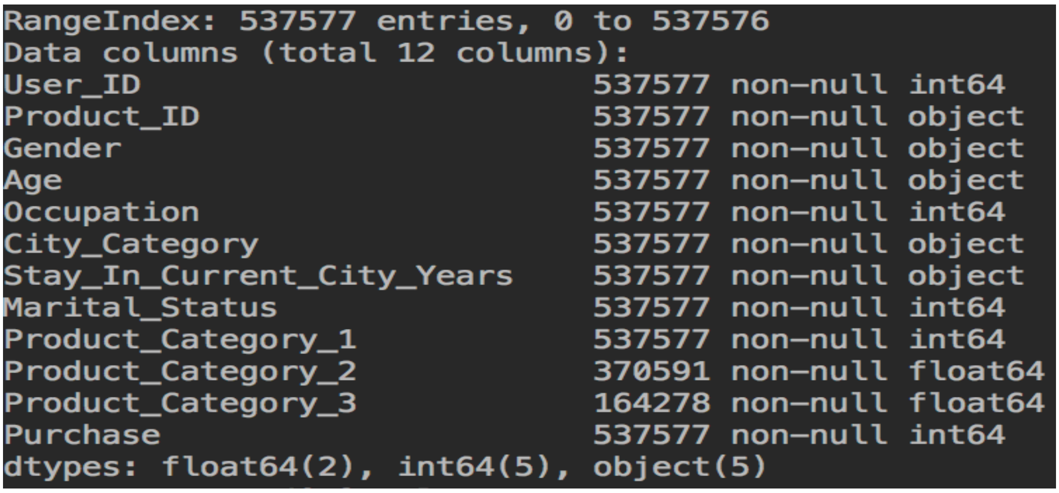
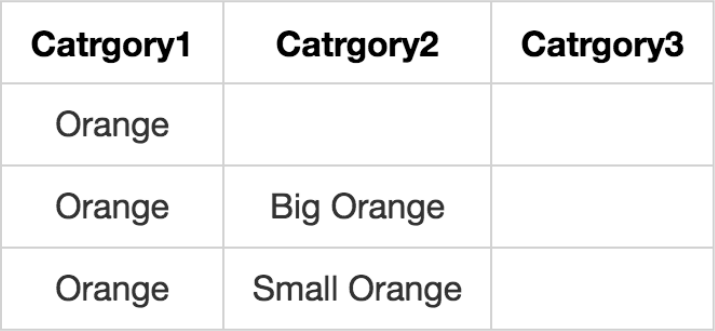
1: Data Reading

The dataset we want to analyze is BlackFriday(from kaggle) As we print the info of the dataset, we find that there are no null values in the data set except Product_Category2 and Product_Category3.



```
RangeIndex: 537577 entries, 0 to 537576
Data columns (total 12 columns):
User_ID                          537577 non-null int64
Product_ID                       537577 non-null object
Gender                           537577 non-null object
Age                              537577 non-null object
Occupation                       537577 non-null int64
City_Category                    537577 non-null object
Stay_In_Current_City_Years       537577 non-null object
Marital_Status                   537577 non-null int64
Product_Category_1               537577 non-null int64
Product_Category_2               370591 non-null float64
Product_Category_3               164278 non-null float64
Purchase                         537577 non-null int64
dtypes: float64(2), int64(5), object(5)
```

When we want to fill the null values of these two columns, it comes to us that, Product2 may be the subset of Product1 and Product3 may be the subset of Product2. Just like we have three kinds of oranges, big; normal and small, we know that big and small are subsets of orange, but normal orange is the subset of itself, so we do not have to fill Product_Category2. Maybe the picture below will give us a clearer understanding



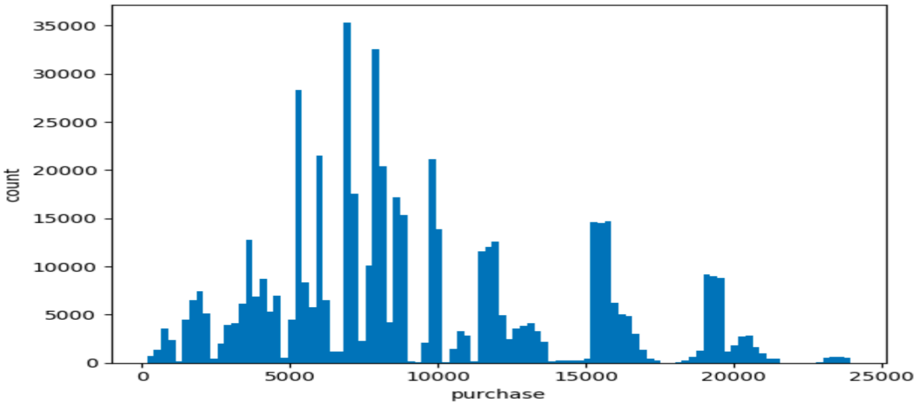
Catrgory1	Catrgory2	Catrgory3
Orange		
Orange	Big Orange	
Orange	Small Orange	

(We should know normal orange is orange)

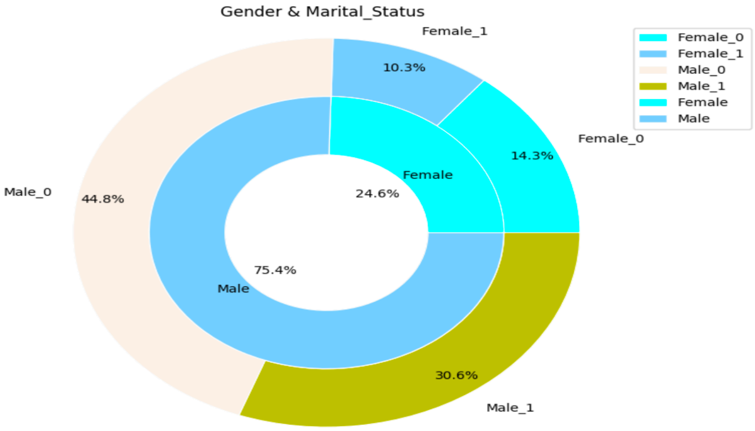
So we decide to delete these two columns for the reason that they are of little use. (Or fill with 0)

2: Data Exploring & Analyze

After data processing, we found that the user's consumption is generally between 0-15000

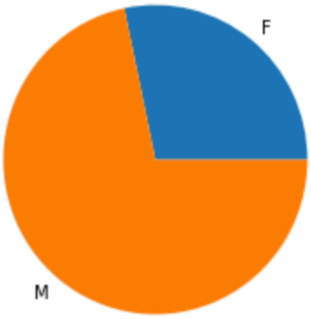


And we want to know gender and marriage impact on purchases:

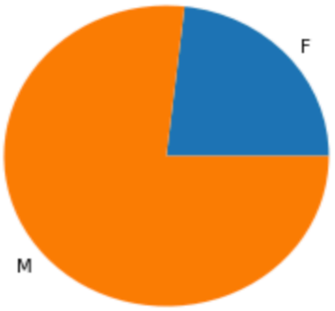


We can know that either Male or Female, unmarried people buy more than married people, but we don’ t know whether men consume more than women. So we conducted a comparison of the number and consumption of men and women. Results are as follows

The ratio of female to male is 1:2.5.



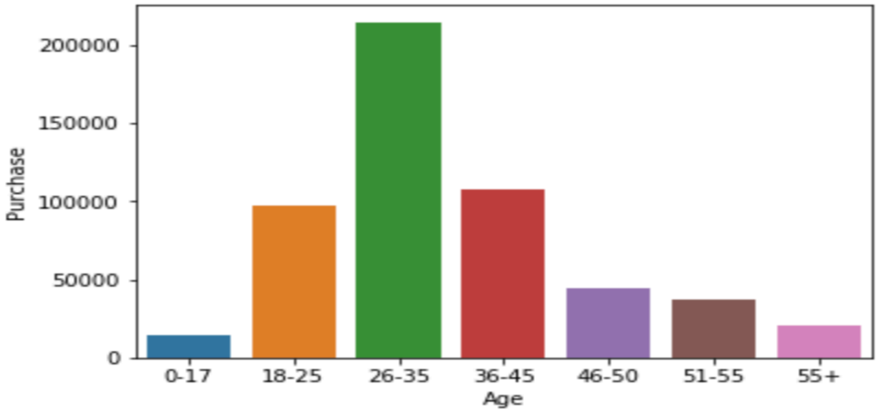
And the ratio of consumption is 1:3.3.



So we can conclude that Men’ s purchasing power is stronger than women’ s

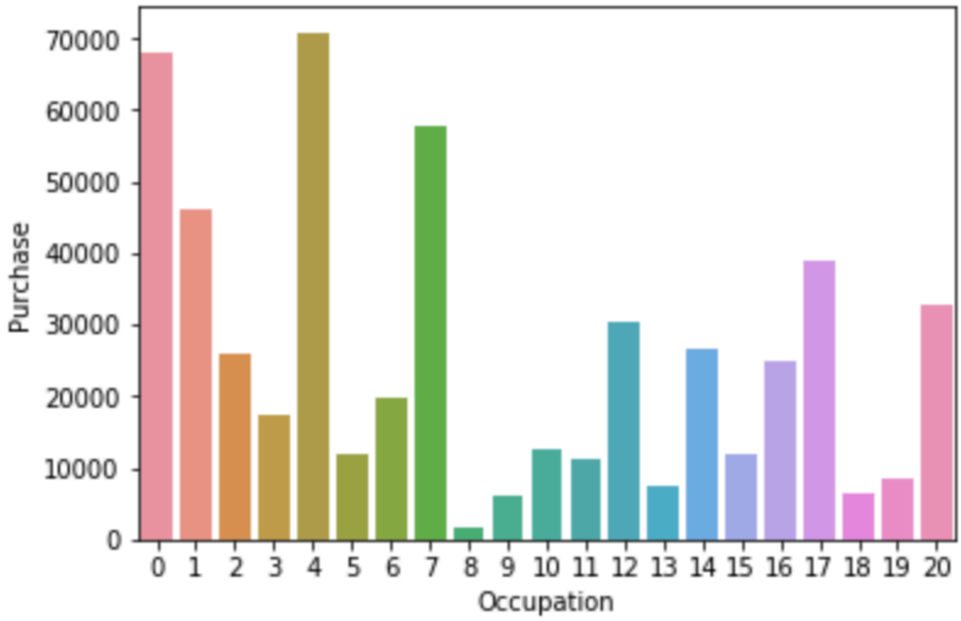
Then the Age:

When we compare the purchases of different ages, we find that the main group purchased is concentrated in the middle-aged group of 18-45 years old, showing a similar normal distribution pattern, among which the youth groups of 26-35 years old contribute the most.



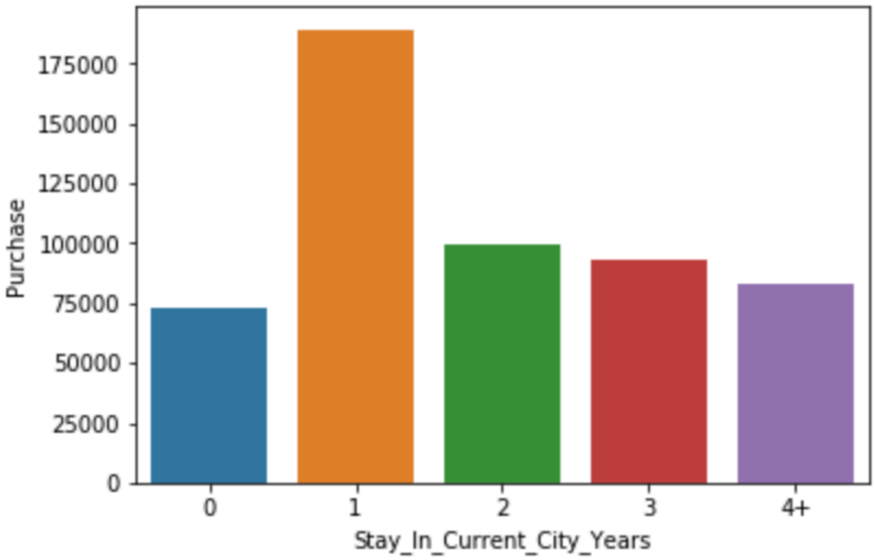
Occupation:

When we compare the purchases of different occupations, we find that the group with occupation code 0, 4, 7 has more purchases, and 8, 13, 18 are the least.



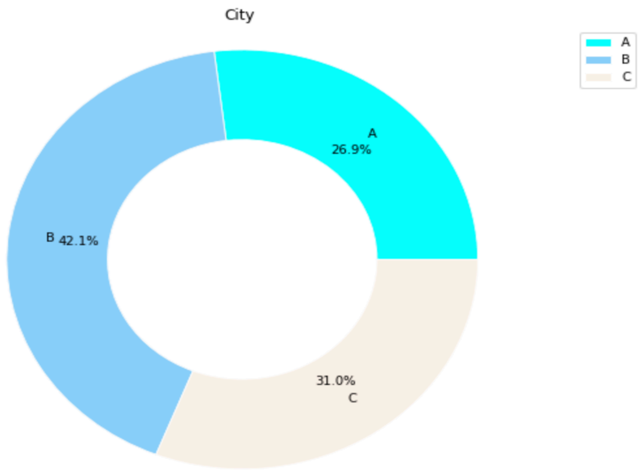
Stay_years:

When we compare the purchases of different residence times, we find that the group that lives for one year buys more, and the rest of the purchases are closer.



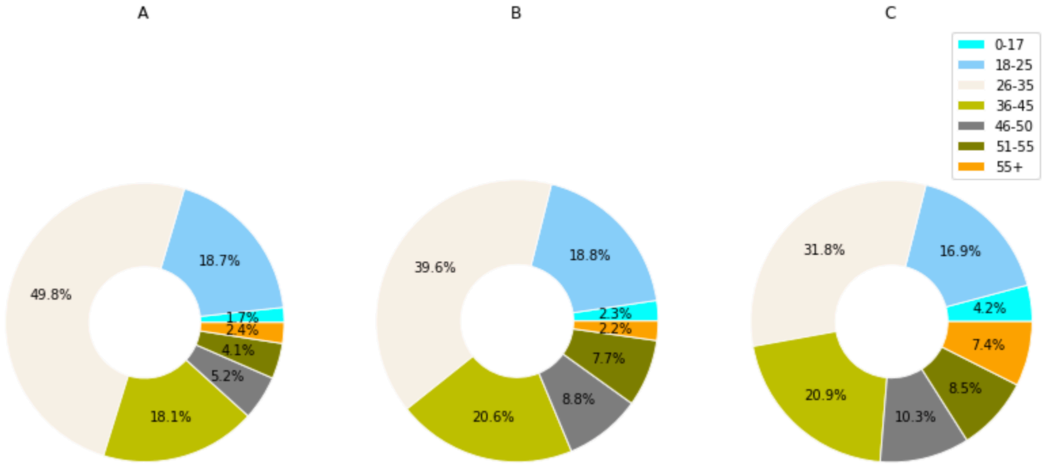
City

When we compare the purchases of different cities, we find that the number of groups purchased in B City is more, C City is second, and A city is the least.

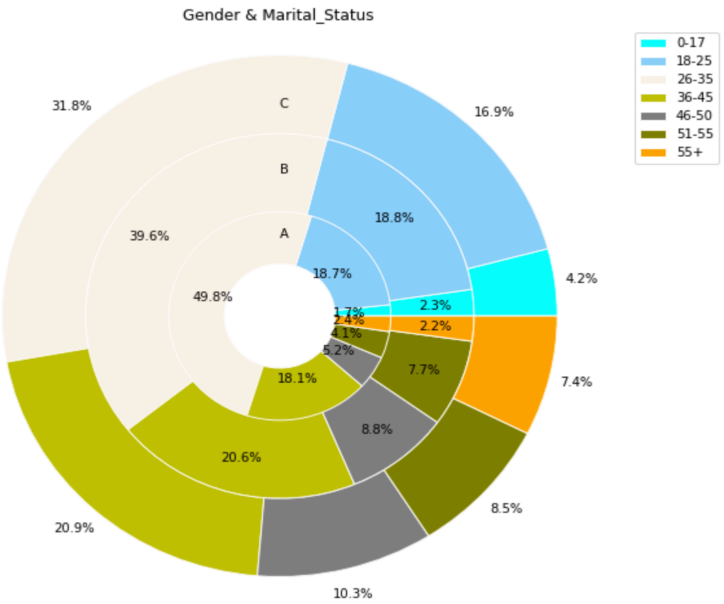


Population composition of each city：

When we compare the age groups of different cities, we find that A city is close to 50% is 26-35 group, 18-45 year old group is 86.6%, but A city's purchasing power is the lowest, indicating that A city is A very young city, the economic strength is not very strong, but it has great development and consumption potential; B City is a relatively robust city with strong purchasing power and young people occupying the majority of the population; C City is a city that is close to aging. The population over 46 years old accounts for 26.2. Considering that its 0-17 age group accounts for 4.2%, the highest among the three, the demand for baby products and juvenile books is higher, so the purchasing power ranks first. two.



We can also show it by Ring diagram


Gender & Marital_Status
C
B
A