1. BlackFriday EDA

* 整理資料 Data Preparation
  1. 將字串及數值混合的欄位(Age)改為數值資料：將年齡(Age)分成6組(Age\_group)

Change Data type

* 1. 利用python的套件製作EDA報表

Use Python Package to make an EDA report

<http://localhost:8888/view/EDA_BlackFriday_report.html>

* 分析資料 Analyze Data
* 資料基本檢視 Data Description

1. 共有537,577筆交易資料，13個變數

537,577 transactions, 13 variables

1. 只有產品種類2有166,986筆缺值及產品種類3有373,299筆缺值，其餘變數皆無缺值

There are no null values in the dataset except product\_catergory2 has 166,986 null values and product\_catergory3 has 373,299 null values.

1. 資料中沒有重複的交易紀錄

There is no duplicated data in this dataset.

* 用戶資料分析：先取出不重複的用戶ID，再看用戶的各項變數，以防止購物超過一次的用戶增加資料數

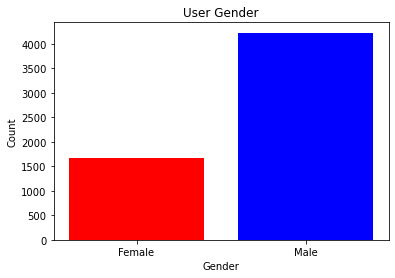
User’s data analysis: retrieve distinct UserID, and analyze each variable to prevent the same User from adding up data.

1. 總共有5891個用戶

There are 5891 users in this dataset

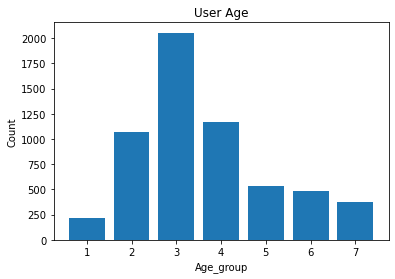
1. 男性用戶共4225個，女性用戶共1666個，**男性**用戶多於女性用戶

There are 4225 male users and 1666 female users, Male users are more than female



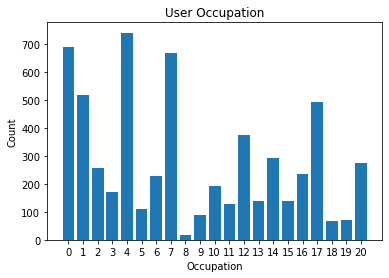
1. 用戶年齡大多為**26-35歲(第3組)**

Most Users’ age is around 26-35 years old (group3)

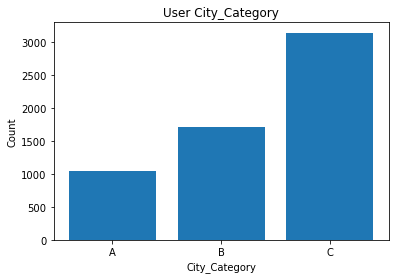


1. 職業代碼為4的用戶最多，8為最少

Most Users’ occupation is categorized in No.4, and No.8 is the less

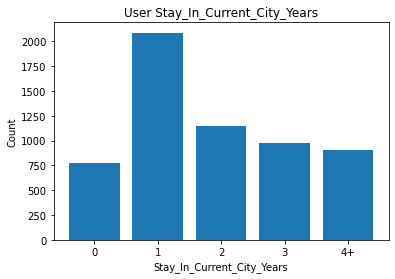


1. C城市的用戶最多 Most Users live in City C



1. 住在目前城市1年的用戶最多

Most users stay in their current city for one year



1. **未婚**用戶多於已婚用戶

Not married users are more than married user



|  |  |  |  |
| --- | --- | --- | --- |
| 用戶代號  UserID | 總消費金額  Total Purchase Amount | 消費次數  Purchase count | 平均每次消費金額  Average Purchase amount |
| 1004277 | 10536783 | 978 | 10773.80675 |
| 1001680 | 8699232 | 1025 | 8487.05561 |
| 1002909 | 7577505 | 718 | 10553.62813 |
| 1001941 | 6817493 | 898 | 7591.863029 |
| 1000424 | 6573609 | 695 | 9458.430216 |
| 1001181 | 6387899 | 861 | 7419.162602 |
| 1000889 | 5499812 | 822 | 6690.768856 |

1. 用戶消費金額排名 Rank of User’s Purchase amount

用戶平均每次消費金額排行 Rank of User’s Average Purchase amount

|  |  |  |
| --- | --- | --- |
| 用戶代號  UserID | 平均每次消費金額  Average Purchase amount | 消費次數  Purchase count |
| 1005069 | 19278.375000 | 16 |
| 1003902 | 18777.247312 | 93 |
| 1005999 | 18345.944444 | 18 |
| 1001349 | 18162.739130 | 23 |
| 1000101 | 17511.369231 | 65 |

由上方兩個表格，可以看出總消費額最高的用戶不是單次消費最高的人，而是**頻繁購物**的用戶

According to the tables above, the user who has the highest purchase amount is not the one who spends the most at one time but the one who frequently purchases.

* 交易資料分析 Overall Transaction Data Analysis

1. 產品銷售量排名 Rank of Product’s sales

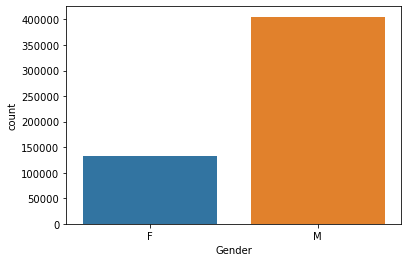
|  |  |
| --- | --- |
| 產品代號 ProductID | 銷售量 sales |
| P00265242 | 1858 |
| P00110742 | 1591 |
| P00025442 | 1586 |
| P00112142 | 1539 |
| P00057642 | 1430 |

1. 總共有3623項產品

There are 3623 products

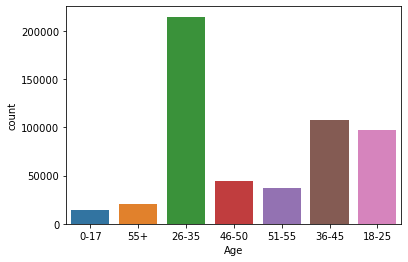
1. **男性**交易次數多於女性

Males’ transactions are more than female’s



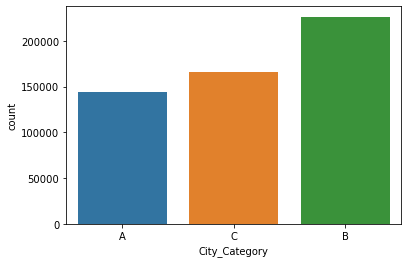
1. **26-35歲**的用戶交易量最多

Ages 26-35 have the most transaction



1. **B城市**為交易量最大的地區

City B has the most transaction



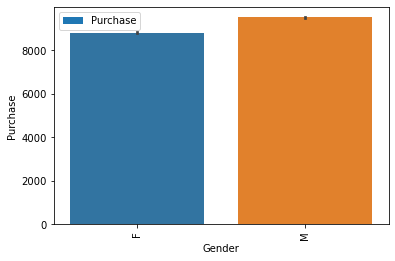
1. **未婚**用戶的交易量比已婚用戶的交易量多

Not married user’s transactions are more than married user’s



1. **男性**用戶平均單次消費金額較女性用戶略高

Male users’ average purchase amount is more than female users’



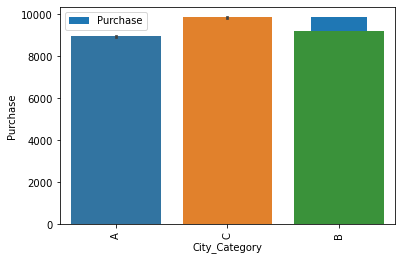
1. 各年齡層平均單次消費金額無顯著差異

There is no significant difference in the purchase amount between each age group



1. 住在**C城市**的用戶單次消費金額較高

Users who live in City C have the highest purchase amount at one time



* 變數間相關性 Correlation between variables

1. 產品種類正相關性高，產品種類2與產品種類1及3最有相關性

Product categories have a high correlation with each other, product category 2 has the most correlation with 1 and 3.

1. 年齡與婚姻狀態有正相關

Age has a positive correlation with Marital Status

1. 年齡相較其他變數，與職業、婚姻狀態、產品種類有正相關性

Age has a relatively positive correlation with occupation, marital status, and product category

1. 用戶消費金額與產品種類1為負相關

The user’s purchase amount has a negative correlation with product category 1

