

TETRIS PROGRAM

UMKM Di Jawa Timur

Maulana Yusuf Ikhsan Robbani
maulanayusufikhsanrobbani@gmail.com

#StackYourSkill

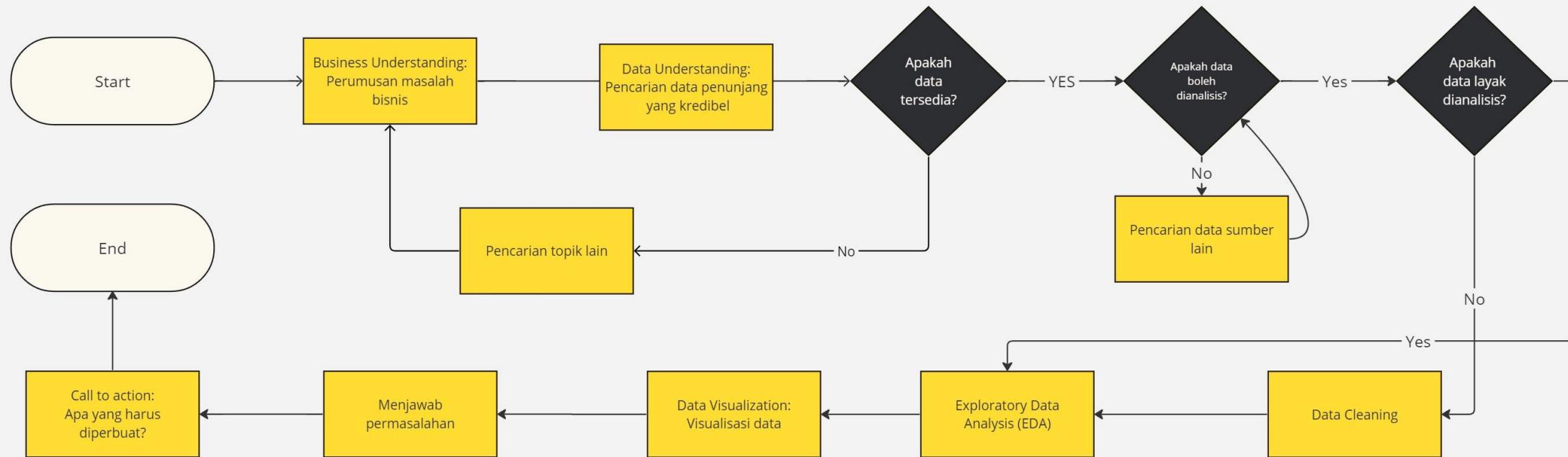




Outline Pertanyaan Penting

1. Daerah mana yang memiliki jumlah usaha terbanyak?
2. Berapa banyak UMKM yang belum pernah mendapatkan kredit?
3. Bagaimana porsi pelaku UMKM berdasarkan jenis kelamin dan kebutuhan pinjaman usaha?
4. Bagaimana korelasi umur dengan omset pelaku usaha di Jawa Timur?
5. Bagaimana korelasi tahun mulai usaha dengan omset ?
6. Bagaimana korelasi gaji karyawan dengan omset?
7. Bagaimana korelasi latar belakang Pendidikan dengan omset pelaku usaha di Jawa Timur?
8. Bagaimana omset usaha tiap daerah?
9. Bagaimana distribusi pelaku UMKM berdasarkan umur, jumlah tenaga kerja, omset biaya produksi, dan gaji karyawan?
10. Bagaimana perubahan omset keseluruhan dari 2019-2021?
11. Seberapa banyak UMKM yang dikategorikan sebagai mikro, ultra mikro, kecil 1, kecil 2?

Langkah-langkah



TETRIS PROGRAM

Step 1 - Data Collection & Data Integration

The image displays three screenshots of the Octoparse software interface, illustrating the process of data collection and integration:

- Screenshot 1:** Shows the 'Data Preview' window for a table on the 'Satu Data - DINAS KOPERASI DAN UKM JATIM' website. It highlights the 'Extract Data' button and a dropdown menu with options like 'Set up pagination', 'Next page button', 'Load more button', 'Infinite scrolling', and 'Scrape linked subpages'.
- Screenshot 2:** Shows the 'New Task' configuration window. A 'Task Flow' diagram is visible, starting with 'Go to Webpage', followed by 'Click Item1' (with a screenshot of a 'DATA UKMK' form), 'Click Item3', and finally 'Extract Data'. The 'Matching XPath' field contains the path: //table[@id='employeeList']/tbody/tr/td/button.
- Screenshot 3:** Shows the task completion screen with a summary: '103 Data Extracted', 'Task completed', 'Duplicates: 0 times', 'Time Spent: 27s', and 'Avg. Speed: 221 lines/min'. A 'Run Completed!' message box is open, stating 'Task: Database UKM', 'Time spent: 27s', 'Data extracted: 103 lines 0 duplicates', and 'Rate your experience?'. Buttons for 'Export Later' and 'Export' are present.

1. Akses aplikasi web scrapping: Octoparse. Dilanjutkan dengan mengakses Alamat website Satu Data Jawa Timur

https://data.diskopukm.jatimprov.go.id/satu_data

2. Buka tab Data UKM, aktifkan opsi data terbaru. Tekan tombol Detail. Kemudian akan muncul table diatas. Pilih kolom tersebut kemudian tekan run.

Website tersebut adalah website open-source dan memiliki data yang valid karena diambil dari banyak acara

3. Setelah selesai, unduh data dengan format csv.

TETRIS PROGRAM

Step 2 - Data Cleansing

The image displays two screenshots illustrating the data cleansing process. On the left, a screenshot of the Dbeaver database interface shows several SQL queries being run against a database named 'dataumkm'. The queries are focused on creating a new table 'data_umkm_kel' from an Excel file, handling categorical and numerical columns, and dealing with multiple years of data. On the right, a screenshot of the Linkr Data Steward interface shows a list of data deduplication results. The table includes columns for ID, Score, Merge status, Confirm status, Date, Name, and Company. The data shows various entries for different individuals and companies, with some entries marked for merging.

Pembersihan data menggunakan Dbeaver dengan menjalankan query yang akan membuat tabel-tabel yang sudah siap untuk dianalisis.

Untuk mengetahui data yang duplikat dengan menggunakan Linkr. Linkr akan memberikan file yang terdiri dari golden record sehingga kita dapat mengekstrak ID yang ingin kita hapus menggunakan query

TETRIS PROGRAM

Step 2 - Data Cleansing

DBeaver 23.3.4 - data_umkm_ke4

File Edit Navigate Search SQL Editor Database Window Help

localhost - localhost:3306

Databases

Tables

Find/Replace

Find: IMADUDDIN

Replace with:

Direction: Forward

Options: Case sensitive, Whole word, Incremental, Regular expressions

Grid

L	RBC KAB_KOTA	RBC KECAMATAN	RBC JENIS_KEGIATAN	RBC NAMA PEMILIK	RBC NAMA_U
1728	8.221 [18] NGANJUK	TANJUNGANOM	KONVEksi,	AFTONUR ROSYAD	KOPERASI KO
1729	18.028 [73] KOTA MALANG	LOWOKWARU	-	SABARUDIN BASO, S.H.	KOPERASI KO
1730	22.440 [07] MALANG	LOWOKWARU	-	SABARUDIN BASRO, S.H.	KOPERASI KO
1731	12.809 [08] LUMAJANG	SUKODONO	TELUR ASIN ASAP	ULUMUDDIN	KOPERASI KO
1732	21.528 [73] KOTA MALANG	KEDUNGKANDANG	BERDAGANG	AHMAD SHALIHIN	KOPERASI KO
7.494	[73] KOTA MALANG	KEDUNGKANDANG	MAUL HAYAT, ALHAYAT MART, VERMIHAYATI	SHALIHIN, S.P.D.I, MM	KOPERASI KO
8.170	[15] SIDOARJO	GEDANGAN	PERDAGANGAN RETAIL	PURWOKO	KOPERASI KO
8.169	[78] KOTA SURABAYA	WONOCOLO	PERDAGANGAN BESAR DAN ECERAN - REP MC DIDIK SISWANTO	KOPERASI KO	
13.106	[15] SIDOARJO	WARU	BENGKEL MOBIL	ACHMAD NUR FALAKHUDIN	KOPERASI KO
21.304	[15] SIDOARJO	WARU	JUAL BELI SPAREPART MOBIL	MAKFUD	KOPERASI KO
17.089	[19] MADIUN	MADIUN	SIMPAN PINJAM	ILHAM IMADUDDIN	KOPERASI KO
8.299	[19] MADIUN	WONOASRI	SIMPAN PINJAM SYARIAH	ILHAM IMADUDDIN	KOPERASI KO
25.865	[19] MADIUN	GEGER	SIMPAN PINJAM PEMBIAYAAN SYARIAH	BPK ILHAM IMADUDDIN, S.E, MM	KOPERASI KO
21.450	[19] MADIUN	SARADAN	SIMPAN PINJAM SYARIAH	ILHAM IMADUDDIN	KOPERASI KO
8.287	[19] MADIUN	BALEREO	PEMBIAYAAN SYARIAH	ILHAM IMADUDDIN	KOPERASI KO
8.293	[19] MADIUN	MEJAYAN	SIMPAN PINJAM SYARIAH	ILHAM IMADUDDIN SEST.MM	KOPERASI KO
21.461	[19] MADIUN	DAGANGAN	SIMPAN PINJAM SYARIAH	ILHAM IMADUDDIN, S.E, MM	KOPERASI KO
8.294	[19] MADIUN	JIWAN	SIMPAN PINJAM SYARIAH	ILHAM IMADUDDIN, S.E, MM	KOPERASI KO
17.081	[19] MADIUN	MADIUN	SIMPAN PINJAM	ILHAM IMADUDDIN, S.E, MM	KOPERASI KO
17.079	[19] MADIUN	PILANGKENCENG	SIMPAN PINJAM SYARIAH	ILHAM IMADUDDIN, S.E., MM,,	KOPERASI KO
21.462	[19] MADIUN	KEBONSARI	SIMPAN PINJAM SYARIAH	ILHAM IMADUDDIN, SE, MM,	KOPERASI KO
8.302	[19] MADIUN	DOLOPO	KOPERASI SIMPAN PINJAM SYARIAH	ILHAM IMADUDDIN, SE, MM,	KOPERASI KO
17.095	[19] MADIUN	GEGER	SIMPAN PINJAM SYARIAH	ILHAM IMADUDDIN, SE,MM,	KOPERASI KO

Namun, ditemukan data yang masih terlihat kembar. Sehingga memerlukan data deduplicating dengan software lanjut Openrevine.

Cluster and edit column "NAMA PEMILIK"

Method: Nearest neighbor

Distance function: Levenshtein

Radius: 1.0

Block chars: 6

23 clusters found

Choices in cluster	Rows in cluster	Average length of choices	Length variance of choices
SAMININGSIH (3 rows)	2 — 4	2 — 13	0 — 0.5
SARININGSIH			
GUSMIATI (4 rows)			
SUSMIATI			
SRI MURJIATI (2 rows)			
SRI MURNIATI			
EKO WAHYUNI (3 rows)			
EKO WAHYUDI			
SUHARTINI (5 rows)			
SUPARTINI			
SUMARTIN (3 rows)			
SUMARTI (2 rows)			
NUR HASANAH (9 rows)			
NUR CHASANAH (2 rows)			
NUR HASANAH			

Select all Deselect all Export clusters Merge selected & re-cluster Merge selected & Close Close

Pada openrevine terdapat banyak algoritma klasterisasi dan memungkinkan pengguna memilih data mana yang akan jadi golden record. Sehingga, data yang akan dihasilkan murni tanpa adanya data kembar

TETRIS PROGRAM

Step 2 - Data Cleansing

DBeaver 23.3.4 - <localhost> CapstoneProject-DataCleaning

File Edit Navigate Search SQL Editor Database Window Help

localhost - localhost:3306

Databases

dataumkm

Tables

Views

Indexes

Procedures

Triggers

Events

minimart

ms_people

sakila

sys

training

world

Users

Administrator

System Info

postgres - localhost:5432

<localhost> ... <localhost> ... ms_harga_harian ... <localhost> ... data_umkm_ke... localhost data_umkm_ke... 31

```
'RETRNO SETYOWATI', 'REVIN RADIANA', 'REZA AULIA RACHMAN', 'RIKA ANGELINA', 'RIMA LOESTANTI',  
'RINI NURUL INDAWATI', 'RINI SUSANTI', 'RINI WIDIASTUTI', 'RIRIN SUSANTI', 'RISKA', 'RISKI FAHRUNNISA',  
'RISVICO WAHYU SURKAWATI', 'RITA SUMALA', 'RITA WIDJAYANTI', 'ROBIATUN', 'RODIYAH AGUSTINA',  
'RODIAH DAMAYANTI', 'RODLIYAH HABIBI M', 'ROMLAR', 'ROMLAN', 'RR.TIWUNG DYAN EKAWATI', 'RUSTIKA DEWI WIJAYANTI',  
'SABARUDIN BASO', 'S.H.', 'SAMINTINGSIH', 'SAMSI MA'ARIF', 'SANT RACHMAWATI', 'SANTOSO', 'SARINGNINGSHI',  
'SATRIO RACHMAD', 'SHINTA NIAWATI', 'SHINTA WIDYA R', 'SIFWAIR RIF'AH', 'SILVIA IRMAWATI', 'SISTYANA',  
'SITI AISAH', 'SITI AMINAH', 'SITI CHOLIMAH', 'SITI FATIMAH', 'SITI MELYSA', 'SITI MUTAMIMMAH',  
'SITI NGAIASAH', 'SITI NGAIASAH', 'SITI NORCHOLIFAH', 'SITI NUR ARIFAH', 'SITI QOWIJULANA', 'SITI RUBIAINTI',  
'SOLICHATUR ROCIMAH', 'SOVIA YULIA ASTUTI', 'SRI AGUSTINI', 'SRI ASTUTIK', 'SRI MARIYANI', 'SRI MURJIATI',  
'SRI RAHAYU', 'SRI RAHAYU SETYOWATI', 'SRI WARYUNI', 'SUBUR', 'SUDARWI YULININGSIH', 'SUGIANTO', 'SUGIARTO',  
'SUKADI', 'SULISTYOWATI', 'SUMAIYAH', 'SUMARTI', 'SUMARTIN', 'SUPRIYAN ARTI', 'SUPRAPTI', 'SUSI HENDRAWATI',  
'SUSA INDAH NOVANTI', 'SUSIANA', 'SUTATIK', 'SUWARTO', 'SUWANTO', 'SUYONO S.E', 'SYIFA'UDDIN',  
'SYIFAU'L MUTTAQIN', 'TAFRIKHAN', 'TATIK MUJIAINTINGSIH', 'TITIN IRAWATI', 'TITING WULANDARI',  
'TIWUK HERMANAWATI', 'TUTI KUSMIAH', 'TUTIK RAHAYU', 'ULFA ARINALISTYANI', 'ULFA NURANI',  
'ULIA HAQUE USSIANTINI', 'UMI SALAMAH', 'UMYATI', 'UNITA ROSALINI', 'VERONIKA DWI SULISTYANI',  
'VITA AGUSTRIANA', 'SAPUTRI', 'WAHYU WISMA PRIYANI', 'WAHYUDI', 'WARLIATI WARIS', 'WASTRI PUTRI KARTININGTYAS',  
'WENDI HERMANTO', 'WIDODO', 'WINARSIH', 'WISANG R WIJAYA S. PSI', 'YENI IRAWATI', 'YENNI TRI ASTUTI',  
'YOSIB ERMANW', 'YULIANA', 'YUNANTO TRIATMOKO', 'YUNI ASTUTI', 'YUNI SUWANTI ASIH', 'ZAENAL ABIDIN',  
'ZAINAL EFENDI', 'ZAMRONI');
```

#Menghapus baris-baris menyesuaikan ID_DT_BINAAN_EXCEL yang sudah dicatat dalam notepad saya

```
delete from data_umkm_ke4_openrevine  
where ID_DT_BINAAN_EXCEL in (  
8983, 19818, 25465, 25790, 7174, 18253, 20114, 23201, 11627, 15368, 7694, 24190, 16601, 13297, 10728, 11396, 9637, 12576, 25857  
);
```

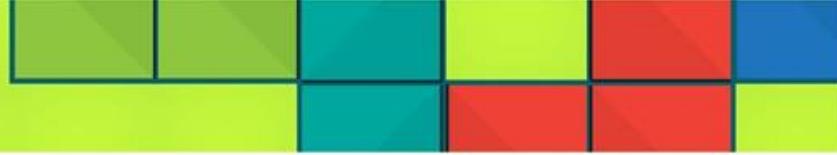
#Hasilnya adalah 3446 data 'unik' dan siap untuk di ekstrak sebagai bahan Exploratory Data Analysis di Colab

data.umkm.ke4.openrevine 1

ID_DT_BINAAN_EXCEL KAB_KOTA KECAMATAN JENIS_KEGIATAN NAMA_PEMILIK Value

ID_DT_BINAAN_EXCEL	KAB_KOTA	KECAMATAN	JENIS_KEGIATAN	NAMA_PEMILIK	Value
17,346	[19] MADIUN	WUNGU	TATIK PUDJIASTUTI SI		17346
14,580	[05] BLITAR	SANANWETAN	SIMPAN PINJAM	-	
14,700	[77] KOTA MADIUN	MANGUHARJO	PEMBUATAN VIDEO	-	
15,761	[13] PROBOLINGGO	MARON	TOKO BAJU	-	
16,410	[02] PONOROGO	BABADAN	MEMBUAT BUNGA IKAT	-	
16,422	[20] MAGETAN	SUKOMORO	-	-	
14,666	[19] MADIUN	-	-	-	
15,688	[22] BOJONEGORO	PADANGAN	PERDAGANGAN	ABDUL GHOFUR	
14,632	[71] KOTA KEDIRI	MOJOROTO	PEMBIAYAAN PERALATAN USAHA ANGGOTA	ADI SURYANTO	
15,728	[14] PASURUAN	BANGIL	KATERING	ALFIA RIZKY FIRDAUSI	

Akhirnya, kita mendapatkan table yang sudah siap untuk diproses dalam tahap Exploratory Data Analysis



Step 3 - Data Exploration & Data Visualisation

The screenshot shows a Google Colab notebook interface. The title bar reads "Data Analysis Project - UMKM Jawa Timur". The left sidebar contains a search bar, a list of recent notebooks, and a collapsed section titled "Penjelasan mengenai dataset". The main content area displays text about the dataset and its source, followed by a code cell starting with "#Dimulai dengan import library dari python".

Data yang digunakan diambil pada website Pemprov Jawa Timur yang mana dalam data tersebut menampilkan beragam informasi penting dari setiap responden mulai dari nama usaha, tempat usaha, omset, sampai kontak berupa email dan nomor telepon.

Berikut website yang dijadikan sumber data utama

https://data.diskopukm.jatimprov.go.id/satu_data/

```
[ ] #Dimulai dengan import library dari python
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
```

Exploratory Data Analysis dilakukan dengan menggunakan Google Colab

https://colab.research.google.com/drive/1h4u8PMjPRJxwa3E7pFf77yL5_qCVpyF8?usp=sharing

TETRIS PROGRAM

Step 3 - Data Exploration & Data Visualisation

The screenshot shows a Jupyter Notebook interface with the title "Data Analysis Project - UMKM Jawa Timur". The code cell contains the following Python code:

```
[ ] #Membaca dokumen csv
tabel = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/Capstone Project Tetris/data_umkm_ke4_openrevine_202402171115.csv')
tabel
```

The resulting table displays data from the CSV file, including columns such as ID_DT_BINAAN_EXCEL, KAB_KOTA, KECAMATAN, JENIS_KEGIATAN, NAMA_PEMILIK, NAMA_USAHA, JENIS_KELAMIN, PENDIDIKAN, STATUS_USAHA, JENIS_LAPANGAN_USAHA, OMSET_2020, and OMSET_2021. The data includes various business details like location, activity type, and financial performance.

Pertama-tama kita lakukan import csv file yang sudah kita lakukan data cleaning. File tersebut diupload pada google drive pribadi.

TETRIS PROGRAM

Step 3 - Data Exploration & Data Visualisation



```
[ ] # membagi kolom numerik dan kolom kategorik yang penting sebagai bahan analisis
numerik = ['ID_DT_BINAAN_EXCEL', 'TAHUN_MULAT_USAHA', 'UMUR',
           'JML_TENAGAKERJA_2019', 'JML_TENAGAKERJA_2020',
           'JML_TENAGAKERJA_2021', 'OMSET_2019', 'OMSET_2020',
           'OMSET_2021', 'BIAYA_PRODUKSI_2019', 'BIAYA_PRODUKSI_2020',
           'BIAYA_PRODUKSI_2021', 'GAJI_KARYAWAN_2019',
           'GAJI_KARYAWAN_2020', 'GAJI_KARYAWAN_2021', 'score', 'row_num']

kategorikal = ['KAB_KOTA', 'KECAMATAN', 'JENIS_KEGIATAN', 'NAMA PEMILIK',
               'NAMA_USAHA', 'JENIS_KELAMIN', 'PENDIDIKAN', 'STATUS_USAHA',
               'JENIS_LAPANGAN_USAHA', 'SERTIFIKASI', 'Klasifikasi_USAHA',
               'DAPAT_KREDIT', 'PERLU_PINJAMAN_PIHAK_LUAR']

[ ] tabel[numerik].describe()

  ID_DT_BINAAN_EXCEL TAHUN_MULAT_USAHA UMUR JML_TENAGAKERJA_2019 JML_TENAGAKERJA_2020 JML_TENAGAKERJA_2021 OMSET_2019 OMSET_2020 OMSET_2021 BIAYA_PRODUKSI_2019 BIAYA_PRODUKSI_2020 BIAYA_PRODUKSI_2021
  count      3388.000000    3300.000000 3310.000000     3376.000000    3375.000000   3374.000000 2.851000e+03 3.028000e+03 3.075000e+03 2.672000e+03
  mean       15188.284238   2011.812424 40.816918      6.624111     6.628741    7.067279 1.299647e+08 1.209466e+08 4.861800e+07 6.786672e+07
  std        6250.492033    50.565450 11.117372     32.016875    25.823806   26.012972 2.976720e+08 2.977378e+08 1.521429e+08 1.975291e+08
  min         4301.000000    0.000000 3.000000      0.000000     0.000000    0.000000 1.084900e+04 1.155800e+04 1.140900e+04 1.010700e+04
  25%        9820.500000   2010.000000 33.000000     1.000000     1.000000    2.000000 9.000000e+06 6.000000e+06 2.000000e+06 4.000000e+06
  50%        15224.000000   2017.000000 41.000000     2.000000     2.000000    3.000000 3.000000e+07 2.300000e+07 5.000000e+06 1.100000e+07
  75%        20465.500000   2019.000000 49.000000     5.000000     5.000000    5.000000 1.000000e+08 8.000000e+07 2.500000e+07 4.000000e+07
  max        25984.000000   2104.000000 82.000000     978.000000    672.000000   672.000000 2.147484e+09 2.147484e+09 2.147484e+09 2.147484e+09

  [ ] tabel[kategorikal].describe()

  KAB_KOTA KECAMATAN JENIS_KEGIATAN NAMA PEMILIK NAMA_USAHA JENIS_KELAMIN PENDIDIKAN STATUS_USAHA JENIS_LAPANGAN_USAHA SERTIFIKASI Klasifikasi_USAHA DAPAT_KREDIT PERLU_PINJAMAN_PIHAK_LUAR
  count      3388      3388      3387      3388      3387      3388      3388      3388      3388      3388      1294      3388      3388
  unique       40        709      2319      2760      2966          2          6          9          15          32          7          2
  PERSEORANGAN                               ULTRA MIKRO:
```

Kita lakukan pendefinisian kolom kategorik yang berisi data non-numerik serta kolom kategorik yang berisi angka-angka

TETRIS PROGRAM

Step 3 - Data Exploration & Data Visualisation

```
▶ # density plot untuk kolom-kolom numerik
plt.figure(figsize=(24,24))

features = numerik
for i in range(1, len(features)):
    plt.subplot(6, len(features)//6 + 1, i+1)
    sns.kdeplot(x=tabel[features[i]], color='skyblue')
    plt.xlabel(features[i])
    plt.grid(True)
    plt.tight_layout()
```

Kita telusuri distribusi data numerik.

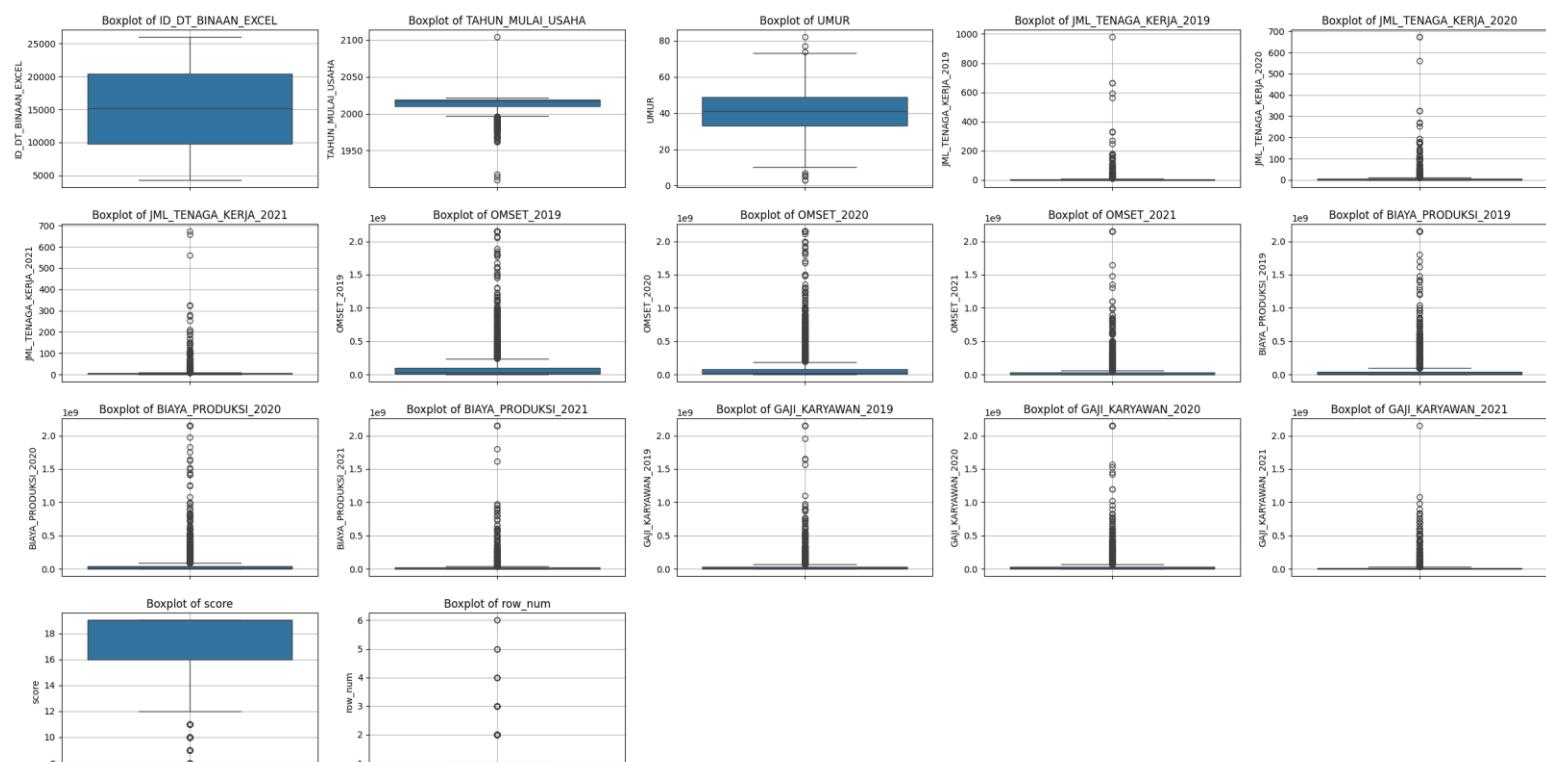
1. Terlihat bahwa banyak kolom yang memiliki data tail (data yang ekstrim).
2. Contoh seperti jumlah tenaga kerja yang tembus sejumlah lebih dari 400 orang. Contoh lain adalah tahun mulai usaha yang lebih dari tahun 2019
3. Jika data yang ekstrim dihapus maka distribusi seluruh kolom mendekati distribusi normal



Step 3 - Data Exploration & Data Visualisation

```
plt.figure(figsize=(24,12))
for i in range(len(numerik)):
    #plt.title('Boxplot of '+tabel[numerik[i]], fontsize=12)
    plt.subplot(4, len(numerik)//4+1, i+1)
    sns.boxplot(y=tabel[numerik[i]])
    plt.title('Boxplot of '+numerik[i], fontsize=12)
    plt.grid(True)

plt.tight_layout()
#plt.xticks(rotation=90)
plt.show()
```



Dilakukan visualisasi boxplot untuk mengetahui outlier

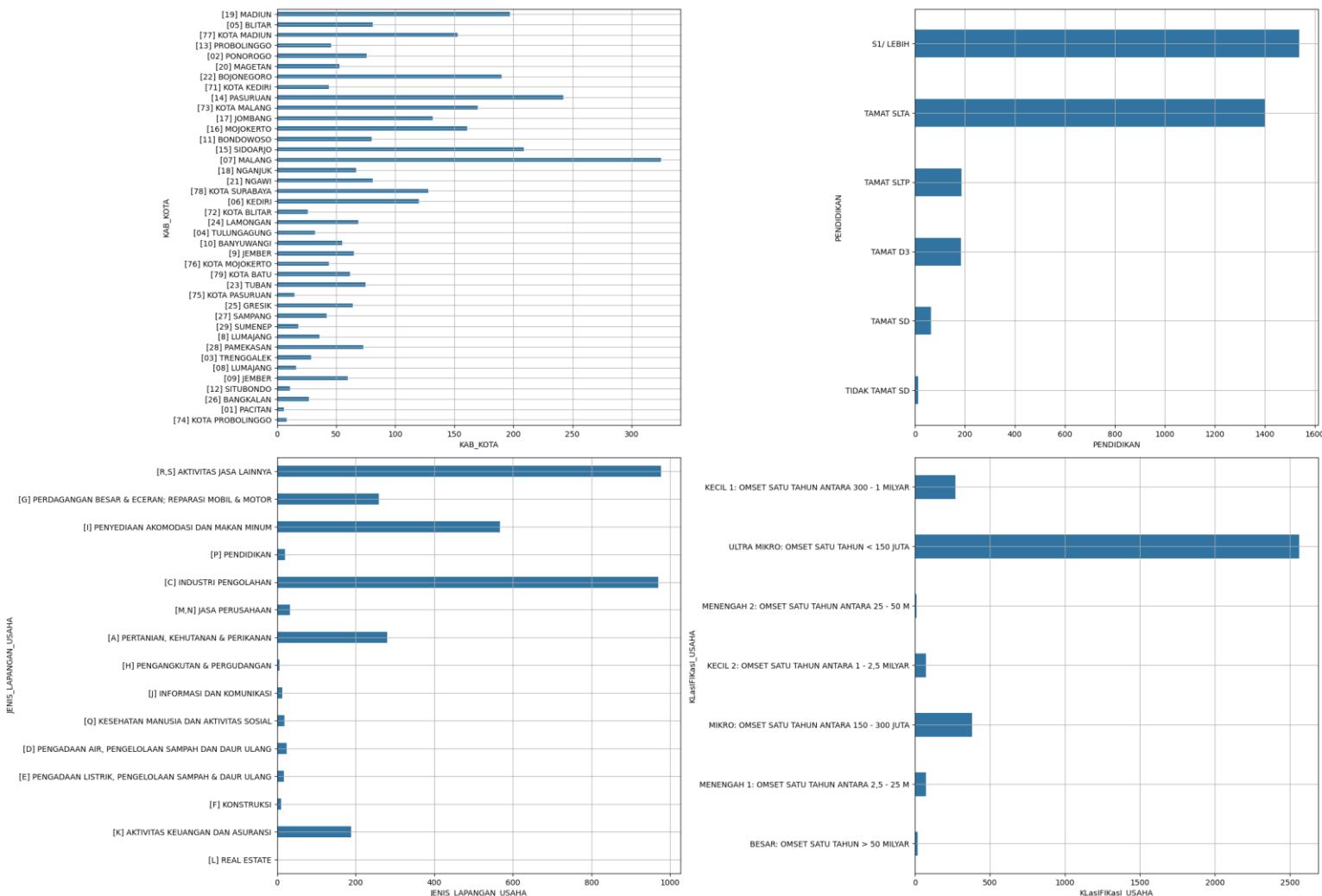
Terlihat bahwa banyak outlier yang memiliki jarak yang cukup jauh dibandingkan mayoritas data yang terpusat untuk semua kolom
(Ditandai dengan banyaknya titik-titik)

TETRIS PROGRAM

Step 3 - Data Exploration & Data Visualisation

```
#Kelompok kan kolom yang penting untuk dihitung
features = ['KAB_KOTA' , 'PENDIDIKAN',
            'JENIS_LAPANGAN_USAHA', 'Klasifikasi_USAHA'
        ]

plt.figure(figsize=(24,32))
for i in range(0, len(features)):
    plt.subplot(4, len(features)//4 + 1, i+1)
    sns.countplot(y=tabel[features[i]],width = 0.4 )
    #plt.xticks()
    plt.grid(True)
    plt.xlabel(features[i])
    plt.tight_layout()
```





Step 3 - Data Exploration & Data Visualisation

Data Kabupaten/Kota

Terlihat bahwa kabupaten Malang menjadi kabupaten/kota dengan UMKM terdata terbanyak di Jawa Timur dengan jumlah 300 lebih. Sedangkan kabupaten Pacitan yang paling sedikit dibandingkan kabupaten/kota lain

Data Latar Belakang Pendidikan

Pelaku UMKM di Jawa Timur didominasi oleh lulusan S1/lebih sejumlah 1500 lebih. Sedangkan porsi pelaku yang tidak tamat SD adalah yang paling sedikit

Data Jenis Lapangan Usaha

Pelaku usaha kebanyakan terpusat pada industri pengolahan. Sedangkan pengangkutan dan pergudangan yang paling kecil jumlahnya

Data Klasifikasi Usaha

UMKM terbagi beragam macam kategori tergantung dari banyaknya omset yang dimiliki. UMKM dengan omset per tahun kurang dari 150 juta rupiah mendominasi. Sedangkan UMKM beromset menengah yaitu 25 M - 50 M yang paling sedikit. Sedikit sekali pengusaha yang memiliki omset lebih dari 2,5 M bahkan semakin besar omsetnya, semakin sedikit jumlahnya.

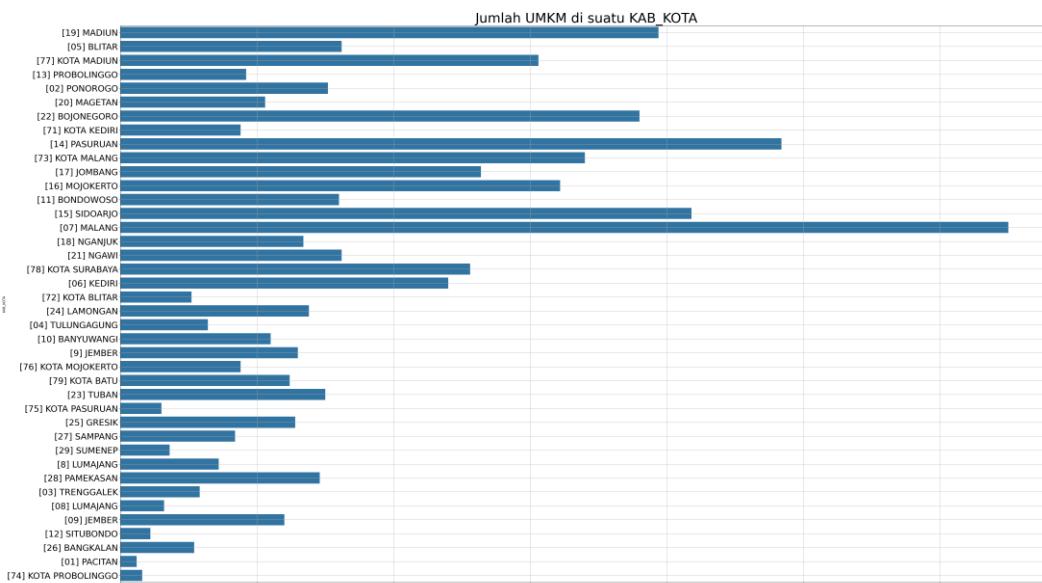
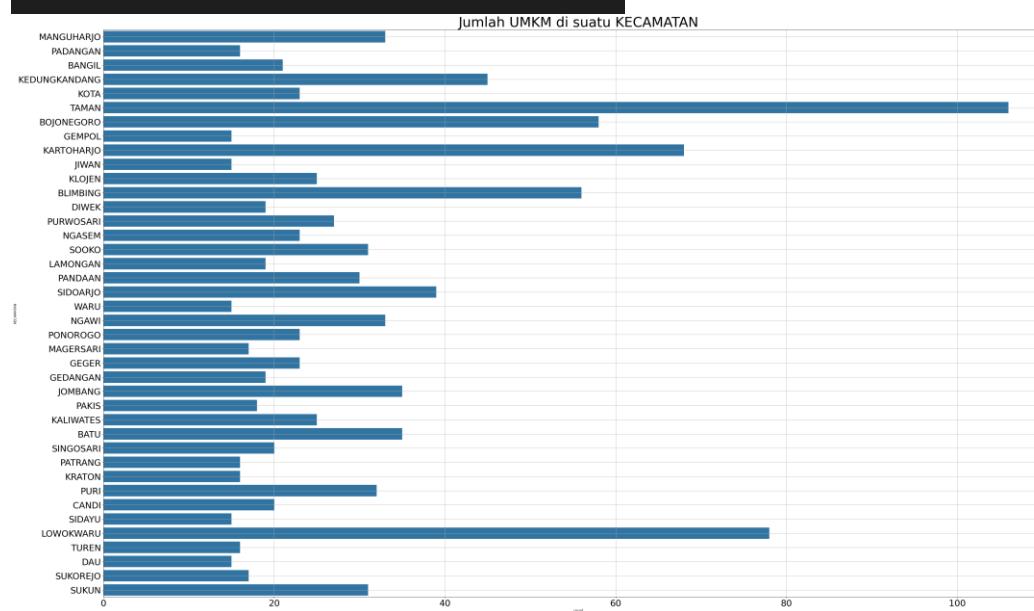
Perlu diperhatikan bahwa kriteria UMKM adalah memiliki omset per tahun sebanyak 500 juta atau kurang. Sehingga dimungkinkan ada responden tidak relevan dikatakan UMKM.

Step 3 - Data Exploration & Data Visualisation

Permasalahan 1:

Daerah mana yang memiliki jumlah usaha terbanyak?

```
feature_1 = ['KAB_KOTA', 'KECAMATAN']
for i in range (0,len(feature_1)):
    plt.figure(figsize=(50, 30))
    if feature_1[i] == 'KECAMATAN':
        top_40_kecamatan = tabel['KECAMATAN'].value_counts().index[:40]
        data_to_plot = tabel[tabel['KECAMATAN'].isin(top_40_kecamatan)]
    else:
        data_to_plot = tabel
    sns.countplot(data=data_to_plot, y=data_to_plot[feature_1[i]])
    plt.title('Jumlah UMKM di suatu '+feature_1[i], fontsize=40)
    plt.grid(True)
    plt.yticks(fontsize = 26)
    plt.xticks(fontsize = 26)
    plt.show()
```



Step 3 - Data Exploration & Data Visualisation

▼ Permasalahan 2:

Berapa banyak UMKM yang belum pernah mendapatkan kredit?

```
[ ] count = tabel['DAPAT_KREDIT'].value_counts()['TIDAK']
print(f"Jumlah UMKM yang tidak mendapatkan kredit adalah {count}")
```

Jumlah UMKM yang tidak mendapatkan kredit adalah 2389

Bagaimana porsi pelaku UMKM berdasarkan jenis kelamin?

```
❶ dual_features = ['DAPAT_KREDIT', 'PERLU_PINJAMAN_PIHAK_LUAR', 'JENIS_KELAMIN']
plt.figure(figsize=(16,8))

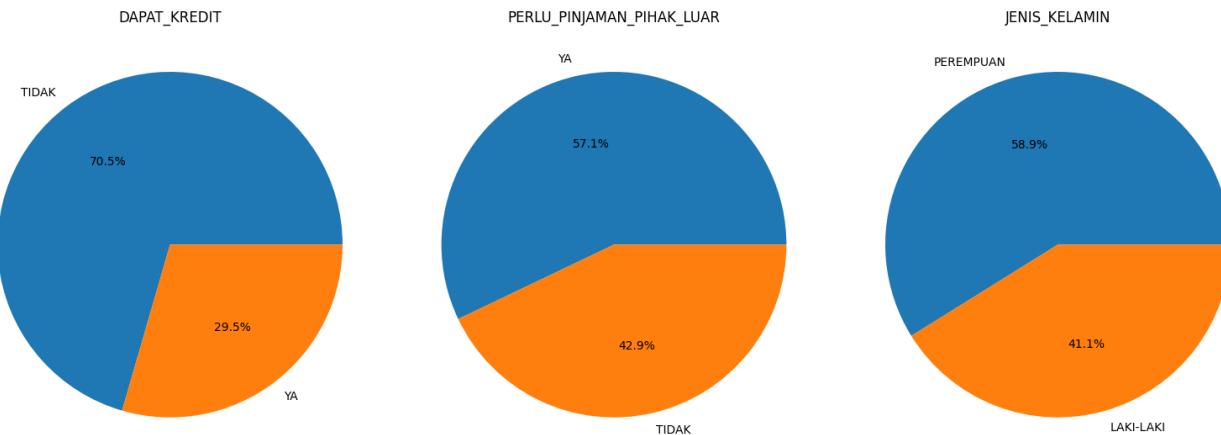
for i in range(0, len(dual_features)):
    plt.subplot(1, len(dual_features), i+1)

    # Get the counts of each unique value in the current feature
    counts = tabel[dual_features[i]].value_counts()

    # Create a pie chart of the counts
    plt.pie(counts, labels=counts.index, autopct='%.1f%%')

    plt.title(dual_features[i])
    plt.tight_layout()

plt.show()
```

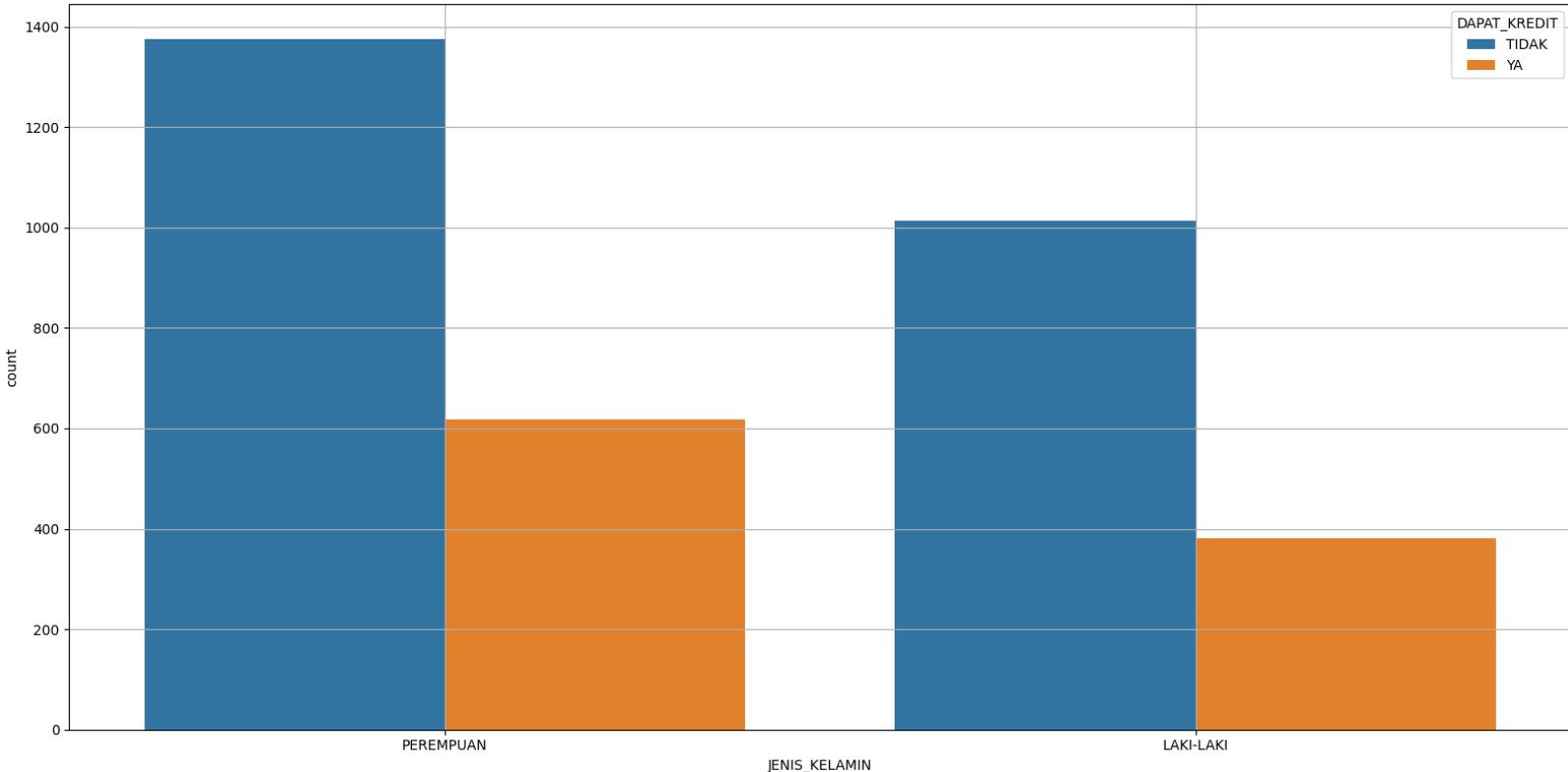


Ternyata pelaku UMKM di Jawa Timur didominasi oleh jenis kelamin perempuan yang mana porsinya sebesar 58.9% atau lebih dari setengah pelaku UMKM
Setengah pelaku UMKM sebesar 57.1% nya ternyata membutuhkan pinjaman usaha



Step 3 - Data Exploration & Data Visualisation

```
plt.figure(figsize=(16,8))
sns.countplot(x=tabel['JENIS_KELAMIN'], hue=tabel['DAPAT_KREDIT'])
plt.grid(True)
plt.xlabel('JENIS_KELAMIN')
plt.tight_layout()
plt.show()
```



Ternyata, jumlah pelaku UMKM yang perempuan dan tidak dapat kredit usaha sebanyak 1300 an usaha. Bisa dikatakan, 2 dari 3 pelaku UMKM wanita tidak mendapatkan kredit usaha

Step 3 - Data Exploration & Data Visualisation

Permasalahan 3:

1. Bagaimana korelasi umur dengan omset ?
2. Bagaimana korelasi tahun mulai usaha dengan omset ?
3. Bagaimana korelasi jumlah tenaga kerja dengan omset ?
4. Bagaimana korelasi biaya produksi dengan omset ?
5. Bagaimana korelasi gaji karyawan dengan omset ?

Ini akan digunakan scatter plot dan heat map untuk menunjukkan korelasi antar data numerik

The screenshot shows a Jupyter Notebook cell with Python code for calculating average values across multiple years for various metrics. Below the code is a partial view of a Pandas DataFrame with columns including ID, location, and various demographic and economic variables.

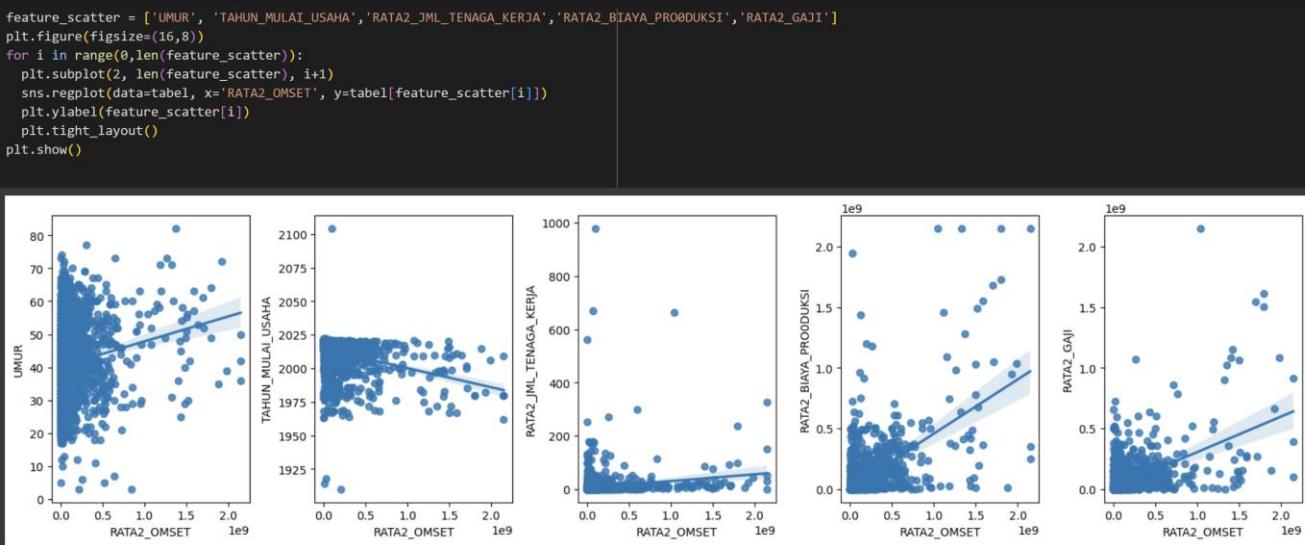
ID	DT_BINAAN_EXCEL	KAB_KOTA	KECAMATAN	JENIS_KEGIATAN	NAMA PEMILIK	NAMA_USAHA	JENIS_KELAMIN	PENDIDIKAN	STATUS_USAHA	JENIS_LAPANGAN_USAHA	...
0	17346	[19] MADIUN	WUNGU	NaN	PUDJIASTUTI, SE, MM	NaN	PEREMPUAN	S1/ LEBIH	KOPERASI	[R,S] AKTIVITAS JASA LAINNYA	...
1	14580	[05] BLITAR	SANANWETAN	SIMPAN PINJAM	-	-	PEREMPUAN	TAMAT SLTA	KOPERASI	[R,S] AKTIVITAS JASA LAINNYA	...
2	14700	[77] KOTA MADIUN	MANGUHARJO	PEMBUATAN VIDEO	-	-	LAKI-LAKI	TAMAT SLTA	PERSEORANGAN (BELUM ADA LEGALITAS)	[R,S] AKTIVITAS JASA LAINNYA	...
3	15761	PROBOLINGGO	[13]	MARON	TOKO BAJU	-	-	LAKI-LAKI	TAMAT SLTP	PERSEORANGAN (BELUM ADA LEGALITAS)	[G] PERDAGANGAN BESAR & ECERAN; REPARASI MOBIL...
4	16410	PONOROGO	[02]	BABADAN	MEMBUAT BUNGA IKAT	-	-	PEREMPUAN	TAMAT SLTA	PERSEORANGAN (BELUM ADA LEGALITAS)	[R,S] AKTIVITAS JASA LAINNYA
...
3383	10435	[17] JOMBANG	JOGOROTO	MEMBUAT KRIPIK PISANG KEPOK, SUKUN, TALAS	ULFA NURANI	ZIRI MAKMUR	PEREMPUAN	S1/ LEBIH	UD	[C] INDUSTRI PENGOLAHAN	...
3384	5734	PONOROGO	[02]	PONOROGO	BATIK	TUMINI	ZOMALI BATIK	PEREMPUAN	TAMAT D3	PERSEORANGAN (BELUM ADA LEGALITAS)	[C] INDUSTRI PENGOLAHAN

Tahap ini dilakukan feature engineering. Feature engineering adalah proses membuat data/kolom baru dari hasil olah data yang sudah ada.

Dibuatlah kolom baru yaitu:
Rata-rata jumlah tenaga kerja,
rata-rata biaya produksi, rata-rata
gaji, rata-rata omset

Step 3 - Data Exploration & Data Visualisation

```
feature_scatter = ['UMUR', 'TAHUN_MULAI_USAHA', 'RATA2_JML_TENAGA_KERJA', 'RATA2_BIAYA_PRODUKSI', 'RATA2_GAJI']  
plt.figure(figsize=(16,8))  
for i in range(0,len(feature_scatter)):  
    plt.subplot(2, len(feature_scatter), i+1)  
    sns.regplot(data=tabel, x='RATA2_OMSET', y=tabel[feature_scatter[i]])  
    plt.ylabel(feature_scatter[i])  
    plt.tight_layout()  
plt.show()
```



Terlihat bahwa korelasi umur dan jumlah tenaga kerja tidak berkorelasi besar terhadap rata-rata omset suatu usaha. Namun, rata-rata biaya produksi dan rata-rata gaji karyawan cukup berkorelasi terhadap rata-rata omset dari suatu usaha.

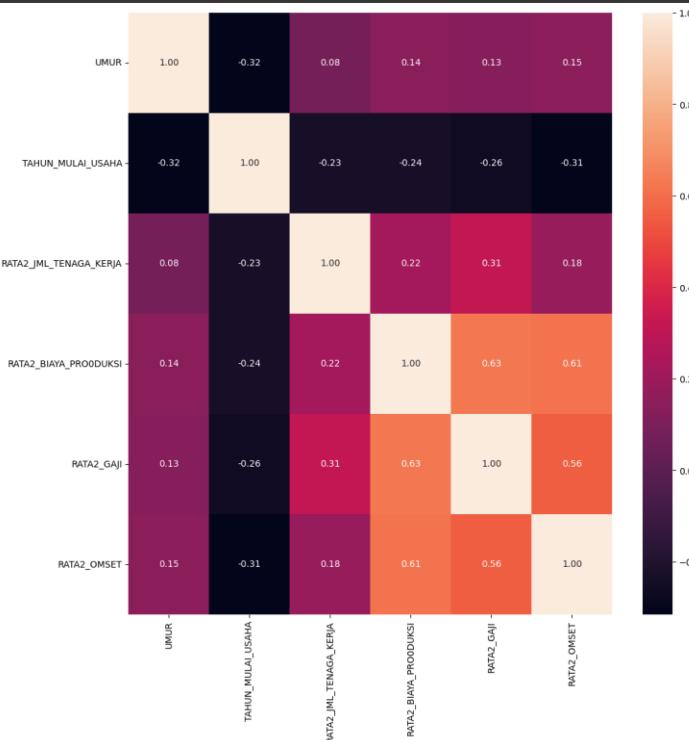
Artinya, semakin besar omset suatu usaha dimungkinkan semakin besar gaji karyawan dan biaya produksi dari suatu usaha tersebut. Dapat dikatakan untuk menilai usaha beromset besar atau tidak dapat melihat dari besar kecilnya gaji karyawan atau besar kecilnya biaya produksi tersebut.

Malah, jumlah tenaga kerja dan umur pelaku usaha tidaklah berpengaruh besar terhadap omset suatu usaha

Mari kita telusuri bagaimana umur, jumlah tenaga kerja, biaya produksi, gaji dan omset berkorelasi satu sama lain dengan heatmap.

Step 3 - Data Exploration & Data Visualisation

```
feature_scatter2 = tabel[['UMUR', 'TAHUN_MULAI_USAHA','RATA2_JML_TENAGA_KERJA','RATA2_BIAYA_PRODUKSI','RATA2_GAJI', 'RATA2_OMSET']]  
plt.figure(figsize=(12,12))  
correlation = feature_scatter2.corr()  
sns.heatmap(correlation, annot=True, fmt=".2f")
```



Terlihat bahwa :

rata-rata gaji karyawan dengan rata-rata biaya produksi,
rata-rata gaji karyawan dengan rata-rata omset,
rata-rata biaya produksi dengan rata-rata omset,
berkorelasi positif lebih dari 0.5 dan dikatakan sebagai korelasi sedang.
Artinya, semakin tinggi salah satu variabel tersebut maka semakin tinggi variabel lain.

Namun berbeda dengan variabel lain yang mana menunjukkan nilai korelasi dibawah 0.5 atau bahkan minus.

Artinya, variabel seperti tahun mulai usaha, dan rata-rata jumlah tenaga kerja tidak benar-benar mempengaruhi variabel keseluruhan



Step 3 - Data Exploration & Data Visualisation

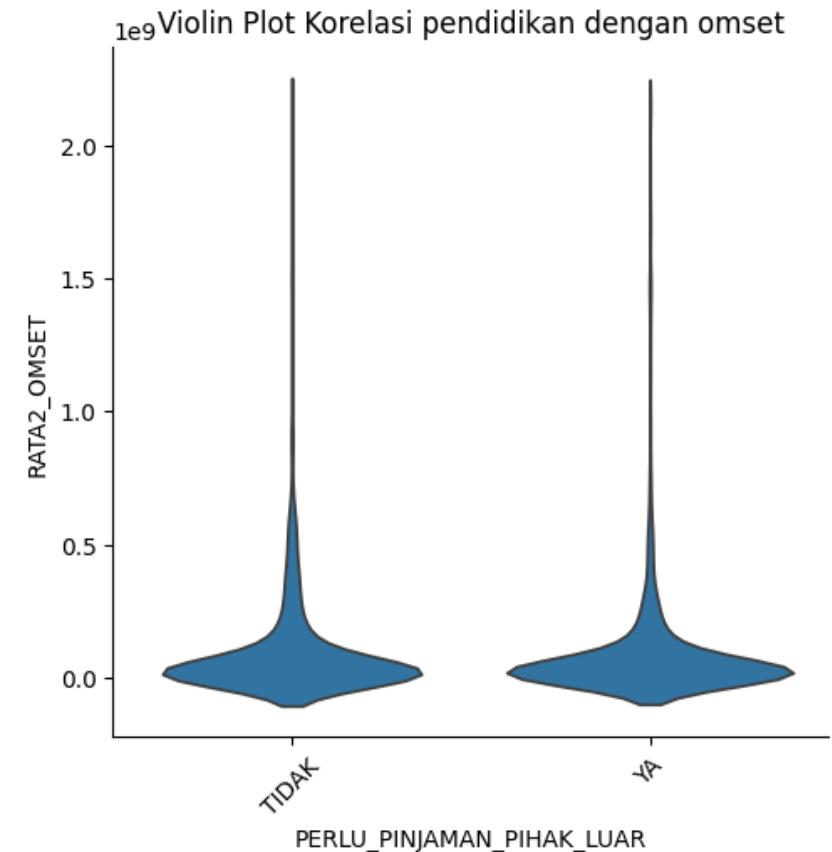
Permasalahan 4:

Bagaimana korelasi latar belakang Pendidikan dengan omset pelaku usaha di Jawa Timur? (Violin plot)

```
#sns.kdeplot(data=tabel, x="PENDIDIKAN", hue="time", multiple = "stack")
plt.figure(figsize=(64,24))
sns.catplot(data=tabel, x="PENDIDIKAN", y="RATA2_OMSET", kind="violin", inner=None)

plt.title('Violin Plot Korelasi pendidikan dengan omset')
plt.xticks(rotation =45)
plt.show()
```

Terlihat bahwa latar belakang pendidikan berpengaruh terhadap omset usaha. Semakin tinggi latar belakang pendidikan, semakin besar omset usaha nya. Namun, tidak menutup kemungkinan bahwa pendidikan yang tinggi menjamin omset usaha yang besar. Dapat dilihat frekuensi omset yang dibawah 0.5 lulusan S1/lebih kurang lebih sama dengan jenjang lain.



Step 3 - Data Exploration & Data Visualisation

✓ Permasalahan 5:

Bagaimana omset usaha tiap daerah? ?

```
df_melt = pd.melt(tabel, id_vars='KAB_KOTA', value_vars=['OMSET_2019', 'OMSET_2020', 'OMSET_2021'], var_name='Year', value_name='Omset')

plt.figure(figsize=(18,18))

# Now we can create the barplot
sns.barplot(data=df_melt, y='KAB_KOTA', x='Omset', hue='Year')

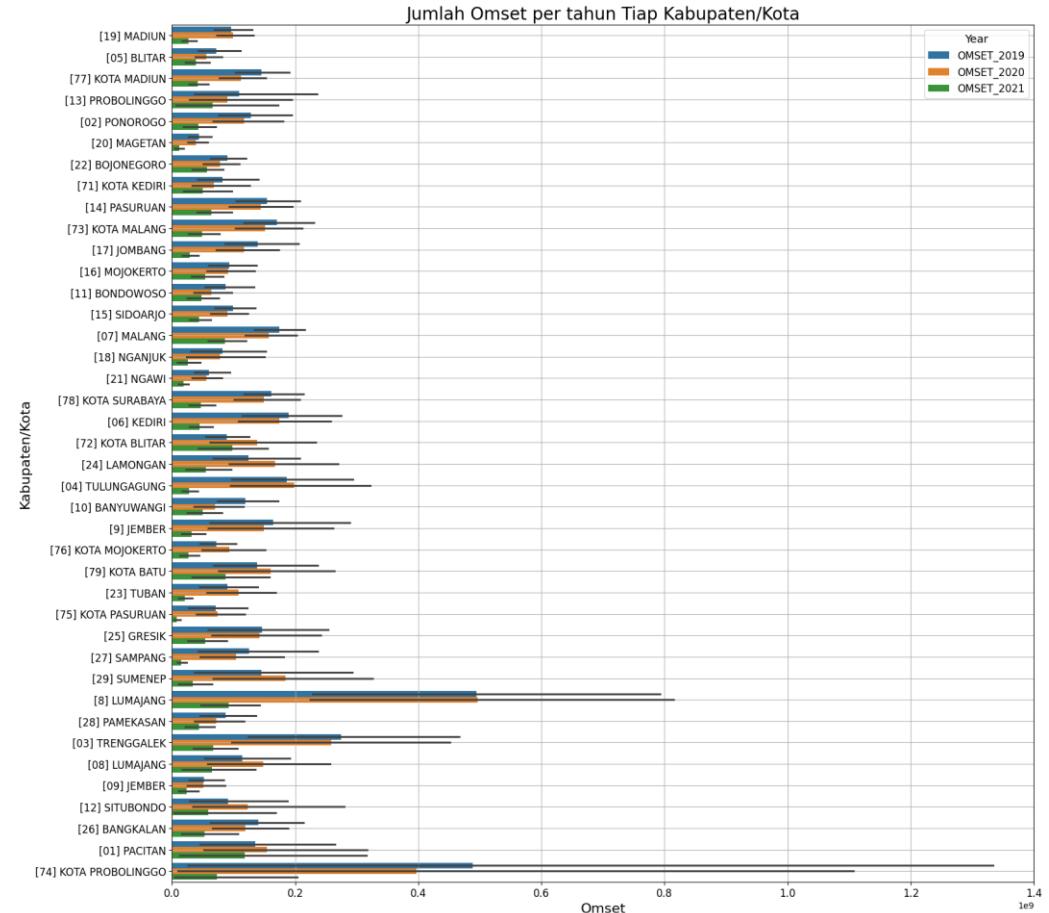
plt.title('Jumlah Omset per tahun Tiap Kabupaten/Kota', fontsize=20)
plt.xlabel('Omset', fontsize=16)
plt.ylabel('Kabupaten/Kota', fontsize=16)
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.grid(True)
plt.legend(title='Year', title_fontsize='13', fontsize='12')

plt.show()
```

Tiap daerah cenderung mengalami penurunan omset dari tahun 2019 ke 2021. Kemungkinan besar diakibatkan pandemi pada tahun tersebut sehingga aktivitas perekonomian yang terbatas mengakibatkan penurunan omset

Pada tahun 2019 Lumajang menjadi daerah dengan omset rata-rata tertinggi di Jawa Timur sampai tahun 2020 dan pada akhirnya tahun 2021 beralih ke Pacitan.

Jika dilihat seksama, Kota Probolinggo memiliki ketimpangan omset yang jauh besar terlihat selisih nilai maksimum dari omset pelaku UMKM disana dengan rata-rata omset tahunan disana sangat besar



TETRIS PROGRAM

Step 3 - Data Exploration & Data Visualisation

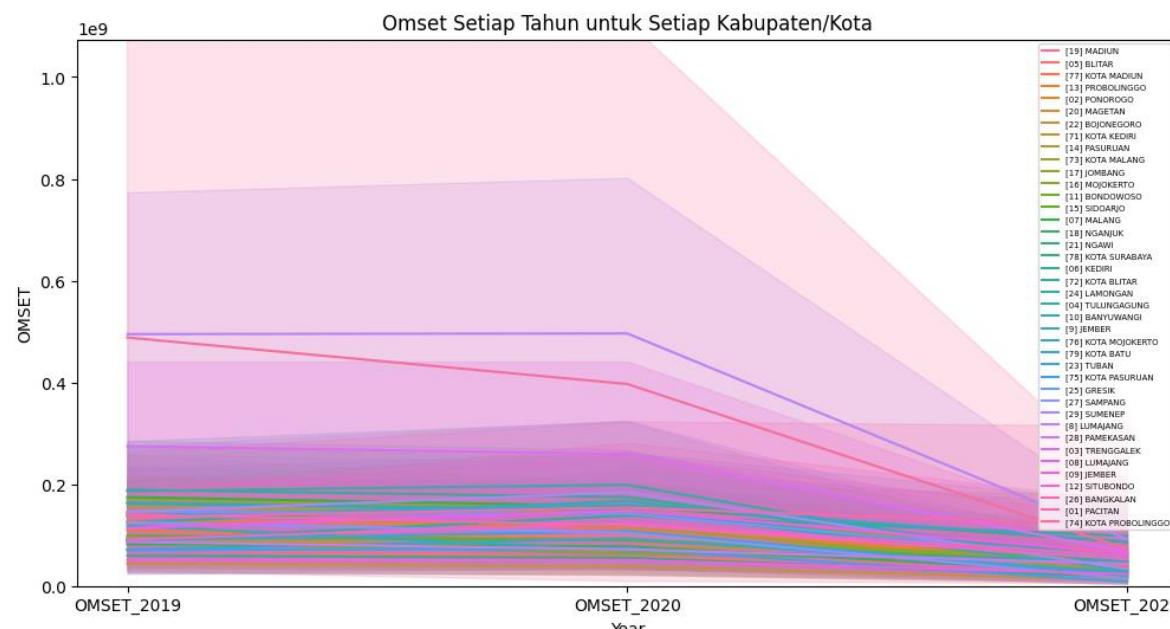
▼ Permasalahan 6:

Bagaimana perubahan omset keseluruhan dari 2019-2021?

▶ Line Plot

```
tbl_melted = tabel.melt(id_vars='KAB_KOTA', value_vars=['OMSET_2019', 'OMSET_2020', 'OMSET_2021'], var_name='Year', value_name='OMSET')

plt.figure(figsize=(12,6))
sns.lineplot(data=tbl_melted, x='Year', y='OMSET', hue='KAB_KOTA')
plt.ylim(0, tbl_melted['OMSET'].max() * 0.5)
plt.legend(prop={'size': 5})
plt.title('Omset Setiap Tahun untuk Setiap Kabupaten/Kota')
plt.show()
```



#StackYourSkill

Step 3 - Data Exploration & Data Visualisation

```
tbl_melted = tabel.melt(id_vars='KAB_KOTA', value_vars=['OMSET_2019', 'OMSET_2020', 'OMSET_2021'], var_name='Year', value_name='OMSET')

total_OMSET = tabel_melted.groupby('KAB_KOTA')['OMSET'].sum()

# Mengambil top 5 KAB_KOTA
top3_KAB_KOTA = total_OMSET.nlargest(3).index

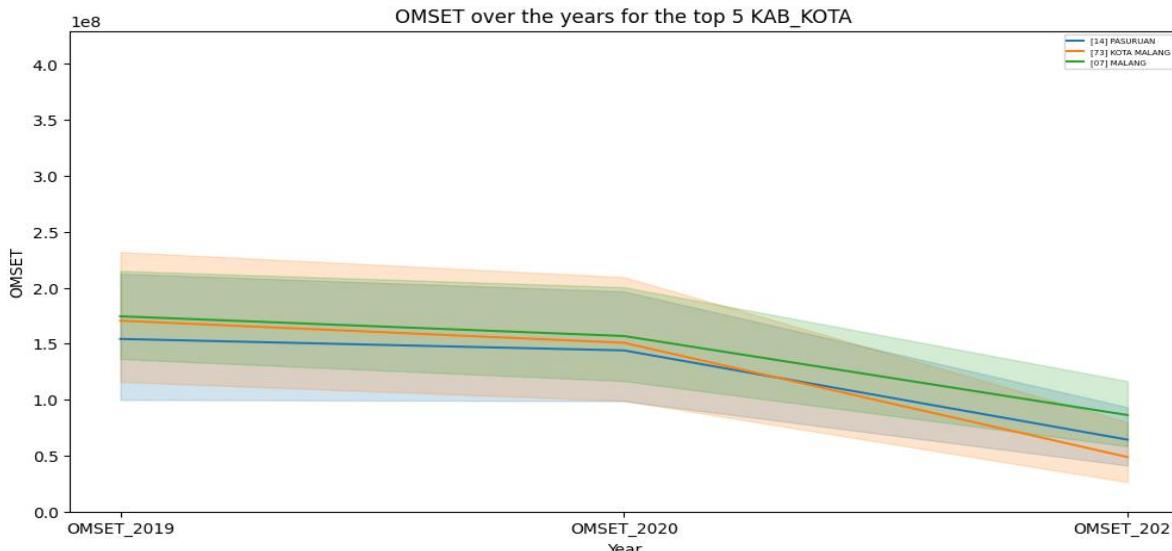
# Filter df_melted ke top 5 KAB_KOTA
tbl_melted_top3 = tabel_melted[tbl_melted['KAB_KOTA'].isin(top3_KAB_KOTA)]

plt.figure(figsize=(12,6))
sns.lineplot(data=tbl_melted_top3, x='Year', y='OMSET', hue='KAB_KOTA')

# Melebarkan y-axis
plt.ylim(0, tbl_melted_top3['OMSET'].max() * 0.2) # Sesuaikan mau distretch seberapa

# Menyesuaikan ukuran legenda
plt.legend(prop={'size': 5}) # sesuaikan mau seberapa besar

plt.title('OMSET per tahun tertinggi 3 KAB_KOTA')
plt.show()
```



Terlihat bahwa keseluruhan omset pelaku UMKM mengalami penurunan dari tahun 2019 ke 2021. Efek pandemi tidak dapat dipungkiri menjadi penyebab penurunan omset.

3 wilayah dengan omset tertinggi adalah kabupaten Pasuruan, kabupaten Malang, dan Kota Malang

Step 3 - Data Exploration & Data Visualisation

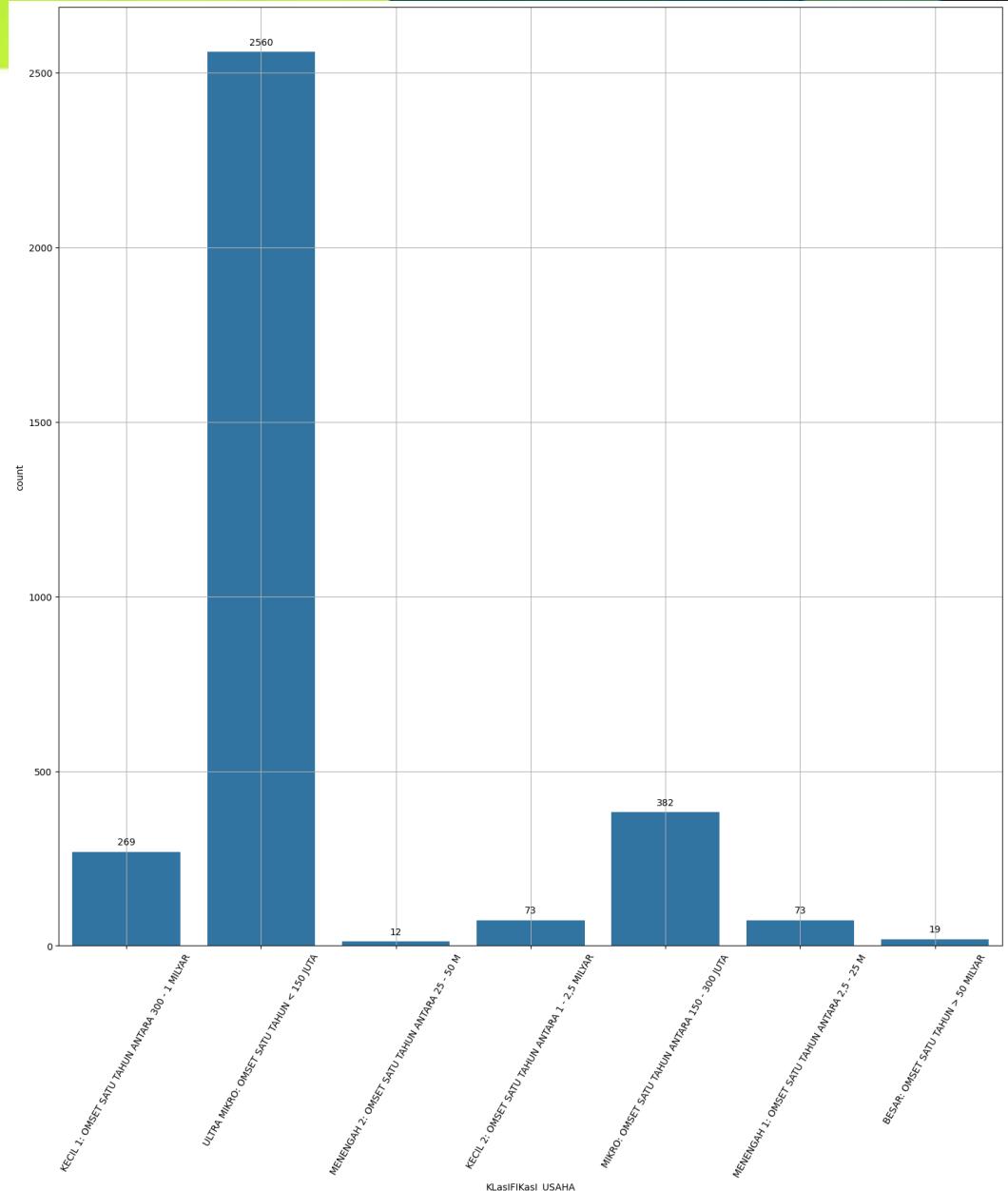
▼ Permasalahan 7:

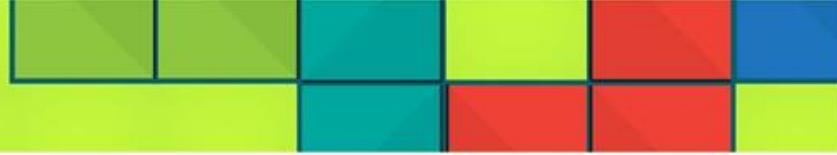
Seberapa banyak UMKM yang dikategorikan sebagai mikro, ultra mikro, kecil 1, kecil 2?

```
▶ column = 'KlasIFIkasI_USAHA'  
plt.figure(figsize=(18,18))  
ax = sns.countplot(x=tabel[column])  
  
# Menambah angka diatas bar  
for p in ax.patches:  
    ax.annotate(format(p.get_height(), '.0f'),  
                (p.get_x() + p.get_width() / 2., p.get_height()),  
                ha = 'center',  
                va = 'center',  
                xytext = (0, 10),  
                textcoords = 'offset points')  
plt.xticks(rotation=60)  
plt.grid(True)  
plt.show()
```

Kategori UMKM yang mendominasi di Jawa Timur adalah UMKM dengan kategori ultra mikro atau **yang beromset kurang dari 150 juta rupiah**

Diikuti dengan UMKM kategori mikro yang beromset 150-300 Juta rupiah dan kecil yang beromset 300 Juta – 100 Miliar





Step 4 – Machine Learning Deployment

▼ Proses Pelatihan model matematika Hist Gradient Boosting Regressor

```
▶ numerik = ['ID_DT_BINAAN_EXCEL', 'TAHUN_MULAI_USAHA', 'UMUR',
    'JML_TENAGA_KERJA_2019', 'JML_TENAGA_KERJA_2020',
    'JML_TENAGA_KERJA_2021', 'OMSET_2019', 'OMSET_2020',
    'OMSET_2021', 'BIAYA_PRODUKSI_2019', 'BIAYA_PRODUKSI_2020',
    'BIAYA_PRODUKSI_2021', 'GAJI_KARYAWAN_2019',
    'GAJI_KARYAWAN_2020', 'GAJI_KARYAWAN_2021', 'score', 'row_num']
kategorikal = ['KAB_KOTA', 'KECAMATAN', 'JENIS_KEGIATAN', 'NAMA PEMILIK',
    'NAMA_USAHA', 'JENIS_KELAMIN', 'PENDIDIKAN', 'STATUS_USAHA',
    'JENIS_LAPANGAN_USAHA', 'SERTIFIKASI', 'KLASIFIKASI_USAHA',
    'DAPAT_KREDIT', 'PERLU_PINJAMAN_PIHAK_LUAR']
# split data untuk training model dan test
from sklearn.model_selection import train_test_split

#feature = tabel.drop(columns='purchased')
numerik_train = ['RATA2_JML_TENAGA_KERJA', 'RATA2_BIAYA_PRODUKSI', 'RATA2_GAJI']
feature = tabel[numerik_train]
target = tabel[['RATA2_OMSET']]

feature_purchase_train, feature_purchase_test, target_purchase_train, target_purchase_test = train_test_split(feature, target, test_size=0.20, random_state=42)

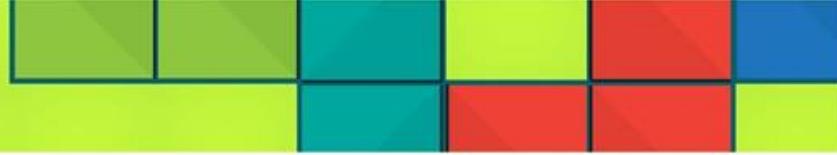
[ ] from sklearn.experimental import enable_hist_gradient_boosting # noqa
from sklearn.ensemble import HistGradientBoostingRegressor
from sklearn.model_selection import train_test_split
from sklearn.impute import SimpleImputer

numerik_train = ['RATA2_JML_TENAGA_KERJA', 'RATA2_BIAYA_PRODUKSI', 'RATA2_GAJI']
feature = tabel[numerik_train]
target = tabel[['RATA2_OMSET']]
```

Dilakukan pelatihan model Hist Gradient Boosting Regressor yang ditujukan untuk memprediksi omset suatu UMKM jika diketahui jumlah tenaga kerja, biaya produksi, dan gaji karyawan.

Alasan penggunaan model ini karena tipe data yang dijadikan sebagai bahan training bertipe numerik dan membutuhkan hasil output. Kemudian, karena dari data yang digunakan terdapat data kosong sehingga perlu menggunakan model ini.

Source <https://medium.com/chat-gpt-now-writes-all-my-articles/a-faster-ensemble-model-method-in-sklearn-histogram-based-gradient-boosting-7033ff170bc0>



Step 4 – Machine Learning Deployment

```
[ ] # Membuat imputer agar nilai kosong terisi di suatu feature
imputer_feature = SimpleImputer(strategy='mean') # or median, most_frequent, constant
feature_imputed = imputer_feature.fit_transform(feature)

[ ] # Sama seperti diatas namun untuk data test
imputer_target = SimpleImputer(strategy='mean') # or median, most_frequent, constant
target_imputed = imputer_target.fit_transform(target)

[ ] # Split data
feature_train, feature_test, target_train, target_test = train_test_split(feature_imputed, target_imputed, test_size=0.20, random_state=42)

[ ] # Pemanggilan model
model = HistGradientBoostingRegressor()

▶ # Train model
model.fit(feature_train, target_train.ravel())

[ ] # Model yang sudah siap memprediksi
predictions = model.predict(feature_test)
```



Step 4 – Machine Learning Deployment

```
[ ] # Mengukur akurasi model dengan mae, nilai mutlak
from sklearn.metrics import mean_absolute_error
mae = mean_absolute_error(target_test, predictions)
print("Mean Absolute Error (MAE):", mae)

Mean Absolute Error (MAE): 83448355.55028255

[ ] # mengukur akurasi model dengan R2 score (coefficient of determination)
from sklearn.metrics import r2_score
r2 = r2_score(target_test, predictions)
print("R2 Score:", r2)

R2 Score: 0.3305351191045067

▶ # Mendefinisikan data point
# Misalkan rata2 jumlah tenaga kerja sejumlah 7 orang, rata2 biaya produksi sebesar Rp. 5 jt, rata2 gaji karyawan sebesar 15 jt maka berapa rata2 omset usahanya
new_data = [[7, 5000000, 15000000]]

# Use the imputer to handle any potential missing values
new_data_imputed = imputer_feature.transform(new_data)

# Use the model to predict 'RATA2_OMSET' for the new data point
new_prediction = model.predict(new_data_imputed)

print("Prediksi nilai rata-rata omset adalah:", new_prediction[0])

/usr/local/lib/python3.10/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but SimpleImputer was fitted with feature names
  warnings.warn(
The predicted value of 'RATA2_OMSET' is: 32525550.724482648
```

Imputan 7 orang karyawan, Rp. 5 jt rata-rata biaya produksi, Rp. 15 jt rata-rata gaji karyawan/bulan diperkirakan beromset Rp. 32.525.550,72



Step 4 – Machine Learning Deployment

- ▼ Klasifikasi UMKM berdasarkan rata-rata omset yang telah diraih

```
[ ] tabel_encoded = tabel
# Mengisi nilai kosong di 2 kolom ini
tabel_encoded['RATA2_OMSET'].fillna(tabel_encoded['RATA2_OMSET'].median(), inplace=True)
tabel_encoded['KlasIFIKasi_USAHA'].fillna(tabel_encoded['KlasIFIKasi_USAHA'].mode()[0], inplace=True)

[ ] # One-hot encoding
one_hot = pd.get_dummies(tabel_encoded['KlasIFIKasi_USAHA'])
# Join the encoded df
tabel_encoded = tabel_encoded.join(one_hot)

▶ from sklearn.preprocessing import LabelEncoder

# Create a label encoder object, mengubah data kategorik jadi numerik
le = LabelEncoder()

# Fit and transform the 'KLASIFIKASI_USAHA' column
tabel_encoded['KlasIFIKasi_USAHA'] = le.fit_transform(tabel_encoded['KlasIFIKasi_USAHA'])

[ ] from sklearn.preprocessing import StandardScaler

# Define the features
features = ['RATA2_OMSET', 'KlasIFIKasi_USAHA']

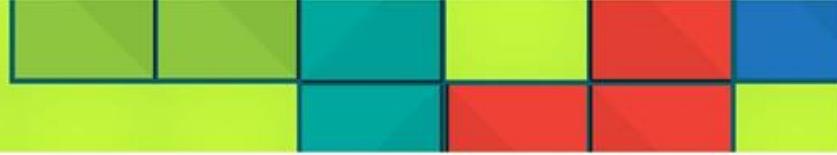
# Scale the features
scaler = StandardScaler()
tabel_scaled = scaler.fit_transform(tabel_encoded[features])
```

Model yang akan digunakan adalah K-means Clustering. Tujuan dari model ini adalah melakukan kategorisasi UMKM berdasarkan rata-rata omset dan klasifikasi usaha



Step 4 – Machine Learning Deployment

```
[ ]  from sklearn.cluster import KMeans  
  
# Create a KMeans object  
kmeans = KMeans(n_clusters=7, random_state=0)  
  
# Fit the model to the data and predict the cluster assignments  
tabel_encoded['Cluster'] = kmeans.fit_predict(tabel_scaled)  
  
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 10 to 'auto' in 1.4. Set the value of  
warnings.warn(  
  
▶ # Print the size of the clusters  
print(tabel_encoded['Cluster'].value_counts())  
  
# Print the mean 'RATA2_OMSET' and 'KLASIFIKASI_USAHA' for each cluster  
print(tabel_encoded.groupby('Cluster')[features].mean())  
  
[?] 0    2486  
6    347  
2    218  
5    162  
4    111  
3    34  
1    30  
Name: Cluster, dtype: int64  
      RATA2_OMSET  KLasIFIkasI_USAHA  
Cluster  
0        2.522263e+07        6.000000  
1        9.745235e+08        2.500000  
2        9.627908e+07        1.610092  
3        1.647978e+09        2.382353  
4        3.414714e+08        5.630631  
5        4.402591e+08        1.104938  
6        9.194610e+07        4.971182
```



Step 3 - Insight Analysis

Kesimpulan

- Kebutuhan Pinjaman Luar yang masih tinggi terlepas pendidikan, jenis usaha, klasifikasi usaha(Berdasarkan jumlah Omset) memiliki porsi lebih dari 50%. Sehingga, perlunya akses yang banyak terhadap pinjaman untuk UMKM
- Kredit masih terbatas dengan terbukti 70% UMKM tidak dapat kredit. Artinya, UMKM perlu disediakan media, dibimbing, atau difasilitasi agar usaha mereka layak mendapatkan kredit atau pinjaman untuk mendukung usaha
- Rata-rata omset dengan rata-rata biaya produksi & rata-rata gaji berkorelasi sedang daripada kolom lain sehingga pengamat UMKM dapat mempertimbangkan omset dengan 2 hal tersebut
- Melihat pendidikan berkorelasi dengan besarnya omset, sehingga dapat mempertimbangkan pendidikan lebih tinggi untuk para pelaku UMKM
- Penurunan omset dari 2019-2021 menandakan perlunya UMKM agar mengembalikan kondisi awal omsetnya sebelum pandemi
- UMKM dengan omset kurang dari 150 juta pertahun mendominasi. Artinya, stakeholder perlu lebih fokus dalam bagaimana pelaku UMKM tersebut bisa berubah menjadi UMKM dengan omset tinggi dan berlaba tinggi

DPLab

AYO #STACKYOURSKILL SEKARANG

dan Persiapkan Diri Menjadi Praktisi Data!

