CSC1109 - Supermarket Sales dataset analysis using Apache Pig and Hive

JiaXiang Chen - 21363061

November 2024

Github Repo here

1 Introduction

In a world full of large data, analysing large-scale data demands robust distributed computing solutions. Apache Hadoop addresses this with a framework that enables parallel processing across computer clusters. Within Hadoop, Apache Pig and Hive are essential tools for data analysis: Pig provides a high-level platform for creating MapReduce programs for data processing, while Hive offers a SQL-like interface for querying large datasets.

The supermarket sales dataset that I have chosen includes transaction records across branches, detailing sales, products, customer types, and performance metrics. The approach is as follows:

1. **Data Cleaning with Pig:** Initial loading and cleaning of data, removing invalid entries and standardising formats.
2. **Basic Analysis with Pig and Hive:** Running identical queries in both tools to compare execution and results, focusing on basic sales metrics and customer patterns.
3. **Complex Analysis with Hive:** Advanced queries using Hive-specific functions, JOIN operations, and sampling techniques for insights into customer payment patterns.

2 Cleaning the data

I performed several steps to prepare the data using Apache Pig, focusing on ensuring data quality for accurate analysis. I began by loading the CSV file using PigStorage, defining appropriate data types for each field including transaction details, customer information, and financial metrics. To maintain data quality, I filtered out records with null InvoiceID values and excluded transactions with invalid quantities or total values.

```
-- Load the dataset
sales_data = LOAD '../Data/Raw_Data/supermarket_sales - Sheet1.csv'
USING PigStorage(',')
AS (InvoiceID:chararray, Branch:chararray, City:chararray, CustomerType:chararray, Gender:chararray, ProductLine:chararray,
UnitPrice:float, Quantity:int, Tax:float, Total:float, Date:chararray, Time:chararray, Payment:chararray, COGS:float,
GrossMarginPercentage:float, GrossIncome:float, Rating:float);

-- Filter the data to remove rows with null InvoiceID or invalid Quantity and Total values
cleaned_sales_data = FILTER sales_data BY InvoiceID IS NOT NULL AND Quantity > 0 AND Total > 0;

-- Process and split the date into Month, Day, and Year fields
split_dates = FOREACH cleaned_sales_data GENERATE
    InvoiceID,
    Branch,
    City,
    CustomerType,
    Gender,
    ProductLine,
    ROUND(UnitPrice * 100.0) / 100.0 AS UnitPrice:float,
    Quantity,
    ROUND(Tax * 100.0) / 100.0 AS Tax:float,
    ROUND(Total * 100.0) / 100.0 AS Total:float,
    Date,
    Time,
    Payment,
    ROUND(COGS * 100.0) / 100.0 AS COGS:float,
    GrossMarginPercentage,
    ROUND(GrossIncome * 100.0) / 100.0 AS GrossIncome:float,
    ROUND(Rating * 100.0) / 100.0 AS Rating:float,
    (chararray)STRSPLIT(Date, '/', 3).$0 AS Month,
    (chararray)STRSPLIT(Date, '/', 3).$1 AS Day,
    (chararray)STRSPLIT(Date, '/', 3).$2 AS Year;

-- Store the cleaned and transformed data
STORE split_dates INTO '../Data/Cleaned_Data/cleaned_sales_data' USING PigStorage(',');
```

The processing focuses on preparing the dataset for analysis. This involved rounding money values for consistency, splitting the date field into month, day, and year components for time-based analysis, and standardising customer ratings. The transformed data was then stored in a new directory named “Cleaned_Data” as a cleaned CSV file which is then ready for querying.

3 Querying the Data

3.1 Average Gross Income by City

Objective: To calculate and compare the average gross income generated from sales across different city branches.

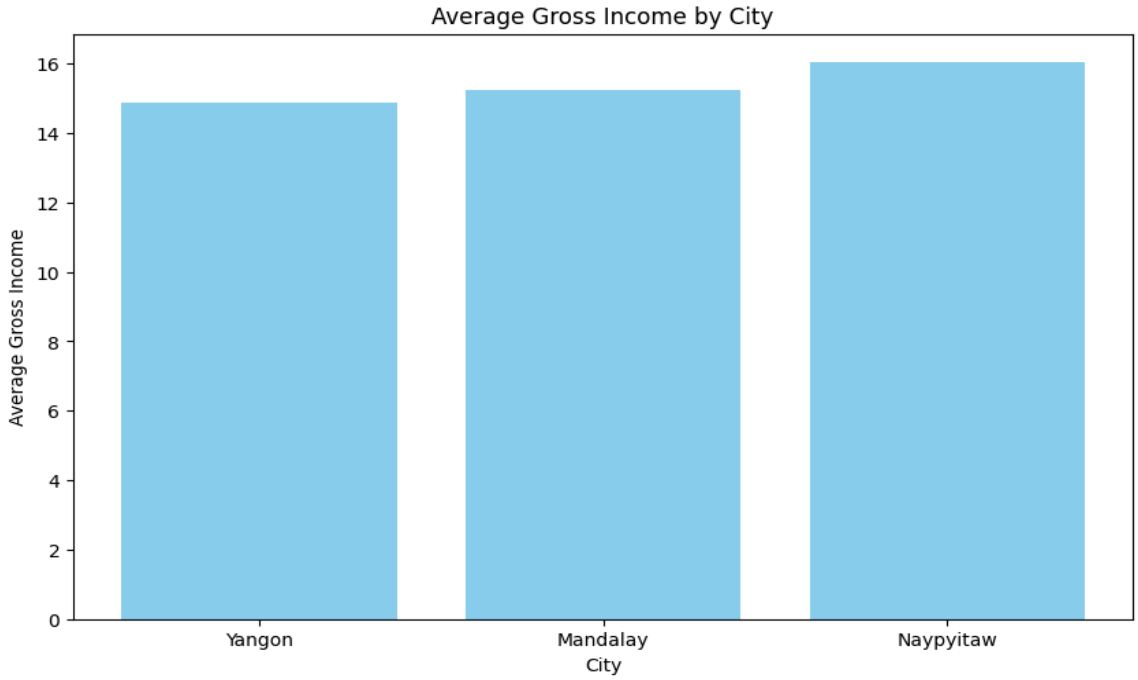
```
Yangon,14.87
Mandalay,15.23
Naypyitaw,16.05
```

(a) Pig query output

```
Mandalay,15.23
Naypyitaw,16.05
Yangon,14.87
```

(b) Hive query output

From the output we can see that the results have matched up meaning that there were no problems with the calculations. We can see that there was a varied level of gross incomes across the three city branches. Naypyitaw leads with the highest average gross income of $16.05 per sale, followed by Mandalay at $15.23, while Yangon shows the lowest at $14.87.



3.2 Total Revenue by Payment Method

Objective: To analyse the distribution of total revenue across different payment methods used by customers.

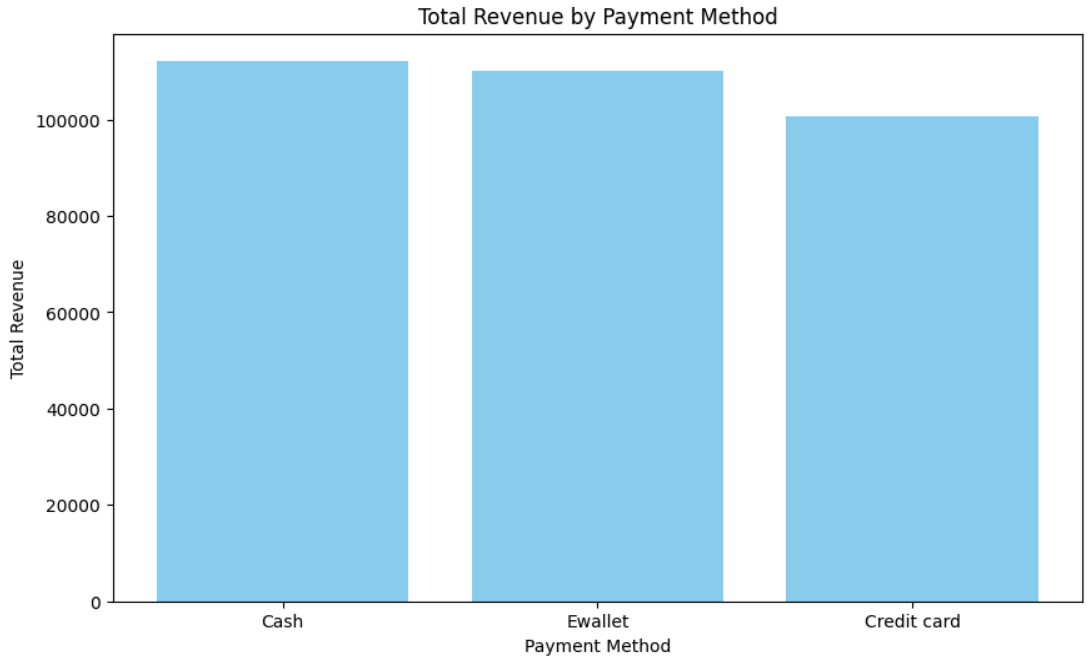
```
Cash,112206.52
Ewallet,109993.16
Credit card,100767.12
```

(a) Pig query output

```
Cash,112206.52
Credit card,100767.12
Ewallet,109993.16
```

(b) Hive query output

We can see that there's also a varied revenue generation across three payment methods. Cash transactions lead with total revenue of $112,206.52, followed by E-wallet payments at $109,993.16, and Credit card transactions at $100,767.12. The relatively even distribution between payment methods suggests customers are comfortable using all available payment options, with a slight preference for cash payments.



This pattern indicates traditional cash payments remain popular, leading by a small margin. There is strong adoption of digital payment methods (E-wallet and Credit card combined), suggesting potential opportunity to encourage more credit card usage through targeted promotions.

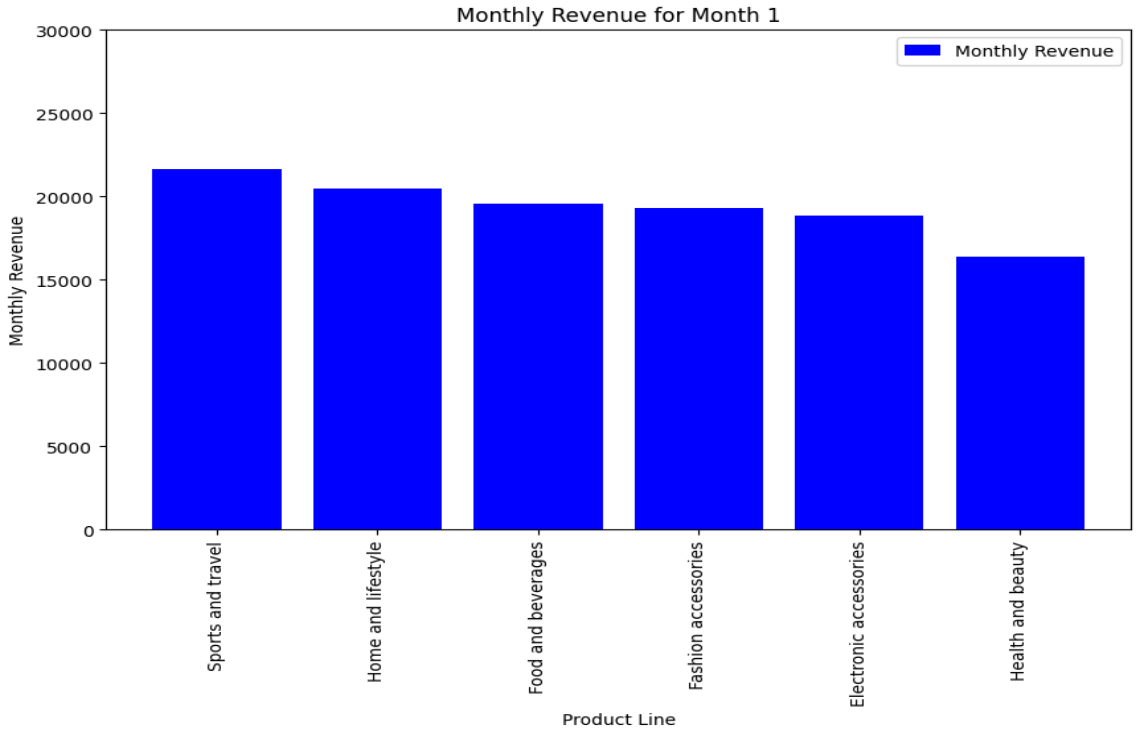
3.3 Monthly Product Revenue and Average Rating

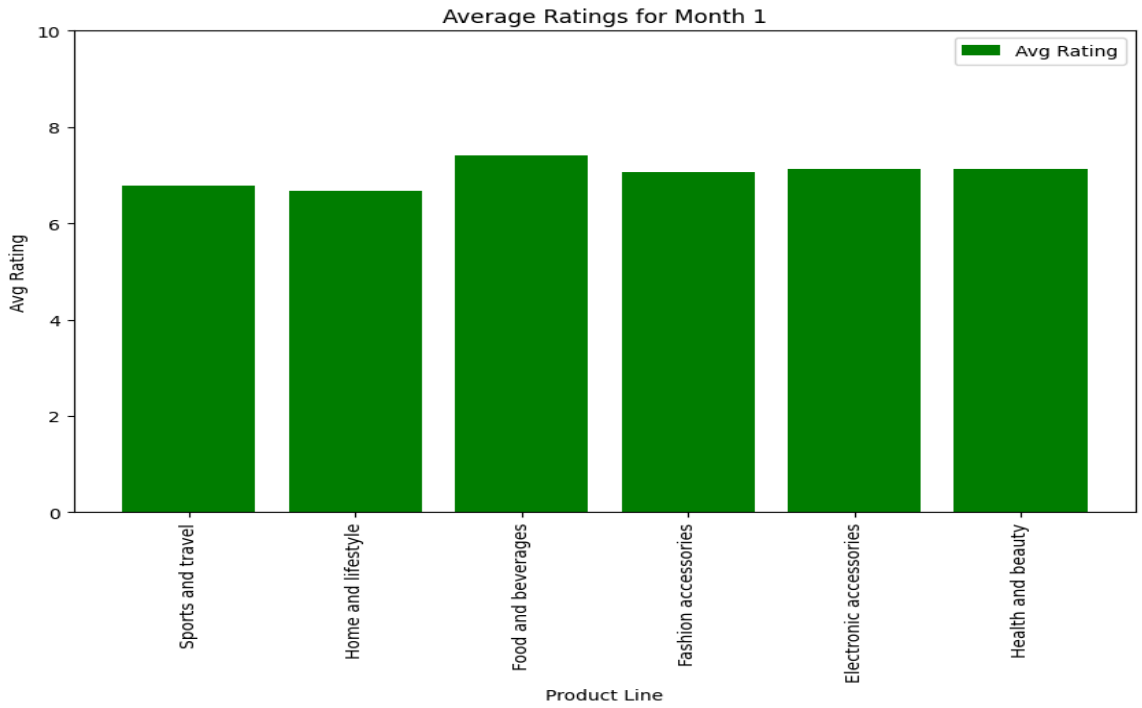
Objective: To analyse monthly revenue performance and customer satisfaction across different product lines, comparing patterns between revenue generation and customer ratings.

```
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2024-11-03 13:53:23,571 Stage-2 map = 0%,  reduce = 0%
2024-11-03 13:53:28,718 Stage-2 map = 100%,  reduce = 0%
2024-11-03 13:53:33,872 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_1730319140840_0030
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   HDFS Read: 154390 HDFS Write: 1077 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   HDFS Read: 9190 HDFS Write: 1115 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1       Sports and travel       21667.040088653564      6.78
1       Home and lifestyle      20494.74010658264       6.67
1       Food and beverages      19570.539932250977      7.41
1       Fashion accessories     19345.100058555603      7.06
1       Electronic accessories  18831.269985198975      7.14
1       Health and beauty       16383.17997932434       7.13
2       Food and beverages      20000.35987472534       7.01
2       Fashion accessories     19009.899995803833      7.08
2       Electronic accessories  17362.859901428223      6.96
2       Health and beauty       14602.25987625122       7.21
2       Sports and travel       13809.629955291748      7.13
2       Home and lifestyle      12434.38000869751       7.07
3       Home and lifestyle      20932.789892196655      6.85
3       Sports and travel       19646.239992141724      6.92
3       Health and beauty       18208.30002593994       6.73
3       Electronic accessories  18143.329977035522      6.7
3       Food and beverages      16573.970142364502      6.93
3       Fashion accessories     15950.909873962402      6.93
Time taken: 50.53 seconds, Fetched: 18 row(s)
```

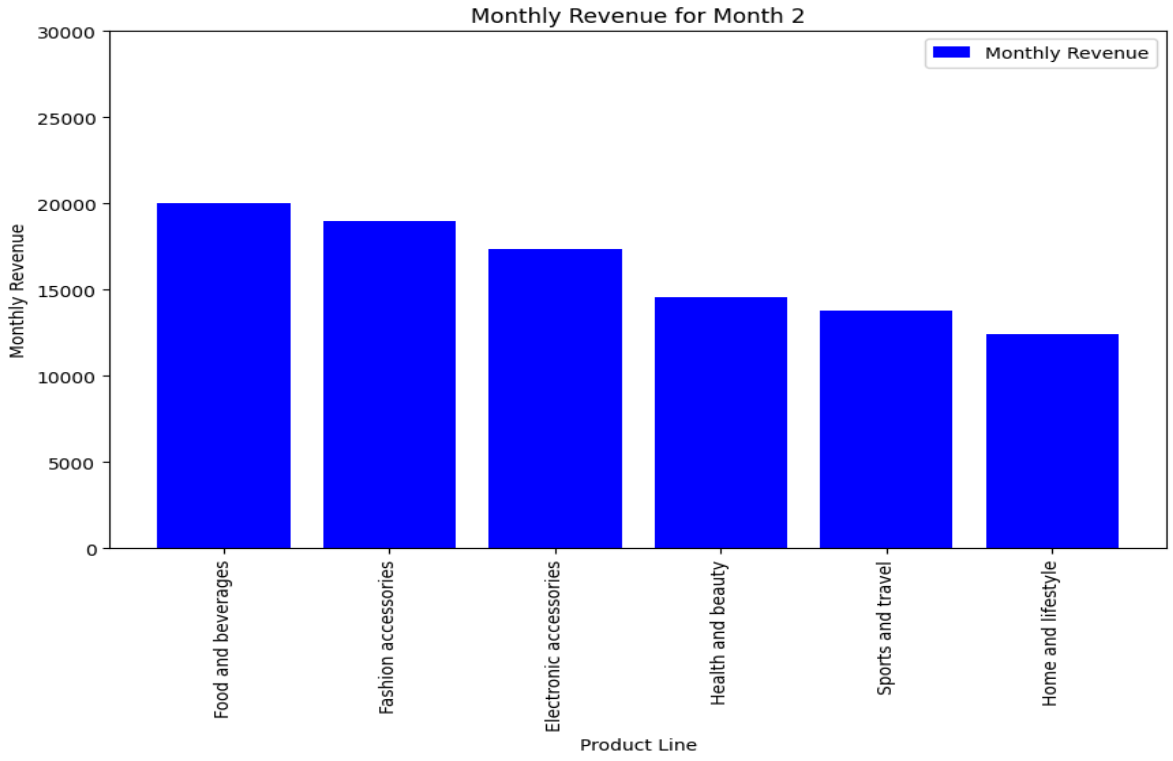
For each of the months I have created visualisations for the monthly revenue and also the customer rating to help us better visualise the output and spot any trends that might be present.

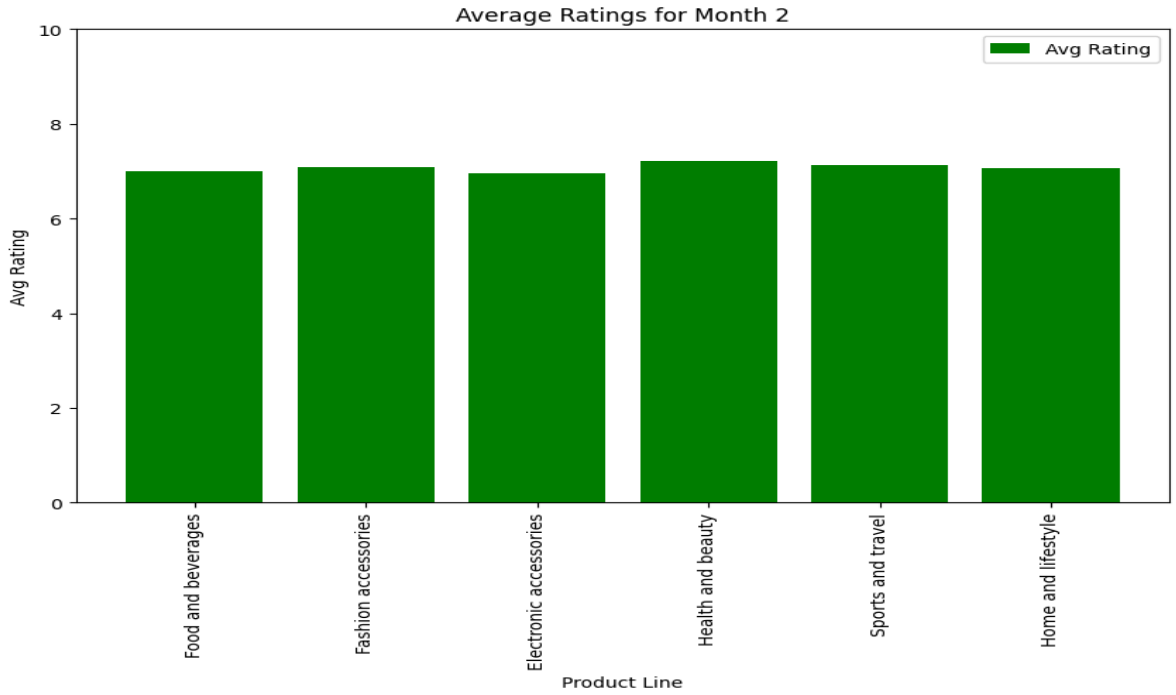
Month 1: Sports and travel leads revenue ($21,667), though with a lower rating (6.78). Food and beverages showed strong customer satisfaction (7.41) despite moderate revenue ($19,570). Health and beauty has the lowest revenue ($16,383) but maintains good ratings (7.13).





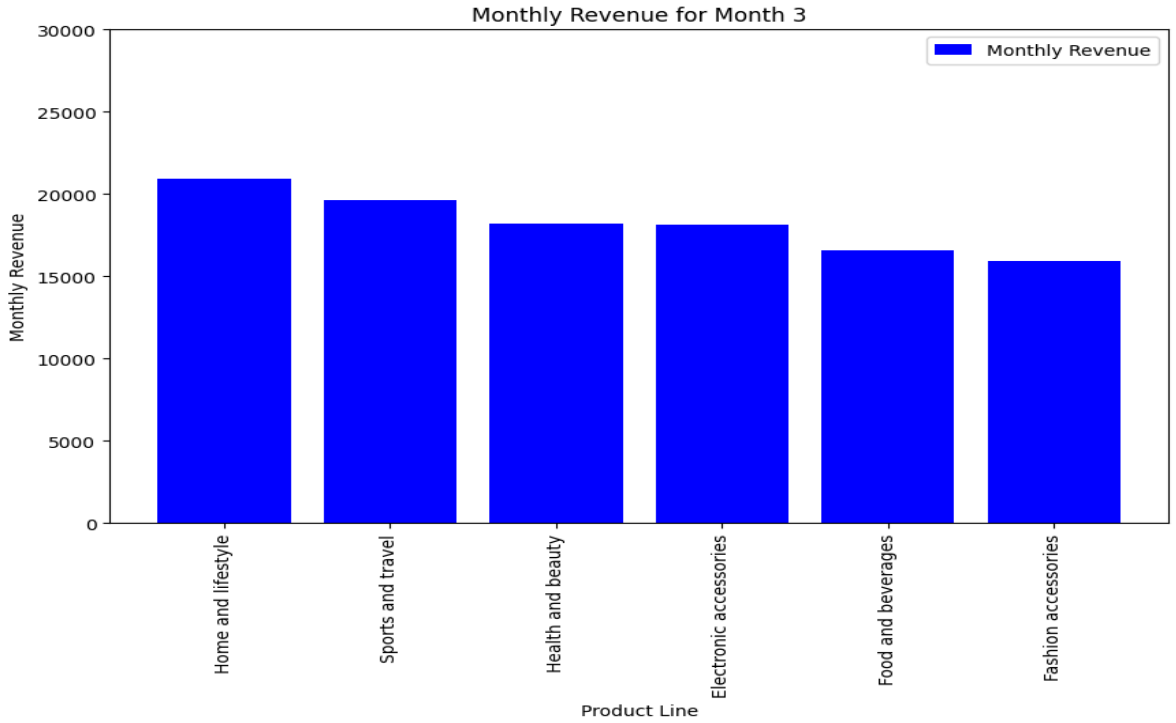
Month 2: Food and beverages takes the lead in revenue ($20,000), maintaining high satisfaction (7.01). Fashion accessories follows with $19,009 in revenue and 7.08 rating. Home and lifestyle shows the lowest revenue ($12,434) with an average rating of 7.07.

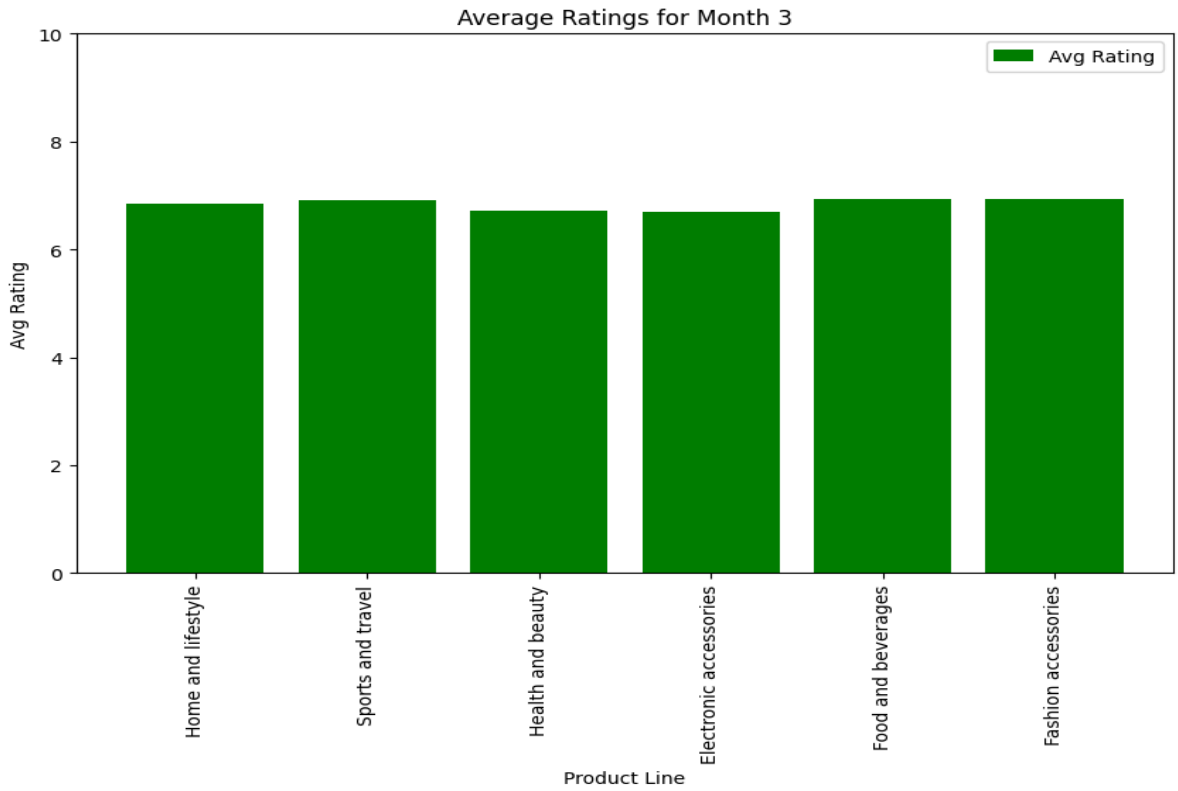




Month 3: Home and lifestyle recovers to lead revenue ($20,932) despite lower ratings (6.85). Sports and travel performs consistently ($19,646) with stable ratings (6.92). Fashion accessories

show the lowest revenue ($15,950) with average ratings (6.93)





Product line performance fluctuates significantly month to month, while higher revenue doesn't necessarily correlate with higher customer satisfaction. Food and beverages consistently maintains high customer satisfaction, and most product lines maintain ratings between 6.5-7.5 across months.

Complex Query 2: Sampled Customer Payment Analysis

Objective: To analyse the relationship between payment methods, customer types, and revenue across different product lines, using a sampled dataset to understand purchasing patterns.

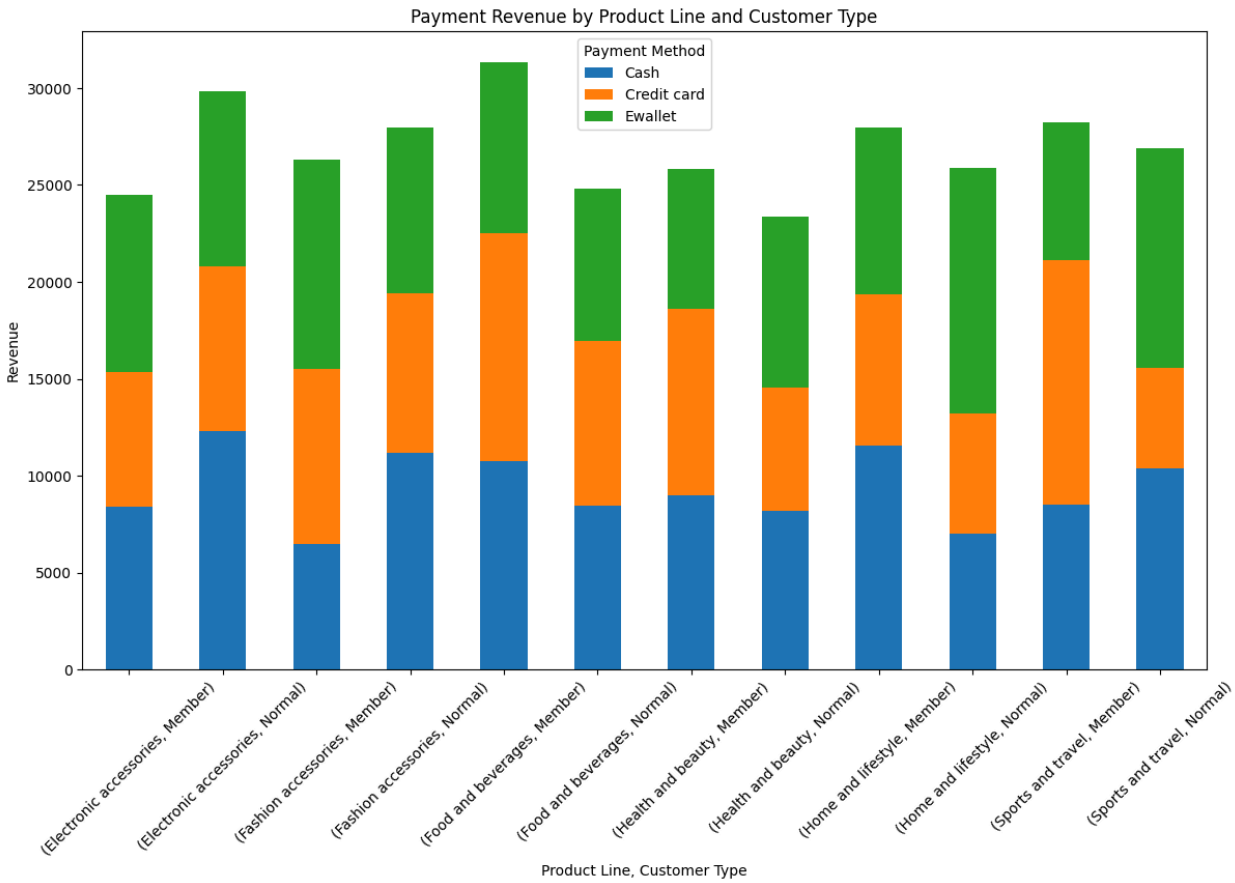
```
Stage-Stage-1: Map: 1  Reduce: 1   HDFS Read: 153780 HDFS Write: 2250 SUCCESS
Stage-Stage-6: Map: 1  Reduce: 1   HDFS Read: 152299 HDFS Write: 363 SUCCESS
Stage-Stage-9: Map: 1   HDFS Read: 8725 HDFS Write: 2538 SUCCESS
Stage-Stage-4: Map: 1  Reduce: 1   HDFS Read: 22719 HDFS Write: 2826 SUCCESS
Stage-Stage-5: Map: 1  Reduce: 1   HDFS Read: 11966 HDFS Write: 2813 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Electronic accessories  Normal  Cash    12324.439975738525      22.68   7.06
Electronic accessories  Member  Ewallet 9157.869983673096       16.85   6.34
Electronic accessories  Normal  Ewallet 9021.659900665283       16.6    7.0
Electronic accessories  Normal  Credit card     8492.909997940063       15.63   6.53
Electronic accessories  Member  Cash    8405.479991912842       15.47   6.76
Electronic accessories  Member  Credit card     6935.10001373291        12.76   7.9
Fashion accessories     Normal  Cash    11157.5199508667        20.55   7.22
Fashion accessories     Member  Ewallet 10811.080028533936      19.91   7.12
Fashion accessories     Member  Credit card     9056.56995010376        16.68   6.97
Fashion accessories     Normal  Ewallet 8545.969972610474       15.74   7.25
Fashion accessories     Normal  Credit card     8278.489953041077       15.24   6.77
Fashion accessories     Member  Cash    6456.280073165894       11.89   6.66
Food and beverages      Member  Credit card     11759.089958190918      20.94   7.17
Food and beverages      Member  Cash    10758.010028839111      19.16   7.35
```

Electronic Accessories: Normal customers prefer cash ($12,324) with highest ratings (7.06), while members show stronger preference for E-wallet ($9,157). Credit card usage is lowest among both customer types.

Fashion Accessories: Normal customers favour cash payments ($11,157), while members prefer E-wallet transactions ($10,811). Both groups maintain consistent ratings (6.66-7.25).

Food and Beverages: Members show high credit card usage ($11,759) and strong cash transactions ($10,758), while normal customers have more balanced payment distribution.

Home and Lifestyle: E-wallet leads for normal customers ($12,654), while members prefer cash payments ($11,564). Credit card usage is lowest across both groups.



Members generally prefer digital payment methods, while normal customers tend toward traditional cash payments. Payment preferences vary significantly by product category, while customer ratings remain relatively consistent across payment methods. The highest revenue contribution comes from normal customers using cash in electronic accessories (22.68%).

Complex Query 3: Top City Product Sales and Revenue Contribution

Objective: To analyse the distribution of product line revenues across different cities and customer types, identifying key revenue contributors and geographical patterns.

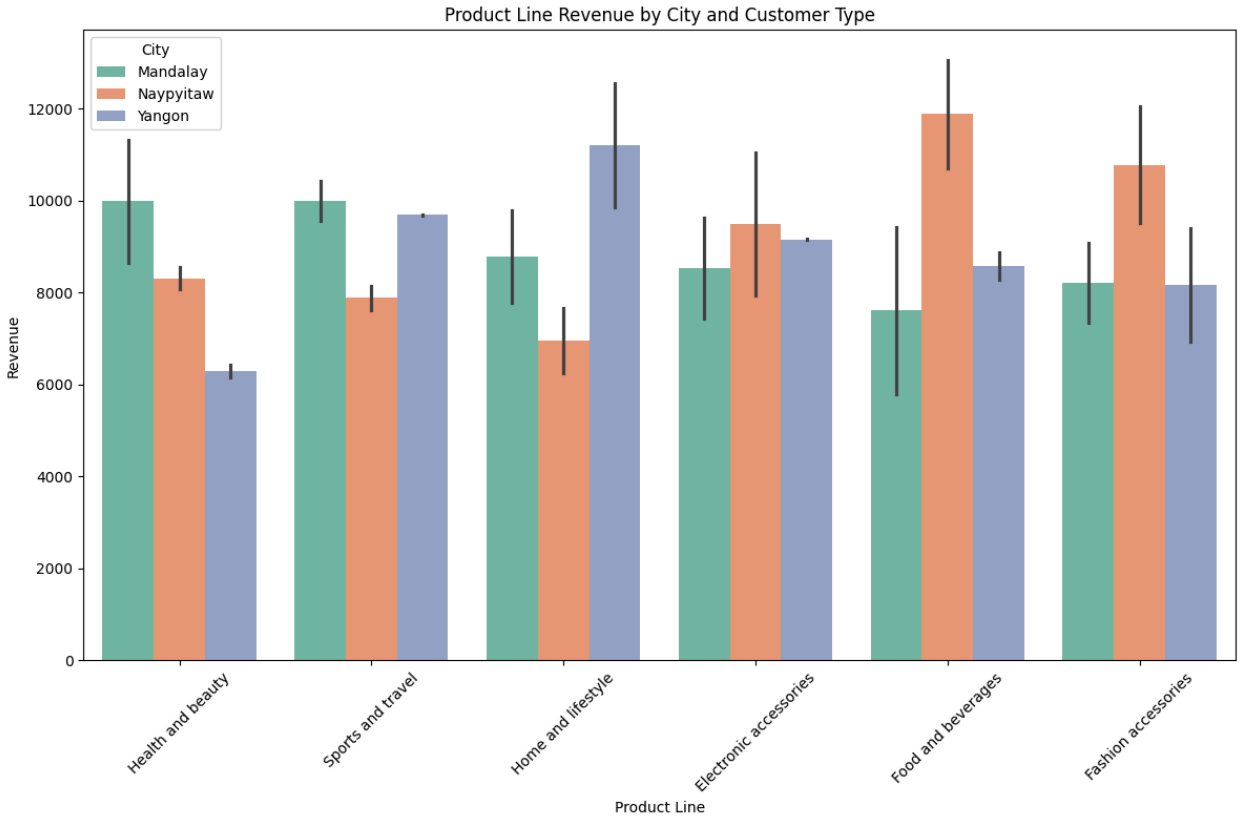
```
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2024-11-03 13:53:23,571 Stage-2 map = 0%,  reduce = 0%
2024-11-03 13:53:28,718 Stage-2 map = 100%,  reduce = 0%
2024-11-03 13:53:33,872 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_1730319140840_0030
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   HDFS Read: 154390 HDFS Write: 1077 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   HDFS Read: 9190 HDFS Write: 1115 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1       Sports and travel       21667.040088653564      6.78
1       Home and lifestyle      20494.74010658264       6.67
1       Food and beverages      19570.539932250977      7.41
1       Fashion accessories     19345.100058555603      7.06
1       Electronic accessories  18831.269985198975      7.14
1       Health and beauty       16383.17997932434       7.13
2       Food and beverages      20000.35987472534       7.01
2       Fashion accessories     19009.899995803833      7.08
2       Electronic accessories  17362.859901428223      6.96
2       Health and beauty       14602.25987625122       7.21
2       Sports and travel       13809.629955291748      7.13
2       Home and lifestyle      12434.38000869751       7.07
3       Home and lifestyle      20932.789892196655      6.85
3       Sports and travel       19646.239992141724      6.92
3       Health and beauty       18208.30002593994       6.73
3       Electronic accessories  18143.329977035522      6.7
3       Food and beverages      16573.970142364502      6.93
3       Fashion accessories     15950.909873962402      6.93
Time taken: 50.53 seconds, Fetched: 18 row(s)
```

From the output we can see that for the cities:

Mandalay: Health and beauty leads with highest revenue ($11,327) for members, with sports and travel showing strong performance ($10,420) for members. Electronic accessories perform well ($9,626) among normal customers, while food and beverages has the lowest revenue contribution (5.45%) for normal customers.

Naypyitaw: Food and beverages dominates with highest revenue ($13,057) from members, with fashion accessories following closely ($12,041) for members. Electronic accessories show strong performance ($11,040) among normal customers, while home and lifestyle has the lowest contribution (5.65%) from normal customers.

Yangon: Home and lifestyle leads revenue ($12,556) for members, with sports and travel showing consistent performance across both customer types. Health and beauty shows lowest performance ($6,158) among normal customers.



Members consistently generate higher revenue across most categories, and each city shows distinct product preferences. Revenue contribution varies significantly by product line and city, while normal customers show different purchasing patterns compared to members.

4 Conclusion

This analysis demonstrated the effectiveness of Apache Pig and Hive in processing and analysing large-scale data. Through the combination of these tools, we discovered that: Naypyitaw leads in gross income performance at $16.05 per sale; cash remains the preferred payment method while digital payments show strong adoption; and product performance varies significantly by location and customer type. Food and beverages consistently maintained high customer satisfaction despite revenue fluctuations, suggesting that higher revenue doesn't necessarily correlate with customer satisfaction. The use of Pig for data cleaning and basic analysis, complemented by Hive for complex queries, proved efficient for handling large data at scale. These insights could

directly inform business strategies around product mix optimisation, payment method promotions, and customer segment targeting.