A Minor Project Report

on

Demographic Customer Segmentation

carried out as part of the course CS1634 Submitted by

Akhil Sambaraju

189301175

VI-CSE

in partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY

In

Computer Science & Engineering

MANIPAL UNIVERSITY
JAIPUR

Department of Computer Science & Engineering,
School of Computing and IT,
Manipal University Jaipur,

May 2021

ACKNOWLEDGEMENT

We had a great experience working on this project and we got to learn a plethora of new skills through this project. However, it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them. We are highly indebted to the teachers and especially Dr. Rishi Gupta for their guidance and constant supervision as well as providing necessary information regarding the project and for their support in completing the project. We would like to express our gratitude towards our parents and friends for their kind cooperation and encouragement which help us in the completion of the project.

Place: Jaipur Akhil Sambaraju (189301175)

Date:14-06-2021

CERTIFICATE

This is to certify that the project entitled "***Demographic Customer Segmentation***" is a bona fide work carried out as part of the course ***Minor Project (CS1634)***, under my guidance by ***Akhil Sambaraju***, student of ***B. TECH*** at the Department of Computer Science & Engineering, Manipal University Jaipur, during the academic semester ***6th***, in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Computer Science & Engineering, at MUJ, Jaipur.

Place: Jaipur, Rajasthan
Date: 14-06-2021

Signature of the Instructor (s)

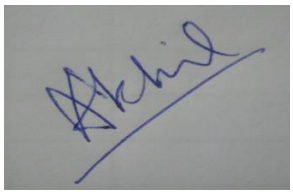
DECLARATION

I hereby declare that the project entitled "***Demographic Customer Segmentation***" submitted as part of the partial course requirements for the course ***Minor Project (CS1634),*** for the award of the degree of Bachelor of Technology in Computer Science & Engineering at Manipal University Jaipur during the ***6th Semester, April 2021*** semester, has been carried out by me. I declare that the project has not formed the basis for the award of any degree, associate ship, fellowship, or any other similar

Further, I declare that I will not share, re-submit, or publish the code, idea, framework and/or any publication that may arise out of this work for academic or profit purposes without obtaining the prior written consent of the Course Faculty Mentor and Course Instructor.

Place: Jaipur, Rajasthan

Date: 14-06-2021



Signature of the Student

Abstract:

With the ever-increasing population size, comes an ever-increasing diversity in tastes and preferences. Catering to each of these nearly 7 billion preferences individually is an unimaginable task. Whereas providing the same service to whole population would nullify the meaning of ‘preferences’.

This is where customer segmentation acts as a middle ground. Customer segmentation is a way to cater to tastes and preferences of groups of individuals rather than individuals itself. Although, the individuals in these groups might not have the exact same preferences, but they lie in the same ballpark, making them more similar to each other than the individuals of other groups.

Segmentation is the first step in ‘targeted marketing’, which is followed my targeting and eventually by positioning. One way of performing said segmentation is by manually segregating customers one by one, be it by using MS Excel or any query language. But this way is very cumbersome and error prone, it is also very time inefficient.

Therefore, machine learning algorithms are used for big data sets. This not only eliminates the above problems, but it also increases the scope of analysis through data manipulation and visualization. The most common machine learning algorithms used for customer segmentation are the unsupervised clustering algorithms. We are going to perform one of these, k-prototype, and look at how it performs when it comes to customer segmentation.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Scope of the Work

There are several ways a business tries to attract new customers all the while trying to retain its current customers. Up until a few decades ago, these businesses practised what was known as mass marketing - in which, companies tried to sell the most popular product to all its customers. Or they practised product differentiation – in this, companies offered a variety of products to a large market.

But, as technology rapidly evolved, companies moved to a newer approach – personalising the products and targeting it to a specific market segment. Customer segmentation (or Market segmentation as it is widely known) is the most approachable method for obtaining the above result. Understanding their customers and the market has allowed businesses to service each of its customers with care and personalisation and proved the value of focusing on them.

Since the birth of ecommerce, sellers are constantly looking for ways to expand their reach while also maintaining a stable relationship with their frequent customers. And, in the age of booming social media, consumer data is as readily available as daily bread. Companies like Facebook, Google sell billions of dollars' worth consumer data to corporations which is then used as a basis of market segmentation. These ecommerce giants like Amazon, Flipkart target each segment with personalised promotional offers as well as personalised advertisements driving in clicks and eventually large amounts of profits.

The aim of this desktop app is to gather demographic data directly from the customer, then assign them one of the segments formed after clustering the dataset. This way each customer will be offered deals and promotions according to their properties. This makes them more likely to buy a product (of their choice) from the app since it has been discounted.

1.2 Product Scenarios

The applications of customer segmentation are irreplaceable. It has now become the heart of product marketing and strategy in any industry. It is an indispensable tool for organizations to understand the market, whom to target with what product, and how to optimize the marketing strategy.

With the increasing popularity of social media, consumer data is readily available for businesses to use and profile the costumers, to understand them better and act accordingly.

Let us say a website has a new user. The user would be asked to register an account on the website, for which they would be offered a discount coupon. Instead of offering every new user the same coupon, the website could - with the help of customer segmentation – give out targeted discount coupons. This will increase the likelihood of that user to buy a product rather than window shop.

2. Requirement Analysis

2.1 Functional Requirements

The software for demographic customer segmentation has the following functional requirements:

- **Login/Signup**
  The user should be able to register a new account or login to an existing account. This login info should be stored in a sql database.

- **Add to cart**
  Inside the products screen (where every product is listed under their respective categories), there should be an option to add the product to cart under every product.

- **Fill a survey**
  The customer would be asked to fill a survey about their general information, which includes – name, age, gender, income, education, number of dependents, and marital status.
  They would be incentivized with a discount coupon if they successfully fill the survey. This discount coupon would be based on the segment they are assigned after running the clustering algorithm using their survey info.
  The data entered by the customer will also be stored in a sql database.

- **View Cart**
  This allows the customer to view all their cart products along with their prices, edit their quantity.
  If the customer has filled the survey, it should show the appropriate discount coupon applied and the total discount obtained.
  At the end, it should show the total price (with discount) the customer has to pay.

2.2 Non-Functional Requirements

The software for demographic customer segmentation has the following non-functional requirements

- The user interface of the application should be user-friendly.
- Any changes made by the user should be brought with immediate effect.
- The login/signup system should be secure.

2.3 Use Case Scenarios

Use Case Diagram for the Application:

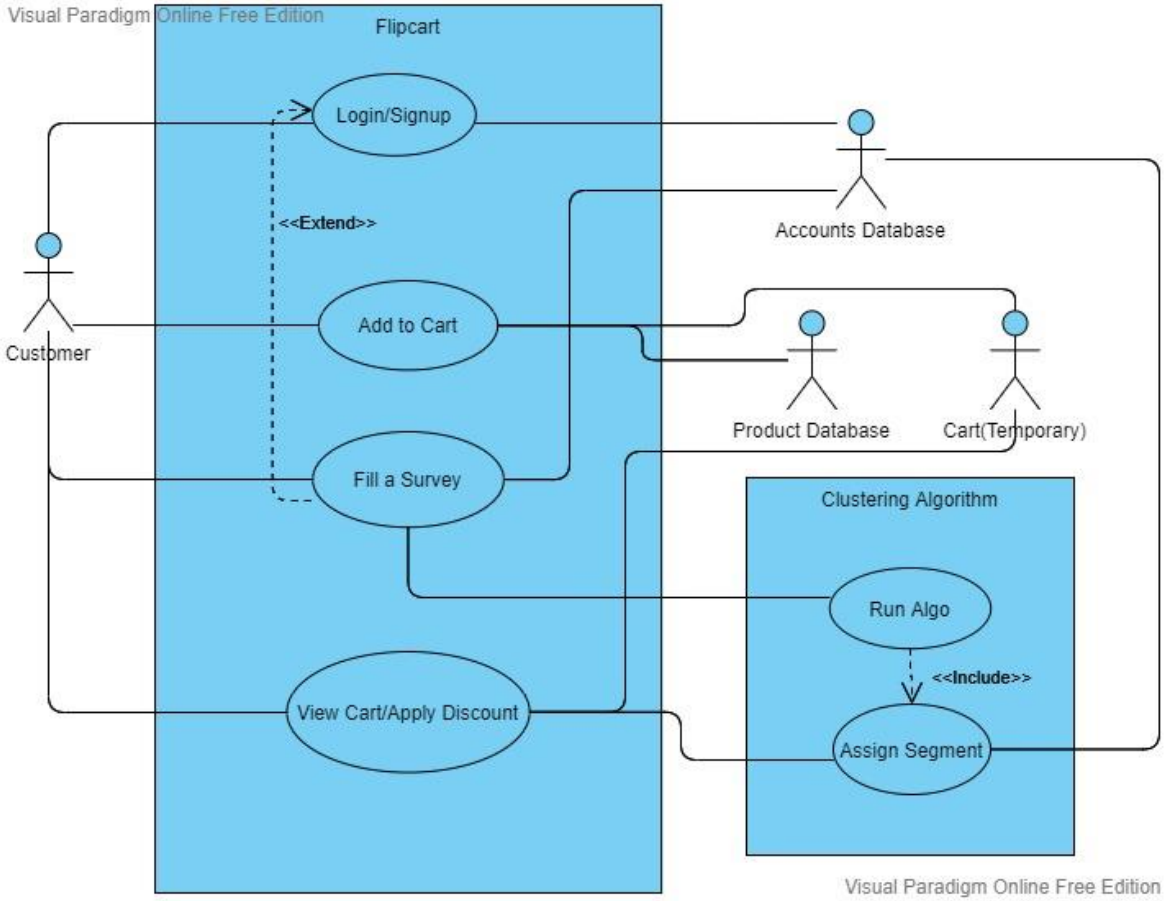


Figure 1: Use Case Diagram of the Application

2.4 Other Software Methodologies

Method	Working	Advantage	Disadvantage
Business Rule-Based [2]	Traditional targeting by filtering in Excel	Easy to apply, Use database query	Tends to rely on heuristics developed over time and slow to adapt to changes [6]
Supervised Clustering with decision tree [1]	Uses a dependent variable to predict differences in independent variables	Classify customers according to target	Use one variable to cluster
k-means clustering [1]	Uses unlabeled data to find significant number of clusters	Use any number of customer attributes	Speed of computation depends on k values
Hierarchical clustering [4]	Initially treats each observation as a cluster, then repeatedly merges two similar clusters	Relatively straightforward to program, no need to specify number of clusters required	Very high time complexity compared to k -means
PAM clustering (k-medoid) [4]	Finds a sequence of objects called medoids that are centrally located in clusters, clusters are constructed by assigning each observation to the nearest medoid.	Effectively deals with the noise and outliers present in data	Since the first k-medoids are chosen randomly, different results may be obtained on the same dataset.
k-prototype [7]	Combines working of k-means (for numerical values) and k-modes (for categorical values) algorithms	Can be applied to dataset with mixed data types, whereas k-means can be applied to only numerical data types.	Unclear what weights have to be given to categorical variables.

Table 1: Other Software Methodologies

3. System Design

3.1 Design Goals

- The system must be designed so that it is easy to use, easy to test and easy to maintain.
- The system structure's design should be in modular form and ensure loose coupling between modules and high cohesion within modules.
- The design should be able to reduce the complex connection between modules and the database.
- The data must be easily accessible to a more significant number of users.

3.2 System Architecture

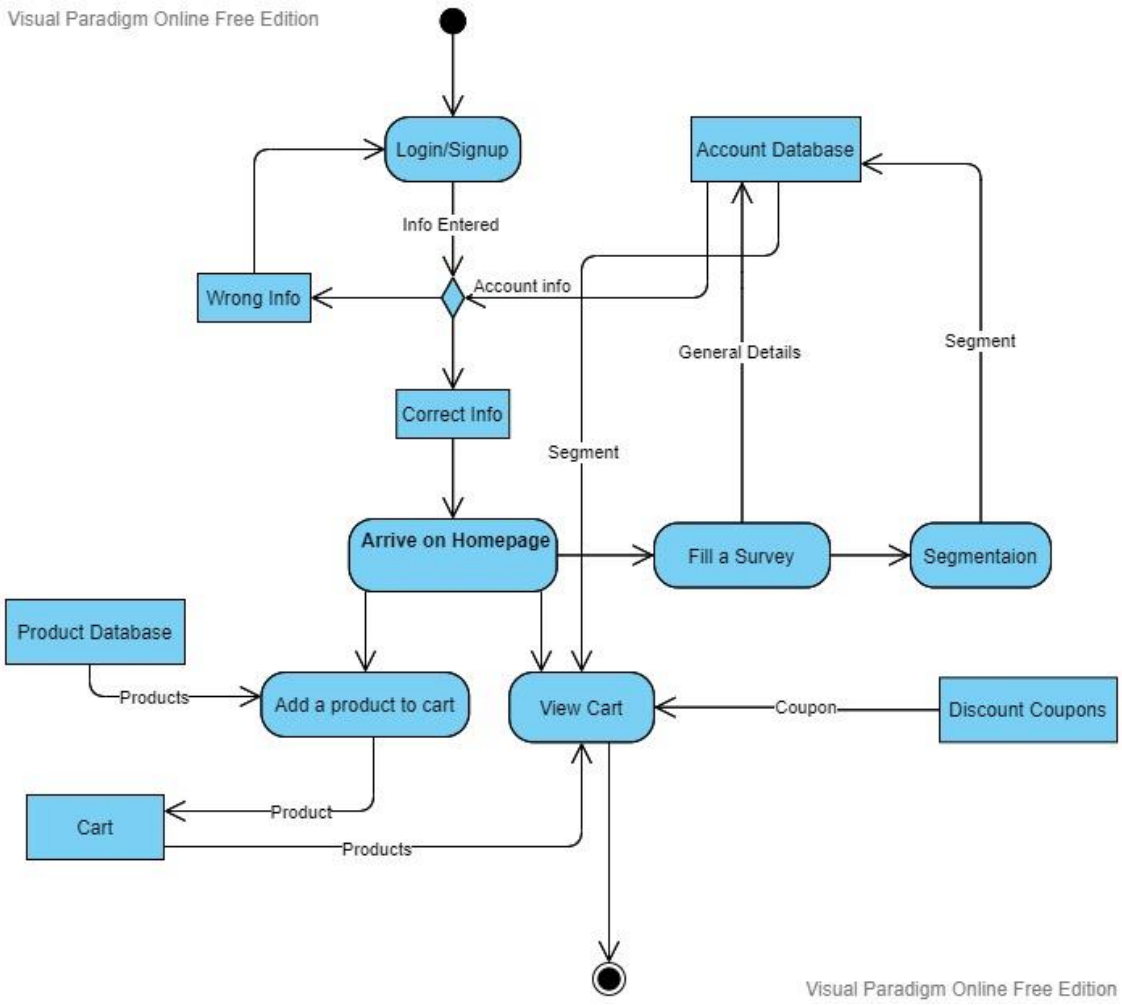


Figure 2: Project Architecture

3.3 Detailed Design Methodology

1. Data Preparation and Analysis:

    - Converting the csv file obtained into a pandas data frame.
    - Cleaning the data by removing null values and outliers.
    - Performing Exploratory Data Analysis through graphical plots using matplotlib and seaborn.
    - Using a sklearn scalar, we scale the two numeric attributes – age and dependent count so that the clustering algorithm doesn't exaggerate their importance in clustering over the non-numeric attributes.

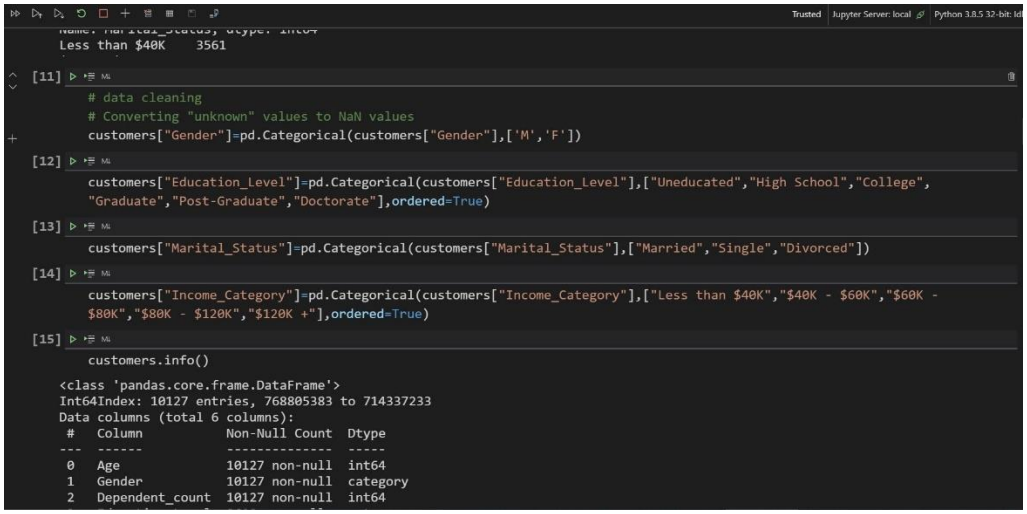


Figure 3: Converting string valued attributes to categorical type.

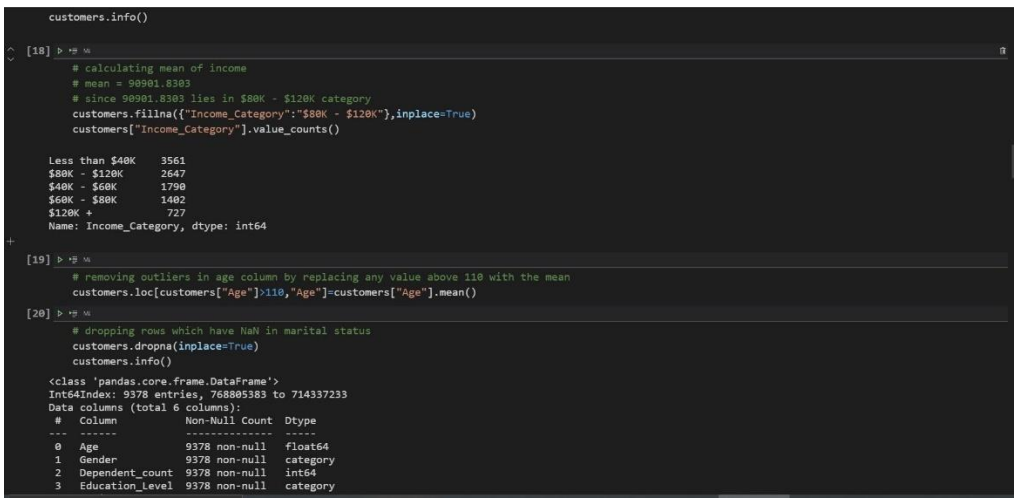


Figure 4: Removing Null values and outliers.

2. Hyperparameter tuning [8]:
    - Using the elbow method [9], we figure out the optimal number of clusters to be formed for the dataset.
    - The elbow method calculates the cost - the sum distance of all points to their respective cluster centroids – for values of k ranging from 1 to 7.
    - The ‘elbow’ lies on 3. Therefore, optimal value of ‘k’ is 3.

```
cost = []
for k in range(1,8):
        kproto=KPrototypes(n_jobs=-1,n_clusters=k,init='Cao',random_state=0)
        kproto.fit_predict(customers,categorical=[1,3,4,5])
        cost.append(kproto.cost_)
```

Figure 5. Algorithm for elbow method

3. Clustering:
    - Using the k-prototype algorithm we perform clustering on the data frame.
    - Since the elbow lies on 3, we input k as 3.
    - We create a new column called as ‘cluster’ in the data frame and assign each data entry its respective cluster, which were obtained as result from the clustering.

```
kproto = KPrototypes(n_jobs=-1,n_clusters=3, init='Cao')
clusters = kproto.fit_predict(customers2, categorical=[1,3,4,5])
```

Figure 6. Algorithm for k-prototype

4. User Interface:
    - Using Tkinter we develop a Graphical User Interface through which users will be able to:
        - Login/Signup
        - Add to cart/View cart
        - Fill a Survey – This will use the k-prototype algorithm to assign one of the 3 clusters to the user, and also provide them with a targeted discount coupon.

4. Work Done

4.1 Development Environment

The project is developed using python 3 and its libraries – Numpy, Pandas, Matplotlib, Seaborn.

For the User Interface, we use Tkinter.

The environment used was Visual Studio Code.

4.2 Results and Discussions

1) Data Frame after Pre-processing:

The data set obtained from Kaggle had 23 columns and 10127 entries (customers); after data pre-processing, we were left with 6 attributes – age, gender, income, education, number of dependents, and marital status and 9378 customers.

```
Age              9378 non-null
Gender           9378 non-null
Dependent_count  9378 non-null
Education_Level  9378 non-null
Marital_Status   9378 non-null
Income_Category  9378 non-null
```

Figure 7. Dataset after pre-processing

2) Exploratory Data Analysis:

After **Exploratory Data Analysis** we obtained the following plots and inferences:

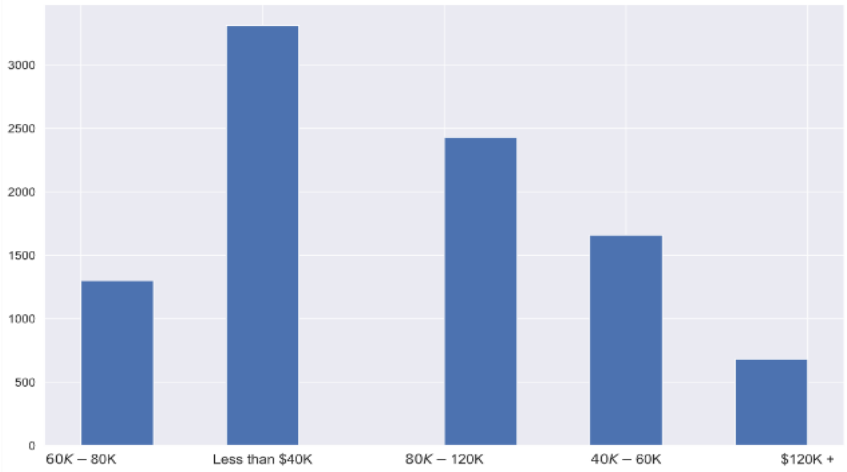


Figure 8. Histogram showing income levels of customers.

- From this we can infer that customers are more likely to earn less than $40k and least likely to earn more than $120k.

Figure 9. Scatterplot of genders against age and number of dependents.

- This graph shows that males aged 30-55 have higher dependents than any other group and that people who have 0 dependents are spread across all groups equally, i: e all ages and both genders.

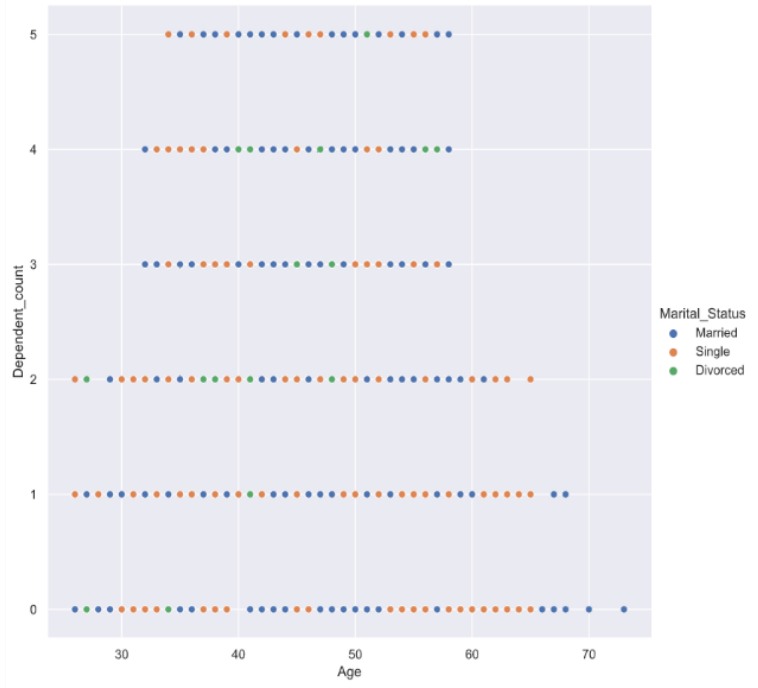


Figure 10. Scatterplot of marital status against age and number of dependents.

- We can infer that middle aged married people tend to have the highest dependents and single senior citizens tend have the lowest, as they are likely dependent on someone else.

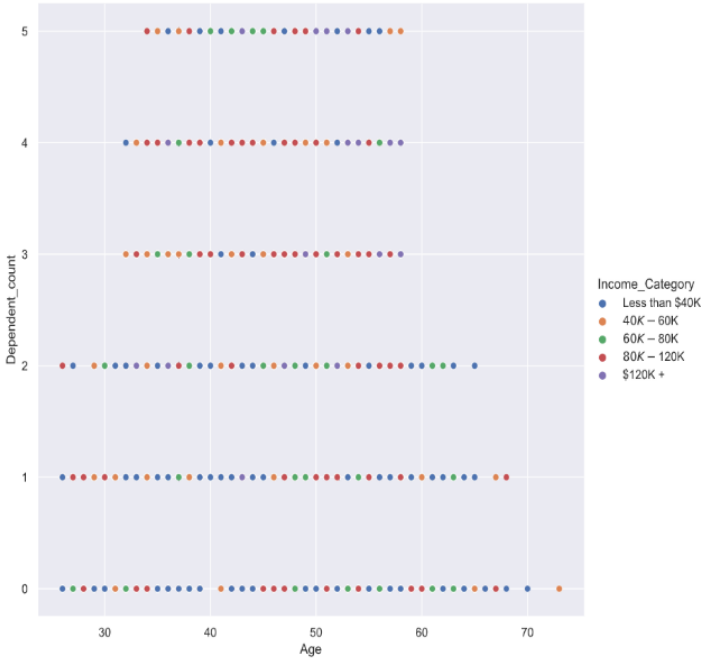


Figure 11. Scatterplot of income against age and number of dependents.

- It is evident that as the income increases, people are able to provide for more people than themselves, thus the dependent count increases irrespective of age.

3) Hyperparameter Tuning (elbow method):

From the **elbow method** we got value of k (optimal number of clusters) as 3. This can also be backed up by graphs from EDA, as we can see that there are 3 groups that share some similarities. These segments get more evident as we plot the clusters obtained after clustering.

Using matplotlib, we plot the costs against number of clusters to find the 'elbow'.

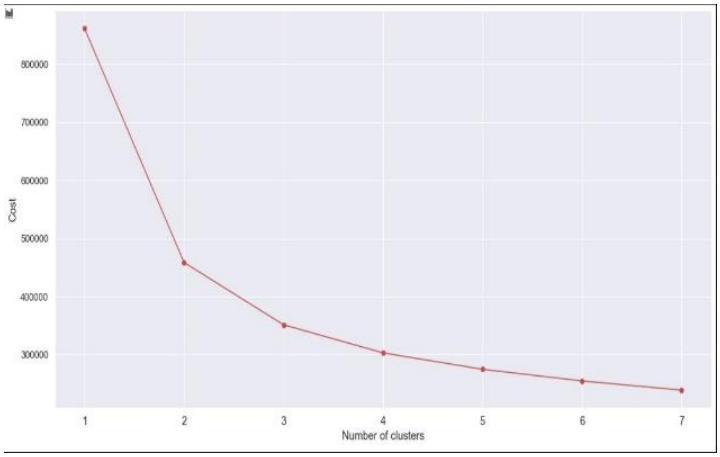


Figure 12. Elbow graph for k values

Since the 'elbow' lies at 3 number of clusters, optimal k = 3

4) Segmentation (Clustering):

We obtain the following results after clustering:

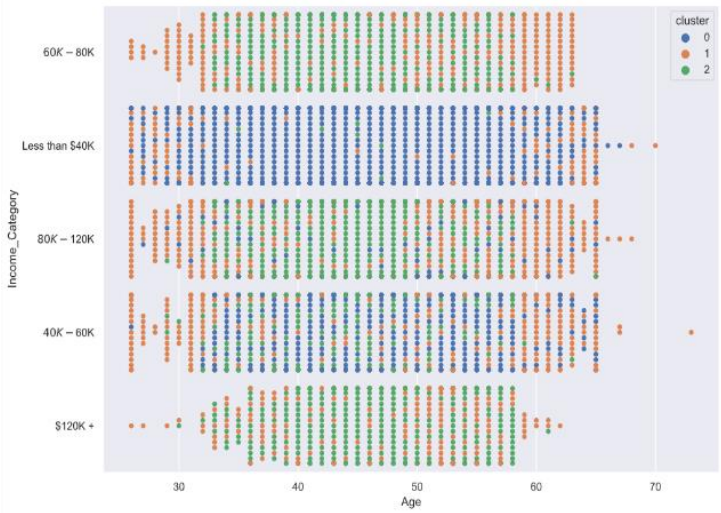


Figure 13. Scatterplot of clusters mapped on income and age.

- As we can see here segment 0 customers have low income, segment 1 customers have moderate income, and segment 2 customers have high income.

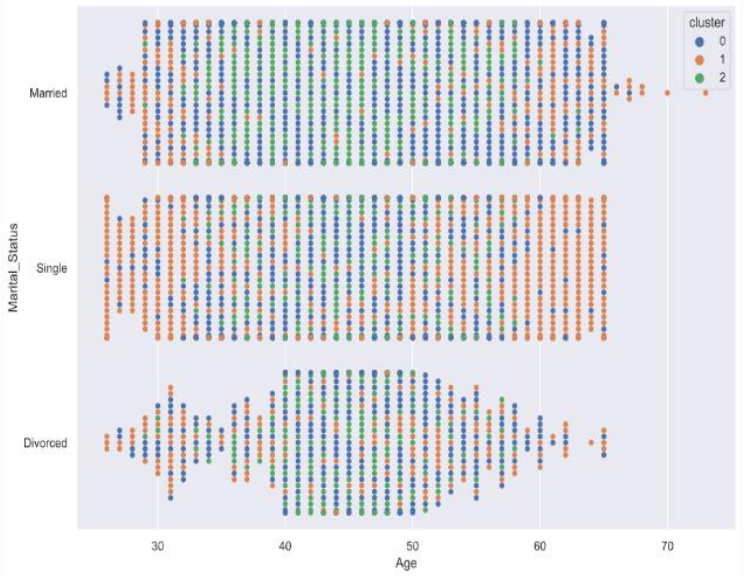


Figure 14. Scatterplot of clusters mapped on marital status and age.

- We can see that segment 1 customers lie on extreme ends of age, and therefore are more likely to be single. Whereas segment 0 and segment 2 customers are spread out evenly across married and divorced.

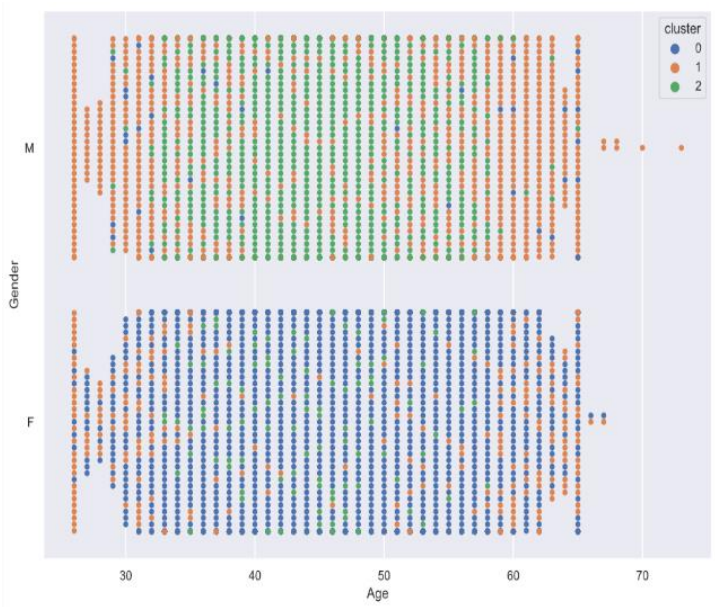


Figure 15. Scatterplot of clusters mapped on gender and age.

- This is a prime example of why demographic customer segmentation is so precise. We can clearly see a stark contrast when we map segments on basis of their gender. Segment 0 (blue) is mostly female while segment 1 and 2 almost no females, this enables us to cater to the female customers to their needs and males to theirs by targeting them with specific product recommendations.

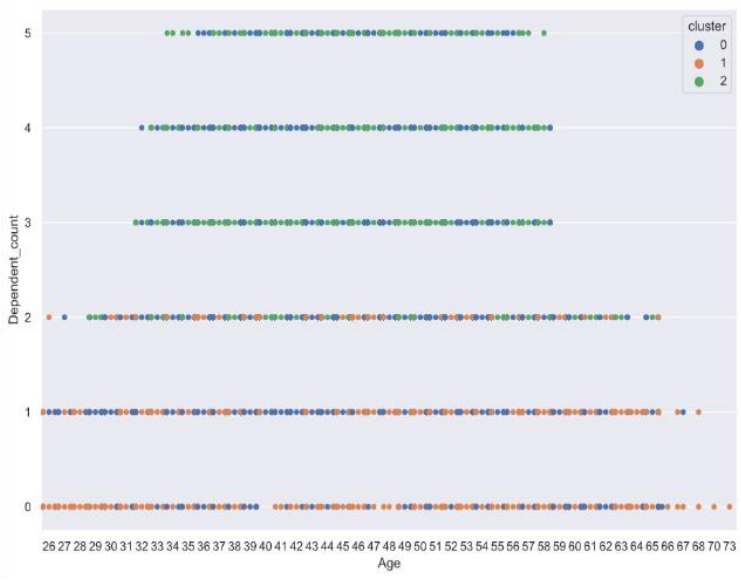


Figure 16. Scatterplot of clusters mapped on number of dependents and age.

- After looking at these graphs we can say that the algorithm has successfully segmented customers into 3 groups. These customers share a lot with their counterparts of the same group. We can conclude some attributes about the segments:
- ***Segment 0 (blue)*** has customers with low income who are very likely to be female and have at least 2-3 dependents. These customers can be categorised as ***stay at home mothers***.
- ***Segment 1 (orange)*** has customers with moderate income, no or less dependents as they are most likely single. This also agrees with their age as they lie on extremes. These customers could be ***unmarried youngsters or widowed senior citizens***, explaining their lack of dependents.
- ***Segment 2 (green)*** has customers who have a stable high income, more likely to be male, middle-aged, married and have high dependent counts. This is very fitting as these are ***middle-aged*** *men* who have secure high paying jobs who have to take care of kids, spouse and even old parents- leading to more dependents.

- As we establish the attributes of the segments, we head on to create discount coupons for **Flipcart (GUI)** tailored to them predicting their purchasing behaviour and spending power.

5) GUI (Flipcart):

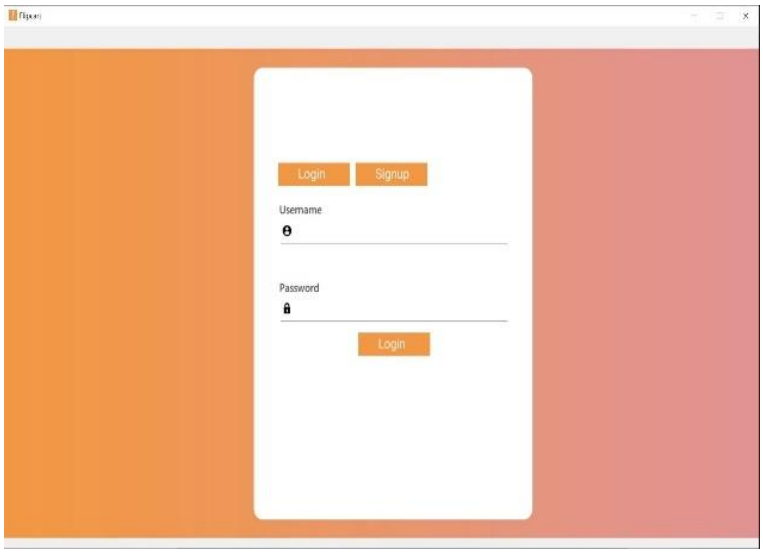


Figure 17. Flipcart's login screen

- After logging in the users land on the products screen which contains 9 categories of products – appliances, clothes, cosmetics, electronics, groceries, medicines, stationery, toys, vehicles.

Figure 18. Flipcart's survey screen

- Upon registration of a new account the user is asked to fill a survey, for which they will be rewarded a discount coupon. The survey asks for their age, gender and the other 4 attributes and predicts which customer segment they belong to.

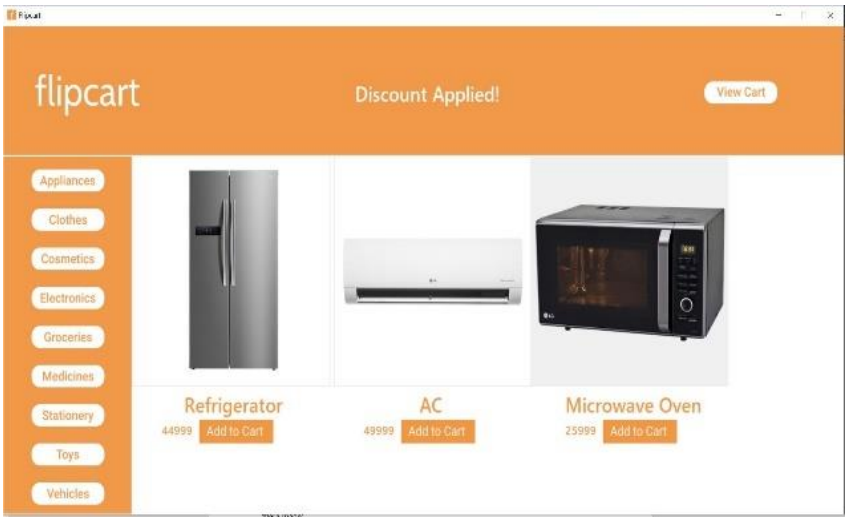


Figure 19. Flipcart's product screen

- To utilise the complete power of market segmentation it is essential to know how to market to those segments. After forming the segments and assessing their living conditions, the following discount coupons were made:
- For Segment 0 - ***"15% off on Cosmetics, Clothes, Groceries"*** as stay at home mothers are more likely to buy those items.
- For Segment 1 - ***"10% off on Stationery, Electronics, Medicines"*** as young students are more likely to buy stationery and electronics and senior citizens are more likely to buy medicines.
- For Segment 2 - ***"5% off on Appliances, Vehicles, Toys"*** as middle-

aged married people tend to buy home appliances, vehicles, and toys for their kids.

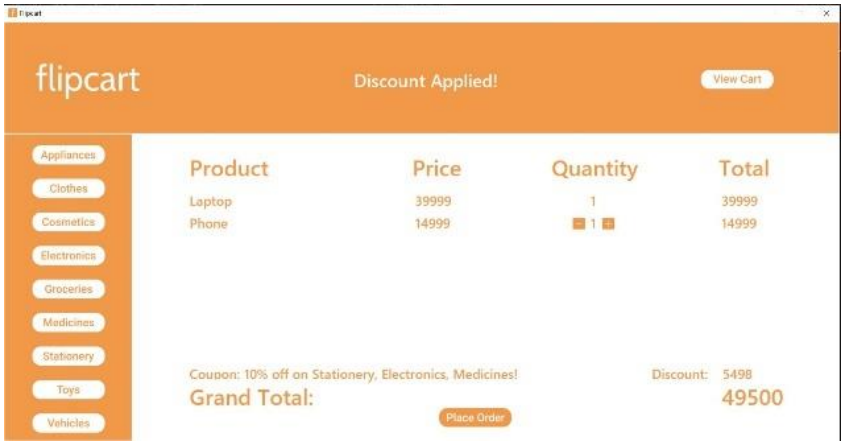


Figure 20. Flipcart's view cart screen

4.3 Individual contribution of project members

The project was done by a single individual.

5. Conclusion and Future Plans:

In this paper, we discussed how customer segmentation has taken over marketing in past decade, why it is as popular it is, and why it works so well.

We obtained a data set of customers of a bank, cleaned it, and used a machine learning algorithm to segment those customers based on their demographic attributes – age, gender, income, education, marital status, and number of dependents. There were 3 segments formed – mothers, students and senior citizens, parents. This segment is distinct from each other but customers belonging to the same segment are similar to each other. This shows that the unsupervised clustering algorithm k-prototype worked well with the data set.

Flipcart, an ecommerce desktop application was created which mimicked how giants like flipkart and amazon use market segmentation to their advantage. The application offers customers targeted discounts if they fill a survey, which in turn clusters them in one of the three segments.

Thus, we can conclude that with advent of social media and big data, customer segmentation will prevail even more and new forms will emerge as technology advances.

The work done on this research was only restricted to one dataset and one algorithm. Since there are many types of companies which offer different types of services, their customers and their attributes might differ from the one we have chosen. And since there

are several machine learning algorithms suitable for customer segmentation the outcomes might differ with the algorithm we have used.

This gives scope to future study on the comparison of efficiency of different machine learning algorithm applied on various types of datasets.

There is also scope to improve the Flipcart application by improving the UI and adding functionalities like multiple carts.

6. References

[1] J. Sari and L. Nugroho and R. Ferdiana and P. Santosa, Review on Customer Segmentation Technique on Ecommerce. 1st ed. Indonesia: American Scientific Publishers, 2011.

[2] Elizabeth, Customer Segmentation: Rules-based vs. K-Means Clustering. 1st ed. Denmark: d3mlabs, 2019.

[3] C. Iyim, Customer Segmentation with Machine Learning. 1st ed. towards data science, 2020

[4] S. Kamande and K. Mirthi and E. Ahishakiye, Consumer Segmentation and Profiling using Demographic Data and Spending Habits Obtained through Daily Mobile Conversations. 1st ed. International Journal of Computer Applications, 2018.

[5] Syviaw. "BankChurners DataSet", https://www.kaggle.com/syviaw/bankchurners (February 25th,2021)

[6] Elizabeth, Customer Segmentation: Rules-based vs. K-Means Clustering. 1st ed. Denmark: d3mlabs, 2019.

[7] Audhi Aprilliant. "The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical)" https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb (June 4th,2021)

[8] Jeremy Jordan. "Hyperparameter tuning for machine learning models.", https://www.jeremyjordan.me/hyperparameter-tuning/ (June 4th,2021)

[9] "Elbow method (clustering)" Wikipedia, Wikimedia Foundation, 11 December 2020, en.wikipedia.org/wiki/Elbow_method_(clustering)