**FLIPKART DATA ANALYSIS**

**Database Foundation for Business Analytics (BUAN6320 Group Project]**

**GROUP 14**

**Sumer Pariani (SXP2103004]**

**Sai Aditya Pappu (SXP210198]**

**Pavan Katady (VXK220055]**

**Karan Reddy Patlola (SXP21232]**

**Akshitha Patchala (AXP21223]**

**Harshita Bhamidipati (]**

1. **Choosing a Dataset**

The dataset chosen for this project has been taken from Kaggle. It is that of the e-commerce website Flipkart (a subsidiary of Walmart in India]. This website is a competitor for Amazon India, and functions on a similar basis. This dataset belongs to the e-commerce domain, which is largely dominant these days, and can provide us some significant insights into the company revenues and product sales. The link for the dataset is given below:

<https://www.kaggle.com/datasets/kiranbudati/mobile-prices-flipkart>

**Business Understanding**

The dataset contains the information of the sale of various products during their ‘Big Billion Day Sale’ (similar to the Amazon Prime Day sale]. It contains various variables that are categorical, numerical, string, and date format in nature.

They have the following variable names, with their respective meaning mentioned as below:

**u\_id** = unique id for the item that acts as a unique identifier of the product

**name** = name of the product

**offer\_price** = sale price of the product after the sale day discount

**original\_price** = marked price of the product mentioned before the discount is applicable

**off\_now** = discount applicable on the product on the day of sale. It is in %

**total\_rating** = total number of customer ratings available for the specific product

**total\_reviews** = total number of customer reviews that have been given for the product

**rating** = overall rating of the product out of 5 (the average of the customer ratings]

**description** = description of the product

**created\_at** = date of the sale

**category** = categories of various products that have been sold

**screen** = if the product has a screen or not

The following are the categories for the below variables:

Category – laptops, mobiles, AC, fridge, smartwatch, washing machine, water purifier, smartwatch, water purifier, earbuds, hdd, memory cards, monitor, pen drives, power bank, ssd, wired headset, tablet, TVs

This is mainly structured data, with few duplicated values that we have managed to remove as part of data cleansing. We have created a total of 9 tables related to each other via foreign keys.

This data has been gathered to gain a deeper understanding into the various products belonging to various categories, along with their prices, offers, ratings, and reviews. We can look at various factors belonging to these areas of interest to develop business solutions aimed at improving the overall sales and revenue in the next Big Billion Day Sale.

The following questions can be answered by looking deeper into this data. They are as follows:

* What is the highest sold category?
* What is the category with highest average price?
* Which category has higher number of customer ratings?
* Do expensive product categories have more customer ratings?
* Does higher average rating mean higher product sales?
* Are products with a screen being sold more than those without?
* What range of discounts have higher sales for each category (zero-thirty, thirty-sixty, sixty plus]

1. **Data Understanding**

The following is the information for each of the variables present in the dataset

**u\_id** = varchar with alpha-numeric values

**name** = varchar with

**offer\_price** = integer

**original\_price** = integer

**off\_now** = integer

**total\_rating** = integer

**total\_reviews** = integer

**rating** = integer

**description** = varchar

**created\_at** = datetime

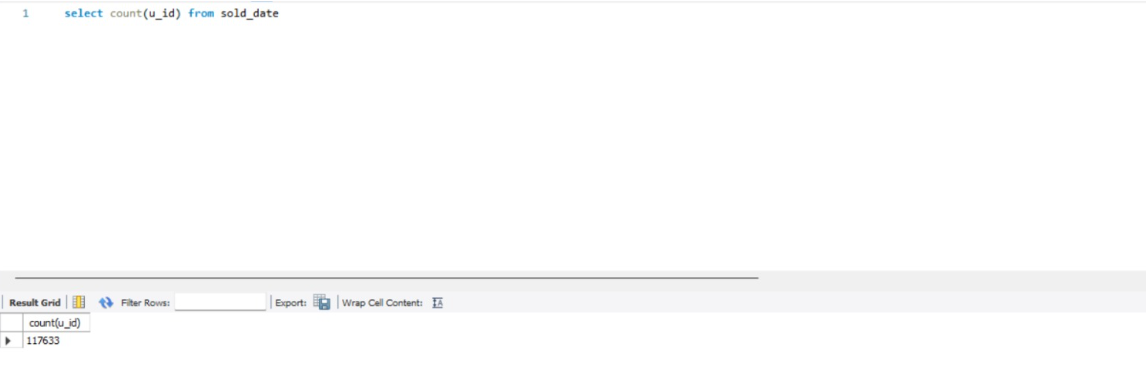
**category** = varchar

**screen** = integer (categorical variable comprising of zeros and ones]

**Insights** -

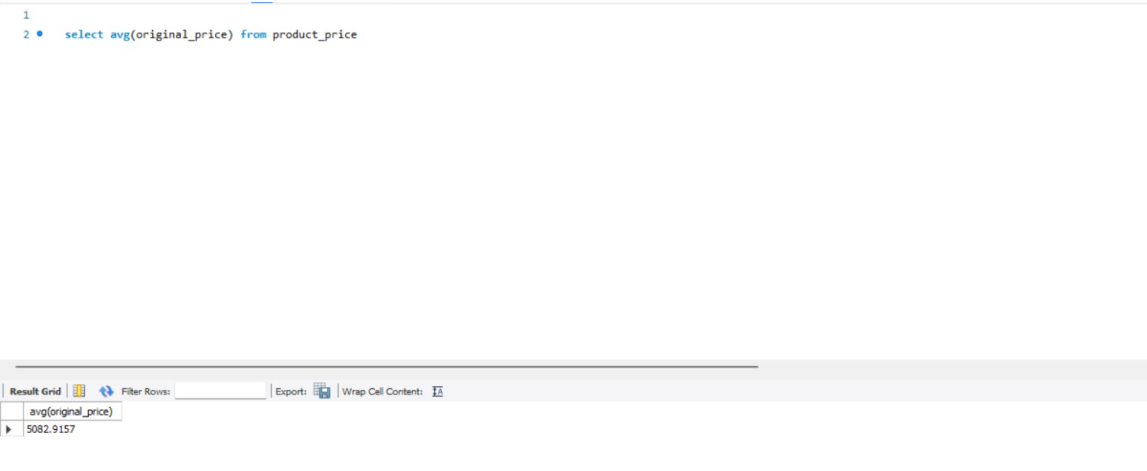
1.Total number of products sold during BIG billion days in 2022 = 177633

Query Screen shot-

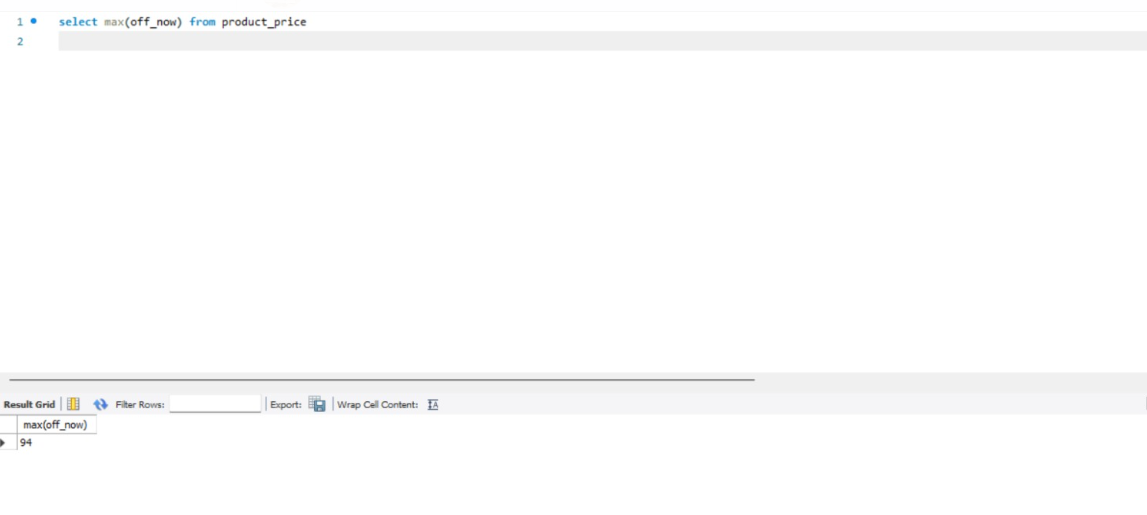


2.Average cost of goods, which includes all categories and all models

=5082.92



3.If there is maximum discount on any product, the max discount is =94% off



Database Design

Database which we have designed is from a Business perspective that mean owner of Flipkart would like to use this database to get insights of his Business .

Before we normalise our tables we have only 3 tables .

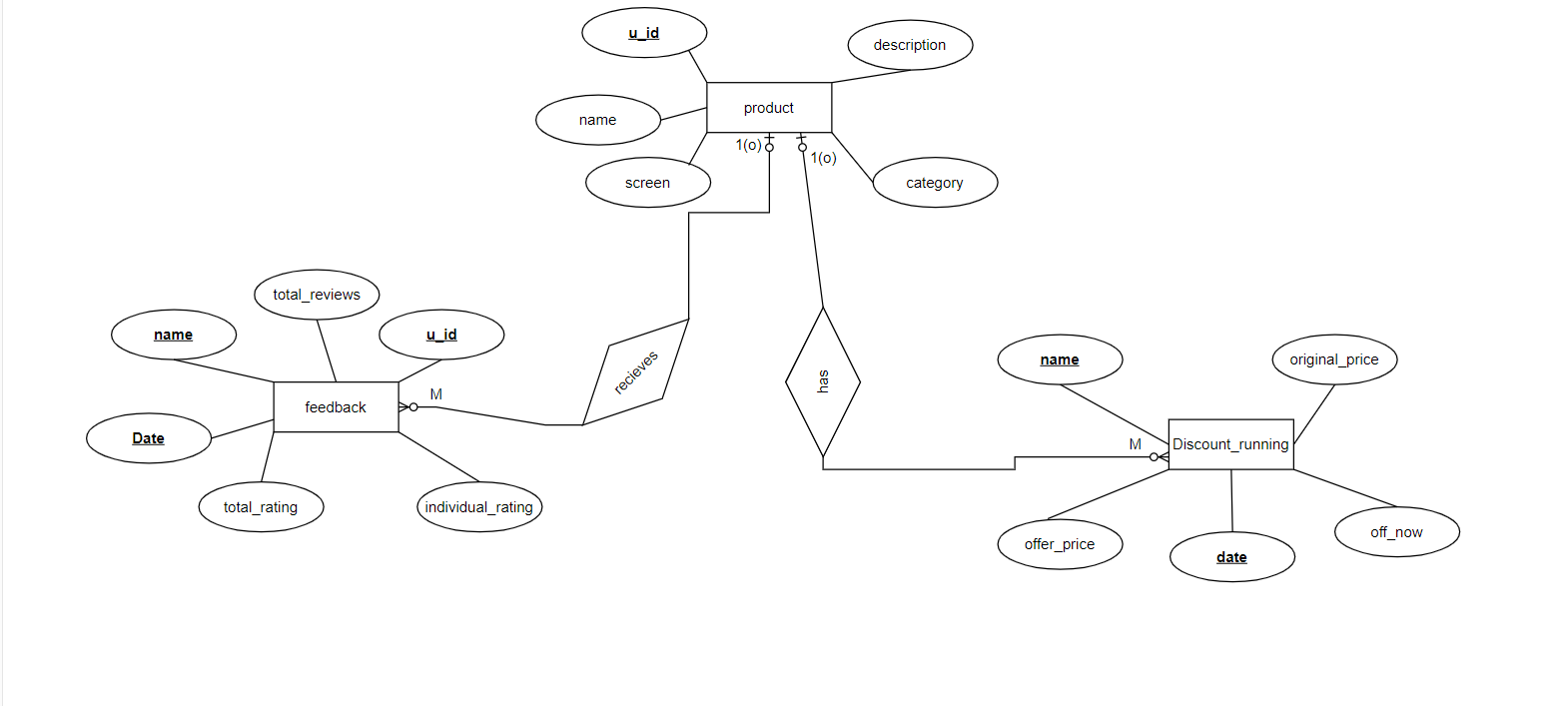
ERD diagram of the following structure

We initial had

u\_id as the Primary key for product table

name, date, u\_id as a composite primary key for feedback table

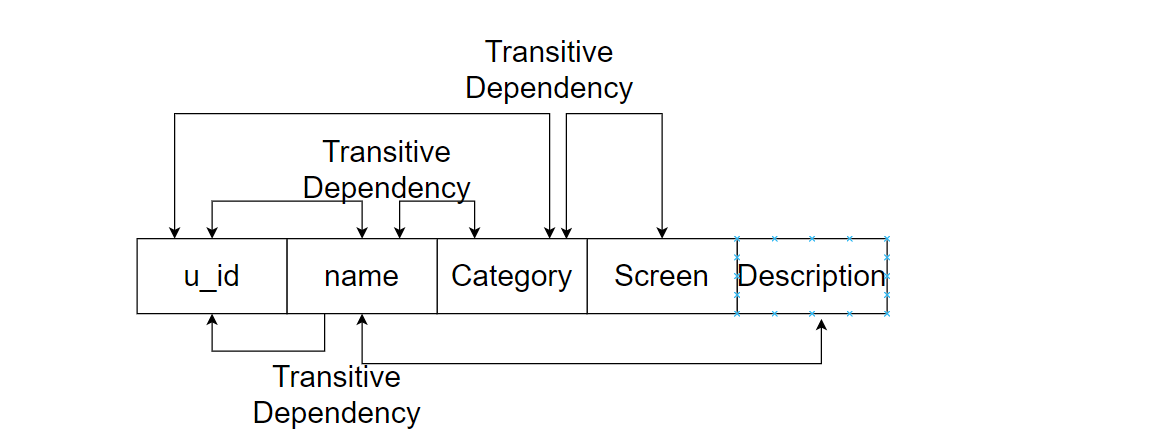
name and date as composite primary key for discount table.



**Normalization and logical Schema**

Product table

**BEFORE**



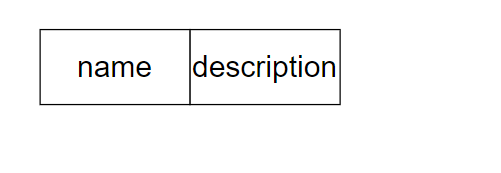
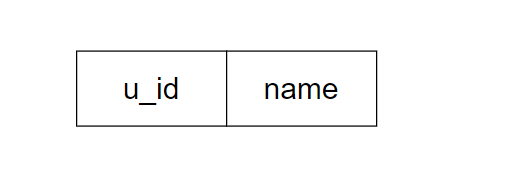
After **Normalization 3NF**

Product\_Name

Primary key – u\_id

Product\_description

Primary\_key - name



Product\_category Screen\_Equipment

Primary\_key - Category

Primary key –name

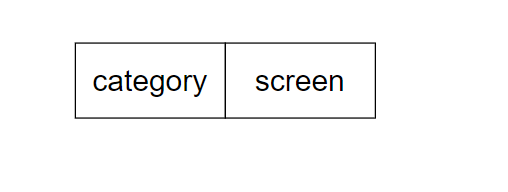
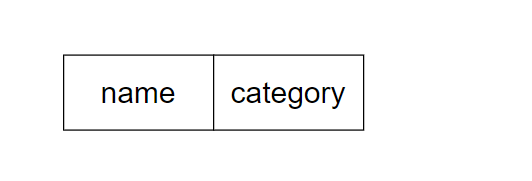
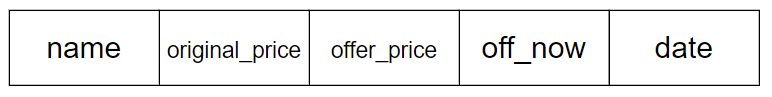


Table Discounts



Here we see full dependency

name and date are the primary keys

and offer price is a derived attribute.

Table feedback

Before normalization

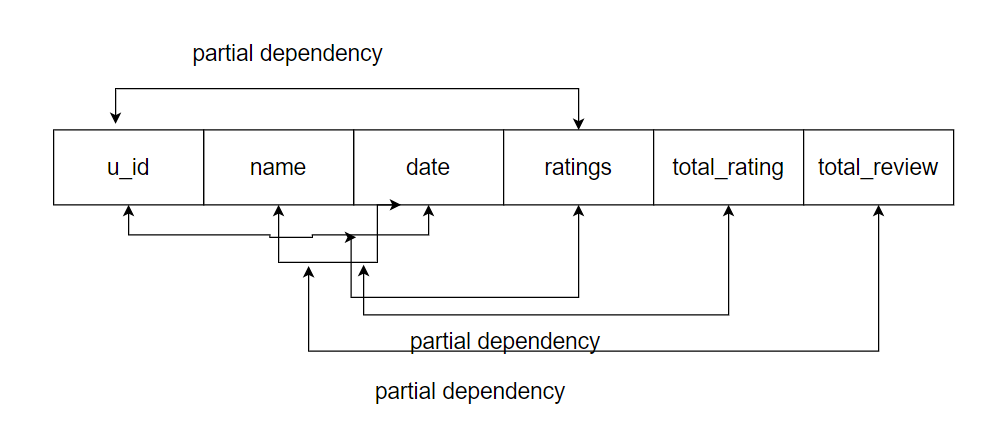
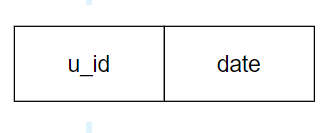


Table date\_sold

Primary key- u\_id



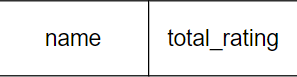
Individual\_item\_rating

Primary key- u\_id



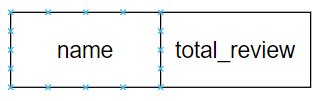
Product\_rating

Primary key- name



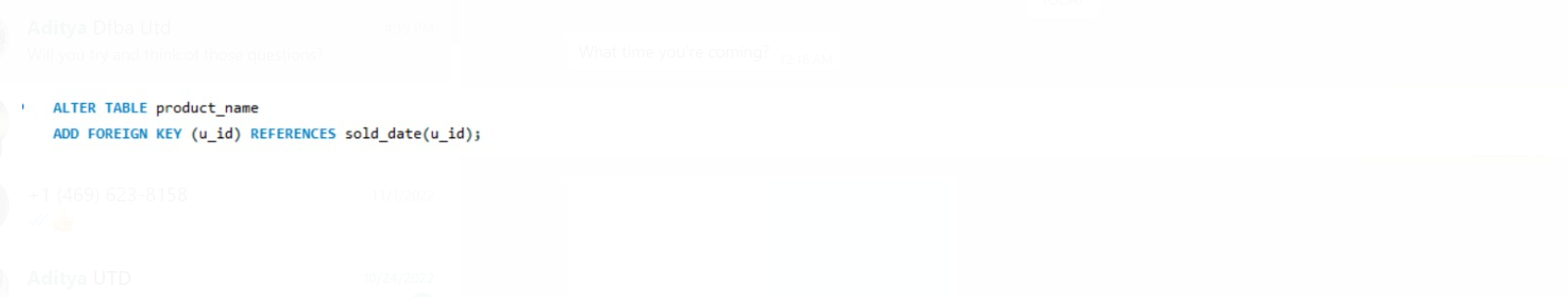
Product review

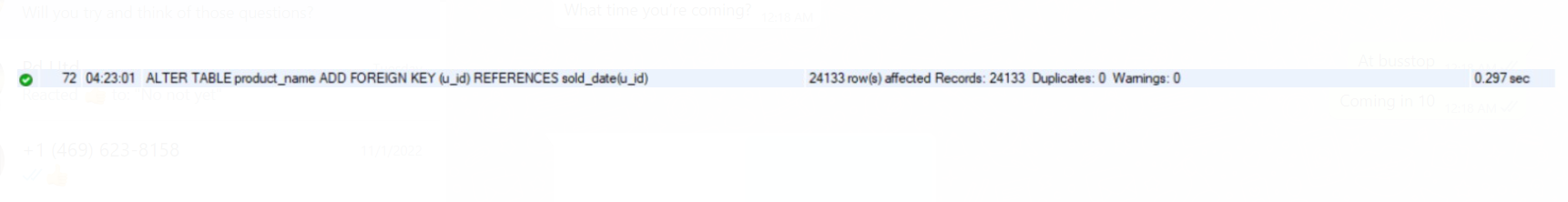
Primary key- name



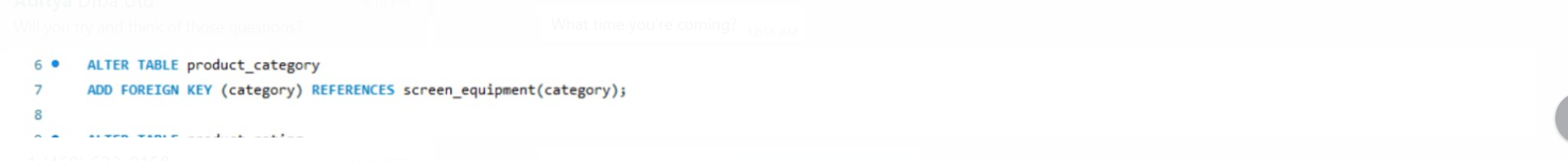
We then added foreign keys to our database for relations and references

Product\_name and sold\_date



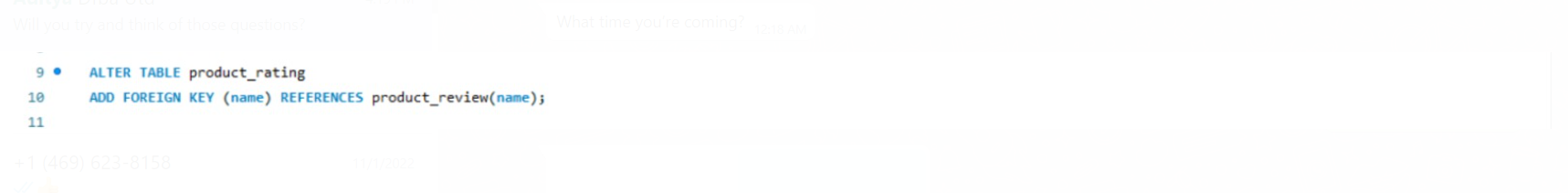


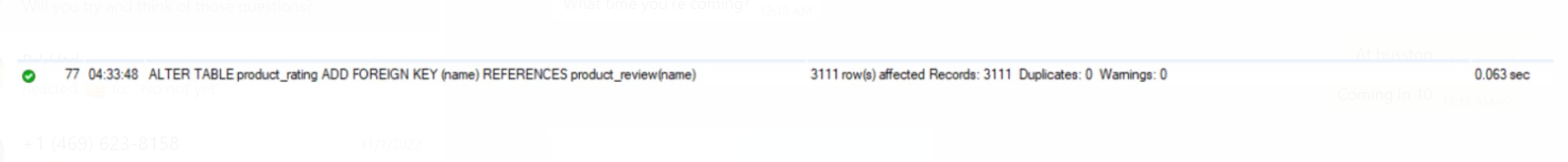
Product\_category and Screen\_equipments



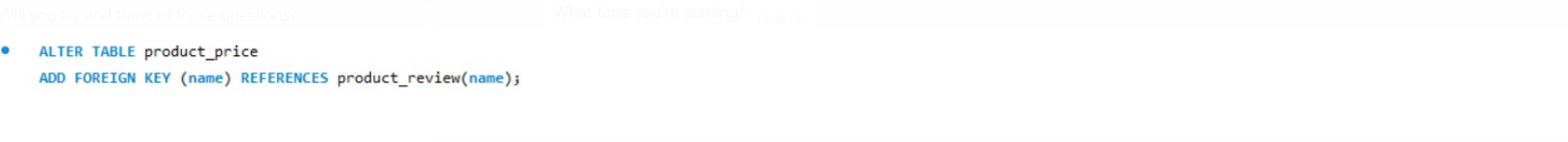


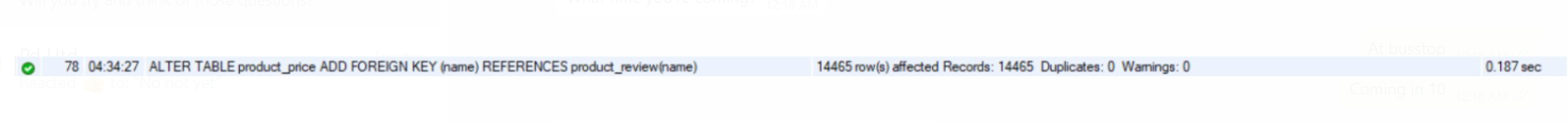
Product\_rating and reviews





Product\_price and product review





Data cleaning and database Testing

Cleaning

Duplicates for each table duplicates were removed

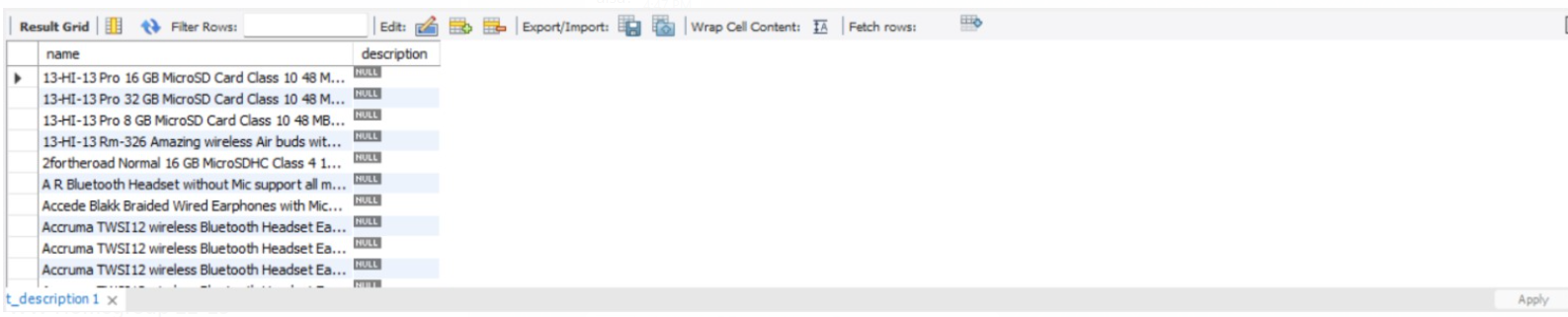
Product category

Duplicates were removed as name is taken as primary key



Product description

Duplicates were removed as name is taken as primary key



Screen equipments

Duplicates were removed as category is taken as primary key

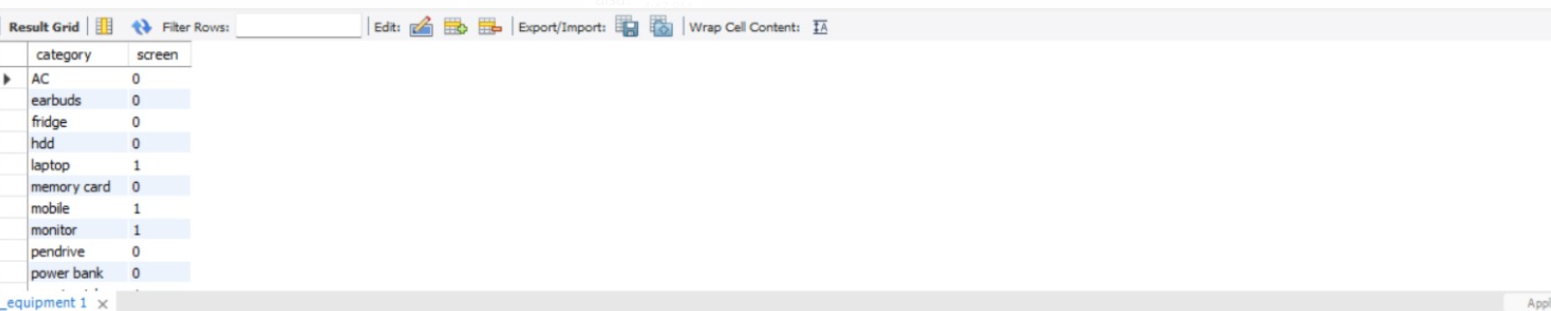
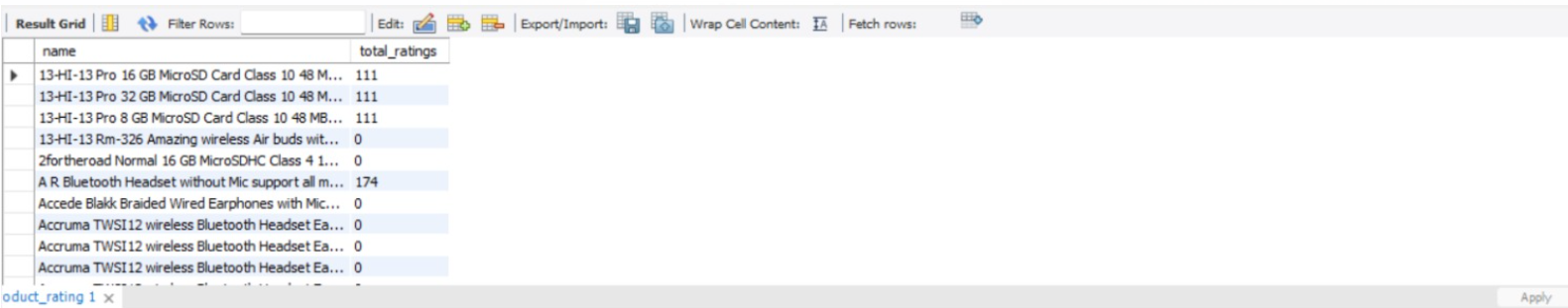
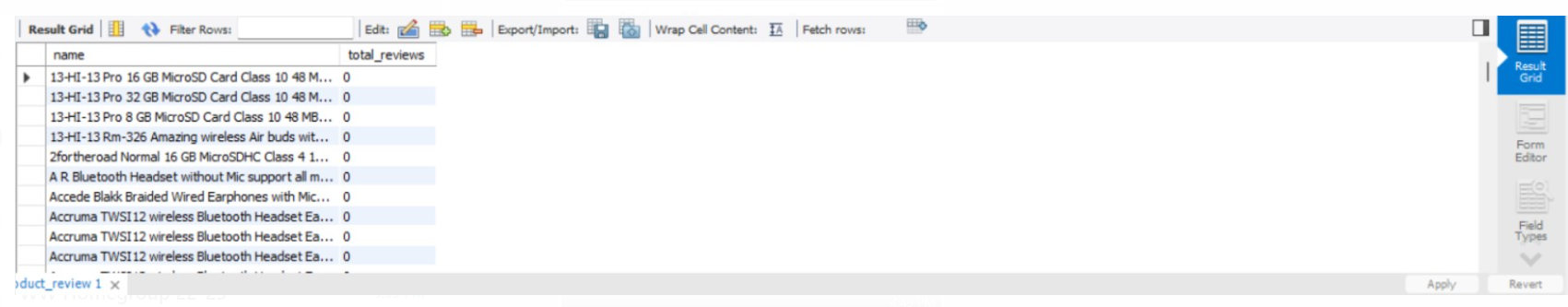


Table discounts had duplicates because name is take as primary key

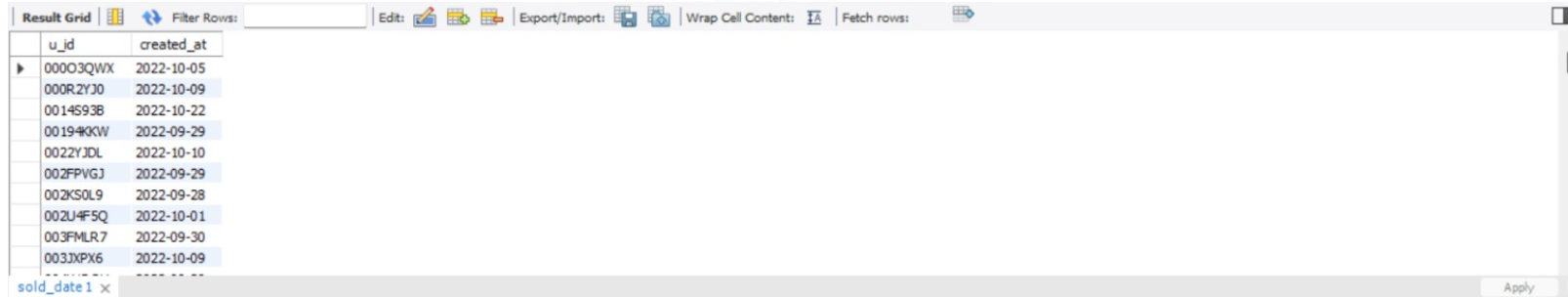


Duplicates from product\_ratings and product\_reviews were removed as anme is the primary key.

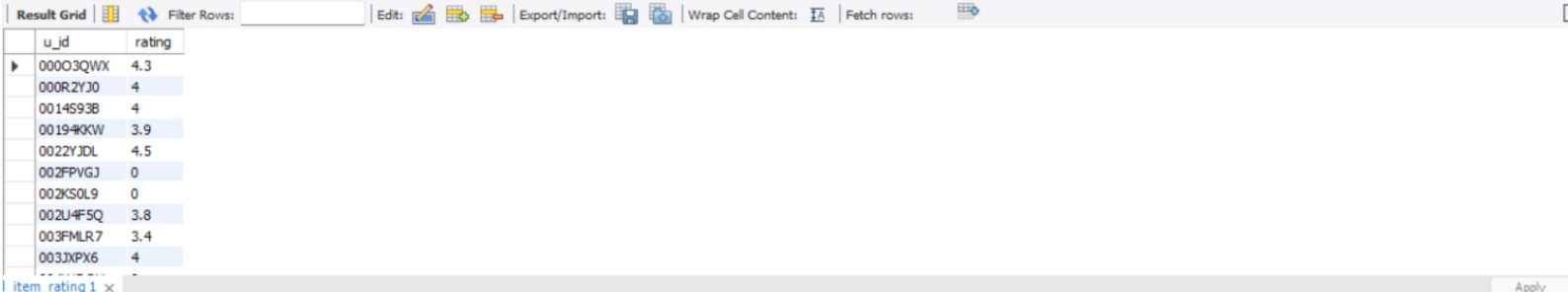




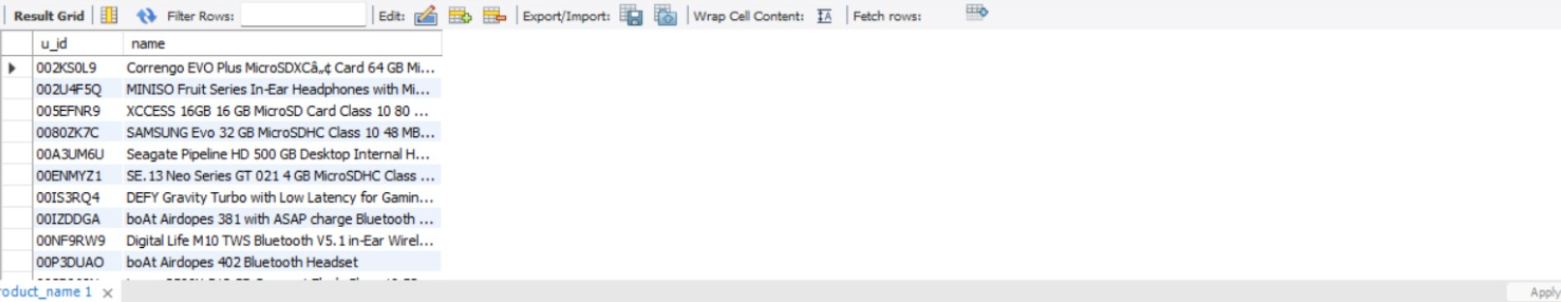
Sold\_date does not need to be cleaned as u\_id was the primary key



Ratings does not need to be cleaned as u\_id was the primary key



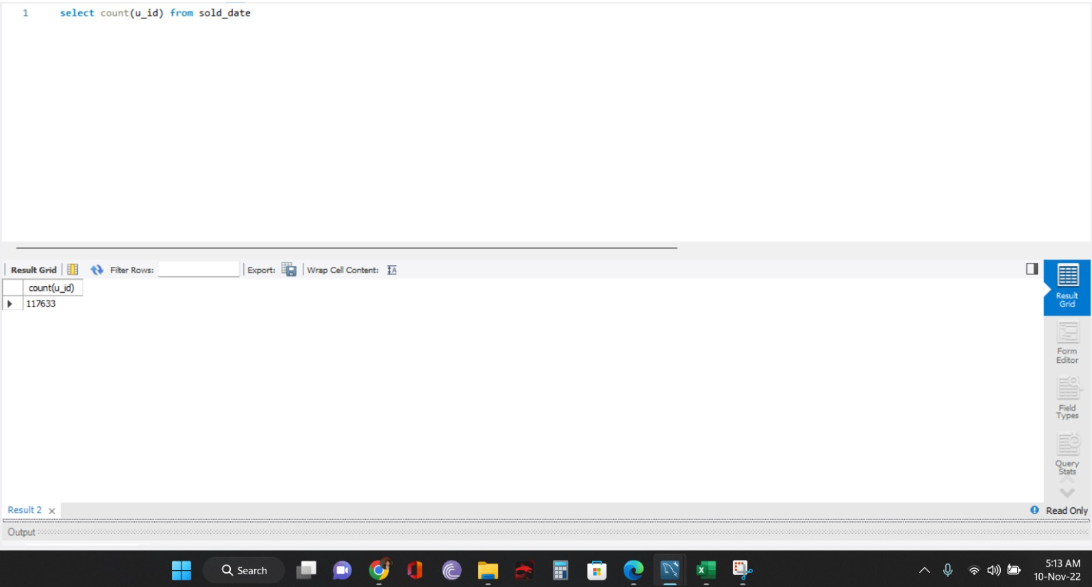
Even product\_name dint had duplicates



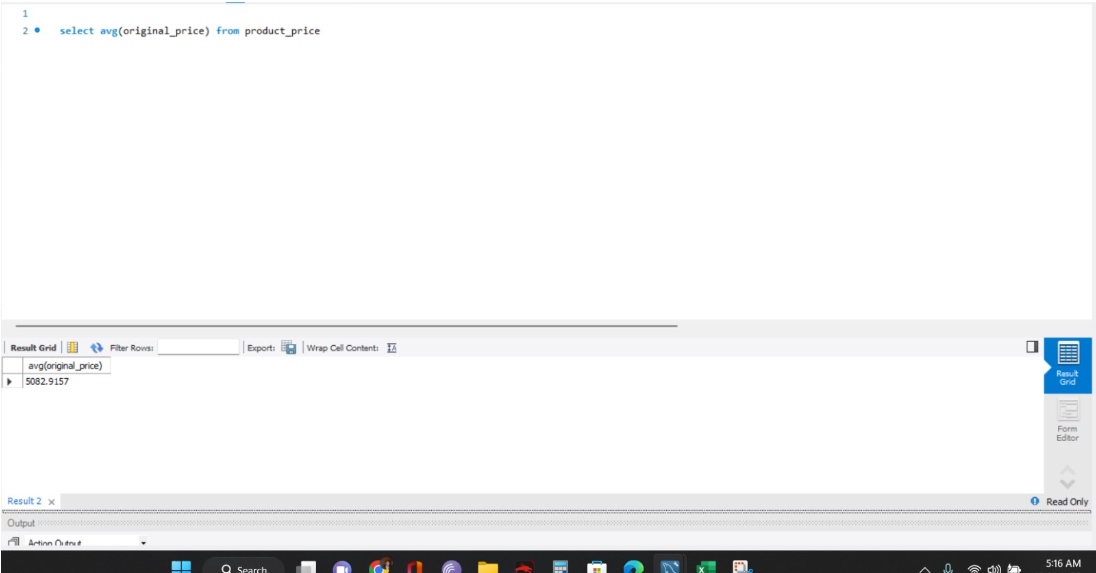
Database testing

Database was tested to answer some basic questions

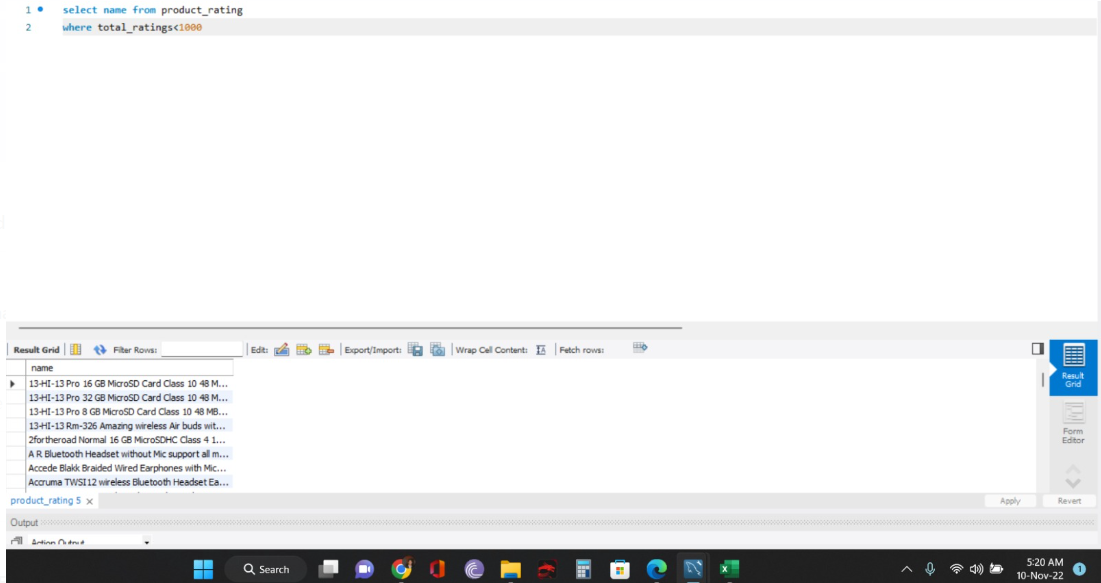
Total number of products sold during the sale in 2022

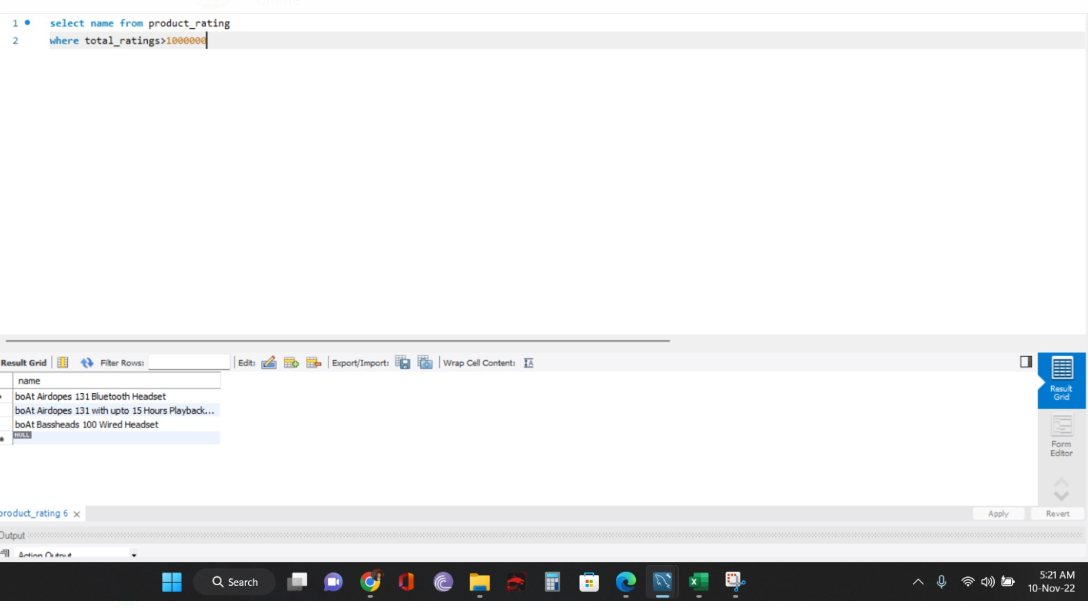


2. average price of all products available in the sale



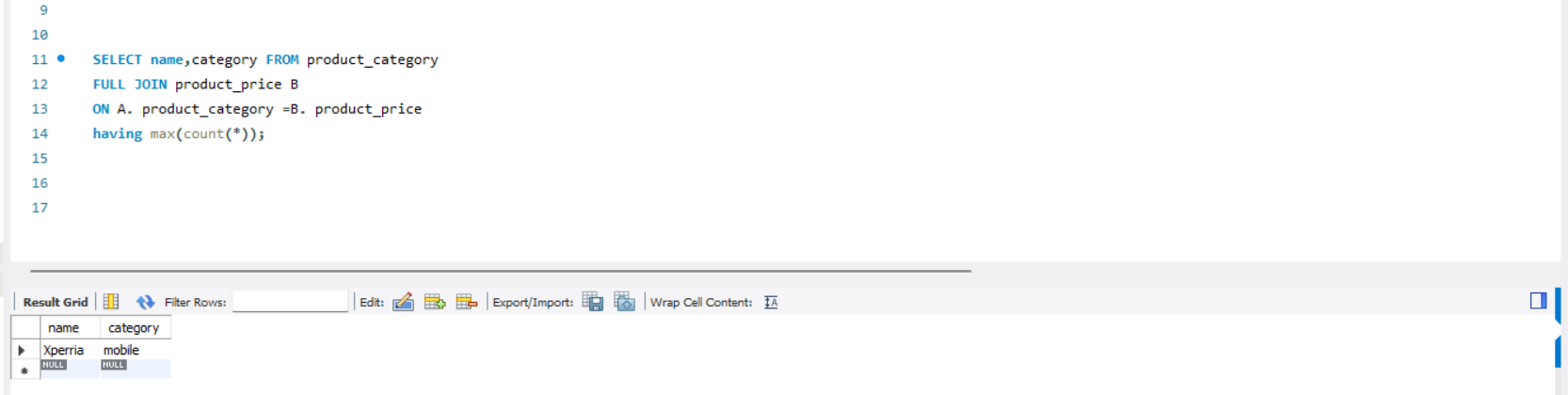
3.sorting the products based on their ratings





Important queries using foreign keys and joins

1. The highest sold category and product



1. The best product sold according to the rating during the sale

