Sample

The dataset links is: <https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store?select=2019-Nov.csv>

We are using the datasets for eCommerce behavior data from multiple category stores. The dataset comes with many columns including category, product\_id, brand, event\_time, event\_type etc. From the dataset, we can analyze a few things that may help with sales decisions and how to improve the user experience as below:

1. To figure out the percentage of purchase from view and get the top category of products.

select count(\*) from project\_table group by event\_type;

Text

Description automatically generated with medium confidence

1. To analyze which brand has the best chance to be purchased, and why?

select brand, count(\*) cnt

from project\_table

where brand is not null and brand != ''

group by brand

order by cnt desc

limit 5;

Text

Description automatically generated

1. To analyze if there are relations between prices and purchase rate?

select max(price) from project\_table;

2574.07

select min(price) from project\_table;

0.0

select avg(price) from project\_table;

292.46

select price\_range, count(\*) as num

from(

select case

when price <= 50 then '0-50'

when price > 50 and price <= 100 then '50-100'

when price > 100 and price <= 200 then '100-200'

when price > 200 and price <= 300 then '200-300'

when price > 300 and price <= 400 then '300-400'

when price > 400 and price <= 500 then '400-500'

when price > 500 and price <= 1000 then '500-1000'

else 'over 1000'

end as price\_range

from project\_table

) as price\_summary

group by price\_range;

this query cannot be completed for 2 hours so I wrote python codes to forloop the data to get the answer:

{

      '0-50': 152856,

      '50-100': 114018,

      '100-200': 244631,

      '200-300': 134998,

      '300-400': 63640,

      '400-500': 44665,

      '500-1000': 116348,

      '1000+': 45783

    }

1. To analyze the time frames that the purchase rate is higher than other times.

select hour(event\_time) event\_hour, count(\*)

from project\_table

where event\_type = 'purchase' and event\_time is not null

group by hour(event\_time);

This query cannot be completed for 2 hours. So I wrote python codes to forloop the data to get the answer:

{

        0: 448734,

        1: 843397,

        2: 1676249,

        3: 2399278,

        4: 3066909,

        5: 3444239,

        6: 3554901,

        7: 3584742,

        8: 3689021,

        9: 3653758,

        10: 3627788,

        11: 3582860,

        12: 3588308,

        13: 3908344,

        14: 4335113,

        15: 4451101,

        16: 4511068,

        17: 4415833,

        18: 3266804,

        19: 2375711,

        20: 1431589,

        21: 825518,

        22: 480585,

        23: 340129

      }

1. To generate some data visualization with figures.

Open price-range.html and purchase-time.html in the Chrome Browser to see the charts.

Chart, pie chart

Description automatically generated

Chart

Description automatically generated

1. To make better decisions for the sales department.

According to charts, customers would like to make purchase from 14:00 to 18:00 so we could make a discount or put ads during this time frame.

And customers are most likely to make purchases with prices 100-200 so we could put focus on this price range.