Data Analytics Midterm

Chris Thai

Nov 21 2024

Part I

Data Analytics Overview and Planning

Data Analytics Overview

Data Analytics for Retail - Collecting, processing, analyzing data to inform decisions to improve the business

Example Decisions

- Items to sell and at what price
- Incentives to keep employees motivated
- Potential customers to target
- Business locations
- Areas to reduce costs

Types and Quality of Data

Customer Data

- 100 entries of customers
- Names, Email, Phone Numbers, and Address
- All Strings except for customer number
- Very bad. Lots of missing or erroneous entries

Product Data

- 100 entries of different products
- Product ID, price, and amt in stock are numerical
- Product Name and Catagory are strings
- Better, missing data, but have 74/100 entries with complete data

Sales Data

- 150 entries of sales
- Sales ID, Customer ID, Product ID, Sale Date, and Sale Amount
- Date is a date, but everything else is number
- Better, missing data, but have 129/150 complete entries
- Interesting: only 3 customers and 3 products account for all of the sales

Part II

Data Cleaning and Transformation in Excel

Cleaning

Customer Data

- Left empty customer names blank
- Completed emails if they left out the “.com”
- Removed all non digits from phone numbers
- Pulled out state and zip codes from address

Product Data

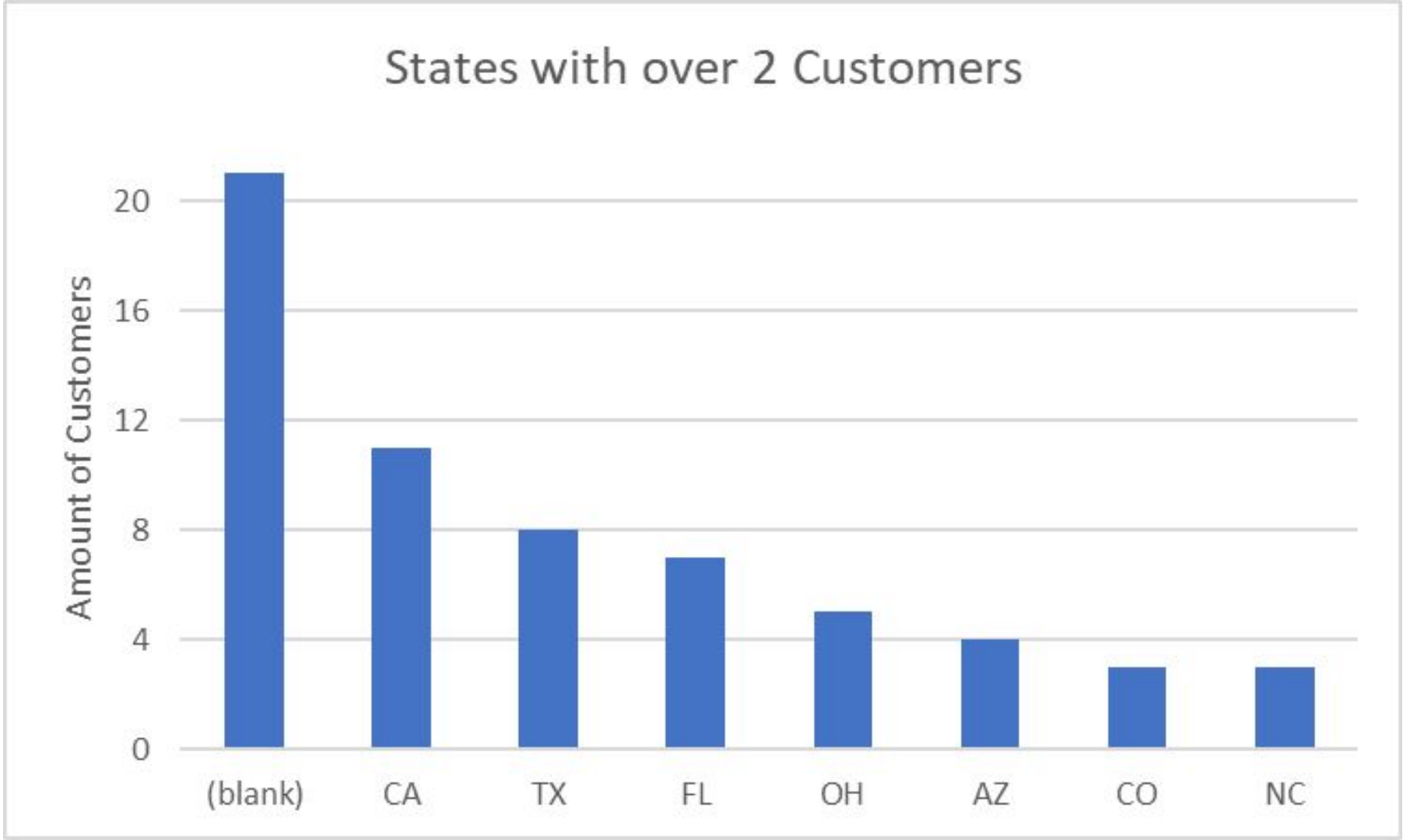
- 100 entries of different products
- Product ID, price, and amt in stock are numerical
- Product Name and Catagory are strings
- Removed spaces so categories matched
- Better, missing data, but have 74/100 entries with complete data

Sales Data

- 150 entries of sales
- Sales ID, Customer ID, Product ID, Sale Date, and Sale Amount
- Date is a date, but everything else is number
- Better, missing data, but have 129/150 complete entries
- Interesting: only 3 customers and 3 products account for all of the sales

Part III

Exploratory Data Analysis in Excel


States with over 2 Customers
Amount of Customers
20
16
12
8
4
0
(blank)
CA
TX
FL
OH
AZ
CO
NC

Product Categories	Count of Products	Average of Price
Electronics	16	$ 232.49
Furniture	12	$ 178.24
Kitchen	11	$ 40.27
(blank)	10	$ 106.59
Appliances	7	$ 392.85
Outdoor	5	$ 130.59
Home Appliances	4	$ 51.49
Home Decor	4	$ 43.49
Fashion	3	$ 56.66
Musical Instruments	2	$ 399.99
Travel	2	$ 139.99
Home Automation	2	$ 75.49
Pet Supplies	2	$ 64.99
Home Organization	2	$ 19.99
Fitness	1	$ 599.99
Tools	1	$ 99.99
Auto	1	$ 59.99
Home Safety	1	$ 19.99
Grand Total	86	$ 156.99

Customer ID	Product 1 Quantity	Product 2 Quantity	Product 3 Quantity		Grand Total
1	1450	600	60		2110
2	100	40	1110		1250
3	750	1080	30		1860
Grand Total	**2300**	**1720**	**1200**		**5220**

Part IV

Automate Data Manipulation in Excel

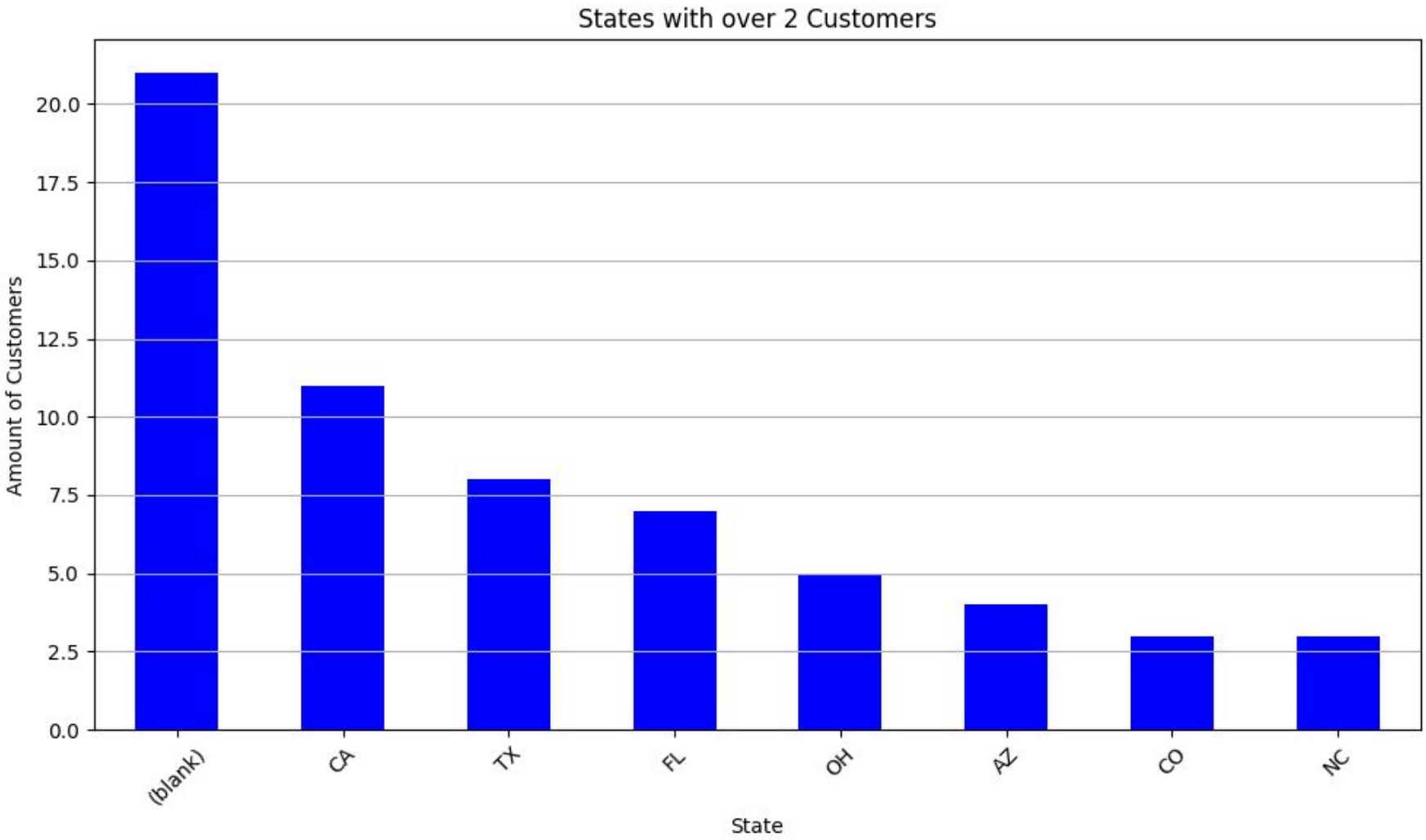
Created a Macro called “format”

1. Selects all in a worksheet
2. Auto sizes columns to fit
3. Creates a filter
4. Keeps everything highlighted and ends

Macro stored in workbook.

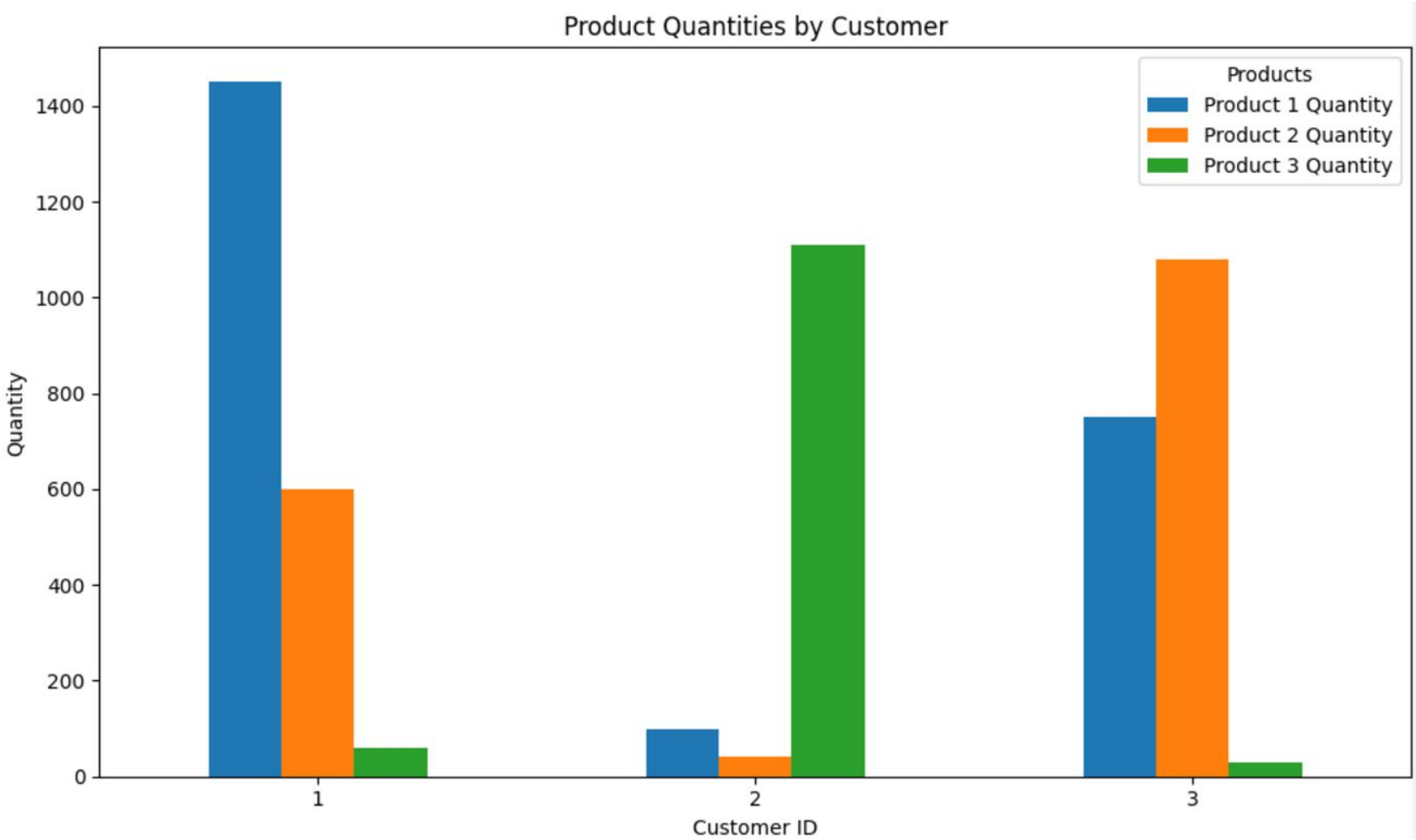
Part V

Data Analysis in Python


States with over 2 Customers
Amount of Customers
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
(blank)
CA
TX
FL
OH
AZ
CO
NC
State

Product Categories	Count of Products	Average of Price
Grand Total	86	$150.73
Electronics	16	$232.49
Furniture	12	$178.24
Kitchen	11	$40.27
(blank)	10	$106.59
Appliances	7	$392.85
Outdoor	5	$130.59
Home Appliances	4	$51.49
Home Decor	4	$43.49
Fashion	3	$56.66
Travel	2	$139.99
Pet Supplies	2	$64.99
Home Organization	2	$19.99
Home Automation	2	$75.49
Musical Instruments	2	$399.99
Auto	1	$59.99
Fitness	1	$599.99
Home Safety	1	$19.99
Tools	1	$99.99

Customer ID	Product 1 Quantity	Product 2 Quantity	Product 3 Quantity	Grand Total
1	1450	600	60	2110
2	100	40	1110	1250
3	750	1080	30	1860
Grand Total	2300	1720	1200	5220


Product Quantities by Customer
Products
Product 1 Quantity
Product 2 Quantity
Product 3 Quantity
Quantity
0
200
400
600
800
1000
1200
1400
1
2
3
Customer ID

	ProductID	ProductName	Category	Price	StockQuantity	InventoryValue
0	1	Laptop	Electronics	999.99	20	19999.80
1	2	Smartphone	Electronics	499.99	30	14999.70
2	3	Desk	Furniture	199.99	15	2999.85

Part VI

Database Management and SQL

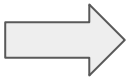
1. Identifying Best-Selling Products

Input

```
ALTER TABLE NewTable RENAME TO product_data_cleaned;


SELECT
    p.ProductName,
    SUM(s.SaleAmount) AS TotalQuantitySold
FROM
    sales_data_cleaned s
JOIN
   product_data_cleaned p ON s.ProductID = p.ProductID
GROUP BY
    p.ProductID, p.ProductName
ORDER BY
    TotalQuantitySold DESC
LIMIT 5; -- Top 5 best-selling products
```

Output



ProductName	TotalQuantitySold
Laptop	4,600
Smartphone	3,440
Desk	2,400

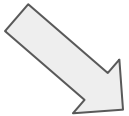
Notes:

1. **For some reason, the products table kept on reverting to the name “NewTable”, so we had to write a line of code to keep its new name**
2. **From the Sales table, it’s clear that only 3 products were sold**

2. Understanding Customer Data

Input

```sql
SELECT
    c.CustomerID,
    c.CustomerName,
    SUM(CASE WHEN p.ProductName = 'Laptop' THEN s.SaleAmount ELSE 0 END) AS LaptopsPurchased,
    SUM(CASE WHEN p.ProductName = 'Smartphone' THEN s.SaleAmount ELSE 0 END) AS SmartphonesPurchased,
    SUM(CASE WHEN p.ProductName = 'Desk' THEN s.SaleAmount ELSE 0 END) AS DesksPurchased,
    SUM(s.SaleAmount * p.Price) AS TotalSpent
FROM
    sales_data_cleaned s
JOIN
    customer_data_cleaned c ON s.CustomerID = c.CustomerID
JOIN
    product_data_cleaned p ON s.ProductID = p.ProductID
GROUP BY
    c.CustomerID, c.CustomerName
ORDER BY
    TotalSpent DESC;
```



Output

CustomerID	CustomerName	LaptopsPurchased	SmartphonesPurchased	DesksPurchased	TotalSpent
1	John Doe	2,900	1,200	120	3,523,957.8
3		1,500	2,160	60	2,591,962.8
2	Jane Smith	200	80	2,220	683,975

3. Sales Trends Over Time

Input

```sql
SELECT
    SaleYear,
    SaleMonth,
     SUM(CASE WHEN p.ProductName = 'Laptop' THEN s.SaleAmount ELSE 0 END) AS LaptopsPurchased,
    SUM(CASE WHEN p.ProductName = 'Smartphone' THEN s.SaleAmount ELSE 0 END) AS SmartphonesPurchased,
    SUM(CASE WHEN p.ProductName = 'Desk' THEN s.SaleAmount ELSE 0 END) AS DesksPurchased,
    SUM(s.SaleAmount * p.Price) AS TotalSales
FROM
    sale_data_cleaned s
JOIN
    product_data_cleaned p ON s.ProductID = p.ProductID
GROUP BY
    SaleYear, SaleMonth
ORDER BY
    SaleYear, SaleMonth;
```

Output

SaleYear	SaleMonth	LaptopsPurchased	SmartphonesPurchased	DesksPurchased	TotalSales
2,023	1	500	320	300	719,988.8
2,023	2	500	400	240	747,988.6
2,023	3	500	400	360	771,987.4
2,023	4	500	400	300	759,988
2,023	5	500	400	300	759,988
2,023	6	500	400	300	759,988
2,023	7	400	400	240	647,989.6
2,023	8	400	320	120	583,991.6
2,023	9	300	160	180	415,993.6
2,023	10	500	240	60	631,992

Part VII

Data-Driven Decision Making

Insights

- Only 3 products are selling:Laptops, Smartphones, and Desks
  - Laptops have the highest sales and volume sold

- All other stocked quantities should be liquidated to focus on the 3 selling items

- Inventory for Laptops, Smartphones, and Desks should greatly increase as any one purchase greatly exceeds inventory

- Only 7 states have more than 2 customers, with the California having the most at 11
  - But, the highest category is “blank” with 21 customers. The company needs to improve its customer record keeping