# SELECTING THE TARGET AUDIENCE USING SAMPLING TECHNIQUES

*A data science project report of the course "Independent Study in Data Science - (DSC 3263)" presented by*

**PERERA, A.P.A.C. (S/17/445)**

**RATHNAWEERA, R.P.R.M. (S/17/465)**

**WEERASINGHE, W.G.K.D. (S/17/509)**

*With the supervision of the company,*

**ZONE 24x7**

**DEPARTMENT OF STATISTICS & COMPUTER SCIENCE**
**FACULTY OF SCIENCE**
**UNIVERSITY OF PERADENIYA**
**SRI LANKA**
**2023**

# DECLARATION

I hereby declare that the Project Summary Report entitled **"Selecting The Target Audience Using Sampling Techniques"** is an authentic record of our own work as a requirement of the three-months project under the course of '**Independent study in Data Science (DSC3263)**' during the period from 28/11/2022 to 31/03/2023 for the award of the degree of B.Sc. Honors in Data Science from Department of Statistics and Computer Science Faculty of Science University of Peradeniya, under the guidance of Dr.Sachith Abeysundara, Prof. Roshan D. Yapa, Miss. B.R Pavithra M. Basnayake and Mr.Prasan Rathnayake.

W.G.K.D. Weerasinghe
S/17/509

Date: 31/03/2023

## Certified by:

1. **Supervisor:** Mr. Prasan Rathnayake

Signature: ______________________ Date: ______________________

2. **Head of the Department:** Dr. Sachith Abeysundara

Signature: ______________________ Date: ______________________

**Department Stamp:**

# ABSTRACT

This project focuses on the importance of identifying the target audience for businesses to succeed in the market using the most effective sampling method. It is crucial for companies to understand the desires, core values, and preferences of their target market to avoid the grave mistake of trying to be everything to everyone. A personalized experience is expected by the majority of customers, and the proper distribution of the target audience ensures the efficiency of the company with minimum time and money. This project aims to identify the target audience by considering variables such as age, gender, and location, including city, state, and region. Involving customers in the early stages of product development can help companies identify key design and functional features, increase customer retention, and ultimately improve their reputation. Clustering and sampling techniques, including simple random sampling, stratified sampling, cluster sampling, and systematic sampling, will be used to determine the target audience accurately. Non-probability sampling techniques will not be used in this project. Because they are associated with case study research design and qualitative research which tend to focus on small samples and are intended to examine a real life phenomenon, not to make statistical inferences.

# ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my project supervisors Mr. Prasan Rathnayake, Prof. Roshan D. Yaya and Miss.Pavithra Basnayake for their invaluable guidance and support throughout this project. Their dedication and expertise have been instrumental in shaping my project to its form. I would also like to extend my thanks to all the lecturers of the Department of Statistics and Computer Science, University of Peradeniya for their great support which have been critical to the success of this project.

Finally, I would like to express my sincere appreciation my group members Mr. Ravindu Rathnaweera and Mr.Avishka Perea for support and dedication throughout the project. Without their contributions, this project would not have been possible.

Once again, thank you to everyone who has contributed to this report in one way or another. Your efforts and support have been invaluable and I am deeply appreciative of everything you have done.

# Table of Contents

# List of Figure

# List of Tables

# CHAPTER 01

# BACKGROUND

## 1.1. Background of the Company

Zone 24x7 is headquartered in San Jose, California, Zone24x7 Inc is a digital transformation partner to a wide range of organizations from technology startups to Fortune 500 organizations. Established in 2003 by founder Llavan Fernando and CFO Saw-Chin Fernando, the company has successfully served leading Tier 1 retailers and hi-tech companies offering best-in-class technology solutions backed by a strong commitment to customer satisfaction.

In 2004, Llavan and Saw-Chin founded Zone24x7 (Private) Limited, an advanced technology center in Colombo, Sri Lanka as a separate entity. The advanced technology center offers highly skilled engineering specialists and related technology services to Zone24x7 Inc.

Zone24x7 specializes in offering end-to-end technology consulting and engineering services encompassing both hardware and software. The services portfolio includes Enterprise Software Applications, Big Data & Data Science Engineering, Embedded Systems Engineering, Remote Monitoring & IoT, Machine Learning, Cognitive Vision, Robotics, and Innovation Services with Technology Proof of Concept Development. Zone24x7 promotes a culture of customer-centric innovation, continuous learning, professionalism, caring for oneself and others, and integrity to create a diverse, inclusive, and thriving workplace for all its associates.

## 1.2. Organizational Structure

Zone24x7 employs a functional organizational structure that prioritizes cross-functional cooperation. The company is divided into various departments, each responsible for a specific field of expertise, such as software development, research and development, and quality assurance. The significant departments at Zone24x7 include Engineering, Research and Development, Quality Assurance, Sales and Marketing, and Human Resources.

The Engineering department handles software development, testing, and deployment of the company's solutions. The Research and Development department is responsible for exploring and developing emerging technologies like artificial intelligence, machine learning, and robotics. The Quality Assurance department ensures that all products and solutions adhere to the company's stringent standards of quality and dependability. The Sales and Marketing department is tasked with promoting and selling the company's products and services to customers, while the Human Resources department manages the company's talent acquisition, employee engagement, and professional development programs.

Zone24x7 also has a Project Management Office (PMO) that supervises the delivery of projects and ensures that they are delivered on time, within budget, and to the client's satisfaction. The company's dedicated innovation lab focuses on exploring new technologies and developing innovative solutions for clients. Zone24x7's functional structure allows the company to effectively deliver high-quality products and solutions to its clients. Cross-functional cooperation and innovation are at the forefront of the company's focus, which enables them to stay ahead of emerging technologies and meet the changing needs of their clients.

## 1.3. Overview of the project

### 1.3.1. Scope

Finding the right target audience for a project is crucial for its success. By using sampling techniques, we can carefully select a group of individuals that are representative of the larger population we are interested in. This allows us to gain valuable insights into the preferences and behaviors of our target audience, which we can use to tailor our project to their needs and interests. For example, we can modify or introduce products according to the specific needs, behaviors, and concerns of different types of customers.

Not only is this important for effectively reaching and engaging our audience, but it also increases customer satisfaction, which is crucial for a business. As customer satisfaction increases, the company can gain a good reputation within the market and ultimately make it easier to generate profits by attracting more customers.

### 1.3.2. Objective

Our main objective is to use sampling techniques on data and see which sampling technique would determine the target audience more accurately. Since this area has a very wide scope as a group we will be focusing on some certain areas in this particular project. We are more into carrying out a customer purchase patterns analysis in order to determine the target audience. When it comes to this kind of an analysis, we have a limitation as to how the feedback is being collected in the dataset. For an instance it could be an in-shop customer feedback or an online customer feedback collection. Furthermore, when it comes to sampling techniques, we will be using 4 techniques in order to determine the best one. Apart from that, the results which we gain could only be valid to the particular dataset which we have used in this project. For example, another dataset from another domain might suggest some other particular sampling technique to be used in determining the target audience.

### 1.3.3. Timeline

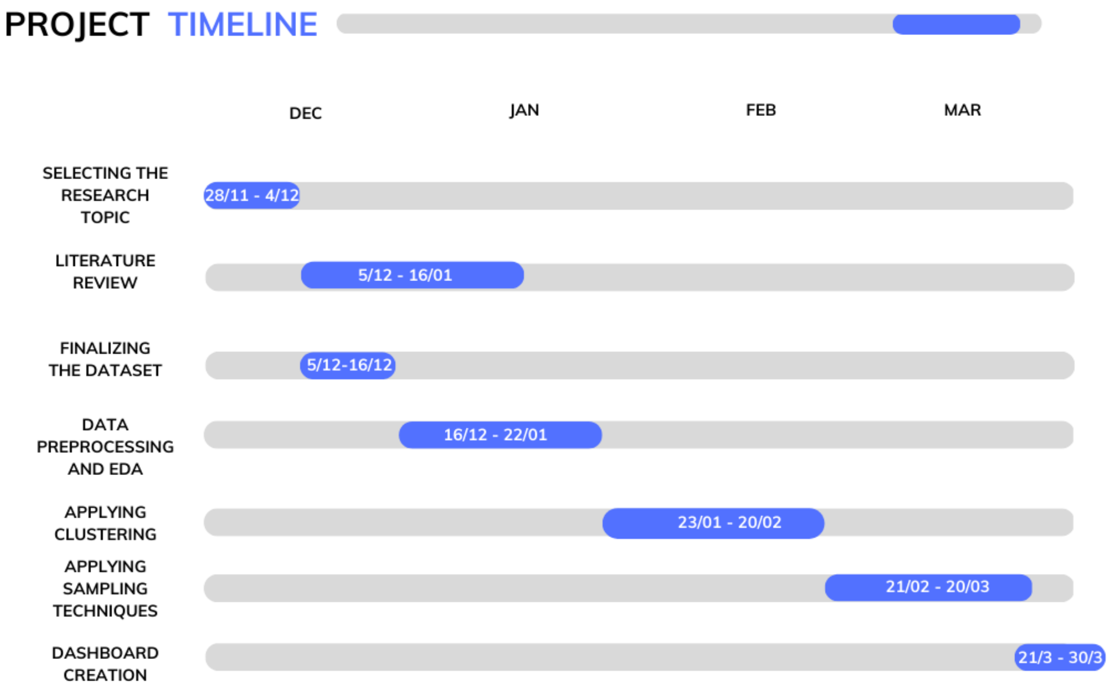

*Figure 1: Project Timeline*

### 1.3.4. Deliverables

Project benefits

- Find the gaps between the markets
- Assess the viability of the product or service
- Better understanding of the customer needs and wants
- Assist in creating a marketing campaign for the company
- Reduced in marketing costs for the company
- Being able to utilize the company resources properly
- Increased customer satisfaction
- Good reputation for the company
- Extra support for the workforce
- Improve business strategy

## 1.4. Team and Contribution

### 1.4.1. Project Team

S/17/445
A.P.A.C. Perera

S/17/465
R.P.R.M. Rathnaweera

S/17/509
W.G.K.D. Weerasinghe

### 1.4.2. Individual Contribution

Mainly I have done the following:

- Preparation of project proposal
- Searched for datasets
- Done some parts in literature review
- Initial EDA to check whether the selected dataset can be used for analysis or not
- Full EDA on dataset
- Worked on K-means clustering
- Worked on K-modes clustering
- Applied Cluster sampling technique
- Contributed to create the dashboard
- Preparation of project final report

Other than these, I have worked on combining the parts that we have done individually to a single result.

# CHAPTER 02

# INTRODUCTION & LITERATURE REVIEW

## 2.1. Introduction to the project

Trying to be everything to everyone is one of the gravest mistakes any business can make. Furthermore, failure to understand the desires, core values and preferences of your target market can backfire tremendously. It is being reported that around 68 percent of customers expect all experiences to be personalized. Therefore, it is crucial that every company should have their target audience identified if they want to perform better as a company within the market.

Here our main idea is to identify which sampling technique is the best in determining the target audience of the dataset called “Online Sales in USA” which was selected for our study from Kaggle. The dataset is about the online sales made in United States of America for the year 2020-2021. It has 36 variables and 286000+ observations. When it comes to sampling techniques will be using four sampling techniques to choose the best sampling method. They are,

- Simple Random Sampling
- Stratified Sampling
- Cluster Sampling
- Systematic Sampling

## 2.2. Literature Review

There is large amount of data being spread in the present world. but the information does not reach the correct audience. Proper distribution of the target audience ensures the efficiency of the company, the minimum time and money.(Smirnova, 2020). These advantages push companies to pay more attention on the specific customer wants and needs. In order to do that there are several factors that need to be considered. Such as Customer demographics (i.e., Age, Gender, Education), Geographic (i.e., city, state) and Psychographic (i.e., Psychological and lifestyle features.). In this project we are making sure that above mentioned variables are also available in our dataset to carry out a successful project. We are mainly focused on the variables such as age, gender and location variables (i.e., city, state, region) in our project.

The other best thing about target audience is that they can also be involved in product development process to develop products based on customer preferences so that there will be huge sales when they are being sold in the market. Involving consumers in the early stages of product development can help companies identify the key design and functional features of a product from the consumer's perspective. (Leahy, 2013) Educated consumer involvement ensures that the product being developed is useful, needed, and wanted by consumers at a price they are willing to pay. (Leahy, 2013)This clearly suggests that involvement of customer would definitely make them feel valued which is really important when it comes to customer retention. This would ultimately increase the company reputation as well. Clustering would also be used prior to applying sampling techniques. Having similar data points within the same cluster helps the particular company to use targeted marketing.

Sampling techniques are methods of selecting a subset of individuals or units from a population of interest for research purposes. Sampling techniques can be classified into two broad categories: probability and non-probability sampling. Probability sampling methods ensure that every element in the population has a known and non-zero chance of being included in the sample, whereas non-probability sampling methods do not have this property. Probability sampling methods are generally preferred for conducting systematic literature reviews, as they allow researchers to make valid inferences about the population based on the sample (Berndt, 2020)

Among the various probability sampling methods, four common ones are simple random sampling, stratified sampling, cluster sampling and systematic sampling. Simple random sampling involves selecting a sample from the population by using a random mechanism, such as a random number generator or a lottery. This method ensures that every element has an equal chance of being selected and that the sample is representative of the population. However, simple random sampling may not be feasible or efficient when the population is large or dispersed, or when some subgroups of interest are rare or overrepresented (*Sampling Methods | Types, Techniques & Examples*, n.d.)

Stratified sampling involves dividing the population into homogeneous groups or strata based on some relevant characteristics, such as age, gender, location, etc., and then selecting a simple random sample from each stratum. This method ensures that the sample reflects the proportion and diversity of the population and that each stratum is adequately represented. However, stratified sampling requires prior knowledge of the population characteristics and may not be applicable when there are too many strata or when some strata are very small (*Sampling Methods | Types, Techniques & Examples*, n.d.)

Cluster sampling involves dividing the population into heterogeneous groups or clusters based on some geographical or administrative criteria, such as regions, districts, schools, etc., and then selecting a simple random sample of clusters and including all or a subset of elements within each cluster in the sample. This method reduces the cost and complexity of sampling when the population is large or dispersed and allows researchers to study the variation within and between clusters. However, cluster sampling may introduce bias and error if the clusters are not representative of the population or if there is high intra-cluster correlation (*Sampling Methods | Types, Techniques & Examples*, n.d.)

Systematic sampling involves selecting a sample from the population by using a fixed interval or a skip pattern, such as every 10th or 20th element. This method is simple and convenient to implement and can produce a representative sample if the population is ordered randomly or cyclically. However, systematic sampling may introduce bias and error if the population is ordered in a way that creates a periodic pattern that coincides with the sampling interval (*Sampling Methods | Types, Techniques & Examples*, n.d.)

Non probability sampling is often associated with case study research design and qualitative research. Case studies tend to focus on small samples and are intended to examine a real life phenomenon, not to make statistical inferences in relation to the wider population.(Taherdoost, 2016) Therefore Non Probability sampling techniques would not be used in this project.

The choice of sampling method depends on several factors, such as the research objectives, questions, design, budget, time frame, availability of data sources, etc. Researchers should consider the advantages and disadvantages of each method and justify their choice based on their research context and criteria. Moreover, researchers should report their sampling methods in a transparent and detailed manner to enable readers and reviewers to understand and evaluate their review results (Berndt, 2020)

# CHAPTER 03

# METHODOLOGY

## 3.1. Data Collection

For this project, the data set was found on Kaggle, a popular platform for finding and sharing datasets. The dataset contains information on online sales in the USA in 2021, including order status, geographical information about customers, their spending habits, payment methods, and other relevant details.

The data was collected from an online sales dataset of various products, including several merchandise and electronic items across different states in the USA. With the rise of online shopping, retailers are increasingly interested in exploring ways to increase their profits. Therefore, large retailers are utilizing techniques such as sales analysis to gain insights into their customers' purchasing behavior and patterns. One important method used in sales analysis is market basket analysis, which involves examining collections of items to identify relationships between items that are frequently purchased together within the business context. The dataset was collected in order to apply market basket analysis to the sales data and identify any meaningful associations between different products. The data set is reliable and accurate, as it was collected using standardized methods and procedures.

In conclusion, the use of the online sales dataset from Kaggle was a valuable resource for this project. The dataset provided detailed and comprehensive information on online sales in the USA, which allowed for a thorough analysis of the topic. The careful collection, cleaning, and analysis of the data set helped to ensure the accuracy and reliability of the research findings.

## 3.2. Data pre-processing

Data pre-processing is an essential step in any data analysis project as it helps ensure that the data is of good quality, consistent, and suitable for analysis. In this section, we will describe the various steps taken to pre-process the data used in our analysis.

The first step we took was to check for null values in the dataset. Null values can significantly affect the analysis by reducing the number of observations available for analysis. Therefore, we carefully examined the data for null values and took appropriate steps to handle them. We did not have to drop or replace any missing values since there were no missing values in the dataset.

Next, we checked for duplicate order IDs in the dataset. Duplicate records can cause discrepancies in the analysis and result in an overestimation of the sample size. Therefore, we thoroughly scanned the dataset for duplicate records. There were no any duplicates. Following this, we checked the number of unique customers in the dataset. It is essential to know the number of unique customers as it can help in identifying customer trends and patterns. The number of unique customers in the dataset helps in estimating the total number of customers who made online orders, and it provides a more accurate picture of the customer base.

In addition, we also checked the number of customers who responded to discounts. This information is critical for analysis as it helps them in identifying the effectiveness of their discounting strategies. Knowing the proportion of customers who respond to discounts can help the business in making informed decisions about their pricing strategies and promotions. Finally, we checked the number of customers who completed the orders after receiving discounts. This information is important as it helps in identifying the effectiveness of discount strategies in generating sales. It also helps us in identifying customer behavior patterns and preferences.

Based on the above results, we decided to take the dataset into our analysis. We were satisfied that the data was of good quality and suitable for analysis. However, it is important to note that data pre-processing is an ongoing process, and we continued to monitor the data throughout the analysis to ensure that the data remained consistent and reliable. The pre-processing steps we took in this project helped us in obtaining accurate and reliable results, which can help us in making informed decisions about their strategies and operations.

## 3.3. Data Analysis

### 3.3.1. Exploratory Data Analysis

The exploratory data analysis is a critical component of any research study. It is the initial step that researchers take to gain insights into the data, identify trends and patterns, and develop a deeper understanding of the relationships between variables. In this section, we present the EDA that was performed on the dataset using Python in Google Colab, with the help of various libraries such as Pandas, NumPy, Plotly, Seaborn, and Matplotlib.

To begin with, we started by checking for the correlation between each type of variable present in the dataset. The aim was to identify any variables that were highly correlated with each other, as this could lead to multicollinearity issues during modeling. To achieve this, we used the Pandas and NumPy libraries to create correlation matrices, which were then visualized using heatmaps. The theoretical background for this approach is that correlation measures the degree to which two variables are related to each other. A correlation coefficient of +1 indicates a perfect positive correlation, while a coefficient of -1 indicates a perfect negative correlation.

Next, we analyzed the customers' data by gender, state, region, and city. This was done to identify any demographic patterns that could influence customer behavior, such as purchasing habits, preferences, and loyalty. We used the Pandas and Seaborn libraries to create bar charts and frequency tables to visualize the distribution of customers across different demographic categories. This approach is to find that demographic factors can have a significant impact on consumer behavior, as they may influence the needs, wants, and preferences of customers.

We also checked for the age distribution and average customer age in each category. This was done to identify any age-related patterns in customer behavior, such as purchasing power, decision-making, and brand loyalty. We used the Pandas and Matplotlib libraries to create histograms and box plots to visualize the distribution of customers by age. The theoretical background for this approach is that age can be an important factor in consumer behavior, as it can affect the perception of value, quality, and price.

Furthermore, we checked for the order count by order status. This was done to identify any issues related to order fulfillment, such as delays, cancellations, or returns. We used the Pandas and Plotly libraries to create stacked bar charts to visualize the distribution of orders by status. Order fulfillment is a critical aspect of customer satisfaction, and any issues related to this can lead to negative customer experiences. Additionally, we checked the sales in each state to identify any geographical patterns in customer behavior, such as regional preferences, buying power, and market saturation. We created choropleth maps to visualize the distribution of sales across different states. Geography can be an important factor in consumer behavior, as it can influence the availability of products, services, and information.

We also analyzed the sales distribution within the period to identify any temporal patterns in customer behavior, such as seasonality, trends, and fluctuations. We created line charts and time-series plots to visualize the sales over time. The time can be an important factor in consumer behavior, as it can affect the demand for products, services, and information. Finally, we checked for insights into revenue, discounts, and interest in categories with regard to the appliances, gender, and geo-locations. This was done to identify any cross-sectional patterns in customer behavior, such as purchasing habits, preferences, and loyalty.

### 3.3.2. Clustering

Clustering is a powerful technique used in data analysis and machine learning to group similar objects or data points together. It is an unsupervised learning method that involves partitioning data into groups or clusters based on their similarity. Clustering is widely used in various fields, including data mining, pattern recognition, image processing, and bioinformatics. It is a valuable tool for exploratory data analysis, allowing researchers to uncover hidden patterns and relationships in data. In this section, we will describe the clustering technique used in our study and how it was applied to our dataset to gain insights into customer behavior and purchasing patterns.

Mainly we have used k-means clustering as our clustering algorithm. K-means clustering is a widely used technique in data analysis and machine learning for partitioning data into groups or clusters based on their similarity. It is an unsupervised learning method that involves assigning each data point to one of K clusters, where K is a predetermined number of clusters. The goal of k-means clustering is to minimize the sum of squared distances between data points and their assigned cluster centroids. In this section, we will describe the k-means clustering algorithm and its application to our dataset.(*K-Means Clustering Algorithm in Python - The Ultimate Guide*, n.d.)

We had both numerical and categorical variables in our dataset. K-means algorithm works only on numerical data. So, we had to encode categorical variables to numerical. We used one hot encoding technique for that. One-hot encoding is a common technique used to convert categorical variables into numerical variables, which can be used as input in various machine learning algorithms, including clustering. The basic idea behind one-hot encoding is to create new binary variables for each category in the categorical variable. One-hot encoding will create k binary variables, where each variable represents a category. For each observation in the dataset, only one of the binary variables will have a value of 1, and all other binary variables will have a value of 0. This allows us to represent each category in the categorical variable as a unique numerical value.

Our next step was determining the optimal number of clusters. We used ‘Elbow method’ which helps to identify the number of clusters by looking for the "elbow" or bend in a plot of the within-cluster sum of squares (WCSS) against the number of clusters. The within-cluster sum of squares is a measure of the total distance between each data point and its assigned cluster centroid. The idea behind the elbow method is to plot the WCSS against the number of clusters and look for the point where the rate of decrease in WCSS starts to level off. This point is often referred to as the "elbow" point, and it corresponds to the optimal number of clusters for the dataset.
To apply the elbow method, we first run k-means clustering for a range of values of k and calculate the WCSS for each value of k. We then plot the WCSS against the number of clusters k and look for the point where the rate of decrease in WCSS starts to level off. This point is typically chosen as the optimal number of clusters for the dataset.

Then we applied k-means clustering. The k-means clustering algorithm can be expressed mathematically as follows:

Given a dataset X = {x1, x2, ..., xn}, where xi is a d-dimensional vector representing the i-th data point, the goal is to partition the dataset into K clusters. Let C = {c1, c2, ..., cK} be the set of centroids representing the K clusters. The objective function of k-means clustering is to minimize the sum of squared distances between data points and their assigned cluster centroids, which can be expressed as:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{K} ||x^{i} - \mu_{j}||^{2} \tag{1}$$

The k-means algorithm is widely used in various applications, including customer segmentation, market analysis, and image processing. In our study, we applied the k-means algorithm to our dataset of online sales in the USA to identify customer segments based on their purchasing behavior. We first preprocessed the data by removing null values, checking for duplicate order IDs, and selecting relevant features such as customer ID, product category, and purchase amount.

Apart from K-means clustering, we also used K-modes clustering to cluster the dataset based on the category of items. But that result was not used much in our analysis. Just extracted a dataset after clustering and applied cluster sampling. K-modes is a clustering algorithm that is similar to k-means but is specifically designed for categorical data. Unlike k-means, which is designed for continuous data, k-modes operates on categorical data by clustering objects into groups of similar categorical values. It is commonly used in data mining and machine learning to identify patterns in categorical data and to group similar data into clusters.

K-modes is an iterative algorithm that starts by randomly selecting k initial cluster centroids. The algorithm then assigns each object to the nearest centroid based on the distance between the categorical values. The distance between two categorical values is measured using a simple matching distance function, which counts the number of mismatches between the categories. Once the objects have been assigned to their nearest centroid, the algorithm updates the centroids by selecting the mode of each categorical attribute in each cluster. The mode is simply the most frequent value of a categorical variable in a cluster.

The k-modes algorithm continues to iterate until the centroids no longer change significantly or a maximum number of iterations is reached. The final centroids represent the optimal k clusters for the categorical dataset.

### 3.3.3. Sampling Techniques

The process of selecting a sample is known as sampling. The quality of the results obtained from a research study depends on the quality of the sample selected. Therefore, it is important to use appropriate sampling techniques to ensure that the sample is representative of the population and the results obtained from the sample can be generalized to the population. In this section, we will describe the sampling techniques that were used in this study to select the sample for analysis. First, we determined the sample sizes for different confidence intervals (90%, 95%, 99%) using following formula.

Infinite Population:

$$n = \frac{z^2 \times \hat{\rho}(1 - \hat{\rho})}{\varepsilon^2} \tag{2}$$

Finite Population:

$$n' = \frac{n}{1 + \frac{z^2 \times \hat{\rho}(1 - \hat{\rho})}{\varepsilon^2 N}} \tag{3}$$

where:

- n is the sample size
- z is the z-score
- p̂ is the population proportion
- ε is the margin of error
- N is the population size

We have applied four probability sampling methods. They are Simple random sampling, Systematic sampling, Cluster sampling and Stratified sampling. Simple random sampling is a popular probability sampling technique in which each member of a population has an equal chance of being selected for the sample. This technique is widely used in research studies because it is easy to implement and produces a sample that is representative of the population. In simple random sampling, each member of the population has an equal probability of being selected, which means that the probability of selecting any individual member is the same as the probability of selecting any other individual member. The sample obtained using simple random sampling is unbiased and representative of the population, which makes it an ideal technique for our study.(Taherdoost, 2016)

Then we used systematic sampling. Systematic sampling is a probability sampling technique that involves selecting members of a population at a fixed interval. In systematic sampling, the first member of the sample is selected randomly from the population and then every nth member is selected after that, where n is the interval between selections. This technique is widely used in research studies when the population is too large to select a simple random sample or when a systematic pattern is present in the population.(*Sample Selection in Systematic Literature Reviews — Methodspace*, n.d.)

Our next approach was cluster sampling. Cluster sampling is a probability sampling technique that involves dividing a population into clusters or groups and then selecting a random sample of clusters to represent the population. This technique is often used when it is difficult or impractical to obtain a complete list of all the members of the population. Cluster sampling is commonly used in survey research, where a large geographic area or a population of people is involved.

The last sampling technique we used was stratified sampling. Stratified sampling is a sampling technique that involves dividing the population into subgroups or strata based on some characteristic or variable, and then selecting samples from each subgroup proportionally to their size in the population. This technique is useful when the population is heterogeneous with respect to the characteristic of interest. In our study, we used stratified sampling based on gender to ensure that our sample was representative of the population in terms of gender.

To compare the accuracy of the four sampling techniques, we calculated the absolute error between the population mean and the sample mean for each technique. Absolute error is calculated as the absolute difference between the population mean and the sample mean. The smaller the absolute error, the more accurate the sampling technique is considered to be. After calculating the absolute errors for each sampling technique, we compared the results and selected the best technique for selecting the target audience. The technique with the smallest absolute error and the narrowest confidence interval at the desired level of confidence was considered to be the most accurate.

In this way, we could ensure that the sample selected using the chosen sampling technique would accurately represent the population and provide the most reliable results for our analysis. By carefully selecting the best sampling technique, we could improve the accuracy of our study and make better decisions based on the data we collected. Overall, the process of comparing and selecting the best sampling technique is an important part of any research study. By carefully evaluating the accuracy and precision of each technique, we can ensure that the sample selected is representative of the population and provides reliable results. This can help us make better decisions and draw more accurate conclusions from our data.(*Sampling Methods | Types, Techniques & Examples*, n.d.)

# CHAPTER 04

# RESULTS & DISCUSSION

## 4.1. Results and Discussion of EDA

This Exploratory Data Analysis aims to investigate insights into the data, identify trends and patterns, and develop a deeper understanding of the relationships between variables in our dataset. In this section we would like to present the main findings related to our EDA. Since our main focus is on the number of unique customers in the dataset, we found out that there are 64428 unique customers who have made purchases. Among them 32011 are female customers and 32237 are male customers.

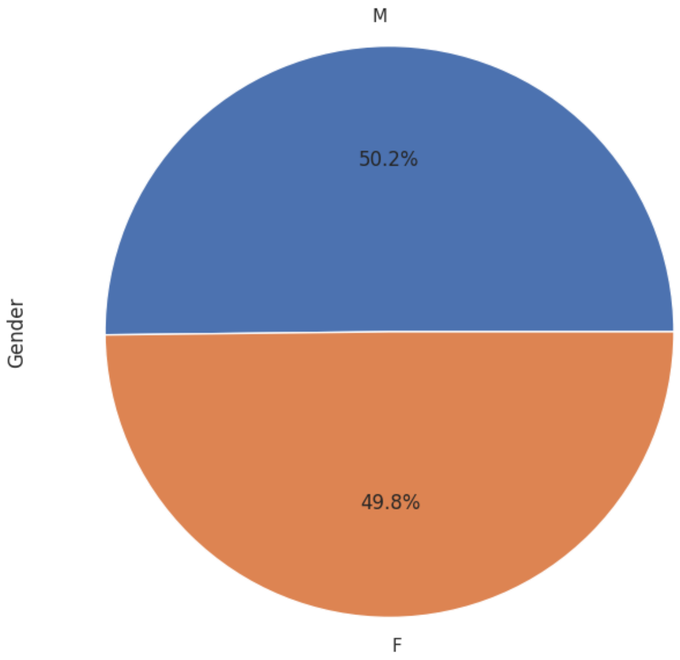

*Figure 2: Male and Female Customer Percentages*

Figure 2 shows that there is not much of a difference between the number of male and female customers.

It is also important to find the number of customers in each geographical location. When it comes to locations there are mainly four regions as shown in Table 1.

*Table 1: Number of customers in each region*

| Region | Customer Count |
|---|---|
| Midwest | 17467 |
| Northeast | 11461 |
| South | 23846 |
| West | 11474 |

Table 1 shows that the majority of customers are from the South region whereas the Northeast has the minority.

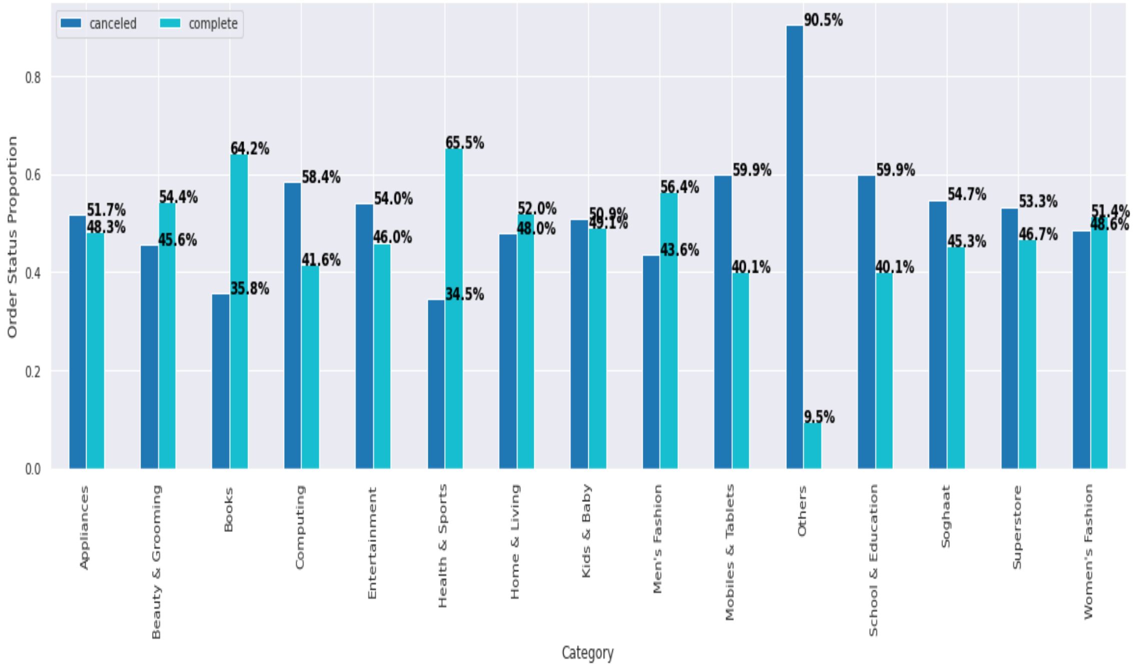

*Figure 3: Percentage of Canceled and Completed Orders in each Category*

Figure 3 shows that highest completed orders have come from the health and sports category whereas the lowest completed orders are in the other category. Furthermore, we could see that highest canceled orders have come from the other category whereas the lowest canceled orders are in the Health and Sports category. This suggests that more customers are interested in Health and Sports products compared all other products available in the market.

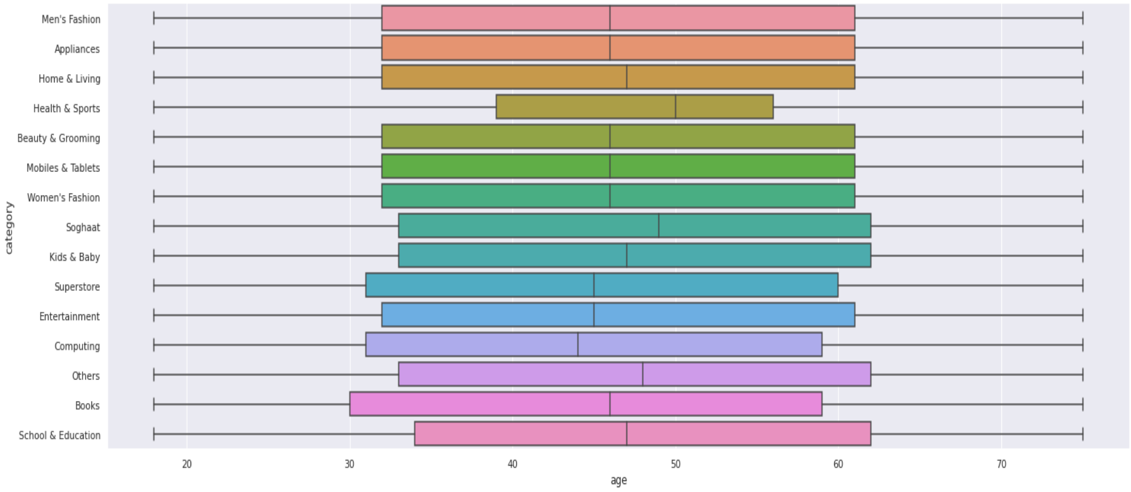

*Figure 4: Age Distribution of Completed Orders*

Figure 4 shows that the highest median age is in the category of Health and Sports whereas the lowest median age is in the computing category. Overall we could see that the age is around the minimum and maximum values of age 30 and 60 respectively.

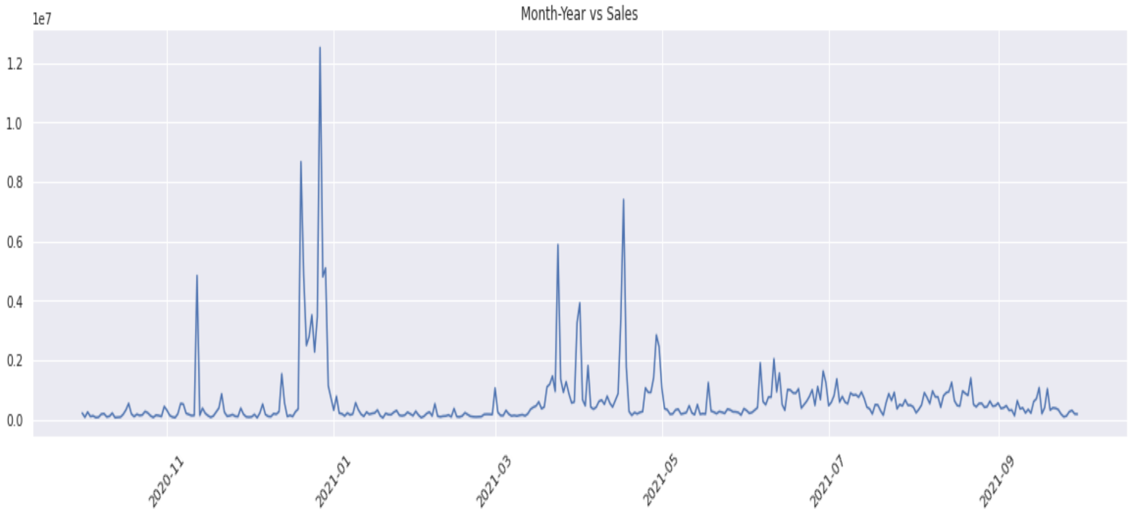

*Figure 5: Trend of Sales during the year 2020-2021*

Figure 5 shows that the sales have peaked during the period of 2020 end and start of 2023. This may have happened mainly due to the Christmas and New year season.

Apart from above results we also found out that the number of customers who have received discounts are 17567.Furthermore the number of customers who completed the order out of customers who received discounts is 10500. In other words, 60% of customers completed the orders out of the customers who received discounts.

Prior to applying sampling techniques, we have also looked at applying clustering.

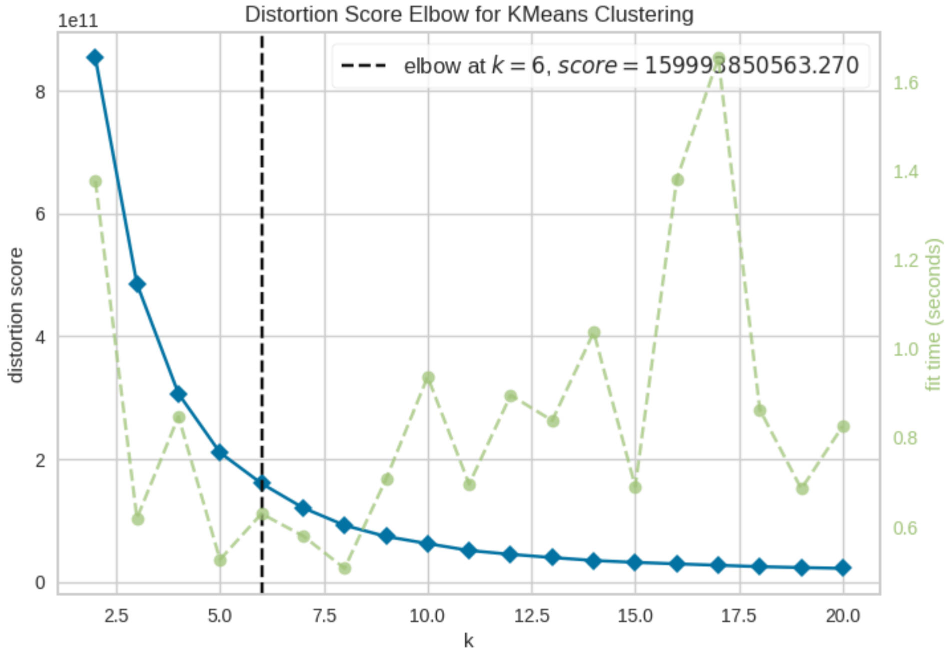

*Figure 6: Results of Elbow Method*

Once we applied k-means clustering, by the use of the Elbow method we managed to find out that the optimal number of clusters are six. This has assisted us in the next part which is the application of sampling techniques.

## 4.2. Results and Discussion of Sampling

In our project we have been able to find the absolute values of the population. Therefore, in that case the comparison of population absolute value of mean with the sample absolute value of mean is clearly possible. This would assist us in finding the best technique in an accurate manner.

This study aimed to determine the most effective sampling technique for selecting the target audience. Four sampling techniques were used, namely simple random sampling, systematic sampling, cluster sampling, and stratified sampling. The results of the study are presented below.

*Table 2: Final result of application of sampling techniques*

| Sampling Technique | Confidence Interval | Sample Size | Absolute Error | Standard Error |
|---|---|---|---|---|
| Simple Random Sampling | 90% | 264 | 61.985049 | 135.110808 |
| Simple Random Sampling | 95% | 370 | 175.533966 | 131.182983 |
| Simple Random Sampling | 99% | 621 | 33.105934 | 91.157292 |
| Systematic Sampling | 90% | 264 | 152.233022 | 112.161991 |
| Systematic Sampling | 95% | 370 | 51.971806 | 88.791998 |
| Systematic Sampling | 99% | 621 | 61.379790 | 100.187781 |
| Cluster Sampling | 90% | 264 | 801.435124 | 68.954945 |
| Cluster Sampling | 95% | 370 | 837.772860 | 64.112406 |
| Cluster Sampling | 99% | 621 | 776.133827 | 48.161081 |
| Stratified Sampling | 90% | 264 | 188.461193 | 114.689339 |
| Stratified Sampling | 95% | 370 | 102.425656 | 136.394127 |
| Stratified Sampling | 99% | 621 | 45.363194 | 89.864470 |

Table 2 shows the absolute errors, standard errors, confidence intervals, and sample sizes for each of the four sampling techniques at three different confidence levels, namely 90%, 95%, and 99%. The absolute error represents the difference between the population mean and the sample mean, while the standard error is a measure of the variability of the sample mean. The confidence interval is the range of values within which the population mean is expected to fall with a specified degree of confidence.

This result is not entirely surprising, as simple random sampling is a widely used and generally effective sampling technique. It involves selecting individuals from the population at random, ensuring that each individual has an equal chance of being selected. This makes it more likely that the sample will be representative of the population as a whole, which in turn makes it more likely that the results obtained from the sample will be accurate.

Stratified sampling also performed reasonably well, particularly for the 90% confidence interval, where it had the second-lowest absolute error. This technique involves dividing the population into subgroups, or strata, based on specific characteristics, and then selecting individuals from each subgroup to create a sample. This approach can be particularly useful when the population being studied is heterogeneous, as it ensures that the sample is representative of all subgroups.

Systematic sampling had mixed results, with relatively high absolute errors for the 90% and 99% confidence intervals, but a relatively low absolute error for the 95% confidence interval. This technique involves selecting individuals from the population at regular intervals, which can be useful in situations where the population is arranged in some predictable order. However, if the population is not ordered, this technique may not produce a representative sample.

Cluster sampling performed poorly across all three confidence intervals, producing the highest absolute errors. This technique involves dividing the population into clusters, or groups, and then selecting clusters at random to create a sample. This approach can be useful in situations where the population is geographically dispersed or where there are logistical constraints, but it can also lead to a biased sample if the clusters are not representative of the population as a whole.

## CHAPTER 05

# CONCLUSION

The selection of a sampling technique is a crucial aspect of any research study aimed at identifying a target audience. Our analysis shows that the best sampling technique varies depending on the desired confidence interval.

For a 90% confidence interval, simple random sampling produced the lowest absolute error, indicating that it is the best sampling technique for this interval. This result is consistent with previous studies that have shown the effectiveness of simple random sampling in producing reliable results in a variety of contexts.

For a 95% confidence interval, the best sampling technique was found to be systematic sampling, which produced the lowest absolute error in this case. Systematic sampling is a useful technique when the population is ordered, and the research question requires a representative sample that captures the ordered structure of the population.

For a 99% confidence interval, simple random sampling was found to produce the lowest absolute error. This result is consistent with the findings for the 90% confidence interval and further underscores the effectiveness of simple random sampling in producing accurate and representative results.

The other thing we should keep in mind is that here the selection of best sampling technique would be dataset dependent. Furthermore, as a limitation we could see that when it comes to a sampling technique like simple random sampling it may not be the best idea to use it for comparison purposes since it generates samples in a random manner each time. But in our case, we decided to use it by setting a fixed seed to generate the same random sample.

While these results provide valuable insights into the effectiveness of different sampling techniques, it is important to recognize that different techniques may be more appropriate in specific situations. Stratified sampling may be useful when studying a heterogeneous population, and cluster sampling may be necessary when the population is geographically dispersed. Overall, our study highlights the importance of selecting an appropriate sampling technique to ensure the accuracy and representativeness of the results obtained. Researchers should evaluate the strengths and limitations of different sampling techniques and choose the most appropriate method based on the characteristics of the population being studied and the research question at hand. By selecting an appropriate sampling technique, researchers can improve the validity of their results and contribute to the advancement of knowledge in their field.

# APPENDIX

1. Dashboard: https://samplingtechnique.streamlit.app/
   Pre-processed dataset:
   https://drive.google.com/file/d/1S3WYsn6wlyYAxi50Ik5D193n0e9JQten/view?usp=sharing

2. Work Folder:
   https://drive.google.com/drive/folders/1C4rh33KCfsg6yF7c7tXBy5XxWfIyOBWf?usp=sharing

# REFERENCES

1. Berndt, A. E. (2020). Sampling Methods. *Journal of Human Lactation*, *36*(2), 224–226. https://doi.org/10.1177/0890334420906850
2. *K-Means Clustering Algorithm in Python—The Ultimate Guide*. (n.d.). Retrieved 30 March 2023, from https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/
3. Leahy, J. (2013). Targeted Consumer Involvement: An Integral Part of Successful New Product Development. *Research-Technology Management*, *56*(4), 52–58. https://doi.org/10.5437/08956308X5603102
4. *Sample Selection in Systematic Literature Reviews—Methodspace*. (n.d.). Retrieved 30 March 2023, from https://www.methodspace.com/blog/sample-selection-in-systematic-literature-reviews
5. *Sampling Methods | Types, Techniques & Examples*. (n.d.). Retrieved 30 March 2023, from https://www.scribbr.com/methodology/sampling-methods/
6. Smirnova, M. (2020). *What is the target audience?* 3–5.
7. Taherdoost, H. (2016). Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3205035