By Ligasyah
Arnanda Primadana

DEEP : Data End-to-End Portfolio Project

The Dataset

Telco Dataset

The dataset in question is an exemplary choice for your portfolio project due to its rich and diverse content, which provides a comprehensive view of customer behavior and service usage. It includes critical information such as customer churn, service subscriptions, account details, and demographic data. This allows for a multifaceted analysis, enabling you to develop focused customer retention programs. By leveraging this dataset, you can showcase your ability to perform sophisticated data analyses and predictive modeling, thereby demonstrating your proficiency in deriving actionable insights that are crucial for strategic decision-making in the telecommunications industry.

Here is a concise explanation of the key business processes:

1. Customer Retention Management: Analyze customer behavior to identify churn factors and develop retention strategies.
2. Service Subscription Analysis: Evaluate customer service subscriptions to inform service enhancements and bundling.
3. Billing and Payment Processing: Manage account details and billing practices to ensure customer satisfaction and revenue collection.
4. Demographic Analysis: Use demographic data to segment customers and tailor marketing efforts.
5. Customer Support and Service Enhancement: Improve customer service by analyzing tech support and service issues.

These processes are essential for enhancing customer experience and business growth.

RevoLeague : Telecommunications customer loyalty

Business Context

RevoLeague is a prominent player in the telecommunications industry, renowned for its innovative services and customer-centric approach. As part of its strategic initiative to enhance customer satisfaction and loyalty, RevoLeague has embarked on an analytical journey to predict customer behavior with the ultimate goal of retaining its valued clientele.

DARCI

Decision-Maker	Accountable	Responsible	Consulted	Informed
Chief Marketing Officer (CMO)	Customer Experience Manager	• Data Analyst Team • Customer Relationship Management (CRM)	• Product Development Team • Sales Department • Customer Support Supervisor • Marketing Strategist	• Marketing Department • Customer Support • Finance Department • Human Resource

Project Scope

The project will utilize a comprehensive dataset that includes variables such as customer churn status, service subscriptions (e.g., phone, internet), account information (e.g., tenure, contract type), and demographic details (e.g., gender, age range). The focus areas of analysis will include identifying predictors of customer churn, evaluating the impact of current services on retention rates, and developing targeted strategies to enhance customer engagement. Exclusions from this scope include any analysis related to new product development outside existing services offered by RevoLeague and external market comparisons with competitors.

Problem Statment

RevoLeague aims to reduce customer churn by 15% over the next six months by analyzing customer data to identify key factors contributing to churn and implementing targeted retention strategies based on these insights.

SMART

Specific	Measurable	Achievable	Relevant	Time-bound
The objective of reducing customer churn	The target of a 15% reduction in churn provides a quantifiable metric that can be tracked and assessed throughout the project duration.	RevoLeague's access to comprehensive customer data and analytical resources	Reducing customer churn is directly aligned with RevoLeague's strategic objectives of enhancing customer satisfaction and loyalty	The inclusion of a six-month timeframe establishes urgency and allows for timely evaluation of progress

Objective and Analysis Purpose

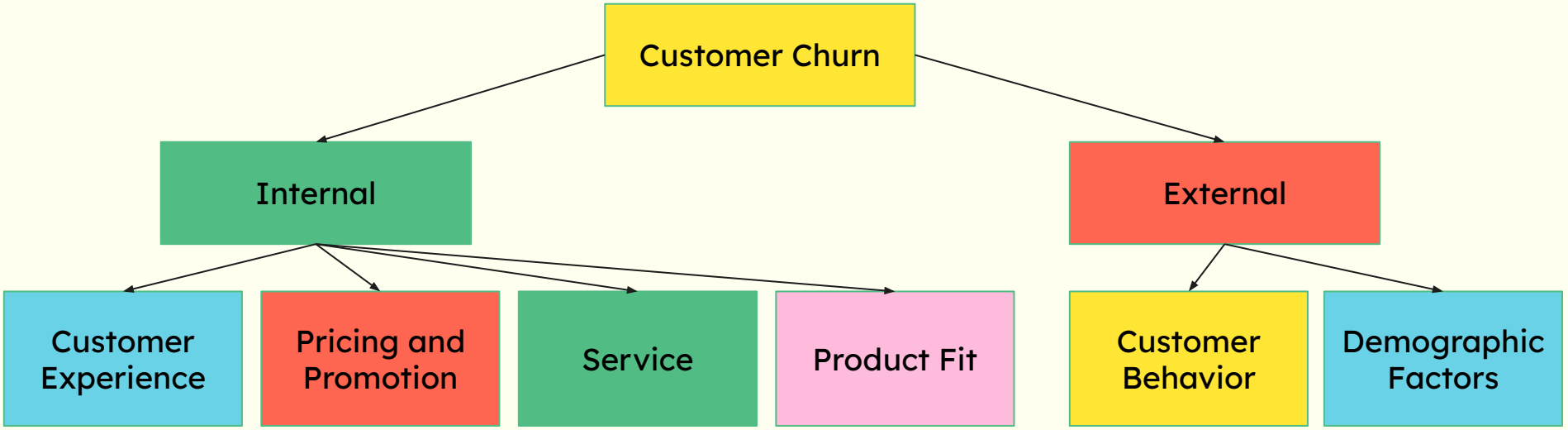
Objective

The primary objective is to develop a predictive model that accurately identifies customers at risk of churning, enabling RevoLeague to implement targeted retention strategies aimed at reducing churn rates by 15% within six months.

Analysis Purpose

The purpose of conducting a narrative analysis is to understand customer experiences and sentiments that contribute to churn.

MECE (Mutually Exclusive, Collectively Exhaustive)



- Inefficient Billing Process
- Technical Issue

- Lack of Competitive Price
- Insufficient promotional offers

- Poor customer service interactions

- Inadequate service and customer needs

- Customers usage product behavior
- Family members

- Demographic segment

Hypothesis

1. Machine Learning and EDA: If we utilize machine learning in conjunction with exploratory data analysis (EDA), then we will enhance our ability to predict customer churn by identifying key factors influencing retention.

2. Personalized Recommendations: If customers receive personalized service recommendations, then their likelihood of churning will significantly decrease.

3. Customer Support Experience: If customers have positive experiences with customer support, then their churn rates will be reduced, enabling us to implement targeted strategies for at-risk customers.

Key Performance Indicators (KPIs)

Churn Rate	Customer Retention Rate
Tracks the percentage of customers who discontinue their services within a given timeframe, providing insight into potential issues leading to customer loss.	Measures the percentage of customers who continue their subscriptions over a specified period, indicating overall satisfaction and loyalty.

Analysis Methods

Exploratory Data Analysis (EDA)

Conducted initially to understand data distributions and relationships among variables through visualizations such as histograms, box plots, and correlation matrices; EDA helps identify patterns or anomalies in customer behavior that may contribute to churn.

Customer Churn Prediction using Machine Learning

Implementing various machine learning algorithms such as logistic regression, decision trees, random forests, or gradient boosting machines to build predictive models that classify customers into "churn" or "not churn" categories based on historical data.

Milestone 2 : Data Cleaning and EDA

Data Cleaning

Here the Google Colab :

The Code

Dataset Overview

This dataset contains 7,043 entries and 21 columns, capturing customer demographics and service usage information. Most columns are of type object, representing categorical data such as gender, partner status, and payment method. Numeric columns like SeniorCitizen and tenure are correctly typed as int64, while MonthlyCharges is appropriately stored as a float64. However, the column TotalCharges is currently of type object, which is incorrect because it represents a numeric value (total amount charged). Therefore, I plan to convert the TotalCharges column to the correct numeric type (float64) to enable proper numerical analysis and calculations.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   customerID        7043 non-null   object 
 1   gender            7043 non-null   object 
 2   SeniorCitizen     7043 non-null   int64  
 3   Partner           7043 non-null   object 
 4   Dependents        7043 non-null   object 
 5   tenure            7043 non-null   int64  
 6   PhoneService      7043 non-null   object 
 7   MultipleLines     7043 non-null   object 
 8   InternetService   7043 non-null   object 
 9   OnlineSecurity    7043 non-null   object 
 10  OnlineBackup      7043 non-null   object 
 11  DeviceProtection  7043 non-null   object 
 12  TechSupport       7043 non-null   object 
 13  StreamingTV       7043 non-null   object 
 14  StreamingMovies   7043 non-null   object 
 15  Contract          7043 non-null   object 
 16  PaperlessBilling  7043 non-null   object 
 17  PaymentMethod     7043 non-null   object 
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object 
 20  Churn             7043 non-null   object 
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

Data Cleaning Step

Checking Data Type

The only issue found is with the TotalCharges column, which was incorrectly stored as an object (string) but has now been converted to the correct numeric type (float64). All other columns have appropriate data types.

Checking Missing Value

There are no missing values in any of the columns; all 7,043 entries are complete across all features.

Checking Duplicates

The dataset contains no duplicate rows, ensuring that each record is unique and reliable for analysis.

Checking Typo

No typos or inconsistent categorical values were detected in any of the string/object columns, indicating clean and consistent categorical data.

After Change DType

After converting, the TotalCharges column is now correctly set to float64, matching the numeric type of MonthlyCharges. The dataset still contains 7,043 entries and 21 columns, with most categorical features as object types and numeric columns like SeniorCitizen, tenure as int64. This change ensures accurate numerical analysis on the total charges data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   customerID        7043 non-null   object 
 1   gender            7043 non-null   object 
 2   SeniorCitizen     7043 non-null   int64  
 3   Partner           7043 non-null   object 
 4   Dependents        7043 non-null   object 
 5   tenure            7043 non-null   int64  
 6   PhoneService      7043 non-null   object 
 7   MultipleLines     7043 non-null   object 
 8   InternetService   7043 non-null   object 
 9   OnlineSecurity    7043 non-null   object 
 10  OnlineBackup      7043 non-null   object 
 11  DeviceProtection  7043 non-null   object 
 12  TechSupport       7043 non-null   object 
 13  StreamingTV       7043 non-null   object 
 14  StreamingMovies   7043 non-null   object 
 15  Contract          7043 non-null   object 
 16  PaperlessBilling  7043 non-null   object 
 17  PaymentMethod     7043 non-null   object 
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7032 non-null   float64
 20  Churn             7043 non-null   object 
dtypes: float64(2), int64(2), object(17)
memory usage: 1.1+ MB
```

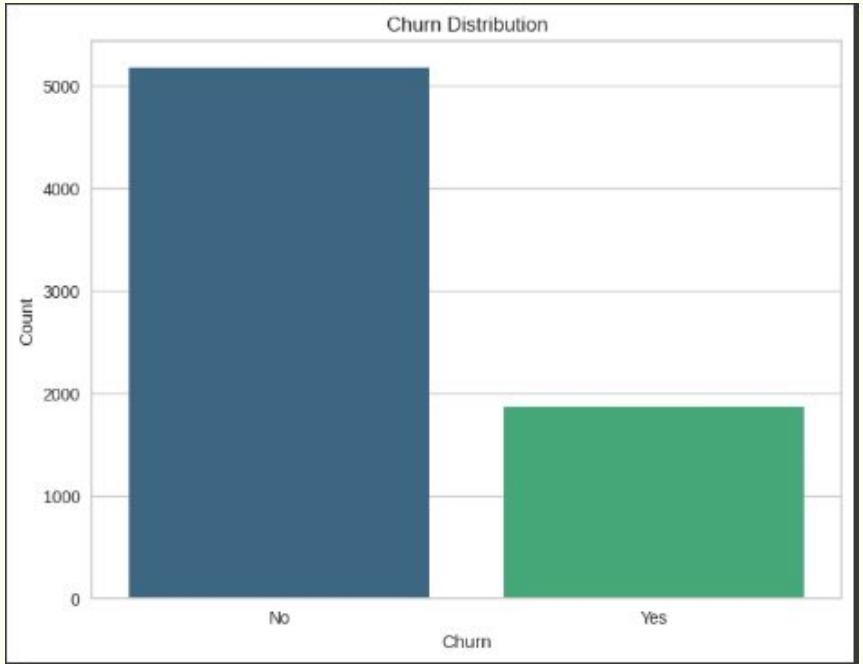

EDA

Understanding Context

In the context of the telecom industry, customer churn is a critical challenge as it directly impacts revenue and growth. Telecom companies often face moderate to high churn rates due to intense competition, service dissatisfaction, pricing issues, or contract flexibility. Our analysis reflects typical industry patterns where customers with shorter tenure, month-to-month contracts, higher monthly charges, and non-automatic payment methods are more likely to churn. Retaining customers through longer-term contracts and improving service value can significantly reduce churn rates. Understanding these factors helps telecom providers tailor strategies for customer loyalty and sustainable business performance.

Churn Distribution

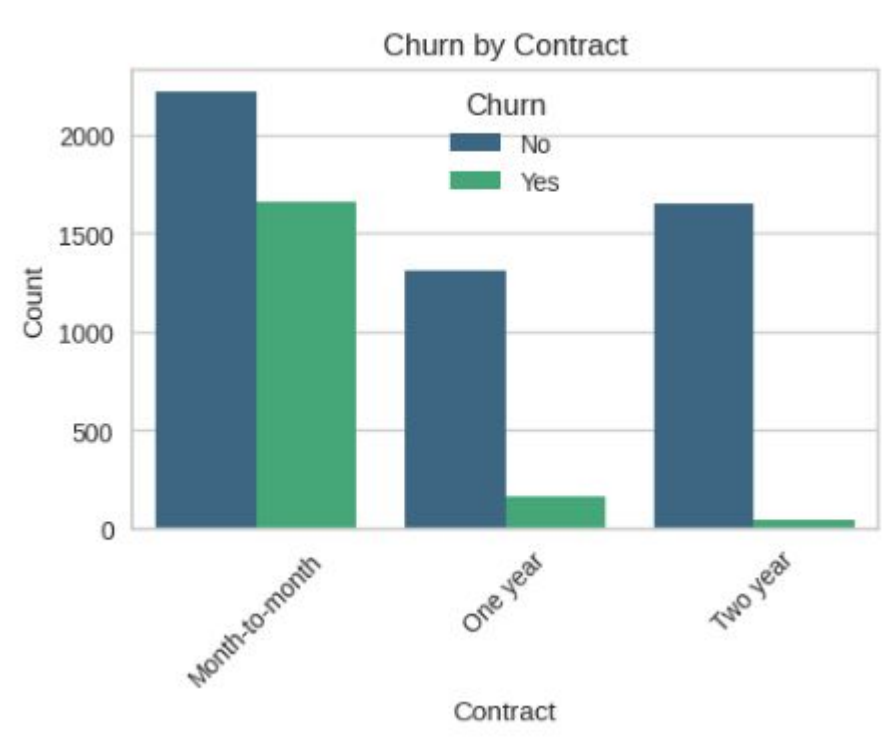
This bar chart illustrates the distribution of customer churn in the dataset. It shows that approximately 73.46% of customers did not churn, while about 26.54% did churn. The taller blue bar represents the majority who stayed, and the shorter green bar represents those who left, indicating a moderate churn rate within the customer base.




```
Churn
No     73.463013
Yes    26.536987
Name: count, dtype: float64
```

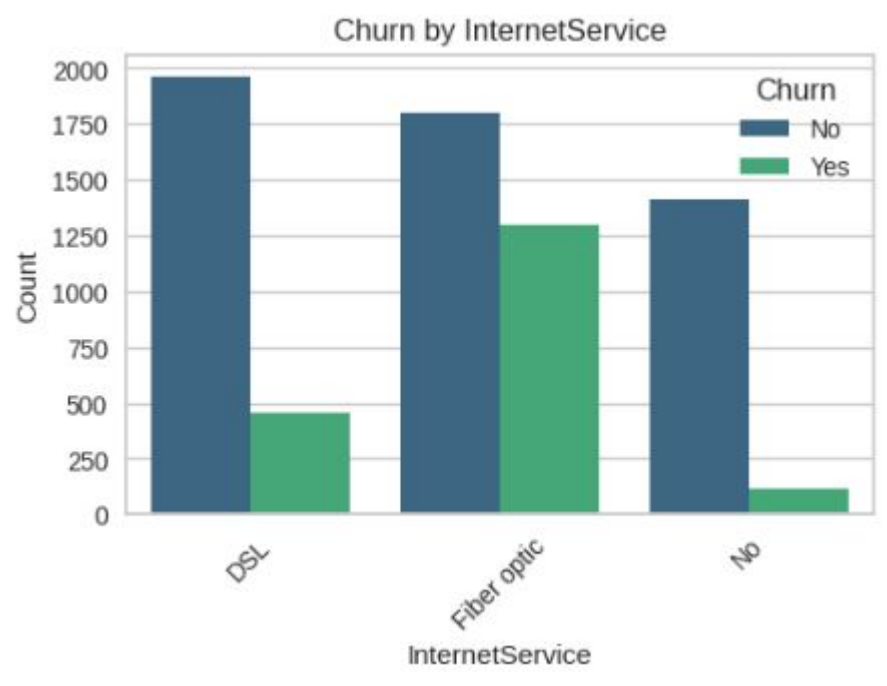
Churn by Contract

The Contract graph reveals that customers with month-to-month contracts have a substantially higher churn rate compared to those with one- or two-year contracts, indicating contract length strongly influences customer retention.

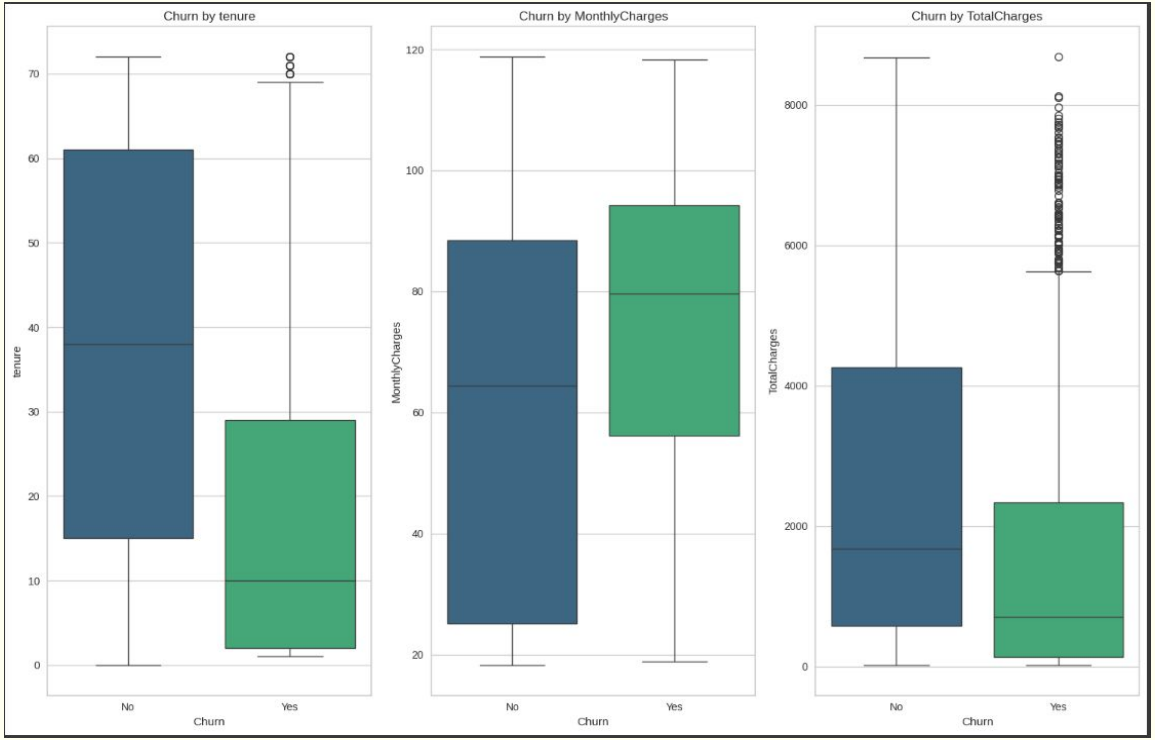


Churn by InternetService

The InternetService graph shows that customers using fiber optic service exhibit notably higher churn rates than those using DSL or no internet service, suggesting that service type impacts customer loyalty.

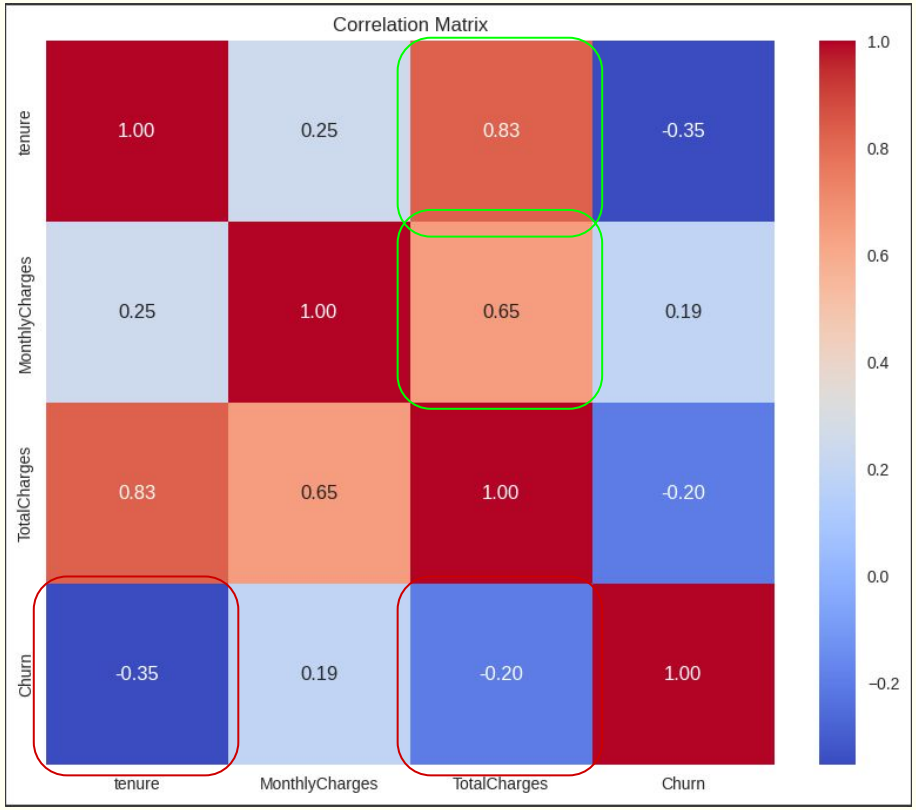


Churn by Numerical



The boxplots reveal clear differences between customers who churned and those who did not. Customers who churned tend to have significantly shorter tenure, indicating they leave earlier in their subscription period. Additionally, churned customers generally incur higher monthly charges, suggesting that higher costs may contribute to attrition. Lastly, total charges are lower for churned customers overall, likely reflecting their shorter tenure and fewer accumulated payments. These patterns highlight tenure and pricing as important factors influencing customer churn.

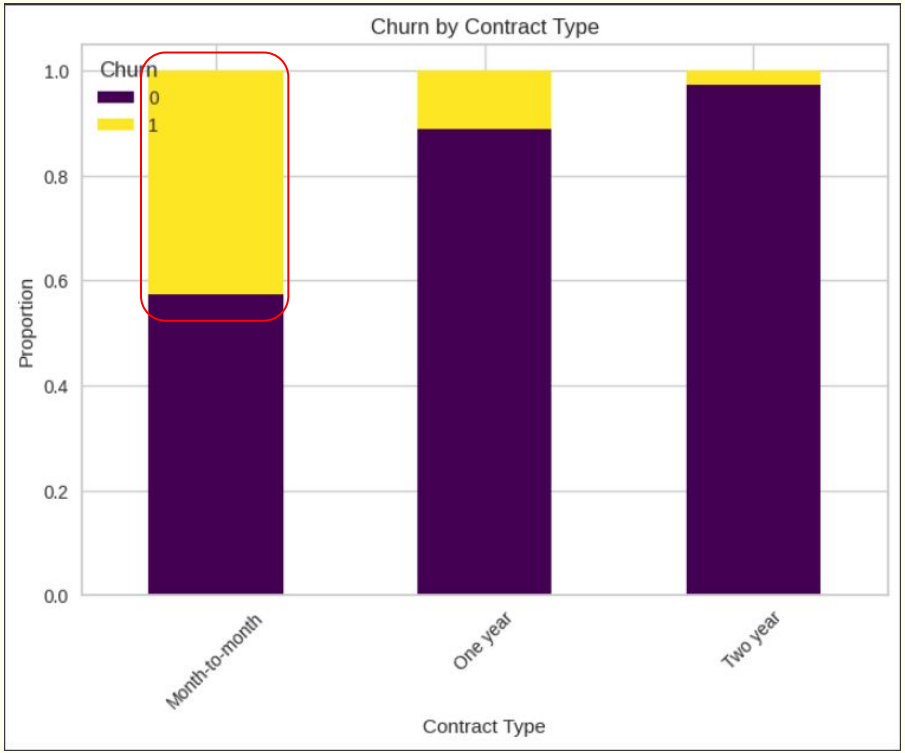
Churn by Numerical



This correlation matrix shows relationships between key numerical variables and churn. Notably, tenure has a moderate negative correlation with churn (-0.35), indicating that customers with longer tenure are less likely to churn. MonthlyCharges has a weak positive correlation (0.19) with churn, suggesting higher monthly fees slightly increase the chance of churn. TotalCharges also shows a weak negative correlation (-0.20) with churn, reflecting that customers who have paid more overall tend to stay longer. Strong positive correlations exist between tenure and total charges (0.83), as well as between monthly charges and total charges (0.65), which is expected since total charges accumulate over time based on monthly fees.

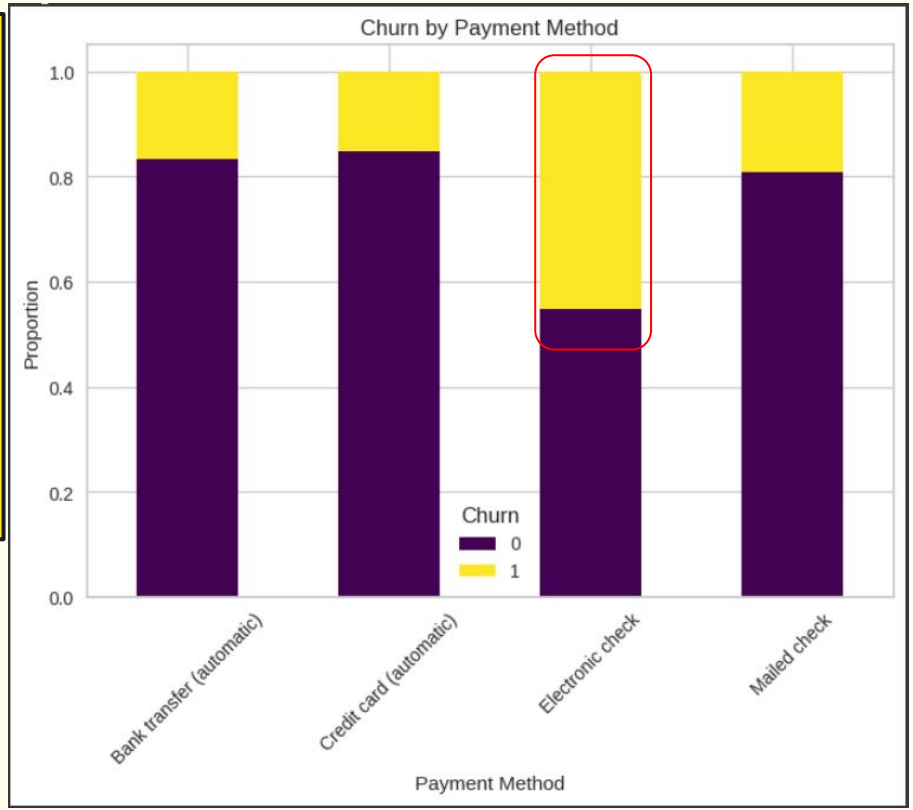
Churn by Contract Type

This stacked bar chart shows the proportion of customers who churned (yellow) versus those who stayed (purple) across different contract types. It clearly indicates that customers with month-to-month contracts have a much higher churn rate compared to those with one-year or two-year contracts, where churn is minimal. This highlights that longer-term contracts are associated with better customer retention.



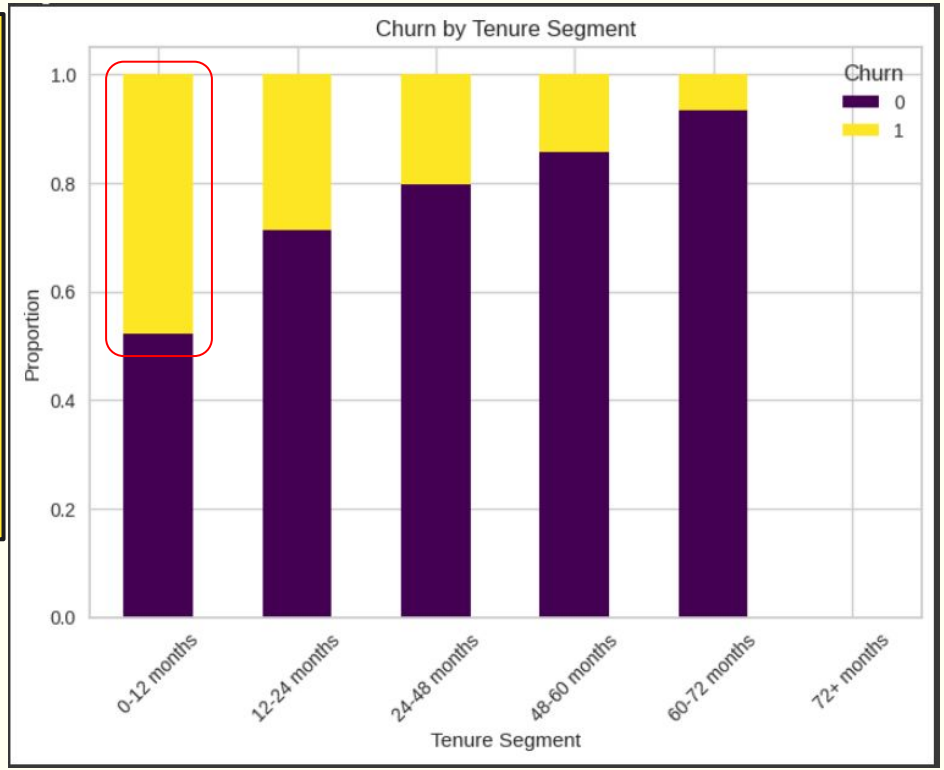
Churn by Payment Method

This bar chart illustrates the churn rates across different payment methods. Customers who pay via electronic check exhibit a noticeably higher churn rate compared to those using bank transfer, credit card, or mailed check. The lower churn proportions for automatic payment methods (bank transfer and credit card) suggest that automated payments may contribute to better customer retention. This insight highlights the importance of encouraging customers to adopt automatic payment options to reduce churn.



Churn by Numerical

This bar chart shows that the churn rate decreases as customer tenure increases. Customers in the 0-12 months segment have the highest proportion of churn, while those with longer tenures, especially beyond 60 months, exhibit much lower churn rates. This indicates that customers who stay longer are less likely to leave, emphasizing the importance of early retention efforts to reduce churn.



Here the Tableau Public Dashboard :

The Dashboard

Churn by Contract

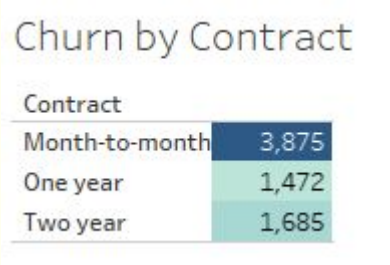
Insights from the Table

- Month-to-month contracts have the highest churn number (3,875), indicating these customers are the most at risk.
- One-year and two-year contracts have significantly lower churn numbers (1,472 and 1,685 respectively), suggesting longer contracts improve retention.

Business Recommendations

- Focus Retention Efforts on Month-to-Month Customers

1. Develop targeted offers or incentives to encourage these customers to switch to longer-term contracts.
2. Use predictive modeling to identify which month-to-month customers are most likely to churn and proactively engage them.



Churn by Service

Insights from the Service Table

1. Customers with DSL internet and no phone service or no streaming services tend to have lower counts compared to those with multiple services.
2. Fiber optic customers with multiple services (phone, streaming TV, streaming movies) show higher numbers, indicating these customers might be more engaged.
3. Customers with no internet service but phone service have the highest count (1,180), which could be a unique segment to analyze further.
4. Bundling services like streaming TV and movies with internet service seems common and could influence retention.

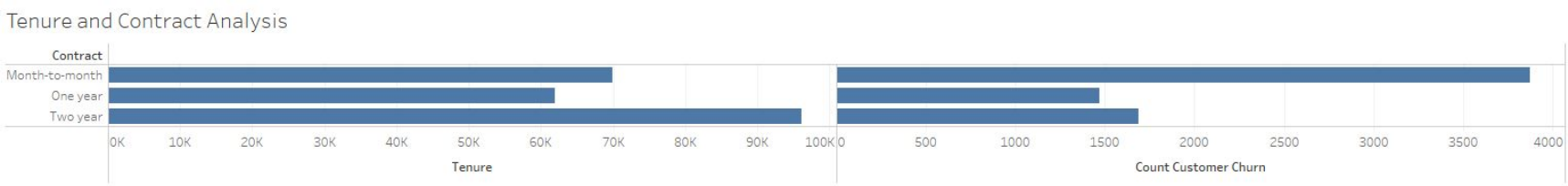
Business Recommendations

- Encourage Service Bundling
- Promote bundles combining internet (especially fiber optic) with streaming and phone services to increase customer stickiness.
- Use targeted offers to upsell streaming or phone services to DSL or no-internet customers.

Service

Internet Se..	Phone Servi..	Multiple Lin..	Streaming TV	Streaming Movies	
DSL	No	No phone service	No	No	299
				Yes	98
			Yes	No	83
				Yes	200
	Yes	No	No	No	588
				Yes	139
			Yes	No	139
				Yes	181
		Yes	No	No	225
				Yes	114
			Yes	No	102
				Yes	248
Fiber optic	Yes	No	No	No	461
				Yes	173
			Yes	No	175
				Yes	349
		Yes	No	No	444
				Yes	268
			Yes	No	265
				Yes	961
No	Yes	No	No internet service	No internet service	1,180
		Yes	No internet service	No internet service	340

Churn by Tenure and Contract



Insights from the Chart

1. Month-to-month customers have the highest churn count and relatively lower total tenure, indicating they are more likely to leave quickly.
2. Customers with two-year contracts have the highest total tenure and lower churn counts, showing longer contracts help retain customers longer.
3. One-year contracts fall in between, with moderate tenure and churn.

Business Recommendations

- Leverage Tenure Data
- Use tenure as a key feature in your churn prediction model to identify customers who might be nearing a churn risk period.
- Tailor retention efforts based on how long a customer has been with the company.

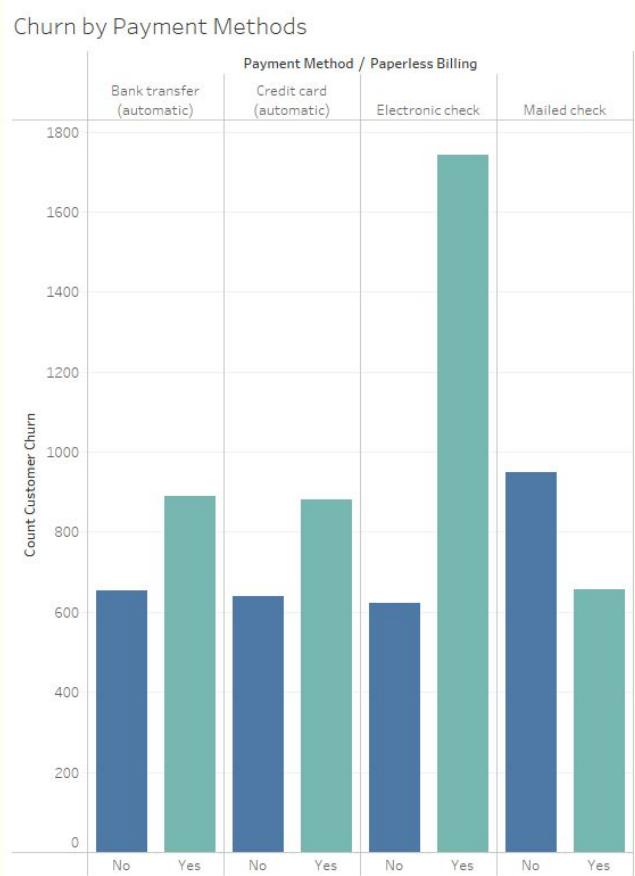
Churn by Payment Methods

Insights from the Chart

1. Customers using Electronic Check with paperless billing have the highest churn count, indicating this group is at significant risk.
2. Other payment methods like bank transfer and credit card (automatic) show lower churn rates, especially when paperless billing is not enabled.
3. Paperless billing seems to increase churn risk for some payment methods, particularly electronic checks.

Business Recommendations

- Target Electronic Check Users with Paperless Billing
- Investigate why this group has higher churn—could be payment issues or dissatisfaction.
- Offer incentives or alternative payment options to reduce churn risk.



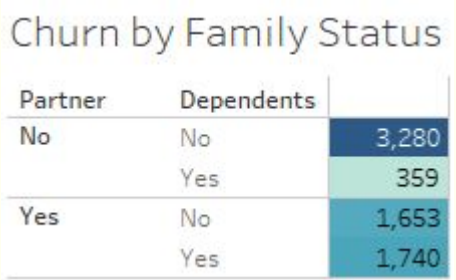
Churn by Family Status

Insights from the Data

1. Customers without a partner and without dependents have the highest churn count (3,280), indicating they are the most likely to leave.
2. Customers with either a partner or dependents (or both) show significantly lower churn numbers, suggesting family status may contribute to customer loyalty.

Business Recommendations

- Target Single Customers Without Dependents
- Develop personalized retention offers or engagement programs for customers who are single and without dependents, as they are at higher risk.
- Consider loyalty incentives or service bundles tailored to their needs.



Soon Machine Learning