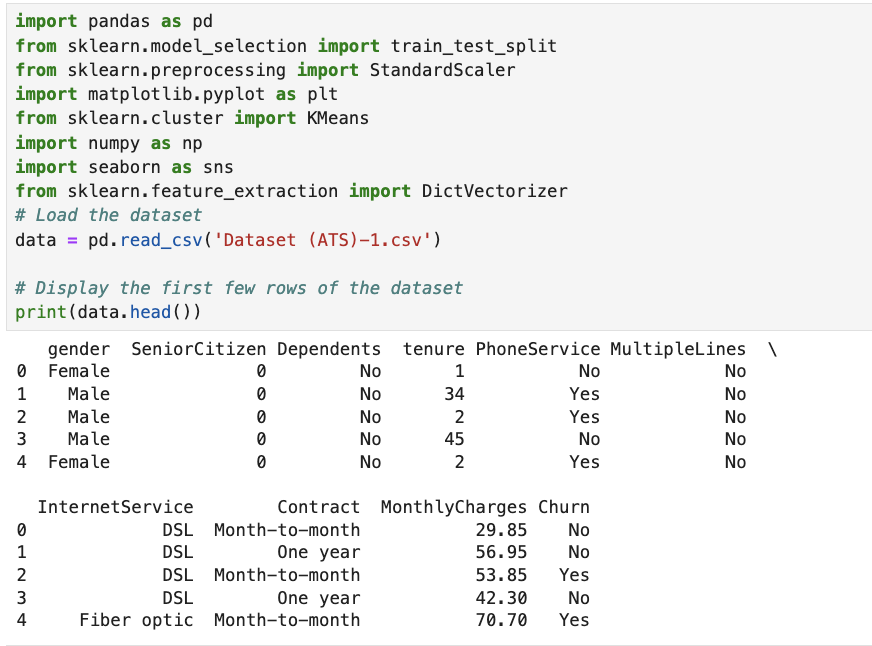
#### Report: Data Preparation

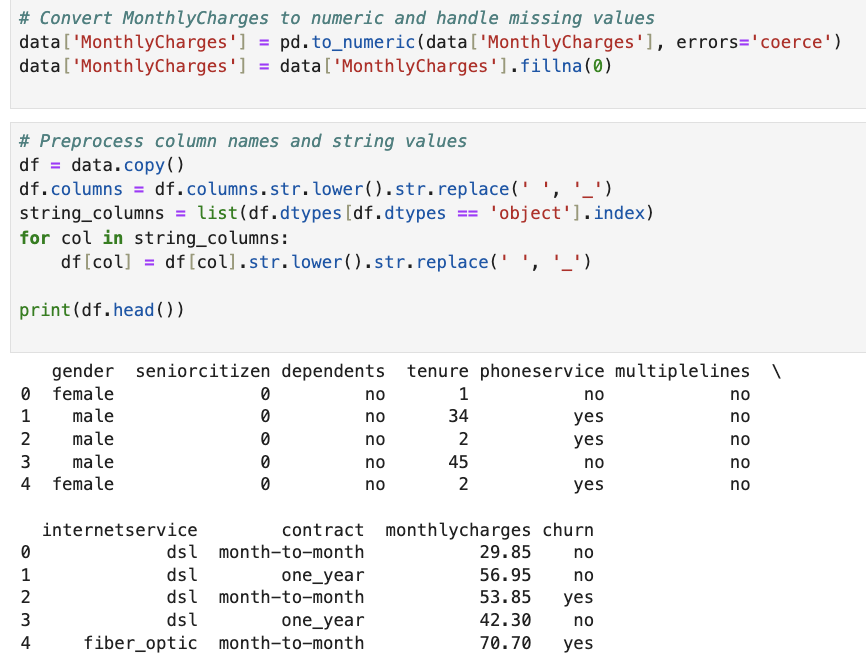
#### Initial Data Inspection

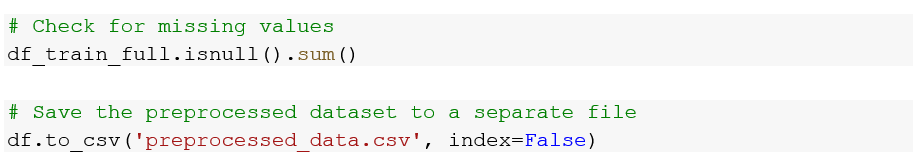
Upon loading the dataset and displaying the first few rows, we observe various columns such as 'gender', 'SeniorCitizen', 'Dependents', 'tenure', 'PhoneService', 'MultipleLines', 'InternetService', 'Contract', 'MonthlyCharges', and 'Churn'. This initial inspection helps us understand the structure and the type of data we are dealing with, highlighting both categorical and numerical features.



**Handling Missing Values and Data Preprocessing**

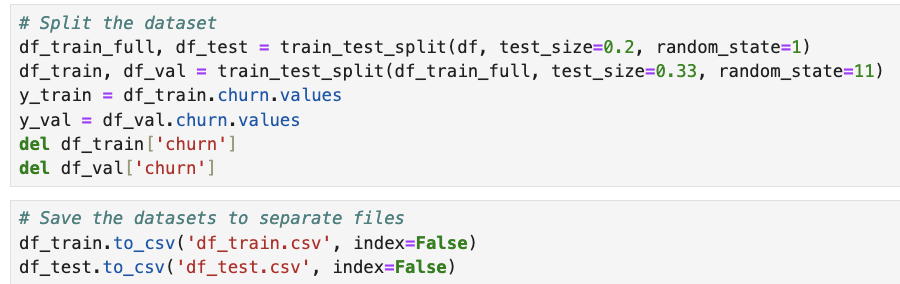
The 'MonthlyCharges' column is converted to numeric values, with non-numeric entries coerced into NaN and subsequently filled with 0. This ensures there are no missing values in this critical feature, which could otherwise affect model performance. Column names and string values are standardized to lowercase with underscores replacing spaces, ensuring consistency and ease of manipulation in further processing steps.





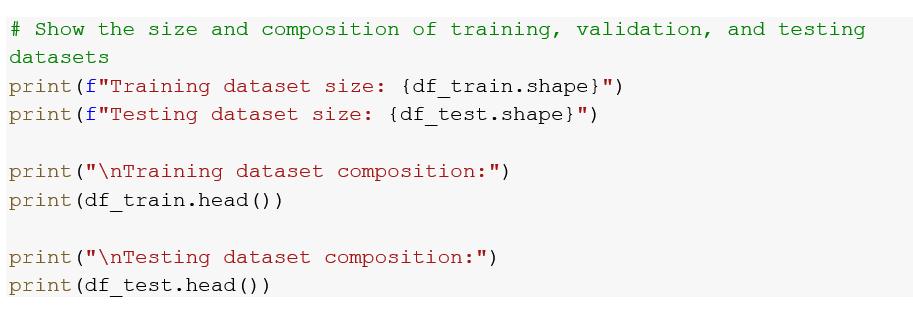
**Dataset Splitting**

The dataset is split into training (80%) and testing (20%) sets, with the training set further divided into training and validation subsets. This split ensures that we have distinct datasets for model training, validation, and final testing, thereby preventing overfitting and enabling an unbiased evaluation of the model.



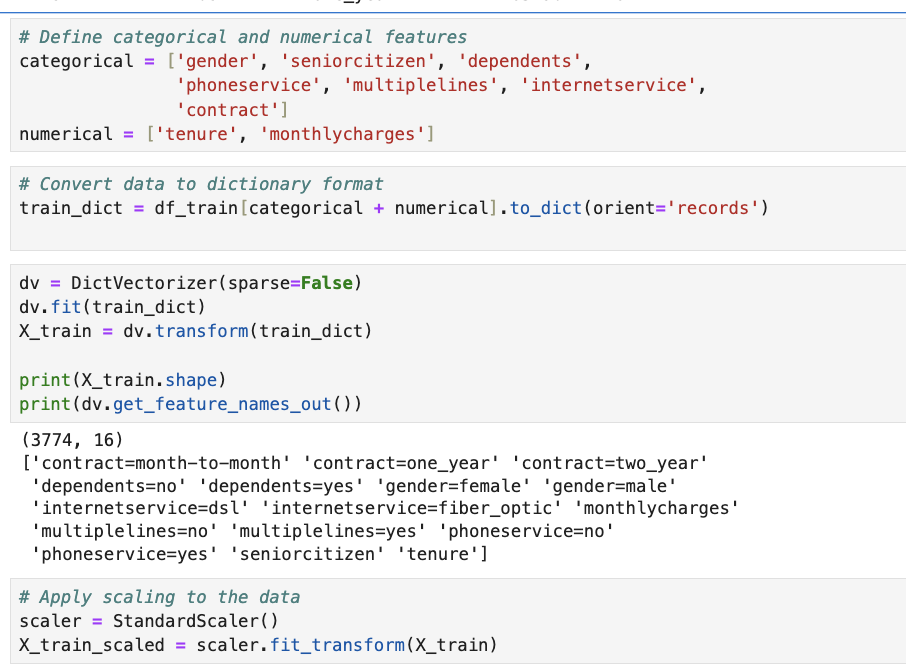
**Size and Composition:**

The Training dataset contains 3,774 samples with a diverse mix of gender, age (senior citizen status), dependents, tenure, phone service, multiple lines, internet service, contract types, and monthly charges. The Testing dataset consists of 1,409 samples with similar diversity in customer profiles, ensuring a representative evaluation set. It includes the 'churn' column for assessing model performance.



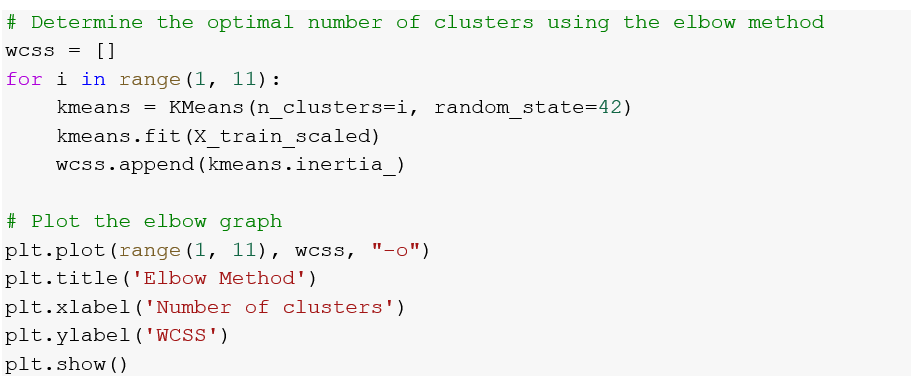
**Feature Vectorization and Scaling**

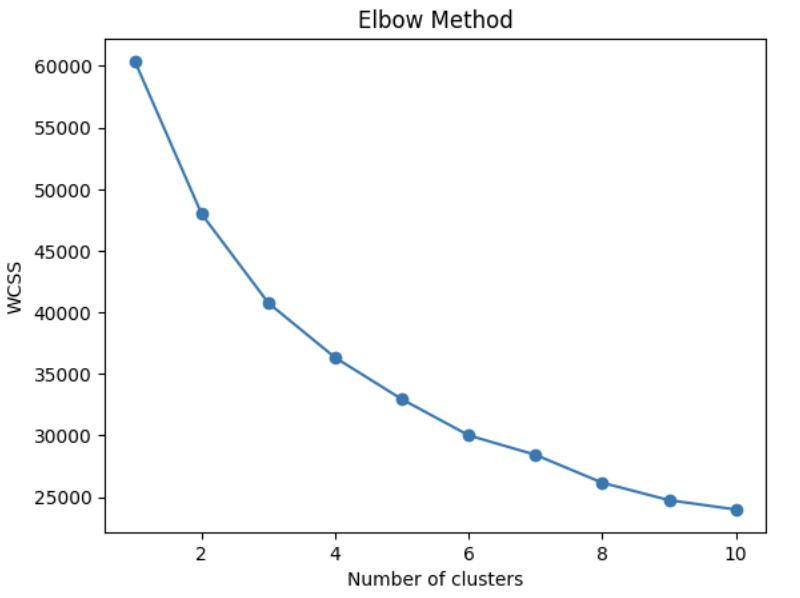
Categorical features are transformed into numerical representations using DictVectorizer, and the resulting matrix of features is standardized using StandardScaler. The scaling technique used in this analysis is **Standard Scaling**. This is implemented using the StandardScaler from the sklearn—preprocessing module. Standard scaling standardizes the features by removing the mean and scaling to unit variance, which ensures that all features contribute equally to the distance calculations in the K-Means clustering algorithm.

****

**Optimal Number of Clusters: The Elbow Method**

The graph created using the Elbow Method displays the Within-Cluster Sum of Squares (WCSS) in relation to the total number of clusters. WCSS falls with an increase in the number of clusters. It suggests the ideal number of clusters at the "elbow" point, where the rate of decrease sharply slows. The elbow in this instance can be seen at approximately four clusters.

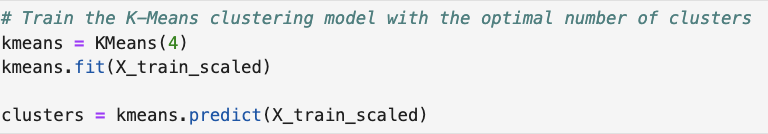




**Clustering**

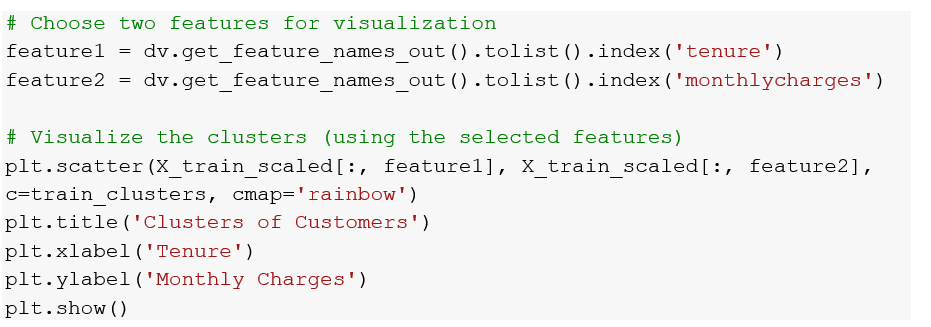
**K-Means Clustering**

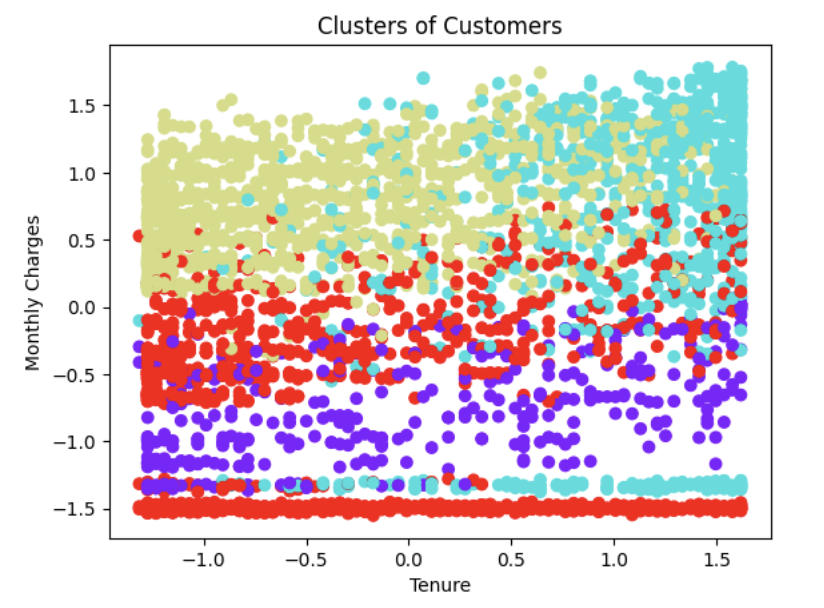
The K-Means model is trained with 4 clusters, as indicated by the Elbow Method. Each data point is assigned to one of these clusters, grouping similar data points based on the features.



**Cluster Visualization**

The scatter plot visualizes the resulting clusters of customers based on 'tenure' and 'monthly charges', segmented using the K-Means clustering algorithm into four distinct groups. Cluster 1 (Red) consists of customers with lower monthly charges and varying tenures, likely on basic plans. Cluster 2 (Yellow) includes customers with medium charges and a wide range of tenures, possibly on standard plans. Cluster 3 (Green) represents high-charge customers with diverse tenures, indicating premium plan users. Cluster 4 (Cyan) has medium to high charges and varying tenures, suggesting a mix of high-tier service users. This segmentation aids in targeting marketing strategies, designing retention programs, and identifying opportunities for service upgrades, thereby enhancing customer satisfaction and retention.





**Insights and Conclusions**

Standardizing features and handling missing values are critical steps that ensure the quality of the input data. The Elbow Method effectively determines the optimal number of clusters, balancing the trade-off between model complexity and accuracy. The K-Means clustering provides meaningful segmentation of customers, allowing for targeted marketing strategies and service improvements. Visualizing the clusters helps in understanding customer behavior and identifying distinct groups within the customer base. Overall, this analysis provides a comprehensive approach to data preprocessing, optimal cluster determination, and customer segmentation using K-Means clustering, offering valuable insights for business decision-making.