Faculty of Engineering
Computer Engineering Department

Phase 2 Report

Big Data

Semester Project

Presented by

Name	Section	BN
Mohamed Yehia Alahmad	2	14
Moustafa Mohammed Elsayed	2	23
Mostafa Hani Mostafa	2	24
Mennatallah Ahmed Moustafa	2	25

Project Idea – Predicting Product Sales in E-Commerce

I. Brief problem description

This project aims to predict the performance of products on an online shopping website. The system will provide accurate performance forecasts by looking at factors like price, ratings, and sales per week. This will help vendors better manage their inventory and marketing planning.

Dataset

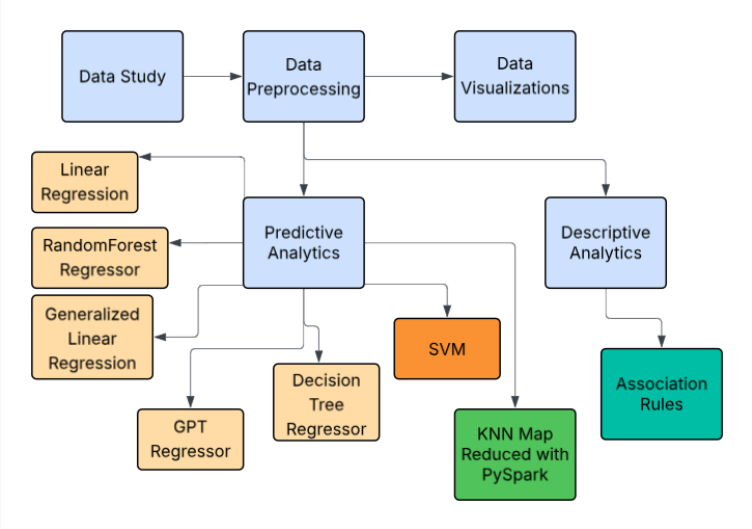
The **BaSalam dataset** (products file) contains detailed information on products from the BaSalam online marketplace

Dataset Description

#	Field Name	Description
0	_id	Product ID
1	_score	A metric for the product performance
2	sales_count_week	Integer (nullable = true), weekly sales count
3	name	Product name
4	price	Current product price
5	status_id	Product status ID (0 or 1)
6	status_title	Product status (e.g., available or null)
7	stock	Number of items stored for the product
8	photo_MEDIUM	URL of product picture (medium size)
9	photo_SMALL	URL of product picture (small size)
10	rating_average	Average rating of the product
11	rating_count	Number of ratings for the product
12	rating_signals	Number of signals contributing to the product rating
13	primaryPrice	Original price of the product
14	preparationDays	Number of days before the product starts shipping
15	weight	Weight of the product
16	categoryId	0 if product doesn't have delivery option, 1 otherwise
17	has_delivery	0 if product has no variations (e.g., sizes/colors), 1 otherwise

18	has_variation	Name of the vendor
19	new_categoryId	Vendor availability status (e.g., always available)
20	navigation_id	City of the vendor
21	vendor_name	Indicates if the product has free shipping
22	vendor_identifier	True if available (then `status_title` = "available"), otherwise false
23	vendor_statusId	(No description provided — you might want to clarify this field)
24	vendor_freeShippingToIran	Price threshold: if the product price > this amount, shipping is free **nationwide (Iran)**
25	vendor_freeShippingToSameCity	Price threshold: if the product price > this amount, shipping is free **within the same city**
26	vendor_cityId	ID of the city where the vendor is located
27	vendor_provinceId	ID of the province where the vendor is located
28	vendor_has_delivery	1 if the vendor offers delivery, 0 if not
29	vendor_score	Vendor performance
30	vendor_id	Unique identifier for the vendor
31	vendor_status_id	id of status (0 = not available, 1 = available)
32	vendor_status_title	vendor status (e.g., “available”, “not available”)
33	vendor_owner_city	Vendor owner city
34	vendor_owner_id	Vendor owner id
35	isFreeShipping	Boolean or computed flag indicating if this product qualifies for free shipping
36	IsAvailable	True if the product is currently available (in stock)
37	IsSaleable	Boolean indicating if the product is allowed to be sold
38	mainAttribute	ignored(most of the data is nulls)
39	categoryTitle	Title of the product’s category
40	published	Boolean indicating if the product is published and visible to users
41	video_ORIGINAL	URL of a video demonstrating the product
42	promotions	ignored(most of the data is nulls)

II. Project pipeline

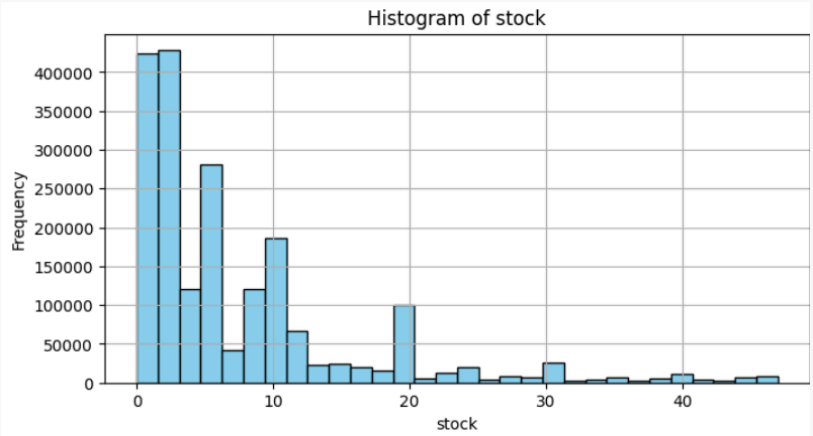


III. Analysis and solution of the problem

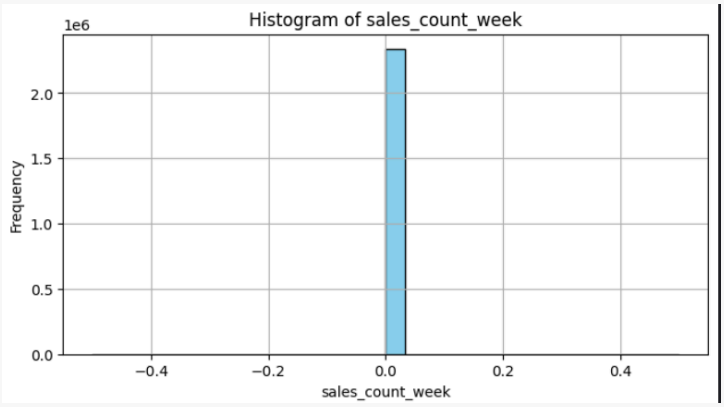
- **Data Study**
  - Discovering the type of each column in data: we have categorical, numerical, binary, boolean
  - Displaying Null counts for each column
  - Histograms for numerical columns
- **Data preprocessing**
  - Dropping columns with high percent of null values (> 0.8)
  - Splitting data into train, test
  - Removing columns that won't affect the model such as ids, photos urls ['_id','photo_MEDIUM','vendor_identifier', 'photo_SMALL','vendor_id','vendor_provinceId', 'vendor_owner_id','vendor_status_id','vendor_cityId','vendor_statusId','categoryId','new_categoryId','status_id']
  - For association rules, we used dropna to drop all rows that have nulls after that we would have
    - Original row count: 2411358
    - Rows after removing nulls: 2262039
    - Removed 149319 rows (6.19% of data)
  - But for the predictive models, we found that the remaining columns have a low percent of nulls (0.2) thus we decided to replace these nulls with a median for numerical columns and mode for categorical, binary columns
  - Dropping categorical columns with very high cardinality before one hot encoding

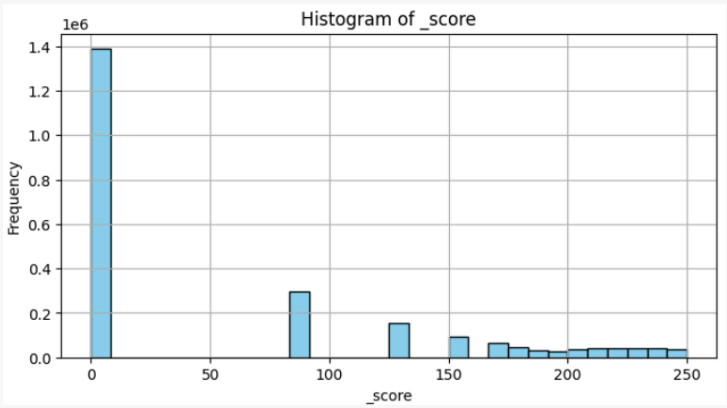
- One hot encoding categorical columns
- Drop very highly correlated columns (if the model requires that)
- Now all columns are numerical, we Calculate the correlation with each column and target and select the most correlated columns with the target

- **Data visualization.**
  - High stock is infrequent

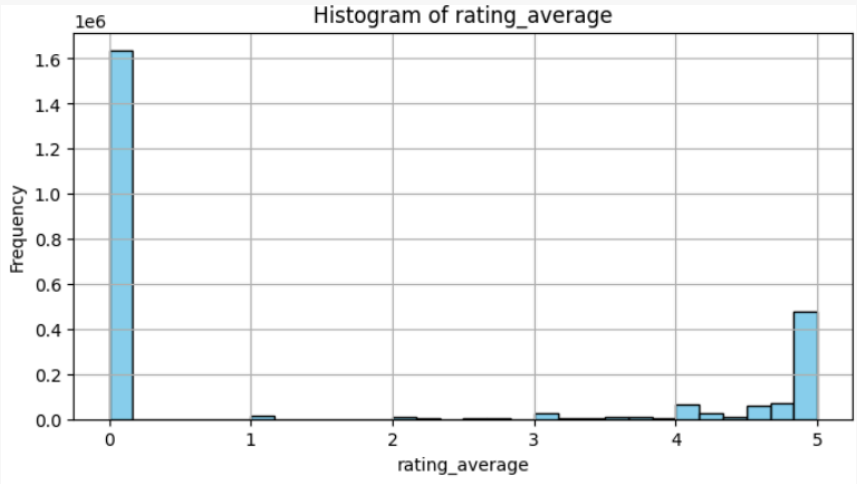


- As we can notice the distribution and values of score in the data are better than sales_count_week

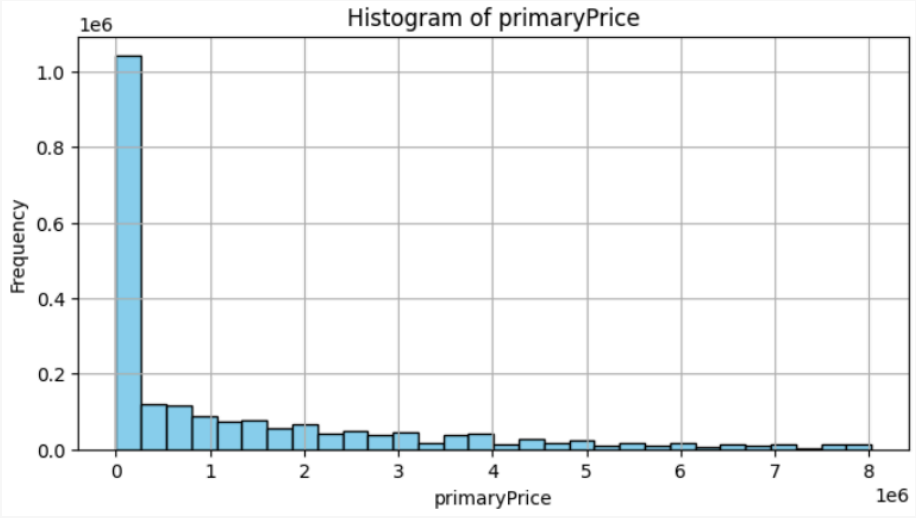




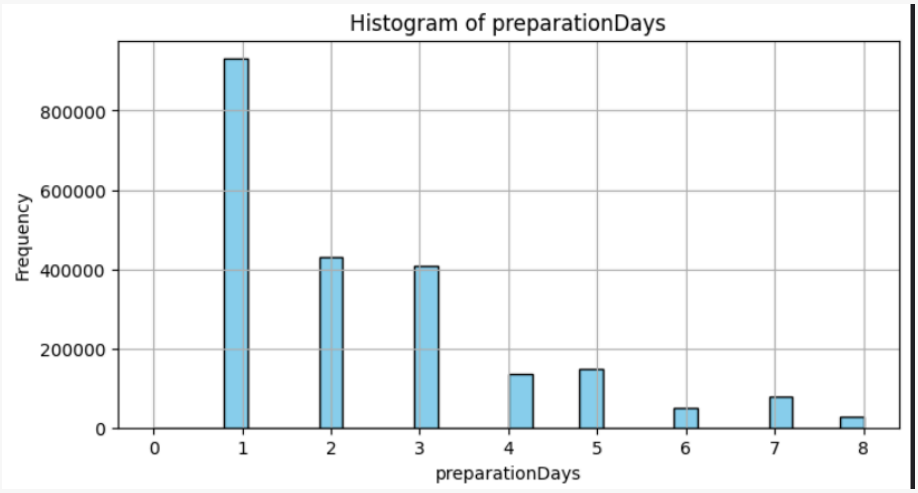
- 0 indicates no rating as we can notice that a high percentage of products have no ratings, but the rating average is accumulated between 4 and 5



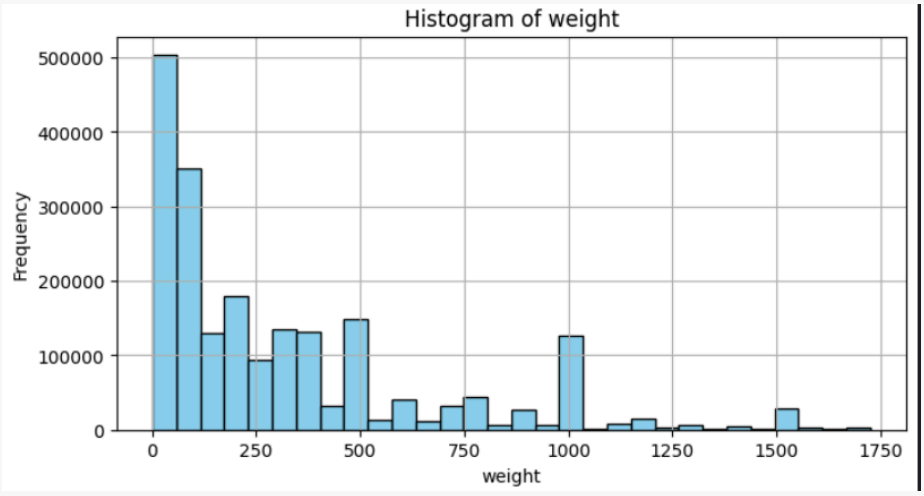
- Low primary price is more frequent



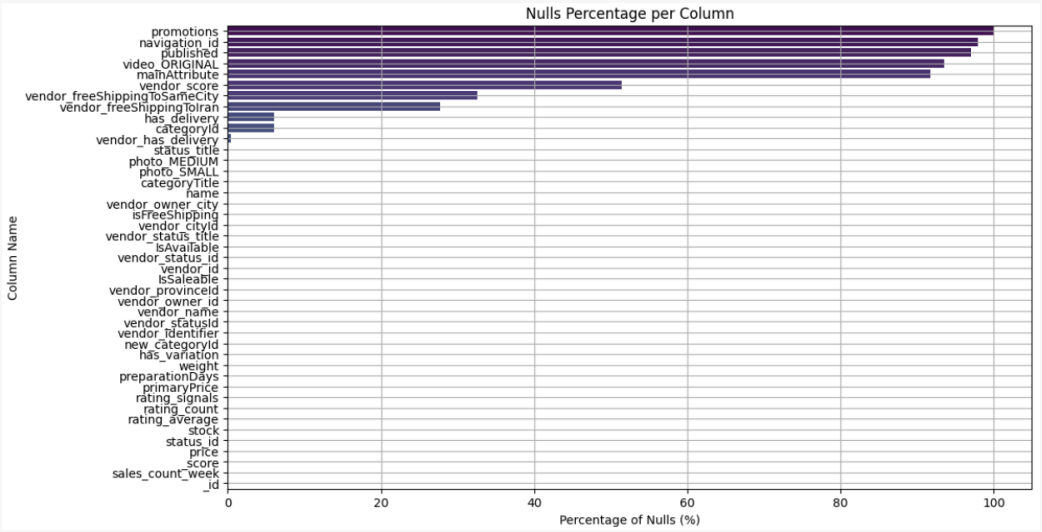
- Most of the products have 1 to 3 preparationDays



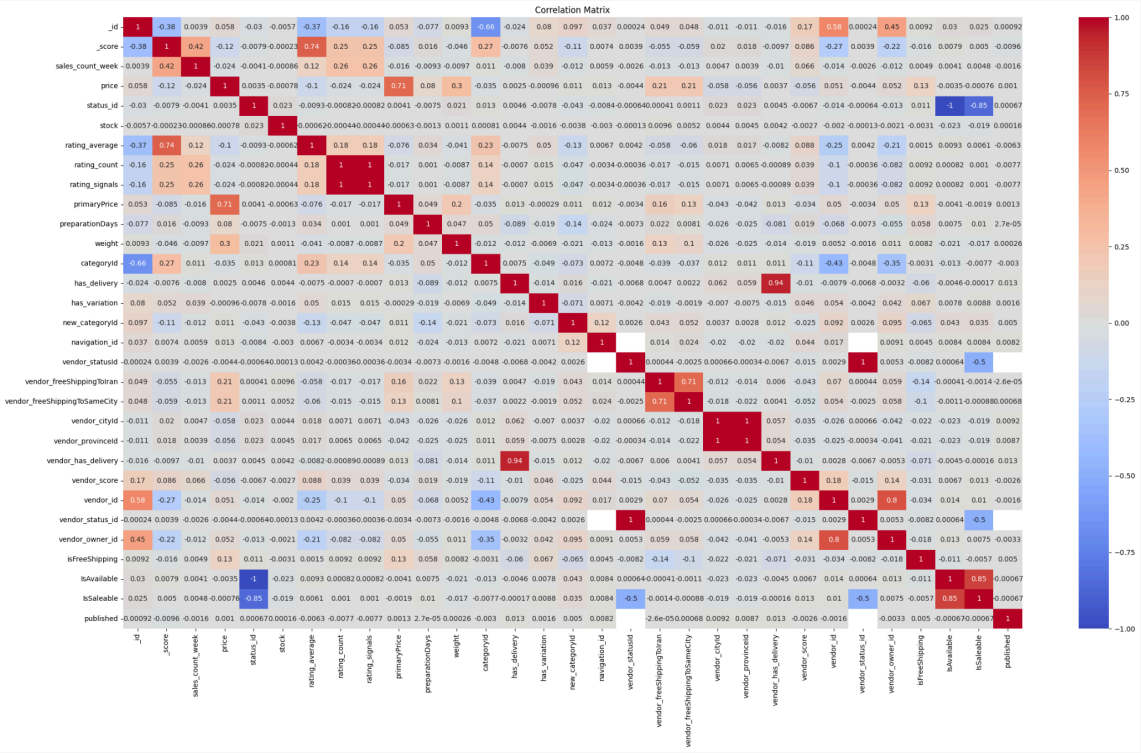
- Most of the data is low-weighted



- Null Percentage in each column



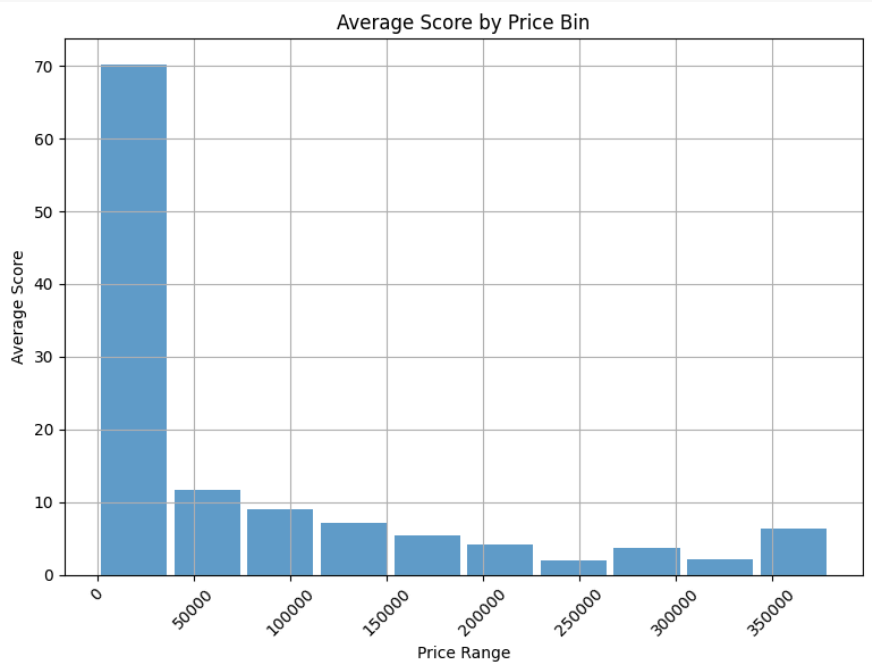
- The correlation of features



Finally, from what we have seen above about distribution and values of score and sales_count_week columns, the score is also correlated with the price which is weakly associated with sales_count_week. Also the score has a high correlation with the rating_average and rating_count and sales_count_week Thus the score represents the overall performance of the product not just the sales (which helps vendors more), so we will consider the score as our target column.

- **Extracting insights from data**

#	Insight	Columns Used	Status	Useful
1	Analyze the relationship between _score, stock, and price. -> Determine if high scores and low stock levels lead to price adjustments. -> Determine if price affects score(Are cheaper products selling more?)	`stock, price, _score`	DONE	YES



- Canceled stock vs price as they aren't highly correlated -0.00023
- Price vs score correlation: -0.12

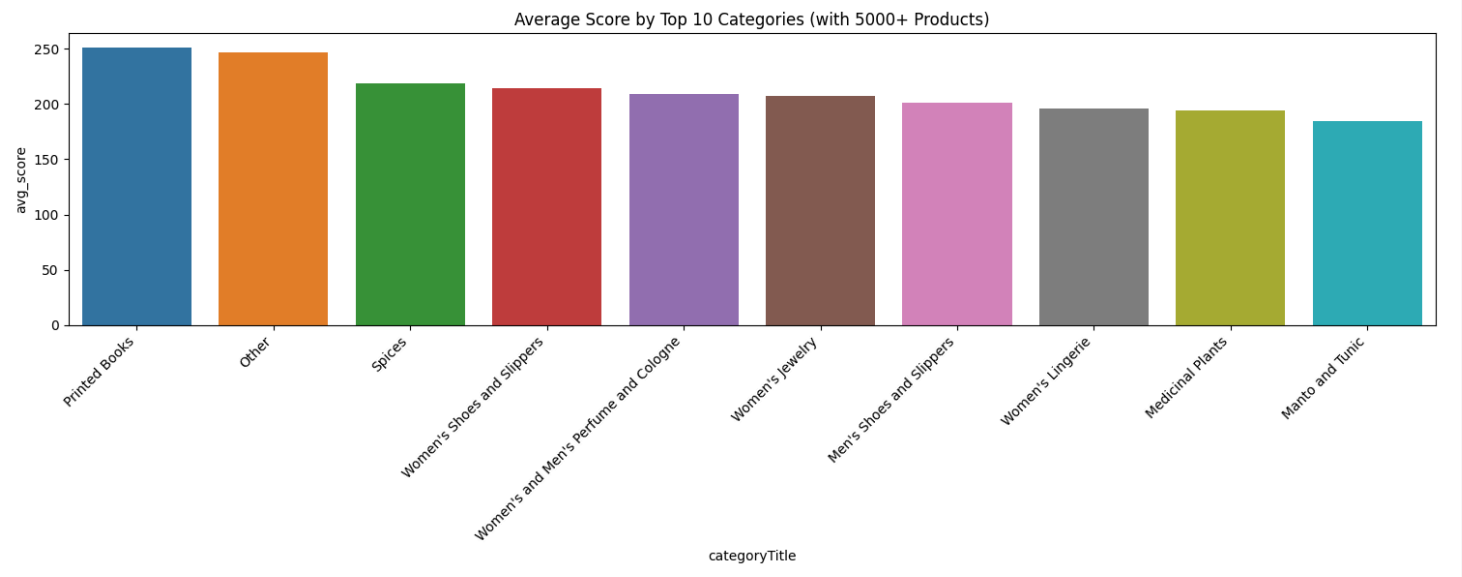
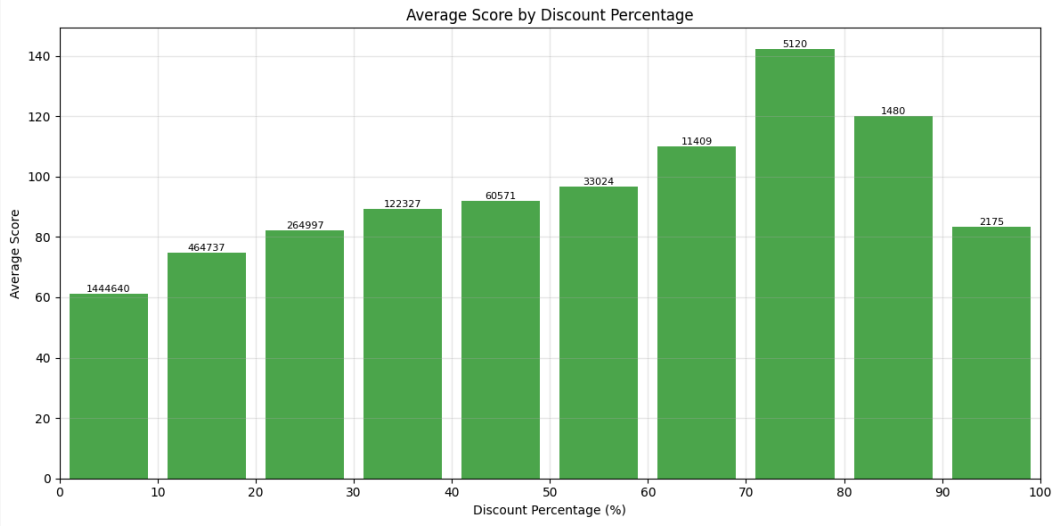
Conclusion: **Cheaper products may receive higher scores.**

#	Insight	Columns Used	Status	Useful
2	Investigate how free shipping affects product scores. -> Compare the scores of products with and without free shipping to see if it's a major selling point.	`isFreeShipping, _score`	DONE	NO

isFreeShipping	average_score	product_count
false	70.06915064404055	1962935
true	66.14196883203742	447545

Conclusion: **Free shipping doesn't affect the score much**

#	Insight	Columns Used	Status	Useful
3	Analyze product performance by category. Which categories are performing best? Does a particular category consistently outperform others in terms of score?	`categoryTitle, _score`	DONE	YES

#	Insight	Columns Used	Status	Useful
	 Conclusion: **Having a product in one of these categories leads to a better score**			
4	Analyze how the gap between the primary price (original price) and the new price (discounted price) influences product performance.	**`price`, `primaryPrice`, `_score`**	**DONE**	**YES**
	Discount vs score ● Increasing discounts make products appear more valuable ("good deal") ● Maximum satisfaction around 70-80% discounts ● Extreme discounts (especially 90-100%) trigger skepticism: ○ Customers suspect products are defective/expired, "Too good to be true" mentality 			
5	Check how products with variations perform compared to those without. Are customized or multi-option products more popular?	**`has_variation`, `_score`**	**DONE**	**YES**

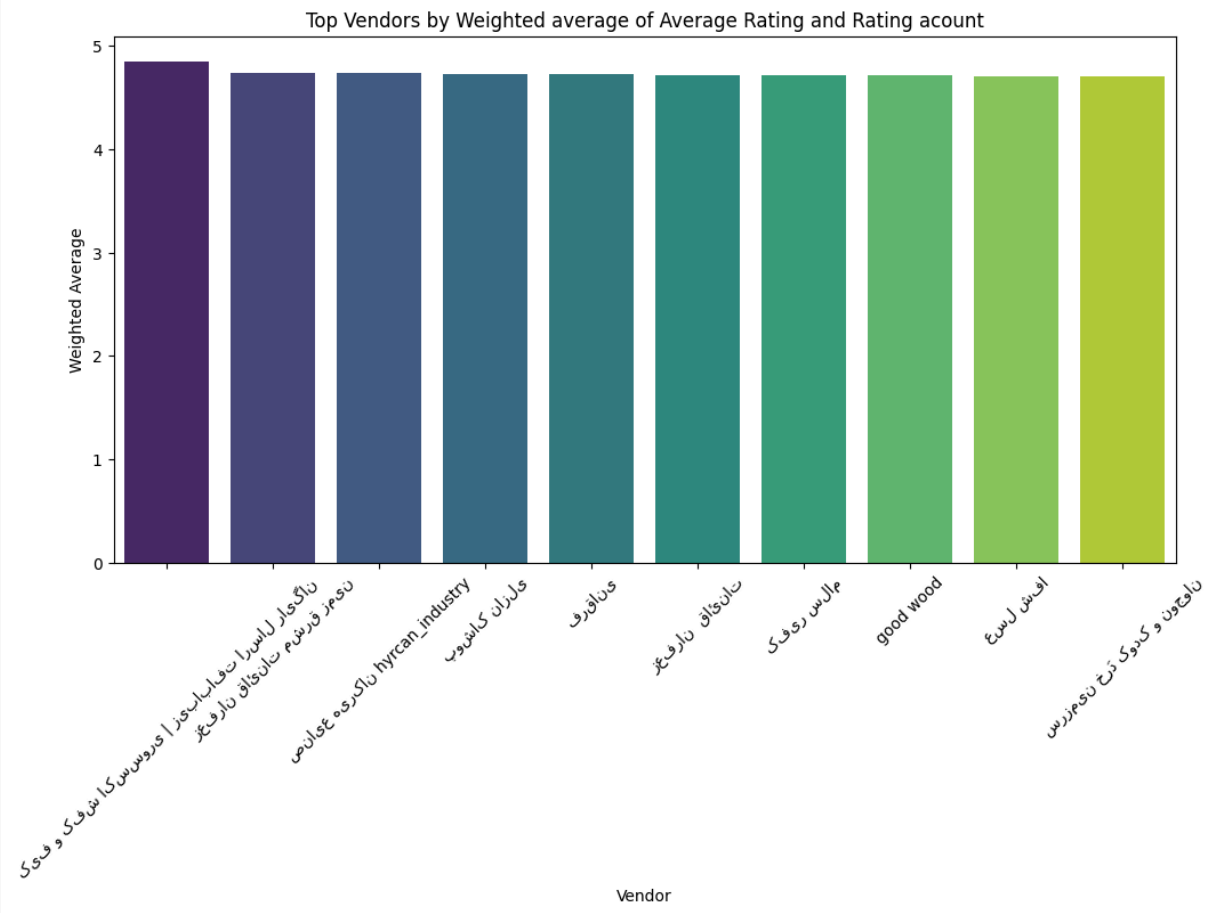
#	Insight	Columns Used	Status	Useful

has_variation	average_score	product_count
false	68.16980545988014	2282435
true	90.1991154324632	128045

Conclusion: **The average score increases when the product has variations of different sizes/colors**

6	Determine which vendors receive the highest customer satisfaction **ratings**. Are there specific vendors that consistently receive positive feedback from customers?	**`rating_average`, `rating_count`, `vendor_name`**	**DONE**	**YES**

The top 10 vendors



Conclusion: **if the products belong to one of these vendors, then the average rating of the product tends to be high**

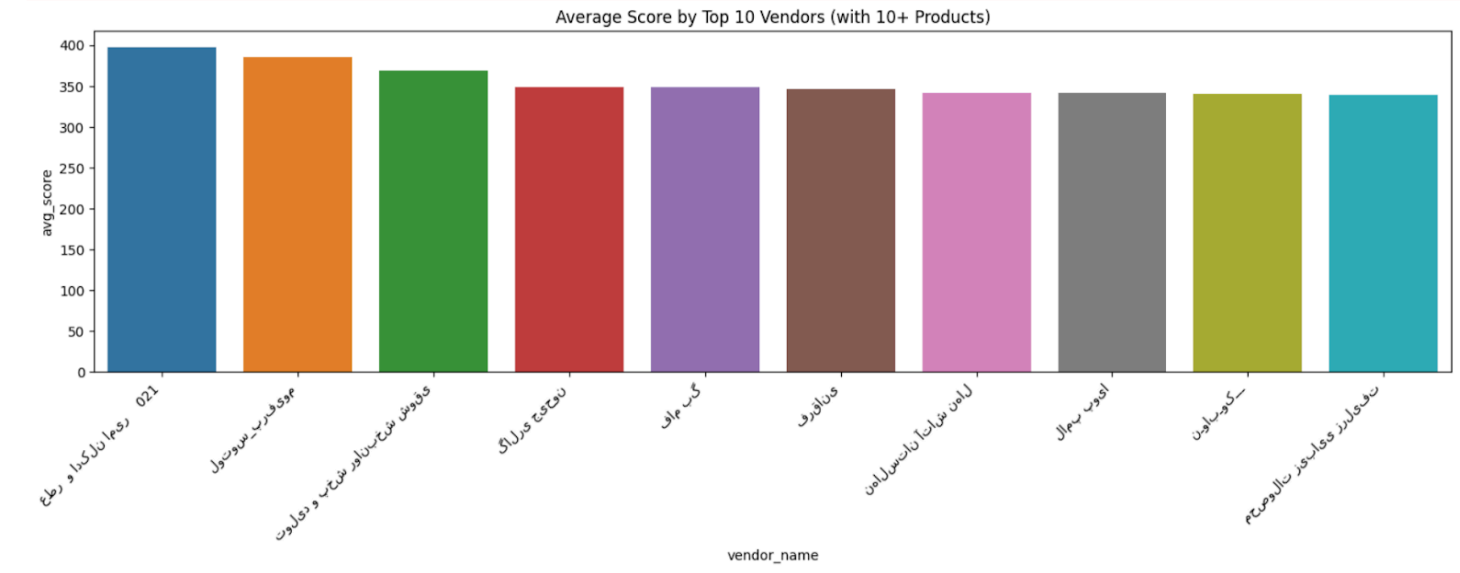
7	Investigate whether product weight affects the likelihood of offering free shipping (heavier items could be harder to ship).	**`weight`, `isFreeShipping`**	**DONE**	**NO**

#	Insight	Columns Used	Status	Useful

isFreeShipping	average_weight	average_price	product_count
false	2639.84652	2719.0726984	1962935
true	3182.76735	7034.4152267	447545

Conclusion: **High weight almost has a high price which probably has Free shipping**

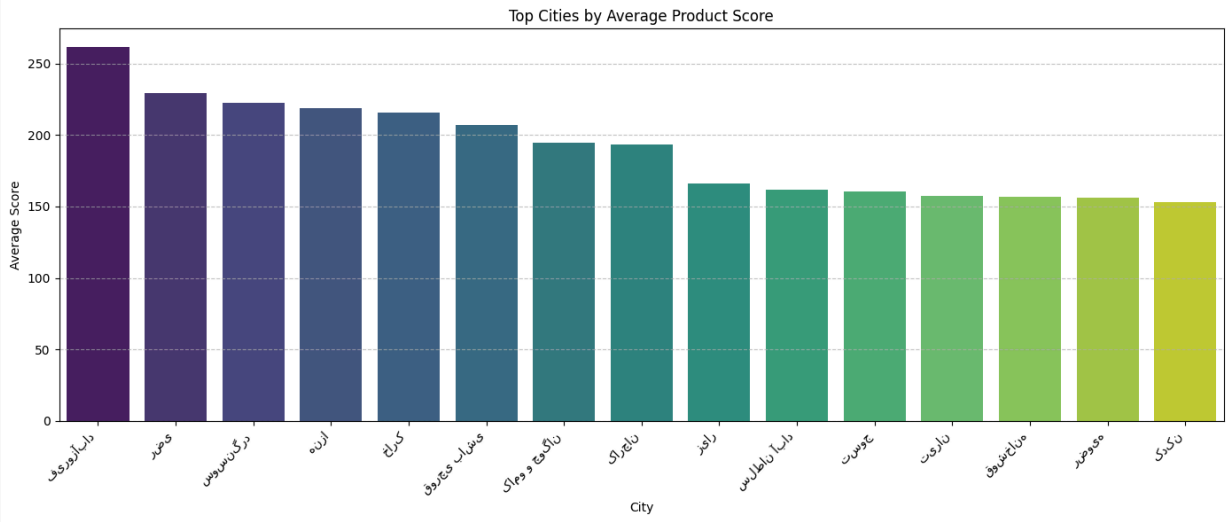
8	Identify the vendors with the highest **scores**. Are there particular vendors that consistently achieve top performance?	**`vendor_name, _score`**	**DONE**	**YES**



Conclusion: **if the products belong to one of these vendors, then the score might be higher**

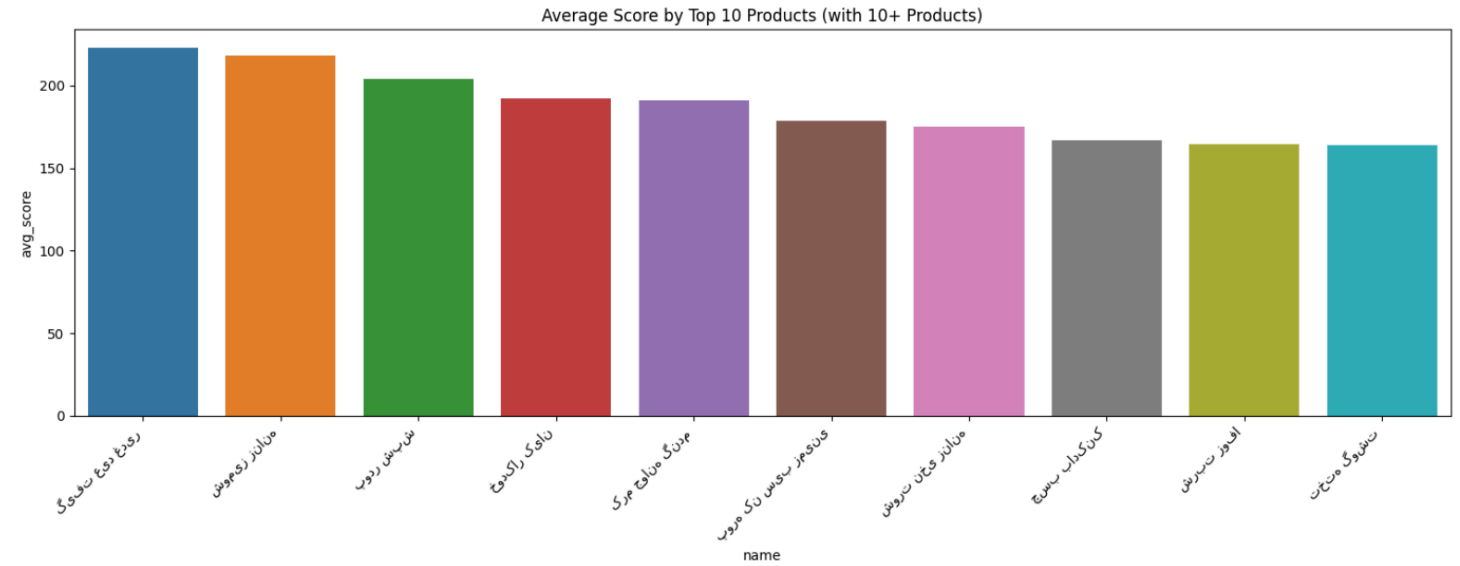
9	Analyze how different free shipping thresholds (vendor_freeShippingToIran and vendor_freeShippingToSameCity) affect _score.	**`vendor_freeShippingToIran, vendor_freeShippingToSameCity, _score`**	**NO CORRELATION**	
10	Analyze the vendor city influence on _score (Does being in a certain city make the vendor perform better?)	**`vendor_owner_city, _score`**	**DONE**	**YES**

#	Insight	Columns Used	Status	Useful



Conclusion: **If the vendor belongs to one of these cities, then the score might be higher**

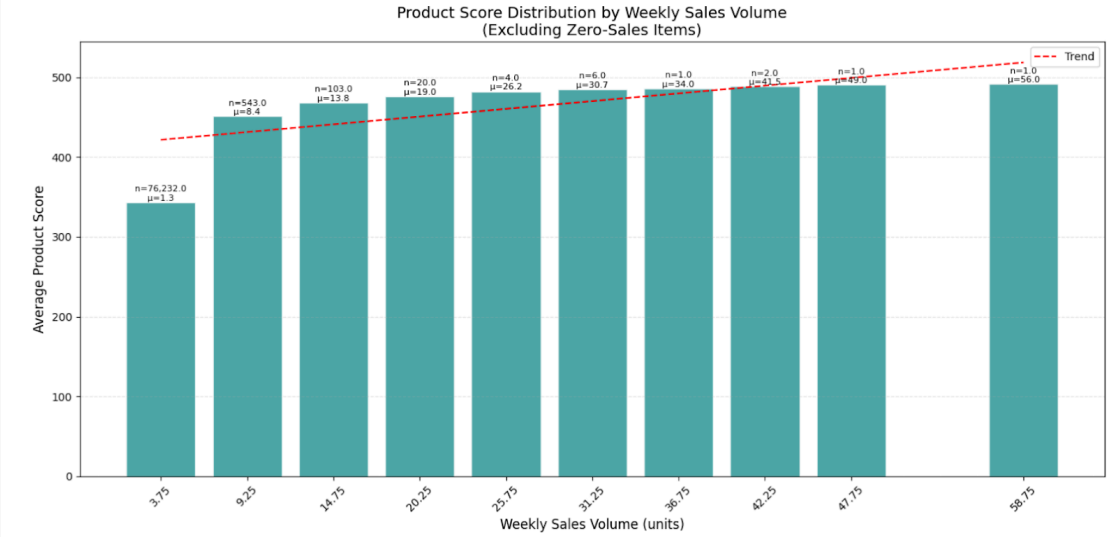
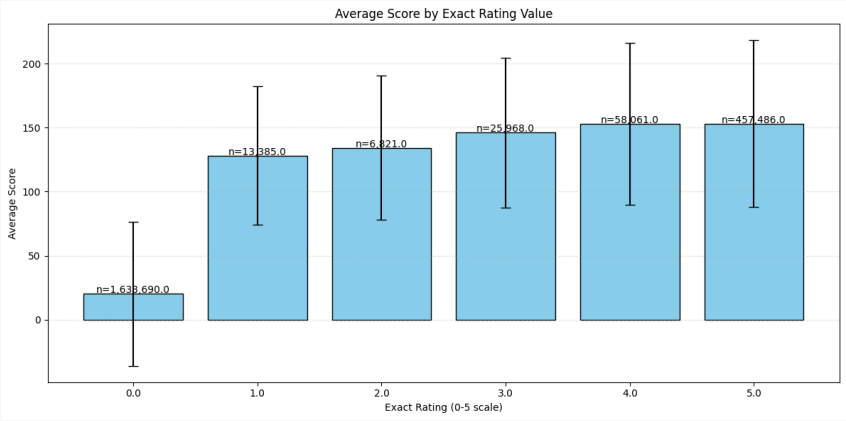
11	Identify the products with the highest score	`_score, name`	**DONE**	**YES**



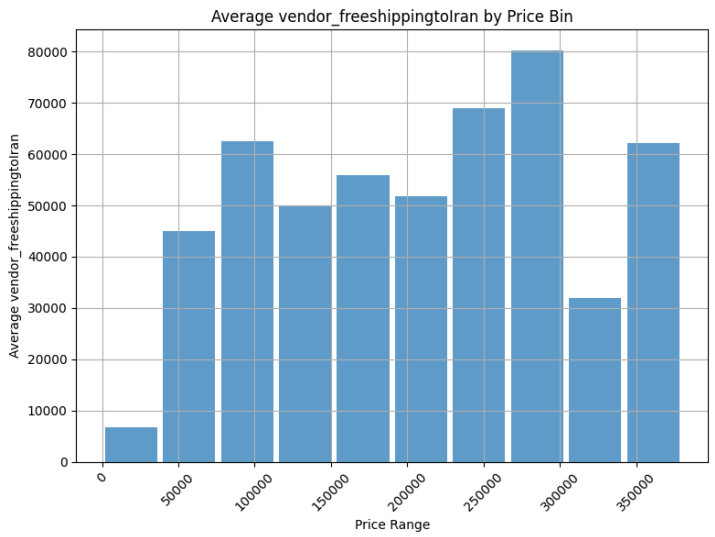
Conclusion: **The following are the products that achieved the highest performance**

12	Analyze how having delivery (has_delivery) affects _score. Determine if products that offer delivery options experience higher scores	`has_delivery, _score,`	**DONE**	**NO**

has_delivery	average_score	product_count
0	65.73936782550204	984490
1	71.83594698220772	1426868

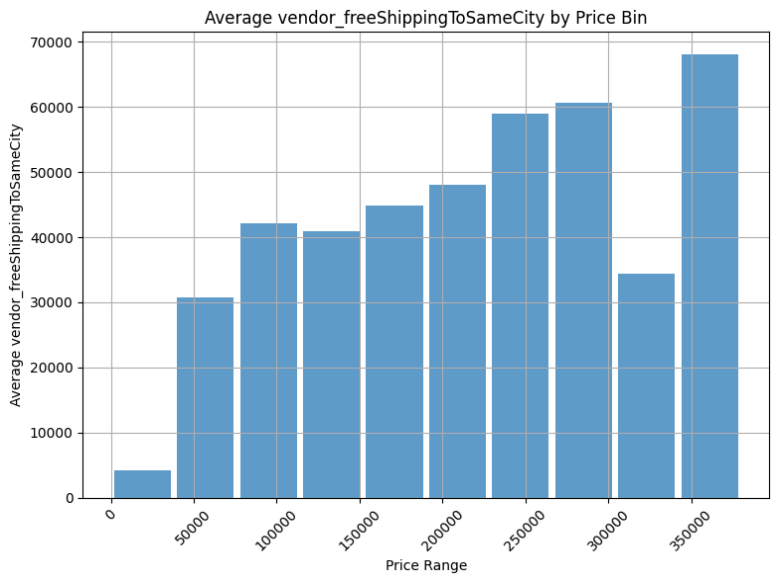
#	Insight	Columns Used	Status	Useful
Conclusion: **Having delivery doesn't affect the average score much**				
13	Analyze how preparation time affects the product score (Do products with less preparation time perform better?)	`preparationDays`, `_score`	NO CORRELATION	
14	Analyze how vendor_has_delivery and vendor_status_title (available vs. not available) affect _score. Vendors offering delivery might cause increased scores.	`vendor_has_delivery`, `vendor_status_title`, `_score`	NO CORRELATION	
15	Score vs sales		DONE	YES
				
Conclusion: **Higher weekly sales correlate with better performance scores.**				
16	Score vs Rating average	***score***, ***average_rating***	DONE	YES
				
Conclusion: **Better ratings strongly lead to higher scores.**				

#	Insight	Columns Used	Status	Useful
17	Price vs vendor_freeshippingtoIran threshold	`price, vendor_freeshippingtoIran`	DONE	YES



Conclusion: **More expensive items often have higher free shipping thresholds.**

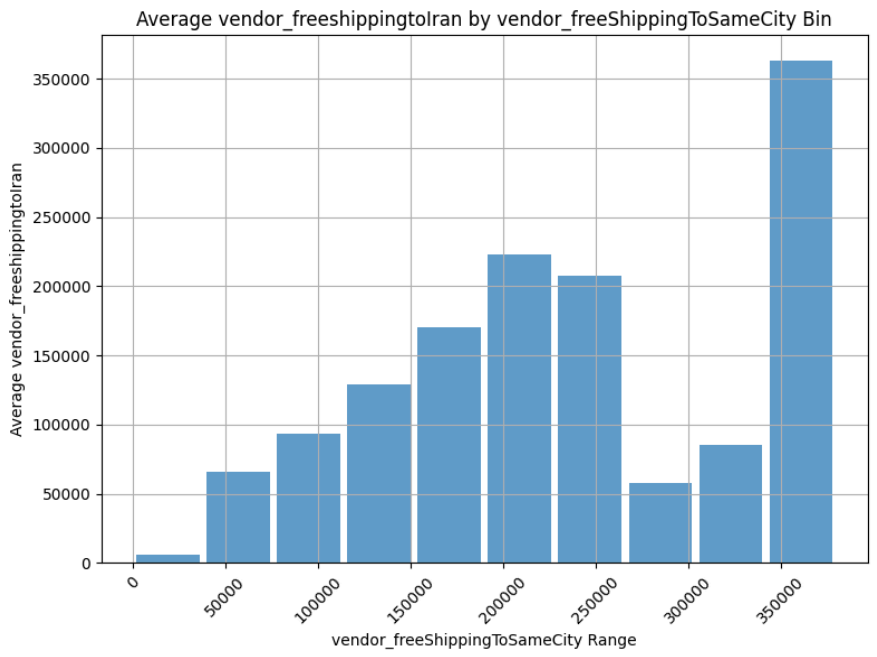
18	Price vs vendor_freeShippingToSameCity threshold	`price, vendor_freeShippingToSameCity`	DONE	YES



Conclusion: **More expensive items often have higher free shipping thresholds.**

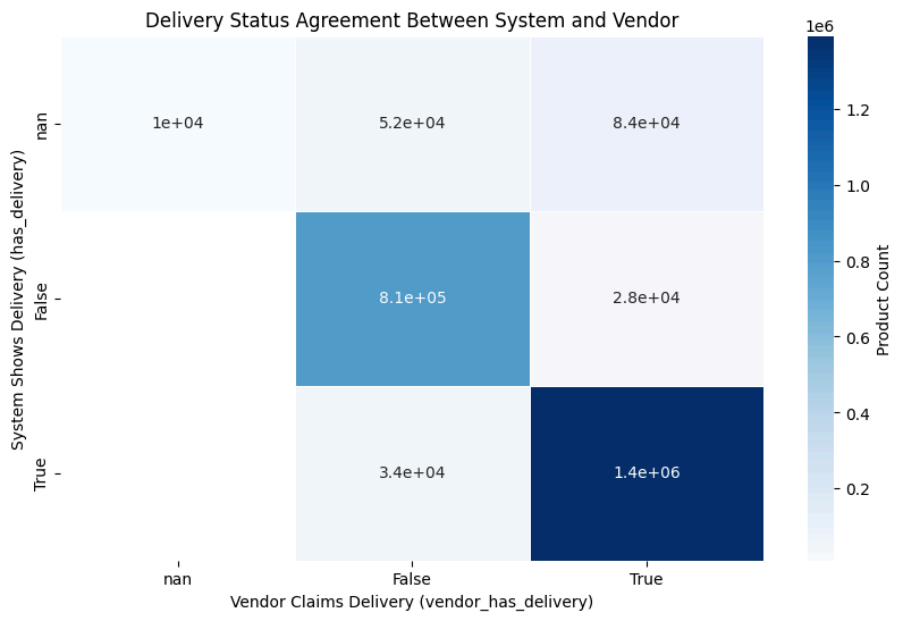
19	Vendor_freeshippingtoIran vs vendor_freeShippingToSameCity	`vendor_freeshippingtoIran,`	DONE	YES

#	Insight	Columns Used	Status	Useful
		`vendor_freeShippingToSameCity`		



Conclusion: **The thresholds for free shipping to same city and to iran changes with each other**

20	has_delivery vs vendor_has_delivery	`has_delivery, vendor_has_delivery`	DONE	YES



Conclusion: **If the vendor has delivery it is more likely that the product will have delivery too**

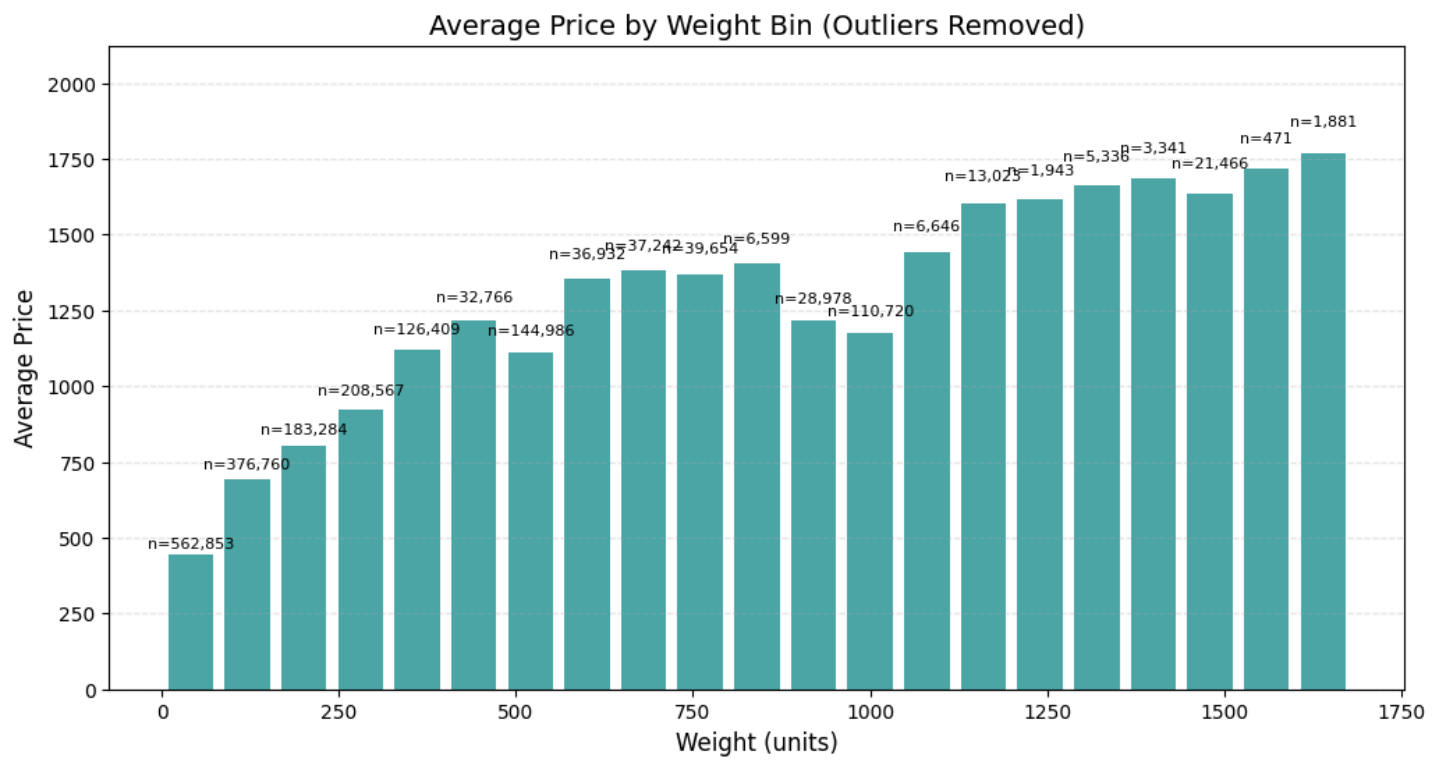
21	IsFreeShipping vs price	`IsFreeShipping, price`	DONE	YES

#	Insight	Columns Used	Status	Useful

isFreeShipping	average_price	product_count
false	2719.0726984141443	1962935
true	7034.415226796704	447545

Conclusion: **Free shipping usually happens when the product price is high**

22	Weight vs price		**DONE**	**YES**



#

Conclusion: **Heavier products tend to be more expensive.**

23	High-visibility vendors (multiple photos/videos) gets higher score		**DONE**	**YES**

has_photo	has_video	average_product_score	product_count
false	false	18.222862316793893	524
false	true	13.888889333333333	6
true	false	68.5869924810835	2134613

#	Insight	Columns Used	Status	Useful
true	true	89.28224372144084	150089	

Conclusion: **Having a picture and a video for the product achieves the highest scores**

IV. Model/Classifier training, Results and Evaluation

Model accuracy on train, test, and validation data.

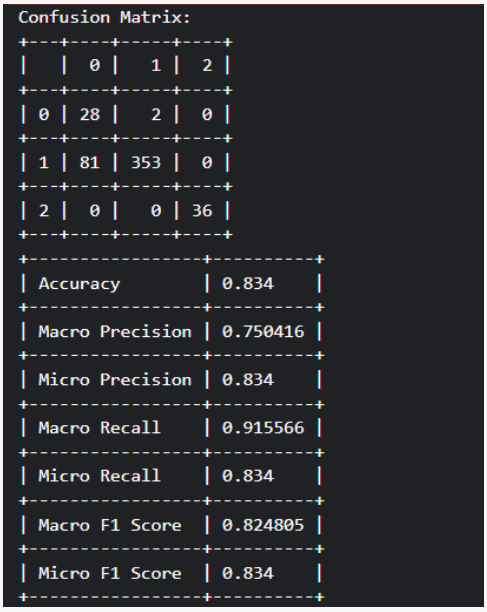
Regression Models

	Train Metrics		Test Metrics	
	RMSE	**R2**	**RMSE**	**R2**
Linear Regression	54.39	0.666	54.53	0.664
Random Forest Regressor	46.606	0.754	46.548	0.756
Decision Tree Regressor	41.52	0.805	41.37	0.807
Generalized Linear Regression	54.39	0.666	54.53	0.6649
GBT Regressor	40.90	.811	40.76	.8127
Isotonic Regression	79.708	0.283	79.76	0.283

- **SVM Model**

```
Validation accuracy for each parameter combination:
MaxIter: 10, RegParam: 0.01, Validation Accuracy: 0.8677
MaxIter: 10, RegParam: 0.1, Validation Accuracy: 0.8677
MaxIter: 10, RegParam: 1.0, Validation Accuracy: 0.8677
MaxIter: 50, RegParam: 0.01, Validation Accuracy: 0.8677
MaxIter: 50, RegParam: 0.1, Validation Accuracy: 0.8677
MaxIter: 50, RegParam: 1.0, Validation Accuracy: 0.8677
MaxIter: 100, RegParam: 0.01, Validation Accuracy: 0.8677
MaxIter: 100, RegParam: 0.1, Validation Accuracy: 0.8677
MaxIter: 100, RegParam: 1.0, Validation Accuracy: 0.8677
Best parameter combination:
MaxIter: 10, RegParam: 0.01, Validation Accuracy: 0.8677
Training Accuracy: 0.8673096628188802
Validation Accuracy: 0.8675115536215187
Test Accuracy: 0.8684010181241871
```

- **Map Reduce KNN**
  1. **Data Preparation**
     - Split data into training and testing datasets.
     - Categorize the score column into 3 classes according to the range of score 0 -150 class 0, 150-300 class 1, and > 300 class 2 as the max score value is 400
  2. **Similarity Computation**
     - For a given test point, compute the cosine similarity between the test point and each training point.
     - Use cosine similarity to measure how close two points are.
  3. **Partition Mapping**
     - In a distributed environment **(e.g., PySpark RDD**), we process data partitions separately, each partition finds its top-most similar neighbors.
  4. **Neighbor Aggregation (reduce)**
     - Combine neighbors from all partitions.
     - Sort combined neighbors based on similarity scores and select the top-k global neighbors.
  5. **Weighted Voting Prediction**
     - Each neighbor votes for its class label.
     - Votes are weighted by similarity scores and inverse class frequency to reduce bias toward frequent classes.
     - Predict the label with the highest total weighted vote.
  6. **Model Evaluation**
     - Apply KNN to part of the test samples to generate predictions.
     - Compute the **confusion matrix** comparing actual vs. predicted labels.
     - Calculate **evaluation metrics**



Confusion Matrix:

	0	1	2
0	28	2	0
1	81	353	0
2	0	0	36

Accuracy	0.834
Macro Precision	0.750416
Micro Precision	0.834
Macro Recall	0.915566
Micro Recall	0.834
Macro F1 Score	0.824805
Micro F1 Score	0.834

- **Association Rules**

  Step 1. Convert numerical columns to categorical

  Step 2. Prepare dataframe: Append column name at the beginning of each item in the column to help differentiate values after creating the itemsets

  Step 3. Concatenate columns to create itemsets

Example row output

[rating_average_4.9|has_delivery_1|has_variation_0|vendor_has_delivery_1|isFreeShipping_0|IsSaleable_1|score_category_High|price_category_Very Low|stock_category_Very Low|rating_count_category_Low|primaryPrice_category_Very Low|sales_count_week_category_Low|preparationDays_category_Very Low|weight_category_Very Low]

  Step 4. Create the Association Rules from the item sets

Association Rules sorted by Support:

antecedent	consequent	confidence	lift	support
[stock_category_Very Low]	[rating_count_category_Very Low]	0.9999593270464414	0.9999999989750774	0.9999341302249872
[rating_count_category_Very Low]	[stock_category_Very Low]	0.9999748004705681	0.9999999989750773	0.9999341302249872
[sales_count_week_category_Very Low]	[stock_category_Very Low]	0.9999748000137937	0.9999999985182915	0.9999160049848831
[stock_category_Very Low]	[sales_count_week_category_Very Low]	0.9999412020077967	0.9999999985182914	0.9999160049848831
[sales_count_week_category_Very Low]	[rating_count_category_Very Low]	0.999964631598307	1.0000053030844678	0.9999058371672637
[rating_count_category_Very Low]	[sales_count_week_category_Very Low]	0.999946506262083	1.0000053030844678	0.9999058371672637
[sales_count_week_category_Very Low, rating_count_category_Very Low]	[stock_category_Very Low]	0.9999747991224789	0.9999999976269541	0.9998806386627286
[sales_count_week_category_Very Low, stock_category_Very Low]	[rating_count_category_Very Low]	0.9999646307070013	1.0000053021931259	0.9998806386627286
[rating_count_category_Very Low, stock_category_Very Low]	[sales_count_week_category_Very Low]	0.999946504914032	1.0000053017363375	0.9998806386627286
[stock_category_Very Low]	[IsSaleable_1]	0.9995490680297191	0.9999999886320627	0.9995238808879953
[IsSaleable_1]	[stock_category_Very Low]	0.9999747901278141	0.9999999886320627	0.9995238808879953
[rating_count_category_Very Low]	[IsSaleable_1]	0.9995490610522705	0.9999999816514664	0.9995084081220527
[IsSaleable_1]	[rating_count_category_Very Low]	0.9999593103817349	0.9999999816514663	0.9995084081220527
[sales_count_week_category_Very Low]	[IsSaleable_1]	0.9995490528784131	0.9999999734739216	0.9994902828819485
[IsSaleable_1]	[sales_count_week_category_Very Low]	0.9999411769648995	0.9999999734739217	0.9994902828819485
[IsSaleable_1, stock_category_Very Low]	[rating_count_category_Very Low]	0.999959309355929	0.9999999806256188	0.9994832096175177
[rating_count_category_Very Low, stock_category_Very Low]	[IsSaleable_1]	0.9995490496885349	0.9999999702826043	0.9994832096175177
[IsSaleable_1, rating_count_category_Very Low]	[stock_category_Very Low]	0.9999747891019922	0.999999987606215	0.9994832096175177
[IsSaleable_1, sales_count_week_category_Very Low]	[stock_category_Very Low]	0.9999747886448056	0.9999999871490168	0.9994650843774134
[IsSaleable_1, stock_category_Very Low]	[sales_count_week_category_Very Low]	0.9999411754819408	0.9999999719008759	0.9994650843774134

only showing top 20 rows

Association Rules sorted by Confidence:

antecedent	consequent	confidence
[score_category_Medium, has_delivery_0, isFreeShipping_0, has_variation_0, preparationDays_category_Very Low, weight_category_Very Low, sales_count_week_category_Very Low]	[stock_category_Very Low]	1.0
[vendor_has_delivery_1, rating_average_0.0, isFreeShipping_0, stock_category_Very Low]	[sales_count_week_category_Very Low]	1.0
[rating_average_5.0, score_category_Medium, isFreeShipping_0, has_variation_0, preparationDays_category_Very Low, price_category_Very Low, weight_category_Very Low, primaryPrice_category_Very Low, sales_count_week_category_Very Low, stock_category_Very Low]	[rating_count_category_Very Low]	1.0
[isFreeShipping_1, vendor_has_delivery_1, rating_average_0.0, score_category_Low, has_variation_0, preparationDays_category_Very Low, price_category_Very Low, primaryPrice_category_Very Low, sales_count_week_category_Very Low, stock_category_Very Low]	[rating_count_category_Very Low]	1.0
[isFreeShipping_1, vendor_has_delivery_1, has_delivery_1, rating_average_0.0, score_category_Low, has_variation_0, price_category_Very Low, weight_category_Very Low, primaryPrice_category_Very Low, sales_count_week_category_Very Low]	[stock_category_Very Low]	1.0
[vendor_has_delivery_1, rating_average_0.0, isFreeShipping_0, stock_category_Very Low]	[rating_count_category_Very Low]	1.0
[isFreeShipping_1, rating_average_0.0, score_category_Low, weight_category_Very Low, primaryPrice_category_Very Low, sales_count_week_category_Very Low]	[rating_count_category_Very Low]	1.0
[score_category_Medium, vendor_has_delivery_0, isFreeShipping_0, preparationDays_category_Very Low, rating_count_category_Very Low, stock_category_Very Low]	[sales_count_week_category_Very Low]	1.0
[rating_average_5.0, score_category_Medium, vendor_has_delivery_1, has_delivery_1, has_variation_0, price_category_Very Low, weight_category_Very Low, stock_category_Very Low]	[rating_count_category_Very Low]	1.0
[isFreeShipping_1, vendor_has_delivery_1, has_delivery_1, preparationDays_category_Very Low, weight_category_Very Low, IsSaleable_1]	[stock_category_Very Low]	1.0
[vendor_has_delivery_0, rating_average_0.0, isFreeShipping_0, primaryPrice_category_Very Low, sales_count_week_category_Very Low]	[rating_count_category_Very Low]	1.0
[has_delivery_0, vendor_has_delivery_0, score_category_Low, isFreeShipping_0, preparationDays_category_Very Low, price_category_Very Low, weight_category_Very Low, stock_category_Very Low]	[rating_count_category_Very Low]	1.0
[vendor_has_delivery_1, has_delivery_1, rating_average_0.0, score_category_Low, preparationDays_category_Very Low, price_category_Very Low, weight_category_Very Low, IsSaleable_1, sales_count_week_category_Very Low, stock_category_Very Low]	[rating_count_category_Very Low]	1.0
[score_category_Medium, has_delivery_1, has_variation_0, preparationDays_category_Very Low, weight_category_Very Low, IsSaleable_1, rating_count_category_Very Low]	[sales_count_week_category_Very Low]	1.0
[has_delivery_0, rating_average_0.0, score_category_Low, isFreeShipping_0, has_variation_0, preparationDays_category_Very Low, price_category_Very Low, primaryPrice_category_Very Low]	[sales_count_week_category_Very Low]	1.0
[rating_average_5.0, score_category_Medium, vendor_has_delivery_1, has_variation_0, price_category_Very Low, weight_category_Very Low, primaryPrice_category_Very Low, rating_count_category_Very Low, stock_category_Very Low]	[sales_count_week_category_Very Low]	1.0
[has_delivery_0, vendor_has_delivery_0, rating_average_0.0, score_category_Low, isFreeShipping_0, has_variation_0, price_category_Very Low, IsSaleable_1, sales_count_week_category_Very Low]	[rating_count_category_Very Low]	1.0
[isFreeShipping_1, has_delivery_1, rating_average_0.0, score_category_Low, has_variation_0, preparationDays_category_Very Low, weight_category_Very Low, primaryPrice_category_Very Low, sales_count_week_category_Very Low, stock_category_Very Low]	[rating_count_category_Very Low]	1.0
[score_category_Medium, has_delivery_0, vendor_has_delivery_0, stock_category_Very Low]	[sales_count_week_category_Very Low]	1.0
[score_category_Low, has_variation_0, price_category_Very Low, weight_category_Very Low, primaryPrice_category_Very Low, sales_count_week_category_Very Low]	[rating_count_category_Very Low]	1.0

Association Rules sorted by Lift:

[Stage 284:==>   (8 + 3) / 11]

antecedent	consequent	confidence	lift	support
[isFreeShipping_1, has_delivery_0, has_variation_0]	[vendor_has_delivery_0]	0.9850907402157982	2.6402071025637732	0.07220520954766917
[isFreeShipping_1, has_delivery_0, has_variation_0, stock_category_Very Low]	[vendor_has_delivery_0]	0.9850907402157982	2.6402071025637732	0.07220520954766917
[isFreeShipping_1, has_delivery_0, has_variation_0, rating_count_category_Very Low, stock_category_Very Low]	[vendor_has_delivery_0]	0.985090110738498	2.640205415459863	0.07220211499448065
[isFreeShipping_1, has_delivery_0, has_variation_0, rating_count_category_Very Low]	[vendor_has_delivery_0]	0.985090110738498	2.640205415459863	0.07220211499448065
[isFreeShipping_1, has_delivery_0, has_variation_0, sales_count_week_category_Very Low]	[vendor_has_delivery_0]	0.9850900208088302	2.64020517443339	0.07220167291545371
[isFreeShipping_1, has_delivery_0, has_variation_0, sales_count_week_category_Very Low, stock_category_Very Low]	[vendor_has_delivery_0]	0.9850900208088302	2.64020517443339	0.07220167291545371
[isFreeShipping_1, has_delivery_0, has_variation_0, IsSaleable_1, stock_category_Very Low]	[vendor_has_delivery_0]	0.9850893912707795	2.640203487166659	0.07219857836226519
[isFreeShipping_1, has_delivery_0, has_variation_0, sales_count_week_category_Very Low, rating_count_category_Very Low]	[vendor_has_delivery_0]	0.9850893912707795	2.640203487166659	0.07219857836226519
[isFreeShipping_1, has_delivery_0, has_variation_0, sales_count_week_category_Very Low, rating_count_category_Very Low, stock_category_Very Low]	[vendor_has_delivery_0]	0.9850893912707795	2.640203487166659	0.07219857836226519
[isFreeShipping_1, has_delivery_0, has_variation_0, IsSaleable_1]	[vendor_has_delivery_0]	0.9850893912707795	2.640203487166659	0.07219857836226519
[isFreeShipping_1, has_delivery_0, has_variation_0, IsSaleable_1, rating_count_category_Very Low]	[vendor_has_delivery_0]	0.9850887616795652	2.6402017997574405	0.07219548380907667
[isFreeShipping_1, has_delivery_0, has_variation_0, IsSaleable_1, rating_count_category_Very Low, stock_category_Very Low]	[vendor_has_delivery_0]	0.9850887616795652	2.6402017997574405	0.07219548380907667
[isFreeShipping_1, has_delivery_0, has_variation_0, IsSaleable_1, sales_count_week_category_Very Low]	[vendor_has_delivery_0]	0.9850886717336229	2.640201558687348	0.07219504173004974
[isFreeShipping_1, has_delivery_0, has_variation_0, IsSaleable_1, sales_count_week_category_Very Low, stock_category_Very Low]	[vendor_has_delivery_0]	0.9850886717336229	2.640201558687348	0.07219504173004974
[isFreeShipping_1, has_delivery_0, has_variation_0, IsSaleable_1, sales_count_week_category_Very Low, rating_count_category_Very Low]	[vendor_has_delivery_0]	0.9850880420816418	2.640199871115265	0.07219194717686123
[isFreeShipping_1, has_delivery_0, has_variation_0, IsSaleable_1, sales_count_week_category_Very Low, rating_count_category_Very Low, stock_category_Very Low]	[vendor_has_delivery_0]	0.9850880420816418	2.640199871115265	0.07219194717686123
[isFreeShipping_1, has_delivery_0, has_variation_0, weight_category_Very Low, stock_category_Very Low]	[vendor_has_delivery_0]	0.985078682624606	2.6401747862129983	0.07211679374228296
[isFreeShipping_1, has_delivery_0, has_variation_0, weight_category_Very Low]	[vendor_has_delivery_0]	0.985078682624606	2.6401747862129983	0.07211679374228296
[isFreeShipping_1, has_delivery_0, has_variation_0, weight_category_Very Low, rating_count_category_Very Low, stock_category_Very Low]	[vendor_has_delivery_0]	0.9850780518735469	2.640173095695203	0.07211369918909444
[isFreeShipping_1, has_delivery_0, has_variation_0, weight_category_Very Low, rating_count_category_Very Low]	[vendor_has_delivery_0]	0.9850780518735469	2.640173095695203	0.07211369918909444

only showing top 20 rows

Interesting Rules:

Products with no ratings tend to have low engagement metrics

- Items with rating_average_0.0 almost always have very low sales counts, stock levels, and rating counts
- This suggests new or unpopular products need specific attention to gain attraction

Extremely low rating counts and low sales are consistently linked.

Products that have low rating counts tend to have lower sales as the products don't have feedback so people don't trust them as much

Low-performing products share a profile of

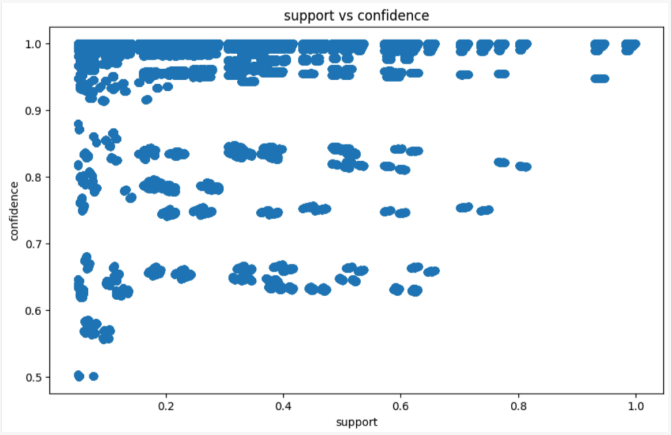
- Low Stock
- Low Weekly Sales

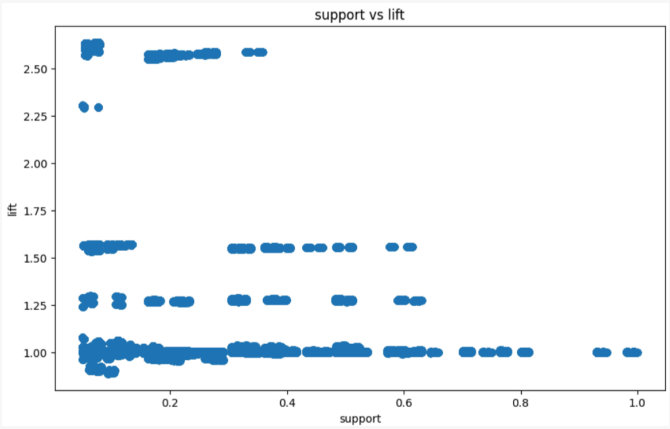
Well-stocked products share a profile of

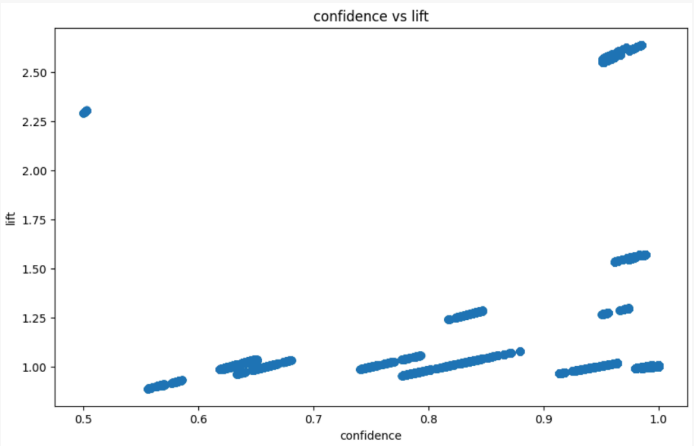
- High-priced
- Without free shipping
- Without variation

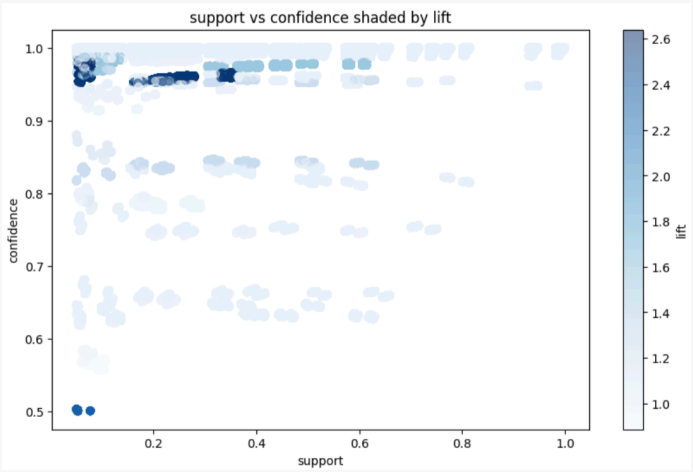
Possibly because they're premium, standardized items.

Relationships Between Support, Confidence, and Lift









V. Unsuccessful trials that were not included in the final solution

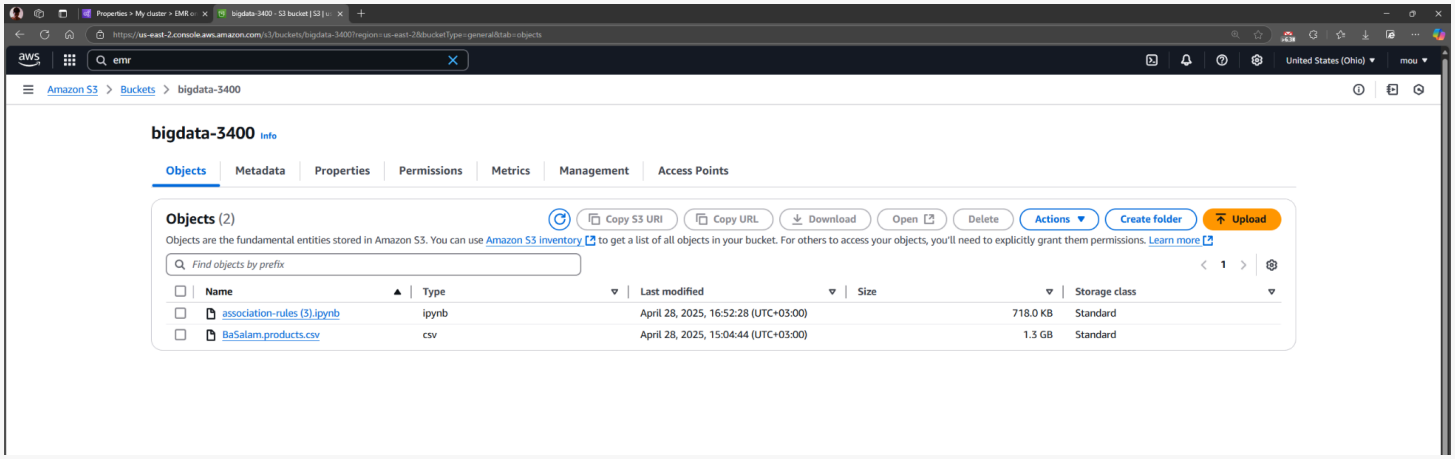
- We experimented with reintroducing some previously dropped columns during model training to evaluate their impact on accuracy and validate our feature selection process. However, these columns provided no benefit reducing the model's accuracy.
- We also attempted to conduct all Exploratory Data Analysis (EDA) using PySpark DataFrames. However, the results were less effective compared to pandas visualization. Thus, we switched to using pandas for visualization of some parts.

VI. Any Enhancements and future work.

- Rather than dropping columns with a high number of unique values(constant columns), explore grouping them using clustering techniques or consolidate rare values under an "Other" category to preserve useful information.

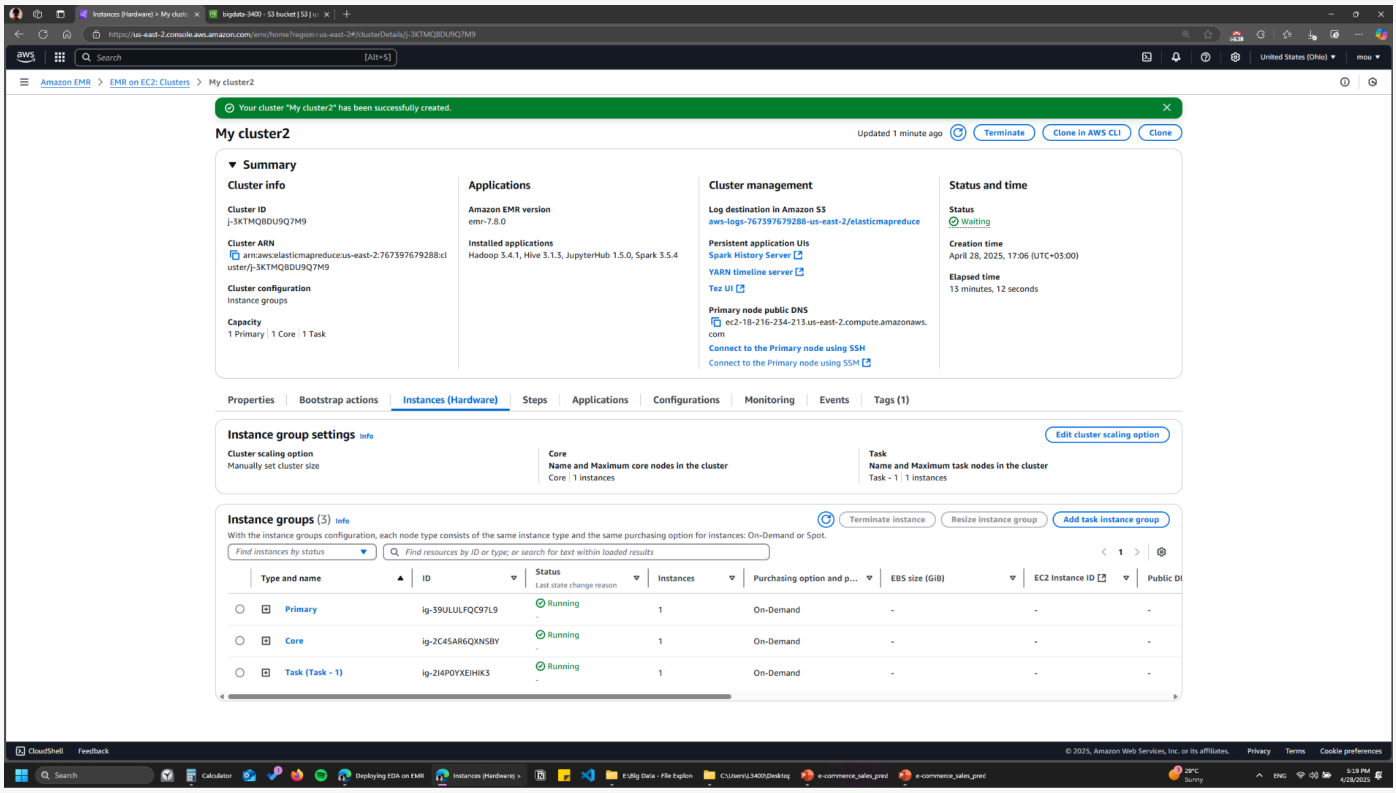
Running on AWS with Fully distributed mode

Step 1. Uploaded the dataset and notebook files to Amazon S3



Step 2. Created an Amazon EMR Cluster with the following configurations:

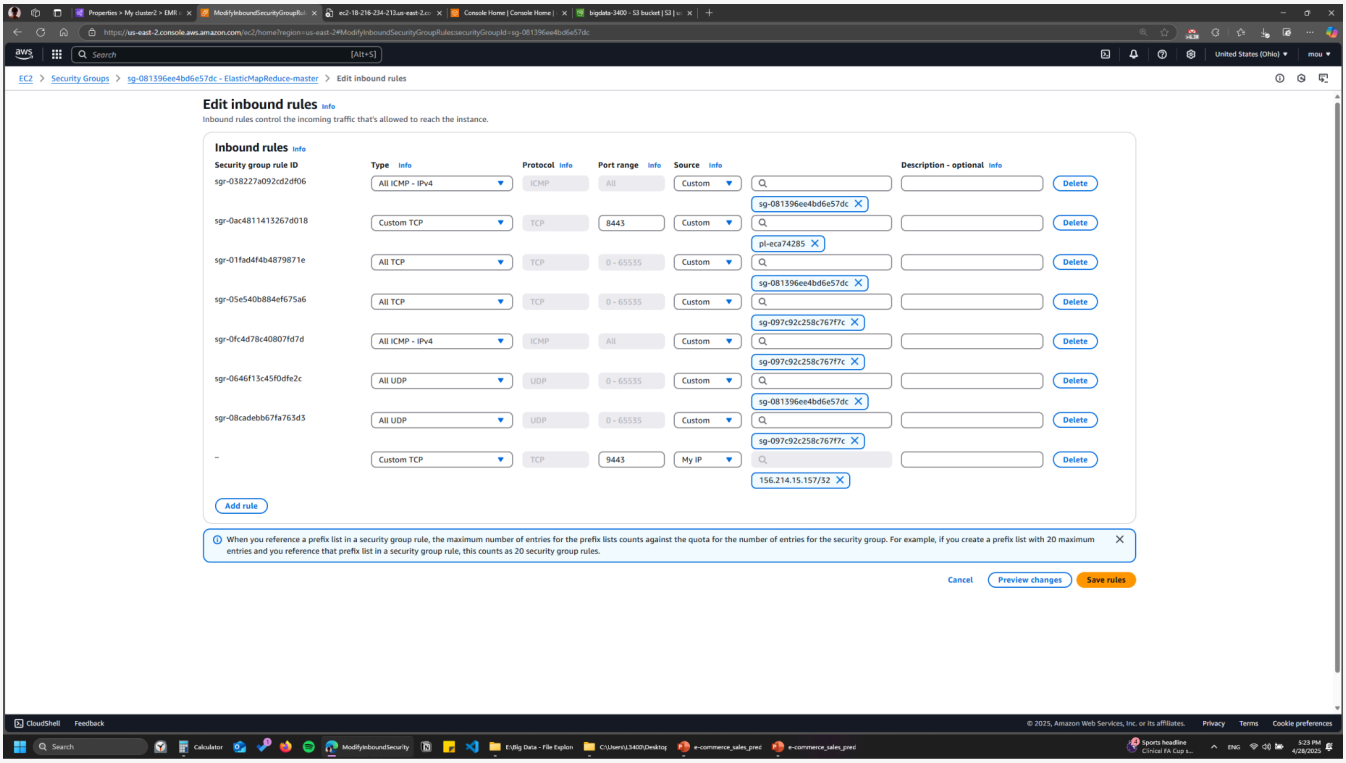
- Application Bundle
  - Hadoop 3.4.1
  - JupyterHub 1.5 (for jupyter notebooks)
  - Spark 3.5.4
- Instances:
  - **1 Primary node** -> m4.large
  - **1 Core node** -> m4.large
  - **1 Task node** -> m4.large



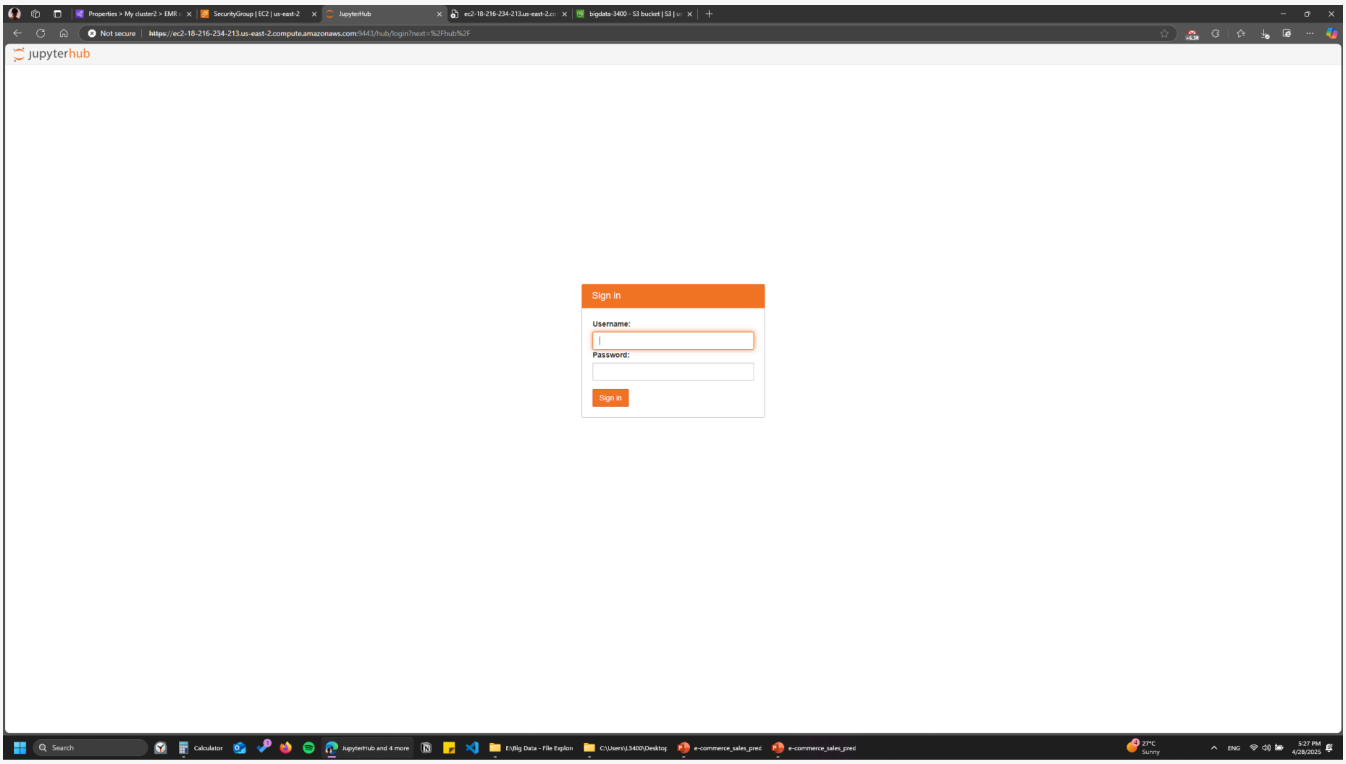
Cluster running successfully

Step 3. Connecting to the EMR Cluster

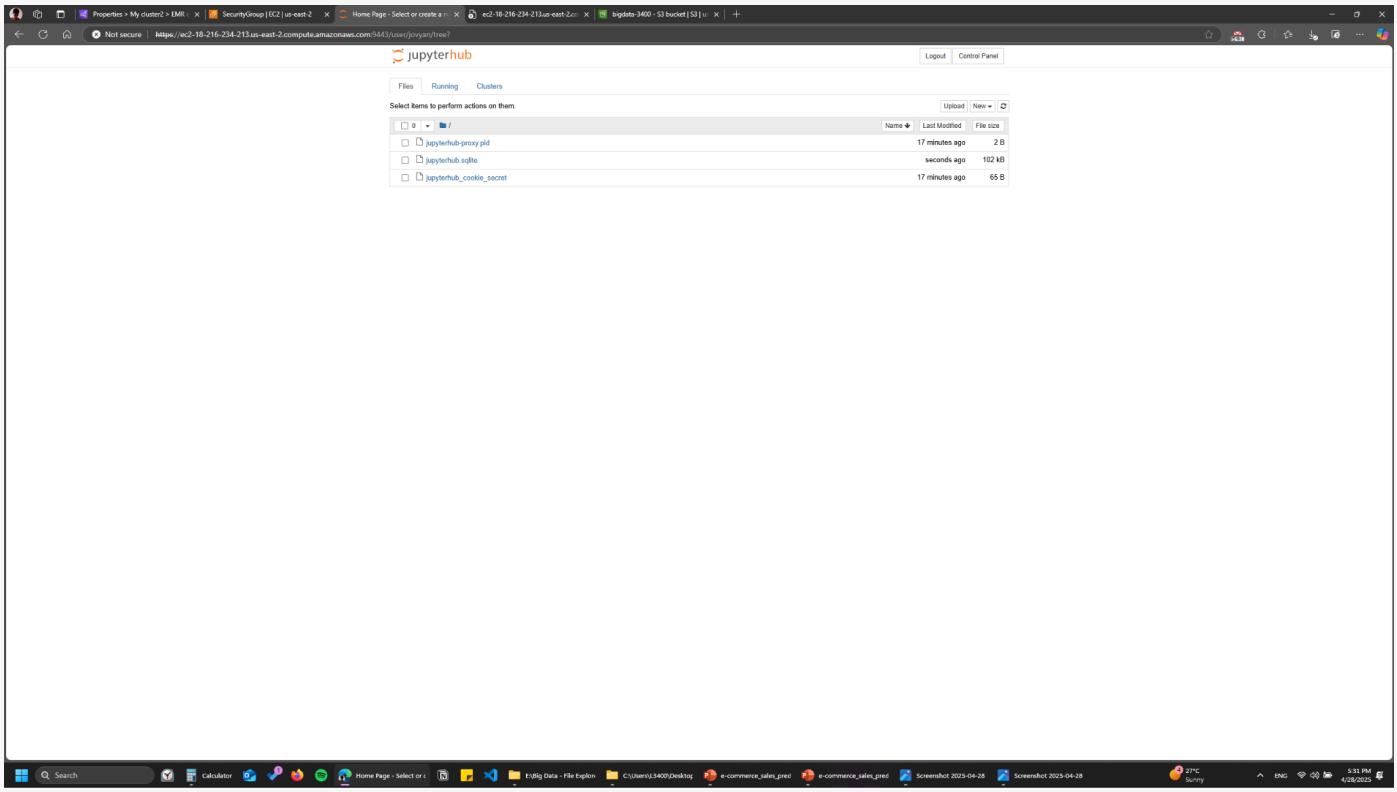
1. Connect to the EMR cluster master node from my computer:
   Add port 9443 to the security group to allow traffic from my IP to the cluster DNS



2. Then navigating to the url https://<master-node-public-dns>:9443 will open the jupyter hub in the master node
   cluster dns: ec2-18-216-234-213.us-east-2.compute.amazonaws.com

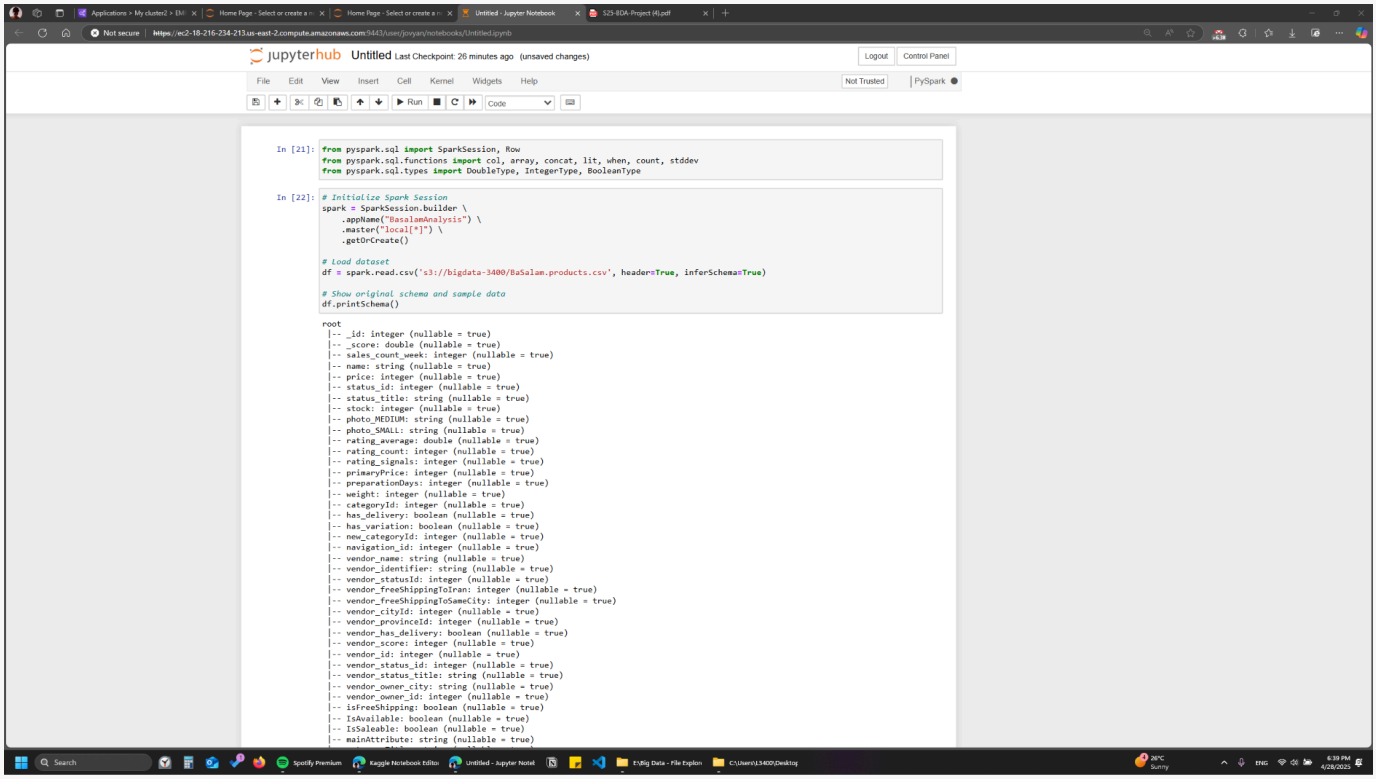


Entered the default username **jovyan** and password **jupyter** for the jupyer hub to open



Jupyterhub opened successfully

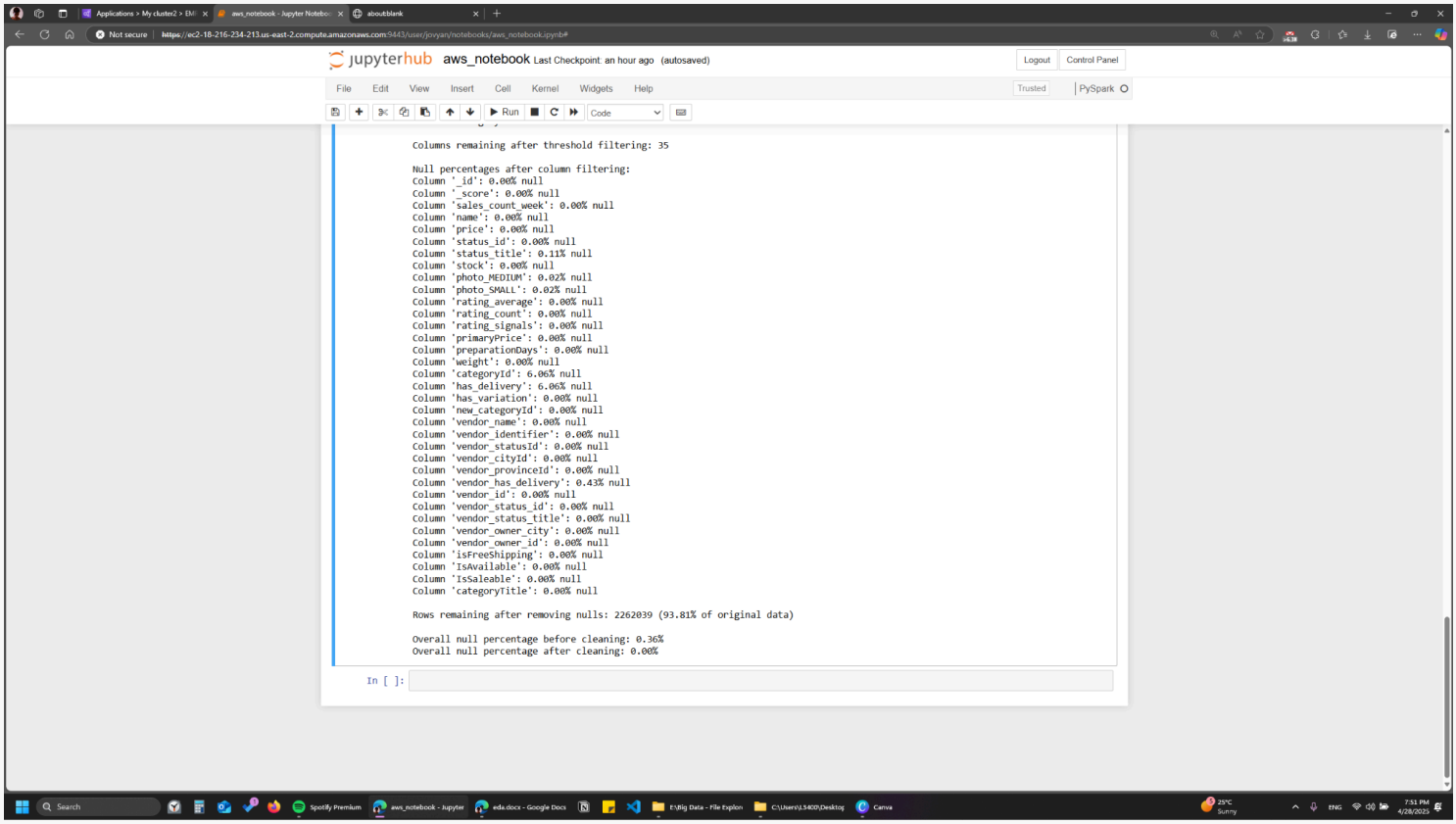
Step 4. Created Jupyter Notebook with Pyspark as a kernel **for a fully distributed mode**
Pulling the dataset from S3 (s3://bigdata-3400/BaSalam.products.csv)



Successfully pulled the .csv file

Now the notebook is ready for use

Testing: Running the preprocessing step on the notebook (removing the nulls):



Notebook opened successfully