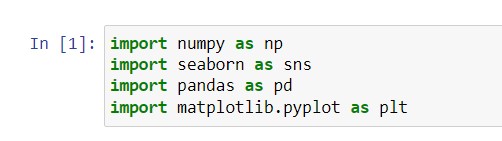
PART A (EXPLORATORY DATA ANALYSIS (EDA) AND PRE-PROCESSING)

1. Importing requisite libraries 

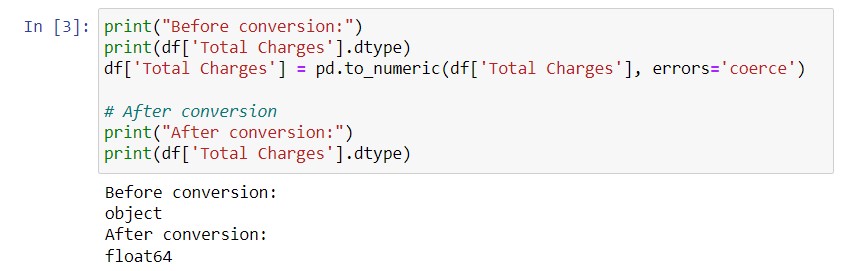
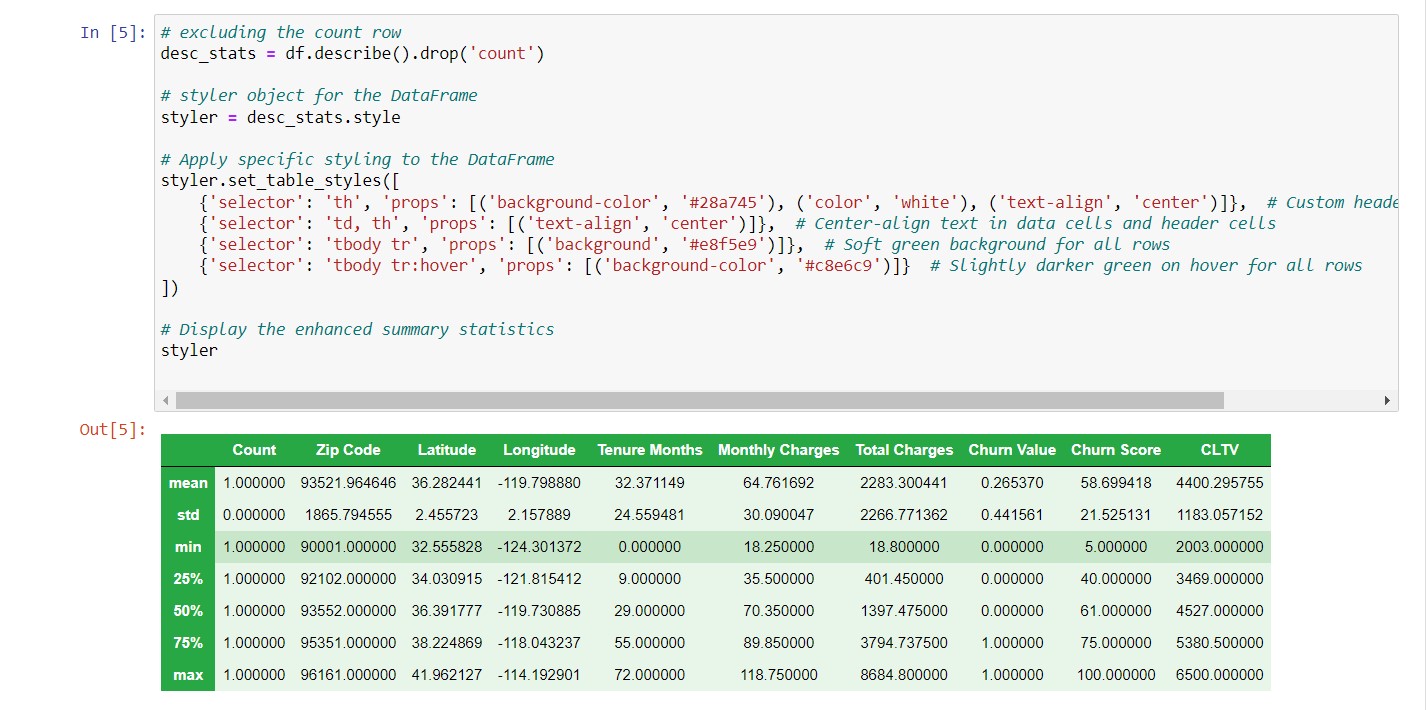
**NumPy:** The name “Numpy” stands for “Numerical Python. It is a machine learning library that supports large matrices and multi-dimensional data. It consists of in-built mathematical functions for easy computations.

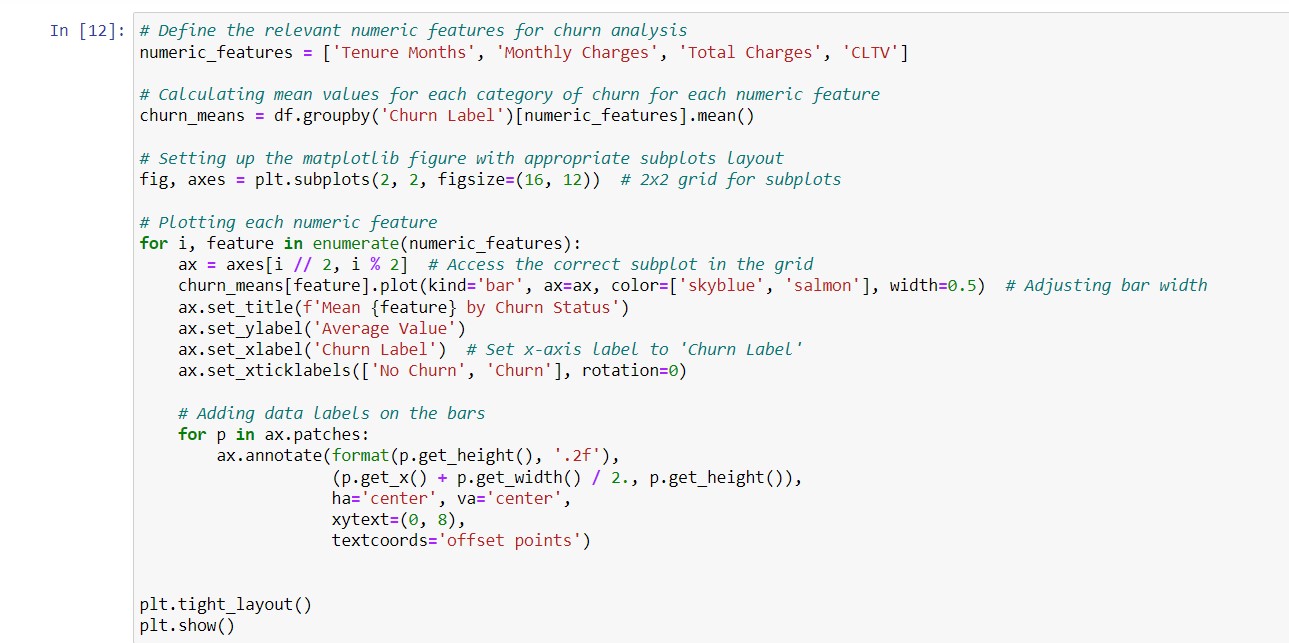
**Seaborn:** Seaborn is another open-source Python library, one that is based on Matplotlib (which focuses on plotting and data visualization) but features Pandas’ data structures. Seaborn is often used in ML projects because it can generate plots of learning data. Of all the Python libraries, it produces the most aesthetically pleasing graphs and plots, making it an effective choice if you also use it for marketing and data analysis.

**Pandas:** It is an open-source machine learning library that provides flexible high-level data structures and a variety of analysis tools. It eases data analysis, data manipulation, and cleaning of data. Pandas support operations like Sorting, Re-indexing, Iteration, Concatenation, Conversion of data, Visualizations, Aggregations, etc.

**Matplotlib:** This library is responsible for plotting numerical data. And that’s why it is used in data analysis. It is also an open-source library and plots high-defined figures like pie charts, histograms, scatterplots, graphs, etc.

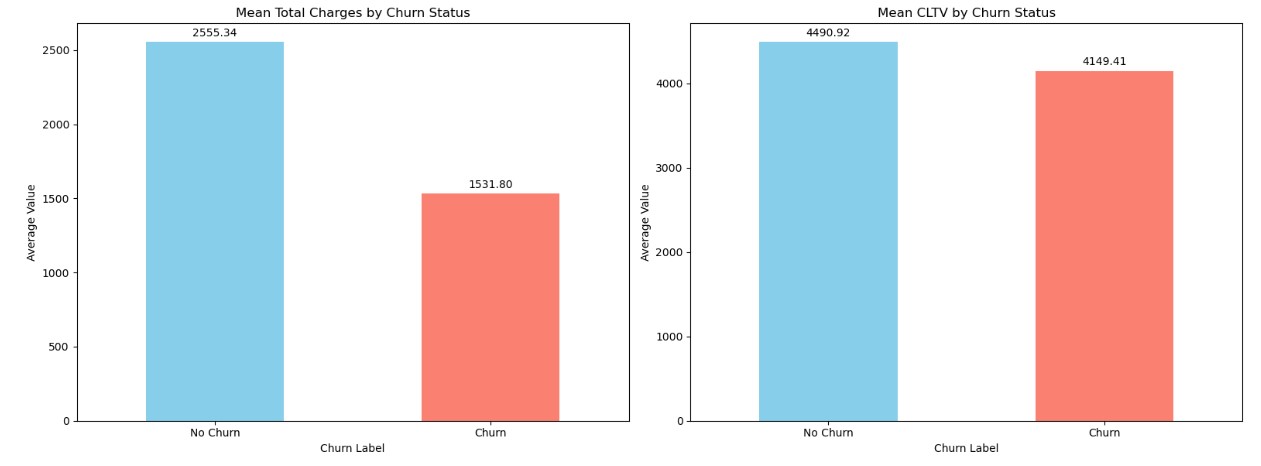
1. Loading the churn.csv dataset into python data frame and checking contents A screenshot of a computer

   Description automatically generated
2. Total Charges to numeric 
3. Summary of descriptive statistics (styling: cantered text, background gradient, custom header colours)

e) Analysing customer churn reason 

A comparison of a bar chart

Description automatically generated with medium confidence



1. Insights:
2. **Tenure Months**: (Tenure is the number of months or years a customer has been using a mobile phone service is known as tenure. Churn is the percentage of customers who cancel their mobile phone service each month).
   * Customers who churn have a significantly lower average tenure (17.98 months) compared to those who do not churn (37.57 months).
   * This suggests that longer tenure is associated with a lower likelihood of churn, indicating that customer loyalty and retention tend to increase with longer tenure.
3. **Monthly Charges**:
   * Churned customers tend to have higher average monthly charges (74.44) compared to non-churned customers (61.27).
   * Higher monthly charges may contribute to customer dissatisfaction or be indicative of premium services, potentially influencing churn rates.
4. **Total Charges**:
   * The average total charges for customers who churned (1531.80) are notably lower than those who did not churn (2555.34).
   * Lower total charges among churned customers could suggest shorter service duration or lower overall spending, impacting churn behaviour.
5. **Customer Lifetime Value (CLTV)**:

(CLTV measures the total revenue a company can expect from a customer over the duration of their relationship with the company.)

* + Both churned and non-churned customers show relatively high CLTV values, with non-churned customers having a slightly higher average CLTV (4490.92) compared to churned customers (4149.41).
  + Higher CLTV typically indicates greater value derived from customers over their lifetime, with slightly higher values associated with lower churn rates.

**Summary Insights**:

* **Tenure**: Longer tenure is associated with reduced churn, highlighting the importance of customer retention strategies over time.
* **Monthly Charges**: Higher monthly charges may contribute to increased churn rates, warranting analysis of pricing strategies and service offerings.
* **Total Charges**: Lower total charges among churned customers may indicate different spending patterns or service usage behaviour.
* **CLTV**: CLTV is relatively high for both churned and non-churned customers, showing the potential value of retaining high-value customers.

1. Pie Chart

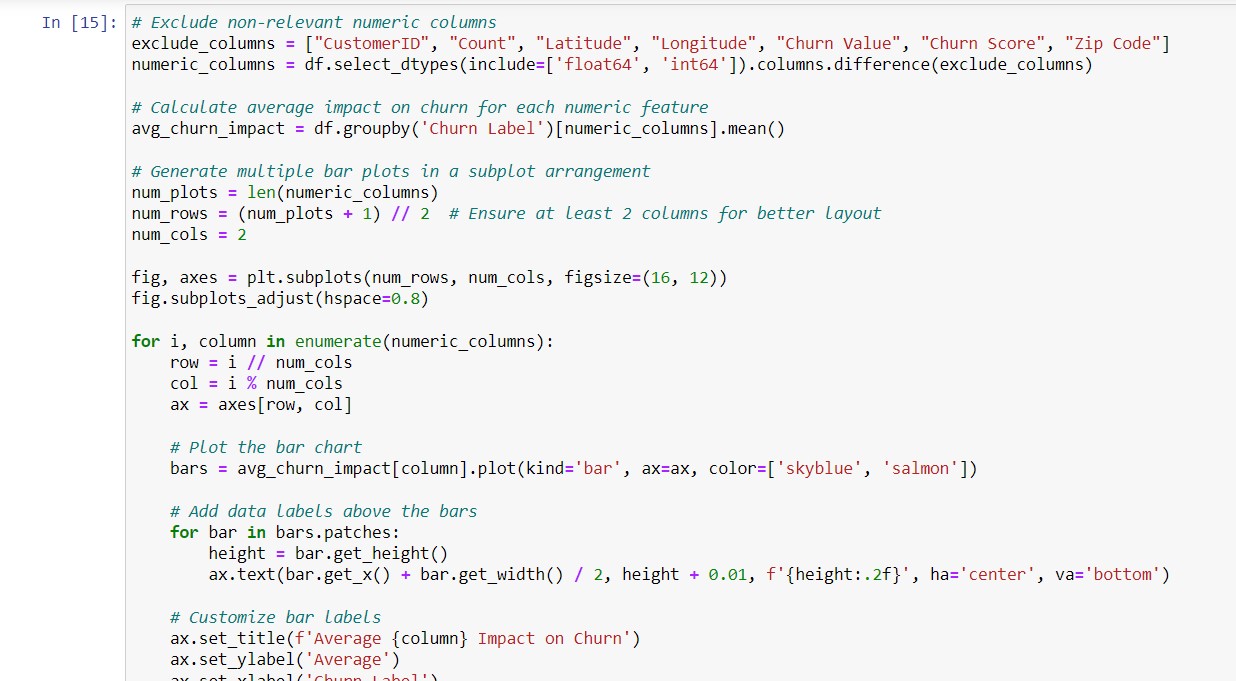
A screenshot of a graph

Description automatically generated

figure(): This function is used to create a new figure, which is a container that holds all the plot elements (like axes, labels, title, etc.).

figsize=(7, 5): This argument specifies the size of the figure in inches. In this case, the figure will be 7 inches wide and 5 inches tall.

1. Set of count plot visualizations for categorical variables:



A computer screen shot of a code

Description automatically generated

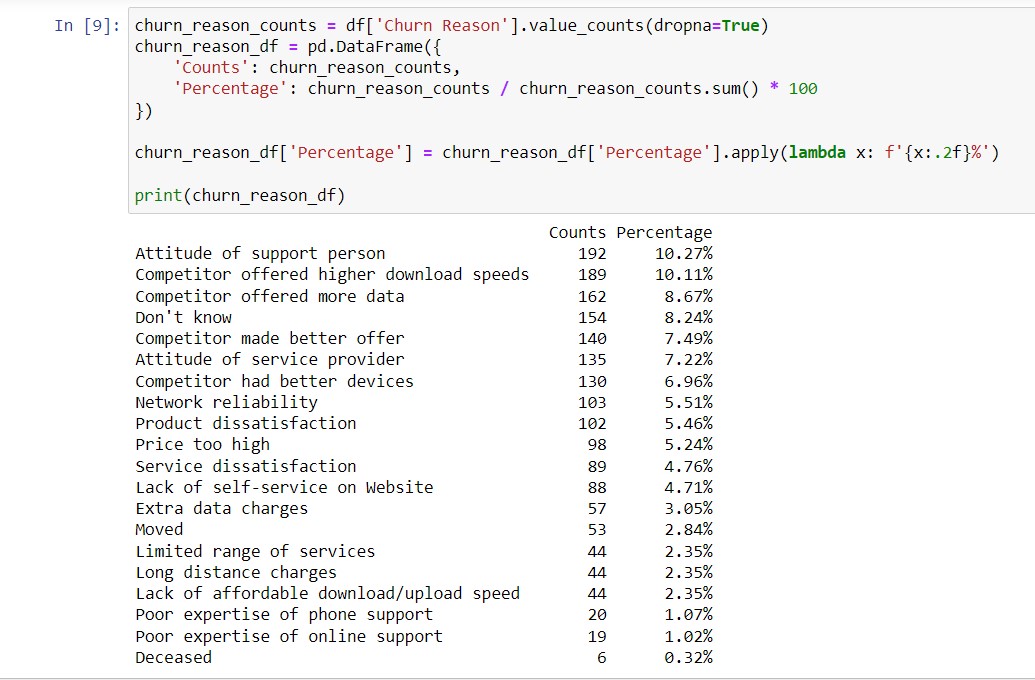
A comparison of a graph

Description automatically generated with medium confidence

A comparison of a graph

Description automatically generated with medium confidence

1. Frequency and percentage of churn reason:



1. TOP 3 Churn Influencers:
   1. Attitude of Support Person (10.27%)

Insight: The behaviour of support staff significantly impacts customer churn rates, indicating that negative customer service experiences contribute to dissatisfaction.

Intervention: Implement specialized training programs aimed at enhancing the interpersonal skills and empathy of support personnel. Prioritize positive customer interactions and swift issue resolution to improve customer satisfaction.

* 1. Competitor Offered Higher Download Speeds (10.11%)

Insight: Customers are drawn to competitors offering faster download speeds, underscoring the importance of network performance in retaining customers.

Intervention: Conduct a thorough analysis of network performance and invest in infrastructure enhancements to boost download speeds. Communicate these improvements effectively to customers to showcase the competitive advantage.

* 1. Competitor Offered More Data (8.67%)

Insight: Competitors offering generous data packages influence churn rates, indicating that customers value data allowances.

Intervention: Review current data plans and explore options for competitive data packages. Implement personalized promotions and offers to retain customers seeking higher data limits.

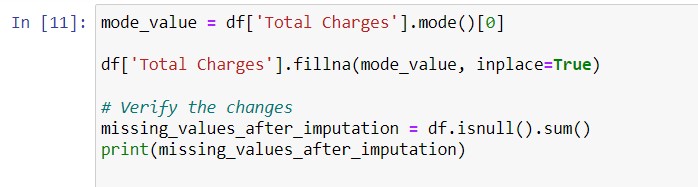
PART B: DATA PREPROCESSING

Identifying missing valuesA screenshot of a computer

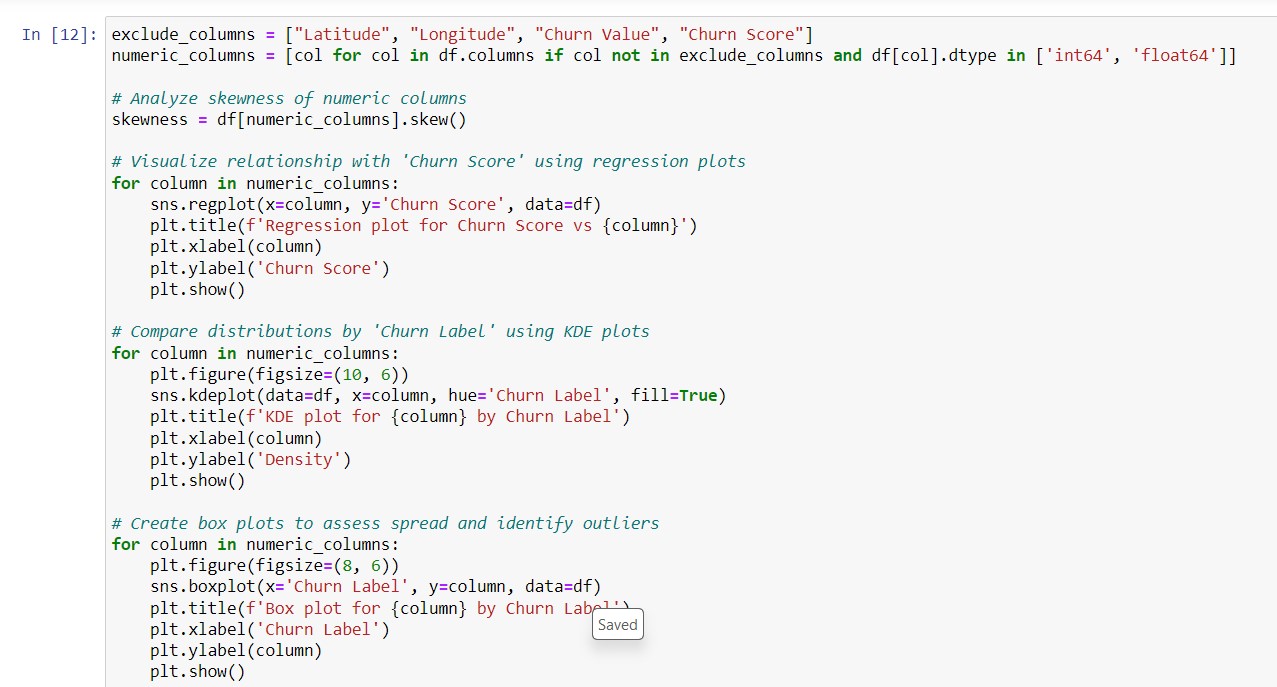
Description automatically generated

A close-up of a number

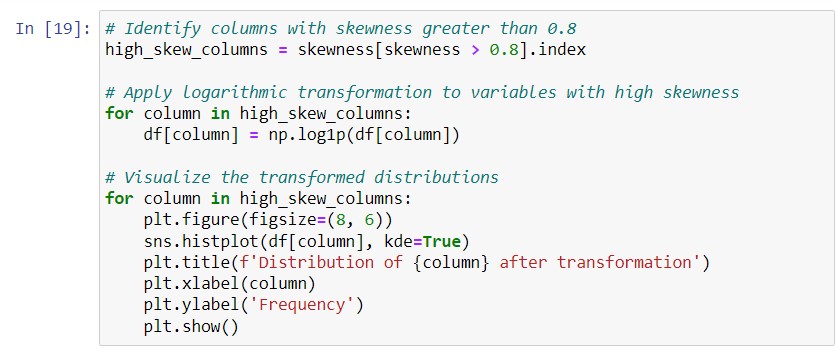
Description automatically generated

b) Using mode to remove missing values A screenshot of a computer

Description automatically generated

1. Analysing skewness of numeric columns
2. outliers:

Tenure Months and Total charges have outliers where churn = YES. Total charges has more outliers than Tenure.

1. Improving skewness 

A graph showing the distribution of a number of charges

Description automatically generated

PART C: Hypothesis testing



A screenshot of a computer

Description automatically generated

Conclusions

i) **Phone Service and Churn Relationship:**

* Null Hypothesis (H0): There is no significant relationship between Phone Service and Churn.
* Alternative Hypothesis (H1): There is a significant relationship between Phone Service and Churn.
* Chi-square statistic: 0.9150329892546948
* p-value: 0.3387825358066928
* **Conclusion:** The p-value (0.3387825358066928) is greater than the conventional significance level of 0.05. Therefore, we fail to reject the null hypothesis. This suggests that there is insufficient evidence to conclude that Phone Service has a significant relationship with Churn. Management can infer that offering or not offering phone service to customers may not directly impact churn rates based on this analysis.

ii) **Type of Contract and Likelihood of Churn:**

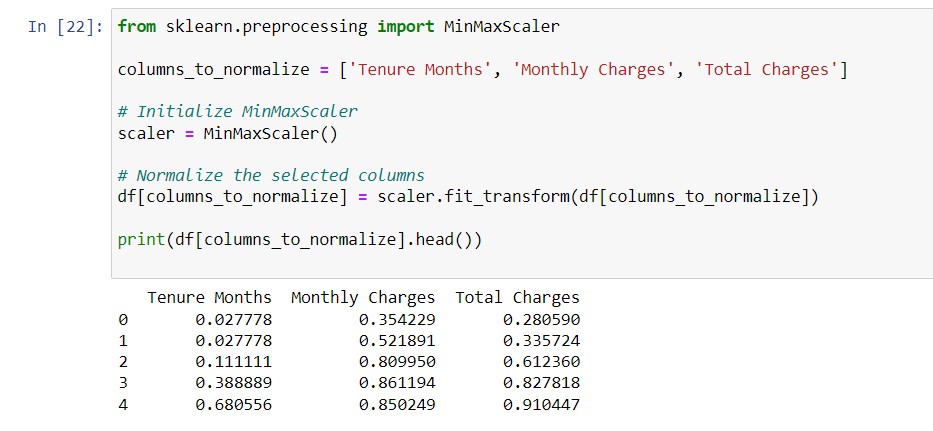
* Null Hypothesis (H0): The type of contract does not affect the likelihood of churn.
* Alternative Hypothesis (H1): The type of contract significantly influences the likelihood of churn.
* Chi-square statistic: 1184.5965720837926
* p-value: 5.863038300673391e-258
* **Conclusion:** The very low p-value (5.863038300673391e-258) is significantly less than the conventional significance level of 0.05. Therefore, we reject the null hypothesis, indicating that the type of contract has a significant influence on the likelihood of churn. Management should consider contract types when assessing and managing churn rates, as certain contract types may be more prone to churn than others.

iii) **Senior Citizen Status and Likelihood of Churn:**

* Null Hypothesis (H0): Senior Citizen status does not affect the likelihood of churn.
* Alternative Hypothesis (H1): Senior Citizen status significantly influences the likelihood of churn.
* Chi-square statistic: 159.42630036838742
* p-value: 1.510066805092378e-36
* **Conclusion:** The obtained p-value (1.510066805092378e-36) is significantly less than the conventional significance level of 0.05. Therefore, we reject the null hypothesis, indicating that Senior Citizen status has a significant impact on the likelihood of churn. Management should recognize that senior citizens may exhibit different churn behaviours compared to non-senior customers, requiring tailored retention strategies.

PART D: PREDICTIVE model development

* + 1. Normalization



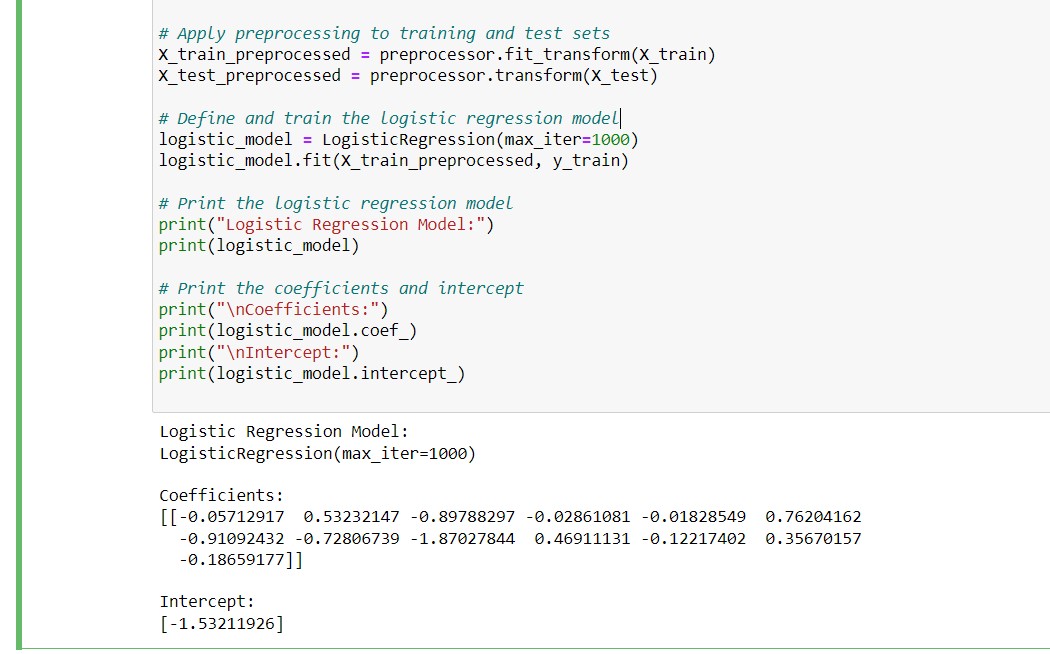
* + 1. Splitting the dataset into training and test sets



* + 1. Training Logistic Regression model

1. A computer screen shot of a computer code

   Description automatically generated



The target variable seems to be binary, as it's labelled as "Churn Label", which suggests it's indicating whether a customer churned or not (churned=1, not churned=0). In binary logistic regression, the model estimates the probability that an instance belongs to a particular class (e.g., churned or not churned).

1. **Logistic Regression Model**: LogisticRegression(max\_iter=1000)

This line indicates the type of model used, which is logistic regression. The parameters specified for the logistic regression model are **max\_iter=1000**, which sets the maximum number of iterations for the solver to converge. This means that the optimization algorithm will iterate up to 1000 times to find the optimal solution.

* **logistic\_model.fit(X\_train\_preprocessed, y\_train)**: This line trains the logistic regression model using the preprocessed training data (**X\_train\_preprocessed**) and the training labels (**y\_train**).

1. **Coefficients**: The coefficients represent the weights assigned to each feature by the logistic regression model. There's a single row of coefficients. Each element in this array corresponds to the weight of a specific feature. For example:
   * The coefficient for the first feature is approximately **-0.057**.
   * The coefficient for the second feature is approximately **0.532**.
   * The coefficient for the third feature is approximately **-0.898**, and so on.
2. **Intercept**: The intercept (or bias term) is added to the decision function of the logistic regression model. It represents the log-odds when all features are zero. In this output, the intercept is approximately **-1.532**.

Together, these coefficients and intercept define the logistic regression model's decision boundary. During prediction, the logistic regression model calculates the weighted sum of the input features (using the coefficients) and adds the intercept to produce a log-odds value. This log-odds value is then transformed into a probability using the logistic function, determining the likelihood of the target variable belonging to a specific class.

1. Classification report A screenshot of a computer

   Description automatically generated

* **No Class:**
  + **Precision (0.82):** Out of all the instances predicted as "No churn" by the model, 82% were actually "No churn".
  + **Recall (0.90):** Out of all the actual "No churn" instances, the model correctly identified 90%.
  + **F1-Score (0.86):** The harmonic mean of precision and recall, indicating a high overall performance for the "No churn" class.
* **Yes Class:**
  + **Precision (0.68):** Out of all the instances predicted as "Churn" by the model, 68% were “Churn”.
  + **Recall (0.51):** Out of all the actual "Churn" instances, the model correctly identified 51%.
  + **F1-Score (0.58):** The harmonic mean of precision and recall, indicating a moderate performance for the "Churn" class.
* **Overall Performance:**
  + **Accuracy (0.79):** The overall accuracy of the model, indicating that 79% of the total predictions were correct.
  + **Macro Average:**
    - **Precision (0.75):** The average precision across both classes, treating each class equally.
    - **Recall (0.70):** The average recall across both classes.
    - **F1-Score (0.72):** The average F1-score across both classes.
  + **Weighted Average:**
    - **Precision (0.78):** The precision weighted by the number of true instances for each class.
    - **Recall (0.79):** The recall weighted by the number of true instances for each class.
    - **F1-Score (0.78):** The F1-score weighted by the number of true instances for each class.

**Summary:**

* The model performs well in identifying customers who will not churn, with high precision, recall, and F1-score.
* The model's performance in identifying customers who will churn is moderate, with lower precision, recall, and F1-score.
* The overall accuracy of the model is good at 79%, but improving recall for the "Churn" class could enhance the model's usefulness for churn prediction.
  + 1. Confusion Matrix

A computer screen shot of a code

Description automatically generated

A graph showing the difference between confusion and confusion matrix

Description automatically generated

A confusion matrix is a table that is often used to describe the performance of a classification model. It compares the actual target values with those predicted by the model.

1. **True Negatives (TN = 913)**:
   * The model correctly predicted 913 instances as NO when they were actually NO.
2. **False Positives (FP = 96)**:
   * The model incorrectly predicted 96 instances as YES when they were actually NO.
   * These are also known as Type I errors.
3. **False Negatives (FN = 198)**:
   * The model incorrectly predicted 198 instances as NO when they were actually YES.
   * These are also known as Type II errors.
4. **True Positives (TP = 202)**:
   * The model correctly predicted 202 instances as YES when they were actually YES.