D208 Exploratory Predictive Data Modeling PA

NBM2 Task 1: Multiple Regression for Predictive Modeling

Sean Simmons

WDU Data Analytics

MSDA D208

January 2023

Part I: Research Question

A. Describe the purpose of this data analysis by doing the following:

1. Summarize one research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using multiple regression.

2. Define the objectives or goals of the data analysis. Ensure that your objectives or goals are reasonable within the scope of the data dictionary and are represented in the available data.

1. "What causes customers to have a high Tenure?" is my research question. This is relevant to the real-world organizational setting because Tenure is defined as the length of time the customer has been with the organization. This is an important metric related to customer retention, a vital aspect to this organization's needs. Understanding why a customer stays with the organization for an extended period of time will help the organization identify what needs to be done, worked on, changed, or further analyzed to prevent the customer leaving. Tenure is our target variable and can be analyzed with multiple linear regression involving other variables in the dataset related to customer information.
2. The objective and goal of the data analysis here is to use multiple linear regression to identify the variables that have a strong relationship with Tenure. How long a customer will stay with the organization, or what variables affect it and can be used to increase the length of Tenure, is a key metric that can be used to inform the organization's decisions

and further data analyst projects. Multiple linear regression can be used to predict the relationship between Tenure and other independent variables in the data set.

Part II: Method Justification

B. Describe multiple regression methods by doing the following:

1. Summarize the assumptions of a multiple regression model.

2. Describe the benefits of using the tool(s) you have chosen (i.e., Python, R, or both) in support of various phases of the analysis.

3. Explain why multiple regression is an appropriate technique to analyze the research question summarized in Part I.

1. The assumptions of a multiple regression model are summarized as follows:
    a. There is a linear relationship between the dependent and independent variables.
    b. The independent variables are not strongly correlated with each other.
    c. The observations are independent from one another.
    d. The Residuals have a constant variance, that is to be normally distributed with a mean of 0.
2. The benefits of using Python and jupyter notebook are numerous. To begin, Jupyter notebook is an interface system for python that provides a user-friendly coding environment, visualization abilities, and code extraction. Python can handle large data sets faster and easier than R, and I am much more familiar with Python. Python can import a large data set, provide the summary, and house the data cleaning tools through packages and functions (such as missingno). Python and its packages can provide the visualizations for every step of this project: univariate, bivariate, and multiple linear

regression modeling as well as our backwards stepwise reduction and VIF analysis. Finally, Python will be able to use packages to perform the multiple linear regression model using the ols function, reduce the variables, and perform the residual plot visualization. To summarize, Python has all of the tools necessary to complete each step of this data analysis project in a quick and efficient manner (Prasanna 2021).

3. Multiple linear regression is an appropriate technique to analyze the research question because it can be used to identify a linear relationship between our continuous target variable and several other independent variables. The organization can benefit from multiple regression because it provides an indication as to which variables have a strong relationship with our target, Tenure, thereby informing their retention and best practices methods (Yadav 2021).

Part III: Data Preparation

C. Summarize the data preparation process for multiple regression analysis by doing the following:

1. Describe your data preparation goals and the data manipulations that will be used to achieve the goals.

2. Discuss the summary statistics, including the target variable and *all* predictor variables that you will need to gather from the data set to answer the research question.

3. Explain the steps used to prepare the data for the analysis, including the annotated code.

4. Generate univariate and bivariate visualizations of the distributions of variables in the cleaned data set. Include the target variable in your bivariate visualizations.

5. Provide a copy of the prepared data set.

1. My data preparation goals are to bring the data into my coding environment and check if it is cleaned. This involves looking for missing values and data types to examine the overall structure of the dataset. The data manipulations I will do are to rename survey response items to be more descriptive and create numeric columns for all of the categorical columns that were informed from the last step. Now we can check if it is cleaned, so I checked for missing values to impute with central tendency. Next, I will use univariate analysis to create visuals for each of the variables I intend to look at. After that, bivariate analysis using scatterplots of numerous independent variables with Tenure will be done.

   No missing values:

```
Out[42]: CaseOrder               0
         Customer_id             0
         Interaction             0
         UID                     0
         City                    0
         State                   0
         County                  0
         Zip                     0
         Lat                     0
         Lng                     0
         Population              0
         Area                    0
         TimeZone                0
         Job                     0
         Children                0
         Age                     0
         Income                  0
         Marital                 0
         Gender                  0
         Churn                   0
         Outage_sec_perweek      0
         Email                   0
         Contacts                0
         Yearly_equip_failure    0
         Techie                  0
         Contract                0
         Port_modem              0
         Tablet                  0
         InternetService         0
         Phone                   0
         Multiple                0
         OnlineSecurity          0
         OnlineBackup            0
         DeviceProtection        0
         TechSupport             0
         StreamingTV             0
         StreamingMovies         0
         PaperlessBilling        0
         PaymentMethod           0
         Tenure                  0
         MonthlyCharge           0
         Bandwidth_GB_Year       0
         item1_responses         0
         item2_fixes             0
         item3_replacements      0
         item4_reliability       0
         item5_options           0
         item6_respectfulness    0
         item7_courteous         0
         item8_listening         0
         dtype: int64
```

2. The target variable is Tenure, a continuous dependent variable. The predictor variables are both categorical and continuous. The predictor variables are as follows (on the left of the first column). Please see the next 3 charts to see the type of data each variable is and an example of what the data looks like in the data set.

   Children -0.366423

   Age 0.040279

   Income-0.000001

   Outage_sec_perweek 0.009168

Email 0.000752

Contacts -0.023848

Yearly_equip_failure -0.027181

Bandwidth GB Year 0.011966

MonthlyCharge 0.018156

iteml_responses 0.057365

item2 fixes -0.054017

item3_replacements 0.024902

item4_reliability -0.009846

item5_options -0.038197

item6_respectfulness -0.014343

item7 courteous -0.005482

item8_listening -0.053401

Churn numeric 1. 347461

Techie numeric -0.053166

Port- modem- numeric -0.043823

Tablet numeric 0.004083

Phone numeric -0.014566

Multiple_numeric 1.380790

OnlineSecurity_numeric 0.998428

OnlineBackup_numeric 1.460700

DeviceProtection numeric 1.171649

TechSupport_numeric 0.349225

StreamingTV_numeric	3.228143
StreamingMovies_numeric	3.126839
PaperlessBilling_numeric	-0.076023
InternetService numeric	-0.333164
Contract numeric	-0.021971
Gender numeric	0.603957

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 7 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Lat                 10000 non-null  float64
 1   Lng                 10000 non-null  float64
 2   Income              10000 non-null  float64
 3   Outage_sec_perweek  10000 non-null  float64
 4   Tenure              10000 non-null  float64
 5   MonthlyCharge       10000 non-null  float64
 6   Bandwidth_GB_Year   10000 non-null  float64
dtypes: float64(7)
memory usage: 547.0 KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   CaseOrder             10000 non-null  int64
 1   Zip                   10000 non-null  int64
 2   Population            10000 non-null  int64
 3   Children              10000 non-null  int64
 4   Age                   10000 non-null  int64
 5   Email                 10000 non-null  int64
 6   Contacts              10000 non-null  int64
 7   Yearly_equip_failure  10000 non-null  int64
 8   Item1                 10000 non-null  int64
 9   Item2                 10000 non-null  int64
 10  Item3                 10000 non-null  int64
 11  Item4                 10000 non-null  int64
 12  Item5                 10000 non-null  int64
 13  Item6                 10000 non-null  int64
 14  Item7                 10000 non-null  int64
 15  Item8                 10000 non-null  int64
dtypes: int64(16)
memory usage: 1.2 MB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 27 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Customer_id       10000 non-null  object
 1   Interaction       10000 non-null  object
 2   UID               10000 non-null  object
 3   City              10000 non-null  object
 4   State             10000 non-null  object
 5   County            10000 non-null  object
 6   Area              10000 non-null  object
 7   TimeZone          10000 non-null  object
 8   Job               10000 non-null  object
 9   Marital           10000 non-null  object
 10  Gender            10000 non-null  object
 11  Churn             10000 non-null  object
 12  Techie            10000 non-null  object
 13  Contract          10000 non-null  object
 14  Port_modem        10000 non-null  object
 15  Tablet            10000 non-null  object
```

```
 16  InternetService   10000 non-null  object
 17  Phone             10000 non-null  object
 18  Multiple          10000 non-null  object
 19  OnlineSecurity    10000 non-null  object
 20  OnlineBackup      10000 non-null  object
 21  DeviceProtection  10000 non-null  object
 22  TechSupport       10000 non-null  object
 23  StreamingTV       10000 non-null  object
 24  StreamingMovies   10000 non-null  object
 25  PaperlessBilling  10000 non-null  object
 26  PaymentMethod     10000 non-null  object
dtypes: object(27)
memory usage: 2.1+ MB
None
   CaseOrder Customer_id                           Interaction  \
0          1     K409198  aa90260b-4141-4a24-8e36-b04ce1f4f77b
1          2     S120509  fb76459f-c047-4a9d-8af9-e0f7d4ac2524

                                UID         City State                 County  \
0  e885b299883d4f9fb18e39c75155d990  Point Baker    AK  Prince of Wales-Hyder
1  f2de8bef964785f41a2959829830fb8a  West Branch    MI                 Ogemaw

     Zip       Lat        Lng  ...  MonthlyCharge Bandwidth_GB_Year Item1  \
0  99927  56.25100 -133.37571  ...     172.455519        904.536110     5
1  48661  44.32893  -84.24080  ...     242.632554        800.982766     3

  Item2  Item3  Item4  Item5 Item6 Item7 Item8
0     5      5      3      4     4     3     4
1     4      3      3      4     3     4     4

[2 rows x 50 columns]
```

Additionally, there are no outliers that need removal, as all datapoints should be kept to preserve the integrity of the data set. Finally, the measures of central tendency for the variables are as follows: Please note all of the categorical values have been changed to numeric and are represented by their original title_numeric. Yes = 0, No = 1.

Mean

```
CaseOrder              5000.500000
Zip                   49153.319600
Lat                      38.757567
Lng                     -90.782536
Population             9756.562400
Children                  2.087700
Age                      53.078400
Income                39806.926771
Outage_sec_perweek       10.001848
```

```
Email                    12.016000
Contacts                  0.994200
Yearly_equip_failure        0.398000
Tenure                   34.526188
MonthlyCharge            172.624816
Bandwidth_GB_Year         3392.341550
item1_responses           3.490800
item2_fixes              3.505100
item3_replacements         3.487000
item4_reliability          3.497500
item5_options             3.492900
item6_respectfulness        3.497300
item7_courteous            3.509500
item8_listening            3.495600
Churn_numeric             0.735000
Area_numeric              1.000000
Marital_numeric           2.017500
Gender_numeric            0.571800
Contract_numeric          1.034000
PaymentMethod_numeric       1.700300
InternetService_numeric      0.772100
Techie_numeric            0.832100
Port_modem_numeric          0.516600
Tablet_numeric            0.700900
Phone_numeric             0.093300
Multiple_numeric          0.539200
OnlineSecurity_numeric      0.642400
OnlineBackup_numeric        0.549400
DeviceProtection_numeric     0.561400
TechSupport_numeric         0.625000
StreamingTV_numeric         0.507100
StreamingMovies_numeric      0.511000
PaperlessBilling_numeric     0.411800
dtype: float64
```

<u>Median</u>

```
CaseOrder              5000.500000
Zip                48869.500000
Lat                  39.395800
Lng                 -87.918800
Population             2910.500000
Children               1.000000
Age                 53.000000
Income              33170.605000
Outage_sec_perweek        10.018560
Email                12.000000
Contacts              1.000000
Yearly_equip_failure        0.000000
Tenure               35.430507
MonthlyCharge          167.484700
Bandwidth_GB_Year        3279.536903
```

```
item1_responses            3.000000
item2_fixes                4.000000
item3_replacements         3.000000
item4_reliability          3.000000
item5_options              3.000000
item6_respectfulness       3.000000
item7_courteous            4.000000
item8_listening            3.000000
Churn_numeric              1.000000
Area_numeric               1.000000
Marital_numeric            2.000000
Gender_numeric             1.000000
Contract_numeric           1.000000
PaymentMethod_numeric      2.000000
InternetService_numeric    1.000000
Techie_numeric             1.000000
Port_modem_numeric         1.000000
Tablet_numeric             1.000000
Phone_numeric              0.000000
Multiple_numeric           1.000000
OnlineSecurity_numeric     1.000000
OnlineBackup_numeric       1.000000
DeviceProtection_numeric   1.000000
TechSupport_numeric        1.000000
StreamingTV_numeric        1.000000
StreamingMovies_numeric    1.000000
PaperlessBilling_numeric   0.000000
dtype: float64
```

3. The steps to prepare the data for analysis are inside of the annotated code file below and summarized as follows:

    i. Import dataset to Python

    ii. Rename columns of survey to easily recognizable descriptions (ex: "Item1" to "item1_responses")

    iii. Get a description of the data set, structure (columns & rows) & data types.

    iv. View summary statistics

v. Check for records with missing data & impute missing data with meaningful measures of central tendency (mean, median or mode) or simply remove outliers that are several standard deviations above the mean. This step might not be necessary if it is determined we will not be removing outliers.

vi. Create numeric variables in order to encode categorical, yes/no data points into 1/0 numerical values.

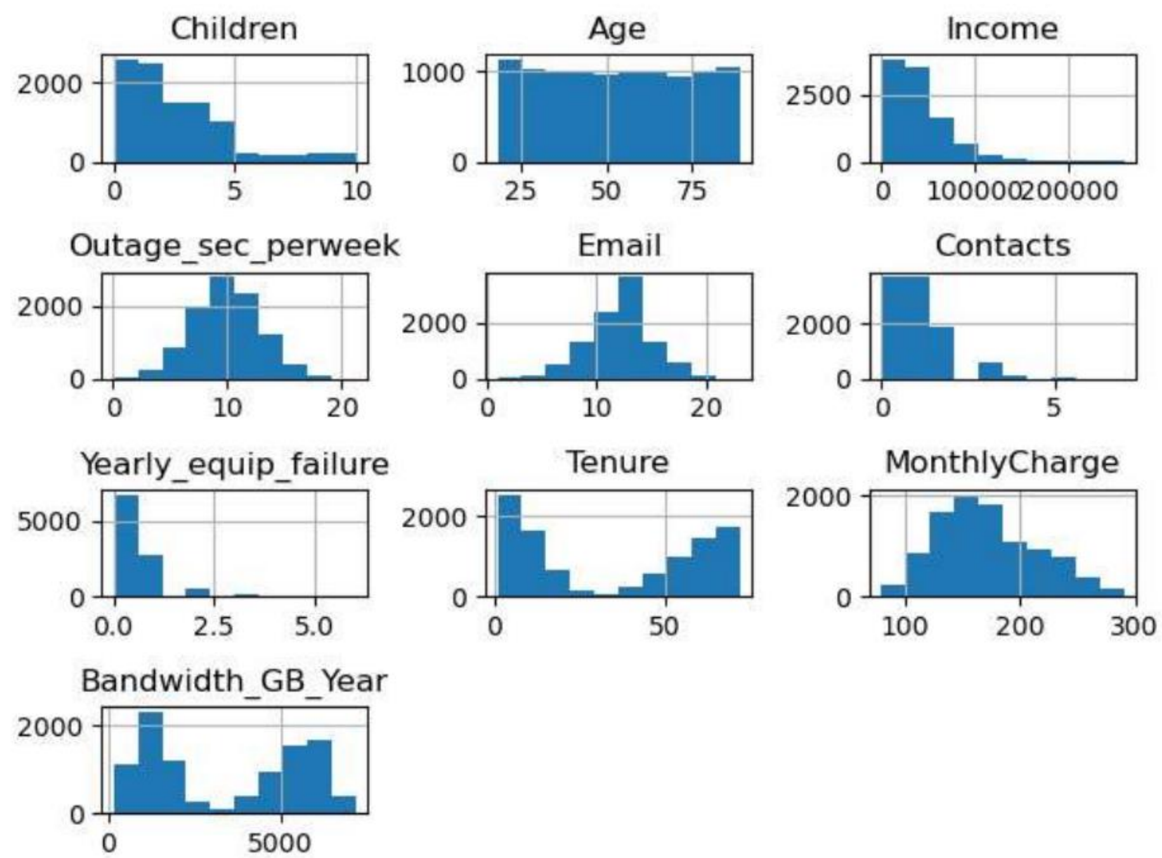
vii. View univariate & bivariate visualizations.

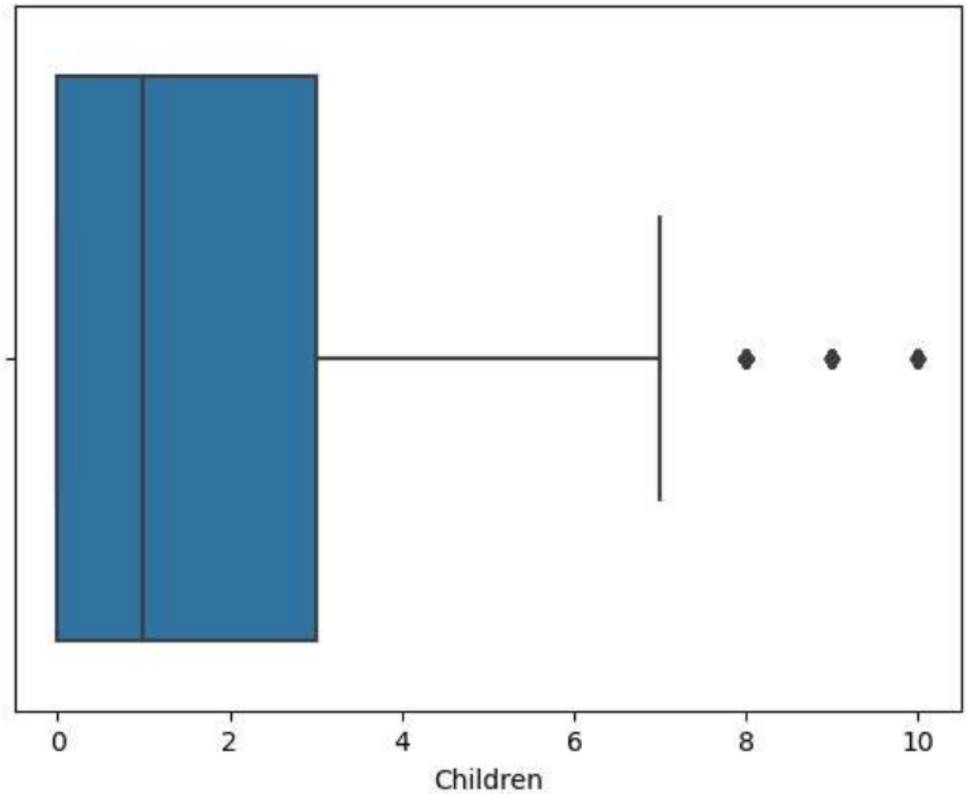
viii. Finally, the prepared dataset will be extracted & provided as "churn_Task1.csv"

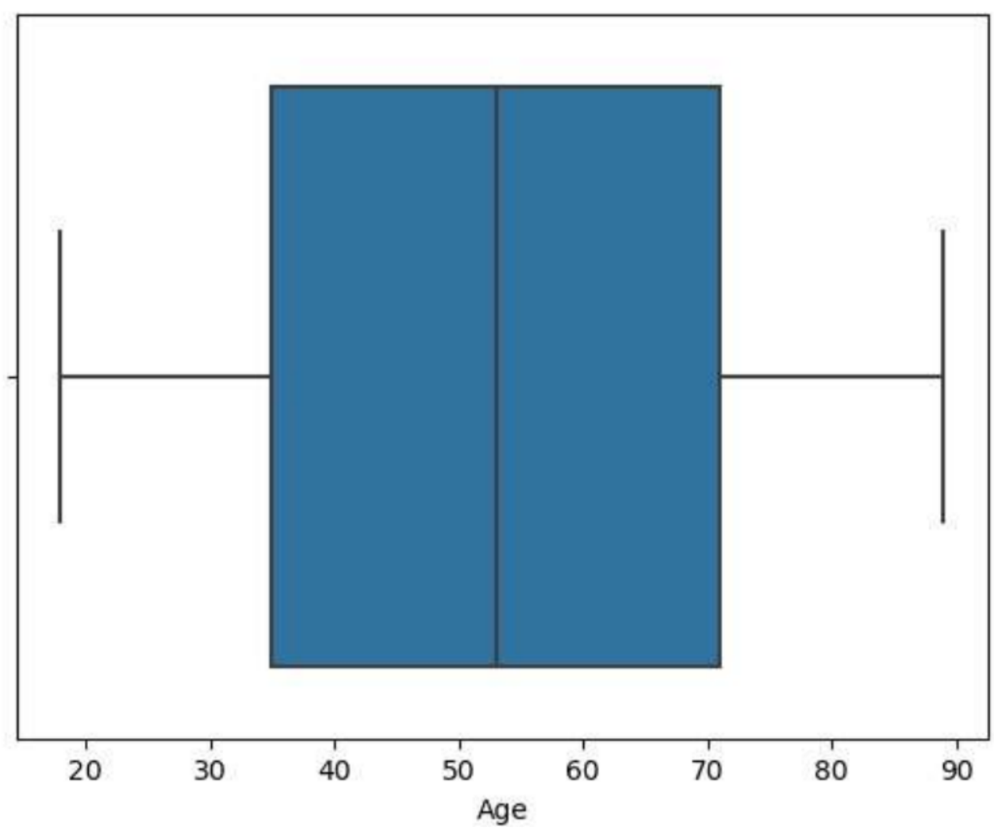
b. The annotated code can be found in “PA_D208_Code_Task1”

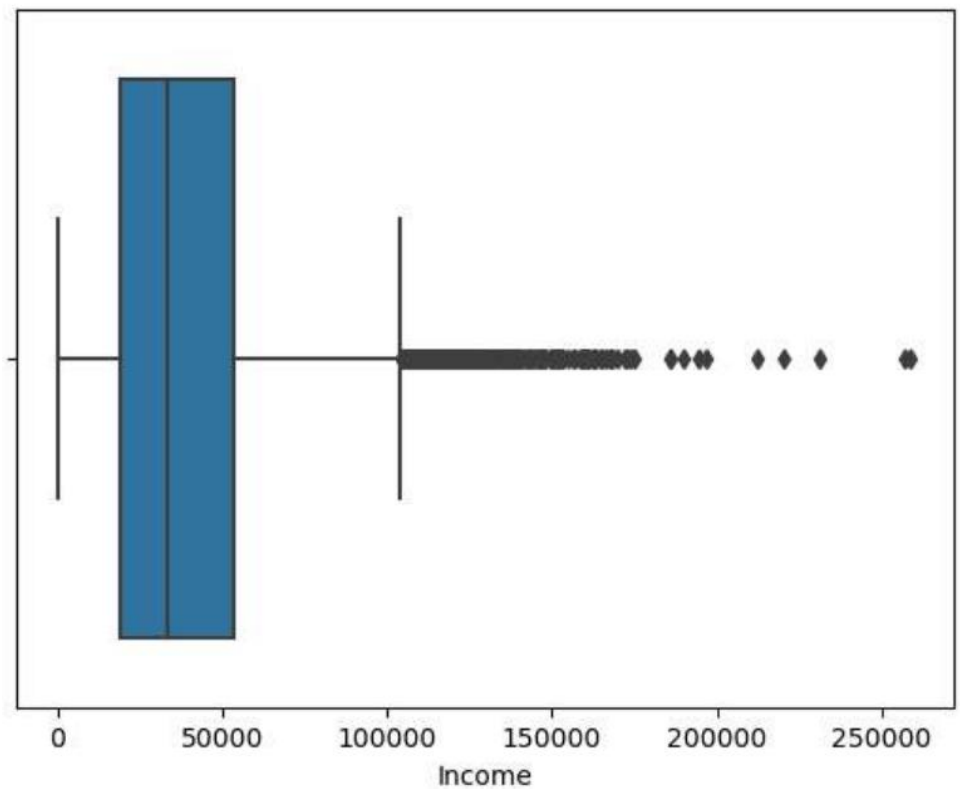
4. The visualizations are as follows:

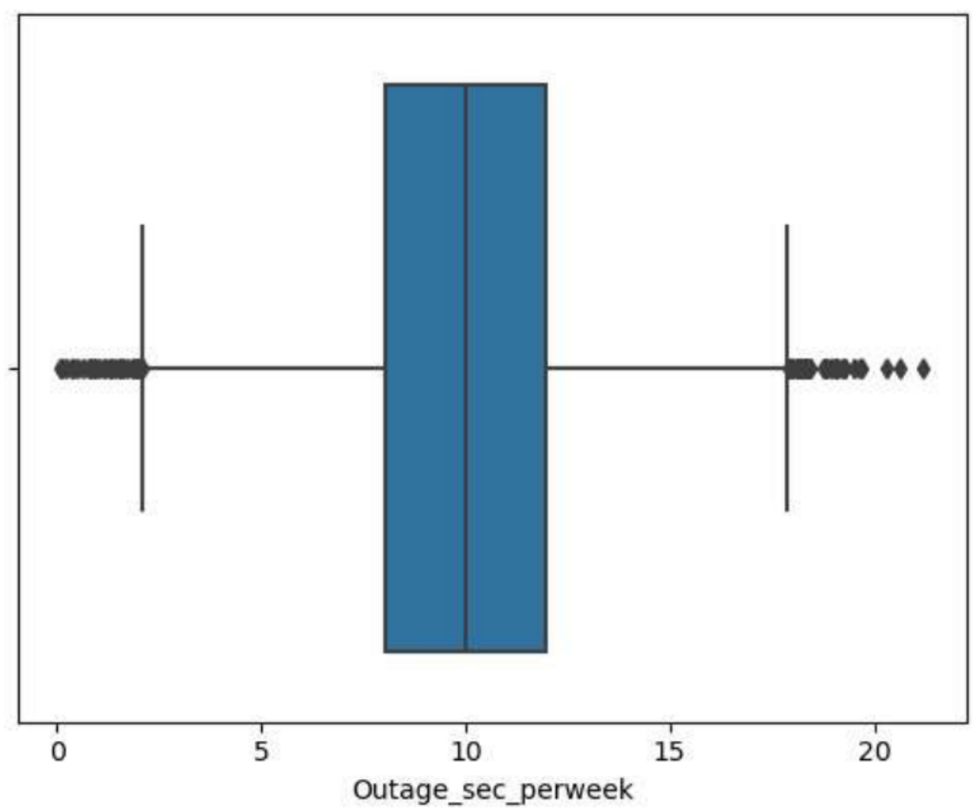
a. Univariate

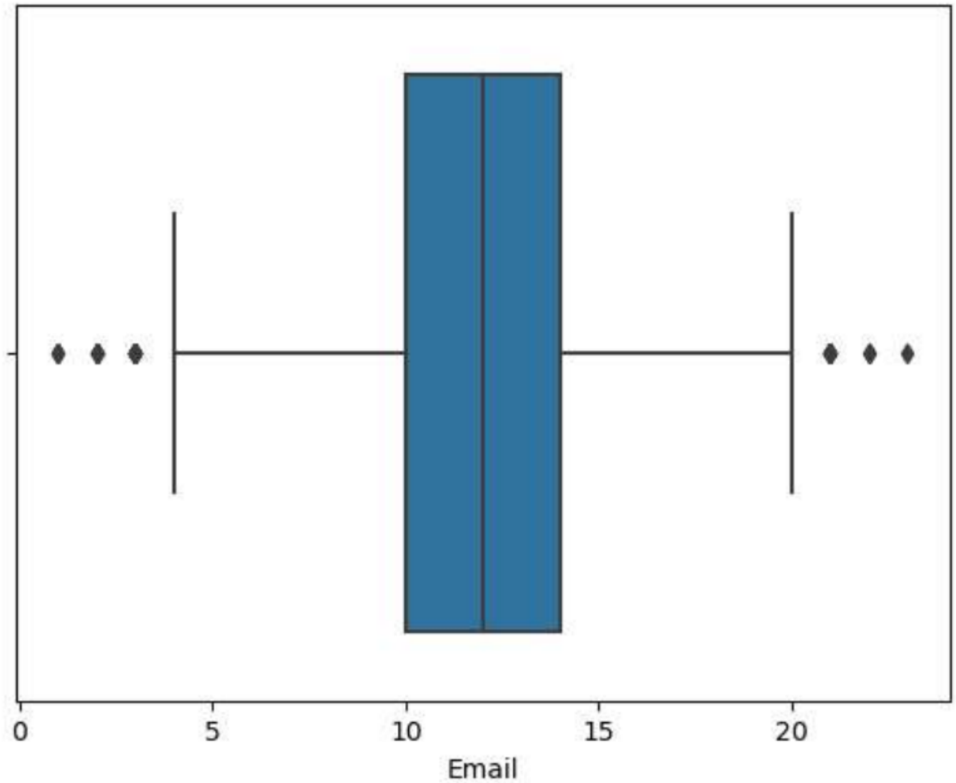

Children
Age
Income
Outage_sec_perweek
Email
Contacts
Yearly_equip_failure
Tenure
MonthlyCharge
Bandwidth_GB_Year

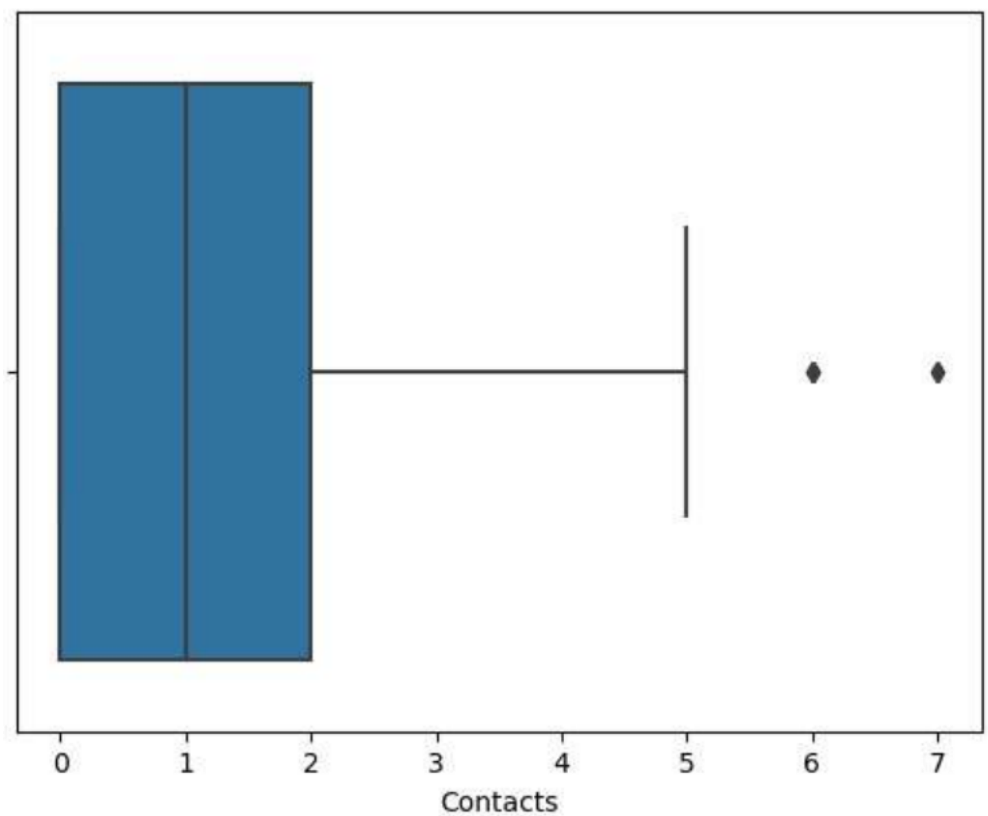

0
2
4
6
8
10
Children

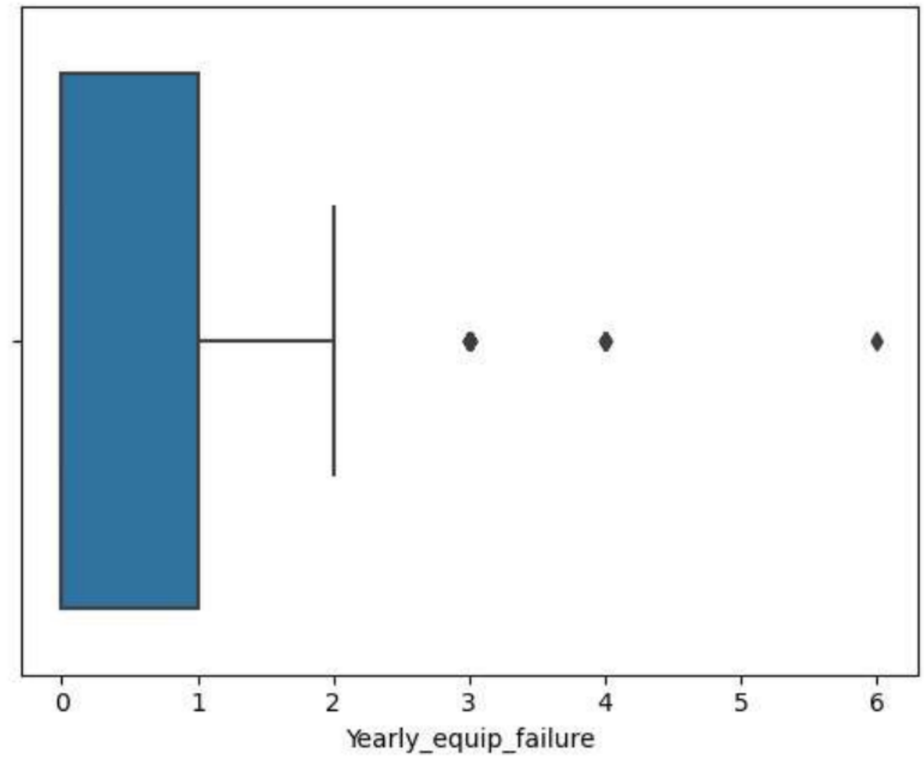

20
30
40
50
60
70
80
90
Age

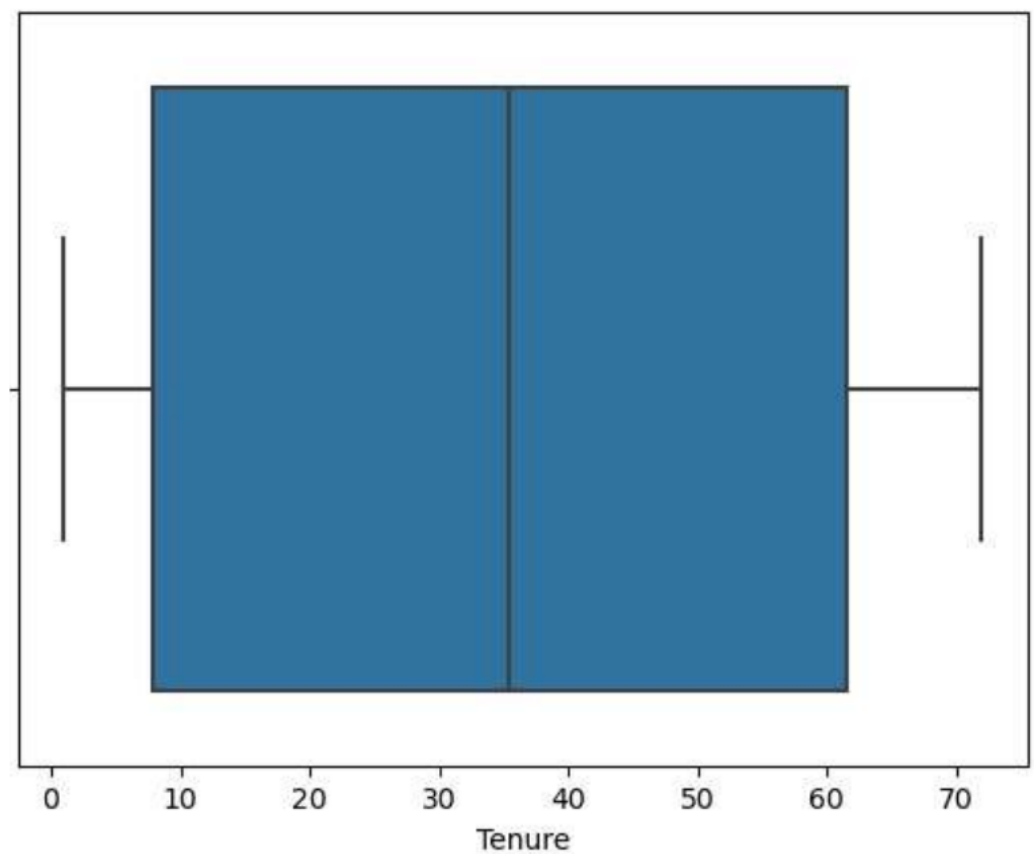

0
50000
100000
150000
200000
250000
Income

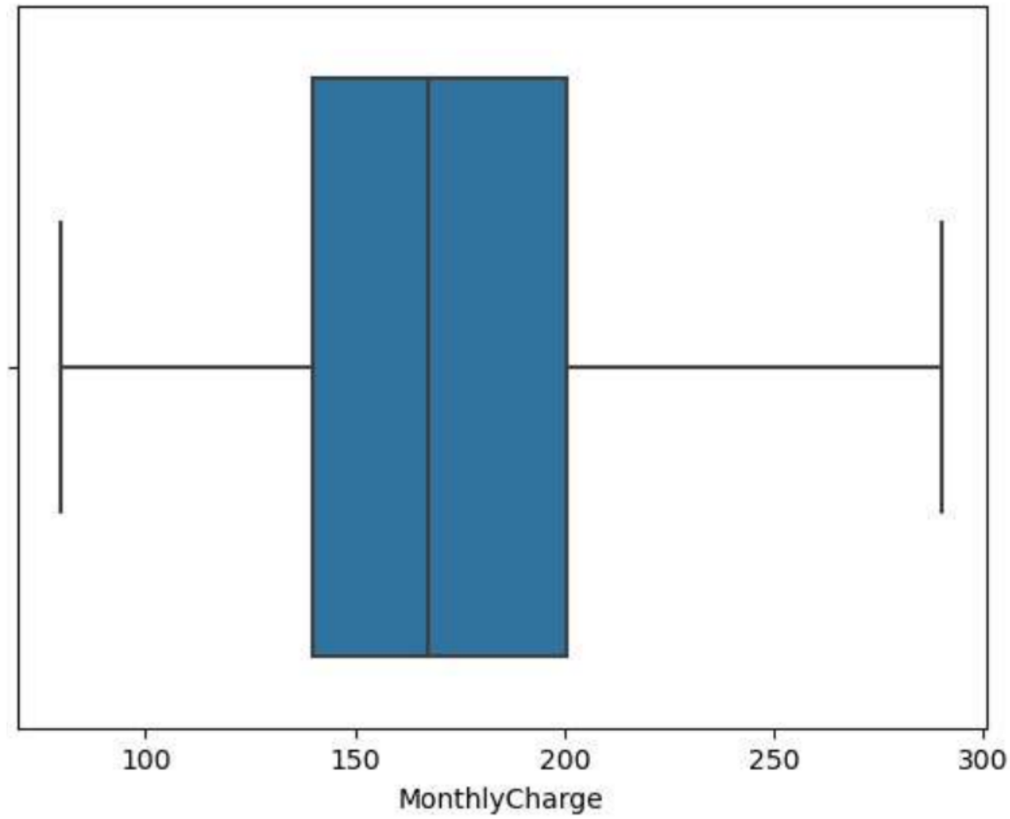

0
5
10
15
20
Outage_sec_perweek

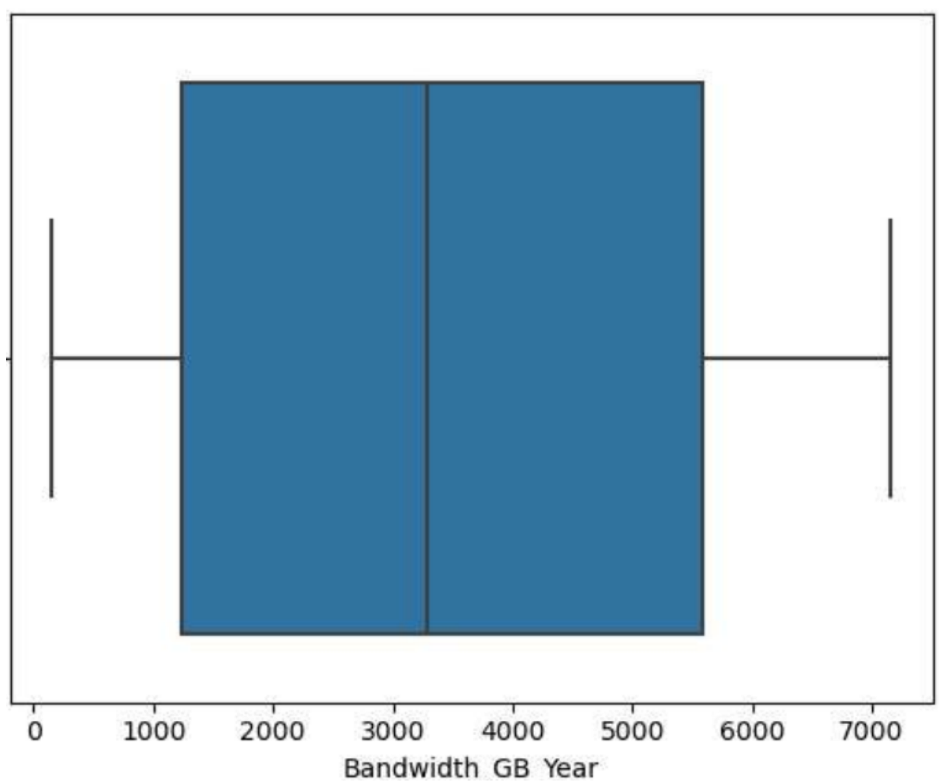

0
5
10
15
20
Email


0
1
2
3
4
5
6
7
Contacts

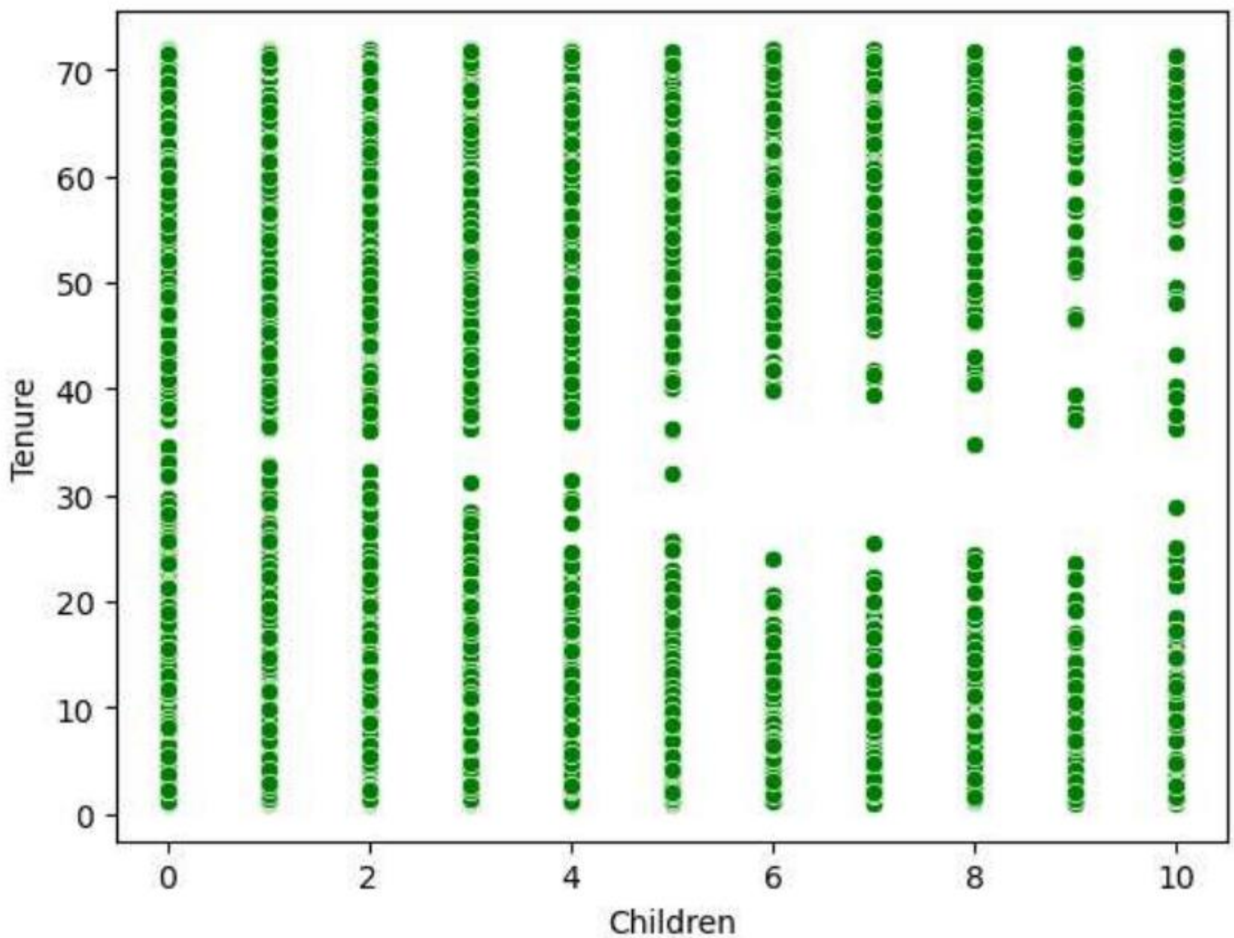

0
1
2
3
4
5
6
Yearly_equip_failure

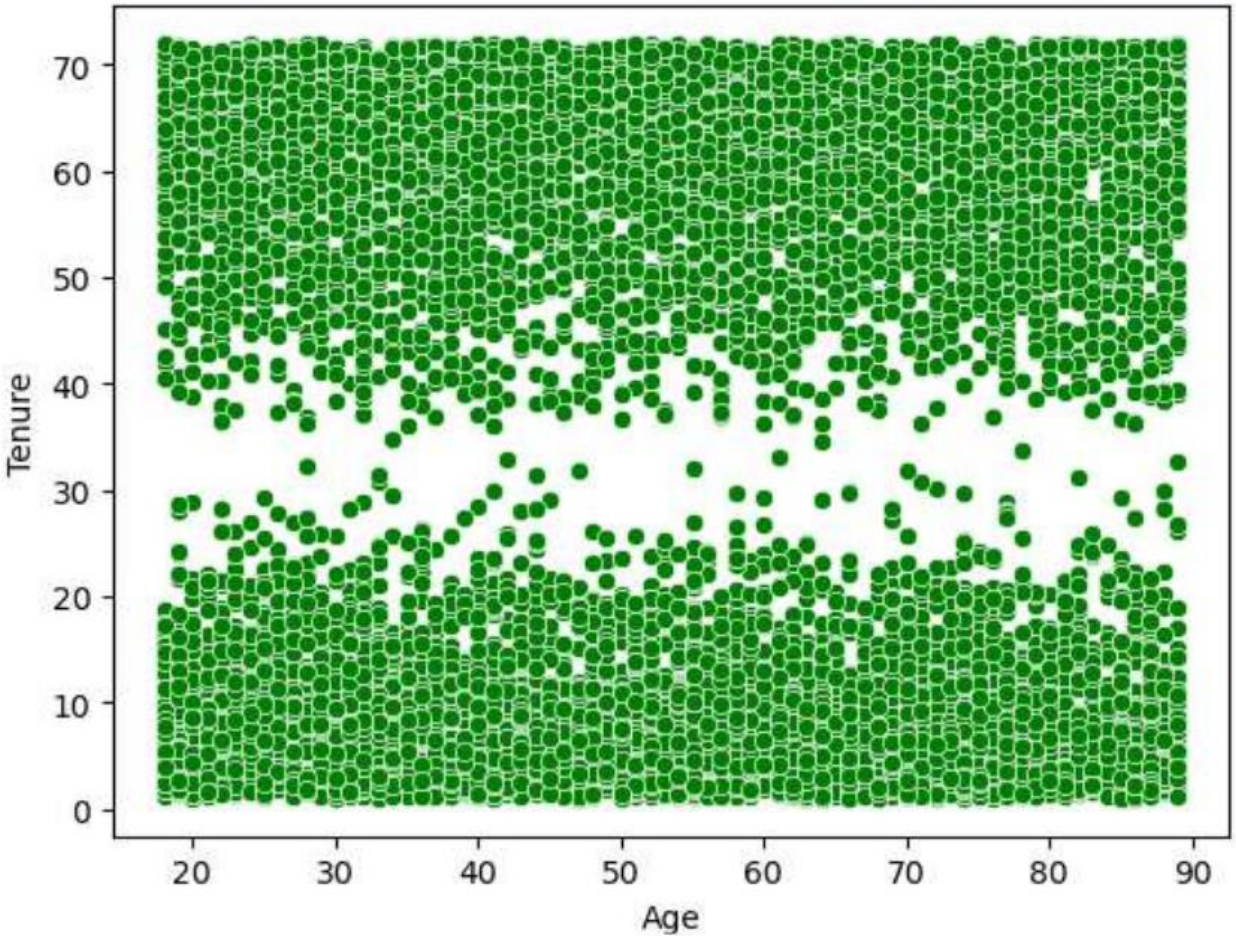

0
10
20
30
40
50
60
70
Tenure

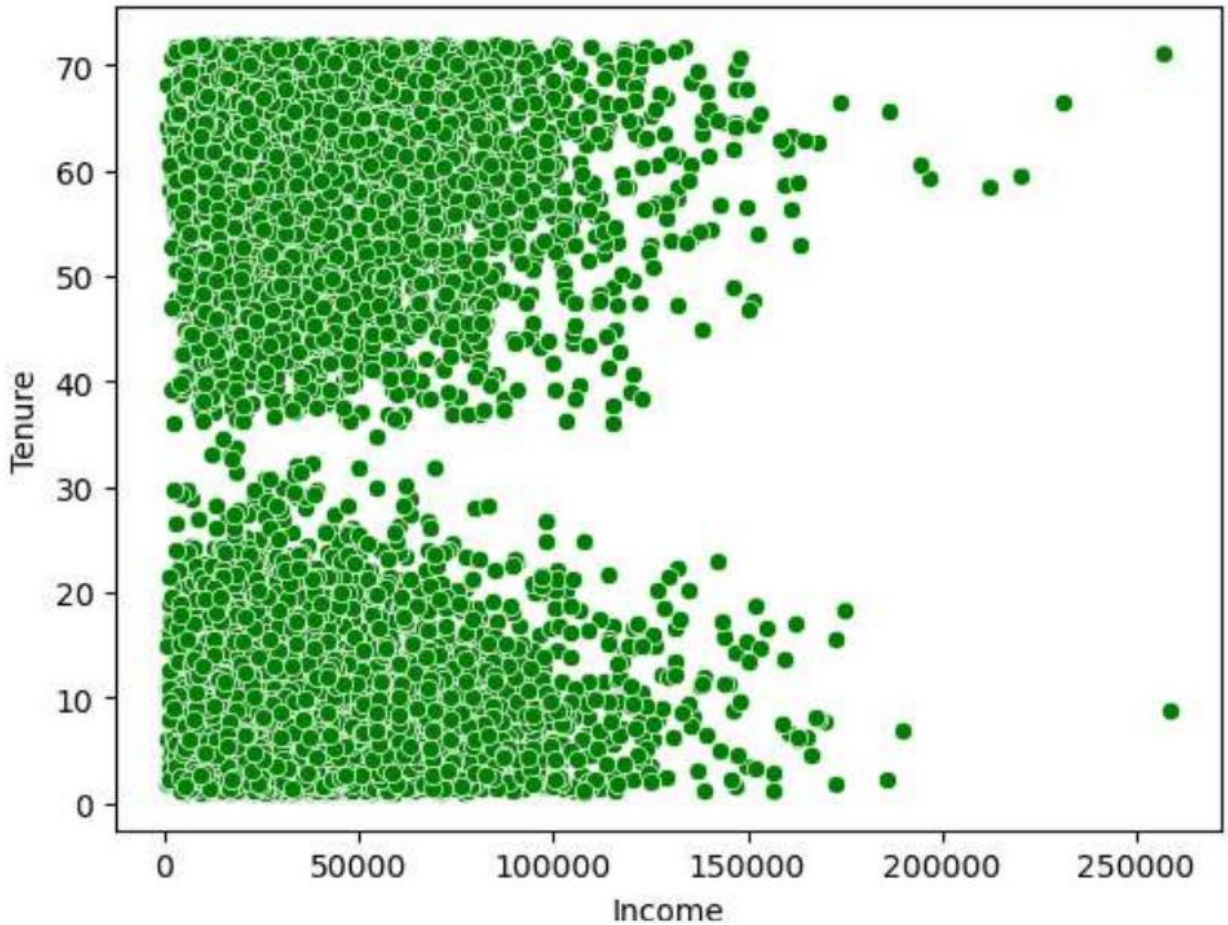

100
150
200
250
300
MonthlyCharge

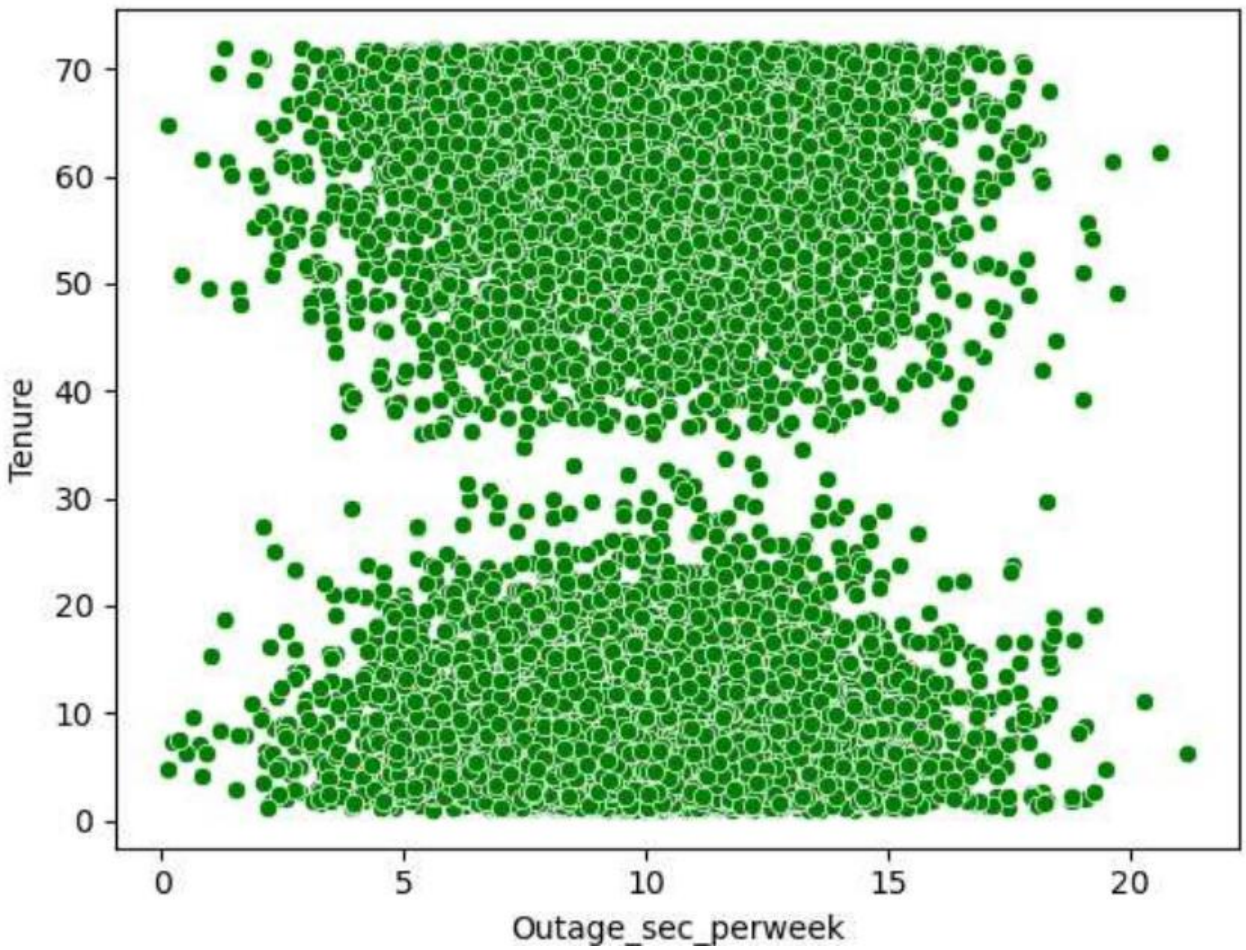

0
1000
2000
3000
4000
5000
6000
7000
Bandwidth_GB_Year

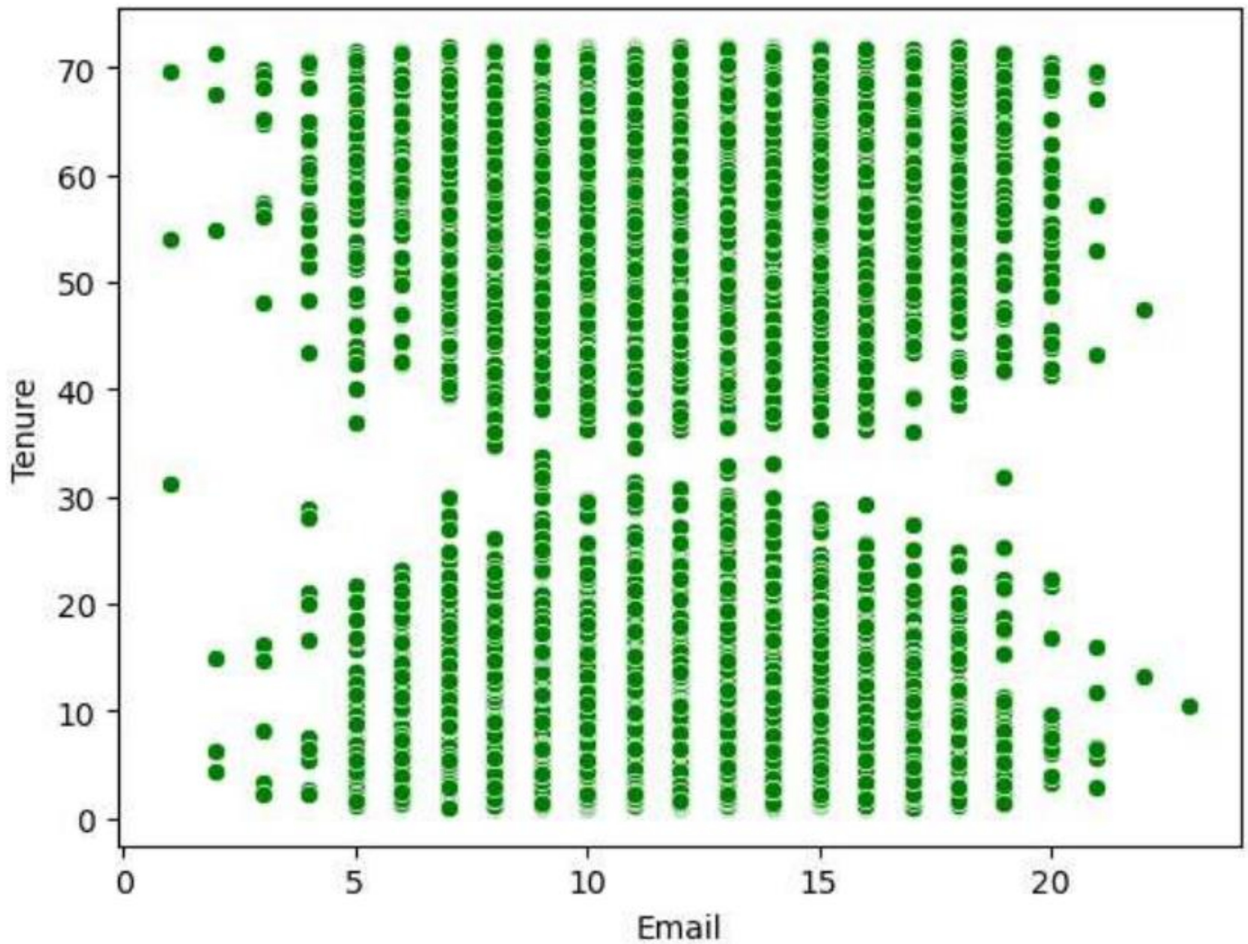
b. Bivariate

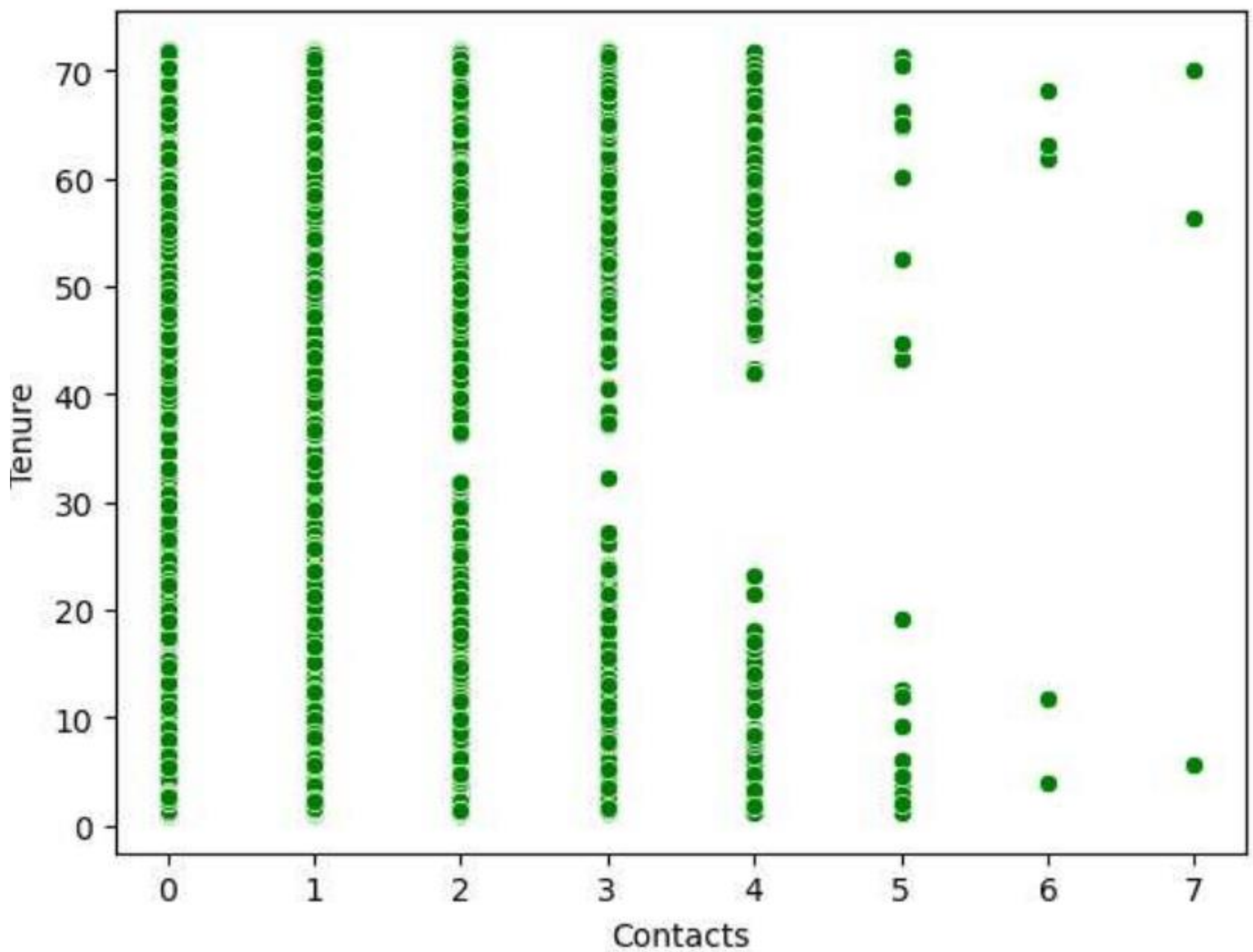


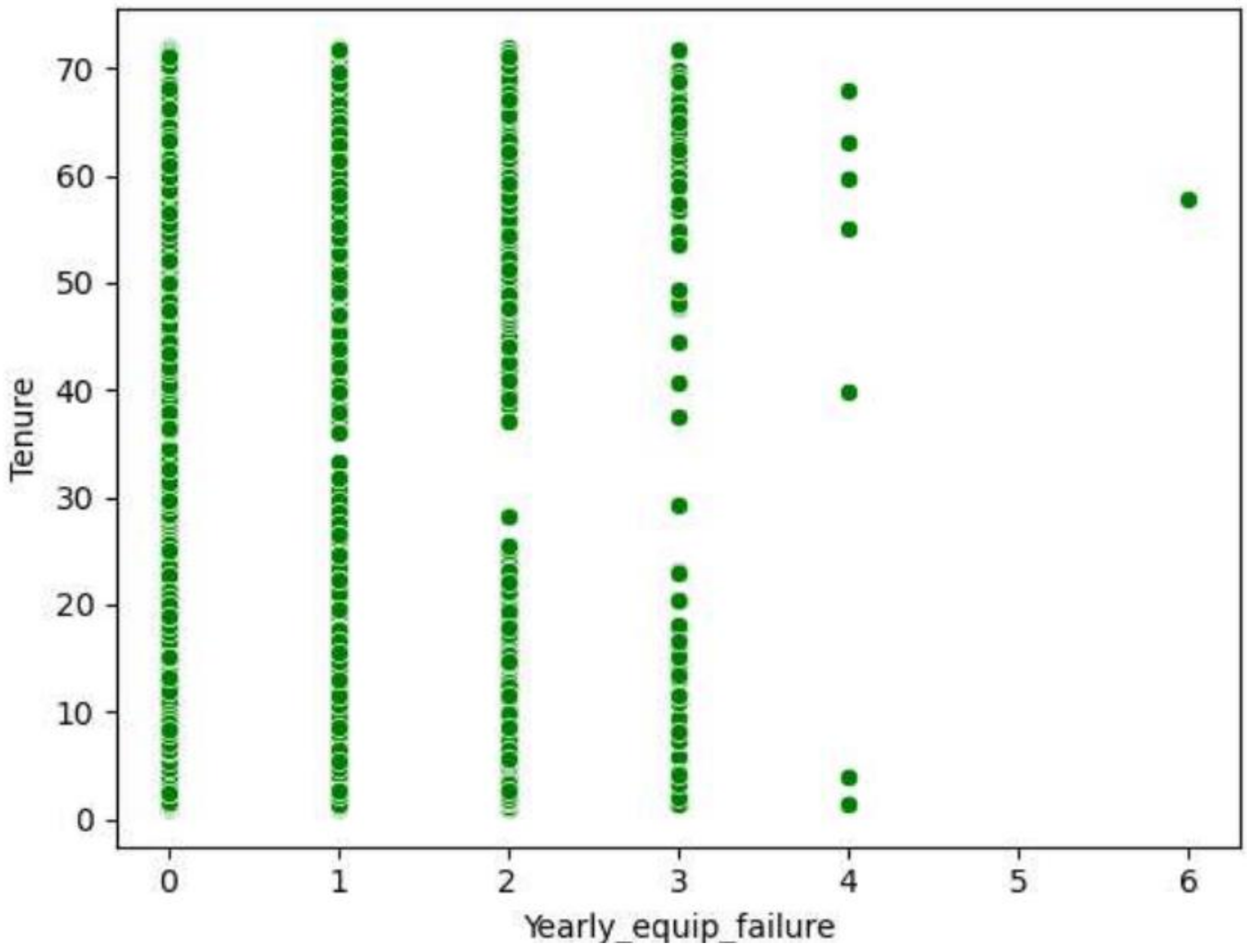

Tenure
Age
70
60
50
40
30
20
10
0
20
30
40
50
60
70
80
90

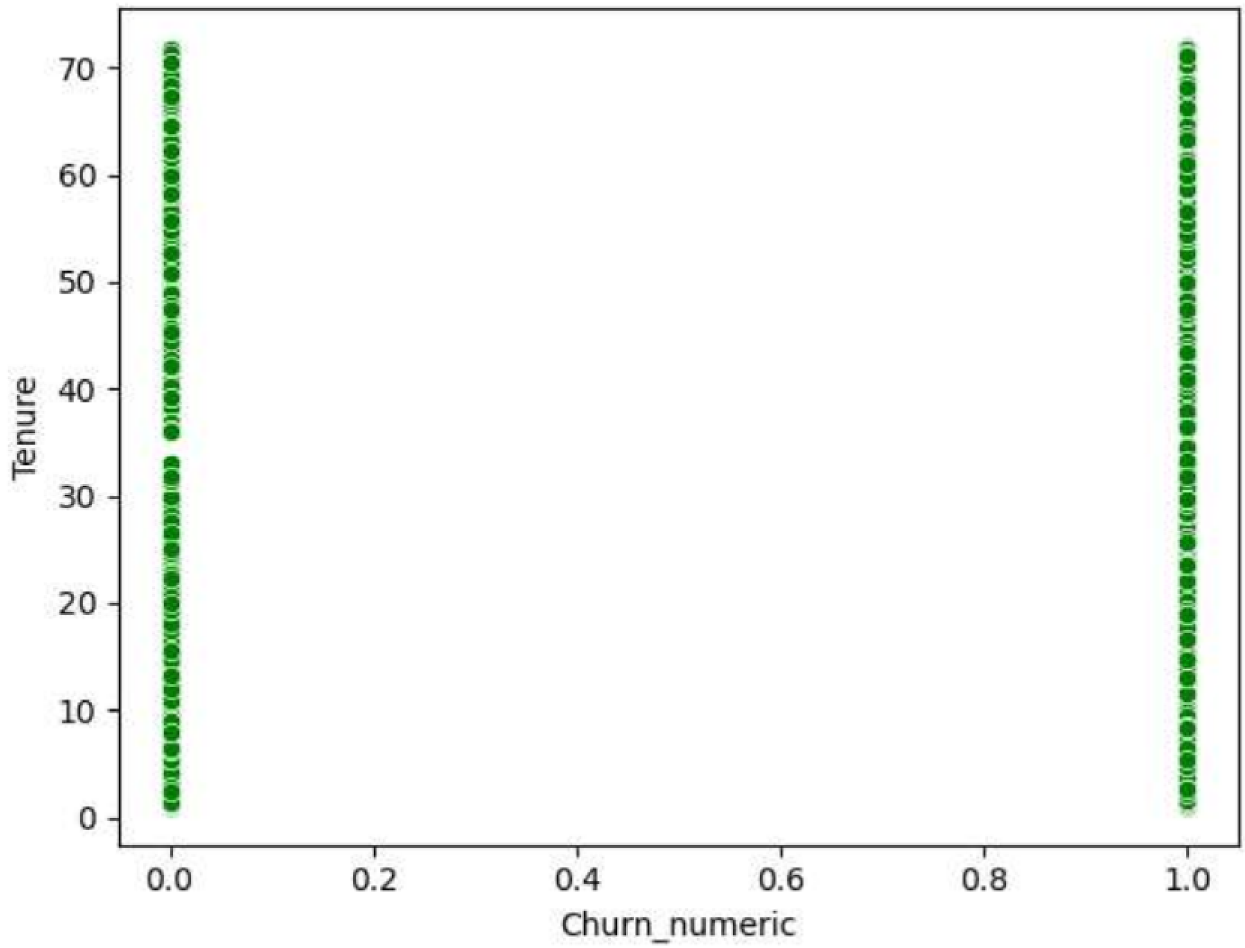

Tenure
Income
0
10
20
30
40
50
60
70
0
50000
100000
150000
200000
250000

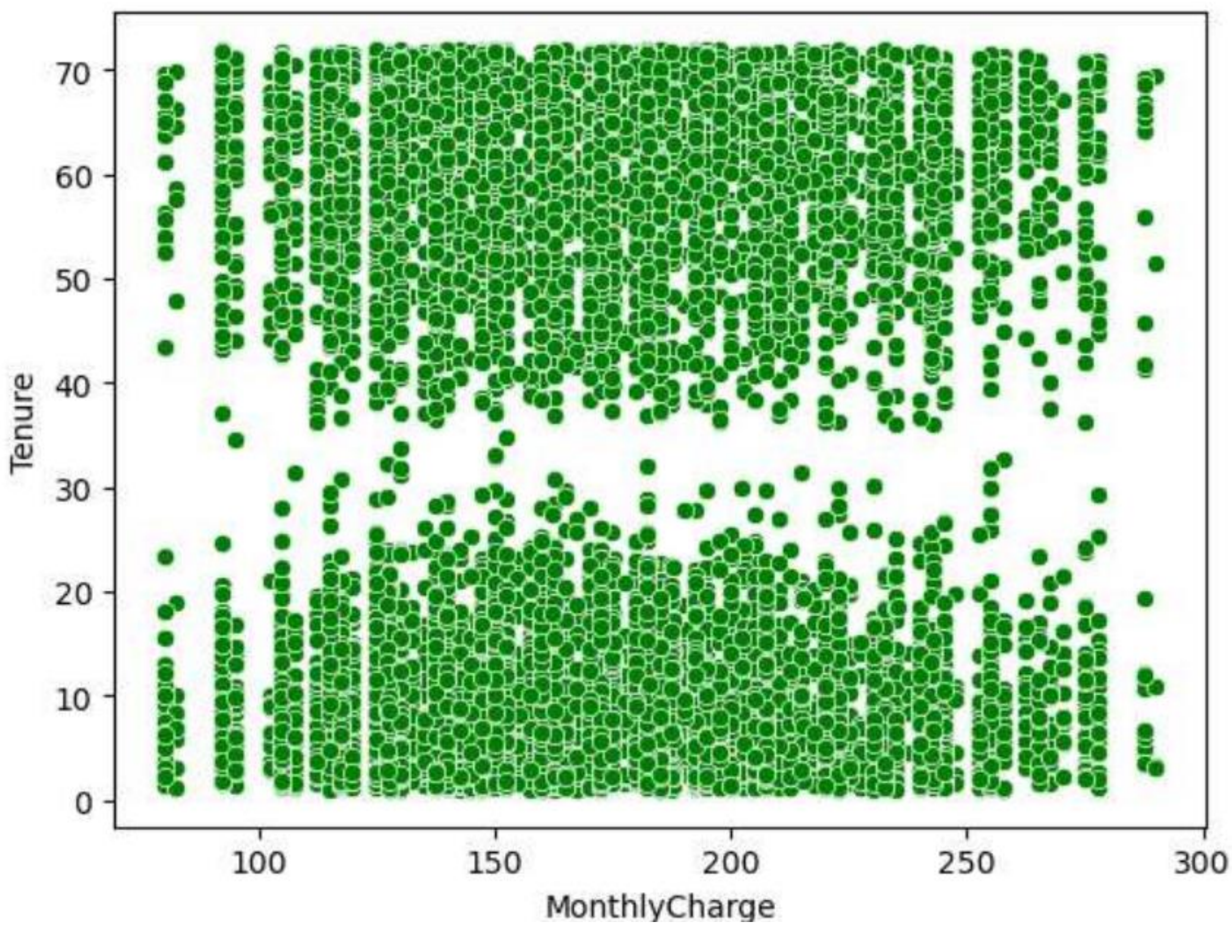

Tenure
Outage_sec_perweek
0
10
20
30
40
50
60
70
0
5
10
15
20

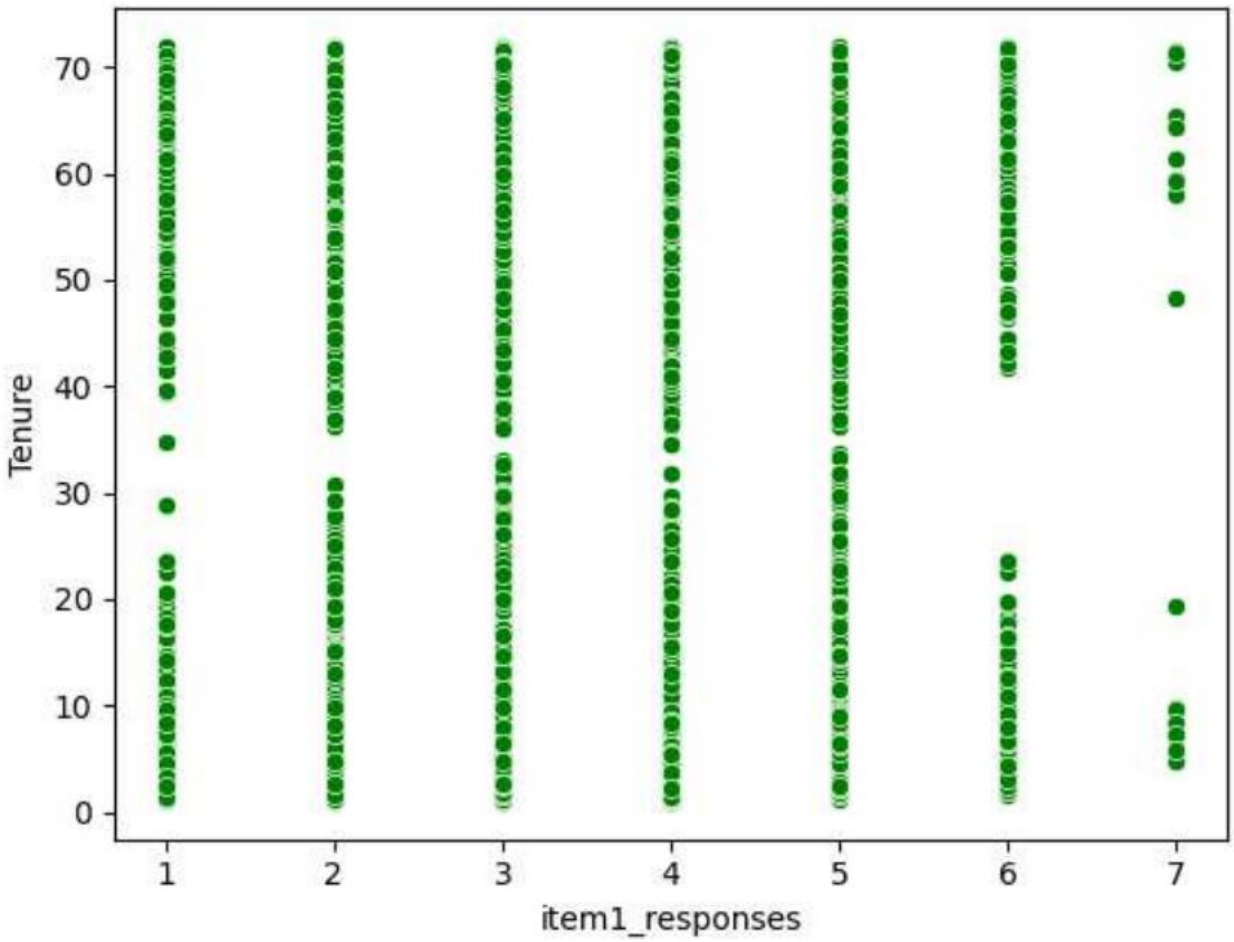

Tenure
Email
70
60
50
40
30
20
10
0
0
5
10
15
20

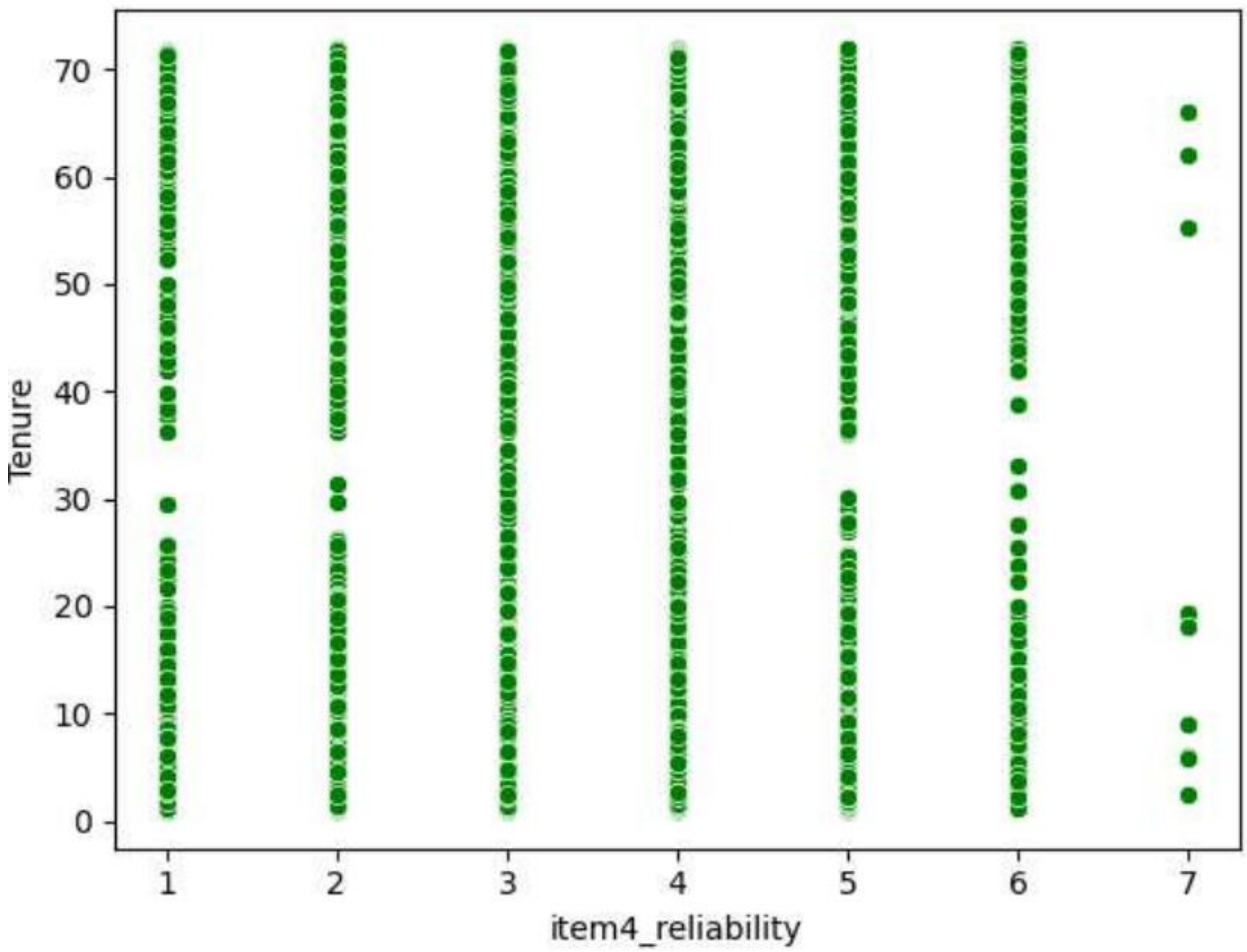

Tenure
Contacts
0
10
20
30
40
50
60
70
0
1
2
3
4
5
6
7

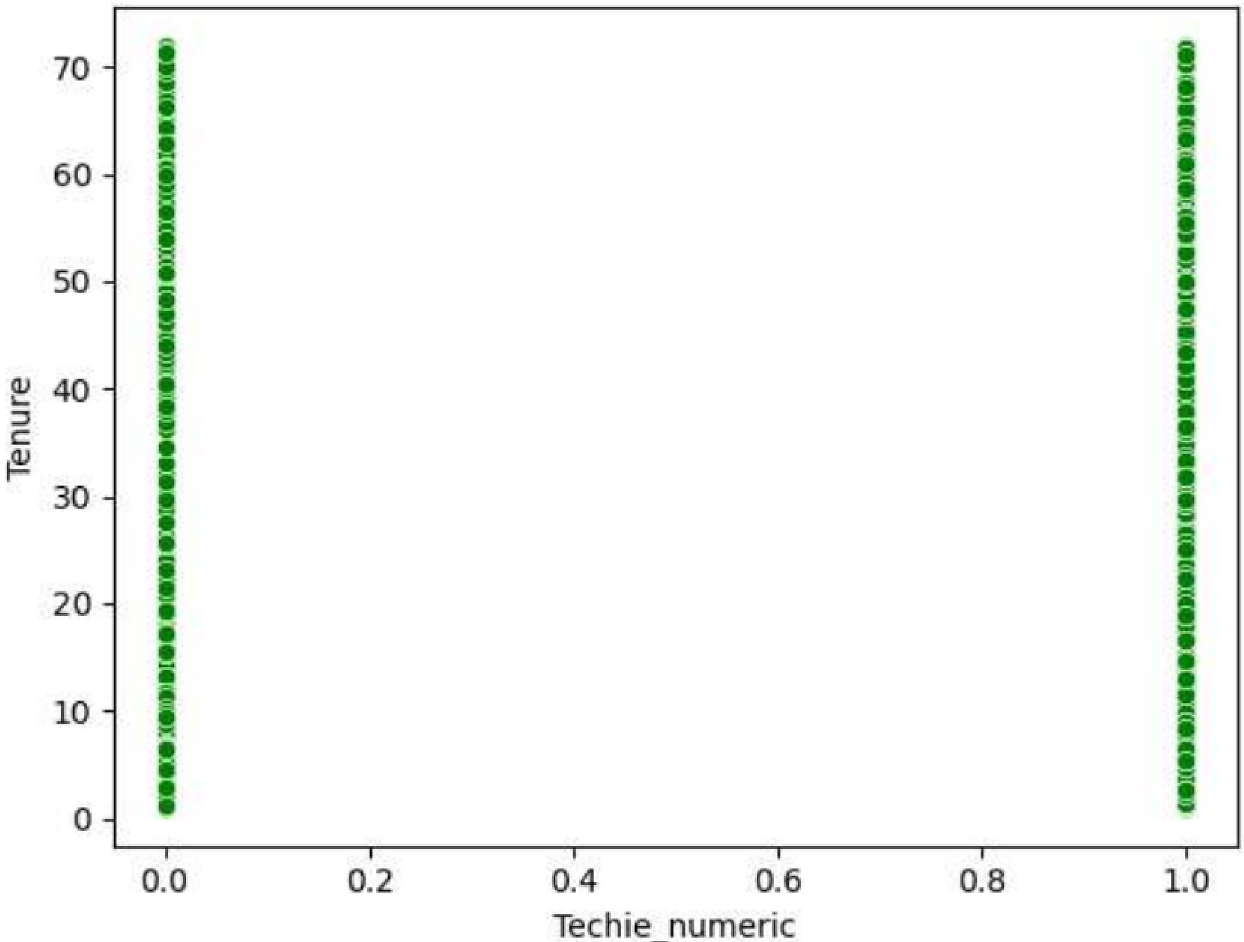

Tenure
Yearly_equip_failure
0
10
20
30
40
50
60
70
0
1
2
3
4
5
6

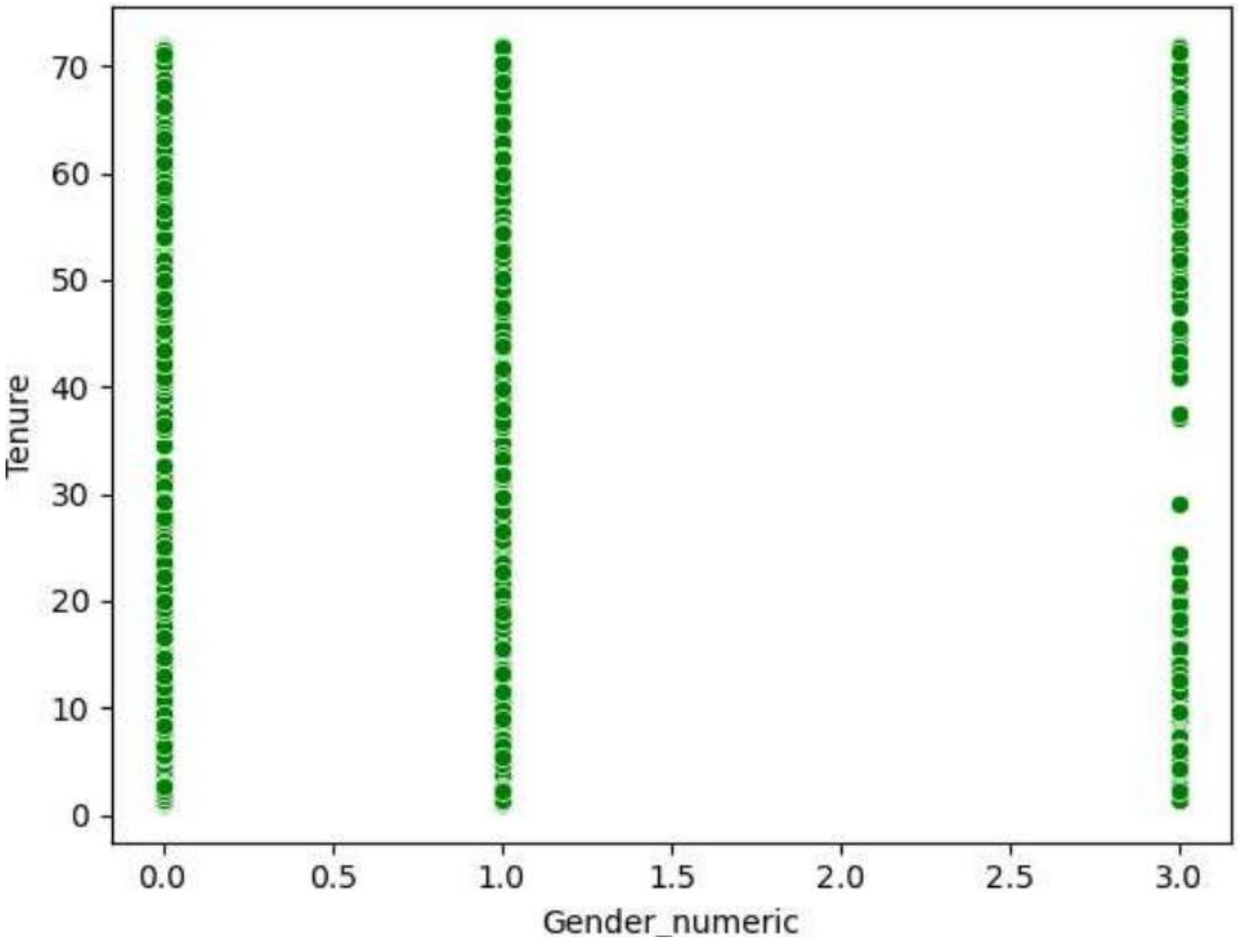

Tenure
Churn_numeric
70
60
50
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1.0

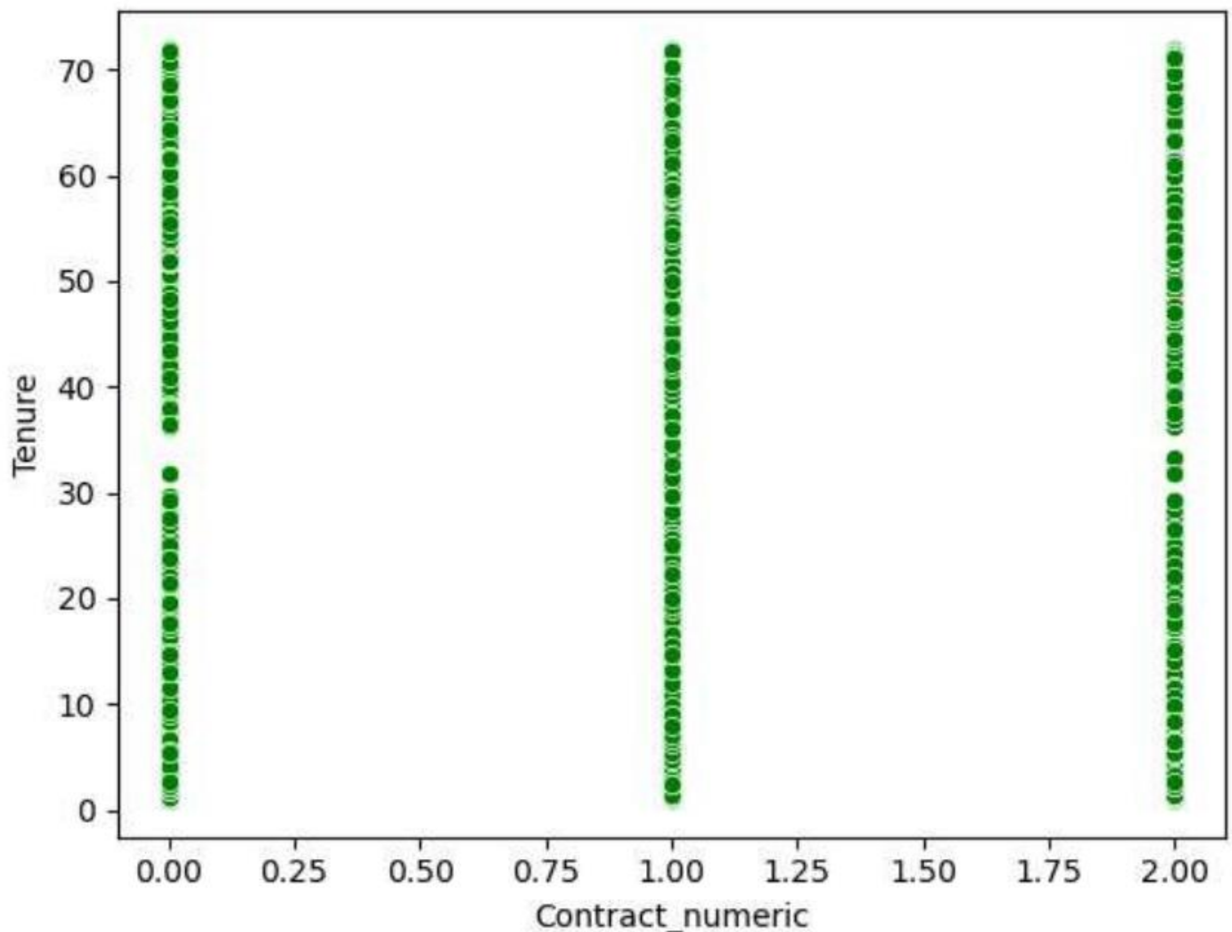

Tenure
MonthlyCharge
0
10
20
30
40
50
60
70
100
150
200
250
300

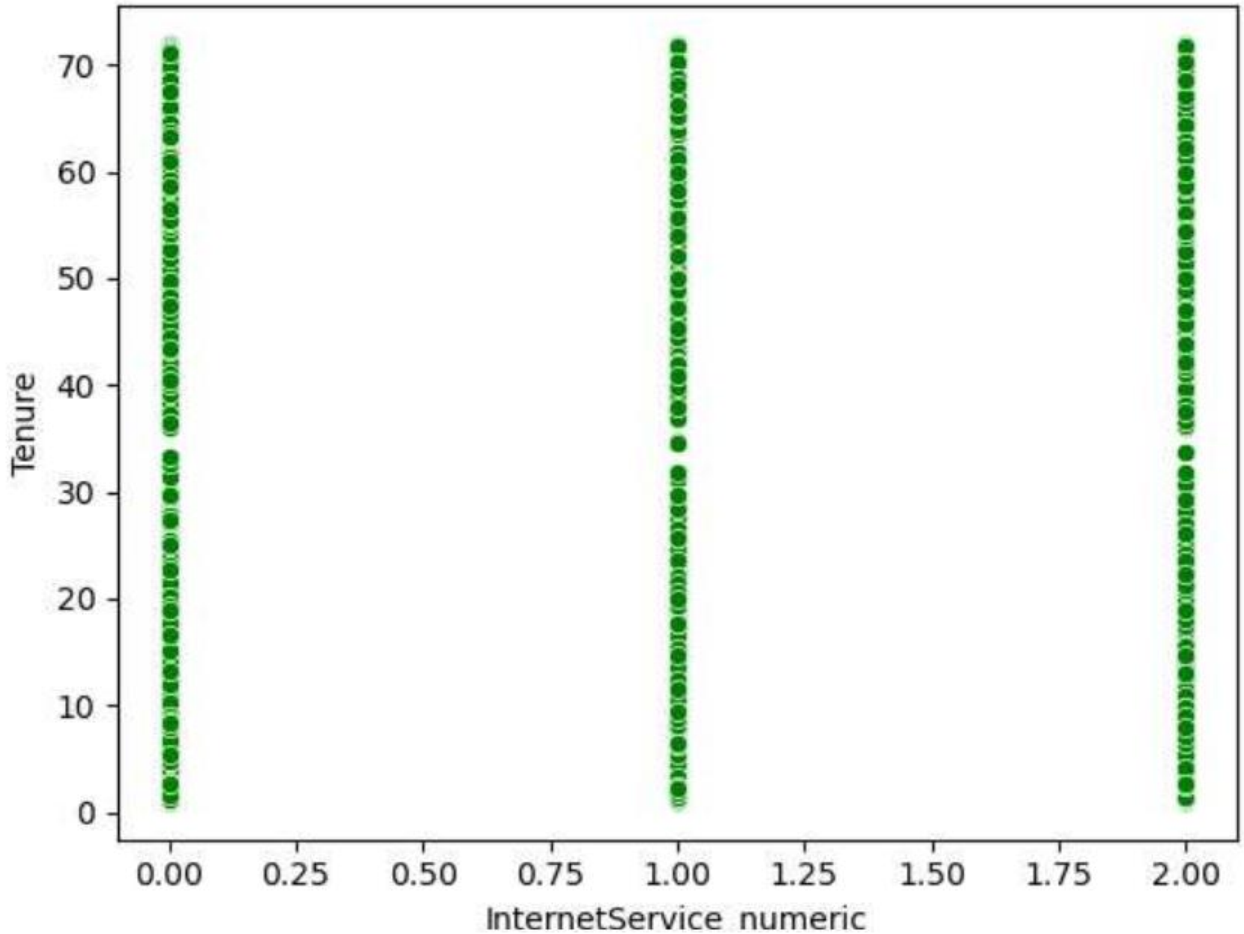

Tenure
item1_responses
0
10
20
30
40
50
60
70
1
2
3
4
5
6
7


Tenure
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
item4_reliability


Tenure
70
60
50
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1.0
Techie_numeric


Tenure
Gender_numeric
0
10
20
30
40
50
60
70
0.0
0.5
1.0
1.5
2.0
2.5
3.0


70
60
50
40
30
20
10
0
Tenure
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Contract_numeric



5. The prepared data set is included as “D208_cleaned_task1”

Part IV: Model Comparison and Analysis

D. Compare an initial and a reduced multiple regression model by doing the following:

> **1. Construct an initial multiple regression model from *all* predictors that were identified in Part C2.**
>
> **2. Justify a statistically based variable selection procedure and a model evaluation metric to reduce the initial model in a way that aligns with the research question.**
>
> **3. Provide a reduced multiple regression model that includes *both* categorical and continuous variables.**

1. Initial multiple regression model from all predictors:

```
Intercept                     -17.523751
Children                       -0.366423
Age                             0.040279
Income                         -0.000001
Outage_sec_perweek              0.009168
Email                           0.000752
Contacts                       -0.023848
Yearly_equip_failure           -0.027181
Bandwidth_GB_Year               0.011966
MonthlyCharge                   0.018156
item1_responses                 0.057365
item2_fixes                    -0.054017
item3_replacements              0.024902
item4_reliability              -0.009846
item5_options                  -0.038197
item6_respectfulness           -0.014343
item7_courteous                -0.005482
item8_listening                -0.053401
Churn_numeric                   1.347461
Techie_numeric                 -0.053166
Port_modem_numeric             -0.043823
Tablet_numeric                  0.004083
Phone_numeric                  -0.014566
Multiple_numeric                1.380790
OnlineSecurity_numeric          0.998428
OnlineBackup_numeric            1.460700
DeviceProtection_numeric        1.171649
TechSupport_numeric             0.349225
StreamingTV_numeric             3.228143
StreamingMovies_numeric         3.126839
PaperlessBilling_numeric       -0.076023
InternetService_numeric        -0.333164
Contract_numeric               -0.021971
Gender_numeric                  0.603957
dtype: float64
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Tenure   R-squared:                       0.992
Model:                            OLS   Adj. R-squared:                  0.992
Method:                 Least Squares   F-statistic:                 3.916e+04
Date:                Wed, 25 Jan 2023   Prob (F-statistic):               0.00
Time:                        17:07:14   Log-Likelihood:                -22576.
No. Observations:               10000   AIC:                         4.522e+04
Df Residuals:                    9966   BIC:                         4.546e+04
Df Model:                          33
Covariance Type:            nonrobust
=====================================================================================
```

```
=======
                              coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------------
-------
Intercept                 -17.5238      0.779    -22.507      0.000     -19.050
-15.998
Children                   -0.3664      0.011    -33.887      0.000      -0.388
-0.345
Age                         0.0403      0.001     35.916      0.000       0.038
0.042
Income                  -1.247e-06   8.23e-07     -1.514      0.130   -2.86e-06
3.67e-07
```



```
Outage_sec_perweek          0.0092      0.008      1.175      0.240      -0.006
0.024
Email                       0.0008      0.008      0.098      0.922      -0.014
0.016
Contacts                   -0.0238      0.023     -1.016      0.310      -0.070
0.022
Yearly_equip_failure       -0.0272      0.036     -0.745      0.456      -0.099
0.044
Bandwidth_GB_Year           0.0120   1.25e-05    960.548      0.000       0.012
0.012
MonthlyCharge               0.0182      0.003      6.969      0.000       0.013
0.023
item1_responses             0.0574      0.033      1.726      0.084      -0.008
0.123
item2_fixes                -0.0540      0.031     -1.734      0.083      -0.115
0.007
item3_replacements          0.0249      0.029      0.872      0.383      -0.031
0.081
item4_reliability          -0.0098      0.026     -0.385      0.700      -0.060
0.040
item5_options              -0.0382      0.027     -1.440      0.150      -0.090
0.014
item6_respectfulness       -0.0143      0.027     -0.525      0.599      -0.068
0.039
item7_courteous            -0.0055      0.026     -0.212      0.832      -0.056
0.045
item8_listening            -0.0534      0.025     -2.173      0.030      -0.102
-0.005
Churn_numeric               1.3475      0.068     19.962      0.000       1.215
1.480
Techie_numeric             -0.0532      0.062     -0.854      0.393      -0.175
0.069
Port_modem_numeric         -0.0438      0.046     -0.944      0.345      -0.135
0.047
Tablet_numeric              0.0041      0.051      0.080      0.936      -0.095
0.104
Phone_numeric              -0.0146      0.080     -0.182      0.855      -0.171
0.142
Multiple_numeric            1.3808      0.096     14.312      0.000       1.192
1.570
OnlineSecurity_numeric      0.9984      0.049     20.399      0.000       0.902
1.094
OnlineBackup_numeric        1.4607      0.075     19.513      0.000       1.314
1.607
```

```
DeviceProtection_numeric     1.1716      0.057     20.559      0.000      1.060
1.283
TechSupport_numeric          0.3492      0.058      6.033      0.000      0.236
0.463
StreamingTV_numeric          3.2281      0.119     27.107      0.000      2.995
3.462
StreamingMovies_numeric      3.1268      0.144     21.763      0.000      2.845
3.408
PaperlessBilling_numeric    -0.0760      0.047     -1.612      0.107     -0.168
0.016
InternetService_numeric     -0.3332      0.053     -6.271      0.000     -0.437
-0.229
Contract_numeric            -0.0220      0.034     -0.637      0.524     -0.090
0.046
Gender_numeric               0.6040      0.038     16.091      0.000      0.530
0.678
```

e/Documents/WGU/D208/PA_D208_Code_Task1.ipynb

PA_D208_Code_Task1

```
==============================================================================
Omnibus:                    24779.448   Durbin-Watson:                   1.960
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              999.182
Skew:                          -0.434   Prob(JB):                    1.07e-217
Kurtosis:                       1.718   Cond. No.                     1.70e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spec
ified.
[2] The condition number is large, 1.7e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Initial Multiple Linear Regression Model

y = -17.523751 + (Children * -.366423) + (Age * .040279) + (Income * -.000001) + (Outage_sec_perweek * 0.009168) + (Email * .000752) + (Contacts * -.023848) + (Yearly_equip_failure * -.027181) + (Bandwidth_GB_Year * .011966) + (MonthlyCharge * .018156) + (item1_responses * .057365) + (item2_fixes * -.054017) + (item3_replacements * .024902) + (item4_reliability * -.09846) + (item5_options * -.038197) + (item6_respectfulness *

-.014343) + (item7_curteous * -.005482) + (item8_listening * -.053166) + (Churn_numeric * 1.347461) + (Techie_numeric * -.053401) + (Port_modem_numeric * -.043823) + (Tablet_numeric * 0.004083) + (Phone_numeric * -.014566) + (Multiple_numeric * 1.380790) + (OnlineSecurity_numeric * .998428) + (OnlineBackup_numeric * 1.460700) + (DeviceProtection_numeric * 1.171649) + (TechSupport_numeric * .349225) + (StreamingTV_numeric * 3.228143) + (StreamingMovies_numeric * 3.126839) + (PaperlessBilling_numeric * -.076023) + (InternetService_numeric * -.333164) + (Contract_numeric * -.021971) + (Gender_numeric * .0603957)

Based on an R squared value = .992, 99.2% of the variation is explained by this model. The condition number is large which suggests strong multicollinearity. It appears that we can reduce the number of variables. To do this, we will use Backwards stepwise reduction for regression models and Variance Inflation Factors (VIF) in Python.

2. The statistically based variable selection procedure is backwards stepwise reduction removing all variables whose P value is larger than .05. I will also use Variance Inflation Factor (VIF) to aid in this. When we look at the initial regression model, the variables whose P value are greater than .05 and therefore according the idea of backwards stepwise reduction, were removed from the dataset are:
    - Income
    - Outage_sec_perweek
    - Email
    - Contacts
    - Yearly_equip_failure
    - Item1_responses

- Item2_fixes
- Item3_replacements
- Item4_reliability
- Item5_options
- Item6_respectfulness
- Item7_courteous
- Techie_numeric
- Port_modem_numeric
- Tablet_numeric
- Phone_numeric
- PaperlessBilling_numeric
- Contract_numeric

Next, I performed a VIF and removed all of the variables that had above a 3 VIF to hold the dataset to a strict cutoff. The following variables that were removed from this method are:

- MonthlyCharge
- Multiple_numeric
- StreamingTV_numeric
- StreamingMovies_numeric
- InternetService_numeric

The choice was informed from the following output:

	VIF	variable
0	1129.060742	Intercept

	VIF	variable
1	1.003919	Children
2	1.003478	Age
3	1.003404	Income
4	1.003788	Outage_sec_perweek
5	1.003559	Email
6	1.003144	Contacts
7	1.002883	Yearly_equip_failure
8	1.380259	Bandwidth_GB_Year
9	23.310180	MonthlyCharge
10	2.216297	item1_responses
11	1.934009	item2_fixes
12	1.605377	item3_replacements
13	1.278476	item4_reliability
14	1.375332	item5_options
15	1.482451	item6_respectfulness
16	1.314521	item7_courteous

	VIF	variable
17	1.189466	item8_listening
18	1.652964	Churn_numeric
19	1.008479	Techie_numeric
20	1.001800	Port_modem_numeric
21	1.004704	Tablet_numeric
22	1.005038	Phone_numeric
23	4.307308	Multiple_numeric
24	1.024982	OnlineSecurity_numeric
25	2.583752	OnlineBackup_numeric
26	1.489515	DeviceProtection_numeric
27	1.462672	TechSupport_numeric
28	6.602124	StreamingTV_numeric
29	9.607575	StreamingMovies_numeric
30	1.003549	PaperlessBilling_numeric
31	3.163871	InternetService_numeric
32	1.003407	Contract_numeric

	VIF	variable
33	1.006126	Gender_numeric

The final reduced regression equation will include the continuous variable for Children, Age, Bandwidth_GB_Year, and the categorical variables of Churn_numeric, OnlineBackup_numeric, OnlineSecurity_numeric, DeviceProtection_numeric, Gender_numeric, and item8_listening.

3. The output of the final reduced multiple regression model that includes both categorical and continuous variables is as follows on the next page:

```
Intercept              -10.863561
Children               -0.356664
Age                    0.039634
Bandwidth_GB_Year          0.011727
item8_listening          -0.047615
Churn_numeric            3.344532
OnlineBackup_numeric       0.897706
OnlineSecurity_numeric      0.961640
DeviceProtection_numeric    0.901045
Gender_numeric            0.596107
dtype: float64
                      OLS Regression Results
==============================================================================
Dep. Variable:               Tenure   R-squared:                       0.989
Model:                          OLS   Adj. R-squared:                  0.989
Method:               Least Squares   F-statistic:                 9.809e+04
Date:              Sun, 29 Jan 2023   Prob (F-statistic):               0.00
Time:                      12:31:24   Log-Likelihood:                -24475.
No. Observations:             10000   AIC:                         4.897e+04
Df Residuals:                  9990   BIC:                         4.904e+04
Df Model:                         9
Covariance Type:          nonrobust
============================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept           -10.8636      0.151    -72.128      0.000     -11.159     -10.568
Children             -0.3567      0.013    -27.342      0.000      -0.382      -0.331
```

```
Age                   0.0396     0.001     29.287     0.000      0.037      0.042
Bandwidth_GB_Year          0.0117   1.43e-05    818.463     0.000      0.012      0.012
item8_listening          -0.0476     0.027     -1.750     0.080     -0.101      0.006
Churn_numeric             3.3445     0.071     47.072     0.000      3.205      3.484
OnlineBackup_numeric       0.8977     0.056     15.894     0.000      0.787      1.008
OnlineSecurity_numeric      0.9616     0.058     16.460     0.000      0.847      1.076
DeviceProtection_numeric    0.9010     0.057     15.941     0.000      0.790      1.012
Gender_numeric            0.5961     0.045     13.172     0.000      0.507      0.685
==============================================================================
Omnibus:                  680.102   Durbin-Watson:               1.967
Prob(Omnibus):              0.000   Jarque-Bera (JB):           375.255
Skew:                      -0.323   Prob(JB):                  3.27e-82
Kurtosis:                   2.305   Cond. No.                  2.22e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.22e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

E. Analyze the data set using your reduced multiple regression model by doing the following:

1. **Explain your data analysis process by comparing the initial and reduced multiple regression models, including the following elements:**
   - **the logic of the variable selection technique**
   - **the model evaluation metric**
   - **a residual plot**
2. **Provide the output and *any* calculations of the analysis you performed, including the model's residual error.**

Note: The output should include the predictions from the refined model you used to perform the analysis.

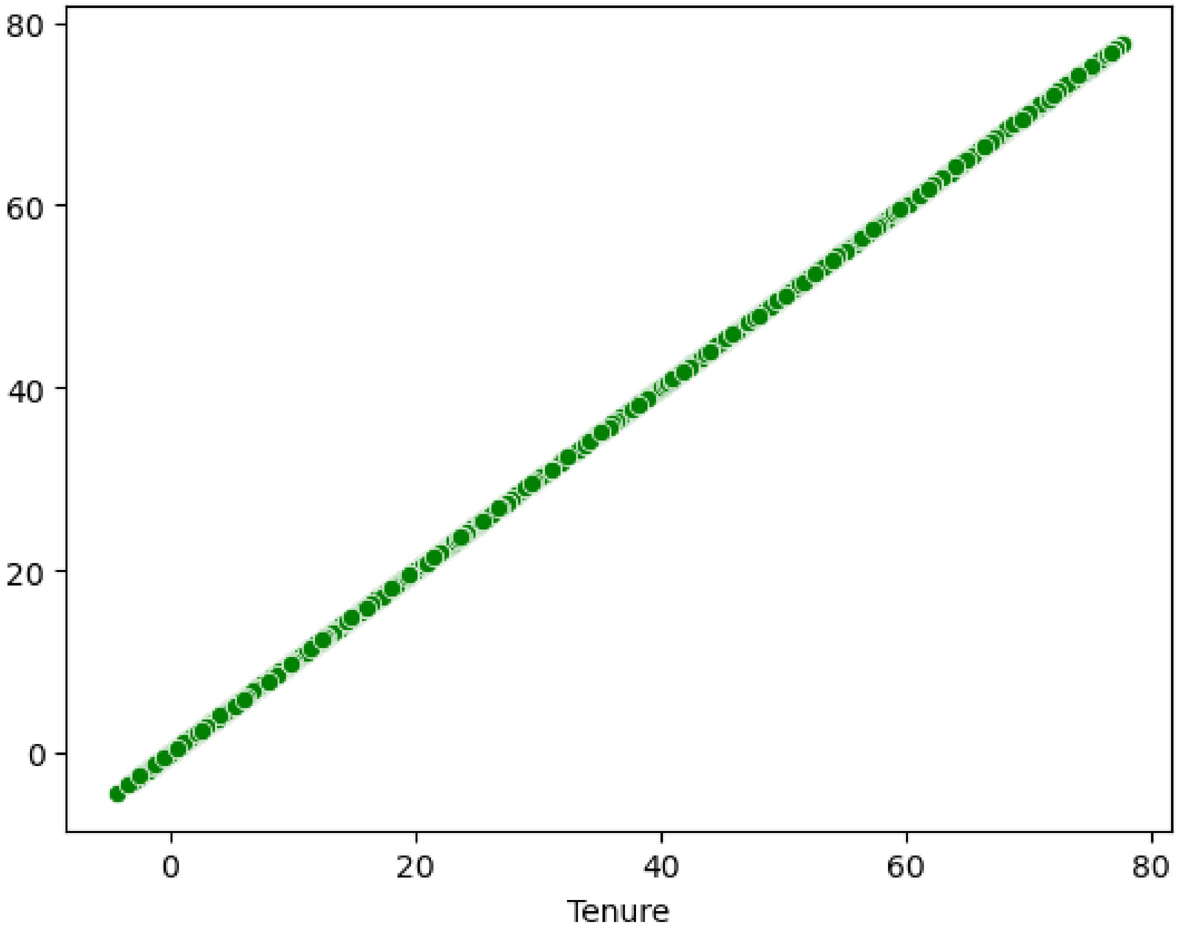
3. Provide the code used to support the implementation of the multiple regression models.

1. The data analysis summary

   a. Logic of the variable selection technique: Tenure is our target variable and the predictor variables initially included were all of the possible variables from the dataset. After running the initial regression, the statistically based variable selection procedure I used to reduce the model was backwards stepwise reduction removing all variables whose P value is larger than .05. I also used Variance Inflation Factor (VIF) to aid in this and further reduce the data.

   b. Model evaluation metric: Backwards Stepwise reduction and VIF. After reduction, the difference is remarkable. With the reduced selection, there was a decrease to .989 R squared value from .992, which indicates further analysis is needed as that is highly irregular. We have reduced variables to a set using valid statistical methods and we have reduced our R squared, indicating there is something going on, more data and analysis is needed, and/or the dataset is dirty in some way we have not detected. The error is still given “The condition number is large, 2.22e+04. This might indicate that there are strong multicollinearity or other numeric problems,”

   c. Residual plot:



2. The output of any and all calculations of the analysis have been placed throughout the report in images, and can be found in the code "PA_D208_Code_Task1" or "PA_D208_Code_Task1_Backup."

    The R value for our final model is .989, this suggests a 98.9% strength to predict. The R squared value being .989 is very high and a good output by itself, but as described before, the decrease we have seen is indicative of a need for more analysis.

    The estimated standard residual error of the reduced model is 2.798474402739518. This means our prediction model, on average, is off by that amount.

```
# Here we are calculating our residual error from the reduced model

print(np.sqrt(LM_Reduced_Tenure.mse_resid))
```

```
2.798474402739518
```

3. The code used to support implementation of the multiple regression models is found in "PA_D208_Code_Task1" and annotated to the respective part of this report.

Part V: Data Summary and Implications

F. Summarize your findings and assumptions by doing the following:

1. **Discuss the results of your data analysis, including the following elements:**
   - **a regression equation for the reduced model**
   - **an interpretation of coefficients of the statistically significant variables of the model**
   - **the statistical and practical significance of the model**
   - **the limitations of the data analysis**
2. **Recommend a course of action based on your results.**

1. The results of the analysis:
   a. Regression equation for the reduced model: y = -10.863561 + (-.356664 * Children) + (0.039634 * Age) + (0.011727 * Bandwidth_GB_Year) + (-.047615 * item8_listening) + (3.344532 * Churn_numeric) + (0.897706 * OnlineBackup_numeric) + (.961640 * OnlineSecurity_numeric) + (.901045 * DeviceProtection_numeric) + (.596107 * Gender_numeric)

b. Interpretation of coefficients of the statistically significant variables of the previous model. The coefficients suggest that per each unit of the following, Tenure will increase or decrease by the following per one unit of the variable.

- Children, Tenure will decrease 0.356664
- Age, Tenure will increase 0.039634
- Bandwidth_GB_Year, Tenure will increase 0.011727
- item8_listening, Tenure will decrease 0.047615
- Churn_numeric, Tenure will increase 3.344532
- OnlineBackup_numeric, Tenure will increase 0.897706
- OnlineSecurity_numeric, Tenure will increase 0.961640
- DeviceProtection_numeric, Tenure will increase 0.901045
- Gender_numeric, Tenure will increase 0.596107

c. Statistical and significance of the model: As described, the condition number is large, indicating there might be strong multicollinearity or other numerical problems. The P value for all of the variables in the reduced model are statistically significant at 0.00. The R value is .989, indicating a strong prediction strength, but again, has lowered after reduction, indicating further analysis and reevaluation of the data acquisition and cleaning phase of this dataset is needed. Finally, with standard residual error of 2.798474402739518 our model is fairly accurate, but whether it is acceptable error or not is up to the organization based off of their specific needs.

d. Limitations of data analysis: There is not enough customer data to clearly identify trends. The data analysis indicates strong multicollinearity or other numerical problems, indicating the dataset is dirty in some way not tested for. More data analysis beyond a multiple linear regression model is needed and the model indicates a need to reevaluate the data acquisition and cleaning phase for this organization. Also, the R squared value lowering after a reduction of non-correlated variables indicates further analysis is needed and there is more going on with the data than meets the eye. Finally, the residual error of 2.798474402739518 could be a limitation of the model's output if the organization deems it too large.

2. The course of action I recommend is based on the strong linear relationship between Tenure and Churn. The organization should firstly begin an analysis into why Tenure and Churning are related, why do customers at certain linear Tenure lengths leave the organization within the last month. Is there a way to give the customers incentives or contacts to prevent churning and specific Tenure lengths? The organization should also investigate the Bandwidth_GB_Year and item8_listening in relation to what they can do to increase Tenure with their service model. Also, the organization can use the demographic variables found in our reduced model to predict the tenure length of customers and adjust their business model to increase the predicted Tenure of these customers. Finally, due to the limitations and the suspicious R value, I would recommend further data collection to get a larger data set and to perform more analysis on the topic.

Part VI: Demonstration

G. Provide a Panopto video recording that includes *all* of the following elements:

- **a demonstration of the functionality of the code used for the analysis**
- **an identification of the version of the programming environment**
- **a comparison of the two multiple regression models you used in your analysis**

- **an interpretation of the coefficients.**

Link to the panopto presentation:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=70b8f409-c65a-434e-8548-af9b0153973f

H. List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable.

The source to inform my code regarding the multiple regression model was the following:

Linear Regression: Residual Standard Error in Python – Data Science Concepts. (n.d.). Retrieved January 31, 2023, from https://www.datascienceconcepts.com/tutorials/python-programming-language/linear-regression-residual-standard-error-in-python/

Yadav, H. (2021, May 8). *Multiple Linear Regression Implementation in Python*. Machine Learning with Python. https://medium.com/machine-learning-with-python/multiple-linear-regression-implementation-in-python-2de9b303fc0c

Zach. (2020, July 20). *How to Calculate VIF in Python*. Statology. https://www.statology.org/how-to-calculate-vif-in-python/

I. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

The sources to inform my python choice, multiple regression, and VIF information are as follows:

Prasanna, M. (2021, October 21). *How to Efficiently Handle Large Datasets for Machine Learning and Data Analysis Using Python*. Medium. https://python.plainenglish.io/working-with-large-datasets-for-machine-learning-d8da0dd802fb

Yadav, H. (2021, May 8). *Multiple Linear Regression Implementation in Python*. Machine Learning with Python. https://medium.com/machine-learning-with-python/multiple-linear-regression-implementation-in-python-2de9b303fc0c

J. Demonstrate professional communication in the content and presentation of your submission.

This aspect cannot be summarized; however, I hope it has shown through in all aspects of this report.