section G) Third Party Source Code

Jain, S. (2024, July 11). *pandas.crosstab() function in Python*. GeeksforGeeks. Retrieved October 17, 2024, from https://www.geeksforgeeks.org/pandas-crosstab-function-in-python/

Jain, S. (2024, March 19). *Python - Pearson Correlation Test Between Two Variables*. GeeksforGeeks. Retrieved October 31, 2024, from https://www.geeksforgeeks.org/python-pearson-correlation-test-between-two-variables/

Pipis, G. (2020, October 24). *How to Run the Chi-Square Test in Python*. Medium. Retrieved October 31, 2024, from https://medium.com/swlh/how-to-run-chi-square-test-in-python-4e9f5d10249d

Section H) Sources

Bivariate Analysis: What is it, Types + Examples. (n.d.). QuestionPro. Retrieved November 2, 2024, from https://www.questionpro.com/blog/bivariate-analysis/

Bobbitt, Z. (2022, April 6). *How to Perform a Correlation Test in Python (With Example)*. Statology. Retrieved November 2, 2024, from https://www.statology.org/correlation-test-in-python/

Chi-Square Test of Independence. (n.d.). StatsTest.com. Retrieved November 2, 2024, from https://www.statstest.com/chi-square-test-of-independence/

Jain, S. (2024, March 19). *Python - Pearson Correlation Test Between Two Variables*. GeeksforGeeks. Retrieved October 31, 2024, from https://www.geeksforgeeks.org/python-pearson-correlation-test-between-two-variables/

Pipis, G. (2020, October 24). *How to Run the Chi-Square Test in Python*. Medium. Retrieved October 31, 2024, from https://medium.com/swlh/how-to-run-chi-square-test-in-python-4e9f5d10249d

Simplilearn. (2023, February 20). *What Is P-Value in Statistical Hypothesis?* Simplilearn.com. Retrieved November 2, 2024, from https://www.simplilearn.com/tutorials/statistics-tutorial/p-value-in-statistics-hypothesis

Standard Deviation Formula and Uses vs. Variance. (n.d.). Investopedia. Retrieved November 2, 2024, from https://www.investopedia.com/terms/s/standarddeviation.asp

Tate, A. (2023, September 29). *How To Use Univariate Analysis in Your Data Exploration*. Hex. Retrieved October 17, 2024, from https://hex.tech/blog/univariate-analysis-data-exploration/

Turney, S. (2022, May 23). *Chi-Square (X^2) Tests | Types, Formula & Examples*. Scribbr. Retrieved November 2, 2024, from https://www.scribbr.com/statistics/chi-square-tests/

Western Governors University. (n.d.). Churn Data Consideration and Dictionary.pdf