TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO GIỮA KÌ**

**HỆ THỐNG THƯƠNG MẠI THÔNG MINH**

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

*Người thực hiện*: **ĐÀM THỊ PHƯỢNG – 52100834**

**PHẠM PHƯƠNG NGỌC – 52100566**

Khoá  **: 25**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO GIỮA KÌ**

**HỆ THỐNG THƯƠNG MẠI THÔNG MINH**

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

*Người thực hiện*: **ĐÀM THỊ PHƯỢNG – 52100834**

**PHẠM PHƯƠNG NGỌC – 52100566**

Khoá  **: 25**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

**LỜI CẢM ƠN**

Chúng em xin trân trọng cảm ơn giảng viên Dương Hữu Phúc- người đã trực tiếp chỉ bảo, hướng dẫn nhóm chúng em trong quá trình hoàn thành dự án giữa kì này.Mặc dù đã có những đầu tư nhất định trong quá trình làm bài song cũng khó có thể tránh khỏi những sai sót, em kính mong nhận được ý kiến đóng góp của quý thầy cô để bài dự án được hoàn thiện hơn.

Em xin chân thành cảm ơn!

**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 Thầy 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)*

*Phượng*

*Đàm Thị Phượng*

*Ngọc*

*Phạm Phương Ngọc*

**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)

**TÓM TẮT**

Báo cáo giữa kỳ dựa trên dữ liệu bán hàng từ trang Amazon.com. Báo cáo được chia làm hai phần chính.

Phần đầu là phân tích dữ liệu EDA, bao gồm dữ liệu tiền xử lý và xác định sản phẩm bán chạy nhất và bán chậm nhất

Phần thứ hai liên qua đến việc tăng cường dữ liệu theo trục x,y và giải quyết vấn đề liên quan tới chiết khấu sản phẩm

**MỤC LỤC**

# CHƯƠNG 1 –XỬ LÝ DỮ LIỆU

1. **Làm sạch dữ liệu**

* Ngôn ngữ lập trình : Python
* Thư viện: pandas
* import pandas as pd
* # Function to clean the dataset
* def clean\_dataset(dataset\_full):
* # Load the dataset
* data = pd.read\_csv(dataset\_full, encoding='utf-8')
* # Kiểm tra cấu trúc, định dạng dữ liệu
* data.info()
* data.head()
* # Kiểm tra và xóa dòng trùng lặp(nếu có)
* if data.duplicated().sum() > 0:
* data = data.drop\_duplicates()
* print("Sum of duplicated rows deleted: ", data.duplicated().sum())#In ra số lượng dòng trùng đã xóa
* # Chuyển 'no\_of\_ratings' sang interger
* data['no\_of\_ratings'] = data['no\_of\_ratings'].str.replace(',', '', regex=True)#bỏ dấu "," thay giá trị không hợp lệ NaN
* data['no\_of\_ratings'] = pd.to\_numeric(data['no\_of\_ratings'], errors='coerce')
* # Convert 'ratings' column to float, replacing invalid values with NaN
* data['ratings'] = pd.to\_numeric(data['ratings'], errors='coerce')
* # Thay giá trị in colum "ratings"sang giá trị TB của cột
* mean\_rating = data['ratings'].mean()
* data['ratings'].fillna(mean\_rating, inplace=True)
* # Tạo tập con invalid\_ratings để lưu các dòng trong cột ratings có giá trị NaN(dòng không thể chuyển qua kiểu float)
* invalid\_ratings = data[data['ratings'].isna()]
* # Loại bỏ ký tự đặc biệt và dấu"," ở cột 'actual\_price' and 'discount\_price' và chuyển sang kiểu float thay giá trị không hợp lệ bằng NaN
* data['actual\_price'] = data['actual\_price'].str.replace("₹", '').str.replace(",", '')
* data['discount\_price'] = data['discount\_price'].str.replace("₹", '').str.replace(",", '')
* data['actual\_price'] = pd.to\_numeric(data['actual\_price'], errors='coerce')
* data['discount\_price'] = pd.to\_numeric(data['discount\_price'], errors='coerce')
* # Điền giá trị NaN của 'actual\_price' and 'discount\_price' bằng giá trị trung bình của cột
* data['actual\_price'].fillna(data['actual\_price'].mean(), inplace=True)
* data['discount\_price'].fillna(data['discount\_price'].mean(), inplace=True)
* # Fill NaN values in 'no\_of\_ratings' bằng giá trị trung bình của cột
* mean\_no\_of\_ratings = data['no\_of\_ratings'].mean()
* data['no\_of\_ratings'].fillna(mean\_no\_of\_ratings, inplace=True)
* # Lưu files mới (định dạng: tên file cũ + \_cleaned)
* output\_file = dataset\_full.replace('.csv', '\_cleaned.csv')
* data.to\_csv(output\_file, index=False)
* #in rhông báo dữ liệu đã được lưu ở đâu
* print("Cleaned data saved to:", output\_file)
* return data, output\_file #trả về dữ liệu đã làm sạch + tên file mới
* # Đường dẫn files:
* dataset\_full = 'D:\\BI\\midterm-dataset\\dataset\_full.csv'  # Ensure the path is correct
* cleaned\_data, new\_dataset\_file = clean\_dataset(dataset\_full)

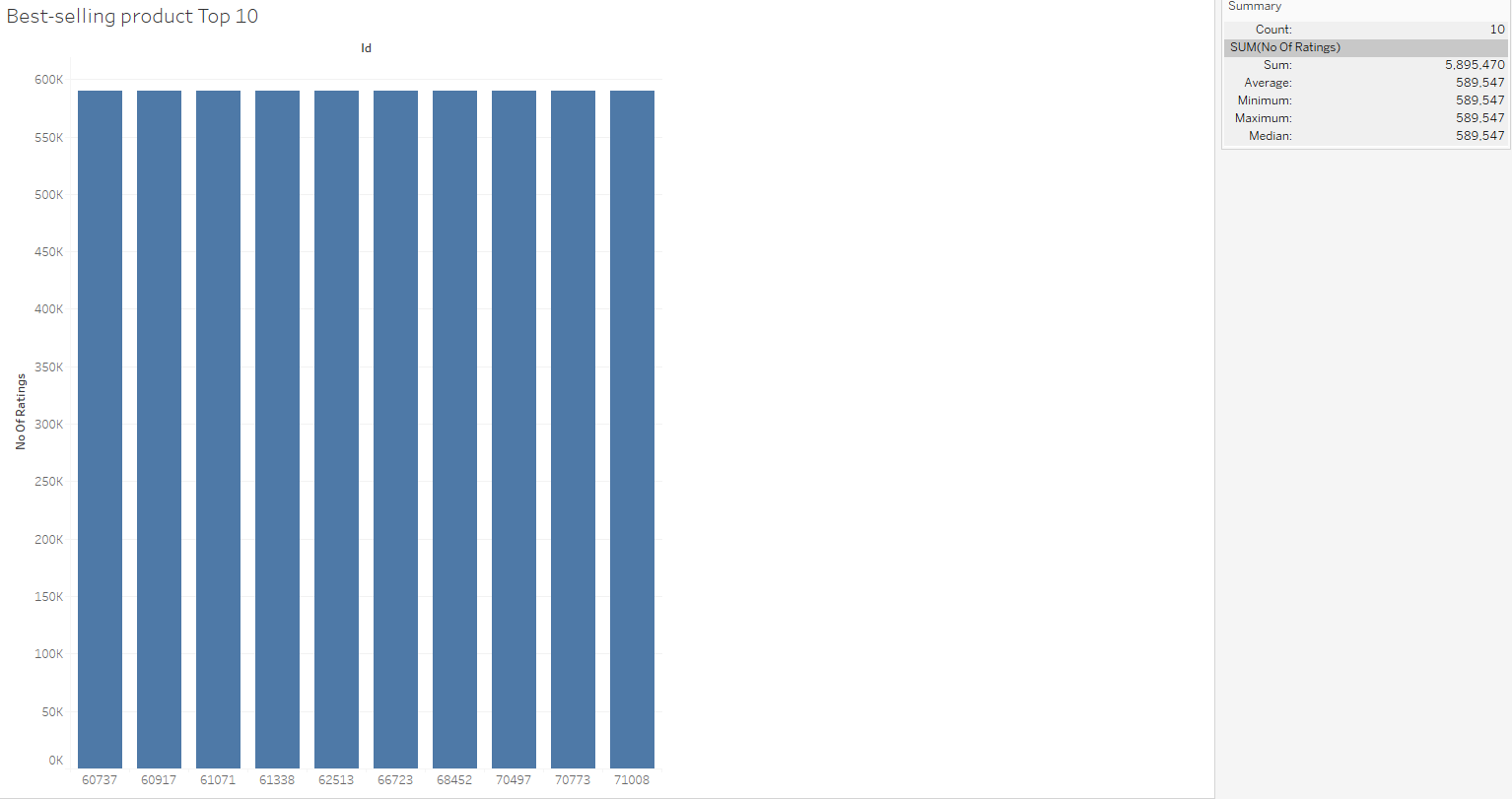
- Mô tả: Thêm thư viện pandas và dữ liệu từ têpk csv vào DataFrame có tên là “data”. Sử dụng hàm info (),để kiểm tra, hiển thị thông tin về dữ liệu bao gồm: số hàng, số cột và kiểu dữ liệu mỗi cột. Sau đó sử dụng hàm head () để xem 5 hàng đầu tiên của dữ liệu, nhằm kiểm tra cấu trúc và dữ liệu mẫu. Đoạn mã tiếp theo kiểm tra và loại bỏ hàng trùng lặp trong tệp dữ liệu (nếu có) và in ra số hàng, cột đã xóa. Cột “ no\_of\_ratings” được chuyển đổi bằng cách xoá dấu phẩy và chuyển đổi thành kiểu “interger” để chuẩn hóa dữ liệu. Cột “ratings” được chuyển đổi thành kiểu “float”, với những giá trị không hợp lệ được thay thế bằng NaN(). Đoạn tiếp theo, tính toán giá trị trung bình cột ”ratings” bằng cách xóa các giá trị NaN và thay bằng giá trị trung bình đã được tính toán. Mã kiểm tra các hàng trong tệp dữ liệu của cột “ratings” không biển đổi được thành kiểu”float” và lưu trữ chúng trong biển ‘invalid\_ratings’. Các cột ‘actual\_price’ và ‘discount\_price’ được chuyển đổi bằng cách xóa ký hiệu tiền tệ và dấu phẩy, sau đó chuyển thành kiểu float. Mã tính giá trị trung bình của các cột “actual\_price” và “discount\_price” thay bằng giá trị NaN trong 2 cột này bằng giá trị trung bình tương ứng. Giá trị NaN trong cột “no\_of\_ratings” được thay thế bằng giá trị trung bình của cột này. Cuối cùng, dữ liệu đã xử lý được lưu vào một tệp csv mới có tên “dataset\_full\_cleaned.csv”.

1. **Đánh giá mức độ phức tạp**

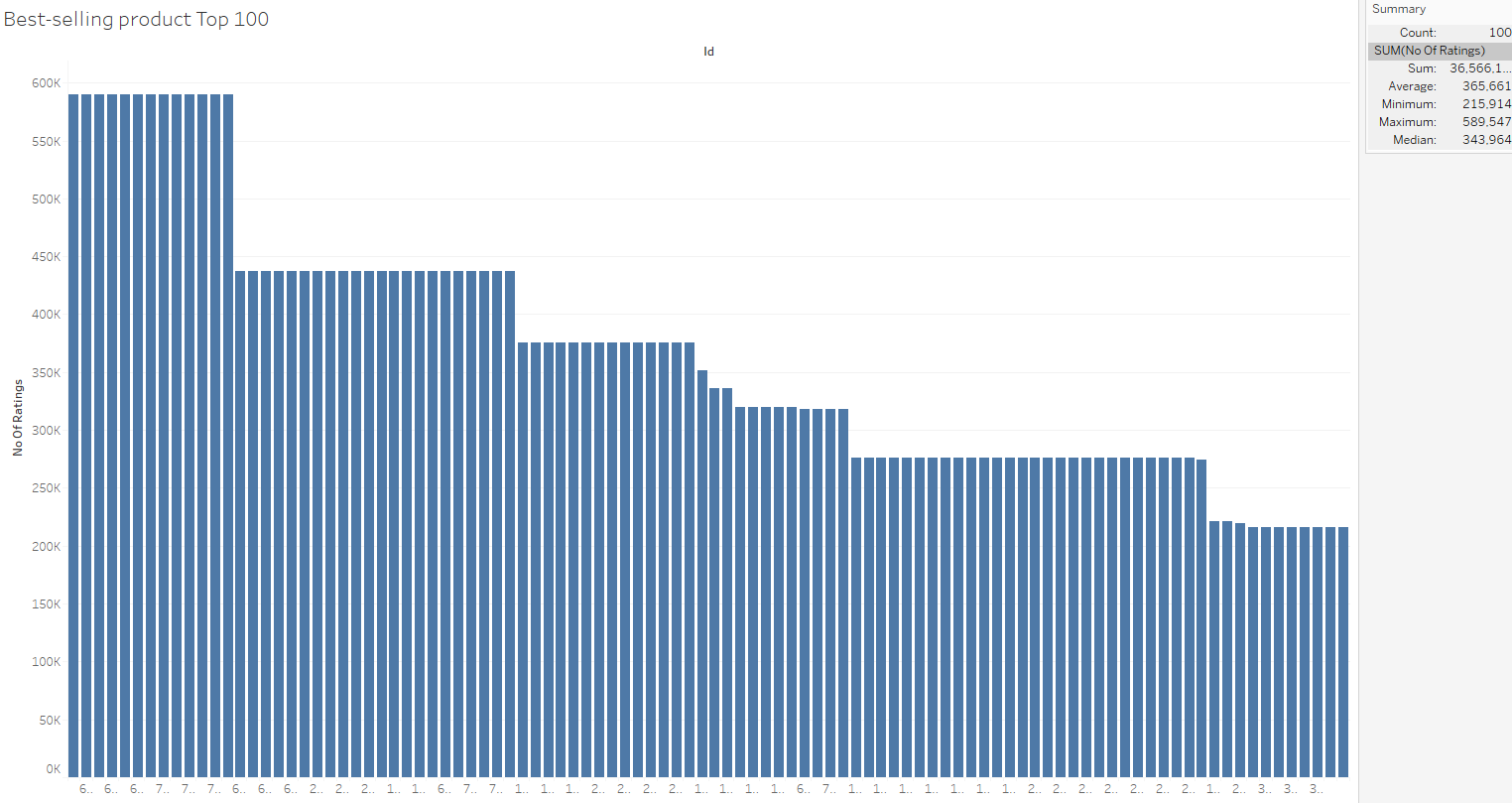
* Mức độ phức tập cho dataset này ở mức trung bình với 12 cột và 551.585 dòng. Với tập dữ liệu lớn, kiểu dữ liệu đa dạng bao gồm kiểu dữ liệu số, văn bản, và giá trị bị thiếu trong một cột, nên cần chú ý chi tiết và để đảm bảo điền dữ liệu đúng. Các bước thực hiện bao gồm xóa ký hiệu tiền tệ, chuyển đổi kiểu dữ liệu, điền các giá trị còn thiếu bằng giá trị trung bình, xử lý giá trị trùng lặp, dữ liệu không quá phức tạp nhưng cần chú ý chi tiết và phân tích xử lý chính xác

# CHƯƠNG 2 – PHÂN TÍCH DỮ LIỆU

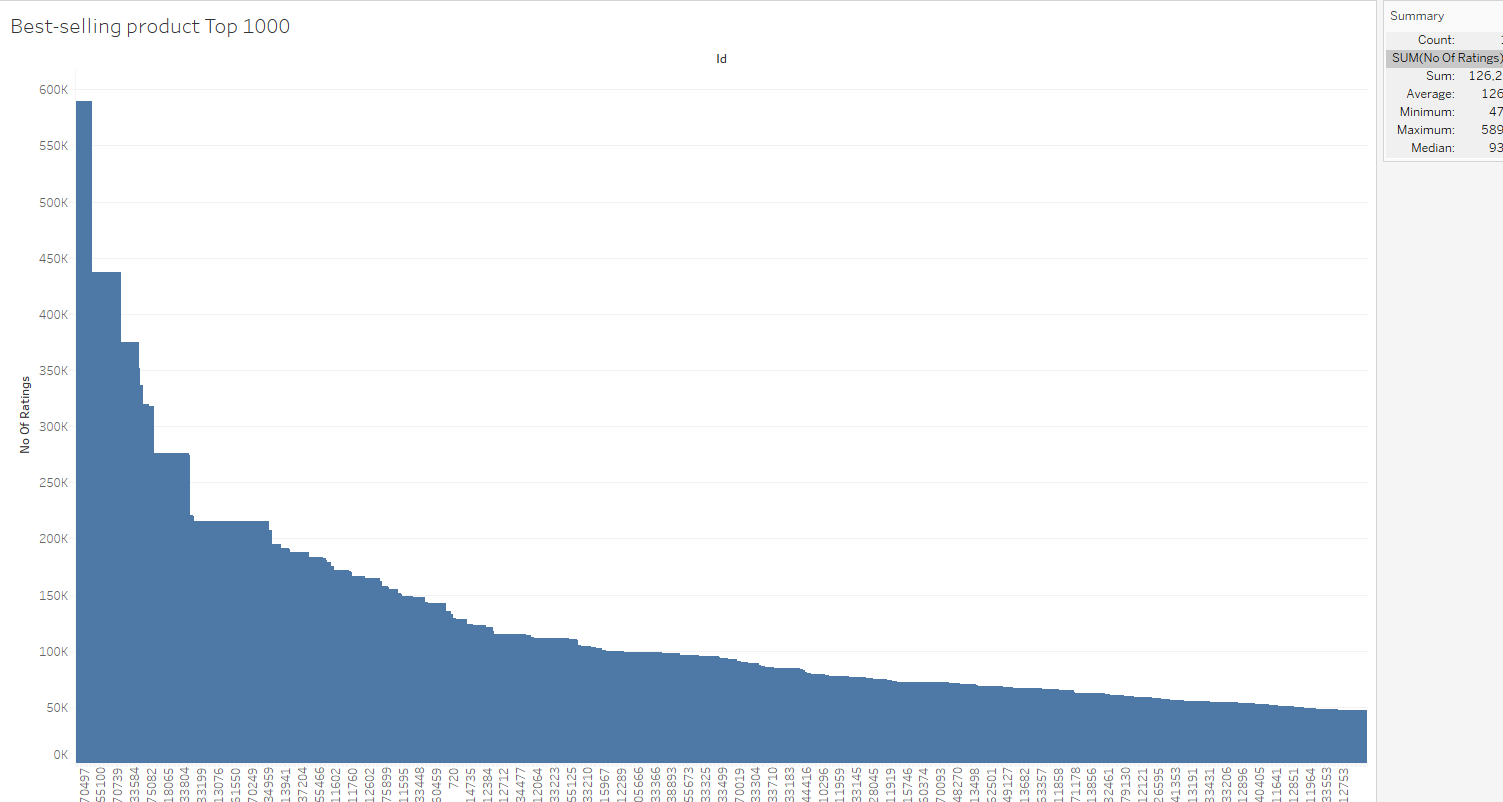
* 1. **Tìm sản phẩm bán chạy nhất theo thang điểm 10, 100, 1K, 10K mặt hàng và hình dung chúng**



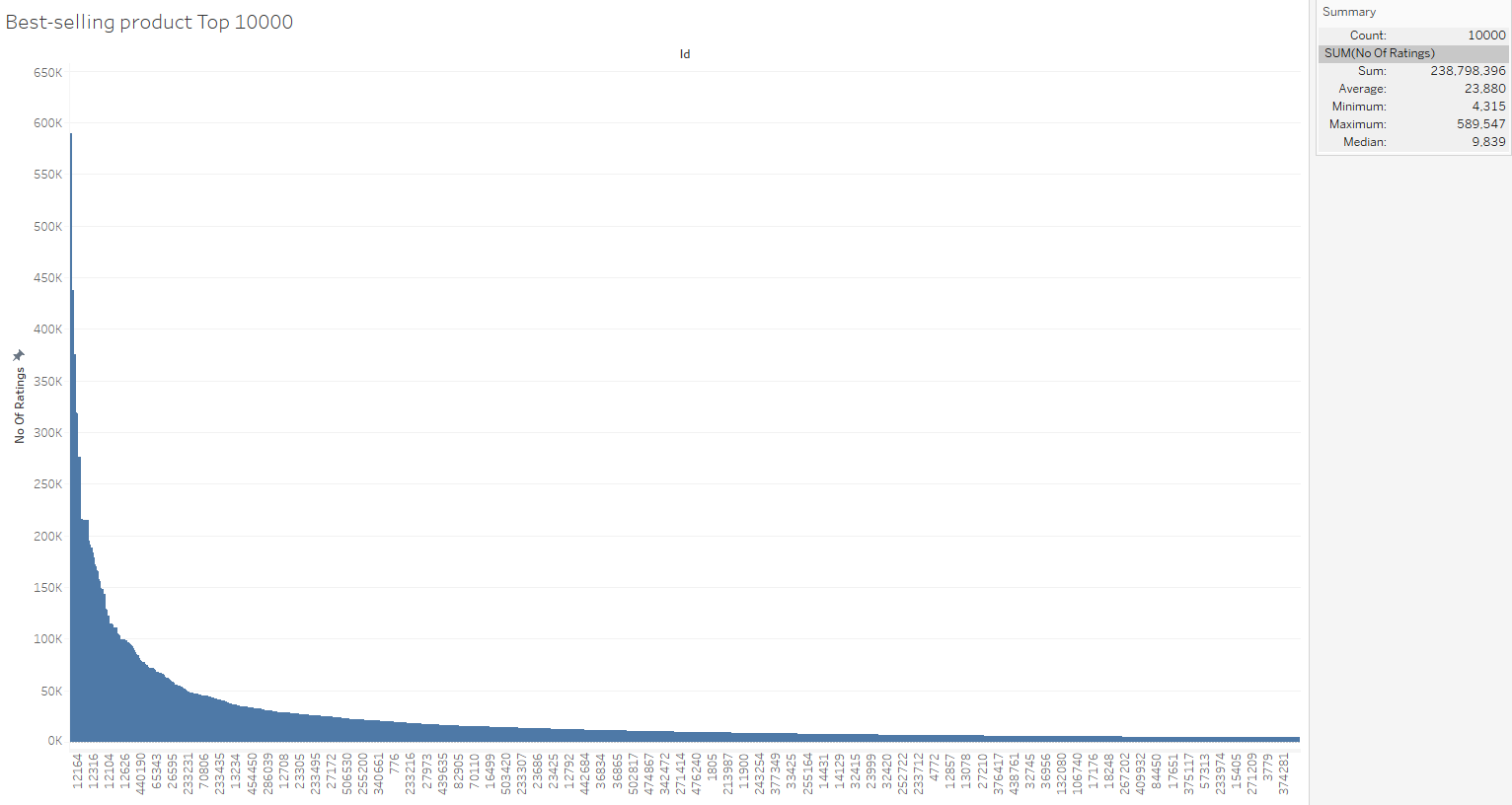
Hình 2.1.1: Hiển thị 10 sản phẩm bán chạy nhất



Hình 2.1.2: Hiển thị 100 sản phẩm bán chạy nhất

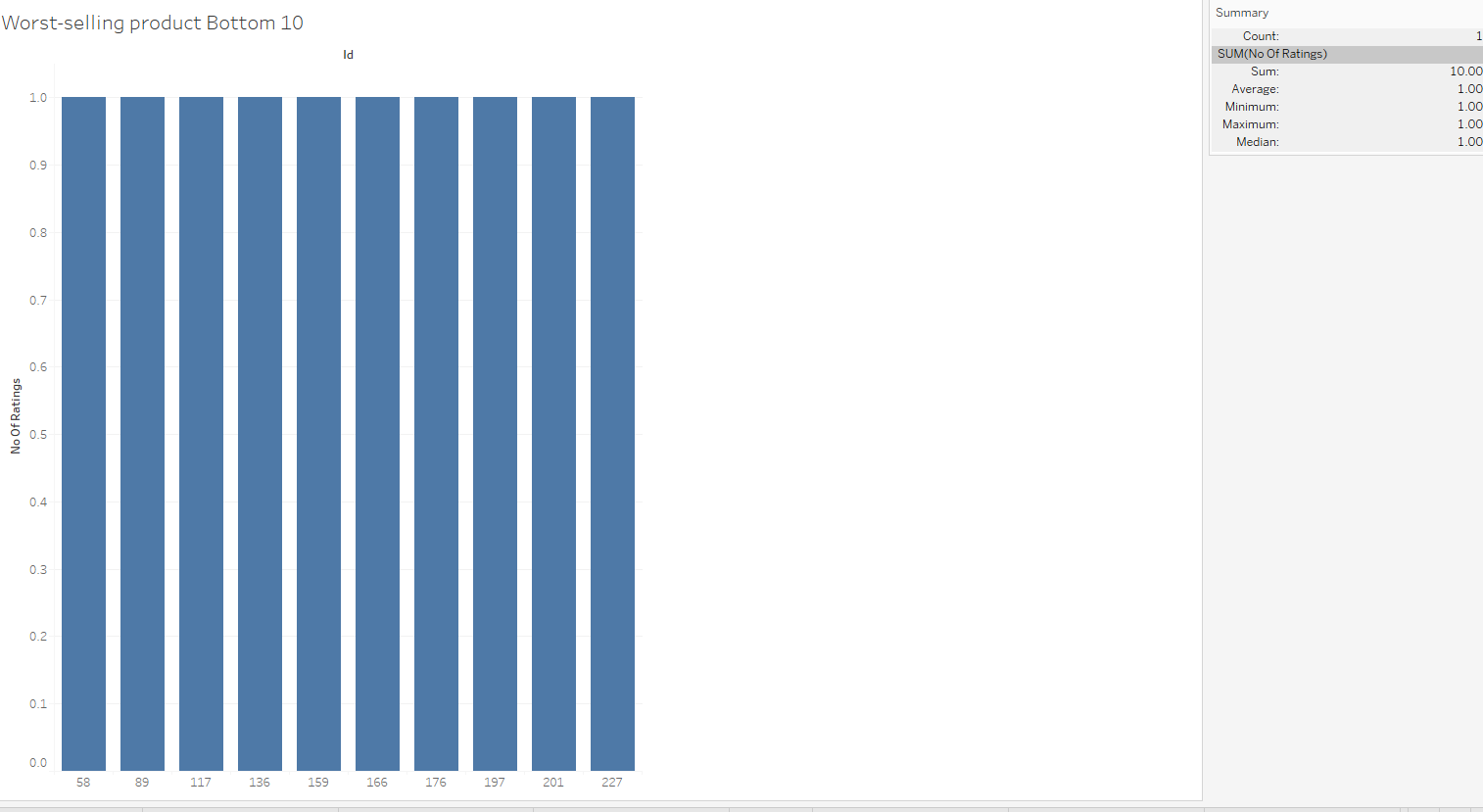


Hình 2.1.3: Hiển thị 1000 sản phẩm bán chạy nhất

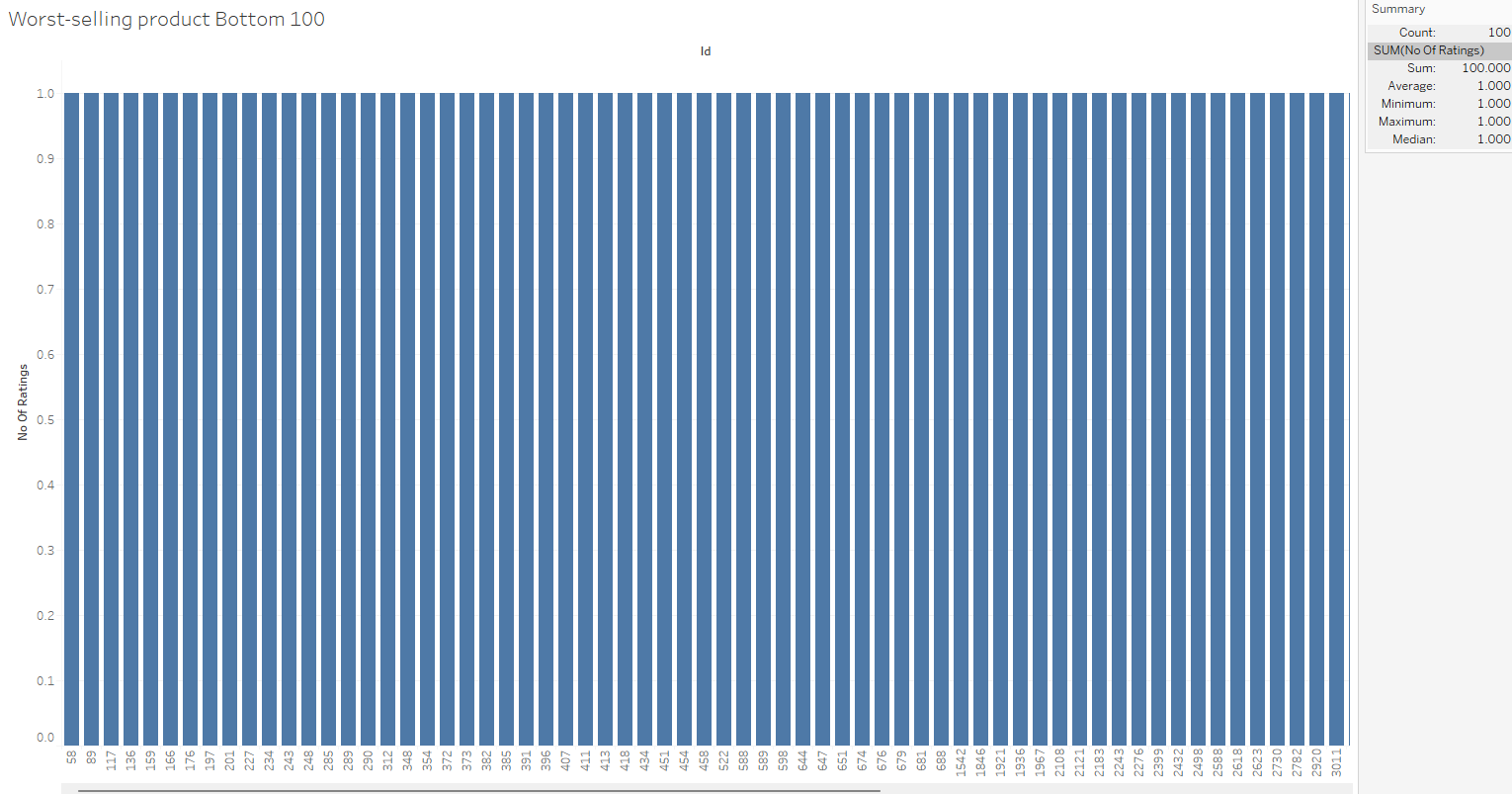


Hình 2.1.4: Hiển thị 10000 sản phẩm bán chạy nhất

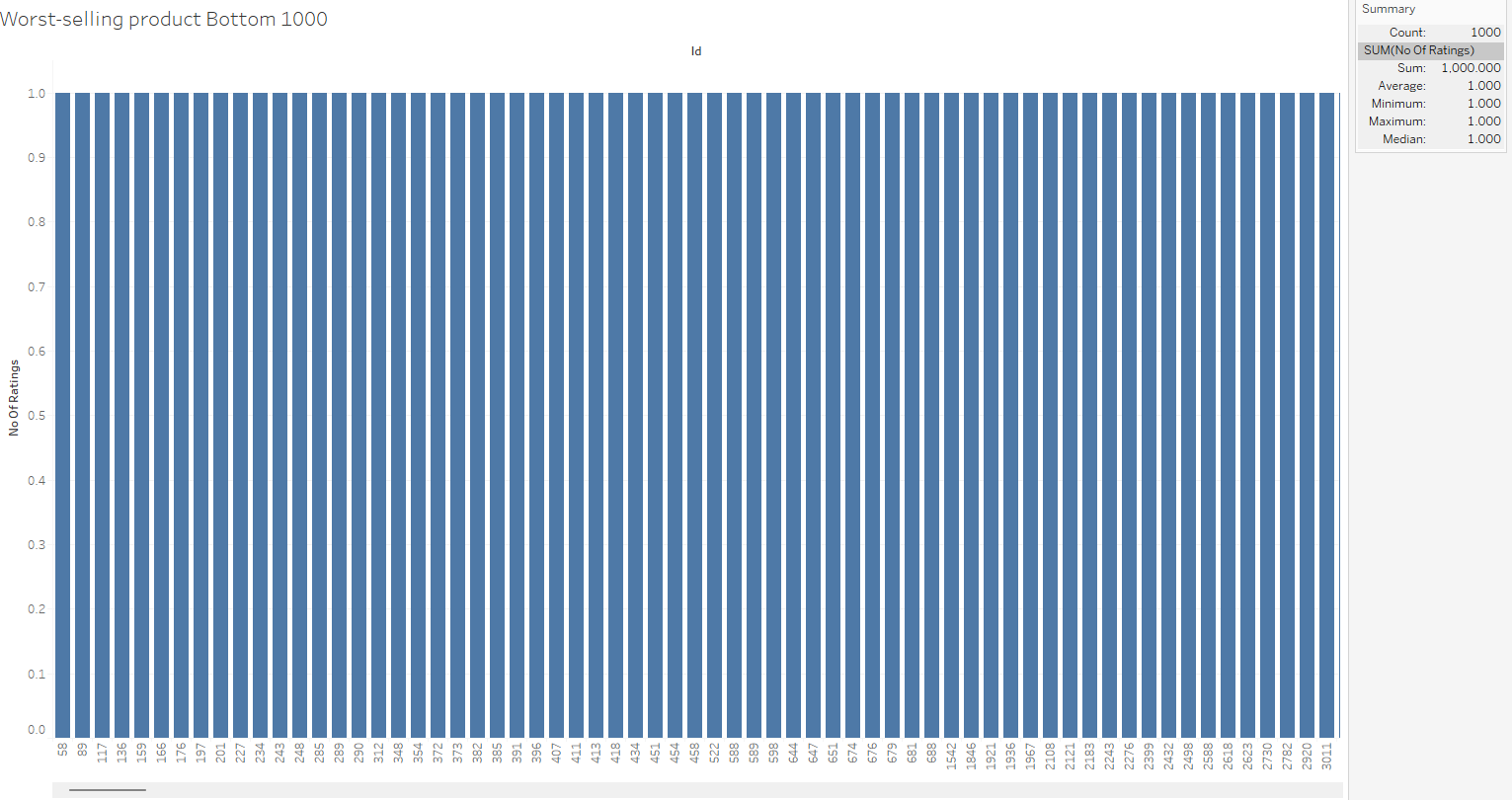
* Nhận xét về biểu đồ: Biểu đồ cho chúng ta thấy sự phân bố không đồng đều của tệp dữ liệu giữa các biển thể. Có sự chênh lệch lớn về doanh số bán ra giữa các sản phẩm(có sản phẩm có doanh số rất cao và có sản phẩm doanh số rất thấp)
  1. **Tìm những sản phẩm bán tệ nhất theo thang điểm 10, 100, 1K, 10K mặt hàng và hình dung chúng**

****

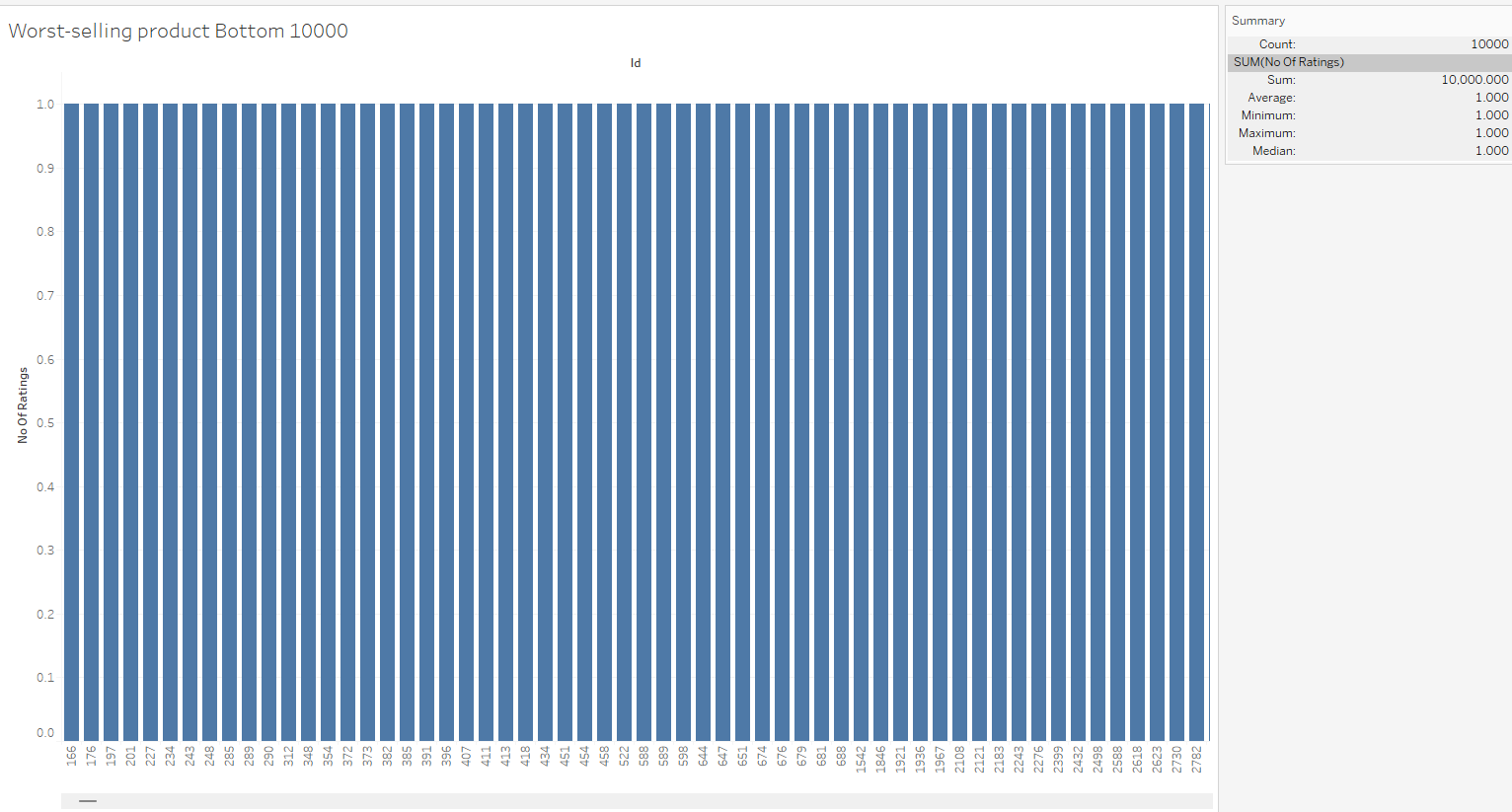
Hình 2.2.1: Hiển thị 10 sản phẩm bán tệ nhất

****

Hình 2.2.2: Hiển thị 100 sản phẩm bán tệ nhất

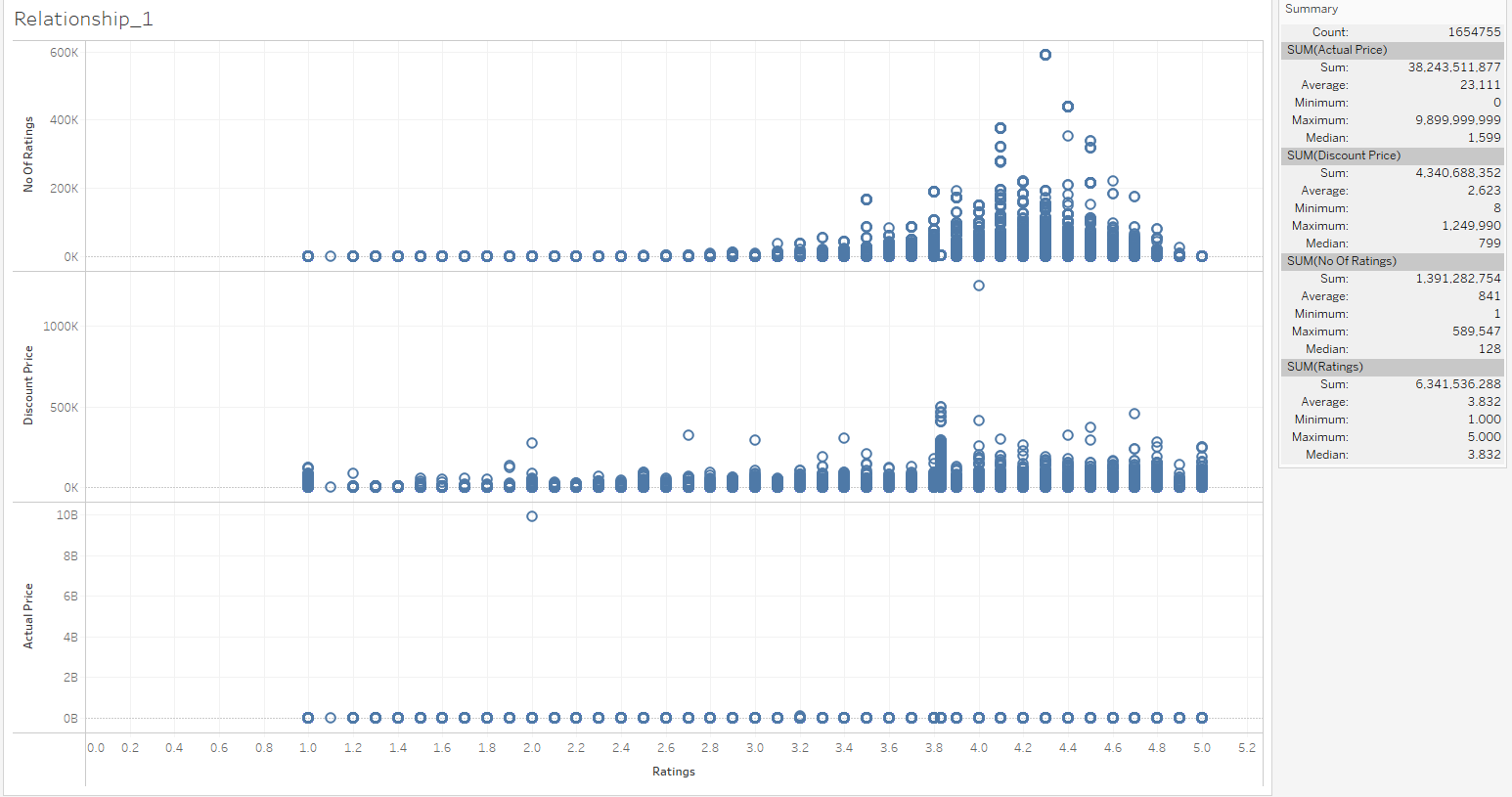


Hình 2.2.3: Hiển thị 1000 sản phẩm tệ nhất



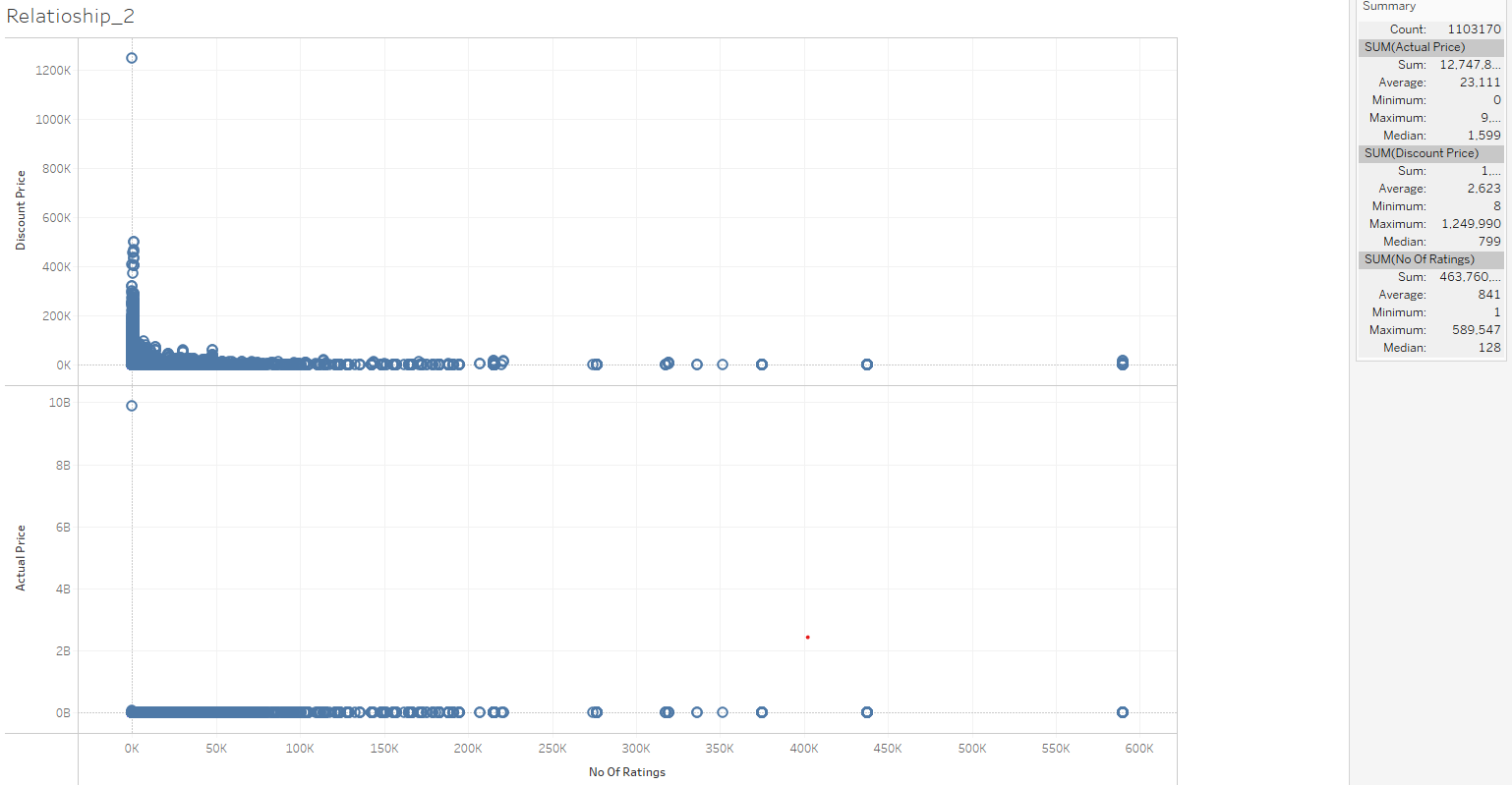
Hình 2.2.4: Hiển thị 10000 sản phẩm tệ nhất

* Nhận xét biểu đồ: Biểu đồ cho thấy sự đồng đều giữa các giá trị của những sản phẩm bán tệ. Người bán cần xem xét và tìm hiểu nguồn gốc vấn đề để có thể cải thiện lợi nhuận
  1. **Tìm và giải thích mối quan hệ các biến của tập dữ liệu**

****

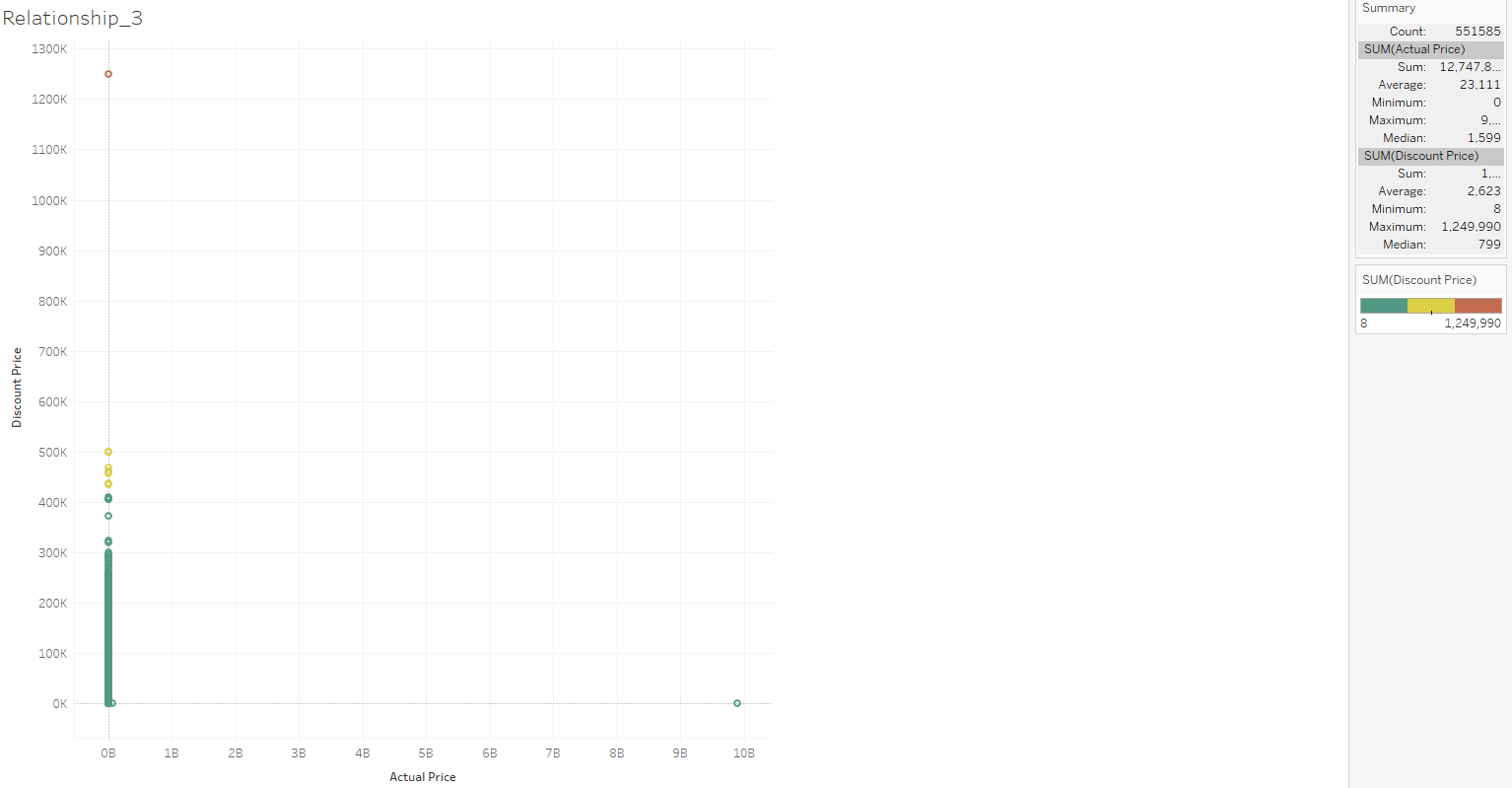
Hình 2.3.1: Biểu đồ tương quan: No of rating & Ratings /Discount price & Ratings/Actual price & Ratings

* Nhận xét No of rating & Ratings: Biểu đồ cho thấy mối quan hệ giữa xếp hạng và No of rating, Một sản phẩm càng có nhiều no of ratings thì xếp hạng càng cao, chứng khi người mua không đánh giá thì sản phẩm đáp ứng được nhu cầu và sự hài lòng của họ.
* Nhận xét Discount price & Ratings: Biểu đồ cho thấy mỗi quan hệ giữa xếp hạng và giảm giá, khi một sản phẩm giảm giá tăng thì xếp hạng cũng tăng. Thể hiện rằng người mua rất yêu thích các sản phẩm giảm giá.
* Nhận xét Actual price & Ratings:Biểu đồ cho thấy mối quan hệ giữa giá thực tế và xếp hạng, có thể thấy giá cả cao hay thấp không ảnh hưởng đến xếp hạng, chúng ta có thể nhìn thấy một sản phẩm có giá thực tế rất cao nhưng vẫn có đánh giá thấp, ngược lại có những sản phẩm giá thấp nhưng xếp hạng vẫn cao. Chứng tỏ xếp hạng cao hay thấp phụ thuộc vào chất lượng sản phẩm, mức độ hài lòng của khách hàng khi nhận sản phẩm



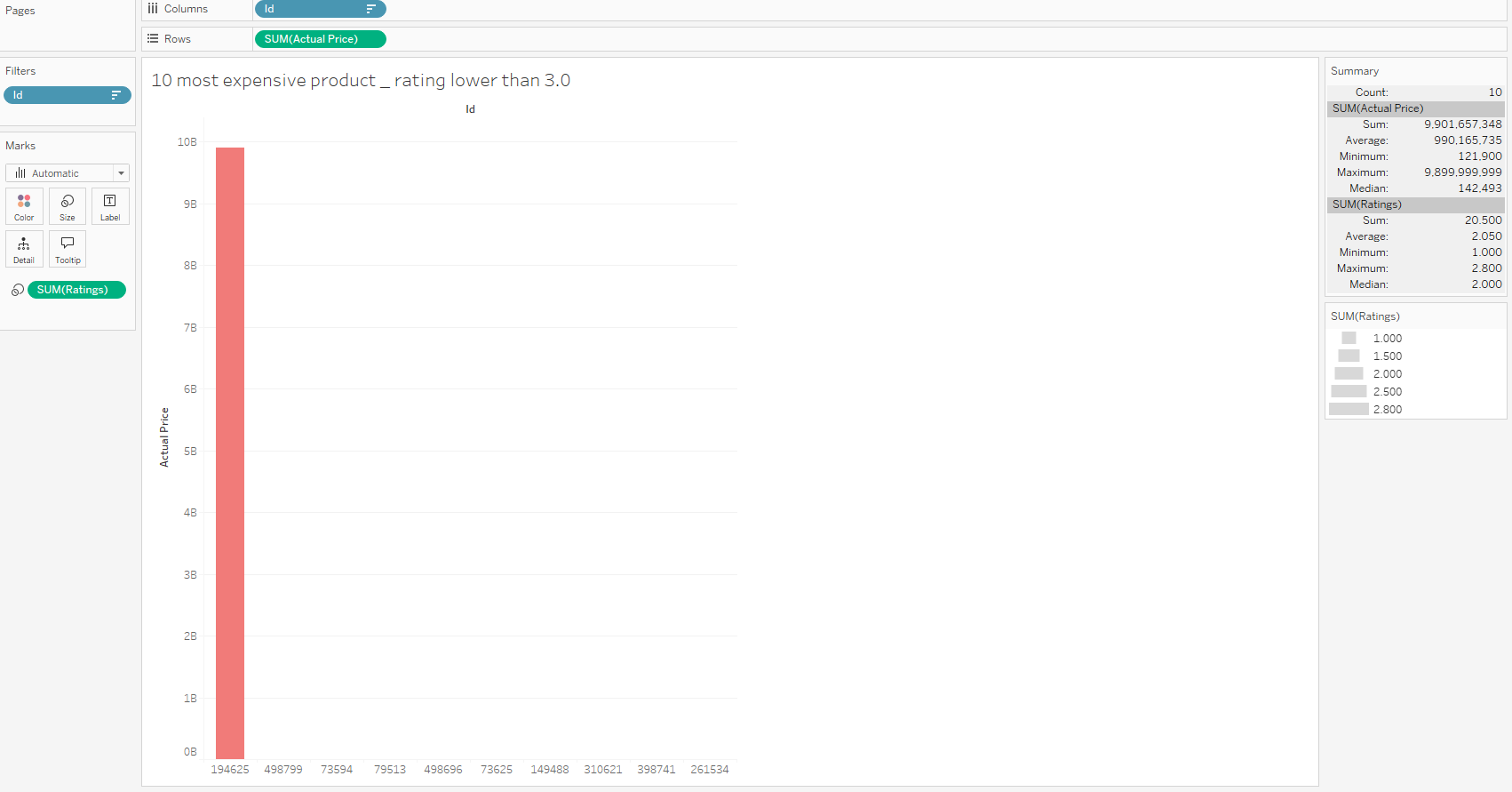
Hình 2.3.2: Biểu đồ tương quan: No of ratings & Discount Price / No of ratings & Actual price

* Nhận xét No of rating & Discount Price: Biểu đồ cho thấy giảm giá (discount price) càng nhiều thì no of ratings càng ít, ngược lại giảm giá càng ít thì no of ratings càng nhiều. Số liệu cho thấy công ty giảm giá nhiều với những sản phẩm có ít no of ratings để nhằm mục đích thu hút người mua nhưng không hiệu quả
* Nhận xét No of rating & Actual price: Biểu đồ cho thấy giá thực tế càng cao thì no of ratings càng ít, ngược lại giá càng thấp no of rating càng nhiều. Số liệu cho thấy những sản phẩm có giá thấp thu hút người nhiều người mua và quan tâm hơn.

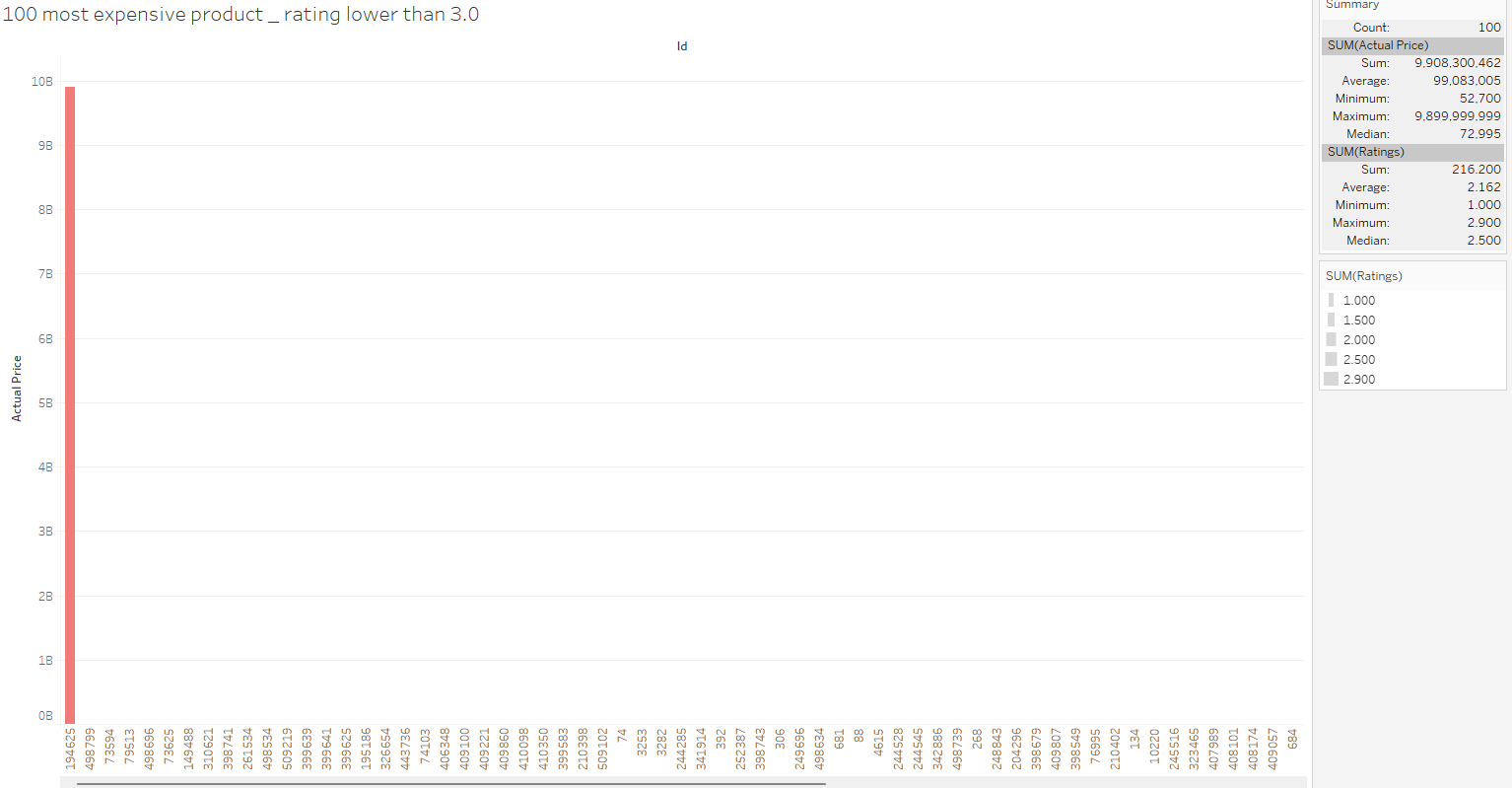
****

Hình 2.3.3: Biểu đồ tương quan: Discount Price & Actual price

* Nhận xét Discount Price & Actual price: Biểu đồ cho thấy mối quan hệ giữa giá thực tế và chiết khấu, những sản phẩm có giá thực tế cao thường sẽ giảm giá ít và ngược lại những sản phẩm có giá thực tế thấp thường sẽ giảm giá nhiều (giá thực tế - chiết khấu). Từ số liệu có thể thấy sự chênh lệch của sản phẩm có giá thực tế cao và giá sau giảm chênh lệch ít hơn so với những sản phẩm có giá thực tế thấp.
* Giải pháp: Từ những biểu đồ trên cho thấy công ty nên chú trọng tới chất lượng sản phẩm và đề xuất ra giá thực tế hợp lý với chất lượng sản phẩm, đồng thời cần có những chính sách ưu đãi giảm giá tri ân khách hàng phù hợp cho từng sản phẩm , đồng thời nên quảng bá sản phẩm rộng rãi để thu hút được những khách hàng tiềm năng nhưng vẫn đáp ứng được nhu cầu và chất lượng sử dụng của sản phẩm.
  1. **Tìm những sản phẩm đắt nhất nhưng có xếp hạng thấp hơn 3.0, ở thang điểm 10, 100, 1K mục**

****

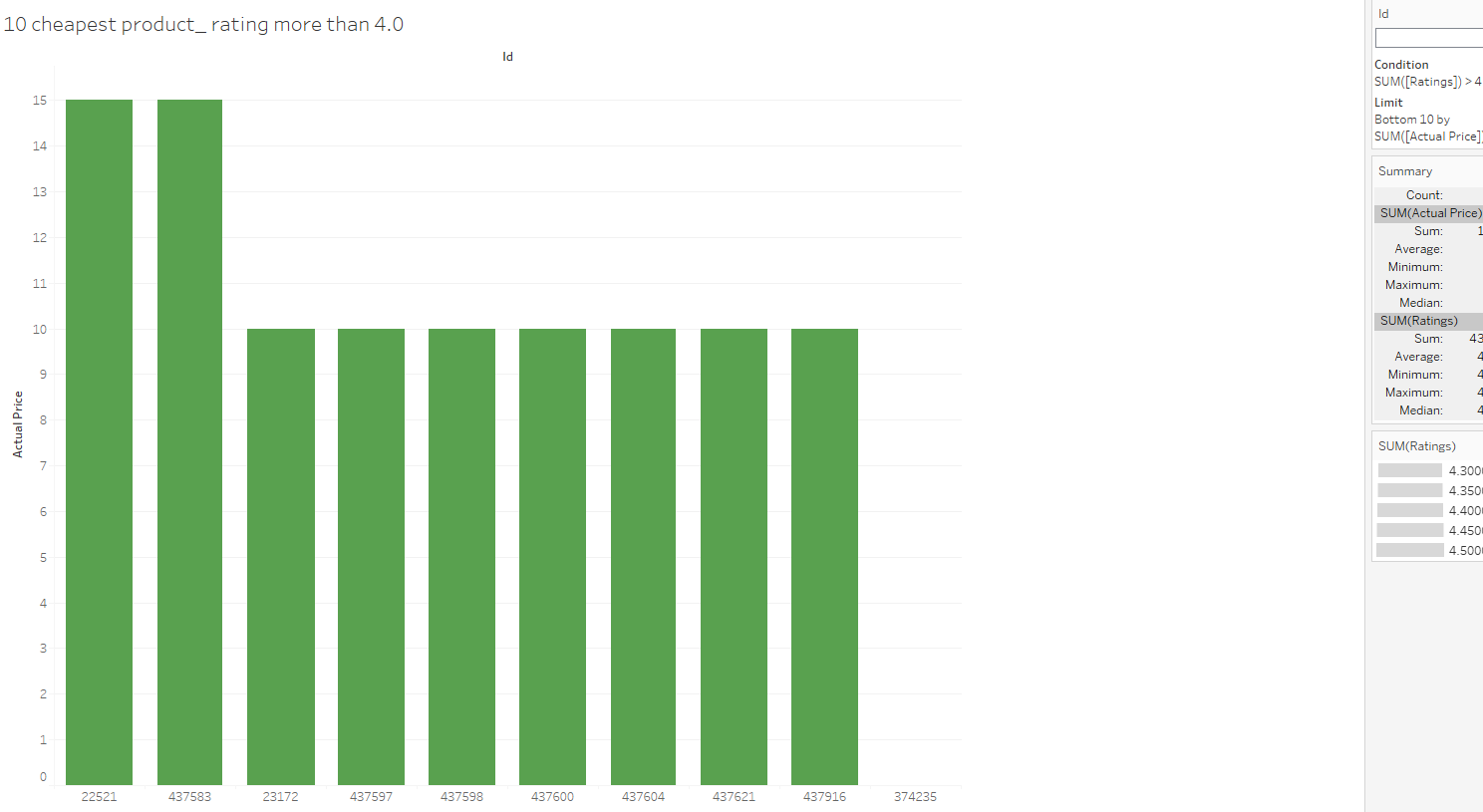
Hình 2.4.1: Hiển thị 10 sản phẩm đắt nhất có xếp hạng nhỏ hơn 3.0



Hình 2.4.2: Hiển thị 100 sản phẩm đắt nhất có xếp hạng nhỏ hơn 3.0

 Hình 2.4.3: Hiển thị 1000 sản phẩm đắt nhất có xếp hạng nhỏ hơn 3.0

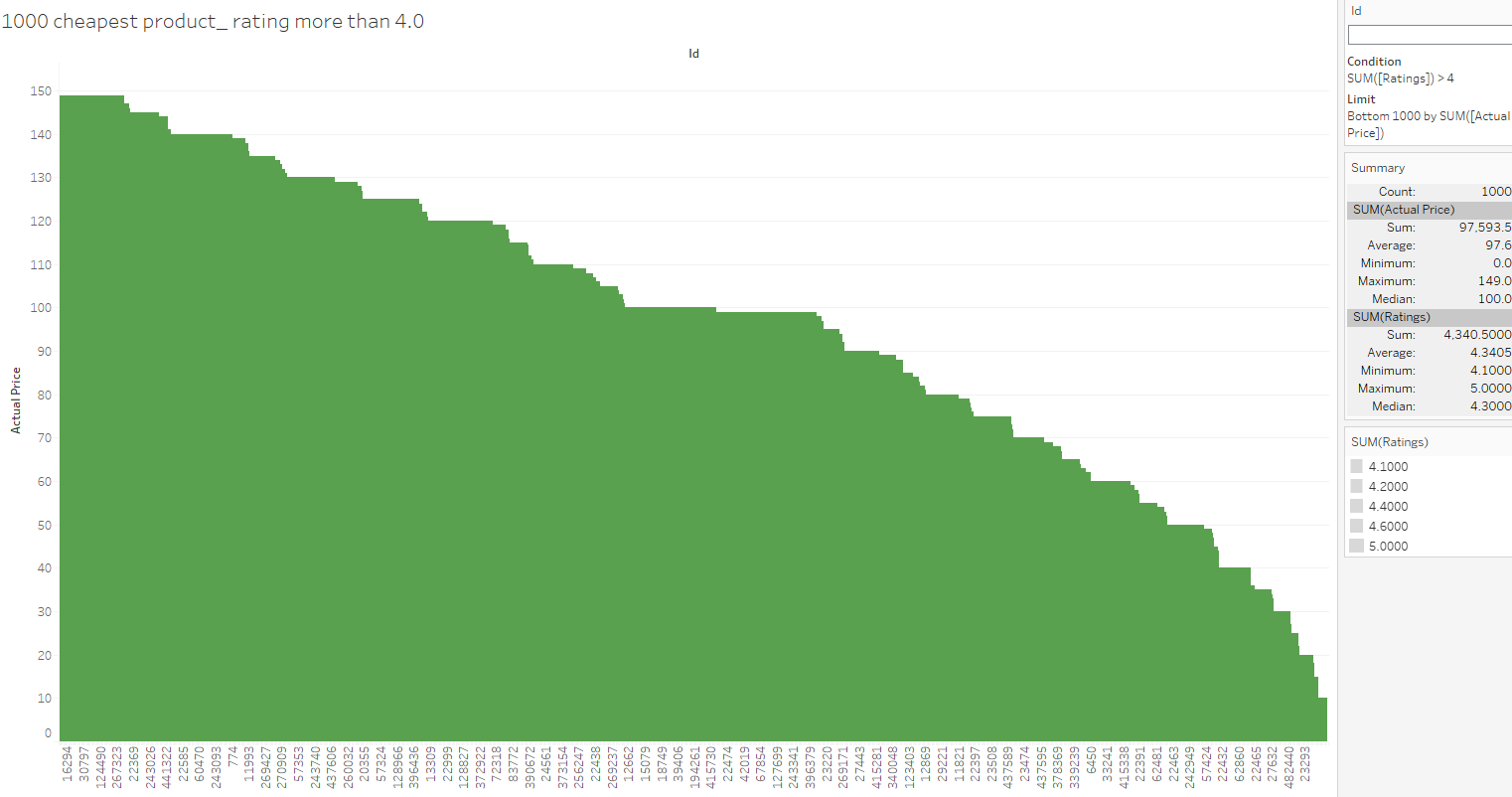
* Nhận xét: Dựa vào số liệu cho thấy những sản phẩm có giá rất cao nhưng lại có xếp hạng rất thấp đều nhỏ hơn 3.0, chỉ ra rằng người dùng không hài lòng với chất lượng sản phẩm so với giá tiền phải bỏ ra, đồng thời ảnh hưởng xấu đến uy tín của sản phẩm, gây thất vọng cho người mua vì không đáng tin cậy.
  1. **Tìm sản phẩm rẻ nhất và có xếp hạng hơn 4.0, trong thang điểm 10, 100, 1K mục**

****

Hình 2.5.1: Hiển thị 10 sản phẩm rẻ nhất có xếp hạng lớn hơn 4.0

****

Hình 2.5.2: Hiển thị 100 sản phẩm rẻ nhất có xếp hạng lớn hơn 4.0

****

Hình 2.5.3: Hiển thị 1000 sản phẩm rẻ nhất có xếp hạng lớn hơn 4.0

* Nhận xét: Biểu đồ cho thấy những sản phẩm có giá rẻ nhất nhưng lại có xếp hạng rất cao đều hơn 4.0, cho sự hài lòng của người dùng khi mua, đồng thời có thể phản ánh được chất lượng sản phẩm rất tốt so với giá. Đồng thời phản ánh mức độ tin cậy của người mua đến những sản phẩm giá rẻ nhưng lại mang lại chất lượng tốt hơn kỳ vọng, cho nên người dùng xếp hạng những sản phẩm này cao và đáng tin cậy
  1. **Tìm sản phẩm có mức giá giảm giá lớn nhất**
* Sử dụng LOD( mức độ chi tiết): {FIXED : MAX ([Discount Price])} để có giá trị tối đa chiết khấu.
* Có thể thấy chiết khấu tối đa của sản phẩm là: 1.249.990



Hình 2.6.1: Hiển thị các sản phẩm có mức giá giảm giá lớn nhất

* Sử dụng phép tính: ([Actual Price]) - ([Discount Price]) để tính giá sản phẩm sau chiết khấu
* Có thể thấy giá sản phẩm cao nhất sau chiết khấu là: 9.899.999.410



Hình 2.6.2: Hiển thị các sản phẩm có mức giá cao nhất sau giảm giá

**CHƯƠNG 3 - TĂNG CƯỜNG DỮ LIỆU**

1. **Tăng cường tập dữ liệu**

* Ngôn ngữ lập trình : Python
* Thư viện: csv, random, pandas

import csv

import random

import pandas as pd

# Lấy danh sách các thành phố

with open('us\_cities.txt', 'r') as city\_file:

    cities = [line.strip() for line in city\_file]

# Tạo 100 lần ra 100 file khác nhau

files = []

for i in range(100):

    # Mỗi lần gen ra file có định dạng:

    new\_csv\_filename = f'new\_data\_{i+1}.csv'

    # Mở file csv đã cleaned để xử lý

    with open('dataset\_cleaned.csv', 'r', encoding='utf-8') as original\_csv, open(new\_csv\_filename, 'w', newline='', encoding='utf-8') as new\_csv:

        reader = csv.reader(original\_csv)

        writer = csv.writer(new\_csv)

        # Add thêm profit và city ở đầu

        header = next(reader)

        header.extend(["profit", "city"])

        writer.writerow(header)

        # Đọc từ dòng rồi in ra profit random từ 1->5000 và city chọn ngẫu nhiên từ danh sách

        for row in reader:

            profit = random.randint(1, 5000)

            city = random.choice(cities)

            row = [cell.replace('\ufffd', '') for cell in row]

            row.extend([profit, city])

            writer.writerow(row)

    files.append(new\_csv\_filename)

    print(f"Created a new file '{new\_csv\_filename}' with 2 columns: profit and city.")

final\_df = pd.read\_csv(files[0])

# Đọc các file từ 1 -> 100 rồi hợp nhất lại

for file in files[1:]:

    df = pd.read\_csv(file)

    final\_df = pd.concat([final\_df, df], ignore\_index=True)

# Tạo thành file final\_merged\_file

merged\_file = 'final\_merged\_file.csv'

final\_df.to\_csv(merged\_file, index=False)

print(f"All files merged into {merged\_file}")

df = pd.read\_csv(merged\_file)

# Order theo id giảm dần

sorted\_df = df.sort\_values(by="id", ascending=True)

# Thay thế id từ 0 đến tăng dần

sorted\_df['id'] = range(len(sorted\_df))

# Lưu lại thành file final\_merged\_file\_sort.

output\_file = "final\_merged\_file\_sort.csv"

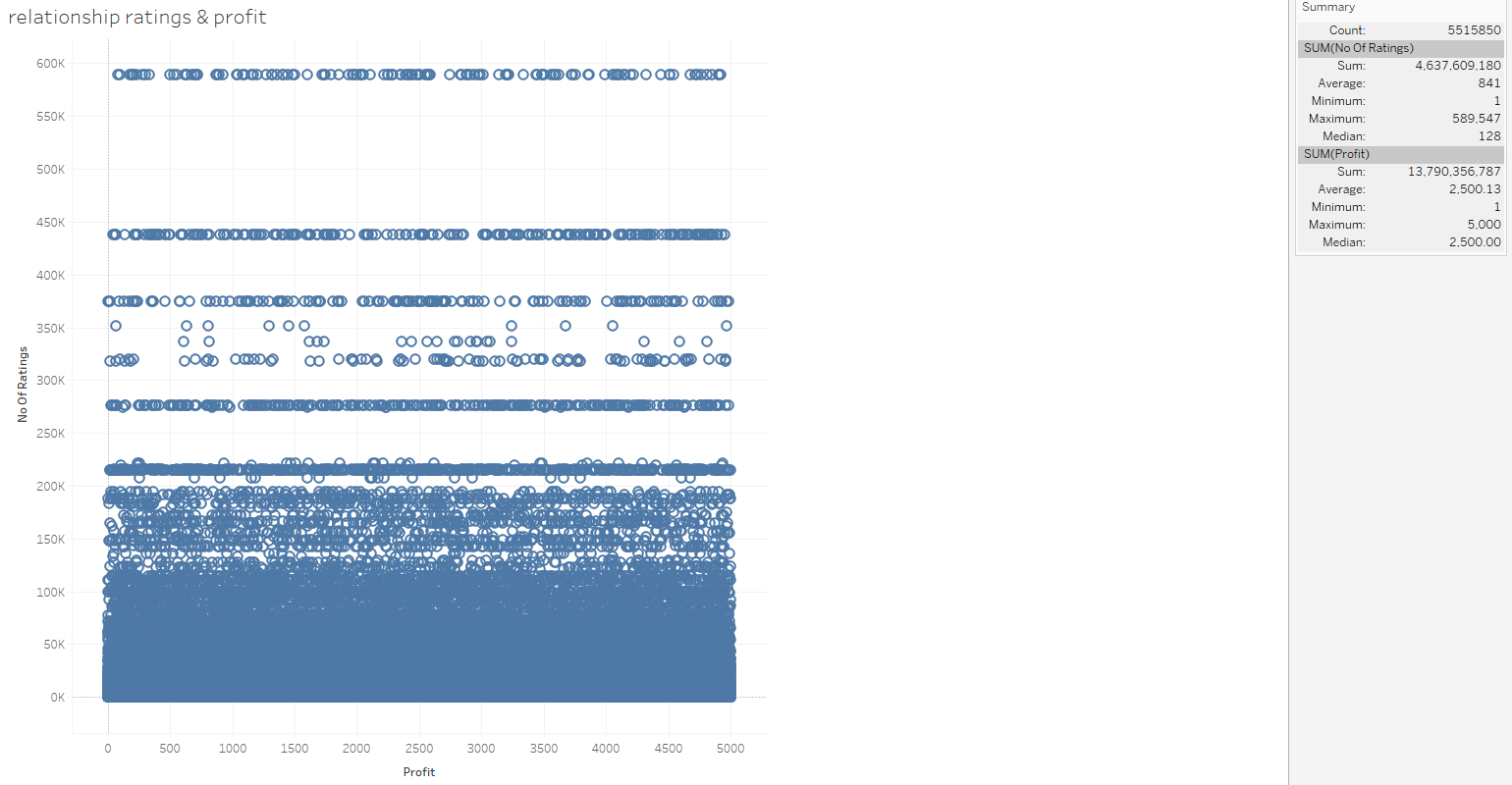
sorted\_df.to\_csv(output\_file, index=False)

print(f"File sorted and 'id' column updated. Saved as {output\_file}")

* Mô tả: Thêm thư viện csv để đọc file csv, thêm thư viện random tạo giá trị ngẫu nhiên, thư viện pandas để xử lý dữ liệu trong python.
* Đầu tiên mở file ở chế độ đọc và loại bỏ khoảng trống và dấu xuống dòng ở mỗi dòng(line.strip()) duyệt qua từng dòng trong file lấy danh sách các thành phố
* Tạo 100 với định dạng “new\_data\_{i+1}.csv”
* Xử lý file gốc và tạo file mới thêm 2 cột mới ”profit” và “city”
* Thêm dữ liệu ngẫu nhiên vào từng dòng trong khoảng từ 1 đến 5000 và chọn ngẫu nhiên một thành phố từ danh sách cities và loại bỏ ký tự không hợp lệ ra khỏi dữ liệu và thêm profit & city vào dòng hiện tại sau đó ghi dòng đã chỉnh vào file mới
* Lưu tên file đã tạo và hiển thị thông báo
* Đọc các file từ 1 đến 100 và hợp nhất tất cả, đọc file đầu tiên vào DataFrame sau đó đọc từng file csv khác và kết hợp các DataFrame với nhau làm mới mục index
* Sau đó lưu file csv hợp nhất vào 'final\_merged\_file.csv'
* Tiếp theo cập nhập lại id tăng dần từ 0 và lưu lại thành file "final\_merged\_file\_sort.csv"

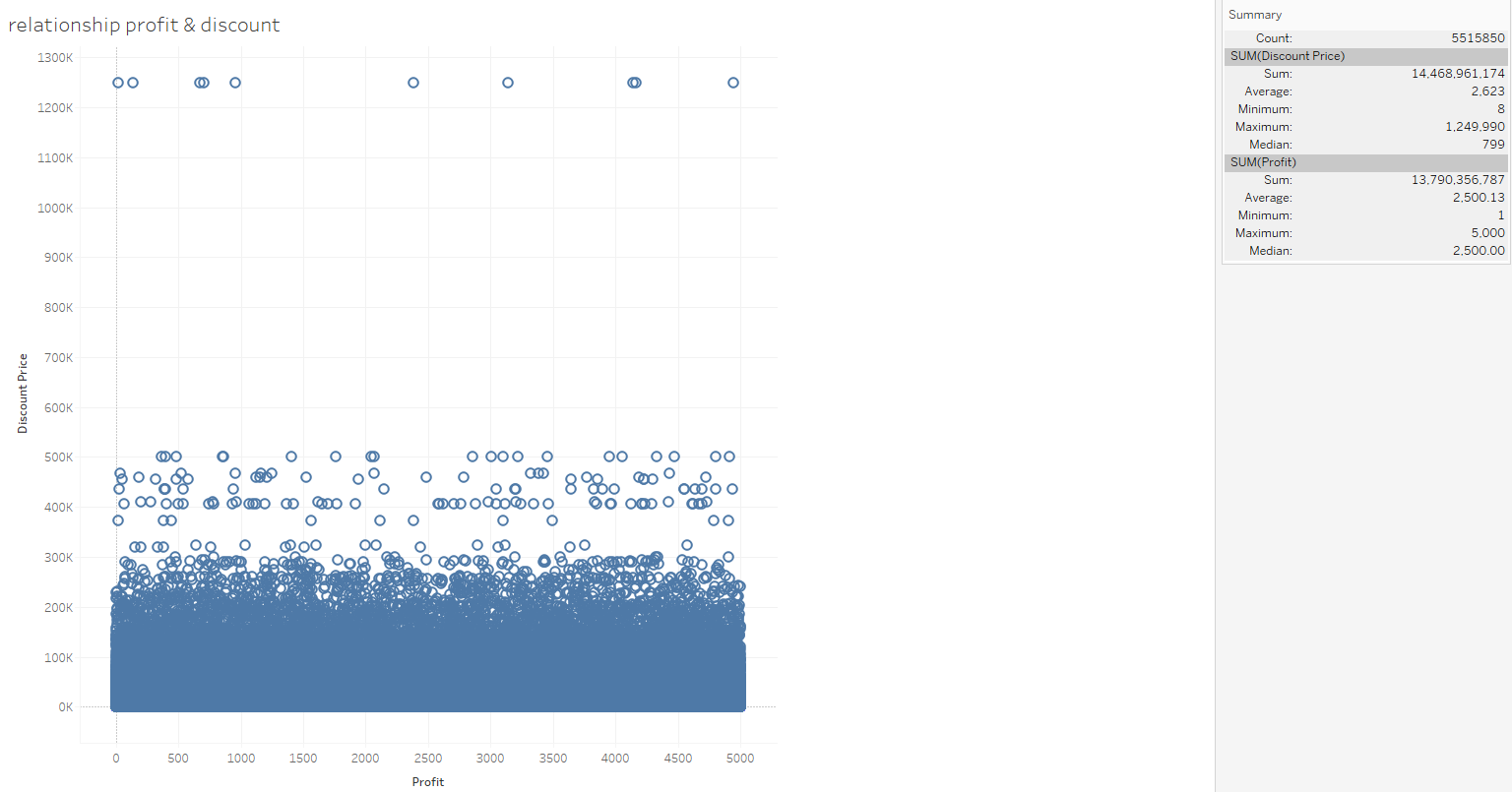
**CHƯƠNG 4 – HOẠT ĐỘNG RA QUYẾT ĐỊNH**

* 1. **Tìm và giải thích mối quan hệ giữa các biến trong tập dữ liệu**

****

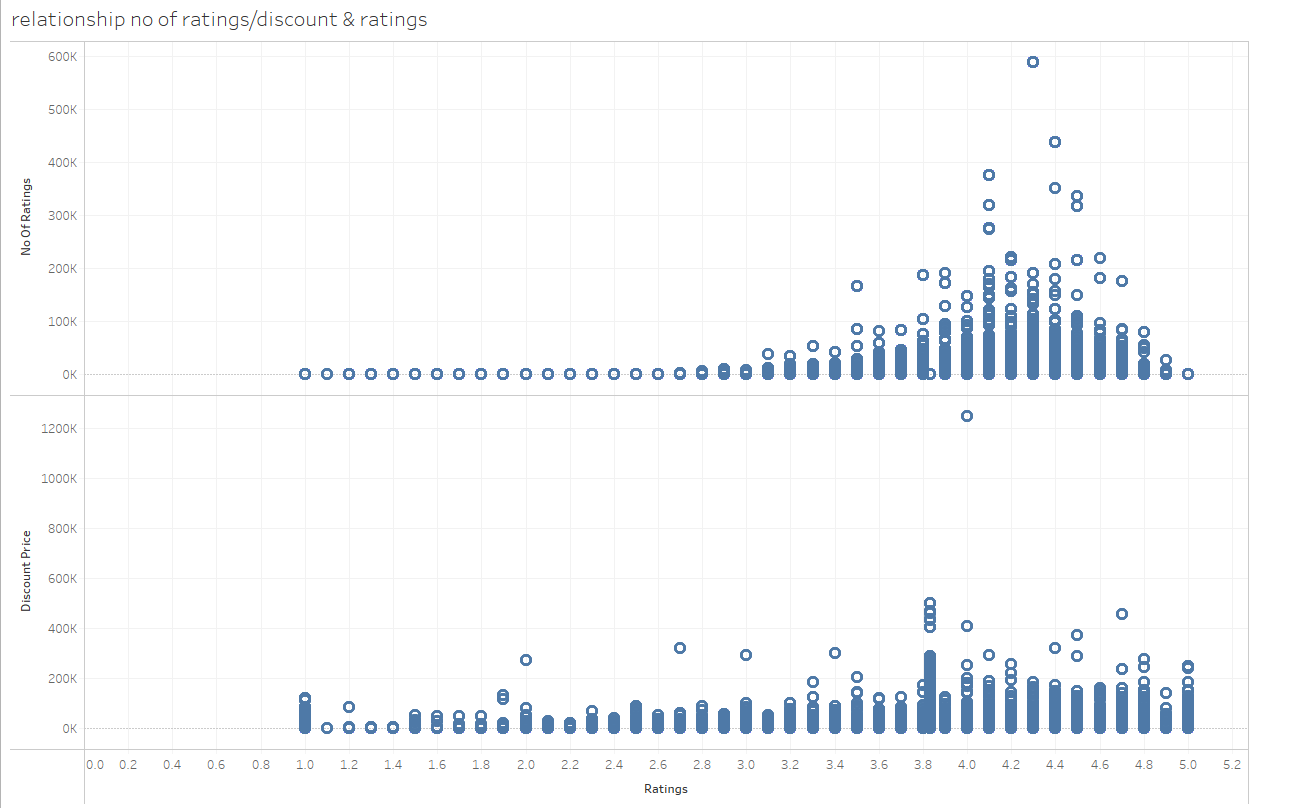
Hình 4.1.1: Mối quan hệ profit & no of ratings

* Nhận xét profit & no of ratings: Từ biểu đồ cho thấy tất cả lợi nhuận cao/thấp đều có đa số lượng no of ratings trong khoảng từ 0 -> 200k. Chỉ một số ít sản phẩm lớn hơn 200k no of ratings



Hình 4.1.2: Mối quan hệ profit & discount

* Nhận xét: Qua biểu đồ cho thấy sự đồng đều của các sản phẩm, không phân biệt những sản phẩm có lợi nhuận cao hay thấp công ty đều đưa ra những giảm giá đều nhau. Từ đó dẫn đến việc những sản phẩm có lợi nhuận thấp nhưng lại có chiết khấu cao có thể dẫn đến việc công ty bị lỗ
* Giải pháp: Công ty nên tiết chế giảm chiết khấu của những sản phẩm có lợi nhuận thấp ở mức ngang bằng nhằm giúp cân bằng lợi nhuận và giúp những sản phẩm có lợi nhuận cao không phải gánh những sản phẩm có lợi nhuận thấp

****

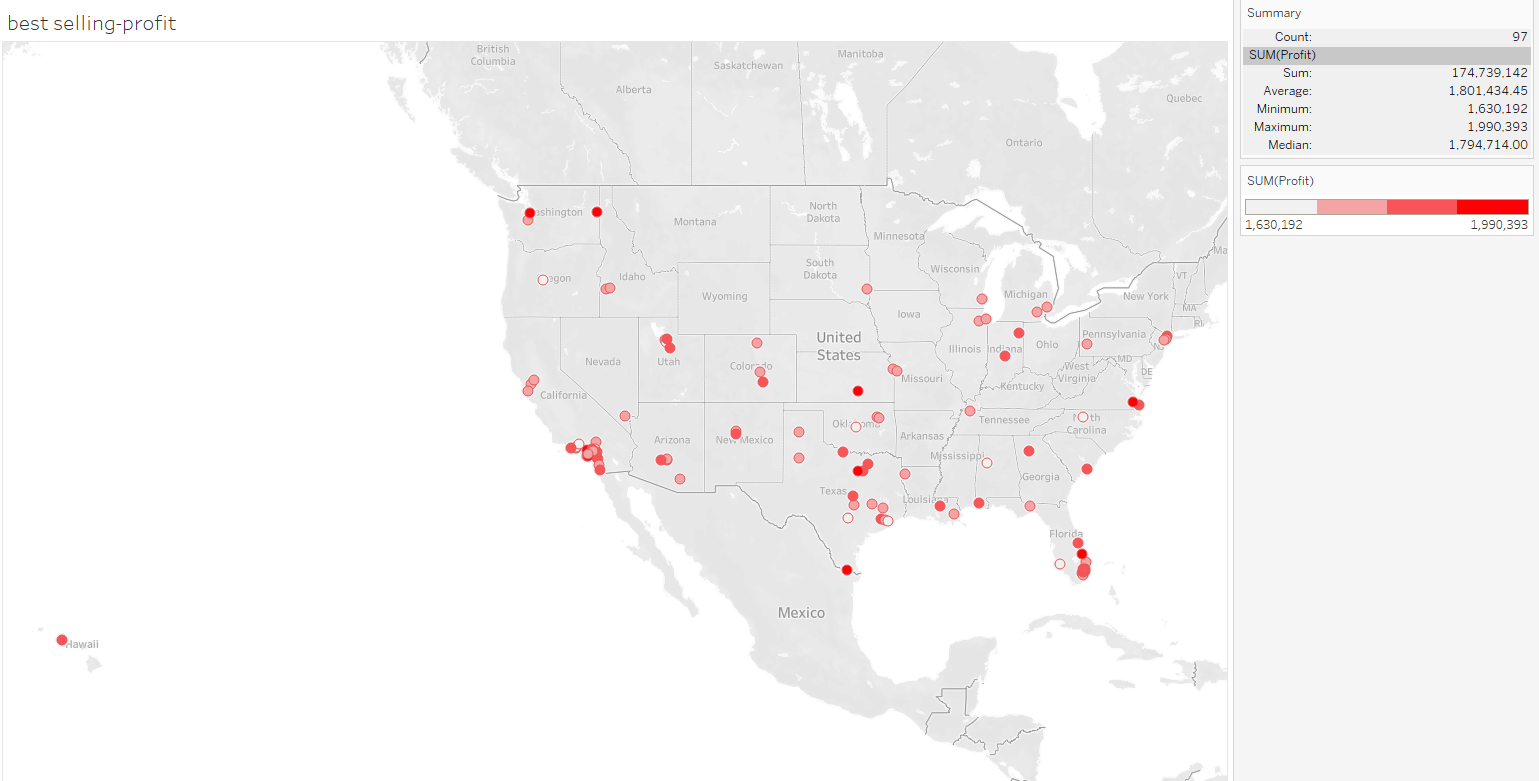
Hình 4.1.3: No of rating & Ratings /Discount price & Ratings

* Nhận xét No of rating & Ratings: Biểu đồ cho thấy mối quan hệ giữa xếp hạng và No of rating, Một sản phẩm càng có nhiều no of ratings thì xếp hạng càng cao đồng thời sản phẩm đáp ứng được nhu cầu và sự hài lòng của họ. Ngược lại sản phẩm có ít no of ratings thì có xếp hạng thấp, chứng tỏ sản phẩm không đạt chất lượng, khiến người dùng thất vọng
* Nhận xét Discount price & Ratings: Biểu đồ cho thấy mỗi quan hệ giữa xếp hạng và chiết khấu, khi một sản phẩm có chiết khấu tăng thì xếp hạng cũng tăng. Thể hiện rằng người mua rất yêu thích các sản phẩm chiết khấu cao.
  1. **Xác định mức lợi nhuận**

****

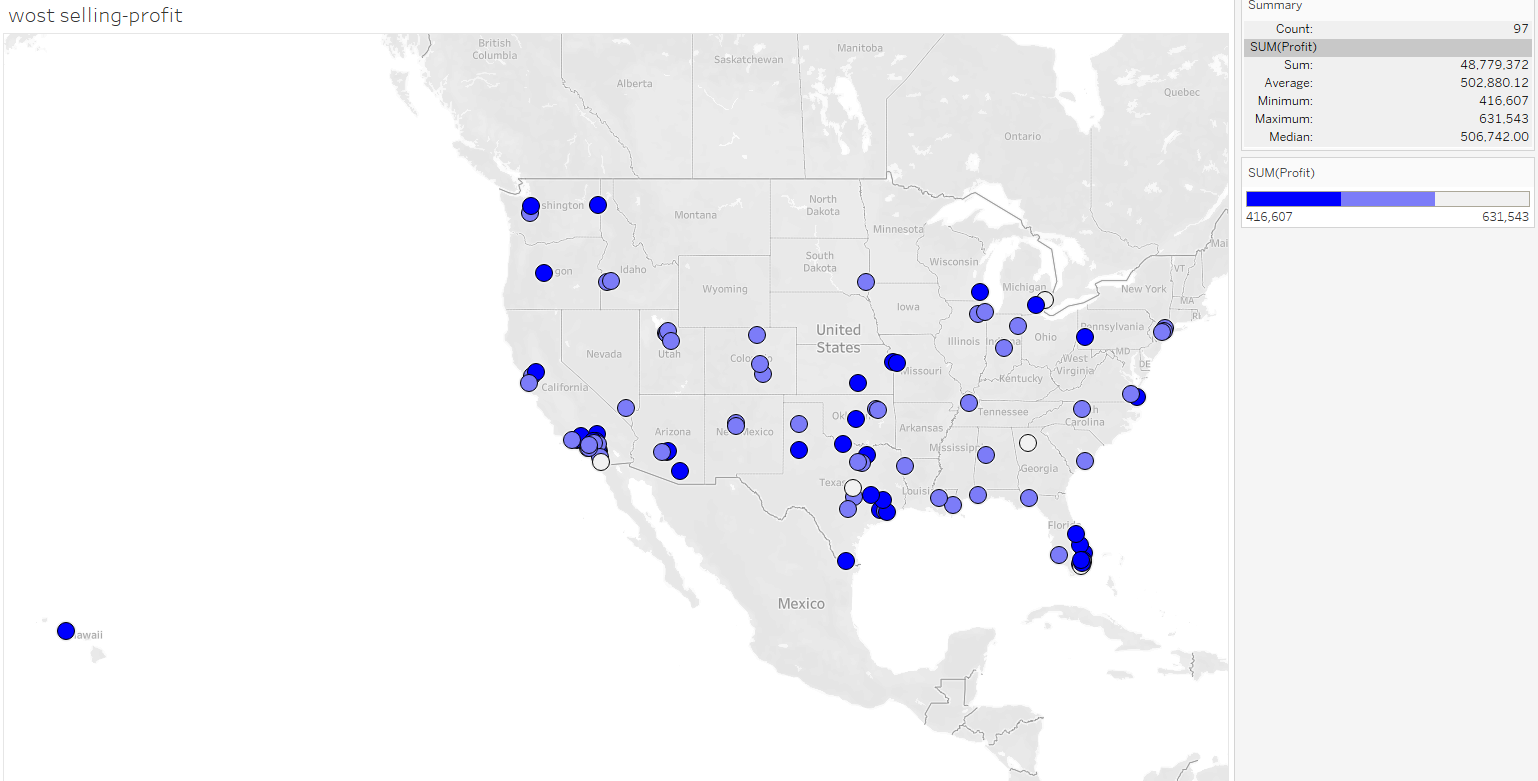
Hình 4.2.1: Lợi nhuận sản phẩm

* Điều kiện: [Profit]>=4000 THEN "good Produc" ELSE "bad Produc"
* Điều kiện này đánh giá sản phẩm có proifit phải từ 4000 trở lên mới có thể trở thành sản phẩm có lợi nhuận tốt.
* Những sản phẩm có nhiều lợi nhuận cao(màu đỏ) có ít hơn những sản phẩm có lợi nhuận thấp(màu xanh)
* Dữ liệu cho thấy các sản phẩm có lợi nhuận cao dù số lượng ít hơn nhưng đóng góp nhiều hơn vào lợi nhuận tổng thể.Các sản phẩm có lợi nhuận thấp có hiệu suất không đồng đều, với nhiều sản phẩm có lợi nhuận thấp hoặc gần bằng 0.
* Nhà đầu tư nên mở rộng sản phẩm có lợi nhuận tốt và nên có kế hoạch cải thiện những sản phẩm có lợi nhuận thấp.
  1. **Hình dung lợi nhuận bán chạy nhất và tệ nhất trên bản đồ thế giới**



Hình 4.3.1: Lợi nhuận bán chạy nhất

* Lợi nhuận trên biểu đồ chạy từ 1.630.192$ đến 1.990.393$
* Lợi nhuận cao của sản phẩm trên biểu đồ được thể hiện bằng các chấm đỏ
* Lợi nhuận cao được phân bố đồng đều trên nước Mỹ
* Nhận xét :Sự phân bố khá đều trên khắp nước Mỹ cho ta thấy rằng các sản phẩm này được người tiêu dùng đón nhận dẫn đến việc có lợi nhuận cao và đồng đều trên khắp khu vực nước Mỹ. Đồng thời thể hiện sự đa dạng, phổ biến và có sức hút đối với khách hàng ở nhiều khu vực, địa hình của sản phẩm có lợi nhuận cao trên toàn quốc

 Hình 4.3.2: Lợi nhuận bán tệ nhất

* Phạm vi lợi nhuận chạy từ 416.607$ đến 631.543$
* Lợi nhuận bán tệ của sản phẩm trên bản đồ thể hiện bằng các chấm xanh
* Lợi nhuận bán tệ phân bố khá đồng đều trên khắp nước Mỹ.
* Nhận xét: Sự phân bố khá đồng đều của các sản phẩm có lợi nhuận bán tệ trên mọi khu vực, địa hình của nước Mỹ cho chúng ta thấy những sản phẩm đó cần phải được cải thiện về chất lượng, công dụng, giá cả, quảng cáo,... để phù hợp với nhu cầu sử dụng của người dân nơi đây hơn và có thể tiếp cận với người dân nhiều hơn nếu muốn cải thiện lợi nhuận bán hàng.