Summary of Data Exploration and Preprocessing

# Project Title:

Customer Churn Prediction and Analysis

# Key Insights from Data Exploration

1. Understanding the Dataset:  
 - Data loaded from a CSV file named 'churn.csv'.  
 - Initial inspection suggests it includes customer-related attributes relevant to churn behavior.  
  
2. Visual Exploration:  
 - Feature distributions were visualized to detect:  
 - Skewed distributions.  
 - Outliers (using boxplots).  
 - Correlation patterns (using heatmaps).  
  
3. Target Variable:  
 - The variable 'Churn' is the primary focus for prediction.  
 - Class imbalance may have been considered for handling.

# Features in the Dataset

Numerical Features (Scaled & Handled for Outliers):

* - age  
  - days\_since\_last\_login  
  - avg\_time\_spent  
  - avg\_transaction\_value  
  - avg\_frequency\_login\_days  
  - points\_in\_wallet

Binary Categorical Features (Label Encoded):

* - gender  
  - joined\_through\_referral  
  - used\_special\_discount  
  - offer\_application\_preference  
  - past\_complaint

Multi-class Categorical Features (One-Hot Encoded):

* - region\_category  
  - membership\_category  
  - medium\_of\_operation  
  - internet\_option  
  - complaint\_status  
  - feedback  
  - preferred\_offer\_types

❌ Dropped Feature:

* - security\_no (non-informative or identifier column)

# Preprocessing Decisions Made

1. Encoding Categorical Variables:  
 - Binary features were label-encoded.  
 - Multi-class features were one-hot encoded with first category dropped to avoid multicollinearity.  
  
2. Scaling Numerical Features:  
 - StandardScaler was applied to normalize numerical features.  
  
3. Handling Skewness:  
 - Box-Cox transformation applied where necessary.  
  
4. Missing Values:  
 - Numerical features filled with the median.  
 - Categorical features filled with the mode.  
  
5. Outlier Treatment:  
 - Used IQR method to cap outliers in selected numerical features.  
  
6. Correlation Analysis:  
 - Heatmaps likely used to identify and possibly drop highly correlated features.

# ✅ Conclusion

This thorough preprocessing pipeline ensures the data is clean, numerically stable, and appropriately encoded for model training.  
Such preparation significantly boosts the accuracy and reliability of churn prediction models.