**PROJECT REPORT**

**ANALYSIS OF ONLINE SHOPHING**

**TEAM DETAILS:**

Y RAGHAVENDRA PES2UG19CS464

SRIVATSAN N PES2UG19CS408

SUDIPTA PAHARI PES2UG19CS411

BILAL RIZVI PES2UG19CS423

**ABSTRACT:**

For various industries world wide ,recent years have been defined by the remarkable growth of e-commerce. Enabled by the internet, retailers can reach more customers,spread much further in the distribution chain, and optimize there resources.In the new market environment customer experience has become a source of competitive advantage. A quantitative methodology is used ,which incorporated a feedback option by which we can increase our sales by looking thorough our customers queries

**INTRODUCTION:**

Clothes are a daily neccessity in our day to day life .We are here analysing a dataset on online shopping of clothes which we got from kaggle website.Increasing online sales is the primary goal of countless businesses , large and small alike.Whether you run a mom-and pop retail business or work for a vast ecommerce giant like Amazon ,increasing sales through online channels is a lot hard than we think is.

Fortunately ,there are some ways you can make more sales online,so we are here to find a way to increase our sales by analysing our dataset.

**DATASET:**

This is an analysis of online clothing sales.In this data set there are the title of the clothes ,their price,the ratings they got,the numbers of pieces sold .The description of the product such as color and the size of the product.The dataset also has the shipping details like countries shipped to and the option details.It also has columns containing the details of the merchant such as merchant name and his id.

Number of rows : 1573

Number of columns:43

Percentage of empty values :3%

**DATA CLEANING:**

Data cleaning is the initial process which needs to be done on a dataset as it might consists of many inconsistent, irrelevant data as well as outliers. We have to analyse each columns properly to figure out what cleaning process should be used.

Initially we remove the columns urgency\_banner and urgency\_text as many rows of these columns were empty(50%).

Then we figure out a way to remove rows for colums which are null(NaN) such as rating or product size or merchant rating as these products lack the crucial information, so they are not recommended. We use a function that finds the empty rows or the provided column and eliminate the rows.

The next step is to remove all the remove all the duplicate rows to purify the dataset further.

The final process in the cleaning is to remove the outliers and irrelevant data in the column product\_variation\_size\_id which contains the size of the given product. We first convert all of rows to upper cases then remove the unwanted data. We are limiting to products which have sizes .

XXXXS,XXXS,XXS,XS,S,M,L,XL,XXL,XXXL,XXXXL

Table

Description automatically generated

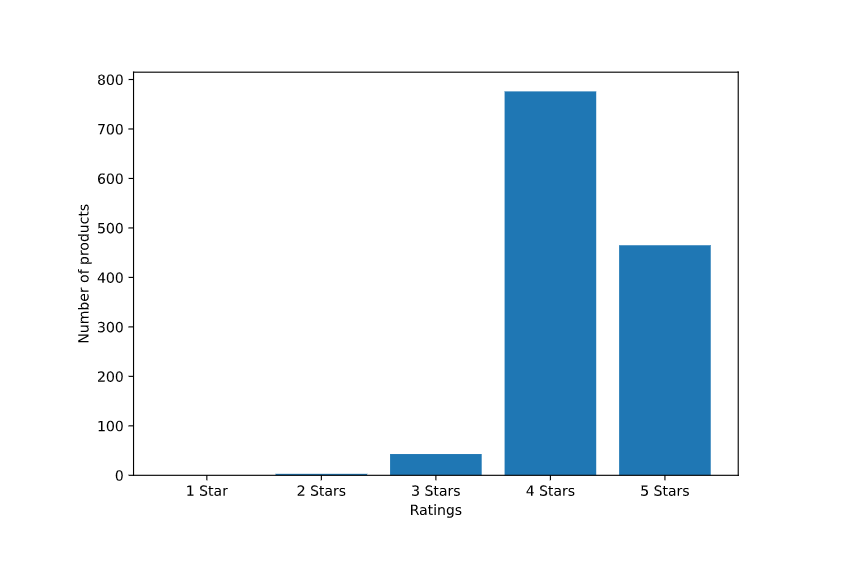
Graphical user interface, text, application

Description automatically generated

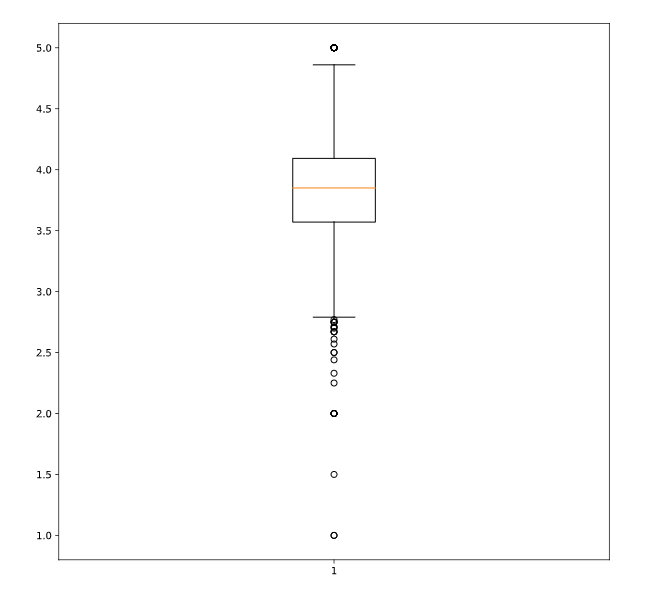
**Graph visualization:**

The idea here is to derive meaningful insights from the dataset chosen. This aids the better understanding of data and deriving meaningful insights which can help in an in-depth analysis. Also, we can have a wider look at the data in a moment by looking at the graphs.

Since the data deals with clothing sold for summer, it would be a good insight to know how people rate the clothing. Hence, the first plot was the number of merchandises to the rating given by the customers. This allows us to understand what the customers feel about the clothing.

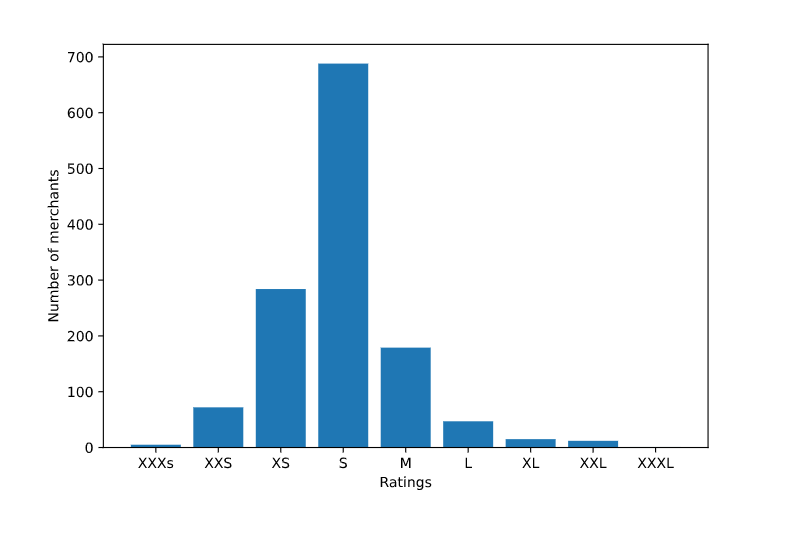


To understand the rating further, we plotted a box plot to understand if there were any outliers and understand where the median lies and how the values spread around it.



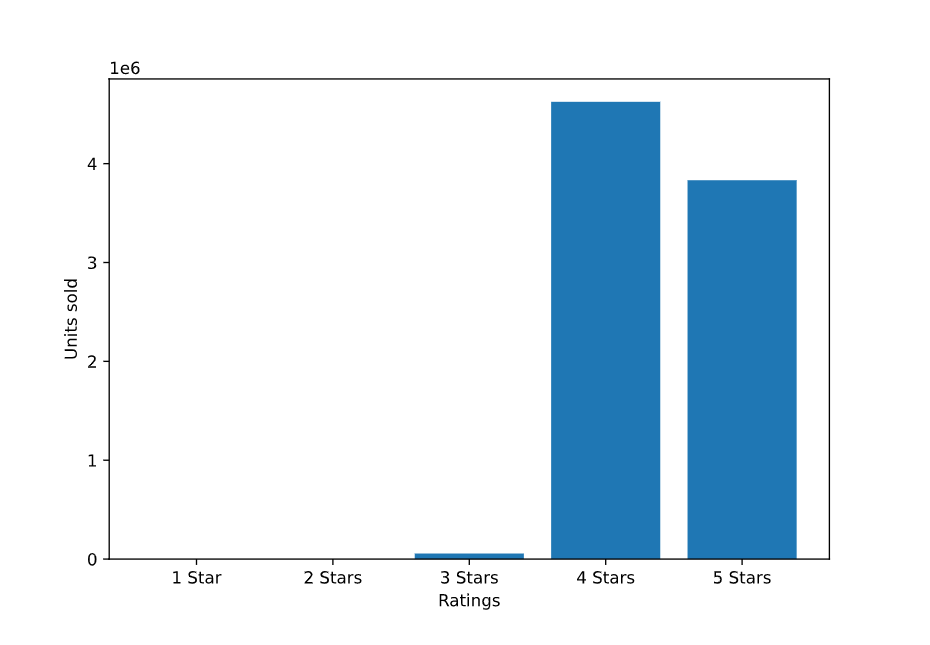
From the above plots, we can understand that the average rating for the products is around 3.8 stars.

Clothing comes in a variety of sizes. So, it would be great to understand the sizes of clothes popular amongst the public. This would also give us an idea about what kind of population it is.



From the above plot, we can derive that most of the clothing sold was of small size. The next popular size is extra small and the next popular one is medium size.

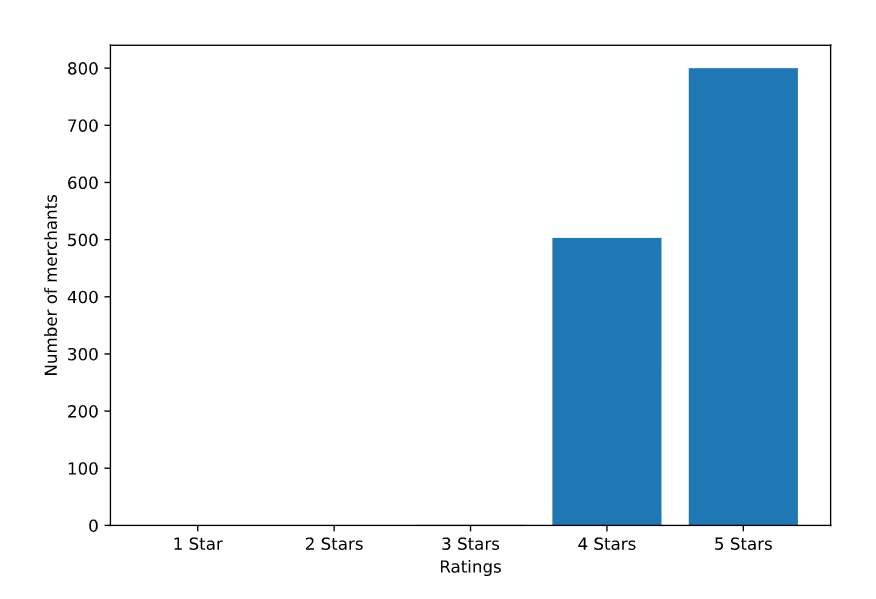
Next we also compared the sales of the products based on the rating.



This graph tells us that majority of the buyers bought merchandise that was rated 4 starts and the next highest selling merchandise was rated 5 stars.

Our dataset also contains information about the merchants and they too have been awarded ratings by the customers.

A plot of this will help us understand what the customers have to say about the sellers.



This plot tells us that majority of the population has rated the merchants 5 stars and a few of them 4 stars.

Through all these different types of plots, we can monitor sales of the clothing and help us improve the sales as we get overall data of the products in a quick instance.

# Normalization and Standardization

# Normalization and Standardization are techniques used to scale different ranges of data present in different columns of the dataset.

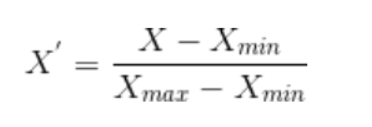
**Normalization** scales different ranges of data between zero and one based on the maximum and minimum values of the feature.

Normalization is done so that all the input variables have same treatment in the dataset.

It also helps to keep the data clearer ,easier to maintain and also convenient to change

Standardization is useful when data has varying scales. It assumes data Gaussian(bell curve).

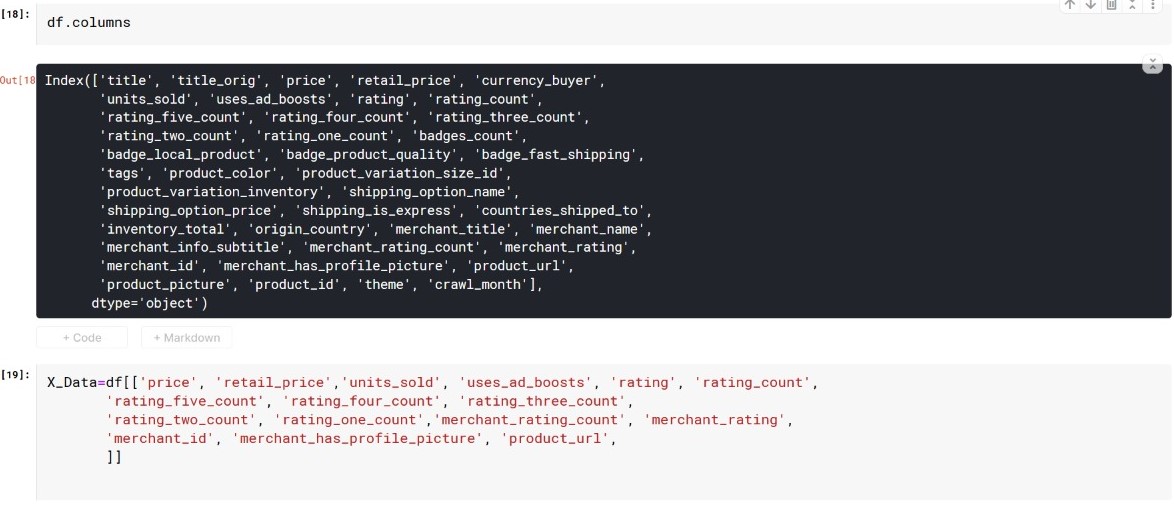
Custom lambda function is used to normalize the data set



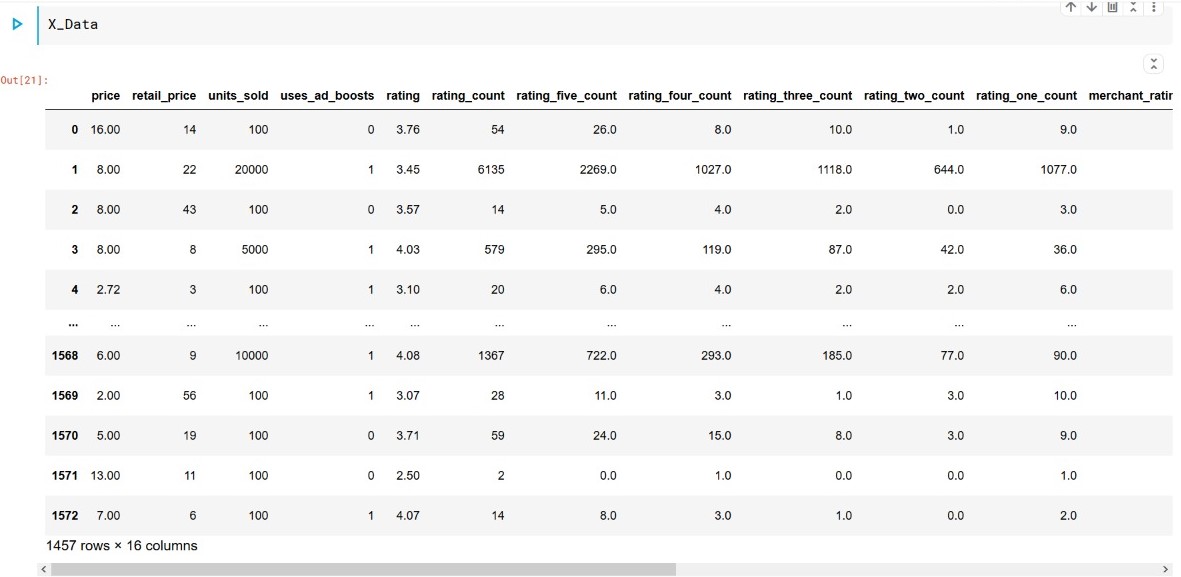
**Data before Normalization :**

# 

# Data for Normalization :



***Data After Normalization :***



**Hypothesis:**

A hypothesis test produces a number between 0 and 1. **Hypothesis is an assumption about population Parameter.**

Steps for Hypothesis testing of a given dataset

Step 1: State the Null Hypothesis

Step 2: State the Alternative Hypothesis

Step 3: Set α(0.05 generally)

Step 4: Collect Data

Step 5: Calculate a test statistic

Step 6: Construct Acceptance / Rejection regions

Step 7: Based on steps 5 and 6, draw a conclusion about H0

Hypothesis 1(Units Sold)

H0: µ=5000

H1:µ≠5000

α=0.05(two-tail test)

Test statistic:- Z=(x̄-µ)/(σ/

Z=-0.016

Z-score=(-1.96,1.96)

-1.96<-0.016<1.96

=> we accept null hypothesis avg units sold=5000

We have picked up number of units sold from the dataset to be our population parameter. We set our null hypothesis H0: µ=5000 and H1:µ≠5000. The level of significance was set to 5%. We pick up a random sample from the given population by using the random function and calculate the approriate sample parameters(mean and standard deviation) and substitute it in the test statistic formula mentioned above.We use the random.seed function to make sure that the same sample is being picked at all times when the code is run in order to maintain consistency.Since significance level or α is 0.05,we divide it by 2 as it's a two tail test and from the z table,we can conclude that the acception region should be between the intervals -1.96 and 1.96.The Z-value was found and compared with the corresponding Z-score value. Given that it is a two-tail test, the z value lies in the acceptance region. Hence we accept our null hypothesis H0: µ=5000.(The average number of units sold in the given population is 5000)

Text

Description automatically generated

CORRELATION:

A correlation coefficient is a numerical measure of some type of correlation ,meaning a statastical relationship between two variables.The variables maybe two columns of a given data set of observations,often called a sample, or two components of a multivariate random variable with a known distribution.

POSITIVE CORRELATION:

It is a relationship between two variables in which both variables move in tandem-that is ,in the same direction. A positive correlation exists when one variable decreases ,or one variables increases while the other increases.

NEGATIVE CORRELATION:

It is a relationship between two variables in which one variables increases as the other decreases ,and vice versa.A perfect negative correlation means the relationship that exists between two variables is negative 100% of the time.

A picture containing text

Description automatically generated