UNIT:1 INTRODUCTION TO DATA ANALYTICS

1.Building an ETL pipeline for Real-world data

Dataset:

women_clothing_ecommerce_sales.csv

```
order_id,order_date,sku,color,size,unit_price,quantity,revenue
1,2022/6/1 16:05:00,708,Dark Blue,2XL,298,1,298
1,2022/6/1 16:05:00,89,Dark Blue,2XL,258,1,258
2,2022/6/10 17:03:00,bobo,Cream,One Size,228,1,228
3,2022/6/10 19:59:00,799,Dark Blue,XL,264,1,264
4,2022/6/11 11:42:00,799,Dark Blue,M,264,1,264
5,2022/6/11 14:46:00,799,Dark Blue,XL,264,1,264
6,2022/6/11 19:23:00,799,Dark Blue,XL,264,1,264
7,2022/6/11 19:54:00,708,Dark Blue,3XL,278,1,278
8,2022/6/11 21:09:00,708,Dark Blue,XL,278,1,278
9,2022/6/11 22:53:00,239,Black,,258,1,258
10,2022/6/12 17:32:00,708,Dark Blue,XL,278,1,278
11,2022/6/12 21:16:00,708,Dark Blue,3XL,278,1,278
12,2022/6/12 8:47:00,708,Dark Blue,3XL,278,1,278
13,2022/6/13 10:11:00,799,Dark Blue,XL,264,1,264
14,2022/6/14 0:07:00,799,Dark Blue,M,264,1,264
15,2022/6/14 14:38:00,bobo,Navy Blue,One Size,228,1,228
15,2022/6/14 14:38:00,bobo,Blue,One Size,228,1,228
15,2022/6/14 14:38:00,bobo,Light Gray,One Size,228,1,228
16,2022/6/14 22:31:00,708,Dark Blue,XL,278,1,278
17,2022/6/15 10:46:00,708,Dark Blue,2XL,278,1,278
18,2022/6/15 13:36:00,77,B,,258,1,258
18,2022/6/15 13:36:00,77,C,,258,1,258
18,2022/6/15 13:36:00,77,D,,258,1,258
18,2022/6/15 13:36:00,77,E,,258,1,258
18,2022/6/15 13:36:00,77,A,,288,1,288
19,2022/6/15 20:42:00,799,Dark Blue,XL,264,1,264
20,2022/6/15 21:22:00,799,Dark Blue,XL,264,1,264
21,2022/6/15 5:50:00,799,Dark Blue,XL,264,1,264
```

Program:

data.py

```python
import pandas as pd
file_path = "women_clothing_ecommerce_sales.csv"
df = pd.read_csv(file_path)
df['size'].fillna("Unknown", inplace=True)
df['order_date'] = pd.to_datetime(df['order_date'], errors='coerce')
df['total_sales'] = df['unit_price'] * df['quantity']
transformed_df = df[(df['order_date'].notna()) & (df['revenue'] > 0)]
output_path = "cleaned_women_clothing_sales.csv"
transformed_df.to_csv(output_path, index=False)

print(f"ETL process complete. Cleaned data saved at: {output_path}")
```

Output:

data.py | cleaned_women_clothing_sales.csv | women_clothing_ecommerce_sales.csv

cleaned_women_clothing_sales.csv

```csv
order_id,order_date,sku,color,size,unit_price,quantity,revenue,total_sales
1,2022-06-01 16:05:00,708,Dark Blue,2XL,298,1,298,298
1,2022-06-01 16:05:00,89,Dark Blue,2XL,258,1,258,258
2,2022-06-10 17:03:00,bobo,Cream,One Size,228,1,228,228
3,2022-06-10 19:59:00,799,Dark Blue,XL,264,1,264,264
4,2022-06-11 11:42:00,799,Dark Blue,M,264,1,264,264
5,2022-06-11 14:46:00,799,Dark Blue,XL,264,1,264,264
6,2022-06-11 19:23:00,799,Dark Blue,XL,264,1,264,264
7,2022-06-11 19:54:00,708,Dark Blue,3XL,278,1,278,278
8,2022-06-11 21:09:00,708,Dark Blue,XL,278,1,278,278
9,2022-06-11 22:53:00,239,Black,Unknown,258,1,258,258
10,2022-06-12 17:32:00,708,Dark Blue,XL,278,1,278,278
11,2022-06-12 21:16:00,708,Dark Blue,3XL,278,1,278,278
12,2022-06-12 08:47:00,708,Dark Blue,3XL,278,1,278,278
13,2022-06-13 10:11:00,799,Dark Blue,XL,264,1,264,264
14,2022-06-14 00:07:00,799,Dark Blue,M,264,1,264,264
15,2022-06-14 14:38:00,bobo,Navy Blue,One Size,228,1,228,228
15,2022-06-14 14:38:00,bobo,Blue,One Size,228,1,228,228
15,2022-06-14 14:38:00,bobo,Light Gray,One Size,228,1,228,228
16,2022-06-14 22:31:00,708,Dark Blue,XL,278,1,278,278
17,2022-06-15 10:46:00,708,Dark Blue,2XL,278,1,278,278
18,2022-06-15 13:36:00,77,B,Unknown,258,1,258,258
18,2022-06-15 13:36:00,77,C,Unknown,258,1,258,258
18,2022-06-15 13:36:00,77,D,Unknown,258,1,258,258
18,2022-06-15 13:36:00,77,E,Unknown,258,1,258,258
18,2022-06-15 13:36:00,77,A,Unknown,288,1,288,288
19,2022-06-15 20:42:00,799,Dark Blue,XL,264,1,264,264
20,2022-06-15 21:22:00,799,Dark Blue,XL,264,1,264,264
21,2022-06-15 05:50:00,799,Dark Blue,XL,264,1,264,264
```

UNIT 1 ,2.Role and Responsibility of a data analyst
case study:

Dataset:

stock_details_5_years.csv

```csv
Date,Open,High,Low,Close,Volume,Dividends,Stock Splits,Company
2018-11-29 00:00:00-05:00,43.829760572993,43.8633538041636,42.6395935832266,43.0835075378418,167080000,0,0,AAPL
2018-11-29 00:00:00-05:00,104.769074332185,105.519257086357,103.534594914971,104.636131286621,28123200,0,0,MSFT
2018-11-29 00:00:00-05:00,54.1764984130859,55.0074996948242,54.0999984741211,54.7290000915527,31004000,0,0,GOOGL
2018-11-29 00:00:00-05:00,83.7494964599609,84.4994964599609,82.6165008544922,83.6784973144531,132264000,0,0,AMZN
2018-11-29 00:00:00-05:00,39.6927840259795,40.0649038762231,38.7351954599368,39.0378532409668,54917200,0.04,0,NVDA
2018-11-29 00:00:00-05:00,135.919998168945,139.990005493164,135.660003662109,138.679992675781,24238700,0,0,META
2018-11-29 00:00:00-05:00,23.1333332061768,23.1666679382324,22.6366672515869,22.7446670532227,46210500,0,0,TSLA
2018-11-29 00:00:00-05:00,106.370277872685,108.796587770742,106.065833939132,107.938613891602,4688300,0,0,LLY
2018-11-29 00:00:00-05:00,135.9730590448,135.982718463696,134.059447051785,134.436370849609,8751500,0,0,V
2018-11-29 00:00:00-05:00,33.5207140725629,33.8916926825545,33.4500495547723,33.5030479431152,7056600,0,0,TSM
2018-11-29 00:00:00-05:00,260.294856793325,264.95443422482,260.183477437836,262.262634277344,4177800,0,0,UNH
2018-11-29 00:00:00-05:00,197.318661129289,200.990960046212,196.046824787412,198.590484619141,1792000,0,0,AVGO
2018-11-29 00:00:00-05:00,19.5540687513432,19.6990401048253,19.4602632795213,19.5796508789063,2778200,0,0,NVO
2018-11-29 00:00:00-05:00,94.7523159828525,95.2077375502605,94.2109747237482,94.5718688964844,11144300,0,0,JPM
2018-11-29 00:00:00-05:00,89.1271596636379,89.8712687567747,88.8699395361231,89.3751983642578,6241300,0,0,WMT
2018-11-29 00:00:00-05:00,60.5674791851337,61.5818393056204,60.5442504659789,61.2179069519043,10255200,0,0,XOM
2018-11-29 00:00:00-05:00,196.521321775761,196.521321775761,190.938379684799,191.27880859375,5447700,0,0,MA
2018-11-29 00:00:00-05:00,127.187295218457,128.567299105769,126.506029332323,127.388191223145,6900000,0,0,JNJ
2018-11-29 00:00:00-05:00,81.8204016834982,82.2080947295993,81.6441812383863,81.78515625,6126900,0,0,PG
2018-11-29 00:00:00-05:00,44.5746723969215,44.5746723969215,43.7086058414661,44.1324272155762,17052100,0,0,ORCL
2018-11-29 00:00:00-05:00,157.362337604526,157.780666164991,155.689023362667,156.347671508789,4329300,0,0,HD
2018-11-29 00:00:00-05:00,246.360000610352,252.25,244.309997558594,249.089996337891,3723500,0,0,ADBE
2018-11-29 00:00:00-05:00,163.608448089733,164.291217945268,162.119640536522,162.650680541992,1079700,0,0,ASML
2018-11-29 00:00:00-05:00,94.6614419348185,96.0899962176464,94.4929059733475,95.3837432861328,6656900,0,0,CVX
2018-11-29 00:00:00-05:00,214.395101307791,216.819915837262,213.795919063334,216.267547607422,1478700,0,0,COST
2018-11-29 00:00:00-05:00,121.519996643066,122.319999694824,121.379997253418,121.779998779297,125600,0,0,TM
2018-11-29 00:00:00-05:00,63.2170479818917,64.2735343409732,63.1515328490186,63.8067169189453,11429174,0,0,MRK
2018-11-29 00:00:00-05:00,42.4934146755008,42.622758774507,42.2261083619453,42.2347297668457,11564300,0.39,0,KO
```

Program:

data.py | cleaned_women_clothing_sales.csv | kar.py | stock_details_

kar.py

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
file_path = "stock_details_5_years.csv"
df = pd.read_csv(file_path)
df['Date'] = pd.to_datetime(df['Date'], errors='coerce')
cleaned_df = df.dropna(subset=['Date'])
print("Basic Statistics:\n", cleaned_df.describe())
print("Missing Values:\n", cleaned_df.isnull().sum())
plt.figure(figsize=(12, 6))
for company in ['AAPL', 'MSFT', 'GOOGL', 'AMZN', 'NVDA']:
    subset = cleaned_df[cleaned_df['Company'] == company]
    plt.plot(subset['Date'], subset['Close'], label=company)
plt.title("Stock Closing Prices Over Time")
plt.xlabel("Date")
plt.ylabel("Closing Price")
plt.legend()
plt.show()
plt.figure(figsize=(10, 5))
sns.histplot(cleaned_df['Volume'], bins=50, kde=True)
plt.title("Distribution of Trading Volume")
plt.show()
plt.figure(figsize=(8, 6))
sns.heatmap(cleaned_df.corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
print("EDA process complete.")
```

Output:

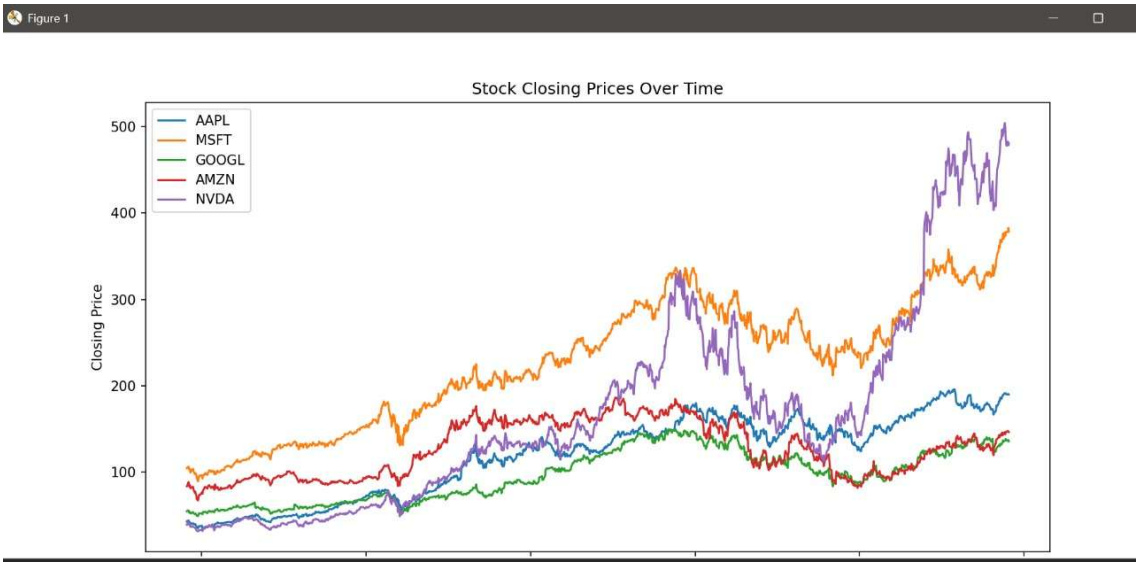
Figure 1
Stock Closing Prices Over Time
AAPL
MSFT
GOOGL
AMZN
NVDA
Closing Price
500
400
300
200
100

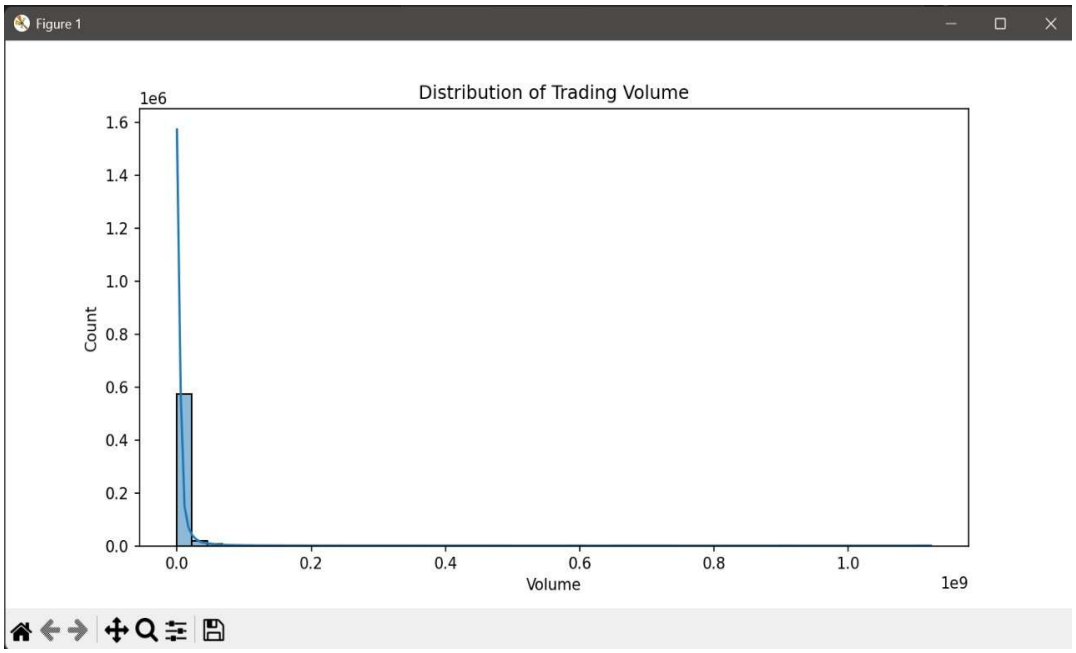
Figure 1
Distribution of Trading Volume
1e6
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Count
0.0
0.2
0.4
0.6
0.8
1.0
Volume
1e9