
# Debre Birhan University
# College of Computing Department of Software Engineering

## Course: Fundamentals of Big Data Analytics and BI

## Title: Building an End-to-End Data Pipeline

GITHUB LINK: https://github.com/abelfentaw/bigdata.git

POWER BI LINK : https://app.powerbi.com/groups/me/reports/d21618d6-42cf-462d-8de8-31e4570b0411/d960b31c84c08bb29d00?ctid=1695066a-e388-40d1-8ed5-5d0b28ba9f80&experience=power-bi&bookmarkGuid=a1ca2b68980d88796007

Submited By: **Abel Fentaw** **ID: 4166/13**

Submited to : **Derbew Felasman(MSc)**

Submited date : **02/06/17**

# Objective

This assignment focuses on building a complete data pipeline from **data extraction to visualization**. The key objectives include:

- Extracting and loading an **e-commerce dataset** (CSV format)
- Cleaning and transforming data using **Pandas in JupyterLab**
- Storing processed data in a **PostgreSQL database**
- Visualizing insights using **Power BI**

# Data Extraction

## Dataset Selection

A dataset with at least **1 million rows** related to e-commerce sales was selected. The dataset contains:

- **Transaction details** (order ID, price, quantity, total revenue)
- **Product information** (SKU, category, discounts)
- **Customer information** (customer ID, payment method, sales commission)
- **Time-based data** (order date, fiscal year, month, customer since date)

to extract the dataset using pandas and show 5 heads

[18]:
```
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\Downloads\archive\Pakistan Largest Ecommerce␣ ↪Dataset.csv"
#  Read the CSV file into a pandas
df = pd.read_csv(file_path)

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

[18]:

| | item_id | status | created_at | \ |
|---|---|---|---|---|
| 0 | 211131.0 | complete | 7/1/2016 | |
| 1 | 211133.0 | canceled | 7/1/2016 | |
| 2 | 211134.0 | canceled | 7/1/2016 | |
| 3 | 211135.0 | complete | 7/1/2016 | |
| 4 | 211136.0 | order_refunded | 7/1/2016 | |

| | sku | price | qty_ordered | \ |
|---|---|---|---|---|
| 0 | kreations_YI 06-L | 1950.0 | 1.0 | |
| 1 | kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo… | 240.0 | 1.0 | |
| 2 | Ego_UP0017-999-MR0 | 2450.0 | 1.0 | |

| | | | | | |
|---|---|---|---|---|---|
| 3 | | | kcc_krone deal | 360.0 | 1.0 |
| 4 | | | BK7010400AG | 555.0 | 2.0 |

| | grand_total | increment_id | category_name_1 | sales_commission_code | ... | \ |
|---|---|---|---|---|---|---|
| 0 | 1950.0 | 100147443 | Women's Fashion | \N | … | |
| 1 | 240.0 | 100147444 | Beauty & Grooming | \N | … | |
| 2 | 2450.0 | 100147445 | Women's Fashion | \N | .. | |
| 3 | 60.0 | 100147446 | Beauty & Grooming | R-FSD-52352 | .. | |
| 4 | 1110.0 | 100147447 | Soghaat | \N | .. | |

| | Month | Customer Since | M-Y | FY | Customer ID | Unnamed: 21 | Unnamed: 22 | \ |
|---|---|---|---|---|---|---|---|---|
| 0 | 7.0 | 2016-7 | 7-2016 | FY17 | 1.0 | NaN | NaN | |
| 1 | 7.0 | 2016-7 | 7-2016 | FY17 | 2.0 | NaN | NaN | |
| 2 | 7.0 | 2016-7 | 7-2016 | FY17 | 3.0 | NaN | NaN | |
| 3 | 7.0 | 2016-7 | 7-2016 | FY17 | 4.0 | NaN | NaN | |
| 4 | 7.0 | 2016-7 | 7-2016 | FY17 | 5.0 | NaN | NaN | |

| | Unnamed: 23 | Unnamed: 24 | Unnamed: 25 |
|---|---|---|---|
| 0 | NaN | NaN | NaN |
| 1 | NaN | NaN | NaN |
| 2 | NaN | NaN | NaN |
| 3 | NaN | NaN | NaN |
| 4 | NaN | NaN | NaN |

[5 rows x 26 columns]

to show the shape of dataset like row and column

[2]:
```
df.shape
```

[2]: (1048575, 26)

to show the column

[19]:
```
df.columns
```

[19]:
```
Index(['item_id', 'status', 'created_at', 'sku', 'price', 'qty_ordered',
       'grand_total', 'increment_id', 'category_name_1',
       'sales_commission_code', 'discount_amount', 'payment_method',
       'Working Date', 'BI Status', ' MV ', 'Year', 'Month', 'Customer Since',
       'M-Y', 'FY', 'Customer ID', 'Unnamed: 21', 'Unnamed: 22', 'Unnamed: 23',
       'Unnamed: 24', 'Unnamed: 25'],
      dtype='object')
```

s]

[21]:
```
df.describe()
```

[21]:

| | item_id | price | qty_ordered | grand_total | \ |
|---|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| count | 584524.000000 | 5.845240e+05 | 584524.000000 | 5.845240e+05 |
| mean | 565667.074218 | 6.348748e+03 | 1.296388 | 8.530619e+03 |
| std | 200121.173648 | 1.494927e+04 | 3.996061 | 6.132081e+04 |
| min | 211131.000000 | 0.000000e+00 | 1.000000 | -1.594000e+03 |
| 25% | 395000.750000 | 3.600000e+02 | 1.000000 | 9.450000e+02 |
| 50% | 568424.500000 | 8.990000e+02 | 1.000000 | 1.960400e+03 |
| 75% | 739106.250000 | 4.070000e+03 | 1.000000 | 6.999000e+03 |
| max | 905208.000000 | 1.012626e+06 | 1000.000000 | 1.788800e+07 |

| | discount_amount | Year | Month | Customer ID | \ |
|---|---|---|---|---|---|
| count | 584524.000000 | 584524.000000 | 584524.000000 | 584513.000000 | |
| mean | 499.492775 | 2017.044115 | 7.167654 | 45790.511965 | |
| std | 1506.943046 | 0.707355 | 3.486305 | 34414.962389 | |
| min | -599.500000 | 2016.000000 | 1.000000 | 1.000000 | |
| 25% | 0.000000 | 2017.000000 | 4.000000 | 13516.000000 | |
| 50% | 0.000000 | 2017.000000 | 7.000000 | 42856.000000 | |
| 75% | 160.500000 | 2018.000000 | 11.000000 | 73536.000000 | |
| max | 90300.000000 | 2018.000000 | 12.000000 | 115326.000000 | |

| | Unnamed: 21 | Unnamed: 22 | Unnamed: 23 | Unnamed: 24 | Unnamed: 25 |
|---|---|---|---|---|---|
| count | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| mean | NaN | NaN | NaN | NaN | NaN |
| std | NaN | NaN | NaN | NaN | NaN |
| min | NaN | NaN | NaN | NaN | NaN |
| 25% | NaN | NaN | NaN | NaN | NaN |
| 50% | NaN | NaN | NaN | NaN | NaN |
| 75% | NaN | NaN | NaN | NaN | NaN |
| max | NaN | NaN | NaN | NaN | NaN |

## 2. Data Cleaning & Transformation

### 2.1 Handling Missing Values

[6]:
```
df.isnull().sum()
```

[6]:
```
item_id                 1
status                 16
created_at              1
sku                    21
price                   1
qty_ordered             1
grand_total             1
increment_id            1
category_name_1       165
sales_commission_code 137179
discount_amount         1
payment_method          1
```

| | |
|---|---|
| Working Date | 1 |
| BI Status | 1 |
| MV | 1 |
| Year | 1 |
| Month | 1 |
| Customer Since | 12 |
| M-Y | 1 |

## 2.2 Removing Duplicates

[20]:
```
df = df.drop_duplicates()
df
```

[20]:

| | item_id | status | created_at | \ |
|---|---|---|---|---|
| 0 | 211131.0 | complete | 7/1/2016 | |
| 1 | 211133.0 | canceled | 7/1/2016 | |
| 2 | 211134.0 | canceled | 7/1/2016 | |
| 3 | 211135.0 | complete | 7/1/2016 | |
| 4 | 211136.0 | order_refunded | 7/1/2016 | |
| … | … | … | … | |
| 584520 | 905205.0 | processing | 8/28/2018 | |
| 584521 | 905206.0 | processing | 8/28/2018 | |
| 584522 | 905207.0 | processing | 8/28/2018 | |
| 584523 | 905208.0 | processing | 8/28/2018 | |
| 584524 | NaN | NaN | NaN | |

| | | sku | price | \ |
|---|---|---|---|---|
| 0 | | kreations_YI 06-L | 1950.0 | |
| 1 | kcc_Buy | 2 Frey Air Freshener & Get 1 Kasual Bo… | 240.0 | |
| 2 | | Ego_UP0017-999-MR0 | 2450.0 | |
| 3 | | kcc_krone deal | 360.0 | |
| 4 | | BK7010400AG | 555.0 | |
| … | | … | … | |
| 584520 | | MATHUA5AF70A7D1E50A | 35599.0 | |
| 584521 | | MATSAM5B6D7208C6D30 | 129999.0 | |
| 584522 | | MATSAM5B1509B4696EA | 87300.0 | |
| 584523 | | MATSAM5B10F91A9B6AB | 108640.0 | |
| 584524 | | NaN | NaN | |

| | qty_ordered | grand_total | increment_id | category_name_1 | \ |
|---|---|---|---|---|---|
| 0 | 1.0 | 1950.0 | 100147443 | Women's Fashion | |
| 1 | 1.0 | 240.0 | 100147444 | Beauty & Grooming | |
| 2 | 1.0 | 2450.0 | 100147445 | Women's Fashion | |
| 3 | 1.0 | 60.0 | 100147446 | Beauty & Grooming | |
| 4 | 2.0 | 1110.0 | 100147447 | Soghaat | |
| … | … | … | … | … | |

| | | | | | |
|---|---|---|---|---|---|
| 584520 | 1.0 | 35899.0 | 100562386 | Mobiles | & Tablets |
| 584521 | 2.0 | 652178.0 | 100562387 | Mobiles | & Tablets |
| 584522 | 2.0 | 652178.0 | 100562387 | Mobiles | & Tablets |
| 584523 | 2.0 | 652178.0 | 100562387 | Mobiles | & Tablets |
| 584524 | NaN | NaN | NaN | | NaN |

| | sales_commission_code | … | Month | Customer Since | M-Y | FY | \ |
|---|---|---|---|---|---|---|---|
| 0 | \N | … | 7.0 | 2016-7 | 7-2016 | FY17 | |
| 1 | \N | … | 7.0 | 2016-7 | 7-2016 | FY17 | |
| 2 | \N | … | 7.0 | 2016-7 | 7-2016 | FY17 | |
| 3 | R-FSD-52352 | … | 7.0 | 2016-7 | 7-2016 | FY17 | |
| 4 | \N | … | 7.0 | 2016-7 | 7-2016 | FY17 | |
| … | … | … | … | … | … | … | |
| 584520 | NaN | … | 8.0 | 2018-8 | 8-2018 | FY19 | |
| 584521 | NaN | … | 8.0 | 2018-7 | 8-2018 | FY19 | |
| 584522 | NaN | … | 8.0 | 2018-7 | 8-2018 | FY19 | |
| 584523 | NaN | … | 8.0 | 2018-7 | 8-2018 | FY19 | |
| 584524 | NaN | … | NaN | NaN | NaN | NaN | |

| | Customer ID | Unnamed: 21 | Unnamed: 22 | Unnamed: 23 | Unnamed: 24 | \ |
|---|---|---|---|---|---|---|
| 0 | 1.0 | NaN | NaN | NaN | NaN | |
| 1 | 2.0 | NaN | NaN | NaN | NaN | |
| 2 | 3.0 | NaN | NaN | NaN | NaN | |
| 3 | 4.0 | NaN | NaN | NaN | NaN | |
| 4 | 5.0 | NaN | NaN | NaN | NaN | |
| … | … | … | … | … | … | |
| 584520 | 115326.0 | NaN | NaN | NaN | NaN | |
| 584521 | 113474.0 | NaN | NaN | NaN | NaN | |
| 584522 | 113474.0 | NaN | NaN | NaN | NaN | |
| 584523 | 113474.0 | NaN | NaN | NaN | NaN | |
| 584524 | NaN | NaN | NaN | NaN | NaN | |

| | Unnamed: 25 |
|---|---|
| 0 | NaN |
| 1 | NaN |
| 2 | NaN |
| 3 | NaN |
| 4 | NaN |
| … | … |
| 584520 | NaN |
| 584521 | NaN |
| 584522 | NaN |
| 584523 | NaN |
| 584524 | NaN |

[584525 rows x 26 column

df.shape

[7]: (584525, 26)

## 2.3 Data Type Conversions

[25]
```
df["order_date"] = pd.to_datetime(df["created_at"]) # Convert to datetime
```
[26]
```
df["price"] = df["price"].astype(float) # Convert price to float
df
```

[25]:

| | item_id | status | created_at | \ |
|---|---|---|---|---|
| 0 | 211131.0 | complete | 2016-07-01 | |
| 1 | 211133.0 | canceled | 2016-07-01 | |
| 2 | 211134.0 | canceled | 2016-07-01 | |
| 3 | 211135.0 | complete | 2016-07-01 | |
| 4 | 211136.0 | order_refunded | 2016-07-01 | |
| … | … | … | … | |
| 584520 | 905205.0 | processing | 2018-08-28 | |
| 584521 | 905206.0 | processing | 2018-08-28 | |
| 584522 | 905207.0 | processing | 2018-08-28 | |
| 584523 | 905208.0 | processing | 2018-08-28 | |
| 584524 | NaN | NaN | NaT | |

| | sku | price | \ |
|---|---|---|---|
| 0 | kreations_YI 06-L | 1950.0 | |
| 1 | kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo… | 240.0 | |
| 2 | Ego_UP0017-999-MR0 | 2450.0 | |
| 3 | kcc_krone deal | 360.0 | |
| 4 | BK7010400AG | 555.0 | |
| … | … | … | |
| 584520 | MATHUA5AF70A7D1E50A | 35599.0 | |
| 584521 | MATSAM5B6D7208C6D30 | 129999.0 | |
| 584522 | MATSAM5B1509B4696EA | 87300.0 | |
| 584523 | MATSAM5B10F91A9B6AB | 108640.0 | |
| 584524 | NaN | NaN | |

| | qty_ordered | grand_total | increment_id | category_name_1 | \ |
|---|---|---|---|---|---|
| 0 | 1.0 | 1950.0 | 100147443 | Women's Fashion | |
| 1 | 1.0 | 240.0 | 100147444 | Beauty & Grooming | |
| 2 | 1.0 | 2450.0 | 100147445 | Women's Fashion | |
| 3 | 1.0 | 60.0 | 100147446 | Beauty & Grooming | |
| 4 | 2.0 | 1110.0 | 100147447 | Soghaat | |
| … | … | … | … | … | |
| 584520 | 1.0 | 35899.0 | 100562386 | Mobiles & Tablets | |
| 584521 | 2.0 | 652178.0 | 100562387 | Mobiles & Tablets | |
| 584522 | 2.0 | 652178.0 | 100562387 | Mobiles & Tablets | |
| 584523 | 2.0 | 652178.0 | 100562387 | Mobiles & Tablets | |
| 584524 | NaN | NaN | NaN | NaN | |

| | sales_commission_code | … | M-Y | FY | Customer ID | unnamed_21 | \ |
|---|---|---|---|---|---|---|---|
| 0 | \N | … | 7-2016 | FY17 | 1.0 | NaN | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | \N | … | 7-2016 | FY17 | 2.0 | NaN | |
| 2 | \N | … | 7-2016 | FY17 | 3.0 | NaN | |
| 3 | R-FSD-52352 | … | 7-2016 | FY17 | 4.0 | NaN | |
| 4 | \N | … | 7-2016 | FY17 | 5.0 | NaN | |
| … | … | … | … | … | … | … | |
| 584520 | NaN | … | 8-2018 | FY19 | 115326.0 | NaN | |
| 584521 | NaN | … | 8-2018 | FY19 | 113474.0 | NaN | |
| 584522 | NaN | … | 8-2018 | FY19 | 113474.0 | NaN | |
| 584523 | NaN | … | 8-2018 | FY19 | 113474.0 | NaN | |
| 584524 | NaN | … | NaN | NaN | NaN | NaN | |

| | unnamed_22 | unnamed_23 | unnamed_24 | unnamed_25 | working_date | order_date |
|---|---|---|---|---|---|---|
| 0 | NaN | NaN | NaN | NaN | 2016-07-01 | 2016-07-01 |
| 1 | NaN | NaN | NaN | NaN | 2016-07-01 | 2016-07-01 |
| 2 | NaN | NaN | NaN | NaN | 2016-07-01 | 2016-07-01 |
| 3 | NaN | NaN | NaN | NaN | 2016-07-01 | 2016-07-01 |
| 4 | NaN | NaN | NaN | NaN | 2016-07-01 | 2016-07-01 |
| … | … | … | … | … | … | … |
| 584520 | NaN | NaN | NaN | NaN | 2018-08-28 | 2018-08-28 |
| 584521 | NaN | NaN | NaN | NaN | 2018-08-28 | 2018-08-28 |
| 584522 | NaN | NaN | NaN | NaN | 2018-08-28 | 2018-08-28 |
| 584523 | NaN | NaN | NaN | NaN | 2018-08-28 | 2018-08-28 |
| 584524 | NaN | NaN | NaN | NaN | NaT | NaT |

[584525 rows x 28 columns]

[27]:
```
df["total_revenue"] = df["price"] * df["qty_ordered"] df
```

[27]:

| | item_id | status | created_at | \ |
|---|---|---|---|---|
| 0 | 211131.0 | complete | 2016-07-01 | |
| 1 | 211133.0 | canceled | 2016-07-01 | |
| 2 | 211134.0 | canceled | 2016-07-01 | |
| 3 | 211135.0 | complete | 2016-07-01 | |
| 4 | 211136.0 | order_refunded | 2016-07-01 | |
| … | … | … | … | |
| 584520 | 905205.0 | processing | 2018-08-28 | |
| 584521 | 905206.0 | processing | 2018-08-28 | |
| 584522 | 905207.0 | processing | 2018-08-28 | |
| 584523 | 905208.0 | processing | 2018-08-28 | |
| 584524 | NaN | NaN | NaT | |

| | sku | price | \ |
|---|---|---|---|
| 0 | kreations_YI 06-L | 1950.0 | |
| 1 | kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo… | 240.0 | |

| | | |
|---|---|---|
| 2 | Ego_UP0017-999-MR0 | 2450.0 |
| 3 | kcc_krone deal | 360.0 |
| 4 | BK7010400AG | 555.0 |
| … | … | … |
| 584520 | MATHUA5AF70A7D1E50A | 35599.0 |
| 584521 | MATSAM5B6D7208C6D30 | 129999.0 |
| 584522 | MATSAM5B1509B4696EA | 87300.0 |
| 584523 | MATSAM5B10F91A9B6AB | 108640.0 |
| 584524 | NaN | NaN |

| | qty_ordered | grand_total | increment_id | category_name_1 | \ |
|---|---|---|---|---|---|
| 0 | 1.0 | 1950.0 | 100147443 | Women's Fashion | |
| 1 | 1.0 | 240.0 | 100147444 | Beauty & Grooming | |
| 2 | 1.0 | 2450.0 | 100147445 | Women's Fashion | |
| 3 | 1.0 | 60.0 | 100147446 | Beauty & Grooming | |
| 4 | 2.0 | 1110.0 | 100147447 | Soghaat | |
| … | … | … | … | … | |
| 584520 | 1.0 | 35899.0 | 100562386 | Mobiles | & Tablets |
| 584521 | 2.0 | 652178.0 | 100562387 | Mobiles | & Tablets |
| 584522 | 2.0 | 652178.0 | 100562387 | Mobiles | & Tablets |
| 584523 | 2.0 | 652178.0 | 100562387 | Mobiles | & Tablets |
| 584524 | NaN | NaN | NaN | | NaN |

| | sales_commission_code | … | Customer ID | unnamed_21 | unnamed_22 | \ |
|---|---|---|---|---|---|---|
| 0 | \N | … | 1.0 | NaN | NaN | |
| 1 | \N | … | 2.0 | NaN | NaN | |
| 2 | \N | … | 3.0 | NaN | NaN | |
| 3 | R-FSD-52352 | … | 4.0 | NaN | NaN | |
| 4 | \N | … | 5.0 | NaN | NaN | |
| … | … | … | … | … | … | |
| 584520 | NaN | … | 115326.0 | NaN | NaN | |
| 584521 | NaN | … | 113474.0 | NaN | NaN | |
| 584522 | NaN | … | 113474.0 | NaN | NaN | |
| 584523 | NaN | … | 113474.0 | NaN | NaN | |
| 584524 | NaN | … | NaN | NaN | NaN | |

| | unnamed_23 | unnamed_24 | unnamed_25 | working_date | order_date | \ |
|---|---|---|---|---|---|---|
| 0 | NaN | NaN | NaN | 2016-07-012016-07-01 | | |
| 1 | NaN | NaN | NaN | 2016-07-012016-07-01 | | |
| 2 | NaN | NaN | NaN | 2016-07-012016-07-01 | | |
| 3 | NaN | NaN | NaN | 2016-07-012016-07-01 | | |
| 4 | NaN | NaN | NaN | 2016-07-012016-07-01 | | |
| … | … | … | … | … | … | |
| 584520 | NaN | NaN | NaN | 2018-08-282018-08-28 | | |
| 584521 | NaN | NaN | NaN | 2018-08-282018-08-28 | | |
| 584522 | NaN | NaN | NaN | 2018-08-282018-08-28 | | |
| 584523 | NaN | NaN | NaN | 2018-08-282018-08-28 | | |

```
584524          NaN          NaN          NaN          NaT          NaT

       customer_since total_revenue
0          2016-07-01        1950.0
1          2016-07-01         240.0
2          2016-07-01        2450.0
3          2016-07-01         360.0
4          2016-07-01        1110.0
…                 …             …
584520     2018-08-01       35599.0
584521     2018-07-01      259998.0
584522     2018-07-01      174600.0
584523     2018-07-01      217280.0
584524            NaT           NaN

[584525 rows x 30 columns]
```

[28]:
```
df.columns
```

[28]:
```
Index(['item_id', 'status', 'created_at', 'sku', 'price', 'qty_ordered',
       'grand_total', 'increment_id', 'category_name_1',
       'sales_commission_code', 'discount_amount', 'payment_method',
       'Working Date', 'BI Status', ' MV ', 'Year', 'Month', 'Customer Since',
       'M-Y', 'FY', 'Customer ID', 'unnamed_21', 'unnamed_22', 'unnamed_23',
       'unnamed_24', 'unnamed_25', 'working_date', 'order_date',
       'customer_since', 'total_revenue'],
      dtype='object')
```

to clean the dataset if have the null value

[29]:
```
df.to_csv("cleaned_bigdata.csv", index=False)
df
```

[29]:
```
          item_id              status created_at    \
0        211131.0            complete 2016-07-01
1        211133.0            canceled 2016-07-01
2        211134.0            canceled   2016-07-01
3        211135.0            complete   2016-07-01
4        211136.0    order_refunded 2016-07-01
…             …                   …          …
584520   905205.0          processing 2018-08-28
584521   905206.0          processing 2018-08-28
584522   905207.0          processing 2018-08-28
584523   905208.0          processing 2018-08-28
584524         NaN                 NaN        NaT
```

| | | |
|---|---|---|
| 0 | kreations_YI 06-L | 1950.0 |
| 1 | kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo… | 240.0 |
| 2 | Ego_UP0017-999-MR0 | 2450.0 |
| 3 | kcc_krone deal | 360.0 |
| 4 | BK7010400AG | 555.0 |
| … | … | … |
| 584520 | MATHUA5AF70A7D1E50A | 35599.0 |
| 584521 | MATSAM5B6D7208C6D30 | 129999.0 |
| 584522 | MATSAM5B1509B4696EA | 87300.0 |
| 584523 | MATSAM5B10F91A9B6AB | 108640.0 |
| 584524 | NaN | NaN |

| | qty_ordered | grand_total | increment_id | category_name_1 | \ |
|---|---|---|---|---|---|
| 0 | 1.0 | 1950.0 | 100147443 | Women's Fashion | |
| 1 | 1.0 | 240.0 | 100147444 | Beauty & Grooming | |
| 2 | 1.0 | 2450.0 | 100147445 | Women's Fashion | |
| 3 | 1.0 | 60.0 | 100147446 | Beauty & Grooming | |
| 4 | 2.0 | 1110.0 | 100147447 | Soghaat | |
| … | … | … | … | … | |
| 584520 | 1.0 | 35899.0 | 100562386 | Mobiles | & Tablets |
| 584521 | 2.0 | 652178.0 | 100562387 | Mobiles | & Tablets |
| 584522 | 2.0 | 652178.0 | 100562387 | Mobiles | & Tablets |
| 584523 | 2.0 | 652178.0 | 100562387 | Mobiles | & Tablets |
| 584524 | NaN | NaN | NaN | | NaN |

| | sales_commission_code | … | Customer ID | unnamed_21 | unnamed_22 | \ |
|---|---|---|---|---|---|---|
| 0 | \N | … | 1.0 | NaN | NaN | |
| 1 | \N | … | 2.0 | NaN | NaN | |
| 2 | \N | … | 3.0 | NaN | NaN | |
| 3 | R-FSD-52352 | … | 4.0 | NaN | NaN | |
| 4 | \N | … | 5.0 | NaN | NaN | |
| … | … | … | … | … | … | |
| 584520 | NaN | … | 115326.0 | NaN | NaN | |
| 584521 | NaN | … | 113474.0 | NaN | NaN | |
| 584522 | NaN | … | 113474.0 | NaN | NaN | |
| 584523 | NaN | … | 113474.0 | NaN | NaN | |
| 584524 | NaN | … | NaN | NaN | NaN | |

| | unnamed_23 | unnamed_24 | unnamed_25 | working_date | order_date | \ |
|---|---|---|---|---|---|---|
| 0 | NaN | NaN | NaN | 2016-07-012016-07-01 | | |
| 1 | NaN | NaN | NaN | 2016-07-012016-07-01 | | |
| 2 | NaN | NaN | NaN | 2016-07-012016-07-01 | | |
| 3 | NaN | NaN | NaN | 2016-07-012016-07-01 | | |
| 4 | NaN | NaN | NaN | 2016-07-012016-07-01 | | |
| … | … | … | … | … | … | |
| 584520 | NaN | NaN | NaN | 2018-08-282018-08-28 | | |
| 584521 | NaN | NaN | NaN | 2018-08-282018-08-28 | | |

| | | | | | |
|---|---|---|---|---|---|
| 584522 | NaN | NaN | NaN | 2018-08-28 | 2018-08-28 |
| 584523 | NaN | NaN | NaN | 2018-08-28 | 2018-08-28 |
| 584524 | NaN | NaN | NaN | NaT | NaT |

| | customer_since | total_revenue |
|---|---|---|
| 0 | 2016-07-01 | 1950.0 |
| 1 | 2016-07-01 | 240.0 |
| 2 | 2016-07-01 | 2450.0 |
| 3 | 2016-07-01 | 360.0 |
| 4 | 2016-07-01 | 1110.0 |
| … | … | … |
| 584520 | 2018-08-01 | 35599.0 |
| 584521 | 2018-07-01 | 259998.0 |
| 584522 | 2018-07-01 | 174600.0 |
| 584523 | 2018-07-01 | 217280.0 |
| 584524 | NaT | NaN |

[584525 rows x 30 columns]

ok let see the shape of dataset

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

[31]:
```
df.shape
```

[31]: (584525, 30)

to change the name and prepare for load to postgres

[32]:
```
# Rename unnamed columns if needed
df.rename(columns={"Unnamed: 21": "unnamed_21", "Unnamed: 22": "unnamed_22",
                   "Unnamed: 23": "unnamed_23", "Unnamed: 24": "unnamed_24",
                   "Unnamed: 25": "unnamed_25"}, inplace=True)

# Convert dates
df["created_at"] = pd.to_datetime(df["created_at"])
df["working_date"] = pd.to_datetime(df["Working Date"], errors='coerce')
df["order_date"] = pd.to_datetime(df["order_date"])
df["customer_since"] = pd.to_datetime(df["Customer Since"], errors='coerce')

# Convert numeric columns
df["price"] = pd.to_numeric(df["price"], errors="coerce")
df["qty_ordered"] = pd.to_numeric(df["qty_ordered"], errors="coerce")
df["grand_total"] = pd.to_numeric(df["grand_total"], errors="coerce")
df["discount_amount"] = pd.to_numeric(df["discount_amount"], errors="coerce")
df["total_revenue"] = pd.to_numeric(df["total_revenue"], errors="coerce")

# Remove duplicates
```

[32]:

| | item_id | status | created_at | \ |
|---|---|---|---|---|
| 0 | 211131.0 | complete | 2016-07-01 | |
| 1 | 211133.0 | canceled | 2016-07-01 | |
| 2 | 211134.0 | canceled | 2016-07-01 | |
| 3 | 211135.0 | complete | 2016-07-01 | |
| 4 | 211136.0 | order_refunded | 2016-07-01 | |

| | sku | price | qty_ordered | \ |
|---|---|---|---|---|
| 0 | kreations_YI 06-L | 1950.0 | 1.0 | |
| 1 | kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo… | 240.0 | 1.0 | |
| 2 | Ego_UP0017-999-MR0 | 2450.0 | 1.0 | |
| 3 | kcc_krone deal | 360.0 | 1.0 | |
| 4 | BK7010400AG | 555.0 | 2.0 | |

| | grand_total | increment_id | category_name_1 | sales_commission_code | … | \ |
|---|---|---|---|---|---|---|
| 0 | 1950.0 | 100147443 | Women's Fashion | \N | … | |
| 1 | 240.0 | 100147444 | Beauty & Grooming | \N | … | |
| 2 | 2450.0 | 100147445 | Women's Fashion | \N | … | |
| 3 | 60.0 | 100147446 | Beauty & Grooming | R-FSD-52352 | … | |
| 4 | 1110.0 | 100147447 | Soghaat | \N | … | |

| | Customer ID | unnamed_21 | unnamed_22 | unnamed_23 | unnamed_24 | unnamed_25 | \ |
|---|---|---|---|---|---|---|---|
| 0 | 1.0 | NaN | NaN | NaN | NaN | NaN | |
| 1 | 2.0 | NaN | NaN | NaN | NaN | NaN | |
| 2 | 3.0 | NaN | NaN | NaN | NaN | NaN | |
| 3 | 4.0 | NaN | NaN | NaN | NaN | NaN | |
| 4 | 5.0 | NaN | NaN | NaN | NaN | NaN | |

| | working_date | order_date | customer_since | total_revenue |
|---|---|---|---|---|
| 0 | 2016-07-01 | 2016-07-01 | 2016-07-01 | 1950.0 |
| 1 | 2016-07-01 | 2016-07-01 | 2016-07-01 | 240.0 |
| 2 | 2016-07-01 | 2016-07-01 | 2016-07-01 | 2450.0 |
| 3 | 2016-07-01 | 2016-07-01 | 2016-07-01 | 360.0 |
| 4 | 2016-07-01 | 2016-07-01 | 2016-07-01 | 1110.0 |

# 3. Data Storage in PostgreSQL

## 3.1 Creating a Table in PostgreSQL

```
CREATE TABLE sales_data (
    item_id SERIAL PRIMARY KEY,
    status VARCHAR(255),
    created_at TIMESTAMP,
    sku VARCHAR(255),
    price NUMERIC(10,2),
    quantity INT,
    grand_total NUMERIC(10,2),
    category_name_1 VARCHAR(255),
```

```
    discount_amount NUMERIC(10,2),
    payment_method VARCHAR(255),
    order_date TIMESTAMP,
    customer_id VARCHAR(255)
);
```

## 3.2 Inserting Data into PostgreSQL using Pandas

[33]:
```
# Database connection
                engine = create_engine("postgresql://postgres:1221@localhost:5432/dataset")

# Store cleaned DataFrame into PostgreSQL
df.to_sql("sales_data", 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

[35]:
```
query = "SELECT * FROM sales_data LIMIT 5;"
df_sql = pd.read_sql(query, engine) df_sql.head()
```

[35]:

| | item_id | status | created_at | \ |
|---|---|---|---|---|
| 0 | 211131.0 | complete | 2016-07-01 | |
| 1 | 211133.0 | canceled | 2016-07-01 | |
| 2 | 211134.0 | canceled | 2016-07-01 | |
| 3 | 211135.0 | complete | 2016-07-01 | |
| 4 | 211136.0 | order_refunded | 2016-07-01 | |

| | sku | price | qty_ordered | \ |
|---|---|---|---|---|
| 0 | kreations_YI 06-L | 1950.0 | 1.0 | |
| 1 | kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo… | 240.0 | 1.0 | |
| 2 | Ego_UP0017-999-MR0 | 2450.0 | 1.0 | |
| 3 | kcc_krone deal | 360.0 | 1.0 | |
| 4 | BK7010400AG | 555.0 | 2.0 | |

| | grand_total | increment_id | category_name_1 | sales_commission_code | ... | \ |
|---|---|---|---|---|---|---|
| 0 | 1950.0 | 100147443 | Women's Fashion | \N | … | |
| 1 | 240.0 | 100147444 | Beauty & Grooming | \N | … | |
| 2 | 2450.0 | 100147445 | Women's Fashion | \N | … | |
| 3 | 60.0 | 100147446 | Beauty & Grooming | R-FSD-52352 | … | |
| 4 | 1110.0 | 100147447 | Soghaat | \N | … | |

| | Customer ID | unnamed_21 | unnamed_22 | unnamed_23 | unnamed_24 | unnamed_25 | \ |
|---|---|---|---|---|---|---|---|
| 0 | 1.0 | None | None | None | None | None | |
| 1 | 2.0 | None | None | None | None | None | |
| 2 | 3.0 | None | None | None | None | None | |
| 3 | 4.0 | None | None | None | None | None | |
| 4 | 5.0 | None | None | None | None | None | |

| | working_date | order_date | customer_since | total_revenue |
|---|---|---|---|---|
| 0 | 2016-07-01 | 2016-07-01 | 2016-07-01 | 1950.0 |
| 1 | 2016-07-01 | 2016-07-01 | 2016-07-01 | 240.0 |
| 2 | 2016-07-01 | 2016-07-01 | 2016-07-01 | 2450.0 |
| 3 | 2016-07-01 | 2016-07-01 | 2016-07-01 | 360.0 |
| 4 | 2016-07-01 | 2016-07-01 | 2016-07-01 | 1110.0 |

[5 rows x 30 columns]

finaly show the data shape after clean and load a data

[36]:
```
df.shape
```

(584525, 30)

# 4. Data Visualization using Power BI

## 4.1 Connecting PostgreSQL to Power BI

Open **Power BI Desktop**

Click **"Get Data"** → "PostgreSQL Database"

Enter:

**Server:** localhost

**Database:**dataset

**Port:** 5432

**User:** postgres

**Password:** ****

Click **"Load"** to import the sales_data table

### 4.2 Charts Created in Power BI

**1.colomun chart: Sales profit Over category**

**X-axis: category_name**

**Y-axis:** average total_revenue

**Legend: year**

**Insight:** Visualizes average revenue .

**2. Area Chart: Top-Selling Categories**

**X-axis:** created at day

**Y-axis:** sum of price

**Insight:** Shows best-selling product.

**3. Donut Chart: total revenue of the product in category**

**Legend:** category_name

**Values:** sum of total revenue

**Insight:** Shows the revenue on product .

**4. Table: Monthly Sales Performance**

**Column:** sum of price

Product name

**Insight:** Highlights the price of product.

# 5. Conclusion

This assignment successfully demonstrates an end-to-end ETL (Extract, Transform, Load) pipeline for an e-commerce dataset. Key takeaways:

Pandas was used for data extraction, cleaning, and transformation.

PostgreSQL was used to store the structured data.

Power BI was used for visualizing insights through interactive dashboards.