MAKALAH DATA WRANGLING
ANALISIS PENGARUH FAKTOR DEMOGRAFIS TERHADAP TINGKAT PENGANGGURAN TERBUKA (TPT) DI INDONESIA

Dosen Pengampu Mata Kuliah:

Disusun Oleh:

Darista Wardhani	24031554146
Septiani Amalia Wulandari	1314623069

SAINS DATA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI SURABAYA

2025

BAB 1
PENDAHULUAN

1.1. Latar Belakang

Pengangguran merupakan salah satu permasalahan ekonomi dan sosial yang masih dihadapi oleh banyak negara, termasuk Indonesia. Tingginya tingkat pengangguran tidak hanya mencerminkan ketidakseimbangan antara permintaan dan penawaran tenaga kerja, tetapi juga menjadi indikator ketidakoptimalan pembangunan ekonomi dan kualitas sumber daya manusia. Badan Pusat Statistik (BPS) melalui indikator Tingkat Pengangguran Terbuka (TPT) menggambarkan persentase jumlah penganggur terhadap total angkatan kerja. Meskipun TPT Indonesia menunjukkan tren menurun dalam beberapa tahun terakhir, angka ini masih menunjukkan adanya ketimpangan yang perlu dikaji lebih mendalam dari berbagai aspek, terutama faktor-faktor demografis.

Berdasarkan proyeksi penduduk oleh BPS, jumlah penduduk Indonesia akan melonjak hingga 300 jiwa. Dari besarnya jumlah tersebut, sebagian besar merupakan kelompok dengan usia produktif (15–64 tahun). Keadaan tersebut menandakan bahwa Indonesia berada dalam era bonus demografi, yaitu ketika jumlah usia produktif melampaui usia non-produktif. Bonus demografi yang terjadi sebenarnya dapat memberikan dampak positif jika dimanfaatkan dengan baik, sebab dominasi usia produktif akan berdampak dengan meningkatnya pembangunan. Namun, bonus demografi juga berpotensi menjadi tantangan dan beban perekonomian dan peningkatan angka pengangguran jika tidak dimanfaatkan dengan baik (Priastiwi, 2018).

Penelitian yang dilakukan oleh Endah Novianti (2018) berjudul "Kesenjangan Gender Tingkat Pengangguran Terbuka di Indonesia" menunjukkan hasil analisis bahwa secara parsial, peluang perempuan tidak memiliki pekerjaan lebih tinggi sebesar 30,61% dibandingkan laki-laki. Selain itu, usia memiliki pengaruh yang negatif pada kesenjangan gender tingkat pengangguran terbuka di Indonesia. Namun, pada perempuan, tingkat pengangguran meningkat seiring dengan pertumbuhan usia, khususnya pada usia 25 tahun ke atas.

Penelitian oleh R.BG. Miko Oktavio Wijaya dan Efri Diah Utami (2021) menyoroti bahwa persentase pengangguran dengan pendidikan terakhir SMK pada tahun 2020 memiliki nilai yang tinggi, yaitu sebesar 8,49%. Tingginya persentase pengangguran dalam penelitian tersebut juga dipengaruhi oleh variabel jenis kelamin, gangguan disabilitas, keahlian, tahun kelulusan, serta pelatihan. Berdasarkan penelitian tersebut, variabel-variabel tersebut berpengaruh signifikan terhadap lulusan SMK yang masih menganggur.

Penelitian terdahulu menunjukkan bahwa faktor demografis berperan dalam meningkatnya tingkat pengangguran terbuka di Indonesia. Melihat situasi bonus demografis yang sedang terjadi di Indonesia, diperlukan strategi terbaik untuk mengoptimalisasikan situasi tersebut. Optimalisasi bonus demografis memerlukan pendekatan yang dapat menyatukan kebijakan demografis, pendidikan, serta perencanaan karir sehingga bonus

demografis dapat memberikan dampak yang signifikan terhadap pertumbuhan ekonomi (Manik et al., 2025).

1.2. Tujuan

1. Menganalisis hubungan antara faktor demografis (pendidikan, umur, gender) terhadap TPT.
2. Mengetahui kelompok demografis mana yang paling rentan terhadap pengangguran.
3. Memberikan masukan bagi pemerintah untuk perumusan kebijakan ketenagakerjaan yang inklusif.

1.3. Rumusan Masalah

1. Bagaimana pengaruh tingkat pendidikan terhadap TPT di Indonesia?
2. Apakah kelompok umur tertentu memiliki tingkat pengangguran yang lebih tinggi?
3. Adakah perbedaan tingkat pengangguran antara laki-laki dan perempuan?

1.4. Manfaat

1. Mendorong kesadaran masyarakat terhadap ketimpangan gender dan usia dalam dunia kerja, sehingga muncul dukungan terhadap kebijakan yang lebih inklusif.
2. Membantu mengidentifikasi kelompok rentan terhadap pengangguran, seperti pemuda atau perempuan, sehingga intervensi sosial dapat lebih tepat sasaran.
3. Secara tidak langsung berkontribusi terhadap penurunan tingkat kemiskinan dan peningkatan kesejahteraan masyarakat melalui pengurangan pengangguran.

BAB II
PEMBAHASAN

2.1 Teknik pengambilan dan integritas data

2.1.1 Teknik pengambilan

Teknik pengambilan (sering disebut sebagai *data extraction technique*) adalah cara atau metode yang digunakan untuk mengambil, memilih, menggabungkan, atau mendapatkan data tertentu dari suatu sumber data baik itu berasal dari tabel, kolom, database, file CSV, maupun kumpulan dataset yang lebih besar.

```
import pandas as pd
import os

base_path = "/content"
files = os.listdir(base_path)

csv_files = [f for f in files if f.endswith('.csv')]
excel_files = [f for f in files if f.endswith('.xlsx')]

print("CSV ditemukan:", csv_files)
print("Excel ditemukan:", excel_files)
```

```
CSV ditemukan: ['Data_pengganggguran.csv', 'Data_Pendidikan.csv', 'TPT_Gender.csv', '
Excel ditemukan: []
```

```
import pandas as pd

df1 = pd.read_csv("Data_Pendidikan.csv")
df2 = pd.read_csv("Data_pengganggguran.csv")
df3 = pd.read_csv("TPT_Age.csv")
df4 = pd.read_csv("TPT_Gender.csv")

print("=== Tabel Data 1 ===")
print(df1.head())
print("\n=== Tabel Data 2 ===")
print(df2.head())
print("\n=== Tabel Data 3 ===")
print(df3.head())
print("\n=== Tabel Data 4 ===")
print(df4.head())
```

Pada kode ini menggunakan teknik pengambilan data dari csv yang menggunakan library pandas. Pada library ini digunakan untuk membaca, memproses, dan menganalisis data berbentuk tabel. Selain itu, fungsi pd.read_csv() digunakan untuk membaca file csv yang berbeda. Setiap pemanggilan fungsi tersebut membuka file csv yang disebutkan, membaca seluruh isi file baris demi baris, lalu mengubahnya menjadi sebuah DataFrame, yaitu struktur data mirip tabel seperti di Excel. DataFrame ini kemudian disimpan dalam variabel df1, df2, df3 dan df4 sehingga setiap variabel tersebut mewakili dataset yang berbeda data pendidikan, data pengangguran, data TPT berdasarkan umur, dan data TPT berdasarkan gender. Setelah

datanya dibaca kode tersebut menampilkan lima baris pertama pada masing-masing frame dengan metode .head, yang memang fungsinya menunjukkan contoh beberapa baris awal dari sebuah tabel.

2.1.2 Integrasi Data

Integrasi data adalah proses menggabungkan beberapa sumber data yang berbeda menjadi satu dataset yang lengkap, konsisten, dan mudah dianalisis. Tujuannya adalah agar data yang tersebar di berbagai file atau tabel dapat dipadukan sehingga memberikan gambaran yang utuh.

```
import pandas as pd

tpt_age = pd.read_csv("TPT_Age.csv")
pendidikan = pd.read_csv("Data_Pendidikan.csv")
tpt_gender = pd.read_csv("TPT_Gender.csv")
pengangguran = pd.read_csv("Data_pengganggguran.csv")
tpt_age["TPT"] = pd.to_numeric(tpt_age["TPT"], errors="coerce")
tpt_age_pivot = (
    tpt_age
    .pivot_table(index="Year", columns="Kelompok_Umur", values="TPT")
    .reset_index()
)
for col in pendidikan.columns:
    if col not in ["Periode", "Bulan"]:
        pendidikan[col] = pd.to_numeric(pendidikan[col], errors="coerce")
pendidikan_mean = (
    pendidikan
    .groupby("Periode")
    .mean(numeric_only=True)
    .reset_index()
    .rename(columns={"Periode": "Year"})
)
tpt_gender["Laki - Laki"] = pd.to_numeric(tpt_gender["Laki - Laki"], errors="coerce")
tpt_gender["Perempuan"] = pd.to_numeric(tpt_gender["Perempuan"], errors="coerce")

pengangguran["Year"] = pd.to_datetime(pengangguran["Date"], errors='coerce').dt.year
pengangguran["Value"] = (
    pengangguran["Value"]
    .astype(str)
    .str.replace("%", "", regex=False)
    .astype(float)
)
pengangguran = pengangguran[["Year", "Value"]]
```

Pada kode ini menggunakan import pandas karena fungsi ini digunakan untuk membaca, membersihkan dan mengolah data dalam bentuk tabel. Selain itu, empat data tersebut dibaca menggunakan pd.read_csv() yaitu data pengangguran, age, gender, dan pendidikan. pada dataset TPT_Age menggunakan pd.to_numeric() karena kolom TPT dikonversi ke dalam bentuk angka. Konversi ini berguna karena ada kemungkinan nilai pada kolom tersebut tersimpan dalam bentuk format teks atau berisi karakter yang tidak dapat langsung dihitung. penggunaan errors=’coerce’ memastikan bahwa nilai yang tidak bisa dikonversi akan diganti dengan NaN sehingga tidak menimbulkan error. Setelah dikonversi

kolom tersebut diubah menjadi pivot sehingga setiap kelompok umur menjadi kolom tersendiri sementara tahun menjadi baris.

pada dataset pendidikan semua kolom ditelusuri kecuali kolom “periode” dan “bulan” dan mengkonversikan ke dalam numerik. Hasilnya adalah sebuah tabel baru yang berisi rata-rata pendidikan per tahun, dan kolom “Periode” diubah namanya menjadi “Year” agar konsisten dengan dataset lain. Pada dataset TPT_Gender, dua kolom utama yaitu “Laki - Laki” dan “Perempuan” juga dikonversi ke dalam bentuk numerik. Ini memastikan bahwa jika ada simbol, spasi, atau format yang salah, nilainya tetap bisa diproses. Setelah itu melakukan pembersihan pada dataset pengangguran Kolom “Date” diubah menjadi format tanggal, lalu hanya bagian tahunnya yang diambil dan disimpan sebagai kolom “Year”. Kolom “Value”, yang kemungkinan berisi angka dalam bentuk teks dan mungkin diikuti tanda persen, dibersihkan dengan cara menghapus tanda persen dan mengkonversinya menjadi tipe data float.

```
merged = (
    tpt_age_pivot
    .merge(pendidikan_mean, on="Year", how="outer")
    .merge(tpt_gender, on="Year", how="outer")
    .merge(pengangguran, on="Year", how="outer")
)

merged = merged[(merged["Year"] >= 2015) & (merged["Year"] <= 2022)]
```

Pada bagian kode ini melakukan proses penyatuan dari empat dataset berbeda menjadi satu tabel besar. Penggabungan dilakukan dengan menggunakan fungsi .merge() dari pandas. Setiap merge menggunakan kolom "Year" sebagai kunci penggabungan. Setelah seluruh dataset berhasil digabungkan, baris merged digunakan untuk memilih hanya data yang berada dalam rentang tahun 2015 hingga 2022.

```
def combine_cols(df, cols, new_name):
    existing = [c for c in cols if c in df.columns]
    if len(existing) > 1:
        df[new_name] = df[existing].bfill(axis=1).iloc[:, 0]
        df.drop(columns=existing, inplace=True)
    elif len(existing) == 1:
        df.rename(columns={existing[0]: new_name}, inplace=True)

combine_cols(merged, ["60-above", "60 keatas"], "60 Keatas")
combine_cols(merged, ["Mean", "Rata-Rata"], "Rata-Rata")
```

Fungsi ini digunakan untuk menggabungkan atau menyatukan beberapa kolom yang sebenarnya berisi data yang sama tetapi memiliki nama berbeda. Parameter df adalah DataFrame, cols adalah daftar nama kolom yang ingin dicek atau digabung, dan new_name adalah nama kolom baru yang akan dipakai sebagai hasil akhirnya. Baris existing digunakan untuk memeriksa kolom mana saja dari daftar cols yang benar-benar ada di dalam

DataFrame. Jika lebih dari satu kolom ditemukan, artinya ada duplikasi nama kolom untuk data yang sama dan perlu digabung.Metode bfill (backward fill) digunakan untuk mengisi nilai kosong dari kolom pertama dengan nilai kolom kedua, dan seterusnya. Setelah nilai sudah dipindahkan ke kolom baru, kolom-kolom lama yang duplikat dihapus dari DataFrame agar tidak membuat kebingungan.Jika hanya satu kolom yang ditemukan, berarti tidak perlu menggabungkan apa pun.
Kolom tersebut cukup di-*rename* menjadi nama baru (new_name), agar nama kolom konsisten di seluruh dataset.

```
cols = list(merged.columns)

if "55-59" in cols and "60 Keatas" in cols:
    cols.insert(cols.index("55-59") + 1, cols.pop(cols.index("60 Keatas")))

if "Rata-Rata" in cols and "60 Keatas" in cols:
    cols.insert(cols.index("60 Keatas") + 1, cols.pop(cols.index("Rata-Rata")))

merged = merged[cols]
merged.to_csv("Data_Terintegrasi_2015_2022.csv", index=False)
print("Integrasi selesai! Kolom 60 Keatas & Rata-Rata sudah diatur posisinya.")
print(merged.head(10))
```

Kode tersebut mengatur ulang urutan kolom pada DataFrame dengan memindahkan kolom **“60 Keatas”** agar berada tepat setelah kolom **“55-59”**, serta memindahkan kolom **“Rata-Rata”** agar berada setelah **“60 Keatas”**. Setelah urutannya sesuai, DataFrame disusun ulang berdasarkan aturan tersebut, lalu disimpan ke file CSV, dan ditampilkan 10 baris awal sebagai hasil akhirnya.

2.2 Data cleaning

Data cleaning (pembersihan data) adalah proses memperbaiki data agar lebih akurat, konsisten, dan siap digunakan untuk analisis. Secara lebih spesifik, data cleaning mencakup kegiatan seperti menghapus data yang tercatat dua kali, mengisi atau menangani nilai yang kosong, serta memperbaiki format yang salah.

```python
import pandas as pd
import numpy as np

df = pd.read_csv("Data_Terintegrasi_2015_2022.csv")

print("=== Data Awal ===")
print(df.head(), "\n")


df.columns = df.columns.str.strip().str.replace('[^A-Za-z0-9_]', '_', regex=True)
for col in df.columns:
    if col != 'Year':
        df[col] = pd.to_numeric(df[col], errors='coerce')
print("=== Jumlah Missing Value per Kolom ===")
print(df.isna().sum(), "\n")

df = df.fillna(0)
df = df.drop_duplicates()

df = df[(df['Year'] >= 2015) & (df['Year'] <= 2022)]
df = df.round(2)

print("=== Statistik Ringkas Setelah Cleaning ===")
print(df.describe(), "\n")

df.to_csv("Data_Terintegrasi_2015_2022_CLEAN.csv", index=False)
print(" Data berhasil dibersihkan dan disimpan sebagai 'Data_Terintegrasi_2015_2022_CLEAN.csv'")
```

Kode tersebut melakukan proses data cleaning dengan beberapa langkah penting untuk memastikan data siap digunakan dalam analisis. Pertama, dataset dibaca dan ditampilkan untuk melihat kondisi awal data. Nama kolom kemudian dibersihkan dengan menghapus spasi dan mengganti karakter yang tidak valid menggunakan regex agar lebih konsisten dan mudah diproses. Selanjutnya, semua kolom kecuali “Year” dikonversi menjadi tipe numerik; jika terdapat nilai yang tidak dapat diubah, nilai tersebut otomatis diubah menjadi NaN. Setelah itu, jumlah missing value dihitung, kemudian seluruh nilai NaN diisi dengan angka 0 sebagai bentuk simple imputation.

Data duplikat dihapus untuk memastikan tidak ada baris yang sama persis. Data kemudian difilter agar hanya mencakup tahun 2015 hingga 2022, sesuai rentang analisis yang diinginkan. Seluruh nilai numerik dibulatkan menjadi dua desimal agar data lebih rapi dan konsisten. Setelah proses cleaning selesai, statistik ringkas ditampilkan untuk memeriksa kondisi akhir data, dan dataset yang telah dibersihkan disimpan kembali dalam file CSV baru. Secara keseluruhan, metode yang digunakan meliputi pembersihan nama kolom, konversi tipe data, imputasi missing value, penghapusan duplikasi, filtering data, dan pembulatan nilai.

```
for col in df.columns:
    if col != "Year":
        df[col] = pd.to_numeric(df[col], errors="coerce")
if "total" in df.columns:
    df = df.drop(columns=["total"])
if "Value" in df.columns:
    df = df.rename(columns={"Value": "Data_Pengangguran"})

df = df.fillna(0)
df = df[(df['Year'] >= 2015) & (df['Year'] <= 2022)]

def add_unit_label(x):
    if x >= 1_000_000:
        new_val = round(x / 1_000_000, 2)
        return f"{new_val} (dalam juta)"
    if x < 10:
        return f"{x} (satuan)"
    elif x < 100:
        return f"{x} (puluhan)"
    elif x < 1000:
        return f"{x} (ratusan)"
    else:  # 1.000 - 999.999
        return f"{x} (ribuan)"
```

Kode tersebut melakukan proses pembersihan data dengan mengonversi seluruh kolom numerik (kecuali *Year*) menjadi tipe angka, menghapus kolom *total* apabila ada, serta mengganti nama kolom *Value* menjadi *Data_Pengangguran* agar lebih sesuai dengan konteks analisis. Nilai hilang kemudian diganti dengan nol, dan dataset difilter sehingga hanya mencakup tahun 2015–2022.

Selain itu, kode mendefinisikan fungsi *add_unit_label()* yang menambahkan label satuan pada setiap nilai berdasarkan besar kecilnya angka. Nilai yang mencapai satu juta dikonversi menjadi satuan "dalam juta", sedangkan nilai lainnya diberi label "satuan", "puluhan", "ratusan", atau "ribuan". Fungsi ini memastikan bahwa setiap nilai numerik disajikan dengan skala yang jelas, sehingga data lebih mudah diinterpretasikan pada tahap analisis selanjutnya.

2.3 Data eksplorasi

Exploratory Data Analysis (EDA) merupakan proses awal dalam analisis data yang bertujuan untuk mengidentifikasi pola, struktur, dan elemen penting dalam data sebagai dasar sebelum melanjutkan ke tahap analisis statistik atau prediksi berikutnya. Eksplorasi data biasanya dapat dilakukan dengan memahami data secara mendalam seperti melihat melihat distribusi tiap variabel, memeriksa kecenderungan atau tren, mengidentifikasi outlier, serta mengevaluasi kemungkinan hubungan antarvariabel melalui visualisasi maupun statistik deskriptif.

2.3.1 Eksplorasi Tabel

Pada tahap ini dilakukan eksplorasi awal untuk melihat struktur dan karakteristik *dataset* menggunakan fungsi dari *library* pandas. Berikut ini merupakan kode yang digunakan untuk eksplorasi tabel.

```
print(data.info())
print("\nBaris dan Kolom :", data.shape)
print("\nNama Kolom:", list(data.columns))
print(data.describe())
```

- data.info(): digunakan untuk menampilkan ringkasan struktur *dataset*, seperti ringkasan baris dan kolom, tipe data setiap kolom, dan jumlah nilai non-null per kolom.
- data.shape: digunakan untuk menampilkan jumlah baris dan kolom pada dataset.
- data.columns: digunakan untuk menampilkan nama-nama setiap kolom pada dataset.
- data.describe(): digunakan untuk menampilkan statistik deskriptif pada kolom numerik, fungsi tersebut akan menampilkan nilai minimum dan maksimum, mean, standar deviasi, dan kuartil pada dataset.

Eksplorasi tabel berikutnya yang dilakukan adalah *aggregating data,* yaitu proses menggabungkan dan merangkum data. Tujuannya untuk mendapatkan informasi ringkas data sehingga analisis lebih mudah untuk dilakukan. Kode yang digunakan di bawah ini, akan menampilkan fungsi statistik, seperti nilai minimum, nilai maksimum, rata-rata (*mean*), dan nilai tengah (*median*). *Output* yang akan dihasilkan pada data ini menampilkan gambaran besar setiap variabelnya.

```
faktor_pendidikan = ['Tidak_belum_pernah_sekolah',  'Tidak_belum_tamat_SD', 'SD','SLTP','SLTA_Umum_SMU','SLTA_Kejuruan_SMK',
        'Akademi_Diploma','Universitas']
data[faktor_pendidikan].agg(['min', 'max', 'mean', 'median'])

faktor_usia = ['15_19', '20_24', '25_29', '30_34',  '35_39',  '40_44',  '45_49',  '50_54',  '55_59',  '60_Keatas']
data[faktor_usia].agg(['min', 'max', 'mean', 'median'])

faktor_gender = ['Laki___Laki', 'Perempuan']
data[faktor_gender].agg(['min', 'max', 'mean', 'median'])
```

2.3.2 Pengecekan *Missing Values* dan *Duplicates*

Eksplorasi berikutnya yang dilakukan adalah pengecekan *missing values* dan *duplicates*. Pengecekan *missing values* dilakukan untuk memastikan kelengkapan data, jika ditemukan ada nilai yang kosong pada setiap kolomnya, maka akan dilakukan proses imputasi maupun pembersihan data. Untuk pengecekan *duplicates* digunakan untuk menghitung jumlah baris yang nilainya sama lebih dari sekali, sebab *duplicates* dapat memengaruhi hasil analisis berikutnya. Berikut ini fungsi-fungsi yang digunakan untuk pengecekan *missing values* dan *duplicates data*.

```
print("\nJumlah Missing Values Tiap Kolom:")
print(data.isna().sum())

print("\nJumlah Data Duplikat:", data.duplicated().sum())
```

2.3.4 Pengecekan *Outlier*

Outlier atau nilai pencilan perlu dideteksi dalam proses eksplorasi data, sebab nilai yang menyimpang jauh dari sebaran data dapat memengaruhi hasil analisis dan berpotensi menghasilkan kesimpulan yang tidak akurat. Pengecekan outlier dilakukan dengan dengan fungsi yang tersedia dalam *library* seaborn untuk visualisasi statistik dan matpolib.pylot untuk menampilkan grafik.

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.boxplot(data=data[['Akademi_Diploma','Universitas']])
plt.title("Boxplot for Outlier Detection")
plt.show()

sns.boxplot(data=data[['20_24','25_29','30_34']])
plt.title("Boxplot for Outlier Detection")
plt.show()
```

Kode di atas digunakan untuk mendeteksi outlier faktor demografis pendidikan tinggi yaitu variabel ‘Akademi_Diploma’ dan ‘Universitas’ dan kelompok usia 20-34 tahun. Keduanya digunakan untuk mengetahui nilai pengangguran yang menyimpang jauh dari mayoritas data setiap variabelnya.

2.3.5 Mengidentifikasi Korelasi

Korelasi merupakan ukuran statistik yang digunakan untuk mengetahui hubungan antara variabel (positif, negatif, atau tidak ada hubungannya. Korelasi biasanya juga digunakan untuk memahami pola hubungan antar variabel.

```
age_cols = ['15_19','20_24','25_29','30_34','35_39','40_44',
            '45_49','50_54','55_59','60_Keatas']

data[age_cols].corr()

edu_cols = ['Tidak_belum_pernah_sekolah','Tidak_belum_tamat_SD','SD','SLTP',
            'SLTA_Umum_SMU','SLTA_Kejuruan_SMK','Akademi_Diploma','Universitas']

data[edu_cols].corr()

gender_cols = ['Laki___Laki','Perempuan']

data[gender_cols].corr()
```

Kode di atas digunakan untuk mencari korelasi setiap variabel dengan tipe numerik. Perhitungan korelasi dilakukan dengan menggunakan fungsi data.corr() yang tersedia pada library pandas. Fungsi tersebut akan menghitung korelasi dengan default korelasi Pearson dan menghasilkan matriks korelasi dengan rentang -1 sampai +1.

2.3.6 Visualisasi

Visualisasi dalam EDA digunakan untuk memperkuat analisis dengan mengubah data menjadi gambar yang bermakna sehingga dapat menghasilkan pemahaman yang lebih mendalam. Visualisasi dalam EDA umumnya disajikan dalam grafik maupun diagram, visualisasi tersebut dapat menyajikan pola, tren, serta hubungan antar variabel dalam data. Seluruh visualisasi menggunakan *library* pandas, numpy, matplotlib.pylot, dan seaborn. Berikut ini kode yang digunakan untuk visualisasi dalam analisis data eksploratif.

- **Line Chart:** digunakan untuk melihat tren waktu.

```
# Tren Jumlah Pengangguran (2015-2022)
plt.figure(figsize=(10,5))
sns.lineplot(data=data, x='Year', y='Value', marker='o', color='teal')
plt.title("Tren Tingkat Pengangguran (2015–2022)")
plt.ylabel("Tingkat Pengangguran (%)")
plt.xlabel("Tahun")
plt.show()
```

```
# Analisis Jenis Kelamin terhadap Pengangguran
plt.figure(figsize=(7,5))
sns.lineplot(data=data, x='Year', y='Laki___Laki', label='Laki-laki', marker='o')
sns.lineplot(data=data, x='Year', y='Perempuan', label='Perempuan', marker='o')
plt.title("Tingkat Pengangguran Berdasarkan Jenis Kelamin (2015–2022)")
plt.ylabel("Tingkat Pengangguran (%)")
plt.xlabel("Tahun")
plt.legend()
plt.show()
```

```
# Analisis Kelompok Umur terhadap Pengangguran
umur_cols = ['15_19', '20_24', '25_29', '30_34',  '35_39',  '40_44',  '45_49',  '50_54',  '55_59',  '60_Keatas']

plt.figure(figsize=(10,6))
for col in umur_cols:
    plt.plot(data["Year"], data[col], marker='o', label=col)
plt.title("Analisis Kelompok Umur terhadap Pengangguran")
plt.xlabel("Tahun")
plt.ylabel("Persentase Pengangguran (%)")
plt.legend(title="Kelompok Umur", bbox_to_anchor=(1.05,1), loc='upper left')
plt.show()
```

```
# Analisis Tingkat Pendidikan Tertinggi terhadap Pengangguran
pendidikan_cols = ['Tidak_belum_pernah_sekolah',  'Tidak_belum_tamat_SD',
                   'SD','SLTP','SLTA_Umum_SMU','SLTA_Kejuruan_SMK','Akademi_Diploma','Universitas']

plt.figure(figsize=(10,6))
for col in pendidikan_cols:
    plt.plot(data["Year"], data[col], marker='o', label=col)
plt.title("Analisis Tingkat Pendidikan Tertinggi terhadap Pengangguran")
plt.xlabel("Tahun")
plt.ylabel("Jumlah Pengangguran (orang)")
plt.legend(title="Pendidikan Tertinggi", bbox_to_anchor=(1.05,1), loc='upper left')
plt.show()
```

Kode-kode di atas digunakan untuk menampilkan visualisasi line chart pada dataset. Pertama grafik akan diatur ukurannya agar sesuai dan proporsional, berikutnya dengan fungsi sns.lineplot akan dibentuk garis tren untuk setiap variabelnya. Kode berikutnya yaitu plt.title akan menambahkan judul pada grafik lalu plt.legend() akan menampilkan legenda dan plt.show() akan menampilkan grafik line chart.

- **Bar Chart:** digunakan untuk membandingkan nilai kategori tertentu pada periode tertentu.

```
# Analisis Kelompok Umur terhadap Pengangguran (2022)
df_2022 = data[data["Year"] == 2022]
umur_cols = ['15_19', '20_24', '25_29', '30_34',  '35_39',  '40_44',  '45_49',  '50_54',  '55_59',  '60_Keatas']

plt.figure(figsize=(10,5))
sns.barplot(x=umur_cols, y=df_2022[umur_cols].iloc[0])
plt.title("Tingkat Pengangguran per Kelompok Umur (Tahun 2022)")
plt.xlabel("Kelompok Umur")
plt.ylabel("Tingkat Pengangguran (%)")
plt.xticks(rotation=45)
plt.show()
```

```
# Analisis Tingkat Pendidikan Terakhir terhadap Pengangguran (2022)
df_2022 = df[df["Year"] == 2022]
pendidikan_cols = ["SD", "SLTP", "SLTA_Umum_SMU", "SLTA_Kejuruan_SMK",
                   "Akademi_Diploma", "Universitas"]
# Plot bar chart
plt.figure(figsize=(10,5))
sns.barplot(x=pendidikan_cols, y=df_2022[pendidikan_cols].iloc[0])
plt.title("Tingkat Pengangguran per Tingkat Pendidikan (Tahun 2022)")
plt.xlabel("Tingkat Pendidikan")
plt.ylabel("Jumlah Pengangguran (Orang)")
plt.xticks(rotation=30)
plt.grid(axis='y', linestyle='--', alpha=0.6)
plt.show()
```

```
# Analisis Jenis Kelamin terhadap Pengangguran (2022)
df_2022 = data[data["Year"] == 2022]
gender_cols = ['Laki___Laki', 'Perempuan']

plt.figure(figsize=(10,5))
sns.barplot(x=gender_cols, y=df_2022[gender_cols].iloc[0])
plt.title("Tingkat Pengangguran per Jenis Kelamin (Tahun 2022)")
plt.xlabel("Jenis Kelamin")
plt.ylabel("Tingkat Pengangguran (%)")
plt.xticks(rotation=45)
plt.show()
```

Kode-kode di atas digunakan untuk menampilkan visualisasi bar chart pada dataset. Pertama akan dipilih tahun yang akan digunakan untuk visualisasi, pada kode di atas dipilih tahun 2022 yang akan divisualisasikan, berikutnya akan diambil daftar kelompok usia, pendidikan terakhir, dan jenis kelamin yang akan divisualisasikan. Seperti pada visualisasi sebelumnya, kode di atas juga dimulai dengan mengatur ukuran grafik lalu kode `sns.barplot()` digunakan untuk membuat bar chart. Berikutnya akan ditambahkan judul, label sumbu, dan memiringkan label umur agar tidak bertumpuk, dan terakhir `plt.show()` digunakan untuk menampilkan grafik.

- **Stacked Bar Chart:** digunakan untuk melihat perbandingan antar kategori dalam bentuk proporsi.

```
pendidikan = [
    'Tidak_belum_pernah_sekolah', 'Tidak_belum_tamat_SD', 'SD','SLTP','SLTA_Umum_SMU','SLTA_Kejuruan_SMK',
        'Akademi_Diploma','Universitas'
]

data_plot = data[['Year'] + pendidikan].set_index('Year')

data_plot.plot(
    kind='bar',
    stacked=True,
    figsize=(12,7)
)

plt.title('Komposisi Rata-rata Pengangguran berdasarkan Pendidikan Tertinggi per Tahun')
plt.xlabel('Tahun')
plt.ylabel('Rata-rata Pengangguran')
plt.legend(
    title='Tingkat Pendidikan',
    loc='center left',
    bbox_to_anchor=(1, 0.5)
)
plt.show()
```

```
gender = ['Laki___Laki',  'Perempuan']

data_plot = data[['Year'] + gender].set_index('Year')

data_plot.plot(
    kind='bar',
    stacked=True,
    figsize=(12,7)
)

plt.title('Komposisi Rata-rata Pengangguran berdasarkan Jenis Kelamin per Tahun')
plt.xlabel('Tahun')
plt.ylabel('Rata-rata Pengangguran')
plt.legend(
    title='Jenis Kelamin',
    loc='center left',
    bbox_to_anchor=(1, 0.5)
)
plt.show()
```

```
age = ['15_19', '20_24', '25_29', '30_34',  '35_39',  '40_44',  '45_49',  '50_54',  '55_59',  '60_Keatas']

data_plot = data[['Year'] + age].set_index('Year')

data_plot.plot(
    kind='bar',
    stacked=True,
    figsize=(12,7)
)

plt.title('Komposisi Rata-rata Pengangguran berdasarkan Kelompok Umur per Tahun')
plt.xlabel('Tahun')
plt.ylabel('Rata-rata Pengangguran')
plt.legend(
    title='Kelompok Umur',
    loc='center left',
    bbox_to_anchor=(1, 0.5)
)
plt.show()
```

Kode-kode di atas digunakan untuk menampilkan visualisasi stacked bar chart pada dataset. Pertama akan dipilih kategori yang akan diplot dan disimpan dalam dataframe. Berikutnya, akan diambil kolom tahun dan data frame dengan kolom 'Year' sebagai index. Kode berikutnya yaitu data_plot.plot() digunakan untuk membuat stacked bar chart. Kode berikutnya digunakan untuk membuat judul dan label dan memindahkan legenda ke luar grafik. Tahap terakhir grafik akan ditampilkan dengan fungsi plt.show().

- **Heatmap:** digunakan untuk melihat korelasi antar variabel dalam bentuk visualisasi matriks berwarna.

```
import seaborn as sns
age_cols = ['15_19','20_24','25_29','30_34','35_39','40_44',
            '45_49','50_54','55_59','60_Keatas']

sns.heatmap(data[age_cols].corr(), vmin=-1, vmax=1, annot=True)

edu_cols = ['Tidak_belum_pernah_sekolah','Tidak_belum_tamat_SD','SD','SLTP',
            'SLTA_Umum_SMU','SLTA_Kejuruan_SMK','Akademi_Diploma','Universitas']

sns.heatmap(data[edu_cols].corr(), vmin=-1, vmax=1, annot=True)

gender_cols = ['Laki___Laki','Perempuan']

sns.heatmap(data[gender_cols].corr(), vmin=-1, vmax=1, annot=True)
```

Kode-kode di atas digunakan untuk menampilkan visualisasi heatmap pada dataset. Pertama akan meng-*import library* seaborn kemudian akan diambil subset dari dataset per faktor demografis, yaitu kelompok umur, pendidikan terakhir, dan jenis kelamin. Berikutnya sns.heatmap() digunakan untuk menampilkan korelasi dalam bentuk matriks berwarna dengan rentang -1 sampai 1, annot=True digunakan untuk menampilkan nilai korelasinya.

BAB III
HASIL IMPLEMENTASI

3.1 Teknik pengambilan dan integrasi data

3.1.1 Teknik pengambilan

```
CSV ditemukan: ['TPT_Age.csv', 'Data_pengganguran.csv', 'Data_Pendidikan.csv', 'TPT_Gender.csv']
Excel ditemukan: []
```

Interpretasi : Teknik pengambilan data yang ditampilkan menunjukkan bahwa sistem berhasil membaca empat file CSV yang berisi data pendidikan, pengangguran, usia, dan gender.

```
=== Tabel Data 1 ===
   Periode     Bulan  Tidak/belum pernah sekolah  Tidak/belum tamat SD  \
0     2006  Februari                      234465                614960
1     2006   Agustus                      170666                611254
2     2007  Februari                      145750                520316
3     2007   Agustus                       94301                438519
4     2008  Februari                       79764                448431

        SD     SLTP  SLTA Umum/SMU  SLTA Kejuruan/SMK  Akademi/Diploma  \
0  2675459  2860007        2842876            1204140           297185
1  2589699  2730045        2851518            1305190           278074
2  2753548  2643062        2630360            1114675           330316
3  2179792  2264198        2532204            1538349           397191
4  2216748  2166619        2204377            1165582           519867

   Universitas     Total
0       375601  11104693
1       395554  10932000
2       409890  10547917
3       566588  10011142
4       626202   9427590

=== Tabel Data 2 ===
        Date  Value
0  31-Dec-82  3.00%
1  31-Dec-83    NaN
2  31-Dec-84    NaN
3  31-Dec-85  2.20%
4  31-Dec-86  2.70%

=== Tabel Data 3 ===
  Kelompok_Umur    TPT  Year
0         15-19  17.71  2015
1         20-24  12.86  2015
2         25-29  10.65  2015
3         30-34   8.86  2015
4         35-39   7.86  2015

=== Tabel Data 4 ===
   Year  Laki - Laki  Perempuan
0  2015         6.07       6.37
1  2016         5.70       5.45
2  2017         5.53       5.44
3  2018         5.34       5.25
4  2019         5.24       5.22
```

Interpretasi : Pada hasil output ini terlihat bahwa sistem membaca dan menampilkan beberapa tabel. Hasilnya menampilkan beberapa tabel: data pendidikan per tahun, tingkat pengangguran tahunan, tingkat pengangguran berdasarkan umur, serta perbedaan pengangguran antara laki-laki dan perempuan

3.1.2 Integrasi data

```
Integrasi selesai! Kolom 60 Keatas & Rata-Rata sudah diatur posisinya.
    Year  15-19  20-24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  ...  \
33  2015  17.71  12.86  10.65   8.86   7.86   7.59   6.97   5.93   6.05  ...
34  2016  28.09  15.80   7.08   3.63   2.21   2.05   1.35   1.66   1.55  ...
35  2017  27.54  16.62   6.76   3.40   2.45   1.86   1.51   1.54   1.73  ...
36  2018  26.93  16.79   6.97   3.44   2.48   1.80   1.58   1.39   1.25  ...
37  2019  26.12  15.64   7.19   3.52   2.25   2.06   1.81   1.65   1.30  ...
38  2020  24.34  18.71   9.77   5.75   4.32   3.92   3.54   3.61   3.21  ...
39  2021  23.91  17.73   9.26   5.43   4.02   3.42   3.30   2.18   1.98  ...
40  2022  29.08  17.02   7.13   3.70   2.65   2.43   2.33   2.38   2.37  ...

           SD          SLTP  SLTA Umum/SMU  SLTA Kejuruan/SMK  \
33  1162676.5  1.512153e+06   2.021220e+06          1372028.0
34  1127342.5  1.304149e+06   1.748662e+06          1434438.0
35  1098397.5  1.277828e+06   1.731862e+06          1502212.0
36   941944.5  1.203794e+06   1.809214e+06          1598790.5
37   915709.5  1.186197e+06   1.849281e+06          1568453.0
38  1208640.5  1.436435e+06   2.205639e+06          1885060.5
39  1306493.0  1.559768e+06   2.388976e+06          2100237.5
40  1259740.0  1.487278e+06   2.402635e+06          1733215.0

    Akademi/Diploma  Universitas      Total  Laki - Laki  Perempuan  Value
33    252926.500000     609494.0  7507794.5         6.07       6.37   4.51
34    234549.000000     631269.5  7027973.5         5.70       5.45   4.30
35    246321.000000     612848.5  7005262.0         5.53       5.44   3.78
36    264100.000000     771997.0  7018421.0         5.34       5.25   4.39
37    246665.500000     801104.0  7001610.0         5.24       5.22   3.59
38    286422.000000     903057.5  8346620.0         7.46       6.46   4.26
39    235240.500000     924100.0  8924030.0         6.74       6.11   3.83
40    184779.666667     743913.0  8418005.0         5.93       5.75   3.46

[8 rows x 23 columns]
```

Interpretasi : Pada data diatas output yang dihasilkan berhasil dilakukan, di mana kolom dari berbagai sumber data (umur, pendidikan, gender, dan TPT) telah digabung menjadi satu tabel.
Isi tabel mencakup:

- Kolom umur (15–19, 20–24, 25–29, dst.)
- Kolom pendidikan (SD, SLTP, SLTA, SMK, Akademi/Diploma, Universitas, Total)
- Kolom gender (Laki-laki, Perempuan)
- Nilai TPT (Value)

Setiap baris merepresentasikan tahun 2015–2022, dengan data yang sudah tersusun rapi dan tidak ada kolom yang salah posisi. Tabel bawah foto pertama memperlihatkan jumlah data kosong (missing values) di setiap kolom. Semua bernilai 0, artinya data sudah lengkap dan tidak ada nilai hilang.

3.2 Cleaning data

```
Tidak_belum_pernah_sekolah    0
Tidak_belum_tamat_SD          0
SD                            0
SLTP                          0
SLTA_Umum_SMU                 0
SLTA_Kejuruan_SMK             0
Akademi_Diploma               0
Universitas                   0
Total                         0
Laki___Laki                   0
Perempuan                     0
Value                         0
dtype: int64

=== Statistik Ringkas Setelah Cleaning ===
             Year      15_19      20_24      25_29     30_34     35_39  \
count     8.00000   8.000000   8.000000   8.000000  8.000000  8.000000
mean   2018.50000  25.465000  16.396250   8.101250  4.716250  3.530000
std       2.44949   3.598877   1.738595   1.536177  1.918101  1.928626
min    2015.00000  17.710000  12.860000   6.760000  3.400000  2.210000
25%    2016.75000  24.232500  15.760000   7.052500  3.500000  2.400000
50%    2018.50000  26.525000  16.705000   7.160000  3.665000  2.565000
75%    2020.25000  27.677500  17.197500   9.387500  5.510000  4.095000
max    2022.00000  29.080000  18.710000  10.650000  8.860000  7.860000

          40_44     45_49     50_54     55_59  ...            SD  \
count  8.000000  8.000000  8.000000  8.000000  ...  8.000000e+00
mean   3.141250  2.798750  2.542500  2.430000  ...  1.127618e+06
std    1.956012  1.876261  1.543704  1.596988  ...  1.401428e+05
min    1.800000  1.350000  1.390000  1.250000  ...  9.157095e+05
25%    2.002500  1.562500  1.622500  1.487500  ...  1.059284e+06
50%    2.245000  2.070000  1.920000  1.855000  ...  1.145010e+06
75%    3.545000  3.360000  2.687500  2.580000  ...  1.221415e+06
max    7.590000  6.970000  5.930000  6.050000  ...  1.306493e+06

               SLTP  SLTA_Umum_SMU  SLTA_Kejuruan_SMK  Akademi_Diploma  \
count  8.000000e+00   8.000000e+00       8.000000e+00         8.000000
mean   1.370950e+06   2.019686e+06       1.649304e+06    243875.521250
std    1.457100e+05   2.796015e+05       2.448723e+05     29200.744081
min    1.186197e+06   1.731862e+06       1.372028e+06    184779.670000
25%    1.259320e+06   1.794076e+06       1.485268e+06    235067.625000
50%    1.370292e+06   1.935250e+06       1.583622e+06    246493.250000
75%    1.493497e+06   2.251473e+06       1.771176e+06    255719.875000
max    1.559768e+06   2.402635e+06       2.100238e+06    286422.000000

         Universitas         Total  Laki___Laki  Perempuan     Value
count       8.000000  8.000000e+00     8.000000   8.000000  8.000000
mean   749722.937500  7.656214e+06     6.001250   5.756250  4.015000
std    125082.106868  7.874083e+05     0.757165   0.497822  0.397312
min    609494.000000  7.001610e+06     5.240000   5.220000  3.460000
25%    626664.250000  7.015131e+06     5.482500   5.392500  3.732500
50%    757955.000000  7.267884e+06     5.815000   5.600000  4.045000
75%    826592.375000  8.364466e+06     6.237500   6.175000  4.322500
max    924100.000000  8.924030e+06     7.460000   6.460000  4.510000

[8 rows x 23 columns]

 Data berhasil dibersihkan dan disimpan sebagai 'Data_Terintegrasi_2015_2022_CLEAN.csv'
```

Interpretasi : pada hasil ini output menampilkan statistik ringkas (descriptive statistics) setelah proses pembersihan data.
Statistik yang ditampilkan mencakup:count (jumlah data)

- mean (rata-rata)
- std (standar deviasi)

- min – max (nilai minimum dan maksimum)
- 25%, 50%, 75% (kuartil)

Statistik ini ditampilkan untuk seluruh kolom, baik kolom umur, pendidikan, gender, maupun nilai TPT.Di bagian paling bawah terdapat pesan bahwa data yang sudah dibersihkan telah disimpan ke dalam file: Data_CLEAN.csv

- Cleaning kedua

```
...     Year           15_19           20_24           25_29          30_34  \
    0  2015  17.71 (puluhan)  12.86 (puluhan)  10.65 (puluhan)  8.86 (satuan)
    1  2016  28.09 (puluhan)   15.8 (puluhan)    7.08 (satuan)  3.63 (satuan)
    2  2017  27.54 (puluhan)  16.62 (puluhan)    6.76 (satuan)   3.4 (satuan)
    3  2018  26.93 (puluhan)  16.79 (puluhan)    6.97 (satuan)  3.44 (satuan)
    4  2019  26.12 (puluhan)  15.64 (puluhan)    7.19 (satuan)  3.52 (satuan)

              35_39          40_44          45_49          50_54          55_59  \
    0  7.86 (satuan)  7.59 (satuan)  6.97 (satuan)  5.93 (satuan)  6.05 (satuan)
    1  2.21 (satuan)  2.05 (satuan)  1.35 (satuan)  1.66 (satuan)  1.55 (satuan)
    2  2.45 (satuan)  1.86 (satuan)  1.51 (satuan)  1.54 (satuan)  1.73 (satuan)
    3  2.48 (satuan)   1.8 (satuan)  1.58 (satuan)  1.39 (satuan)  1.25 (satuan)
    4  2.25 (satuan)  2.06 (satuan)  1.81 (satuan)  1.65 (satuan)   1.3 (satuan)

       ...                SD              SLTP      SLTA_Umum_SMU  \
    0  ...  1.16 (dalam juta)  1.51 (dalam juta)  2.02 (dalam juta)
    1  ...  1.13 (dalam juta)   1.3 (dalam juta)  1.75 (dalam juta)
    2  ...   1.1 (dalam juta)  1.28 (dalam juta)  1.73 (dalam juta)
    3  ...  941944.5 (ribuan)   1.2 (dalam juta)  1.81 (dalam juta)
    4  ...  915709.5 (ribuan)  1.19 (dalam juta)  1.85 (dalam juta)

       SLTA_Kejuruan_SMK    Akademi_Diploma       Universitas             Total  \
    0  1.37 (dalam juta)  252926.5 (ribuan)  609494.0 (ribuan)  7.51 (dalam juta)
    1  1.43 (dalam juta)  234549.0 (ribuan)  631269.5 (ribuan)  7.03 (dalam juta)
    2   1.5 (dalam juta)  246321.0 (ribuan)  612848.5 (ribuan)  7.01 (dalam juta)
    3   1.6 (dalam juta)  264100.0 (ribuan)  771997.0 (ribuan)  7.02 (dalam juta)
    4  1.57 (dalam juta)  246665.5 (ribuan)  801104.0 (ribuan)   7.0 (dalam juta)

         Laki___Laki      Perempuan Data_Pengangguran
    0  6.07 (satuan)  6.37 (satuan)     4.51 (satuan)
    1   5.7 (satuan)  5.45 (satuan)      4.3 (satuan)
    2  5.53 (satuan)  5.44 (satuan)     3.78 (satuan)
    3  5.34 (satuan)  5.25 (satuan)     4.39 (satuan)
    4  5.24 (satuan)  5.22 (satuan)     3.59 (satuan)

    [5 rows x 23 columns]
    Data berhasil dibersihkan dengan satuan lengkap.
    /tmp/ipython-input-2231280398.py:46: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
      df[numeric_cols] = df[numeric_cols].applymap(add_unit_label)
```

Data yang ditampilkan mencakup berbagai kelompok variabel, antara lain kelompok umur, tingkat pendidikan, jenis kelamin, serta total populasi. Seluruh nilai numerik telah diformat ulang sehingga disertai dengan keterangan satuan, seperti “(satuan)”, “(puluhan)”, “(ratusan)”, “(ribuan)”, dan “(dalam juta)”. Penambahan label ini bertujuan untuk memberikan kejelasan mengenai skala angka sehingga mempermudah proses interpretasi dan perbandingan antarvariabel. Secara keseluruhan, tampilan ini menunjukkan bahwa proses pembersihan data berhasil dilakukan dengan baik: kolom sudah rapi, nilai-nilai numerik telah dikonversi dengan benar, data tahun di luar 2015–2022 telah disaring, dan setiap angka kini memiliki penanda unit yang konsisten. Hasil ini mempermudah interpretasi lanjutan dan memastikan bahwa dataset siap digunakan untuk analisis statistik berikutnya.

3.3 Eksplorasi data

Exploratory Data Analysis (EDA) merupakan proses awal dalam analisis data yang bertujuan untuk mengidentifikasi pola, struktur, dan elemen penting dalam data sebagai dasar sebelum melanjutkan ke tahap analisis statistik atau prediksi berikutnya.

- Eksplorasi Tabel

  Berdasarkan output, tabel memiliki 23 kolom dan 8 baris. Selain itu nama dari setiap kolom nya yaitu, `['Year', '15_19', '20_24', '25_29', '30_34', '35_39', '40_44', '45_49', '50_54', '55_59', '60_Keatas', 'Tidak_belum_pernah_sekolah', 'Tidak_belum_tamat_SD', 'SD', 'SLTP', 'SLTA_Umum_SMU', 'SLTA_Kejuruan_SMK', 'Akademi_Diploma', 'Universitas', 'Total', 'Laki___Laki', 'Perempuan', 'Value']`.

  Untuk ringkasan dari setiap kolomnya seperti mean, standar devasi, nilai minimum dan maksimum, serta kuartil ditampilkan dalam gambar di bawah ini.

	Year	15_19	20_24	25_29	30_34	35_39
count	8.00000	8.000000	8.000000	8.000000	8.000000	8.000000
mean	2018.50000	25.465000	16.396250	8.101250	4.716250	3.530000
std	2.44949	3.598877	1.738595	1.536177	1.918101	1.928626
min	2015.00000	17.710000	12.860000	6.760000	3.400000	2.210000
25%	2016.75000	24.232500	15.760000	7.052500	3.500000	2.400000
50%	2018.50000	26.525000	16.705000	7.160000	3.665000	2.565000
75%	2020.25000	27.677500	17.197500	9.387500	5.510000	4.095000
max	2022.00000	29.080000	18.710000	10.650000	8.860000	7.860000

	40_44	45_49	50_54	55_59	...	SD
count	8.000000	8.000000	8.000000	8.000000	...	8.000000e+00
mean	3.141250	2.798750	2.542500	2.430000	...	1.127618e+06
std	1.956012	1.876261	1.543704	1.596988	...	1.401428e+05
min	1.800000	1.350000	1.390000	1.250000	...	9.157095e+05
25%	2.002500	1.562500	1.622500	1.487500	...	1.059284e+06
50%	2.245000	2.070000	1.920000	1.855000	...	1.145010e+06
75%	3.545000	3.360000	2.687500	2.580000	...	1.221415e+06
max	7.590000	6.970000	5.930000	6.050000	...	1.306493e+06

	SLTP	SLTA_Umum_SMU	SLTA_Kejuruan_SMK	Akademi_Diploma
count	8.000000e+00	8.000000e+00	8.000000e+00	8.000000
mean	1.370950e+06	2.019686e+06	1.649304e+06	243875.521250
std	1.457100e+05	2.796015e+05	2.448723e+05	29200.744081
min	1.186197e+06	1.731862e+06	1.372028e+06	184779.670000
25%	1.259320e+06	1.794076e+06	1.485268e+06	235067.625000
50%	1.370292e+06	1.935250e+06	1.583622e+06	246493.250000
75%	1.493497e+06	2.251473e+06	1.771176e+06	255719.875000
max	1.559768e+06	2.402635e+06	2.100238e+06	286422.000000

	Universitas	Total	Laki___Laki	Perempuan	Value
count	8.000000	8.000000e+00	8.000000	8.000000	8.000000
mean	749722.937500	7.656214e+06	6.001250	5.756250	4.015000
std	125082.106868	7.874083e+05	0.757165	0.497822	0.397312
min	609494.000000	7.001610e+06	5.240000	5.220000	3.460000
25%	626664.250000	7.015131e+06	5.482500	5.392500	3.732500
50%	757955.000000	7.267884e+06	5.815000	5.600000	4.045000
75%	826592.375000	8.364466e+06	6.237500	6.175000	4.322500
max	924100.000000	8.924030e+06	7.460000	6.460000	4.510000

  Eksplorasi tabel terakhir yang dilakukan adalah aggregating data yaitu proses pengelompokkan dan menghitung ringkasan statistik seperti nilai minimum dan maksimum, mean, serta median.

Aggregating berdasarkan pendidikan terakhir

	Tidak_belum_pernah_sekolah	Tidak_belum_tamat_SD	SD	SLTP	SLTA_Umum_SMU	SLTA_Kejuruan_SMK	Akademi_Diploma	Universitas
min	18421.33000	387031.5000	915709.5	1.186197e+06	1.731862e+06	1.372028e+06	184779.67000	609494.0000
max	89928.50000	588023.0000	1306493.0	1.559768e+06	2.402635e+06	2.100238e+06	286422.00000	924100.0000
mean	49400.47875	447848.0625	1127618.0	1.370950e+06	2.019686e+06	1.649304e+06	243875.52125	749722.9375
median	38312.00000	433173.5000	1145009.5	1.370292e+06	1.935250e+06	1.583622e+06	246493.25000	757955.0000

Aggregating berdasarkan kelompok umur

	15_19	20_24	25_29	30_34	35_39	40_44	45_49	50_54	55_59	60_Keatas
min	17.710	12.86000	6.76000	3.40000	2.210	1.80000	1.35000	1.3900	1.250	0.6100
max	29.080	18.71000	10.65000	8.86000	7.860	7.59000	6.97000	5.9300	6.050	4.7400
mean	25.465	16.39625	8.10125	4.71625	3.530	3.14125	2.79875	2.5425	2.430	2.0425
median	26.525	16.70500	7.16000	3.66500	2.565	2.24500	2.07000	1.9200	1.855	1.6100

Aggregating berdasarkan jenis kelamin

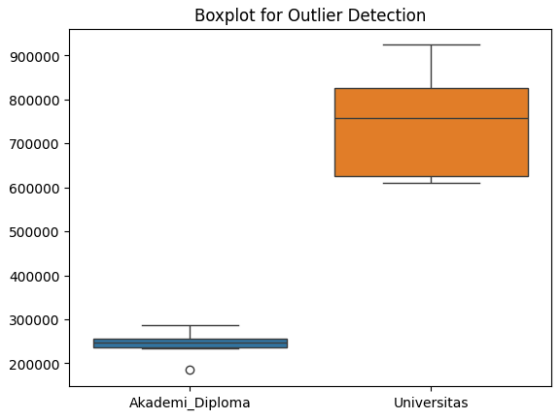
	Laki___Laki	Perempuan
min	5.24000	5.22000
max	7.46000	6.46000
mean	6.00125	5.75625
median	5.81500	5.60000

- Pengecekan Missing Value dan Duplicates

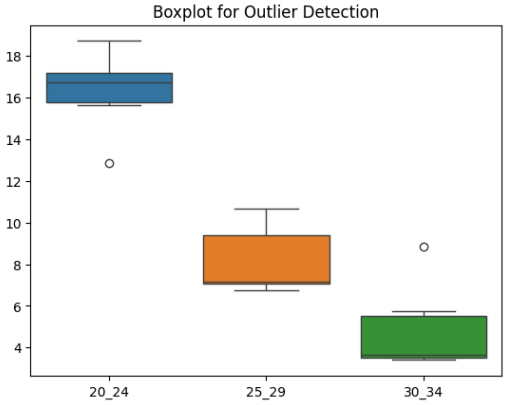
  Berdasarkan output, tidak ditemukan missing values dan duplicates pada data, hal tersebut karena data sebelumnya sudah dilakukan proses cleaning

- Pengecekan Outlier

  Pengecekan outlier dilakukan pada variabel akademi diploma dan universitas dengan menggunakan boxplot, didapatkan bahwa variabel akademi diploma memiliki satu pencilan dan variasi pada variabel tersebut cenderung kecil. Namun pada variabel universitas tidak ditemukan pencilan dan variasi pada variabel tersebut cukup besar.



Selain dua variabel di atas, pengecekan outlier juga dilakukan pada kelompok umur 20-34 tahun. Dari visualisasi boxplot, didapatkan bahwa kelompok usia 20-24 tahun dan 30-34 tahun memiliki pencilan/outlier, sedangkan kelompok usia 25-29 tidak memiliki outlier.



- Mengidentifikasi Korelasi

	15_19	20_24	25_29	30_34	35_39	40_44	45_49	50_54	55_59	60_Keatas
15_19	1.000000	0.547551	-0.895975	-0.958531	-0.956932	-0.945225	-0.942859	-0.878509	-0.862485	-0.704088
20_24	0.547551	1.000000	-0.195756	-0.479279	-0.523162	-0.560738	-0.517466	-0.509441	-0.559597	-0.497382
25_29	-0.895975	-0.195756	1.000000	0.945861	0.912663	0.902732	0.911599	0.874466	0.830796	0.711699
30_34	-0.958531	-0.479279	0.945861	1.000000	0.993616	0.991453	0.988141	0.958349	0.945864	0.832228
35_39	-0.956932	-0.523162	0.912663	0.993616	1.000000	0.994334	0.993070	0.963346	0.961324	0.845805
40_44	-0.945225	-0.560738	0.902732	0.991453	0.994334	1.000000	0.994285	0.978130	0.973690	0.857825
45_49	-0.942859	-0.517466	0.911599	0.988141	0.993070	0.994285	1.000000	0.970342	0.963696	0.866661
50_54	-0.878509	-0.509441	0.874466	0.958349	0.963346	0.978130	0.970342	1.000000	0.991781	0.832556
55_59	-0.862485	-0.559597	0.830796	0.945864	0.961324	0.973690	0.963696	0.991781	1.000000	0.867297
60_Keatas	-0.704088	-0.497382	0.711699	0.832228	0.845805	0.857825	0.866661	0.832556	0.867297	1.000000

Interpretasi: berdasarkan kelompok umur, TPT memiliki korelasi kuat hubungan baik positif maupun negatif. Korelasi positif menunjukkan jika kelompok umur tertentu naik, maka TPT kelompok umur lain juga akan meningkat. Namun, jika korelasi menunjukkan negatif maka jika TPT kelompok umur tertentu maka kelompok umur lainnya akan meningkat. Korelasi negatif salah satunya ditunjukkan pada kelompok umur 15-19 dengan 25-29 dan korelasi positif salah satunya ditunjukkan pada kelompok umur 25-29 dengan 30-34.

	Tidak_belum_pernah_sekolah	Tidak_belum_tamat_SD	SD	SLTP	SLTA_Umum_SMU	SLTA_Kejuruan_SMK	Akademi_Diploma	Universitas
Tidak_belum_pernah_sekolah	1.000000	0.179546	-0.193062	-0.163330	-0.665173	-0.808214	0.231360	-0.847899
Tidak_belum_tamat_SD	0.179546	1.000000	0.356956	0.295748	0.191862	-0.347971	-0.779991	-0.585094
SD	-0.193062	0.356956	1.000000	0.925743	0.785476	0.549486	-0.349201	0.219548
SLTP	-0.163330	0.295748	0.925743	1.000000	0.835394	0.490962	-0.266640	0.234620
SLTA_Umum_SMU	-0.665173	0.191862	0.785476	0.835394	1.000000	0.765863	-0.363981	0.606688
SLTA_Kejuruan_SMK	-0.808214	-0.347971	0.549486	0.490962	0.765863	1.000000	-0.018671	0.902418
Akademi_Diploma	0.231360	-0.779991	-0.349201	-0.266640	-0.363981	-0.018671	1.000000	0.213400
Universitas	-0.847899	-0.585094	0.219548	0.234620	0.606688	0.902418	0.213400	1.000000

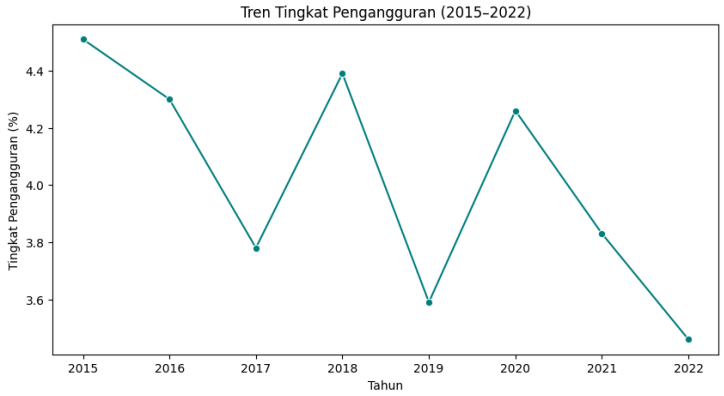
Interpretasi: berdasarkan tingkat pendidikan terakhir, TPT memiliki korelasi kuat hubungan baik positif maupun negatif. Korelasi positif menunjukkan jika kelompok umur tertentu naik, maka TPT kelompok umur lain juga akan meningkat. Namun, jika korelasi menunjukkan negatif maka jika TPT

kelompok umur tertentu maka kelompok umur lainnya akan meningkat. Korelasi negatif salah satunya ditunjukkan pada tingkat pendidikan SD dengan Akademi/Diploma dan korelasi positif salah satunya ditunjukkan pada tingkat pendidikan SLTP dengan SD.

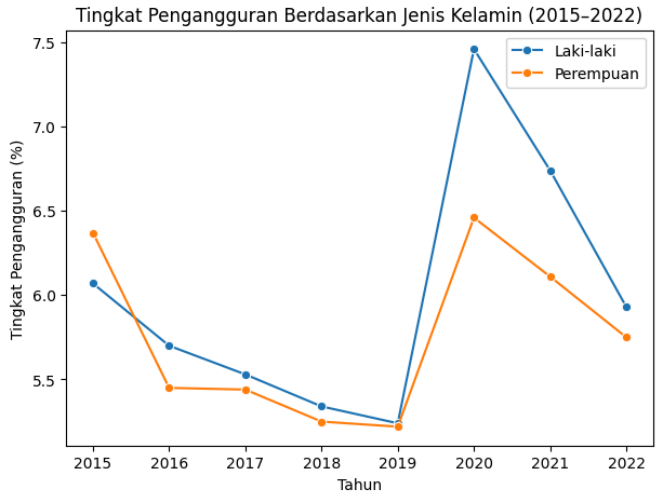
	Laki___Laki	Perempuan
Laki___Laki	1.00000	0.87732
Perempuan	0.87732	1.00000

Interpretasi: berdasarkan jenis kelamin, TPT memiliki korelasi kuat hubungan positif pada kedua variabel. Artinya, jika TPT pada laki-laki meningkat maka TPT pada perempuan juga cenderung meningkat.

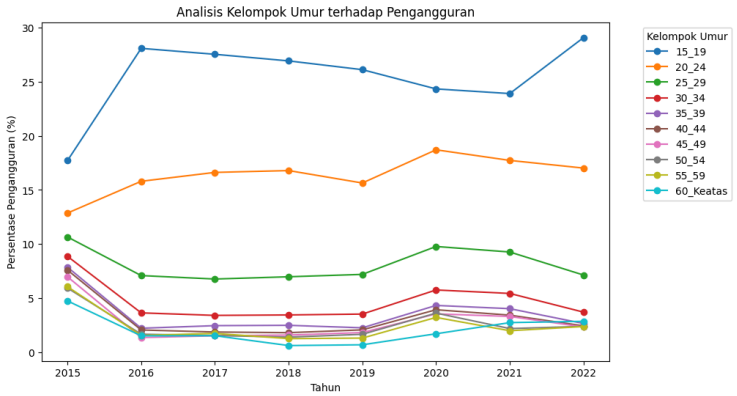
- Visualisasi
  - Line Chart



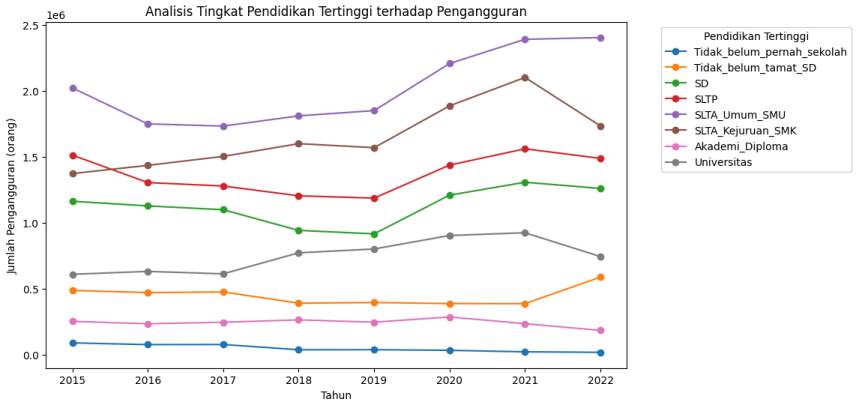
Interpretasi: berdasarkan gambar di atas dapat diketahui bahwa tren tingkat pengangguran cenderung mengalami banyak fluktuasi di setiap tahunnya. Pengangguran tertinggi terjadi pada tahun 2015 dan mengalami tingkat pengangguran terendah pada tahun 2022.



Interpretasi: berdasarkan gambar di atas, dapat diketahui bahwa tingkat pengangguran berdasarkan jenis kelamin mengalami lonjakan tingkat pengangguran pada tahun 2020 pada kedua variabel dan terus menurun hingga tahun 2022.



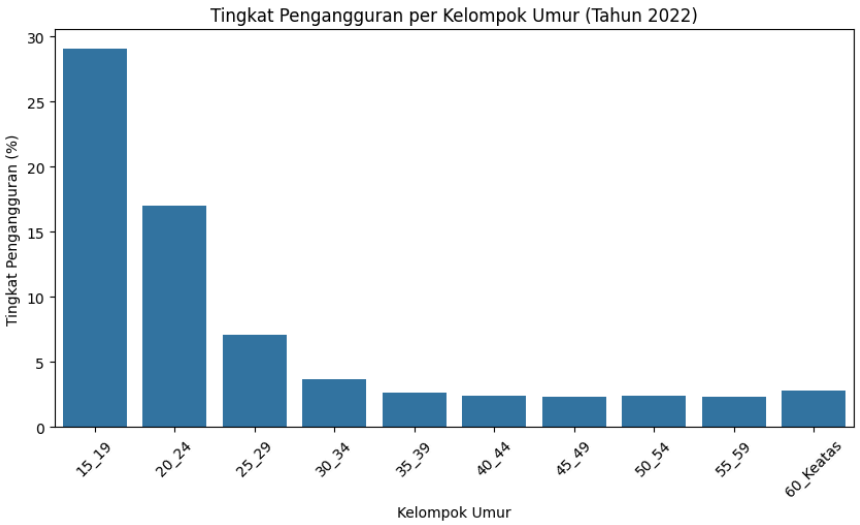
Interpretasi: berdasarkan gambar di atas, dapat diketahui bahwa tingkat pengangguran berdasarkan kelompok umur memiliki persentase pengangguran pada kelompok umur 15-19 tahun setiap tahunnya, disusul dengan usia 20-24 tahun dan 25-29 tahun.



Interpretasi: berdasarkan gambar di atas, dapat diketahui bahwa tingkat pengangguran berdasarkan tingkat pendidikan terakhir memiliki jumlah pengangguran tertinggi pada variabel SLTA Umum/SMU dan SLTA Kejuruan/SMK setiap tahunnya pada periode 2015-2022.

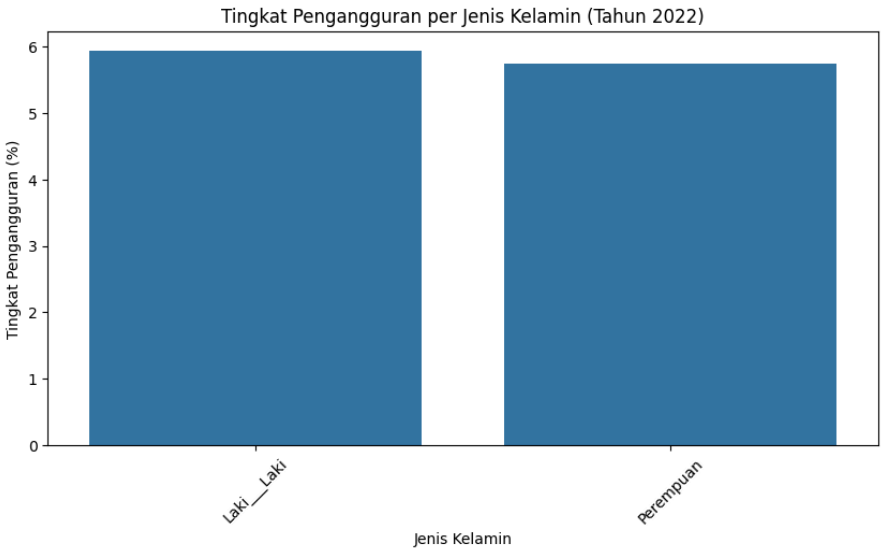
.

- Bar Chart



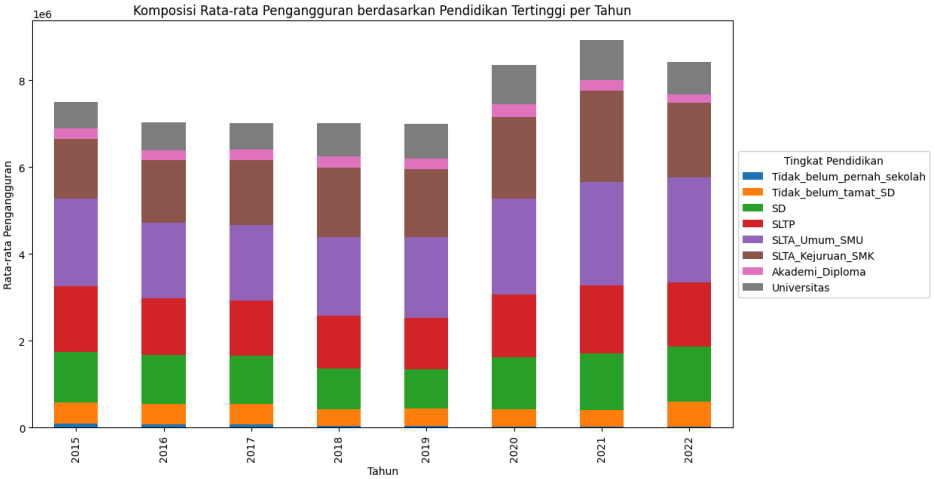
Interpretasi: berdasarkan gambar di atas dapat diketahui bahwa pada tahun 2022, tingkat pengangguran tertinggi terjadi pada kelompok umur 15-19 tahun kemudian diikuti oleh kelompok umur 20-24 tahun.

Interpretasi: berdasarkan gambar di atas dapat diketahui bahwa pada tahun 2022, tingkat pengangguran tertinggi terjadi pada lulusan SLTA Umum/SMU dan tingkat pengangguran terjadi pada lulusan akademi diploma.

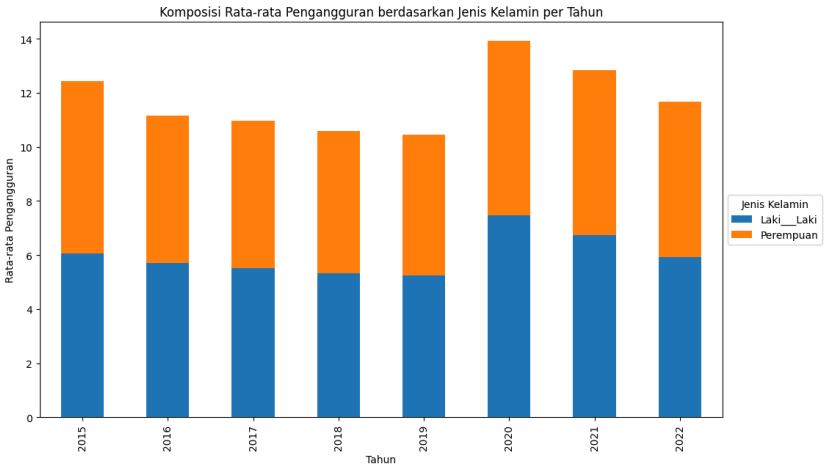


Interpretasi: berdasarkan gambar di atas dapat diketahui bahwa pada tahun 2022, kedua variabel memiliki selisih yang kecil. Namun, variabel laki-laki menjadi variabel dengan tingkat pengangguran tertinggi.

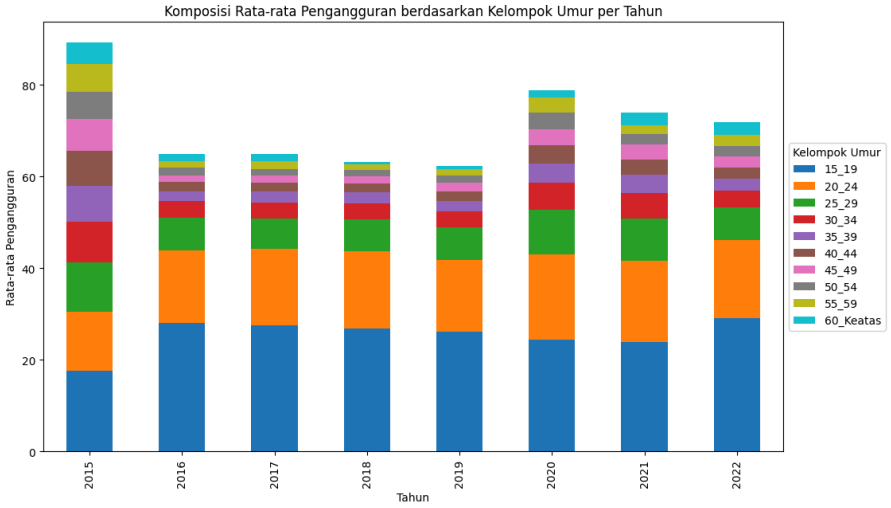
- Stacked Bar Chart



Interpretasi: berdasarkan gambar di atas dapat diketahui bahwa pada setiap tahunnya pada periode 2015-2022, lulusan SLTA Umum/SMU dan SLTA Kejuruan/SMK memiliki proporsi tingkat pengangguran yang tinggi dibandingkan dengan lulusan tingkat pendidikan lainnya.

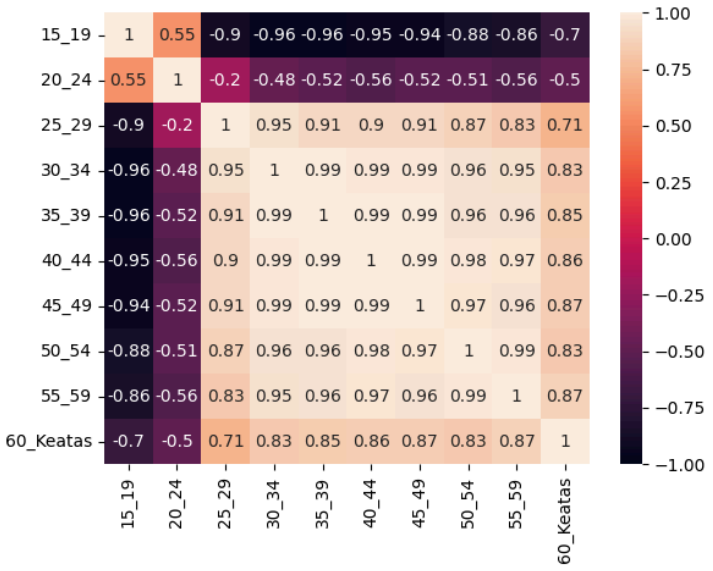


Interpretasi: berdasarkan gambar di atas dapat diketahui bahwa pada setiap tahunnya pada periode 2015-2022, kedua jenis kelamin memiliki selisih proporsi yang cenderung kecil. Namun, jika diamati kembali tingkat pengangguran pada jenis kelamin laki-laki cenderung memiliki nilai yang lebih tinggi dibandingkan tingkat pengangguran pada jenis kelamin perempuan.

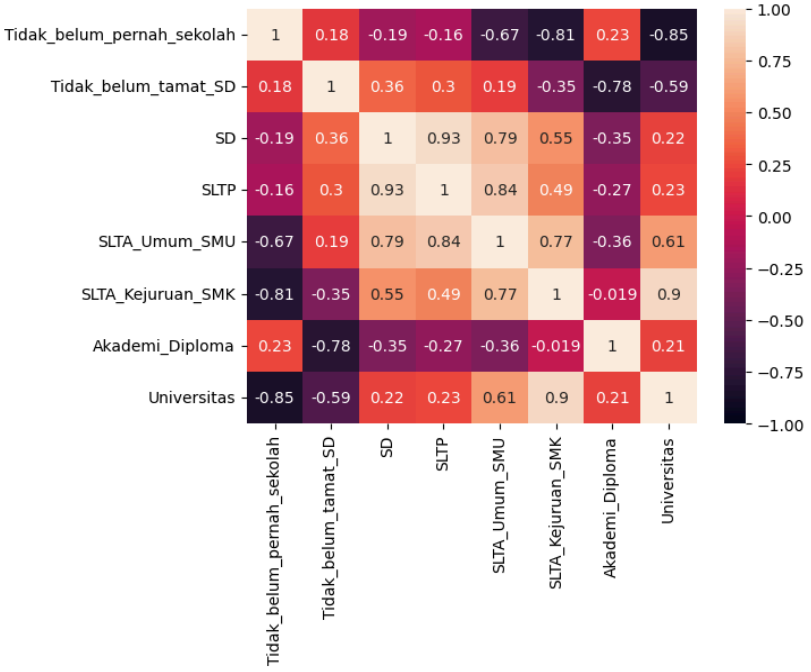


Interpretasi: berdasarkan gambar di atas dapat diketahui bahwa pada setiap tahunnya pada periode 2015-2022, kelompok umur 15-19 tahun dan 20-24 tahun memiliki tingkat pengangguran yang lebih tinggi dibandingkan kelompok umur lainnya.

- Heatmap

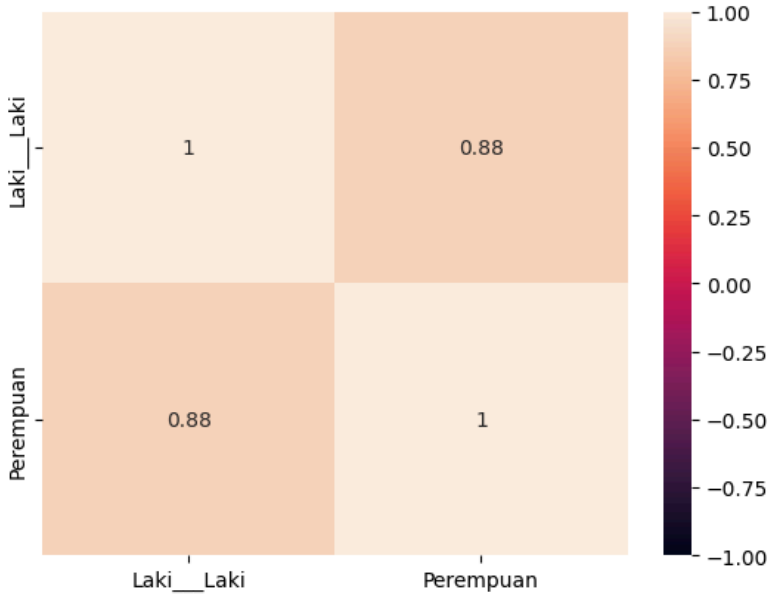


Interpretasi: berdasarkan kelompok umur, TPT memiliki korelasi kuat hubungan baik positif maupun negatif. Korelasi positif menunjukkan jika kelompok umur tertentu naik, maka TPT kelompok umur lain juga akan meningkat. Namun, jika korelasi menunjukkan negatif maka jika TPT kelompok umur tertentu maka kelompok umur lainnya akan meningkat. Korelasi negatif salah satunya ditunjukkan pada kelompok umur 15-19 dengan 25-29 dan korelasi positif salah satunya ditunjukkan pada kelompok umur 25-29 dengan 30-34.



Interpretasi: berdasarkan tingkat pendidikan terakhir, TPT memiliki korelasi kuat hubungan baik positif maupun negatif. Korelasi positif menunjukkan jika kelompok umur tertentu naik, maka TPT kelompok umur lain juga akan meningkat. Namun, jika korelasi menunjukkan negatif maka jika TPT

kelompok umur tertentu maka kelompok umur lainnya akan meningkat. Korelasi negatif salah satunya ditunjukkan pada tingkat pendidikan SD dengan Akademi/Diploma dan korelasi positif salah satunya ditunjukkan pada tingkat pendidikan SLTP dengan SD.



Interpretasi: berdasarkan jenis kelamin, TPT memiliki korelasi kuat hubungan positif pada kedua variabel. Artinya, jika TPT pada laki-laki meningkat maka TPT pada perempuan juga cenderung meningkat.

2. 4 Data publishing

Seluruh file data yang digunakan dalam penelitian ini, seperti source code pemrosesan, hasil proses wrangling, serta dokumentasi proses wrangling, dipublikasikan secara terbuka melalui repositori GitHub yang dapat diakses melalui link berikut: Projek Akhir Data Wrangling Kelompok 3_2024C.

2. 4.1 Raw data

- Dataset Tingkat Pengangguran di Indonesia

  Dataset ini didapatkan dari situs web YCharts yang menyajikan tingkat pengangguran di Indonesia (tahunan). Data dalam situs web ini disajikan dalam bentuk persentase per tahun dan terdiri dari dua kolom yaitu “Date/Year” dan “Value”. Dataset yang digunakan yaitu dari periode 2015–2022 dan dataset dapat diakses melalui link berikut: Indonesia Unemployment Rate.

- Dataset Tingkat Pengangguran Terbuka berdasarkan kelompok umur

  Dataset ini didapatkan dari Badan Pusat Statistik yang menyajikan persentase pengangguran yang dikelompokkan dalam rentang usia tertentu. Data yang dihasilkan dari web ini berbentuk deret waktu tahunan. Data ini terdiri dari kolom tahun dan kelompok umur (15–19 tahun, 20–24 tahun, 25–29 tahun, 30–34 tahun, 35–39 tahun,

40–44 tahun, 45–49 tahun, 50–54 tahun, 55–59 tahun, dan 60 tahun ke atas). Dataset yang digunakan yaitu dari periode 2015–2022 dan dataset dapat diakses melalui link berikut: TPT Berdasarkan Kelompok Umur.

- Dataset Tingkat Pengangguran Terbuka berdasarkan jenis kelamin

  Dataset ini didapatkan dari Badan Pusat Statistik yang menyajikan persentase pengangguran berdasarkan jenis kelamin. Kolom yang ditampilkan dalam dataset ini adalah tahun, laki-laki, dan perempuan. Dataset yang digunakan yaitu dari periode 2015–2022 dan dataset dapat diakses melalui link berikut: TPT Berdasarkan Jenis Kelamin.

- Dataset Tingkat Pengangguran Terbuka berdasarkan tingkat pendidikan

  Dataset ini didapatkan dari situs web Kaggle yang menyajikan jumlah pengangguran di Indonesia berdasarkan jenjang pendidikan tertinggi. Kolom yang ditampilkan dalam dataset ini adalah tahun, bulan, dan jenjang pendidikan (tidak/belum pernah sekolah, tidak/belum tamat, SD, SLTP, SLTA Umum/SMU, SLTA Kejuruan/SMK, Akademi/Diploma, dan Universitas. Dataset yang digunakan yaitu dari periode 2015–2022 dan dataset dapat diakses melalui link berikut: Dataset Pengangguran Indonesia Tingkat Pendidikan.

2.4.2 Penjelasan Fitur

- Year

  Kolom ini menunjukkan tahun pengamatan pada dataset hasil proses wrangling, kolom ini terdiri dari delapan tahun pengamatan. Tipe data pada kolom ini adalah numerik (float64).

- Kelompok umur

  Kolom-kolom pada kelompok umur menggambarkan jumlah Tingkat Pengangguran Terbuka (TPT) berdasarkan kelompok usia tertentu yaitu dari rentang umur 15–60 ke atas. Semua tipe data pada kolom ini adalah numerik (float64). Kolom-kolom pada kelompok umur sebagai berikut:
  - 15_19: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) pada usia 15–19 tahun.
  - 20_24: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) pada usia 20–24 tahun.
  - 25_29: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) pada usia 25–29 tahun.
  - 30_34: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) pada usia 30–34 tahun.
  - 35_39: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) pada usia 35–39 tahun.
  - 40_44: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) pada usia 40–44 tahun.
  - 45_49:kolom ini berisi Tingkat Pengangguran Terbuka (TPT) pada usia 45–49 tahun.
  - 50_54: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) pada usia 50–54 tahun.
  - 55_59: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) pada usia 55–59 tahun.
  - 60_Keatas: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) pada usia 60 tahun ke atas

- Pendidikan terakhir

  Kolom-kolom pada kelompok umur menggambarkan jumlah Tingkat Pengangguran Terbuka (TPT) berdasarkan tingkat pendidikan terakhir seseorang. Semua tipe data pada kolom ini adalah numerik (float64). Kolom-kolom pada pendidikan terakhir sebagai berikut:
  - Tidak_belum_pernah_sekolah: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) yang tidak atau belum pernah bersekolah.
  - Tidak_belum_tamat_SD: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) yang tidak atau belum belum menyelesaikan pendidikan SD.
  - SD: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) dengan pendidikan terakhir Sekolah Dasar (SD).

- SLTP: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) dengan pendidikan terakhir SMP/SLTP.
- SLTA_Umum_SMU: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) dengan pendidikan terakhir SMA/SLTA Umum.
- SLTA_Kejuruan_SMK: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) dengan pendidikan terakhir SMK/SLTA Kejuruan.
- Akademi_Diploma: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) dengan pendidikan terakhir Diploma (D1-D3).
- Universitas: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) dengan pendidikan terakhir Perguruan Tinggi (S1/S2/S3).

- Jenis Kelamin

  Kolom-kolom pada kelompok umur menggambarkan jumlah Tingkat Pengangguran Terbuka (TPT) berdasarkan jenis kelamin seseorang. Semua tipe data pada kolom ini adalah numerik (float64). Kolom-kolom pada jenis kelamin sebagai berikut:
  - Laki___Laki: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) pada jenis kelamin laki-laki
  - Perempuan: kolom ini berisi Tingkat Pengangguran Terbuka (TPT) pada jenis kelamin perempuan.

- Data_Pengangguran

  Kolom ini berisi jumlah total pengangguran di Indonesia setiap tahunnya, yaitu dari 2015–2022. Tipe data pada kolom ini adalah numerik (float64).

2.4.3 Pipeline

BAB IV
KENDALA DAN RENCANA

4.1 Kendala

- Perubahan ide dan data yang digunakan

  Pada awal eksplorasi ide kami sudah menentukan judul dan data yang akan digunakan. Namun, dalam pengumpulan data terjadi perubahan judul dan beberapa data yang akan digunakan, hal tersebut disebabkan oleh terbatasnya data yang mendukung judul sebelumnya. Dari kendala tersebut, kami memutuskan mengganti judul yang digunakan dan menyesuaikannya dengan data yang ditemukan.

- Keterbatasan sumber data tingkat provinsi

  Pada judul sebelumnya, kami menggunakan data berdasarkan provinsi di Indonesia, tapi dalam proses pengumpulan datanya, sumber yang menyediakan sebagian besar hanya tersedia pada satu sumber utama, yaitu BPS. Dari kendala tersebut, data berdasarkan provinsi kami ganti menjadi data berdasarkan data historis, sebab data historis lebih mudah ditemukan selain dari BPS.

- Ketidaklengkapan data historis

  Dalam proses pengumpulan data, seringkali data yang ditemukan memiliki rentang historis yang berbeda-beda. Pada sumber seperti BPS, data historis biasanya lebih lengkap untuk seluruh tahun yang dibutuhkan. Namun pada sumber lain seperti Kaggle, data historis terbatas sehingga dapat menyulitkan proses integrasi karena memungkinkan muncul missing values pada data. Dari kendala tersebut, dipilih rentang tahun yang memiliki kesediaan data paling lengkap untuk seluruh variabel, yaitu dari tahun 2015–2022, sehingga proses selanjutnya memungkinkan untuk dilakukan dan tidak mengandung missing values yang besar.

- Kesulitan dalam menggabungkan data

  Salah satu kendala utama dalam proses wrangling adalah perbedaan struktur kolom pada dataset dari kata kunci yang berbeda. Misalnya, dua data memiliki kolom provinsi yang sama tapi kolom satunya tidak punya. jadi, dengan perbedaan kolom yang berbeda membuat kita kesulitan dalam menggabungkan hingga perlu mengganti beberapa dataset. Perbedaan ini menyebabkan proses penggabungan (merge/concat) menjadi sulit karena data tidak dapat disejajarkan secara otomatis.

- Kesulitan dalam mencari *dataset*

  Salah satu kendala yang muncul dalam proses pengumpulan data adalah sulitnya menemukan *dataset* yang konsisten dan lengkap dari berbagai sumber. Setiap sumber data, seperti BPS, Kaggle, dan lain lain seringkali memiliki format yang berbeda. Meskipun banyak mendapatkan data dari sumber yang berbeda tapi juga sulit buat digabungkan kalau datanya tidak sama.

4.2 Rencana

- Melakukan Data Validation (Pemeriksaan Kembali Hasil Wrangling)

  Pada saat semua proses wrangling dilakukan kita perlu mengecek apakah data itu sudah benar tidak ada missing value yang tertinggal, tipe data sudah benar (numeric, category, datetime), kategori seperti pendidikan/umur/gender sudah sesuai, hasil merge tidak menimbulkan duplikasi atau baris hilang. Ini memastikan kualitas data benar-benar terjamin.
- Melakukan Feature Engineering

  Pada tahap ini dilakukan pembuatan fitur-fitur baru dari kolom yang sudah ada untuk memperkaya informasi dalam dataset. Proses ini mencakup membuat kategori umur, mengelompokkan tingkat pendidikan, membuat rasio atau perbandingan antar variabel, serta melakukan transformasi sederhana seperti normalisasi.
- Melakukan Transformasi Data

  Pada tahap ini dilakukan penyesuaian data agar format dan skala antar fitur lebih seragam. Transformasi dapat berupa normalisasi, standarisasi, atau pengubahan distribusi variabel tertentu agar lebih stabil.
- Melakukan Forecasting (ARIMA)

  Tahap ini akan dilakukan peramalan menggunakan model ARIMA *(AutoRegressive Integrated Moving Average)*. Analisis ini dipilih untuk melihat gambaran Tingkat Pengangguran Terbuka (TPT) di Indonesia di masa mendatang. Hasil peramalan nantinya akan dibandingkan antara data historis dengan data prediksi dengan visualisasi plot.
- Melakukan Analisis Regresi

  Analisis lanjutan lainnya yang bisa dilakukan dari data hasil proses wrangling kami adalah analisis regresi. Analisis ini dipilih untuk melihat pengaruh dari faktor-faktor demografis terhadap Tingkat Pengangguran Terbuka di Indonesia. Pada analisis ini, variabel “Data_Pengangguran” akan digunakan sebagai variabel dependen (Y) sedangkan variabel demografis seperti kelompok umur, tingkat pendidikan terakhir, dan jenis kelamin digunakan sebagai variabel independen (X).

BAB IV
KESIMPULAN

Berdasarkan hasil analisis terhadap data Tingkat Pengangguran Terbuka (TPT) di Indonesia dari tahun 2015–2022, dapat disimpulkan bahwa faktor demografis seperti umur, tingkat pendidikan, dan gender memiliki pengaruh yang signifikan terhadap variasi tingkat pengangguran. Data menunjukkan bahwa kelompok umur muda, khususnya usia 15–24 tahun, cenderung memiliki tingkat pengangguran yang lebih tinggi dibandingkan kelompok umur lainnya. Hal ini mengindikasikan bahwa keterbatasan pengalaman kerja dan keterampilan yang belum matang menjadi penyebab utama tingginya pengangguran pada usia muda. Selain itu, tingkat pendidikan juga menjadi faktor yang menentukan, di mana lulusan menengah seperti SMA dan SMK cenderung memiliki tingkat pengangguran yang lebih tinggi akibat ketidaksesuaian antara kompetensi lulusan dengan kebutuhan pasar kerja.

Perbedaan gender turut memberikan kontribusi terhadap variasi tingkat pengangguran. Dalam beberapa tahun, laki-laki cenderung menunjukkan tingkat pengangguran yang relatif lebih tinggi dibandingkan perempuan, kondisi tersebut berkaitan dengan persaingan yang lebih ketat pada sektor-sektor pekerjaan yang didominasi tenaga kerja laki-laki dan pergeseran kebutuhan kompetensi. Secara keseluruhan, hasil analisis menunjukkan bahwa dinamika pengangguran di Indonesia tidak hanya dipengaruhi faktor ekonomi, tetapi juga sangat dipengaruhi oleh karakteristik demografis, sehingga kebijakan penanggulangan pengangguran perlu disesuaikan dengan kelompok demografis tertentu untuk mencapai efektivitas yang lebih optimal.

DAFTAR PUSTAKA

Astuti, W. I., Ratnasari, V., & Wibowo, W. (2017). Analisis Faktor yang Berpengaruh Terhadap Tingkat Pengangguran Terbuka di Provinsi Jawa Timur Menggunakan Regresi Data Panel. *Jurnal Sains dan Seni ITS*, *6*(1), 144–149. https://doi.org/10.12962/j23373520.v6i1.22977

Manik, C. W., Giting, H. N. B., Aini, L., & Hidayat, N. (2025). *Analisis Bonus Demografi Ditengah Tingginya Pengangguran Terdidik di Indonesia*.

Novianti, E. (2018). *KESENJANGAN GENDER TINGKAT PENGANGGURAN TERBUKA DI INDONESIA*. Universitas Negeri Yogyakarta.

Priastiwi, D. (2018). *ANALISIS PENGARUH JUMLAH PENDUDUK, PENDIDIKAN, UPAH MINIMUM, DAN PDRB TERHADAP TINGKAT PENGANGGURAN TERBUKA DI PROVINSI JAWA TENGAH*. Universitas Diponegoro.

Saragih, R. C., Eva Sriwiyanti, & Vitryani Tarigan. (2021). Pengaruh Faktor Demografi (Usia, Jenis Kelamin Dan Tingkat Pendidikan) Terhadap Kepatuhan Wajib Pajak UMKM Di Kecamatan Siantar Barat. *Jurnal Ilmiah Accusi*, *3*(2), 117–123. https://doi.org/10.36985/jia.v3i2.130

Wijaya, M. O., & Utami, E. D. (2021). Determinan Pengangguran Lulusan SMK di Indonesia Tahun 2020. *Seminar Nasional Official Statistics*, *2021*(1), 801–810. https://doi.org/10.34123/semnasoffstat.v2021i1.1048