Internship Group 2

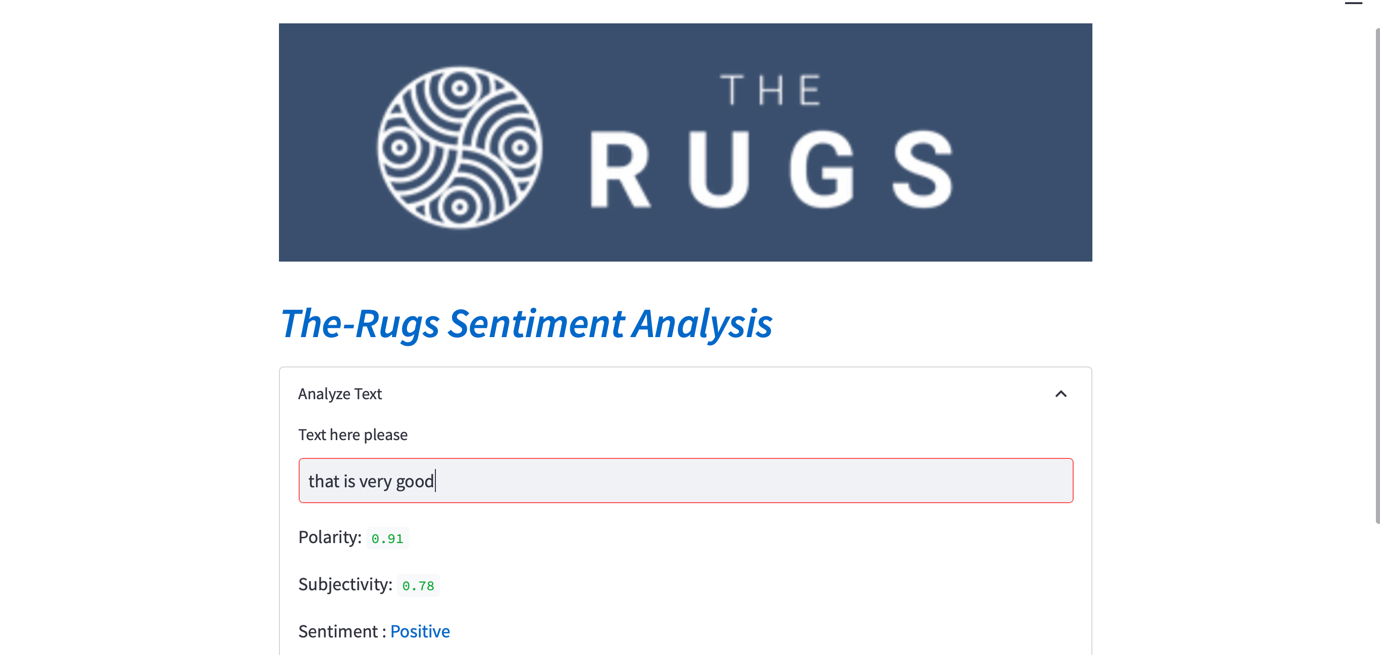
Sentiment Analysis APP



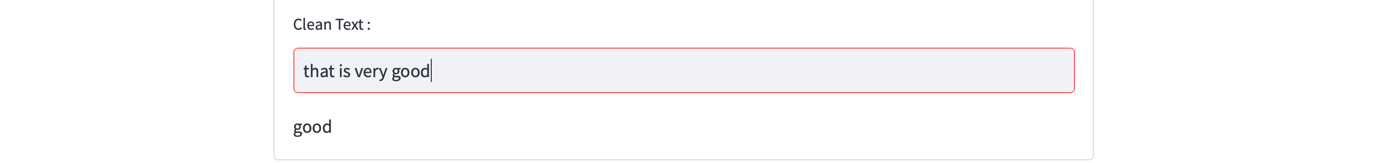
We have an interface in this way.

We can do 3 operations here.

1. If a sentence is entered manually, it analyzes the sentence.



And if you are wondering which words are actually more important for our application, and which words are evaluated according to a manually entered text, this can be viewed on the 'Clean Text' line.



1- When we add the EXCEL file that contains the comments collectively, a WordCloud is created immediately according to the words in these comments.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

1- And it analyzes all the comments one by one and gives the sentiment results and scores to them.metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

In this way, negative comments can be seen immediately and an immediate reaction can be given. In this way, customer satisfaction is increased.

This is how the APP looks when a file is dropped. I shrunk the screen to fit them all in the screenshot. That's why some of the comments are not visible. But in normal working time, it is perfectly fine and the comment on each line can be read regardless of its length.metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

This way you can see the results of the comments and if desired, these results can be downloaded as a CSV file.

Parts of the APP Project that need to be completed

1- I used a ready-made library called Textblob for Sentiment analysis. But in the text tests I made afterwards, I saw that he made obvious analysis mistakes in some places.

As a solution to this, I want to include Bert model from Hugging Face in App instead of Textblob.

2- We can only export Excel files to the APP.

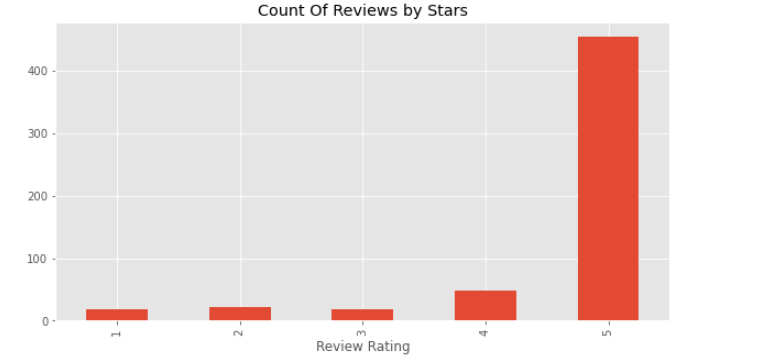
I want to write another function here and read and evaluate whichever Excel and CSV formats are discarded.

3- We can only download the results in CSV format if desired. I want to add one more option here. If desired, it should be downloadable as Excel.

**Sentiment Analysis with Vader and Roberta**

In this Sentiment Analysis we implemented a Vader and Roberta pretrained models on the ‘Comments’ data set. The results are shown as it follows.

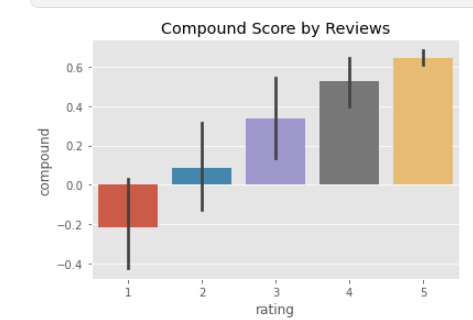
**Amount of Ratings**

****

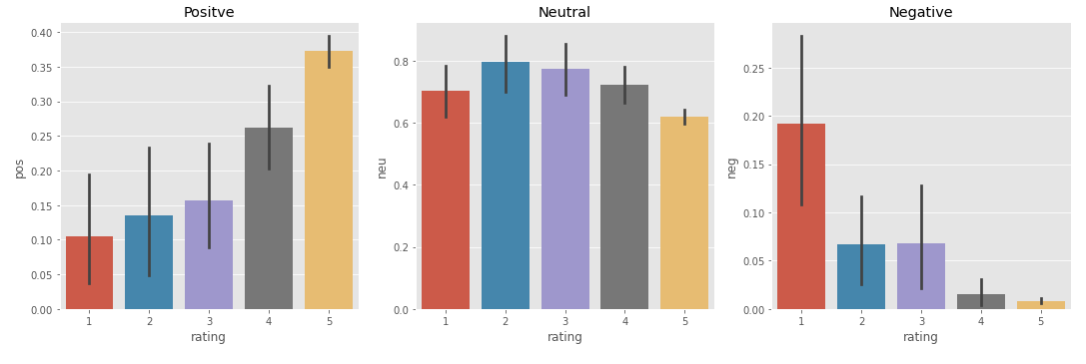
As it is seen above the most of the ratings accumulated on 5. That means we have lots of ‘5’ ratings.

**Vader Sentiment Analysis**

Vader is pre-trained ‘NLP’ library which is used to analyze text or sentence. The main purpose of Vader is Classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, sentence, or entity feature/aspect is positive, negative, or neutral—is a fundamental task in sentiment analysis. Advanced sentiment classification " polarity" considers emotional states such as ‘positive’,’negative’ , ‘neutral and ‘compound’.



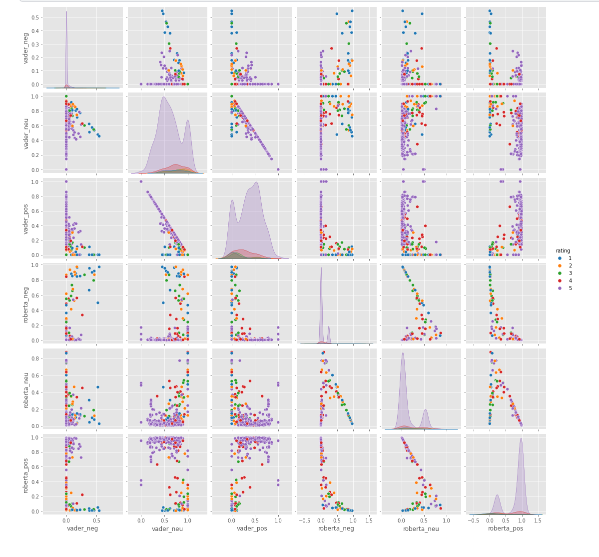
As it is shown above, the data set is the data set consists of relatively positive comments according to the vader library.



The data above has proved, the sentiment classification sentiment shows that the classification is working quite properly.

**Sentiment Analysis with Roberta**

Roberta is one of the pre-trained ‘NLP’ models to implement sentiment analysis on text and Word. Roberta is similar to Vader in several points much more better than Vader.



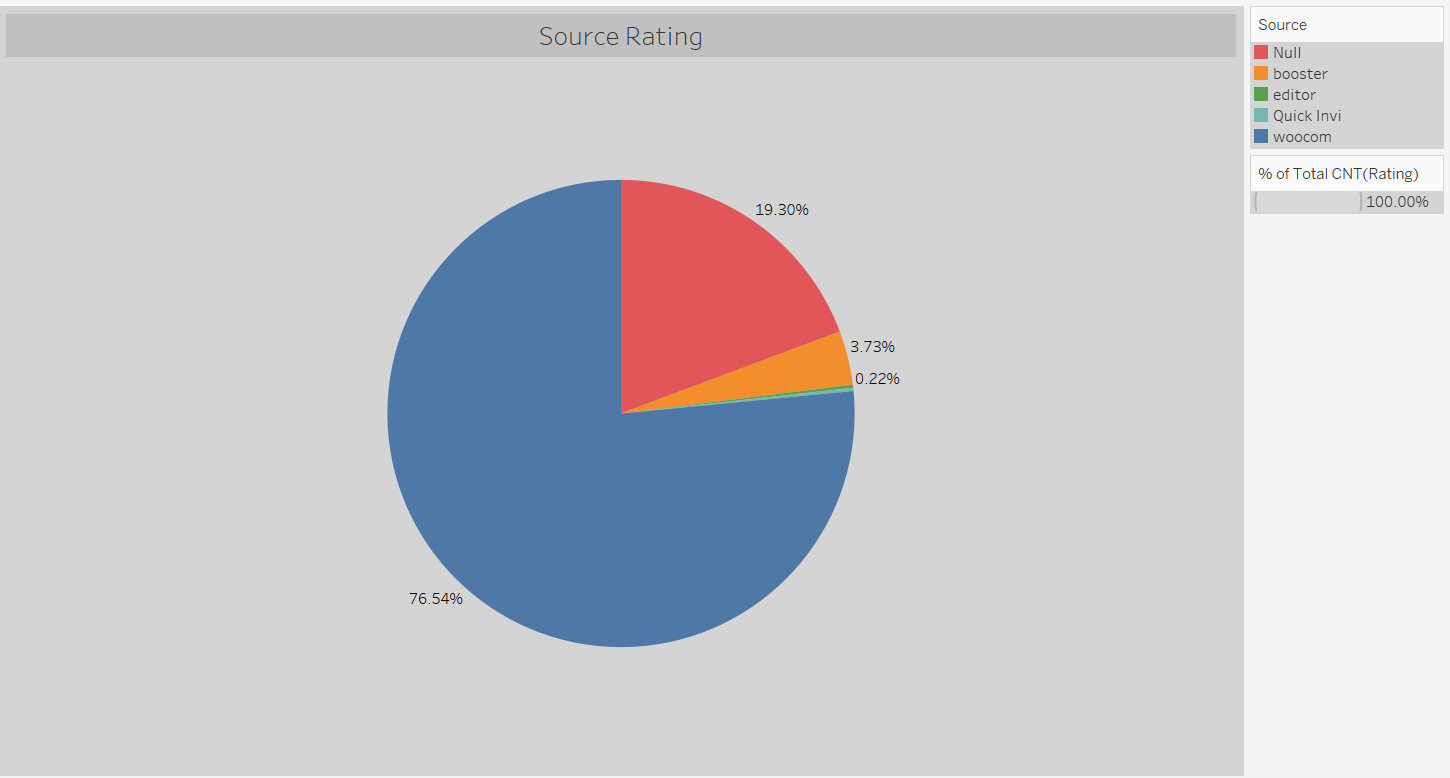
Comparing to Vader Roberta gave us better results on sentiment analysis.

**Recommendations**

The most important suggestion to be made to the company in this section is to update the tabs for website comments and make it mandatory for customers to comment in the comment tabs.

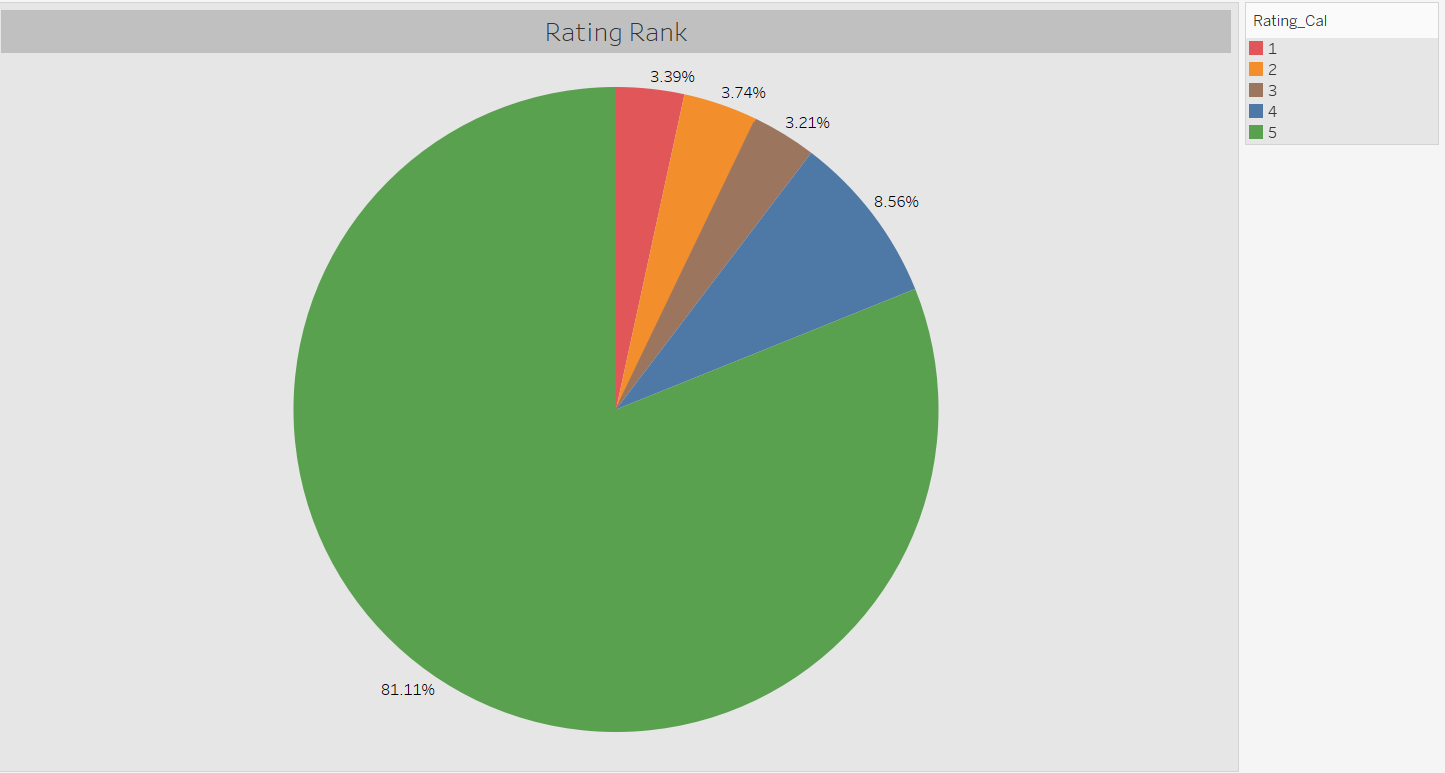
**CUSTOMER ANALYSIS**

**Graph-1**



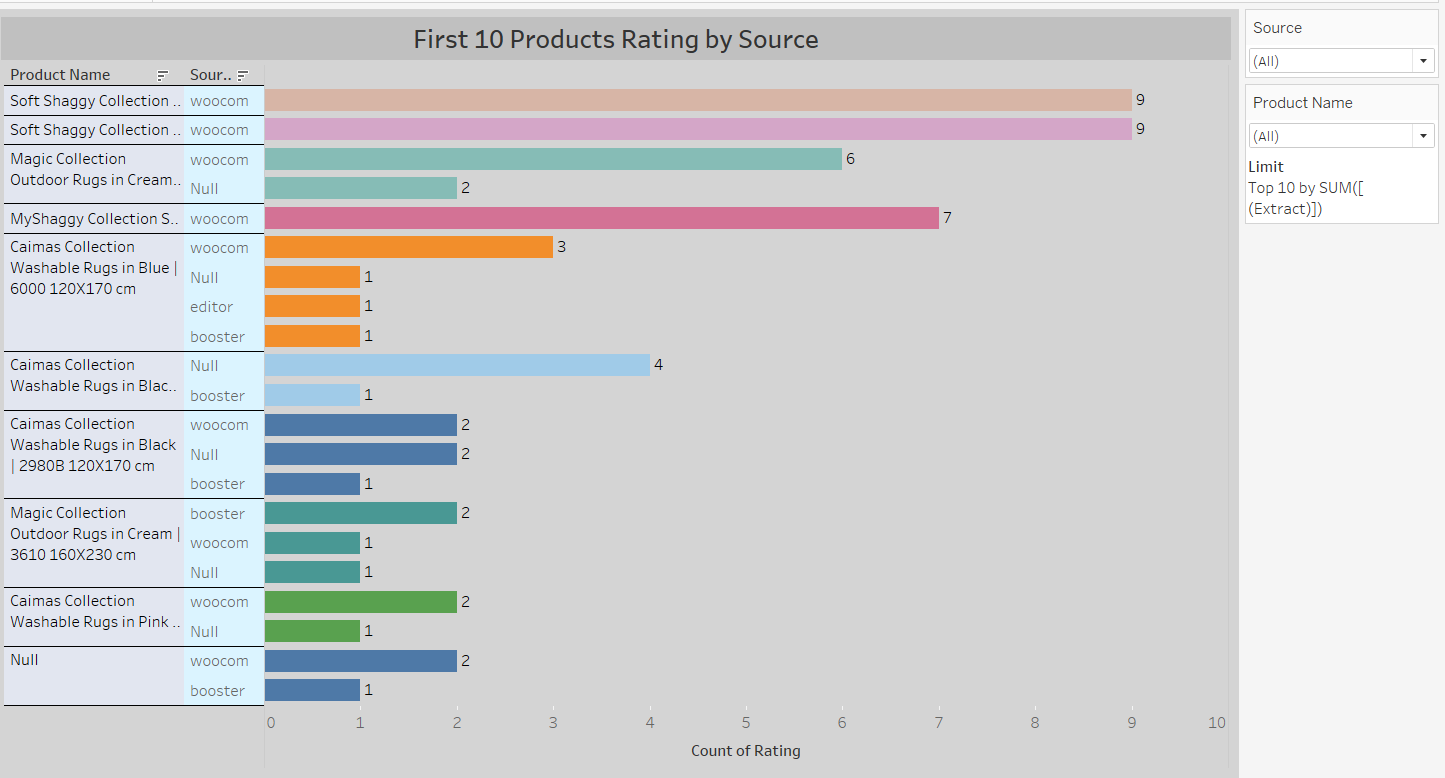
The highest rating was made from the woocommerce platform with 76.54%.

**Graph-2**



It is seen that the highest rating is made as “5” with a rate of 81.11%. At the rate of 3.39%, the lowest (1) rating was taken.

**Graph-3**



When the 10 product categories with the highest ratings are examined, it is seen that the “Soft Shaggy Collection” brand is the product with the highest rating (9) from the woocommers platform. When the whole analysis is examined, although the products with high ratings seem to be in the first place, in fact, it is seen that the products with a low rating are in the first place.

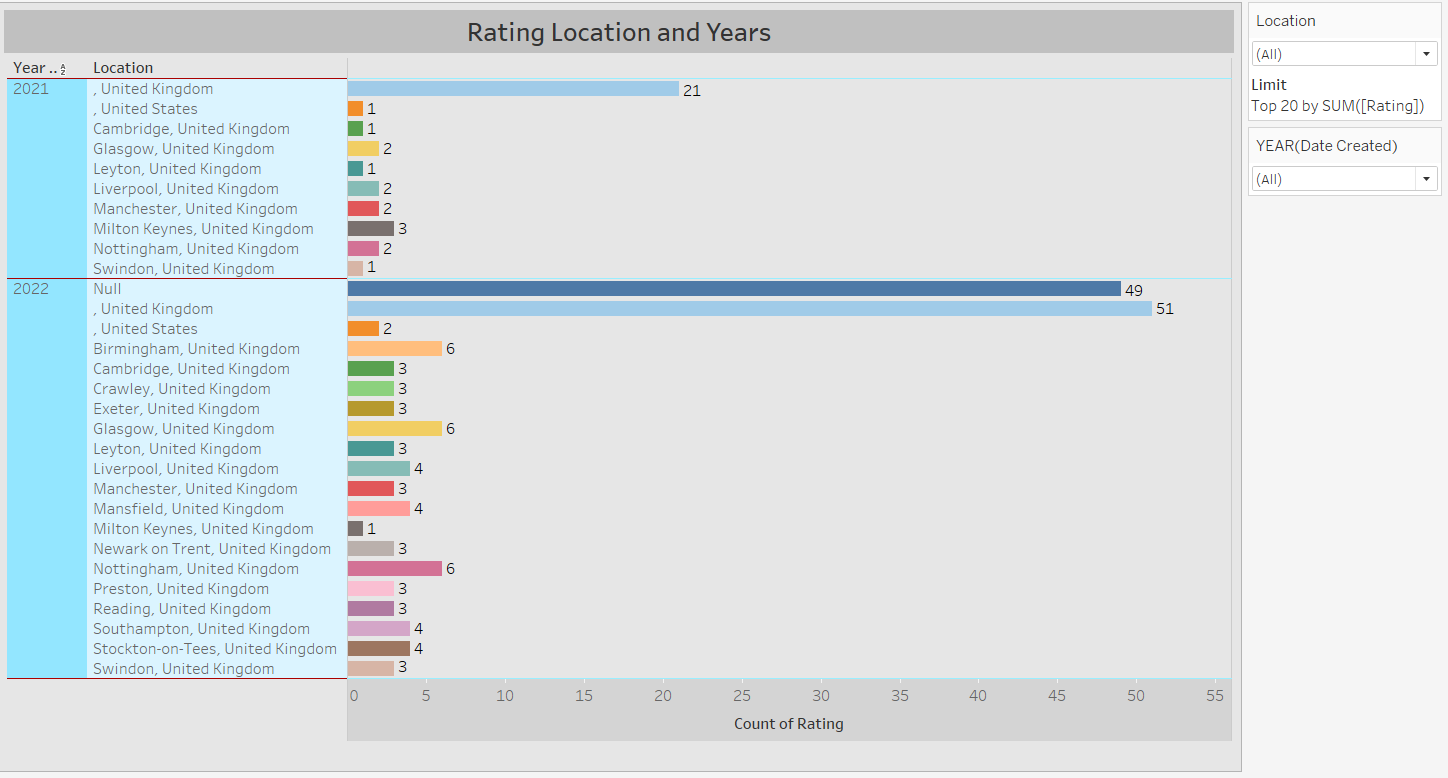
**Graph-4**

harita içeren bir resim

Açıklama otomatik olarak oluşturuldu

When the rating situation on the basis of countries is examined, the highest rating is seen in the UK. This is similar to the order numbers.

**Graph-5**



Considering the number of ratings according to years and location, it is seen that more ratings are received in 2022 than in other years. In addition, it is understood that the highest rating is received from United Kingdom (London).

**Graph-6**

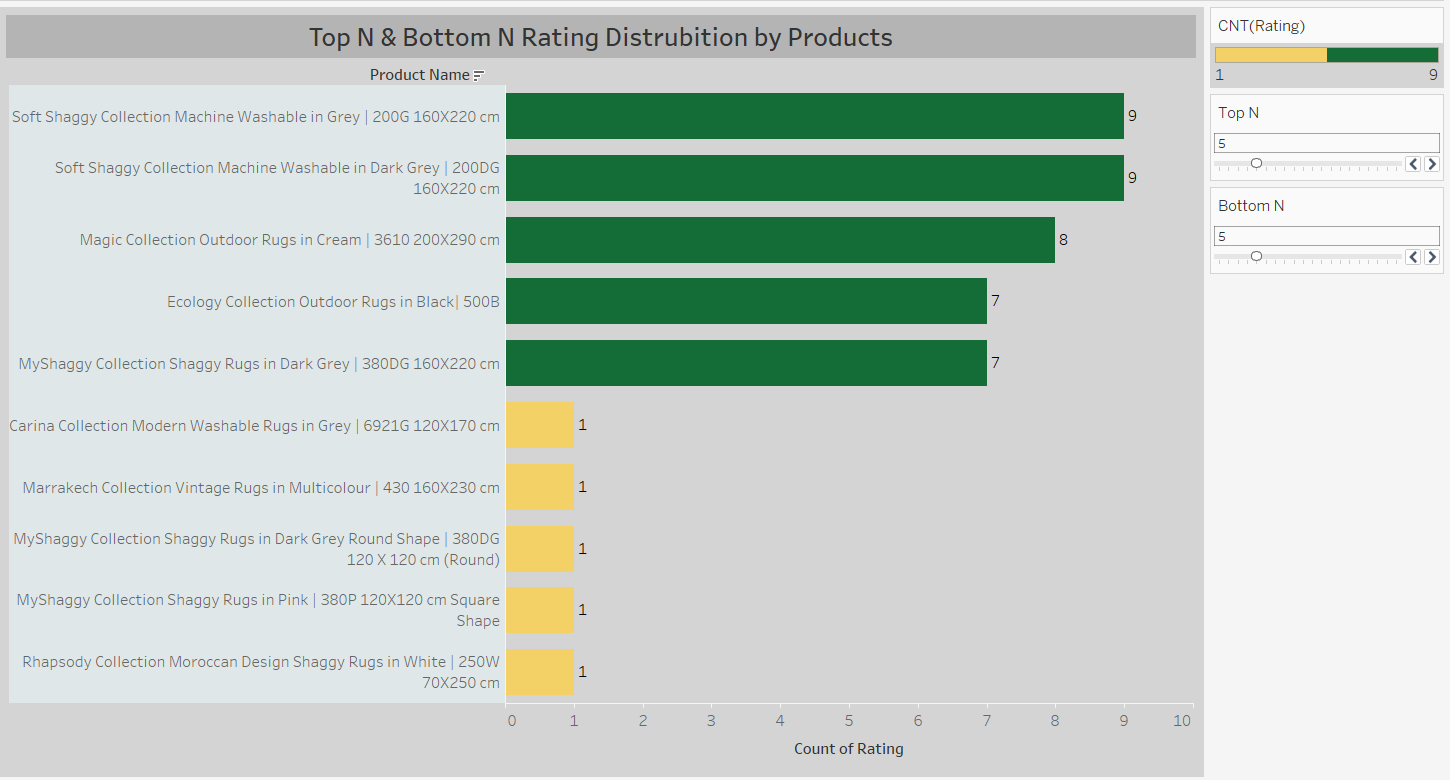


When the rating status of the first 10 products according to their rating is examined; “Washable Shaggy Collection in Red | It is seen that the 181 120X170 cm” product has 3 ratings at the 5 level and is in the first place.

Again, “Washable Shaggy Collection in Red Round Shape| 181 100X100 cm (Round)” product, on the other hand, is seen to be in the top 10 product category, despite receiving a rating of 1 at the level of 2.

Although there is a rating of 2 for the product, it is seen that the order of the product is not affected much.

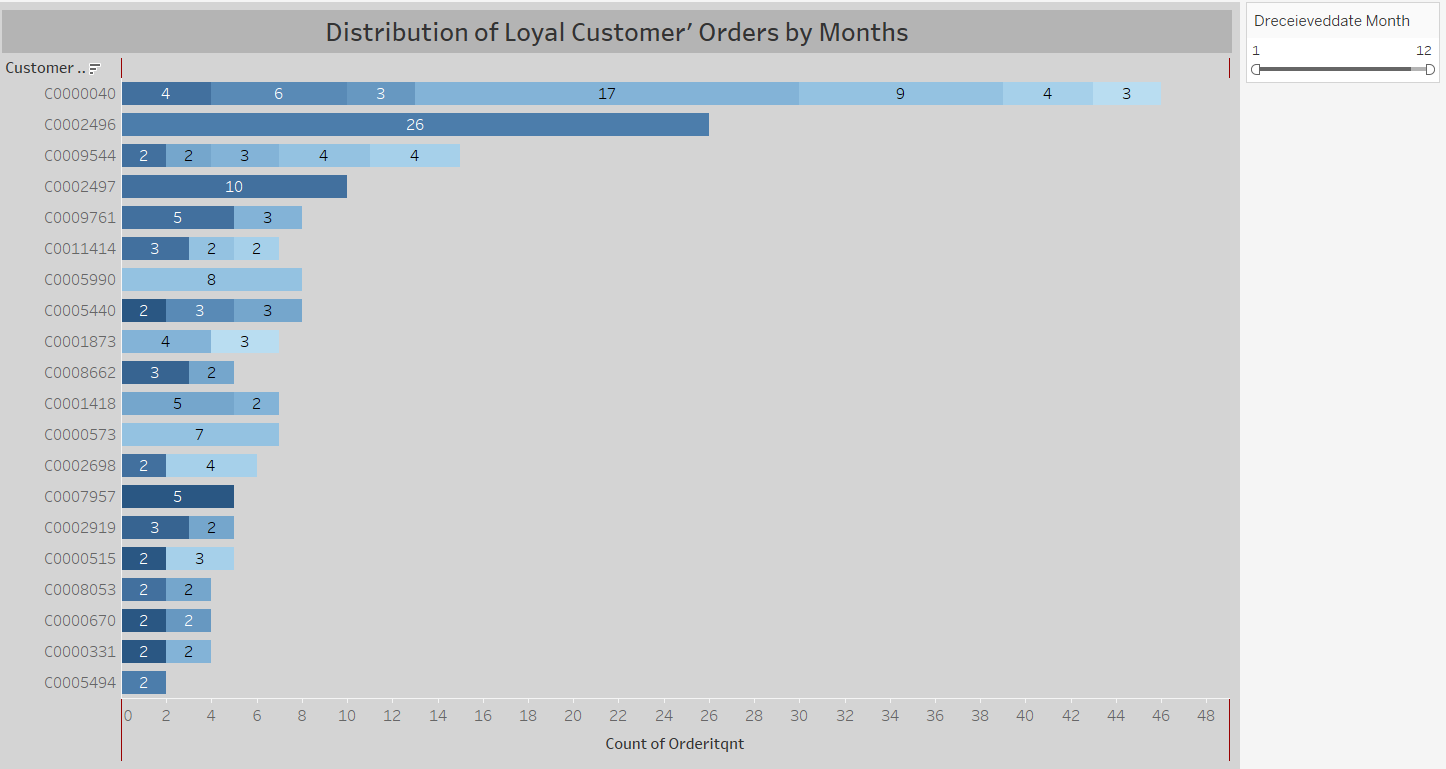
**Graph-7**



According to the rating status; when the first and last 5 products are examined, “Soft Shaggy Collection Machine Washable in Gray | 200G 160X220 cm” and “Soft Shaggy Collection Machine Washable in Dark Gray | It is seen that 200DG 160X220 cm” is the product with the highest rating, with 9 ratings.

“Rhapsody Collection Moroccan Design Shaggy Rugs in White | 250W 70X250 cm” and “MyShaggy Collection Shaggy Rugs in Pink | It is understood that 380P 120X120 cm Square Shape is the product with the least rating with 1 rating.

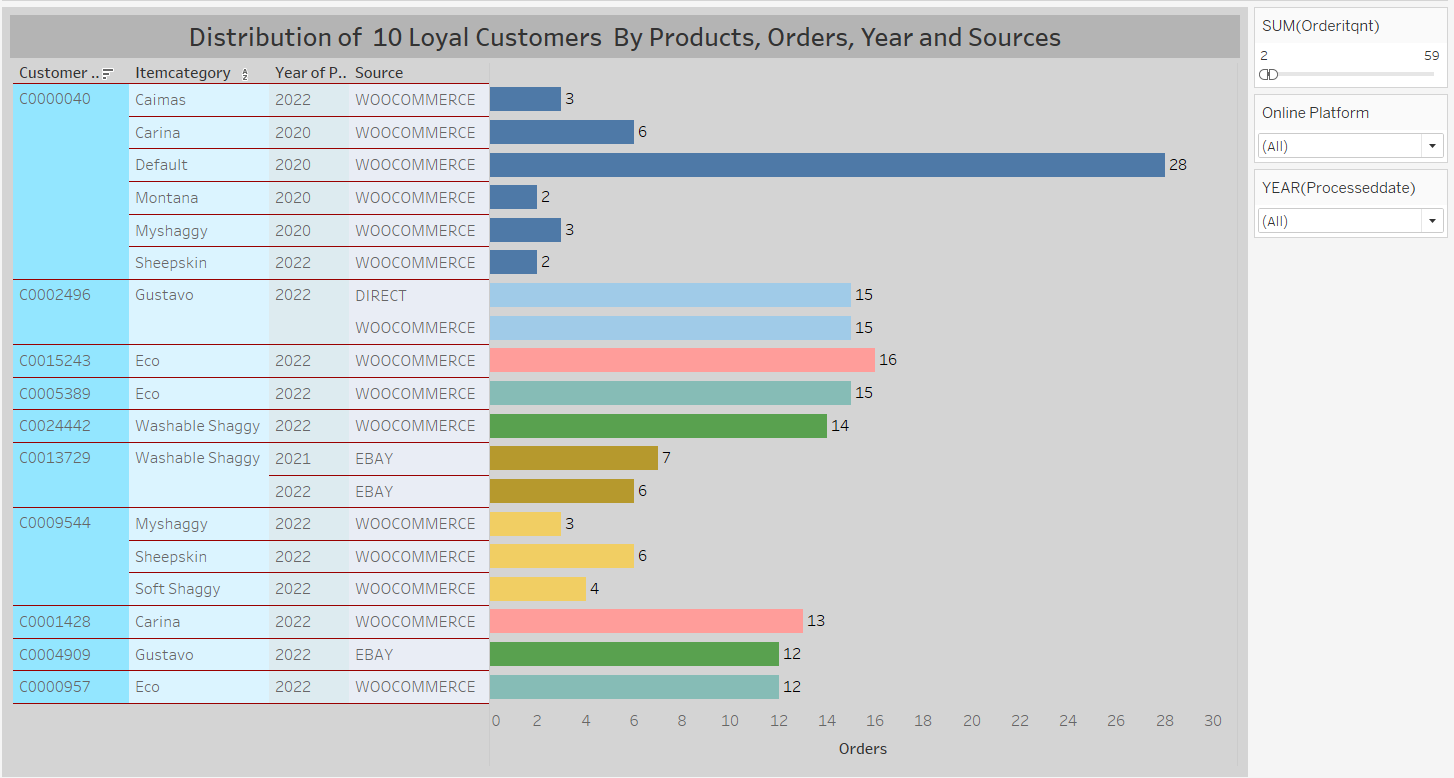
**Graph-8**



When the first 20 loyal customers are examined, it is seen that the first customer (C0000040) placed a total of 46 orders at 7 different times. It is understood that the second customer (C0002496) ordered 26 pieces at once. The reason why 1 person orders so much in a product category is not well understood.

\*Amazon customers are not included in the loyal customer list.

**Graph-9**

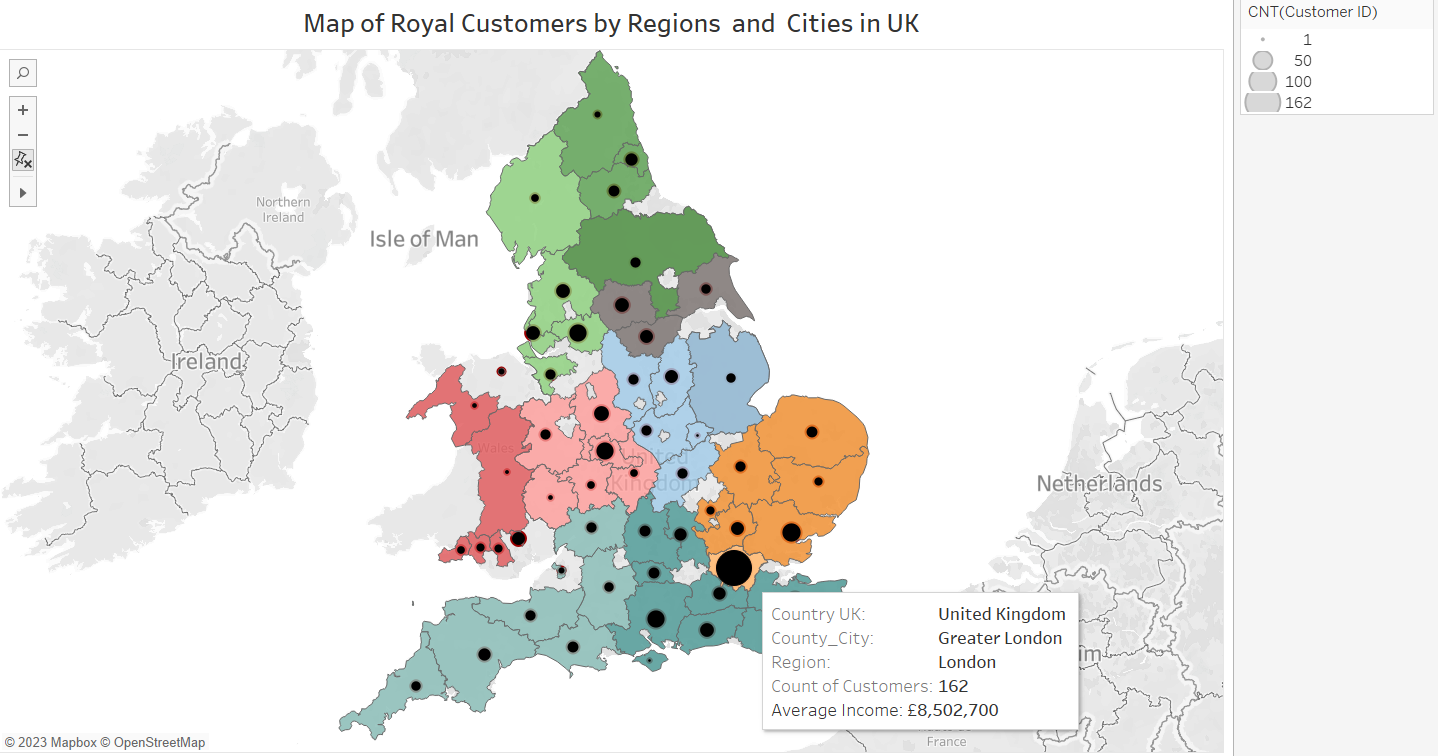


The first customer, in 2020 and 2022, using the woocommers platform; A total of 44 orders were placed in the “caimas, carina, myshaggy, montana, sheepskin and default” models. It is noteworthy here that he did not place an order in 2021. In addition, it is seen that the customer orders all these models from the same platform.

If the second customer (C0002496); It is seen that in 2022, 30 orders were placed in total from woocommerce and direct platforms.

It is understood that loyal customers place their orders from different platforms and different carpet models.

**Graph-10**



Looking at the number of loyal customers in the regions and cities on the map, it is seen that the most loyal customers are in London and then in the south east region.

In addition, when the relationship between the number of loyal customers and their average income is examined, it can be said that there is no difference.

**Clustering Of Customers**

Under this title, we made analysis according to the tax amount which paid by customers. The main logic of this clustering analysis to divide customers after ‘Subtotal’ and ‘Tax’ amount.

**Realization of Clustering Analysis**

At first the data set is prepared to analysis. We named it as ‘df\_subtotal’.

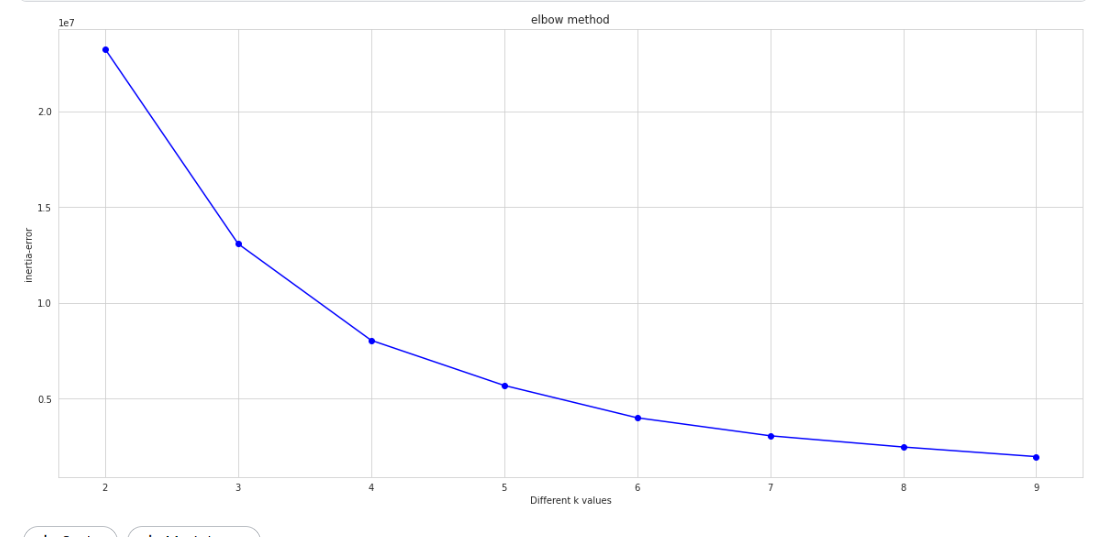
df2.set\_index('CUSTOMER\_ID',inplace=True)  
df\_subtotal = df2[['SUBTOTAL','TAX']]

In this step the target variable ‘Tax’ is dropped.

X = df\_subtotal.drop('TAX',axis=1)

In this step, the k means model determines the best clustering value. At the end of the clustering the elbow is

**from** sklearn.cluster **import** KMeans  
ssd = []  
K = range(2,10)  
**for** k **in** K:  
 model = KMeans(n\_clusters =k, random\_state=42)  
 model.fit(X)  
 ssd.append(model.inertia\_)

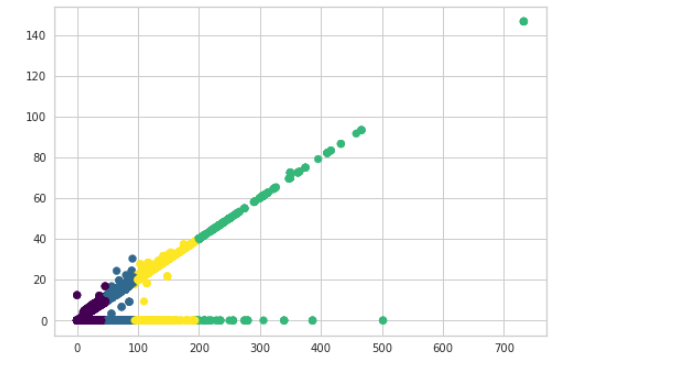


With this code below the model predicted the ‘Tax’ class of the customers.

**from** sklearn.cluster **import** KMeans  
  
K\_means\_model = KMeans(n\_clusters=4, random\_state=42)  
K\_means\_model.fit\_predict(X)

With the code below and the graph it is seen that the more paid the customers that paid much more tax.

plt.scatter(df\_subtotal["SUBTOTAL"],df\_subtotal['TAX'],c = df\_subtotal.predicted\_clusters,cmap='viridis');



**COHORT ANALYSIS WITH PYTHON AND POWER BI**

**What is Cohort Analaysis?**

Cohort analysis is a kind of [behavioral analytics](https://en.wikipedia.org/wiki/Behavioral_analytics) that breaks the data in a [data set](https://en.wikipedia.org/wiki/Data_set) into related groups before analysis. These groups, or [cohorts](https://en.wikipedia.org/wiki/Cohort_(statistics)), usually share common characteristics or experiences within a defined time-span.Cohort analysis allows a company to "see patterns clearly across the life-cycle of a customer (or user). By seeing these patterns of time, a company can adapt and tailor its service to those specific cohorts.

**Types of Cohorts**

1- Time-Based Cohorts

Time-based cohorts are customers who signed up for a product or service during a particular time frame. Analyzing these cohorts shows the customers’ behavior depending on the time they started using a company’s products or services. The time may be monthly or quarterly, depending on the sales cycle of a company.

2- Segment-Based Cohorts

It groups customers by the type of product or level of service they signed up for. Customers who signed up for basic level services might have different needs than those who signed up for advanced services. Understanding the needs of the various cohorts can help a company design tailor-made services or products for particular segments.

3- Size -Based Cohorts

Size-based cohorts refer to the various sizes of customers who purchase a company’s products or services. The customers may be small and startup businesses, middle-sized businesses, and enterprise-level businesses. Comparing the different categories of customers based on their size reveals where the largest purchases come from.

**Cohort Analysis with and without Amazon Sales**

In this analysis, we will initially use the entire sales data. However, since it includes Amazon sales and represents all Amazon customers as one (C0000001), we will eliminate Amazon sales and customer "C0000001" in our second analysis.

**1- Customers who have purchased multiple times:**

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Insights:**

1. Out of 92249 records, 54672 orders come from customer "C0000001" representing Amazon customers.

2. The second best customer is "C0000040". But when we check the customer "C0000040" from the original Sales data, almost half of the company columns have "Student". So, it is a group of people but not a specific customer.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

**2- Our first customers (in November 2019):**

**metin içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**Insights:**

1. In the beginning, we had only three different customers: “C0000001” , “C0001525 ”, “C0005444 ” and one of them represents Amazon customers.

2. As seen above, there are a few Amazon records in the first month. So, if we exclude Amazon sales, we will lose only one customer (C0000001) in Cohort Analysis in November 2019. Other months will not be affected because the first record for each customer is included in this analysis. Apart from that, Cohort Analysis without Amazon Sales will not make any difference and will resemble the tables and graphs below.

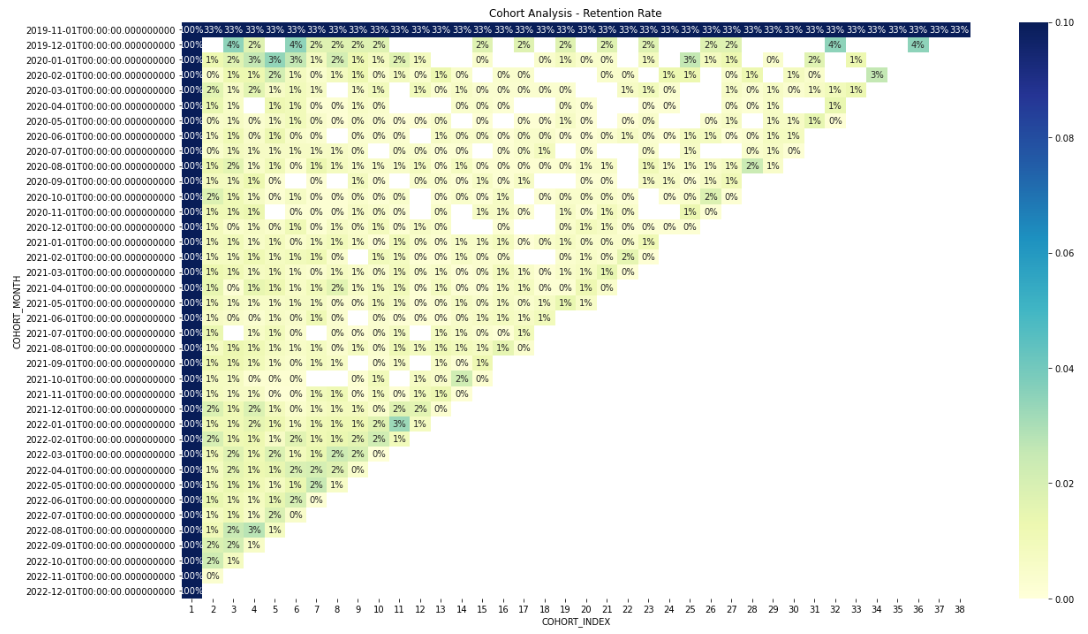
**3- Create a Cohort Index for 38 months to see customer life-cycle**

**metin içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**Insights:**

As seen above, the company had only 3 customers in the beginning and it could achieve to retain only one customer among them, namely, Amazon customers (C0000001). In other words, other customers were one-time customers and left the company in the following month. So, the retention rate is 33 % and the churn rate is 66 % for November 2019 customers.

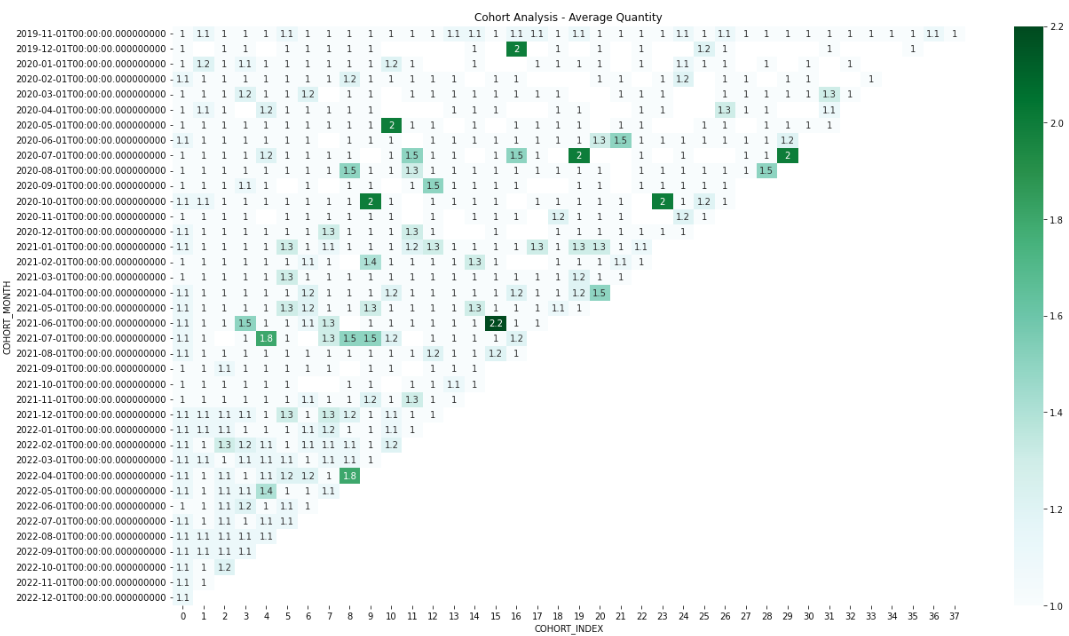


**Insights:**

Retention rates in general are very low regardless of month and year. Rarely, 3% and 4% of our customers continue shopping. The majority of these are customers who bought products from the company in December 2019. The customers who met the company during this period seem to be the most loyal customers.

One reason why the company has low retention rates or high churn rates is the lack of variety of products sold by the company since it is a rug company. Also, they sell a product (rugs and carpets) that customers can use for many years when they buy them. So it takes a long time for a customer to return. As a result, Cohort Analysis does not give valuable insights for the company unless it diversifies the products it sells.

**4- Average Quantity Sold by These Customers**

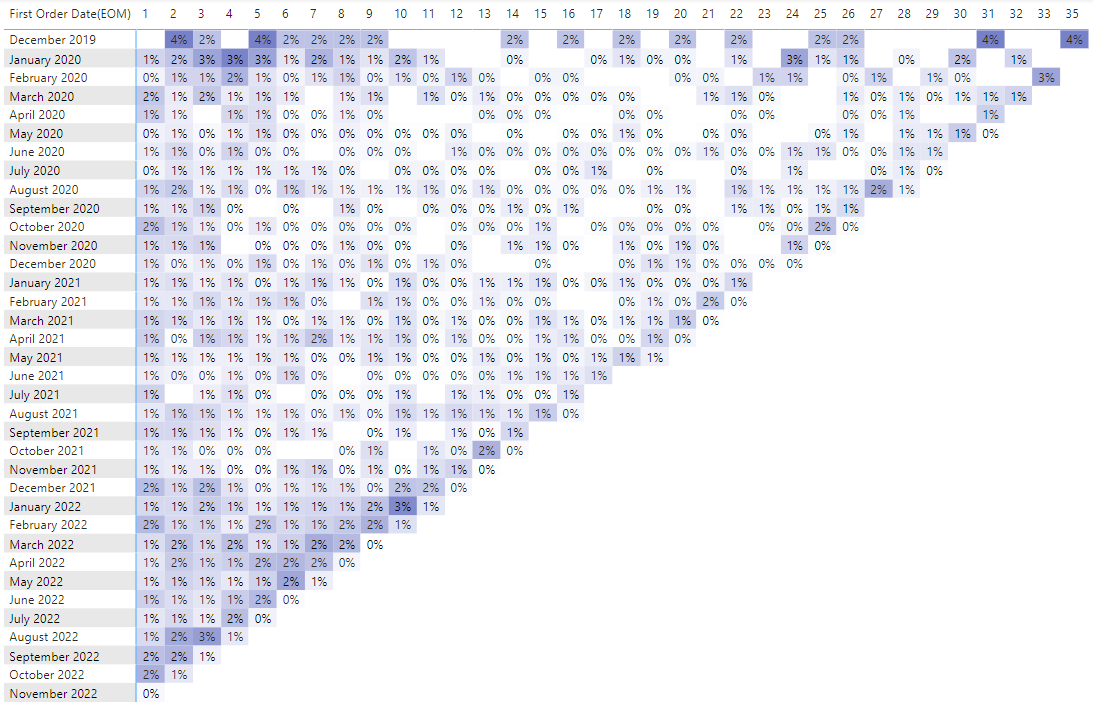
****

|  |
| --- |
|  |

**Insights:**

The graph above reveals that loyal customers buy between 1 - 2.2 items on average. The reason is that a customer can buy a limited number of rugs and use them for a long time.

**5- Cohort Analysis without Amazon Sales by using Power BI**

****

**Insights:**

1. When we exclude Amazon Sales and the first month the customers buy for the first time in our company, we see that November 2019 customers disappear. The reason is that no other customer continues shopping except Amazon customers.

2. When we check the months for above 1 %, we see that customers return mostly around November. Campaigns before Christmas and Black Friday may cause to return customers during this time of the year. The cell before the last one for each month represents November and we can detect a cross line for November across the graph.

3. Lastly, December 2019 and January 2020 customers seem the most loyal customers.

# Word Cloud

Expressions containing unnecessary characters and numbers in product comments have been deleted.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure 1.1.

The number of times each word was used was checked.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure 1.2.

A word cloud was created using a mask. A speech bubble was preferred because the comments were visualised.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure 1.3.

The words used in the comments are mostly words with positive connotations. This is an indication that customers are satisfied with the products they buy.

# Customer Segmentation by RFM

One of the most popular, easy-to-use, and effective segmentation methods to enable marketers to analyze customer behavior is RFM analysis. [RFM](https://clevertap.com/rfm/) stands for Recency, Frequency, and Monetary value, each corresponding to some key customer trait. These RFM metrics are important indicators of a customer’s behavior because frequency and monetary value affects a [customer’s lifetime value](https://clevertap.com/blog/customer-lifetime-value/), and recency affects retention, a measure of engagement.

RFM factors illustrate these facts:

* the more recent the purchase, the more responsive the customer is to promotions
* the more frequently the customer buys, the more engaged and satisfied they are
* monetary value differentiates heavy spenders from low-value purchasers

RFM values were calculated for each customer.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure 2.1.

**Frequency** and **Recency** is low because the products sold by the company are products that are used for a long time, such as carpets. For example, you can’t expect a customer to purchase a rug on a monthly basis. In this case, a marketer could give more weight to **Monetary** and **Recency** aspects rather than. Therefore, we will not use **Frequency** scores when calculating the RFM score. Instead, customers with a Frequency value greater than 1 were included in the analysis.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure 2.2.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure 2.3.

According to RFM scores, customers were divided into 10 different segments.

These are *can’t loose them, loyal customers, champions, at risk, need attention, potential loyalists, hibernating, about to sleep, promising*and*new customers.*

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure 2.4.

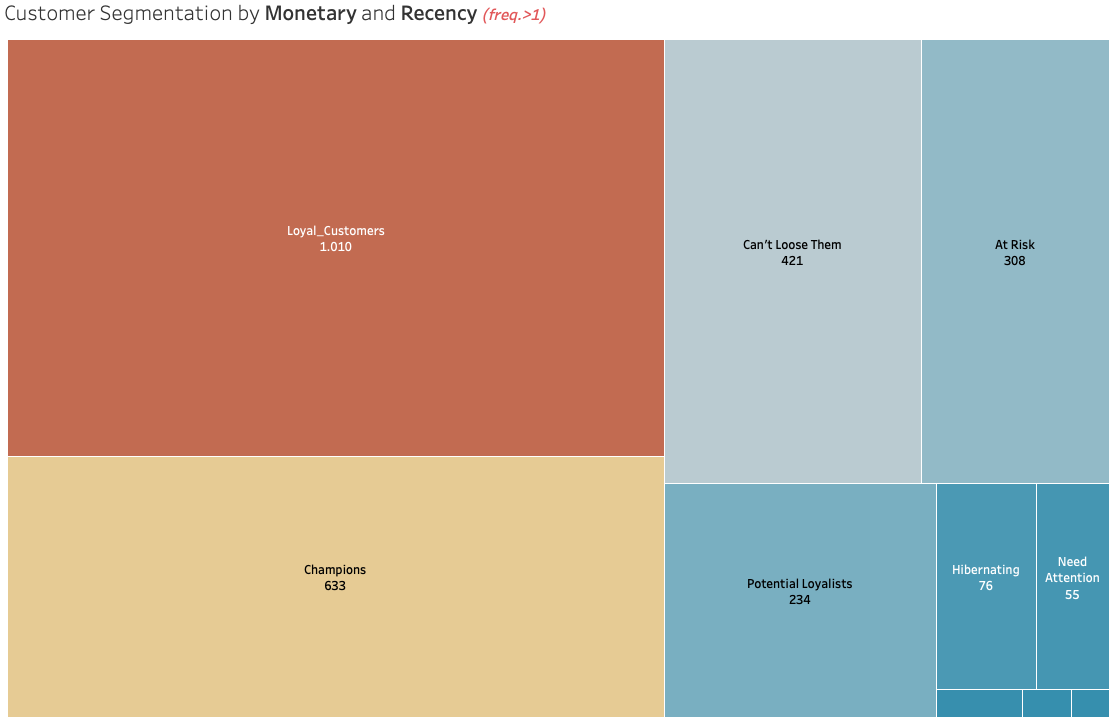


Figure 2.5.

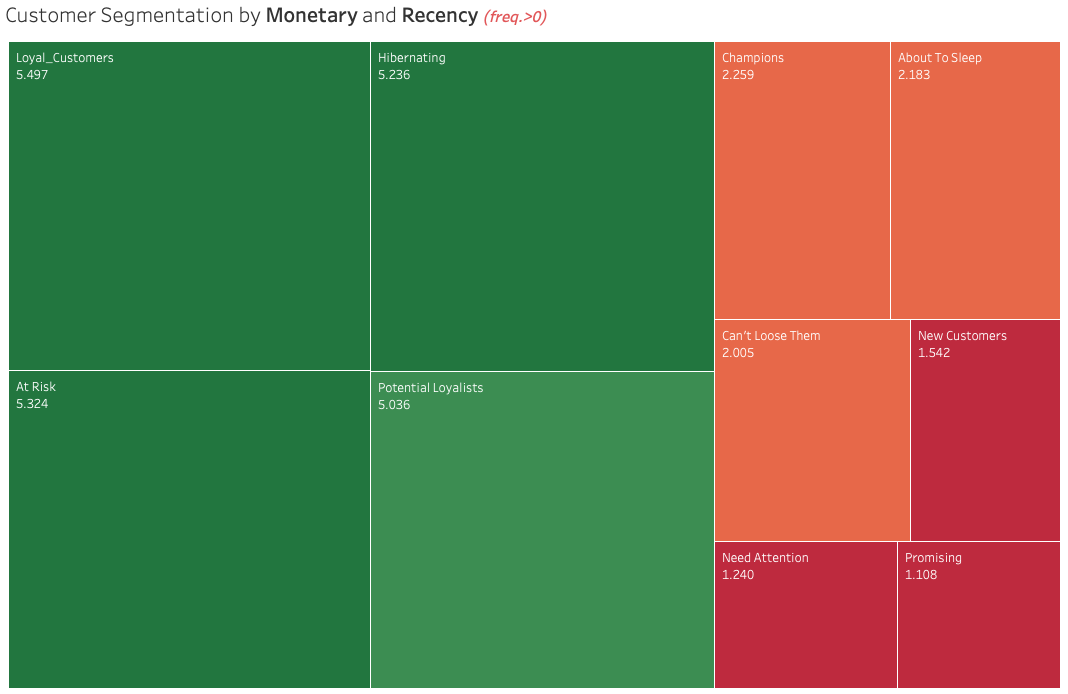


Figure 2.6.

All customers are included in the final picture.

**Recommendations**

* Different complementary products that can be sold with carpets can be added to the product list.
* For customer segmentation, information such as age, gender and zip code can be obtained from a customer.
* It is suggested that the location of the ‘Subscribe’ button can be placed at the top of the website page.
* Since most loyal customers are in the Greater London and South East region and because of the similar average income; The company can apply more advertisements in these regions.

The recommendations for the customers are composed as follows:

* **Champions** are your best customers, who bought most recently, most often, and are heavy spenders. Reward these customers. They can become early adopters for new products and will help promote your brand.
* **Potential Loyalists** are your recent customers with average frequency and who spent a good amount. Offer membership or loyalty programs or recommend related products to upsell them and help them become your Loyalists or Champions.
* **New Customers** are your customers who have a high overall RFM score but are not frequent shoppers. Start building relationships with these customers by providing onboarding support and special offers to increase their visits.
* **At Risk Customers** are your customers who purchased often and spent big amounts, but haven’t purchased recently. Send them personalized reactivation campaigns to reconnect, and offer renewals and helpful products to encourage another purchase.
* **Can’t Lose Them** are customers who used to visit and purchase quite often, but haven’t been visiting recently. Bring them back with relevant promotions, and run surveys to find out what went wrong and avoid losing them to a competitor.