CSI_4_DMA

DISCRETE MATHEMATICS COURSE WORK

Student ID: 4238064
SUBMITTED TO: AARBAZ ALAM

Table of Contents

1. Overview 2
2. Importing Libraries 2
3. Dataset Used 2
4. Methods 4
   a. analysis techniques 4
   b. sort information 4
   c. region bar chart 7
   d. Number of item ordered 7
   e. genders pie chart 8
   f. categories pie chart 9
   g. price and value scatter plot 10
5. Conclusion 11
6. Bibliography 11

List of figures

Figure 1_import libraries 2
Figure 2_read dataset 3
Figure 3_head function 3
Figure 4_tail function 4
Figure 5_ duplicated 4
Figure 6_check unique values 5
Figure 7_nunique function 5
Figure 8_drop function 6
Figure 9_describe function 6
Figure 10_bar chart of region 7
Figure 11_histogram number of item in order 8
Figure 12_pie chart of genders 9
Figure 13_pie chart of categories 10
Figure 14_scatter plot price and value 11

1. Overview

A data set with details of orders and the customers information was given to me as a data analyst. This dataset can give valuable information related to the order, which is price, category, status, … and related to the customers like address and name.

The data appears to be the data given is very complete and detailed about the orders, and there seems to be nothing confusing based on the initial analysis. It is clearly that this data was gathered in 2021 from February to September and consisted of both numerical data about the value of item and a lot of categorical variables. I will analysis the data and pointed out which kind of item was the bestseller and what types of customers bought the most product.

2. Importing Libraries

To analysis data, first step must be importing the library into Python. Data manipulation and visualisation will be easier to understand and break down with the functions of libraries like NumPy, Pandas and Matplotlib. They will help you to read data quicker by doing computations, making charts and graphs, and with that summarized information, analysis will be easier.

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
```

Figure 1_import libraries

3. Dataset Used

The data set contains details of orders and the info about the customer who purchased them. This includes necessary characteristic about the order like what has been bought, the quantity of the item, total price and who bought them, the report of the customers. The data given was neat and make no confusing, the collected information is more than enough to analysis.

```
In [2]: order = pd.read_csv("4238064_dataset.csv")
        order
```

Out[2]:

	order_id	order_date	status	item_id	sku	qty_ordered	price	value	discount_amount	total	..
0	100445581	16/02/2021	complete	730939	APPANE5A584FF945176	2	306.9	306.9	0.0	306.9	..
1	100445581	16/02/2021	complete	730940	HALCLI5A1EA4DECC6C9	2	34.7	34.7	0.0	34.7	..
2	100445581	16/02/2021	complete	730941	HALDYN5A00299858EA6	2	35.0	35.0	0.0	35.0	..
3	100446643	22/02/2021	complete	732849	APPPHI59FADFA46C15E	2	249.8	249.8	0.0	249.8	..
4	100446643	22/02/2021	complete	732850	APPANE59D4925C945DB	2	992.6	992.6	0.0	992.6	..
...	...	...	...	...	...	...	...	...	...	...	..
1214	100446474	21/02/2021	received	732548	MEFKEN5A604F71E202A	2	0.0	0.0	0.0	0.0	..
1215	100446474	21/02/2021	received	732549	MEFKEN5A604F6F07118	2	0.0	0.0	0.0	0.0	..
1216	100446474	21/02/2021	received	732550	MEFKEN5A604F6F8EB71	2	0.0	0.0	0.0	0.0	..
1217	100446474	21/02/2021	received	732551	MEFKEN5A604F73ABEBA	2	0.0	0.0	0.0	0.0	..
1218	100446474	21/02/2021	received	732552	MEFKEN5A604F715DE52	2	0.0	0.0	0.0	0.0	..

1219 rows × 26 columns

Figure 2_read dataset

It appears that this data set was collected in 8 months since February 2021, and it show that several kinds of items had been ordered. The categories mentioned in the data set are order id, order date, status, item id, stoke keeping unit, quantity order, price, value, discount amount, total, category, payment method, bi_st, customer id, year, month, name prefix, gender, age, customer since, county, city, state, zip, region, discount percentage. Most of the variables are categorical, while minority is numerous.
Show the first five rows of the dataset.

```
In [3]: order.head()
```

Out[3]:

	order_id	order_date	status	item_id	sku	qty_ordered	price	value	discount_amount	total	...	N P
0	100445581	16/02/2021	complete	730939	APPANE5A584FF945176	2	306.9	306.9	0.0	306.9	...	
1	100445581	16/02/2021	complete	730940	HALCLI5A1EA4DECC6C9	2	34.7	34.7	0.0	34.7	...	
2	100445581	16/02/2021	complete	730941	HALDYN5A00299858EA6	2	35.0	35.0	0.0	35.0	...	
3	100446643	22/02/2021	complete	732849	APPPHI59FADFA46C15E	2	249.8	249.8	0.0	249.8	...	
4	100446643	22/02/2021	complete	732850	APPANE59D4925C945DB	2	992.6	992.6	0.0	992.6	...	

5 rows × 26 columns

Figure 3_head function

Show last five rows of the dataset.

```
In [4]: order.tail()
```

Out[4]:

	order_id	order_date	status	item_id	sku	qty_ordered	price	value	discount_amount	total	...
1214	100446474	21/02/2021	received	732548	MEFKEN5A604F71E202A	2	0.0	0.0	0.0	0.0	...
1215	100446474	21/02/2021	received	732549	MEFKEN5A604F6F07118	2	0.0	0.0	0.0	0.0	...
1216	100446474	21/02/2021	received	732550	MEFKEN5A604F6F8EB71	2	0.0	0.0	0.0	0.0	...
1217	100446474	21/02/2021	received	732551	MEFKEN5A604F73ABEBA	2	0.0	0.0	0.0	0.0	...
1218	100446474	21/02/2021	received	732552	MEFKEN5A604F715DE52	2	0.0	0.0	0.0	0.0	...

5 rows × 26 columns

Figure 4_tail function

4. Methods

a. analysis techniques

There are a lot of methods to examining and evaluating data. For this data set, the techniques below will be applied to analysis:

- Data visualization: This is a technique that representation of data through use common graphics, such as charts, plots, infographics and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.
- Descriptive analysis: this method will be used to reflect the analysis. By sorting, manipulating, and interpreting raw data from a variety of source to answer the question what happened.

b. sort information

Code checks the duplicated data in a row. If any data is duplicated it will show true, otherwise false.

```
In [5]: order.duplicated()
```

```
Out[5]: 0       False
        1       False
        2       False
        3       False
        4       False
                ...
        1214    False
        1215    False
        1216    False
        1217    False
        1218    False
        Length: 1219, dtype: bool
```

Figure 5_ duplicated

Code counts the number of unique values in a data frame and return the sum.

```
In [6]: num_unique_values = order.value_counts().sum()
        num_unique_values
Out[6]: 1219
```

Figure 6_check unique values

Code to show how many are there in each category.

```
In [8]: order.nunique()
Out[8]: order_id            839
        order_date          103
        status                7
        item_id            1219
        sku                 748
        qty_ordered          23
        price               419
        value               471
        discount_amount      24
        total               476
        category             15
        payment_method       10
        bi_st                 3
        cust_id             340
        year                  1
        month                 8
        Name Prefix           7
        Gender                2
        age                  58
        Customer Since      332
        County              260
        City                321
        State                51
        Zip                 339
        Region                4
        Discount_Percent     39
        dtype: int64
```

Figure 7_nunique function

From the unique function its concluded that category year has only one value corresponding for all the rows, so it makes no effect to the data analysis, that is the reason I am deleting it with data drop function.

```
In [9]: order.drop("year", axis = 1, inplace = True)
```

```
In [10]: order
```

Out[10]:

	order_id	order_date	status	item_id	sku	qty_ordered	price	value	discount_amount	total	..
0	100445581	16/02/2021	complete	730939	APPANE5A584FF945176	2	306.9	306.9	0.0	306.9	..
1	100445581	16/02/2021	complete	730940	HALCLI5A1EA4DECC6C9	2	34.7	34.7	0.0	34.7	..
2	100445581	16/02/2021	complete	730941	HALDYN5A00299858EA6	2	35.0	35.0	0.0	35.0	..
3	100446643	22/02/2021	complete	732849	APPPHI59FADFA46C15E	2	249.8	249.8	0.0	249.8	..
4	100446643	22/02/2021	complete	732850	APPANE59D4925C945DB	2	992.6	992.6	0.0	992.6	..
...	...	...	...	...	...	...	...	...	...	...	..
1214	100446474	21/02/2021	received	732548	MEFKEN5A604F71E202A	2	0.0	0.0	0.0	0.0	..
1215	100446474	21/02/2021	received	732549	MEFKEN5A604F6F07118	2	0.0	0.0	0.0	0.0	..
1216	100446474	21/02/2021	received	732550	MEFKEN5A604F6F8EB71	2	0.0	0.0	0.0	0.0	..
1217	100446474	21/02/2021	received	732551	MEFKEN5A604F73ABEBA	2	0.0	0.0	0.0	0.0	..
1218	100446474	21/02/2021	received	732552	MEFKEN5A604F715DE52	2	0.0	0.0	0.0	0.0	..

1219 rows × 25 columns

Figure 8_drop function

There are 1219 sets of measurements in this data cover the following variables: order_id, item_id, qty_ordered, price, value, discount_amount, total, cust_id, age, Zip, Discount_Percent. Because each order and each item have their own id, then the statistics of these section was not use for calculation. Similarly, customer id was use for identifying the customer. On the other hand, the standard deviation of quantity ordered was 6.70 and the mean was 4.05. With the standard deviation of 1658.88, the mean of price was 575.27. The standard deviation was 2090.34 and the mean was 655.4 for value section. 38.08 and 5.09 were the data of discount amount, respectively. Total had the standard validation at 2089.45 and an average of 650.30. With the standard validation of 17.62, the average age in this data set was 42.19. The standard validation of Zip was 23265.78 and the mean for it was 46239.18. Finally, the discount percentage had 5.91 at standard validation and 1.10 at mean.

```
In [11]: order.describe()
```

Out[11]:

	order_id	item_id	qty_ordered	price	value	discount_amount	total	cust_id	1
count	1.219000e+03	1219.000000	1219.000000	1219.000000	1219.000000	1219.000000	1219.000000	1219.000000	
mean	1.004646e+08	758506.909762	4.059885	575.276830	655.401170	5.099363	650.301806	75877.688269	
std	3.156216e+04	45032.730046	6.701263	1658.884924	2090.343957	38.083541	2089.458452	22487.581058	
min	1.004456e+08	730939.000000	1.000000	0.000000	0.000000	0.000000	0.000000	675.000000	
25%	1.004459e+08	731824.000000	2.000000	24.432500	26.600000	0.000000	26.600000	85377.000000	
50%	1.004463e+08	732292.000000	2.000000	72.000000	89.900000	0.000000	89.900000	85428.000000	
75%	1.004703e+08	770458.000000	3.000000	183.550000	271.350000	0.000000	259.950000	85517.000000	
max	1.005613e+08	903204.000000	38.000000	15249.900000	20000.000000	530.604000	20000.000000	85630.000000	

Figure 9_describe function

c. region bar chart

The bar chart below illustrates the number of orders from four different region: South, Northeast, Midwest, and West in the year of 2021. The quantity of orders come from the South was the biggest while the West had the smallest number of orders in 2021. The South's number of orders was 488 orders. In the second place was the figure from Midwest, which is less than the figure from the West 36 orders and finally the number of orders from customers live in the Northeast was nearly reach 200. Finally, 89 was the number of orders come from the West.

```
In [21]: plt.hist(order['Region'])
         bb = ["South","Northeast","Midwest","West"]
         cc = [488,194,452,89]
         plt.bar (bb,cc)
         for i in range(len(bb)):
             plt.text(i, cc[i]+1, str(cc[i]), ha='center', va='bottom')
         plt.title('Region histogram')
         plt.xlabel('region')
         plt.ylabel('order')

         plt.show()
```

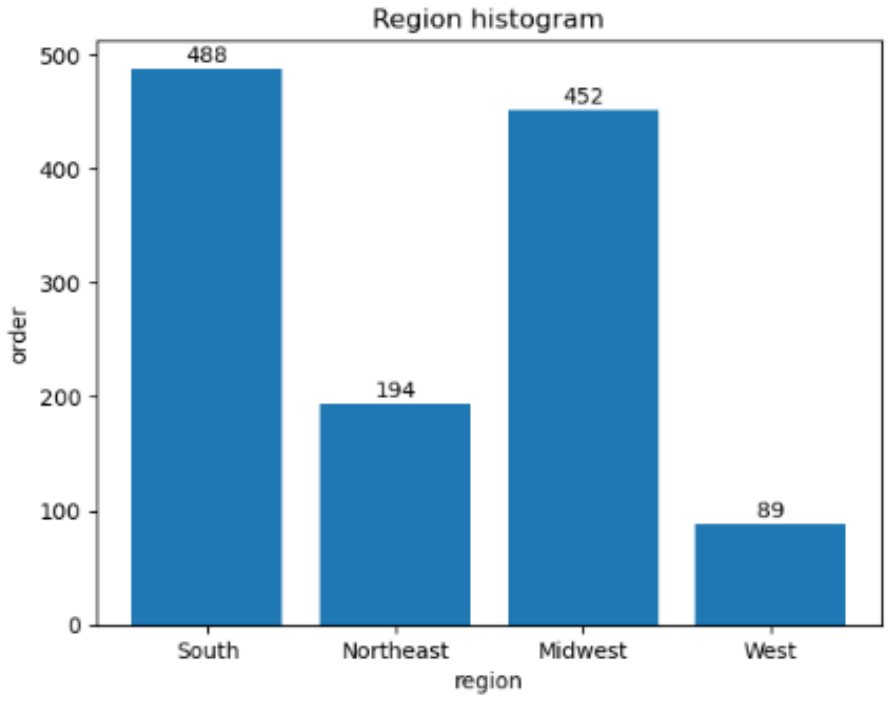

Figure 10_bar chart of region

d. Number of item ordered

The histogram below demonstrated number of item had been ask and how many order with that amount of item. Overall, it is clearly seen that most of the customer ordered an 2 items order and the order with highest item of number is 38. Customer prefer order a pair because it had the highest figure at 882 orders while 185 orders with 3 items was the second most. There were only seven types of order that have more than 10 number of order respectively order with one, six, eleven, twelve, twenty-nine, thirty and thirty-eight. The rest of the histogram were around one to seven.

```
In [60]: number_item = ["1", "2", "3", "4", "5", "6", "7", "9", "10", "11", "12", "13", "15", "16",
                        "21", "22", "23", "24", "29", "30", "31", "37", "38"]
         quantity = [17,882,185,3,4,18,2,3,1,18,17,7,3,1,2,1,1,1,13,13,5,1,21]
         plt.bar(number_item, quantity)
         for i in range(len(number_item)):
             plt.text(i, quantity[i]+1, str(quantity[i]), ha='center', va='bottom')
         plt.xlabel('number of item ordered')
         plt.ylabel('Number of order')
         plt.title('Histogram of number item')
         plt.tight_layout()
         plt.show()
```

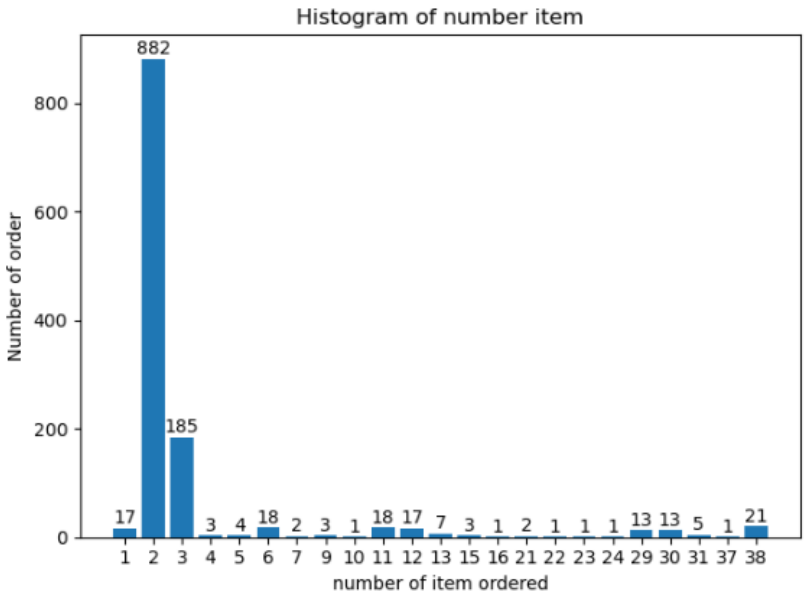

Figure 11_histogram number of item in order

e. genders pie chart

The pie chart below shows the percentage of male and female in order list. Overall, the percentage for each figure does not have a big difference, there were 56.6% of the customers are male and 43.4% is the percentage of female customer in the year of 2021. These percentage could give a comment that the customer gender was balance well.

```
In [22]: my_data = [690,529]
         my_labels = ["male", "female"]

         plt.pie(my_data, labels=my_labels, autopct="%1.1f%%")
         plt.title("Gender")
         plt.axis("equal")
         plt.show()
```

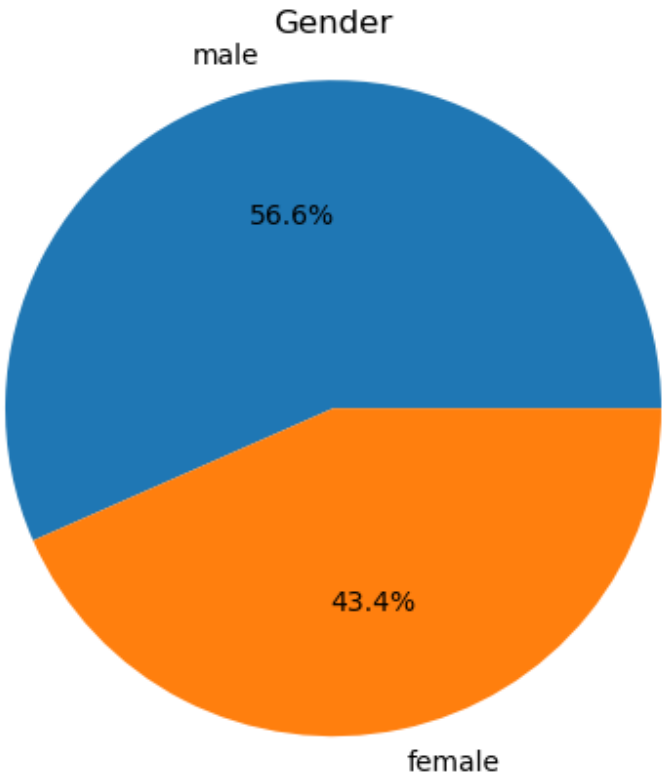

Figure 12_pie chart of genders

f. categories pie chart

The pie chart shows the percentage of each category that have been order from February to September in 2021. There are 15 types of categories respectively Kids & Baby, Men's Fashion, Mobiles & Tablets, Others, School & Education, Soghaat, Superstore, Women's Fashion, Appliances, Beauty & Grooming, Books, Computing, Entertainment, Health & Sports, Home & Living. At first, Men's Fashion and Women's Fashion are the best-selling categories based on the part shown in the pie chart. Besides, item in the Mobiles & Tablets section has also been ordered a lot and standing right after that is the order from category named others. Nearly 50% of the remaining were belong to other 11 sections, which Beauty & Grooming category accounts for most of them. It is easy to see that books orders took up the smallest part of all.

```
In [36]: cc =order.groupby("category")["category"].count()

         plt.pie(cc, labels=cc.index)
         plt.title("categories")
         plt.show()
```

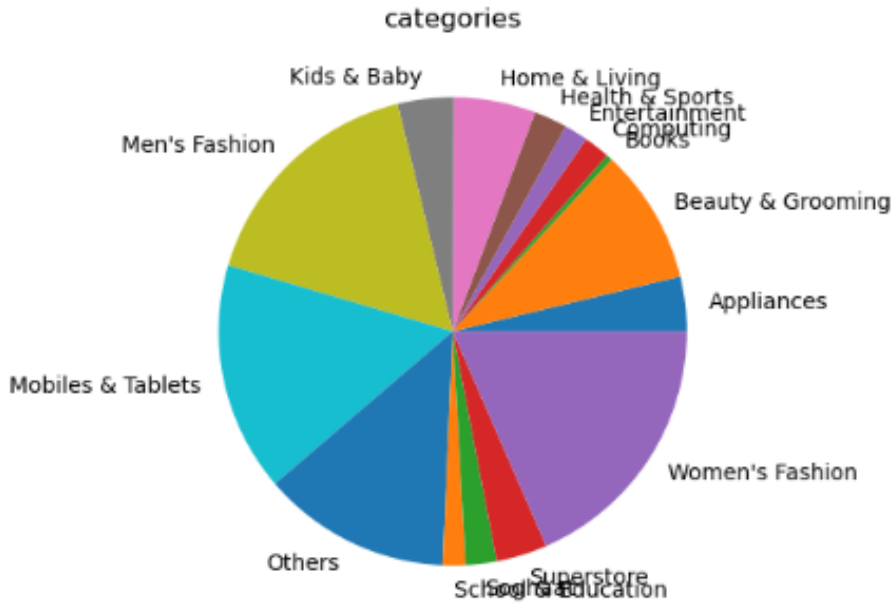

Figure 13_pie chart of categories

g. price and value scatter plot

This scatter plot represents the relationship between price of the order and the value of them. The order with higher price is located to the right and the order with higher value are placed higher up in the graph. On the bottom left side, there are a lot of dots located there and this demonstrate that the amount of order which have low price and low value were take up a large part of total. And on the top right of the plot, the 2 dots are the order with high value and price, and this just take a small amount.

```
In [23]: plt.scatter(corr['price'], corr['value'])
         plt.title('Price and value')
         plt.xlabel('price')
         plt.ylabel('value')
Out[23]: Text(0, 0.5, 'value')
```

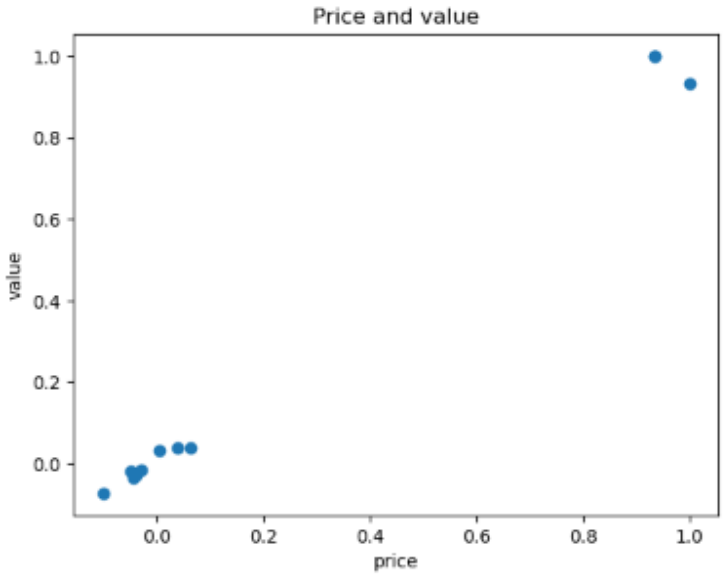

Figure 14_scatter plot price and value

5. Conclusion

As a data analysis, after examining this data set, I came to the following conclusions:

The dataset reveal the top selling categories that the fashion are the most selling categories for both men and women, and this also balance customer gender which make sales better because of the profit gain from both male customer and female customer. Moreover, it also clarify where most of there customers come from and the company can shifts the direction of serving to the South and the Midwest.

The number of item ordered is showing their ordering behavior of purchasing a pair instead of other quantity, the company can base on this information and give a promotion about buying many pair at once. Furthermore, the price and value of the order seem to leaning towards cheap items. This is the signal that the customer are likely to purchase low price item rather than the expensive one.

These conclusions were reached after analysis the dataset through the several charts. In addition, more information may be available based on the application of other methods.

6. Bibliography

Hudabai Soomro (January 17, 2023). Essential types of data analysis methods and processes for business success. Data analysis process, descriptive analysis technique. Available at: https://datasciencedojo.com/blog/data-analysis-methods/#

Data to Fish (no date). Python tutorials - Matplotlib. Avaiable at: https://datatofish.com/python-tutorials/