Black Friday Sales Analysis

1. *Background*

To improve sales effectiveness and make informed data-backed decisions, we need to conduct sales analysis regularly. Sales data analysis provides intelligence about sales strategy, the performance of marketing team and much more, we can even mining the sales data to evaluate the performance of sales against its goals.

We will use SQL to analyze sales by demographic Analysis of customers eg city, age, gender .. The goal of this process is to give more information about our data so that the marketing team prepares to intensify the efficiency based on the data and information we will provide.

2. *Data Resource*

https://www.kaggle.com/girishkurup/new-blackfriday-dataset

3. *Research Content*

A. Overall sales
B. Consumer characteristics

- Percentage of Consumers by Gender and their Purchasing Power
- Percentage of Consumers by Gender and their Purchasing Power in different Product Categories
- Percentage of Consumers by Age and their Purchasing Power
- Percentage of Consumers by Industry and their Purchasing Power
- Product Category Preferences of Consumers by Marital Status
- Purchases of Consumers by City and Years of Residence

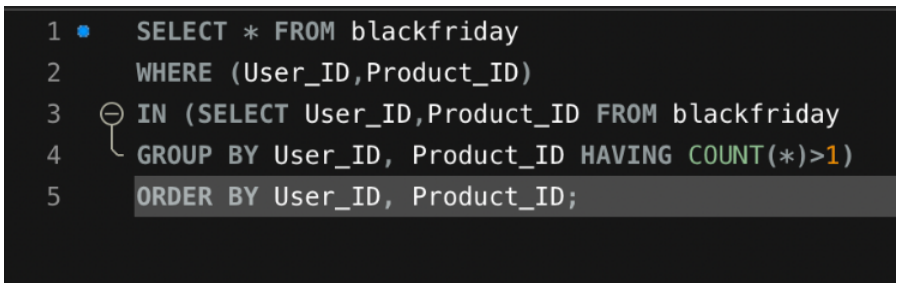
C. Merchandise Basket size Coefficient Analysis

- Overall factors of basket size for Black Friday promotion.
- Basket size of coefficients for different product categories
- Basket size of coefficients by city

4. *Data Understanding*

This dataset contains 12 fields with a total of 550068 rows of data as below:
User_ID, Gender, Age, Occupation, CIty_Category, Stay_In_Current_City_Years, Marital_Status, Product_Category_1, Product_Category_2, Product_Category_3, Purchase.

5. *Data Cleaning*



```
SELECT * FROM blackfriday
WHERE (User_ID,Product_ID)
IN (SELECT User_ID,Product_ID FROM blackfriday
GROUP BY User_ID, Product_ID HAVING COUNT(*)>1)
ORDER BY User_ID, Product_ID;
```

The query shows that the dataset does not contain duplicate values, and the primary key and product category categories are not missing, so data cleansing is unnecessary.

6. *Data Analysis*

A. Overall Sales

```
SELECT COUNT(DISTINCT User_ID) AS 'user_count',
COUNT(product_ID) AS 'product_count',
SUM(purchase) AS 'sum_purchase',
ROUND(SUM(purchase)/COUNT(DISTINCT User_ID),2) AS 'avg_purchase'
FROM blackfriday;
```

100% 33:3

Result Grid Filter Rows: Search Export:

user_count	product_count	sum_purchase	avg_purchase
5891	550068	5095812742	865016.59

The analysis shows that there were 5,891 buyers, 55,068 items purchased, a total purchase amount of $50,9581,2742, and a per capita consumption of $865,016.59.

B. Consumer Characteristics

- Percentage of Consumers by Gender and their Purchasing Power

```
SELECT Gender as 'gender',count(*),
CONCAT(ROUND(COUNT(Gender)/(SELECT COUNT(Gender)FROM blackfriday)*100,2),'%') AS 'share',
SUM(Purchase)
FROM blackfriday
GROUP BY Gender;
```

100% 17:5

Result Grid Filter Rows: Search Export:

gender	count(*)	share	SUM(Purchase)
F	135809	24.69%	1186232642
M	414259	75.31%	3909580100

From the query results, male consumers take the major place, accounting for 75.31% of the total, while female consumers account for only 24.59% of the total. In order to avoid Simpson's Paradox[1], we will break down the percentage of purchasing power of consumers in different product categories.

- Percentage of Consumers by Gender and their Purchasing Power in different Product Categories

[1] Simpson's paradox: A phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.

```sql
SELECT Product_Category_1 AS 'category',
CONCAT (ROUND(SUM(IF(Gender = 'F',Purchase,0))/(SELECT SUM(Purchase)FROM blackfriday)*100,2),'%') AS 'F',
CONCAT (ROUND(SUM(IF(Gender = 'M',Purchase,0))/(SELECT SUM(Purchase)FROM blackfriday)*100,2),'%') AS 'M',
CONCAT (ROUND(SUM(Purchase)/(SELECT SUM(Purchase)FROM blackfriday)*100,2),'%') AS 'share'
FROM blackfriday
GROUP BY Product_Category_1
ORDER BY substring(share,1,length(share)-1)+0 DESC;
```

100% 52:7

Result Grid Filter Rows: Search Export:

category	F	M	share
1	6.63%	30.86%	37.48%
5	5.19%	13.29%	18.48%
8	4.94%	11.83%	16.77%
6	1.40%	4.97%	6.36%
2	1.27%	4.00%	5.27%
3	1.21%	2.80%	4.00%
16	0.69%	2.16%	2.85%
11	0.43%	1.80%	2.23%
10	0.45%	1.53%	1.98%
15	0.30%	1.52%	1.82%
7	0.30%	0.89%	1.20%
4	0.18%	0.36%	0.54%
14	0.17%	0.22%	0.39%
18	0.02%	0.16%	0.18%
9	0.02%	0.10%	0.13%
17	0.01%	0.10%	0.12%
12	0.04%	0.06%	0.10%
13	0.02%	0.06%	0.08%
20	0.01%	0.01%	0.02%
19	0.00%	0.00%	0.00%

Result Grid Form Editor Field Types Query Stats Execution Plan

The query divides consumers by gender and calculates the percentage of purchases in each product category. Out of a total of 20 categories, all product categories have higher purchases by men than by women, except for category 20, which has equal purchases by men and women, and category 19, for which no data is available, so Simpson's paradox can be eliminated. Among them, products in label 1, 5, and 8 occupy the top three spots in the Black Friday sales list.

- Percentage of Consumers by Age and their Purchasing Power

```sql
SELECT Age as'Age',COUNT(*) AS 'amount',
CONCAT(ROUND(COUNT(Age)/(SELECT COUNT(Age) FROM blackfriday)*100,2),'%')AS 'share',SUM(Purchase)
FROM blackfriday
GROUP BY Age
ORDER BY substring(share,1,length(share)-1)+0 DESC;
```

100% 1:6

Result Grid Filter Rows: Search Export:

Age	amount	share	SUM(Purchase)
26-35	219587	39.92%	2031770578
36-45	110013	20.00%	1026569884
18-25	99660	18.12%	913848675
46-50	45701	8.31%	420843403
51-55	38501	7.00%	367099644
55+	21504	3.91%	200767375
0-17	15102	2.75%	134913183

Result Grid Form Editor Field

The data divides age into seven groups as shown in the figure, and we can see that the age of Black Friday purchasers is mainly concentrated between 18 and 45 years old, with the age group of 26-35 accounting for 39.92% and the total shopping amount of 20,3177,0578 USD. The age group of 36-45 ranks second, accounting for 20%, and the age group of 18-25 ranks third, accounting for 18.12%. Therefore, the store's customers are mainly young and middle-aged, while children and the elderly are excluded.

- Percentage of Consumers by occupation and their Purchasing Power

```
SELECT Occupation as'Occupation',COUNT(*) AS 'amount',
CONCAT(ROUND(COUNT(Occupation)/(SELECT COUNT(Occupation) FROM blackfriday)*100,2),'%')AS 'share',SUM(Purchase)
FROM blackfriday
GROUP BY Occupation
ORDER BY substring(share,1,length(share)-1)+0 DESC LIMIT 10;
```

100% 1:7

Result Grid | Filter Rows: Search | Export:

Occupation	amount	share	SUM(Purchase)
4	72308	13.15%	666244484
0	69638	12.66%	635406958
7	59133	10.75%	557371587
1	47426	8.62%	424614144
17	40043	7.28%	393281453
20	33562	6.10%	296570442
12	31179	5.67%	305449446
14	27309	4.96%	259454692
2	26588	4.83%	238028583
16	25371	4.61%	238346955

Result Grid | Form Editor | Field Types

We only selected top 10 occupation out of the 20 industries included in the dataset. The results pointed out that the largest percentage of customers mainly from three occupation: 4, 0, and 7, respectively. Further analysis is not possible due to limited information.

- Product Category Preferences of Consumers by Marital Status

```
SELECT Marital_Status AS'Marital_Status',COUNT(*) AS 'amount',
CONCAt(ROUND(COUNT(Marital_Status)/(SELECT COUNT(Marital_Status)FROM blackfriday)*100,2),'%')as 'share',
SUM(Purchase)
FROM blackfriday
GROUP BY Marital_Status
ORDER BY substring(share,1,length(share)-1)+0 DESC;
```

100% 1:7

Result Grid | Filter Rows: Search | Export:

Marital_Status	amount	share	SUM(Purchase)
0	324731	59.03%	3008927447
1	225337	40.97%	2086885295

Result Grid

Marital status in the data set is 0 for unmarried and 1 for married. The results indicate that unmarried consumers consume more than married consumers, with 59.03% of all consumers unmarried and 40.97% of all consumers married.

```
SELECT COALESCE(Product_Category_1,'sum')AS 'category',
CONCAT(ROUND(SUM(IF(Marital_Status='0',Purchase,0))/(SELECT SUM(Purchase)FROM blackfriday)*100,2),'%')AS 'Married',
CONCAT(ROUND(SUM(IF(Marital_Status='1',Purchase,0))/(SELECT SUM(Purchase)FROM blackfriday)*100,2),'%')AS 'Unmarried'
FROM blackfriday
GROUP BY Product_Category_1;
```

100% 29:5

Result Grid | Filter Rows: Search | Export:

category	Married	Unmarried
3	2.43%	1.57%
1	22.50%	14.98%
12	0.05%	0.05%
8	9.61%	7.16%
5	10.95%	7.53%
4	0.33%	0.21%
2	3.10%	2.17%
6	3.78%	2.58%
14	0.22%	0.18%
11	1.36%	0.87%
13	0.04%	0.03%
15	1.05%	0.77%
7	0.66%	0.54%
16	1.65%	1.19%
18	0.10%	0.09%
10	1.07%	0.91%
17	0.06%	0.06%
9	0.08%	0.05%
20	0.01%	0.01%
19	0.00%	0.00%

Result Grid | Form Editor | Field Types | Query Stats | Execution Plan

Both married and unmarried consumers have a preference for products in the 1, 5, and 8 from categories. Marital status does not reflect a significant difference in product preference.

- Purchases of Consumers by City and Years of Residence

```
SELECT COALESCE(S.Stay_In_Current_City_Years,'SUM')AS 'Years',
CONCAT(ROUND(SUM(IF(S.City_Category='A',sum_purchase,0))/(SELECT SUM(Purchase)FROM blackfriday)*100,2),'%')AS 'A',
CONCAT(ROUND(SUM(IF(S.City_Category='B',sum_Purchase,0))/(SELECT SUM(Purchase)FROM blackfriday)*100,2),'%')AS 'B',
CONCAT(ROUND(SUM(IF(S.City_Category='C',sum_Purchase,0))/(SELECT SUM(Purchase)FROM blackfriday)*100,2),'%')AS 'C'
FROM (SELECT DISTINCT User_ID, Stay_In_Current_City_Years,City_Category,SUM(Purchase)AS sum_purchase FROM blackfriday
GROUP BY User_ID, Stay_In_Current_City_Years, City_Category) S
GROUP BY Stay_In_Current_City_Years
```

100% 1:8

Result Grid | Filter Rows: Search | Export:

Years	A	B	C
2	4.77%	7.54%	6.32%
4+	3.87%	6.24%	5.31%
3	4.34%	7.70%	5.33%
1	8.59%	15.03%	11.57%
0	4.26%	5.01%	4.13%

Result Grid | Form Editor

In terms of period of resident , consumers who have lived in the city for about one year have the highest purchasing power. Overall, consumers in city B contribute most of the sales, followed by those in city C and third in city A. However, this statistic does not indicate that consumers in city B have more purchasing power, so we will continue to analyze the purchasing power per capita of each city.

```
1  SELECT City_Category AS'city_category',
2  ROUND(SUM(sum_Purchase)/COUNT(User_ID),2)AS'PPP'
3  FROM(SELECT DISTINCT User_ID,City_Category,
4  SUM(Purchase)AS sum_purchase FROM blackfriday GROUP BY User_ID,City_Category)C
5  GROUP BY City_Category
6  ORDER BY City_Category;
```

100% 24:6

Result Grid Filter Rows: Search Export:

city_category	PPP
A	1259781.49
B	1239328.42
C	530043.80

The results of the analysis indicate that the city with the highest purchasing power per capita is city A and city B is in second place, and the assumption that consumers in city B have more punching power has been confirm a Simpson's paradox.

C. Merchandise Basket size[2] Analysis

- Overall factors of basket size for Black Friday promotion.

```
1  SELECT COUNT(Product_ID) AS 'Count_Of_Products',
2  COUNT(DISTINCT User_ID) AS 'Count_Of_orders',
3  COUNT(Product_ID)/COUNT(DISTINCT USER_ID) AS 'basket size'
4  FROM blackfriday;
```

100% 18:4

Result Grid Filter Rows: Search Export:

Count_Of_Products	Count_Of_orders	basket size
550068	5891	93.3743

From the statistics, the basket size of the Black Friday sales is approximately 93, which means that each customer purchased about 93 items averagely .

- Basket size of different product categories

[2] Basket size Factor = total number of items sold in a given time period / total number of basket sizes in a given time period, which indicates the average number of items purchased per customer at one time.

```sql
SELECT Product_Category_1 AS 'Product_Category',
COUNT(Product_ID) AS 'Count_Of_Products',
COUNT(DISTINCT User_ID) AS 'Count_Of_orders',
COUNT(Product_ID)/COUNT(DISTINCT USER_ID) AS 'basket_size'
FROM blackfriday
GROUP BY Product_Category_1;
```

100% 29:6

Result Grid Filter Rows: Search Export:

Product_Category	Count_Of_Products	Count_Of_orders	basket_size
1	140378	5767	24.3416
2	23864	4296	5.5549
3	20213	3838	5.2665
4	11753	3361	3.4969
5	150933	5751	26.2447
6	20466	4085	5.0100
7	3721	1461	2.5469
8	113925	5659	20.1316
9	410	410	1.0000
10	5125	2328	2.2015
11	24287	3583	6.7784
12	3947	1567	2.5188
13	5549	2272	2.4423
14	1523	971	1.5685
15	6290	2440	2.5779
16	9828	3130	3.1399
17	578	426	1.3568
18	3125	1284	2.4338
19	1603	1603	1.0000
20	2550	2550	1.0000

The top three basket size coefficients for each category are label 1, 5, and 8. Unlike the label 1 takes the top position of spending list, the highest basket size coefficient listed in label.

- Basket size of different city

```sql
SELECT City_Category AS 'City',
COUNT(Product_ID)AS'Count_Of_Products',
COUNT(DISTINCT User_ID)AS'Count_Of_Orders',
COUNT(Product_ID)/COUNT(DISTINCT User_ID)AS'basket_size'
FROM blackfriday
GROUP BY City_Category;
```

100% 24:6

Result Grid Filter Rows: Search Export:

City	Count_Of_Products	Count_Of_Orders	basket_size
A	147720	1045	141.3589
B	231173	1707	135.4265
C	171175	3139	54.5317

According to the results, the basket size coefficient of both city A and city B is much higher than the overall basket size of 93, while the basket size of city C is lower than the overall basket size coefficient, and much lower than that of both cities A and B.

Consumer Characteristics.

There were 5,891 customers who made transactions in the Black Friday sale, with the number of items traded reaching 550,068, the total purchase amount of US$50,9581,2742, and the per capita spending of US$865,016.59. Among them, mainly male customers, with no clear preference in product categories by gender. Meanwhile, the store's customers were mainly young and middle-aged, with the 26-35 age group having the highest number of customers and spending the highest amount of money. Marital status had no influence on customers' product category preferences. The customers with the largest spending amount in this campaign were from the 4, 0, and 7 industry categories, respectively.

Suggestion

It is recommended to refine the classification of customers, distinguish between new customers, returning customers and loyal customers, and adopt different preferential strategies for different customers, especially to improve the retention rate of customers with higher spending power, and increase promotion efforts for potential customers.
Merchandise basket size coefficient analysis.
Products are categorized into top sellers, ordinary products, and marginal products.
Bestsellers: The basket size coefficients and the number of basket size of products in categories 1, 5, and 8 are much higher than the average, and are the main source of sales.

Common products: 2, 3, 4, 6, 11, 16 and other categories of products have higher than average number of shopping baskets, but the value of the basket coefficient is almost the same as the average or lower than the average, so the sales performance is not obvious.
Marginal products: 7, 9, 10, 12, 13, 14, 15, 18 and other categories are all below the average shopping basket coefficient and number of shopping baskets, and are not sold well during the event.
Suggestions: Pay attention to the products with high basket size and number of baskets, and put them in a conspicuous position or increase the number of containers when displaying the products. For general merchandise with low basket size, attention should be paid to improving the correlation between merchandise, such as buy one get one free or bundled sales. For marginal merchandise, increase promotions and increase the number of sales through bundled sales.