

Exploratory Data Analysis - Presentation

- Suhaana Kapali

MSC DATA AND ARTIFICIAL
INTELLIGENCE ·
ACCA7015 · DATA MINING
& VISUALISATION

Aims and Objectives

To clean and prepare the telecom customer dataset for analysis.

To handle missing values, duplicates and outliers to improve data quality and churn.

To identify key trends, patterns and relationships between customer features and churn.

To use visual and statistical techniques to compare customer who churned with those who stayed.

To understand the findings and highlight key insights about customer behavior and churn risk.

Introduction

This study examines customer churn in the telecommunications industry.

The dataset contains 100,000 customer records.

It includes customer usage details such as minutes of use, average monthly usage, equipment days, monthly charges, and churn status.

Customer churn is an important issue because losing customers can reduce company revenue and increase replacement costs.

Predicting churn is difficult because customer behavior depends on several usage and billing factors.

The goal of this study is to analyze usage patterns and identify factors that may influence whether a customer churns or stays.

Exploratory data analysis and visual techniques are used to understand trends, relationships, and churn behavior.

Data Overview

```
import pandas as pd

csv_file_path = '/content/drive/MyDrive/DH/abc.csv'
df = pd.read_csv(csv_file_path)
display(df.head())
```

	eqpdays	mou_Mean	avg6mou	avg3mou	avg3qty	mou_cvce_Mean	totmrc_Mean	mou_peav_Mean	mou_opkv_Mean	mou_rvce_Mean	churn
0	361	219.25	322	272	116	90.643333	22.500	132.600000	55.220000	97.176667	1
1	240	482.75	477	305	158	189.396667	37.425	75.333333	169.343333	55.280000	0
2	1504	10.25	11	12	7	5.426667	16.990	5.193333	0.233333	0.000000	1
3	1812	7.50	50	8	3	8.410000	38.000	3.380000	5.450000	0.413333	0
4	434	570.50	586	558	191	285.233333	71.980	173.476667	218.086667	106.330000	0

Dataset contains 100,000 telecom customer records

11 attributes describing customer usage, billing, equipment, and churn status

Includes customer behavior metrics such as minutes of use, monthly charges, and equipment age

Real-world customer data used for customer churn analysis and prediction

Target variable: Churn (0 = Active, 1 = Churned)

Source: Telecom Customer Churn Dataset

Variables

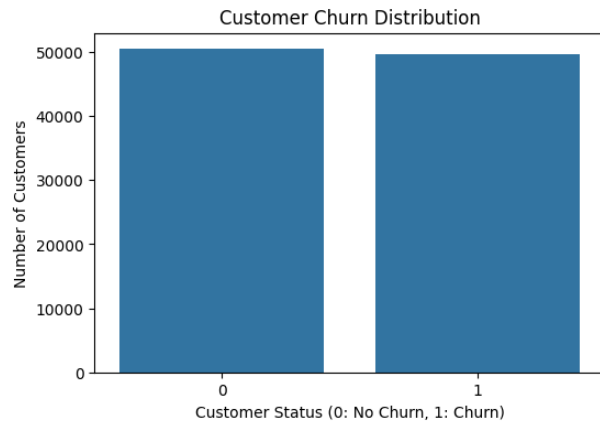
- ▶ The analysis focuses on the following key variables:
- ▶ **eqpdays** – Number of days the customer has used their current equipment.
- ▶ **mou_Mean** – Average minutes of use.
- ▶ **avg3mou / avg6mou** – Average monthly usage over 3 and 6 months.
- ▶ **avg3qty** – Average number of calls over the last 3 months.
- ▶ **totmrc_Mean** – Average monthly recurring charge.
- ▶ **Customer Usage Variables** – Behavioural indicators describing service usage patterns.

Reason for selection:

These variables were selected because they best represent customer engagement, service usage, and billing behaviour, which are important factors related with churn.

Feature	Value
Total Records	100,000
Selected Variables	11
Missing Values	0
Target Variable	Churn
Active Customers	50.4%
Churned Customers	49.6%
Data Type	Primarily Numerical

This dataset contains information about Telecom Customer Churn, including:



Customer ID: A unique identifier assigned to each customer.

eqpdays: The number of days the customer has been using their current equipment.

mou_Mean: The average number of minutes used by the customer.

avg3mou: The average monthly minutes of use over the last 3 months.

avg6mou: The average monthly minutes of use over the last 6 months.

avg3qty: The average number of calls made over the last 3 months.

totmrc_Mean: The average monthly recurring charge paid by the customer.

Monthly Charges: The customer's monthly service charges.

Customer Usage Variables: Numerical measures describing customer calling and service usage patterns.

Churn: A binary indicator showing whether the customer left the telecom company (1 = Churned, 0 = active).

Research Methodology

Data Collection:

- Telecom Customer Churn dataset containing 100,000 customer records and 11 variables.
- Data includes customer usage, billing information, equipment details, and churn status.

Data Pre Processing:

- Checked for missing values, duplicate records and inconsistencies.
- The numerical variables were checked to confirm they were suitable for analysis. Selected 11 relevant variables related to usage, billing, equipment age, and churn.
- Grouped customers into active(0) and Churned (1) to compare customer behaviour. Examined outliers using boxplots before developing visual insights.

Exploratory Data Analysis:

- Created summary statistics and visualizations.
- Examined the relationships between customer usage, monthly charges, equipment age, and customer churn..
- Recognized trends, patterns, and potential outliers.

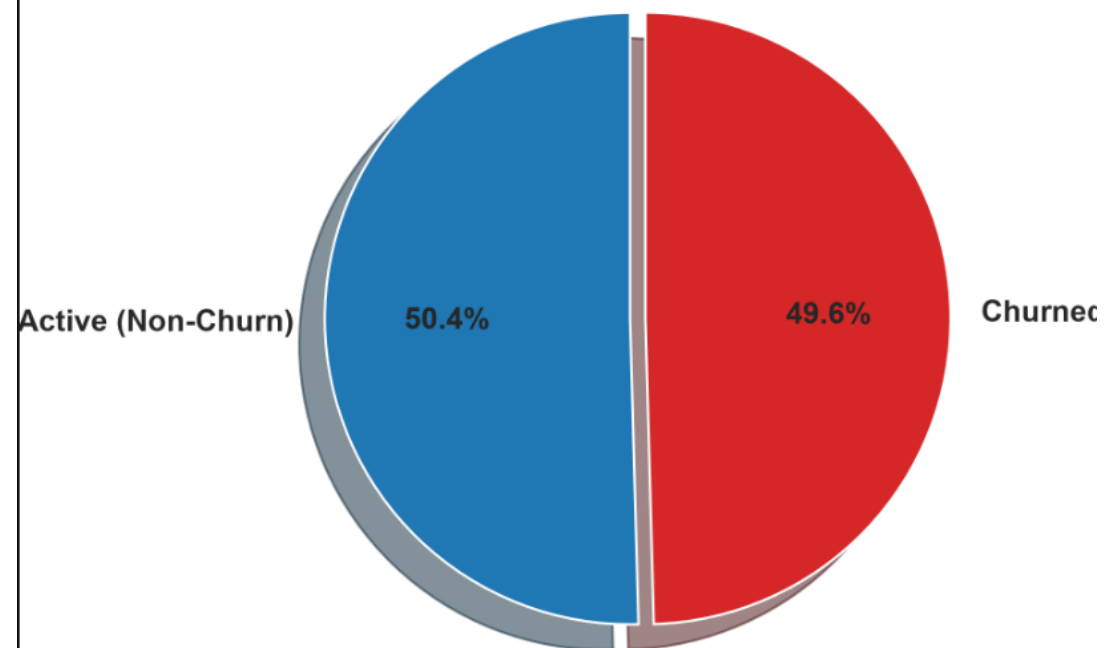
Univariate Analysis:

- ▶ Customer churn distribution.
- ▶ Balanced dataset suitable for classification.

KEY FINDINGS

- ▶ 50.4% customers remained active.
- ▶ 49.6% customers churned.
- ▶ Balanced classes reduce model bias.
- ▶ Churn is a significant business issue.
- ▶ Customers are divided into two groups: Active customers (0) and Churned customers (1). The dataset is almost evenly balanced, with approximately 50.4% active customers and 49.6% churned customers. This means the dataset provides a fair basis for comparing customers who stayed with those who left.

Customer Churn Distribution



Univariate Analysis : Box plots for Numerical Features

Distribution Explanation:

► These boxplots show the distribution, spread, and variability of the numerical features in the dataset.

► Variables such as **minutes of use**, **average monthly usage**, **equipment days**, and **monthly recurring charges** show different levels of spread.

► Most numerical variables are positively skewed, meaning that the majority of customers have moderate service usage, with relatively few customers recording very high usage levels. This indicates that customer behaviour is not uniform across the dataset.

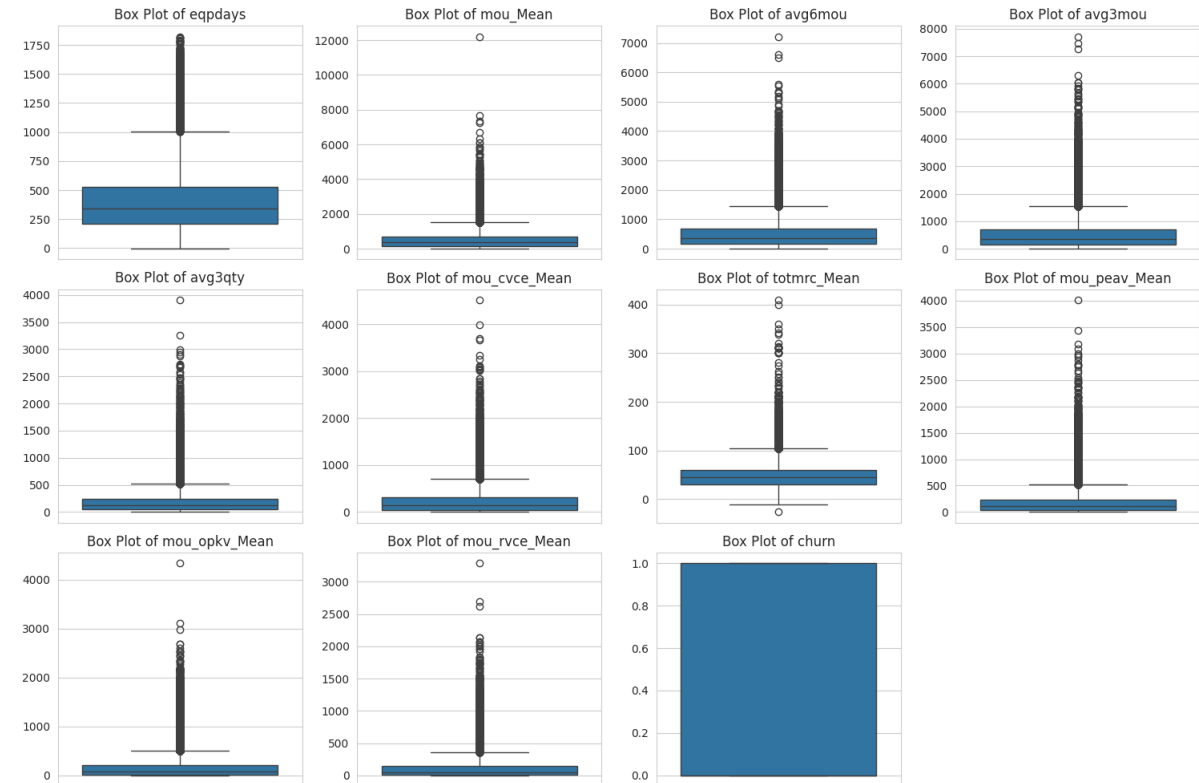
Outlier Discussion:

► Several numerical variables contain outliers, especially usage-related variables such as **mou_Mean**, **avg3mou**, and **avg6mou**.

► These outliers were reviewed rather than automatically removed.

► In this context, outliers may represent genuine customer behaviour, such as heavy users or customers with unusually high service usage.

► Retaining these values preserves genuine customer behaviour and provides a more accurate representation of the dataset.

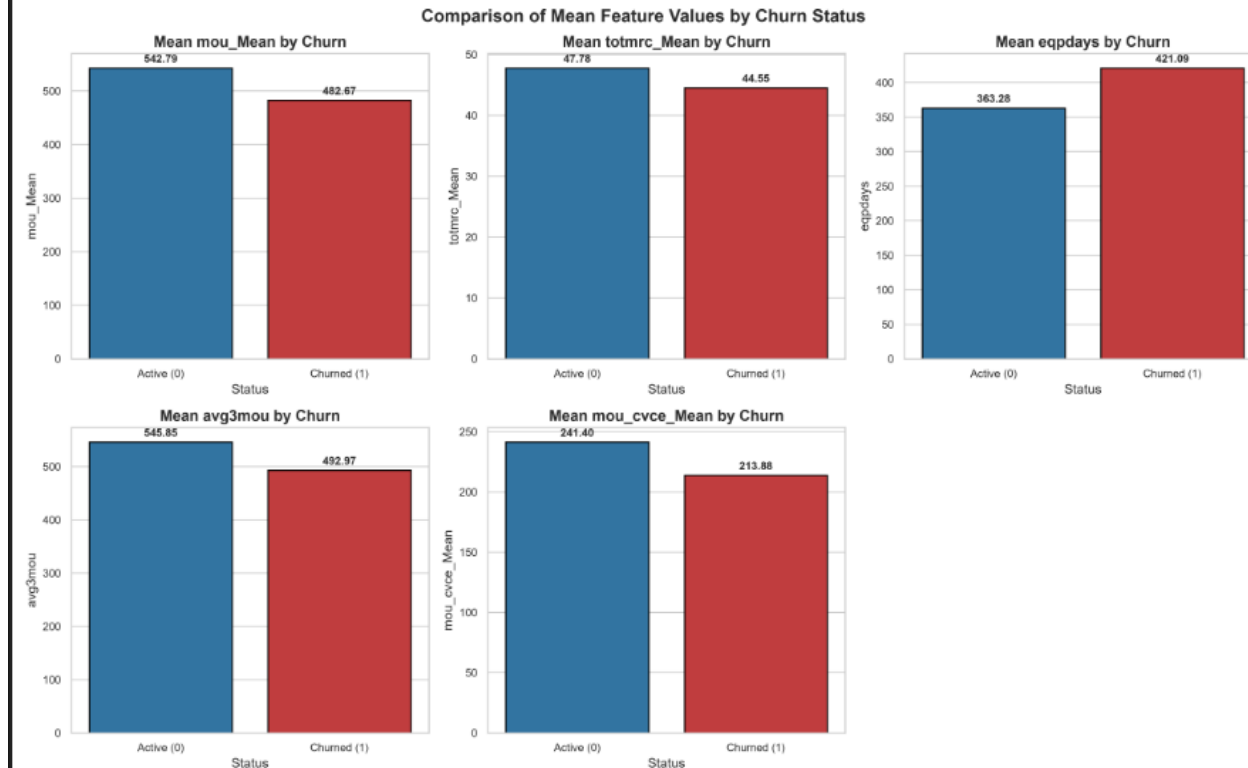


Multivariate Analysis

Comparison of average customer behaviour.

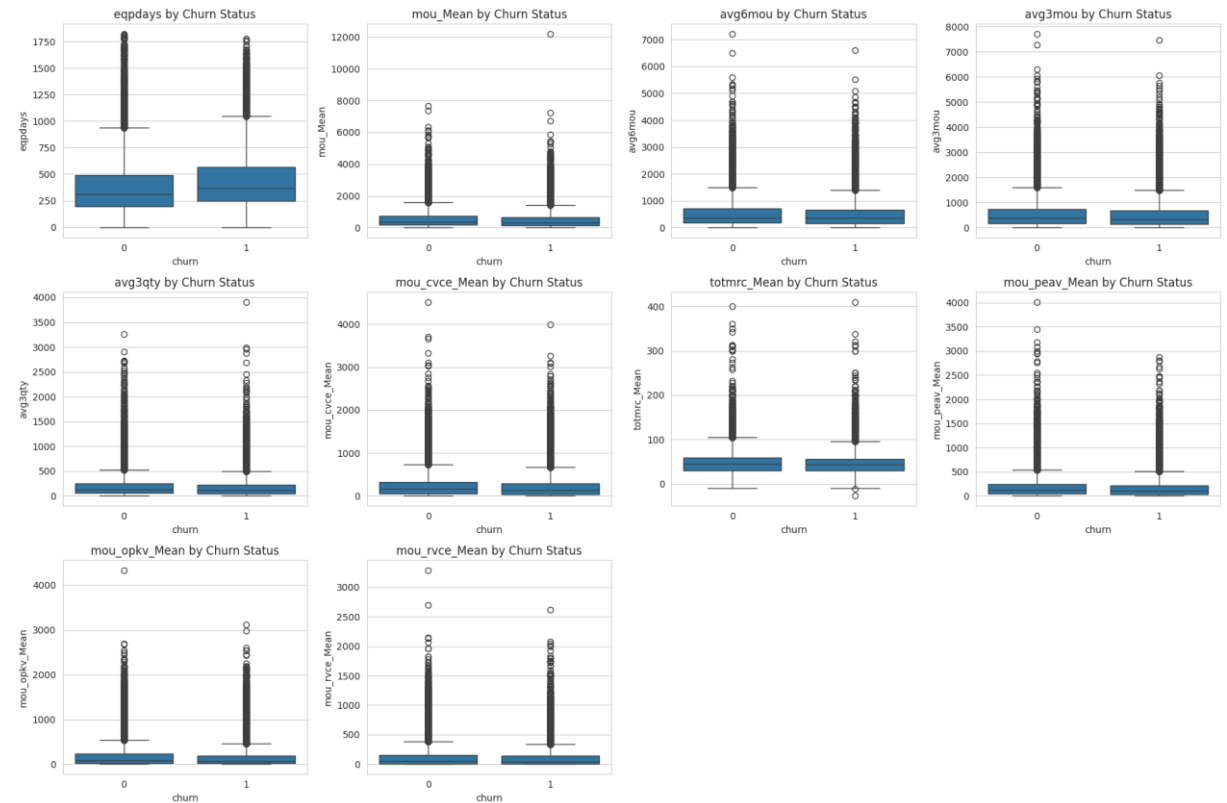
Business Interpretation

- Customers who use telecom services more frequently appear more likely to remain with the provider.
- Older equipment may indicate customers who are out of contract or considering changing providers, increasing the possibility of churn.
- The relatively small difference in monthly charges suggests that pricing alone is not the major reason customers leave.
- Overall, customer engagement and equipment age appear to be stronger indicators of churn than billing alone.



Multivariate Analysis : Box plots for Comparison Graphs with Churn

- ▶ Churned customers have a higher median equipment age (**eqpdays**), showing that customers with older devices are more likely to leave the service.
- ▶ Active customers generally show higher median values for usage related variables such as **mou_Mean**, **avg3mou**, and **avg6mou**, which suggests greater customer engagement.
- ▶ Monthly recurring charges (totmrc_Mean) shows **similar** distributions for both customer groups, showing that pricing alone does not clearly separate churned customers.
- ▶ Several variables contain high value outliers, particularly the usage related features, representing customers with extremely high service usage.

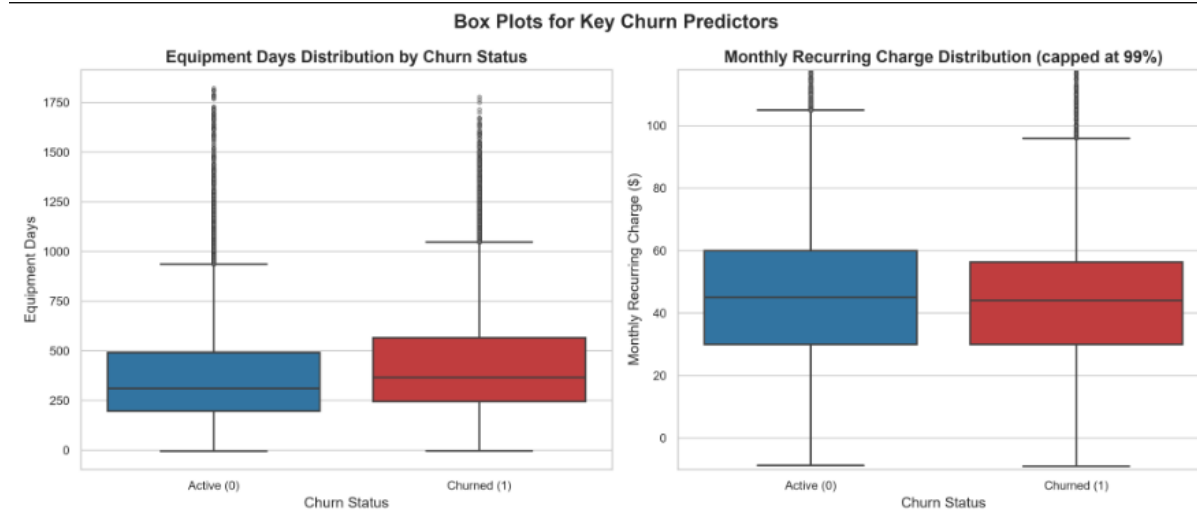


Multivariate Analysis

- Distribution of equipment age and monthly charges.
- Older equipment is related with churn.
- Monthly charges overlap significantly.
- Equipment improvements may reduce churn.

Insight

The analysis indicates that telecom providers should focus on customers with older devices by offering upgrade programmes and monitoring customer engagement, as usage behaviour appears to have a greater influence on churn than pricing alone.



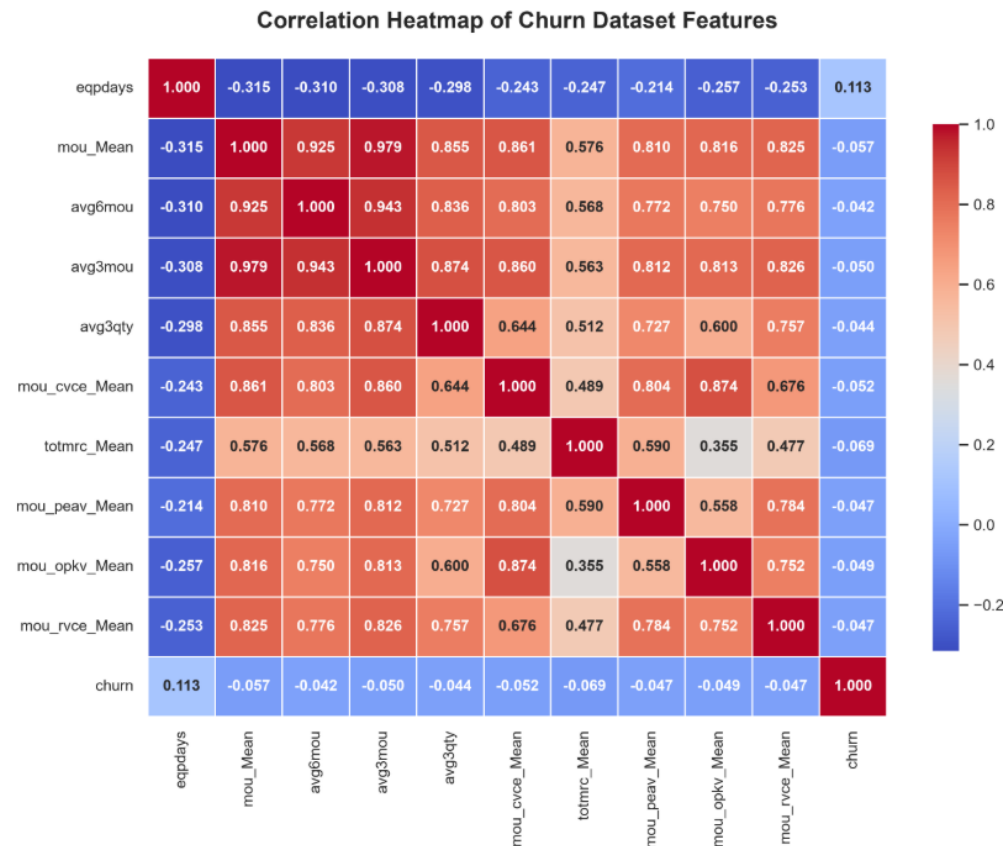
Correlation Analysis

The correlation heatmap was used to examine the strength and direction of relationships between the numerical variables and identify which features are most closely related with customer churn.

► Strong connections exist among customer usage variables, reflecting consistent customer behaviour over time.

► No individual variable has a strong direct relationship with churn.

► Customer churn is a difficult business problem that requires analysing multiple behavioural factors together rather than be dependent on on a single predictor.



Practical Data Handling – Cleaning, Grouping and Visual Insight Development

Data Cleaning

- ▶ Assessed missing values and confirmed the selected variables were complete.
- ▶ Checked duplicate records to avoid repeated customer observations.
- ▶ Verified numerical data types for usage, billing, and equipment variables.
- ▶ Used boxplots to detect outliers and review extreme customer behaviour.

Data Grouping

- ▶ Grouped customers into **Active (0)** and **Churned (1)**.
- ▶ Selected features related to usage, monthly charges, equipment age, and churn.
- ▶ Compared customer behaviour across churn groups to identify patterns.

Visual Insight Development

- ▶ **Pie chart:** showed churn distribution.
- ▶ **Boxplots:** identified spread and outliers.
- ▶ **Mean comparison chart:** compared active vs churned customers.
- ▶ **Heatmap:** examined correlations between numerical variables.

Executive Summary

The results provide valuable understandings into customer behaviour and support informed business decisions. Telecommunications companies can use these understandings to enhance customer satisfaction, reduce churn, and improve sustainable profitability.

Key Findings

The dataset showed a balanced distribution between active (50.4%) and churned (49.6%) customers, enabling meaningful comparisons.

Active customers consistently demonstrated higher service usage, while churned customers had older equipment and lower engagement levels.

Strong connections were observed among usage related variables, whereas individual variables showed only weak relationships with churn.

The analysis indicates that customer churn is influenced by multiple behavioural factors rather than a single characteristic.



Recommendations

1. Implement Device Upgrade Campaigns

- ▶ Target customers with older equipment by offering discounted upgrades or contract renewal incentives.
- ▶ This may improve customer satisfaction and reduce churn.

2. Monitor Customer Engagement

- ▶ Track customers whose usage declines over time.
- ▶ Use automated alerts to identify risky customers before they leave.

3. Provide Personalised Retention Offers

- ▶ Offer personalized pricing plans, loyalty rewards, or promotional packages based on customer behaviour.
- ▶ Personalised helps are more effective than common maintenance campaigns.

4. Support Future Predictive Analytics

- ▶ The understandings gained from this EDA can be used to develop future machine learning models for early churn prediction.
- ▶ Combining behavioural variables may improve prediction accuracy and allow proactive customer retention.

Conclusion



- ▶ The exploratory data analysis successfully identified key behavioural patterns associated with customer churn.
- ▶ Active customers consistently demonstrated higher service usage and customer engagement.
- ▶ Churned customers generally had older equipment, suggesting equipment age may influence customer retention.
- ▶ Monthly charges alone showed only a weak relationship with customer churn.
- ▶ Strong correlations were observed among customer usage variables, indicating consistent usage behaviour over time.
- ▶ Data cleaning and preprocessing ensured a reliable dataset for meaningful analysis.
- ▶ Combining multiple behavioural features provides better insight into customer churn than relying on a single variable.
- ▶ The findings support analytical customer retention strategies and informed business planning
- ▶ The analysis provides a strong foundation for future predictive analytics and customer churn management.

Customer churn is influenced by multiple behavioural factors rather than a single characteristic. By implementing exploratory data analysis, telecom companies can better understand customer behaviour, identify potential churn customers early, and implement targeted retention strategies to improve customer loyalty and continuous business performance.



► Thankyou