



REVA
UNIVERSITY

Bengaluru, India

A Project Report on
Customer Base Analytics
In E-Commerce

Submitted in Partial Fulfilment for Award of Degree of
Master of Business Administration
In Business Analytics

Submitted By
Sanjeev Kumar Jha
R19MBA06

Under the Guidance of
Dr. J. B. Simha
CTO, Abiba Technologies/Chief Mentor, RACE

REVA Academy for Corporate Excellence - RACE
REVA University
Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064
race.reva.edu.in

September, 2020



Candidate's Declaration

I, Sanjeev Kumar Jha hereby declare that I have completed the project work towards the Master in Business Administration at, REVA University on the topic entitled **Customer Base Analytics In E-Commerce** under the supervision of **Dr. J. B. Simha**, Chief Mentor. This report embodies the original work done by me in partial fulfilment of the requirements for the award of degree for the academic year 2020.

Place: Bengaluru

Name of the Student: Sanjeev Kumar Jha

Date: 12th Dec 2020

Signature of Student



Certificate

This is to Certify that the PROJECT work entitled **Customer Base Analytics in E-Commerce** carried out by **Sanjeev Kumar Jha** with SRN **R19MBA06**, is a bonafide student of REVA University, is submitting the project report in fulfilment for the award of Master in Business Administration in Business Analytics during the academic year 2020. The Project report has been tested for plagiarism, and has passed the plagiarism test with the similarity score less than 15%. The project report has been approved as it satisfies the academic requirements in respect of Customer Base Analytics in E-Commerce work prescribed for the said Degree.

Signature of the Guide
Dr. J. B. Simha
Guide

Signature of the Director
Dr. Shinu Abhi
Director

External Viva

Names of the Examiners

1. Indrajit Kar, Head of AI-Chief Architect & Data Scientists, Siemens
2. Pradeepta Mishra, Associate Principal & Head of AI, LTI-Larsen and Toubro

Place: Bengaluru

Date: 12th Dec 2020



Acknowledgement

Please acknowledge the role of your mentors, trainers, classmates, program office members, family and friends who have directly and indirectly supported you in this work.

Please acknowledge the support provided by Hon'ble Chancellor, Dr. P Shayma Raju, Vice Chancellor, Dr. K. Mallikharjuna Babu, and Registrar, Dr. M. Dhanamjaya, as a standard protocol.

Place: Bengaluru

Date: 12th Dec 2020



Similarity Index Report

This is to certify that this project report titled **Customer Base Analytics in E-Commerce** was scanned for similarity detection. Process and outcome is given below.

Software Used: **Turnitin**

Date of Report Generation: **12th Dec 2020**

Similarity Index in %: **5 %**

Total word count: **7903**

Name of the Guide: **Dr. J.B.Simha**

Place: Bengaluru

Date: 12th Dec 2020

Verified by: M N Dincy Dechamma

Name of the Student: Sanjeev Kumar

Jha Signature of Student

Signature

Dr. Shinu Abhi,

Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	RFM	Recency, Frequency, Monetary
2	RFMT	Recency, Frequency, Monetary, Tenure
3	K-Means	K-Means Clustering

List of Figures

No.	Name	Page No.
4.1	Customer Segmentation	15
5.1	Project Plan	16
5.2	Project Pipeline	17
8.1	Cleaning State Column	20
8.2	Removing negative value	21
8.3	Top 10 States Sales in Million	21
8.4	State wise Map	22
8.5	Top 10 cities Invoiced in Million	22
8.6	Top 10 cities gross margin in Million	23
8.7	Total gross margin Vs. Decile	23
8.8	Total Monetary Vs Decile	24
11.1	Decile Analysis	29
11.2	Average Monetary Vs Decile	30
11.3	Average Frequency Vs Decile	30
11.4	Average Recency Vs Decile	31
11.5	Customer Profile per Cluster	31
11.6	RFM Graph based on Clusters	32
11.7	Customer Base Analysis Result	32

Abstract

A world driven by the internet must give rise to a more convenient way of conducting business. This proved to be a reality when Amazon worked painstakingly to tap into a customer base with their online presence. This was the birth of the e-commerce industry, which has now grown into one of the biggest industries in the economy. E-commerce is efficient, cheaper, encourages impulsive shopping and retail therapy, and all this just a click away. It also provides a more global reach which physical stores or traditional businesses could never aspire for.

E-commerce has enabled even small businesses to market their products worldwide, strengthened by the force that is social media.

The lockdown proved fatal for many, forcing most to shut down their businesses. However, those that could afford it were able to adapt by creating an online presence. Since most firms are choosing e-commerce as their mode of business and transactions, it has created competition. People are now spoiled with choices which makes them disloyal customers.

This is there is a need for customer base analysis. E-commerce may make getting customers easier but the task of retaining the same customers is difficult. Customer base analytics in e-commerce is a topic which needs more attention in the current scenario. A study of the patterns of behaviour of a customer provides fodder for creativity and helps understand and formulate effective strategies to retain the same.

This research deals with the process of segmentation. Segmentation is one way of deriving customer base analytics by grouping those customers who share the same demography, taste, background among other things. Segmentation of customers and storing the necessary data helps organizations to be better at CRM - customer relations management. It enables a firm to live up to customer expectations, hence making it easier to retain the same.

To generate better results, an RFM approach has been used along with clustering to categorize customer data. For the creation of customer segments, a method has been derived from K-Means Clustering. The variational approximation is used, to estimate the given results from segmentation more efficiently. The proposed procedure leads to a generation of better results than the RFM and the research seeks to prove the same.

The findings of the research will underscore the importance of segmentation as a mode of customer base analytics and establish the difference between the data of customers and their privacy.

Patterns of the transaction, choice of products, product categories, wish lists: these form the core of customer base analysis. Consumer behaviour has always been of the utmost importance for organizations so it is only correct that one must have a database of the online customer base.

Keywords: RFM, Customer Segmentation, K-Means Clustering

Contents

Candidate's Declaration.....	2
Certificate.....	3
Acknowledgement	4
Similarity Index Report.....	5
List of Abbreviations	6
List of Figures	6
Abstract	7
Chapter 1: Introduction	9
Chapter 2: Literature Review	11
Chapter 3: Problem Statement	14
Chapter 4: Objectives of the Study	15
Chapter 5: Project Methodology	16
Chapter 6: Business Understanding	18
Chapter 7: Data Understanding.....	19
Chapter 8: Data Preparation.....	20
Chapter 9: Data Modeling.....	25
Chapter 9: Data Evaluation.....	27
Chapter 10: Deployment	28
Chapter 11: Analysis and Results	29
Chapter 12: Conclusions and Recommendations for future work	33
Appendix.....	37
Plagiarism Report.....	37
Publications in a Journal/Conference Presented/White Paper	38
Any Additional Details	38

Chapter 1: Introduction

1.1 Introduction:

The world has seen a considerable change in the last few years, owing to the internet revolution and the rising popularity of social media. This has propelled a series of changes in the manner of how business is conducted and the relations between a customer and a business. Online transactions make for a better option for everyone, and most have opted for it today, as India has seen a rise in the usage of the internet since Jio's launch. Social media has gained enough momentum to change the way marketing works, including social media marketing under the umbrella term of digital marketing, as opposed to the obsolete traditional modes.

Since the world is changing at a fast pace, it is advisable to keep up with it. The shifting changes brought about by the internet makes for difficulty in retaining a loyal customer base. While social media and the internet have become the vehicle through which an organization can easily get a global reach, it has also given rise to mercurial behaviour of existing customers. People are spoilt for choices and most conduct simple research before resorting to any transaction. This builds pressure on firms to hold on to the same customer base and the graph can just as easily blow up and have a steep fall. Electronic commerce has boosted competition where many are selling the same products or rendering the same service, but better and perhaps cheaper. It is important to be unique but just as important to have an insight into what drives customers.

This is where customer base analysis plays an important role for it helps to understand the patterns of customer behaviour. This includes the trends in shopping across multiple categories of products and services, tentative wish lists, the time of purchase and the choice of products. Analysis and research have gained importance since firms have launched themselves on online platforms.

The lockdown has catalysed the same, with sellers choosing to go online instead of shutting down businesses completely. In the last two years, e-commerce has expanded exponentially and the future holds bright promises for the same. Jeff Bezos, the founder of the e-commerce giant Amazon is today the richest man in the world and his wealth has only increased in the lockdown.

It is well known that e-commerce giants retain the data of customers, but to what extent does it stretch? Data mining is an important concept, one that is on the rise since it helps with analytics. Thus, have the smaller companies followed in the same footsteps, by using a host of resources to track customer's behaviour and data. The time between purchases is calculated as is the product categories of the customer's baskets. The world of internet has started running on cookies which makes similar content pop up everywhere, to imprint the idea on the minds of customers. This has been made possible due to customer base analysis in e-commerce and is helpful in lead generation.

Segmentation is one such process for customer base analytics in the field of e-commerce. It involves the grouping of similar customers in such a way that they fall under the same demographics, patterns and choices. The difference between the two groups is considerable and helps understand patterns and the factors that these patterns depend upon.

The particular technique of segmentation of the customer base has been popularized in the past years and led to developments in information and the business of electronic commerce. It has particularly helped with database management systems and data mining. Big data is easily available on the internet today and holds much value in the market. The inefficiency of traditional techniques that involve statistics on a large amount of data has forced researchers to adopt tools of segmentation that fare better and are more effective to pry into the information regarding customers.

KD or knowledge discovery and DM, that is, data mining has proved to be a possible solution. Marketing research analysts find these to be of singular interest, especially in the application of said technologies in problems related to marketing such as forecasting among others.

The aim of the research also involves finding out the efficiency of segmentation tools concerning future use. This is mostly done to find profitable quotients. The model RFM (recency, frequency and monetary) has been used by the majority as a tool of segmentation and has had quite a long history in traditional marketing. However, studies today show that the RFM model can be improved with the inclusion of such additional variables during the prediction of customer behaviour. It is found that firms have limitations in the use of the RFM model: a distinction between long term and short-term customers is rendered quite impossible.

The above-stated reason is why adding a T, that stands for tenure to the existing RFM model makes a world of difference. The tenure or time in e-commerce helps in the identification of valuable customers.

The number of clusters or groups must be determined before the usage of the RFMT model of consumer base analytics. One widely used to mean of clustering, also known as the K-Means of clustering is a widely used strategy that is used for the segmentation of customers when the RFM model is applied for lesser data. This can help managers with their specialization in marketing in the easy recognition and precision in customer segments while comparing market maps and monitoring responses of the market.

The data that has been used in the research comprises 8671 customers who are unique and who visited an e-commerce site. The dates for the same or the tenure is between the first of January, 2019 right up to the thirty first of December, 2019.

The given profiles are inclusive of the following: order numbers of customers, the dates of order, the city to which the customers belong, SKU Codes, Sizes, Colours, the Quantities ordered, Invoices, the Gross margin and the duration between the first visit to the website and the last known visit to the same along with the frequency of the visits.

Chapter 2: Literature Review

2.1 Segmentation

On an average, business to consumer (B2C) e-commerce is increasing by as much as 20% each year (2014), with this year bringing a pivotal change in how businesses are conducted. This has propelled firms to up their research analysis game since online business is a risky deal. There is no saying when it might rise beyond expectations or pummel down the graph. This is why companies are now investing in big data mining - this much has been established in the abstract and introduction of the research itself.

Big data can be referred to as the data that is amassed online by e-commerce companies in their transactions. It can also include offline data which is collected by surveillance cameras or CCTV cameras and traffic monitors. This research has to do with the collection of online data, that is, the big data that is mined from such sources as websites and search engines which give considerable insight as to the customer's identity, the kind of products they are most likely to buy among other things. This data is invaluable and provides fodder for research, as it is used by analysts to predict certain outcomes. Considering the constantly shifting customer base, it is imperative that analysts gain some perspective regarding future trends. Firms are now getting the better of their competitors by being ahead of each other in customer base analytics.

Segmentation is one such process that helps with the same and in this research, it will be seen how segmentation of customers help to achieve better results.

Any business depends on the quality of service rendered or the product delivered. This is dependent on the expectations of the customer or the wishes of the customer. It is often said that the "customer is king." It is in accordance with the similarities found in the choices of customers that helps to separate them into different groups. This segregation of customers into different groups is known as segmentation in customer base analysis. Segmentation is an easier way to identify the requirements of each and hence helps with easier calculations. In addition to this, Schneider also calls for market segmentation, wherein, in the same manner, potential customers are accommodated in one group.

The process of segmentation may seem like an "extra", and not essential. But it is just as important. This segregation helps to identify potential customers and also helps in the offering of choice, customized products and services, along with the identification of the customers that are the most profitable.

The data required for customer segregation can be availed from a host of different sources. This eclectic behaviour can be broken down into two main sources: external and internal. Internal data is the data that is availed from purchase history, customer registration on the website as well as details on the customer's profile. External data is more varied and can be obtained from market surveys and research, census, mass media, social media, cookies, search options and web history.

2.2 Potential Methods of Segmentation in Customer Base Analytics

Four methods have been identified as having potential to segregate customers. These are as follows:

The Business Rule: Classification based on the availability of the demographic data such as age, gender, class, income and education. Another possibility is the grouping of customers in accordance with the interaction between the customer and the company. This is based on three factors: recency, frequency and monetary. This will be taken up in detail in the research as it is one of the suggested methods.

Quantile Membership: This again uses the RFM approach. This method helps to identify the difference between high value customers and low value customers.

Supervised Clustering along with a Decision Tree: This technique involves the usage of a target that is specific or has a variable that is positively dependable. It is done by drawing out the differences between dependable variables and independent variables. The history of purchase pattern as well as demographic data becomes the modes on which supervised clustering depends. The algorithm which is used is called the decision tree wherein the nodes are targeted.

Unsupervised Clustering: A more mathematical approach which is formulaic in nature. Any attributes of the customer that gives rise to any semblance of similarity is used by the analyst.

The following table shows all the possible methods of customer segregation. These have been proposed by famous names in the field of customer analytics, with respect to the industry of electronic commerce.

2.3 Cluster Analysis: Introduction and Function

Clusters are the groups created to accommodate customers having similar tendencies. They are the very basis of segmentation. The cluster analysis can be performed in various ways. They depend upon a lot of factors which must be taken into account.

The function of it is the evaluation of the ecommerce company. It evaluates the performance and predicts the upcoming performances based on the number of customers that visit or interact with the site. The function is also to understand how the mindset of customers work so as to influence the same and create need where there is no need.

Cluster analysis is one of the most popular ways in which a company performs customer base analytics. In this, the people are clustered together so as to determine common factors influencing all. It is easier to group people together rather than individually calculate.

2.3 The Two Models: RFM and RFMT

The RFM or recency, frequency and monetary model is a popular choice to derive the value of customers in customer base analytics. This is an important method of analysis and is a behaviour-based model, that is, it depends solely on customer behaviour. B2B and B2C

companies both store customer data in their database management systems. This data enables them to perform analysis with the help of the RFM mode.

The R refers to the recency of purchase, as in the last time a customer shopped from a particular site. The F refers to the frequency of purchase done or services availed, that is, the number of times a customer has shopped from a particular website. And M stands for monetary, which simply means the amount of money a customer spends on that particular site every time they make a purchase.

The method or technique of applying the RFM approach is by separating them into quintiles. Recency is divided into a certain number of periods, for instance from 0 days to a total of maybe 730 days. It is then labelled from the alphabet A to E. The letter signifies or denotes the customer who provides maximum value and the letter E denotes the low value customer who is less likely to be profitable to the firm.

The result inferred from this helps to categorize into different sets such as the good frequency customer, the good monetary customer or the poor recency customer. On this basis strategies are developed to lure old customers back to the same site. Promotion strategies are applied here and marketing analysts try their best creative methods to reel in the old customer base.

An important aim of this research is to lay emphasis on the inclusion of the T or tenure approach instead of just using the RFM approach. The T or tenure signifies the duration from the first time of purchase by a customer.

2.4 The K Means of Approach in Cluster Analysis

The most widely used method is the K means of approach in cluster analysis. Previous researches used to incorporate the hierarchical mode of cluster analysis in segmentation. This however was faulty. The structural analysis of a dendrogram led to imprecision in calculation. This is because when the dendrogram reaches a certain level, cutting it off for partition gives rise to imprecision.

Furthermore, all the methods that are non-hierarchical or partitional have their basis on the presumption that the number of clusters and all initial cluster points are pre-defined. This has an effect on the solution.

The K means is an algorithm which groups items into subgroups or clusters of k.

Chapter 3: Problem Statement

The purpose of this research has been established but the problems posed by the same are manifold. The most important question raised is the inability of brands to retain the same customer base. This can have a number of reasons which may not be answered through merely segmentation. The reasons can be varied.

Some of the questions posed that this research seeks to answer are:

Is competition the only reason that affects the customer base? For the reason could be a difficult user interface, a lagging site, or perhaps incompetent payment gateways. It could also depend on the products supplied: a person who has recently purchased a costly dress may not do so again very soon.

If so, then what can segmentation or clustering actually solve? It can help group customers, yes. It can also help to understand the type of customers one has and the income capacity of the same. Segmentation also helps to understand the changes that are required to improve the ecommerce industry.

A disloyal consumer base is a difficult situation but the human mind is just as capable of problem solving. The computation may result in better strategies and can help the company profit.

Chapter 4: Objectives of the Study

To understand how the background of customers affect the dynamics of ecommerce. There is a very specific reason as to the need for analysts in companies. The need to understand customer base is foremost in any business, something at which Mukesh Ambani is excellent. He has invested in all the right boxes to ensure the growth of his e-commerce businesses.

The objective of the study is to see which method works best during customer base analytics and why. This also gives rise to several mathematical computations and need for algorithms to understand the underlying forces that propel a customer to buy. Money is obviously at the centre of it as is want and need. It is the work of the marketing head to ensure that the want of a customer gets changed into need. The aim is also to understand whether customer segmentation is a necessity or an option.

This is to churn customers that is acquire new customers while retaining the old at the same time.

Cluster	Average Recency	Average Frequency	Average Gross Margin	Average Monetary	Count	% Pop	% Rev	% GM
1	32.7	1.7	174.5	1098.5	3321	0.38	0.08	0.05
2	249.6	2	623.1	1932	1022	0.12	0.15	0.18
3	331.5	1.9	488.6	1726	892	0.1	0.13	0.15
4	88.7	1.9	350.1	1501	1994	0.23	0.12	0.1
5	174.4	2	468.8	1822	1261	0.15	0.14	0.14
6	74.9	6.7	1263.8	4933.4	180	0.02	0.38	0.38

Figure 4.1

Chapter 5: Project Methodology

With regard to the methodology of the research work, secondary data had been derived from existing sources. This data is based on an apparel company that produces textiles with fashion brands as its customers.

Define: The very first idea of the project was to define the meaning of segmentation and customer base analytics both of which are the foundation. These were introduced along with cluster analysis in order to clarify the understanding of even a layman.

Plan: The planning involved two main events: research on what customer base analytics meant for an e-commerce industry and the importance of it in present situations. The data was sourced from online resources and the patterns of shopping observed closely, amongst both family and friends. This gave rise to the idea that similar groups must be having similar tendencies or patterns. Hence segmentation was chosen as the topic.

Design: The research must be simple, comprehensive and analytical which gave rise to concrete inferences. Hence the RFMT mode or procedure to analyse segmentation of a customer base of an e-commerce company.

Analysis: Data analysed and computed with the means reported in the research. The findings would only solidify the purpose of the project — segmentation or cluster analysis is an effective means of figuring out types of customers.

Delivery: The presentation of the insights gained to clients illustrated with the means of reports, graphics among others. Techniques have been noted down as well as summarised.

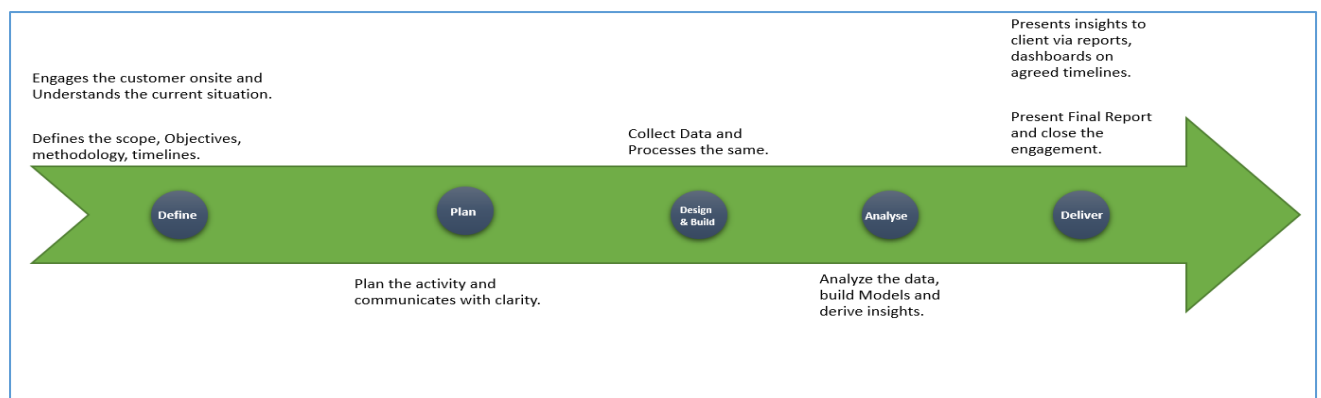


Figure 5.1

Project Pipeline

The project pipeline refers to the time line of the research journal. It gives insight regarding the time that was required to conduct the research by collecting data, and surveying customers. It involves the understanding of the business that is the ecommerce industry. This is followed closely by understanding the data that was collected. The time taken to prepare the data, the segmentation and the resulting evaluation in the project differs.

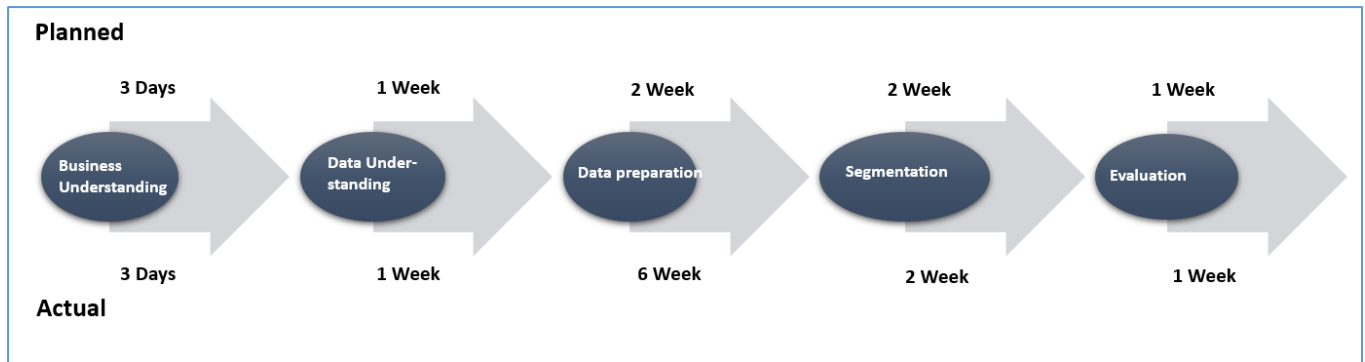


Figure 5.2

There is a difference between the originally planned procedure and the actual procedure.

The image gives a rough idea about the methodology of the project and the pipeline. CRISP DM methodology has been used which is as follows:

Understanding of the Project: The objective of the project had to be established, the focus of it determined and specified. The business had to be analysed by predicting the consequences of the customer base analytics: would an e-commerce company profit or lose by conducting a customer base analytics through segmentation? This question was of considerable importance since any kind of research requires time and in a business: time is money.

Understanding the Data: Since the data was secondary in nature, that is, obtained from sources online, it took time to understand the specifics. The familiarisation with the collected data is an important part of analytics or the entire process could go wrong because of a lack of better understanding. This would only result in lost time, effort and money. To optimize all three, it is necessary to study the data closely, point out the discrepancies and disparities, if any, beforehand.

Preparation of Collected Data: This involves arrangement, the use of algorithms to create subsets and a proper database with the raw data that has been collected. This also involves selecting the procedures that shall be used for analysis and evaluation.

Actual Segmentation or Clustering of the Customers: The grouping of the customers in accordance with the k means of clustering. This is the process that forms the foundation of the research and hence is time consuming.

Evaluation: The results have been evaluated and rechecked and an interpretation released of the same. This is the final step in the research after which a conclusion shall be drawn with a possibility of future recommendations or suggestions.

Chapter 6: Business Understanding

The background of the business is described here. The business chosen is a company that deals in apparel. This is a multinational profit-making organization that has a turnover of about one billion dollars. The company which shall be called 'A' has franchises or retail shops all over the world: 40 shops precisely.

Like all fast fashion companies such as HM, Zara, Mango, Guess and Topshop, even A has switched over to the online platform. It has collaborations with famous e-marketplaces such as Amazon and Flipkart in India, along with an online presence of its own. The profit is sourced from its own website and the company's tie ups with brands.

The clients are all high profile, especially Victoria's Secret and Calvin Klein. The client base also covers Speedo, Triumph and Marks and Spencer's among others.

The business understanding of this company involves two major things. As a new member of the e marketplace, it must build trust in customers with regards to its online services. Along with this, it is also important to establish credibility with customers. A smooth optimization is one of the key factors of running an e-commerce industry so as to probe easy for all transactions.

The problem posed is that this company is unable to retain its customers. This could be due to a number of problems which are mentioned in Chapter 3. The result should be such that they get to know the reason of disloyal customers. The analysis will help them to come up with answers and strategy to reel back old customers.

Chapter 7: Data Understanding

This section gives a description of the data that has been collected.

Sales data: from March 2018 to February 2020.

Repeat Purchase Modelling: data of one year, from first of January, 2019 to December 2019.

All the aspects of the data collected are given below as follows:

The Order Number - a unique number, automatically generated by the system to mark an order.

The External Order Number - number sent to the customer.

The Purchase Date - the date of the order: on which day it was placed and at what time.

Payment type: Cashless or cash on delivery. Whether payment gateways used, debit or credit cards used or net banking or mobile wallets.

Status of Shipment: whether order was cancelled, shipped, received or not dispatched.

Name of the Customer: under whose name was the order made and whether the person has done it for their own self or for someone else.

Country: shipped only in India for the moment.

Email ID: which email address was used.

SKU code: the code under which the customer has placed an order.

Category: the main category of the product: western attire or Indian, demographic

Sub category: the specifics of the order whether jeans or dress or salwar or saree or socks.

Size: the size of the order (gives the basis of an identity of the customer).

Colour: the preferred colour of choice.

Quantity: number of orders per SKU number.

Return Quantity: number of orders returned.

Currency: INR or Indian Rupee.

Shipping cost: cost based on location.

Packing cost: cost of packaging, includes plastic, bubble wrap, paper and tape.

Discount: discount availed in the cart.

Invoiced: the invoice given to customer

COGS: the approximate cost of goods sold. Invoiced in base currency

Margin of profit: gross margin of profit made. Percentage of Gross margin

Primary vendor: the primary seller of product

On hold status: status of product

Replacement order: requests for replacement of orders

Chapter 8: Data Preparation

For any business, data preparation is the most important step as it improves the data quality and increases the overall productivity.

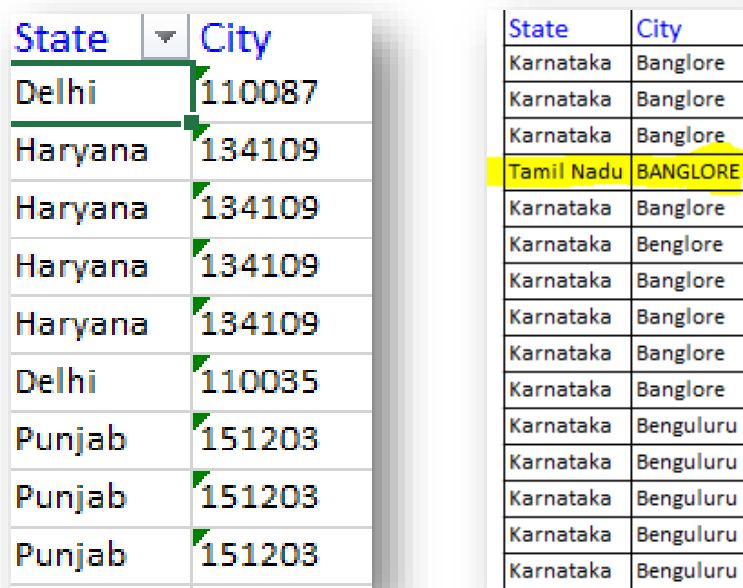
Below are the few techniques which has been used for the data cleaning:

Creating Unique Order No.: Dataset contain multiple Order number for same product purchase (as it's taking one row for every unit purchase) Solution to this is creating a new order number by matching Order No and SKU Code

Removing Outlier: Based on Invoiced amount removed the outlier (these are retailer. Will treat them separately)

The State Column must be cleaned.

Cleaning of cities which may have confusion regarding similar names such as Bombay and Mumbai, Bangalore, Bengaluru. The address may have been entered or just the pin code. Some cities like Bangalore is tagged to Tamil Nadu, Delhi etc.



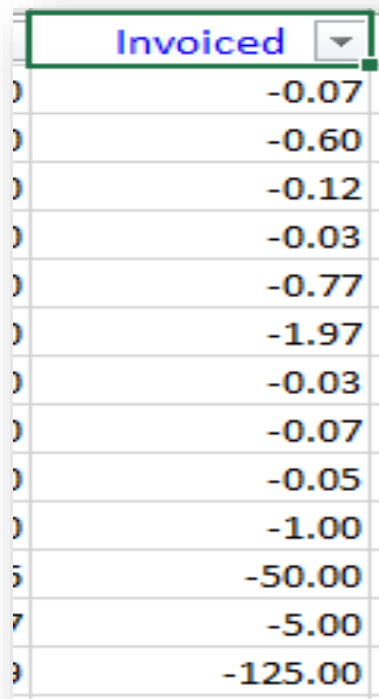
State	City
Delhi	110087
Haryana	134109
Haryana	134109
Haryana	134109
Haryana	134109
Delhi	110035
Punjab	151203
Punjab	151203
Punjab	151203

State	City
Karnataka	Banglore
Karnataka	Banglore
Karnataka	Banglore
Tamil Nadu	BANGLORE
Karnataka	Banglore
Karnataka	Benglore
Karnataka	Banglore
Karnataka	Banglore
Karnataka	Banglore
Karnataka	Banglore
Karnataka	Benguluru
Karnataka	Benguluru
Karnataka	Benguluru
Karnataka	Benguluru
Karnataka	Benguluru

Figure 8.1

Invoiced feature has some negative values: - Upon check with business, we found that these are few dissatisfied customers who has been rewarded with free gift items.

Removing negative value from invoiced has been suggested.



A screenshot of a data table with a column header 'Invoiced' highlighted in blue. The table contains 15 rows of data, all of which are negative values. The values range from -0.07 to -125.00. The table is displayed in a web browser interface.

	Invoiced
0	-0.07
0	-0.60
0	-0.12
0	-0.03
0	-0.77
0	-1.97
0	-0.03
0	-0.07
0	-0.05
0	-1.00
5	-50.00
7	-5.00
9	-125.00

Figure 8.2

Exploratory Data Analysis: Finding out the correlation between the features using Heatmap

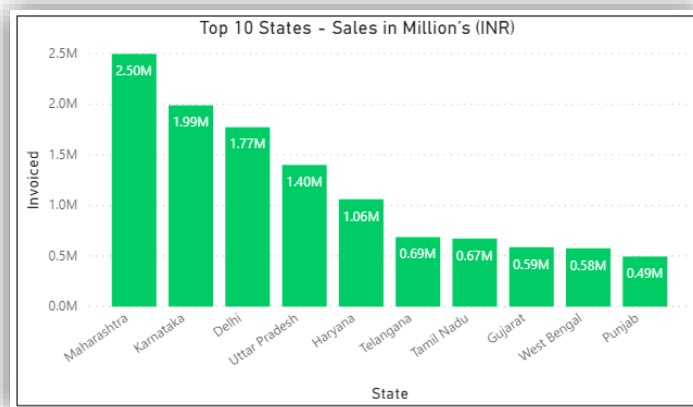


Figure 8.3

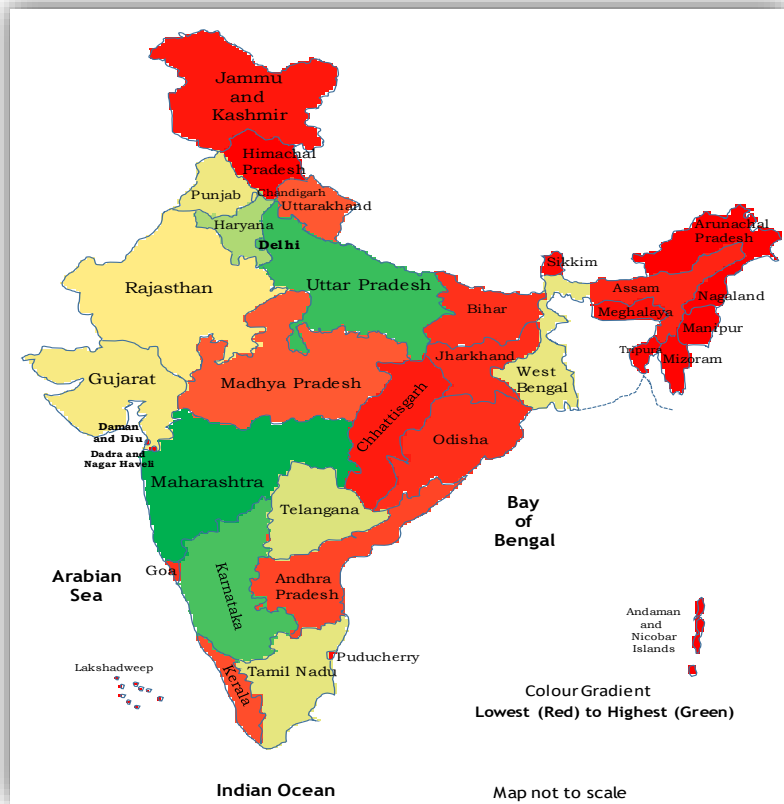


Figure 8.4

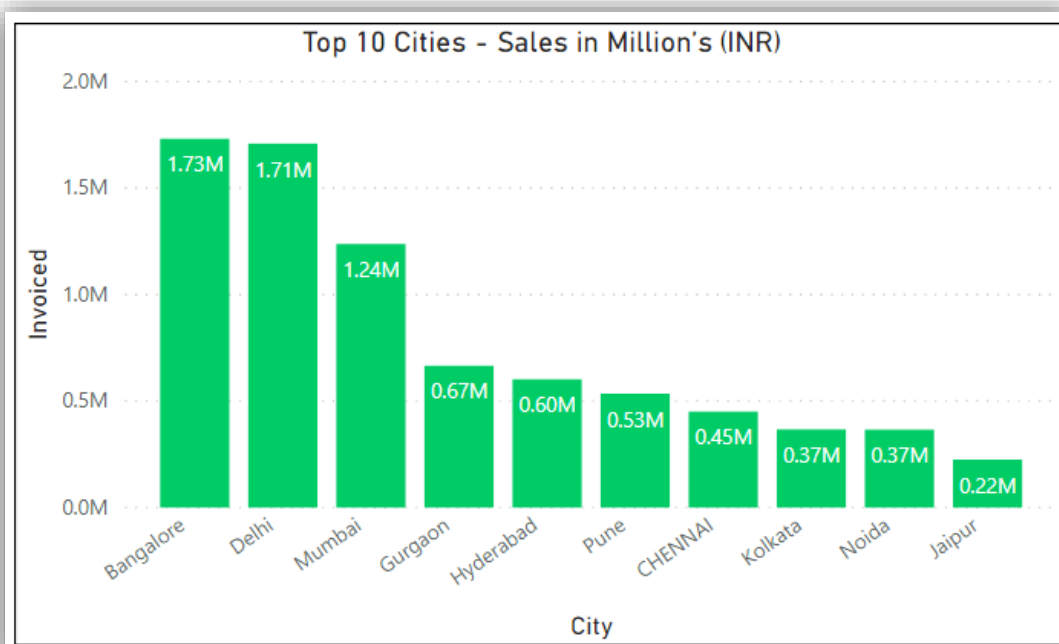


Figure 8.5

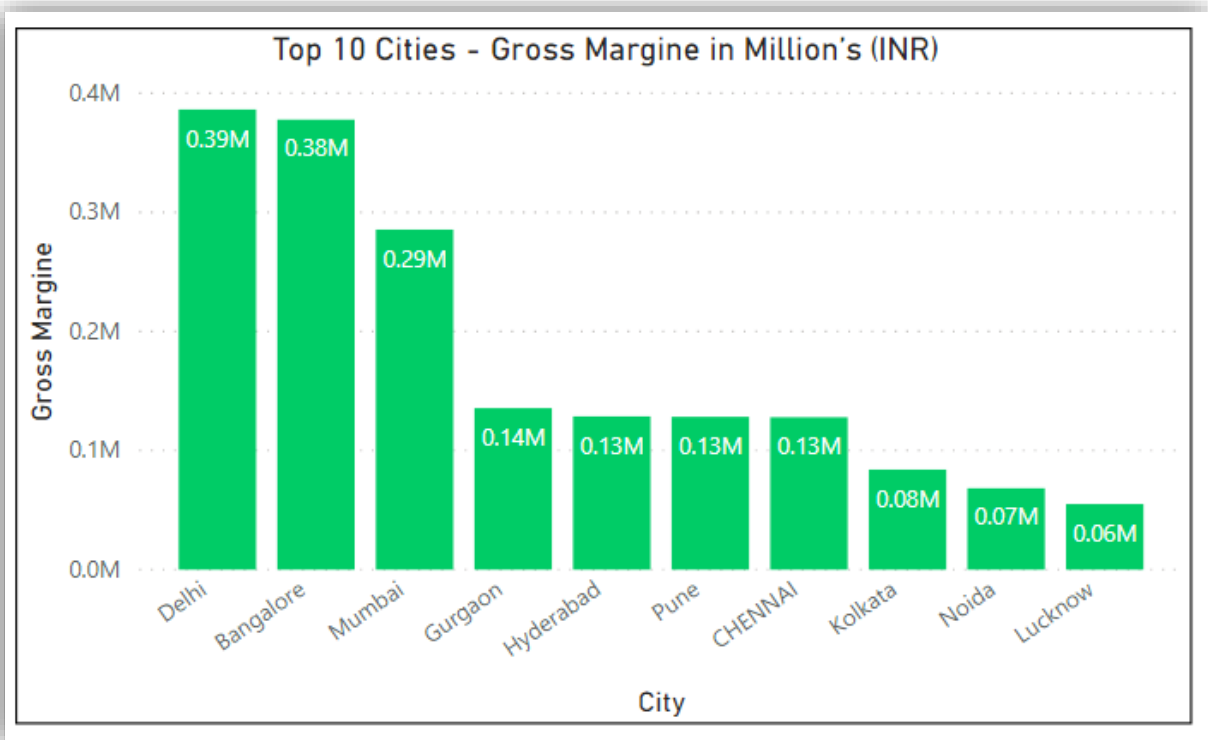


Figure 8.6

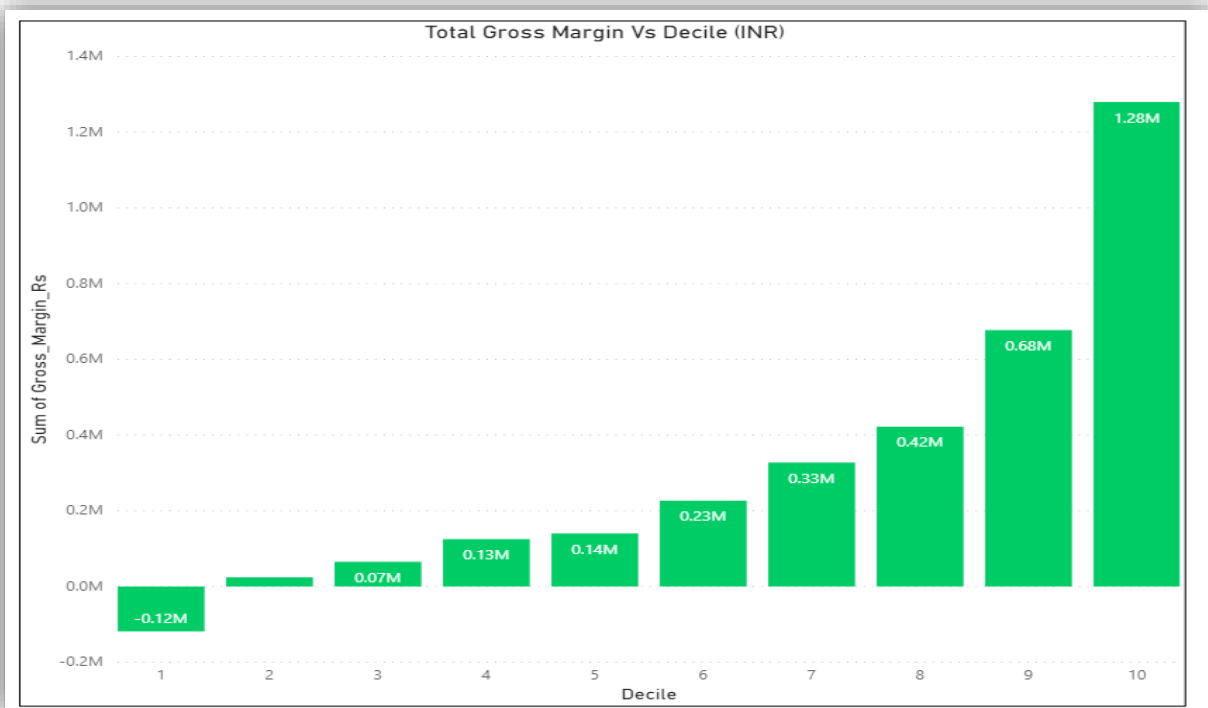


Figure 8.7

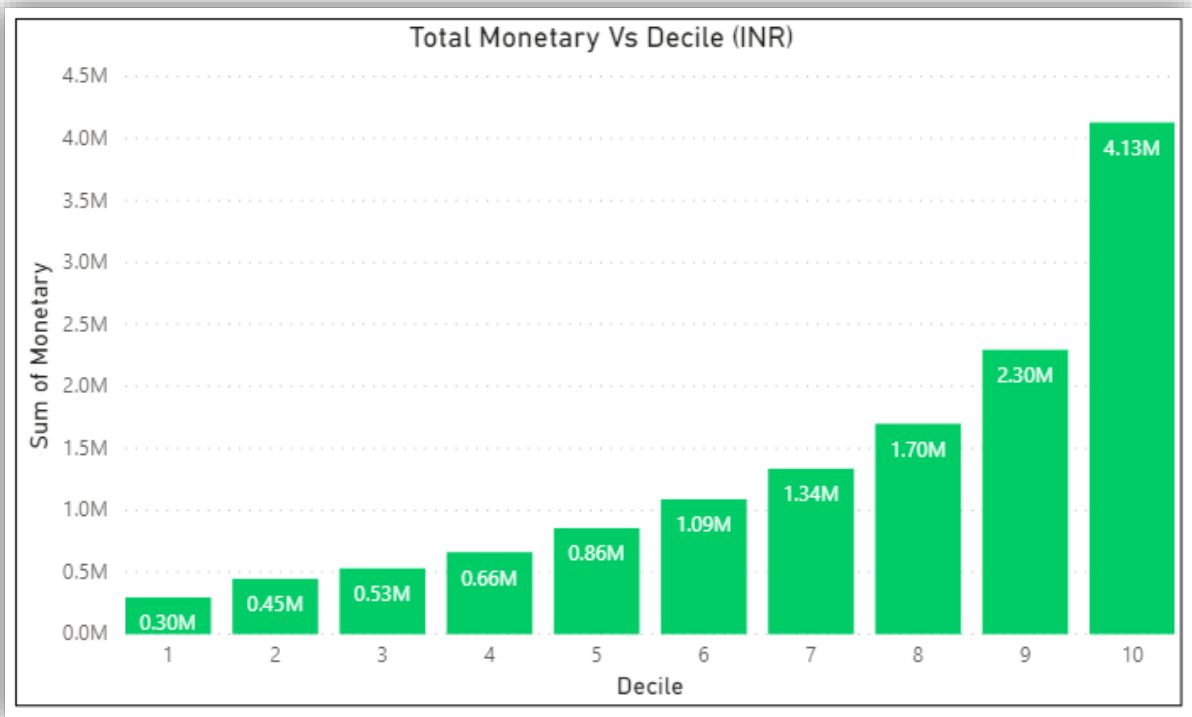


Figure 8.8

Chapter 9: Data Modeling

Evaluation of Data

Data modelling is the manner in which data is evaluated. An effective evaluation requires preciseness. This is why three means have been taken into account. These include the procedures of the above-mentioned K Means, RFM analysis in Segmentation and Hierarchical Clustering.

The aim is to understand the underlying efficiency of all these and any problems which occur will result in the negation of the same. This would lead to a narrower process of analysis.

The idea is to expand the differences between two groups, that is increase between-group distinctions. And to reduce to a minimal difference within the same group. The following techniques are detailed reviews of the procedures.

The post hoc approach of customer segmentation and cluster analysis includes the following procedures:

RFM:

Instead of sticking to the RFM analysis adopted by the companies, the inclusion of Tenure or T becomes a necessary element. This is the most appropriate method.

The Hierarchical Algorithm:

This leads to the building of a hierarchy of clusters, like a ladder, according to the generated data. This is also known as the HCA or the hierarchical cluster analysis technique. It can be further divided into the following sub categories: the divisive approach to hierarchy and the agglomerative approach towards hierarchy. In the former, all the observations that have been jotted down, are supposed to start from one cluster. The further one moves down the hierarchy, the more splits occur in a recursive manner in the clusters. The entire data collection is treated as one huge cluster and the rest of the process includes further division and subdivision of the same cluster into tinier pieces.

The latter is a more individualistic approach. Each of the given observation point starts in a separate cluster. These clusters are paired up together as one moves in an upward direction of the hierarchical structure. These begin with individual points of data and follow a bottom-up approach. This approach requires one to climb down the figurative ladder.

All the steps that are needed for the creation of the Hierarchical Clustering method using agglomeration are as follows:

The beginning of every data must be taken as only one cluster. This means that at every beginning the number of clusters that one has is equal to k. Here, K refers to an integer which is indicative of the various data points.

A cluster has to be formed by the union of two of the closest or nearest points of data. This becomes the K-1 cluster.

More clusters are formed in a similar manner. The next two closest points of data will unite to form the clusters that shall be referred to as K-2.

These steps will have to be constantly repeated until the formation of one humongous cluster. Post formation of said cluster, the dendrogram is used and this gets divided into a multiplicity of clusters: this is based on the problem.

Clustering that is based on the centroid based form:

All members sharing the same features, tendencies, trends, choices among others will naturally belong to the same cluster. The analysis will be conducted amidst the same to underscore the commonality and repetitiveness in patterns. Distance is measured between both clusters, that is, inter clusters or intra clusters. The data which results shall be separated into groups of equal variances. With the shifting of the mean that exists in inter clusters, the minimisation of inertia takes place. This inertia refers to the errors regarding sum of squares that can exist within the clusters. Then the required number of clusters will have to be specified. The method used shall be described below. It is known as the infamous elbow method of analysis.

All the steps that are required in order to create the k means clustering have been given as follows:

A set or group of samples shall give to be divided into disjointed clusters. These can be described as the mean or average of all the samples of clusters. These means are more commonly known as the centroids. It must be noted that these centroids may not generally be the points from the same space even if they do live in the same space. The proposed k- means of algorithm aims at choosing only those centroids that result in the minimisation of inertia within the same clusters.

What has been discovered is that there are a lot of disadvantages also that accompany the k-means clustering. These disadvantages are listed below in order to make it more comprehensive:

An optimal result with regards to the global aspect may or may not be achieved.

All the clusters will have to be selected prior to the use, or selected beforehand. This can be done by either opting for the elbow analysis or the silhouette analysis.

This method, that is, the k-means, has one limitation. It is bounded by the linearity of cluster boundaries.

Since it is based on distance, this method will be an unfortunate procedure when taking into account a very large sample.

The Elbow Method as mentioned above has been described below:

This method helps in the determination of the optimal cluster in the k-means. The metrics that are used are intrinsic. It is known that when k increases, the average rate of distortion decreases. One example is the inertia which results from the error of sum of squares.

Chapter 9: Data Evaluation

The evaluation of the same has been performed keeping in mind the pros and cons of every procedure. The proposed model has been run across the different methods for evaluation purpose. There are three machine learning algorithms which have been used, and as mentioned above they are the RFM analysis approach, the K means of clustering and Hierarchical approach. Of these, the most suitable has resulted in optimal results.

Chapter 10: Deployment

In this project, we have built up a customer segmentation model. The model has been built and tested across multiple algorithms like RFM Analysis, Hierarchical and K-means to find most accurate solution. For the time being, we haven't deployed the model.

Chapter 11: Analysis and Results

This part of the research enumerates the result and its analysis: of the repercussions of the achieved result. This is inclusive of the outcome of the proposed methodology, and covers both the numerical and descriptive work.

The outcome was best when the K means was adopted as the procedure as it showed a better performance in the field of segmentation in customer bases analysis.

Decile Analysis: Out of the entire sales revenue that was generated, 30 percent of it comes from 20 percent of the company's customer base. The topmost is thirteen times better than the bottom and twice as good as the second decile. Only ten percent of sub bases will give rise to 30% revenue.

The customer base analysis resulted in three segments even if the topmost decile is not recent. All the recent transactions are of low value which is an unfortunate circumstance since it shows that there has been a considerable decline in old customers.

The following graphics are accurate illustrations of the achieved result.

Decile	Average Recency	Average Frequency	Average Monetary	Count of Customers
1	66.51	1.12	346.89	867
2	72.61	1.06	515.46	867
3	98.71	1.1	609.72	867
4	110.39	1.2	763.69	867
5	106.79	1.59	988.9	867
6	136.65	1.73	1255.48	867
7	160.74	1.89	1538.72	867
8	152.91	2.35	1952.95	867
9	145.66	2.8	2646.76	867
10	149.07	4.93	4736.64	867

Figure 11.1

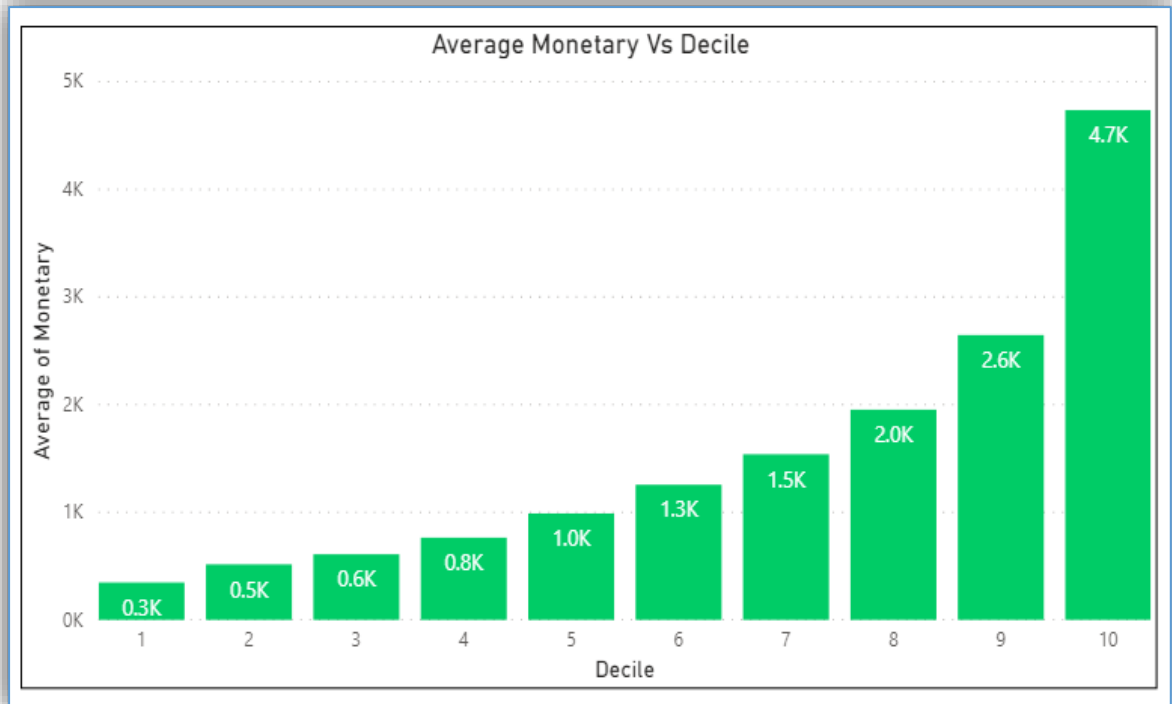


Figure 11.2

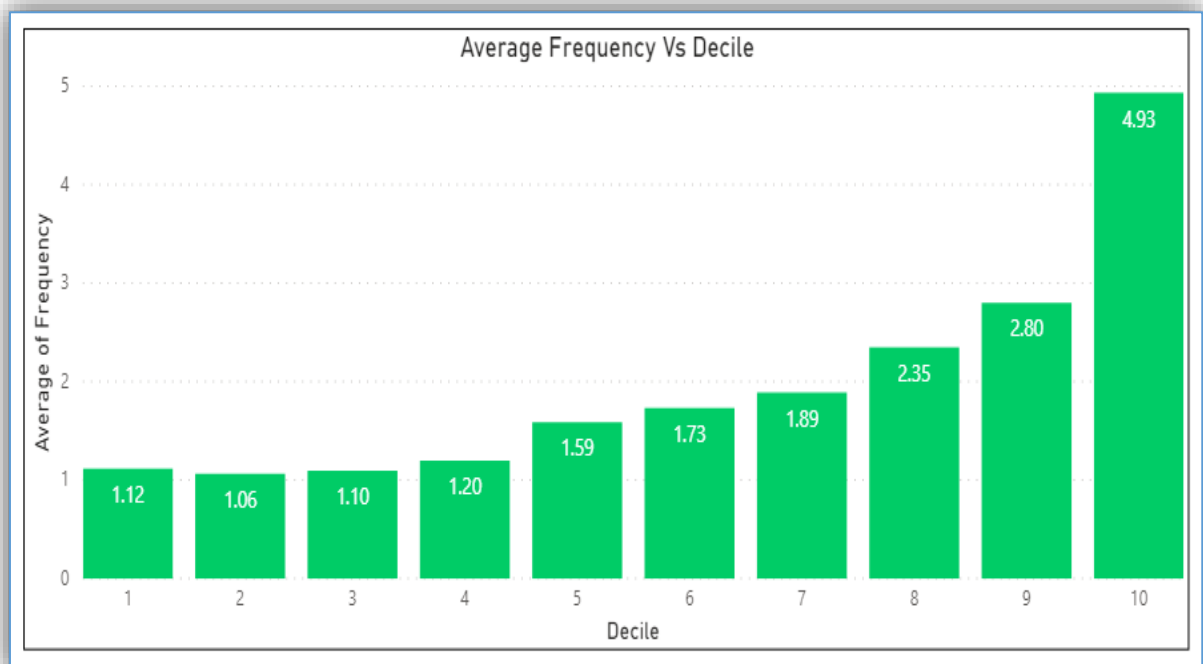


Figure 11.3

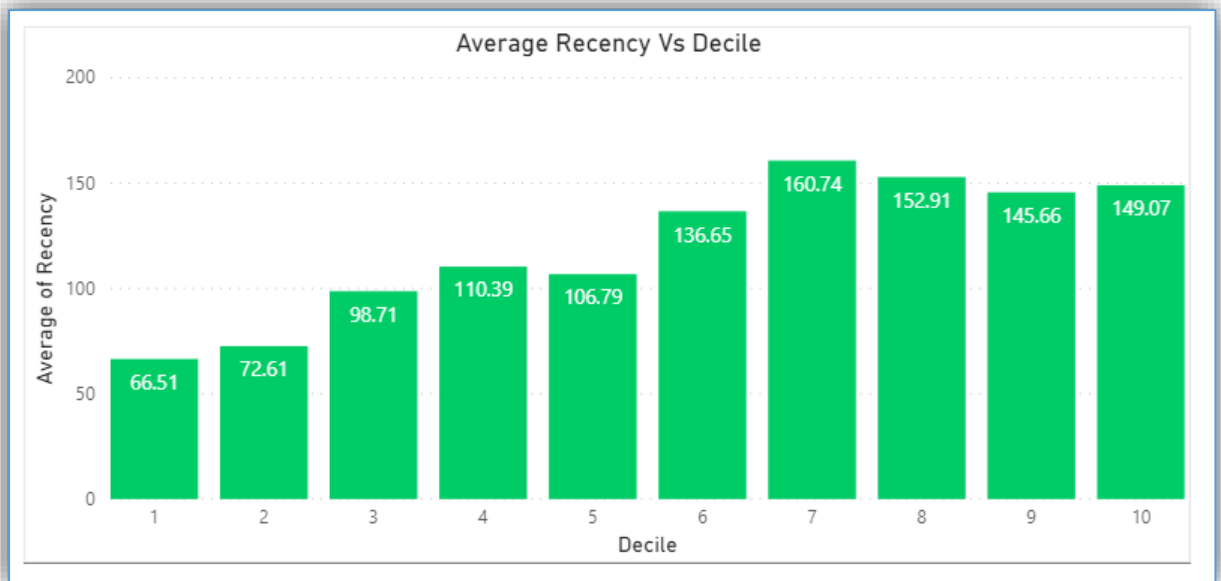


Figure 11.4

Profiles of All Customers and their Clusters:

The totality of the customers crossed the 8000 mark: 8670 to be precise. Of these, the average recency to the e-commerce websites have been one hundred and twenty-three days. The average frequency per year, that is the number of times the customers bought was found to be 1.97 times. On an average, one person would buy a total of two products. The cluster six is not very recent at all. Yet in spite of this fact, they are the most valuable. They create the most profits for the apparel company.

Cluster	Average Recency	Average Frequency	Average Gross Margin	Average Monetary	Count	% Pop	% Rev	% GM
1	32.7	1.7	174.5	1098.5	3321	0.38	0.08	0.05
2	249.6	2	623.1	1932	1022	0.12	0.15	0.18
3	331.5	1.9	488.6	1726	892	0.1	0.13	0.15
4	88.7	1.9	350.1	1501	1994	0.23	0.12	0.1
5	174.4	2	468.8	1822	1261	0.15	0.14	0.14
6	74.9	6.7	1263.8	4933.4	180	0.02	0.38	0.38

Figure 11.5

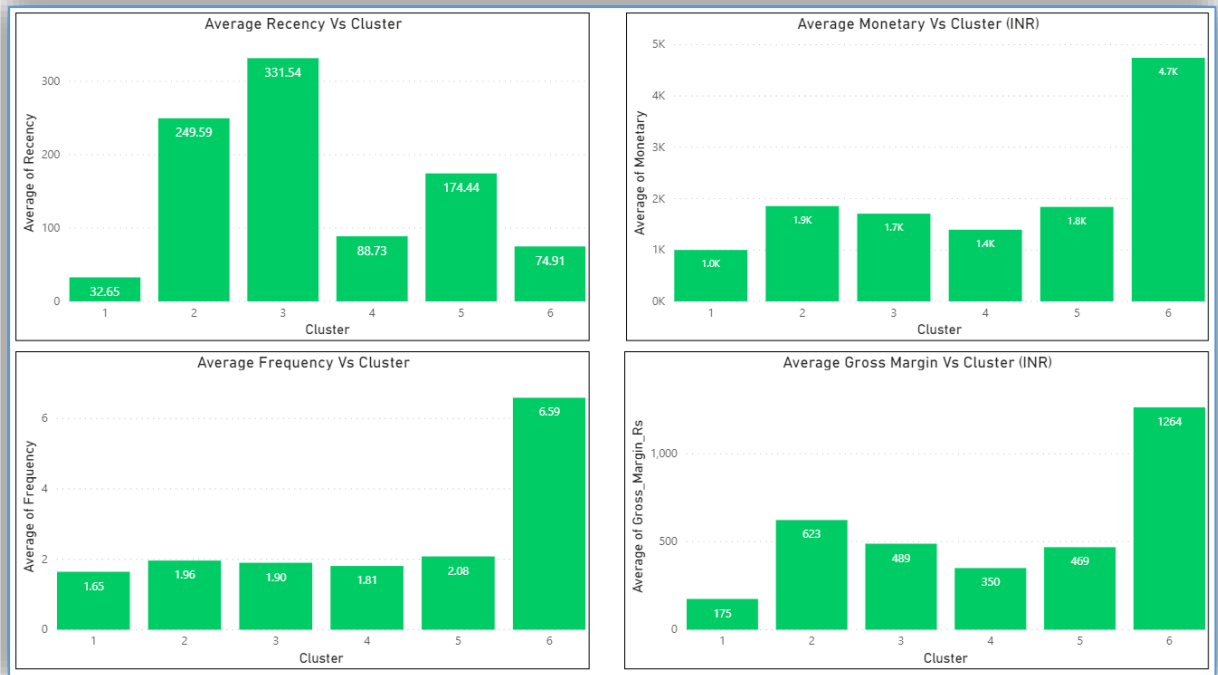


Figure 11.6

The results of customer base analysis are as follows:

The recency is indicative of behaviour of the customers. The monetary aspect has been calculated with respect to the Indian Rupee.

		Recency			Description
		L	M	H	
Monetary	H	522	200	235	Stars
	M	1895	765	1087	Cash Cows
	L	1001	940	2027	Potential - One Timers

Figure 11.7

Chapter 12: Conclusions and Recommendations for future work

Humans thrive on their skill of adaptability. Without the flexibility to change, there cannot be progress or profit in the human world. It is only right to be ready for upcoming changes as that is the only non-variable in life. So, must e-commerce companies adapt and adopt new modes of transaction and find a way to retain customers.

What the research proves is that customer base analytics in e-commerce is a necessity, it is no longer an option. Predictive approaches are compulsory for optimal growth. Customer acquisition is an important part of any business but retaining the said customer is more important: to keep them coming back for more.

There are a few things that every company, whether B2B or B2C, even though for the latter it is of the utmost importance, must have. Social media engagement is one of them and a well an entourage of dedicated analysts. Studies have shown that Facebook and Instagram have been an incredible source of opportunity, as several small businesses operate without websites only on social media platforms. India's digital revolution, the introduction of mobile wallets and UPI have been an integral part of the change.

It must be noted that the more customers switch to anything digital, the easier it gets to leave an imprint on the internet. Hence the data available is hefty and if used appropriately, it can yield good results. It can also be concluded that data mining is crucial and there is no such thing as customer privacy. Everything is known to the data management system: addresses, email Ids, phone numbers. Therefore, it can be safely assumed that privacy is a myth. It does not exist in the modern day.

The findings of the research also help to conclude that the growth in the number of customers does not equate to the growth of revenue. While customer acquisition is still the most important, retaining the same should be given more priority. The topmost, that is the top 10% accounts for almost 30% of the revenue. Three segments or clusters were made possible owing to the amount of revenue generated. All segments of high value customers have low recency- therein lies the problem for the company.

The apparel brand obviously needs an intervention. It needs a strategy that is custom made to retain customers. The problems must be looked into. There is a potential for upselling in all middle value segments. These need the backing of specifically targeted campaign proposals and offers such as discounts and sales. Offers often help generate customer growth, especially the promise of offers: such as those provided by Google Pay.

The conclusion therefore comes to this: all e-commerce sites must be adept at analysing the behaviour of their customer base. They must be intuitive enough so as to stop or avoid losses. In a constantly changing world of overnight glories and condemnation, it is well advised to keep up with trends. The clever will strategize to incorporate psychographic behaviour of the customer. Marketing will be boosted by this research and hence it cannot be ignored. The world has seen an explosion of data and so the clever will be using the same to control their losses and profits. In the end it comes down to just that: the control you have over your customers. There is a need to engage with the same and since a website is unable to do that, this is where social media comes in. Social media must be the answer to customer relations management. Several small companies such as The Big Book Box on Instagram function with the help of

social media to boost sales. They also engage with customers by taking in feedback and actively asking for information regarding the customers' expectations and choices.

The world runs on marketing today. Everything is essentially a marketing gimmick. But marketing alone cannot stand when there has been no research regarding the target audience. The response of the audience or market must be noted, especially the response to offers, promotions and contests. These are active ways of retaining customers. Since there is so much competition, only the unique will survive. And it is true that every unique product loses its charm after a while. This is the reason that one must know how to sugar coat a product as something new.

Another thing that will help with CRM or customer relations management is the user interface of websites. There is a need to check out the competition, and also to understand what works for them.

E-commerce is here to stay. And since it is a fairly new market place, everyone is busy to have a go. This has boosted the careers of such people who did not dream of careers. There is place for everyone here. Customer base analytics is the only measure with which one can dominate the industry. It is also necessary to add new varieties, but in a manner that is appealing. For instance, HM has been at the receiving end of social media censure for its unethical approach of fast fashion. Fast fashion has been responsible for a considerable amount of environmental waste. The brand's environmentally conscious line has gathered censure because they are only 5% of the total of the brand's clothing lines. Hence proved that variety is not everything since the very concept of fast fashion is variety.

Bibliography / References

1. <https://openviewpartners.com/blog/customer-segmentation/#.X68JFh5X6Nw>
2. Review on Customer Segmentation Technique on Ecommerce (© American Scientific Publishers)
3. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-6S, April 2019 Implementation of Customer Segmentation using Integrated Approach
4. Customer Segmentation based on Behavioural Data in E-marketplace Andrew Aziz (research journal)
5. University of Pennsylvania ScholarlyCommons Research Journal : Three Essays on Big Data Consumer Analytics in ECommerce
6. Review on Customer Segmentation Technique on Ecommerce Article in Journal of Computational and Theoretical Nanoscience · October 2016
7. Agarwal, R., Dhar, V., 2014. Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. *Information Systems Research* 25, 443-448.
8. Agarwal, R., Weill, P., 2012. The Benefits of Combining Data With Empathy. *MIT Sloan Management Review* 54, 35.
9. Allen, B., Bresnahan, J., Childers, L., Foster, I., Kandaswamy, G., Kettimuthu, R., Kordas, J., Link, M., Martin, S., Pickett, K., Tuecke, S., 2012. Software as a Service for Data Scientists. *Communications of the ACM* 55, 81-88.
10. Ann Keller, S., Koonin, S.E., Shipp, S., 2012. Big data and city living - what can it do for us? *Significance* 9, 4-7. Bankston, K.S., Soltani, A., 2014. Tiny constables and the cost of surveillance: Making cents out of United States V. Jones.
11. Yale Law Journal Online 123. Barney, J., 1991. Firm resources and sustained competitive advantage. *Journal of management* 17, 99-120. Barrett, M., Davidson, E., Prabhu, J., Vargo, S.L., 2015. Service Innovation in the Digital Age. *MIS Quarterly* 39, 135-154. Barton, D., 2012.
12. Making Advanced Analytics Work For You. *Harvard business review* 90, 78-83, 128. Barton, D., Court, D., 2012. Making advanced analytics work for you. *Harvard business review* 90, 78.
13. Beath, C., Becerra-Fernandez, I., Ross, J., Short, J., 2012. Finding Value in the Information Explosion. *MIT Sloan Management Review* 53, 18-20. Benedettini, O., Neely, A., 2012. Complexity in services: an interpretative framework, *POMS 23rd Annual Conference*. Bennett, P., Giles, L., Halevy, A., Han, J., Hearst, M., Leskovec, J., 2013. Channeling the deluge: research challenges for big data and information systems, *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, pp. 2537-2538.
14. Bouhaddou, O., Bennett, J., Cromwell, T., Nixon, G., Teal, J., Davis, M., Smith, R., Fischetti, L., Parker, D., Gillen, Z., 2011. The Department of Veterans Affairs, Department of Defense, and Kaiser Permanente Nationwide Health Information Network Exchange in San Diego: Patient Selection, Consent, and Identity Matching, *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, p. 135.
15. Boyd, D., Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society* 15, 662-679.

16. Bragge, J., Sunikka, A., Kallio, H., 2012. An Exploratory Study on Customer Responses to Personalized Banner Messages in the Online Banking Context. *JITTA : Journal of Information Technology Theory and Application* 13, 5-18.
17. Braun, V., Clarke, V., 2006. Using Thematic Analysis in Psychology Qualitative Research in Psychology, 3, 77-101. Bristol: University of the West of England. Brown, B., Chul, M., Manyika, J., 2011. Are you ready for the era of 'big data'? *McKinsey Quarterly*, 24-27+30-35.
18. Bughin, J., Chui, M., Manyika, J., 2010. Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. *McKinsey Quarterly*, 26-43. Bughin, J., Livingston, J., Marwaha, S., 2011. Seizing the potential of 'big data'.
19. *McKinsey Quarterly*, 103-109. Chandrasekaran, S., Levin, R., Patel, H., Roberts, R., 2013. Winning with IT in consumer-packaged goods:

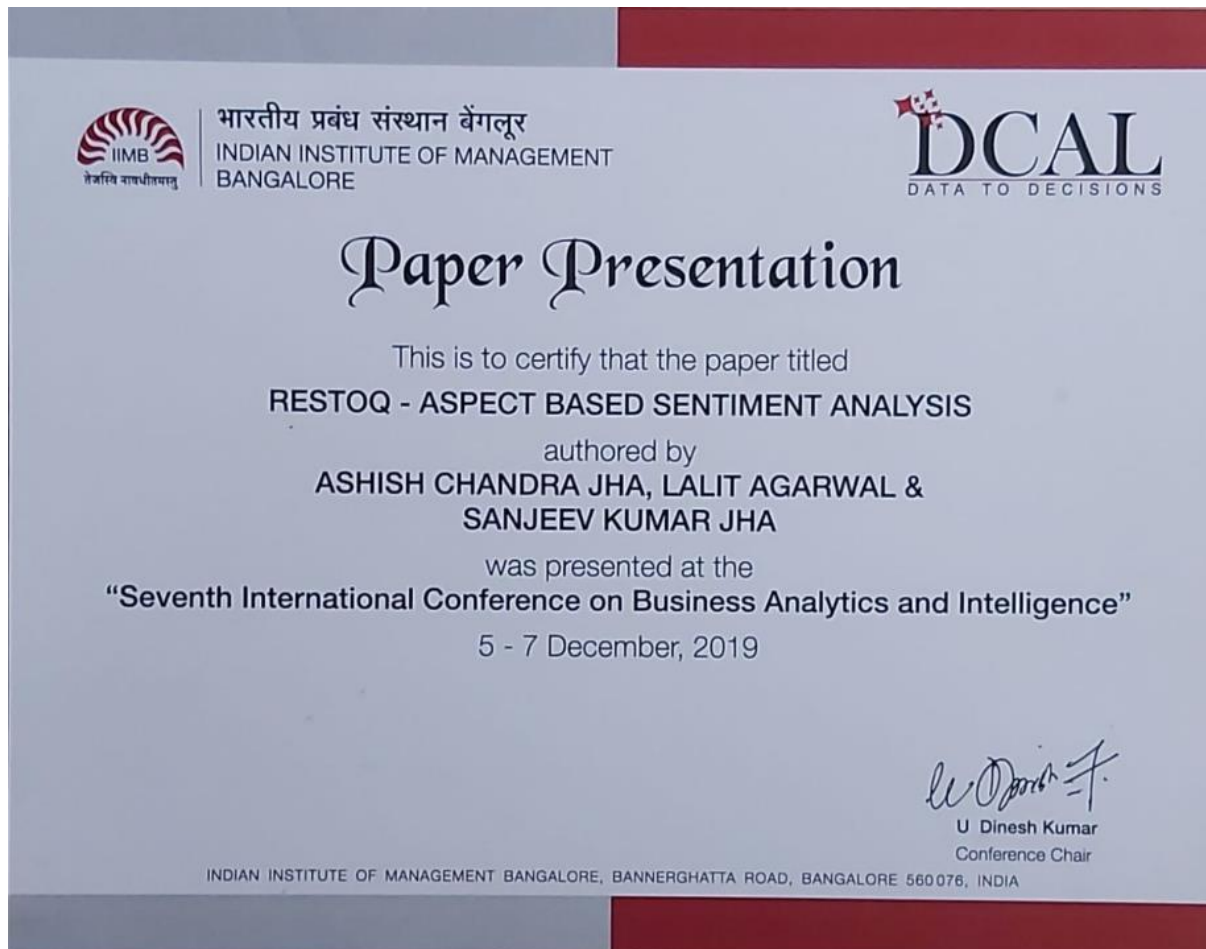
Appendix

Plagiarism Report¹

Customer based analytics in ecommerce			
ORIGINALITY REPORT			
5%	4%	1%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS
PRIMARY SOURCES			
1	www.wholesalehandbagjewelry.com Internet Source	2%	
2	Submitted to Indian Institute of Management Student Paper	1%	
3	Submitted to Sogang University Student Paper	1%	
4	www.termpaperwarehouse.com Internet Source	<1%	
5	www.ijitee.org Internet Source	<1%	
6	Submitted to Lovely Professional University Student Paper	<1%	
7	www.inmybangalore.com Internet Source	<1%	
8	www.scribd.com Internet Source	<1%	
9	open.uct.ac.za Internet Source	<1%	
10	uu.diva-portal.org Internet Source	<1%	
11	eprints.kfupm.edu.sa Internet Source	<1%	
<div> <div>Exclude quotes</div> <div>On</div> <div>Exclude matches</div> <div>< 10 words</div> </div> <div> <div>Exclude bibliography</div> <div>On</div> </div>			

¹ Turnitn report to be attached from the University.

Publications in a Journal/Conference Presented/White Paper²



Any Additional Details

² URL of the white paper/Paper published in a Journal/Paper presented in a Conference/Certificates to be provided.

RestoQ – Aspect Based Sentiment Analysis

Ashish Chandra Jha ^{1,†}
REVA University
Bangalore, India
Ashish.ba05@reva.edu.in

Lalit Agarwal
REVA University
Bangalore, India
Lalit.ba05@reva.edu.in

Sanjeev Kumar Jha ^{2,†}
REVA University
Bangalore, India
Sanjeev.ba05@reva.edu.in

[†]A.C.J. and S.K.J. contributed equally to this work.

Abstract – People love experimenting on their food with different tastes. And when it's about visiting the restaurant or ordering food online, they will definitely look for the reviews which will talk about the aspects like services, ambience and cost along with food quality. This food problem is not a single day problem, it's getting repeated everyday. Some end up with positive reviews and few end up with negative or neutral reviews. In this work a framework is developed called 'RestoQ', which uses text analytics for sentiment analysis at the aspect level to discover and rank the restaurants. The framework analyzes the reviews for the sentiments across four aspects – price, food quality, service quality and ambience. Unsupervised lexicon-based classifier and a naïve Bayesian classifier are used to evaluate and score the sentiments at aspect level. The final score will be a combined sum of each score for the review, which requires further work rank the aspects based on reviews. Surprisingly unsupervised method out performs the supervised method. It is proposed to extend the work with context based methods using word2vect and LSTM.

I. INTRODUCTION

Before ordering food or booking a table in any restaurant consumers generally check the reviews of the places. Online food ordering sites like Zomato, Food Panda and UberEats do sentiment analysis of the reviews given by the customers and give a rating for these restaurants. The majority of current sentiment analysis approaches try to detect the overall polarity of the reviews or sentence regardless of the target entities (e.g. restaurants) and their aspects (e.g. services, ambience and cost along with food quality). Aspect Based Sentiment Analysis is fine grained sentiment analysis. A sentence may contain multiple opinions about different entities and we need to find each of them. This has to be analysed by model and should give insights. In this work, the research findings of such a system are presented.

II. LITERATURE SURVEY

Sentiment analysis is one of the fastest growing research areas in computer science, making it challenging to keep track of all the activities in the area. It is a case of natural language processing which could mark the emotion or mood of the people about any specific product by analysis. It is a process of automatic extraction of features by mode of notions of others about specific product, services or experience. [1]

Customers as well as ecommerce companies (online food ordering in this case) are looking for the reviews of the restaurants to order food or to check their customer satisfaction ratio. A lot of research has been done on Sentiment analysis on restaurants and their reviews. Reviews are considered to be positive, negative or neutral on the overall score of the sentence. To some extent it is very useful and many customers are using it before ordering their food on daily basis [2][3].

Unlike document level sentiment classification task, aspect based sentiment analysis is a more fine-grained classification task. It aims at identifying the sentiment polarity (e.g. positive, negative and neutral) of one specific

aspect in its context sentence. For example, given a sentence "great food but the service was dreadful" the sentiment polarity for aspects "food" and "service" are positive and negative respectively [4].

Aspect Based Sentiment Analysis (ABSA) was introduced as a shared task for the first time in the context of SemEval in 2014; SemEval2014 Task 41 (SE-ABSA14) provided datasets of English reviews annotated at the sentence level with aspect terms (e.g., "mouse", "pizza") and their polarity for the laptop and restaurant domains, as well as coarser aspect categories (e.g., "food") and their polarity only for restaurants (Pontiki et al., 2014). SemEval-2015 Task 122 (SE-ABSA15) built upon SE-ABSA14 and consolidated its subtasks into a unified framework in which all the identified constituents of the expressed opinions (i.e., aspects, opinion target expressions and sentiment polarities) meet a set of guidelines and are linked to each other within sentence-level tuples (Pontiki et al., 2015) [5][6][7].

Aspect Based Sentiment Analysis poses several challenges in processing text data, is a popular area of research in this direction. Several challenges which has not been addressed and people are trying to do some research are implicit aspect detection, mapping aspect words to categories, resolving anaphora references etc. Researchers combine techniques from common sense rules, unsupervised supervised and semi supervised techniques to perform these tasks.

Aspect Based Sentiment Analysis has been done for this particular topic by various researchers [8] [9] [10]. In this paper, the four aspects depending on which the comments will be reviewed. It has been seen that there are lots of aspects which affects the overall sentiment of the review. For example, in restaurants, people give review based on food quality, services, ambience and price. In this work restaurants will categorized based on the customer reviews. The goal is to determine the sentiment expressed toward each aspect on restaurant of Bangalore in English language.

The problem of aspect-based sentiment analysis deals with classifying sentiments (negative, neutral, positive) for a given aspect in a sentence. A traditional sentiment classification task involves treating the entire sentence as a text document and classifying sentiments based on all the words [11].

Labeling of data is a little difficult task to perform automatically. Most of the researcher who are working on new dataset used to label the data manually. The lack of labeled data has led to several researchers to explore unsupervised learning techniques to learn both aspects and their sentiments expressed in plain text. Particularly the fact that aspects are normally described by opinion words and opinion words in turn will have a target aspect can be used to iteratively expand the sentiment and aspect lexicon. The expansion is done with the help of rules to associate aspects and sentiment [12][13].

In this paper, we are trying to do aspect based sentiment analysis on restaurant reviews data from an online food delivery site (Zomato). We have introduced a system based on Text Analytics on the reviews using Supervised Machine Learning with a Naïve Bayes algorithm and unsupervised Machine Learning with Lexicon based algorithm for scoring sentiments

III. RESEARCH METHODOLOGY

In this section, we will explore the different techniques, methods, and features used in this experiment. We will divide the section into two sections: data exploration and pre-processing and model building. Model building is further divided into supervised ML and Unsupervised ML.

Data access: 2000 restaurant across Bangalore along with their reviews has been collected. The data is from an online food ordering company i.e. Zomato. Labeling of data is the hard part of any new research and is done manually. Here the reviews have been labeled based on restaurants name, aspects and sentiment. Positive, negative and neutral sentiments have been used as the three classes.

Data Exploration: Data has the solution to every problem. But one must know how to use that data. Data exploration gives the ability to summarize the main characteristics of a data set, including its size, accuracy, initial patterns, null values, outlier values and missing values. It can use a combination of manual methods and automated tools such as data visualizations, charts, and initial reports to explore the data.

Data Preprocessing: It is the most vital part of any analysis. Considering few important preprocessing steps, below mentioned techniques have been used

- **Stopwords Removal** - Stopwords are the meaningless and repeated words which do not contribute to the semantic of the statement. It should be removed.
- **Symbol Removal** – Reviews generally contain symbols like @, #, \$ with no contribution towards analyzing the sentiments. So, it should be removed.
- **Contractions and Annotation Removal** – The contractions and annotation like shouldn't should be removed with 'should not'
- **Normalization** - Normalization stands for making the word or sentence case insensitive. Data should be normalized.
- **Exploration** – It is to check the word frequency of the corpus. It gives the idea of what the document is about. We check the word frequency by TF-IDF model. Words with high frequency can be seen using word cloud. In addition conditional exploration, based on sentiments and aspects word cloud has been made.
- For sentiments three word cloud, one for positive, one negative and one for neutral emotions. For aspect four word cloud has been made based on food quality, services, ambience and cost.

Model Building: The model will be created four times with different strategies. Here combination of 90%-10%, 80%-20%, 70%-30% and 60%-40% train/test split along with 10 fold cross validation has been used.

Starting with the supervised learning model then tried unsupervised learning model has been created to compare which algorithm will perform better.

In Supervised Machine Learning stage, label data is used for building classifier using Naive Bayes algorithm. Naive Bayesian algorithm is a probabilistic ML algorithm, which assumes independence among the features.

In Unsupervised Machine Learning stage, label data is used for building classifier using Lexicon based algorithm. The lexicon based approach is based on the assumption that the contextual sentiment orientation is the sum of the sentiment orientation of each word or phrase.

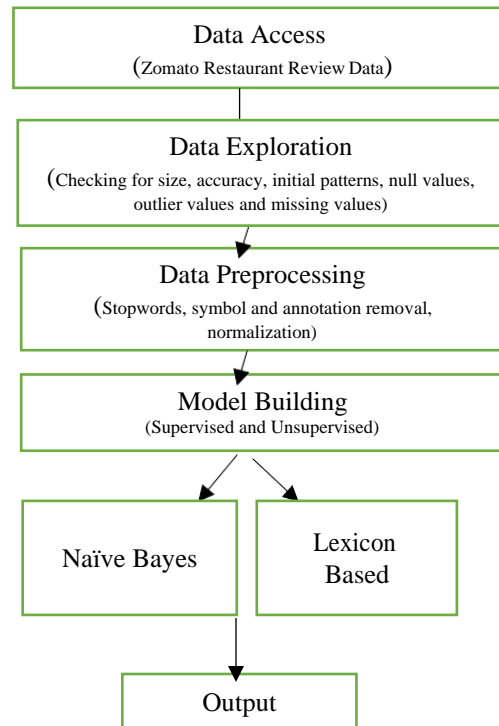


FIG.1: PROPOSED APPROACH

IV. RESULTS AND DISCUSSION

Supervised (Lexicon based model) and Unsupervised (Naive Bayes model) has been created at different train test split. For each instant, accuracy has been captured and report is mentioned below:

Table 1. Experimental results at review level

Training	Test	Accuracy	
		Lexicon	Naive Bayes
90%	10%	71%	72%
80%	20%	68%	65%
70%	30%	70%	65%
60%	40%	68%	63%

The classification accuracy of all the models are consistent with the results published in literature and hence support the methodology used in this research. From the result, it clearly shows, the unsupervised learning lexicon-based model performs better than supervised learning technique using Naive Bayes. Accuracy at different train test split given an idea of optimum split scoring highest accuracy.

V. CONCLUSIONS

This paper covers the Aspect Based Sentiment Analysis on restaurants reviews dataset for Bangalore restaurants. The ABSA task consists of four aspects namely food quality, services, ambience and cost. For each aspect, sentiments have been analyzed. Supervised and Unsupervised machine learning has been used.

The proposed approaches achieved very good results. The algorithm successfully able to analyze the aspects of the sentiments. Further the restaurants are ranked based on the over all score and the positive score, which can be used by consumers for selection of restaurants. It is proposed to carry out a context-based analysis of the sentiments using word2vec and LSTM to test the improvements in the accuracies.

REFERENCES

- [1] "A Study on Sentiment Analysis: Methods and Tools by Abhishek Kaushik1, Anchal Kaushik2, Sudhanshu Naithani3 <https://pdfs.semanticscholar.org/c151/dfad8c1bf88b0afc716758c77d533ded7dd0.pdf>"
- [2] "Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews <https://arxiv.org/ftp/arxiv/papers/1709/1709.08698.pdf>"
- [3] "SENTIMENT ANALYSIS OF RESTAURANT REVIEWS USING HYBRID CLASSIFICATION METHOD

- http://iraj.in/journal/journal_file/journal_pdf/4-54-140014488817-23.pdf
- [4] "Sethia, A., & Bhattacharyya, P. Aspect Based Sentiment Analysis-A Survey. Accessed on August 2019; http://www.cfil.itb.ac.in/resources/surveys/aspect-based-sentiment-analysis_survey.pdf"
- [5] "Bhoi, A., & Joshi, S. (2018). Various Approaches to Aspect-based Sentiment Analysis. arXiv preprint arXiv:1805.01984"
- [6] "Brychcin, T., Konkol, M., & Steinberger, J. (2014, August). Uwb: Machine learning approach to aspect-based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 817-822) <http://alt.qcri.org/semeval2014/cdrom/pdf/SemEval2014145.pdf>
- [7] "Aspect Level Sentiment Classification with Attention-over-Attention Neural Networks Binxuan Huang, Yanglan Ou and Kathleen M. Carley <https://arxiv.org/pdf/1804.06536.pdf>"
- [8] 8. "Pontiki, M., Galanis, D., Papageorgiou, H., Androutopoulos, I., Manandhar, S., Mohammad, A. S., & Hoste, V. (2016, June). Semeval-2016 task 5: Aspect based sentiment analysis. In Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016) (pp. 19-30). <https://www.aclweb.org/anthology/S15-2082>"
- [9] "SemEval-2016 Task 5: Aspect Based Sentiment Analysis Maria Pontiki*1, Dimitrios Galanis1, Haris Papageorgiou1, Ion Androutopoulos1,2, Suresh Manandhar3, Mohammad AL-Smadi4, Mahmoud Al-Ayyoub4, Yanyan Zhao5, Bing Qin5, Orphée De Clercq6, Véronique Hoste6, Marianna Apidianaki7, Xavier Tannier7, Natalia Loukachevitch8, Evgeny Kotelnikov9, Nuria Bel10, Salud Maria Jiménez-Zafra11, Gülşen Eryigit12 <https://www.aclweb.org/anthology/S16-1002.pdf>"
- [10] "SemEval-2015 Task 12: Aspect Based Sentiment Analysis Maria Pontiki*, Dimitrios Galanis*, Haris Papageorgiou*, Suresh Manandhar±, Ion Androutopoulos∅ * <https://www.aclweb.org/anthology/S15-2082.pdf>"
- [11] "Various Approaches to Aspect-based Sentiment Analysis Amlaan Bhoi Department of Computer Science University of Illinois at Chicago Chicago, IL, USA abhoi3@uic.edu Sandeep Joshi Department of Computer Science University of Illinois at Chicago Chicago, IL, USA sjoshi37@uic.edu . <https://arxiv.org/pdf/1805.01984.pdf>"
- [12] "http://UWB: Machine Learning Approach to Aspect-Based Sentiment Analysis Toma's Brychcin in NTIS – New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Univerzita 8, 306 14 Plzeň Czech Republic brychcin@kiv.zcu.cz Michal Konkol NTIS – New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Univerzita 8, 306 14 Plzeň Czech Republic konkol@kiv.zcu.cz Josef Steinberger Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Univerzita 8, 306 14 Plzeň Czech Republic jstein@kiv.zcu.cz <https://www.aclweb.org/anthology/S14-2145.pdf>"
- [13] "Semi-supervised Aspect Based Sentiment Analysis for Movies using Review Filtering Deepa Ananda Deepan Naorema <https://www.sciencedirect.com/science/article/pii/S1877050916300850>"