



---

# REPORT

IT-300

## Business Intelligence and Database Management Systems

---

### Business Intelligence Research Shopping Trends in USA

---

*Authors:*

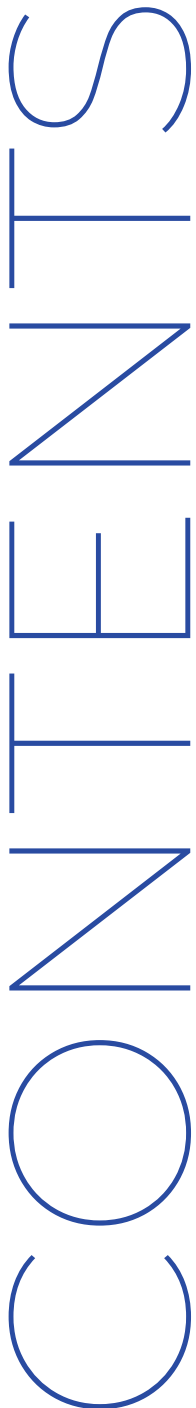
Maamoun Ghnimi  
Wacel Rahal  
Mehdi Benothmen

*Submitted to:*

Prof. Manel Abdelkader

2023-2024

# Table of Contents



## **01.**

### **Introduction**

## **02.**

### **Implementation**

#### 2.1

Data Gathering

#### 2.2

Data Preparation

#### 2.3

Data Storage

##### 2.3.1

Storage

##### 2.3.2

Fact

##### 2.3.3

Dimensions

#### 2.4

Data Visualization

## **03.**

### **Conclusion**

Understanding shopping trends is crucial for businesses to stay competitive and responsive to customer preferences. This business intelligence research aims to shed light on shopping trends in the USA, utilizing a comprehensive dataset that encompasses various customer demographics, purchasing patterns, and product preferences.

The primary objective of this project is to analyze and derive actionable insights from the provided dataset. By examining customer information, purchase details, and reviewing patterns, we aim to uncover trends that can inform strategic decision-making for businesses operating in the retail sector.

Specifically, our focus includes:

1. **Demographic Analysis:** Understanding the age, gender, and location distribution of customers to identify potential target markets and tailor marketing strategies.
2. **Product Preferences:** Examining the types of items customers purchase, their preferred categories, sizes, colors, and seasonal inclinations, providing valuable information for inventory management and product assortment.
3. **Purchase Behavior:** Analyzing the frequency of purchases, preferred payment methods, and shipping choices to optimize the overall shopping experience and streamline operational processes.
4. **Customer Satisfaction:** Exploring customer review ratings to gauge satisfaction levels and identify areas for potential improvement in products or services.

By accomplishing these objectives, businesses can gain a deeper understanding of their customer base, enabling them to refine marketing strategies, optimize inventory management, and enhance overall customer satisfaction – ultimately contributing to sustained business growth in the competitive retail landscape

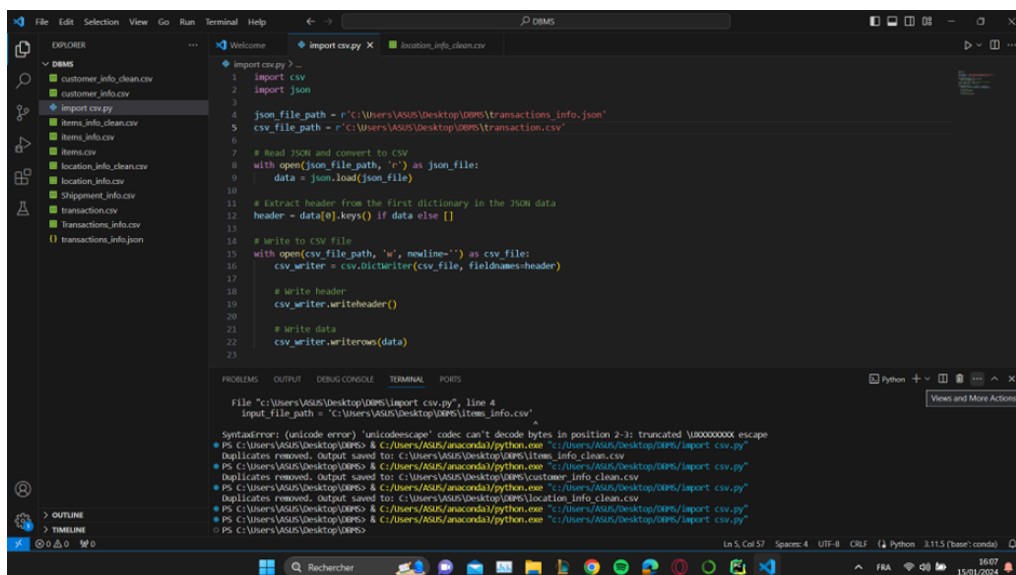
## 02. IMPLEMENTATION

### 2.1 DATA GATHERING

We extracted the Shopping Trends ratings dataset in USA from kaggle. This is a link to the [dataset](#)

### 2.2 DATA PREPARATION

For data preparation, Python was employed to streamline the manipulation and configuration of our dataset for integration into the data warehouse. The dataset existed in two formats: JSON and CSV. Utilizing Python, we seamlessly converted the JSON format into CSV, facilitating smoother data manipulation.



```
import csv
import json

json_file_path = r"C:\Users\ASUS\Desktop\DWPS\transactions_info.json"
csv_file_path = r"C:\Users\ASUS\Desktop\DWPS\transaction.csv"

# Read JSON and convert to CSV
with open(json_file_path, 'r') as json_file:
    data = json.load(json_file)

# Extract header from the first dictionary in the JSON data
header = data[0].keys() if data else []

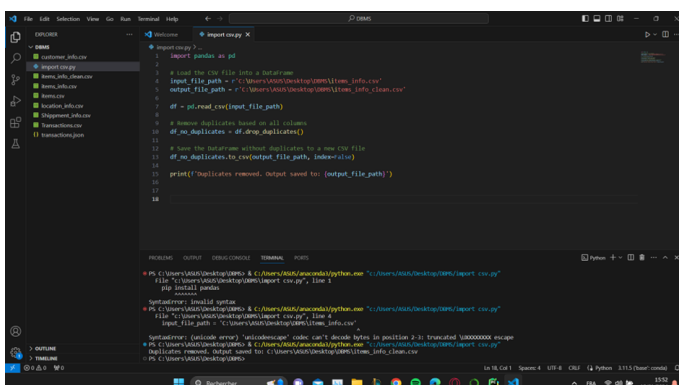
# Write to CSV file
with open(csv_file_path, 'w', newline='') as csv_file:
    csv_writer = csv.DictWriter(csv_file, fieldnames=header)

    # Write header
    csv_writer.writeheader()

    # Write data
    csv_writer.writerows(data)
```

#### Transitioning into the transformation phase:

-A critical observation unveiled the presence of duplicated data. To address this challenge, we leveraged Python to identify and resolve duplicate rows in the dataset.



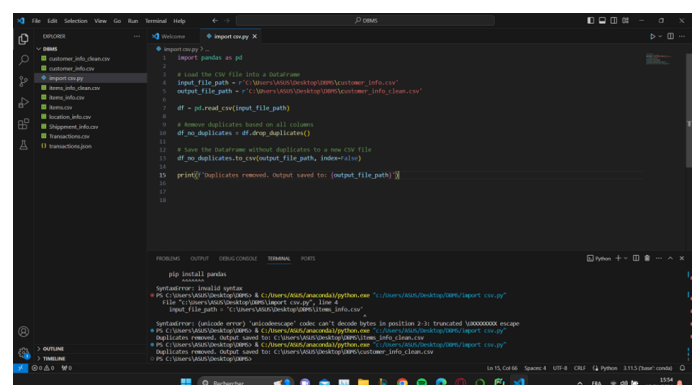
```
import pandas as pd

# Load the CSV file into a dataframe
input_file_path = r"C:\Users\ASUS\Desktop\DWPS\transaction.csv"
df = pd.read_csv(input_file_path)

# Remove duplicates based on all columns
df_no_duplicates = df.drop_duplicates()

# Save the dataframe without duplicates to a new CSV file
df_no_duplicates.to_csv(output_file_path, index=False)

print("Duplicates removed. Output saved to: (output_file_path)")
```



```
import pandas as pd

# Load the CSV file into a dataframe
input_file_path = r"C:\Users\ASUS\Desktop\DWPS\customer_info.csv"
df = pd.read_csv(input_file_path)

# Remove duplicates based on all columns
df_no_duplicates = df.drop_duplicates()

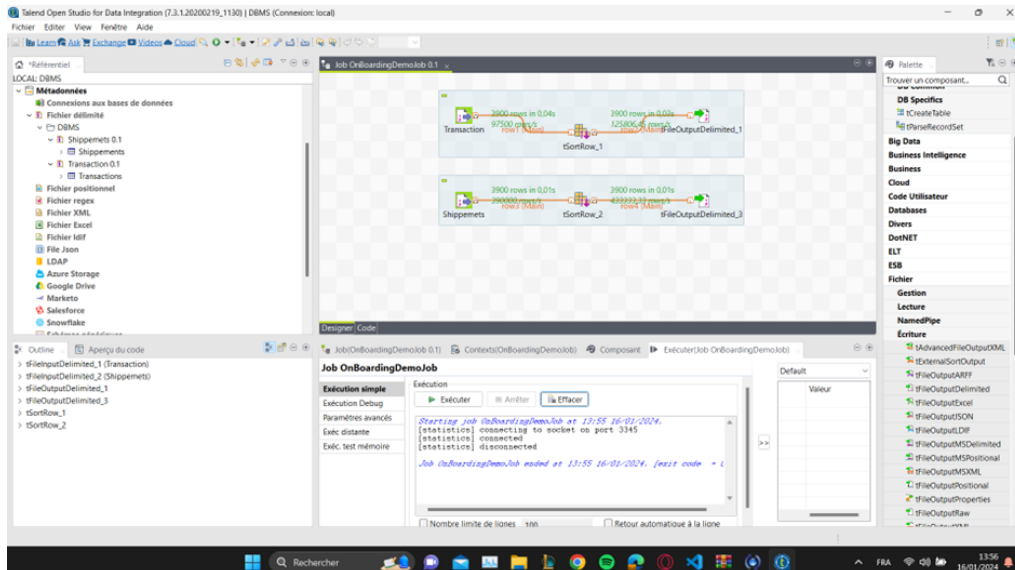
# Save the dataframe without duplicates to a new CSV file
df_no_duplicates.to_csv(output_file_path, index=False)

print("Duplicates removed. Output saved to: (output_file_path)")
```

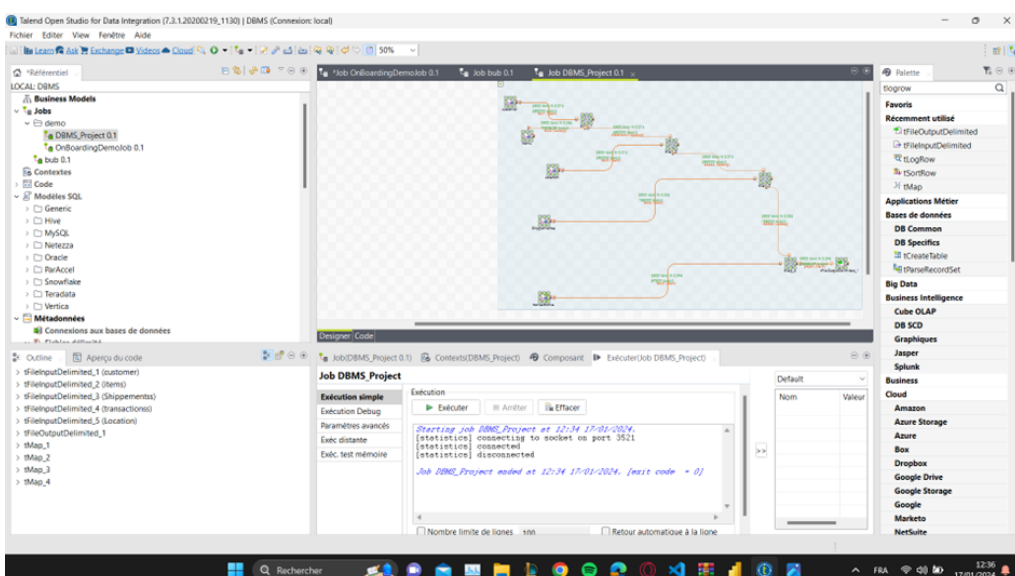
## 2.3 Data Storage :

### 2.3.1.Storage:

In this phase, we imported the modified data into Talend using Python. Prior to commencing the mapping phase, it was necessary to enhance the data organization by sorting it. We accomplished this task using the 'tSortRow' component in Talend, ensuring a well-ordered dataset for subsequent processes.



We initiated the mapping process for various datasets, including 'customers.csv,' 'shipment.csv,' 'transaction.csv,' 'items.csv,' and 'info.csv.' Employing the 'tMap' functionality, we combined these diverse datasets to create a unified table, primed for the subsequent visualization phase.



## 2.3 Data Storage :

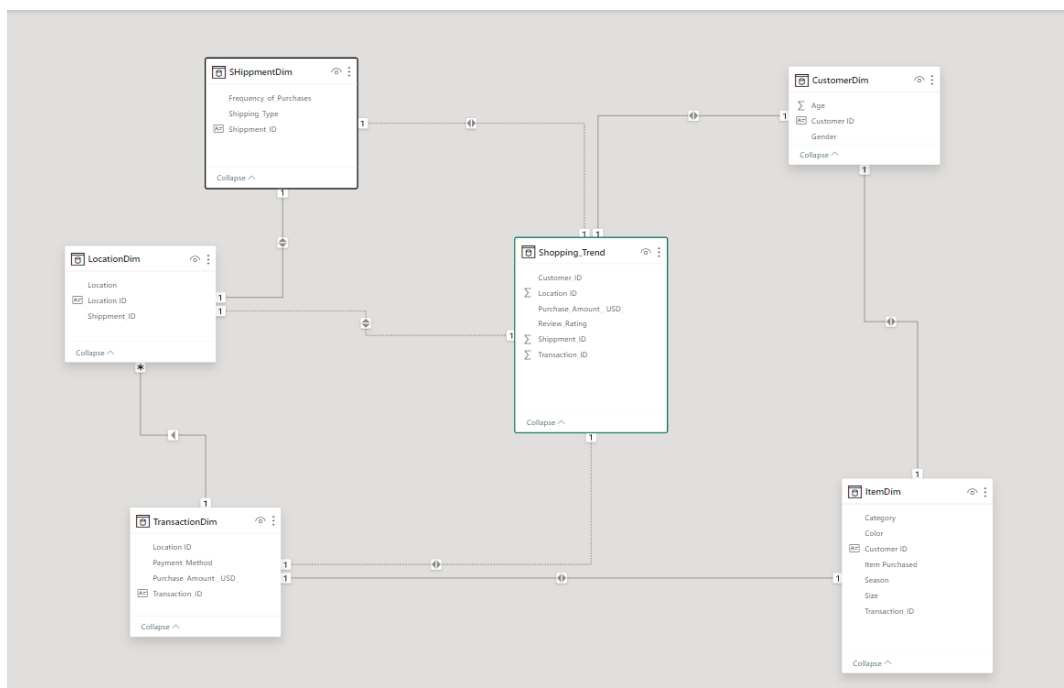
### 2.3.1.Storage:

*In the Loading Phase*, we imported the output files from Talend into pgAdmin4, an SQL tool. This step was essential for thorough data exploration in Power BI, facilitating the creation of comprehensive reports.

```
1 SELECT * FROM public.location_info_clean1
2 ORDER BY "Location ID" ASC
```

	Location ID [PK] integer	Location character varying	Shipment_ID integer
1	1	Kentucky	1
2	2	Maine	2
3	3	Massachusetts	3
4	4	Rhode Island	4
5	5	Oregon	5
6	6	Wyoming	6
7	7	Montana	7
8	8	Louisiana	8
9	9	West Virginia	9

In this phase, we explore the data modeling process, emphasizing the structuring of our dataset to facilitate efficient analysis and reporting. The key elements of our data model encompass the Fact Table, Dimensions, Measures, and the selected schema.



## 2.3 Data Storage :

### 2.3.2.Fact:

The central table in our data model is **"Shopping\_Trends"**. This fact table holds key information about customer reviews and purchase amount. This includes details like Customer ID, Location ID, Shippment ID , Transaction ID , review scores and purchase amounts. Together, these elements make up the central focus of our analysis, offering valuable insights into customer feedback and shopping trends.

### 2.3.2.Dimensions:

To enrich our analysis, we employ various Dimension tables, each providing a distinct perspective on the data . These dimensions include :

**-Custmer Dimension :** Represents the Customer information.

**-Items Dimension :** Represents the items' information. It includes many features such as category, size, color, and item purchased.

**-Location Dimension:** Represents the location of the transaction.

**-Transactions Dimension:** Represents the transactions' information. It includes the payment method.

**-Shippment Dimension:** Represents the shippment information.

\*In the end, We opted for the Snowflake Schema to improve data organization and minimize redundancy in our database.By breaking down dimension tables into related tables, updates and modifications can be made more efficiently. Although it may introduce some additional complexity in queries due to the need for joins, the overall benefits include better scalability, easier maintenance, and improved performance in certain scenarios.

## 2.4 Data Visualization

### Data Visualization Summary:

## 1. Customer Demographics:

- Men constitute the majority of customers, with **2,652** individuals, representing 68% of the total.

## 2.Customer Age:

- The average age of customers is 44 years.

### 3. Purchase Statistics:

- The total sum of all purchases amounts to approximately **\$233,081**.

#### 4. Payment Preferences:

- o Credit card emerges as the most preferred payment method, accounting for **696** transactions.

### 5. Popular Items:

- Top three items purchased include pants, a blouse, and jewellery.

## 6. Clothing Size Trends:

- The most frequent clothing size is M.

### 7. Profitable and Least Profitable Items:

- Pants prove to be the most profitable item, generating **\$10,501**, while jeans rank as the least profitable with **\$7,326**.

### 8. Average Customer Rating:

- o The average customer rating is **3.75**, ranging between **2.9** and **5.0**.

**9. Preferred Category:**

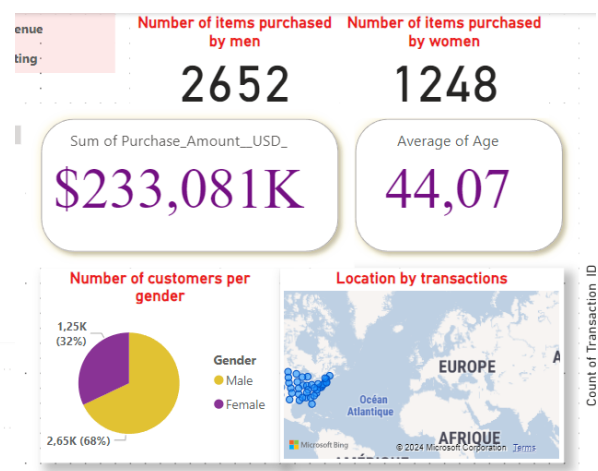
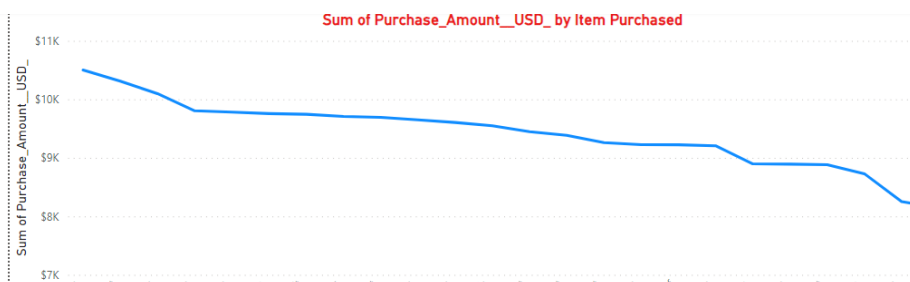
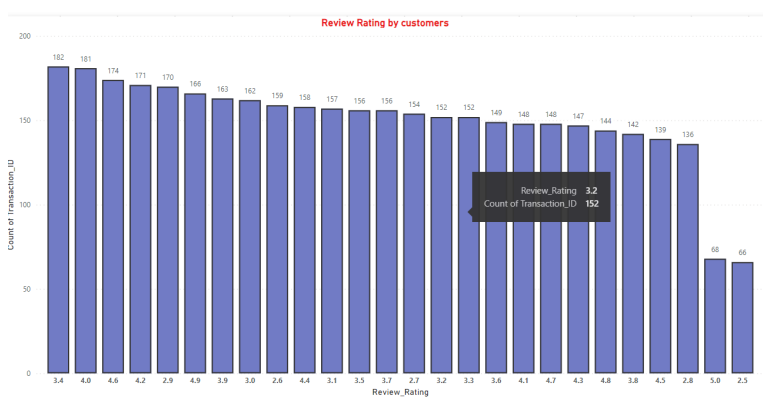
- Clothing emerges as the favorite category among customers.

## 10. Shipping Insights:

- Free shipping is the most favored shipping method.
- Shipping frequency peaks around every 3 months.

## 11. Seasonal Transactions:

- Transactions exhibit consistent volumes across each season.





### 3. Conclusion

In summary, the analysis of shopping trends in the USA reveals a dynamic and robust consumer market. With men comprising the majority of customers and an average customer age of 44 years, there is a diverse demographic to target. The substantial total sum of purchases (\$233,081) reflects a healthy financial landscape, emphasizing the potential for business growth. Credit card transactions dominate, emphasizing the importance of secure and convenient payment options.

Top-selling items such as pants, blouses, and jewellery provide valuable insights for inventory management and marketing strategies. Understanding the most profitable items, like pants, and identifying the least profitable, such as jeans, allows for strategic merchandising decisions.

Customer satisfaction, with an average rating of 3.75, suggests a positive sentiment, providing a foundation for building and maintaining a loyal customer base. The preference for clothing as the favorite category underscores the significance of offering a diverse and appealing range of clothing items.

Shipping preferences, particularly the popularity of free shipping and the highest shipping frequency occurring every 3 months, can inform shipping policies and promotional initiatives. Additionally, the consistent transaction volumes across seasons indicate a stable demand, allowing for effective resource allocation and inventory planning.

In conclusion, this comprehensive analysis provides valuable insights for tailoring marketing, inventory, and customer service strategies. Leveraging these insights will contribute to a customer-centric approach, enhancing the overall shopping experience and ensuring competitiveness in the USA retail market.

# Challenges we faced:

We encountered several challenges throughout our project:

## **1. Research Phase Difficulty:**

- The research phase posed challenges as we struggled to find the ideal dataset for our project. Locating suitable and comprehensive data proved to be a significant hurdle.

## **2. Data Quality Issues:**

- Dealing with numerous duplicates and misleading data added complexity to the project. Ensuring data accuracy and reliability became a crucial task.

## **3. Mapping Complexity in Talend:**

- The mapping phase in Talend presented challenges, particularly when linking all five subtables using the tMap function. To overcome this, we adopted a stepwise approach, initially linking two subtables and progressively incorporating the others into the output, eventually achieving the final comprehensive output.

## **4. Power BI Range Limitation:**

- Challenges arose in Power BI due to the relatively narrow data range. To address this limitation, we implemented the use of additional measures to generate a more meaningful and insightful final report.

Navigating these challenges required creative problem-solving and iterative approaches to ensure the successful completion of our data management and visualization project.