Retail Sales Data Analysis

ECE 552- Big Data Technology

Wenzhe Luo

Askash Aundhkar

Team 18

Spring 2023

George Mason University

Dataset source comes from:
https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

Table of Contents

Dataset source comes from:
https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

Abstract

The epidemic in the past three years has brought many economic blows to many countries. Especially for the retail industry, many shopping malls are facing the impact of short-term revenue and profit setbacks. But with the publicity of vaccinations and people traveling and shopping again, business is expected to explode. As a bridge city in Eurasia, Istanbul attracts many tourists to stop and travel. Our project is to analyze the purchase records of customers in the top ten shopping malls in Istanbul from 2021 to 2023, analyze the shopping situation in Istanbul in the past three years, and understand the differences between men and women in shopping, and realize the future through MLlib compatible with Spark Prediction of shopping trends for major commodities. For this method, we will use Zeppelin as the platform, and use some libraries in the PySpark API, such as PySpark SQL, PySpark DataFrame, to do some analysis using the Spark-specific language. After analysis, we found that women shop the most, and clothing is the mainstream shopping trend. Our project has more analysis results, and we want to provide these analysis results to merchants in Istanbul or tourists who love shopping around, so that they can refer to them.

Key words: Virtual Machine, Microsoft Azure, Zeppelin, Apache Spark, MLlib

Introduction

Background

The history of human shopping can be traced back to the use of shells by primitive tribes to exchange items needed by both parties. Istanbul is a place where the ancient Silk Road passed, and it is also a necessary place for the railway network between Europe and the Middle East, and the sea route between the Black Sea and the Mediterranean Sea, which makes Istanbul's strategic position very important, and thus nurtures an eclectic population and culture. It is located on the coast of the Bosphorus in northwestern Turkey, between the Sea of Marmara and the Black Sea, across the Eurasian continent, with its economic and historical center on the European side, and one-third of the population living in Asia side. With a population of 14.4 million, it is the largest urban agglomeration in Europe.

Shopping is a common behavior in human society, which has many meanings for people. There are many large malls in Istanbul, which are an important part of the city's economy and a place for shopping and entertainment for many people. The pandemic has had a profound impact on these malls; however, many malls have taken proactive measures such as offering online shopping services, enhanced safety and

Dataset source comes from:
https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

sanitation measures, and negotiating lease agreements with tenants to meet these challenges and keep malls operating.

Data analysis is of great significance in the statistics of customer shopping records in shopping malls. Shopping malls can use data analysis to understand customer needs, predict sales trends, optimize operation management, and improve customer satisfaction, thereby increasing sales efficiency and profits. Data analysis can help shopping malls formulate better product and service strategies, recommend relevant products, monitor employee performance, improve shopping environment and experience, and increase customer loyalty and brand recognition.

It is a very good plan to use Spark to analyze the customer consumption records of Istanbul's top ten shopping malls. Spark is a fast, general-purpose, distributed computing engine that can process large-scale data and supports various data sources and data processing tasks. By using Spark for data analysis, data can be processed faster, and more accurate analysis results can be obtained. At the same time, Spark also supports distributed computing, which can perform data analysis on multiple computing nodes at the same time, thereby improving computing efficiency. Ultimately, by analyzing the consumption records of the mall, we can better understand customer needs and preferences, predict sales trends, optimize operation management, and improve customer satisfaction, thereby improving the sales efficiency and profits of the mall.

Project Objective

The purpose of this research is to apply big data techniques to analyze real datasets existing on Kaggle, to demonstrate the flexible application of the knowledge learned in this semester and to overcome the challenges associated with processing large datasets. The purpose of this study was to identify the following three factors:

1. **Comparing the two groups of men and women, which type of product is more purchased by people?**
2. **Between 2021 and 2023, which quarter of the year is the shopping season?**

To analyze whether a quarter is a shopping season, the following indicators usually need to be considered:

Sales: The sales in a quarter are an important indicator to measure whether the quarter is a shopping season. If sales increase significantly in the quarter, it may indicate that the quarter is a shopping season.

Customer unit price: The customer unit price refers to the average consumption amount of each customer. If the unit price per customer increases significantly in this quarter, it may indicate that this quarter is a shopping season.

Purchases: Purchases are the average number of purchases per customer during the quarter. If the number of purchases increases significantly in that quarter, it may indicate that it is a shopping season.

Dataset source comes from:
https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

Product sales ratio: The sales ratio of different products in a quarter can also be used to judge whether the quarter is a shopping season. If the percentage of sales of seasonal products (such as holiday gifts, seasonal clothing, etc.) increases significantly during the quarter, it may indicate that it is a shopping season.

Number of customers: If there is a significant increase in the number of customers during the quarter, it may indicate that it is a shopping season. The increase in the number of customers may be related to factors such as festivals and promotions.

3. **Which payment method is the most popular?**

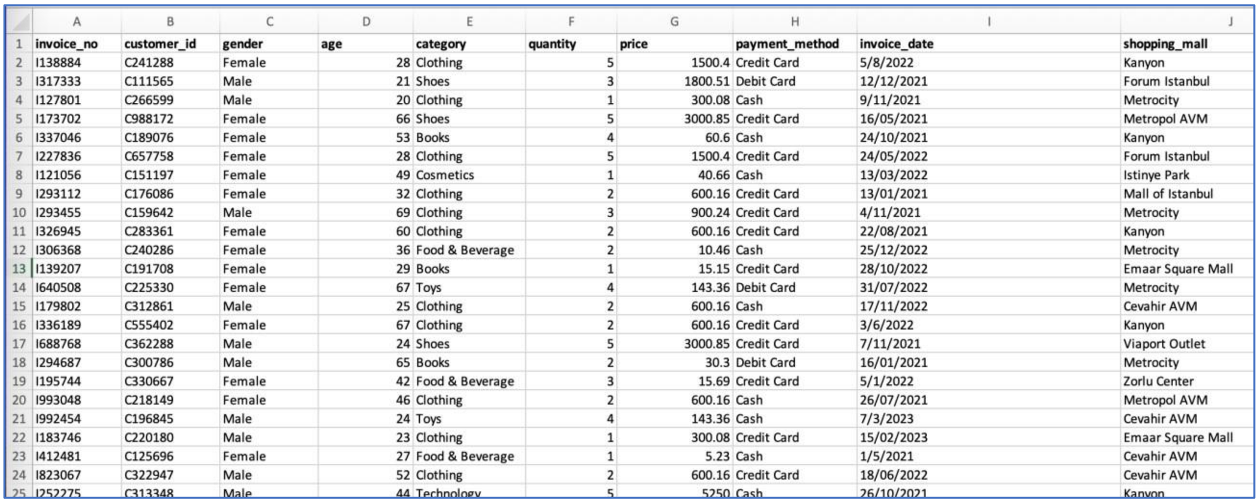
Select the payment method used by the most people and analyze the reasons to understand the range of the payment amount under this type of consumption payment mode?

4. **Through the overall data analysis, use spark's unique ML library for data cleaning, training, and testing, and finally predict the customer consumption level of major shopping malls in the next few years.**

With the above research objects, we objectively make assumptions based on the above factors.

Data Source

The dataset we collect from Kaggle [1] contains shopping information from 10 different shopping malls between 2021 and 2023. The author has gathered data from various age groups and genders to provide a comprehensive view of shopping habits in Istanbul (Contains 99,457 rows * 10 columns). The dataset includes essential information such as invoice numbers, customer IDs, age, gender, payment methods, product categories, quantity, price, order dates, and shopping mall locations. Figure 1 shows shows a screenshot of a very small portion of the dataset.



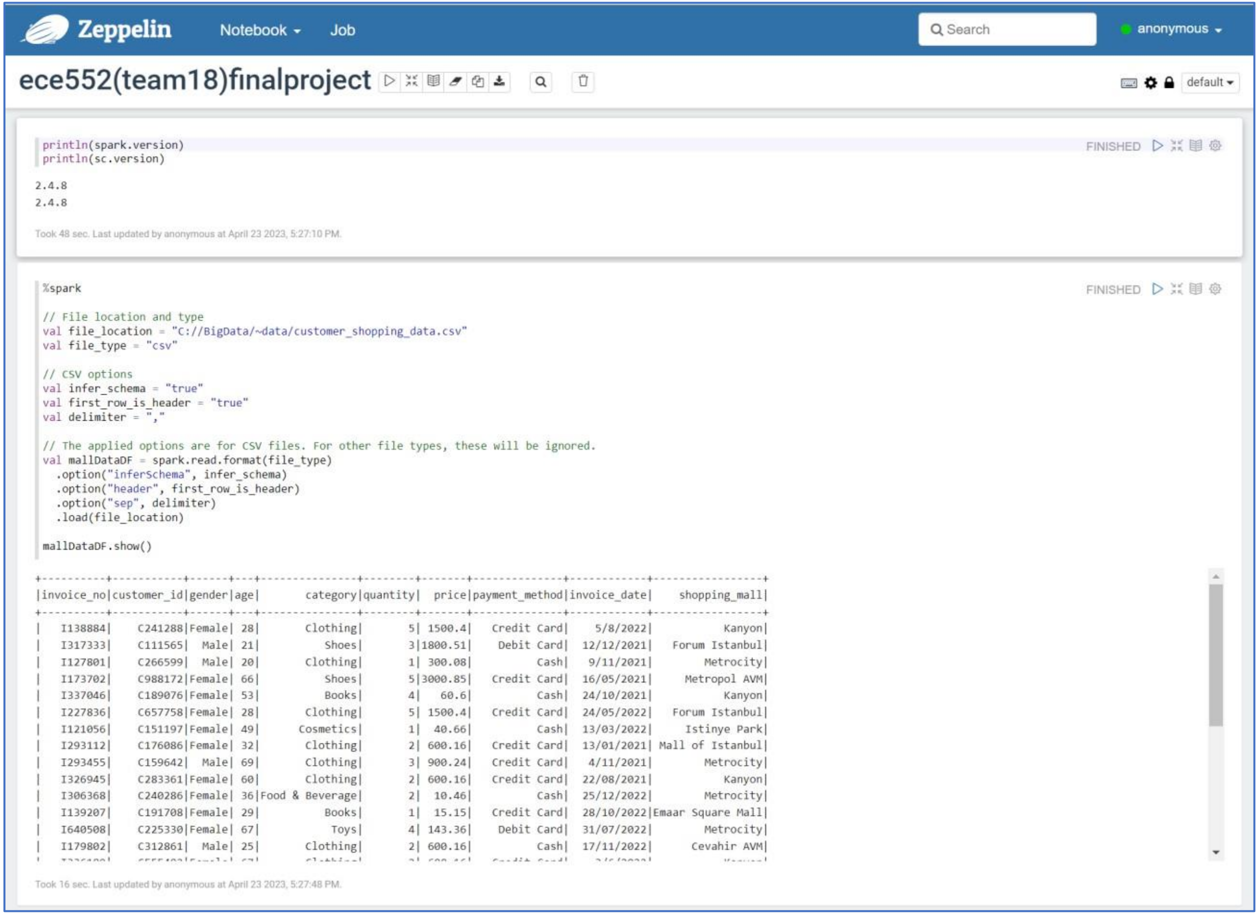
	A	B	C	D	E	F	G	H	I	J
1	invoice_no	customer_id	gender	age	category	quantity	price	payment_method	invoice_date	shopping_mall
2	I138884	C241288	Female	28	Clothing	5	1500.4	Credit Card	5/8/2022	Kanyon
3	I317333	C111565	Male	21	Shoes	3	1800.51	Debit Card	12/12/2021	Forum Istanbul
4	I127801	C266599	Male	20	Clothing	1	300.08	Cash	9/11/2021	Metrocity
5	I173702	C988172	Female	66	Shoes	5	3000.85	Credit Card	16/05/2021	Metropol AVM
6	I337046	C189076	Female	53	Books	4	60.6	Cash	24/10/2021	Kanyon
7	I227836	C657758	Female	28	Clothing	5	1500.4	Credit Card	24/05/2022	Forum Istanbul
8	I121056	C151197	Female	49	Cosmetics	1	40.66	Cash	13/03/2022	Istinye Park
9	I293112	C176086	Female	32	Clothing	2	600.16	Credit Card	13/01/2021	Mall of Istanbul
10	I293455	C159642	Male	69	Clothing	3	900.24	Credit Card	4/11/2021	Metrocity
11	I326945	C283361	Female	60	Clothing	2	600.16	Credit Card	22/08/2021	Kanyon
12	I306368	C240286	Female	36	Food & Beverage	2	10.46	Cash	25/12/2022	Metrocity
13	I139207	C191708	Female	29	Books	1	15.15	Credit Card	28/10/2022	Emaar Square Mall
14	I640508	C225330	Female	67	Toys	4	143.36	Debit Card	31/07/2022	Metrocity
15	I179802	C312861	Male	25	Clothing	2	600.16	Cash	17/11/2022	Cevahir AVM
16	I336189	C555402	Female	67	Clothing	2	600.16	Credit Card	3/6/2022	Kanyon
17	I688768	C362288	Male	24	Shoes	5	3000.85	Credit Card	7/11/2021	Viaport Outlet
18	I294687	C300786	Male	65	Books	2	30.3	Debit Card	16/01/2021	Metrocity
19	I195744	C330667	Female	42	Food & Beverage	3	15.69	Credit Card	5/1/2022	Zorlu Center
20	I993048	C218149	Female	46	Clothing	2	600.16	Cash	26/07/2021	Metropol AVM
21	I992454	C196845	Male	24	Toys	4	143.36	Cash	7/3/2023	Cevahir AVM
22	I183746	C220180	Male	23	Clothing	1	300.08	Credit Card	15/02/2023	Emaar Square Mall
23	I412481	C125696	Female	27	Food & Beverage	1	5.23	Cash	1/5/2021	Cevahir AVM
24	I823067	C322947	Male	52	Clothing	2	600.16	Credit Card	18/06/2022	Cevahir AVM
25	I252275	C313348	Male	44	Technology	5	5250	Cash	26/10/2021	Kanyon

Figure 1. Portion of the dataset (Customer shopping Retail Sales Data)

Dataset source comes from:
https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

Data Preparation

In our research, we used data downloaded from Kaggle for analysis. We downloaded the "*customer_shopping_data.csv*" dataset from Kaggle. In order to better analyze the shopping trends of the male group and the female group, when analyzing this task, we use Spark.SQL to divide the data into 'male' and 'female' at the initial stage. In order to complete the first task mentioned above more clearly. As you can see in Figure 2 below, we use Spark Dataframes to clean our dataset for our analysis requirements. We are loading CSV (.csv) file into Apache Spark Dataframe on the Zeppelin platform. Figure 3 shows the processing code to get statistics of the overall Data. After whole Dataframe preparation. We create temporary view in order to perform Spark.SQL on Data. As Figure 4 shows.



```
println(spark.version)
println(sc.version)
```

```
2.4.8
2.4.8
```

Took 48 sec. Last updated by anonymous at April 23 2023, 5:27:10 PM.

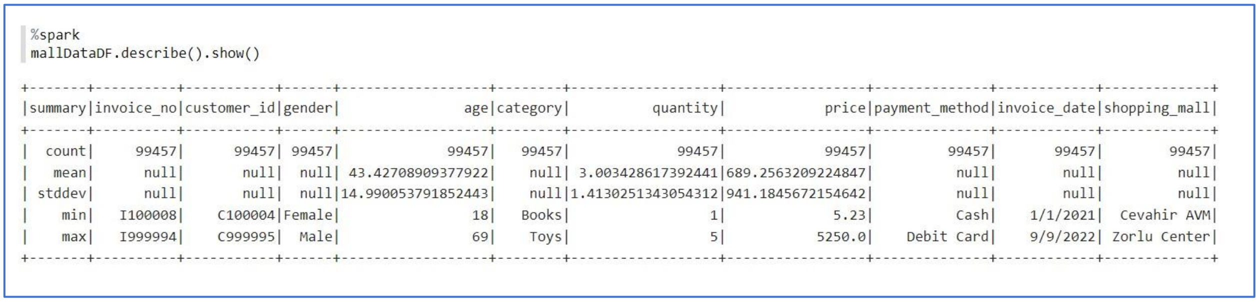
```
%spark

// File location and type
val file_location = "C://BigData/~data/customer_shopping_data.csv"
val file_type = "csv"

// CSV options
val infer_schema = "true"
val first_row_is_header = "true"
val delimiter = ","

// The applied options are for CSV files. For other file types, these will be ignored.
val mallDataDF = spark.read.format(file_type)
  .option("inferSchema", infer_schema)
  .option("header", first_row_is_header)
  .option("sep", delimiter)
  .load(file_location)

mallDataDF.show()
```

```
+----------+-----------+------+---+---------------+--------+-------+--------------+------------+-----------------+
|invoice_no|customer_id|gender|age|       category|quantity|  price|payment_method|invoice_date|    shopping_mall|
+----------+-----------+------+---+---------------+--------+-------+--------------+------------+-----------------+
|   I138884|    C241288|Female| 28|       Clothing|       5| 1500.4|   Credit Card|    5/8/2022|           Kanyon|
|   I317333|    C111565|  Male| 21|          Shoes|       3|1800.51|    Debit Card|  12/12/2021|   Forum Istanbul|
|   I127801|    C266599|  Male| 20|       Clothing|       1| 300.08|          Cash|   9/11/2021|        Metrocity|
|   I173702|    C988172|Female| 66|          Shoes|       5|3000.85|   Credit Card|  16/05/2021|     Metropol AVM|
|   I337046|    C189076|Female| 53|          Books|       4|   60.6|          Cash|  24/10/2021|           Kanyon|
|   I227836|    C657758|Female| 28|       Clothing|       5| 1500.4|   Credit Card|  24/05/2022|   Forum Istanbul|
|   I121056|    C151197|Female| 49|      Cosmetics|       1|  40.66|          Cash|  13/03/2022|     Istinye Park|
|   I293112|    C176086|Female| 32|       Clothing|       2| 600.16|   Credit Card|  13/01/2021| Mall of Istanbul|
|   I293455|    C159642|  Male| 69|       Clothing|       3| 900.24|   Credit Card|   4/11/2021|        Metrocity|
|   I326945|    C283361|Female| 60|       Clothing|       2| 600.16|   Credit Card|  22/08/2021|           Kanyon|
|   I306368|    C240286|Female| 36|Food & Beverage|       2|  10.46|          Cash|  25/12/2022|        Metrocity|
|   I139207|    C191708|Female| 29|          Books|       1|  15.15|   Credit Card|  28/10/2022|Emaar Square Mall|
|   I640508|    C225330|Female| 67|           Toys|       4| 143.36|    Debit Card|  31/07/2022|        Metrocity|
|   I179802|    C312861|  Male| 25|       Clothing|       2| 600.16|          Cash|  17/11/2022|      Cevahir AVM|
```

Took 16 sec. Last updated by anonymous at April 23 2023, 5:27:48 PM.

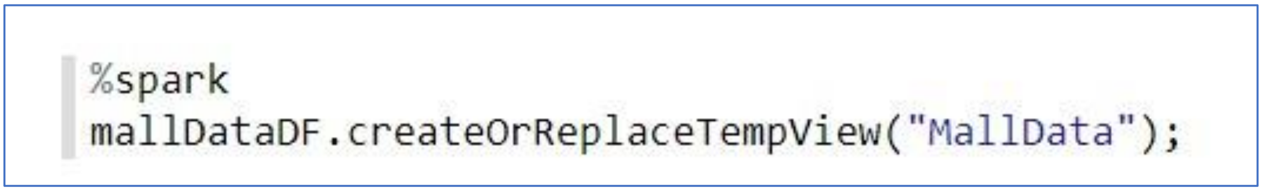
Figure 2. using spark to Load Data in Dataframe ("mallDataDF")

Dataset source comes from:
https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset



```
%spark
mallDataDF.describe().show()

+-------+----------+-----------+------+------------------+--------+------------------+-----------------+--------------+------------+-------------+
|summary|invoice_no|customer_id|gender|               age|category|          quantity|            price|payment_method|invoice_date|shopping_mall|
+-------+----------+-----------+------+------------------+--------+------------------+-----------------+--------------+------------+-------------+
|  count|     99457|      99457| 99457|             99457|   99457|             99457|            99457|         99457|       99457|        99457|
|   mean|      null|       null|  null| 43.42708909377922|    null| 3.003428617392441|689.2563209224847|          null|        null|         null|
| stddev|      null|       null|  null|14.990053791852443|    null|1.4130251343054312|941.1845672154642|          null|        null|         null|
|    min|   I100008|    C100004|Female|                18|   Books|                 1|             5.23|          Cash|    1/1/2021|  Cevahir AVM|
|    max|   I999994|    C999995|  Male|                69|    Toys|                 5|           5250.0|    Debit Card|    9/9/2022| Zorlu Center|
+-------+----------+-----------+------+------------------+--------+------------------+-----------------+--------------+------------+-------------+
```

Figure 3. Get Statistics of Data



```
%spark
mallDataDF.createOrReplaceTempView("MallData");
```

Figure 4. Temporary View to “MallData”

Methodology

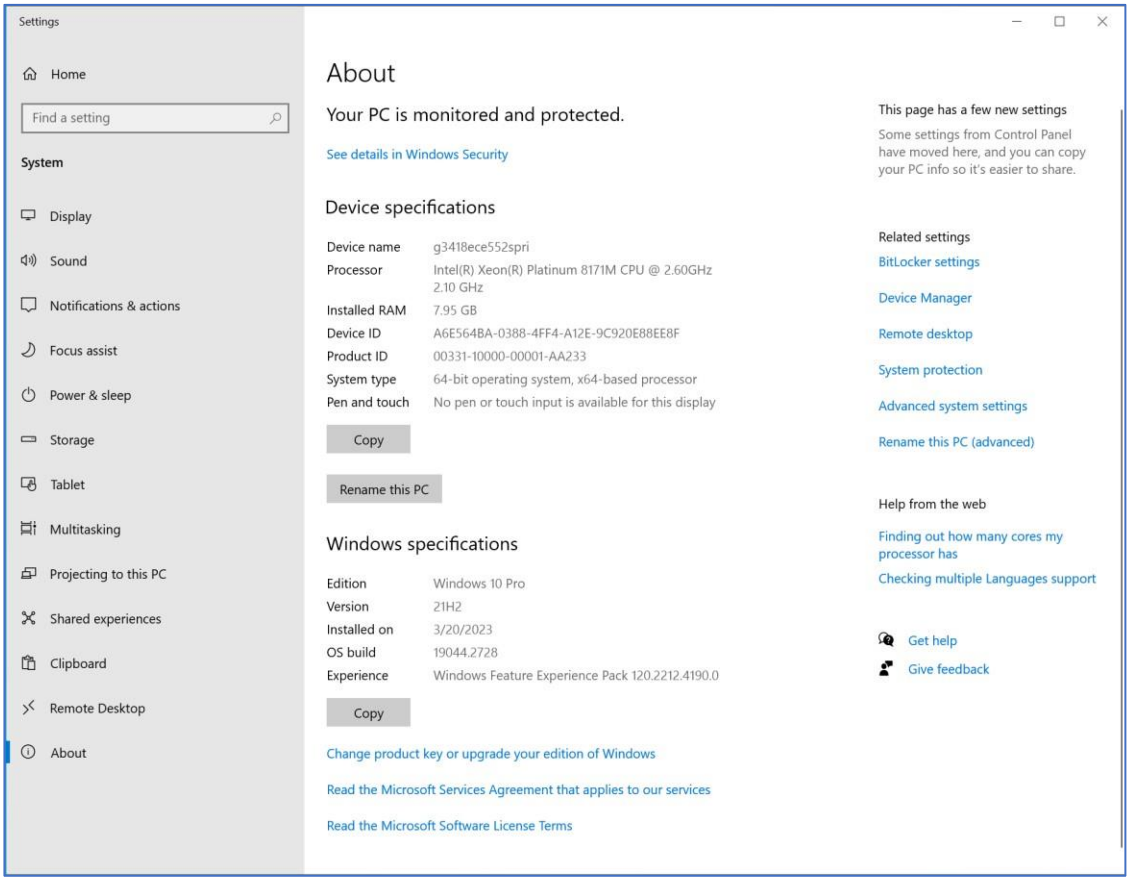
This section will provide all the methods we used during the analysis as well as our project ideas. Section 1 describes our system architecture, the tools we will use, and the environment in which we work. Section 2 provides a description of data exploration, which mentions why we specifically added two new columns and the meaning of some important attributes. Section 3 will provide the process of analyzing what library we use from Apache Spark and why we use this library, and why we use Zeppelin.

1. Architecture and tools

In the choice of virtual machine, we use the same scale of Microsoft Azure virtual machine as we have been learning and applying throughout the semester to perform analysis, (in the Windows 10 64-bit operating system environment with standard D4s v3 (4 vcpus, 16 GiB memory)). We download the specified datasets in an Azure virtual machine and use the following applications to process our analysis:

- Open-source cluster computing framework: Apache Spark
- Web GUI for encoding: Zeppelin, version xxxxxx.xxxx
- API: PySpark, Spark.sql
- Visualization: Python libraries (Pandas, Matplotlib, MLlib)

Dataset source comes from:
https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset



About

Your PC is monitored and protected.

See details in Windows Security

Device specifications

Device name g3418ece552spri
Processor Intel(R) Xeon(R) Platinum 8171M CPU @ 2.60GHz 2.10 GHz
Installed RAM 7.95 GB
Device ID A6E564BA-0388-4FF4-A12E-9C920E88EE8F
Product ID 00331-10000-00001-AA233
System type 64-bit operating system, x64-based processor
Pen and touch No pen or touch input is available for this display

Windows specifications

Edition Windows 10 Pro
Version 21H2
Installed on 3/20/2023
OS build 19044.2728
Experience Windows Feature Experience Pack 120.2212.4190.0

Change product key or upgrade your edition of Windows

Read the Microsoft Services Agreement that applies to our services

Read the Microsoft Software License Terms

This page has a few new settings

Some settings from Control Panel have moved here, and you can copy your PC info so it's easier to share.

Related settings

Help from the web

Finding out how many cores my processor has

Checking multiple Languages support

Figure 5. overview about MS Azure Virtual Machine Device Specifications

The described system will run in Zeppelin notebook. The dataset in CSV format was downloaded from Kaggle. Everything runs on MS Azure Virtual Machine. As you can see in Figure 6 below, it shows how our tools are connected and work together.

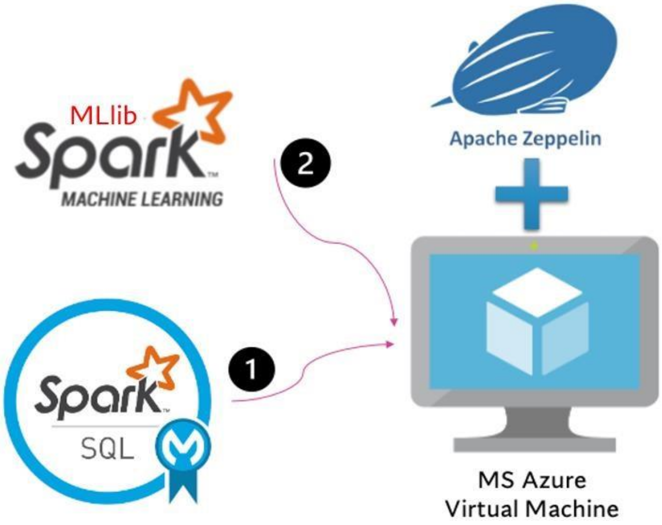


Figure 6. Workflow and Architecture

Dataset source comes from:
https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

2. Data Exploration

As mentioned in the above, the Data Source section contains shopping information from 10 different shopping malls between 2021 and 2023. The author has gathered data from various age groups and genders to provide a comprehensive view of shopping habits in Istanbul (Contains 99,457 rows * 10 columns). The following content will explain in detail the meanings of the different columns in Figure 1:

invoice_no: Invoice number. Nominal. A combination of the letter 'I' and a 6-digit integer uniquely assigned to each operation.

customer_id: Customer number. Nominal. A combination of the letter 'C' and a 6-digit integer uniquely assigned to each operation.

gender: String variable of the customer's gender.

age: Positive Integer variable of the customer's age.

category: String variable of the category of the purchased product.

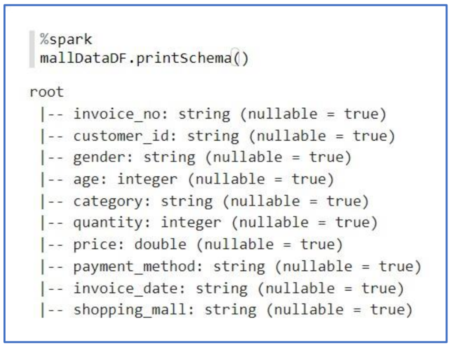
quantity: The quantity of each product (item) per transaction. Numeric.

price: Unit price. Numeric. Product price per unit in Turkish Liras

(TL).

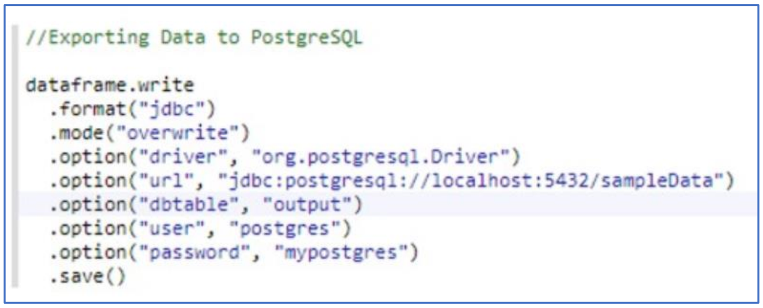
payment_method: String variable of the payment method (cash, credit card or debit card) used for the transaction.

invoice_date: Invoice date. The day when a transaction was generated. **shopping_mall**: String variable of the name of the shopping mall where the transaction was made.



```
%spark
mallDataDF.printSchema()

root
 |-- invoice_no: string (nullable = true)
 |-- customer_id: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- category: string (nullable = true)
 |-- quantity: integer (nullable = true)
 |-- price: double (nullable = true)
 |-- payment_method: string (nullable = true)
 |-- invoice_date: string (nullable = true)
 |-- shopping_mall: string (nullable = true)
```

Figure 7. Schema of Dataframe



```
//Exporting Data to PostgreSQL

dataframe.write
  .format("jdbc")
  .mode("overwrite")
  .option("driver", "org.postgresql.Driver")
  .option("url", "jdbc:postgresql://localhost:5432/sampleData")
  .option("dbtable", "output")
  .option("user", "postgres")
  .option("password", "mypostgres")
  .save()
```

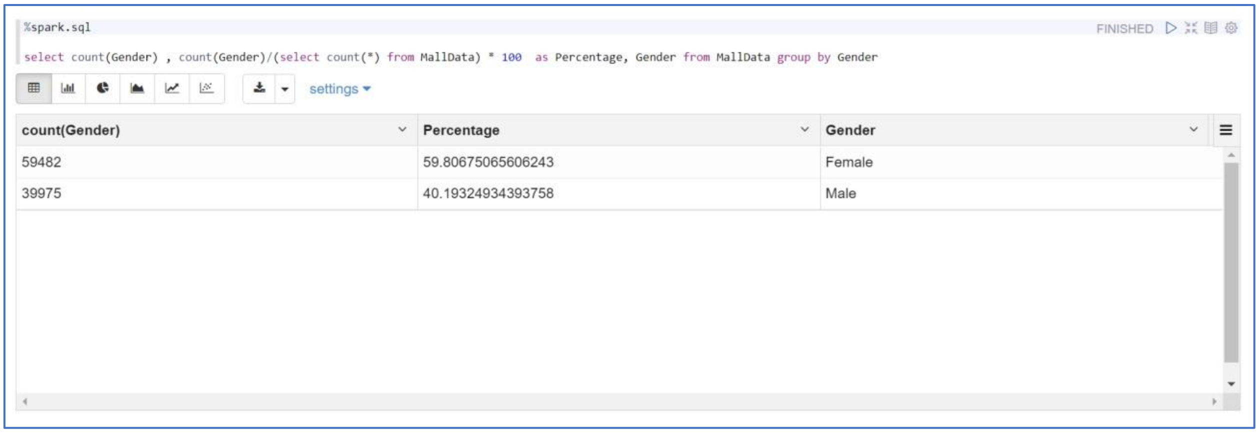
Figure 8. Exporting Data to PostgreSQL

As Figure 8 shows above, in terms of Data Exploration, our group converted the csv file and added data to PostgreSQL to make sure that the dataset is in parquet form.

Parquet is a columnar storage format that stores data by columns instead of rows. Parquet format is an open data storage format widely supported and used in various big data processing frameworks and tools. Due to the columnar storage nature of Parquet format, it typically achieves higher compression ratios, reducing data storage space. This is particularly important in scenarios that require storing large amounts of data, as it helps save storage costs and provides better scalability.

Dataset source comes from:
https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

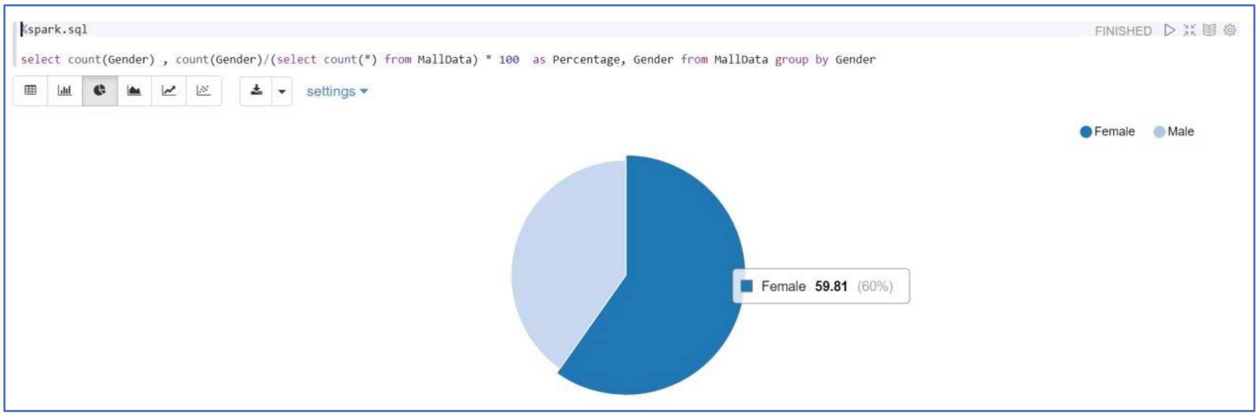
3. Data Analysis



```
%spark.sql
select count(Gender) , count(Gender)/(select count(*) from MallData) * 100  as Percentage, Gender from MallData group by Gender
```

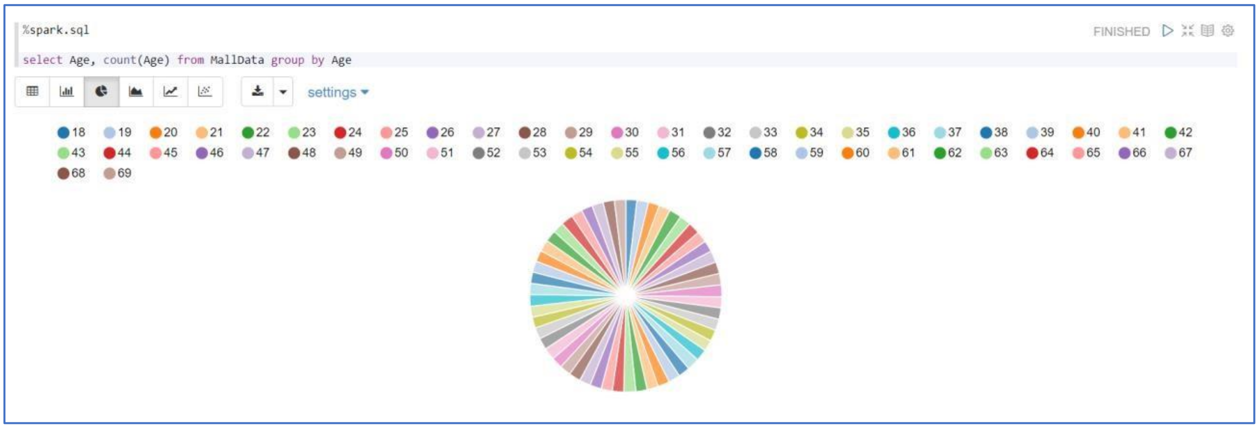
count(Gender)	Percentage	Gender
59482	59.80675065606243	Female
39975	40.19324934393758	Male

Figure 9 - 1. Ratio of Female and Male



```
%spark.sql
select count(Gender) , count(Gender)/(select count(*) from MallData) * 100  as Percentage, Gender from MallData group by Gender
```

Figure 9 - 2. Ratio of Female and Male (Pie Chart)



```
%spark.sql
select Age, count(Age) from MallData group by Age
```

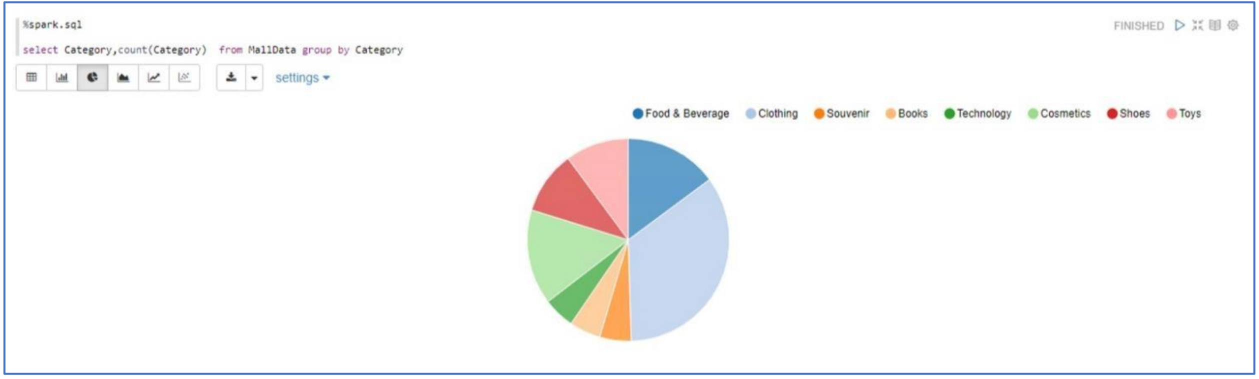
Figure 10. Visualization of Age Distribution

Dataset source comes from:
https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

MAIN TARGETS AND CONCLUSION

Three Main Targets

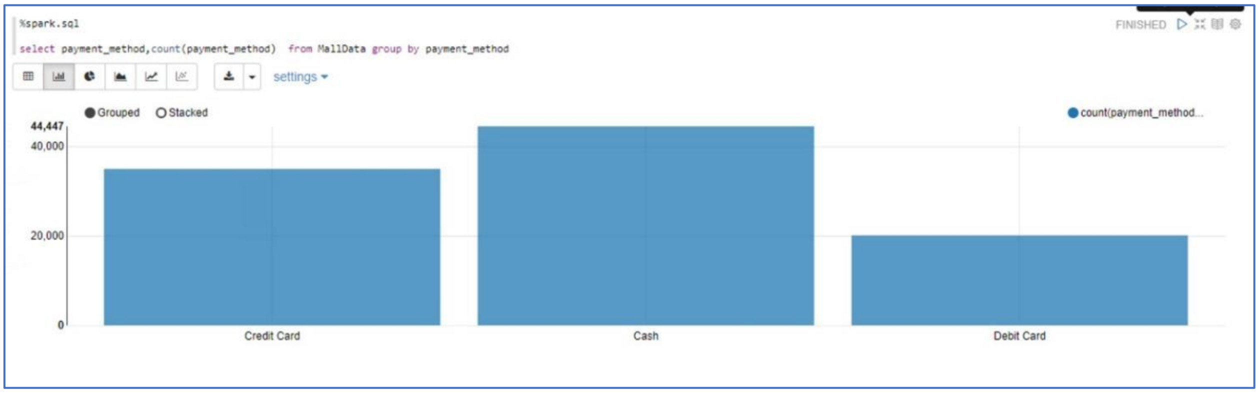
I. Which types of product is most purchased by people?



```
%spark.sql
select Category,count(Category) from MallData group by Category
```

Figure 11. Result about the most popular product all the customers want to buy II.

Which payment method is the most popular?



```
%spark.sql
select payment_method,count(payment_method) from MallData group by payment_method
```

Figure 12. Cash is the most popular payment method

III. Through the overall data analysis, use spark's unique ML library for data cleaning, training, and testing, and finally predict the total amount of bills a store can get from each customer, which can be used to predict revenue for a day.

Linear regression is a type of supervised learning algorithm that is commonly used in machine learning for prediction and data analysis. It is used to model the relationship between a dependent variable and one or more independent variables. In data analysis, linear regression is used to predict the value of a dependent variable based on the values of one or more independent variables. For example, a company may use linear regression to predict future sales based on historical sales data and other relevant factors such as marketing spend, pricing, and economic indicators. Linear regression is particularly useful in data analysis because it provides a simple, interpretable model that can be used to make predictions and draw insights from the data.

Dataset source comes from:
https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

Additionally, linear regression allows for the identification of the most important independent variables that are driving the relationship with the dependent variable.

The above few reasons support our team pick Linear regression model to train and test the customer shopping dataset. Figure 12 shows the data flow for the entire project, including ML prediction by using Linear regression.

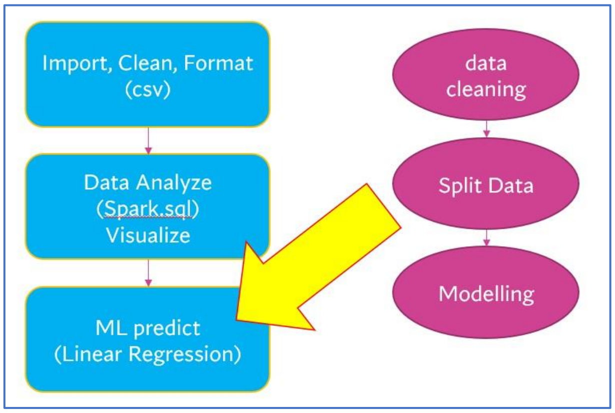
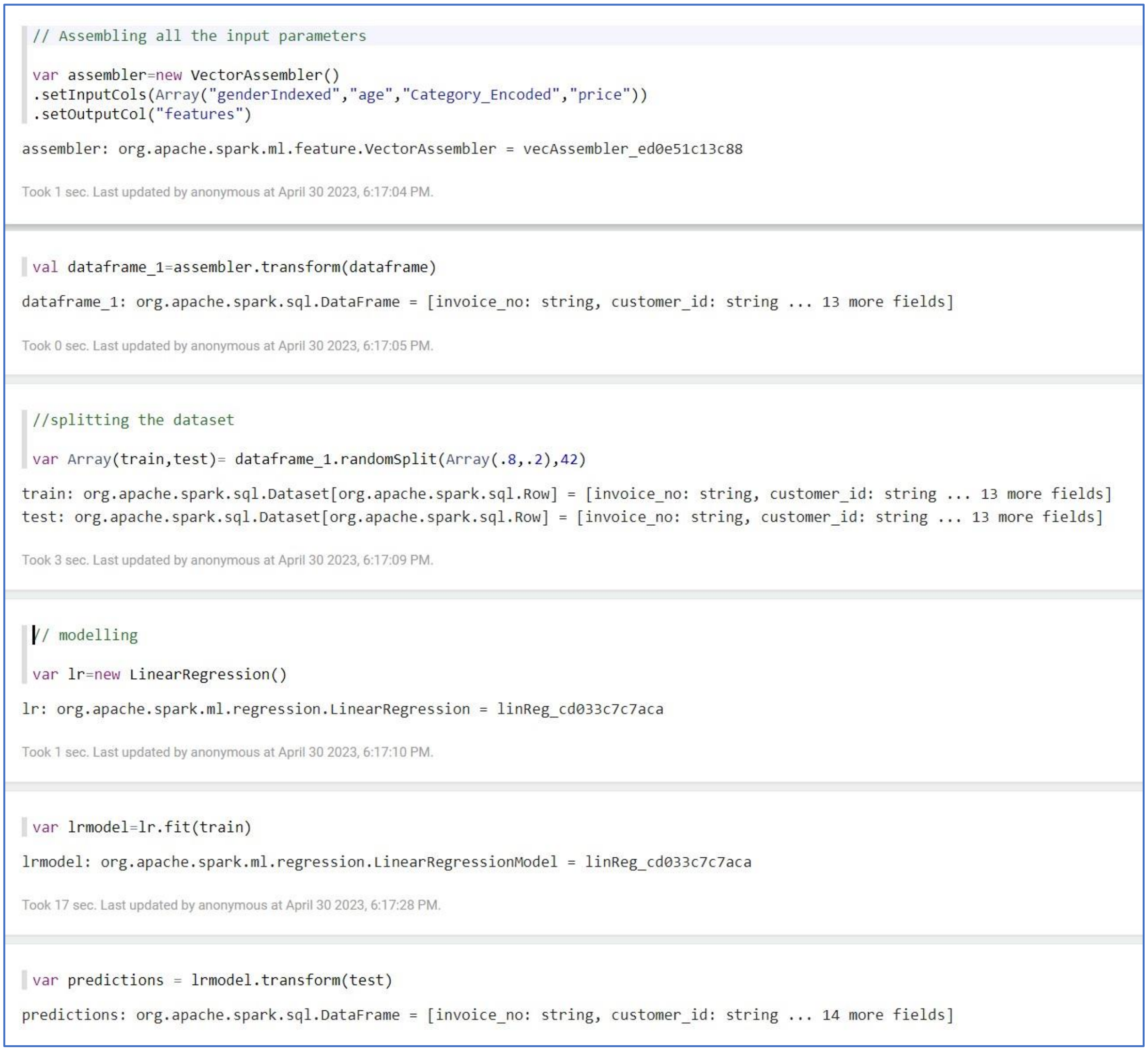

Import, Clean, Format (csv)
Data Analyze (Spark.sql) Visualize
ML predict (Linear Regression)
data cleaning
Split Data
Modelling

Figure 13. Data Flow for the entire project

When building a machine learning model, we first split the source data, use some of it to train the model, and keep some to test the trained model. In this project, we use 80% of the data for training and keep 20% of the data for testing. Then we import and call the model of the dataset, as shown in Figure 13 and 14.



```
// Assembling all the input parameters

var assembler=new VectorAssembler()
.setInputCols(Array("genderIndexed","age","Category_Encoded","price"))
.setOutputCol("features")

assembler: org.apache.spark.ml.feature.VectorAssembler = vecAssembler_ed0e51c13c88

Took 1 sec. Last updated by anonymous at April 30 2023, 6:17:04 PM.
```

```
val dataframe_1=assembler.transform(dataframe)

dataframe_1: org.apache.spark.sql.DataFrame = [invoice_no: string, customer_id: string ... 13 more fields]

Took 0 sec. Last updated by anonymous at April 30 2023, 6:17:05 PM.
```

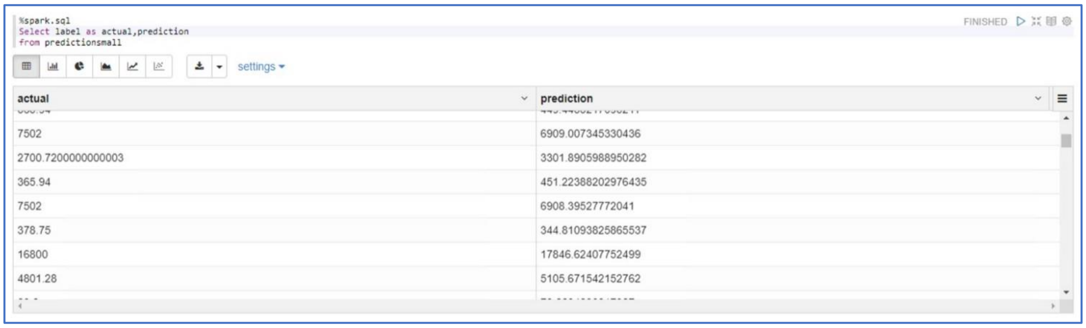
```
//splitting the dataset

var Array(train,test)= dataframe_1.randomSplit(Array(.8,.2),42)

train: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [invoice_no: string, customer_id: string ... 13 more fields]
test: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [invoice_no: string, customer_id: string ... 13 more fields]

Took 3 sec. Last updated by anonymous at April 30 2023, 6:17:09 PM.
```

```
// modelling

var lr=new LinearRegression()

lr: org.apache.spark.ml.regression.LinearRegression = linReg_cd033c7c7aca

Took 1 sec. Last updated by anonymous at April 30 2023, 6:17:10 PM.
```

```
var lrmodel=lr.fit(train)

lrmodel: org.apache.spark.ml.regression.LinearRegressionModel = linReg_cd033c7c7aca

Took 17 sec. Last updated by anonymous at April 30 2023, 6:17:28 PM.
```

```
var predictions = lrmodel.transform(test)

predictions: org.apache.spark.sql.DataFrame = [invoice_no: string, customer_id: string ... 14 more fields]
```

Figure 14. partition of coding for ML prediction



```
%spark.sql
Select label as actual,prediction
from predictionsmall
```

actual	prediction
7502	6909.007345330436
2700.7200000000003	3301.8905988950282
365.94	451.22388202976435
7502	6908.39527772041
378.75	344.81093825865537
16800	17846.62407752499
4801.28	5105.671542152762

Figure 15. The result of ML prediction

Dataset source comes from:
https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

Conclusion

According to the above visualization conclusions, we found that although the economy has been affected by the epidemic all the time in the past two years. However, we found that the number of people shopping has not dropped sharply. We also checked the purchase status of different products and learned that the top three are clothing, food and cosmetics. In addition, we performed machine learning training on the dataset through a linear regression model and classified the dataset into 80% training dataset and 20% testing dataset. From the results, we derived the predicted values shown in Figure 14.

In addition, after taking this course, we know more about sing big data technologies like Apache Spark and Zeppelin on Microsoft's virtual machines provides several advantages, including scalability, speed, flexibility, cost savings, and collaboration. These technologies can handle massive amounts of data and process them quickly, while using virtual machines allows you to easily scale your resources up or down as needed. Additionally, you can avoid the expense of purchasing and maintaining your own hardware and collaborate with other team members on big data projects using Zeppelin notebooks. Overall, this combination provides a powerful and flexible platform for processing and analyzing large datasets.

Dataset source comes from:
https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

Reference

[1] https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

[2] Armbrust, M., Das, T., Davidson, A., Ghodsi, A., Hall, D., Kang, M., ... & Xin, R. (2015). Spark SQL: Relational data processing in spark. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (pp. 1383-1394).

[3] Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., ... & Franklin, M. J. (2016). MLlib: Machine learning in Apache Spark. Journal of Machine Learning Research, 17(1), 1235-1241.

[4] Chambers, C., & Zaharia, M. (2018). Spark: The definitive guide: big data processing made simple. O'Reilly Media, Inc.

[5] Karau, H., Konwinski, A., Wendell, P., & Zaharia, M. (2015). Learning Spark: Lightning-fast big data analysis. O'Reilly Media, Inc.

Dataset source comes from:
https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset