**The George Washington University**

**School of Business**

## 

## 

***Ecommerce***

***ISTM 6217: Internet of Things Management***

***Spring 2023 - Prof. Anu Juneja***

May 10, 2023

## 

## 

**Group 7 :**

**SHRIYA SHARMA**

**KOMAL PATIL**

**CHRISTINE**

Table Of Contents

[**I. Executive Summary 3**](#_3c3abpci9xs5)

[**II. Data Description 3**](#_lbjnck8g1xuk)

[Ecommerce Dataset 3](#_6xuu5da7ho2b)

[**III. Research Questions 4**](#_49xbj32kedz2)

* [Project Goal: 4](#_fuccucgpqfw6)
* [Analyze the purchase trend and manage the supply chain operations to have an effective inventory with accurate predictions. 6](#_q6s33oe2108w)
* [Analyze the data and determine optimal pricing strategies for products. 7](#_d6v2l2d9n9bz)
* [Building a recommendation engine for customers based on their behavior and preferences and having targeted marketing. 8](#_yyf102gwfxjc)
* [Analyze the purchase data and to predict whether a customer will make a purchase or not. 9](#_z80kmyyxq38)
* [Analyze the suspicious activity and take prevention measures against fraudulent users. 10](#_8mtjo9qg8wae)

[**IV. Methodology 10**](#_qaa7ruipoxl4)

[1. Jupyter Notebook 10](#_8eitcxf84e6b)

[2. PySpark's DataFrame API 10](#_z42icys5un05)

[3. Logistic Regression 10](#_150uud2dm24y)

[4. Confusion Matrix 11](#_hmm1qad8r68x)

[5. Linear Regression 11](#_qr5pz8nls6lh)

[6. SQL 11](#_sp98qokph8kh)

[7. ALS(Alternating Least Squares) 11](#_fbe47ze63lc9)

[**V. Model Analysis And Result 12**](#_h2u7b92xwxh6)

* [Jupyter Notebook 12](#_9ysog4nhuxf7)
* [Trend Analysis based on Sales 12](#_y2wvhua5ermy)
* [Logistic Regression 13](#_w38jsbe2n0al)
* [Confusion Matrix 13](#_qchdryw39koo)
* [Linear Regression 13](#_ufg9pfwrluqk)
* [SQL 14](#_7m902phjs7q)
* [ALS(Alternating Least Squares) 14](#_gk4k9odfmu)
* [Suspicious Activity Analysis 15](#_uih53iqify0i)

[**VI. Conclusion 15**](#_e6h3zicl90rf)

[**VII. Future Goals 15**](#_34eop1zcekhx)

## **Executive Summary**

This project aims to analyze eCommerce behavior data from a multi-category store and provide insights for optimizing sales and revenue, building a recommendation engine for customers, analyzing supply chain operations, determining optimal pricing strategies, and detecting suspicious activity or behavior for fraud prevention.

The dataset includes information about customer activity such as browsing history, add-to-cart behavior, and purchase history, as well as product information such as category, brand, and price. This project's analysis includes a wide range of issues, such as daily sales trends, hourly total sales, category-based analysis, user analysis, optimal pricing analysis and targeted recommendation system. With the insights gained from the analysis and predictions, the online retail company may be able to improve its marketing strategy and more effectively target its audience.

This study also underlines the need of big data science and mining in utilizing the massive amount of data generated by the internet. Overall, the findings of this analysis can assist eCommerce businesses in improving their processes, increasing income, and providing a better customer experience.

## **Data Description**

### **Ecommerce Dataset**

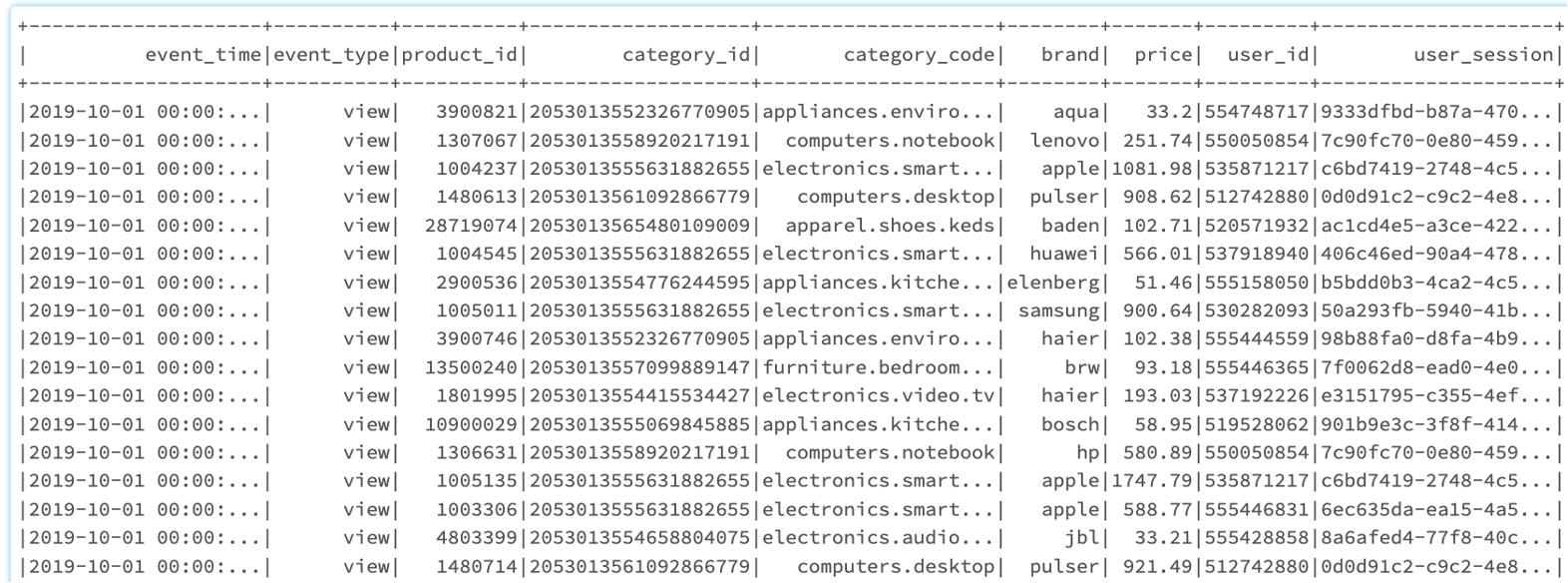
The dataset "E-Commerce Behavior Data from Multi-Category Store" includes information about the online behavior of customers on an e-commerce platform in Brazil. “[*https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store*](https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store)”. This data was collected during the period of September 2019 to August 2020. The dataset contains different CSV files containing9 columns and 22 million rows and the size of data is 15 GB. The dataset consists of various features, such as event time, event type, product ID, category ID, category code, brand, price, user ID, and user session. By analyzing these features, we can gain insights into various aspects of the e-commerce business, such as sales trends, popular categories, popular brands, and user behavior.

For this project we decided to use data from 2 CSV files of October 2019 and November 2019. The eCommerce dataset includes the following variables:

1. event\_time: The time when the event (e.g., view, cart, purchase) occurred.
2. event\_type: The type of the event, which can be "view", "cart", "purchase", or "remove\_from\_cart".
3. product\_id: The ID of the product.
4. category\_id: The ID of the category that the product belongs to.
5. category\_code: The code of the category that the product belongs to.
6. brand: The brand of the product.
7. price: The price of the product.
8. user\_id: The ID of the user.
9. user\_session: The session ID of the user.

In this project, we will use various data analysis techniques and tools such as Jupyter Notebook, pySpark, and SQL to analyze the dataset and derive meaningful insights. The analysis will be presented in the form of visualizations such as graphs, charts, and tables to provide a better understanding of the results. Overall, this project aims to provide valuable insights into e-commerce behavior and to improve business operations, increase revenue, optimize prices and provide a better customer experience.

*Summary of e-Commerce Data*



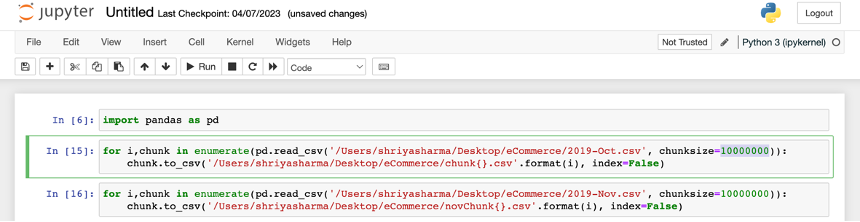
## **Research Questions**

### **Project Goal:**

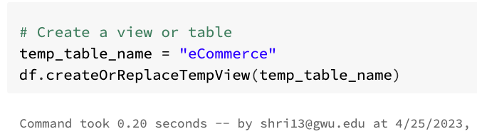
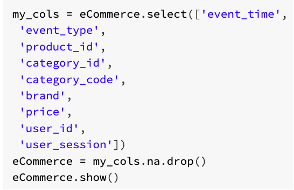
Our eCommerce business analysis aims to optimize sales and revenue by implementing a recommendation engine based on customer behavior. To have efficient supply chain operations. Also check with the inventory management, data analysis for pricing strategies, and fraud prevention measures through dataset analysis.

To choose our research questions we first started with data cleaning and preprocessing. We could understand that the size of the dataset is significant, and it includes a diverse range of data types, including customer behavior, product information, and time data. As a result, it may pose challenges in terms of data processing and analysis. The dataset was incomplete, it had inaccuracies, which might affect the reliability and precision of our analysis or modeling. As it is 15 GB data, integrating and combining diverse data sources was challenging. Data mapping and transformation was one of the major challenges as we had more timestamps and string data columns.

* Utilized python to split the data of 2 months into 12 equal parts with 10,000,000 lines and 1.33 GB each.
* We combined all the 12 parts using the “union” function.
* Removed all the rows with Null values.
* Created a temporary view table for working on SQL Queries.
* Added and removed columns as required for analysis.







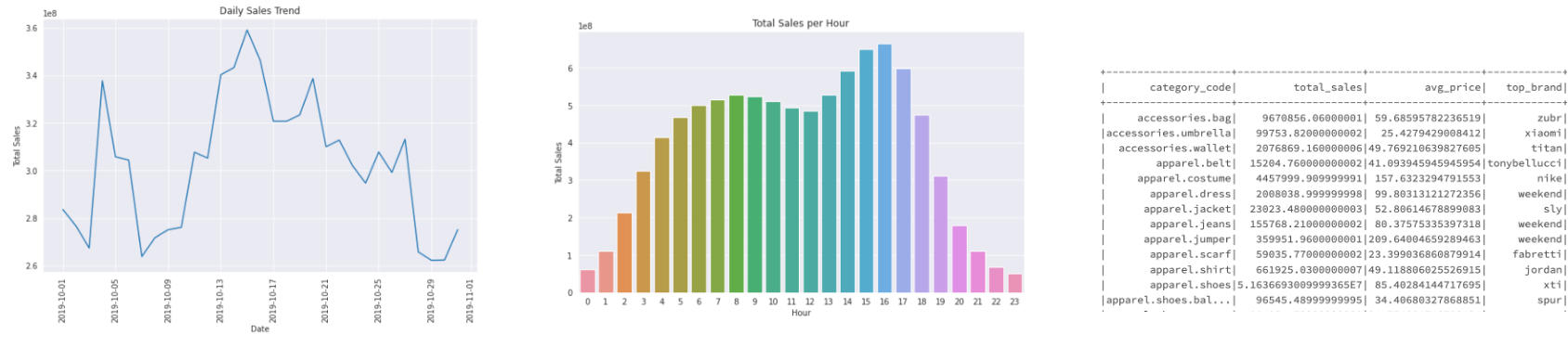
### **Analyze the purchase trend and manage the supply chain operations to have an effective inventory with accurate predictions.**

We conducted a number of trend studies on our data to get insights about consumer behavior and product popularity. First, we used a line chart to assess the daily sales trend. Using PySpark's "to\_date" function, we changed the "event\_time" column into a "DateType" column, grouped the data by date, and determined the total sales for each day. We next changed the resulting PySpark DataFrame into a Pandas DataFrame and used Seaborn's "line plot" function to visualize the daily sales trend. According to our study, sales reached their high between 10/13/2019 and 10/17/2019 with a total of $36,000. According to our study, sales are at their strongest between 2 pm to 5 pm, peaking at 4 pm. Sales can go up and customer service can grow better by scheduling promotions during busy times and staffing accordingly.

To find out which product categories were most popular, we then conducted a category-based analysis. By using the "category\_code" to group the data, we were able to get the total sales and average price for each category. According to our data, the "apparel.jumper" category had the highest average price, at $209, while the "accessories.bag" category had the most sales, of $9670856.

In order to identify which categories and brands were most well-liked by our customers, we studied customer interactions and brand perceptions. We sorted the resulting DataFrame by interactions in descending order after grouping the data by "category\_code" and counting the quantity of interactions for each category. With a total of 25,000 interactions, our study showed that the category "electronics.smartphone" had the most interactions. Additionally, we divided the data into "brand" groups and tallied the number of views for each brand before filtering the data by "view" event type. We then sorted the resulting DataFrame by views in descending order. According to our data, "samsung" is the most viewed brand.

Overall, our trend research offered insightful information about consumer preferences and product popularity that might guide future marketing and sales plans.



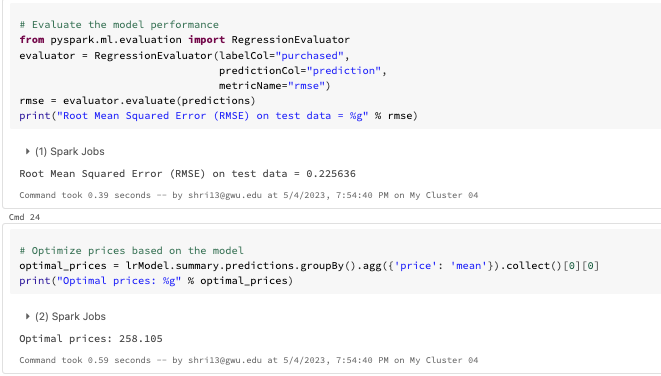


## **Analyze the data and determine optimal pricing strategies for products.**

For price optimization we can use various techniques such as linear regression, decision trees, or neural networks. Here we are using linear regression to optimize prices based on the features that are available in our dataset.

Linear regression is a statistical technique used to describe the relationship between a dependent variable (in this case, price) and one or more independent variables (such as product attributes, rival prices, and so on). Linear regression determines the best fit line which indicates the connection between the dependent variable and the independent variable(s). For price optimization, linear regression was used to identify the factors that have the greatest impact on price and to determine the optimal price for a given product. The independent variables could include product features, competitor prices, customer demographics, etc. The dependent variable would be the price of the product.

This analysis was utilized to predict the best price for a new product based on its specifications and other pertinent criteria. Therefore, in our study we are using this strategy to optimize prices in order to maximize earnings and acquire a competitive advantage in the market.

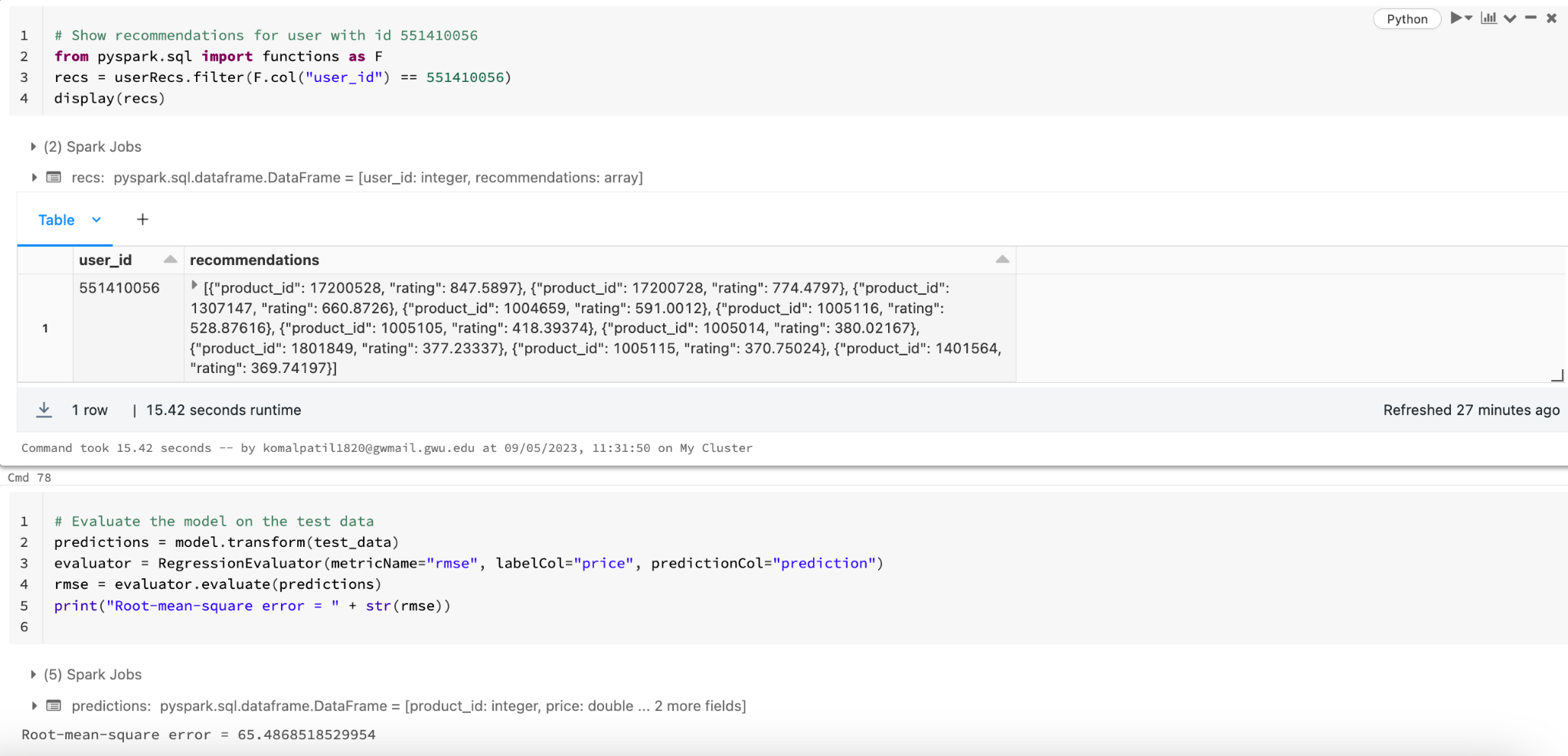


### **Building a recommendation engine for customers based on their behavior and preferences and having targeted marketing.**

The Alternating Least Squares (ALS) algorithm was used in the recommendation system's development to forecast which products a user is likely to buy based on their prior behavior and preferences. This analysis was made using PySpark, for processing massive datasets, and it was trained using data on user interactions with products.

To assess the effectiveness of the recommendation system, the dataset was separated into training and testing datasets after being preprocessed to eliminate any invalid or missing data. The training dataset was utilized to train the model using the ALS algorithm, and parameters like rank, maxIter, and regParam were changed to enhance the model's performance. Based on each user's prior encounters with products, suggestions were then generated for them using the model. The user was shown the top recommendations in order of anticipated likelihood of purchase, which were rated recommendations. Using the root-mean-square error (RMSE) metric, which gauges the discrepancy between anticipated and actual ratings, the accuracy of the recommendation system was assessed. Several methods, including hyperparameter tweaking, adding more characteristics to the model, and experimenting with other recommendation algorithms, such as user-based or item-based collaborative filtering, were used to enhance the performance of the recommendation system.

We assume that the recommendation system performed well as it is recommending products based on the purchase price history of the user, and RMSE value is calculated based on user price preferences. If we had a column for rating/review available, then RMSE would have been more appropriate. By implementing more sophisticated approaches like deep learning or reinforcement learning, the system can be made even better.



### **Analyze the purchase data and to predict whether a customer will make a purchase or not.**

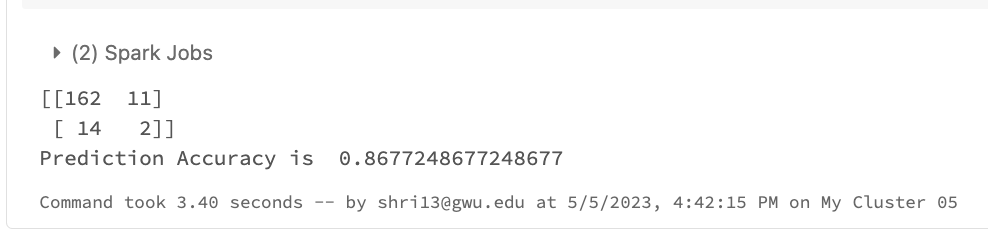
As we had learned about the Logistic Regression analysis in our course, we wanted to try this regression model in our eCommerce dataset to see if a customer would make a purchase or not based on the “event\_type” column. In this analysis we created a new column named “purchased”. This column consisted of binary values and the query is as mentioned below.

SELECT \*

FROM df\_reg\_final\_data

WHERE (event\_type = 'purchase' OR event\_type = 'cart') AND purchased =1

OR event\_type = 'view' AND purchased = 0



The results of the same can be found in the Model Analysis and Results section.

### **Analyze the suspicious activity and take prevention measures against fraudulent users.**

In this question we tracked the users who make purchases that are much larger or smaller than the typical purchase, or who make purchases at unusual times based on the user session information. We also listed users who made zero purchases and had long user sessions.

## **Methodology**

### **Jupyter Notebook**

We used the Jupyter Notebook to divide the dataset into small chunks to upload to DataBricks. We used the pandas library in Jupyter Notebook to read the original dataset and split it into multiple smaller dataframes then saved these dataframes as separate CSV files. This code reads the original dataset and splits it into smaller chunks of 100,000 rows each. It then saves each chunk as a separate CSV file with a filename in the format "chunk\_i.csv", where i is the index of the chunk. We adjusted the chunk size as needed based on the size of your dataset and for our analysis model.

### **PySpark's DataFrame API**

In-depth data transformation and manipulation operations, like filtering, grouping, aggregating, and joining, were carried out using PySpark's DataFrame API. Additionally, we made use of built-in PySpark functions like hour() and to\_date() to extract the hour from the "event\_time" column and change the "event\_time" column to a DateType column, respectively.We plotted various graphs and charts using the Matplotlib and Seaborn libraries for data visualization. In order to assess the daily sales trend and the total sales per hour, we created line charts and bar charts, respectively. To make the charts more illuminating and aesthetically pleasing, we made use of the different modification options offered by these libraries.

### **Logistic Regression**

We used Logistic regression to predict whether a customer will make a purchase or not based on their behavior and other relevant factors. We split the dataset into 0.7 and 0.3 randomly to create the training and test data. We considered seed = 123. For this analysis we created a new column named “Purchased” and then stored the binary values based on column “event type” = "view" the Purchased column is "0" when “event type” = "purchased" and "cart" then Purchased column is "1". Our model gave Prediction Accuracy of 86.7% on the customer purchase analysis.

### **Confusion Matrix**

It summarizes the counts of true positives, false positives, true negatives, and false negatives produced by applying the model to a set of test data with known true labels. In the matrix, each row represents examples in an actual class, whereas each column represents instances in a forecast class. The correctly identified cases are represented by the diagonal of the matrix, whereas the wrongly classified occurrences are represented by the off diagonal. We can calculate measures like accuracy, precision, and recall by evaluating the confusion matrix.

### **Linear Regression**

The goal of linear regression analysis is to find the best-fit line that describes how to predict the price of a product. An RMSE value of 0.2256 is a relatively low error, indicating that the model is performing well in predicting the prices of products. With predictions of this model, we tried to find the optimal price for this chunk of dataset to be $262.61.

### **SQL**

We used SQL queries to create a new column “Purchased” and to create a temporary table for viewing the data. Later we also used SQL for knowing the optimal pricing for each brand and to know which brand has the least priced item.

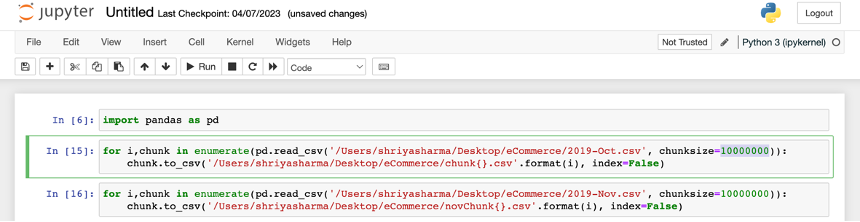
### **ALS (Alternating Least Squares)**

A frequent collaborative filtering technique in recommendation systems is the ALS (Alternating Least Squares) model. It is based on the matrix factorization method, which divides a big user-item interaction matrix into two more manageable matrices—one for users and the other for items—to better represent the user and item perspectives.

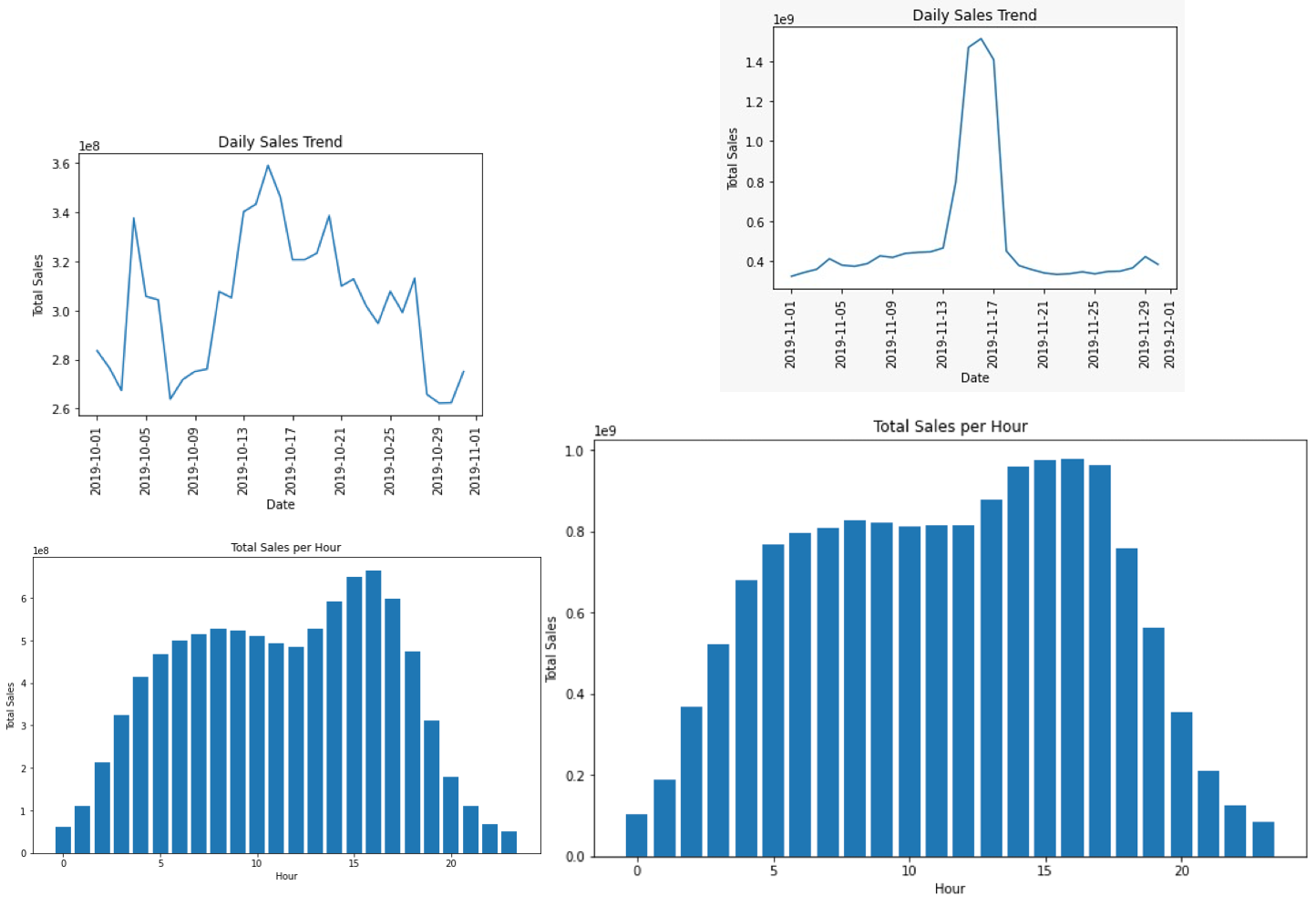
The ALS model was utilized in our project with the hyperparameters rank=10, maxIter=10, and regParam=0.1. We chose the rank parameter to be 10, which specifies how many latent factors to employ in the matrix factorization. As a reasonable compromise between speed and calculation time, we selected 10 for the maxIter parameter—the ALS algorithm's maximum number of iterations. Our investigation led us to set the regParam parameter, which acts as a regularization parameter to avoid overfitting, to 0.1. Our recommendation model gave rmse value as 65.48.

## Model Analysis and Result

### **Jupyter Notebook**



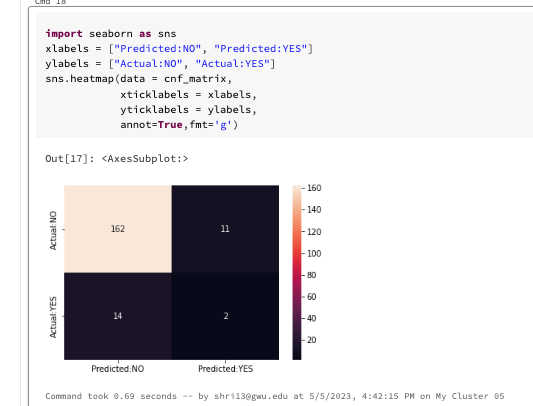
### **Trend Analysis based on Sales**



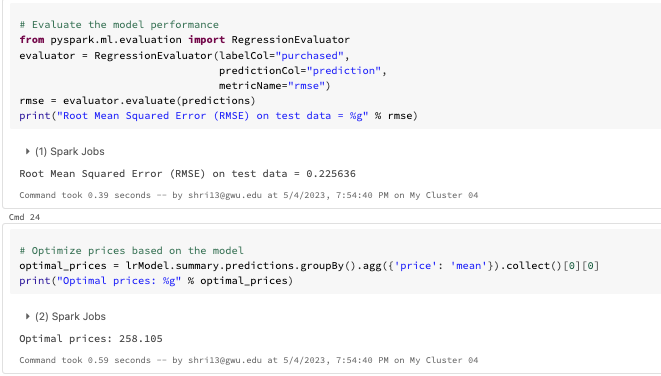
### **Logistic Regression**

### 

### **Confusion Matrix**



### **Linear Regression**



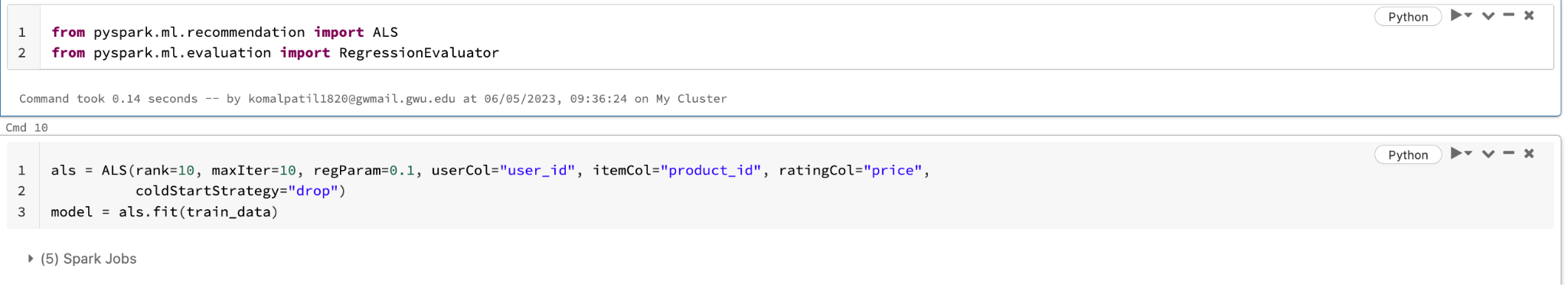


### **SQL**

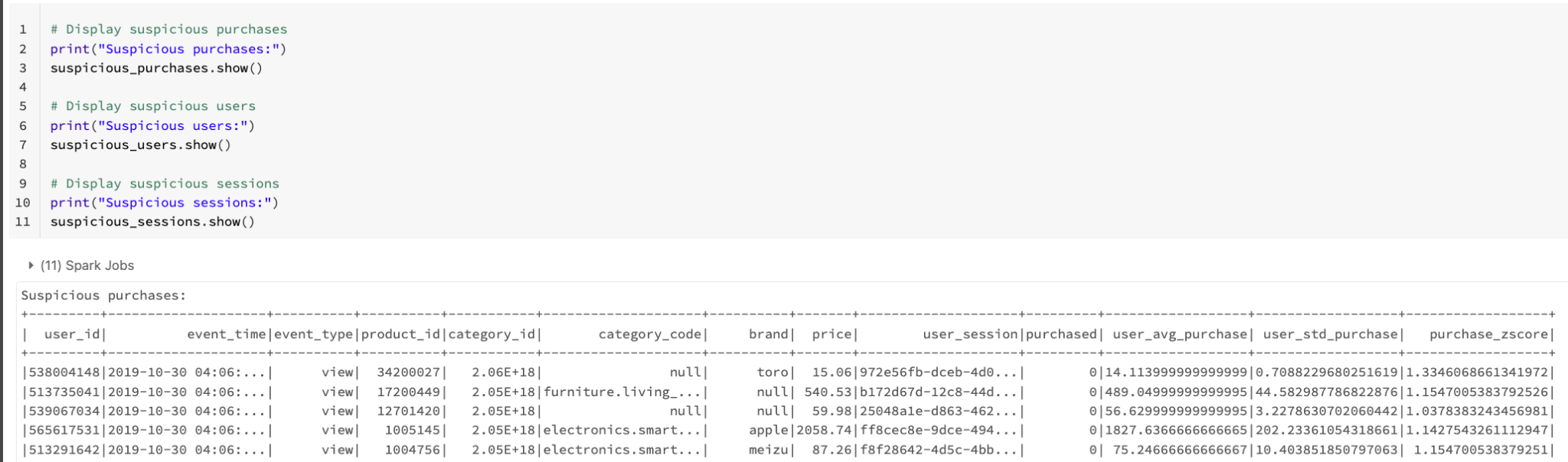
#### 

#### 

### **ALS(Alternating Least Squares)**



### **Suspicious Activity Analysis**



## **Conclusion**

The analysis covers a wide range of topics, including daily sales trends, hourly total sales, category-based analysis, user analysis, and market analysis. These insights are used to optimize sales and revenue, build a recommendation engine for customers, analyze supply chain operations, determine optimal pricing strategies, and try to detect suspicious activity or behavior for fraud prevention.

The study emphasizes the significance of big data science and mining in making use of the huge amount of data. The data and projections would help eCommerce organizations improve their processes, increase revenue, and provide a better customer experience. The linear regression model for price optimization's RMSE score of 0.2256 suggests that the model is relatively accurate in predicting prices based on the input variables.On the other hand the logistic regression analysis predicts the purchase analysis with 86.7% accuracy. The Confusion Matrix helps us to predict the True Negatives and positives & False Negative and Positives with the help of a heat map which looks good.The Recommendation system also helps in understanding he user purchase price range to recommend products in that range and save time, make better decisions and improve overall user experience.

Overall, the project provides valuable insights for businesses in the eCommerce industry to improve their operations, marketing strategies, and overall performance by leveraging the power of big data analytics.

## **Future Goals**

Our project's future goal is assessing how recommendations affect user engagement, putting a feedback loop in place, utilizing A/B testing to compare various recommendation approaches, and expanding the system to handle additional kinds of data. By fulfilling these objectives, we hope to create a strong recommendation system that can improve user experience, stop fraud, and increase business revenue.

We also intend to expand our study by including more information from sources including customer demographics, location, and product reviews. We also plan to research cutting-edge machine learning techniques like clustering and classification in order to discover client categories and predict future sales patterns. Additionally, in order to make our findings more understandable and practical for business stakeholders, we plan to develop interactive dashboards and visualizations. Finally, we will keep an eye on trends and performance metrics to make sure that our suggestions hold up over time. We have listed these promising future goals and are determined to fulfill them as shopping is what we girls like the most in our free time, why not make it a good experience.