Exploratory Data Analysis

February 14, 2024

A1. Research Question Is there a significant difference in the monthly charges between customers who churn and those who do not churn?

A2. Benefits to Stakeholders Benefits: Analyzing the data to answer this question can provide valuable insights for the organization's stakeholders. For example:

- Management can understand the impact of pricing on customer retention and make informed decisions about pricing strategies.
- Marketing can identify segments at higher risk of churning and target them with promotional offers.
- Customer Service can prioritize engagement with customers who have higher monthly charges and might be at risk of churning.
- Product Development can use these insights to tailor service offerings that might better meet the needs of different customer segments and thus reduce churn.

A3. Relevant Data The below are the relevant variables for my research question:

- Churn (Qualitative): Churn status. Example: 'No'
- InternetService (Qualitative): Type of internet service. Example: 'Fiber Optic'
- MonthlyCharge (Quantitative): Monthly charge. Example: 171.449762
- Bandwidth_GB_Year (Quantitative): Annual bandwidth usage. Example: 904.536110

B1 Analysus of Variables

[2]:
```
# see attach codes

import pandas as pd
from scipy import stats

# Load the dataset
df = pd.read_csv(r'C:\Users\Hien Ta\OneDrive\WGU\MSDA\D207\churn_clean.csv')

# Grouping data by churn status and extracting monthly charges
churned = df[df['Churn'] == 'Yes']['MonthlyCharge']
not_churned = df[df['Churn'] == 'No']['MonthlyCharge']

# Performing a t-test
statistic, pvalue = stats.ttest_ind(churned, not_churned, equal_var=False)
```

```
# Printing the results
print(f"T-test: Statistic={statistic}, p-value={pvalue}")

# D207 T-Test-Python-pdf. (2023)
```

```
T-test: Statistic=39.28778644007045, p-value=1.7823941678632952e-290
```

B2 Analyis Results The T-test results yielded a statistic of approximately 39.29 and a p-value of 1.78e-290, indicating a statistically significant difference in monthly charges between the two groups.

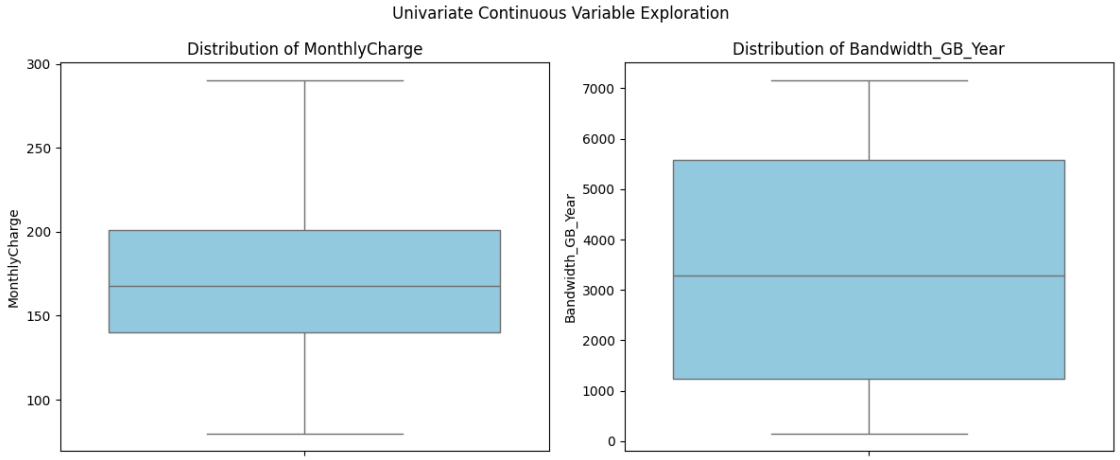
B3 Justification for Analysis Technique A T-test was selected for this analysis as it is appropriate for comparing the means of two independent samples. In this case, it helps us determine if the monthly charges for churned customers significantly differ from those who didn't churn. I believe that this is crucial for understanding factors influencing customer churn.

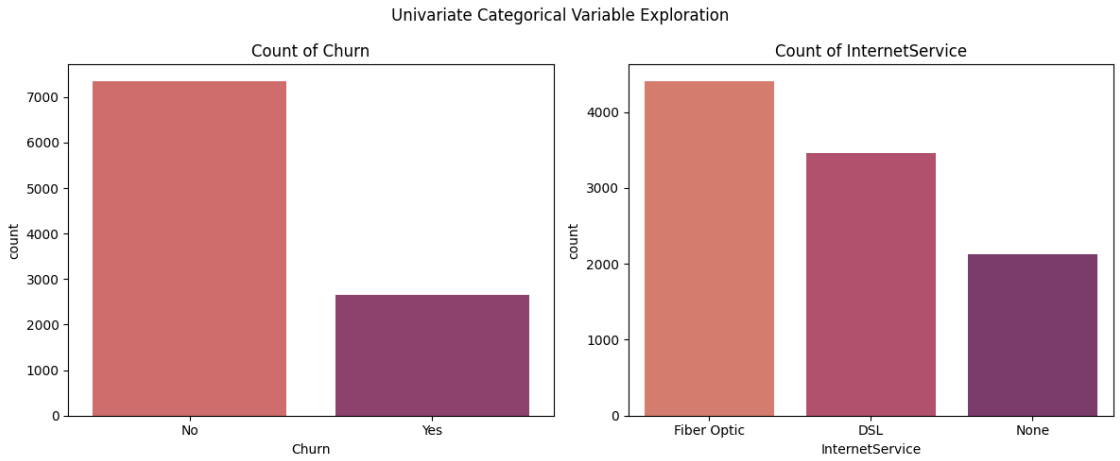
C1 Univariate Statistics of Continuous and Categorical Variables

[3]:
```
# see attached codes

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
df = pd.read_csv(r'C:\Users\Hien Ta\OneDrive\WGU\MSDA\D207\churn_clean.csv')

# Replace 'nan' with 'None' in the 'InternetService' column
df['InternetService'] = df['InternetService'].fillna('None')

# Checking for unique values in the 'InternetService' column to verify the␣
 ↪presence of 'None'
unique_internet_services = df['InternetService'].unique()
print(unique_internet_services)

# Selecting two continuous and two categorical variables for univariate analysis
continuous_vars = ['MonthlyCharge', 'Bandwidth_GB_Year']
categorical_vars = ['Churn', 'InternetService']

# Define the color palette for the plots from seaborn
color_palette = "flare"

# Plotting boxplots for continuous variables
plt.figure(figsize=(12, 5))
plt.suptitle("Univariate Continuous Variable Exploration")

for i, var in enumerate(continuous_vars, 1):
    plt.subplot(1, 2, i)
```

```
    sns.boxplot(y=df[var], color='skyblue')
    plt.title(f'Distribution of {var}')

plt.tight_layout()
plt.show()

# Plotting count plots for categorical variables with the new color palette
plt.figure(figsize=(12, 5))
plt.suptitle("Univariate Categorical Variable Exploration")

for i, var in enumerate(categorical_vars, 1):
    plt.subplot(1, 2, i)
    sns.countplot(x=var, data=df, hue=var, palette=color_palette, legend=False)
    plt.title(f'Count of {var}')

plt.tight_layout()
plt.show()


# Sewell, W. (2023)
```

```
['Fiber Optic' 'DSL' 'None']
```





The above plots shows a distributions of a variable from the Churn dataset. These plots cover 4 variables:

- 'MonthlyCharge' (continuous)
- 'Bandwidth_GB_Year' (continuous)
- 'Churn' (categorical)
- 'InternetService' (categorical)

[25]:
```
df.Churn.value_counts()
```

[25]:
```
Churn
No     7350
Yes    2650
Name: count, dtype: int64
```

[27]:
```
df.InternetService.value_counts().sort_index()
```

[27]:
```
InternetService
DSL            3463
Fiber Optic    4408
None           2129
Name: count, dtype: int64
```

[28]:
```
df.MonthlyCharge.describe()
```

[28]:
```
count    10000.000000
mean       172.624816
std         42.943094
min         79.978860
25%        139.979239
50%        167.484700
75%        200.734725
```

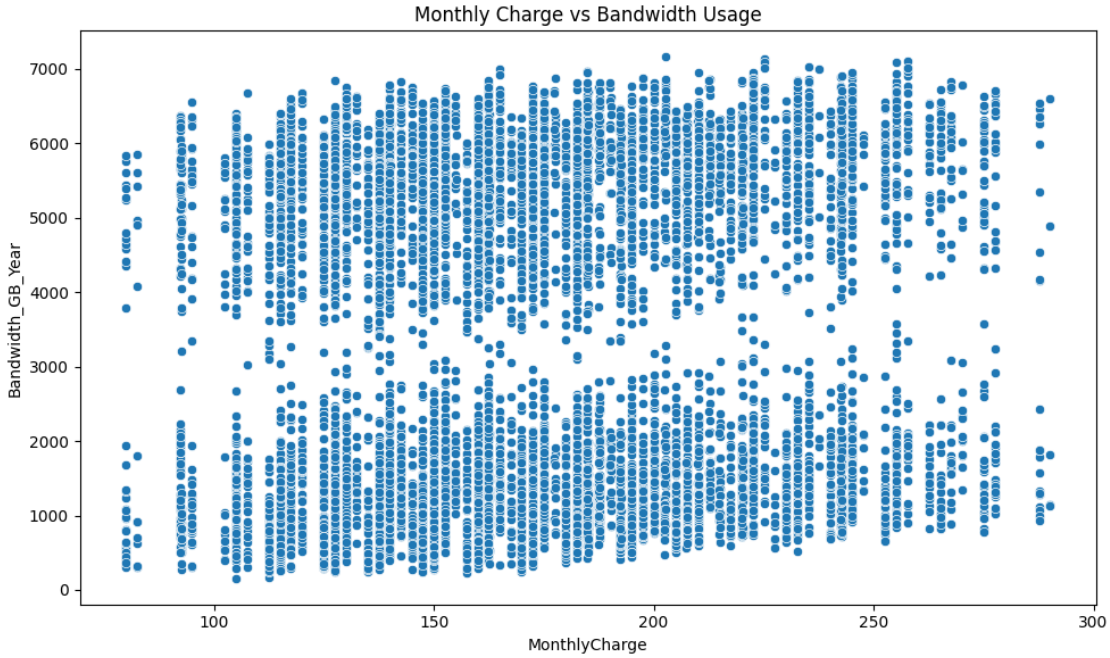
```
max        290.160419
Name: MonthlyCharge, dtype: float64
```

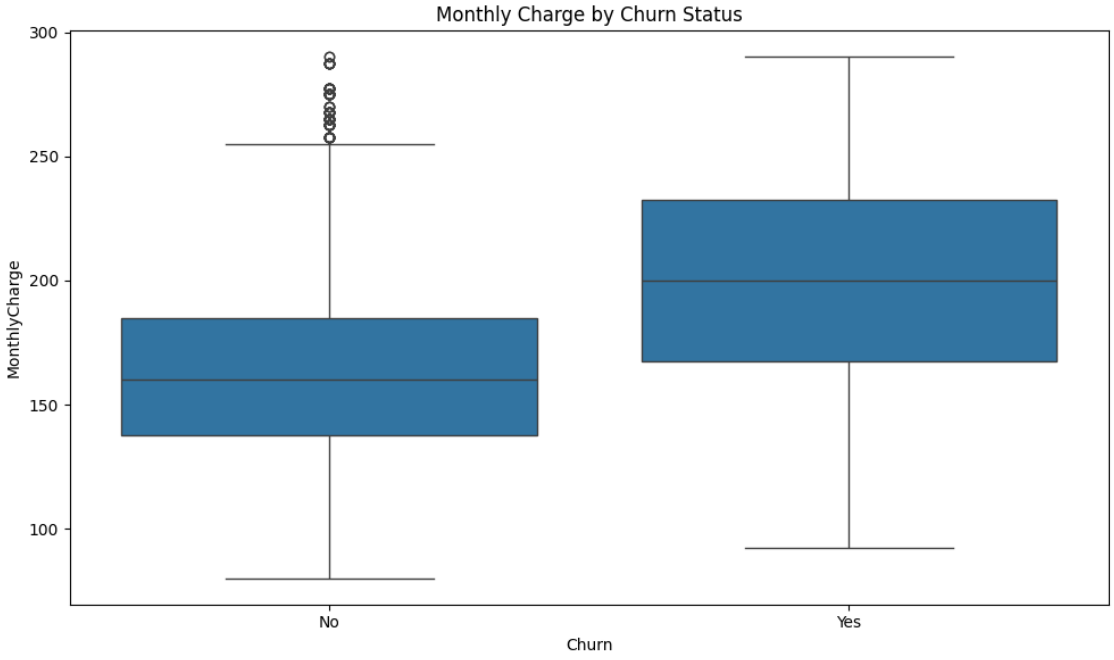
[29]:
```
df.Bandwidth_GB_Year.describe()
```

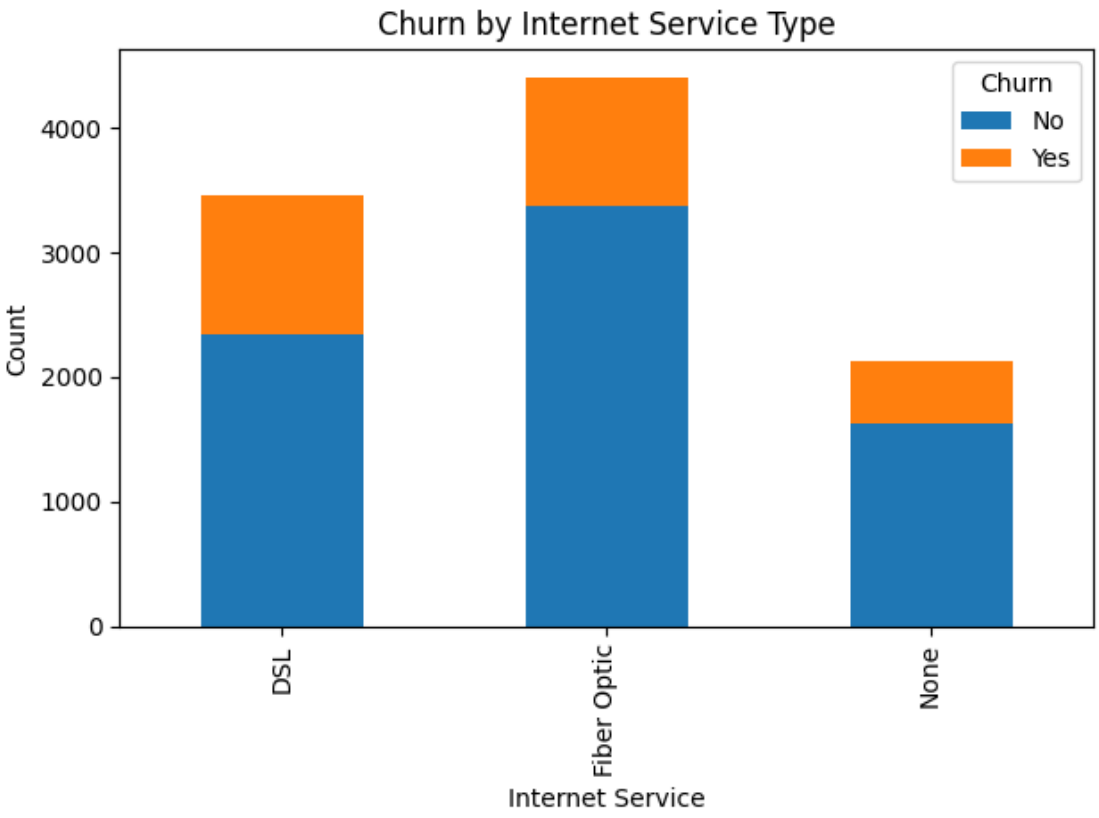
[29]:
```
count    10000.000000
mean      3392.341550
std       2185.294852
min        155.506715
25%       1236.470827
50%       3279.536903
75%       5586.141370
max       7158.981530
Name: Bandwidth_GB_Year, dtype: float64
```

D1 Bivariate Statistics of Continuous and Categorical Variables

[8]:
```
# see attached codes

# Bivariate Analysis: Continuous vs. Continuous
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='MonthlyCharge', y='Bandwidth_GB_Year')
plt.title('Monthly Charge vs Bandwidth Usage')
plt.tight_layout()
plt.show()

# Bivariate Analysis: Continuous vs. Categorical
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='Churn', y='MonthlyCharge')
plt.title('Monthly Charge by Churn Status')
plt.tight_layout()
plt.show()

# Bivariate Analysis: Categorical vs. Categorical
ct = pd.crosstab(df['InternetService'], df['Churn'])
ct.plot(kind='bar', stacked=True)
plt.title('Churn by Internet Service Type')
plt.xlabel('Internet Service')
plt.ylabel('Count')
plt.tight_layout()
plt.show()

# Sewell, W. (2023)
```


Monthly Charge vs Bandwidth Usage
Bandwidth_GB_Year
MonthlyCharge
0
1000
2000
3000
4000
5000
6000
7000
100
150
200
250
300


Monthly Charge by Churn Status
MonthlyCharge
Churn
100
150
200
250
300
No
Yes


Churn by Internet Service Type
Churn
No
Yes
Count
4000
3000
2000
1000
0
DSL
Fiber Optic
None
Internet Service

[ ]: