D208: Predictive Modeling Task 1

Justin Huynh

Student ID: 012229514

M.S. Data Analytics

A1. Research Question

My research question for this project is “What factors most significantly contribute to the monthly charge of the customers?”

A2. Goals

The goals of the data analysis are to identify what key factors most significantly contribute to the monthly charge, use multiple linear regression to understand the relationship between the predictors and the ‘MonthlyCharge’ continuous variable, assess the magnitude of the impact of these factors on monthly charges to understand their importance, and provide actionable insights or recommendations for how to improve pricing strategies to enhance customer retention.

B1. Summary of Assumptions

Here are 4 assumptions of a multiple linear regression model:

- There must be a linear relationship between a dependent variable and an independent variable, which suggests that independent variables are associated with the changes in the dependent variable.
- Observations of the variables are independent of each other, which implies that any errors that may exist within the variables are not correlated with each other.
- The model’s residuals, or variance from the line, are consistent across all values of the independent variables.
- The model’s residuals are also normally distributed, which is required to use multiple linear regression.

B2. Tool Benefits

The 2 benefits of using Python, which I will be using for this project's analysis, are the comprehensive libraries and the visualization capabilities. The libraries such as Pandas, NumPy, and Statsmodels are great at facilitating statistical analyses and making data manipulation more efficient. The visualizations that we can create from libraries such as Matplotlib and Seaborn are great at creating graphs and plots that we need to not only visualize our data but to also identify trends and understand data distributions. This makes communicating results to stakeholders a lot simpler and effective from a technical perspective.

B3. Appropriate Technique

Multiple linear regression is an appropriate technique because of the nature of the dependent variable 'MonthlyCharge' as it is continuous and well-suited for modeling relationships. This technique also provides a way to quantify the impact of each independent variable on the dependent variable. The predictive capabilities also allows us to predict monthly charges based on the values of the independent variables. Lastly, this technique provides statistical tests such as R-squared that helps assess the validity of our findings during our analysis.

C1. Data Cleaning

Here are all the data cleaning goals used in this analysis:

- Remove any missing values to ensure they won't bias the analysis using the isnull().sum() method
- Ensure all data types are correct by verifying we have the correct data types using .info()
- Remove any possible duplicates in our data set using .drop_duplicates() method

- Convert the categorical variables into numerical so they are suited for regression analysis using one-hot encoding

See code attached in WGU_D208_Task_1.ipynb.

C2. Summary Statistics

The dependent variable we'll be analyzing is 'MonthlyCharge' and the continuous independent variables are 'Age', 'Income', 'Bandwidth_GB_Year', 'Outage_sec_perweek', 'Contacts', and 'Yearly_equip_failure'. The categorical variables are 'Gender', 'Marital', 'Techie', 'Contract', 'InternetService', 'PaperlessBilling', and 'PaymentMethod'. These are the relevant variables to answer our research question.

```
In [46]: # summary statistics for the dependent variable and independent continuous variables
summary_stats = df[['MonthlyCharge', 'Age', 'Income', 'Bandwidth_GB_Year', 'Outage_sec_perweek', 'Contacts', 'Yearly
print(summary_stats)
```

```
       MonthlyCharge           Age         Income  Bandwidth_GB_Year  \
count   10000.000000  10000.000000   10000.000000       10000.000000
mean      172.624816     53.078400   39806.926771        3392.341550
std        42.943094     20.698882   28199.916702        2185.294852
min        79.978860     18.000000     348.670000         155.506715
25%       139.979239     35.000000   19224.717500        1236.470827
50%       167.484700     53.000000   33170.605000        3279.536903
75%       200.734725     71.000000   53246.170000        5586.141370
max       290.160419     89.000000  258900.700000        7158.981530

       Outage_sec_perweek      Contacts  Yearly_equip_failure
count        10000.000000  10000.000000          10000.000000
mean            10.001848      0.994200              0.398000
std              2.976019      0.988466              0.635953
min              0.099747      0.000000              0.000000
25%              8.018214      0.000000              0.000000
50%             10.018560      1.000000              0.000000
75%             11.969485      2.000000              1.000000
max             21.207230      7.000000              6.000000
```

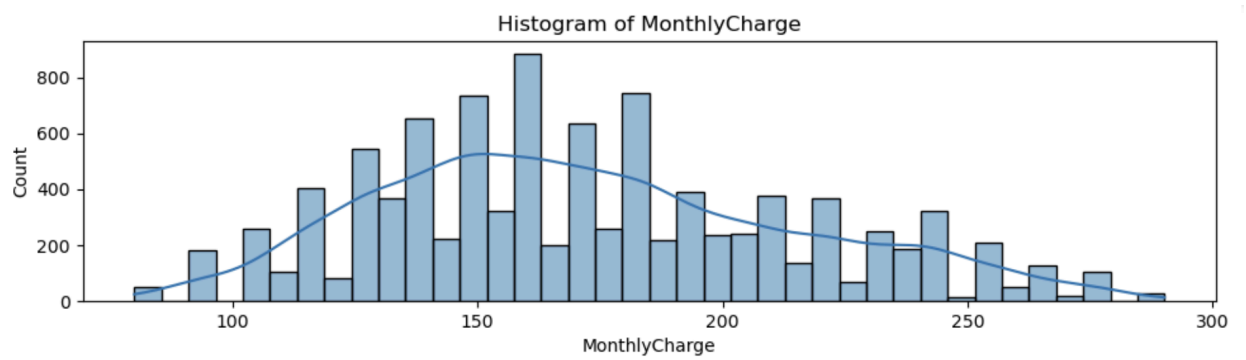
```
In [48]: # summary statistics for categorical variables
categorical_summary_stats = df[original_categorical_vars].describe()
print(categorical_summary_stats)
```

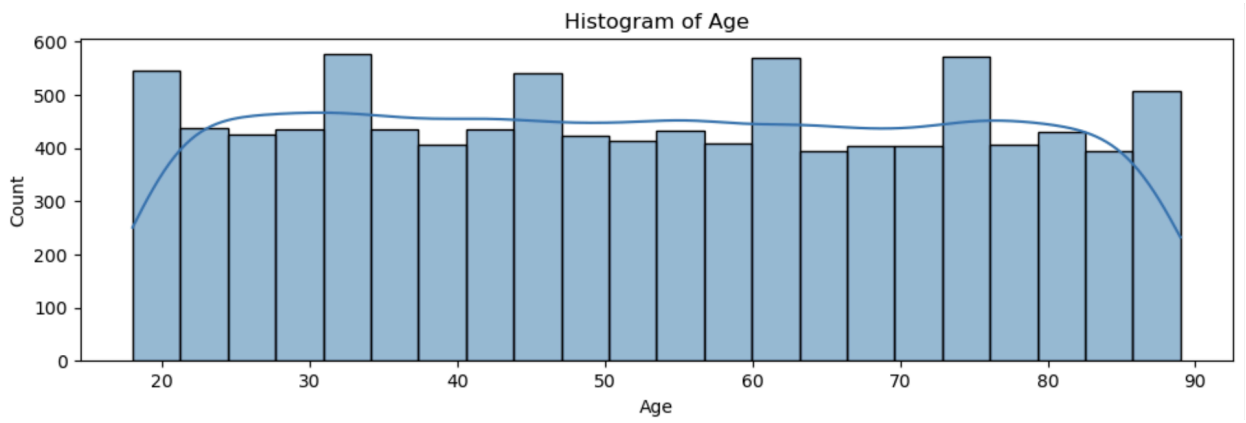
```
        Gender   Marital Techie        Contract InternetService  \
count    10000     10000  10000           10000           10000
unique       3         5      2               3               3
top     Female  Divorced     No  Month-to-month     Fiber Optic
freq      5025      2092   8321            5456            4408

       PaperlessBilling     PaymentMethod
count             10000             10000
unique                2                 4
top                 Yes  Electronic Check
freq               5882              3398
```

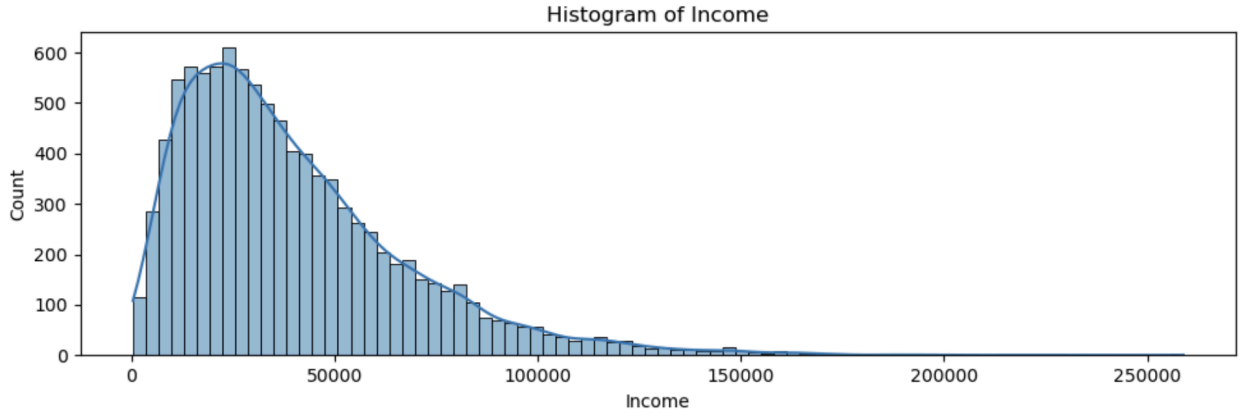
See code attached in WGU_D208_Task_1.ipynb.

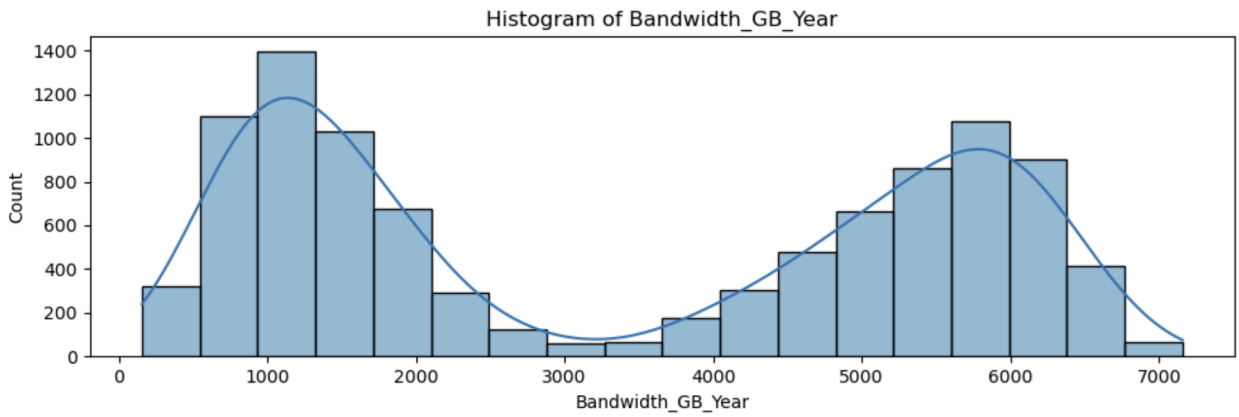
C3. Visualizations

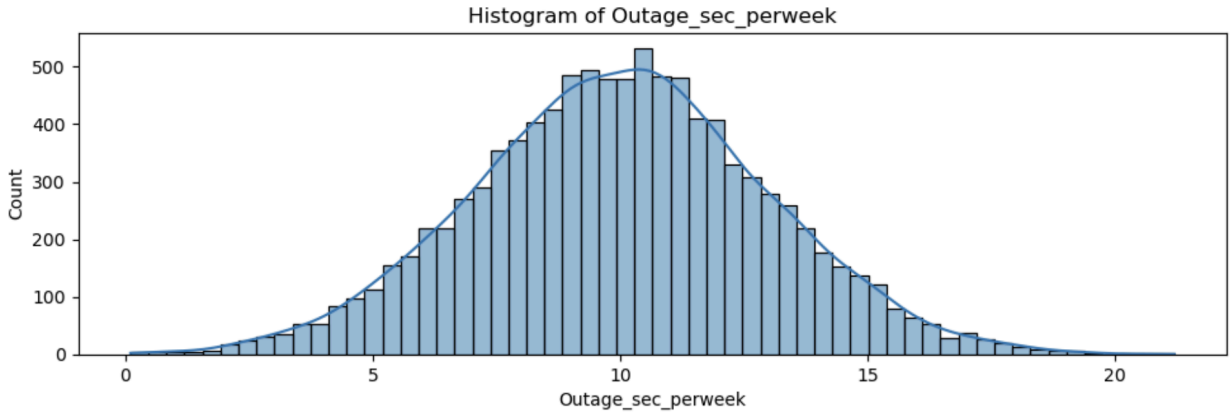
Univariate Visualization

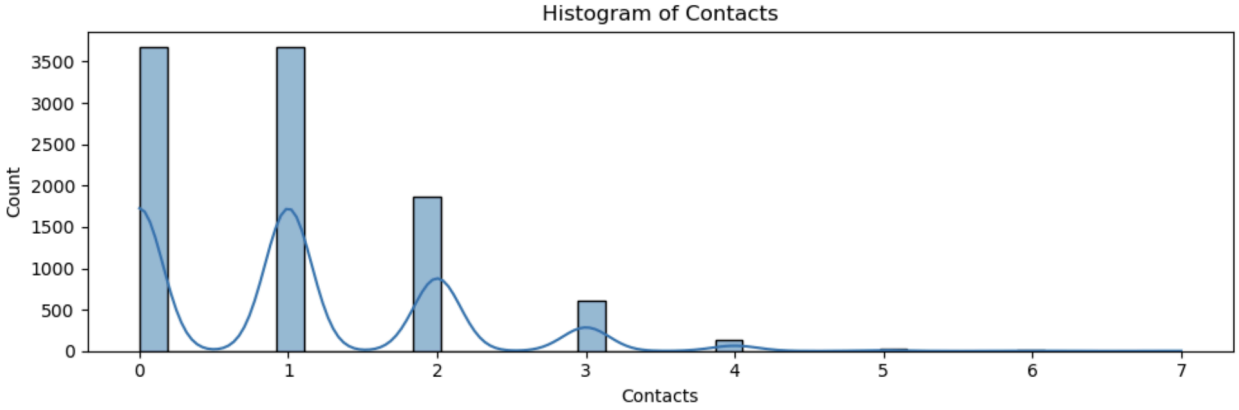


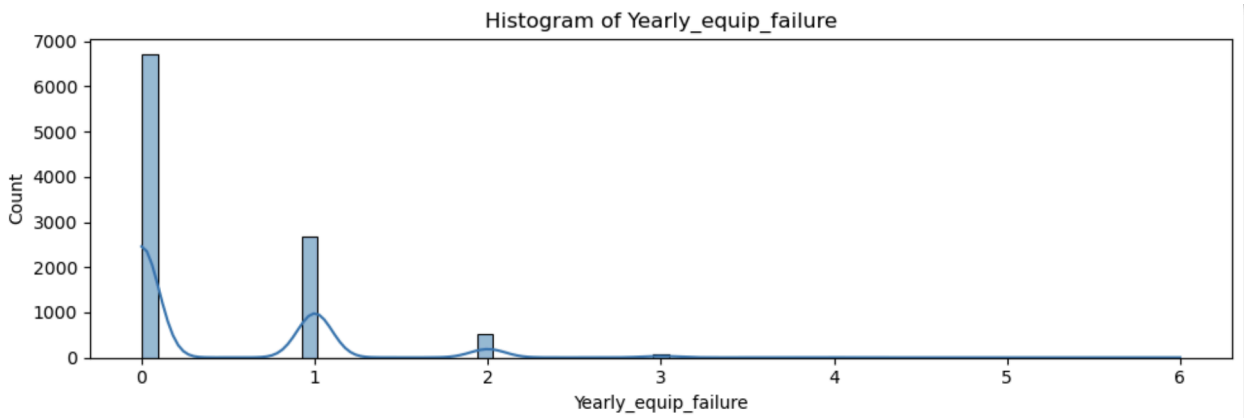


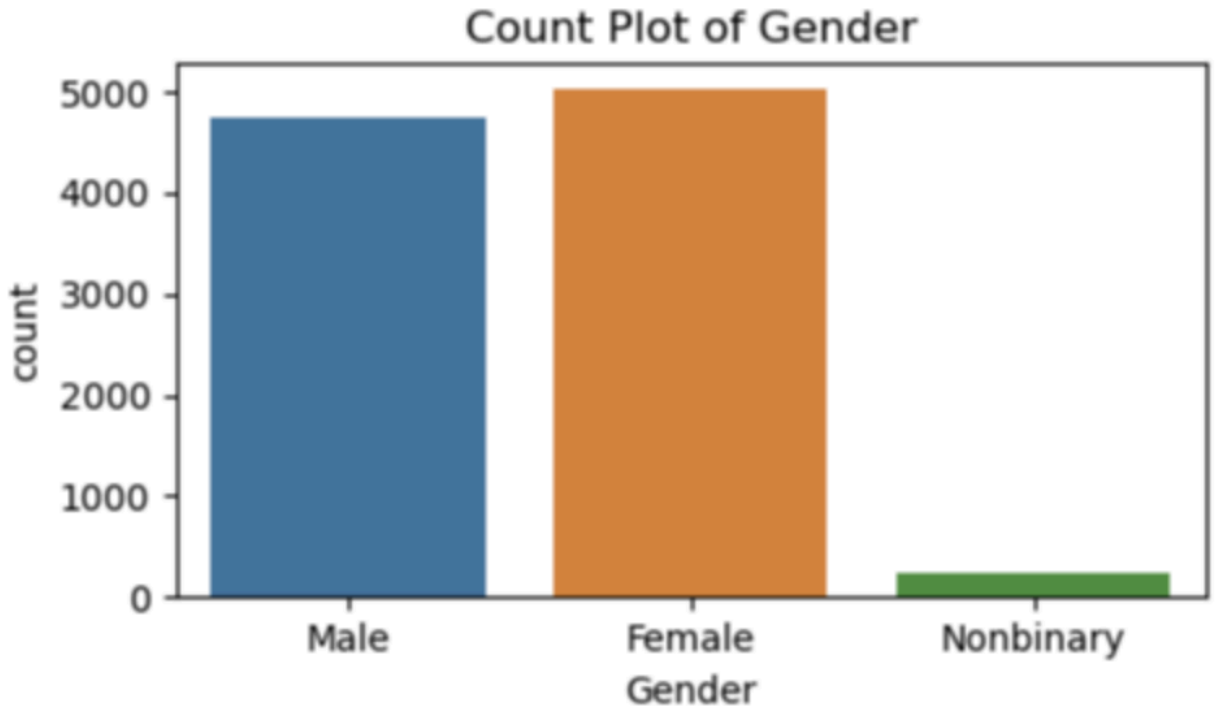


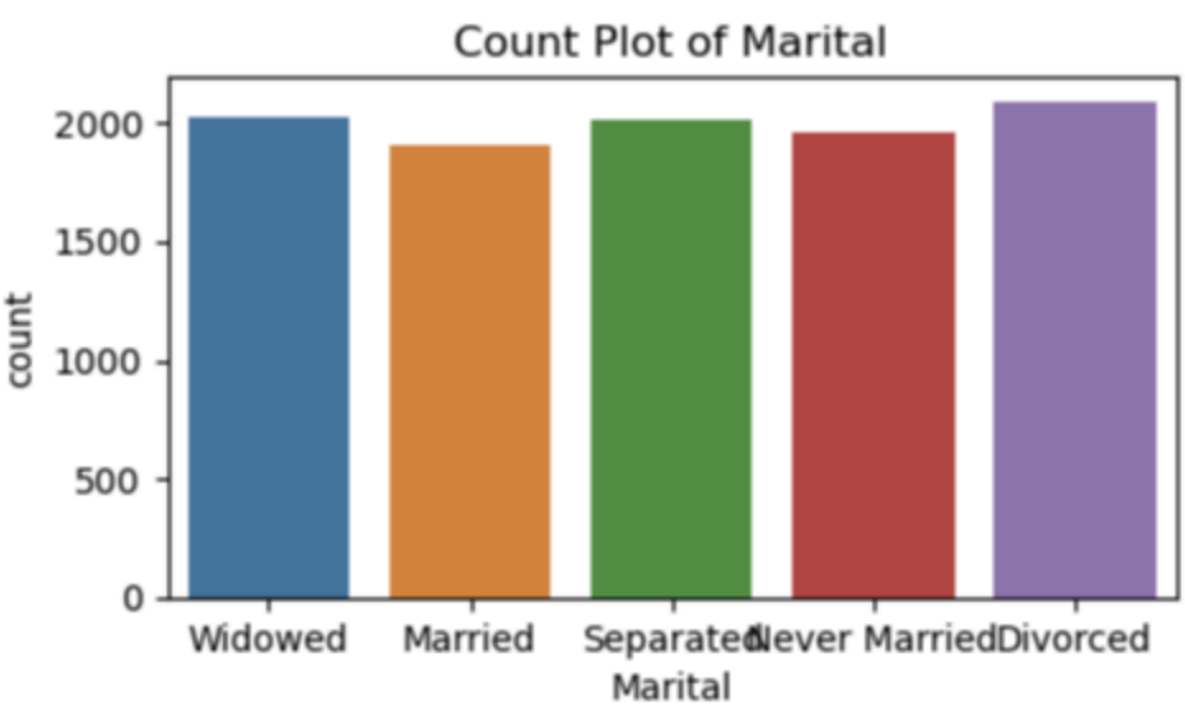

Histogram of Bandwidth_GB_Year
Count
Bandwidth_GB_Year
0
200
400
600
800
1000
1200
1400
0
1000
2000
3000
4000
5000
6000
7000

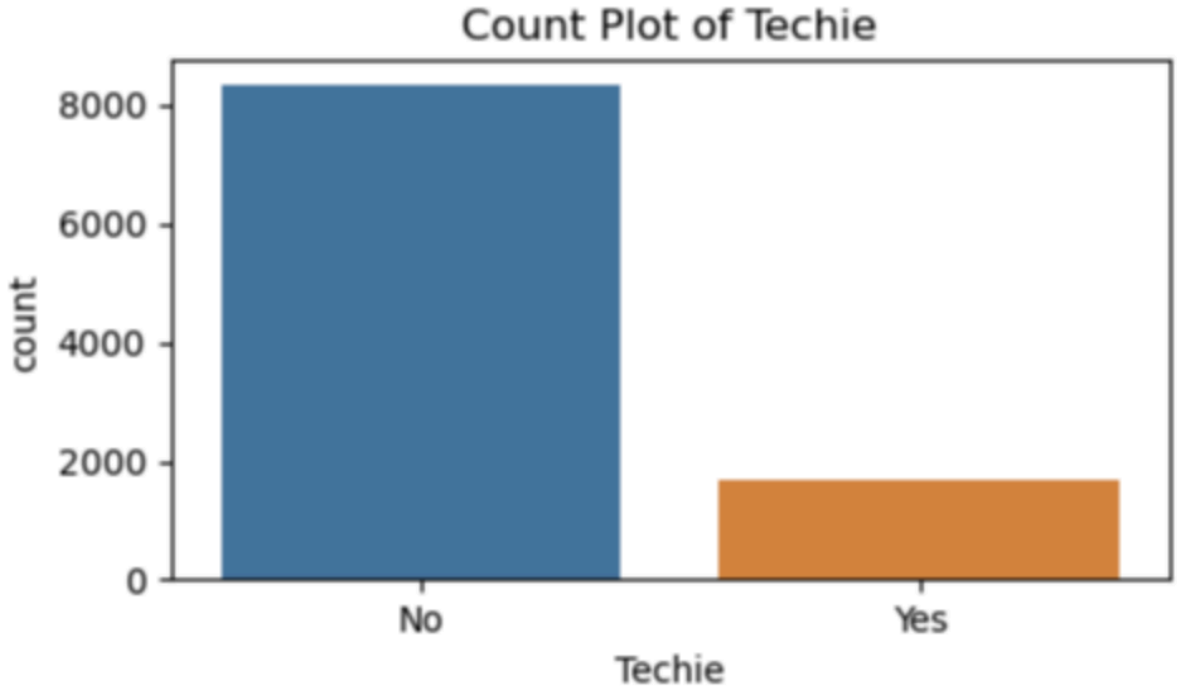

Histogram of Outage_sec_perweek
Count
Outage_sec_perweek
0
100
200
300
400
500
0
5
10
15
20

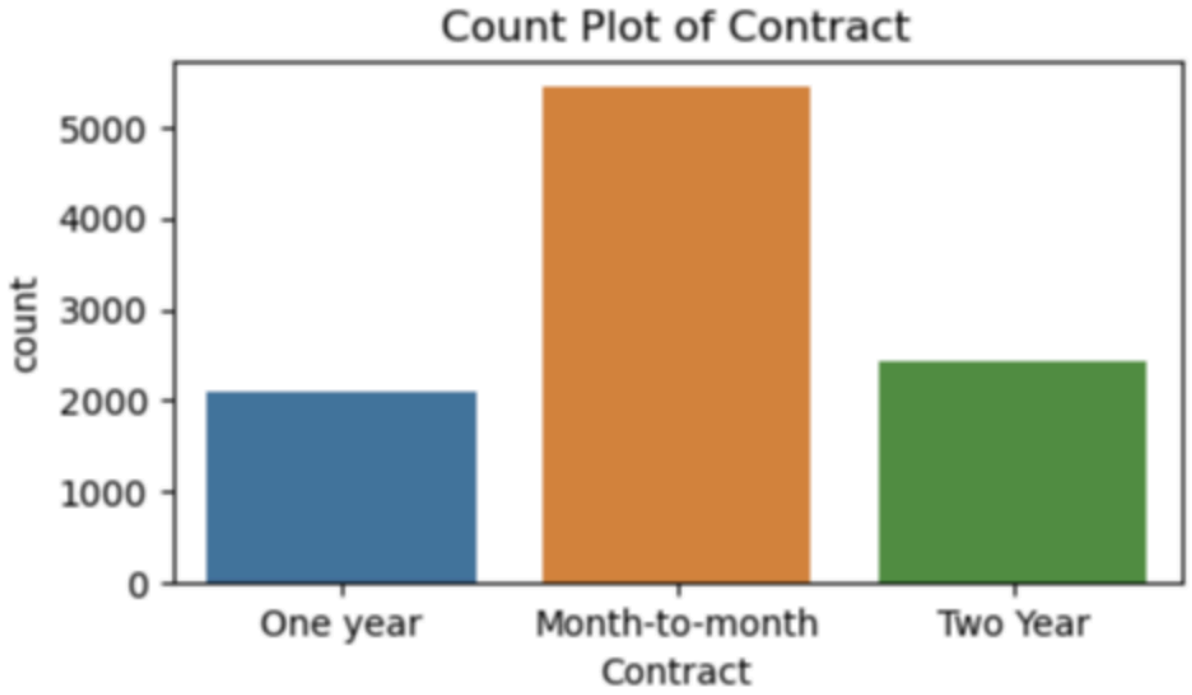

Histogram of Contacts
Count
Contacts
0
500
1000
1500
2000
2500
3000
3500
0
1
2
3
4
5
6
7

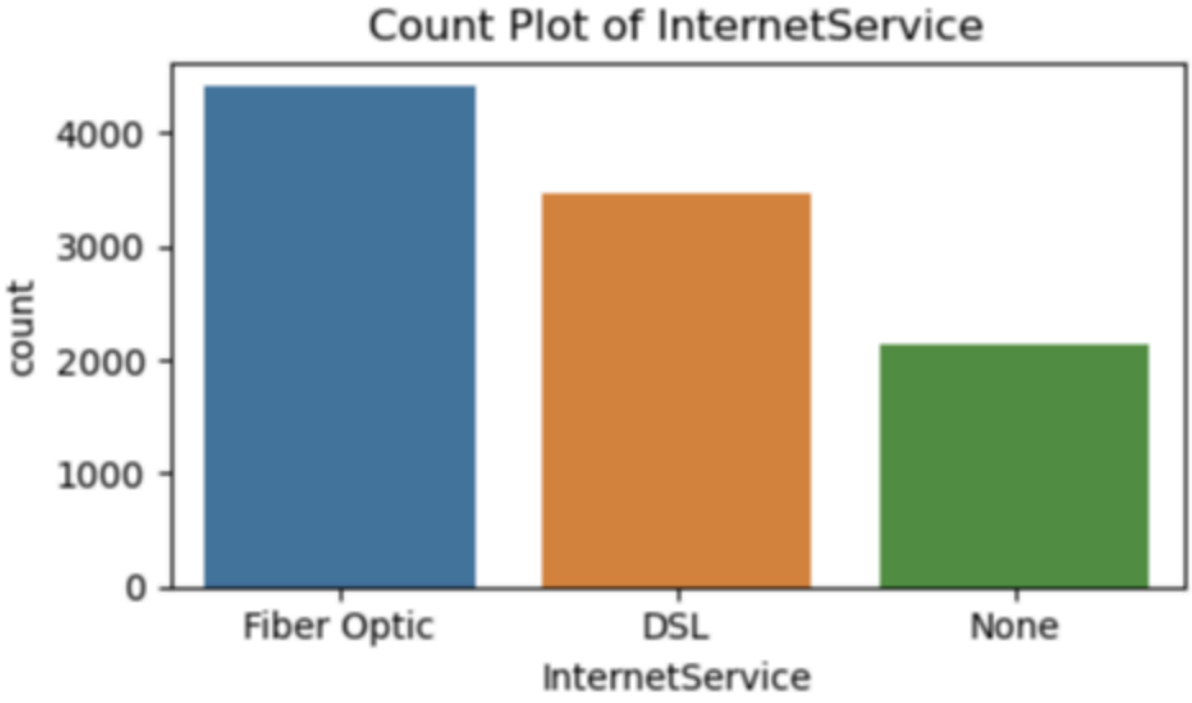

Histogram of Yearly_equip_failure
Count
Yearly_equip_failure
0
1000
2000
3000
4000
5000
6000
7000
0
1
2
3
4
5
6

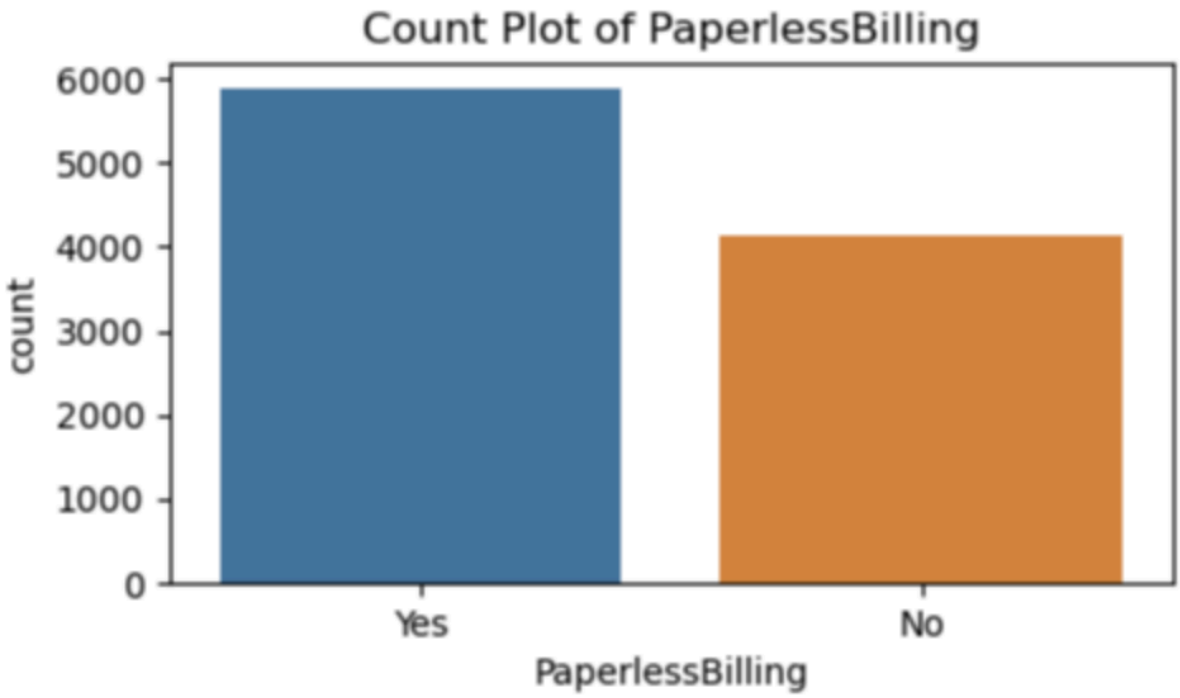

Count Plot of Gender
count
Gender
0
1000
2000
3000
4000
5000
Male
Female
Nonbinary

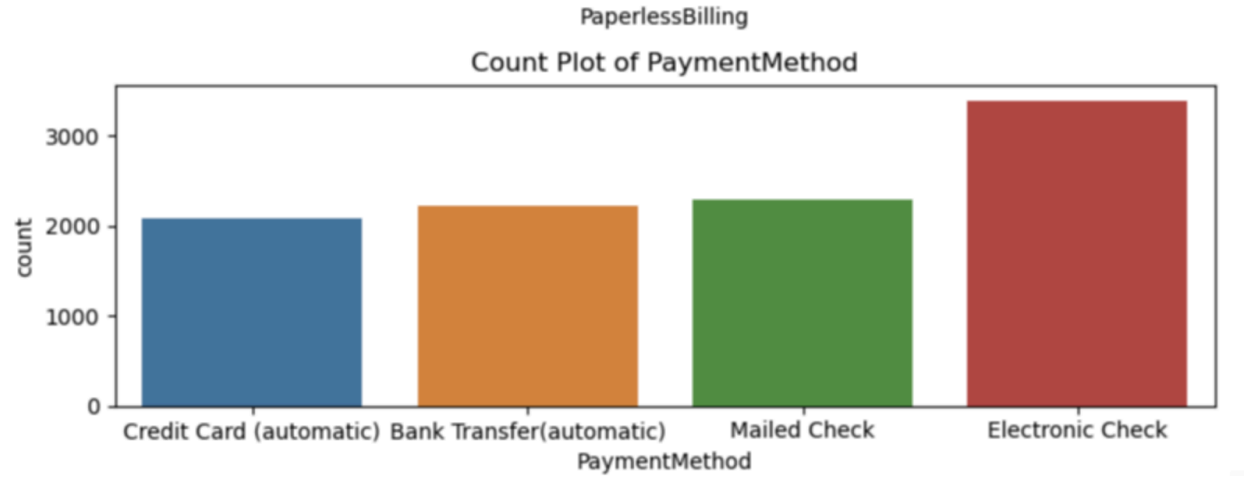

Count Plot of Marital
count
2000
1500
1000
500
0
Widowed
Married
Separated
Never Married
Divorced
Marital


Count Plot of Techie
count
8000
6000
4000
2000
0
No
Yes
Techie

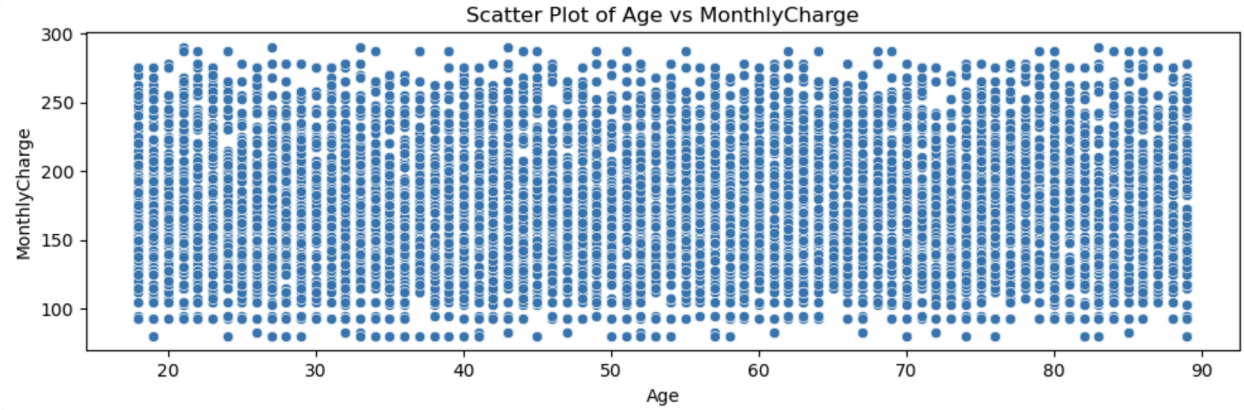

Count Plot of Contract
count
5000
4000
3000
2000
1000
0
One year
Month-to-month
Two Year
Contract

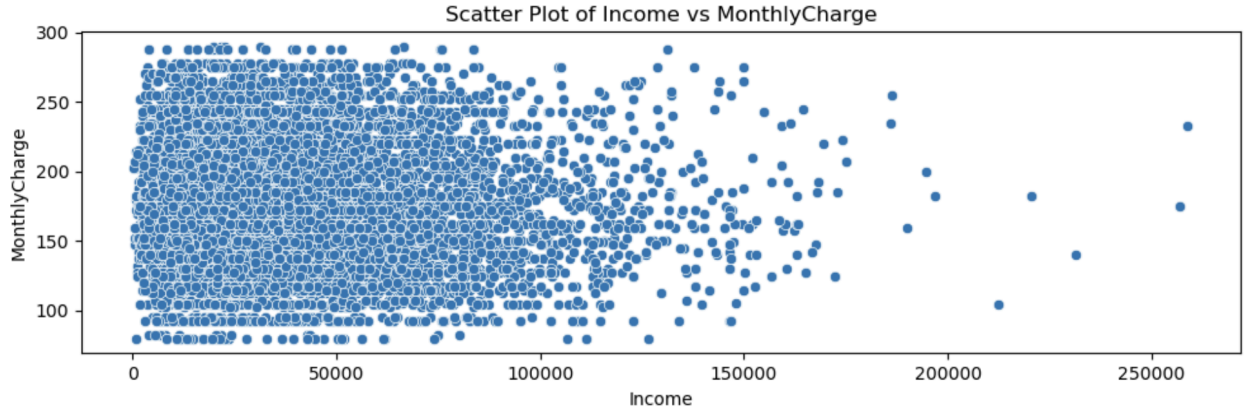

Count Plot of InternetService
count
4000
3000
2000
1000
0
Fiber Optic
DSL
None
InternetService

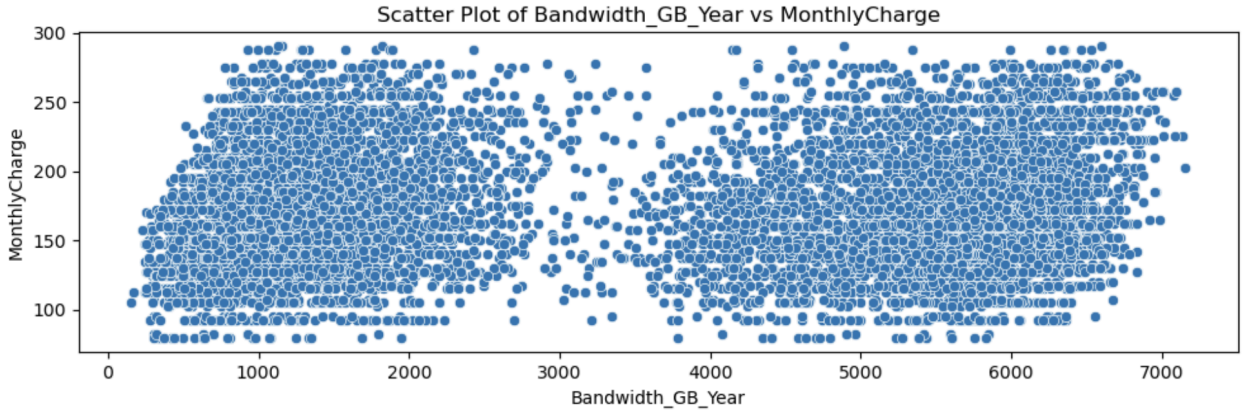


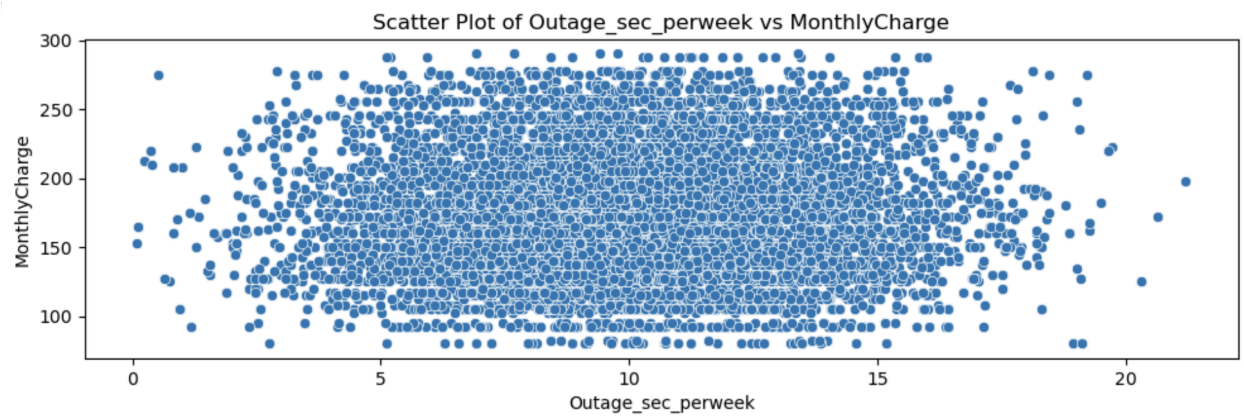


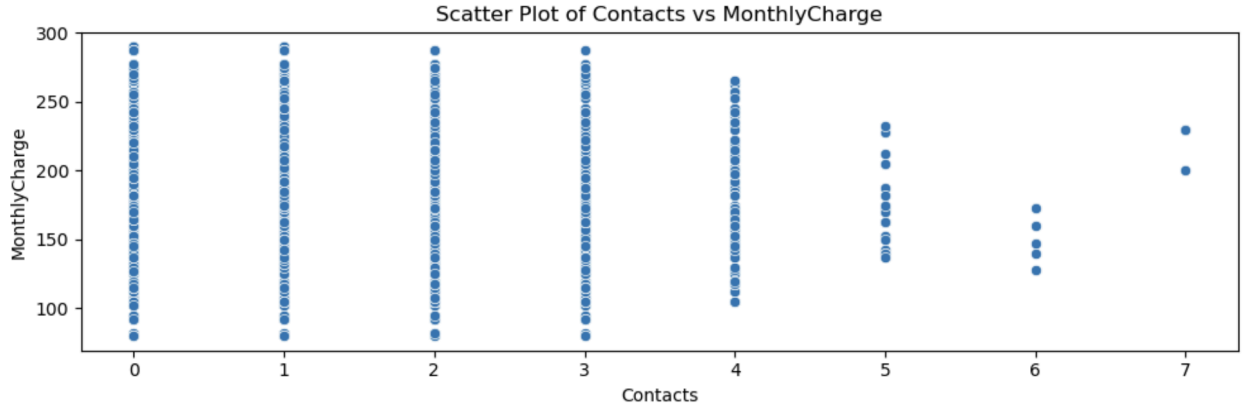
Bivariate Visualization

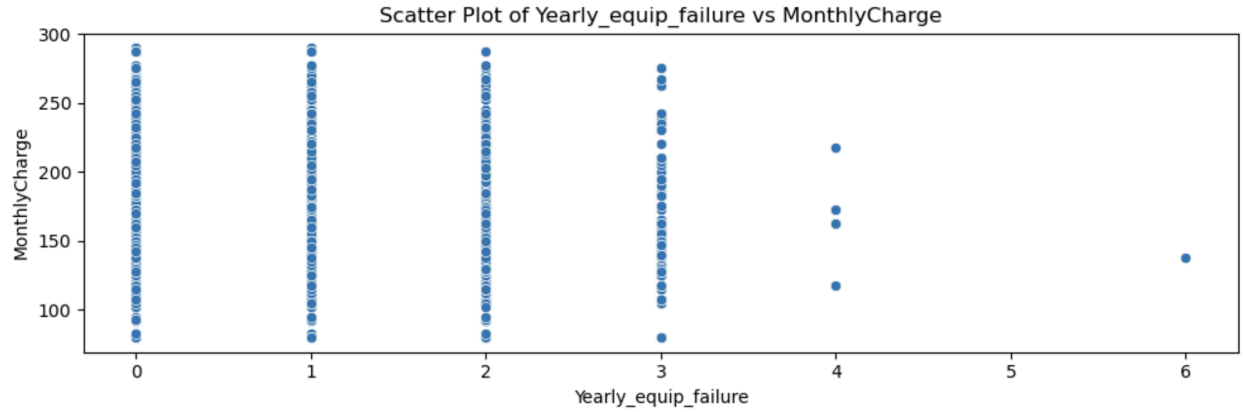

Scatter Plot of Age vs MonthlyCharge
MonthlyCharge
Age
300
250
200
150
100
20
30
40
50
60
70
80
90

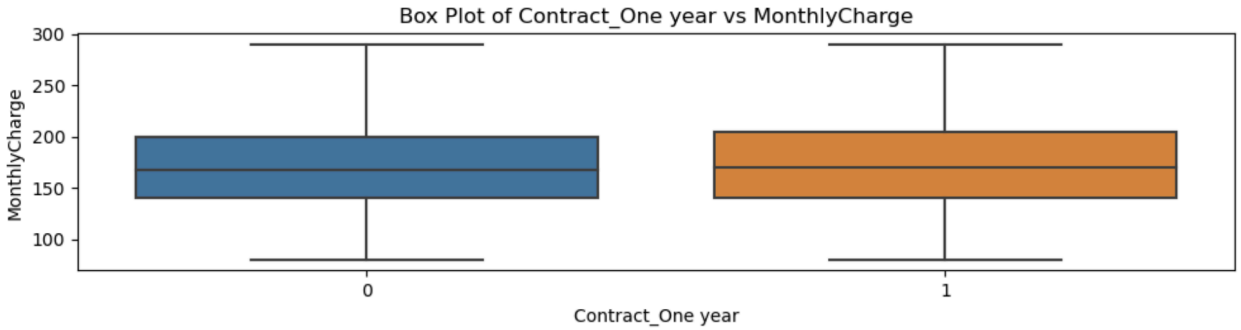

Scatter Plot of Income vs MonthlyCharge
MonthlyCharge
Income
300
250
200
150
100
0
50000
100000
150000
200000
250000

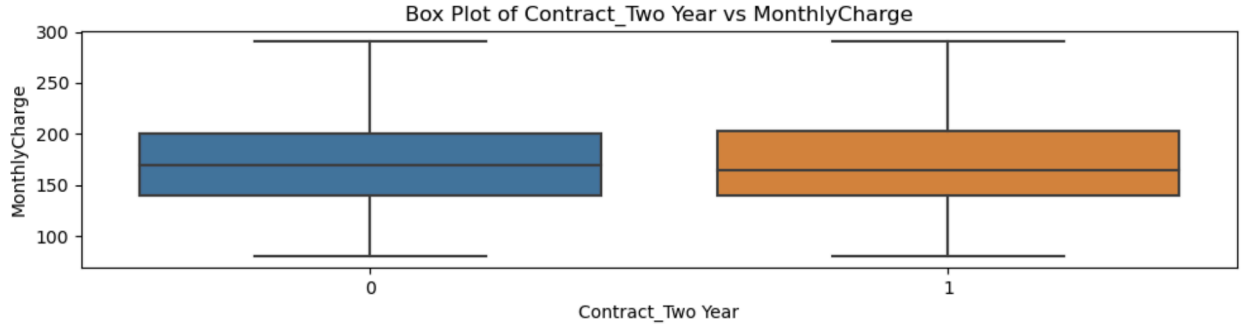

Scatter Plot of Bandwidth_GB_Year vs MonthlyCharge
MonthlyCharge
Bandwidth_GB_Year
300
250
200
150
100
0
1000
2000
3000
4000
5000
6000
7000

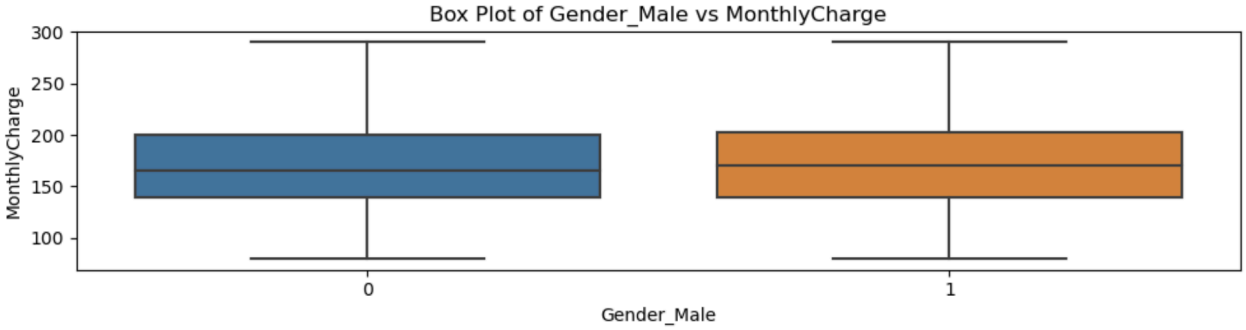

Scatter Plot of Outage_sec_perweek vs MonthlyCharge
MonthlyCharge
Outage_sec_perweek
300
250
200
150
100
0
5
10
15
20

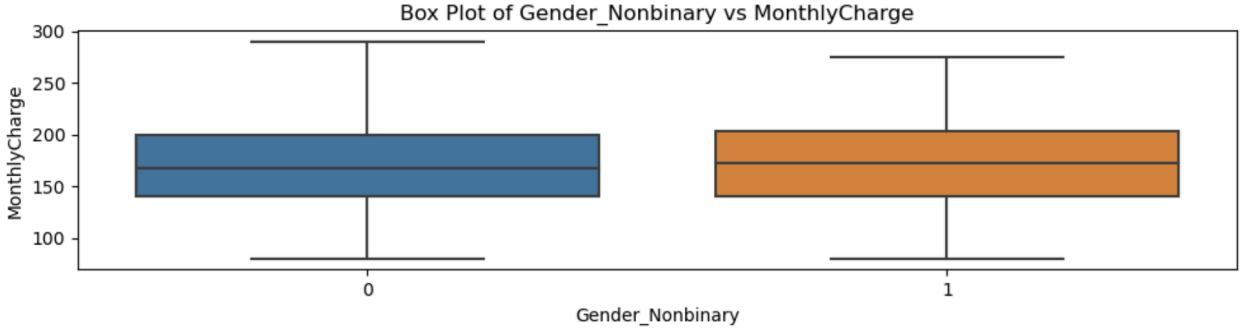

Scatter Plot of Contacts vs MonthlyCharge
MonthlyCharge
Contacts
300
250
200
150
100
0
1
2
3
4
5
6
7

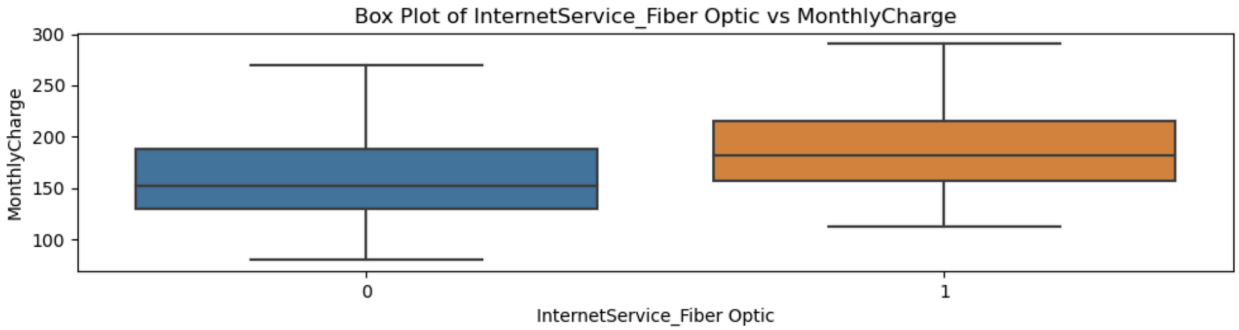

Scatter Plot of Yearly_equip_failure vs MonthlyCharge
MonthlyCharge
Yearly_equip_failure
300
250
200
150
100
0
1
2
3
4
5
6

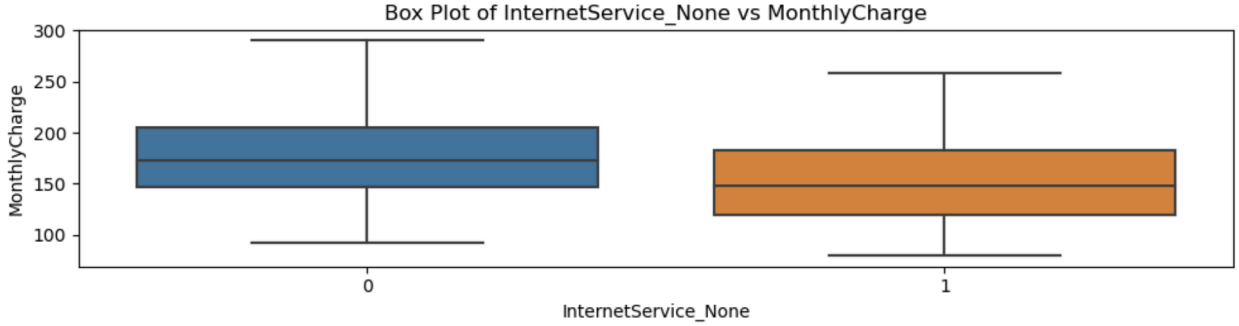

Box Plot of Contract_One year vs MonthlyCharge
MonthlyCharge
300
250
200
150
100
0
1
Contract_One year

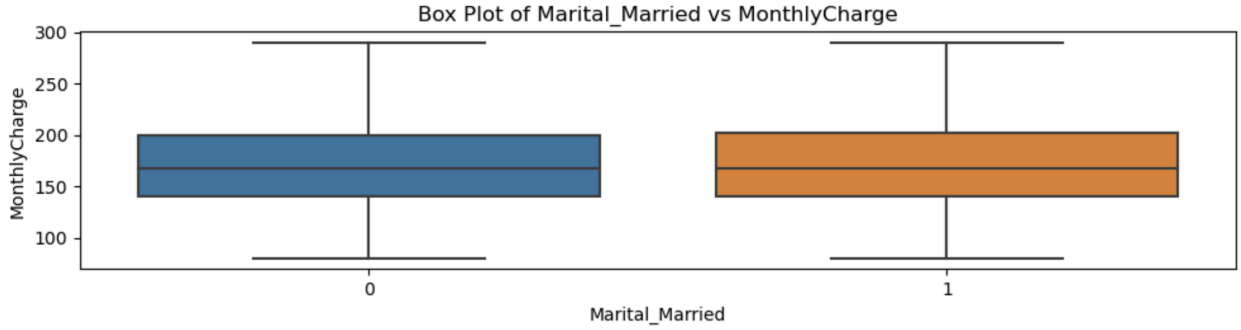

Box Plot of Contract_Two Year vs MonthlyCharge
MonthlyCharge
300
250
200
150
100
0
1
Contract_Two Year

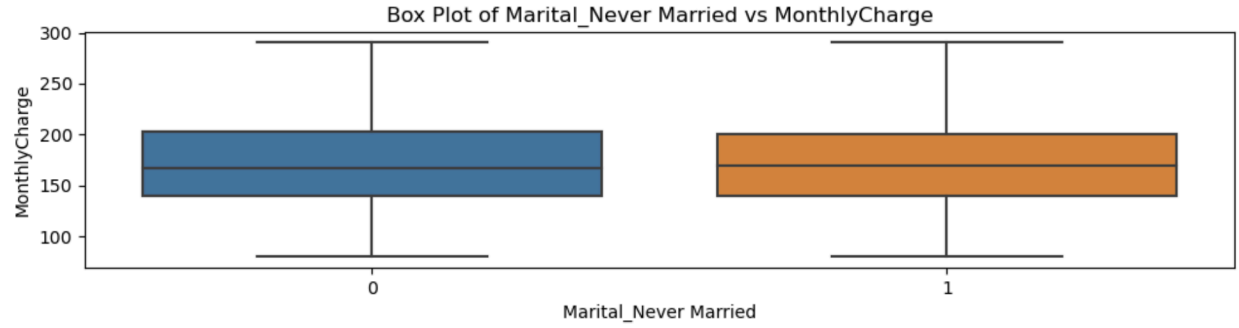

Box Plot of Gender_Male vs MonthlyCharge
MonthlyCharge
300
250
200
150
100
0
1
Gender_Male

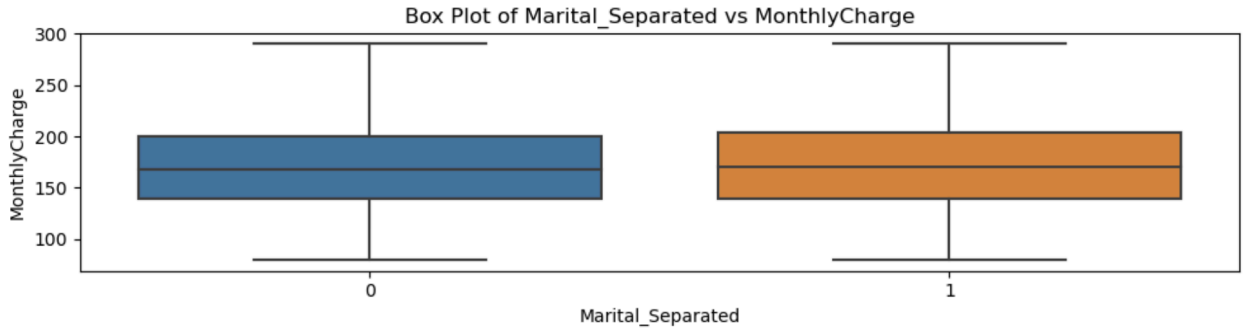

Box Plot of Gender_Nonbinary vs MonthlyCharge
MonthlyCharge
300
250
200
150
100
0
1
Gender_Nonbinary

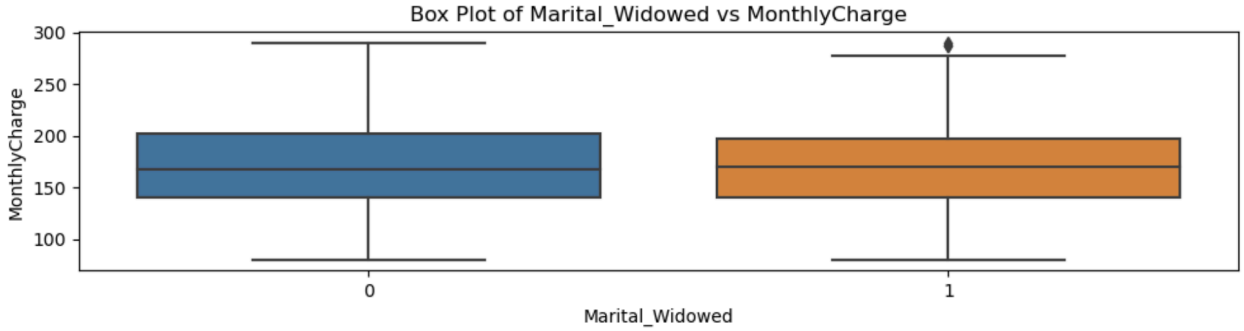

Box Plot of InternetService_Fiber Optic vs MonthlyCharge
MonthlyCharge
300
250
200
150
100
0
1
InternetService_Fiber Optic

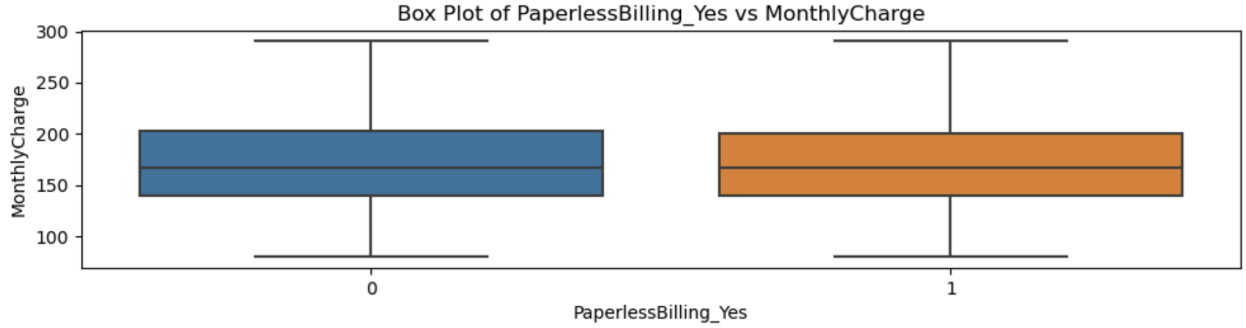

Box Plot of InternetService_None vs MonthlyCharge
MonthlyCharge
300
250
200
150
100
0
1
InternetService_None

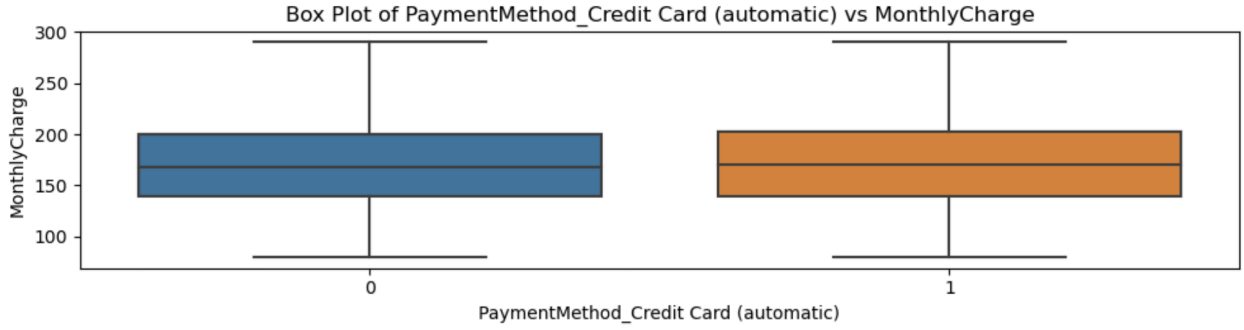

Box Plot of Marital_Married vs MonthlyCharge
MonthlyCharge
300
250
200
150
100
0
1
Marital_Married

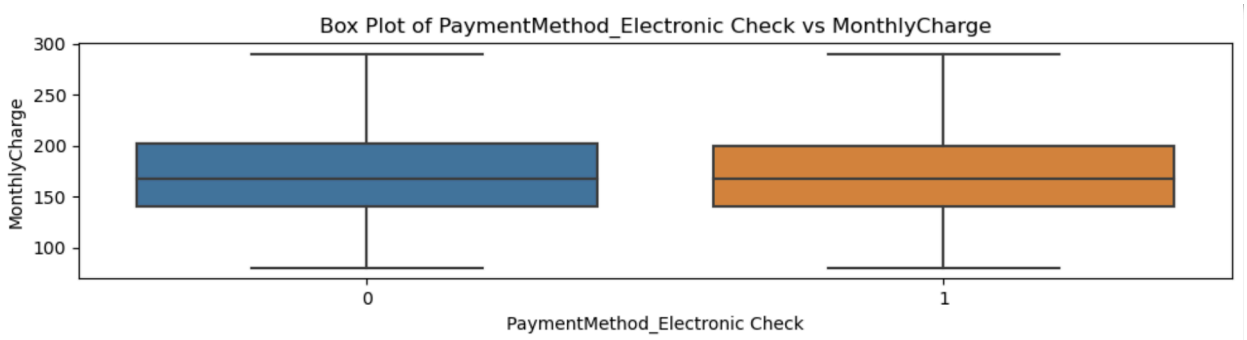

Box Plot of Marital_Never Married vs MonthlyCharge
MonthlyCharge
300
250
200
150
100
0
1
Marital_Never Married

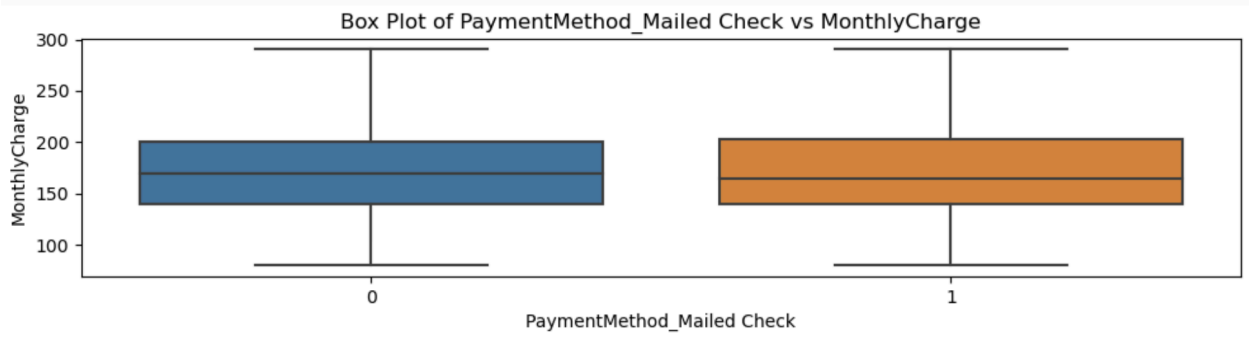

Box Plot of Marital_Separated vs MonthlyCharge
MonthlyCharge
300
250
200
150
100
0
1
Marital_Separated

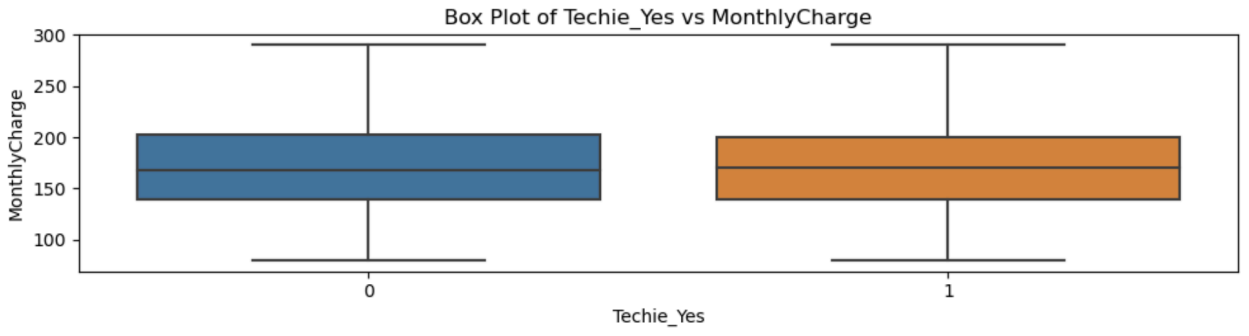

Box Plot of Marital_Widowed vs MonthlyCharge
MonthlyCharge
300
250
200
150
100
0
1
Marital_Widowed


Box Plot of PaperlessBilling_Yes vs MonthlyCharge
MonthlyCharge
300
250
200
150
100
0
1
PaperlessBilling_Yes


Box Plot of PaymentMethod_Credit Card (automatic) vs MonthlyCharge
MonthlyCharge
300
250
200
150
100
0
1
PaymentMethod_Credit Card (automatic)







See code attached in WGU_D208_Task_1.ipynb.

C4: Data Transformation

The data transformation process first prepares the data by handling any missing values to ensure our data from the relevant variables are clean and prepared. Then we convert the categorical values into numerical format by using one-hot encoding as our next step. We then ensure all the data types are correct so they can be inputted into our regression model. Lastly, we'll generate a summary statistic for the encoded variables. In summary, the steps we'll take are loading the data -> select relevant variables ->

handle any missing values -> encode the categorical variables -> generate summary statistics for encoded variables.

See code attached in WGU_D208_Task_1.ipynb.

C5: Prepared Data Set

A copy of the fully prepared data set will be submitted as ‘prepared_data.csv’.

D1:Initial Model

OLS Regression Results

Dep. Variable:	MonthlyCharge	R-squared:	0.102
Model:	OLS	Adj. R-squared:	0.100
Method:	Least Squares	F-statistic:	54.18
Date:	Sat, 15 Jun 2024	Prob (F-statistic):	5.28e-215
Time:	11:16:19	Log-Likelihood:	-51248.
No. Observations:	10000	AIC:	1.025e+05
Df Residuals:	9978	BIC:	1.027e+05
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	157.5819	2.508	62.822	0.000	152.665	162.499
Age	0.0255	0.020	1.292	0.196	-0.013	0.064
Income	1.775e-06	1.45e-05	0.123	0.902	-2.66e-05	3.01e-05
Bandwidth_GB_Year	0.0014	0.000	7.267	0.000	0.001	0.002
Outage_sec_perweek	0.2901	0.137	2.117	0.034	0.022	0.559
Contacts	0.0191	0.412	0.046	0.963	-0.789	0.827
Yearly_equip_failure	-0.4991	0.641	-0.778	0.436	-1.756	0.758
Gender_Male	0.9169	0.825	1.111	0.267	-0.701	2.535
Gender_Nonbinary	1.5643	2.743	0.570	0.568	-3.813	6.941
Marital_Married	-1.0012	1.290	-0.776	0.438	-3.529	1.526
Marital_Never Married	-0.7097	1.283	-0.553	0.580	-3.224	1.804
Marital_Separated	0.3558	1.272	0.280	0.780	-2.138	2.850
Marital_Widowed	-0.9493	1.271	-0.747	0.455	-3.440	1.542
Techie_Yes	0.7941	1.091	0.728	0.467	-1.344	2.932
Contract_One year	2.0284	1.046	1.939	0.053	-0.022	4.079
Contract_Two Year	0.2845	0.992	0.287	0.774	-1.661	2.230
InternetService_Fiber Optic	20.0585	0.930	21.573	0.000	18.236	21.881
InternetService_None	-13.4176	1.126	-11.912	0.000	-15.626	-11.210
PaperlessBilling_Yes	-0.1575	0.829	-0.190	0.849	-1.782	1.467
PaymentMethod_Credit Card (automatic)	0.1644	1.242	0.132	0.895	-2.271	2.599
PaymentMethod_Electronic Check	-0.5861	1.111	-0.528	0.598	-2.764	1.592
PaymentMethod_Mailed Check	-0.4382	1.214	-0.361	0.718	-2.817	1.941

```
==============================================================================
Omnibus:                      514.348   Durbin-Watson:                   1.959
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              456.891
Skew:                           0.459   Prob(JB):                    6.13e-100
Kurtosis:                       2.497   Cond. No.                     3.38e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.38e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

See code attached in WGU_D208_Task_1.ipynb.

D2: Justification of Model Reduction

To align with our research question, we're going to use backward elimination for feature selection. This method will test the model and iteratively remove the least significant variables, which are the ones with the highest p-value that's greater than the significance level of 0.05. This process is repeated until all variables are statistically significant. By maintaining this process of removing p-values > 0.05, we ensure the model only has statistically significant variables retained. Backward elimination also helps maintain model simplicity by reducing the number of predictors, making it easier to interpret. Lastly, the Adjusted R-square value helps validate how good the fit of the model is, adjusting for the number of predictors.

D3: Reduced Linear Regression Model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          MonthlyCharge   R-squared:                       0.101
Model:                            OLS   Adj. R-squared:                  0.101
Method:                 Least Squares   F-statistic:                     282.0
Date:                Sat, 15 Jun 2024   Prob (F-statistic):          3.96e-230
Time:                        11:33:34   Log-Likelihood:                -51253.
No. Observations:               10000   AIC:                         1.025e+05
Df Residuals:                    9995   BIC:                         1.026e+05
Df Model:                           4
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          159.1447      1.675     95.025      0.000     155.862     162.428
Bandwidth_GB_Year                0.0014      0.000      7.233      0.000       0.001       0.002
Outage_sec_perweek               0.2908      0.137      2.125      0.034       0.023       0.559
InternetService_Fiber Optic     20.0389      0.929     21.579      0.000      18.219      21.859
InternetService_None           -13.4241      1.125    -11.930      0.000     -15.630     -11.218
==============================================================================
Omnibus:                      513.679   Durbin-Watson:                   1.960
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              459.230
Skew:                           0.462   Prob(JB):                    1.90e-100
Kurtosis:                       2.500   Cond. No.                     1.76e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.76e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

See code attached in WGU_D208_Task_1.ipynb.

E1: Model Comparison

The data analysis process for this research question involves constructing an initial multiple linear regression model with all the identified independent variables relevant to our research question. This is followed by reducing the model via backward elimination to retain only the statistically significant variables, which are determined by their respective p-values. When employing backward elimination, we're iteratively removing the least significant variables, p-values > 0.05, to ensure the final model only carries statistically significant variables. Once that is completed, the model evaluation metric we're using is the Adjusted R-squared. This metric adjusts the R-squared value based on the number of predictors in the model. It's used to evaluate how good the fit of the model is by adjusting for the number of predictors and measures how much the

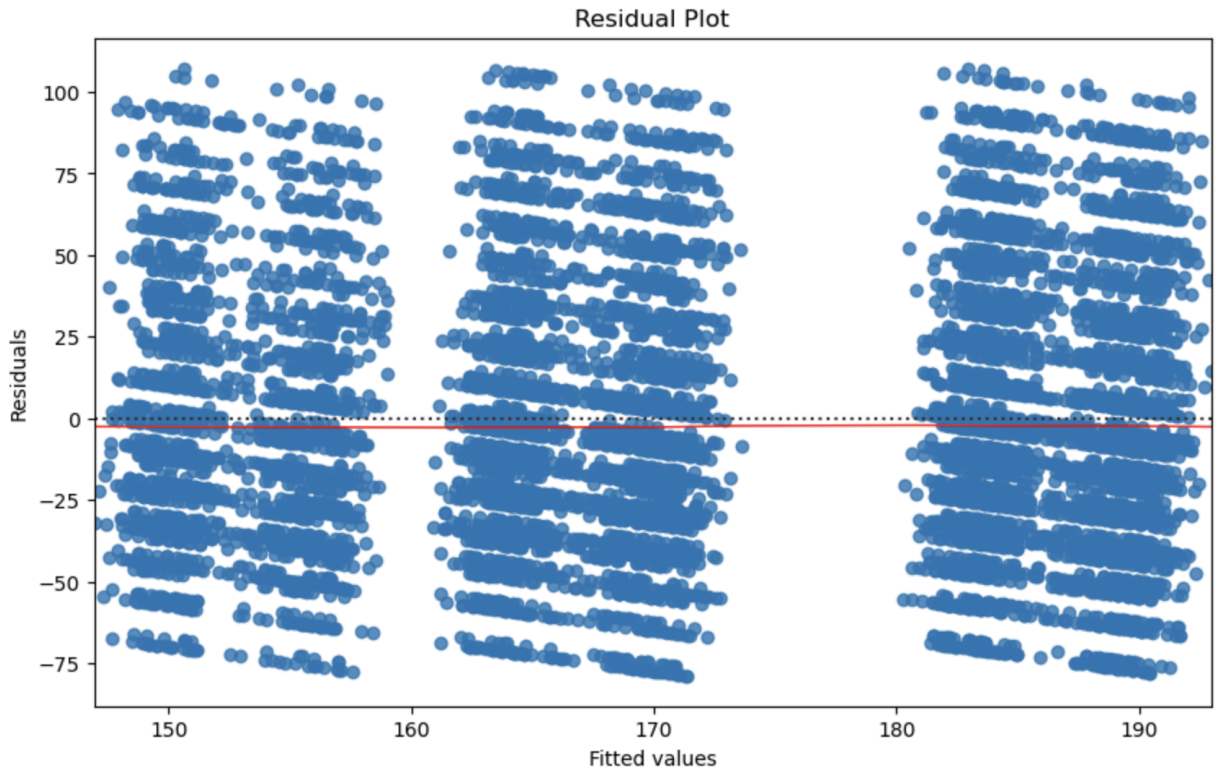
independent variables affect the variances in the dependent variable, which gives us more accuracy in our model. In the initial model, the R-squared value is 0.1024 and the Adjusted R-squared is 0.1005. The reduced model metrics have an R-squared value of 0.1014 and Adjusted R-squared of 0.1011. The R-squared value means the initial value has a 10.24% variance in 'MonthlyCharge' that can be explained by the significant independent variables vs. the reduced value having a 10.14% variance. The reduction in R-squared is minimal, indicating the excluded variables did not contribute significantly to the model's explanatory power. The Adjusted R-squared, however, has an initial value of 10.05% vs the reduced value of 10.11%. The slight increase of the reduced value suggests that the reduced model, despite having fewer variables, has a better fit when adjusting for the number of predictors. This also suggests that the removed variables were not only insignificant, but potentially adding noise to the model.

E2: Output and Calculations

In [66]:
```
# define dependent variable
y = df_encoded['MonthlyCharge']

# define the relevant independent variables from C2
X = df_encoded.drop(columns=['MonthlyCharge'])

# add a constant to the model (intercept)
X = sm.add_constant(X)

# fit the initial regression model
initial_model = sm.OLS(y, X).fit()

# display initial model summary
print(initial_model.summary())
```

In [67]:
```python
# perform backward elimination for a more iterative approach
def backward_elimination(X, y, significance_level=0.05):
    while True:
        # fit the model
        model = sm.OLS(y, X).fit()

        # get p-values
        p_values = model.pvalues

        # get max p-value
        max_p_value = p_values.max()

        # check if the max p-value is greater than 0.05
        if max_p_value > significance_level:
            # get variable with the max p-value
            excluded_var = p_values.idxmax()

            # drop variable with the max p-value
            X = X.drop(columns=[excluded_var])
        else:
            break

    return model

# apply backward elimination
reduced_model = backward_elimination(X, y)

# display reduced model summary
print(reduced_model.summary())
```

In [69]:
```python
# residual plot for the reduced model
plt.figure(figsize=(10, 6))
sns.residplot(x=reduced_model.fittedvalues, y=reduced_model.resid, lowess=True, line_kws={'color': 'red', 'lw': 1})
plt.xlabel('Fitted values')
plt.ylabel('Residuals')
plt.title('Residual Plot')
plt.show()
```




```
In [70]: # Residual Standard Error (RSE)
         rse = reduced_model.mse_resid ** 0.5
         print(f"Residual Standard Error (RSE): {rse}")

         Residual Standard Error (RSE): 40.71531103844988
```

See code attached in WGU_D208_Task_1.ipynb.

E3: Code

See code attached in WGU_D208_Task_1.ipynb.

F1: Results

- The regression equation for the reduced model is as follows:

  MonthlyCharge =

  $\beta 0+\beta 1 \cdot$ Bandwidth_GB_Year$+\beta 2 \cdot$ Outage_sec_perweek$+\beta 3 \cdot$ InternetService_Fibe

  r Optic$+\beta 4 \cdot$ InternetService_None$+\epsilon$

If we use the coefficient values, the equation would look like this:

$$\text{MonthlyCharge} = 159.1447 + 0.0014 \cdot \text{Bandwidth\_GB\_Year} + 0.2908 \cdot \text{Outage\_sec\_perweek} + 20.0389 \cdot \text{InternetService\_Fiber Optic} - 13.4241 \cdot \text{InternetService\_None} + \epsilon$$

- Each coefficient in the reduced model represents the expected change in the dependent variable ‘MonthlyCharge’ for a one-unit charge in the corresponding independent variable, holding all other variables constant.
    - Intercept ($\beta_0 = 159.1447$): this is the estimated monthly charge when all independent variables are zero.
    - Bandwidth_GB_Year ($\beta_1 = 0.0014$): for each additional gigabyte of bandwidth used per year, the monthly charge increases by $0.0014, holding all other variables constant.
    - Outage_sec_perweek ($\beta_2 = 0.2908$): for each additional second of outage per week, the monthly charge increases by $0.2908, holding all other variables constant.
    - InternetService_Fiber Optic ($\beta_3 = 20.0389$): using fiber optic internet service is associated with an increase of $20.0389 in monthly charge compared to the reference category, holding all other variables constant.
    - InternetService_None ($\beta_4 = 13.4241$): not having internet service is associated with a decrease of $13.4241 in monthly charge compared to the reference category, holding all other variables constant.
- The statistical significance of the reduced model is the p-values indicate whether the relationships between the independent variables and the dependent variable

‘MonthlyCharge’ are statistically significant, with a p-value < 0.05 suggesting the corresponding independent variables are significant predictors to the changes of ‘MonthlyCharge’. The practical significance would consider the real-world impacts the coefficients would have from a business perspective. (i.e. a large coefficient for Bandwidth_GB_Year suggests customers with high data usage also have higher monthly charges.)

- There are several limitations with this analysis. One would be omitted variable bias since other important variables that could potentially influence ‘MonthlyCharge’ may not be included in the data set or model, which leads to bias estimates. Another limitation is data quality, which assumes the data is accurate and reliable, which means any errors or biases in the data collection process can affect the results differently. Lastly, the model assumption that there is a linear relationship between the dependent and independent variables means non-linear relationships may not be captured by this model.

F2: Recommendations

Based on the results of the research, there are several recommendations for an organization to take. They could focus on retention towards their higher value customers based on significant predictors such as Bandwidth_GB_Year by offering additional incentives. Another recommendation would be to create customized plans based on variables such as age, income, data usage, etc. Having custom plans will help meet customers’ needs and improve retention. Lastly, the organization could review and adjust their pricing based on different payment methods if they discover certain methods like electronic checks could significantly affect monthly charges.

G: Panopto Video

The URL link for the Panopto video can be found here. It will also be submitted in the Performance Assessment task submission.

H: Sources of Third Party Code

1. “One Hot Encoding in Machine Learning.” Retrieved from https://www.geeksforgeeks.org/ml-one-hot-encoding/
2. “How to Create a Residual Plot in Python.” Retrieved from https://scales.arabpsychology.com/stats/how-to-create-a-residual-plot-in-python/
3. “Linear Regression. Residual Standard Error in Python (Jupyter).” Retrieved from https://www.youtube.com/watch?v=QxYmj-E3Ud4
4. “Multiple Linear Regression (Backward Elimination Technique).” Retrieved from https://www.geeksforgeeks.org/ml-multiple-linear-regression-backward-elimination-technique/

I: Web Sources

1. “Python: Intro to MLR/OLS in statsmodels.api.” Retrieved from https://www.youtube.com/watch?v=0-fkgpK2knA&list=PLe9UEU4oeAuV7RtCbL76hca5ELO_IELk4&index=9
2. “Building A Logistic Regression in Python, Step by Step.” Retrieved from https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8
3. “Feature Selection.” Retrieved from https://scikit-learn.org/stable/modules/feature_selection.html