```python
# Filename:Lab2.py
# Date: 07/20/2022
# Qiang Liu
# NCSU-GTI Summer 2022 Data Science
#------------------import library--------------------
import sys
import math
import os,io,csv

#------------------Data Science library---------------
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sn

#------------------global variables-------------------
dash = "-" * 80
print(dash)

def main():
    print("Start Program.....")
    #a. Import data(VandelaySales2015.csv) into a Pandas dataframe
    FILENAME = "VandelaySales2015.csv"
    df = pd.read_csv(FILENAME,encoding="utf-8")
    print('VandelaySales2015 data are following:\n',df)
    print(dash)
    # check if file has missing values
    print("\n Any null values?")
    print(df.isnull().any())
    print(dash)
    print(df.isnull().sum())
    print(dash)
    print(df.info())
    print(dash)
    #a. print head for top 10 record only for the dataframe call it "df"
    print('A. Top 10 record of VandelaySales2015 are following:\n',df.head(10))
    print(dash)


    #b. Find the sum, mean, max, min value of total_product_price (COLUMN I) of VandelaySales2015.csv file.
    print('B. The sum, mean, max, min value of total_product_price are following:\n',df['total_product_price'].describe())
    print(dash)

    #c. Show correlation
    print('C. The correlations between variables are following:\n',df.corr())
    print(dash)
    # show corr in heatmap
    dataplot = sn.heatmap(df.corr(),
                          cmap='YlGnBu',
                          annot=True)
    plt.show()

    #d. Show distribution analysis Histogram for item_product_price(COLUMN H) of VandelaySales2015.csv file.
    plt.hist(df['item_product_price'],
             bins=15,
             color='steelblue',
             edgecolor='k',
             label='Histogram')
    plt.title("D. Distribution analysis Histogram for item_product_price")
    plt.ylabel("Frequency for item_product_price")
    plt.xlabel("Bins")
    plt.show()

#starting point... launch
# __name__ is predefined class attribute
if __name__ == '__main__':
    main()
```

```
C:\Users\Liuqiang\AppData\Local\Programs\Python\Python310\python.exe C:/Users/Liuqiang/PycharmProjects/GTISummer2022/Lab2.py
--------------------------------------------------------------------------------
Start Program.....
VandelaySales2015 data are following:
        order_id  customer_id  ... category_name product_color
0          59929        11914  ...      T-Shirts          Blue
1          59929        11914  ...   Sweatshirts        Purple
2          59966        14644  ...   Sweatshirts        Yellow
3          59973        23186  ...         Jeans         Green
4          59973        23186  ...         Visors        Purple
...          ...          ...  ...            ...           ...
47296      31061        12976  ...      T-Shirts        Yellow
47297      31061        12976  ...          Hats         Green
47298      31033         1759  ...         Socks        Orange
47299      31033         1759  ...      T-Shirts        Yellow
47300      31033         1759  ...   Sweatshirts        Purple

[47301 rows x 13 columns]
--------------------------------------------------------------------------------

 Any null values?
order_id               False
customer_id            False
order_date             False
order_short_date       False
promo_code             False
referral_source        False
product_quantity       False
item_product_price     False
total_product_price    False
product_name           False
vendor_name            False
category_name          False
product_color          False
dtype: bool
--------------------------------------------------------------------------------
order_id               0
customer_id            0
order_date             0
order_short_date       0
promo_code             0
referral_source        0
product_quantity       0
item_product_price     0
total_product_price    0
product_name           0
vendor_name            0
category_name          0
product_color          0
dtype: int64
--------------------------------------------------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 47301 entries, 0 to 47300
Data columns (total 13 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   order_id             47301 non-null  int64
 1   customer_id          47301 non-null  int64
 2   order_date           47301 non-null  object
 3   order_short_date     47301 non-null  object
 4   promo_code           47301 non-null  object
 5   referral_source      47301 non-null  object
 6   product_quantity     47301 non-null  int64
 7   item_product_price   47301 non-null  float64
 8   total_product_price  47301 non-null  float64
 9   product_name         47301 non-null  object
 10  vendor_name          47301 non-null  object
 11  category_name        47301 non-null  object
 12  product_color        47301 non-null  object
dtypes: float64(2), int64(3), object(8)
memory usage: 4.7+ MB
None
--------------------------------------------------------------------------------
```

```
--------------------------------------------------------------------------------
A. Top 10 record of VandelaySales2015 are following:
    order_id  customer_id  ... category_name product_color
0      59929        11914  ...      T-Shirts          Blue
1      59929        11914  ...   Sweatshirts        Purple
2      59966        14644  ...   Sweatshirts        Yellow
3      59973        23186  ...         Jeans         Green
4      59973        23186  ...        Visors        Purple
5      59935         7437  ...      T-Shirts        Orange
6      59935         7437  ...      T-Shirts          Blue
7      59980        23193  ...      T-Shirts          Blue
8      59916          360  ...      T-Shirts           Red
9      59916          360  ...        Boxers        Orange

[10 rows x 13 columns]
--------------------------------------------------------------------------------
B. The sum, mean, max, min value of total_product_price are following:
 count    47301.000000
mean        91.612787
std        100.521693
min         13.530000
25%         25.720000
50%         57.390000
75%        124.320000
max        893.280000
Name: total_product_price, dtype: float64
--------------------------------------------------------------------------------
C. The correlations between variables are following:
                     order_id  ...  total_product_price
order_id             1.000000  ...            -0.013055
customer_id          0.218983  ...             0.360153
product_quantity    -0.010340  ...             0.792602
item_product_price  -0.003488  ...             0.465881
total_product_price -0.013055  ...             1.000000

[5 rows x 5 columns]
--------------------------------------------------------------------------------
```

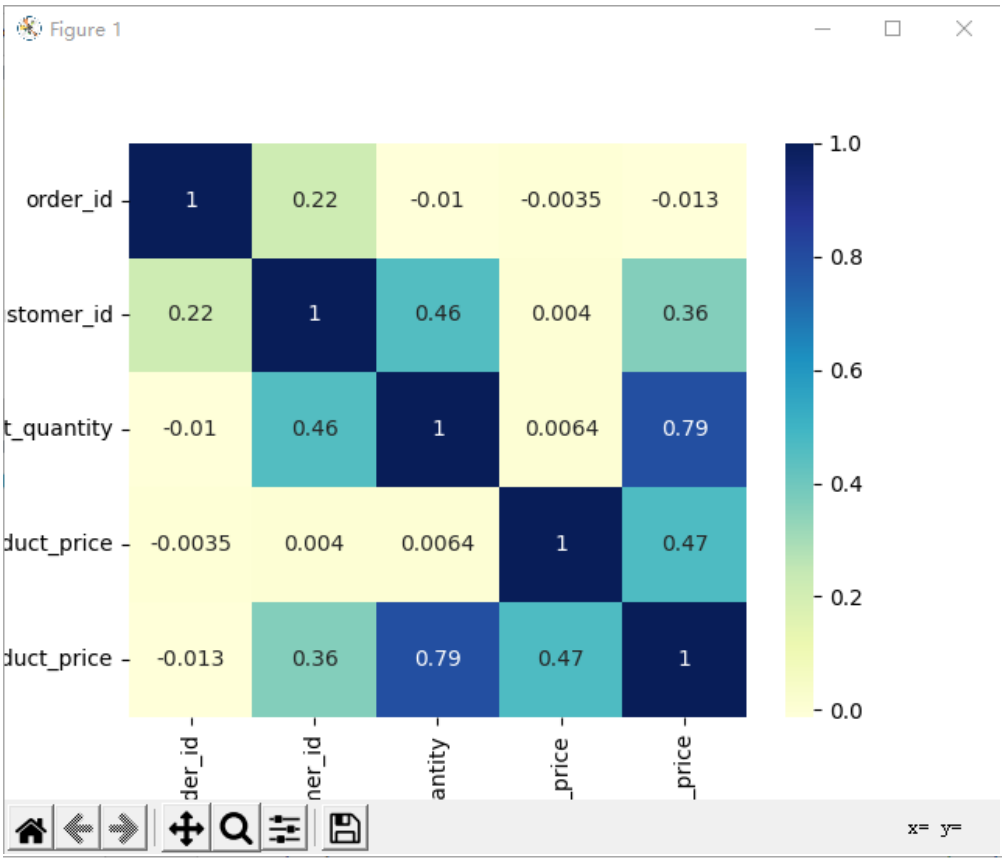
Figure 1
order_id
stomer_id
t_quantity
duct_price
duct_price
1
0.22
-0.01
-0.0035
-0.013
0.22
1
0.46
0.004
0.36
-0.01
0.46
1
0.0064
0.79
-0.0035
0.004
0.0064
1
0.47
-0.013
0.36
0.79
0.47
1
1.0
0.8
0.6
0.4
0.2
0.0
der_id
ner_id
antity
_price
_price
x= y=

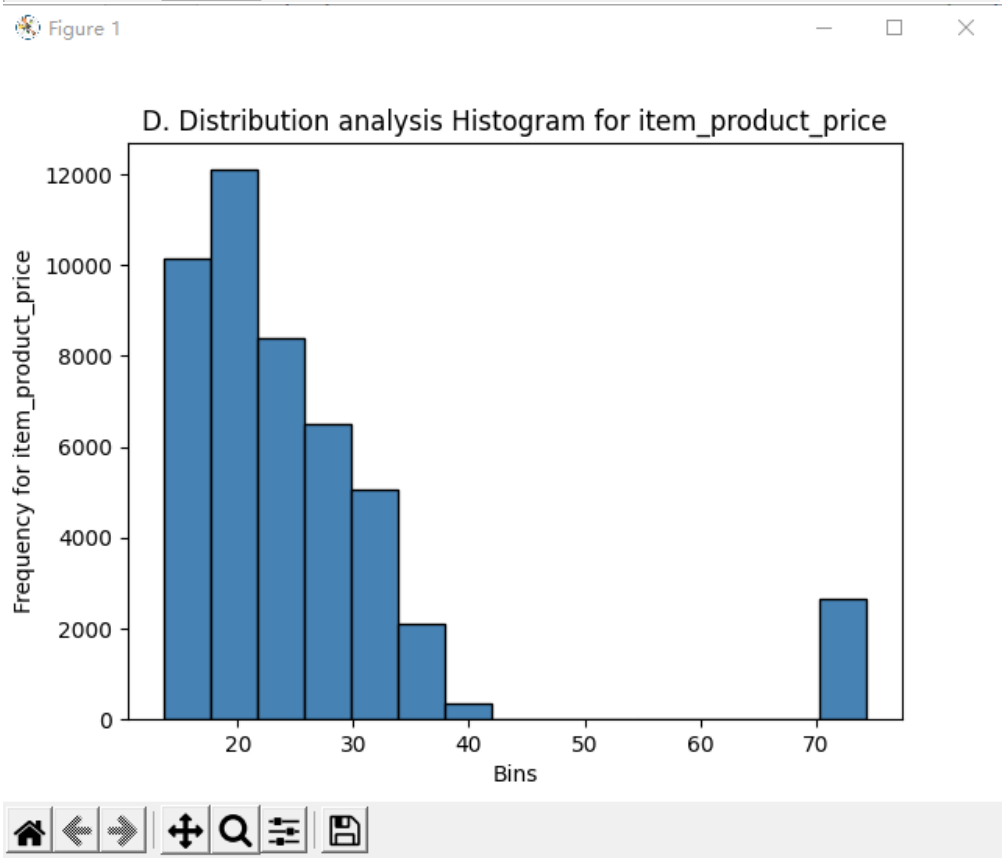
Figure 1
D. Distribution analysis Histogram for item_product_price
Frequency for item_product_price
12000
10000
8000
6000
4000
2000
0
20
30
40
50
60
70
Bins