bigdata

February 14, 2025

to extract the dataset using pandas and show 5 heads

[1]:
```
import pandas as pd
from sqlalchemy import create_engine


# Load the CSV file (Change the file path to match your actual file)
file_path = r"C:\Users\Postlab\Desktop\Aliexpress\data\raw_csv.csv"

# Read the CSV file into a pandas DataFrame
df = pd.read_csv(file_path)

# Display first 5 rows
df.head()
```

[1]:

	id	storeId	storeName	title	rating	lunchTime	category	postCategory	sold	price	discount	shippingCost	imageUrl
0	1005002508947205	900250272	IDEACH Official Store	Wireless Handheld Vacuum Cleaner 10kPa 150W Po…	4.8	2021-04-19 00:00:00	NaN	608	1487 sold	464.16	76	0.0	//ae01.alicdn.com/kf/S29307438fb224285b2981d71…
1	1005004546160959	5379102	Factory Direct Collected Store	Portable Water Drop Humidifier USB Desktop Ind…	4.8	2022-07-19 00:00:00	NaN	625	5348 sold	22.28	54		
2	1005004967489874	911794975	HOMDD Specialty Store	Portable Desktop Instant Automatic Multi-Speed…	0.0	2022-11-17 00:00:00	NaN	610	1 sold	251.76	25		
3	1005003601452827	912652146	XINGHUO ONLINE Store	Portable 420ml Air Humidifier Aroma Oil Humidi…	4.7	2021-11-24 00:00:00	NaN	625	2979 sold	8.95	84		
4	33022569739	4991203	ABIR Official Store	ABIR X6 Robot Vacuum Cleaner Visual Navigatio…	4.7	2019-05-24 00:00:00	NaN	608	2103 sold	775.35	59		

```
1           0.0  //ae01.alicdn.com/kf/S8d4807ce10bd4152850ec872…
2           0.0  //ae01.alicdn.com/kf/Sdae1fa83cf1a482397b6e061…
3           0.0  //ae01.alicdn.com/kf/Se200f7a404974c87b5941587…
4           0.0  //ae01.alicdn.com/kf/Scfe6fa94f0644aaabff79c0f…

                              storeUrl    category_name  category_id     type
0  //www.aliexpress.com/store/900250272  home-appliances            6  natural
1    //www.aliexpress.com/store/5379102  home-appliances            6  natural
2  //www.aliexpress.com/store/911794975  home-appliances            6  natural
3  //www.aliexpress.com/store/912652146  home-appliances            6  natural
4    //www.aliexpress.com/store/4991203  home-appliances            6       ad
```

to show the shape of dataset like row and column

[28]:
```
df.shape
```

[28]:
```
(711424, 18)
```

to show the column

[29]:
```
df.columns
```

[29]:
```
Index(['id', 'storeId', 'storeName', 'title', 'rating', 'lunchTime',
       'category', 'postCategory', 'sold', 'price', 'discount', 'shippingCost',
       'imageUrl', 'storeUrl', 'category_name', 'category_id', 'type',
       'totla-revenue'],
      dtype='object')
```

to delete the duplicate of datas in dataset

[30]:
```
df = df.drop_duplicates()
df
```

[30]:
```
                     id     storeId  \
0       1005002508947205   900250272
1       1005004546160959     5379102
2       1005004967489874   911794975
3       1005003601452827   912652146
4            33022569739     4991203
…                    …           …
864207  1005004659411453  1102181212
864208  1005004840722028     2985025
864209  1005002636959670   910326412
864268  1005004954707268      117712
864269  1005004476618259   911459011

                                          storeName  \
0                                IDEACH Official Store
```

```
1                        Factory Direct Collected Store
2                                HOMDD Specialty Store
3                                 XINGHUO ONLINE Store
4                                  ABIR Official Store
…                                                  …
864207                          Xsplendor Battery Store
864208                               Wildcard 365 Store
864209                     ROGRAPO Global Online Store
864268  Tianchang Meide Electronic Technology Co.  Ltd.
864269                     HengChiTai components Store

                                                    title  rating  \
0       Wireless Handheld Vacuum Cleaner 10kPa 150W Po…     4.8
1       Portable Water Drop Humidifier USB Desktop Ind…     4.8
2       Portable Desktop Instant Automatic Multi-Speed…     0.0
3       Portable 420ml Air Humidifier Aroma Oil Humidi…     4.7
4       ABIR X6 Robot Vacuum Cleaner  Visual Navigatio…     4.7
…                                                     …     …
864207  1800mAh ICOM BP-264 BP264 Battery for IC-F3101…     5.0
864208  Remote Control Suitable for Samsung SMART TV B…     0.0
864209  New 6500mAh Li-ion Replacement Battery For Xia…     5.0
864268  New Replacement CD Remote Control For DENON RC…     0.0
864269  NEW FBS-232P0-9F PLC Programming Cable RS232 P…     0.0

                 lunchTime  category  postCategory       sold   price  \
0       2021-04-19 00:00:00       NaN           608  1487 sold  464.16
1       2022-07-19 00:00:00       NaN           625  5348 sold   22.28
2       2022-11-17 00:00:00       NaN           610     1 sold  251.76
3       2021-11-24 00:00:00       NaN           625  2979 sold    8.95
4       2019-05-24 00:00:00       NaN           608  2103 sold  775.35
…                       …       …           …         …     …
864207  2022-08-18 00:00:00       NaN         52805     1 sold   68.07
864208  2022-10-13 00:00:00       NaN           623     2 sold    4.48
864209  2021-05-14 00:00:00       NaN         52801    18 sold   84.87
864268  2022-11-14 00:00:00       NaN           623          0   14.94
864269  2022-06-30 00:00:00       NaN     100000356          0   42.72

        discount  shippingCost  \
0             76          0.00
1             54          0.00
2             25          0.00
3             84          0.00
4             59          0.00
…          …           …
864207        20         71.18
864208        51           NaN
864209        22         81.66
```

```
864268        10          9.03
864269        12         32.19

                                                    imageUrl  \
0       //ae01.alicdn.com/kf/S29307438fb224285b2981d71…
1       //ae01.alicdn.com/kf/S8d4807ce10bd4152850ec872…
2       //ae01.alicdn.com/kf/Sdae1fa83cf1a482397b6e061…
3       //ae01.alicdn.com/kf/Se200f7a404974c87b5941587…
4       //ae01.alicdn.com/kf/Scfe6fa94f0644aaabff79c0f…
…                                                     …
864207  //ae01.alicdn.com/kf/S30eacf3e0f834c18a69003a5…
864208  //ae01.alicdn.com/kf/Sb70f27a7c4ce44e5bb7150e0…
864209  //ae01.alicdn.com/kf/S7a6ecce2e99b4a0186895d82…
864268  //ae01.alicdn.com/kf/S5395dd9f321e48d282c128f3…
864269  //ae01.alicdn.com/kf/Sa125d5752c0040659b9f0fe5…

                                      storeUrl         category_name  \
0        //www.aliexpress.com/store/900250272       home-appliances
1          //www.aliexpress.com/store/5379102       home-appliances
2        //www.aliexpress.com/store/911794975       home-appliances
3        //www.aliexpress.com/store/912652146       home-appliances
4          //www.aliexpress.com/store/4991203       home-appliances
…                                          …                    …
864207  //www.aliexpress.com/store/1102181212  consumer-electronics
864208     //www.aliexpress.com/store/2985025  consumer-electronics
864209   //www.aliexpress.com/store/910326412  consumer-electronics
864268      //www.aliexpress.com/store/117712  consumer-electronics
864269   //www.aliexpress.com/store/911459011  consumer-electronics

        category_id     type  totla-revenue
0                 6  natural       35276.16
1                 6  natural        1203.12
2                 6  natural        6294.00
3                 6  natural         751.80
4                 6       ad       45745.65
…               …       …             …
864207           44  natural        1361.40
864208           44  natural         228.48
864209           44  natural        1867.14
864268           44  natural         149.40
864269           44  natural         512.64

[711424 rows x 18 columns]
```

[31]:
```
df.describe()
```

[31]:
```
                 id       storeId         rating  category  postCategory  \
count  7.114240e+05  7.114240e+05  711424.000000       0.0  7.114240e+05
mean   9.056908e+14  5.203403e+08       2.246521       NaN  1.228978e+08
std    2.995966e+14  4.956373e+08       2.402573       NaN  8.925647e+07
min    1.008900e+04  1.000300e+04       0.000000       NaN  1.450000e+02
25%    1.005003e+15  4.247012e+06       0.000000       NaN  3.806100e+05
50%    1.005004e+15  9.104400e+08       0.000000       NaN  2.000003e+08
75%    1.005005e+15  9.125403e+08       5.000000       NaN  2.000039e+08
max    1.010000e+15  1.102528e+09       5.000000       NaN  2.018986e+08

                price       discount  shippingCost   category_id  totla-revenue
count  711424.000000  711424.000000  6.512240e+05  7.114240e+05   7.114240e+05
mean      227.596012      30.649160  9.747764e+01  1.389763e+08   4.323534e+03
std      2467.899156      20.023911  6.799883e+03  8.227431e+07   2.040220e+04
min         0.010000       0.000000  0.000000e+00  5.000000e+00   0.000000e+00
25%        14.060000      15.000000  0.000000e+00  1.000016e+08   1.656000e+02
50%        38.570000      32.000000  1.220000e+01  2.000021e+08   7.740800e+02
75%       111.460000      47.000000  2.255000e+01  2.002153e+08   2.679500e+03
max    976428.200000      99.000000  2.896137e+06  2.060891e+08   2.491002e+06
```

[45]:
```
df['totla-revenue']=df['price'] + df['shippingCost']
df
```

[45]:
```
                       id     storeId  \
0       1005002508947205   900250272
1       1005004546160959     5379102
2       1005004967489874   911794975
3       1005003601452827   912652146
4            33022569739     4991203
…                    …           …
864207  1005004659411453  1102181212
864208  1005004840722028     2985025
864209  1005002636959670   910326412
864268  1005004954707268      117712
864269  1005004476618259   911459011

                                          storeName  \
0                                IDEACH Official Store
1                        Factory Direct Collected Store
2                                HOMDD Specialty Store
3                                 XINGHUO ONLINE Store
4                                  ABIR Official Store
…                                                  …
864207                         Xsplendor Battery Store
864208                             Wildcard 365 Store
864209                       ROGRAPO Global Online Store
864268  Tianchang Meide Electronic Technology Co.  Ltd.
```

```
864269                    HengChiTai components Store

                                                    title  rating  \
0       Wireless Handheld Vacuum Cleaner 10kPa 150W Po…     4.8
1       Portable Water Drop Humidifier USB Desktop Ind…     4.8
2       Portable Desktop Instant Automatic Multi-Speed…     0.0
3       Portable 420ml Air Humidifier Aroma Oil Humidi…     4.7
4       ABIR X6 Robot Vacuum Cleaner  Visual Navigatio…     4.7
…                                                     …     …
864207  1800mAh ICOM BP-264 BP264 Battery for IC-F3101…     5.0
864208  Remote Control Suitable for Samsung SMART TV B…     0.0
864209  New 6500mAh Li-ion Replacement Battery For Xia…     5.0
864268  New Replacement CD Remote Control For DENON RC…     0.0
864269  NEW FBS-232P0-9F PLC Programming Cable RS232 P…     0.0

                lunchTime  category  postCategory       sold   price  \
0       2021-04-19 00:00:00       NaN           608  1487 sold  464.16
1       2022-07-19 00:00:00       NaN           625  5348 sold   22.28
2       2022-11-17 00:00:00       NaN           610     1 sold  251.76
3       2021-11-24 00:00:00       NaN           625  2979 sold    8.95
4       2019-05-24 00:00:00       NaN           608  2103 sold  775.35
…                       …       …           …         …     …
864207  2022-08-18 00:00:00       NaN         52805     1 sold   68.07
864208  2022-10-13 00:00:00       NaN           623     2 sold    4.48
864209  2021-05-14 00:00:00       NaN         52801    18 sold   84.87
864268  2022-11-14 00:00:00       NaN           623          0   14.94
864269  2022-06-30 00:00:00       NaN     100000356          0   42.72

        discount  shippingCost  \
0             76          0.00
1             54          0.00
2             25          0.00
3             84          0.00
4             59          0.00
…           …           …
864207        20         71.18
864208        51           NaN
864209        22         81.66
864268        10          9.03
864269        12         32.19

                                                 imageUrl  \
0       //ae01.alicdn.com/kf/S29307438fb224285b2981d71…
1       //ae01.alicdn.com/kf/S8d4807ce10bd4152850ec872…
2       //ae01.alicdn.com/kf/Sdae1fa83cf1a482397b6e061…
3       //ae01.alicdn.com/kf/Se200f7a404974c87b5941587…
4       //ae01.alicdn.com/kf/Scfe6fa94f0644aaabff79c0f…
```

```
…                                                      …
864207  //ae01.alicdn.com/kf/S30eacf3e0f834c18a69003a5…
864208  //ae01.alicdn.com/kf/Sb70f27a7c4ce44e5bb7150e0…
864209  //ae01.alicdn.com/kf/S7a6ecce2e99b4a0186895d82…
864268  //ae01.alicdn.com/kf/S5395dd9f321e48d282c128f3…
864269  //ae01.alicdn.com/kf/Sa125d5752c0040659b9f0fe5…

                                      storeUrl         category_name  \
0        //www.aliexpress.com/store/900250272       home-appliances
1          //www.aliexpress.com/store/5379102       home-appliances
2        //www.aliexpress.com/store/911794975       home-appliances
3        //www.aliexpress.com/store/912652146       home-appliances
4          //www.aliexpress.com/store/4991203       home-appliances
…                                          …                   …
864207  //www.aliexpress.com/store/1102181212  consumer-electronics
864208     //www.aliexpress.com/store/2985025  consumer-electronics
864209   //www.aliexpress.com/store/910326412  consumer-electronics
864268      //www.aliexpress.com/store/117712  consumer-electronics
864269   //www.aliexpress.com/store/911459011  consumer-electronics

        category_id     type  totla-revenue
0                 6  natural         464.16
1                 6  natural          22.28
2                 6  natural         251.76
3                 6  natural           8.95
4                 6       ad         775.35
…               …       …             …
864207           44  natural         139.25
864208           44  natural            NaN
864209           44  natural         166.53
864268           44  natural          23.97
864269           44  natural          74.91

[711424 rows x 18 columns]
```

to show the null value if have in dataset

[46]:
```
df.isnull().sum()
```

[46]:
```
id                   0
storeId              0
storeName            0
title                0
rating               0
lunchTime            0
category        711424
postCategory         0
```

```
sold                 0
price                0
discount             0
shippingCost     60200
imageUrl             0
storeUrl             0
category_name        0
category_id          0
type                 0
totla-revenue    60200
dtype: int64
```

[47]:
```
df.shape
```

[47]:
```
(711424, 18)
```

to convert the datatype of datetime and float

to clean the dataset if have the null value of

[48]:
```
df.to_csv("cleaned_bigdata.csv", index=False)
df
```

[48]:
```
                      id     storeId  \
0       1005002508947205   900250272
1       1005004546160959     5379102
2       1005004967489874   911794975
3       1005003601452827   912652146
4            33022569739     4991203
…                    …           …
864207  1005004659411453  1102181212
864208  1005004840722028     2985025
864209  1005002636959670   910326412
864268  1005004954707268      117712
864269  1005004476618259   911459011

                                              storeName  \
0                                  IDEACH Official Store
1                         Factory Direct Collected Store
2                                  HOMDD Specialty Store
3                                  XINGHUO ONLINE Store
4                                   ABIR Official Store
…                                                     …
864207                         Xsplendor Battery Store
864208                              Wildcard 365 Store
864209                       ROGRAPO Global Online Store
864268  Tianchang Meide Electronic Technology Co.  Ltd.
864269                       HengChiTai components Store
```

```
                                                    title  rating  \
0       Wireless Handheld Vacuum Cleaner 10kPa 150W Po…     4.8
1       Portable Water Drop Humidifier USB Desktop Ind…     4.8
2       Portable Desktop Instant Automatic Multi-Speed…     0.0
3       Portable 420ml Air Humidifier Aroma Oil Humidi…     4.7
4       ABIR X6 Robot Vacuum Cleaner  Visual Navigatio…     4.7
…                                                     …     …
864207  1800mAh ICOM BP-264 BP264 Battery for IC-F3101…     5.0
864208  Remote Control Suitable for Samsung SMART TV B…     0.0
864209  New 6500mAh Li-ion Replacement Battery For Xia…     5.0
864268  New Replacement CD Remote Control For DENON RC…     0.0
864269  NEW FBS-232P0-9F PLC Programming Cable RS232 P…     0.0

                  lunchTime  category  postCategory       sold   price  \
0       2021-04-19 00:00:00       NaN           608  1487 sold  464.16
1       2022-07-19 00:00:00       NaN           625  5348 sold   22.28
2       2022-11-17 00:00:00       NaN           610     1 sold  251.76
3       2021-11-24 00:00:00       NaN           625  2979 sold    8.95
4       2019-05-24 00:00:00       NaN           608  2103 sold  775.35
…                       …       …             …         …     …
864207  2022-08-18 00:00:00       NaN         52805     1 sold   68.07
864208  2022-10-13 00:00:00       NaN           623     2 sold    4.48
864209  2021-05-14 00:00:00       NaN         52801    18 sold   84.87
864268  2022-11-14 00:00:00       NaN           623          0   14.94
864269  2022-06-30 00:00:00       NaN     100000356          0   42.72

        discount  shippingCost  \
0             76          0.00
1             54          0.00
2             25          0.00
3             84          0.00
4             59          0.00
…          …            …
864207        20         71.18
864208        51           NaN
864209        22         81.66
864268        10          9.03
864269        12         32.19

                                                  imageUrl  \
0       //ae01.alicdn.com/kf/S29307438fb224285b2981d71…
1       //ae01.alicdn.com/kf/S8d4807ce10bd4152850ec872…
2       //ae01.alicdn.com/kf/Sdae1fa83cf1a482397b6e061…
3       //ae01.alicdn.com/kf/Se200f7a404974c87b5941587…
4       //ae01.alicdn.com/kf/Scfe6fa94f0644aaabff79c0f…
…                                                     …
```

```
864207  //ae01.alicdn.com/kf/S30eacf3e0f834c18a69003a5…
864208  //ae01.alicdn.com/kf/Sb70f27a7c4ce44e5bb7150e0…
864209  //ae01.alicdn.com/kf/S7a6ecce2e99b4a0186895d82…
864268  //ae01.alicdn.com/kf/S5395dd9f321e48d282c128f3…
864269  //ae01.alicdn.com/kf/Sa125d5752c0040659b9f0fe5…

                                    storeUrl         category_name  \
0         //www.aliexpress.com/store/900250272       home-appliances
1           //www.aliexpress.com/store/5379102       home-appliances
2         //www.aliexpress.com/store/911794975       home-appliances
3         //www.aliexpress.com/store/912652146       home-appliances
4           //www.aliexpress.com/store/4991203       home-appliances
…                                         …                 …
864207  //www.aliexpress.com/store/1102181212  consumer-electronics
864208     //www.aliexpress.com/store/2985025  consumer-electronics
864209   //www.aliexpress.com/store/910326412  consumer-electronics
864268      //www.aliexpress.com/store/117712  consumer-electronics
864269   //www.aliexpress.com/store/911459011  consumer-electronics

        category_id     type  totla-revenue
0                 6  natural         464.16
1                 6  natural          22.28
2                 6  natural         251.76
3                 6  natural           8.95
4                 6       ad         775.35
…               …       …             …
864207           44  natural         139.25
864208           44  natural            NaN
864209           44  natural         166.53
864268           44  natural          23.97
864269           44  natural          74.91

[711424 rows x 18 columns]
```

ok let see the shape of dataset

the dataset loss above 400 thousend of dataset because of clean data

[49]:
```
df.shape
```

[49]:
```
(711424, 18)
```

to change the name and prepare for load to postgres

load the dataset to postgres

[50]:
```
# Database connection
engine = create_engine("postgresql://postgres:1221@localhost:5432/
 ↪big_assignment")
```

```
# Store cleaned DataFrame into PostgreSQL
df.to_sql("products", engine, if_exists="replace", index=False)

print("Cleaned data successfully stored in PostgreSQL!")
```

```
Cleaned data successfully stored in PostgreSQL!
```

read a dataset from database and show the head of 5

[51]:
```
query = "SELECT * FROM products LIMIT 5;"
df_sql = pd.read_sql(query, engine)
df_sql.head()
```

[51]:

	id	storeId	storeName	title	rating	lunchTime	category	postCategory	sold	price	discount	shippingCost	imageUrl	storeUrl	category_name	category_id
0	1005002508947205	900250272	IDEACH Official Store	Wireless Handheld Vacuum Cleaner 10kPa 150W Po…	4.8	2021-04-19 00:00:00	None	608	1487 sold	464.16	76	0.0	//ae01.alicdn.com/kf/S29307438fb224285b2981d71…	//www.aliexpress.com/store/900250272	home-appliances	6
1	1005004546160959	5379102	Factory Direct Collected Store	Portable Water Drop Humidifier USB Desktop Ind…	4.8	2022-07-19 00:00:00	None	625	5348 sold	22.28	54	0.0	//ae01.alicdn.com/kf/S8d4807ce10bd4152850ec872…	//www.aliexpress.com/store/5379102	home-appliances	6
2	1005004967489874	911794975	HOMDD Specialty Store	Portable Desktop Instant Automatic Multi-Speed…	0.0	2022-11-17 00:00:00	None	610	1 sold	251.76	25	0.0	//ae01.alicdn.com/kf/Sdae1fa83cf1a482397b6e061…	//www.aliexpress.com/store/911794975	home-appliances	6
3	1005003601452827	912652146	XINGHUO ONLINE Store	Portable 420ml Air Humidifier Aroma Oil Humidi…	4.7	2021-11-24 00:00:00	None	625	2979 sold	8.95	84	0.0	//ae01.alicdn.com/kf/Se200f7a404974c87b5941587…	//www.aliexpress.com/store/912652146	home-appliances	6
4	33022569739	4991203	ABIR Official Store	ABIR X6 Robot Vacuum Cleaner Visual Navigatio…	4.7	2019-05-24 00:00:00	None	608	2103 sold	775.35	59	0.0	//ae01.alicdn.com/kf/Scfe6fa94f0644aaabff79c0f…	//www.aliexpress.com/store/4991203	home-appliances	6

	type	totla-revenue
0	natural	464.16
1	natural	22.28
2	natural	251.76
3	natural	8.95
4	ad	775.35

finaly show the data shape after clean and load a data

[52]:
```
df.shape
```

[52]: (711424, 18)