Product Listing From Amazon India

Abhishek Mishra	PES2UG19CS009	abhishekmis40@gmail.com
Abhishek V	PES2UG19CS012	v.abhishek1347@gmail.com
Anush P Upadhya	PES2UG19CS053	anushpupadhya@gmail.com

Links
Colab notebook:
https://colab.research.google.com/drive/107wwi3RH1nrsN8JCz9fJus7ENzNqCyiM?usp=sharing
Github repository: Amazon Data Analysis

ABSTRACT:

The data set mainly focuses on product listing on amazon.com : A dataset was selected from the internet with all the necessary requirements like necessary categorical variables, null/missing values or outliers to name a few. Henceforth, we clean the data by handling missing data or duplicate or irrelevant data or any outliers if present. The next step was exploratory data analysis (EDA) which includes a set of techniques to display data in such a way that interesting features will become apparent. This is followed by data visualization of the current data in the form of various graphs.we have represented line graph heat maps bar charts histograms using statics results obtained . Penultimately, we end it with normalization and standardization and discuss why it is needed. We conclude this with hypothesis testing and correlation.

INTRODUCTION:

An Amazon product listing is the product page for each of the items you sell on Amazon. It is made up of the information you enter when you list your product including its title, images, description, and price. Shoppers on Amazon use product listing pages to make a purchase, i.e. the Add to Cart button is on all product listing pages. As a result, getting the product listing right will determine the success of selling your products on Amazon.

An Amazon product listing performs several functions, but the two main ones are:
i) Enables your products to be found in Amazon searches
ii) Encourages shoppers to purchase your product more easily
It also helps them determine which product category is the most selling and which is the least and helps them make a business strategy accordingly.

We wanted to analyse which category of products were sold the most and how and what kind of factors affect the sale of products.

DATASET:

This dataset was obtained from kaggle (Product Listing From Amazon). This dataset originally includes around 30,000 records in it. It roughly had around 6-7% of NULL values.This dataset includes the different products bought in the month of october 2019 at (which was an offer period) different times of the day.

The following where the features available in the dataset :

→ Unique Id,
→ Crawl Timestamp,
→ Category,
→ Product Title,
→ Product Description,
→ Brand,
→ Pack Size Or Quantity,
→ MRP,
→ Price,
→ Offers,
→ Combo Offers,
→ Stock Availability

The variables Category, Brand , Stock Availability where the categorical variable. The variable MRP , price, Offers where the numerical variables. The rest of the variables were raw strings.

PREPROCESSING OR DATA CLEANING:

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. In our dataset we had used used basic string manipulation to extract important information from it (In our case we extracted the brand name from the product description).The other techniques used by us for the numerical column was ‘np.nanmean()’ function which basically fills the empty value with the mean of the column and ‘np.nanmedian()’ function which also works in the similar way but instead replaces the empty value with the median. Since our data was divided into 6 categories we had found mean and median values individually for each category. We also used the datetime method to convert our date column to a timestamp object which was initially as a string.

DATA AFTER CLEANING:

```
[63] amzn.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 30000 entries, 0 to 29989
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Uniq Id               30000 non-null  object 
 1   Crawl Timestamp       30000 non-null  object 
 2   Category              30000 non-null  object 
 3   Product Title         30000 non-null  object 
 4   Product Description   30000 non-null  object 
 5   Brand                 30000 non-null  object 
 6   Pack Size Or Quantity 30000 non-null  object 
 7   Mrp                   30000 non-null  float64
 8   Price                 30000 non-null  float64
 9   Offers                30000 non-null  object 
 10  Combo Offers          30000 non-null  object 
 11  Stock Availibility    30000 non-null  int64  
dtypes: float64(2), int64(1), object(9)
memory usage: 4.2+ MB
```

```
amzn.isna().sum()

Uniq Id                 0
Crawl Timestamp         0
Category                0
Product Title           0
Product Description     0
Brand                   0
Pack Size Or Quantity   0
Mrp                     0
Price                   0
Offers                  0
Combo Offers            0
Stock Availibility      0
dtype: int64
```

EXPLORATORY DATA ANALYSIS:

Exploratory data analysis is the process of performing investigations on data so as to discover patterns, spot anomalies and to test for hypotheses with the help of summary statistics and graphical visualizations.

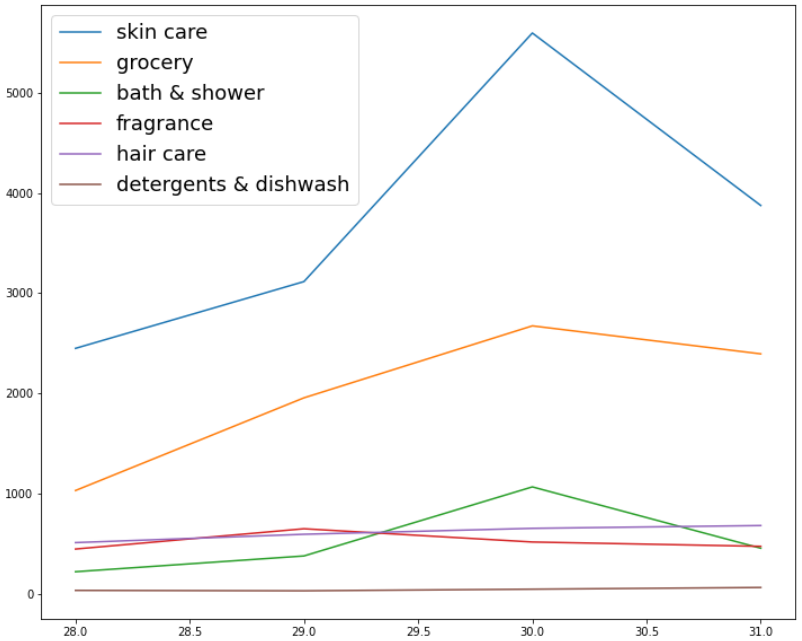
- The Correlation matrix

  Correlation matrix is used to get insights about the relationships between the variables of the dataset

From the correlation matrix, we can infer that "MRP" has a strong correlation with "Price" whereas "MRP" and "Stock Availability" have a weak correlation. This correlation is graphically represented using a heat map.

- Line Graph

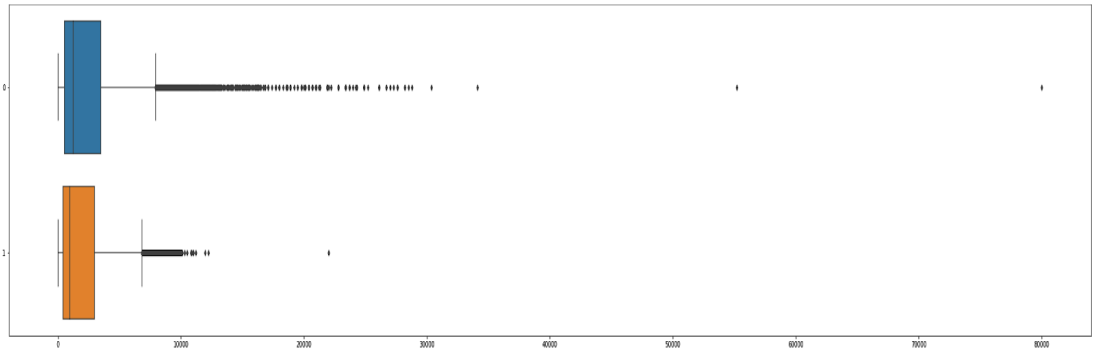
  The lineplot from seaborn library measures change over time by plotting individual data points connected by straight lines.



From the above line graph we can infer that Skin care products have a higher sales margin than other products listed. It is also evident that Skin Care products sales decreases in the last few days of the month.
The sales of Detergents & Dishwashing products remain constant at the lower end throughout the month.

- Boxplot

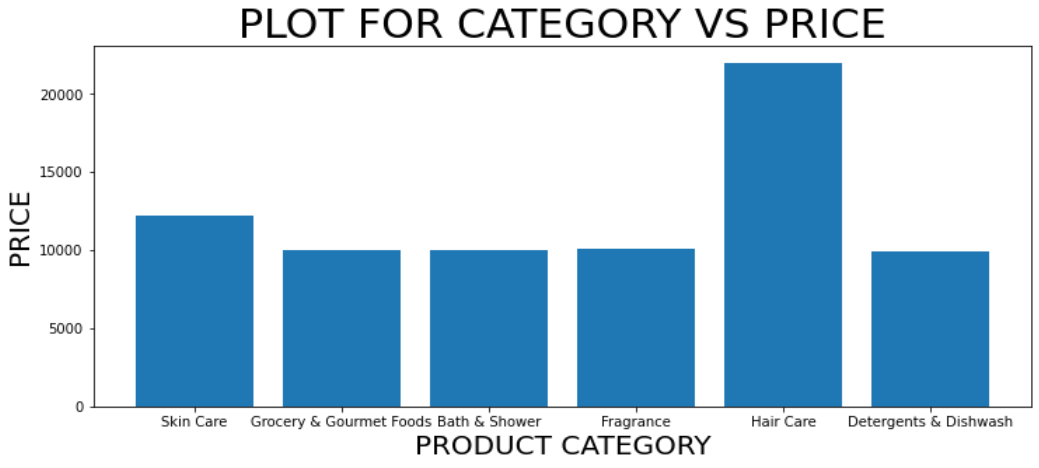
  A box plot (or box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable. The box shows the quartiles of the dataset while the whiskers extend to show the rest of the distribution, except for points that are determined to be outliers.
  A data point that is more or less than 1.5 times the interquartile range is considered as an outlier.



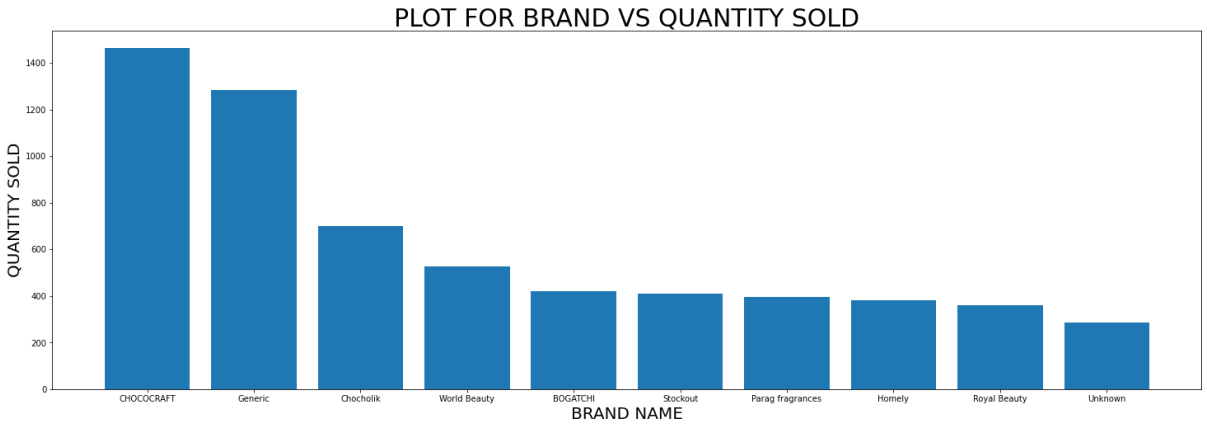
Boxplot to compare the summary of MRP and Price of listed products.

- Bar Graph

  The bar plot from matplotlib.pyplot library helps plot data in rectangular bins that represent the total amount of observations in the data for that category.



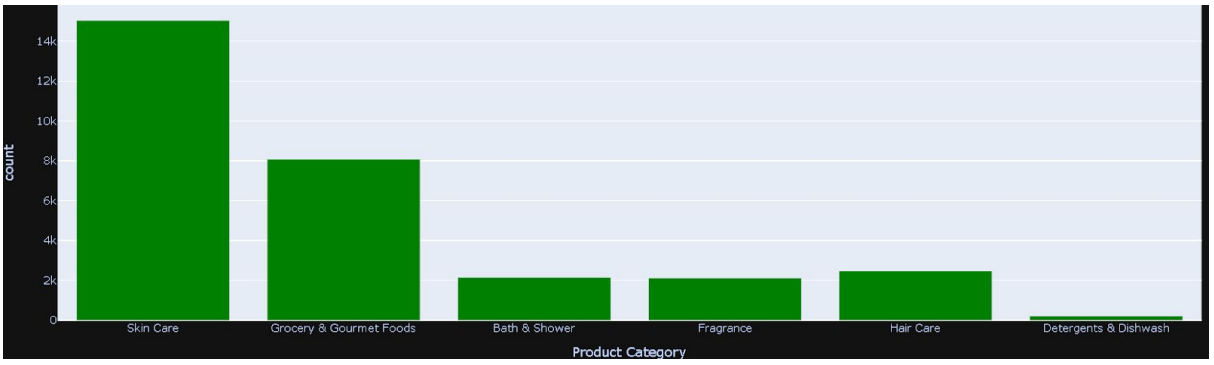
From the bar graph “Product category” vs “Price”, it is evident that hair care products are more expensive than the other products.



In another bar graph representing the number of items sold by top brands, we can infer that “Chococraft” is the best selling brand in the month.

- Histogram

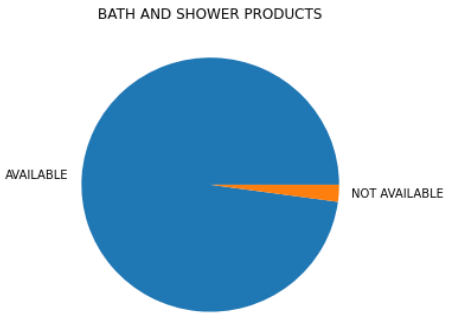
  Histograms provide a visual interpretation of numerical data by indicating the number of data points that lie within a range of values.



From the histogram representing the number of products sold in each category, one can infer that Skin Care products are the highest selling products in the month with over 15,000 units sold and Detergents & Dishwash are the least selling products with less than 200 units sold in the month.

- Pie Chart

  A pie chart is a circular chart that shows how data sets relate to one another. The arc length of each section is proportional to the quantity it represents, usually resulting in a shape similar to a slice of pie



From the pie chart representing the stock availability of products under different categories, it is evident that the majority of the products are on stock and very less number of products are out of stock. Products under the "Bath and Shower" category has slight higher percentage of unavailability.

Hence, from the above analysis we can conclude that skin care products are the highest grossing products in the month with a large stock availability and more offers.

HYPOTHESIS TESTING:

The general idea of hypothesis testing involves: Making an initial assumption. Collecting evidence (data). Based on the available evidence (data), deciding whether to reject or not reject the initial assumption.

Hypothesis: To check whether the total offers increases with time
We assume the significance level to be 5% = 0.05
Average offer price on 28th: x
Average offer price on 31st: mu0
Null hypothesis H0: mu0 <= x
Alternate hypothesis: mu0 > x

Formula: $z = (x - \mu) / (\sigma / \sqrt{n})$

We obtain the value of P as 0.184 which is greater than the significance level(0.05). Hence we accept the null hypothesis.
We fail to reject the Null Hypothesis i. e. we cannot say that the offers have increased from 28th to 31st.

RESULT AND CONCLUSION:

Since this data was collected during one of the offer and festival seasons , it was observed that Skin care products were the highest selling products with high stock availability and a lot of offers.

We were also able to conclude based on the analysis that the sale of a product depends mainly on two factors that are :

→ Stock availability of the products

→ The offers available on the product.