D208

Performance Assessment Part 1 – Laurie Narcisse

Contents

[**Part I: Research Question** 2](#_Toc152860607)

[A.  Describe the purpose of this data analysis by doing the following: 2](#_Toc152860608)

[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 linear regression in the initial model. 2](#_Toc152860609)

[2.  Define the goals of the data analysis. 2](#_Toc152860610)

[**Part II: Method Justification** 3](#_Toc152860611)

[B.  Describe multiple linear regression methods by doing the following: 3](#_Toc152860612)

[1.  Summarize four assumptions of a multiple linear regression model. 3](#_Toc152860613)

[2.  Describe two benefits of using Python or R in support of various phases of the analysis. 4](#_Toc152860614)

[3.  Explain why multiple linear regression is an appropriate technique to use for analyzing the research question summarized in part I. 5](#_Toc152860615)

[**Part III: Data Preparation** 5](#_Toc152860616)

[C.  Summarize the data preparation process for multiple linear regression analysis by doing the following: 5](#_Toc152860617)

[1.  Describe your data cleaning goals and the steps used to clean the data to achieve the goals that align with your research question including your annotated code. 5](#_Toc152860618)

[2.  Describe the dependent variable and all independent variables using summary statistics that are required to answer the research question, including a screenshot of the summary statistics output for each of these variables. 6](#_Toc152860619)

[3.  Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables, including the dependent variable in your bivariate visualizations. 11](#_Toc152860620)

[4.  Describe your data transformation goals that align with your research question and the steps used to transform the data to achieve the goals, including the annotated code. 13](#_Toc152860621)

[5.  Provide the prepared data set as a CSV file. 14](#_Toc152860622)

[Part IV: Model Comparison and Analysis 15](#_Toc152860623)

[D.  Compare an initial and a reduced linear regression model by doing the following: 15](#_Toc152860624)

[1.  Construct an initial multiple linear regression model from all independent variables that were identified in part C2. 15](#_Toc152860625)

[2.  Justify a statistically based feature selection procedure or a model evaluation metric to reduce the initial model in a way that aligns with the research question. 15](#_Toc152860626)

[3.  Provide a reduced linear regression model that follows the feature selection or model evaluation process in part D2, including a screenshot of the output for each model. 20](#_Toc152860627)

[E.  Analyze the data set using your reduced linear regression model by doing the following: 21](#_Toc152860628)

[1.  Explain your data analysis process by comparing the initial multiple linear regression model and reduced linear regression model, including the following element: 21](#_Toc152860629)

[2.  Provide the output and *all* calculations of the analysis you performed, including the following elements for your reduced linear regression model: 22](#_Toc152860630)

[3.  Provide an executable error-free copy of the code used to support the implementation of the linear regression models using a Python or R file. 24](#_Toc152860631)

[**Part V: Data Summary and Implications** 24](#_Toc152860632)

[F.  Summarize your findings and assumptions by doing the following: 24](#_Toc152860633)

[1.  Discuss the results of your data analysis, including the following elements: 24](#_Toc152860634)

[2.  Recommend a course of action based on your results. 24](#_Toc152860635)

[**Part VI: Demonstration** 25](#_Toc152860636)

[G.  Provide a Panopto video recording that includes the presenter and a vocalized demonstration of the functionality of the code used for the analysis of the programming environment, including the following elements: 25](#_Toc152860637)

[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. 25](#_Toc152860638)

# **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 linear regression in the initial model.

**Research question:** What is the impact of various customer characteristics and behaviors on the monthly charges incurred by customers, focusing on the role of the selected features in predicting the monthly charge?

### 2.  Define the goals of the data analysis.

In my pursuit to unravel the intricate web of factors influencing monthly charges, the initial step involves identifying and comprehending the significant elements at play. Delving into this investigation requires a meticulous examination of various variables that might contribute to the fluctuations in monthly charges. It is essential to scrutinize not only the apparent factors but also those that might be hidden beneath the surface, influencing the overall dynamics of the monthly charges.

Embarking on the journey of data analysis, my focus shifts towards the development of a robust predictive model utilizing multiple linear regression. This statistical technique enables a comprehensive exploration of the relationships between the diverse independent variables and the elusive monthly charges. By harnessing the power of this method, I aim to create a predictive framework that not only accurately reflects the existing patterns but also possesses the potential to anticipate future variations in monthly charges.

As I navigate through the data, my objective is to unravel the nuanced relationships between the independent variables and the elusive monthly charges. This involves deciphering how each variable interplays with the others, contributing to the overall dynamics of the charges incurred monthly. By understanding these intricate connections, I laid the groundwork for a comprehensive analysis that goes beyond surface-level correlations, providing a deeper insight into the underlying mechanisms governing monthly charges.

The goal of this analytical endeavor is to offer actionable insights that can be employed to optimize monthly charges effectively. Armed with a nuanced understanding of the factors at play, I seek to provide practical recommendations and strategies for fine-tuning and streamlining the monthly charges. These insights not only serve as a guide for immediate optimization but also empower decision-makers with the tools to adapt and respond to evolving factors that may impact monthly charges in the future. In essence, the culmination of this analysis is not just the identification of influencing factors but the provision of a roadmap for proactive and strategic management of monthly charges based on a thorough and insightful analysis.

# **Part II: Method Justification**

## B.  Describe multiple linear regression methods by doing the following:

### 1.  Summarize **four** assumptions of a multiple linear regression model.

In the context of a multiple linear regression model, several key assumptions underlie the reliability and validity of the model's results. Firstly, the assumption of linearity posits that the relationship between the independent variables and the dependent variable is best characterized by a linear function. This implies that any change in the predictor variables is associated with a constant change in the response variable, ensuring a consistent and interpretable association.

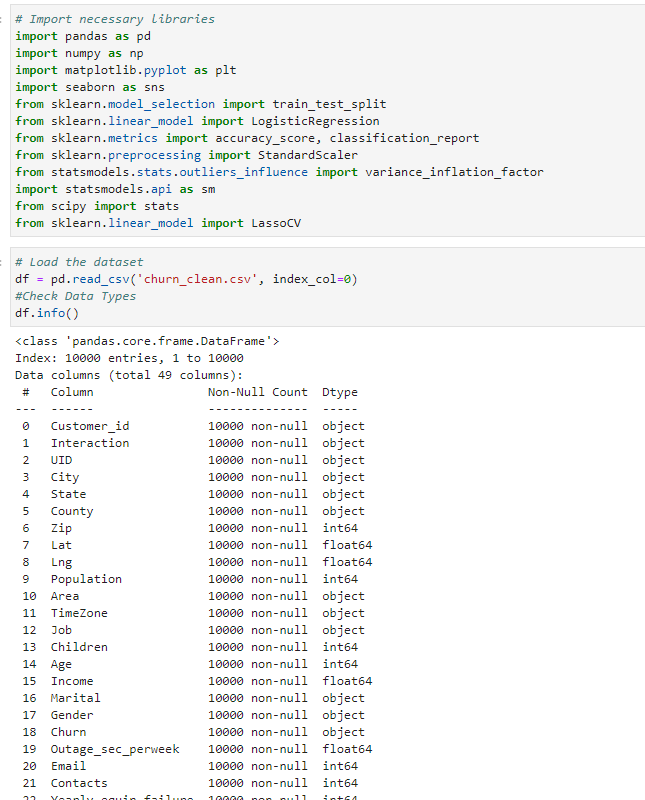
The assumption of independence is crucial, asserting that the residuals, or the differences between observed and predicted values, are independent of each other. This ensures that the occurrence of one residual does not provide information about the occurrence of another, preventing the introduction of bias into the model.

Homoscedasticity, another critical assumption, implies that the residuals exhibit constant variance across all levels of the independent variables. This assumption safeguards against the presence of systematic patterns in the residuals, ensuring that the spread of these differences remains consistent, contributing to the model's stability and reliability.

Lastly, the normality of residuals assumption posits that the distribution of residuals should ideally follow a normal distribution. While the model itself does not require normality, this assumption facilitates the application of inferential statistical techniques. Addressing violations of these assumptions is paramount to ensuring the accuracy and generalizability of the multiple linear regression model, thereby enhancing its utility in extracting meaningful insights from the data.

### 2.  Describe **two** benefits of using Python or R in support of various phases of the analysis.

The use of Python in multiple linear regression analysis offers several advantages, aligning with the specific needs of data analysts. Python's Pandas and NumPy libraries provide robust tools for efficient data manipulation, with Pandas offering flexible data structures such as data frames and NumPy excelling in numerical operations. Visualization is facilitated through libraries like Matplotlib and Seaborn, allowing for the creation of a diverse range of plots to aid in data exploration. Python's capacity to handle large datasets is enhanced by Pandas' optimized memory usage and support for parallel computing. Furthermore, Python's Statsmodels library offers comprehensive statistical modeling capabilities, including multiple linear regression, providing functionalities for model estimation and diagnostics. This comprehensive ecosystem, coupled with Python's widespread use in the machine learning domain through libraries like scikit-learn, positions Python as a versatile and powerful tool for conducting multiple linear regression analysis, meeting the demands of data analysts in various domains.



A white background with many text

Description automatically generated with medium confidence

### 3.  Explain why multiple linear regression is an appropriate technique to use for analyzing the research question summarized in part I.

Multiple linear regression is a suitable statistical technique for analyzing the impact of various customer characteristics and behaviors on monthly charges. In exploring the research question about the impact of various customer characteristics and behaviors on monthly charges, I have opted for multiple linear regression as the statistical technique. This choice is driven by the need to understand how selected features collectively influence monthly charges and to quantify the individual contributions of diverse customer-related factors.

First and foremost, my research question involves delving into the influence of several customer characteristics and behaviors on monthly charges. Multiple linear regression aligns well with this inquiry because it allows me to incorporate more than one predictor variable. This flexibility enables the inclusion of a range of customer-related features, such as demographics and usage patterns, providing a comprehensive analysis of potential factors contributing to variations in monthly charges.

Moreover, the technique empowers me to assess the strength and direction of the relationships between each independent variable and the dependent variable. By estimating coefficients associated with each predictor, I can discern the expected change in monthly charges for a unit change in each characteristic, while keeping other variables constant. This not only helps identify significant factors but also provides insights into the magnitude and direction of their impact on monthly charges.

Furthermore, multiple linear regression provides a holistic view of the combined influence of multiple predictors on the dependent variable. This aspect is crucial in understanding how customer characteristics and behaviors may interact or have overlapping effects on monthly charges. The technique allows for the exploration of these collective impacts, contributing to a nuanced understanding of the underlying dynamics.

Additionally, the statistical evaluation capabilities of multiple linear regression are valuable. I can employ tests to determine whether the selected features collectively contribute significantly to explaining the variability in monthly charges. This step ensures that the relationships observed are not merely due to random chance, adding a layer of confidence to the reliability of the analysis.

In essence, the choice of multiple linear regression is driven by its ability to handle multiple variables, provide quantitative insights into individual and collective effects, and allow for a rigorous statistical evaluation of the overall model performance. It enables a thorough examination of the impact of various customer characteristics and behaviors on monthly charges, contributing to a more comprehensive and nuanced understanding of the research question.

# **Part III: Data Preparation**

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

### 1.  Describe your data cleaning goals and the steps used to clean the data to achieve the goals that align with your research question including your annotated code.

The dataset supplied by WGU is initially described as clean; however, upon closer examination, it is evident that there are certain imperfections. To address this, I plan to leverage codes from a prior class's Cleaning performance assessment. These codes will assist in handling missing values, detecting outliers, formatting the dataset, and converting categorical columns, ensuring a more accurate and reliable foundation for subsequent analysis.

In the data cleaning process, I took several steps to ensure the dataset is prepared for meaningful analysis. Initially, I identified and addressed missing values across columns, distinguishing between numerical and categorical features. For numerical columns, I opted to fill missing values with the mean, aiming to maintain the overall statistical properties of the data. Meanwhile, for categorical columns, I used the mode, representing the most frequent value, as a suitable replacement for missing entries.

To handle potential outliers, I implemented a z-score-based approach, defining a function to identify and remove data points beyond a specified threshold. This outlier treatment was specifically applied to numerical columns, enhancing the robustness of the dataset.

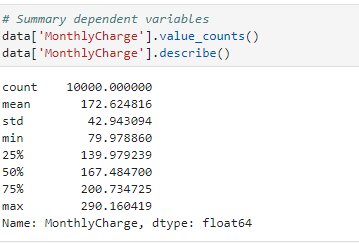
Next, I converted categorical columns into a numeric format, facilitating further analysis. Specifically, I mapped 'Yes' to 1 and 'No' to 0 in relevant columns. Additionally, I performed various formatting adjustments, such as filling missing values in the 'Age' column with zeros and converting it to an integer type. The 'Lng' column underwent similar treatment, being filled with zeros for missing values and converted to its absolute value.

Furthermore, certain categorical columns were mapped to numeric values using predefined dictionaries. For instance, 'Area,' 'Marital,' 'Gender,' 'Contract,' 'InternetService,' and 'PaymentMethod' were transformed to numeric representations, aligning them with the analytical requirements.

Finally, the cleaned dataset, incorporating all these adjustments, was saved to a new CSV file named 'cleaned\_churn\_data.csv.' This meticulous data cleaning process ensures that the dataset is now well-structured and free from issues such as missing values and outliers, setting the stage for a robust and reliable analysis of customer churn in the telecommunications industry.

### 2.  Describe the dependent variable and all independent variables using summary statistics that are required to answer the research question, including a screenshot of the summary statistics output for each of these variables.

Dependent Variable:

* MonthlyCharge: The amount charged to the customer monthly
  + The "MonthlyCharge" column has no missing values, with a mean of 172.62 and a standard deviation of 42.94. This indicates variability in monthly charges among customers, ranging from a minimum of 79.98 to a maximum of 290.16.
  + 

Independent Variables:

* Age
  + The "Age" column has no missing values and a mean of 53.08, with a standard deviation of 20.70, highlighting a moderate spread in customer ages. The minimum age is 18, and the maximum is 89. The data's quartiles show a fairly even distribution across age groups, with the median (50th percentile) at 53.
  + A screenshot of a computer

    Description automatically generated
* Gender
  + The column "Gender" has no missing values. The mean of 0.525 suggests a relatively balanced distribution between male (coded as 0) and female (coded as 1) customers. The standard deviation of 0.499 indicates moderate variability, reflecting the diversity in gender representation among customers.
  + A screenshot of a computer

    Description automatically generated
* StreamingMovies: Whether the customer has streaming movies (yes, no)
  + The "StreamingMovies" column is binary, representing whether customers have streaming movies (1) or not (0). All 10,000 entries are non-null, and the mean of 0.489 suggests that a slightly lower proportion of customers have streaming movies compared to those who do not.
  + A screenshot of a computer program

    Description automatically generated
* Churn: Whether the customer discontinued service within the last month (yes, no) - This remains the dependent variable .
  + A screenshot of a computer code

    Description automatically generated
* Tenure: Number of months the customer has stayed with the provider
  + The "Tenure" column, representing the number of months a customer has stayed with the provider, has no missing values. The mean of 34.53 months and a standard deviation of 26.44 suggest variability in customer tenure, with a minimum of 1 month and a maximum of 71.99 months.
  + A screenshot of a computer

    Description automatically generated
* Outage\_sec\_perweek: Average number of seconds per week of system outages in the customer’s neighborhood
  + The "Outage\_sec\_perweek" column has no missing values, with a mean of 10.00 seconds per week and a standard deviation of 2.98. The data reflects variability in weekly outage seconds, ranging from a minimum of 0.10 seconds to a maximum of 21.21 seconds.
  + A screenshot of a computer code

    Description automatically generated
* Email: Number of emails sent to the customer in the last year
  + A screenshot of a computer

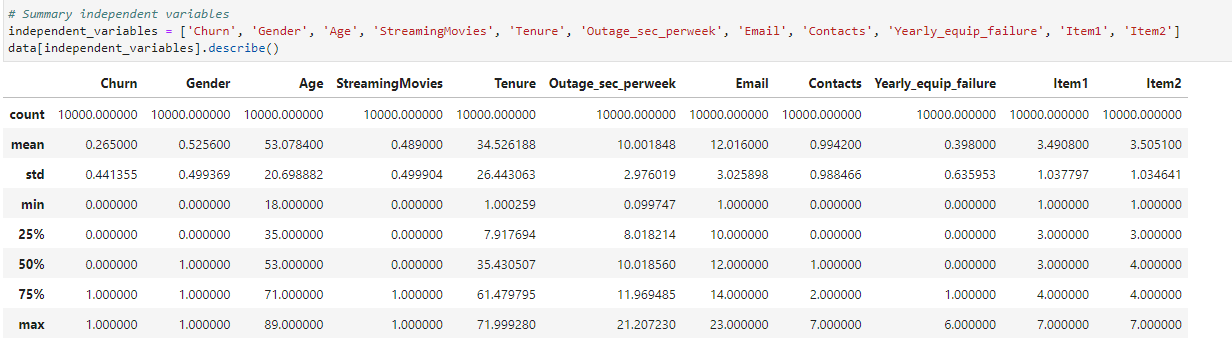
    Description automatically generated
* Contacts: Number of times customer contacted technical support
  + A screenshot of a computer

    Description automatically generated
* Yearly\_equip\_failure: The number of times customer’s equipment failed and had to be reset/replaced in the past year
  + A screenshot of a computer code

    Description automatically generated
* Email, Contacts, Yearly\_equip\_failure:
  + - These columns have no missing values. "Email" has a mean of 12.02, "Contacts" has a mean of 0.99, and "Yearly\_equip\_failure" has a mean of 0.40. These metrics suggest varying degrees of customer interactions and equipment failures.
* Item1 to Item2: Responses to an eight-question survey asking customers to rate the importance of various factors on a scale of 1 to 8
  + The columns "Item1" and "Item2" represent responses to survey questions, with means of 3.49 and 3.51, respectively. The standard deviations of 1.04 indicate moderate variability in customer ratings. The data range from a minimum of 1 to a maximum of 7, reflecting the diverse responses to the survey items
  + A screenshot of a computer

    Description automatically generated A screenshot of a computer

    Description automatically generated



### 3.  Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables, including the dependent variable in your bivariate visualizations.

Histograms:

A screenshot of a graph

Description automatically generated

Churn (Yes/No)

Gender( Female/Male)

A screenshot of a graph

Description automatically generated

A graph with blue lines

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph of data and a chart of data

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

### 4.  Describe your data transformation goals that align with your research question and the steps used to transform the data to achieve the goals, including the annotated code.

I will provide screenshots below and an attachment.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

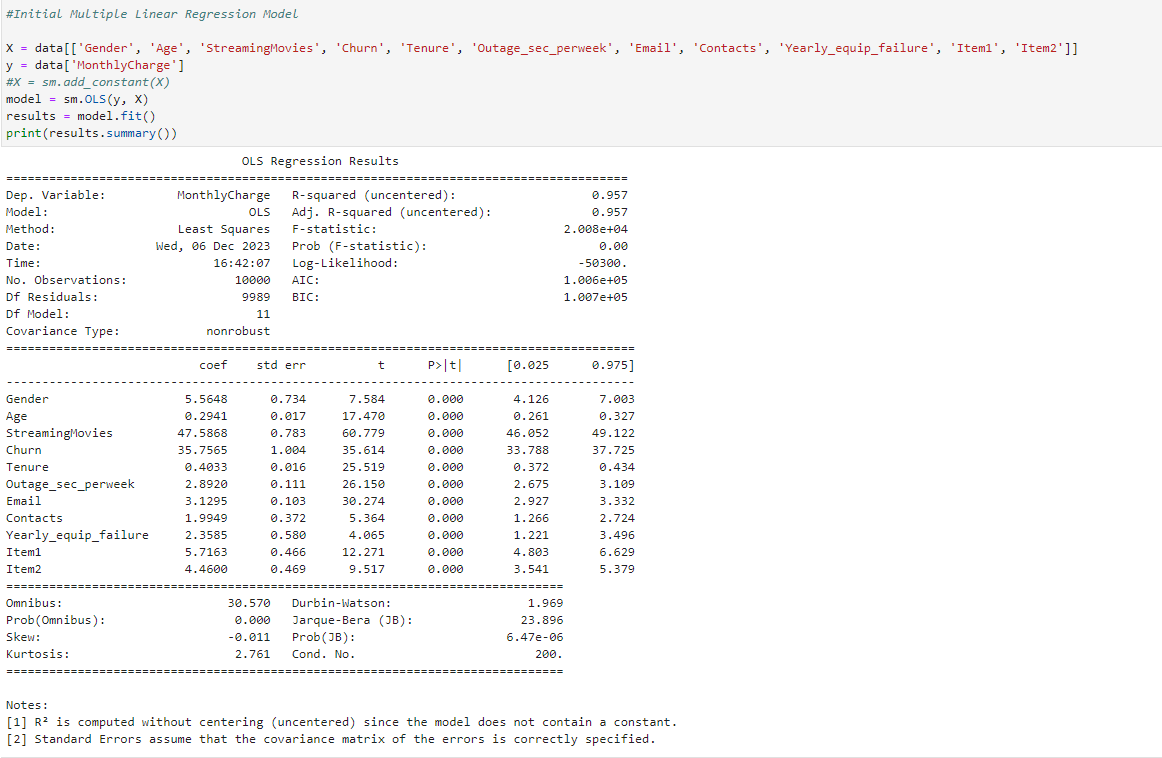
### 5.  Provide the prepared data set as a CSV file.

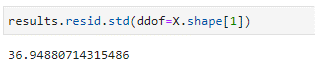
The file is labeled as cleaned\_churn\_data.csv.

# **Part IV: Model Comparison and Analysis**

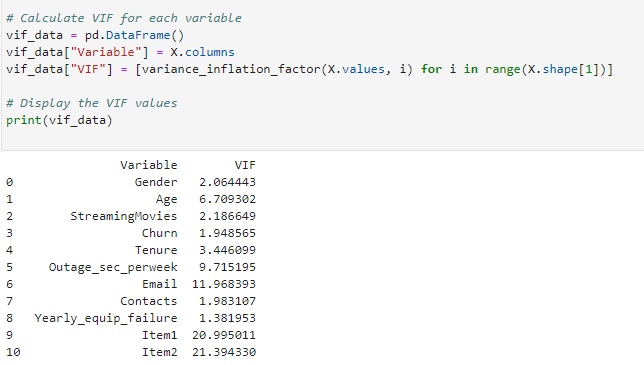
## D.  Compare an initial and a reduced linear regression model by doing the following:

### 1.  Construct an initial multiple linear regression model from all independent variables that were identified in part C2.





### 2.  Justify a statistically based feature selection procedure or a model evaluation metric to reduce the initial model in a way that aligns with the research question.



Calculating Variance Inflation Factor (VIF): In this step, I calculated the VIF for each variable in the multiple linear regression model. The VIF measures the degree of multicollinearity among independent variables. High VIF values, generally above 10, indicate strong multicollinearity. This process helps identify variables that might need further attention due to collinearity.



A computer screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer error

Description automatically generated

VIF-Based Feature Selection with Lasso Regression: The calculate\_vif function is used to iteratively drop variables with high VIF values. The threshold for dropping variables is set at 11. The process involves calculating VIF, dropping the variable with the highest VIF, and repeating until all VIF values are below the threshold. This is followed by fitting a Lasso regression model on the filtered features, which performs regularization and helps in feature selection.

Removing variables that have a high p-value.

The high VIF for 'const' indicates multicollinearity.

A computer screen shot of a computer code

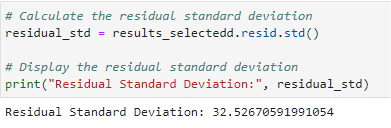
Description automatically generated

A screenshot of a computer

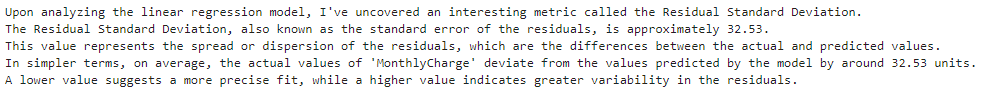
Description automatically generated

I've modified the linear regression model by removing the variables 'Age,' 'Contacts,' 'Outage\_sec\_perweek,' and 'Item1,' which appeared to be statistically insignificant based on their p-values and had high VIF values. The focus is now on 'StreamingMovies,' 'Churn,' and 'Tenure' as the main predictors of 'MonthlyCharge.'

Updated Multiple Linear Regression Model with Selected Variables: After VIF-based feature selection, I updated the multiple linear regression model with the selected variables. The summary statistics, including R-squared and coefficients, are displayed. This process helps evaluate the impact of feature selection on the model's performance and interpretability.

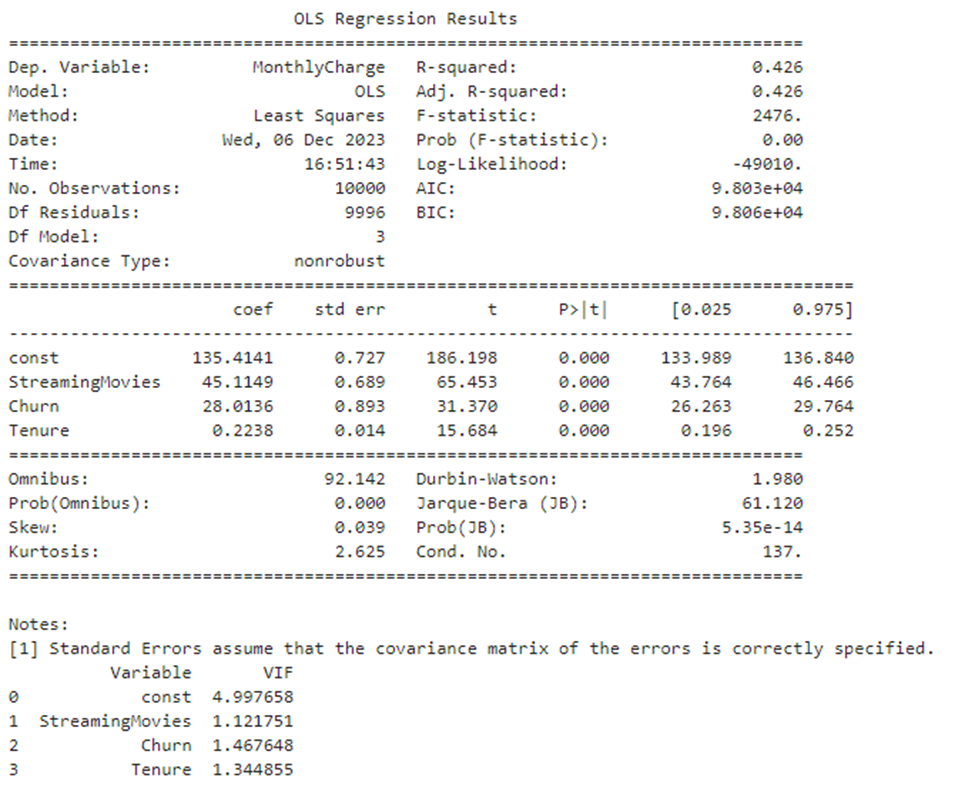


Calculating Residual Standard Deviation:



### 3.  Provide a reduced linear regression model that follows the feature selection or model evaluation process in part D2, including a screenshot of the output for each model.

The variables for Gender, Age, Outage\_sec\_perweek, Email, Contacts, Item1, Item2, Yearly\_equip\_failure were removed.

Below is the model: 

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

### 1.  Explain your data analysis process by comparing the initial multiple linear regression model and reduced linear regression model, including the following element:

Comparison of Initial and Reduced Models:

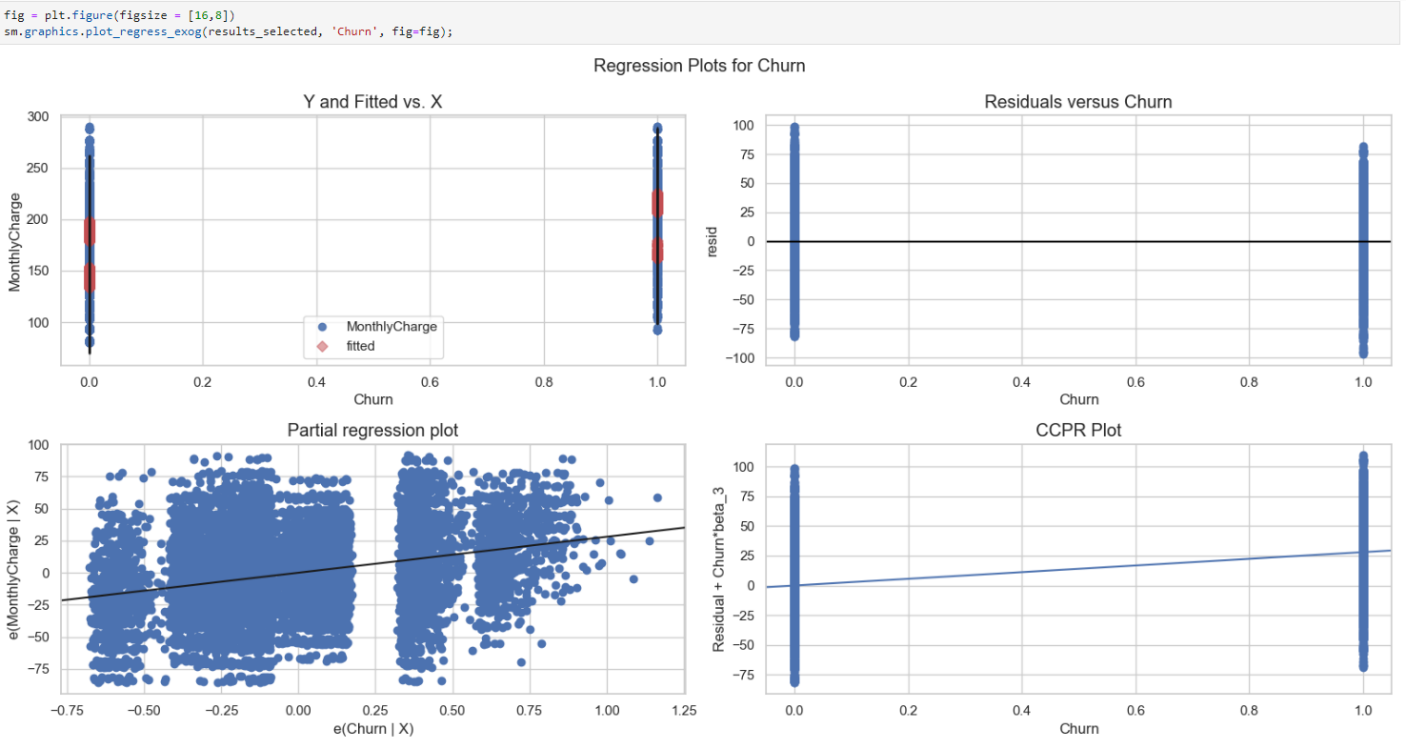
* Initially, I conducted a multiple linear regression analysis with several independent variables. To enhance the model's interpretability and address multicollinearity, I employed variance inflation factor (VIF) analysis and removed variables with VIF above a threshold (11) and removed high p values.
* The reduced model includes the variables StreamingMovies, Churn, and Tenure.
* To evaluate the model, I used the residual standard deviation as a metric
  + A screenshot of a computer code

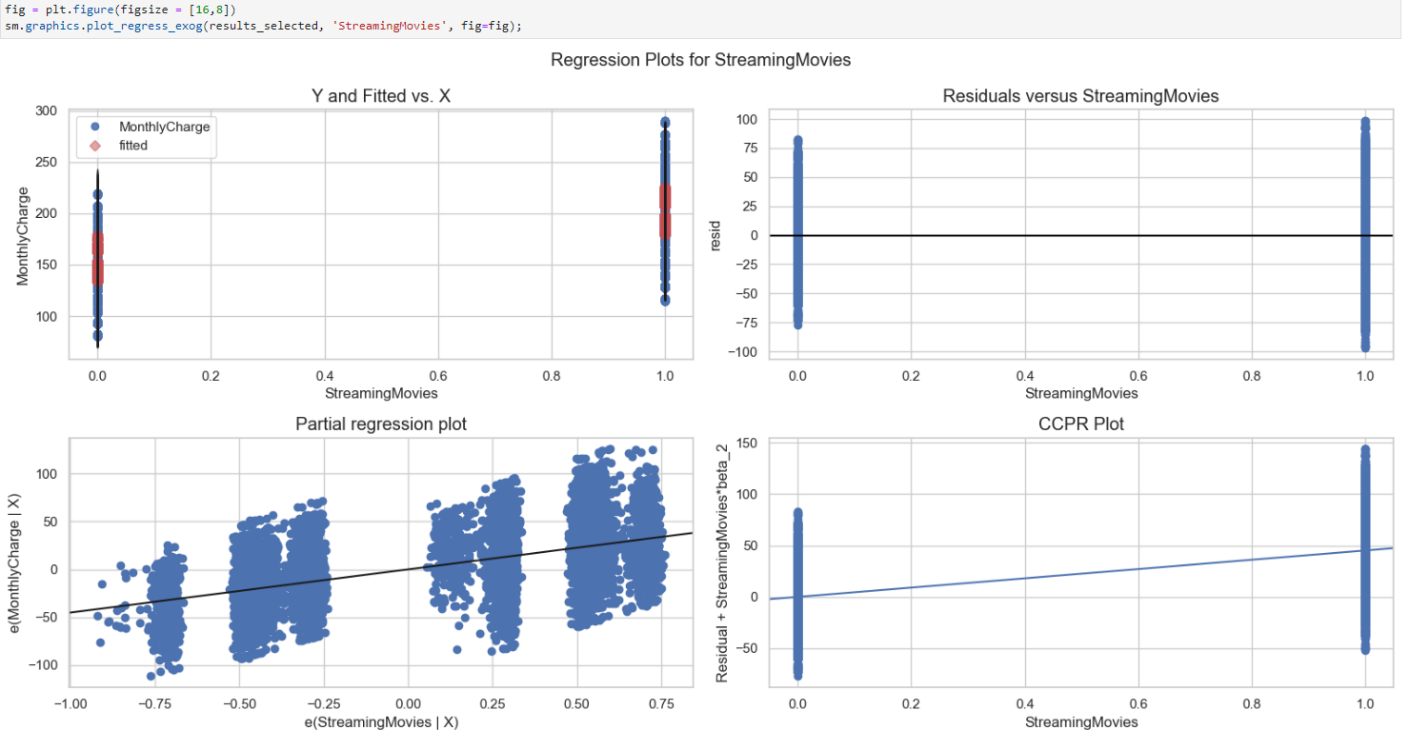
    Description automatically generated

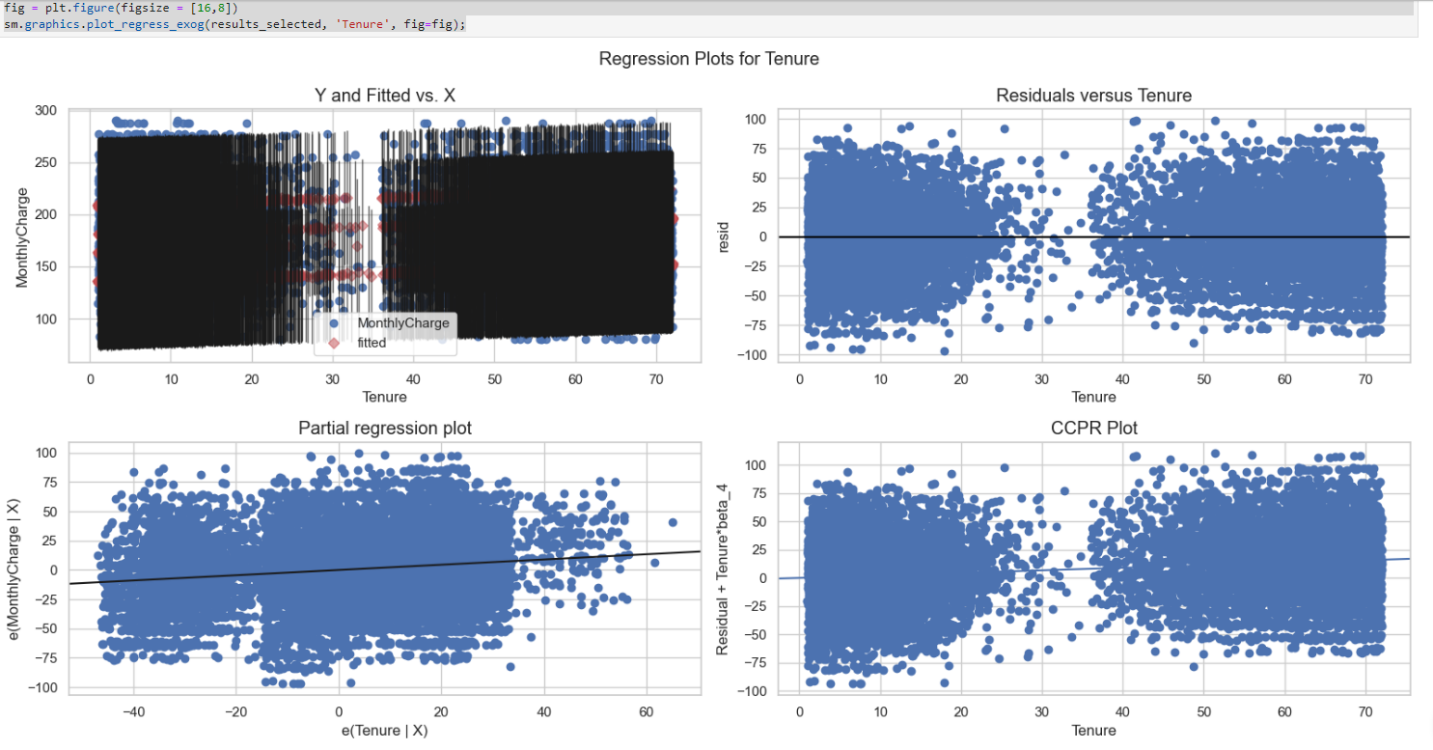
### 2.  Provide the output and all calculations of the analysis you performed, including the following elements for your reduced linear regression model:

Please see attached to see all input and out calculations. The video will also provided all outputs as well.

Residual Plot:







* Residual Standard Error:
  + The residual standard deviation for the reduced model is approximately 32.52670591991054

### 3.  Provide an executable error-free copy of the code used to support the implementation of the linear regression models using a Python or R file.

Please see PA1.pynb attached.

# **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

* The regression equation for the reduced model can be derived from the coefficients provided in the results summary. For example, if the reduced model has the equation:
  + MonthlyCharge=135.4141+45.1149×StreamingMovies+28.0136×Churn+0.2238×Tenure
  + This equation represents the relationship between the independent and the dependent variable.

#### •   an interpretation of the coefficients of the reduced model

* Intercept (const): 135.4141 - This represents the estimated monthly charge when all other predictors are zero.
* StreamingMovies: 45.1149 - For each additional unit increase in StreamingMovies, the monthly charge is estimated to increase by 45.1149 units.
* Churn: 28.0136 - If a customer has churned, the monthly charge is estimated to increase by 28.0136 units.
* Tenure: 0.2238 - For each additional month of tenure, the monthly charge is estimated to increase by 0.2238 units.

#### •   the statistical and practical significance of the reduced model

* Statistical Significance: All coefficients have p-values less than 0.05, indicating that they are statistically significant.
* Practical Significance: The practical impact of a one-unit increase in Tenure might be small compared to StreamingMovies.

#### •   the limitations of the data analysis

* Causation: My analysis identifies associations, but it doesn't establish causation. Remember, correlation does not imply causation.
* Model Assumptions: I assume linearity, independence of errors, homoscedasticity, and normally distributed errors. If these assumptions are violated, it might affect the results.
* Explanatory Power: The R-squared value is 0.426, indicating that our model explains 42.6% of the variance. However, there might be other important factors not considered.

### 2.  Recommend a course of action based on your results.

In light of the OLS regression results, several recommendations emerge to guide further investigation and strategic decision-making based on the identified factors influencing monthly charges.

Firstly, it is crucial to delve into the specific aspects of StreamingMovies that contribute to higher monthly charges. Given its significant impact as indicated by the positive coefficient, understanding the components of StreamingMovies could inform targeted strategies.

Addressing churn management is also imperative, considering the positive coefficient associated with it. Implementing strategies to mitigate customer churn may positively influence monthly charges, emphasizing the importance of customer retention initiatives.

The positive relationship between monthly charges and Tenure suggests that longer-tenured customers tend to have higher monthly charges. Investigating the reasons behind this trend is key, and considering loyalty programs or incentives for long-term customers may be a strategic approach.

To ensure the model's reliability and generalizability, validating it on additional datasets is recommended. Additionally, explore opportunities for model improvement by incorporating other relevant variables that might influence monthly charges.

Although the Variance Inflation Factor (VIF) values suggest low multicollinearity among the variables, continued monitoring for changes in the dataset that might impact multicollinearity is advised.

Lastly, the residual standard deviation provides insights into the variability of the residuals. Continuous monitoring of this metric is essential to assess the model's predictive accuracy over time.

In conclusion, these recommendations collectively aim to enhance the understanding of the factors influencing monthly charges and provide a strategic framework for decision-making in response to the research question.

# **Part VI: Demonstration**

G.  Provide a Panopto video recording that includes the presenter and a vocalized demonstration of the functionality of the code used for the analysis of the programming environment, including the following elements:

Please see attached.

## 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.

*How to Create a Residual Plot in Python*. (2023, December 4). *How to Create a Residual Plot in Python*. <https://www.geeksforgeeks.org/how-to-create-a-residual-plot-in-python/>

*How to Create a Residual Plot in Python*. (2023, December 4). *How to Create a Residual Plot in Python*. <https://www.statology.org/residual-plot-python/>

*p-value Basics with Python Code*. (2023, December 4). *p-value Basics with Python Code*. <https://towardsdatascience.com/p-value-basics-with-python-code-ae5316197c52>

*Explaining P-Value and its Interpretation with Examples in Python*. (2023, December 4). *Explaining P-Value and its Interpretation with Examples in Python*. <https://python.plainenglish.io/explaining-p-value-and-its-interpretation-with-examples-in-python-with-chatgpt-acdc8cb55576>

*Python Statistics – Python p-Value, Correlation, T-test, KS Test*. (2023, December 4). *Python Statistics – Python p-Value, Correlation, T-test, KS Test*. <https://data-flair.training/blogs/python-statistics/>

*ML | Multiple Linear Regression using Python*. (2023, December 4). *ML | Multiple Linear Regression using Python*. <https://www.geeksforgeeks.org/ml-multiple-linear-regression-using-python/>

*PA from Data Cleaning D206*

*Multiple Linear Regression Implementation in Python*. (2023, December 4). *Multiple Linear Regression Implementation in Python*. <https://medium.com/machine-learning-with-python/multiple-linear-regression-implementation-in-python-2de9b303fc0c>

*Python Machine Learning Multiple Regression*. (2023, December 4). *Python Machine Learning Multiple Regression*. <https://www.w3schools.com/python/python_ml_multiple_regression.asp>

*How to Calculate VIF in Python*. (2023, December 4). *How to Calculate VIF in Python*. <https://www.statology.org/how-to-calculate-vif-in-python/>

*Variance Inflation Factor in Python*. (2023, December 4). *Variance Inflation Factor in Python*. Stackoverflow. <https://stackoverflow.com/questions/42658379/variance-inflation-factor-in-python>

*Detecting Multicollinearity with VIF - Python*. (2023, December 4). *Detecting Multicollinearity with VIF - Python*. <https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/>