Project Report: Customer Shopping Trends Analysis

1. Introduction

This project presents a structured analysis of customer shopping behavior using a combination of **data preprocessing**, **feature engineering**, and **SQL-based Exploratory Data Analysis (EDA)**. The dataset, containing demographic variables, purchase information, product attributes, review ratings, subscription details, and behavioral indicators, is examined to uncover meaningful patterns that may support business decision-making.
The workflow incorporates **Python (Pandas)** for data cleaning and preparation, **PostgreSQL** for analytical querying, and **Power BI** for visualization.

2. Data Preprocessing

Prior to analysis, the dataset underwent a series of preprocessing steps to improve data quality, analytical consistency, and SQL compatibility.

2.1 Handling Missing Values

A total of **37 missing values** were detected in the *Review Rating* column.

Instead of filling these with a global median, which could distort the distribution across product categories, the missing ratings were imputed using the **median rating within each category**:

```
df['Review Rating'] = df.groupby('Category')['Review Rating'].transform(lambda x: x.fillna(x.median()))
```

This category-wise median imputation maintains rating relevance and reduces the impact of outliers.

2.2 Standardization of Column Names

To avoid case-sensitivity issues in SQL and improve naming consistency:

- All column names were converted to lowercase.
- Spaces were replaced with underscores.
- The column purchase_amount_(usd) was renamed to purchase_amount.

```
df.columns = df.columns.str.lower().str.replace(' ', '_')

df = df.rename(columns={'purchase_amount_(usd)':'purchase_amount'})
```

This ensures smoother queries and enhances readability.

2.3 Feature Engineering: Age Segmentation

To analyze behavioral differences across age groups, the continuous age variable was converted into **quantile-based segments** using pandas.qcut.
The resulting buckets were:

- Young Adult
- Adult
- Middle-aged
- Senior

```
labels = ['Young Adult', 'Adult', 'Middle-aged', 'Senior']

df['age_group'] = pd.qcut(df['age'], q=4, labels=labels)
```

This segmentation supports demographic insights and aligns with business marketing practices.

Chart:
Bar Chart – Average Purchase Amount by Age Group
Provides a clear view of which demographic contributes most to revenue.

2.4 Encoding Textual Frequencies

Certain behavioral attributes, such as **frequency of purchases**, were textual (e.g., "Monthly", "Quarterly").
To make them analytically meaningful, they were mapped to equivalent **frequency in days**:

```
frequency_mapping = {
    'Fortnightly': 14, 'Weekly': 7, 'Monthly': 30,
    'Quarterly': 90, 'Bi-Weekly': 14, 'Annually': 365,
    'Every 3 Months': 90
}
df['purchase_frequency_days'] = df['frequency_of_purchases'].map(frequency_mapping)
```

This allows quantitative comparison of buying patterns.

Chart:
Scatter Plot – Purchase Frequency vs Purchase Amount

2.5 Removal of Redundant Features

The variables *discount_applied* and *promo_code_used* contained identical values for every record:

```
(df['discount_applied'] == df['promo_code_used']).all()   # True
```

Thus, **promo_code_used** was removed as redundant:

```
df = df.drop('promo_code_used', axis=1)
```

Streamlining features reduces noise and simplifies SQL analysis.

3. SQL-Based Exploratory Data Analysis

Using SQL queries (see *customer_trends.sql*), Exploratory Data Analysis was performed to uncover the following insights:

1. Revenue distribution by gender
2. Identification of high-value discount users
3. Top 5 products with highest average review rating
4. Average purchase amounts across Standard vs Express shipping types
5. Subscription-status-based comparison of average spend and total revenue
6. Products with the highest discount utilization percentages
7. Customer segmentation into New, Returning, and Loyal groups
8. Top 3 most-purchased products within each product category
9. Subscription behavior among repeat buyers (>5 previous purchases)
10. Revenue contribution across engineered age groups

4. Key Insights

- **Category-wise imputation** improved the quality of review data.
- **Young Adults and Middle-aged** groups contributed the highest revenue.
- **Subscribed customers** spent more and had a higher total revenue impact.
- **Discount-driven items** tended to align with top-selling products.

- Product ratings and purchase frequency provided valuable signals for product strategy.

- Loyal customers formed a significant high-value segment.

- Category-wise analysis revealed distinct performance patterns and helped identify strong product clusters.

5. Tools and Technologies

- **Python:** Pandas, NumPy

- **SQL:** PostgreSQL(for EDA)

- **Jupyter Notebook:** Preprocessing and validation

- **Power BI Dashboard:** Interactive visualization

- **CSV Dataset:** Customer shopping behavior file

6. Conclusion

This project demonstrates a complete data analytics pipeline, from preprocessing and feature engineering to SQL-driven analysis and dashboard visualization.
By cleaning, enriching, and analyzing the data, we uncover actionable insights into customer behavior, demographic influence, shipping preferences, discount patterns, and overall purchasing trends.