This report will present a comprehensive analysis of an ecommerce product dataset, with particular focus on multiple hypotheses. Hypothesis 1: Higher-rated products have a higher number of reviews. Hypothesis 2: The price of the product is correlated with its rating. Hypothesis 3: Certain categories of products have higher average ratings than others. Hypothesis 4: There is a seasonal trend in the number of reviews; specifically there are more reviews during holiday seasons. I also used the Pearson Correlation Coefficient to understand any linear relationship between two variables.

The dataset was cleaned and analysed to provide insights into various aspects such as product ratings, prices, number of reviews, and their distributions. The report includes statistical summaries and visualizations generated using R and Python. The initial dataset was loaded, and a summary was created to understand its structure. The data was then cleaned by removing any missing values.

The dataset summary provides several insights into various attributes of the products in the dataset. There were key points observed within each attribute. ProductID: ranges from 1 to 1000, indicating the dataset contains 1000 unique products. ProductName: is a character data type with 1000 unique entries, indicating each product belongs to a specific category. Category: was also character data type containing 1000 unique entries, indicating each product belongs to a specific category. Price: The minimum price is $10.11, and the maximum price is $499.74. The mean price is $253.78, with a median of $251.31. Prices are evenly distributed with the 1st quartile at $133.09 and the 3rd quartile at $375.83. Rating: ratings range from 1.0 to 5.0. The mean rating is 3.026, with a median of 3.1. Ratings are distributed with the 1st quartile at 2.1 and the 3rd quartile at 4.0. NumReviews: number of reviews ranges from 3 to 4994. The mean number of reviews is 2499, with a median of 2476. Reviews are distributed with the 1st quartile at 1202 and the 3rd quartile at 3798. StockQuantity: stock quantity ranges from 0 to 993 units. The mean stock quantity is 494 with a median of 505. Stock levels are distributed with the 1st quartile at 242 and the 3rd quartile at 744. Discount: discount rates range from 0% to 50%. The mean discount rate is 25.16%, with a median of 25%. Discounts are distributed with the 1st quartile at 13% and the 3rd quartile at 38%. Sales: sales range from 0 to 1997 units. The mean number of sales is 1011, with a median of 998. Sales are distributed with the 1st quartile at 502 and the 3rd quartile at 1540. DateAdded: dates range from June 14, 2023, to June 11, 2024. The mean date added is December 10, 2023, with a median of December 5, 2023. Dates are distributed with the 1st quartile at September 9, 2023 and the 3rd quartile at March 9, 2024. City: is a character data type with 1000 unique entries, indicating each product is associated with a specific city.

Overall, this summary provides a snapshot of the product dataset, giving insights into the distribution and range of various attributes such as price, rating, reviews, stock, discounts, and sales. It also provides information about the dates products were added and their associated cities.

ProductID ProductName Category Price

Min. : 1.0 Length:1000 Length:1000 Min. : 10.11

1st Qu. : 250.8 Class :character Class :character 1st Qu.:133.09

Median : 500.5 Mode :character Mode :character Median :251.31

Mean : 500.5 Mean :253.78

3rd Qu. : 750.2 3rd Qu.:375.83

Max. :1000.0 Max. :499.74

Rating NumReviews StockQuantity Discount Sales

Min. :1.000 Min. : 3 Min. : 0.0 Min. :0.0000 Min. : 0

1st Qu.:2.100 1st Qu.:1202 1st Qu.:241.8 1st Qu.:0.1300 1st Qu.: 502

Median :3.100 Median :2476 Median :505.0 Median :0.2500 Median : 998

Mean :3.026 Mean :2499 Mean :495.4 Mean :0.2516 Mean :1011

3rd Qu.:4.000 3rd Qu.:3798 3rd Qu.:743.5 3rd Qu.:0.3800 3rd Qu.:1540

Max. :5.000 Max. :4994 Max. :993.0 Max. :0.5000 Max. :1997

DateAdded City

Min. :2023-06-14 Length:1000

1st Qu.:2023-09-09 Class :character

Median :2023-12-05 Mode :character

Mean :2023-12-10

3rd Qu.:2024-03-09

Max. :2024-06-11

I created two histograms, one of product ratings and the other of product prices. This will help in understanding how ratings and prices are spread across the dataset. The “Distribution of Ratings” histogram shows that most products have a rating between 3 and 4 stars, indicating a generally positive reception. The “Distribution of Prices” histogram can reveal whether the products are mostly low-priced, mid-priced, or high-priced. This can be used in pricing strategies.

A graph of a distribution of ratings

Description automatically generated

A graph of a distribution of prices

Description automatically generated

Boxplots and violin plots were created to analyse the number of reviews across different product ratings. These visualizations will help identify if higher-rated products have more reviews. This suggests that satisfied customers are more likely to leave reviews. Conversely, a lack of reviews for high-rated products indicate an opportunity to encourage more customer feedback.

A graph of a number of reviews

Description automatically generated

A graph showing a number of reviews

Description automatically generated

Boxplots were used to visualize the distribution of ratings and the number of reviews across different product categories. These visualizations are used to help understand if certain categories receive higher ratings and more reviews. Categories with consistently high ratings and a high number of reviews represent popular or high-quality products. Identifying these categories can help in focusing marketing and inventory efforts.

A graph with purple bars

Description automatically generated

A bar graph with red and black bars

Description automatically generated

The dataset was then analysed to observe trends over time. Line plots were used to show the number of reviews and average prices over time. These visualizations will identify any temporal trends, such as seasonal variations. An increase in the number of reviews during certain periods (e.g., holiday seasons) can indicate seasonal trends. Understanding these trends helps in planning marketing campaigns and stock management. Average price trends over time can indicate changes in pricing strategy or market conditions.

A graph with a line going up

Description automatically generated

A graph showing the average price over time

Description automatically generated

I chose to use bar plots to identify seasonal trends in the number of reviews and seasonal trends in average prices. These visualizations will highlight periods with more activity, which can be useful for marketing and inventory planning. Spikes in the number of reviews during holiday seasons suggest increased customer engagement during these times. Seasonal price variations can indicate periods of discounts or promotions. These insights should help in planning promotional strategies.

A graph of different colored bars

Description automatically generated

A graph of different colored bars

Description automatically generated

A correlation matrix was computed to understand the relationship between numerical variables and displays the Pearson correlation coefficients. Pearson correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It can greatly aid in understanding if an increase in one variable tends to correspond with an increase or decrease in another variable. The value of the Pearson correlation coefficient ranges from -1 to 1. One represents a perfect positive linear relationship (as one variable increases, the other variable increases proportionally). Negative one represents a perfect negative linear relationship (as one variable increases, the other variable decreases proportionally). Zero represents no linear relationship (the variables show no linear pattern). For example, a positive correlation between sales and the number of reviews would suggest that more reviews can drive higher sales. The correlation matrix helps in identify key factors that influence product performance.

ProductID Price Rating NumReviews StockQuantity

ProductID 1.0000000000 -0.040180334 -0.0002887686 0.043875378 -0.015074422

Price -0.0401803341 1.000000000 -0.0064621904 0.012781943 -0.006409298

Rating -0.0002887686 -0.006462190 1.0000000000 -0.009832483 0.000575592

NumReviews 0.0438753784 0.012781943 -0.0098324833 1.000000000 -0.020957835

StockQuantity -0.0150744218 -0.006409298 0.0005755920 -0.020957835 1.000000000

Discount -0.0167460051 -0.045133213 0.0244951626 -0.015579497 -0.007659338

Sales 0.0020034377 0.029076056 0.0084753829 0.056586375 -0.001457783

Discount Sales

ProductID -0.016746005 0.002003438

Price -0.045133213 0.029076056

Rating 0.024495163 0.008475383

NumReviews -0.015579497 0.056586375

StockQuantity -0.007659338 -0.001457783

Discount 1.000000000 0.027413470

Sales 0.027413470 1.000000000

A graph of blue squares

Description automatically generated

Additional, correlation tests were performed for pairs of numeric variables to validate the significance of their correlations. The p value included below is used to assess statistical significance. Low or non-significant correlations suggest that these variables do not have a strong linear relationship. Significant correlations can help in the understanding how changes in one variable might affect another. This is useful for predictive modelling and strategic decisions.

variable1 variable2 correlation p\_value

cor ProductID Price -0.0401803341 0.20424952

cor1 ProductID Rating -0.0002887686 0.99272320

cor2 ProductID NumReviews 0.0438753784 0.16562675

cor3 ProductID StockQuantity -0.0150744218 0.63398468

cor4 ProductID Discount -0.0167460051 0.59685405

cor5 ProductID Sales 0.0020034377 0.94954745

cor6 Price Rating -0.0064621904 0.83827618

cor7 Price NumReviews 0.0127819434 0.68642489

cor8 Price StockQuantity -0.0064092975 0.83958185

cor9 Price Discount -0.0451332125 0.15381662

cor10 Price Sales 0.0290760562 0.35835201

cor11 Rating NumReviews -0.0098324833 0.75614347

cor12 Rating StockQuantity 0.0005755920 0.98549601

cor13 Rating Discount 0.0244951626 0.43907726

cor14 Rating Sales 0.0084753829 0.78894200

cor15 NumReviews StockQuantity -0.0209578347 0.50797845

cor16 NumReviews Discount -0.0155794968 0.62266235

cor17 NumReviews Sales 0.0565863749 0.07367708

cor18 StockQuantity Discount -0.0076593378 0.80884979

cor19 StockQuantity Sales -0.0014577834 0.96327715

cor20 Discount Sales 0.0274134699 0.38650741

A python visualization was created to show the correlation between ratings and the number of reviews by category. Categories with a strong positive correlation between ratings and reviews indicate that customer satisfaction strongly drives review behavior in these categories Identifying these categories can help in prioritizing efforts to improve product quality and customer satisfaction.

A graph of a bar graph

Description automatically generated with medium confidence

This analysis provides valuable insights into the distribution and relationships between product attributes in the e-commerce dataset. The visualizations highlight key trends and correlations that can be leveraged for strategic decisions in e-commerce platforms. For instance, understanding the weak correlation between price and rating can help in pricing strategies, and recognizing seasonal trends in reviews can inform marketing campaigns.