Weekly Assignment - Data Cleaning, Preprocessing & EDA

Project Overview

You've been hired as a Junior Data Analyst at EXL. As part of your onboarding, you are asked to work on a realistic data cleaning and exploration task based on a simulated customer dataset. This is a 1-week long final assignment to test your understanding of the training sessions.

Dataset Context

You are provided with a fictional dataset named customer_data.csv. It contains transactional and demographic data of EXL's retail customers across regions. This data needs cleaning, preprocessing, and analysis before being used for modeling or reporting.

Column Descriptions

- CustomerID: Unique identifier
- Name: Customer's full name
- Age: Customer age (may contain missing values)
- Gender: Male/Female
- Region: Region of customer (North, South, East, West)
- AnnualIncome: Annual income in INR (some missing)
- MembershipLevel: Loyalty tier (Silver, Gold, Platinum)
- Purchases: Total number of purchases
- LastPurchaseDate: Date of last purchase
- TotalRevenue: Lifetime total revenue from the customer

Day 1: Data Loading & Initial Exploration

1. Load the data into a DataFrame using pandas.
2. Display first 10 rows.
3. Use df.info() and df.describe() to get a structural overview.
4. Count unique values in columns like Region, Gender, MembershipLevel.

Bonus: Convert LastPurchaseDate column to datetime.

Day 2: Handling Missing Values

1. Identify columns with missing values using isnull().sum().
2. Drop rows where Age is missing.
3. Fill missing AnnualIncome by group median based on MembershipLevel.
4. Fill missing Gender with mode.
5. Create a summary DataFrame showing before vs after missing values.

Day 3: Outlier Detection and Treatment

1. Detect outliers in TotalRevenue using:
   - Z-Score method
   - IQR method

Weekly Assignment - Data Cleaning, Preprocessing & EDA

2. Remove or cap/floor the outliers.
3. Show comparison using boxplot before and after.

Day 4: Data Normalization

1. Normalize AnnualIncome and TotalRevenue using:
   - Min-Max Scaling
   - Z-score Standardization
2. Store scaled versions as new columns.
3. Compare original vs scaled values using histplot or lineplot.

Day 5: Feature Engineering

1. Create new column RevenuePerPurchase = TotalRevenue / Purchases.
2. Extract CustomerTenureMonths = today - LastPurchaseDate.
3. Categorize customers based on income brackets.
4. Convert MembershipLevel into numerical values using label encoding.

Day 6: Visualizations & Insights

1. Region-wise Revenue: Bar plot
2. Revenue over time: Line chart
3. Product-wise share: Pie chart (simulate product column if needed)
4. Scatter plot: Purchases vs TotalRevenue

Day 7: Final Report

1. Summarize insights:
   - Which regions perform better?
   - What income groups are more profitable?
   - Does age or gender show any pattern in revenue?
2. Export cleaned dataset.
3. Submit PDF report with graphs, key findings, and cleaning steps.

Code Sample: Median Imputation

```
df['AnnualIncome'] = df.groupby('MembershipLevel')['AnnualIncome'].transform(
    lambda x: x.fillna(x.median())
)
```