Kristianstad University
SE-291 88 Kristianstad
+46 44-250 30 00
www.hkr.se

Big Data Analytics Project Report

Students:
Nedim Kanat & Omar Zarifa

01.01.2025

Table of Contents

1. Introduction

Data plays an important role in terms of success or failure for a business [1]. E-commerce platforms usually capture large quantities of data which may or may not include data capturing user behaviour, which would be interactions with the platform such as how often users view the cart page/section, or how often users add items to the cards and how often do users purchase items which have been added to the carts. Having insight into such data is important for businesses because it allows for optimizing experience for customers, increasing sales and profits [2].

The purpose of this project is to learn, use and understand the importance of big data tools and in this case we utilize Hadoop for storage with PySpark for processing and analyzing data to get insights from the e-commerce dataset which we have obtained from the well known online platform, kaggle. The dataset contains approximately 42 million records and includes data which captures user behavior in relation to products.

2. Dataset

2.1 Dataset Overview

The dataset, “Ecommerce Behavior Data from Multi-category Store” contains data which captures user behavior. It has been collected for over a seven month time period, mainly during the year 2019 and the data belongs to an online retail store. Each row in the dataset represents an event which captures a type of interaction between a user and a product. The events are divided into types such as “viewing”, “adding to cart” or “purchasing”. [3].

2.2 Data Preprocessing

In the preprocessing stage, data has been loaded with the use of PySpark framework which produces a DataFrame object that we are able to interact with in similar ways to how one would interact with a database.

After loading the data, we looked for missing values in order to handle them appropriately. We found that columns “brand” and “category_code” had missing values, however the amount of values missing were too large to consider dropping the corresponding rows, 13512262 missing in **category_code**, and 6112126 missing in **brand.** For that reason, we decided that it would be a better option to map null values with “unknown_brand” and “unknown_ctg” depending on column name.

Apart from that, we found that there were **30 220** duplicate rows and these were dropped since they don’t make up a significant amount of the total data.

There were also rows whose product price was equal to 0, and these were a tiny amount of outliers as we observed, **68670 rows**, making up an insignificant portion of the total data. These rows were also dropped from the dataset.

As a final step in data cleaning, we removed “category_id” and “user_session” columns because these provided no useful data. We thought that this was reasonable because numerical values make little sense to humans. Perhaps it would be useful in a scenario where the organization would benefit from knowing which user_id’s are most recurring in terms of sessions in order to perhaps provide bonuses/points, but for this project, we don’t exactly find a use-case for such data.

2.3 Data Analysis

In order to get an overview of the statistics for the price column, we made use of the **describe** method which displays a table that contains minimum, maximum, average, total and std deviation. We observed that the minimum cost that occurred in the dataset was 0.77 and maximum ~2574. By Looking at these values we came to the conclusion that there were no extreme outliers.

By using group and aggregation methods, we found that there were 742 773 “purchase” actions, 40 708 807 view actions and 898 294 cart actions which involved adding an item to the app’s virtual cart. This shows that the most recurring event type was the viewing of products which makes sense because generally users view more than they purchase or add to the cart.

As for the cart to purchase, meaning the ratio of how many items which are added to the virtual cart are purchased, we found that about 82.7% of items which were added to the cart were purchased. This gives the insight that users are highly likely to purchase items which they add to their carts.

With the use of grouping and aggregation methods, we made the observation that the total amount spent by a user was almost 2 million dollars, (1 993 636). Such users might be representing large organisations or businesses that order many products, in larger quantities than the average individual. However, the data gives no way of distinguishing between the two types of users.

Aside from that, we were able to find out which types of products were purchased most. Smartphones have a top revenue of $5.42 billion dollars, followed by notebooks having a total revenue of approximately $808 million dollars, and third being TVs having a total

revenue of approximately $491 million dollars. As for the categories which we were unable to identify, “unknown_ctg”, had a total revenue of $2.5 billion (Figure 2).

In terms of brand, Apple appears to be leading in sales: the data shows that Apple products produced a revenue of $3.43 billion, followed by Samsung with a total revenue of $1.74 billion and Xiaomi with about $617 million in total revenue. As for the unknown brands, they had in total $2.5 billion in revenue (Figure 3).

And for the final analysis, we observed that the busiest day was Tuesday, having almost 7 million visits.

3. Working Mechanism and Principle

The working mechanism for this project was divided into 6 stages, as can be seen in the following flowchart (**Figure 1**):

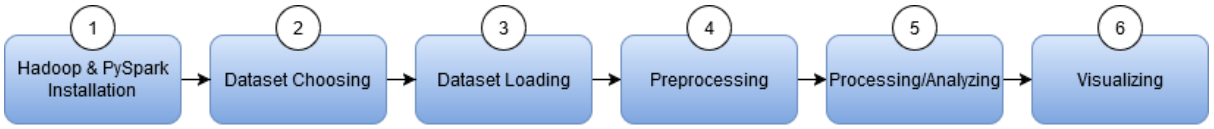

Figure 1: Project Workflow

As shown in the flowchart [Figure 1], the process began with the installation and setup/configuration of Hadoop and PySpark. After that, the dataset was selected since it had a large quantity of data to analyze user behaviour on an e-commerce platform.

Thereafter, with the use of PySpark, the dataset was loaded into the program. During the preprocessing stage; missing values, duplicates and outliers were handled in certain columns.

In the analysis phase, we focused on gaining insights into user behaviour from the dataset in order to find answers to questions such as:

- Which brands brought in the most profit for the company?
- Which users spent the most money?
- What types of products were purchased most?
- Which is the busiest day, in terms of platform traffic?

.. which involved analyzing user behavior based on event types, figuring out what types of aggregations would be necessary to approach these questions and how we could implement these in code.

The final stage involved visualizing the findings. Key insights were presented through multiple types of charts, which are identified as bar- and line charts.

Despite PySpark the optimization that PySpark framework provides, the execution of processing a large dataset such as the selected one consumed a lot of time.

4. Results and Visualization

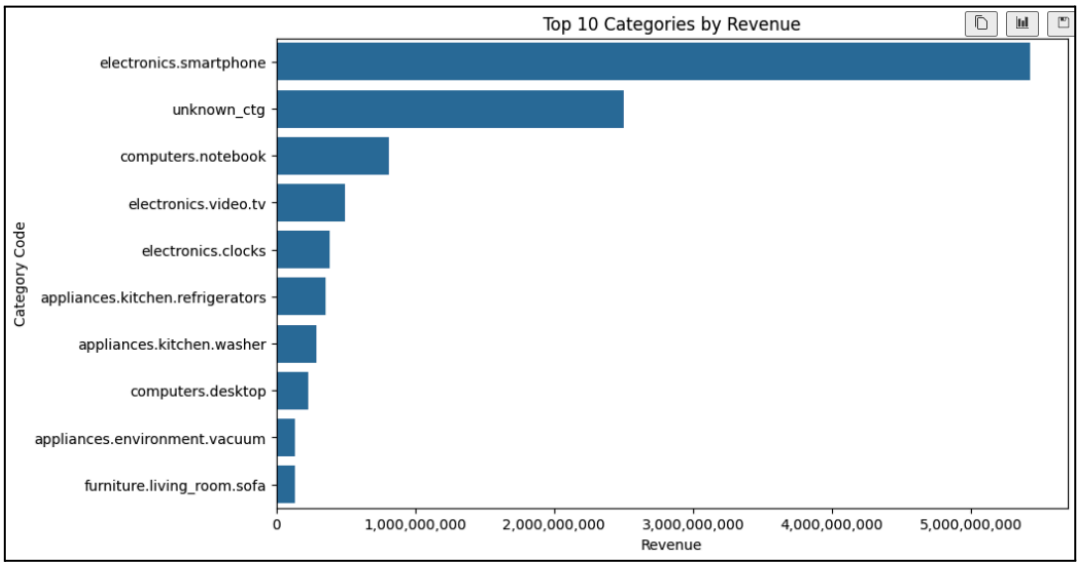

Figure 2: Top 10 categories, sorted by total revenue in descending order.

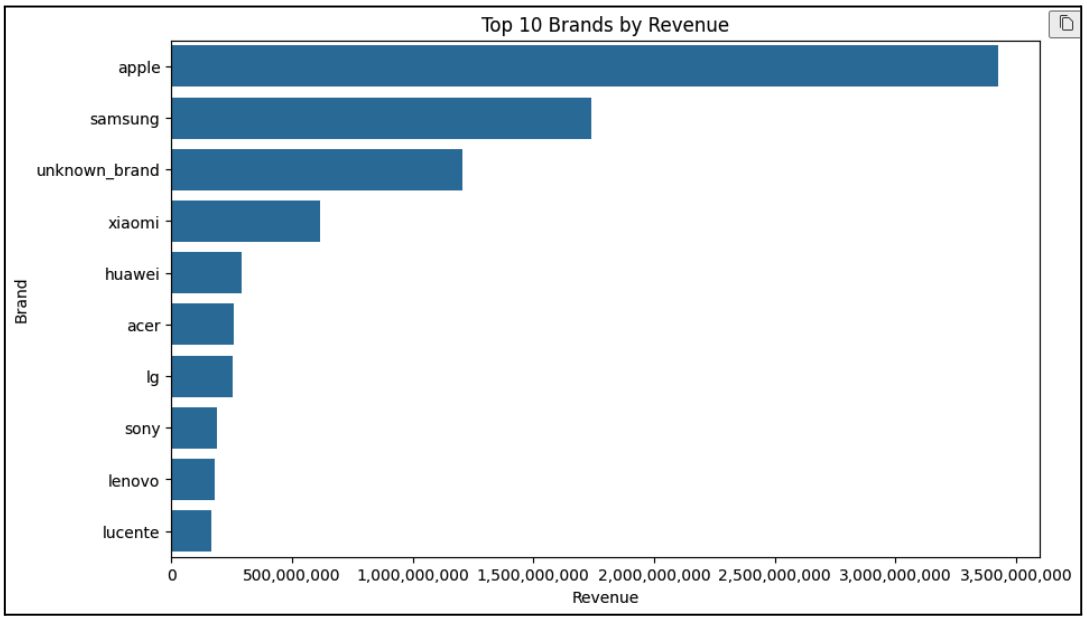

Figure 3: Top 10 brands, sorted by total revenue in descending order.

5. Evaluation and Lessons Learned

In this project, we had the opportunity to learn hadoop & pyspark utilization, preprocessing a large-scale dataset, cleaning it, analyzing interesting aspects and visualising the results. However, We faced challenges related to the installation of Hadoop and PySpark at the beginning.

Key learnings from this project include:

- Learning how to handle large-scale datasets.
- Learning how to use Hadoop and PySpark.
- Gaining practical experience in data analysis and visualization.
- Enhancing our understanding of user behavior in e-commerce settings.

6. References

[1] Kumar B, Roy S, Sinha A, Iwendi C, Strážovská Ľ. (2023). “E-Commerce Website Usability Analysis Using the Association Rule Mining and Machine Learning Algorithm”. doi: https://doi.org/10.3390/math11010025.

[2] Alrumiah S S and Hadwan M. (2021). “Implementing Big Data Analytics in E-Commerce: Vendor and Customer View,". doi: https://doi.org/10.1109/ACCESS.2021.3063615.

[3] Kechinov M. (2020). “eCommerce behavior data from multi category store”. Available from: https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store/data