D208: Predictive Modeling Task 2

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 whether customers will churn or not?”

A2. Goals

The goals of the data analysis are to identify the significant predictors that most contribute to customer churn, develop a logistic regression model to estimate the probability of a customer churning, and provide actionable insights or recommendations for how to improve customer churn by addressing those key predictors in this analysis.

B1. Summary of Assumptions

Here are 4 assumptions of a logistic regression model:

- Logistic regression requires the dependent variable to be dichotomous, meaning it should only have 2 possible outcomes (churn vs. no churn).
- Observations of the variables are independent of each other, which implies that any outcome for one observation should not influence the outcome of another.
- Logistic regression assumes a linear relationship between the logit (log-odds) of the dependent variable and the independent variables.
- The independent variables should not be highly correlated with each other as it would make identifying the individual effect of each variable more difficult.

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

Logistic regression is an appropriate technique because of the nature of the dependent variable ‘churn’ as it is binary and well-suited for this model. This technique also provides a way to not only identify significant predictors of churn but also estimate the probability of a customer churning. Lastly, this technique is capable of handling both continuous and categorical variables, which allows us to add in customer attributes (age, income, contract, etc.) to the model.

C1. Data Cleaning

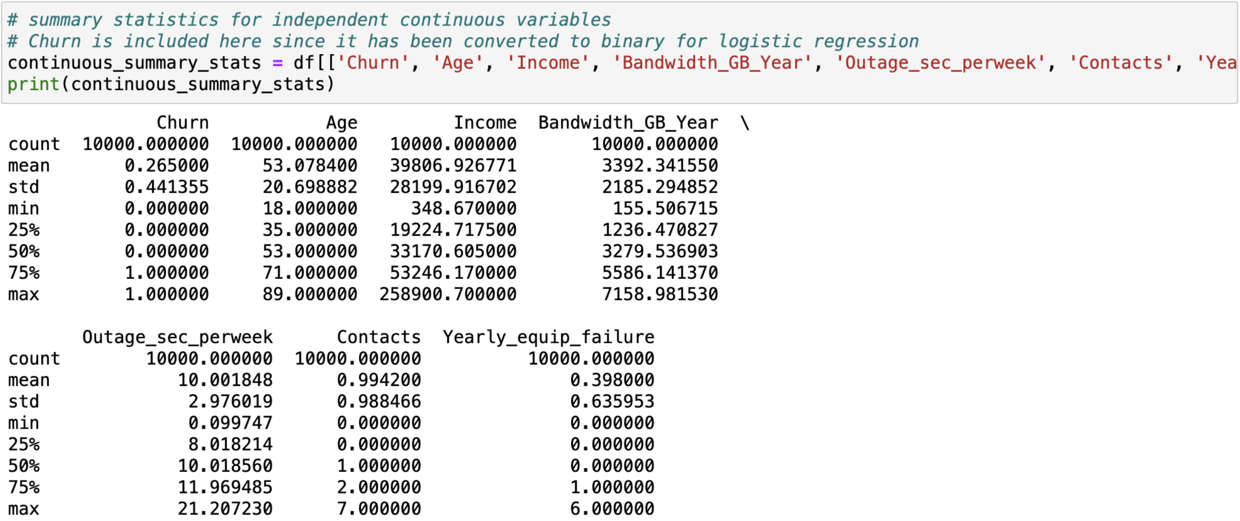
Here are all the data cleaning goals used in this analysis:

- 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_2.ipynb.

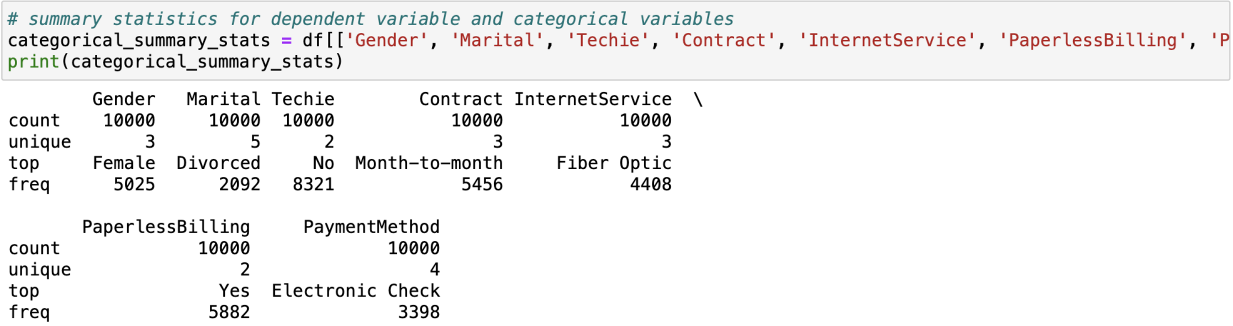
C2. Summary Statistics

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



```
# summary statistics for independent continuous variables
# Churn is included here since it has been converted to binary for logistic regression
continuous_summary_stats = df[['Churn', 'Age', 'Income', 'Bandwidth_GB_Year', 'Outage_sec_perweek', 'Contacts', 'Yea
print(continuous_summary_stats)
```

```
              Churn           Age          Income  Bandwidth_GB_Year  \
count  10000.000000  10000.000000    10000.000000       10000.000000   
mean       0.265000     53.078400    39806.926771        3392.341550   
std        0.441355     20.698882    28199.916702        2185.294852   
min        0.000000     18.000000      348.670000         155.506715   
25%        0.000000     35.000000    19224.717500        1236.470827   
50%        0.000000     53.000000    33170.605000        3279.536903   
75%        1.000000     71.000000    53246.170000        5586.141370   
max        1.000000     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  
```



```
# summary statistics for dependent variable and categorical variables
categorical_summary_stats = df[['Gender', 'Marital', 'Techie', 'Contract', 'InternetService', 'PaperlessBilling', 'P
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  
```

Continuous Variables:

- Churn:
  - Mean: 0.265 indicates that 26.5% of customers churned.
  - Standard Deviation: 0.441 shows the spread of churn occurrences.

    - Range: Min is 0 and Max is 1, reflecting binary classification.
- Age:
    - Mean: 53.08 years, indicating the average customer age.
    - Standard Deviation: 20.70 years, showing significant age variability.
    - Range: Min is 18, Max is 89, with ages distributed across adult age groups.
- Income:
    - Mean: $39,806.93, indicating the average income level of customers.
    - Standard Deviation: $28,199.92, suggesting substantial income diversity.
    - Range: Min is $348.67, Max is $258,900.70, with incomes varying widely.
- Bandwidth_GB_Year:
    - Mean: 3392.34 GB, indicating average yearly bandwidth usage.
    - Standard Deviation: 2185.29 GB, reflecting wide variability in data usage.
    - Range: Min is 155.51 GB, Max is 7158.98 GB, indicating a broad spectrum of data consumption.
- Outage_sec_perweek:
    - Mean: 10.00 seconds, indicating average weekly outage duration.
    - Standard Deviation: 2.98 seconds, showing variability in outages.
    - Range: Min is 0.10 seconds, Max is 21.21 seconds, with most customers experiencing minimal outages.
- Contacts:
    - Mean: 0.99, indicating the average number of customer service contacts.

    - Standard Deviation: 0.99, reflecting the spread in customer contact frequency.
    - Range: Min is 0, Max is 7, showing that most customers contacted support fewer times.
- Yearly_equip_failure:
    - Mean: 0.40 failures, indicating the average yearly equipment failures.
    - Standard Deviation: 0.64, suggesting some customers experience frequent failures.
    - Range: Min is 0, Max is 6, with most customers facing few failures.

Categorical Variables:

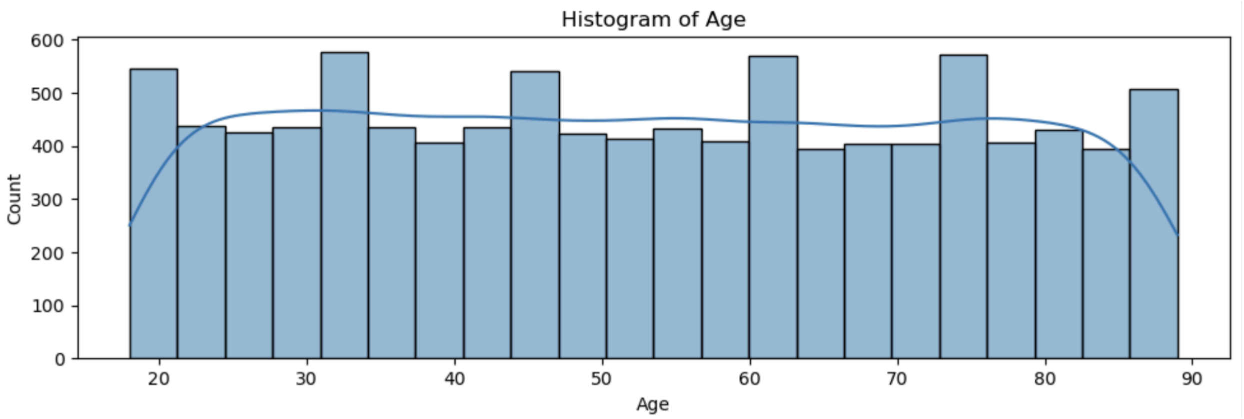
- Gender:
    - Categories: Female, Male, Nonbinary.
    - Most Common: Female, with 5025 occurrences (50.25%).
- Marital:
    - Categories: Divorced, Married, Separated, Single, Widowed.
    - Most Common: Divorced, with 2092 occurrences (20.92%).
- Techie:
    - Categories: Yes, No.
    - Most Common: No, with 8321 occurrences (83.21%).
- Contract:
    - Categories: Month-to-month, One year, Two year.
    - Most Common: Month-to-month, with 5456 occurrences (54.56%).
- InternetService:

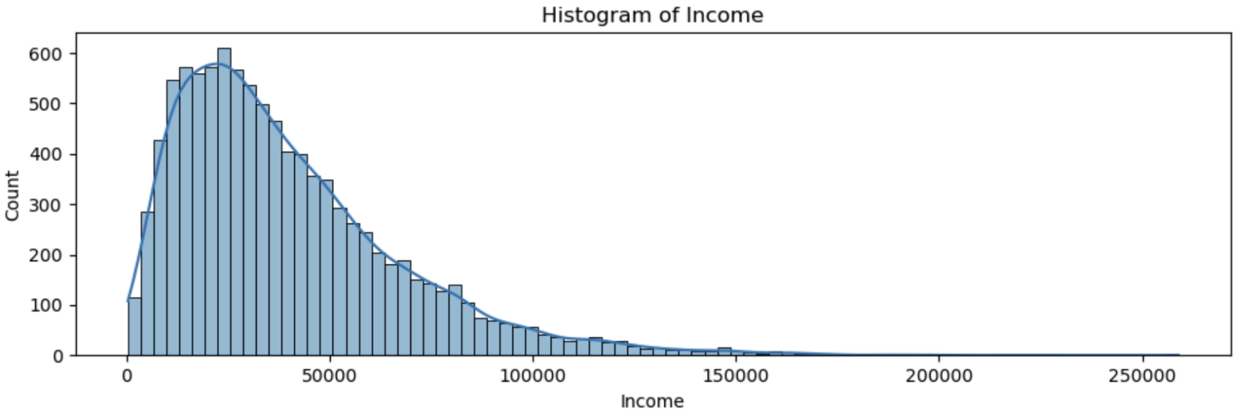
  - Categories: DSL, Fiber Optic, None.
  - Most Common: Fiber Optic, with 4408 occurrences (44.08%).
- PaperlessBilling:
  - Categories: Yes, No.
  - Most Common: Yes, with 5882 occurrences (58.82%).
- PaymentMethod:
  - Categories: Bank Transfer, Credit Card, Electronic Check, Mailed Check.
  - Most Common: Electronic Check, with 3398 occurrences (33.98%).

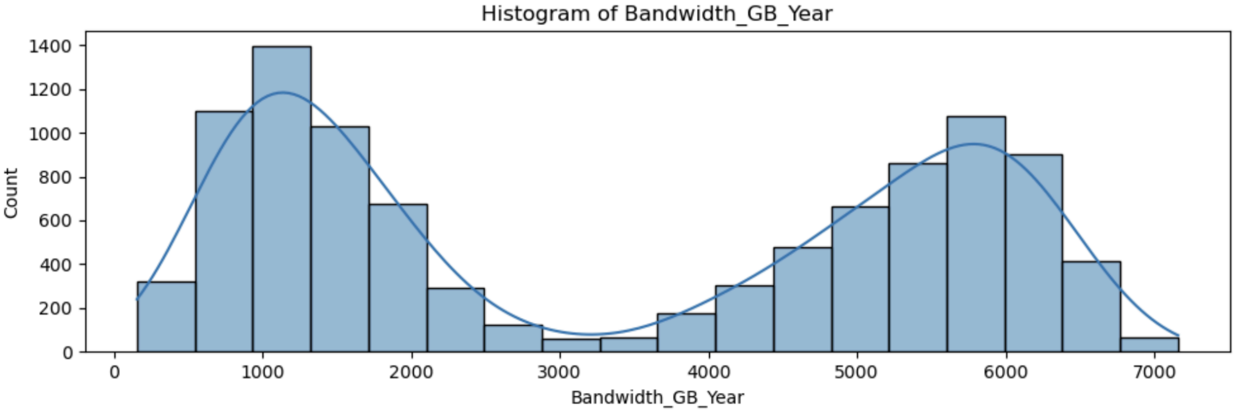
See code attached in WGU_D208_Task_2.ipynb.

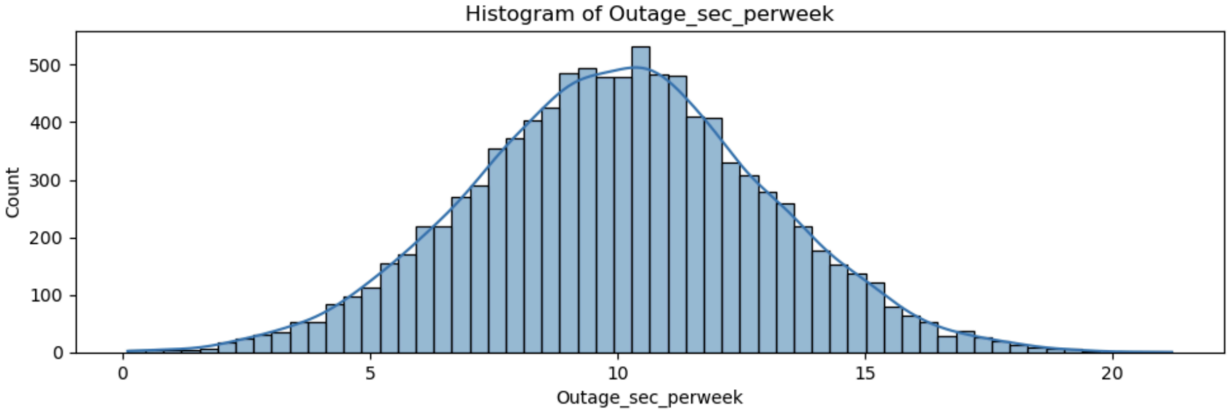
C3. Visualizations

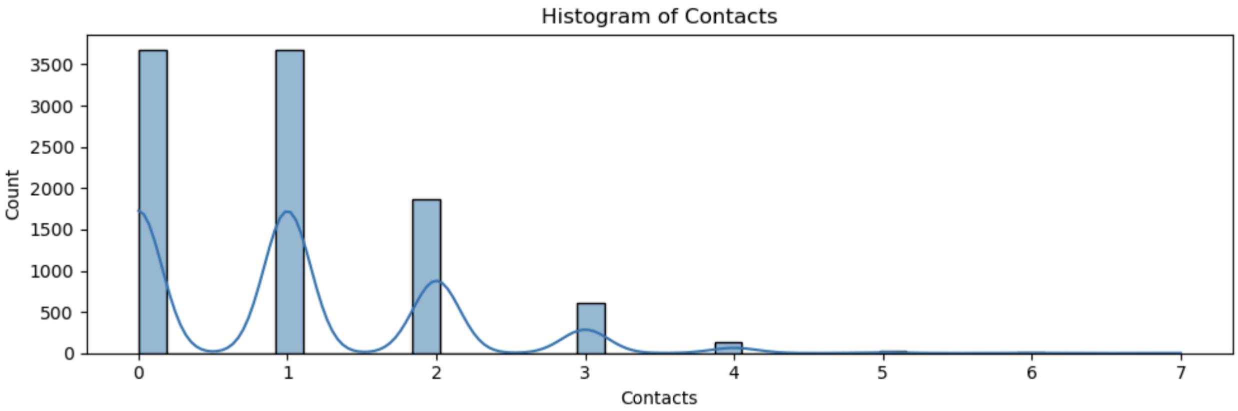
Univariate Visualization

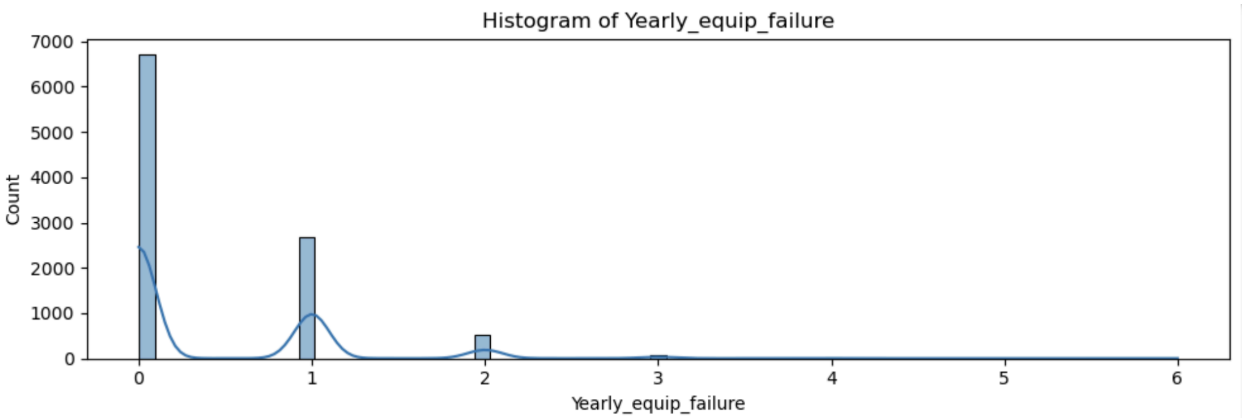


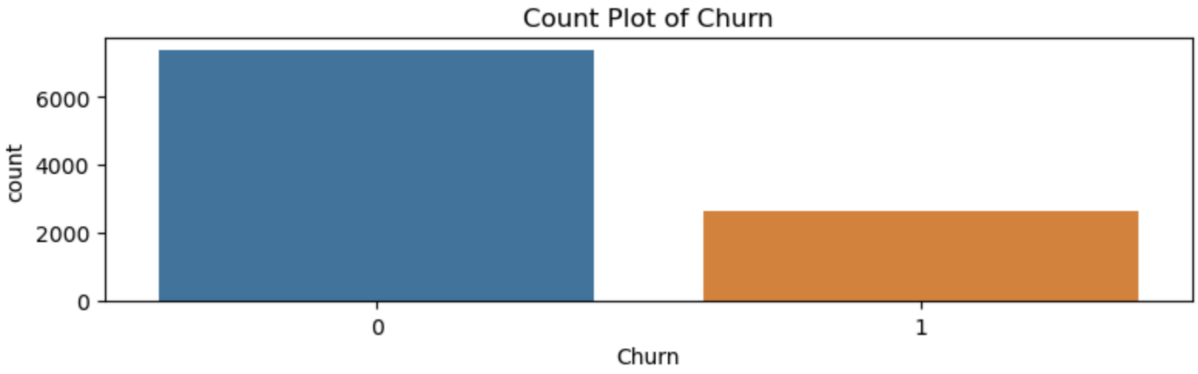

Histogram of Income
Count
Income
0
100
200
300
400
500
600
0
50000
100000
150000
200000
250000

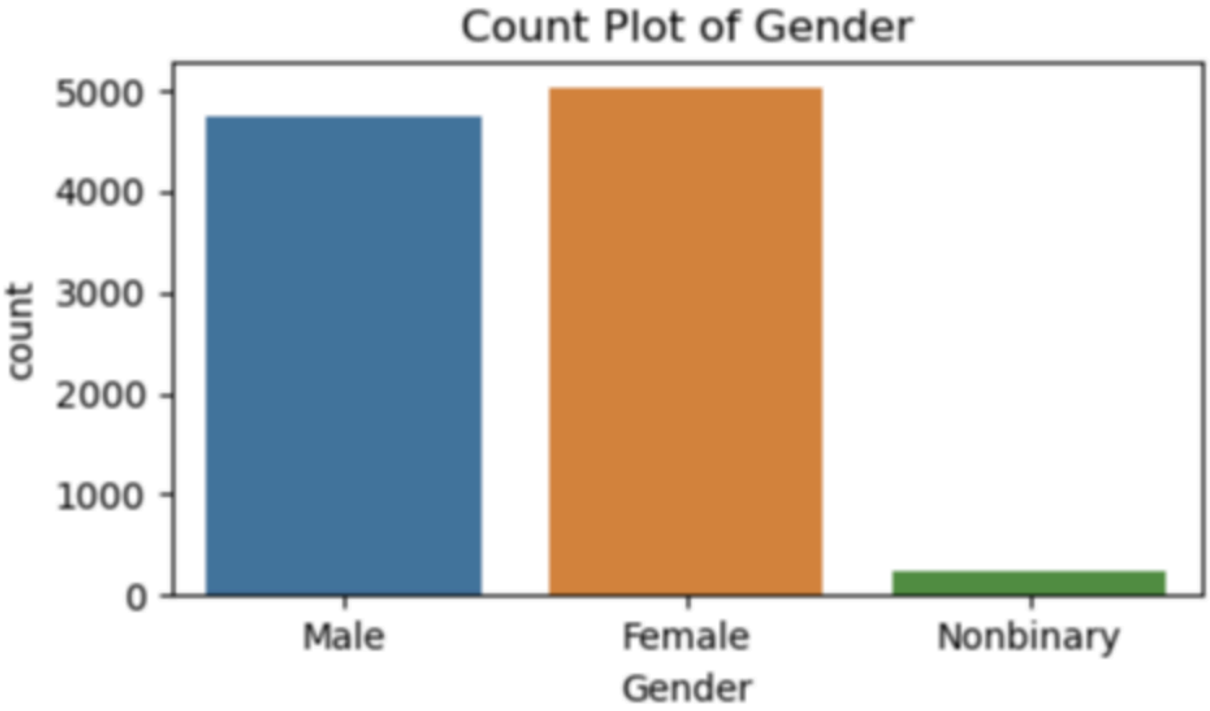

Histogram of Bandwidth_GB_Year
Count
Bandwidth_GB_Year
0
200
400
600
800
1000
1200
1400
0
1000
2000
3000
4000
5000
6000
7000

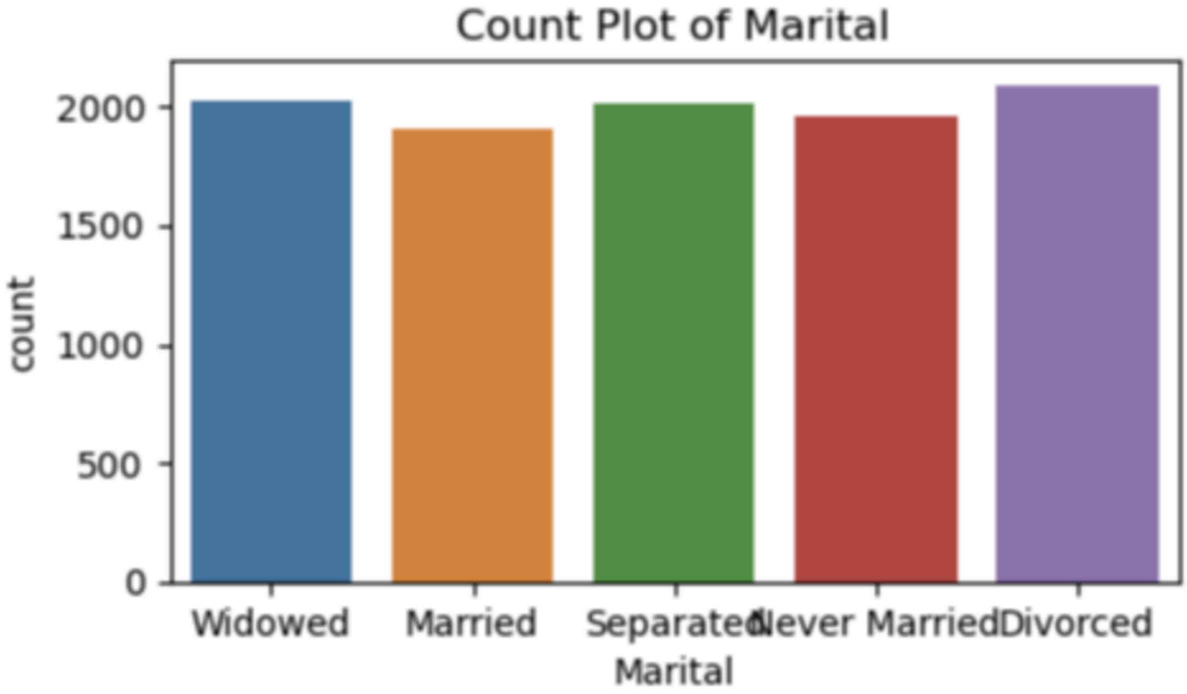

Histogram of Outage_sec_perweek
Count
Outage_sec_perweek
0
100
200
300
400
500
0
5
10
15
20

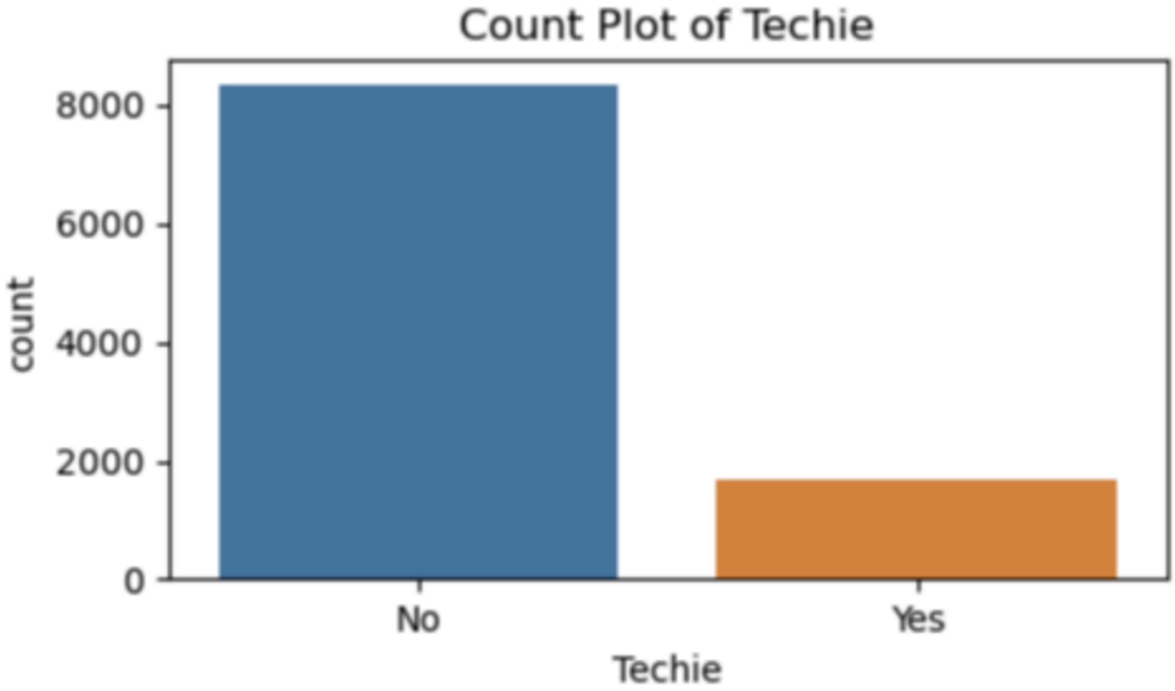

Histogram of Contacts
Count
3500
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
Contacts

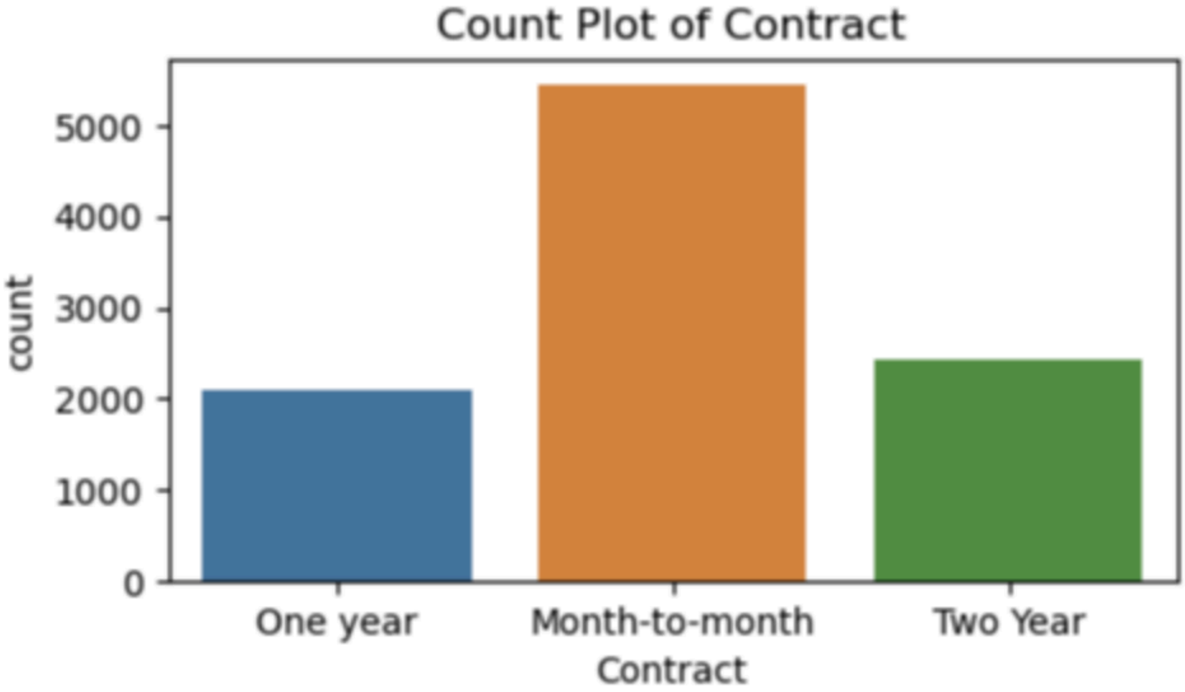

Histogram of Yearly_equip_failure
Count
7000
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
Yearly_equip_failure

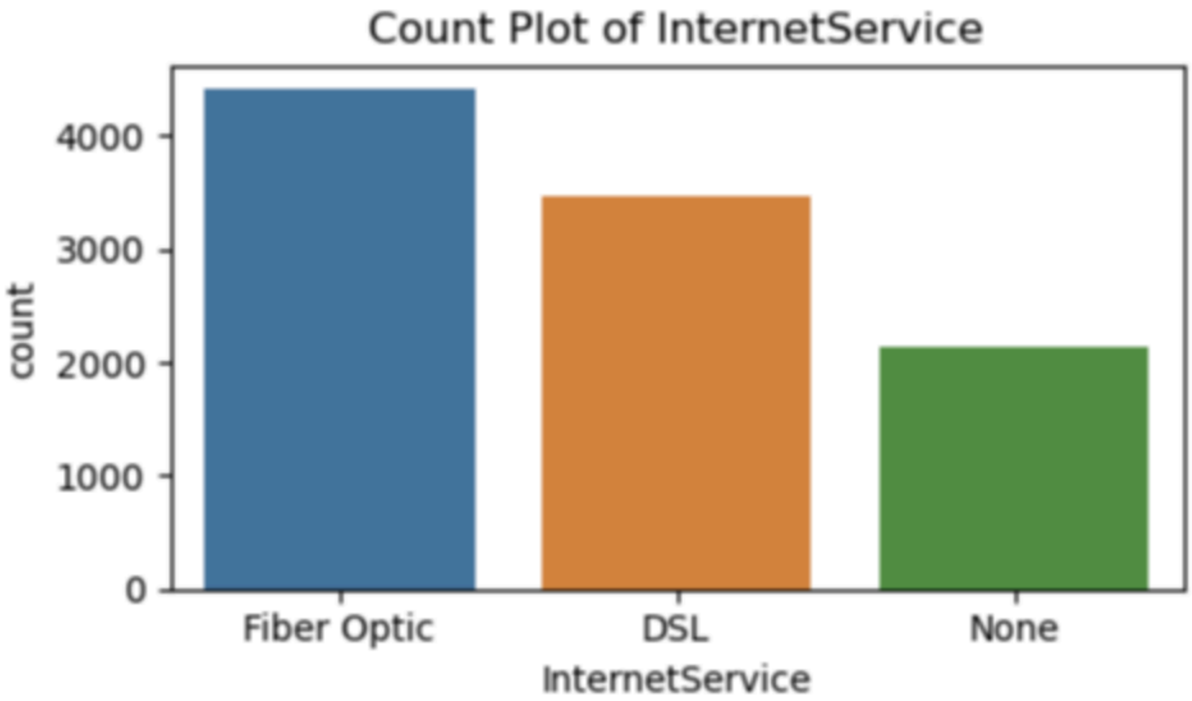

Count Plot of Churn
count
6000
4000
2000
0
0
1
Churn

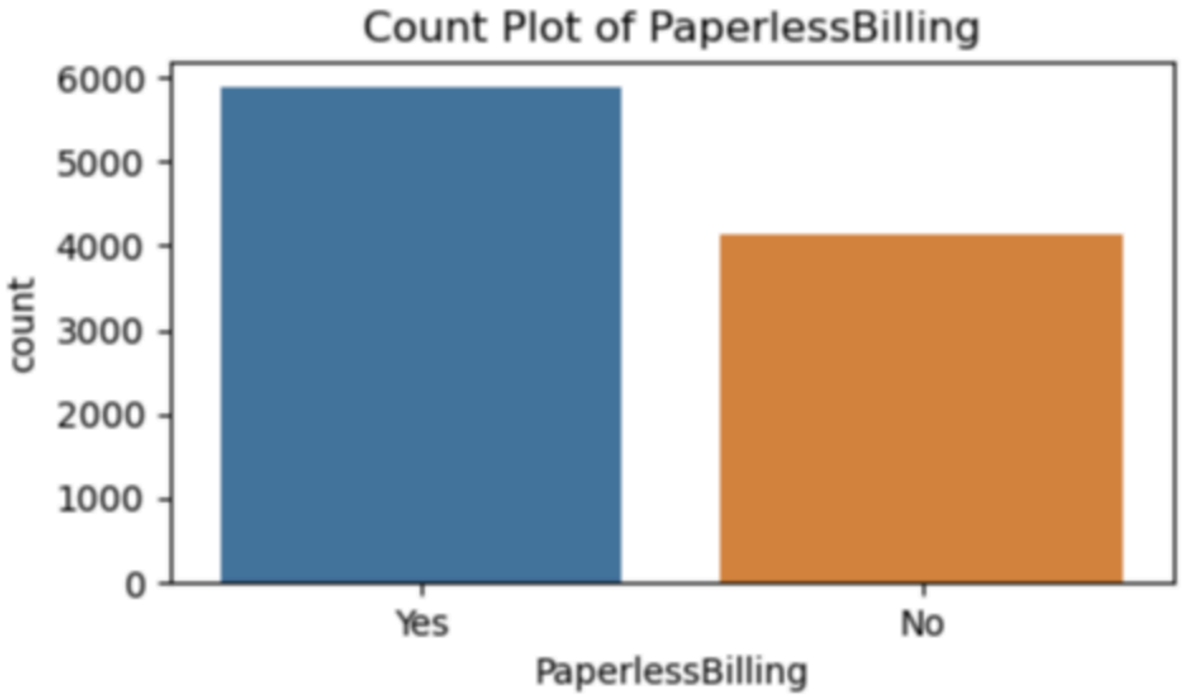

Count Plot of Gender
count
5000
4000
3000
2000
1000
0
Male
Female
Nonbinary
Gender

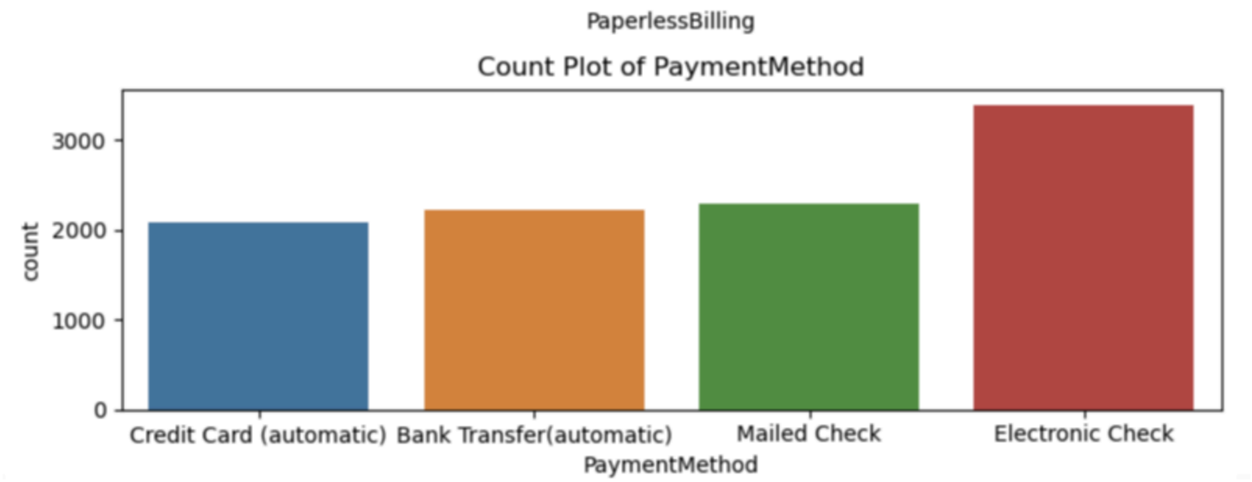

Count Plot of Marital
count
2000
1500
1000
500
0
Widowed
Married
Separated
Never Married
Divorced
Marital


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

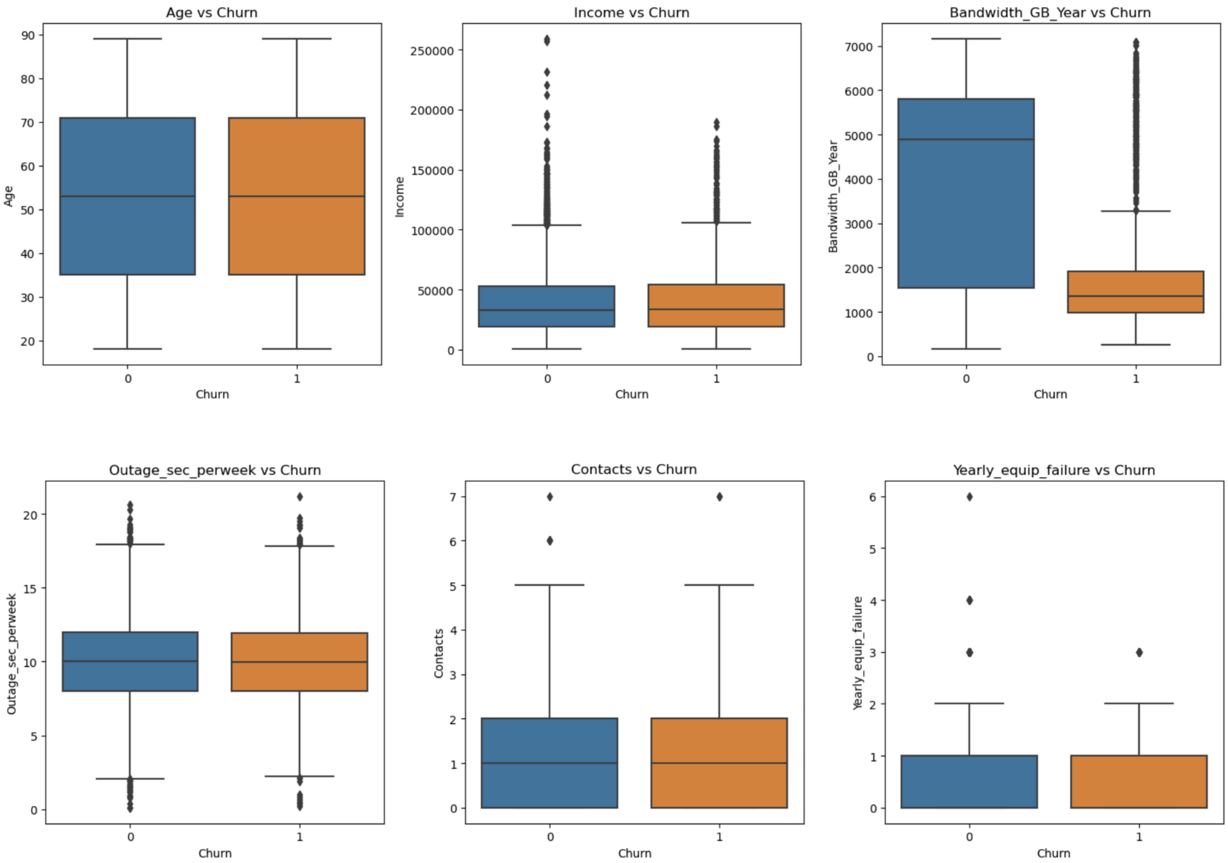

Count Plot of Contract
5000
4000
3000
2000
1000
0
count
One year
Month-to-month
Two Year
Contract

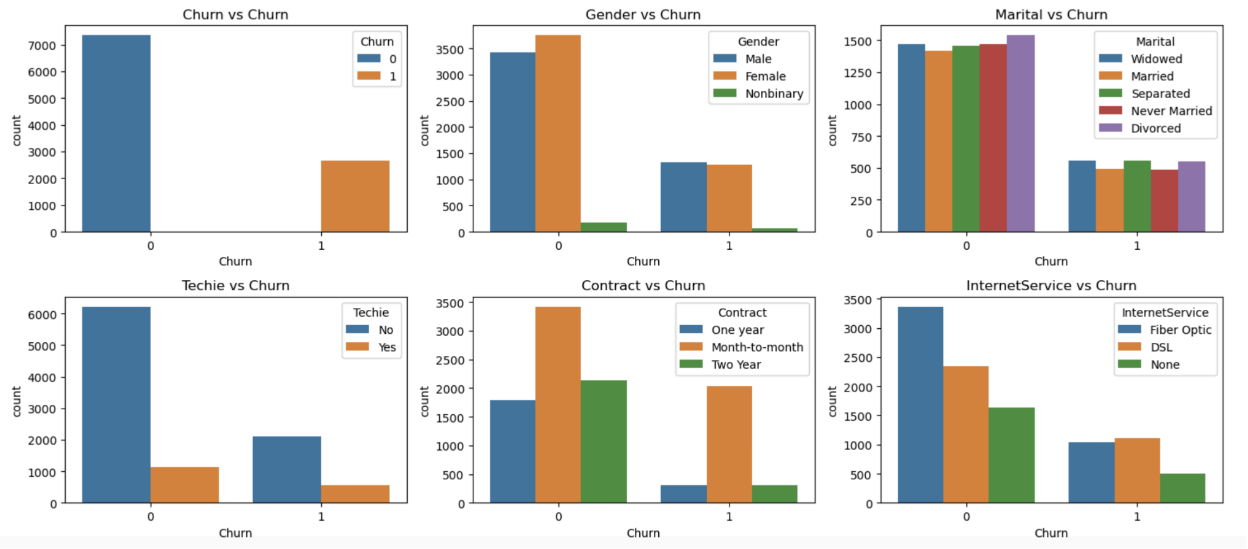

Count Plot of InternetService
count
4000
3000
2000
1000
0
Fiber Optic
DSL
None
InternetService

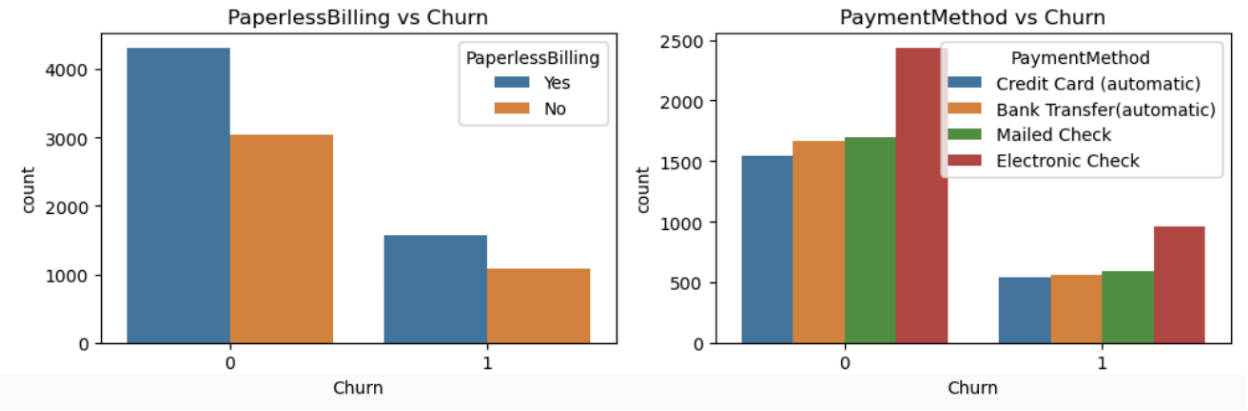

Count Plot of PaperlessBilling
count
6000
5000
4000
3000
2000
1000
0
Yes
No
PaperlessBilling



Bivariate Visualization







See code attached in WGU_D208_Task_2.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 ->

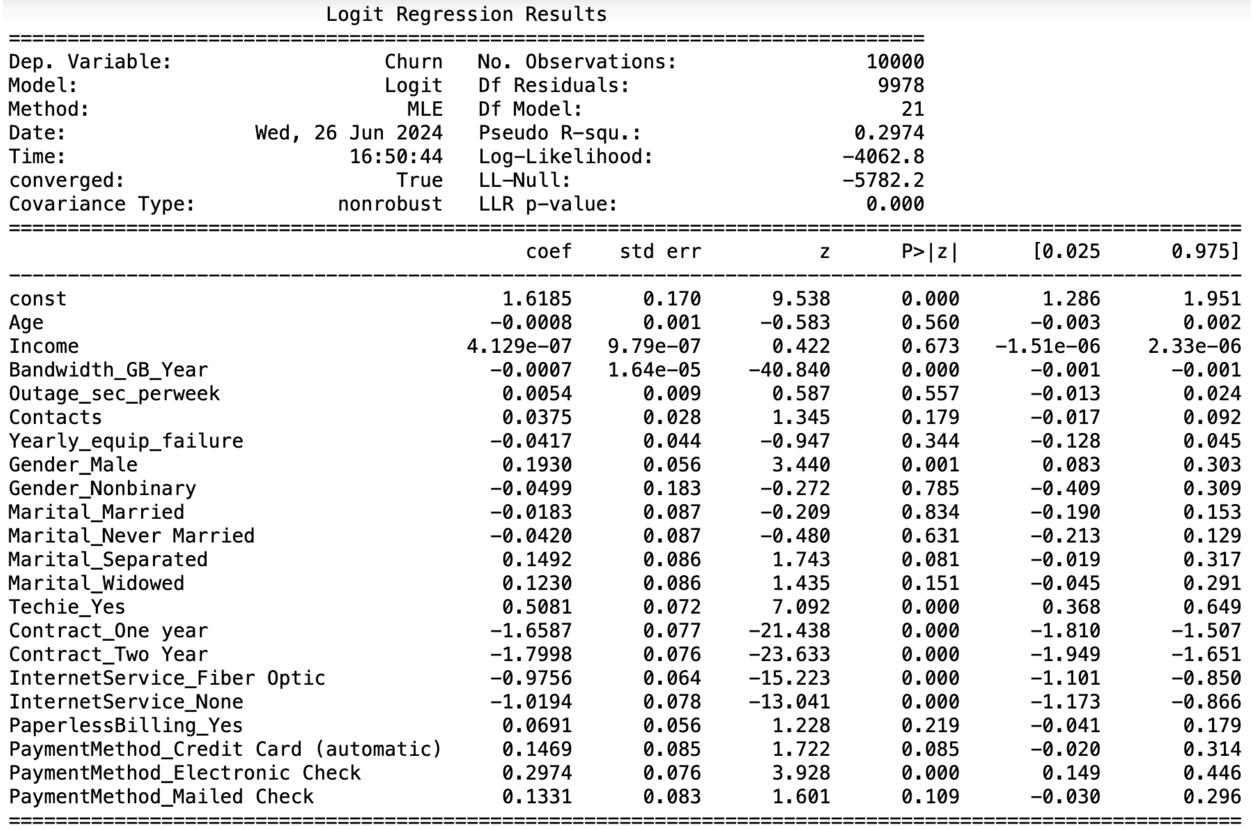
handle any missing values -> encode the categorical variables -> generate summary statistics for encoded variables.

See code attached in WGU_D208_Task_2.ipynb.

C5: Prepared Data Set

A copy of the fully prepared data set will be submitted as ‘prepared_data_d208_task2.csv’.

D1:Initial Model



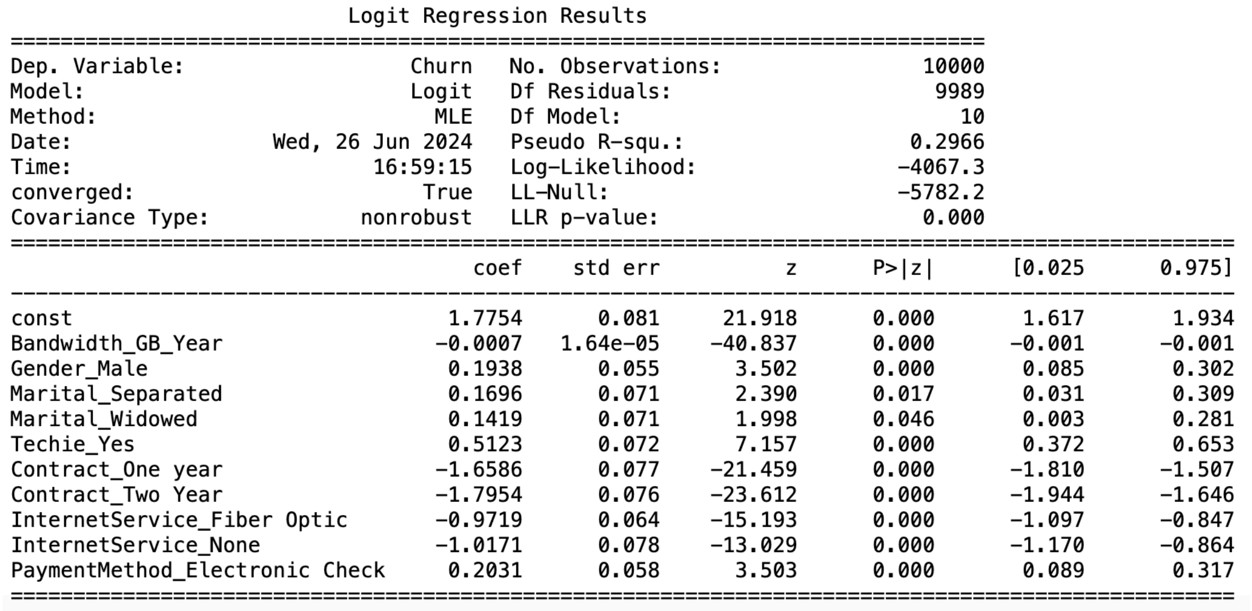
```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  Churn   No. Observations:                10000
Model:                          Logit   Df Residuals:                     9978
Method:                           MLE   Df Model:                           21
Date:                Wed, 26 Jun 2024   Pseudo R-squ.:                  0.2974
Time:                        16:50:44   Log-Likelihood:                -4062.8
converged:                       True   LL-Null:                       -5782.2
Covariance Type:            nonrobust   LLR p-value:                     0.000
=========================================================================================================
                                            coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                     1.6185      0.170      9.538      0.000       1.286       1.951
Age                                      -0.0008      0.001     -0.583      0.560      -0.003       0.002
Income                                 4.129e-07   9.79e-07      0.422      0.673   -1.51e-06    2.33e-06
Bandwidth_GB_Year                        -0.0007   1.64e-05    -40.840      0.000      -0.001      -0.001
Outage_sec_perweek                        0.0054      0.009      0.587      0.557      -0.013       0.024
Contacts                                  0.0375      0.028      1.345      0.179      -0.017       0.092
Yearly_equip_failure                     -0.0417      0.044     -0.947      0.344      -0.128       0.045
Gender_Male                               0.1930      0.056      3.440      0.001       0.083       0.303
Gender_Nonbinary                         -0.0499      0.183     -0.272      0.785      -0.409       0.309
Marital_Married                          -0.0183      0.087     -0.209      0.834      -0.190       0.153
Marital_Never Married                    -0.0420      0.087     -0.480      0.631      -0.213       0.129
Marital_Separated                         0.1492      0.086      1.743      0.081      -0.019       0.317
Marital_Widowed                           0.1230      0.086      1.435      0.151      -0.045       0.291
Techie_Yes                                0.5081      0.072      7.092      0.000       0.368       0.649
Contract_One year                        -1.6587      0.077    -21.438      0.000      -1.810      -1.507
Contract_Two Year                        -1.7998      0.076    -23.633      0.000      -1.949      -1.651
InternetService_Fiber Optic              -0.9756      0.064    -15.223      0.000      -1.101      -0.850
InternetService_None                     -1.0194      0.078    -13.041      0.000      -1.173      -0.866
PaperlessBilling_Yes                      0.0691      0.056      1.228      0.219      -0.041       0.179
PaymentMethod_Credit Card (automatic)     0.1469      0.085      1.722      0.085      -0.020       0.314
PaymentMethod_Electronic Check            0.2974      0.076      3.928      0.000       0.149       0.446
PaymentMethod_Mailed Check                0.1331      0.083      1.601      0.109      -0.030       0.296
=========================================================================================================
```

See code attached in WGU_D208_Task_2.ipynb.

D2: Justification of Model Reduction

To align with our research question, we’re going to use backward elimination based on the p-values of the predictors. 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. This feature selection will help us identify significant predictors of customer churn.

D3: Reduced Logistic Regression Model



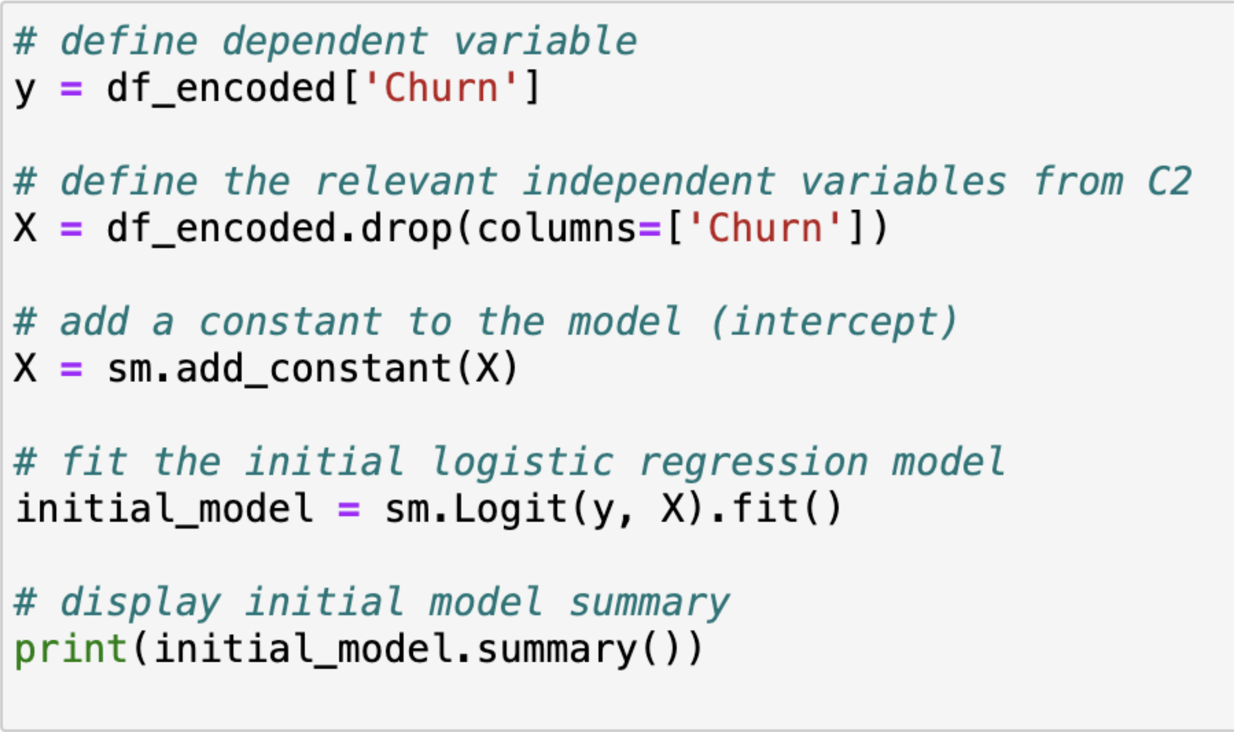
```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  Churn   No. Observations:                10000
Model:                          Logit   Df Residuals:                     9989
Method:                           MLE   Df Model:                           10
Date:                Wed, 26 Jun 2024   Pseudo R-squ.:                  0.2966
Time:                        16:59:15   Log-Likelihood:                -4067.3
converged:                       True   LL-Null:                       -5782.2
Covariance Type:            nonrobust   LLR p-value:                     0.000
==================================================================================================
                                     coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                              1.7754      0.081     21.918      0.000       1.617       1.934
Bandwidth_GB_Year                 -0.0007   1.64e-05    -40.837      0.000      -0.001      -0.001
Gender_Male                        0.1938      0.055      3.502      0.000       0.085       0.302
Marital_Separated                  0.1696      0.071      2.390      0.017       0.031       0.309
Marital_Widowed                    0.1419      0.071      1.998      0.046       0.003       0.281
Techie_Yes                         0.5123      0.072      7.157      0.000       0.372       0.653
Contract_One year                 -1.6586      0.077    -21.459      0.000      -1.810      -1.507
Contract_Two Year                 -1.7954      0.076    -23.612      0.000      -1.944      -1.646
InternetService_Fiber Optic       -0.9719      0.064    -15.193      0.000      -1.097      -0.847
InternetService_None              -1.0171      0.078    -13.029      0.000      -1.170      -0.864
PaymentMethod_Electronic Check     0.2031      0.058      3.503      0.000       0.089       0.317
==================================================================================================
```

See code attached in WGU_D208_Task_2.ipynb.

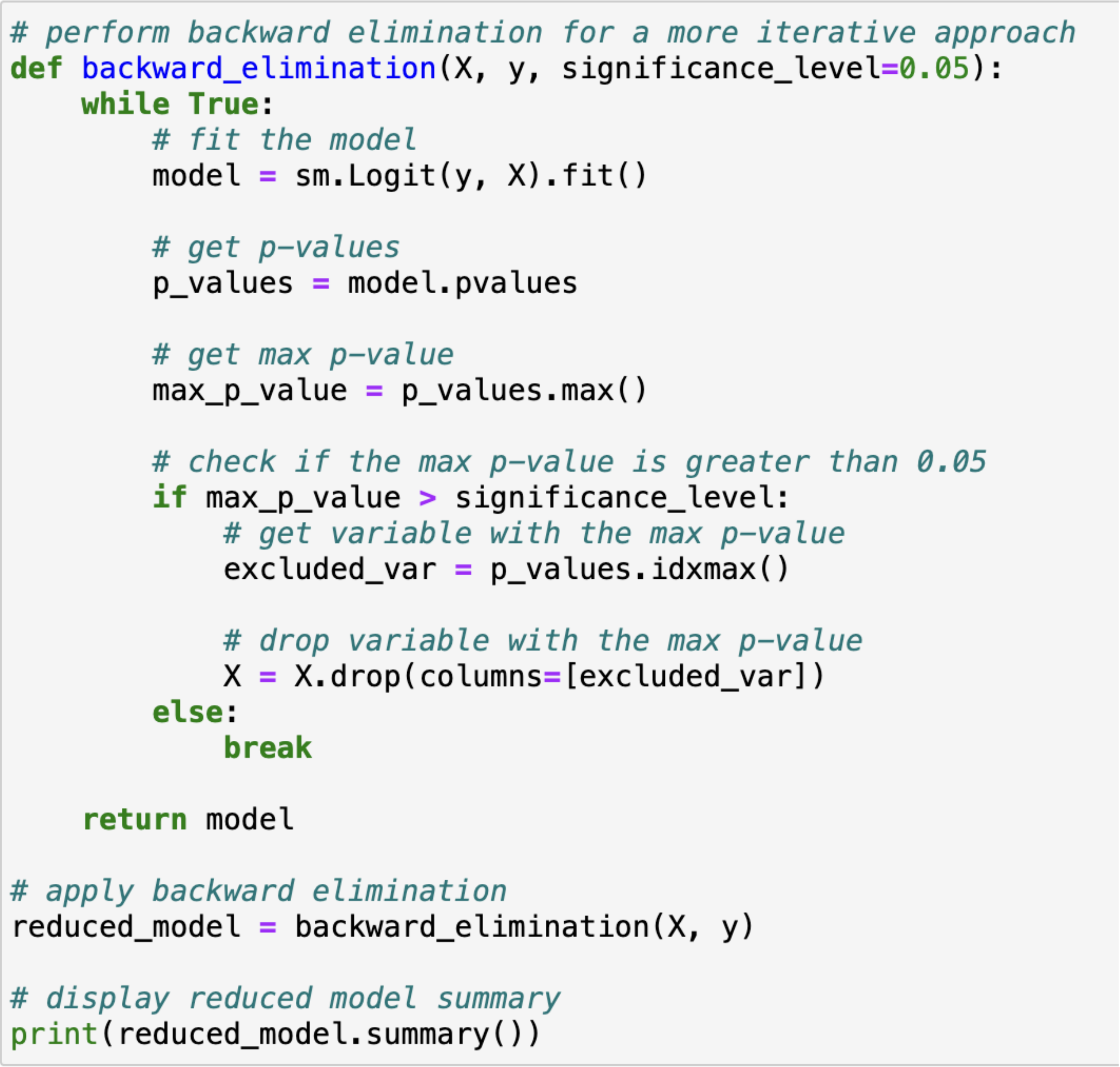
E1: Model Comparison

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 Pseudo R-squared. This metric indicates how well the model explains the variability of the data, with higher values meaning better explanatory power. We see that the initial Pseudo R-squared value is 0.2974 but once we apply backward elimination to retain only the significant variables, we see the value decreased to 0.2966 in the reduced model. The difference between the Pseudo R-squared values is minimal, which means the reduced model still captures most of the variability of the data. Although the initial has a slightly higher Pseudo R-squared value, it is still preferred to use the reduced model as it simplifies the analysis by having fewer predictors, reducing the chance of overfitting, and making the model more interpretable. The trade-off is still advantageous towards the reduced model despite having a marginally lower Pseudo R-squared value.

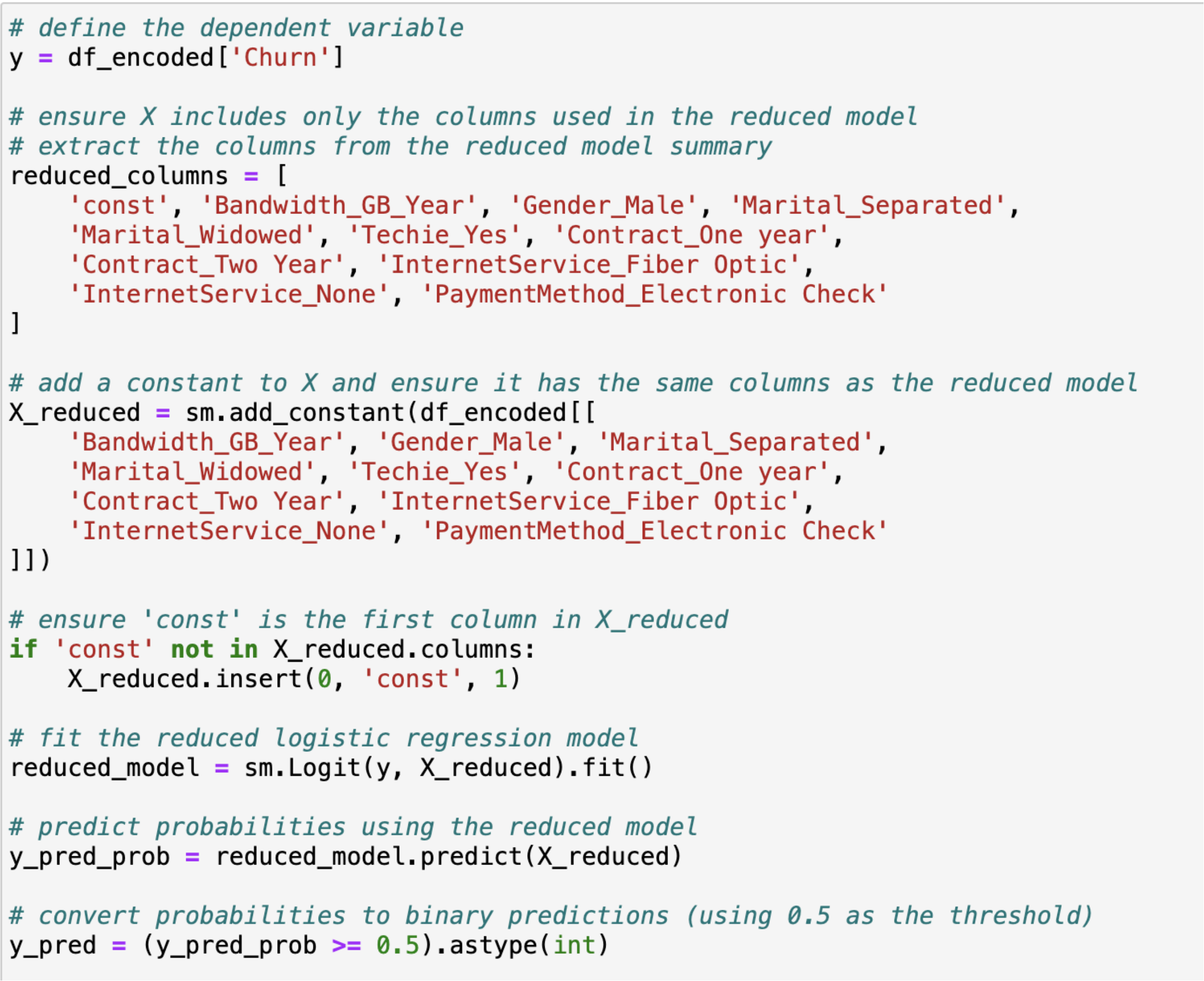
E2: Output and Calculations



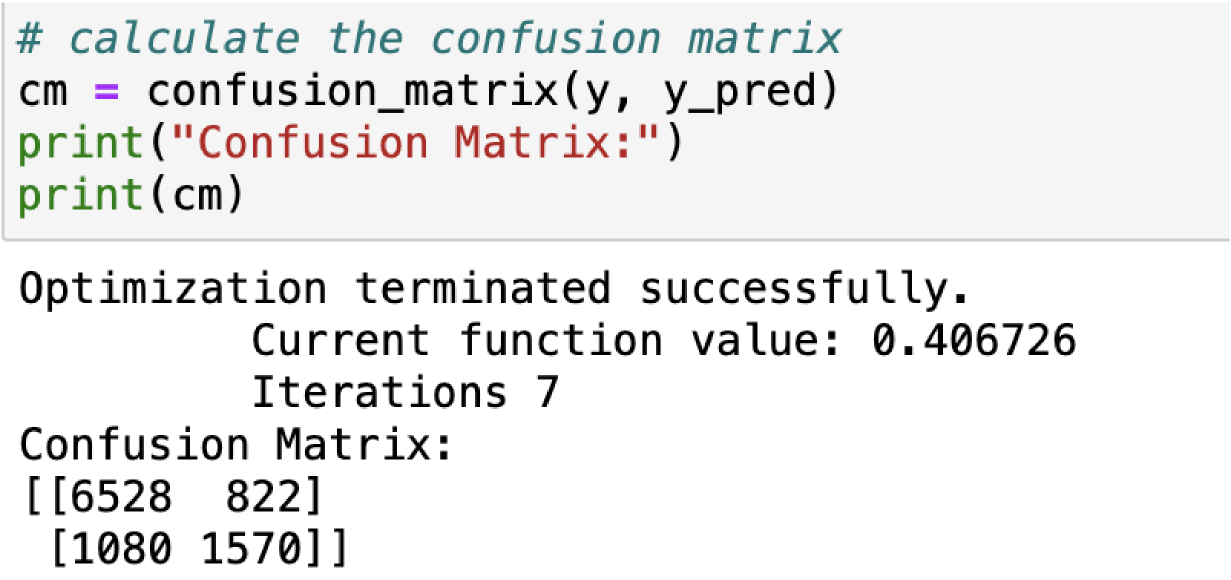
```
# define dependent variable
y = df_encoded['Churn']

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

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

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

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



```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.Logit(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())
```

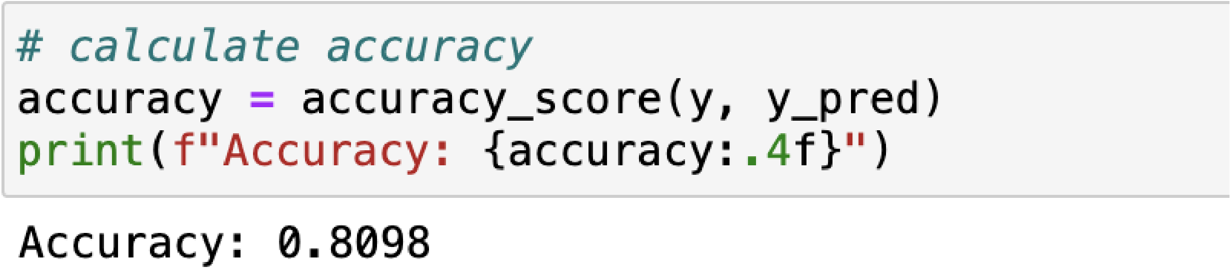


```python
# define the dependent variable
y = df_encoded['Churn']

# ensure X includes only the columns used in the reduced model
# extract the columns from the reduced model summary
reduced_columns = [
    'const', 'Bandwidth_GB_Year', 'Gender_Male', 'Marital_Separated',
    'Marital_Widowed', 'Techie_Yes', 'Contract_One year',
    'Contract_Two Year', 'InternetService_Fiber Optic',
    'InternetService_None', 'PaymentMethod_Electronic Check'
]

# add a constant to X and ensure it has the same columns as the reduced model
X_reduced = sm.add_constant(df_encoded[[
    'Bandwidth_GB_Year', 'Gender_Male', 'Marital_Separated',
    'Marital_Widowed', 'Techie_Yes', 'Contract_One year',
    'Contract_Two Year', 'InternetService_Fiber Optic',
    'InternetService_None', 'PaymentMethod_Electronic Check'
]])

# ensure 'const' is the first column in X_reduced
if 'const' not in X_reduced.columns:
    X_reduced.insert(0, 'const', 1)

# fit the reduced logistic regression model
reduced_model = sm.Logit(y, X_reduced).fit()

# predict probabilities using the reduced model
y_pred_prob = reduced_model.predict(X_reduced)

# convert probabilities to binary predictions (using 0.5 as the threshold)
y_pred = (y_pred_prob >= 0.5).astype(int)
```



```
# calculate the confusion matrix
cm = confusion_matrix(y, y_pred)
print("Confusion Matrix:")
print(cm)
```

```
Optimization terminated successfully.
         Current function value: 0.406726
         Iterations 7
Confusion Matrix:
[[6528  822]
 [1080 1570]]
```



```
# calculate accuracy
accuracy = accuracy_score(y, y_pred)
print(f"Accuracy: {accuracy:.4f}")
```

```
Accuracy: 0.8098
```

See code attached in WGU_D208_Task_2.ipynb.

E3: Code

See code attached in WGU_D208_Task_2.ipynb.

F1: Results

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

  logit(P(Churn)) =

  β0+β1 · Bandwidth_GB_Year+β2 · Gender_Male+β3 · Marital_Separated+β4 · Marital_Widowed+β5 · Techie_Yes+β6 · Contract_One_year+β7 · Contract_Two_Year+β8 · InternetService_Fiber_Optic+β9 · InternetService_None+β10 · PaymentMethod_Electronic_Check

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

$$\text{logit}(P(\text{Churn})) = 1.7754+(-0.0007)\cdot\text{Bandwidth\_GB\_Year}+0.1938\cdot\text{Gender\_Male}+0.1696\cdot\text{Marital\_Separated}+0.1419\cdot\text{Marital\_Widowed}+0.5123\cdot\text{Techie\_Yes}+(-1.6586)\cdot\text{Contract\_One\_year}+(-1.7954)\cdot\text{Contract\_Two\_Year}+(-0.9719)\cdot\text{InternetService\_Fiber\_Optic}+(-1.0171)\cdot\text{InternetService\_None}+0.2031\cdot\text{PaymentMethod\_Electronic\_Check}$$

- Interpretation of the coefficients of the reduced model:
    - Intercept (β0=1.7754): the baseline log-odds of churn when all predictors are zero.
    - Bandwidth_GB_Year (β1=-0.0007): for each additional GB of bandwidth used per year, the log-odds of churn decrease by 0.0007, holding all other variables constant.
    - Gender_Male (β2=0.1938): being male increases the log-odds of churn by 0.1938 compared to being female, holding all other variables constant.
    - Marital_Separated (β3=0.1696): being separated increases the log-odds of churn by 0.1696 compared to being married, holding all other variables constant.
    - Marital_Widowed (β4=0.1419): being widowed increases the log-odds of churn by 0.1419 compared to being married, holding all other variables constant.
    - Techie_Yes (β5=0.5123): being a techie increases the log-odds of churn by 0.5123 compared to not being a techie, holding all other variables constant.

  - Contract_One_year (β6=-1.6586): having a one-year contract decreases the log-odds of churn by 1.6586 compared to month-to-month contract, holding all other variables constant.
  - Contract_Two_year (β7=-1.7954): having a two-year contract decreases the log-odds of churn by 1.7954 compared to month-to-month contract, holding all other variables constant.
  - InternetService_Fiber_Optic (β8=-0.9719): having fiber optic internet decreases the log-odds of churn by 0.9719 compared to DSL, holding all other variables constant.
  - InternetService_None (β9=-1.0171): not having internet service decreases the log-odds of churn by 1.0171 compared to DSL, holding all other variables constant.
  - PaymentMethod_Electronic_Check (β10=0.2031): using electronic check as a payment method increases the log-odds of churn by 0.2031 compared to automatic bank transfer, holding all other variables constant.
- The statistical significance of the reduced model is the p-values for most of the coefficients have a p-value < 0.05 suggesting the corresponding independent variables are significant predictors of ‘Churn’. The practical significance would be some of the coefficients have a larger practical impact on the likelihood of churn (i.e. one or two year contracts significantly reduces churn but being a techie increases likelihood churn, which could be an area that requires more attention).
- There are several limitations with this analysis. One would be the model assumption that a logistic regression assumes a linear relationship between the

log-odds of the dependent variable and the independent variables, which may not always be true. Another limitation is assuming the data is clean and accurately represents the underlying population but any data quality issue could affect the end results. Lastly, another limitation is feature selection since the backward elimination method may not identify the most optimal predictors as it depends on the initial model and the significance level.

F2: Recommendations

Based on the results of the research, there are several recommendations for an organization to take. They could potentially offer longer term contracts since one and two year contracts have shown to reduce churn significantly. Another action they could take is to focus on higher quality internet service since customers with fiber optic or no internet service have lower churn rates compared to those with DSL. Lastly, they could encourage customers to use more stable payment methods since those who use electronic checks have higher churn rates, so an organization could further investigate why that is while focusing on getting customers to use other forms of payment.

G: Panopto Video

The URL link for the Panopto video 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 Fix: pandas data cast to numpy dtype of object. Check input data with np.asarray(data).” Retrieved from

https://www.statology.org/pandas-data-cast-to-numpy-dtype-of-object-check-input-data-with-np-asarraydata/

3. “D208: Predictive Modeling Task 1.” Retrieved from WGU_D208_Task1_Justin_Huynh.pdf

I: Web Sources

1. “Confusion Matrix in Machine Learning.” Retrieved from https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
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
4. “Machine Learning - Confusion Matrix.” Retrieved from https://www.w3schools.com/python/python_ml_confusion_matrix.asp