Task 1: Linear Regression Modeling – D208

Abhishek Aern

Western Governor University

Contents

Part I - Research Question

A1: Question for analysis

What all factors influence customer tenure in the telecom company?

Can we develop the model to predict tenure in terms of how many months somebody will stay with the telecom company based on its relationship with the dependent variable?

A2: Goal of analysis

The goal is to identify the factors that contribute to customers staying with a telecom company over some time. This research question can help telecom company stakeholders to understand the drivers of customer loyalty and develop strategies to improve customer retention.

Understanding these factors can help companies develop targeted interventions to address issues related to customer tenure, such as improving customer service, enhancing the quality of their products and services, or developing innovative loyalty programs that increase customer satisfaction.

Part II - Method Justification

B1: Summary of Assumptions

Here is a list of the assumptions of the multiple linear regression model (avcontentteam, 2016):

1. The relationship between the dependent variable and the independent variables should be linear.
2. The absence of multicollinearity is expected in the model, meaning that independent variables are not too highly correlated.
3. The observations are selected independently and randomly from the population.
4. The variance of the errors should be constant across all values of the independent variables. This means that the variability of the residuals should be the same for all levels of the independent variables.

5. The errors should be normally distributed. This means that the residuals should have a bell-shaped curve when plotted on a histogram.

B2: Tool Benefits

I am using Python and Jupiter notebook for my analysis. Python provides a flexible and powerful platform for regression analysis across various phases. It provides several libraries like Scikit-learn, Statsmodels, and TensorFlow that offer powerful tools for building regression models. These libraries enable users to build, train, and test multiple linear regression models, as well as evaluate their performance.

(Jain, 2017) Python provides libraries like Scikit-learn that offer powerful tools for feature selection and engineering. These tools can help data analysts identify the most important features that have the most impact on the dependent variable and discard irrelevant features. The libraries like Statsmodels offer powerful tools for interpreting regression models. These tools help data analysts to understand the relationships between variables and the effect of each variable on the dependent variable.

B3: Appropriate Technique

(The Ultimate Guide to Linear Regression) Multiple linear regression allows for the analysis of the relationship between multiple independent variables and a single dependent variable, which is ideal for understanding the factors that influence customer tenure. Telecom companies typically have multiple factors that could impact customer tenure and multiple linear regression provides a way to quantify the strength of the relationship between the independent variables and the dependent variable through the use of regression coefficients and R-squared values.

Multiple linear regression also provides a way to test the significance of the independent variables in explaining the variation in the dependent variable through hypothesis testing. This allows for the identification of the most important factors that influence customer tenure and the rejection of factors that do not have a significant impact on customer tenure (The Ultimate Guide to Linear Regression).

Part III - Data Preparation

The data preparation process for multiple linear regression analysis involves careful cleaning and selection of variables, data transformation, analysis of multicollinearity, creation of dummy variables, and data splitting to ensure accurate and reliable results.

C1: Data Cleaning

The data preparation process for multiple linear regression analysis involves several steps. My approach includes -

Data cleaning:

- Drop the variables which are not needed to answer the research question.
- Renames the last 8 survey columns for a better description of variables.
- Examine the Null values using isnull().any() method. No null value is present.
- Examine missing values using isna().any() method. No missing value was found.
- Examine the outliers using histograms and z-score.

Variable selection:

To answer the research question Tenure will be my dependent variable and after dropping many variables initially and based on the VIF score, I have the following features left as an independent variable to help answer my research question.

Children, Age, Income, Outage_sec_perweek, Contacts, Yearly_equip_failure, Mul tiple, PaperlessBilling, Bandwidth_GB_Year, Reliability, Options, Contract and Int ernetService.

Data Wrangling (Dummy variable creation) :

I am using 4 categorical variables (Contract, InternetService, Multiple, and PaperlessBilling) that needs to be converted for use in the multiple linear regression model.

The Multiple and PaperlessBilling have Yes/No values so they can be converted to 1/0 using Label encoding. The Contract variable is an ordinal categorical variable so using LabelEncoder class from the sklearn.preprocessing module. The InternetS ervice variable is a nominal categorical variable so one hot encoding is used to crea te the dummy variable for each category of internet service. There are 3 values pres ent so 3 dummy variables are created but I am dropping 1 variable as regression needs a K-1 dummy variable so 2 in this case.

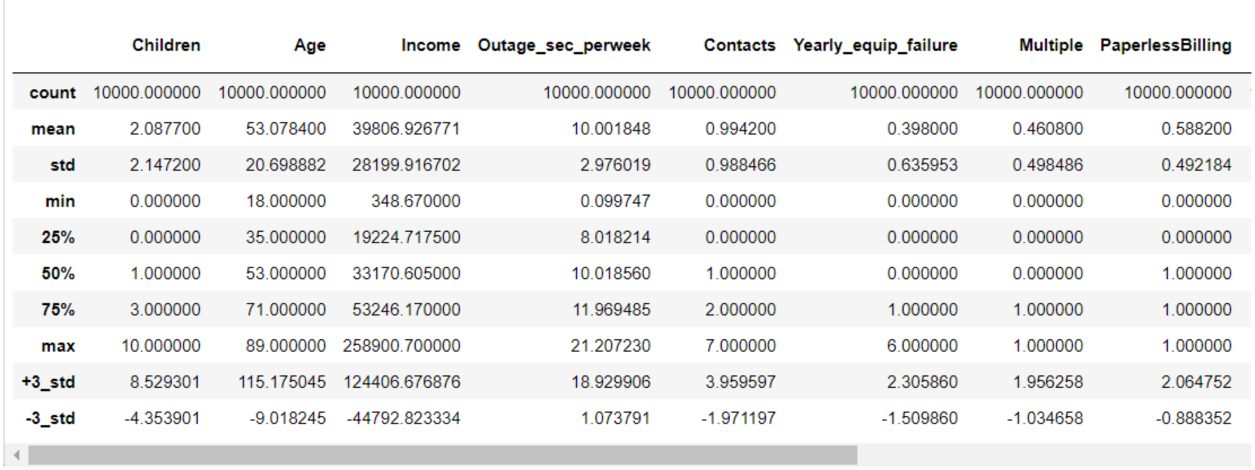
Multicollinearity analysis:

The first thing I created is a correlation matrix using the corr() function; this will create a matrix with each variable having its correlation calculated for all the other variables. Also plotted the correlation heatmap to better visualize the correlation between independent variables.

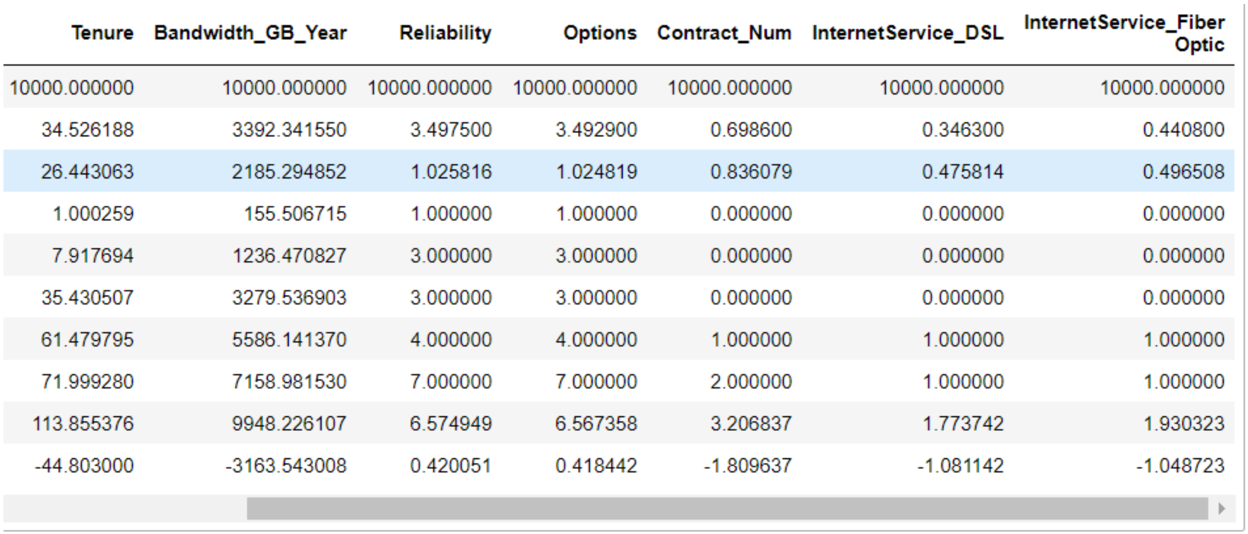
As the number of variables is high so sometimes it is hard to find the correlated variables, so I am using a more systematic approach to calculate the variance inflation factor(VIF) to find if significant multicollinearity exists and drop the variables if VIF score > 10.

C2: Summary Statistics

The summary statistics for the dependent variable and all independent variables are created using describe() method. I am adding a few more metrics to the summary data frame so that it now includes a metric for three standard deviations below and above the mean.



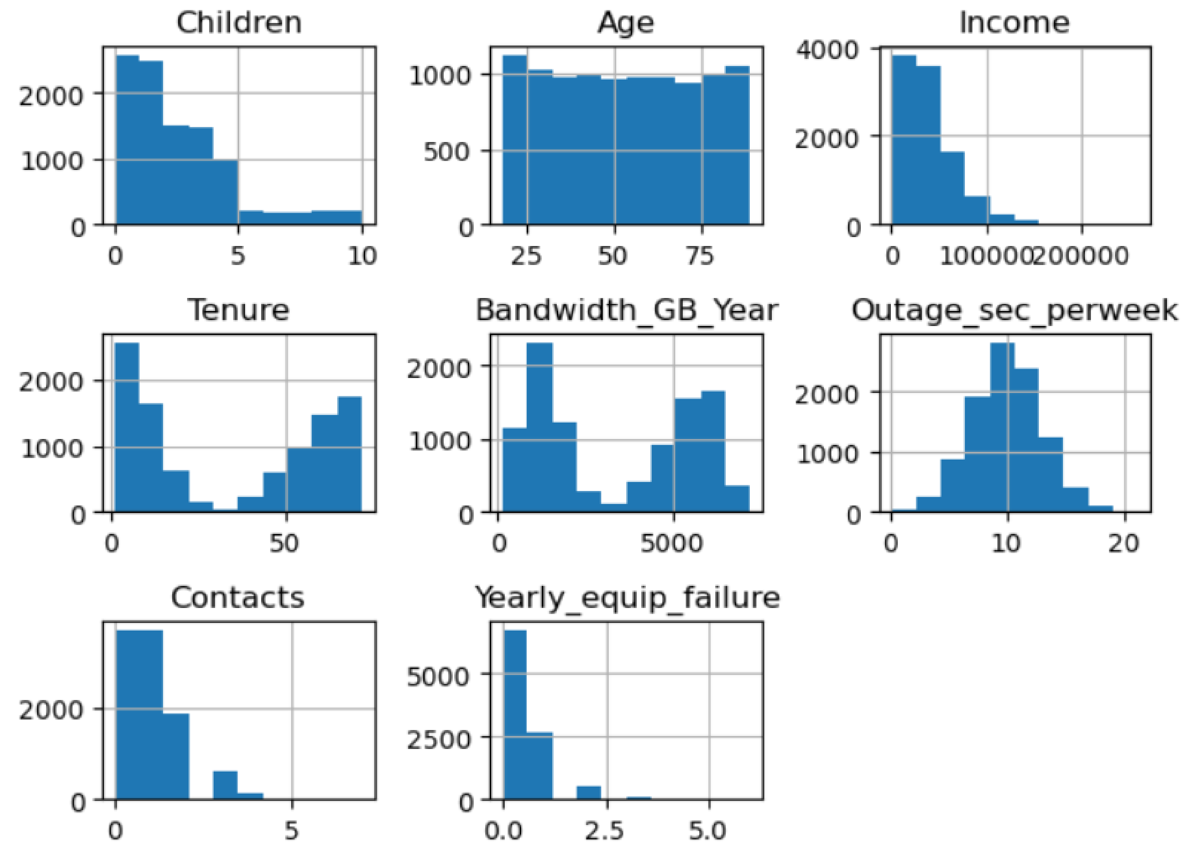
	Children	Age	Income	Outage_sec_perweek	Contacts	Yearly_equip_failure	Multiple	PaperlessBilling
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	2.087700	53.078400	39806.926771	10.001848	0.994200	0.398000	0.460800	0.588200
std	2.147200	20.698882	28199.916702	2.976019	0.988466	0.635953	0.498486	0.492184
min	0.000000	18.000000	348.670000	0.099747	0.000000	0.000000	0.000000	0.000000
25%	0.000000	35.000000	19224.717500	8.018214	0.000000	0.000000	0.000000	0.000000
50%	1.000000	53.000000	33170.605000	10.018560	1.000000	0.000000	0.000000	1.000000
75%	3.000000	71.000000	53246.170000	11.969485	2.000000	1.000000	1.000000	1.000000
max	10.000000	89.000000	258900.700000	21.207230	7.000000	6.000000	1.000000	1.000000
+3_std	8.529301	115.175045	124406.676876	18.929906	3.959597	2.305860	1.956258	2.064752
-3_std	-4.353901	-9.018245	-44792.823334	1.073791	-1.971197	-1.509860	-1.034658	-0.888352



Tenure	Bandwidth_GB_Year	Reliability	Options	Contract_Num	InternetService_DSL	InternetService_Fiber Optic
10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
34.526188	3392.341550	3.497500	3.492900	0.698600	0.346300	0.440800
26.443063	2185.294852	1.025816	1.024819	0.836079	0.475814	0.496508
1.000259	155.506715	1.000000	1.000000	0.000000	0.000000	0.000000
7.917694	1236.470827	3.000000	3.000000	0.000000	0.000000	0.000000
35.430507	3279.536903	3.000000	3.000000	0.000000	0.000000	0.000000
61.479795	5586.141370	4.000000	4.000000	1.000000	1.000000	1.000000
71.999280	7158.981530	7.000000	7.000000	2.000000	1.000000	1.000000
113.855376	9948.226107	6.574949	6.567358	3.206837	1.773742	1.930323
-44.803000	-3163.543008	0.420051	0.418442	-1.809637	-1.081142	-1.048723

C3: Visualizations

Univariate visualizations



Observation using the Histograms :

Children – Almost 75% of customers have children between 0-3.

Age – The age group is very uniformly distributed.

Income – More than 70% of customer Income is less than 55K.

Tenure – This is a dependent variable. 75% of the customer's tenure is less than 62 months.

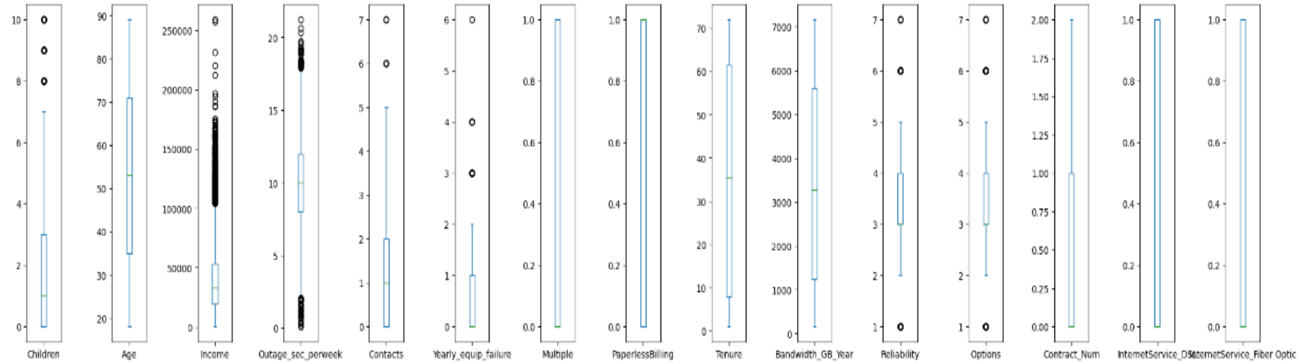
Bandwidth_GB_Year – Around 70% of customers are within the 5000GB range.

Outage_sec_perweek – This is a normal distribution. A mean outage is 10 sec per week.

Contacts – 75 % of customers contacted 2 times or less.

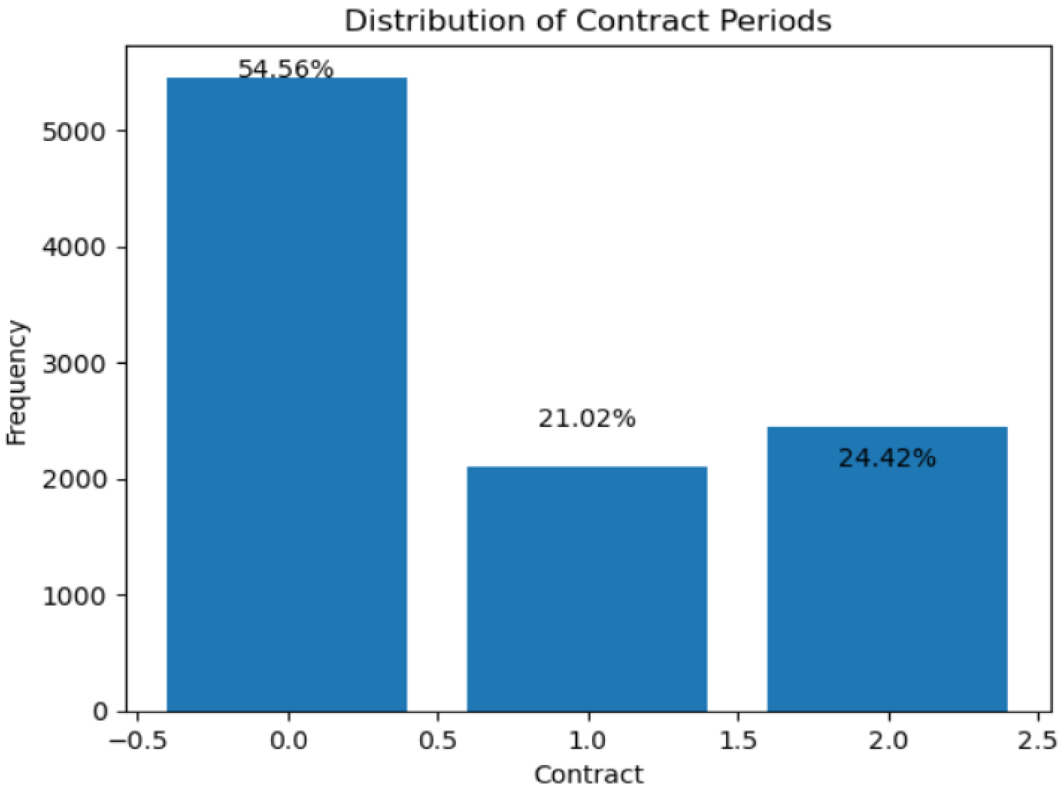
Yearly_equip_failure – Almost 75% of customers are having only 1 times equipment failure in the whole year. Very few customers reported a higher number of failures.

Boxplot: Boxplot is showing some of the outliers for Children, Income, Outage_sec_perweek, Email, Contacts, and all survey variables. I am not going to delete any of the entries as it doesn't look like error records to me.

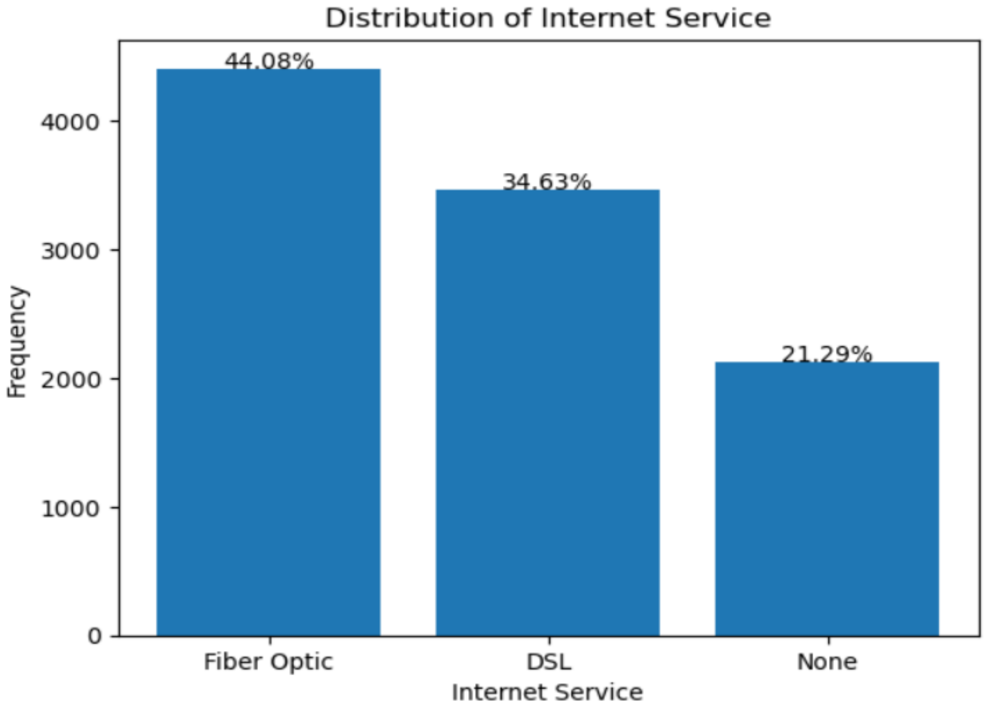


Categorical variable distribution :

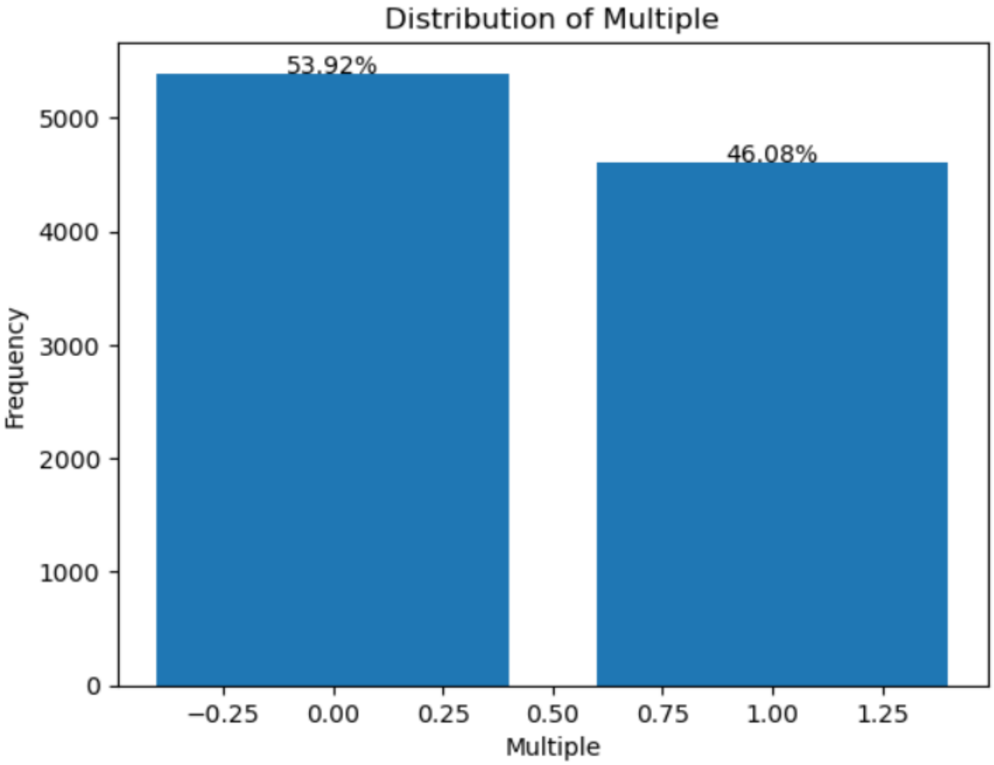
Contract type – Most of the customers seem to have a month-to-month contract with the telecom company. Also, there are more or less equal proportions of customers for 1-year and 2-year contract types.



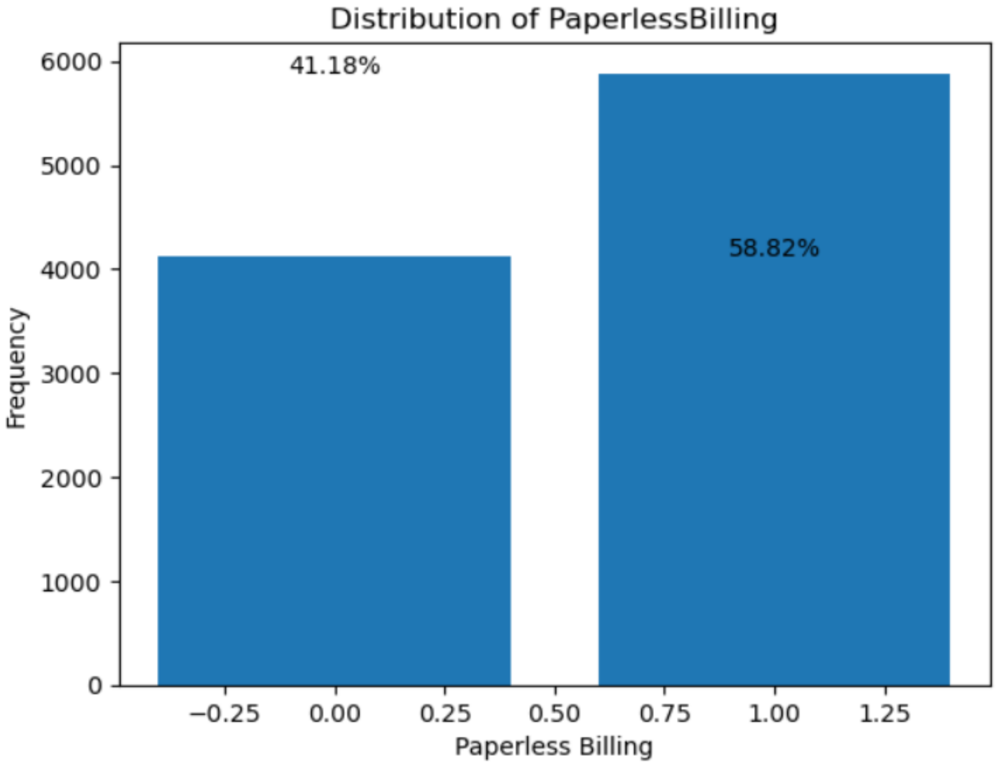
Almost half of the customers have opted for internet service with Fiber Optic(44%) and 35% are with DSL.



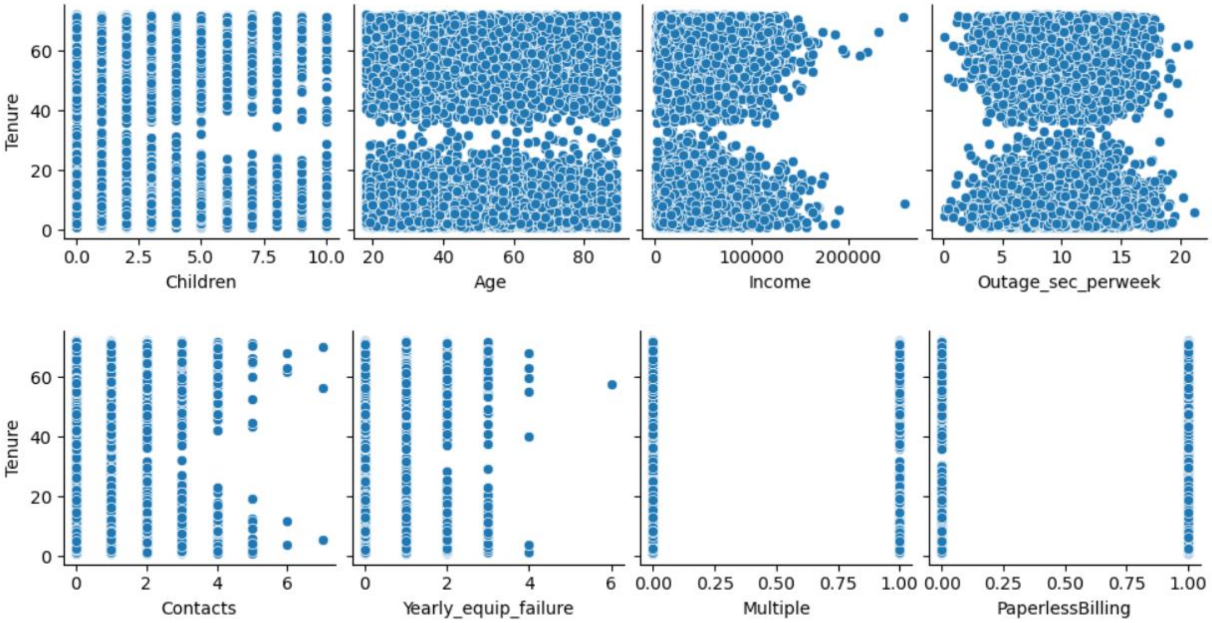
Multiple - More than 50% of customers did not opt for multiple services.

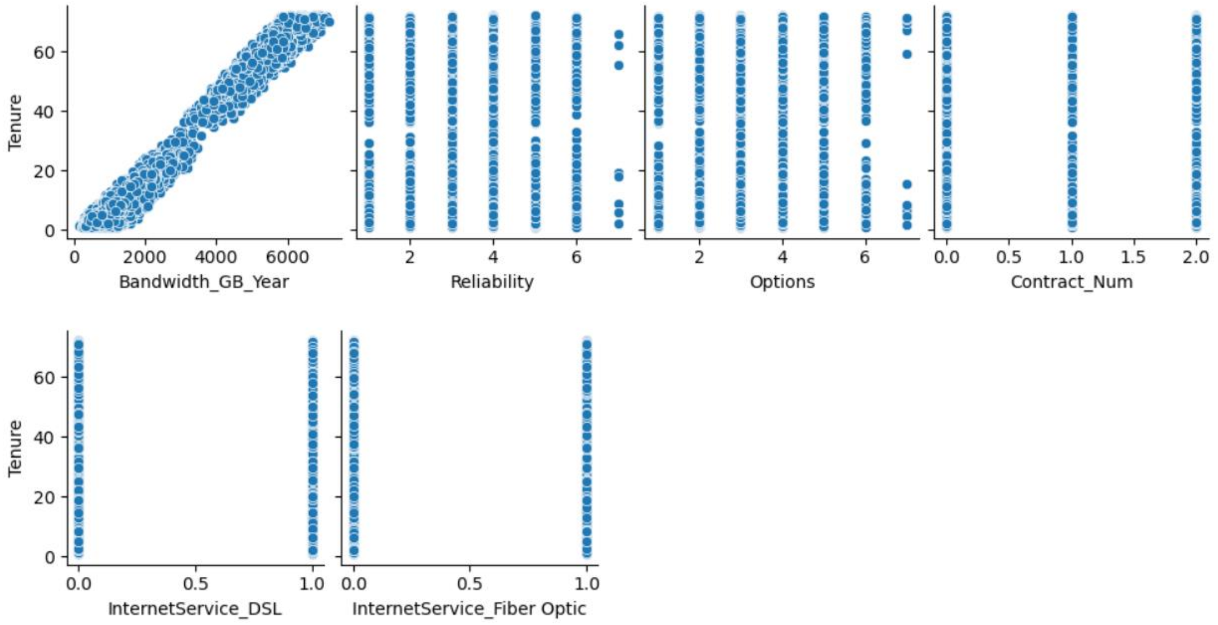


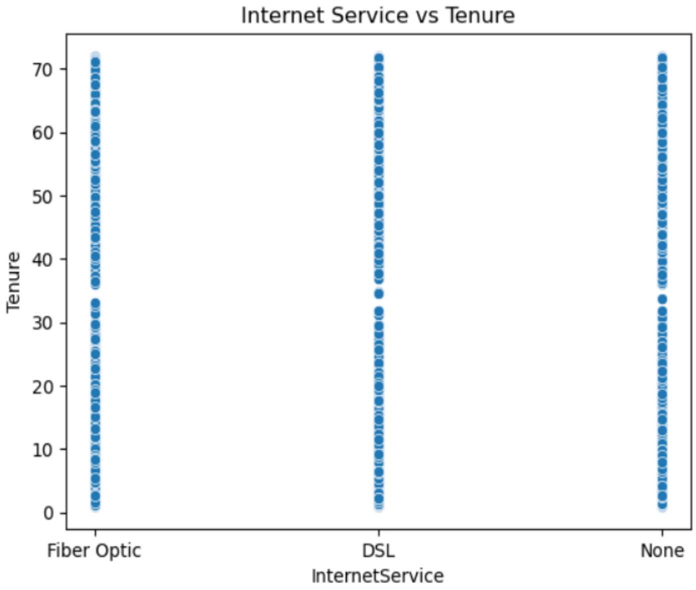
Around 58% of customers are using paperless billing.



Bivariate visualizations: All independent variables with dependent variable Tenure







All the variables are distributed normally for tenure. The Bandwidth_GB_Year and tenure are co-related to each other.

C4: Data Transformation

As described in the C1 section, the 4 categorical variables Contract, InternetService, Multiple, and PaperlessBilling are converted to numerical so that they can be used in the multiple linear regression model.

After creating the dummy variable for InternetService the output data type was uint8 so that was converted to int format using the applymap () method to apply the astype() function to each element in the DataFrame.

Some of the variables are dropped while checking for multicollinearity using VIF(variance_inflation_factor). The variance_inflation_factor() function is applied to each independent variable in the dataframe to calculate its corresponding VIF value and the values are stored in a new dataframe vif_data and displayed using the print() function.

A VIF value of 1 indicates that there is no multicollinearity, while values above 1 indicate increasing levels of multicollinearity. A commonly used threshold for detecting multicollinearity is a VIF value of 5 or above. I am dropping the following variables where VIF > 10.

The following variables are dropped using this method –

'MonthlyCharge', 'Timely_Response', 'Timely_Fixes', 'Respectfulness', 'Timely_Replacements', 'Courteous_Exchange', 'Active_Listening'

C5: Prepared Dataset

The prepared clean data is provided in a CSV file.

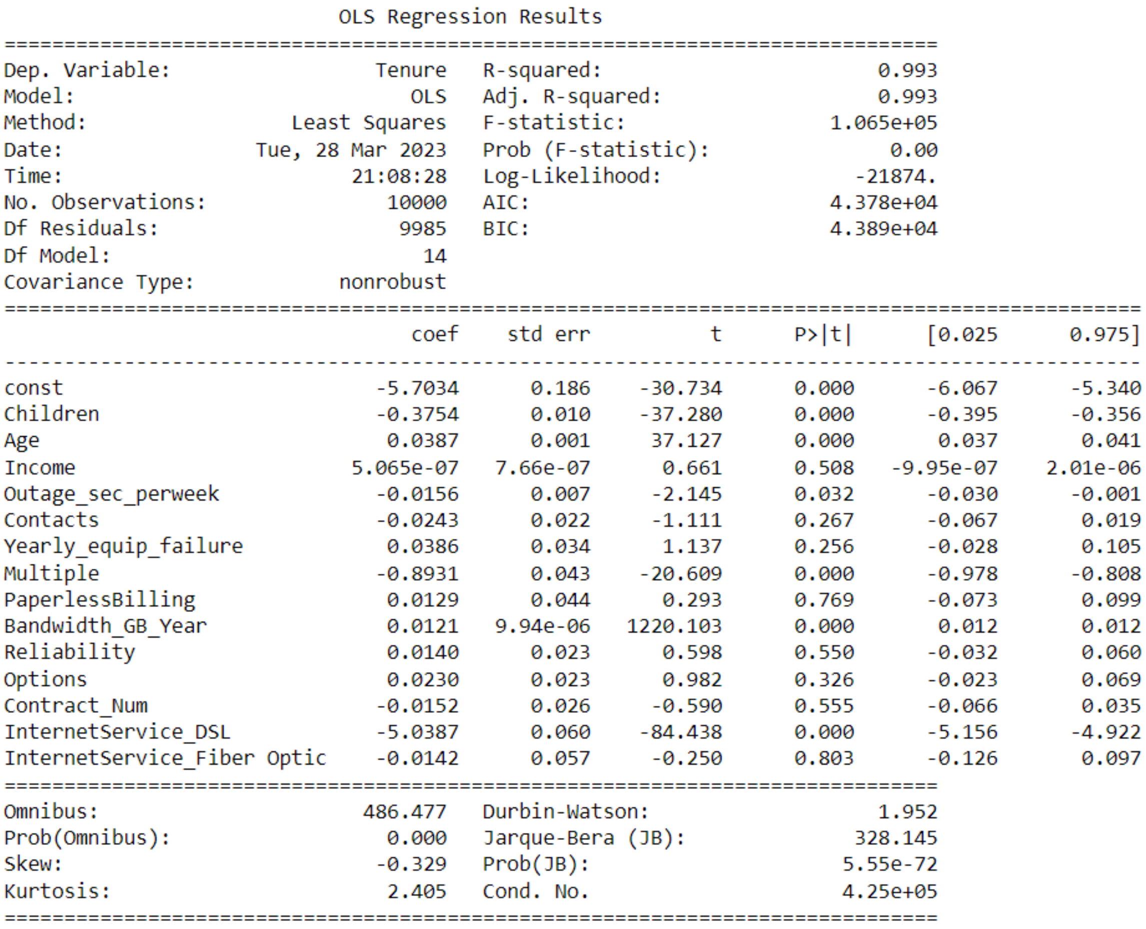
Part IV - Model Comparison and Analysis

(Bevans, 2020) I am using the Statsmodels which is a powerful package for regression analysis and provides a wide range of statistical tests and diagnostic tools such as p-values, confidence intervals, hypothesis tests, and goodness-of-fit measures.

I am creating the constant variable to the independent variable X, then creating and fitting the OLS model using the fit() method.

D1: Initial Model

Initial Model summary -



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Tenure   R-squared:                       0.993
Model:                            OLS   Adj. R-squared:                  0.993
Method:                 Least Squares   F-statistic:                 1.065e+05
Date:                Tue, 28 Mar 2023   Prob (F-statistic):               0.00
Time:                        21:08:28   Log-Likelihood:                -21874.
No. Observations:               10000   AIC:                         4.378e+04
Df Residuals:                    9985   BIC:                         4.389e+04
Df Model:                          14
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                          -5.7034      0.186    -30.734      0.000      -6.067      -5.340
Children                       -0.3754      0.010    -37.280      0.000      -0.395      -0.356
Age                             0.0387      0.001     37.127      0.000       0.037       0.041
Income                       5.065e-07   7.66e-07      0.661      0.508   -9.95e-07    2.01e-06
Outage_sec_perweek             -0.0156      0.007     -2.145      0.032      -0.030      -0.001
Contacts                       -0.0243      0.022     -1.111      0.267      -0.067       0.019
Yearly_equip_failure            0.0386      0.034      1.137      0.256      -0.028       0.105
Multiple                       -0.8931      0.043    -20.609      0.000      -0.978      -0.808
PaperlessBilling                0.0129      0.044      0.293      0.769      -0.073       0.099
Bandwidth_GB_Year               0.0121   9.94e-06   1220.103      0.000       0.012       0.012
Reliability                     0.0140      0.023      0.598      0.550      -0.032       0.060
Options                         0.0230      0.023      0.982      0.326      -0.023       0.069
Contract_Num                   -0.0152      0.026     -0.590      0.555      -0.066       0.035
InternetService_DSL            -5.0387      0.060    -84.438      0.000      -5.156      -4.922
InternetService_Fiber Optic    -0.0142      0.057     -0.250      0.803      -0.126       0.097
==============================================================================
Omnibus:                      486.477   Durbin-Watson:                   1.952
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              328.145
Skew:                          -0.329   Prob(JB):                     5.55e-72
Kurtosis:                       2.405   Cond. No.                     4.25e+05
==============================================================================
```

This summary includes important information about the regression model, including the coefficients, standard errors, p-values, and R-squared values.

Observations :

(1) Adj. R-squared: 0.993 (adjusted for the number of variables in the regression)

(2) Prob (F-statistic): 0.00 (provide overall significance of the model)

(3) The p-values of variables Income, Contacts, PaperlessBilling, Reliability, Options, Yearly_equip_failure, Contract_Num, and InternetService_Fiber Optic are greater than 0.05 which indicates that this is not significant to the dependent variable Tenure.

(4) The coefficient estimates are showing how these features influence the output(dependent) variable and the sign indicates the direction positive or negative.

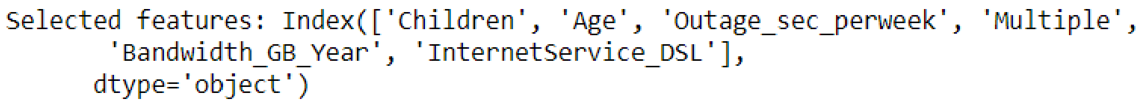
D2: Justification of Model Reduction

(Verma, 2020) I am using the Backward stepwise elimination feature selection method that is commonly used in multiple linear regression. In backward stepwise elimination, we start with all the features in the model and iteratively remove the least significant feature until a stopping criterion is met. I am using scikit-learn for this.

I am using a while loop for backward stepwise elimination. In each iteration of the loop, we get the indices of the selected features then fit a linear regression model using only those features and get the p-values of the coefficients. After that find the index of the feature with the highest p-value (excluding the intercept), and if the p-value is greater than 0.05, remove that feature from the support array. This process is continued until all the remaining features have p-values less than or equal to 0.05.

Finally, printing the names of the selected features using the columns attribute of the X DataFrame and the support array (Verma, 2020).

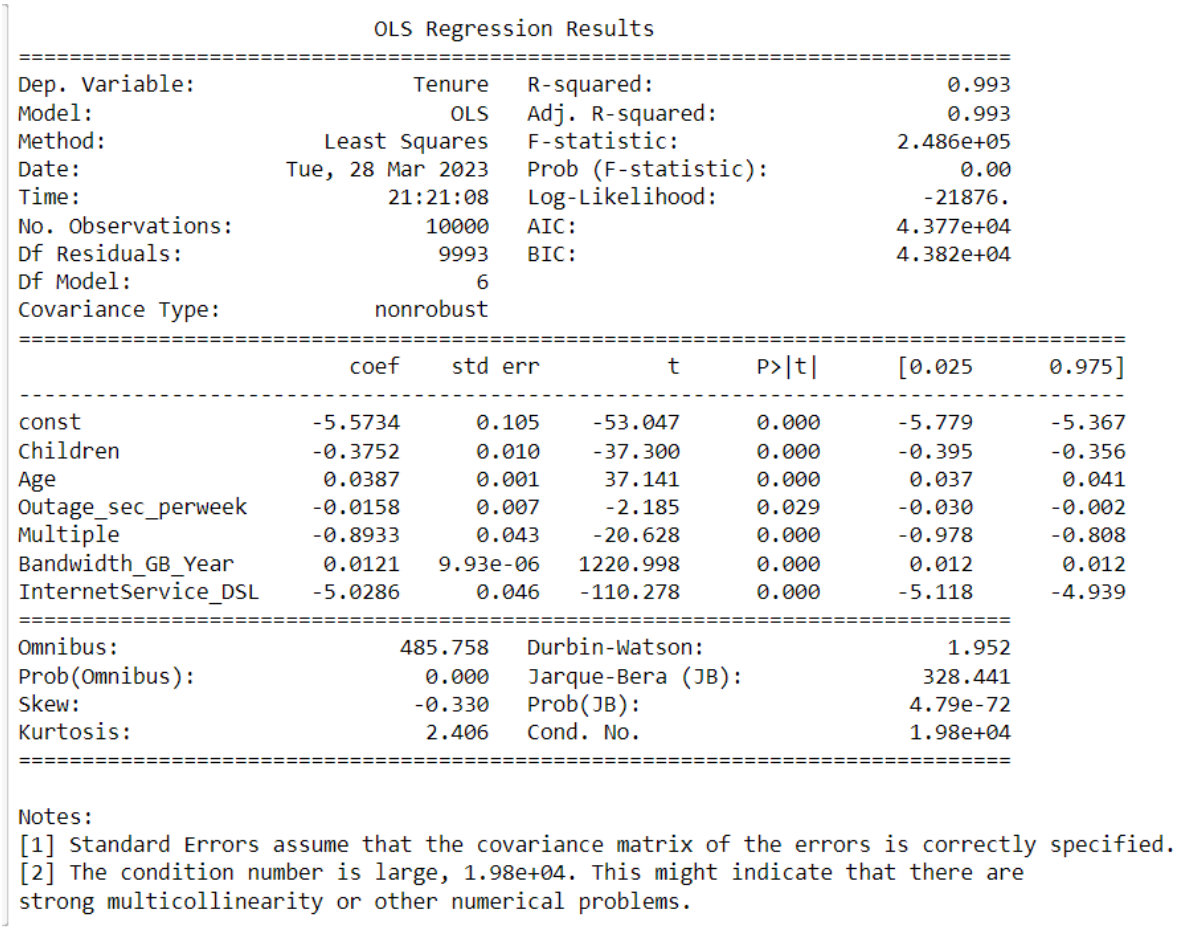
The following features are selected after Backward stepwise elimination –



```
Selected features: Index(['Children', 'Age', 'Outage_sec_perweek', 'Multiple',
       'Bandwidth_GB_Year', 'InternetService_DSL'],
      dtype='object')
```

D3: Reduced Linear Regression Model

Reduced linear regression model -



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Tenure   R-squared:                       0.993
Model:                            OLS   Adj. R-squared:                  0.993
Method:                 Least Squares   F-statistic:                 2.486e+05
Date:                Tue, 28 Mar 2023   Prob (F-statistic):               0.00
Time:                        21:21:08   Log-Likelihood:                -21876.
No. Observations:               10000   AIC:                         4.377e+04
Df Residuals:                    9993   BIC:                         4.382e+04
Df Model:                           6
Covariance Type:            nonrobust
=========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                  -5.5734      0.105    -53.047      0.000      -5.779      -5.367
Children               -0.3752      0.010    -37.300      0.000      -0.395      -0.356
Age                     0.0387      0.001     37.141      0.000       0.037       0.041
Outage_sec_perweek     -0.0158      0.007     -2.185      0.029      -0.030      -0.002
Multiple               -0.8933      0.043    -20.628      0.000      -0.978      -0.808
Bandwidth_GB_Year       0.0121   9.93e-06   1220.998      0.000       0.012       0.012
InternetService_DSL    -5.0286      0.046   -110.278      0.000      -5.118      -4.939
==============================================================================
Omnibus:                      485.758   Durbin-Watson:                   1.952
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              328.441
Skew:                          -0.330   Prob(JB):                     4.79e-72
Kurtosis:                       2.406   Cond. No.                     1.98e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.98e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Reduced model matrix :

(1) Adj. R-squared: 0.993 This is not decreased after removing many other variables !! This indicates that 99.3 % of the variance in the tenure can be explained by the number of children, customer age, outage sec per week, bandwidth used per year, whether the customer has multiple services or not, and which service providers they are with.

(2) Prob (F-statistic): 0.00 (provide overall significance of the model)

E1: Model Comparison

(Zach, How to Read and Interpret a Regression Table, 2019) When comparing the initial multiple linear regression model and the reduced linear regression model, there are several factors to consider :

The reduced model should have a similar or better performance than the initial model in terms of metrics such as R-squared, adjusted R-squared, and root mean squared error. My initial model has R-squared, adjusted R-squared is 0.993 and root mean squared error is 2.1581009146776284. The R-squared, and adjusted R-squared are the same in my reduced model, and the root mean squared error is reduced to 2.157720865373507. This is very less decrement, but reduction is always better when compared to the initial model.

The reduced model is including only the most important variables which are more interpretable and easier to explain than the initial model. Also, the reduced model with fewer variables is less complex than the initial model, which can make it easier to work with and less prone to overfitting.

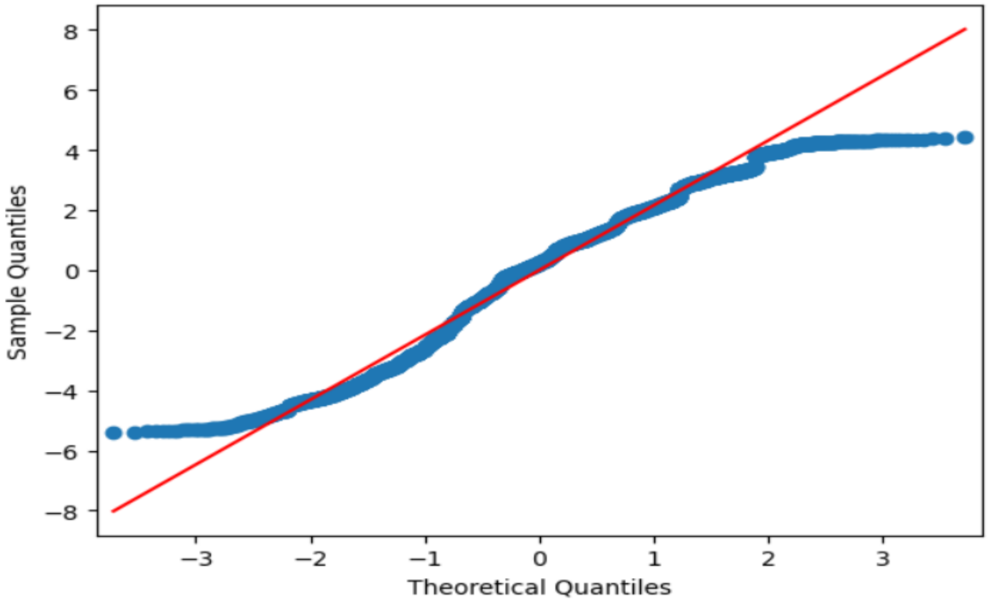
Prob (F-statistic) provide the collective significance of all the variable to the dependent variable. This is 0.00 in my initial and reduced model which indicates that the model is statically significant. (Zach, How to Read and Interpret a Regression Table, 2019)

E2: Output and Calculations

Residual plot of the reduced model –

I am using the QQ pplot method. This would give us the confidence in our assumption that the residuals are normally distributed.

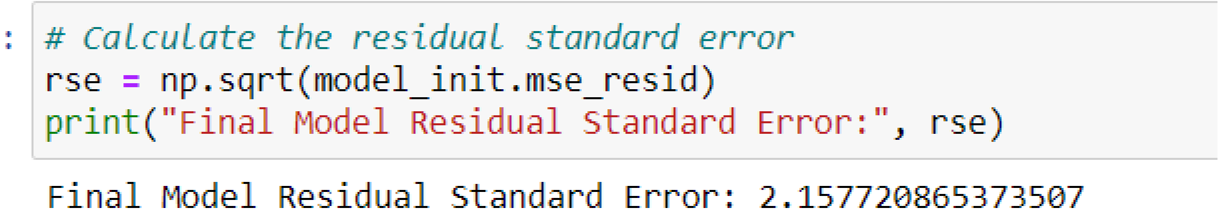
Additionally, I am checking another assumption that the mean of the residuals should be equal to zero. In my case mean this is very close to zero(-0.0000000000006298).



```
The mean of the residuals is -6.298e-13
```

This Q-Q plot shows a deviation from normality at the beginning and end which means the normality assumption is violated and the residuals are not normally distributed. This indicates that the model may not be capturing all the relevant information in the data, and further investigation is needed.

Reduced model residual standard error -

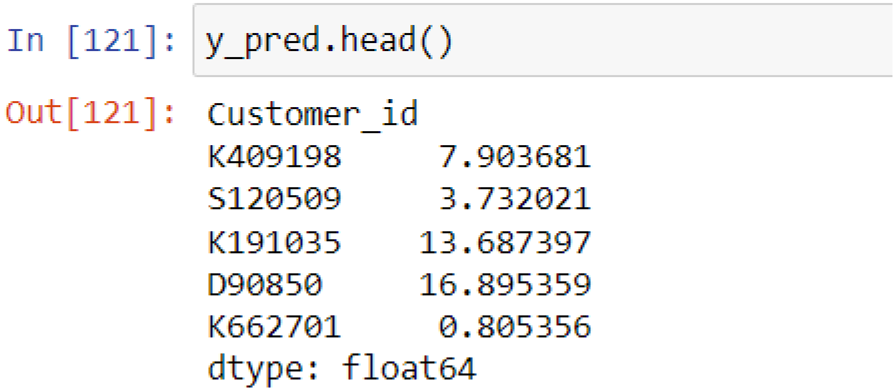


```
# Calculate the residual standard error
rse = np.sqrt(model_init.mse_resid)
print("Final Model Residual Standard Error:", rse)
```

```
Final Model Residual Standard Error: 2.157720865373507
```

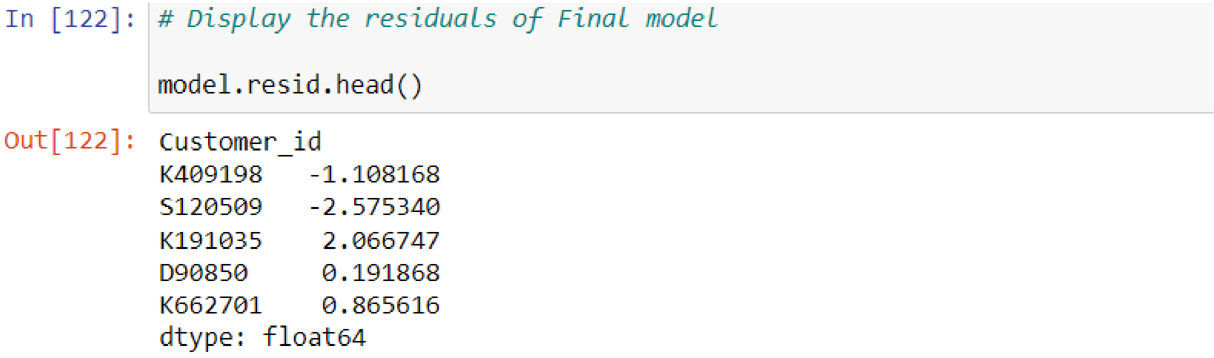
A smaller RSE indicates a better fit of the model to the data, as it means that the residuals are smaller in magnitude and the predicted values are closer to the actual values. My final RSE is not very high which indicates a good overall fit of a regression model.

The predicted value of the y :



```
In [121]: y_pred.head()

Out[121]: Customer_id
          K409198     7.903681
          S120509     3.732021
          K191035    13.687397
          D90850     16.895359
          K662701     0.805356
          dtype: float64
```

Residuals of Final model :



```
In [122]: # Display the residuals of Final model

          model.resid.head()

Out[122]: Customer_id
          K409198   -1.108168
          S120509   -2.575340
          K191035    2.066747
          D90850     0.191868
          K662701    0.865616
          dtype: float64
```

E3: Code

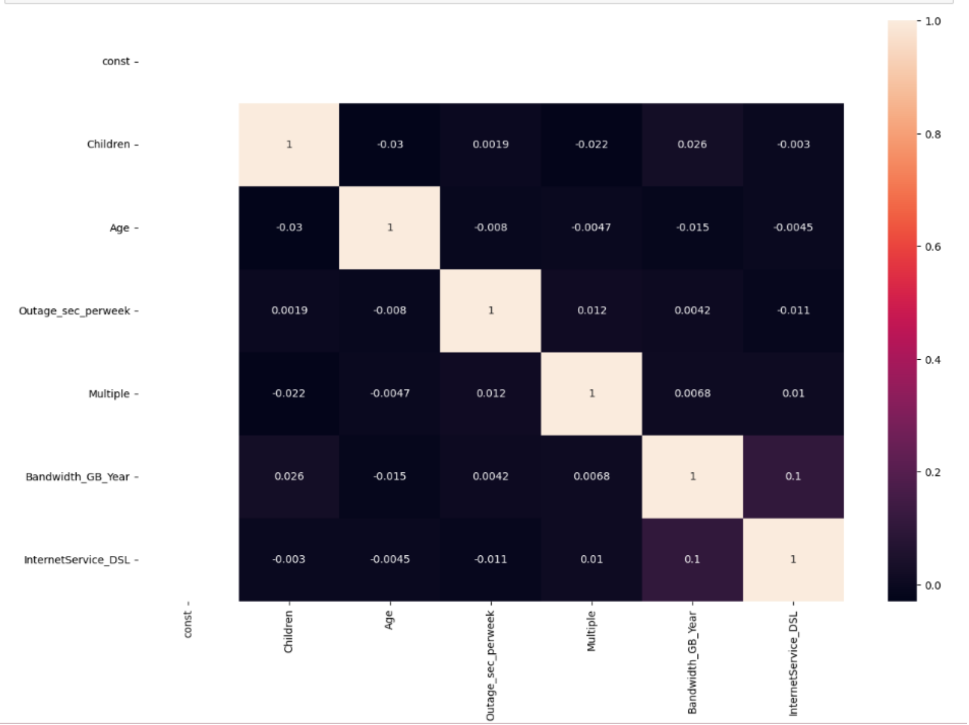
All the code is present in the attached .ipynb file provided in the attachment.

Here is the code for linear regression models using Python –

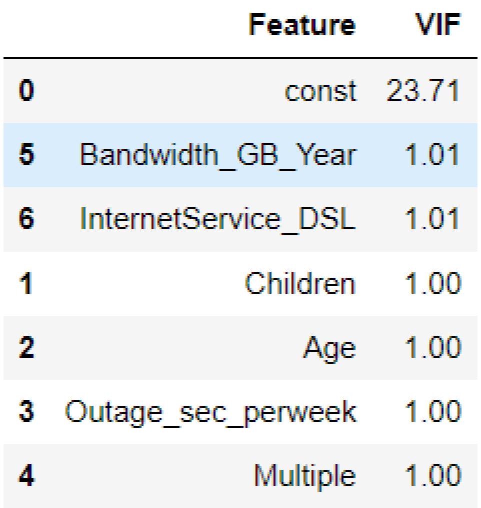
```
# Define the dependent variable (y) and independent variables (X)
y = df_churn['Tenure']
X = df_churn.drop(columns=['Tenure'])

# Add a constant term to the independent variables
X = sm.add_constant(X)

# Fit the multiple linear regression model
model = sm.OLS(y, X).fit()

# Print the model summary
print(model.summary())
```

Part V- Data Summary and Implications

(kassambara, 2018) **Check for Multicollinearity:** Multicollinearity does not exist in the final reduced model –



All the variable VIF is less than 5 for all the independent variables.



	Feature	VIF
0	const	23.71
5	Bandwidth_GB_Year	1.01
6	InternetService_DSL	1.01
1	Children	1.00
2	Age	1.00
3	Outage_sec_perweek	1.00
4	Multiple	1.00

Check for the mean of the residuals -

Additionally, I am checking another assumption that the mean of the residuals should be equal to zero. In my case, this is very close to zero (-0.0000000000006298).

The Residual Sum of Squares(RSS) is a measure of the sum of the squared differences between the predicted values and the actual values of the dependent variable in a regression model. I got RSS = 46504 which indicates that the model is unable to explain 46504 units of variance in the dependent variable. The larger the value of RSS, the greater the amount of unexplained variation in the dependent variable and the poorer the fit of the model to the data

Check for Heteroscedasticity :

Heteroscedasticity is a common issue in statistical analysis, and it occurs when the variance of the residuals (i.e., the differences between the predicted values and the actual values) is not constant across the range of the independent variable. One way to check for heteroscedasticity in Python is to use the statsmodels library.

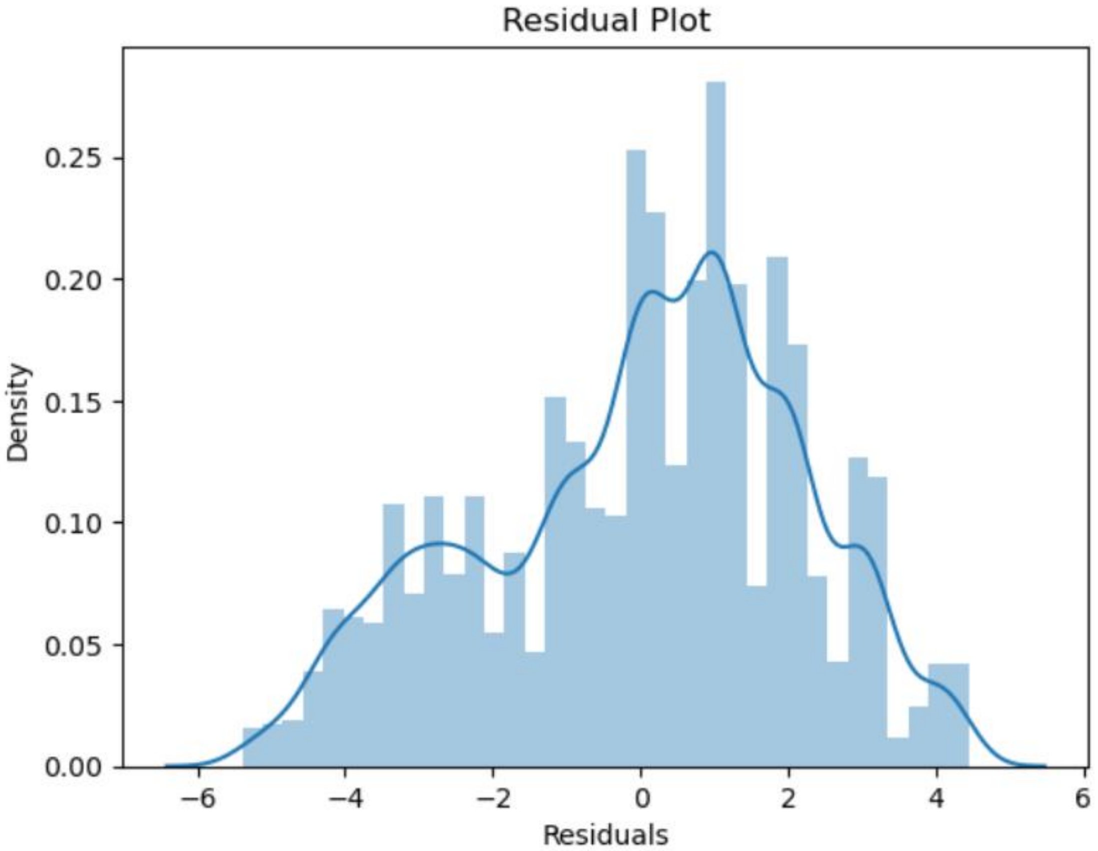
I am using the het_breuschpagan() function to perform the Breusch-Pagan test for heteroscedasticity. The test returns a p-value, and if the p-value is less than the significance level (usually 0.05), it indicates that there is evidence of heteroscedasticity in the data.

In my test, the p-value is 0.028 which is less than the significant level and it shows that my final reduced model is not able to satisfy the regression assumptions for Heteroscedasticity (kassambara, 2018).

Check the distribution of residuals :

(Zach, What Are Residuals in Statistics?, 2020) This allows us to verify 2 assumptions of a linear regression model

- Distribution of residuals are approximately normally distributed and
- Distribution of residuals is centered around zero



This is not perfectly normally distributed but doesn't have heavy tails so I can say that the linear regression model may be a good fit for the data. Also, the residuals don't have a mean of zero but are very close to zero which indicates that the model is not systematically over- or under-predicting the target variable (Zach, What Are Residuals in Statistics?, 2020).

F1: Results

The regression equation for the reduced model –

Tenure(y) = -5.5734 -0.3752 * Children + 0.0387 * Age - 0.0158 * Outage_sec_perweek - 0.8933 * Multiple + 0.0121 * Bandwidth_GB_Year - 5.0286 * InternetService_DSL

The regression coefficient interpretation associated with Tenure

The y-intercept is -5.5734 which means the estimated value of customer tenure whe n all the independent variables in the model are equal to zero is -5.5734.

Keeping all things constant, one unit increase in Children is associated with a 0.3752 unit decrease in Tenure.

Keeping all things constant, one unit increase in Age is associated with a 0.0387 unit increase in Tenure.

Keeping all things constant, one unit increase in Outage_sec_perweek is associated with a 0.0158 unit decrease in Tenure.

The tenure of the customers will decrease by 0.8933 units if they have multiple services as compared to the customers who have single service (kassambara, 2018).

Keeping all things constant, one unit increase in Bandwidth_GB_Year is associated with a 0.0121 unit increase in Tenure.

The tenure of the customers will decrease by 5.0286 units for DSL service providers as compared to the Non-DSL service providers (kassambara, 2018).

The statistical and practical significance of the reduced model

Statistical significance refers to whether the relationship between the independent variable(s) and the dependent variable is statistically significant, which is typically assessed using a p-value or confidence interval. All the independent variable in my reduced model has low p-value which denotes that all variables are significant to the dependent variable ‘Tenure’. Also reduced model Prob (F-statistic)is 0.00 which means the overall model is good. The multicollinearity among the independent variables is also very less so based on this I can conclude that this reduced model is Statistically significant.

The Practical significance can be measured using Effect size which is typically assessed using a coefficient estimate. The coefficient estimate of my reduced model is very low which means the change in the dependent variable is relatively small compared to the change in the independent variable.
Also as described above that the reduced model failed to satisfy the regression assumption for Homoscedasticity. Based on all the reasons I think my reduced model is not practically significant.

The limitations of the data analysis – here are some of the limitations

Regression analysis relies on several assumptions, including linearity, independenc e, normality, and homoscedasticity, among others. Violations of these assumptions affect the accuracy and reliability of the results and may require additional analysis.

Due to the presence of heteroscedasticity in my reduced model, the coefficient esti mates may be less precise, and additional work is required to determine if the mode l is significant or not.

Multiple linear regression assumes that the errors are normally distributed. In my results, the errors are not perfectly normally distributed. however, in practice, this may not always be the case, leading to inaccurate predictions and biased estimates.

Statsmodels is primarily designed for statistical modeling and may not be as effective as machine learning models for certain types of data, such as large datasets or complex non-linear relationships

Statsmodels is primarily designed for linear regression models and may not be as effective for non-linear regression models or models with complex interactions.

Backward stepwise elimination considers each predictor variable separately and eliminates the least significant one at each step. However, it may fail to capture important interactions between the predictor variables that affect the response.

F2: Recommendations

This regression analysis is to identify the key factors influencing Tenure which provides a starting point for developing effective retention strategies and specific courses of action in the organization.

Due to the lower coefficient estimate(size of the effect) which makes the model practically less significant I would recommend Re-evaluating the variable and exploring whether there are interactions between the independent variable with other variables that are not being considered in this analysis.

Part VI - Demonstration

G. Provide a Panopto recording

Panopto recording is provided.

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=4bbdceab-8d9e-46d4-9dbd-afd80167ffd6

H. Third-Party Code References

Alex. (2019, April 28). Multiple Regression Analysis in Python. https://www.youtube.com/watch?v=M32ghIt1c88&t=4s

Moffitt, C. (2017, February 06). Guide to Encoding Categorical Values in Python. pbpython.com: https://pbpython.com/categorical-encoding.html

Zach. (2020, JULY 20). How to Create a Q-Q Plot in Python. www.statology.org: https://www.statology.org/q-q-plot-python/

sklearn.preprocessing.LabelEncoder. (2013). https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

I. References

avcontentteam. (2016, July 14). Going Deeper into Regression Analysis with Assumptions, Plots & Solutions. https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/

Eubank, N. (n.d.). Using and Interpreting Indicator (Dummy) Variables. https://www.unifyingdatascience.org/html/interpreting_indicator_vars.html

Jain, K. (2017, Sept 12). Python vs. R vs. SAS – which tool should I learn for Data Science? https://www.analyticsvidhya.com/blog/2017/09/sas-vs-vs-python-tool-learn/

kassambara. (2018, November 03). Regression with Categorical Variables. http://www.sthda.com/english/articles/40-regression-analysis/163-regression-with-categorical-variables-dummy-coding-essentials-in-r/#:~:text=subtraction%20is%20reversed.-,Categorical%20variables%20with%20more%20than%20tw

Verma, V. (2020, October 24). A comprehensive guide to Feature Selection using Wrapper methods in Python. https://www.analyticsvidhya.com/blog/2020/10/a-comprehensive-guide-to-feature-selection-using-wrapper-methods-in-python/

Zach. (2019, MARCH 20). How to Read and Interpret a Regression Table. www.statology.org: https://www.statology.org/read-interpret-regression-table/

Zach. (2020, December 7). What Are Residuals in Statistics? https://www.statology.org/residuals/

Bevans, R. (2020, February 19). Simple Linear Regression. Retrieved from www.scribbr.com: https://www.scribbr.com/statistics/simple-linear-regression/

The Ultimate Guide to Linear Regression. (n.d.). https://www.graphpad.com/guides/the-ultimate-guide-to-linear-regression