TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN GIỮA KÌ MÔN

HỆ THỐNG THƯƠNG MẠI THÔNG MINH

MIDTERM PROJECT

Người hướng dẫn: **ThS DƯƠNG HỮU PHÚC**

Người thực hiện: **PHẠM NGUYỄN – 52000092**

NGÔ HOÀNG KHÔI – 52000676

TRẦN TRUNG HIẾU – 52000888

Lớp **: 20050261**

Khoá **: 24**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN GIỮA KÌ MÔN

HỆ THỐNG THƯƠNG MẠI THÔNG MINH

MIDTERM PROJECT

Người hướng dẫn: **ThS DƯƠNG HỮU PHÚC**

Người thực hiện: **PHẠM NGUYỄN – 52000092**

NGÔ HOÀNG KHÔI – 52000676

TRẦN TRUNG HIẾU – 52000888

Lớp **: 20050261**

Khoá **: 24**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023

LỜI CẢM ƠN

Chúng em xin chân thành gửi lời cảm ơn này đến thầy Dương Hữu Phúc giảng viên phụ trách giảng dạy bộ môn Hệ thông thương mai thông minh. Nhờ có sự tận tình giảng dạy, truyền đạt kiến thức của quý thầy mà chúng em mới đủ kiến thức để hoàn thành đồ án giữa kỳ này.

Song song với đó, chúng em cũng xin gửi lời cảm ơn đến Khoa Công Nghệ Thông Tin, trường Đại học Tôn Đức Thắng vì đã tạo điều kiện cho chúng em học tập, nghiên cứu trong suốt quá trình học tập môn học này nói riêng và cả quá trình học tại môi trường Đại học nói chung. Một lần nữa chúng em xin gửi lời cảm ơn chân thành đến mọi người và chúc tất cả thật nhiều sức khỏe.

ĐỒ ÁN ĐƯỢC HOÀN THÀNH
TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi / chúng tôi và được sự hướng dẫn của ThS Dương Hữu Phúc;. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày tháng năm

Tác giả

(ký tên và ghi rõ họ tên)

Nguyễn

Phạm Nguyễn

Khôi

Ngô Hoàng Khôi

Hiếu

Trần Văn C

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

Phần xác nhận của GV hướng dẫn

_____

_____

_____

_____

_____

_____

_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

Phần đánh giá của GV chấm bài

_____

_____

_____

_____

_____

_____

_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

SUMMARY

The midterm project questions will be organized into two types

- QUESTION 1: Exploratory data analysis (EDA).

  The midterm project questions will be organized into two types

- QUESTION 2: Augumented the dataset.

  The midterm project questions will be organized into two types

MỤC LỤC

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

CÁC KÝ HIỆU

f *Tần số của dòng điện và điện áp (Hz)*

p *Mật độ điện tích khối (C/m3)*

CÁC CHỮ VIẾT TẮT

EDA Exploratory data analysis

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

DANH MỤC HÌNH

DANH MỤC BẢNG

QUESTION 1 – EXPLORATORY DATA ANALYSIS (EDA)

1.1 Topic

Exploratory data analysis (EDA) [3] is used to analyze and investigate datasets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. Based on the idea of EDA, using Tableau software to answer the following questions:

a. Find best-selling products in the scale of 10, 100, 1K, 10K items, and visualize them.
b. Find worst-selling products in the scale of 10, 100, 1K, 10K items, and visualize them.
c. Find and explain the relationships between dataset variables.
d. Find products which are most expensive but have ratings lower than 3.0, in the scale of 10, 100, 1K items.
e. Find products which are cheapest and have ratings more than 4.0, in the scale of 10, 100, 1K items.
f. Find products which have the largest discounted price.

1.2 Answer

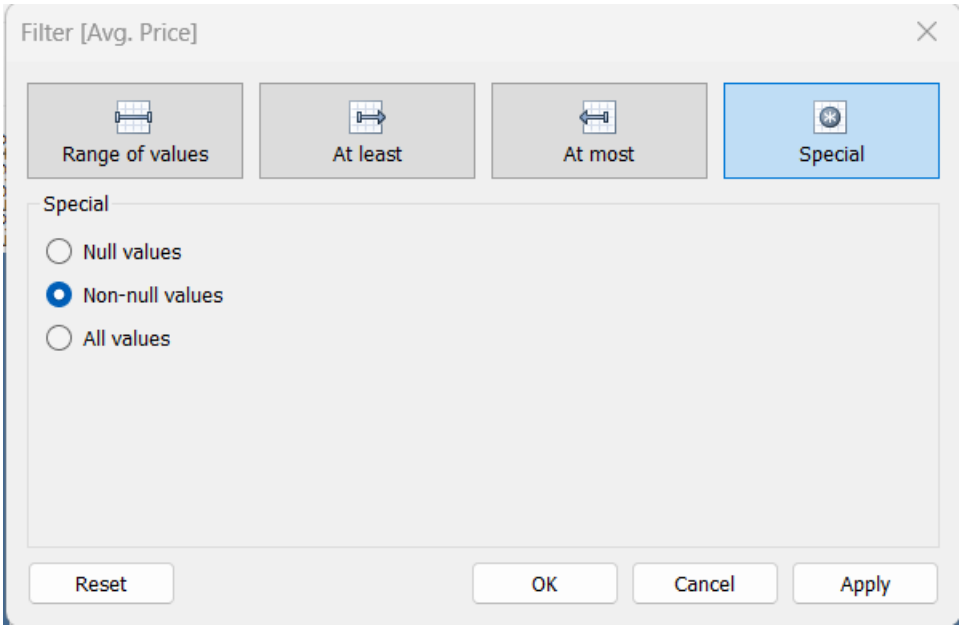
This chapter presents the reason for choosing the topic, purpose, object and scope of research, scientific and practical significance of the topic; Scientific basis for choosing topics...;

1.2.1 Hypothesis:

According to my group's hypothesis:

– Because there are products that appear many times in the data set, we should be objective. We'll use the average price of those occurrences as that product's price and remove the currency character for ease of calculation, then remove the null value for that reason are products that have not yet been brought to market.

Price = INT(RIGHT([Actual Price],LEN([Actual Price])-1))



Hình 1. 1 Filter [Avg.Price]

– Because we do not have a representative value for purchases, we use the numerical value of ratings as a substitute. After making a purchase, people can provide feedback, so the number of purchases and the number of ratings are proportional to each other.

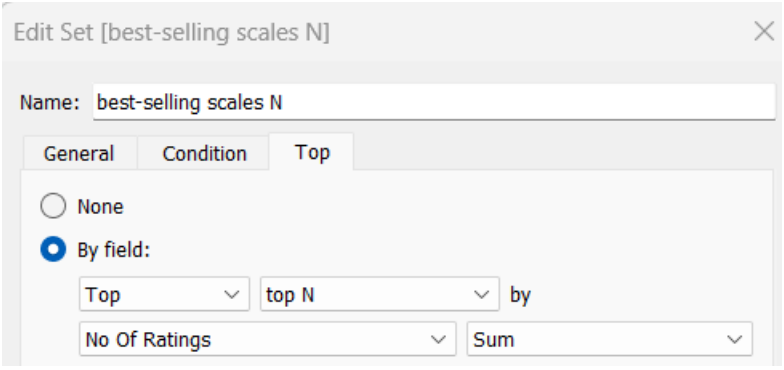
Our team hypothesizes that the 'Discount Price' column represents the price of the product after a discount. Therefore, we have:

Discount amount = AVG([Price]) - [Discounted Price (AVG)]

Discount(%) = ((AVG([Price]) - [Discounted Price (AVG)])*100)/AVG([Price])

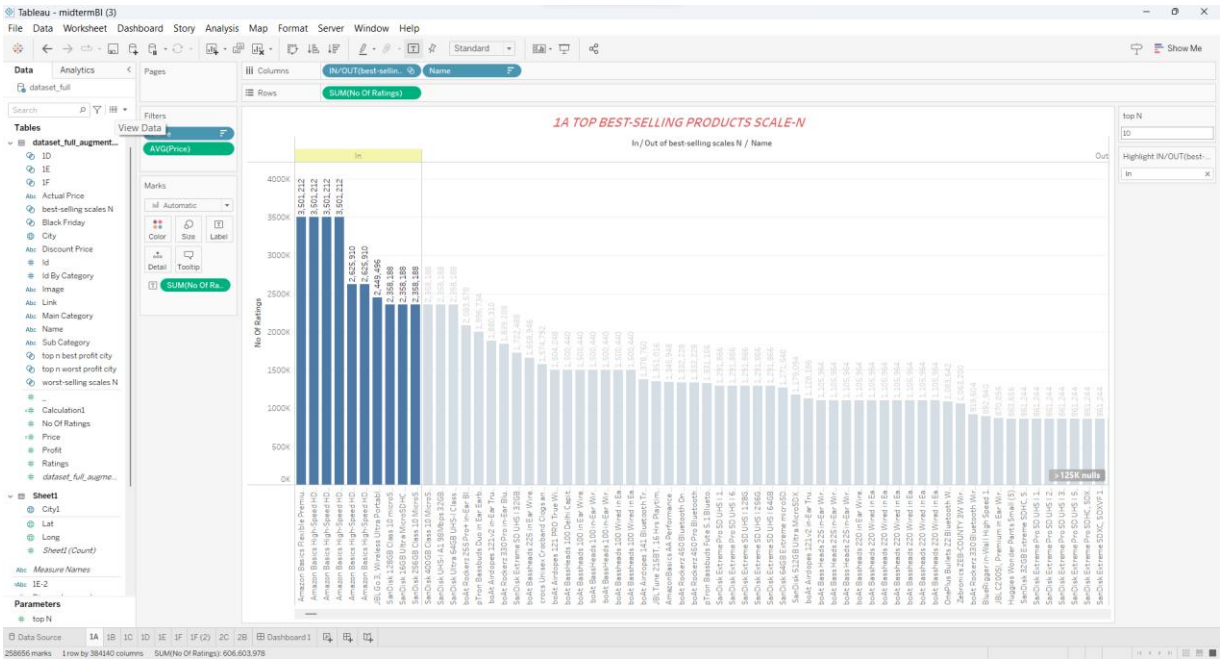
1.2.2 Problem a.

We hypothesize that the products with the highest number of purchases are also the ones with the highest number of reviews:

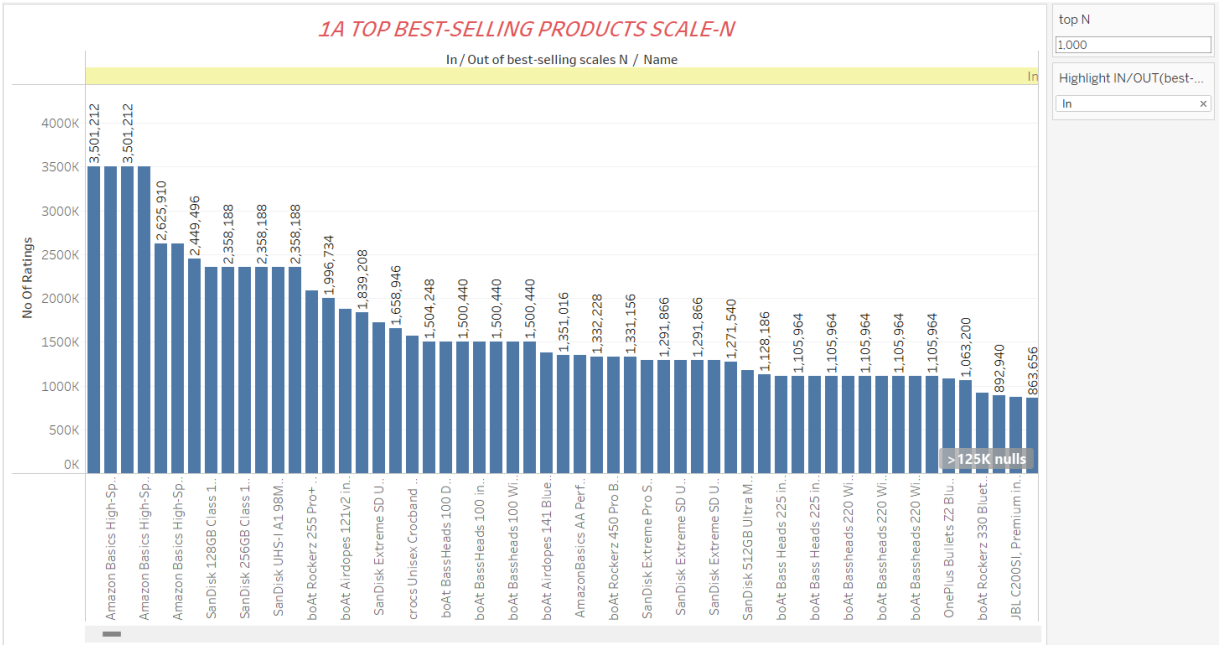


Hình 1. 2 Set [best-selling scales N]

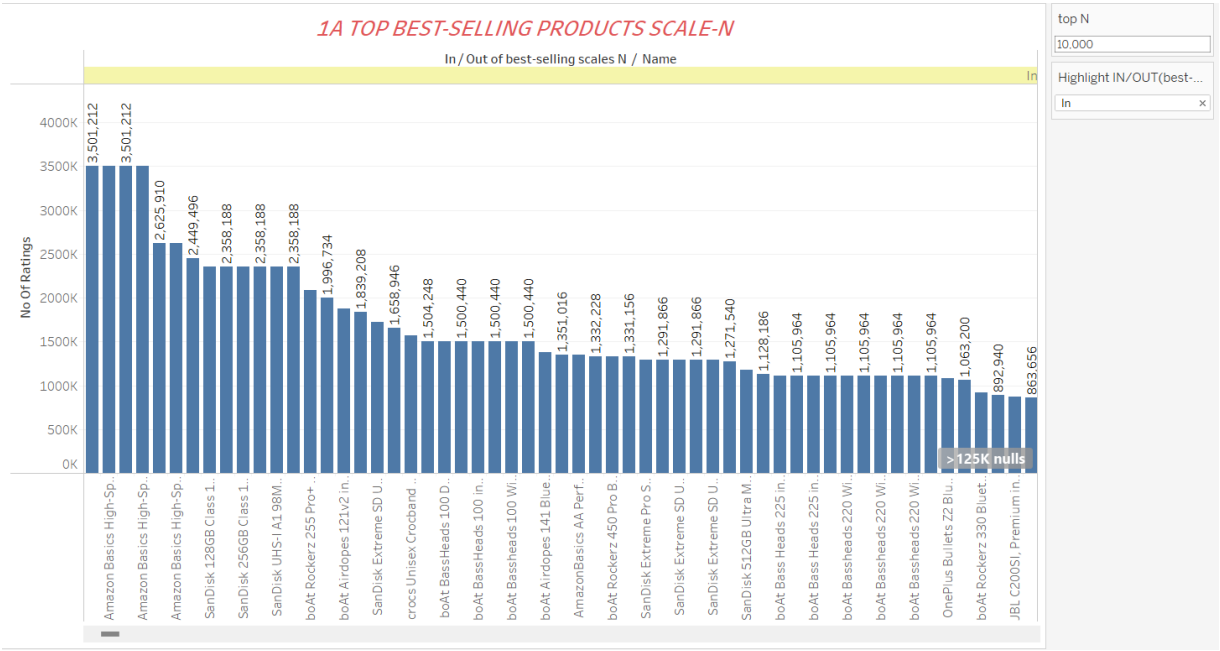
The following will be a presentation of data visualization using Tableau. Here, we will create a set to be able to count the number of best-selling products in the top 10, 100, 1000, 10000



Hình 1. 3 Top 10 best selling.



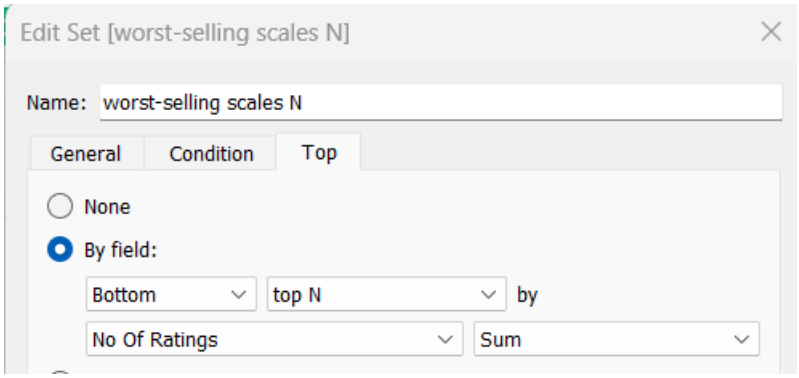
Hình 1. 4 Top 1000 best selling.



Hình 1. 5 Top 10000 best selling.

1.2.3 Problem b.

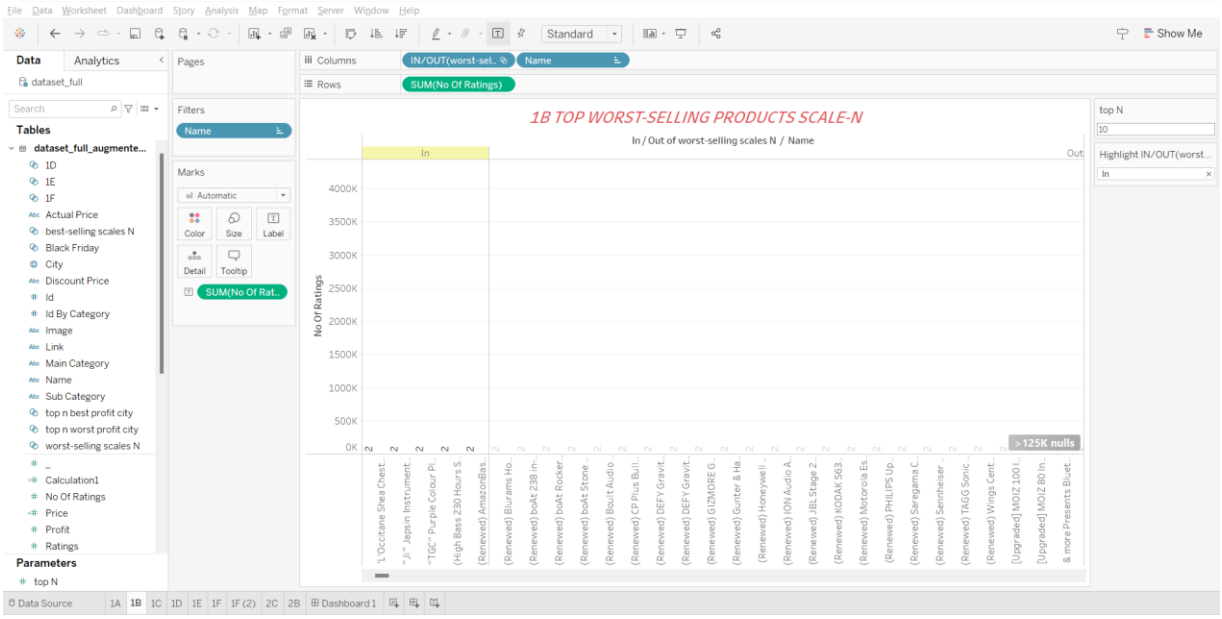
We hypothesize that the products with the lowest number of purchases are also the ones with the lowest number of reviews.



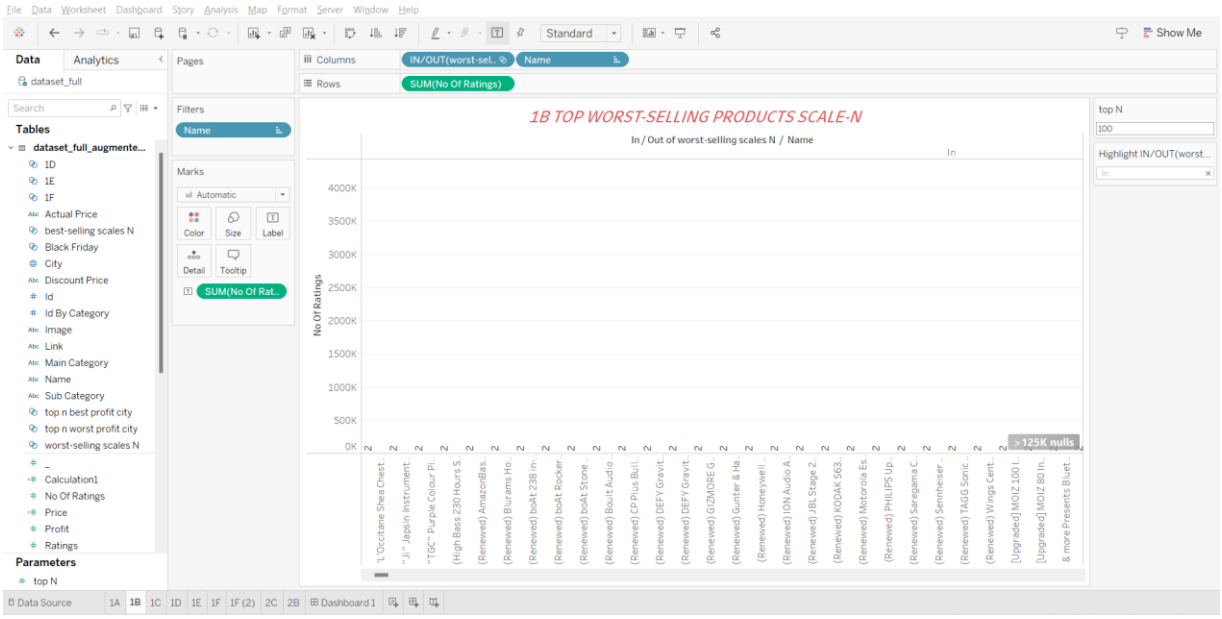
Hình 1. 6 Set [worst-selling scales N]

The following will be a presentation of data visualization using Tableau. Here, we will create a set to be able to count the number of worst-selling products in the top 10, 100,

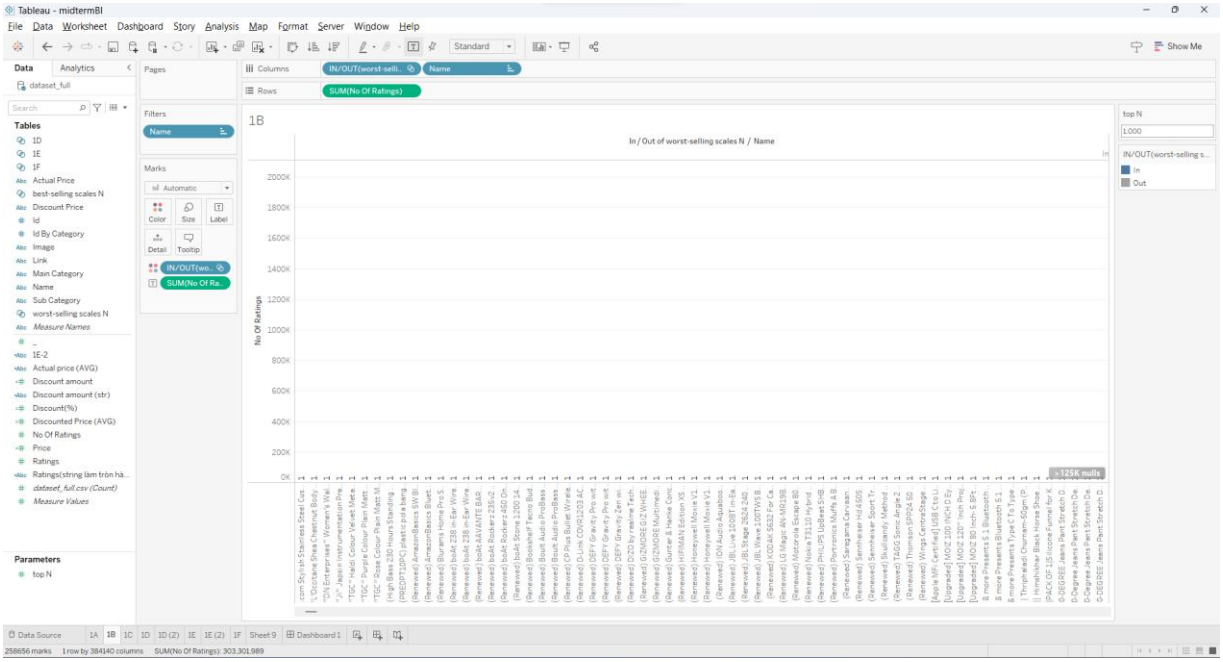
1000,10000.



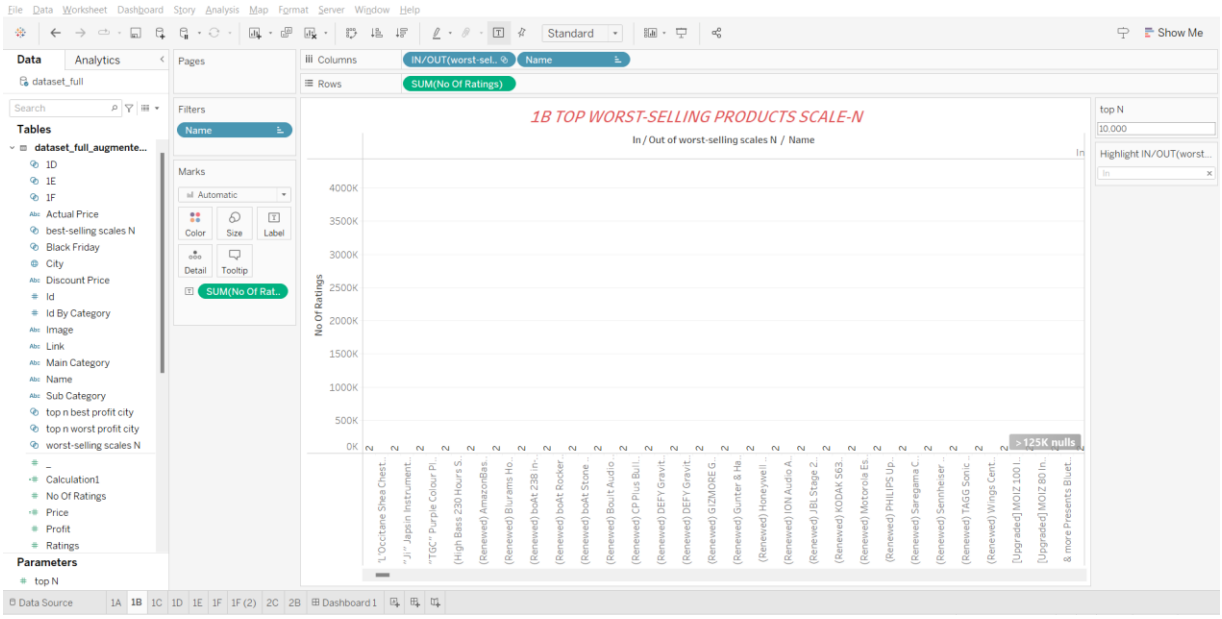
Hình 1. 7 Top 10 worst-selling



Hình 1. 8 Top 100 worst-selling.



Hình 1. 9 Top 1000 worst-selling.



Hình 1. 10 Top 1000 worst-selling.

1.2.4 Problem c.

Variables

Attribute	Variable	Description
Name	numerical variable	The name of the product
Main_category	categorical variable	The main category of the product belong
Sub_category	categorical variable	The sub category of the product belong
Image	No numerical variable	
Link	No numerical variable	
ratings	numerical variable	Ratings score
No of ratings	numerical variable	the number of ratings
Discount_price	numerical variable	the price after discounted
Actual_price	numerical variable	Origin price

Table 1. Dataset variables 1

Find and explain relationship

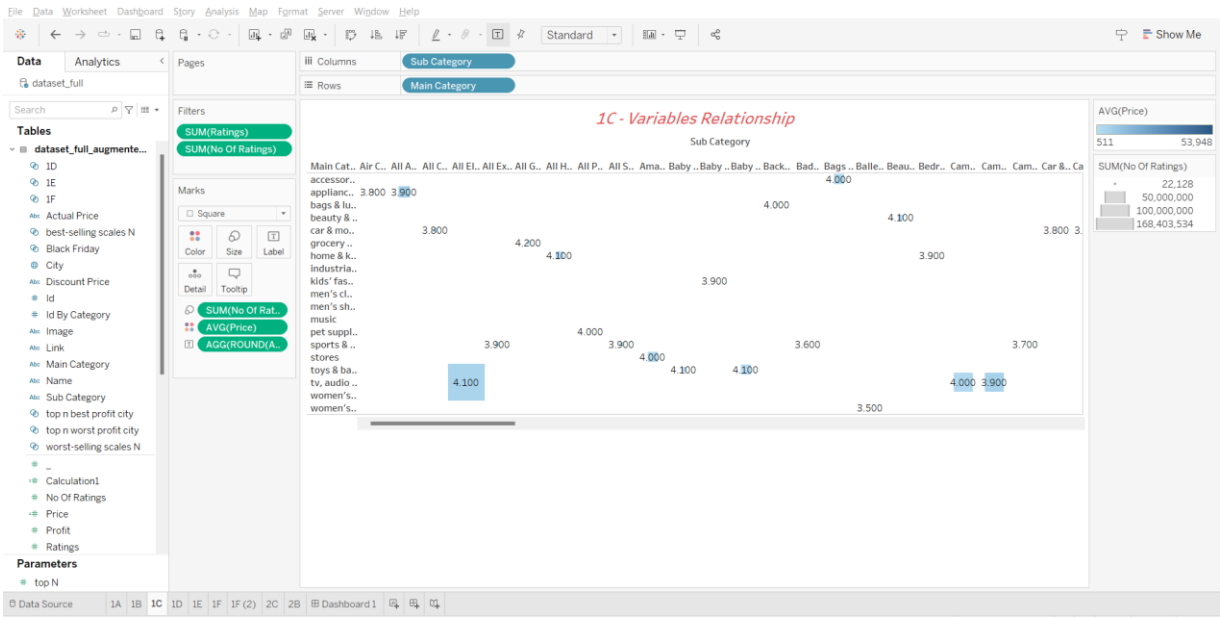
1. Name (numerical variable) and Main_category (categorical variable):
   The variable "Name" represents the name of the product, while "Main_category" indicates the main category to which the product belongs. This relationship allows us to understand the association between the product's name and its main category. Each product has a unique name, and it is categorized into a specific main category.

2. Main_category (categorical variable) and Sub_category (categorical variable):
   The variable "Main_category" represents the main category of the product, while "Sub_category" denotes the sub-category to which the product belongs. This relationship helps us understand the hierarchical structure of the categories. Each product falls under a main category and can further be classified into a specific sub-category within that main category.

3. Ratings (numerical variable) and No of ratings (numerical variable):
   The variable "Ratings" represents the score or rating given to the product, and "No of ratings" indicates the number of ratings received by the product. This relationship enables us to analyze the correlation between the rating score and the number of ratings. Generally, higher-rated products tend to attract more ratings, indicating their popularity or customer satisfaction.

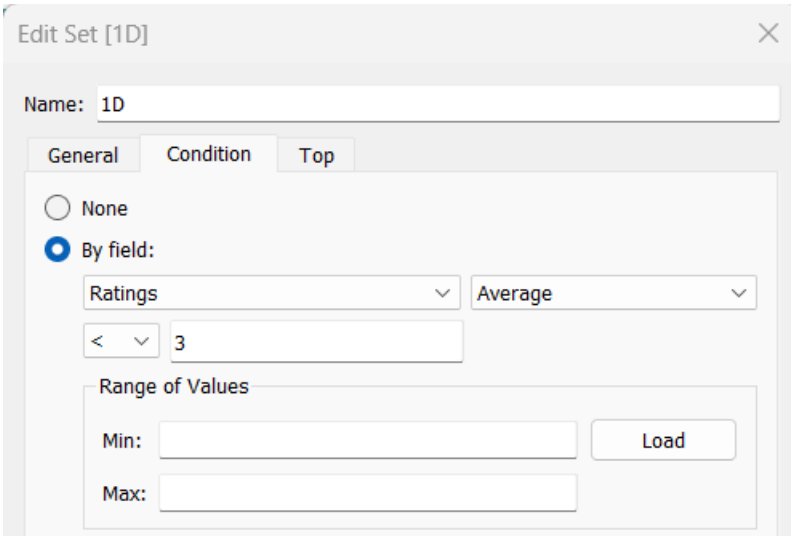
4. Discount_price (numerical variable) and Actual_price (numerical variable):
   The variable "Discount_price" represents the price of the product after applying a discount, while "Actual_price" indicates the original price of the product. This relationship helps us understand the impact of discounts on the product's pricing. The

discount price is derived from the original price by reducing it based on the discount offered.

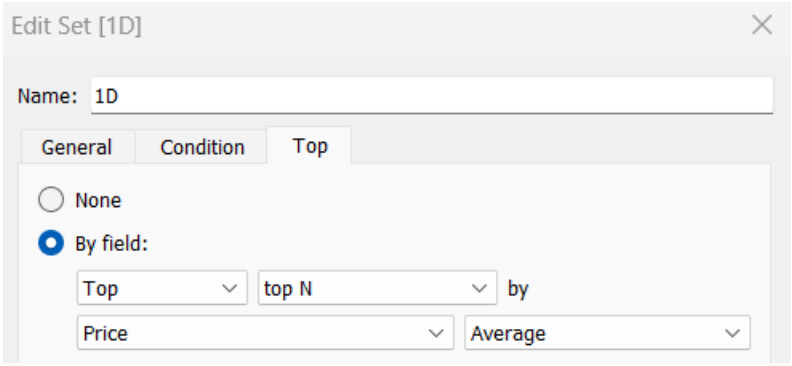


Hình 1. 11 Relationships between variable in dataset.

1.2.5 Problem d.

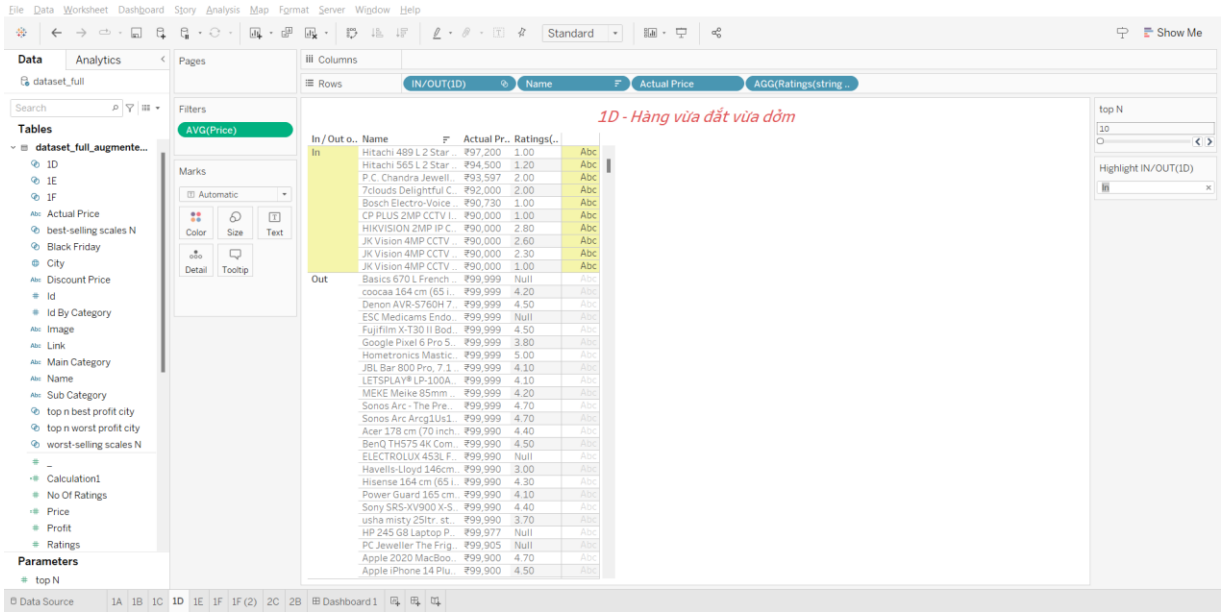


Hình 1. 12 Condition of set 1D.

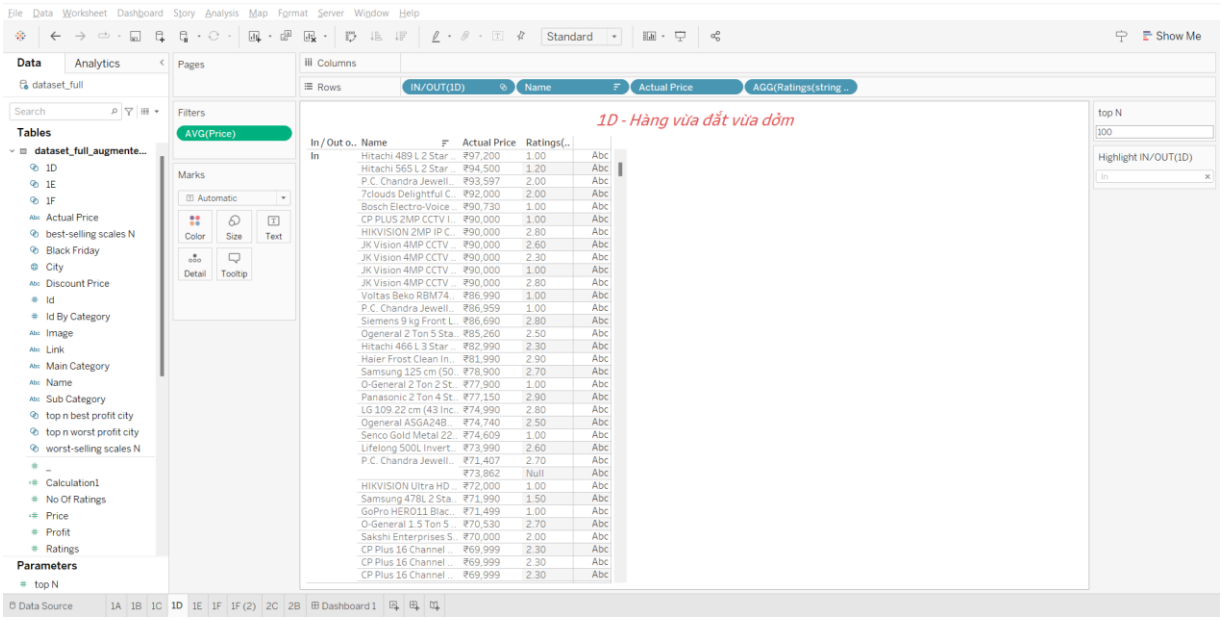


Hình 1. 13 Top of set 1D.

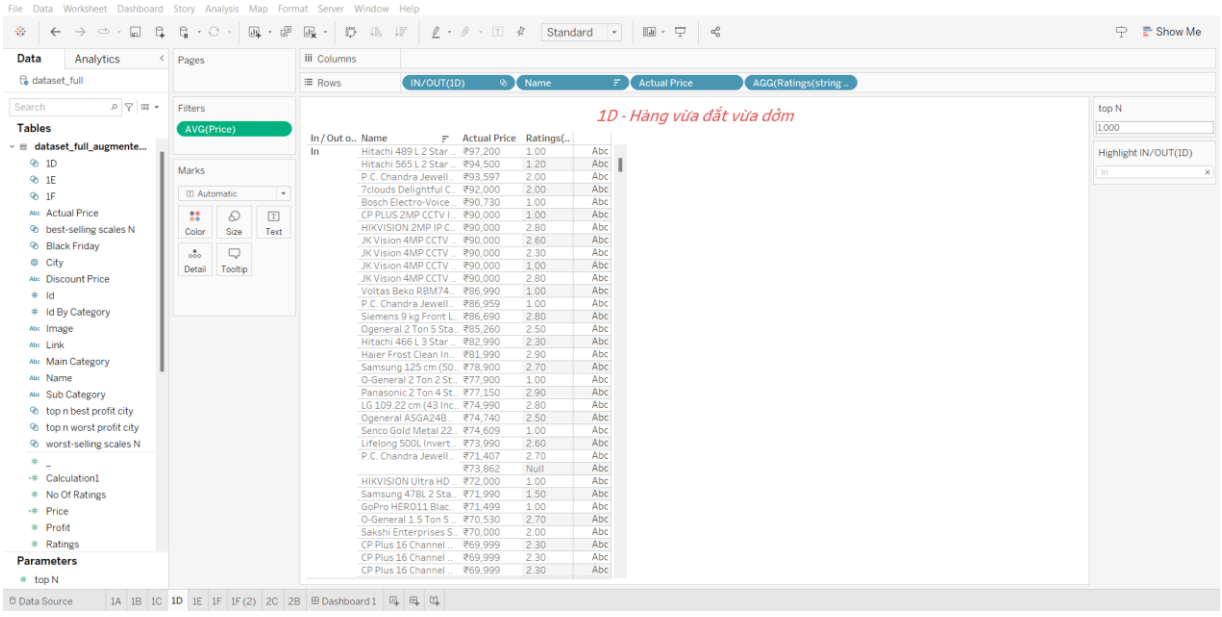
To find products that are most expensive and have a rating below 3.0. Use the Name and Ratings columns from the given data then create a set to set the conditions and list the scale of items.



Hình 1. 14 Top 10 product which are most expensive and ratings lower than 3.0

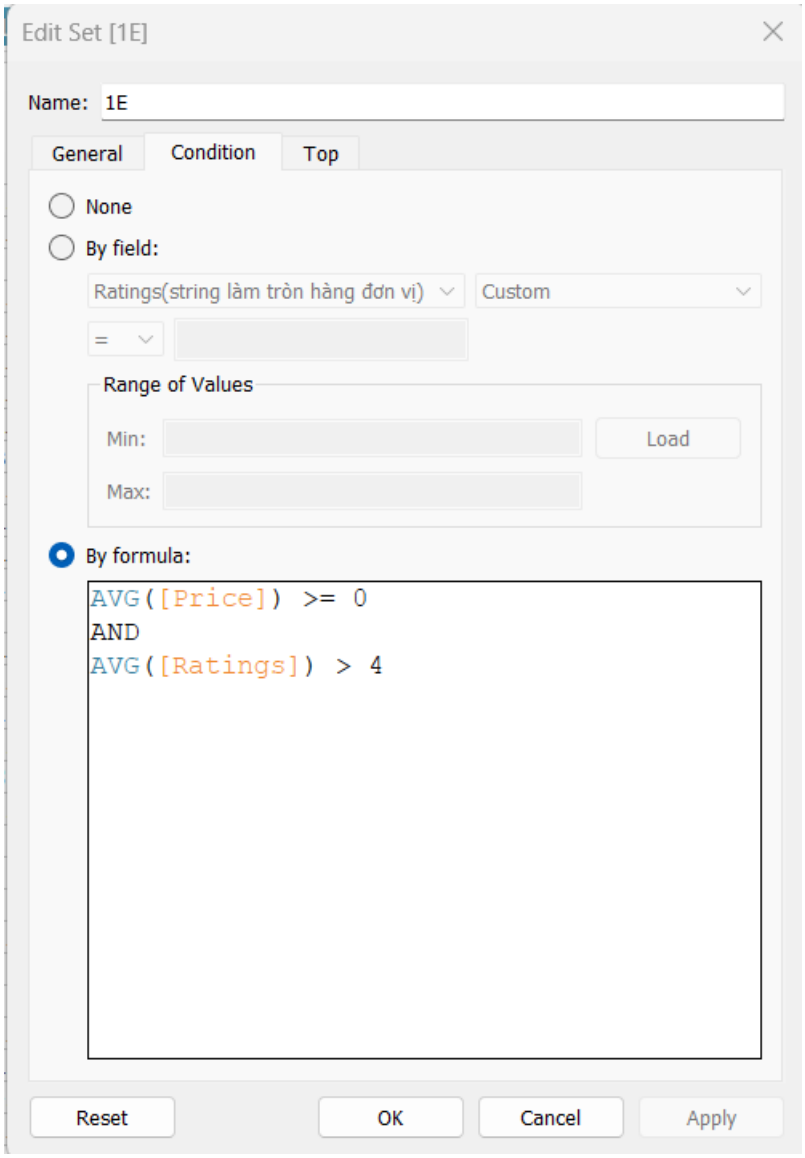


Hình 1. 15 Top 100 product which are most expensive and ratings lower than 3.0

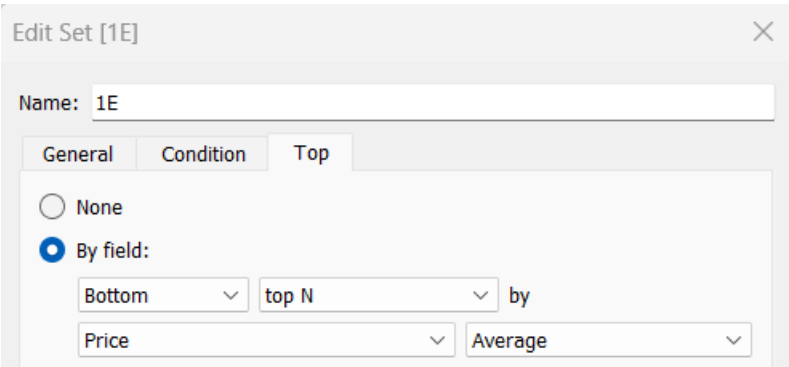


Hình 1. 16 Top 1000 product which are most expensive and ratings lower than 3.0

1.2.6 Problem e.

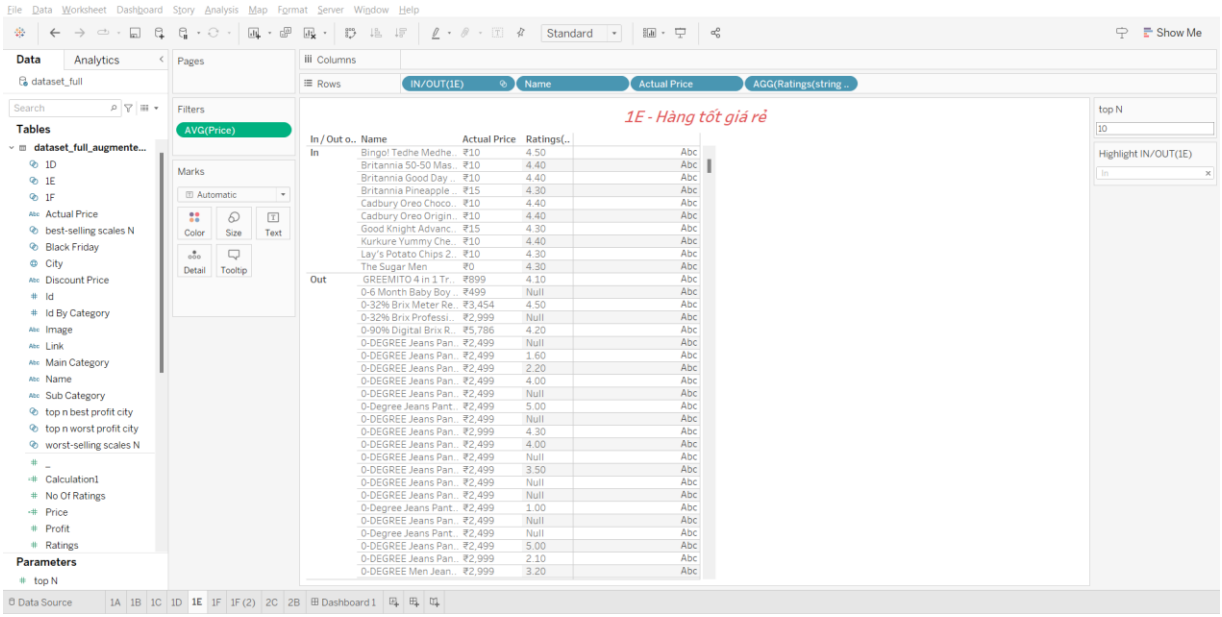


Hình 1. 17 Condition of Set 1E

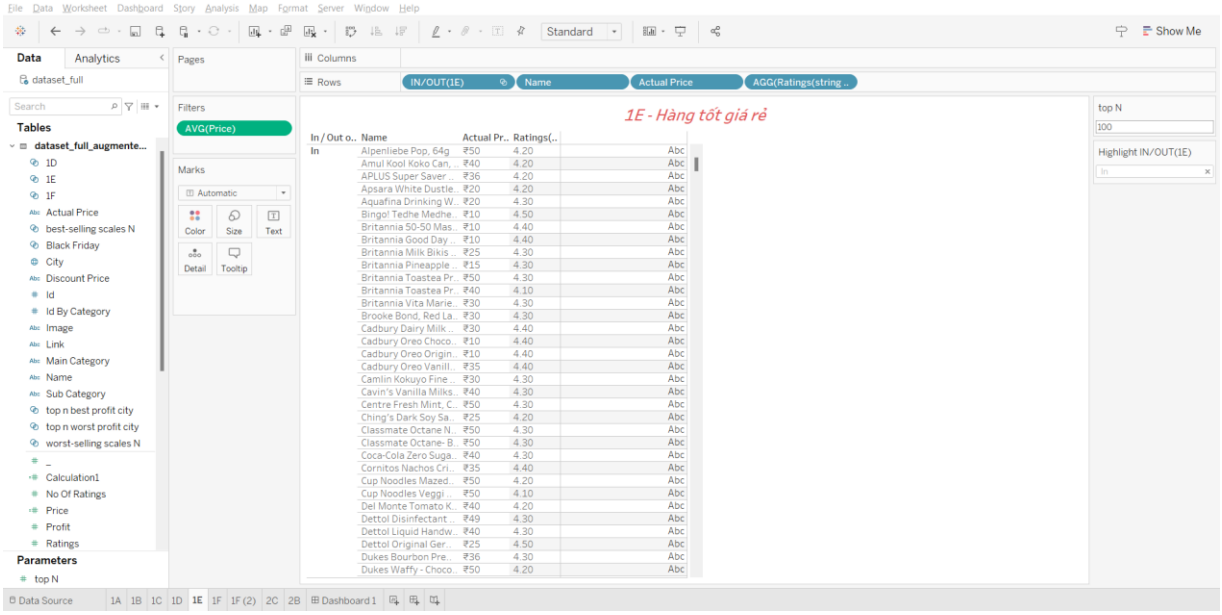


Hình 1. 18 Top of 1E

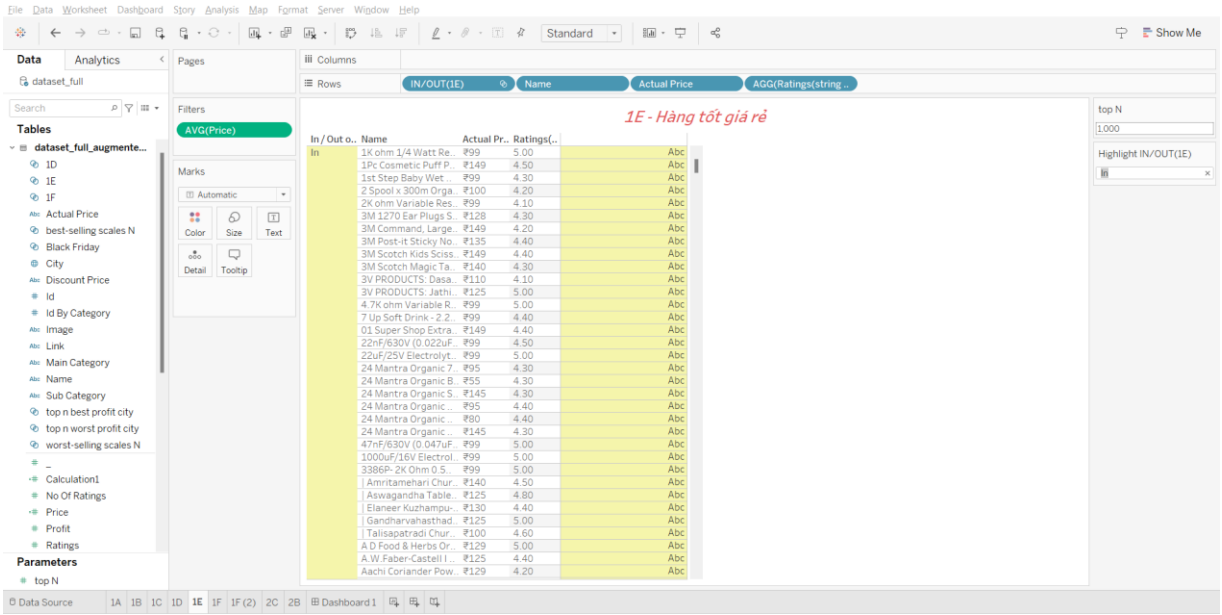
The cheapest products have a rating above 4.0. Use 2 data columns: Ratings and Name for processing.



Hình 1. 19 Top 10 the cheapest product with ratings more than 4.0

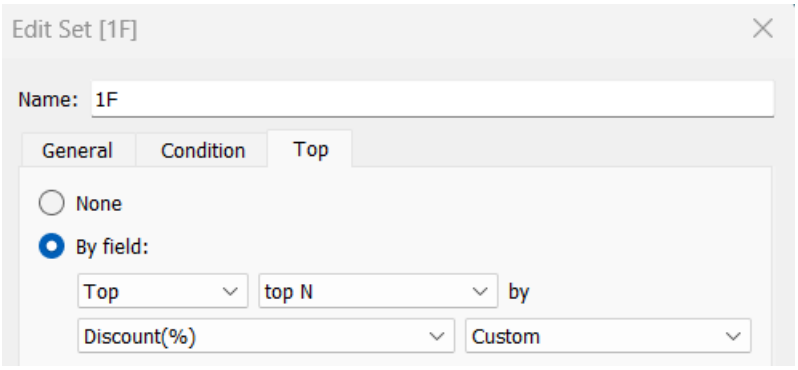


Hình 1. 20 Top 100 the cheapest product with ratings more than 4.0



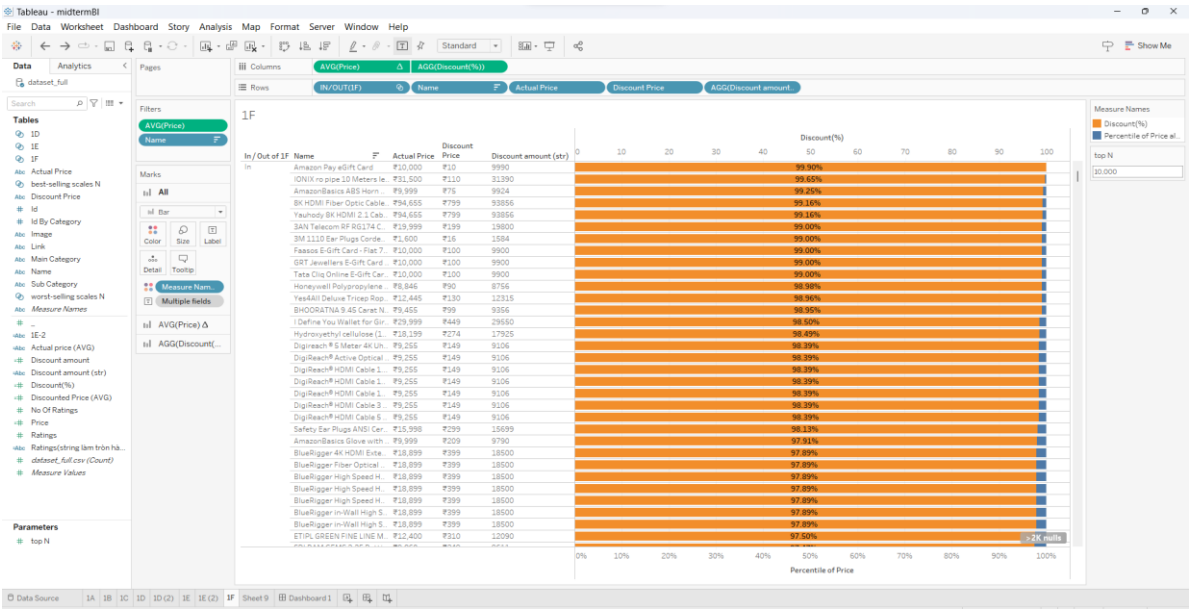
Hình 1. 21 Top 1000 the cheapest product with ratings more than 4.0

1.2.7 Problem f.



Hình 1. 22 Top of set 1F

Products with the greatest preferential prices.



Hình 1. 23 Products with the greatest preferential prices.

QUESTION 2 – AUGUMENTED THE DATASET

2.1 Topic

To employ more power of Tableau software, in this section, you need to augment the given dataset in both *x* and *y* axes. For *x* axis, you will insert two new columns, ie.. "profit" and "city", wheren"profit" indicates the number of products which were sold in the corresponding "city" of US. For *y* axis, it's quite easy to imagine, you need to duplicate each instance of the dataset to at least 100 new ones, and randomly generate the "profit" and "city" values for each of them. Table 2 illustratesthe new dataset after augmenting in both *x* and *y* axes. You should note that the "id" and "id by category" columns need to be altered to make sure these identification numbers are unique.

Table 2. Illustrate a new dataset after augmenting *x* and *y* axes

id	*name*	*...*	*profit*	*city*
0	Lloyd 1.5 Ton 3 Star Inverter Split	…	126	South Bend
1	Lloyd 1.5 Ton 3 Star Inverter Split	…	32	Springfield
2	Lloyd 1.5 Ton 3 Star Inverter Split	…	147	Savannah
3	Lloyd 1.5 Ton 3 Star Inverter Split	…	296	Knoxville
4	Lloyd 1.5 Ton 3 Star Inverter Split	…	13	Salt Lake City
5	Lloyd 1.5 Ton 3 Star Inverter Split	…	58	Cincinnati

The dataset after augmenting is now having enough information to support the decision-making activities. You need to answer the following questions:

a. Find and explain the relationships between dataset variables
b. Get a list of products that should be included in a promotion campaign based on the sellingprofits by city. For example, you can define a measure such as "if a product has a goodaselling profit at a city, it should be discounted as *x* percent to improve its profit", or "we should discount that jacket's price because it is going to be outdated soon". Whatever the measure you define, you need to define a threshold, for example, to determine whether aproduct has a good selling profit or not.
c. Visualize the best-and-worst-selling-profit on world map.

2.2 Answer

2.2.1 Problem a

Variables

Attribute	Variable	Description
Name	numerical variable	The name of the product
Main_category	categorical variable	The main category of the product belong
Sub_category	categorical variable	The sub category of the product belong
Image	No numerical variable	
Link	No numerical variable	
ratings	numerical variable	Ratings score
No of ratings	numerical variable	the number of ratings
Discount_price	numerical variable	the price after discounted
Actual_price	numerical variable	Origin price
Profit	numerical variable	The profit of the product
City	numerical variable	The city where the product be sold

Table 2. Dataset variables 2.

Find and explain the relationship:

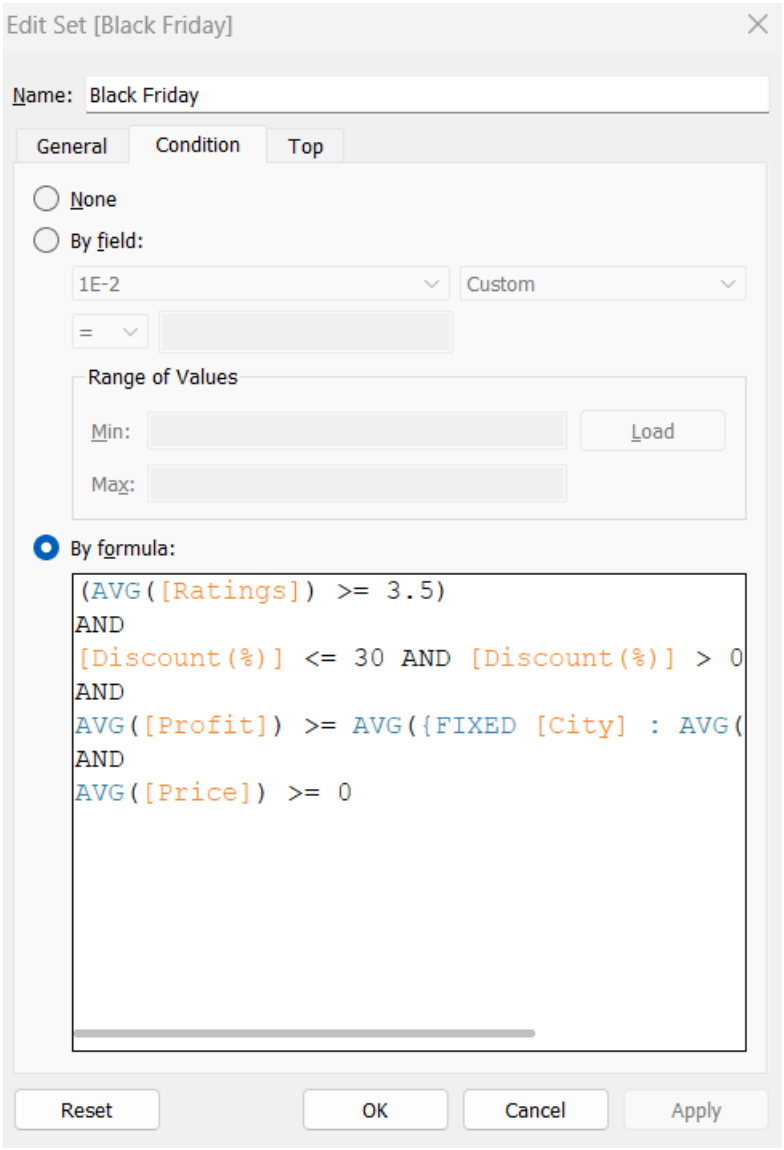
1. Profit and Main_category: Different main categories may have varying profitability levels.
2. Profit and Sub_category: Profitability can differ across sub-categories within a main category.
3. Profit and Ratings: Higher ratings may indicate increased profitability, but other factors also influence it.
4. Profit and No of ratings: More ratings may suggest higher profitability, considering other factors.
5. Profit and Discount_price: Discounts can attract more customers, affecting profitability.
6. Profit and City: Profitability can vary across cities due to market factors and preferences.

7. City and Main_category: Different cities may have varying preferences for specific product categories.
8. City and Sub_category: Sub-categories may have different popularity levels across cities.
9. City and Ratings: Customer ratings may vary across different cities, indicating regional sentiment.
10. City and No of ratings: The number of ratings may differ based on city population and market activity.

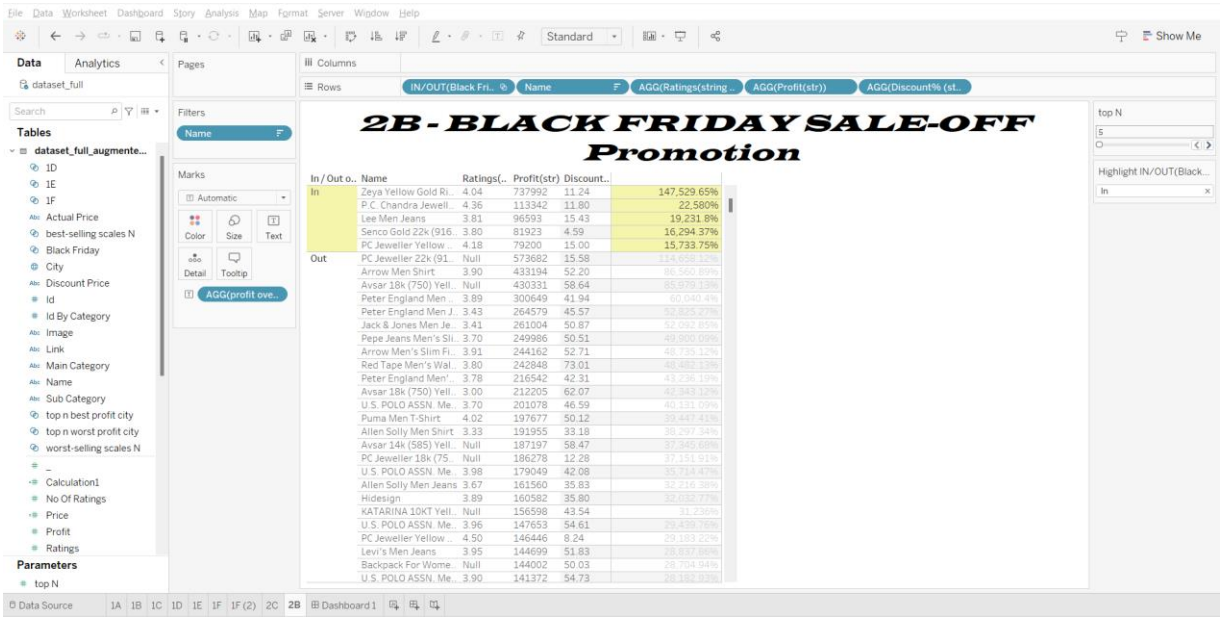
2.2.2 Problem b

Promotion details:

- Name: Black Friday.
- Condition: The products with ratings higher than 3.5, a discount rate equal to or less than 30%, and profits exceeding the average profit of the city where they are sold.
- Describe: Discounting popular and high-profit products in cities while ensuring that these products have not been discounted by more than 30% of their original value.

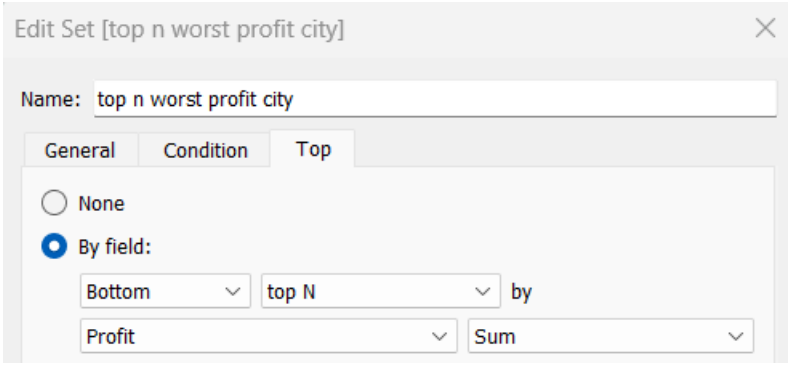


Hình 1. 24 The condition of promotion.

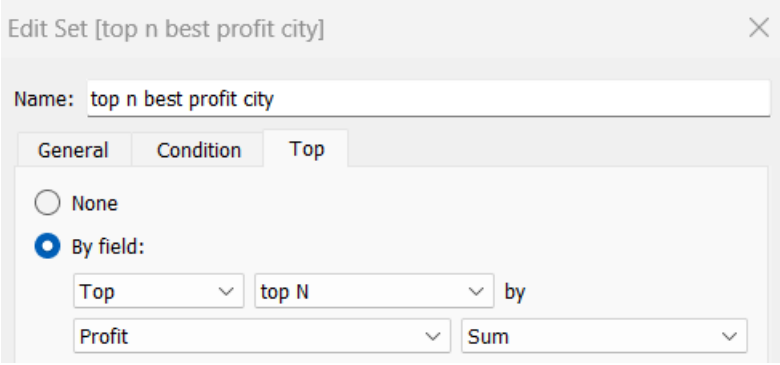


Hình 1. 25 Top products should be appear in a promotion.

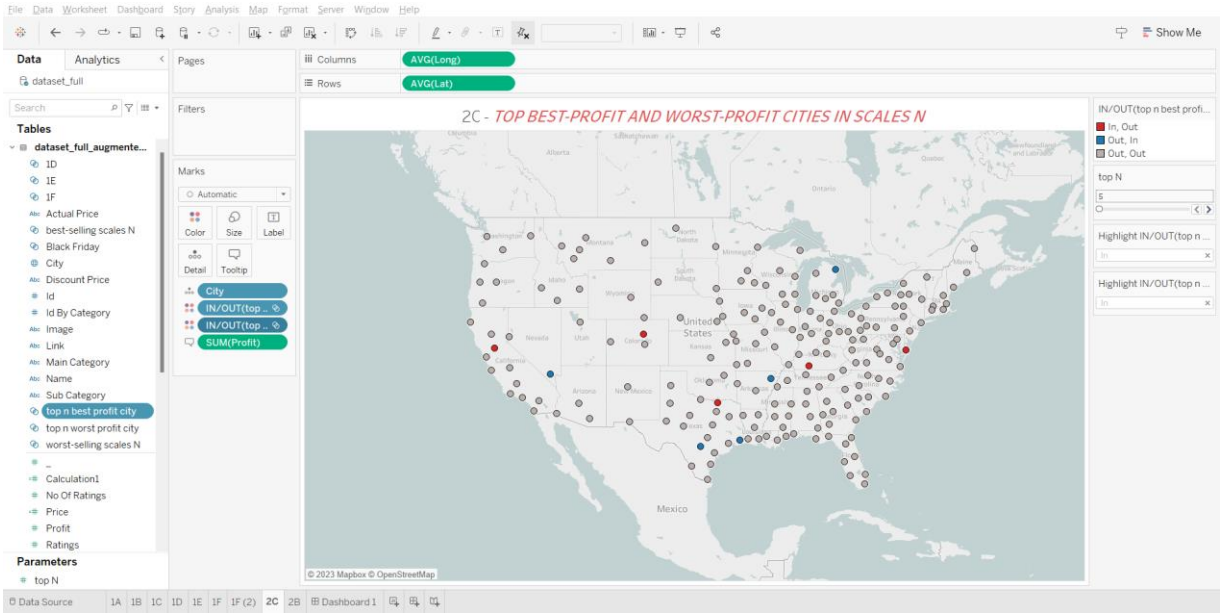
2.2.3 Problem c



Hình 1. 26 Top n worst profit city



Hình 1. 27 Top n best profit city



Hình 1. 28 Top best-profit and worst-profit cities in scales N

TÀI LIỆU THAM KHẢO

Tiếng Anh

[1] Midterm Project Description - https://cs504049.fastai.dev/midterm.html

[2] pandas.read_csv-https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html

[3] Exploratory data analysis - https://en.wikipedia.org/wiki/Exploratory_data_analysis