Customer Segmentation Analysis of E-Commerce Big Data

Indivar Shaik, Tryambak Hiwarkar, Srinivas Nalla

Abstract**:** ***E-Commerce is a major area of application for generating Big Data and the data generated by E-Commerce is increasing rapidly. Now a days, almost all companies selling their products on E-Commerce platforms. The objective of every company is to improve their business and maximize their profits. As tough competition has become the norm of the day, every company works relentlessly to sustain their existing customer base and to acquire new customers, which intern increases their sales volume. To do so, each transaction generated by the company should be recorded so that the data can be further user for analysis. Analyzing such data has become a primary resource for organizations to improve upon their business. The current study analyzes the Customer Segmentation on a sample of Big Data generated by an E-Commerce firm. Customer Segmentation is performed based on the revenue and number of invoices generated on monthly, weekly and on a particular point of time in a day basis. Further this paper focuses on different Segmentation approaches proposed by different researches by analyzing the sample of an E-Commerce Big Data.***

Index Terms**:** ***E-Commerce, Segmentation, Big Data, Transaction.***

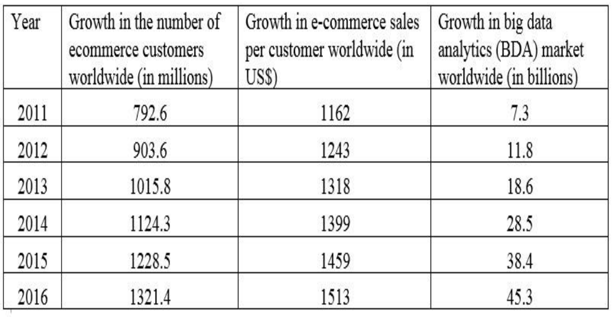
I. INTRODUCTION

Big Data can be defined as the huge amounts of different types of data. This day may be structured, semi-structured or unstructured or even combination of the mentioned. And this data changes and updates frequently over time. Processing this type of data is really a challenging task. While dealing with this type of huge data, organizations generally face difficulties to create, manage and manipulate the data. [1]

E-commerce can be defined as, selling and buying of goods and services online for a profit. These online transactions can be a one-time transaction (like Flipkart, SnapDeal, Amazon, etc.) or can also be continuous transactions (like Amazon Prime, LinkedIn). [2] These transactions generate different types of data like text, images, audios, videos, and many more. This continuous process of recording every transaction generates huge data and analyzing such data is an important task for organizations to improve their business. By attracting more customers organizations can improve their business. To attract more customers' organizations should provide the different goods or services customers are exactly looking for. Once organizations understand customers' requirements, they can exactly provide those goods or services. [3]

Every customer is unique; everyone has their own preferences and criteria. So, analyzing customers is a big task. One way of analyzing customers is Segmentation. Segmentation can be defined as grouping all the customers who are having similar requirements, preferences and characteristics. This can be treated as Customer Segmentation or Market Segmentation. [3]

Table 1: Global growth rate of E-Commerce



Year	Growth in the number of ecommerce customers worldwide (in millions)	Growth in e-commerce sales per customer worldwide (in US$)	Growth in big data analytics (BDA) market worldwide (in billions)
2011	792.6	1162	7.3
2012	903.6	1243	11.8
2013	1015.8	1318	18.6
2014	1124.3	1399	28.5
2015	1228.5	1459	38.4
2016	1321.4	1513	45.3

Source: Adapted from emarketer (2013) and (Piatetsky, 2014)

Customer segmentation is an activity to divide customers into groups that have the same characteristics. Customer Segmentations enables organization to match with customers' wishes and provide similar products to them. By using Customer Segmentation organizations can identify who are the profitable for the organization. [1]

The remaining sections of this paper are organized as follows. Section II provides the details of various research works related to E-Commerce based Customer Segmentation. Section III introduces the methodology used in this research work.

II. RELATED WORKS

Baer et al., used Customer Segmentation Intelligence to improve marketing by offering products or services that meet the needs of each customer group. As a part of their research Customer Segmentation Intelligence, they used internal data by looking at the demographic data from customer profile and purchase history. [4]

The following are the customer segmentation methods proposed by the Baer:

i) Business Rules: in this method, customers are grouped into specific groups based on a predetermined class, [5] such as:

a) Grouping based on demographic data, such as age, gender, income and education, etc.

Manuscript published on 30 June 2019.

***** Correspondence Author (s)
Indivar Shaik*, Research Scholar, Dept. of Computer Science and Engineering, SSSUTMS, Sehore, Madhya Pradesh, India.
Dr. Tryambak Hiwarkar, Professor, Dept. of Computer Science and Engineering, SSSUTMS, Sehore, Madhya Pradesh, India.
Dr. Srinivas Nalla, Principal, Sahaja Institute of Technology & Science for Women, KarimNagar, Telangana, India.

b) Grouping based on customer interaction with the company based on data purchase pattern such as the type of product or service provided or RFM data, where R is Recency (when customer last shopped), F is Frequency (how frequently the customer purchases) and M is Monetary (how much amount the customer spends every time).

ii) Creation of Quantile Membership Segments: This method uses data Recency, Frequency, and Monetary. Here are the quantile membership methods: [5]

a) Recency(R), Frequency(F) and Monetary(M) are divided into five groups of intervals, for example, starting from 0 days up to 365 days then classify it with label A B C D and E, where A is very valuable customer and E is low-value customer. When 3 RFM is combined, there is a label AAA until EEE.

b) Map two components of RFM to a table.

c) Divided into two groups A, B with the classification most valuable customer and two groups D, E to the classification of least valuable customer. C is average value customer.

d) The result can be concluded as, for example better frequency (A or B), moderate monetary (A or B) but bad Recency (D or E).

iii) Supervised Clustering with decision tree: This method uses a specific target, or dependent variable and target would predict differences in independent variables (input). Data utilized in this method is previous purchase pattern data and customer demographic data. The algorithm that is used is decision tree with the target on their nodes. According to Baer, even though this method connects the target with the other customer characteristics, it focuses on only one aspect of customer behavior. [5]

iv) *Unsupervised Clustering*: This method uses any number of customer attributes then measure the similarity among customers, each customer attribute then uses Euclidean distance method and then cluster the customers by using k-means clustering. If the distance is the shortest distance between customer data and cluster, then customer is included in that cluster. [5]

Magento et al., categorizes the data into internal data and external Data. Customer registration, customer profile, and purchase history are the internal data obtained from the database of an ecommerce. While external data are census data, media browsing, surveys and market search cookies, web and social media analysis. Information about customer lifestyle, attitude, activity and shopping preferences are obtainable through surveys and market search and social media. Browsing history can be seen from server log or cookies. [6]

Magento et al., used several variables in their research. They are:

a) *Profit Potential:* Identify the potential customers by using frequency of the transactions, last purchase date, average value of the order and their lifetime. [7]

b) *Past Purchases:* Analyzing the product's purchase history using the type of product, price of the product, method used for payment and shipping the product, and finally feedback and reviews about the product (satisfaction with the product).

c) *Demographic:* Analyzing location of the customer (like country, state, region), age of the customer, gender of the customer, income of the customer, ethnicity, educational qualification, profession, device used for browsing (like smartphone, PC, Laptop, Tablet and their model).

d) *Psychographic:* Analyzing the hobbies and interests of customer, customer's activities, and affiliations of customer (like professional, religious, political, cultural, and institutional).

e) *Behavior:* using the variable of pages viewed, responses to offers and promotions, participation in reward programs, channel management. [8]

Collica et al., described the segmentation as the process of categorizing or classifying items into groups those are having similar characteristics. In CRM (Customer Relationship Management) segmentation is used to classify the customer based on some similarities by segmenting the records of customer database. They used the customer database and purchase history on customer segmentation methods. [9]

III. METHODOLOGY

The aim of this research is to analyze the Customer Segmentation process based on different variables. Secondary data were collected and relevant literatures were referred. The dataset (Online Retail Dataset) for conducting this research was collected from UCI Machine Learning Repository which has 541909 records of the transactions. And the Customer Segmentation was performed and analyzed using R programming language.

In developing Customer Segmentation in E-Commerce organizations, for the collected dataset Segmentation was done based on the point of time like monthly analysis, weekly analysis and daily analysis. Customer Segmentation Process was done as follows. First Market Identification, Identifying Potential needs of Customer, after that Segmentation of customers and finally Evaluation of Customer Segment Behavior.

In the first step of Customer Segmentation, an E-Commerce organization need to identify the top most countries where the highest revenues are generated and more transactions recorded. Once Customer Segmentation is done, identifying potential needs of customer falls into place. Here, the products with highest sales and the potential customers who generate high business for the organization are to be identified and analyzed. So that organization can analyze requirements, wishes and characteristics of potential customers. After that, using Data Mining, a customer profile can be viewed and his requirements, wishes and characteristics can be analyzed. Based on the results of the Data mining, Organizations can divide customers into different segments based upon their criteria. [7][10]

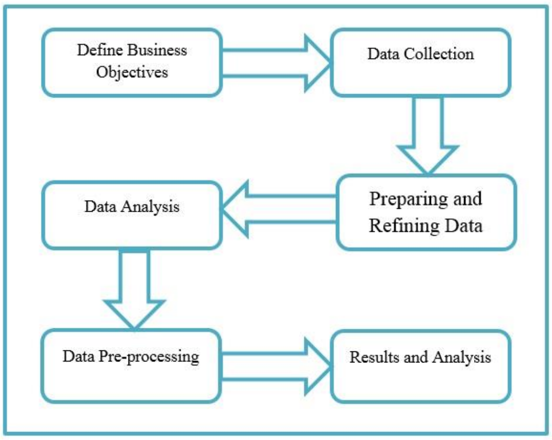
Some of the key categories used in this research are as follows:

- *Objective of the Customer:* Identifying the purpose of the customer vising the website. It can be done by tracking the past transactions of the customer.
- *Date of the Month:* Every customer is unique. Some of the customers prefer to shop during the first week of the month, because generally salaried employees get salaries during the first week of the week. This helps to target the particular segment of customers.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

- *Day of the Week:* Some customers prefer to shop during the weekends while some customers prefers weekdays. It can be identified by using the transaction history of the customers.
- *Time of the Day:* If a customer visiting the website during the late hours it can be assumed that that particular customer is an employee working from morning to evening. So it is better to target such customers during the late hours.

Figure 1: Customer Segmentation Process



Customer Profiling in Segmentation process was done base on the Type of Customer (Whether the customer is new or old), Date of the Month (whether customer purchases products throughout the month or purchases on a specific day in a month), Day of the Week (whether he shops during weekdays or weekends), Time of the Day (particular time period of day like morning hours 8AM-10AM, or late hours 20PM-22PM) and Purchase History of customer (what the customer purchased for a particular period like purchases made in a month).

IV. EMPERICAL ANALYSIS AND RESULTS

This experiment was conducted by using R programming language and R-Studio. As already mentioned earlier, the dataset was collected from UCI Machine Learning Repository which consists of 541909 records of the transactions. As explained in the previous section, First Business Objectives of customers are defined. The purpose of conducting this research is to identify the Revenue generated on a particular point of time (in a day, hour and week etc.), countries with highest revenue, average order values and transaction in a particular point of time.Once the dataset is gathered, it is refined (removing duplicates, removing missing values, etc.) and made it ready for the analysis. As part of Data Refining, to have a better view of data, the dataset was glimpsed to know the data type of the columns and partial view of the dataset. There are different types datatypes are available in the dataset. In data pre-processing different constraints are applied on data to extract the required results.

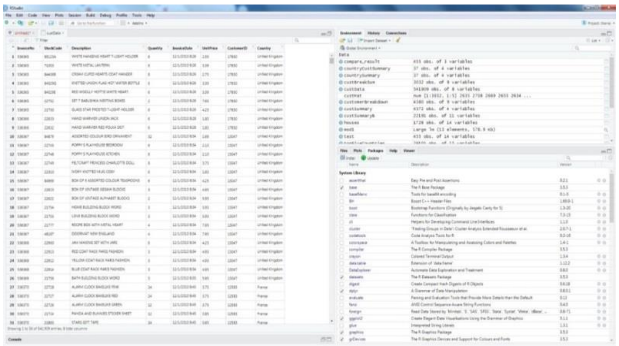


Figure 2: Contents of Dataset

As part of the data refinement first missing data from the dataset was retrieved and removed missing values from the dataset were removed as there is no use with that data. Among the available fields in the dataset, 24.93% values of Customer ID fields are missing. Initially there were 541,909 entries available in the dataset. Once the missing values were removed the numbers of entries were 406829.

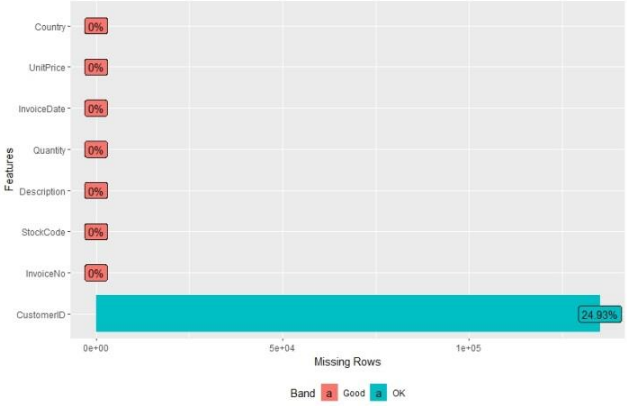
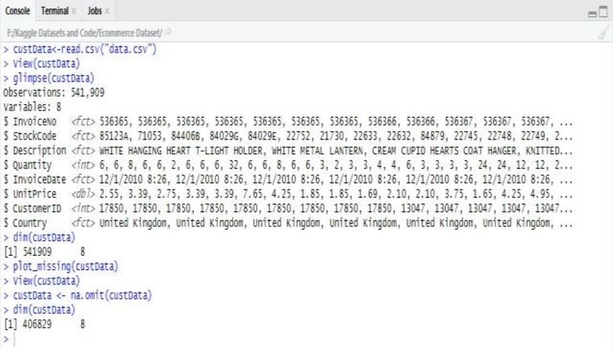


Figure 3: Missing Values

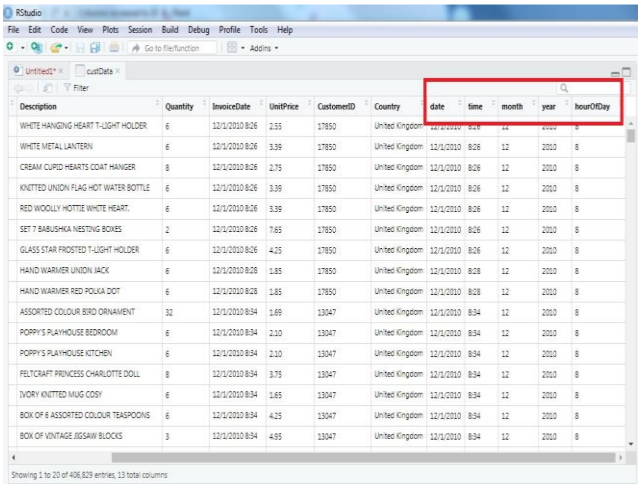


```
> custData<-read.csv("data.csv")
> View(custData)
> glimpse(custData)
Observations: 541,909
Variables: 8
$ InvoiceNo   <fct> 536365, 536365, 536365, 536365, 536365, 536365, 536365, 536366, 536366, 536367, 536367, 536367, ...
$ StockCode   <fct> 85123A, 71053, 84406B, 84029G, 84029E, 22752, 21730, 22633, 22632, 84879, 22745, 22748, 22749, 2...
$ Description <fct> WHITE HANGING HEART T-LIGHT HOLDER, WHITE METAL LANTERN, CREAM CUPID HEARTS COAT HANGER, KNITTED...
$ Quantity    <int> 6, 6, 8, 6, 6, 2, 6, 6, 6, 32, 6, 6, 8, 6, 6, 3, 2, 3, 3, 4, 4, 6, 3, 3, 3, 3, 24, 24, 12, 12, 2...
$ InvoiceDate <fct> 12/1/2010 8:26, 12/1/2010 8:26, 12/1/2010 8:26, 12/1/2010 8:26, 12/1/2010 8:26, 12/1/2010 8:26, ...
$ UnitPrice   <dbl> 2.55, 3.39, 2.75, 3.39, 3.39, 7.65, 4.25, 1.85, 1.85, 1.69, 2.10, 2.10, 3.75, 1.65, 4.25, 4.95, ...
$ CustomerID  <int> 17850, 17850, 17850, 17850, 17850, 17850, 17850, 17850, 17850, 13047, 13047, 13047, 13047, 13047...
$ Country     <fct> United Kingdom, United Kingdom, United Kingdom, United Kingdom, United Kingdom, United Kingdom, ...
> dim(custData)
[1] 541909      8
> plot_missing(custData)
> View(custData)
> custData <- na.omit(custData)
> dim(custData)
[1] 406829      8
> |
```

Figure 4: New Dataset after removing Missing Values

As mentioned earlier the main focus of this research is to find out the segments of customers who generate maximum business for the organizations on particular point of time such as date, time and day. In the given dataset all the segments such as date, time and day were incorporated into one field. So we have separated invoice date field into separate fields date, time, month, year and hour of the day for better analysis.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

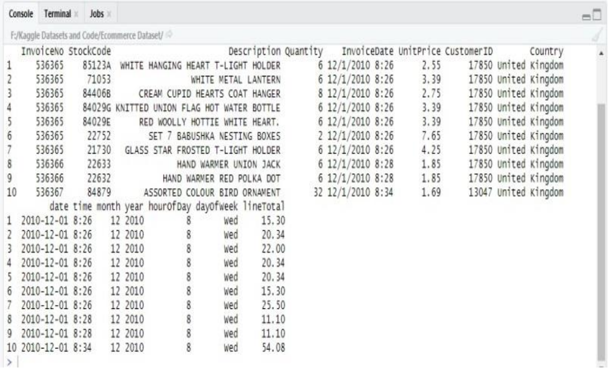

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Untitled1*
custData
Filter

Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	date	time	month	year	hourOfDay
WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom	12/1/2010	8:26	12	2010	8
WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom	12/1/2010	8:26	12	2010	8
CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom	12/1/2010	8:26	12	2010	8
KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom	12/1/2010	8:26	12	2010	8
RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom	12/1/2010	8:26	12	2010	8
SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850	United Kingdom	12/1/2010	8:26	12	2010	8
GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850	United Kingdom	12/1/2010	8:26	12	2010	8
HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850	United Kingdom	12/1/2010	8:28	12	2010	8
HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850	United Kingdom	12/1/2010	8:28	12	2010	8
ASSORTED COLOUR BIRD ORNAMENT	32	12/1/2010 8:34	1.69	13047	United Kingdom	12/1/2010	8:34	12	2010	8
POPPY'S PLAYHOUSE BEDROOM	6	12/1/2010 8:34	2.10	13047	United Kingdom	12/1/2010	8:34	12	2010	8
POPPY'S PLAYHOUSE KITCHEN	6	12/1/2010 8:34	2.10	13047	United Kingdom	12/1/2010	8:34	12	2010	8
FELTCRAFT PRINCESS CHARLOTTE DOLL	8	12/1/2010 8:34	3.75	13047	United Kingdom	12/1/2010	8:34	12	2010	8
IVORY KNITTED MUG COSY	6	12/1/2010 8:34	1.65	13047	United Kingdom	12/1/2010	8:34	12	2010	8
BOX OF 6 ASSORTED COLOUR TEASPOONS	6	12/1/2010 8:34	4.25	13047	United Kingdom	12/1/2010	8:34	12	2010	8
BOX OF VINTAGE JIGSAW BLOCKS	3	12/1/2010 8:34	4.95	13047	United Kingdom	12/1/2010	8:34	12	2010	8

Showing 1 to 20 of 406,829 entries, 13 total columns

Figure 5: Separating invoice date into date, time, month, year and hour of day fields

We have created a variable, i.e., day of week which provides information of sales transactions on the particular day of the week. Later we have created one more variable, i.e., line total which calculates the values by multiplying the Quantity with the Unit Price for each entry.



```
   InvoiceNo StockCode                         Description Quantity     InvoiceDate UnitPrice CustomerID        Country
1     536365    85123A  WHITE HANGING HEART T-LIGHT HOLDER        6 12/1/2010 8:26      2.55      17850 United Kingdom
2     536365     71053                 WHITE METAL LANTERN        6 12/1/2010 8:26      3.39      17850 United Kingdom
3     536365    84406B      CREAM CUPID HEARTS COAT HANGER        8 12/1/2010 8:26      2.75      17850 United Kingdom
4     536365    84029G KNITTED UNION FLAG HOT WATER BOTTLE        6 12/1/2010 8:26      3.39      17850 United Kingdom
5     536365    84029E      RED WOOLLY HOTTIE WHITE HEART.        6 12/1/2010 8:26      3.39      17850 United Kingdom
6     536365     22752        SET 7 BABUSHKA NESTING BOXES        2 12/1/2010 8:26      7.65      17850 United Kingdom
7     536365     21730   GLASS STAR FROSTED T-LIGHT HOLDER        6 12/1/2010 8:26      4.25      17850 United Kingdom
8     536366     22633              HAND WARMER UNION JACK        6 12/1/2010 8:28      1.85      17850 United Kingdom
9     536366     22632           HAND WARMER RED POLKA DOT        6 12/1/2010 8:28      1.85      17850 United Kingdom
10    536367     84879       ASSORTED COLOUR BIRD ORNAMENT       32 12/1/2010 8:34      1.69      13047 United Kingdom
         date time month year hourOfDay dayOfWeek lineTotal
1  2010-12-01 8:26    12 2010         8       Wed     15.30
2  2010-12-01 8:26    12 2010         8       Wed     20.34
3  2010-12-01 8:26    12 2010         8       Wed     22.00
4  2010-12-01 8:26    12 2010         8       Wed     20.34
5  2010-12-01 8:26    12 2010         8       Wed     20.34
6  2010-12-01 8:26    12 2010         8       Wed     15.30
7  2010-12-01 8:26    12 2010         8       Wed     25.50
8  2010-12-01 8:28    12 2010         8       Wed     11.10
9  2010-12-01 8:28    12 2010         8       Wed     11.10
10 2010-12-01 8:34    12 2010         8       Wed     54.08
> 
```

Figure 6: Sample of first 10 entries after Data Refinement

A) INVOICE ANALYSIS:

1. Monthly Analysis:

We have analyzed the invoices for the years 2010 and 2011 based on months in a year and compared the results. For the year 2010 all the invoices were made in the month of December alone. For the year 2011 highest invoices were made in the month of November followed by October and September. Based on this analysis we can group the customers who made the invoices in the months of November to December as a segment.

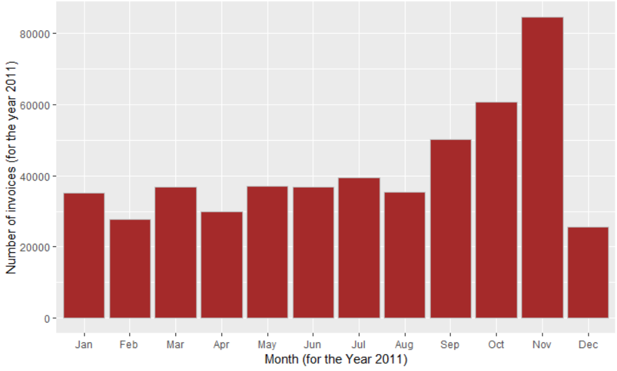


Figure 7: Invoices for the year 2011 by Month

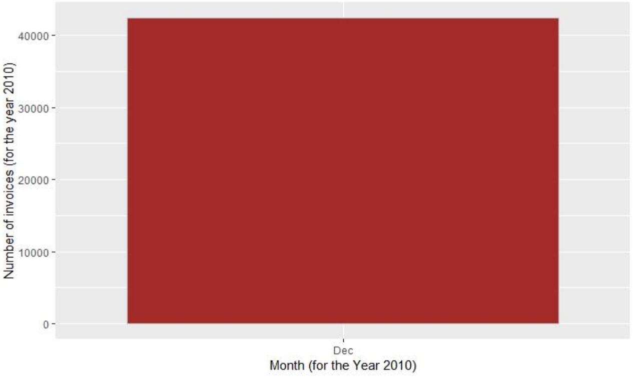


Figure 8: Invoices for the year 2010 by Month

2. Weekday Analysis:

We have analyzed the invoices for the years 2010 and 2011 based on a day of a week in a year and compared the results. For the year 2010 highest invoices were made on Friday followed by Monday. For the year 2011 highest invoices were made on Thursday and Tuesday. Based on this analysis we can group the customers who made the invoices Tuesday, Thursday and Friday as a segment.

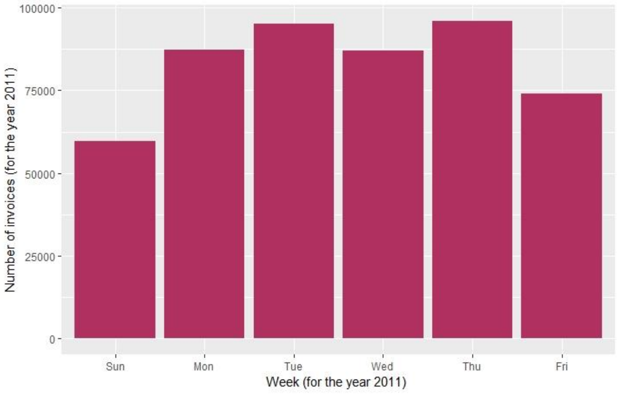


Figure 9: Invoices for the year 2011 by day

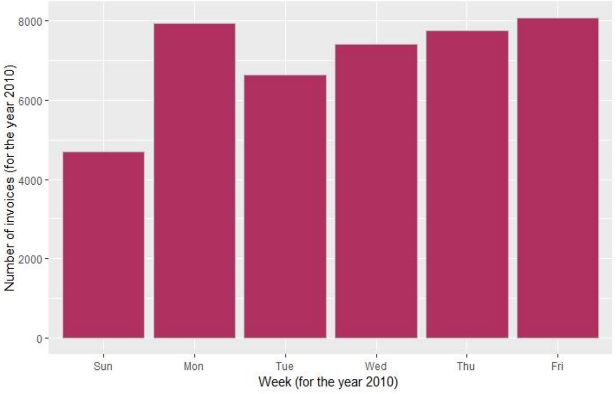


Figure 10: Invoices for the year 2010 by weekday

3. Based on a particular point of time in a day:

We have analyzed the invoices for the years 2010 and 2011 based on a particular point of time in a day and compared the results. Based on this analysis most of the invoices were made during the time 10AM-3PM.

So we can group the customers who made the invoices during that particular time period as a segment.

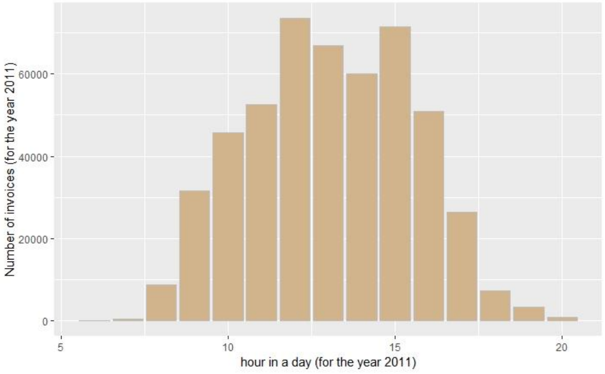


Figure 11: Invoices for the year 2011 by Hour

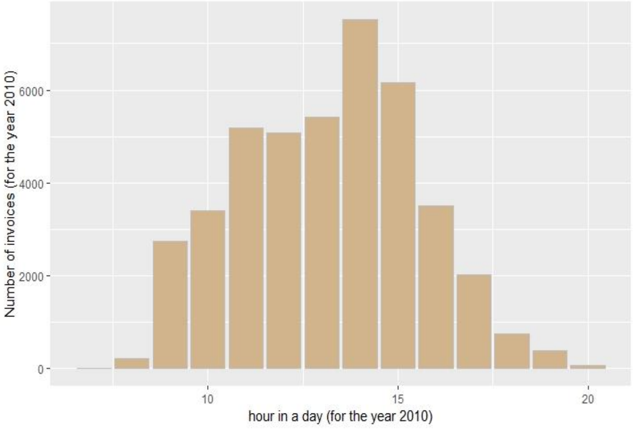


Figure 12: Invoices for the year 2010 by Hour

B) REVENUE ANALYSIS:

1. *Monthly Analysis:*

We have analyzed the revenue generated for the years 2010 and 2011 based on months in a year and compared the results. For the year 2010, as already analyzed as above all the invoices were made in the month of December alone, so all the revenue was generated in the month of December. For the year 2011 highest revenue was generated in the month of November followed by October and September. It is also similar to the year 2010, as the numbers of invoices in the month of November were high, so the revenue generated by the month of November is also high. Based on this analysis we can group the customers who made the invoices in the months of September to December as a segment.

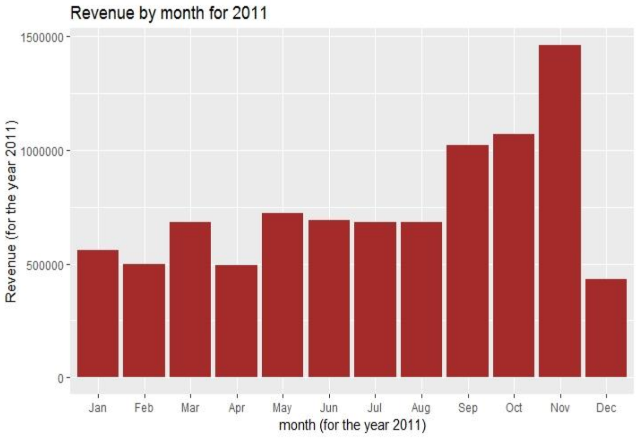


Figure 13: Revenue for the year 2010 by Month

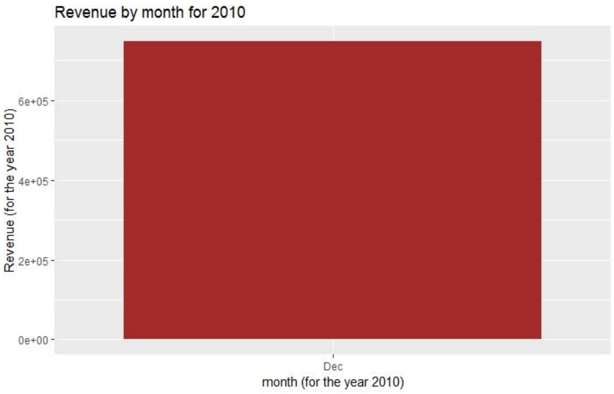


Figure 14: Revenue for the year 2010 by Month

2. *Weekday Analysis:*

We have analyzed the revenue for the years 2010 and 2011 based on a day of a week in a year and compared the results. For the year 2010 highest revenue was generated on Thursday followed by Friday. For the year 2011 highest revenue was generated on Thursday followed by Tuesday. Based on this analysis we can group the customers who made the invoices Thursday and Friday as a segment.

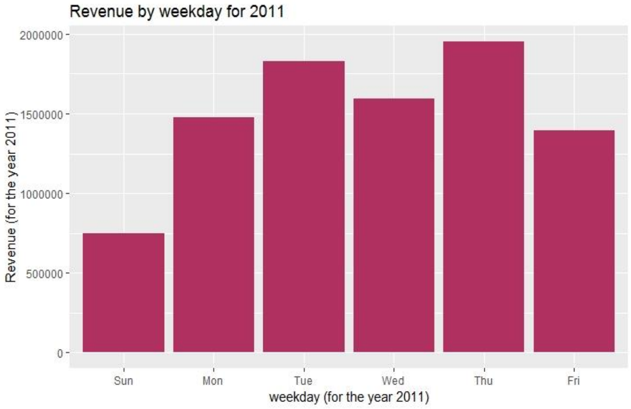


Figure 15: Revenue for the year 2011 by weekday

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

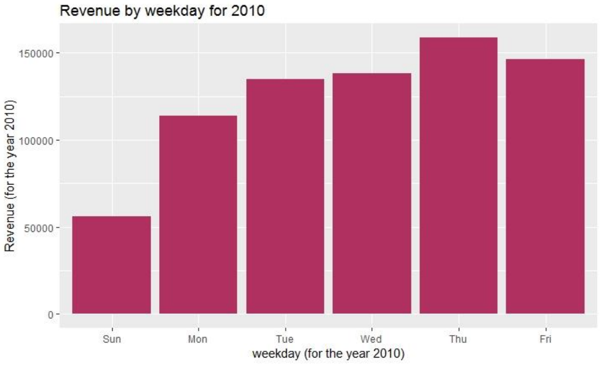


Figure 16: Revenue for the year 2010 by weekday

3. Based on a particular point of time in a day:

We have analyzed the revenue generated for the years 2010 and 2011 based on a particular point of time in a day and compared the results. Based on this analysis highest revenue was generated during the time 10AM-3PM. So we can group the customers who made the invoices during that particular time period as a segment.

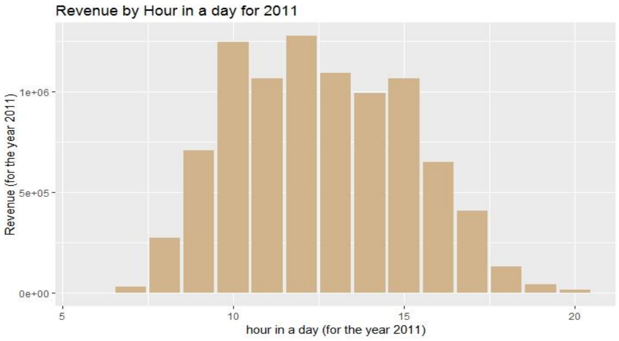


Figure 17: Invoices for the year 2011 by Hour

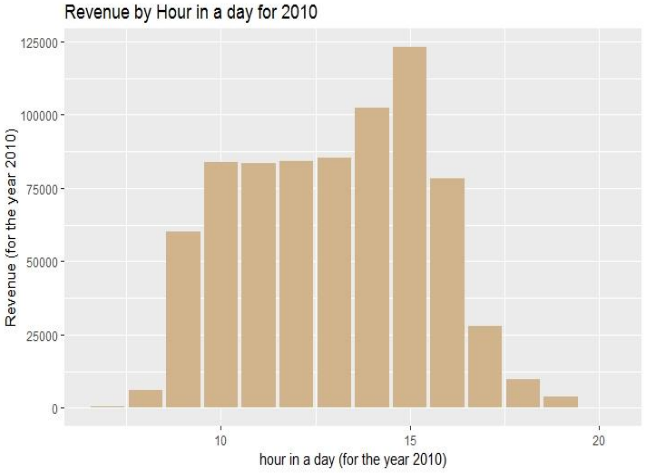


Figure 18: Invoices for the year 2010 by Hour

Summary:
From the above analysis it can be summarized customers who are shopping between 10am-3pm on Thursday and Friday in the month of November and December can be grouped as a Segment.

Table 2: Summary of Analysis

	Invoices		Revenue	
	2010	2011	2010	2011
Monthly Analysis	December	November	December	November
Weekday Analysis	Friday	Thursday	Thursday	Thursday
Hourly Analysis	10am-to-3pm	10am-to-3pm	10am-to-3pm	10am-to-3pm

V. CONCLUSION

As the data generated by E-Commerce organizations growing rapidly, identifying the potential customers is a challenging task. Customer Segmentation is a way of identifying and grouping customers depending different criteria's. In this paper we have analyzed, compared and summarized the Customer Segmentation process of an E-commerce firm. This paper focused revenue generated by the customers and number of invoices made by the customers and analyzed the results based on Monthly, Weekday and Hourly analysis. From results analysis, it can be concluded as most of the times customers who are having more number of invoices are generating more revenue. It can be further analyzed by taking different other criteria's such as past purchases, average orders made by the customers, countries with highest number of invoices and also based RFM analysis.

REFERENCES

1. Vinodhini, M & Manju, A. (2016) " A Survey on Big Data Analytics in E-Commerce". International Journal on Innovations in Engineering Sciences and Technology (IJIEST), Volume 1 Issue 1 June.
2. Frost, R., Strauss, J., (2014). "e-Marketing, 7e". Pearson Higher Education.
3. Kumbogulu. (2007). "Cluster Analysis for Segmentation". *Quintessence Int*, *38*(91), 92–98.
4. Baer, D. (2012). SAS Global Forum 2012 "Customer Intelligence CSI : Customer Segmentation Intelligence for Increasing Profits. *SAS Global Forum"*, 1–13. Retrieved from http://support.sas.com/resources/papers/proceedings12/103-2012.pdf
5. Sari, J. N., Nugroho, L. E., Ferdiana, R., & Santosa, P. I. (2016). "Review on customer segmentation technique on ecommerce". *Advanced Science Letters*, *22*(10), 3018–3022. https://doi.org/10.1166/asl.2016.7985
6. Genius, C., & Fisk, P. (n.d.). An introduction to Customer Genius.
7. Ballestar, M. T., Grau-Carles, P., & Sainz, J. (2018). "Customer segmentation in e-commerce: Applications to the cashback business model". *Journal of Business Research*, *88*(November 2017), 407–414. https://doi.org/10.1016/j.jbusres.2017.11.047
8. Liu, Y., Li, H., Peng, G., Lv, B., & Zhang, C. (2015). "Online purchaser segmentation and promotion strategy selection: evidence from Chinese E-commerce market". *Annals of Operations Research*, *233*(1), 263–279. https://doi.org/10.1007/s10479-013-1443-z
9. Randall, S., "Customer Segmentation and Clustering Using SAS" Enterprise Miner, Third Edition.
10. Sundjaja, A. M. (2013). "Analysis of customer segmentation in Bank XYZ using data mining technique". *Asian Journal of Information Technology*, *12*(1), 39–44. https://doi.org/10.3923/ajit.2013.1.39

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)