Datamining Project Report

Nouman Ali

2024

Github profile link: <https://github.com/Noumzz>

**Data used:**

 For this project, we used the data set online retail.xlsc provided in the class by instructors. This data contained 541910 rows and 8 columns. To import this data to MySQL workbench I used table data import wizards. For that, I created the schema based on the columns of the Excel file. However, due to some null values, I guess the data is not being imported to the MySQL workbench. So, I changed the data type of the columns to text and then imported the data. Next, I changed the data types again to the desired type and this time when I got an error while I checked for null values and deleted the rows containing these values. Now there are less than 500000 rows but the data is clean. I tried to clean the data using Python but due to the large size of the data set and the poor computational power of my laptop, I was unable to do that.

Simple Queries:

1. Meta data:

To check the meta data we ran the query “Describe onlineRetail” here onlineretail is name of the table we are performing our queries, Before this query we ran use retail to select the database.

A screenshot of a computer

Description automatically generated

1. What is the distribution of order values across all customers in the dataset?

We can use the aggregate to find the distribution values in a dataset. Here I used minimum purchase, maximum purchase, total purchases made, and total amount spent. One interesting thing which I saw in my result was a negative value for the total amount spent. After analyzing the data, I realized that it was because of the returns that came negative. However, does make sense for the total spending to be negative because you can’t return more than you purchased. So, the conclusion was that this was happening due to the deleted rows in the data cleaning process. I checked for the customer's id which was negative and deleted them.

The result of the query before deletion of rows of customer ID with negative total purchases.

A screenshot of a computer

Description automatically generated

The count of rows before the deletion operation to eliminate the customer with negative spending’s.

A screenshot of a computer

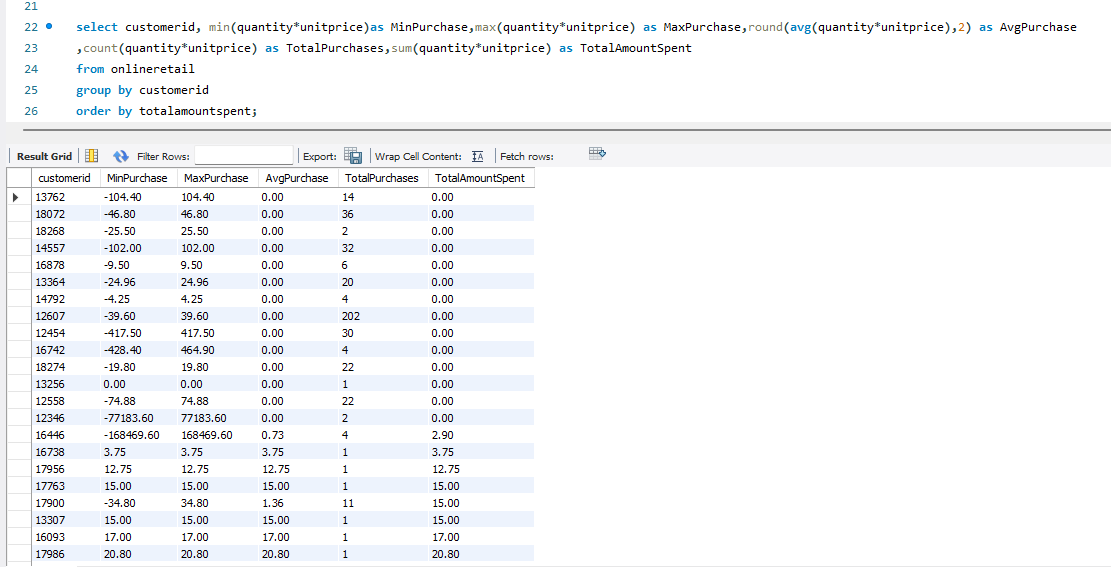
Description automatically generated

The count of rows after the deletion operation.

A screenshot of a computer program

Description automatically generated

Result of the query after performing deletion:



1. How many unique products has each customer purchased?

Here we will make a group based on customerid and count(distinct stockcode). This will tell how many unique products each customer has bought.

A screenshot of a computer

Description automatically generated

1. Which customers have only made a single purchase from the company?

In the above query we found the number of distinct products each customer have bought here we just need to find the customer having that count as 1.

A screenshot of a computer

Description automatically generated

1. Which products are most commonly purchased together by customers in the dataset?

To do this task we need to do self-join the table on condition t1.invoiceno=t2.invoiceno and t1.stockcode>t2.stockcode. Also we will use where on condition t1.customerid=t2.customerid and group them by t1.stockcode=t2.stock code. In the join clause to avoid the same pairs were as using > instead of <> (a,b <>b,a) so we will get same pair twice.

Also, I used a smaller version of data as my laptop was unable to perform this query with the original data set.

A screenshot of a computer

Description automatically generated

Advanced Quereis:

1. Group customers into segments based on their purchase frequency, such as high, medium, and low frequency customers.

To achieve this task we would use case function and make segments based on purchasing frequency and group based on customer id.

A screenshot of a computer

Description automatically generated

1. Calculate the average order value for each country to identify where your most valuable customers are located.

Here we would find the average sale and group them by country. This will give the average order value by each country.

A screenshot of a computer

Description automatically generated

1. Identify customers who haven't made a purchase in a specific period (e.g., last 6 months) to assess churn.

Here we are dealing with the datetime data type and we will find the difference of most recent purchase with the most recent purchase of the customer. We will use dateddiff() function and in having clause we will give the condition based on days. In the mention screen shot we check for greater than 180 which is approximately 6 months.

A screenshot of a computer

Description automatically generated

1. Determine which products are often purchased together by calculating the correlation between product purchases.

To find correlation first we would find the count for the purchased together items. Then check the co occurrence of products. Now we would use the formula for correlation and compute it. In screenshot whole code might not be visible please look into the sql file which I submit on lms.

Again, I am using the smaller dataset for this problem as well.

A screenshot of a computer

Description automatically generated

1. Explore trends in customer behavior over time, such as monthly or quarterly sales patterns

Again, here we are dealing with the datetime data type. We will use year, month, and quarter function of MySQL to get the required part of datetime format. Then see the spending’s of customers over different year, month, and quarter.

A screenshot of a computer

Description automatically generated