Factors Impacting Bandwidth Usage

Gerald Burke

Western Governors University

D208: Predictive Modeling

Dr. Taylor Jensen

October 23, 2024

[A - Purpose 4](#_Toc180910451)

[A1 – Research Question 4](#_Toc180910452)

[A2 – Goal 4](#_Toc180910453)

[B – Multiple Linear Regression 6](#_Toc180910454)

[B1 - Assumptions 6](#_Toc180910455)

[B2 – Benefits of Python in Linear Regression 6](#_Toc180910456)

[B3 – Multiple Linear Regression in Hypothesis Testing 7](#_Toc180910457)

[C – Data Preparation 7](#_Toc180910458)

[C1 – Data Cleaning 7](#_Toc180910459)

[Duplicate Values 7](#_Toc180910460)

[Missing Values 8](#_Toc180910461)

[Outliers 10](#_Toc180910462)

[Evaluation 11](#_Toc180910463)

[C2 – Variables 17](#_Toc180910464)

[Dependent Variable 17](#_Toc180910465)

[Independent Variables 18](#_Toc180910466)

[C3 – Univariate and Bivariate Analyses 25](#_Toc180910467)

[Univariate Analysis 25](#_Toc180910468)

[Bivariate Analysis 37](#_Toc180910469)

[C4 – Data Transformation 47](#_Toc180910470)

[C5 – Prepared Data 53](#_Toc180910471)

[D – Initial vs. Reduced Models 53](#_Toc180910472)

[D1 – Initial Multiple Linear Regression 53](#_Toc180910473)

[D2 – Feature Selection 56](#_Toc180910474)

[D3 – Reduced Linear Regression 80](#_Toc180910475)

[E – Linear Regression Analysis 82](#_Toc180910476)

[E1 - Process 82](#_Toc180910477)

[E2 - Output 82](#_Toc180910478)

[E3 - Code 108](#_Toc180910479)

[F – Data Summary and Implications 108](#_Toc180910480)

[F1 - Results 108](#_Toc180910481)

[F2 – Course of Action 110](#_Toc180910482)

[G – Panopto Video 110](#_Toc180910483)

[H – Code Sources 110](#_Toc180910484)

[I - Sources 110](#_Toc180910485)

# A - Purpose

## A1 – Research Question

What factors, if any, provided in the data set impact the rates of bandwidth usage?

## A2 – Goal

Bandwidth is a finite resource for a telecommunications company. High usage rates can impact network performance, cause service disruptions, and burden field technicians. My goal is to probe the dataset to look for independent variables that have an impact on the dependent variable(Bandwidth\_GB\_Year). Bandwidth\_GB\_Year is defined as “The average amount of data used, in GB, in a year by the customer” (Churn Data Consideration and Dictionary). The variable is continuous, represented by a floating-point value.

For the analysis, I will be examining the following explanatory variables:

State

* Categorical: Customer state of residence
* Do customers in certain states use more or less data?

Population

* Discrete: Population within a mile radius of the customer
* Do customers in higher or lower population areas use more or less data?

Area

* Categorical: Area type based on census data
* Do customers in different areas use more or less data?

Children

* Discrete: Number of children in customer’s household
* Do customers with more or fewer children use more or less data?

Age

* Continuous/Discrete(based on treatment): Age of the customer
* Do customers of certain ages use more or less data?

Income

* Continuous: Annual income of the customer
* Does the annual income of the customer impact their data use?

Marital

* Categorical: Marital status of the customer
* Does the marital status of the customer impact their data use?

Gender

* Categorical: Gender identification of the customer
* Does the customer’s gender identification impact their data use

Email

* Discrete: The number of emails sent to the customer in the last year
* Does the number of emails sent to the customer impact their data use?

Contacts

* Discrete: The number of times the customer contacted technical support
* Does the number of times the customer contacted technical support impact their data use?

Yearly\_equip\_failure

* Discrete: The number of times the customer’s equipment failed
* Does the number of time the customer’s equipment failed impact their data use?

Techie

* Categorical: Whether the customer considers themselves technically inclined
* Does the customer’s estimation of their own technical skills impact their data use?

Contract

* Categorical: The contract term of the customer
* Does the customer’s contract term impact their data use?

Port\_modem

* Categorical: Whether the customer has a portable modem
* Does having a portable modem impact customer data use?

Tablet

* Categorical: Whether the customer owns a tablet
* Does owning a tablet impact customer data use?

InternetService

* Categorical: Customer’s internet service provider
* Does internet service provider impact customer data use?

Phone

* Categorical: Whether the customer has phone service
* Does having phone service impact customer data use?

Multiple

* Categorical: Whether the customer has multiple lines
* Does having multiple lines impact customer data use?

OnlineSecurity

* Categorical: Whether the customer has online security add-on
* Does having an online security add-on impact customer data use?

OnlineBackup

* Categorical: Whether the customer has the online backup add-on
* Does having an online backup add-on impact customer data use?

StreamingTV

* Categorical: Whether the customer uses streaming TV
* Does using streaming TV impact customer data use?

StreamingMovies

* Categorical: Whether the customer uses streaming movies
* Does using streaming movies impact customer data use?

Tenure

* Continuous: Number of months the customer has stayed with the provider
* Does the length of customer tenure impact their data use?

# B – Multiple Linear Regression

## B1 – Assumptions

Multiple Regression requires a number of assumptions in order to be applied to the data set for analysis. These are:

1. There is a linear relationship between the dependent variable and the independent variables. This can be verified through bivariate visualizations between the dependent and independent variables.
2. The independent variables are not too highly correlated with each other. This can be verified using the Variance Inflation Factor, with those values having a VIF above 10 having a high likelihood of multicollinearity.
3. Observations are selected independently and randomly from the population. This can be verified by looking for distinct patterns within a residuals plot.
4. Homoscedasticity – Variance of the sequence of variables should be homogenous. This is usually determined by finding patterns in a residual plot that indicate high degrees of variance in errors(can often be ‘cone’ shaped).

## B2 – Benefits of Python in Linear Regression

I chose to approach this project using python. My two primary reasons are:

1. Python is a language I am both professionally and academically experienced in.
2. Incredibly robust and powerful tools and libraries are available to assist in performing data analysis in Python.

The libraries I primarily used for this task are:

* Pandas – A framework for managing data utilizing powerful and robust objects called data frames.
* NumPy – A library of mathematical functions that expand the mathematical analysis available in vanilla Python.
* Statsmodels – A library full of tools for performing statistical tests and analyses, includes the OLS function that enabled my linear regression analysis.
* SciKit Learn – A library of tools for scientific analysis. Used primarily in this project for data scaling and variable re-expression.
* MatPlotLib – A library of tools for plotting various graphs and charts.
* Seaborn – An extension of MatPlotLib designed to provide more specialized graphs and charts for MatPlotLib users.
* SciPy – Another library of tools for scientific analysis. Used particularly for this project were the statistical analysis tools available in the stats package.
* MissingNo – A highly specialized tool for visualizing missing data within data frames

## B3 – Multiple Linear Regression in Hypothesis Testing

Multiple linear regression was the most appropriate tool for this analysis as it is designed to compare a continuous value with one or more continuous or categorical variables. Given I was testing what independent variables impacted the continuous variable, Bandwidth\_GB\_Year, the approach fit the hypothesis being tested. Another benefit was my ability to cast a large net and test a large number of independent variables for the sake of developing a model that could provide insight into action for the project stakeholders.

# C – Data Preparation

## C1 – Data Cleaning

My cleaning plan primarily focused on rooting out duplicate records, missing values, and outliers.

### Duplicate Values

To find duplicate records, I simply used the command: print(df.duplicated().value\_counts()). The output was:

False 10000

Name: count, dtype: int64

This suggested that no duplicate records were found in the set, so no further action was required.

### Missing Values

My next action was to examine the set for missing values using the command *print(df.isna().any()).* This gave me a high level view of which columns contain missing, NA, or NaN values in *any* of their records. The results of the analysis were:

CaseOrder False

Customer\_id False

Interaction False

UID False

City False

State False

County False

Zip False

Lat False

Lng False

Population False

Area False

TimeZone False

Job False

Children False

Age False

Income False

Marital False

Gender False

Churn False

Outage\_sec\_perweek False

Email False

Contacts False

Yearly\_equip\_failure False

Techie False

Contract False

Port\_modem False

Tablet False

InternetService True

Phone False

Multiple False

OnlineSecurity False

OnlineBackup False

DeviceProtection False

TechSupport False

StreamingTV False

StreamingMovies False

PaperlessBilling False

PaymentMethod False

Tenure False

MonthlyCharge False

Bandwidth\_GB\_Year False

Item1 False

Item2 False

Item3 False

Item4 False

Item5 False

Item6 False

Item7 False

Item8 False

With missing values found in InternetService, I decided to examine the column a little more closely with *print(df[‘InternetService’].isna().value\_counts()).* The results were:

False 7871

True 2129

Name: count, dtype: int64

The missing values were found to be significant enough to require mitigation, though they fell short of the threshold for removing the column completely. My initial instinct was to fill the values with the mode, as the column represented a dichotomous categorical variable. I did this with *df.fillna({‘InternetService’ : df[‘InternetService’].mode()[0]}, inplace=True).* This method did work to mitigate the missing values, as expressed below:

InternetService

False 10000

Name: count, dtype: int64

The real problem with this method comes with examining the distribution of values before and after imputation. Below is the result of a report generated to examine the phenomenon:

Percentage before imputation: InternetService

Fiber Optic 56.003049

DSL 43.996951

Name: count, dtype: float64

Count before imputation: InternetService

Fiber Optic 4408

DSL 3463

Name: count, dtype: int64

Percentage after imputation: InternetService

Fiber Optic 65.37

DSL 34.63

Name: count, dtype: float64

Count after imputation: InternetService

Fiber Optic 6537

DSL 3463

Name: count, dtype: int64

The distribution of values changed pretty dramatically. To determine a better fill method, I performed a missing number matrix analysis.

A close-up of a bar code

Description automatically generated

Given the even distribution of missing values, I decided to try a forward fill with the following code: *df[‘InternetService’] = df[‘InternetService’].ffill()*

The result of the analysis ended up being much closer to the original distribution:

Percentage before imputation: InternetService

Fiber Optic 56.003049

DSL 43.996951

Name: count, dtype: float64

Count before imputation: InternetService

Fiber Optic 4408

DSL 3463

Name: count, dtype: int64

Percentage after imputation: InternetService

Fiber Optic 56.08

DSL 43.92

Name: count, dtype: float64

Count after imputation: InternetService

Fiber Optic 5608

DSL 4392

Name: count, dtype: int64

The new percentage fall within a reasonable margin of error and accurately represent the original distributions.

### Outliers

My method for detecting outliers was a combination of three techniques. I printed a report that featured minimum, maximum, range, and mean of all the numeric variables. This report also features an evaluation of values with z-scores above 3 and below -3. It also examines values above Q4 and below Q1, with a minimum, maximum, and range of outlier values outside of the Q1-Q4 range.

The other two techniques involve plotting the variables for visual evaluation of outliers. The plots chosen for the analysis were histograms and box plots. Below is my evaluation of each independent numeric variable.

### Evaluation

I have organized my evaluation into three sections per variable.

1. The Python generated report of values and ranges
2. The generated plots
3. Conclusions from the analysis and justification for chose mitigation method

#### Population

##### Report

++++=========Population========++++

Minimum value: 0

Maximum value: 111850

Value range: 111850

Mean Value: 9756.5624

Values with a z-score over 3: 219

Values with a z-score under -3: 0

Values below IQR lower threshold: 0

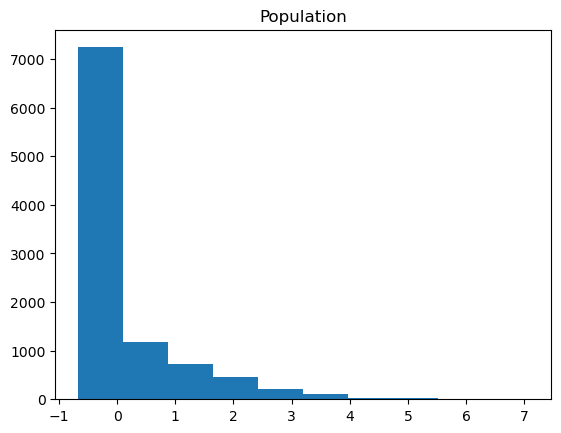
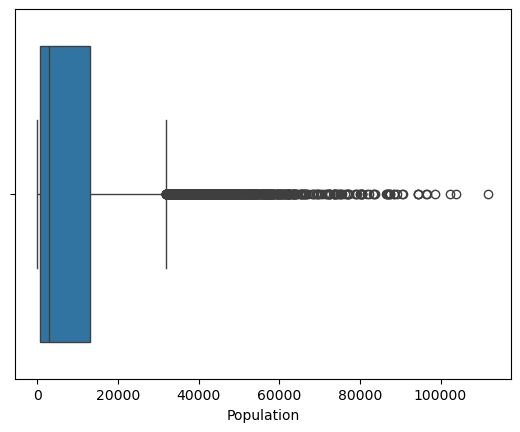
Values above IQR upper threshold: 937

Minimum Outlier Value: 31816

Maximum Outlier Value: 111850

Range of Outlier Values: 80034

##### Graphs

##### Conclusion and justification

While the mean value is ~9,756, the outliers found do not represent values that could be considered outside of the norm. A maximum value of 111,850 does not appear erroneous. For these reasons I’ve decided to retain outliers in this case.

#### Children

##### Report

++++=========Children========++++

Minimum value: 0

Maximum value: 10

Value range: 10

Mean Value: 2.0877

Values with a z-score over 3: 191

Values with a z-score under -3: 0

Values below IQR lower threshold: 0

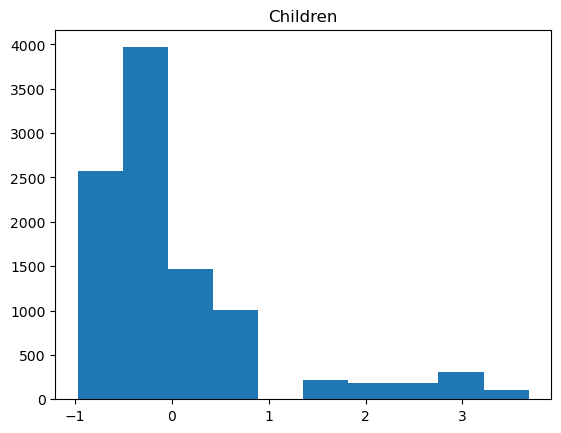
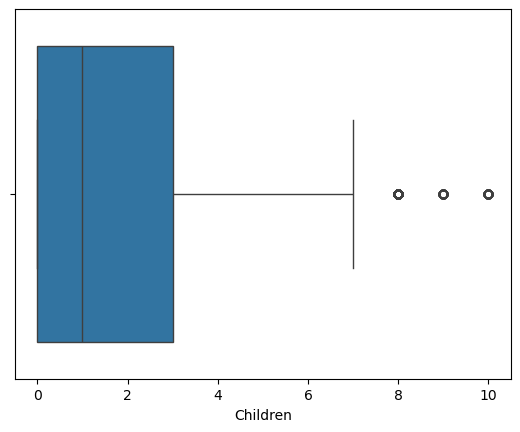
Values above IQR upper threshold: 401

Minimum Outlier Value: 8

Maximum Outlier Value: 10

Range of Outlier Values: 2

##### Graphs

##### Conclusion and justification

The histogram featured shows the isolated columns generally associated with outliers. The count of values above Q4 is 401, or roughly 0.04% of the set. The rang of values is small with a minimum of 0 and a maximum of 10. The values that are above the Q4 threshold represent ~0.04% of the set. With these considerations, I decided to retain the found outliers.

#### Age

##### Report

++++=========Age========++++

Minimum value: 18

Maximum value: 89

Value range: 71

Mean Value: 53.0784

Values with a z-score over 3: 0

Values with a z-score under -3: 0

Values below IQR lower threshold: 0

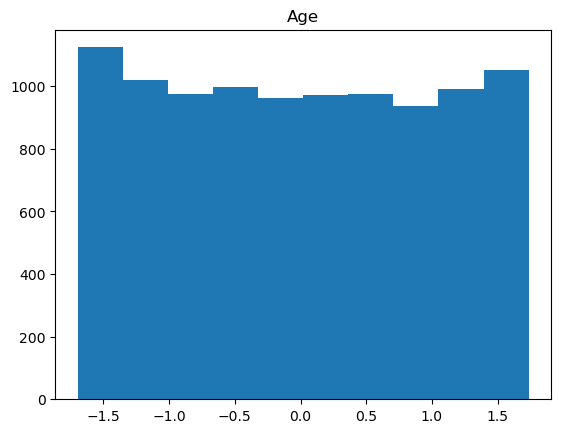
Values above IQR upper threshold: 0

Minimum Outlier Value: nan

Maximum Outlier Value: nan

Range of Outlier Values: nan

##### Graphs

##### Conclusion and justification

No outliers were found in the Age column, so no further mitigation or analysis was necessary.

#### Income

##### Report

++++=========Income========++++

Minimum value: 348.67

Maximum value: 258900.7

Value range: 258552.03

Mean Value: 39806.926771

Values with a z-score over 3: 145

Values with a z-score under -3: 0

Values below IQR lower threshold: 0

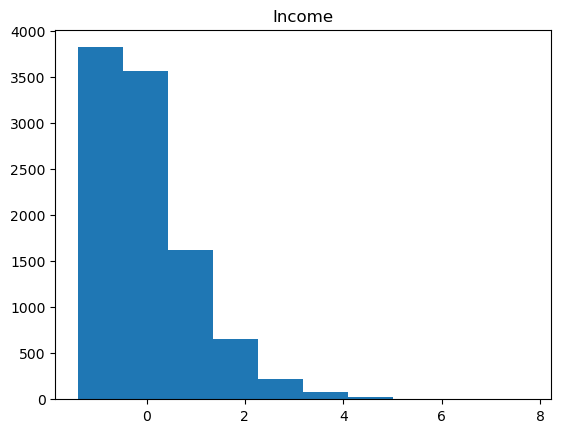
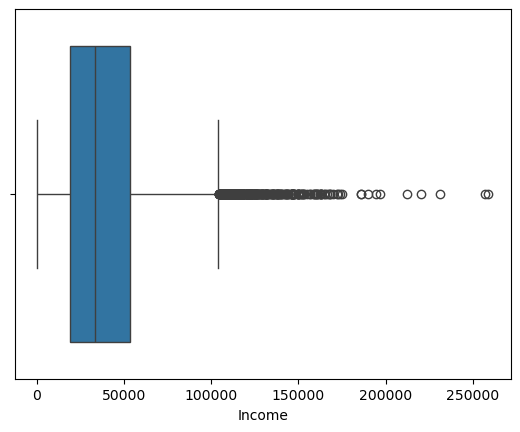
Values above IQR upper threshold: 336

Minimum Outlier Value: 104362.5

Maximum Outlier Value: 258900.7

Range of Outlier Values: 154538.2

##### Graphs

##### Conclusion and justification

Outliers were found within the column. These were expected within income and represent reasonable values. None of the values appear to be erroneous. For those reasons my chosen course of treatment was to retain the values.

#### Email

##### Report

++++=========Email========++++

Minimum value: 1

Maximum value: 23

Value range: 22

Mean Value: 12.016

Values with a z-score over 3: 3

Values with a z-score under -3: 9

Values below IQR lower threshold: 23

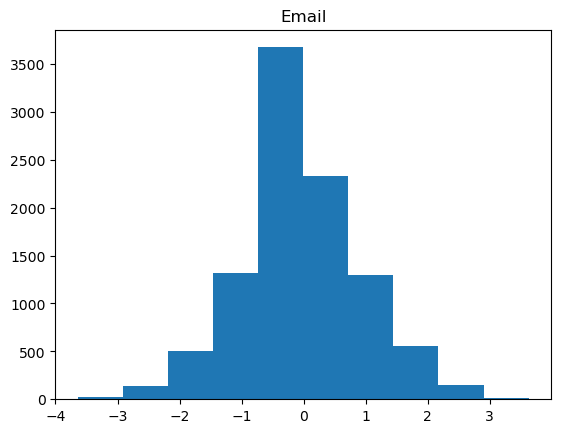
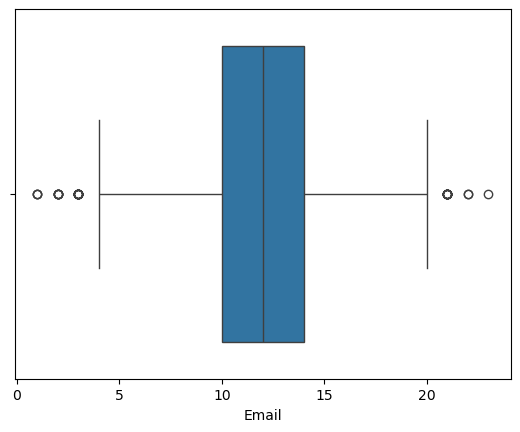
Values above IQR upper threshold: 15

Minimum Outlier Value: 1

Maximum Outlier Value: 23

Range of Outlier Values: 22

##### Graphs

##### Conclusion and justification

Email appears to have a small number of outliers below Q1 and above Q4. The range of total values is limited between 1 and 23. The values outside of range also represent a negligible portion of the data, ~0.0045% of the set. For these reasons, I decided to retain the outliers as my treatment.

#### Contacts

##### Report

++++=========Contacts========++++

Minimum value: 0

Maximum value: 7

Value range: 7

Mean Value: 0.9942

Values with a z-score over 3: 165

Values with a z-score under -3: 0

Values below IQR lower threshold: 0

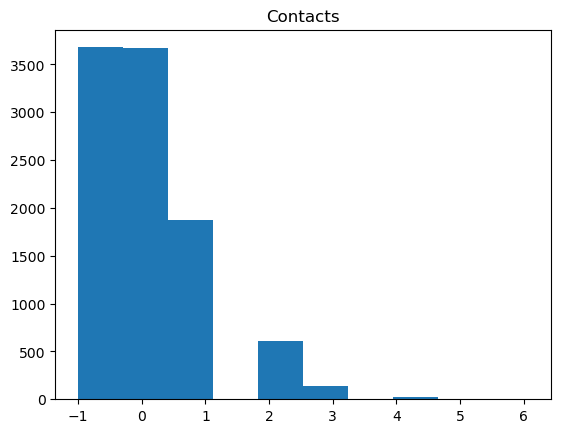
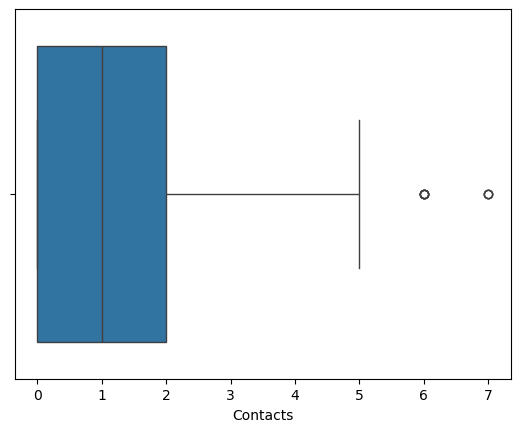
Values above IQR upper threshold: 8

Minimum Outlier Value: 6

Maximum Outlier Value: 7

Range of Outlier Values: 1

##### Graphs

##### Conclusion and justification

Contacts features a small number of outliers above Q4. The range of total value is 0 – 7, with the rang of outliers being 6 – 7. The total number of values above Q4 is 8, which represents a negligible percentage of the 10,000 records in the column. For these reasons, I have decided the appropriate treatment would be to retain the outliers.

#### Yearly\_equip\_failure

##### Report

++++=========Yearly\_equip\_failure========++++

Minimum value: 0

Maximum value: 6

Value range: 6

Mean Value: 0.398

Values with a z-score over 3: 94

Values with a z-score under -3: 0

Values below IQR lower threshold: 0

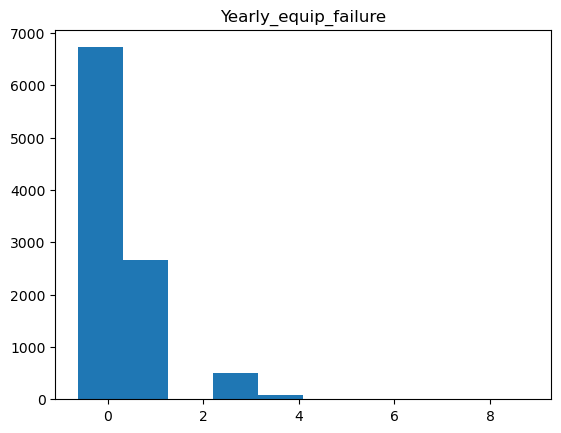
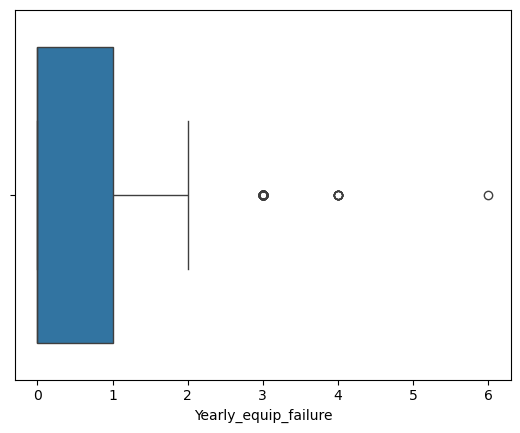
Values above IQR upper threshold: 94

Minimum Outlier Value: 3

Maximum Outlier Value: 6

Range of Outlier Values: 3

##### Graphs

##### Conclusion and justification

Yearly\_equip\_failure represents a particularly small range of values: 0-6. There were 94 values found above the Q4 threshold. These values represent a small portion of the data set at 0.009%. Given these factors I decided to retain the values as my treatment of the outliers.

#### Tenure

##### Report

++++=========Tenure========++++

Minimum value: 1.00025934

Maximum value: 71.99928

Value range: 70.99902066

Mean Value: 34.5261880889938

Values with a z-score over 3: 0

Values with a z-score under -3: 0

Values below IQR lower threshold: 0

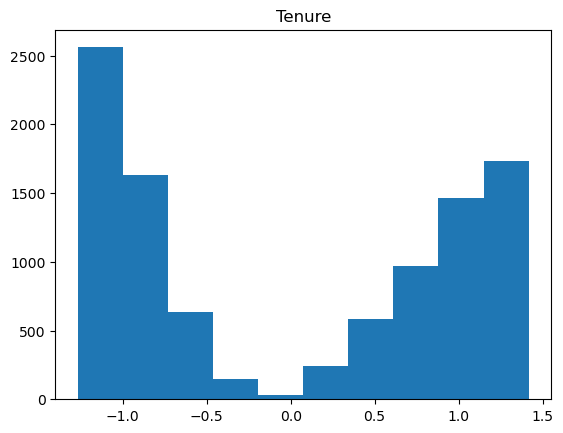
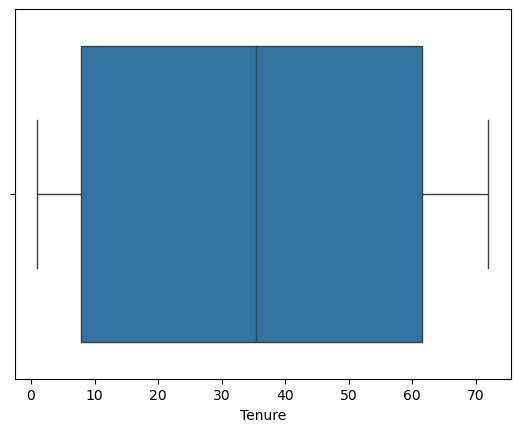
Values above IQR upper threshold: 0

Minimum Outlier Value: nan

Maximum Outlier Value: nan

Range of Outlier Values: nan

##### Graphs

##### Conclusion and justification

No outliers were found in the Tenure column, so no further mitigation or analysis was necessary.

## C2 – Variables

For each variable in the set, I generated a report of relevant statistical information using the below formula:

associated\_variables = ['Bandwidth\_GB\_Year', 'State', 'Population', 'Area', 'Children', 'Age', 'Income', 'Marital', 'Gender', 'Email', 'Contacts',

'Yearly\_equip\_failure', 'Techie', 'Contract', 'Port\_modem', 'Tablet', 'InternetService', 'Phone', 'Multiple', 'OnlineSecurity',

'OnlineBackup', 'StreamingTV', 'StreamingMovies', 'Tenure']

#Print reports for associated variables

for variable in associated\_variables:

print(f'========Begin Report========')

print(f'========{variable}========')

print(df[variable].describe())

if pd.api.types.is\_numeric\_dtype(df[variable]):

print(f'Median: {df[variable].median()}')

else:

print(f'Unique Values: {df[variable].unique()}')

print(f'Unique Value Counts: {df[variable].value\_counts()}')

print(f'========End Report========')

Each report leverages the *describe()* function to get some basic statistics. Further statistics were generated based on whether the values in the column were numeric or categorical. For this I use *pd.api.types.is\_numeric\_dtype()* to distinguish between the two. For numeric values I also included the median in the report. For the categorical values, I included a list of the unique values present in the column and a count of each unique value.

### Dependent Variable

Bandwidth\_GB\_Year:

========Bandwidth\_GB\_Year========

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

Median: 3279.536903

### Independent Variables

State:

========State========

count 10000

unique 52

top TX

freq 603

Name: State, dtype: object

Unique Values: ['AK' 'MI' 'OR' 'CA' 'TX' 'GA' 'TN' 'OK' 'FL' 'OH' 'PA' 'PR' 'IA' 'ME'

'IL' 'WI' 'NC' 'AL' 'NM' 'VT' 'MD' 'NY' 'WA' 'CT' 'NJ' 'DC' 'ND' 'LA'

'NE' 'WV' 'AZ' 'MO' 'WY' 'MT' 'VA' 'KY' 'MN' 'KS' 'MA' 'IN' 'SC' 'NH'

'DE' 'MS' 'ID' 'AR' 'SD' 'CO' 'HI' 'UT' 'RI' 'NV']

Unique Value Counts: State

TX 603

NY 558

PA 550

CA 526

IL 413

OH 359

FL 324

MO 310

VA 285

NC 280

IA 279

MI 279

MN 264

WV 247

IN 241

GA 238

KY 238

WI 228

OK 203

KS 195

NJ 190

TN 185

AL 181

NE 181

AR 176

WA 175

MA 172

CO 155

LA 141

MS 126

SC 124

MD 123

ND 118

NM 114

OR 114

AZ 112

ME 112

SD 101

MT 96

NH 85

VT 84

ID 81

AK 77

CT 71

UT 66

NV 48

WY 43

PR 40

HI 35

DE 21

RI 19

DC 14

Name: count, dtype: int64

Population:

========Population========

count 10000.000000

mean 9756.562400

std 14432.698671

min 0.000000

25% 738.000000

50% 2910.500000

75% 13168.000000

max 111850.000000

Name: Population, dtype: float64

Median: 2910.5

Area:

========Area========

count 10000

unique 3

top Suburban

freq 3346

Name: Area, dtype: object

Unique Values: ['Urban' 'Suburban' 'Rural']

Unique Value Counts: Area

Suburban 3346

Urban 3327

Rural 3327

Name: count, dtype: int64

Children:

========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

Median: 1.0

Age:

========Age========

count 10000.000000

mean 53.078400

std 20.698882

min 18.000000

25% 35.000000

50% 53.000000

75% 71.000000

max 89.000000

Name: Age, dtype: float64

Median: 53.0

Income:

========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

Median: 33170.604999999996

Marital:

========Marital========

count 10000

unique 5

top Divorced

freq 2092

Name: Marital, dtype: object

Unique Values: ['Widowed' 'Married' 'Separated' 'Never Married' 'Divorced']

Unique Value Counts: Marital

Divorced 2092

Widowed 2027

Separated 2014

Never Married 1956

Married 1911

Name: count, dtype: int64

Gender:

========Gender========

count 10000

unique 3

top Female

freq 5025

Name: Gender, dtype: object

Unique Values: ['Male' 'Female' 'Nonbinary']

Unique Value Counts: Gender

Female 5025

Male 4744

Nonbinary 231

Name: count, dtype: int64

Email:

========Email========

count 10000.000000

mean 12.016000

std 3.025898

min 1.000000

25% 10.000000

50% 12.000000

75% 14.000000

max 23.000000

Name: Email, dtype: float64

Median: 12.0

Contacts:

========Contacts========

count 10000.000000

mean 0.994200

std 0.988466

min 0.000000

25% 0.000000

50% 1.000000

75% 2.000000

max 7.000000

Name: Contacts, dtype: float64

Median: 1.0

Yearly\_equip\_failure:

========Yearly\_equip\_failure========

count 10000.000000

mean 0.398000

std 0.635953

min 0.000000

25% 0.000000

50% 0.000000

75% 1.000000

max 6.000000

Name: Yearly\_equip\_failure, dtype: float64

Median: 0.0

Techie:

========Techie========

count 10000

unique 2

top No

freq 8321

Name: Techie, dtype: object

Unique Values: ['No' 'Yes']

Unique Value Counts: Techie

No 8321

Yes 1679

Name: count, dtype: int64

Contract:

========Contract========

count 10000

unique 3

top Month-to-month

freq 5456

Name: Contract, dtype: object

Unique Values: ['One year' 'Month-to-month' 'Two Year']

Unique Value Counts: Contract

Month-to-month 5456

Two Year 2442

One year 2102

Name: count, dtype: int64

Port\_modem:

========Port\_modem========

count 10000

unique 2

top No

freq 5166

Name: Port\_modem, dtype: object

Unique Values: ['Yes' 'No']

Unique Value Counts: Port\_modem

No 5166

Yes 4834

Name: count, dtype: int64

========End Report========

Tablet:

========Tablet========

count 10000

unique 2

top No

freq 7009

Name: Tablet, dtype: object

Unique Values: ['Yes' 'No']

Unique Value Counts: Tablet

No 7009

Yes 2991

Name: count, dtype: int64

InternetService:

========InternetService========

count 7871

unique 2

top Fiber Optic

freq 4408

Name: InternetService, dtype: object

Unique Values: ['Fiber Optic' 'DSL' nan]

Unique Value Counts: InternetService

Fiber Optic 4408

DSL 3463

Name: count, dtype: int64

Phone:

========Phone========

count 10000

unique 2

top Yes

freq 9067

Name: Phone, dtype: object

Unique Values: ['Yes' 'No']

Unique Value Counts: Phone

Yes 9067

No 933

Name: count, dtype: int64

Multiple:

========Multiple========

count 10000

unique 2

top No

freq 5392

Name: Multiple, dtype: object

Unique Values: ['No' 'Yes']

Unique Value Counts: Multiple

No 5392

Yes 4608

Name: count, dtype: int64

OnlineSecurity:

========OnlineSecurity========

count 10000

unique 2

top No

freq 6424

Name: OnlineSecurity, dtype: object

Unique Values: ['Yes' 'No']

Unique Value Counts: OnlineSecurity

No 6424

Yes 3576

Name: count, dtype: int64

OnlineBackup:

========OnlineBackup========

count 10000

unique 2

top No

freq 5494

Name: OnlineBackup, dtype: object

Unique Values: ['Yes' 'No']

Unique Value Counts: OnlineBackup

No 5494

Yes 4506

Name: count, dtype: int64

StreamingTV:

========StreamingTV========

count 10000

unique 2

top No

freq 5071

Name: StreamingTV, dtype: object

Unique Values: ['No' 'Yes']

Unique Value Counts: StreamingTV

No 5071

Yes 4929

Name: count, dtype: int64

StreamingMovies:

========StreamingMovies========

count 10000

unique 2

top No

freq 5110

Name: StreamingMovies, dtype: object

Unique Values: ['Yes' 'No']

Unique Value Counts: StreamingMovies

No 5110

Yes 4890

Name: count, dtype: int64

Tenure:

========Tenure========

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

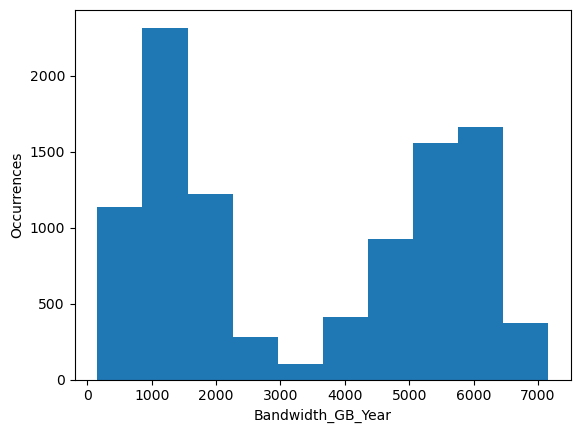
Median: 35.430506995

## C3 – Univariate and Bivariate Analyses

### Univariate Analysis

#### Dependent Variable

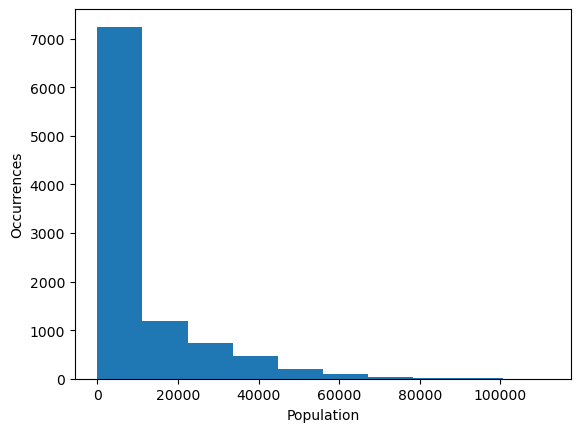
##### Bandwidth\_GB\_Year



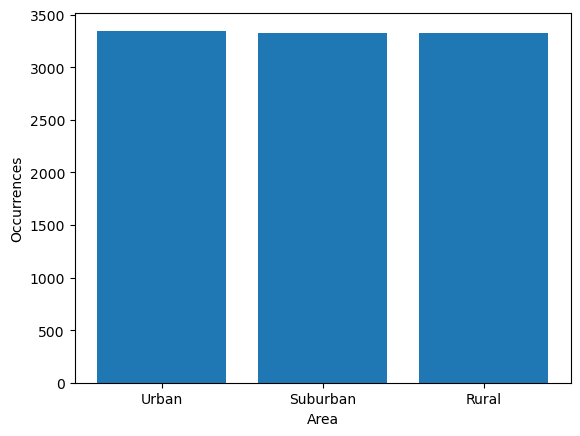
#### Independent Variables

##### State

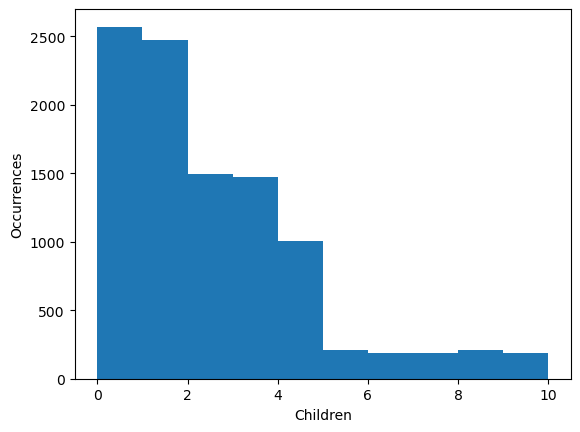
##### Population



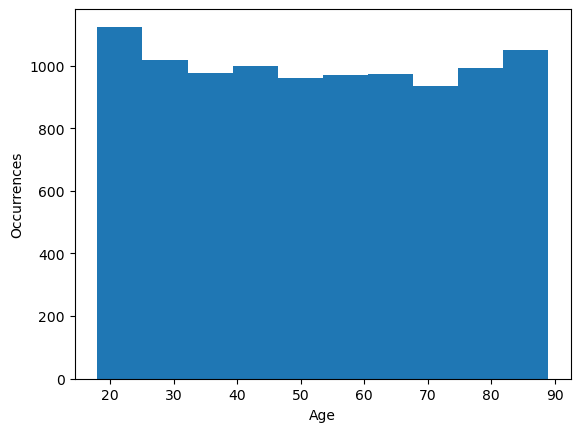
##### Area



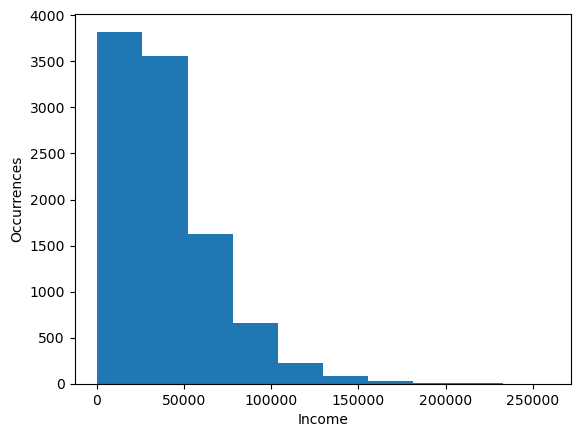
##### Children



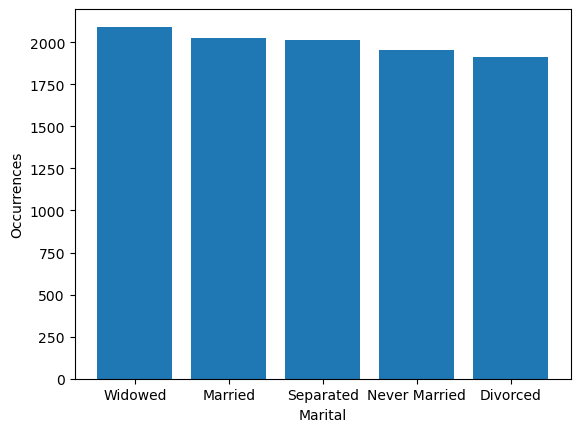
##### Age



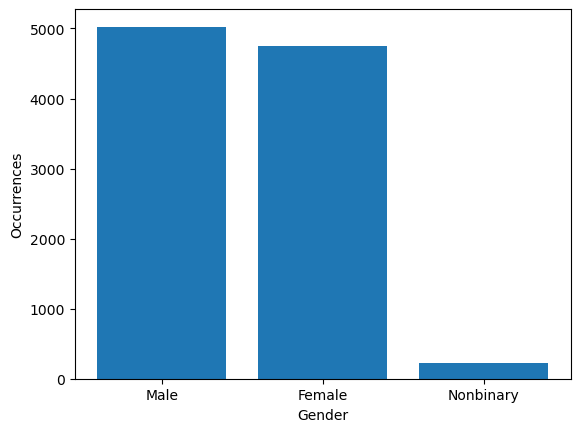
##### Income



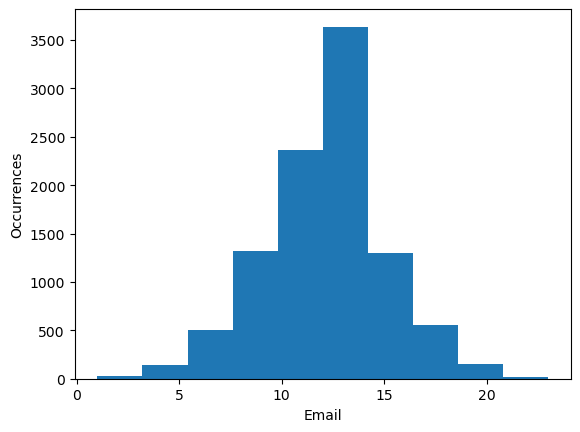
##### Marital



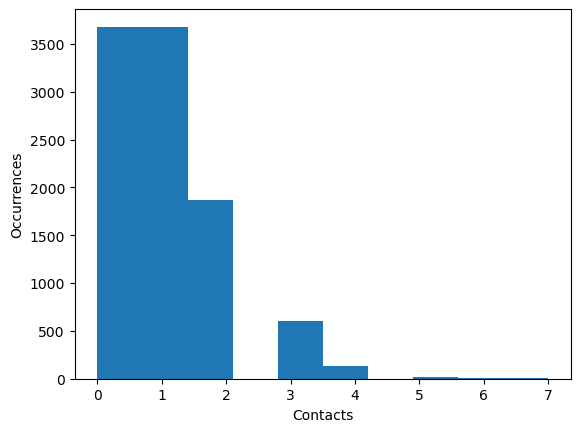
##### Gender



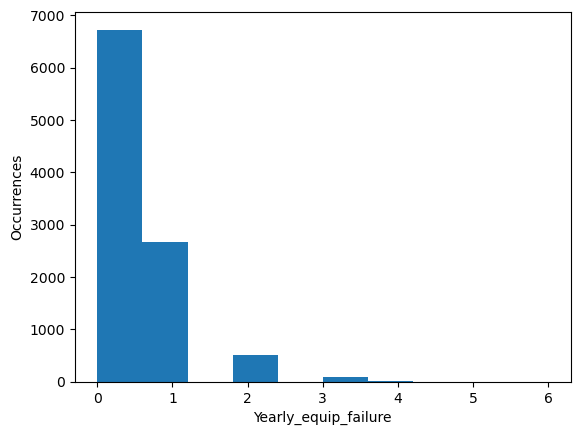
##### Email



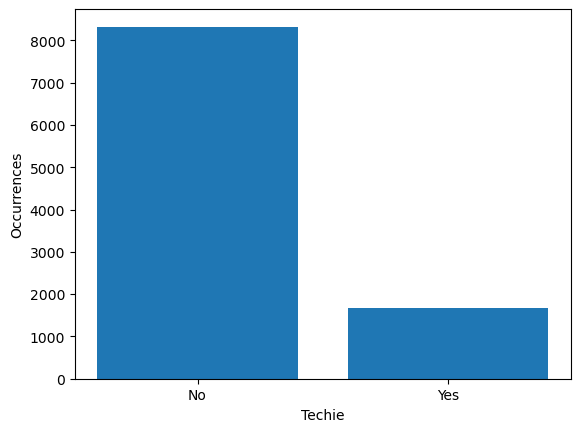
##### Contacts



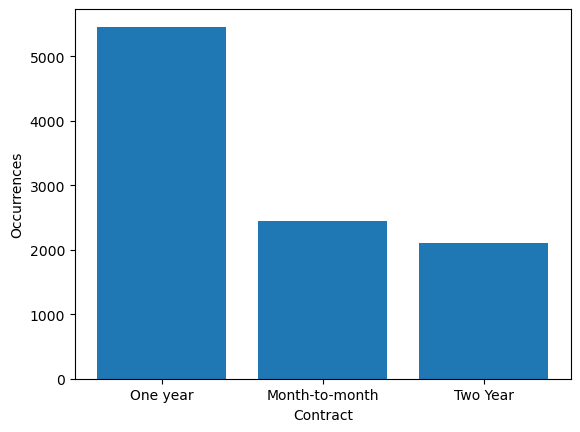
##### Yearly\_equip\_failure



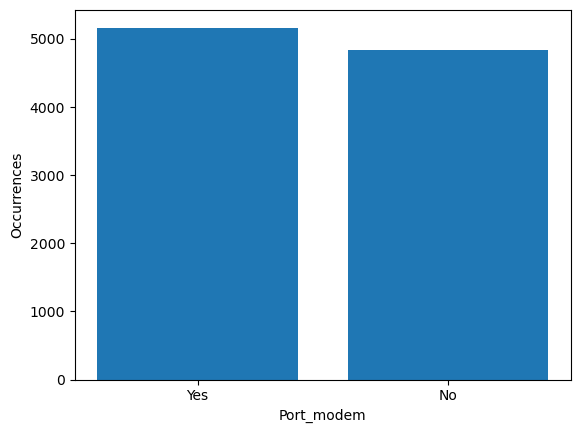
##### Techie



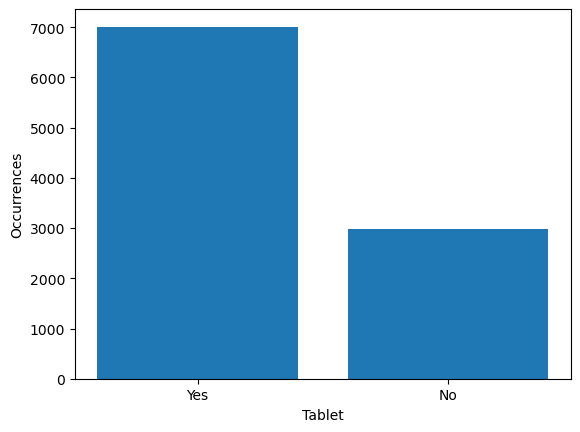
##### Contract



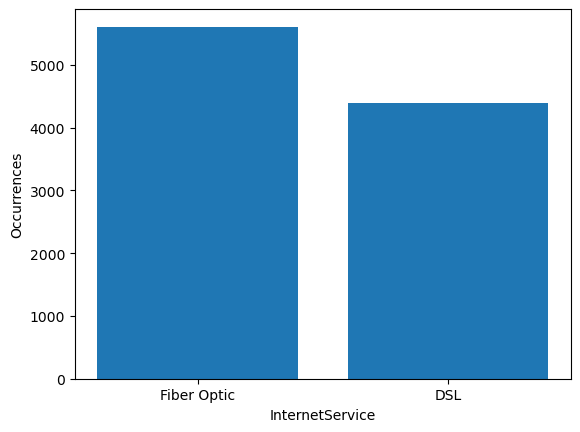
##### Port\_modem



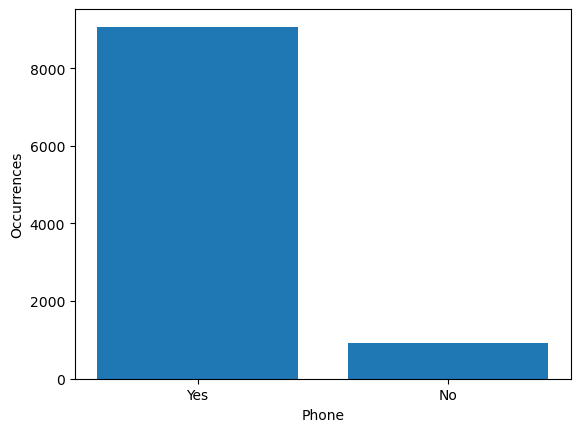
##### Tablet



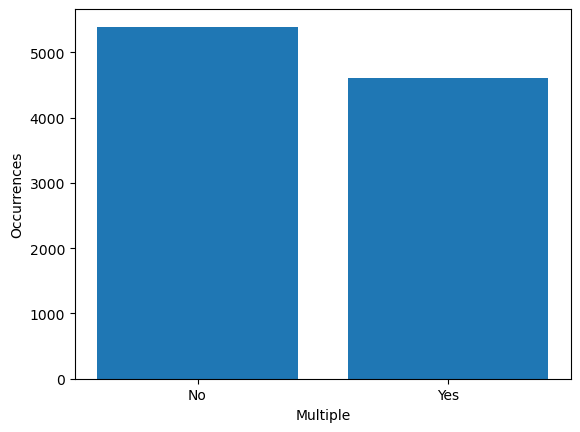
##### InternetService



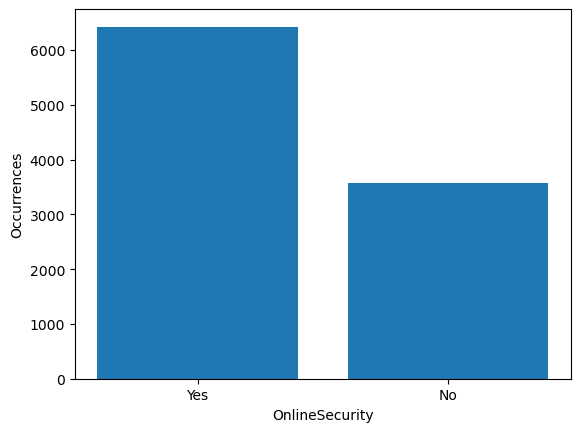
##### Phone



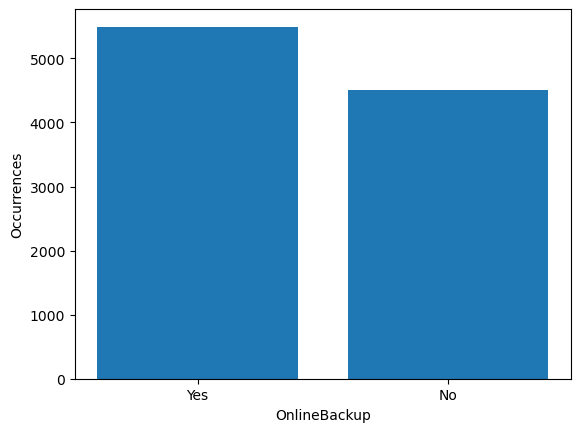
##### Multiple



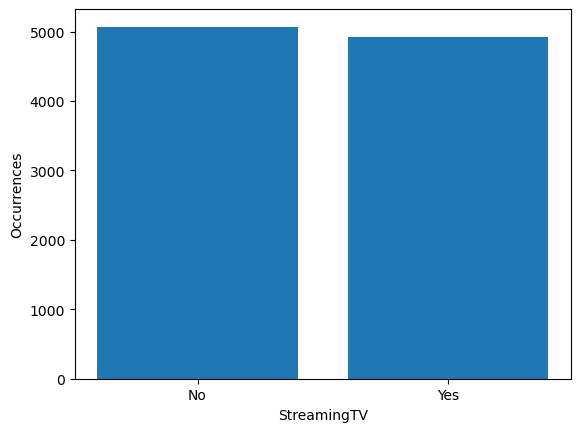
##### OnlineSecurity



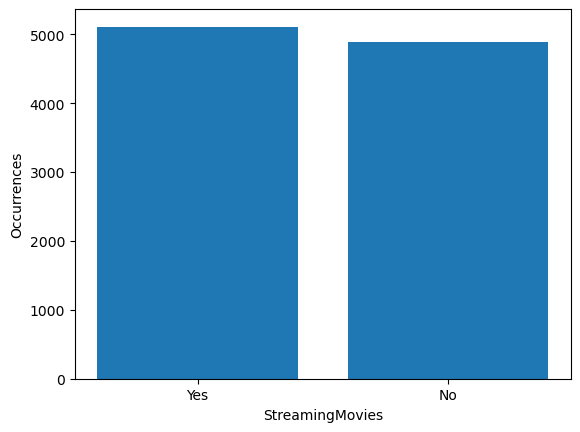
##### OnlineBackup



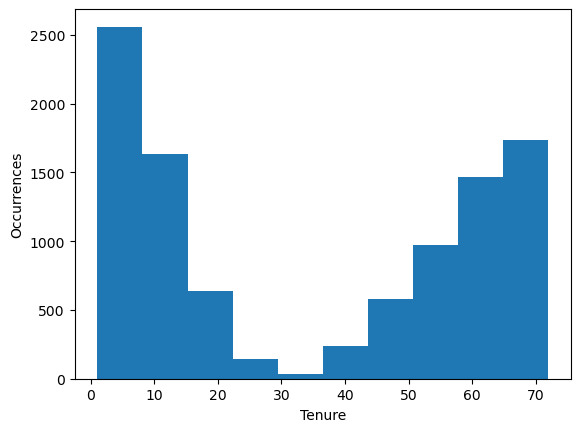
##### StreamingTV



##### StreamingMovies



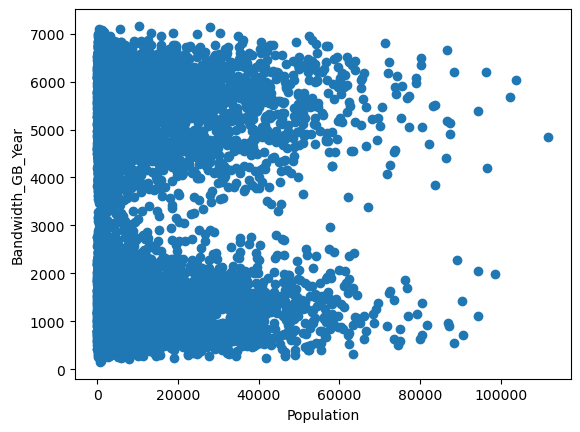
##### Tenure



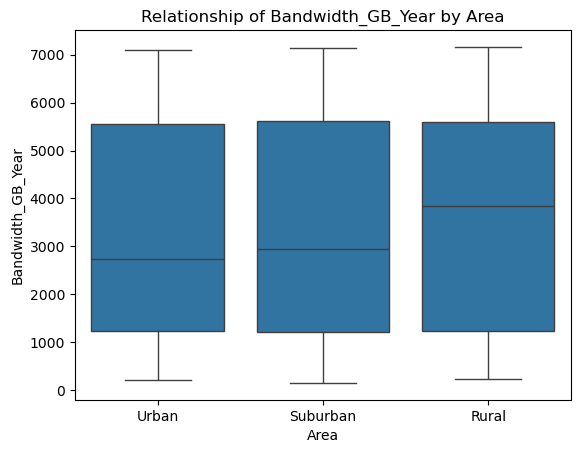
### Bivariate Analysis

#### Bandwidth\_GB\_Year vs. State

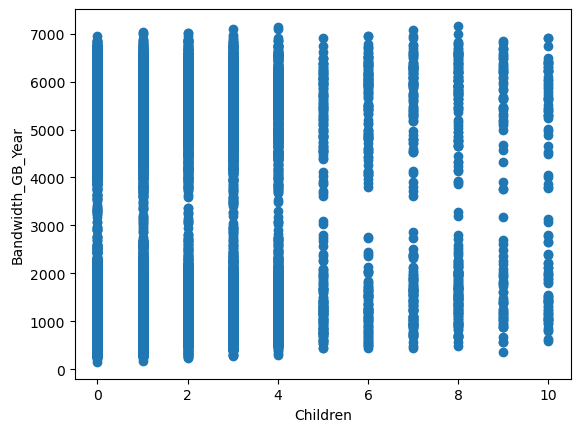
#### Bandwidth\_GB\_Year vs. Population



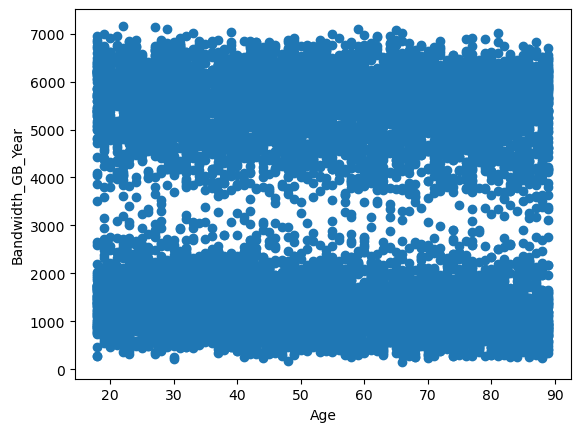
#### Bandwidth\_GB\_Year vs. Area



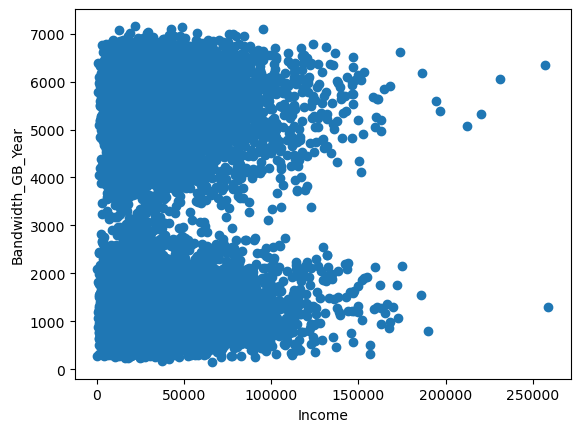
#### Bandwidth\_GB\_Year vs. Children



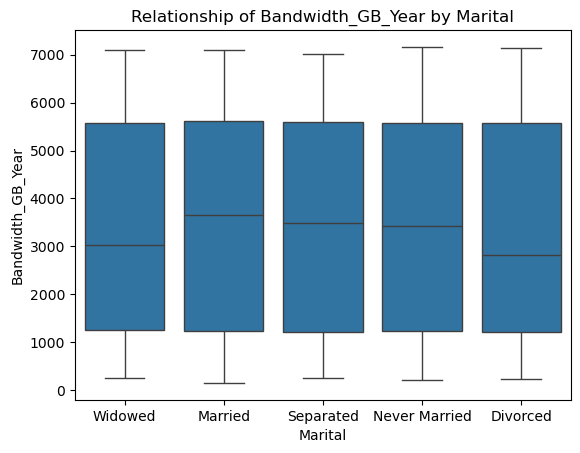
#### Bandwidth\_GB\_Year vs. Age



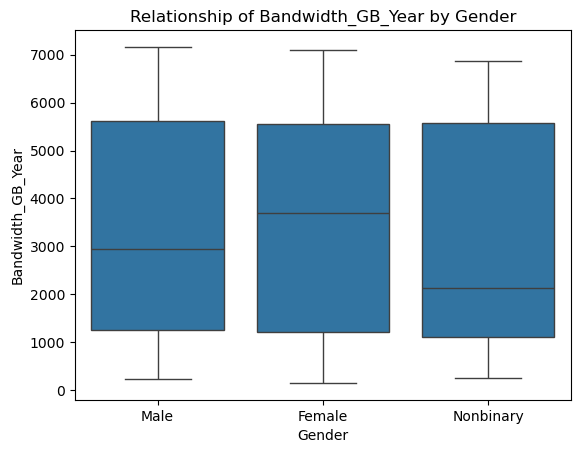
#### Bandwidth\_GB\_Year vs. Income



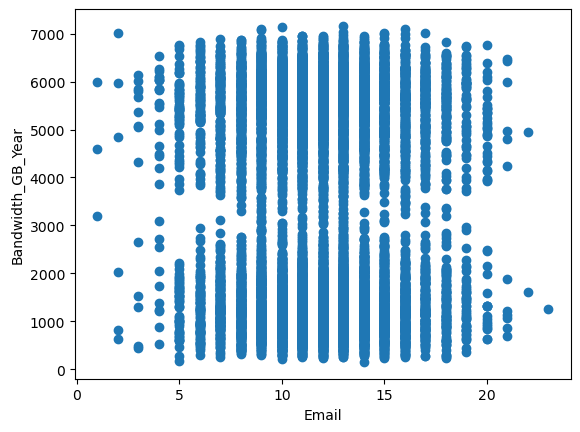
#### Bandwidth\_GB\_Year vs. Marital



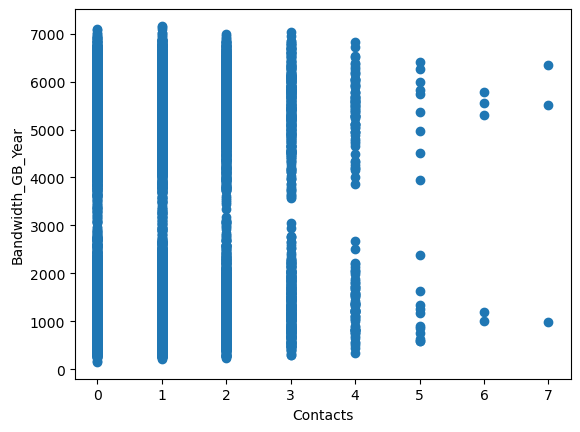
#### Bandwidth\_GB\_Year vs. Gender



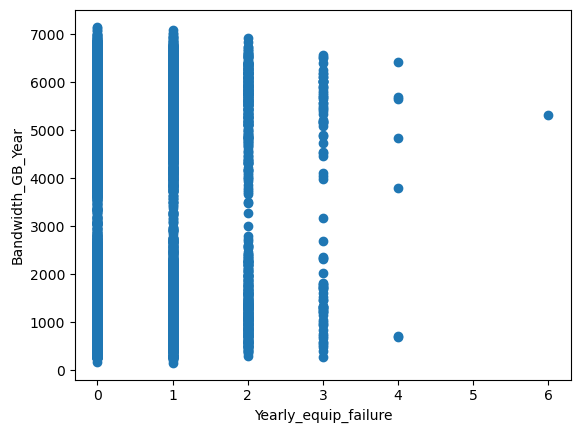
#### Bandwidth\_GB\_Year vs. Email



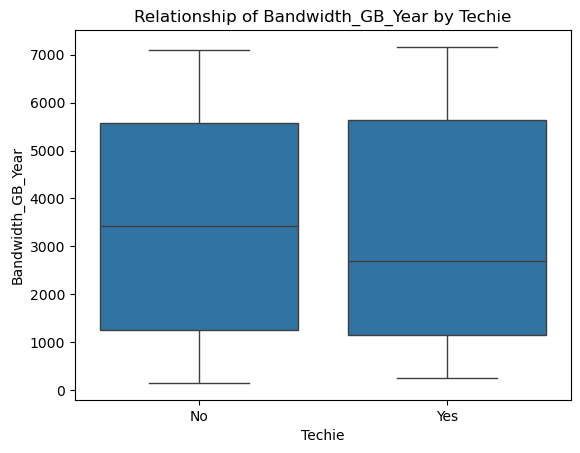
#### Bandwidth\_GB\_Year vs. Contacts



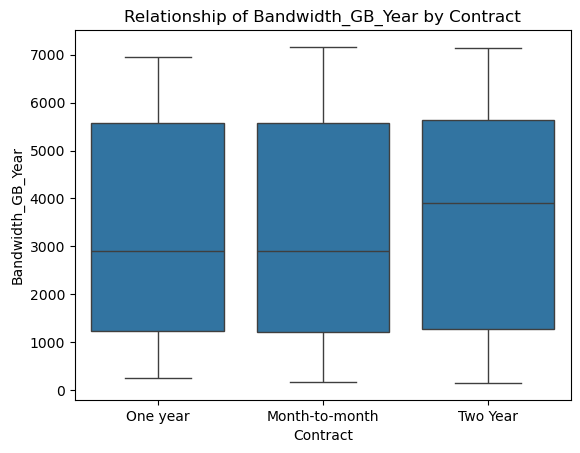
#### Bandwidth\_GB\_Year vs. Yearly\_equip\_failure



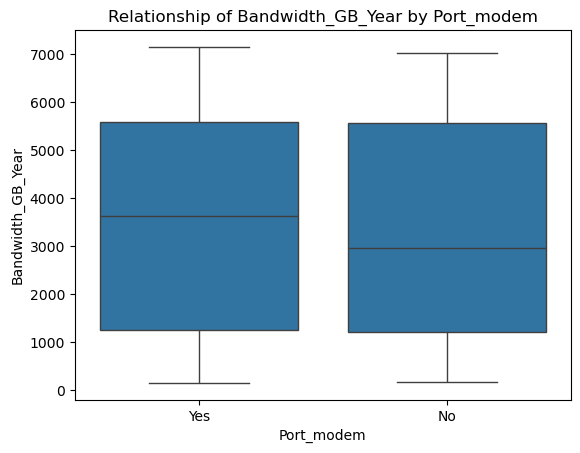
#### Bandwidth\_GB\_Year vs. Techie



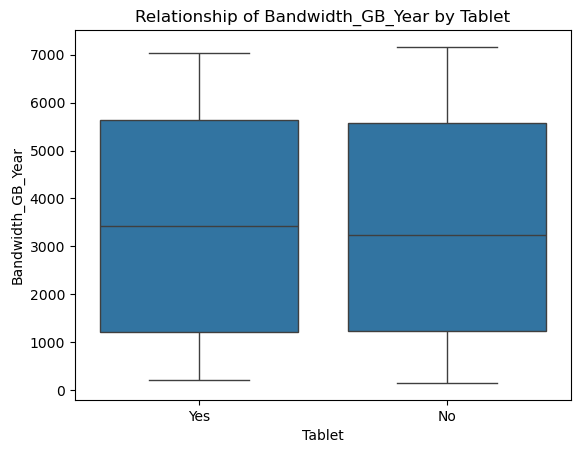
#### Bandwidth\_GB\_Year vs. Contract



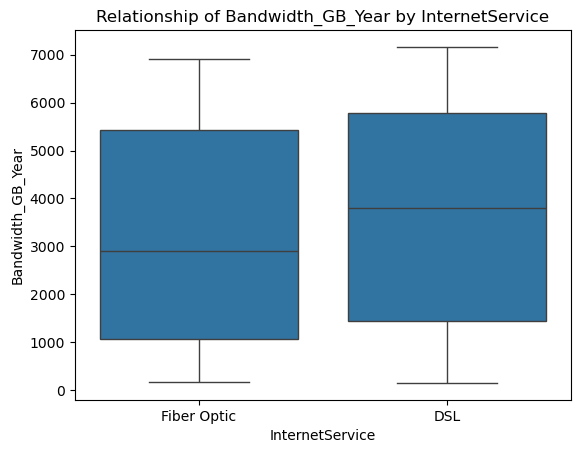
#### Bandwidth\_GB\_Year vs. Port\_modem



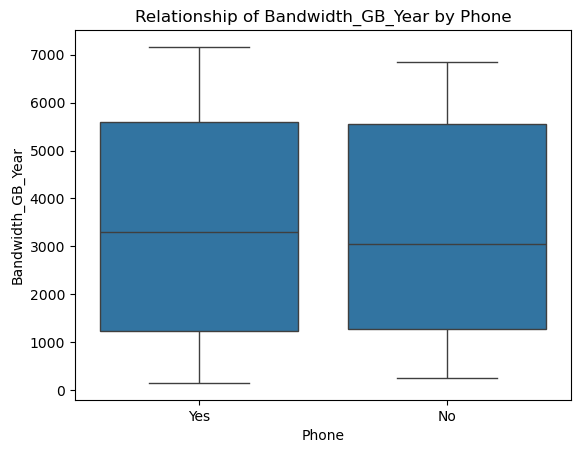
#### Bandwidth\_GB\_Year vs. Tablet



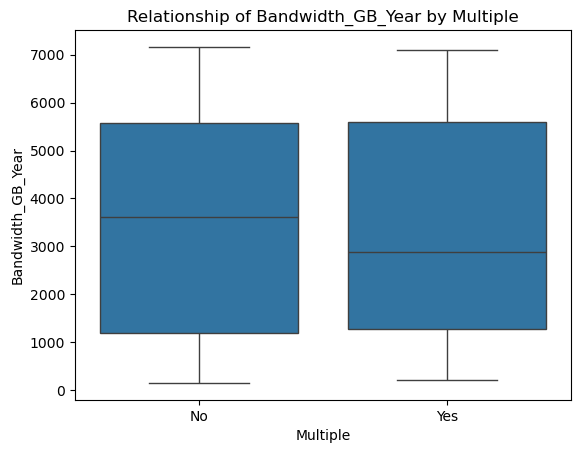
#### Bandwidth\_GB\_Year vs. InternetService



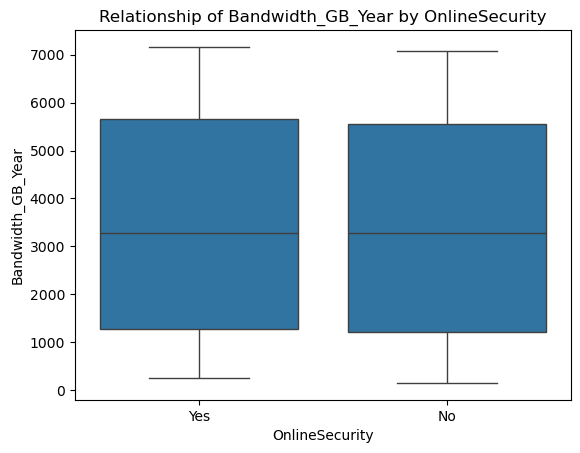
#### Bandwidth\_GB\_Year vs. Phone



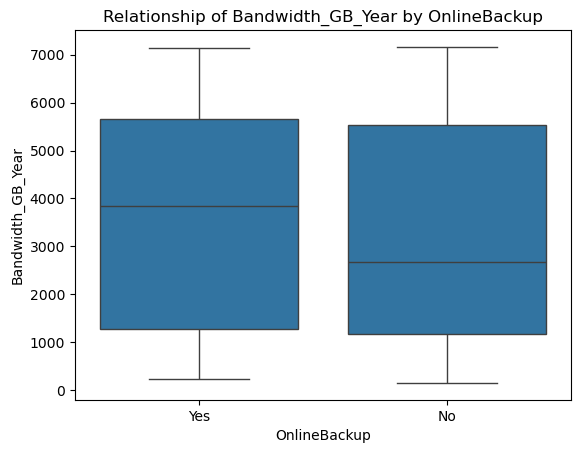
#### Bandwidth\_GB\_Year vs. Multiple



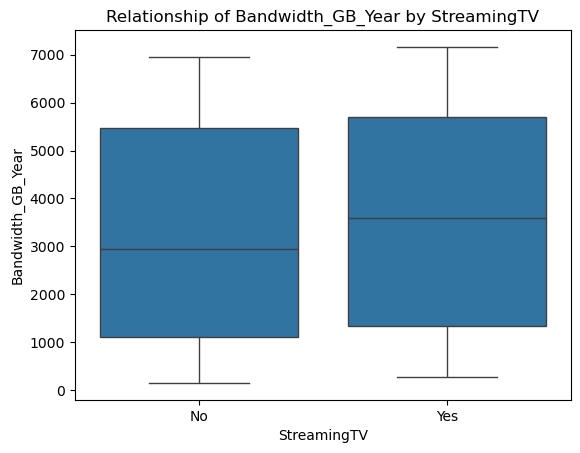
#### Bandwidth\_GB\_Year vs. OnlineSecurity



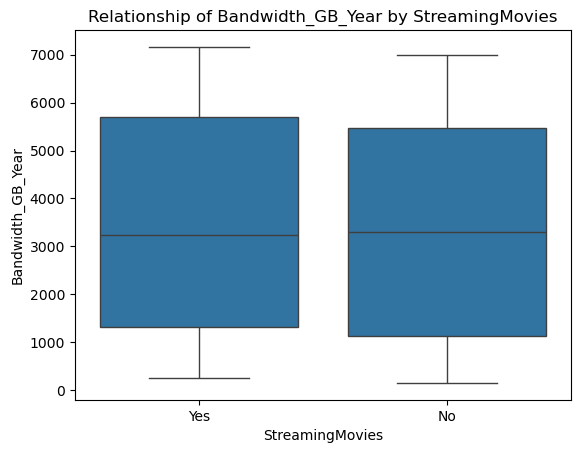
#### Bandwidth\_GB\_Year vs. OnlineBackup



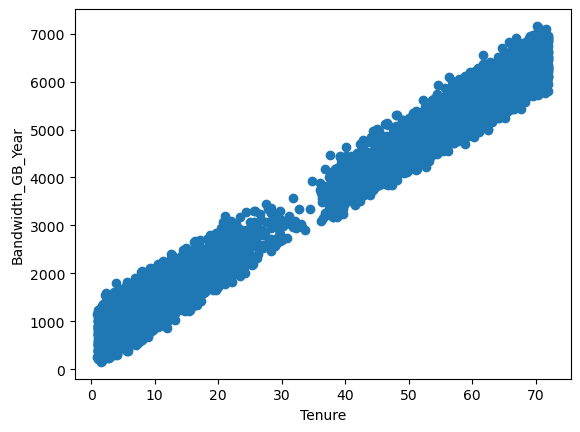
#### Bandwidth\_GB\_Year vs. StreamingTV



#### Bandwidth\_GB\_Year vs. StreamingMovies



#### Bandwidth\_GB\_Year vs. Tenure



## C4 – Data Transformation

My initial model includes a number of categorical variables with non-numeric values. For each of these I had to go through and make the decision on treatment. A cursory analysis using:

#Examine values present in categorical\_variables to determine treatment

for variable in categorical\_variables:

print(f’{variable} : {df[variable].unique()})

Yielded the following report:

Area : ['Urban' 'Suburban' 'Rural']

Marital : ['Widowed' 'Married' 'Separated' 'Never Married' 'Divorced']

Gender : ['Male' 'Female' 'Nonbinary']

Techie : ['No' 'Yes']

Contract : ['One year' 'Month-to-month' 'Two Year']

Port\_modem : ['Yes' 'No']

Tablet : ['Yes' 'No']

InternetService : ['Fiber Optic' 'DSL']

Phone : ['Yes' 'No']

Multiple : ['No' 'Yes']

OnlineSecurity : ['Yes' 'No']

OnlineBackup : ['Yes' 'No']

StreamingTV : ['No' 'Yes']

StreamingMovies : ['Yes' 'No']

My treatment for each of the variables was as follows:

Area – One-hot encoding based on Urban, Suburban, and Rural

Marital – One-hot encoding based on Widowed, Married, Separated, Never Married, Divorced

Gender – One-hot encoding based on Male, Female, Nonbinary

Techie – Ordinal encoding based on 0 = False, 1 = True

Contract – My initial thought was one-hot encoding, because the variable appeared nominal to me. Upon further inspection, I determined it to be ordinal given the following logic:

A one year contract is a longer term than month to month, a two year contract is a longer term than one year. Given this, I used the ordinal encoding: 0 = Month-to-month, 1 = One year, 2 = Two year.

Port\_modem - Ordinal encoding based on 0 = False, 1 = True

Tablet - Ordinal encoding based on 0 = False, 1 = True

InternetService – One-hot encoding based on Fiber Optic, DSL

Phone - Ordinal encoding based on 0 = False, 1 = True

Multiple - Ordinal encoding based on 0 = False, 1 = True

OnlineSecurity - Ordinal encoding based on 0 = False, 1 = True

OnlineBackup - Ordinal encoding based on 0 = False, 1 = True

StreamingTV - Ordinal encoding based on 0 = False, 1 = True

StreamingMovies - Ordinal encoding based on 0 = False, 1 = True

I used the following code to perform the ordinal encoding:

#Separate reports

print('##### Before Encoding ####')

#Examine values present in categorical\_variables to determine treatment

for variable in categorical\_variables:

print(f'{variable} : {df[variable].unique()}')

#Variables I'll apply one-hot encoding to

one\_hot\_columns = ['Area', 'Marital', 'Gender', 'InternetService']

#Variables I'll apply ordinal encoding to

ordinal\_columns = ['Techie', 'Port\_modem', 'Tablet', 'Phone', 'Multiple', 'OnlineSecurity', 'OnlineBackup', 'StreamingTV', 'StreamingMovies']

#Separate reports

print('##### After Encoding ####')

#I went back and forth with Contract, but determinied it was ordinal given that one year is a longer term than month-to-month

#And two years is a longer term than one year

df['Contract'] = df['Contract'].map({'Month-to-month' : 0, 'One year' : 1, 'Two Year' : 2})

print(f'Contract : {df['Contract'].unique()}')

for column in ordinal\_columns:

df[column] = df[column].map({'No' : 0, 'Yes' : 1})

print(f'{column} : {df[column].unique()}')

Most of the ordinal values were dichotomous, but I did have a special case in Contract, so I performed its treatment separately. Below is a follow-up report using *series.unique()* to ensure encoding was performed successfully:

Contract : [1 0 2]

Techie : [0 1]

Port\_modem : [1 0]

Tablet : [1 0]

Phone : [1 0]

Multiple : [0 1]

OnlineSecurity : [1 0]

OnlineBackup : [1 0]

StreamingTV : [0 1]

StreamingMovies : [1 0]

To perform the one-hot encoding, I brought in the OneHotEncoder from Scikit-learn. Below is my code for the treatment of the nominal variables:

#Perform One-Hot Encoding using Scikit-learn's OneHotEncoder

#https://www.geeksforgeeks.org/ml-one-hot-encoding/

#Initialize new encoder

encoder = OneHotEncoder(sparse\_output=False, drop=’first’)

#Encode values

one\_hot\_encoded = encoder.fit\_transform(df[one\_hot\_columns])

#Create a new data frame from encoded values

one\_hot\_df = pd.DataFrame(one\_hot\_encoded, columns=encoder.get\_feature\_names\_out(one\_hot\_columns))

#Verify new data frame

print('#### One-Hot Encoded Data Frame ####')

print(one\_hot\_df)

#Concatenate the original data frame with the one-hot encoded data frame

df = pd.concat([df, one\_hot\_df], axis=1)

#Drop the original columns

df = df.drop(one\_hot\_columns, axis=1)

To follow the k-1 rule, I used the *drop* option in OneHotEncoder to drop the first column. This allowed me to reduce the number of variables in the model, without losing any of the data in the process.

This is a snapshot of the resulting dataframe:

Area\_Suburban Area\_Urban Marital\_Married Marital\_Never Married \

0 0.0 1.0 0.0 0.0

1 0.0 1.0 1.0 0.0

2 0.0 1.0 0.0 0.0

3 1.0 0.0 1.0 0.0

4 1.0 0.0 0.0 0.0

... ... ... ... ...

9995 0.0 0.0 1.0 0.0

9996 0.0 0.0 0.0 0.0

9997 0.0 0.0 0.0 1.0

9998 0.0 1.0 0.0 0.0

9999 0.0 1.0 0.0 1.0

Marital\_Separated Marital\_Widowed Gender\_Male Gender\_Nonbinary \

0 0.0 1.0 1.0 0.0

1 0.0 0.0 0.0 0.0

2 0.0 1.0 0.0 0.0

3 0.0 0.0 1.0 0.0

4 1.0 0.0 1.0 0.0

... ... ... ... ...

9995 0.0 0.0 1.0 0.0

9996 0.0 0.0 1.0 0.0

9997 0.0 0.0 0.0 0.0

9998 1.0 0.0 1.0 0.0

9999 0.0 0.0 1.0 0.0

InternetService\_Fiber Optic

0 1.0

1 1.0

2 0.0

3 0.0

4 1.0

... ...

9995 0.0

9996 1.0

9997 1.0

9998 1.0

9999 1.0

[10000 rows x 9 columns]

## C5 – Prepared Data

*Prepared data set included as churn\_clean\_final.csv*

# D – Initial vs. Reduced Models

## D1 – Initial Multiple Linear Regression

To perform my initial regression, I used the following code:

#Initial Regression

import pandas as pd

import numpy as np

import statsmodels.api as sm

df = pd.read\_csv('churn\_clean\_final.csv')

independent\_variables = ['Population', 'Area\_Suburban', 'Area\_Urban', 'Children', 'Age', 'Income', 'Marital\_Married', 'Marital\_Never Married',

'Marital\_Separated', 'Marital\_Widowed', 'Gender\_Male', 'Gender\_Nonbinary', 'Contacts', 'Yearly\_equip\_failure', 'Techie',

'Contract', 'Port\_modem', 'Tablet', 'InternetService\_Fiber Optic', 'Phone', 'Multiple', 'OnlineSecurity',

'OnlineBackup', 'StreamingTV', 'StreamingMovies', 'Tenure', ‘Email’]

#Define dependent variable

y = df['Bandwidth\_GB\_Year']

#Define independent variables

x = df[independent\_variables].assign(const = 1)

#Perform OLS regression

model = sm.OLS(y,x)

#Store regression results

model\_result = model.fit()

model\_result = model.fit()

print(model\_result.summary())

The resulting output was:

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 1.170e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:52:37 Log-Likelihood: -62278.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9972 BIC: 1.248e+05

Df Model: 27

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Population 7.742e-06 8.52e-05 0.091 0.928 -0.000 0.000

Area\_Suburban 0.2200 3.007 0.073 0.942 -5.675 6.115

Area\_Urban 2.9830 3.013 0.990 0.322 -2.923 8.889

Children 30.3100 0.573 52.930 0.000 29.187 31.432

Age -3.2207 0.059 -54.220 0.000 -3.337 -3.104

Income 5.995e-05 4.36e-05 1.376 0.169 -2.55e-05 0.000

Marital\_Married -2.0321 3.888 -0.523 0.601 -9.653 5.589

Marital\_Never Married -3.1597 3.865 -0.818 0.414 -10.735 4.416

Marital\_Separated -2.5039 3.835 -0.653 0.514 -10.021 5.013

Marital\_Widowed -7.9492 3.830 -2.076 0.038 -15.457 -0.442

Gender\_Male 65.4739 2.489 26.309 0.000 60.596 70.352

Gender\_Nonbinary -27.6246 8.271 -3.340 0.001 -43.838 -11.411

Contacts 2.2100 1.243 1.778 0.075 -0.227 4.647

Yearly\_equip\_failure 3.0497 1.933 1.578 0.115 -0.739 6.839

Techie 6.5274 3.288 1.985 0.047 0.081 12.973

Contract 1.1269 1.470 0.766 0.443 -1.755 4.009

Port\_modem -0.8264 2.458 -0.336 0.737 -5.645 3.993

Tablet -2.8245 2.686 -1.052 0.293 -8.089 2.440

InternetService\_Fiber Optic -325.7427 2.475 -131.606 0.000 -330.594 -320.891

Phone -6.2049 4.225 -1.468 0.142 -14.487 2.078

Multiple 77.1220 2.465 31.282 0.000 72.289 81.955

OnlineSecurity 79.9195 2.566 31.151 0.000 74.891 84.949

OnlineBackup 96.9996 2.471 39.258 0.000 92.156 101.843

StreamingTV 230.7226 2.457 93.885 0.000 225.905 235.540

StreamingMovies 211.2536 2.458 85.948 0.000 206.436 216.072

Tenure 81.9185 0.046 1762.150 0.000 81.827 82.010

Email -0.0017 0.406 -0.004 0.997 -0.798 0.795

const 500.2724 9.043 55.320 0.000 482.546 517.999

==============================================================================

Omnibus: 2900.550 Durbin-Watson: 1.970

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7192.825

Skew: -1.608 Prob(JB): 0.00

Kurtosis: 5.629 Cond. No. 3.85e+05

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 3.85e+05. This might indicate that there are

strong multicollinearity or other numerical problems.

Residual Standard Error: 122.59269611296408

## D2 – Feature Selection

My approach to feature selection was to perform VIF for multicollinearity and backwards stepwise elimination with an alpha of 0.05. To make the code less repetitive, I wrote a loop to perform the analysis and make a report at each step of the elimination process.

To limit the impact of multicollinearity on my model, I performed VIF on the defined independent variables using the following code:

#Separate independent variables

independent\_variables = ['Population', 'Area\_Suburban', 'Area\_Urban', 'Children', 'Age', 'Income', 'Marital\_Married', 'Marital\_Never Married',

'Marital\_Separated', 'Marital\_Widowed', 'Gender\_Male', 'Gender\_Nonbinary', 'Email', 'Contacts', 'Yearly\_equip\_failure',

'Techie', 'Contract', 'Port\_modem', 'Tablet', 'InternetService\_Fiber Optic', 'Phone', 'Multiple', 'OnlineSecurity',

'OnlineBackup', 'StreamingTV', 'StreamingMovies', 'Tenure']

#Create a new data frame

vif\_data = pd.DataFrame()

#Add a 'feature' column with column values from df[independent\_variables]

vif\_data['feature'] = df[independent\_variables].columns

#Perform VIF for each column in the set

vif\_data['VIF'] = [variance\_inflation\_factor(df[independent\_variables].values, i) for i in range(len(df[independent\_variables].columns))]

#Print the resulting data frame

print(vif\_data)

The following data frame was the result:

feature VIF

0 Population 1.451374

1 Area\_Suburban 1.948945

2 Area\_Urban 1.947030

3 Children 1.912114

4 Age 6.709893

5 Income 2.871709

6 Marital\_Married 1.843415

7 Marital\_Never Married 1.857115

8 Marital\_Separated 1.880700

9 Marital\_Widowed 1.888451

10 Gender\_Male 1.916043

11 Gender\_Nonbinary 1.046518

12 Email 11.800665

13 Contacts 1.982403

14 Yearly\_equip\_failure 1.383085

15 Techie 1.197232

16 Contract 1.685945

17 Port\_modem 1.910048

18 Tablet 1.422338

19 InternetService\_Fiber Optic 2.226076

20 Phone 8.805998

21 Multiple 1.825112

22 OnlineSecurity 1.542042

23 OnlineBackup 1.798065

24 StreamingTV 1.939186

25 StreamingMovies 1.927313

26 Tenure 2.626055

With a threshold of 10, Email ended up being the only variable that needed culling for the sake of multicollinearity. Phone was close, but given the defined standards it should be acceptable to keep it in the model. I performed a second round of VIF calculations without Email included and got the following results:

feature VIF

0 Population 1.443818

1 Area\_Suburban 1.928521

2 Area\_Urban 1.924815

3 Children 1.893368

4 Age 6.326383

5 Income 2.825946

6 Marital\_Married 1.804786

7 Marital\_Never Married 1.818512

8 Marital\_Separated 1.839740

9 Marital\_Widowed 1.849458

10 Gender\_Male 1.896519

11 Gender\_Nonbinary 1.046019

12 Contacts 1.967013

13 Yearly\_equip\_failure 1.380763

14 Techie 1.195813

15 Contract 1.679311

16 Port\_modem 1.892423

17 Tablet 1.419813

18 InternetService\_Fiber Optic 2.202741

19 Phone 8.035883

20 Multiple 1.812105

21 OnlineSecurity 1.539187

22 OnlineBackup 1.790157

23 StreamingTV 1.921621

24 StreamingMovies 1.912981

25 Tenure 2.599486

Phone and Age were still high, but nothing rose above the threshold after the second round of calculations.

The following is my code for backwards stepwise analysis:

#Backwards stepwise elimination

#Establish significance level

significance\_level = 0.05

#Loop until model reaches exepected significance for all values

while True:

#Get current/updated list of independent variables

x = df[independent\_variables].assign(const = 1)

#Perform OLS regression

model = sm.OLS(y, x)

#Store regression results

model\_result = model.fit()

#Get the p-values

p\_values = model\_result.pvalues.drop('const')

#Find the highest p-value

max\_p\_value = p\_values.max()

#Get the variable with the highest p-value

highest\_variable = p\_values.idxmax()

#If the variable with the highest p-value is above significance, remove it, otherwise end elimination

if max\_p\_value > significance\_level:

independent\_variables.remove(highest\_variable)

print(f"I have decided to remove '{highest\_variable}' as it has a p-value {max\_p\_value}")

print(model\_result.summary())

else:

break

Below are the reports for each feature that was removed during the process:

I have decided to remove 'Area\_Suburban' as it has a p-value 0.9416484308612554

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 1.215e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62278.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9973 BIC: 1.248e+05

Df Model: 26

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Population 7.736e-06 8.51e-05 0.091 0.928 -0.000 0.000

Area\_Suburban 0.2201 3.007 0.073 0.942 -5.674 6.114

Area\_Urban 2.9831 3.013 0.990 0.322 -2.923 8.889

Children 30.3099 0.573 52.933 0.000 29.188 31.432

Age -3.2207 0.059 -54.223 0.000 -3.337 -3.104

Income 5.996e-05 4.36e-05 1.376 0.169 -2.55e-05 0.000

Marital\_Married -2.0323 3.888 -0.523 0.601 -9.653 5.588

Marital\_Never Married -3.1598 3.864 -0.818 0.414 -10.735 4.415

Marital\_Separated -2.5041 3.834 -0.653 0.514 -10.020 5.012

Marital\_Widowed -7.9493 3.830 -2.076 0.038 -15.456 -0.442

Gender\_Male 65.4738 2.488 26.313 0.000 60.596 70.351

Gender\_Nonbinary -27.6243 8.271 -3.340 0.001 -43.837 -11.412

Contacts 2.2100 1.243 1.778 0.075 -0.227 4.647

Yearly\_equip\_failure 3.0498 1.933 1.578 0.115 -0.738 6.838

Techie 6.5276 3.288 1.985 0.047 0.083 12.973

Contract 1.1270 1.470 0.766 0.443 -1.755 4.009

Port\_modem -0.8265 2.458 -0.336 0.737 -5.645 3.992

Tablet -2.8244 2.685 -1.052 0.293 -8.088 2.439

InternetService\_Fiber Optic -325.7427 2.475 -131.613 0.000 -330.594 -320.891

Phone -6.2047 4.225 -1.469 0.142 -14.487 2.077

Multiple 77.1221 2.465 31.283 0.000 72.290 81.954

OnlineSecurity 79.9198 2.565 31.163 0.000 74.893 84.947

OnlineBackup 96.9998 2.471 39.263 0.000 92.157 101.842

StreamingTV 230.7226 2.457 93.891 0.000 225.906 235.539

StreamingMovies 211.2536 2.458 85.952 0.000 206.436 216.071

Tenure 81.9185 0.046 1762.417 0.000 81.827 82.010

const 500.2515 7.575 66.038 0.000 485.403 515.100

==============================================================================

Omnibus: 2900.550 Durbin-Watson: 1.970

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7192.830

Skew: -1.608 Prob(JB): 0.00

Kurtosis: 5.629 Cond. No. 3.43e+05

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 3.43e+05. This might indicate that there are

strong multicollinearity or other numerical problems.

I have decided to remove 'Population' as it has a p-value 0.9271845316813481

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 1.264e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62278.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9974 BIC: 1.248e+05

Df Model: 25

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Population 7.78e-06 8.51e-05 0.091 0.927 -0.000 0.000

Area\_Urban 2.8727 2.608 1.101 0.271 -2.240 7.985

Children 30.3097 0.573 52.936 0.000 29.187 31.432

Age -3.2206 0.059 -54.228 0.000 -3.337 -3.104

Income 5.997e-05 4.36e-05 1.376 0.169 -2.54e-05 0.000

Marital\_Married -2.0315 3.887 -0.523 0.601 -9.651 5.588

Marital\_Never Married -3.1585 3.864 -0.817 0.414 -10.733 4.416

Marital\_Separated -2.5039 3.834 -0.653 0.514 -10.020 5.012

Marital\_Widowed -7.9459 3.829 -2.075 0.038 -15.452 -0.440

Gender\_Male 65.4728 2.488 26.314 0.000 60.595 70.350

Gender\_Nonbinary -27.6235 8.270 -3.340 0.001 -43.835 -11.412

Contacts 2.2098 1.243 1.778 0.075 -0.227 4.646

Yearly\_equip\_failure 3.0501 1.932 1.578 0.115 -0.738 6.838

Techie 6.5257 3.288 1.985 0.047 0.081 12.970

Contract 1.1271 1.470 0.767 0.443 -1.755 4.009

Port\_modem -0.8270 2.458 -0.336 0.737 -5.645 3.991

Tablet -2.8232 2.685 -1.051 0.293 -8.087 2.440

InternetService\_Fiber Optic -325.7442 2.475 -131.624 0.000 -330.595 -320.893

Phone -6.2059 4.225 -1.469 0.142 -14.487 2.076

Multiple 77.1223 2.465 31.285 0.000 72.290 81.954

OnlineSecurity 79.9213 2.564 31.166 0.000 74.895 84.948

OnlineBackup 96.9985 2.470 39.265 0.000 92.156 101.841

StreamingTV 230.7224 2.457 93.896 0.000 225.906 235.539

StreamingMovies 211.2546 2.458 85.958 0.000 206.437 216.072

Tenure 81.9184 0.046 1762.637 0.000 81.827 82.010

const 500.3616 7.424 67.396 0.000 485.809 514.914

==============================================================================

Omnibus: 2900.504 Durbin-Watson: 1.970

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7192.573

Skew: -1.608 Prob(JB): 0.00

Kurtosis: 5.629 Cond. No. 3.40e+05

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 3.4e+05. This might indicate that there are

strong multicollinearity or other numerical problems.

I have decided to remove 'Port\_modem' as it has a p-value 0.7370882888788902

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 1.316e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62278.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9975 BIC: 1.248e+05

Df Model: 24

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Area\_Urban 2.8702 2.608 1.101 0.271 -2.241 7.982

Children 30.3093 0.573 52.939 0.000 29.187 31.432

Age -3.2206 0.059 -54.232 0.000 -3.337 -3.104

Income 5.993e-05 4.36e-05 1.376 0.169 -2.55e-05 0.000

Marital\_Married -2.0311 3.887 -0.523 0.601 -9.651 5.588

Marital\_Never Married -3.1542 3.864 -0.816 0.414 -10.728 4.419

Marital\_Separated -2.5059 3.834 -0.654 0.513 -10.021 5.009

Marital\_Widowed -7.9467 3.829 -2.075 0.038 -15.452 -0.441

Gender\_Male 65.4707 2.488 26.315 0.000 60.594 70.348

Gender\_Nonbinary -27.6295 8.270 -3.341 0.001 -43.840 -11.419

Contacts 2.2102 1.243 1.778 0.075 -0.226 4.647

Yearly\_equip\_failure 3.0493 1.932 1.578 0.115 -0.738 6.837

Techie 6.5224 3.287 1.984 0.047 0.079 12.966

Contract 1.1290 1.470 0.768 0.443 -1.753 4.011

Port\_modem -0.8251 2.458 -0.336 0.737 -5.643 3.992

Tablet -2.8228 2.685 -1.051 0.293 -8.086 2.440

InternetService\_Fiber Optic -325.7455 2.475 -131.634 0.000 -330.596 -320.895

Phone -6.2028 4.224 -1.468 0.142 -14.484 2.078

Multiple 77.1219 2.465 31.286 0.000 72.290 81.954

OnlineSecurity 79.9241 2.564 31.171 0.000 74.898 84.950

OnlineBackup 97.0009 2.470 39.270 0.000 92.159 101.843

StreamingTV 230.7209 2.457 93.902 0.000 225.905 235.537

StreamingMovies 211.2533 2.457 85.964 0.000 206.436 216.070

Tenure 81.9184 0.046 1762.745 0.000 81.827 82.009

const 500.4344 7.381 67.801 0.000 485.966 514.902

==============================================================================

Omnibus: 2900.512 Durbin-Watson: 1.970

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7192.665

Skew: -1.608 Prob(JB): 0.00

Kurtosis: 5.629 Cond. No. 3.35e+05

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 3.35e+05. This might indicate that there are

strong multicollinearity or other numerical problems.

I have decided to remove 'Marital\_Married' as it has a p-value 0.6019100883469823

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 1.374e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62278.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9976 BIC: 1.248e+05

Df Model: 23

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Area\_Urban 2.8778 2.608 1.104 0.270 -2.233 7.989

Children 30.3072 0.572 52.941 0.000 29.185 31.429

Age -3.2207 0.059 -54.238 0.000 -3.337 -3.104

Income 6.015e-05 4.36e-05 1.381 0.167 -2.52e-05 0.000

Marital\_Married -2.0277 3.887 -0.522 0.602 -9.647 5.591

Marital\_Never Married -3.1453 3.863 -0.814 0.416 -10.718 4.428

Marital\_Separated -2.5038 3.834 -0.653 0.514 -10.019 5.011

Marital\_Widowed -7.9382 3.829 -2.073 0.038 -15.443 -0.433

Gender\_Male 65.4595 2.488 26.314 0.000 60.583 70.336

Gender\_Nonbinary -27.6328 8.269 -3.342 0.001 -43.842 -11.423

Contacts 2.2110 1.243 1.779 0.075 -0.225 4.647

Yearly\_equip\_failure 3.0459 1.932 1.576 0.115 -0.742 6.833

Techie 6.5375 3.287 1.989 0.047 0.095 12.980

Contract 1.1312 1.470 0.770 0.442 -1.750 4.013

Tablet -2.8238 2.685 -1.052 0.293 -8.087 2.439

InternetService\_Fiber Optic -325.7497 2.474 -131.643 0.000 -330.600 -320.899

Phone -6.2053 4.224 -1.469 0.142 -14.486 2.075

Multiple 77.1241 2.465 31.289 0.000 72.292 81.956

OnlineSecurity 79.9219 2.564 31.172 0.000 74.896 84.948

OnlineBackup 97.0026 2.470 39.273 0.000 92.161 101.844

StreamingTV 230.7213 2.457 93.906 0.000 225.905 235.537

StreamingMovies 211.2490 2.457 85.967 0.000 206.432 216.066

Tenure 81.9183 0.046 1762.888 0.000 81.827 82.009

const 500.0432 7.288 68.612 0.000 485.757 514.329

==============================================================================

Omnibus: 2899.996 Durbin-Watson: 1.970

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7190.176

Skew: -1.608 Prob(JB): 0.00

Kurtosis: 5.629 Cond. No. 3.34e+05

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 3.34e+05. This might indicate that there are

strong multicollinearity or other numerical problems.

I have decided to remove 'Marital\_Separated' as it has a p-value 0.6470427156874505

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 1.436e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62278.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9977 BIC: 1.248e+05

Df Model: 22

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Area\_Urban 2.8502 2.607 1.093 0.274 -2.260 7.960

Children 30.3057 0.572 52.941 0.000 29.184 31.428

Age -3.2204 0.059 -54.237 0.000 -3.337 -3.104

Income 6e-05 4.36e-05 1.378 0.168 -2.54e-05 0.000

Marital\_Never Married -2.1779 3.389 -0.643 0.520 -8.821 4.466

Marital\_Separated -1.5363 3.355 -0.458 0.647 -8.113 5.040

Marital\_Widowed -6.9694 3.348 -2.082 0.037 -13.533 -0.406

Gender\_Male 65.4692 2.487 26.320 0.000 60.593 70.345

Gender\_Nonbinary -27.6266 8.269 -3.341 0.001 -43.835 -11.418

Contacts 2.2108 1.243 1.779 0.075 -0.225 4.647

Yearly\_equip\_failure 3.0489 1.932 1.578 0.115 -0.738 6.836

Techie 6.5236 3.287 1.985 0.047 0.081 12.966

Contract 1.1332 1.470 0.771 0.441 -1.748 4.015

Tablet -2.8227 2.685 -1.051 0.293 -8.085 2.440

InternetService\_Fiber Optic -325.7537 2.474 -131.650 0.000 -330.604 -320.903

Phone -6.2317 4.224 -1.475 0.140 -14.511 2.048

Multiple 77.1338 2.465 31.295 0.000 72.302 81.965

OnlineSecurity 79.9446 2.563 31.186 0.000 74.920 84.970

OnlineBackup 96.9884 2.470 39.271 0.000 92.147 101.830

StreamingTV 230.7306 2.457 93.916 0.000 225.915 235.546

StreamingMovies 211.2584 2.457 85.976 0.000 206.442 216.075

Tenure 81.9181 0.046 1763.015 0.000 81.827 82.009

const 499.0884 7.054 70.751 0.000 485.261 512.916

==============================================================================

Omnibus: 2900.939 Durbin-Watson: 1.970

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7194.777

Skew: -1.609 Prob(JB): 0.00

Kurtosis: 5.630 Cond. No. 3.31e+05

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 3.31e+05. This might indicate that there are

strong multicollinearity or other numerical problems.

I have decided to remove 'Marital\_Never Married' as it has a p-value 0.6029103855466915

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 1.505e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62278.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9978 BIC: 1.248e+05

Df Model: 21

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Area\_Urban 2.8524 2.607 1.094 0.274 -2.257 7.962

Children 30.3059 0.572 52.944 0.000 29.184 31.428

Age -3.2202 0.059 -54.237 0.000 -3.337 -3.104

Income 6.022e-05 4.36e-05 1.383 0.167 -2.52e-05 0.000

Marital\_Never Married -1.6634 3.197 -0.520 0.603 -7.931 4.604

Marital\_Widowed -6.4556 3.154 -2.046 0.041 -12.639 -0.272

Gender\_Male 65.4750 2.487 26.324 0.000 60.599 70.351

Gender\_Nonbinary -27.6284 8.269 -3.341 0.001 -43.836 -11.420

Contacts 2.2163 1.243 1.783 0.075 -0.220 4.652

Yearly\_equip\_failure 3.0606 1.932 1.584 0.113 -0.726 6.847

Techie 6.5225 3.286 1.985 0.047 0.080 12.965

Contract 1.1232 1.470 0.764 0.445 -1.758 4.004

Tablet -2.8416 2.684 -1.059 0.290 -8.103 2.420

InternetService\_Fiber Optic -325.7655 2.474 -131.667 0.000 -330.615 -320.916

Phone -6.2135 4.223 -1.471 0.141 -14.492 2.065

Multiple 77.1320 2.465 31.295 0.000 72.301 81.963

OnlineSecurity 79.9403 2.563 31.186 0.000 74.916 84.965

OnlineBackup 96.9906 2.470 39.273 0.000 92.150 101.832

StreamingTV 230.7200 2.457 93.920 0.000 225.905 235.535

StreamingMovies 211.2568 2.457 85.979 0.000 206.440 216.073

Tenure 81.9180 0.046 1763.088 0.000 81.827 82.009

const 498.5543 6.957 71.665 0.000 484.918 512.191

==============================================================================

Omnibus: 2901.438 Durbin-Watson: 1.970

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7196.915

Skew: -1.609 Prob(JB): 0.00

Kurtosis: 5.630 Cond. No. 3.31e+05

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 3.31e+05. This might indicate that there are

strong multicollinearity or other numerical problems.

I have decided to remove 'Contract' as it has a p-value 0.44861313438602957

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 1.580e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62279.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9979 BIC: 1.248e+05

Df Model: 20

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Area\_Urban 2.8405 2.607 1.090 0.276 -2.269 7.950

Children 30.3067 0.572 52.947 0.000 29.185 31.429

Age -3.2203 0.059 -54.243 0.000 -3.337 -3.104

Income 6.013e-05 4.36e-05 1.381 0.167 -2.52e-05 0.000

Marital\_Widowed -6.0472 3.055 -1.979 0.048 -12.036 -0.059

Gender\_Male 65.4937 2.487 26.335 0.000 60.619 70.369

Gender\_Nonbinary -27.6337 8.268 -3.342 0.001 -43.841 -11.426

Contacts 2.2112 1.243 1.779 0.075 -0.225 4.647

Yearly\_equip\_failure 3.0531 1.932 1.580 0.114 -0.734 6.840

Techie 6.5502 3.286 1.993 0.046 0.109 12.991

Contract 1.1135 1.470 0.758 0.449 -1.767 3.994

Tablet -2.8366 2.684 -1.057 0.291 -8.098 2.425

InternetService\_Fiber Optic -325.7729 2.474 -131.677 0.000 -330.622 -320.923

Phone -6.1677 4.222 -1.461 0.144 -14.444 2.109

Multiple 77.1146 2.464 31.292 0.000 72.284 81.945

OnlineSecurity 79.9246 2.563 31.183 0.000 74.900 84.949

OnlineBackup 96.9873 2.470 39.274 0.000 92.147 101.828

StreamingTV 230.7234 2.456 93.925 0.000 225.908 235.539

StreamingMovies 211.2713 2.457 85.994 0.000 206.455 216.087

Tenure 81.9179 0.046 1763.156 0.000 81.827 82.009

const 498.1307 6.909 72.102 0.000 484.588 511.673

==============================================================================

Omnibus: 2901.786 Durbin-Watson: 1.970

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7199.120

Skew: -1.609 Prob(JB): 0.00

Kurtosis: 5.631 Cond. No. 3.31e+05

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 3.31e+05. This might indicate that there are

strong multicollinearity or other numerical problems.

I have decided to remove 'Tablet' as it has a p-value 0.29255608466583954

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 1.663e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62279.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9980 BIC: 1.247e+05

Df Model: 19

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Area\_Urban 2.8285 2.606 1.085 0.278 -2.281 7.938

Children 30.3171 0.572 52.982 0.000 29.195 31.439

Age -3.2208 0.059 -54.253 0.000 -3.337 -3.104

Income 6.004e-05 4.35e-05 1.379 0.168 -2.53e-05 0.000

Marital\_Widowed -6.0369 3.055 -1.976 0.048 -12.025 -0.048

Gender\_Male 65.5099 2.487 26.343 0.000 60.635 70.385

Gender\_Nonbinary -27.6280 8.268 -3.342 0.001 -43.835 -11.421

Contacts 2.2183 1.243 1.785 0.074 -0.217 4.654

Yearly\_equip\_failure 3.0501 1.932 1.579 0.114 -0.736 6.837

Techie 6.5419 3.286 1.991 0.047 0.101 12.983

Tablet -2.8253 2.684 -1.053 0.293 -8.087 2.436

InternetService\_Fiber Optic -325.7718 2.474 -131.679 0.000 -330.621 -320.922

Phone -6.1436 4.222 -1.455 0.146 -14.420 2.133

Multiple 77.1482 2.464 31.312 0.000 72.318 81.978

OnlineSecurity 79.9578 2.563 31.201 0.000 74.934 84.981

OnlineBackup 96.9769 2.469 39.271 0.000 92.136 101.818

StreamingTV 230.7467 2.456 93.944 0.000 225.932 235.561

StreamingMovies 211.2555 2.457 85.992 0.000 206.440 216.071

Tenure 81.9186 0.046 1763.478 0.000 81.828 82.010

const 498.8292 6.847 72.856 0.000 485.408 512.250

==============================================================================

Omnibus: 2902.760 Durbin-Watson: 1.970

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7203.810

Skew: -1.609 Prob(JB): 0.00

Kurtosis: 5.632 Cond. No. 3.30e+05

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 3.3e+05. This might indicate that there are

strong multicollinearity or other numerical problems.

I have decided to remove 'Area\_Urban' as it has a p-value 0.2824871678101414

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 1.756e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62279.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9981 BIC: 1.247e+05

Df Model: 18

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Area\_Urban 2.8013 2.606 1.075 0.282 -2.308 7.910

Children 30.3166 0.572 52.981 0.000 29.195 31.438

Age -3.2205 0.059 -54.249 0.000 -3.337 -3.104

Income 5.977e-05 4.35e-05 1.373 0.170 -2.56e-05 0.000

Marital\_Widowed -5.9797 3.055 -1.958 0.050 -11.967 0.008

Gender\_Male 65.5357 2.487 26.355 0.000 60.661 70.410

Gender\_Nonbinary -27.6288 8.268 -3.342 0.001 -43.836 -11.422

Contacts 2.2347 1.242 1.799 0.072 -0.201 4.670

Yearly\_equip\_failure 3.0411 1.932 1.574 0.115 -0.745 6.828

Techie 6.5022 3.286 1.979 0.048 0.062 12.943

InternetService\_Fiber Optic -325.7563 2.474 -131.675 0.000 -330.606 -320.907

Phone -6.2568 4.221 -1.482 0.138 -14.530 2.017

Multiple 77.2011 2.463 31.340 0.000 72.372 82.030

OnlineSecurity 79.9328 2.563 31.193 0.000 74.910 84.956

OnlineBackup 96.9853 2.469 39.274 0.000 92.145 101.826

StreamingTV 230.6974 2.456 93.940 0.000 225.884 235.511

StreamingMovies 211.2008 2.456 85.989 0.000 206.386 216.015

Tenure 81.9185 0.046 1763.468 0.000 81.827 82.010

const 498.0861 6.810 73.137 0.000 484.736 511.436

==============================================================================

Omnibus: 2902.225 Durbin-Watson: 1.970

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7200.590

Skew: -1.609 Prob(JB): 0.00

Kurtosis: 5.631 Cond. No. 3.30e+05

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 3.3e+05. This might indicate that there are

strong multicollinearity or other numerical problems.

I have decided to remove 'Income' as it has a p-value 0.16946678292377645

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 1.859e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62280.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9982 BIC: 1.247e+05

Df Model: 17

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Children 30.3138 0.572 52.976 0.000 29.192 31.435

Age -3.2200 0.059 -54.242 0.000 -3.336 -3.104

Income 5.984e-05 4.35e-05 1.374 0.169 -2.55e-05 0.000

Marital\_Widowed -5.9542 3.055 -1.949 0.051 -11.942 0.033

Gender\_Male 65.5662 2.487 26.368 0.000 60.692 70.440

Gender\_Nonbinary -27.6771 8.268 -3.347 0.001 -43.884 -11.470

Contacts 2.2371 1.243 1.800 0.072 -0.198 4.673

Yearly\_equip\_failure 3.0231 1.932 1.565 0.118 -0.763 6.810

Techie 6.5110 3.286 1.982 0.048 0.070 12.952

InternetService\_Fiber Optic -325.7412 2.474 -131.670 0.000 -330.591 -320.892

Phone -6.3190 4.220 -1.497 0.134 -14.592 1.954

Multiple 77.2134 2.463 31.345 0.000 72.385 82.042

OnlineSecurity 79.9017 2.562 31.182 0.000 74.879 84.925

OnlineBackup 97.0334 2.469 39.300 0.000 92.194 101.873

StreamingTV 230.6745 2.456 93.934 0.000 225.861 235.488

StreamingMovies 211.1957 2.456 85.986 0.000 206.381 216.010

Tenure 81.9179 0.046 1763.576 0.000 81.827 82.009

const 499.0473 6.751 73.918 0.000 485.813 512.281

==============================================================================

Omnibus: 2902.715 Durbin-Watson: 1.970

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7203.189

Skew: -1.609 Prob(JB): 0.00

Kurtosis: 5.632 Cond. No. 3.30e+05

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 3.3e+05. This might indicate that there are

strong multicollinearity or other numerical problems.

I have decided to remove 'Phone' as it has a p-value 0.13379977039667143

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 1.975e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62281.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9983 BIC: 1.247e+05

Df Model: 16

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Children 30.3216 0.572 52.990 0.000 29.200 31.443

Age -3.2203 0.059 -54.245 0.000 -3.337 -3.104

Marital\_Widowed -5.9573 3.055 -1.950 0.051 -11.945 0.030

Gender\_Male 65.4869 2.486 26.342 0.000 60.614 70.360

Gender\_Nonbinary -27.6445 8.268 -3.343 0.001 -43.852 -11.437

Contacts 2.2401 1.243 1.803 0.071 -0.196 4.676

Yearly\_equip\_failure 3.0370 1.932 1.572 0.116 -0.749 6.824

Techie 6.5348 3.286 1.989 0.047 0.094 12.976

InternetService\_Fiber Optic -325.8032 2.474 -131.711 0.000 -330.652 -320.954

Phone -6.3284 4.221 -1.499 0.134 -14.602 1.945

Multiple 77.2136 2.463 31.343 0.000 72.385 82.043

OnlineSecurity 79.8675 2.562 31.169 0.000 74.845 84.890

OnlineBackup 97.0128 2.469 39.291 0.000 92.173 101.853

StreamingTV 230.6646 2.456 93.926 0.000 225.851 235.478

StreamingMovies 211.1896 2.456 85.980 0.000 206.375 216.004

Tenure 81.9181 0.046 1763.504 0.000 81.827 82.009

const 501.5218 6.507 77.074 0.000 488.767 514.277

==============================================================================

Omnibus: 2902.555 Durbin-Watson: 1.970

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7201.960

Skew: -1.609 Prob(JB): 0.00

Kurtosis: 5.631 Cond. No. 456.

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

I have decided to remove 'Yearly\_equip\_failure' as it has a p-value 0.11314506913198433

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 2.106e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62282.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9984 BIC: 1.247e+05

Df Model: 15

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Children 30.3226 0.572 52.988 0.000 29.201 31.444

Age -3.2211 0.059 -54.258 0.000 -3.337 -3.105

Marital\_Widowed -5.9471 3.055 -1.947 0.052 -11.935 0.041

Gender\_Male 65.4774 2.486 26.337 0.000 60.604 70.351

Gender\_Nonbinary -27.6998 8.269 -3.350 0.001 -43.908 -11.491

Contacts 2.2302 1.243 1.795 0.073 -0.206 4.666

Yearly\_equip\_failure 3.0606 1.932 1.584 0.113 -0.726 6.847

Techie 6.5490 3.286 1.993 0.046 0.108 12.990

InternetService\_Fiber Optic -325.8016 2.474 -131.702 0.000 -330.651 -320.952

Multiple 77.2004 2.464 31.336 0.000 72.371 82.030

OnlineSecurity 79.8631 2.563 31.165 0.000 74.840 84.886

OnlineBackup 97.0261 2.469 39.294 0.000 92.186 101.866

StreamingTV 230.7109 2.456 93.946 0.000 225.897 235.525

StreamingMovies 211.2005 2.456 85.979 0.000 206.385 216.016

Tenure 81.9178 0.046 1763.398 0.000 81.827 82.009

const 495.8090 5.275 93.986 0.000 485.468 506.150

==============================================================================

Omnibus: 2903.121 Durbin-Watson: 1.969

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7207.851

Skew: -1.609 Prob(JB): 0.00

Kurtosis: 5.634 Cond. No. 455.

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

I have decided to remove 'Contacts' as it has a p-value 0.07419950974593911

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 2.257e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62283.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9985 BIC: 1.247e+05

Df Model: 14

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Children 30.3295 0.572 52.998 0.000 29.208 31.451

Age -3.2203 0.059 -54.242 0.000 -3.337 -3.104

Marital\_Widowed -5.8853 3.055 -1.927 0.054 -11.873 0.103

Gender\_Male 65.4972 2.486 26.343 0.000 60.624 70.371

Gender\_Nonbinary -27.3596 8.267 -3.310 0.001 -43.564 -11.155

Contacts 2.2189 1.243 1.786 0.074 -0.217 4.655

Techie 6.5052 3.286 1.980 0.048 0.064 12.947

InternetService\_Fiber Optic -325.7742 2.474 -131.684 0.000 -330.624 -320.925

Multiple 77.2099 2.464 31.338 0.000 72.380 82.039

OnlineSecurity 79.7967 2.562 31.141 0.000 74.774 84.820

OnlineBackup 96.9776 2.469 39.274 0.000 92.137 101.818

StreamingTV 230.7195 2.456 93.943 0.000 225.905 235.534

StreamingMovies 211.1848 2.457 85.967 0.000 206.369 216.000

Tenure 81.9188 0.046 1763.434 0.000 81.828 82.010

const 496.9529 5.226 95.091 0.000 486.709 507.197

==============================================================================

Omnibus: 2908.068 Durbin-Watson: 1.969

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7232.507

Skew: -1.611 Prob(JB): 0.00

Kurtosis: 5.641 Cond. No. 455.

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

I have decided to remove 'Marital\_Widowed' as it has a p-value 0.055006810696007505

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 2.430e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62285.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9986 BIC: 1.247e+05

Df Model: 13

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Children 30.3083 0.572 52.966 0.000 29.187 31.430

Age -3.2188 0.059 -54.216 0.000 -3.335 -3.102

Marital\_Widowed -5.8629 3.055 -1.919 0.055 -11.852 0.126

Gender\_Male 65.5053 2.487 26.344 0.000 60.631 70.380

Gender\_Nonbinary -27.3392 8.268 -3.307 0.001 -43.545 -11.133

Techie 6.5366 3.286 1.989 0.047 0.095 12.979

InternetService\_Fiber Optic -325.7669 2.474 -131.667 0.000 -330.617 -320.917

Multiple 77.1377 2.464 31.309 0.000 72.308 81.967

OnlineSecurity 79.8471 2.563 31.159 0.000 74.824 84.870

OnlineBackup 97.0218 2.469 39.290 0.000 92.181 101.862

StreamingTV 230.7218 2.456 93.934 0.000 225.907 235.536

StreamingMovies 211.2329 2.457 85.983 0.000 206.417 216.049

Tenure 81.9190 0.046 1763.249 0.000 81.828 82.010

const 499.0684 5.091 98.038 0.000 489.090 509.047

==============================================================================

Omnibus: 2911.384 Durbin-Watson: 1.970

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7246.619

Skew: -1.613 Prob(JB): 0.00

Kurtosis: 5.644 Cond. No. 455.

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

I have decided to remove 'Techie' as it has a p-value 0.05146367520619489

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 2.631e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62287.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9987 BIC: 1.247e+05

Df Model: 12

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Children 30.3087 0.572 52.960 0.000 29.187 31.431

Age -3.2182 0.059 -54.200 0.000 -3.335 -3.102

Gender\_Male 65.5374 2.487 26.354 0.000 60.663 70.412

Gender\_Nonbinary -27.4537 8.268 -3.320 0.001 -43.662 -11.246

Techie 6.4007 3.286 1.948 0.051 -0.041 12.842

InternetService\_Fiber Optic -325.7713 2.475 -131.651 0.000 -330.622 -320.921

Multiple 77.1702 2.464 31.319 0.000 72.340 82.000

OnlineSecurity 79.8472 2.563 31.155 0.000 74.823 84.871

OnlineBackup 96.9744 2.470 39.267 0.000 92.134 101.815

StreamingTV 230.7082 2.457 93.916 0.000 225.893 235.524

StreamingMovies 211.2905 2.457 86.001 0.000 206.475 216.106

Tenure 81.9190 0.046 1763.014 0.000 81.828 82.010

const 497.8445 5.051 98.561 0.000 487.943 507.746

==============================================================================

Omnibus: 2915.815 Durbin-Watson: 1.970

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7268.556

Skew: -1.614 Prob(JB): 0.00

Kurtosis: 5.649 Cond. No. 455.

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS Regression Results

==============================================================================

Dep. Variable: Bandwidth\_GB\_Year R-squared: 0.997

Model: OLS Adj. R-squared: 0.997

Method: Least Squares F-statistic: 2.870e+05

Date: Wed, 06 Nov 2024 Prob (F-statistic): 0.00

Time: 07:53:10 Log-Likelihood: -62289.

No. Observations: 10000 AIC: 1.246e+05

Df Residuals: 9988 BIC: 1.247e+05

Df Model: 11

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

Children 30.3009 0.572 52.940 0.000 29.179 31.423

Age -3.2191 0.059 -54.209 0.000 -3.336 -3.103

Gender\_Male 65.4824 2.487 26.330 0.000 60.607 70.357

Gender\_Nonbinary -27.6530 8.269 -3.344 0.001 -43.862 -11.444

InternetService\_Fiber Optic -325.7578 2.475 -131.628 0.000 -330.609 -320.907

Multiple 77.1463 2.464 31.305 0.000 72.316 81.977

OnlineSecurity 79.7711 2.563 31.125 0.000 74.747 84.795

OnlineBackup 96.9995 2.470 39.272 0.000 92.158 101.841

StreamingTV 230.7259 2.457 93.911 0.000 225.910 235.542

StreamingMovies 211.2715 2.457 85.982 0.000 206.455 216.088

Tenure 81.9181 0.046 1762.841 0.000 81.827 82.009

const 499.0671 5.013 99.561 0.000 489.241 508.893

==============================================================================

Omnibus: 2914.651 Durbin-Watson: 1.969

Prob(Omnibus): 0.000 Jarque-Bera (JB): 7262.284

Skew: -1.614 Prob(JB): 0.00

Kurtosis: 5.647 Cond. No. 455.

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Residual Standard Error: 122.7218455907163

# E – Linear Regression Analysis

## E1 – Process

In the end the variables removed from the initial model were: ‘Area’, ‘Population’, ‘Port\_modem’, ‘Marital’, ‘Contract’, ‘Tablet’, ‘Income’, ‘Phone’, ‘Yearly\_equip\_failure’, ‘Contacts’, ‘Techie’, and ‘Email’. Most of these were removed for having p-values above the designated threshold of 0.05. ‘Email’ was removed due to having a VIF score above 10, indicating high probability of multicollinearity. The remaining model consists of ‘Children’, ‘Age’, ‘Gender’, ‘InternetService’, ‘Multiple’, ‘OnlineSecurity’, ‘OnlineBackup’, ‘StreamingTV’, ‘StreamingMovies’, and ‘Tenure’. Given the broad range of starting factors, this is a considerable reduction.

Most model comparison metrics stayed roughly consistent between the initial and reduced model. The residual standard error was very slightly higher in the reduced model, indicating a slightly less appropriate fit for the data. However, this is offset by the principle benefit of the reduced model: the reduction in the number of factors. This reduction narrows the focus of the analysis and helps to come to a decision on action by cutting the model down to the fewest number of components that support the model.

Heteroscedasticity was not found in the below generated residual plots.

## E2 – Output

Below are residual plots printed for each stage of the model analysis using the following code:

#Visualize residual plots

for variable in independent\_variables:

sm.graphics.plot\_regress\_exog(model\_result, variable, fig=plt.figure(figsize=[12,12]))

plt.show()

Below are the plots generated:

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## No description has been provided for this image

## E3 – Code

*Code is included in D208-Task1.ipynb*

# F – Data Summary and Implications

## F1 – Results

The equation for the reduced model is as follows:

y = 499.0671 + 30.3009Children – 3.2191Age + 65.4824Gender\_Male – 27.6530Gender\_Nonbinary – 325.7578InternetService\_Fiber Optic + 77.1463Multiple + 79.7711OnlineSecurity + 96.9995OnlineBackup + 230.7259StreamingTV + 211.2715StreamingMovies + 81.9181Tenure

The following interpretation of the coefficients determines:

* Keeping all things constant, one unit increase in Children is associated with a 30.3009 increase in bandwidth usage
* Keeping all things constant, one unit increase in Age is associated with a 3.2191 decrease in bandwidth usage
* Keeping all things constant, one unit increase in Gender\_Male is associated with a 65.4824 increase in bandwidth usage
* Keeping all things constant, one unit increase in Gender\_Nonbinary is associated with a 27.6530 decrease in bandwidth usage
* Keeping all things constant, one unit increase in InternetService\_Fiber Optic is associated with a 325.7578 increase in bandwidth usage
* Keeping all things constant, one unit increase in OnlineSecurity is associated with a 79.7711 increase in bandwidth usage
* Keeping all things constant, one unit increase in OnlineBackup is associated with a 96.9995 increase in bandwidth usage
* Keeping all things constant, one unit increase in StreamingTV is associated with a 230.7259 increase in bandwidth usage
* Keeping all things constant, one unit increase in StreamingMovies is associated with a 211.2715 increase in bandwidth usage
* Keeping all things constant, one unit increase in Multiple is associated with a 77.1463 increase in bandwidth usage
* Keeping all things constant, one unit increase in Tenure is associated with a 81.9181 increase in bandwidth usage

The model was found to be statistically significant with a Prob (F-statistic) of 0.00 and all features having a p value of 0.001 or lower.

In terms of practical significance, factors like age, the number of children, and gender are not likely to yield any actionable results. However, service types and levels were found to be significant and there are avenues for action with those, as pricing structures for these things are in control of the stakeholders.

The primary limitations to the analysis come to outliers and the reliance on observed correlation. There were outliers present in the data set. While many were deemed to be retainable, there is a non-zero chance the outliers could have had an impact on the results of the analysis. Taking action based solely on observed correlation is tricky because you cannot assume causation (x causes y) from correlation (x is related to y) alone. Lurking variables (factors outside of our examined data set that have an impact on our dependent variable) could potentially have an unmeasured collinearity that renders the analysis invalid. The only way to prove causation is to stage an experiment where all variables are tightly controlled.

What this analysis found was patterns in data that alluded to a relationship between the dependent variable and the independent variables. This analysis is in no way an assertion that any of the independent variables have a causal relationship with the dependent variable.

## F2 – Course of Action

There are a number of options available given the results. Tenure is the strongest factor found in the analysis and is a consistent predictor of bandwidth usage. While the obvious answer would be to increase data consumption rates based on tenure, squeezing longstanding customers is likely not a good approach, given the larger issue of customer retention. My recommendation would be to take a closer look at services like streaming, online security, and online backup. Examine the current rates and potential build a new structure for charges based on usage quotas. This would ensure that high consumption users of these services could offset some of the resource loss in their administration.

# G – Panopto Video

*The video is included in the project files*

# H – Code Sources

[Statsmodel Docs](https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html)

# I – Sources

[Heteroscedasticity](https://www.vexpower.com/brief/homoskedasticity)

[Feature Selection](https://towardsdatascience.com/feature-selection-techniques-in-regression-model-26878fe0e24e)

Datacamps

Dr. Middleton’s Getting Started with D208 Part I & II