**PART 2: WORKING WOTH PYTHON PANDAS DATAFRAMES**

1. **Purchased products (purchase count and quantity purchased) per month**

import pandas as pd

# Load CSV files into pandas DataFrames

purchase\_df = pd.read\_csv('C://Users//ranja//OneDrive//Documents//Interviews//Copeland//Dataset//Data Engineer Take-Home//CustomerPurchaseHistorySampleData.csv')

product\_df = pd.read\_csv('C://Users//ranja//OneDrive//Documents//Interviews//Copeland//Dataset//Data Engineer Take-Home//ProductSampleData.csv')

print(product\_df)

# Convert 'PurchaseDate' column to datetime object

purchase\_df[' PurchaseDate'] = pd.to\_datetime(purchase\_df[' PurchaseDate'])

# Merge CustomerPurchaseHistory and Product DataFrames

merged\_df = pd.merge(purchase\_df, product\_df, on=' ProductID')

# Extract month from 'PurchaseDate' column

merged\_df['purchase\_month'] = merged\_df[' PurchaseDate'].dt.month

# Group by 'purchase\_month' and aggregate purchase count and quantity purchased

result\_df = merged\_df.groupby('purchase\_month').agg({

' ProductID': 'count',

' Quantity': 'sum'

}).reset\_index()

# Rename columns for clarity

result\_df = result\_df.rename(columns={'ProductID': 'PurchaseCount', 'Quantity': 'TotalQuantityPurchased'})

# Sort by 'purchase\_month' column

result\_df = result\_df.sort\_values(by='purchase\_month')

print(result\_df)

**OUTPUT**:

purchase\_month ProductID Quantity

0 1 167 86964

1 2 170 86131

2 3 185 86447

3 4 148 70199

4 5 167 80074

1. 6 163 87466
2. **Average age of customer per product sold.**

import pandas as pd

# Load CSV files into pandas DataFrames

purchase\_df = pd.read\_csv('C://Users//ranja//OneDrive//Documents//Interviews//Copeland//Dataset//Data Engineer Take-Home//CustomerPurchaseHistorySampleData.csv')

customer\_df = pd.read\_csv('C://Users//ranja//OneDrive//Documents//Interviews//Copeland//Dataset//Data Engineer Take-Home//CustomerDemographicsSampleData.csv')

# Merge CustomerPurchaseHistory and CustomerDemographics DataFrames

merged\_df = pd.merge(purchase\_df, customer\_df, left\_on='Customer',right\_on='CustomerID')

print(merged\_df)

# Group by ProductID and calculate the average age of customers

average\_age\_per\_product = merged\_df.groupby(' ProductID')['Age'].mean().reset\_index()

# Optionally, round the average age for readability

average\_age\_per\_product['AverageAge'] = average\_age\_per\_product['Age'].round(2)

print(average\_age\_per\_product)

|  | **ProductID** | **Age** | **AverageAge** |
| --- | --- | --- | --- |
| **0** | 1 | 60.500000 | 60.50 |
| **1** | 2 | 64.400000 | 64.40 |
| **2** | 3 | 57.166667 | 57.17 |
| **3** | 4 | 72.166667 | 72.17 |
| **4** | 5 | 58.133333 | 58.13 |
| **...** | ... | ... | ... |
| **95** | 96 | 71.500000 | 71.50 |
| **96** | 97 | 38.857143 | 38.86 |
| **97** | 98 | 63.777778 | 63.78 |
| **98** | 99 | 60.210526 | 60.21 |
| **99** | 100 | 74.857143 | 74.86 |

1. ws × 3 columns
2. **Which products are purchased most by age group?**

import pandas as pd

# Load CSV files into pandas DataFrames

purchase\_df = pd.read\_csv('C://Users//ranja//OneDrive//Documents//Interviews//Copeland//Dataset//Data Engineer Take-Home//CustomerPurchaseHistorySampleData.csv')

customer\_df = pd.read\_csv('C://Users//ranja//OneDrive//Documents//Interviews//Copeland//Dataset//Data Engineer Take-Home//CustomerDemographicsSampleData.csv')

product\_df = pd.read\_csv('C://Users//ranja//OneDrive//Documents//Interviews//Copeland//Dataset//Data Engineer Take-Home//ProductSampleData.csv')

# Merge CustomerPurchaseHistory and CustomerDemographics DataFrames

merged\_df = pd.merge(purchase\_df, customer\_df, left\_on='Customer', right\_on='CustomerID')

# Merge with Product DataFrame to get ProductName

merged\_df = pd.merge(merged\_df, product\_df, on=' ProductID')

# Define age groups (you can customize the age groups as needed)

age\_bins = [ 18, 28, 38, 48, 58, float('inf')]

age\_labels = ['19-28', '29-38', '39-48', '49-59', '60+']

merged\_df['AgeGroup'] = pd.cut(merged\_df['Age'], bins=age\_bins, labels=age\_labels, right=False)

# Group by AgeGroup and ProductName, counting the number of purchases

purchase\_count\_by\_age\_group = merged\_df.groupby(['AgeGroup', 'ProductName']).size().reset\_index(name='PurchaseCount')

# Find the product with the highest purchase count for each age group

idx = purchase\_count\_by\_age\_group.groupby(['AgeGroup'])['PurchaseCount'].transform(max) == purchase\_count\_by\_age\_group['PurchaseCount']

most\_purchased\_products\_by\_age\_group = purchase\_count\_by\_age\_group[idx]

print(most\_purchased\_products\_by\_age\_group)

|  | **AgeGroup** | **ProductName** | **PurchaseCount** |
| --- | --- | --- | --- |
| **81** | 18-28 | 'ProductName82' | 6 |
| **111** | 29-38 | 'ProductName19' | 4 |
| **191** | 29-38 | 'ProductName91' | 4 |
| **197** | 29-38 | 'ProductName97' | 4 |
| **286** | 39-48 | 'ProductName87' | 5 |
| **353** | 49-59 | 'ProductName57' | 8 |
| **465** | 60+ | 'ProductName68' | 14 |

**4.Repeat customers?**

import pandas as pd

# Load CSV file into pandas DataFrame

purchase\_df = pd.read\_csv('C://Users//ranja//OneDrive//Documents//Interviews//Copeland//Dataset//Data Engineer Take-Home//CustomerPurchaseHistorySampleData.csv')

# Group by customer and count the number of purchases

customer\_purchase\_count = purchase\_df.groupby('Customer')[' PurchaseDate'].count().reset\_index()

# Filter customers who have made more than one purchase

repeat\_customers = customer\_purchase\_count[customer\_purchase\_count[' PurchaseDate'] > 1]

print(repeat\_customers)

|  | **Customer** | **PurchaseDate** |
| --- | --- | --- |
| **0** | 1 | 10 |
| **1** | 2 | 15 |
| **2** | 3 | 14 |
| **3** | 4 | 13 |
| **4** | 5 | 15 |
| **...** | ... | ... |
| **95** | 96 | 9 |
| **96** | 97 | 5 |
| **97** | 98 | 12 |
| **98** | 99 | 9 |
| **99** | 100 | 11 |

99 rows × 2 columns

​

**Based on the data set provide some other information that could be useful to the business.**

Popular Products: Identify the top-selling products to understand customer preferences and trends. This information can help in inventory management, marketing strategies, and product development.

Customer Segmentation: Analyze customer demographics and purchase behavior to segment customers based on age, location, or purchasing patterns. This segmentation can help tailor marketing campaigns, promotions, and product offerings to specific customer groups.

Average Purchase Amount: Calculate the average purchase amount per transaction to understand customer spending habits. This information can guide pricing strategies and help in setting sales targets.

Repeat Purchase Rate: Determine the percentage of repeat customers compared to one-time buyers. Repeat purchase rate indicates customer loyalty and satisfaction with the products and services offered.

Seasonal Trends: Identify seasonal trends in product purchases to anticipate demand fluctuations throughout the year. Understanding seasonal variations can aid in adjusting inventory levels, scheduling promotions, and optimizing marketing efforts.

Customer Lifetime Value (CLV): Calculate the CLV for each customer to estimate the potential revenue generated over the entire customer relationship. CLV helps prioritize customer acquisition and retention strategies, as well as identify high-value customer segments.

Customer Feedback Analysis: Collect and analyze customer feedback and reviews to gain insights into product satisfaction, identify areas for improvement, and address customer concerns promptly. Monitoring customer feedback helps maintain positive brand perception and customer loyalty.

Competitive Analysis: Monitor competitors' product offerings, pricing strategies, and promotional activities to stay informed about market trends and identify opportunities for differentiation and innovation.