

SUPERSTORE SALES

FINAL PROJECT



AGENDA

- Data Extraction
- Data Cleaning
- Data Transformation
- Data Loading
- Data Analysis
- Data Visualization



DATA EXTRACTION

- Superstore Sales
- Dataset from Kaggle

<https://www.kaggle.com/datasets/rohitsahoo/sales-forecasting>

DATA EXTRACTION

- Databricks Community Edition
- Load the data

```
# Load superstore sales data from a CSV file
superstore_sales_data_url = "/FileStore/tables/train-8.csv" # Ensure this is the correct path
df = spark.read.format("csv") \
    .option("header", "true") \
    .option("inferSchema", "true") \
    .load(superstore_sales_data_url)
```

- Using PySpark

DATA

```
(13) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [Row_ID: integer, Order_ID: string ..., 16 more fields]

root
|-- Row_ID: integer (nullable = true)
|-- Order_ID: string (nullable = true)
|-- Order_Date: date (nullable = true)
|-- Ship_Date: date (nullable = true)
|-- Ship_Mode: string (nullable = true)
|-- Customer_ID: string (nullable = true)
|-- Customer_Name: string (nullable = true)
|-- Segment: string (nullable = true)
|-- Country: string (nullable = true)
|-- City: string (nullable = true)
|-- State: string (nullable = true)
|-- Postal_Code: integer (nullable = true)
|-- Region: string (nullable = true)
|-- Product_ID: string (nullable = true)
|-- Category: string (nullable = true)
|-- Sub-Category: string (nullable = true)
|-- Product_Name: string (nullable = true)
|-- Sales: string (nullable = true)
```

SUPERSTORE SALE DATA

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----												
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----												
	1	CA-2017-152156	2017-11-08	2017-11-11	Second Class	CG-12520	Claire Gute	Consumer	United States		Hen	
derson	Kentucky	42420	South	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Col...	261.96				
	2	CA-2017-152156	2017-11-08	2017-11-11	Second Class	CG-12520	Claire Gute	Consumer	United States		Hen	
derson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabric...	731.94				
	3	CA-2017-138688	2017-06-12	2017-06-16	Second Class	DV-13045	Darrin Van Huff	Corporate	United States		Los A	
ngeles	California	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive Add...	14.62				
	4	US-2016-108966	2016-10-11	2016-10-18	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Laud		
erdale	Florida	33311	South	FUR-TA-10000577	Furniture	Tables	Bretford CR4500 S...	957.5775				
	5	US-2016-108966	2016-10-11	2016-10-18	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Laud		
erdale	Florida	33311	South	OFF-ST-10000760	Office Supplies	Storage	Eldon Fold 'N Rol...	22.368				
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----												
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----												

only showing top 5 rows

DATA CLEANING

- *Column Names :*

Replace spaces with underscores and remove special characters

- Remove Duplicates

- Null values

DATA TRANSFORMATION

- Transform the data
- Add new columns

```
df = df.withColumn('Order_Month', month(col('Order_Date')))
```

```
df = df.withColumn('Order_Year', year(col('Order_Date')))
```

- Save the data into Tables

```
df.write.format("delta").mode("overwrite").save("/FileStore/delta/superstore_transformed")
```


DATA TRANSFORMATION



► (12) Spark Jobs

```
► df: pyspark.sql.dataframe.DataFrame = [Row_ID: integer, Order_ID: integer, Order_Date: date, Ship_Date: date, Ship_Mode: string, Customer_ID: string, Customer_Name: string, Segment: string, Country: string, City: string, State: string, Postal_Code: integer, Region: string, Product_ID: string, Category: string, Sub-Category: string, Product_Name: string, Sales: float, Order_Month: integer, Order_Year: integer]
```

root

- |-- Row_ID: integer (nullable = true)
- |-- Order_ID: integer (nullable = true)
- |-- Order_Date: date (nullable = true)
- |-- Ship_Date: date (nullable = true)
- |-- Ship_Mode: string (nullable = true)
- |-- Customer_ID: string (nullable = true)
- |-- Customer_Name: string (nullable = true)
- |-- Segment: string (nullable = true)
- |-- Country: string (nullable = true)
- |-- City: string (nullable = true)
- |-- State: string (nullable = true)
- |-- Postal_Code: integer (nullable = true)
- |-- Region: string (nullable = true)
- |-- Product_ID: string (nullable = true)
- |-- Category: string (nullable = true)
- |-- Sub-Category: string (nullable = true)
- |-- Product_Name: string (nullable = true)
- |-- Sales: float (nullable = true)
- |-- Order_Month: integer (nullable = true)
- |-- Order_Year: integer (nullable = true)

DATA ANALYSIS

- Total sales By ProductName

```
top_products: pyspark.sql.dataframe.DataFrame = [Product_Name: string, Total_Sales: double]
```

Product_Name	Total_Sales
Canon imageCLASS ...	61599.822265625
Fellowes PB500 El...	27453.384033203125
Cisco TelePresenc...	22638.48046875
HON 5400 Series T...	21870.57550048828
GBC DocuBind TL30...	19823.478515625
GBC Ibimaster 500...	19024.500244140625
Hewlett Packard L...	18839.685913085938
"HP Designjet T52...	18374.895263671875
GBC DocuBind P400...	17965.06787109375
High Speed Automa...	17030.311767578125

DATA ANALYSIS

- Total sales by Region

```
regional_performance: pyspark.sql.dataframe.DataFrame = [Region: string, Total_Sales: double]
```

```
+-----+-----+
| Region| Total_Sales|
+-----+-----+
|  South|386413.13934862614|
|Central|489321.39007872343|
|   East| 663043.8557248116|
|   West| 698354.7733091116|
+-----+-----+
```

DATA ANALYSIS

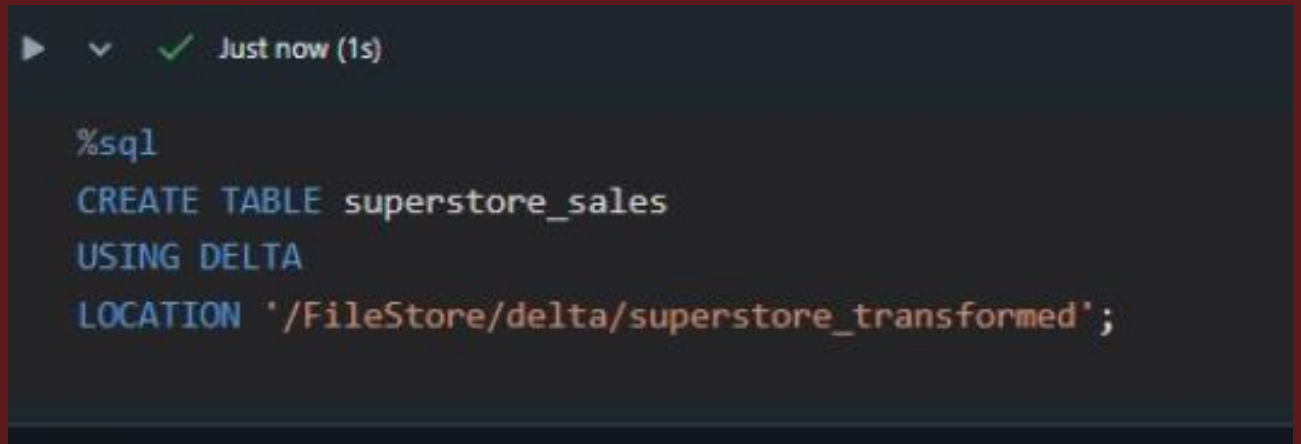
- Total sales by YEAR

2015	1	14130.160984992981
2015	2	4119.8159548044205
2015	3	55040.987458229065
2015	4	27751.07089471817
2015	5	23630.68287038803
2015	6	34298.34748792648
2015	7	33336.022790551186
2015	8	26811.580191135406
2015	9	81342.98330289125
2015	10	31394.94074845314
2015	11	77622.53452861309
2015	12	68001.27428495884
2016	1	17977.997522592545
2016	2	11924.271917104721
2016	3	32234.358570575714
2016	4	32599.74250805378
2016	5	29209.414824724197
2016	6	23461.765960991383
2016	7	28377.822920084
2016	8	36300.93817472458

+-----+-----+-----+

DATA LOADING

- CREATE SQL Table
- Load the data into TABLES

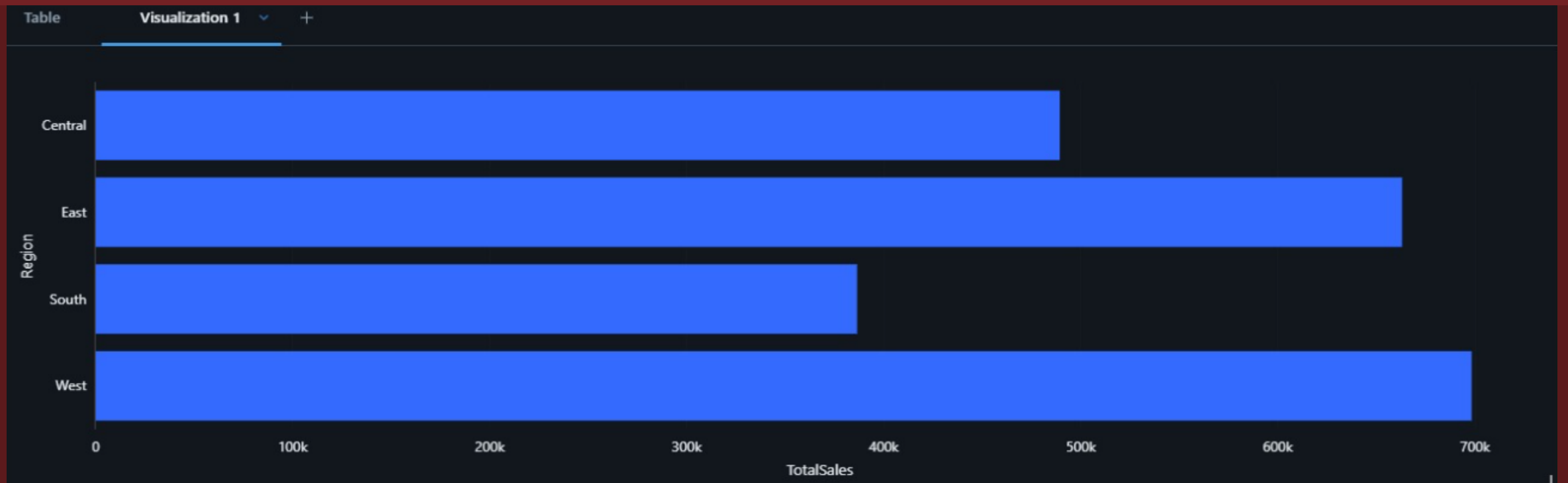


A screenshot of a SQL execution interface. At the top, there is a status bar with a play button, a dropdown arrow, a green checkmark, and the text "Just now (1s)". Below this, the SQL query is displayed in a monospaced font with syntax highlighting: "%sql" in light blue, "CREATE TABLE" in blue, "superstore_sales" in white, "USING DELTA" in blue, and "LOCATION '/FileStore/delta/superstore_transformed';" in white. The background is a dark blue-grey.

```
%sql
CREATE TABLE superstore_sales
USING DELTA
LOCATION '/FileStore/delta/superstore_transformed';
```

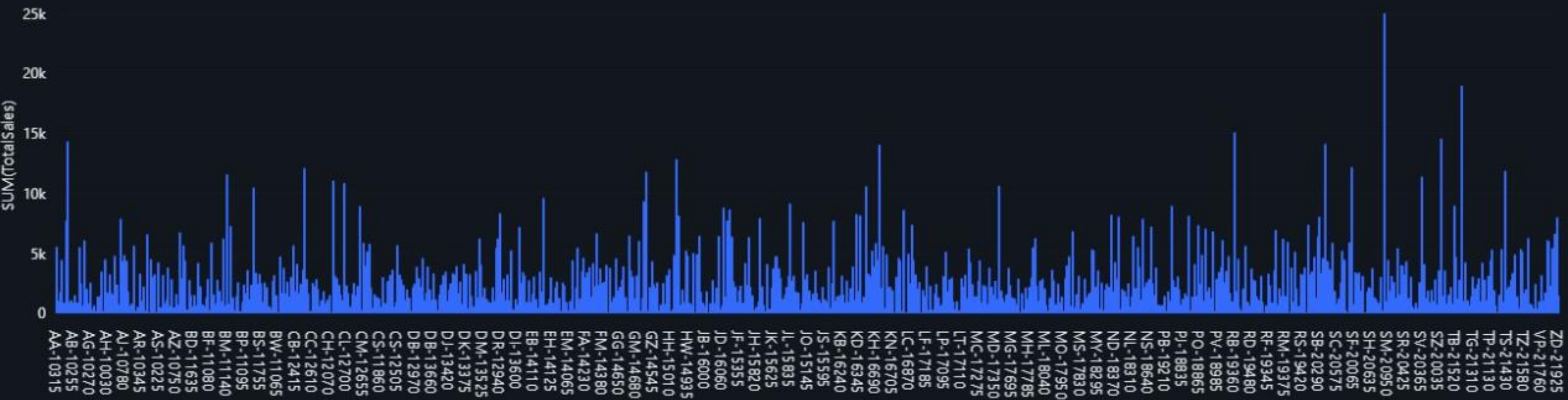
DATA VISUALIZATION

- Total sales by Region



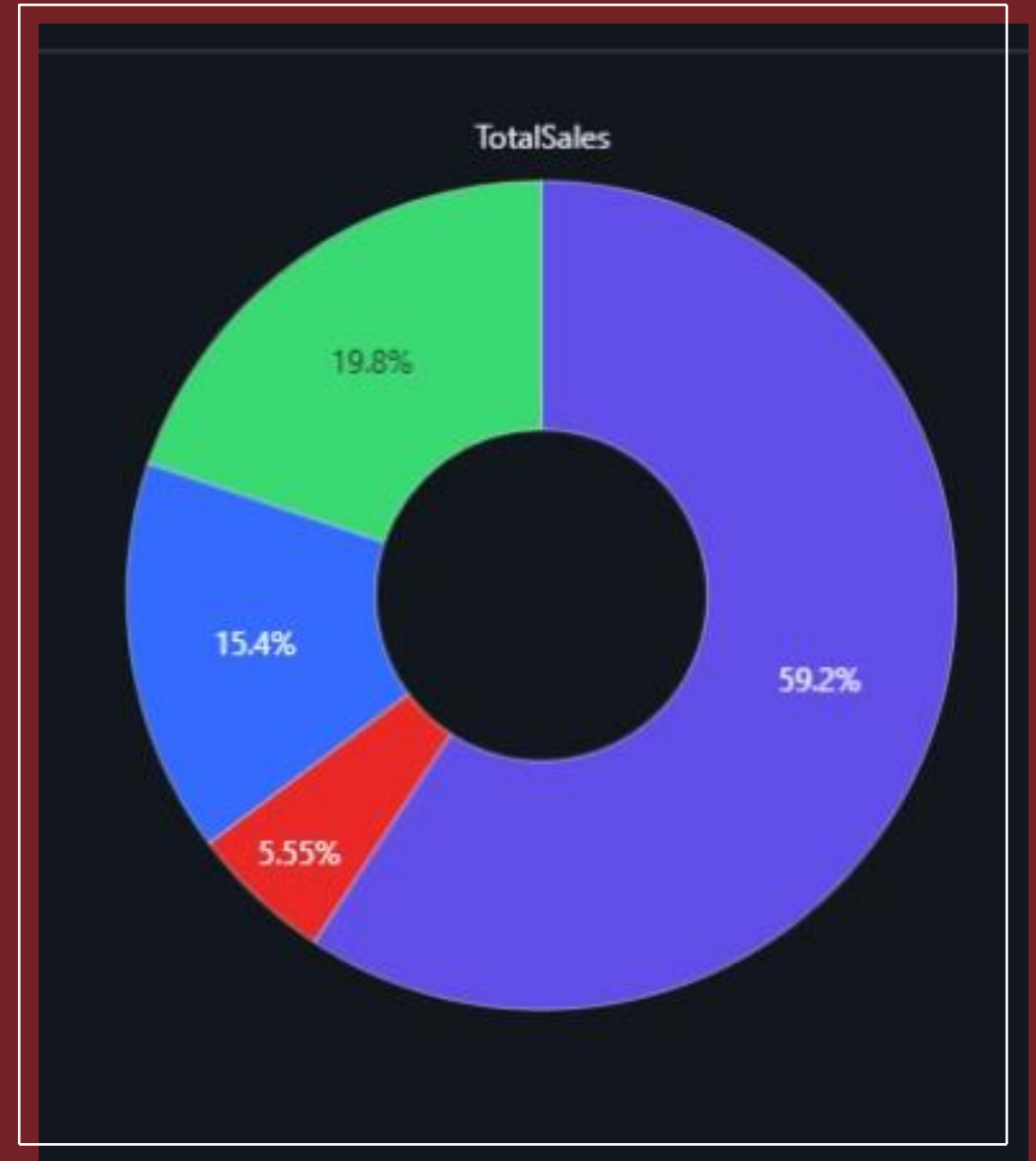
DATA VISUALIZATION

- Total sales by customer



DATA VISUALIZATION

- Total Sales by shipmode



DATA VISUALIZATION

- Total sales by productname



THANK YOU

