A

1)

Is there a relationship between number of contacts to tech support, and churn?

2)

The stakeholders in the organization will benefit from an analyis of the data because if there is a relationship between these two variables, the organization can take actions to reduce the number of contacts to tech support.

3)

The relevant data used to answer the question in A1 will be the 'Contacts' column and the 'Churn' column from the Churn data set.

B)

1)

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from scipy.stats import ttest_ind
# Replace 'file_path.csv' with the path to your CSV file
df = pd.read_csv('churn_clean.csv')



# Separate number of contacts column into two different groups by churn value
seriesYes = pd.Series(df.loc[df['Churn'] == 'Yes', 'Contacts'])
seriesNo = pd.Series(df.loc[df['Churn'] == 'No', 'Contacts'])
# Perform t-test
t_statistic, p_value = ttest_ind(seriesYes, seriesNo)
```

```
# Print results
print("t-statistic:", t_statistic)
print("p-value:", p_value)
```

```
t-statistic: 0.8566219322168955
p-value: 0.3916743913251065
```

In [ ]:

2)

Null hypotheses = There is no difference in mean of the number of 'Contacts' between the two groups separated by 'Churn'.

Alternative hypotheses = There is a difference in mean of the number of 'Contacts' between the two groups separated by 'Churn'.

Alpha = 0.05

Result = fail to reject null hypotheses. p-value greater than alpha. 0.3916 > 0.05.

There is not a meaningful difference between the mean number of contacts to tech support between two groups separated by 'Churn'.

3)

I chose this analysis technique because I wanted to test the relationship betwen only two groups. Also my predicter variable is continuous. If I wanted to test more than two groups I would use anova test.

C)

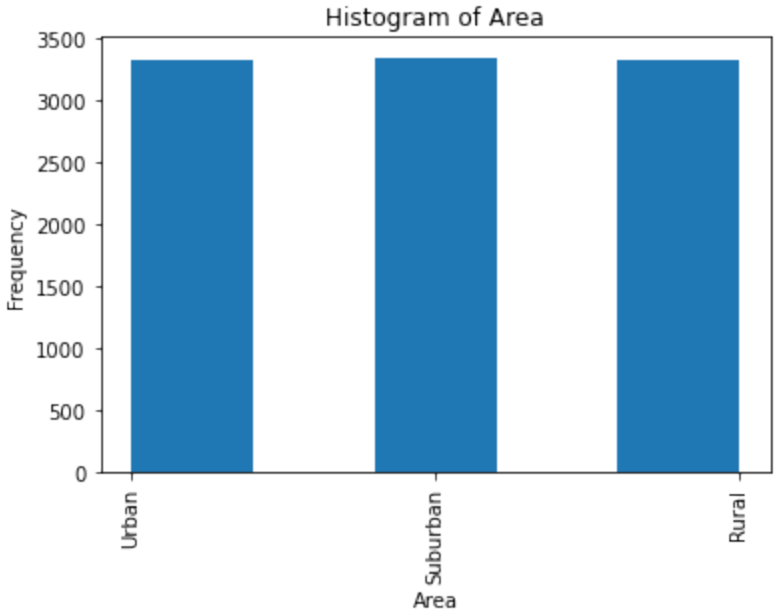
1)

In [2]:

```
#helper functions
#function to plot histogram
def plot_hist(col_name, num_bins, do_rotate=False):
    plt.hist(df[col_name], bins=num_bins)
    plt.xlabel(col_name)
    plt.ylabel('Frequency')
```

```
    plt.title(f'Histogram of {col_name}')
    if do_rotate:
        plt.xticks(rotation=90)
    plt.show()
#function to describe column
def print_desc(col_name):
    print(df[col_name].describe())
```

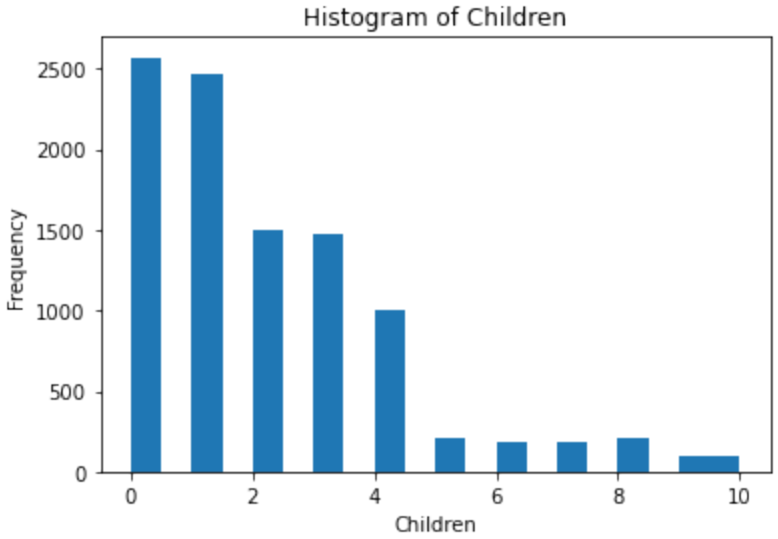
In [3]:
```
#categorical
plot_hist("Area",5,True)
print_desc("Area")
print(df["Area"].mode())
```




```
count        10000
unique           3
top       Suburban
freq          3346
Name: Area, dtype: object
0    Suburban
dtype: object
```

The mode is 'Suburban'

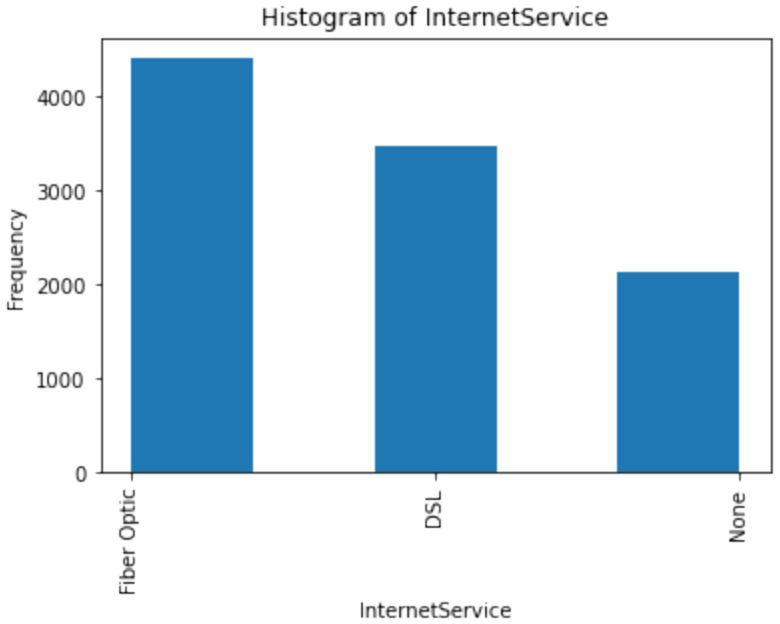
In [4]:
```
#continuous
plot_hist("Children",20)
print_desc("Children")
```




```
count    10000.0000
mean         2.0877
std          2.1472
min          0.0000
25%          0.0000
50%          1.0000
75%          3.0000
max         10.0000
Name: Children, dtype: float64
```

The distribution of 'Children' looks like it is positively skewed.

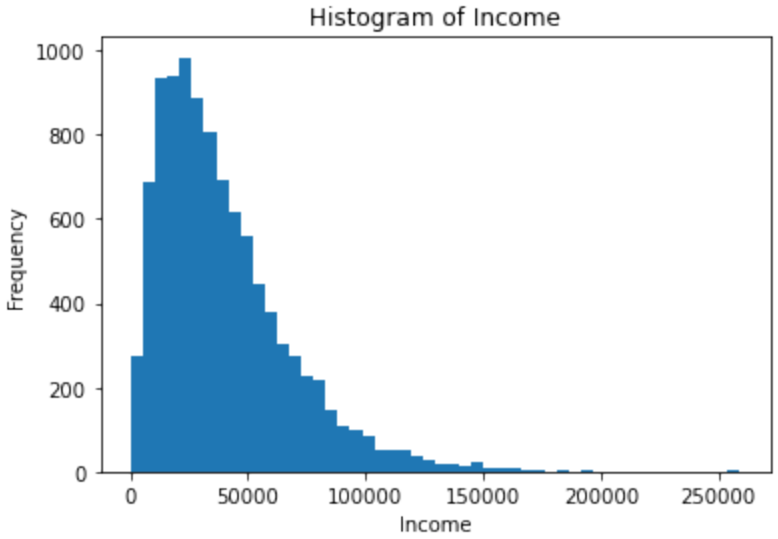
In [5]:
```
#categorical
plot_hist("InternetService",5,True)
print_desc("InternetService")
print(df["InternetService"].mode())
```




```
count           10000
unique              3
top       Fiber Optic
freq             4408
Name: InternetService, dtype: object
0    Fiber Optic
dtype: object
```

The mode of 'InternetService' is Fiber Optic.

In [6]:
```
#continuous.
plot_hist("Income",50)
print_desc("Income")
```



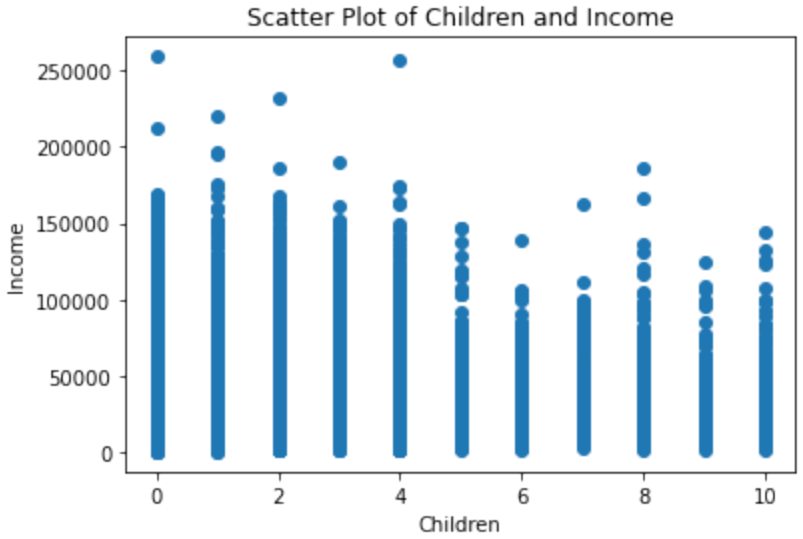

```
count     10000.000000
mean      39806.926771
std       28199.916702
min         348.670000
25%       19224.717500
50%       33170.605000
75%       53246.170000
max      258900.700000
Name: Income, dtype: float64
```

The distribution of 'Income' is positively skewed.

D)

1)

In [18]:
```
plt.scatter(df['Children'], df['Income'])
plt.xlabel('Children')
plt.ylabel('Income')
plt.title('Scatter Plot of Children and Income')
# Show plot
plt.show()
summary_stats = df[['Children', 'Income']].describe()
print(summary_stats)
```



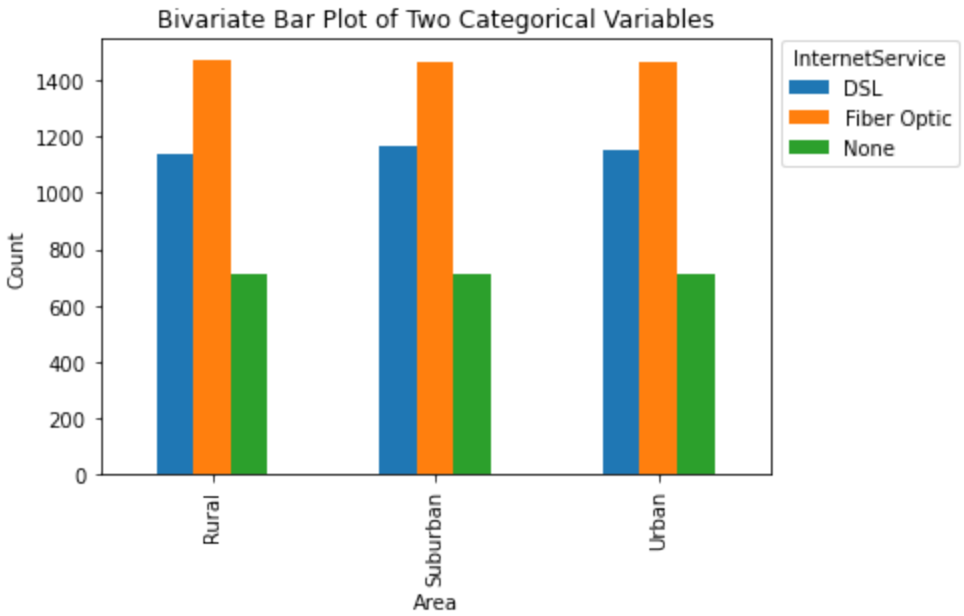
	Children	Income
count	10000.0000	10000.000000
mean	2.0877	39806.926771
std	2.1472	28199.916702
min	0.0000	348.670000
25%	0.0000	19224.717500
50%	1.0000	33170.605000
75%	3.0000	53246.170000
max	10.0000	258900.700000

This distribution looks positively skewed. This graph shows that people with less children have higher income.

In [23]:
```
# create crosstab bar graph
cross_tab = pd.crosstab(df['Area'], df['InternetService'])

# Plot the bivariate bar plot
cross_tab.plot(kind='bar')

# Add labels and title
plt.xlabel('Area')
plt.ylabel('Count')
plt.title('Bivariate Bar Plot of Two Categorical Variables')
plt.legend(title='InternetService', bbox_to_anchor=(1, 1.02),
           loc='upper left')
# Show plot
plt.show()
contingency_table = pd.crosstab(df['Area'], df['InternetService'])
print(contingency_table)
```



InternetService	DSL	Fiber Optic	None
Area			
Rural	1142	1477	708
Suburban	1170	1465	711
Urban	1151	1466	710

By looking at the graph and contingency table we can see that in each 'Area' category that 'Fiber Optic' is the most frequent, followed by 'DSL', and 'None' has the lowest frequency. This pattern repeats across each 'Area' category.

I used this source for the contingency table code.

Gfg. (2024, March 21)

E)

1)

There is no correlation between contacts to tech support and customer churn. I failed to reject the null hypothesis. In my hypothesis test, because the p value was greater than alpha of 0.05, there is no difference in mean contacts to tech support between groups of people who canceled service and those who did not based on churn.

2)

The limits of the data analysis done with hypothesis testing is that the results could be different if we had a larger sample size. Also the t-test works best on data with a normal distribution. Other than that I think it is a pretty good analysis.

3)

Based on the results of the hypothesis test I think the organization should focus their resources on other aspects of their service provison in order to reduce customer churn. This is based on the observation that customers who canceled service or 'Churn' and those who did not, don't have a difference in mean numbers of contacts to tech support or 'Contacts'. In other words 'Contacts' does not seem to be correlated to 'Churn'.

G)

citations

G. (2024, March 21). Contingency Table in Python. GeeksforGeeks. https://www.geeksforgeeks.org/contingency-table-in-python/

GfG (2022) Using pandas crosstab to create a bar plot, GeeksforGeeks. Available at: https://www.geeksforgeeks.org/using-pandas-crosstab-to-create-a-bar-plot/ (Accessed: 07 April 2024).

Barbara Illowsky & OpenStax et al. (no date) Introduction to statistics, Lumen. Available at: https://courses.lumenlearning.com/introstats1/chapter/null-and-alternative-hypotheses/#:~:text=In%20a%20hypothesis%20test%2C%20we,the%20hypothesis%20test%20shows%20otherwise. (Accessed: 07 April 2024).

Residentmario (2018) Bivariate plotting with Pandas, Kaggle. Available at: https://www.kaggle.com/code/residentmario/bivariate-plotting-with-pandas (Accessed: 07 April 2024).

(PDF) http://www.scirp.org/journal/paperinformation.aspx?paperid=19966. (n.d.). https://www.researchgate.net/publication/273338628_httpwwwscirporgjournalPaperInformationaspxPaperID19966

In [ ]: