D207 – Exploratory Data Analysis

Performance Assessment

Bernard James Connelly III

Master of Science, Data Analytics, Western Governors University

Dr. William Sewell

December 09, 2024

D207 – Exploratory Data Analysis Performance Assessment

A. Description of a real-world organizational situation with the churn dataset

A1. Research Question

To a telecommunications company, customer retention is integral to the business's overall success. This is because a subscription-based service that maintains more customers yields more income. Still, it also costs significantly less money to keep a current customer than it does to replace them with a new one. As a result, the research question most pertinent to the company includes the variable customer churn as a part of the analysis. Additionally, since multiple complex technology services are offered, a technologically capable person's perception of those services could directly affect their interest in maintaining their contract with the company. As a result, the research question "Are customers who identify as being technologically capable more likely to remain with the telecommunications company or leave it?" was selected for this project. To identify a method to reduce customer churn, a chi-squared test will be conducted with the "Churn" and "Techie" variables to determine if there is a relevant relationship.

A2. Benefits of Analysis

As already noted, whether or not a customer leaves the company or stays has a direct profit effect on a company, and the costs of maintaining a customer are significantly lower than the overhead of acquiring new ones. Additionally, since the WGU Telecommunications company offers many services, if there is a significant connection between whether or not loyal customers prefer high-tech options, the company executives would know where to invest development money to increase retention. There are also clear benefits to targeting specific areas of the country whose local infrastructure or population leans more into the use of technology as opposed to using more minimalist or traditional demographics.

A3. Relevant Data and Variables

There are two primary variables to review in this analysis, “Churn” and “Techie.” Details of the variables are listed below:

name	environment data type	data type	example	Description / Notes
Churn	object	Qualitative Nominal (Boolean)	Yes	Service Details - Did the customer terminate service within one calendar month?
Techie	object	Qualitative Nominal (Boolean)	Yes	Customer Self-reported demographics - Is the customer technically inclined?

To analyze the relationship between the above variables, a chi-square test will be conducted to identify how closely these variables relate to one another to predict a relevant business relationship. Factors for the test are as follows:

H_0 = There is **not** a statistically significant relationship between customers self-identifying as “Techies” and customer churn

H_A = a statistically significant relationship exists between customers self-identifying as “Techies” and customer churn.

α = 0.05, a standard p-Value

B. Details of Statistical the Test

B1 & B2. Code for chi-square test and results of the calculations

Confirmation that the relevant fields were cleaned and completed:

```
print(df['Techie'].value_counts())

print ("\n")

print(df['Churn'].value_counts())
```

```
Techie
No     8321
Yes    1679
Name: count, dtype: int64


Churn
No     7350
Yes    2650
Name: count, dtype: int64
```

Creation of a contingency table to complete the analysis:

```
table = pd.crosstab(df.Churn, df.Techie)

print(table)
```

```
Techie    No   Yes
Churn
No      6226  1124
Yes     2095   555
```

Code and results of the chi-square test:

```
chi2, p, dof, expected = stats.chi2_contingency(table)

print(f"Chi-Squared statistic: {chi2}")

print(f"P-value: {p}")
```

```
Chi-Squared statistic: 44.11479393861451
P-value: 3.096716355509661e-11
```

B3. Justification of analysis

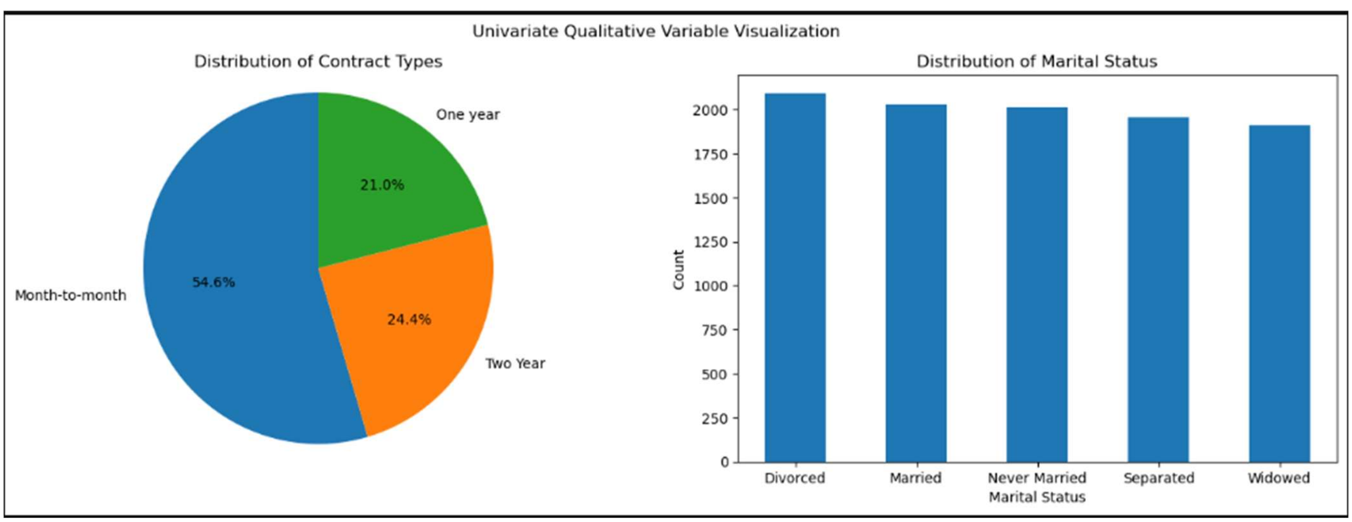
A chi-square test was utilized as it compares two categorical variables, so it was the most relevant test to perform on this type of data – a T-test or ANOVA is designed for numerical values and would not be appropriate for categorical data only. Additionally, after profiling the data, a normal distribution was not observed, which further justifies using a chi-square test over other alternatives. Finally, as the p-value identified was approximately 3.09^{-11}, this value was extremely close to 0, indicating a significant statistical correlation between the two variables, further justifying that chi-square was an appropriate test to review.

C. Univariate Statistical Analysis

```python
#Univariate Comparison of Qualitative Variables
plt.figure(figsize = [17,5])
plt.suptitle("Univariate Qualitative Variable Visualization")

#Left plot is Pie Chart of "Contract" a Qualitative Ordinal Variable
plt.subplot(1, 2, 1)
plt.title("Distribution of Contract Types")
plt.pie(df['Contract'].value_counts(), labels=df['Contract'].value_counts().index, autopct='%1.1f%%', startangle=90)
plt.axis('square');

#Right plot is a Bar Chart of "Marital" a Qualitative Nominal Variable
plt.subplot(1, 2, 2)
plt.title("Distribution of Marital Status")
marital_count = df['Marital'].value_counts()
marital_labels = ['Divorced', 'Married', 'Never Married', 'Separated', 'Widowed']
plt.bar(marital_labels, marital_count, width=0.5)
plt.xlabel('Marital Status')
plt.ylabel('Count')
plt.show()
```



```
#Describing Qualitative Variables
df['Contract'].value_counts()
```

```
Contract
Month-to-month    5456
Two Year          2442
One year          2102
Name: count, dtype: int64
```

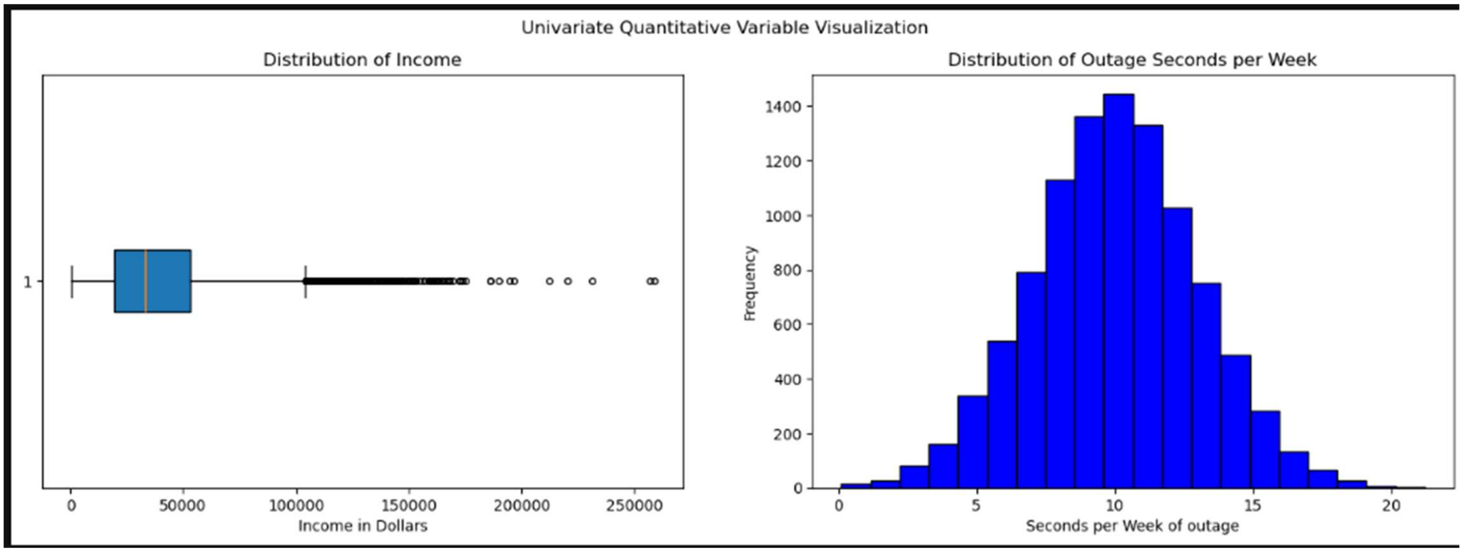
```
df['Marital'].value_counts()
```

```
Marital
Divorced         2092
Widowed          2027
Separated        2014
Never Married    1956
Married          1911
Name: count, dtype: int64
```

```
#Univariate Comparison of Quantitative Variables
plt.figure(figsize = [17,5])
plt.suptitle("Univariate Quantitative Variable Visualization")

#Left Plot is a Box Plot of "Income" a Quantitative Continous variable
plt.subplot(1, 2, 1)
plt.title("Distribution of Income")
plt.boxplot(df['Income'], vert=False, patch_artist=True,
            flierprops=dict(marker='o', color='red', markersize=4))
plt.xlabel('Income in Dollars')
plt.show

#Right Plot is a Histogram of "Outate_sec_perweek" a Quantitative continuous variable
plt.subplot(1, 2, 2)
plt.title("Distribution of Outage Seconds per Week")
plt.hist(df['Outage_sec_perweek'], bins=20, color='blue', edgecolor='black')
plt.xlabel("Seconds per Week of outage")
plt.ylabel("Total Frequency")
plt.show()
```




```
#Calculating the IQR, Upper Bound and total count of outliers from Income Boxplot
Q1 = df['Income'].quantile(0.25)
Q3 = df['Income'].quantile(0.75)
IQR = Q3 - Q1
upper_bound = Q3 + 1.5 * IQR
total_outliers = (df['Income'] > upper_bound).sum()
print(f"Upper Bound Value is {upper_bound}")
print(f"Total count of outliers is {total_outliers}")

Upper Bound Value is 104278.34875
Total count of outliers is 336
```

```
#Describing categorical variables
df['Income'].describe()

count     10000.000000
mean      39806.926771
std       28199.916702
min         348.670000
25%       19224.717500
50%       33170.605000
75%       53246.170000
max      258900.700000
Name: Income, dtype: float64
```

```
df['Outage_sec_perweek'].describe()

count    10000.000000
mean        10.001848
std          2.976019
min          0.099747
25%          8.018214
50%         10.018560
75%         11.969485
max         21.207230
Name: Outage_sec_perweek, dtype: float64
```

Variables utilized for Univariate statistical analysis are described as follows:

Name	environment data type	data type	example	Description / Notes
Contract	object	Qualitative Ordinal	Month-to-month	Service Details - Type of contract the customer has (monthly, annual, bi-annual)
Marital	object	Qualitative Nominal	Married	Customer Self-reported Demographics - Marital status
Income	float64	Quantitative Continuous	21704.77	Customer Self-reported Demographics - Annual income earned
Outage_sec_perweek	float64	Quantitative Continuous	12.01454108	Service Details - Average seconds per week of outage in the customer's neighborhood

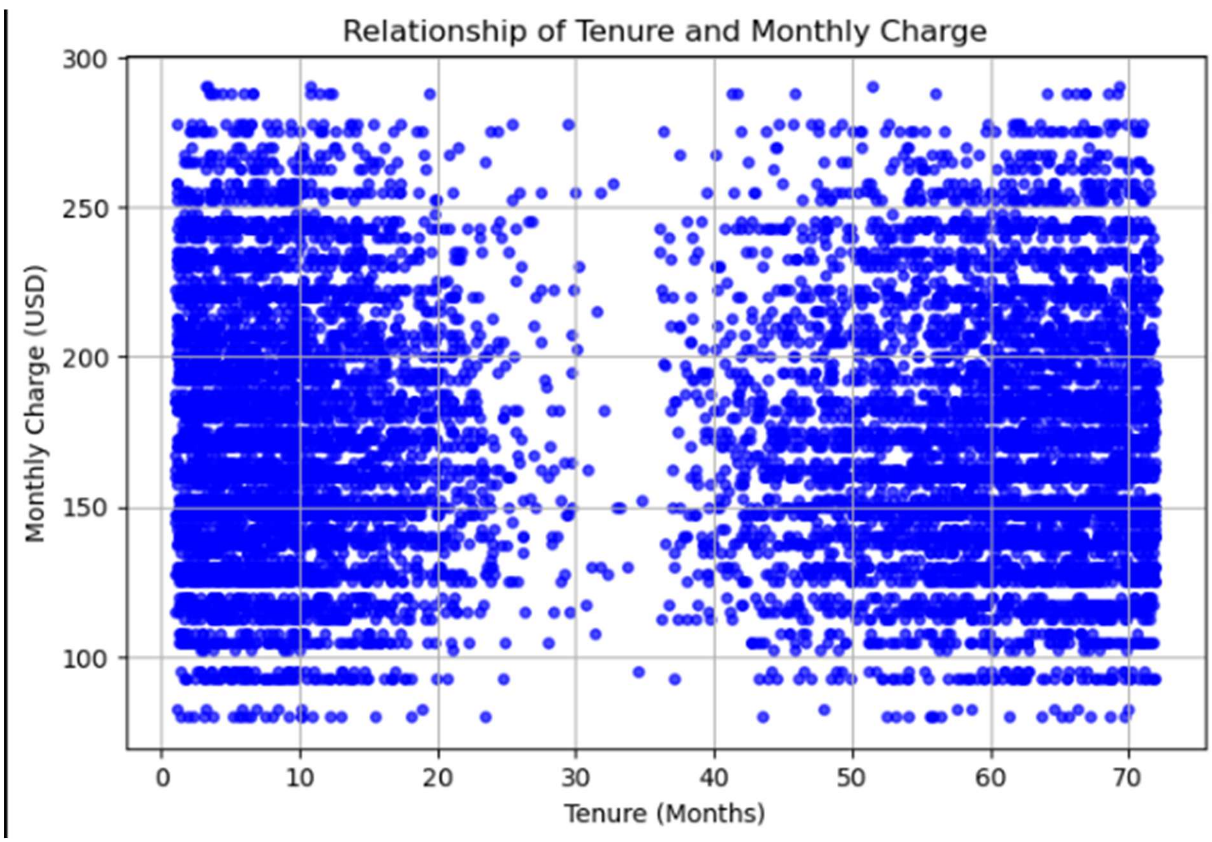
Most of the analyzed variables exhibited atypical or unexpected distributions or details. The distribution of contracts had a higher modality for month-to-month compared to annual, but this is expected in an industry with a relatively high rate of churn and competitive options. Marital produced some unusual results, with mean and median appearing much closer than expected, as all categories seemed reasonably close to one another in terms of overall count. This is surprising as only 1 in 5 customers are married, significantly lower than the US average of 54.6 (Census, 2023). Income represented some expected statistics, with the mean of customers being $39,806 annually. Further analysis was conducted to ensure the outliers in the model were low, with only 336 or a little over 3.3% of customers being significant outliers outside of the upper bound of the boxplot. Finally, the weekly outage demonstrated a reasonably normal distribution across the dataset.

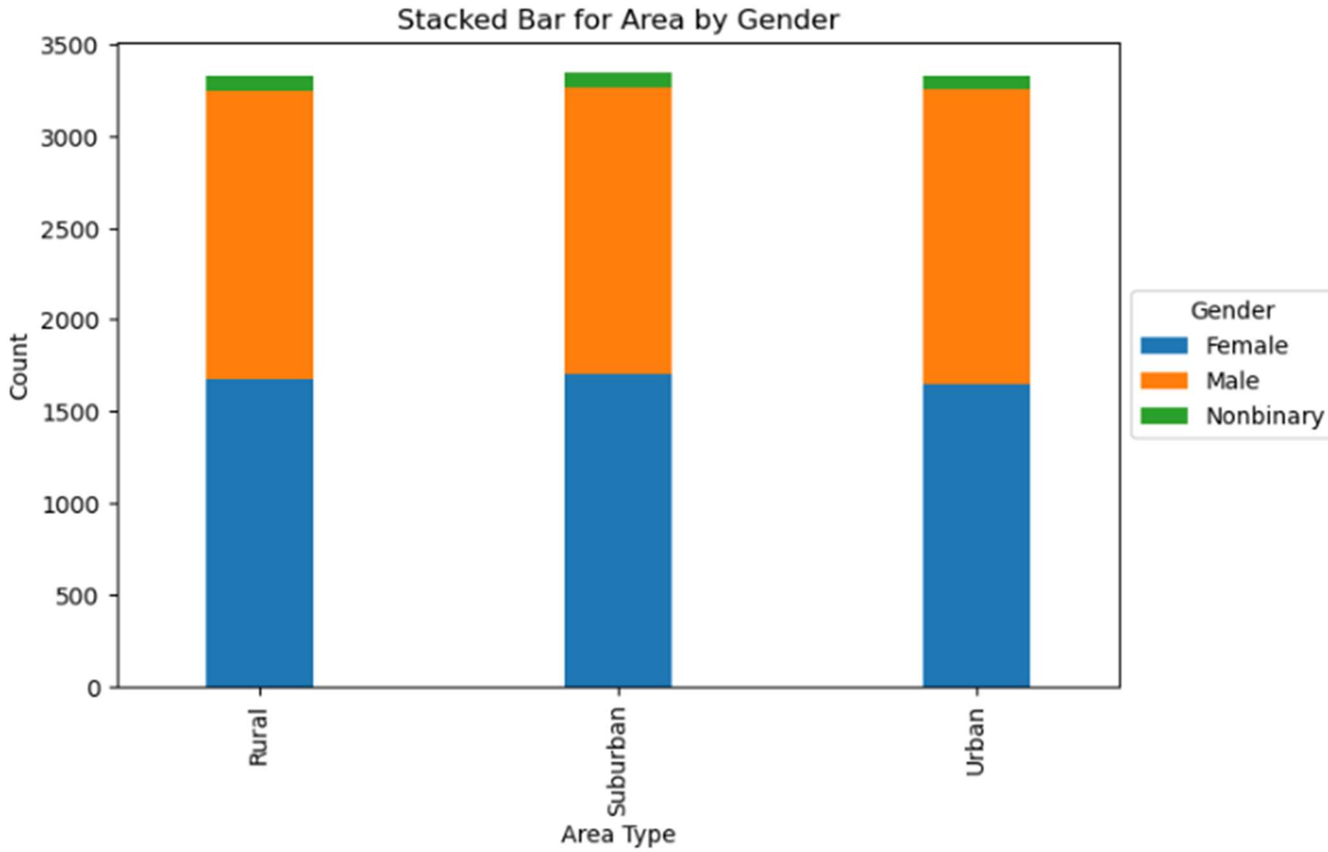
D. Bivariate Statistical Analysis

```
# Bivariate Comparison of Variables
plt.figure(figsize=[17, 5])  # Create a figure of the correct size
plt.suptitle("Bivariate Variable Visualization")  # Title for the entire figure

# Left plot is a Scatterplot of "Tenure" and "MonthlyCharge"
plt.subplot(1, 2, 1)
plt.title("Relationship of Tenure and Monthly Charge")
plt.scatter(df['Tenure'], df['MonthlyCharge'], color='blue', alpha=0.7, s=15)
plt.xlabel("Tenure (Months)")
plt.ylabel("Monthly Charge (USD)")
plt.grid(True)

# Right plot is a Stacked Bar for "Area" and "Gender"
plt.subplot(1, 2, 2)
plt.title('Stacked Bar for Area by Gender')
counts = df.groupby(['Area', 'Gender']).size().unstack(fill_value=0)
ax = counts.plot(kind='bar', stacked=True, width=0.3, ax=plt.gca())
plt.xlabel('Area Type')
plt.ylabel('Count')
plt.legend(title='Gender', loc='center left', bbox_to_anchor=(1, 0.5))
plt.show()
```






```
#Identifying correlation between Tenure and MonthlyCharge
correlation = df['Tenure'].corr(df['MonthlyCharge'])
print(f"Correlation: {correlation}")
##Very low correlation coefficient

Correlation: -0.0033368104134518864
```

```
#Describing Variables
print(df['Tenure'].describe())

count    10000.000000
mean        34.526188
std         26.443063
min          1.000259
25%          7.917694
50%         35.430507
75%         61.479795
max         71.999280
Name: Tenure, dtype: float64
```

```
print(df['MonthlyCharge'].describe())

count    10000.000000
mean       172.624816
std         42.943094
min         79.978860
25%        139.979239
50%        167.484700
75%        200.734725
max        290.160419
Name: MonthlyCharge, dtype: float64
```

```
#Viewing Proportional Distribution of variables in normalized Contingency Table
contingency_table = pd.crosstab(df['Gender'], df['Area'], normalize = 'columns')
print(contingency_table)

Area          Rural  Suburban     Urban
Gender
Female     0.502855  0.508966  0.495642
Male       0.473700  0.466826  0.482717
Nonbinary  0.023445  0.024208  0.021641
```

```
#Checking statistical significance of qualitative variables
contingency_table = pd.crosstab(df['Gender'], df['Area'])
chi2, p, dof, expected = chi2_contingency(contingency_table)
print(f"Chi-square statistic: {chi2}")
print(f"p-value: {p}")
##High p-Value demonstrates insufficient evidence to reject null hypothesis.

Chi-square statistic: 1.9852728650526417
p-value: 0.7384677628593668
```

Variables utilized for Bivariate statistical analysis are described as follows:

Name	environment data type	data type	example	Description / Notes
Tenure	float64	Quantitative Continuous	1.156680997	Service Details - Number of months the customer has been with the provider
MonthlyCharge	float64	Quantitative Continuous	242.9480155	Service Details - Average value of the customer's monthly charges
Gender	object	Qualitative Nominal	Female	Customer Self-reported Demographics - Gender identification (male, female, non-binary)
Area	object	Qualitative Nominal	Urban	Customer Demographics - Census Data - Classification of Area Type (Urban vs. Rural vs. Suburban)

Bivariate statistical analysis produced no relevant insights or relationships between the selected variables. The scatterplot for Tenure and Monthly Charge yielded an extremely low correlation coefficient, and no visual relationship could be identified as there was a clear gap in values in the center of the tenure values, with a clustering towards the higher ends of customer tenure. This indicates no relevant relationship between these two variables. Comparing the categorical Gender and Area variables returned a stacked bar chart with little variability in distribution across all categories. No significant difference can be seen in the visualization. The normalized contingency table represents an even distribution of variables across all combinations available. A conclusion may be drawn that nonbinary individuals were more prone to suburban areas, but this is a slight difference (0.024 to 0.023 and 0.021). A chi-square test was conducted on Area and Gender and returned a p-value too high to reject a standard null hypothesis (0.73), meaning there is no statistical significance between these two variables.

E. Summary of Results and Actionable Items

E1. Hypothesis Test Results

The α value was set at 0.05 or 95% certainty in testing the null hypothesis that no relationship exists between customers identifying as techies and leaving the company. The p-value identified in the chi-square test was 3.09^{-11}, which is extremely close to 0. The low p-value falls well under the α value threshold, and as a result, the alternative hypothesis that there is a statistically significant relationship between customers self-identifying as "Techies" and customer churn is accepted.

E2. Limitations

Even though there is a statistically significant correlation between the two chosen variables, the analysis has several limitations worth noting.

- Causation is not implied when a correlation is identified between two variables – additional analysis and comparisons must be completed before causation can be determined.
- The analysis only uses two variables from the 50 available in the dataset. To draw more relevant and overarching conclusions, statistical tests should be run between other variables and compared to those with statistically significant results.
- Both the churn and techie variables were not normal distributions of responses. Both variables skewed pretty heavily towards "no" responses.

E3. Recommendations

A relationship between customers being technologically savvy and customer churn exists, as evidenced by the chi-square test, which was conducted in the analysis. As these variables have a relationship, the company would benefit from targeting their retention strategies towards customers who self-identify as techies. By exploring additional options to maintain these customers, churn could be reduced overall. More specifically, offering enhanced technical support and tech-related promotional and marketing strategies could reduce the churn rate overall. Expanding on the company's overall technological repertoire could also reduce churn, as techie customers would be more interested in more advanced technology items and may be willing to maintain their contract if the offerings are superior to those offered by competitors. Finally, additional analysis should be conducted between other variables in the dataset and underlying causes of churn to more directly identify areas where customers are interested in maintaining their service.

References

Census.gov. (09/17/2023) *Unmarried and Single Americans Week: September 17-23, 2023.*

https://www.census.gov/newsroom/stories/unmarried-single-americans-week.html

Bobbitt, Zach. (10/07/2021). *How to Change the Position of a Legend in Matplotlib.*

https://www.statology.org/matplotlib-legend-position/