Shivani Ray

2023JULB01050

Customer Churn Rate

**Business Problem:**

Customer churn negatively impacts businesses by eroding revenue, increasing costs associated with acquiring new customers, and diminishing market share. For many organizations, understanding why customers leave and addressing those issues proactively remains a challenge. Without a systematic approach to identifying at-risk customers and analyzing churn trends, businesses struggle to retain valuable clients and maintain profitability.

This project addresses this business problem by leveraging Power BI Desktop and Excel to provide predictive churn modeling, customer segmentation, and actionable insights. These capabilities enable businesses to identify potential churn risks early, tailor retention strategies, and improve overall customer satisfaction and loyalty. The ultimate goal is to reduce customer loss, enhance decision-making, and foster long-term business sustainability.

**Data Requirements for Customer Churn Analysis and Their Importance**

To build a comprehensive model for predicting customer churn, a wide range of data is required. Each data type serves a specific purpose in understanding customer behavior, identifying at-risk customers, and implementing retention strategies. Here's a breakdown of why each data category is important:

**1. Customer Demographics**

* **Data Required**: Age, Gender, Location, Income Level, Marital Status
* **Why it's required**:  
  Demographic information helps identify patterns and trends within different customer segments. For example, younger customers may have different churn behavior compared to older ones, or customers from certain regions may show higher churn rates. This data can help personalize retention efforts and allow businesses to tailor offers, services, and communication strategies to specific groups.

**2. Customer Account Information**

* **Data Required**: Account Creation Date, Account Type, Tenure
* **Why it's required**:  
  Understanding how long customers have been with the company (tenure) helps determine whether long-term customers are at risk of leaving or if newer customers are more likely to churn. The account type (e.g., basic plan vs. premium plan) can also provide insights into the correlation between churn and service usage or subscription level. Early churn typically occurs within the first few months, so knowing when a customer joined is crucial to flagging potential issues early.

**3. Service Usage Data**

* **Data Required**: Frequency of Service Usage, Features/Services Used, Usage Patterns
* **Why it's required**:  
  A decrease in service usage or engagement is often a strong indicator of potential churn. Monitoring how often customers use a product or service, which features they engage with, and when they use them can reveal dissatisfaction or a shift in needs. For example, a customer who used to engage frequently with a service but now uses it infrequently may be signaling an intention to churn. Analyzing usage patterns can highlight these risks early.

**4. Interaction and Support Data**

* **Data Required**: Customer Support Tickets, Feedback/Complaints, Satisfaction Scores
* **Why it's required**:  
  Customer support interactions are often directly tied to satisfaction and retention. A high number of support tickets, unresolved complaints, or negative feedback can signal frustration, which often leads to churn. Satisfaction scores such as Net Promoter Score (NPS) or Customer Satisfaction (CSAT) can quantify the likelihood of a customer leaving. Identifying customers with low satisfaction scores can help in proactively addressing their issues before they decide to churn.

**5. Billing and Payment Data**

* **Data Required**: Payment Methods, Billing History, Refunds or Dispute Records
* **Why it's required**:  
  Financial data can highlight potential churn indicators, such as missed payments, overdue balances, or disputes over charges. A customer who regularly misses payments or has issues with billing is at risk of leaving. Conversely, customers who are timely with payments may be less likely to churn. Identifying financial friction points and resolving them can prevent churn, especially if financial issues are a major source of dissatisfaction.

**6. Behavioral Data**

* **Data Required**: Website/App Usage Metrics, Purchase Behavior
* **Why it's required**:  
  Behavioral data shows how customers interact with a business's digital properties or services. For example, customers who frequently visit a website or app but do not complete a purchase might show signs of disengagement, which could lead to churn. Understanding these behaviors allows businesses to create personalized strategies to re-engage customers. Additionally, a drop in purchase frequency or transaction value can indicate declining interest or satisfaction, making this data critical for churn prediction.

**7. Churn Labels (if available)**

* **Data Required**: Indicator of Churn, Date of Churn
* **Why it's required**:  
  Churn labels (whether a customer has left or is still active) are essential for supervised learning models in predictive analytics. These labels allow businesses to train machine learning algorithms to distinguish between factors that lead to churn and those that do not. By analyzing historical churn data, businesses can predict future churn behavior, enabling more effective retention strategies. The date of churn provides context and allows businesses to identify the lead time before a customer churns, helping with early intervention.

**8. External Data (Optional)**

* **Data Required**: Market Trends, Competitor Data, Macroeconomic Indicators
* **Why it's required**:  
  External factors can significantly impact customer behavior and churn. For example, shifts in the economy or industry trends may affect a customer's purchasing ability or decision to stay with a provider. Competitor data can help understand if customers are leaving for better offers or services elsewhere. Incorporating external data provides a broader context and enhances the predictive power of churn models, ensuring that businesses stay ahead of market dynamics that may influence customer decisions.

Each category of data plays a pivotal role in identifying the root causes of churn and providing actionable insights. By combining demographic, behavioral, financial, and interaction data, businesses can gain a 360-degree view of their customers. This comprehensive data set enables predictive models to forecast which customers are at risk and why, allowing companies to take preemptive actions to retain them, ultimately improving customer satisfaction and reducing churn.

The dataset contains information on 7,043 customers and 23 variables.

**Data Collection Overview**

This dataset provides comprehensive information about customers, their behavior, service usage, and churn status. The data is structured with the following key attributes:

**Columns and Their Description:**

1. **customerID**: Unique identifier for each customer.
   * **Purpose**: Ensures each record is distinct and trackable.
2. **gender**: Gender of the customer (Male/Female).
   * **Purpose**: Useful for understanding demographic patterns in churn.
3. **SeniorCitizen**: Indicates if the customer is a senior citizen (0 = No, 1 = Yes).
   * **Purpose**: Helps identify if age influences churn behavior.
4. **Partner**: Indicates if the customer has a partner (Yes/No).
   * **Purpose**: Understands if relationships affect customer loyalty.
5. **Dependents**: Indicates if the customer has dependents (Yes/No).
   * **Purpose**: Analyzes household dynamics affecting retention.
6. **tenure**: Number of months the customer has stayed with the company.
   * **Purpose**: Determines loyalty trends and retention duration.
7. **PhoneService**: Indicates if the customer has a phone service (Yes/No).
   * **Purpose**: Explores the relationship between service types and churn.
8. **MultipleLines**: Indicates if the customer has multiple phone lines (Yes/No).
   * **Purpose**: Assesses how product bundling impacts churn.
9. **InternetService**: Type of internet service used (DSL, Fiber optic, None).
   * **Purpose**: Helps analyze service-specific churn trends.
10. **OnlineSecurity**: Indicates if online security services are enabled (Yes/No).
    * **Purpose**: Determines whether additional services impact customer retention.
11. **OnlineBackup**: Indicates if online backup services are enabled (Yes/No).
    * **Purpose**: Explores the influence of ancillary services on churn.
12. **DeviceProtection**: Indicates if device protection is enabled (Yes/No).
    * **Purpose**: Examines the role of supplementary features in retention.
13. **TechSupport**: Indicates if technical support services are subscribed (Yes/No).
    * **Purpose**: Highlights the importance of support in customer satisfaction.
14. **StreamingTV**: Indicates if streaming TV is subscribed (Yes/No).
    * **Purpose**: Evaluates entertainment services’ contribution to loyalty.
15. **StreamingMovies**: Indicates if streaming movies are subscribed (Yes/No).
    * **Purpose**: Similar to streaming TV for entertainment preferences.
16. **Contract**: Type of contract (Month-to-month, One year, Two year).
    * **Purpose**: Captures the impact of contract length on churn rates.
17. **PaperlessBilling**: Indicates if billing is paperless (Yes/No).
    * **Purpose**: Studies customer preferences for convenience and its link to retention.
18. **PaymentMethod**: Method of payment (Electronic check, Mailed check, etc.).
    * **Purpose**: Identifies preferred payment options and their correlation with churn.
19. **MonthlyCharges**: Monthly charges incurred by the customer.
    * **Purpose**: Helps detect pricing sensitivity and its role in churn.
20. **TotalCharges**: Total charges incurred during the tenure.
    * **Purpose**: Aggregated measure to understand long-term spending behavior.
21. **numAdminTickets**: Number of administrative support tickets.
    * **Purpose**: Reflects customer satisfaction and potential issues.
22. **numTechTickets**: Number of technical support tickets.
    * **Purpose**: Measures technical challenges impacting customer retention.
23. **Churn**: Whether the customer has churned (Yes/No).
    * **Purpose**: The target variable for churn prediction analysis.

**Data Understanding and Insights**

**General Observations:**

* **Demographics**: Variables like gender, SeniorCitizen, Partner, and Dependents help profile customer segments.
* **Behavioral Patterns**: Service-related data like tenure, MonthlyCharges, and InternetService provide insights into usage trends.
* **Customer Interactions**: Columns like numAdminTickets and numTechTickets highlight customer touchpoints with the company.

**Potential Insights:**

1. Customers with **month-to-month contracts** may exhibit higher churn compared to those with long-term contracts.
2. Customers with multiple support tickets might churn due to unresolved issues.
3. Senior citizens or customers with high MonthlyCharges might be more price-sensitive, leading to churn.

**Data Collection for Understanding Customer Churn**

Data collection is a foundational step in understanding customer churn. It involves gathering relevant data from multiple sources, ensuring the dataset comprehensively captures all aspects of customer interactions, behavior, and satisfaction. Here’s an in-depth overview of how data collection should be approached for customer churn analysis:

**1. Identifying Data Sources**

Data can come from various internal and external sources. Identifying and consolidating these data sources ensures a holistic understanding of customer churn patterns.

**Internal Sources:**

* **CRM Systems**: Provide customer demographics, account details, and interaction history.
* **Billing Systems**: Contain financial data, including payment methods, billing cycles, and refunds.
* **Service Usage Logs**: Capture how customers use the product or service, including frequency and engagement levels.
* **Support Databases**: Record customer support tickets, complaints, and resolution times.
* **Survey Tools**: Gather customer feedback, satisfaction scores, and NPS.
* **Website/App Analytics**: Offer insights into behavioral data, such as browsing patterns, click rates, and time spent on platforms.

**External Sources:**

* **Market Research Reports**: Provide insights into industry trends affecting customer retention.
* **Competitor Analysis**: Identify offers or services from competitors that could be driving churn.
* **Macroeconomic Data**: Include indicators like inflation rates, which may influence purchasing decisions.

**2. Defining Data Requirements**

It is essential to clearly define the data attributes needed to analyze customer churn. These should align with the business problems and the specific objectives of the churn analysis.

**Key Considerations:**

* **Relevance**: Ensure all collected data is pertinent to churn analysis.
* **Accuracy**: Data should reflect the true state of customer behavior and interactions.
* **Timeliness**: Data must be up-to-date to capture recent trends and behaviors.
* **Completeness**: Avoid missing values to ensure reliable analysis.

**3. Data Collection Methods**

The methods used for data collection depend on the type of data and its source. Below are common techniques:

**Automated Data Pipelines:**

* Use tools like **ETL (Extract, Transform, Load)** processes to pull data from various databases and systems.
* Example: Connecting Power BI to CRM or billing systems using APIs for real-time data extraction.

**Manual Data Entry:**

* For data that isn’t digitized, such as surveys or offline interactions, manual data entry may be required.
* Ensure quality checks to minimize errors.

**Third-Party Integrations:**

* Integrate external datasets, such as competitor analysis, using APIs or purchased datasets.

**Customer Surveys and Feedback:**

* Conduct surveys to understand customer satisfaction and gather qualitative insights.

**4. Structuring the Data**

Once the data is collected, it should be structured into a unified format to ensure consistency and ease of analysis.

**Common Formats:**

* **Flat Files**: CSV or Excel sheets for raw data.
* **Relational Databases**: SQL databases for structured storage.
* **Cloud Storage**: Platforms like Azure or AWS for scalability.

**Key Steps in Structuring:**

* Define unique identifiers (e.g., Customer ID) to link data across sources.
* Categorize data into meaningful segments (e.g., demographic, financial, behavioral).
* Ensure a consistent format for dates, currencies, and other standardized fields.

**5. Validating Data During Collection**

Data validation is crucial to ensure the quality and reliability of the collected data.

**Steps to Validate Data:**

* **Consistency Checks**: Verify that data is consistent across sources (e.g., matching customer IDs in CRM and billing systems).
* **Duplicate Removal**: Identify and eliminate duplicate entries to avoid skewed analysis.
* **Outlier Detection**: Highlight unusual values that could indicate errors.
* **Cross-Verification**: Match data against known benchmarks or external sources for accuracy.

**6. Tools for Data Collection and Integration**

* **Power BI**: Allows seamless integration of data from multiple sources using built-in connectors.
* **Excel**: Useful for manual entry and preliminary data organization.
* **Database Management Tools**: SQL or NoSQL databases for storing structured data.
* **ETL Tools**: Tools like Talend or SSIS for automating data extraction and transformation.

**7. Challenges in Data Collection**

* **Data Silos**: Information scattered across departments or systems, making integration difficult.
* **Incomplete Data**: Missing values that reduce the dataset's usability.
* **Bias in Data**: Skewed datasets that may not represent the entire customer base.
* **Data Privacy**: Ensuring compliance with regulations like GDPR or CCPA when collecting personal data.

**8. Ensuring Data Security and Privacy**

* **Encryption**: Protect data during transfer and storage.
* **Access Control**: Limit data access to authorized personnel only.
* **Anonymization**: Remove personally identifiable information to protect customer privacy.

**9. Iterative Data Collection**

Data collection should be an ongoing process. Regular updates and refinements ensure the dataset remains relevant for future churn analysis.

Effective data collection is the cornerstone of understanding customer churn. By integrating data from diverse sources and ensuring its accuracy, reliability, and relevance, businesses can gain deep insights into customer behavior. This understanding lays the groundwork for creating predictive models and actionable strategies to reduce churn and improve customer satisfaction.

**Data Validation for Customer Churn Analysis**

Data validation is a critical step in ensuring the quality, accuracy, and reliability of data for analysis. Proper validation minimizes errors, reduces bias, and ensures that insights drawn from the data are trustworthy. Below is a detailed explanation of the data validation process for the provided dataset.

**1. Purpose of Data Validation**

The primary goals of data validation are:

1. Ensuring data accuracy and consistency.
2. Detecting and handling missing or invalid data points.
3. Minimizing biases to improve analysis and predictive model performance.
4. Guaranteeing data reliability for meaningful decision-making.

**2. Data Validation Process**

**Step 1: Data Integrity Checks**

Data integrity ensures the structure and logical consistency of the dataset.

* **Duplicate Records**:
  + Check for duplicate rows by verifying the customerID field. Each customerID must be unique.
  + Handle duplicates by retaining only one record per customer, if duplicates exist.
* **Missing Data**:
  + Identify missing values in critical columns such as TotalCharges, tenure, and MonthlyCharges.
  + Methods for handling missing data:
    - **Imputation**: Replace missing numerical values with the mean, median, or mode. For example, missing TotalCharges can be computed based on tenure and MonthlyCharges.
    - **Removal**: If a row contains too many missing values, consider removing it.
* **Invalid Data**:
  + Check for invalid entries in categorical fields (e.g., gender, InternetService). Ensure all values are within expected categories.
  + For numerical fields, verify that values are within realistic ranges. For instance, MonthlyCharges should not be negative.

**Step 2: Data Type Validation**

Data types must align with the nature of the variables.

* **Conversion of TotalCharges**:
  + TotalCharges is currently an object but should be numeric. Convert the column to a numeric type after addressing non-numeric values (e.g., blanks).
  + Handle errors during conversion by flagging invalid entries and either imputing or removing them.
* **Categorical Variables**:
  + Columns like gender, Partner, and Churn should remain as categorical data. Encode them for analysis if required.

**Step 3: Logical Consistency Checks**

Logical consistency ensures that relationships between variables make sense.

* **Relationship Between tenure and TotalCharges**:
  + Validate that TotalCharges is proportional to MonthlyCharges and tenure. For example, a customer with a tenure of 0 should have minimal or no total charges.
* **Contract Type and Tenure**:
  + Customers with longer contracts (e.g., one year, two years) should have a minimum tenure consistent with their contract length.

**Step 4: Outlier Detection**

Outliers are extreme values that may distort analysis.

* **Numerical Variables**:
  + Use statistical methods like the Z-score or Interquartile Range (IQR) to identify outliers in MonthlyCharges and TotalCharges.
  + Investigate outliers to determine if they are valid (e.g., high spenders) or errors requiring correction.
* **Categorical Variables**:
  + Check for rare or unexpected categories in fields like InternetService and PaymentMethod. Rare categories might indicate errors in data entry.

**Step 5: Bias and Fairness Validation**

Data bias can lead to unfair or misleading conclusions.

* **Demographic Bias**:
  + Assess whether certain demographics, such as gender or SeniorCitizen, are disproportionately represented. For instance, if senior citizens form a small fraction of the dataset, models may not predict their behavior accurately.
* **Churn Label Distribution**:
  + Check for imbalance in the Churn variable. If one class (e.g., "No") is overrepresented, resampling methods like oversampling, undersampling, or Synthetic Minority Oversampling Technique (SMOTE) may be necessary.

**Step 6: Cross-Validation with Business Rules**

Validate the dataset against known business rules to ensure reliability.

* Customers with No InternetService should have "No" for all related services (e.g., OnlineSecurity, OnlineBackup).
* Customers with No PhoneService should have "No phone service" for MultipleLines.
* Ensure that churned customers have valid data for all periods before their churn date.

**Step 7: Statistical Validation**

Perform basic statistical analyses to confirm data consistency.

* **Distribution Analysis**:
  + Review the distributions of numerical variables (tenure, MonthlyCharges) to ensure they align with expected patterns.
* **Correlation Analysis**:
  + Identify relationships between variables. For example, validate that tenure has a negative correlation with churn likelihood.

**3. Tools and Methods for Data Validation**

* **Power BI**:
  + Use data profiling tools in Power BI to check for missing values, duplicates, and data types.
  + Create data quality dashboards to visualize anomalies and inconsistencies.
* **Python/Pandas**:
  + Use Python libraries like Pandas and NumPy to automate data validation checks.
* **SQL Queries**:
  + Run SQL queries for consistency checks across large datasets.

**4. Common Challenges in Data Validation**

1. **Incomplete Data Documentation**: Without proper documentation, it’s difficult to verify field definitions and valid ranges.
2. **High Proportion of Missing Data**: Large-scale imputation or removal may lead to biased results.
3. **Balancing Data Quality and Volume**: Over-cleaning may remove valuable data, while under-cleaning risks analysis errors.

**5. Summary of Data Validation Steps**

1. Perform **integrity checks** for duplicates and missing data.
2. Validate **data types** and convert as needed.
3. Ensure **logical consistency** between related fields.
4. Detect and address **outliers** in numerical data.
5. Assess **bias** and ensure fairness in representation.
6. Cross-validate with **business rules** for alignment.
7. Use **statistical analysis** to confirm reliability.

By rigorously validating the data, we ensure it is accurate, consistent, and ready for analysis. This improves the quality of insights drawn from the dataset, leading to more reliable churn predictions and actionable business strategies.

**In-Depth Data Cleaning in Power BI for Customer Churn Analysis**

Data cleaning is an essential step to ensure that your dataset is accurate, consistent, and ready for meaningful analysis. Power BI provides a robust set of tools to clean and transform your data using the Power Query Editor. Here's an extensive guide to cleaning the **Customer Churn Dataset** step-by-step in Power BI.

**1. Load the Data into Power BI**

1. **Open Power BI Desktop**:  
   Launch Power BI Desktop to begin the process of data cleaning.
2. **Import Data**:
   * Click on **Home > Get Data > Excel**.
   * Select the **Customer Churn Dataset** file and load the appropriate sheet (in this case, 01 Churn-Dataset).
   * Once the data is loaded, click on **Transform Data** to open the **Power Query Editor**, where data cleaning will take place.

**2. Handle Missing Values**

**Missing values** can significantly impact analysis and predictive modeling, as missing data can lead to incorrect conclusions. Here's how to deal with them:

1. **Identify Missing Values**:
   * In Power Query Editor, you can use the **Column Quality** feature (found under the **View** tab) to check the **null** or missing values in each column. This feature visually shows how much of each column is filled with valid data.
   * Use **Remove Duplicates** and **Remove Empty** options to clear out empty values in key fields like TotalCharges, tenure, or MonthlyCharges.
2. **Handle Missing Data in Numerical Columns**:
   * For fields like TotalCharges, which may have missing or incorrect values (e.g., blank cells in some rows), you can compute them based on other columns (such as tenure multiplied by MonthlyCharges).
   * To replace missing values, right-click the column (e.g., TotalCharges), choose **Replace Values**, and insert a calculated value.
   * Alternatively, you can replace missing data in a numerical column with the **median** or **mean**. For example, you can fill missing MonthlyCharges with the median value of that column.

**Example Steps**:

* + Select TotalCharges > **Transform > Fill Down** (to propagate values from the rows above).
  + Alternatively, for median imputation, you could calculate the median and replace missing entries manually.

1. **Handle Missing Data in Categorical Columns**:
   * Categorical variables like Partner, Dependents, PhoneService, and Churn can have missing values, too. These can be imputed by replacing missing data with the **mode** (the most frequent value).
   * To do this, right-click on the column (e.g., Partner) and choose **Replace Values** to substitute missing values with the most common value.
2. **Remove Rows with Excessive Missing Values**:
   * For rows where a large portion of columns have missing values, consider removing them from the dataset.
   * Use the **Remove Rows** function under **Home** and filter rows with excessive missing values by specifying a percentage threshold (e.g., 50% missing data).

**3. Standardize Data Types**

Ensuring that data is stored in the correct format is crucial for analysis. In Power BI, you can change the data types directly in Power Query Editor.

1. **Converting Columns to Appropriate Data Types**:
   * **Numerical Columns**: For instance, TotalCharges may be incorrectly stored as an object type (string). Convert it to a **Decimal Number**.
     + Right-click on the column and choose **Change Type > Decimal Number**.
   * **Date Columns**: If you have any date fields (e.g., customer subscription start date), make sure they are properly formatted as **Date**.
   * **Categorical Columns**: Fields like gender, Partner, and Churn should be set to **Text**. You can also convert them into **Boolean** (True/False) or **binary (0/1)** for machine learning purposes.

**Steps for Conversion**:

* + For TotalCharges: Right-click the column > **Change Type > Decimal Number**.
  + For categorical fields: Right-click the column > **Change Type > Text**.

**4. Remove Duplicate Records**

Duplicates can distort analysis by causing overrepresentation of some customers, leading to incorrect conclusions.

1. **Detect and Remove Duplicates**:
   * Use the **Remove Duplicates** option to ensure that each customer appears only once in the dataset.
   * Right-click the column containing unique identifiers, such as customerID, and select **Remove Duplicates**.

**Steps**:

* + Select the customerID column, then click on **Remove Duplicates** under the **Home** tab.

**5. Handling Outliers**

Outliers are extreme values that could skew the analysis, especially in numerical columns like MonthlyCharges and TotalCharges. Identifying and handling outliers is critical to building accurate models.

1. **Visualizing Outliers**:
   * Use visualizations such as **box plots** or **scatter plots** in Power BI to identify potential outliers in MonthlyCharges, TotalCharges, and other numerical variables.
   * You can create a box plot visualization to visually detect data points that are far away from the rest of the distribution.
2. **Filter or Adjust Outliers**:
   * Once outliers are identified, you can either **remove** them or **cap** them to a maximum threshold value. For instance, if MonthlyCharges exceeds an expected range, you can limit it to a certain value.

**Steps for Removing Outliers**:

* + Use **Transform > Filter Rows** to apply filters for columns like MonthlyCharges or TotalCharges to remove values outside a logical range.

**6. Standardizing Text Data**

Inconsistent text data can lead to misinterpretation during analysis. This is especially relevant in categorical columns like PaymentMethod, InternetService, or Contract.

1. **Normalize Text Case**:
   * To ensure consistency in text columns like PaymentMethod or InternetService, standardize the case (lowercase or uppercase) for all values.
   * You can do this by selecting the column and using the **Transform > Format > Uppercase/Lowercase** function.
2. **Remove Extra Spaces**:
   * Extra spaces in text fields (e.g., leading/trailing spaces) can cause issues during analysis. You can remove these by selecting **Transform > Format > Trim**.

**7. Logical Inconsistencies and Validation**

Power BI allows you to apply logical checks to ensure that values across different columns are consistent with one another. For example, a customer with No InternetService should not have OnlineSecurity or StreamingTV.

1. **Cross-Column Validation**:
   * For customers with No InternetService, set all related fields like OnlineSecurity, OnlineBackup, and StreamingTV to "No".
   * Create a new column using **Add Column > Conditional Column** to validate and ensure that logical relationships hold true across columns.

**Example Rule**:

plaintext

Copy code

If [InternetService] = "No" Then "No" Else [OnlineSecurity]

1. **Address Logical Errors**:
   * Use **Conditional Column** to create custom rules. For instance, if a customer has "No PhoneService", the MultipleLines column should also be set to "No phone service".

**8. Recode and Categorize Data**

For better analysis, particularly when preparing for predictive modeling, you may want to group continuous variables (e.g., tenure) into categories or encode categorical variables as numerical values.

1. **Bin Continuous Variables**:
   * Group customers into categories such as "New Customer", "Loyal Customer", or "At Risk" by binning tenure into ranges.
   * To bin tenure, select the column, go to **Add Column > Conditional Column**, and define rules to categorize tenure (e.g., 0-12 months = "New", 13-24 months = "Loyal").
2. **Encode Categorical Variables**:
   * You may convert categorical variables like gender or Churn into numerical values for machine learning purposes.
   * Use **Transform > Replace Values** to map categories like "Male" = 1, "Female" = 0, and "Yes" = 1, "No" = 0 for binary encoding.

**9. Save the Cleaned Data**

After applying all cleaning steps:

1. Click **Close & Apply** in Power Query Editor to load the cleaned data back into Power BI.
2. You can now create visualizations, dashboards, and perform detailed analysis using the cleaned dataset.

**10. Summary of Power BI Cleaning Tools**

1. **Transform**: Change data types, fill missing values, and replace data.
2. **Remove Duplicates**: Remove duplicate records using unique identifiers like customerID.
3. **Conditional Columns**: Apply business logic to ensure data consistency.
4. **Column Quality**: Visualize missing, invalid, or incomplete data.
5. **Data Profiling**: Assess the quality and distribution of each column.

By following the detailed steps outlined above, you ensure that the **Customer Churn Dataset** is free of inconsistencies, duplicates, missing values, and errors. After cleaning, the dataset will be well-prepared for analysis and predictive modeling, leading to more accurate churn predictions and actionable insights to improve customer retention.

**Tool Selection: Power BI**

**Power BI** is selected for this customer churn analysis due to its user-friendly interface, powerful data transformation capabilities, and the ability to create dynamic visualizations. Here's why Power BI is ideal for this analysis:

1. **Data Cleaning & Transformation**: Power BI's Power Query Editor allows for seamless data cleaning, including handling missing values, correcting data types, removing duplicates, and transforming data for analysis.
2. **Interactive Visualizations**: Power BI provides various visual tools, such as bar charts, line charts, scatter plots, and more, that allow users to interact with data dynamically and drill down into insights.
3. **Data Integration**: Power BI supports easy integration with multiple data sources like Excel, SQL databases, and web-based services, which allows for flexible and scalable analysis.
4. **Dashboarding**: Power BI makes it easy to create dashboards that display multiple key performance indicators (KPIs) and trends, giving a comprehensive view of the churn analysis.

**Graphs and Charts Used**

Power BI provides various visualization types that help tell a detailed story about customer churn. Here's a breakdown of different charts and graphs used for **univariate**, **bivariate**, and **multivariate** analysis:

**Univariate Analysis (Single variable analysis)**

1. **Bar Charts**:
   * Used to visualize the frequency distribution of categorical variables like gender, Partner, Churn, and PaymentMethod. This helps in identifying the distribution and proportions of different groups.
2. **Histograms**:
   * Used for numerical variables like MonthlyCharges, TotalCharges, and tenure. Histograms help understand the distribution of these variables, highlighting trends like pricing sensitivity, length of customer engagement, and overall spending patterns.
3. **Pie Charts**:
   * Used to show the proportion of churned vs. non-churned customers. This provides an immediate understanding of the churn rate in the customer base.

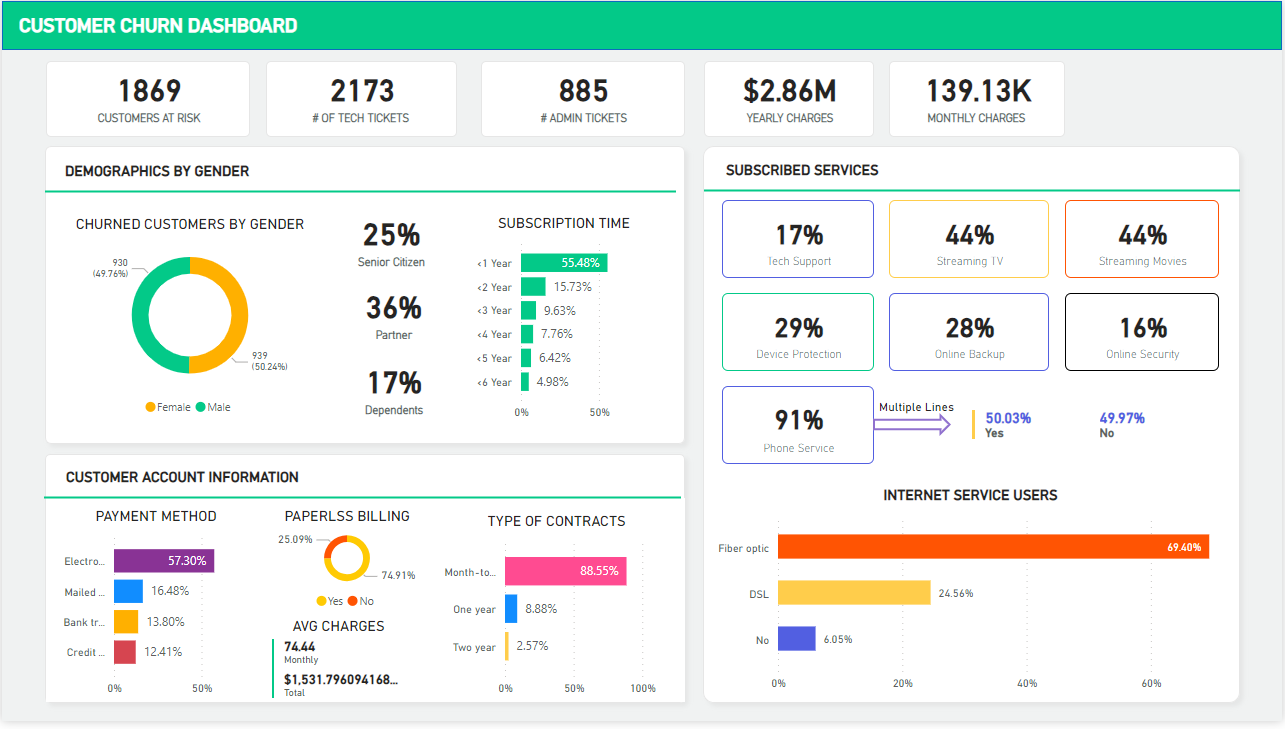
**Bivariate Analysis (Two-variable relationship analysis)**

1. **Stacked Bar Charts**:
   * Used to compare the relationship between categorical variables (e.g., Churn vs. Partner, Churn vs. InternetService). This helps identify how different categories (e.g., customers with a partner) are more or less likely to churn.
2. **Scatter Plots**:
   * Used to visualize relationships between two numerical variables, such as MonthlyCharges vs. tenure. This can help identify trends such as higher churn rates among customers with higher monthly charges or shorter tenure.
3. **Box Plots**:
   * Used to compare the distribution of numerical variables (e.g., MonthlyCharges, TotalCharges) across categorical variables (e.g., Churn). This reveals whether customers who churned have significantly different spending patterns compared to those who stayed.

**Multivariate Analysis (Analysis involving more than two variables)**

1. **Heatmaps**:
   * Used to show correlations between multiple numerical variables, such as tenure, MonthlyCharges, and TotalCharges. A heatmap makes it easy to identify relationships and trends across a wide range of variables.
2. **Clustered Bar Charts**:
   * These are used to show the relationship between several categorical variables, like Contract, Churn, and PaymentMethod. This can help identify patterns like whether month-to-month contracts result in higher churn compared to one-year or two-year contracts.
3. **Bubble Charts**:
   * Used to visualize the relationship between three variables, like MonthlyCharges, TotalCharges, and tenure. The size of the bubbles can indicate the number of customers in each segment, while the axes show their respective values.

**Customer Churn Dashboard**



**What Story Does the Dashboard Tell?**

The dashboard tells a comprehensive and data-driven story about customer churn, focusing on factors influencing retention and highlighting strategies for improvement. Here’s a detailed breakdown of the narrative the dashboard communicates:

**1. The Churn Rate:**

* The first key element of the story is the **overall churn rate**, which shows how many customers have left the service. A pie chart reveals the percentage of customers who have churned versus those who have remained.
* **Insight**: A high churn rate might indicate problems with customer satisfaction, pricing, or service quality. This is the starting point for investigating further.

**2. Customer Demographics and Behavior:**

* The dashboard shows a breakdown of demographic variables such as gender, SeniorCitizen, and Partner to analyze how these groups are related to churn.
* **Insight**: For instance, if senior citizens have a higher churn rate, it might suggest that they face difficulties with the service or are less tech-savvy. Customers without partners may also show higher churn, potentially due to changes in household dynamics or pricing sensitivity.

**3. Pricing Sensitivity and Contract Type:**

* The **MonthlyCharges** distribution and **Contract Type** analysis are crucial. Stacked bar charts compare churn rates across different contract types (Month-to-month, One year, Two year) and monthly charge ranges.
* **Insight**: A common trend could be that customers on month-to-month contracts have higher churn rates compared to those on long-term contracts. Additionally, customers with lower MonthlyCharges may have a lower tendency to churn, suggesting pricing plays a role in retention.

**4. Length of Customer Relationship (Tenure):**

* A **histogram** or **box plot** for tenure compares the length of time customers have stayed with the company. This shows that customers who stay longer are less likely to churn.
* **Insight**: Customers with shorter tenures (e.g., under 6 months) might be at a higher risk of churn, which suggests that early-stage engagement and onboarding strategies could be critical.

**5. Usage of Additional Services:**

* A comparison of customers who subscribed to services like **TechSupport**, **StreamingTV**, or **OnlineSecurity** reveals that those using more services tend to have a lower churn rate.
* **Insight**: Offering additional services or creating product bundles might increase customer loyalty and reduce churn.

**6. Customer Support Interactions:**

* **Number of Support Tickets**: A key part of the story is the relationship between customer support interactions and churn. The dashboard shows that customers who have raised more support tickets (both technical and administrative) are more likely to churn.
* **Insight**: High numbers of support tickets, particularly technical ones, might indicate dissatisfaction with the service, suggesting that improving customer support and addressing common issues could reduce churn.

**7. Financial Aspects and Churn:**

* A scatter plot or box plot of MonthlyCharges vs. Churn might reveal that customers with higher charges have a higher tendency to churn, particularly those with **lower tenure**.
* **Insight**: High monthly charges may be a significant factor driving churn. Pricing strategies, offering discounts, or introducing more affordable plans could help retain these customers.

**8. Actionable Insights and Strategies for Retention:**

* The final part of the dashboard story focuses on **actionable insights**:
  + **For high churn segments**: Target customers on month-to-month contracts with offers to lock in long-term plans or provide loyalty incentives.
  + **For at-risk groups**: Customers with low usage of services like TechSupport or StreamingTV may benefit from personalized engagement strategies to improve service adoption and satisfaction.

The dashboard, by combining **univariate**, **bivariate**, and **multivariate analysis**, presents a holistic view of the customer churn landscape. It not only identifies the factors contributing to churn but also provides actionable insights to improve customer retention and reduce churn rates. Through interactive charts and visuals, Power BI helps users dive deeper into the data, uncovering patterns and trends that drive the story of churn and retention.