DQLab

TETRIS PROGRAM

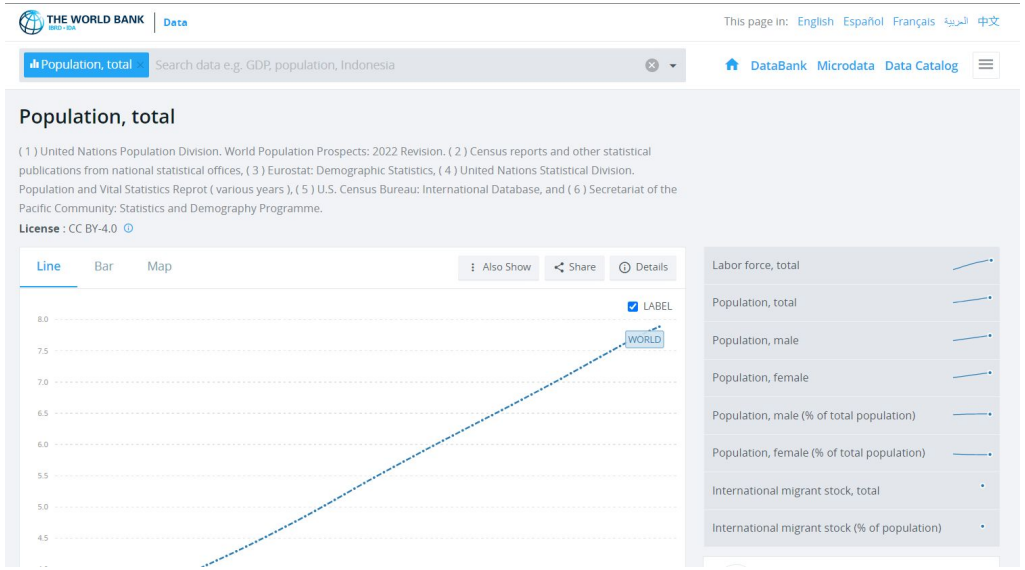
Analisis Demografi dan Karakteristik Setiap Provinsi di Indonesia

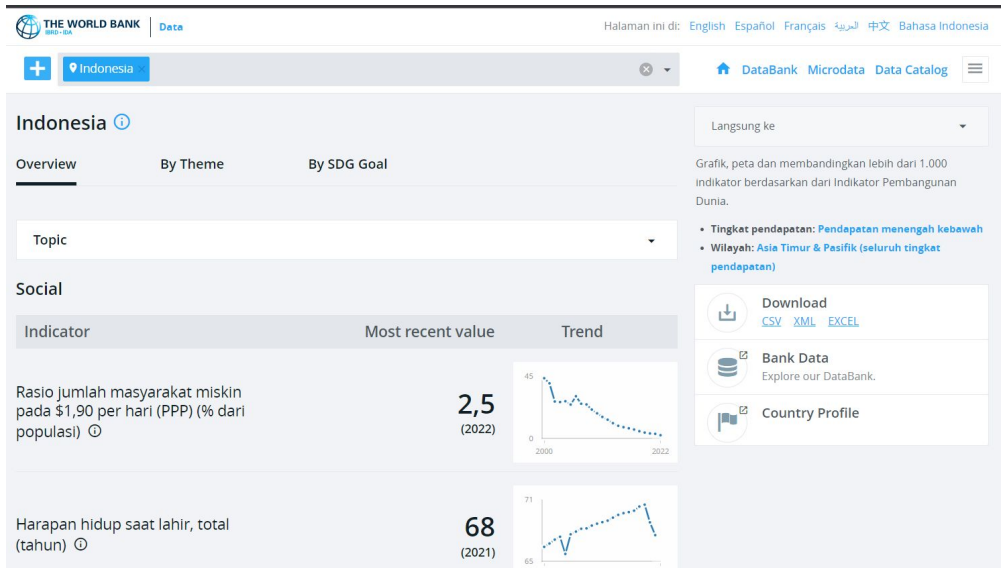
Ghana Ahmada Yudistira
ghanamada@gmail.com

#StackYourSkill

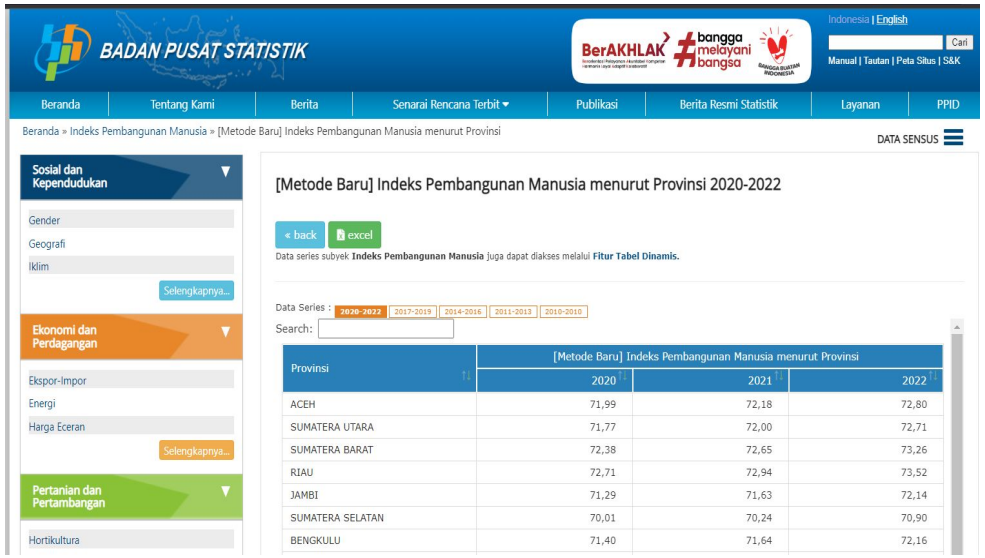
Step 1 - Data Collection & Data Integration

1. Mengambil data sekunder mengenai total populasi seluruh negara di dunia dan atribut demografi di Indonesia dari website [WorldBank]

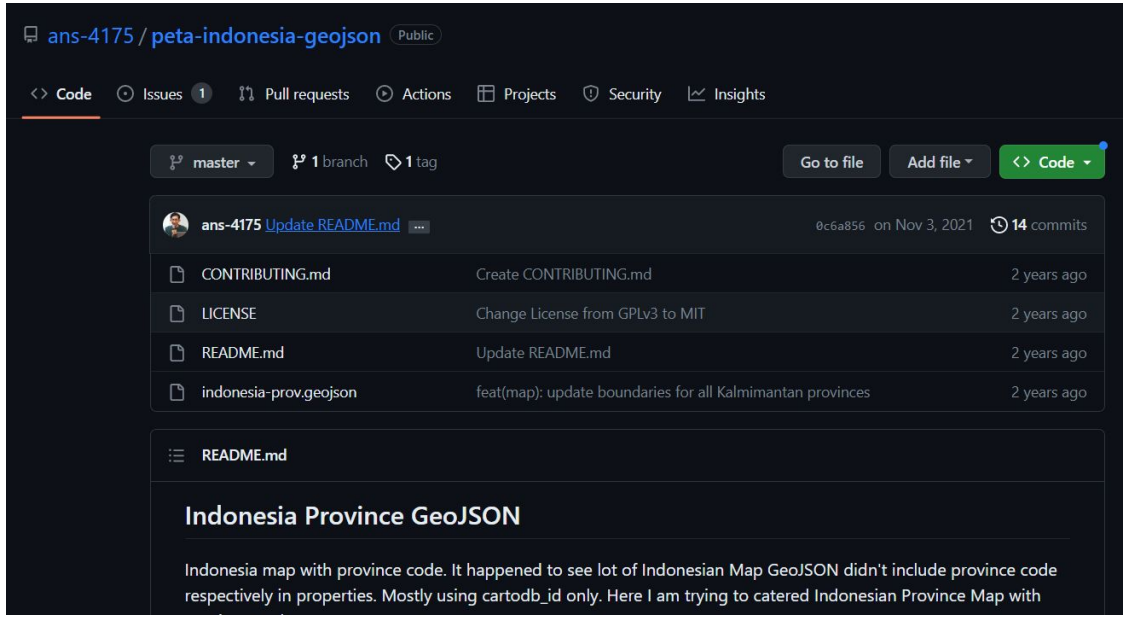




2. Mengambil data sekunder mengenai data umur penduduk, faktor sosial ekonomi, dan pembangunan dari website [Badan Pusat Statistik (BPS)]



3. Mengambil data pendukung dalam format .geojson untuk keperluan visualisasi peta Indonesia menggunakan plotly dari [github]



Step 2 - Data Cleansing

+ Code + Markdown

1. Data Populasi dan Atribut Demografi Indonesia

```
data_pop = pd.read_csv("raw_data/data_penduduk/population_worldwide.csv")
data_growth_pop = pd.read_csv("raw_data/data_penduduk/growth_population.csv").iloc[:, :-1]
indo_stats = pd.read_csv("raw_data/data_penduduk/statistik_indo.csv")
```

[5] ✓ 0.6s

```
tr_indo_stats = indo_stats.iloc[:, 2:-1].T.drop(["Indicator Name", "Indicator Code"]).iloc[40:, :]
tr_indo_stats.columns = indo_stats["Indicator Name"].values
tr_indo_stats = tr_indo_stats.reset_index()
```

[6] ✓ 0.0s

```
def transform_pop_data(df: pd.DataFrame, countries: str, value_col_name):
    new_df = pd.DataFrame()
    for country in countries:
        population_df = df[df["Country Name"] == country].iloc[:, 4:]
        reformed_pop = {"negara":[country for _ in population_df.columns],
                        "tahun":[tahun for tahun in population_df.columns],
                        value_col_name:[population_df[tahun].values[0] for tahun in population_df.columns]}
        new_df = pd.concat([new_df, pd.DataFrame(reformed_pop)])
    return new_df
```

[7] ✓ 0.0s

```
indo_population = transform_pop_data(data_pop, countries=["Indonesia"], value_col_name="total_penduduk").iloc[40:, :].reset_index(drop=True)

data_nasional = pd.merge(indo_population, tr_indo_stats, left_on="tahun", right_on="index").drop(["index"], axis=1)
data_nasional = data_nasional.dropna(axis=1)

# for col in data_nasional.columns:
#     data_nasional[col] = data_nasional[col].astype(float)
lst_2022 = ["Indonesia", "2022", 275361267.0] + [np.nan for _ in range(len(data_nasional.columns)-3)]
data_nasional.loc[len(data_nasional)] = lst_2022
data_nasional
```

[8] ✓ 0.1s

Transformasi data populasi dan atribut demografi di Indonesia. Lalu simpan dalam file csv untuk keperluan analisis

```
rank_pop = transform_pop_data(data_pop, countries=["Indonesia", "Pakistan", "United States", "China", "India"], value_col_name="total_penduduk")
growth_pop = transform_pop_data(data_growth_pop, countries=["Indonesia", "Pakistan", "United States", "China", "India"], value_col_name="pertumbuhan_penduduk").dropna()

rank_pop_21 = rank_pop[rank_pop["tahun"] == "2021"].reset_index(drop=True)
```

✓ 0.2s

```
data_nasional.to_csv("preprocessed_data/data_nasional.csv", index=False)
rank_pop_21.to_csv("preprocessed_data/peringkat_total_penduduk_dunia.csv", index=False)
growth_pop.to_csv("preprocessed_data/data_pertumbuhan_penduduk.csv", index=False)
```

TETRIS PROGRAM

2. Data Distribusi Umur Nasional

```
data_umur_0910 = pd.read_excel("raw_data\data_umur\Jumlah Penduduk Usia 15 tahun ke Atas Menurut Golongan Umur 2009-2010.xlsx")
data_umur_1112 = pd.read_excel("raw_data\data_umur\Jumlah Penduduk Usia 15 tahun ke Atas Menurut Golongan Umur 2011-2012.xlsx")
data_umur_1314 = pd.read_excel("raw_data\data_umur\Jumlah Penduduk Usia 15 tahun ke Atas Menurut Golongan Umur 2013-2014.xlsx")
data_umur_1516 = pd.read_excel("raw_data\data_umur\Jumlah Penduduk Usia 15 tahun ke Atas Menurut Golongan Umur 2015-2016.xlsx")
data_umur_1718 = pd.read_excel("raw_data\data_umur\Jumlah Penduduk Usia 15 tahun ke Atas Menurut Golongan Umur 2017-2018.xlsx")
data_umur_1920 = pd.read_excel("raw_data\data_umur\Jumlah Penduduk Usia 15 tahun ke Atas Menurut Golongan Umur 2019-2020.xlsx")
data_umur_2122 = pd.read_excel("raw_data\data_umur\Jumlah Penduduk Usia 15 tahun ke Atas Menurut Golongan Umur 2021-2022.xlsx")

used_df = [data_umur_0910, data_umur_1112, data_umur_1314, data_umur_1516, data_umur_1718, data_umur_1920, data_umur_2122]
```

[47] ✓ 0.1s

```
start_year = 2009
dict_umur = {"tahun":[i for i in range(start_year, 2023)], "usia_dibawah_15":[], "usia_produktif":[], "usia_diatas_60":[]}

for df in used_df:
    # print(start_year)
    dict_umur["usia_dibawah_15"].append(data_nasional[data_nasional["tahun"] == str(start_year)]["total_penduduk"].values[0] - int(df.iloc[-1, 1]))
    start_year += 1
    dict_umur["usia_dibawah_15"].append(data_nasional[data_nasional["tahun"] == str(start_year)]["total_penduduk"].values[0] - int(df.iloc[-1, 4]))
    start_year += 1

    dict_umur["usia_produktif"].append(sum(df.iloc[2:-2, 1].astype(int).values))
    dict_umur["usia_produktif"].append(sum(df.iloc[2:-2, 4].astype(int).values))

    dict_umur["usia_diatas_60"].append(int(df.iloc[-2, 1]))
    dict_umur["usia_diatas_60"].append(int(df.iloc[-2, 4]))
```

[48] ✓ 0.1s

```
df_umur = pd.DataFrame(dict_umur)
df_umur["total"] = df_umur["usia_dibawah_15"] + df_umur["usia_produktif"] + df_umur["usia_diatas_60"]
df_umur["dependency_ratio"] = (df_umur["usia_dibawah_15"] + (df_umur["usia_diatas_60"] * 0.6)) / df_umur["usia_produktif"]
```

[49] ✓ 0.0s

```
✓ # dict_umur = {"tahun":[2018, 2019, 2020, 2021, 2022], "usia_dibawah_15":[], "usia_produktif":[], "usia_diatas_60":[]} ···
```

```
df_umur.to_csv("preprocessed_data/data_klasifikasi_umur2.csv", index=False)
```

[16] ✓ 0.0s

Transformasi untuk klasifikasi umur penduduk berdasarkan kategori berikut:

a. kurang dari 15 tahun
b. 15 - 60 tahun
c. lebih dari 60 tahun

Setelah itu, data disimpan dalam file .csv untuk keperluan analisis

TETRIS PROGRAM

3. Data penduduk tiap provinsi

```
penduduk_prov_18_19 = pd.read_excel("raw_data/data_penduduk//Jumlah Penduduk Menurut Provinsi di Indonesia 2018-2019.xlsx").i
penduduk_prov_20_22 = pd.read_excel("raw_data/data_penduduk//Jumlah Penduduk Menurut Provinsi di Indonesia 2020-2022.xlsx").i

old_prov = ["Kep. Bangka Belitung", "DKI Jakarta", "Kep. Riau", "DI Yogyakarta"]
new_prov = ["Bangka Belitung", "Jakarta Raya", "Riau", "Yogyakarta"]

penduduk_prov = penduduk_prov_18_19.merge(penduduk_prov_20_22, on="Provinsi di Indonesia")
```

22] ✓ 0.1s

```
old_prov = ["Kep. Bangka Belitung", "Aceh", "Kep. Riau", "DI Yogyakarta", "Nusa Tenggara Barat"]
new_prov = ["Bangka Belitung", "DI. ACEH", "Riau", "DAERAH ISTIMEWA YOGYAKARTA", "NUSATENGGARA BARAT"]

penduduk_prov["Provinsi di Indonesia"] = penduduk_prov["Provinsi di Indonesia"].replace(old_prov, new_prov)
penduduk_prov["Provinsi di Indonesia"] = penduduk_prov["Provinsi di Indonesia"].apply(lambda x: x.upper())
penduduk_prov.rename(columns={"Provinsi di Indonesia":"Provinsi"}, inplace=True)
```

23] ✓ 0.0s

```
penduduk_prov.to_csv("preprocessed_data\penduduk_per_provinsi_processed.csv", index=False)
```

24] ✓ 0.0s

Transformasi data penduduk tiap provinsi dan pembersihan nama provinsi untuk mengikuti ketentuan nama pada file peta .geojson.

Setelah itu, data disimpan dalam file .csv untuk keperluan analisis

TETRIS PROGRAM

4. Data clustering

```
data_ekonomi = pd.read_excel("raw_data/data_demografi/[Seri 2010] Produk Domestik Regional Bruto Per Kapita.xlsx")
data_humdev = pd.read_excel("raw_data/data_demografi/Indeks Pembangunan Manusia Menurut Provinsi, 2022.xlsx", decimal=',')
data_kemiskinan = pd.read_excel("raw_data/data_demografi/Jumlah dan Persentase Penduduk Miskin Menurut Provinsi, 2022.xlsx", decimal=',').dropna(axis=1)
data_faskes = pd.read_excel("raw_data/data_demografi/Jumlah Rumah Sakit Umum, Rumah Sakit Khusus, Puskesmas, Klinik Pratama, dan Posyandu Menurut Provinsi,
data_fasped = pd.read_excel("raw_data/data_demografi/Kelurahan yang Memiliki Fasilitas Sekolah Menurut Provinsi, 2021.xlsx")
data_luas = pd.read_excel("raw_data/data_demografi\Luas Daerah dan Jumlah Pulau Menurut Provinsi, 2021.xlsx")

data_ekonomi = data_ekonomi[["Provinsi", "hb_2022", "hk_2022"]]
data_faskes = data_faskes.iloc[:, :-2].replace("...", 0)
data_fasped.columns = ["Provinsi", "kelurahan_jumlah_sd", "kelurahan_jumlah_smp",
                       "kelurahan_jumlah_sma", "kelurahan_jumlah_smk", "kelurahan_jumlah_pt"]
```

[43] ✓ 0.0s

```
old_prov_upper = ["ACEH", "KEP. BANGKA BELITUNG", "NUSA TENGGARA BARAT", "DI YOGYAKARTA", "KEP. RIAU"]
old_prov_cap = ["Aceh", "Kepulauan Bangka Belitung", "Nusa Tenggara Barat", "DI Yogyakarta"]

new_prov_geo = ["DI. ACEH", "BANGKA BELITUNG", "NUSATENGGARA BARAT", "DAERAH ISTIMEWA YOGYAKARTA", "KEPULAUAN RIAU"]

data_ekonomi["Provinsi"] = data_ekonomi["Provinsi"].replace(old_prov_upper, new_prov_geo)
data_humdev["Provinsi"] = data_humdev["Provinsi"].replace(old_prov_cap, new_prov_geo[:-1]).apply(lambda x: x.upper())
data_kemiskinan["Provinsi"] = data_kemiskinan["Provinsi"].replace(old_prov_cap, new_prov_geo[:-1]).apply(lambda x: x.upper())
data_faskes["Provinsi"] = data_faskes["Provinsi"].replace(old_prov_cap, new_prov_geo[:-1]).apply(lambda x: x.upper())
data_fasped["Provinsi"] = data_fasped["Provinsi"].replace(old_prov_cap, new_prov_geo[:-1]).apply(lambda x: x.upper())
data_luas["Provinsi"] = data_luas["Provinsi"].replace(old_prov_cap, new_prov_geo[:-1]).apply(lambda x: x.upper())
```

[44] ✓ 0.0s

```
cluster_df = data_ekonomi.copy()

for df in [data_humdev, data_faskes, data_fasped,
           data_kemiskinan[["Provinsi", "Persentase Penduduk Miskin - Maret"]],
           data_luas[["Provinsi", "Luas Wilayah (km2)"]],
           penduduk_prov[["Provinsi", "2022"]]]:
    cluster_df = cluster_df.merge(df, on="Provinsi")

for col in cluster_df.columns:
    if col != "Provinsi" and cluster_df[col].dtypes == cluster_df["Provinsi"].dtypes:
        cluster_df[col] = cluster_df[col].apply(lambda x: str(x).replace(' ','').replace(',','.'))
    if col != "Provinsi":
        cluster_df[col] = cluster_df[col].astype(float)

cluster_df.to_csv("preprocessed_data/cluster_data.csv", index=False)
```

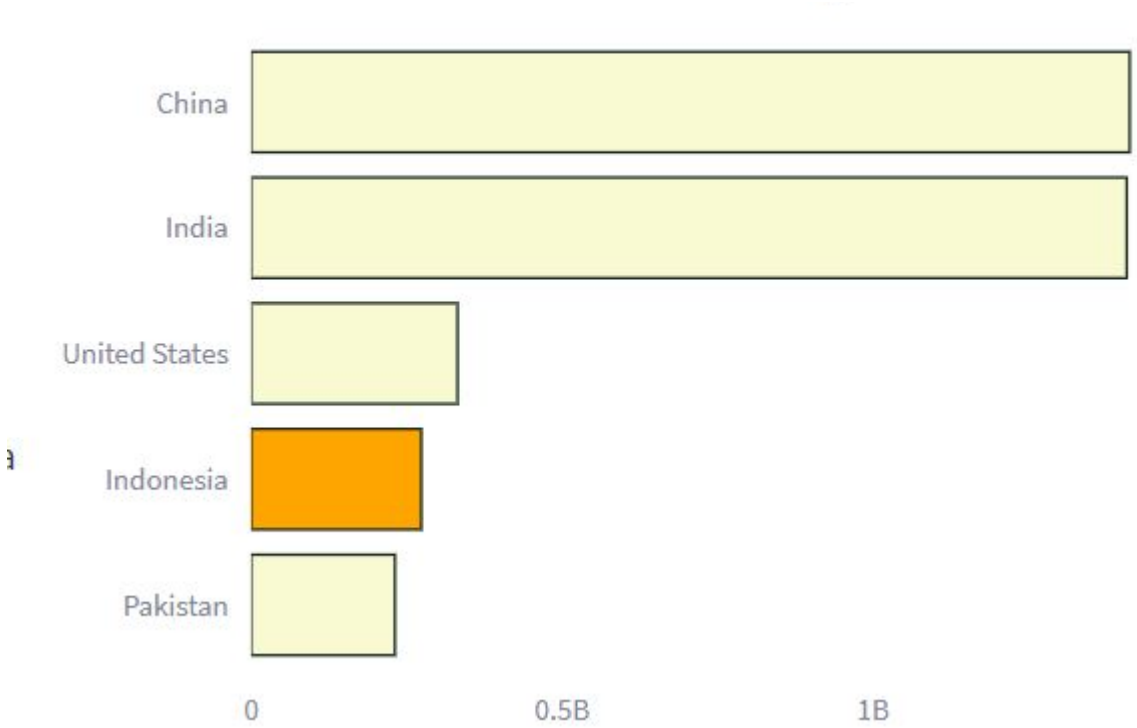
[45] ✓ 0.0s

Penggabungan beberapa data faktor sosial ekonomi dan pembangunan serta diberlakukannya pembersihan nama provinsi untuk mengikuti ketentuan nama pada file peta .geojson.

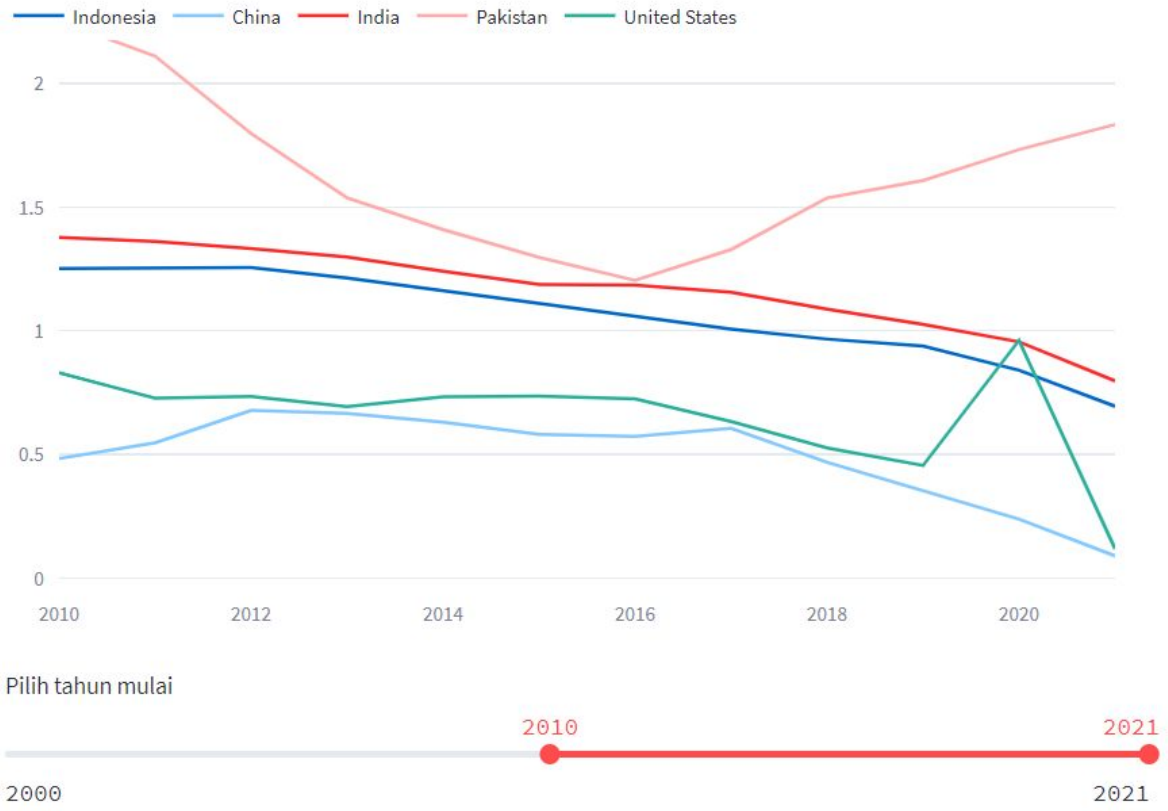
Setelah itu, data disimpan dalam file .csv untuk keperluan analisis

Step 3 - Data Exploration & Data Visualisation

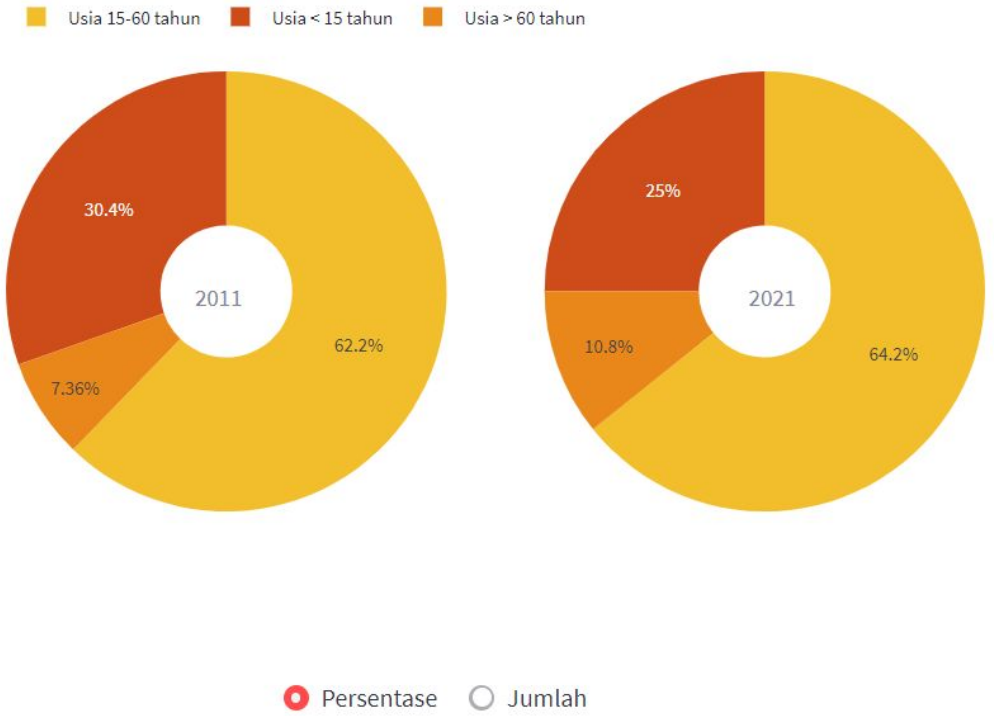
- Peringkat Indonesia dalam konteks total populasi di seluruh dunia pada tahun 2021



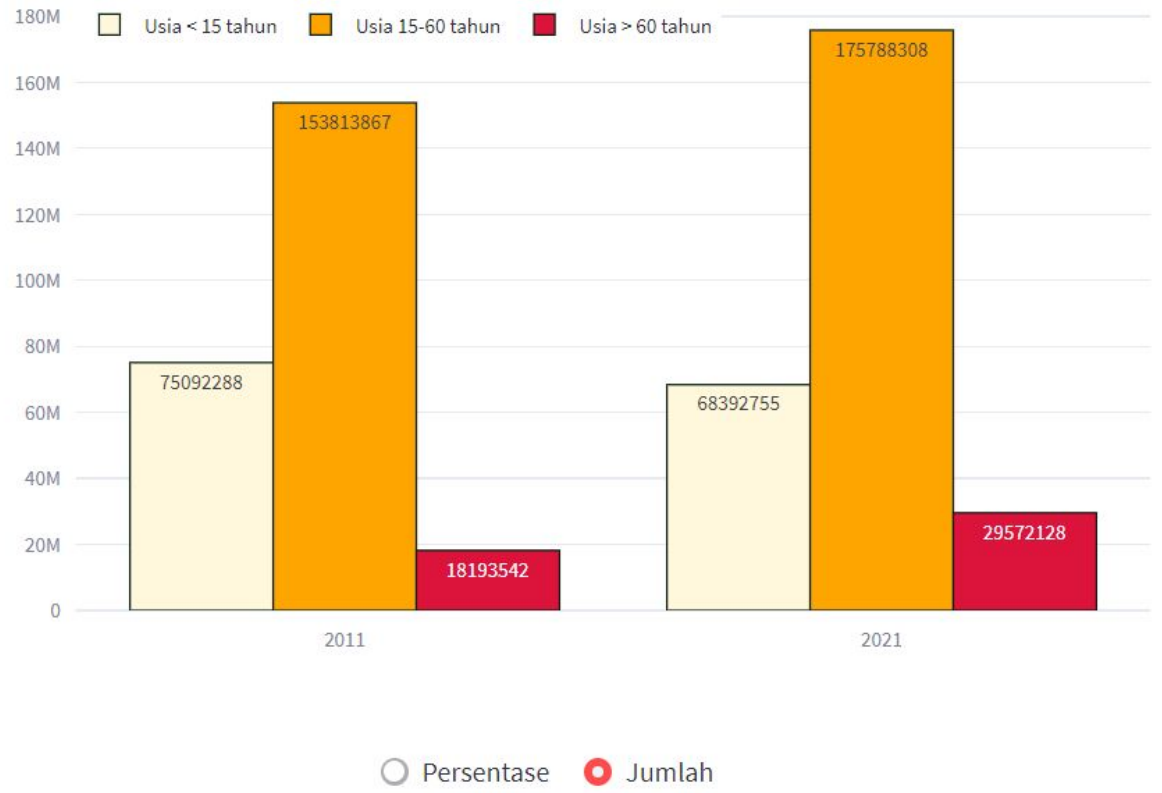
- Membandingkan persentase pertumbuhan penduduk di Indonesia dengan 4 negara dengan jumlah penduduk terbanyak lainnya



- Membandingkan komposisi kelompok usia di Indonesia tahun 2011 dan 2021 dalam bentuk persentase



- Membandingkan komposisi kelompok usia di Indonesia tahun 2011 dan 2021 dalam bentuk kuantitas

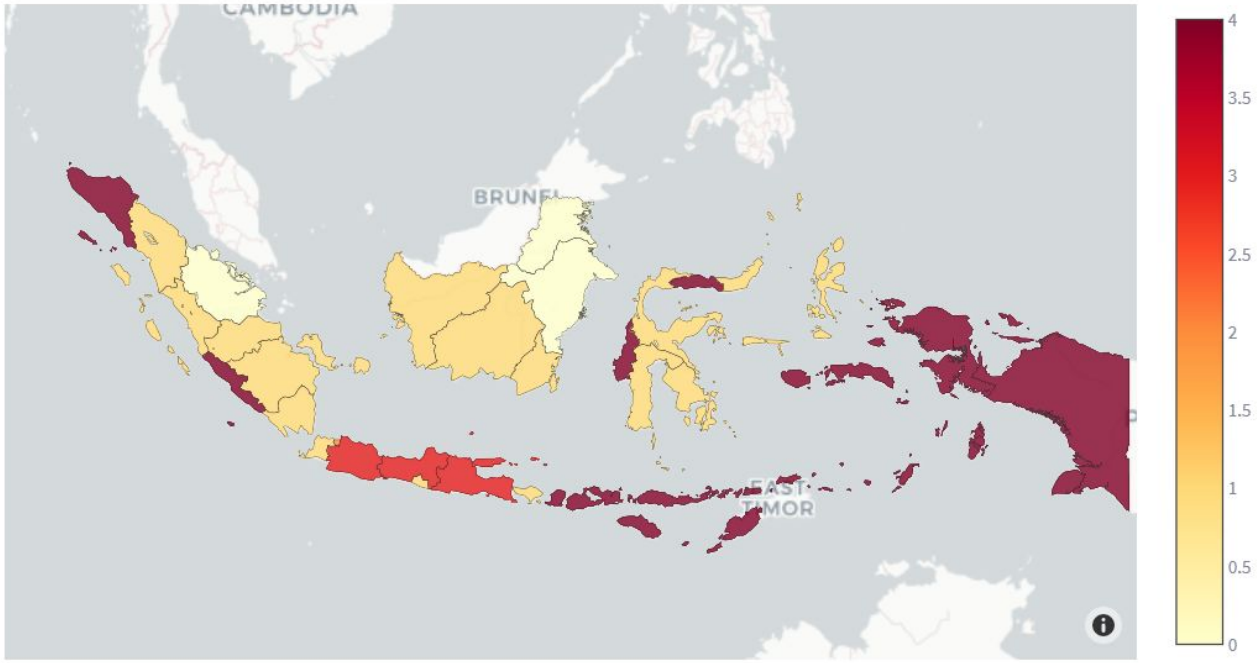


- Melihat persebaran penduduk pada setiap provinsi di Indonesia

- Melakukan *clustering* berdasarkan variabel faktor sosial ekonomi dan pembangunan pada setiap provinsi di Indonesia

Plot Map

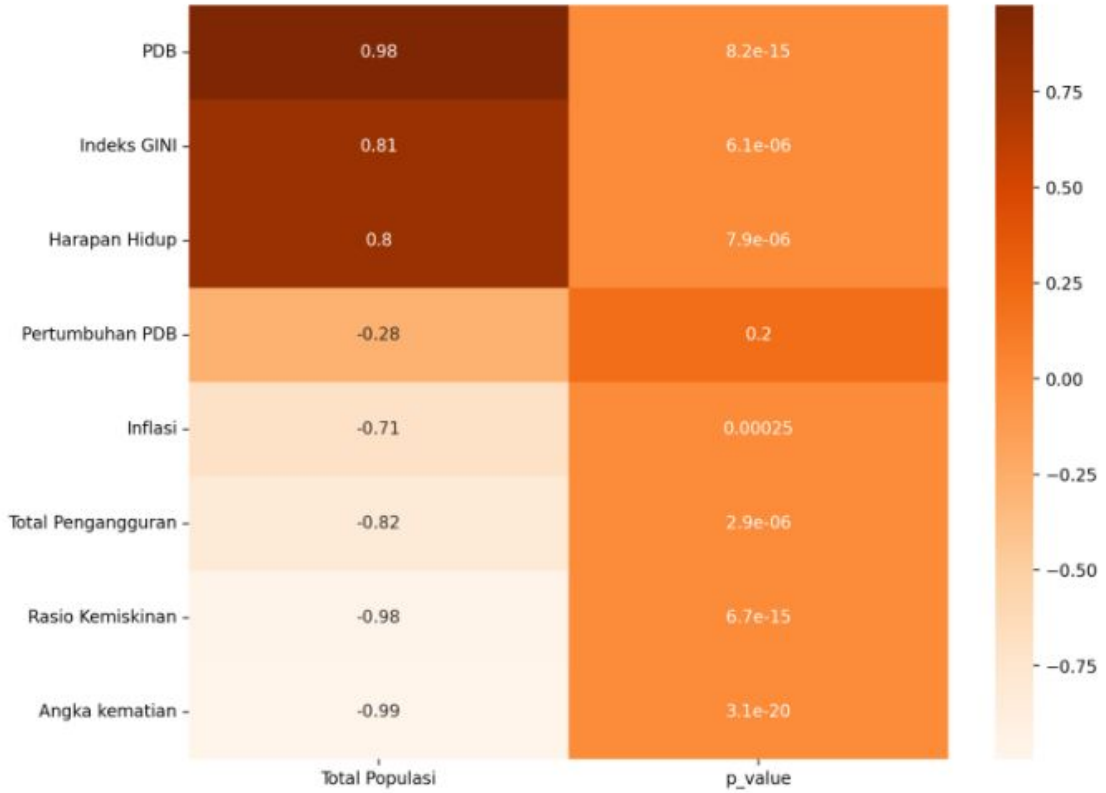
Clustering



Deskripsi Cluster

- **Cluster 0:** pertumbuhan ekonomi [**TINGGI-182K**], IPM [SEDANG-74], rasio kemiskinan [SEDANG-6.6%], fasilitas kesehatan dan pendidikan [SEDANG-2836]
- **Cluster 1:** pertumbuhan ekonomi [SEDANG-61K], IPM [SEDANG-72], rasio kemiskinan [SEDANG-7.8%], fasilitas kesehatan dan pendidikan [SEDANG-3833]
- **Cluster 2:** pertumbuhan ekonomi [**TINGGI SEKALI-293K**], IPM [**TINGGI-82**], rasio kemiskinan [RENDAH-4.6%], fasilitas kesehatan dan pendidikan [SEDANG-1597]
- **Cluster 3:** pertumbuhan ekonomi [SEDANG-52K], IPM [SEDANG-72], rasio kemiskinan [SEDANG-9.7%], fasilitas kesehatan dan pendidikan [**TINGGI-17124**]
- **Cluster 4:** pertumbuhan ekonomi [**RENDAH-41K**], IPM [RENDAH-68], rasio kemiskinan [**TINGGI-17.1%**], fasilitas kesehatan dan pendidikan [SEDANG-3095]

- Melihat pearson korelasi dan p-value dari variabel-variabel lainnya dengan jumlah total populasi di Indonesia



Step 4 - Insight Analysis

1. Persentase pertumbuhan penduduk di Indonesia selalu mengalami penurunan dan sejak tahun 2018 sudah dibawah angka 1%

2. Jumlah penduduk dengan usia lebih dari 60 tahun meningkat 11 ribu dari tahun 2011 ke tahun 2021

3. 55% dari total penduduk di Indonesia menempati di Pulau Jawa

4. Masih terdapat ketimpangan ekonomi dan faktor sosial lainnya antara beberapa provinsi di Indonesia

5. Terdapat korelasi total penduduk di Indonesia dengan beberapa aspek seperti, Produk Domestik Bruto (PDB), indeks GINI, rasio kemiskinan, dan total pengangguran. Namun, korelasi ini tidak bisa diambil secara mentah-mentah karena masih *general* dan hanya dilihat dari 1 sudut pandang saja, yaitu total penduduk.

DQLab

AYO #STACKYOURSKILL SEKARANG

dan Persiapkan Diri Menjadi Praktisi Data!