Introduction

The Internet has revolutionized the way we buy products. In the retail e-commerce world of online marketplace, where experiencing products are not feasible. Also, in today’s retail marketing world, there are so many new products are emerging every day. Therefore, customers need to rely largely on product reviews to make up their minds for better decision making on purchase. However, searching and comparing text reviews can be frustrating for users. Hence we need better numerical ratings system based on the reviews which will make customers purchase decision with ease.

During their decision making process, consumers want to find useful reviews as quickly as possible using rating system. Therefore, models able to predict the user rating from the text review are critically important. Getting an overall sense of a textual review could in turn improve consumer experience. Also, it can help businesses to increase sales, and improve the product by understanding customer’s needs.

The Electronics & ModCloth dataset for electronics products were considered & cloth products were consisdered. The reviews and ratings given by the user to different products as well as reviews about user’s experience with the product(s) were also considered.

# Problem Statement

The goal is to develop a model to predict user rating, usefulness of review and **recommend** most similar items to users based on collaborative filtering.

## **Code and Resources Used :**

Python Version : '1.0.1'

Libraries Used : pandas, numpy, seaborn, plotly, datetime, matplotlib

## **The project consists of two parts :**

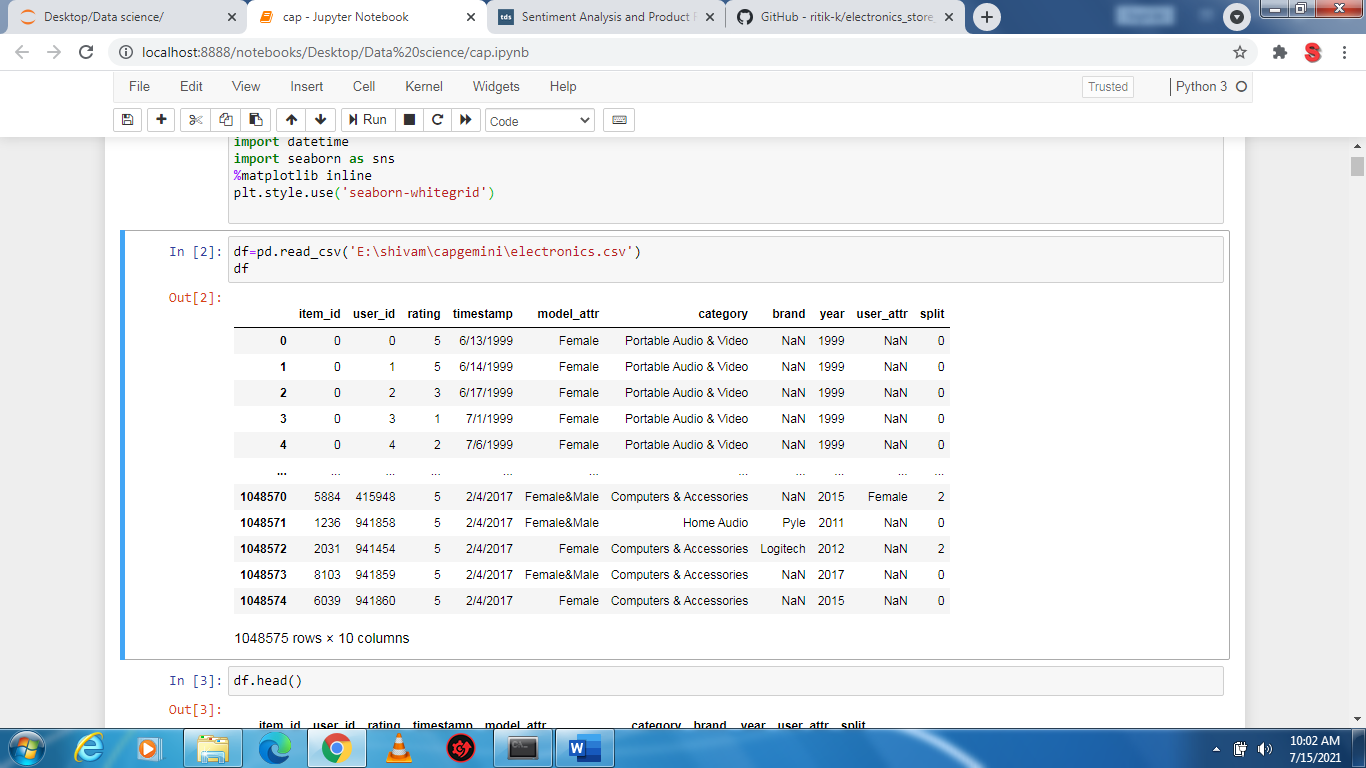
* Data Cleaning
* Data Analysis

# Data Collection

The electronics dataset consists of reviews and product information. This dataset includes reviews (ratings, item\_id,user\_id) and product metadata

(category,timestamp,model\_attr,…….).

**Dataset view: -**



Each row corresponds to product and includes the following variables:

Item\_id:- ID of item

User\_id:- ID of user

Rating:- rating of product

Model\_attr:- it contain the model gender

Category:- contain the product category

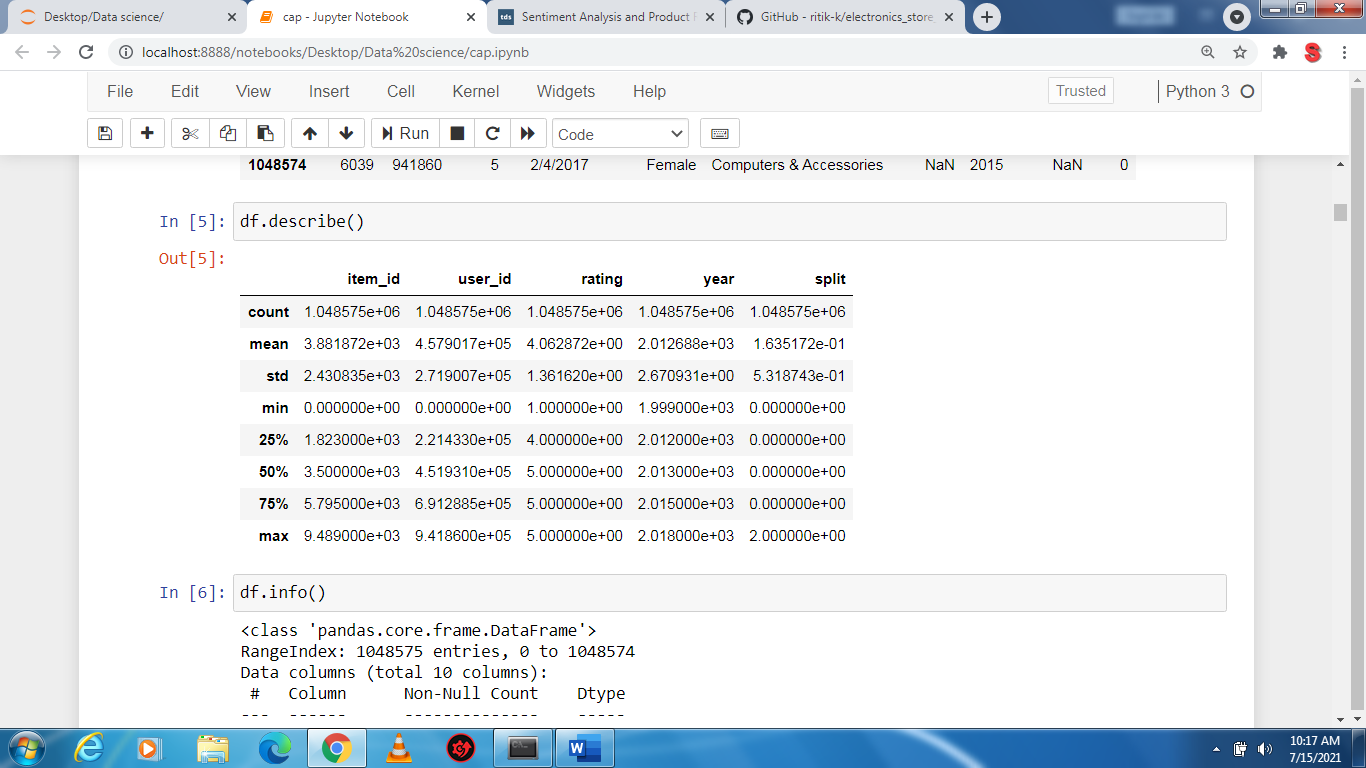
Brand:- name of brand

Year: - contain the year

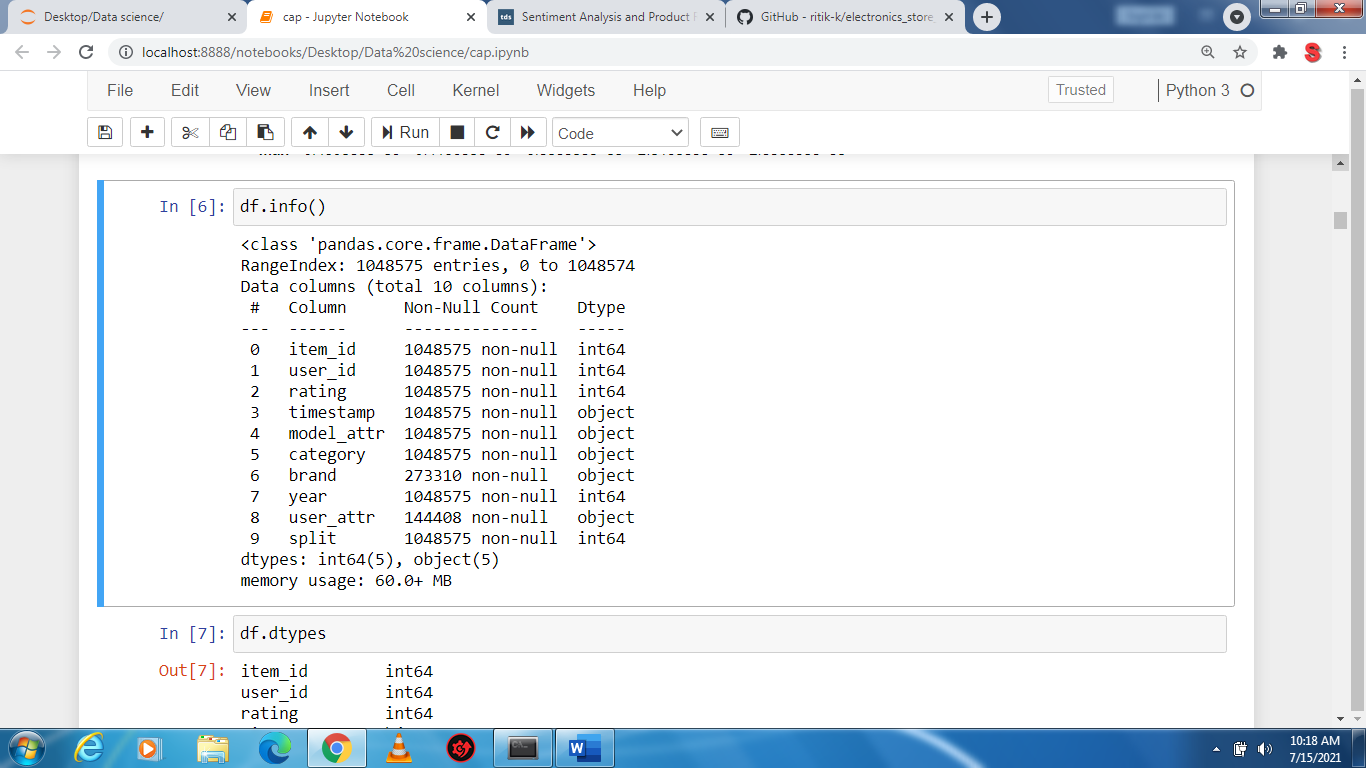
User\_attr:- contain the user gender

## **Descriptive Statistics**

The following summary statistics was obtained:-



**Data info:-**



NULL Values:-

item\_id 0

user\_id 0

rating 0

timestamp 0

model\_attr 0

category 0

brand 775265

year 0

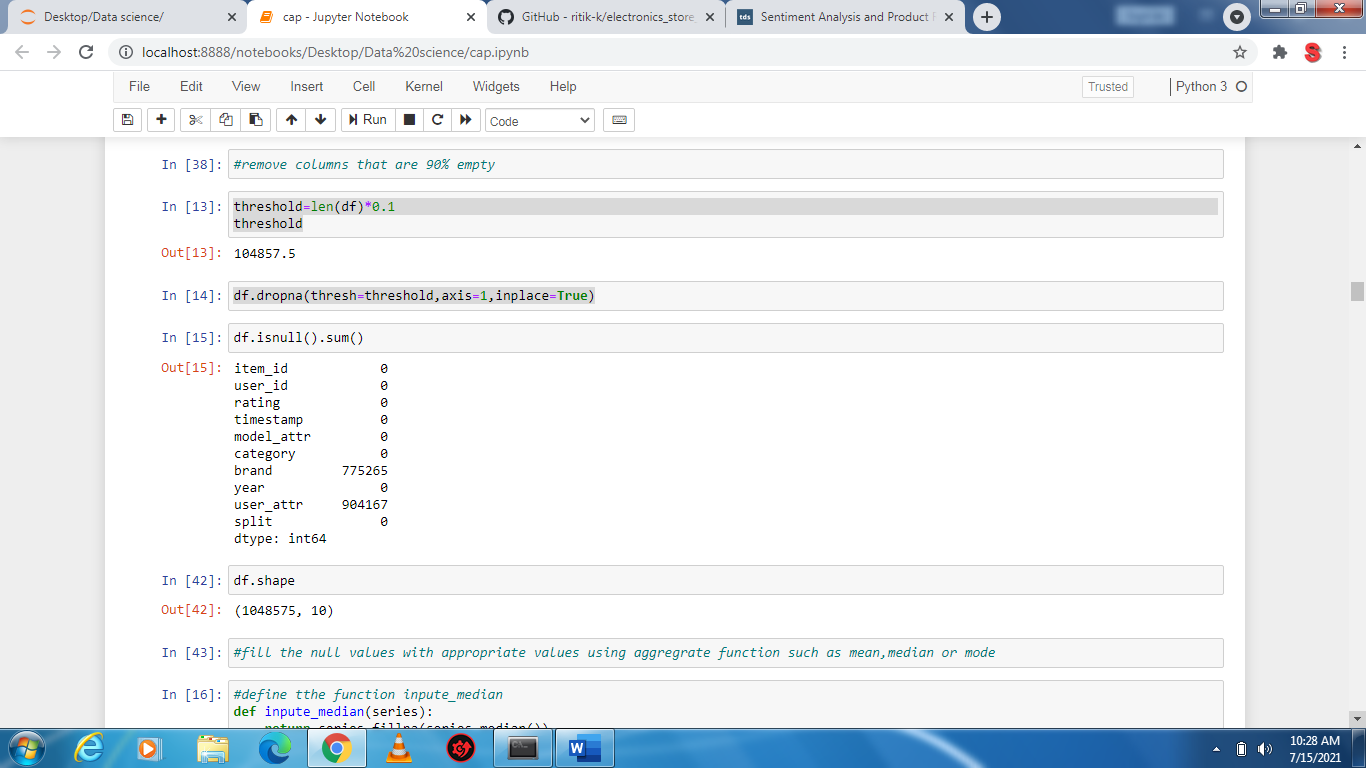
user\_attr 904167

split 0

dtype: int64

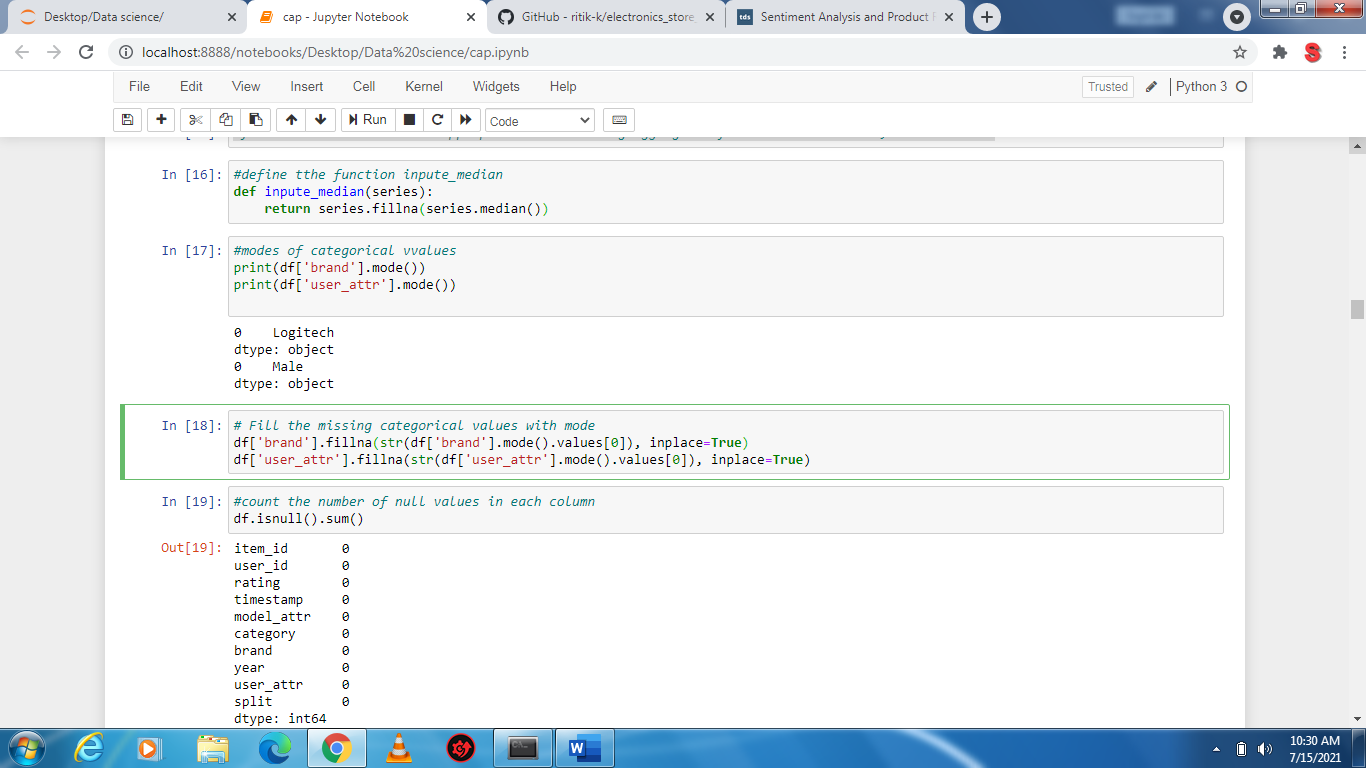
**Data Cleaning: -**

Remove the column that are 90% empty:-



There are no column that are 90% empty.

Now, fill the null values with appropriate values using aggregate function such as mean, median or mode

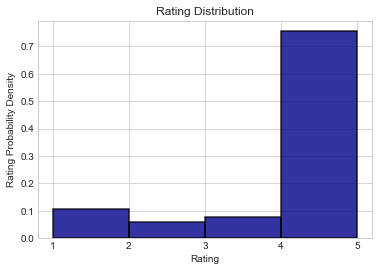


Now the dataset contain zero(0) null values.

## **Data Analysis :**

### *Here we explore 4 high level business questions related to our data*.

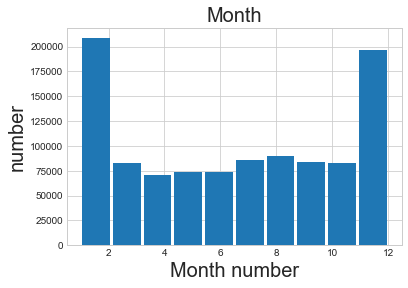
### **Whether the rating are Positive or Negative?**



**The histogram is clearly skewed to the left. The majority of reviews are *positive.***

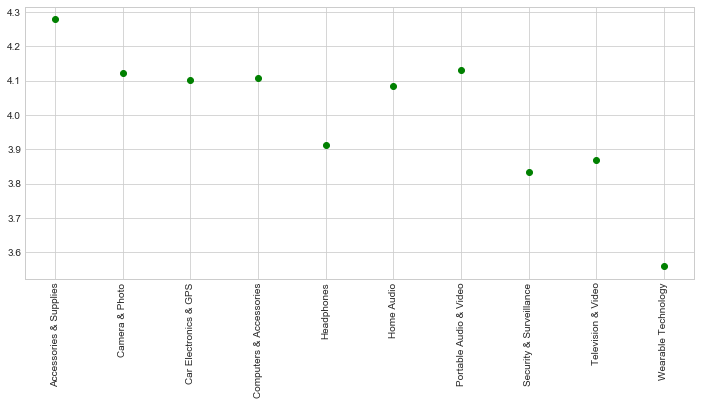
***This show that the happier customer is more than the sad customer. And we have to improve the quality of our products, this also good for business growth as well as image.***

### **What were the best months for sales? How much was earned in those month?**

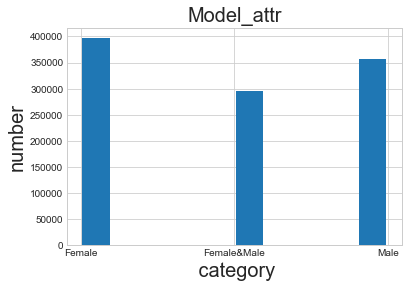


### **We can see that March & April is the worst month for Sales and January is the best month and this month contain New year sale also. We also see that sales keep increasing till April after which they start decreasing again; this is interesting and further analysis can be done on this, we can try to a correlation to the marketing and advertisement budget. It is seen that the months of December report good sales and it is expected as this months contain most of the holiday such as Christmas. People usually buy a lot of gifts during these holidays.**

### **Which city had the best category of products?**

**Here we see that Accessories & supplies are sold more frequently as compared to other expensive products.. So, we have to more focus on these products. And the wearable technology are less sold that’s why we have to focus on the quality of these product to increases the growth of these technology business .**

### **Which model\_attr are most used? Why?**

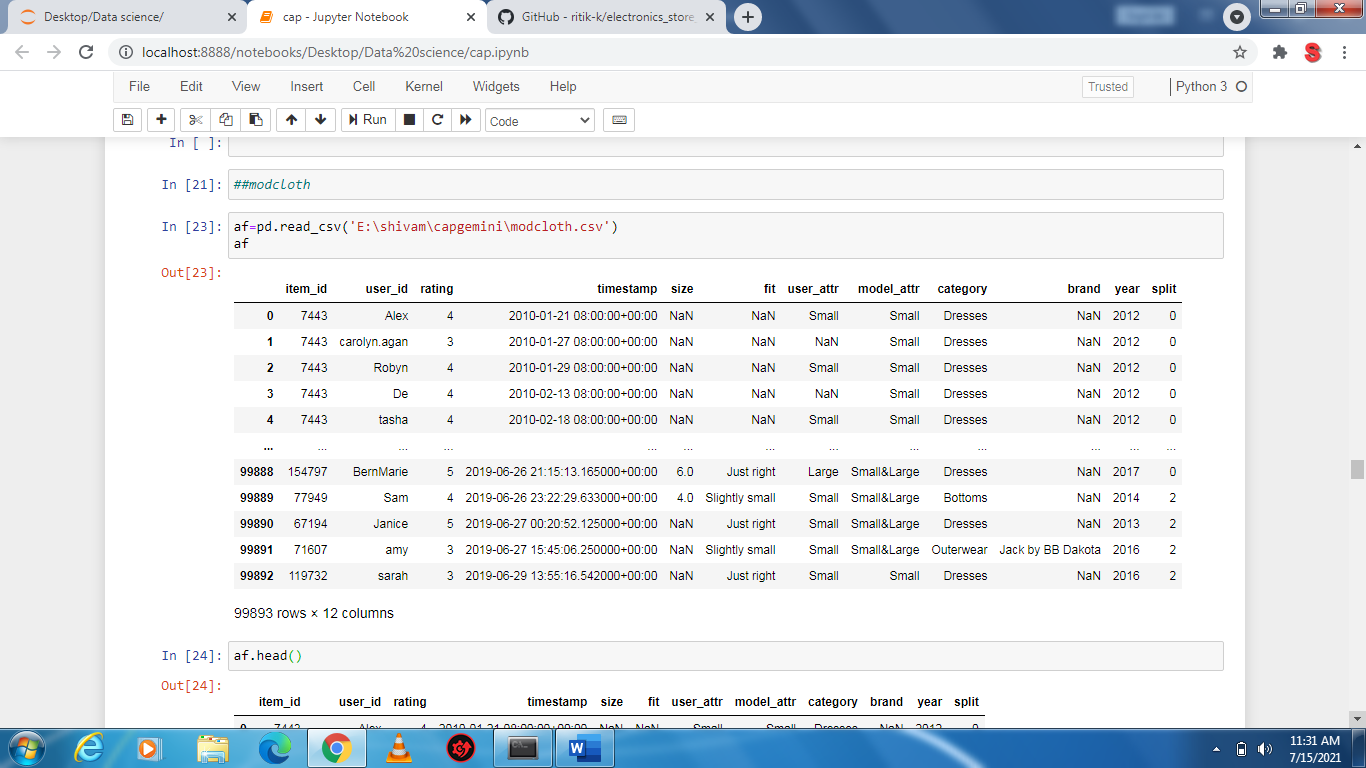


**Here we see that the number of Female model are most as compared with other category.**

**This is because more the number of Female users in the dataset that’s why we have to more focus on female model as compared with male model for the growth of business.**

**ModCloth**

**Dataset view: -**



Each row corresponds to product and includes the following variables:

Item\_id:- contains the ID of item

User\_id:- contains the ID of user

Rating: - rating of product

Timestamp: - contains the time and date

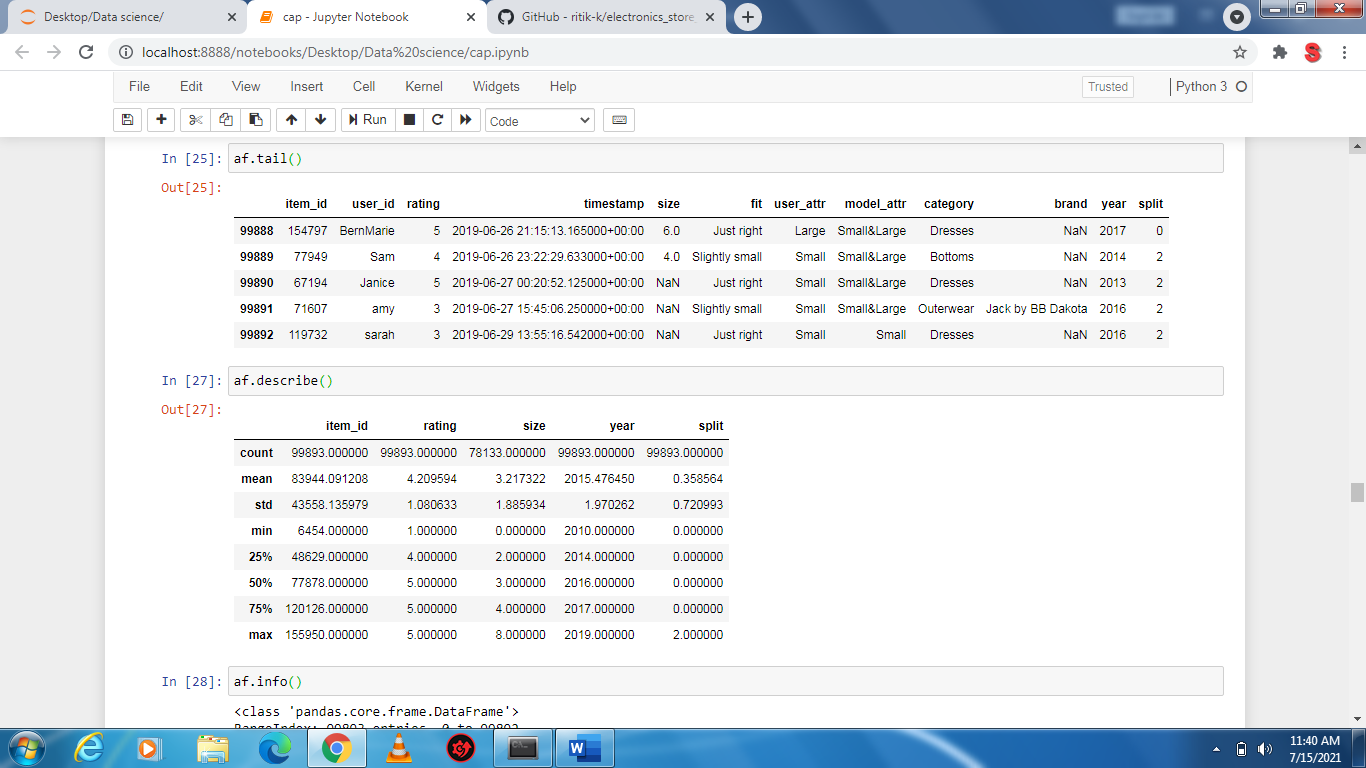
Model\_attr:- contains the model gender

User\_attr:- contains the user gender

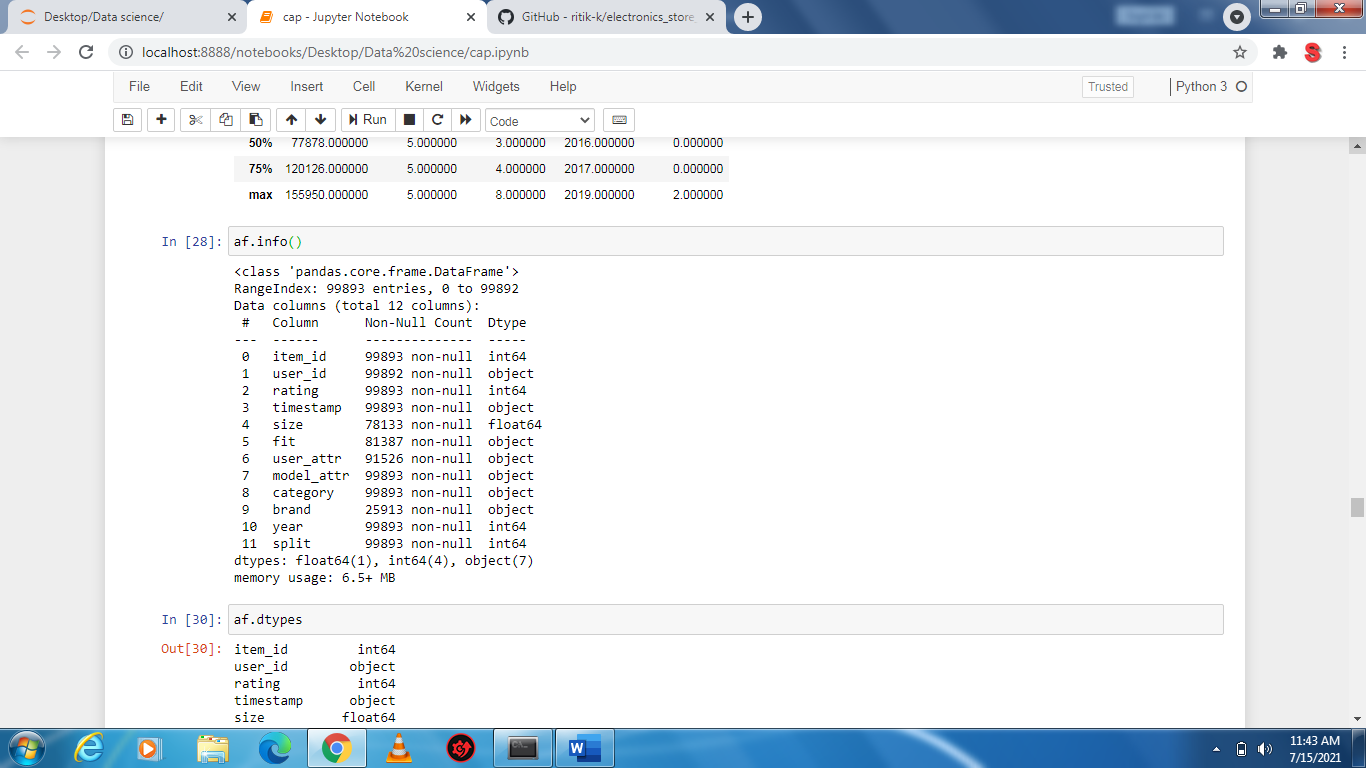
Year: - contains the year

## **Descriptive Statistics**

The following summary statistics was obtained:-



**Data info:-**



NULL Values:-

item\_id 0

user\_id 0

rating 0

timestamp 0

model\_attr 0

category 0

brand 775265

year 0

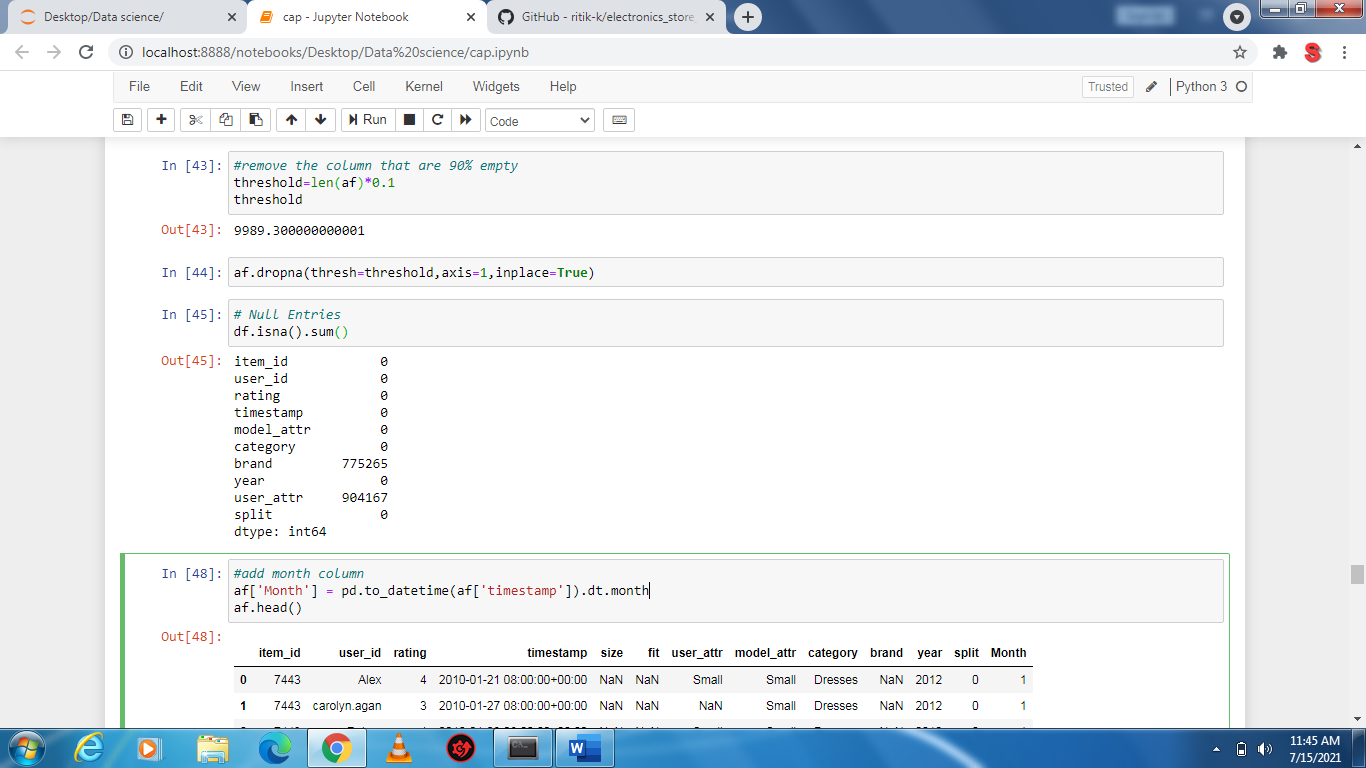
user\_attr 904167

split 0

dtype: int64

**Data Cleaning: -**

Remove the column that are 90% empty:-

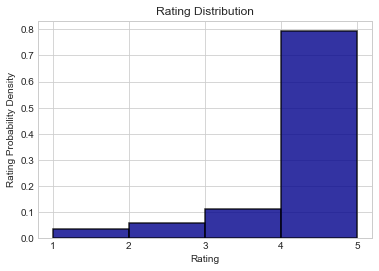


There are no column that are 90% empty.

## **Data Analysis :**

### *Here we explore 6 high level business questions related to our data*.

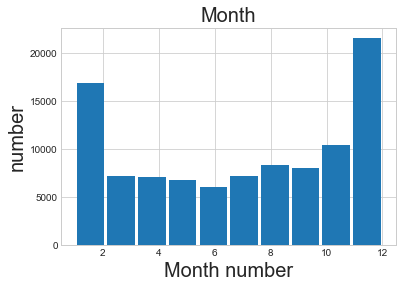
### **Whether the rating are Positive or Negative?**



**The histogram is clearly skewed to the left. The majority of reviews are *positive.***

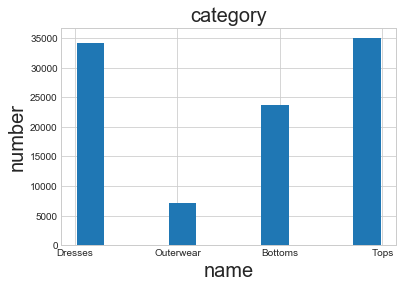
***This show that the happier customer is more than the sad customer. And this also good for business growth.***

### **What were the best months for sales? How much was earned in those month?**



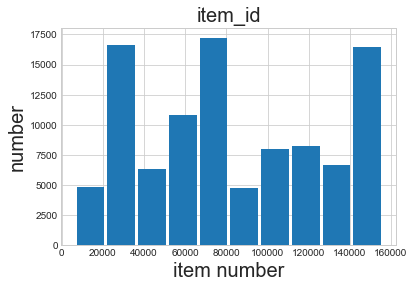
### **We can see that June is the worst month for Sales and December is the best month and this month contain Christmas sale also. We also see that sales keep increasing till April after which they start decreasing again; this is interesting and further analysis can be done on this, we can try to a correlation to the marketing and advertisement budget. It is seen that the months of January report good sales and it is expected as this months contain most of the holiday such as New Year. People usually buy a lot of gifts during these holidays.**

### **What category are most often sold together?**



**As we see that Tops has the category that are most often sold and the outwear lies on the category that are less sold. The above analysis show that we have to more focus on Tops(high quality) as well as outerwear(low quality) for the growth of Business.**

### **What item\_id sold the most product?**



**Here we see that the item\_id that are lies between the 6k to 8k are mostly sold due to high quality whereas till 2k the item are less sold due to low quality and need replacements.**

### **What products are most often sold together?**

ModCloth 15071 times

Chi Chi London 2940 times

Retrolicious 1951 times

Steve Madden 1222 times

Out of Print 757 times

Ryu 599 times

Emily and Fin 445 times

Banned 408 times

Kin Ship 377 times

Jack by BB Dakota 355 times

Miss Candyfloss 343 times

Collectif 286 times

Hell Bunny 251 times

Pepaloves 208 times

Wrangler 167 times

Pink Martini 143 times

Compania Fantastica 87 times

Sugarhill Boutique 65 times

Daisey Natives 58 times

Miss Patina 49 times

Alice's Pig 25 times

Effie's Heart 17 times

Yumi 16 times

BB Dakota 16 times

Closet London 12 times

Mata Traders 11 times

Wendy Bird 9 times

Eliza J 8 times

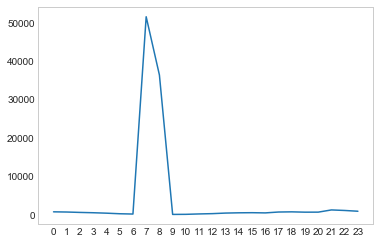
Louche 7 times

Blue Platypus 6 times

Rolla's 4 times

### ***This data is very useful as using this we can create offers or promotions which can increase our sales further.*** ***When a person buys a products at a discounted price which will push more customers to buy these products.***

### **What time should we display advertisements to maximize the likelihood of customers buying the product? ( Best Time for Sales )?**



### **Here we can see that 6PM and 9PM are the two hours during which we have the highest sales. This makes sense because these are the hours during which people are generally going to or returning from work. Thus if we ran an advertisement during these 2 hours, it will get maximum visibility and in turn increase the chance of customers buying the product.**

#### **This project is best viewed in a notebook viewer. In this notebook, you will find a walk through of the work done, detailed analysis and the respective code.**

#### https://drive.google.com/file/d/1m-OQ7f3vNyvinJt5qyvYkmNPCIwnUWqi/view?usp=sharing