Customer Shopping Behavior Analysis

Transforming raw customer data into actionable business insights.

Project Overview: From Data to Decisions

This project delves into customer shopping behavior, aiming to uncover key purchasing patterns, segment customers, and identify primary revenue drivers. It simulates a comprehensive data analyst workflow, integrating various tools and techniques.

Our journey spans data cleaning in Python, business analysis with SQL, and culminates in an interactive Power BI dashboard.

Python (Jupyter)

Data Cleaning & EDA

SQL (PostgreSQL)

Business Analysis

Power BI

Interactive Dashboard

The Dataset: A Glimpse into Customer Habits

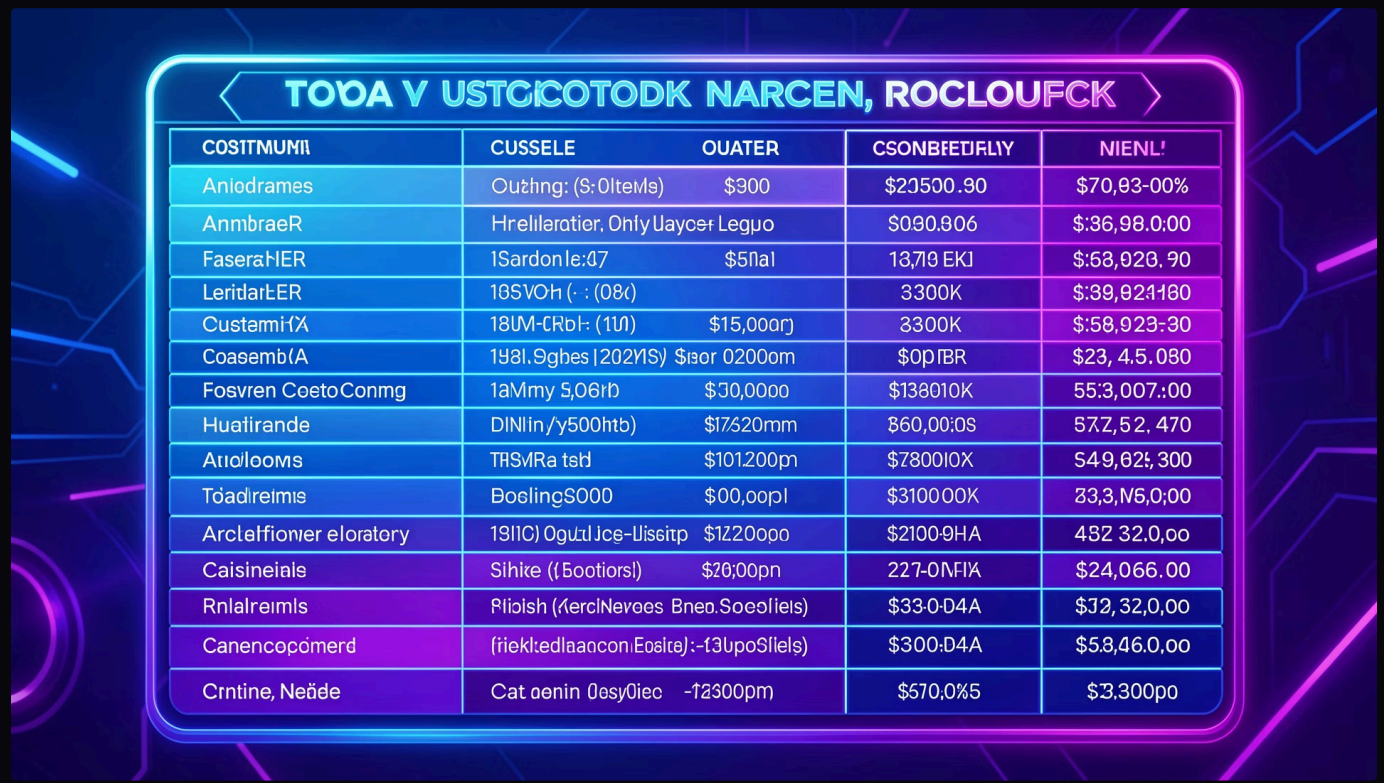
Our analysis is built upon a rich dataset detailing various aspects of customer shopping behavior.

Dataset Metrics:

- **Rows:** 4000 entries
- **Columns:** 18 distinct attributes

Key Data Points:

- Customer ID, Age, Gender
- Item Purchased, Category, Purchase Amount
- Location, Size, Color, Season
- Review Rating, Subscription Status
- Shipping Type, Discount Applied, Promo Code Used
- Previous Purchases, Payment Method, Frequency of Purchases



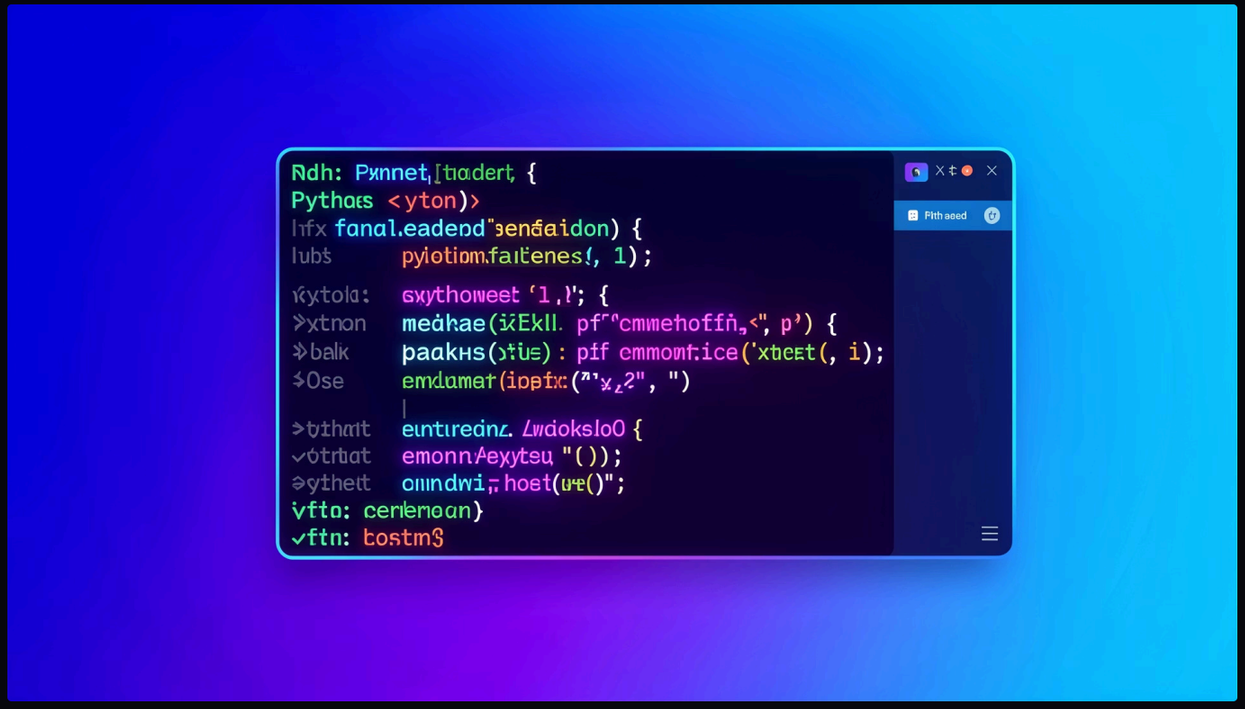
Exploratory Data Analysis (EDA) with Python

Leveraging Python's data manipulation capabilities, we initiated our analysis by performing comprehensive EDA to understand the dataset's structure and identify initial patterns.

Data Loading & Initial Inspection

```
import pandas as pd
df =
pd.read_csv("customer_data.csv")
print(df.head())
```

The initial steps involved loading the data and using `df.head()` to preview the first few rows, ensuring correct data ingestion.



Data Information & Summary Statistics

```
df.info()
df.describe(include='all')
```

We utilized `df.info()` to check data types and non-null counts, and `df.describe(include='all')` for a statistical summary, revealing distributions and potential issues.

Data Cleaning & Feature Engineering

To ensure data quality and prepare for deeper analysis, several cleaning and transformation steps were performed.

Handling Missing Values

```
df['review_rating'] = df['review_rating'].fillna(4.1)
```

Identified 37 missing values in 'Review Rating' which were imputed with 4.1 to maintain data integrity for subsequent analysis.

Standardizing Column Names

```
df.columns = df.columns.str.lower().str.replace(' ','_')
df = df.rename(columns=
{'purchase_amount_(usd)':'purchase_amount'})
```

Column names were converted to lowercase and spaces replaced with underscores for consistency and easier access.

Creating Age Groups

```
labels = ['Young Adult','Adult','Middle aged','Senior']
df['age_group'] = pd.qcut(df['age'],q=4,labels=labels)
```

A new 'age_group' column was engineered to categorize customers into segments like 'Young Adult' and 'Senior' based on age quartiles.

Numerical Purchase Frequency

```
frequency_mapping = {'Fortnightly': 14, 'Weekly': 7, ...}
df['purchase_frequency_days'] =
df['frequency_of_purchases'].map(frequency_mapping)
```

Transformed categorical purchase frequencies into numerical days (e.g., 'Weekly' to 7) for quantitative analysis.

Redundancy Elimination: Discount & Promo Codes

An important step in data cleaning involves identifying and eliminating redundant data to streamline the dataset and prevent analytical errors.

Identifying Redundancy

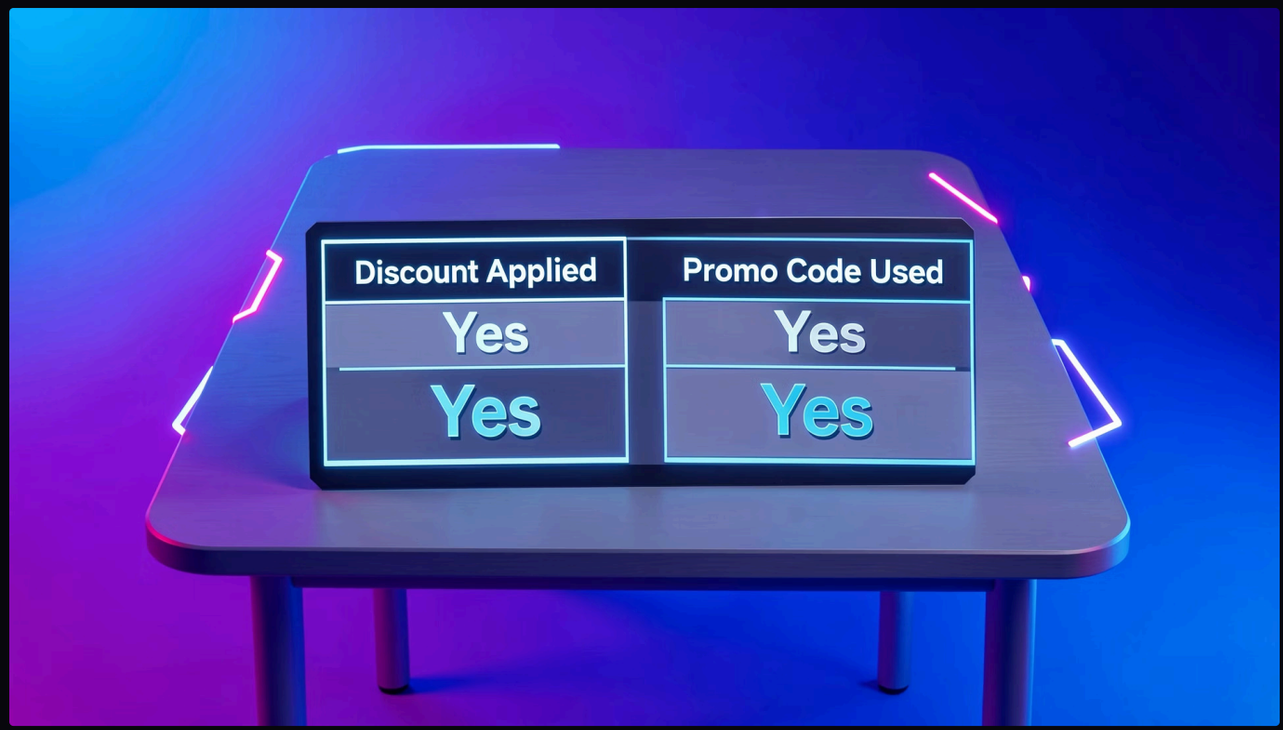
```
(df['discount_applied'] ==
df['promo_code_used']).all()

# Output: np.True_
```

We discovered that the 'Discount Applied' and 'Promo Code Used' columns contained identical values for all entries, indicating a perfect correlation and redundancy.

Streamlining the Dataset

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

To optimize our dataset, the 'promo_code_used' column was removed, as its information was fully captured by 'discount_applied'.



Data Ingestion: From Pandas to PostgreSQL

Once cleaned and prepared, the dataset was seamlessly loaded into a PostgreSQL database, setting the stage for advanced SQL queries and business analysis.

Establishing Database Connection

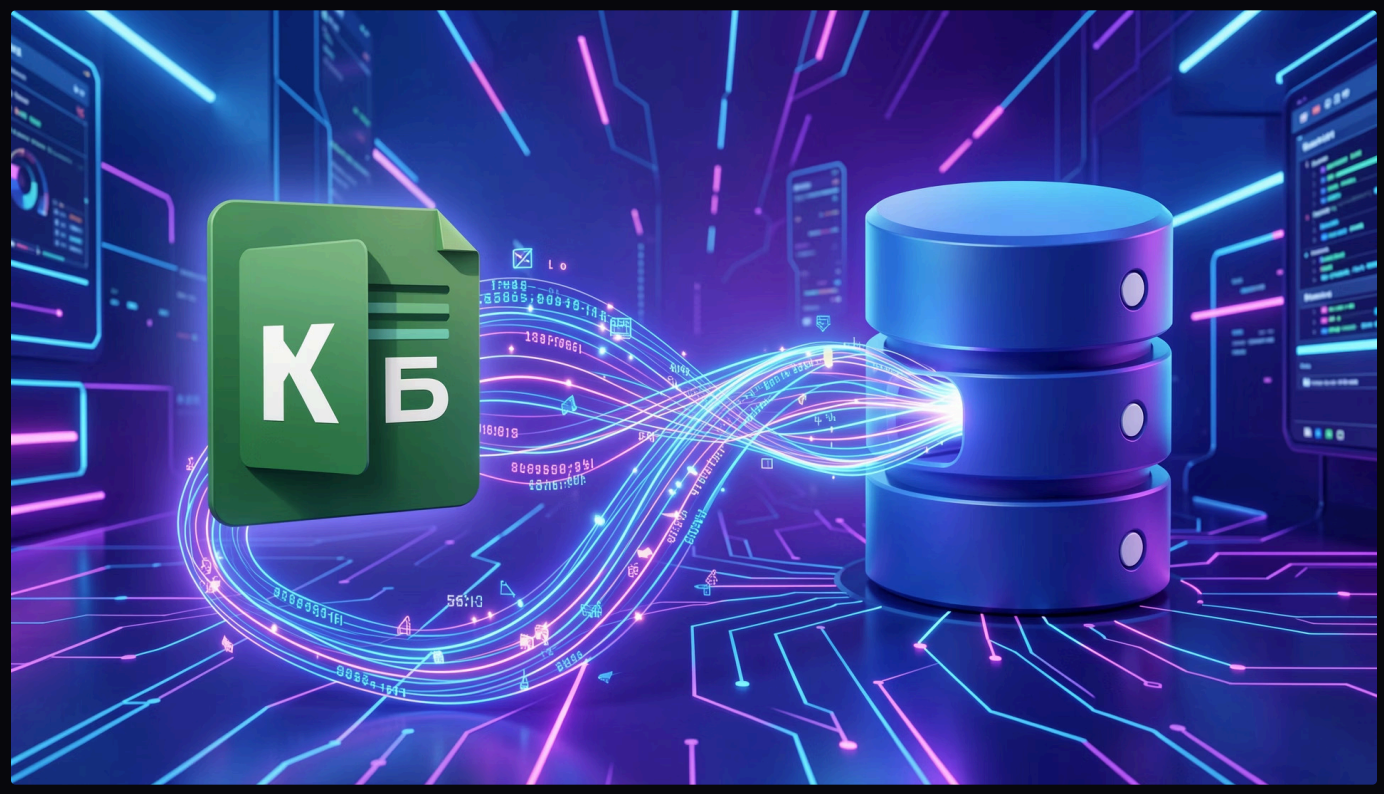
```
from sqlalchemy import
create_engine
engine =
create_engine('postgresql://postgre
s:password@localhost:5433/custom
er_database')
```

A robust connection was established using `create_engine` from SQLAlchemy, ensuring secure and efficient data transfer.

Loading Dataframe to SQL

```
table_name = "customer"
df.to_sql(table_name, engine,
if_exists="replace", index=False)
print(f"Data successfully loaded into
table '{table_name}'")
```

The processed DataFrame was then loaded into a new table named "customer" in the PostgreSQL database, ready for SQL-based business intelligence.



SQL Insights: Unlocking Business Intelligence

Using SQL, we performed structured queries to answer critical business questions, revealing key trends and customer preferences.

Gender-Based Revenue

Males contributed $157,890 in revenue, significantly higher than females at $75,191.

Discount User Spending

Identified 9 customers who used discounts yet spent above the average purchase amount, highlighting high-value, discount-sensitive segments.

Top 5 Reviewed Products

Gloves, Sandals, and Boots emerged as products with the highest average review ratings, indicating strong customer satisfaction.

Shipping Type vs. Spend

Express shipping users had a slightly higher average purchase amount ($60.48) compared to standard shipping ($58.46).

Deep Dive: Subscription, Loyalty, and Age Segmentation

Further SQL analysis provided granular insights into customer loyalty, subscription impact, and age-group contributions to revenue.

Subscriber vs. Non-Subscriber Spending

Subscribers (1053 customers) showed an average spend of $59.49 and total revenue of $62,645. Non-subscribers (2847 customers) had a slightly higher average spend of $59.87 and a total revenue of $170,436.

Product Discount Analysis

Hats, Sneakers, and Coats had the highest percentage of purchases with discounts applied, suggesting these items are frequently part of promotional offers.

Customer Loyalty Segmentation

Customers were segmented into 'New' (83), 'Returning' (701), and 'Loyal' (3116) based on their previous purchase history, with Loyal customers forming the largest group.

Age Group Revenue Contribution

Young Adults generated the highest total revenue at $62,143, followed by Middle-aged customers with $59,197.

Dashboard & Recommendations: Actionable Insights

The Power BI dashboard visualizes these insights, leading to strategic business recommendations.

Key Insights:

- Non-subscribers represent significant revenue potential.
- Young adults demonstrate the highest purchase frequency.
- Clothing category consistently dominates sales volume.
- Discount usage, while boosting conversion, impacts margins.
- Express shipping customers typically exhibit higher spending.

Strategic Recommendations:

Targeted Subscription Campaigns

Focus efforts on converting high-spending non-subscribers.

Enhanced Loyalty Programs

Reward repeat buyers to foster long-term engagement.

Optimized Discount Strategies

Balance conversion with profitability.

Promote High-Rated Products

Leverage positive reviews for increased sales.