FINAL PROJECT
DATA ANALYSIS - PYTHON

By: Endah Rakhmawati

Intensive Bootcamp Data Analysis Batch 18 | MySkill

MySkill

Overview

Project

Develop Python code for analyzing numerical data with NumPy, Tabular data with Pandas, data visualization Matplotlib or Seaborn, and Exploratory data analysis including manipulate data using dataframes and summarize data.

Dataset

The dataset used is sales data from Tokopedia (not real data). It consists of 4 tables in the period 2021 to 2022.

Dataset

order_detail:

1. id → unique number of order / id_order
2. customer_id → unique number of customer
3. order_date → date when transaction was made
4. sku_id → unique number of product (sku is stock keeping unit)
5. price → price listed on price tag
6. qty_ordered → number of items purchased by customer
7. before_discount → total price value of product (price * qty_ordered)
8. discount_amount → total product discount value
9. after_discount → total price value of product when reduced by discount
10. is_gross → indicates customer has not paid for order
11. is_valid → indicates customer has made payment
12. is_net → indicates transaction is complete
13. payment_id → unique number of payment method

Dataset

sku_detail:

1. id → unique number of the product (can be used for key when joining)
2. sku_name → name of the product
3. base_price → price of goods listed on the price tag / price
4. cogs → cost of goods sold / total cost to sell 1 product
5. category → product category

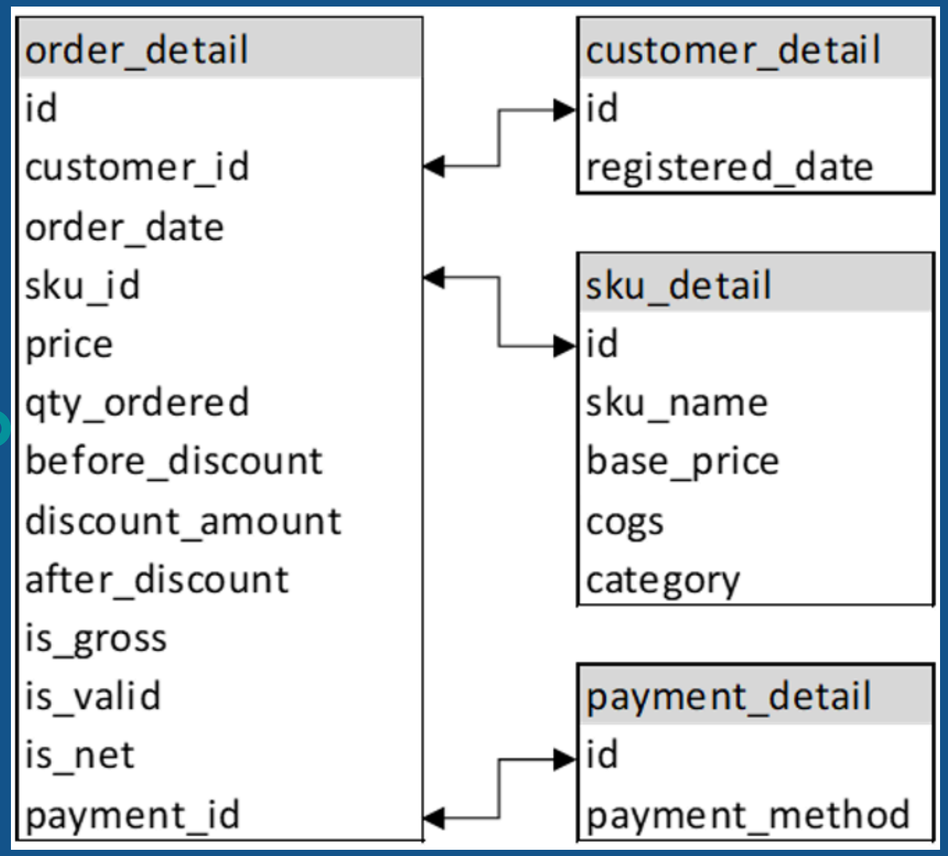
customer_detail:

1. id → unique number of the customer
2. registered_date → date the customer started registering as a member

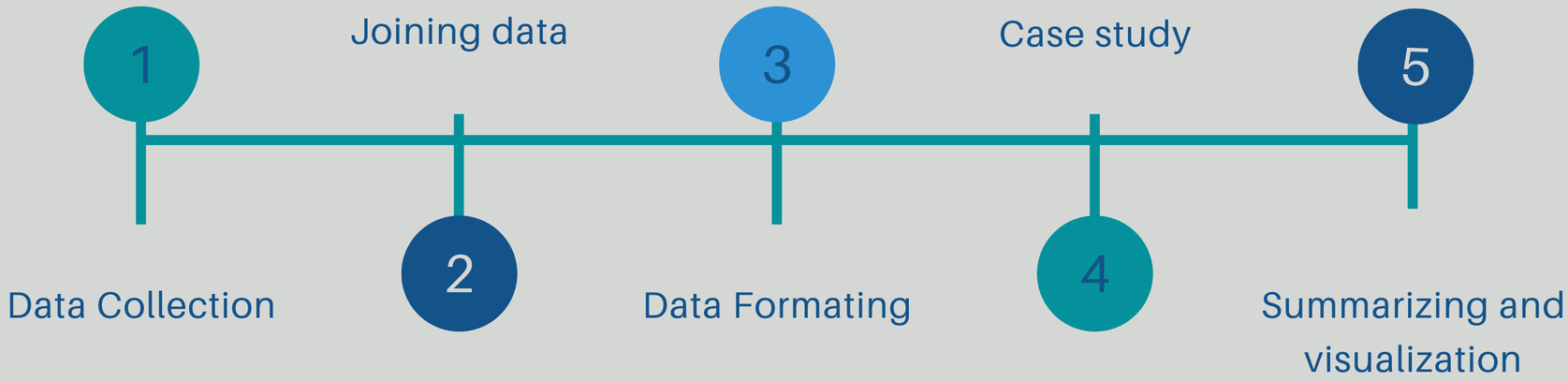
payment_detail:

1. id → unique number of payment method
2. payment_method → payment method used

Schema


order_detail
id
customer_id
order_date
sku_id
price
qty_ordered
before_discount
discount_amount
after_discount
is_gross
is_valid
is_net
payment_id
customer_detail
id
registered_date
sku_detail
id
sku_name
base_price
cogs
category
payment_detail
id
payment_method

Highlights


1
Data Collection
Joining data
2
3
Data Formating
Case study
4
5
Summarizing and visualization

Data Collection

```
#Sumber data yang digunakan
path_od = "https://raw.githubusercontent.com/dataskillsboost/FinalProjectDA11/main/order_detail.csv"
path_pd = "https://raw.githubusercontent.com/dataskillsboost/FinalProjectDA11/main/payment_detail.csv"
path_cd = "https://raw.githubusercontent.com/dataskillsboost/FinalProjectDA11/main/customer_detail.csv"
path_sd = "https://raw.githubusercontent.com/dataskillsboost/FinalProjectDA11/main/sku_detail.csv"
df_od = pd.read_csv(path_od)
df_pd = pd.read_csv(path_pd)
df_cd = pd.read_csv(path_cd)
df_sd = pd.read_csv(path_sd)
```

Extract dataset

Import dataset using url path

Dataframe

Creating dataframes for each dataset

Joining Data

Connecting to sqlite3

Create tables from dataframes the in-memory databases by specifying the database path as :memory:

```
#Menjalankan SQL di Colab
from sqlite3 import connect
conn = connect(':memory:')
df_od.to_sql('order_detail', conn, index=False, if_exists='replace')
df_pd.to_sql('payment_detail', conn, index=False, if_exists='replace')
df_sd.to_sql('sku_detail', conn, index=False, if_exists='replace')
df_cd.to_sql('customer_detail', conn, index=False, if_exists='replace')
```

Create final dataframe

Creating final dataframe from joining tables using SQL.

```
#Query SQL untuk menggabungkan data
df = pd.read_sql("""
SELECT
    order_detail.*,
    payment_detail.payment_method,
    sku_detail.sku_name,
    sku_detail.base_price,
    sku_detail.cogs,
    sku_detail.category,
    customer_detail.registered_date
FROM order_detail
LEFT JOIN payment_detail
    on payment_detail.id = order_detail.payment_id
LEFT JOIN sku_detail
    on sku_detail.id = order_detail.sku_id
LEFT JOIN customer_detail
    on customer_detail.id = order_detail.customer_id
""", conn)
```

Data Formating

```
#Mengubah tipe data agar mudah dilakukan pengolahan data
df = df.astype({"before_discount":'int', "discount_amount":'int', \
                "after_discount":'int',"base_price":'int'})
```

```
#Mengubah tipe kolom Date menjadi Datetime
df['order_date']= pd.to_datetime(df['order_date'])
df['registered_date']= pd.to_datetime(df['registered_date'])
df.dtypes
```

Change float to int

Columns : before_discount, discount_amount, after_discount, base_price

Change object to datetime

Columns : order_date, registered_date

Case Study

01

Dear Data Analyst,
At the end of this year, the company will give prizes to customers who win the Year-End Festival competition. The Marketing Team needs help to determine the estimated prizes that will be given to the winners of the competition later. The prizes will be taken from the TOP 5 Products from the Mobiles & Tablets Category during 2022, with the highest sales quantity (valid = 1). Please help, to send the data before the end of this month to the Marketing Team.
Thank you for your assistance.
Regards,
Marketing Team

Python

01

df_filter

- add new column 'year' extracted from order_date
- filter values : is_valid=1, category='Mobiles & Tablets', year=2022

```
df['year'] = pd.to_datetime(df['order_date']).dt.year
df_filter = df[['id','order_date','year','sku_name','category','qty_ordered']]\
 [(df['is_valid']==1) & (df['category']=='Mobiles & Tablets') & (df['year']==2022)]
df_filter.head()
```

	id	order_date	year	sku_name	category	qty_ordered
20	ODR2268957100j	2022-04-16	2022	Samsung_Galaxy_S8_Plus_Black	Mobiles & Tablets	1
300	ODR4269164386x	2022-09-06	2022	IDROID_BALRX7-Gold	Mobiles & Tablets	1000
334	ODR2855118495m	2022-07-03	2022	IDROID_BALRX7-Jet black	Mobiles & Tablets	26
350	ODR1542623352b	2022-07-23	2022	cc_samsung_G935F-Blue	Mobiles & Tablets	3
355	ODR4709500777n	2022-07-26	2022	Samsung-Galaxy-S8-G955-Plus-Black	Mobiles & Tablets	2

df1_answer

- group by : sku_name
- value : sum of qty_ordered

```
df1_answer = df_filter.groupby('sku_name').agg(totalqty=('qty_ordered','sum'))
df1_answer.sort_values(by=['totalqty'], ascending=False).head()
```

sku_name	totalqty
IDROID_BALRX7-Gold	1000
IDROID_BALRX7-Jet black	31
Infinix Hot 4-Gold	15
samsung_Grand Prime Plus-Black	11
infinix_Zero 4-Grey	10

Insights

01

01 The list of top 5 sku_names in the Mobiles & Tablets category for 2022, with the highest sales quantity, has been obtained.

02 All of them are based on Android operating system. Android is open-source, which allows a wide range of manufacturers to use it without licensing fees. This has led to a vast selection of Android devices across various price points, making it accessible to a global audience with different budgets.

Case Study

02

Dear Data Analyst,
Following up on the joint meeting of the Werehouse Team and Marketing Team, we found that the availability of product stock with the Others Category at the end of 2022 was still high.

- We ask for your assistance in checking the sales data for this category with 2021 in terms of sales quantity. Our temporary suspicion is that there has been a decrease in sales quantity in 2022 compared to 2021. (Please also display data for the 15 categories)
- If there is indeed a decrease in sales quantity in the Others category, we ask for your assistance in providing data on the TOP 20 product names that experienced the highest decrease in 2022 compared to 2021. We will use this as discussion material at the next meeting.

Please help to send the data no later than 4 days from today. Thank you for the assistance provided.
Regards,
Werehouse Team

Python

02.A

df_filter

- create list 'tahun' which contains of years, 2021 and 2022
- filter values : is_valid=1, year=tahun

```
tahun = [2021,2022]
df_filter = df[['id','order_date','year','sku_name','category','qty_ordered']]\
 [(df['is_valid']==1) & (df['year'].isin(tahun))]
df_filter.head()
```

	id	order_date	year	sku_name	category	qty_ordered
0	ODR9939707760w	2021-11-19	2021	RB_Dettol Germ Busting Kit-bf	Others	200
3	ODR3378927994s	2021-11-22	2021	dawlance_Inverter 30	Appliances	1
4	ODR4904430099k	2021-11-21	2021	Dawlance_Inverter-45 2.0 ton	Appliances	1
6	ODR7610732813d	2022-12-01	2022	mitsubhisi_1.0 Ton - SRK-13CMK-CS	Appliances	1
7	ODR4415476736l	2022-12-01	2022	lenovo_80HR00AKUE	Computing	1

df2_answer

- based on df_filter, index=category, columns=year, value=sum of qty_ordered
- add new column 'growth'= value of year 2022-value of year 2021

```
df2_answer = pd.pivot_table(df_filter, index='category',
                            columns='year', values='qty_ordered',
                            aggfunc='sum',
                            fill_value=0
)
df2_answer['growth'] = df2_answer[2022]-df2_answer[2021]
df2_answer.sort_values(by='growth', ascending=True)
```

Insights

02.A

01 It is true that the Others category will experience the largest decrease in sales quantity in 2022 by 163 units

02 If the decrease in sales quantity is due to a shift in priority needs, it often reflects customers reallocating their spending to essentials or higher-priority items over discretionary purchases

year category	2021	2022	growth
Others	426	263	-163
Soghaat	759	612	-147
Men Fashion	237	175	-62
Beauty & Grooming	168	153	-15
Appliances	124	148	24
Books	171	195	24
Health & Sports	173	200	27
Computing	109	153	44
School & Education	184	237	53
Home & Living	193	250	57
Kids & Baby	170	227	57
Entertainment	77	150	73
Superstore	327	536	209
Women Fashion	140	489	349
Mobiles & Tablets	107	1154	1047

Python

02.B

df_filter2

- based on df_filter
- filter values :
  category='Others'

```
df_filter2 = df_filter[['id','order_date','year','sku_name','category','qty_ordered']]\
 [(df_filter['category']=='Others')]
df_filter2.head()
```

	id	order_date	year	sku_name	category	qty_ordered
0	ODR9939707760w	2021-11-19	2021	RB_Dettol Germ Busting Kit-bf	Others	200
79	ODR5050363774l	2022-06-16	2022	Voucher 9000	Others	1
125	ODR3678705048c	2022-04-05	2022	Voucher 6000	Others	1
129	ODR8294249799k	2022-06-15	2022	Saylani_Health-Contribution-Package-1	Others	2
130	ODR9190651304u	2022-08-13	2022	Charizma_DG35	Others	1

df2_answer

- based on df_filter2, index=sku_name, columns=year, value=sum of qty_ordered
- add new column 'growth'= value of year 2022-value of year 2021
- display top 20 of the sku_name, which experienced the largest decrease in sales quantity

```
df2_answer = pd.pivot_table(df_filter2, index='sku_name',
                            columns='year', values='qty_ordered',
                            aggfunc='sum',
                            fill_value=0
)
df2_answer['growth'] = df2_answer[2022]-df2_answer[2021]
df2_answer.sort_values(by='growth', ascending=True).head(20)
```

Insights

02.B

01 RB_Dettol Germ Busting Kit-bf is a product from the Others category which experienced the largest decrease in sales quantity in 2022 of 155 units.

02 Check if a recent price increase could be causing the decrease, especially if it affects the affordability or perceived value of the product.

03 Analyze customer demographics and behaviors to see if there are any shifts, like preferences for different products, brands, or even competitor offerings.

year	2021	2022	growth
sku_name			
RB_Dettol Germ Busting Kit-bf	200	45	-155
Dawlance_MD 10 + DWB 600	23	0	-23
Telemall_MM-DR-HB-L	23	2	-21
iu_Tickets General Enclosure-Islamabad	20	0	-20
RS_Rehmat-e-Shereen Mix Mithai	13	0	-13
kansai_NeverWet	10	1	-9
sindbad_Sindbad Gift Card-3	7	0	-7
emart_00-1	7	1	-6
Vouch 365 2016	5	0	-5
Am-PTV_ATS-004-M	5	0	-5
duma_4561253300294	4	0	-4
sockoye_QG in Quarter Grey	4	0	-4
The Vitamin Company Kojic Acid Whitening Cream 40GM	4	0	-4
aw_Octane Booster-12oz./354ml	3	0	-3
MEGUIAR_G12711	4	1	-3
Trans2_LW 999	3	0	-3
MEGUIAR_G19216	2	0	-2
JBS_IFAM-009	2	0	-2
MEGUIAR_X1030EU	2	0	-2
sstop_Universallensclipkit	2	0	-2

Case Study

03

Dear Data Analyst,
Regarding the company's anniversary in the next 2 months, the Digital Marketing Team will provide promotional information for customers at the end of this month. The customer criteria that we will need are those who have checked out but have not made a payment (is_gross = 1) during 2022. The data we need is Customer ID and Registered Date.
Please help, to send the data before the end of this month to the Digital Marketing Team.
Thank you for the assistance provided.
Regards,
Digital Marketing Team

Python

03

df_filter

- filter values : is_gross=1, is_valid=0, is_net=0, year=2022

```
#Memfilter data dengan gross = 1, valid = 0, net = 0, transaksi selama 2022
df_filter = df[['id','customer_id','registered_date','order_date','year','sku_name','is_gross','is_valid']]\
 [(df['is_gross']==1) & (df['is_valid']==0) & (df['is_net']==0) & (df['year']==2022)]
df_filter.head()
```

	id	customer_id	registered_date	order_date	year	sku_name	is_gross	is_valid
9	ODR9699658949w	C246762L	2022-05-08	2022-05-21	2022	iPhone7Plus-Red-256GB	1	0
18	ODR1965502162e	C848774L	2021-11-07	2022-05-20	2022	iPhone7Plus-Red-256GB	1	0
19	ODR8450052777q	C693415L	2022-04-12	2022-04-15	2022	Samsung_Galaxy_S8_Plus_Gray	1	0
21	ODR7673587024b	C180595L	2022-04-22	2022-04-17	2022	Samsung_Galaxy_S8_Plus_Black	1	0
22	ODR7333927150n	C587425L	2022-03-22	2022-12-04	2022	Samsung_Galaxy_S8_Plus_Black	1	0

df3_answer

- copying df_filter
- add columns : customer_id and registered_date

```
#Buat dataframe baru yang berisi customer_id beserta registered_datenya
df3_answer = df_filter.copy()
df3_answer = df3_answer[['customer_id','registered_date']]
df3_answer.head()
```

	customer_id	registered_date
9	C246762L	2022-05-08
18	C848774L	2021-11-07
19	C693415L	2022-04-12
21	C180595L	2022-04-22
22	C587425L	2022-03-22

Python

03

drop_duplicates

- subset : customer_id, registered_date
- keep = 'first', keep the first row
- updated the dataframe at once

df3_answer

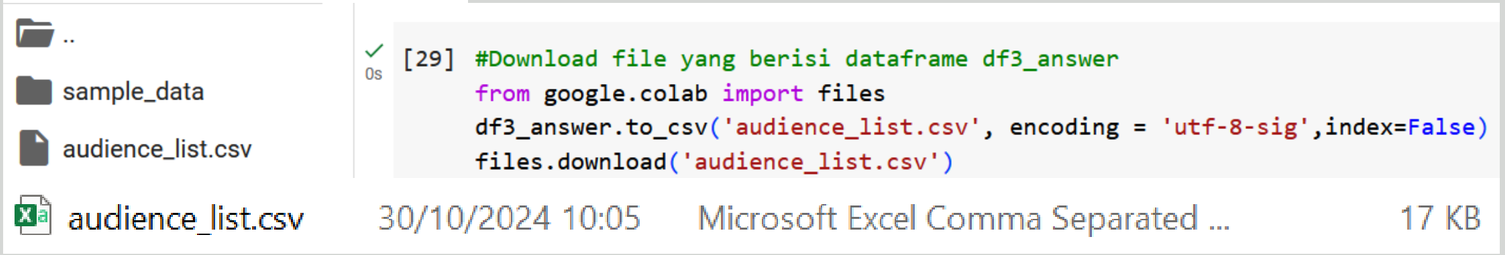
- there are 820 customers in list
- download as a csv file named 'audience_list'

```
#Hapus baris yang duplikat, pertahankan baris pertama saja.
df3_answer.drop_duplicates(subset=['customer_id', 'registered_date'], keep='first', inplace=True)
```

df3_answer

	customer_id	registered_date
9	C246762L	2022-05-08
18	C848774L	2021-11-07
19	C693415L	2022-04-12
21	C180595L	2022-04-22
22	C587425L	2022-03-22
...	...	...
5855	C653797L	2022-04-03
5856	C394076L	2021-10-12
5859	C248585L	2022-07-10
5865	C471304L	2022-05-13
5881	C265450L	2022-02-17

820 rows × 2 columns



Insights

03

01 During 2022, there are 820 customers provided promotional information in the end of this month. They are who have checked out but have not made a payment.

02 Complicated checkouts can discourage purchases. If possible, reduce steps, offer guest checkout, and make sure the process is mobile-friendly. Also we can consider the other additional services such as : Shipping Costs and Delivery Options, Payment Flexibility, etc.

03 We can calculate how long the customer has been shopping in our store since they registered. Also calculate how much the total transaction has been completed. Creating terms and conditions to be able to provide personal promotions. So they don't hesitate to make purchases.

Case Study

04

Dear Data Analyst,
From October to December 2022, we conducted a campaign every Saturday and Sunday. We want to assess whether the campaign had a sufficient impact on increasing sales (before_discount). Please help us display the following data:

- Average daily sales for weekends (Saturday and Sunday) vs. average daily sales for weekdays (Monday-Friday) per month. Is there an increase in sales in each of these months?
- Average daily sales for weekends (Saturday and Sunday) vs. average daily sales for weekdays (Monday-Friday) for the entire 3 months. Please help us send the data no later than next week.

Thank you for your assistance.
Regards,
Campaign Team

Python 04

```
df['month_year'] = pd.to_datetime(df['order_date']).dt.to_period('M')
df['day'] = pd.to_datetime(df['order_date']).dt.day_name()
weekend = ['Saturday','Sunday']
df['daysofweek'] = np.where(df['day'].isin(weekend), 'weekend', 'weekday')
df.head()
```

t_method	sku_name	base_price	cogs	category	registered_date	year	month_year	day	daysofweek
azzwallet	RB_Dettol Germ Busting Kit-bf	26100	18270	Others	2021-07-07	2021	2021-11	Friday	weekday
azzwallet	PS4_Slim-500GB	1971942	1321182	Entertainment	2021-11-20	2021	2021-11	Friday	weekday
Payaxis	Changhong Ruba 55 Inches UD55D6000i Ultra HD T...	7482000	5162580	Entertainment	2021-11-19	2021	2021-11	Thursday	weekday
azzwallet	dawlance_Inverter 30	3593680	3054628	Appliances	2021-11-03	2021	2021-11	Monday	weekday
Payaxis	Dawlance_Inverter-45 2.0 ton	4413220	3177472	Appliances	2021-07-05	2021	2021-11	Sunday	weekend

```
df['month_year'] = df['month_year'].astype(str)
```

```
bulan = ['2022-10','2022-11','2022-12']
df_filter = df[['id','month_year','daysofweek','before_discount']]\
 [(df['is_valid']==1) & (df['month_year'].isin(bulan))]
df_filter.head()
```

	id	month_year	daysofweek	before_discount
6	ODR7610732813d	2022-12	weekday	2697000
7	ODR4415476736l	2022-12	weekday	2533672
34	ODR3138948564v	2022-11	weekend	1195902
35	ODR6438394533v	2022-11	weekend	918952
36	ODR1691826218q	2022-11	weekend	762062

Add columns

- month_year : year and month extracted from order_date
- day : day name extracted from order_date
- dayofweek : mention the order_date whether is in weekend or weekday

df_filter

- change data type of month_year as string
- create list of month_year, named 'bulan'
- filter values : is_valid=1, month_year=bulan

Python

04.A

df4_answer

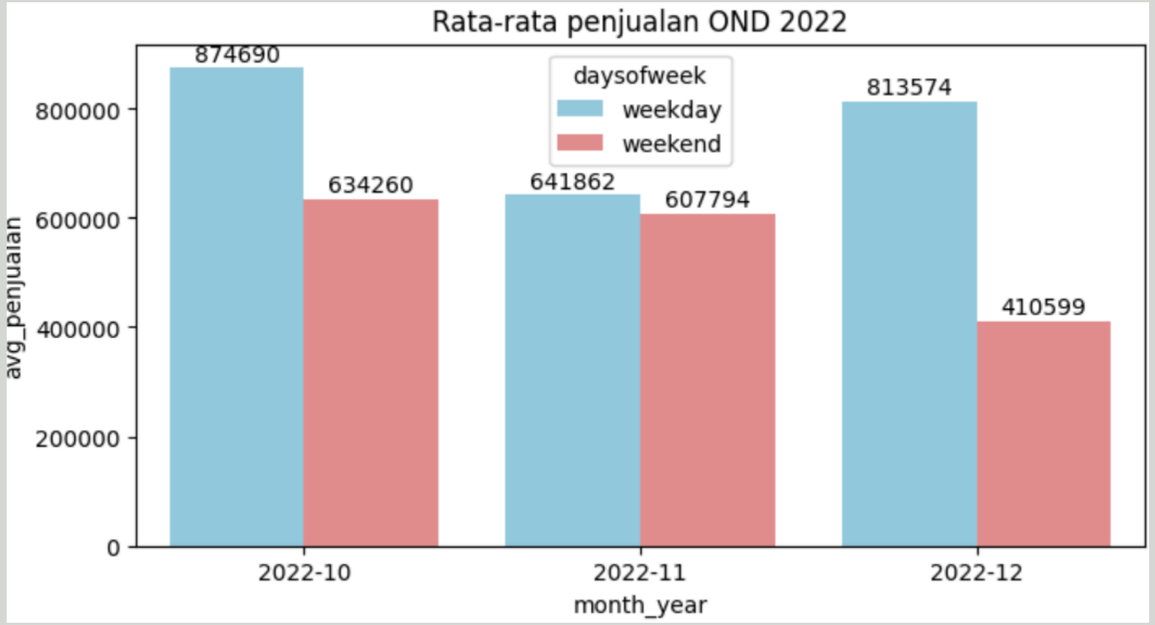
- group by : month_year, dayofweek
- value : average of before_discount
- sort ascending by month_year

```
df4_answer = df_filter.groupby(['month_year','daysofweek'])\
.agg(avg_penjualan=('before_discount','mean'))
df4_answer.sort_values(by=['month_year'], ascending=True)
```

month_year	daysofweek	avg_penjualan
2022-10	weekday	874690.266667
	weekend	634260.074074
2022-11	weekday	641862.000000
	weekend	607794.210526
2022-12	weekday	813574.285714
	weekend	410599.400000

Bar chart of df4_answer

- Oct 2022 is the highest average of sales during 3 months



Python

04.B

df4_answer

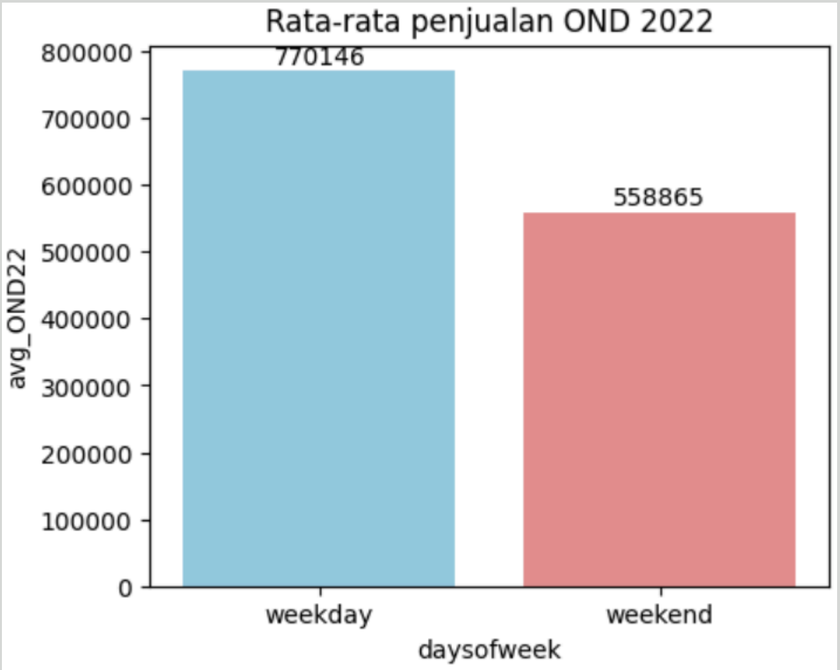
- group by : dayofweek
- value : average of before_discount
- sort ascending by dayofweek

Bar chart of df4_answer

- during OND 2022, weekday obtained higher average of sales than weekend

```
df4_answer = df_filter.groupby(['daysofweek'])\
.agg(avg_OND22=('before_discount','mean'))
df4_answer.sort_values(by=['daysofweek'], ascending=True)
```

daysofweek	avg_OND22
weekday	770146.012048
weekend	558865.151515



Insights

04

01 During the 3 months there tends to be a decline in sales, especially in November. October has the highest average sales.

02 During those 3 months, weekdays had higher average daily sales compared to weekends. Weekday had a sufficient impact for campaign on increasing sales. Many people prefer weekends for leisure and activities, so they might do more online shopping on weekdays for convenience.

03 When weekday sales outperform weekend sales, it can often reflect customer shopping patterns. Businesses often run weekday promotions, such as "Monday deals" or "mid-week specials," to drive sales when consumers might need extra motivation. These promotions can make weekdays an attractive time for shoppers.

THANK YOU!

Connect with me

endahen12@gmail.com

www.linkedin.com/in/endahrakhmawati