Dash data visualization demo description document

Name	Student ID
Qiao Liang	1853572

Data Analysis task

Objectives

The dataset I selected for visualization is the **Black Friday dataset**. The purpose of this dataset is to find some usable information and patterns by analyzing the data in different dimensions, such as finding the shopping patterns, shopping tendencies, etc. on Black Friday. Here I mainly analyze the huge dataset from the user's perspective.

Characteristics of the dataset

First is the table header for this dataset:

No	Field Name	Data Type	Field Description
1	User_ID	Int	The ID of customer
2	Product_ID	String	The ID of the Product
3	Gender	String	Gender of the customer
4	Age	String	Age of the customer
5	Occupation	Int	ID of the customer's occupation
6	City_Category	String	Categories of cities
7	Stay_In_Current_City_Years	String	Number of years in current city
8	Marital_Stayus	Int	Marital Status
9	Product_Category_1	Int	Product Category 1
10	Product_Category_2	String	Product Category 2
11	Product_Category_3	String	Product Category 3
12	Purchase	Int	Amount spent (in USD)

We can also see that there are a total of 3 categories of products, and through some statistical calculations can also be concluded that there are a total of 5891 customers (of which 4225 are men and 1666 are women, with a male to female ratio of 2.54:1), 3623 kinds of products. Product categories 2 and 3 have a 31% and 69% missing rate respectively, which is very high, so we will not deal with them here (the analysis of product categories is not covered below) and will study them later when we have time. The average amount of each transaction is $9,333, equivalent to RMB 63,746. Among them, the lowest one transaction consumed $185 (¥ 1,263), and the highest one consumed 23,961, equivalent to RMB 163,658.

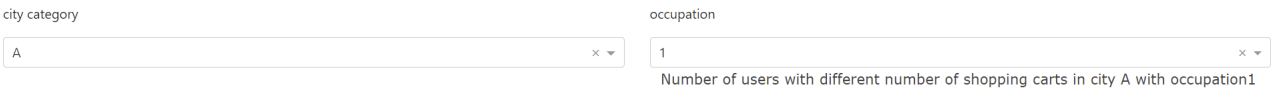
Due to the large data set, here I will analyze the data for customers' gender, marital status, city, occupation, sales and purchase amount.

Designed Layout

This dashboard uses a grid layout in general, starting with a conditional filter module, followed by two rows and two columns of data charts, and finally the corresponding original data set filtered according to the filtering criteria.

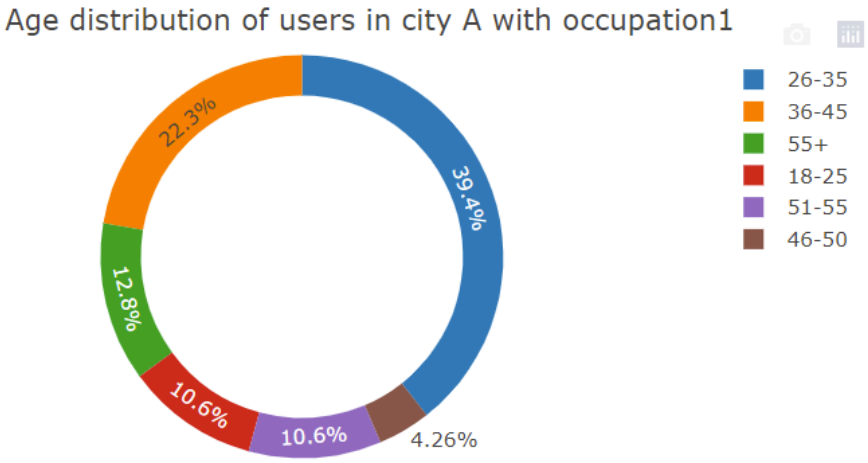
Design of data visualization

Since the focus of my analysis is on user profiling, due to the large amount of data, I selected City and Occupation as the main filtering criteria to filter the data before visualization in order to make the analysis more geographically and occupationally diverse. As shown below:



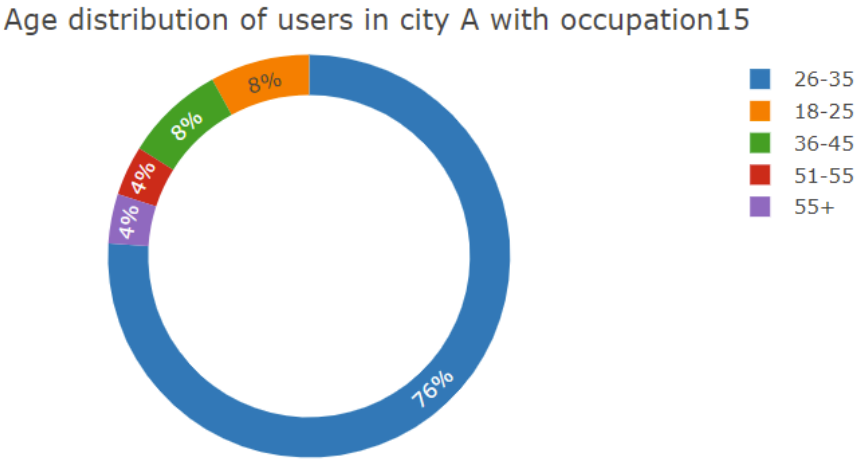
Pie chart of customer age structure by city and occupation

After completing the filtering of the data, I first analyzed the age distribution of the purchasers, using a pie chart to represent:



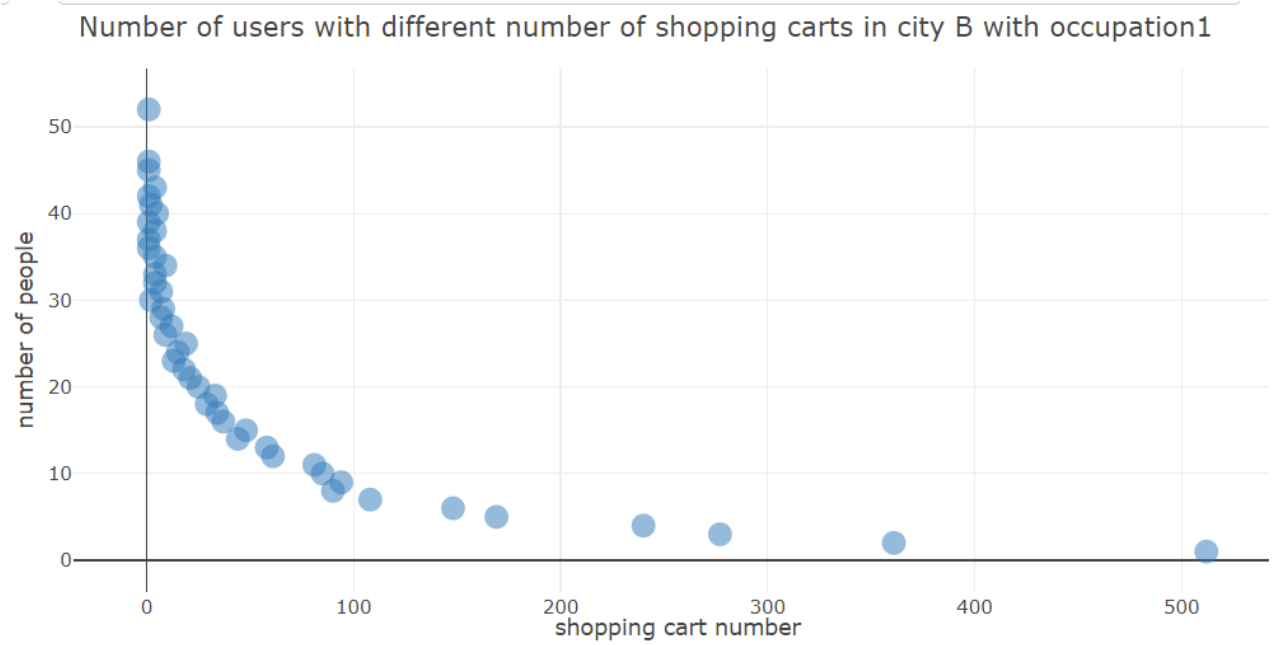
For example, for city A, with occupation 1, we found that the majority of customers are 26-35 years old, accounting for about 20% of the total number of customers.

But this result is closely related to the occupation. For example, for occupation 15, customers aged 26-35 make up the majority：



Scatter plot of the distribution of the number of goods purchased by different cities and occupations

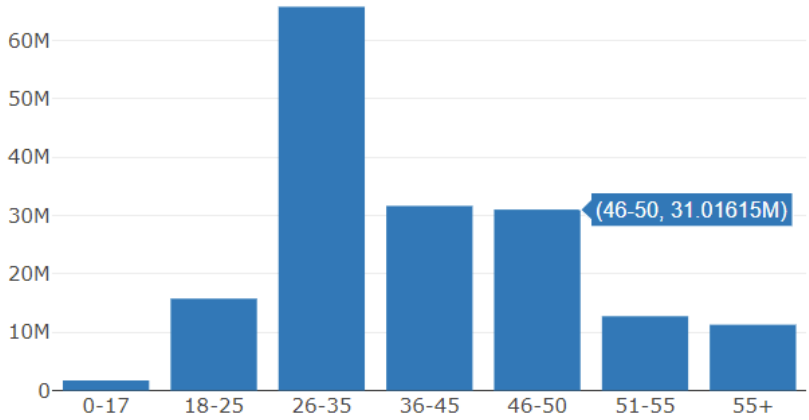
In response, I found an overall hyperbolic trend: the vast majority of shoppers purchased a number of items of 1, but a few purchased a number of items of 200 or more.



And the law of attribution is applicable to all professions.

Bar chart of purchasing power of different age groups

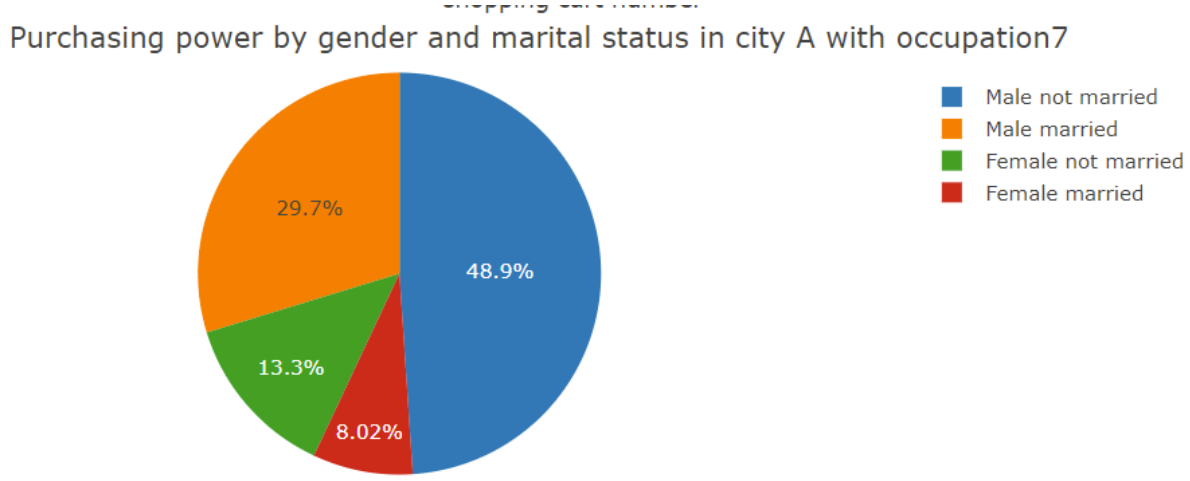
For the age structure, we need to analyze the actual purchasing power in addition to the number of customers of different age groups. Therefore I plotted the total amount of purchase spending of different age groups in different cities with different occupations:



We can find that the 26-35 age group has the greatest purchasing power, and they represent the younger generation.

Pie chart of purchasing power of customers by gender and marital status

Marital status and gender are also among the decisive factors influencing customers' purchasing power. Here we have analyzed the purchasing power of customers of different gender and marital status separately. The four types of customers are: male married, male unmarried, female married and female unmarried.



For example, for the customers of occupation 7 in city A as above, we found that overall men have stronger purchasing power than women, and unmarried men have the strongest desire to buy. **And such a pattern is independent of city and occupation.**

Source data table

Part of source data

User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
1000015	P00334242	M	26-35	7	A	1	0	1	8		19653
1000015	P00247542	M	26-35	7	A	1	0	8	16		5958
1000015	P00338442	M	26-35	7	A	1	0	1	16		11415
1000015	P00275142	M	26-35	7	A	1	0	5	8		5380
1000015	P00333042	M	26-35	7	A	1	0	5	8		3594
1000015	P00166242	M	26-35	7	A	1	0	8			4209
1000015	P00161942	M	26-35	7	A	1	0	5	8		5407
1000015	P00348242	M	26-35	7	A	1	0	8			7803
1000015	P00042142	M	26-35	7	A	1	0	1	2	6	11458
1000024	P00346642	F	26-35	7	A	3	1	8			2230
1000024	P00205642	F	26-35	7	A	3	1	5	8		6940
1000024	P00248942	F	26-35	7	A	3	1	1	6	14	15774
1000142	P00192042	M	26-35	7	A	2	0	5	9	14	5267
1000142	P00002542	M	26-35	7	A	2	0	8	14		4072
1000142	P00028842	M	26-35	7	A	2	0	6	8		20214

« ‹ 1 / 1054 › »

To ensure the authenticity of the results, the raw data corresponding to the data visualization charts are also shown here in separate pages, which can be viewed in ascending and descending order for each field.

Patterns I have found

In general, I have found the following four patterns:

1. The number of customers in different age groups varies greatly with occupation, but in general, the highest number of customers are in the youth group such as 16-35.

2. The number of products purchased by customers of different occupations in different cities is concentrated in 1-3, and the number of customers decreases significantly as the number of products purchased increases. The overall trend is hyperbolic.
3. Customers in the 26-35 age group have the strongest purchasing power and are not influenced by occupation or city.
4. Among customers of different genders and ages, unmarried men have the highest purchasing power.

How to Run the code

1. First create a virtual environment using anaconda, and then install the following libraries：

    1. `dash`
    2. `plotly`
    3. `pandas`
    4. `numpy`
    5. `dash_bootstrap_components`

2. run `main.py`