Data Analysis with PySpark: Exploring Trends and Insights in Retail Sales

Section Number: 44517-03

Team Name: DataXplorers

Team Members:
1. Venkatalahari Ravipati
2. Chiranjeevi Ankaiah
3. Vikas Allam

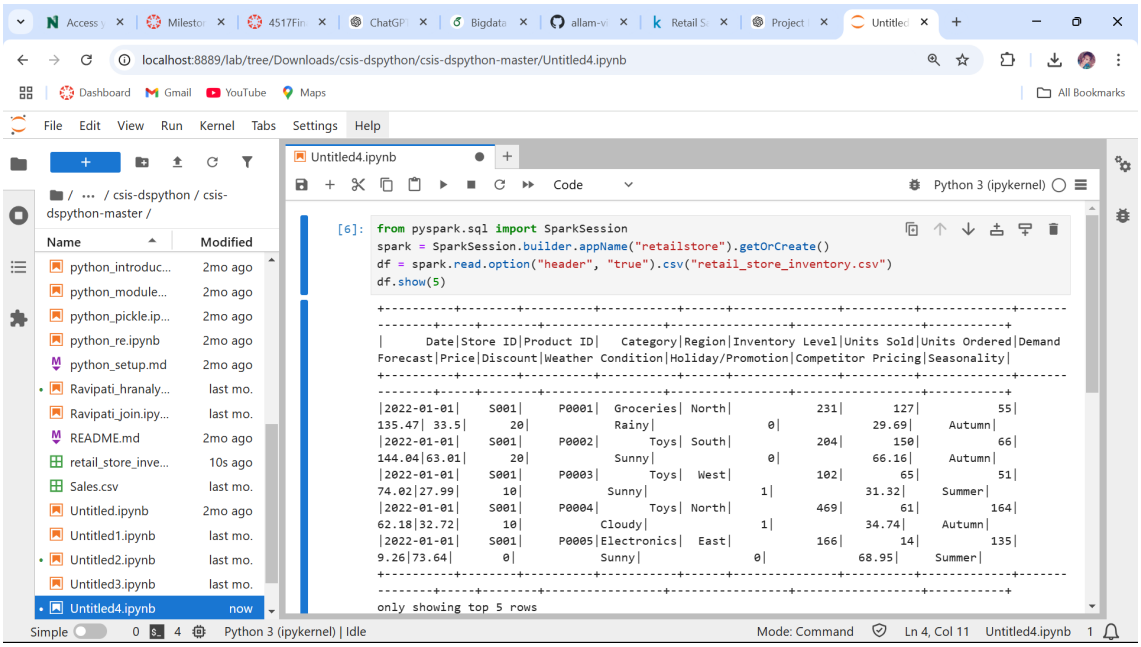
Abstract

This project involves analyzing sales data using PySpark. The primary objectives are to clean and preprocess the data, gain insights into sales performance by region and product, analyze sales trends over time, visualize key findings, and evaluate profitability for strategic decision-making.

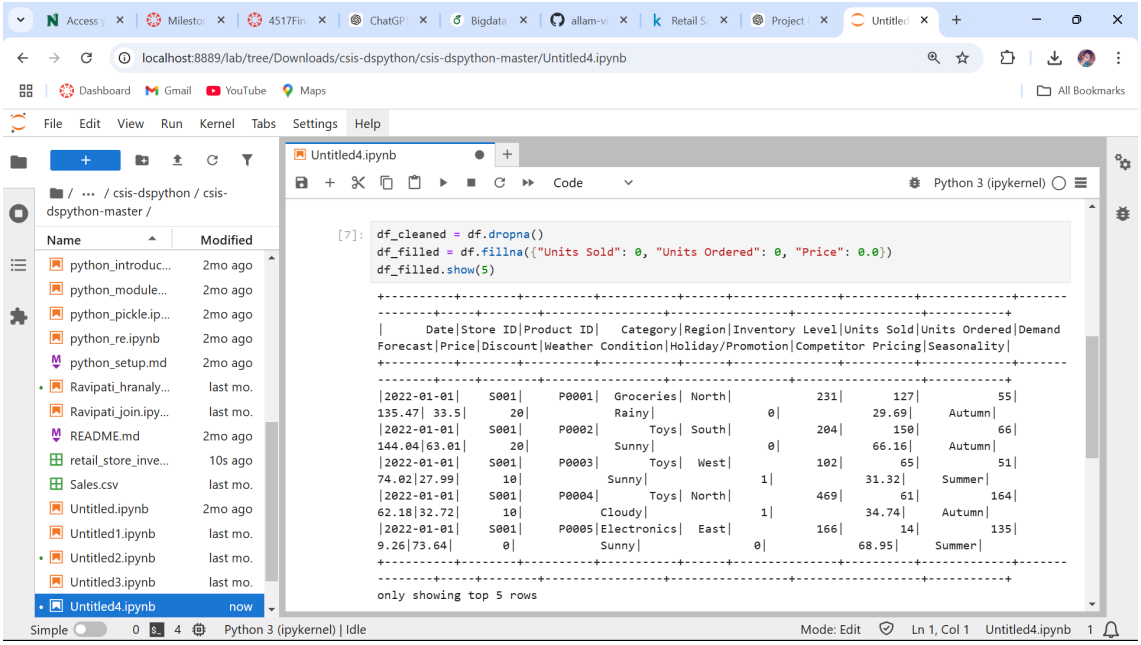
1. Implementation Steps

1.1. Goal 1: Data Cleaning and Preparation

1. **Load Data:** The dataset is loaded into a PySpark DataFrame using the `spark.read.csv` function.



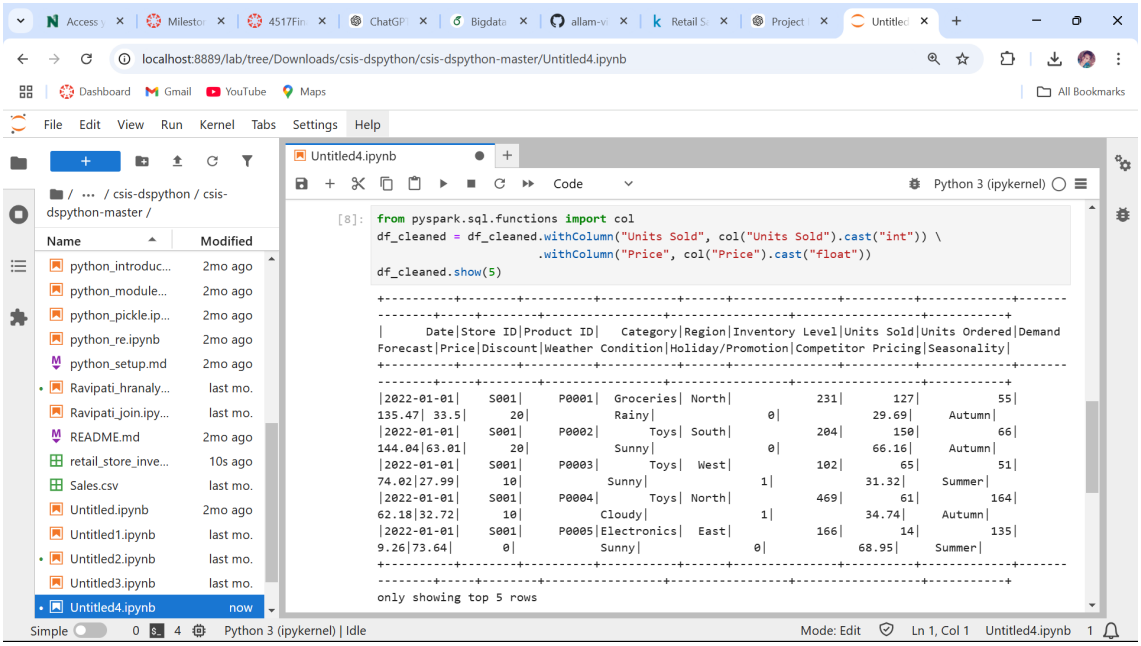
2. **Handle Missing Values:** Rows with missing values are either dropped or filled with default values.



```python
df_cleaned = df.dropna()
df_filled = df.fillna({"Units Sold": 0, "Units Ordered": 0, "Price": 0.0})
df_filled.show(5)
```

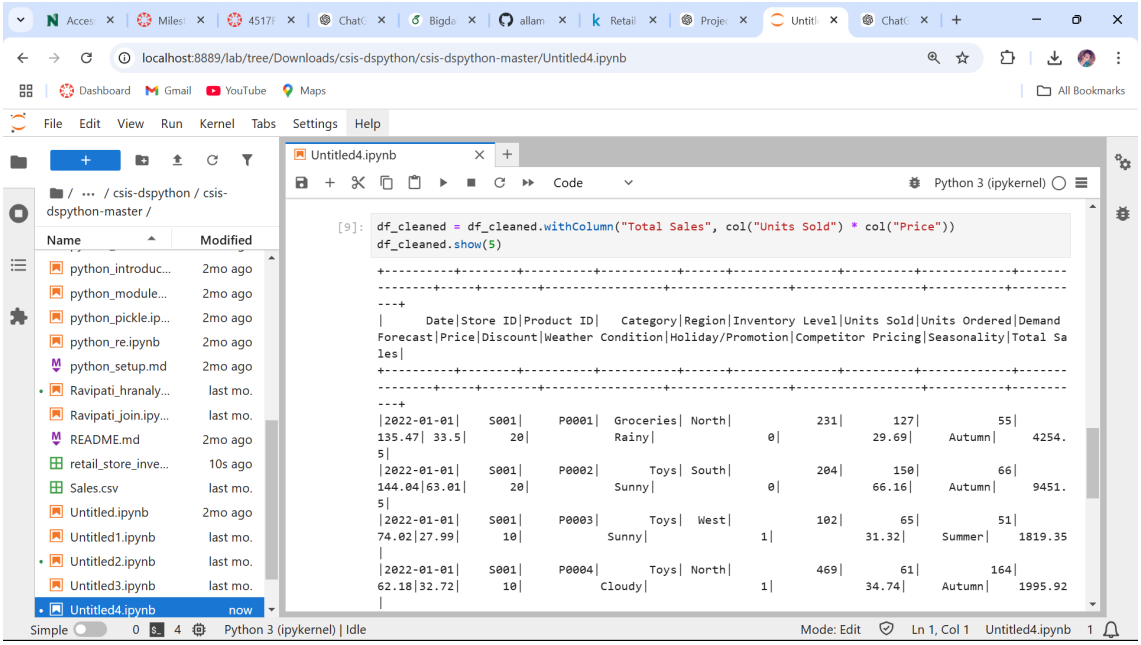
```
+----------+--------+----------+-----------+------+---------------+----------+-------------+-------
--------+-----+--------+-----------------+-----------------+------------------+-----------+
|      Date|Store ID|Product ID|   Category|Region|Inventory Level|Units Sold|Units Ordered|Demand
Forecast|Price|Discount|Weather Condition|Holiday/Promotion|Competitor Pricing|Seasonality|
+----------+--------+----------+-----------+------+---------------+----------+-------------+-------
--------+-----+--------+-----------------+-----------------+------------------+-----------+
|2022-01-01|    S001|     P0001|  Groceries| North|            231|       127|           55|
135.47| 33.5|      20|            Rainy|                0|             29.69|     Autumn|
|2022-01-01|    S001|     P0002|       Toys| South|            204|       150|           66|
144.04|63.01|      20|            Sunny|                0|             66.16|     Autumn|
|2022-01-01|    S001|     P0003|       Toys|  West|            102|        65|           51|
74.02|27.99|      10|            Sunny|                1|             31.32|     Summer|
|2022-01-01|    S001|     P0004|       Toys| North|            469|        61|          164|
62.18|32.72|      10|           Cloudy|                1|             34.74|     Autumn|
|2022-01-01|    S001|     P0005|Electronics|  East|            166|        14|          135|
9.26|73.64|       0|            Sunny|                0|             68.95|     Summer|
+----------+--------+----------+-----------+------+---------------+----------+-------------+-------
--------+-----+--------+-----------------+-----------------+------------------+-----------+
only showing top 5 rows
```

3. **Standardize Column Types:** Numeric columns are converted to the appropriate data types.

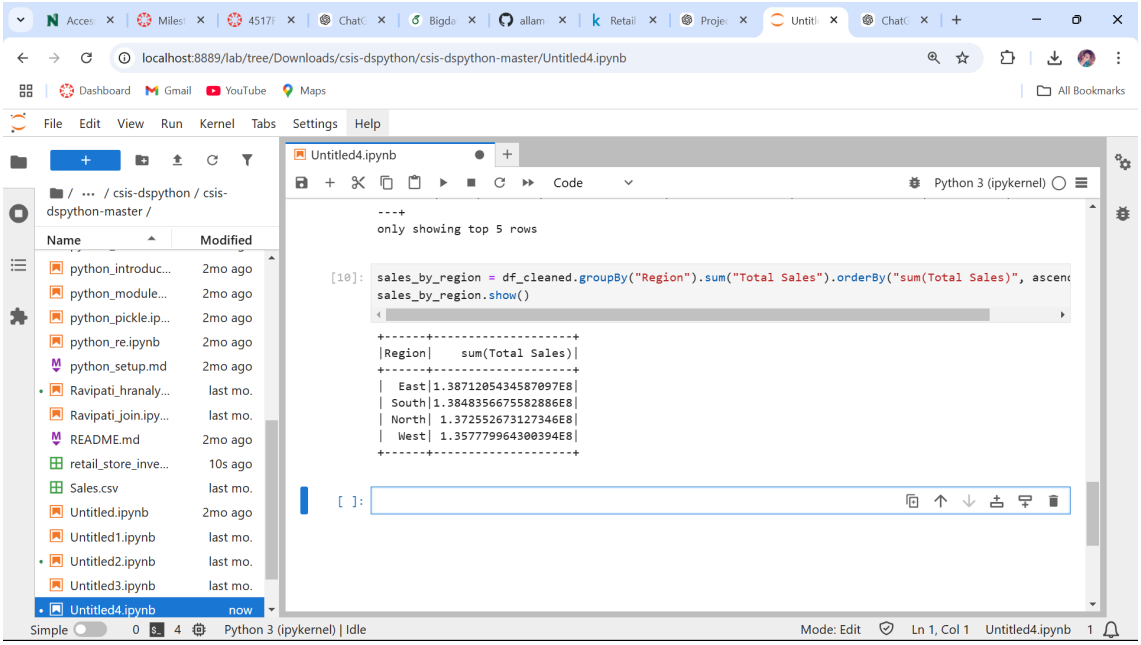


1.2. Goal 2: Sales Insights by Region

1. Calculate total sales by multiplying `Units Sold` with `Price`.

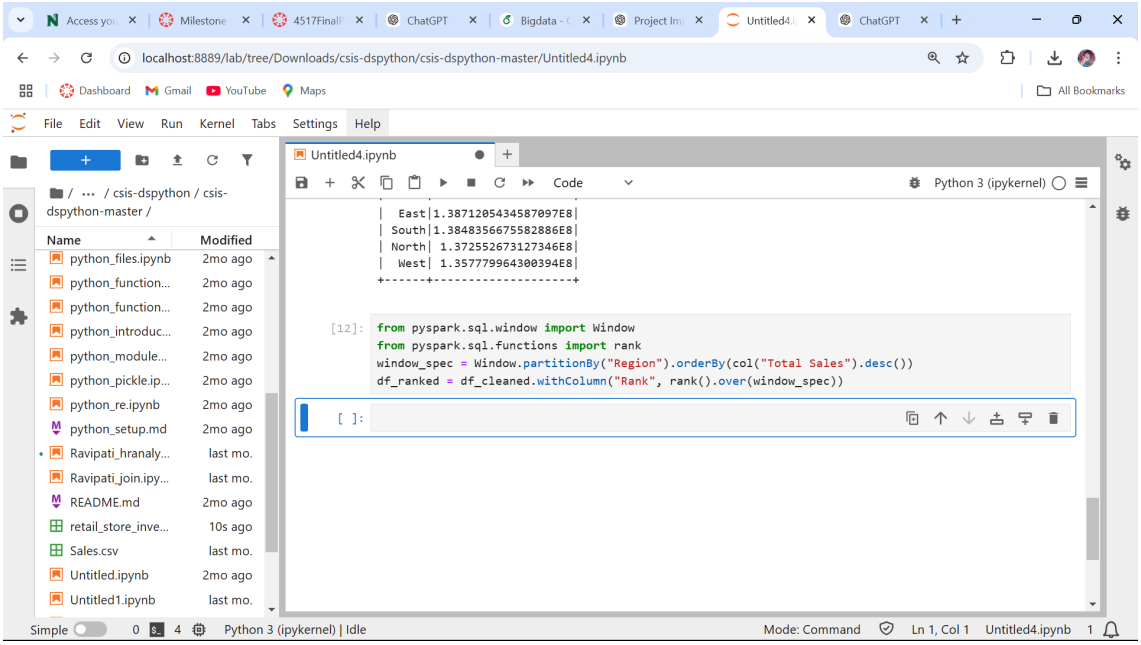


2. Group by `Region` and calculate total sales.

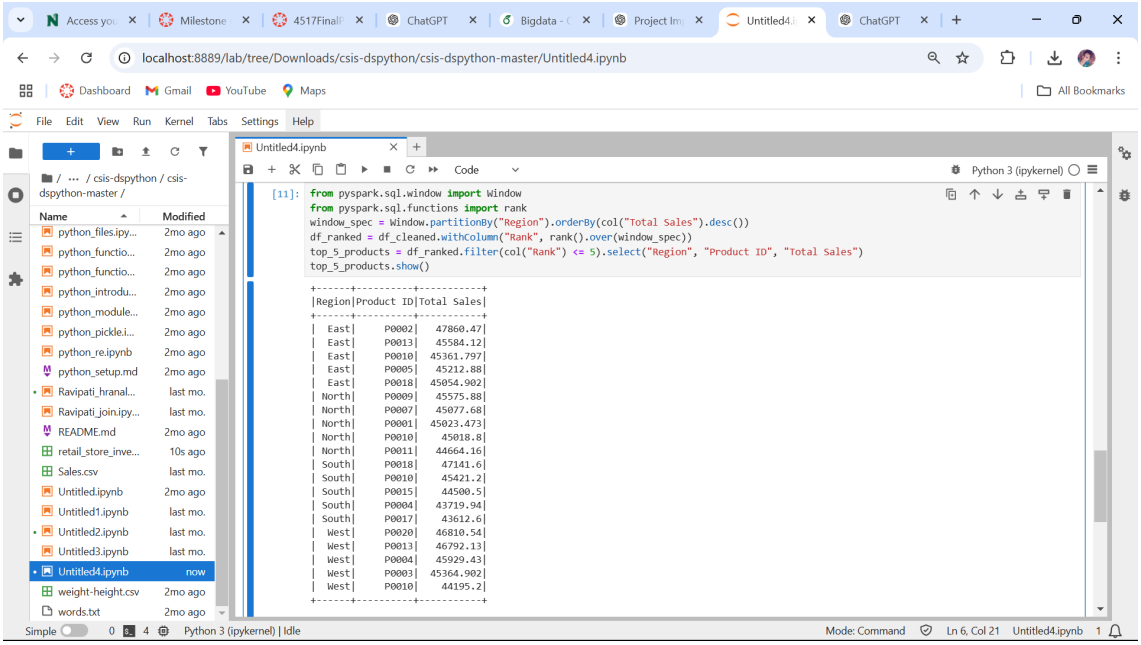


1.3. Goal 3: Top Products by Region

1. Use a window function to rank products by sales within each region.

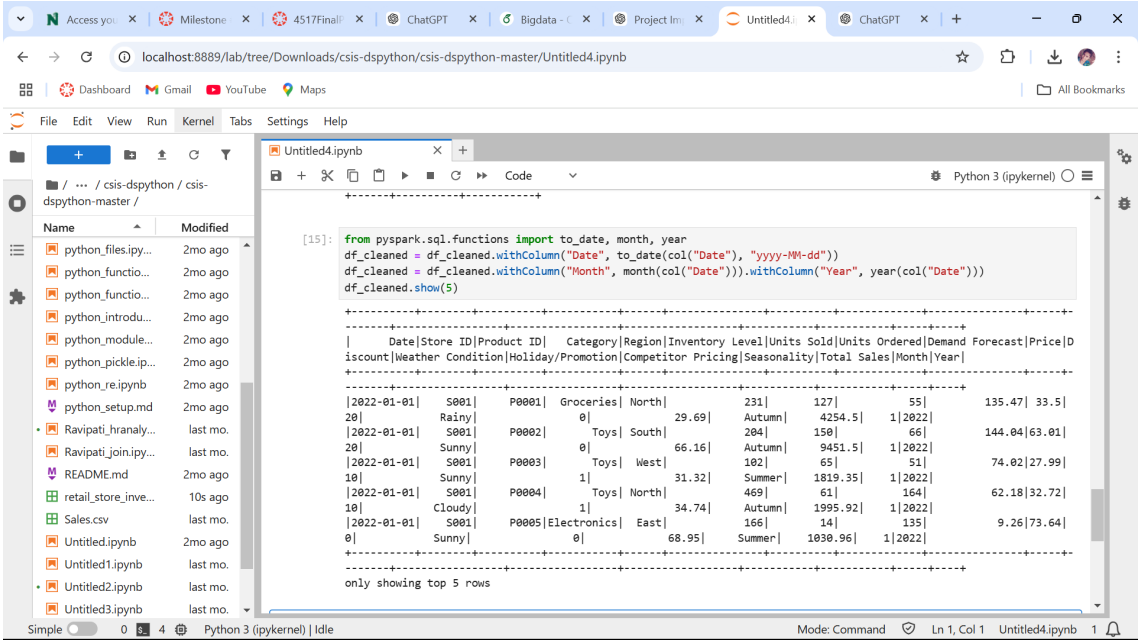


2. Filter the top 5 products for each region.

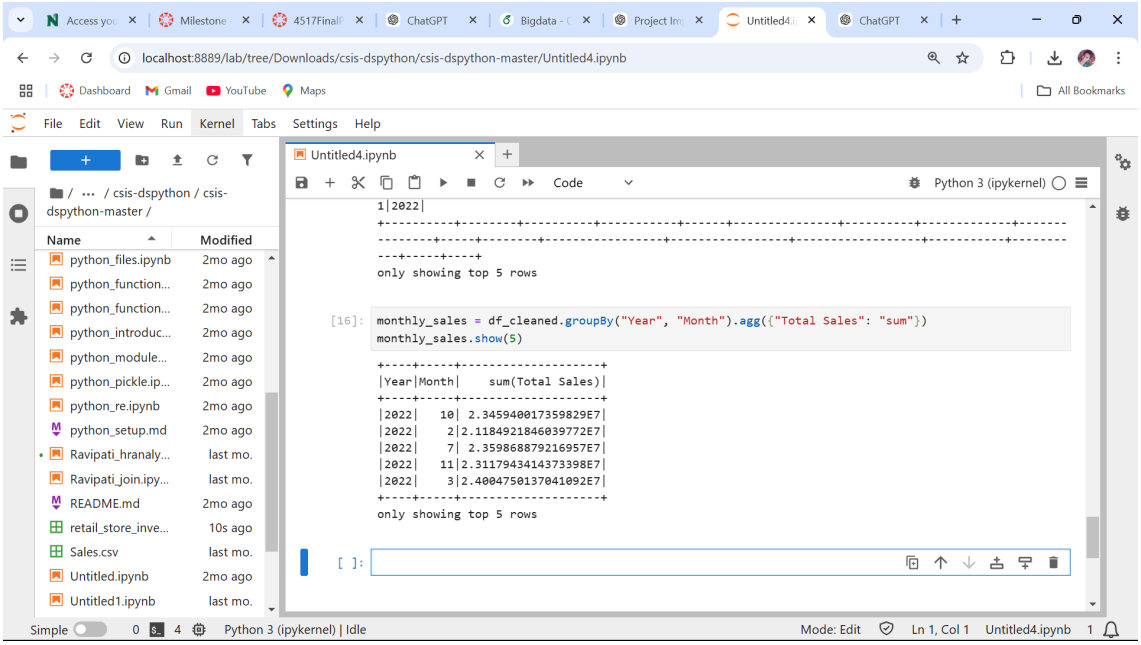


1.4. Goal 4: Sales Trend Analysis

1. Extract `Month` and `Year` from the `Date` column.



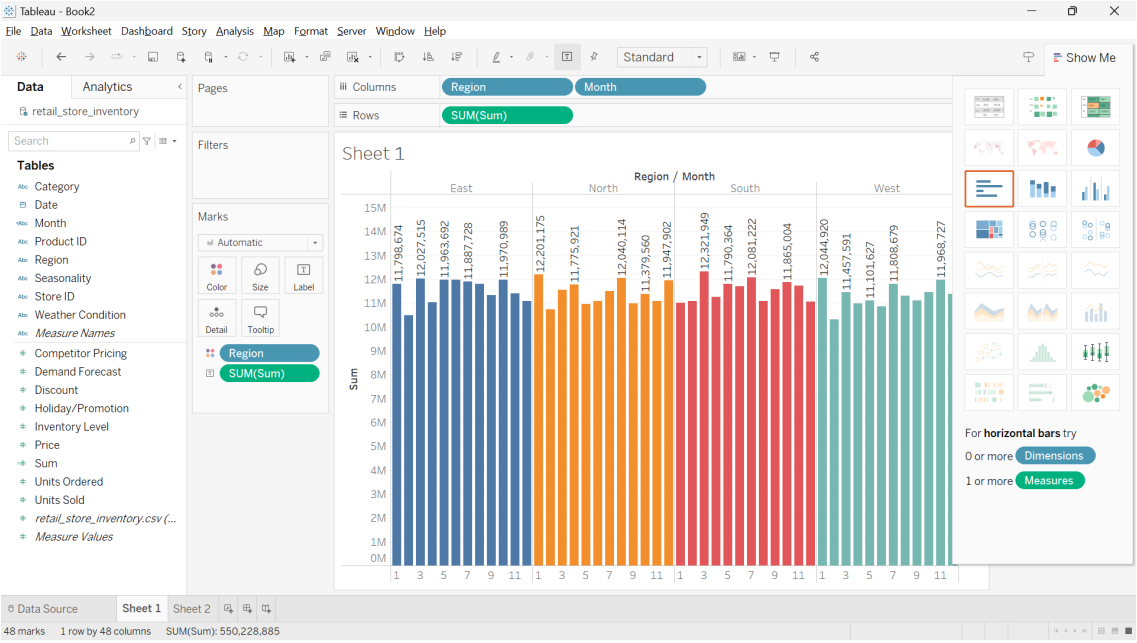
2. Group data by `Month` and `Year` and calculate total sales.



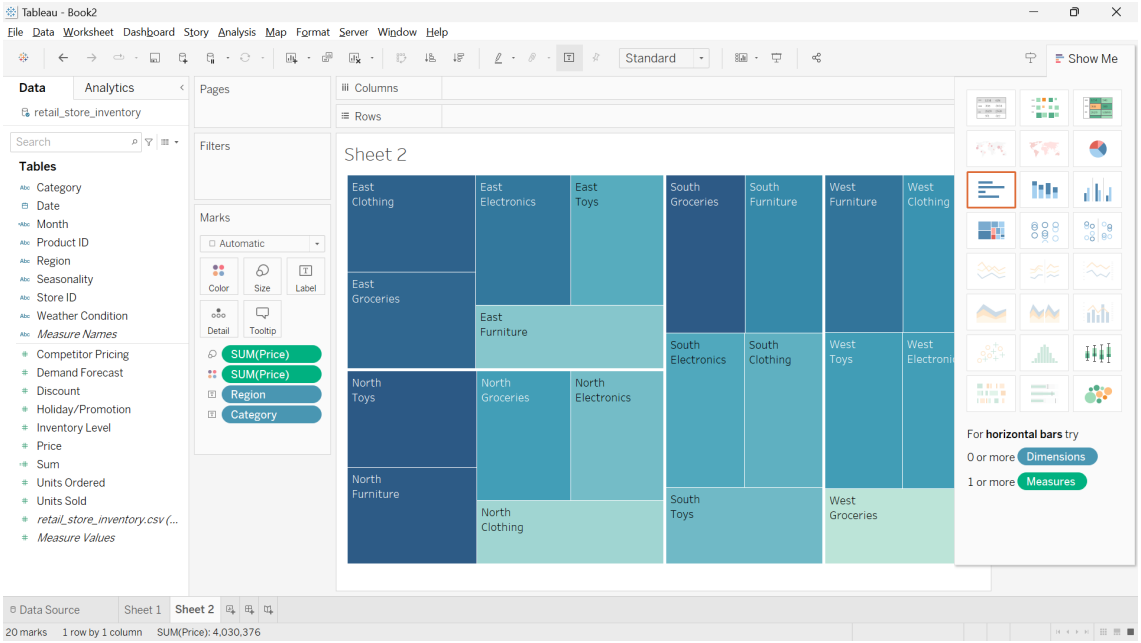
1.5. Goal 5: Visualization of Insights

1. Convert the PySpark DataFrame to Pandas and use Matplotlib for visualizations.

2. Calculating total sales of each region in every month.

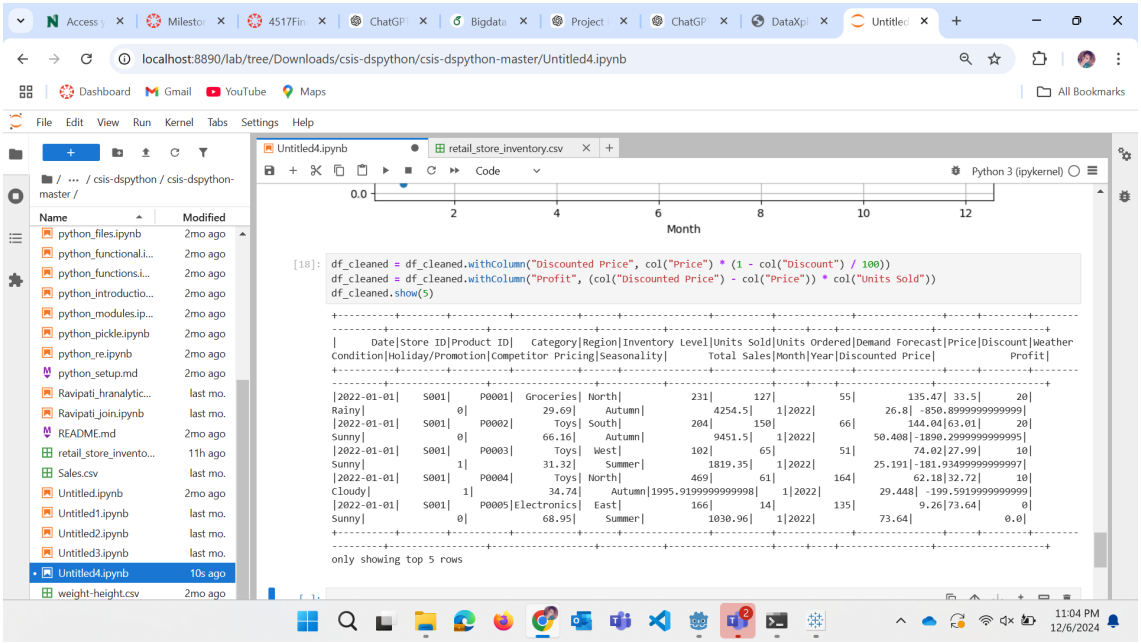


3. Calculating product price in each region.

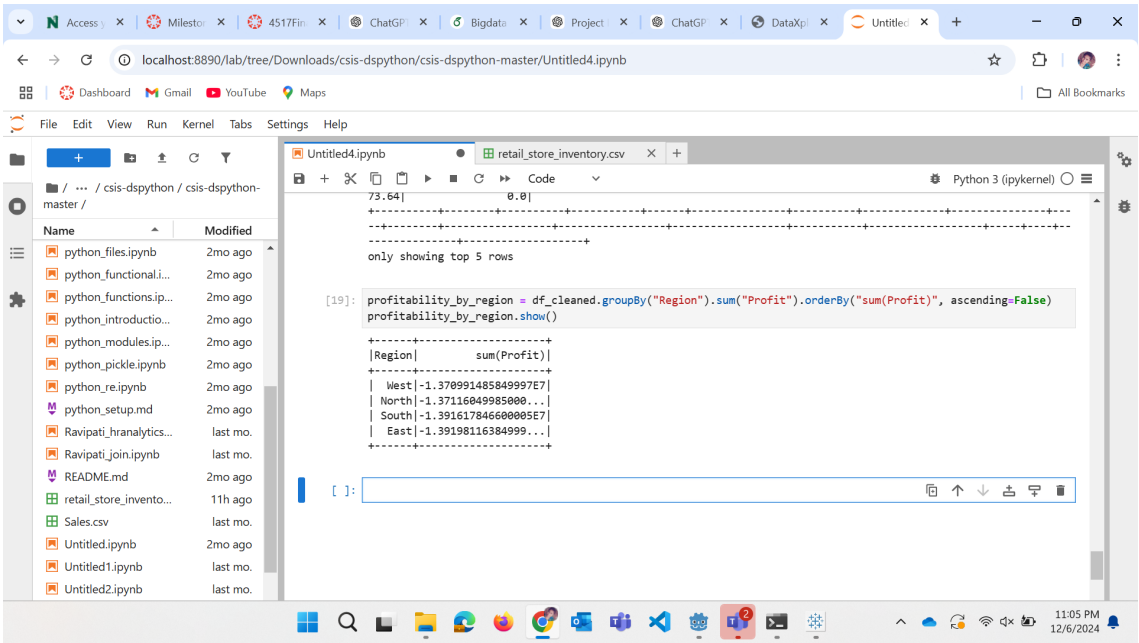


1.6. Goal 6: Profitability Analysis

1. Calculate `Profit` based on `Discounted Price`.



2. Group by `Region` to calculate total profit.



2. Detailed Discussion of Results Achieved for each goal

2.1. Goal 1: Data Cleaning and Preparation

The data was successfully cleaned and standardized, ensuring no missing values or inconsistent column types. This improved the dataset's quality and usability.

2.2. Goal 2: Sales Insights by Region

Regions were ranked based on total sales, identifying the highest and lowest performers. This insight helps target regions for marketing or operational improvements.

2.3. Goal 3: Top Products by Region

The top 5 products for each region were identified, highlighting the most successful products by sales volume in different areas.

2.4. Goal 4: Sales Trend Analysis

Monthly sales trends revealed seasonal patterns. For instance, sales increased during specific months, likely due to holidays or promotions.

2.5. Goal 5: Visualization of Insights

The visualizations provided a clear representation of sales trends and regional performance, making it easier for stakeholders to interpret the results.

2.6. Goal 6: Profitability Analysis

Profitability analysis highlighted regions and products with the highest and lowest profit contributions. These insights can guide pricing and marketing strategies.

3. Conclusion

This project demonstrates the power of PySpark for analyzing large datasets. By cleaning the data, generating insights, and visualizing key findings, we provided actionable insights into sales performance and profitability. These results can help organizations optimize their operations and strategies.

4. Citations

1. Dataset Link: `https://www.kaggle.com/datasets/anirudhchauhan/retail-store-inventory-forecasting-dataset`

5. GitHub Repository

The complete project, including source code and visualizations, is available at: `https://github.com/your-repo-url`