Impact of Data Analysis on Amazon Mobile Dataset

Shubham Kumar
Department of Information Technology Oriental Institute of Science & Technology Bhopal, India.
shubhamsingh3765@gmail.com

Shadab Pasha Khan
Department of Information Technology Oriental Institute of Science & Technology Bhopal, India.
shadabpasha@gmail.com

Abstract—These days every technical stuff is based upon data. Everyday IT companies are evolving & wrapping up with new technology that meets the current demand of numerous clients. Many IT companies came into picture like Amazon, eBay, Walmart, Myntra, Flipkart, Snapdeal but Amazon came and captured the market through its flawless business model and big data analysis and providing Items and services for clients using two-way revenue. It fulfils all the customer's requirements at a very satisfactory level and widely used over the globe.

This paper aims to study Amazon's impeccable Business Model. Several kinds of company terms are noted in this paper for understanding the strategies running behind the Amazon E-commerce service. With proper tools and analysis methods we examine the dataset at various different levels and find out the impacts of Brand image, rating, total reviews, prices etc by analysing the Amazon mobile dataset.

Keywords- Amazon, digital Marketplace, python, two-way revenue, analysing tools, business model;

I. INTRODUCTION

Data Analysis is the process in which data is transformed, cleaned so that one can make business decisions for future. The most required fuel of the twenty-first century is nothing but data. There are several IT companies which generate lots of data (in huge amounts like exabytes). Every day at each point of time the data is generated and counted up to various references. In E-Commerce people solely depend on the reviews & ratings given by the customers who already bought and use the products. Online surveys on shopping sites are important to understand customer needs and feedback to upgrade the product quality and outcomes (D.R. Kumar Raja et. al, 2017) [14].

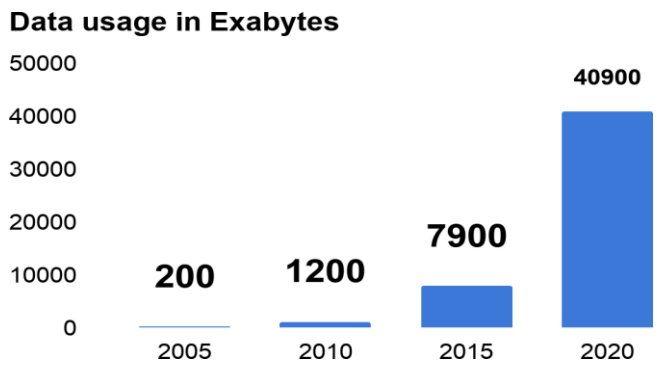

Fig 1: Data Usage from 2005 to 2020

About Amazon Business Model Amazon has the most valuable company in today's World. The rise of companies such as Google, Amazon, Facebook and Apple on the Internet has facilitated information technologies. These businesses are interlinked and shaping the business ecosystem (Juan Carlos Miguel,2016) [17]. Advantage of Amazon's Business Model is that: -

A. Big Data

Amazon clusters has several servers which are working together to perform different tasks on distributed databases on different servers parallelly. Amazon services are used in big data analysis to boost business efficiency (Ankush Verma,2018) [8].

Amazon Knows the requirement of Customers that varies from place to place with the help of large amounts of data stored as per there buying Experience It Started Its own selling brands which are more effective in all the ways. The different brands are like Symbol for clothes and Amazon Basic for electronic products. There's 5Ws of data dimension What's the data quality, Why the data occurred, from where the data is coming, When the data occurred, who got the data and How the data was transmitted (Janson Zhang,2013) [19].

Predictive Shipping: - It stores the product to the nearest warehouses before the orders to be done. This is done by the data stored and analysed and makes it faster in the market.

Amazon as a logistics: - It partners with the local business group and serves them as a delivery merchant as a result it gets to know what is the requirement of the people of that local area which it feeds to use for the future predictive Shopping.

B. Cutting Edge services

Amazon has something exceptional like no other IT companies have. The Amazon Web Services is the largest cloud server in the world. Most of the web sites use Amazon Web Services to run online and for which the revenue gets generated which are high. Companies like Netflix use the services AWS.

C. Cost Advantage

Amazon has a model called Marketplace Model which is responsible for its advantage over the costing. The revenues that are generated over here has different means. Like It charges high commission and at the same time it provides various types of offers to the buyer so that the online purchasing gets increased day by day and the seller intends to sell its product more by this means. One other way of revenue is that Advertising, it shows the product of those sellers who have already paid it extra then the other one so their ranking is good and the traffic over them is always high.

Published By:
International Journal of Engineering Research in Current Trends

AWS Pricing Philosophy

Although the number and types of services provided by AWS have drastically expanded, our pricing policy has not changed: you just pay for the products you use. In the AWS pricing theory, the main tenets are:

- Pay as you go
- Pay less when you reserve
- Pay even less per unit by using more
- Pay even less as AWS grows
- Custom pricing

Companies like Walmart founded by Sam Wilton in the year of 1962 has given the priority to the profit rather than growth which is just opposite to the philosophy of the working engine in Amazon when after the internet age has arrived in order to compete with the veterans Walmart has invested $1.5billion to the technical infrastructure (Manuel Rivera et.al *;2015*) [9].

competition will now get tougher because other than the initial predictions at the dawn of the internet era, and mortar stores are far from current market realities (Doherty and Ellis-Chadwick, 2010) [10].

Data Mining by Amazon: A great strategy of Marketing

Having thousands of online customers throughout the world, Amazon has become the greatest store offering amazing products and services. They have a giant customer database and they are using this data to build strong relationships with their new as well as old connected customers. By analysing and comprising They plan their effective campaigns about promotions and goods with valuable consumer knowledge. The idea of data mining from the supply chain to marketing operations has been implemented. (Dholakia,2013) [1]. Not only for this purpose, but Amazon still uses data mining in multiple ways to sell its products in order to have a strategic edge from its diverse rivals. Because of increased social media interventions, consumers want personalization from the firms they buy goods from mainly online businesses. The consumer shopping background at amazon lets them classify customers choices and preferences (Clufia, Bunzel, Snuggs.2014) [2].

In 2004 Amazon.com, Inc entered with all its services in China but there is a high domination of China's native company Alibaba, Amazon cut its root from China and entered solely in India around 2005 with all its products and services into the second largest densely populated country which proved to be the right for right philosophy resulted in online shopping increased exponentially from

$35million in 2014 to $15 billion in the year of 2016 (Anubha Vashisht et. al ;2017) [3]

Ideology

Once Bezos explained his strategy "If everything you do needs to work on a three-year time horizon, then you're competing against a lot of people. But if you're willing to invest on a seven-year time horizon, you're now competing against a fraction of those people… Just by lengthening the time horizon, you can engage in endeavours that you could never otherwise pursue". [1]. Analysis of sentiment or opinion mining is a field of research that analyses the feelings, attitudes, or emotions of people against certain individuals (Xing Fang et. al, 2015) [15].

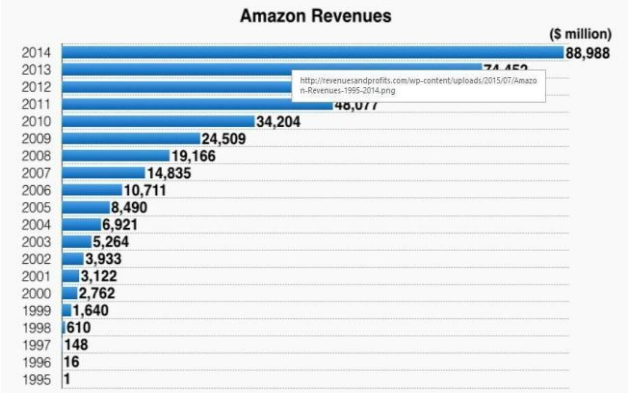

Fig 2: Revenue Graph of Amazon.com from 1996 to 2014 $ in million.

II. LITERATURE SURVEY

Dholakia et. al [1] emphasised on "the design and powerful strategies regarding promotions and products. It has incorporated the concept of data mining from supply chain to marketing operations, Jess Ciufia et. al [2] derived conclusions "data mining may be the unwelcome snoop on our consumption habits, but for now, we can rest easy that this spying will more likely help us rather than hurt us", Anubha Vashisht et. al [3] analysed "growth rate on internet users is highest, around 6M users are joining every month. Online shopping market in India hiked $35M to $15B in 2014-2016", Nanda Kumar et. Al [4] emphasised on how the Recommendations and consumer reviews were acknowledged as features of a business-to-consumer website significantly. PA Dabholkar et al [5] concluded "consumer choice of Rating Web Sites as online, interactive, decision aids, that credibility of the Rating Web Site is the most important attribute, followed by the opportunity for customization of information". Zhenhui Jiang et. al [6] evaluated "functional mechanisms (namely, vividness and interactivity) influence consumers' intentions to return to a website and their intentions to purchase products". Bernard Marr [7] derived the conclusion "put something that someone might like in front of them and they may well be overcome by a burning desire to buy it – regardless of whether or not it will fulfil any real need." Ankush Verma et. al [8] concluded that "Amazon clusters have several servers which are working together to perform different tasks on distributed databases on different servers parallelly. Amazon services are used in

big data analysis to boost business efficiency". Manuel Rivera et. al [9] analysed "Companies like Walmart founded by Sam Wilton in the year of 1962 has given priority to the profit rather than growth". Doherty et. Al evaluated that competition will now get tougher because other than the initial predictions at the dawn of the internet era".

III. TOOLS FOR DATA ANALYSIS

The growing demand and importance of data analysis in the market have generated many openings worldwide. The standard open-source data analysis tools are more popular, user-friendly and performance oriented than the paid version. There are several open-source applications that do not need much coding and are able to achieve greater outcomes than paid ones, e.g., R programming in data mining and Tableau, Python in data visualization. Here are a

few standard tools for data analysis that can be open source as well as paid tools, but are more useful for data analysis and visualization of data.

1) R-Programming
2) Tableau Public
3) SAS
4) Microsoft Excel
5) Python :- It is recommended because of its vast libraries like pandas for data preparation, matplot & seaborn for visualisation, scikit-learn for machine learning and pyspark for big data (I. Stanč in et. al, 2019) [11].

There are few several tools which are as used as these tools are used but in term of complexity and dataset management these tools are found to be less powerful. Here is the list of the following: -

I. Splunk
II. QlikView
III. Knime
IV. Apache Spark
V. Rapidminer

IV. DATASET ANALYSIS

Some important descriptions of the dataset through code are the following are some of the essential functions which must be used after importing the dataset and libraries as well. This makes pretty nice ideology about the dataset and sets the path for further analysis in which direction the operation goes. The review count has gradually increased over the years (Monika Mishra, 2019) [20].

Ratings help in making decisions on buying products personalized, Product level ratings gives the generic results where feature level rating provides particular results (K. R. Jerripothula et. al, 2020)[12]. It would be much easier to go through hundreds of reviews on this vibrant machine learning day if a model were used to
polarize these reviews and learn from them. (Tanjim Ul Haque et.al, 2018) [18].

For the latest product customers' opinion available can be in the thousands. It gets tough for the customers to read all the reviews and if he reads only a few of those reviews, then he might get a biased view about it. The feature-based summarization systems implemented are more generic as well as static in behaviour (Kushal Bafna, et. al)[13]

Here is the list of few of them: -

A. df.shape()

This is used for getting the information that how many rows and columns are there present in the dataset and can be used meanwhile the analysis so that either the rows and columns are added or dropped in the complete analysis. The provided dataset has- (792, 9)

B. df.info()

It gives the information of the provided dataset that is the class of dataset, number of columns present, its datatype, its memory usage. This information function can be called any time in the total process of analysis.

Data columns (total 9 columns):

asin	792 non-null object
brand	792 non-null object
title	792 non-null object
url	792 non-null object
image	792 non-null object
rating	792 non-null float64 reviewUrl
	792 non-null object totalReviews
	792 non-null int64
prices	577 non-null object

C. df.describe()

It tells a lot of things about the provided dataset. It gives the overall idea about the dataset its mean values, min, max, std, its total count and so on.

	rating	totalReviews
count	792.000000	792.000000
mean	3.607576	104.231061
std	0.668730	166.242503
min	1.000000	1.000000
25%	3.200000	7.000000
50%	3.700000	31.500000
75%	4.000000	122.250000
Max	5.000000	984.000000

D. df.head()

It gives only the first five rows information index starting from 0 to 4.This is basically used to check whether the data is loaded in variable or not.

E. df.tail()

It gives the last five rows information to the programmer by showing the bottom five rows used to check whether it loaded up to the last or not.

F. df.isna()

This gives the idea where the values are not present in the dataset so4 that is get removed by some algorithmic code. The particular function that is used in this is **df.isna().count()** for the total count on the dataset. asin

	792
brand	792
title	792
url	792
image	792
rating	792
reviewUrl	792
totalReviews	792
prices	792

G. sns.heatmap()

This function is used to detect where the actual missing values are and it checks throughout the dataset each column of every rows to give the detail. For this there is another function which is passed as an argument in this and that is sns.heatmap(df.isnull(), yticklabels=False, cbar=False, cmap='viridis'),here df.isnull() is an argument function,yticklabels are false so that nothing is written in y-axis.mini yellow bars show the places where values are missing

Published By:
International Journal of Engineering
Research in Current Trends

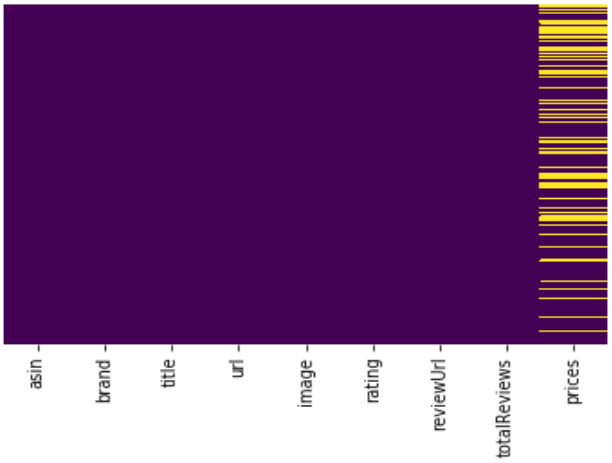

Fig 3: Heatmap showing gaps in values of prices another function is passed in sns.heatmap for the information about the correlation or interdependency between the variables of the dataset and that is **sns.heatmap(df.corr()).**

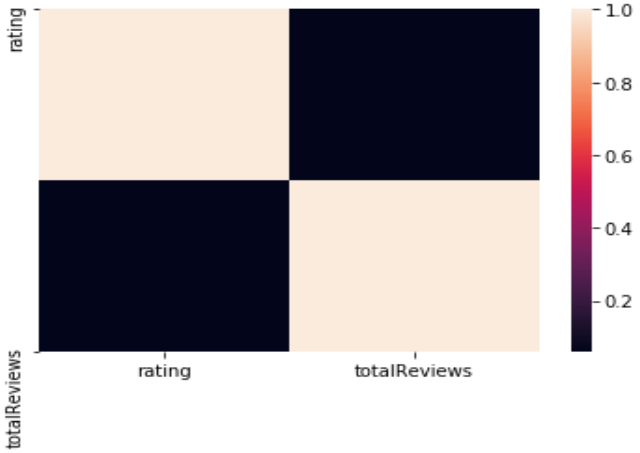

Fig 4: correlated columns of graph

Here, the two columns of the given dataset are completely numerical so the operation is performed only over the two columns named as rating and total Reviews.
one of the most important function is used that is **sns.pairplot(df)** it gives us two different type of graph one is bar graph and another one is graph using dot

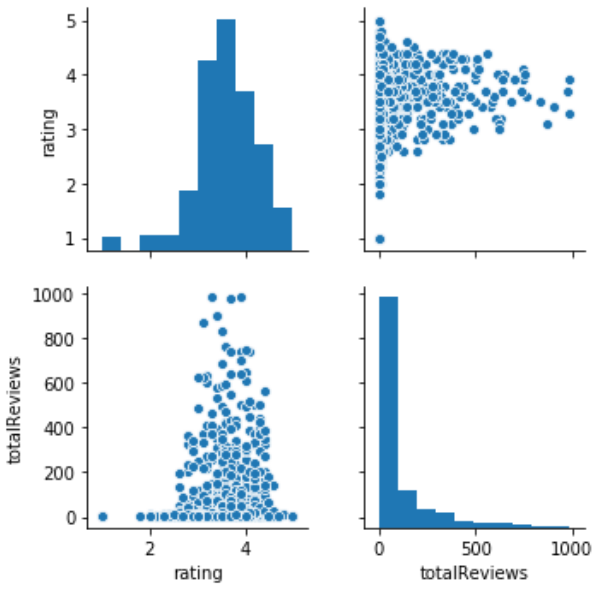

Fig 5: pairplots between the columns of dataset

These are the graph related functions which are used to describe the dataset in a sequence manner that provides the basic information and decides the direction of analysis, where to go. There are several more functions which are used for various descriptive information analysis.
The most important Libraries required here are matplotlib.pyplot as well as seaborn. These are widely used in all types of graphical stated analysis.
Replacing all the missing values of dataset
Now let us see some of the analysis done over the provided dataset. As we know that the dataset contains 792 rows and 9 columns which is detected by df.shape where few of the entries has missing value so by using NumPy library the missing values are filled by taking the medians of all the presented values in that particular columns. There is a unique function which is called for this kind of work that is impute_median
#here is the sort of code def impute_median(series):
return series.fillna(series.medians())

dataframe.column_name=dataframe['column_name'].tran
sform(impute_medians)
now this is used in name of any particular variables and all the missing values are filled at once. All the values are updated for detecting it we can further call the function dataframe.isnull().sum()
Dataset analysis Unique values
In the given dataset if to find all the unique values of that particular column there is a function which is used so
that all the unique values are extracted and their total count can also be measured.
print(df["brand"].nunique())
df["brand"].unique() 10
array *(['Nokia', 'Motorola', 'Sony', 'Samsung',*
'HUAWEI', 'Apple','OnePlus', 'Google', 'ASUS', 'Xiaomi'])
Here, In the provided dataset which is kept in a variable called df. The total unique numbers of brand and the total count of it is given by using unique() and nunique() function. The dtype is also been shown by this function.
Boxplot Function
there is a function which is used to get the overall idea of maximum traffic of that column. For example, if one has to check what set of people gives more reviews to the mobile phone. It can be checked by using
df.boxplot("totalReviews")

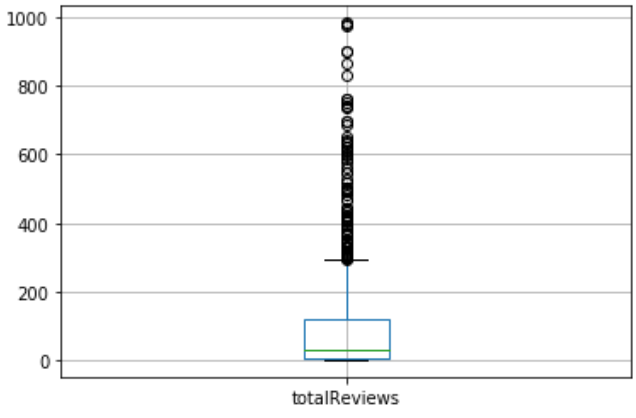

Fig 6: boxplot of total reviews
It means that at every movie set there are mostly 300-680 people who gave their reviews by using boxplot() function it

can be done for any other column of the dataset variable.
Number of phones having same price
In the given dataset there are a total 792 rows means there are a total 792 different mobiles set to represent this is from the dataset Now If one has to find that the total number of datasets that has similar price to another handset then the function is applied here.
df["prices"].nunique()
which gives the result 447 means that 345 mobile sets have the same price, they have the matching price but 447 mobile sets are there present who have a unique price to buy.
Finding the maxima and minima in the dataset variable
To check the maximum of any column like the stuff which has maximum rating or the mobile phone which has maximum number of total reviews over it so that the phone is awarded by user best choice this is done by max () function.
print(df["brand"][df["totalReviews"]
= df["totalReviews"].max()].nunique())
print(print(df["brand"][df["totalReviews"]
= df["totalReviews"].max()].unique()))
1
['Google']
here the output is one that is only one brand has max. rating which is presented in the array and value is Google means Google mobile set has maximum reviews over it
If to check which brand has the maximum rating then this can also be done over here by using max() function.
print(df["brand"][df["rating"] == df["rating"].max()].nunique())
print(print(df["brand"][df["rating"] == df["rating"].max()].unique()))
now the answer is:- 7
['Samsung' 'Motorola' 'Google' 'Sony' 'Xiaomi' 'Apple' 'ASUS']
There is total 7 different mobile sets which has maximum number of ratings and are presented in the sort of array. Here one can make conclusion over it that which brand he should prefer to buy a mobile set over amazon.com

If to check minima of the similar condition It can be done similarly by using min() function for example to check the brand which has lowest Total reviews than this could be done by
print(df["brand"][df["totalReviews"] == df["totalReviews"].min()].nunique())
print(print(df["brand"][df["totalReviews"] == df["totalReviews"].min()].unique()))
['Sony']
Here the Sony brand has minimum reviews by people who buy it. Similarly, we can also check the rating of mobile brands over the amazon this can also be done similarly.
print(df["brand"][df["rating"] == df["rating"].min()].nunique())
print(print(df["brand"][df["rating"] == df["rating"].min()].unique())) 4
['Motorola' 'Samsung' 'Apple' 'OnePlus']
There are 4 different brands who fall in the less categories zone here and their names are present in the array. Now these types of analysis results can be used by the new user to make a good and satisfactory decision about which type of mobile set he should buy.
one more analysis that can be done over here for pricing purpose that
df[["brand",prices"]][df["totalReviews"]=df["totalReviews"].max()]

	brand	prices
352	Google	$107.70

Amazon is a big data tech, that is why it wants to look at the company in his second post of his series on how specific organizations use big data. Since we all know Amazon was a leader in e-commerce in so many different areas, but maybe one of the greatest advances was the Individual Reviews Framework – based on big data, so it collects from its millions of consumer purchases and much more. Psychologists often say of the influence of suggestion—"put anything that anyone may want in front of them, and they could well be overwhelmed with a raging impulse to purchase it—regardless of whether or not it fulfils some specific need."(Bernard Marr,22015) []
Creating and dropping the data variables over different analysis
As per the analysis we can drop the unwanted rows of the dataset by a particular function as well as we can add new column to the dataset regarding the further analysis of the dataset let us take an example by creating a new column in the given dataset that is rating by review column which tells the ratio that number of rating done per total reviews of a mobile set and for this the required function is
df["rating/review"] = df["rating"]/df["totalReviews"]
now by to check it we can see the first five dataset rows and column or by using df.shape we can check this now by using df.shape the answer is (792,10)

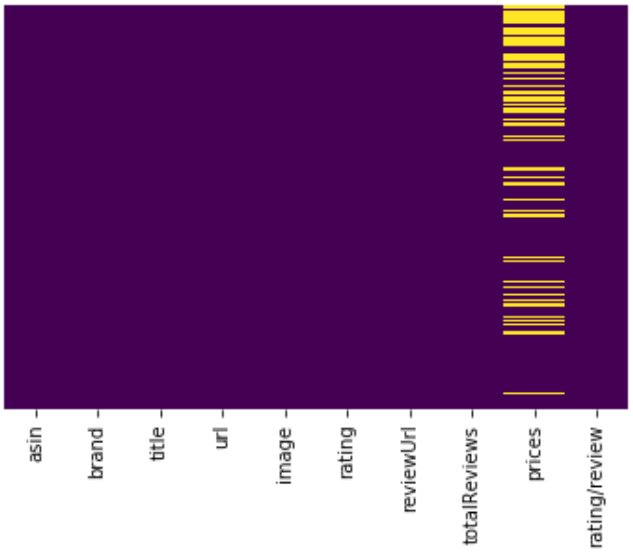

Fig 7: heatmap showing no null values of new column (last column)
Here this heatmap shows the null value in yellow colours like prices shows the gap in the dataset column now rating/review has no yellow bars this depicts that there is no blank value for the newly created column this is proved to be true because no value is missing from the columns rating and reviews.
Now let us see another analysis done by using the newly

Published By:
International Journal of Engineering Research in Current Trends

created column in the provided dataset that
df["brand"][df["rating/review"]== df["rating/review"].max()].unique()
array (['Samsung', 'Sony', 'Xiaomi', 'ASUS', 'Motorola', 'Google'])
This is the classification of brands that which
brand has maximum rating per review and the answer is shown in the array above. This subsection will introduce new independent variables to take into account such as the reviewer and product information as well as the bag of opinion model. Li, Liu, Jin, Zhao, Yang and Zhu (2011) bring a new perspective to the review rating prediction task.
Graphs and their description The groupby() function
This is a very special function which is used in plotting the graph under various conditions in which the data is distributed among the various rows and columns of the dataset now, this function collects the data over the given parameter and grouped it to one single group and count is as well. For example, in the given dataset there
are several mobile sets of Apples and the rating over it is distributed whole over the dataset now for detecting the total number of reviews done over the dataset group by function is called and passed Brand as parameter and saved under a new data variable.
df1 = df.groupby(by='brand').count() df1

TABLE 1(A): - Details of Columns Brand wise

Brand	**title**	**url**	**image**	**rating**
ASUS	13	13	13	13
Apple	101	101	101	101
Google	33	33	33	33
Huawei	36	36	36	36
Motorola	100	100	100	100
Nokia	49	49	49	49
OnePlus	7	7	7	7
Samsung	397	397	397	397
Sony	29	29	29	29
Xiaomi	27	27	27	27

TABLE 1(B): -Details of Last 4 Columns Brand wise

Brand	**reviewUrl**	**prices**	**rating/review**
ASUS	13	11	13
Apple	101	94	101
Google	33	26	33
HUAWEI	36	29	36
Motorola	100	69	100
Nokia	49	31	49
OnePlus	7	5	7
Samsung	397	264	397
Sony	29	21	29
Xiaomi	27	27	27

To understand this let's view the row of Samsung Brand in which there are total 397 different title, images, rating, total review, review url, but 264 different values of prices all other are same so this is the work of groupby function. It operates on all kind of datatypes whether it is int, float or string.
By analysing the dataset over various kinds of parameter there are several of graphs which are formed by using a special library called matplotlib and particularly its sub-library i.e. pyplot together called as
matplotlib.pyplot called as plt basically it depicts the knowledge or information about the goodness of dataset analysing and gives a very clear view of the dataset and there are various different conclusion based upon this and these result affects the real-time working of number of different things. There are so many decisions made over these dataset because using this library data visualisation is done completely and this visualisation gives a clear fact and Figure which is used after and after whenever it is needed.
By the help of these graphs and data visualisation several multinational companies sell as well as buys different products to whether the seller is a client or another companies. The graphs may be in the form of Bar graph, Histogram, Pie Charts, Dot graphs and so on.
Now there are some graph plotted over the provided dataset and these graphs gives a bundle of information about the dataset and data visualisation is completely done. There can be some conclusion made over these graphs result similar work is done in real time environment. These graphs may be Bar graph or the another one based upon the column or data variable of dataset.
Now if there is a query that which brand has a maximum number of reviews and detect its count also.
df1 = df.groupby(by='brand').count()
df1["totalReviews"].plot(kind='bar') plt.title("Graph On Brand vs total reviews") plt.ylabel("total reviews")
plt.xlabel("Brand")
plt.show()

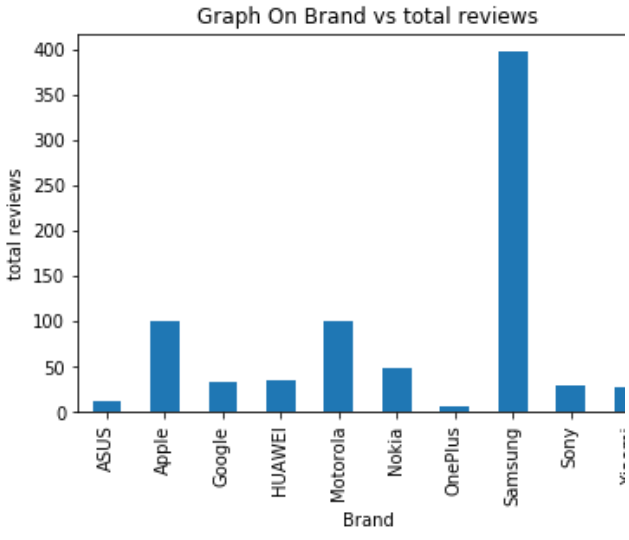

Fig 8: bar graph on Brand vs total reviews

This graph is plotted between Brand and total reviews where brand is in x-axis and total review is in y-axis the bar in the graph shows the relative position of the x and y axis. The title of the graph is "Graph on Brand vs total reviews " which gives a clear idea about so many things. The colour of the bar is blue which can be changed under plt.plot().
Here It is clear after this graph that most of the people reviewed their Samsung Brand mobile phone very large in number in comparison to the other Brands. By the help of this graph, it is clearly concluded that Samsung, Apple and Motorola are the top 3 brands in terms of review and
the Brand which falls in bottom 3 in terms of review are ASUS, One Plus and Sony. The gap between Samsung brand to any other brand is very large in comparison to others.
Brands like Apple and Motorola are in competition in terms of total reviews and so is for Sony and Xiaomi. Nokia is little above in total reviews then Huawei and Google.
It can be a crucial way to get ahead in the competitive e-commerce marketplace to understand the significance of feedback and how to exploit them to improve your brand, positioning yourself miles ahead of the competition (Leigh-Anne Truitt)[16].
Another Query which gives the graph between Rating and Total Reviews.

```
sns.lineplot(x=df["rating"], y=df["totalReviews"])
plt.title("Graph On Rating vs total reviews")
plt.ylabel("total reviews")
plt.xlabel("Ratings") plt.show()
```

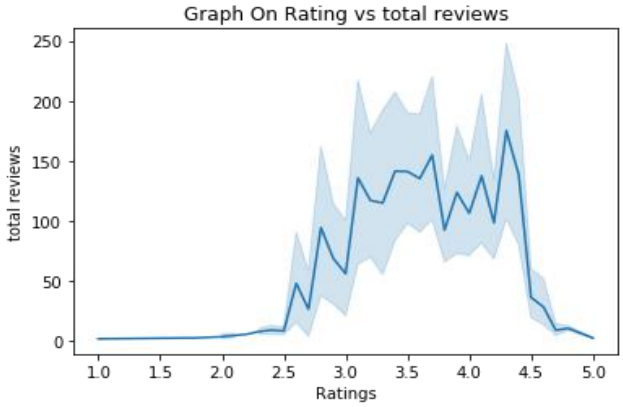

Fig 9: Line graph between rating vs total reviews

This graph is very important and inclusive in context
to the analysis of data of the dataset that is provided. This graph is plotted between Rating and total Review which does not has any relation with brand and all the other column of the given dataset. Titled as versus graph between rating and total review. X-axis denotes different Rating and Y-axis denotes different count of reviews. This is an interactive graph which is under the Axes Subplot.
The conclusion made through this graph is: -

a. Most of the reviews are given to the mobiles whose rating is in between 2.5 to 5.0.
b. Highest reviews is given to the mobile whose rating is 4.4 no matter the mobile set is of Asus, Samsung, apple or any of the brand.
c. The graph is drastically downed from 4.4 to 4.5 means that the mobile set whose rating is 4.5 is very less reviewed.
d. Reviews of mobile rating 2.5 and 4.4 are the same in number.
e. Reviews of mobile rating between 0 to 2.5 is near about zero.
f. Top three most reviewed mobile sets are 4.4, 3.6 and 2.8. These are the ratings of mobile sets on which the maximum number of reviews are written by the people.

Now based upon these results one can predict the further work related to it. Various different decisions should be made upon this so that further upcoming results are beneficial.
To check the average value of ratings and reviews of the Brand we use groupby(by=' '). mean() this gives the average or mean value of the numeric columns of the given dataset.

```
df1 = df.groupby(by='brand').mean() df1
```

brand	rating	totalReviews	rating/review
ASUS	3.776923	38.769231	0.778401
Apple	3.527723	118.039604	0.444820
Google	3.763636	122.090909	0.354766
HUAWEI	4.019444	82.555556	0.755683
Motorola	3.528000	88.150000	0.571945
Nokia	3.322449	117.428571	0.233329
OnePlus	3.342857	80.428571	0.270446
Samsung	3.573300	104.937028	0.548093
Sony	3.731034	116.689655	0.625766
Xiaomi	4.337037	109.185185	0.353336

from all the rows and column of the dataset all the separated brand is grouped in their particular group and the average value of rating total rating and rating/review is calculated throughout the dataset and various conclusion can be made through this table. There can be a separate Graph plotted to each of the column and data visualisation can be done more clearly.
Here is the graph plotted over the different columns of the dataset giving the clear visualisation: -
graph between Brands and there Average Rating

```
df1["rating"].plot(kind='bar') plt.xlabel("Brands")
plt.ylabel("Average Ratings ") plt.title("Brands vs Average Ratings ") plt.show()
```

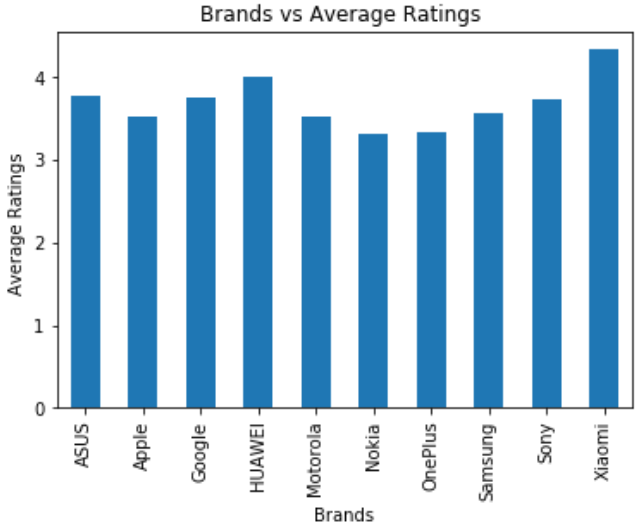

Fig10: Bar graph on brand vs average ratings Brands

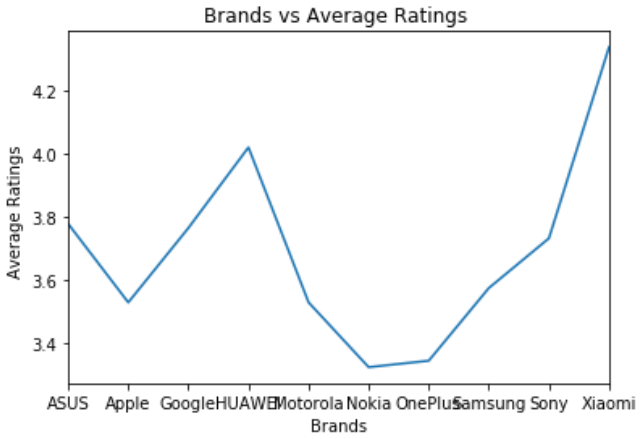

Fig 11: Line graph on brand vs average ratings

This is the graph which clearly tells that which mobile set Brand on amazon has how much average Rating. In this it is shown that Xiaomi Brand has maximum average rating throughout all the other brands present in the dataset. The top three brands who are Xiaomi, Huawei and Asus.
graph between Brands and their Average Total Reviews
df1["totalReviews"].plot(kind='bar')
plt.xlabel("Brands")1
plt.ylabel("Average Total reviews ")
plt.title("Average Total Review vs Brands ")
plt.show()

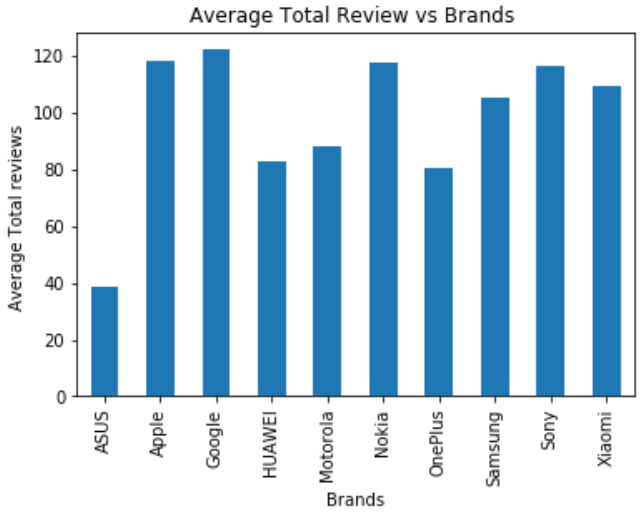

Fig 12: Bar graph of Avg Total reviews vs Brands

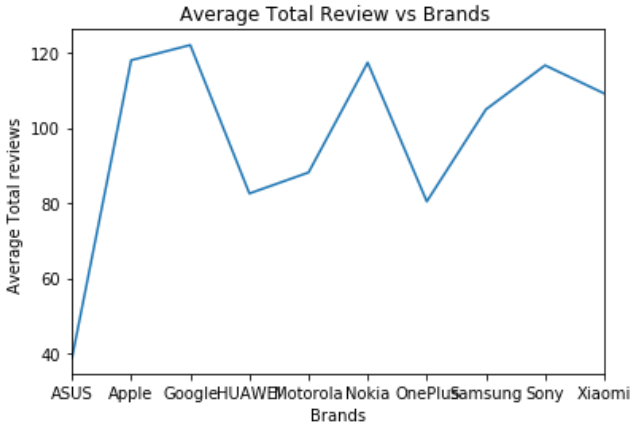

Fig 13: Line graph of Avg Total reviews vs Brands

This is the graph plotted between different Brands and average or mean value of Total Review in which Brand is in X-axis and average of total Reviews is on Y-axis. Here from both the graph it is clearly visualised that the Brands like Google, Apple and Nokia has the maximum reviews wrote on it by its particular user.

df1["rating/review"].plot(kind='bar')
plt.xlabel("Brands")
plt.ylabel("Average Total rating/review ")
plt.title("Average Total rating/Review vs Brands ")
plt.show()

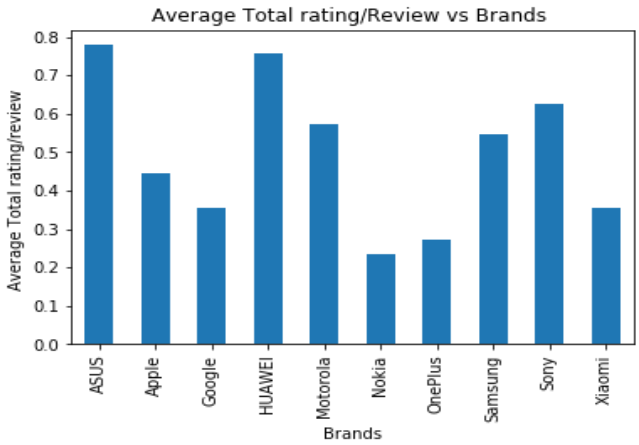

Fig:-14 Bar graph of Avg Total reviews/reviews vs Brands

Fig 15: Line graph of Avg Total reviews/reviews vs Brands

these two graphs are the data visualisation in which Average rating/review is associated this is combined form of the above two graph which gives the clear visualisation over the various parameters of the dataset.
The customer reviews present on a website has been shown to improvise customer thought of the usefulness and social presence of the website (Kumar and Benbasat 2006) [4]. Reviews have the power to attract consumer online visits, increase the time spent on the site (cause more traffic on the site), and create an environment of community among frequent shoppers. However, as the available capacity of customer reviews, the strategic tendency shifts from the mere activity of customer reviews to the customer evaluation and review uses. Online marketers should get the right opportunity to offer online information that consumers believe is relevant, and places like eOpinions and Amazon.com can post detailed recommendations for writing reviews. Giving a better option more conveniently is the key explanation why users use the reviews website. (Dabholkar

2006) [5], and the perceived diagnostic of website information positively affects consumers' attitudes toward shopping online (Jiang and Benbasat 2007) [6].

sns.heatmap(df.corr())

Once again, the heatmap is used to describe the interrelation between the columns of the dataset.

This gives the inter-relation between the three numeric columns of the given dataset. This shows how each of the columns of the dataset which takes part in the dataset analysis are interdependent to each other.

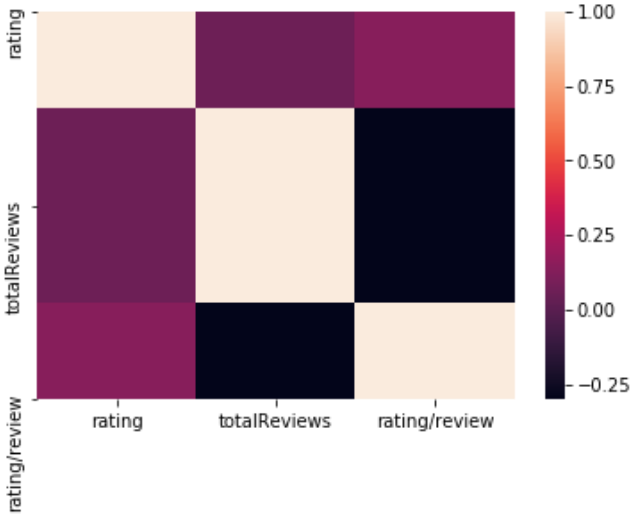

Fig 16: Correlated Graph between 3 columns of dataset

sns.pairplot(df)

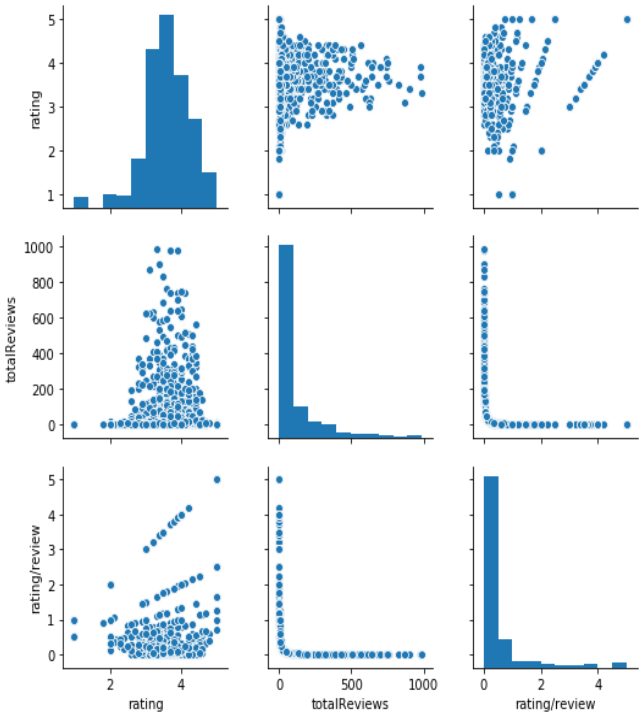

Fig 17(a): pairplot of the newly created dependent columns

This pairplot graph is plotted between the rating, reviews and total rating/reviews

sns.pairplot(df1)

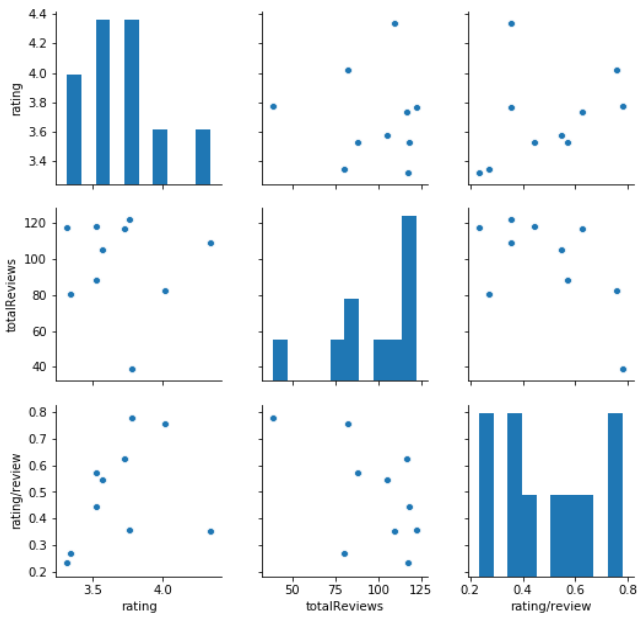

Fig 17(b): pairplot of the newly created avg of dependent columns

This graph is plotted between the average rating, average total review and average rating/totalreview

By this data visualisation one can clearly make several decisions which are proved to be beneficial for the firm as well. With the help of graph one can easily detect the cons and flaws of different mobile set marketing brands and so many further analyses can be done by meaningful use of this data analysis.

CONCLUSIONS

Amazon Business Model is one of the unique business models that plays a key role in online shopping among everyone. We can not ignore the strength of model when it comes to accessibility, wide product range and the purchasing capacity of the people or cost effectiveness feature.

In this paper we have identified various parameters like brand, image, totalReviews, prices based on these factors after analysis using Python, we found that in the period of 2005 to 2020 the data usage hiked up to 40,200 exabytes.

Following are the conclusions drawn from the above graphs:-

A. Mostly 300-680 people gave reviews to the mobile phones as shown in fig 6.
B. Most of the people reviewed Samsung Brand phones, very large in number as compared to the other brands.
C. Most of the reviews are given to the mobiles whose rating is in between 2.5 to 5.0.
D. Highest reviews are given to the mobile whose rating is 4.4 no matter the mobile set is of ASUS, Samsung, apple or any of the brand.
E. The graph is drastically downed from 4.4 to 4.5 means that the mobile set whose rating is 4.5 is very less reviewed.
F. Reviews of mobile rating 2.5 and 4.4 are the same in number.
G. Reviews of mobile rating between 0 to 2.5 is near about zero.

H. Top three most reviewed mobile sets are 4.4, 3.6 and 2.8. These are the ratings of mobile sets on which the maximum number of reviews are written by the people.

REFERENCES

[1] Dholakia, R., & Dholakia, N. ,"Scholarly Research in Marketing : Trends and Challenges in the Era of Big Data"(2004), available at:- https://web.uri.edu/business/files/Encycl-Communication-Data Mining-n-Marketing-.pdf

[2] J Clufia , D Bunzel and S Scruggs,"Consumer Marketing digging too deep with Data Mining and Digital wave", 2014, DOI: - http://digitalmediaix.com/consumer-marketers-digging-too-deep-w ith-data-mining/#.YBkhKugzZPY

[3] Wadhwa, Bharti, Anubha Vashisht, and Nidhi Phutela. "Business model of amazon India-A case study." *South Asian Journal of Marketing & Management Research* 10.1 (2020): 32-40.

[4] Kumar, Nanda, and Izak Benbasat. "Research note: the influence of recommendations and consumer reviews on evaluations of websites." *Information Systems Research* 17.4 (2006): 425-439.

[5] Dabholkar, P. A. (2006). Factors influencing consumer choice of a" rating Website": An experimental investigation of an online interactive decision aid. Journal of Marketing Theory and Practice, 14(4), 259-273.

[6] Zhenhui Jiang, Izak Benbasat,"Research Note—Investigating the Influence of the Functional Mechanisms of Online Product Presentations" 2007, Information Systems Research, Vol. 18, No. 4,published online, available at:- https://doi.org/10.1287/isre.1070.0124

[7] Marr, B. "Big Data case study collection." (2015)

[8] A. Verma, N. Sethi and N. Jai, "Beyond Hadoop for e-commerce Big Data Analysis through Amazon," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), Bhopal, India, 2018, pp. 1-4, DOI: 10.1109/ICACAT.2018.8933660.

[9] Manuel Rivera, Yunfeng Pi, Samuel Gray, "The Economist Case Study Contest Amazon vs Walmart A Case for Proven Reliability Over Unbridled Enthusiasm", 2015 University of Portland available at:https://www.economist.com/sites/default/files/universityofportla nd_ws.pdf

[10] Doherty, Neil & Ellis-Chadwick, Fiona, "Internet retailing: The past, the present and the future," (2010) International Journal of Retail &amp Distribution Management. 38. DOI: - 0.1108/09590551011086000.

[11] I. Stančin and A. Jović, "An overview and comparison of free Python libraries for data mining and big data analysis," 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp. 977-982, doi: 10.23919/MIPRO.2019.8757088.

[12] K. R. Jerripothula, A. Rai, K. Garg and Y. S. Rautela, "Feature-Level Rating System Using Customer Reviews and Review Votes," in IEEE Transactions on Computational Social Systems, vol. 7, no. 5, pp. 1210-1219, Oct. 2020, doi: 10.1109/TCSS.2020.3010807.

[13] Kushal Bafna, Durga Toshniwal, "Feature based Summarization of Customers' Reviews of Online Products," 2013, Procedia Computer Science, Volume 22, Pages 142-151, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2013.09.090.(https://www.sciencedi rect.com/science/article/pii/S1877050913008831).

[14] D.R. Kumar Raja, S. Pushpa, "Feature level review table generation for E-Commerce websites to produce qualitative rating of the products," 2017, Future Computing and Informatics Journal, Volume 2, Issue 2, Pages 118-124, ISSN 2314-7288, https://doi.org/10.1016/j.fcij.2017.09.002,(https://www.sciencedire ct.com/science/article/pii/S2314728817300326)

[15] Fang, X., Zhan, J. ,"Sentiment analysis using product review data," Journal of Big Data 2, 5 (2015). https://doi.org/10.1186/s40537-015-0015-2

[16] Leigh-Anne Truitt ,"he Inside Scoop on Ecommerce Reviews: Why They Matter and How to Make the Most of Them", available:- https://www.bigcommerce.com/blog/online-reviews/#who-is-rea ding-online-reviews

[17] Miguel J.C., Casado M.Á.," GAFAnomy (Google, Amazon, Facebook and Apple): The Big Four and the b-Ecosystem." In: Gómez-Uranga M., Zabala-Iturriagagoitia J., Barrutia J. (eds),(2016), Dynamics of Big Internet Industry Groups and Future https://doi.org/10.1007/978-3-319-31147-0_4

[18] T. U. Haque, N. N. Saber and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," 2018 IEEE International Conference on Innovative Research and Development (ICIRD), Bangkok, 2018, pp. 1-6, DOI: 10.1109/ICIRD.2018.8376299.

[19] J. Zhang and M. L. Huang, "5Ws Model for Big Data Analysis and Visualization," 2013 IEEE 16th International Conference on Computational Science and Engineering, Sydney, NSW, 2013, pp. 1021-1028, DOI: 10.1109/CSE.2013.149.

[20] Mishra, M., Chopde, J., Shah, M., Parikh, P., Babu, R. C., & Woo,J. (2019). Big data predictive analysis of amazon product review. In KSII the 14th Asia Pacific International Conference on Information Science and Technology, APIC-IST, KSII, Beijing, China.

[21] Jeff Bezos Quote at AZ Quotes available at: - (Source:- https://www.azquotes.com/quote/698930).