Project Documentation

Data description:

This data is behavior data for E-commerce customers for two months (October and November 2019) from a medium cosmetics online store hosted by REES46 niche-specific personalization engine Platform.

Each row in the data represents an event.

All events are related to products and users. Each event is like many\_to-many relations between products and users.

A session can have multiple purchase events if it's a single order.

Columns:

| **event\_time** | Time when event happened at (in UTC). |
| --- | --- |
| **event\_type** | **View**: a user viewed a product  **Cart**: a user added a product to the cart  **remove\_from\_cart**: a user removed a product from shopping cart  purchase - a user purchased a product |
| **product\_id** | ID of a product |
| **category\_id** | Product's category ID |
| **category\_code** | Product's category taxonomy (code name) if it was possible to make it. Usually present for meaningful categories and skipped for different kinds of accessories. |
| **brand** | Downcased string of brand name. |
| **price** | Float price of a product. Present. |
| **user\_id** | Permanent user ID. |
| **user\_session** | Temporary user's session ID. Same for each user's session. Is changed every time user come back to online store from a long pause. |

1-Data Validation:

> Columns datatypes:

‘event\_time’ : object --> (Datetime)

- ‘product-id’ , ‘category\_id’ , ‘user\_id’: (int64) --> (Object)

> Validating each column that it consists of the right data:

Event\_time:

- Is correct it is between Oct and Nov for the year 2019.

User\_id and product\_id:

- Does not have a specific length or pattern but it’s correct.

Coategory\_id:

- It is a fixed number of 19 digits, all correct.

Price:

- There are negative values, it may be due to returned items (analyze more)

- Negative values are in front of Purchase events only

- There are 0 values and they are in front of (view, cart, remove\_from\_cart ) events only (it may not be available for purchase, still need more investigation)

- Most of the values are outliers, High prices may be due to high-end products that are actually expensive (analyze more)

User\_session:

- There are some null values, that may be due to an error in the system, at this time it didn’t catch the session\_id maybe because of load on the platform.

After doing some search we have reached that there are more than one way to group and make user\_sessions here in our data the user\_session follows the User-based grouping method

User-based grouping: Sessions are grouped based on user ID to identify sessions performed by the same user. This grouping can help analyze the behavior of individual users, their preferences, and their lifetime value.

Brand:

- There are null values

Category\_code:

- Lots of null values

2-Data Cleaning:

> Duplicates:

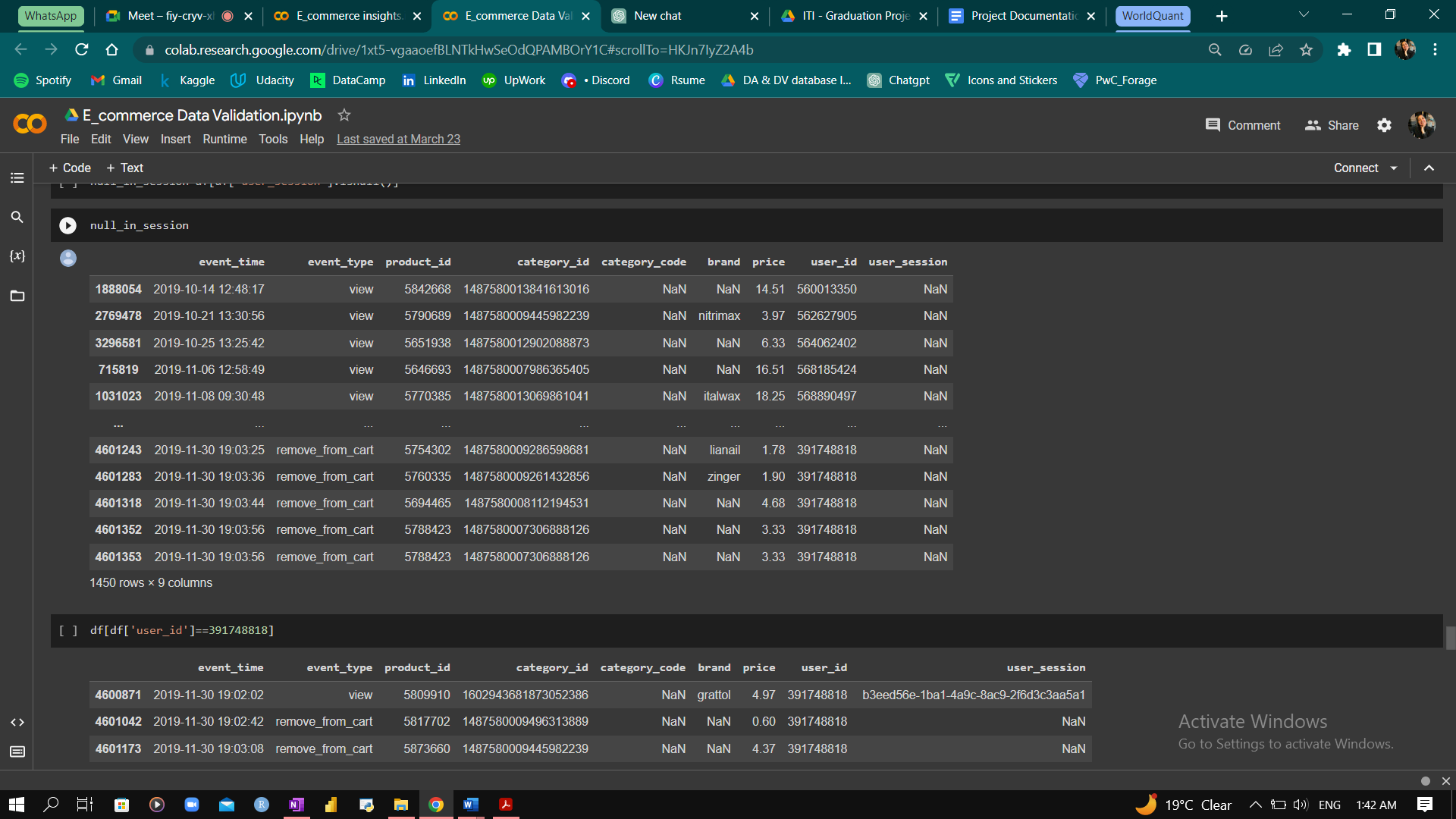
- There are 459,848 full duplicate rows in the data but turns out these values were not duplicated actually it has a meaning, this data is collected by a machine which is the platform this means that for this exact time one user purchased a certain product with a quantity > 1

- This also applies to the events of cart , remove\_from\_cart as the user may select more than item at the same time and remove/add it to the cart so it still makes sense

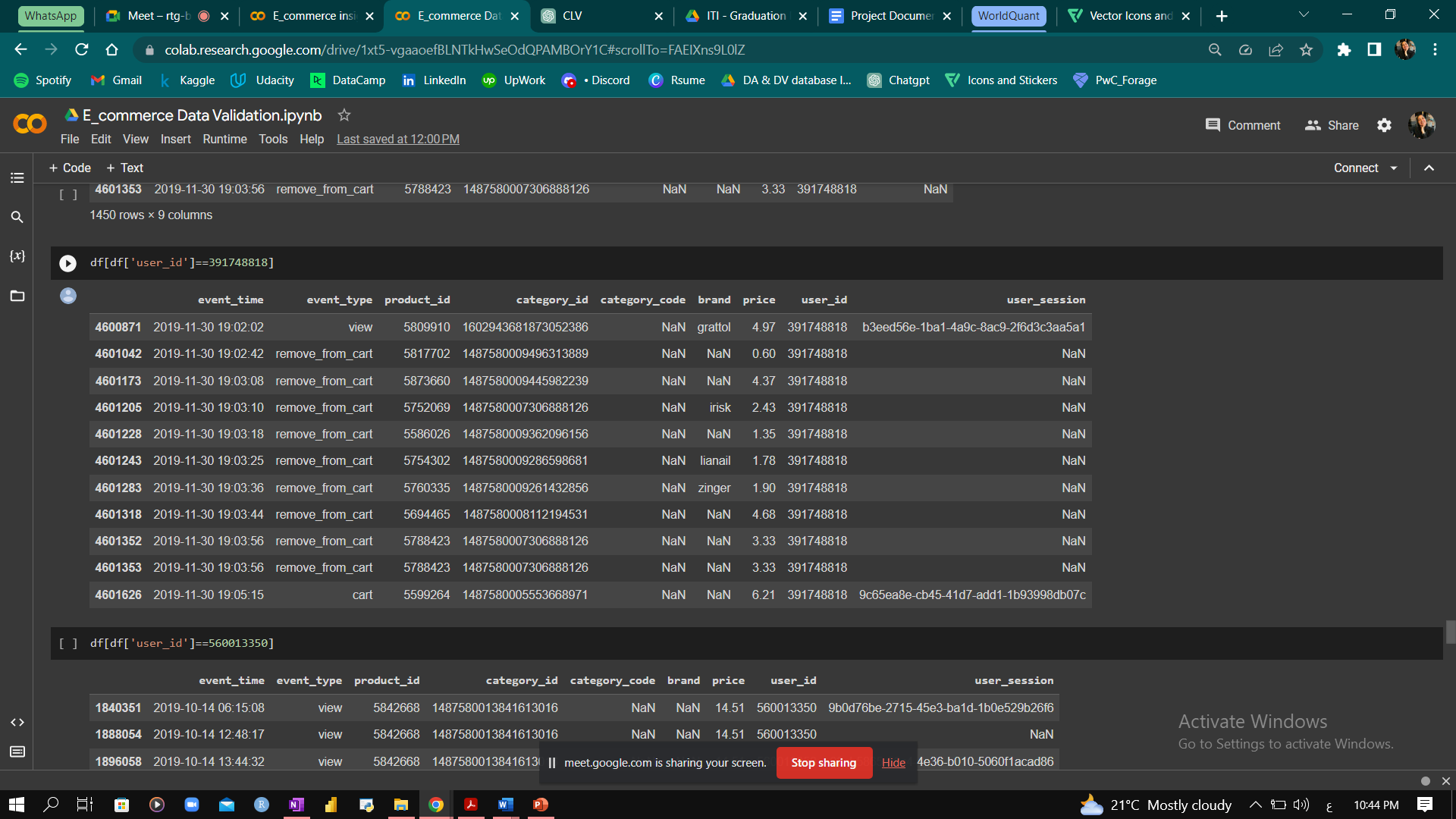
- But when it comes to view it doesn’t make sense if it is viewed in the exact time more than once most probably it’s because the user refreshed the page multiple times or due to error but either way counting it will be misleading and that is why we dropped the duplicates in the view event

> Missing values:

- user\_session: as we noticed in the validation process the user\_session missing values are due to system error this clear here



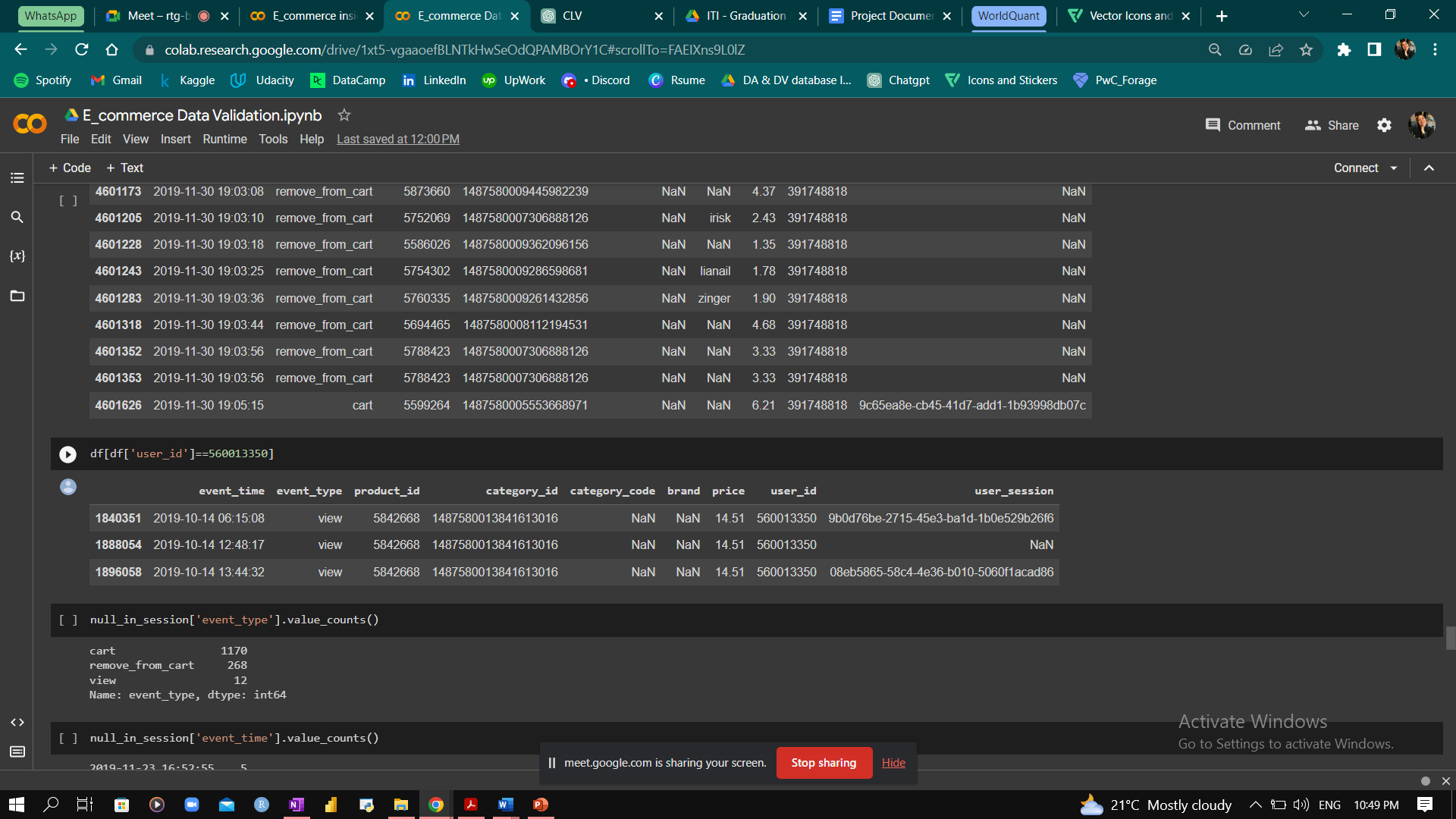
we then checked for a user



at exact time the user\_session was not recorded for this action

we can not be sure which user\_session to fill the nan values with because it may be the first session or the last and there isn’t enough evidence to make the decision

Here is another user to prove the point



At the end we decided to leave it as is.

- Brand: the brand column had 3,645,191 missing values

we checked if we can fill any

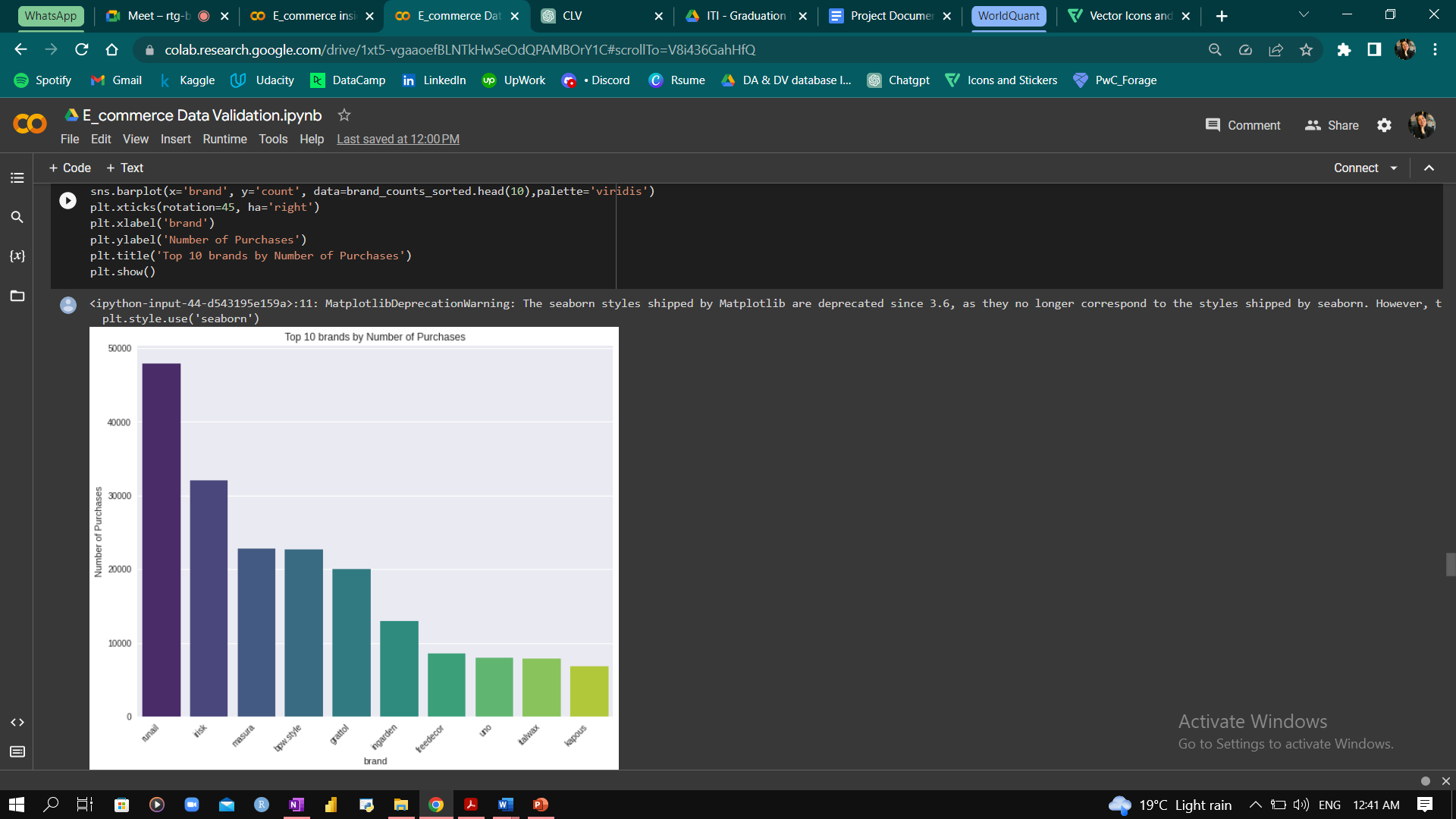
to do so we first checked if the missing values in the brand column exists in all the appearances of the product for example, the product may have missing values in their brand for instance but it exist for another instance

and turns out that this was true, so we filled those values and there were 9,094 rows filled

leaving us with 3,636,097 nan values that can not be imputed

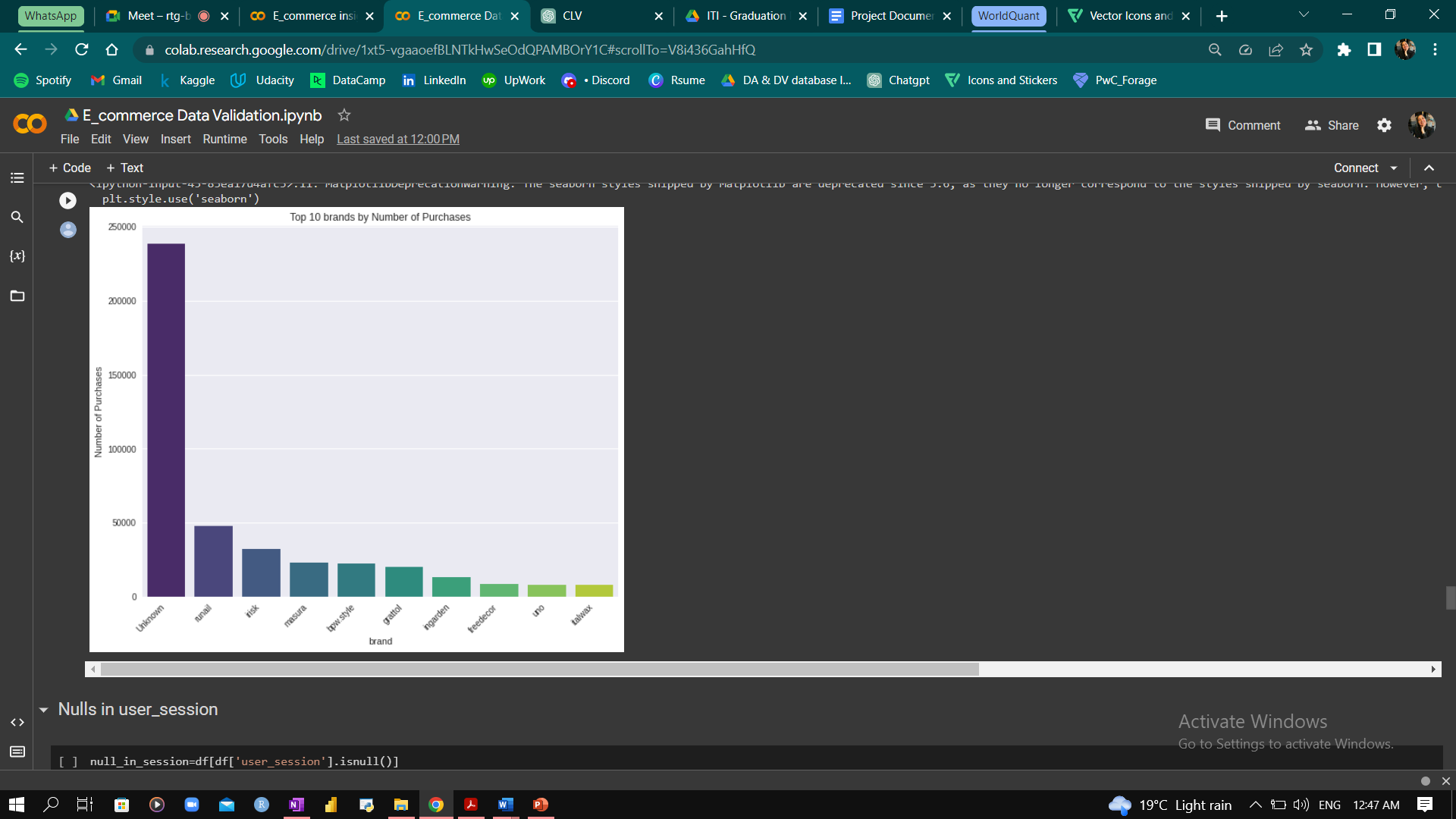
we then decided to give them the label unknown to see if that makes a difference or affects our insights in any way

Here is the top 10 brands customers purchase from before filling the missing values with unknown



as the graph shows the most brand customers purchase from is **Runail**

and here is after filling the missing values with unknown

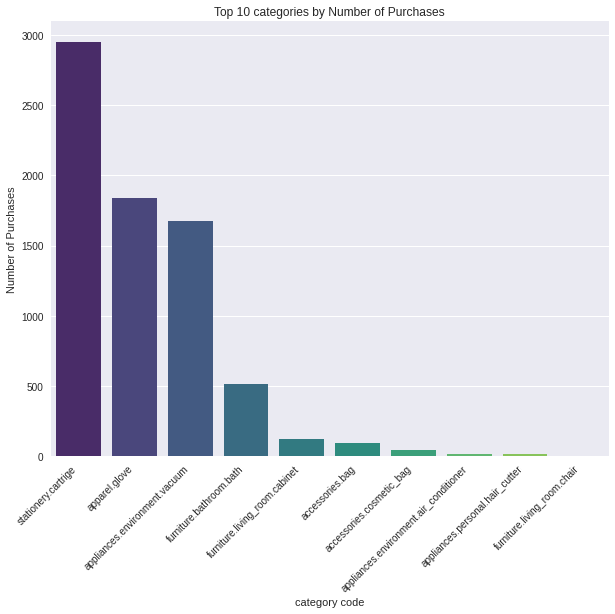


This shows that the most brands customers purchase from are unknown brands, so to benefit from this insight our business needs to identify each unknown brand to know the actual highest purchased brand.

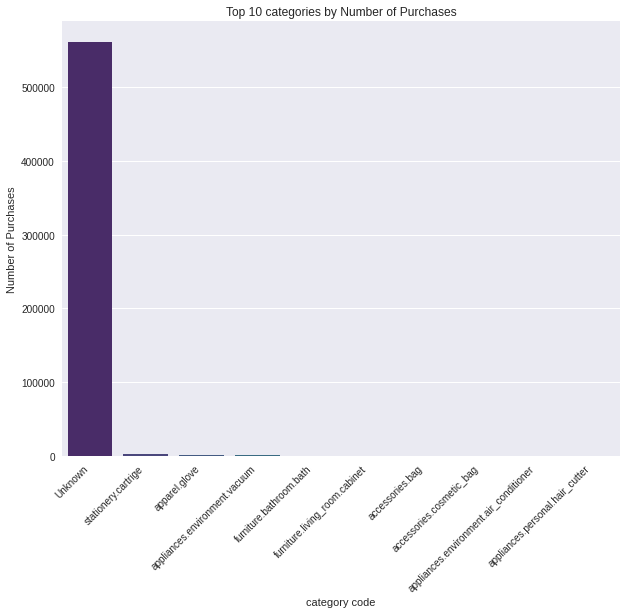
Category\_code: there are alot of missing values in the category\_code column we tried to check if we can fill any of it by the same way we did with the brand but turns out there were no appearance for a category\_code for some instance and didn’t for another

so there were no way of filling them,

before filling with unknown, the top categories are



After filling the missing values with unknown



had a huge effect on the insights as it shows that the most categories customers purchase from are unknown categories

So again to benefit from this insight our business needs to identify each unknown category\_code to know the actual highest categories customers purchase.

> Outliers:

Price

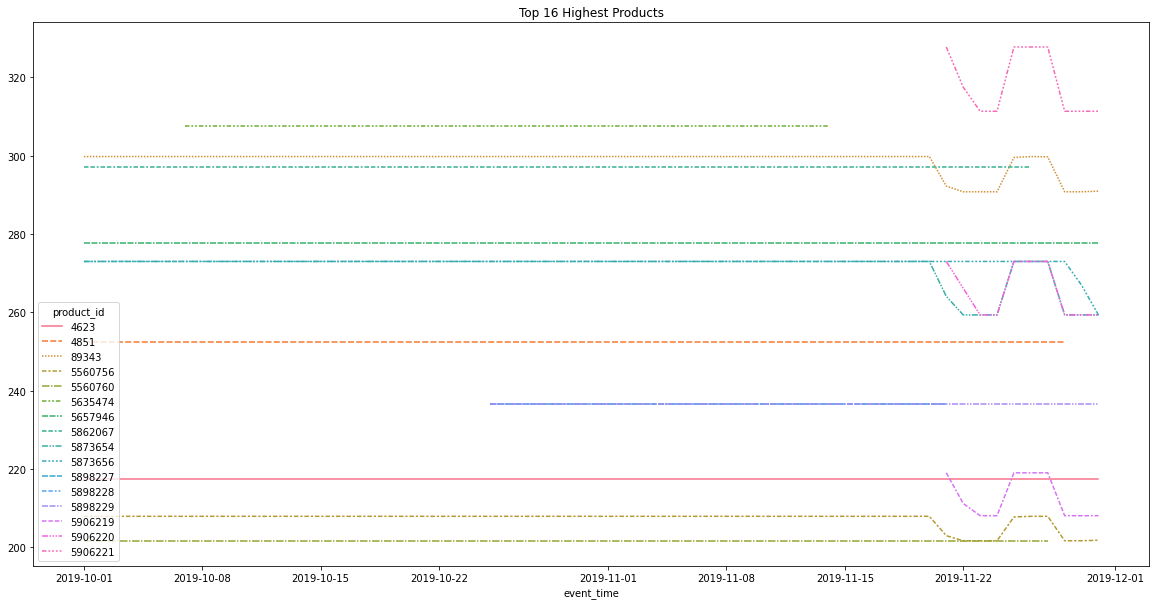
- Negative values were because when a customer returns a product or cancels an order, the platform issues a refund for the original purchase amount. Refunds are typically recorded as negative values in the price column, to indicate that money is leaving the company rather than being received.

- The 0 values that were in front of (view, cart, remove\_from\_cart ) events only, most probably these products were being used as promotional items, for example as part of a "buy one get one free" offer. In this case, the "purchase" event may have been associated with the product that was actually paid for, rather than the free product.

Also a customer may have added a product to their cart during a sale, resulting in a discounted price being displayed.

- We made sure that the very high prices in the data were actually the price for these products and not just an outlier

and to do so we checked to see if this was it’s price since the beginning of its appearance or not



turns out that this is the actual price for these products.

--------------------------------------------

**Analysis and insights**

Customer analysis

We started our analysis with the customer to better understand their behavior on our website

We first wanted to see what is the actions our customers do most on the website

View: 45.1 %

Purchase: 6.5%

Remove from cart: 19.3%

Cart: 29.1%

We then focused our analysis more on the purchase event since it is the action that makes a viewer an actual buyer or (customer) in the two months of our data

And based on that we split the customers to

Regular customers: customers who purchased more than once

New customers: customers who purchased only once

Then we focused on each type of user to better understand them and to offer them best promotions based on their preferences.

First: New Customers

We want to see what products new customers purchase the most?

What is the top category they purchase from?

And on which day? Turns out most of the new customers visit our website at the end of the month, This may be because most of the employees receive their salaries by this time.

This analysis gives us the insight to what products to give promotions on and on which time

And this to try to turn the new customers to regular customers and make them come back and purchase from our website again.

Second: Regular Customers

We analyzed the behavior of the regular customers to know what products they purchase the most and on which day? Turns out the day with highest purchasing by regular customers is 22 of November

And the purpose behind that was to try and keep them as regular customers on my website by offering them promotions based on their preferences.

After knowing more about the behavior of each of the two groups we wanted to analyze more about the time of visiting and purchasing

Which days do customers visit my website more?

And on which days they purchase more? In which hour?

Turns out

- Highest hours customers purchase at are two intervals from (10 am to 12 pm) and from (6 pm to 8 pm)

- The daily views on our website were on Thursday

- Daily purchases were on Friday

- This makes a high conversion rate from Thursday to Friday

Which means that customers enter our website on Thursday navigate and view products then decide on buying them on Friday

What is meant by conversion rate is when a user turns from being a viewer to being a buyer, and this is what happens from Thursday to Friday

From this analysis we knew that Friday was the highest day of week with purchasing since our data consist of the customers behavior in two months October and November and it is known globally that at the last Friday of November there is a huge sale happens by lots of brands which is Black Friday so we analyzed more about the revenue generated at this day

Turns out that

- The total revenue in the two months is around 2M $

- and total sales on black Friday was 3% of the total sales which is around 102K $

Product analysis:

We then analyzed more about the products that generated high revenue

Turns out

- Highest product by revenue has this product\_id 5560754 this product generated high revenue because it has a high price of 194.44 $ even though it wasn’t purchased with a lot

We analyzed more about this product and turns out that there was a discount on this product on the Black Friday and its price decreased to 188.6 $ and it was the highest product that generated revenue on the Black Friday

This can tell us that customers buy high priced products more when there is a discount on it

We can use this insight to do more discounts on high-end products to generate more revenue.

We then looked at

- The highest brands by revenue the highest was (Runail)

-The highest categories by revenue and highest was (appliances.environment.vacuum)

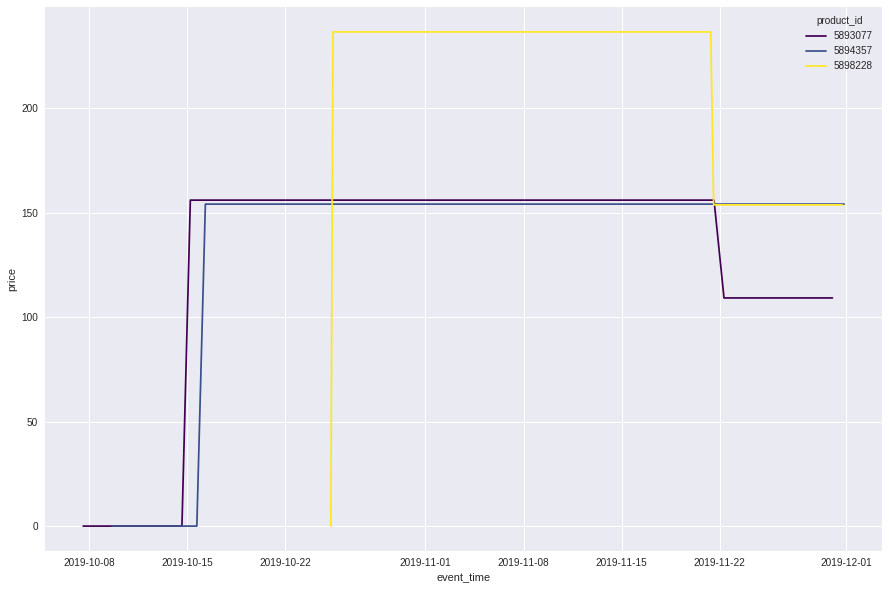
We also wanted to know if the highest viewed products are also the most purchased?

Turns out yes, it is. This means that what customers view a lot there is a high chance they will purchase this product.

The discount in the product price raised the question, Are there any other trends in the price of other products?

We looked at the products that had a price increase by 50% of its minimum price

We found that there were products whose first appearance in the data was with price 0 which we explained before because of (buy one get one free offer) then these products returned to their original price and around the time of the Black Friday a small discount happened in these products prices.



Also there was a general trend in most of the products prices which is the discount around the Black Friday then returns to its original price after it ends.

Campaign launching:

After understanding the customer behavior we now have a big picture of the preferences of our customers

but we wanted to feature more products that perform well in different areas to appeal to a variety of customers.

That’s why we looked at

- Products with the highest conversion rate (from viewing to purchasing) and feature them in a campaign.

- Most popular products among customers based on number of purchases.

- Products that generate the highest revenue.

By this we will be sure that we covered a variety of products which meet the different needs of our customers.

Customer segmentation:

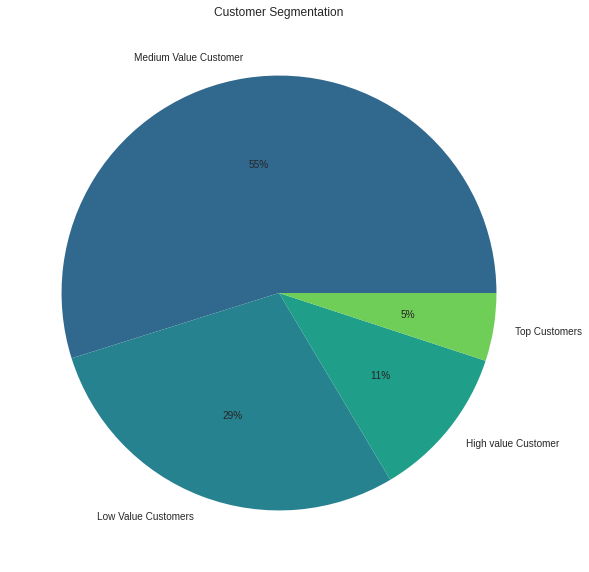
To launch better campaigns that target the right customers it was important to segment our customers

to do so we used the RFM model

we gave weights to each rank based on what we wanted to bring out the most to calculate the RFM score

monetary had the most weight since we wanted our valuable customers to be the one who brings us more money

the results was



Top Customers 5% : We can reward them by making giveaways and they will have the priority of winning them, they can also be promoters to the brands, Most likely to send referrals.

High value customers 11% : Offer membership to keep them engaged, offer them personalized recommendations.

Medium value customers 55% : Offer coupons and personalized recommendations so we don’t lose them to competition

Low value customers 29% : Provide helpful resources on the site and, Send them personalized emails.

Lost customers 0% : Revive interest with reach out campaign.

The lost customers segment did not appear in the months of October and November.

CLV to measure customers loyalty:

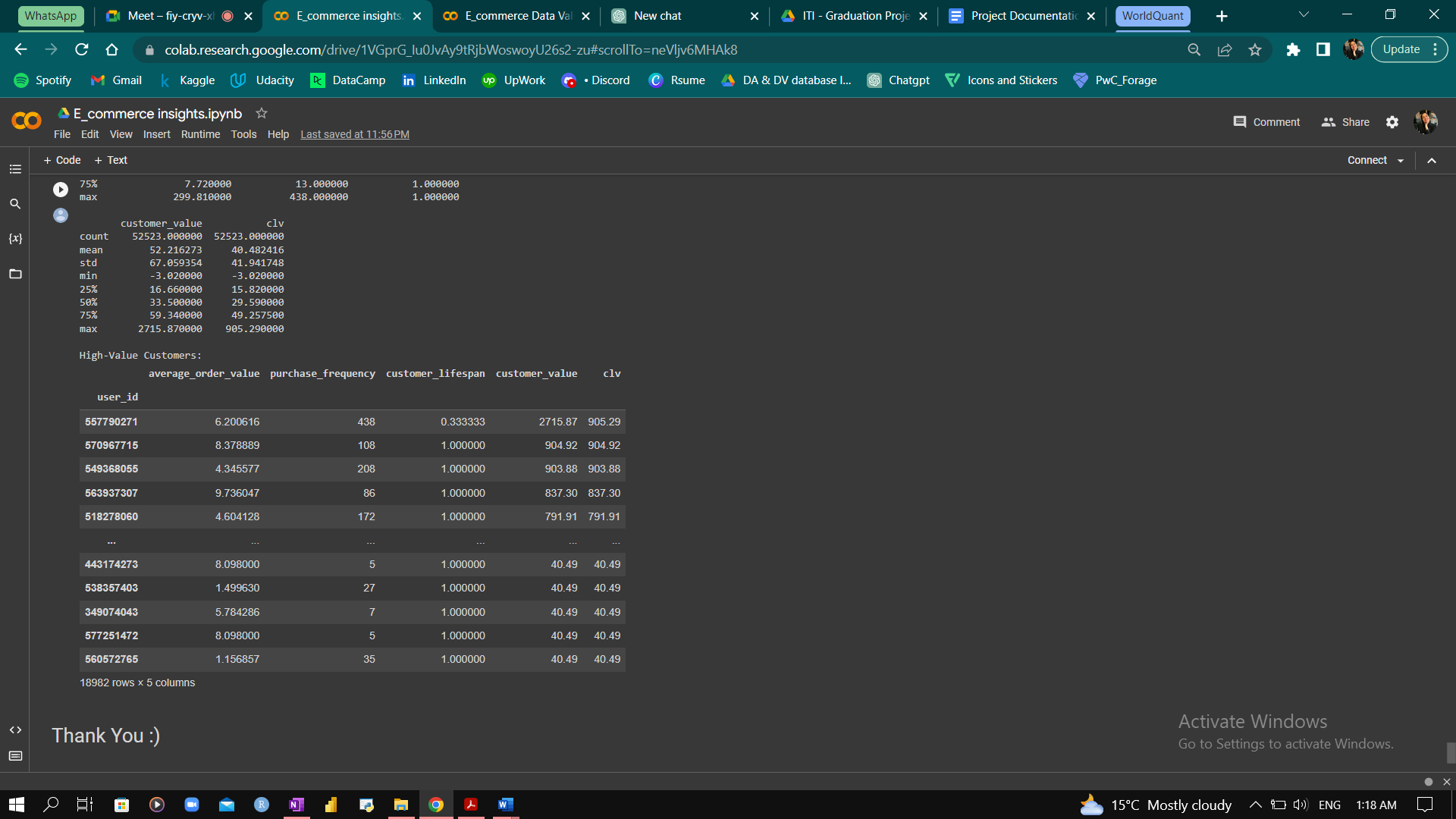
CLV is a metric used to estimate the total value a customer will bring to a business over their entire lifetime. There are several ways to calculate CLV, but one common approach is to use the following formula:

CLV = (average order value) x (number of repeat purchases) x (customer lifetime)

Here, the average order value is the average amount of money spent per order, the number of repeat purchases is the number of times a customer makes a purchase, and the customer lifetime is the length of time a customer continues to make purchases.

To calculate the average order value, you can sum the total revenue from all orders and divide by the total number of orders. To calculate the number of repeat purchases, you can count the number of orders made by each customer and then calculate the average number of orders per customer. To calculate the customer lifetime, you can use a variety of methods, such as the average time between purchases or the average time a customer remains active.

Once you have calculated the three components (average order value, number of repeat purchases, and customer lifetime), you can multiply them together to get the CLV.



Recommendation System:

Building a simple recommendation system will bring more benefits to our business because it will help make the experience on our website more personalized.

To do so we used the Association Rule and implemented it using the Apriori algorithm and FP growth algorithm

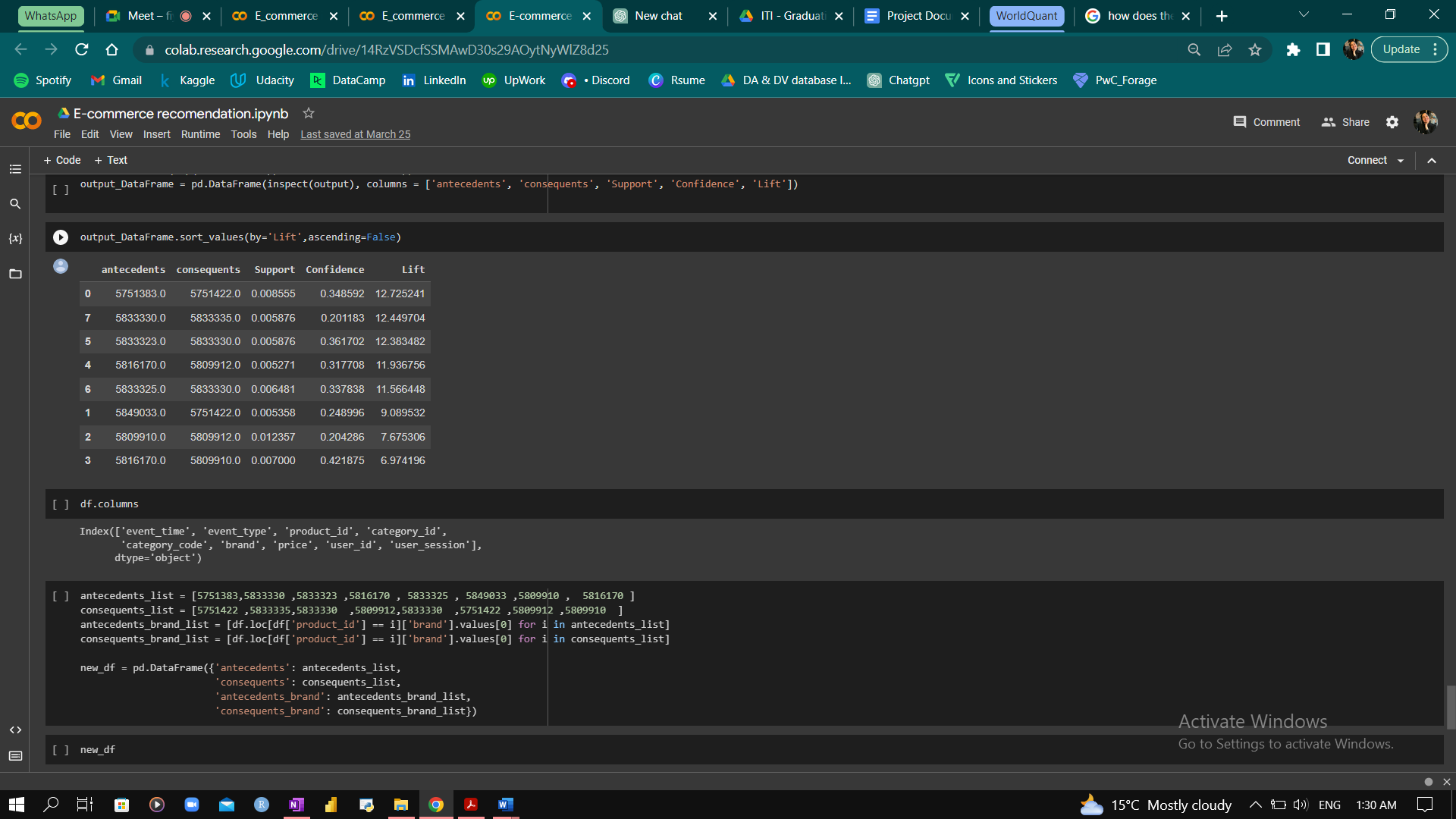
Apriori:

we started with the Apriori algorithm, we first gave it the entire data it didn’t perform well and it was hard for it to recognize any patterns

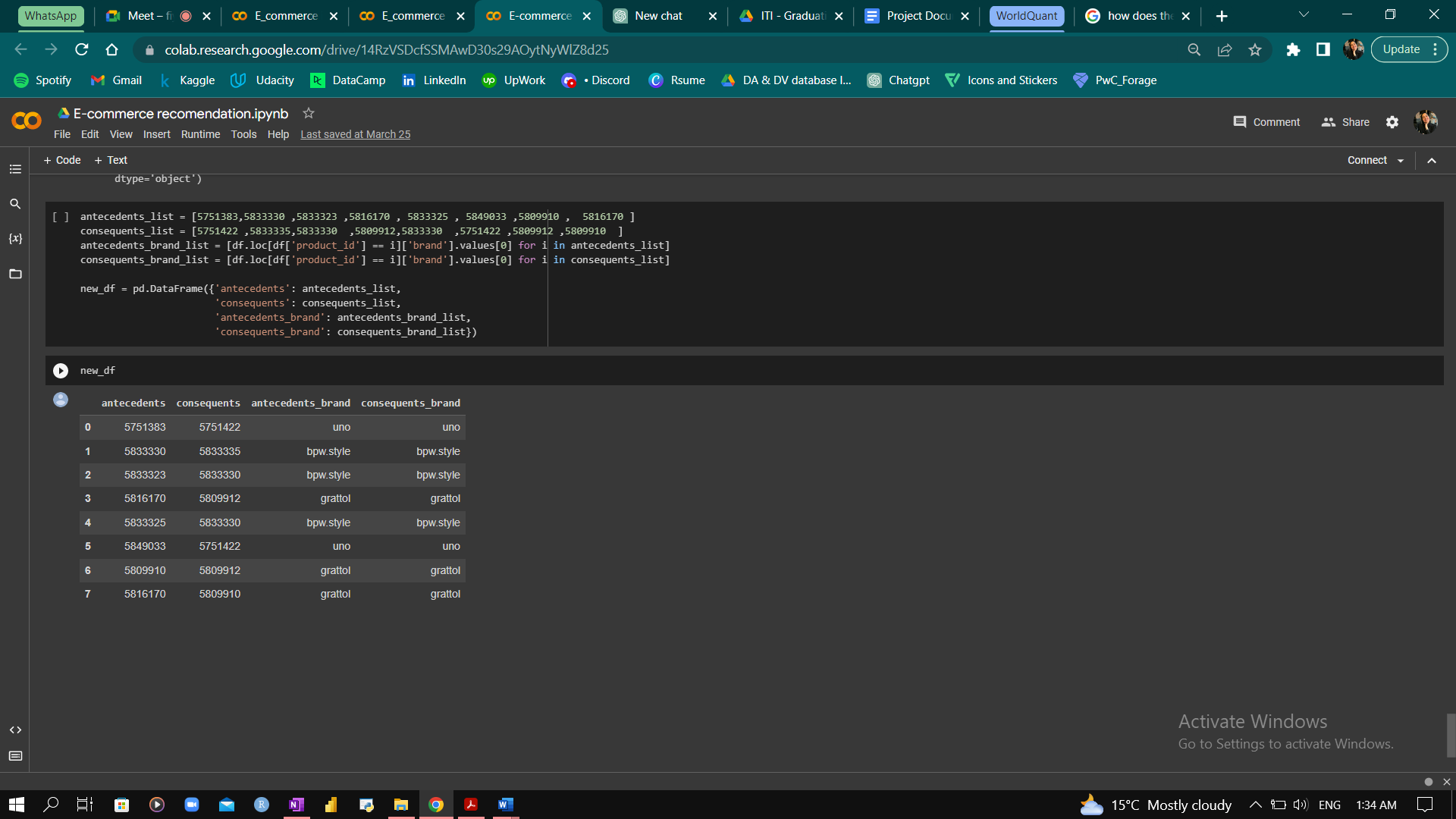
then we only took the data from 15th of November to 24th of November because at this time the purchasing rate was high

and turns out that this was a good decision, it performed much better

Results:



the products that appeared here turns out that each pair is from the same brand



FP Growth:

we tried the FP growth algorithm next, and it almost returned the same results

