DATA VISUALIZATION

Group 8: Hye-Jin Cho-Drugeon

Data: Supermarket

À PROPOS

Plainly, I analyze the data frame. I would like to investigate which variables are capably explaining this data's causality.

There are 17 columns.

```
df.head()
```

Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1
226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29	Cash	76.40	4.761905	3.8200	9.6
631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit card	324.31	4.761905	16.2155	7.4
123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	20:33	Ewallet	465.76	4.761905	23.2880	8.4
373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.2085	5.3

THE TYPE OF VARIABLES

THE CATEGORY OF PRODUCT LINE

```
df['Product line'].value_counts().head(10)
```

```
Fashion accessories        178
Food and beverages         174
Electronic accessories     170
Sports and travel          166
Home and lifestyle         160
Health and beauty          152
Name: Product line, dtype: int64
```

VARIABLE TYPES

```
df.dtypes
```

```
Invoice ID                 object
Branch                     object
City                       object
Customer type              object
Gender                     object
Product line               object
Unit price                float64
Quantity                    int64
Tax 5%                    float64
Total                     float64
Date                       object
Time                       object
Payment                    object
cogs                      float64
gross margin percentage   float64
gross income              float64
Rating                    float64
dtype: object
```

DESCRIBE

```
df.describe()
```

	Unit price	Quantity	Tax 5%	Total	cogs	gross margin percentage	gross income	Rating
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000	1.000000e+03	1000.000000	1000.00000
mean	55.672130	5.510000	15.379369	322.966749	307.58738	4.761905e+00	15.379369	6.97270
std	26.494628	2.923431	11.708825	245.885335	234.17651	6.131498e-14	11.708825	1.71858
min	10.080000	1.000000	0.508500	10.678500	10.17000	4.761905e+00	0.508500	4.00000
25%	32.875000	3.000000	5.924875	124.422375	118.49750	4.761905e+00	5.924875	5.50000
50%	55.230000	5.000000	12.088000	253.848000	241.76000	4.761905e+00	12.088000	7.00000
75%	77.935000	8.000000	22.445250	471.350250	448.90500	4.761905e+00	22.445250	8.50000
max	99.960000	10.000000	49.650000	1042.650000	993.00000	4.761905e+00	49.650000	10.00000

DISTRIBUTION HAS THE VARIATY. FOR EXAMPLE, STANDARD DEVIATION IS FROM MINIMAL VALUES SUCH AS 0 (GROSS INCOME), 1 (QUANTITY).

RATING IS NOT UNIFORMLY DISTRIBUTED AND CONDENSED CLOSE TO 6.

PRESENTATION DU PRODUIT

REPLACE

Make the category variables to Boolean types or numeric values for the correlation check.

NULL VARIABLES

Relative low attention on null variables except for one abnormal variable

SIGNIFICANT PRCING INFO.

There is a tax info. which ought to be corresponding to unit prices and amounts.

PREDICTION

Prices are required for prediction analysis for next saling.

REPLACES FOR DATA

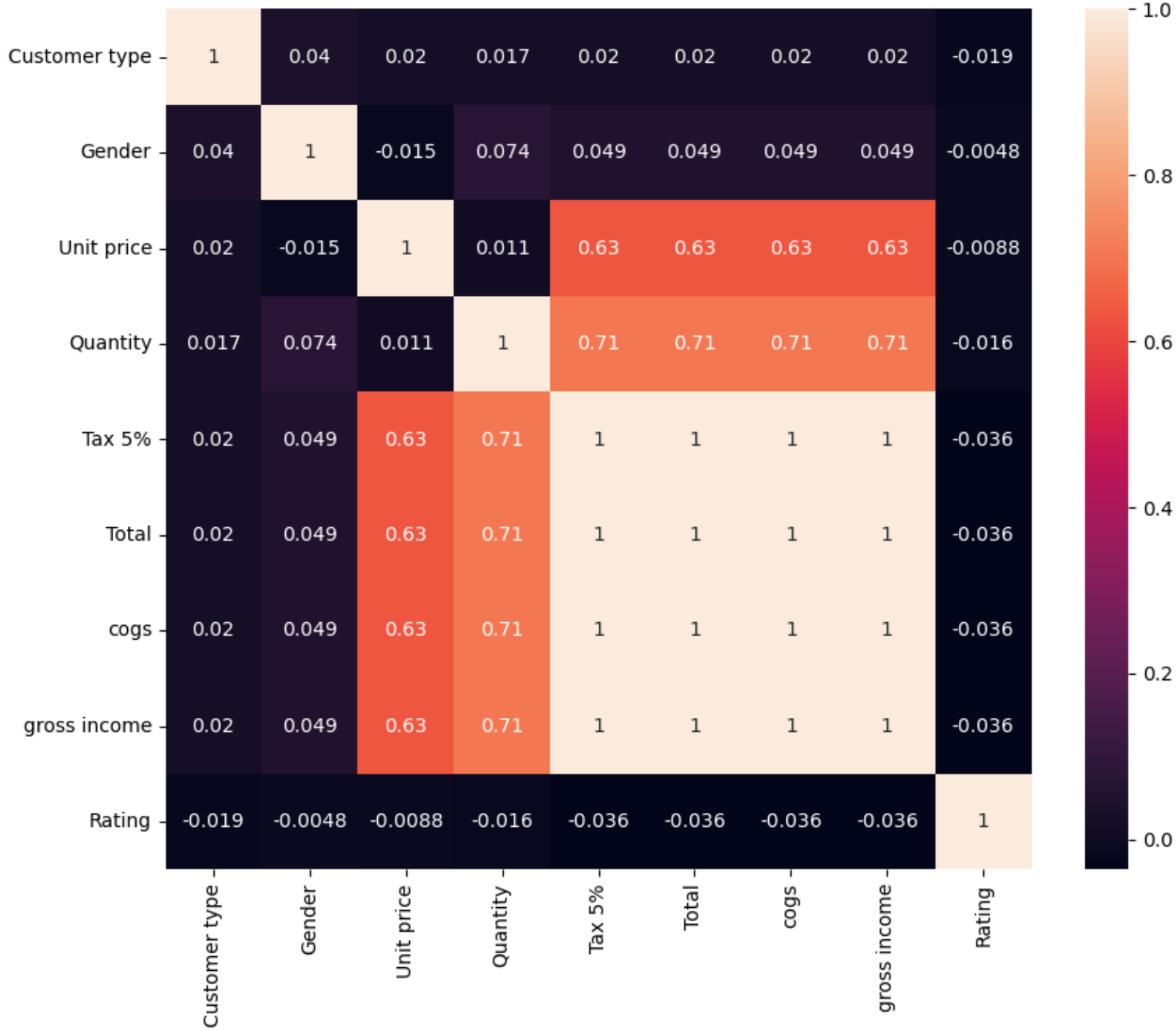
```
# converting the health column to string instead of integer in existing column:
df = df.replace({
    'Gender': {
        'Female': 1,
        'Male': 0
    }
})
```

```
# converting the health column to string instead of integer in existing column:
df = df.replace({
    'Customer type': {
        'Member': 1,
        'Normal': 0
    }
})
```

```
print(df['Payment'].value_counts() )
```

```
Ewallet        345
Cash           344
Credit card    311
Name: Payment, dtype: int64
```

```
# converting the health column to string instead of integer in existing column:
df = df.replace({
    'Payment': {
        'Ewallet': 1,
        'Cash': 2,
        'Credit card': 3
    }
})
```


Customer type
Gender
Unit price
Quantity
Tax 5%
Total
cogs
gross income
Rating
1 0.04 0.02 0.017 0.02 0.02 0.02 0.02 -0.019
0.04 1 -0.015 0.074 0.049 0.049 0.049 0.049 -0.0048
0.02 -0.015 1 0.011 0.63 0.63 0.63 0.63 -0.0088
0.017 0.074 0.011 1 0.71 0.71 0.71 0.71 -0.016
0.02 0.049 0.63 0.71 1 1 1 1 -0.036
0.02 0.049 0.63 0.71 1 1 1 1 -0.036
0.02 0.049 0.63 0.71 1 1 1 1 -0.036
0.02 0.049 0.63 0.71 1 1 1 1 -0.036
-0.019 -0.0048 -0.0088 -0.016 -0.036 -0.036 -0.036 -0.036 1
1.0
0.8
0.6
0.4
0.2
0.0