Data Analytics Interview Question

Data analytics plays a key role for understanding customers and products in retailing systems. Companies such as Amazon, Hudson's Bay, and Target strongly relay on data to design better services and improve their day-to-day operations.

You are provided with data reporting product sales of one of the largest worldwide retailers. This is a real dataset where the company name and products were anonymized for confidentiality. Each row corresponds to a product, and the data contains the following columns:

- sale_price: price at which product is sold on the website (in $);
- retail_price: regular price in the market or other stores (empty if the same as price); - units_sold: number of product units sold at Wish.com;
- uses_ad_boosts: if the seller paid for product ads or better placement in the platform;
- rating: average product rating (minimum is 1, maximum is 5);
- merchant_rating: merchant's average rating (minimum is 1, maximum is 5);
- page_views: number of page views the product received on the platform

Answer the following questions using R or Python:

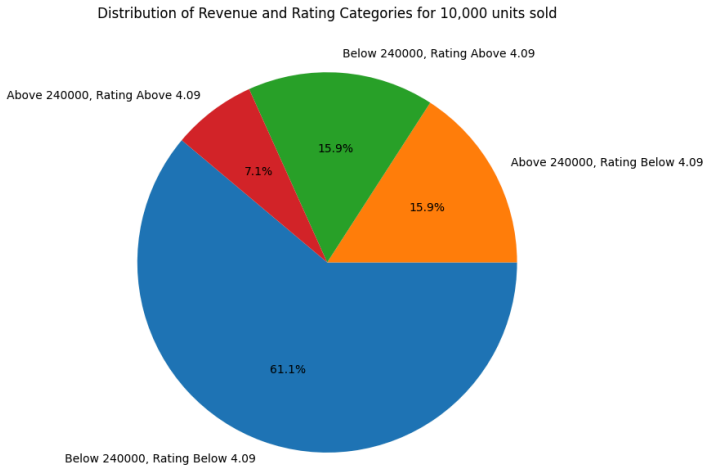
1. Clean the columns sale_price, units_sold, rating, and page_views. What are the issues that you identify?

Potential problem	How I fixed it
Checking to see if their are null and n/a values in the data set	Page_views section had 1 null value and 1 n/a value. I removed the row that had the value, as the data set is large and removing one row

	wouldn't change the results drastically my law of large numbers.
Checking if sales price has dollars for all values	Used a lambda function and apply function to make sure all sales price values have dollar symbol
Check all values in rating are between 1 and 5	All values in rating where between 1 and 5
Checking if units sold are in an acceptable range	I checked to see if the maximum number of units sold (100000) and I compared it to the mean (7641.4857142857145) and the min (100) I found there were 6 values with 100000 as the max and 23 with 100 as the min. I computed the interquartile range for the values and I got Q1: 1000.0 Q3: 10000.0 iqr:9000.0. The upper bound of this data is 23500 so the max is an outlier and the lower bound is contained within the data set implying it is not an outlier. Whether to delete the maximum depends on gathering more specific information
Checking if page_views are in an acceptable range	Mean , Min, Max for page views (10031,13705,7011) the IQR range and values Q1: 9353.5 Q3: 10696.5 iqr:1343.0 this implies that both of the max and min are potential outliers. Whether to delete the maximum depends on gathering more specific information

2. A product is considered popular if it sells more than 10,000 units. Are popular products necessarily the ones with highest revenues and better ratings?

For this I decided to segment the rating value about quartile 3 as high ratings so a rating above 4.09 and revenue above quartile 3 as high revenue so revenue above 240000. I created a pie chart to represent the data



From this we can see that the units that sold above 10000 units tend to underperform in rating and revenue, so popular products are not necessary he ones with the highest revenue and better ratings

3. A manager said that the current ad strategy is effective in increasing page views, which in turn helps increase the number of units sold of a product. Are those indeed true?

To answer this I need to answer two different questions:

Does ad strategy increase page views ?

From completing the analysis we see that ad boost gets 30 more viewers on average. The products that don't use a boost tend to have a higher variance than those that do. Although the statement is true the difference ads bring to page views is marginal.

Do page views increase the total unit sold of a product?

Since we have showcased that ad boost does increase page views lets see how does this affect the total unit sold. By measuring the mean units sold of ad boosted and non ad boosted mean units sold we can get a view into whether this relationship does exist. Mean ad not used 7657.699805 Mead ad used 7625.068871. This shows that using ad boost whilst it does increase the amount of page views doesn't increase the amount of units sold. However the percentage of ad boost that becomes popular is 35.5% whereas no ad boost popular is 33.92%, thus ad boosting influences popularity.

4. Design a classification model to decide whether a product will become popular or not. Discuss the features that you selected in detail, the quality of your prediction, and how you would implement your model in practice.

Based on this I want to create a binary classification with the following features: revenue below 240000, does the product have a rating below 4.09 and checking if ad boosts are used. To ensure accuracy in the model I made sure to separate the data in test cases and training case's this was a necessary step when it comes to computing the cost function. I used the Randomforrest classifier as it provides high accuracy and ensures that overfitting is minimized.

	precision	recall	f1-score	support
Unpopular	0.88	1.00	0.94	179
popular	1.00	0.40	0.57	40
accuracy			0.89	219
macro avg	0.94	0.70	0.75	219
weighted avg	0.90	0.89	0.87	219

The model I created was very accurate with a score of 0.89.From the accuracy table we can see the model was very good at identifying when something was going to be popular as its precision was perfect, the popularity table was weaker in the recall section. The unpopular model performed very well, indicated by the f1 score, there may be issues in the future when it comes to fitting to new data.