T.C. BAHÇEŞEHİR UNIVERSITY

FACULTY OF ENGINEERING AND NATURAL SCIENCES

CAPSTONE FINAL REPORT
BEHAVIOR ANALYSIS OF CUSTOMER FOR A COMPANY - 1010390

Student Name, Student No:, & Departments.

Hussein Aldahesh - 1801959 - Software Engineering

Shukatullah Anwari - 1900904 - Software Engineering

Ufuk Atmaca - 1903141- Management Engineering

Nejat Kerem Eralp - 1804381 - Management Engineering

Hammam Hraesha - 1906874 - Software Engineering

Emre Kütükçü – 1902595 - Management Engineering

Advisors:

Alper Camcı - Management Engineering

Pınar Sarısaray - Software Engineering

ISTANBUL, Dec 2022

STUDENT DECLARATION

By submitting this report, as partial fulfillment of the requirements of the Capstone course, the students promise on penalty of failure of the course that

- they have given credit to and declared (by citation), any work that is not their own (e.g. parts of the report that is copied/pasted from the Internet, design or construction performed by another person, etc.);
- they have not received unpermitted aid for the project design, construction, report or presentation;
- they have not falsely assigned credit for work to another student in the group, and not takecredit for work done by another student in the group.

ABSTRACT

This capstone project aimed to analyze the impact of the COVID-19 pandemic on online shopping behavior from both a software engineering and management perspective. The project utilized Python programming language for data preprocessing, analysis, and visualization. K-means clustering, an unsupervised learning algorithm, was applied to identify patterns and clusters within the datasets. Additionally, a survey was conducted to gather responses from 199 individuals, providing insights into their online shopping habits during and after the pandemic

The findings showed that internet shopping significantly increased throughout the COVID-19 period, with consumers depending increasingly on online platforms. Trust in payment methods and customized product recommendations were found to be important determinants of online purchasing behavior. These findings have significant business ramifications, emphasizing the necessity to comprehend consumer behavior and raise satisfaction. They can improve software systems, optimize decision-making procedures, and guide marketing strategies. The project promotes the integration of data analysis, customer-centric strategies, and cutting-edge marketing techniques in the area of online retail as well as laying the foundation for future study.

By thoroughly examining the effects of the COVID-19 pandemic on online shopping behavior, this study helps to the understanding of the online commerce landscape. It highlights how crucial data-driven insights and customer-centered strategies are for navigating the changing internet market. The project's ultimate goal is to help firms adapt to and thrive in the quickly evolving e-commerce industry.

The project successfully made use of sophisticated programming tools and methodologies from a software engineering standpoint. The Jupyter Notebook IDE, the Pandas, NumPy, Plotly, and Matplotlib libraries, as well as the Python programming language, provide a powerful and adaptable environment for data manipulation, visualization, and clustering analysis. A deeper knowledge of consumer behavior was made possible by the introduction of k-means clustering, which allowed for the identification of various patterns and clusters within the datasets.

The project's management focus was on customer happiness, marketing tactics, and decision-making procedures. The poll, which had 199 respondents, provided insightful information about consumer preferences and behavior both during and after the COVID-19 period. The study of the survey responses provided insight into the variables affecting online shopping, such as the impact of personalized product recommendations and trust in payment systems.

The combined findings from the two angles showed a definite tendency toward increasing online commerce during the COVID-19 pandemic and continuing reliance on online platforms even after the epidemic. Businesses now have the chance to modify their marketing plans, upgrade their software, and improve the overall online buying experience due to the change in consumer behavior. Higher customer satisfaction and loyalty can be attained by comprehending consumer preferences, developing trust, and making tailored recommendations.

Although there were certain drawbacks, such as tiny dataset sizes and the need for more study, the project nonetheless offered helpful advice. More thorough and precise results might be obtained by increasing the dataset size and using different analysis methods like dimensionality reduction or manifold learning. The analysis would be improved, and a more complete understanding of online shopping behavior would result from streamlining the feature selection process and gathering more information.

The project's findings can help companies better understand customer behavior and make educated decisions that will improve their marketing tactics and increase customer happiness. The analysis's findings highlight the significance of fostering consumer confidence in payment options, offering individualized product recommendations, and adjusting to the changing online retail environment.

In conclusion, the COVID-19 pandemic's effects on consumers' online shopping behavior were successfully examined by this capstone project from both the software engineering and management angles. A thorough investigation of the subject was made possible by the combination of data analysis, programming tools, and customer-centric tactics. The results highlight the value of consumer satisfaction, data-driven decision making, and adaptability in the world of online shopping. Businesses can better understand their clients, optimize their marketing initiatives, and enhance their software systems to match the changing needs of online shoppers by utilizing the information gathered through this study.

Overall, this capstone project contributes to the study of online consumer behavior and provides insightful advice for companies looking to improve their online marketing strategies, especially in light of the COVID-19 epidemic and its effects on customer behavior.

Key Words: Customer, Analysis, Behavior,

TABLE OF CONTENTS

List of Tables

List of Figures

1. OVERVIEW

1.1. Identification of the need

A customer relationship is an important part of any company's existence. It is the primary way that companies generate income and gain new customers. Therefore, the way a company approaches their customer relationship is essential to their success.
When a company wants to improve the quality of its services, it should involve a behavior analyst. A behavior analyst suggests that companies focus on four areas to improve their services. Companies should address customer retention, quality of service delivery and communication with their customers. Focusing on these areas helps companies satisfy the needs of their customers. Customer service is an industry where spending money does not bring instant results. The way in which a company handles customer service affects its success. A company with poor customer service may lose clients and revenue due to poor quality of services. In contrast, a company with excellent customer service may be able to keep all of its clients. Therefore, companies should identify ways to better satisfy their customers' needs before they purchase their products or services.

1.2. Definition of the problem

The marketing concept of marketing psychology or behavioral marketing is gaining increasing importance in the business world. Marketing now includes the concept of customer analysis, where marketing personnel study customers to understand how to design products and services that match their needs. This behavior analysis approach is also known as customer understanding or human analysis. In this approach, several human factors are involved to achieve the best results. Since marketing is the process of getting the right information to the right people at the right time. In other words, marketing is creating a plan to communicate with the public. Companies use marketing to create and sell products or services that help them meet a need. Marketing can be conducted through verbal communication- for example, when a company calls its target market- or through other forms of communication, such as email or social media. Marketing can be expensive and time-consuming, but it's essential if companies want to stay competitive and create loyal customers. In recent years, with the increasing popularity of internet and buying products online, A concept called online marketing and online shopping has also entered our lives.

Some customers are using online shops but there are a huge number of people that are not using online shops, however with the recent pandemic SARS-CoV-2 also known as Covid-19, influenced online shopping experience. According to US data, %69 of people buy electronics online rather than offline (Rypáková et al., 2015). These numbers differ in every category, some of them are: %67 books, %63 clothes, %38 household items are sold online(Rypáková et al., 2015). With this, there are also problems on online shopping, related to money and transactions. The traditional understanding of barter, money and trade, which has been around for centuries, has become easier and changed in the last century in a way never seen before in human history. Of course, a change in such a deep-rooted issue has not been easy for some age groups, especially older people. Older people are a significant part of the online shopping customer base for the

reason being that older people tend to have a lower debt ratio compared to rest and more money to spend on their free time Hernández et al. (2010). When it comes to the older age group however most companies tend to neglect them and their problems. One of their problems is, payment problem. They usually do not understand the convenience and reality of digital payment. This intellectual conflict became a generational problem and widened the gap between the two generations. Individuals with different acceptance status caused a problem regarding the online-digital payment.

For this problem we will try to understand how covid-19 affected the online shopping in terms of numbers, how significant are the changes. Also, what were customers expecting from an online shop, what were the attractive things that an online shop can offer, what were the customers expecting pre-covid and what are the attractive things an online shop can offer, what are the customers expecting from online shops post-covid. In addition to this we want to look at the thoughts behind older consumers and what makes them buy the product rather than just skim over the products and leave. How they get affected by the online payment systems. To understand these problems, we will create a survey and try to understand how online platforms attracted current customers, what are the reasons those who do not use online stores are not buying from online stores and how can a company change these peoples shopping behavior.

Companies face a never-ending battle when trying to meet changing customer needs and expectations. However, managing customer relationships is much easier when you understand how to interpret customer behavior data effectively. Whether addressing current or potential customers, always remember that you're making life easier for the people who make your living off of you!

Customer behavior can be analyzed by looking at past trends to predict future behavior. For example, if customers frequently purchase pizzas online from restaurants A and B but rarely purchase them from restaurant C may consider investing in new products or marketing strategies to increase sales. According to research findings, by analyzing past trends, analysts can also predict which countries are most likely to purchase pizzas online in the near future. In addition, analyzing data from social media platforms like Facebook or Twitter allows companies to gain insights into how influencers (individuals with large followings) behave and what impact their behaviors have on consumer behavior. Analyzing this data can help companies identify celebrities who are likely to purchase certain brands of plush toys, fragrances or luxury goods and present them as brand influencers on social media platforms.

By evaluating customer behavior from a cost-benefit perspective, companies can better understand which strategies are most effective at winning new customers. For example, if a company sells pizzas online, it may want to convince its current customers that purchasing pizzas online is both healthier and more cost-effective than eating at its restaurants. To do this, it can analyze customer behavior regarding cost-benefit considerations by contacting businesses that satisfy both existing and potential customers' needs through supply chains or partnerships. Analyzing data from social media platforms allows companies to evaluate whether influencers - particularly celebrities - are satisfied with their current brands or are considering purchasing particular brands for their personal use. Using this information, companies can identify strategies that will increase sales among both current customers and influencers alike.

In order to better understand and analyze present online purchasing behavior of a customer and forecast future demands, the project will use machine learning. To do this, data analysis software will be used. Data analysis is the process of exploring, organizing and interpreting data. Data analysis has various applications- in business, science and engineering. Essentially, data analysis is the process of analyzing data. Data analysis involves a number of steps such as data collection, data processing, feature selection and machine learning algorithms. Essentially, data analysis is a way to analyze information or data.

1.2.1. Functional requirements

Software companies and product developers are constantly looking for ways to improve their product features and software applications. The primary objective behind customer analysis is to identify and fix functional issues with the software application. Through customer analysis, developers can generate ideas for enhancing their software applications and reducing development costs. The characteristics of an ideal software application include ease of use, flexibility, and the ability to perform as expected. Developers must analyze their software applications from the perspective of end users in order to determine what these requirements are. Identifying these requirements enables developers to create an application that meets the expectations of end users. Analyzing end users involves asking them questions about their job functions, businesses, personal situations, hobbies and other factors to understand their needs better. Doing so helps companies tailor their services or products to meet the needs of their target market.

Actor 1- Data Analyst

The Data Analyst shall be able to make a private account to open the database.
The database shall send an email to the Data Analyst to confirm their identities ofwhoever logins into the system for surveillance.
The Data Analyst shall be able to create/update/delete data from the database usingthe system.
The Data Analyst shall be able to check their work history; hence, the analyst will beable to compare the present data to the previous data.

Actor 2 - Data Scientist

The Data Scientist shall be able to send analysis of the data to the companymanagement.
The Machine Learning System shall confirm to the Data Scientist that their analysishas been sent.
The Data Scientist shall be able to restrict access of other users into their analysisfiles.
The Data Scientist shall be able to use live data from the database.

Actor 3 - Machine Learning Engineer (MLE)

The MLE shall be able to change and optimize models and algorithms.
The MLE shall be able to deploy AI processes into production.
The MLE shall be able to test different machine learning models and algorithms.
The MLE shall be able to ensure good data flow between the databases and backendsystems.

1.2.2. Performance requirements

2. Performance:

- Reaction Time: The framework of the management system gives affirmation in less than 3seconds once the analyst analyzes or files a complaint.
- The ML system must display pages within at most 1 second.
- The system must display statistics within at most 10 seconds.
- The ML system must send data to analysts within at most 1-2 seconds.
- The ML system must proceed the data within at most 10 seconds.
- The ML system interprets data within at most 5 seconds.

3. Maintainability:

- The ML system offers different time ranges of backup plans for the data that is being storedsuch as daily, weekly, monthly or yearly backup.
- The ML system offers bug/other update at least 1 every three to six months

4. Usability:

- Analysts should be able to understand the flow of Machine Learning (ML) systems easily.
- Analysts should be able to use the ML system without any guidelines or any help.
- Analysts should be able to use the ML system in multiple languages.

5. Security and Safety:

- The ML system should require analysts to enter their passwords every 30 days.
- analysts should create a strong password that includes numbers, capital, and small letters; otherwise, the password will not be accepted.
- All the data should be secured and be encrypted in the database.
- Analyst should link their email account to the ML system to insure their account safety
- The ML system must have separated DBMS parts.
- The DBMS must backup data within 10 seconds.

6. Availability:

- The ML system should be available on Mac and Windows/PC.

Measures of effectiveness

- If the companies were satisfied
- If the Stakeholders were satisfied
- If the ML system launched within a fix budget and assigned period
- If the Retention Rate would be between 40%-70%
- If the team members were up to task and performances

1.2.2. Constraints

1. Economic:

- The customer behavior analysis must provide comprehensible data that are valid and verified in order to build business ideals and increase profitability and competitivenessin the market.

2. Environmental:

- The ML system must be environmentally friendly by efficiently collecting and storingdata in the database.
- The ML system must optimize ML techniques to conduct accurate customer behavior

3. Social impact

- The ML system must identify the customers' needs in order to improve the services of the company.
- The ML system must be highly secured to prevent data theft and data leakage.

1.3. Conceptual solutions

Marketing can cause changing customers' economic activities by analyzing data from social media networks. Businesses can determine whether influencers are happy with their existing brands or are thinking about acquiring particular brands for their own usage. Companies can use this data to determine marketing tactics that will boost sales among current clients and influential people alike.

Customers' behavior has significantly shifted as a result of Covid 19, moving them from traditional retail to internet shopping. The project will make use of machine learning to more thoroughly comprehend and examine a customer's current online shopping behavior as well as predict future requirements. Software for data analysis will be employed in this. The process of investigating, arranging, and interpreting data is known as data analysis. There are many uses for data analysis in business, science, and engineering. Data analysis is essentially the process of analyzing data. Data gathering, data processing, feature selection, and machine learning algorithms are a few of the procedures that go into data analysis.

1.3.1. Literature Review

What Is Machine Learning:

Numerous technical tools and applications have been developed to address and analyze customer behavior. In order to give their customers, the best services possible, businesses needed to understand their customers and their needs better, which led to the design, development, and use of ML technique tools and methods. A deeper comprehension of the crucial functions that ML systems and technology analysts may play would be advantageous to companies. The issues associated with ML technological tools, technology design, development, and implementation are discussed in this article along with how developing ML technology tools are being used by companies to better analyze their customer behavior. Additionally, it provides ideas and insights on how technology analysts and ML systems could aid businesses in analyzing client behavior and expectations.
The machine learning method is used in this article to analyze the technical solutions currently available for analyzing consumer behavior, as well as the challenges and opportunities for ML technique tools and technology analyst.[1] Unsupervised learning, which identifies hidden patterns or intrinsic structures in input data, and supervised learning, which builds a model using known input and output data to predict future outputs, are the two types of approaches used in machine learning [1].

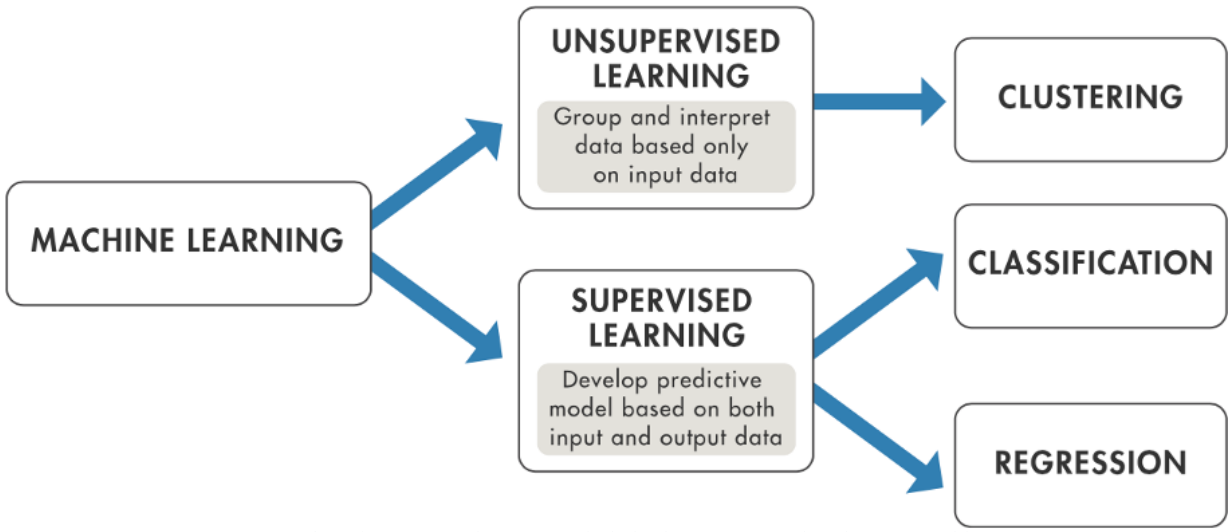


Figure 1 Machine learning techniques include both unsupervised and supervised learning.

The above figure illustrates how machine learning works and how it is divided into two types. Each of these types has different functions and features, such as clustering for unsupervised learning and classification and regression for supervised learning, respectively.

Online shopper behavior: Influences of online shopping decision

The Internet has changed our daily life significantly, it enables people to communicate, play games, send mail, buy products, and search for things. In recent years people accepted online shopping more than ever and with this change, online shopping companies managed to collect more information about people's behavior through their shopping. Whether the consumer is online or offline, the decision-making process is very similar [2]. The consumer decision model shows that decision to purchase a product begins with: Recognizing needs then searching for information, evaluating alternatives, making a purchasing decision and finally, there is post-purchase behavior [2].

In these processes online consumers are able to search for more things with the help of search engines and with this they are able to compare many products in little time. It doesn't end with comparing products, after comparing the product, online offers so many different prices for the same product. Other than this some websites offer customer auctions, with this online shopping almost becoming a game for customers [2].

Online shopping provides many things but because consumers can't touch and look closely at the product, the consumer takes a risk by online shopping [2]. Online payment is also risky for consumers because they need to give personal information to the company [2]. Online consumers tend to buy the product from the websites they trust. In terms of online shopping companies' success, trust is one of the most critical issues [2].

Even after all these risks online shopping is still very popular and some of the reasons that online shopping is very popular are:

The consumer can buy products online without going anywhere and they can do it anytime. Consumers also can avoid face-to-face interactions with cashiers. Empirical research gives us the information that consumers buy online because of the convenience of the internet and online shopping [2]. Online stores are open 24 hours, 7 days. Research shows us %58 percent of consumers choose online over offline because they can buy products after the regular store closes [2]. Also %61 percent of consumers selected online shopping over offline to avoid crowded stores and long waiting lines [2]. Consumers also look for services and some companies provide 24-hour operating online customer services [2].

Online Shopping Customer Behaviour Analysis using centrality measures:

In the current state of the world, every activity generates an enormous amount of data, and it is called Big Data [3]. As we mentioned before, online shopping is growing significantly and companies have become more competitive than ever before. There is a big competition between companies to attract more consumers to buy the product, and to do so these companies are tryingto offer more to consumers [3]. To achieve this target, shopping datasets are the key element.
Datasets can be obtained from various different data, but first, we need to collect data from customer's online movies and timings, with the help of collected data about the customer, we canpredict customer shopping behavior [3]. To predict customer behavior, we can use centrality measures [3]. In the process we intend to find important customers, firstly we use different attributes and with them, we create a dataset. We select the attributes of important customers with using different methods, and all these methods include mathematical results [3]. In all methods of centrality measures, the Attribute list method has been the most accurate method, while having many attributes [3]. After all, centrality measure methods may help companies to understand their customer behavior much more than before. With this method, they can find the important methods without much effort [3].

Analysis and modeling of changes in online shopping behavior due to Covid-19 pandemic:

Online shopping has been growing since the early 2000s, as we mentioned before online shopping offered customers many things that offline shopping cannot offer. With the pandemic, these offers became much more important than before. These days we consider if the pandemic completely changed the online and offline customer behavior [4]. Research using the NHTS 2017 and 216 different data with the help of descriptive analysis and choice modeling shows that: For grocery stores, the pandemic didn't change shopping behavior significantly, most of thecustomers continued to buy from offline stores [4]. For household needs, the situation is not likethe grocery store, the number of consumers using online grocery stores has increased with the pandemic and after the pandemic, it still stands at close levels [4]. From here we can say that after the pandemic there will be some increase in the number of online and household goods shopping [4].
After grocery and household numbers, when we look at the technological products, we can see that while the pandemic happens more people tried online stores [4]. After the pandemic people tend to use online stores more than before the pandemic happened, only 20% of consumers say that they will buy from offline stores when the pandemic is over [4]. This %20 people want to goback to offline shopping because they want to touch and see the product visually [4]. NHTS datashows us, with pandemic online shopping trends grew from the pre-pandemic area, but it did notgrow significantly [4]. In a survey to understand consumer behavior, %59 consumers said they will increase the online shopping after the pandemic, %38 percent said they have no intention to change their shopping after the pandemic, while %3 percent said they intend to use less online shopping stores [4]. All these people have their own reasons to answer questions in this way.
These reasons have a huge scale from crowd avoidance to spending less time while shopping [4].From these findings, we can say that wherever we look at the situation, and which method we use does not matter. The pandemic changed online shopping in an effective way. The reason online shopping numbers have changed is because customer behavior has changed and if we canunderstand the after-pandemic consumer behavior, we can use it in many fields.

Marketing Identity:

Customer behavior is how the consumer reacts and thinks while trying to buy a new product or using, rating the product [5]. The center of behavioral choices is the decision-making process ofthe consumer [5]. There are 3 different types of consumer behavior: Rational, Stereotypical, andImpulsive [5]. With these behaviors, consumers decide if they want to buy the product or not.
This buying process begins with recognition of the need and ends with the decision to buy [5]. These behavior types are used in offline shopping, but with the internet so many things have changed. The relationship between customer to customer and customer to company has changedsignificantly [5]. Online shopping has been growing in the number of people using it, %96 of Slovak people has used online shopping at least once in their lifetime [5]. Online shopping happens in seconds and it needs to be a smooth experience [5]. If it is not smooth, then other customers may send bad reviews, and affect possible customers [5]. Online Customers are not passive buyers, they use the communication and searching methods [5].
According to US data, %69 of people buys electronics online rather than offline [5]. these numbers differ in every category, some of them are: %67 books, %63 clothes, %38 household items are sold online rather than offline [5]. These data may differ from country to country. Fromthis perspective, if we look at the Slovak data. from a total %100 for each item, %20 clothing,
%19 service, %18 electronic items are sold online [5]. These numbers mean something but if we understand why people buy from the online store, and why they behave this way, it would explain more. When we look at the studies, they show us why, %80 of consumers chooses onlinebecause of free shipping, other numbers of why they choose are: %66 1-day shipping, %64 free return, %41 same day shipping, but this changes from country to country as well. For Slovakia
%31 of consumers chooses online because of price, %18 chooses because of shipping conditions(4). Except from these data`s, if they need to open an account before buying an item what happens [5]. In the US %54 percent of consumers directly leave without buying [5]. Only %25 percent open an account and continue the shopping [5]. In Slovakia %62 of consumers directly leaves the process, only %34 of consumers opens an account [5]. This data shows us behavior changes from country to country and if we need to understand the consumer behavior, then we need to look at the consumer behavior from country-to-country consumer behavior rather than the world.

Customer Behavior Analysis Towards Online Shopping using Data Mining:

Online retailers are currently one of the most used methods, and their popularity is constantly growing [6]. The public does not need to lose time and spend additional funds on transportation. Confirming the clients' requirements and working to meet them. In addition, the customer's financial situation will be considered, along with his decision over whether to purchase or not. Nowadays, there is fierce competition on the market, every business invests more money in websites that market their products effectively [6]. For this study, they worked at websites that sell hundreds of home decoration goods, including covers, bedspreads, and kitchen necessities, The selection of this database was motivated by the fact that it is a functioning website with a significant daily customer traffic volume, offering a large dataset for analyzing consumer behavior as well as [6]. The number of customers and overall sales are an accurate representationof the site's user base and financial resources. Additionally, it may choose which things on the site are sold the most and the least, also which cities provide the most sales. Through query

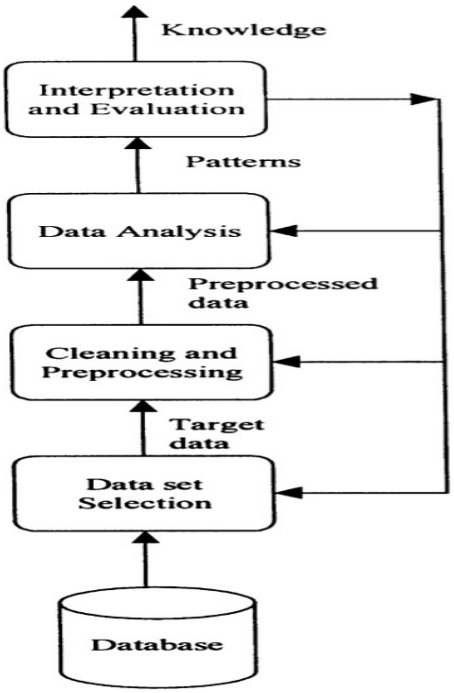
processing on the database of the relevant site the names of the cities are retrieved. Market basket analysis and the most popular internet pages are two examples of association mining [6].

Business data mining — a machine learning perspective:

Data mining is the practice of identifying intriguing patterns in datasets that may be used to inform decisions. By taking use of the possibilities of huge data warehouses, data mining may provide a company a considerable competitive edge. Due to the extensive usage of networking and computer technology, this activity's scope has evolved [7]. The study of computer techniques for automating knowledge acquisition from examples is known as machine learning. Finding a pattern in a training data set is a typical tactic. The behavior of additional cases is then classified or predicted using this pattern. A test data set is then applied to the resultant decision tree for assessment [7].

Input nodes are connected to output nodes through hidden nodes in a conventional neural network (NN), which comprises these nodes. Utilizing examples from previously classified data, a NN is trained. In order to employ NN as symbolic rules in machine learning, researchers are creating extraction methods [7]. Data mining applications in business have certain characteristics that affect the performance of machine learning techniques. Understanding these characteristics and their impacts on machine learning is useful in selecting an appropriate technique for an application. The predictive accuracy of a data mining technique strongly nuances its effectiveness [7]. Web data mining has several applications, but one of the most crucial ones is identifying web usage trends. For the purpose of understanding web user access patterns, log file data has been analyzed using data mining. Finding new users and enhancing website accessibility are also possible uses for this data [7]. Fig 1. offers a summary of this procedure. When data is retrieved from several sources, a severe issue is caused by databases' lack of consistency. The data that has been retrieved from various sources is cleaned to get rid of conflicts and inconsistencies. Data can be obtained from several active databases or from a single source like a data warehouse.

Figure 2 Database Process



Prediction of Customer Behavior using Machine Learning: A Case Study:

All business operations are impacted by the digital revolution. Information systems can be used to gather consumer and product data. By utilizing technologies from data science, these data may be employed to comprehend the customer's intention more thoroughly. In machine learning, the issues might be viewed as supervised or unsupervised models. A Bayesian approach for learning rule sets to resolve a classification problem was put forth [8].
In this research, we take into account the issue of predicting client intentions using the information [8]. Instead of breaking the dataset up into five folders, our strategy is to deal with it directly [8]. With the aid of cutting-edge technology, the environment for improving customer behavior prediction has expanded in numerous ways. This research represents a novel strategy that enhances the accuracy of numerous prediction models. To enhance performance and predict with calculated accuracy, the data is explored using a variety of preprocessing models [8]. With an acceptance rate of almost 50%, the data is relatively balanced. The type of data must be numerical in order to test the models. The technique we employ is called min max scaling, and it entails rescaling the range of features to scale the range in [0,1] [8]. Predictive modeling is the process of approximating the mapping function from the input variables to the discrete output variables. We examined the effectiveness of many classification models, including logistic regression, feedforward neural networks (MLP), support vector machines (SVM), decision trees, and random forests [8].

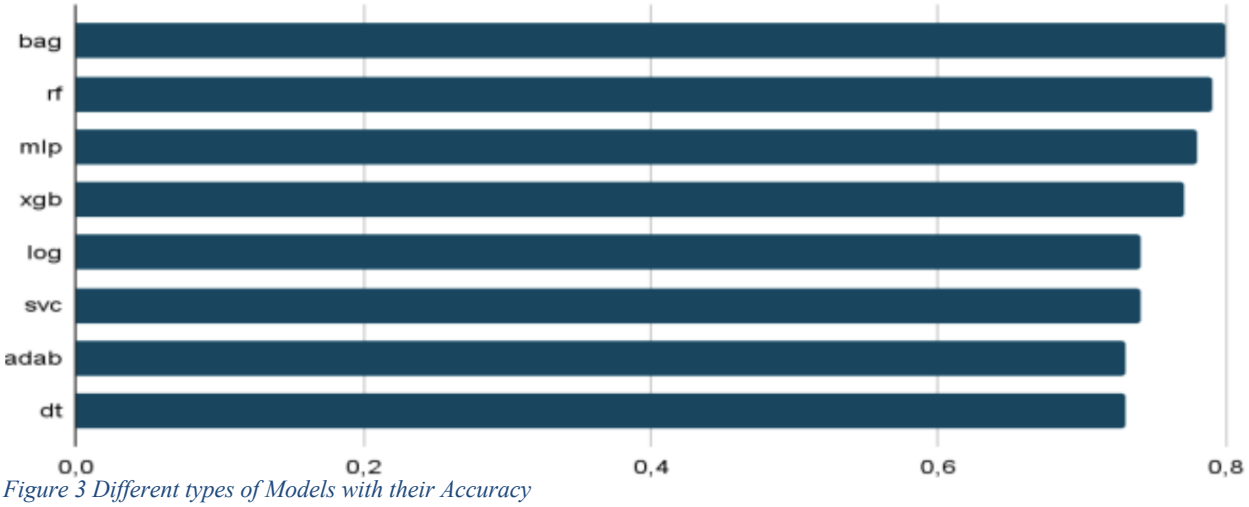


Figure 3 Different types of Models with their Accuracy

FIG1: With 80% accuracy, the bagging model is the most accurate. Random forest is the model with the second-highest rating. Therefore, this study may be viewed as improving the predictability of marketers. For the dataset and the general challenge of predicting consumers' future purchase intentions, we advocate the use of ensemble learning techniques [8].

Consumer behavior analysis for business development. Aggression and Violent Behavior

Three unique roles are played by the customer's behavior: buyer, payer, and consumer. Knowing how consumers choose a product helps close a market gap and aid advertising in understanding the preferences of their target market. The more individuals respond to your marketing tactics, the more probable it is that they will purchase your product. More marketers from other disciplines transition to the profession laterally as marketing continues to diversify. For marketers with advanced technology and digital skills, addressing the absence of fundamental marketing competence is a challenge. Social networking is a helpful tool for the business to draw in new clients and get insightful data. In order to study customer behavior for company growth, the Adaptive Hybridized Intelligent Computational Model (AHICM) has been created. In an economy that prioritizes marketing, the analysis of consumer behavior describes the behavioral economy with the intricacies of real-world customer choice [9]. In order to provide long-term value for a customer, market, or partnership organization, a business must be developed. To grow the company beyond its current condition, it is crucial to focus on one or more of these areas. Programs for business growth can be given the planned marketing budget. Companies look into consumer psychology and services following customer purchasing behavior to generate marketing campaigns, new goods, and increasing profitability. Using this knowledge, businesses may assess their present course to see whether it best appeals to their target audience. They benefit from a growth in sales, earnings, and profitability. The primary goal of marketing a product is to satisfy the wants and desires of the consumer [9]. Analyzing good purchasing habits aids business owners practically in marketing and production plan adaptation. The performance and substance of several strategies associated with high-tech enterprises pursuing growth are identified using cluster analysis. The procedure draws attention to the connections between organizational objectives, intended outcomes, and outputs. Marketing analytics have been applied in a variety of methods to understand profitability. Based on past customer behavior, predictive analysis aids in identifying those market sectors that are the most lucrative. According to PwC, the AHICM approach improves the efficiency ratio when compared to other current methods [9]

Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing:

Over a century has passed since the beginning of marketing research. The development of digital data opens us a world of possibilities for forecasting and understanding consumer behavior [10]. For more than 60 years, marketing experts' research has been nearly solely based on survey data. Digital data is being used more often by marketing academics to assess the long-term effects of marketing initiatives. Marketers may learn from digital data how specific customers react to different product and service offers on the market. Additionally, it may be used to create standards for gauging how different online pricing and advertising messages are received [10]. Research methodologies from the fields of computer and data science are now attracting the interest of marketing researchers. The study and use of machine learning algorithms is expanding. Since they are only as good as the data they are based on, data quality is a key issuein their use. As more divergent digital data sources are combined, it becomes more difficult to reliably verify data quality. Marketing researchers that use machine learning appear to dispute data quality issues considerably less than those in computer science and information systems [10]. A data type is considered sparse if just a tiny portion of it includes pertinent information. Missing values can cause data sparsity, however this problem usually arises during the data generation process. It makes it more difficult to recognize the distinctive characteristics of the material and its semantic patterns [10]. Text, pictures, and audio comprise between 80 and 90 percent of all unstructured data worldwide. The analysis of the sentiment and meaning of social media postings uses information extraction from text. Language-related aspects such as semantic, visual, and grammatical descriptions can be extracted from images. For the marketing industry, unstructured data have enormous promise. Optimization strategies are used in computation time reduction efforts that are commonly seen in high performance computing. However, the prioritizing and curation of such material are made more difficult by the data structure's [10]. Testing well-established ideas of consumer and business-to-business behavior has been the motto of marketing scholars. The development of big data and machine learning techniques has raised questions about this approach. Achieving a degree of knowledge that can be utilized to make accurate forecasts requires understanding the why [10]. Machine learning models must, according to computer experts, offer a human-understandable explanation of how they work. An artificial intelligence of this kind might facilitate decision-makers' acceptance of the results and provide information about probable causative pathways [10]. Predictive modeling, which may offer helpful guidance for theory creation, may help prediction-driven explanation, which has been shown to be uncontroversial in the physical sciences. The inherent limitations of induction and the efficacy of machine learning approaches for hypothesis testing should be understood by marketing researchers. Causal diagrams have an explanatory component since they display the operationalization of conceptual factors and the connections between the variables that researchers are attempting to explain. The extraordinary internal validity of lab experiments makes it possible to deduce causal relationships from the study's results. It emphasizes prediction heavily when estimating a model as a causal-predictive technique of structural equation modeling (PLS-SEM). The ability of marketing researchers to perceive marketing issues and challenges should be improved.by utilizing the growing amount of digital data that is available [10].

Using big data to improve customer experience and business performance:

Offering a superior client experience is essential to competing in a market where communications service providers (CSPs) have more and more of the same service offerings and devices. Key differentiators for CSPs include solutions with the capacity to emphasize what truly influences customer happiness and give actionable insights from their extensive customer, network, and service data [11]. Communications service providers (CSPs) hold vast amounts of data about their customers, but are not effectively managing it. Each organization has a partial view of its own customer touch points, but lacks a comprehensive end-to-end view across all of the customers' touch points with the CSP [11]. Measuring CSP customer experience holistically is a "big data" problem, and to be effective, analytics tools must incorporate big data. Gartner defines big data as having three attributes: high volume, high velocity, and high variety. A big data solution for CEM needs to incorporate these technologies. As customer-centric data is generated, it needs to be rapidly stored and processed. New data-management technologies such as Apache Hadoop and NoSQL databases have been introduced. Streaming analytics technology can be used to detect preset conditions and trigger actions in real time [11]. Another crucial step in turning data into insights, and insights into actions, is choosing the customer experience use cases that will give the service provider with the best financial return—those that are immediately commercially viable or are regarded as mission-critical—and starting with the most straightforward with the most convincing evidence [11].
When employing analytics to transform customer data into insights and subsequently into actions, an average CSP may possibly create $10 per subscriber per year in additional profit, according to a Gartner study [11]. 55% of the total revenue and cost savings come from six use cases:

- Churn propensity scoring.
- Improved self-service.
- Customer segmentation.
- Best next offer.
- Prepaid recharge.
- Tracking dissatisfied customers.

Customer experience management (CEM) strategies must be driven by a top-down executive who is charged with creating an enterprise-wide CEM strategy. In some countries, market regulations create silos by separating service provider operations into separate businesses such as wholesale, retail, or infrastructure. Service providers have begun to make this change and are appointing experienced officers and others in similar roles to tackle the problem [11].
The paper uses a unique methodology to select the right key quality indicators, build accurate key business objective "formulas," and predict customer behavior. The Customer Experience Manager (CEM) is the person who provides the customer with actionable insights and a 360-degree view of each customer's behavior, issues, and desires. CSPs must enable a positive perception of their company coupled with quality, customer care, and ease of doing business that satisfies or delights them [11].

Customer behavior analysis using real-time data processing: A case study of digital signage-based online stores:

Understanding how consumers react to innovative goods or services is made easier by studying consumer behavior. Customer behavior analysis has grown in importance across many businesses due to the favorable effects it has on profitability and sales. Understanding consumerbehavior enables marketers to identify the where, when, how, why, and elements that influence consumer behavior [12].

This paper's goal is to suggest consumer behavior analysis using real-time data processing and association rules for an online business that uses digital signage. To manage the enormous volume of customer behavior data, real-time data processing based on big data technology (suchas NoSQL MongoDB and Apache Kafka) is used [12].
Customers' browsing and transactional data from digital signage (DS) could be used as the inputfor making decisions in order to extract customer behavior patterns. For clients to experience product browsing and purchase, the authors created a DSOS and placed it in various locations. The real-time data processing system captured browsing and transaction data from clients as it occurred [12].

Digital signage-based online store (DSOS) should be used to improve the shopping experiencefor customers because retailers seek to increase customer satisfaction. In order to increase customer satisfaction, management must consider customer behavior carefully. Since customerbehavior plays a significant role in management decision making to enhance customer satisfaction. A study has proposed customer behavior analysis based on real-time data processing. For real-time data processing, large data technologies like NoSQL MongoDB and Apache Kafka were used. The suggested system tracked consumer browsing and purchase information from DSOS and used it as a decision-making input. Additionally, from consumer behavior and purchase data, the association rule was employed to obtain insightful information[12].

The proposed model's scalability was demonstrated, and it was determined that when the DS device's customer base grew, the proposed system could efficiently process enormous amounts of input data. Our findings showed a connection between a high volume of visitors to a certain product and its sales. The outcome showed that as a customer's number of product browses increased, so did the likelihood that a consumer would buy a certain product. Additionally, whenindividuals shopped for longer periods of time, there were more items bought concurrently overall. Retail managers might highlight the product categories with the highest sales growth onthe front page and develop a loop of content that features top-selling items and lasts a few minutes, which would increase future sales revenue. Additionally, this study demonstrated that the association rules from the frequent transaction pattern were attained by merging client purchase and browsing data. The suggested approach is expected to aid management in making decisions on enhancing the design of DS and providing customers with better product suggestions and recommendations, particularly for green products [12].

Customers' relationships with the retailer grew stronger as interaction and the caliber of shoppingthrough DSOS increased, which aided in boosting the company's income. In this manner, the

primary goal of customer relationship management was achieved, strengthening the relationship between the client and management through the usage of DSOS. The outcome showed that the suggested system could readily handle an increase in the DSOS's customer base while processinga large amount of incoming data [12]. Additionally, the suggested association rule can be used torecommend products in the DSOS, assisting customers in getting the goods they want while alsoincreasing future sales revenue for the business. A better and more convincing virtual shopping environment was also made possible by good collaboration amongst various fields, including human computer interface, marketing, and commerce [12]. In subsequent work, we will combinethe big data platform with eye tracking, facial recognition of the user, and browser history to predict client behavior more effectively. This will enable us to comprehend client feedback on the store better, and store managers will be better able to display suggested products in-store.
Additionally, comparing various data mining techniques is essential for examining consumerbehavior [12].

Improvement:

- Eye tracking,
- Customer facial recognition
- Browsing history

Customer Behavior Analysis Using Rough Set Approach:

By examining client demands and behaviors, the customer relationship management (CRM) business model helps businesses develop long-term lucrative consumers. By selecting significant attributes from the client database, the behavior of the customers is examined. The customers are then divided into groups based on the values of their attribute. To describe the customers in each category, rule induction techniques are used to construct the rules. These guidelines can be used by business owners to anticipate the behavior of potential clients and alter the way they attract current clients [13].

The ideas of rough set theory have been used to develop a novel rule algorithm that has been proposed in this research [13]. The LEM2 method, an existing rough set-based rule induction technique, has been used to compare its performance. Analysis is done using a real data set of client transactions. For the purpose of evaluating client data, the qualities chosen are Recency (R), Frequency (F), Monetary (M), and Payment (P) [13].

Customer management activities and corporate performance are mediated by customer relationship management (CRM) technology. In order for a company to maximize its profit by acting in accordance with the customer's characteristics, it helps industries obtain insight into the behavior of customers and their worth. Customer identification, customer attraction, customer retention, and customer development are the four philosophies that make up CRM.
Data mining is a group of methods for quickly and automatically finding new, accurate, practical, and intelligible patterns in huge databases. These patterns are applied to decision-making inside an organization [13]. Clustering, association rule mining, rule induction, and classification are among the tasks that can be carried out in data mining [13].

- Data with similar properties can be grouped using the unsupervised classification method known as clustering. For the supplied input data, it creates clusters where datain one cluster is more comparable to data in other clusters [13].
- Based on the occurrence of other attributes in the database, dependency rules createdby association rule mining will anticipate the existence of an attribute [13].
- Rule induction is a type of supervised learning that produces rules by identifyingpatterns in data that have already been sorted into groups [13].
- On the other hand, classification also produces rules for describing the data ineach cluster [13].

Customer Relationship Management is a collection of operational procedures and enablers that support a company's long-term, lucrative customer relationship strategy [13]. In order to increase customer value, loyalty, and retention and to adopt customer-centric tactics, this company operating philosophy is used [13]. Because every organization is focused on the client, this technology is crucial. A thorough grasp of enterprise clients is necessary for customer identification [13].

By enhancing client relationships, customer relationship management is a technology that aids entrepreneurs in growing their firm. In this study, client segments were created using the clustering technique for data mining, and each segment's consumer behavior was described using rule induction. Therefore, assigning a customer to a cluster is crucial in CRM. The customer behavior in each cluster must be well defined for an effective rule induction method to predict the new customers to the correct cluster. The rule induction algorithm's prediction accuracy is used to determine the performance evaluation criteria [13].

Keywords:
- LEM2 (Learning from Examples Module, version 2)
- Customer relationship management (CRM)
- Clustering
- Rule mining
- Rule induction
- Classification

Tools for Data Analysis used in Data Science, ML and Big Data:

Innovative technological tools have been developed to analyze customer behavior, including Microsoft Excel, KNIME, Splink, and many others. These tools can help companies understand their customers better through analytics, data science, machine learning, and big data [14]. ML method tools can help discover a company's ideal customer, track customer preferences, boost revenue, and aid in the development of new goods and services.

Furthermore, Big Data Analytics will be utilized to ensure that companies maintain track of their customers' needs and wishes, ensuring that each client will be supplied the proper product for their home, office, travels, family, or other everyday needs [14]. Additionally, it will motivate businesses to monitor the expansion of consumer needs and demand for services, deliver products on time, and gradually enhance their performance and understanding of customer behavior. Big data analytics has been used successfully in many large and small businesses to assist clients and offer prompt service by analyzing store visits, shopping purchases, and discount cards.

Customer Segmentation and Strategy Development Based on Customer Lifetime Value: A Case Study:

The proliferation and usage of the internet and mobile phones on a worldwide scale have led to the growth of digital payments [15]. Despite the enormous potential of digital payment systems, there is a paucity of research that provides a complete synthesis and analysis of the elements that influence their usage, adoption, and acceptability. This research intends to solve this deficiency by conducting a thorough analysis of the relevant literature gathered from the Scopus and Web of Science databases [15]. A final sample of 193 research publications was discovered and analyzed using a systematic procedure [15]. The findings demonstrate that no one explanation adequately explains the complexity of electronic payment uptake [15]. Existing theories are severely limited by their failure to account for the impact of social and cultural factors in the acceptance of new technologies [15]. In business studies, literature reviews are a common activity, but there are few reviews that use the systematic review approach, which accumulates information using well defined methods and criteria [15]. This is the first comprehensive review on the adoption of electronic payment methods, and it organizes the current information and suggests areas for future study [15].

Predicting Customer Behavior by Analyzing Clickstream Data:

The objective of the research is to identify the elements that impact the behavior of consumers toward shopping malls so that appropriate action can be taken[16]. In conjunction with this research, efforts have been undertaken to analyze the many factors that impact the behavior of customers toward shopping malls[16]. This assessment of the relevant literature revealed that consumers had a preference towards shopping malls. Customers do not just travel to the shopping mall for the purpose of shopping; rather, they also go there for amusement[16]. According to the findings of a variety of research, there are a great many aspects that contributeto the attraction of shopping malls to customers[16]. This work will undoubtedly enable shopkeepers or merchants to make modifications (if any) in the mall in order to attract clients and meet their requirements and desires, which is crucial for the growth of a mall[16]. Additionally, this work is important for the development of a mall[16]. After reviewing one hundred papers on shopper attitudes toward shopping malls, the author came to the conclusion that the shopping environment, the ease of shopping, the availability of various products, the showbiz that is offered at malls, the parking facility, the good product quality, discount and sales promotion are the factors that convince shoppers to visit shopping malls with entertainment[16].

Age, gender and income: do they really moderate online shopping behaviour:

From the start of trading, throughout history and currently, attracting customers has always been the number one priority for most companies. Most companies look at supplier-focused problems and for this moment, our focus will be on customers. In this paper Kim et al. (2005) talks about the LTV model (Life time Value) and CRM (Customer relationship Management) and how it became the prime scheme of work in today's aggressive marketplace. The author talks about CRM as a method, a way of understanding and administering the customer relationship with the added perspective of constantly developing technologies and operating procedures of a company. CRM is a great way of tracking and like we mentioned, administering customer information we gather from various marketing strategies. This information can help segment our buyers and customers by their demographic and psychographic. In addition to that, CRM helps us with targeting. The author continues by talking about the four advantages of CRM: [1] Keeping the customers from detaching as well as increasing faithfulness, [2] increased cost-effectiveness, [3] Formation of value in favor of consumers and [4] Adjustments related to the companies' output and products. Kim et al. (2005) also mentions how in the context of today's market, with the rising motion of focusing on pleasing the consumers, most companies try to use this to their advantage and increase the loyalty aspect. The author talks about how with a strong CRM, determining the value of a consumer is possible and how it is necessary to know the values so the company can issue an overall better marketing plan. People have values and learning these values makes it easier to target and segment consumers more efficiently. The author also talks about how the values of consumers, like the potential value, support the segmentation process by helping the marketing department, while also giving the company the opportunity to cross-sell products. The author touches on the importance of the brand image, especially with young adults, and how it can attract more consumers. Kim et al. (2005) continue by saying rather than changing the brand image completely for the reason being that the image and the product should go parallel, the company may be able to implement new marketing ways with excluding the brand image.

The impact of the customer satisfaction, switching costs and trust on customer relationship commitment:

With the increase in customer size when it comes to online shopping, all companies want to start analyzing this data for their marketing framework, which the information we are talking about is close to 2 million and 5 hundred Terabytes worth of data fabricated every single day and this data is called Big Data. In this paper, Gumber et al. (2021) talks about clickstream data and how it is used for predicting the behavior of an online customer on a company website. The author continues by stating that, for this data they use a technique named Extreme Gradient Boosting, for the reason being that this technique is one of the most accurate ones for predicting customer behavior. Within this research it gives an exact %85.9 accuracy. Like we mentioned, the way we gather this data is clickstream data, which is basically the data that is put into the data log when a consumer clicks something on a website. This data can be just a customer skimming through the products of the company, or it can be adding new products to the basket but not buying, etc. Gumber et al. (2021) talks about their findings from a popular website with various categories of products and the data of the 3.5 million distinctive customers with close to 70 million recordings in the time span of 30 days of November 2019. The author states their findings with: %4.49 users put products into their basket, but they do not buy the said product, %94.15 of the customers just skim and view the website without buying anything or putting anything in the basket and the remaining %1.36 buy the products they add to the basket. The author also adds that on Saturday and Sunday there is a rise in the customer numbers that visit the company's website. The author continues by mentioning the technique used in this research also contains the fusion of multiple decision trees, and click streaming is the best way for the companies to cooperate with the customers and get into their heads for predicting what the customers do. According to Gumber et al. (2021), Extreme Gradient boosting is the best way to predict consumer behavior. If we are to make a comment on this, besides the methods used in this research we can say that most online customers on a website do not purchase the product even if they put the product into their basket. With this knowledge about the website, we can optimize the clicking order of the customers and suggest better options regardingtheir choices on the website.

CONSUMER BEHAVIOR TOWARDS SHOPPING MALLS: A SYSTEMATIC NARRATIVE REVIEW:

Since the 90s marketers believed that the socioeconomic variables, which are for example income they get yearly, the education they have or their gender, have affected the buying phase and the customer behaviors immensely. While some of it might be true, it might not be as important as they make it seem. In this study of Hernández et al. (2010) consumers who interact with online marketing and do any form of online shopping are being examined and regarding our previous knowledge, in customer segmentation, we know factors like demographic and psychographic give directions to our marketing focus. However, distinctly Hernández et al. (2010) states that when it comes to targeting segmented consumers, the company should place their focus elsewhere other than socioeconomic variables, for the reason being that they lose its credibility under customer behavior at certain conditions. The author explains, in the beginning, the "developing" phase of the information technology, it was likely that outnumbered consumers in the public like low-income groups, would affect the consumer behavior. The author continues by stating that however when it comes to information technology, as the technology improves and after a certain knowledge and experience is gathered about it, the "I do not know how to do this" thinking disappears and once that happens, the behaviors of the customers become the same. Hernández et al. (2010) also talks about the importance of the older users in online shopping and how they are a valuable and highly profitable part of the customer base for the reason being that in most cases older people tend to have a lower debt ratio and more money to spend on their free time. The author continues by stating that while at first the older generation tend to have a hard time purchasing, as they get used to it when segmenting the customer based on their age, they no longer become a different section from the rest of the customers. Hernández et al. (2010) adds, while socioeconomic factors do not play a role on the customer behavior in developed countries, they can be a somewhat important factor in underdeveloped countries that will affect the customer behavior. The author discusses that this research is done with ease-of-use and self-efficacy kept in mind, meaning the experience users gather while online shopping like before that we mentioned previously and factors like gender, income level, the age of the customer, do not play a role once the shopper is experienced with online shopping.

The Evolving Research of Customer Adoption of Digital Payment: Learning from Content and Statistical Analysis of the Literature:

One of the most important aspects of the factors that affect their behavior is customer needs. We also know customer needs need to be answered immediately with a positive impact. However, formany people, rather than the needs of the customer and the satisfaction of the customer, the most decisive factor in market competition is usually the price of the product or the service that is given.In this study, Gan and Li (2013), talk about correlation and regression-based analysis, the focus of this paper is to analyze the relationship between the satisfaction of the consumer, consumer trust and the firm's strategy on changing the prices with overall customer commitment. Customer loyalty to a firm is a very important aspect. Strong loyalty will ensure that the consumer will buy from that brand again, will not choose other brands over theirs and will talk positively about the brand to their friends and family which also means a good mouth-to-mouth reputation. For this reason, a company must understand this aspect for the longer life span of the company. In this paper Gan and Li (2013) mentions about their findings from other papers and talk about the hypothesis of satisfaction of the customer is a positive factor that influence the consumer loyalty and changing the price of the product influences consumer loyalty positively. The author continues about their data which contains 300 surveys with %93 acceptable ones and for the reliability of the data authors use Cronbach. The author also states that they used the KMO method which is used for if the data can be used for factor analysis and KMO's of this data was bigger than 0.5, which makes it a positive. This shows that there is a correlation between the loyalty of the consumer with other aspects like we mentioned, price and satisfaction. Now Gan and Li (2013) talk about their hypothesis and if it is true, and with Pearson coefficient, they confirm that there is a positive relationship. The author continues that trusting the brand, the satisfaction they get and the changein price forms the %71 of the customers loyalty to the company. The author also adds that while the changes in price of the product and the trust they have to the company strongly influences the loyalty of the customer, from the data, satisfaction of the customer and the loyalty, while it is evident, it is not strongly connected. The meaning we can get from this study is that customer satisfaction is an important part of the loyalty the consumer shows, nevertheless it is not a factor that can keep the customer loyal to the company when other factors like trust and the price createa stronger bond. Gand and Li (2013) also add that trust is not only with the brand, but it is also with the employees and the product. The author concludes with switching cost or the changing cost, which is the cost of changing from one brand to another, is a type of "investment".

Consumer Behaviour Through Neuromarketing Approach:

The advancement of neuroscience made it possible for neuromarketing to use neuroimaging tools, whether for marketing purposes or to study people's day-to-day behavior[21]. This overview showed how neuroscientific tools can be effectively used to understand individuals' decision-making processes[21]. They described the period of neuromarketing development and its application in evaluating perception of marketing stimuli[21]. After that, they discussed the tools for measuring brain activity and non-brain activity, the pros and cons, what it measures, and when to use each device[21]. Their paper also discussed a number of publications related toneuromarketing[21]. In addition, their paper discussed the ethical issues raised by using these tools to assess human behavior in purchasing decisions[21]. Finally, the challenges of this area and possible future scenarios are discussed.

A detailed behavioral analysis on consumer and customer changing behavior with respect to social networking sites:

The main purpose of the study was to analyze the changing behavior of customers towards different products through customer ratings. The analysis was performed in three main steps, e.g. B. Customer Review Quality Calculation, Customer Behavior Quality Calculation and Review Comparison [22]. Behavior analysis is a type of science that helps to study and understand people's behavior. It examines the factors that influence the behavior of living and nonliving things on a larger aspect. A particular focus is on understanding, describing, predicting and changing behavior. Understanding the behavior of individuals is a tedious task and hence the study used the chi-square method to assess the final result [22]. The newspaper uses customer reviews as a tool to collect primary and secondary data. These reviews inform, inform and explain the benefits of using customer reviews in data collection [22]. It is one of the best and easiest ways to formulate the comparison score between the customer profile and their rating. The results of the study showed that the quality of the product varied depending on the customer profile [22]. In addition, high rates of efficiency and accuracy are observed in collecting information about the products via customer profiles and reviews [22]. The study validates an integrative approach to explaining factors manipulating consumer awareness, observation and sensitivity to product value [22]. A deep analysis was performed to identify the purchase decision, brand behavior, price influence, buying behavior and customer opinion [22]. However, factors such as high quality, social media networks, and customer profiles had a greater impact on customer buying behavior [22].

Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: A machine learning approach:

Customer loyalty is one of the biggest challenges in the telecommunications industry. Organizations may find that predicting customer churn is critical to the success of their business, as careful analysis of churn can be a crucial vehicle for customer retention [23]. Among potentially a multitude of factors affecting churn, it is crucial to identify the most influential ones to target customer retention efforts on. In the article, they compared the performance of different churn prediction models based on real data obtained from a partner company. The prediction models included logistic regression, support vector machines, random forest and decision tree [23]. Additionally, the push-pull mooring (PPM) framework was used to study the impact of features on customer churn behavior from push, pull, and mooring perspectives [23]. A partial least squares (PLS) regression was used to perform the PPM analysis. Furthermore, the behavior of dropouts and non-dropouts was analyzed. The results showed that logistic regression had the highest predictive accuracy, and the percentage push factor of churn emerges as one of the most influential factors influencing customer churn, as churners are more sensitive to service quality than non-churners [23].

1.3.2. Concepts

Shopping is an essential part of daily life. Many people enjoy the thrill of browsing for new items or services. However, many people have reduced their shopping due to the growth of online shopping. Traditional retail stores struggle to compete with the convenience and accessibility of web transactions. That's because web transactions are both cheaper and easier than traditional transactions.

Both online and traditional shopping have their advantages. The most obvious pro of shopping isthat it saves time. Consumers no longer need to travel to different stores or wait in line to purchase items. Plus, retailers no longer need to close down during peak hours to increase their profits. Instead, they can remain open all day and night for shoppers' convenience. Traditional shopping is great for saving time, but online shopping is even better in this regard.

	Physical Retail		Online Retail	
	For Consumers	For Companies	For Consumers	For Companies
Cost	Medium	High	Medium	Low
Complexity	Low	High	High	Medium
Performance	High	Medium	Medium	High
Utility	High	Medium	Medium	High
Availability	Medium	Medium	High	High

Table 1 Comparison of the two conceptual solutions

This table compares the characteristics of physical retail and online retail for both consumers and companies. The cost of physical retail for consumers is listed as medium, while for companies it is listed as high. Online retail is listed as medium for consumers and low for companies. The complexity of physical retail is listed as low for consumers and high for companies, while online retail is listed as high for consumers and medium for companies. The performance of physical retail is listed as high for consumers and medium for companies, while online retail is listed as medium and high respectably. The utility of physical retail is listed as high for consumers and medium for companies, while online retail is listed as medium and high. The availability of physical retail is listed as medium, while online retail is listed as high for both consumers and companies.

2. WORK PLAN

2.1. Work Breakdown Structure (WBS)

Work Breakdown Structure shows the single steps to achieve and form this project and bring it to life. We breakdown the project to make it easier to understand and achieve assigned responsibilities.

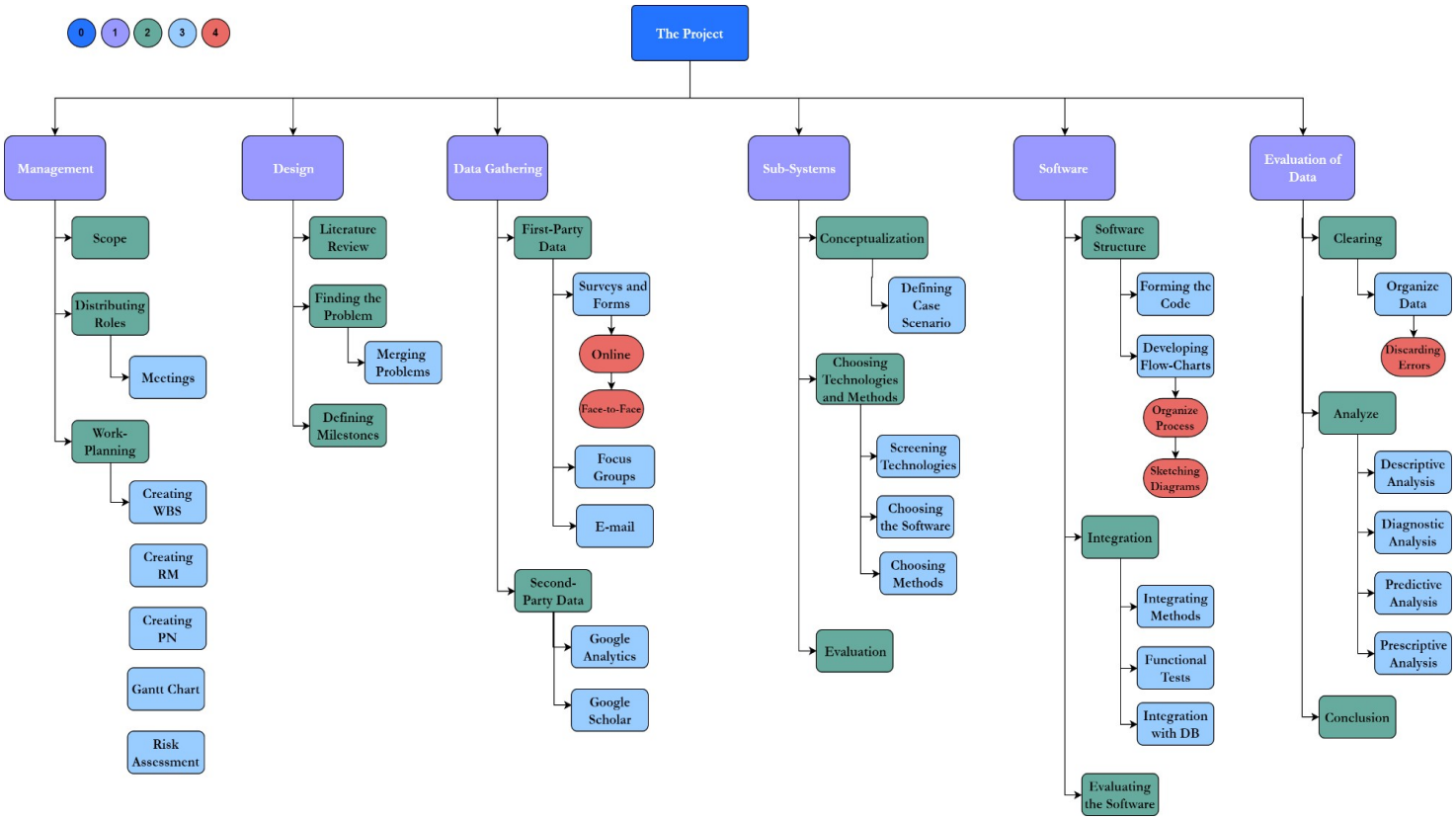


Figure 4 Work Breakdown Structure

2.2. Responsibility Matrix

Responsibility Matrix shows how the team distributed its workforce to certain tasks in theproject. Our project's Responsibility Matrix is shown at the bottom. It shows our team distribution to tasks.

Task	Shukat	Hammam	Hussein	Kerem	Ufuk	Emre
Work Plan				S	R	S
Research	R	S	S	S	S	S
Problem Identification	S	S	R	S	S	S
Data Gathering				S	S	R
Conceptualizing	S	R	S			
Method Choosing	R	S	S			
Analyzing Data	S	R	S			
Evaluation of Data				S	R	S
Integration	S	S	R			
Final Evaluation				S	S	R
Summary and Conclusion	S	S	S	R	S	S
R = Responsible; S = Support						

Figure 5 Responsibility Matrix

2.3. Project Network (PN)

The Project Network provides information to the project members by informing them about the steps of the project and what is required for a step to begin, along with the critical path of the project.

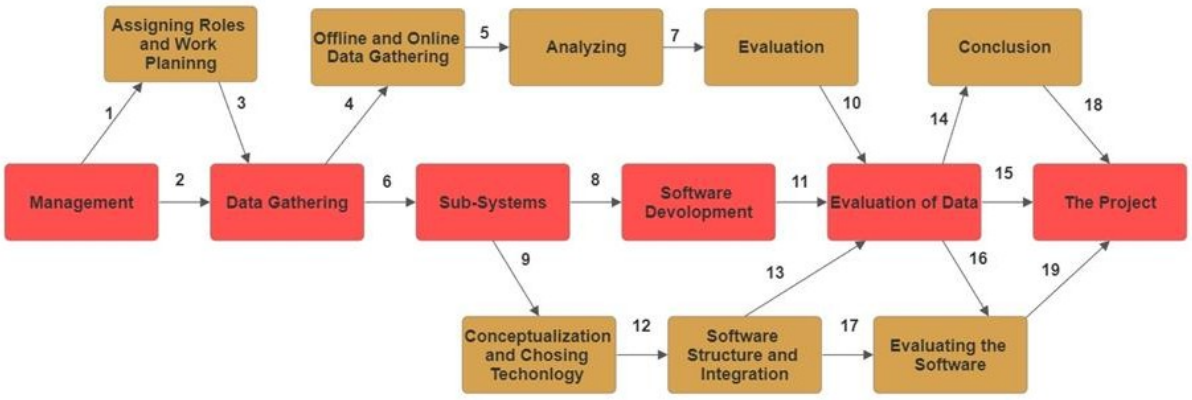
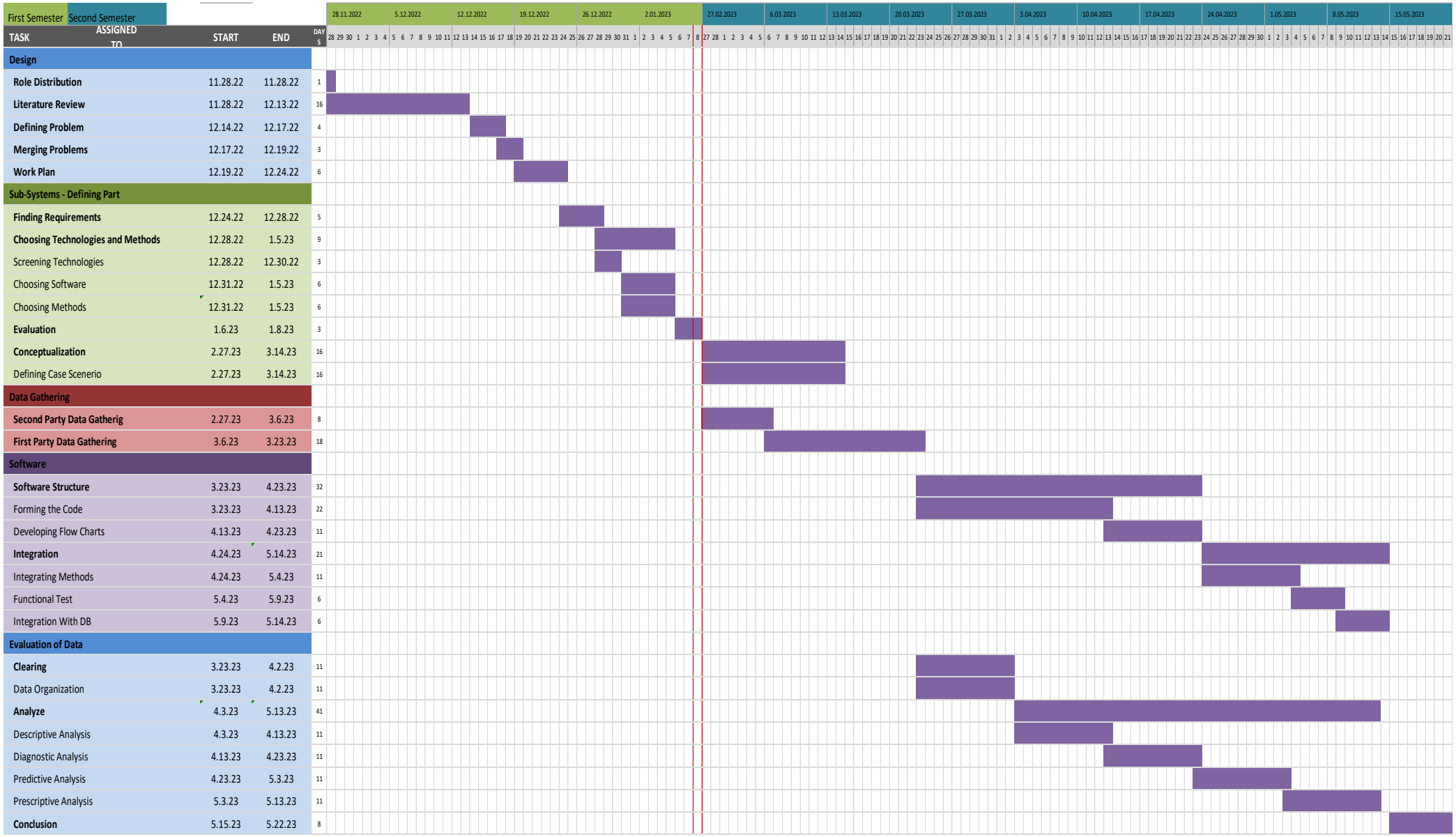


Figure 6 Project Network

2.4. Gantt Chart

Gantt Chart shows deadlines for each task in the project. It helps team members to follow the deadlines and finish the project in time. Our Gantt chart is shown at the bottom. Our team will follow these deadlines.

Figure 7 Gantt Chart of the processes



2.5. Costs

This table provides a detailed breakdown of the costs associated with the project and allows for better management of the budget. It also shows the relationship between the total cost and total value of the project, which can be used to evaluate the project's financial performance.

Software and Mechanical	
Cloud Datasbase	46,72 ₺ / month
UI/UX Design	36,44 ₺ / month
Additional	216 ₺
Total	299,16 ₺
Management	
Materials	57.44 ₺
Travel	109.10 ₺
Utilities	28 ₺
Other	192 ₺
Total	386,54 ₺

Figure 8 Cost of Software, Mechanical, and Management

2.6. Risk Assessment

2.6.1. Risk Matrix

Severity Matrix shows the severity of the problems that may happen in the project. We showed the possible problems and their severity that may occur in our project.

Risk Level		Severity Of The Possible Event			
		Minor	Moderate	Major	
Probability that the event will take place.	Unlikely	Lack of Communication With Advisors	Intra Group Conflict	Poor Code Quality	Very Low / Low
	Possible	In Team Scheduling Problems	failure to find suitable persons for the survey	Saved ML System Data Lost	Medium
	Likely	Unforseen Requirements for Project	Poor Provided to Software Engineers	System Overload	High / Very High

Figure 9 Risks Matrix of the Project

2.6.2. Risk Assessment

At the risk assessment we are able to look at our risks at the workplace to identify those things that may cause harm and find a way to prevent them from happening.

Risk Identified	*Probability*	*Severity*	*Risk Level*	*Action Plan*
Lack of Communication with Advisors	Unlikely	Minor	Very Low	Meeting professors face-to-face and visiting during office hours in order to prevent disruptions in online communication tools due to lack of time and delay in response.
Intra Group Conflict	Unlikely	Moderate	Low	Planning through both online and face-to-face to avoid misunderstandings and increasing the clear usage of English.
Poor Code Quality	Unlikely	Major	Medium	Doing the necessary coding tests. Also, a review and feedback can be taken from an experienced coder.
In team Scheduling Problems	Possible	Minor	Low	Attending meetings regularly and scheduling these meetings at right times for everyone. Also, choosing the right subjects for everyone.
Failure to Find Suitable People for the Survey	Possible	Moderate	Medium	Making preparations by observing people in the right place. Giving this responsibility to the team member who knows environment the most.
Saved ML System Data Lost	Possible	Major	High	Storing the saves into more than one place and keeping them with different people (cloud storage, physical storage etc.).
Unforeseen Requirements for Project	Likely	Minor	Medium	Being foresighted in the future in order to carry out crisis management well and to prevent it in a possible situation.
Poor Data Provided to Software Engineers	Likely	Moderate	High	Completing the survey on time, with the right people, in the right place, and making machine learning synchronization correctly.
System Overload	Likely	Major	Very High	Predicting a possible overload by operating on high-capacity devices and taking action accordingly.

Figure 10 Risk Assessment for the Project

3. SUB-SYSTEMS

3.1. Software Engineering

This sub-system is associated with the data analysis conducted by the software engineering team on the customers' behavior in order to help companies better understand their customers and improve their business strategies.

A data analysis system is a software program that aids companies in comprehending the performance and customer behavior. Data centers can use the analysis to give them the knowledge they need to increase sales, either manually or automatically. Companies use this information in a variety of ways, such as to develop new products and services, educate staff,and deliver product samples to influential customers. By analyzing sales data, companies findways to improve business operations and make better products.

3.1.1. Requirements

ActorName	Description of an Actor
Customer	**A customer is a person who purchases an item from a seller, vendor,or provider in return for cash or another valuable item.**
Data Analyst	**A data analyst is a person whose role it is to collect and analyze datain order to address a certain issue. In addition to spending a lot of time with data, the profession also requires communicating results.**
Data Scientist	**Data scientists decide what questions their team should be asking andthen work out how to use data to respond to those questions. They frequently create predictive models for forecasting and theorizing.**
MachineLearning Engineer	**A machine learning engineer (ML engineer) is a professional in the field of information technology who specializes in the development ofself-contained artificial intelligence (AI) systems that automate the usage of prediction models.**

Table 2 Actor and Description of the Software Application

3.1.2. Use-case Glossary

Use Case Number	*Participating Actor Name*	*Use Case Name /Attribute*	*Description*
U1	*Customer*	*Account*	*Able to create an account in the company for the system and provide a login id and password.*
U2	*Customer*	*Retrieve Account*	*Retrieval of an Account if it is lost or forgotten.*
U3	*Customer*	*Retrieve Password*	*Retrieval of a password if it is lost or forgotten.*
U4	*Customer*	*Log an issue/complain*	*Customer be able to log an issue orcomplain into the system*
U5	*Customer*	*Search for a Product*	*Able to search for products.*
U6	*Customer*	*Order/Buy Product*	*Able to place an order or buy a product.*
U7	*Customer*	*Cancel a Product*	*Able to place cancel orders or refund a product.*
U8	*Data Analyst*	*Sign Up*	*Able to create a private account to sign in into the database.*
U9	*Data Analyst*	*Login*	*Able to log into their account.*
U10	*Data Analyst*	*Logout*	*Able to log out of their account.*

U11	***Data Analyst***	***Monitoring users***	***Able to send an email to the Data Analyst to confirm users' identities of whoever logins into the system for surveillance.***
U12	***Data Analyst***	***Click Data***	***Able to view data from the database using the system.***
U13	***Data Analyst***	***Click Create***	***Able to create data into the database using the system.***
U14	***Data Analyst***	***Click Update***	***Able to update data from the database using the system.***
U15	***Data Analyst***	***Click Delete***	***Able to delete data from the database using the system.***
U16	***Data Analyst***	***Click History***	***Able to check work history***
U17	***Data Analyst***	***Gather data***	***Able to collect data. This can entail completing surveys, monitoring website visitor demographics, or purchasingdatasets from data collection experts.***
U18	***Data Analyst***	***Clean data***	***Duplicate, incorrect, or erroneous data may be present in raw data. Cleaning the data refers to keeping the quality of the data so that your interpretations won't be inaccurate or distorted.***
U19	***Data Analyst***	***Model data***	***This requires developing and designing a database's structural elements. You may decide which data kinds to save and gather, how to tie different data categories to one another, and how the data will actually look.***
U20	***Data Analyst***	***Interpret data***	***Process procedure or trends in the data that could address the current query is a necessary step in data interpretation.***

U21	***Data Analyst***	***Present***	***A significant portion of your job will involve presenting the outcomes of your research. You accomplish this by assembling visual aids like graphs and charts, producing written reports, and delivering information to interested companies.***
U22	***Data Scientist***	***Send Data***	***Able to send analysis of the data to the company management.***
U23	***Data Scientist***	***Confirm Message***	***The System shall confirm to the Data Scientist that their analysis has been sent.***
U24	***Data Scientist***	***Access Button***	***Able to restrict access of other users into their analysis files.***
U25	***Data Scientist***	***Live Data***	***Able to use and view live data from the database***
U26	***Data Scientist***	***View Dataset***	***To gain insights, look for patterns and trends in datasets.***
U27	***Data Scientist***	***Create Model***	***Create data models and algorithms topredict outcomes.***
U28	***Data Scientist***	***Edit Model***	***Edit data models and algorithms to predict outcomes.***
U29	***Data Scientist***	***Upload Model***	***Upload data models and algorithms to predict outcomes.***
U30	***Data Scientist***	***Save Model***	***Save data models and algorithms to predict outcomes.***
U31	***Data Scientist***	***Delete Model***	***Delete data models and algorithms to predict outcomes.***

U32	*Data Scientist*	*Feedback*	*Discuss your recommendations with the senior staff and other teams.*
U33	*Machine Learning Engineer*	*View Model*	*Able to View models and algorithms.*
U34	*Machine Learning Engineer*	*Optimize Model*	*Able to change and optimize models and algorithms.*
U35	*Machine Learning Engineer*	*Test Model*	*Able to test different machine learning models and algorithms.*
U36	*Machine Learning Engineer*	*AI Processes*	*Able to deploy AI processes into production.*
U37	*Machine Learning Engineer*	*Data Control*	*Able to ensure good data flow between the databases and backend*
U38	*Machine Learning Engineer*	*Admin Access*	*Able to limit features access into the system*
U39	*Machine Learning Engineer*	*Search*	*Able to search in the system.*
U40	*Machine Learning Engineer*	*Select Data*	*Choosing relevant data sets.*
U41	*Machine Learning Engineer*	*Select Methods*	*Choosing the best data representation methods.*
U42	*Machine Learning Engineer*	*Examine Data*	*Examining the data's quality.*
U43	*Machine Learning Engineer*	*Analysis Data*	*Analyzing data statistically.*

U44	***Machine Learning Engineer***	***Test System***	***Doing tests for machine learning.***
U45	***Machine Learning Engineer***	***Edit System***	***Systems for training and retraining as necessary.***
U46	***Machine Learning Engineer***	***Upload Features***	***expanding the libraries for machine learning in the system***

Table 3 Use-Case/Attributes of the Software Application

3.1.3. Use-case Scenarios

Table 4 Use-case Scenarios

Use Case ID:	U1
Use Case Name	Account for the application
Use Case Overview	Able to create an account in the company for the system and providea login id and password.
Actor(s)	Customer
Preconditions	System is available, Database is available
Trigger	Select Account button
Steps	1. Access to System 2. Login to the system 2. Select Create Account Button 3. Enter details

Post Conditions	1. Personal details are entered into the user database table In the system 2. new account is created 3. Get feedback on the screen once a new account has been created
Exception Flow	1. If there is already an account, the message "Account alreadyexists" is presented. 2. The warning "Insufficient Information" is displayed if informationis lacking.

Use Case ID:	U2
Use Case Name	Retrieve Account
Use Case Overview	Retrieval of an Account if it is lost or forgotten.
Actor(s)	Customer
Preconditions	System is available, Database is available
Trigger	Select Retrieve Account Button
Steps	1. Click the Login button. 2. Click the Forgot Username button. 3. Type in your information as prompted.
Post Conditions	The user receives their username via email.
Exception Flow	Incorrectly entered personal information

Use Case ID:	U3
Use Case Name	Retrieve Password
Use Case Overview	Retrieval of a password if it is lost or forgotten.
Actor(s)	Customer
Preconditions	System is available, Database is available
Trigger	Select Retrieve Password Button
Steps	1. Click the Login button. 2. Click the Forgot Password button. 3. Type in your information as prompted.
Post Conditions	User receives email with password.
Exception Flow	Incorrectly entered personal information

Use Case ID:	U4
Use Case Name	Log an issue/complain
Use Case Overview	Customer be able to log an issue or complain into the system
Actor(s)	Customer
Preconditions	System is available, Database is available
Trigger	Select Complain Button / There has been a problem that has to be logged.
Steps	1. Click Go after choosing the Log an issue button. 2. On the following screen, complete the issue information and click Submit. 3 A problem was noted in the database.
Post Conditions	The first step is to log and commit the issue to the database.
Exception Flow	1. Incorrect information was entered, like a date 2. The computer asks for a new date.

Use Case ID:	U5
Use Case Name	Search for a Product
Use Case Overview	Able to search for products.
Actor(s)	Customer
Preconditions	System is available, Database is available
Trigger	Select Search Bar / The customer wants to look for a product thatthey want to buy.
Steps	1. Click the search icon 2. Type a product's keyword and click Go. 3. Able to see the searched project
Post Conditions	1. Keyword is compared to the product in the database. 2. A list of products that match the keyword is shown.
Exception Flow	1. No keyword matches 2. No products were returned 3. The choice to change your term or stop the search.

Use Case ID:	U6
Use Case Name	Order/Buy Product
Use Case Overview	Able to place an order or buy a product.
Actor(s)	Customer
Preconditions	System is available, Database is available
Trigger	Select Buy/Order Button / The customer wants to buy a product
Steps	1. Select the product. 2. Click on the product order 3. Enter Address and Payment Information.
Post Conditions	1. Order/Purchase has been made 2. Product is on the way
Exception Flow	1. Error during payment 2. Product out of stocks 3. Unable to delivery to Customers address

Use Case ID:	U7
Use Case Name	Cancel a Product
Use Case Overview	Able to place cancel orders or refund a product.
Actor(s)	Customer
Preconditions	System is available, Database is available
Trigger	Select Cancel Order/purchase Button
Steps	1. Click on the order/purchase's product 2. Click on the cancel product button 3. Select on the reason on cancellation and submit
Post Conditions	1. Product is canceled with valid reason 2. Refund has been returned to the account
Exception Flow	1. Invalid canceling reason 2. Product is not refundable

Use Case ID:	U8
Use Case Name	Sign Up
Use Case Overview	Able to create a private account to sign in into the database.
Actor(s)	Data Analyst
Preconditions	System is available, Database is available
Trigger	Click on the Sign-Up button
Steps	1. Access to System 2. Login to the system 3. Select Create Sign-Up Button 4. Enter details
Post Conditions	1. Personal details are entered into the user database table In the system 2. new User Account is created 3. Get feedback on the screen once a new user has been created
Exception Flow	1. If there is already an account, the message "User Account already exists" is presented. 2. If there is not enough information, the message "Insufficient Information" is shown.

Use Case ID:	U9
Use Case Name	Login
Use Case Overview	Able to log into their account.
Actor(s)	Data Analyst
Preconditions	System is available, Database is available
Trigger	Click on the Sign-In button / User information is entered into the program.
Steps	1. Click the Login button. 2. Type in a username 3. Type in a Password 4. Click "Enter"
Post Conditions	User has signed in to the personal area.
Exception Flow	1. Incorrect user login information 2. asked to enter the login information again.

Use Case ID:	U10
Use Case Name	Logout
Use Case Overview	Able to log out of their account.
Actor(s)	Data Analyst
Preconditions	System is available, Database is available
Trigger	Click on the Logout button / User wants to exit the system.
Steps	1. Click on the logout button 2. Confirm the logout by pressing Yes/No 3. Takes back to Sign in page
Post Conditions	User has signed out to the personal area.
Exception Flow	1. Incorrect user Logout information 2. Error during confirmation of logout

Use Case ID:	U11
Use Case Name	Monitoring users
Use Case Overview	Able to send an email to the Data Analyst to confirm users' identitiesof whoever logins into the system for surveillance.
Actor(s)	Data Analyst
Preconditions	System is available, Database is available
Trigger	Analyst wishes to keep tabs on users' work
Steps	1. Pressing the "User States" button 2. Examine the usage of the user account 3. observe user actions
Post Conditions	Analysts receive an email report detailing user actions.
Exception Flow	1. "Unable to send the report through email" error 2. Found no user actions

Use Case ID:	U12
Use Case Name	Click Data
Use Case Overview	Able to view data from the database using the system.
Actor(s)	Data Analyst
Preconditions	System is available, Database is available
Trigger	Analyst desires to access customer information. Click on “Data”button
Steps	1. Choosing a customer id 2. Selects the "customer data" button. 3. View client information
Post Conditions	Analysts have access to all consumer data.
Exception Flow	1. Data on customers not found 2. Invalid customer ID.

Use Case ID:	U13
Use Case Name	Click Create
Use Case Overview	Able to create data into the database using the system.
Actor(s)	Data Analyst
Preconditions	System is available, Database is available
Trigger	Customer data will be created by the analyst. Click on “CreateData” Button
Steps	1. choosing a customer id 2. clicks the button to create customer data 3. Add client information 4. select the "Submit" button.
Post Conditions	1. Analysts who can generate new customer data 2. Database will be updated with new data.
Exception Flow	1. Incorrect Customer id was used to enter customer data. 2. If there is not enough information, the message"Insufficient Information" is shown. 3. Incorrect Customers Data was entered

Use Case ID:	U14
Use Case Name	Click Update
Use Case Overview	Able to update data from the database using the system.
Actor(s)	Data Analyst
Preconditions	System is available, Database is available
Trigger	Analyst wants to Update customer data/Click on “UpdateData”Button
Steps	1. Select a customer id 2. Clicks on update customers data button 3. Upload Customer Data 4. Click on Submit Button
Post Conditions	1. Analyst able to Upload new data of the customers 2. uploaded Data will be add to database
Exception Flow	1. Unable to upload data to customer 2. Invalid customer id. 3. If there is not enough information, the message"Insufficient Information" is shown.

Use Case ID:	U15
Use Case Name	Click Delete
Use Case Overview	Able to delete data from the database using the system.
Actor(s)	Data Analyst
Preconditions	System is available, Database is available
Trigger	Analyst wants to delete customer data/ Click on “delete Data”Button
Steps	choosing a customer id clicks the button to Delete customer dataconfirm delectation select the "Submit" button.
Post Conditions	1. Analyst able to Delete data of the customers 2. Deleted Data will be removed from the database
Exception Flow	1. Unable to delete data to customer 2. Invalid customer id. 3. If there is not enough authority , the message "Insufficient authorization" is shown.

Use Case ID:	U16
Use Case Name	Click History
Use Case Overview	Able to check work history
Actor(s)	Data Analyst
Preconditions	System is available, Database is available
Trigger	Analyst wants to view data history of a customer / Click on “Data History” Button
Steps	1. Select a customer id 2. Clicks on History customers data button 3. view History Customer Data
Post Conditions	1. Analyst able to view history data of the customers. 2. Analyst able to analyze data from customer history.
Exception Flow	1. Unable to view history data to customer 2. Invalid customer id. 3. If there is not enough authority to view history , the message "Insufficient authorization" is shown.

Use Case ID:	U17
Use Case Name	Gather data
Use Case Overview	Able to collect data. This can entail completing surveys, monitoring website visitor demographics, or purchasing datasets from data collection experts.
Actor(s)	Data Analyst
Preconditions	System is available, Database is available
Trigger	Analyst wants to gather data / Click on “Data gather” Button
Steps	1. Select a customer id 2. Clicks on gather customers data button 3. Gather data Customer Data
Post Conditions	1. Analyst able to gather data of the customers. 2. Analysts are able to analyze data from gathered data.
Exception Flow	1. Unable to gather data for customers from the database. 2. Invalid customer id.

Use Case ID:	U18
Use Case Name	Clean data
Use Case Overview	Duplicate, incorrect, or erroneous data may be present in raw data. Cleaning the data refers to keeping the quality of the data so that your interpretations won't be inaccurate or distorted.
Actor(s)	Data Analyst
Preconditions	System is available, Database is available
Trigger	Analyst wants to clean data / Click on “Data Clean” Button
Steps	1. Select a customer id 2. Clicks on Clean customers data button . Select the important data for Customer Data from database
Post Conditions	1. Analyst able to clean data of the customers. 2. Analysts are able to analyze data in order to select important data and clean the data.
Exception Flow	1. Unable to clean data for customer from database 2. Error "No data found" message

Use Case ID:	U19
Use Case Name	Model data
Use Case Overview	This requires developing and designing a database's structural elements. You may decide which data types to save and gather, howto tie different data categories to one another, and how the data will actually look.
Actor(s)	Data Analyst
Preconditions	System is available, Database is available
Trigger	Analyst wants to Model data / Click on “Data Model” Button
Steps	1. Select a model type 2. Clicks on Model data button 3. Select the important model type for Data
Post Conditions	1. Display of Model type for data
Exception Flow	1. Unable to design Model type 2. Error "No Model found" message

Use Case ID:	U20
Use Case Name	Interpret data
Use Case Overview	Process procedure or trends in the data that could address the current query is a necessary step in data interpretation.
Actor(s)	Data Analyst
Preconditions	System is available, Database is available
Trigger	Analyst wants to Interpret data / Click on “Data Interpret” Button
Steps	1. Select a Data to interpret 2. Clicks on Data Interpret button 3. Select the important Data to interpret
Post Conditions	Display of interpreted data
Exception Flow	Unable to interpreted data Error "No data found to interpreted" message

Use Case ID:	U21
Use Case Name	Present
Use Case Overview	A significant portion of your job will involve presenting the outcomes of your research. You accomplish this by assembling visual aids like graphs and charts, producing written reports, anddelivering information to interested companies.
Actor(s)	Data Analyst
Preconditions	System is available, Database is available
Trigger	Analyst wants to present data / Click on “Data present” Button
Steps	1. Select a Data to present 2. Clicks on Data present button 3. Select the important Data to present to company
Post Conditions	Display Data to present to the company
Exception Flow	Unable to present data Error "No data found to present" message

Use Case ID:	U22
Use Case Name	Send Data
Use Case Overview	Able to send analysis of the data to the company management.
Actor(s)	Data Scientist
Preconditions	System is available, Database is available
Trigger	Analyst wants to Send data / Click on “Data Send” Button
Steps	1. Select a Data to send to analysis 2. Clicks on Data Send button 3. Select submit to send the data
Post Conditions	1. Sent data to the company after it has been analyzed
Exception Flow	Unable to send data Error "No data found to send" message

Use Case ID:	U23
Use Case Name	Confirm Message
Use Case Overview	The System shall confirm to the Data Scientist that their analysis has been sent.
Actor(s)	Data Scientist
Preconditions	System is available, Database is available
Trigger	Received message from the system
Steps	1. Received a message 2. Clicks on the message 3. check the message
Post Conditions	Received a confirmation message from the system
Exception Flow	Unable to received message Error "No data has been sent" message

Use Case ID:	U24
Use Case Name	Access Button
Use Case Overview	Able to restrict access of other users into their analysis files.
Actor(s)	Data Scientist
Preconditions	System is available, Database is available
Trigger	Click on “Access Button” Button
Steps	1. select a user 2. Clicks on Access Button 3. restrict access a user into the system 4. Submit Confirmation to restrict access
Post Conditions	User unable to access into the system Ask for authorization in order to access the system
Exception Flow	Unable to restrict an account/userError “user not found” message

Use Case ID:	U25
Use Case Name	Live Data
Use Case Overview	Able to use and view live data from the database
Actor(s)	Data Scientist
Preconditions	System is available, Database is available
Trigger	Click on “Live Data” Button
Steps	1. select a data 2. Clicks on live Button 3. display live data on the system
Post Conditions	Able to view the data live on the system
Exception Flow	Unable to display live data due to restriction Erorr “Live data unavailable at the moment” message

Use Case ID:	U26
Use Case Name	View Dataset
Use Case Overview	To gain insights, look for patterns and trends in datasets.
Actor(s)	Data Scientist
Preconditions	System is available, Database is available
Trigger	Click on “View Dataset ” Button
Steps	1. select a Dataset 2. Clicks on View Dataset Button 3. display Dataset on the system
Post Conditions	Able to view the Dataset on the system
Exception Flow	Unable to display Dataset due to restriction Erorr “Dataset unavailable at the moment” message

Use Case ID:	U27
Use Case Name	Create Model
Use Case Overview	Create data models and algorithms to predict outcomes.
Actor(s)	Data Scientist
Preconditions	System is available, Database is available
Trigger	Click on “Create Model ” Button
Steps	1. Access to System 2. Select Model type in the system 2. Select Create Model Button 3. Enter details
Post Conditions	1. Model details are entered into the user database table in thesystem 2. new Model is created 3. Get feedback on the screen once a new Model has been created
Exception Flow	1. If there is already a similar Model for a data, the message "model already exists" is presented. 2. If there is not enough information, the message "Insufficient Information" is shown.

Use Case ID:	U28
Use Case Name	Edit Model
Use Case Overview	Edit data models and algorithms to predict outcomes.
Actor(s)	Data Scientist
Preconditions	System is available, Database is available
Trigger	Click on “Edit Model” Button
Steps	1. Select a Model 2. Clicks on Edit Model button 3. Edit Model 4. Click on Submit Button
Post Conditions	1. able to Edit Model 2. Edit Model will add to database
Exception Flow	1. Unable to Edit Model 2. Invalid Model 3. If there is not enough information, the message"Insufficient Information" is shown.

Use Case ID:	U29
Use Case Name	Upload Model
Use Case Overview	Upload data models and algorithms to predict outcomes.
Actor(s)	Data Scientist
Preconditions	System is available, Database is available
Trigger	Click on “Upload Model” Button
Steps	1. Select a Model 2. Clicks on update Model button 3. Upload Model 4. Click on Submit Button
Post Conditions	1. Able to Upload Model 2. Uploaded Model will add to database
Exception Flow	1. Unable to upload Model 2. Invalid Model. 3. If there is not enough information, the message"Insufficient Information" is shown.

Use Case ID:	U30
Use Case Name	Save Model
Use Case Overview	Save data models and algorithms to predict outcomes.
Actor(s)	Data Scientist
Preconditions	System is available, Database is available
Trigger	Click on “Save Model” Button
Steps	1. Choosing a Model 2. Selects the "Save Model" button. 3. Confirm the Submit details
Post Conditions	Able to save the model into the system
Exception Flow	1. Unable to save model 2. Invalid Model. 3. If there is not enough information, the message"Insufficient Information" is shown.

Use Case ID:	U31
Use Case Name	Delete Model
Use Case Overview	Delete data models and algorithms to predict outcomes.
Actor(s)	Data Scientist
Preconditions	System is available, Database is available
Trigger	Click on “Delete Model” Button
Steps	choosing a Model clicks on Delete Model buttonconfirm delectation select the "Submit" button.
Post Conditions	1. able to Delete Model 2. Deleted Model will be removed from the database
Exception Flow	1. Unable to delete Model 2. Invalid Model. 3. If there is not enough authority, the message "Insufficient authorization" is shown.

Use Case ID:	U32
Use Case Name	Feedback
Use Case Overview	Discuss your recommendations with the senior staff and other teams.
Actor(s)	Data Scientist
Preconditions	System is available, Database is available
Trigger	Click on “Feedback” Button
Steps	1. Click Go after choosing the Feedback button. 2. On the following screen, complete the Feedback information and click Submit. 3 A Feedback was noted in the database.
Post Conditions	Feedback will be sent to other users
Exception Flow	1. Incorrect information was entered 2. Unable to send feedback 3. If there is not enough authority, the message "Insufficient authorization" is shown.

Use Case ID:	U33
Use Case Name	View Model
Use Case Overview	Able to View models and algorithms.
Actor(s)	Machine Learning Engineer
Preconditions	System is available, Database is available
Trigger	Click on “View Model” Button
Steps	1. select a Model 2. Clicks on View Model Button 3. display Model on the system
Post Conditions	Able to view the Model on the system
Exception Flow	Unable to display Model due to restriction Erorr “Model unavailable at the moment” message

Use Case ID:	U34
Use Case Name	Optimize Model
Use Case Overview	Able to change and optimize models and algorithms.
Actor(s)	Machine Learning Engineer
Preconditions	System is available, Database is available
Trigger	Click on “Optimize Model” Button
Steps	1. select a Model 2. Clicks on Optimize Model Button 3. display Optimize Model on the system
Post Conditions	Able to Optimize the Model on the system
Exception Flow	Unable to display Optimize Model due to restriction Erorr “Optimize Model unavailable at the moment” message

Use Case ID:	U35
Use Case Name	Test Model
Use Case Overview	Able to test different machine learning models and algorithms.
Actor(s)	Machine Learning Engineer
Preconditions	System is available, Database is available
Trigger	Click on “Test Model” Button
Steps	1. select a Model 2. Clicks on Test Model Button 3. Do a Test on Model on the system 4. Submit result of test
Post Conditions	Able to Test the Model on the system
Exception Flow	Unable to Test Model due to restriction Erorr “Test Model unavailable at the moment” message

Use Case ID:	U36
Use Case Name	AI Processes
Use Case Overview	Able to deploy AI processes into production.
Actor(s)	Machine Learning Engineer
Preconditions	System is available, Database is available
Trigger	Click on “AI Processes” Button
Steps	1. Pressing the "AI Processes" button 2. Select data 3. AI process the data 4. Submit the processed data
Post Conditions	AI Able to process the data into the system
Exception Flow	1. "Unable to process the data " error 2. No Data found for AI to process

Use Case ID:	U37
Use Case Name	Data Control
Use Case Overview	Able to ensure good data flow between the databases and backend
Actor(s)	Machine Learning Engineer
Preconditions	System is available, Database is available
Trigger	Click on “Data Control” Button
Steps	1. Pressing the "Data Control" button 2. Examine the usage of the Data 3. observe Data actions
Post Conditions	Able to observe and examine the data and control the access
Exception Flow	1. "Unable to Control the data " error 2. No Data found to observe and examine

Use Case ID:	U38
Use Case Name	Admin Access
Use Case Overview	Able to limit features access into the system
Actor(s)	Machine Learning Engineer
Preconditions	System is available, Database is available
Trigger	Click on “Admin Access” Button
Steps	1. select a user 2. Clicks on Admin Access Button 3. restrict access a feature into the system 4. Submit Confirmation to restrict access
Post Conditions	features unavailable in the system New features are added in the system
Exception Flow	Unable to restrict an features Error “features not found” message

Use Case ID:	U39
Use Case Name	Search
Use Case Overview	*Able to search in the system.*
Actor(s)	Machine Learning Engineer
Preconditions	System is available, Database is available
Trigger	Click on “Search” Button
Steps	1. Click the search icon 2. Type a feature or a keyword and click Go. 3. Able to see the searched project
Post Conditions	1. Keyword is compared to the feature or data in thedatabase. 2. A list of features and data that match the keyword isshown.
Exception Flow	1. No keyword matches 2. No features and data were returned 3. The choice to change your term or stop the search.

Use Case ID:	U40
Use Case Name	Select Data
Use Case Overview	Choosing relevant data sets.
Actor(s)	Machine Learning Engineer
Preconditions	System is available, Database is available
Trigger	Click on “Select Data” Button
Steps	1. Choosing a Data 2. Selects the "Select data" button. 3. View data information
Post Conditions	has access to all data.
Exception Flow	1. Unable to select data at the moment 2. Error “data not found” message

Use Case ID:	U41
Use Case Name	Select Methods
Use Case Overview	Choosing the best data representation methods.
Actor(s)	Machine Learning Engineer
Preconditions	System is available, Database is available
Trigger	Click on “Select Methods” Button
Steps	1. Choosing a Methods 2. Select the "Select Methods" button. 3. View Methods information
Post Conditions	has access to all Methods.
Exception Flow	1. Unable to select Methods at the moment 2. Error “Method not found” message

Use Case ID:	U42
Use Case Name	Examine Data
Use Case Overview	Examining the data's quality.
Actor(s)	Machine Learning Engineer
Preconditions	System is available, Database is available
Trigger	Click on “Examine Data” Button
Steps	1. select a data 2. Clicks on Examine Data Button 3. Do a Examine on Data on the system 4. Submit result of Examine
Post Conditions	Able to Examine the Data on the system
Exception Flow	Unable to “Examine Data due to restriction Error ““Examine Data unavailable at the moment” message

Use Case ID:	U43
Use Case Name	Analysis Data
Use Case Overview	Analyzing data statistically.
Actor(s)	Machine Learning Engineer
Preconditions	System is available, Database is available
Trigger	Click on “Analysis Data” Button
Steps	1. Select a Data 2. Clicks on Analysis Data button 3. Analyze Data 4. Click on Submit Button
Post Conditions	1. Able to Analyze Data 2. Analyze Data will be added to database
Exception Flow	1. Unable to Analyze Data 2. Invalid Data 3. If there is not enough information, the message"Insufficient Information" is shown.

Use Case ID:	U44
Use Case Name	Test System
Use Case Overview	Doing tests for machine learning.
Actor(s)	Machine Learning Engineer
Preconditions	System is available, Database is available
Trigger	Click on “Test System” Button
Steps	1. select a System feature 2. Clicks on Test System Button 3. Do a Test on System feature on the system 4. Submit result of Test
Post Conditions	Able to Test the System feature on the system
Exception Flow	Unable to Test the System feature due to restriction Error “Test System feature unavailable at the moment” message

Use Case ID:	U45
Use Case Name	Edit System
Use Case Overview	Systems for training and retraining as necessary.
Actor(s)	Machine Learning Engineer
Preconditions	System is available, Database is available
Trigger	Click on “Edit System” Button
Steps	1. Select a System feature 2. Clicks on Edit System button 3. Edit System feature 4. Click on Submit Button
Post Conditions	1. able to Edit System feature 2. Edit System feature will be added to database
Exception Flow	1. Unable to Edit System feature 2. Invalid System feature 3. If there is not enough information, the message"Insufficient Information" is shown.

Use Case ID:	U46
Use Case Name	Upload Features
Use Case Overview	expanding the libraries for machine learning in the system
Actor(s)	Machine Learning Engineer
Preconditions	System is available, Database is available
Trigger	Click on “Upload Feature” Button
Steps	1. Select a Feature 2. Clicks on Upload Feature button 3. Upload Feature 4. Click on Submit Button
Post Conditions	1. able to Upload Feature 2. uploaded Feature will be added to database
Exception Flow	1. Unable to upload Feature 2. Invalid Feature. 3. If there is not enough information, the message"Insufficient Information" is shown.

3.1.4. Technologies and methods

For the software engineering team, the analysis will be conducted with python programming language, in Jupyter Notebook with Visual Studio Code environment. We will use multiple python libraries such as Panda, Numpy, Scikit-Learn, and TensorFlow. These libraries will helpus preprocess our dataset, conduct analysis, and visualize our findings. The data to be analyzed will be stored in a database in Microsoft SQL Server. The method we will use is k-means clustering, which is an unsupervised learning algorithm commonly used in market segmentationand search engines. The algorithm classifies a dataset into a number of clusters, k, which is predetermined, then the clusters are positioned as points and all observed data points are associated with the nearest cluster. The observed data points are then computed and adjusted.
This process repeats until a desirable result is reached [24].

This figure is an Entity-Relationship (E-R) diagram. It is used to visually represent the relationship between entities in a database. The diagram includes symbols such as rectangles representing entities, and lines connecting them which indicates the relationship between the entities. Each rectangle represents an entity and contains the entity name and its attributes. The lines connecting entities represents the relationships, such as one-to-one, one-to-many, or many-to-many. The diagram can be used to understand the structure of the data, how the entities are related, and how data is stored and accessed. The E-R diagram is a useful tool for database designers, developers, and administrators to understand and manage the data in a database. It is a way to express the logical structure of databases. It is the foundation for the design of a relational database.

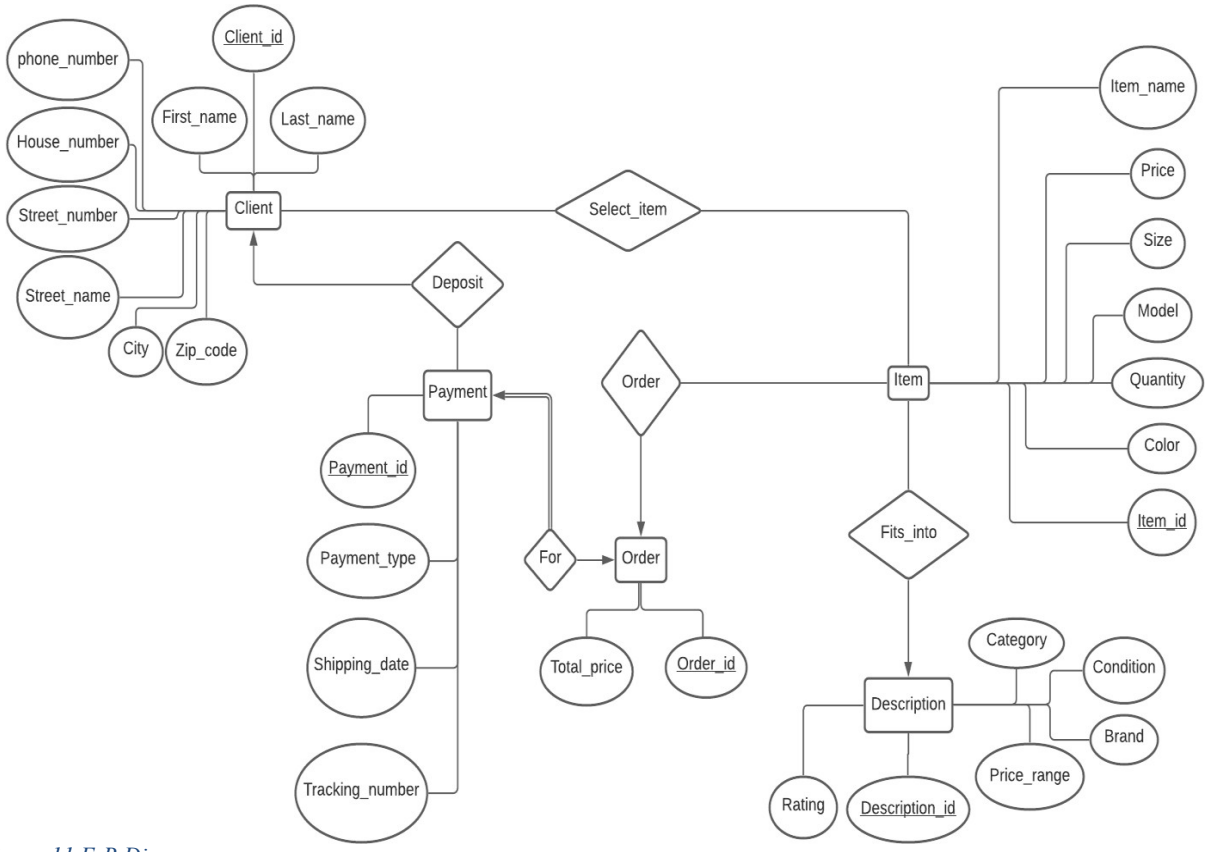


Figure 11 E-R Diagram

This diagram shows the relationship between entities in the databas

This figure is a schema diagram. It is used to visually represent the tables and their relationships in a database. The diagram includes symbols such as rectangles representing tables, and lines connecting them which indicates the relationship between the tables. Each rectangle represents a table and contains the table name and its attributes. The lines connecting tables represents the relationships, such as one-to-one, one-to-many, or many-to-many. The diagram can be used to understand the structure of the database, how the tables are related, and how data is stored and accessed. The schema diagram is a useful tool for database designers, developers, and administrators to understand and manage the database.

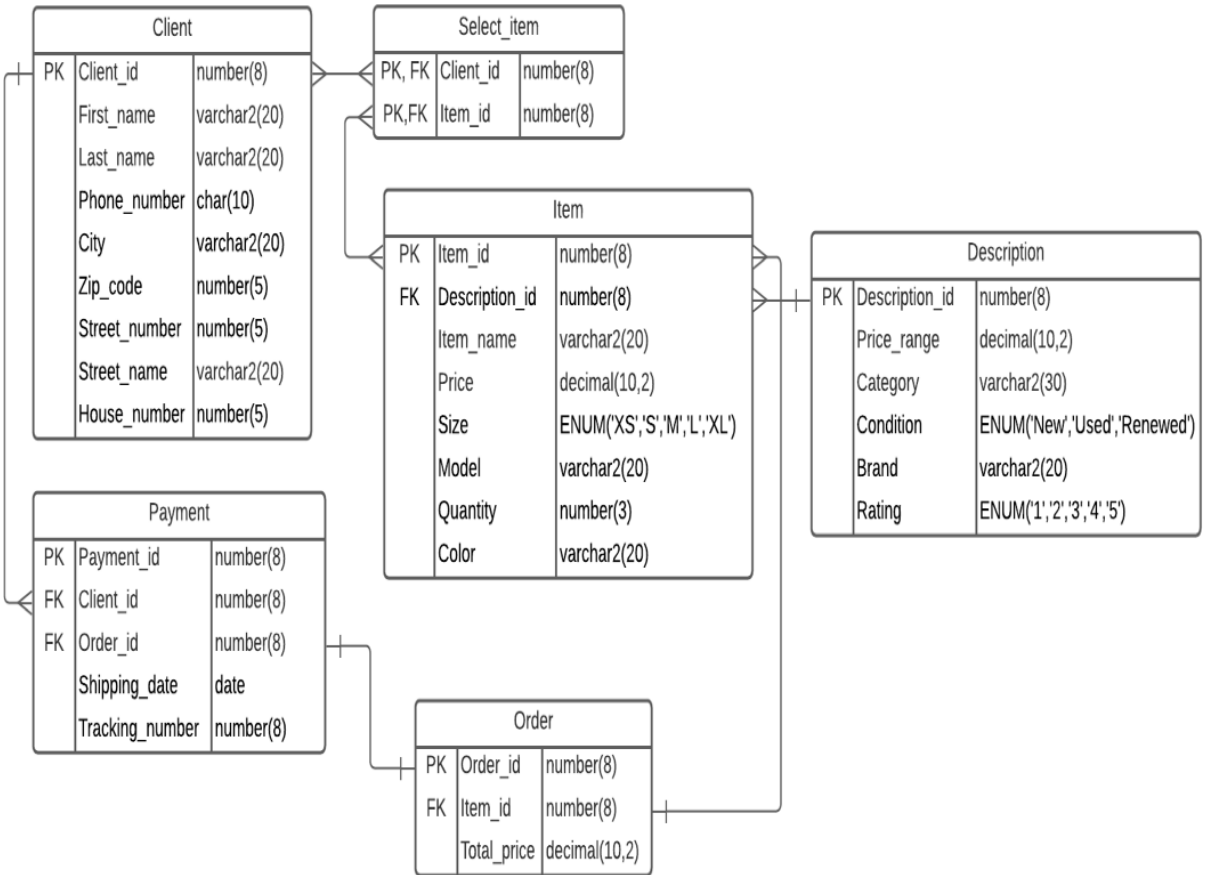


Figure 12 Schema Diagram

This diagram shows the tables and their relationship in the database

3.1.5. Conceptualization

The first diagram depicts the process of a customer registering their contact information into a database. The customer is prompted to enter their contact information, such as their name, phone number, and email address. Once the information is entered, it is then saved into the database. The customer can then be contacted by the company through this contact information. It can also be used for future reference by the company. The diagram shows the process of collecting valuable data and storing it into a database.

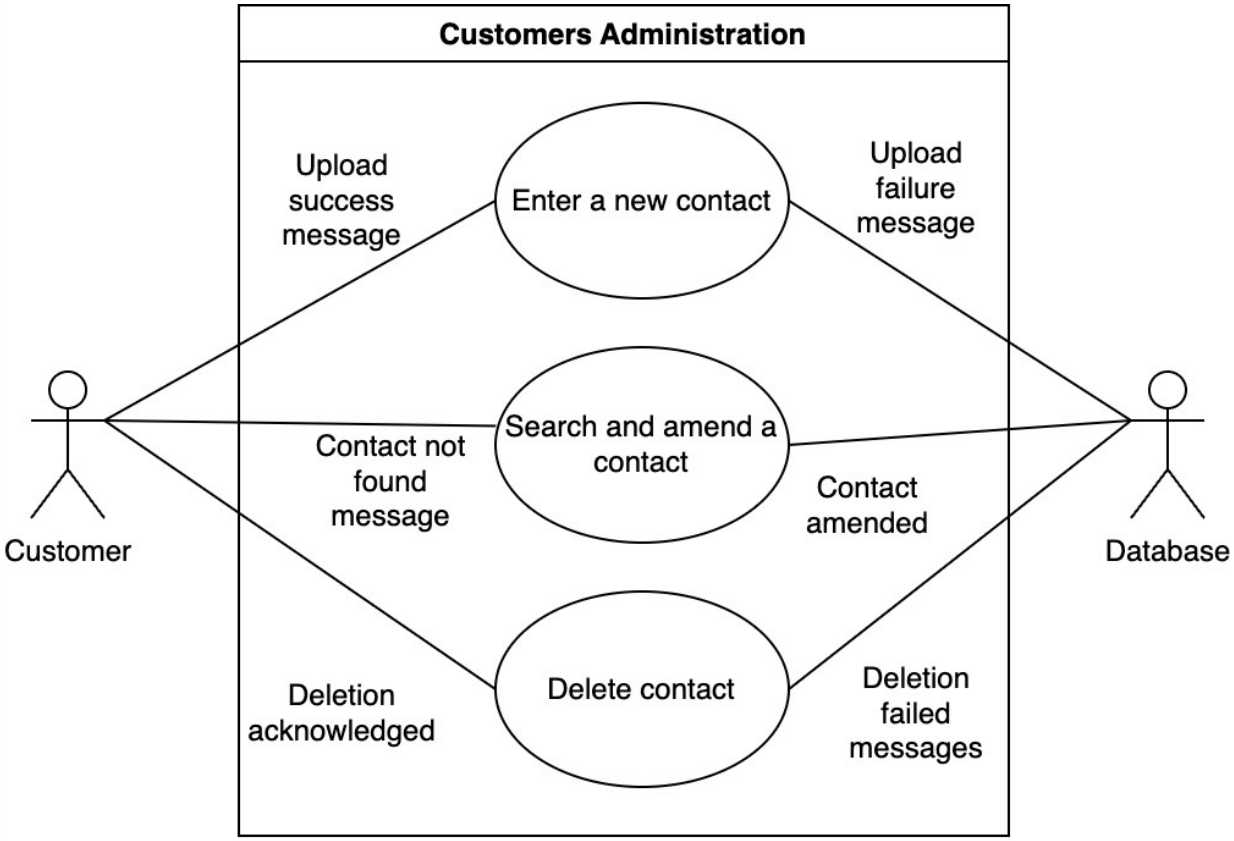


Figure 13 Customers Administration

Customer able to register into the database through the system

The second diagram illustrates the process of a customer creating an order and the order being saved into a database. The customer is shown selecting items from a list, which are then added to their order. Once the order is complete, it is saved into the database. The data analyst will have access to the database and can monitor the customer's order history. This allows the data analyst to analyze customer purchasing patterns and make informed decisions about inventory and sales strategies. The diagram represents how the data collected by the database can be used to improve the company's performance and customer satisfaction.

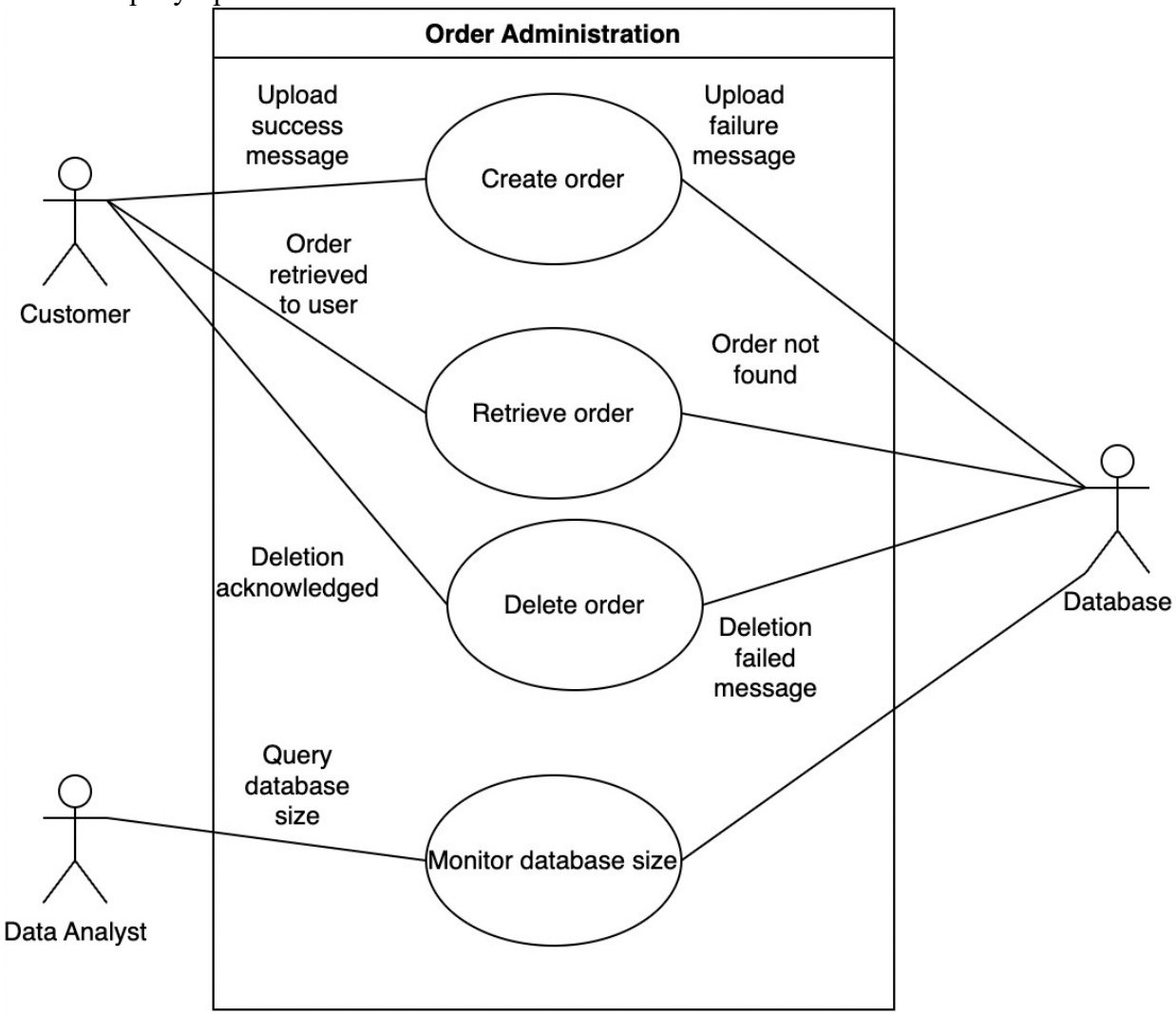


Figure 14 Order Administration

Data Analyst monitors the Customer's activities from the Database through the system

The third diagram depicts a subsystem use case model, which shows how different parts of a system interact with each other. The diagram shows the flow of data between the customer, the company system, and the data analyst and data scientist.
The customer is shown providing data, which is then saved into the company's database. The ML Engineer is able to view the data in the database from the company's system. The data analyst and data scientist are then able to conduct their analysis on the customer data. This allows them to gain insights and make decisions based on customer behavior and preferences.
The diagram illustrates how different parts of a system work together to collect and analyze data, which can be used to improve the performance of the company and provide a better experience for the customer.

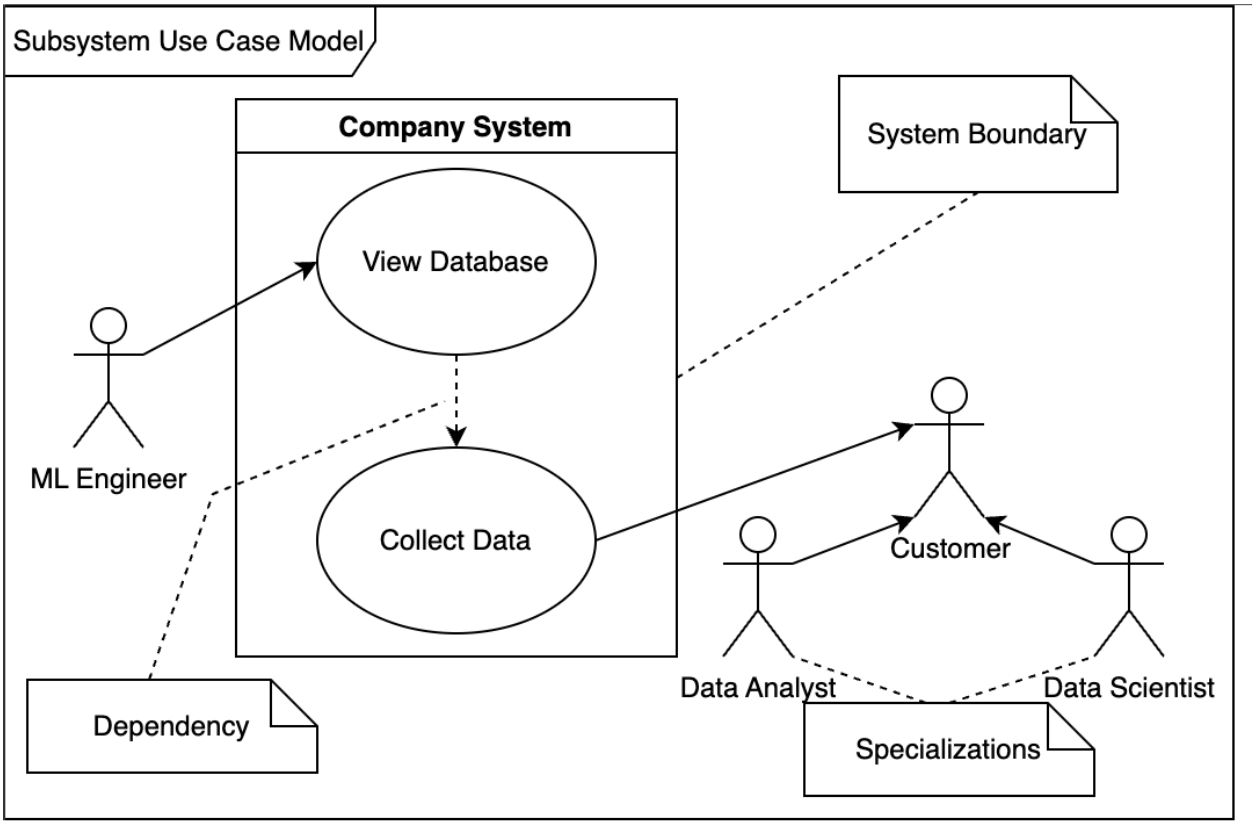


Figure 15 Company System

Data Analyst and Data Scientists analyze and model the data from the Customers while the ML Engineer improves the services of the company based on the data collected

The final diagram shows a subsystem use case model on a macro scale, illustrating how the different parts of an ML system interact within the company's overall system.
The diagram depicts the flow of data between the customer, the company's system, the ML engineer, the data scientist, and the data analyst. The customer is shown providing data, which is then saved into the company's database. The data scientist is able to create visualizations of the data, which can be used to gain insights and make decisions. The ML engineer is able to create models using the data, which can be deployed into production. The data analyst is able to conduct CRUD operations on the customer's data, allowing them to manage, manipulate, and query the data in the database.
The diagram illustrates how the different parts of an ML system work together to collect, analyze, and process data, which can be used to improve the performance of the company and provide a better experience for the customer. It also shows how the ML system is integrated into the company's overall system and how it interacts with other subsystems.

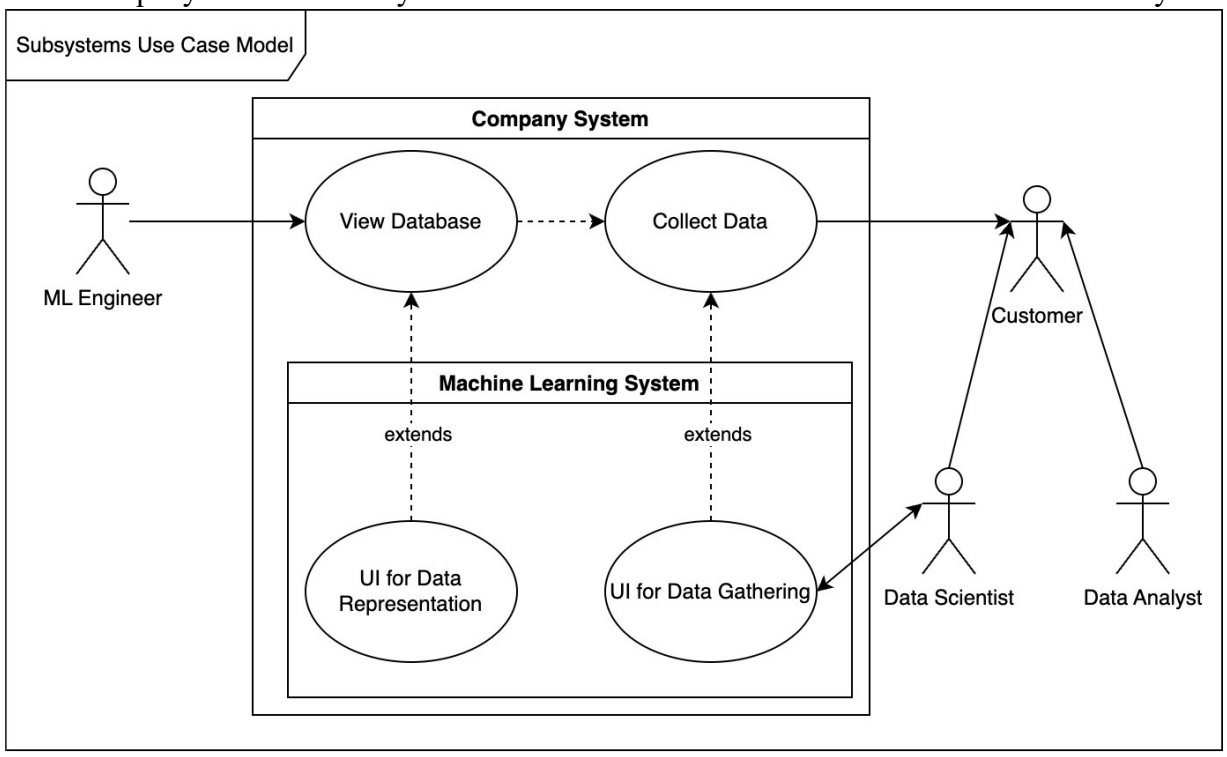


Figure 16 Company Subsystems with ML

Data Scientist will build the visualizations of the analysis, done on the Customer along side the Data Analyst, while the ML Engineer build the subsystem that improves the services of the Company

3.1.6. Materialization

This chart depicts the total revenue generated from the sale of goods for an online business. It shows the overall sales for the business over a specific period of time. The data is represented in the form of a line chart, with the x-axis representing the time frame and the y-axis representing the revenue in currency units. The chart provides a visual representation of the business's sales performance over time and can be used to identify trends, patterns, and changes in revenue.

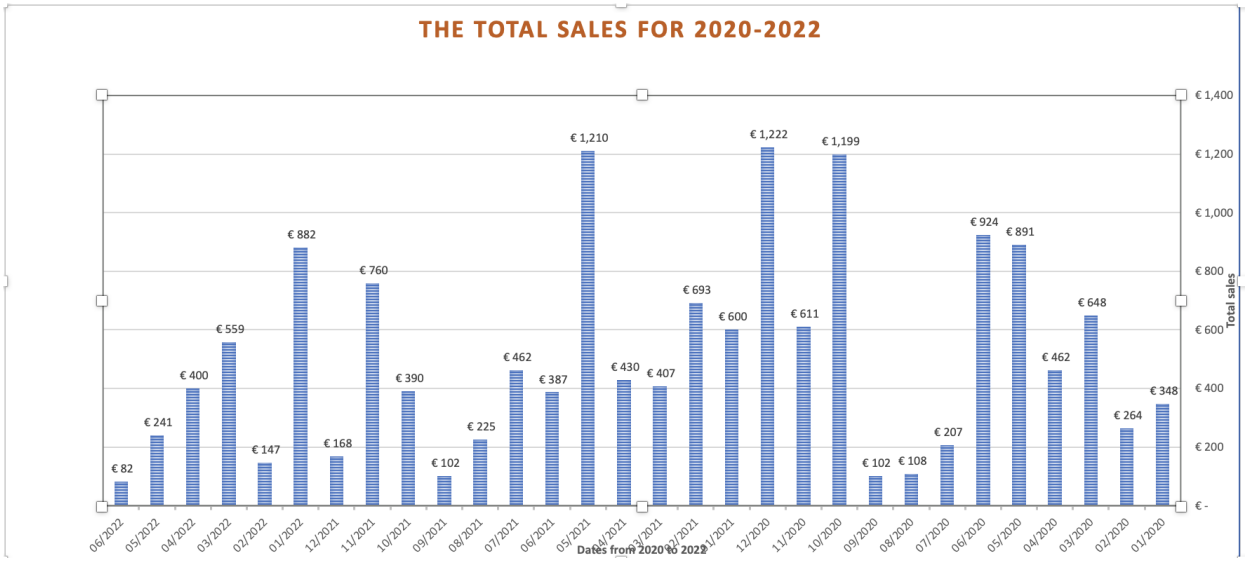
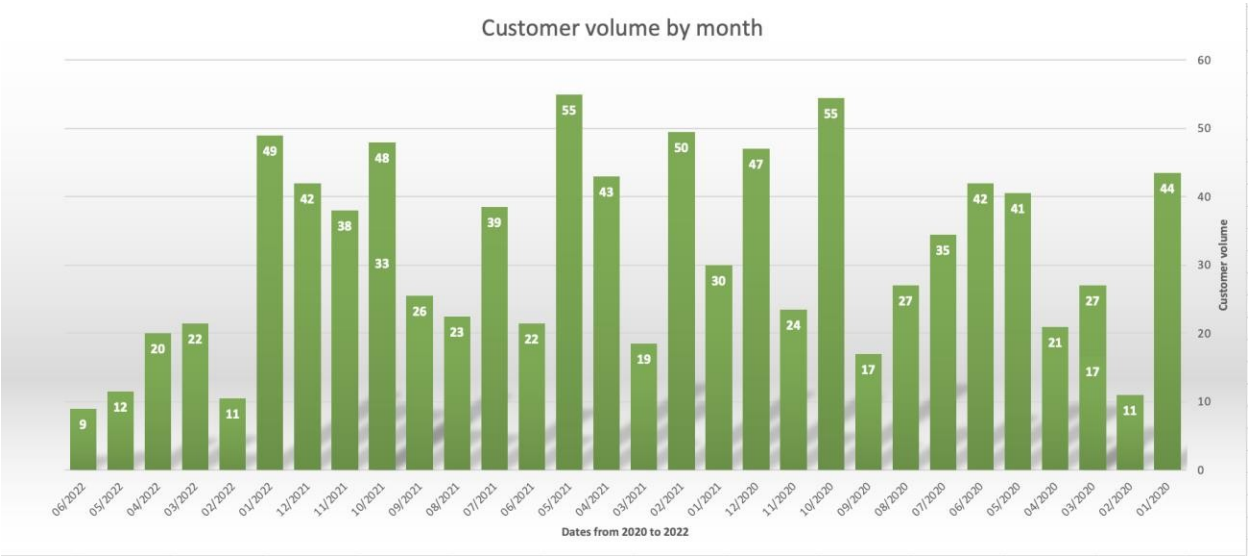


Figure 17 The Total Sales For 2020-2022

This chart is designed to help evaluate data to determine the number of consumers per month, and which month has the highest volume of customers. The data is represented in the form of a bar chart, with the x-axis representing the months and the y-axis representing the number of customers. Each bar in the chart represents the number of customers for a specific month. By comparing the heights of the bars, one can easily determine which month had the highest volume of customers. The chart provides a visual representation of the number of customers per month and can be used to identify trends, patterns, and changes in consumer behavior.

Figure 18 Customer Volume by Month



This chart is designed to help evaluate data to determine the monthly quantity sold and the month with the highest quantity sold. The data is represented in the form of a bar chart, with the x-axis representing the months and the y-axis representing the quantity sold. Each bar in the chart represents the quantity sold for a specific month. By comparing the heights of the bars, one can easily determine which month had the highest quantity sold. The chart provides a visual representation of the quantity sold per month and can be used to identify trends, patterns, and changes in sales. The information can be used to enhance the goods and services offered during the time when the highest quantity is sold.

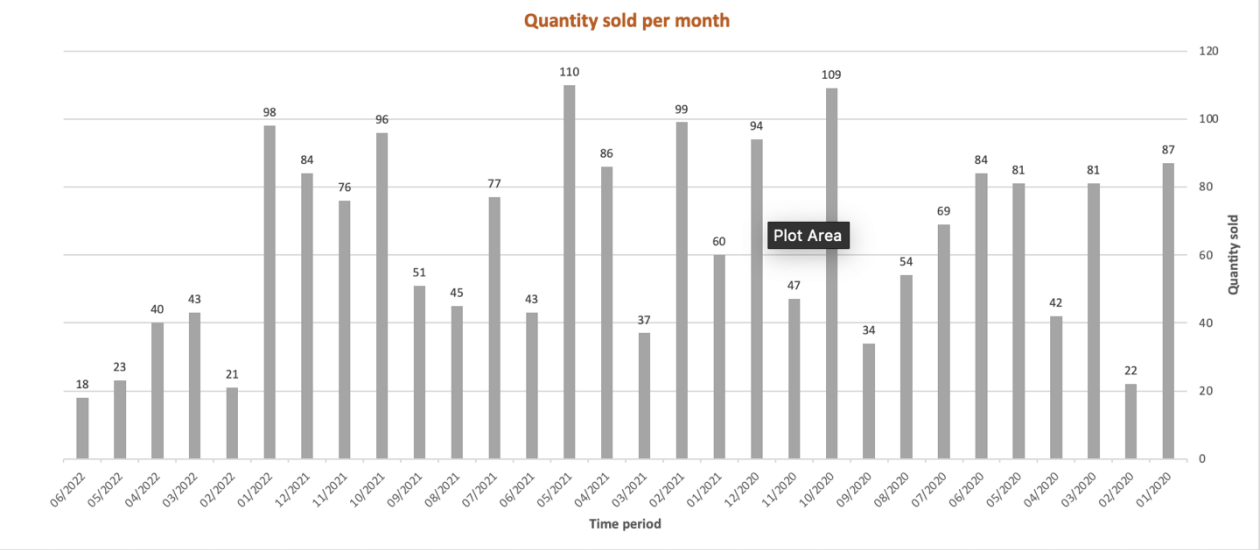
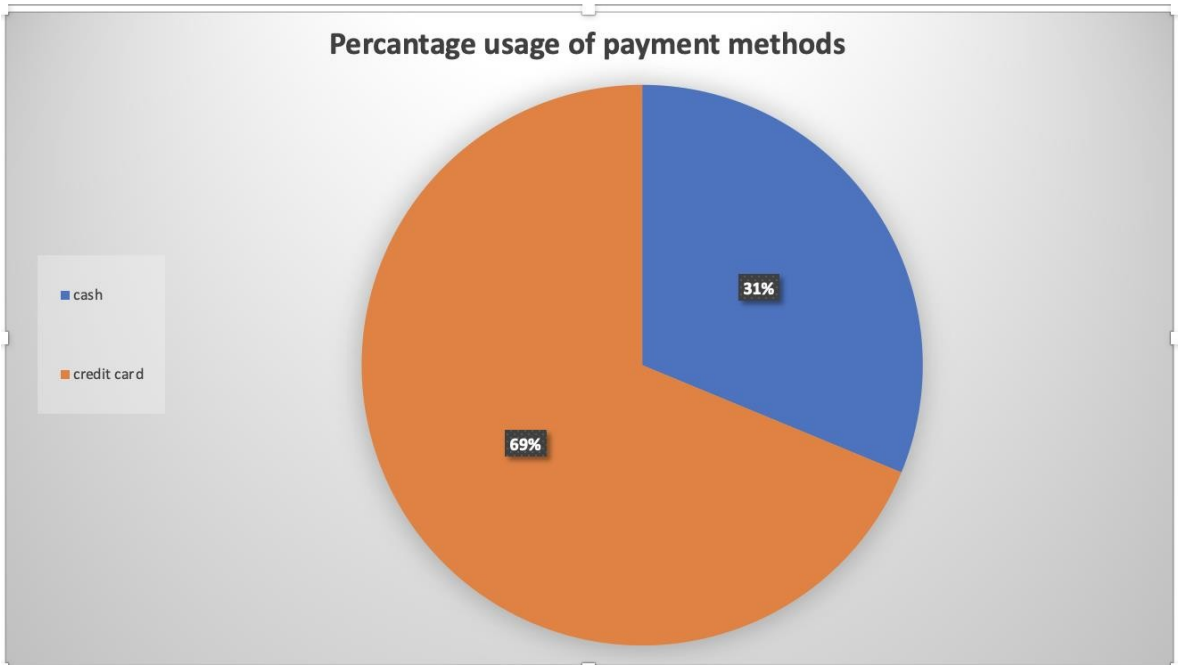


Figure 19 Quantity Sold Per Month

This pie chart is designed to help research customer behavior for the payment methods. It shows the percentage of customers that prefer each method of payment, whether it is cash or online. The chart is divided into different sections, with each section representing a different payment method. The size of each section is proportional to the percentage of customers that prefer that method. By looking at the chart, one can easily determine which payment method is the most popular among customers. This information can be used to make the payment process better for customers by offering the payment methods they prefer.

Figure 20 Percentage Usage of Payment Methods



3.1.7. Evaluation

Customer behavior analysis is a kind of analysis used to comprehend and spot trends in how customers engage with a business, such as The overall revenues, Customer volume, Sales volume, and Payment options

Data mining businesses can use data mining techniques to examine massive amounts of data, such as transaction histories and user interactions with their website, in order to spot trends in their clients' online activity.

In order to track and manage customer interactions with a business, including sales, support, and marketing activities, customer relationship management systems can be employed.
Businesses may spot possibilities to enhance their goods and services by studying consumer behavior. They can also spot possible concerns or problems that clients could be having. This kind of study may be very helpful for businesses trying to maximize their marketing initiatives and comprehend their target market.

Whatever method is employed, it's critical to have a specific end in mind when looking at consumer behavior data. This might be done to enhance revenue, boost customer happiness, or promote client retention. Businesses may utilize data analysis to better understand and serve their consumers by having clear goals and utilizing data to guide decision-making.

The association between consumer purchases and the season may be depicted on a graph, as may the correlation between customer happiness and the volume of customer service contacts. These observations may be useful in identifying the parts of the company that are operating efficiently and those that may require improvement.

Customers desire a simple, efficient payment procedure with a wide range of payment alternatives. Credit cards, debit cards, cash left on the doorstep, electronic payment methods like PayPal, and mobile payment methods like Apple Pay and Google Pay might all be considered.

3.2.1. Review of technologies and methods Data Collection:

There are many ways to collect data, every different data allows you to understand different perspectives. For every data, collector must be sure, the data relates to the problem. To achieve this, data collection must be done properly.

There are 2 types of data, first one is 2nd party data. Which is data you found from other sources rather than collecting yourself. Second one is 1st part data, which is a data collected directly from the you. 1st party data is not enough, in this project we need both of these data types. 1st party data is not enough, 2nd party data is also crucial for the project. We need to compare these two data in the data analysis part to understand the differences.

2nd party data is easy to collect, we can simply find it on the internet. But 1st party data collection is not that easy.

For our project 1st party data will be collected through survey`s. Survey needs to be specifically given the data this project needs, the team must prepare the survey carefully and ask every question needed for the data analysis part. After completing the survey, this must ask to people relevant to the problem. In our case survey need to ask to every age group who uses online shopping or who used online shopping in the pandemic area. Every age group must be equal or close to equal in number.

There will be face-to-face surveys and online surveys, for face-to-face surveys, survey document is needed. For online surveys, website is needed to put survey. In our case we plan to use google forms. It will be easy to manage the survey results.

Data Analysis:

The total data analysis procedure covers various technical skills and abilities, but it mainly involves gathering, categorizing, and analyzing data. From the behavior analysis perspective this analysis gives so much information to companies and in the project, it will give it to us. With a good data analysis, we can predict the future behaviors of the online shopping customers. Also, we can understand why covid changed behavior of the customers. With these predicts most of the post-covid behavior changes becomes clearer. With this analysis results companies might find a solution to hold the customers that used online shopping in the pandemic area. From these we can understand that Data analysis part is one of the most crucial parts of the whole project and needs to be done perfectly.

In this procedure the most needed requirement is data analysis tools, one of the most used tools is excel, but there are many others and, in this project, we will also need R Project, which will allow us to do Correlation analysis, Regression analysis, Cluster Analysis etc.

After completing these requirements, team needs to make analysis out of the raw data from that we collected from surveys. Without making the analysis team cannot move to other steps of the project.

Evaluation and Analyzing Data:
Some data needs to be analyzed with various graphs. This procedure is a must because some data analysis tools don't give the full outcome of the data and to understand more, graph analysis needs to be done. This gives the behavior of the age groups, genders, people who has different income etc. Without this part data analysis part alone is not enough. This procedure does not require many tools, excel is enough to make every possible graph needed. After all these evaluation parts require all the data and comments done in the recent parts.

Correlation Analysis:

What is it: Correlation Analysis, in simple terms, means that similar to other analyses, is used to, as we can gather from its name, form a correlation between variables. There are two types of correlations for this method, the first one is low correlation which are things not very correlated like how you like your coffee and what kind of a car you drive and there is high correlation which are things that are highly correlated like your BMI and your risk of having a heart attack. There is also correlation coefficient which is a number that if it is less than zero there is a negative correlation and if it is positive then there is a positive correlation betwixt different variables.

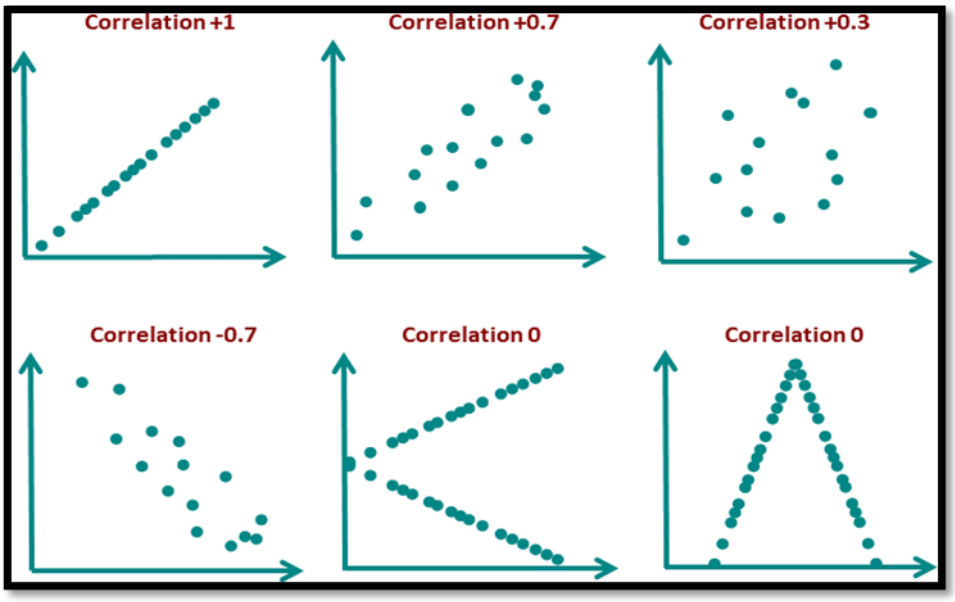


Figure 21 Correlation Graph Types

Why are we using it: While correlation does not cause causation, it will help us get close to create a conclusion. We will be using correlation analysis at finding an if there is a correlation between two of our variables. For example, the correlation between Covid-19 pandemic and the changes in online shopping user count or the age of the participant and the actually purchasing behavior of the customer.

Cluster Analysis:
What is it: In basic terms, Cluster Analysis or Cluster is a part of a learning a machine/computer does in addition to its being unsupervised. By unsupervised, it means that the machine will try to form a connection between variables by their familiar characteristics without us, the user, giving it a pattern, an example, or a right answer. An example would be to give the machine two types of pictures that were shot in fall and spring and without any help it will try to correlate both variables.

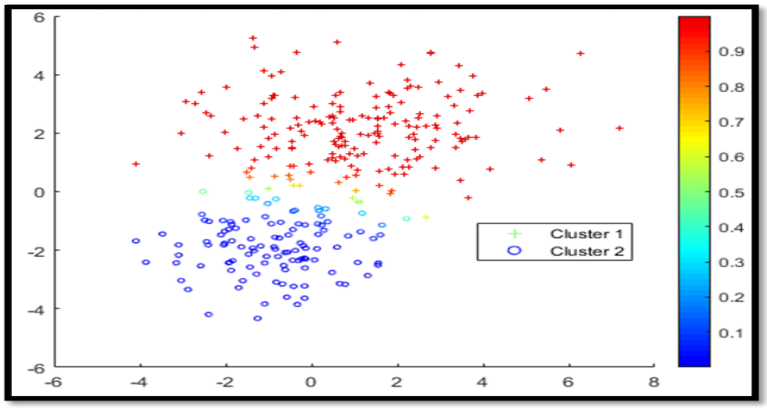


Figure 22 Cluster Graph Example

Why are we using it: For this project as we mentioned, we want to categorize our participants and data into masses or groups and try to find familiar variables that will affect the conclusion and analysis. For our dataset, we will be using a K-Means clustering method as our dataset will not be very big.

For example, we will be gathering such data as:

- *Age*
- *Gender*
- *Income*

and try to create clusters such as:

- *Lower Payers – Older Age*
- *Lower Payers – Younger Age*
- *Higher Payers – Older Age*
- *Higher Payers – Younger Age.*

The main point is to use this data and make a robust connection between variables and try to see the changes in online shopping behavior pre-SARS-CoV-2 and post- SARS-CoV-2.

Regression Analysis

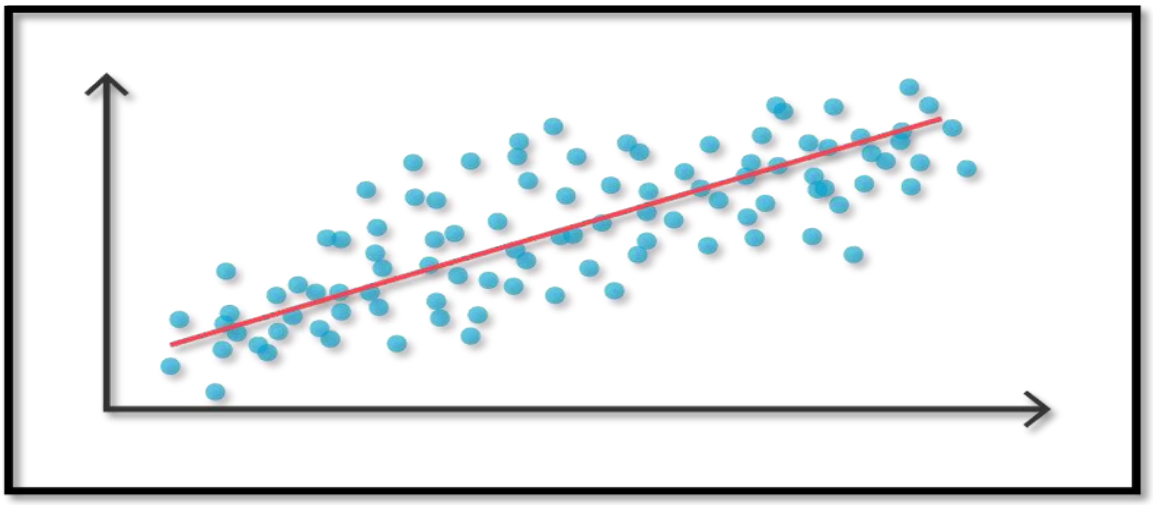


Figure 23 Regression Analysis Display

What is it: Without going into mathematics, in simple terms, Regression analysis is a type of analysis that observes factors and tells us which one's matter to help our cause. Not all factors we include and gather in our research actually affect our journey to reach a solution and if they do in fact affect, not all of them affect at the same level

With this analysis, we will be able to tell that what variable actually matters as this analysis lets us to see in which variables there is a stable and strong connection, so they affect each other enough to form a conclusion.

Why are we using it: In our problem, there are many factors we want to observe like gender, age, race or even income, therefore there will be a lot of factors and variables that will affect our conclusion and analysis. An example would be to connect the age of the individual with the website view time. With this many variables it will be easier for us to see which variables help us by using Regression Analysis.

3.2.2. Proposed Solution Approach

For the main problem of our project, there are some solutions and at the final, there is a merged solution. We can offer this solution to the customers.

At this solution, problem is that people want to see more clearly the product they will buy. If the product is a shirt, they want to make sure it will fit them. In the pandemic area they could not choose between going to the mall or shopping from online, but now they can, and this problem became really big. At this solution we offer 3D videos of the product, if the product is wearable then we will ask customers their height, weight etc. and we will show the video related to that information. With this we will also collect data and in the other buying attempts we aim to show video directly without asking the sizes.

Another solution: a shorter and smoother way to pay.

As mentioned in our problem, the older part of the age spectrum seems to have a difficulty paying or even if they can, if the process takes too long there will be issues like trusting the website if it is not known even though it is safe. Moreover, we know that as the research suggests, older people are a highly profitable part of the customer base for the reason being that in most cases older people tend to have a lower debt ratio and more money to spend on their free time. That is why we want a system that will make it easier to make payments in online shopping and make them shorter and smoother. We will be integrating a system that will collect all consumers cards with their permission of course, and with the click of purchase they will be able to purchase it without going through all the safety issues, making it shorter to pay. One issue with this thing is that although it makes paying easier for the not experienced individuals, this process would be a costly concept as this requires a talk with the banks therefore it is a beneficial solution but not a low cost one.

Another solution is a way to suggest products from the beginning user registration, later it will change with the consumption behavior.

Normally websites show its users a front page covered with products, websites show, chooses these products from analyzing user movements and consumption behavior. We aim to ask some questions at the beginning of the registration process, those questions will be additional, users will choose if they want to answer them or not. The questions will be their age, gender, did they begin shopping online with the pandemic. When they answer the questions we will show them a front page, created with the survey data results. With this we aim to make users' first experience better.

For the final solution, companies need to satisfy all the needs of the customers, to be able to do that, they need to understand the needs, and find many solutions for all the needs. In our case we searched and found some problems, but if a company uses one of these solutions, which may not be enough, to be able to change what customers think about online and how they behave we need to merge all the solutions, because every solution is needed to solve the whole problem. Without implementing and using all solutions, some people that would change their choice after solution would still choose offline stores over online.

4. INTEGRATION AND EVALUATION

4.1. Integration for Software Engineering

4.1.1. Introduction

A particular type of unsupervised learning algorithm called clustering involves assembling comparable data points based on their shared traits. Finding traits in common within a dataset while maintaining dissimilar data points separate while grouping like data points together is the aim of clustering.

4.1.2. Dataset before Covid

4.1.2.1. About Dataset

Customer Data is a vast dataset that offers a thorough study of a business' ideal clients. This dataset offers useful insights that can help a firm better understand its customers by gathering and analyzing information about clients through loyalty cards.

The dataset, which has 2000 records and 8 columns, offers a plethora of details about the clientele of the store. The unique Customer ID, Gender, Age, Annual Income, Spending Score, Profession, Work Experience, and Family Size are all represented by a different column, and each one corresponds to a different characteristic of the customer's profile.

Businesses can obtain important insights into the interests, actions, and spending patterns of their customers by examining this data. To better understand how these variables affect customers' purchase decisions, for instance, businesses can segment clients by age, income, or family size.

- **Consumer ID**: A special number given to each consumer in the dataset. It is employed to identify specific clients and keep tabs on their purchases and other actions.
- **Gender**: The customer's gender, either a man or a woman. Gender can be used to compare gender differences in preferences and purchasing behavior.
- **Age**: The customer's age, typically expressed in years. Customers can be divided according to their age, allowing for the identification of purchasing trends and preferences within various age groups.
- **Annual Income**: The customer's yearly income, typically expressed in dollars or a different currency. Customers can be divided into different income groups based on their annual income, which can be used to determine the buying habits and preferences of people with different income levels.
- **Spending Score**: A rating given by the store depending on how the customer behaves and their propensity to spend. Based on their buying habits, clients can be divided into groups using this score, such as those who tend to make impulse buys, high-spending customers, and low-spending customers.
- **Profession**: The customer's line of work or line of business. Analyzing purchase trends and preferences across various occupations is possible using profession.

- **Work Experience**: The customer's total number of years of employment. Customers can be divided into groups according to their experience level with this functionality, which can be used to spot purchasing trends and preferences across a range of experience levels.
- **Family Size**: The number of family members in the customer's family, typically expressed as a number. When comparing various household sizes, such as households with children versus households without children, family size can be used to examine buying habits and preferences.

4.1.2.2. Set Up

Importing all required modules, setting hyperparameters, and imposing constraints are all crucial steps in getting ready for data loading, preprocessing, and model construction. To guarantee accuracy and consistency in the analysis, these settings will be used consistently across the notebook.

4.1.2.3. Data Loading & Processing

	CustomerID	Gender	Age	Annual Income ($)	Spending Score (1-100)	Profession	Work Experience	Family Size
0	1	Male	19	15000	39	Healthcare	1	4
1	2	Male	21	35000	81	Engineer	3	3
2	3	Female	20	86000	6	Engineer	1	1
3	4	Female	23	59000	77	Lawyer	0	2
4	5	Female	31	38000	40	Entertainment	2	6

Table 5 Customer Data Before Covid

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              2000 non-null   int64
 1   Gender                  2000 non-null   object
 2   Age                     2000 non-null   int64
 3   Annual Income ($)       2000 non-null   int64
 4   Spending Score (1-100)  2000 non-null   int64
 5   Profession              1965 non-null   object
 6   Work Experience         2000 non-null   int64
 7   Family Size             2000 non-null   int64
dtypes: int64(6), object(2)
memory usage: 125.1+ KB
```

Figure 24 Customer Data Before Covid

```
CustomerID              0
Gender                  0
Age                     0
Annual Income ($)       0
Spending Score (1-100)  0
Profession             35
Work Experience         0
Family Size             0
dtype: int64
```

Figure 25 Customer Data Before Covid Missing Data

There are two categorical and five numerical aspects in the data under evaluation. There are no missing values among the numeric features, all but the "profession" property being fully populated. The "profession feature" only has about 35 null values, which is a little amount of missing data. In order to maintain the integrity and correctness of subsequent data analysis, it is

crucial to keep in mind that the presence of null values, especially in a categorical feature, may necessitate imputation or removal.

We will use the mode value to impute the value because this column is categorical.

```
CustomerID                0
Gender                    0
Age                       0
Annual Income ($)         0
Spending Score (1-100)    0
Profession                0
Work Experience           0
Family Size               0
dtype: int64
```

Figure 26 Customer Data Before Covid No Missing Data

Excellent! After successfully removing the null values, we may go on to data visualization and further data processing. It is advisable to undertake data visualization to better understand the underlying patterns and relationships within the dataset before starting with additional data preprocessing techniques.

4.1.2.4. Data Visualization

It is possible to explore, evaluate, and communicate patterns and relationships within data by using data visualization, which is the process of constructing visual representations of data. It is an efficient tool for comprehending complex data and conveying ideas.

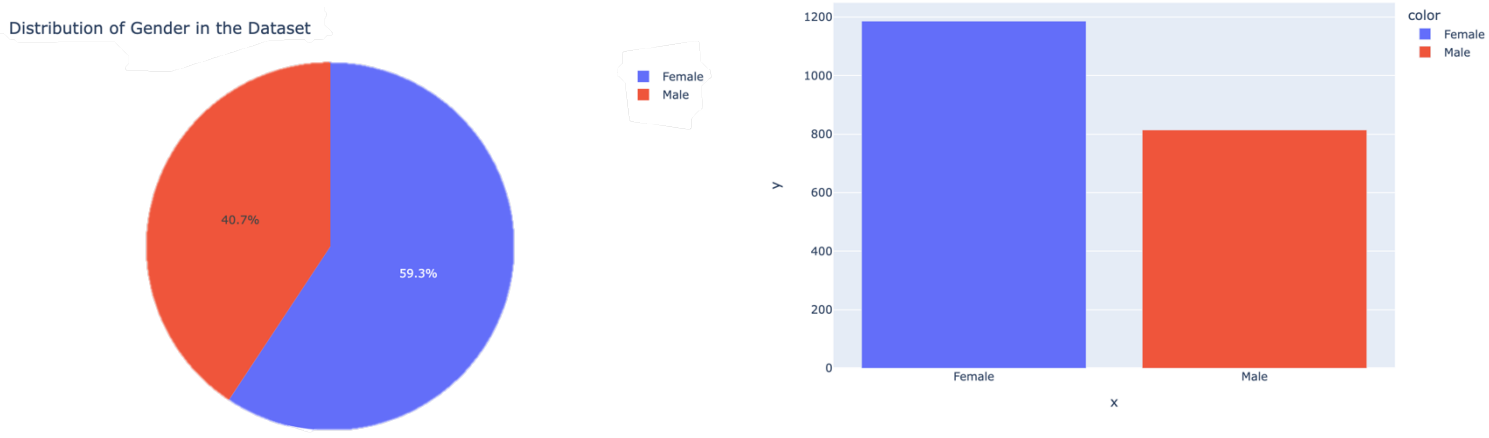
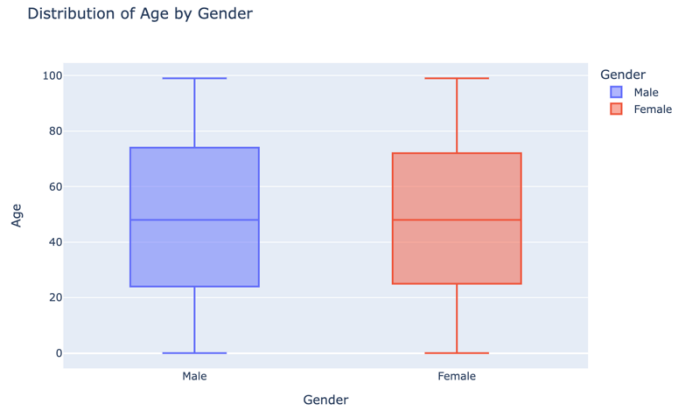


Figure 27 Distribution of Gender in the Dataset

The dataset makes it clear that there is a bias in favor of women because there are much more female candidates than male candidates. Particularly, there are only about 800 male candidates in the sample compared to over 1,200 female candidates.

It is significant to highlight that such a bias may have an effect on how well machine learning models that have been trained on this dataset perform. This is especially relevant when using the dataset to forecast results or make choices that can be influenced by gender.



The age distribution by gender can be clearly seen in the box plot. It is clear from the plot that age distribution does not seem to be greatly impacted by gender. There are no discernible disparities in the age distributions of men and women.

Figure 28 Distribution of Age by Gender

In a similar way to how age is distributed based on gender, so is annual income. The annual income is not significantly impacted by gender. And it's very wonderful to see that.

Figure 29 Distribution of Annual Income by Gender

We may get the conclusion that there is no significant link between the gender distribution and any other variable after performing exploratory data analysis on the "Gender" feature. This shows that when predicting the values of other variables, gender is not a major determinant. This observation is noteworthy from the perspective of machine learning because it suggests that adding gender as a feature to a predictive model may not result in appreciable increases in accuracy. It also suggests that there might be other elements that have a greater impact on the values of these variables. It could be necessary to conduct additional analysis, such as feature engineering or the creation of a correlation matrix, to pinpoint these elements and raise the precision of our models.

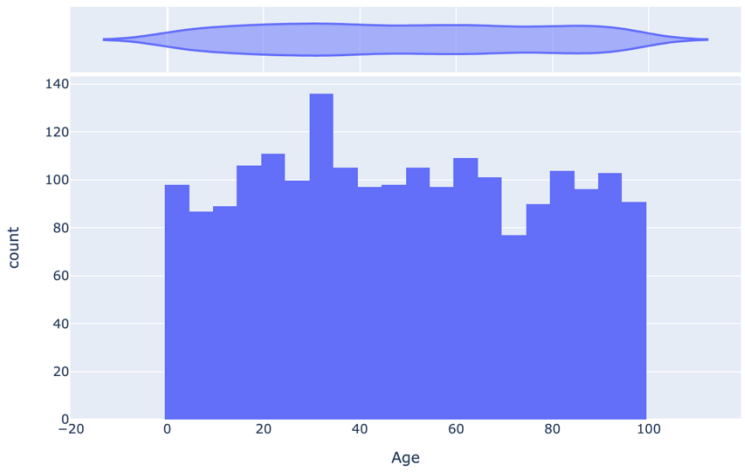


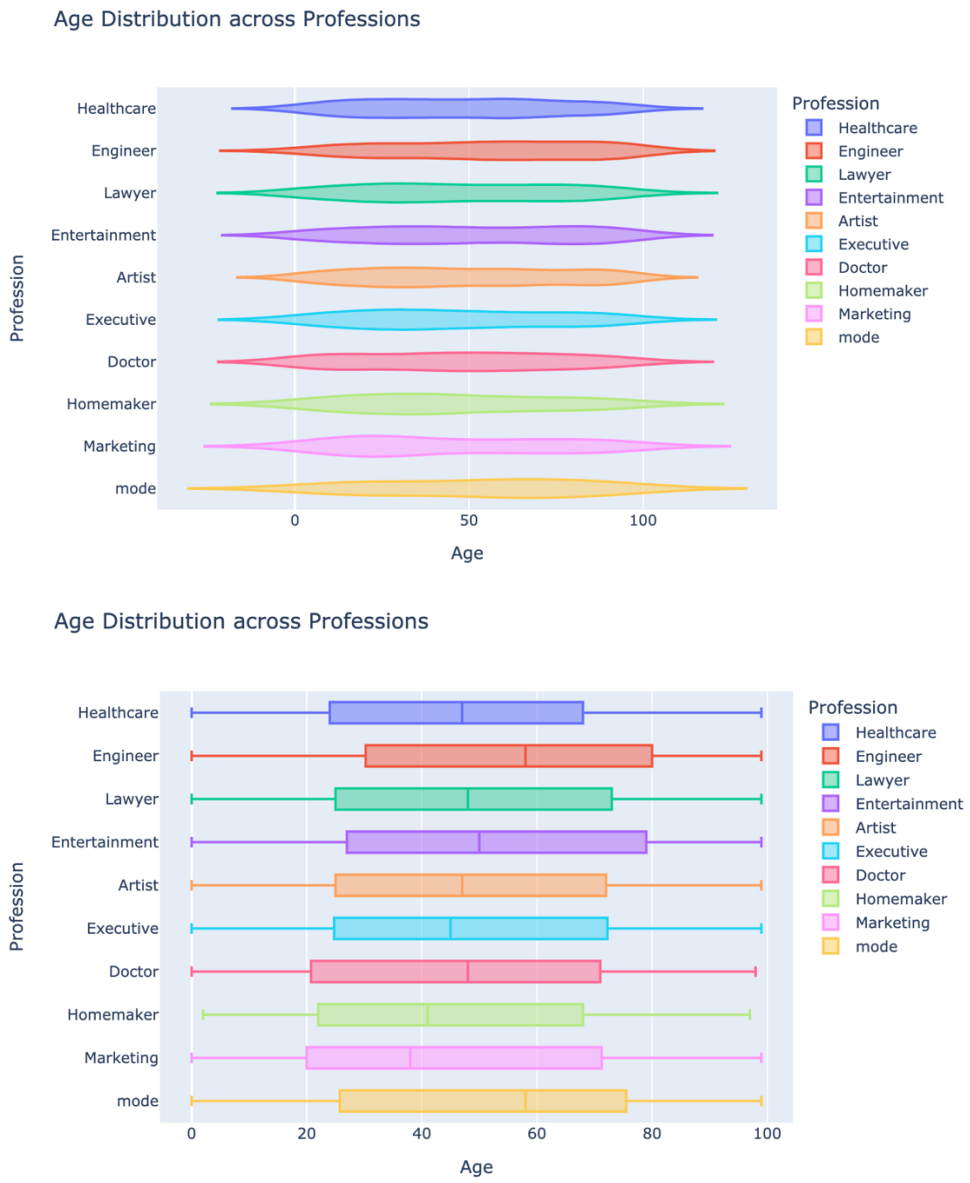
Figure 30 Age Distribution Range 0 - 100

When we examine the dataset's age distribution, which ranges from 0 to 100, we find that it is largely

consistent. It deviates somewhat from a normal distribution due to a modest peak in the 30-34 age group, but it still falls within a suitable range. So, based on statistics, we can say that the data is not prejudiced or slanted in terms of age.

Additionally, from the perspective of deep learning and machine learning, it is crucial to have a balanced dataset that includes data from all age groups in order to guarantee the best model performance. Building precise, reliable models that can generalize effectively to new data can therefore be made easier by the balanced distribution of ages.

Figure 31 Age Distribution Across Professions



We can see that there is no discernible variation in the age distribution among occupations by using a violin plot to analyze the age distribution of various professions. This is unexpected

because one could anticipate that certain professions might favor older workers due to requirements for education and experience. The data, however, argues differently, and it's possible that each occupation is equally represented in the sample regardless of age due to the distribution of clients. This is an intriguing study that raises the possibility that a customer's age may not be a reliable indicator of their job.

In contrast to the violin plot, the box plot suggests that age and career are only tangentially associated. There are differences in the age distribution between professions and the age distribution for various professions is not uniform. These insights can be essential in forecasting a person's career depending on their age from the perspective of machine learning and deep learning.

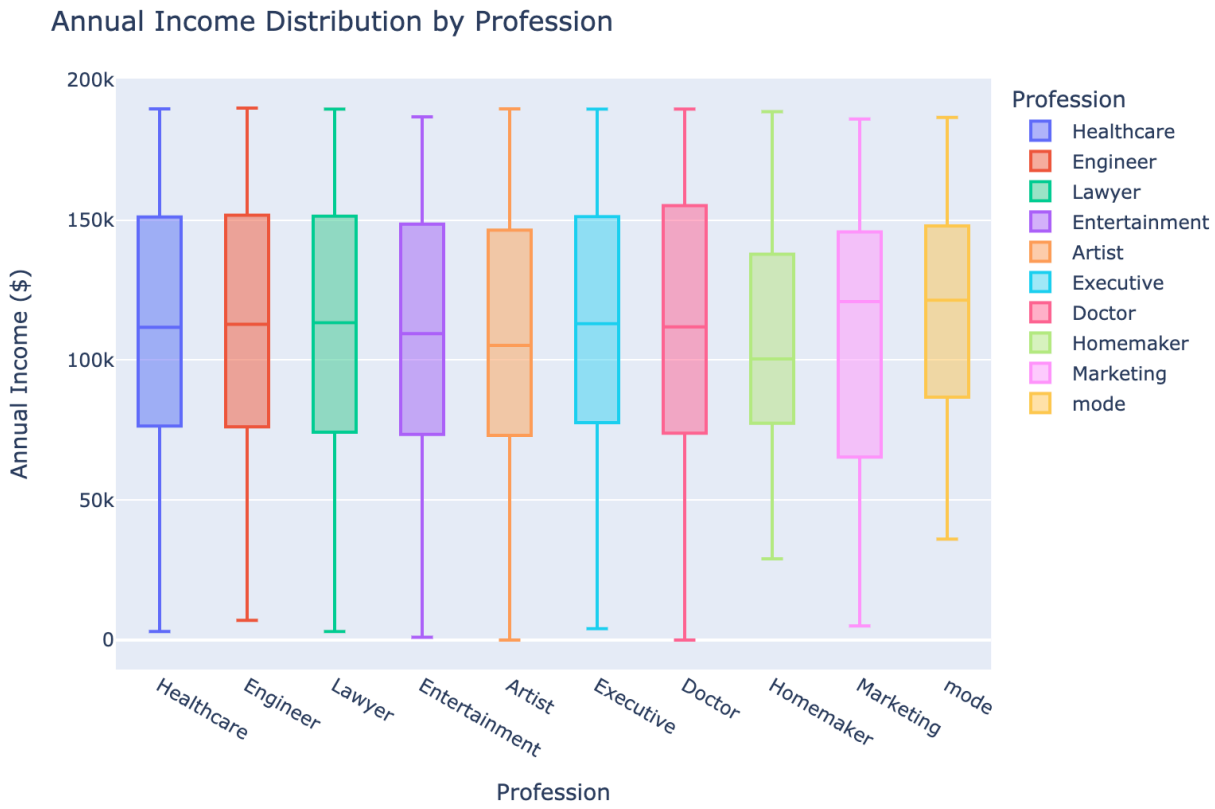
It is also important to remember that different occupations have different median ages of the population. In the case of engineers, the median is moved to the right, indicating that most engineers are in their 60s. In contrast, the mode for homemakers and marketers is pushed to the left, indicating that most people in these fields are in their 40s. This knowledge can be helpful in creating marketing plans for goods or services that are intended for a certain profession or age group.

Job experience does not appear to rise proportionally with age, according to the density contour plot between age and job experience, which displays an interesting trend. This goes against what we would typically anticipate from the actual world, where having more work experience typically increases with age. The dataset can be too perfect and not precisely reflect the distribution in the real world. Alternately, there can be other elements at work that have an impact on work experience independent of age, such as changes in career pathways or educational attainment.

Figure 32 Annual Income Distribution range 0 to 200K

When the annual revenue histogram is examined, a clear pattern can be seen. Notably, the number of people earning between zero and fifty thousand dollars is quite low—just 15. The data indicates that a significant change happens after the 50,000-point threshold, when the count soars to almost 70. Up until it reaches 100,000, this trend is increasing linearly; after that point, some oscillations are seen. These issues continue until an income of around $190,000.

Figure 33 Annual Income Distribution by Profession



Further examination of the association between yearly income and occupation reveals that the distribution of income for some professions, like engineering, law, entertainment, the arts, executive, and medicine, appears to be largely consistent. The distribution of income for those in the homemaking profession has undergone some significant shifts, with the lower values skewing significantly higher.

Additionally, the typical salary for those in the aforementioned professions stays almost the same at $100,000, however it slightly declines for those in the homemaking industry. On the other side, the marketing industry goes through a significant transition, with the total income distribution remaining stable but the median income rising.

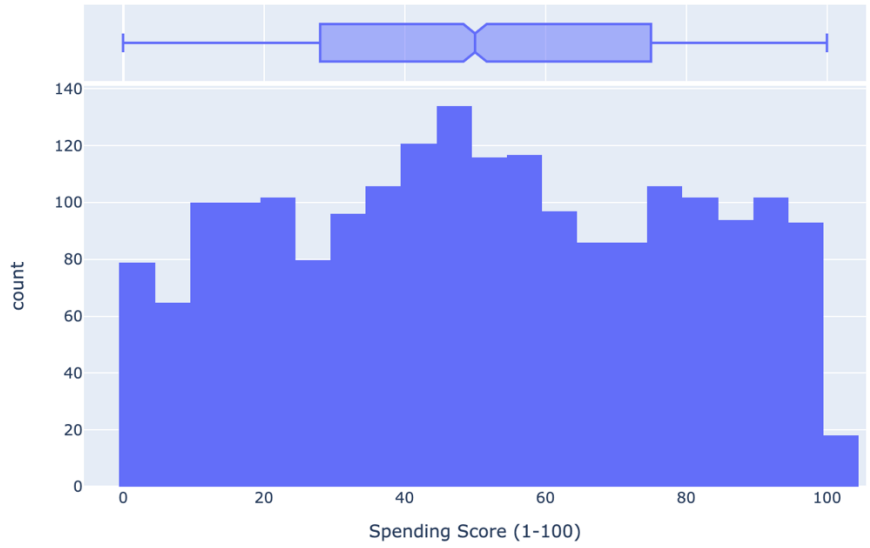


Figure 34 Spending Score 1-100

It is clear from the spending score's histogram that the amount of times counts are essentially constant throughout. However, there is a noticeable surge in the frequency count in the spending score between 45 and 49. This summit has a sloping terrain. The frequency count drastically decreases to just 18 as we get closer to the histogram's finish. The initial frequency count was around 80, peaked at 134, then fell to 18—a considerable change from the original values. Despite these peaks and valleys, the distribution exhibits a steady rising and downward tendency.

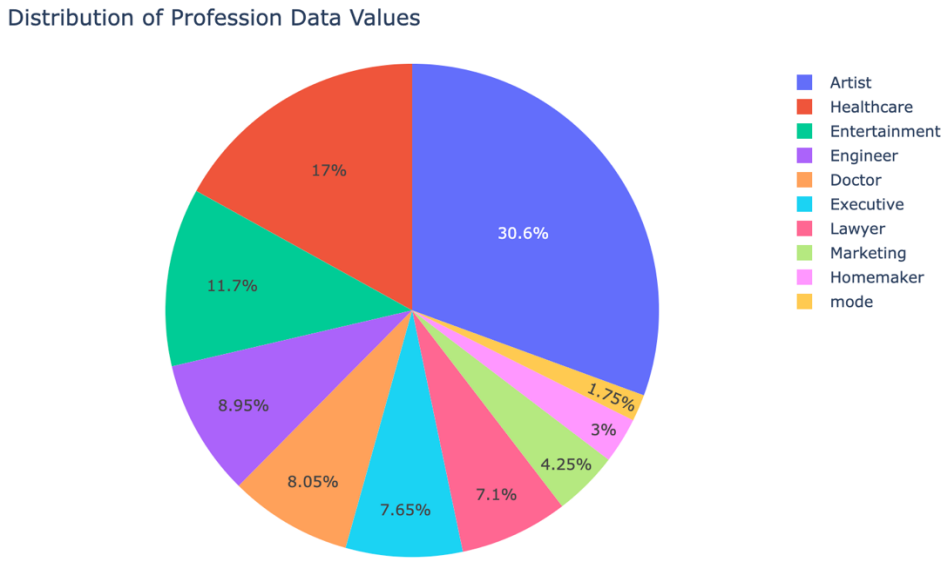


Figure 35 Distribution of Profession Data Values

In the dataset, about 30% of all customers are categorized as artists, which provides a general overview of the distribution of clients across occupations. Around 17% of all clients are healthcare professionals, making them the second-largest group. The entertainment sector

follows closely after, accounting for about 11.7% of all consumers. However, only 3% of all consumers are homemakers, making them the smallest segment of the client base.

Given that homemaking is no longer the principal occupation of women in contemporary culture, it was not surprising to see that the homemaker profession had the lowest count when the distribution of occupations was examined. The artist profession, which accounts for the largest portion of clients at about 30%, experienced an unexpectedly sharp surge. The prevalence of artists in the dataset is notable and shows the growing importance of art and creative expression in contemporary culture, despite the common misconception that engineering careers are more common than others.

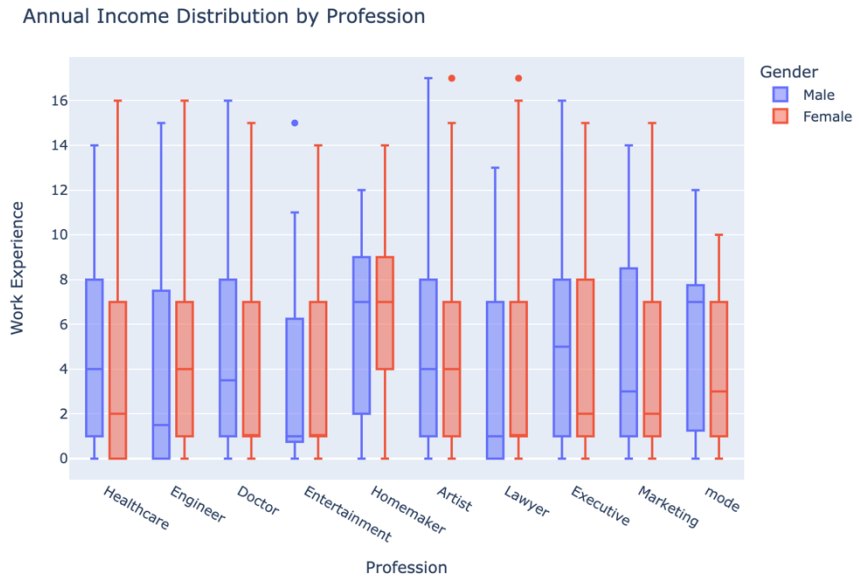


Figure 36 Annual Income Distribution by Profession

Several significant findings can be seen when evaluating the box plot for job experience across different occupations.

- Particularly, compared to other sectors, the healthcare, executive, medical, and marketing departments have a wider total range of job experience. The median job experience in the legal and entertainment industries is only one year, which is a minimal amount of experience. The median values for both sectors are unsatisfactory even if the distribution is respectable, with a low of one year and a high of about seven years.
- As expected considering the nature of their jobs, the median work experience for doctors, executives, and those in the healthcare industry ranges from one year to almost eight years. Although doctors only have a two-year median experience, which is also poor compared to the healthcare and executive divisions.
- Additionally, certain notable outliers can be seen in the data, such as the professional with an astounding 17 years of job experience in the lawyer and arts industries. The housewife profession, which has a substantially higher range of work experience, starting from roughly three years and going up to nine years, is the most important factor that needs to be highlighted. This is understandable because once individuals enter this field,

they are more inclined to quit it permanently. Additionally, the median value is high, with a maximum median value of almost seven years for the entire distribution. This suggests that those who work in this field are likely to stay employed for a minimum of seven years.

- There are observable variances in the data distribution when gender differences are considered. When comparing boys and females, the median value for various professions differs dramatically. For instance, the median value in healthcare is low for women and high for men. This may be because nurses are seen as female and doctors as men, respectively, in their respective professions.
- It's interesting to note that in engineering, women have a substantially higher median experience value than men do. Males only have about one or two years of job experience on average, whereas females have a median of four years. Similar to this, even though the general range is nearly the same, female doctors have one year less job experience than male doctors do.
- Last but not least, for those who work as homemakers, the top values are the same for men and women, but the bottom values are different. Men seem to start out as stay-at-home parents sooner than women do, with an average of two years of job experience compared to four years for women. But for all genders, the median number is the same.

It appears that no noteworthy data distributions or novel insights can be discovered after analysis of the data collection. I've come up with a number of plots, but none of them have turned up any ground-breaking facts.

4.1.2.5. Data Preprocessing

Before beginning any analytical work in the field of data analysis, data pretreatment is a crucial step that must be completed. Prior to beginning the analysis, it is essential to have a thorough grasp of how the data is distributed and the insights it contains. With this knowledge, patterns and connections in the data may be found, which are essential for drawing accurate conclusions. The next stage after gaining this insight is data preparation, which entails cleaning and altering the data to make it ready for analysis.

Addressing the dataset's category columns is one of the first jobs in data preparation, which is a critical step in machine learning. Converting categorical data into numerical values is crucial since the majority of machine learning models operate on numerical data. Using the Scikit-Learn library's Label Encoder is one typical method for accomplishing this. Categorical data may be represented as a series of integers thanks to the Label Encoder, which gives each category in a column a distinct numerical label. This translation method facilitates the processing of the data by machine learning algorithms and enhances the model accuracy.

	CustomerID	Gender	Age	Annual Income ($)	Spending Score (1-100)	Profession	Work Experience	Family Size
0	1	1	19	15000	39	5	1	4
1	2	1	21	35000	81	2	3	3
2	3	0	20	86000	6	2	1	1
3	4	0	23	59000	77	7	0	2
4	5	0	31	38000	40	3	2	6

Table 6 Customer Data Before Covid Prepressing

It is essential to execute feature scaling on the dataset in order to make sure the machine learning model is correct and effective. In order to improve the performance of the model, it is necessary to bring all the feature columns to a common range. Currently, each feature column's range may differ from the others. For scaling numerical data, the standard scaler is a popular technique that is advised.

```
array([[ 1.20706357, -1.05408932, -2.09350095, -0.42833854,  0.77835593,
        -0.79120713,  0.11749744],
       [ 1.20706357, -0.98372287, -1.65613312,  1.07554599, -0.35434734,
        -0.28116224, -0.39005088],
       [-0.82845678, -1.0189061 , -0.54084515, -1.6099621 , -0.35434734,
        -0.79120713, -1.40514752],
       [-0.82845678, -0.91335643, -1.13129172,  0.93231889,  1.53349144,
        -1.04622958, -0.8975992 ],
       [-0.82845678, -0.63189066, -1.59052794, -0.39253176,  0.02322042,
        -0.53618469,  1.13259408]])
```

Figure 37 Customer Data Before Covid Prepressing

Now we can move towards modeling, Since The pre-processing steps have completed.

4.1.2.6. Data Correlation

There is little question that there is a sizable departure from the predicted linearity in the data scatter plot. The plot reveals that the data points have a significant degree of variability, suggesting that the data are spread widely and lack a discernible pattern. This can be as a result of the existence of outliers, measurement mistakes, or additional variables impacting the data gathering process. A strong prediction model that effectively represents the underlying patterns in the data may be difficult to construct due to the variability in the data. To locate and resolve any flaws with the dataset, more research and data cleaning may be needed.
Without a doubt, the dataset's lack of substantial connections is a departure from what is often seen in real-world situations. In reality, characteristics are more frequently connected to one another, and a correlation matrix frequently indicates substantial positive or negative correlations between features. The lack of significant correlations in the dataset shows that there may be underlying variables that are yet unaccounted for and might potentially affect the behavior of the characteristics. It could be essential to perform more exploratory analysis and feature engineering to unearth any concealed patterns or links in the data in order to improve our analysis and obtain a better knowledge of the data. By doing so, we might be able to better understand how the characteristics behave and possibly enhance our prediction models.

4.1.2.7. Principal Component Analysis

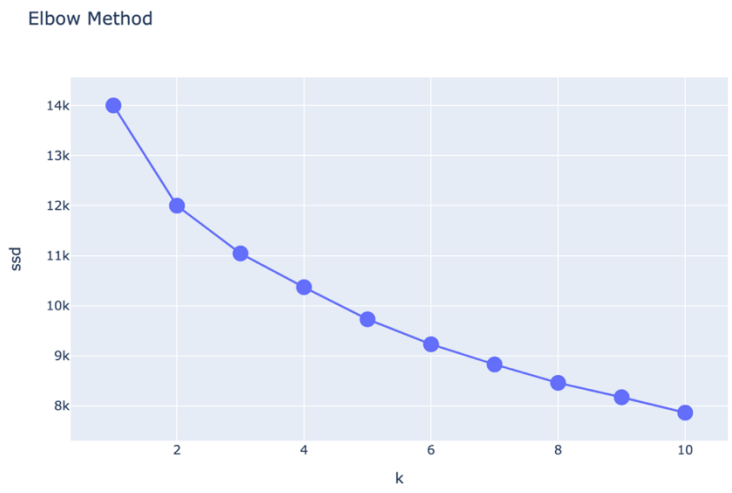
A popular statistical method for decomposing complicated datasets into simpler parts is principal component analysis (PCA). This is achieved by lowering the number of variables while keeping the data's crucial information intact. In data analysis and machine learning, PCA is a useful technique since it aids in feature extraction, dimensionality reduction, and data visualization. PCA can assist to improve the data's interpretability, model performance, and computing complexity by identifying and summarizing the dataset's key properties. In the field of data science, PCA is a potent tool that is used to unearth insightful information and guide decision-making because of its capacity to recognize underlying patterns and structures in data.
Using a new set of variables called principle components, which capture the majority of the volatility in the data, PCA is a statistical method for finding patterns in high-dimensional data. These components, which are linear combinations of the initial variables, are created so that the first component has the highest variance possible and that each succeeding component has the highest variance possible while still having to be orthogonal to the previous components. The data can be projected onto the principal components to create a lower-dimensional representation that can be used for visualization and to reduce the dimensionality before using other machine learning algorithms. This will ultimately improve interpretability, lower overfitting, and boost computational efficiency.

4.1.2.8. K-Means Clustering

Unsupervised machine learning technique K-means clustering combines comparable data points in a dataset by randomly choosing k data points as the initial centroids for the clusters. It assigns the nearest centroid to each data point in the dataset, builds k clusters, calculates the mean of each cluster, and then shifts the centroid to this new position. Until the centroids stop moving or a maximum number of iterations is reached, the algorithm restarts the process of reassigning data points and computing new cluster means. The Within-Cluster Sum of Squares (WCSS), often known as the distance between each data point and its assigned centroid, is the goal of the K-means method. The final centroids and the cluster assignments for each data point are included in the algorithm's output.

Elbow Method

Figure 38 Elbow Method



By displaying the explained variation vs the cluster count and choosing the "elbow"—the point where the curve starts to flatten—the Elbow approach is used to determine the optimal number of clusters for a K-means algorithm. The Within-Cluster Sum of Squares (WCSS), which declines with more clusters, is a metric for explaining variance. The ideal number of clusters may be

discovered when the pace of WCSS decline starts to slow down, which occurs at this time. For selecting the ideal number of clusters in K-means clustering, the Elbow technique offers a helpful heuristic.

Silhouette Method

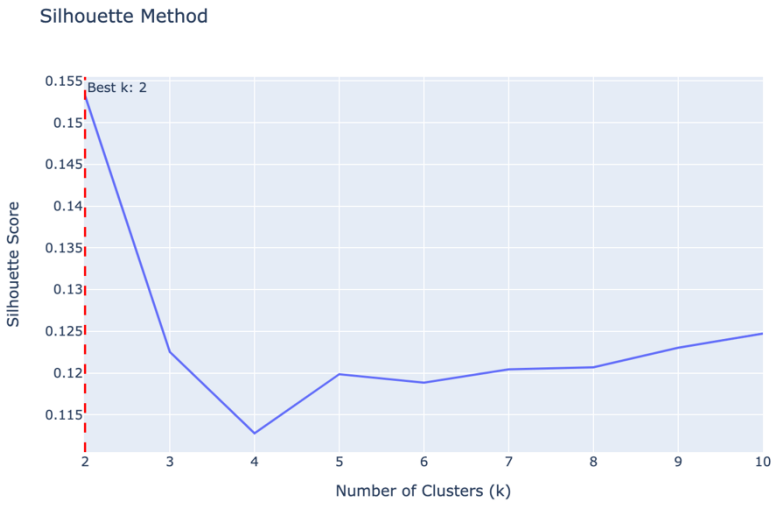


Figure 39 Silhouette Method

By calculating how similar a data point is to its own cluster in comparison to other clusters, the Silhouette Method assesses the performance of clustering. To get the overall Silhouette score for a clustering solution, the approach derives a score for each data point with a range of -1 to 1, then averages them. A score of 1 denotes a perfectly matched data point, a score of 0 denotes a boundary point, and a score of -1 denotes a data point that was incorrectly allocated to a cluster. The greatest Silhouette score is used to determine the ideal number of clusters. To choose the ideal number of clusters for a particular dataset, the Elbow Method and the Silhouette Method can be combined. The average distance between a single data point (i) and every other point in its cluster and the closest neighboring cluster are computed, and these numbers are then used to create a formula to determine the silhouette score.

Calinski-Harabasz Index

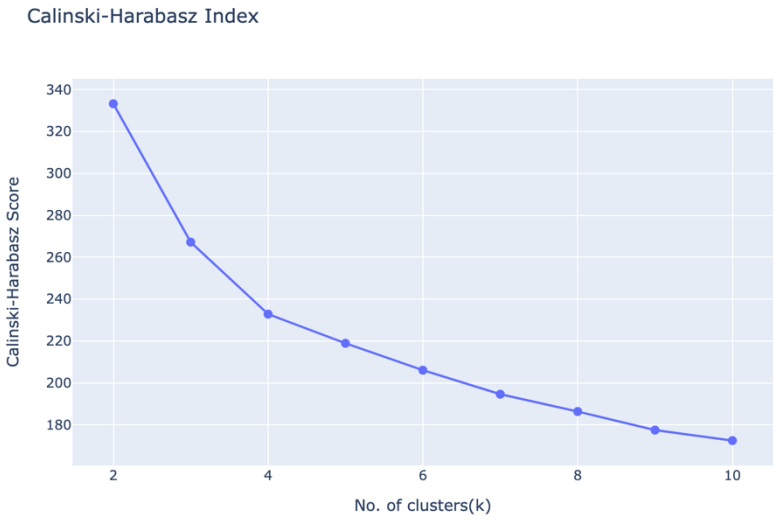


Figure 40 Calinski-Harabasz Index

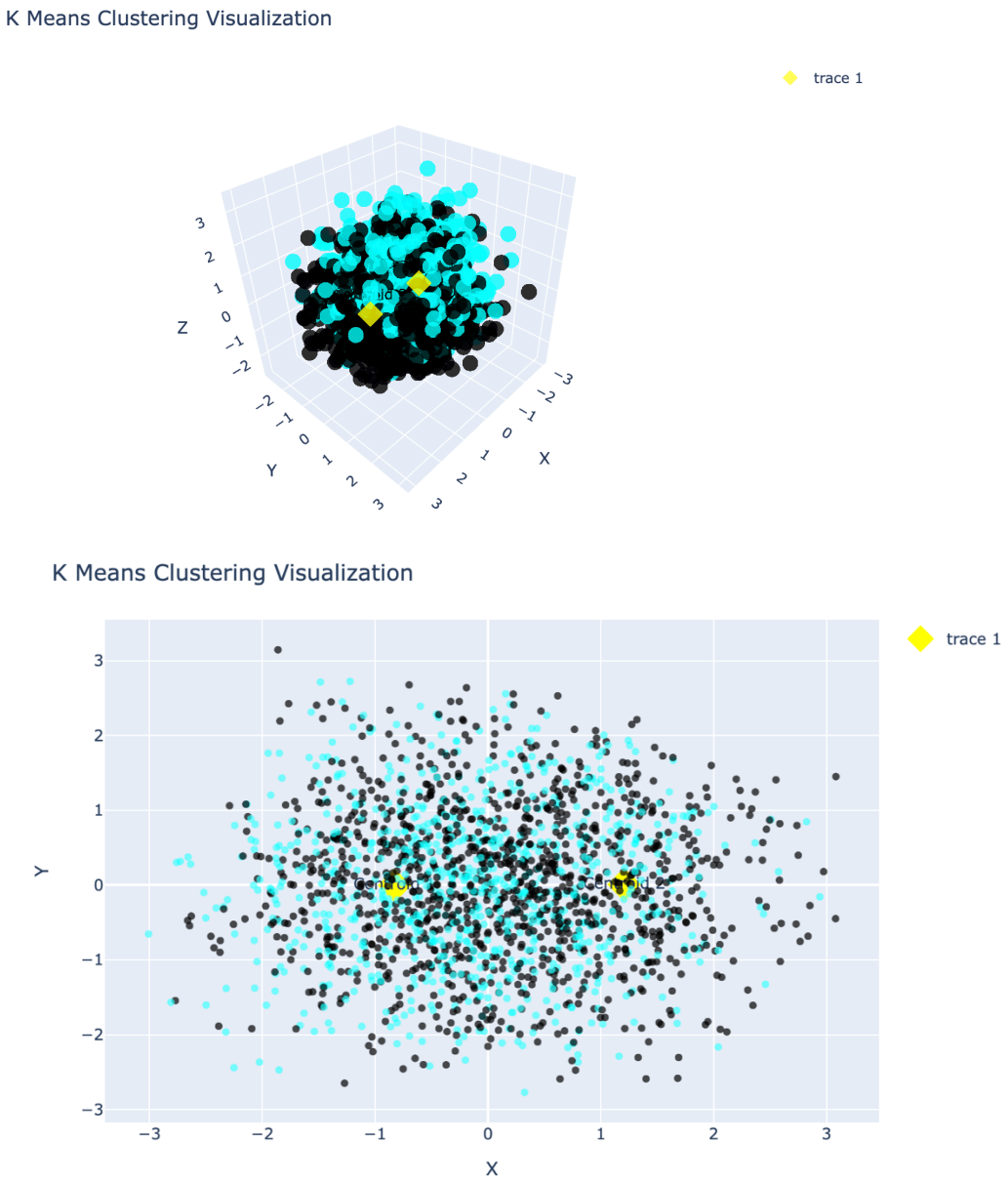
The ratio between the within-cluster dispersion and the between-cluster dispersion is measured by the Calinski-Harabasz Index, a clustering assessment tool. The within-cluster dispersion estimates the average distance between each data point and the cluster centroid, whereas the inter-cluster dispersion determines the average distance between each cluster's centroids.

The formula for the Calinski-Harabasz Index is:

$$CH(k)=B(k)/(k-1)/W(k)/(n-k)$$

For a given set of clusters, k, the Calinski-Harabasz Index, also known as the Variance Ratio Criterion, aims to maximize the ratio of between-cluster to within-cluster dispersion, where B(k) is the between-cluster dispersion, W(k) is the within-cluster dispersion, and n is the total number of data points. A better clustering solution is indicated by a higher value of the Calinski-Harabasz Index. It may be used with other clustering assessment metrics like the silhouette coefficient and the elbow method to figure out the ideal number of clusters given a dataset.

Figure 41 K-Means Clustering Visualization



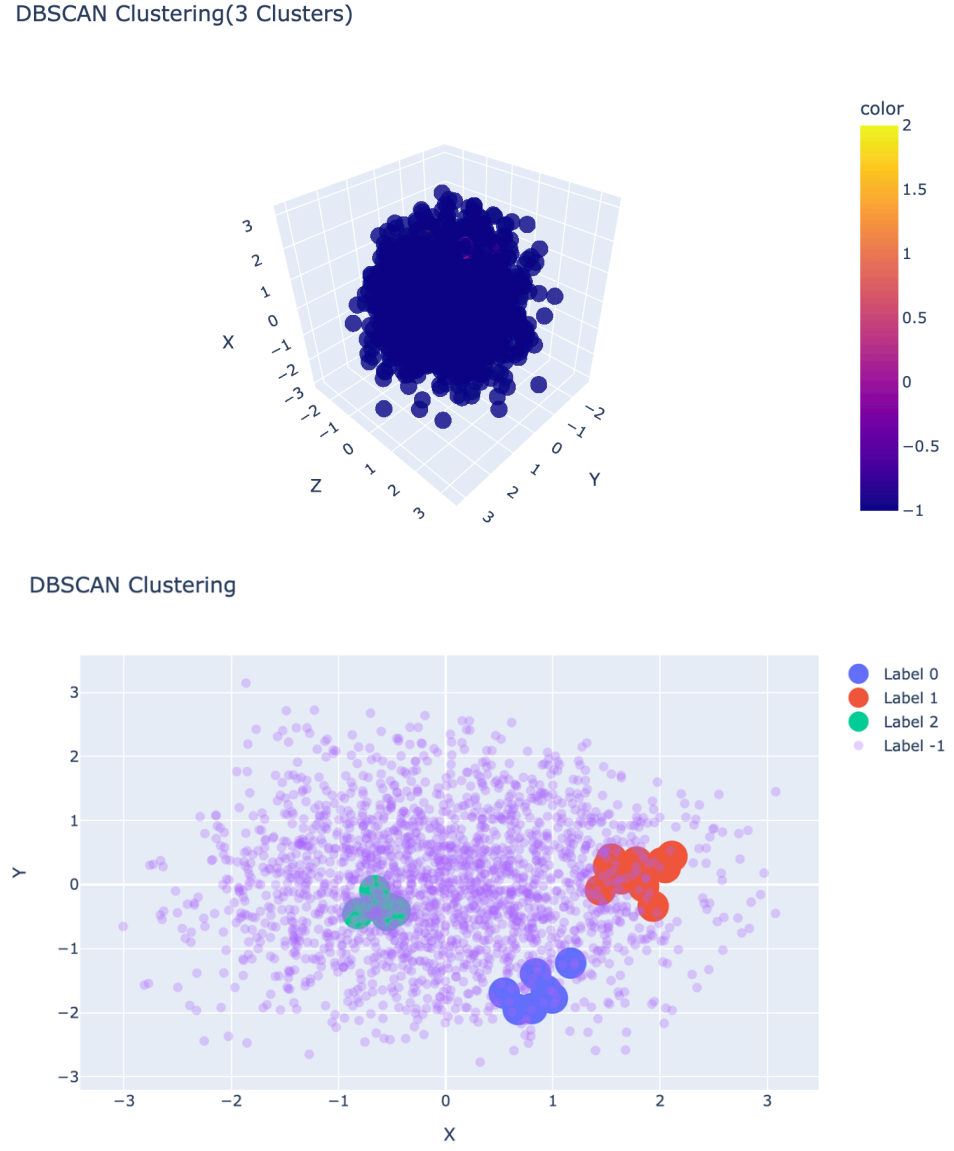
As just two centroids were detected using the K-means method and no clearly defined clusters with distinct borders were produced, the approach performed poorly in terms of clustering. It is clear that the algorithm failed to recognize the underlying structure of the data when viewing the centroids in both 3D and 2D space, where the clusters seem jumbled together. In the higher-

dimensional feature space, there are at least two clusters, albeit the data may not be sufficiently varied to truly reflect reality.

4.1.2.9. DBSCAN

Using a density criteria, the clustering algorithm DBSCAN finds clusters of any form in noisy datasets. It locates "core points" and groups them using the minimal number of neighbors they have within a certain radius. The same cluster is allocated to all of the points that are inside the radius of a core point, while noise is applied to all of the other points. The algorithm includes two crucial inputs: epsilon, which specifies the neighborhood's radius, and min_samples, which establishes the bare minimum of points necessary to constitute a dense region. DBSCAN is beneficial since it is noise- and outlier-resistant and does not require prior specification of the number of clusters.

Figure 42 DBSCAN Clustering



The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method detected three clusters when the parameters were set to DBSCAN(eps=0.7, min_samples=5), although there were few points in each cluster. The Feature Description approach supported the idea that the points were not tightly connected in the higher-dimensional space. The clusters found are not perfect, and the majority of the data is noise, thus this is a frequent issue with ideal datasets. Prior to constructing models, it is essential to deal with the issue of high levels of noise in the data since this might impair model performance and accuracy.

4.1.3. Survey Dataset

4.1.3.1. About Dataset

The dataset, which we created with the help of the public and which we are making available in two languages—English and Turkish—can help all businesses with online storefronts tailor their offerings to the needs of their customers. Additionally, we emphasize the online purchasing before to and following COVID-19, including the prior to and subsequent online buying budget. Covid-19, What elements are vital for the product before and after, how frequently do you use internet shopping each year, and gender.

A dataset with 213 entries and 29 columns can offer important insights into customers' buying choices. Let's investigate some:

Do you shop online:
Many people engage in online shopping, but some remain skeptical due to the inability to physically inspect the product's material before purchase. The need to see and feel the product firsthand is a common concern for those who hesitate to fully trust online shopping.

Have you started to buy products online, that you were buying from store before covid-19:
Customers are more likely to make repeat purchases if they have a positive initial experience with a product or service. When the first impression is favorable, it increases the chances of customer loyalty and encourages them to buy from the same provider again.

Have you started to pay more attention to prices and made research about it more after Covid:
Each customer prioritizes their desired product, and the extent of their research depends on the perceived importance of the item. The level of research conducted by customers varies based on the significance they attach to the product they intend to acquire.

Do you have the chance to examine the products better thanks to online shopping:
In the absence of physically examining products as in brick-and-mortar stores, many companies offer return policies that allow customers to return products if they are dissatisfied with them. This provision enables customers to have a safety net in case they are not pleased with the product they receive.

Do you expect most businesses to have an online shopping system after Covid:
Due to the impact of the Covid-19 pandemic, companies have increasingly adopted online stores as a strategy to boost their sales

Do you think online shopping is more affordable than In-Store shopping:

Customers can save money on transportation expenses and avoid time wastage by opting for online shopping

What are some factors that are important to you when shopping online:
Different customers have varying priorities when it comes to online shopping. While some prioritize shipping speed, others prioritize price or customer service. Each customer has their own preferences

How likely are you to recommend online shopping to a friend or family member:
When individuals come across a product that is both of high quality and offered at an affordable price, they tend to share this information with their friends or family members

What is the primary reason for you to shop online instead of in-person:
Saving time and enjoying convenience are key advantages of online shopping

In which category do you use online shopping the most:
A significant portion of people prefer online shopping for categories such as Clothing & Shoes as well as food items**.**

Which payment method do you usually use when shopping online:
Some people, often tech-savvy and seasoned online buyers, are more likely to trust and utilize other payment methods, such as credit/debit cards, digital wallets, or online payment platforms. Others could be more cautious and choose conventional means, such as bank transfers or cash on delivery, for increased security.

Which website do you use the most, while shopping online:
Customers have a preference for online platforms that offer fast and efficient service, along with excellent customer support, exemplified by popular sites like Amazon and Trendyol.

How often do you use online shopping: The frequency of online shopping varies among individuals, with some using it daily, others using it weekly or monthly, and some choosing not to use it at all. Ultimately, the decision of how often to engage in online shopping is a personal one.

What was your monthly online shopping budget before Covid-19:
Prior to the Covid-19 epidemic, personal financial situations, tastes, and priorities affected people's buying expenditures, which differed from person to person. Some people may have had larger spending limits when shopping, allowing them to spend more freely on a variety of products and services. Others may have adhered to more stringent spending plans, prioritizing necessities while reducing discretionary spending.

What was your monthly online shopping budget after Covid-19:
Due to changes in customer behavior and a greater dependence on e-commerce since the Covid-19 epidemic, internet shopping budgets have undergone considerable adjustments. Due to the fact that internet shopping has become a popular method of obtaining goods and services, many people have changed their spending plans to include more money for online purchases.

During Covid-19, my online shopping rate increased:

Due to the Covid-19 pandemic, many individuals faced restrictions and lockdown measures that limited their ability to go outdoors

After Covid-19, I think people started to do more online shopping:
After a prolonged period of relying on online shopping during the Covid-19 pandemic, it can be challenging for individuals to transition back to traditional shopping methods

On average how long do you wait for your online purchase to arrive:
The duration of product arrival can be influenced by several factors, including the type of website used (local or global) and the specific category of products such as food, clothing, or furniture. These variables play a role in determining the shipping time for the purchased items. The choice of website, as well as the nature of the product, can impact the length of time it takes for the products to be delivered to the customer**.**

I trust the payment methods when shop online:
Depending on their interests and degree of comfort with online transactions, different sorts of consumers trust various online payment methods when they purchase. Some people, often tech-savvy and seasoned online buyers, are more likely to trust and utilize other payment methods, such as credit/debit cards, digital wallets, or online payment platforms. Others could be more cautious and choose conventional means, such as bank transfers or cash on delivery, for increased security.

When I'm shopping online, websites suggest items that I am looking for:
All websites analyze and track customer behavior within their platform.

As of today, I am shopping online less than I did during Covid-19:
Currently, my online shopping frequency has decreased compared to the period during the Covid-19

When I register to an online shopping site for the first time, seeing better product suggestions would increase my satisfaction:
My overall pleasure would increase if, after enrolling on an online purchasing site, I received better product suggestions.

where do you live:
The majority of individuals answering here are Turkish, as they represent a significant portion of the user base seeking information and assistance.

What is your educational background:
The majority of individuals answering here are undergraduate

4.1.3.2. Set Up

Importing all required modules, setting hyperparameters, and imposing constraints are all crucial steps in getting ready for data loading, preprocessing, and model construction. To guarantee accuracy and consistency in the analysis, these settings will be used consistently across the notebook.

4.1.3.3. Data Loading & Processing

This statistic reveals the number of consumers who purchase online, divided into two categories: those who do and those who don't. According to the statistics, 199 of the consumers who were polled said they did their shopping online, while 14 said they did not.

In Here, We will remove the data for consumers whom do not purchase online, afterward we use the data of only for consumers whom purchase online and analyze them.

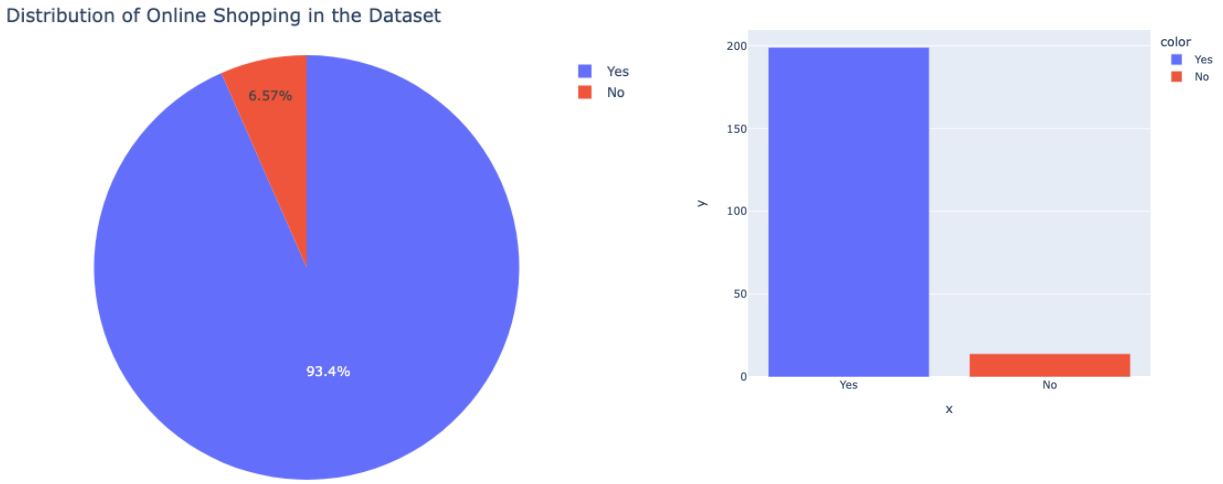


Figure 43 Distribution of Online Shopping in the Dataset After Covid

Count of customers who shop online

This data demonstrates the popularity of internet shopping among the studied consumers, with a sizeable majority choosing this practical way to make purchases. It indicates a clear tendency toward e-commerce as well as the growing acceptance and use of online purchasing platforms.

4.1.3.4. Data Visualization

Count of customers buying online

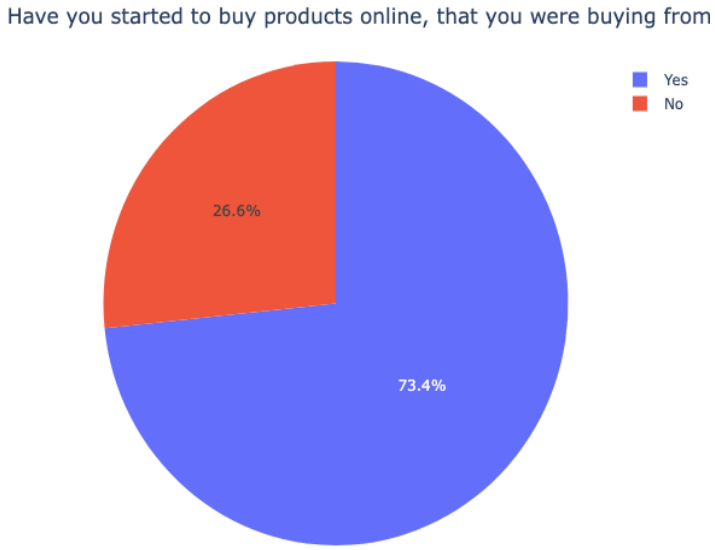


Figure 44 Count of customers buying online

The information presented illustrates the number of clients depending on their online purchasing habits. 53 of the clients who participated in the poll had never purchased goods or services online, compared to 146 who have done so.

This information suggests that a sizeable percentage of the questioned consumers like internet shopping, whereas a smaller percentage favors other options or abstains from making online purchases entirely. Given the popularity of online shopping, e-commerce platforms and the convenience they provide are becoming more and more important. It also suggests that a sizable portion of consumers would still favor conventional brick-and-mortar businesses or other offline channels for their buying requirements.

Count of customers' attention on product

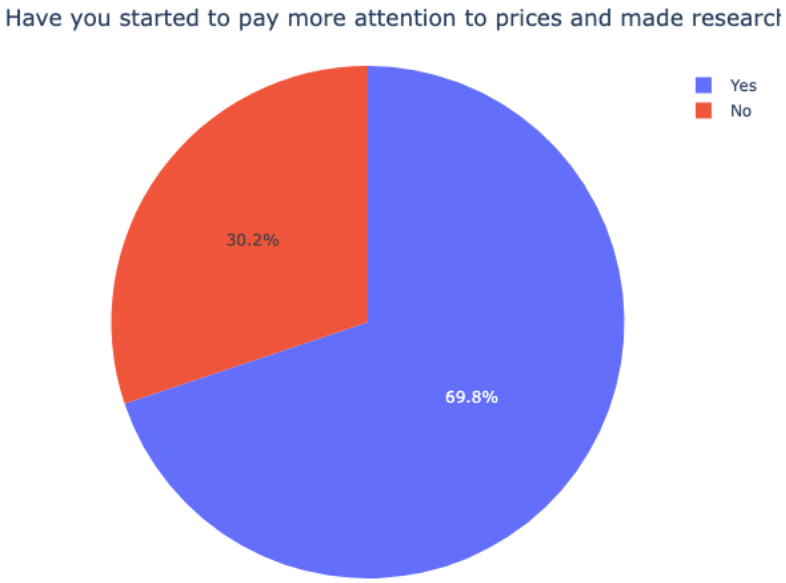


Figure 45 Count of customers' attention on product

According to how much attention people are paying to a product, the data shows the number of customers. 139 respondents to the study of consumers said they pay attention to the product, whereas 60 consumers said they don't.

According to the research, the majority of the questioned consumers exhibit active interest in the product, actively considering and weighing its characteristics, advantages, or allure. During the purchase process, these clients are probably more attentive and involved, making deft choices based on their observations and evaluations.

The existence of clients who are not interested in the goods, on the other hand, may signify a lack of interest, a lack of engagement, or different priorities when it comes to their purchase behavior. To effectively cater to their target audience and maximize their marketing tactics, organizations must recognize and fulfill the wants and preferences of both client segments.

4.1.3.5. Data Visualization

Count of customers' examination of products

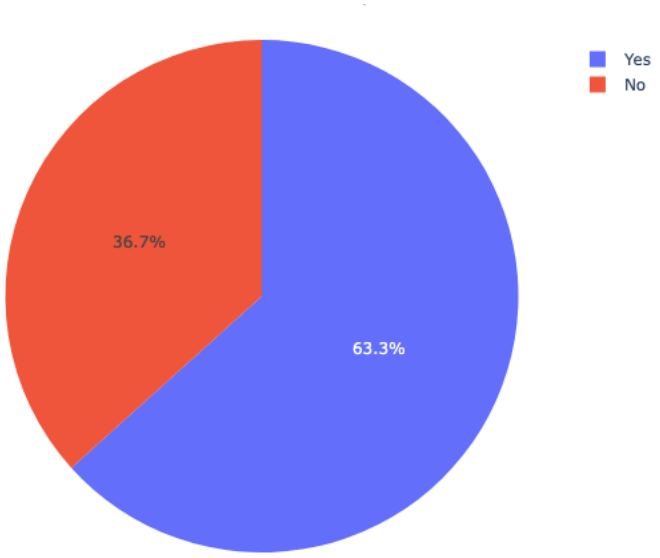


Figure 46 Count of customers' examination of products

According to the statistics given, clients are counted depending on how often they look at items. Among the consumers who responded to the study, 126 people said they do inspect things, while 73 people said they don't.

This information reveals that a sizable proportion of the questioned consumers actively engage in the process of product evaluation. Before making a purchase, these clients are likely to check, analyze, or assess the items' physical features, quality, functioning, or other factors. Their propensity for product inspection suggests a careful and comprehensive approach to guarantee that they make decisions that are informed by their preferences and requirements.

On the other side, the existence of buyers who don't examine the products might indicate more impulsive buying behavior or a dependence on other elements like brand reputation, referrals, or convenience.

Businesses may better adjust their marketing tactics, enhance product displays, and give pertinent information to meet the demands of various client groups by being aware of the diverse preferences and behaviors associated with product evaluation among consumers.

Figure 47 Expectation of Business having online Shopping

Expectation of Business having online Shopping

The information given is a count of answers based on a certain criterion. 179 respondents, or 89.45% of the total, gave favorable replies, while 21 respondents, or 10.55%, gave negative responses.

The larger count and proportion of yes replies in this data suggest that the majority of respondents favor or agree with the stated criterion. It implies a generalized tendency or predisposition toward the subject of the evaluation.

A lesser percentage of negative replies, on the other hand, indicates that some respondents have different opinions or don't fit the criteria. This demonstrates the diversity or varying viewpoints among the respondents to the survey.

Insights about the general attitude or consensus surrounding the criterion in issue may be gained by analyzing and comprehending the distribution of replies. This information might be helpful for making decisions, gauging public sentiment, or figuring out what needs more attention or clarity.

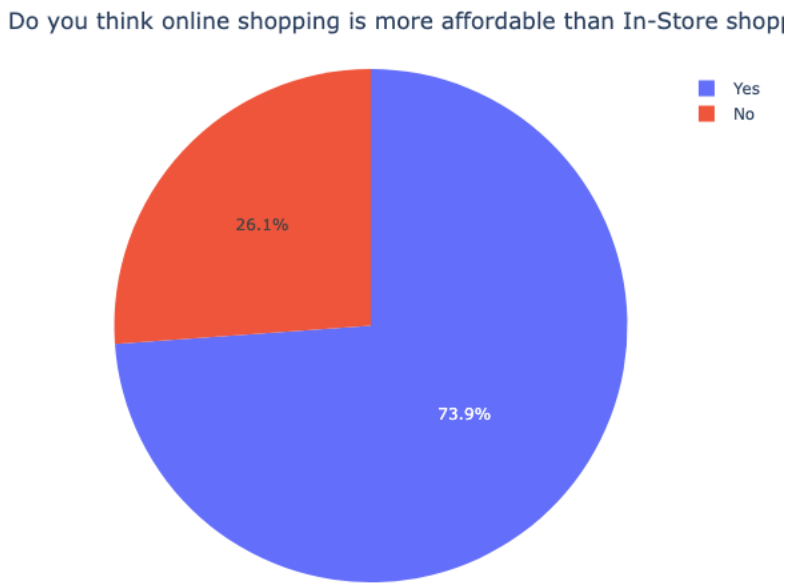


Figure 48 Affordability of Shopping online

Affordability of Shopping online

The data made available displays the total number of replies that met a certain criterion. 52 respondents, or 26.13%, of the total replies received, answered negatively, while 147 respondents, or 73.87% of the total, reacted positively.

According to the findings, a sizable majority of respondents agreed with or supported the stated criterion. The large number and proportion of yes replies point to a significant preference for the question or statement under consideration.

A lesser percentage of negative replies, on the other hand, indicates that some respondents have different opinions or do not fit the criteria. This demonstrates the diversity or varying viewpoints among the respondents to the survey.

The general attitude or consensus surrounding the criterion in question may be learned through analyzing and comprehending the distribution of replies. Making decisions, gauging public opinion, or finding areas that might need more attention or clarity can all benefit from this information.

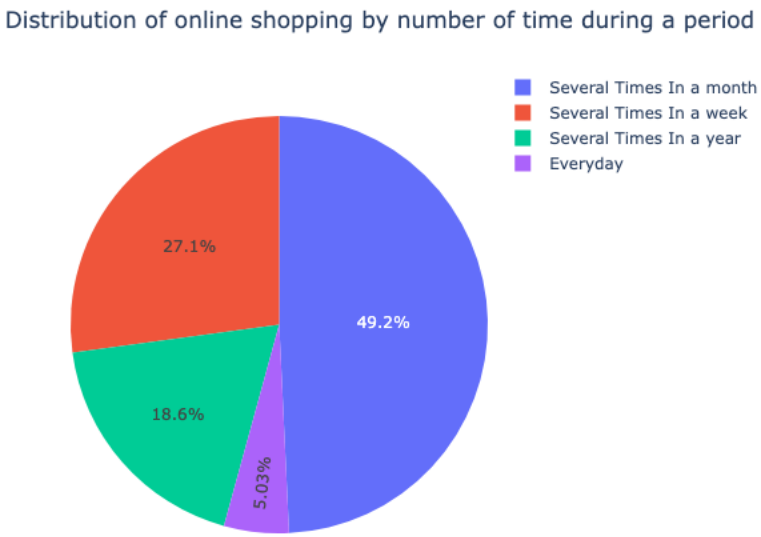


Figure 49 Distribution of Online Shopping by Number of times during a period

Distribution of Online Shopping by Number of times during a period
The information given illustrates how frequently a specific action occurs. The replies are divided into four groups according to when they occur: "Several Times In a Month," "Several Times In a Week," "Several Times In a Year," and "Everyday."

98 respondents, or 49.25% of the total, said that they did the activity "Several Times In a Month." This means that among the people polled, this frequency is the most prevalent.

The second most common frequency indicated is "Several Times In A Week," with 54 respondents (27.14%) expressing their engagement at this frequency.

37 respondents (18.59%) of the total respondents reported doing the activity "Several Times In a Year." This implies that this frequency is less frequent than the latter two groups.

The last group of respondents, 10 people (5.03%), reported doing the activity "Everyday," which indicates a greater level of regularity and frequency compared to the previous groups.

This information sheds light on the distribution of participation rates in the activity that is the subject of the survey. When analyzing customer behavior, creating marketing plans, or improving product or service offerings, it can be helpful to uncover patterns, preferences, and trends connected to the frequency of involvement.

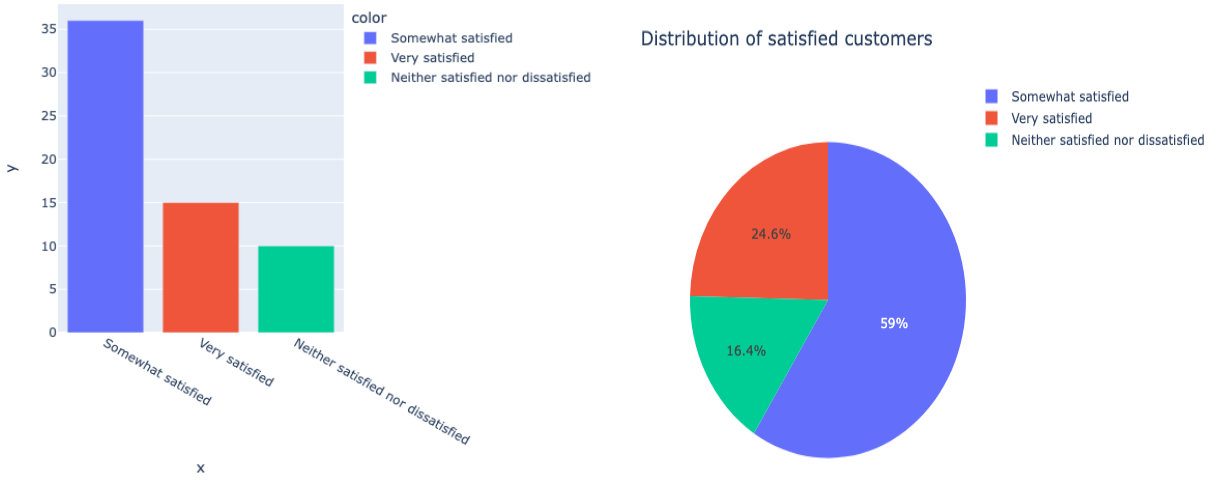


Figure 50 Distribution of Satisfied Customers

Distribution of Satisfied Customers

Based on certain criteria, the presented data shows the respondents' degrees of satisfaction. There are three different categories for satisfaction levels: "Somewhat satisfied," "Very satisfied," and "Neither satisfied nor dissatisfied."

36 people, or 18.09% of the total responses, said that they were "Somewhat satisfied." This suggests that the respondents in this category are generally satisfied.

15 respondents (7.54%), a lesser percentage, said that they were "Very satisfied." This shows that this specific group is more satisfied overall.

In addition, 10 people (5.03%) indicated that they were "neither satisfied nor dissatisfied." This category denotes a middle ground where respondents' attitudes are neither overwhelmingly positive nor negative.

The information reveals how respondents' levels of satisfaction varied, reflecting different levels of satisfaction with the particular evaluation criteria. In order to gauge consumer sentiment, identify areas for development, and develop tactics to raise overall satisfaction levels, it might be helpful to understand these satisfaction levels.

To acquire a complete knowledge of the elements impacting satisfaction levels and to address any possible problems or chances for improvement, it is crucial to take into account the survey's unique criteria and context.

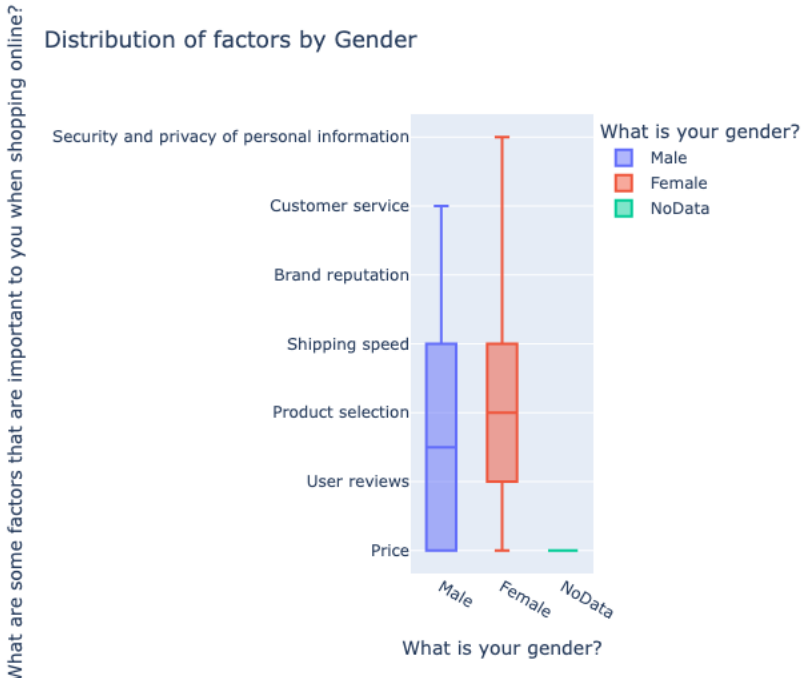


Figure 51 Distribution of Factors by Gender

Distribution of Factors by Gender

In accordance with the information supplied, both male and female respondents' perspectives on the various elements affecting their decision-making are presented. Each component is divided into gender categories, and the matching number of replies is shown.

According to the female respondents, customer service received four mentions, price received six, product selection received four mentions, security and privacy of personal information received two mentions, shipping speed received eight mentions, and user reviews received eight mentions. Brand reputation received two mentions.

Mentioned topics among the male respondents included brand reputation (2 mentions), customer service (3 mentions), pricing (9 mentions), product selection (4 mentions), shipment speed (4 mentions), and user reviews (4 mentions).

It is important to note that there is one case where the response for the pricing category is marked as "NoData," which denotes that a response was either not supplied or could not be established.

This information sheds light on the variables that matter to respondents of both genders when making choices based on the analyzed criteria. It draws attention to the differences between the two genders' priorities and interests. Both men and women appeared to place a lot of importance on factors including company reputation, customer service, pricing, product variety, shipment speed, and user reviews.

Figure 52 Important Factors while Shopping Online

Important Factors while Shopping Online

The information given demonstrates the significance or applicability of numerous elements in respondents' decision-making. The following variables are taken into account: "Price," "Shipping speed," "User reviews," "Product selection," "Customer service," "Brand reputation," and "Security and privacy of personal information." For each factor, the associated response count is given.

With 16 references, "Price" appeared as the most important component in the data, demonstrating its critical role in the decision-making process. With 12 references apiece, the factors "Shipping speed" and "User reviews" clearly had a significant impact on the decision-making process.

A smaller degree than the aforementioned aspects, "product selection" was noted by 8 respondents, demonstrating its relevance. Seven references of "customer service" indicate that it has a limited impact on choice-making.

Indicating a significantly lower degree of relevance for the respondents, the variables "Brand reputation" and "Security and privacy of personal information" garnered 4 and 2 mentions, respectively.

This information offers insightful information on the variables that respondents find important when making decisions. Price, shipment speed, and user reviews stand up as highly important variables, indicating that cost, delivery time, and customer feedback are vital things to take into account. The decision-making process is also significantly influenced by product choice and customer service, but to a much lesser extent.

Understanding these preferences may help organizations better cater to their target market's demands and priorities by streamlining their customer service, increasing brand reputation, and optimizing their product and service offerings.

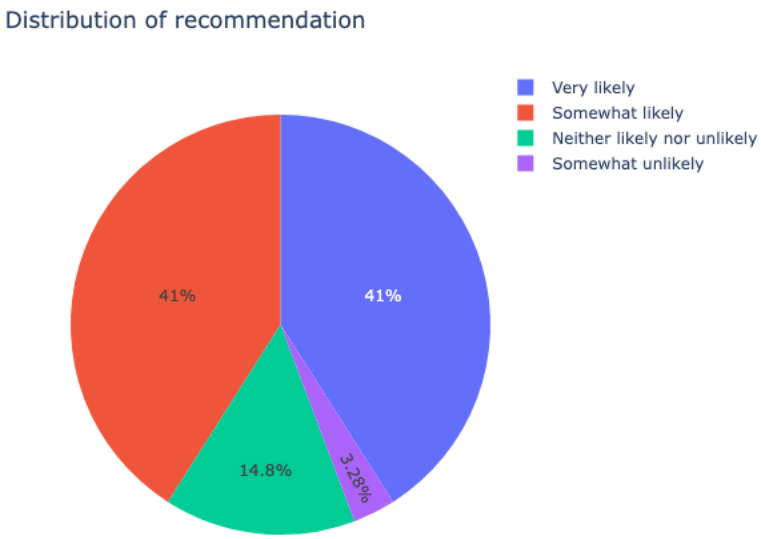


Figure 53 Distribution of Recommendation

Distribution of Recommendation
The information supplied illustrates how likely it is that someone will react to a certain claim or query. There are four levels of likelihood: "Neither likely nor unlikely," "Very likely," "Somewhat likely," and "Somewhat unlikely." There is a response count for each category.

The findings show that 25 respondents gave the statement a "Very likely" response, signifying a high degree of agreement or plausibility. Additionally, 25 respondents said that they were "Somewhat likely," which denotes a modest degree of agreement or possibility.

Nine respondents, a lower proportion, indicated that they were "Neither likely nor unlikely," indicating that they had no strong feelings about the concept.

Only 2 respondents indicated they were "Somewhat unlikely," suggesting a lesser degree of likelihood or agreement.

Based on respondents' propensity to support the proposition being reviewed, this data offers insights into the distribution of replies. A broad inclination towards agreement or positive mood is shown by the presence of a sizable number of answers in the "Very likely" and "Somewhat likely" categories.

Understanding the amount of agreement or probability given by the respondents can be helpful for evaluating viewpoints, forecasting behavior, or making educated decisions.

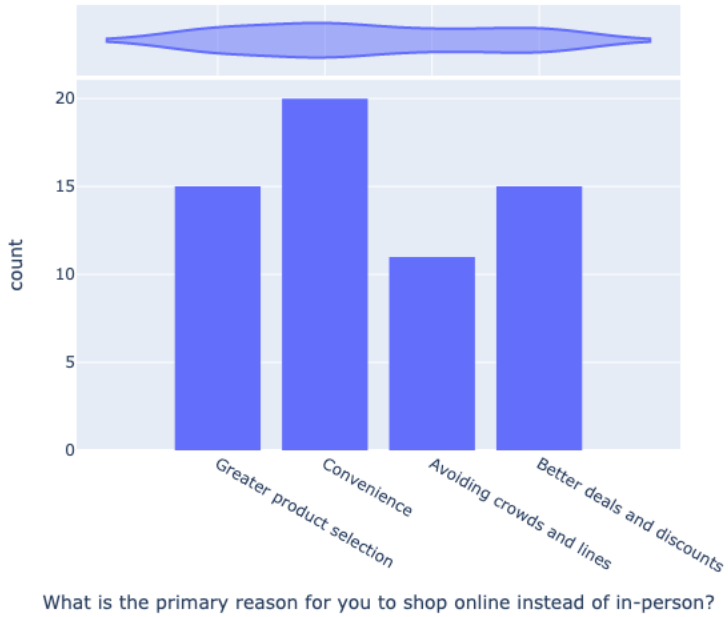


Figure 54 Primary Reason of Shopping Online

Primary Reason of Shopping Online

The information given reflects the motivations or elements affecting respondents' behavior. The following elements are taken into account: "Convenience," "Greater Product Selection," "Better Deals and Discounts," and "Avoiding Crowds and Lines." For each factor, the number of replies is reported.

The findings show that 20 respondents mentioned "Convenience", indicating that this aspect significantly affects how they behave. The ease and pleasure connected with purchasing in a convenient way, whether internet shopping or organized in-store experiences, is probably meant by the convenience factor.
Both "Better deals and discounts" and "Greater product selection" were mentioned by 15 respondents each, showing that a sizeable percentage of the respondents value having a large selection of products to choose from and look for appealing deals or discounts when making their purchasing decisions.

With 11 mentions, the factor "Avoiding crowds and lines" indicates that a sizeable percentage of respondents prefer to buy away from busy areas or lengthy lines. This may be especially important in the context of actual retail locations or times of strong demand, like the holidays.

This information sheds light on the variables that affect respondents' behavior while making purchasing decisions. The respondents identified convenience, a wider range of goods, better prices and discounts, and avoiding crowds and lineups as top priorities.

Understanding these elements may help businesses improve convenience, optimize their product offers and pricing, and provide customers a better shopping experience.

Distribution of Category

The distribution of several categories according to their count and percentage is shown in the pie chart. The relative proportions of each category within the dataset are summarized. Eight categories are represented on the graph, each with a corresponding count and percentage.

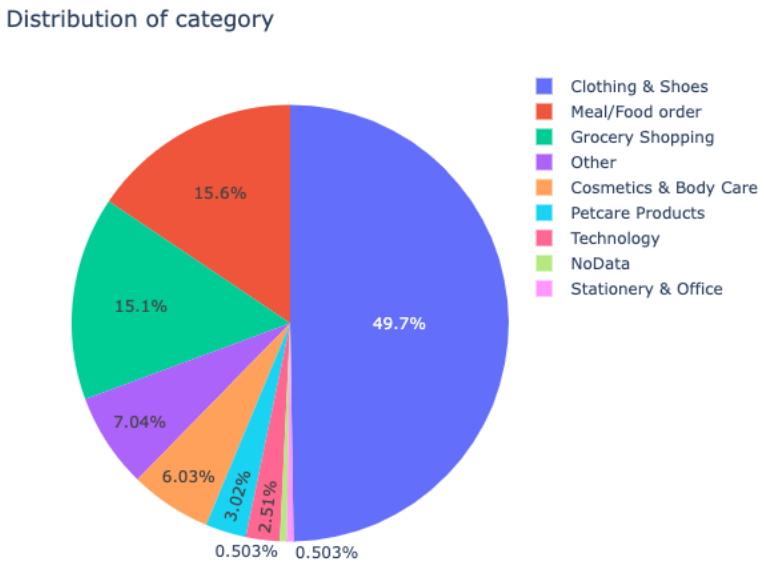


Figure 55 Distribution of Category

"Clothing & Shoes," which comprises 99 items or 49.75% of the total, is the largest category. "Meal/Food order" comes in second place with 31 items, accounting for 15.58% of the distribution. 30 items, or 15.08% of the chart, are categorized as "Grocery Shopping".

14 items fall under the "Other" category, making up 7.04% of the distribution. Following "Cosmetics & Body Care" are 12 entries, or 6.03% of the pie chart. Six items under the heading "Petcare Products" make up 3.02% of the distribution.

The category "Technology" is further down the chart, with five items making up 2.51% of the total. One item each from the "NoData" and "Stationery & Office" categories makes up 0.50% of the distribution.

A rapid and simple explanation of the relative proportions of various categories within the dataset is made possible by this pie chart, which provides a visual depiction of the category distribution.

Category Distribution by Gender

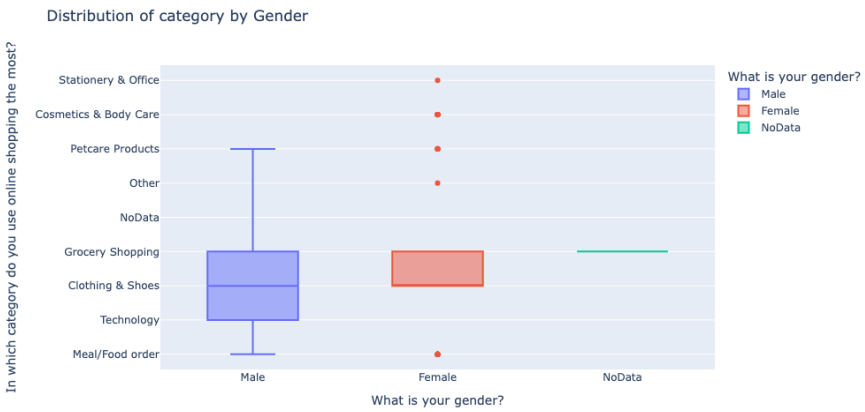


Figure 56 Category Distribution by Gender

The information provided depicts the gender distribution of categories in the context of online buying. The pie chart gives a broad overview of the distribution for each gender based on the number and percentage of each category.

The most popular online shopping category among female respondents is "Clothing & Shoes," with 63 respondents representing 54.78% of the female distribution. "Cosmetics & Body Care" is the second most popular category, with 12 respondents, or 10.43% of the female respondents. Following "Grocery Shopping" are 22 people, or 19.13% of the female distribution. Of the 11 female respondents—9.57% of the total—"Meal/Food order" was their answer. Additionally, the "Other" and "Stationery & Office" categories each had one female respondent, accounting for 0.87% of the distribution in each category. Finally, five female internet shoppers (4.35%) said they choose "Petcare Products".

With 36 respondents, or 43.37% of the male respondents, the category with the greatest count is also "Clothing & Shoes" for men. After that, "Grocery Shopping" came in second with seven replies, making up 8.43% of the male respondents. Twenty individuals, or 24.10% of the male distribution, selected "Meal/Food order." One male responder, or 1.20% of the male respondents, fell into the "NoData" group. Thirteen males, or 15.66% of the male distribution, favor the "Other" group. In the categories of "Petcare Products" and "Technology," there is only one male respondent, which represents 1.20% and 6.02% of the male distribution, respectively.

One respondent, who represents 100% of the distribution in the "NoData" gender group, did not indicate their gender but preferred the category "Grocery Shopping" for online purchasing.

By highlighting the differences in their choices and enabling comparison, this category distribution by gender pie chart sheds light on the preferences of the various genders when it comes to online shopping.

Payment Method Distribution

Figure 57 Payment Method Distribution

The information given reflects the distribution of online purchasing payment options. Instead of a pie chart, a histogram that shows the number and percentage of each payment method is used to represent this distribution.

The histogram displays the frequency or count on the y-axis and the various payment methods on the x-axis. The most popular payment method is represented by the tallest bar on the histogram, while the least popular is shown by the shorter bars.

The results show that "credit card" is the most common payment method, accounting for 75.88% of distribution with 151 instances. The histogram's tallest bar, which represents this strategy, is shown. "Cash," which was used 40 times and accounted for 20.10% of all instances, is the second most common payment method. The bar that corresponds to "cash" is smaller than the bar that corresponds to "credit card."

"Installment shopping," "Other," and "Cash on Delivery," the remaining payment options, have lower counts. With three instances each, "installment shopping" and "Other" account for 1.51% of the distribution. Two instances, or 1.01% of all instances, are for the "Cash on Delivery" option. On the histogram, these payment methods are represented by shorter bars.

The histogram gives a visual depiction of the distribution of payment options so that their frequencies may be quickly compared. It emphasizes how cash payments are the second most popular mode of payment after credit card payments. Installment shopping, "Other," and "Cash on Delivery" are included, indicating that these alternate ways are also used, albeit less frequently.

Shopping Website Distribution

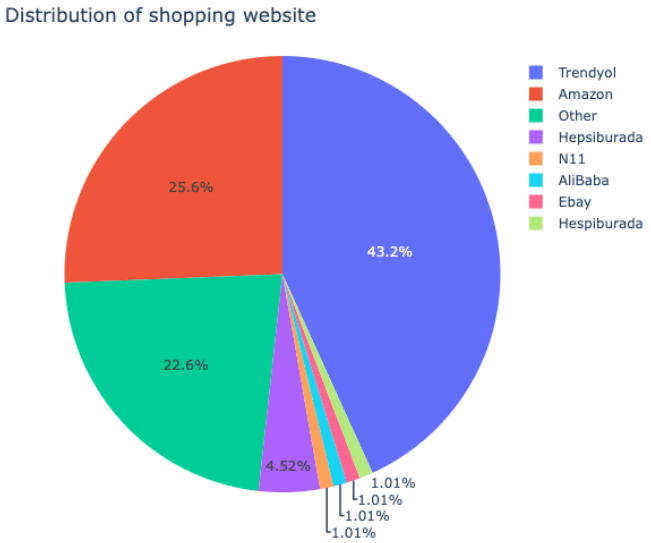


Figure 58 Shopping Website Distribution

The information given reflects the spread of online retailers that people utilize. A table illustrating this distribution and displaying the number and percentage of people who use each website illustrates the distribution.

"Trendyol" stands up as the most popular option among the mentioned websites, being used by 86 users, or 43.22% of the distribution, to do online shopping. Using "Amazon," which is used by 51 people, or 25.63% of the total, is right behind it. 45 people are included in the "Other" group, which accounts for 22.61% of the distribution.

Nine people, or 4.52% of the dataset, choose the website "Hepsiburada" for their online purchasing. Two users each of "N11," "AliBaba," and "Ebay" account for 1.01% of the distribution for each website. The spelling of "Hepsiburada" differs in two places, with the variations "Hespiburada" and "Hespiburada," which could be typographical errors.

The popularity and usage of numerous online shopping websites are revealed by this distribution. Top picks "Amazon" and "Trendyol" stand out since a sizable portion of people use these websites for their online shopping. Minor differences in the spelling of "Hepsiburada," together with the existence of other websites, suggest that the respondents to the survey had a wide variety of tastes.

Annual Income Distribution

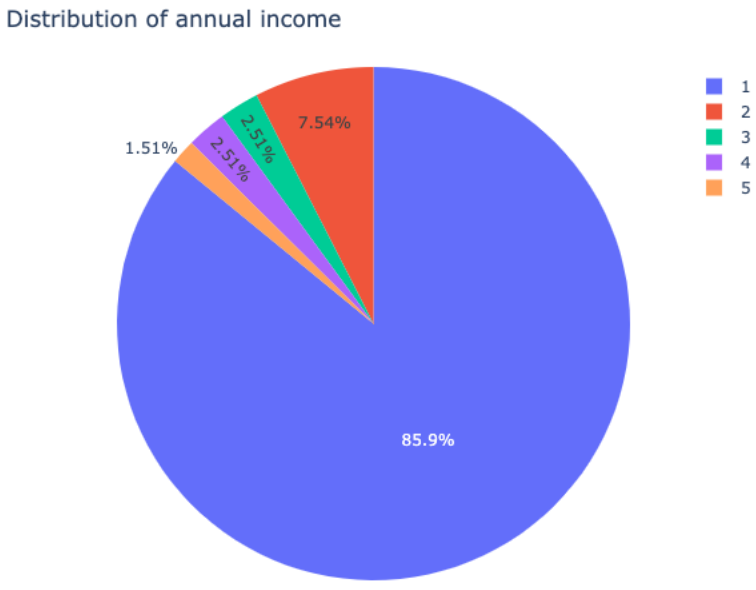


Figure 59 Annual Income Distribution

The information presented shows how different people's annual income levels are distributed. The distribution is shown in a tabular fashion, with the number and proportion of people in each income bracket shown.

The most common income group among those listed is "1.0," which is represented by 171 people and makes up 85.9% of the distribution. This suggests

that the majority of those questioned belong to this income group.

With 15 people, or 7.54% of the total, the "2.0" income category has the second-highest representation. This group reflects people who earn more money than the majority of people.

Five people in each of the income categories of "3.0" and "4.0" make up 2.51% of each category's distribution. A smaller percentage of people in these categories fall into the upper income brackets.

The income category with the lowest representation is "5.0," which has three people and represents 1.51% of the total. The highest-earning members of the population under study fall into this category.

This distribution sheds light on how the respondents' annual income levels were distributed. The bulk of people fall into the "1.0" income group, while smaller percentages of people are spread out over higher income tiers. It highlights the different income levels within the dataset and enables a rapid comparison of the income distribution among the population being polled.

Budget Distribution Before Covid-19 and Budget Distribution After Covid-19

The information shown shows how the funding levels were distributed both before and after the Covid-19 pandemic. The count and proportion of people in each budget category are shown for each dataset in a tabular manner.

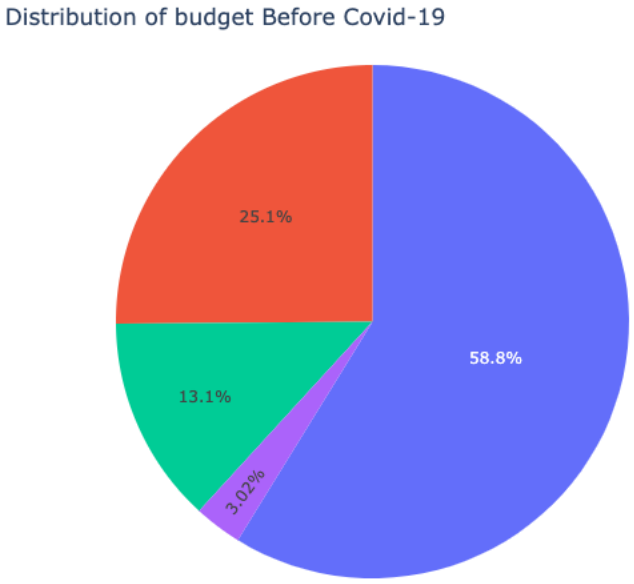


Figure 60 Budget Distribution Before Covid-19

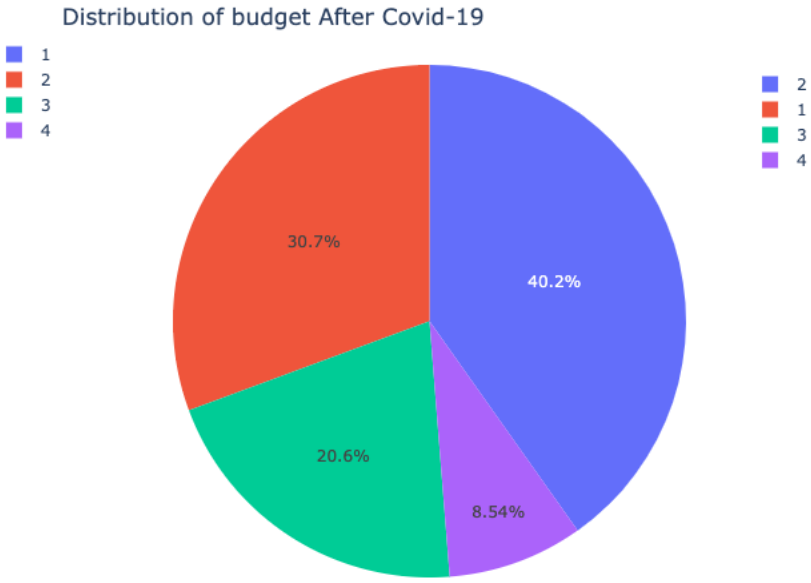


Figure 61 Budget Distribution After Covid-19

Before Covid-19, 117 people accounted for 58.79% of the distribution, with the bulk of them falling into the "1.0" budget bracket. This suggests that a sizable portion of people had tighter finances prior to the outbreak. "2.0" was the second-most represented budget group with 50

participants, or 25.13% of the total. There were 26 people (13.07%) and six people (3.2%) in the "3.0" and "4.0" budget groups, respectively. These groups included people who had larger budgets than the majority did.

The distribution of the money has changed significantly after Covid-19. The most common budget category was "2.0," which had 80 people and accounted for 40.20% of the distribution. This suggests that more people after the pandemic set aside a modest budget. Prior to Covid-19, the "1.0" budget category was the largest; today, it accounts for 61 people, or 30.65% of the total. 41 people (20.60%) fall into the "3.0" budget category, which includes those with a higher income. There are 17 people in the "4.0" budget bracket, making about 8.54% of the distribution.

We can see from comparing the two datasets that the distribution of budget levels changed following the Covid-19 pandemic. Prior to the pandemic, a greater percentage of people (category 1.0) had a lesser budget, however after the outbreak, the percentage fell. After Covid-19, category 2.0's budget allocation saw the most growth, indicating a shift toward a more moderate budget allocation. After the pandemic, there was also a rise in the representation of higher budget categories (3.0 and 4.0).

This comparison shows how the Covid-19 outbreak affected people's budget distribution, shifting toward various budget levels as people changed their spending habits in response to the shifting economic situations.

Budget Distribution Before Covid-19 Grouped by Profession and Educational Background

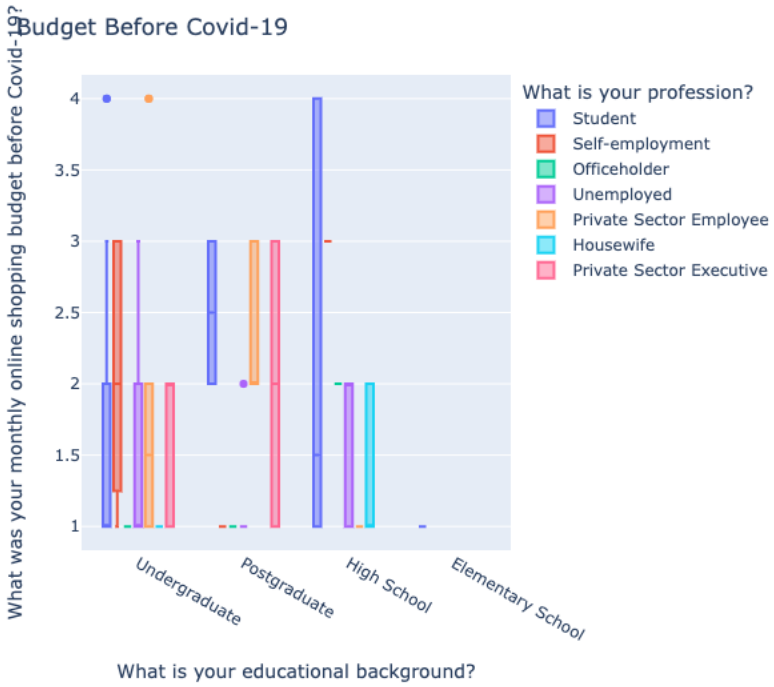


Figure 62 Budget Distribution Before Covid-19 Grouped by Profession and Educational Background

Budget Distribution After Covid-19 Grouped by Profession and Educational Background

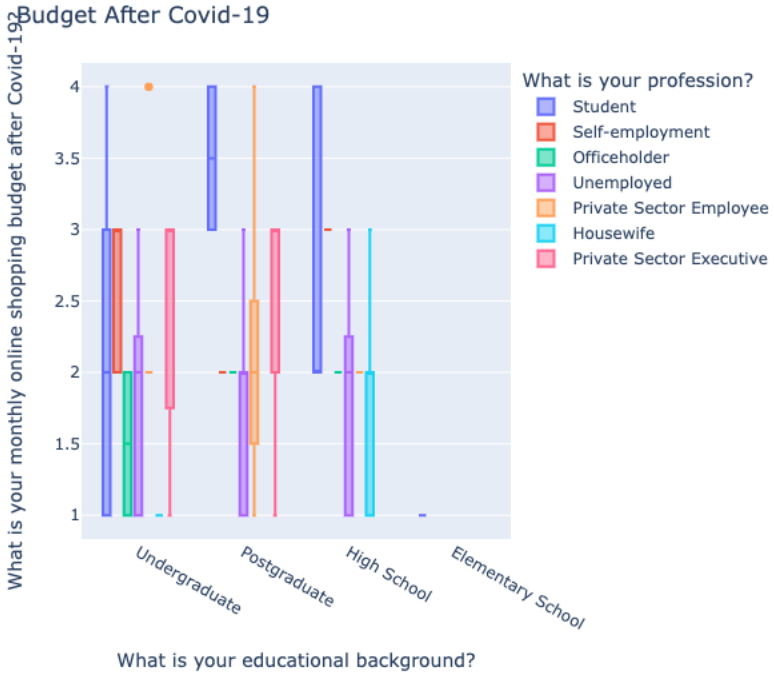


Figure 63 Budget Distribution After Covid-19 Grouped by Profession and Educational Background

The information given includes the budget distribution by occupation and level of education before and after the Covid-19 pandemic. Each dataset is displayed in a tabular manner, with the number of items in each category, their frequency, and whether they are unique or not.

Before Covid-19, certain tendencies could be seen among people from various disciplinary backgrounds and occupations. Most people with "Undergraduate" degrees and "Student" as their occupation had a budget level of "1.0". 102 people fell into this category, making up 54.8% of the distribution for this group. With 12 people, or 6.45% of the distribution for this group, the "Undergraduate" category with a "Private Sector Employee" career and a budget level of "1.0" was the second most common.

There are observable similarities and shifts in the budget distribution across various educational and professional backgrounds following the Covid-19 pandemic. As before, "Undergraduate" people with a "Student" profession and a budget level of "1.0" account for 102 people, or 50.25% of the distribution for this group, making them the most common budget category. There have been some changes in other categories, though. For instance, those with a "High School" education and a "Housewife" occupation are now more prevalent in the "2.0" budget category, accounting for 6 people or 30% of the distribution for this group.

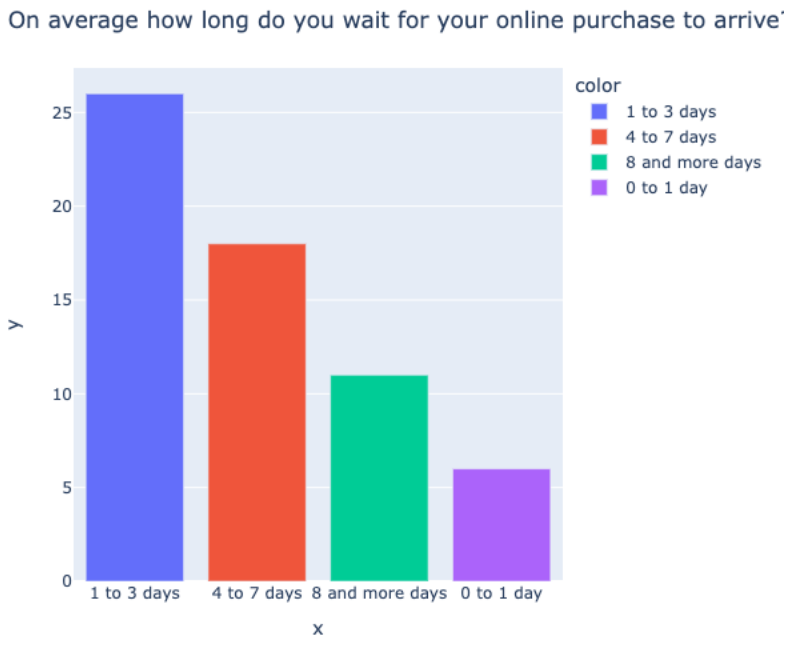
By comparing the two datasets, we can see that, both before and after the Covid-19 pandemic, the prevalence of various budget categories is still largely stable for people with particular educational and occupational backgrounds. Students typically designate a smaller budget, as seen by the fact that those with "Undergraduate" educational backgrounds and the profession "Student" typically fall into the greatest budget group. However, there have been some changes in other categories, suggesting that the pandemic may have altered spending habits. For instance,

after the pandemic, people with a "High School" degree and a "Housewife" occupation are more prevalent in the "2.0" budget group.

Overall, the comparison sheds light on how budgets were distributed among people with various educational and professional backgrounds before and after the Covid-19 outbreak, highlighting both trends and probable changes in spending patterns within particular groups.

Distribution of Waiting Time for Online Purchases

Figure 64 Distribution of Waiting Time for Online Purchases



The information given demonstrates the dispersion of online purchase wait times. It displays the number of people according to various waiting time ranges.

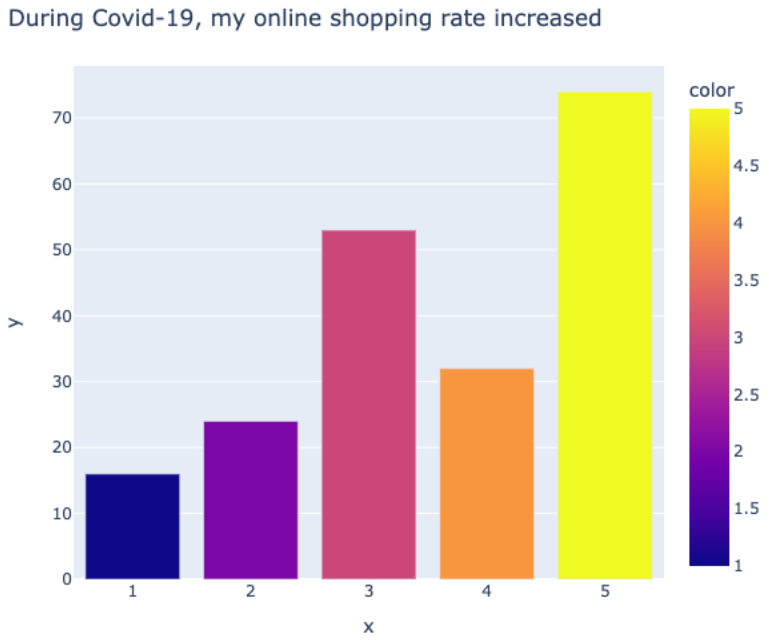
26 respondents reported a waiting period of one to three days for their internet transactions. This period represents the greatest count and shows that a sizeable portion of people received their products during this time. 18 people said that the next most typical waiting time range was between 4 and 7 days. This implies that a sizable proportion of respondents experienced a little delay in the delivery of their online purchases.

Additionally, 11 people reported a wait time of 8 days or longer. This category has a relatively low count, but it reveals that some respondents had to wait longer than average to receive their online purchases. Six people reported waiting times between 0 and 1, indicating that they received their online purchases quickly, either the same day or right away.

The distribution of waiting times for online purchases sheds light on the respondents' perceptions of delivery speed. It demonstrates that the vast majority of people got their products within a week or less, with a sizable percentage getting it within a day or two. A lower percentage of people, though, had to wait longer—more than 7 days—to receive their online purchases.

Distribution of Increased Online Shopping Rate During Covid-19

Figure 65 Distribution of Increased Online Shopping Rate During Covid-19



The information shown shows how the rise in internet sales during the COVID-19 outbreak was distributed. It displays the number of people related to various increases in online shopping.

Among the respondents, 74 people gave the pandemic a grade of 5.0, suggesting a considerable growth in their online purchasing habits. The largest number in this category indicates that a significant number of people depended heavily on online shopping as a result of the epidemic.

53 people gave the next-most prevalent rating, 3.0, as their response. This suggests that these people used online platforms to a greater amount than those in the category with the highest ratings, but only to a moderate extent.

A rating of 4.0 was also reported by 32 people, which shows that the pandemic significantly increased people's internet buying habits. This category has a sizable sample size, which shows that a sizeable proportion of respondents have drastically changed the way they shop by turning to online stores.

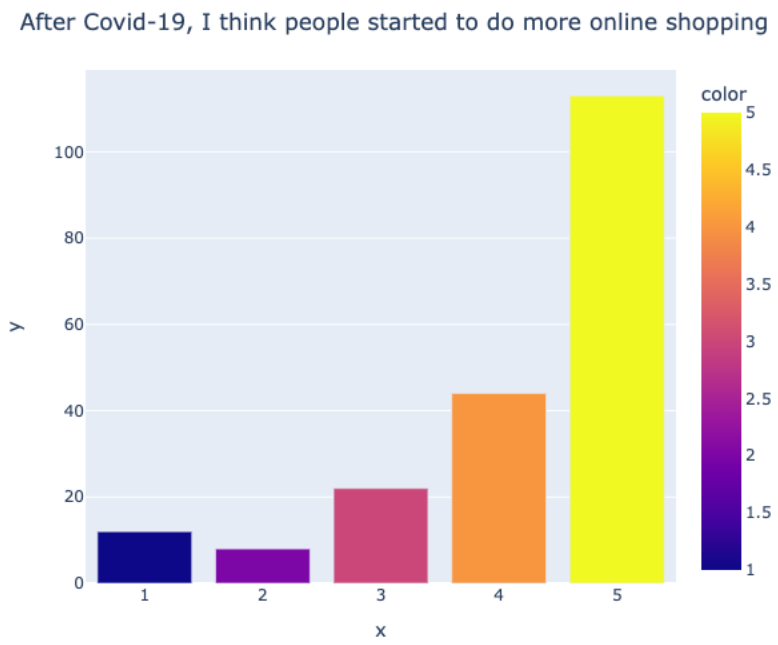
Additionally, 24 people reported a rating of 2.0, indicating a little uptick in online buying. This suggests that compared to the higher rating categories, these people expanded their reliance on online platforms to a lesser level.

In the end, 16 people reported a rating of 1.0, which shows a negligible rise in internet buying during the epidemic. These people most likely kept up their regular buying routines with only minor alterations for online retailers.

The distribution of the higher online shopping rate during COVID-19 sheds light on how customer behavior has changed. It demonstrates that while some people reported moderate or minor increases in internet purchasing, a sizable number of people reported a large increase. These results show how the epidemic affected customer choices and the reliance on online buying platforms.

Distribution of Perception: People Started to Do More Online Shopping After Covid-19

Figure 66 Distribution of Perception: People Started to Do More Online Shopping After Covid-19



The information provided reflects how people see the rise in online shopping following the Covid-19 pandemic. It displays the number of people matching to various levels of perception.

The perception level of 5 was recorded by 113 respondents, who strongly believed that after the epidemic, people had greatly increased their online buying. This area has the highest number of responses, indicating that the majority of respondents strongly believe that internet shopping has increased among the general public.

4 was the next most prevalent impression level, as indicated by 44 people. This shows that a sizable portion of respondents believe that following the pandemic, people's online shopping has moderately grown.

The perception level of 3 was also recorded by 22 people, indicating a moderate perception of increased online purchasing. This area has a sizable number of entries, which implies that some respondents indeed, to a lesser extent than the higher perception levels, believe that consumers have increased their online buying.

Additionally, 12 participants responded that their perception level was 1, which indicates that they had a very limited perception of the rise in internet buying. This shows that these people think that people's internet buying hasn't considerably grown as a result of the pandemic.

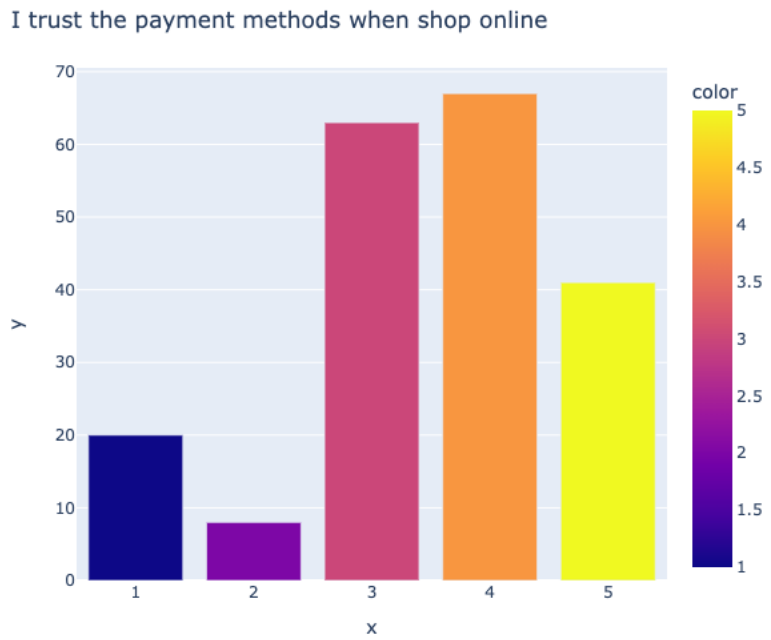
Eight respondents gave their assessment of the growth in online shopping a score of 2, which indicates a mild perception. This suggests that although not as pronounced as the higher perception levels, these people think there has been a minor rise in internet buying among the general public.

The way respondents view the shifts in consumer behavior is revealed by the distribution of perceptions regarding the rise in online shopping following COVID-19. It displays a variety of impressions, ranging from strong perceptions of considerable increases to softer perceptions of

moderate or little changes. These results show how different people have different opinions on how much the pandemic has affected people's internet buying behavior.

Distribution of Trust in Online Payment Methods when Shopping Online

Figure 67 Distribution of Trust in Online Payment Methods when Shopping Online



The information provided illustrates the range of levels of trust that consumers have in online payment options. It displays the number of people according to various levels of trust.

67 respondents gave an average trust score of 4, suggesting a high level of confidence in online payment methods. This area has the highest number of responses, which indicates that the majority of respondents have a high level of confidence in online payment methods while making purchases online.

Three was the next most prevalent trust level, as stated by 63 people. Inferring a respectable level of confidence in the security and dependability of online payment systems, this shows that a sizable portion of respondents have a moderate level of trust in them.

In addition, 41 people reported having a trust level of 5, which is the greatest possible level of confidence in online payment systems. This category has a sizable sample size, which shows that a significant proportion of respondents place a very high value on the safety and dependability of online payment methods.

Additionally, 20 people indicated that their trust level in online payment methods was 1, which is a poor degree of trust. This implies that these people have doubts or worries regarding the dependability and security of online payment methods when they shop online.

Eight respondents gave their trust level a score of 2, indicating a moderate level of mistrust or caution toward online payment methods. This suggests that despite having certain doubts, these people still have a certain amount of faith in the security and dependability of online payment systems.

The distribution of respondents' levels of trust in online payment options while making purchases online reveals how respondents judge the security and dependability of such payment options. It displays a range of trust attitudes, from strong trust to various levels of caution or mistrust. These

results underline how crucial it is to establish and uphold trust in online payment systems in order to increase customer trust in making purchases online.

Distribution of Perception: Websites Suggest Items I'm Looking For When Shopping Online

Figure 68 Distribution of Perception for Websites

The information given reflects how people feel about websites that recommend products to people who are doing online shopping. It displays the number of people matching to various levels of perception.

65 respondents gave a perception score of 4, indicating a strong belief that websites effectively propose products they are seeking for when they conduct online shopping. The fact that this category has the most votes indicates that most survey participants think website suggestions are accurate and relevant.

3 was the next most prevalent perception level, as stated by 57 people. This implies that although there is certainly space for improvement, a sizable portion of respondents have a relatively positive opinion that websites actually propose items they are looking for.
Additionally, 32 people reported having an impression rating of 5, which is the most favorable. This category has a sizable sample size and implies that some respondents have a strong opinion that websites are very good at suggesting products they would like, indicating a high degree of satisfaction with the algorithms used for suggestion.

Furthermore, 18 people gave a perception score of 1, which denotes a low perception of the relevance of website suggestions. This indicates a need for improvement in the precision and relevancy of website suggestions because it demonstrates that these people believe websites are poor at proposing the products they are looking for.

Finally, 27 respondents gave their assessment of websites' ability to propose relevant items a rating of 2, which indicates a modest perception of that effectiveness. This suggests that although these people think websites' suggestion algorithms may use some work, they still perceive some degree of accuracy in the recommendations.

The distribution of opinions on websites that propose products people might be interested in buying when they browse online sheds light on how respondents view the value of recommendation systems. It displays different levels of satisfaction, ranging from high levels of trust in website ideas to lesser levels of perceived correctness. These results underline how crucial it is to improve suggestion algorithms in order to better cater to customer preferences and raise consumer happiness with online purchasing.

Distribution of Shopping Behavior: Currently Shopping Online Less Than During Covid-19

Figure 69 Distribution of Shopping Behavior

The information presented shows how people who currently buy online less frequently than they did during the Covid-19 period distribute their shopping habits. It displays the number of people according to various behavior levels.

Among the respondents, 63 people indicated that they greatly reduced their online buying compared to the time during Covid-19 by reporting a behavior level of 1. The majority of respondents appear to have noticed a significant decline in their online buying, as this category has the greatest count.

The following two behavior levels, both recorded by 44 people, were 3 and 2. Inferring a change in their shopping patterns, this shows that a sizable portion of respondents have moderately decreased their online shopping compared to the peak Covid-19 period.

In addition, 24 people indicated a behavior level of 5, which implies a slight decline in their online buying habits. This shows that these people have only slightly decreased their online purchasing and still do it to a significant degree.

Additionally, 24 people indicated that their behavior level was 4, which implies a little decline in their online buying activity compared to the Covid-19 period. This shows that although not as much as those with the highest behavior level, these people have seen a discernible drop in their online buying behaviors.

The distribution of purchasing behavior among people who currently shop online less frequently than during the Covid-19 period sheds light on how respondents' online shopping behaviors have

changed. It shows various levels of reduction, from sharp drops to smaller changes. These data demonstrate how internet shopping dynamics are evolving as well as the impact of outside variables like the Covid-19 outbreak on consumer preferences and buying behavior.

Distribution of Perception: Impact of Better Product Suggestions on Satisfaction When Registering to an Online Shopping Site

Figure 70 Distribution of Perception for Better Product

The information given reflects how people who have registered on an online retailer feel about how better product recommendations affect their level of happiness. It displays the number of people matching to various levels of perception.

69 respondents gave their assessment of the situation a rating of 5, indicating that they firmly feel that improved product recommendations significantly increase their happiness when using an online shopping site. The fact that this category received the most votes indicates that the majority of respondents place a high value on and appreciate the influence of ideas for better products.

4 was the next most prevalent perception level, as stated by 62 people. This shows that a sizable portion of respondents, albeit not to the same degree as those in the highest assessment level, believe that improved product ideas contribute favorably to their satisfaction.

In addition, 34 people indicated that their perception level was 3, indicating a moderate belief in the influence of suggestions for better products on their pleasure. This shows that while these people value better ideas, they may not view them as being essential to their level of pleasure in general.

Additionally, 18 people rated a perception level of 2, indicating a modest belief in the impact of suggestions for better products on their satisfaction. This suggests that these people might not place a lot of value on individualized product recommendations and might place greater weight on other aspects of happiness.

Last but not least, 16 people indicated that their perception level was 1, which is the lowest total. This shows that these people have little faith in the influence of better product recommendations on their pleasure when signing up for an online store.

The distribution of opinions on how better product recommendations affect customer satisfaction sheds light on how important people find tailored recommendations. It shows different levels of belief in the impact of these recommendations on general satisfaction. These findings draw attention to the potential value of customized product recommendations in boosting user experience and fulfilling customer expectations on online retail websites.

Distribution of Gender in the Dataset

The gender distribution in the dataset is shown in the pie chart. There are three categories in the data: Female, Male, and NoData.

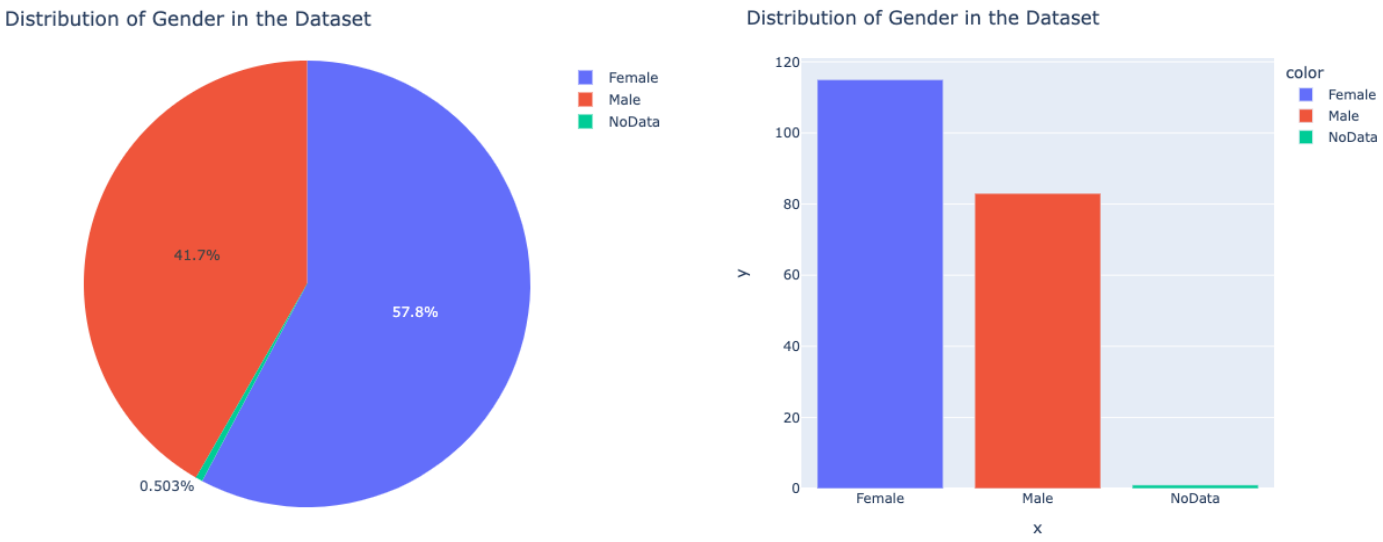


Figure 71 Distribution of Gender in the Survey

Women make up 115 of the respondents, which is the majority of the population. This shows that a sizable share of the sample is made up of females.

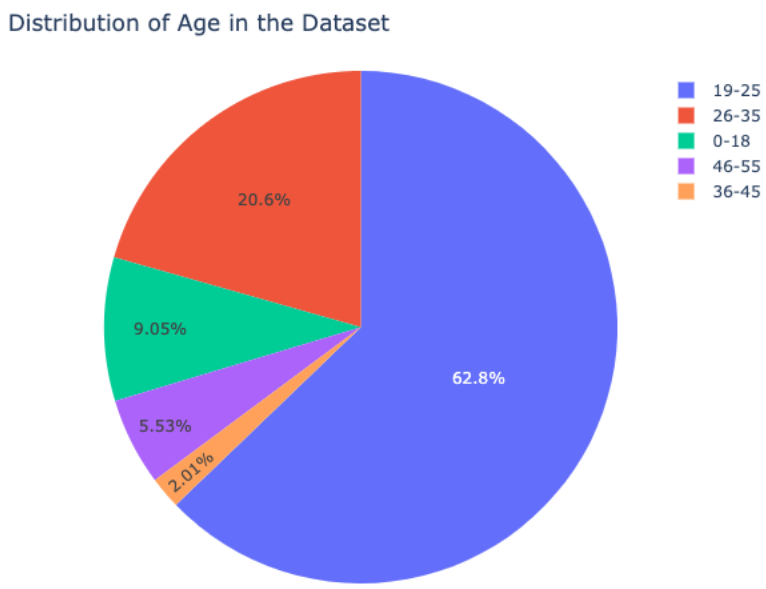
Males make up the second-largest category, with a total of 83. This implies that men are likewise fairly represented in the sample, albeit to a little lesser degree than women.

Last but not least, there is one entry marked as NoData, suggesting that the gender of one person in the dataset is either unavailable or missing.

Overall, the pie chart shows how the gender distribution in the dataset looks visually, with females being the most common gender followed by males.

Distribution of Age in the Dataset

Figure 72 Distribution of Age in the Survey



The categories for the age distribution in the dataset are: 19–25, 26–35, 0–18, 46–55, and 36–45.

The dataset's 19–25 age group, which includes 125 people, is the largest. This indicates that respondents in the early adult to mid-20s age range make up the majority of the dataset.

With 41 people, the age group 26 to 35 is the second-largest. This suggests that respondents who are in their late 20s to early 30s make up a sizable portion of the sample.

There are 18 respondents in the age range 0-18, indicating a smaller but discernible presence of respondents in their teenage or younger years.

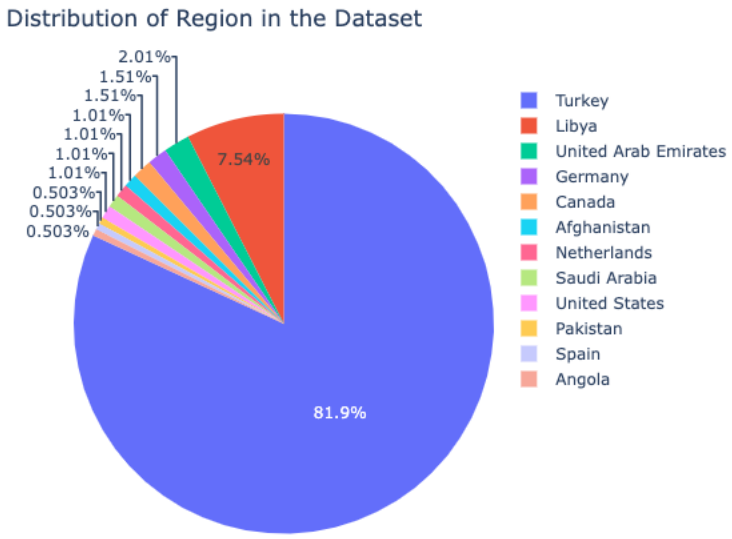
A lesser percentage of respondents in their mid-40s to mid-50s are represented by the 11 people in the 46-55 age group.

Last but not least, there are just 4 respondents in the age range of 36 to 45, indicating a significantly lower presence of respondents in their mid-30s to mid-40s.

Overall, the dataset's age distribution shows that respondents are primarily between the ages of 19 and 25, with lower but still substantial numbers in the 26 to 35, 0 to 18, 46 to 55, and 36 to 45 age ranges.

Distribution of Regions in the Dataset

Figure 73 Distribution of Age in the Survey



Various nations are represented in the dataset's distribution of regions. The regions and accompanying counts are listed below:

1. Turkey: The majority of respondents in the dataset are from Turkey, with 163 individuals representing this region.

2. Libya: There are 15 individuals from Libya in the dataset.

3. United Arab Emirates: Four individuals in the dataset are from the United Arab Emirates.

4. Germany: Three individuals are from Germany.

5. Canada: Three individuals are from Canada.

6. Afghanistan: Two individuals are from Afghanistan.

7. Netherlands: Two individuals are from the Netherlands.

8. Saudi Arabia: Two individuals are from Saudi Arabia.

9. United States: Two individuals are from the United States.

10. Pakistan: There is one individual from Pakistan.

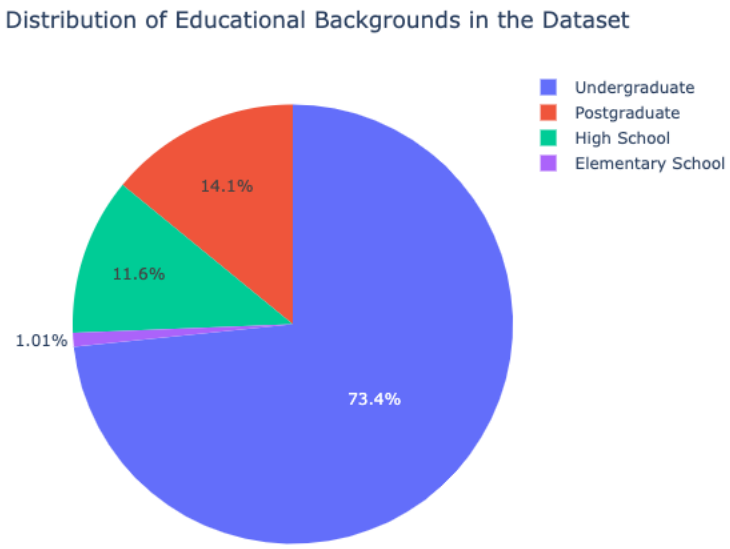
11. Spain: There is one individual from Spain.

12. Angola: There is one individual from Angola.

A majority of respondents came from Turkey, as seen by the distribution of regions in the dataset, with lesser but still significant shares coming from nations like Libya, the United Arab Emirates, Germany, Canada, Afghanistan, the Netherlands, Saudi Arabia, the United States, Pakistan, Spain, and Angola.

Distribution of Educational Backgrounds in the Dataset

Figure 74 Distribution of Education Backgrounds in the Survey



Following is the dataset's distribution of educational backgrounds:

1. Undergraduate: With 146 responders, the dataset's majority of participants had only completed their undergraduate degrees.

28 people have postgraduate degrees, according to the second category.

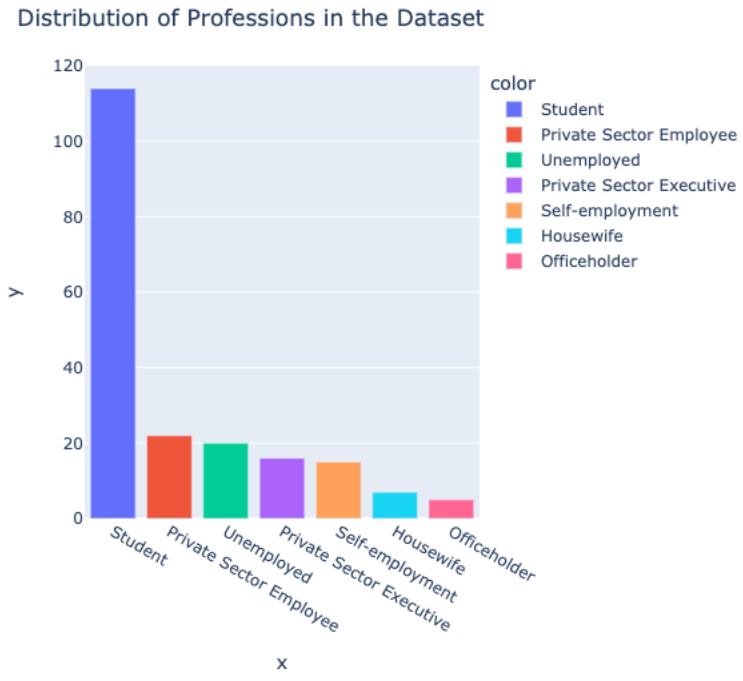
3. High School: 23 people have a high school diploma or equivalent.

4. Elementary School: There are just 2 people who have attended elementary school.

According to the distribution of respondents' educational backgrounds in the dataset, respondents with undergraduate degrees are more prevalent than those with postgraduate degrees and high school diplomas. The percentage of people having only an elementary school education is quite low.

Distribution of Professions in the Dataset

Figure 75 Distribution of Professions in the Survey



As for how the dataset's vocations are distributed, they are as follows:

1. Student: With 114 people, this profession comprises the largest group.

2. Private Sector Worker: There are 22 people employed in the private sector.

3. Unemployed: Of the 20 people in the sample, 20 are unemployed at the moment.

4. Private Sector Executive: There are 16 private sector executives in the world.

5. Self-employment: There are 15 people who work for themselves.

6. Housewife: Seven people describe themselves as being housewives.

7. Officeholder: There are five people who work in offices.

According to the dataset's breakdown of occupations, a sizable portion of respondents are students. There are people who work in the private sector, are unemployed, are executives in the private sector, are self-employed, are housewives, and people who hold office posts.

4.1.3.6. Data Preprocessing

'categorical_columns' and 'numerical_columns' are two lists that are defined in this snippet of code. In order to categorize the columns in a dataset for data preparation, these lists are utilized.

The names of the columns that correspond to categorical variables are listed in the 'categorical_columns' list. Usually, these variables have definite values or categories. The categories columns in this scenario include numerous questions about demographics, online purchasing activity, and preferences, such as "Do you shop online?" and "What is your gender?" as well as "What is your educational background?"

The names of the columns that correspond to numerical variables are listed in the numerical_columns list. These variables frequently have discrete or continuous numerical values. In this instance, responses to comments or thoughts about online shopping are included in the numerical columns. For example, "During CVID-19, my online shopping rate increased," "I trust the payment methods when shop online," and so forth.

This code snippet aids in categorizing and identifying the different sorts of variables in the dataset by grouping the columns into categories and numerical groupings. This knowledge is useful for carrying out particular preprocessing procedures or characterizing the data.

The data preprocessing phase carried out by this code snippet transforms category columns into numerical columns. To carry out this task, it makes use of the LabelEncoder class from the Scikit-Learn library.

The categorical_columns list's column names are iterated over by the code. It completes the following actions for each categorized column:

1. Creates the 'encoder' property of the 'LabelEncoder' object.
2. Apply the transformation by using the DataFrame 'df''s associated column to invoke the 'fit_transform()' method of the 'encoder' object.

3. Puts the values of the altered column back into the DataFrame.

By assigning them distinctive numerical labels, the 'fit_transform()' method of the 'LabelEncoder' encodes the categorical values. Since most machine learning models require numerical inputs, this conversion enables the models to handle and evaluate the data more efficiently.

The categorical columns in the DataFrame 'df' will be replaced with their numerical equivalents after this code has been executed. The initial category values will be converted to labels with the associated numbers.

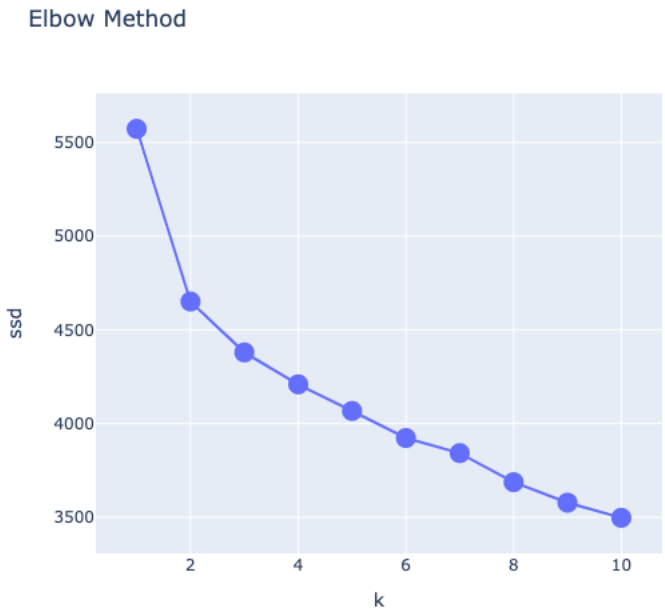
```
# Rescaled data.
df.drop(columns=['Do you shop online?'], inplace=True)
full_data = StandardScaler().fit_transform(df)
```

Figure 76 Data Preprocessing for Survey

Using the StandardScaler class from the scikit-learn library, this code snippet scales the data. Additionally, it deletes a particular column from the DataFrame.

4.1.3.7. K-Means Clustering

Elbow Method



The Elbow Method is a method for figuring out how many clusters in a K-means clustering algorithm are best. Plotting the sum of squared distances (ssd) vs the number of clusters is necessary to determine the value of k at which the ssd reduction begins to noticeably level off. The Elbow Method suggests that 10 clusters are the right number in this situation.

Figure 77 Elbow Method for Survey

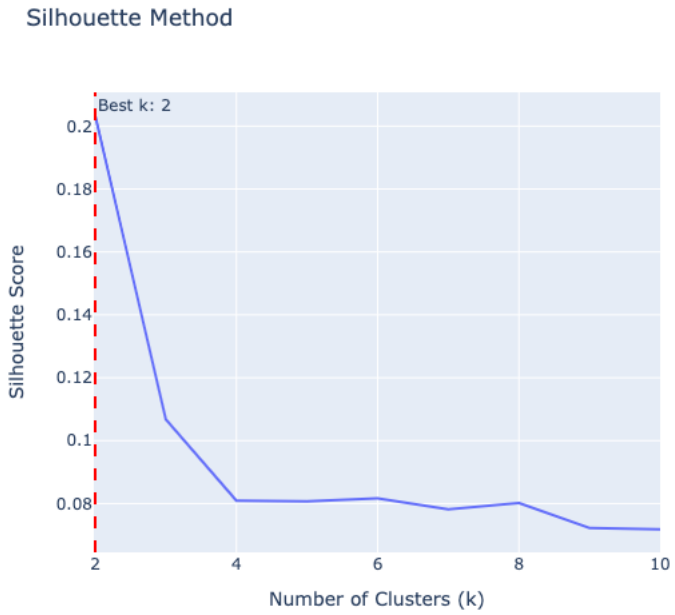
For this ideal number of clusters, the appropriate sum of squared distances (ssd) is 3496.878493851247. The squared separation between each data point and its assigned centroid inside the clusters is represented by the ssd. The clusters are more tightly packed and well-separated the lower the ssd.

We may hone and enhance the K-means clustering algorithm by determining the ideal number of clusters and the corresponding ssd. It makes the data easier to comprehend and organize, which makes it possible to conduct more insightful analyses.

Silhouette Method

Figure 78 Silhouette Method for Survey

The silhouette score in a clustering algorithm evaluates how well each data point fits into the cluster to which it has been assigned. A higher score means the data point is better matched to its own cluster and worse matched to nearby clusters, with a score ranging from -1 to 1.



The silhouette scores in the given data for various values of k (number of clusters) are as follows:

- Silhouette score for k=2: 0.2034
- Silhouette score for k=3: 0.1068
- Silhouette score for k=4: 0.0809
- Silhouette score for k=5: 0.0807
- Silhouette score for k=6: 0.0816
- Silhouette score for k=7: 0.0782
- Silhouette score for k=8: 0.0802
- Silhouette score for k=9: 0.0722
- Silhouette score for k=10: 0.0718

The maximum silhouette score is obtained for k=2, which according to these results shows that the data points are largely distinct and well-separated into two clusters. The silhouette scores fall as the number of clusters rises, indicating that the data points are less well-separated and more overlapping in the clusters.

In order to choose the right number of clusters for the data, the silhouette scores offer information on the compactness and separation of the clusters. Based on the highest silhouette score, a k value of 2 appears to be the best option in this situation.

Calinski-Harabasz Index

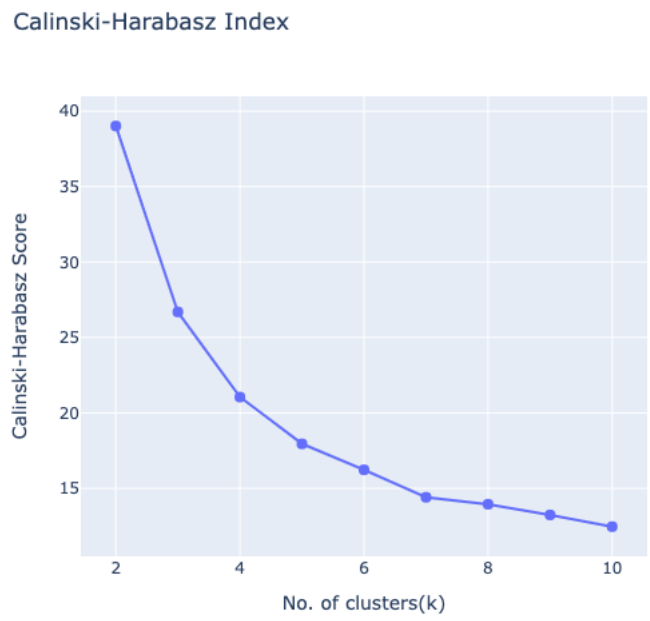


Figure 79 Calinski-Harabasz Index for Survey

Another evaluation tool used to rate the caliber of clustering findings is the Calinski-Harabasz score. To optimize the score for evenly spaced and compact clusters, it measures the ratio of between-cluster to within-cluster dispersion.

The Calinski-Harabasz scores for various values of k (number of clusters) in the provided data are as follows:

Calinski-Harabasz score for k=2: 39.0195
Calinski-Harabasz score for k=3: 26.6938
Calinski-Harabasz score for k=4: 21.0526
Calinski-Harabasz score for k=5: 17.9496
Calinski-Harabasz score for k=6: 16.2287
Calinski-Harabasz score for k=7: 14.4082
Calinski-Harabasz score for k=8: 13.9492
Calinski-Harabasz score for k=9: 13.2394
Calinski-Harabasz score for k=10: 12.4618

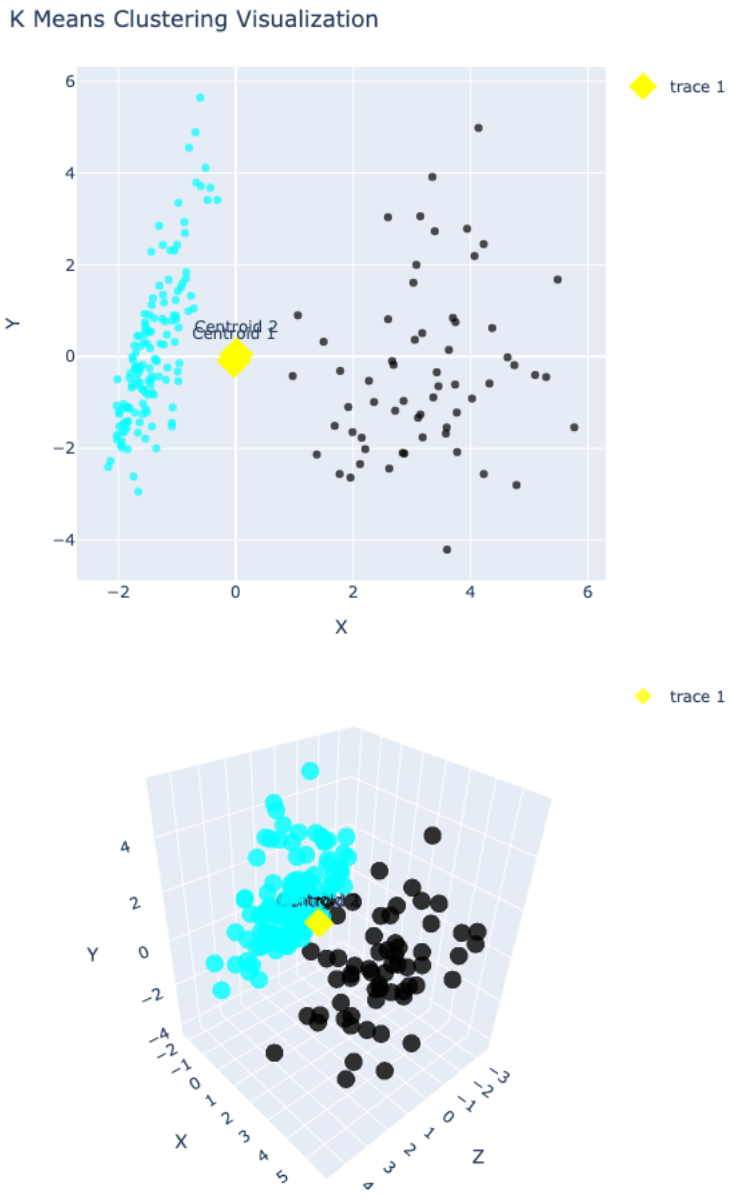
Based on these scores, the highest Calinski-Harabasz score of 2 is attained, indicating that the clusters are compact and distinct, and the data points are well-separated. The score drops as the number of clusters rises, indicating that the clusters become less distinct and more spread.

The clustering quality is indicated by the Calinski-Harabasz score, with higher scores suggesting more compact and well-defined clusters. Based on the highest Calinski-Harabasz score, a k value of 2 seems to be the best option in this situation.

4.1.3.8. K Means Clustering Visualization

A clustering method (like K-means) has been applied to a dataset based on the information given, producing two clusters. The centroids for these clusters have the following coordinates:

Figure 80 K Means Clustering Visualization



Centroid 1: (-0.03, -0.09)
Centroid 2: (0.01, 0.04)

The data points have also been assigned to the appropriate clusters. The following is how the data points were assigned to clusters:

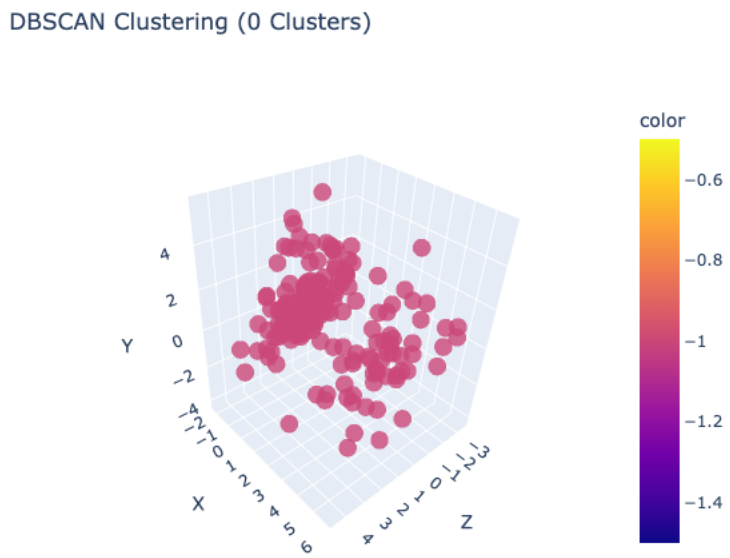
Cluster 0:
- Data Point 1 to Data Point 199

Cluster 1:
- Data Point 62 to Data Point 199

Please be aware that the information provided just lists the data points' cluster assignments; no other details or trends regarding the data or the clustering results are mentioned.

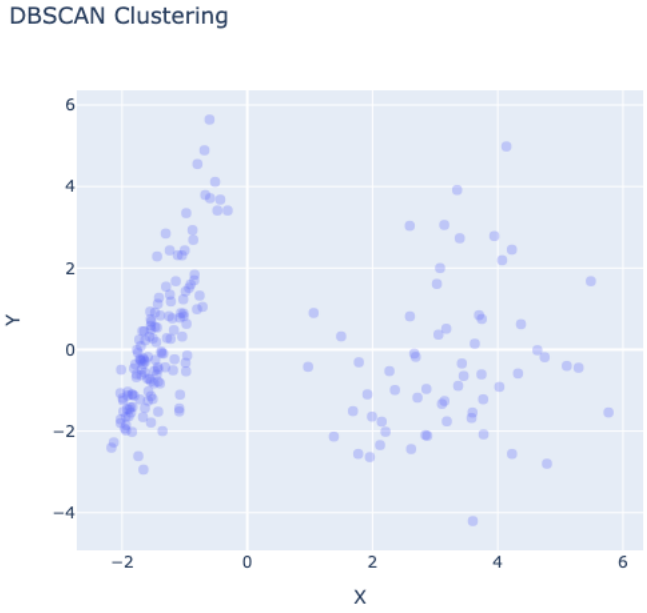
4.1.3.8. DBSCAN

Figure 81 DBSCAN for Survey



Cluster -1

199 data points make up Cluster -1, and they are represented by the indices [0, 1, 2, 3,..., 196, 197, 198]. This cluster lacks a centroid because it is a so-called outlier cluster. The data points in this cluster stand out from the other clusters and might be different from the other data in terms of certain traits or attributes.



Summary for Clustering

Cluster 0:
- Centroid coordinates: (-0.03, -0.09)
- Data points: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198]

Cluster 1:
- Centroid coordinates: (0.01, 0.04)
- Data points: [62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198]

Cluster -1:
- Centroid coordinates: Not applicable (outlier cluster)
- Data points: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198]

Based on the clustering algorithm, the data points assigned to Cluster -1 are considered outliers and don't have a centroid.

4.2. Integration For Managements

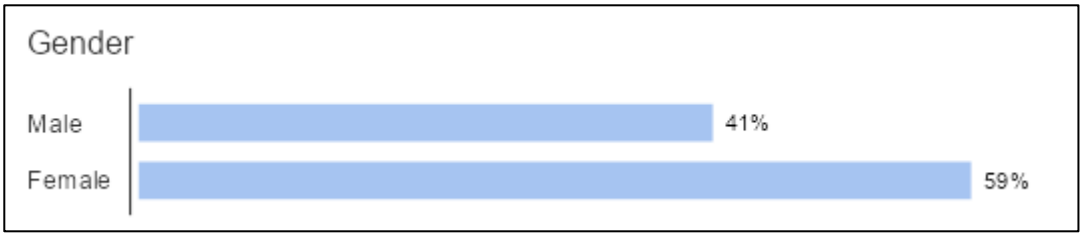


Figure 82 Gender

Our data gives us the participants age distribution as 41% **Male** and 59% **Female**. From this we can say most of the people that participated are female.

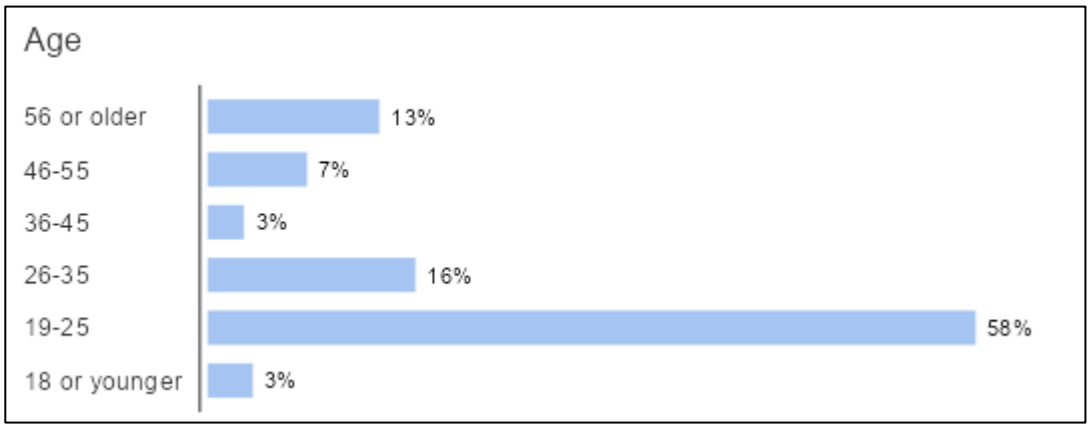


Figure 83 Age

The data we have for the age is **56 and higher** as %13, **between 46 and 55** as 7%, **between 36 and 45** as 3%, **between 26 and 35** as 16%, **between 19 and 25** as 58% and, **18 and lower** as 3%.

Figure 84 Do You Shop Online

We have good portion of people who do shop online at the moment, so it is a great support for a more accurate outcome in our analysis.

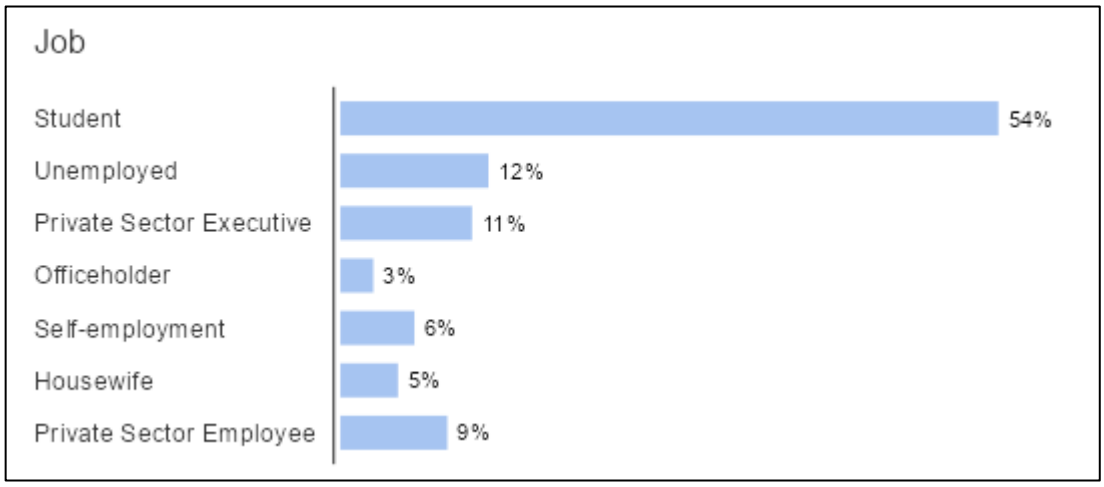


Figure 85 Job

The job status is dominantly **Student** with a 54%, **Unemployed** by 12%, **Private Sector Executive** with 11%, **Officeholder** with 3%, **Self-Employment** with 6%, **Housewife** with 5% and **Private Sector Employee** with 9%.

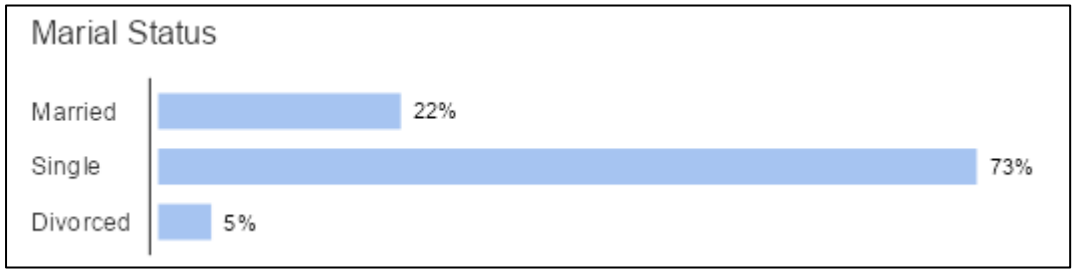


Figure 86 Marital Status

Our participant Marial Status is respectively **Married** with 22%, **Single** with 73% and **Divorced** with 5%.

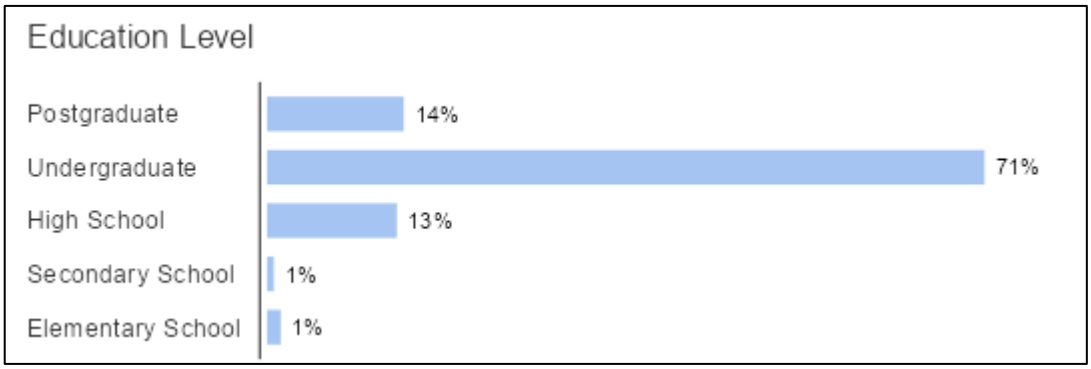


Figure 87 Education Level

Our education level comes to **Postgraduate** with 14%, **Undergraduate** with 71**%, High School** with 13%, **Secondary School** with 1% and **Elementary School** with 1%.

4.2.1. Correlation Analysis

	N	Mean	SD	Sum	Min	Max
During Covid-19, my online shopping rate increased	138	3,69565	1,28784	510	1	5
After Covid-19, I think people started to do more online shopping.	138	4,21739	1,13818	582	1	5
I trust the payment methods when shop online	138	3,53623	1,16619	488	1	5
When I'm shopping online, websites suggest items that I am looking for.	138	3,36232	1,11357	464	1	5
As of today, I am shopping online less than I did during Covid-19	136	2,49265	1,34438	339	1	5
When I register to an online shopping site for the first time, seeing better product suggestions would increase my satisfaction	138	3,85507	1,22356	532	1	5

Table 7 Correlation Analysis for Survey

This table shows us some statistical results of the 6-scale question we have.
Next steps of the analysis will show us more, from here we can see that there are high mean valued questions and low means valued questions. We will put them into the correlation analysis and try to understand the correlations.

		During Covid-19, my online shopping rate increased	After Covid-19, I think people started to do more online shopping.	I trust the payment methods when shop online	When I'm shopping online, websites suggest items that I am looking for.	As of today, I am shopping online less than I did during Covid-19	When I register to an online shopping site for the first time, seeing better product suggestions would increase my satisfaction
During Covid-19, my online shopping rate increased	Pearson Corr.	1	0,54842	0,38649	0,23524	-0,00985	0,25437
	p-value	--	<0.0001	<0.0001	0,00548	0,90937	0,00261
After Covid-19, I think people started to do more online shopping.	Pearson Corr.	0,54842	1	0,47245	0,42116	-0,031	0,41589
	p-value	<0.0001	--	<0.0001	<0.0001	0,72015	<0.0001
I trust the payment methods when shop online	Pearson Corr.	0,38649	0,47245	1	0,54627	-0,06663	0,29017
	p-value	<0.0001	<0.0001	--	<0.0001	0,44091	5,55708E-4
When I'm shopping online, websites suggest items that I am looking for.	Pearson Corr.	0,23524	0,42116	0,54627	1	0,00432	0,40847
	p-value	0,00548	<0.0001	<0.0001	--	0,96016	<0.0001
As of today, I am shopping online less than I did during Covid-19	Pearson Corr.	-0,00985	-0,031	-0,06663	0,00432	1	-0,16539
	p-value	0,90937	0,72015	0,44091	0,96016	--	0,05433
When I register to an online shopping site for the first time, seeing better product suggestions would increase my satisfaction	Pearson Corr.	0,25437	0,41589	0,29017	0,40847	-0,16539	1
	p-value	0,00261	<0.0001	5,55708E-4	<0.0001	0,05433	--

2-tailed test of significance is used

Table 8 Correlation Analysis 2 for Survey

As we can see from the table above, we put all the questions into the correlation analysis and took the Pearson correlation scores and p values. With these values we can understand if there is a correlation or not.
We will show some graphs and Pearson correlation score and try to understand the what people thinks and which answers do or do not have correlation.

Pearson Correlation = 0,54842

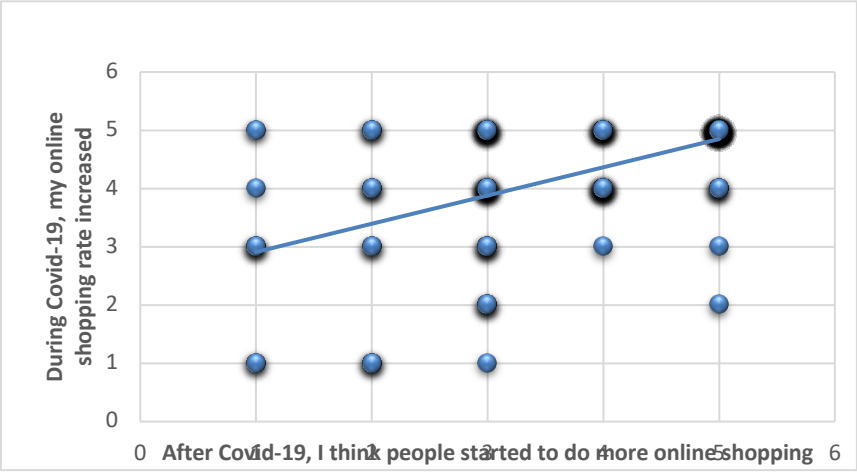


Figure 88 Shopping Rate During & After Covid

Correlation coefficient is between 0,3 – 0,7 it means that there is mid-level of correlation between these questions.
The purpose to ask these questions were to understand how people think and how they actually behave. From the outcomes we can say that most people behave the way they think others do after covid – 19.
This shows that covid – 19 significantly changed our shopping behavior from both ways.

Pearson Correlation = 0,54627

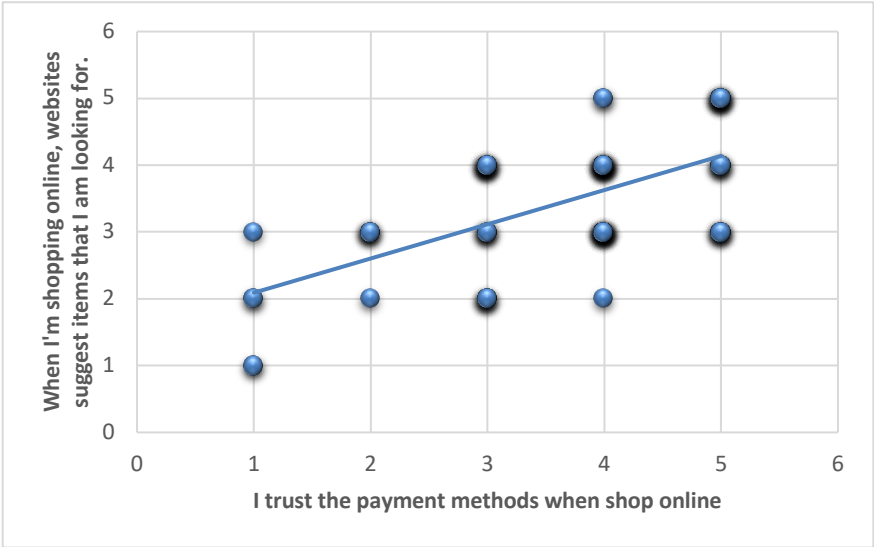


Figure 89 Website Suggest for Survey

Correlation coefficient is between 0,3 – 0,7 it means that there is mid-level of correlation between these questions.
These two answers are mid-level correlated might mean that, people who get good suggestions might not actually taking good suggestions, but they may be convincing themselves that they are getting good suggestion. If this type of behaviors is true, this question gives us a hint about why

it might be true, they might be trusting the website, payment methods etc. and that trust may have convinced them that things about the site were fine. We will try to understand this behavior in the next graph.

Pearson Correlation = 3,67477E-4

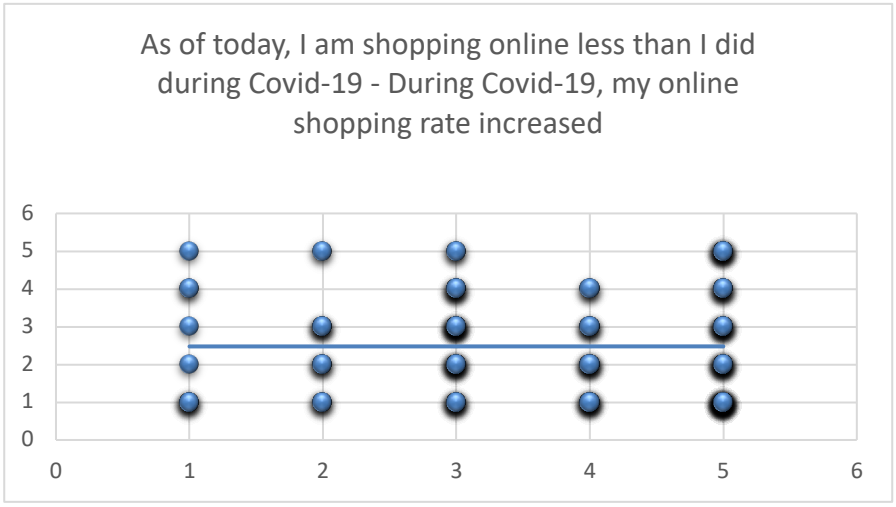


Figure 90 Shopping Online rate as of Now

Correlation Coefficient is lower than 0,3 it means there is a very low correlation.
This result shows that, people who begin to buy more things with covid–19, they did not lower their buying behavior after the covid-19. This is really important because it means that our project is on point.

Pearson Correlation = 0,38649

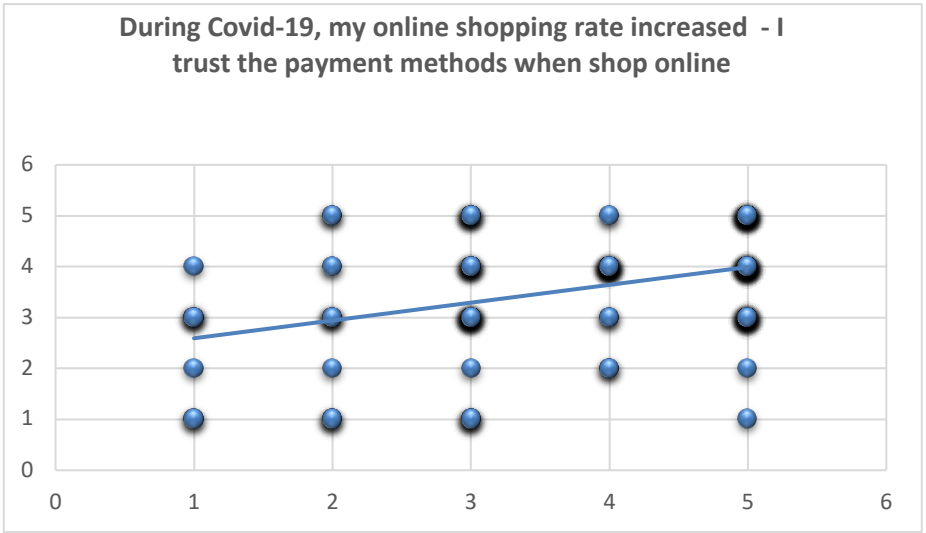


Figure 91 Trust of Payment Methods

Correlation coefficient is between 0,3 – 0,7 it means that there is mid-level of correlation between them.

This result is related to our first problem and really important for us, from here we can see that people who trusts the payment method began to use online shopping more in the covid-19 area. This shows us if a company improves its payment methods and make customers believe its trustable, then they may attract some customer. We can also see that many customers do not really trust the payment methods.

Pearson Correlation = 0,42116

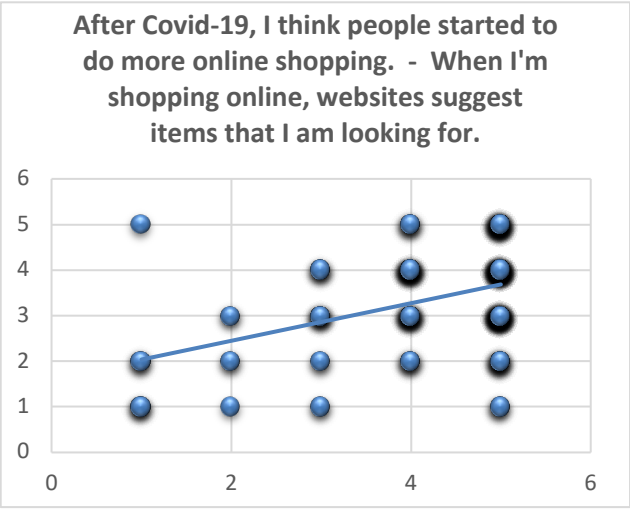


Figure 92 When I'm shopping online, websites suggest items

Pearson Correlation = 0,47245

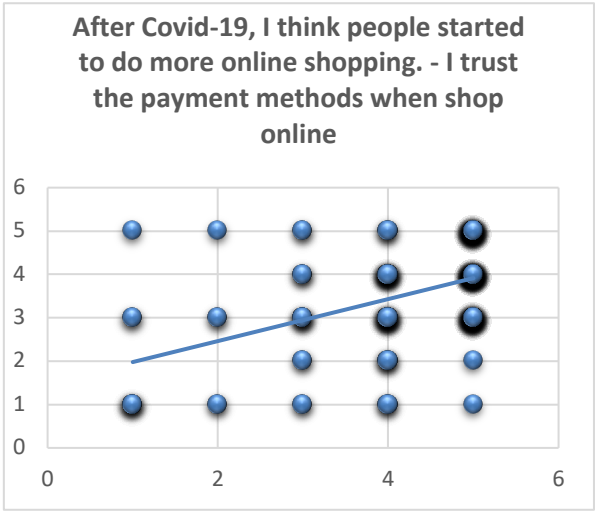


Figure 93 I trust the payment methods when shop online

Correlation coefficient is between 0,3 – 0,7, mid-level of correlation.
These graphs show us that people who trust the payment method or likes the suggestions of the website also thinks that after covid-19 people started to do more shopping. This might be showing the way people think. When they trust or like thinking style might be getting into the positive side.

Pearson Correlation = 0,40847

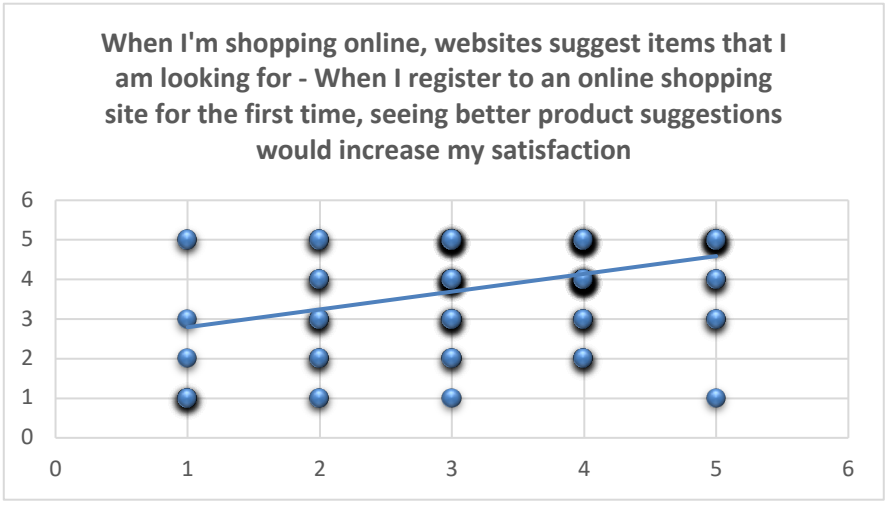


Figure 94 register to an online shopping site for the first time

Correlation coefficient is between 0,3 – 0,7, mid-level of correlation.
This graph shows that who gets good suggestion from online shopping site, also thinks that getting good suggestion at the beginning increase the satisfaction.
This is important for this project, because as we said in the problem part, we thought that if company suggest better at the beginning, they can improve the satisfaction. We will look at all the graphs and results together at the end of this paper but for now we need to say this result directly supports our idea as some other graph directly did.

4.2.2. Regression Analysis

Before beginning, we wanted to show our reliability analysis and our Cronbach's alpha for the reason being that usually below 0,6 is not acceptable or rather questionable data thus our data we are about the analyze might have a smaller sample size, but it is a good data. Regression Analysis is a broad term for analyzing the data we have collected. There are various kinds of variables we need to look at. There are the Ordinal Logistic Regression Analysis and Binomial Logistic Regression Analysis. Therefore, depending on the prediction, we want to make, we will use several types of analysis methods. The things we will take a look will include the p value of the overall model, the model coefficients, the confidence intervals with a %95, some graphs, and a likelihood chart in this model to generate an idea about our outcome. It would not be correct to just look at p values as other "tools" are also important about the data being a good fit, besides being significant and if there is a positive correlation between the outcome and the predictors. We wanted to see the correlation between if there is a relationship between increase in satisfaction with correct product suggestions, age, and the payment method trust. Our goal is to generate an ordinal logistic regression model to predict customer satisfaction based on these independent factors.

Figure 95 Overall Model Test

Overall Model Test		
χ^2	df	p
43.7	9	< .001

Our model has a p value of lower than 0.001 which indicates a strong significance with our dependent and independent variables. For Pseudo-R squared, we have used McFadden's, Cox and Snell's and Nagelkerke's R squared but because there is no "correct" option, we are showing the significance and the estimation with the Odds Ratio and Confidence Interval of %95.
Firstly, gender does not seem to play a significant role in our prediction with (p > .05). If it did, the estimation shows us that the Female represented by "Female", are more likely to be more prone to increase in satisfaction due to good suggestion.

Figure 96 Model Coefficients

Model Coefficients - Good suggestions increase my satisfaction				
Predictor	Estimate	SE	Z	p
Gender:				
Female – Male	0.315	0.328	0.960	0.337
Shopping sites show me what I want:				
2 – 1	1.948	0.803	2.425	0.015
3 – 1	2.723	0.781	3.488	< .001
4 – 1	2.641	0.790	3.342	< .001
5 – 1	2.932	0.864	3.392	< .001
I don't trust payment methods:				
2 – 1	1.767	0.858	2.060	0.039
3 – 1	1.455	0.678	2.147	0.032
4 – 1	1.439	0.683	2.108	0.035
5 – 1	1.948	0.753	2.585	0.010

The shopping site suggestion Likert chart seem to play a significant role in all divided predictors with reference level of "1" which stands for "Strongly Disagree" with positive estimate (B2= 1.94, B3=2.72, B4=2.64 and B=2.93 respectively). The more likely people agree on shopping sites showing good products and services, the more likely they are more satisfied. With peoples' trust on payment methods, people that do not trust them the most are more prone to be affected by good product suggestions with (B5= 1.94 and p value = 0.01) "Strongly Agree" being the most significant out of this group. Beside Gender predictor, all other predictors do not contain "1" between lower and upper intervals. Although if we look at the Likelihood Ratio of this model, the trust in payment methods predictors overall significance is low thus we can say that focusing on the more significant levels like "Strongly Agree" represented by "5", we can affect the outcome.

95% Confidence Interval	
Lower	Upper
0.720	2.61
1.484	35.42
3.326	72.86
3.009	68.29
3.497	105.27
1.122	33.36
1.162	16.92
1.136	16.87
1.650	32.08

Figure 97 Confidence Interval

Figure 98 Predictor

Predictor	χ^2	df	p
Gender	0.921	1	0.337
Shopping sites show me what I want	13.787	4	0.008
I don't trust payment methods	7.753	4	0.101

Figure 99 Overall Model Test 2

Overall Model Test		
χ^2	df	p
36.8	16	0.002

In this case, we have a different situation that can be talked about. In addition to our analysis being a good fit, it has a p value of 0.002 (<.05) which indicates a significant result. Our goal here was to predict who is more susceptible to change their shopping habits to the online side. For the predictor variables, we have taken gender, age, their post covid online shopping budget and their shopping categories. Our Intercept is positive so if all predictors were 0, we would have a positive outcome.

If we look at the model, we can see that, each individual predictor does not cause a significant change however, when we look at the overall model, it is affected significantly. From this model, we can make the commentary that while some predictors (people who have an online shopping budget of 2000₺ - 5000₺ for example), have a significance, the affect they have as a whole matter in predicting on thinking they can change their shopping to online. The one thing we can look at is estimate and Z values. The negative estimate tells us that there is a negative correlation is some shopping categories.

Figure 100 Model Coefficients 2

Model Coefficients - I changed my shopping to online

Predictor	Estimate	SE	Z	p
Intercept	2.0448	1.437	1.4232	0.155
I can analyze products more with online shopping:				
No – Yes	0.6860	0.439	1.5632	0.118
Gender:				
Female – Male	−0.4104	0.498	−0.8242	0.410
Age:				
46-55 – 56 or older	−2.6584	1.600	−1.6618	0.097
36-45 – 56 or older	0.5254	1.353	0.3881	0.698
26-35 – 56 or older	0.5179	0.881	0.5881	0.556
19-25 – 56 or older	0.0790	0.720	0.1097	0.913
18 or younger – 56 or older	2.1359	1.395	1.5311	0.126
Post covid budget:				
Between 2.000₺ - 5.000₺ – Between 0₺ - 2.000₺	−1.4116	0.545	−2.5894	0.010
Between 5.000₺ - 10.000₺ – Between 0₺ - 2.000₺	−0.4303	0.553	−0.7775	0.437
Between 10.000₺ - 50.000₺ – Between 0₺ - 2.000₺	−1.7093	1.139	−1.5007	0.133
Shopping Category:				
Clothing & Shoes – Petcare Products	−2.9472	1.283	−2.2962	0.022
Meal/Food order – Petcare Products	−2.5969	1.339	−1.9399	0.052
Grocery Shopping – Petcare Products	−2.6736	1.372	−1.9492	0.051
Other – Petcare Products	−1.2764	1.371	−0.9312	0.352
Cosmetics & Body Care – Petcare Products	−2.1377	1.545	−1.3841	0.166
Stationery & Office – Petcare Products	15.3432	1455.398	0.0105	0.992

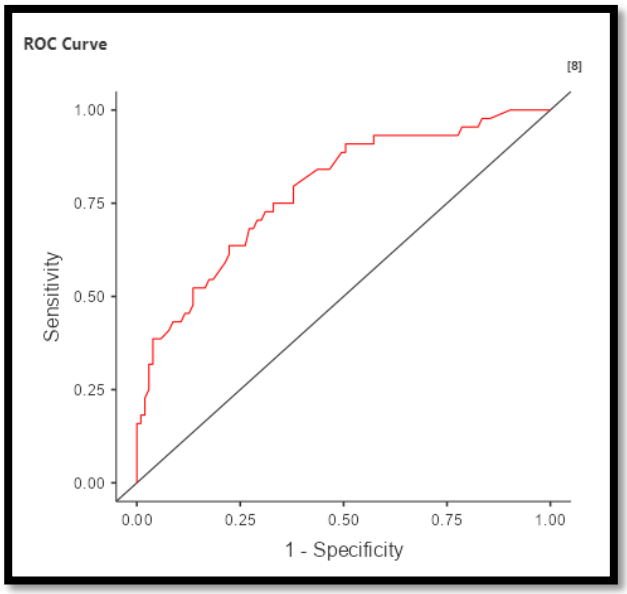


Figure 101 Roc Curve

Our ROC curve can be considered a good model in addition to our predictors which shows a curve than can be considered a good model as it curves. Our accuracy for this model is 0.782 which comes to %78,2 accuracy, meaning it is good. Our specificity is %95,1 which means our model was great at predicting the negative cases in our data and it predicts %95,1 percent of people who choose yes to change their shopping to online as “Evet” is yes and “Hayır” is no.

Figure 102 Overall Model Test 3

Overall Model Test		
χ^2	df	p
91.6	17	< .001

Figure 103 Model Coefficients 3

Model Coefficients - During covid I shopped more				
Predictor	Estimate	SE	Z	p
I think after covid other people online shopped more:				
2 – 1	2.36496	1.038	2.27735	0.023
3 – 1	0.11805	0.858	0.13766	0.891
4 – 1	1.71787	0.781	2.19915	0.028
5 – 1	2.80432	0.778	3.60641	< .001
I don't trust payment methods:				
2 – 1	2.14045	0.904	2.36676	0.018
3 – 1	1.41031	0.654	2.15519	0.031
4 – 1	2.07089	0.701	2.95254	0.003
5 – 1	2.28610	0.743	3.07802	0.002
Gender:				
Female – Male	0.64847	0.367	1.76670	0.077
Age:				
46-55 – 56 or older	0.22379	0.768	0.29145	0.771
36-45 – 56 or older	1.27499	1.323	0.96354	0.335
26-35 – 56 or older	0.32689	0.653	0.50023	0.617
19-25 – 56 or older	0.00180	0.524	0.00343	0.997
18 or younger – 56 or older	−1.11469	1.313	−0.84883	0.396
Post covid budget:				
Between 2.000₺ - 5.000₺ – Between 0₺ - 2.000₺	1.17400	0.416	2.81935	0.005
Between 5.000₺ - 10.000₺ – Between 0₺ - 2.000₺	1.81220	0.490	3.70055	< .001
Between 10.000₺ - 50.000₺ – Between 0₺ - 2.000₺	2.10607	0.766	2.74852	0.006

For this question and the model, we have a p value of < .001 which indicates a significance in addition with pseudo-R squared we have a good fit (> 0.2). The goal in this model is to ask the question whether their perception of other people's shopping behavior, their trust on online shopping payment methods, their gender, age, and their budget for predict their shopping behavior.

For the people that think other people shop online more compared to pre-covid era, we see a significant p values with positive correlation (B2= 2.36, B4=1.71, B5=2.8 respectively). At the strongly agree section represented by "5", being the most dependent variable, "4" and "2" level agreements also dependent (< .05) but not as significant can be seen.

Figure 104 Confidence Interval 2

95% Confidence Interval	
Lower	Upper
1.4385	87.87
0.2095	6.22
1.2259	27.09
3.6657	79.96
1.4738	52.32
1.1547	15.29
2.0328	32.31
2.3355	43.60
0.9329	3.95
0.2776	5.76
0.3347	89.07
0.3860	5.07
0.3546	2.80
0.0240	3.97
1.4399	7.40
2.3866	16.39
1.9617	41.34

Trust of payment methods seem the be dependent for the shopping behavior with p values smaller than 0.05. The correlation suggests that once the forcefulness of the covid pandemic making people stay at home declined, trust issues seem to affect the shopping behavior after covid pandemic ended. Gender and Age does not play a significant role in this model and on our dependent as predictors. The budget of people seems to play a significant role as middle aged budget group represented by the "between 5.000₺ - 10.000₺" translated as between "5.000₺ - 10.000₺". The other budget groups too seem to play a significant role in this model. The Odds Ratio seem to be above one (>1) beside Gender and Age predictors.

Figure 105 Predictor 2

Predictor	χ^2	df	p
I think after covid other people shopped more	28.03	4	< .001
I don't trust payment methods	11.73	4	0.019
Gender	3.13	1	0.077
Age	2.60	5	0.762
Post covid budget	18.41	3	< .001

We wanted to see how being careful about the prices after covid, is affected by the age, gender, and if they shop online less now than before covid. We wanted to see this because if the price research is affecting them, covid made an effect on shopping behavior.

Overall Model Test		
χ^2	df	p
18.3	10	0.050

Figure 106 Overall Model Test 4

As we can see from the overall model test, our p value comes to 0.05 which indicates a value that it is significant model for our dependent variable as 0.05 and 0.001 is considered a significant number.

Figure 107 Model Coefficients 4

Here we have our factors which have been put into factors because they are categorical variables. The "I shop less than during covid" part takes the base value of "1" for its calculation. From the analysis, we do not see a significant effect beside the "3" which stands for neutral, however 0.064 is not very significant even though our data set is not very big. The gender role plays a significant role on our dependent variable with the (p value of 0.013 (< .05) and B value of 0.99). Positive estimate indicates that male part of the survey population seems to more likely play a significant role in researching for prices therefore we can say gender plays a role in this dependent variable. Age again does not seem to play a significant role in this outcome.

Model Coefficients - More careful about the prices				
Predictor	Estimate	SE	Z	p
Intercept	−0.7560	0.382	−1.977	0.048
I online shop less post covid:				
2 – 1	−0.0846	0.502	−0.169	0.866
3 – 1	−1.1101	0.599	−1.854	0.064
4 – 1	−0.9986	0.732	−1.365	0.172
5 – 1	0.0693	0.617	0.112	0.911
Gender:				
Male – Female	0.9990	0.402	2.483	0.013
Age:				
56 or older – 19-25	−1.2538	0.811	−1.546	0.122
46-55 – 19-25	−0.7448	0.865	−0.861	0.389
36-45 – 19-25	−0.4411	1.237	−0.357	0.721
26-35 – 19-25	−0.5640	0.559	−1.009	0.313
18 or younger – 19-25	−0.2546	0.986	−0.258	0.796

Figure 108 Confidence Interval 3

If we look at the confidence intervals, the only one that does one include "1" between lower and upper interval is Gender. From here we can say male referenced to female population represented by "Male" is more careful about the prices when they buy anything online compared to females. At the bottom we can see also the Likelihood Ratio Test of this model. Beside the Gender, all Odds Ratios are below one (< 1) therefore their probability decreases although they do not matter as 1 fall between lower and upper bounds.

95% Confidence Interval	
Lower	Upper
0.2219	0.993
0.3438	2.456
0.1019	1.066
0.0878	1.545
0.3196	3.594
1.2344	5.974
0.0582	1.399
0.0872	2.586
0.0569	7.269
0.1902	1.702
0.1123	5.354

Figure 109 Predictor 3

Predictor	χ^2	df	p
I online shop less post covid	5.90	4	0.206
Gender	6.34	1	0.012
Age	3.96	5	0.555

Figure 110 Overall Model Test 5

Overall Model Test		
χ^2	df	p
23.2	14	0.057

In this model we talked about the trust of the people to the online payment methods by taking the trust Likert chart as dependent and taking Age, Gender, Job, Marriage Status as predictors. The Akaike Information Criterion also known as AIC in our model, while AIC does not have a single parameter, compared to our other models, this model is way is seen as very above the general AIC. Also, the p value is not significant (>.05) thus we cannot act on this model. This can be due to several reasons like our data being too small, they truly do not correlate, etc.

As we can see from the model, no predictor plays a significant role in our dependent variable (p values are bigger than .05), in addition to not being a good fit, we cannot come to a conclusion from this model. The Odds Ratios are also included 0 therefore we can conclude this model does not carry a significant enough information for our hypothesis in this capstone project.

Model Coefficients - I don't trust payment methods				
Predictor	Estimate	SE	Z	p
Age:				
46-55 – 56 or older	0.700	0.806	0.869	0.385
36-45 – 56 or older	1.903	1.059	1.797	0.072
26-35 – 56 or older	−0.524	0.789	−0.664	0.507
19-25 – 56 or older	−0.576	0.866	−0.665	0.506
18 or younger – 56 or older	−2.628	1.184	−2.220	0.026
Gender:				
Female – Male	0.121	0.341	0.355	0.723
Job:				
Unemployed – Student	−0.994	0.812	−1.223	0.221
Private Sector Executive – Student	−1.442	0.766	−1.881	0.060
Officeholder – Student	0.410	1.175	0.349	0.727
Self-employment – Student	−0.553	0.833	−0.664	0.507
Housewife – Student	0.623	0.987	0.631	0.528
Private Sector Employee – Student	−0.496	0.667	−0.744	0.457
Marial Status:				
Single – Married	1.112	0.611	1.820	0.069
Divorced – Married	0.162	0.841	0.193	0.847

Figure 111 Model Coefficients 5

Overall Model Test		
χ^2	df	p
71.4	20	< .001

Figure 112 Overall Model Test 6

Figure 113 Model Coefficients 6

Model Coefficients - I think after covid other people online shopped more				
Predictor	Estimate	SE	Z	p
Gender:				
Female – Male	0.210	0.401	0.523	0.601
Age:				
46-55 – 56 or older	−1.242	0.885	−1.404	0.160
36-45 – 56 or older	−1.929	1.124	−1.715	0.086
26-35 – 56 or older	1.041	0.969	1.075	0.282
19-25 – 56 or older	0.536	0.905	0.592	0.554
18 or younger – 56 or older	−3.473	1.616	−2.149	0.032
I don't trust payment methods:				
2 – 1	2.513	0.894	2.810	0.005
3 – 1	3.451	0.726	4.754	< .001
4 – 1	4.205	0.748	5.620	< .001
5 – 1	4.669	0.823	5.670	< .001
Kid Count:				
3-4 – 0	1.141	1.832	0.622	0.534
1-2 – 0	2.594	0.995	2.607	0.009
5 veya daha fazla – 0	−0.499	2.155	−0.231	0.817
Marial Status:				
Single – Married	1.825	0.725	2.516	0.012
Divorced – Married	0.310	1.143	0.271	0.786
How Often do you online shop:				
Everyday – Never	−4.267	2.045	−2.087	0.037
Several Times In a week – Never	−5.115	1.852	−2.762	0.006
Several Times In a month – Never	−5.417	1.828	−2.963	0.003
Several Times In a year – Never	−6.134	1.898	−3.233	0.001
Online shopping is better than physical shopping:				
No – Yes	−0.329	0.435	−0.756	0.450

In this model, we are having a look on how different predictors affect the outcome of what online shoppers think about other people's online shopping behavior. If we look at the model coefficients and the likelihood chart, we can make some comments. At first Gender (B=0.21, p value= 0.6) does not play a significant role. The age seems to play a significant role however if we look the confidence intervals, they do not cause a great effect on the outcome. The trust in payment do play a significant role and the more people do not trust payment methods, the more they think about other people's shopping behavior with a positive relation (B > 0). If we look at the Kid count and the confidence intervals, we can see only people with 1-2 kids are more significant for the outcome (B= 2.59, p value= 0.009). The Marial Status also has a significant role in this model (p < .05) and a positive correlation but only on the single people represented by "Single". So, we can say, single people are more likely to affect this outcome. How often people shop and people's perception on online shopping compared to physical shopping does not play a significant role for the reason being that while the p values are significant, if we look at the confidence interval, we can say it is not a good fit. As lower and upper intervals include one (1) between, we can make this comment on this model with significance on outcome.

95% Confidence Interval	
Lower	Upper
0.56180	2.7220
0.04872	1.6114
0.01530	1.3458
0.42537	19.7090
0.27915	10.2027
8.47e−4	0.6580
2.22711	75.7291
8.03262	140.8080
16.28828	311.3758
22.59326	579.4003
0.06857	151.1062
1.97512	100.3781
0.00721	51.3155
1.52484	26.9296
0.13999	13.8923
1.23e−4	0.5436
6.94e−5	0.1506
5.18e−5	0.1043
2.29e−5	0.0590
0.30815	1.7093

Figure 114 Confidence Interval 4

Figure 115 Predictor 4

Predictor	χ^2	df	p
Gender	0.274	1	0.601
Age	17.517	5	0.004
I don't trust payment methods	45.561	4	< .001
Kid Count	8.428	3	0.038
Marial Status	6.538	2	0.038
How Often	15.462	4	0.004
Online shopping is better than physical shopping	0.566	1	0.452

4.2.3. Limitations

There were some limitations on our project. Most of the limitations were related to survey, but there were some limitations related to analysis part as well.
First limitation we had was the number of survey participants, there were 153 participants who took our survey and we do think that this number limited us. If more people would solve our survey, we would have more precise answers after implementing data analysis methods. The main problem at there, that caused problem to us was, one ``0`` answer at the Likert Scale questions were changing the analysis drastically and because of that we found less precise answers.
Second limitation we had was We asked too little amount of Likert Scale questions. This caused us to find less outcome from correlation analysis part. Even though we found some good results, we could have found better results. Because of this limited our work.
Third and last limitation we had was At the first "problem" part of the Project we asked a question that covers wide range of people. When we try to ask questions about our first problem, there were so many little problems we needed to cover. Because of we had many little questions, we couldn't focus on one specific part. It caused us to found results that covers long range of problems but in that long range we found less than we would find if we focused on one specific problem.

4.3. Evaluation

We are pleased to introduce our cutting-edge data analysis system, a sophisticated software solution specifically designed to empower businesses in gaining comprehensive insights into their performance and customer behavior. This powerful system serves as a strategic asset for data centers, providing them with invaluable information to drive actionable decisions in alignment with their business goals. Companies can leverage this rich analysis in a multitude of ways, ranging from the strategic development of innovative products and services to the targeted training of their staff, as well as the strategic distribution of product samples to influential clientele. By meticulously analyzing sales data, organizations can unlock opportunities to enhance their operational efficiency and deliver exceptional products that meet the evolving needs of their customers.

4.2.1. Requirements

Functional Requirements

The functional requirements encompass three distinct roles: the Data Analyst, the Data Scientist, and the Machine Learning Engineer (MLE). The Data Analyst played a critical role in the system by creating a secure private account for database access, maintaining data integrity by monitoring login activities, and performing essential data management tasks such as data creation, update, and deletion. Additionally, they had the capability to track their work history for comparative analysis. The Data Scientist, on the other hand, focused on delivering comprehensive data analysis reports to company management, leveraging real-time data from the database, and safeguarding analysis files by restricting access to authorized users. Both relied on the Machine Learning System to confirm the successful transmission of their analysis. Lastly, the

MLE assumed the responsibility of optimizing models and algorithms, deployed AI processes in production environments, conducted rigorous testing of various machine learning models, and ensured smooth data flow between databases and backend systems for seamless operations. These meticulous roles collectively contribute to the technical finesse and precision of the data analysis system.

Performance Requirements

In terms of performance, the data analysis system was striving to meet several crucial requirements. Firstly, the framework was providing timely affirmation to the data analyst, responding within 3 seconds upon analysis or complaint filing. Secondly, the machine learning (ML) system was displaying pages promptly, aiming to render them in 1 second or less. Additionally, the system was presenting statistics within a maximum of 10 seconds to ensure efficient data visualization. Swift data delivery was being prioritized, with the ML system sending data to analysts within 1-2 seconds, while data processing was being completed within 10 seconds. Finally, the ML system was interpreting data swiftly, aiming to accomplish this task in 5 seconds or less.

Maintainability Requirements

To ensure the system's maintainability, several measures were being implemented. The ML system was offering flexible backup plans for data storage, including options such as daily, weekly, monthly, or yearly backups. This ensured data integrity and facilitated easy recovery in case of any mishaps. Regular updates were considered vital for system performance and security. The ML system was providing bug fixes and other updates at least once every three to six months, ensuring that the system remained up-to-date and efficient.

Usability Requirements

The system's usability was of paramount importance to analysts, and specific considerations were being made to enhance user experience. Analysts were able to easily comprehend the flow of the ML system, making it intuitive to navigate and operate. The system was designed in such a way that analysts could use it without requiring extensive guidelines or external assistance. Furthermore, multi-language support was crucial, enabling analysts to utilize the ML system in their preferred language for seamless interaction.

Security and Safety Requirements

Ensuring the security and safety of the data and the system was a top priority. The ML system was enforcing a robust password policy, requiring analysts to change their passwords every 30 days. Passwords were adhering to strong criteria, including a combination of numbers, capital letters, and small letters, to prevent unauthorized access. All data stored in the database was being encrypted, safeguarding sensitive information from unauthorized exposure. Analysts were linking their email accounts to the ML system, adding an extra layer of security to their accounts. The ML system had separate components for the Database Management System (DBMS), ensuring isolation and enhancing security. Additionally, the DBMS was performing timely data backups within 10 seconds to prevent data loss.

Availability Requirements

To ensure wide accessibility, the ML system was compatible with both Mac and Windows/PC platforms, allowing users to access the system irrespective of their operating system.

Measures of Effectiveness:

Several measures were being employed to gauge the effectiveness of the system. Customer satisfaction was a crucial metric, evaluating the satisfaction of companies utilizing the system in their operations. Stakeholder satisfaction was also being considered, reflecting the satisfaction of key stakeholders involved in the system's development and implementation. Project completion within the allocated budget and assigned timeframe was another significant measure, ensuring efficient project management. Retention rate, ranging between 40% and 70%, was providing insight into the system's ability to retain users over time. Lastly, team performance was being evaluated, ensuring that team members were capable of meeting assigned tasks and delivering satisfactory performances throughout the development and maintenance phases of the system.

4.2.2. Analysis programming Langauge

Our analysis was conducted using the Python programming language within the Jupyter Notebook IDE, which was accessed through the Anaconda distribution. Instead of TensorFlow, we utilized Plotly and Matplotlib libraries for data visualization. Our datasets were stored in .csv files for convenient access and management. To facilitate the analysis process, we made use of popular Python libraries such as Pandas, NumPy, Plotly, and Matplotlib. These libraries played a crucial role in data pre-processing, analysis, and generating visual representations of our findings. In terms of the analysis technique, we opted for the k-means clustering method, a widely-used unsupervised learning algorithm for tasks like market segmentation and search engines. This method involves classifying a dataset into a predetermined number of clusters, denoted as 'k'. Each cluster is represented as a point, and data points are assigned to the nearest cluster based on their proximity. The data points are then adjusted and computed within their respective clusters.

By leveraging these technologies and methodologies, we aimed to effectively pre-process our data, gain meaningful insights through k-means clustering, and visualize our results. Our choice of Jupyter Notebook, along with the Plotly and Matplotlib libraries, provided a versatile and efficient environment for conducting our analysis and achieving the objectives of our project**.**

4.2.3. Results

For the first dataset, the clustering method did not exhibit strong performance due to the low number of sample points within each cluster. This indicates that the points were not closely grouped together in the higher dimensional space and lacked correlation. The Feature Description method further supported this finding, showing that the features were mostly independent when analyzed individually.

Encountering such challenges is common when dealing with real-world datasets, where clusters may not be well-defined, and noise dominates a significant portion of the data. This presence of

noise can adversely affect the performance and accuracy of machine learning and deep learning models. Hence, it is crucial to address this issue before proceeding with model development. Based on these observations, it suggests that the data may not be suitable for traditional clustering techniques or that there may be underlying structures not effectively captured by the employed algorithms. In such cases, alternative techniques like dimensionality reduction or manifold learning should be considered to gain a deeper understanding of the data and potentially uncover hidden patterns. Additionally, collecting more data or refining the feature selection process could enhance the data quality and potentially yield improved clustering outcomes.

For the survey dataset, I devised a survey and actively sought to gather responses from as many individuals as possible. In total, I collected 199 data points through the survey. Although this dataset was relatively small in size, I persevered and proceeded with conducting an analysis.

However, I must acknowledge that working with a small dataset posed certain challenges. The limited number of data points within the dataset presented constraints in terms of statistical power and robustness. With a smaller sample size, it becomes more challenging to draw definitive conclusions or generalize findings to a larger population accurately.

Despite these difficulties, I strived to derive meaningful insights and conduct a comprehensive analysis of the dataset I created. I employed appropriate analytical techniques and considered the limitations associated with the data size throughout the process. While the smaller data size may limit the depth and breadth of the analysis, I made every effort to leverage the available data to its fullest potential.

When evaluating the findings from the two datasets, we found that customers are more likely to use online shopping after the pandemic than before. We can see from the dataset that customers' online spending has risen after the pandemic. Their perception of it has changed such that to many, online stores have become the primary platform for users to shop. This observation is consistent with global market trends in recent years where companies are investing far more capital in digital marketing than offline or conventional means of marketing.

5. Summary and Conclusion

5.1. Summary and Conclusion for Software Engineering

Moving forward, it is important to acknowledge that larger and more diverse datasets often yield more reliable and representative results. In future endeavors, we will aim to augment the dataset size to enhance the statistical power and increase the robustness of my analyses. Nonetheless, despite the challenges posed by the small data size, we remain committed to extracting valuable insights and drawing accurate conclusions from the data at hand.

For a business to succeed, improving client relationships is essential. A behavior analyst advises putting special emphasis on customer retention, service excellence, and communication to efficiently satisfy customer wants and prevent revenue loss. Effective marketing tactics must comprehend client behavior through data analysis. By examining historical trends, social media

effects, and cost-benefit analysis, businesses may forecast and satisfy client wants, increasing sales and boosting customer happiness.

Software businesses may make improvements to their products by discovering functional problems and receiving user feedback through customer analysis. For our project, the data scientists, data analysts, and machine learning engineers all play important roles in data analysis. Fast performance, frequent maintenance and upgrades, intuitive usability, reliable security measures, and availability across numerous platforms are essential components of an effective and user-friendly software system are our specification requirements.

Work Breakdown Structure, Responsibility Matrix, Project Network, Gantt Chart, Severity Matrix, and Risk Assessment are a few examples of project management tools that can be used to organize and understand project tasks, assign responsibilities, manage timelines, determine the severity of problems, and mitigate risks for successful project completion.

To undertake data analysis on client behavior, the software engineering team will use Python programming language within the Jupyter Notebook IDE, which was accessed through the Anaconda distribution. Instead of TensorFlow, we utilized Plotly and Matplotlib libraries for data visualization. Our datasets were stored in .csv files for convenient access and management. To facilitate the analysis process, we made use of popular Python libraries such as Pandas, NumPy, Plotly, and Matplotlib. K-means clustering, an unsupervised learning algorithm frequently used for market segmentation, was chosen as the analytical technique. The program will categorize the dataset into specified clusters and repeatedly modify the groupings until a successful outcome is obtained. Companies will be able to better understand their clients and enhance their business plans thanks to the insights gathered from the investigation.

The project's data collection approach includes gathering both first-party and second-party data. First-party data will be gathered using carefully crafted questionnaires that are directed at various age groups that engage in or have engaged in online purchasing. These surveys may be conducted face-to-face or online utilizing tools like Google Forms. The project's most important component is data analysis, where a variety of technical abilities and tools like Excel and R Project will be used to sort, classify, and analyze the data. To explore the relationships between variables, graph analysis and correlation analysis will be employed. Cluster analysis will be used to find patterns and similarities in the data.

The project's ultimate answer requires putting several distinct concepts into practice to cater to various consumer wants. This involves making 3D films available for product customization and visualization, streamlining and speeding up the payment process, and tailoring product recommendations based on information from user registration. Businesses may successfully enhance customers' online buying experiences and alter consumers' attitudes and behaviors regarding online shopping by combining these solutions.

5.2. Summary and Conclusion For Managements

The concept of online shopping, which emerged shortly after the internet came into our lives, started to gain popularity in a brief time. This concept has been gaining popularity day by day

and in today's world online shopping is more popular than ever before. But it was not that popular. In the last 10 years, the rate of online shopping has boomed and brought us to the current point. In this time period there were many reasons behind the increased rate of online shopping. There are many reasons but the part we looked at in this research was the effects of Covid-19 on online shopping.

The Covid-19 era has affected human life. This change was so big that it even affected the habit of shopping online. In order to do understand Covid-19 effects on online shopping, we prepared a survey that was prepared by scientific techniques and shared this survey with 153 different people. Thanks to the survey we conducted, we can see that the majority of people started to shop online much more during this period. We can also see that most of the people who began to use online shopping during the Covid-19 period are still using it. We can understand this part by only looking at the graphs, but it needed to be proved and showed by analytical results. To be able to do that we made some analysis, challenge the employees on a new way of thinking and shape the online shopping experience of the customers. With the help of our software engineer friends and adding the information from what we have learned in our courses as management engineers, we were able to do some analysis with various tools provided. There are variety of ways to analyze the data we have in our dataset, and we have applied multiple analysis methods in addition to our machine learning system and gathering multiple minds for the same goal. The methods and the tools we have used is based on finding the best correlation between our variables and come to a predicting conclusion about the behavior of customers based on before and after timelapse of Covid-19. The correlation analysis and logistic regression analysis methods are used for the reason being that the categorical data we have proven to be most fruitful when these methods are used on the dataset. The analysis of the dataset and the results have been shared and commented with our goal in mind. The methods we have used in R based analysis methods have shown us there are significant correlations between the shopping behavior and the time of Covid-19 time era. Not only there are single connections between two variables it is also important to include multiple factors in predicting the outcome goal. There were three main "problem" we talked about at the beginning of the project. We analyzed our survey to understand these problems.

As a result of all the analysis methods we used, we saw that the covid-19 period contributed positively to online shopping and this contribution continued after covid-19. We can see that the majority of those who started online shopping during the covid-19 period still continue to use it.

Our results show that trusting the payment method also positively affected the online shopping at the covid-19 era. When we look at the analysis part, we can see those trusting payments positive effect got bigger after covid-19, and it has little more impact on online shopping after covid-19. The difference is not big but results of after covid-19 is little, bigger in every point. This can be seen from the correlation and regression analysis part.

For the last part we can say that better product suggestions on website may increase the satisfaction. This part is not as clear as other parts. From some answers we can say that it increases, but on the other hand some answers show us that people believes that it increases other people's satisfaction more than their own. Even though we know this, our results show that having better suggestions positively affected online shopping at covid-19 era and it effects more positively in today's world. Much research is done with this goal in their mind however lacking the idea of adding the machine mind next to a human mind.

The companies can use our research findings in every aspect of their work departments and processes to find new opportunities in their industries. The decision making process for online

shopping should be open to more innovative ideas in order to optimize the enhancement of the marketing practices they have used for a long time.

REFERENCES

1. Mathworks. (2019). What Is Machine Learning? | How It Works, Techniques & Applications. Mathworks.com. https://www.mathworks.com/discovery/machine-learning.html

2. Katawetawaraks, C. & Cheng, L.W. (2011). Online shopper behavior: Influences of online shopping decision. Asian Journal of Business Research, 1(2). Available at SSRN: https://ssrn.com/abstract=2345198.

3. K. H.M, S. Duncan T, P. Ravikumar and V. E, "Online Shopping Customer BehaviourAnalysis using centrality measures," 2019 1st International Conference on Advances inInformation Technology (ICAIT), 2019, pp. 223-227, doi: 10.1109/ICAIT47043.2019.8987252.

4. Alireza Adibfar, Siddhartha Gulhare, Siva Srinivasan, Aaron Costin, Analysis and modeling of changes in online shopping behavior due to Covid-19 pandemic: A Floridacase study, Transport Policy, Volume 126, 2022, Pages 162-176, ISSN 0967-070X, https://doi.org/10.1016/j.tranpol.2022.07.003.

5. By: Rypáková, Martina; Moravčíková, Katarína; Križanová, Anna. Marketing Identity , 2015, Issue part 2, p233-246, 14p. Publisher: University of SS. Cyril & Methodius in Trnava, Slovakia., doi:https://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=119226605&lang=tr&site=eds-live

6. Muqaddas Gull and Arshi Pervaiz.(2018 April 5). Customer Behavior Analysis TowardsOnline Shopping using Data Mini https://www.researchgate.net/publication/327784914

7. Bose, I., & Mahapatra, R. K. (2001). Business data mining — a machine learning perspective. Information & Management, 39(3), 211–225. https://doi.org/10.1016/s0378-7206(01)00091-x

8. Duc, T., Quynh, Thi, H., & Dung, T. (n.d.). Prediction of Customer Behavior using Machine Learning: A Case Study. Retrieved December 13, 2022, from https://ceur- ws.org/Vol-3026/paper18.pdf

9. Zhao, J., Xue, F., Khan, S., & Khatib, S. F. A. (2021). Consumer behavior analysis forbusiness development. Aggression and Violent Behavior, 101591. https://doi.org/10.1016/j.avb.2021.101591

10. Hair, J. F., & Sarstedt, M. (2021). Data, measurement, and causal inferences in machinelearning: opportunities and challenges for marketing. Journal of Marketing Theory and Practice, 29(1), 1–13. https://doi.org/10.1080/10696679.2020.1860683

11. Spiess, J., T'Joens, Y., Dragnea, R., Spencer, P., & Philippart, L. (2014). Using big data to improve customer experience and business performance. Bell Labs Technical Journal, 18(4), 3–17. https://doi.org/10.1002/bltj.21642

12. Ganjar Alfian, Muhammad Fazal Ijaz, Muhammad Syafrudin, M. Alex Syaekhoni, Norma Latif Fitriyani, Jongtae Rhee " Customer behavior analysis using real-time dataprocessing: A case study of digital signage-based online stores " Asia Pacific Journal ofMarketing and Logistics Article publication date: 6 February 2019 https://www.emerald.com/insight/content/doi/10.1108/APJML-03-2018-0088/full/html

13. Dhandayudam, P., & Krishnamurthi, I. (2013). Customer Behavior Analysis UsingRough Set Approach. Journal of Theoretical and Applied Electronic Commerce Research, 8(2), 5–6. https://doi.org/10.4067/s0718-18762013000200003

14. Neto (Zezinho), J. A. R. (2021, September 23). Tools for Data Analysis used in Data Science, ML and Big Data. BIG DATA for EVERYONE. https://medium.com/xnewdata/tools-for-data-analysis-used-in-data-science-ml-and-big- data-87e0
15. Su-Yeon Kim, Tae-Soo Jung, Eui-Ho Suh, Hyun-Seok Hwang,

    Kim, S.-Y. et al. (2006) "Customer segmentation and strategy development based on Customer Lifetime Value: A case study," Expert Systems with Applications, 31(1), pp. 101–107. Available at: https://doi.org/10.1016/j.eswa.2005.09.004.

16. M. Gumber, A. Jain and A. L. Amutha, "Predicting Customer Behavior by Analyzing Clickstream Data," 2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP), 2021, pp. 1-6, doi: 10.1109/ICCCSP52374.2021.9465526

17. Hernández, B., Jiménez, J. and José Martín, M. (2011), "Age, gender and income: do theyreally moderate online shopping behaviour?", Online Information Review, Vol. 35 No. 1, pp. 113-133.

18. Y. Gan and D. -r. Li, "The impact of the customer satisfaction, switching costs and trust on customer relationship commitment," 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering, 2013, pp. 189-192, doi: 10.1109/ICIII.2013.6703545.

19. Nasim, S. (2018). CONSUMER BEHAVIOR TOWARDS SHOPPING MALLS: A SYSTEMATIC NARRATIVE REVIEW. IBT Journal of Business Studies, 14(1), 81–94. https://doi.org/10.46745/ilma.jbs.2018.14.01.07

20. Sahi, A. M., Khalid, H., Abbas, A. F., & Khatib, S. F. A. (2021). The Evolving Research of Customer Adoption of Digital Payment: Learning from Content and Statistical Analysis of the Literature. Journal of Open Innovation: Technology, Market, and Complexity, 7(4), 230. https://doi.org/10.3390/joitmc7040230

21. Ahmed H. Alsharif et al / "Consumer Behaviour Through Neuromarketing Approach"Journal of Contemporary Issues in Business and Government Vol. 27, No. 3, 2021 http://eprints.utm.my/id/eprint/96838/1/RohaizatBaharun2021_ConsumerBehaviou rThroughNeuromarketingApproach.pdf
22. Vinu Sundararaj,M R Rejeesh "A detailed behavioral analysis on consumer and customer changing behavior with respect to social networking sites " Journal of Retailing and Consumer Services by Elsevier January 2021 https://reader.elsevier.com/reader/sd/pii/S0969698920306238?token=6F3BC520938 CA18FC6E0E811A85CA57EFAB4FECDFFBA8EF085FD460E90AA98DA1C853D6 7FD765B0E21F60D9A58FE535B&originRegion=eu-west- 1&originCreation=20221209103247

23. Mohammed Al-Mashraie, Sung Hoon Chung,Hyun Woo Jeon "Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework:A machine learning approach" Computers & Industrial Engineering by Elsevier Date: June 2020 https://www.sciencedirect.com/science/article/abs/pii/S0360835220302102

24. https://www.techopedia.com/definition/32057/k-means-clustering

APPENDIX A

Coding for the Dataset for Customer Data

```
#!/usr/bin/env python
# coding: utf-8

# In[1]:


# Data
import numpy as np
import pandas as pd

# Data Visualization
import plotly.express as px
import plotly.graph_objs as go
import matplotlib.pyplot as plt

# Data preprocessing
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split

# Clustering Models
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score
from sklearn.metrics import calinski_harabasz_score


# In[2]:


# Specify the data path
data_path = 'Customer_Data.csv'

# Read the file
df = pd.read_csv(data_path)

# Quick look at the data
df.head()

df['Gender']
# In[3]:
```

```
df.info()


# In[4]:


df.isnull().sum()


# In[5]:


# Inpute the missing value by the Model Value
df.Profession.fillna('mode', inplace=True)

# Quick check
df.isnull().sum()


# In[6]:



# Obtain the count of each gender in the dataset
gender_count = df['Gender'].value_counts()

# Create a pie chart to visualize the distribution of gender in the dataset
fig = px.pie(values=gender_count, names=gender_count.index)

# Enhance the plot by adding a title and labels
fig.update_layout(title="Distribution of Gender in the Dataset")

# Create a bar chart to visualize the distribution of gender in the dataset
fig2 = px.bar(y=gender_count, x=gender_count.index, color=gender_count.index)

# Display the plot
fig.show()
fig2.show()


# In[7]:


# Create a box plot of Age by Gender
```

```
age_gender_boxplot = px.box(df, x='Gender', y='Age', color='Gender', title='Distribution of Age by Gender')

# Display the plot
age_gender_boxplot.show()


# In[8]:



# Create a box plot of Age by Gender
anual_income_gender_boxplot = px.box(df, x='Gender', y='Annual Income ($)', color='Gender', title='Distribution of Anual Income ($) by Gender')

# Display the plot
anual_income_gender_boxplot.show()


# In[9]:


# Create a histogram of the 'Age' column, and include the Violin plot to show the distribution
fig = px.histogram(df, x='Age', marginal='violin')

# Display the plot
fig.show()


# In[10]:


# Create violin plot for Age versus Profession
fig1 = px.violin(df, x='Age', y='Profession', color='Profession', title='Age Distribution across Professions')

# Create box plot for Age versus Profession
fig2 = px.box(df, x='Age', y='Profession', color='Profession', title='Age Distribution across Professions')

# Display the plots
fig1.show()
fig2.show()


# In[11]:
```

```
# Create a histogram of the 'Age' column, and include the Violin plot to show the distribution
fig = px.histogram(df, x='Annual Income ($)', marginal='violin')

# Display the plot
fig.show()


# In[12]:



# Create a box plot for annual income grouped by profession
fig = px.box(df, y='Annual Income ($)', x='Profession', color="Profession")

# Set the title of the plot
fig.update_layout(title_text='Annual Income Distribution by Profession')

# Show the plot
fig.show()


# In[13]:



# Create a histogram of the 'Age' column, and include the Violin plot to show the distribution
fig = px.histogram(df, x='Spending Score (1-100)', marginal='box')

# Display the plot
fig.show()


# In[14]:


# Extracting the count of each profession from the dataframe and storing in profession_dis
profession_dis = df.Profession.value_counts()

# Extracting the names of each profession from the profession_dis index
names = profession_dis.index

# Creating a pie chart to visualize the distribution of profession data values
fig = px.pie(values=profession_dis, names=names, color=names)
```

```
# Setting the title of the plot
fig.update_layout(title_text='Distribution of Profession Data Values')

# Displaying the plot
fig.show()


# In[15]:



# Create a box plot for annual income grouped by profession
fig = px.box(df, y='Work Experience', x='Profession', color="Gender")

# Set the title of the plot
fig.update_layout(title_text='Annual Income Distribution by Profession')

# Show the plot
fig.show()


# # Data Preprocessing

# In[16]:


# define the categorical columns.
categorical_columns = ['Gender', 'Profession']

# define the remarkable columns.
numerical_columns = ['Age', 'Annual Income ($)', 'Spending Score (1-100)', 'Work Experience',
'Family Size']


# In[17]:


# Convert or categorical columns to numerical columns.
for cat_col in categorical_columns:

    # Initialise label encoder.
    encoder = LabelEncoder()

    # Apply transformation.
    df[cat_col] = encoder.fit_transform(df[cat_col])
```

In[18]:

```
df.head()
```

In[19]:

```
# Rescaled data.
df.drop(columns=['CustomerID'], inplace=True)
full_data = StandardScaler().fit_transform(df)
```

In[20]:

```
# Quick Look
full_data[:5]
```

In[30]:

```
# Instantiate a PCA object with 2 components for 2D data
pca_2D = PCA(n_components=2, random_state=42)

# Fit and transform the data to obtain the 2D projection
data_2D = pca_2D.fit_transform(full_data)

# Instantiate a PCA object with 3 components for 3D data
pca_3D = PCA(n_components=3, random_state=42)

# Fit and transform the data to obtain the 3D projection
data_3D = pca_3D.fit_transform(full_data)
```

K-Means Clustering

In[22]:

```
# create a list to store the sum of squared distances for each k
```

```
ssd = []

# fit KMeans clustering with different values of k
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(full_data)
    ssd.append(kmeans.inertia_)

# create a dataframe with the k values and corresponding ssd
df = pd.DataFrame({'k': range(1, 11), 'ssd': ssd})

# create the line plot using Plotly Express
fig = px.line(df, x='k', y='ssd', title='Elbow Method')
fig.update_traces(mode='markers+lines', marker=dict(size=15))
fig.show()


# In[23]:


# create a list to store the silhouette scores for each k
silhouette_scores = []

# fit KMeans clustering with different values of k
for k in range(2, 11):
    kmeans = KMeans(n_clusters=k, random_state=0)
    kmeans.fit(full_data)
    silhouette_avg = silhouette_score(full_data, kmeans.labels_)
    silhouette_scores.append(silhouette_avg)

# find the k with the highest silhouette score
best_k = np.argmax(silhouette_scores) + 2

# plot the silhouette scores vs k
fig = px.line(x=range(2, 11), y=silhouette_scores, title='Silhouette Method')
fig.update_layout(xaxis_title='Number of Clusters (k)', yaxis_title='Silhouette Score')
fig.add_vline(x=best_k, line_dash='dash', line_color='red', annotation_text=f'Best k: {best_k}')
fig.show()


# In[24]:


# create a list to store the Calinski-Harabasz scores for each k
scores = []
```

```
# fit KMeans clustering with different values of k
for k in range(2, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(full_data)
    score = calinski_harabasz_score(full_data, kmeans.labels_)
    scores.append(score)

# create a dataframe with the k values and corresponding scores
df = pd.DataFrame({'No. of clusters(k)': range(2, 11), 'Calinski-Harabasz Score': scores})

# create the line plot using Plotly Express
fig = px.line(df, x='No. of clusters(k)', y='Calinski-Harabasz Score', title='Calinski-Harabasz
Index')
fig.update_traces(mode='markers+lines', marker=dict(size=8))
fig.show()


# In[25]:


# KMeans Clustering
kmeans = KMeans(n_clusters=2, random_state=42)

# Fit the KMeans model on train_ds
kmeans.fit(full_data)

# Obtain cluster labels and centroids
labels = kmeans.labels_
centroids = kmeans.cluster_centers_


# In[26]:


# Create the 3D scatter plot
fig = px.scatter_3d(
    x=data_3D[:, 0], y=data_3D[:, 1], z=data_3D[:, 2],
    color=labels,
    size_max=5,
    opacity=0.8,
    labels={'x':'X', 'y':'Y', 'z':'Z'},
    color_continuous_scale=['black', 'cyan'])

# Add a trace for the cluster centers
fig.add_trace(
    go.Scatter3d(
```

```
        x=centroids[:,0],
        y=centroids[:,1],
        z=centroids[:,2],
        mode='markers+text',
        text=['Centroid 1', 'Centroid 2'],
        marker=dict(
            size=10,
            color='yellow',
            opacity=0.8,
            symbol='diamond'
        )
    )
)

# Update the layout
fig.update_layout(
    coloraxis_showscale=False,
    title='K Means Clustering Visualization'
)

# Show the plot
fig.show()


# In[27]:


# Create the 2D scatter plot
fig = px.scatter(
   x=data_2D[:, 0], y=data_2D[:, 1],
   color=labels,
   size_max=5,
   opacity=0.7,
   labels={'x':'X', 'y':'Y'},
   color_continuous_scale=['black', 'cyan'])

# Add a trace for the cluster centers
fig.add_trace(
   go.Scatter(
       x=centroids[:,0],
       y=centroids[:,1],
       mode='markers+text',
       text=['Centroid 1', 'Centroid 2'],
       marker=dict(
           size=20,
           color='yellow',
```

```
        opacity=1.0,
        symbol='diamond'
      )
    )
)

# Update the layout
fig.update_layout(
  coloraxis_showscale=False,
  title='K Means Clustering Visualization'
)

# Show the plot
fig.show()


# In[28]:


from sklearn.cluster import DBSCAN

# Perform DBSCAN clustering
model = DBSCAN(eps=0.7, min_samples=5)
model.fit(full_data)

# Obtail labels
labels = model.labels_
n_clusters = len(set(labels)) - (1 if -1 in labels else 0)        # -1 stands for noise in the data i.e.
outliers

# Create the 3D scatter plot
fig = px.scatter_3d(
   x=data_3D[:, 0], y=data_3D[:, 1], z=data_3D[:, 2],
   color=labels,
   color_discrete_sequence=px.colors.qualitative.Alphabet,
   size_max=5,
   opacity=0.8,
   labels={'x':'X', 'y':'Y', 'z':'Z'},
   title=f'DBSCAN Clustering({n_clusters} Clusters)')

# Show the plot
fig.show()


# In[29]:
```

```
# Define the labels and their corresponding opacity values
label_opacity = {
    0: 1.0,       # opacity for label 0
    1: 1.0,       # opacity for label 1
    2: 1.0,       # opacity for label 2
    -1: 0.3       # opacity for label -1
}

# Create separate traces for each label with the corresponding opacity values
traces = []
for label in set(labels):
    opacity = label_opacity[label]
    mask = labels == label
    trace = go.Scatter(
        x=data_2D[mask, 0], y=data_2D[mask, 1],
        mode='markers',
        marker=dict(
            size=5*(opacity*5),
            opacity=opacity
        ),
        name=f'Label {label}'
    )
    traces.append(trace)

# Create the plot
fig = go.Figure(data=traces, layout=go.Layout(
    title='DBSCAN Clustering',
    xaxis_title='X',
    yaxis_title='Y'
))

# Show the plot
fig.show()
```

APPENDIX B

Coding for Survey Dataset

```
#!/usr/bin/env python
# coding: utf-8

# In[1]:


# Data
import numpy as np
import pandas as pd

# Data Visualization
import plotly.express as px
import plotly.graph_objs as go
import matplotlib.pyplot as plt

# Data preprocessing
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split

# Clustering Models
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score
from sklearn.metrics import calinski_harabasz_score


# In[2]:


# Specify the data path for Survey Dataset
data_path = 'Survey_Dataset.csv'

# Read the file
df = pd.read_csv(data_path)

# Quick look at the data
df.head()


# In[3]:
```

```
df['What is your gender?']

# In[4]:

df.info()

# In[5]:

df = df.drop(df.columns[[-1, -2]], axis=1)

# In[6]:

df.head()

# In[7]:

df.info()

# In[8]:

df.isnull().sum()

# In[9]:

df["Have you started to buy products online, that you were buying from store before covid-19"].fillna("NoData", inplace = True)
df["In which category do you use online shopping the most?"].fillna("NoData", inplace = True)
df["Which payment method do you usually use when shopping online?"].fillna("NoData", inplace = True)
df["Which website do you use the most, while shopping online?"].fillna("NoData", inplace = True)
df["What was your monthly online shopping budget before Covid-19?"].fillna("NoData", inplace = True)
```

```
df["What is your monthly online shopping budget after Covid-19?"].fillna("NoData", inplace =
True)
df["During Covid-19, my online shopping rate increased"].fillna("NoData", inplace = True)
df["When I'm shopping online, websites suggest items that I am looking for."].fillna("NoData",
inplace = True)
df["As of today, I am shopping online less than I did during Covid-19"].fillna("NoData", inplace
= True)
df["What is your gender?"].fillna("NoData", inplace = True)
df["What is your approximate annual income?"].fillna("1", inplace = True)


# In[10]:


df.isnull().sum()


# In[11]:


# Obtain the count of each customers who shop online in the dataset
ShopOnline = df['Do you shop online?'].value_counts()

# Create a pie chart to visualize the distribution of online shopping in the dataset
fig = px.pie(values=ShopOnline, names=ShopOnline.index)

# Enhance the plot by adding a title and labels
fig.update_layout(title="Distribution of Online Shopping in the Dataset")

# Create a bar chart to visualize the distribution of Online Shopping in the dataset
fig2 = px.bar(y=ShopOnline, x=ShopOnline.index, color=ShopOnline.index)

# Display the plot
fig.show()
fig2.show()


# In[12]:


df = df[df['Do you shop online?'] != 'No']
df.head()


# In[14]:
```

```
# Obtain the count of each customer buying online in the dataset
Products_before = df['Have you started to buy products online, that you were buying from store before covid-19'].value_counts()

# Create a pie chart to visualize the distribution of customer buying online in the dataset
fig = px.pie(values=Products_before, names=Products_before.index)

# Enhance the plot by adding a title and labels
fig.update_layout(title="Have you started to buy products online, that you were buying from store before covid-19")

# Display the plot
fig.show()

# Display the count of customers buying online in words and numbers
print("Count of customers buying online:")
for index, count in Products_before.items():
    print(f"{index}: {count}")


# In[15]:


# Obtain the count of each customers attention on product the dataset
Attention_Price = df['Have you started to pay more attention to prices and made research about it more after Covid?'].value_counts()

# Create a pie chart to visualize the distribution of customers attention on product in the dataset
fig = px.pie(values=Attention_Price, names=Attention_Price.index)

# Enhance the plot by adding a title and labels
fig.update_layout(title="Have you started to pay more attention to prices and made research about it more after Covid?")


# Display the plot
fig.show()
# Display the count of customers' attention on product in words and numbers
print("Count of customers' attention on product:")
for index, count in Attention_Price.items():
    print(f"{index}: {count}")


# In[16]:
```

```
# Obtain the count of each customers examine on product in the dataset
Examine_product = df['Do you have the chance to examine the products better thanks to online shopping?'].value_counts()

# Create a pie chart to visualize the distribution of each customers examine on product in the dataset
fig = px.pie(values=Examine_product, names=Examine_product.index)

# Enhance the plot by adding a title and labels
fig.update_layout(title="Do you have the chance to examine the products better thanks to online shopping?")


# Display the plot
fig.show()
# Display the count of each customer's examination of products in words and numbers
print("Count of customers' examination of products:")
for index, count in Examine_product.items():
    print(f"{index}: {count}")


# In[17]:


# Obtain the count of business online shopping in the dataset
Busi_Online_shop = df['Do you expect most businesses to have an online shopping system after Covid?'].value_counts()

# Create a pie chart to visualize the distribution of business online shopping in the dataset
fig = px.pie(values=Busi_Online_shop, names=Busi_Online_shop.index)

# Enhance the plot by adding a title and labels
fig.update_layout(title="Do you expect most businesses to have an online shopping system after Covid?")


# Display the plot
fig.show()
# Get the count of business online shopping in words and numbers
count = len(df)
percentage = (Busi_Online_shop / count) * 100

for index, value in Busi_Online_shop.items():
    label = f"{index}: {value} ({percentage[index]:.2f}%)"
    print(label)
```

In[18]:

```
# Obtain the count of each customers affordablity in the dataset
OS_affordable = df['Do you think online shopping is more affordable than In-Store
shopping?'].value_counts()

# Create a pie chart to visualize the distribution of customers affordablity  in the dataset
fig = px.pie(values=OS_affordable, names=OS_affordable.index)

# Enhance the plot by adding a title and labels
fig.update_layout(title="Do you think online shopping is more affordable than In-Store
shopping?")


# Display the plot
fig.show()
# Get the count of customers' affordability in words and numbers
count = len(df)
percentage = (OS_affordable / count) * 100

for index, value in OS_affordable.items():
    label = f"{index}: {value} ({percentage[index]:.2f}%)"
    print(label)
```

In[19]:

```
# Extracting the count of each online shopping by duration from the dataframe and storing in
profession_dis
often_OS = df['How often do you use online shopping?'].value_counts()
# Extracting the names of each profession from the profession_dis index
names = often_OS.index

# Creating a pie chart to visualize the distribution of online shopping by duration data values
fig = px.pie(values=often_OS, names=names, color=names)

# Setting the title of the plot
fig.update_layout(title_text='Distribution of online shopping by number of time during a period')

# Displaying the plot
fig.show()
# Getting the count of each online shopping duration in words and numbers
```

```
count = len(df)
percentage = (often_OS / count) * 100

for index, value in often_OS.items():
    label = f"{index}: {value} ({percentage[index]:.2f}%)"
    print(label)


# In[20]:


# Extracting the count of each satisfied from the dataframe and storing in profession_dis
satisfied_OS = df['How satisfied are you with your online shopping experience
overall?'].value_counts()
# Extracting the names of each satisfied from the profession_dis index
names = often_OS.index


# Create a pie chart to visualize the distribution of satisfied in the dataset
fig = px.pie(values=satisfied_OS, names=satisfied_OS.index)

# Enhance the plot by adding a title and labels
fig.update_layout(title="Distribution of satisfied customers")

# Create a bar chart to visualize the distribution of satisfied in the dataset
fig2 = px.bar(y=satisfied_OS, x=satisfied_OS.index, color=satisfied_OS.index)

# Display the plot
fig.show()
fig2.show()
# Get the count of each customer satisfaction level in words and numbers
count = len(df)
percentage = (satisfied_OS / count) * 100

for index, value in satisfied_OS.items():
    label = f"{index}: {value} ({percentage[index]:.2f}%)"
    print(label)


# In[21]:


# Create a box plot to visualize the distribution of factors by gender
age_gender_boxplot = px.box(df, x='What is your gender?', y='What are some factors that are
important to you when shopping online?', color='What is your gender?', title='Distribution of
factors by Gender')
```

```
# Display the plot
age_gender_boxplot.show()

# Group the data by gender and factors and count the occurrences
factor_gender_count = df.groupby(['What is your gender?', 'What are some factors that are
important to you when shopping online?']).size().reset_index(name='Count')

# Print the count of factors by gender
for _, row in factor_gender_count.iterrows():
    label = f"{row['What is your gender?']}: {row['What are some factors that are important to
you when shopping online?']}: {row['Count']}"
    print(label)


# In[22]:


# Create a histogram with violin plot to visualize the distribution of factors
fig = px.histogram(df, x='What are some factors that are important to you when shopping
online?', marginal='violin')

# Display the plot
fig.show()

# Get the count of factors in words and numbers
factor_count = df['What are some factors that are important to you when shopping
online?'].value_counts()

for index, value in factor_count.items():
    label = f"{index}: {value}"
    print(label)


# In[23]:


# Extracting the count of each recommend from the dataframe and storing in profession_dis
recommendation = df['How likely are you to recommend online shopping to a friend or family
member?'].value_counts()

# Create a pie chart to visualize the distribution of recommend in the dataset
fig = px.pie(values=recommendation, names=recommendation.index)

# Enhance the plot by adding a title and labels
fig.update_layout(title="Distribution of recommendation")
```

```
# Display the plot
fig.show()
# Print the count of recommendations in words and numbers
for index, value in recommendation.items():
    label = f"{index}: {value}"
    print(label)


# In[24]:


# Create a histogram with violin plot to visualize the distribution of primary reasons
fig = px.histogram(df, x='What is the primary reason for you to shop online instead of in-
person?', marginal='violin')

# Display the plot
fig.show()

# Get the count of primary reasons in words and numbers
primary_reason_count = df['What is the primary reason for you to shop online instead of in-
person?'].value_counts()

for index, value in primary_reason_count.items():
    label = f"{index}: {value}"
    print(label)


# In[25]:


# Extracting the count of each category from the dataframe and storing in category
category = df['In which category do you use online shopping the most?'].value_counts()

# Create a pie chart to visualize the distribution of category in the dataset
fig = px.pie(values=category, names=category.index)

# Enhance the plot by adding a title and labels
fig.update_layout(title="Distribution of category")

# Display the plot
fig.show()

# Get the category distribution data in words and numbers
```

```
category_data = category.reset_index().rename(columns={"index": "Category", "In which category do you use online shopping the most?": "Count"})
category_data['Percentage'] = category_data['Count'] / category_data['Count'].sum() * 100
category_data['Percentage'] = category_data['Percentage'].round(2)

# Print the category distribution data
print("Category Distribution:")
print(category_data.to_string(index=False))


# In[26]:


# Create a box plot of category by Gender
age_gender_boxplot = px.box(df, x='What is your gender?', y='In which category do you use online shopping the most?', color='What is your gender?', title='Distribution of category by Gender')

# Display the plot
age_gender_boxplot.show()

# Get the category distribution by gender data in words and numbers
category_gender_data = df.groupby(['What is your gender?', 'In which category do you use online shopping the most?']).size().reset_index(name='Count')
category_gender_data['Percentage'] = category_gender_data.groupby('What is your gender?')['Count'].apply(lambda x: x / x.sum() * 100)
category_gender_data['Percentage'] = category_gender_data['Percentage'].round(2)

# Print the category distribution by gender data
print("Category Distribution by Gender:")
print(category_gender_data.to_string(index=False))


# In[27]:


# Create a histogram of the 'payment method' column with a violin plot
fig = px.histogram(df, x='Which payment method do you usually use when shopping online?', marginal='violin')

# Display the plot
fig.show()

# Get the payment method distribution data in words and numbers
payment_method_data = df['Which payment method do you usually use when shopping online?'].value_counts().reset_index()
```

```
payment_method_data.columns = ['Payment Method', 'Count']
payment_method_data['Percentage'] = payment_method_data['Count'] /
payment_method_data['Count'].sum() * 100
payment_method_data['Percentage'] = payment_method_data['Percentage'].round(2)

# Print the payment method distribution data
print("Payment Method Distribution:")
print(payment_method_data.to_string(index=False))


# In[28]:



# Extracting the count of each website from the dataframe and storing in website
website = df['Which website do you use the most, while shopping online?'].value_counts()

# Create a pie chart to visualize the distribution of website in the dataset
fig = px.pie(values=website, names=website.index)

# Enhance the plot by adding a title and labels
fig.update_layout(title="Distribution of shopping website")

# Display the plot
fig.show()

# Get the website distribution data in words and numbers
website_data = website.reset_index().rename(columns={"index": "Website", "Which website do
you use the most, while shopping online?": "Count"})
website_data['Percentage'] = website_data['Count'] / website_data['Count'].sum() * 100
website_data['Percentage'] = website_data['Percentage'].round(2)

# Print the website distribution data
print("Website Distribution:")
print(website_data.to_string(index=False))




# In[29]:


# Extracting the count of each annual income from the dataframe and storing in annual_income
annual_income = df['What is your approximate annual income?'].value_counts()

# Create a pie chart to visualize the distribution of annual income in the dataset
```

```
fig = px.pie(values=annual_income, names=annual_income.index)

# Enhance the plot by adding a title and labels
fig.update_layout(title="Distribution of annual income")

# Display the plot
fig.show()

# Get the annual income distribution data in words and numbers
income_data = annual_income.reset_index().rename(columns={"index": "Annual Income",
"What is your approximate annual income?": "Count"})
income_data['Percentage'] = income_data['Count'] / income_data['Count'].sum() * 100
income_data['Percentage'] = income_data['Percentage'].round(2)

# Print the annual income distribution data
print("Annual Income Distribution:")
print(income_data.to_string(index=False))


# In[30]:


# Extracting the count of each budget before Covid-19 from the dataframe and storing in
Before_covid
Before_covid = df['What was your monthly online shopping budget before Covid-
19?'].value_counts()

# Create a pie chart to visualize the distribution of budget before Covid-19
fig2 = px.pie(values=Before_covid, names=Before_covid.index)

# Enhance the plot by adding a title and labels
fig2.update_layout(title="Distribution of budget Before Covid-19")

# Display the plot
fig2.show()

# Extracting the count of each budget after Covid-19 from the dataframe and storing in
After_covid
After_covid = df['What is your monthly online shopping budget after Covid-19?'].value_counts()

# Create a pie chart to visualize the distribution of budget after Covid-19
fig1 = px.pie(values=After_covid, names=After_covid.index)

# Enhance the plot by adding a title and labels
fig1.update_layout(title="Distribution of budget After Covid-19")
```

```
# Display the plot
fig1.show()

# Get the budget distribution data before Covid-19 in words and numbers
before_covid_data = Before_covid.reset_index().rename(columns={"index": "Budget", "What
was your monthly online shopping budget before Covid-19?": "Count"})
before_covid_data['Percentage'] = before_covid_data['Count'] /
before_covid_data['Count'].sum() * 100
before_covid_data['Percentage'] = before_covid_data['Percentage'].round(2)

# Print the budget distribution data before Covid-19
print("Budget Distribution Before Covid-19:")
print(before_covid_data.to_string(index=False))

# Get the budget distribution data after Covid-19 in words and numbers
after_covid_data = After_covid.reset_index().rename(columns={"index": "Budget", "What is
your monthly online shopping budget after Covid-19?": "Count"})
after_covid_data['Percentage'] = after_covid_data['Count'] / after_covid_data['Count'].sum() *
100
after_covid_data['Percentage'] = after_covid_data['Percentage'].round(2)

# Print the budget distribution data after Covid-19
print("Budget Distribution After Covid-19:")
print(after_covid_data.to_string(index=False))


# In[31]:


# Create a box plot for budget before Covid-19 grouped by profession and educational
background
fig = px.box(df, y='What was your monthly online shopping budget before Covid-19?', x='What
is your educational background?', color='What is your profession?')

# Set the title of the plot
fig.update_layout(title_text='Budget Before Covid-19')

# Show the plot
fig.show()

# Create a box plot for budget after Covid-19 grouped by profession and educational background
fig1 = px.box(df, y='What is your monthly online shopping budget after Covid-19?', x='What is
your educational background?', color='What is your profession?')

# Set the title of the plot
fig1.update_layout(title_text='Budget After Covid-19')
```

```
# Show the plot
fig1.show()

# Get the budget distribution data before Covid-19 grouped by profession and educational background in words and numbers
budget_before_covid_data = df.groupby(['What is your educational background?', 'What is your profession?'])['What was your monthly online shopping budget before Covid-19?'].describe()
budget_before_covid_data = budget_before_covid_data.reset_index().rename(columns={"count": "Count", "mean": "Mean", "std": "Standard Deviation", "min": "Minimum", "25%": "25th Percentile", "50%": "Median", "75%": "75th Percentile", "max": "Maximum"})

# Print the budget distribution data before Covid-19 grouped by profession and educational background
print("Budget Distribution Before Covid-19 Grouped by Profession and Educational Background:")
print(budget_before_covid_data.to_string(index=False))

# Get the budget distribution data after Covid-19 grouped by profession and educational background in words and numbers
budget_after_covid_data = df.groupby(['What is your educational background?', 'What is your profession?'])['What is your monthly online shopping budget after Covid-19?'].describe()
budget_after_covid_data = budget_after_covid_data.reset_index().rename(columns={"count": "Count", "mean": "Mean", "std": "Standard Deviation", "min": "Minimum", "25%": "25th Percentile", "50%": "Median", "75%": "75th Percentile", "max": "Maximum"})

# Print the budget distribution data after Covid-19 grouped by profession and educational background
print("Budget Distribution After Covid-19 Grouped by Profession and Educational Background:")
print(budget_after_covid_data.to_string(index=False))


# In[32]:


# Extracting the count of each average from the dataframe and storing in average_d
average_d = df['On average how long do you wait for your online purchase to arrive?'].value_counts()

# Create a bar chart to visualize the distribution of average in the dataset
fig2 = px.bar(y=average_d, x=average_d.index, color=average_d.index)
fig2.update_layout(title="On average how long do you wait for your online purchase to arrive?")

# Display the plot
```

```
fig2.show()

# Get the distribution of waiting time for online purchases in words and numbers
waiting_time_data = average_d.reset_index().rename(columns={"index": "Waiting Time", "On
average how long do you wait for your online purchase to arrive?": "Count"})

# Print the distribution of waiting time for online purchases
print("Distribution of Waiting Time for Online Purchases:")
print(waiting_time_data.to_string(index=False))


# In[33]:


# Extracting the count of each increased from the dataframe and storing in increased_SO
increased_SO = df['During Covid-19, my online shopping rate increased'].value_counts()

# Create a bar chart to visualize the distribution of increased in the dataset
fig2 = px.bar(y=increased_SO, x=increased_SO.index, color=increased_SO.index)
fig2.update_layout(title="During Covid-19, my online shopping rate increased")

# Display the plot
fig2.show()

# Get the distribution of increased online shopping rate during Covid-19 in words and numbers
increased_data = increased_SO.reset_index().rename(columns={"index": "Increased", "During
Covid-19, my online shopping rate increased": "Count"})

# Print the distribution of increased online shopping rate during Covid-19
print("Distribution of Increased Online Shopping Rate During Covid-19:")
print(increased_data.to_string(index=False))


# In[34]:


# Extracting the count of each "more" from the dataframe and storing in more_SO
more_SO = df['After Covid-19, I think people started to do more online
shopping.'].value_counts()

# Create a bar chart to visualize the distribution of "more" in the dataset
fig2 = px.bar(y=more_SO, x=more_SO.index, color=more_SO.index)
fig2.update_layout(title="After Covid-19, I think people started to do more online shopping")

# Display the plot
fig2.show()
```

```
# Get the distribution of people's perception regarding increased online shopping after Covid-19 in words and numbers
more_data = more_SO.reset_index().rename(columns={"index": "Perception", "After Covid-19, I think people started to do more online shopping.": "Count"})

# Print the distribution of people's perception regarding increased online shopping after Covid-19
print("Distribution of Perception: People Started to Do More Online Shopping After Covid-19:")
print(more_data.to_string(index=False))


# In[35]:


# Extracting the count of each payment from the dataframe and storing in payment_SO
payment_SO = df['I trust the payment methods when shop online'].value_counts()

# Create a bar chart to visualize the distribution of payment in the dataset
fig2 = px.bar(y=payment_SO, x=payment_SO.index, color=payment_SO.index)
fig2.update_layout(title="I trust the payment methods when shop online")

# Display the plot
fig2.show()

# Get the distribution of people's trust in online payment methods when shopping online in words and numbers
payment_data = payment_SO.reset_index().rename(columns={"index": "Trust Level", "I trust the payment methods when shop online": "Count"})

# Print the distribution of people's trust in online payment methods when shopping online
print("Distribution of Trust in Online Payment Methods when Shopping Online:")
print(payment_data.to_string(index=False))


# In[36]:


# Extracting the count of each suggestion from the dataframe and storing in suggest_SO
suggest_SO = df["When I'm shopping online, websites suggest items that I am looking for."].value_counts()

# Create a bar chart to visualize the distribution of suggestions in the dataset
fig2 = px.bar(y=suggest_SO, x=suggest_SO.index, color=suggest_SO.index)
fig2.update_layout(title="When I'm shopping online, websites suggest items that I am looking for.")
```

```
# Display the plot
fig2.show()

# Get the distribution of people's perception regarding website suggestions of items they are looking for when shopping online in words and numbers
suggest_data = suggest_SO.reset_index().rename(columns={"index": "Perception", "When I'm shopping online, websites suggest items that I am looking for.": "Count"})

# Print the distribution of people's perception regarding website suggestions when shopping online
print("Distribution of Perception: Websites Suggest Items I'm Looking For When Shopping Online:")
print(suggest_data.to_string(index=False))


# In[37]:


# Extracting the count of each "less" from the dataframe and storing in less_SO
less_SO = df['As of today, I am shopping online less than I did during Covid-19'].value_counts()

# Create a bar chart to visualize the distribution of "less" in the dataset
fig2 = px.bar(y=less_SO, x=less_SO.index, color=less_SO.index)
fig2.update_layout(title="As of today, I am shopping online less than I did during Covid-19")

# Display the plot
fig2.show()

# Get the distribution of people's shopping behavior online, indicating whether they are currently shopping less than they did during Covid-19, in words and numbers
less_data = less_SO.reset_index().rename(columns={"index": "Behavior", "As of today, I am shopping online less than I did during Covid-19": "Count"})

# Print the distribution of people's shopping behavior online
print("Distribution of Shopping Behavior: Currently Shopping Online Less Than During Covid-19:")
print(less_data.to_string(index=False))


# In[38]:


# Extracting the count of each register from the dataframe and storing in register_SO
register_SO = df['When I register to an online shopping site for the first time, seeing better product suggestions would increase my satisfaction'].value_counts()
```

```
# Create a bar chart to visualize the distribution of register in the dataset
fig2 = px.bar(y=register_SO, x=register_SO.index, color=register_SO.index)
fig2.update_layout(title="When I register to an online shopping site for the first time, seeing better product suggestions would increase my satisfaction")

# Display the plot
fig2.show()

# Get the distribution of people's perception regarding the impact of better product suggestions on their satisfaction when registering to an online shopping site for the first time, in words and numbers
register_data = register_SO.reset_index().rename(columns={"index": "Perception", "When I register to an online shopping site for the first time, seeing better product suggestions would increase my satisfaction": "Count"})

# Print the distribution of people's perception regarding the impact of better product suggestions on their satisfaction
print("Distribution of Perception: Impact of Better Product Suggestions on Satisfaction When Registering to an Online Shopping Site:")
print(register_data.to_string(index=False))


# In[39]:


# Obtain the count of each gender in the dataset
gender_count = df['What is your gender?'].value_counts()

# Print the distribution of gender in the dataset
print("Distribution of Gender in the Dataset:")
print(gender_count.to_string())

# Create a pie chart to visualize the distribution of gender in the dataset
fig = px.pie(values=gender_count, names=gender_count.index)
fig.update_layout(title="Distribution of Gender in the Dataset")
fig.show()

# Create a bar chart to visualize the distribution of gender in the dataset
fig2 = px.bar(y=gender_count, x=gender_count.index, color=gender_count.index)
fig2.update_layout(title="Distribution of Gender in the Dataset")
fig2.show()


# In[40]:
```

```
# Obtain the count of each age group in the dataset
age_count = df['Which range does your age belong to?'].value_counts()

# Print the distribution of age in the dataset
print("Distribution of Age in the Dataset:")
print(age_count.to_string())

# Create a pie chart to visualize the distribution of age in the dataset
fig = px.pie(values=age_count, names=age_count.index)
fig.update_layout(title="Distribution of Age in the Dataset")
fig.show()


# In[41]:


df["where do you live?"].fillna("Turkey", inplace=True)

# Obtain the count of each region in the dataset
region_count = df['where do you live?'].value_counts()

# Print the distribution of regions in the dataset
print("Distribution of Regions in the Dataset:")
print(region_count.to_string())

# Create a pie chart to visualize the distribution of regions in the dataset
fig = px.pie(values=region_count, names=region_count.index)
fig.update_layout(title="Distribution of Region in the Dataset")

# Display the plot
fig.show()


# In[42]:


# Obtain the count of each educational background in the dataset
educational = df['What is your educational background?'].value_counts()

# Obtain the count of each profession in the dataset
profession = df['What is your profession?'].value_counts()

# Print the distribution of educational backgrounds in the dataset
print("Distribution of Educational Backgrounds in the Dataset:")
print(educational.to_string())
```

```
# Print the distribution of professions in the dataset
print("Distribution of Professions in the Dataset:")
print(profession.to_string())

# Create a pie chart to visualize the distribution of educational backgrounds in the dataset
fig = px.pie(values=educational, names=educational.index)
fig.update_layout(title="Distribution of Educational Backgrounds in the Dataset")
fig.show()

# Create a bar chart to visualize the distribution of professions in the dataset
fig2 = px.bar(y=profession, x=profession.index, color=profession.index)
fig2.update_layout(title="Distribution of Professions in the Dataset")
fig2.show()


# # Data Preprocessing

# In[43]:


# define the categorical columns.
categorical_columns = ['Do you shop online?', 'Have you started to buy products online, that you were buying from store before covid-19', 'Have you started to pay more attention to prices and made research about it  more after Covid?', 'Do you have the chance to examine the products better thanks to online shopping?', 'Do you expect most businesses to have an online shopping system after Covid?', 'Do you think online shopping is more affordable than In-Store shopping?', 'How often do you use online shopping?','How satisfied are you with your online shopping experience overall?', 'What are some factors that are important to you when shopping online?','How likely are you to recommend online shopping to a friend or family member?','What is the primary reason for you to shop online instead of in-person?','In which category do you use online shopping the most?','Which payment method do you usually use when shopping online?','Which website do you use the most, while shopping online?','On average how long do you wait for your online purchase to arrive?','What is your gender?','Which range does your age belong to?','where do you live?','What is your educational background?','What is your profession?']

# define the remarkable columns.
numerical_columns = ['During Covid-19, my online shopping rate increased', 'After Covid-19, I think people started to do more online shopping.', 'I trust the payment methods when shop online', "When I'm shopping online, websites suggest items that I am looking for.", 'As of today, I am shopping online less than I did during Covid-19','When I register to an online shopping site for the first time, seeing better product suggestions would increase my satisfaction','What was your monthly online shopping budget before Covid-19?','What is your monthly online shopping budget after Covid-19?']
```

In[44]:

```
# Convert or categorical columns to numerical columns.
for cat_col in categorical_columns:

    # Initialise label encoder.
    encoder = LabelEncoder()

    # Apply transformation.
    df[cat_col] = encoder.fit_transform(df[cat_col])
```

In[45]:

```
df.head()
```

In[46]:

```
# Rescaled data.
df.drop(columns=['Do you shop online?'], inplace=True)
full_data = StandardScaler().fit_transform(df)
```

In[47]:

```
# Quick Look
full_data[:5]
```

In[48]:

```
# Instantiate a PCA object with 2 components for 2D data
pca_2D = PCA(n_components=2, random_state=42)

# Fit and transform the data to obtain the 2D projection
data_2D = pca_2D.fit_transform(full_data)

# Instantiate a PCA object with 3 components for 3D data
pca_3D = PCA(n_components=3, random_state=42)
```

```
# Fit and transform the data to obtain the 3D projection
data_3D = pca_3D.fit_transform(full_data)


# # K-Means Clustering

# In[49]:



# create a list to store the sum of squared distances for each k
ssd = []

# fit KMeans clustering with different values of k
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(full_data)
    ssd.append(kmeans.inertia_)

# create a dataframe with the k values and corresponding ssd
df = pd.DataFrame({'k': range(1, 11), 'ssd': ssd})

# create the line plot using Plotly Express
fig = px.line(df, x='k', y='ssd', title='Elbow Method')
fig.update_traces(mode='markers+lines', marker=dict(size=15))
fig.show()

# Find the optimal number of clusters (k)
optimal_k = df.loc[df['ssd'].idxmin(), 'k']

# Print the optimal number of clusters and corresponding ssd
print(f"The optimal number of clusters (k) is: {optimal_k}")
print(f"The corresponding sum of squared distances (ssd) is: {df['ssd'].min()}")


# In[50]:



# Create a list to store the silhouette scores for each k
silhouette_scores = []

# Fit KMeans clustering with different values of k
for k in range(2, 11):
    kmeans = KMeans(n_clusters=k, random_state=0)
```

```
    kmeans.fit(full_data)
    silhouette_avg = silhouette_score(full_data, kmeans.labels_)
    silhouette_scores.append(silhouette_avg)

# Find the k with the highest silhouette score
best_k = np.argmax(silhouette_scores) + 2

# Print the silhouette scores in words and numbers
for k, score in zip(range(2, 11), silhouette_scores):
    print(f"Silhouette score for k={k}: {score:.4f}")

# Plot the silhouette scores vs k
fig = px.line(x=range(2, 11), y=silhouette_scores, title='Silhouette Method')
fig.update_layout(xaxis_title='Number of Clusters (k)', yaxis_title='Silhouette Score')
fig.add_vline(x=best_k, line_dash='dash', line_color='red', annotation_text=f'Best k: {best_k}')
fig.show()


# In[51]:


# Create a list to store the Calinski-Harabasz scores for each k
scores = []

# Fit KMeans clustering with different values of k
for k in range(2, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(full_data)
    score = calinski_harabasz_score(full_data, kmeans.labels_)
    scores.append(score)

# Print the Calinski-Harabasz scores in words and numbers
for k, score in zip(range(2, 11), scores):
    print(f"Calinski-Harabasz score for k={k}: {score:.4f}")

# Create a dataframe with the k values and corresponding scores
df = pd.DataFrame({'No. of clusters(k)': range(2, 11), 'Calinski-Harabasz Score': scores})

# Create the line plot using Plotly Express
fig = px.line(df, x='No. of clusters(k)', y='Calinski-Harabasz Score', title='Calinski-Harabasz
Index')
fig.update_traces(mode='markers+lines', marker=dict(size=8))
fig.show()


# In[52]:
```

```
# KMeans Clustering
kmeans = KMeans(n_clusters=2, random_state=42)

# Fit the KMeans model on train_ds
kmeans.fit(full_data)

# Obtain cluster labels and centroids
labels = kmeans.labels_
centroids = kmeans.cluster_centers_


# In[53]:


# Create the 3D scatter plot
fig = px.scatter_3d(
    x=data_3D[:, 0], y=data_3D[:, 1], z=data_3D[:, 2],
    color=labels,
    size_max=5,
    opacity=0.8,
    labels={'x':'X', 'y':'Y', 'z':'Z'},
    color_continuous_scale=['black', 'cyan'])

# Add a trace for the cluster centers
fig.add_trace(
    go.Scatter3d(
        x=centroids[:,0],
        y=centroids[:,1],
        z=centroids[:,2],
        mode='markers+text',
        text=['Centroid 1', 'Centroid 2'],
        marker=dict(
            size=10,
            color='yellow',
            opacity=0.8,
            symbol='diamond'
        )
    )
)

# Update the layout
fig.update_layout(
    coloraxis_showscale=False,
    title='K Means Clustering Visualization'
```

```
)

# Show the plot
fig.show()


# In[54]:



# Create the 2D scatter plot
fig = px.scatter(
    x=data_2D[:, 0], y=data_2D[:, 1],
    color=labels,
    size_max=5,
    opacity=0.7,
    labels={'x':'X', 'y':'Y'},
    color_continuous_scale=['black', 'cyan'])

# Add a trace for the cluster centers
fig.add_trace(
    go.Scatter(
        x=centroids[:,0],
        y=centroids[:,1],
        mode='markers+text',
        text=['Centroid 1', 'Centroid 2'],
        textposition="top center",
        marker=dict(
            size=20,
            color='yellow',
            opacity=1.0,
            symbol='diamond'
        )
    )
)

# Update the layout
fig.update_layout(
    coloraxis_showscale=False,
    title='K Means Clustering Visualization'
)

# Show the plot
fig.show()

# Print the coordinates of the cluster centers
```

```
for i, centroid in enumerate(centroids):
    print(f"Coordinates of Centroid {i+1}: ({centroid[0]:.2f}, {centroid[1]:.2f})")

# Print the cluster assignment for each data point
for i, label in enumerate(labels):
    print(f"Data Point {i+1}: Cluster {label}")


# In[55]:


from sklearn.cluster import DBSCAN


# Perform DBSCAN clustering
model = DBSCAN(eps=0.7, min_samples=5)
model.fit(full_data)

# Obtain labels
labels = model.labels_
n_clusters = len(set(labels)) - (1 if -1 in labels else 0)  # -1 stands for noise in the data i.e. outliers

# Create the 3D scatter plot
fig = px.scatter_3d(
    x=data_3D[:, 0], y=data_3D[:, 1], z=data_3D[:, 2],
    color=labels,
    color_discrete_sequence=px.colors.qualitative.Alphabet,
    size_max=5,
    opacity=0.8,
    labels={'x': 'X', 'y': 'Y', 'z': 'Z'},
    title=f'DBSCAN Clustering ({n_clusters} Clusters)')

# Show the plot
fig.show()

# Get cluster information
cluster_info = {}
for i, label in enumerate(labels):
    if label in cluster_info:
        cluster_info[label].append(i)
    else:
        cluster_info[label] = [i]

# Print cluster information
for cluster, data_points in cluster_info.items():
    print(f"Cluster {cluster}: {len(data_points)} data points")
```

```
    print(data_points)


# In[56]:


# Define the labels and their corresponding opacity values
label_opacity = {
    0: 1.0,       # opacity for label 0
    1: 1.0,       # opacity for label 1
    2: 1.0,       # opacity for label 2
    -1: 0.3       # opacity for label -1
}

# Create separate traces for each label with the corresponding opacity values
traces = []
for label in set(labels):
    opacity = label_opacity[label]
    mask = labels == label
    trace = go.Scatter(
        x=data_2D[mask, 0], y=data_2D[mask, 1],
        mode='markers',
        marker=dict(
            size=5*(opacity*5),
            opacity=opacity
        ),
    )
    traces.append(trace)

# Create the plot
fig = go.Figure(data=traces, layout=go.Layout(
    title='DBSCAN Clustering',
    xaxis_title='X',
    yaxis_title='Y'
))

# Show the plot
fig.show()
```

APPENDIX C

Survey Questions

Başlıksız form

1.Email*

2. Do you shop online?

Yes

No

3. Have you started to buy products online, that you were buying from store beforecovid-19

Yes

No

4. Have you started to pay more attention to prices and made research about it more after Covid?

Yes

No

5. Do you have the chance to examine the products better thanks to online shopping?

Yes

No

6. Do you expect most businesses to have an online shopping system after Covid?

Yes

No

7. Do you think online shopping is more affordable than In-Store shopping?

Yes

No

8. How often do you use online shopping?

Everyday

Several Times In a week

Several Times In a month

Several Times In a year

Never

9. How satisfied are you with your online shopping experience overall?

Very satisfied

Somewhat satisfied

Neither satisfied nor dissatisfied

Somewhat dissatisfied

Very dissatisfied

10. What are some factors that are important to you when shopping online?

Price

Shipping speed

Product selection

Customer service

Brand reputation

User reviews

Security and privacy of personal information

11. How likely are you to recommend online shopping to a friend or family member?

Very likely

Somewhat likely

Neither likely nor unlikely
Somewhat unlikely
Not at all likely

12. What is the primary reason for you to shop online instead of in-person?

Convenience
Better deals and discounts
Greater product selection
Avoiding crowds and lines

13. In which category do you use online shopping the most?

Grocery Shopping
Meal/Food order
Clothing & Shoes
Stationery & Office
Furniture & Decor
Cosmetics & Body Care
Children's Products
Petcare Products
Technology
Other

14. Which payment method do you usually use when shopping online?

cash
credit card
installment shopping
Other

15. Which website do you use the most, while shopping online?

Trendyol
Hespiburada
N11
Amazon
Ebay
AliBaba
Other

16. What is your approximate annual income?

Under $5,000
$5,000 – $15,000
$15,000 – $30,000
$30,000 to $60,000
Over $60,000

17. What was your monthly online shopping budget before Covid-19?

Under $100
$100 - $250
$251 - $500
$501 - $2600
more than $2600

18. What is your monthly online shopping budget after Covid-19?

Under $100
$100 - $250
$251 - $500
$501 - $2600
more than $2600

19. On average how long do you wait for your online purchase to arrive?
Less than 1 day
1 to 3 days
4 to 7 days
More than 7 days
20. During Covid-19, my online shopping rate increased
Strongly Disagree
From 1 to 5
Strongly Agree
21. After Covid-19, I think people started to do more online shopping.
Strongly Disagree
From 1 to 5
Strongly Agree
22. I trust the payment methods when shop online
Strongly Disagree
From 1 to 5
Strongly Agree
23. When I'm shopping online, websites suggest items that I am looking for.
Strongly Disagree
From 1 to 5
Strongly Agree
24. As of today, I am shopping online less than I did during Covid-19
Strongly Disagree
From 1 to 5
Strongly Agree
25. **When I register to an online shopping site for the first time, seeing better product suggestions would increase my satisfaction**
Strongly Disagree
From 1 to 5
Strongly Agree
26. What is your gender?
Male
Female
27. Which range does your age belong to?
18 or younger
19-25
26-35
36-45
46-55
56 or older
28. where do you live?
Pick your country
29. What is your educational background?
Elementary School
Secondary School
High School
Undergraduate
Postgraduate
30. What is your profession?
Student

Housewife
Unemployed
Officeholder
Self-employment
Private Sector Executive
Private Sector Employee

31. What is your marital status?

Single
Married
Divorced

32. How many kids do you have?

0
1-2
3-4
5 or more