

# **Analyzing e-commerce sales data using Hive programming and visualization with R programming.**

Project submitted to the  
SRM University – AP, Andhra Pradesh  
for the partial fulfillment of the requirements to award the degree of

**Bachelor of Technology**

In

**Computer Science and Engineering  
School of Engineering and Sciences**

Submitted by

[Kavya Keerthana Palisetti | AP20110010668](#)

**Chandanarchutha Namburu | AP20110010660**

[Harshavardhan Ganduri | AP20110010689](#)



Under the Guidance of  
**Dr. Sriramulu Bojjagani**  
**SRM University–AP**  
**Neerukonda, Mangalagiri, Guntur**  
**Andhra Pradesh – 522 240**

**[Nov, 2023]**

# Certificate

Date: 13-Nov-23

This is to certify that the work presented in this Project entitled “**Analyzing e-commerce sales data using Hive programming and visualization with R programming.**” has been carried out by **Kavya Keerthana Paliseti, Chandanarchutha Namburu and Harshavardhan Ganduri** under my supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in **School of Engineering and Sciences**.

## Supervisor

(Signature)

Prof. / Dr. Sriramulu Bojjagani

Asst. professor,

Affiliation.

## Acknowledgements

First and foremost, I want to extend my heartfelt thanks to Dr. Sriramulu Bojjagani, who provided invaluable mentorship and feedback throughout the project. Their expertise and encouragement were instrumental in shaping the project's direction and ensuring its success.

We would like to express our appreciation to kaggle for granting us access to the e-commerce sales dataset. The availability of this valuable data has been instrumental in conducting a comprehensive analysis and deriving meaningful insights.

Special thanks go to the open-source community and the developers behind the tools and technologies used in this project, particularly the Apache Hive community. The power and flexibility of these tools have greatly facilitated the execution of complex data analyses and made this project feasible.

# Table of Contents

Abstract.....	4
Abrevations.....	5
List of Tables.....	6
List of figures.....	7
List of graphs.....	8
List of equations.....	9
1. Introduction.....	10
1.1 Motivation	
1.2 Objective	
1.3 Scope	
2. Methodology.....	12
2.1 Data Collection	
2.2 Data Preprocessing	
2.3 Data transformation	
2.4 Customer Segmentation	
2.5 Product popularity Analysis	
3. Discussion.....	20
4. Conclusion.....	25
5. Future Work.....	26

# Abstract

This project focuses on the analysis of e-commerce sales data using Apache Hive, aimed at gaining insights into customer purchasing behavior and product popularity. It involves the collection of e-commerce sales data, transformation and data preparation within Hive, and extensive querying using HiveQL for the extraction of meaningful insights.

The key areas of analysis in this project encompass customer segmentation, product popularity identification, and the evaluation of sales trends over time. Customer segmentation allows us to group customers based on their purchasing behavior, distinguishing frequent buyers from occasional ones. The identification of top-selling products enables us to recognize products that have garnered the most attention and sales. Additionally, the analysis of sales trends reveals the patterns and variations in sales over monthly or quarterly intervals.

The project demonstrates the capabilities of Apache Hive as a robust tool for data analysis and provides valuable insights for businesses in the e-commerce sector. By understanding customer behavior and recognizing product preferences, e-commerce businesses can make informed decisions, optimize their operations, and enhance customer satisfaction.

In summary, e-commerce data analysis is essential for making informed business decisions, improving customer experiences, optimizing operations, and staying competitive in a rapidly evolving online marketplace. It empowers businesses to adapt and thrive in the digital commerce landscape by leveraging data as a strategic asset.

## Abbreviations

SQL      Structured Query Language

## List of Tables

Table1. List of frequent customers.

Table2. Top 10 customers.

Table3.Top selling products.

## List of Figures

Figure1. First five rows of data set.

Figure 2. Null values.

Figure3. Null values removed from Description column.

Figure 4. Null values removed from CustomerID column.

Figure5. Negative quantity values.

Figure6. After removing negative values from the quantity column.

Figure 7. Dataset after cleaning.

Figure8. Sandbox login.

Figure9. Load data into hive table.

## **List of Graphs**

Graph1. Monthly sales data.

Graph2. daily sales data.

Graph3. Hourly sales data.

Graph4. Number of orders in each country (Without UK).



Graph5. Number of orders in each country (With UK).

Graph6. Money spent by different countries (including UK).

Graph7. Money spent by different countries (excluding UK).

## List of Equations

$\text{amount\_spent} = \text{quantity} * \text{unit\_price}$



# 1.Introduction

The e-commerce industry has witnessed a profound transformation in recent years, with online retail becoming a vital part of the modern consumer experience. As the e-commerce landscape continues to evolve, it has become increasingly crucial for businesses to gain a deep understanding of customer behavior, product performance, and sales trends. This project aims to address this imperative by conducting a comprehensive analysis of e-commerce sales data, leveraging the capabilities of Apache Hive.

## 1.1 Motivation:

E-commerce platforms generate vast quantities of transactional data daily, encompassing customer interactions, purchases, and product reviews. This wealth of data presents a unique opportunity to extract valuable insights, optimize strategies, and enhance the overall customer experience. In this context, the motivation for this project lies in harnessing the power of data analysis to answer fundamental questions:

What are the driving factors behind customer purchasing behavior in the e-commerce domain?

Which products resonate most with customers and exhibit the highest sales figures?

Can we discern trends and patterns in sales, helping businesses to make data-informed decisions?

## 1.2 Objectives:

The objectives of the project involve leveraging Apache Hive for analyzing e-commerce transaction data to derive actionable insights. These objectives typically include, but are not limited to, understanding customer behavior, optimizing inventory and pricing strategies, improving conversion rates, and enhancing the overall performance and profitability of the e-commerce business.

Additionally, the project may aim to identify emerging market trends, enhance customer retention strategies, and detect and prevent fraud. The ultimate goal is to use data analysis to inform strategic decision-making and drive business growth within the e-commerce sector.

## 1.3 Scope:

The scope of the project encompasses the entire data analysis process, from data collection and preparation to modeling, analysis, and reporting. It defines the boundaries of what will be included in the analysis, such as the specific data sources to be used, the time period under consideration, and the depth and breadth of analysis.

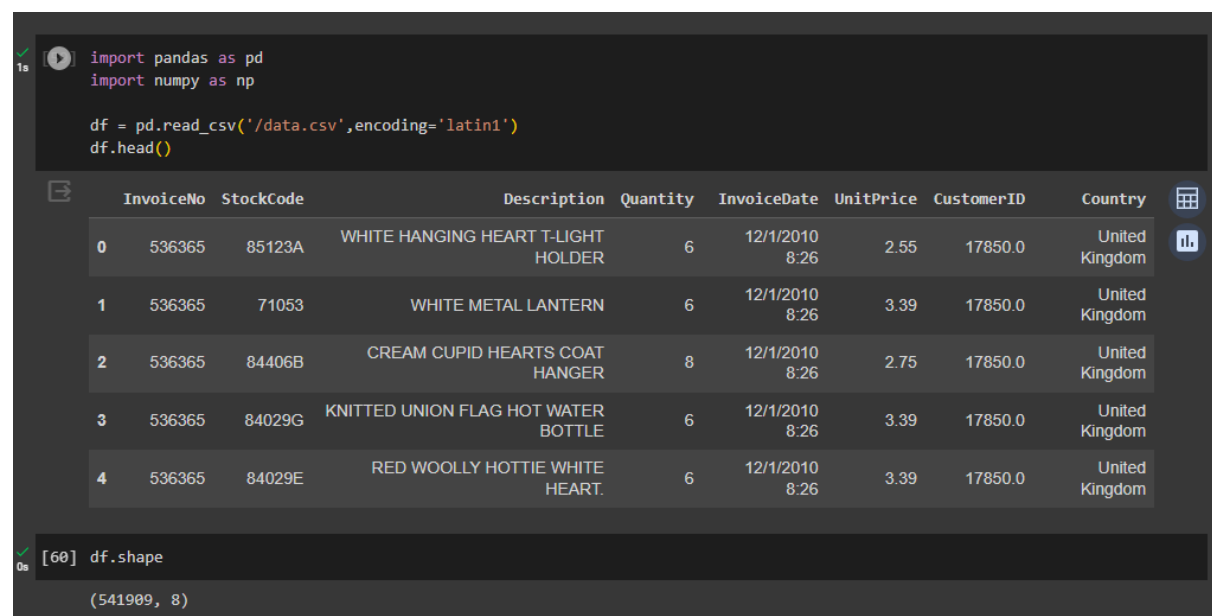
The scope also outlines the tools and technologies to be employed, including Apache Hive for SQL-based data analysis. Moreover, it defines the expected deliverables, which may include detailed reports, visualizations, and recommendations for the e-commerce business based on the analysis results. Clear scoping ensures that the project remains manageable and focused on achieving its intended goals.

## 2. Methodology

### 2.1 Data collection:

The E-commerce dataset is obtained from kaggle. The dataset consists of transactional data from 1st December 2010 to 9th December 2011 of customers in different countries who make purchases from an online retail company based in the United kingdom(UK) that sells unique all-occasion gifts.

The data set contains 8 columns and 541989 rows. Below is a snapshot of what the original data looks like after loading into the jupyter notebook.



The screenshot shows a Jupyter Notebook interface. The top cell contains Python code to load pandas and numpy, read a CSV file, and display the first five rows. The bottom cell shows the output of the shape function, indicating the dataset has 541989 rows and 8 columns.

```
import pandas as pd
import numpy as np

df = pd.read_csv('/data.csv', encoding='latin1')
df.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

```
[60] df.shape

(541989, 8)
```

Figure1. First five rows of data set.

Each column contains details like customer ID, Description of product etc...

Here is a brief description of each column:

**InvoiceNo** (*invoice\_num*): A number assigned to each transaction

**StockCode** (*stock\_code*): Product code

**Description** (*description*): Product name

**Quantity** (*quantity*): Number of products purchased for each transaction

**InvoiceDate** (*invoice\_date*): Timestamp for each transaction

**UnitPrice** (*unit\_price*): Product price per unit

**CustomerID** (*cust\_id*): Unique identifier each customer

**Country** (*country*): Country name

## 2.2 Data Preprocessing:

Data cleaning is a crucial step in the data analysis process. Raw data often contains errors, inconsistencies, and missing values. Cleaning the data helps ensure its accuracy and reliability, which are essential for making sound analyses and decisions. Data cleaning is a fundamental step in the data analysis process that ensures data accuracy, consistency, and reliability. It is necessary to eliminate potential sources of bias or error, making the data suitable for analysis and decision-making.

```
[63] df.isnull().sum()

InvoiceNo      0
StockCode      0
Description    1454
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID    135080
Country        0
dtype: int64
```

Figure 2. Null values

There are some missing values for description and Customer ID as shown in figure2. The rows with any of these missing values will therefore be removed.

```
[64] df.dropna(subset=['Description'], inplace=True)
df.isnull().sum()

InvoiceNo      0
StockCode      0
Description     0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID    133626
Country        0
dtype: int64

[65] df.shape

(540455, 8)
```

Figure3. Null values removed from Description column.

```
df.dropna(subset=['CustomerID'], inplace=True)
df.isnull().sum()

InvoiceNo    0
StockCode    0
Description   0
Quantity     0
InvoiceDate  0
UnitPrice    0
CustomerID   0
Country      0
dtype: int64

[67] df.shape

(406829, 8)
```

Figure 4. Null values removed from CustomerID column.

the null values in the description column and customerID column are removed. and there are 406829 rows in the dataset now.

```
df.describe()
```

	Quantity	UnitPrice	CustomerID
count	406829.000000	406829.000000	406829.000000
mean	12.061303	3.460471	15287.690570
std	248.693370	69.315162	1713.600303
min	-80995.000000	0.000000	12346.000000
25%	2.000000	1.250000	13953.000000
50%	5.000000	1.950000	15152.000000
75%	12.000000	3.750000	16791.000000
max	80995.000000	38970.000000	18287.000000

Figure5. Negative quantity values

By understanding the data in a more descriptive manner, we see that there are negative values in the quantity column. At this stage we'll just remove the quantity with negative values.

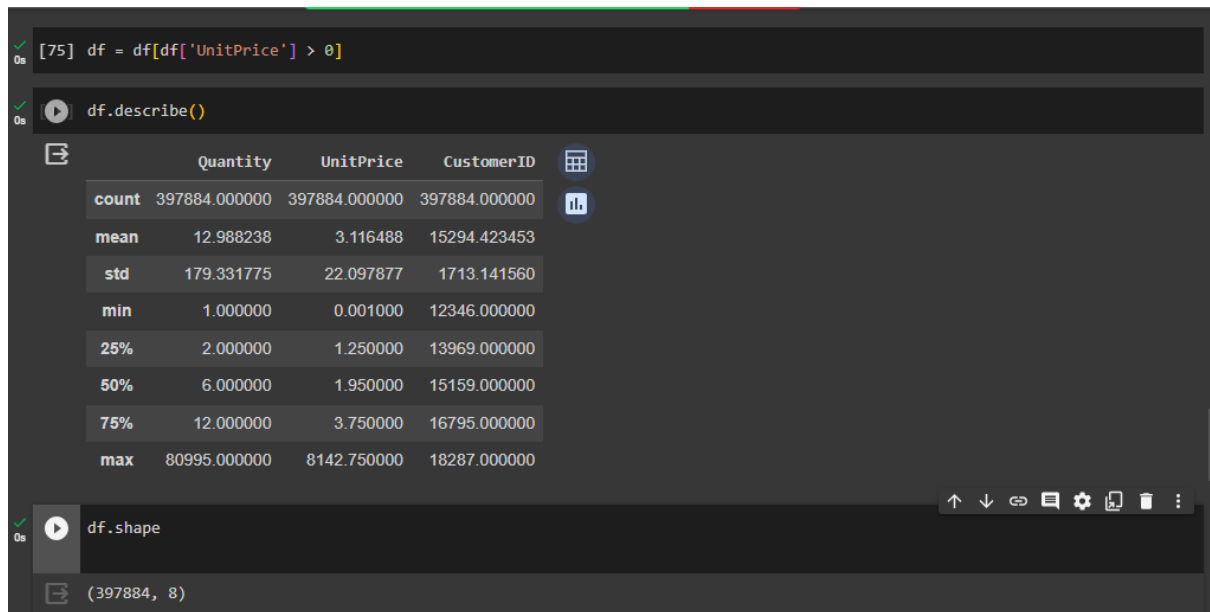


Figure6. After removing negative values from the quantity column.

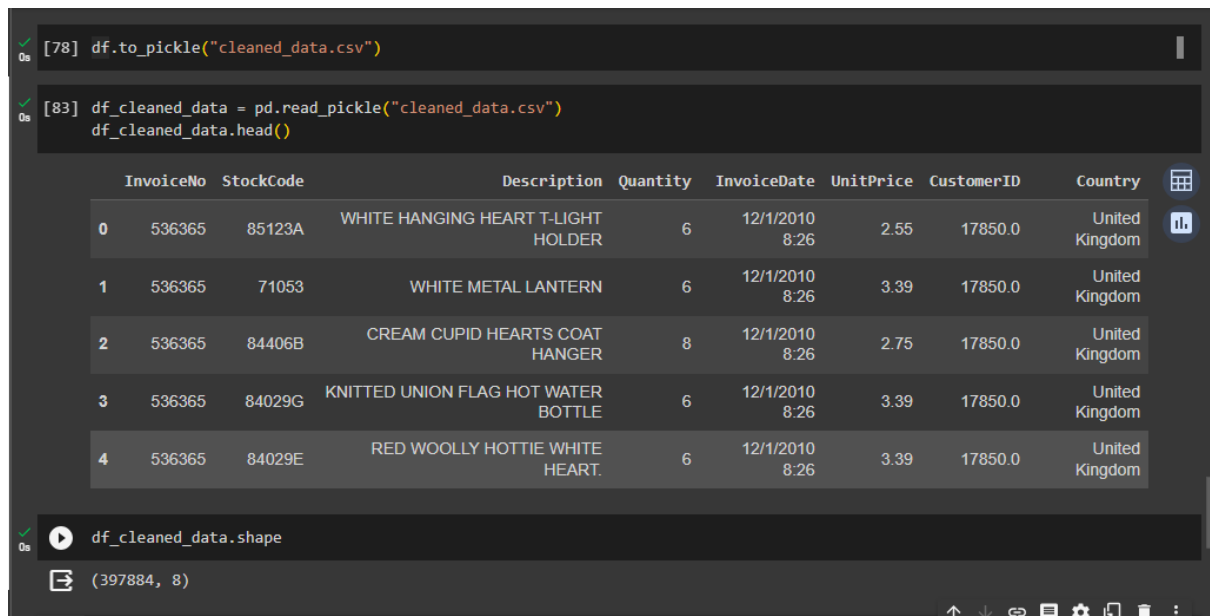


Figure 7. Dataset after cleaning.



### 2.3 Data Transformation:

Create Hive tables that represent the data, ensuring that data types and structures are appropriate for analysis.

```
sandbox login: root
root@sandbox.hortonworks.com's password:
Last login: Mon Nov 13 15:46:43 2023 from 172.17.0.2
[root@sandbox ~]# hdfs dfs -get /POBD/input/sales_data.csv
get: `sales_data.csv': File exists
[root@sandbox ~]# hive
```

Figure8. Sandbox login

```
hive>
>
>
>
>
>
>
> CREATE TABLE sales(InvoiceNo INT, StockCode STRING,Description
> STRING,Quantity INT,InvoiceDate STRING, Time STRING,
> UnitPrice INT, CustomerID INT, Country STRING)
> ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. AlreadyExistsException(message:Table sales already exists)
hive> LOAD DATA LOCAL INPATH 'sales_data.csv' INTO TABLE sales;
Loading data to table default.sales
Table default.sales stats: [numFiles=2, numRows=0, totalSize=67784786, rawDataSize=0]
OK
Time taken: 3.645 seconds
hive> █
```

Figure9. Load data into hive table.

### 2.4 Customer Segmentation:

Segmentation Criteria: Customer segmentation was conducted based on the number of purchases. Customers were categorized as frequent buyers if they had made more than 25 purchases, while others were considered occasional buyers.

query1: frequent customers

frequent_customers.customerid	frequent_customers.order_count
null	632
0	1034
1	122
2	254
3	41
12	83
12471	30
12569	32
12682	31
12720	25
12748	209
12841	25
12901	28
12921	37
12971	86
13018	28
-----	--

Table1. List of frequent customers.

Query2: Top customers

<code>top_customers.customerid</code>	<code>top_customers.total_purchase_amount</code>
18102	447974
14646	370782
17450	336342
14911	205018
14156	176414
12415	175784
12346	148430
17511	118642
16029	118280
14096	113104

Table2. Top 10 customers.

## 2.5. Product Popularity Analysis:

The analysis identified the top-selling products based on the total sales amount.

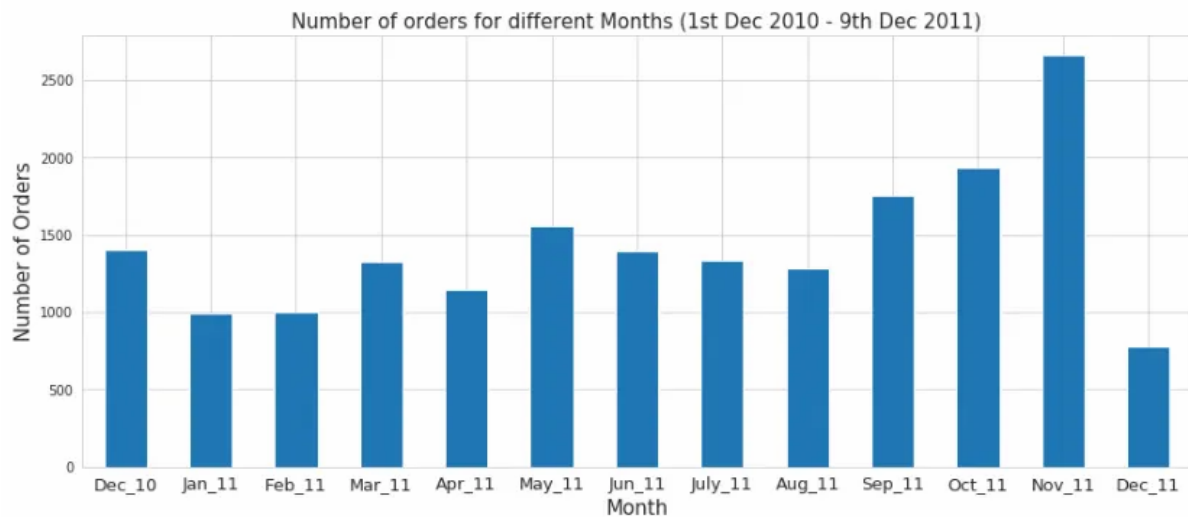
Query 3:

top_selling_products.stockcode	top_selling_products.description	top_selling_products.total_quantity_sold
23166	MEDIUM CERAMIC TOP STORAGE JAR	77916
84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	54319
85099B	JUMBO BAG RED RETROSPOT	46078
85123A	WHITE HANGING HEART T-LIGHT HOLDER	36706
84879	ASSORTED COLOUR BIRD ORNAMENT	35263
21212	PACK OF 72 RETROSPOT CAKE CASES	33670
22197	POPCORN HOLDER	30919
23084	RABBIT NIGHT LIGHT	27153
22492	MINI PAINT SET VINTAGE	26076
22616	PACK OF 12 LONDON TISSUES	25329

Table3.Top selling products.

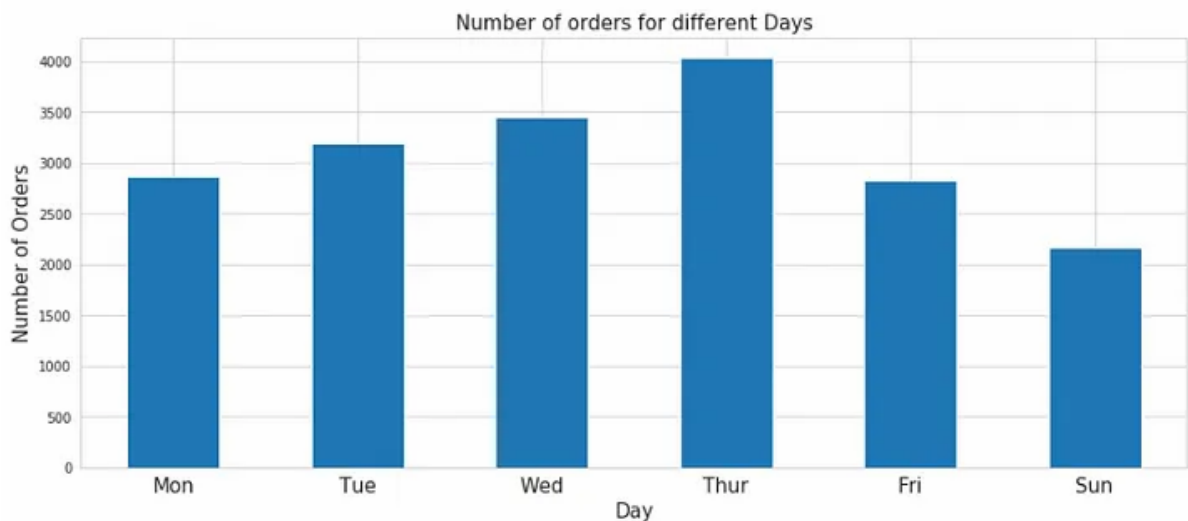
### 3.Discussion

## Sales Trends Analysis:



Graph1. Monthly sales data.

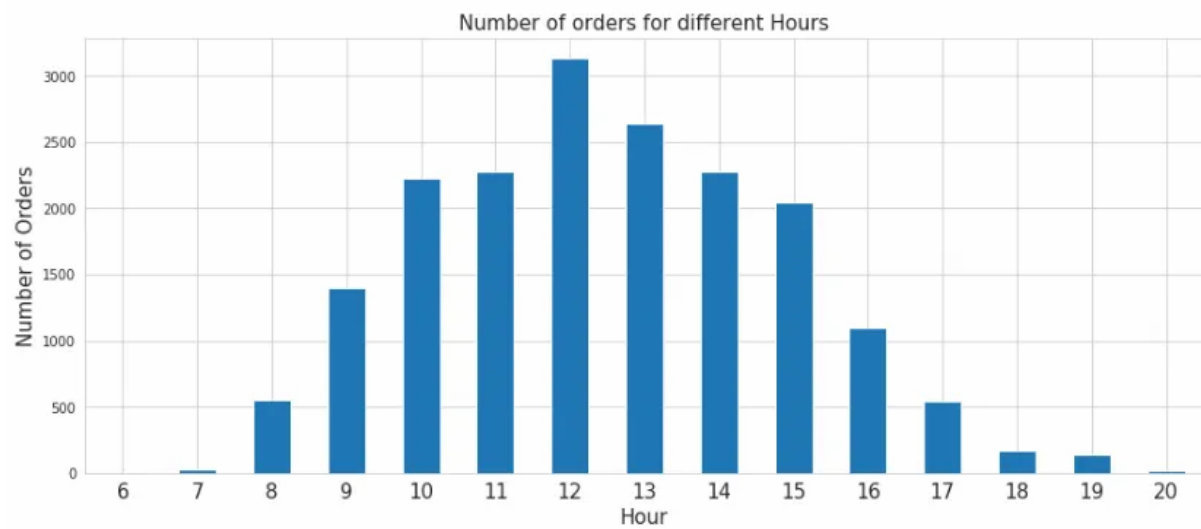
Overall, we consider that the company received the highest number of orders in November 2011 since we do not have the full month of data for December 2011.



Graph2. daily sales data.

The number of orders received by the company tends to increase from monday to thursday and decrease afterward.

Surprisingly, there are no transactions on saturday throughout the whole period (1st dec 2010 - 9th dec 2011). Reasons behind are left for discussion as the dataset and its context are limited.

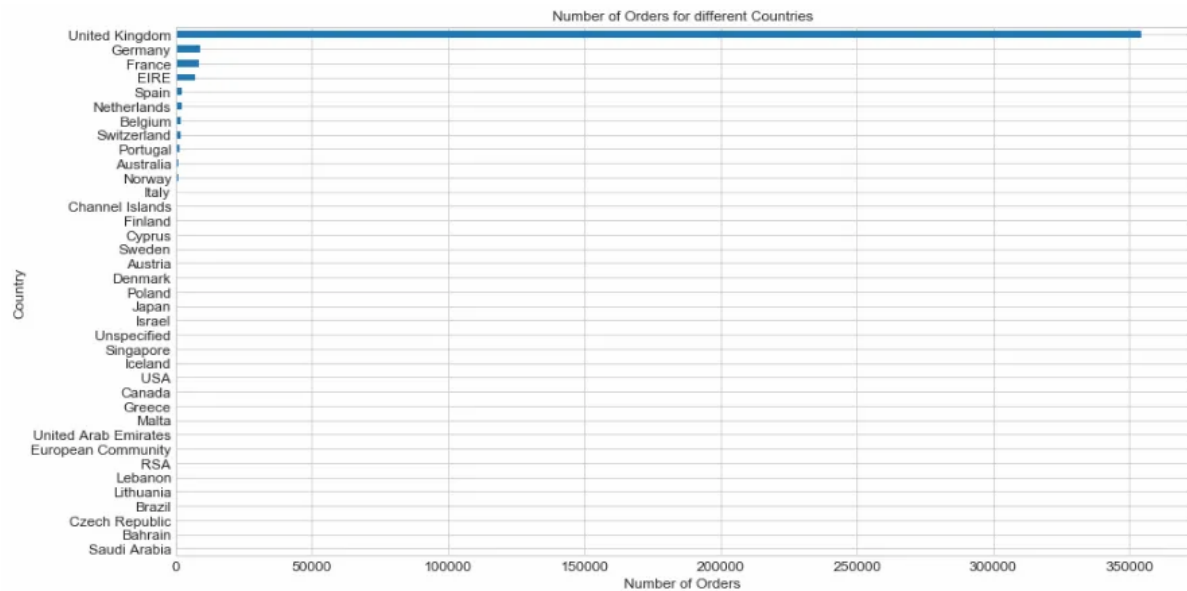


Graph3. Hourly sales data.

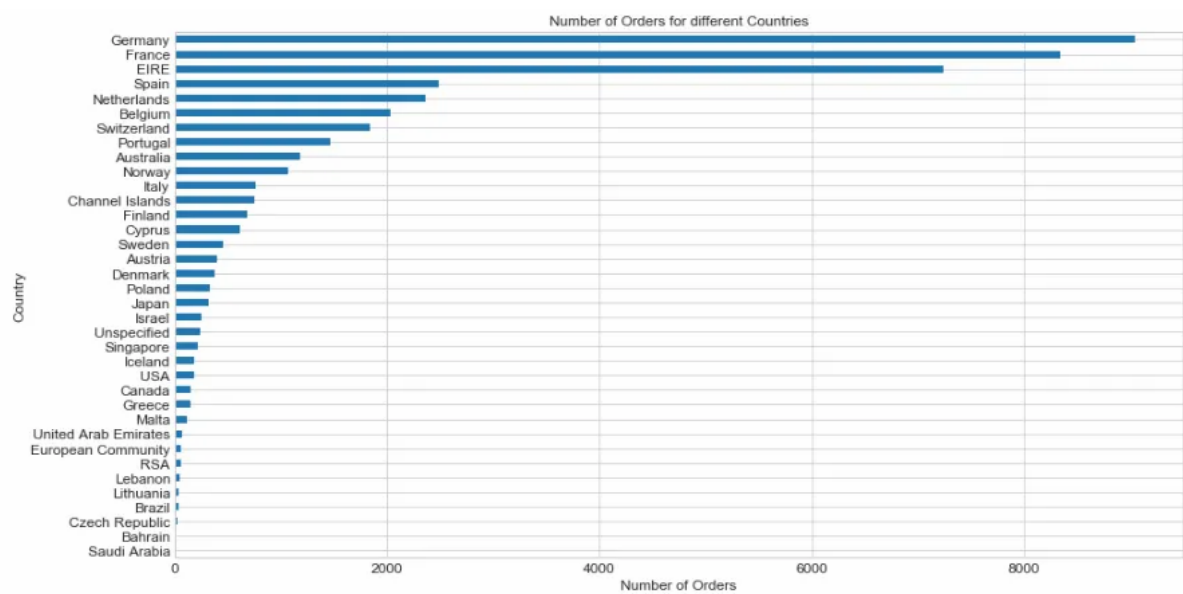
In terms of hours, there are no transactions after 8:00pm until the next day at 6:00am. Besides, we notice that the company receives the highest number of orders at 12:00pm. One of the reasons could be due to the fact that most customers make purchases during lunch hour between 12:00pm — 2:00pm.

## Transactional pattern for each country

### Countries with most number of orders



Graph4. Number of orders in each country (With UK).



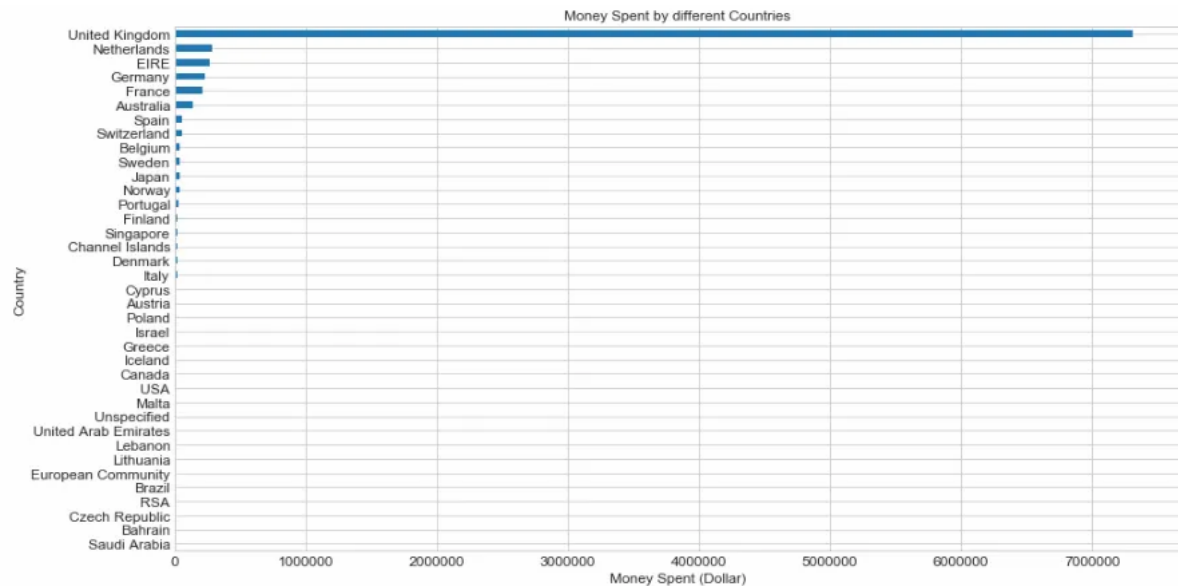
Graph5. Number of orders in each country (Without UK).

As expected, the company receives the highest number of orders in the UK since it's a UK based company. To better discern the trend, the UK is removed for clearer comparison among other countries. The top 5 countries that place the highest number of orders are as below:

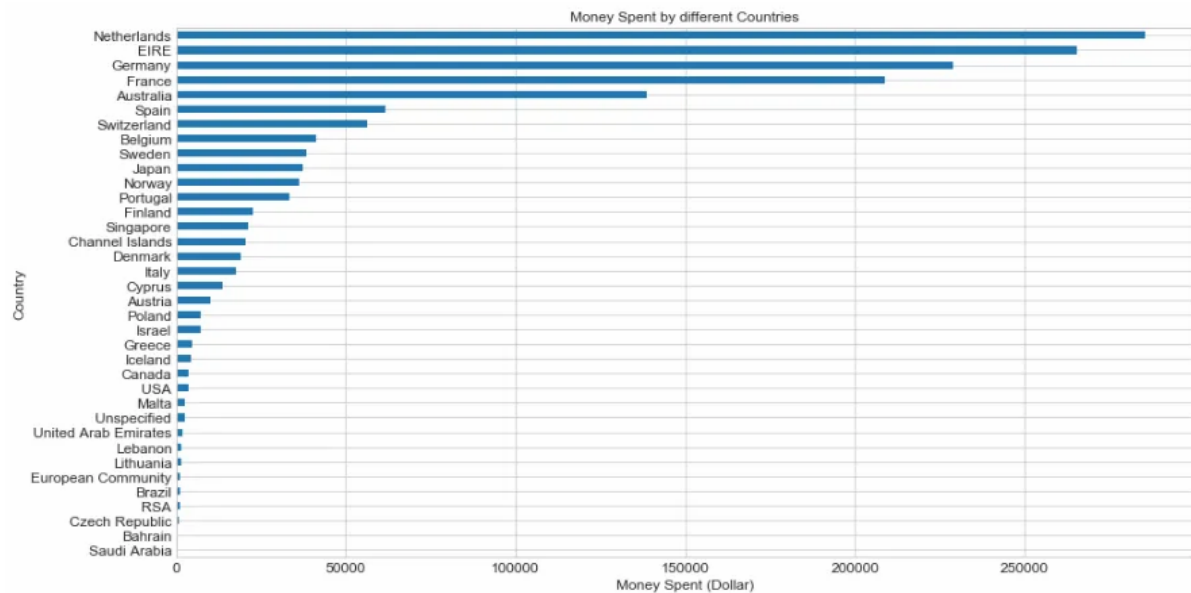
→ United kingdom

- Germany
- France
- Ireland
- Spain

## Countries with highest money spent



Graph6. Money spent by different countries (including UK)



Graph7. Money spent by different countries (excluding UK).



As the company receives the highest number of orders from customers in the UK, it is natural to see that customers in the UK spend the most on their purchases.

Same as before, the UK is removed for clearer comparison among other countries. The top 5 countries that spend the most money on purchases are as below:

- United kingdom
- Netherlands
- Ireland
- Germany
- France

## 4. Conclusion

In conclusion, the e-commerce sales analysis project has provided valuable insights into customer behavior, product popularity, and sales trends. The analysis, conducted using Apache Hive on a comprehensive sales dataset, aimed to enhance the understanding of the e-commerce business dynamics.

Through customer segmentation, we identified two distinct customer groups based on order frequency: Frequent Shoppers and Occasional Shoppers. This categorization enables the business to tailor marketing and engagement strategies to meet the varying needs of these customer segments.

The segmentation of customers assists in optimizing operations, directing resources toward retaining and engaging frequent shoppers while tailoring marketing approaches to attract occasional shoppers.

The analysis unveiled the top-selling products, shedding light on items that significantly contribute to overall sales. Understanding the popularity of specific products allows for strategic inventory management and targeted promotional efforts.

Knowledge of top-selling products guides inventory management, ensuring that popular items are adequately stocked. Additionally, it aids in product development strategies by identifying features or characteristics that resonate with customers.

Monthly sales trends revealed patterns and fluctuations in purchasing behavior over time. Recognizing these trends is essential for proactive decision-making, such as adjusting marketing campaigns or optimizing stock levels to meet demand.

Understanding monthly sales trends facilitates seasonal planning, enabling the business to anticipate peak periods and strategize marketing and inventory efforts accordingly.

## 5. Future work

Moving forward, there are several promising avenues for extending and refining the analysis of e-commerce sales data. One potential area for future exploration involves the integration of machine learning algorithms to predict customer behavior and sales trends with greater accuracy. Leveraging advanced predictive models, such as ensemble methods or deep learning architectures, could enable the development of more robust forecasting capabilities, providing the business with a proactive edge in anticipating market shifts and optimizing inventory management.

Additionally, the incorporation of external data sources, such as socio-economic indicators or industry trends, could enhance the predictive power of the models, offering a more holistic understanding of the factors influencing customer purchasing decisions.