Title of the Project

E-retail factors for customer activation and retention: A case study from Indian e-commerce customers

Submitted by

Name of the Candidate: Muskan Sureka

Internship Batch Number: 34

Supervised by

Name of the SME: Khushboo Garg

Month & Year of Submission: January,2023.

Student's Declaration

I hereby declare that the Project Work with the title (in block letters) "E-retail factors for customer activation and retention: A case study from Indian e-commerce customers"
submitted by me for the project allocated to Internship batch no.: 34 by FLIPROBO TECHNOLOGIES as a part of my internship phase of my PG DIPLOMA COURSE OF DATA SCIENCE AND NEURAL NETWORKS BY DATATRAINED INSITITUE is my original work and has not been submitted earlier to any other Institution for the fulfilment of the requirement for any course of study.

I also declare that no chapter of this manuscript in whole or in part has been incorporated in this report from any earlier work done by others or by me. However, extracts of any literature which has been used for this report has been duly acknowledged providing details of such literature in the references.

Signature: Muskan Sureka

Name: Muskan Sureka

Address: Alipore Residency,
3 Burdwan Road

Place: Kolkata

Internship batch No.: 34

Date: 12/01/2023

ACKNOWLEDGEMENT

The success of the project is the result of hard work and endeavor of not only one but rather numerous people. It is the outcome of the exceptional support of many individuals.

Firstly, I take immense pleasure in thanking DataTrained Institute and Flip Robo Technologies for furnishing me with a chance to conduct a research project by making it a part of the curriculum.

I also express my sincere indebtedness and profound gratitude to my teachers: Dr. Deepika Sharma, Mr. Ravikesh Pandey and my SME: Khushboo Garg.

I would also like to express my deep sense of gratitude to my project SME Khushboo Garg for her continuous guidance and support, which has helped me tremendously to complete this project with the best of my abilities. Without her help, completing this research would not have been possible.

Also, I would like to express my sincere gratitude towards my friends and family for providing me with valuable and significant ideas throughout the course of the project.

I hence express my sincere gratitude and appreciation to all the individuals who have helped in any possible way and contributed in any manner. Their constant support and assistance has been significant and of great value.

INDEX

1. INTRODUCTION

1.1 Introduction

For any business, customers are the basis for its success and revenue and that is why companies become more aware of the importance of gaining customers' satisfaction. Customer relationship management (CRM) supports marketing by selecting target consumers and creating cost-effective relationships with them. CRM is the process of understanding customer behavior in order to support organization to improve customer acquisition, retention, and profitability. Thus, CRM systems utilize business intelligence and analytical models to identify the most profitable group of consumers and target them achieve higher customer retention rates. Those models can predict customers with high probability to churn based on analyzing customers' personal, demographic and behavioral data to provide personalized and customer-oriented marketing campaigns to gain customer satisfaction. The lifecycle of business – customer relationship includes four main stages: 1) identification; 2) attraction; 3) retention; and 4) development.

1) Customer identification/acquisition: This aims to identify profitable customers and the ones that are highly probable to join organization. Segmentation and clustering techniques can explore customers' personal and historical data to create segments/sub-groups of similar customers.

2) Customer attraction: The identified customer segments / sub-groups are analyzed to identify the common features that distinguish customers within a segment. Different marketing techniques can be used to target different customer segments such targeted advertising and/or direct marketing.

3) Customer retention: This is the main objective of CRM as retaining existing customers is at least 5 to 20 times more cost effective than acquiring new ones depending on business domains. Customer retention includes all actions taken by organization to guarantee customer loyalty and reduce customer churn. Customer churn refers to customers moving to a competitive organization or service provider. Churn can be for better quality of service, offers and/or benefits. Churn rate is an important indicator that all organizations aim to minimize. For this sake, churn prediction is an integral part of proactive customer retention plan. Churn prediction includes using data mining and predictive analytical models in predicting the customers with high likelihood to churn/defect. These models analyze personal and behavioral customer data for tailored and customer-centric retention marketing campaigns.

4) Customer development: The main objective of this phase is to increase the amount of customer transactions for more profitability. For this sake, market basket analysis, customer lifetime value, up, and cross selling techniques are used. Market basket analysis tries to analyze customers' behavior patterns to maximize the intensity of transactions. Analyzing customer lifetime value (CLTV) can help identifying the total net income expected from customer. Up and/or Cross selling include activities that increase the transactions of the associated services/products.

Customer retention and churn prediction have been increasingly investigated in many business domains, including, but not limited to, telecommunication, banking, retail and cloud services subscriptions. Different statistical and machine-learning techniques are used to address this

problem. Many attempts have been made to compare and benchmark the used techniques for churn prediction. In a comparison between (Decision trees, Logistic regression and Neural Network) models was performed. The study found that neural network perform slightly higher than the other two techniques. Another comparison between a set of models against their boosted versions are there. This study included two- layer Back-Propagation neural network (BPN), Decision Trees, SVM and Logistic Regression. The study showed that both decision trees and BPN achieved accuracy 94%, SVM comes next with 93% while Logistic Regression failed with accuracy 86%. Additionally, study showed 1-4% performance improvement in the boosted versions. In [68] the study investigated the accuracy of different models (Multi-layer perceptron (MLP) and Decision Tree (C5)). The study showed that MLP achieves accuracy of 95.51%, which outperforms C5 decision tree 89.63%.

Most of comparisons in the literature did not consider a study that covers the various categories of learning techniques. The bulk of the models applied for churn prediction fall into one of the following categories:

1) Regression analysis, 2) Decision tree–based, 3) Support Vector Machine, 4) Bayesian algorithm, 5) Instance – based learning, 6) Ensemble learning, 7) Artificial neural network, and 8) Linear Discriminant Analysis.

But first, let’s talk about customer activation and retention in more details.

1.2 What is Customer Activation and Why is it Important?

Customer activation means motivating customers to move toward the next stage of their customer life cycle faster than they would on their own.
Whether attracting a new customer, re-engaging an inactive one or creating a loyal advocate; every customer has the potential to feel *"attached"* to your brand. By strategizing your marketing efforts carefully, you will be able to activate these customers – as in making them more actionable. They may visit, purchase, recommend your brand more, and so on.

Customer activation is an act of motivating customers to completely realize the benefits of the product or services they are testing, hence increasing their overall engagement. the result could be acquiring a new customer or re-engaging an inactive one.
Activating the customer does really matter in order to stop them from ending their buyer's journey and continue being an active customer. As marketing leaders, you need to identify the factors that stopping your customers from continuing their buyer's journey. *Doing* this may influence customers' movement in their life cycle a step closer to a purchase. Sometimes, customers can move through these stages naturally, but not most of them. This is the reason why marketers keep on strategizing and utilizing customer activation to learn the series of actions that can move customers to the next level of their journey to your business.
In moving your customers through the next stage, it is important to know how to nudge your customer, when to push and when to leave them. To **nudge** a customer means to draw your customer attention and build engagement towards the next level. To **push** is intended for making the conversion. It is a definite action such as customers making purchases, registering for events, signing up for new offerings, or referring a friend. Lastly, to leave them is to avoid your customers from being overwhelmed. Being familiar with the concept of customer activation will teach you which of these three actions to take and when.

What is a customer activation strategy?

A customer activation strategy is a set of steps and actions to take in order to incite your customers to move on to the next stages of their lifecycle as efficiently as possible.

Customer activation strategies target a variety of users, including:

- Currently active customers
- Previously active customers
- Customers who have recently returned

Many times, these customers will naturally move throughout the customer journey without any help. However, a successful customer activation strategy helps move them more efficiently and in the right direction.

For many marketers, setting up a customer activation strategy is about creating an impression on their customers. Nonetheless, an impactful impression is achieved differently according to the attributes, needs and wants of each customer segment. There are three ways to incite your customers movement towards an actionable positive impression and one of the most difficult tasks is determining which move is best fit for which customer :

1. Nudge them
2. Push them
3. Leave them

Three options might sound simple, right? – but when you're dealing with hundreds, thousands, or even millions of customers with each being in a different stage of their lifecycle, with their own personal preferences, and unique experience with your brand, it's not always easy to make to pick the right path.

Therefore, creating an activation strategy is the first step towards taking the best course of actions. A well mapped out strategy, leveraging your multichannel capabilities and connecting at the right touchpoint would help you set off the movement of your customers in the right direction.

Why is customer activation important?

Besides the obvious benefit of generating more revenue as your customers return or upgrade to repeat buyers status, let's look at some of the other advantages that implementing a customer activation strategy can bring to your business.

1. Lower your cost of acquisition

Obtaining new customers costs 5x as much as re-engaging with your existing ones. Indeed, creating a customer activation strategy is a cost-effective way to bring customers back into your sales funnel. Furthermore, it allows you to move these customers through their lifecycle, whether they're active customers or dormant ones. Therefore, adopting such a strategy means that you can spend less time and money worrying about acquiring new customers and leveraging the existing potential.

2. Raise your ROI

When you retarget existing customers with a customer activation strategy, you already have some insights into their behavior and interaction with your brand. Which inevitably makes deploying targeted and personalised offers to these customers a more efficient task. Therefore resulting in higher conversion rates and, ultimately, higher return on investment compared to other marketing tactics.

3. Increase your customer lifetime value

Over time, your existing customers will likely bring more value to your brand and make more purchases compared to new prospects.

Therefore, the key is to keep your active customers engaged through an effective customer activation strategy. Moreover, enriching their customer experience will inevitably increase their lifetime value, prolong their customer lifecycle and bring your company more and more value.

How to create a customer activation strategy?

Activating and re-engaging with your customers is about more than just extracting value from the relationship customers entertain with your brand. It extends to their ability to obtain value from this relationship in return.

Fifty-eight percent of customers will pay higher prices for products and services from brands that offer exceptional customer service – but delivering that high level of value throughout their lifecycle can be challenging. Therefore, the main components of a successful customer activation strategy reside within taking the five steps below:

1. Know your customers

An impactful customer activation strategy starts with getting to know your customers. However, simply obtaining their information is no longer enough, the challenge resides in how to turn your customer data into something actionable?

The answer is quite straightforward: by creating and maintaining buyer personas for your different customer segments.

2. Create your buyer personas

Constructing a buyer persona is an approach through which you can gain exclusive insight into a particular customer segment. Indeed, each buyer persona you create should capture unique traits about the customers' segment it intends to portray (wants, needs and motivations behind a purchase). Which can then be used to engage with and activate these customers behaviors and interactions throughout their customers lifecycles.

For your customer activation strategy, you'll want to create a minimum of three buyer personas targeting the three groups we mentioned earlier: current customers, previous customers, and newly returning customers. Each of these buyer personas must be managed by your marketing team through relevant content to appeal to and engage with the buyer persona in question.

3. Identify customer online behavior

Establishing buyer personas used to rely on obtaining customers' data and identifying behavioral patterns through their interactions with your marketing efforts. However, monitoring online behavior has enhanced marketers' ability to predict their customers upcoming steps within the lifecycle through collected behavioral data . Therefore, making such insight crucial to conducting a successful activation strategy.

Indeed, identifying customers' online behavior allows marketing professionals to detect the most effective channels to reach their customers and spot the type of content they're more likely to interact with. Thus, avoiding negative interactions as well such as unsubscribing or dislike for the brand.

4. Create your customer journey map

Constructing your customer journey map starts with defining an overall goal for your actions such as “customer activation”. A customer journey map should evidently revolve around a single buyer persona for increased efficiency of your marketing activation efforts.

This essential step to every successful customer journey map comes as a second to establishing your buyer personas. Defining the buying process and customer touchpoints beforehand helps to determine which actions your customers are to take, and the channels used to do so. Therefore, it allows you to address their pain points and create a multichannel customer journey map.

5. Identify customer stage and speed

Customer’s stage and speed are two essential elements for a more in depth understanding of your customers and influencing their behavior throughout their life cycles.

Identifying the stage at which your customers are allows you to prompt their movement within the sales funnel whilst taking into consideration their buying habits and frequency. Therefore, allowing you to increase their speed of movement throughout the lifecycle to achieve the status of “advocate”. The best way to obtain such results is by creating content and engaging in marketing campaigns along the customer journey map.

5 ideas and examples of customer activation strategies

Customer activation strategies enable marketers to build and sustain better customer relationships, boost conversions and efficiently manage customer interactions with your brand throughout the lifecycle. Customer activation strategies will vary according to the stage you customer is at. Below are 5 different customer strategies, illustrating the different approaches to be adopted at different customer stages:

1. Welcoming new subscribers

An effective workflow for new subscribers sets the tone for a lasting relationship and lays the foundation for your prospects’ future interactions with your brand. Considered as some of the highest performing brand activation strategies, the opening rate for these campaigns is around 29.7% with a signficantly high average conversion rate. These customer activation scenarios are often times triggered by new subscriptions which instigate:

- A welcome campaign that re-establishes your brand’s value proposition, sets your customers expectations with regard to your content’s quality and frequency
- A promotional email if for example no interactions happened within the three weeks time frame. This could either be a coupon or a bonus piece of content exclusive to your target segment.

2. Engaging with first-time buyers

Existing customers who have successfully made their first purchase can either be the easiest or the hardest to move to the next stage of their lifecycle. Many companies mistakenly concentrate their efforts on long existing customers or on new leads. However, it’s of the utmost importance

to building a sustainable and successful customer relationship to keep the new first-buyers engaged as well.

As per the scenario below, it's very important to keep the conversation going through different interactions at different touch points of the customer lifecycle. Based on the buyer persona and preferences of the customer in question, you may map a successful customer engagement strategy through multi-channel interactions using the most adequate content.

3. Boosting previous customers

Multi-channel marketing can prove to be quite handy when it comes to boosting engagement with previous customers that are yet to reinstate a favorable purchasing behavior towards your brand.

An incentive for coming back might be necessary in such cases according to the buyer persona. However, One thing is sure: a well-thought out customer activation strategy in such cases must include different interactions through different channels providing different advantages (personalisation, interesting content, etc.) at different touch points

4. Endorse recurring customers

Customers who have repeatedly interacted with your brand are proof of a successful activation marketing strategy. However, this success can both be maintained and capitalized on through inciting these customers to become brand advocates.

As shown in the example below, interactions at this stage may include frequent conversations about updated features, on-going interesting deals, and early sneak-peeks into interesting content or releases

5. Uphold loyal customers

Achieving a larger base of loyal customers is every brand activation strategy's end goal (and every marketer's dream). Loyal customers are the best brand advocates, but by no means the least important ones to activate. Indeed, for this customer stage to be sustained in the long term and to grow in terms of adherents, considerable marketing efforts need to be mapped out according to a well thought-out customer activation strategy.

These efforts however are much different then those deployed at other stages of the customer journey.

1.3 What Is Customer Retention? Importance, Metrics & Strategies

Customer retention plays a crucial role in the success and lasting sustainability of a business. When done right, it can also increase a company's profits.

The key to a high customer retention is to determine what's causing customers to leave and then employing strategies that will build a loyal group of buyers who will buy more often and make bigger purchases. By discovering what delights and motivates your most loyal customers, you can then attempt to replicate it.

What Is Customer Retention?

Customer retention is a business's ability to keep existing customers and continue to generate revenue from them. Companies use different tactics to convert first-time buyers into repeat shoppers. In other words, customer retention allows a business to increase the profitability of an existing customer and maximize their lifetime value (LTV).

Customer Retention Explained

Think of customer retention as a process where a business aims to convince existing customers to keep purchasing their products or services. Since a customer has already made a purchase, it's different from lead generation, which is the effort involved in capturing contact information of businesses or individuals who are likely to buy a product or service.

Instead, customer retention is focused on existing customers. The goal is to increase repeat purchases by building customer loyalty through excellent customer service, product value and a distinct advantage over similar products or services.

Why Is Customer Retention Important?

Customer retention is vital in driving repeat purchases and ongoing value from your customer base. One oft-cited rule of thumb is that it costs five times as much to acquire a new customer as it does to retain an existing customer. Two of the most important factors in improving customer retention is understanding your customers' satisfaction and loyalty. Businesses also need to

understand any operations that may turn off potential and existing customers, such as slow or poor customer service or a faulty product.

Customer Acquisition vs. Customer Retention

Customer acquisition refers to the actions or processes designed to help a business gain new customer. This includes any efforts focused on finding new leads or turning prospects into paying customers.

Customer retention, on the other hand, happens after you acquire the customer. Once they make a purchase, you're trying to build loyalty and drive repeat business.

Why Do Customers Leave a Company?

Customers tend to move on for a myriad of reasons, which can include poor customer service, too much friction in the buying process and a lack of perceived value. This is why it's a good idea to map out the customer journey to know where the leaks are. It's also a best practice to solicit customer feedback and incorporate it into the company's larger plans.

When to Focus on Retention

The maturity of your business will determine whether you need to focus on customer retention. For instance, if your company just launched, you should focus on customer acquisition, as there are no existing clients to retain. The focal point at that stage should be developing strategies that will cultivate your initial customer base. That's because you're not getting any sales or customers, making it a moot point in trying to retain them. Tactics can include creating co-branded content, content produced by your business and a non-competitor, or creating a paid ad campaign.

But it doesn't mean you can plant the seeds. Actions such as engaging with consumers and making the purchase process as frictionless as possible helps a business to gain traction. As this happens, there will be more data to look at what's working—perhaps start tactics such as retention email campaigns or surveys. That way your business is working towards encouraging existing and past customers to make additional purchases.

Once you're more established, you can incorporate more customer retention tactics—making it more of a priority than customer acquisition—as you start to generate more consistent sales. As your sales grow at a steady rate and you have a decent-sized customer base, you can shift more time and attention to your customer retention efforts. At this stage, things like loyalty or referral programs make sense, as you'll have a steady (and hopefully loyal) customer base to draw from.

Benefits of Customer Retention

The main benefit of customer retention is the ability to maximize the amount of money you can extract from each customer. There are also other benefits including the following:

- **Increased profits:** Many companies generate the majority of their revenue from existing customers—61% of SMBs said this was the case, per a BIA/Kelsey report—so focusing on this part of your business should be the priority. It will not only increase your revenue, but also your business's profitability.
- **Lower costs:** Retaining an existing customer is anywhere from *5-25 times cheaper* than acquiring a new one, according to Bain & Company, so it's a much more cost-effective strategy in the long run.
- **Increased average order value (AOV):** Repeat customers tend to spend more over time while increasing their average order value. That's why just a 5% increase in retention rate can lead to profits growing 25-95%, per Bain & Company. And loyal customers are 23% more likely to buy again than others, according to a Gallup study.
- **Acquire brand ambassadors:** Word of mouth is one of the best ways to grow your business organically. The more loyal your customers, the more likely that they'll share positive experiences and recommend your company to others.

Customer Retention Statistics

Here are some more statistics that demonstrate why focusing on customer retention is vital to your business:

- Poor customer service would convince 39% of people never to use a company again and 37% to change suppliers, according to research from New Voice Media.
- Temkin Group found that 77% of customers would recommend a business to a friend after having just one positive experience.
- It takes 12 positive customer experiences to make up for one negative experience, according to Ruby Newell-Legner's "Understanding Customers."

Measuring Customer Retention and Key Metrics

Your customer retention strategies should be guided by data beyond sales numbers that can help quantify your efforts. Let's take a look at some of the key metrics you can use to determine your customer retention rate.

Attrition Rate Formula

The attrition rate is the number of customers a company lost in a specific time frame relative to its existing customer base. To calculate the attrition rate, take the number of customers your business lost by the end of a specified period and divide it by the total number of customers at the beginning of the period.

For example, if your business had 2,000 customers at the start of the quarter and 1,300 by the end, you'd apply the above formula as follows:

700 (number of customers lost) / 2,000 (total customers at the start of the quarter) **= 0.35**

This means that your attrition rate, or the percentage of customers that left during that quarter, was 35%.

Customer Retention Rate Formula

To calculate customer retention rate, determine the number of customers you acquired over a specific period. Subtract that figure from the total number of customers at the end of that period. Then you take this number and divide it by the number of customers you had at the beginning of the period.

Customer Retention Rate

(Total customers at end of period – new customers acquired) / Customers at beginning of period * 100

For example, let's say you have 2,000 customers at the start of the quarter and acquired another 400 customers but still have only 1,300 customers at the end of the quarter.

Customer Retention Rate

(1,300 – 400) / 2,000 = **.45**

This means you've kept 45% of your customers over the last three months.

Repeat Customer Rate

The repeat customer rate measures the chances an existing customer will make more than one purchase.

To calculate your repeat customer rate, take the number of customers who made more than one purchase and divide it by the total number of unique customers.

For example, if you have 4,000 customers who made more than one purchase in a month and 7,500 unique customers, your formula would look like this:

4,000 / 7,500 = **0.53**

This means you have a repeat customer rate of 53%.

Purchase Frequency

The purchase frequency formula is related to the repeat customer rate and represents the average number of orders placed by each customer. Take the same period used for your repeat customer rate (such as a month or quarter) and divide the total number of orders by the total number of unique customers.

Let's say your company had 30,000 orders within a month and 7,500 unique customers. You would apply the following formula:

30,000 / 7,500 = **4**

So, the average customer made four purchases.

Average Order Value (AOV)
AOV shows the average amount spent per purchase. Using the same period for the repeat purchase rate, divide your total annual revenue by the number of orders processed.

If your company earned $1,000,000 in revenue on 100,000 orders last year, your formula would look like this:

$$\$1{,}000{,}000 / 100{,}000 = \mathbf{10}$$

The average value of each order was $10.

Customer Retention Examples

Aside from making sure you're taking customer feedback, sending personalized email campaigns and ensuring fast response times, one of the best ways to engage and retain customers is to make them feel like they're a part of the brand. One such example would be to allow ample time to create momentum for a new product line or new features for an existing one.

Let's say you've taken customer feedback and launch a few new features that will significantly improve their user experience. What you can do is to send emails, or other forms of content marketing detailing what these features will help customers accomplish to generate buzz—both to promote an upcoming release and to show existing customers what feature they might miss. The idea is to give ample time for someone to repurchase or recommend your product or service.

Education is another great way to engage with customers. Examples include sending a series of emails that walks customers through a complex product or service after they make a purchase. Or, you can provide an opportunity for customers to schedule an online appointment where someone on your team provides a thorough consultation and answers their questions. Even a more hands-off approach like creating an online course to train new customers on a product can go a long way.

Customer Retention Rate by Industry

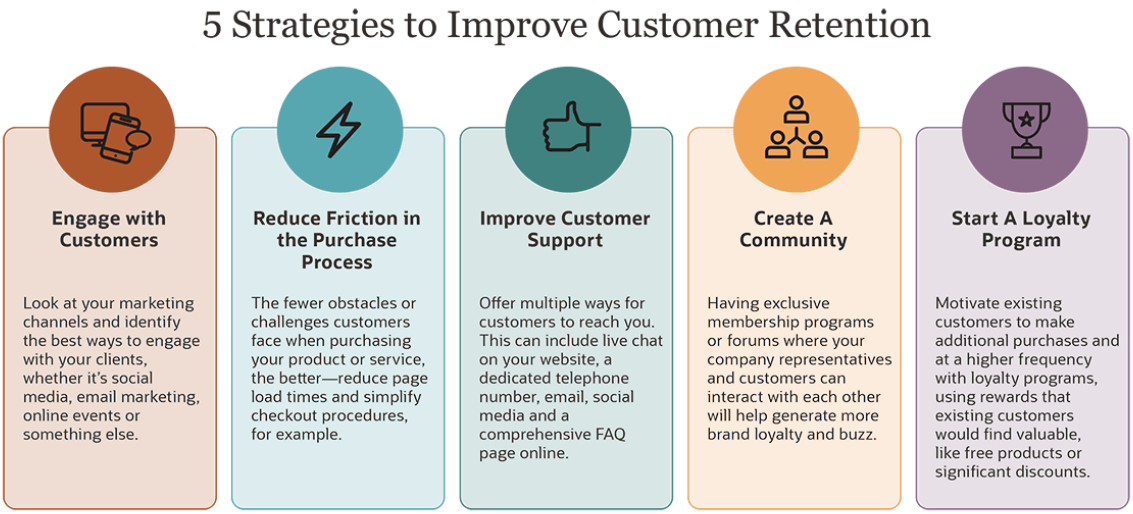
Customer retention rates vary widely by industry, and it's important for businesses to understand what's normal in their industry as they evaluate their own rate. The highest average retention rates are in the media and professional services industries, at 84%, according to the 2018 NPS® & CX Benchmarks report by CustomerGauge. The next highest rates are automotive/transportation and insurance at 83%, then IT services at 81%.

Retail and hospitality, travel and restaurants score the lowest at 63% and 55%, respectively. Clients have many options when it comes to hotels, airlines, dining and purchasing consumer goods, which might explain the lower rates in those industries. However, the report notes almost every category has businesses with a retention rate of less than 50% and higher than 95%, demonstrating that many organizations have opportunities to boost retention rate.

5 Strategies to Improve Customer Retention

Here are some practical ideas for improving customer retention:

1. **Engage with customers:** Look at your marketing channels and identify the best ways to engage with your clients. Do they respond best to social media, email marketing, online events or something else? Let customers weigh in on upcoming products and services, so they feel like they're part of the brand.
2. **Reduce friction in the purchase process:** The fewer obstacles or challenges customers face when purchasing your product or service, the better. When it comes to ecommerce, fast page load times and a fast, simple checkout experience is critical. In a store, eliminate friction by making sure a staff member is always available to help a customer when they're ready to check out.
3. **Improve customer support:** Offer multiple ways for customers to reach you. This can include live chat on your website, a dedicated telephone number, email, social media and a comprehensive FAQ page online. Additionally, you want to ensure fast response rates. Training your staff well and measuring their performance with benchmarks will help you meet customer expectations for communication.
4. **Create a community:** Having exclusive membership programs or forums where your company representatives and customers can interact with each other will help generate more brand loyalty and buzz. Other ideas include giving discount codes to loyal customers and creating referral programs that offer current clients an incentive.
5. **Start a loyalty program:** Loyalty programs can be a great way to motivate existing customers to make additional purchases and at a higher frequency. Ensure that your loyalty program has rewards that existing customers would find valuable, like free products or significant discounts.



How Can a CRM System Help With Customer Retention?

A customer relationship management (CRM) system gives businesses a central view of its customers, with contact information, order history, previous communication and their response to different marketing campaigns or promotions. It offers valuable insights into their behavior and the most effective ways to reach these customers so your company can target them with personalized offers. A CRM system also helps you see where customers are dropping off in the customer acquisition or retention process.

Choosing the Right CRM Solution

Leading CRM software has features such as lead scoring and analytics, empowering you to make better decisions that will increase conversion and customer lifetime value. Additionally, having all this information in one place improves customer service, which in turn helps convert more one-time buyers into repeat customers.

For just about any business, customer retention is an important metric to pay attention to and work to improve, especially as your organization grows. Offering an exceptional product or service is the first step in retaining customers, and once a business has that, it should look for ways to cultivate loyalty and identify new sales opportunities within its client base.

1.4 RESEARCH OBJECTIVES

This research paper has the following objectives:

- Identifying the impact of customer activation and retention of e-commerce Indian customers.
- Factors affecting their buyer behaviour.
- Predicting how many customers will recommend the e-commerce site to their friends/family. This is our target variable as this shows how much retention is the company gaining with customer's word of mouth.

1.5 RESEARCH METHODOLOGY

❖ RESEARCH DESIGN

The research design is the conceptual framework around which the survey is undertaken. Here a part of the research undertaken is a **Exploratory Research** as it is describing the perception of the respondents.

❖ SAMPLE SIZE

The Sample Size is **256**. Data has been collected and analysed on the basis of the responses. The research was conducted with an aim of getting respondents from all across India.

❖ PERIOD OF STUDY

The period of study was more than **a week.**

❖ DATA COLLECTION

1. **Primary Data-** Dataset and excel sheet provided by Flip Robo technologies for project completion.
2. **Secondary Data-**The data includes reference from previously published research papers, journals, books and articles.

❖ METHOD OF ANALYSIS

In this research, tables and various types of graphs such as bar graph and count plot have been used. This helped to present the data in a meaningful way and making it easily understandable.

1.6 RESEARCH LIMITATIONS

- **Non Representative Sample**: The research project was based on the survey conducted of only 256 respondents. Hence, such sample size cannot be said to be genuine and actual representative of the people.

- **Shortage of Time**: The time frame of the study was restricted and limited. It therefore becomes difficult to have detailed study on project work. The period was not enough for the proper study and investigation on the project.

- **Use of only Virtual Research Methods**: The survey was based on the data collected by questionnaire which was circulated virtually having limited questions which in turn resulted in collection of insufficient data due to which there was inadequacy in the research.

- **Lack of Scientific Method**: The absence of use of scientific and logical training in research approach turned out to be a great hindrance in the exploration programme.

2. STEPS USED IN PREDICTING CUSTOMER CHURN:

2.1 Choosing the model

We will use Linear Regression model for this dataset as, we need to predict the customer retention which is a continuous data. Linear regression is a machine learning algorithm which estimates how a model is following a linear relationship between one response variable (denoted by y) and one or more explanatory variables (denoted by X1, X2, X3…. Xn). The response variable will dependent on how the explanatory variables changes and not the other way round. Response variable is also known as target or dependent variable while the explanatory variable is known as independent or predictor variables.

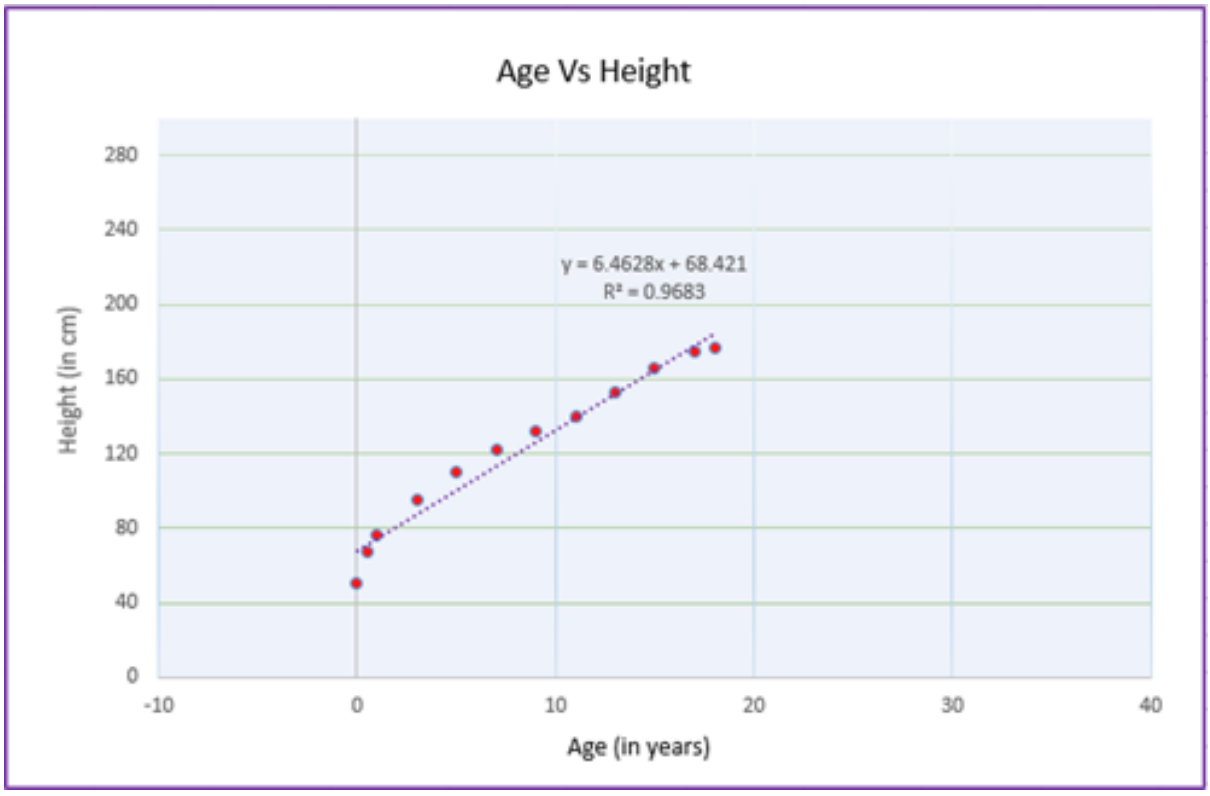


Fig: Simple Linear Regression between Height and age

There are 2 types of linear regression:

1. Simple Linear Regression

2. Multiple Linear Regression

Simple Linear Regression*: It is a type of linear regression model where there is only independent or explanatory variable. For e.g., the above scatter plot follows a simple linear regression with age being an independent variable is responsible for any change in height (dependent variable).*

Multiple Linear Regression*: It is similar to simple linear regression but here we have more than one independent or explanatory variable.*

Linear Regression can be written mathematically as follows:

$$\mathbf{Y = \beta 0 + \beta 1.X1 + \beta 2.\ X2 + \beta 3.\ X3 + \beta 4.\ X4 + \beta 5.\ X5 + \beta 5.\ X6 + \epsilon}$$

$$\mathbf{charges = \beta 0 + \beta 1.bmi + \beta 2.age + \beta 3.sex + \beta 4\ .children + \beta 5.region + \beta 5.smoker + \epsilon}$$

charges= response variable, generally denoted by Y

bmi, age, sex, children, region, smoker=Predictor variables, denoted by X1, X2, X3 and X4 respectively

β0 = Y-intercept (always a constant)

β1, β2, β3, β4, β5 = regression coefficients

ϵ = Error terms (Residuals)

C**omponents of Linear Regression:**

1. Regression Coefficient (or β1):

The Regression Coefficient in the above equation talks about the change in the value of dependent variable corresponding to the unit change in the independent variable. So, for e.g. if X1 increases or decreases by one unit, then Y will increase or decrease by β1 units. An important assumption followed by an ideal linear regression is that any increase or decrease in

one independent variable will not have any corresponding changes in other independent variables.

2. Intercept (or β0):

Intercept is a constant value which tells us at what point in the x-y coordinate graph, should the regression line must start if it follows a linear regression. Since it is a constant value, hence it is not dependent on any change in independent variables. Even if the values of X=0, intercept will have a constant value. If the value of intercept is 0, that means, the line will start at the origin point (0,0).

3. Error Terms or Residuals (ϵ):

It is the difference between the actual and the predicted data point in the x-y coordinate graph

Objective of Linear Regression:

The goal of linear regression is to perform predictive analytics and it is done by making the machine learn the science of generating a trained (best fitted) line that will very well generalize how new and unknown data (test set or new dataset) will be evaluated, and how the fitted line will be able to accurately estimate new or unknown datasets.

S**teps to be followed in Linear Regression Algorithm:**

1. Reading and understanding the data

a. Importing required libraries like pandas & numpy for data analysis and manipulation and seaborn & matplotlib for data visualization

b. Cleaning and manipulating data to make it up to the standards that exploratory data analysis can be performed by treating null values if any, updating to necessary formats, changing data types if needed, removing unwanted rows or columns etc. The raw data in whatever condition you get must be squeaky cleaned of any muck before assessing it for visualization.

2. Visualizing the data (Exploratory Data Analysis)

a. Visualizing numerical variables using scatter or pairplots in order to interpret business /domain inferences.

b. Visualizing categorical variables using barplots or boxplots in order to interpret business/domain inferences.

3. Data Preparation

a. Converting categorical variables with varying degrees of levels into dummy variables (numerical in nature) so that these variables can be represented during model building in order to contribute to the best fitted line for the purpose of better prediction.

4. Splitting the data into training and test sets

a. Splitting the data into two sections in order to train a subset of dataset to generate a trained (fitted) line that will very well generalize how new and unknown data (test set or new dataset) will be evaluated, and how the fitted line will be able to accurately estimate new or unknown datasets. Generally, the train-test split ratio is 70:30 or 80:20.

b. Rescaling the trained model: It is a method used to normalize the range of numerical variables with varying degrees of magnitude. For e.g. height or bmi or age are of different magnitude and units or some feature may have values in 10000s while feature may contain values in the magnitude of 10s or 100s, then the contribution of each feature for the dependent variable will be different

5. Building a linear model

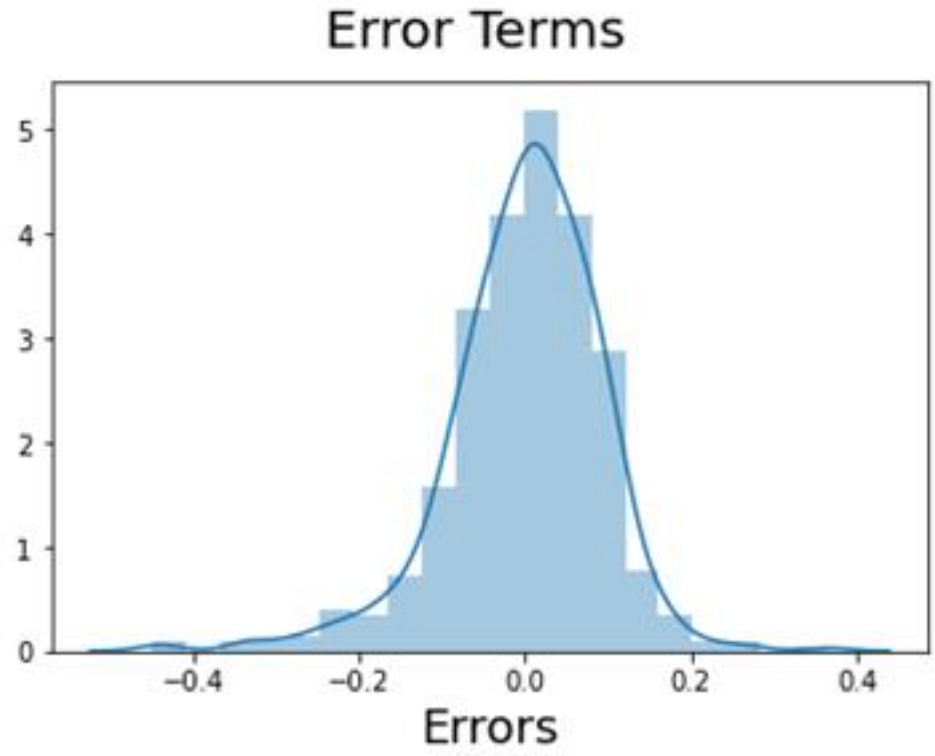
a. Forward Selection: We start with null model and add variables one by one. These variables are selection on the basis of high correlation with target variable. First we select the one, which has highest correlation and then we move on to the second highest and so on.

b. Backward Selection: We add all the variables at once and then eliminate variables based on high multicollinearity (VIF>5) or insignificance (high p- values).

c. RFE or Recursive Feature Elimination is more like an automated version of feature selection technique where we select that we need “m” variables out of “n” variables and then machine provides a list of features with importance level given in terms of rankings. A rank 1 means that feature is important for the model, while a rank 4 implies that we are better off, if we don’t consider the feature.

6. Residual analysis of the train data:

a. It tells us how much the errors (y_actual — y_pred) are distributed across the model. A good residual analysis will signify that the mean is centred around 0.



The residual errors are centered around 0

7. Making predictions using the final model and evaluation:

a. We will predict the test dataset by transforming it onto the trained dataset

b. Divide the test sets into X_test and y_test and calculate r2_score of test set. The train and test set should have similar r2_score. A difference of 2–3% between r2_score of train and test score is acceptable as per the standards.

2.2 Steps in EDA and Preprocessing:

1. Identification of variables and data types
2. Analyzing the basic metrics
3. Non-Graphical Univariate Analysis
4. Missing value treatment
5. Graphical Univariate Analysis
6. Bivariate Analysis
7. Encoding the categorical Data
8. Outlier treatment
9. Variable transformations
10. Correlation Analysis
11. Dimensionality Reduction
12. Scaling of Independent features

2.3 Steps followed in project

1. We need to apply feature engineering/ EDA on the dataset. After that you can split the dataset into train_model and test_model; use this train_model to train your model and test_model to validate your model.

2. After splitting the dataset you need to train at least 4-5 models.

3. Check performance of each model

For regression problem- create regression model and check the r2 score and metrics of each model .

4. Check the cross validation score for each model(for classification as well as for regression model).

5. Choose the model as the best model.

For regression problem- model with least difference between performance parameter and cross validation computed on same performance parameter is the best model. Example- Difference between r2 score and cross validation computed on r2 scoring parameter.

6. Perform hyper parameter tuning on the best model and check performance of the best model.

7. Save the best model.

*3. **ANALYSIS**: DATA FINDINGS AND INTERPRETATION*

3.1 Data shape and Data Types

The shape property returns a tuple containing the shape of the DataFrame.

The shape is the number of rows and columns of the DataFrame

When doing data analysis, it is important to make sure you are using the correct data types; otherwise you may get unexpected results or errors. In the case of pandas, it will correctly infer data types in many cases and you can move on with your analysis without any further thought on the topic.

Despite how well pandas works, at some point in your data analysis processes, you will likely need to explicitly convert data from one type to another. This article will discuss the basic pandas data types (aka dtypes), how they map to python and numpy data types and the options for converting from one pandas type to another.

Pandas Data Types

A data type is essentially an internal construct that a programming language uses to understand how to store and manipulate data. For instance, a program needs to understand that you can add two numbers together like 5 + 10 to get 15. Or, if you have two strings such as “cat” and “hat” you could concatenate (add) them together to get “cathat.”

A possible confusing point about pandas data types is that there is some overlap between pandas, python and numpy. This table summarizes the key points:

Pandas dtype mapping

Pandas dtype	**Python type**	**NumPy type**	**Usage**
object	str or mixed	string_, unicode_, mixed types	Text or mixed numeric and non-numeric values
int64	Int	int_, int8, int16, int32, int64, uint8, uint16, uint32, uint64	Integer numbers
float64	Float	float_, float16, float32, float64	Floating point numbers
bool	Bool	bool_	True/False values

Pandas dtype mapping

Pandas dtype	**Python type**	**NumPy type**	**Usage**
datetime64	NA	datetime64[ns]	Date and time values
timedelta[ns]	NA	NA	Differences between two datetimes
category	NA	NA	Finite list of text values

For the most part, there is no need to worry about determining if you should try to explicitly force the pandas type to a corresponding to NumPy type. Most of the time, using pandas default int64 and float64 types will work. The only reason I included in this table is that sometimes you may see the numpy types pop up on-line or in your own analysis.

For this article, I will focus on the follow pandas types:

- object
- int64
- float64
- datetime64
- bool

The category and timedelta types are better served in an article of their own if there is interest. However, the basic approaches outlined in this article apply to these types as well.

One other item I want to highlight is that the object data type can actually contain multiple different types. For instance, the a column could include integers, floats and strings which collectively are labeled as an object . Therefore, you may need some additional techniques to handle mixed data types in object columns.

Columns and datatypes of it on our dataset:

There are total 71 columns in our dataset and they are of the following datatypes:

1Gender of respondent – integer (int64)
2 How old are you? - integer (int64)
3 Which city do you shop online from? - object
4 What is the Pin Code of where you shop online from? - integer (int64)
5 Since How Long You are Shopping Online ? - integer (int64)
6 How many times you have made an online purchase in the past 1 year? - integer (int64)

7 How do you access the internet while shopping on-line? - integer (int64)
8 Which device do you use to access the online shopping? - integer (int64)
9 What is the screen size of your mobile device? - integer (int64)
10 What is the operating system (OS) of your device? - integer (int64)
11 What browser do you run on your device to access the website? - integer (int64)
12 Which channel did you follow to arrive at your favorite online store for the first time? - integer (int64)
13 After first visit, how do you reach the online retail store? - integer (int64)
14 How much time do you explore the e- retail store before making a purchase decision? - integer (int64)
15 What is your preferred payment Option? - integer (int64)
16 How frequently do you abandon (selecting an items and leaving without making payment) your shopping cart? - integer (int64)
17 Why did you abandon the "Bag", "Shopping Cart"? - integer (int64)
18 The content on the website must be easy to read and understand - integer (int64)
19 Information on similar product to the one highlighted is important for product comparison - integer (int64)
20 Complete information on listed seller and product being offered is important for purchase decision. - integer (int64)
21 All relevant information on listed products must be stated clearly - integer (int64)
22 Ease of navigation in website - integer (int64)
23 Loading and processing speed - integer (int64)
24 User friendly Interface of the website - integer (int64)
25 Convenient Payment methods - integer (int64)
26 Trust that the online retail store will fulfill its part of the transaction at the stipulated time - integer (int64)
27 Empathy (readiness to assist with queries) towards the customers - integer (int64)
28 Being able to guarantee the privacy of the customer - integer (int64)
29 Responsiveness, availability of several communication channels (email, online rep, twitter, phone etc.) - integer (int64)
30 Online shopping gives monetary benefit and discounts - integer (int64)
31 Enjoyment is derived from shopping online - integer (int64)
32 Shopping online is convenient and flexible - integer (int64)
33 Return and replacement policy of the e-tailer is important for purchase decision - integer (int64)
34 Gaining access to loyalty programs is a benefit of shopping online - integer (int64)
35 Displaying quality Information on the website improves satisfaction of customers - integer (int64)
36 User derive satisfaction while shopping on a good quality website or application - integer (int64)
37 Net Benefit derived from shopping online can lead to users satisfaction - integer (int64)
38 User satisfaction cannot exist without trust - integer (int64)
39 Offering a wide variety of listed product in several category - integer (int64)
40 Provision of complete and relevant product information - integer (int64)
41 Monetary savings - integer (int64)
42 The Convenience of patronizing the online retailer - integer (int64)
43 Shopping on the website gives you the sense of adventure - integer (int64)
44 Shopping on your preferred e-tailer enhances your social status - integer (int64)
45 You feel gratification shopping on your favorite e-tailer - integer (int64)

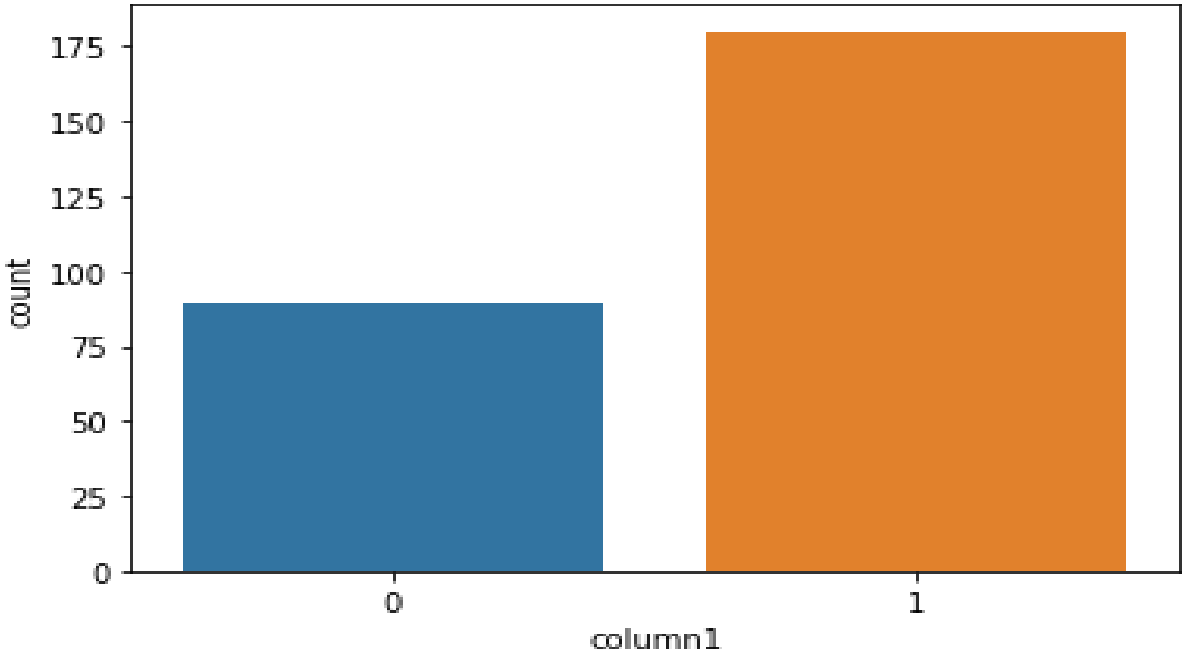
46 Shopping on the website helps you fulfill certain roles - integer (int64)
47 Getting value for money spent - integer (int64)
48 From the following, tick any (or all) of the online retailers you have shopped from - object
49 Easy to use website or application - object
50 Visual appealing web-page layout - object
51 Wild variety of product on offer - object
52 Complete, relevant description information of products - object
53 Fast loading website speed of website and application - object
54 Reliability of the website or application - object
55 Quickness to complete purchase - object
56 Availability of several payment options - object
57 Speedy order delivery - object
58 Privacy of customers' information - object
59 Security of customer financial information - object
60 Perceived Trustworthiness - object
61 Presence of online assistance through multi-channel - object
62 Longer time to get logged in (promotion, sales period) - object
63 Longer time in displaying graphics and photos (promotion, sales period) - object
64 Late declaration of price (promotion, sales period) - object
65 Longer page loading time (promotion, sales period) - object
66 Limited mode of payment on most products (promotion, sales period) - object
67 Longer delivery period - object
68 Change in website/Application design - object
69 Frequent disruption when moving from one page to another - object
70 Website is as efficient as before - object
71 Which of the Indian online retailer would you recommend to a friend? - object

3.2 DATA VISUALIZATION

Firstly, to avoid confusion and errors, we name the columns from column1 to column71. Then, we make the dataframe for the nominal categorial data under our dataset customer.csv. We see that, all the columns except for column 4 which is the Pincode column, the data is spread over certain categories. For example: Gender- Male/Female, City- Delhi, Noida etc. We use countplot for data visualization of the categorial data as it gives the frequency and value counts of specific categories.

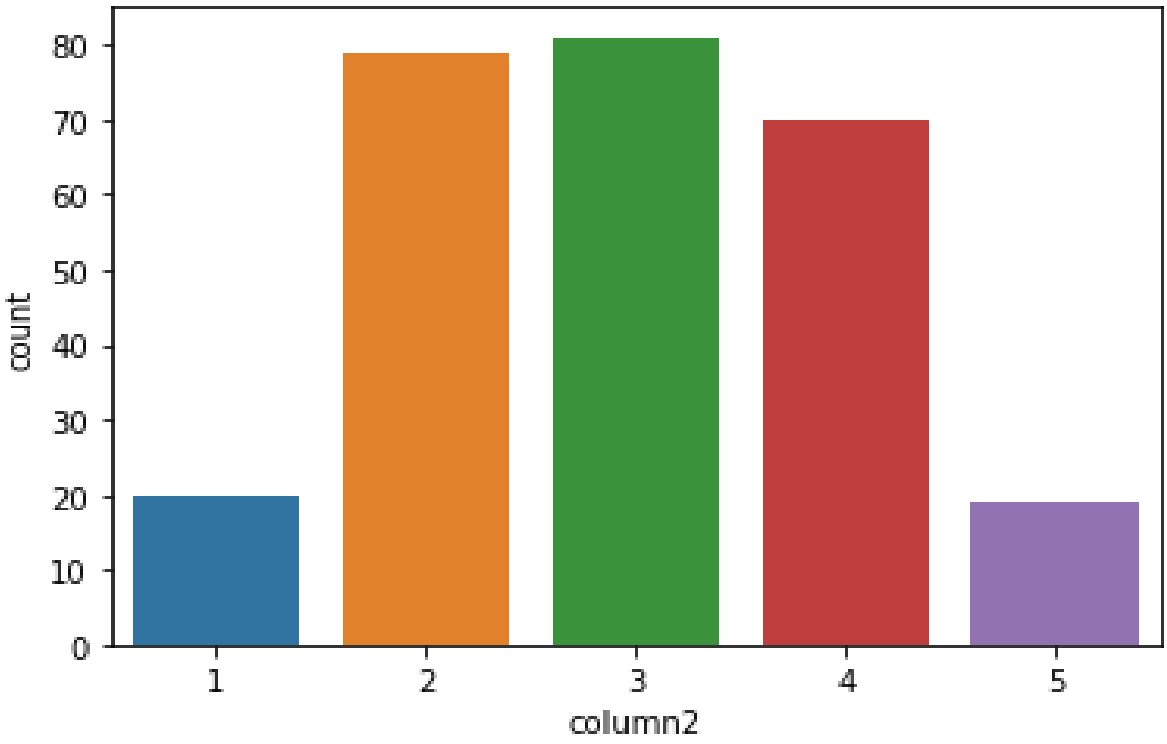
1. Column1-

```
1    180
0     89
```



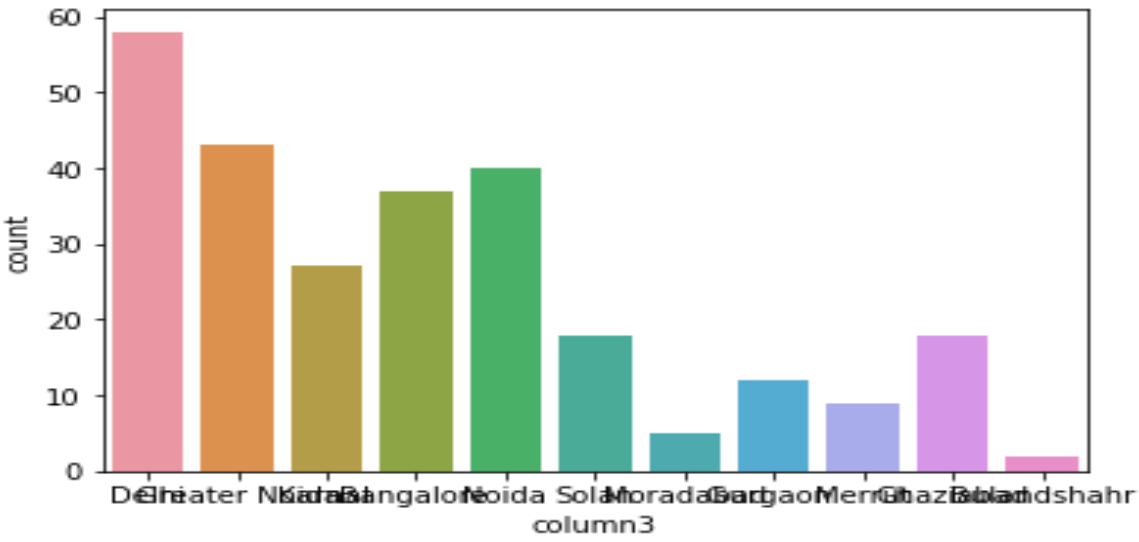
2. Column2-

```
3    81
2    79
4    70
1    20
5    19
```



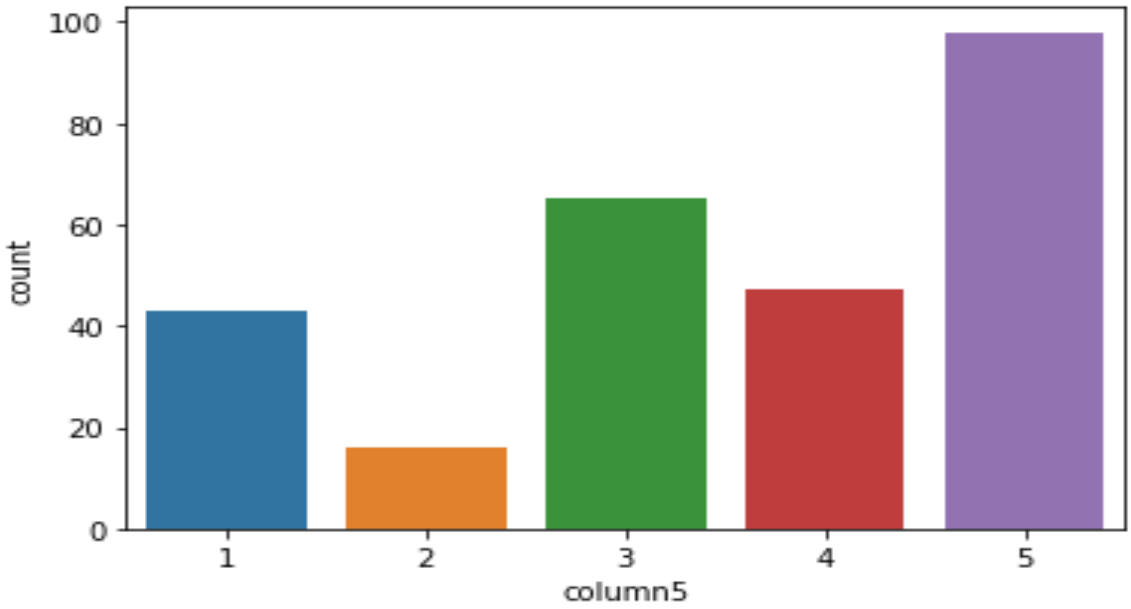
3. Column3-

```
Delhi              58
Greater Noida      43
Noida              40
Bangalore          37
Karnal             27
Solan              18
Ghaziabad          18
Gurgaon            12
Merrut              9
Moradabad           5
Bulandshahr         2
```



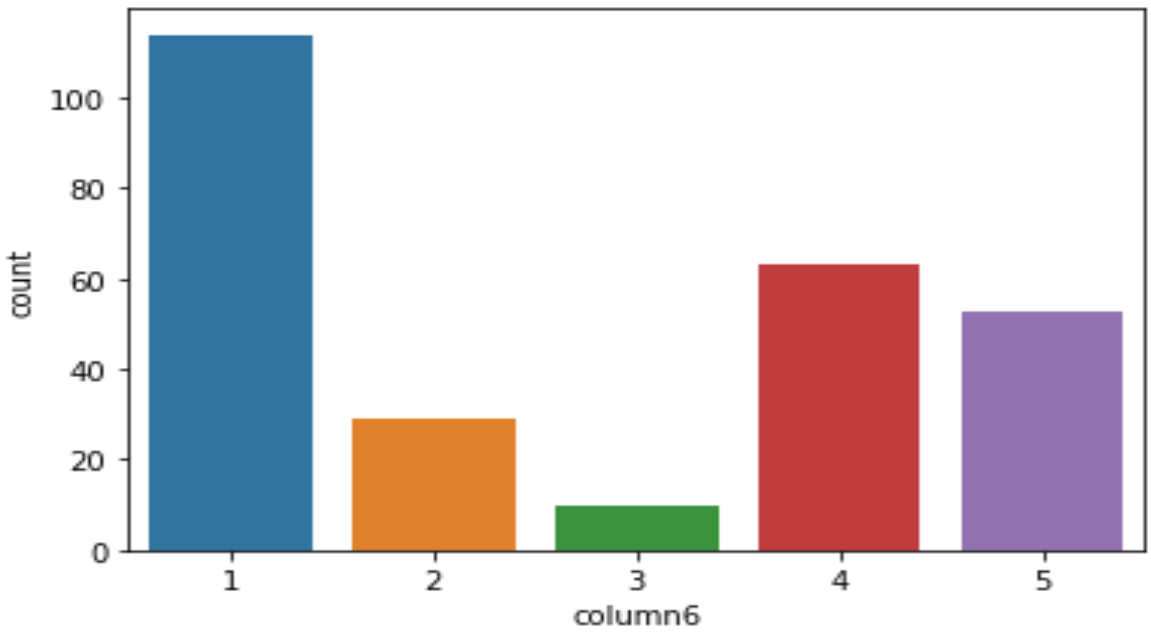
4. Column5-

```
5     98
3     65
4     47
1     43
2     16
```



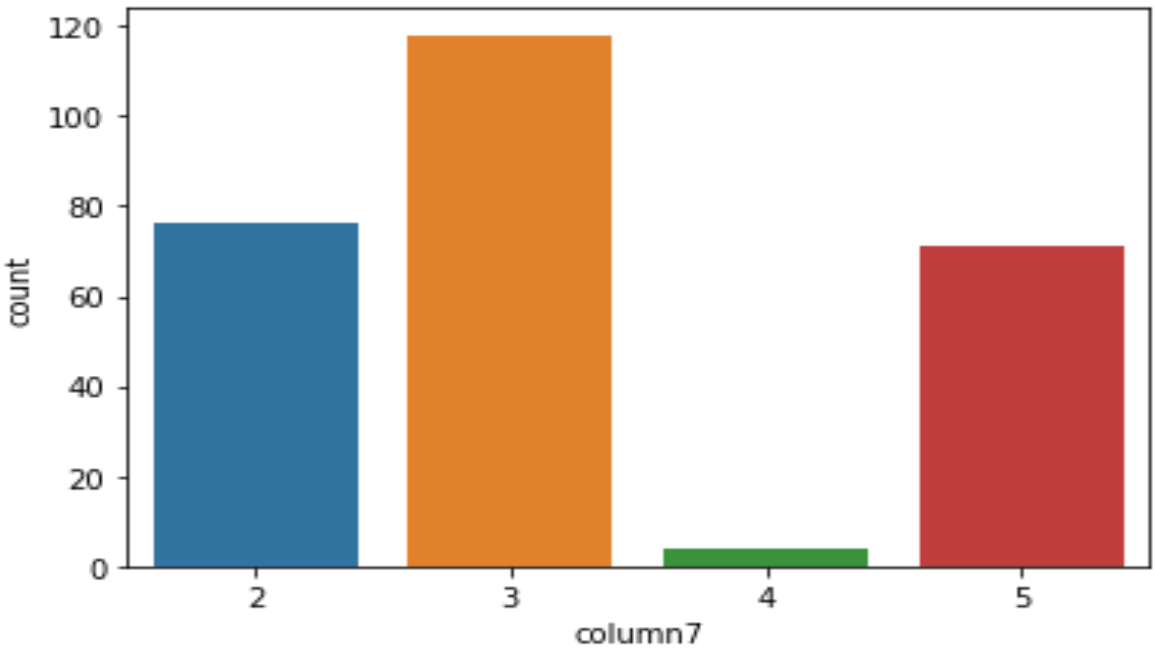
5. Column6-

```
1     114
4      63
5      53
2      29
3      10
```



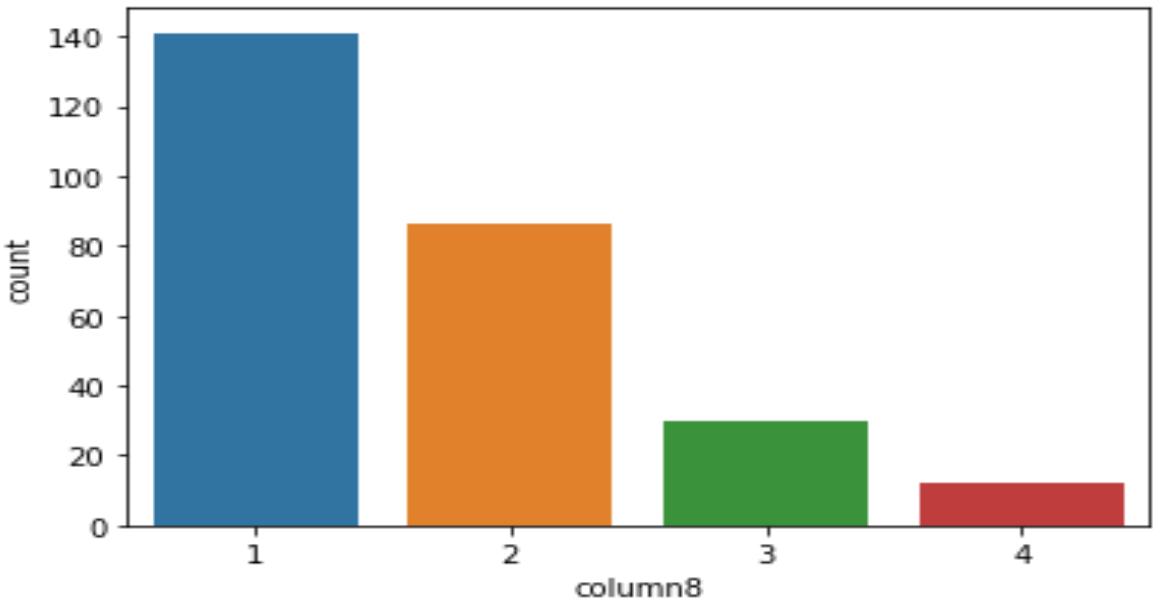
6. Column7-

```
3    118
2     76
5     71
4      4
```



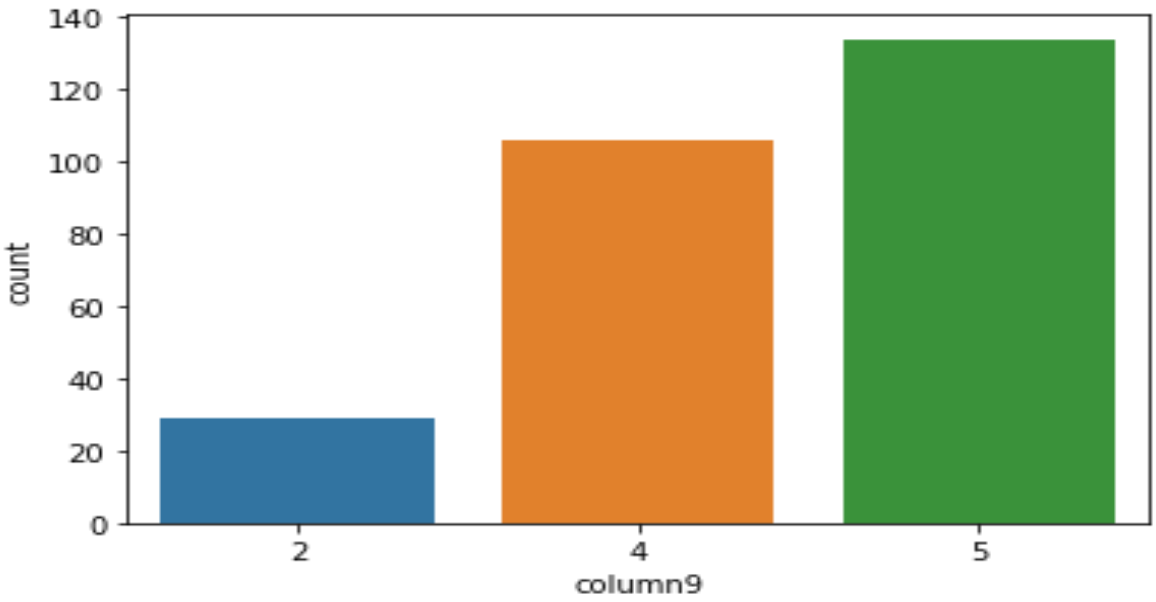
7. Column8-

```
1    141
2     86
3     30
4     12
```



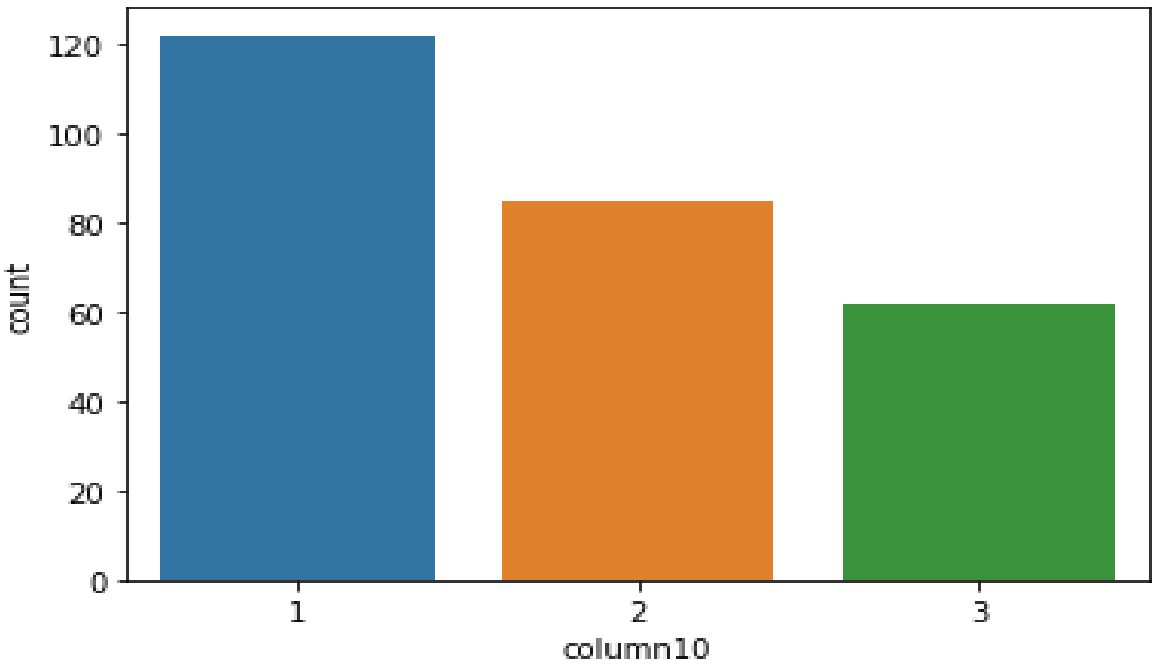
8. Column9-

```
5    134
4    106
2     29
```



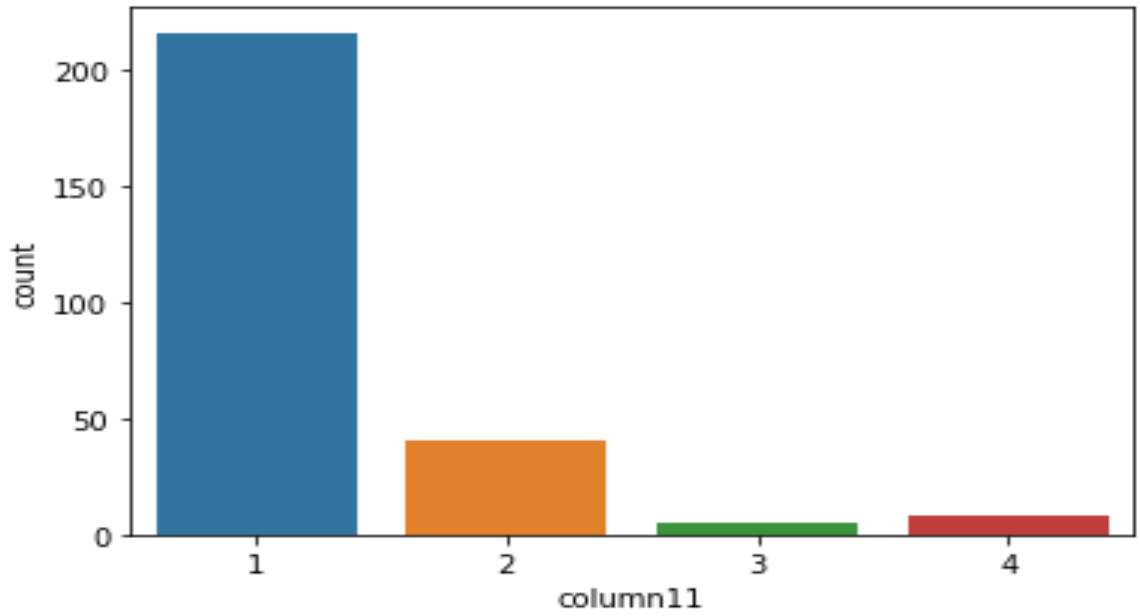
9. Column10-

```
1    122
2     85
3     62
```



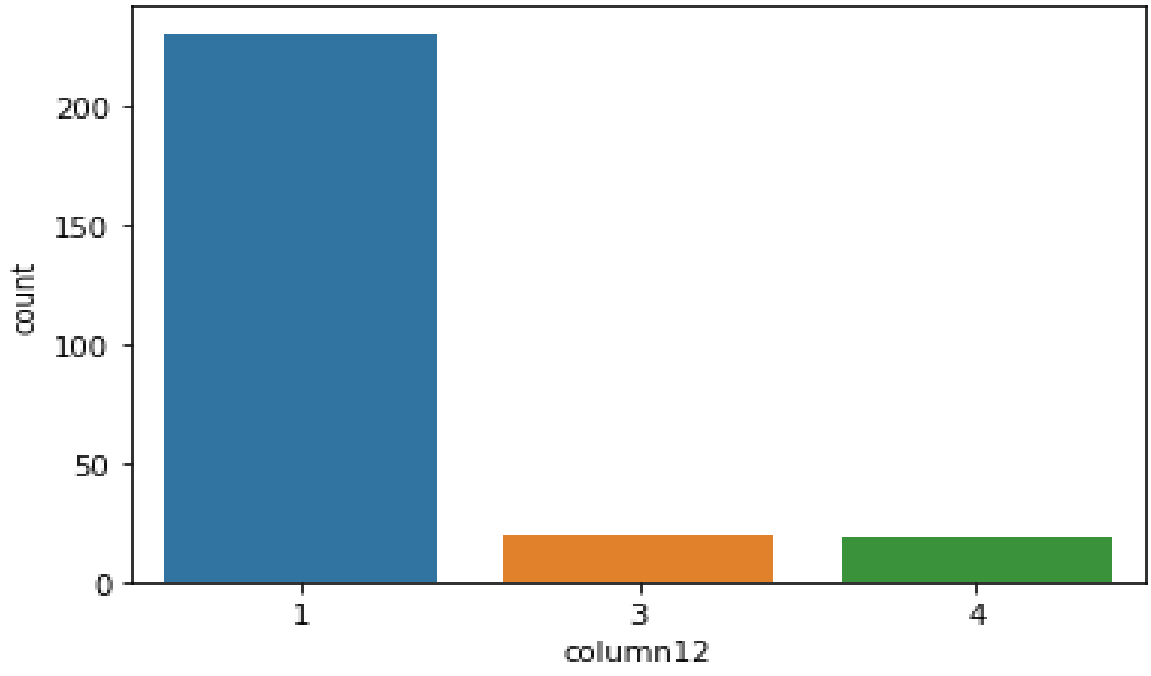
10. Column11-

```
1    216
2     40
4      8
3      5
```



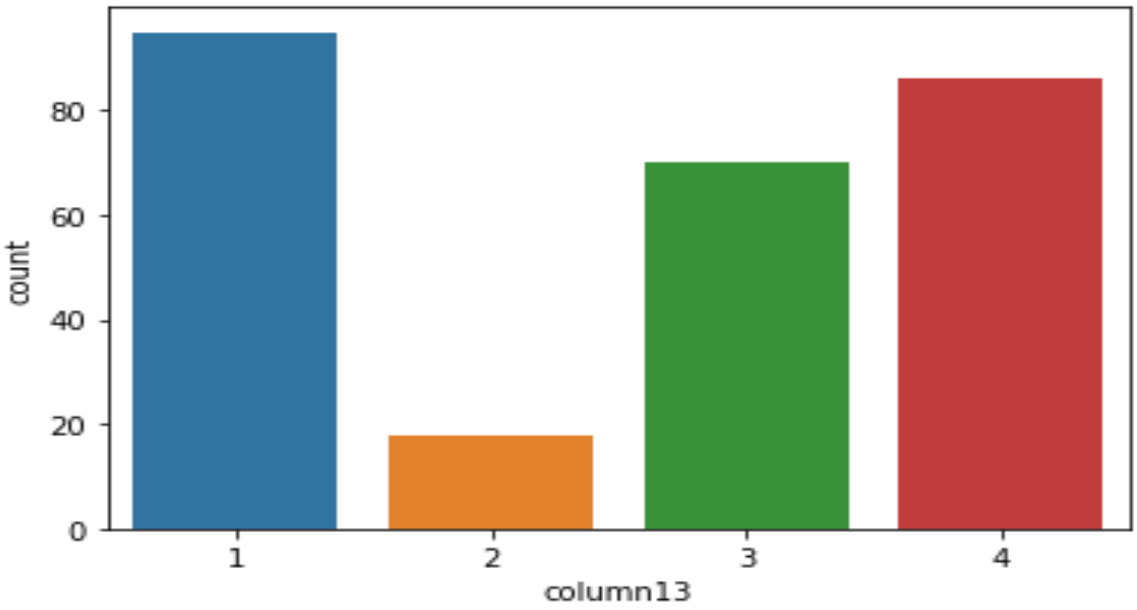
11. Column12-

```
1    230
3     20
4     19
```



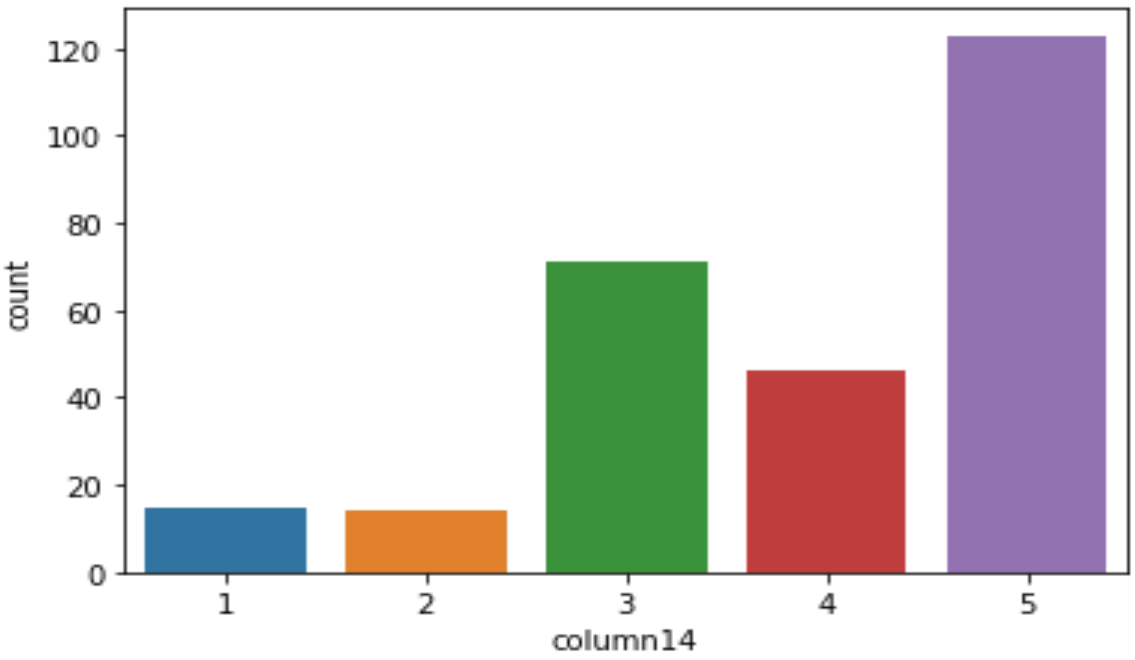
12. Column13-

```
1     95
4     86
3     70
2     18
```



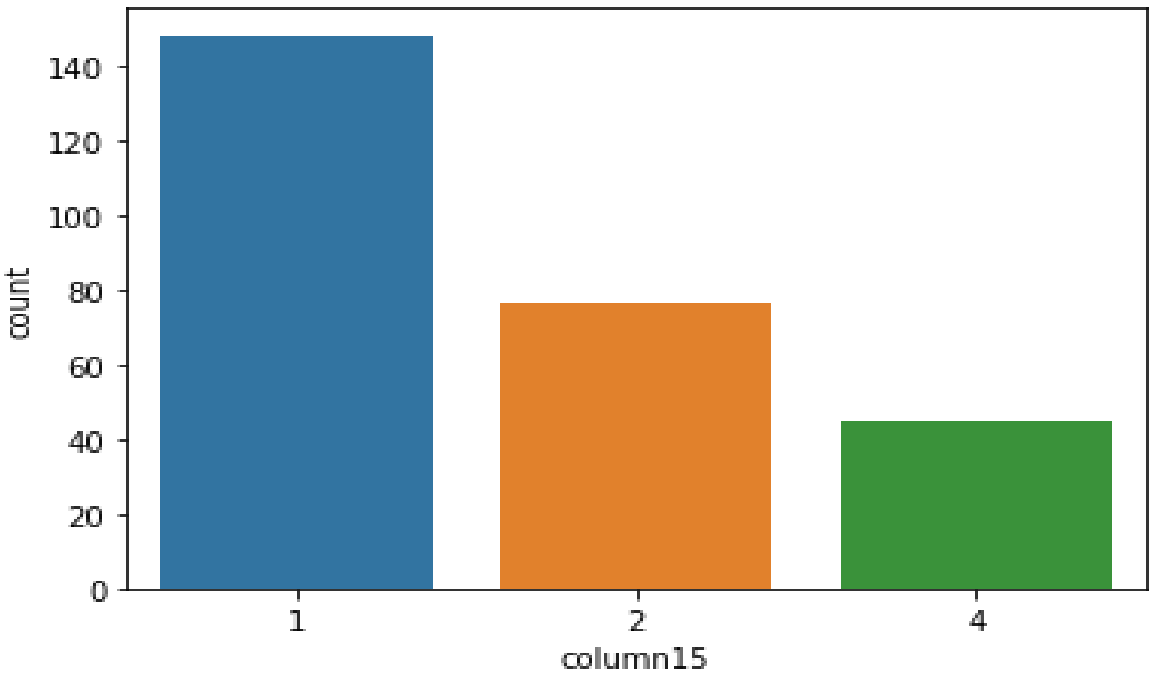
13. Column14-

```
5     123
3      71
4      46
1      15
2      14
```



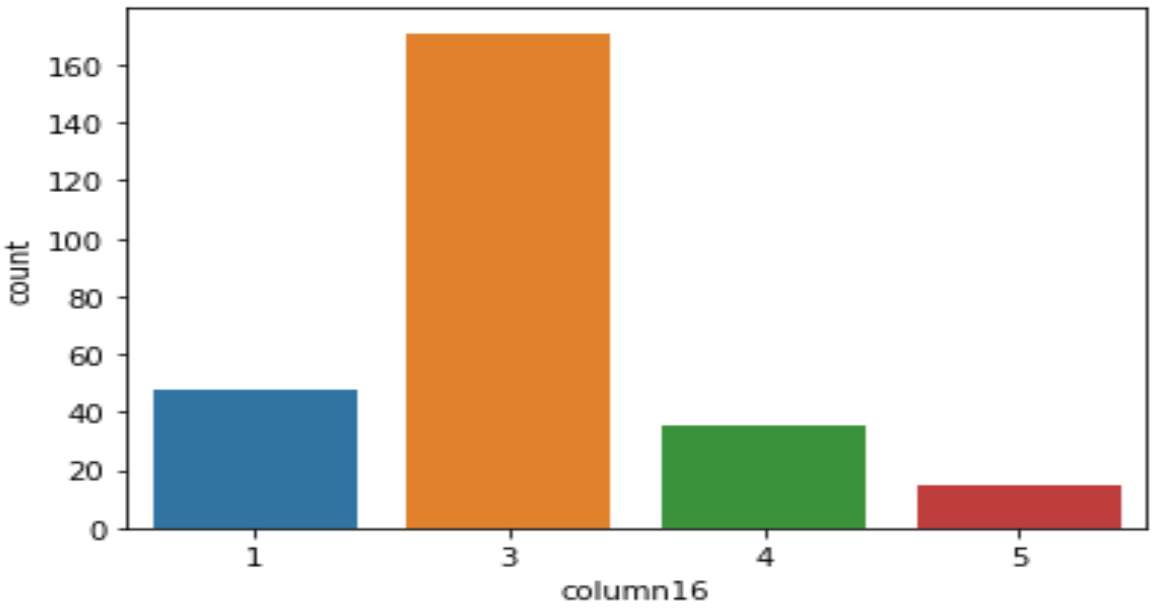
14. Column15-

```
1    148
2     76
4     45
```



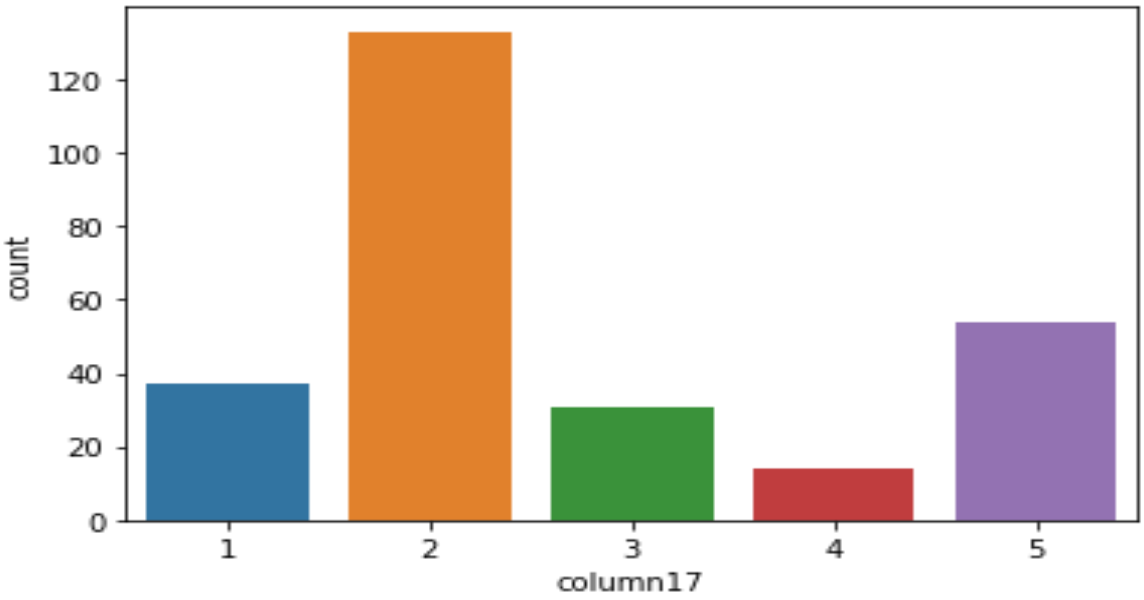
15. Column16-

```
3    171
1     48
4     35
5     15
```



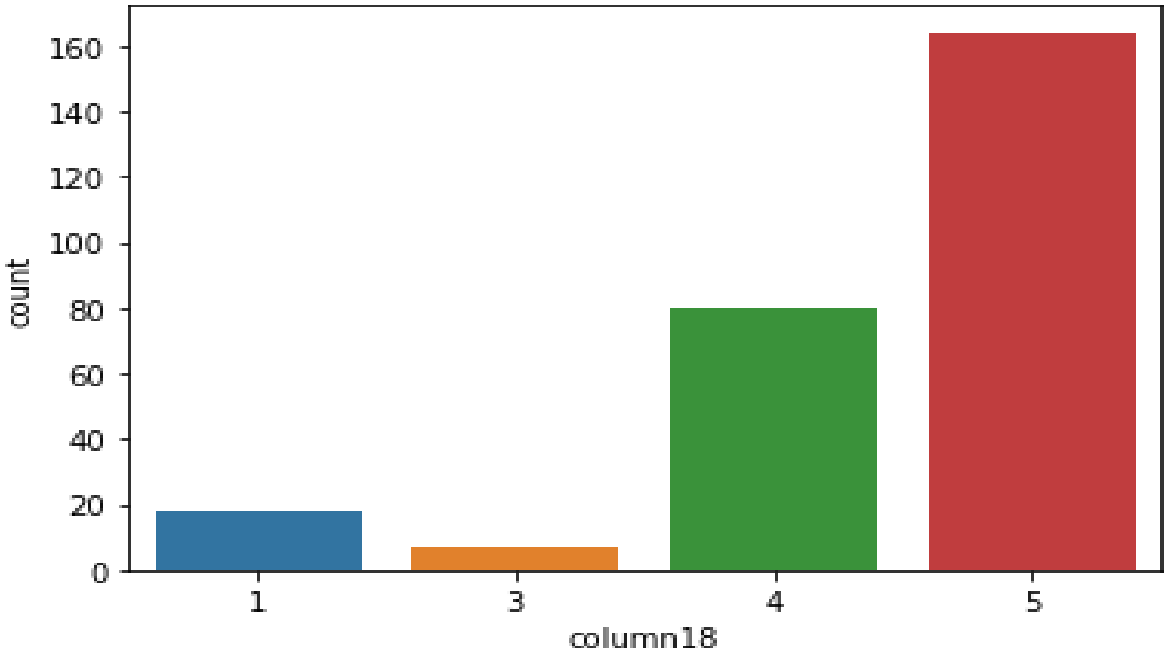
16. Column17-

```
2    133
5     54
1     37
3     31
4     14
```



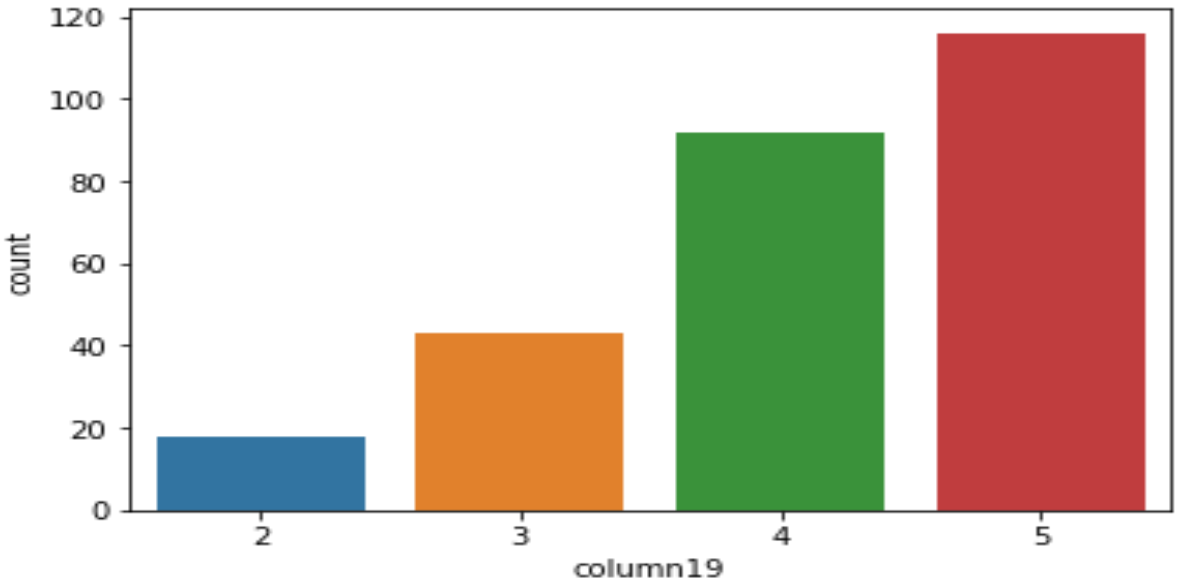
17. Column18-

```
5    164
4     80
1     18
3      7
```



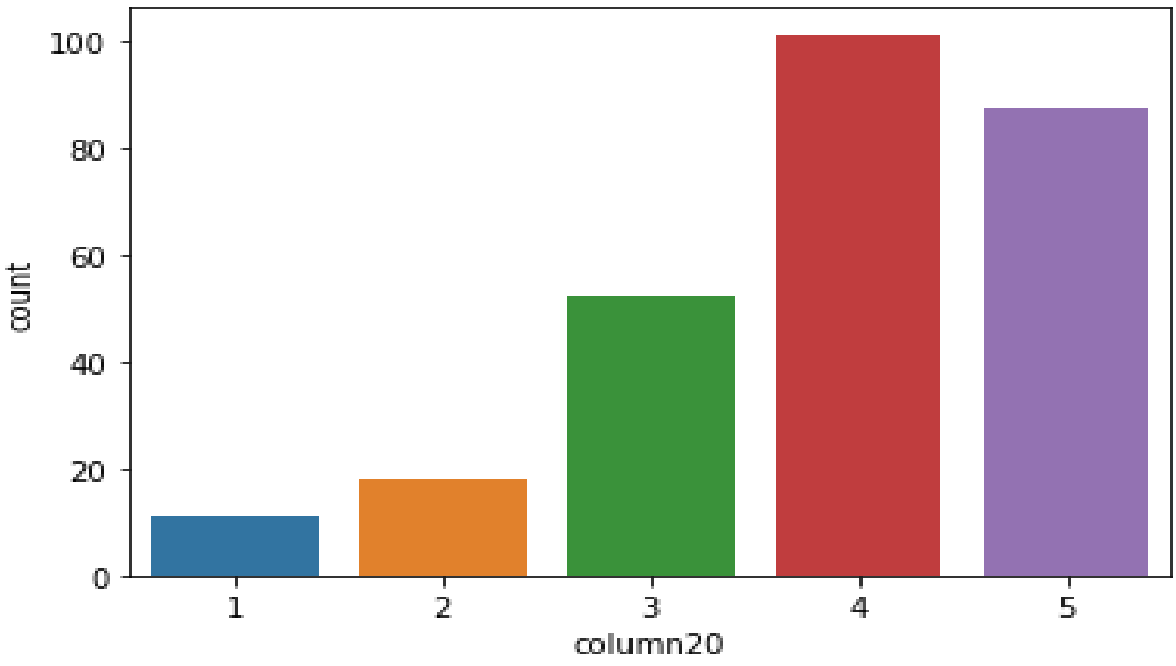
18. Column19-

```
5    116
4     92
3     43
2     18
```



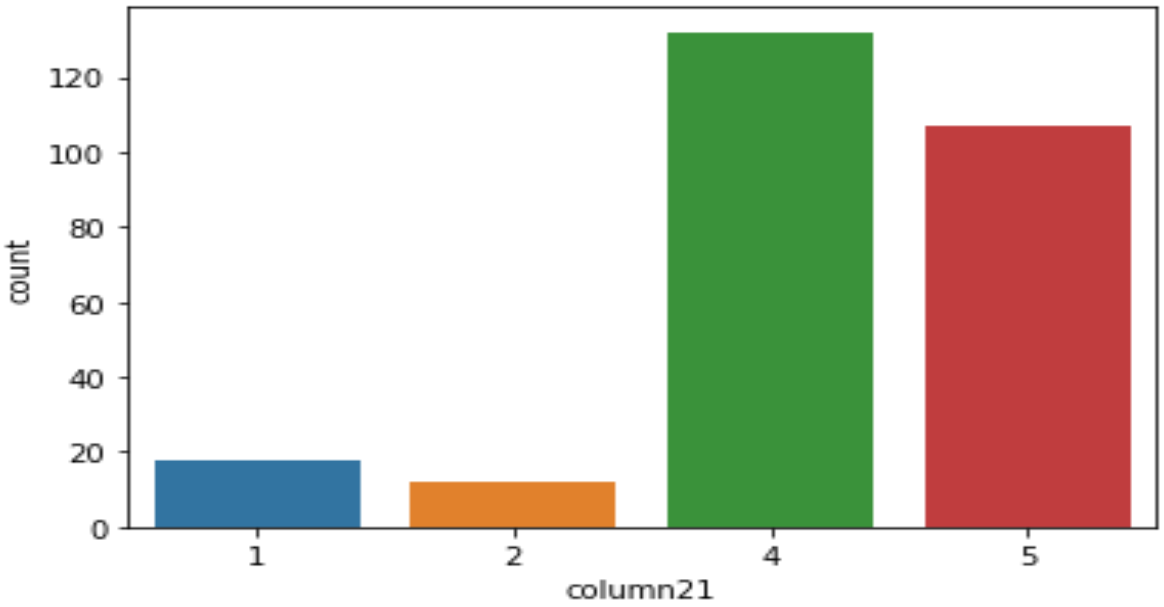
19. Column20-

```
4    101
5     87
3     52
2     18
1     11
```



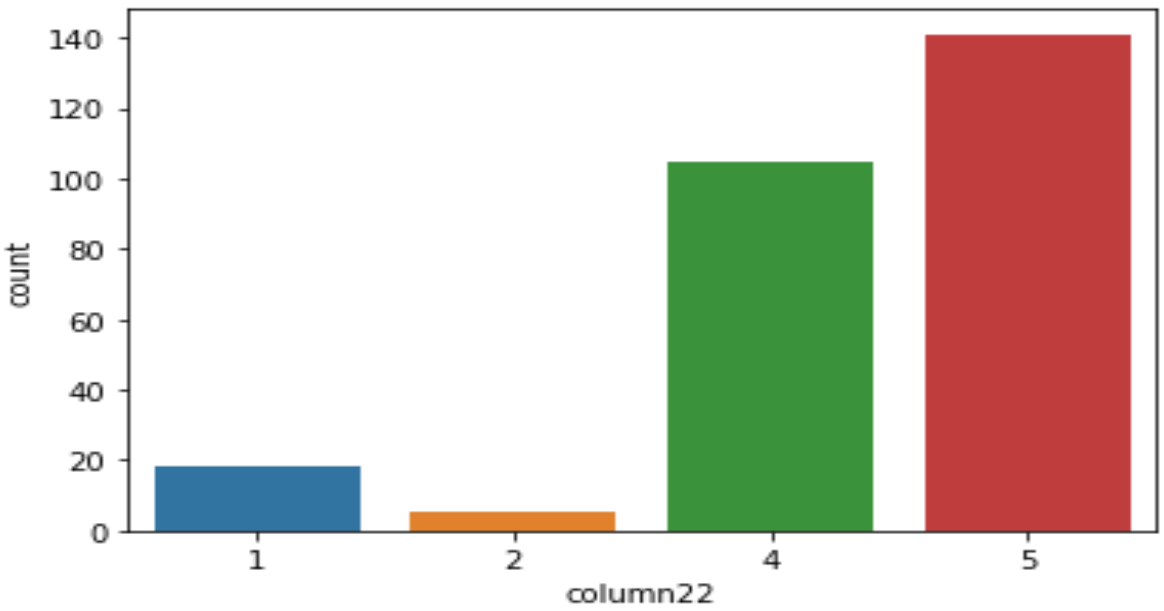
20. Column21-

```
4    132
5    107
1     18
2     12
```



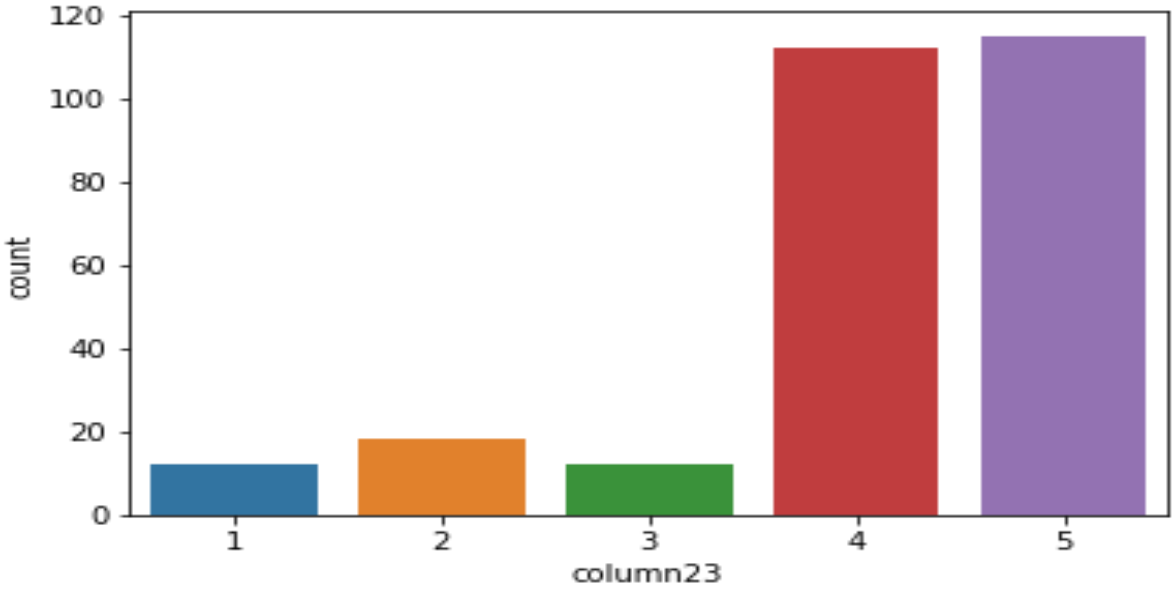
21. Column22-

```
5    141
4    105
1     18
2      5
```



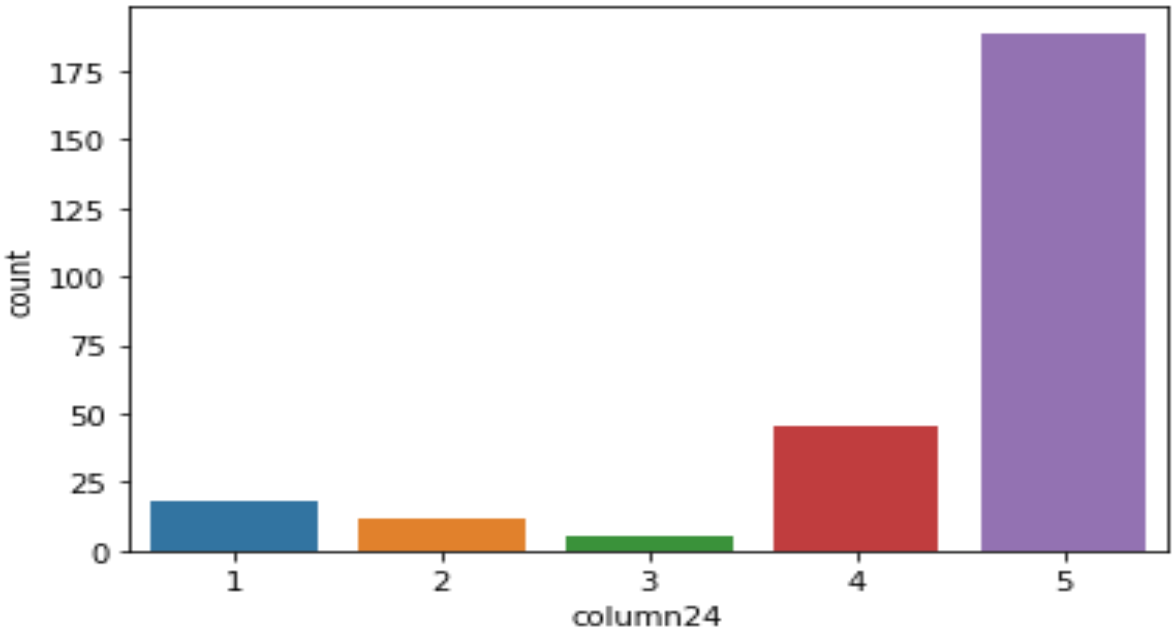
22. Column23-

```
5    115
4    112
2     18
1     12
3     12
```



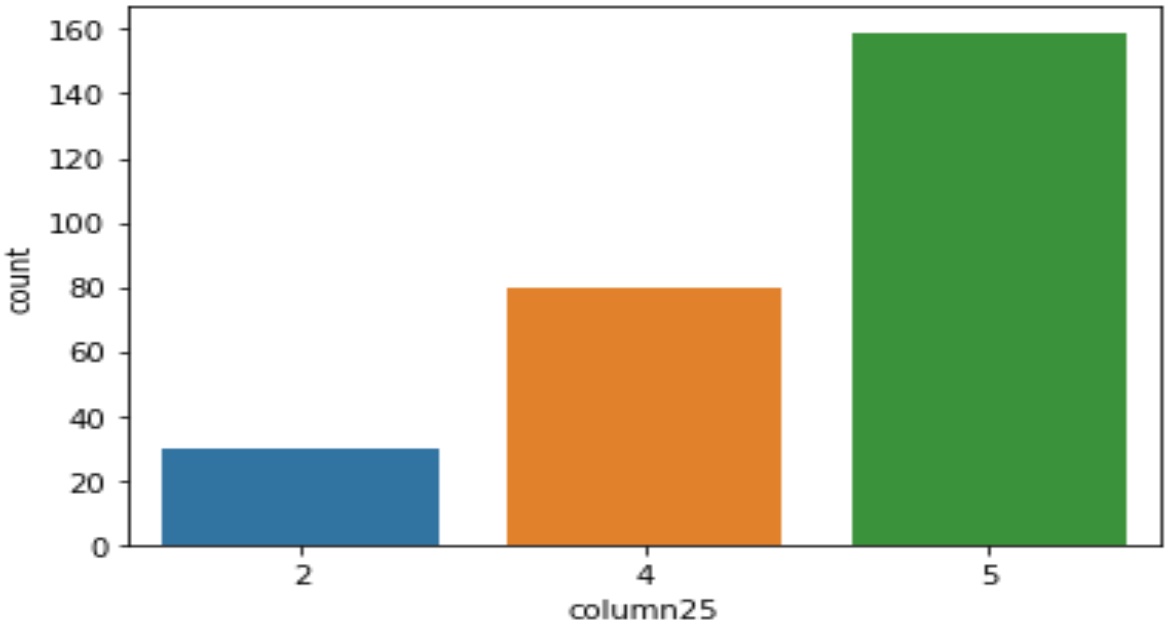
23. Column24-

```
5    189
4     45
1     18
2     12
3      5
```



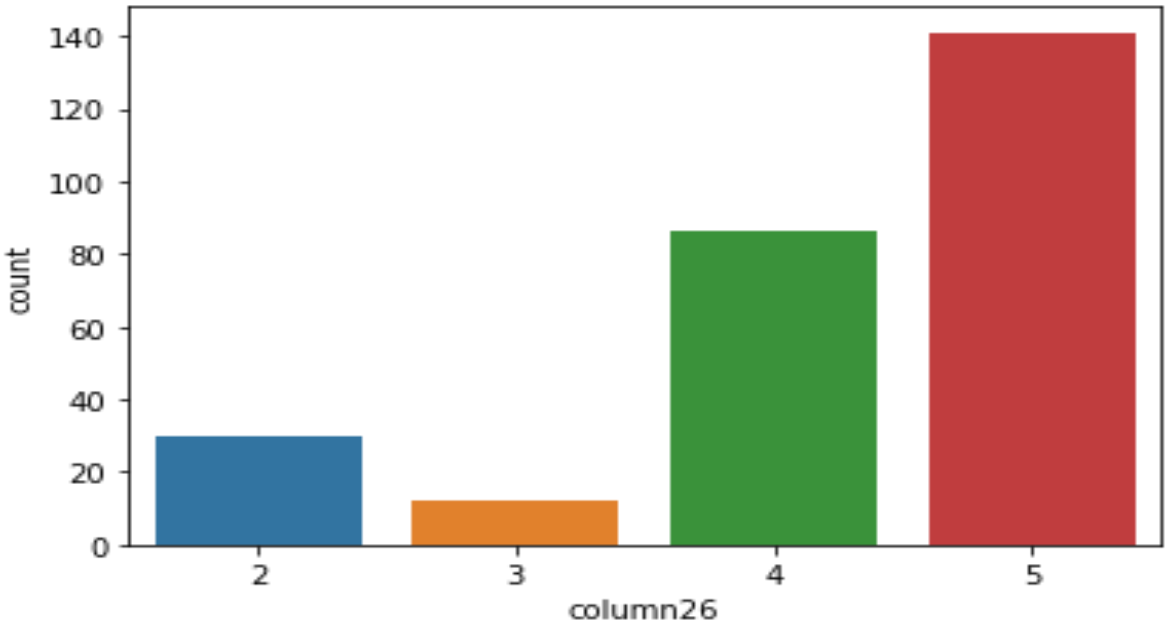
24. Column25-

```
5    159
4     80
2     30
```



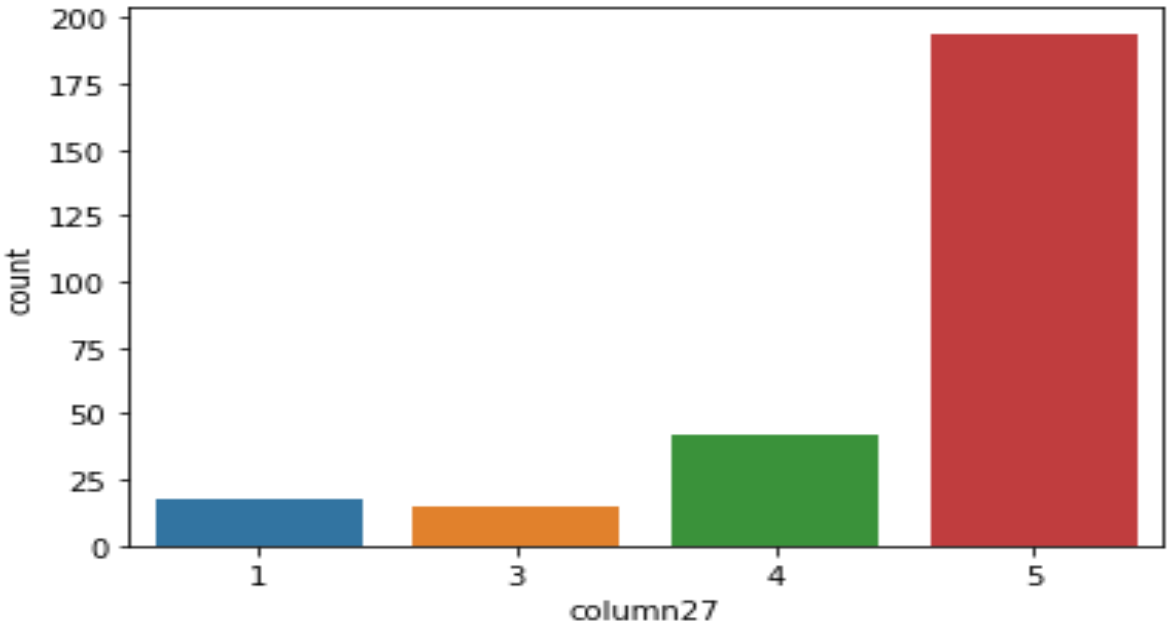
25. Column26-

```
5    141
4     86
2     30
3     12
```



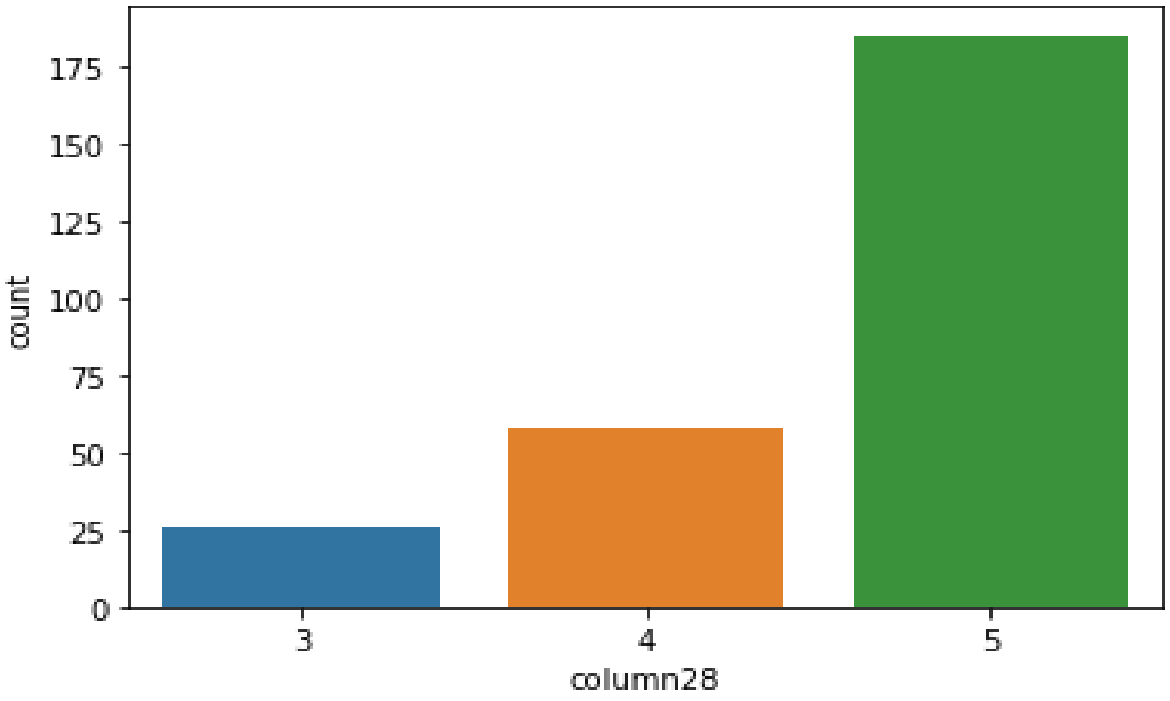
26. Column27-

```
5    194
4     42
1     18
3     15
```



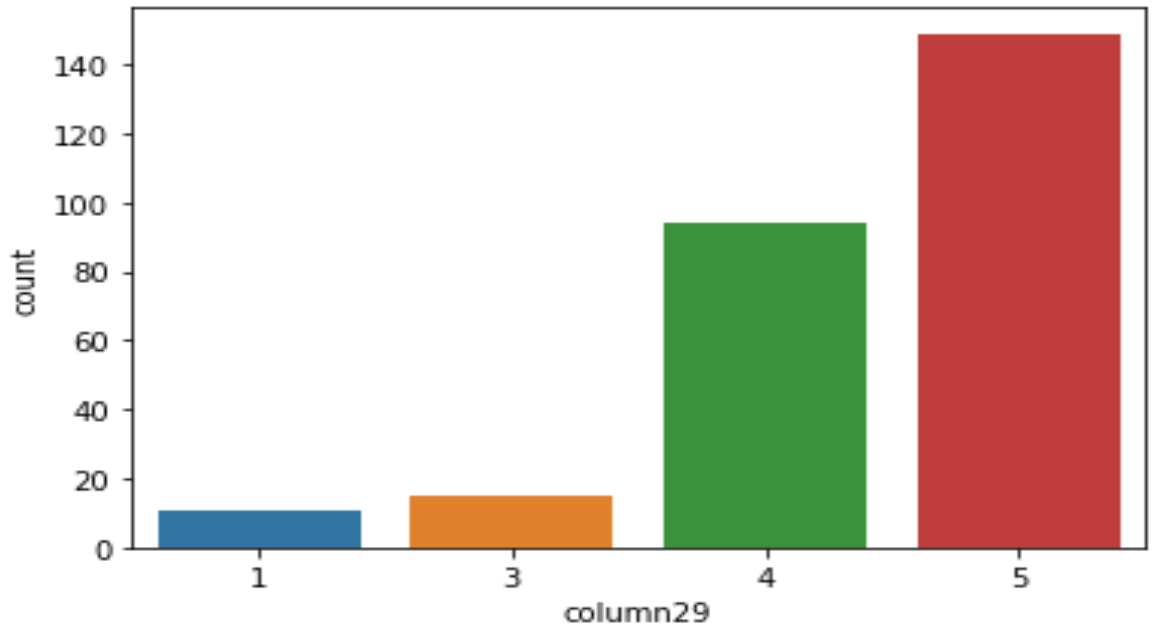
27. Column28-

```
5    185
4     58
3     26
```



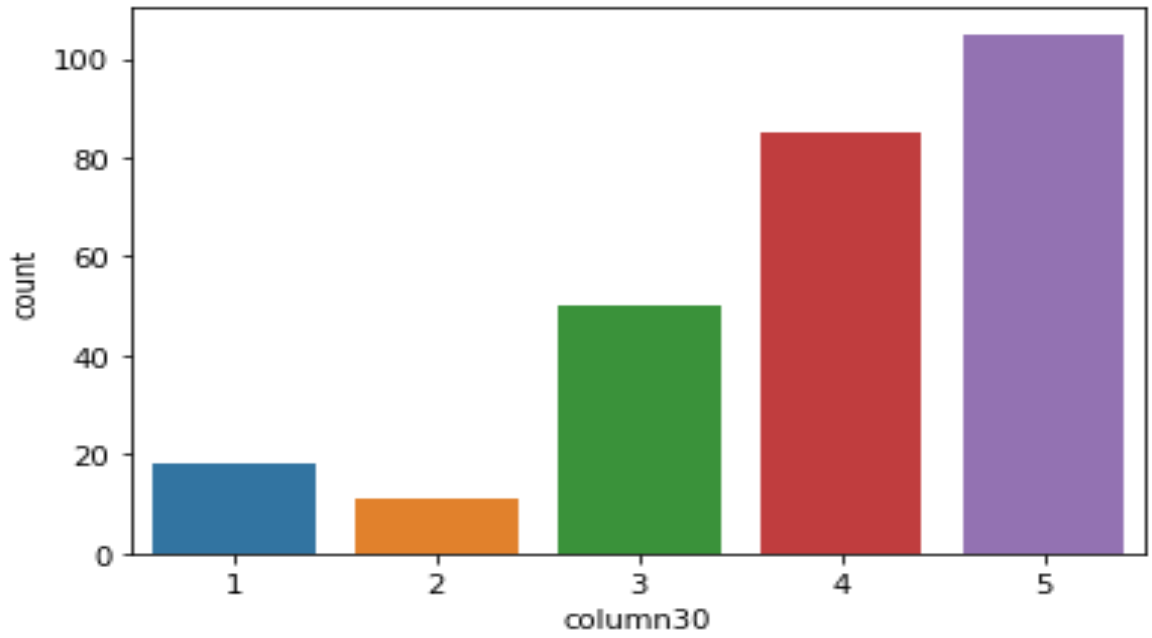
28. Column29-

```
5    149
4     94
3     15
1     11
```



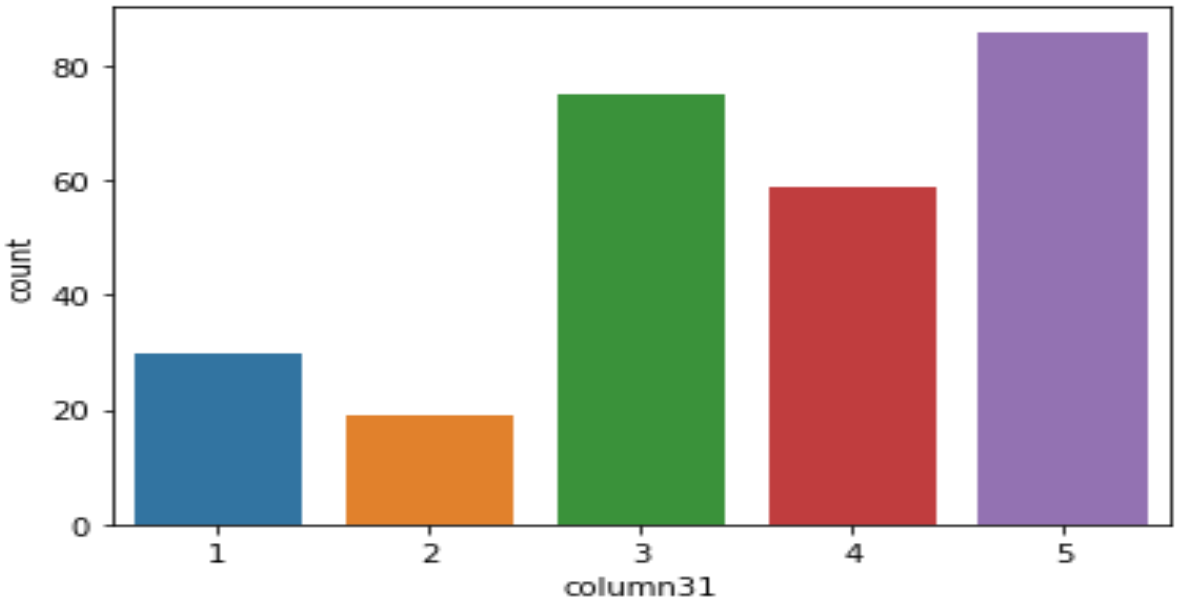
29. Column30-

```
5    105
4     85
3     50
1     18
2     11
```



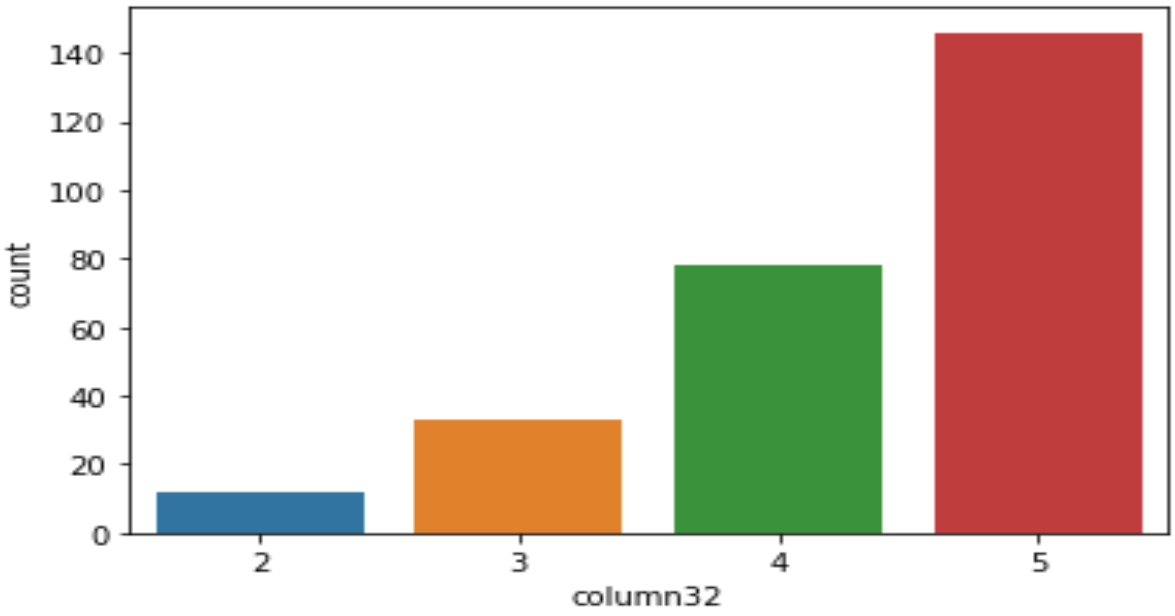
30. Column31-

```
5    86
3    75
4    59
1    30
2    19
```



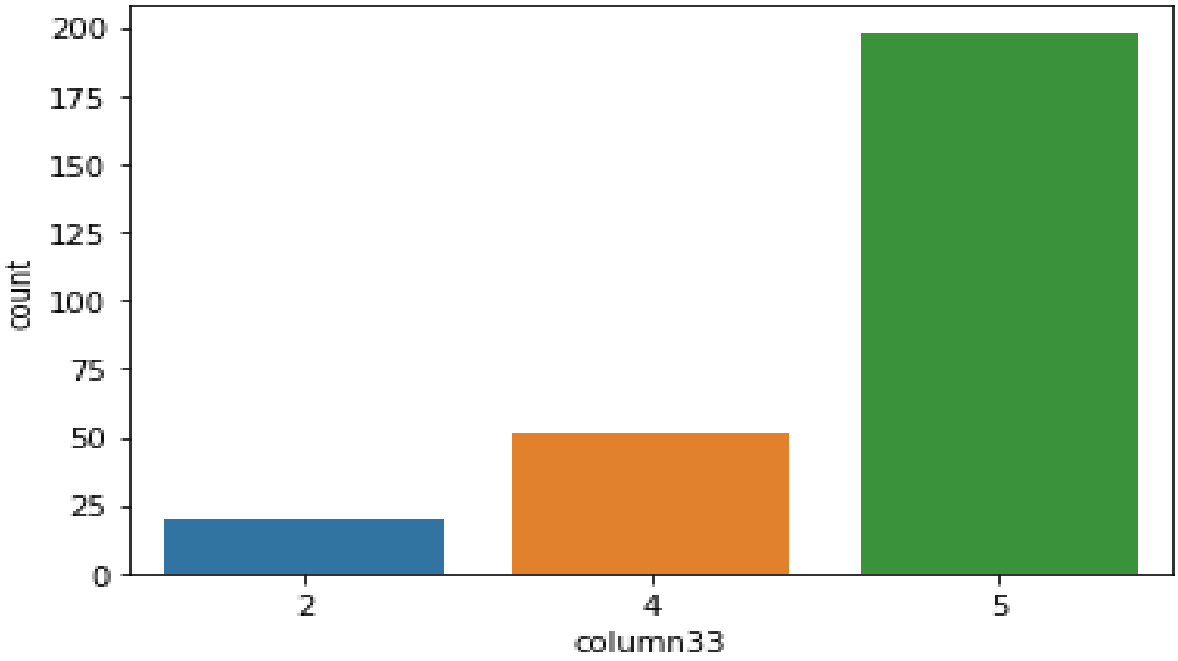
31. Column32-

```
5    146
4     78
3     33
2     12
```



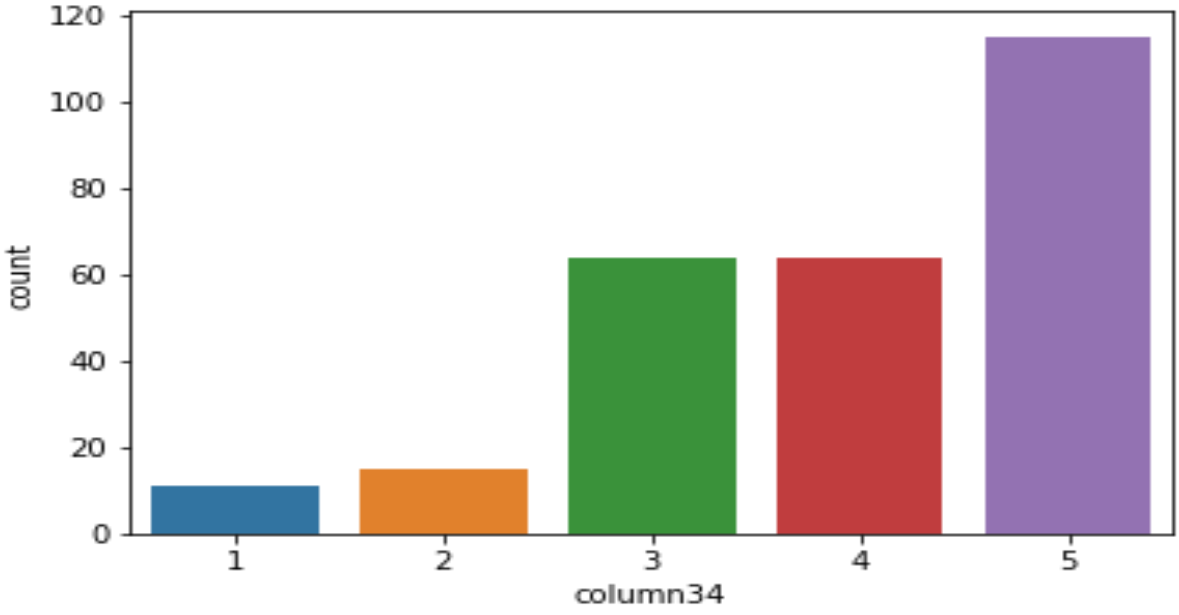
32. Column33-

```
5    198
4     51
2     20
```



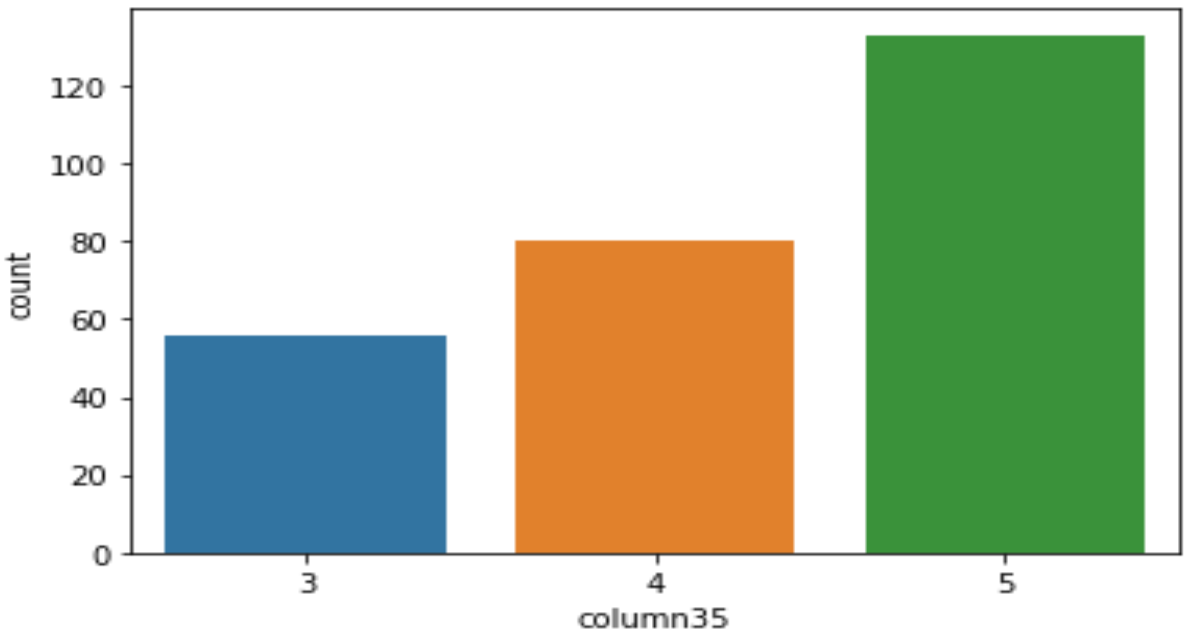
33. Column34-

```
5    115
4     64
3     64
2     15
1     11
```



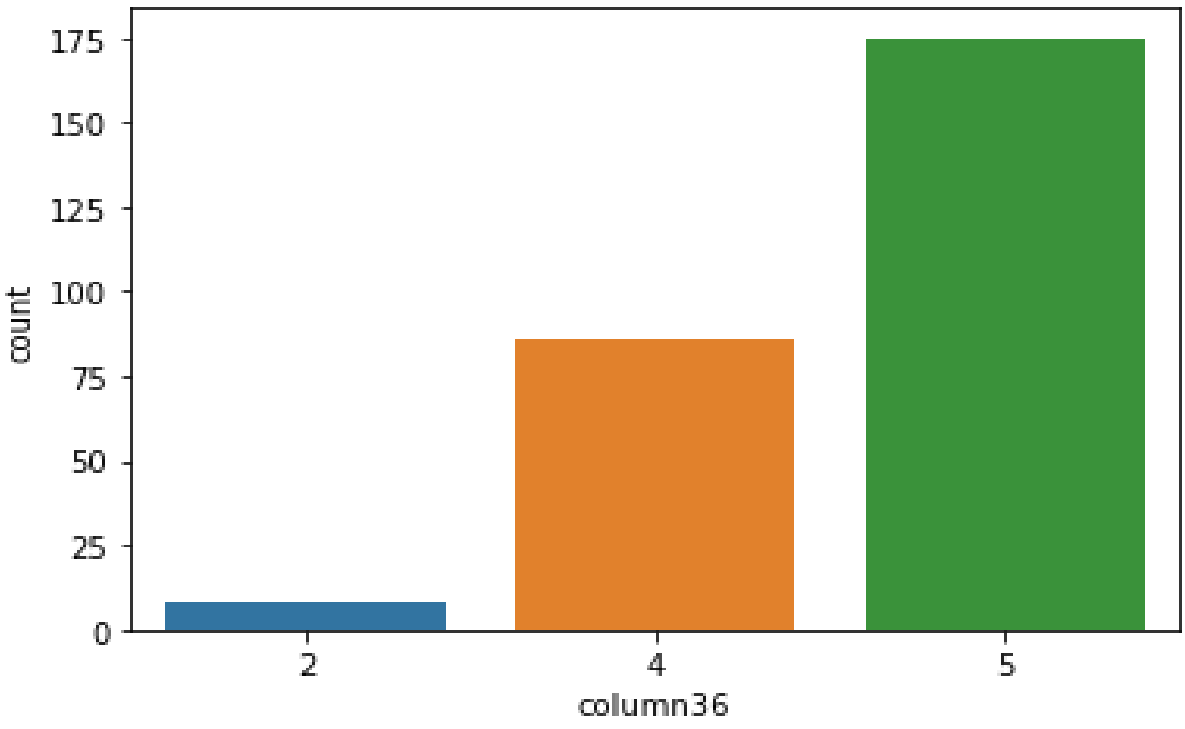
34. Column35-

```
5    133
4     80
3     56
```



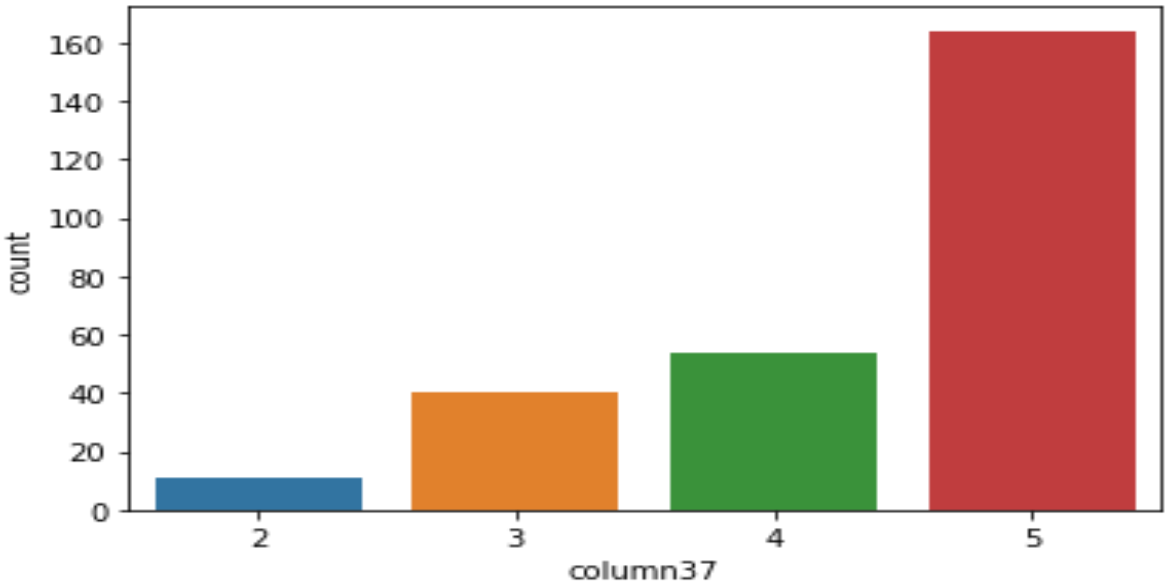
35. Column36-

```
5    175
4     86
2      8
```



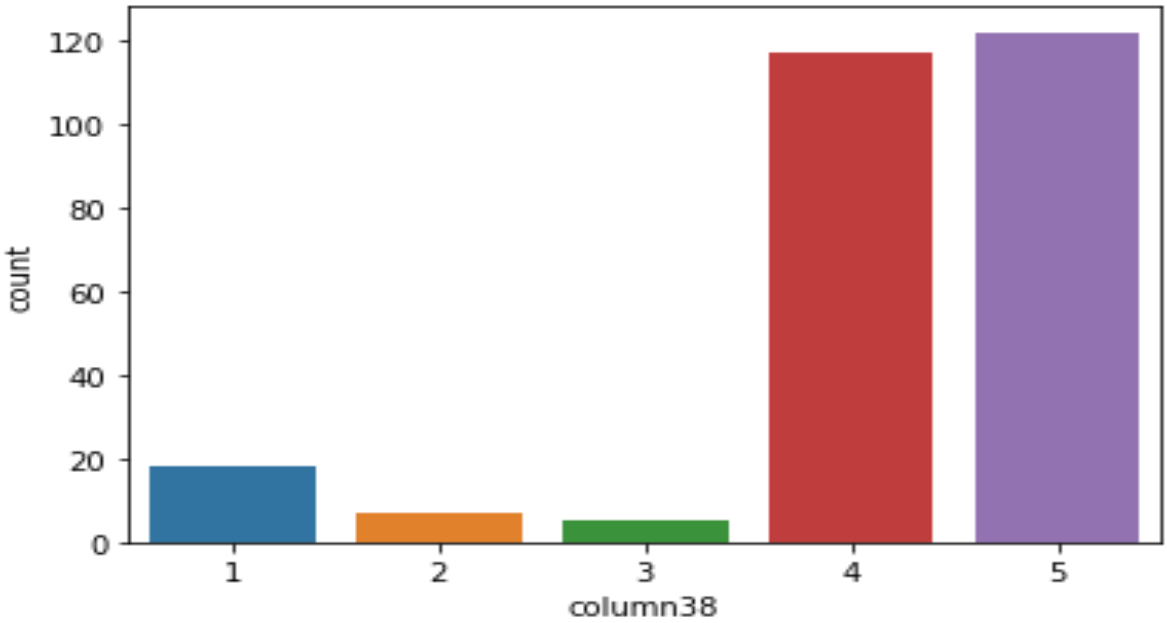
36. Column37-

```
5    164
4     54
3     40
2     11
```



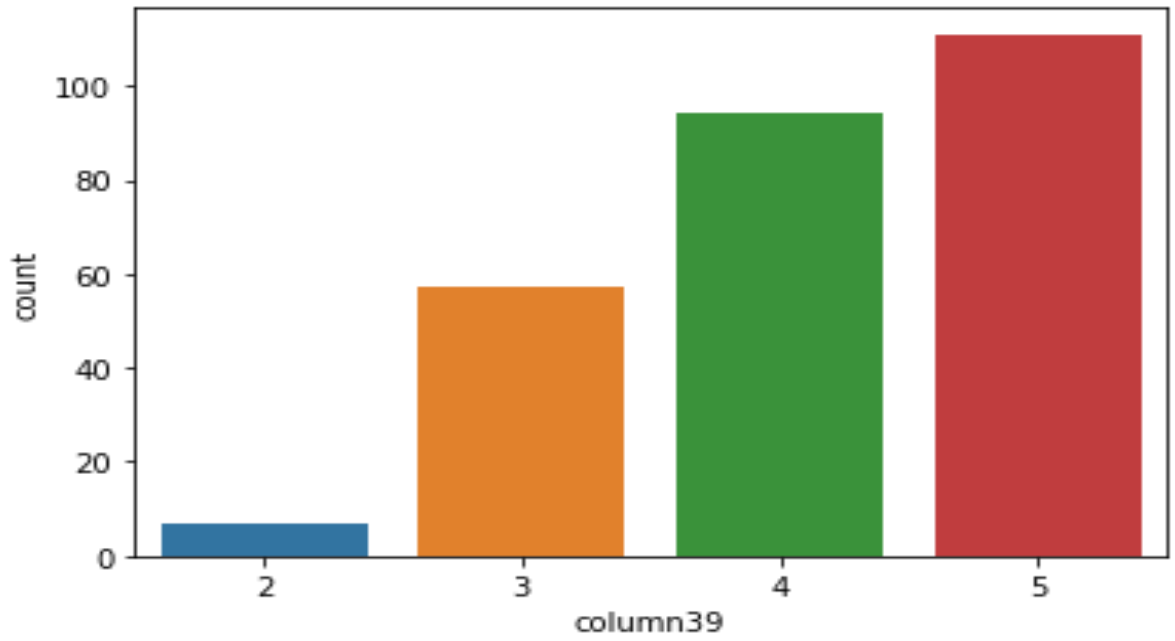
37. Column38-

```
5    122
4    117
1     18
2      7
3      5
```



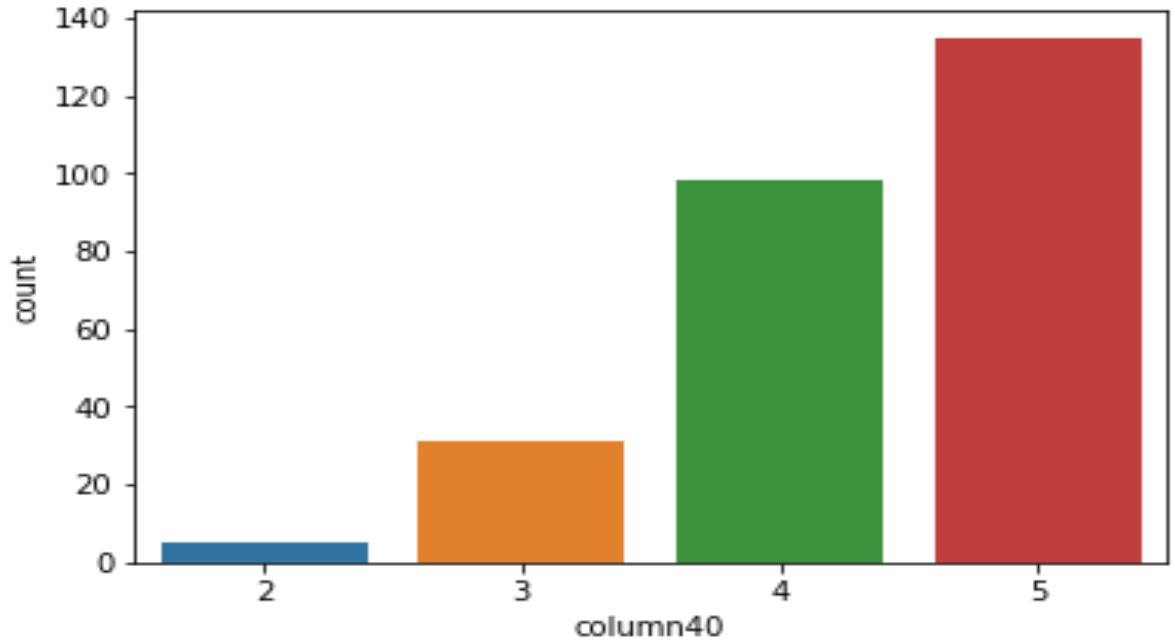
38. Column39-

```
5    111
4     94
3     57
2      7
```



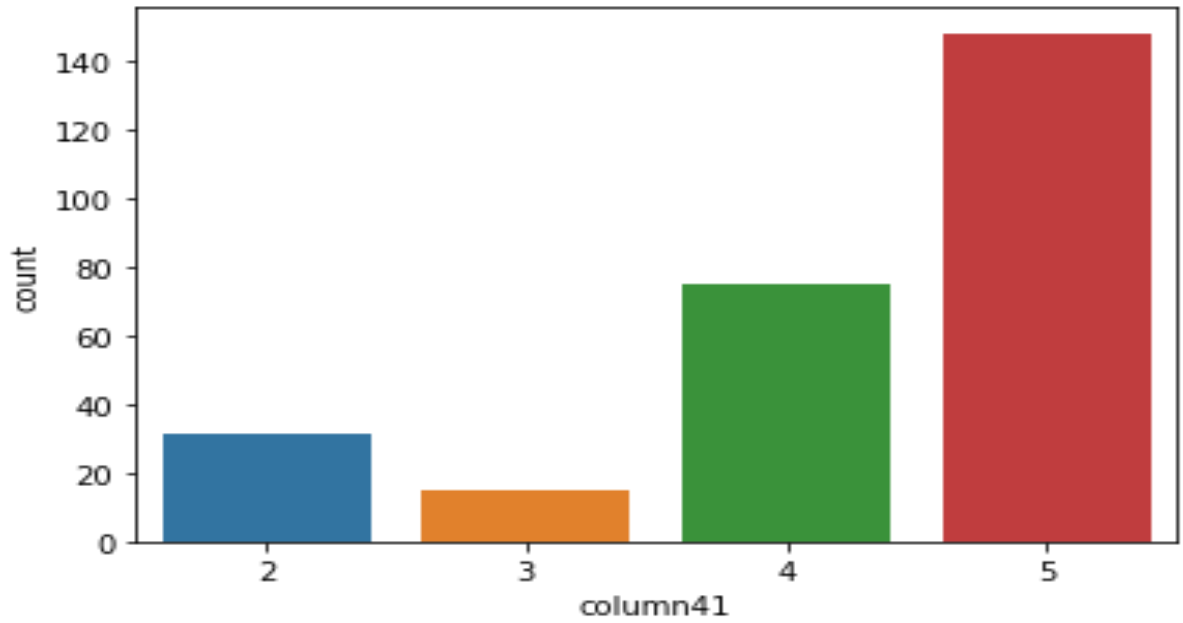
39. Column40-

```
5    135
4     98
3     31
2      5
```



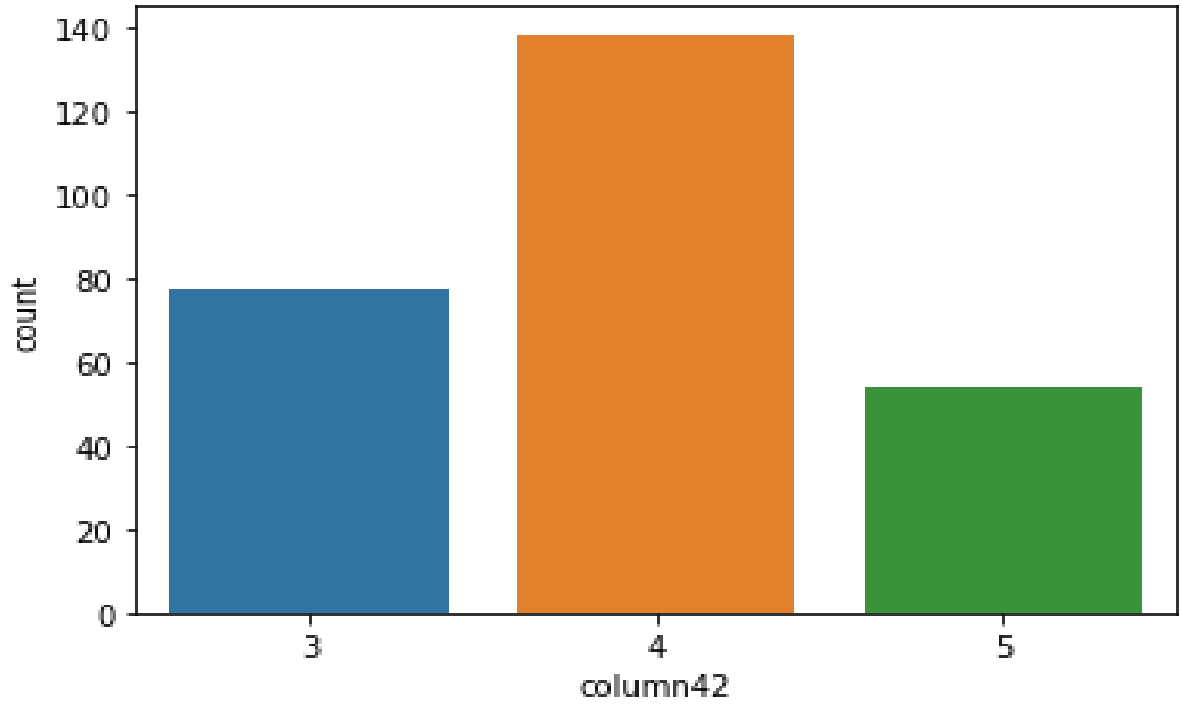
40. Column41-

```
5    148
4     75
2     31
3     15
```



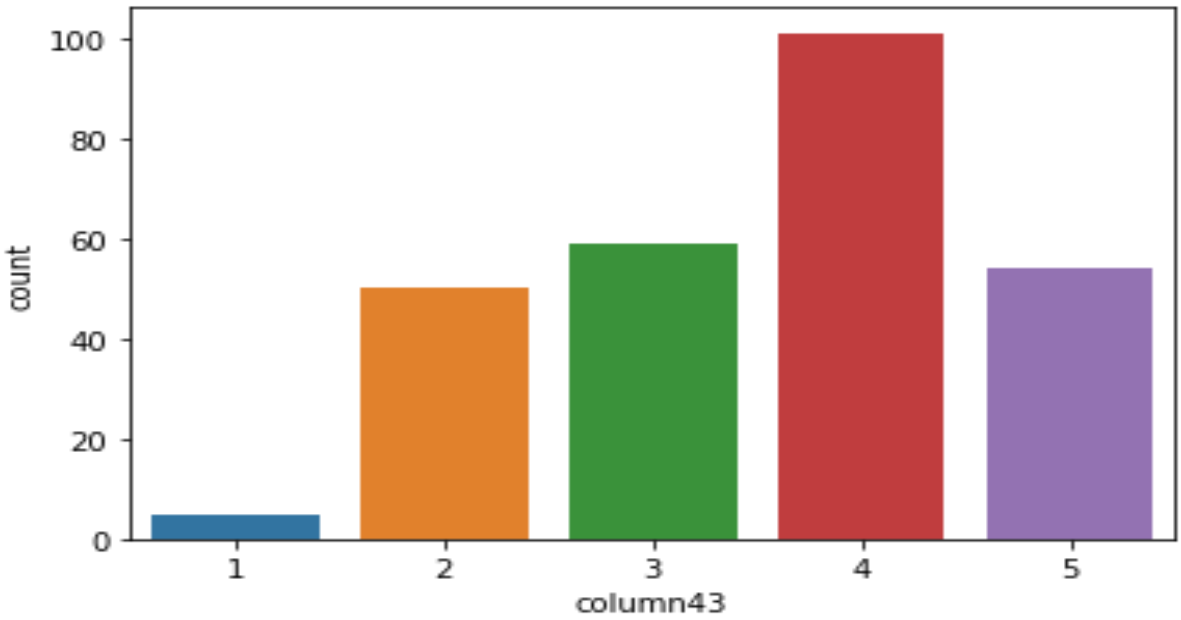
41. Column42-

```
4    138
3     77
5     54
```



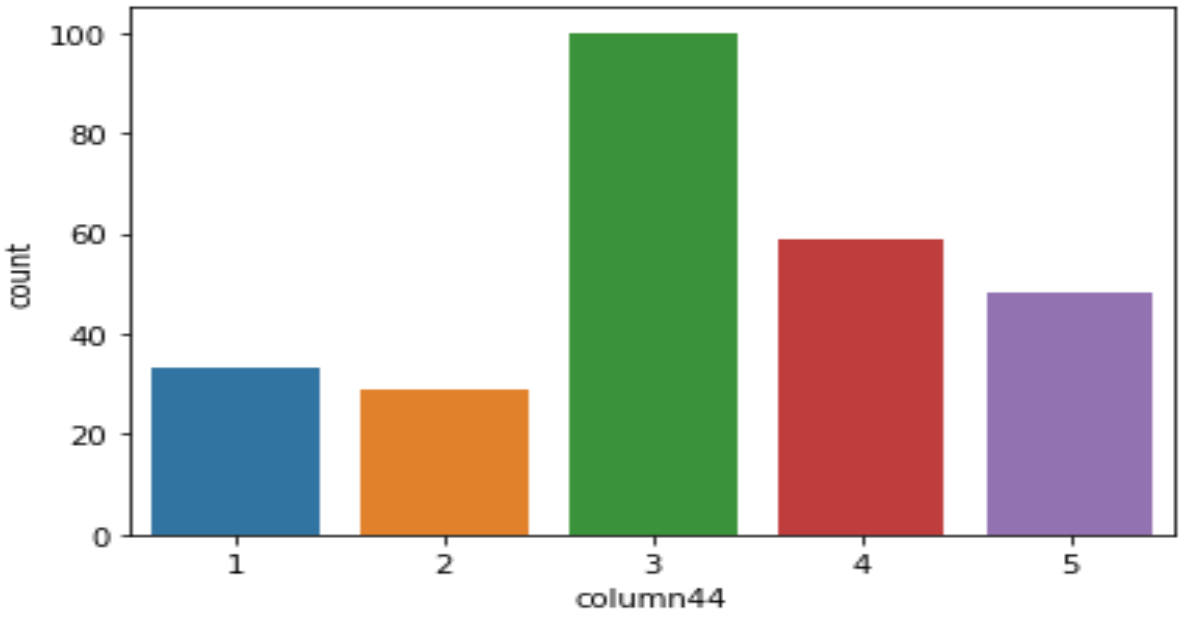
42. Column43-

```
4    101
3     59
5     54
2     50
1      5
```



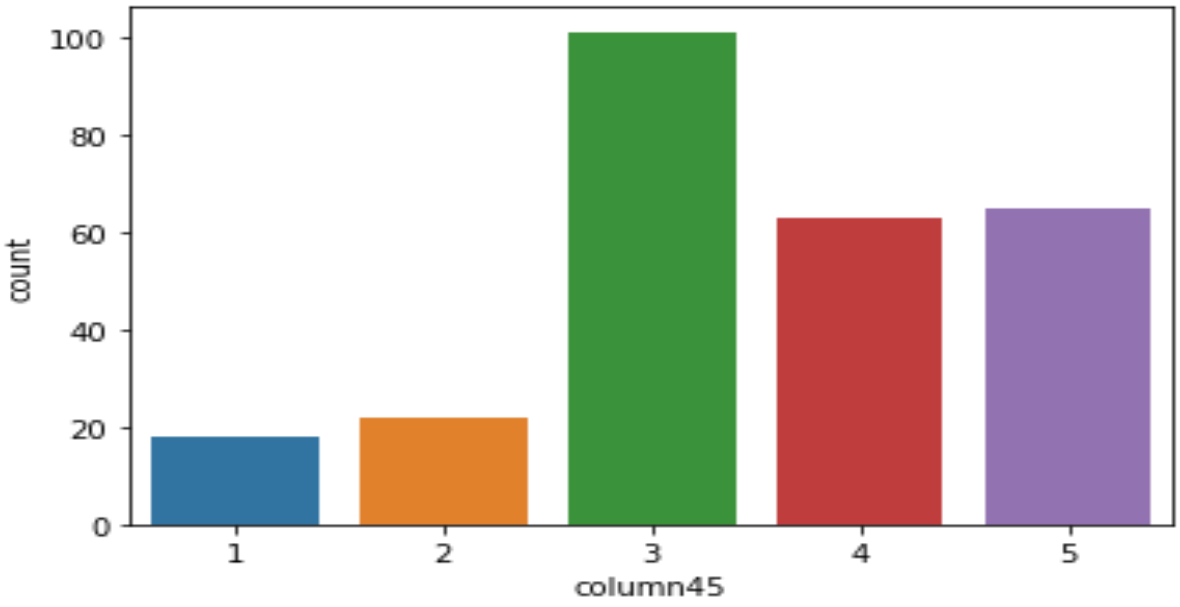
43. Column44-

```
3    100
4     59
5     48
1     33
2     29
```



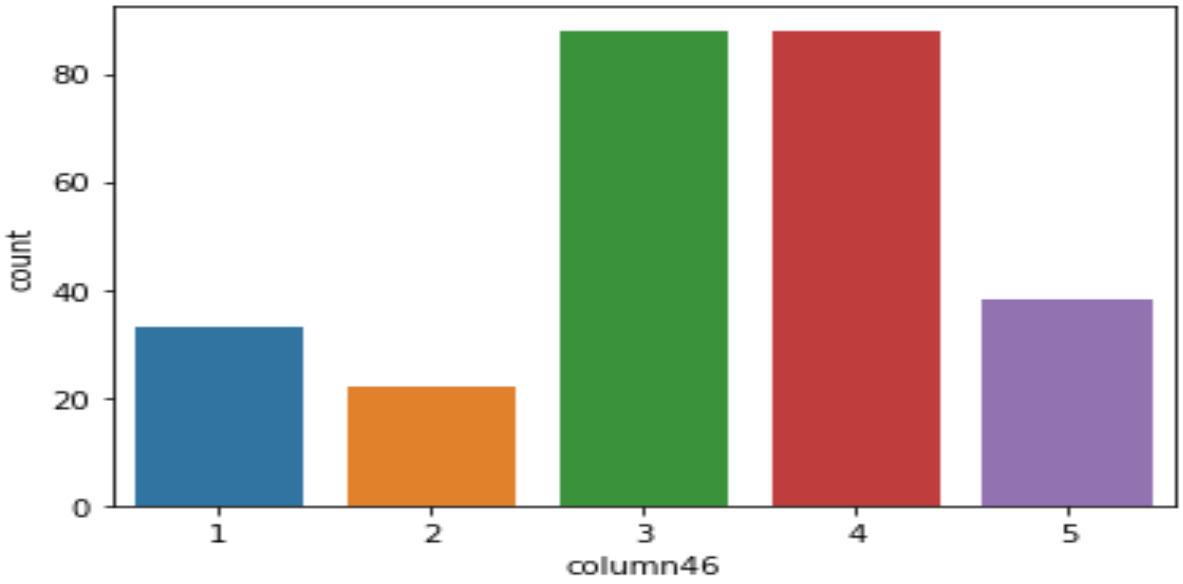
44. Column45-

```
3    101
5     65
4     63
2     22
1     18
```



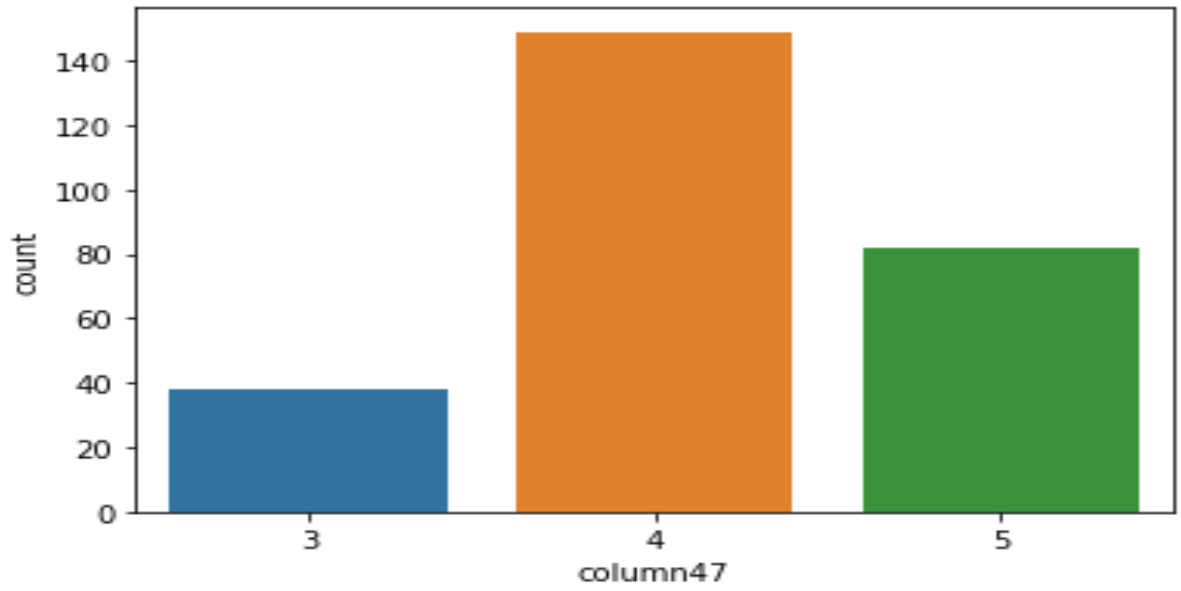
45. Column46-

```
4    88
3    88
5    38
1    33
2    22
```



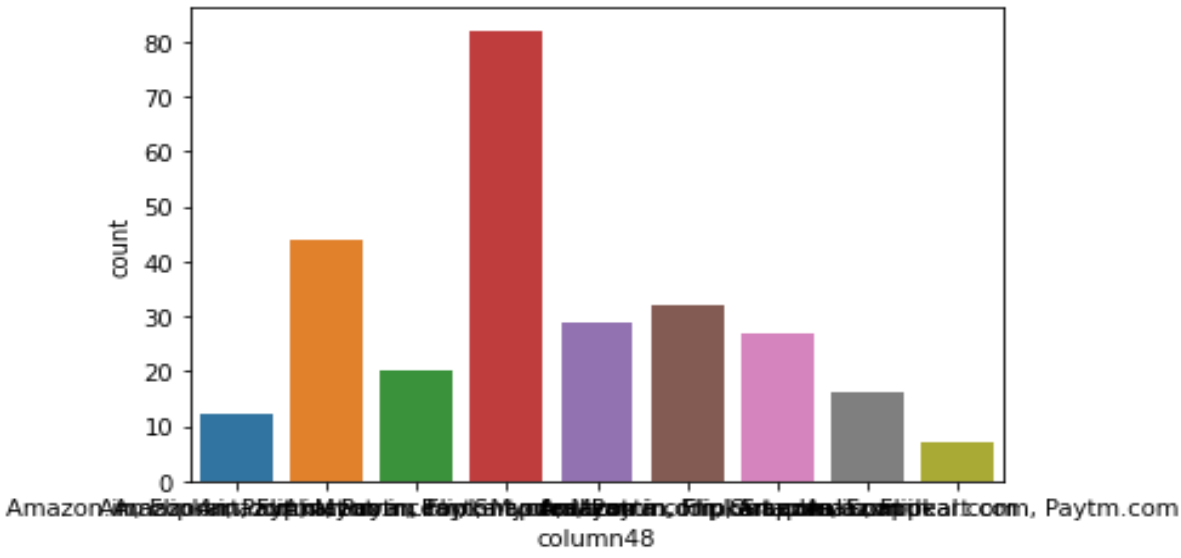
46. Column47-

```
4    149
5     82
3     38
```



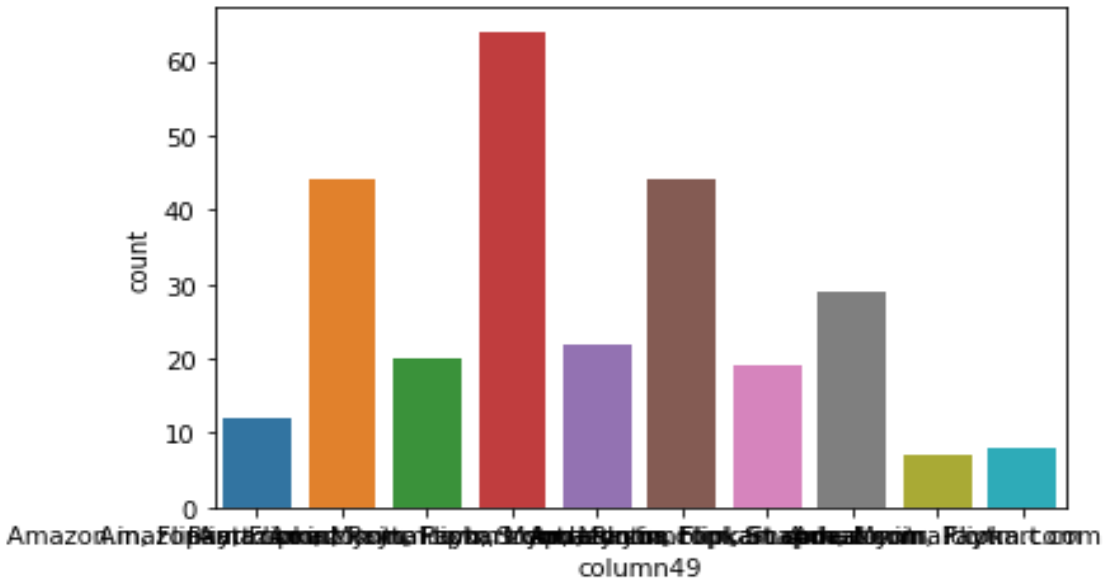
47. Column48-

```
Amazon.in, Flipkart.com, Paytm.com, Myntra.com, Snapdeal.com    82
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com               44
Amazon.in, Flipkart.com                                         32
Amazon.in, Flipkart.com, Paytm.com, Snapdeal.com                29
Amazon.in, Flipkart.com, Snapdeal.com                           27
Amazon.in, Paytm.com, Myntra.com                                20
Amazon.in                                                       16
Amazon.in, Paytm.com                                            12
Amazon.in, Flipkart.com, Paytm.com                               7
```



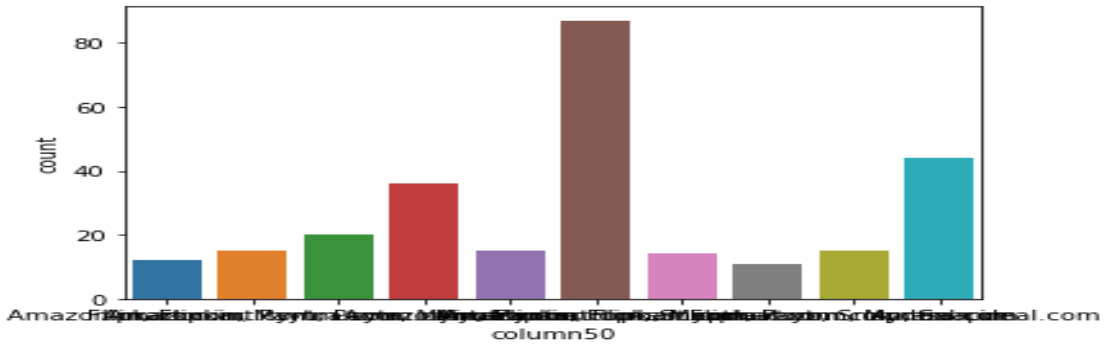
48. Column49-

```
Amazon.in, Flipkart.com, Paytm.com, Myntra.com, Snapdeal.com    64
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com               44
Amazon.in, Flipkart.com                                         44
Amazon.in                                                       29
Amazon.in, Flipkart.com, Paytm.com, Snapdeal.com                22
Amazon.in, Paytm.com, Myntra.com                                20
Amazon.in, Flipkart.com, Myntra.com                             19
Paytm.com                                                       12
Flipkart.com                                                     8
Amazon.in, Paytm.com                                             7
```



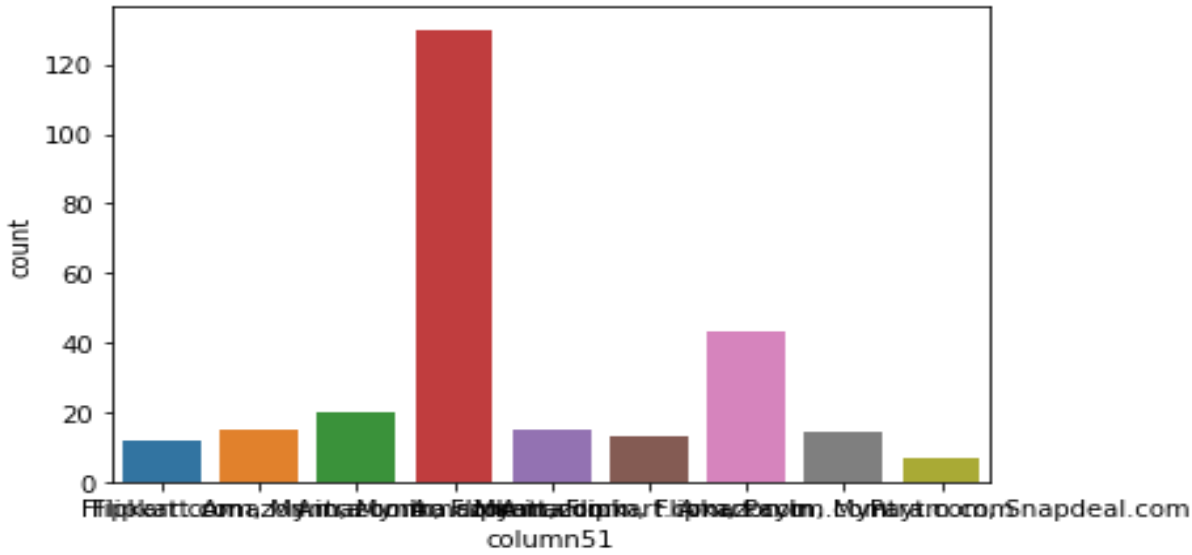
49. Column50-

```
Amazon.in, Flipkart.com                                         87
Amazon.in                                                       44
Amazon.in, Flipkart.com, Paytm.com, Myntra.com, Snapdeal.com    36
Amazon.in, Paytm.com, Myntra.com                                20
Amazon.in, Myntra.com                                           15
Myntra.com                                                      15
Flipkart.com, Myntra.com                                        15
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com               14
Flipkart.com                                                    12
Amazon.in, Flipkart.com, Paytm.com, Snapdeal.com                11
```



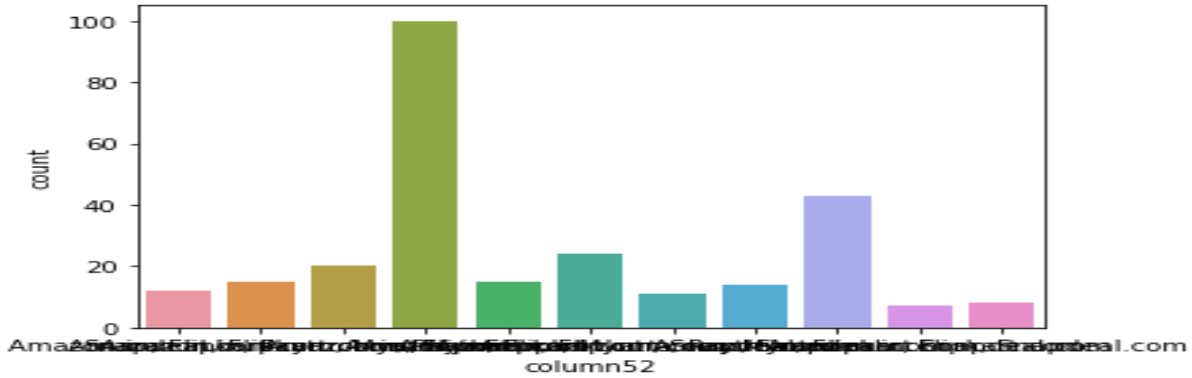
50. Column51-

```
Amazon.in, Flipkart.com                                  130
Amazon.in                                                 43
Amazon.in, Myntra.com                                     20
Flipkart.com, Myntra.com                                  15
Myntra.com                                                15
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com         14
Amazon.in, Flipkart.com, Paytm.com                        13
Flipkart.com                                              12
Paytm.com                                                  7
```



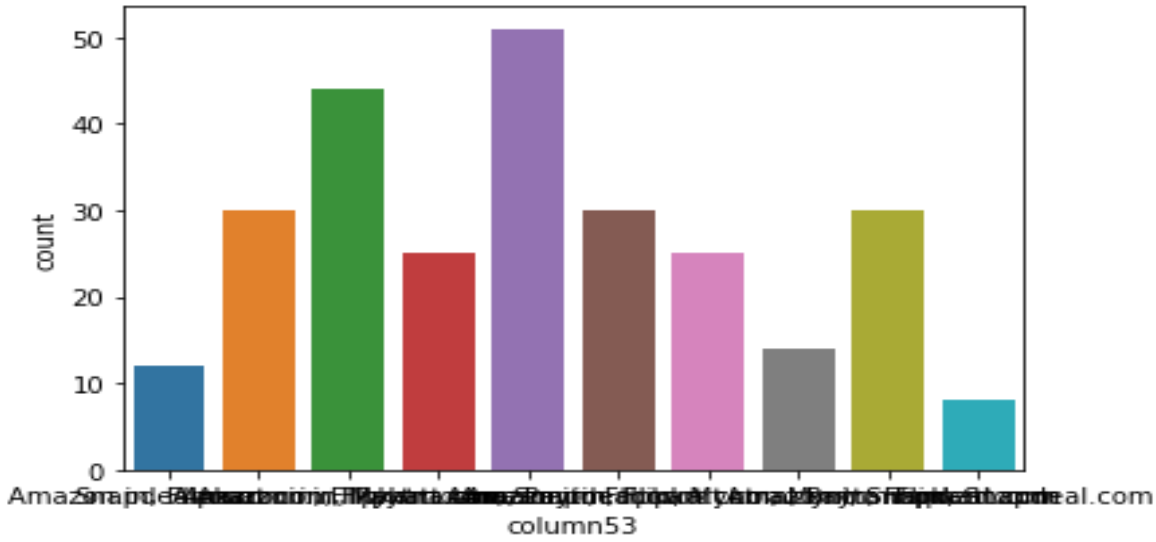
51. Column52-

```
Amazon.in, Flipkart.com                                      100
Amazon.in                                                     43
Amazon.in, Flipkart.com, Paytm.com                            24
Amazon.in, Paytm.com, Myntra.com                              20
Amazon.in, Flipkart.com, Myntra.com                           15
Amazon.in, Flipkart.com, Paytm.com, Myntra.com, Snapdeal.com  15
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com             14
Snapdeal.com                                                  12
Flipkart.com, Snapdeal.com                                    11
Flipkart.com                                                   8
Amazon.in, Flipkart.com, Snapdeal.com                          7
```



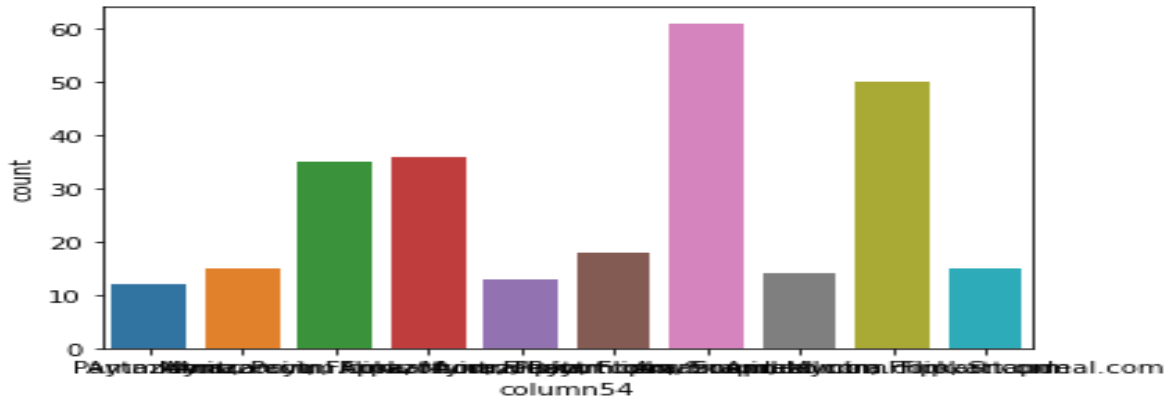
52. Column53-

```
Amazon.in                                                          51
Amazon.in, Paytm.com                                               44
Amazon.in, Flipkart.com, Myntra.com                                30
Amazon.in, Flipkart.com, Paytm.com, Myntra.com, Snapdeal.com       30
Amazon.in, Flipkart.com                                            30
Amazon.in, Flipkart.com, Snapdeal.com                              25
Amazon.in, Flipkart.com, Paytm.com                                 25
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com                  14
Snapdeal.com                                                       12
Flipkart.com                                                        8
```



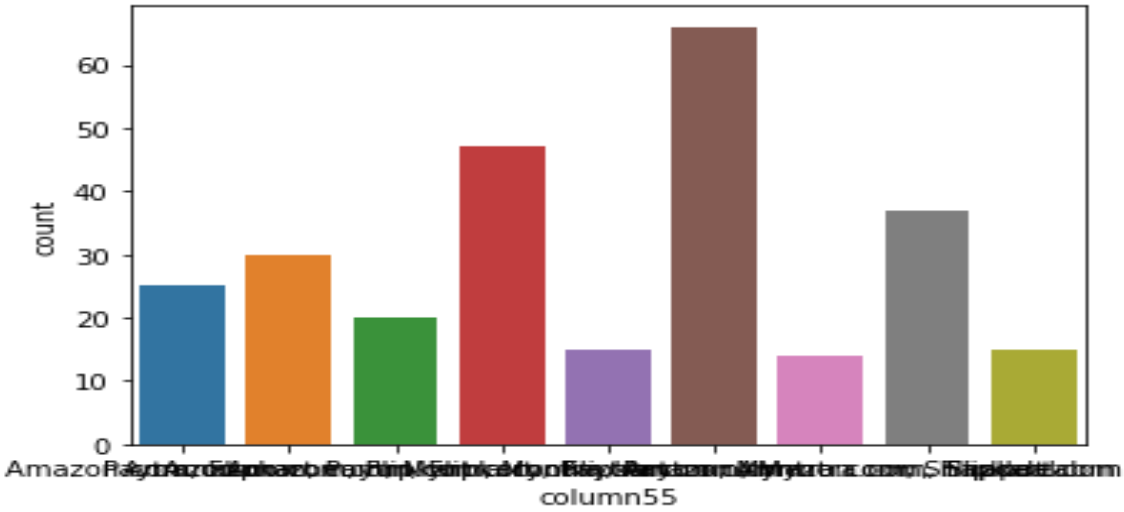
53. Column54-

```
Amazon.in                                                61
Amazon.in, Flipkart.com                                  50
Amazon.in, Flipkart.com, Paytm.com                       36
Amazon.in, Paytm.com, Myntra.com                         35
Amazon.in, Flipkart.com, Snapdeal.com                    18
Myntra.com                                               15
Flipkart.com                                             15
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com        14
Amazon.in, Flipkart.com, Paytm.com, Snapdeal.com         13
Paytm.com                                                12
```



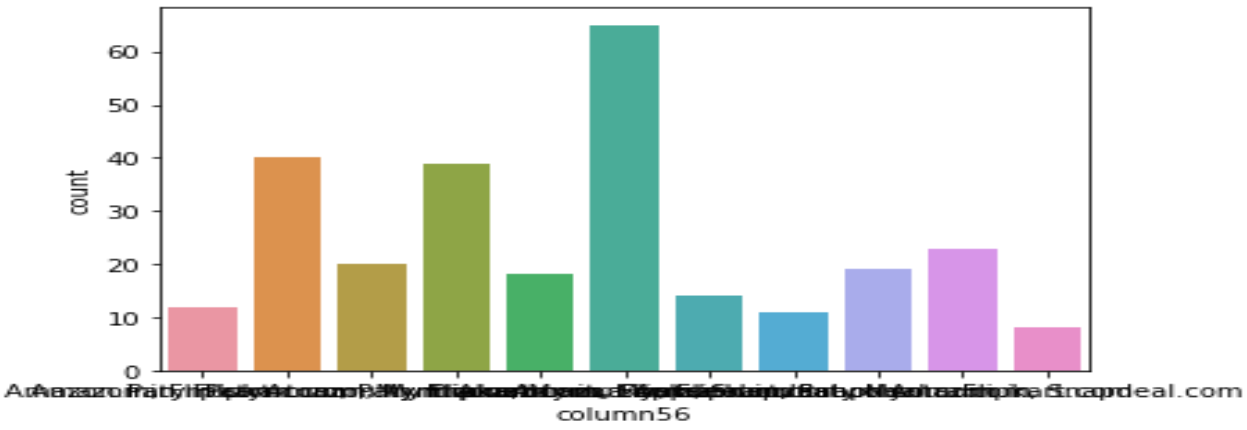
54. Column55-

```
Amazon.com                                                    66
Amazon.com, Flipkart.com, Paytm.com                           47
Amazon.com, Flipkart.com                                      37
Amazon.com, Flipkart.com, Myntra.com                          30
Paytm.com                                                     25
Amazon.com, Paytm.com, Myntra.com                             20
Amazon.com, Flipkart.com, Paytm.com, Myntra.com, Snapdeal     15
Flipkart.com                                                  15
Flipkart.com, Myntra.com, Snapdeal                            14
```



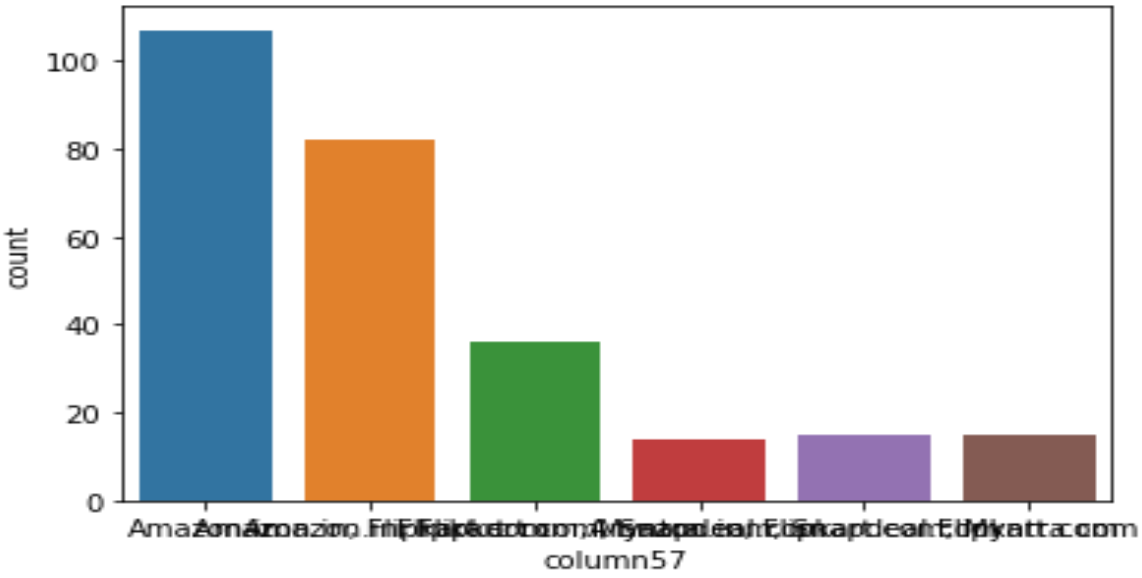
55. Column56-

```
Amazon.in, Flipkart.com                                    65
Amazon.in, Flipkart.com, Myntra.com                        40
Amazon.in, Flipkart.com, Patym.com, Myntra.com, Snapdeal.com    39
Amazon.in                                                  23
Patym.com, Myntra.com                                      20
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com          19
Amazon.in, Flipkart.com, Snapdeal.com                      18
Flipkart.com, Myntra.com, Snapdeal.com                     14
Patym.com                                                  12
Amazon.in, Patym.com                                       11
Flipkart.com                                                8
```



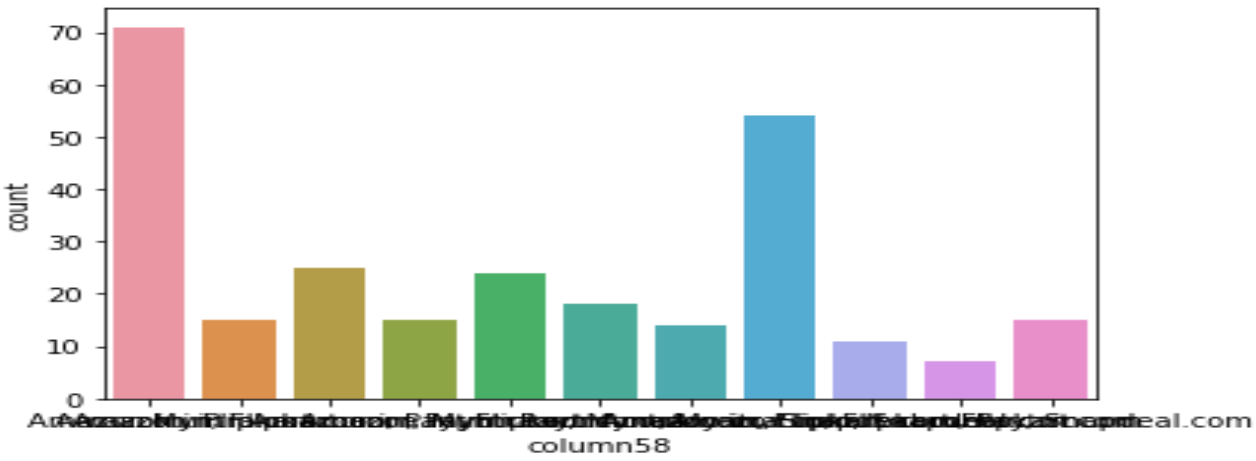
56. Column57-

```
Amazon.in                                    107
Amazon.in, Flipkart.com                       82
Amazon.in, Flipkart.com, Snapdeal.com         36
Amazon.in, Flipkart.com, Myntra.com           15
Flipkart.com                                  15
Flipkart.com, Myntra.com, Snapdeal.com        14
```



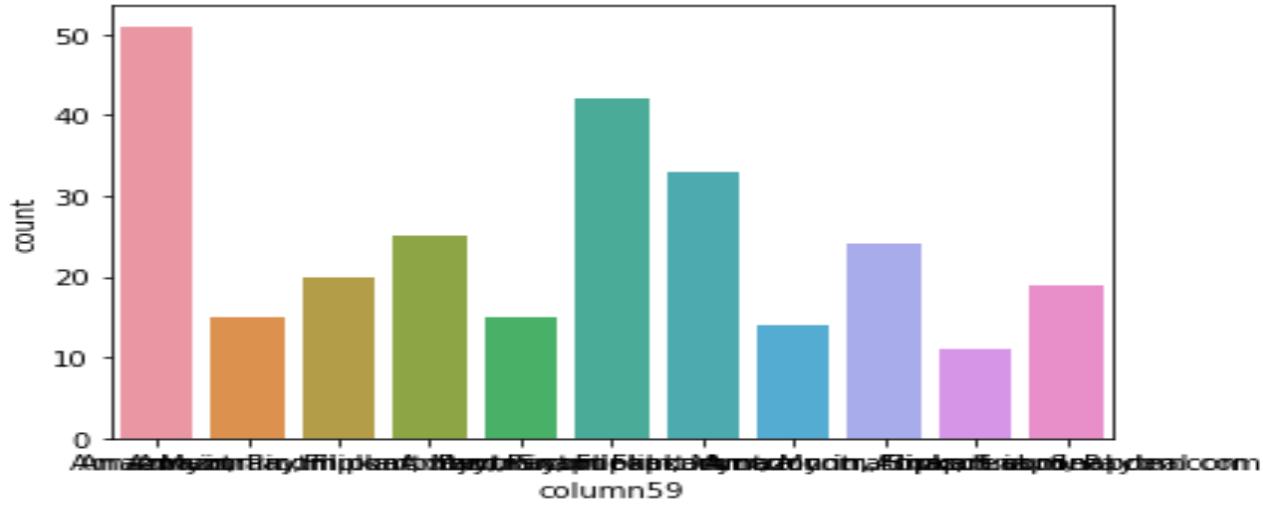
57. Column58-

```
Amazon.in                                                        71
Amazon.in, Flipkart.com                                          54
Amazon.in, Flipkart.com, Myntra.com                              25
Amazon.in, Flipkart.com, Paytm.com, Myntra.com, Snapdeal.com     24
Paytm.com                                                        18
Myntra.com                                                       15
Amazon.in, Paytm.com                                             15
Flipkart.com                                                     15
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com                14
Amazon.in, Flipkart.com, Paytm.com                               11
Amazon.in, Flipkart.com, Snapdeal.com                             7
```



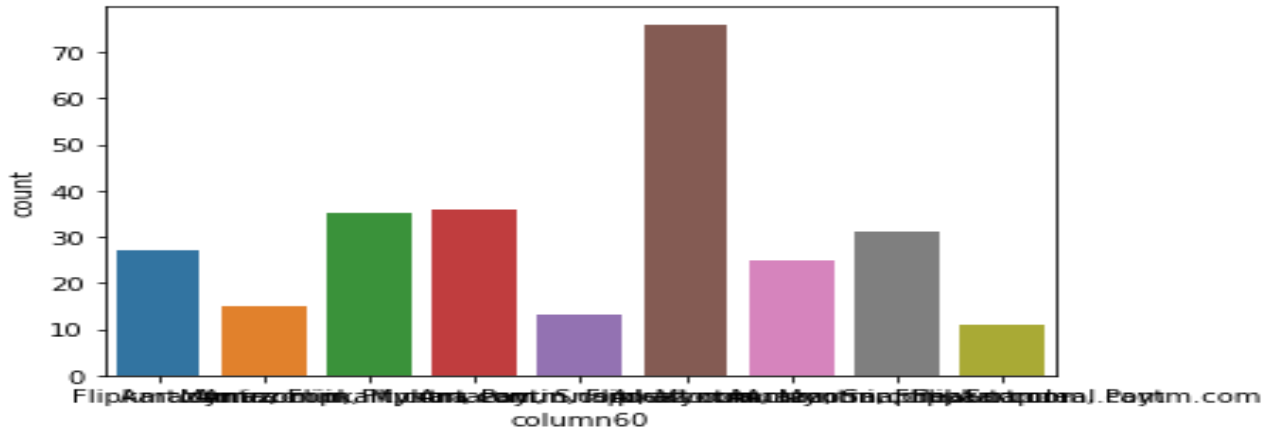
58. Column59-

```
Amazon.in                                                          51
Amazon.in, Flipkart.com, Paytm.com, Myntra.com, Snapdeal.com       42
Flipkart.com                                                       33
Amazon.in, Flipkart.com, Snapdeal.com                              25
Amazon.in, Flipkart.com                                            24
Amazon.in, Paytm.com, Myntra.com                                   20
Amazon.in, Snapdeal.com                                            19
Myntra.com                                                         15
Paytm.com                                                          15
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com                  14
Amazon.in, Flipkart.com, Paytm.com                                 11
```



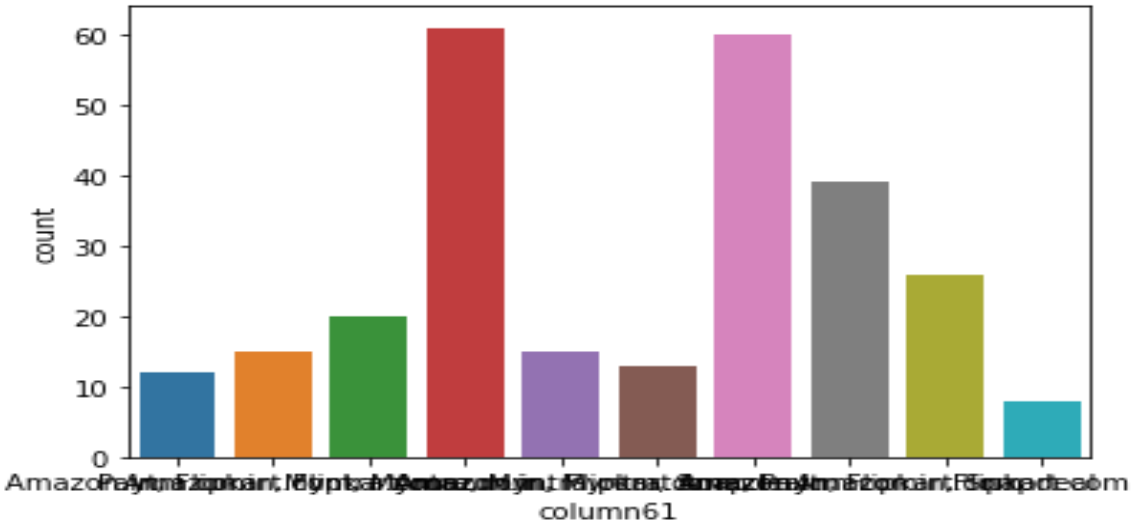
59. Column60-

```
Amazon.in                                                          76
Amazon.in, Flipkart.com, Snapdeal.com                              36
Amazon.in, Myntra.com                                              35
Amazon.in, Flipkart.com                                            31
Flipkart.com                                                       27
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com                  25
Myntra.com                                                         15
Amazon.in, Flipkart.com, Paytm.com, Myntra.com, Snapdeal.com       13
Amazon.in, Flipkart.com, Paytm.com                                 11
```



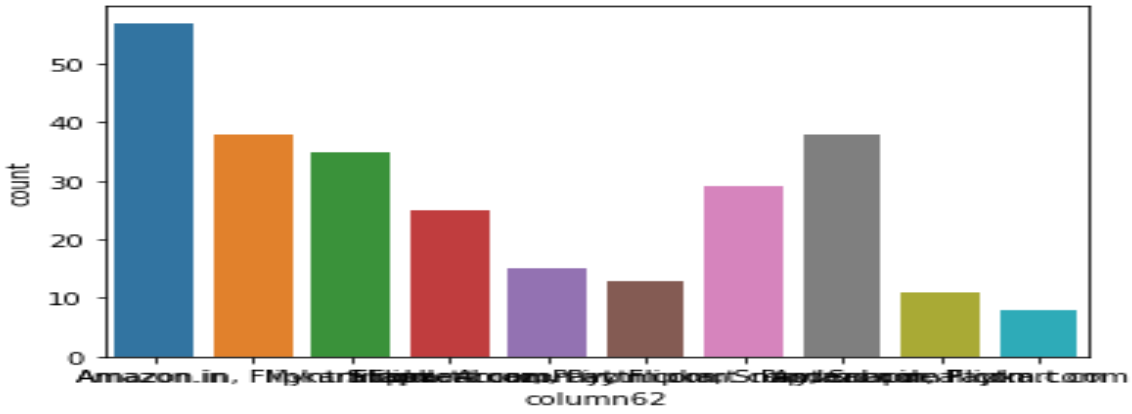
60. Column61-

```
Amazon.in, Flipkart.com, Myntra.com, Snapdeal    61
Amazon.in                                        60
Amazon.in, Flipkart.com                          39
Amazon.in, Snapdeal                              26
Myntra.com                                       20
Amazon.in, Flipkart.com, Myntra.com              15
Amazon.in, Myntra.com                            15
Amazon.in, Flipkart.com, Paytm.com               13
Paytm.com                                        12
Flipkart.com                                      8
```



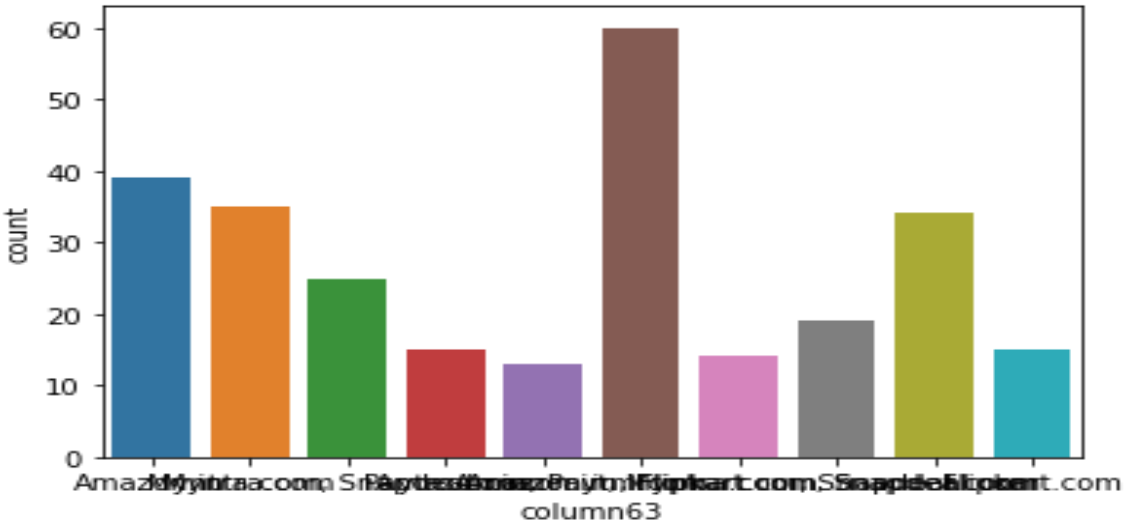
61. Column62-

```
Amazon.in                                57
Amazon.in, Flipkart.com                  38
Paytm.com                                38
Myntra.com                               35
Amazon.in, Flipkart.com, Snapdeal.com    29
Snapdeal.com                             25
Flipkart.com, Paytm.com                  15
Flipkart.com, Paytm.com, Snapdeal.com    13
Amazon.in, Paytm.com                     11
Flipkart.com                              8
```



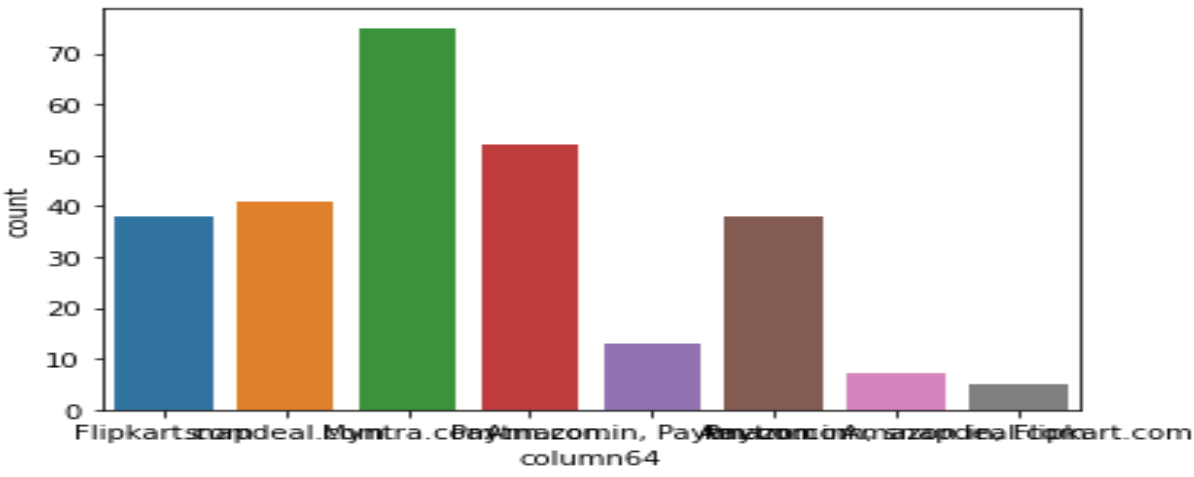
62. Column63-

```
Amazon.in, Flipkart.com                 60
Amazon.in                               39
Myntra.com                              35
Snapdeal.com                            34
Myntra.com, Snapdeal.com                25
Flipkart.com, Snapdeal.com              19
Paytm.com                               15
Flipkart.com                            15
Amazon.in, Myntra.com, Snapdeal.com     14
Amazon.in, Paytm.com                    13
```



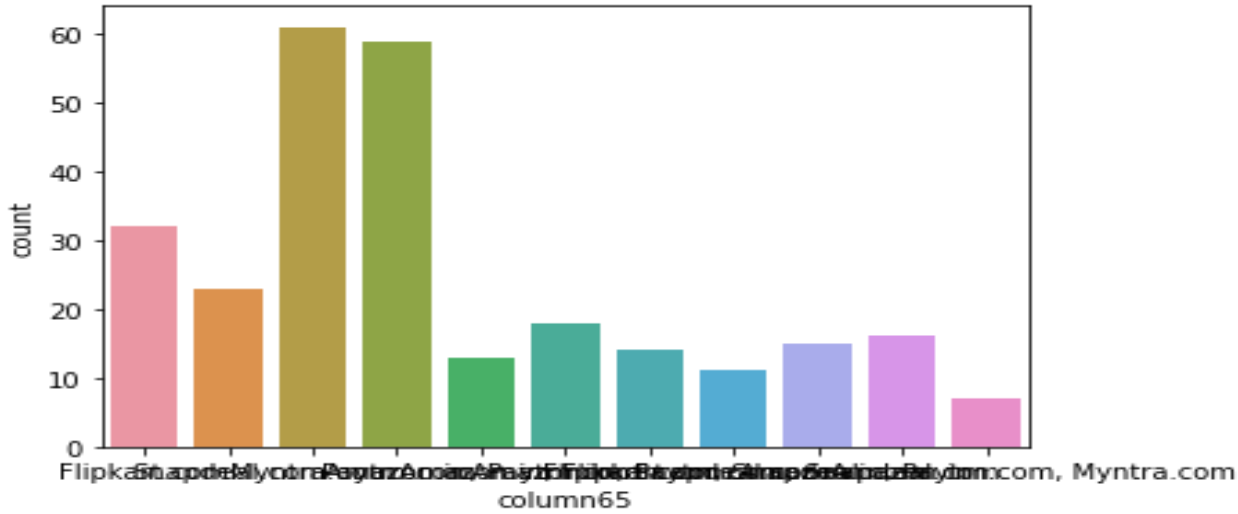
63. Column64-

```
Myntra.com                 75
Paytm.com                  52
snapdeal.com               41
Flipkart.com               38
Amazon.in                  38
Amazon.in, Paytm.com       13
Paytm.com, snapdeal.com     7
Amazon.in, Flipkart.com     5
```



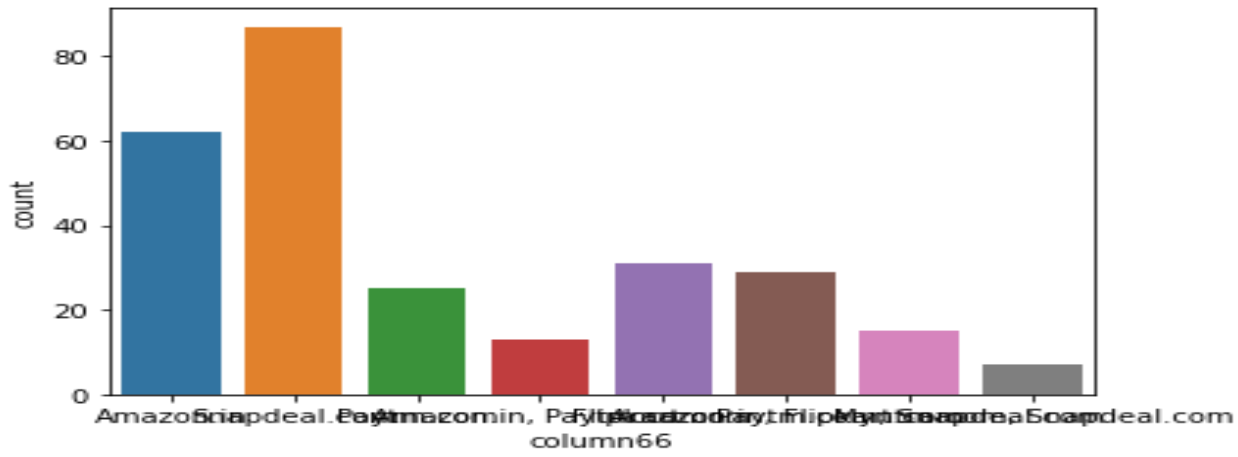
64. Column65-

```
Myntra.com                          61
Paytm.com                           59
Flipkart.com                        32
Snapdeal.com                        23
Amazon.in, Flipkart.com             18
Amazon.in                           16
Paytm.com, Snapdeal.com             15
Amazon.in, Snapdeal.com             14
Amazon.in, Paytm.com                13
Flipkart.com, Snapdeal.com          11
Amazon.in, Paytm.com, Myntra.com     7
```



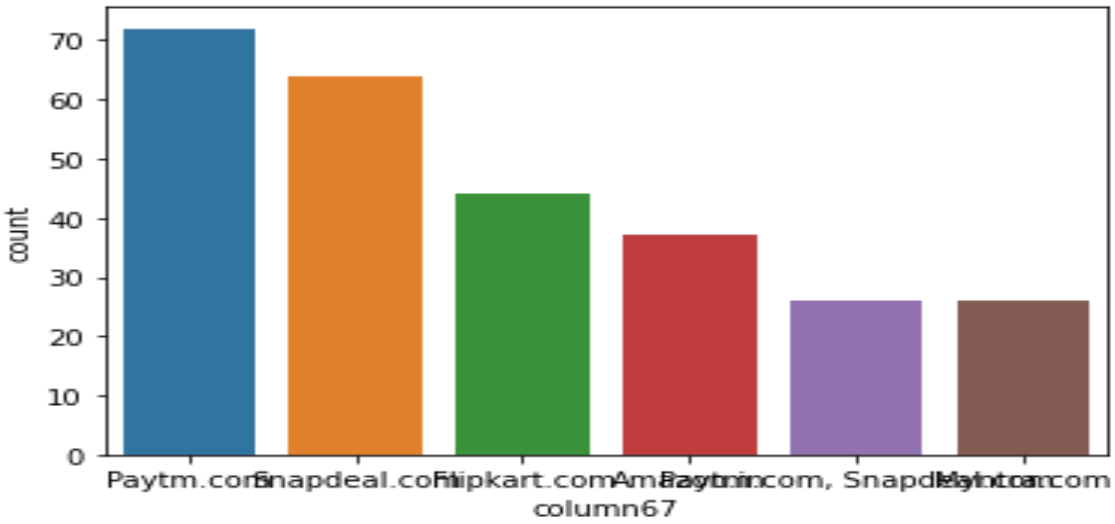
65. Column66-

```
Snapdeal.com                87
Amazon.in                   62
Flipkart.com                31
Amazon.in, Flipkart.com     29
Paytm.com                   25
Paytm.com, Snapdeal.com     15
Amazon.in, Paytm.com        13
Myntra.com, Snapdeal.com     7
```



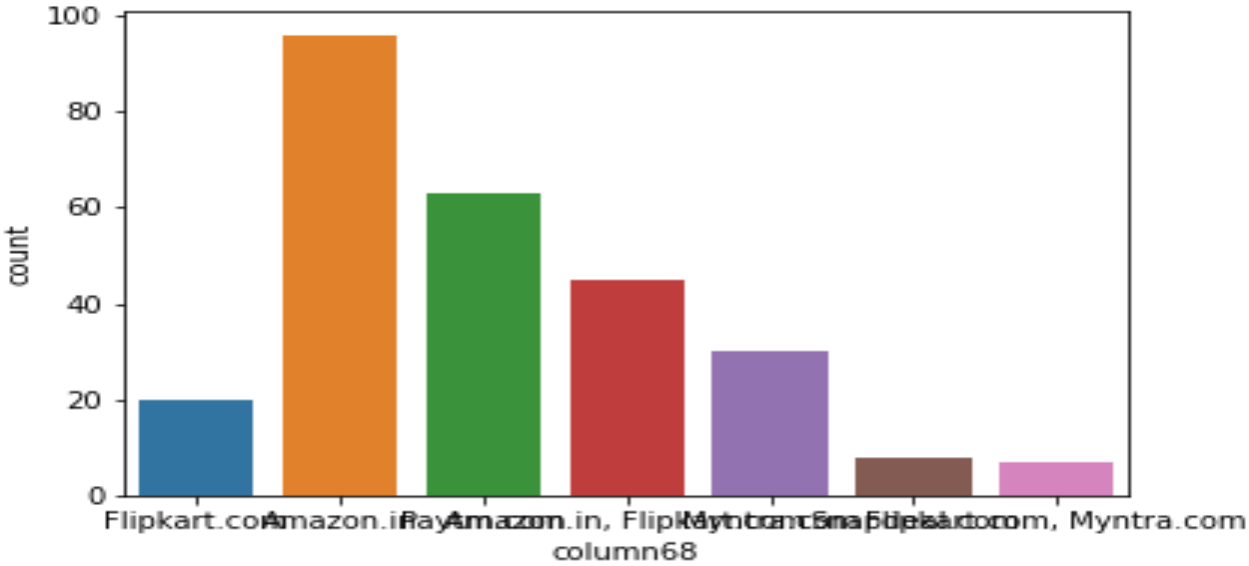
66. Column67-

```
Paytm.com                    72
Snapdeal.com                 64
Flipkart.com                 44
Amazon.in                    37
Paytm.com, Snapdeal.com      26
Myntra.com                   26
```



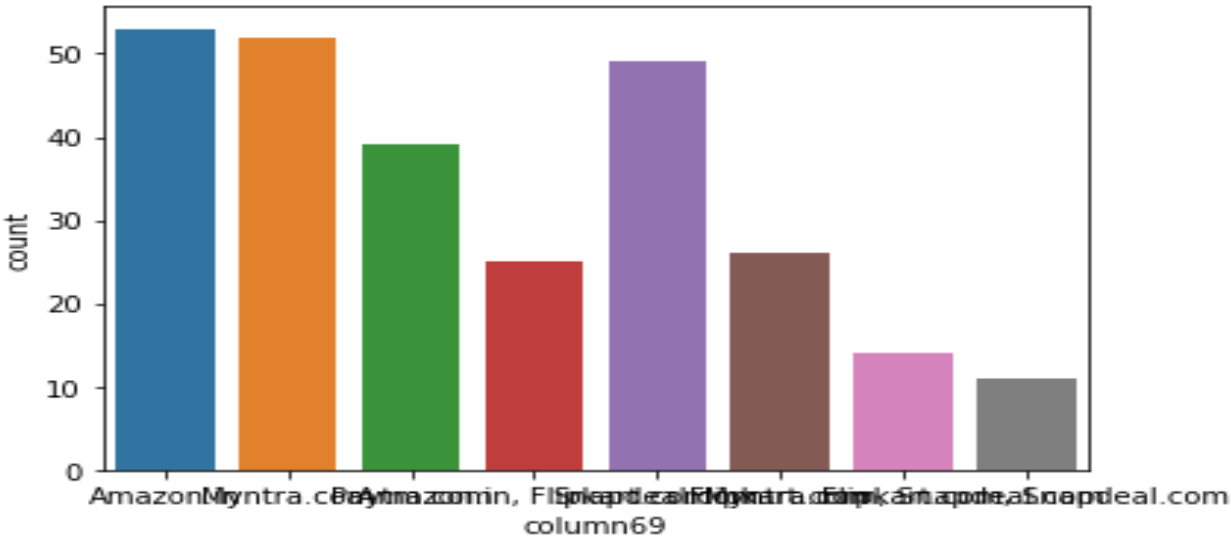
67. Column68-

```
Amazon.in                    96
Paytm.com                    63
Amazon.in, Flipkart.com      45
Myntra.com                   30
Flipkart.com                 20
Snapdeal.com                  8
Flipkart.com, Myntra.com      7
```



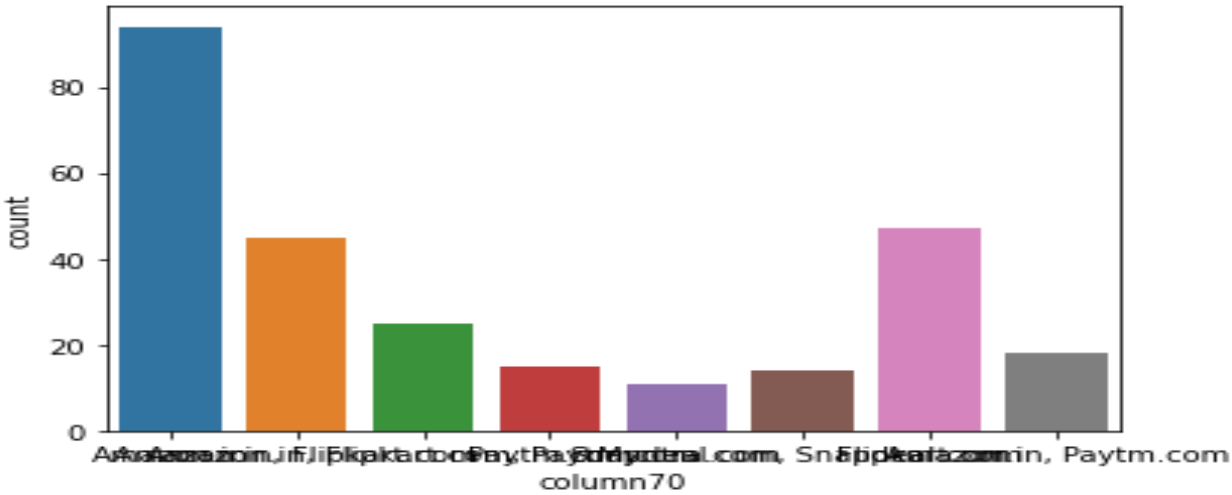
68. Column69-

```
Amazon.in                          53
Myntra.com                         52
Snapdeal.com                       49
Paytm.com                          39
Flipkart.com                       26
Amazon.in, Flipkart.com            25
Myntra.com, Snapdeal.com           14
Flipkart.com, Snapdeal.com         11
```



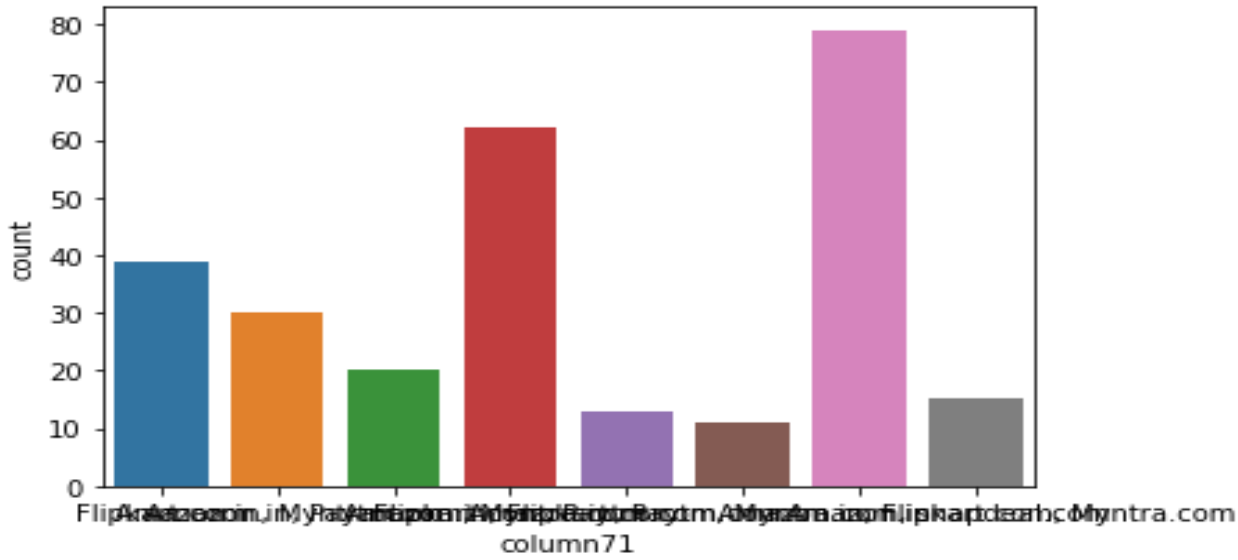
69. Column70-

```
Amazon.in                                94
Flipkart.com                             47
Amazon.in, Flipkart.com                  45
Amazon.in, Flipkart.com, Paytm.com       25
Amazon.in, Paytm.com                     18
Paytm.com                                15
Myntra.com, Snapdeal.com                 14
Snapdeal.com                             11
```



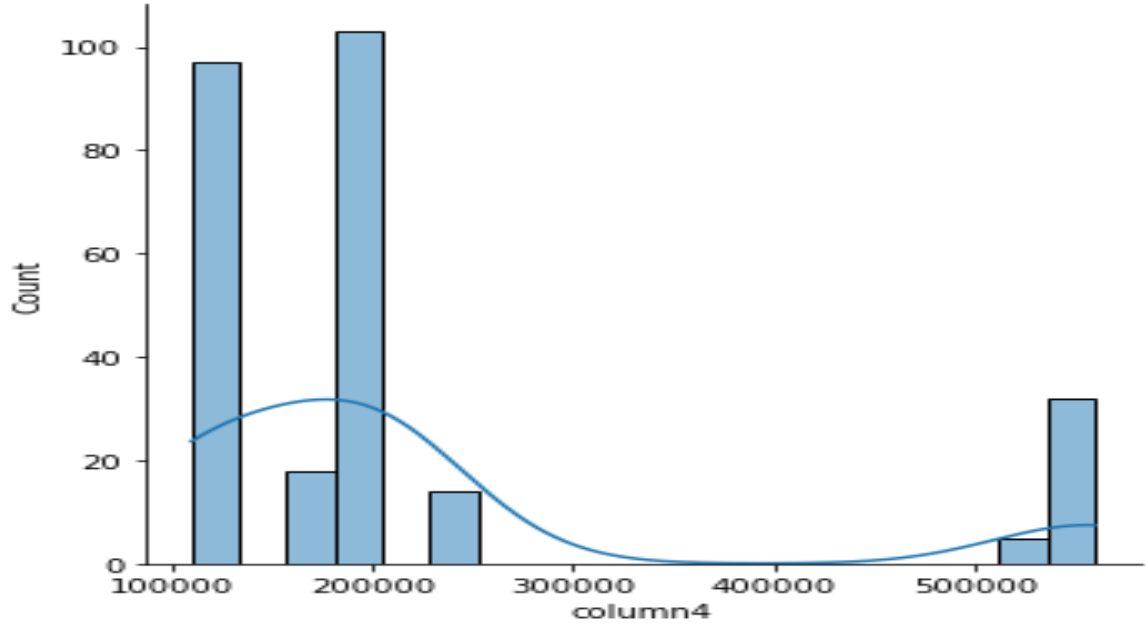
70. Column71-

```
Amazon.in                                                79
Amazon.in, Flipkart.com                                  62
Flipkart.com                                             39
Amazon.in, Myntra.com                                    30
Amazon.in, Paytm.com, Myntra.com                         20
Amazon.in, Flipkart.com, Myntra.com                      15
Amazon.in, Paytm.com                                     13
Flipkart.com, Paytm.com, Myntra.com, snapdeal.com        11
```



We see from the above graphs, that Amazon.in has the highest customer retention rate excelling in most of the categories and maximum customers will recommend the e-commerce site to their friends/family.

For Column4 which is continuous data, we will use distplot.



3.3 ENCODING OF DATAFRAME

Encoding a dataframe means changing the data type of a particular column to the required type as the dataset demands. There are various types on Encoding techniques:

I. Classic Encoders

We started with the most basic techniques, classic encoders. As the name suggests, these encoders are well known and widely used. Their concept is also pretty straight-forward.

1) Ordinal Encoding

The ordinal features are features that have an order. This type of data is also called **ordinal data**. Let's look at the Height column in the data frame. The categories are: *very short, short, normal, tall, very tal*l and it makes sense to put them in increasing/decreasing order. By encoding the columns manually, we can significantly boost the model performance.

2) One-hot encoding

Let's look at the column Type. This is **nominal** data which is the opposite to **ordinal data** in the Height column. The easiest way to turn this column into numerical is to use **one-hot encoding** by following the 2 steps

- Split all the categories in one column to different columns
- Put the checkmark *1* for the appropriate location

The get_dummiesfunction in pandas can achieve this goal

3) Binary Encoding

Imagine that you have 200 different categories. One-hot encoding will create 200 different columns. That a lot of columns will takes up a lot of memory. It the meantime, **binary encoding** only need 8 columns. It takes advantage of the binary system and so there might be multiple ones in a row. The logical explanation behind binary encoding is:

- Going down the column, every time it sees a new category, it gives a number, starting from 1 (and the next one is 2)
- Convert these number into binary
- Place each digit in this binary in a separate column.

4) Frequency Encoding

Give each category the **probability** (occurrence/total event). This means that if there are two categories in a column with the same probability (3 `fire` and 3 `bugs`), you cannot really tell the difference between them after being frequency encoded. The trade-off is no new column will be introduced.

5) Hashing Encoding

Hashing converts categorical variables to a higher dimensional space of integers. I won't comment on the methodology much here since scikit-learn explains it very well.

The `n_feature` is the number of columns you want to add. These new columns distinguish the corresponding category. However, you can adjust to any number. This is like binary encoding on steroids!

Advantage

- Deal with large scale categorical features
- High speed and reduced memory usage

Disadvantage

- No inverse-transformation method

Note: What is a good number of returning features? If m are distinct features, and n_feature=k , then m<2^k

II. Contrast encoders

Contrast encoding allows for recentering of categorical variables such that the intercept of a model is not the mean of one level of a category, but instead, the mean of all data points in the data set.

Many people argue that these encodings are not very effective. However, I will leave them here as references.

6) Helmert (reverse) Encoding

Helmert encoding compares each level of a categorical variable to the mean of the subsequent levels.

7) Backward Difference Encoding

In **backward difference encoding**, the mean of the dependent variable for a level is compared with the mean of the dependent variable for the prior level.

III. Bayesian Target Encoders

The general idea of this method is to take the target into account.

Advantage:

- Require minimal effort, only create one column for any number of categories in that feature
- Most favorite encoding scheme in Kaggle competition

Disadvantage:

- Only work for supervised learning (thus, inherently leaky). This means that when dealing with unsupervised data, it gets worse!

- Need regularization for the previous reason

8) Target Encoding

Target-based encoding basically means using the target to encode categorical features. The formula for it is:

$$TE_i = \frac{\text{total true}(y_i)}{\text{total}(y_i)}$$

where y_i is a category and λ is a smoothing function

In the table below, Legendary is our target. Since there are only 1 out of 3 Fire pokemon that are legendary, its value is 1/3. Think about target encoding as frequency encoding on the target!

Note: There are some different version that multiplies the output by a (Laplace)smoothing value. This is to avoid data leakage.

9) Leave One Out Encoding

Leave One Out Encoding (LOOE) is very similar to target encoding but excludes the current row's target when calculating the mean target for a level to reduce the effect of outliers.

Additionally, you can add some (Gaussian) noise to the data to prevent overfitting by changing the sigma value between 0 and 1.

10) Weight of Evidence Encoding

Weight of Evidence Encoding (WoE) is a measure of how much the evidence supports or undermines a hypothesis.

$$WoE = \left[\ln \left(\frac{\text{Distribution of goods} + adj}{\text{Distribution of Bads} + adj} \right) \right]$$

where adj is the adjacent factor is a function that avoids division by 0.

Advantage:

- Work well with logistic regression since WoE transformation has the same logistic scale.
- Can use WoE to compare across feature since their values are standardized.

Disadvantages:

- May lose information due to some category may have the same WoE
- Does not take into account features correlation
- Overfitting

Note: We can adjust the adj factor by changing regularization. (By default it is 1). When setting it equal to 0. You come back to the original WOE and may encounter division by 0

11) James-Stein Encoding (JSE)

This is target encoding but is more robust. **James-Stein (JS)** is it works best for the feature that has a normal distribution. JS is defined by the formula:

$$JS_i = (1 - B) \cdot \text{mean}(y_i) + B \cdot \text{mean}(y)$$

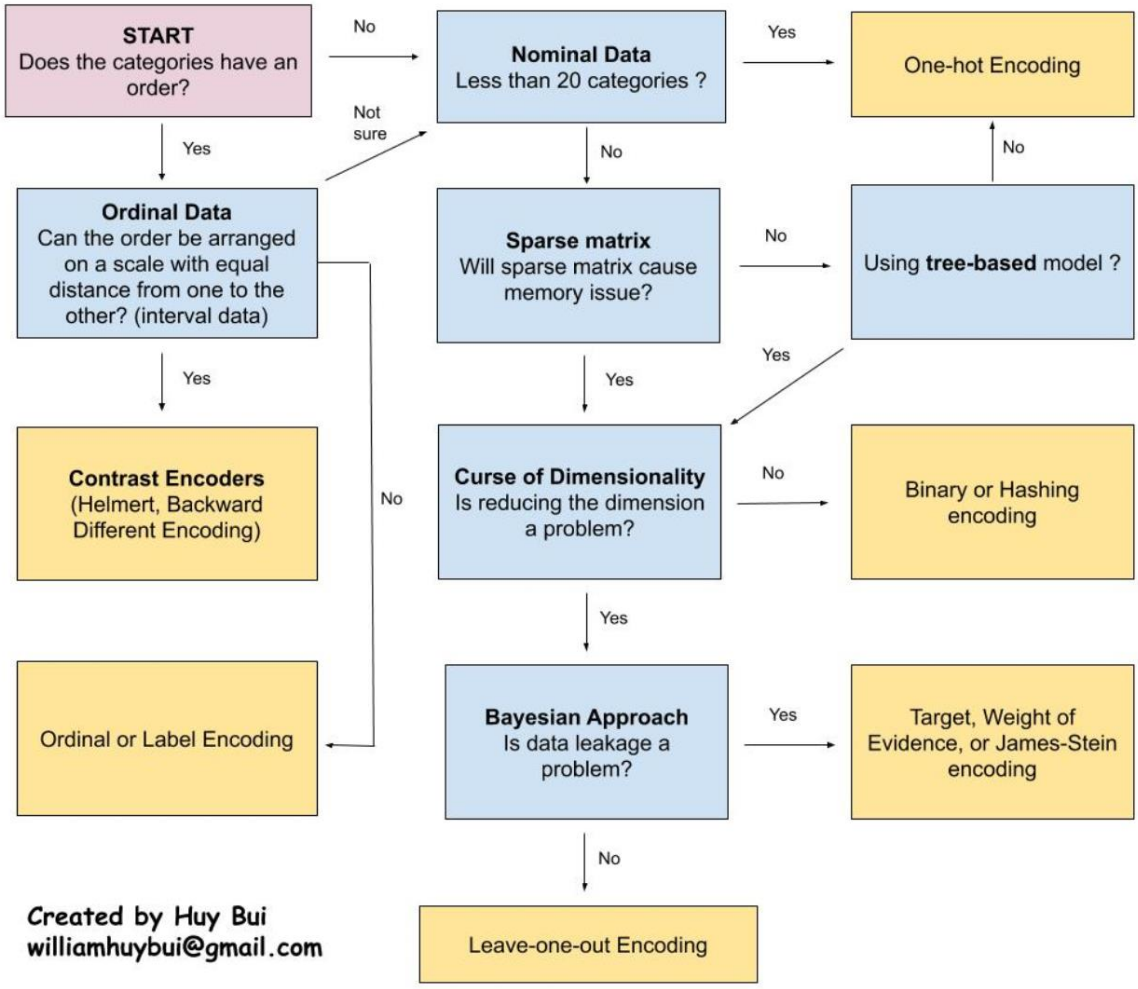
where $\text{mean}(y)$ is the global mean of the target, $\text{mean}(y_i)$ is the mean of the category, and B is the weight.

The weight B depends on the variances *σ(y)* and *σ(yi)*.

12) M-estimator Encoding

M-Estimate encoder is a simplified version of Target Encoder. The stands for maximum likelihood-type. It has only one hyper-parameter m, which represents the power of regularization. The higher the value of m results into stronger shrinking. Recommended values m are in the range of 1 to 100.

There is no single formula for encoding a feature. However, if you understand the 12 encoding techniques I introduced above, you would be able to move fast. Moreover, it always worth tries all the techniques that apply to the feature and decides which one works best. Try to input different regularization coefficient values and see if they increase your score. The cheat-sheet below will help us make some initial decisions.



Categorical encoding cheat sheet

For our dataframe, we use the Ordinal Encoder to change the object data type columns to integers. This will make describing, correlating and model testing much easier and accurate.

3.4 DESCRIBING THE DATASET

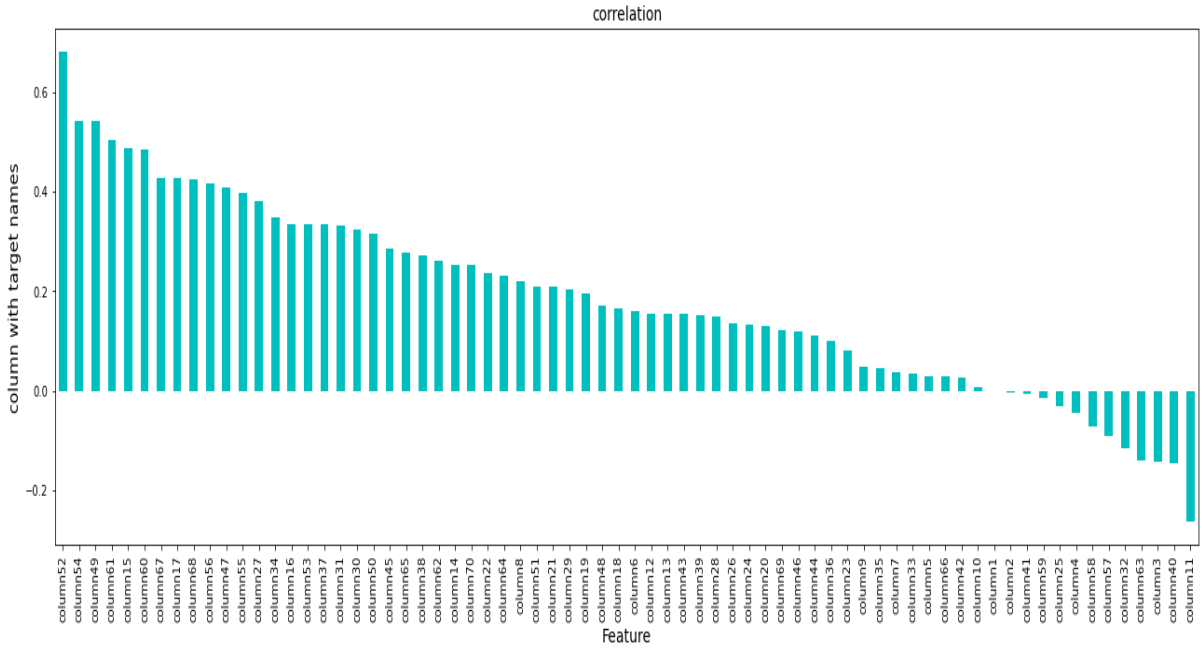
Describing the dataset gives us an understanding of various values like count of the attribute,mean of that attribute, standard deviation, minimum, 25th percentile, median/50th percentile, 75th percentile and the maximum value of that attribute.

We also use heatmap visualization for checking the relation between the described data.

After this, we check the correlation with the target column which is column71- Which e-commerce site will you recommend to your friend?

We also form a heatmap for visualising the correlation between each column with each other.

For checking whether the columns are positively or negatively correlated with the target column, we see the following chart:



Keeping +/-0.5 as the range for skewness, here are the columns which do not lie within the range- column1 column4 column5 column70 column71 ETC. Since no column has skewness, we will not treat them.

3.5 CHECKING FOR OUTLIERS

An **outlier** is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining.

Why outlier analysis?
Most data mining methods discard outliers noise or exceptions, however, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring one and hence, the outlier analysis becomes important in such case.

Detecting Outlier:
Clustering based outlier detection using distance to the closest cluster:
In the K-Means clustering technique, each cluster has a mean value. Objects belong to the cluster whose mean value is closest to it. In order to identify the Outlier, firstly we need to initialize the threshold value such that any distance of any data point greater than it from its nearest cluster identifies it as an outlier for our purpose. Then we need to find the distance of the test data to each cluster mean. Now, if the distance between the test data and the closest cluster to it is greater than the threshold value then we will classify the test data as an outlier.

Algorithm:

1. Calculate the mean of each cluster
2. Initialize the Threshold value
3. Calculate the distance of the test data from each cluster mean
4. Find the nearest cluster to the test data
5. If (Distance > Threshold) then, Outlier

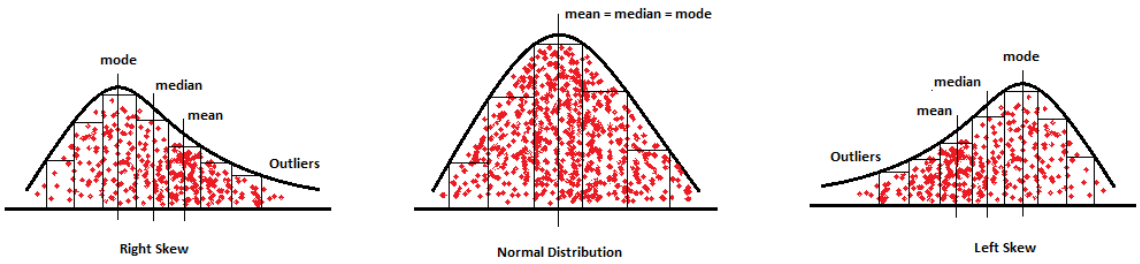
We use boxplot visualization for checking the outliers for the columns.

For removal of outliers we use Z-score.

The Z-score is a way to standardize the data to standard scale i.e. how far the data point is from the mean. The z-score can come positive or negative based on the help of mean and standard deviation values.

The data point away from the mean with some standard deviation is called a z-score.

The z-score can be perfectly found in a normal distribution curve with no left skew and right skew. The below image shows these curves.



Normal Distribution: The normal distribution is a curve in which the data is spread symmetrically on both sides of the mean.

Right Skew: The data is mostly skewed on the right side because most of the data is on the right side. If we talk about outliers they are mostly on the right side too.

Left Skew: The data is mostly skewed on the left side because most of the data is on the left side. If we talk about outliers they are mostly on the left side too.

Mean and Standard Deviation

Mean: The mean is an average value of the data that tells about the center value of the data.

Standard deviation: It is a spread of the data around the mean with one standard deviation.

Application of z-score in Machine learning

- To standardize the data as a part of data pre-processing.
- To compare the z-score values of different standard distributions for better results. Standard scaling is a crucial process in the data pre-processing of the machine learning algorithm.

3.6 SEPARATING COLUMNS INTO FEATURES AND TARGET

In order to make a prediction (in this case, whether a customer will recommend the e-commerce site to a friend or not), one needs to separate the dataset into two components:

- the **dependent** variable or **target** which needs to be predicted
- the **independent** variables or **features** that will be used to make a prediction

In machine learning, the concept of dependent and independent variables is important to understand. In the above dataset, if you look closely, the first 70 columns determine the outcome of the 71st , or last, column (Recommended). Intuitively, it means that the decision to buy a product of a given category is determined by the Gender (Male, Female), Age, and the Buying capacity of the individual. So, we can say that Recommended is the dependent variable, the value of which is determined by the other four variables.

With this in mind, we need to split our dataset into the matrix of independent variables and the vector or dependent variable. Mathematically, Vector is defined as a matrix that has just one column.

For dataset splitting we give the following command:

```
features=df.drop('column71',axis=1)
target=df['column71']
```

3.7 SCALING THE DATASET

Scaling is a method of standardization that's most useful when working with a dataset that contains continuous features that are on different scales, and you're using a model that operates in some sort of linear space (like linear regression or K-nearest neighbors)

Feature scaling transforms the features in your dataset so they have a mean of zero and a variance of one This will make it easier to linearly compare features. Also, this a requirement for many models in scikit-learn.

Feature Scaling is one of the most important transformation we need to apply to our data. Machine Learning algorithms (Mostly Regression algorithms) don't perform well when the inputs are numerical with different scales.

when different features are in different scales, after applying scaling all the features will be converted to the same scale. Let's take we have two features where one feature is measured on a scale from 1 to 10 and the second feature is measured on a scale from 1 to 100,00, respectively. If we calculate the mean squared error, algorithm will mostly be busy in optimizing the weights corresponding to second feature instead of both the features. Same will be applicable when the algorithm uses distance calculations like Euclidian or Manhattan distances, second feature will dominate the result. So, if we scale the features, algorithm will give equal priority for both the features.

There are two common ways to get all attributes to have the same scale: *min-max scaling* and *standardization.*

We will use Min-Max scaling technique for our dataset.

Min-Max scaling, We have to subtract min value from actual value and divide it with max minus min. Scikit-Learn provides a transformer called MinMaxScaler. It has a feature_range hyperparameter that lets you change the range if you don't want 0 to1 for any reason.

class sklearn.preprocessing.MinMaxScaler(feature_range=0,1,*, copy=True, clip=False).

After this, we split the data into train and test data. We now check the training and testing accuracy using random state for loop for the range (0,100).

We now check the r2 score and handle overfitting and underfitting of the data.

Cross validation gives us the training and testing score, model accuracy and we do this using the CV mean using the for loop function.

We now draw the Best Fit line to represent the number of datapoints which shows good fit of our model. Equation of this line is : y=mx+c

3.8 REGULARIZATION OF THE DATASET

Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from overfitting by adding extra information to it.

Sometimes the machine learning model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when deals with unseen data by introducing noise in the output, and hence the model is called overfitted. This problem can be deal with the help of a regularization technique.

This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model.

It mainly regularizes or reduces the coefficient of features toward zero. In simple words, "*In regularization technique, we reduce the magnitude of the features by keeping the same number of features.*"

Techniques of Regularization

There are mainly two types of regularization techniques, which are given below:

- **Ridge Regression**
- **Lasso Regression**

Ridge Regression

- Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions.
- Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as **L2 regularization**.
- In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called **Ridge Regression penalty**. We can calculate it by multiplying with the lambda to the squared weight of each individual feature.

Lasso Regression:

- Lasso regression is another regularization technique to reduce the complexity of the model. It stands for **Least Absolute and Selection Operator.**
- It is similar to the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights.

- Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.
- It is also called as **L1 regularization.**

Key Difference between Ridge Regression and Lasso Regression

- **Ridge regression** is mostly used to reduce the overfitting in the model, and it includes all the features present in the model. It reduces the complexity of the model by shrinking the coefficients.
- **Lasso regression** helps to reduce the overfitting in the model as well as feature selection.

For our dataset, we use the Lasso Regression for regularization giving the following command:

```
from sklearn.model_selection import GridSearchCV

from sklearn.linear_model import Lasso

parameters={'alpha': [.0001, .001, .01, .1, 1, 10],

        'random_state':list(range(0,10))}

ls=Lasso()

clf= GridSearchCV(ls,parameters)

clf.fit(features_train,target_train)

print(clf.best_params_)
```

We get an output stating: {'alpha': 0.0001, 'random_state': 0}

We now do the final model training using the above best parameters and get the test score as 99.99%.

3.9 ENSEMBLE TECHNIQUE

Ensemble method in Machine Learning is defined as the multimodal system in which different classifier and techniques are strategically combined into a predictive model (grouped as Sequential Model, Parallel Model, Homogeneous and Heterogeneous methods etc.) Ensemble method also helps to reduce the variance in the predicted data, minimize the biasness in the predictive model and to classify and predict the statistics from the complex problems with better accuracy.

Ensemble Methods help to create multiple models and then combine them to produce improved results, some ensemble methods are categorized into the following groups:

1. Sequential Methods

In this kind of Ensemble method, there are sequentially generated base learners in which data dependency resides. Every other data in the base learner is having some dependency on previous data. So, the previous mislabeled data are tuned based on its weight to get the performance of the overall system improved.

Example: Boosting

2. Parallel Method

In this kind of Ensemble method, the base learner is generated in parallel order in which data dependency is not there. Every data in the base learner is generated independently.

Example: Stacking

3. Homogeneous Ensemble

Such an ensemble method is a combination of the same types of classifiers. But the dataset is different for each classifier. This will make the combined model work more precisely after the aggregation of results from each model. This type of ensemble method works with a large number of datasets. In the homogeneous method, the feature selection method is the same for different training data. It is computationally expensive.

Example: Popular methods like bagging and boosting comes into the homogeneous ensemble.

4. Heterogeneous Ensemble

Such an ensemble method is the combination of different types of classifiers or machine learning models in which each classifier built upon the same data. Such a method works for small datasets. In heterogeneous, the feature selection method is different for the same training data. The overall result of this ensemble method is carried out by averaging all the results of each combined model.

Below are the technical classification of Ensemble Methods:

1. Bagging

This ensemble method combines two machine learning models i.e. Bootstrapping and Aggregation into a single ensemble model. The objective of the bagging method is to reduce the high variance of the model. The decision trees have variance and low bias. The large dataset is (say 1000 samples) sub-sampled (say 10 sub-samples each carries 100 samples of data). The multiple decision trees are built on each sub-sample training data. While banging the sub-sampled data on the different decision trees, the concern of over-fitting of training data on each decision tree is reduced. For the efficiency of the model, each of the individual decision trees is grown deep containing sub-sampled training data. The results of each decision tree are aggregated to understand the final prediction. The variance of the aggregated data comes to reduce. The accuracy of the prediction of the model in the bagging method depends on the number of decision-tree used. The various sub-sample of a sample data is chosen randomly with replacement. The output of each tree has a high correlation.

2. Boosting

The boosting ensemble also combines different same type of classifier. Boosting is one of the sequential ensemble methods in which each model or classifier run based on features that will utilize by the next model. In this way, the boosting method makes out a stronger learner model from weak learner models by averaging their weights. In other words, a stronger trained model depends on the multiple weak trained models. A weak learner or a wear trained model is one that is very less correlated with true classification. But the next weak learner is slightly more correlated with true classification. The combination of such different weak learners gives a strong learner which is well-correlated with the true

3. Stacking

This method also combines multiple classifications or regression techniques using a meta-classifier or meta-model. The lower levels models are trained with the complete training dataset and then the combined model is trained with the outcomes of lower-level models. Unlike boosting, each lower-level model is undergone into parallel training. The prediction from the lower level models is used as input for the next model as the training dataset and form a stack in which the top layer of the model is more trained than the bottom layer of the model. The top layer model has good prediction accuracy and they built based on lower-level models. The stack goes on increasing until the best prediction is carried out with a minimum error. The prediction of the combined model or meta-model is based on the prediction of the different weak models or lower layer models. It focuses to produce less bias model.

4. Random Forest

The random forest is slightly different from bagging as it uses deep trees that are fitted on bootstrap samples. The output of each tress is combined to reduce variance. While growing each tree, rather than generating a bootstrap sample based on observation in the dataset, we also sample the dataset based on features and use only a random subset of such a sample to

build the tree. In other words, sampling of the dataset is done based on features that reduce the correlation of different outputs. The random forest is good for deciding for missing data. Random forest means random selection of a subset of a sample which reduces the chances of getting related prediction values. Each tree has a different structure. Random forest results in an increase in the bias of the forest slightly, but due to the averaging all the less related prediction from different trees the resultant variance decreases and give overall better performance.

For our dataset, we use the Random Forest Regressor and get the output as:

{'criterion': 'mse', 'max_features': 'auto'}.

Now, we put the output and check the r2 and cross val score of the model. We get an output:

R2 Score: 99.998974488827
Cross val score: 99.99459877146819

We are getting model accuracy and cross validation both above 99% which shows our model is performing very good.

4. CONCLUSION

4.1 SAVING THE MODEL

Pickle is a useful Python tool that allows you to save your ML models, to minimise lengthy re-training and allow you to share, commit, and re-load pre-trained machine learning models. Most data scientists working in ML will use Pickle or Joblib to save their ML model for future use.

Pickle is a generic object serialization module that can be used for serializing and deserializing objects. While it's most commonly associated with saving and reloading trained machine learning models, it can actually be used on any kind of object. Here's how you can use Pickle to save a trained model to a file and reload it to obtain predictions.

To save the ML model using Pickle all we need to do is pass the model object into the dump() function of Pickle. This will serialize the object and convert it into a "byte stream" that we can save as a file called model.pkl. You can then store, or commit to Git, this model and run it on unseen test data without the need to re-train the model again from scratch.

We use the pickle.dump method to save the final model.

4.2 CONCLUSION

To load a saved model from a Pickle file, all you need to do is pass the “pickled” model into the Pickle load() function and it will be deserialized. By assigning this back to a model object, you can then run your original model’s predict() function, pass in some test data and get back an array of predictions.

At the end, we load the saved model and check the predicted and original recommendations done by the customer for the e-commerce site.

We see that our model is performing very good with the accuracy score of 99.99%.

5. BIBLIOGRAPHY

References:

- https://pbpython.com/pandas_dtypes.html
- https://www.datasciencemadesimple.com/encode-decode-column-dataframe-python/#:~:text=encode%20%28%29%20function%20with%20codec%20%E2%80%98base64%E2%80%99%20and%20error,be%20Decode%20a%20column%20of%20dataframe%20in%20python%3A
- https://www.geeksforgeeks.org/machine-learning-outlier/
- https://medium.com/pythoneers/z-distribution-or-z-score-application-in-machine-learning-fbba081cd9fe
- https://campus.datacamp.com/courses/hr-analytics-predicting-employee-churn-in-python/predicting-employee-turnover?ex=2
- https://towardsdatascience.com/data-scaling-for-machine-learning-the-essential-guide-d6cfda3e3d6b
- https://blog.finxter.com/the-complete-guide-to-min-max-scaler-in-machine-learning-with-ease/#:~:text=Min-Max%20scaling%20is%20a%20normalization%20technique%20that%20enables,range%20using%20each%20feature%E2%80%99s%20minimum%20and%20maximum%20value.
- https://www.javatpoint.com/regularization-in-machine-learning
- https://www.educba.com/ensemble-methods-in-machine-learning/
- https://practicaldatascience.co.uk/machine-learning/how-to-save-and-load-machine-learning-models-using-pickle

THE END