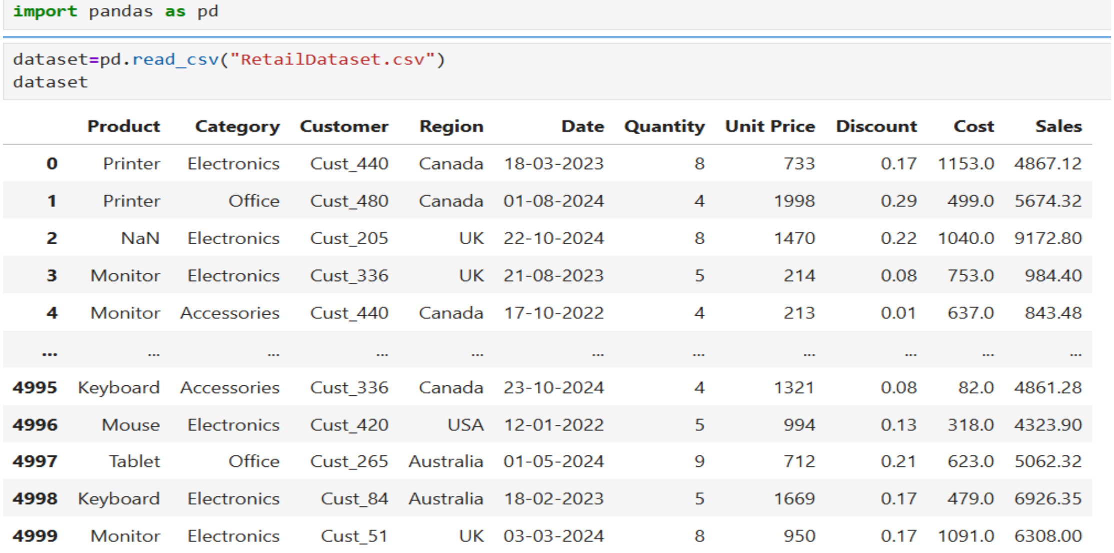
EXPLORATORY DATA ANALYSIS OF PROJECT

Project Description:

This project focuses on building an end-to-end Retail Sales & Profit Analysis pipeline. The workflow includes handling a messy retail dataset, performing data cleaning and exploratory data analysis (EDA) using Python (pandas), extracting business insights using SQL, and developing an interactive dashboard in Power BI.

The dataset contains financial metrics such as Sales, Cost, Profit, Discount, Quantity, and Unit Price along with business dimensions like Product, Category, Customer, Region, and Date. The objective is to uncover revenue trends, profitability drivers, discount impact, and customer behavior to support data-driven decision making.

EDA using Python (pandas):

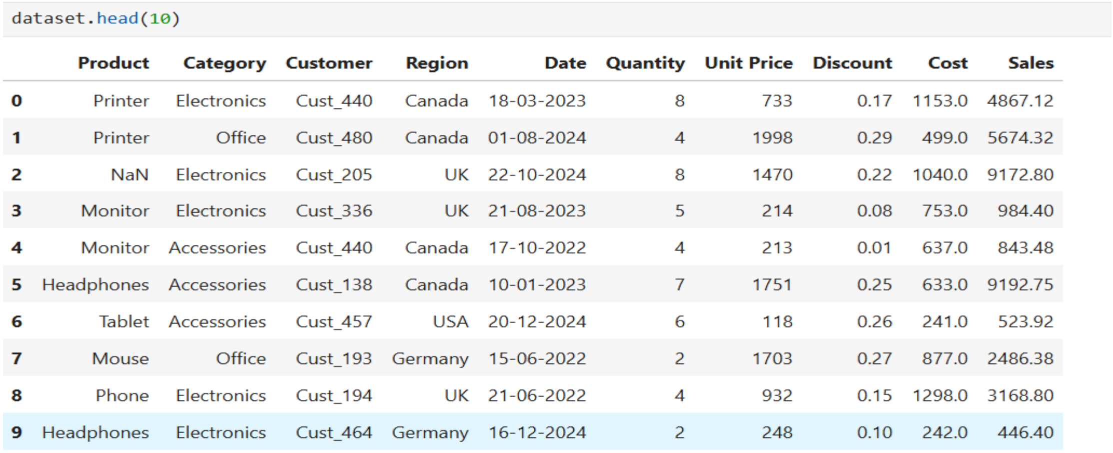


```
import pandas as pd
```

```
dataset=pd.read_csv("RetailDataset.csv")
dataset
```

	Product	Category	Customer	Region	Date	Quantity	Unit Price	Discount	Cost	Sales
0	Printer	Electronics	Cust_440	Canada	18-03-2023	8	733	0.17	1153.0	4867.12
1	Printer	Office	Cust_480	Canada	01-08-2024	4	1998	0.29	499.0	5674.32
2	NaN	Electronics	Cust_205	UK	22-10-2024	8	1470	0.22	1040.0	9172.80
3	Monitor	Electronics	Cust_336	UK	21-08-2023	5	214	0.08	753.0	984.40
4	Monitor	Accessories	Cust_440	Canada	17-10-2022	4	213	0.01	637.0	843.48
...	...	...	...	...	...	...	...	...	...	...
4995	Keyboard	Accessories	Cust_336	Canada	23-10-2024	4	1321	0.08	82.0	4861.28
4996	Mouse	Electronics	Cust_420	USA	12-01-2022	5	994	0.13	318.0	4323.90
4997	Tablet	Office	Cust_265	Australia	01-05-2024	9	712	0.21	623.0	5062.32
4998	Keyboard	Electronics	Cust_84	Australia	18-02-2023	5	1669	0.17	479.0	6926.35
4999	Monitor	Electronics	Cust_51	UK	03-03-2024	8	950	0.17	1091.0	6308.00

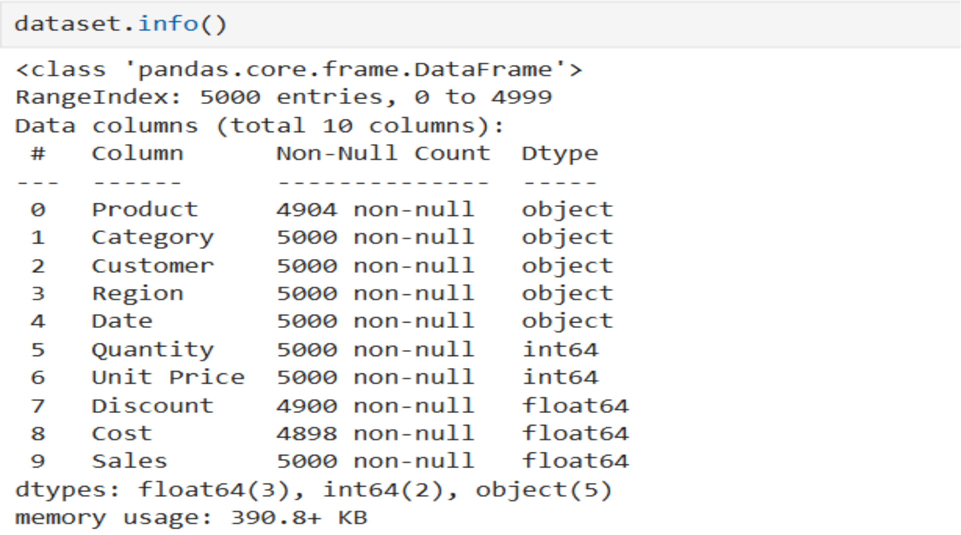
head() – Preview first rows to understand dataset structure.



```
dataset.head(10)
```

	Product	Category	Customer	Region	Date	Quantity	Unit Price	Discount	Cost	Sales
0	Printer	Electronics	Cust_440	Canada	18-03-2023	8	733	0.17	1153.0	4867.12
1	Printer	Office	Cust_480	Canada	01-08-2024	4	1998	0.29	499.0	5674.32
2	NaN	Electronics	Cust_205	UK	22-10-2024	8	1470	0.22	1040.0	9172.80
3	Monitor	Electronics	Cust_336	UK	21-08-2023	5	214	0.08	753.0	984.40
4	Monitor	Accessories	Cust_440	Canada	17-10-2022	4	213	0.01	637.0	843.48
5	Headphones	Accessories	Cust_138	Canada	10-01-2023	7	1751	0.25	633.0	9192.75
6	Tablet	Accessories	Cust_457	USA	20-12-2024	6	118	0.26	241.0	523.92
7	Mouse	Office	Cust_193	Germany	15-06-2022	2	1703	0.27	877.0	2486.38
8	Phone	Electronics	Cust_194	UK	21-06-2022	4	932	0.15	1298.0	3168.80
9	Headphones	Electronics	Cust_464	Germany	16-12-2024	2	248	0.10	242.0	446.40

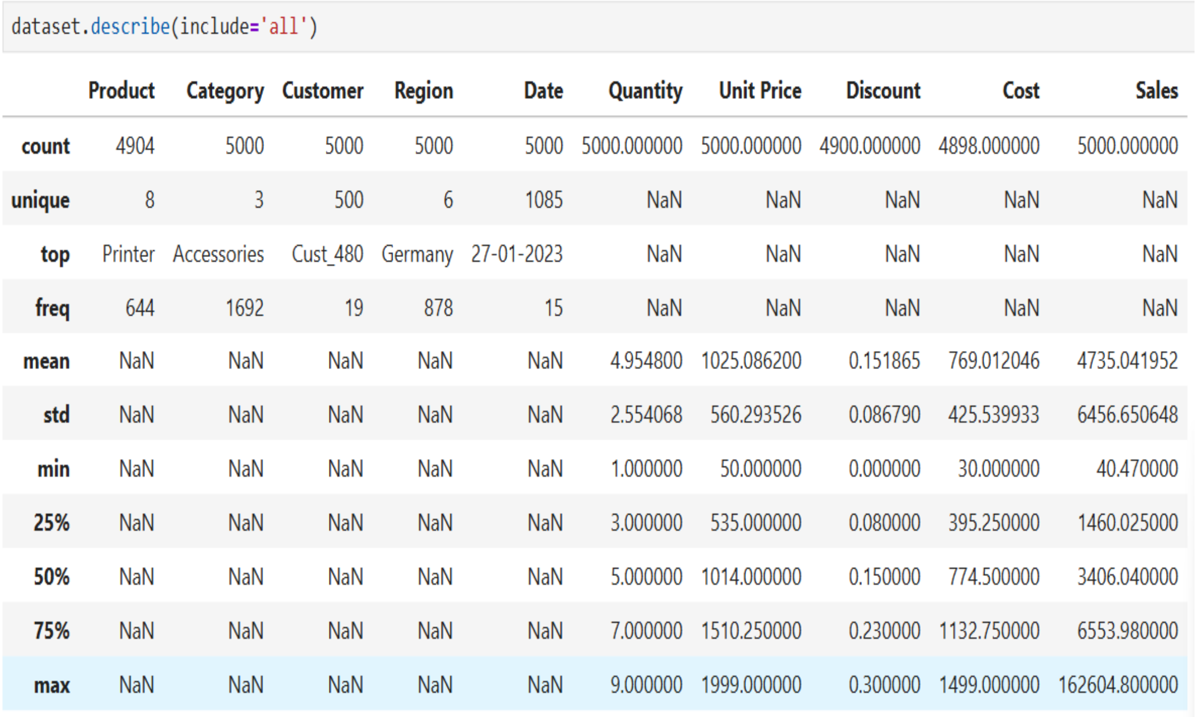
info() – Check data types and missing values.



```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Product     4904 non-null   object 
 1   Category    5000 non-null   object 
 2   Customer    5000 non-null   object 
 3   Region      5000 non-null   object 
 4   Date        5000 non-null   object 
 5   Quantity    5000 non-null   int64  
 6   Unit Price  5000 non-null   int64  
 7   Discount    4900 non-null   float64
 8   Cost        4898 non-null   float64
 9   Sales       5000 non-null   float64
dtypes: float64(3), int64(2), object(5)
memory usage: 390.8+ KB
```

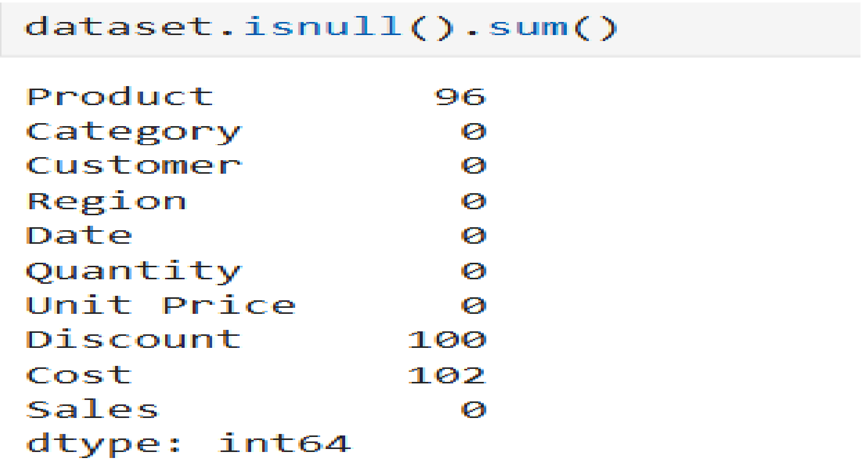
Describe() – Generate statistical summary of numerical columns.



```
dataset.describe(include='all')
```

	Product	Category	Customer	Region	Date	Quantity	Unit Price	Discount	Cost	Sales
count	4904	5000	5000	5000	5000	5000.000000	5000.000000	4900.000000	4898.000000	5000.000000
unique	8	3	500	6	1085	NaN	NaN	NaN	NaN	NaN
top	Printer	Accessories	Cust_480	Germany	27-01-2023	NaN	NaN	NaN	NaN	NaN
freq	644	1692	19	878	15	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	NaN	4.954800	1025.086200	0.151865	769.012046	4735.041952
std	NaN	NaN	NaN	NaN	NaN	2.554068	560.293526	0.086790	425.539933	6456.650648
min	NaN	NaN	NaN	NaN	NaN	1.000000	50.000000	0.000000	30.000000	40.470000
25%	NaN	NaN	NaN	NaN	NaN	3.000000	535.000000	0.080000	395.250000	1460.025000
50%	NaN	NaN	NaN	NaN	NaN	5.000000	1014.000000	0.150000	774.500000	3406.040000
75%	NaN	NaN	NaN	NaN	NaN	7.000000	1510.250000	0.230000	1132.750000	6553.980000
max	NaN	NaN	NaN	NaN	NaN	9.000000	1999.000000	0.300000	1499.000000	162604.800000

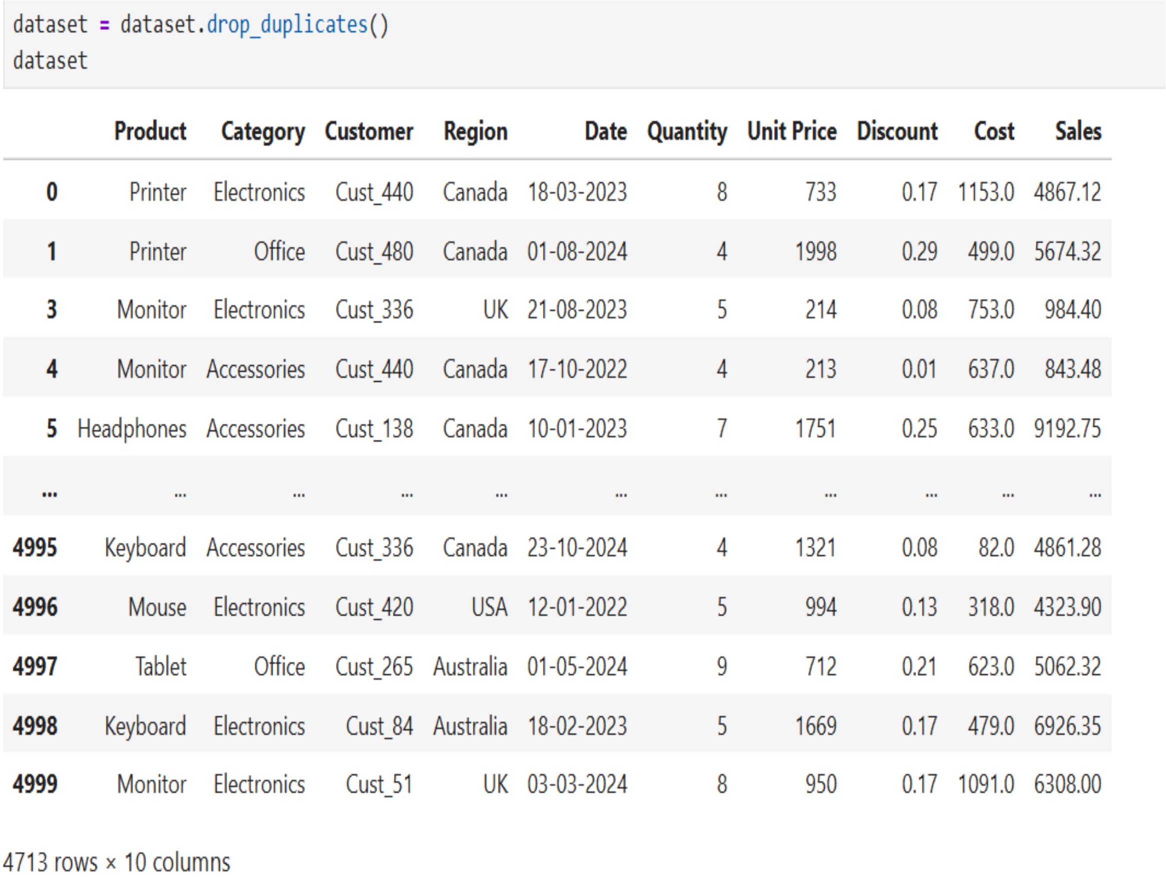
isnull().sum() – Identify null values across columns.



```
dataset.isnull().sum()
```

```
Product        96
Category        0
Customer        0
Region          0
Date            0
Quantity        0
Unit Price      0
Discount      100
Cost          102
Sales           0
dtype: int64
```

drop_duplicates() – Remove duplicate rows for data consistency.



```
dataset = dataset.drop_duplicates()
dataset
```

	Product	Category	Customer	Region	Date	Quantity	Unit Price	Discount	Cost	Sales
0	Printer	Electronics	Cust_440	Canada	18-03-2023	8	733	0.17	1153.0	4867.12
1	Printer	Office	Cust_480	Canada	01-08-2024	4	1998	0.29	499.0	5674.32
3	Monitor	Electronics	Cust_336	UK	21-08-2023	5	214	0.08	753.0	984.40
4	Monitor	Accessories	Cust_440	Canada	17-10-2022	4	213	0.01	637.0	843.48
5	Headphones	Accessories	Cust_138	Canada	10-01-2023	7	1751	0.25	633.0	9192.75
...	...	...	...	...	...	...	...	...	...	...
4995	Keyboard	Accessories	Cust_336	Canada	23-10-2024	4	1321	0.08	82.0	4861.28
4996	Mouse	Electronics	Cust_420	USA	12-01-2022	5	994	0.13	318.0	4323.90
4997	Tablet	Office	Cust_265	Australia	01-05-2024	9	712	0.21	623.0	5062.32
4998	Keyboard	Electronics	Cust_84	Australia	18-02-2023	5	1669	0.17	479.0	6926.35
4999	Monitor	Electronics	Cust_51	UK	03-03-2024	8	950	0.17	1091.0	6308.00

4713 rows × 10 columns

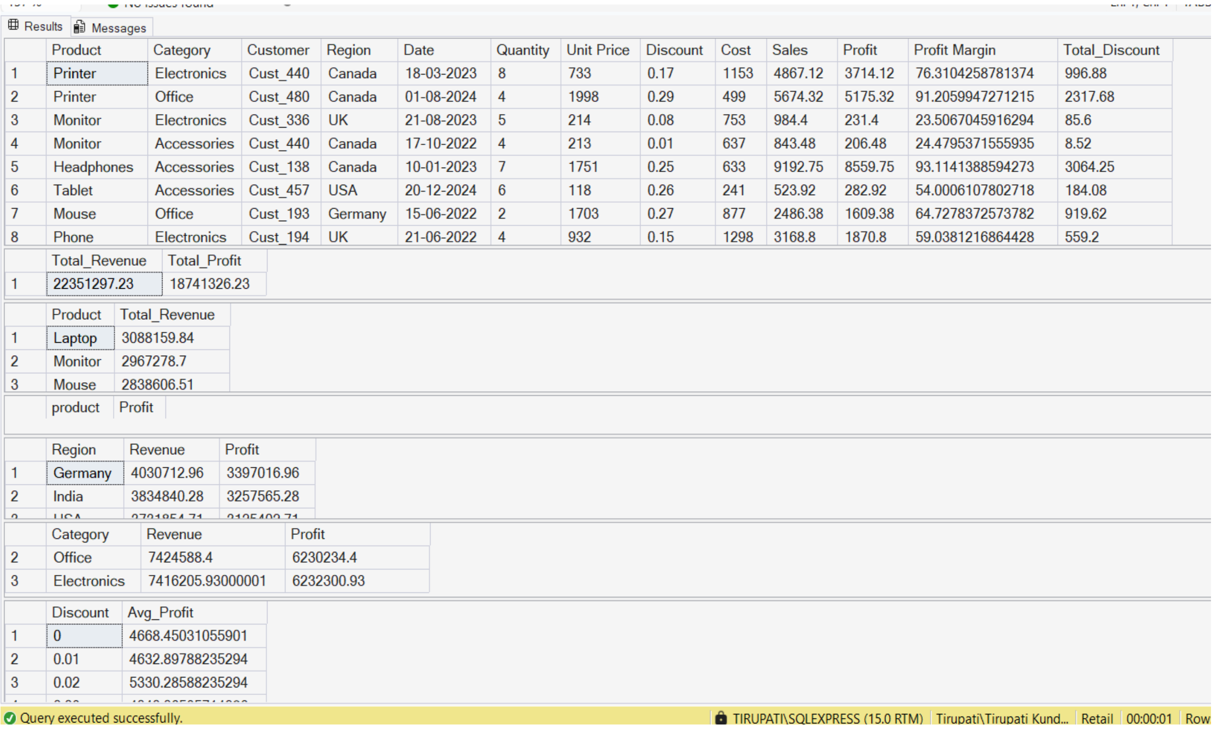
fillna() – Handle missing values such as discount and cost.

groupby() – Aggregate data for category, product, and region analysis.

value_counts() – Identify frequency distribution of categories.

EDA using SQL:

- Aggregation queries were used to calculate total sales, profit, and average discount.
- GROUP BY queries provided category, product, region, and customer performance insights.
- Date-based queries enabled monthly trend analysis.
- Financial comparison queries analyzed sales vs cost vs profit.
- Discount impact queries evaluated profitability effects.



	Product	Category	Customer	Region	Date	Quantity	Unit Price	Discount	Cost	Sales	Profit	Profit Margin	Total_Discount
1	Printer	Electronics	Cust_440	Canada	18-03-2023	8	733	0.17	1153	4867.12	3714.12	76.3104258781374	996.88
2	Printer	Office	Cust_480	Canada	01-08-2024	4	1998	0.29	499	5674.32	5175.32	91.2059947271215	2317.68
3	Monitor	Electronics	Cust_336	UK	21-08-2023	5	214	0.08	753	984.4	231.4	23.5067045916294	85.6
4	Monitor	Accessories	Cust_440	Canada	17-10-2022	4	213	0.01	637	843.48	206.48	24.4795371555935	8.52
5	Headphones	Accessories	Cust_138	Canada	10-01-2023	7	1751	0.25	633	9192.75	8559.75	93.1141388594273	3064.25
6	Tablet	Accessories	Cust_457	USA	20-12-2024	6	118	0.26	241	523.92	282.92	54.0006107802718	184.08
7	Mouse	Office	Cust_193	Germany	15-06-2022	2	1703	0.27	877	2486.38	1609.38	64.7278372573782	919.62
8	Phone	Electronics	Cust_194	UK	21-06-2022	4	932	0.15	1298	3168.8	1870.8	59.0381216864428	559.2

	Total_Revenue	Total_Profit
1	22351297.23	18741326.23

	Product	Total_Revenue
1	Laptop	3088159.84
2	Monitor	2967278.7
3	Mouse	2838606.51

	product	Profit

	Region	Revenue	Profit
1	Germany	4030712.96	3397016.96
2	India	3834840.28	3257565.28

	Category	Revenue	Profit
2	Office	7424588.4	6230234.4
3	Electronics	7416205.93000001	6232300.93

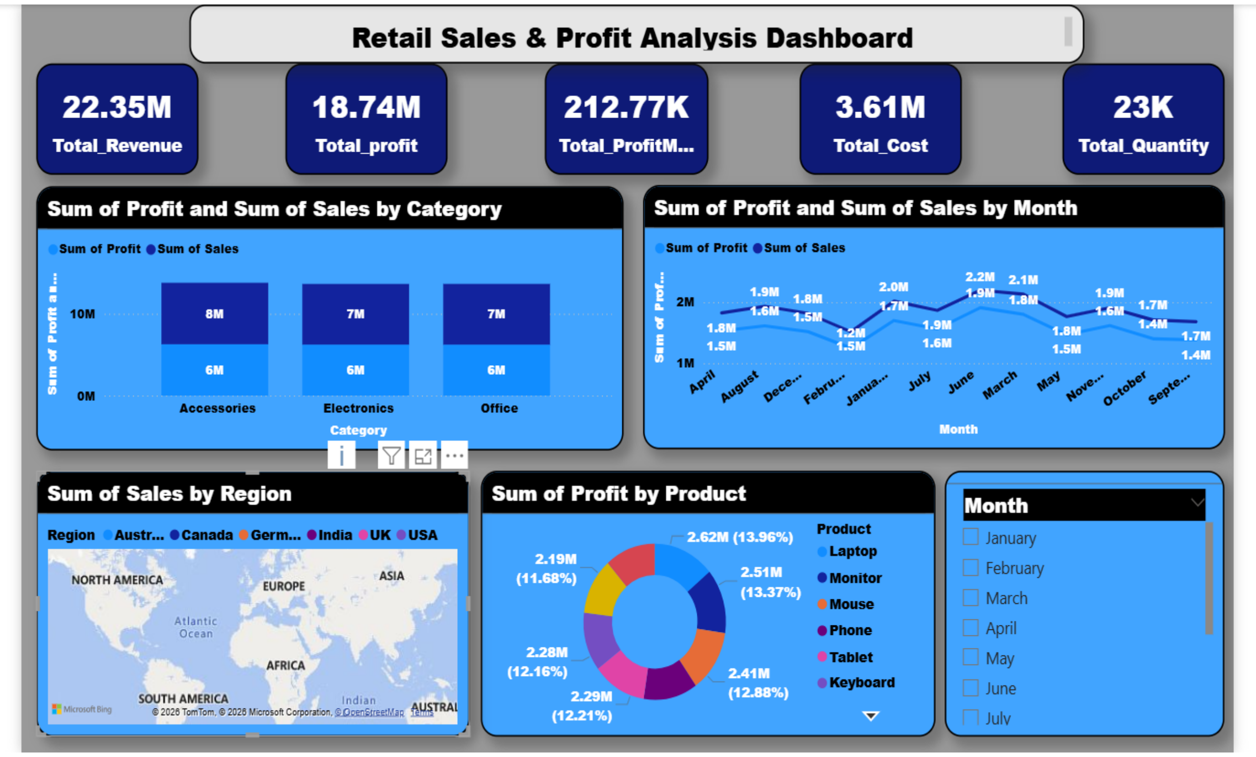
	Discount	Avg_Profit
1	0	4668.45031055901
2	0.01	4632.89788235294
3	0.02	5330.28588235294

Query executed successfully.

TIRUPATI\SQLEXPRESS (15.0 RTM) | Tirupati\Tirupati Kund... | Retail | 00:00:01 | Row

EDA using Power BI:

- KPI cards summarized total revenue, profit, cost, and quantity.
- Line charts identified sales and profit trends over time.
- Bar charts highlighted category and product performance.
- Scatter charts evaluated discount vs profit relationship.
- Regional visuals revealed geographic revenue distribution.
- Tables provided detailed customer and product insights.



Retail Sales & Profit Analysis Dashboard

22.35M Total_Revenue

18.74M Total_profit

212.77K Total_ProfitM...

3.61M Total_Cost

23K Total_Quantity

Month

□ January
□ February
□ March
□ April
□ May
□ June
□ July