**Analyzing the Flipkart Sales Dataset**

***Introduction***

Flipkart is an electronic commerce company headquartered in Bengaluru, India. It was founded in October 2007 by Sachin Bansal and Binny Bansal (no relation). Flipkart has launched its own product range under the name “DigiFlip” with products including tablets, USB flash drives, and laptop bags. As of April 2017, the company was valued at $11.6 billion.

Flipkart has established itself as a renowned E-commerce company in India through TV ads, online branding, and through its presence on social media. Brand activities like the “Big billion day” have really increased the brand recall of the company.

After its 2014 Big Billion Sale, Flipkart carried out a second Big Billion Sale. Where it is reported that they saw a business turnover of $300 million in gross merchandise volume**.**

***Overview***

An analysis of 20,000 products of Flipkart. In this, we analyze the various factors on which the overall rating and product rating of any product depends. This analysis will help the manager to know the factors on which the rating of the Product depends, and with this, they can increase their market value.

***Dataset***

Data has been collected from the Kaggle website. Kaggle is a platform for predictive modeling and analytics competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users. This crowdsourcing approach relies on the fact that there are countless strategies that can be applied to any predictive modeling task, and it is impossible to know beforehand which technique or analyst will be most effective.

***About Dataset***

***Context***

*This is a pre-crawled dataset, taken as subset of a bigger*[*dataset (more than 5.8 million products)*](https://www.promptcloud.com/datastock-access-ready-to-use-datasets/?utm_source=fl-kaggle&utm_medium=referral)*that was created by extracting data from Flipkart.com, a leading Indian eCommerce store.*

**Content**

This dataset has following fields:

* product\_url
* product\_name
* product\_category\_tree
* pid
* retail\_price
* discounted\_price
* image
* is\_FK\_Advantage\_product
* description
* product\_rating
* overall\_rating
* brand
* product\_specifications

***Features***

* 20K products
* With properly leveled catalog information
* Products come with their images! There is a list of urls for each product (Some of the URL's are broken tho)

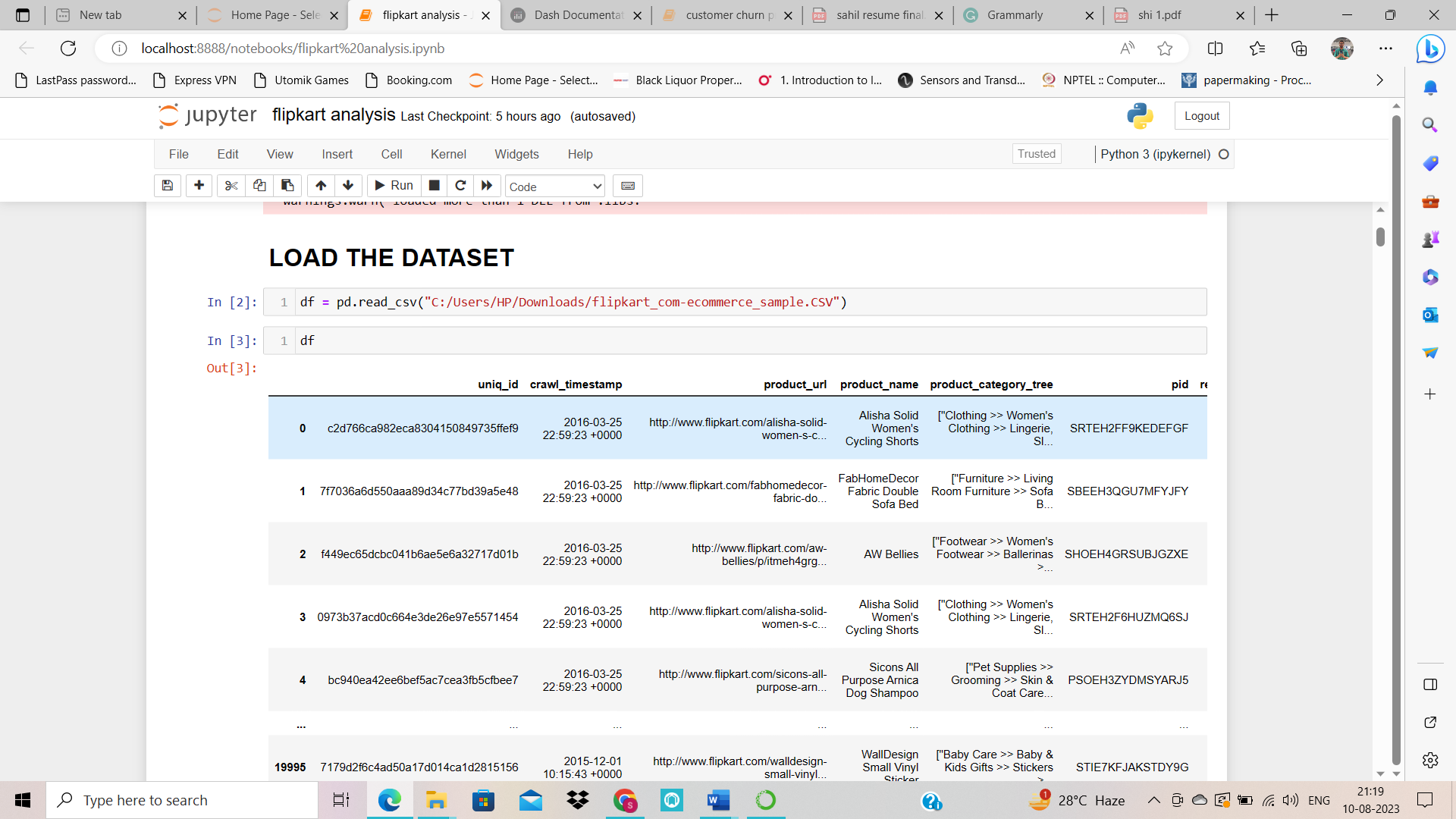
**Let's learn to do a deep analysis of the Flipkart sales dataset using the pandas and Plotly libraries.**

Data The decision-makers may better comprehend how the business data is analyzed to make business choices by visualizing the outcome. Additionally, it emphasizes business analytics to identify trouble spots. As a result, we'll also utilize the Plotly library in addition to pandas.

You should provide a thorough and simple-to-understand summary on the dataset that addresses the following topics:

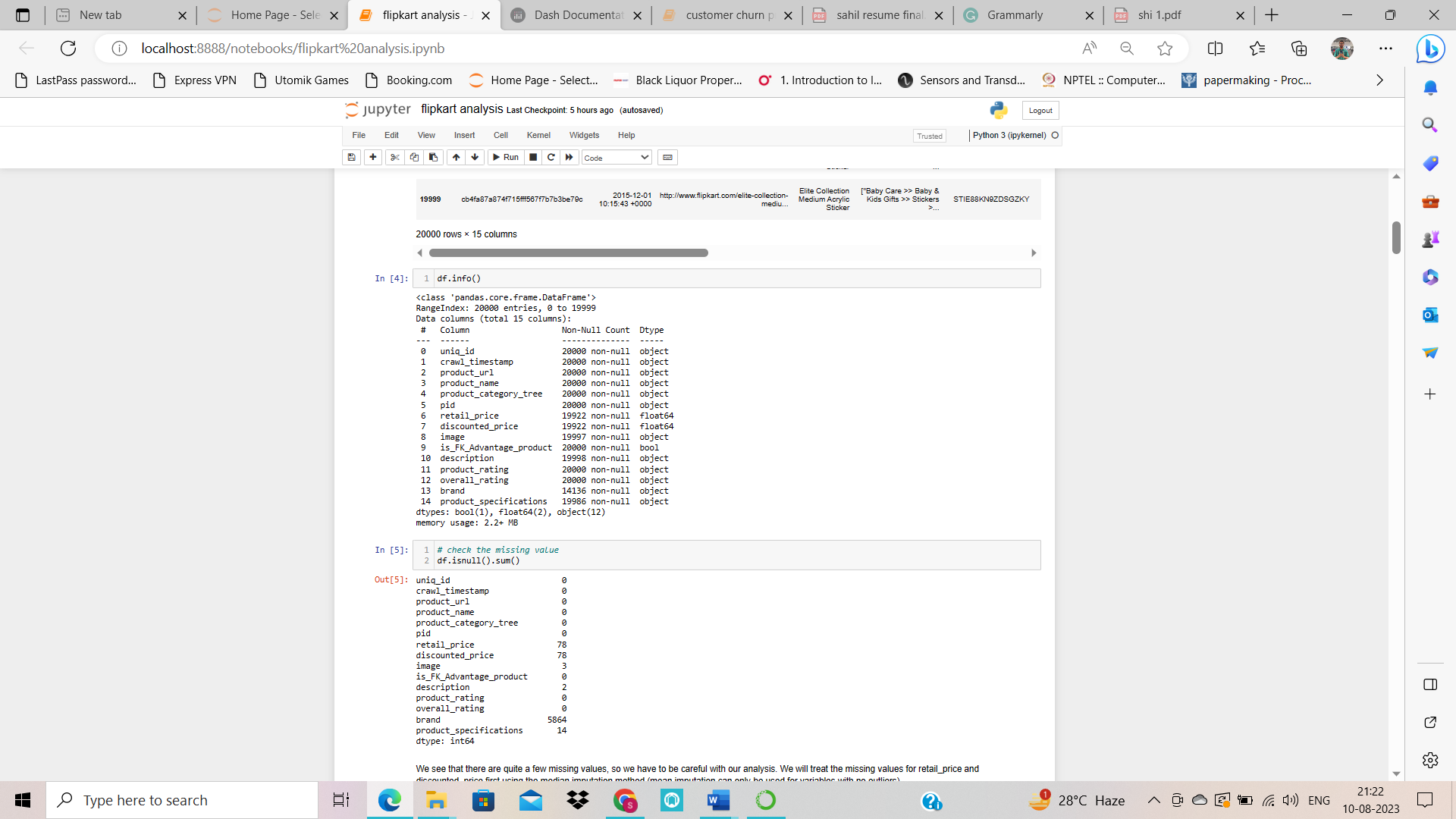
1. What is the top-performing products and brands?
2. Which brands are offering the highest discount?
3. A thorough analysis of product ratings:
   1. What is the proportion of 5-star rated products to the total products?
   2. What are the best and worst-performing products and brands in terms of ratings?
   3. How many numbers of products are sold for different types of ratings?
4. What are some of the target customers that spend the most money on Flipkart?
5. Is there any trend associated with the retail price and discount price over the months?
6. When are customers the most active during the day?

Dataset look like



This dataset has columns for uniq\_id, timestamp, item name, retail price, reduced price, and more, as can be seen.

There are several columns with very accurate and thorough details on each ordered item. This enables a more thorough analysis of the data. Despite the columns' level of information, we may not have all of the cells filled with these helpful data.

Let's see if we have any missing values.

We must be cautious while doing our analysis since there are a lot of missing variables. The median imputation technique will be used to address the missing values for ‘retail\_price’ and ‘discounted\_price’ initially.

Our dataset ‘discount\_percentage’ will get a new column. This will provide the percentage of the discount that various companies are offering.

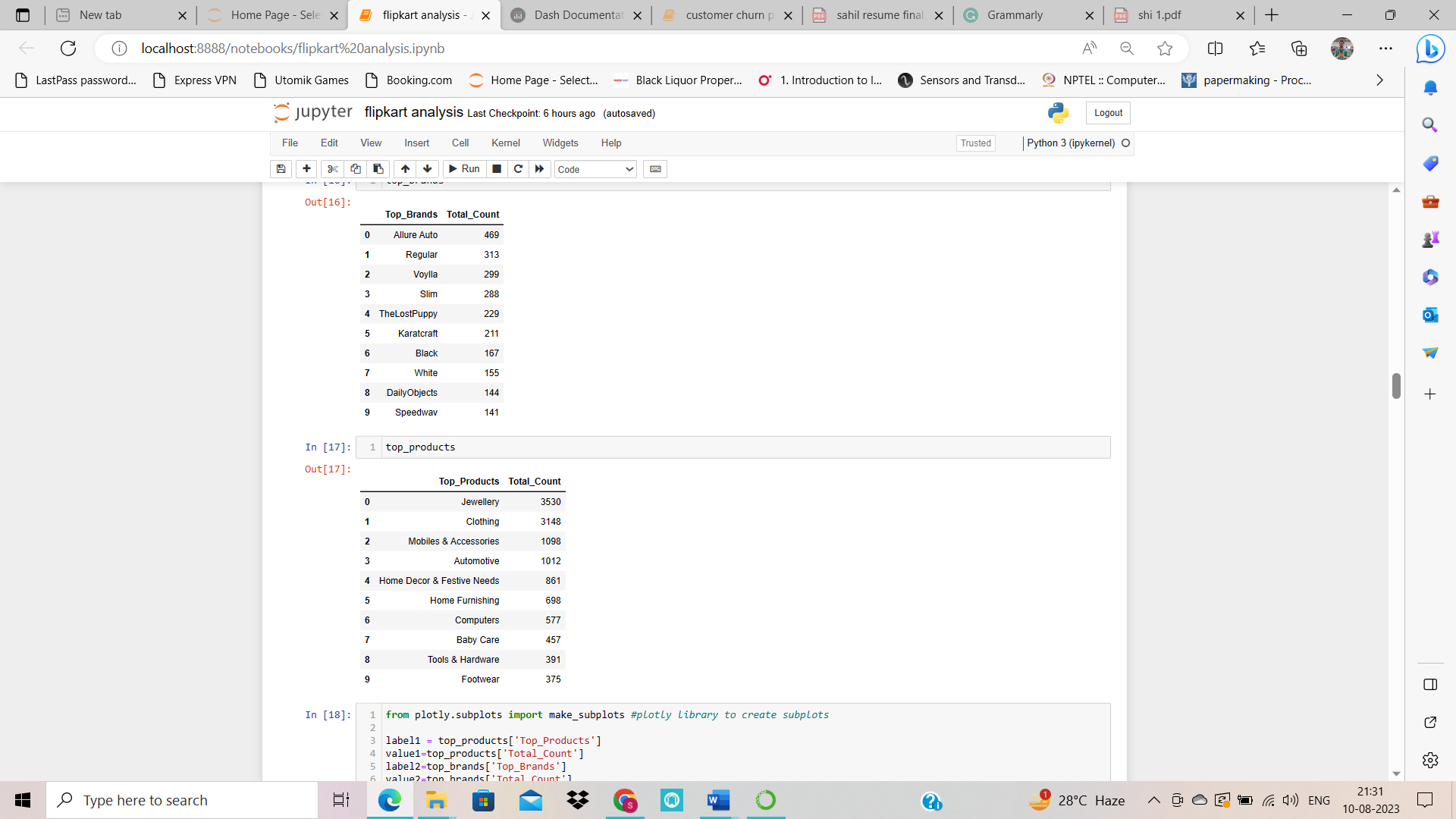
Extract some more columns like ‘Time’, ‘date’,’ year’,’ month’ and lastly ‘main\_category ‘ with the help of the columns ‘crawl\_timestamp’ and ‘product\_category\_tree’ respectively.

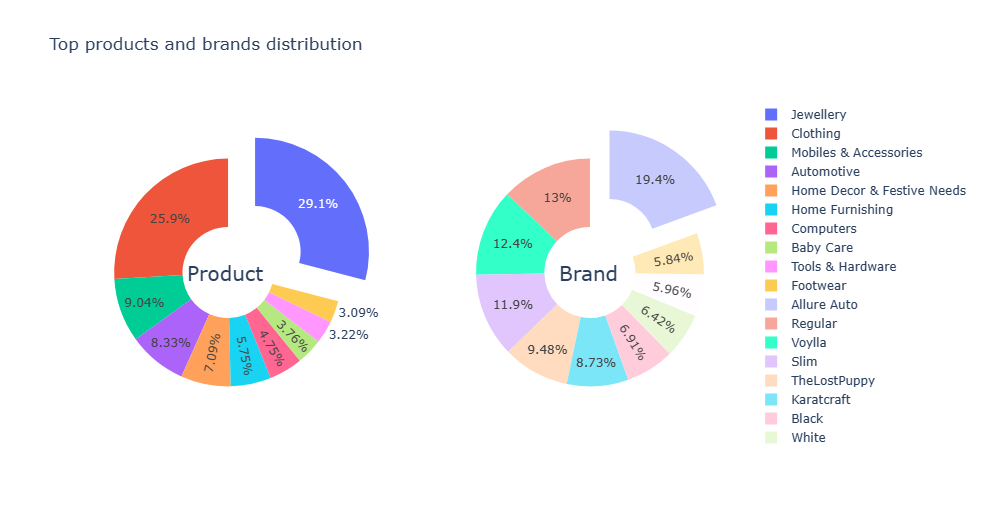
***Analyzing the dataset***

* **What is the top-performing products and brands?**

It's crucial to remember that only 1849 rows out of the total ‘product\_rating’ in our dataset have relevant ratings. No rating is available is the value in the remaining rows. The codes ‘df['product\_rating']’ options. This omits a large amount of information. Because of this, it's crucial to examine all companies and goods, regardless of their ratings.

In the section that follows, we'll examine how ratings affect certain goods and brands.





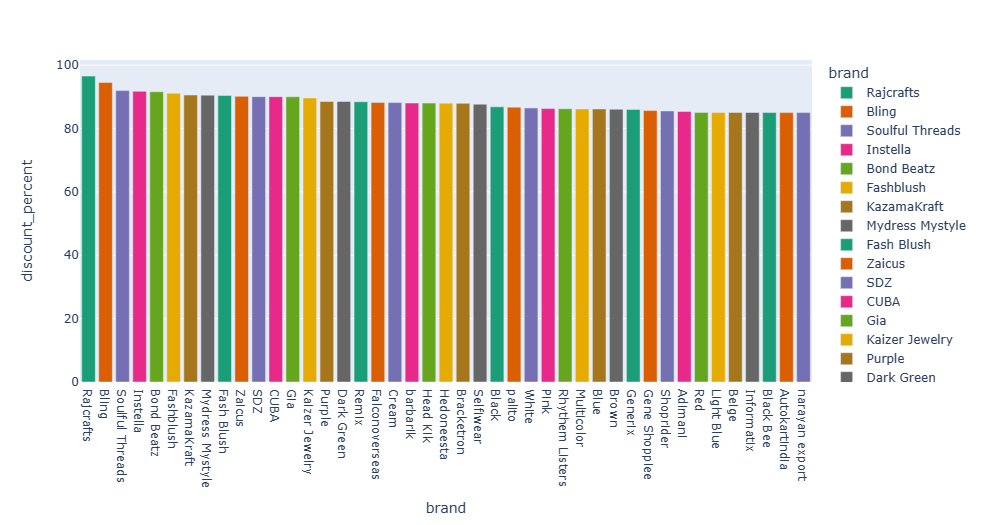
Using a Pie Chart with a Hole (also known as a Donut Chart), Plotly has made it incredibly simple to evaluate and comprehend the outcome. It is simple to analyze the chart and come to meaningful conclusions.

We see that "Clothing" is the top item people are buying, while "Regular" is the top brand. The pie chart shows the other top-performing goods and brands.

* **Which brands are offering the highest discount?**

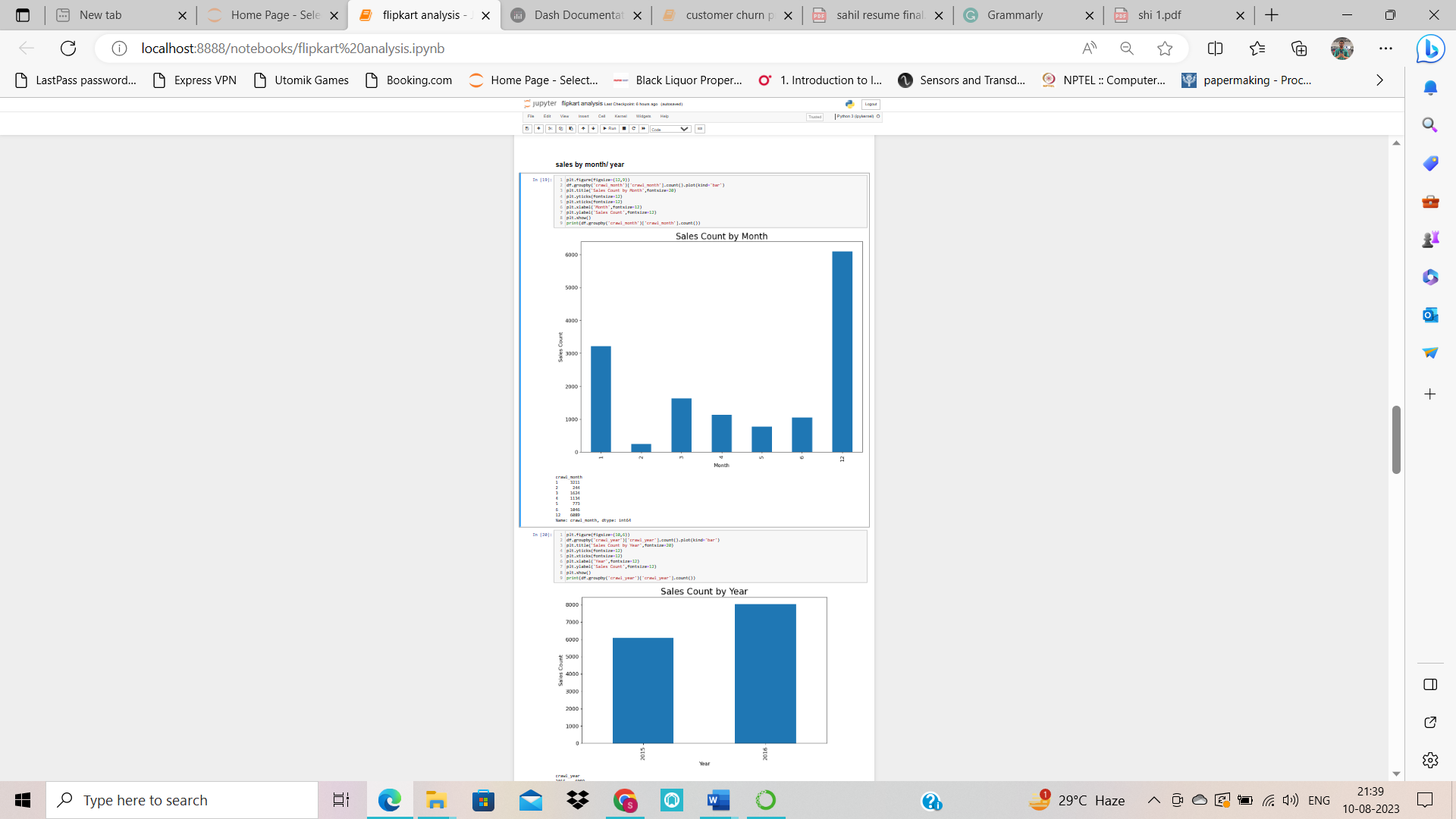
We may promote and suggest these target products to our engaged clients, which is one of the main factors contributing to the importance of the brands delivering the largest discount in our study. Due to the huge discount's allure, more people may visit the purchasing website. When you inform a consumer that they may save money, you are more likely to catch their attention and ultimately boost sales.

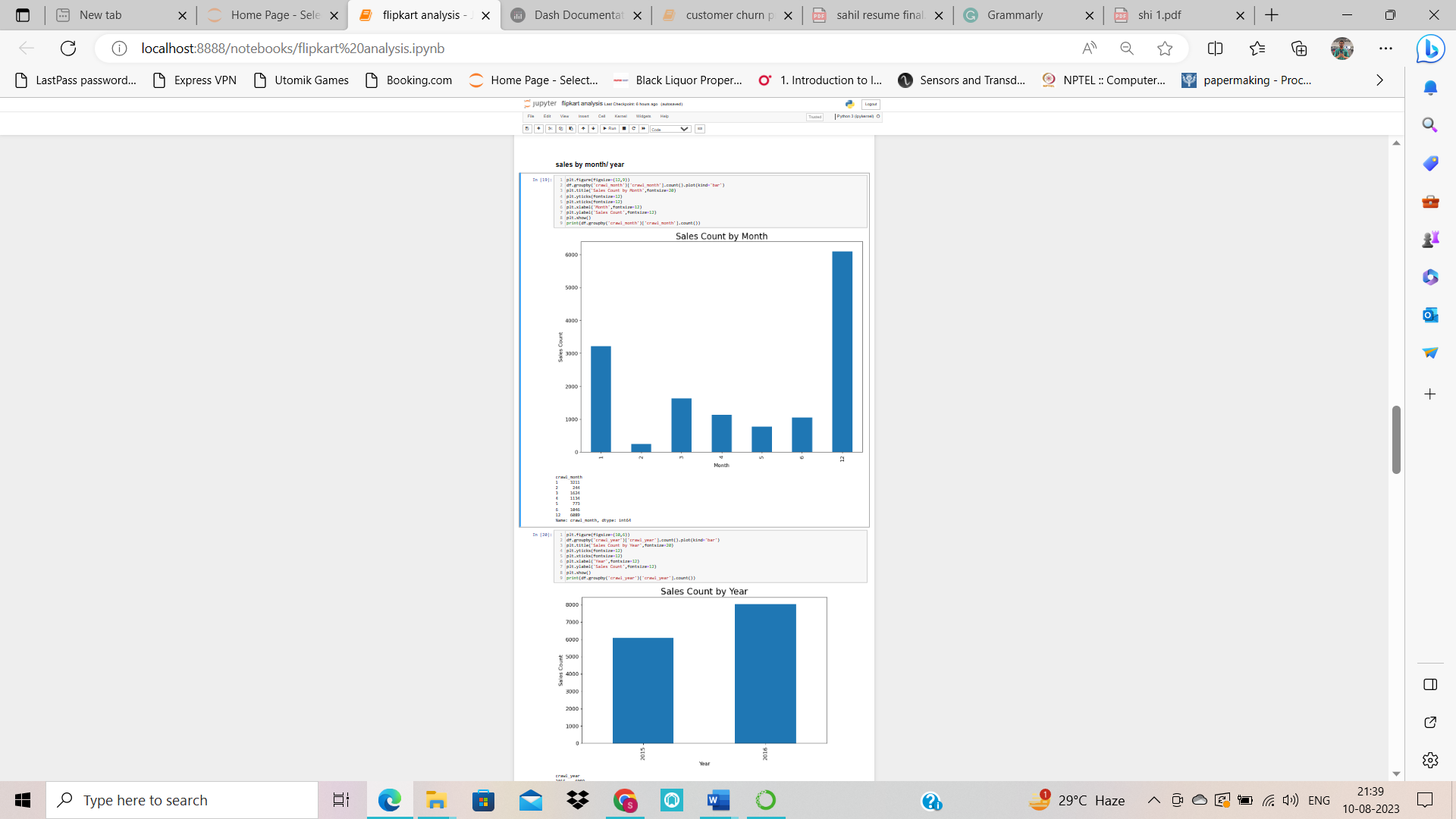
While tackling this issue, the newly generated column ‘discount\_percentage’ will be useful.



The bar graph has returned a very insightful result into the brands offering the highest discounts. Some of these brands are 'Rajcrafts', 'Bling', 'Fash Blush' and more.

We have successfully identified the target brands.

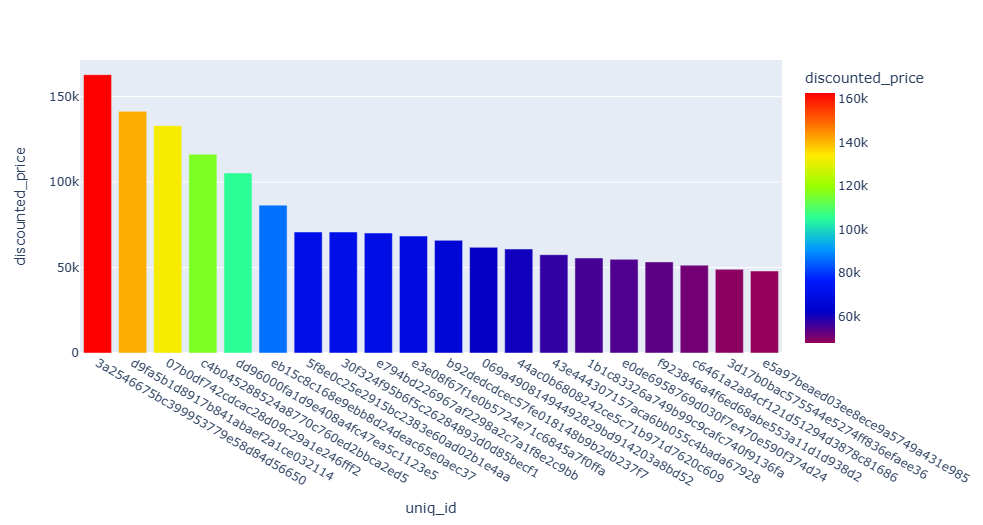
**Sales by month and year**

****

* **What are some of the target customers that spend the most money on Flipkart?**

It is essential to take into account the platform's most engaged users. The clients who have spent the most money on the site are presummably those who purchase at Flipkart the most. Our target market is comprised of these prospective clients.

Our dataset's column ‘uniq\_id’ contains the ‘unique ID’ that the system gives to each individual website user. These distinctive IDs may be used, and we can treat them as distinctive consumers. We may get a list of prospective clients by categorizing these consumers according to the total amount of ‘discount\_price’ they paid.



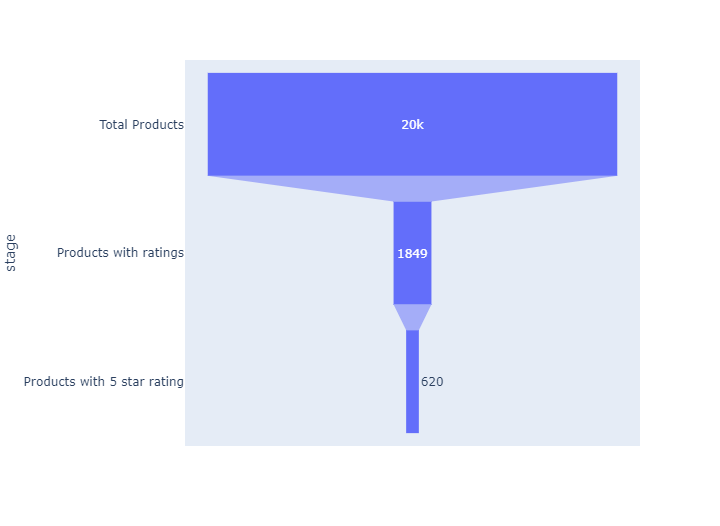
The top 20 clients who spend the most money on Flipkart have been identified. Our emphasis is on the Flipkart accounts connected to these particular IDs.

It's critical to comprehend the benefit of the star rating system. People may rate a product or brand using a number of stars by answering a rating question with a star rating. They establish standards that make it simple for shoppers to compare various goods and brands.

The ranking system may be used to separate the best brands and items from the others, which is a wonderful advantage. It should go without saying that consumers are more likely to advocate and use companies and items with five stars since they are the most dependable.

It's important to remember that not every product has a rating score. A relatively tiny fraction of the population is where the information about the items with various ratings comes from. Let's first comprehend how many of the total items have ratings, and how many of those things have ratings of 5 stars.

* **What is the proportion of 5 star rated products to the total products?**

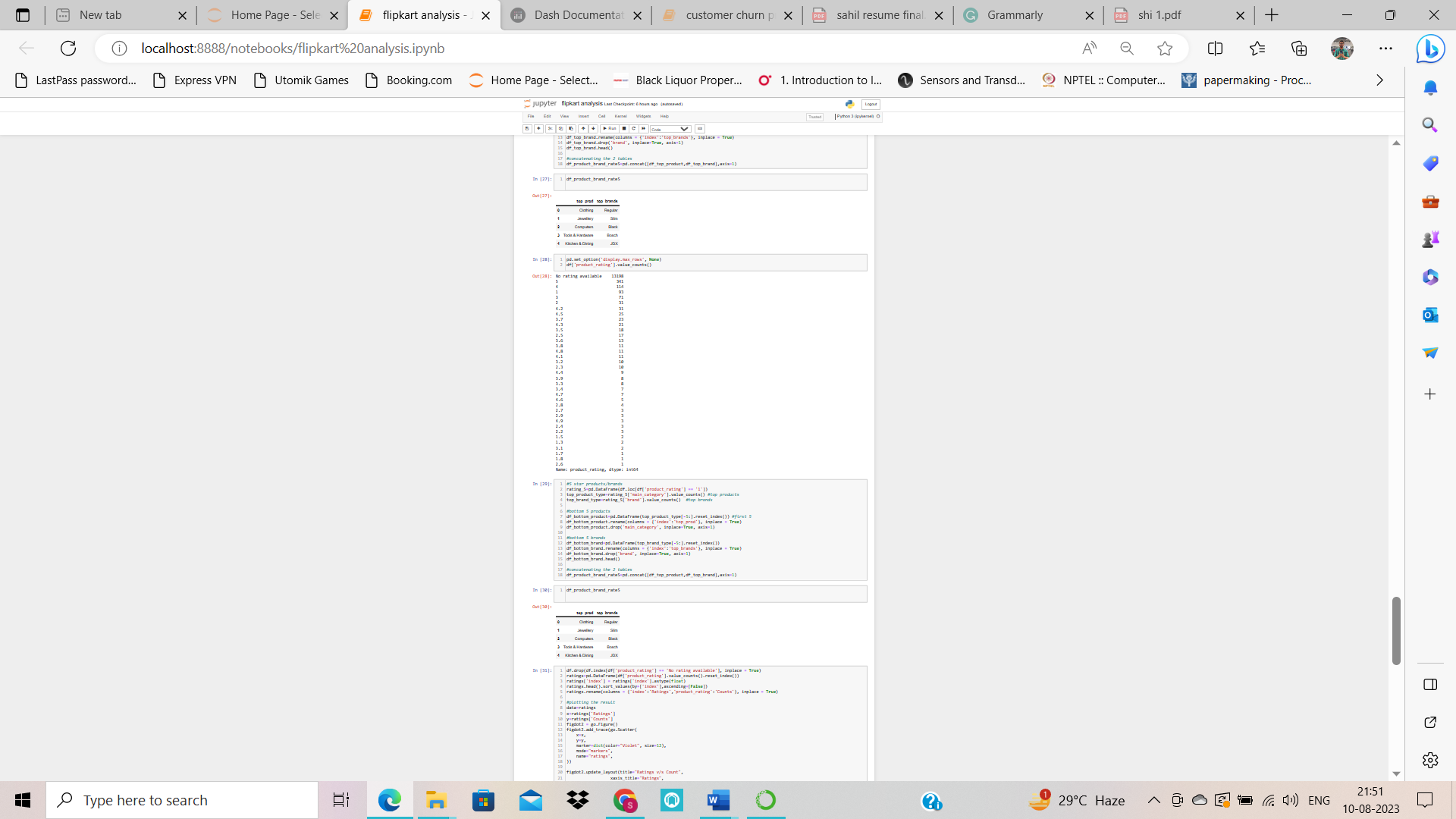


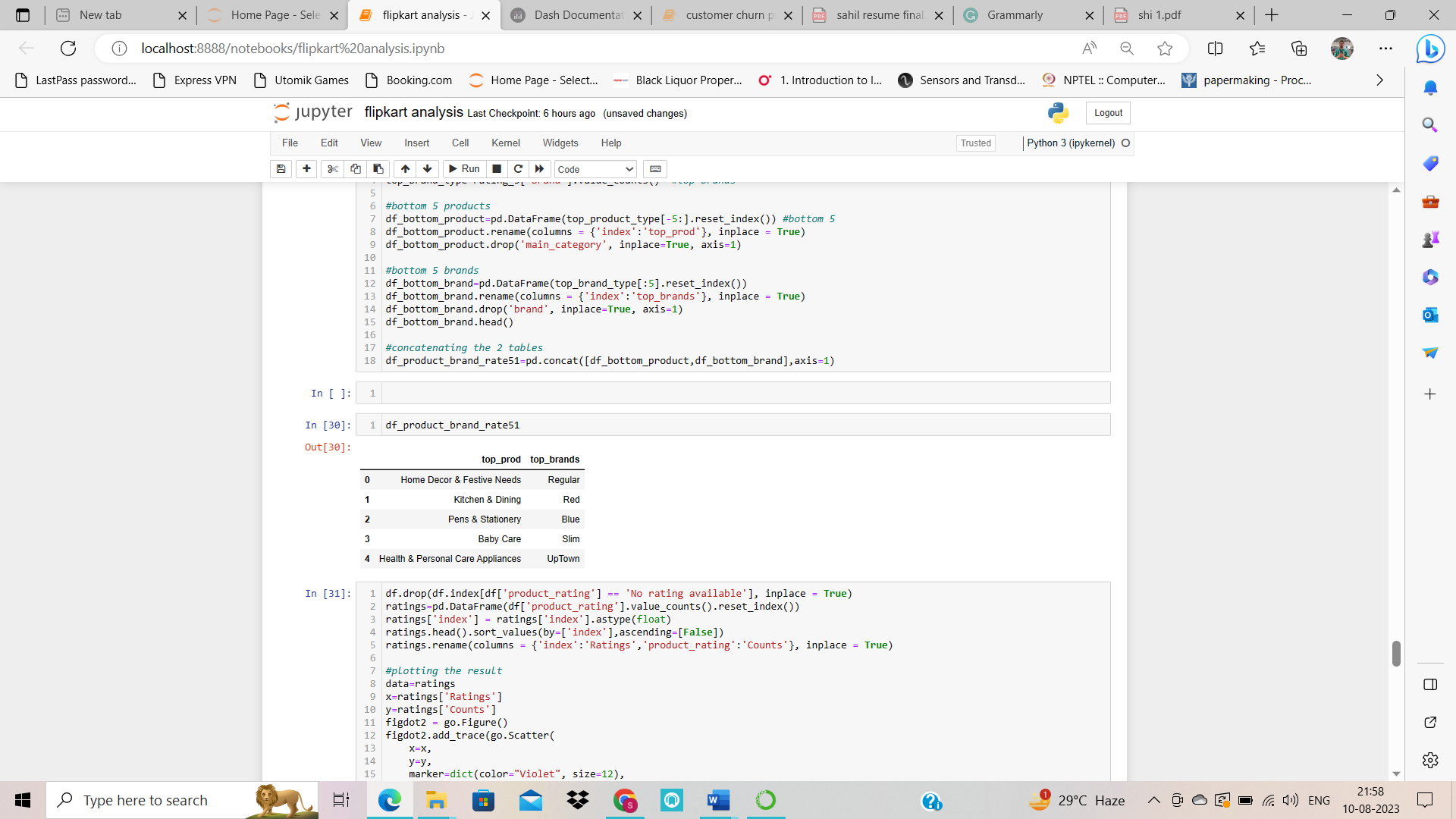
A funnel chart is what is seen in the above graph. Data in various phases of a business process is often represented using funnel charts.

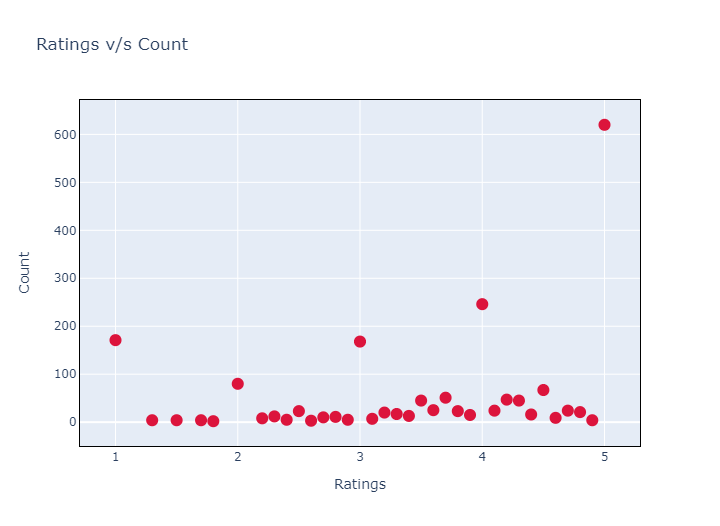
It is quite simple for us to see the various levels, beginning with the total items and finishing with the 5-star products, thanks to the above graph. Only 1849 of the 20,000 goods have ratings, and only 620 of them have five-star reviews. This represents a minuscule part of the whole population.

* **What are the best and worst-performing products and brands in terms of ratings?**

For the best-performing, we will look at the 5-star items and brands, while for the poorest, we will look at the 1-star products and brands. The 1-star items and brands should also be considered since they tend to have issues. To satisfy client needs and raise the caliber of the commodities on the internet, these products/brands may be further enhanced.



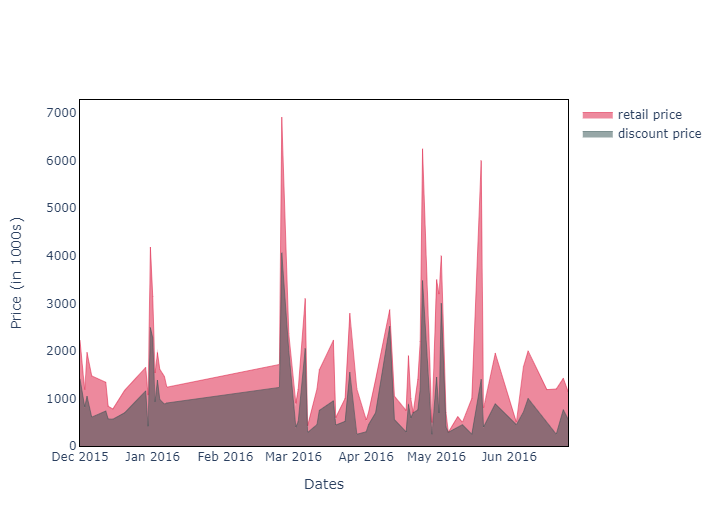


The above is called a Dot Chart. One of the simplest ways to visualize data.

We observe that 5-star products have the highest count, followed by 4-star, 1 star, 3-star, and 2-star. We can look at the count by placing our cursor on different dots.

* **Is there any trend associated with the retail price and discount price over the months?**

Understanding when the prices have surged up or down during the year is one reason to examine the pattern of retail and discount pricing. This study provides us with some useful knowledge about how the various dataset characteristics operate. With the use of this information, we can decide whether to increase or drop product pricing. These choices shouldn't always be based on intuition.



The above plot is an example of an Area Chart. We observe that in the months of December, February, April, and May the prices had spiked up, which can be due to various reasons.

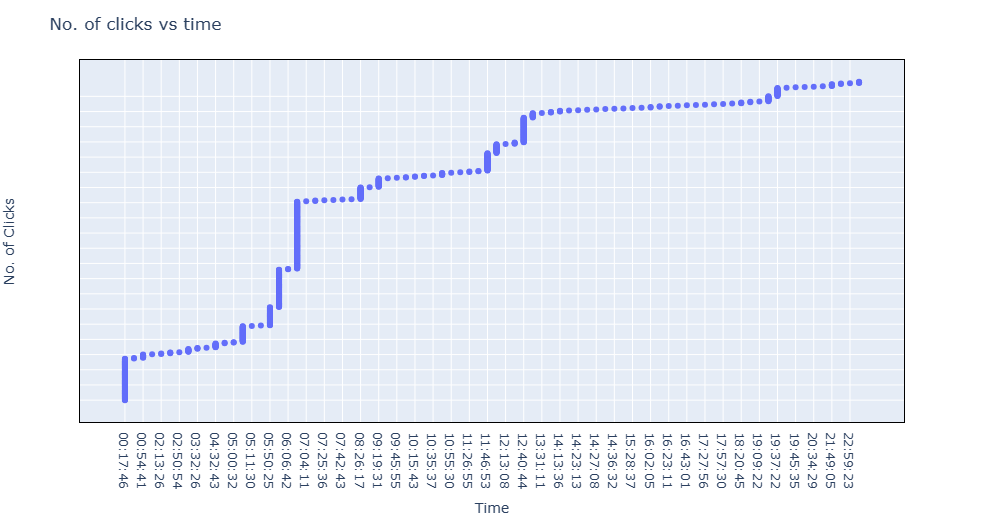
* **When are customers the most active during the day?**

Let's find out when clients are the most active throughout the day.

Using the field product\_id, we can link consumer behavior to the various product id s that were clicked. This will show us the times of day when various URLs are being clicked.

Understanding when rush hours are more prevalent will help us optimize the efficiency of the Flipkart platform during these times, which is one of the reasons why it is crucial to examine this information. Moreover, during these busy hours, advertisements for other products may also run.

For the analysis, we'll utilize the date column with the product\_id. We've created a scatter plot in the example below to help you see the outcome and draw important conclusions from it.



We see that user activity keeps increasing throughout the day

**Conclusion:**

Hence, we conclude that:

* The top product customers are purchasing is 'Clothing', whereas the top brand is 'Regular'.
* Brands giving the highest discounts are 'Rajcrafts', 'Bling', 'Fash Blush', and more.
* Learned how to target active customers.
* Out of 20,000 products, only 620 products have 5-star ratings.
* The product 'Baby Care' and the brand 'Trident' are the lowest-performing product and brands in the 1-star category.
* In the months of December, February, April, and May, the prices spiked up.

**Thank you!!!**