TDS2101 INTRODUCTION TO DATA SCIENCE
SEMESTER 2,
YEAR 2020/2021

PROJECT REPORT

FACTORS CONTRIBUTE TO THE HIGH SALES OF A BEAUTY PRODUCTS

NAME	STUDENT ID
CHANG KAI BOON (Team Leader)	1181101282
ANG KELVIN	1181101297
PRITESH PATEL	1181101645

Table of Content

1.0 Problem Statement

1.1 Background and Problems

Amazon, the world's largest online retailer, has posted the biggest profit in its 26-year history as online sales in July 2020 despite the covid-19 pandemic [1] . Most people will just easily make an assumption that the increasing sales of Amazon is due to the convenience of online shopping. It is true that online shopping is convenient. However, there are still many factors and criteria that play a bigger role in influencing online sales and total revenue of a year.

To understand more about different factors and criteria that might or might not correlate with increasing online sales,the team decided to make use of the available big data. Therefore, the team proposed to use the data science process to gain business insights from the Amazon beauty datasets, and determine the **factors that contribute to the high sales of a beauty product.**

1.2 Questions

There are 6 types of questions in Data Science : Descriptive, Exploratory, Inferential, Causal, Predictive and Mechanistic. We can get useful business insight from the data by asking the right data science questions. This project focuses on the Amazon datasets, and aims to study the following questions using the data science process.

1. Does a high number of good ratings contribute to a lower sales rank and does a high number of bad ratings contribute to a higher sales rank ? (Causal)
2. Which item has the lowest rating and which item has the highest rating? (Descriptive)
3. What are the common words found in the description of highly rated items?(Descriptive)
4. What is the relationship between price and rating of an item? (Exploratory)
5. What is the average price for items that have a high sales rank and what is the average price for items that have low sales rank ? (Descriptive)
6. What is the peak time of the year for a customer to buy beauty products?(Predictive)
7. How does an increase or decrease of price affect the sales rank of a beauty product? (Mechanistic)
8. What are the most common words that contribute to negative reviews. Also, what are the most common words that contribute to positive reviews?(Exploratory)

1.3 Potential Benefits

By using the insights gathered from answering the questions stated above, the Amazon sellers and beauty products entrepreneurs can have a better understanding and guidelines on their products. In this project, the team has come out with the High Sales Factor Model. With this model, the sellers are able to identify which factors that contribute to the high sales of a product, and try to use this insight to boost their sales.

2.0 Datasets Used

The team has used 3 datasets in this project. There are 2 given datasets and 1 supplementary dataset.

2.1 Project datasets

1. **Amazon beauty products reviews data (371345 rows × 4 columns)**

Attribute Name	Explanation
asin	Unique id of product
user	Unique id of user
rating	Rating of the product
timestamp	Time when user give the rating

2. **Amazon beauty products metadata (32892 rows × 29 columns)**

Attribute Name	Explanation
tech 1	Not relevant
description	Description of item
fit	Not relevant
title	Title of item
also_buy	Product bought along with item
image	Image link of item
Tech 2	Not relevant
brand	Brand of item
feature	Not relevant
rank	Sales rank of item
also_view	Products also viewed after viewing item
main_cat	Main category of item

similar_item	Products similar to item
date	Not relevant
price	Price of item
details.Shipping Weight:	Not relevant
details.\n Item Weight: \n	Not relevant
details.Item model number:	Not relevant
details.UPC:	Not relevant
details.\n Product Dimensions: \n	Not relevant
details.Discontinued by manufacturer:	Not relevant
details.Domestic Shipping:	Not relevant
details.International Shipping:	Not relevant
details.Batteries	Not relevant
details.Shipping Advisory:	Not relevant
details.ASIN:	Not relevant

2.2 Supplementary dataset

3. **Amazon reviews data (371345 rows × 25 columns)**
    - Source : https://nijianmo.github.io/amazon/index.html

Attribute	Explanation
overall	Overall rating
verified	Is user verified
reviewTime	When was the review
reviewerID	Unique id of review
asin	Unique id of the item
reviewerName	Name of reviewer
reviewText	Review text

summary	Summary of review
unixReviewTime	When was the review
vote	How many people find this review helpful
style.Format	Not relevant
style.Size	Not relevant
image	Image of reviewed item
style.SizeName	Not relevant
style.Flavor:	Not relevant
style.Style Name:	Not relevant
style.Scent Name:	Not relevant
style.Color:	Not relevant
style.Color Name:	Not relevant
style.Package Quantity:	Not relevant
style.Package Type:	Not relevant
style.Style:	Not relevant
style.Design:	Not relevant
style.Item Package Quantity:	Not relevant
style.Pattern:	Not relevant

3.0 Data Cleaning

Data cleaning is a process of transforming data from “raw” form into an appropriate format that can be conveniently consumed or analysed to generate actionable insights. This process includes filling missing data, smoothing noisy data, identifying and removing outliers and resolving inconsistencies.

We began the cleaning process with the Amazon beauty product metadata’s dataset. We dropped all the irrelevant attributes as they didn’t bring any useful information. Then, we checked for the duplicated product and removed them to ensure that each record is unique. Diagram 1 shows the result of this dataset. Then, we proceed to the next dataset which is the beauty product review dataset.

	description	title	brand	rank	price	asin
0	["Loud 'N Clear Personal Sound Amplifier allow...	Loud 'N Clear™ Personal Sound Amplifier	idea village	2,938,573 in Beauty & Personal Care (		6546546450
1	['No7 Lift & Luminate Triple Action Serum 50ml...	No7 Lift & Luminate Triple Action Serum 50...		872,854 in Beauty & Personal Care (	$44.99	7178680776
2	["No7 Stay Perfect Foundation now stays perfec...	No7 Stay Perfect Foundation Cool Vanilla by No7	No7	956,696 in Beauty & Personal Care (	$28.76	7250468162
3	[]	Wella Koleston Perfect Hair Colour 44/44 Mediu...		1,870,258 in Beauty & Personal Care (		7367905066
4	['Lacto Calamine Skin Balance Daily Nourishing...	Lacto Calamine Skin Balance Oil control 120 ml...	Pirmal Healthcare	67,701 in Beauty & Personal Care (	$12.15	7414204790

Diagram 1

As this dataset consists of the records of the product reviews, each product might appear in the records multiple times. This does not count as duplicated data because each product might receive different reviews from different buyers. One way to process the data is to group the product using their unique ID (asin), sum up all the ratings they received and divide them by the total number of buyers as average rating. Diagram 2 shows the result of this grouping process.

	asin	total_rating	total_user	average_rating
0	0061073717	10.0	2	5.000000
1	0143026860	70.0	17	4.117647
2	014789302X	87.0	20	4.350000
3	0571348351	15.0	3	5.000000
4	0692508988	5.0	1	5.000000

Diagram 2

Then, we merged the product reviews grouping data with the product metadata according to their unique ID (asin) so that we can finally know each product's average rating and the number of buyers who give the review. Next, we do binning on the average rating attribute to group the data into 2 bins. We defined the product as a high rating and highly rated product if its average rating is greater than the mean rating of all records in the original unprocess dataset. Diagram 3 shows the merge data.

	description	title	brand	rank	price	asin	total_rating	total_user	average_rating	average_rating_class	average_rating_class_numeric
0	["Loud 'N Clear Personal Sound Amplifier allow...	Loud 'N Clear™ Personal Sound Amplifier	idea village	2,938,573 in Beauty & Personal Care (	NaN	6546546450	5.0	2.0	2.5	negative	0
1	['No7 Lift & Luminate Triple Action Serum 50ml...	No7 Lift & Luminate Triple Action Serum 50...	NaN	872,854 in Beauty & Personal Care (	$44.99	7178680776	3.0	1.0	3.0	negative	0
2	["No7 Stay Perfect Foundation now stays perfec...	No7 Stay Perfect Foundation Cool Vanilla by No7	No7	956,696 in Beauty & Personal Care (	$28.76	7250468162	5.0	1.0	5.0	positive	1
3	[]	Wella Koleston Perfect Hair Colour 44/44 Mediu...	NaN	1,870,258 in Beauty & Personal Care (	NaN	7367905066	5.0	1.0	5.0	positive	1
4	['Lacto Calamine Skin Balance Daily Nourishing...	Lacto Calamine Skin Balance Oil control 120 ml...	Pirmal Healthcare	67,701 in Beauty & Personal Care (	$12.15	7414204790	66.0	15.0	4.4	positive	1

Diagram 3

Lastly, we extracted the rank and price from the original attributes and converted its data type from string to float. Before we move to the next section, we checked again the result to make sure the data is cleaned enough. Diagram 4 shows the null checking result.

```
description                    0
title                          0
brand                      15570
asin                           0
total_rating                   0
total_user                     0
average_rating                 0
average_rating_class           0
average_rating_class_numeric   0
vote_count                     0
average_vote                   0
high_rated                     0
rank_cleaned                 338
price_cleaned              21139
dtype: int64
```

Diagram 4

From the result above, we can see that the 3 attributes (brand, rank_cleaned,price_cleaned) still contain quite a number of null data. However, not all attributes will be used in EDA,we might use the attribute that is needed only. Thus, we decide to keep them at the moment and clean them only when we need them to be cleaned enough. This is because if we drop them immediately, the size of the dataset will reduce dramatically and might show misleading outcomes in some situations.

Let's move to the next section.

4.0 EDA and Descriptive Statistics

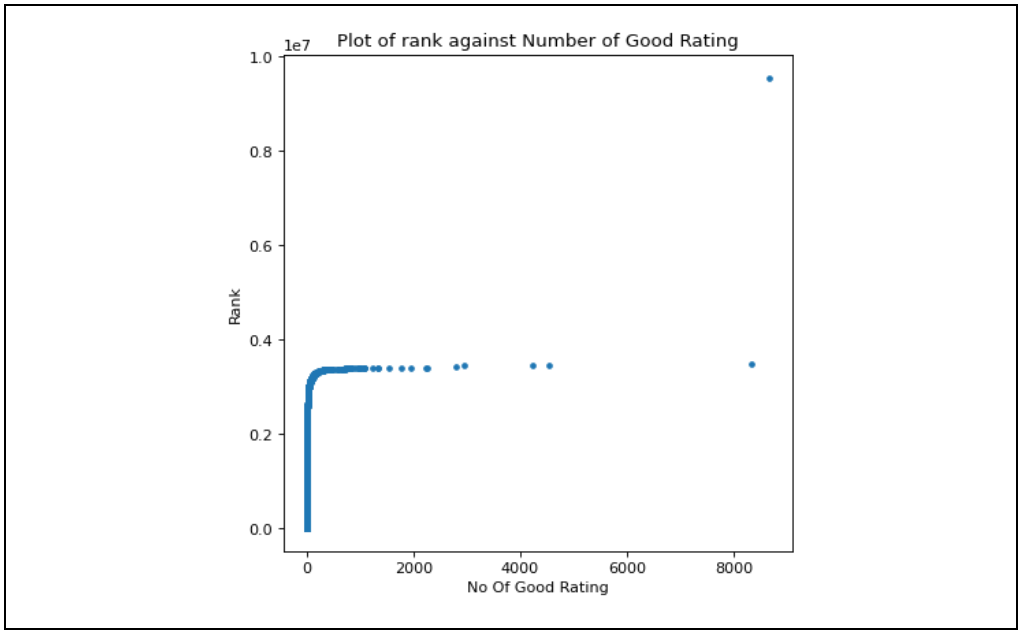
Data Exploratory Analysis (EDA) is a process of knowing the structure, attributes, granularity, scope and faithfulness of our dataset to gain the maximum insight. In this project, we will do the EDA by answering our proposed questions.

4.1 Does a high number of good ratings contribute to a lower sales rank and does a high number of bad ratings contribute to a higher sales rank?

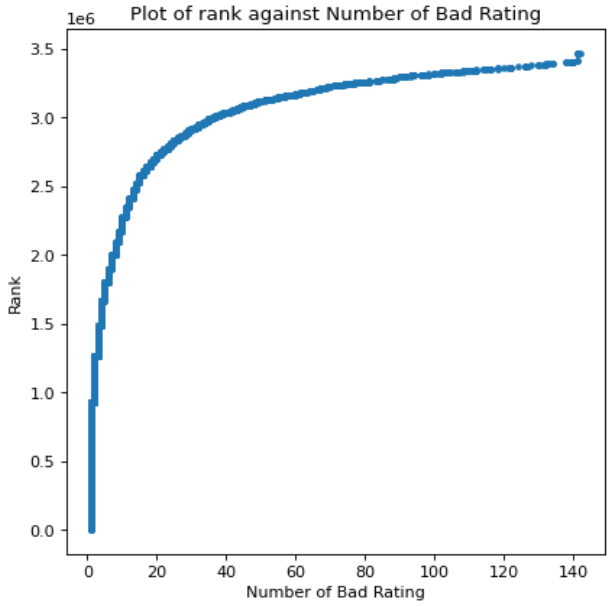
- **Why ?**
  - The purpose of this question is to explore the relationship between sales rank and ratings. By exploring the relationship between sales rank and ratings, the seller will be able to know whether rating is important to increase sales. If rating is indeed important to increase sales, then the seller will need to encourage buyers to rate their item to increase potential sales in the future.

- **How ?**
  - Create a dataframe which consists of cleaned data of rating and rank. Then ,create variables which differentiate good rating class and bad rating class. The team decided to create a function to determine whether a rating is considered bad or good. If the rating is larger than the average rating of the dataset , then the rating is considered as good, else the rating is considered as bad. Plot rank against number of good ratings and plot rank against number of bad ratings respectively.

- **Visualization**
  -



From the plot above, it seems like with good ratings, most of the product rank does not exceed 4 million. But it doesn't seem like a higher number of good ratings causes lower ranking. There is a potential outlier near the top right corner of the plot. The plot is growing logarithmically.

- 

From the plot above, it can be seen that the range of ranking for products with mostly bad ratings is very huge compared with good ratings. The range goes from 0 to 3.5 million. It seems like the higher the number of bad ratings, the higher the rank of the product.

- **Conclusion**
  - The team could not draw a conclusion towards the causality of rating and sales rank because there might be a potential simpson's paradox. Due to the incompleteness of the dataset, the team could not identify the confounding factors causing the simpson's paradox.

4.2 Which item has the lowest rating and which item has the highest rating ?

- **Why?**
  - The purpose of this question is to compare the two items together. The comparison helps to determine how the attributes of the lowest rated item differs from the highest rated item.

- **How ?**
  - Create a dataframe which consists of the cleaned data. Define a function which defines bad and good ratings. The team decided to categorize every rating below the average rating as bad and everything above the average rating as good. With that, create a variable which filters out both good and bad rating. By making use of the total_user_voted attribute and the “max()” function, the team can find out the lowest rated and highest items.

- **Visualization**
  - Highest items

	asin	total_rating_x	total_user_x	average_rating_x	rating_class	rank_cleaned	price_cleaned
1157	B000KNELAW	605.0	121	5.0	Good	2817656.0	NaN

  - Lowest items

	asin	total_rating_x	total_user_x	average_rating_x	rating_class	rank_cleaned	price_cleaned
29456	B01DKQO7YK	9.0	9	1.0	Bad	836937.0	NaN

4.3 What are the common words found in the description of highly rated items?

- **Why?**
  - The purpose of this question is to study if there is any similarity in the description of each highly rated item. If there is a similarity, what word appears the most in the description.

- **How?**
  - Create a dataframe that contains the product's descriptions and its rating class. Next, create a text_clean function to clean the description. Then, create a remove_stopwords function to remove the English Stopwords in the cleaned description. The reason we need to remove them is because they do not add much meaning to a sentence. Lastly, we use the python counter function to count the word and show the top 10 words found in the description of highly rated items.

- **Visualization /Output**

```
High rated : 8996 beauty products
[('skin', 5232),
 ('hair', 3039),
 ('oil', 2256),
 ('use', 1988),
 ('natural', 1709),
 ('body', 1599),
 ('color', 1182),
 ('product', 1142),
 ('made', 1052),
 ('dry', 1001),
```

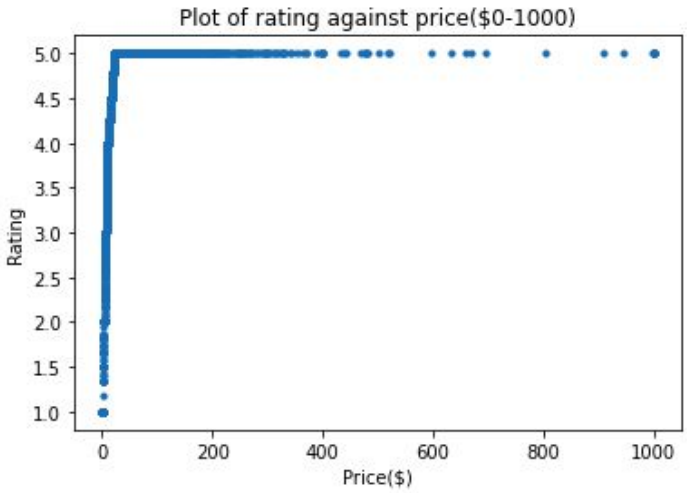
The python counter function shows the most common top 10 words and their appearance in the description. The word 'skin' appears the most in the description, which is 5232 times.

4.4 What is the relationship between price and rating of an item ?

- **Why?**
  - The purpose of this question is to explore the relationship between price and ratings. By exploring the relationship between price and ratings, the sellers or business entrepreneurs will be able to know whether the price is important to get a good rating. If price is indeed important to get a good rating, then the sellers need to think twice before setting the price of the product.

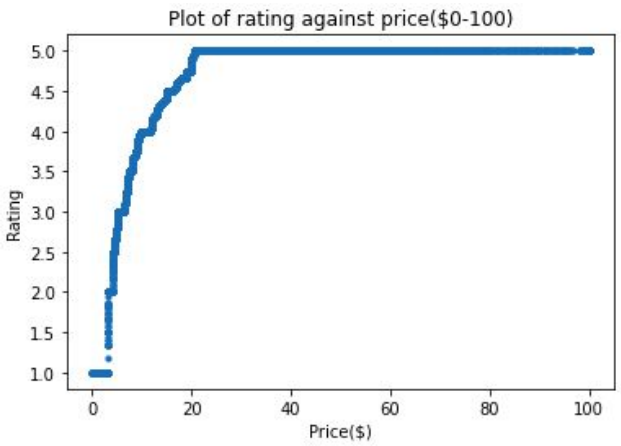
- **How?**
  - Since we are going to use the price attribute, we need to drop all the null data (remember the team decided not to drop them in the data cleaning part). Then plot the ratings against prices on a different scale.

- **Visualization**
  -



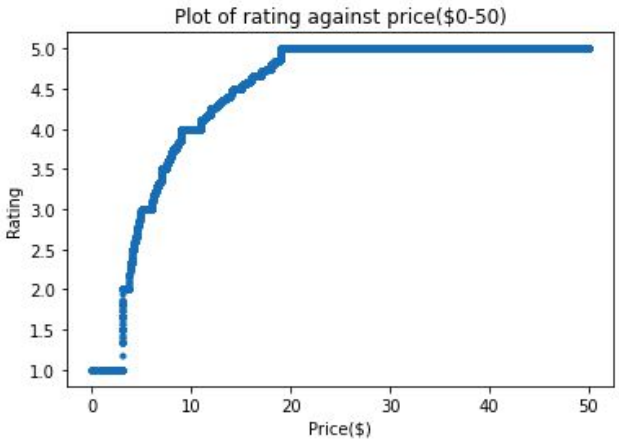
From the plot above, we can see that the plot is growing logarithmically.There are many potential outliers near the top right of the plot. We can observe that most of the data are located at the price range between $0 to $200. This graph is the steepest among 3 graphs that we plotted.

○



By lower down the price scale range to less than $100, we can see that the plot is still growing logarithmically. The slope decreases when the price increasing.

○

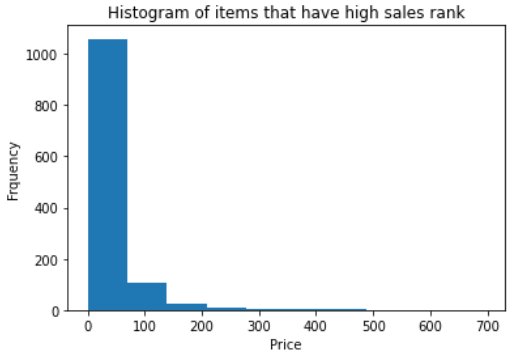


By lowering the price scale range to less than $50, we can see that the graph is still growing logarithmically, but now we can clearly see that it stops growing at price $20. Besides, this graph is the least steep among 3 graphs that we plotted.

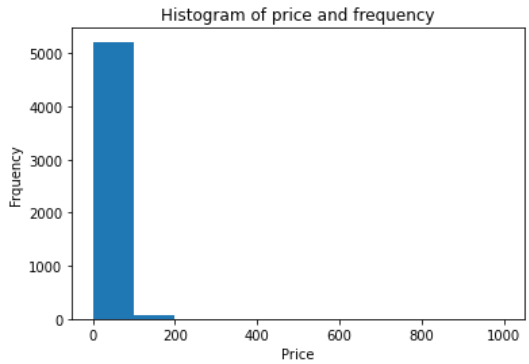
- **Conclusion**
  - Although the rating increases when the price increases until it reaches $20, the team still could not draw a conclusion towards the causality of rating and price because there might be a potential simpson's paradox. However, the team could not identify the confounding factors causing Simpson's paradox.

4.5 What is the average price for items that have a high sales rank and what is the average price for items that have low sales rank ?

- **Why**
  - The purpose of this question is to compare and see the average price for beauty products that are below and above 10000 to discover potential insights of the dataset. The team wanted to test their hypothesis on whether or not beauty products with lower sales rank will have cheaper prices.
- **How ?**
  - Create a dataframe which consists of the cleaned data of rank. Then make use of the function “describe()” to find out where the 25% and the 75% of the data are located. Create a variable which checks whether the rank of the particular row is below the 25% or above the 75%. By using the variable created, we can use the dataframe and the variable together with the “.mean()” pandas function to find the average price of the beauty products.
- **Visualization**
  -



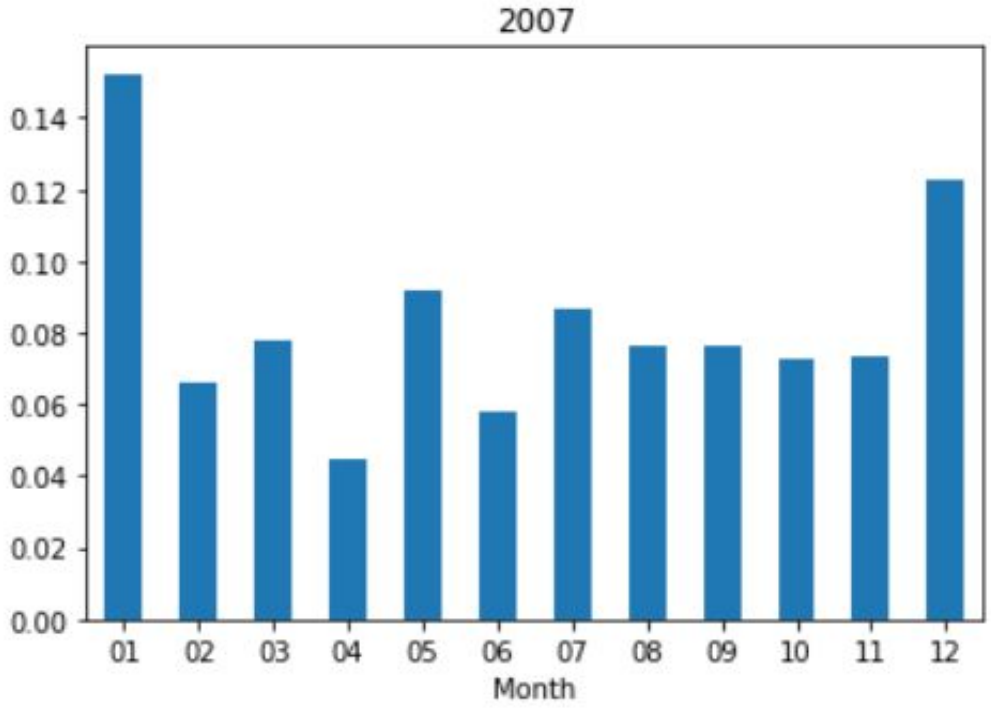
(Figure 1: Histogram of price and frequency of beauty product that are high in rank)



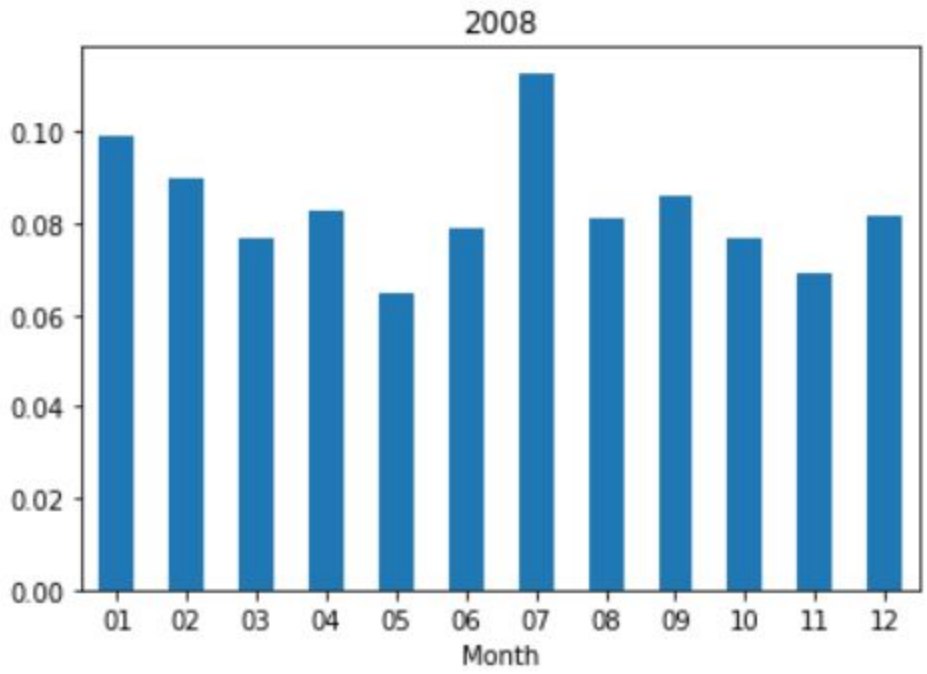
(Figure 2: Histogram of price and frequency of beauty product that are low in rank)

4.6 What is the peak time of the year for a customer to buy beauty products?

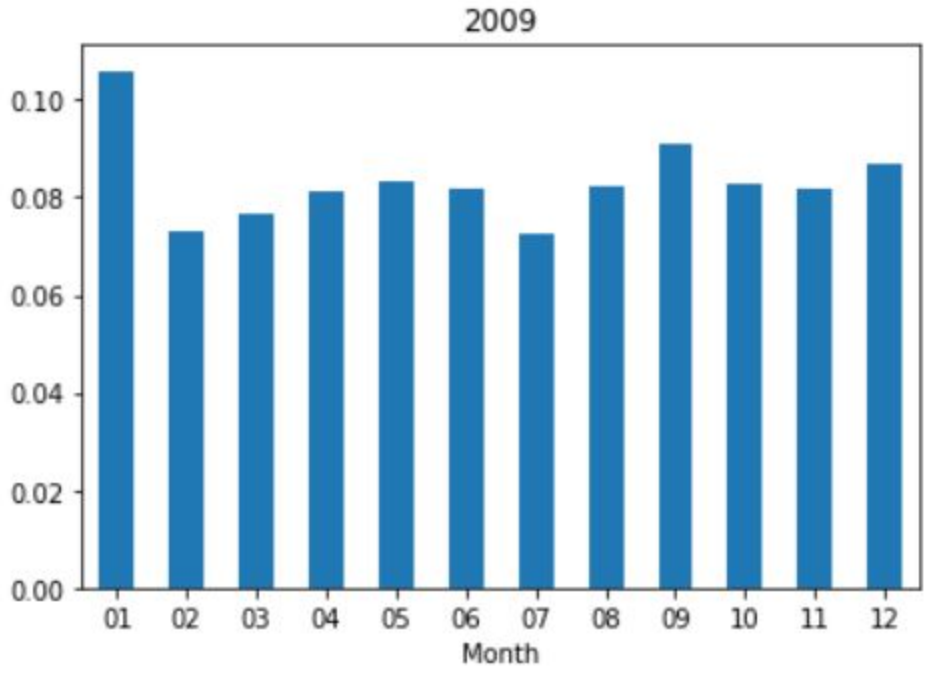
- **Why**
  - Why do we care to find a relationship between time and sales of beauty products? Well, we believe that with this insight, sellers can choose to spend more on advertising or more on stock. If a seller knows in a certain month the consumer is more likely to buy products, the seller can invest in more advertisements which hopefully will further increase sales.
- **How ?**
  - First we cleaned the date column values and expanded it so there is a column for year, month. Then create 2 data frames, one grouped by year and the other grouped by year and month. Only the data from 2007 to 2018 is chosen because the data prior to 2007 is too small. Then normalize the values. Finally plot a bar plot to see the trend of sales with time.
- **Visualization**



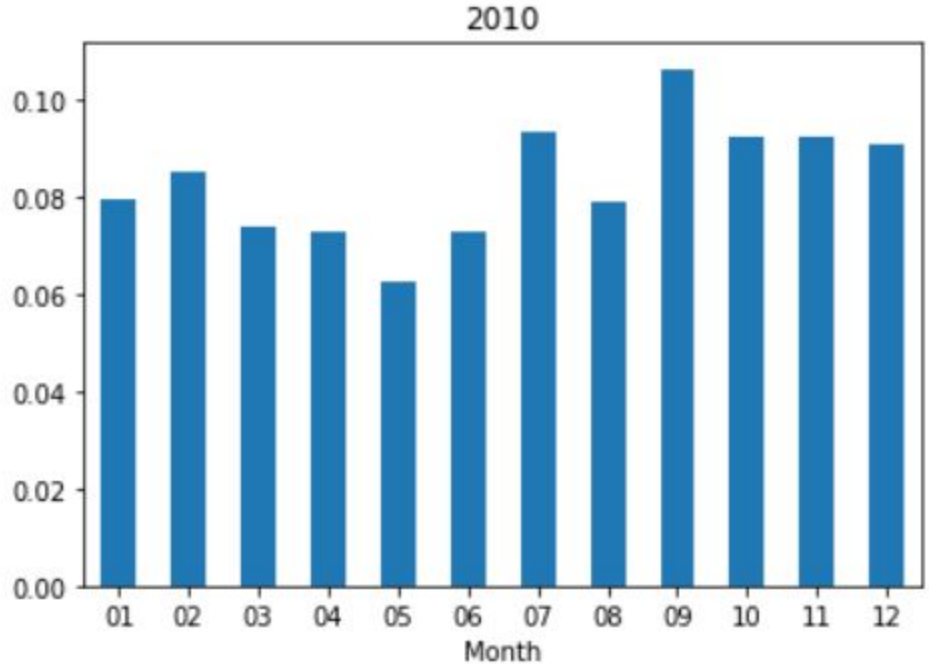
We can see in 2007 the sales peaked in the January and December



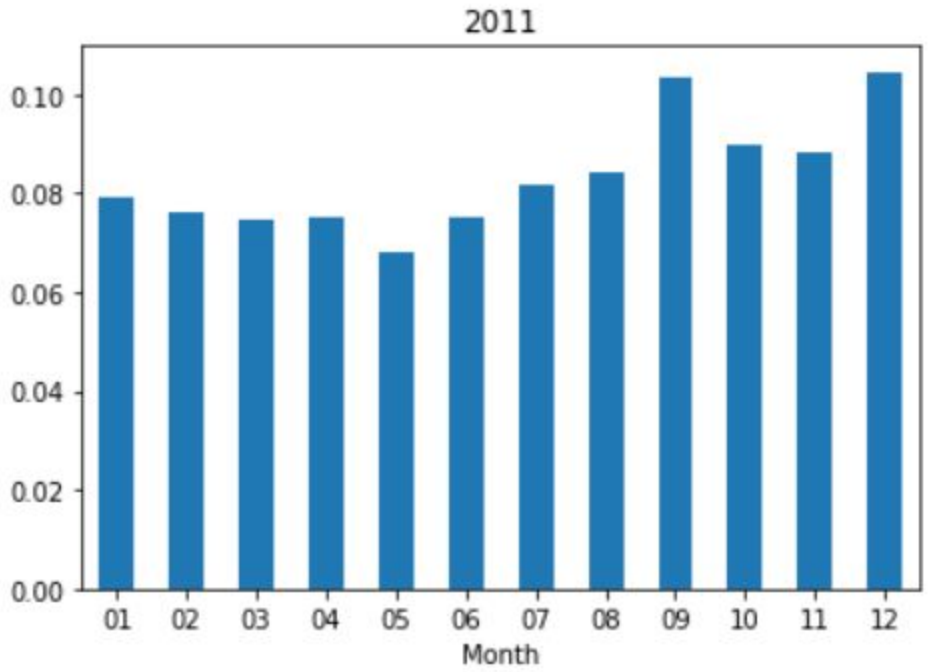
In 2008 we can see that the sales peaked at January and July



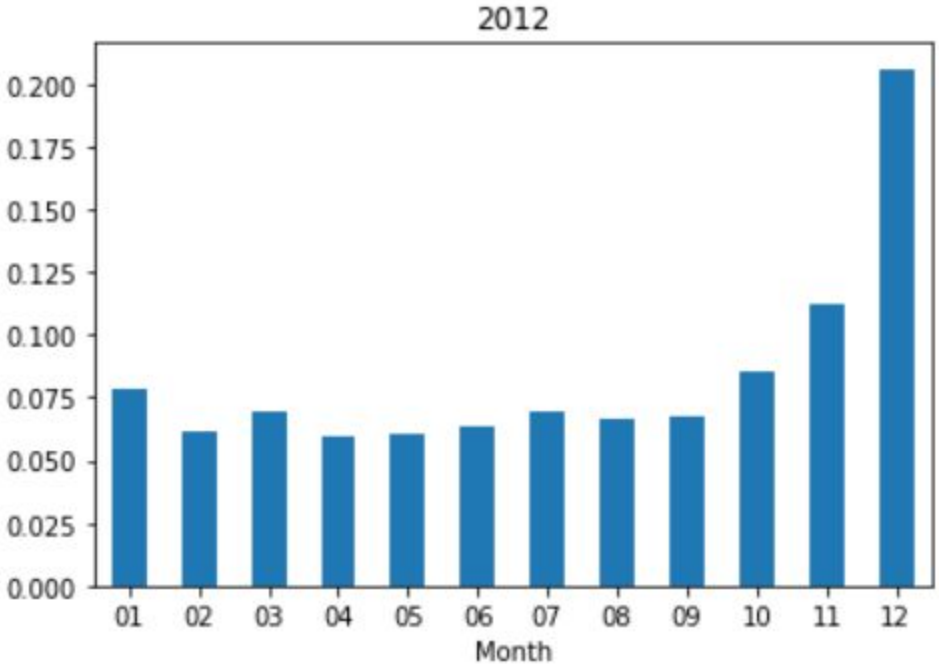
In 2009 we can see sales peaked at January



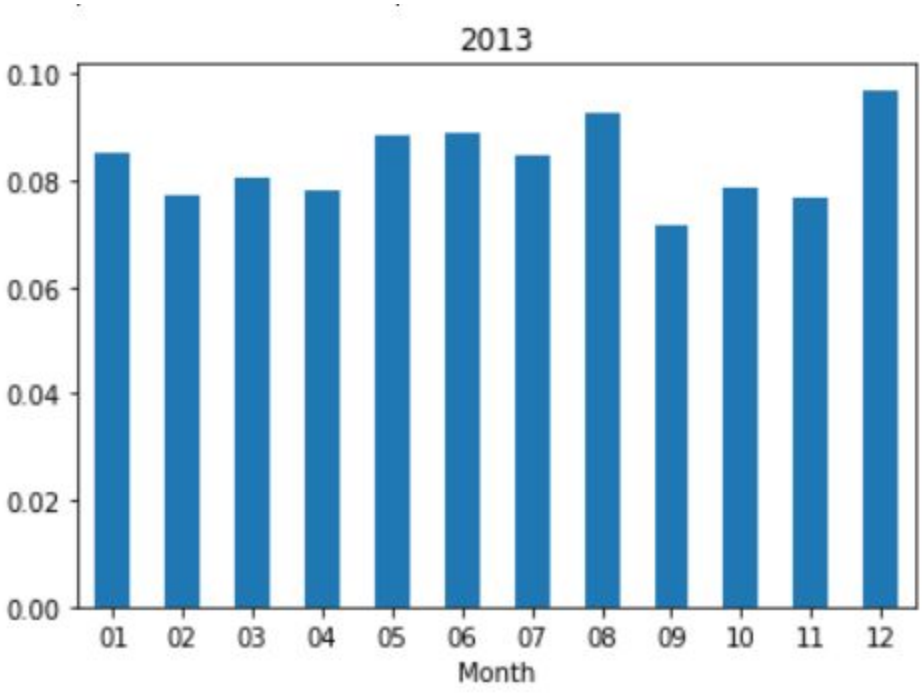
In 2010 sales are generally low in the beginning of the year and slightly higher in the later of the year



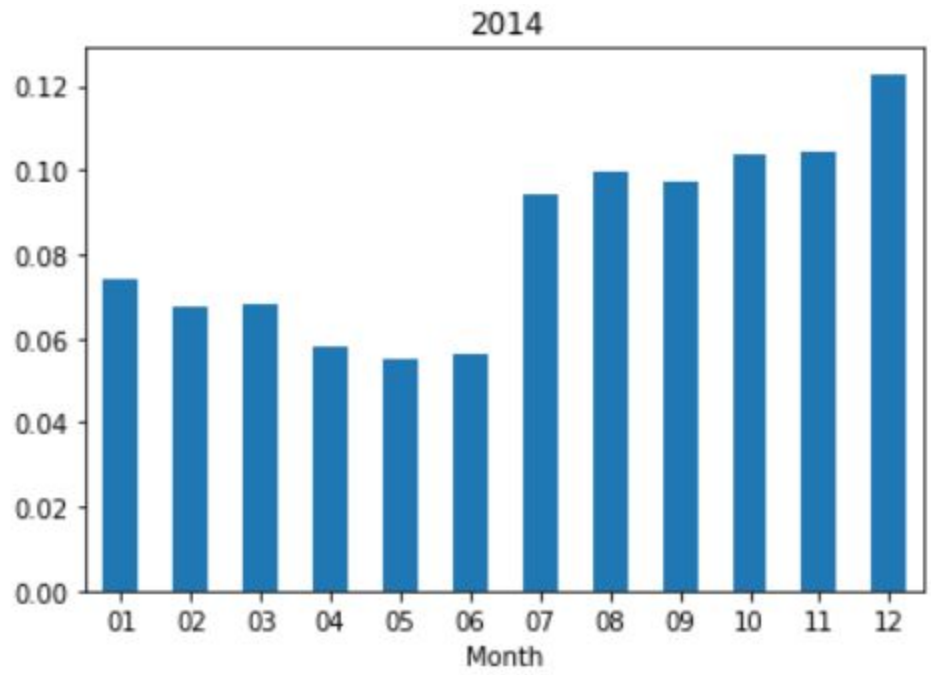
In 2011 sales are generally low in the beginning of the year and slightly higher in the later of the year



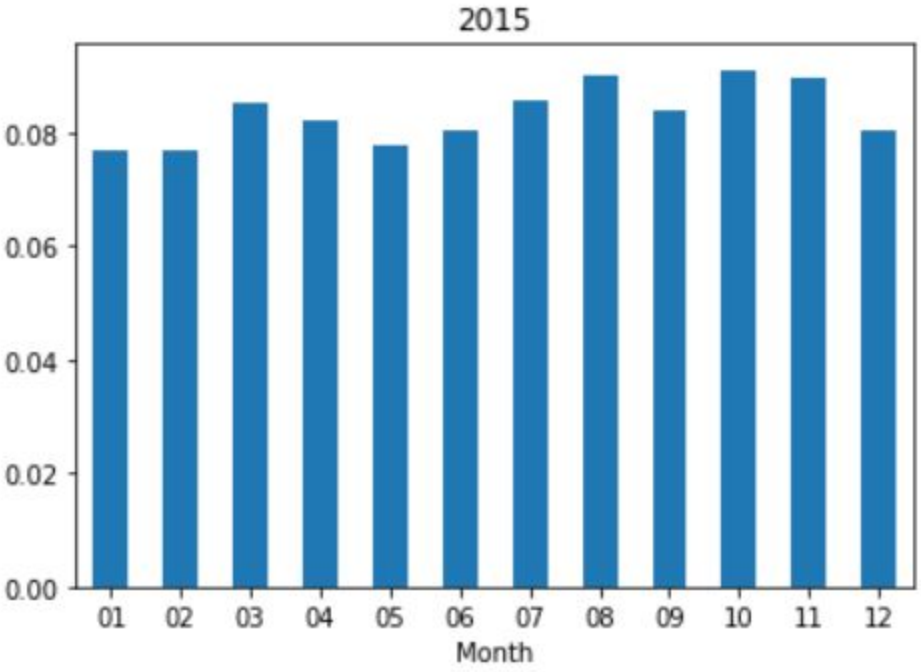
Surprisingly in 2012, sales are all time low except for December.



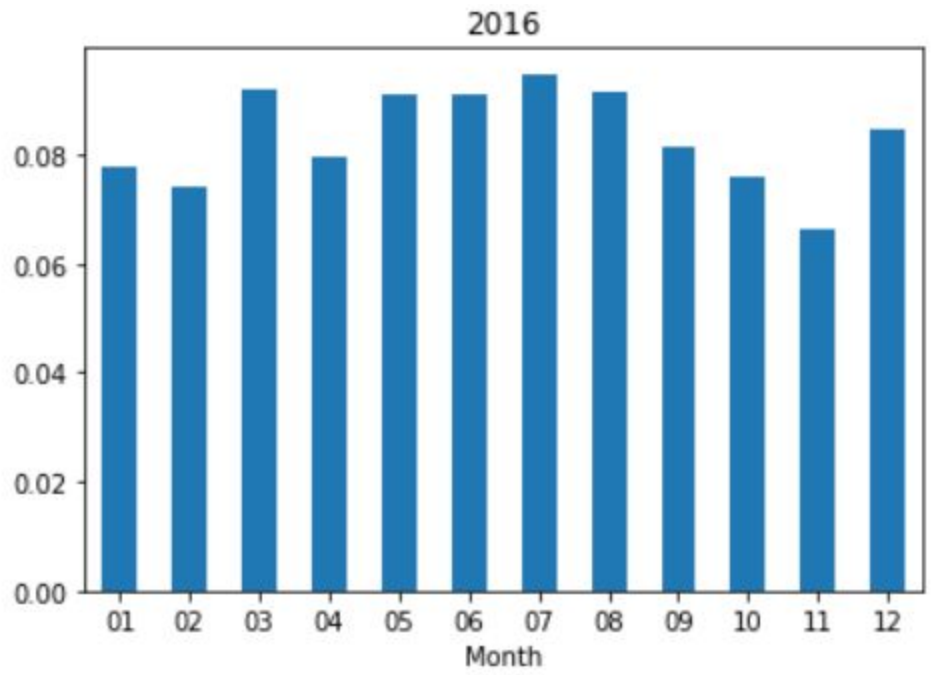
In 2013, there does not seem to be a trend.



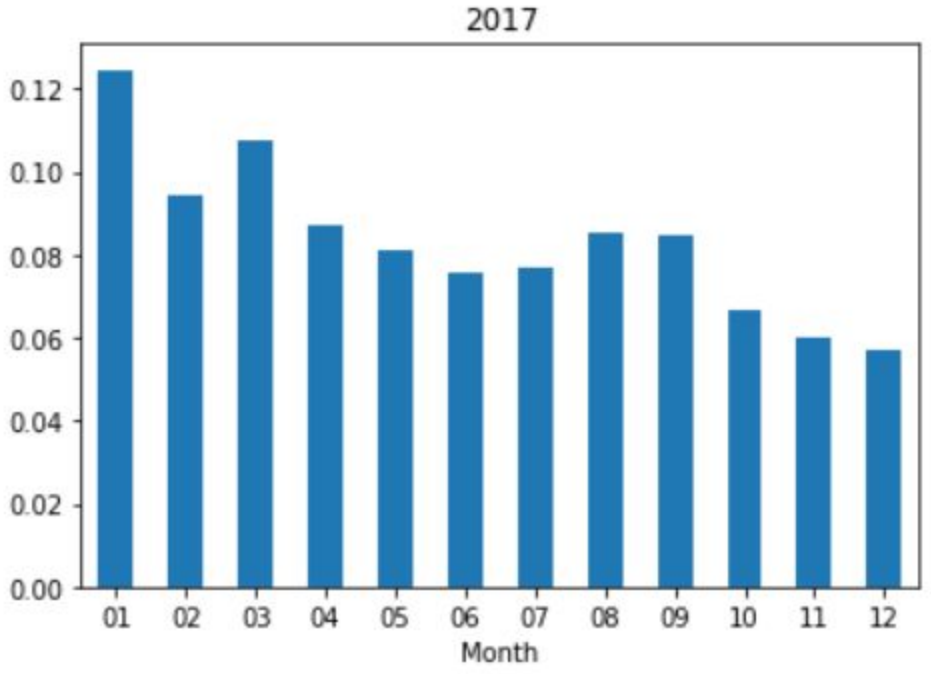
In 2014 sales are generally low in the beginning of the year and slightly higher in the later of the year



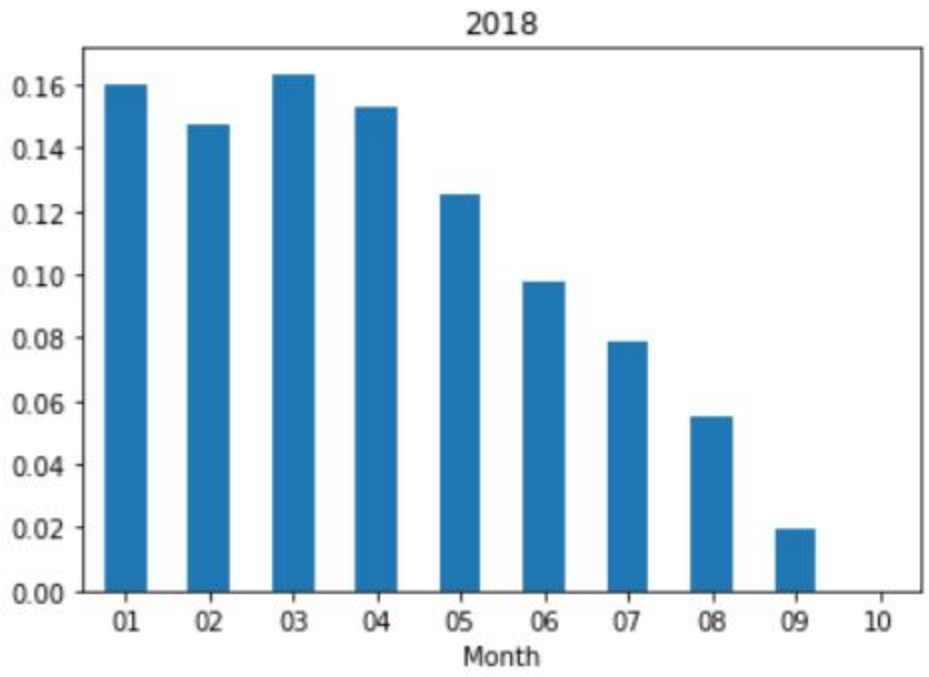
In 2015 there does not seem to be any trend



In 2016 there does not seem to be any trend



In 2017 the beginning of the year has higher sales than later in the year



Due to incompleteness of the data, in 2018 we cannot conclude a trend.

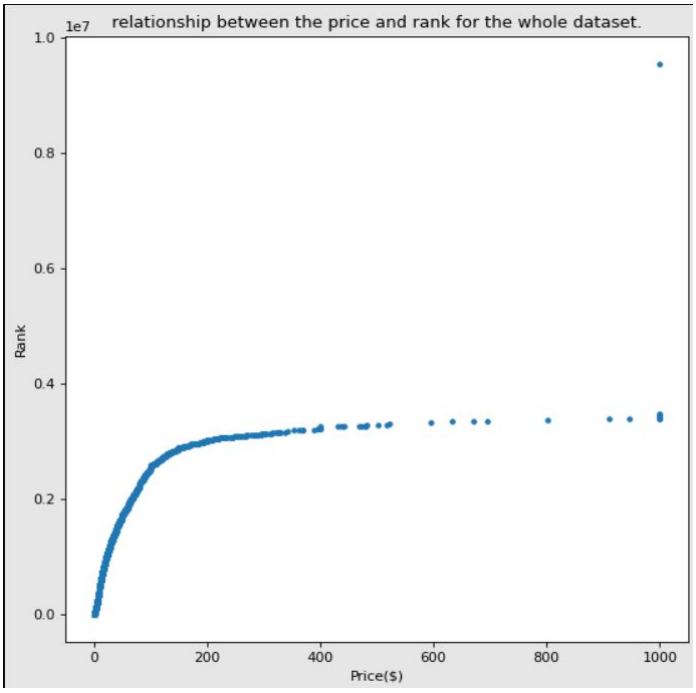
- **Challenges**
  - Challenges faced in this question is due to the fact that we do not have the actual sales data. This data analysis is based on the estimation of the sales. We assume that 1 rating = 1 sale. So we assume the number of ratings = number of sales. Which is not the case.
- **Conclusion**
  - Therefore with this analysis we cannot conclude any insight and any visualization above is misleading as it is only the trend of the estimated sales.

4.7 How does an increase or decrease of price affect the sales rank of a beauty product?

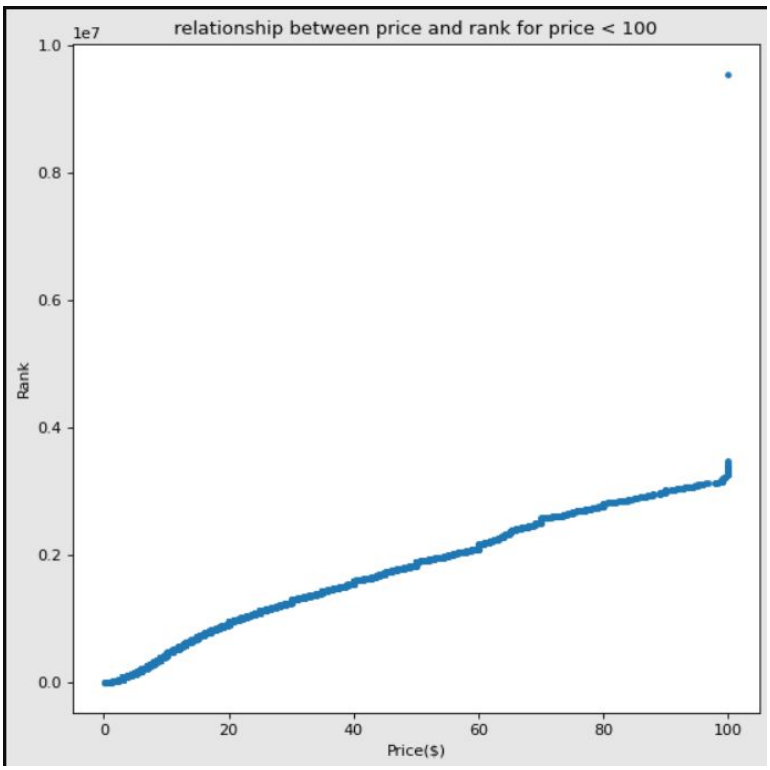
- **Why?**
  - Why is it important to know the relationship of price and sales rank? Sometimes a seller might be too focused on margin of profit which would overprice some items and deviate customers from their product. With this insight sellers are able to generally understand trends on how an item performs based on the given price.

- **How?**
  - First step is creating a dataframe containing the cleaned data of price and rank. Then upon further inspection, we found that the data contain some null values. Since we are trying to find the relationship between rank and price dropping the null values seems to be the best choice as it does not contribute anything. We did not choose to fill it with a statistics value because the insight obtained might be inaccurate. Next we plot the scatter plot x axis as the explanatory variable, price and y axis as the response variable, rank.

- **Visualization**
  -



Trend of the relationship between the price and rank for the whole dataset.

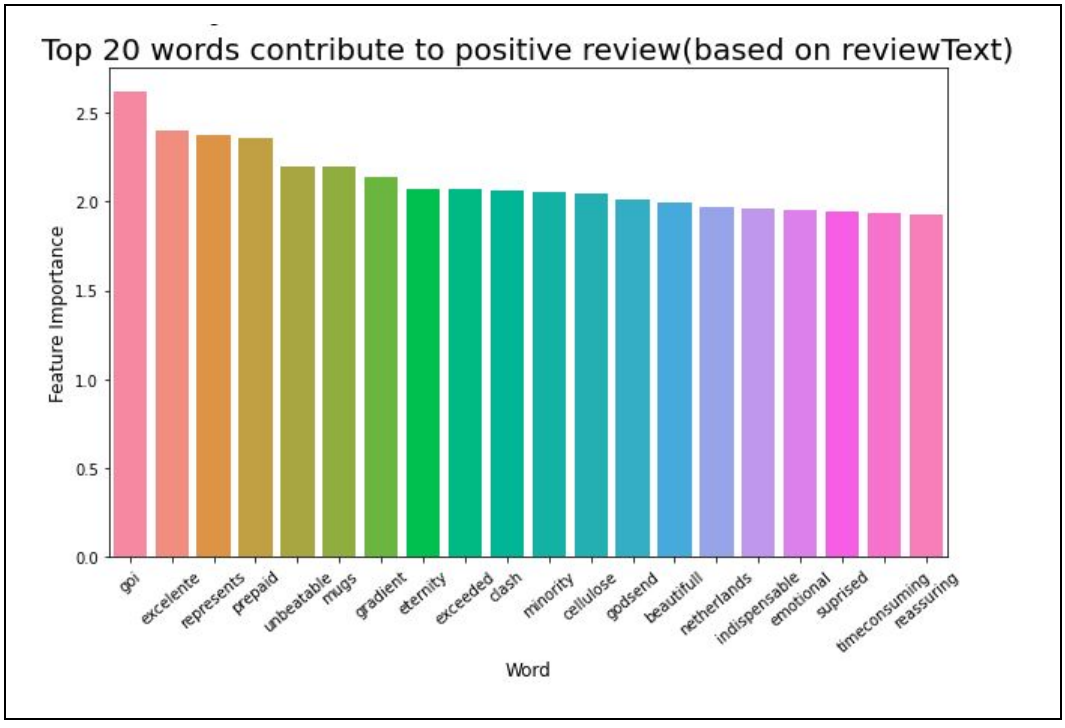


Trend of the relationship between price and rank for price < 100

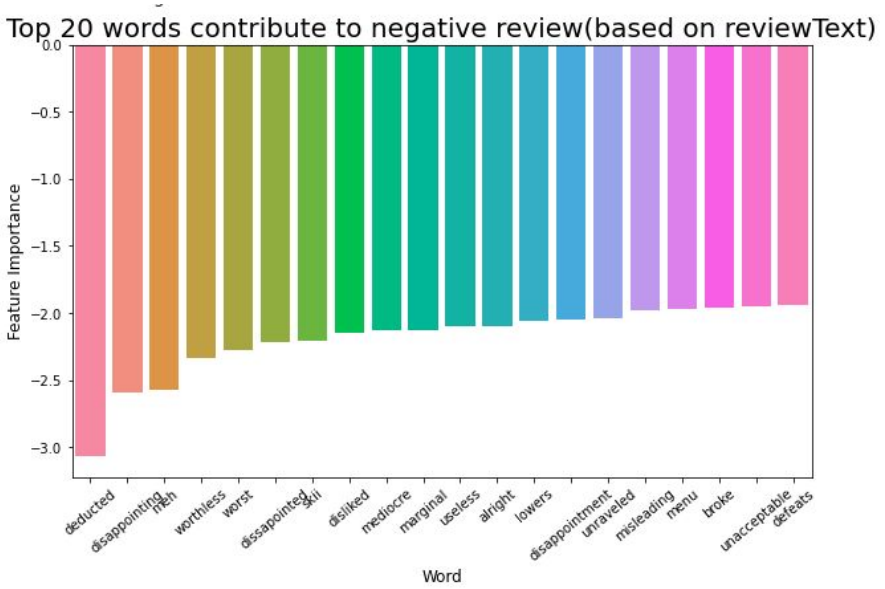
- **Conclusion**
  - From this visualization we can conclude that generally items of cheaper price tend to have overall better performance in sales. There are some outliers in both analyses.

4.8 What are the words that contribute to negative reviews? Also, what are the words that contribute to positive reviews.

- **Why?**
  - The purpose of this question is to study the words that contribute to the negative and positive review. The team uses the word ‘contribute’ because we want to know how a particular word affects the rating of an item. By exploring this question, the sellers are able to know their customer satisfaction and product rating class based on the reviews they gave.
- **How?**
  - The team decided to create a simple Logistic Regression model to study the ‘reviewText’ and ‘summary’ attributes in the supplementary data. First, create a dataframe with reviewText and overall attributes. Next, create a function to determine whether a reviewText is considered positive or negative.If the overall score is larger than the average overall score, then the review is considered as positive, else, it’s considered as negative.Thus, this has become a binary classification problem. ReviewText will be used as independent variable while the rating class as dependent variable.Then,use the text_clean function and remove_stopwords function that we created in Q3 to clean the reviewText. Next, convert the cleaned reviewText into a feature vector using CountVectorizer and split the data into 80-20. Lastly, fit the data into our model to train and plot the feature importance. Repeat the whole process by replacing “reviewText” with “summary” attributes.
- **Visualization**
  - Based on review Text

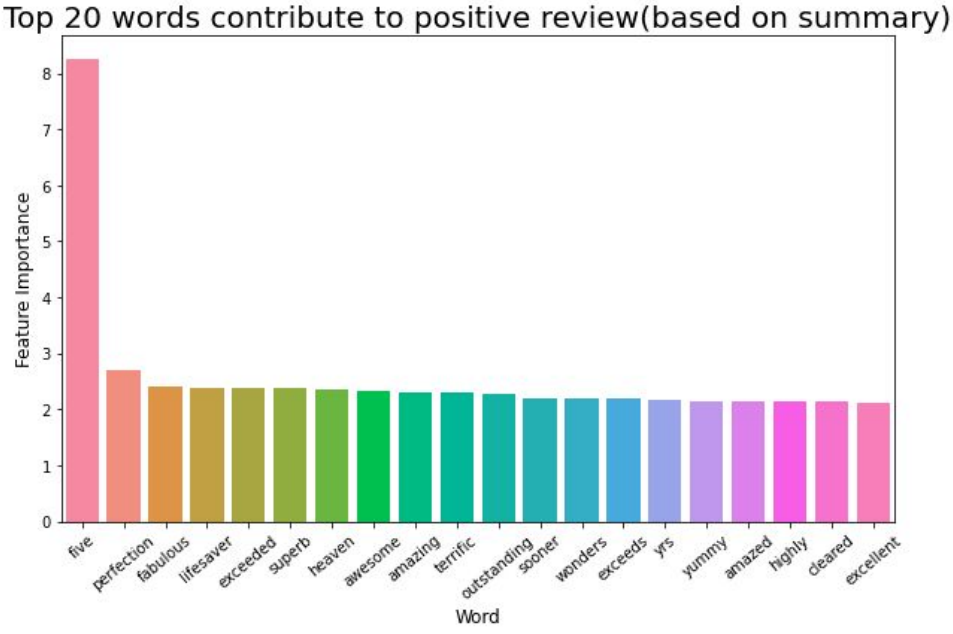


We can see that the words such as ‘goi’, ‘excelente’, ‘represents’, ‘prepaid’,’unbeatable’ are identified as words that contribute to positive review by our model.

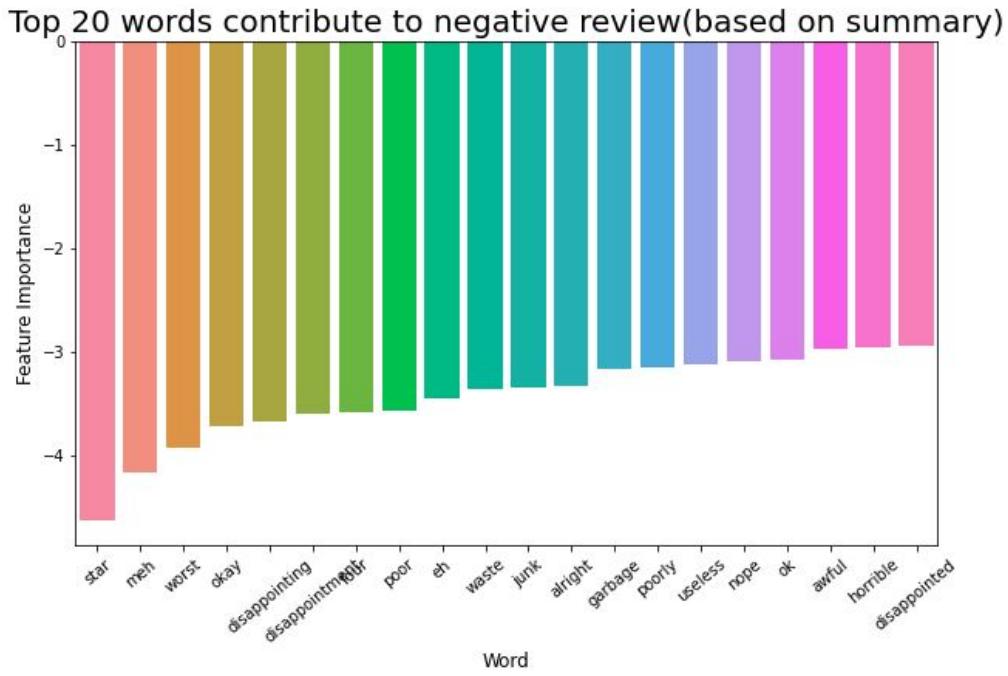


We can see that the words such as ‘deducted’, ‘disappointing’, ‘worthless’, ‘worst’,’disliked’ are identified as words that contribute to negative review by our model.

- Based on summary



We can see that the words such as ‘five’, ‘perfection’, ‘fabulous’, ‘lifesaver’,’exceeded’ are identified as words that contribute to positive review by our model.



We can see that the words such as ‘star’, ‘worst’, ‘disappointing’, ‘disappointment’,’poor’ are identified as words that contribute to negative review by our model.

- **Conclusion**
  - By comparing both visualization based on reviewText and summary, the team found out that there were some similarities among them. For example, the words ‘worst’,’disappointing’ and ‘disappointment’ are identified as words that contribute to negative review in both situations (reviewText and summary). This tells us that these words tend to give us negative ratings. On the other hand, the words that contribute to positive are different in both situations. Thus, the team thinks that there might be a confounding variable because the similar words only appear in negative reviews but not positive. However, the team could not identify the confounding variables.

5.0 Data Mining and Predictive Modelling

As our project aims to study the factors that contribute to the high sales of a beauty product, the team decided to build a machine learning model to discover the pattern and determine the factors.

Before we start building the model, we need to make a final check to make sure that our data can be fit into machine learning algorithms. At this stage, all the unnecessary attributes need to be dropped as they might cause misleading results. Besides, all the NaN data needs to be removed and all categorical variables which in “string” need to be encoded into numerical value as the machine learning algorithm only processes real numbers. In this project, we used sklearn LabelEncoder function to encode the categorical data into numerical data. Last but not least, we need to determine the type of our problem so that we can know which algorithm to be used. Our project falls into the supervised learning classification category.

Since there are no high sales attributes in our dataset, we need to define it by ourselves. We created a function to transform the product rank into high sales. We define a product as high sales if its rank is above the top 25 in overall products, else the product is defined as low sales. At this point, our project becomes a binary classification problem.We decided to use high sales attributes as the dependent variable and other variables as independent variables. Diagram 1 shows the cleaned dataset.

Total data: (7665, 6)

	brand	Review Count	Average Rating	Average Rating Class	Price	High Sales
2	2550	1.0	5.0	1	28.76	0
4	2805	15.0	4.4	1	12.15	1
11	1591	1.0	1.0	0	24.99	0
18	1287	1.0	1.0	0	3.00	0
19	3202	1.0	1.0	0	21.95	0

Diagram 1

By looking at the high sales distribution of the products, we can see that our data is slightly imbalanced. Diagram 2 shows the high sales distribution of the products.

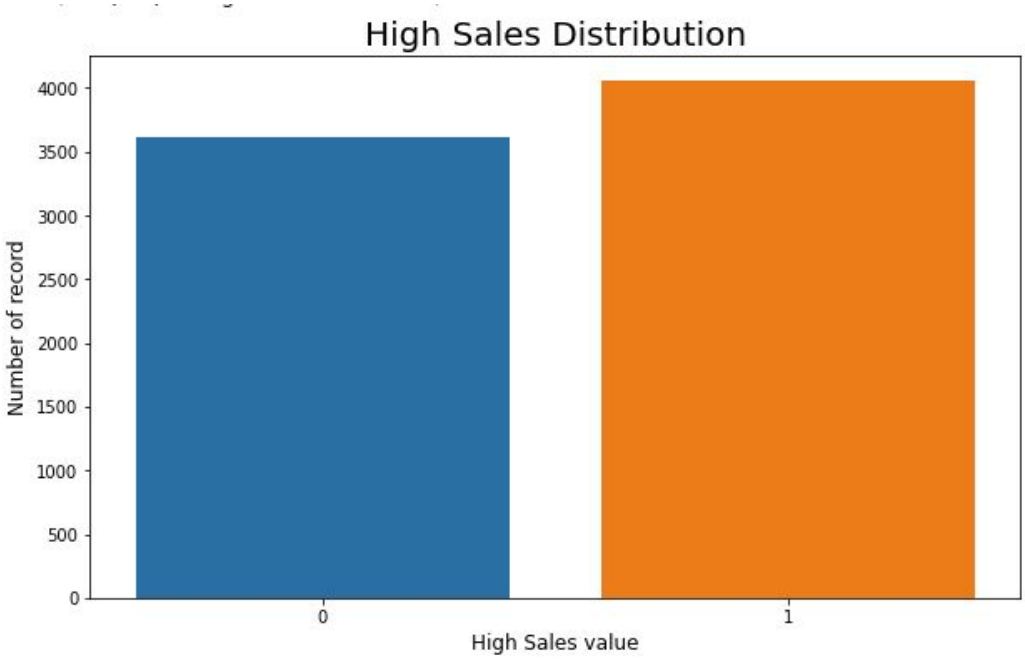


Diagram 2

After that, we split the data into the train set and test set with 80-20 percentages.Diagram 3 shows the result.

```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=42)

#CHECK THE SHAPE
print(X_train.shape)
print(X_test.shape)
print(y_train.shape)
print(y_test.shape)
```

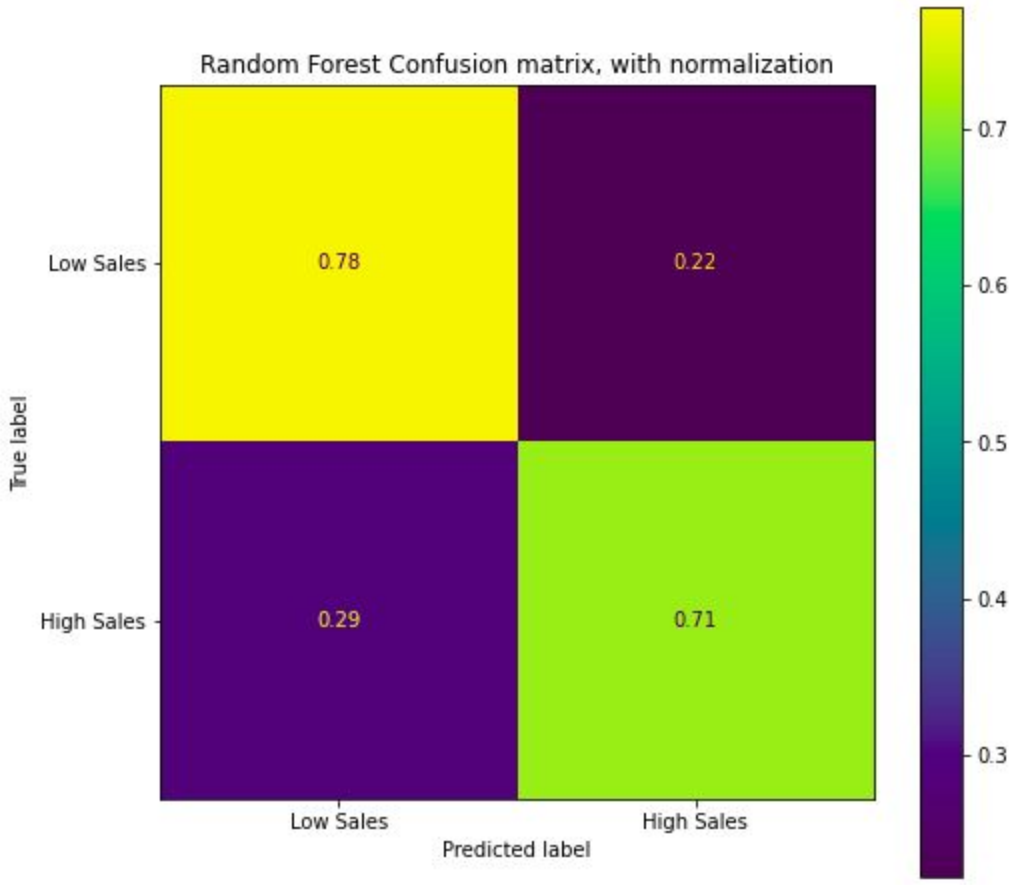
```
(6132, 5)
(1533, 5)
(6132,)
(1533,)
```

Diagram 3

In this project, the team decided to build two models using Random Forest Classification Algorithm and XGBoost Algorithm respectively and compare the result generated by the two models. The primary reason for using these 2 algorithms is for their performance and ability to handle imbalance data, as our data is slightly imbalance [1,2]. Besides, they are easy to implement.

In terms of comparing, the team compares the F1 score and confusion matrix of 2 models because F1 score takes in precision and recall into consideration as our data is imbalanced.

5.1 Random Forest Classification Algorithm

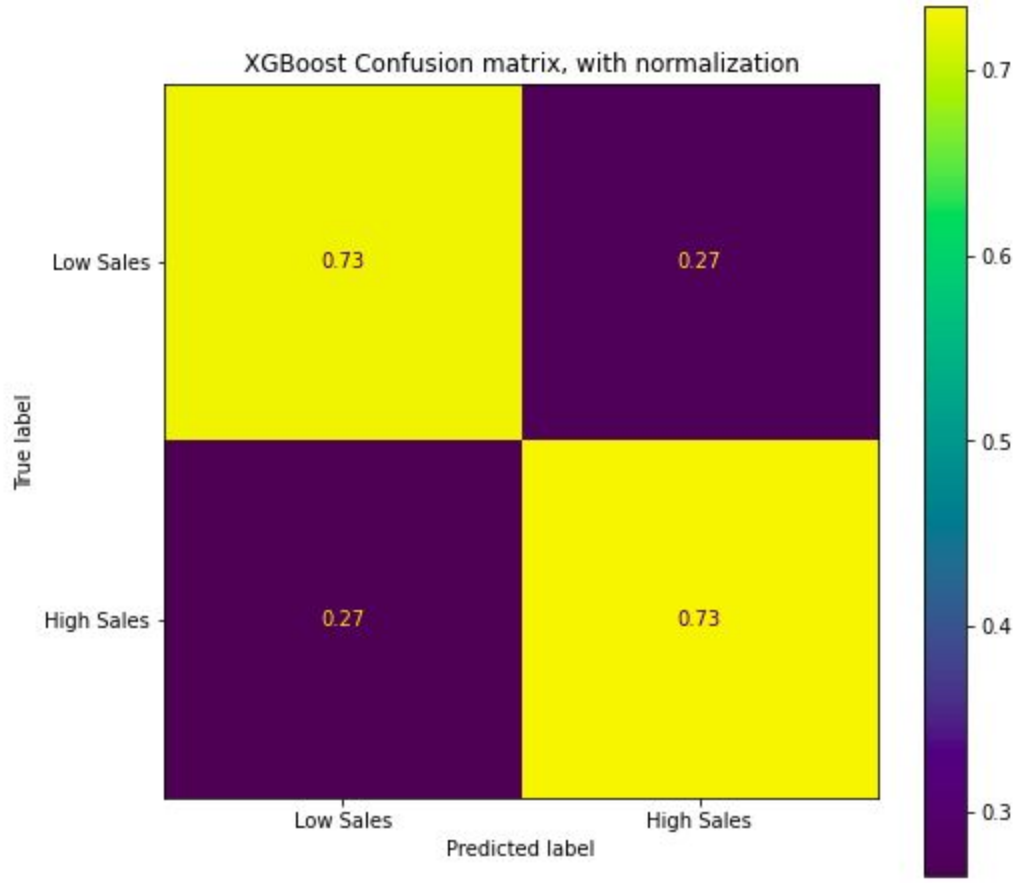


The values of TP,TN,FP and FN are 0.71,0.78,0.29 and 0.22 respectively.

	precision	recall	f1-score	support
0	0.69	0.78	0.73	694
1	0.80	0.71	0.75	839
accuracy			0.74	1533
macro avg	0.74	0.75	0.74	1533
weighted avg	0.75	0.74	0.74	1533

The F1 score is 0.74

5.2 XGBoost Algorithm



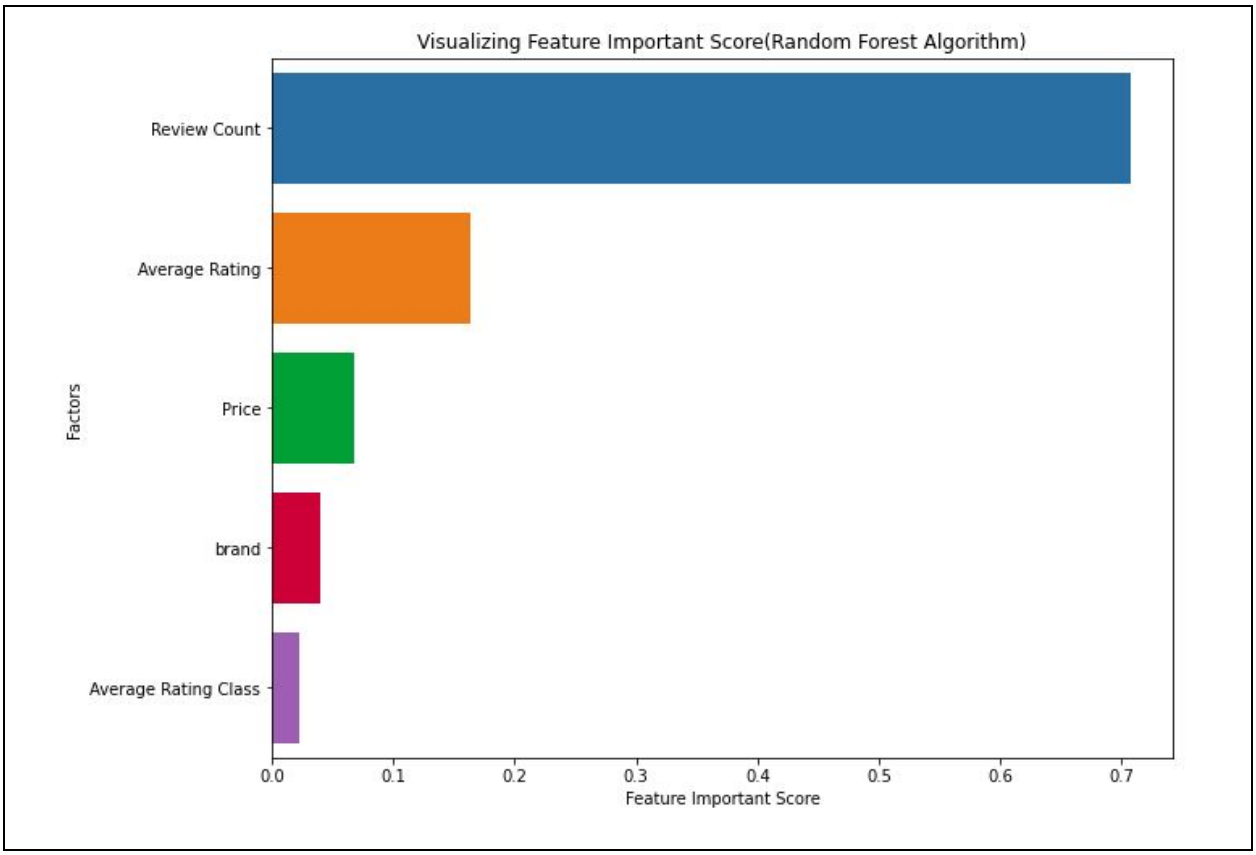
The values of TP,TN,FP and FN are 0.73,0.73,0.27 and 0.27 respectively.

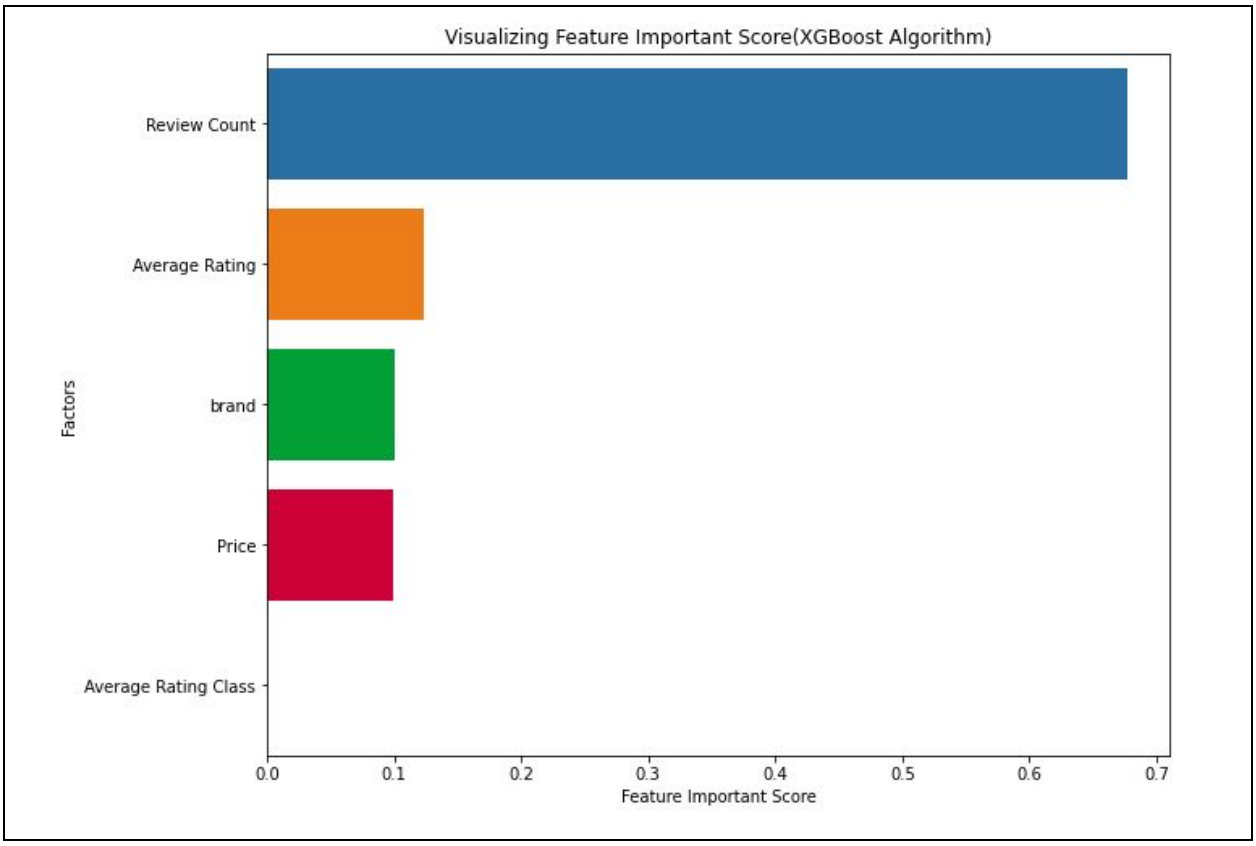
	precision	recall	f1-score	support
0	0.69	0.73	0.71	694
1	0.77	0.73	0.75	839
accuracy			0.73	1533
macro avg	0.73	0.73	0.73	1533
weighted avg	0.73	0.73	0.73	1533

The F1 score is 0.73

6.0 Data Visualization

The feature importance, also known as variable importance, describes which features are relevant to the prediction. It assigns a score to our input features based on how useful they are at predicting the target output [3]. In our case, it can help us have a better understanding on how each independent variable (Brand, Review Count, Average Rating, Average Rating class and Price) in predicting the dependent variable (High Sales). The 'Review Count' attribute obtains the highest feature importance score in both models we built.




Visualizing Feature Important Score(XGBoost Algorithm)
Review Count
Average Rating
brand
Price
Average Rating Class
Factors
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Feature Important Score

7.0 Conclusion

In conclusion, after analysing the amazon dataset, it was hard to conclude any correlation or causality for most of the questions asked. When finding relationships between rank and number of ratings we were unable to conclude causality due to the fact there might be some confounding variables. We managed to find a common pattern in the description of highly rated beauty products. However, we can't conclude much from this as it is just common beauty product words like ‘skin’ or ‘natural’.

Furthermore, when analyzing the relationship between price and rating of an item, we found that a general trend is that as price increases the rating increases. However, we cannot conclude any causal relationship or correlation. Moreover, when trying to find a pattern in the number of sales against time. We did not manage to come to a conclusion as the data was based on an estimation of number of sales and it may/is inaccurate and could be misleading. Analysing the relationship between price and sales rank we found that cheaper products then have lower sales rank (rank < 100,000) but we are unable to say that having lower price causes lower sales rank. When finding words that contribute to a negative or positive review, we manage to find multiple common words. Potentially with this insight we could build a recommendation model for future work.

Lastly, we used random forest classification and xgboost to find the feature importance of the attributes. Both outputs have a bit of variation but however the common for both outputs is that the number of reviews has the highest importance.

8.0 References

[1] Why Random Forest is My Favorite Machine Learning Model by Julia Kho
https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706

[2] Decision Tree vs Random Forest vs Gradient Boosting Machines: Explained Simply by Stephanie Gien
https://www.datasciencecentral.com/profiles/blogs/decision-tree-vs-random-forest-vs-boosted-trees-explained#:~:text=Like%20random%20forests%2C%20gradient%20boosting,one%20tree%20at%20a%20time.

[3] How to Calculate Feature Importance With Python by Jason Brownlee
https://machinelearningmastery.com/calculate-feature-importance-with-python/