# \_\_Module 9 Task 1\_\_

## MSAI7\_20.1\_9\_6

## 7th May 2025

### Joseph Mario Maggi

# Step 2: To call the data file in which the dataset is found

### Pathname in mac is done by pressing option key on the file and it then provides a window from where to choose to copy filename.

### \_\_In this case the filename is: '/Users/josephmaggi/Desktop/Masters AI/Module 9/Mod9/datasets/laptops\_dataset\_final\_600.csv'\_\_

# Step 3: To explore the data set.

## Observations from the data

### 1. Data Review.

#### DATA SET REVIEW: The dataset is made up of 24,113 rows and 7 columns of variables

#### DATA REVIEW SAMPLE: First Few Rows and last few Rows.

##### The review of this data shows that most data is descriptive. You have:

##### A. Product Name: Descriptive: This is self explanatory - the product being reviewed.

##### B. Overall Rating: Quantitative: This is the averaged overall rating across all product reviews for this product.

##### C. No of Ratings: Quantitative: How many ratings were received for this product.

##### D. No of reviews: Quantitative: How many reviews were received on this product

##### E. Rating: Quantitative: This is the each individual rating for each product item reviewed.

##### F. Title: This is the title put in by the reviewer for his review.

##### G. Review: This is the actual review comment. From reviewing the first and last few rows, it is appears that the reviews submitted are detailed and unique.

##### It is also clear that the data was provided already grouped by product name, though not in any particular ascending or descending order.

#### EMPTY CHECK. All datafields are filled in. Therefore now one has to check whether there is duplication of data or not.

#### RUNNING THE UNIQUE DATA CHECKS. When considering the dataset in hand, the following can be observed from our data:

##### A. product\_name 365. This means that we have 365 unique products.

##### B. overall\_rating 16. We have 16 unique values as an overall rating, this currently being read across all products.

##### C. no\_ratings 288. This is the number of unique ratings received in total. This is less than the total number of products which might mean that there are products which had the same number of users who gave in their review or reviews were duplicated by error.

##### D. no\_reviews 147. This is the number of unique reviews received in total. This is less than the total number of products which might mean that there are products which received the same number of reviews were duplicated by error.

##### E. rating 5. We have 5 unqiue values as a rating give (most probably from 1 to 5)

##### F. title 213. 213 different titles put in or chosen by reviewers

##### G. review 11223. Unique reviews. Put in. When one considers that a comment section is a generic field and what low probability it is that one reviewer describes their experience in the same exact words, one starts considering that the fact that perhaps the dataset should be of 11,233 rows representing 11,223 reviews. The difference to 24,113 rows means that the data was duplicated.

#### UNDERSTANDING THE DATA FURTHER.

##### At this stage, one has to understand better whether with the data in hand, to be able to answer the requirement of the task at hand:

#### \_\_Each of you will be asked to select a particular salient data point that is of interest to you (such as which are the factors contributing to a good review for a laptop in general) and explain it in the light of your own theoretical explanation of why this could be the case. Your notebook will provide an outline of different laptop models and reviews for that model from flipcart, indicating what factors governed the likelihood of a good or bad review for a particular laptop model. These factors are to be extracted and clearly explained, with the data backing up the selection of these factors clearly shared by the student in their presentation\_\_

#### NEXT STEPS

##### A: Produce a summary of the data as per unique reviews, to try and understand whether such generic review description is a justifiable or not.

##### B: Produce a summary per product name, reducing the size of the dataset and making it more manageable to correlate any leading factors in relation to good or bad reviews.

##### C: Extract also the laptop brand from the description, to understand whether these contribute to the reviewer's perception and resulting review.

## Step 4: Data cleaning and analysis.

### ##### A: Produce a summary of the data as per unique reviews, to try and understand whether such generic review description is a justifiable or not. The below focuses on understanding the data duplication / or repetition

### OBSERVATION:

#### How can the duplicate reviews irrespective of product be less than the duplicate reviews per product?

##### The 1st check: This checks for duplicate review text across all products. Then it keeps only one instance of each repeated review text.

##### The 2nd check: This checks for duplicate review text per product. It considers the same review text multiple times across multiple products.

##### So if the same review appears once for Product A and once for Product B, its not a duplicate in the first check hbut it is treated as duplicate in the second check, because it matches the same text per product.

##### The purpose above was to be able to understand duplication. However there could be triplication of data in certain cases. However it could be that one user put in the same comment across two products. So one has to be careful not to cancel out completely duplication. Therefore:

##### first and formeost duplication is the wrong term, for it means twice. It should be repetition.

##### Always taking into consideration that we are assuming that the review, being generic, is a rare probability to encounter unless replicated intentionally by the same user.

##### We require to understand how many repetitions we have, a. across all the dataset irrespective of the product - this could be mean the same user/reviewer inputting the same comment on multiple users or the data set was mishandled and reviews were copied, b. across the dataset grouped per product.

#### Aim is now to produce a summary highlighting the repetitions of reviews across the dataset and a summary highlighting the reviews repetitions per product

## OBSERVATION from the Review statistics:

#### There are 5,895 unique reviews. The rest are found repeated. Therefore further investigation on what these duplicated comments are is necessary.

## OBSERVATION:

#### When grouping the review comments together, it becomes clear that whilst there can be repetition of datasets present, one can conclude whether this is data errors or not. For example: Good was written 753 times. One cannot assume that good should be unique. Therefore this check, though necessary shows that trying to clean out repetead data and filter out from reviews is inconclusive.

#### Therefore we have to proceed with grouping the following fields to understand if now one can relate between product name, overall rating, number of ratings received, no of reviews and averaged rating.

### This matches the plan labelled B in the above narrative, to produce a summary per product name, reducing the size of the dataset and making it more manageable to correlate any leading factors in relation to good or bad reviews.

## OBSERVATION: Variation between user rating and overall rating.

#### When one reviews the data, it is immediately apparent that the overall rating is different from the average rating.

#### This could be as a result of multiple factors, where the overall rating is not a direct relation to the data of this dataset but considers under variables such as age of review, number of reviews received (therefore considering statistical sampling confidence criteria) same brand different model reviews etc.

#### A separate evaluation and research would be required to value this. However, a rating discrepency percentage has been calculated to be able to view this difference.

#### From the preliminary review of this data, the % difference varies significantly and no direct relation was observed between overall rating and no of ratings or reviews received. Therefore at this stage this evaluation is also being considered as non conclusive.

#### However the below code saves this information in a specific csv file named product\_summary\_sorted for any future reference that would be required

#### The scatter plot was carried out to visually depict the difference between overall rating and average user rating. As one can see, that whilst there is a significant number of datapoints which are close to the red line which represents when the overall rating and the average rating coincide, there is also othere where the overall rating is diverges considerably from the user rating.

#### To understand the extent of this variation, a histogram between these 365 product\_names and variation was plotted. A normal distribution with a slight right shift was observed, showing that most product names have minimal discrepancy. Only around 7% show discrepencies beyond plus/minus 2 standard deviations.

## OBSERVATION:

#### When one considers the above, this variation between overall and average rating may result in being a key concern for any decision support system for the bias being observed can result in ethically unfair messages. Further considerations and clarifications would be required to be requested to Flipkart, the provider of the dataset to understand better the situation.

#### At this stage, this evaluation of this variation stops here for though there is variation one can still evaluate further with the current dataset whether one can make use the overall rating, which in its intent is to give a 'summed up' review of the strength of the product can be used to understand and mirror the reviewers opinion on the brands and the products they produce.

### Next Stage is Stage C mentioned above. To Extract the laptop brand from the description and compare against the overall rating, to understand whether these contribute to the reviewer's perception and resulting review.

## OBSERVATION

#### One can see from the table that for a good amount of brands have a distinct average. When one compares as the spread between the Max & Min ratings, there are also gaps on which one can distinguish between one brand and another. This is now going to be hereunder as follows:

##### Bar graph showing brand versus average rating

##### Box plot, depicting the overall rating per brand showing the distribution of the ratings.

## Step 5: Analysis Summary

#### The relationship between the brand and the overall rating is a good salient data point to monitor for it is a factor which captures the perception of the public, their hopes and expectations and the resulting good or bad reviews.