

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()

item_id	status	s created_at \
211131.0	complete	7/1/2016
211133.0	canceled	7/1/2016
211134.0	canceled	7/1/2016
211135.0	complete	7/1/2016
211136.0	order_refunded	7/1/2016
	211131.0 211133.0 211134.0 211135.0	211131.0 complete 211133.0 canceled 211134.0 canceled 211135.0 complete

	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 ...
                                                                                             \ \ \ \ \dots
                     240.0
         1
                                100147444
                                                 Beauty & Grooming
         2
                                                                                             \N ...
                    2450.0
                                100147445
                                                  Women's Fashion
         3
                       60.0
                                100147446
                                                  Beauty & Grooming
                                                                                  R-FSD-52352 ...
         4
                    1110.0
                                                            Soghaat
                                                                                             \N ...
                                100147447
                 Month Customer Since
                                                       FY Customer ID
                                                                            Unnamed: 21 Unnamed: 22 \
                                            M-Y
         0
                               2016-7
                7.0
                                           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
       df.shape
 [2]: (1048575, 26)
       to show the column
       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]
```

qty ordered

grand total \

price

[2]:

[19]:

[21]:

[21]:

df.describe()

item id

count	584524.000000	5.845240e+05	584524.000	000 5.845	240e+05
mean	565667.074218	6.348748e+03	1.296	388 8.530	619e+03
std	200121.173648	1.494927e+04	3.996	061 6.132	081e+04
min	211131.000000	0.000000e+00	1.000	000 -1.594	1000e+03
25%	395000.750000	3.600000e+02	1.000	9.450	000e+02
50%	568424.500000	8.990000e+02	1.000	000 1.960	400e+03
75%	739106.250000	4.070000e+03	1.000	000 6.999	000e+03
max	905208.000000	1.012626e+06	1000.000	000 1.788	800e+07
	discount_amount	Yea	ır	Month	Customer ID \
count	584524.000000	584524.00000	0 584524	1.000000 5	84513.000000
mean	499.492775	2017.04411	5 7	.167654 4	5790.511965
std	1506.943046	0.70735	5 3	.486305 3	34414.962389
min	-599.500000	2016.00000	0 1	.000000	1.000000
25%	0.000000	2017.00000	0 4	.000000 1	3516.000000
50%	0.000000	2017.00000	0 7	.000000 4	2856.000000
75%	160.500000	2018.000000	11.000	000 73	536.000000
max	90300.000000	2018.000000	12.000		5326.000000
Шах	70300.000000	2010.000000	12.000	000 11	3320.00000
	Unnamed: 21		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	

price \

sku

[20]:	item_id	status create	d_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	//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

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_to	otal increment_id	cat	egory_name_1	\
0	1.0	1950.0	100147443	Wo	men's Fashion	
1	1.0	240.0	100147444	Beauty	& Grooming	
2	1.0	2450.0	100147445	Wo	men's Fashion	
3	1.0	60.0	100147446	Beauty	& Grooming	
4	2.0	1110.0	100147447		Soghaat	

...

584520	1.0	35899.0	100	562386	Mobiles	& Tab	olets	
584521	2.0	652178.0	100	562387	Mobiles	& Tab	olets	
584522	2.0	652178.0		562387	Mobiles	& Tab		
584523	2.0	652178.0		562387	Mobiles	& Tab		
584524	NaN	NaN		NaN			NaN	
30 132 1	11411	11411		11411			11411	
	sales commission	code	M	onth Cust	tomer Since	M-Y	FY	\
0		\N	7.0	contin e as	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-52		7.0		2016-7	7-2016	FY17	
4	K-1 5D-32	\ 3 . T	7.0		2016-7	7-2016	FY17	
7			7.0			7-2010	1 11/	
584520	 N	 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			8.0		2018-7	8-2018	FY19	
584524	1	NaN	NaN		NaN	NaN	NaN	
	Customer ID	Unnam	ed: 21	Unname	ed: 22 Unnar	ned: 23 Uni	named: 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
	5.0		INGIN		INGIN			INAIN
584520	115326.0	•••	NaN	•••	NaN	NaN		NaN
584521	113474.0		NaN		NaN	NaN		NaN
584522 584522			NaN		NaN	NaN		NaN NaN
584523	113474.0		NaN		NaN	NaN		NaN
584524	NaN		NaN		NaN	NaN	1	NaN
	Unnamed: 25							
0	NaN							
1	NaN							
2	NaN							
3								
3 4	NaN NaN							
	NaN							
 594520	NT.NT							
584520	NaN							
584521	NaN							
584522	NaN							
584523	NaN							
584524	NaN							

df.shape

[584525 rows x 26 column

2.3 Data Type Conversions

0

```
df["order date"] = pd.to datetime(df["created at"]) # Convert to datetime
[25]
[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
                                                   2016-07-01
                                     canceled
                   211134.0
        3
                   211135.0
                                     complete
                                                    2016-07-01
       4
                 211136.0 order refunded 2016-07-01
                    905205.0
                                       processing 2018-08-28
       584520
       584521
                    905206.0
                                       processing 2018-08-28
       584522
                    905207.0
                                       processing 2018-08-28
       584523
                    905208.0
                                       processing 2018-08-28
                                           NaN
       584524
                       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
                                          1950.0
                                                                      Women's Fashion
                             1.0
                                                    100147443
                             1.0
       1
                                          240.0
                                                    100147444
                                                                      Beauty & Grooming
       2
                             1.0
                                         2450.0
                                                                      Women's Fashion
                                                    100147445
       3
                             1.0
                                            60.0
                                                    100147446
                                                                      Beauty & Grooming
       4
                             2.0
                                          1110.0
                                                    100147447
                                                                                Soghaat
                             1.0
                                        35899.0
                                                                       Mobiles & Tablets
       584520
                                                    100562386
                             2.0
                                       652178.0
                                                                       Mobiles & Tablets
       584521
                                                    100562387
       584522
                             2.0
                                       652178.0
                                                    100562387
                                                                       Mobiles & Tablets
       584523
                             2.0
                                       652178.0
                                                    100562387
                                                                       Mobiles & Tablets
       584524
                           NaN
                                           NaN
                                                           NaN
                                                                                   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	unname	ed_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 20	018-08-28 2018-08-28
584523	NaN NaN	NaN	NaN 20	018-08-28 2018-08-28
584524	NaN NaN	NaN	NaN	NaT NaT

[584525 rows x 28 columns]

```
df["total revenue"] = df["price"] * df["qty ordered"] df
[27]:
   [27]:
                                        status created at \
                     item id
         0
                                     complete 2016-07-01
                   211131.0
         1
                   211133.0
                                     canceled 2016-07-01
         2
                   211134.0
                                     canceled
                                                    2016-07-01
         3
                   211135.0
                                     complete
                                                    2016-07-01
                 211136.0 order_refunded 2016-07-01
        4
                    905205.0
                                       processing 2018-08-28
       584520
       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
                                                                                     1950.0
                                                             kreations YI 06-L
         1
                              kcc Buy 2 Frey Air Freshener & Get 1 Kasual Bo...
                                                                                    240.0
```

	Ego_UP0017-999-MR0 kcc_krone deal BK7010400AG					
	 MATHUA5AF70A7D1E50A MATSAM5B6D7208C6D30 MATSAM5B1509B4696EA MATSAM5B10F91A9B6AB NaN					
qty_ordered 1.0 1.0 1.0 1.0 2.0	19 2 24	950.0 240.0 450.0 60.0	100147443 100147444 100147445 100147446 100147447	Wome Beauty o Wome	en's Fashion & Grooming en's Fashion	\
1.0 2.0 2.0 2.0 2.0 NaN	652 652	178.0 178.0	100562386 100562387 100562387 100562387	Mobiles Mobiles Mobiles Mobiles	& Tablets & Tablets & Tablets & Tablets NaN	
_	_ \N \N \N		1.0 2.0 3.0 4.0	NaN NaN NaN NaN	1 unnamed_22 NaN NaN NaN NaN NaN NaN	\
	NaN NaN NaN NaN NaN NaN		115326.0 113474.0 113474.0 113474.0 NaN	NaN NaN NaN NaN NaN	NaN NaN NaN NaN NaN	
unnamed_23 unnan NaN NaN NaN NaN NaN NaN NaN NaN NaN	ned_24 NaN NaN NaN NaN NaN NaN NaN NaN NaN Na	unna	med_25 NaN NaN NaN NaN NaN NaN NaN NaN NaN Na	2016-07-0120 2016-07-0120 2016-07-0120 2016-07-0120 2016-07-0120 2018-08-2820 2018-08-2820	016-07-01 016-07-01 016-07-01 016-07-01 018-08-28 018-08-28	
	1.0 1.0 1.0 1.0 1.0 2.0 1.0 2.0 2.0 2.0 2.0 NaN sales_commission R-FSD unnamed_23 unnar NaN NaN NaN NaN NaN NaN NaN NaN NaN N	1.0 19 1.0 24 1.0 24 1.0 22 1.0 2.0 11 1.0 358 2.0 652 2.0 652 2.0 652 2.0 652 NaN sales_commission_code \N \N \N R-FSD-52352 \N NaN NaN NaN NaN NaN NaN NaN NaN NaN	1.0 1950.0 1.0 240.0 1.0 2450.0 1.0 60.0 2.0 1110.0 1.0 35899.0 2.0 652178.0 2.0 652178.0 2.0 652178.0 2.0 652178.0 NaN NaN sales_commission_code \N aN \NaN	MATHUA5. MATSAM5 MAT	Rec_krone de BK70104002	Rcc_krone_deal 360.0

584524	NaN	NaN	NaN	NaT	NaT
	customer_s	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			
[EQ4E3E	20 1 1				

[584525 rows x 30 columns]

[28]: df.columns

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
                                                2016-07-01
                211134.0
                                 canceled
     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 1 2 3 4 584520 584521 584522 584523	kreations_YI 06-L kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo Ego_UP0017-999-MR0 kcc_krone deal BK7010400AG MATHUA5AF70A7D1E50A MATSAM5B6D7208C6D30 MATSAM5B1509B4696EA MATSAM5B10F91A9B6AB						240.0 0 2450.0 al 360.0 G 555.0 A 35599.0 0 129999.0 A 87300.0 B 108640.0))))
584524						Nal	N NaN	
	qty ordered	g	rand to	tal increment	id	categor	y name 1	\
0	1.0	_	50.0	100147443	_		's Fashion	
1	1.0	2	40.0	100147444	4	Beauty &	Grooming	
2	1.0	24	50.0	100147445	5	Women	's Fashion	
3	1.0		60.0	100147446	5	Beauty & Grooming		
4	2.0	11	10.0	100147447	7		Soghaat	
•••				•••		•••		
584520	1.0	358	99.0	100562386	5	Mobiles	& Tablets	
584521	2.0	6521	78.0	100562387	7	Mobiles	& Tablets	
584522	2.0	6521	78.0	100562387	7	Mobiles	& Tablets	
584523	2.0	6521	78.0	100562387	7	Mobiles	& Tablets	
584524	NaN		NaN	N	aN		NaN	
	calas commission	aada		Custos	m on ID	yamamad 21	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	\
0	sales_commission_	_code \N	•••		.0	unnamed_21 NaN	unnamed_22 NaN	\
1		\N			2.0	NaN	NaN	
2		\N			3.0	NaN	NaN	
3	R-FSD-52				i.0	NaN	NaN	
4	K-1 5D-5.	\N			5.0	NaN	NaN	
			•••	<i>3</i>		11411	11411	
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		Na		NaN	NaN	
	unnamed_23 unname	ed_24	unnan	ned_25		ting_date	order_date	\
0	NaN	NaN		NaN		16-07-0120		
1		NaN		NaN		16-07-0120		
2		NaN		NaN		16-07-0120		
3		NaN		NaN		16-07-0120		
4	NaN	NaN		NaN	20	16-07-0120	16-07-01	
			•••		•••			
584520		NaN		NaN		18-08-2820		
584521	NaN	NaN		NaN	20	18-08-2820	18-08-28	

584522	NaN	NaN	NaN	2018-08-28 20	18-08-28
584523	NaN	NaN	NaN	2018-08-28 20	18-08-28
584524	NaN	NaN	NaN	NaT	NaT
	customer_s	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	1		
•••	•••	•••			
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
          211134.0
                           canceled
                                          2016-07-01
     3
                                          2016-07-01
          211135.0
                           complete
     4 211136.0 order refunded 2016-07-01
                                                                          price
                                                                                     qty ordered \
                                                                 sku
     0
                                                   kreations YI 06-L
                                                                         1950.0
                                                                                            1.0
       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 ...
                                                                                       \N ...
     1
                 240.0
                                            Beauty & Grooming
                            100147444
     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
                                             NaN
                                                           NaN
                                                                         NaN
                   1.0
                               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
            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
       query = "SELECT * FROM sales data LIMIT 5:"
[35]:
       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
                                    canceled 2016-07-01
              211134.0
         3
              211135.0
                                   complete 2016-07-01
        4
              211136.0
                                order refunded 2016-07-01
                                                                                        qty ordered \
                                                                     sku
                                                                              price
         0
                                                                            1950.0
                                                                                                1.0
                                                       kreations YI 06-L
         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
                                                                                          \N ...
                   2450.0
                                                Women's Fashion
                               100147445
         3
                      60.0
                                                Beauty & Grooming
                                                                               R-FSD-52352 ...
                               100147446
         4
                   1110.0
                               100147447
                                                          Soghaat
                                                                                          \ \ \ \ \dots
                        Customer ID unnamed 21 unnamed 22 unnamed 23 unnamed 24 unnamed 25 \
         0
                       1.0
                                 None
                                               None
                                                             None
                                                                           None
                                                                                         None
                       2.0
         1
                                 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.