Assignemnt 1 - Amazon Sales Analysis

Mohammad Hossein Basoulii

June 6, 2025

Abstract

*In this study we focus on a sales dataset gathered from **Amazon**. By this, we are looking towards finding patterns and relationships between different factors in our dataset and their contribution to our sales. We have used Exploratory Data Analysis (EDA) to get initial insights and form hypotheses about the data. And then, we used statistical testing as a tool for testing our hypotheses. Some of the key findings of this study showcase a independence of **discount price** and **ratings**, dependence of **product category** and **ratings** and non-normality of **rating counts** for different products.*

Introduction

Background: Analysis of sales' data of e-commerce businesses has been a matter of great interest, since the very beginning of this era, for all of the companies which care about growing their benefit from the continuous change of demands that cutomers have over time. So the importance of this study is obvious.

Objective: We first, start off by forming some hypotheses about our data, e.g., does **discount price** affect **rating**?. Then we try to find the correct way to test our hypotheses through Exploratory Data Analysis (EDA) section. And at the end, we actually go for Hypothesis Testing section, which would allow us to formally justify our hypotheses or reject them.

Initial hypotheses:

- Does the discounted price significantly impact product ratings?
- Are product categories and high/low ratings independent?

- Is there a significant difference in ratings between high-discount and low-discount products?
- Do more expensive products receive higher ratings on average?
- Does the distribution of rating counts follow a normal distribution?

Data

Data Source: The dataset comes from a sales' product record from **Amazon**. It has 1465 rows(sale records) and 16 different columns(features).

Features: numerical features include **discounted_price**, **actual_price**, **discount_percentage**, **rating** and **rating_count**. The only important categorical feature is **category**.

Data Preprocessing: We find out that there is only two rows that have missing values. More specifically, these rows are missing values only on their **rating_count** column. We decide to just drop these rows, simply because missing them is insignificant, compared to the magnitude of the entire dataset. We also need to apply a simple transformation to clean the data in columns **discounted**, **actual_price**, **discount_percentage**, **rating** and **rating_count** and convert their initial data type to numeric to be able to work with them.

Exploratory Data Analysis(EDA):

Univeriate Analysis:

Distributions of discounted_price and rating:

Analysis: By looking at the Figure 1, we can see that the distribution of both variables is skewed. Thus performing ***Pearson Correlation Test*** wouldn't be appropriate for examining the relationship between these two variables, since assumption of normality of the variables is violated.

Relevance: This analysis would be importatnt when we want to decide whether if there is a relationship between these two variables.

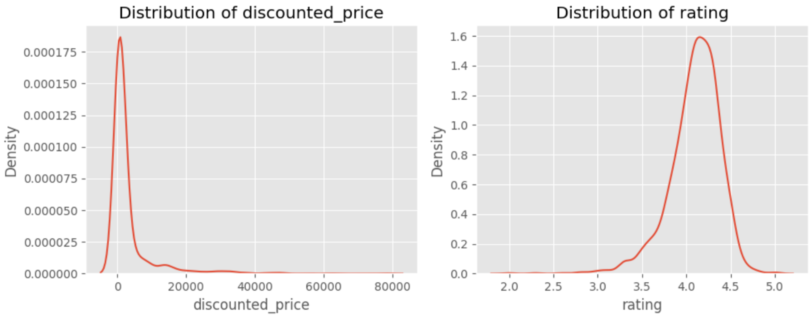

Figure 1: Distributions of **discounted_price** and **rating**

Distribution of actual_price:

Analysis: By looking at Figure 2, we find out that distribution of **actual_price** is skewed highly skewed as well. This could tell us that this variable is not normally distributed.

Relevance: This analysis would be important when we want to decide on what range of prices are low, medium and high later.

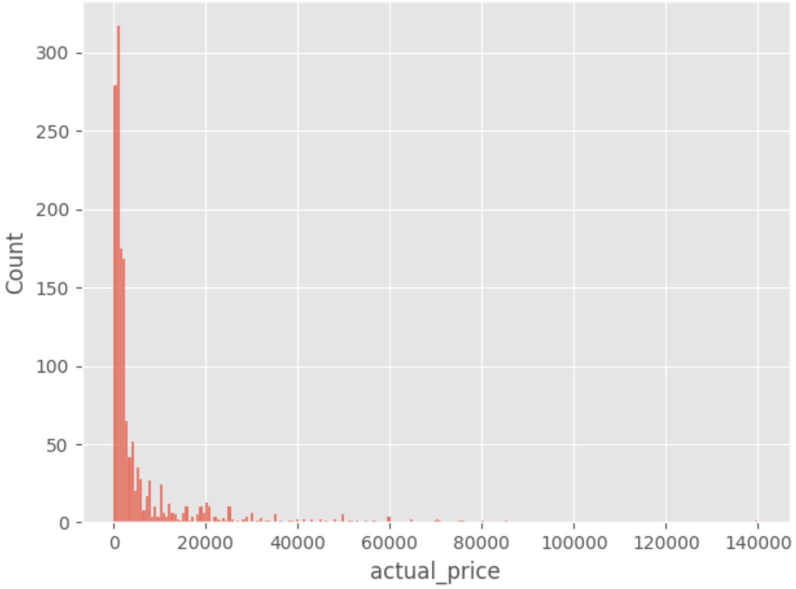

Figure 2: Distribution of **actual_price**

Distribution of rating_count:

Analysis: By looking at Figure 3, we find out that this variable is highly skewed to the right. Meaning that **rating_count** of distinct products is not normally distributed.

Relevance: This analysis would be important when we want to test our hypothesis about the normality of this variable.

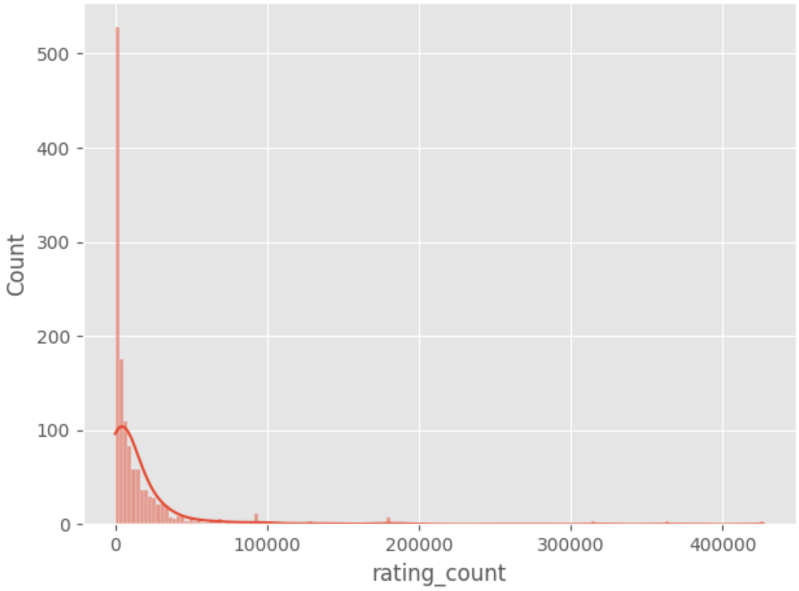

Figure 3: Distribution of **rating_count** for distinct products

Multivariate Analysis:

Scatter plot of discounted_price and rating:
Analysis: By looking at Figure 4, we figure out that there doesn't seem to be any monotonic relationship between these two variables. Thus we say that these two variables are uncorrelated.

Relevance: This analysis is connected with our prior Univeriate analysis about the distribution these two variables and would be investigated further in Hypothesis Testing section as well.

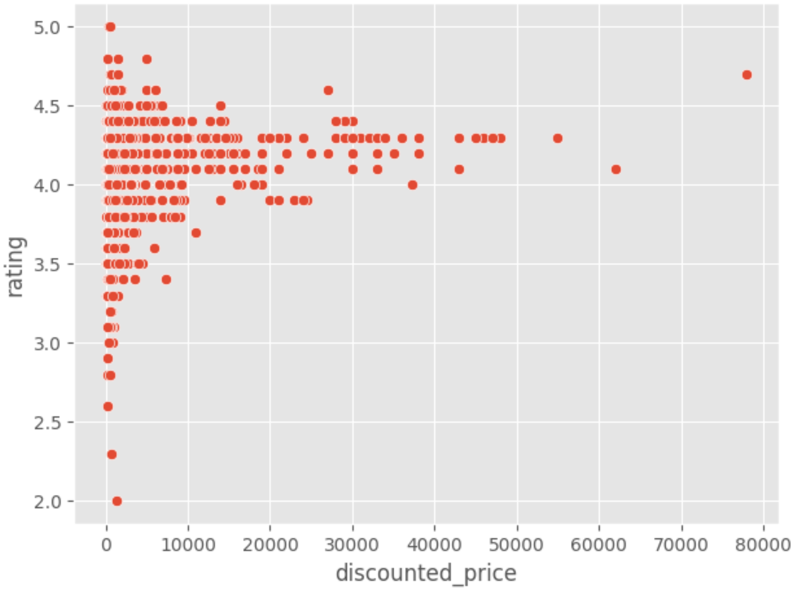

Figure 4: Scatter Plot of **discounted_price** and **rating**

Contingency Table of rating vs. category:
Analysis: By Looking at Figure 5, we find out that there are some **categories**, which are different from others in terms of distribution of **rating_levels**, namely, low and high. This could tell us that **category** and **rating** are dependent.

Considerations: We have dropped the records which are in a product **category** that has appeared less than 5 times in our entire dataset. This has been done, simple because the anaylsis of **rating_level** within these groups would be totally meaningless because of their low frequency of appearance.

Relevance: This analysis would be used as an evidence to our future hypothesis testings of examining the dependence of these two variables.

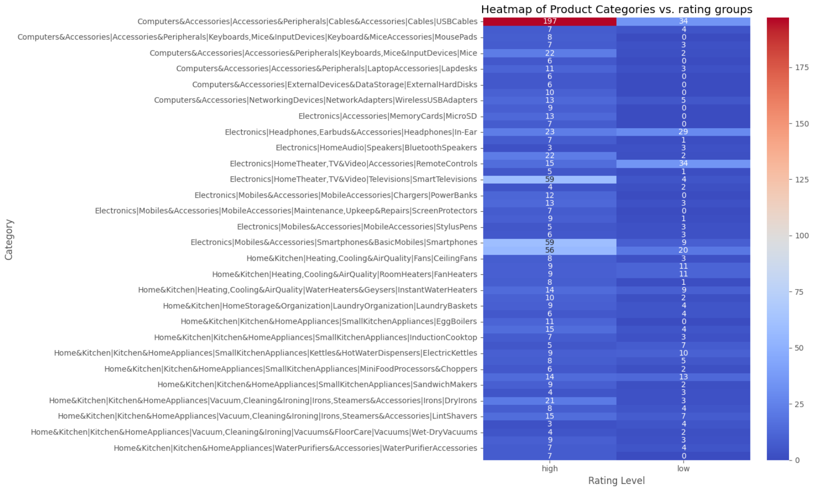

Figure 5: Contingency Table of **category** and **rating_level**

Distribution of rating for low/high discount_percentages:
Analysis: By looking at Figure 6, we figure out that there doesn't seem to be a significant difference between the means of these two levels of **discount_percentage** in terms of **rating**.

Relevance: This analysis would be used in hypothesis testing section, when we want to examine, more formally the difference between these two levels of **discount_percentage**.

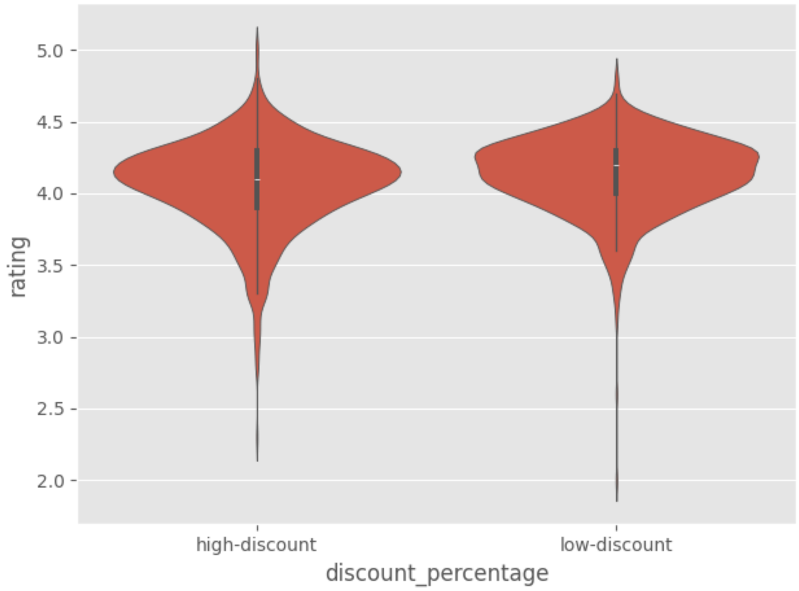

Figure 6: Violin Plot of **rating** for different levels of **discount_percentage**

Distribution of rating for different price groups:
Analysis: By having a look at Figure 7, we can say that different levels of **price**, namely, low, medium and high, don't seem to significantly differ in terms of **rating**. Also it looks that the distribution of all of the levels is approximately normmal.

Relevance: This anaylsis would be continued further, in Hypothesis Testing section, in order to examine the difference between these three different levels of **price** in terms of **rating**.

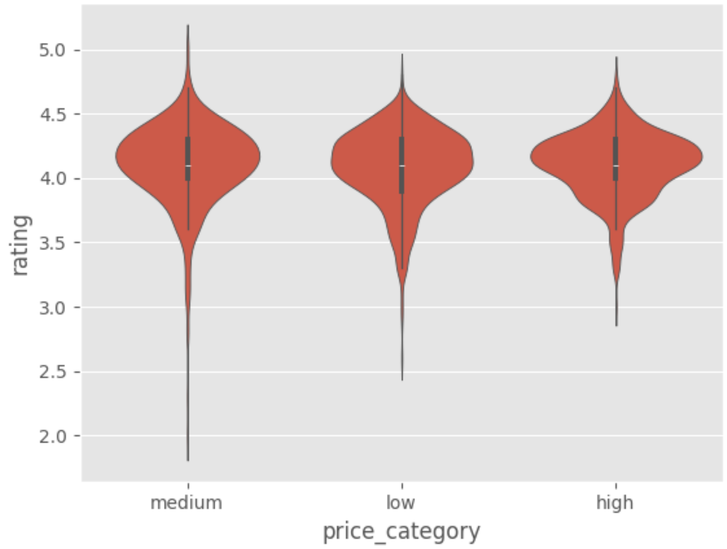

Figure 7: Violin Plot of **rating** for different levels of **price**

Hypothesis Testing:

1. Does the discounted price significantly impact product ratings?

Test used: ***Spearman Correlation Test***

Reason for why we used this test: In our earlier analysis in Exploratory Data Analysis Section, we saw that the variables are not normally distributed. Thus the normality assumption of ***Pearson Correlation Test*** is violated and we can't use it.

Result: Spearman Correlation: 0.080, P-value: 0.002

Interpretation & Discussion: We both see an exteremly low p-value meaning that their is a significant monotonic relationship between these two variables as well as correlation coefficient close to zero, meaning that there isn't a relationship between these two. It seems to be a contradiction. At the end we would say that these two variables aren't correlation because it matches with out prior analysis better.

2. Are product categories and high/low ratings independent?

Test used: $\chi^2 - test$.

Result: Chi-Squared Statistic: 213.369, P-value: 0.000.

Interpretation & Discussion: We have obtained a low p-value which matches perfectly with what we saw in our prior analysis in Exploratory Data Analysis Section. Thus we accept that there is a significant dependency between these two variables.

3. Is there a significant difference in ratings between high-discount and low-discount products?

Test used: student's two sample $t - test$.

Result: T-statistic: -4.261, P-value: 0.000

Interpretation & Discussion: This result, actually is in contradiction with our prior analysis in Exploratory Data Analysis(EDA) scetion. Thus we should have a closer look at the distribution of **rating** for different levels of **discount_percentage**. By having a look at Figure 8, we could see the difference that the hypothesis test has found(there at the picture, we could clearly see a difference that's about 0.4) and convince ourselves that our analysis was wrong. Thus we accept a significant relationship between **discount_percentage** and **rating**.

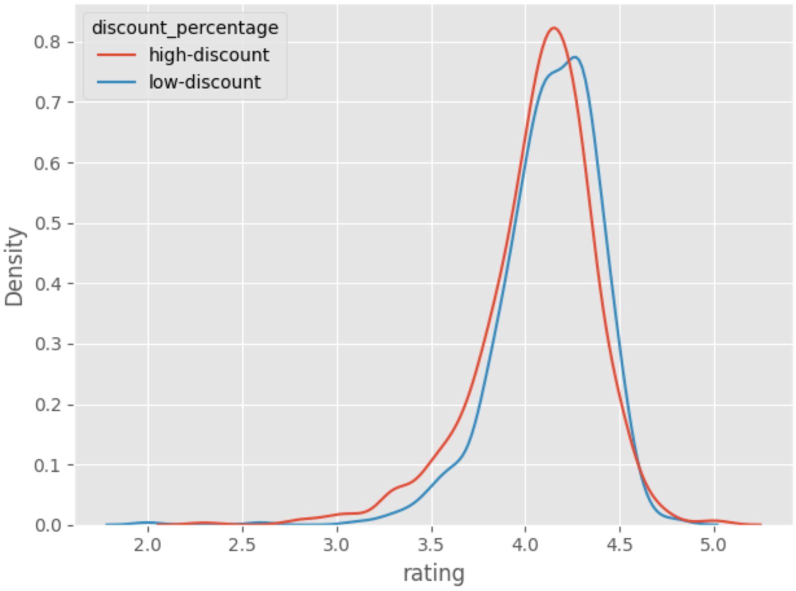

Figure 8: Distribution of **rating** for different **discount_percentage** levels

4. Do more expensive products receive higher ratings on average?

Test used: ANOVA one-way test.

Considerations: We have used 33rd and 66th percentiles to divide the range of **price** into three different levels, namely, low, medium and high. We used this method for division because the data is skewed and this methid is robus to skewness of the data.

Result: F-statistic: 0.907, P-value: 0.404

Interpretation & Discussion: This matches with prior analysis from the Exploratory Data Analysis section. Thus we conclude that there isn't a significant differnece between different levels of **price** in terms of **rating**.

5. Does the distribution of rating counts follow a normal distribution?

Test used: Shapiro-Wilk's test

Result: Shapiro-Wilk Test Statistic: 0.407, P-value: 0.000.

Interpretation & Discussion: This tells us that **rating_count** of distinct products is not normally distributed and it actually matches with our prior analysis in Exploratory Data Analysis section.