

eda

October 4, 2024

0.1 # Tugas 1 Data Mining - EDA

List Anggota Kelompok 9 * Cahaya Aulia Firdausyah (2006304) * Anderfa Jalu Kawani (2102671)
* Sabila Rosad (2106000)

0.2 Import Library

```
[ ]: %matplotlib inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from google.colab import drive
```

0.3 Load Dataset

Import dan load dataset yang akan digunakan dari GDrive

```
[ ]: drive.mount('/content/drive')

df_jml_hotel = pd.read_excel("/content/drive/MyDrive/Data Mining/Tugas 1/
↳disparbud-od_15356_jml_hotel_bintang_non_bintang__jenis_hotel_kabupat_v1_data.
↳xlsx")
df_jml_pengunjung_pariwisata = pd.read_excel("/content/drive/MyDrive/Data_
↳Mining/Tugas 1/
↳disparbud-od_15361_jml_pengunjung_kawasan_pariwisata__jenis_wisatawan_data.
↳xlsx")
df_jml_pariwisata = pd.read_excel("/content/drive/MyDrive/Data Mining/Tugas 1/
↳disparbud-od_15362_jml_kawasan_pariwisata__kabupatenkota_v1_data.xlsx")
df_jml_pendapatan = pd.read_excel("/content/drive/MyDrive/Data Mining/Tugas 1/
↳disparbud-od_15380_jml_pendapatan_asli_drh_bidang_pariwisata__sektor_wisa_data.
↳xlsx")
df_jml_potensi_odtw = pd.read_excel("/content/drive/MyDrive/Data Mining/Tugas 1/
↳disparbud-od_15387_jml_ptns_obyek_daya_tarik_wisata_odtw__jenis_kabup_v2_data.
↳xlsx")
df_jml_rumah_makan = pd.read_excel("/content/drive/MyDrive/Data Mining/Tugas 1/
↳disparbud-od_15393_jumlah_rumah_makan_berdasarkan_kabupatenkota_v1_data.
↳xlsx")
```

```

df_jml_pengunjung_perkemahan = pd.read_excel("/content/drive/MyDrive/Data Mining/Tugas 1/
↳disparbud-od_16111_jml_pengunjung_perkemahan__jenis_wisatawan_kabupat_v2_data.
↳xlsx")
df_jml_pengunjung_homestay = pd.read_excel("/content/drive/MyDrive/Data Mining/
↳Tugas 1/
↳disparbud-od_17834_jml_pengunjung_homestay__jenis_wisatawan_kabupaten_v2_data.
↳xlsx")
df_jml_tk_pariwisata = pd.read_excel("/content/drive/MyDrive/Data Mining/Tugas 1/
↳disparbud-od_jml_tk_kawasan_pariwisata__jk_data.xlsx")
df_luas_pariwisata = pd.read_excel("/content/drive/MyDrive/Data Mining/Tugas 1/
↳disparbud-od_kawasan_pariwisata_berdasarkan_luas_data.xlsx")
df_jml_kawasan_pariwisata = pd.read_excel("/content/drive/MyDrive/Data Mining/
↳Tugas 1/disparbud-od_15362_jml_kawasan_pariwisata__kabupatenkota_v1_data.
↳xlsx")

```

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

0.4 Drop Atribut

Melakukan Drop (hapus) pada atribut yang tidak digunakan, seperti: * id * kode_provinsi * nama_provinsi * satuan

```

[ ]: df_jml_hotel = df_jml_hotel.drop(columns=["id", "kode_provinsi",
↳"nama_provinsi", "satuan"])
df_jml_pengunjung_pariwisata = df_jml_pengunjung_pariwisata.drop(columns=["id",
↳"kode_provinsi", "nama_provinsi", "satuan"])
df_jml_pariwisata = df_jml_pariwisata.drop(columns=["id", "kode_provinsi",
↳"nama_provinsi", "satuan"])
df_jml_pendapatan = df_jml_pendapatan.drop(columns=["id", "kode_provinsi",
↳"nama_provinsi", "satuan"])
df_jml_potensi_odtw = df_jml_potensi_odtw.drop(columns=["id", "kode_provinsi",
↳"nama_provinsi", "satuan"])
df_jml_rumah_makan = df_jml_rumah_makan.drop(columns=["id", "kode_provinsi",
↳"nama_provinsi", "satuan"])
df_jml_pengunjung_perkemahan = df_jml_pengunjung_perkemahan.drop(columns=["id",
↳"kode_provinsi", "nama_provinsi", "satuan"])
df_jml_pengunjung_homestay = df_jml_pengunjung_homestay.drop(columns=["id",
↳"kode_provinsi", "nama_provinsi", "satuan"])
df_jml_tk_pariwisata = df_jml_tk_pariwisata.drop(columns=["id",
↳"kode_provinsi", "nama_provinsi", "satuan"])
df_luas_pariwisata = df_luas_pariwisata.drop(columns=["id", "kode_provinsi",
↳"nama_provinsi", "satuan"])
df_jml_kawasan_pariwisata = df_jml_kawasan_pariwisata.drop(columns=["id",
↳"kode_provinsi", "nama_provinsi", "satuan"])

```

0.5 Memfilter tiap dataset

Melakukan filterasi agar interval tiap data set sama dari tahun 2018 sampai 2023.

```
[ ]: con1 = (df_jml_hotel['tahun'] >= 2018) & (df_jml_hotel['tahun'] <= 2023)
df_jml_hotel = df_jml_hotel[con1].reset_index(drop=True)
df_jml_hotel.to_csv('filtered_data.csv', index=False)

con1 = (df_jml_pengunjung_pariwisata['tahun'] >= 2018) &
↳(df_jml_pengunjung_pariwisata['tahun'] <= 2023)
df_jml_pengunjung = df_jml_pengunjung_pariwisata[con1].reset_index(drop=True)
df_jml_pengunjung.to_csv('filtered_data.csv', index=False)

con1 = (df_jml_pariwisata['tahun'] >= 2018) & (df_jml_pariwisata['tahun'] <=
↳2023)
df_jml_pariwisata = df_jml_pariwisata[con1].reset_index(drop=True)
df_jml_pariwisata.to_csv('filtered_data.csv', index=False)

con1 = (df_jml_pendapatan['tahun'] >= 2018) & (df_jml_pendapatan['tahun'] <=
↳2023)
df_jml_pendapatan = df_jml_pendapatan[con1].reset_index(drop=True)
df_jml_pendapatan.to_csv('filtered_data.csv', index=False)

con1 = (df_jml_potensi_odtw['tahun'] >= 2018) & (df_jml_potensi_odtw['tahun']
↳<= 2023)
df_jml_potensi_odtw = df_jml_potensi_odtw[con1].reset_index(drop=True)
df_jml_potensi_odtw.to_csv('filtered_data.csv', index=False)

con1 = (df_jml_rumah_makan['tahun'] >= 2018) & (df_jml_rumah_makan['tahun'] <=
↳2023)
df_jml_rumah_makan = df_jml_rumah_makan[con1].reset_index(drop=True)
df_jml_rumah_makan.to_csv('filtered_data.csv', index=False)

con1 = (df_jml_pengunjung_perkemahan['tahun'] >= 2018) &
↳(df_jml_pengunjung_perkemahan['tahun'] <= 2023)
df_jml_pengunjung_perkemahan = df_jml_pengunjung_perkemahan[con1].
↳reset_index(drop=True)
df_jml_pengunjung_perkemahan.to_csv('filtered_data.csv', index=False)

con1 = (df_jml_pengunjung_homestay['tahun'] >= 2018) &
↳(df_jml_pengunjung_homestay['tahun'] <= 2023)
df_jml_pengunjung_homestay = df_jml_pengunjung_homestay[con1].
↳reset_index(drop=True)
df_jml_pengunjung_homestay.to_csv('filtered_data.csv', index=False)

con1 = (df_jml_tk_pariwisata['tahun'] >= 2018) & (df_jml_tk_pariwisata['tahun']
↳<= 2023)
```

```
df_jml_tk_pariwisata = df_jml_tk_pariwisata[con1].reset_index(drop=True)
df_jml_tk_pariwisata.to_csv('filtered_data.csv', index=False)

con1 = (df_jml_kawasan_pariwisata['tahun'] >= 2018) &
↳ (df_jml_kawasan_pariwisata['tahun'] <= 2023)
df_jml_kawasan_pariwisata = df_jml_kawasan_pariwisata[con1].
↳ reset_index(drop=True)
df_jml_kawasan_pariwisata.to_csv('filtered_data.csv', index=False)
```

0.6 Mengganti nama atribut

Mengganti nama atribut agar tidak memiliki kesamaan nama saat melakukan penggabungan (merge)

```
[ ]: df_jml_pengunjung_perkemahan.rename(columns={"jumlah_pengunjung":
↳ "jumlah_pengunjung_perkemahan"}, inplace=True)
df_jml_pengunjung_homestay.rename(columns={"jumlah_pengunjung":
↳ "jumlah_pengunjung_homestay"}, inplace=True)
df_jml_pengunjung_pariwisata.rename(columns={"jumlah_pengunjung":
↳ "jumlah_pengunjung_pariwisata"}, inplace=True)
df_jml_pariwisata.rename(columns={"jumlah_kawasan":
↳ "jumlah_kawasan_pariwisata"}, inplace=True)
df_jml_potensi_odtw.rename(columns={"jumlah_odtw": "jumlah_potensi_odtw"},
↳ inplace=True)
df_jml_tk_pariwisata.rename(columns={"jumlah_tenaga_kerja":
↳ "jumlah_tk_pariwisata"}, inplace=True)
```

0.7 Data Exploration

Mengeksplorasi setiap dataset yang digunakan

Mengeksplorasi dataset jumlah hotel

```
[ ]: df_jml_hotel.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 945 entries, 0 to 944
Data columns (total 5 columns):
#   Column                Non-Null Count  Dtype
---  -
0   kode_kabupaten_kota    945 non-null    int64
1   nama_kabupaten_kota    945 non-null    object
2   jenis_hotel            945 non-null    object
3   jumlah_hotel           945 non-null    int64
4   tahun                 945 non-null    int64
dtypes: int64(3), object(2)
memory usage: 37.0+ KB
```

Mengekplorasi dataset jumlah pengunjung pariwisata

```
[ ]: df_jml_pengunjung_pariwisata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 484 entries, 0 to 483
Data columns (total 5 columns):
#   Column                      Non-Null Count  Dtype
---  ---
0   kode_kabupaten_kota        484 non-null    int64
1   nama_kabupaten_kota        484 non-null    object
2   jenis_wisatawan            484 non-null    object
3   jumlah_pengunjung_pariwisata 483 non-null    float64
4   tahun                      484 non-null    int64
dtypes: float64(1), int64(2), object(2)
memory usage: 19.0+ KB
```

Mengekplorasi dataset jumlah pariwisata

```
[ ]: df_jml_pariwisata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 162 entries, 0 to 161
Data columns (total 4 columns):
#   Column                      Non-Null Count  Dtype
---  ---
0   kode_kabupaten_kota        162 non-null    int64
1   nama_kabupaten_kota        162 non-null    object
2   jumlah_kawasan_pariwisata  157 non-null    float64
3   tahun                      162 non-null    int64
dtypes: float64(1), int64(2), object(1)
memory usage: 5.2+ KB
```

Mengekplorasi dataset jumlah pendapatan asli daerah bidang pariwisata

```
[ ]: df_jml_pendapatan.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541 entries, 0 to 540
Data columns (total 5 columns):
#   Column                      Non-Null Count  Dtype
---  ---
0   kode_kabupaten_kota        541 non-null    int64
1   nama_kabupaten_kota        541 non-null    object
2   sektor_wisata              541 non-null    object
3   jumlah_pendapatan          541 non-null    int64
4   tahun                      541 non-null    int64
dtypes: int64(3), object(2)
memory usage: 21.3+ KB
```

Mengeksplorasi dataset jumlah potensi Obyek Daya Tarik Wisata (ODTW) dan jenisnya

```
[ ]: df_jml_potensi_odtw.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 486 entries, 0 to 485
Data columns (total 5 columns):
#   Column                Non-Null Count  Dtype
---  -
0   kode_kabupaten_kota    486 non-null    int64
1   nama_kabupaten_kota    486 non-null    object
2   jenis_odtw              486 non-null    object
3   jumlah_potensi_odtw    486 non-null    int64
4   tahun                  486 non-null    int64
dtypes: int64(3), object(2)
memory usage: 19.1+ KB
```

Mengeksplorasi dataset jumlah rumah makan

```
[ ]: df_jml_rumah_makan.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 162 entries, 0 to 161
Data columns (total 4 columns):
#   Column                Non-Null Count  Dtype
---  -
0   kode_kabupaten_kota    162 non-null    int64
1   nama_kabupaten_kota    162 non-null    object
2   jumlah_rumah_makan     162 non-null    float64
3   tahun                  162 non-null    int64
dtypes: float64(1), int64(2), object(1)
memory usage: 5.2+ KB
```

Mengeksplorasi dataset jumlah pengunjung perkemahan

```
[ ]: df_jml_pengunjung_perkemahan.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 324 entries, 0 to 323
Data columns (total 5 columns):
#   Column                Non-Null Count  Dtype
---  -
0   kode_kabupaten_kota    324 non-null    int64
1   nama_kabupaten_kota    324 non-null    object
2   jenis_wisatawan        324 non-null    object
3   jumlah_pengunjung_perkemahan  324 non-null    int64
4   tahun                  324 non-null    int64
dtypes: int64(3), object(2)
```

memory usage: 12.8+ KB

Mengekplorasi dataset jumlah pengunjung homestay

```
[ ]: df_jml_pengunjung_homestay.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 324 entries, 0 to 323
Data columns (total 5 columns):
#   Column                      Non-Null Count  Dtype
---  -
0   kode_kabupaten_kota         324 non-null    int64
1   nama_kabupaten_kota         324 non-null    object
2   jenis_wisatawan             324 non-null    object
3   jumlah_pengunjung_homestay  324 non-null    int64
4   tahun                       324 non-null    int64
dtypes: int64(3), object(2)
memory usage: 12.8+ KB
```

Mengekplorasi dataset jumlah tenaga kerja pariwisata

```
[ ]: df_jml_tk_pariwisata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 0 entries
Data columns (total 5 columns):
#   Column                      Non-Null Count  Dtype
---  -
0   kode_kabupaten_kota         0 non-null      int64
1   nama_kabupaten_kota         0 non-null      object
2   jenis_kelamin               0 non-null      object
3   jumlah_tk_pariwisata        0 non-null      int64
4   tahun                       0 non-null      int64
dtypes: int64(3), object(2)
memory usage: 124.0+ bytes
```

Mengekplorasi dataset luas kawasan pariwisata

```
[ ]: df_luas_pariwisata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 107 entries, 0 to 106
Data columns (total 4 columns):
#   Column                      Non-Null Count  Dtype
---  -
0   kode_kabupaten_kota         107 non-null    int64
1   nama_kabupaten_kota         107 non-null    object
2   luas_kawasan                107 non-null    float64
3   tahun                       107 non-null    int64
```

```
dtypes: float64(1), int64(2), object(1)
memory usage: 3.5+ KB
```

Mengeplorasi dataset jumlah kawasan pariwisata

```
[ ]: df_jml_kawasan_pariwisata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 162 entries, 0 to 161
Data columns (total 4 columns):
#   Column                Non-Null Count  Dtype
---  -
0   kode_kabupaten_kota    162 non-null   int64
1   nama_kabupaten_kota    162 non-null   object
2   jumlah_kawasan         157 non-null   float64
3   tahun                  162 non-null   int64
dtypes: float64(1), int64(2), object(1)
memory usage: 5.2+ KB
```

0.8 Menggabungkan dataset

Menggabungkan setiap dataset yang digunakan

```
[ ]: df_merged = pd.merge(df_jml_hotel, df_jml_pengunjung_pariwisata,
    ↪on=['kode_kabupaten_kota', 'nama_kabupaten_kota', 'tahun'], how='outer')
df_merged = pd.merge(df_merged, df_jml_pariwisata, on=['kode_kabupaten_kota',
    ↪'nama_kabupaten_kota', 'tahun'], how='outer')
df_merged = pd.merge(df_merged, df_jml_pendapatan, on=['kode_kabupaten_kota',
    ↪'nama_kabupaten_kota', 'tahun'], how='outer')
df_merged = pd.merge(df_merged, df_jml_potensi_odtw, on=['kode_kabupaten_kota',
    ↪'nama_kabupaten_kota', 'tahun'], how='outer')
df_merged = pd.merge(df_merged, df_jml_rumah_makan, on=['kode_kabupaten_kota',
    ↪'nama_kabupaten_kota', 'tahun'], how='outer')
df_merged = pd.merge(df_merged, df_jml_pengunjung_perkemahan,
    ↪on=['kode_kabupaten_kota', 'nama_kabupaten_kota', 'tahun'], how='outer')
df_merged = pd.merge(df_merged, df_jml_pengunjung_homestay,
    ↪on=['kode_kabupaten_kota', 'nama_kabupaten_kota', 'tahun'], how='outer')
df_merged = pd.merge(df_merged, df_jml_tk_pariwisata,
    ↪on=['kode_kabupaten_kota', 'nama_kabupaten_kota', 'tahun'], how='outer')
df_merged = pd.merge(df_merged, df_luas_pariwisata, on=['kode_kabupaten_kota',
    ↪'nama_kabupaten_kota', 'tahun'], how='outer')
df_merged = pd.merge(df_merged, df_jml_kawasan_pariwisata,
    ↪on=['kode_kabupaten_kota', 'nama_kabupaten_kota', 'tahun'], how='outer')

df_merged = df_merged.loc[:, [
    'kode_kabupaten_kota',
    'nama_kabupaten_kota',
```



```

'tahun',
'jumlah_hotel',
'jumlah_pengunjung_pariwisata',
'jumlah_pendapatan',
'jumlah_rumah_makan',
'jumlah_pengunjung_perkemahan',
'jumlah_pengunjung_homestay',
'jumlah_kawasan_pariwisata',
'jumlah_potensi_odtw',
'jenis_odtw',
'jumlah_tk_pariwisata',
]]

```

```
df_merged.head(10)
```

```

[ ]:  kode_kabupaten_kota  nama_kabupaten_kota  tahun  jumlah_hotel  \
0      3201      KABUPATEN BOGOR      2014      NaN
1      3201      KABUPATEN BOGOR      2014      NaN
2      3201      KABUPATEN BOGOR      2015      NaN
3      3201      KABUPATEN BOGOR      2015      NaN
4      3201      KABUPATEN BOGOR      2016      NaN
5      3201      KABUPATEN BOGOR      2016      NaN
6      3201      KABUPATEN BOGOR      2017      NaN
7      3201      KABUPATEN BOGOR      2017      NaN
8      3201      KABUPATEN BOGOR      2018      0.0
9      3201      KABUPATEN BOGOR      2018      0.0

    jumlah_pengunjung_pariwisata  jumlah_pendapatan  jumlah_rumah_makan  \
0              0.0              NaN              NaN
1              0.0              NaN              NaN
2              0.0              NaN              NaN
3              0.0              NaN              NaN
4              0.0              NaN              NaN
5              0.0              NaN              NaN
6              0.0              NaN              NaN
7              0.0              NaN              NaN
8      4411967.0              0.0              0.0
9      4411967.0              0.0              0.0

    jumlah_pengunjung_perkemahan  jumlah_pengunjung_homestay  \
0              NaN              NaN
1              NaN              NaN
2              NaN              NaN
3              NaN              NaN
4              NaN              NaN
5              NaN              NaN
6              NaN              NaN

```

7	NaN	NaN
8	0.0	0.0
9	0.0	0.0

	jumlah_kawasan_pariwisata	jumlah_potensi_odtw	jenis_odtw \
0	NaN	NaN	NaN
1	NaN	NaN	NaN
2	NaN	NaN	NaN
3	NaN	NaN	NaN
4	NaN	NaN	NaN
5	NaN	NaN	NaN
6	NaN	NaN	NaN
7	NaN	NaN	NaN
8	86.0	55.0	ALAM
9	86.0	55.0	ALAM

	jumlah_tk_pariwisata
0	NaN
1	NaN
2	NaN
3	NaN
4	NaN
5	NaN
6	NaN
7	NaN
8	NaN
9	NaN

0.9 Cek value null

Memeriksa apakah terdapat data yang kosong atau tidak

```
[ ]: df_merged.isnull().sum()
```

```
[ ]: kode_kabupaten_kota      0
nama_kabupaten_kota          0
tahun                        0
jumlah_hotel                  215
jumlah_pengunjung_pariwisata 1909
jumlah_pendapatan             1834
jumlah_rumah_makan            215
jumlah_pengunjung_perkemahan  215
jumlah_pengunjung_homestay    215
jumlah_kawasan_pariwisata     3095
jumlah_potensi_odtw            215
jenis_odtw                    215
jumlah_tk_pariwisata          79595
```

dtype: int64

0.10 Visualisasi

0.11 Bar chart

untuk memvisualisasi banyaknya hotel dengan jumlah pengunjung pariwisata

```
[64]: # Set ukuran plot
plt.figure(figsize=(10, 6))

# Buat bar plot jumlah hotel vs jumlah pengunjung pariwisata
sns.barplot(x='nama_kabupaten_kota', y='jumlah_hotel', data=df_merged,
            color='blue', label='Jumlah Hotel')
sns.barplot(x='nama_kabupaten_kota', y='jumlah_pengunjung_pariwisata',
            data=df_merged, color='red', label='Jumlah Pengunjung Pariwisata')

# Rotasi label kabupaten/kota untuk keterbacaan
plt.xticks(rotation=90)

# Tambahkan judul dan label
plt.title('Jumlah Hotel dan Pengunjung Pariwisata per Kabupaten/Kota')
plt.xlabel('Kabupaten/Kota')
plt.ylabel('Jumlah')

# Tampilkan legenda
plt.legend()

# Tampilkan plot
plt.tight_layout()
plt.show()
```

