D207 Exploratory Data Analysis – Task 1

Eric D. Otten

Student ID: 011399183

A. Research Question

My research question is: “Does marital status have a statistically significant impact on a customer’s tenure with the internet service provider?”

A2. Justification of Research Question

This question is relevant to an organization for marketing segmentation purposes; the marketing department of an organization would want to identify which demographics to target for marketing campaigns.

Additionally, this information can be used for customer retention purposes; knowing which groups are more likely to stay longer or leave sooner allows the service provider to develop tailored retention strategies such as loyalty benefits.

Lastly, an organization can use this information to forecast customer churn and retention. For example, if an organization receives a higher percentage of customers who are married and knows that these customers have a longer tenure, it can anticipate an increase in earnings.

A3. Relevant Data

The following columns are relevant to answering my research question:

- Marital (independent variable): Marital status of the customer.
- Tenure (dependent variable): Number of months the customer has been with the service provider.

B1. Data Set Analysis

Shown below is an analysis of Tenure and Marital status, calculated using analysis of variance (ANOVA) and implemented in the Python programming language using pandas and scipy.

Code

```
from scipy.stats import f_oneway
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
from scipy.stats import chi2_contingency

df = pd.read_csv("churn_clean.csv", keep_default_na=False, na_values=["NA"])

# B1: Data Set Analysis
```

```
table = pd.crosstab(df.Tenure, df.Marital)
grouped_tenures = df.groupby("Marital")["Tenure"].apply(list).to_dict()

tenures_married = grouped_tenures["Married"]
tenures_separated = grouped_tenures["Separated"]
tenures_divorced = grouped_tenures["Divorced"]
tenures_widowed = grouped_tenures["Widowed"]
tenures_never_married = grouped_tenures["Never Married"]

anova_result = f_oneway(
    tenures_married,
    tenures_separated,
    tenures_divorced,
    tenures_widowed,
    tenures_never_married,
)

print("F-statistic:", anova_result.statistic)
print("P-value:", anova_result.pvalue)
```

B2. Analysis Output

Output
F-statistic: 0.21046519425379842 P-value: 0.9327428319416796

B3. Justification of Technique

I chose the analysis of variance (ANOVA) technique because it is used to compare the means of three or more groups with continuous data. Walde (1977) defined, "Analysis of Variance partitions the observed variance based on explanatory (independent) variables."

I originally planned to use the chi-square technique, however, I realized this was not a suitable approach because the chi-square test requires two categorical variables. In my initial analysis, I treated each unique tenure value as a categorical value, which yielded a p-value of 0.5. This error arose due to the fact that I treated a continuous numerical variable (tenure) as a categorical variable.

C. Distribution Using Univariate Statistics

I chose to analyze the following continuous variables using univariate statistics:

- Tenure: Understanding the distribution of customer tenure can help identify different segments based on how long customers have been with the company. When graphed, customer tenure appears to be bi-modal, indicating that customers can be segmented into short-term and long-term customers.

- Income: Understanding the distribution of customer income can be used to create targeted marketing campaigns and personalized offers that cater to different income groups. When graphed, customer income appears to be right-skewed, indicating that the mean income is affected by several outliers, i.e., high-earning customers.

Additionally, I chose to analyze the following categorical variables using univariate statistics:

- Marital Status: Knowing the distribution of marital status among a business' customers allows the organization to create targeted marketing campaigns. For example, offers for family plans might appeal more to married customers, while single customers might be interested in individual packages or lifestyle-related services.

- Area: Knowing the distribution of area type among a business' customers allows the business to create targeted marketing campaigns. For example, tight-knit communities may benefit more heavily from referral incentives due to the typically higher trust factor found in rural communities when compared to urban ones.

The univariate statistics for these variables were calculated using the code below. The univariate statistics of the categorical variables were calculated using the `.describe()` method which lists the count, mean, standard deviation, minimum, 25%, 50%, and 75% percentiles, as well as the maximum value. The `.describe()` method was used on the Tenure and Income variables.

The univariate statistics for the categorical variables were calculated using the `.value_counts()` method once without any parameters, and then a second time using the normalize=True argument. These lines of code list out the total count and relative frequency of the values within the categorical variables Marital and Area.

Code

```
# C. Distribution Using Univariate Statistics
# Calculate basic statistics for 'Tenure' and 'Income'
tenure_stats = df['Tenure'].describe()
income_stats = df['Income'].describe()

print("Tenure Statistics:")
print(tenure_stats)

print("\nIncome Statistics:")
print(income_stats)

# Calculate frequencies and proportions for 'Marital' and 'Area'
marital_counts = df['Marital'].value_counts()
marital_proportions = df['Marital'].value_counts(normalize=True)

area_counts = df['Area'].value_counts()
area_proportions = df['Area'].value_counts(normalize=True)

print("\nMarital Status Distribution:")
```

```
print(marital_counts)
print(marital_proportions)

print("\nArea Distribution:")
print(area_counts)
print(area_proportions)
```

Result

```
Tenure Statistics:
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

Income Statistics:
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
Marital Status Distribution:
Marital
Divorced         2092
Widowed          2027
Separated        2014
Never Married    1956
Married          1911
Name: count, dtype: int64
Marital
Divorced         0.2092
Widowed          0.2027
Separated        0.2014
Never Married    0.1956
Married          0.1911
Name: proportion, dtype: float64

Area Distribution:
Area
Suburban    3346
Urban       3327
Rural       3327
Name: count, dtype: int64
Area
```

```
Suburban    0.3346
Urban       0.3327
Rural       0.3327
Name: proportion, dtype: float64
```

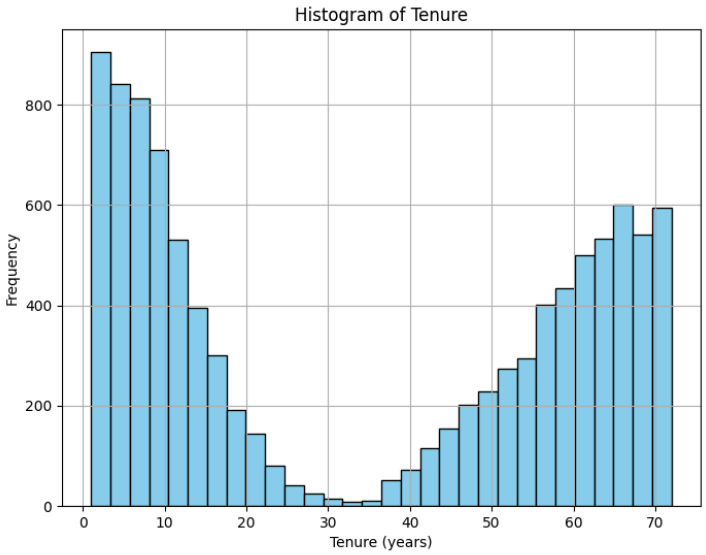
C1. Visual of Findings

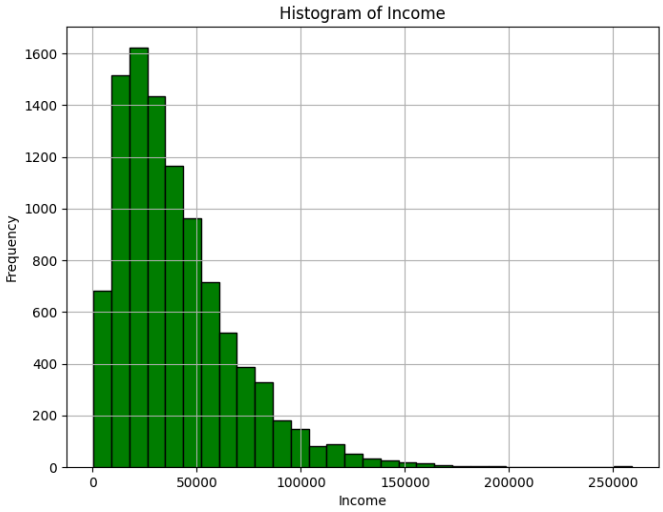
The continuous variables were graphed as histograms using the code below.

Code

```
# Histogram for Tenure
plt.figure(figsize=(8, 6))
plt.hist(df["Tenure"], bins=30, color="skyblue", edgecolor="black")
plt.title("Histogram of Tenure")
plt.xlabel("Tenure (years)")
plt.ylabel("Frequency")
plt.grid(True)
plt.show()

# Histogram for Income
plt.figure(figsize=(8, 6))
plt.hist(df["Income"], bins=30, color="green", edgecolor="black")
plt.title("Histogram of Income")
plt.xlabel("Income")
plt.ylabel("Frequency")
plt.grid(True)
plt.show()
```



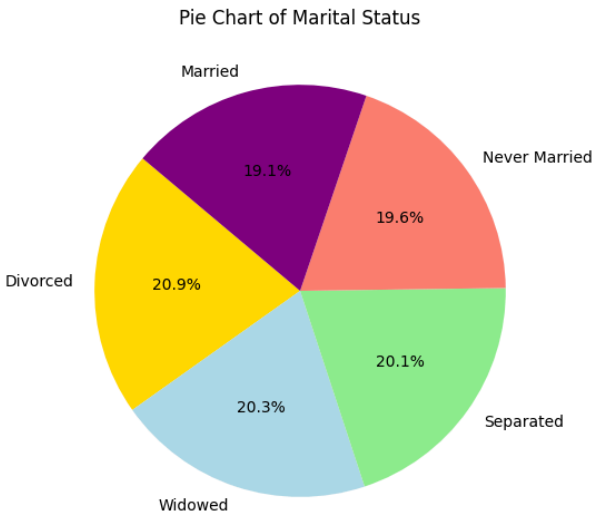


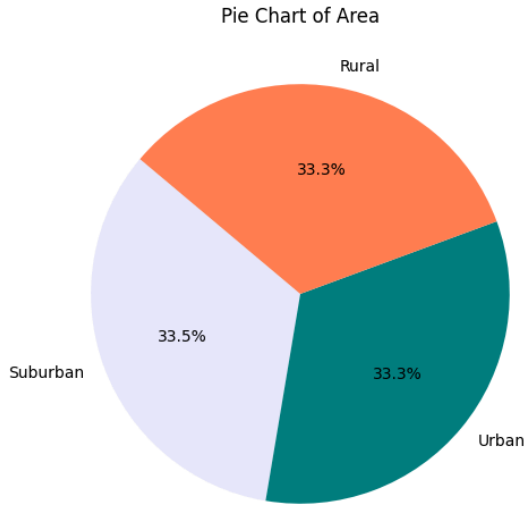
The categorical variables were graphed as pie charts using the code below.

Code

```
# Pie Chart for Marital Status
marital_counts = df["Marital"].value_counts()
plt.figure(figsize=(8, 6))
plt.pie(
    marital_counts,
    labels=marital_counts.index,
    autopct="%1.1f%%",
    startangle=140,
    colors=["gold", "lightblue", "lightgreen", "salmon", "purple"],
)
plt.title("Pie Chart of Marital Status")
plt.show()

# Pie Chart for Area
area_counts = df["Area"].value_counts()
plt.figure(figsize=(8, 6))
plt.pie(
    area_counts,
    labels=area_counts.index,
    autopct="%1.1f%%",
    startangle=140,
    colors=["lavender", "teal", "coral"],
)
plt.title("Pie Chart of Area")
plt.show()
```





D. Distribution Using Bivariate Statistics

I chose to analyze the following pairs of continuous variables using bivariate statistics:

- Income vs Monthly Charge: Analyzing this relationship can reveal how much customers are willing to pay for services based on their income level. This can help in understanding pricing strategies or customer segmentation based on economic status. The analysis shows no linear relationship between these two variables.

- Tenure and Bandwidth_GB_Year: Analyzing this relationship can reveal customer usage patterns and help decision makers plan the network capacity of the service provider. The analysis shows a positive linear relationship between a customer's tenure and their usage, indicating that customers use more bandwidth the longer they stay with the company, or the type of customer who uses more bandwidth is also the type of customer to stay with the company longer. A further analysis including churn rates could provide insight into which explanation is more accurate.

Additionally, I chose to analyze the following pairs of categorical variables using bivariate statistics:

- Internet Service vs Contract Type: Analyzing this relationship can reveal preferences or tendencies in the type of service chosen by those who commit to longer or shorter contract terms. This can assist in tailoring service offerings or promotional strategies.
- Churn vs. Payment Method: Analyzing this relationship can reveal patterns that indicate higher risk of churn based on payment method. The analysis shows that customers who pay by electronic check are overrepresented among customers who churn. An organization can use this information by funneling customers towards payment methods designed to reduce churn.

The bivariate statistics for these variables were calculated using the code below. The bivariate statistics for the categorical variables were calculated as a correlation coefficient using the `.corr()` method, which was used to find the correlation between Income vs. MonthlyCharge and Tenure vs. Bandwidth_GB_Year.

The bivariate statistics for the categorical variables were calculated using contingency tables via the `pd.crosstab()` method and chi-squared tests were performed using the `chi2_contingency()` method.

Code

```
# D. Distribution Using Bivariate Statistics
# Calculate correlation coefficient for Income and MonthlyCharge
income_monthly_charge_corr = df['Income'].corr(df['MonthlyCharge'])
print(f"Correlation coefficient between Income and Monthly Charge:
{income_monthly_charge_corr:.2f}")

# Calculate correlation coefficient for Tenure and Bandwidth_GB_Year
tenure_bandwidth_corr = df['Tenure'].corr(df['Bandwidth_GB_Year'])
print(f"Correlation coefficient between Tenure and Bandwidth_GB_Year:
{tenure_bandwidth_corr:.2f}")

# Contingency Table for Internet Service vs Contract Type
internet_contract_table = pd.crosstab(df['InternetService'], df['Contract'])
print("\nContingency Table for Internet Service vs Contract Type:")
```

```
print(internet_contract_table)

# Perform Chi-squared test
chi2_stat, p_val, dof, expected = chi2_contingency(internet_contract_table)
print(f"\nChi-squared Test results for Internet Service vs Contract Type:")
print(f"Chi-squared Statistic: {chi2_stat}, P-value: {p_val}")

# Contingency Table for Churn vs Payment Method
churn_payment_table = pd.crosstab(df['Churn'], df['PaymentMethod'])
print("\nContingency Table for Churn vs Payment Method:")
print(churn_payment_table)

# Perform Chi-squared test
chi2_stat, p_val, dof, expected = chi2_contingency(churn_payment_table)
print(f"\nChi-squared Test results for Churn vs Payment Method:")
print(f"Chi-squared Statistic: {chi2_stat}, P-value: {p_val}")
```

Result

```
Correlation coefficient between Income and Monthly Charge: -0.00
Correlation coefficient between Tenure and Bandwidth_GB_Year: 0.99

Contingency Table for Internet Service vs Contract Type:
Contract         Month-to-month  One year  Two Year
InternetService
DSL                        1878       734       851
Fiber Optic                2419       897      1092
None                       1159       471       499

Chi-squared Test results for Internet Service vs Contract Type:
Chi-squared Statistic: 3.481478725512579, P-value: 0.4807001277448576

Contingency Table for Churn vs Payment Method:
PaymentMethod  Bank Transfer(automatic)  Credit Card (automatic)  Electronic Check  Mailed Check
Churn
No                                1671                     1543              2435          1701
Yes                                558                      540               963           589

Chi-squared Test results for Churn vs Payment Method:
Chi-squared Statistic: 9.437373459430551, P-value: 0.02400702004497883
```

The output of this code shows a correlation of roughly 0.0 for Income and MonthlyCharge, which indicates there is no linear relationship between these two variables and suggests that variations in a customer's income do not correspond to changes in their monthly charges.

Additionally, the output of this code shows a correlation of roughly 0.99 for Tenure and Bandwidth_GB_Year, which indicates a very strong linear relationship and suggests that as the tenure of a customer increases, their bandwidth usage over the year also increases almost proportionally.

From the contingency table for Internet Service vs Contract Type, we see that both DSL and Fiber Optic are most commonly subscribed on a month-to-month basis, which indicates a preference for less commitment among customers using these services.

The chi-squared value for these two variables is roughly 0.48, well above the conventional threshold of 0.05, which suggests no significant association between the type of Internet Service and Contract Type.

The contingency table for Churn vs. Payment method shows that Electronic Checks seem to have a higher count in both the churned and not churned categories, suggesting a possible preference or policy that encourages this payment method.

The chi-squared value for these two variables is roughly 0.02, which is below the conventional threshold of 0.05 and suggests an association between these two variables. This means that the likelihood of churn may be influenced by the payment method, with a notably higher proportion of churn among those who pay by Electronic Check.

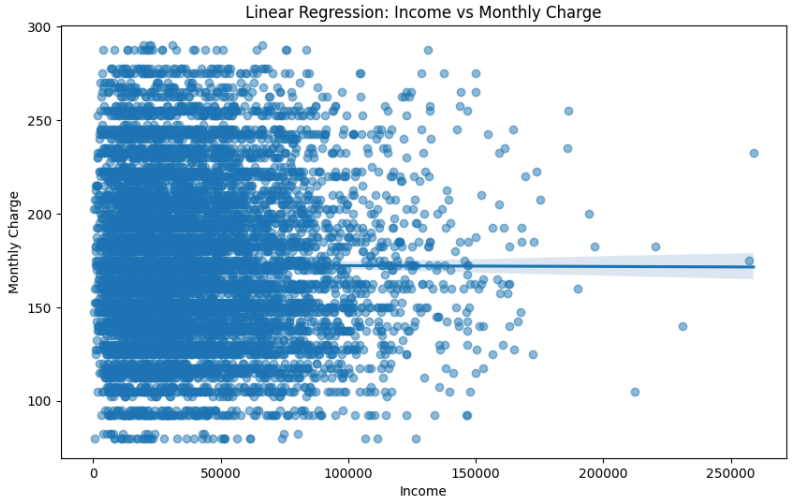
D1. Visual of Findings

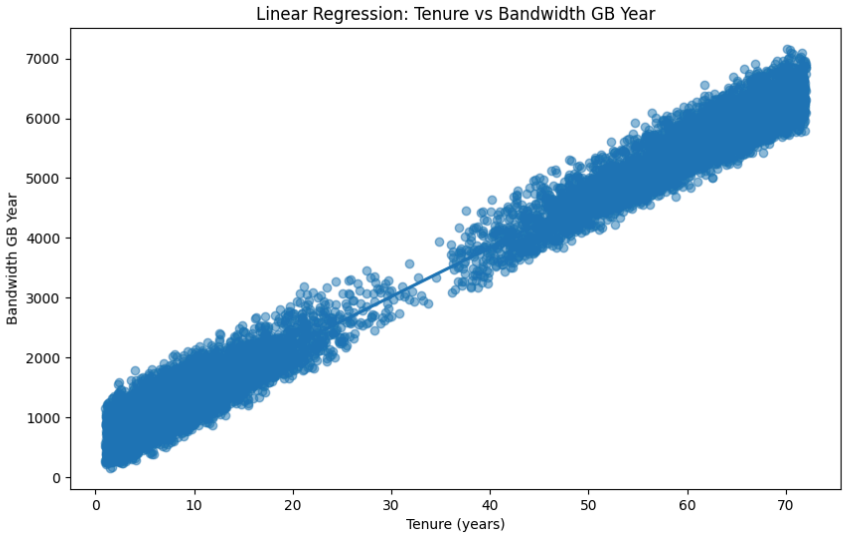
These variables were graphed as linear regression plots using the code below.

Code

```
# Plot for Income vs. MonthlyCharge
plt.figure(figsize=(10, 6))
sns.regplot(x="Income", y="MonthlyCharge", data=df, scatter_kws={"alpha": 0.5})
plt.title("Linear Regression: Income vs Monthly Charge")
plt.xlabel("Income")
plt.ylabel("Monthly Charge")
plt.show()

# Plot for Tenure vs. Bandwidth_GB_Year
plt.figure(figsize=(10, 6))
sns.regplot(x="Tenure", y="Bandwidth_GB_Year", data=df, scatter_kws={"alpha": 0.5})
plt.title("Linear Regression: Tenure vs Bandwidth GB Year")
plt.xlabel("Tenure (years)")
plt.ylabel("Bandwidth GB Year")
plt.show()
```





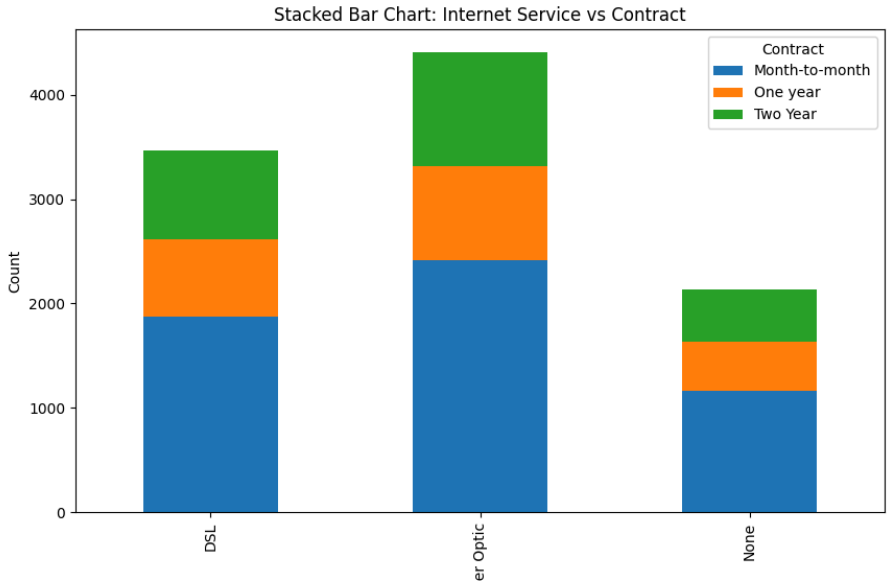
These variables were graphed as stacked bar charts using the code below.

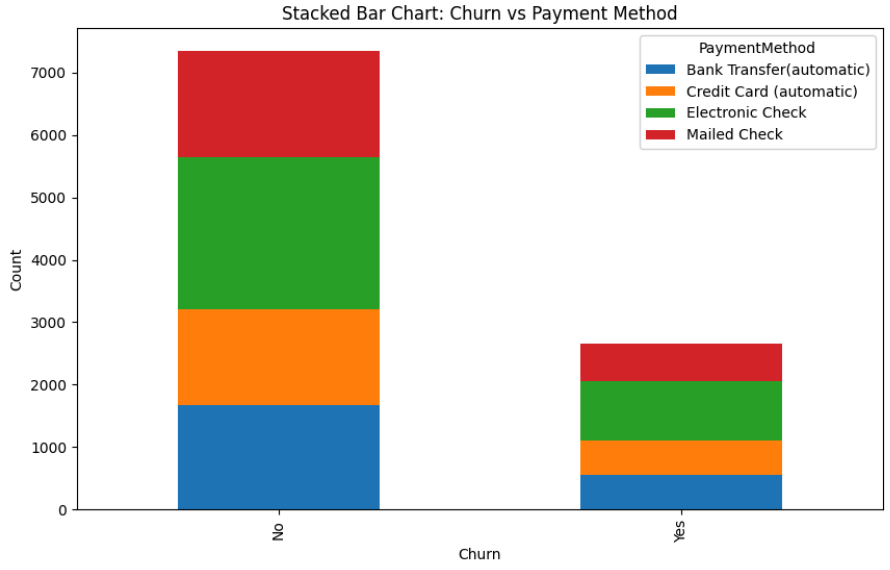
Code

```
# Data preparation for InternetService vs. Contract
internet_contract = pd.crosstab(df['InternetService'], df['Contract'])
internet_contract.plot(kind='bar', stacked=True, figsize=(10, 6))
plt.title('Stacked Bar Chart: Internet Service vs Contract')
plt.xlabel('Internet Service')
plt.ylabel('Count')
plt.show()

# Data preparation for Churn vs. PaymentMethod
```

```
churn_payment = pd.crosstab(df['Churn'], df['PaymentMethod'])
churn_payment.plot(kind='bar', stacked=True, figsize=(10, 6))
plt.title('Stacked Bar Chart: Churn vs Payment Method')
plt.xlabel('Churn')
plt.ylabel('Count')
plt.show()
```





E1. Results of Hypothesis Test

The output of the ANOVA test includes a p-value of approximately 0.93. In most research contexts, an alpha level (p-value threshold) of 0.05 is used to determine statistical significance. Therefore, there is not enough evidence to reject the null hypothesis which states that there is no correlation between marital status and customer tenure. Walde (1977) concluded, "If H0 is rejected, we conclude that not all the μ's are equal."

The f-statistic is low: approximately 0.21. This suggests that the mean tenure of each marital status is similar.

E2. Limitations of Analysis

One of the key assumptions of ANOVA is that all groups (marital statuses) have roughly equal variances in the dependent variable (tenure). However, different marital statuses may inherently have different levels of variability due to household needs. For example, suppose the data looked like this:

- Married: Tenure values are mostly around 5-10 years with little variation.
- Never Married: Tenure values range widely from 1 year to 15 years.

Since the variance in tenure for married customers is small and the variance for never married customers is large, the assumption of equal variances (homoscedasticity) is violated.

E3. Recommended Course of Action

From a business perspective, these results suggest that any efforts to tailor services or marketing strategies based on marital status may not influence customer tenure. Resources might be better spent on other demographic or psychographic factors that show a stronger correlation with tenure.

G. Web Sources

The following web sources were consulted:

Analysis of Variance (ANOVA): Types and Limitations
https://www.analyticssteps.com/blogs/analysis-variance-anova-types-and-limitations

scipy.stats.f_oneway documentation
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html

H. Works Consulted

Walde, J. (1977). Analysis of Variance. Lecture Notes.