PROJEK AKHIR DATA WRANGLING 2025/2026

ANALISIS PENGARUH FAKTOR IKLIM DAN KEPADATAN PENDUDUK TERHADAP KASUS DEMAM BERDARAH DENGUE (DBD) MENURUT PROVINSI DI INDONESIA TAHUN 2019-2020

Dosen Pengampu:

Ulfa Siti Nuraini, S.Stat., M.Stat.

Disusun oleh:

Kelompok 9 – 2024 C

Siti Fadilah Nurkhotimah (1314623019 – UNJ / 251155009 – UNESA)

Laili Nurrohmatul Fadhila Zulfa (24031554093 – UNESA)

Kelas: Sains Data 2024 C

UNIVERSITAS NEGERI SURABAYA – UNIVERSITAS NEGERI JAKARTA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

PROGRAM STUDI SAINS DATA – PROGRAM STUDI STATISTIKA

2025

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Deman Berdarah Dengue (DBD) adalah salah satu penyakit menular yang masih menjadi masalah di Indonesia. Menurut Kementerian kesehatan RI Tahun 2020, kasus DBD pada tahun 2019 tercatat lebih dari 138.000 kasus dan tersebar hampir di seluruh provinsi. Sedangkan, pada 2020 jumlah kasus turun menjadi 95.994 kasus. Kondisi ini menunjukkan DBD bukan hanya masalah ringan tetapi masalah endemis yang dipengaruhi faktor lingkungan dan demografis.

Faktor lain seperti kepadatan penduduk juga menjadi risiko penting dalam penyebaran DBD. Semakin tinggi kepadatan penduduk akan lebih banyak terjadi interaksi manusia yang menyebabkan risiko penularan lebih tinggi. Data BPS menunjukkan bahwa kepadatan penduduk nasional berada di kisaran 141 jiwa/km ^{2} berdasarkan Sensus penduduk 2020.

Faktor iklim seperti suhu, curah hujan, dan kelembapan memengaruhi kecepatan berkembang biak nyamuk Aedes dan menyebabkan juga terjadi kecepatan penyebaran virus dengue. Menurut BMKG Tahun 2021, peningkatan curah hujan dan kelembapan dapat meningkatkan potensi genangan air di lingkungan. Hasil penelitian oleh Keman (2022) menunjukkanbahwa puncak kasus DBD sering terjadi setelah periode penigkatan curah hujan dan suhu yang relatif tinggi.

Berdasarkan beberapa faktor tersebut, kami memutuskan untuk menganalisis mengenai pengaruh faktor iklim dan kepadatan penduduk terhadap kasus DBD menurut provinsi tahun 2019-2020. Proses wrangling data menjadi penting untuk mengintegrasikan berbagai sumber data sebelum dilakukan analisis lebih lanjut. Oleh karena itu, penelitian ini berfokus pada proses pengambilan, pembersihan, integrasi, eksplorasi, dan publikasi data terkait.

1.2 Rumusan Masalah

1. Bagaimana proses pengambilan dan integrasi data iklim, kepadatan penduduk, dan kasus DBD dari tiga sumber berbeda?
2. Bagaimana proses pembersihan (cleaning) yang diperlukan agar ketiga dataset dapat dianalisis bersama?

3. Bagaimana gambaran eksplorasi data terhadap faktor iklim, kepadatan penduduk, dan kasus DBD per provinsi?
4. Bagaimana hasil akhir data yang telah melalui proses wrangling dan siap untuk analisis lanjutan?

1.3 Tujuan Penelitian

1. Melakukan proses *data wrangling* pada tiga dataset berbeda.
2. Melakukan pembersihan, standarisasi, dan integrasi ketiga dataset menjadi satu dataset analisis.
3. Menyajikan hasil eksplorasi awal terkait hubungan faktor iklim, kepadatan penduduk, dan kasus DBD.
4. Menghasilkan dataset bersih yang siap digunakan untuk analisis statistika lanjutan.

1.4 Manfaat Penelitian

Secara akademis:

- Membantu memahami alur lengkap proses wrangling data dari berbagai sumber.
- Menjadi contoh implementasi nyata integrasi data kesehatan, iklim, dan sosial.

Secara praktis:

- Menyediakan dataset terintegrasi yang dapat dimanfaatkan untuk analisis epidemiologi dan public health.
- Memberikan dasar untuk rekomendasi pengendalian DBD berdasarkan pola iklim dan kepadatan penduduk.

BAB II
TINJAUAN PUSTAKA

2.1 Demam Berdarah Dengue (DBD)

Demam Berdarah Dengue (DBD) merupakan penyakit infeksi yang disebabkan oleh virus dengue dan ditularkan melalui gigitan nyamuk *Aedes aegypti* dan *Aedes albopictus*. Menurut World Health Organization, Dengue merupakan infeksi virus yang menyebar dari nyamuk ke manusia. DBD menjadi salah satu masalah kesehatan masyarakat di berbagai negara tropis dan subtropis, termasuk Indonesia. Penyakit DBD cenderung meningkat pada musim hujan dan wilayah lingkungan padat penduduk.

Menurut Kementerian Kesehatan, kasus DBD di Indonesia cenderung berfluktuasi setiap tahun. Tingginya angka kasus DBD dipengaruhi oleh berbagai faktor, antara lain karakteristik lingkungan, kondisi iklim, dan faktor sosial-demografis yang berdampak pada dinamika populasi nyamuk vektor. Kondisi ini menjadikan DBD sebagai penyakit yang erat kaitannya dengan faktor iklim dan kepadatan penduduk.

2.2 Pengaruh Faktor Iklim terhadap DBD

Faktor iklim berperan penting dalam menentukan perkembangan vektor dan penyebaran virus dengue. Variabel-variabel iklim, seperti suhu, kelembaban, dan curah hujan memiliki keterkaitan yang kuat dengan siklus hidup nyamuk. Suhu udara mempengaruhi kecepatan perkembangan larva, tingkat gigitan nyamuk, serta masa inkubasi virus dalam tubuh vektor. Kelembaban udara yang tinggi dapat memperpanjang umur nyamuk dewasa, sehingga meningkatkan peluang penyebaran virus. Curah hujan yang tinggi dapat menciptakan lebih banyak tempat perindukan bagi nyamuk, seperti genangan air di lingkungan pemukiman. Namun, curah hujan yang ekstrem justru dapat mengurangi populasi vektor karena larva bisa hanyut.

2.3 Kepadatan Penduduk dan Kaitannya dengan Penyebaran DBD

Kepadatan penduduk merupakan faktor demografis yang sering dikaitkan dengan tingginya penyebaran dan penularan DBD. Lingkungan dengan jumlah penduduk yang padat, terutama kawasan perkotaan, cenderung memiliki kondisi yang mendukung berkembangnya nyamuk Aedes, seperti rumah-rumah yang berdekatan, wadah air yang tidak tertutup, atau ventilasi kurang optimal. Kepadatan penduduk yang tinggi membuat jarak terbang nyamuk yang relatif pendek (sekitar 50–100 meter) menjadi lebih efektif dalam menjangkau banyak manusia dalam area terbatas. Wilayah padat penduduk sering

mengalami tantangan dalam pengelolaan lingkungan, seperti penumpukan sampah dan sanitasi yang buruk dapat menjadi tempat berkembang biaknya vektor (nyamuk pembawa penyakit). Hal tersebut meningkatkan probabilitas kontak antara manusia dan nyamuk yang terinfeksi (membawa virus dengue) dan menyebarkan virus DBD.

2.4 Konsep Data Wrangling

Data wrangling, atau dikenal juga sebagai data munging, adalah proses mengubah, membersihkan, dan menyusun data mentah menjadi format yang lebih terstruktur sehingga siap digunakan untuk analisis lebih lanjut. Dalam penelitian data kuantitatif, terutama yang melibatkan integrasi berbagai sumber data seperti iklim, kepadatan penduduk, dan kasus penyakit, proses data wrangling menjadi tahap yang sangat penting. Hal ini disebabkan oleh karakteristik data mentah yang umumnya masih berantakan, tidak konsisten antar sumber, memiliki *missing value*, dan memerlukan penyesuaian struktur agar dapat digabungkan.

Data wrangling meliputi beberapa langkah sistematis. Tahapan utamanya adalah pengambilan data, yaitu proses memperoleh data dari berbagai sumber seperti Kaggle untuk data iklim, Badan Pusat Statistik (BPS) untuk kepadatan penduduk, dan Kementerian Kesehatan untuk data kasus DBD. Tahap ini mencakup identifikasi format file, validitas sumber, dan pemahaman struktur data awal.

Tahapan kedua adalah data cleaning, yaitu proses menghapus duplikasi, memperbaiki format kolom, menangani missing values, menyamakan nama provinsi atau kategori, serta memastikan tidak ada anomali yang mengganggu integrasi data. Tahapan ketiga, transformasi dan integrasi data, yakni menggabungkan beberapa dataset yang berbeda menjadi satu dataset analisis, termasuk proses normalisasi, penggabungan berdasarkan *key* tertentu (misalnya nama provinsi), dan penyesuaian periode waktu.

Tahapan selanjutnya adalah data eksplorasi (*exploratory data analysis*/EDA), yang dilakukan untuk memahami pola awal, menampilkan visualisasi dari analisis antar variabel, serta untuk mendapatkan korelasi antar variabel melalui eksplorasi data. Tahap terakhir adalah publikasi data (data publishing). Pada tahap ini, peneliti menghasilkan tiga bentuk keluaran, yaitu raw data yang telah disimpan sebagaimana diperoleh dari sumber awal, data hasil proses wrangling yang siap dianalisis, dan dokumentasi pipeline wrangling yang menjelaskan langkah-langkah teknis secara jelas.

Dengan demikian, data wrangling menjadi fondasi utama dalam penelitian berbasis data. Tanpa proses wrangling yang baik, data dapat menghasilkan kesimpulan yang bias atau salah interpretasi.

BAB III
METODOLOGI PENELITIAN

3.1 Sumber Data

Laporan ini dibuat menggunakan tidak dataset utama yang berasal dari sumber resmi untuk mendukung analisis pengaruh faktor iklim dan kepadatan penduduk terhadap kasus Demam Berdarah Dengue (DBD) di Indonesia tahun 2019 dan 2020.

Data mengenai kasus DBD per provinsi diperoleh dari laman resmi Kementerian Kesehatan Republik Indonesia melalui kategori Profil Kesehatan. Dokumen tersebut menyediakan laporan tahunan mengenai jumlah kasus dan distribusi DBD berdasarkan provinsi di seluruh Indonesia. Data tahun 2019–2020 digunakan untuk menggambarkan tren kasus DBD menjelang dan pada awal pandemi COVID-19, sehingga dapat dianalisis apakah terdapat pengaruh yang signifikan dari kondisi iklim dan kepadatan penduduk terhadap jumlah kasus DBD.

Komponen iklim yang berhubungan kejadian DBD adalah suhu, kelembapan udara, dan curah hujan Data iklim yang digunakan dalam analisis bersumber dari Kaggle, yaitu dataset Indonesia Climate, berisi informasi iklim per provinsi Indonesia, yang terdiri dari tigas file csv, yaitu *climate_data, station_detail.csv*, dan *province_detail.csv*. Data ini digunakan untuk mengidentifikasi pola iklim pada tahun 2019–2020 dan melihat keterkaitannya dengan kasus DBD.

Data kepadatan penduduk yang digunakan dalam laporan ini diambil dari Badan Pusat Statistik (BPS) melalui tabel Population Density by Province berisi kepadatan penduduk (jiwa/km^2) seluruh provinsi di Indonesia, yang memuat informasi jumlah penduduk per kilometer persegi. Data ini digunakan untuk menilai sejauh mana kepadatan penduduk mempengaruhi jumlah kasus DBD antar provinsi di Indonesia.

3.2 Teknik Pengambilan Data

Pengambilan data untuk analisis ini menggunakan teknik pengumpulan data sekunder, yang diperoleh dari sumber-sumber yang telah dipublikasikan oleh pihak ketiga. Proses ini melibatkan tiga sumber data utama.

3.2.1 Dataset Kasus Demam Berdarah Dengue (DBD)

Dataset Kasus Demam Berdarah Dengue (DBD), data diambil dari dokumen Profil Kesehatan Indonesia tahun 2019 dan tahun 2020 yang diterbitkan oleh Kementerian Kesehatan (Kemenkes) sebagai Laporan Tahunan. Data kasus DBD per

provinsi berbentuk dokumen dalam format PDF, tahap pengumpulan ini dilanjutkan dengan proses ekstraksi atau scraping data menggunakan Python di Google Colab untuk mengubah tabel dari format PDF menjadi data terstruktur (CSV) agar memudahkan analisis selanjutnya.

3.2.2 Dataset Iklim Indonesia

Untuk Dataset Iklim Indonesia, data diunduh langsung dalam bentuk CSV dari platform Kaggle. Dataset ini berisi data iklim (kemungkinan curah hujan, suhu, kelembaban, dll.) yang sudah dikumpulkan. Dataset pada Kaggle ini terdiri dari 3 *file* berformat CSV yang berbeda, yaitu *climate_data, station_detail.csv*, dan *province_detail.csv*. Tahap pengambilan ini harus diikuti oleh proses integrasi data untuk menggabungkan informasi iklim menjadi satu dataset yang komprehensif menggunakan kolom kunci yang sama (misalnya, berdasarkan nama provinsi).

3.2.3 Dataset Kepadatan Penduduk menurut Provinsi (jiwa/km^2)

Untuk Dataset Kepadatan Penduduk, data diakses dari tabel statistik di situs resmi Badan Pusat Statistik (BPS). Pengambilan data dari BPS dilakukan melalui mengunduh (*download*) dengan format CSV. Secara keseluruhan, teknik ini berfokus pada pengambilan data kepadatan penduduk (jiwa/km^2) seluruh provinsi di Indonesia yang sudah ada untuk tahun 2019 dan tahun 2020.

3.3 Teknik Integrasi Data

Teknik integrasi data digunakan untuk menggabungkan beberapa dataset yang berbeda menjadi satu dataset yang bisa digunakan untuk analisis data. Dalam penelitian ini, proses integrasi data dilakukan untuk menggabungkan data iklim, data kepadatan penduduk, dan data kasus DBD untuk rentan waktu 2019-2020. Berikut integrasintegrasi data yang diigunakan:

3.3.1 Penyamaan Nama Provinsi

Dalam dataset province_detail terdapat provinsi yang berbeda dengan dataset lain yaitu “Nanggro Aceh darussalam” dirubah menjadi “Aceh”. Format string menjadi title, mengganti “Dki Jakarta” dan “Di Yogyakarta” menjadi “DKI Jakarta” dan “ DI yogyakarta” untuk semua dataset.

3.3.2 Standarisasi Nama Kolom

Perlu dilakukan standarisasi nama kolom agar bisa dilakukan penggabungan antar dataset. Pada dataset province_detail kolom “province_name” diganti dengan “provinsi”. Dalam dataset Kepadatan Penduduk – 2019 kolom “Unnamed” diganti dengan “kepadatan 2019” dan kolom “38 Provinsi” diganti dengan “provinsi” begitu

pula dilakukan pada dataset Kepadatan Penduduk – 2020. Standarisasi nama kolom untuk dataset DBD pada kolom “Provinsi” , “Jumlah Penduduk”, “Jumlah Kasus”, Incidence Rate per\n100.000 Penduduk”, “meninggaal”, “CFR (%)” menjadi “provinsi”, “jumlah penduduk 2019”, “jumlah kasus 2019”, “incidence rate per 100.000 penduduk 2019”, “meninggal 2019”, “CFR (%) 2019”, begitupun untuk dataset 2020.

3.3.3 Penggabungan Dataset

Pada penggabungan dataset ini dilakukan untuk beberapa kali yaitu:

1. Penggabungan Dataset Iklim

   Pertama, pada dataset iklim climate_data.csv dan dataset stasiun station_detail.csv menggunakan merge untuk menggabungkan 2 dataset dan menambah 2 kolom yaitu station_id dan province_id. Kemudian melakukan merge lagi terhadap dataset province_detail untuk menggabungkan nama provinsi sesuai provinsi_id lalu di simpan dalam csv.

2. Penggabungan Dataset Kepadatan Penduduk

   Penggabungan selanjutnya yaitu pada dataset kepadatan penduduk tahun 2019 dan 2020 yang kemudian disimpan dalam csv. Penggabungan ini menggunakan merge nama privinsi sehingga menghasilkan 3 kolom yaitu provinsi, kepadatan penduduk 2019, dan kepadatan penduduk 2020.

3. Penggabungan Dataset DBD

   Sama seperti kepadatan penduduk, penggabungan ini menggunakan merge dataset dbd 2019 dan 2020 terhadap provinsi.

4. Penggabungan Seluruh Dataset

   Terakhir menggabungkan ketiga dataset yaitu dataset iklim 2019-2020, kepadatan penduduk 2019-2020, dan dataset DBD 2019-2020.

3.4 *Data Cleaning*

Cleaning data dilakukan pada beberapa dataset:

1. Cleaning Dataset Iklim

   Cleaning pada dataset iklim berupa perubahan NaN menjadi angka 0 dan standarisasi nama provinsi dari “province_name” menjadi “provinsi”.

2. Cleaning Dataset Kepadatan Penduduk

   Cleaning dataset kepadatan penduduk berupa standarisasi nama kolom “unnamed: 1” menjadi “kepadatan 2019” dan nama kolom “38 Provinsi” menjadi “ provinsi”. Cleaning yang kedua yaitu menghilangkan baris NaN yang berada pada 2 baris setelah nama kolom. Cleaning yang ketiga yaitu merubah tipe kolom

“kepaadatan 2019” menjadi numerik. Cleaning yang terakhir yaitu menghapus baris yang tidak berisi atau tidak ada nilai kepadatannnya. Cleaning pertama sampai terakhir juga dilakukan untuk dataset kepadatan peduduk 2020.

3. Cleaning Dataset DBD

Pada cleaning dataset DBD pertama yang dilakukan adalah penghapusan baris index 0 dan 35 karena bukan data yang akan digunakan. Kemudian dilakukan cleaning berupa merubah tipe data dari object menjadi sesuai format bertipe numerik untuk kolom tertentu. Menghapus titik pemisah ribuan dan koma ke titik, membersihkan kolom integer, Membersihkan kolom float.

3.5 Eksplorasi Data

Eksplorasi Data (Exploratory Data Analysis / EDA) adalah tahap metodologi yang dilakukan setelah data selesai dikumpulkan dan dibersihkan (data wrangling), tetapi sebelum analisis statistik formal (inferensial) atau pengujian hipotesis dilakukan. Tujuan eksplorasi data adalah untuk memahami struktur, pola, dan karakteristik dasar dataset sebelum dilakukan pengolahan lebih lanjut. Pada tahap eksplorasi data (*data exploration*) ini dilakukan proses pemahaman awal terhadap karakteristik data kasus Demam Berdarah Dengue (DBD), indikator iklim, serta kepadatan penduduk pada setiap provinsi di Indonesia untuk periode tahun 2019–2020.

Eksplorasi statistik deskriptif dilakukan terhadap variabel-variabel penelitian, meliputi nilai minimum, maksimum, rata-rata, maupun standar deviasi untuk masing-masing variabel per provinsi. Eksplorasi visual dilakukan menggunakan grafik atau plot, seperti heatmap korelasi, boxplot, bar chart, ataupun scatterplot untuk melihat pola hubungan awal antara faktor iklim dan jumlah kasus DBD. Visualisasi ini membantu mengidentifikasi kemungkinan variabel yang memiliki hubungan dengan peningkatan jumlah kasus DBD dan incidence rate per 100.000 penduduk.

3.5.1 Analisis Univariat

Analisis univariat dilakukan untuk mendapatkan gambaran karakteristik variabel variabel penelitian. Dikarenakan data pada penelitian ini merupakan data numerik maka analisis yang dilakukan meliputi rata-rata, nilai median, nilai maksimum dan minimum, serta standar deviasi. Penyajian hasil analisis univariat menggunakan tabel dan grafik.

3.5.2 Analisis Bivariat

Uji bivariat dilakukan sebagai pendekatan analitik untuk mengevaluasi korelasi antara variabel independen dan variabel dependen. Sebelum pengujian tersebut

dilaksanakan, perlu dilakukan uji terhadap karakteristik sebaran data melalui uji normalitas untuk menentukan pendekatan statistik saat analisis korelasi. Adanya korelasi yang signifikan secara statistik antara dua variabel ditandai dengan nilai p yang berada di bawah atau setara 0,05. Kekuatan dan arah hubungan ditandai dengan r. Hubungan lemah atau bahkan tidak ada hubungan ditandai nilai r berada di rentang 0,00-0,25. Hubungan sedang ditandai nilai r pada angka 0,26-0,50. Sementara itu, korelasi yang kuat memiliki nilai r 0,51-0,75. Nilai r yang berada pada rentang 0,76-1,00 menandakan adanya korelasi yang sangat kuat. Semakin mendekati 1, hubungan semakin kuat. Nilai r yang positif menunjukan bahwa hubungan memiliki arah positif, sebaliknya nilai r negatif menunjukan hubungan dengan arah negatif. Hubungan dengan arah positif menandakan nilai suatu variabel berbanding lurus dengan nilai variabel lain. Hubungan negatif menandakan hubungan yang berlawanan, semakin besar nilai suatu variabel semakin kecil nilai variabel yang lain.

3.5.3 Uji Normalitas

Sebelum menganalisis data numerik, harus melakukan uji normalitas untuk mengetahui bagaimana distribusi normalitas data tersebut. Hasil uji normalitas digunakan untuk menentukan jenis uji yang akan dipakai dalam analisis bivariat. Jika hasil uji menunjukan p value >0,05 maka data tersebut memiliki distribusi normal sehingga uji bivariat yang akan digunakan adalah uji korelasi Pearson. Jika hasil uji normalitas menunjukan p value < 0,05, maka data tersebut terdistribusi tidak normal sehingga uji bivariat yang digunakan adalah korelasi Spearman.

3.6 *Data Publishing*

Pada penellitian ini, data publishing mencakup penyajian tiga komponen utama yaitu: raw data, processed data, dan dokumentasi pipeline wrangling. Publishing untuk penelitian ini dikumpulkan dan didokumentasikan pada GitHub melalui *link* berikut. *SitiFadilahNurkhotimah/9_Laili-Nurrohmatul-Fadhila-Zulfa_Siti-Fadilah-Nurkhotimah_2024C_Projek_Akhir_Data_Wrangling: Analisis Pengaruh Faktor Iklim dan Kepadatan Penduduk terhadap Kasus Demam Berdarah Dengue (DBD) Menurut Provinsi di Indonesia Tahun 2019–2020*

1. Raw Data

   Raw data merupakan 3 dataset awal yang diperoleh dari sumber yang berbeda yaitu dataset iklim, dataset kepadatan penduduk tahun 2019 dan 2020, dan dataset

kasus DBD yang diperoleh dari pdf kementerian kesehatan. Ketiga dataset disimpan tanpa modifikasi sehingga dapat dijadikan acuan ketika dilakukan replikasi penelitian.

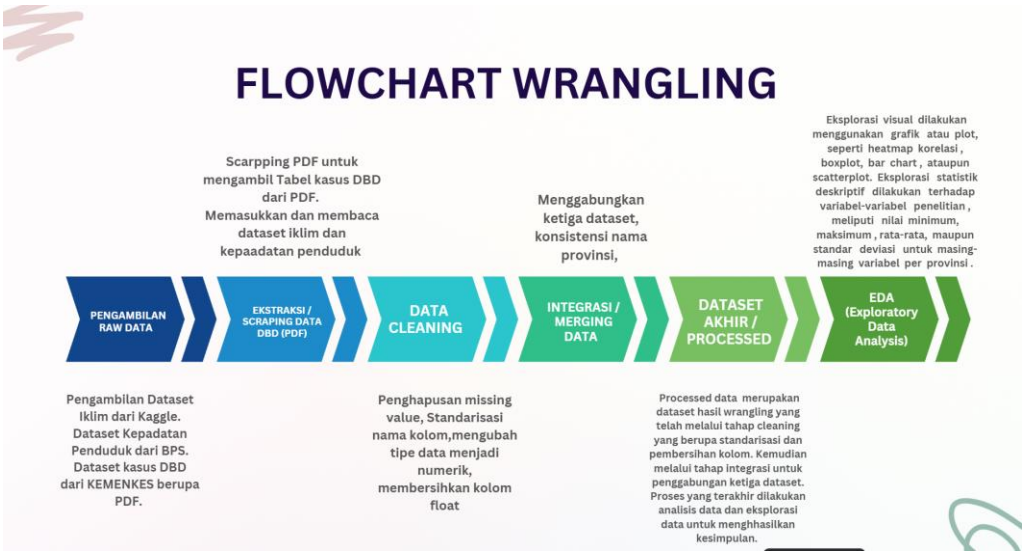
2. *Processed Data*

Processed data merupakan dataset hasil wrangling yang telah melalui tahap cleaning yang berupa standarisasi dan pembersihan kolom. Kemudian melalui tahap integrasi untuk penggabungan ketiga dataset. Proses yang terakhir dilakukan analisis data dan eksplorasi data untuk menghhasilkan kesimpulan.

3. Dokumentasi Pipeline Wrangling

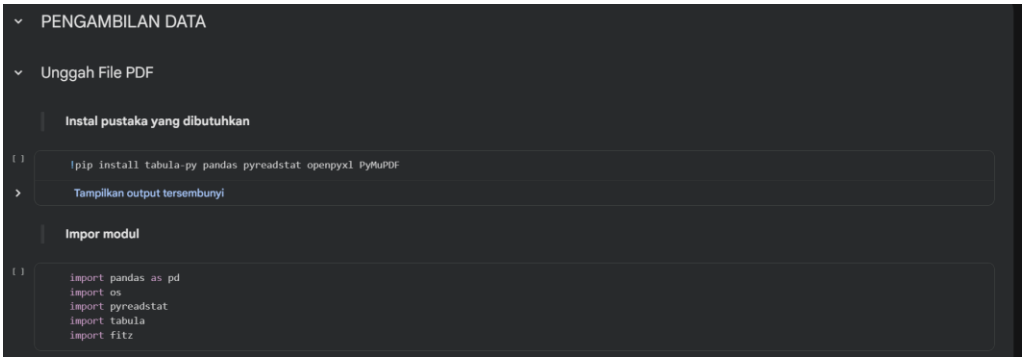
Pipeline wrangling adalah rangkaian langkah teknis yang dilakukan untuk mengubah *raw data* menjadi dataset bersih. Dokumentasi pipeline pada penelitian ini mencakup:

- Flowchart yang menggambarkan alur kerja



- Kode Python

a. Pengambilan Data



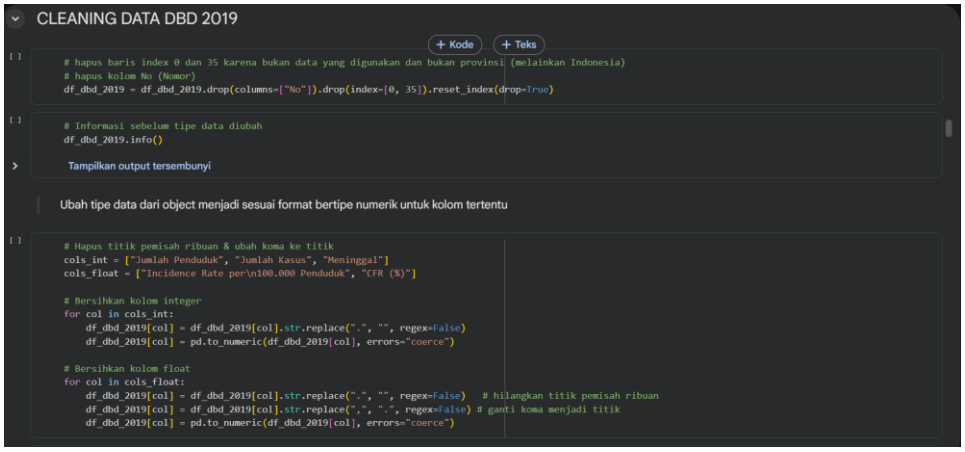
b. Cleaning Data

- Pembersihan Data Iklim (df_final)

- Pembersihan Data Kepadatan Penduduk
  - 2019

- 2020

- Pembersihan Data DBD
  - 2019

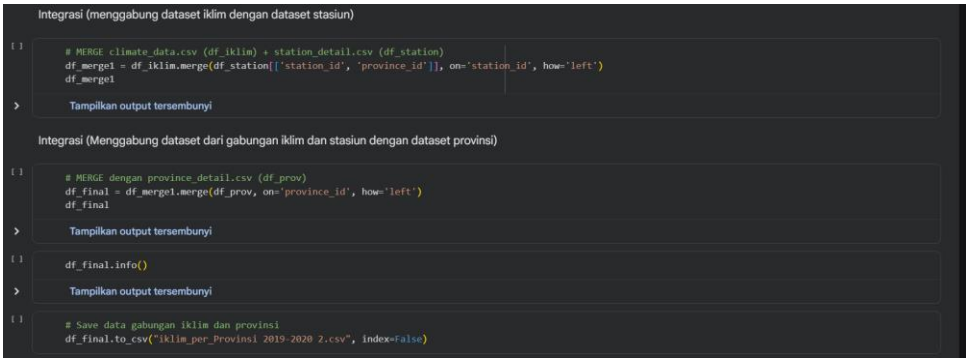


  - 2020

c. Integrasi Data

- Integrasi Dataset Provinsi Iklim

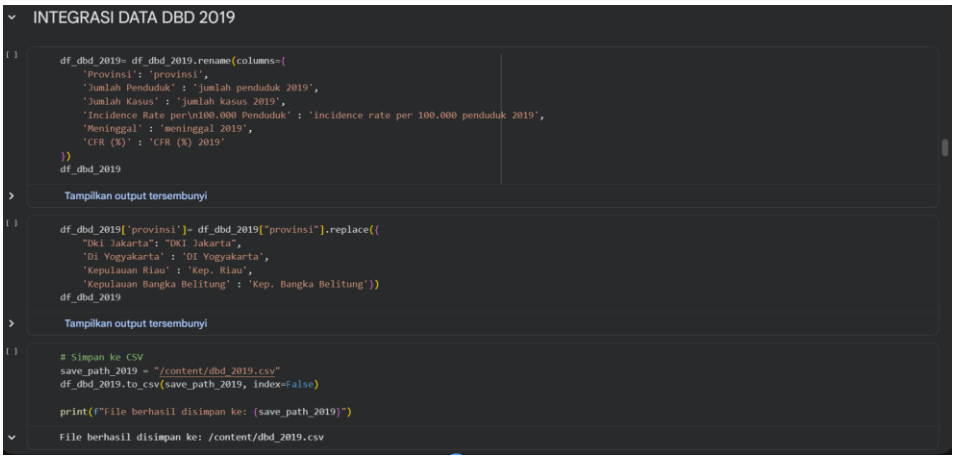
- Sebelum melakukan merge dilakukan standarisasi nama provinsi agar dapat di gabungkan.
  - Integrasi Dataset Iklim dengan Dataset Stasiun



    - Data hasil wrangling untuk df_final yang memuat data iklim disimpan dengan nama “iklim_per_Provinsi 2019-2020 2.csv”.
  - Integrasi Dataset Kepadatan penduduk tahun 2019 dan dataset penduduk 2020

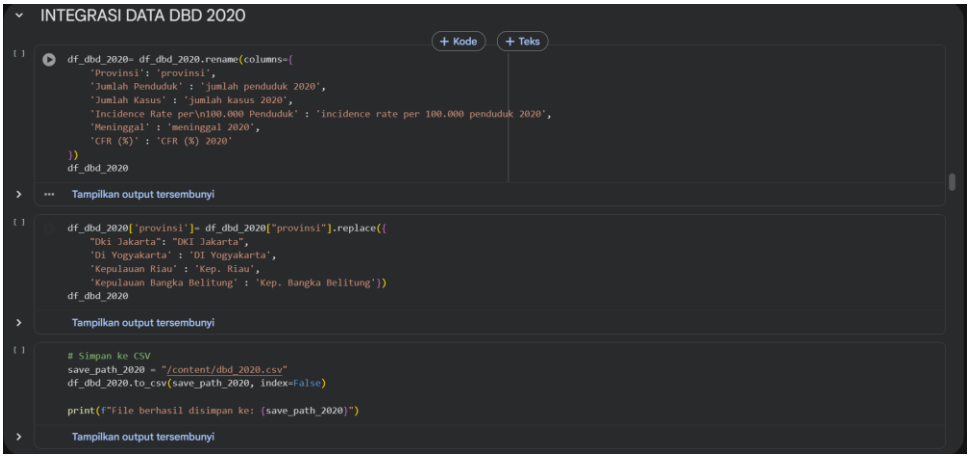
    - Data hasil wrangling untuk df_fix yang memuat data kepadatan penduduk 2019 dan 2020 disimpan dengan nama “kepadatan penduduk 2019-2020.csv”.

- Integrasi Dataset DBD tahun 2019



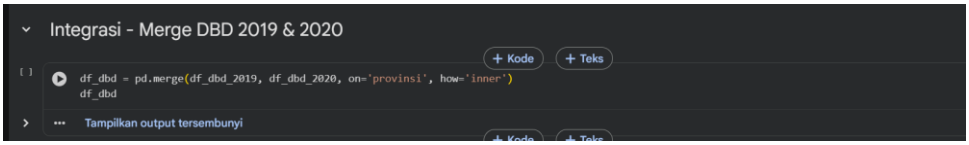
- Sebelum melakukan merge dilakukan pergantian nama kolom dan standarisasi nama provinsi. Kemudian data hasil nya disimpan dengan nama “dbd_2019.csv”.

- Integrasi Dataset DBD tahun 2020



  - Sebelum melakukan merge dilakukan pergantian nama kolom dan standarisasi nama provinsi. Kemudian data hasil nya disimpan dengan nama “dbd_2020.csv”.

- Integrasi Dataset DBD tahun 2019 dan Dataset DBD tahun 2020



  - Data hasil wrangling untuk df_dbd disimpan dengan nama “Kasus DBD tahun 2019-2020.csv”.

- Integrasi Dataset Gabungan (3 dataset)

  - Data hasil wrangling untuk df_gabungan disimpan dengan nama "Gabungan (all dataset).csv".

- Cek missing value akhir pada df_gabungan untuk memastikan data bersih dan siap digunakan dalam tahapan dan analisis lebih lanjut, misalnya untuk visualisasi (EDA).

d. Analisis Data Ekplorasi (*Data Exploration Analysis* (EDA))

- Statistik Deskriptif
  - Rata-rata

```
## Rata-rata (mean)
kolom = [
    "Tavg",          # Rata-rata Suhu
    "RH_avg",        # Rata-rata Kelembaban Relatif
    'RR',            # Curah Hujan
    'ss',             # Durasi sinar matahari
    'ff_avg',        # Kecepatan Angin Rata-rata
    'jumlah kasus 2019',    # Jumlah Kasus DBD 2019
    'jumlah kasus 2020',    # Jumlah Kasus DBD 2020
    'kepadatan 2019', # Kepadatan Penduduk 2019
    'kepadatan 2020' # Kepadatan Penduduk 2020
]

rata_rata = df_gabungan[kolom].mean()
print(rata_rata)
```

- Standar Deviasi

```
## Standar Deviasi
kolomstd = [
    "Tavg",          # Rata-rata Suhu
    "RH_avg",        # Rata-rata Kelembaban Relatif
    'RR',            # Curah Hujan
    'ss',             # Durasi sinar matahari
    'ff_avg',        # Kecepatan Angin Rata-rata
    'jumlah kasus 2019',    # Jumlah Kasus DBD 2019
    'jumlah kasus 2020',    # Jumlah Kasus DBD 2020
    'kepadatan 2019', # Kepadatan Penduduk 2019
    'kepadatan 2020' # Kepadatan Penduduk 2020
]

standar_deviasi = df_gabungan[kolomstd].std()
print(standar_deviasi)
```

- Nilai Minimum

```
# nilai minimum incidence rate per 100.000 penduduk tahun 2019
idx_min = df_gabungan["incidence rate per 100.000 penduduk 2019"].idxmin()
df_gabungan.loc[idx_min, ["provinsi", "tahun", "incidence rate per 100.000 penduduk 2019"]]
```

```
# nilai minimum jumlah kasus DBD tahun 2019
idx_minn = df_gabungan["jumlah kasus 2019"].idxmin()
df_gabungan.loc[idx_minn, ["provinsi", "tahun", "RR", "Tavg", "RH_avg", "ff_avg", "kepadatan 2019",
                           "incidence rate per 100.000 penduduk 2019", "jumlah penduduk 2019", "jumlah kasus 2019"]]
```

```
# nilai minimum jumlah kasus DBD tahun 2020
idx_mini = df_gabungan["jumlah kasus 2020"].idxmin()
df_gabungan.loc[idx_mini, ["provinsi", "tahun", "RR", "Tavg", "RH_avg", "ff_avg", "kepadatan 2020",
                           "incidence rate per 100.000 penduduk 2020", "jumlah penduduk 2020", "jumlah kasus 2020"]]
```

- Nilai Maksimum

```
# maksimum berdasarkan RR (curah hujan)
idx_max = df_gabungan["RR"].idxmax()
df_gabungan.loc[idx_max, ["provinsi", "tahun", "RR"]]
```

```
# nilai maksimum jumlah kasus DBD tahun 2019
idx_maks = df_gabungan["jumlah kasus 2019"].idxmax()
df_gabungan.loc[idx_maks, ["provinsi", "tahun", "RR", "Tavg", "RH_avg", "ff_avg", "kepadatan 2019",
                           "incidence rate per 100.000 penduduk 2019", "jumlah penduduk 2019", "jumlah kasus 2019"]]
```

```
# nilai maksimum jumlah kasus DBD tahun 2020
idx_maxx = df_gabungan["jumlah kasus 2020"].idxmax()
df_gabungan.loc[idx_maxx, ["provinsi", "tahun", "RR", "Tavg", "RH_avg", "ff_avg", "kepadatan 2020",
                           "incidence rate per 100.000 penduduk 2020", "jumlah penduduk 2020", "jumlah kasus 2020"]]
```

- Analisis Eksploratif Visual (Visualisasi EDA)
  - Time Series Rata-rata Suhu (Tavg) dan Kelembaban (RH_avg)

```
# Time Series Rata-rata Suhu (Tavg) dan Kelembaban (RH_avg)

# Agregasi harian rata-rata di seluruh provinsi
df_ts = df_gabungan.groupby('date')[['Tavg', 'RH_avg']].mean().reset_index()

# Plotting dengan twin axes
fig, ax1 = plt.subplots(figsize=(14, 6))

color_temp = 'tab:red'
ax1.set_xlabel('Tanggal')
ax1.set_ylabel('Rata-rata Suhu (°C)', color=color_temp)
ax1.plot(df_ts['date'], df_ts['Tavg'], color=color_temp, label='Suhu Rata-rata')
ax1.tick_params(axis='y', labelcolor=color_temp)

ax2 = ax1.twinx()  # Sumbu Y kedua untuk Kelembaban
color_hum = 'tab:blue'
ax2.set_ylabel('Kelembaban Rata-rata (%)', color=color_hum)
ax2.plot(df_ts['date'], df_ts['RH_avg'], color=color_hum, label='Kelembaban Rata-rata')
ax2.tick_params(axis='y', labelcolor=color_hum)

plt.title('Rata-rata Harian Suhu dan Kelembaban di Seluruh Provinsi')
fig.tight_layout()

plt.savefig("timeseries_tavg_rhavg.png", dpi=300, bbox_inches='tight')
# dpi=300 biar kualitas bagus # bbox_inches='tight' biar gak kepotong margin

plt.show()
plt.close()
```

- Bar Chart Total Kasus DBD per Provinsi (2019 vs 2020)

```python
# Bar Chart Total Kasus DBD per Provinsi (2019 vs 2020)

# Agregasi data kasus tahunan per provinsi (menggunakan .mean() karena nilai duplikat per hari)
df_prov_cases = df_gabungan.groupby('provinsi')[[
    'jumlah kasus 2019',
    'jumlah kasus 2020'
]].mean().sort_values(by='jumlah kasus 2020', ascending=False)

# Plotting
df_prov_cases.plot(kind='bar', figsize=(14, 7))
plt.title('Perbandingan Total Kasus DBD per Provinsi (2019 vs 2020)')
plt.ylabel('Jumlah Kasus')
plt.xlabel('Provinsi')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()

plt.savefig("bar_chart kasus dbd19&20 per provinsi.png", dpi=300, bbox_inches='tight')
# dpi=300 biar kualitas bagus # bbox_inches='tight' biar gak kepotong margin

plt.show()
plt.close()
```

- Agregasi untuk Korelasi

```python
# Agregasi data iklim dan DBD/Kepadatan per provinsi untuk KEDUA TAHUN
df_corr_all = df_gabungan.groupby(['provinsi', 'tahun']).agg({
    'Tavg': 'mean',  # Rata-rata suhu harian tahunan
    'RR': 'mean',  # Rata-rata curah hujan tahunan
    'incidence rate per 100.000 penduduk 2019': 'mean',  # IR 2019
    'incidence rate per 100.000 penduduk 2020': 'mean'   # IR 2020
}).reset_index()

# Buat kolom 'Incidence_Rate' tunggal berdasarkan tahun
df_corr_all['Incidence_Rate'] = np.where(  #np.where(KONDISI, NILAI_JIKA_TRUE, NILAI_JIKA_FALSE)
    df_corr_all['tahun'] == 2019,  # memilih salah satu tahun sebagai kondisi penentu
    df_corr_all['incidence rate per 100.000 penduduk 2019'],  # TRUE: Ambil data 2019 (_x)
    df_corr_all['incidence rate per 100.000 penduduk 2020']   # FALSE: Ambil data 2020 (_y)
)

df_corr_all
```

- Scatter Plot Suhu (Tavg) vs Incidence Rate DBD 2019 dan 2020

```python
# Scatter Plot Tavg vs Incidence Rate DBD
plt.figure(figsize=(10, 6))

# Loop untuk memplot data 2019 dan 2020 dengan warna berbeda
for year, group in df_corr_all.groupby('tahun'):
    plt.scatter(
        group['Tavg'],
        group['Incidence_Rate'],
        label=f'Tahun {year}',
        alpha=0.7,
        s=100  # Ukuran titik
    )

    # Menambahkan Label Provinsi
    for i in range(len(group)):
        provinsi_name = group.iloc[i]['provinsi']

        # Penyesuaian label agar tidak bertumpuk (optional, tergantung penyebaran data)
        # Contoh: Jika tahun 2019, label diletakkan sedikit di kiri titik.
        x_offset = -0.015 * max(group['Tavg']) if year == 2019 else 0.005 * max(group['Tavg'])
        y_offset = 2

        plt.text(
            group.iloc[i]['Tavg'] + x_offset,
            group.iloc[i]['Incidence_Rate'] + y_offset,
            provinsi_name,
            fontsize=8,
            alpha=0.8,
            ha='center',
            color=colors[year]
        )

    # --- Menghitung dan Memplot Garis Regresi ---
    X = group['Tavg']
    Y = group['Incidence_Rate']

    z = np.polyfit(X, Y, 1)
    p = np.poly1d(z)

    plt.plot(
        X,
        p(X),
        color=colors[year],
        linestyle=line_styles[year],
        label=f'Tren {year} ($y={z[0]:.2f}x+{z[1]:.2f}$)'
    )

plt.title('Hubungan Tavg vs Incidence Rate DBD (2019 vs 2020)')
plt.xlabel('Rata-rata Suhu Harian (°C)')
plt.ylabel('Incidence Rate per 100.000 Penduduk')
plt.legend(title='Tahun')
plt.grid(True, linestyle='--', alpha=0.5)
plt.tight_layout()
plt.show()
```

- Scatter Plot Curah Hujan (RR) vs Incidence Rate DBD 2019 dan 2020

```
# Scatter Plot Curah Hujan (RR) vs Incidence Rate DBD 2019 dan 2020
plt.figure(figsize=(14, 8)) # Ukuran diperbesar agar label tidak bertumpuk
colors = {2019: 'blue', 2020: 'red'}
line_styles = {2019: '--', 2020: '-'}

# Loop untuk memplot data scatter, garis regresi, dan label untuk setiap tahun
for year, group in df_corr_all.groupby('tahun'):
    # Scatter Plot
    plt.scatter(
        group['RR'],
        group['Incidence_Rate'],
        label=f'Data {year}',
        alpha=0.6,
        s=100,
        color=colors[year]
    )
```

```
'''
    # Menambahkan Label Provinsi
    for i in range(len(group)):
        provinsi_name = group.iloc[i]['provinsi']

        # Penyesuaian label agar tidak bertumpuk (optional, tergantung penyebaran data)
        # Contoh: Jika tahun 2019, label diletakkan sedikit di kiri titik.
        x_offset = -0.015 * max(group['RR']) if year == 2019 else 0.005 * max(group['RR'])
        y_offset = 2

        plt.text(
            group.iloc[i]['RR'] + x_offset,
            group.iloc[i]['Incidence_Rate'] + y_offset,
            provinsi_name,
            fontsize=8,
            alpha=0.8,
            ha='center',
            color=colors[year]
        )
'''
```

```
'''
    # --- Menghitung dan Memplot Garis Regresi ---
    X = group['RR']
    Y = group['Incidence_Rate']

    z = np.polyfit(X, Y, 1)
    p = np.poly1d(z)

    plt.plot(
        X,
        p(X),
        color=colors[year],
        linestyle=line_styles[year],
        label=f'Tren {year} ($y={z[0]:.2f}x+{z[1]:.2f}$)'
    )
'''
plt.title('Hubungan RR (Rainfall) vs Incidence Rate DBD (dengan Label Provinsi)')
plt.xlabel('Total Curah Hujan Tahunan (mm)')
plt.ylabel('Incidence Rate per 100.000 Penduduk')
plt.legend(title='Keterangan')
plt.grid(True, linestyle='--', alpha=0.5)
plt.tight_layout()

plt.savefig("scatter_RR_IR.png", dpi=300, bbox_inches='tight')
# dpi=300 biar kualitas bagus # bbox_inches='tight' biar gak kepotong margin

plt.show()
plt.close()
```

- Uji Normalitas Shapiro-Wilk

```
from scipy.stats import shapiro

kolom_uji = ["Tavg", "RH_avg", "RR", "ss", "ff_avg",
             "incidence rate per 100.000 penduduk 2019", "incidence rate per 100.000 penduduk 2020",
             "jumlah kasus 2019", "jumlah kasus 2020",
             "kepadatan 2019", "kepadatan 2020"]

hasil = {}

for kolom in kolom_uji:
    # sampling biar tidak error jika >5000
    data = df_gabungan[kolom].dropna().sample(5000, random_state=1)
    stat, p = shapiro(data)
    hasil[kolom] = p
```

```
import pandas as pd
pd.DataFrame(hasil, index=["p-value"]).T
```

Sebelum menganalisis data numerik, harus melakukan uji normalitas untuk mengetahui bagaimana distribusi normalitas data tersebut. Hasil uji normalitas digunakan untuk menentukan jenis uji

yang akan dipakai dalam analisis bivariat. Jika hasil uji menunjukan p value >0,05 maka data tersebut memiliki distribusi normal sehingga uji bivariat yang akan digunakan adalah uji korelasi Pearson. Jika hasil uji normalitas menunjukan p value < 0,05, maka data tersebut terdistribusi tidak normal sehingga uji bivariat yang digunakan adalah korelasi Spearman.

- Korelasi Spearman

  Korelasi Spearman digunakan karena metode ini tidak mengasumsikan hubungan linear dan tidak mensyaratkan data berdistribusi normal, sehingga lebih sesuai untuk data epidemiologi dan iklim yang sering bersifat non-linear, memiliki outlier, serta skala yang tidak selalu interval. Selain itu, Spearman mengukur hubungan monotonic—apakah suatu variabel cenderung naik ketika variabel lain naik—sehingga lebih robust terhadap nilai ekstrem dan pola hubungan yang tidak sepenuhnya linier, yang umum ditemukan pada data kasus DBD dan faktor lingkungannya.

```
kolom = [
    "jumlah kasus 2019",
    "jumlah kasus 2020",
    "Tavg",  #suhu rata-rata
    "RH_avg",  #kelembaban
    "RR",  #curah hujan
    "ss",  # durasi sinar matahari
    "ff_avg",  #kecepatan angin rata-rata
    "incidence rate per 100.000 penduduk 2019",
    "incidence rate per 100.000 penduduk 2020",
    "kepadatan 2019",
    "kepadatan 2020"
]

corr_spearman = df_gabungan[kolom].corr(method='spearman')
corr_spearman
```

- Heatmap Matriks Korelasi Tahun 2019

```
# Agregasi data iklim (Tavg, RH_avg, RR) dan data DBD/Kepadatan (Incidence Rate, Kepadatan) per provinsi untuk tahun 2019
df_corr_2019 = df_gabungan[df_gabungan['tahun'] == 2019].groupby('provinsi').agg({
    'Tavg': 'mean',  # Rata-rata suhu harian tahunan pada tahun 2019
    'RH_avg': 'mean',  # Rata-rata kelembaban harian tahunan pada tahun 2019
    'RR': 'sum',     # Total curah hujan tahunan pada tahun 2019
    'incidence rate per 100.000 penduduk 2019': 'mean',
    'kepadatan 2019': 'mean'
}).reset_index()
```

```
# Asumsi: df_corr_2019 sudah diagregasi per provinsi untuk tahun 2019

# Kolom yang akan dikorelasikan
corr_cols = [
    'Tavg',
    'RH_avg',
    'RR',
    'kepadatan 2019',
    'incidence rate per 100.000 penduduk 2019'
]

# Hitung Matriks Korelasi (Pearson)
# Pastikan semua kolom yang digunakan adalah tipe data numerik
df_corr_matrix = df_corr_2019[corr_cols].corr(method="spearman")

# Plot Heatmap
plt.figure(figsize=(9, 8))
sns.heatmap(
    df_corr_matrix,
    annot=True,        # Tampilkan nilai koefisien korelasi
    cmap='coolwarm',   # Skema warna yang baik untuk menunjukkan korelasi positif/negatif
    fmt=".2f",         # Format angka hingga 2 desimal
    linewidths=.5,     # Garis pemisah antar sel
    cbar_kws={'label': 'Koefisien Korelasi'}
)
```

```
# Sesuaikan label agar lebih ringkas di plot
plt.title('Matriks Korelasi Antar Variabel Kunci (Tahun 2019)')
plt.yticks(rotation=0)
plt.xticks(rotation=45, ha='right')
plt.tight_layout()

plt.savefig("heatmap_corr_2019.png", dpi=300, bbox_inches='tight')
# dpi=300 biar kualitas bagus # bbox_inches='tight' biar gak kepotong margin

plt.show()
```

- Heatmap Antar Variabel Iklim vs Jumlah Kasus DBD

```
import seaborn as sns

cols_corr = ["jumlah kasus 2019", "jumlah kasus 2020", "Tavg", "RH_avg", "RR", "Tx", "Tn", "ff_avg"]
corr = df_gabungan[cols_corr].corr(method="spearman")

plt.figure(figsize=(10,7))
sns.heatmap(corr, annot=True, cmap="coolwarm", fmt=".2f") # cmap bisa diganti coolwarm, plasma, inferno, magma, atau YlGnBu
plt.title("Heatmap Korelasi Faktor Iklim dan Kasus DBD")

plt.savefig("heatmap_corr_iklim_dbd.png", dpi=300, bbox_inches='tight')
# dpi=300 biar kualitas bagus # bbox_inches='tight' biar gak kepotong margin

plt.show()
```

- Boxplot Faktor Iklim

```
plt.figure(figsize=(10,6))
df_gabungan[["Tavg", "RH_avg", "RR", "ss"]].boxplot()
plt.title("Boxplot Distribusi Variabel Iklim")

plt.savefig("boxplot_iklim.png", dpi=300, bbox_inches='tight')
# dpi=300 biar kualitas bagus # bbox_inches='tight' biar gak kepotong margin

plt.show()
```

- Scatter Plot Kepadatan Penduduk vs Incidence Rate per 100.000 Pddk
  - 2019

```
plt.figure(figsize=(7,5))
plt.scatter(df_gabungan["kepadatan 2019"], df_gabungan["incidence rate per 100.000 penduduk 2019"], alpha=0.5)
plt.title("Kepadatan Penduduk vs Incidence Rate per 100.000 penduduk (2019)")
plt.xlabel("Kepadatan Penduduk (/km²)")
plt.ylabel("Incidence Rate per 100.000 penduduk (2019)")
plt.grid()

plt.savefig("scatter_kepadatan_IR_2019.png", dpi=300, bbox_inches='tight')
# dpi=300 biar kualitas bagus # bbox_inches='tight' biar gak kepotong margin

plt.show()
```

  - 2020

```
plt.figure(figsize=(7,5))
plt.scatter(df_gabungan["kepadatan 2020"], df_gabungan["incidence rate per 100.000 penduduk 2020"], alpha=0.5)
plt.title("Kepadatan Penduduk vs Incidence Rate per 100.000 penduduk (2020)")
plt.xlabel("Kepadatan Penduduk (/km²)")
plt.ylabel("Incidence Rate per 100.000 penduduk (2020)")
plt.grid()

plt.savefig("scatter_kepadatan_IR_2020.png", dpi=300, bbox_inches='tight')
# dpi=300 biar kualitas bagus # bbox_inches='tight' biar gak kepotong margin

plt.show()
```

- Grafik Tren Musiman Kasus DBD per Bulan
  - 2019

```
df_gabungan["bulan"] = df_gabungan["date"].dt.month

df_musim = df_gabungan.groupby("bulan")["jumlah kasus 2019"].sum()

plt.figure(figsize=(8,5))
df_musim.plot(marker='o')
plt.title("Tren Musiman Kasus DBD Tahun 2019")
plt.xlabel("Bulan")
plt.ylabel("Jumlah Kasus")
plt.grid()

plt.savefig("tren musiman kasus dbd 2019.png", dpi=300, bbox_inches='tight')
# dpi=300 biar kualitas bagus # bbox_inches='tight' biar gak kepotong margin

plt.show()
```

➢ 2020

```
df_gabungan["bulan"] = df_gabungan["date"].dt.month

df_musim = df_gabungan.groupby("bulan")["jumlah kasus 2020"].sum()

plt.figure(figsize=(8,5))
df_musim.plot(marker='o')
plt.title("Tren Musiman Kasus DBD Tahun 2020")
plt.xlabel("Bulan")
plt.ylabel("Jumlah Kasus")
plt.grid()

plt.savefig("tren musiman kasus dbd 2020.png", dpi=300, bbox_inches='tight')
# dpi=300 biar kualitas bagus # bbox_inches='tight' biar gak kepotong margin

plt.show()
```

- Rata-Rata Suhu harian di provinsi tertentu

```
# Provinsi bisa disesuaikan
prov = "Aceh"

df_plot = df_gabungan[df_gabungan["provinsi"] == prov]

plt.figure(figsize=(10,5))
plt.plot(df_plot["date"], df_plot["Tavg"])
plt.title(f"Rata-rata Suhu Harian di {prov}")
plt.xlabel("Tanggal")
plt.ylabel("Tavg")
plt.grid()

plt.savefig(f"rata-rata suhu harian di {prov}.png", dpi=300, bbox_inches='tight')
# dpi=300 biar kualitas bagus # bbox_inches='tight' biar gak kepotong margin

plt.show()
```

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil Proses Wrangling

Proses data wrangling dilakukan untuk memastikan bahwa seluruh data yang digunakan dalam penelitian ini berada dalam kondisi yang bersih, konsisten, dan siap untuk dianalisis. Pada tahap ini, data dari berbagai sumber—meliputi data iklim, kepadatan penduduk, dan kasus DBD—digabungkan dan diproses melalui serangkaian langkah seperti pembersihan data, standarisasi format, penanganan nilai hilang, serta transformasi variabel yang diperlukan. Hasil dari proses wrangling ini menghasilkan dataset final yang terintegrasi dengan struktur yang lebih rapi dan mudah diolah, sehingga dapat mendukung analisis pengaruh faktor iklim dan kepadatan penduduk terhadap kasus DBD pada tingkat provinsi secara lebih akurat dan terarah.

4.1.1 Pengambilan Data

Pengambilan data dalam penelitian ini dilakukan dari tiga sumber utama, yaitu data iklim, data kepadatan penduduk, dan data kasus Demam Berdarah Dengue (DBD). Setelah proses pengambilan data dilakukan, seluruh data dari ketiga sumber siap digunakan untuk proses pembersihan serta tahapan dan analisis lebih lanjut.

a. Data Iklim

   Data iklim diperoleh dari dataset *Indonesia Climate* yang tersedia di Kaggle, kemudian diunduh dan dibaca menggunakan fungsi pd.read_csv() untuk tiga file utama, yaitu *climate_data.csv*, *station_detail.csv*, dan *province_detail.csv*. Langkah ini memastikan seluruh informasi parameter iklim per provinsi dapat terintegrasi dalam satu kerangka data.

   - Fitur -fitur untuk *climate_data.csv*
     - Tn = Suhu minimum (°C)
     - Tx = Suhu maksimum (°C)
     - Tavg = Suhu rata-rata (°C)
     - RH_avg = kelembaban rata-rata (%)
     - RR = curah hujan (mm)
     - ss = durasi sinar matahari (jam)
     - ff_x = kecepatan angin maksimum (m/s)
     - ddd_x = arah angin pada kecepatan maksimum (°)

- ff_avg = kecepatan angin rata-rata (m/s)
- ddd_car = arah angin dominan (°)

- Fitur-fitur untuk station_detail.csv
  - station_id
  - station_name
  - region_name
  - latitude
  - longitude
  - region_id
  - province_id
- Fitur-fitur untuk province_detail.csv
  - province_id
  - province_name

b. Data Kepadatan Penduduk

Data kepadatan penduduk tahun 2019 dan 2020 diambil dari publikasi Badan Pusat Statistik (BPS) dalam bentuk file CSV, yang masing-masing dimuat menggunakan pd.read_csv() untuk menghasilkan dataframe df19 untuk menyimpan data kepadatan penduduk tahun 2019 dan df20 untuk menyimpan data kepadatan penduduk tahun 2020. Data kepadatan penduduk untuk tahun 2019 dan 2020 masih mengikuti format CSV awal dan belum bersih dalam format penamaan, seperti hasil output berikut.

```
1 df19
```

	38 Provinsi	Unnamed: 1
0	NaN	Kepadatan Penduduk menurut Provinsi (jiwa/km2)
1	NaN	2019
2	ACEH	93
3	SUMATERA UTARA	200

```
1 df20
```

	38 Provinsi	Unnamed: 1
0	NaN	Kepadatan Penduduk menurut Provinsi (jiwa/km2)
1	NaN	2020
2	ACEH	91
3	SUMATERA UTARA	203

c. Data Kasus Demam Berdarah Dengue (DBD) Menurut Provinsi di Indonesia

Data kasus DBD diperoleh dari dokumen *Profil Kesehatan Indonesia 2019* dan *Profil Kesehatan Indonesia 2020* yang berformat PDF. Proses ekstraksi atau scraping tabel pada file PDF dilakukan melalui instalasi pustaka tambahan seperti tabula-py, pandas, pyreadstat, openpyxl, dan PyMuPDF. Setiap halaman PDF dipindai menggunakan metode find_tables() untuk mendeteksi seluruh tabel yang berisi informasi jumlah kasus DBD per provinsi. Tabel yang ditemukan kemudian diekstraksi menjadi dataframe dengan memanfaatkan modul PyMuPDF (fitz) dan dikonversi ke dalam format terstruktur berupa dataframe menggunakan pandas.

- Tabel yang ditargetkan untuk dataset kasus DBD tahun 2019 mengenai Kasus Demam Berdarah Dengue (DBD) Menurut Provinsi di Indonesia Tahun 2019 berada pada halaman 458, maka proses ekstraksi data difokuskan hanya pada halaman tersebut yang diambil dari Profil-Kesehatan-Indonesia-2019.pdf
- Tabel yang ditargetkan untuk dataset kasus DBD tahun 2020 mengenai Kasus Demam Berdarah Dengue (DBD) Menurut Provinsi di Indonesia Tahun 2020 berada pada halaman 453, maka proses ekstraksi data difokuskan hanya pada halaman tersebut yang diambil dari Profil-Kesehatan-Indonesia-2020.pdf
- Fitur-fitur pada data df_dbd_2019 dan df_dbd_2020
  - No
  - Provinsi
  - Jumlah Penduduk
  - Jumlah Kasus
  - Incidence Rate per 100.000 Penduduk
  - Meninggal
  - CFR (%)

4.1.2 Cleaning Data

Bagian ini merupakan proses pembersihan data yang dilakukan utnuk tiga dataset yaitu dataset iklim, kepadatan pendduduk, dan kasus DBD. Cleaning data ini dilakukan untuk memastikan data dalam kondisi rapi dengan menghilangkan ketidakkonsistenan, kesalahan penulisan, menangani nilai yang hilang, menghapus duplikasi, serta membuang variabel yang tidak relevan.

1. Standarisasi Nama Provinsi

   Ketiga dataset digunakan dari sumber berbeda sehingga format nama provinsi tidak seragam seperti huruf kapital yang berbeda, dan variasi penulisan namma provinsi yang berbeda. Proses cleaning yang dilakukan adalah mengubah seluruh nama provinsi menjadi Title Case, kemudian mengubah penamaan "Dki Jakarta" dan "Di yogyakarta" menjadi "DKI Jakarta" dan "DI yogaykarta". Hasil yang diperoleh berupa seluruh dataset memiliki 33 provinsi dengan format yang sama sehingga dapat di-merge dengan benar.

2. Konversi Tipe Data

   Beberapa kolom pada dataset bertipe object padahal merupakan angka atau tanggal. Perbaikan pertama yang dilakukan pada kolom "date" dikonversi ke "datetime". Perbaikan kedua yaitu kolom kolom tahun dibuat baru menggunakan dt.year dari kolom tanggal iklim. Perbaikan yang ketiga pada kolom numerik yang dikonversi ke integer. Hasil yang diperoleh yaitu sesluruh dataset sudah sesuai tipe datanya sehingga dapat dihitung, diagregasi, dan dianalisis.

3. Menangani Missing Value

   Setiap dataset memiliki nilai kosong di beberapa variabel. Pada dataset iklim missing value terletak pada suhu dan kelembapan dengan penanganan berupa penggantian menjadi angka 0. Pada dataset kepadatan penduduk beberapa provinsi tidak memiliki nilai kepadatan akhirnya di drop baris provinsi. Dataset DBD terkadang nilai kematian kosong atau diisi 0 dengan begitu menggunakan fill 0 untuk kolom kematian yang kosong. Hasil yang diperoleh, yaitu tidak ada missing value pada dataset final sehingga hasil analisis lebih stabil dan tidak bias.

4. Seleksi Tahun

   Karena fokus penelitian pada rentan waktu 2019-2020 maka dataset iklim yang mencakup banyak tahun difilter menjadi tahun 2019-2020. Hasilnya berupa dataset yang lebih ringkas, fokus, dan sesuai kebutuhan penelitian.

4.1.3 Integrasi Data

Pada bagian ini membahas kualitas, konsistensi, kesesuaian data setelah proses wrangling, untuk memastikan bahwa data yang dihasilkan benar-benar layak dan dapat diandalkan untuk analisis selanjutnya.

1. Konsistensi Antar Dataset

   Setelah cleaning kemudian dilakukan konsistensi terhadap nama provinsi dimana semua dataset harus memiliki 33 provinsi yang sama. Kemudian konsistensi jumlah baris pertahun dan jumlah data iklim yang digabungkan dengan data DBD per provinsi. Menghasilkan data yang sesuai, dari jumlah provinsi atau tahun antar dataset semua sama. Sehingga proses merging berhasil.

Nama Provinsi Awal	Nama Provinsi Akhir
Nanggroe Aceh Darussalam	Aceh
Dki Jakarta	DKI Jakarta
Di Yogyakarta	DI Yogyakarta
Kepulauan Riau	Kep. Riau
Kepulauan Bangka Belitung	Kep. Bangka Belitung

2. Merging Data

   Proses penggabungan data dilakukan melalui kolom provinsi dan tahun yaitu 33 provinsi dan tahun 2019-2020. Integrasi menghasilkan dataset akhir yang disimpan dalam dataframe df_gabungan dan file CSV “Gabungan (all dataset).csv dengan struktur kolom provinsi, iklim, kepadatan penduduk, dan kasus DBD.

Merge Semua Dataset (DBD, Iklim, dan Kepadatan Penduduk) Tahun 2019 - 2020

[75] 0 d
```
df_gabungan = pd.merge(df_dbd, df_IP, on='provinsi', how='inner')
df_gabungan
```

	provinsi	jumlah penduduk 2019	jumlah kasus 2019	incidence rate per 100.000 penduduk 2019	meninggal 2019	CFR (%) 2019	jumlah penduduk 2020	jumlah kasus 2020	incidence rate per 100.000 penduduk 2020	meninggal 2020	...	ss	ff_x
0	Aceh	5371532	2386	44.42	6	0.25	5459891	891	0.0	1	...	8.5	4.0
1	Aceh	5371532	2386	44.42	6	0.25	5459891	891	0.0	1	...	9.0	5.0
2	Aceh	5371532	2386	44.42	6	0.25	5459891	891	0.0	1	...	6.5	4.0
3	Aceh	5371532	2386	44.42	6	0.25	5459891	891	0.0	1	...	7.5	4.0
4	Aceh	5371532	2386	44.42	6	0.25	5459891	891	0.0	1	...	8.0	4.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
114365	Papua	3379302	597	17.67	6	1.01	3435430	172	5.0	1	...	1.4	4.0
114366	Papua	3379302	597	17.67	6	1.01	3435430	172	5.0	1	...	3.0	12.0
114367	Papua	3379302	597	17.67	6	1.01	3435430	172	5.0	1	...	6.5	5.0
114368	Papua	3379302	597	17.67	6	1.01	3435430	172	5.0	1	...	2.4	7.0
114369	Papua	3379302	597	17.67	6	1.01	3435430	172	5.0	1	...	5.8	7.0

114370 rows × 27 columns

4.2 Hasil Eksplorasi Data (EDA)

Exploratory Data Analysis (EDA) merupakan tahapan penting dalam proses pengolahan data untuk memahami karakteristik awal, distribusi variabel, pola hubungan, serta potensi anomali dalam dataset. Pada penelitian ini, EDA dilakukan terhadap tiga jenis

data utama, yaitu data iklim, data kepadatan penduduk, dan data kasus Demam Berdarah Dengue (DBD) tahun 2019–2020. Tahap ini menggunakan *library* pandas, numpy, matplotlib.pyplot, dan seaborn.

Proses eksplorasi ini bertujuan untuk memberikan gambaran menyeluruh mengenai kondisi iklim antar provinsi, kepadatan penduduk Indonesia menurut provinsi, variasi kasus DBD, serta potensi keterkaitan antar variabel. Melalui analisis deskriptif, visualisasi grafik, serta uji statistika pendukung, EDA memberikan dasar yang kuat sebelum melakukan analisis lanjutan. Berbagai metode eksplorasi, seperti perhitungan statistik dasar, analisis tren, scatter plot, heatmap korelasi, dan pengujian normalitas digunakan untuk memahami dinamika data tahun 2019 dan 2020 secara lebih komprehensif untuk mendukung penentuan metode analisis statistik yang paling sesuai.

4.2.1 Statistik Deskriptif

4.2.1.1 Rata-Rata

Hasil perhitungan statistik deskriptif menunjukkan bahwa rata-rata suhu udara (Tavg) di seluruh provinsi Indonesia pada periode 2019–2020 berada pada kisaran 24,24°C. Nilai ini menggambarkan bahwa Indonesia berada pada zona iklim tropis dengan suhu yang relatif stabil dan hangat sepanjang tahun, kondisi yang ideal bagi perkembangan nyamuk Aedes aegypti. Kelembaban udara rata-rata (RH_avg) tercatat 73,70%, menunjukkan tingkat kelembaban tinggi yang dapat memperpanjang umur nyamuk dewasa serta meningkatkan peluang penularan virus dengue. Curah hujan (RR) rata-rata mencapai 6,52 mm, sementara durasi penyinaran matahari (ss) memiliki rata-rata 5,42 jam per hari. Kedua faktor ini berperan dalam membentuk lingkungan yang memungkinkan terbentuknya tempat perindukan nyamuk, terutama ketika curah hujan cukup tinggi tetapi tidak ekstrem. Kecepatan angin rata-rata (ff_avg) sebesar 1,76 m/s, nilai yang relatif rendah sehingga tidak banyak mengganggu aktivitas terbang nyamuk.

Untuk variabel epidemiologis, jumlah kasus DBD rata-rata antar provinsi tercatat 4.739 kasus pada tahun 2019 dan menurun menjadi 3.443 kasus pada tahun 2020. Penurunan ini dapat disebabkan oleh berbagai faktor, termasuk perubahan perilaku masyarakat pada awal pandemi COVID-19, peningkatan kesadaran kesehatan, atau kemungkinan keterbatasan pelaporan kasus. Dari aspek demografi, rata-rata kepadatan penduduk mencapai 492,42 jiwa/km² pada tahun 2019 dan 491,51 jiwa/km² pada tahun 2020, menunjukkan bahwa Indonesia memiliki variasi

tingkat kepadatan antar provinsi yang cukup kontras. Provinsi dengan kepadatan tinggi umumnya memiliki risiko penularan DBD lebih besar karena nyamuk Aedes memiliki jarak terbang yang pendek, sehingga lingkungan padat mempercepat kontak antara nyamuk dan manusia. Rata-rata variabel tersebut memberikan gambaran awal mengenai kondisi iklim, demografi, dan situasi kasus DBD sebelum dilakukan analisis lebih lanjut.

4.2.1.2 Standar Deviasi

Variabel iklim seperti suhu rata-rata (Tavg), kelembapan (RH_avg), curah hujan (RR), durasi penyinaran matahari (ss), dan kecepatan angin (ff_avg) memiliki standar deviasi yang relatif kecil. Ini menunjukkan bahwa kondisi iklim antar provinsi di Indonesia cenderung tidak terlalu bervariasi pada tahun 2019–2020. Jumlah kasus DBD tahun 2019 dan 2020 memiliki standar deviasi yang cukup besar (±5600 dan ±4400), menandakan adanya perbedaan kasus yang sangat besar antar provinsi—ada provinsi yang kasusnya tinggi dan ada yang sangat rendah. Kepadatan penduduk juga menunjukkan variasi yang besar antar provinsi (±1783 jiwa/km²). Ini menggambarkan bahwa distribusi penduduk Indonesia tidak merata, dengan provinsi tertentu sangat padat dan yang lain sangat jarang penduduk.

4.2.1.3 Nilai Minimum

a. Nilai minimum incidence rate per 100.000 penduduk tahun 2019

	95399
provinsi	Maluku
tahun	2019
incidence rate per 100.000 penduduk 2019	13.09

Provinsi Maluku memiliki nilai incidence rate DBD paling rendah pada tahun 2019, yaitu sebesar 13,09 kasus per 100.000 penduduk. Angka ini menunjukkan bahwa tingkat penularan DBD di Maluku relatif rendah dibandingkan provinsi lain. Rendahnya nilai ini dapat dipengaruhi oleh jumlah kasus yang sedikit (236 kasus) serta kepadatan penduduk yang rendah, sehingga peluang kontak antara manusia dan nyamuk pembawa virus menjadi lebih kecil.

b. Nilai minimum jumlah kasus DBD tahun 2019

	95399
provinsi	Maluku
tahun	2019
RR	1.9
Tavg	0.0
RH_avg	0.0
ff_avg	1.0
kepadatan 2019	38.0
incidence rate per 100.000 penduduk 2019	13.09
jumlah penduduk 2019	1802870
jumlah kasus 2019	236

	95399
provinsi	Maluku
tahun	2019
RR	1.9
Tavg	0.0
RH_avg	0.0
ff_avg	1.0
kepadatan 2020	39.0
incidence rate per 100.000 penduduk 2020	4.2
jumlah penduduk 2020	1831880
jumlah kasus 2020	77

Pada tahun 2019, provinsi dengan jumlah kasus DBD terendah adalah Maluku, yaitu 236 kasus. Rendahnya kasus pada wilayah ini diduga berkaitan dengan kepadatan penduduk yang rendah (38 jiwa/km²), sehingga potensi interaksi antara vektor dan manusia lebih kecil. Selain itu, beberapa variabel iklim seperti curah hujan (RR 1.9 mm) dan parameter iklim lain yang tercatat rendah dapat turut memengaruhi terbatasnya tempat berkembang biaknya nyamuk.

Pada tahun 2020, Maluku kembali menjadi provinsi dengan kasus DBD terendah yaitu 77 kasus, mengalami penurunan signifikan dari tahun sebelumnya. Kepadatan penduduk tetap rendah (39 jiwa/km²), sehingga pola risiko penularan masih relatif kecil. Tingkat kejadian (incidence rate) juga menurun menjadi 4,2 per 100.000 penduduk, menunjukkan bahwa persebaran penyakit semakin terkendali di wilayah tersebut.

4.2.1.4 Nilai Maksimum

a. Maksimum berdasarkan RR (curah hujan)

	64860
provinsi	Kalimantan Barat
tahun	2019
RR	325.7

Provinsi Kalimantan Barat tercatat memiliki curah hujan (RR) tertinggi, yaitu 325,7 mm pada tahun 2019. Curah hujan yang sangat tinggi seperti ini biasanya berpotensi meningkatkan tempat perindukan nyamuk (genangan air), meskipun curah hujan ekstrem juga dapat mengurangi larva pada kondisi tertentu.

b. Maksimum jumlah kasus DBD tahun 2019 dan 2020

	31257
provinsi	Jawa Barat
tahun	2019
RR	23.5
Tavg	20.2
RH_avg	94.0
ff_avg	1.0
kepadatan 2019	1394.0
incidence rate per 100.000 penduduk 2019	47.62
jumlah penduduk 2019	49316712
jumlah kasus 2019	23483

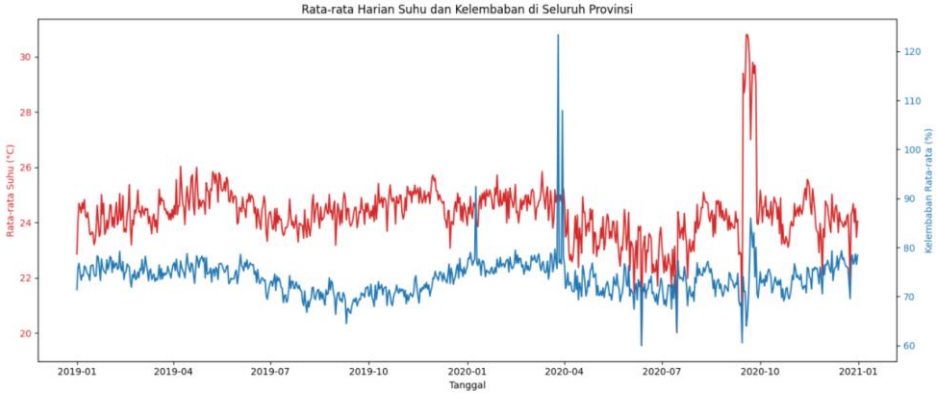
	31257
provinsi	Jawa Barat
tahun	2019
RR	23.5
Tavg	20.2
RH_avg	94.0
ff_avg	1.0
kepadatan 2020	1365.0
incidence rate per 100.000 penduduk 2020	45.3
jumlah penduduk 2020	49935858
jumlah kasus 2020	22613

Pada tahun 2019, provinsi dengan jumlah kasus DBD tertinggi adalah Jawa Barat, yaitu mencapai 23.483 kasus. Tingginya angka kasus ini sejalan dengan jumlah penduduk yang sangat besar, yaitu lebih dari 49 juta jiwa, serta kepadatan penduduk yang tinggi (1.394 jiwa/km²). Kondisi ini memungkinkan penularan DBD lebih cepat karena banyaknya kontak antara manusia dan nyamuk vektor. Selain itu, kelembaban udara yang tinggi (94%) dan curah hujan yang cukup (23,5 mm) turut mendukung tumbuhnya habitat perkembangbiakan nyamuk Aedes.

Pada tahun 2020, provinsi Jawa Barat tetap menjadi daerah dengan jumlah kasus DBD tertinggi, yaitu sebanyak 22.613 kasus. Meskipun terjadi sedikit penurunan dari tahun 2019, angka ini masih menjadi yang terbesar secara nasional. Faktor seperti jumlah penduduk sangat besar (49,9 juta jiwa) dan kepadatan penduduk yang tinggi (1.365 jiwa/km²) tetap menjadi penyebab utama tingginya risiko penularan. Kondisi iklim yang relatif serupa dengan tahun sebelumnya di mana kelembaban tinggi dan curah hujan sedang turut mendukung keberlangsungan vektor nyamuk.

4.2.2 Analisis Eksploratif Visual (Visualisasi EDA)

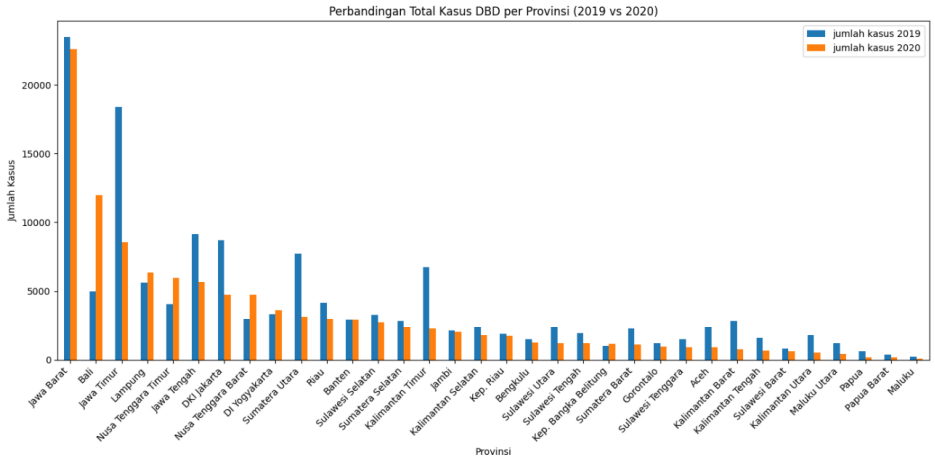
4.2.2.1 Time Series Suhu Rata-Rata (Tavg) dan Kelembaban Rata-Rata (RH_Avg)



Gambar timeseries_tavg_rhavg.png

Grafik *time series* suhu (garis merah) dan kelembaban (garis biru) harian dari 2019 hingga 2021 menunjukkan pola iklim tropis yang relatif stabil, di mana suhu umum berkisar antara 23°C hingga 26°C dan kelembaban antara 70% - 85%. Secara keseluruhan, terlihat korelasi negatif yang jelas, dengan peningkatan suhu sering diikuti oleh penurunan kelembaban relatif. Namun, pola ini diselingi oleh dua anomali ekstrem di tahun 2020, yaitu awal 2020 ditandai dengan lonjakan kelembaban ekstrem (di 100%) dan penurunan suhu tajam, mengindikasikan peristiwa cuaca basah yang intens. Sedangkan akhir 2020 menunjukkan lonjakan suhu ekstrem (mencapai lebih dari 30°C diikuti oleh fluktuasi kelembaban tinggi. Dalam konteks penyakit DBD, periode dengan kelembaban tinggi yang berkelanjutan dan suhu optimal 25°C – 27°C, menciptakan kondisi ideal untuk perkembangbiakan nyamuk *Aedes aegypti*, di mana kelembaban tinggi cenderung memperpanjang umur nyamuk, sedangkan suhu yang stabil mendukung proses perkembangbiakan dan aktivitas vektor.

4.2.2.2 Bar Chart Total Kasus DBD per Provinsi (2019 vs 2020)



Gambar bar_chart kasus dbd19&20 per provinsi.png

Bar Chart yang membandingkan Total Kasus Demam Berdarah Dengue (DBD) untuk 34 Provinsi pada tahun 2019 (biru) dan 2020 (oren), dengan provinsi diurutkan berdasarkan jumlah kasus pada tahun 2020. Sebagian besar provinsi mengalami penurunan jumlah kasus dari 2019 ke 2020, tetapi terdapat beberapa pengecualian yang justru mengalami kenaikan, salah satunya Provinsi Bali. Pada grafik terlihat jelas bahwa Bali mengalami peningkatan yang mencolok dari sekitar lima ribu kasus pada 2019 menjadi sekitar dua belas ribu kasus pada 2020. Disusul

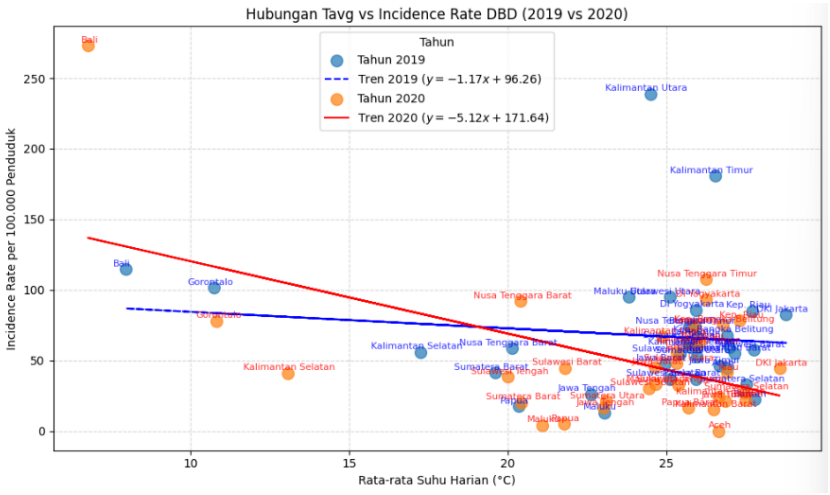
provinsi lainnya yang mengalami peningkatan jumlah kasus di 2020, seperti Lampung, NTT, NTB, DI Yogyakarta, Banten, dan Kep. Bangka Belitung.

Sementara itu, provinsi dengan beban kasus tertinggi seperti Jawa Barat tetap menempati posisi teratas pada kedua tahun, meskipun jumlah kasusnya turun cukup signifikan pada 2020. Beberapa provinsi lainnya yang mengalami penurunan jumlah kasus DBD adalah Jawa Timur, Jawa Tengah, DKI Jakarta, Sumetera Utara, Aceh, Kalimantan Tengah, Kalimantan Barat, Sulawesi Utara, dan berbagai provinsi lainnya. Secara keseluruhan, grafik ini menunjukkan bahwa tren nasional kasus DBD menurun pada tahun 2020, dengan variasi antarprovinsi yang dipengaruhi oleh kepadatan penduduk, kondisi lingkungan, serta efektivitas pengendalian vektor di masing-masing daerah.

4.2.2.3 Agregasi untuk Korelasi

Agregasi dan penyatuan data perlu dilakukan agar analisis korelasi antara iklim (suhu & curah hujan) dan insidensi DBD dapat dilakukan dalam satu kolom dan satu struktur data yang rapi. Agregasi dengan groupby(['provinsi', 'tahun']) digunakan untuk memperoleh nilai rata-rata iklim per provinsi per tahun, sehingga setiap baris mewakili satu provinsi pada satu tahun tertentu. Karena data incidence rate DBD awalnya dipisah menjadi dua kolom (IR 2019 dan IR 2020), maka dibuat kolom baru Incidence_Rate yang memilih nilai insidensi sesuai tahunnya. Dengan cara ini, kamu mendapatkan satu kolom metrik DBD yang konsisten, sehingga analisis korelasi antara variabel iklim dan insidensi bisa dilakukan lebih mudah, tanpa harus memisahkan tahun atau menulis kode terpisah untuk tiap kolom.

4.2.2.4 Scatter Plot Suhu (Tavg) vs Incidence Rate DBD 2019 dan 2020



Gambar scatter_tavg_IR.png

Scatter plot tersebut menunjukkan hubungan antara rata-rata suhu harian (Tavg) dan angka insidensi DBD per 100.000 penduduk di seluruh provinsi Indonesia untuk tahun 2019 dan 2020. Secara umum, titik-titik data tersebar luas tanpa membentuk pola linier yang jelas, sehingga hubungan antara suhu dan insidensi DBD tampak lemah atau tidak konsisten pada kedua tahun.

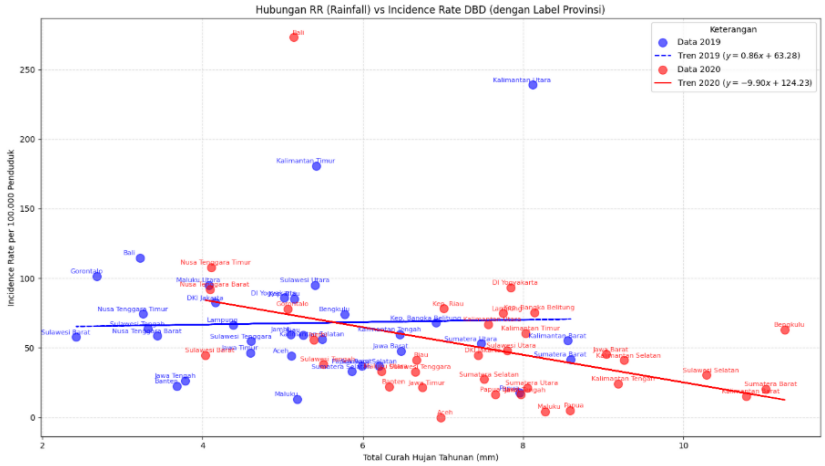
Titik data pada tahun 2019 yang memiliki *Incidence Rate* sangat tinggi cenderung tersebar pada rentang suhu yang lebih luas (sekitar 7°C hingga 25°C). Terdapat beberapa kasus tinggi yang ekstrem pada suhu dingin. Terdapat satu kasus dengan *Incidence Rate* tertinggi (sekitar 270) pada suhu yang sangat dingin (sekitar 6°C). Namun, sebagian besar titik oranye (2020) menunjukkan Incidence Rate yang lebih rendah dibandingkan titik biru (2019) di rentang suhu optimal 25°C). Ini mengindikasikan bahwa secara agregat, kasus DBD mungkin menurun di banyak provinsi pada tahun 2020 di rentang suhu optimal tersebut.

Suhu menjadi komponen penting dalam penyebaran DBD. Nyamuk Aedes membutuhkan suhu optimal pada rentang 25-30°C untuk mendukung perkembangannya. Pada suhu di atas 32°C hingga 35°C, nyamuk dapat mati, umur nyamuk memendek, dan laju gigitan nyamuk menurun. Pada suhu di bawah 18°C, nyamuk *Aedes aegypti* tidak dapat bereproduksi atau virus dengue tidak dapat bereplikasi. Namun, misalnya pada provinsi Bali dan Aceh untuk perbandingan. Bali pada tahun 2019 dan 2020 memiliki suhu cenderung rendah, tetapi jumlah kasus DBD berdasarkan visualisasi bar chart menunjukkan Bali termasuk salah satu provinsi dengan jumlah kasus DBD yang tinggi. Sementara Aceh, meskipun berada pada suhu di rentang optimal, Aceh tidak termasuk ke dalam salah satu provinsi penghasil kasus DBD terbanyak di indonesia.

Grafik menampilkan hubungan non-linier antara Rata-rata Suhu Harian (Tavg) dan *incidence rate* DBD, meskipun garis tren linier menunjukkan korelasi negatif yang lemah, di mana garis biru ($y = –1.17x + 96.26$ pada 2019) dan garis merah ($y = –5.12x + 171.64$ pada 2020), yang ditunjukkan oleh kedua garis tren yang miring ke bawah. Korelasi negatif ini utamanya didorong oleh outlier pada suhu rendah (di bawah 15°C), seperti Bali dan Gorontalo, yang menunjukkan *incidence rate* sangat tinggi (mencapai > 250 per 100.000 penduduk) meskipun suhunya dingin. Hal ini mengindikasikan bahwa pada suhu non-optimal, faktor lain seperti kelembaban ekstrem menjadi pemicu wabah yang lebih kuat. Sementara itu, pada rentang suhu optimal tropis sebaran kasus padat, tetapi sebagian besar provinsi menunjukkan

incidence rate yang relatif lebih rendah dan banyak titik menurun dari tahun 2019 ke 2020. Hal ini menunjukkan bahwa suhu udara (Tavg) dengan *incidence rate* DBD tidak terdapat hubungan signifikan di tahun 2019 dan 2020. Hasil ini sesuai dengan studi yang dilakukan oleh Wulandari et al. (2023) yang melaporkan bahwa antara suhu udara dengan kejadian Demam Berdarah Dengue, tidak ditemukan adanya korelasi signifikan.

4.2.2.5 Scatter Plot Curah Hujan (RR) vs Incidence Rate DBD 2019 dan 2020



Gambar scatter_RR_IR.png

Grafik di atas membandingkan Total Curah Hujan Tahunan (RR - Rainfall) dan Insidence Rate DBD per 100.000 Penduduk (IR) untuk tahun 2019 dan 2020 untuk mengetahui hubungan di antaranya, di mana tahun 2019 (biru) dan tahun 2020 (merah). Kedua tahun menunjukkan pola korelasi yang berlawanan, yang diwakili oleh garis tren linier. Garis biru putus-putus sedikit ke dengan y = 0.86x + 63.28 menyiratkan bahwa pada tahun 2019, peningkatan total Curah Hujan Tahunan (RR) sedikit berhubungan dengan peningkatan Tingkat Insidensi (IR). Sementara garis merah turun tajam ke bawah (tren 2020) dengan y = -9.90x + 124.23 menunjukkan bahwa pada tahun 2020, peningkatan total Curah Hujan Tahunan (RR) sangat berhubungan dengan penurunan Tingkat Insidensi (IR), artinya korelasi antara RR dan IR tahun 2020 berupa korelasi negatif yang kuat.

Curah hujan juga berperan dalam perkembangan nyamuk Aedes. Curah hujan yang tinggi menciptakan genangan air yang dapat digunakan nyamuk sebagai tempat perindukan. Banyaknya genangan air dan tempat untuk bertelur, memudahkan nyamuk berkembangbiak. Provinsi dengan total curah hujan tahunan yang lebih tinggi cenderung memiliki IR yang lebih rendah, kemungkinan karena efek *flushing* yang membersihkan tempat perkembangbiakan nyamuk. Benedum et

al (2018) juga menyebut fenomena di mana curah hujan lebat yang membanjiri tempat perkembangbiakan nyamuk sehingga meluap dan menyebabkan larva hanyut dan mati sebagai fenomena *flushing*. Sebaliknya, IR tertinggi pada 2020 terkonsentrasi di provinsi dengan RR rendah hingga moderat (RR ≈ 3 hingga 5 mm), yang ditandai oleh lonjakan ekstrem pada Bali. Hal ini konsisten dengan penelitian Wulandari et al. (2023) yang juga menunjukan tidak ada hubungan antara curah hujan dengan IR DBD. Hasil penelitian ini juga memiliki kesesuaian dengan penelitian Khairinnisa et al. (2025) yang juga menyatakan tidak ditemukan asosiasi signifikan antara curah hujan dengan kasus DBD di Provinsi Bengkulu.

4.2.2.6 Uji Normalitas Shapiro-Wilk

	p-value
Tavg	1.429509e-78
RH_avg	4.818778e-76
RR	1.497819e-80
ss	1.411423e-38
ff_avg	6.679958e-52
incidence rate per 100.000 penduduk 2019	3.998841e-64
incidence rate per 100.000 penduduk 2020	2.949753e-69
jumlah kasus 2019	3.007503e-70
jumlah kasus 2020	1.352320e-70
kepadatan 2019	2.819699e-90
kepadatan 2020	2.563162e-90

Gambar hasil uji normalitas Shapiro Wilk

Berdasarkan uji normalitas Shapiro–Wilk yang dilakukan terhadap variabel-variabel penelitian tahun 2019–2020, seluruh variabel memiliki nilai p-value < 0,05. Hal ini menunjukkan bahwa distribusi data tidak normal pada semua variabel, baik variabel kasus DBD, faktor iklim (Tavg, RH_avg, RR, ss, ff_avg), maupun variabel kepadatan penduduk. Dengan demikian, analisis korelasi dalam penelitian ini tidak dapat menggunakan korelasi Pearson yang mensyaratkan distribusi normal. Oleh karena itu, uji korelasi yang digunakan adalah Korelasi Spearman, yang lebih sesuai untuk data non-parametrik.

4.2.2.7 Korelasi Spearman

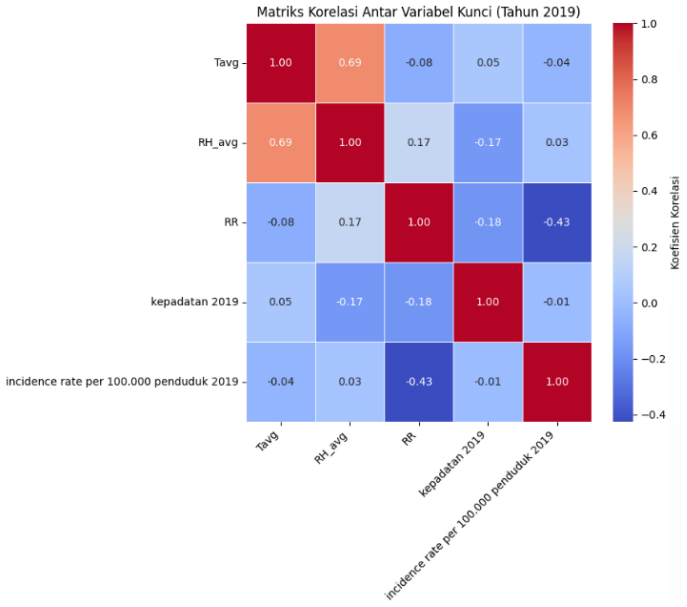
	jumlah kasus 2019	jumlah kasus 2020	Tavg	RH_avg	RR	ss	ff_avg	incidence rate per 100.000 penduduk 2019	incidence rate per 100.000 penduduk 2020	kepadatan 2019	kepadatan 2020
jumlah kasus 2019	1.000000	0.909809	0.074894	-0.141178	-0.079267	0.075812	-0.007214	0.102067	0.218070	0.751434	0.754502
jumlah kasus 2020	0.909809	1.000000	0.041950	-0.192439	-0.114730	0.132908	0.081753	0.214982	0.490830	0.839426	0.843808
Tavg	0.074894	0.041950	1.000000	-0.083637	-0.212328	0.193395	0.126276	0.001447	-0.004951	0.051646	0.051281
RH_avg	-0.141178	-0.192439	-0.083637	1.000000	0.375707	-0.268461	-0.294609	-0.001904	-0.096730	-0.178664	-0.179680
RR	-0.079267	-0.114730	-0.212328	0.375707	1.000000	-0.369877	-0.192577	-0.049592	-0.095831	-0.101307	-0.101375
ss	0.075812	0.132908	0.193395	-0.268461	-0.369877	1.000000	0.199927	0.087534	0.145841	0.099491	0.098982
ff_avg	-0.007214	0.081753	0.126276	-0.294609	-0.192577	0.199927	1.000000	-0.016843	0.083797	0.076226	0.074795
incidence rate per 100.000 penduduk 2019	0.102067	0.214982	0.001447	-0.001904	-0.049592	0.087534	-0.016843	1.000000	0.813259	0.033233	0.024154
incidence rate per 100.000 penduduk 2020	0.218070	0.490830	-0.004951	-0.096730	-0.095831	0.145841	0.083797	0.813259	1.000000	0.306827	0.303471
kepadatan 2019	0.751434	0.839426	0.051646	-0.178664	-0.101307	0.099491	0.076226	0.033233	0.306827	1.000000	0.999309
kepadatan 2020	0.754502	0.843808	0.051281	-0.179680	-0.101375	0.098982	0.074795	0.024154	0.303471	0.999309	1.000000

Gambar hasil korelasi Spearman

Korelasi Spearman adalah pilihan yang tepat untuk menganalisis data ini karena sifat hubungan antara iklim dan penyakit vektor yang non-linier. Matriks korelasi menunjukkan nilai koefisien korelasi Spearman (ρ) antara berbagai variabel (iklim, kasus, dan kepadatan) untuk tahun 2019 dan 2020. Nilai ρ berkisar dari -1 (korelasi negatif sempurna) hingga +1 (korelasi positif sempurna). Spearman bekerja berdasarkan peringkat data alih-alih nilai mentah, lebih tahan terhadap *outlier* ekstrem (seperti kasus Bali pada suhu rendah) dan tidak memerlukan asumsi distribusi normal, sehingga memberikan ukuran asosiasi yang lebih andal dan realistis untuk data penduduk dan iklim.

Analisis matriks korelasi Spearman mengonfirmasi bahwa faktor pendorong utama Jumlah Kasus DBD Absolut adalah Kepadatan Penduduk, ditunjukkan oleh korelasi positif yang sangat kuat ($\rho \approx 0.84$ pada tahun 2020). Artinya, provinsi dengan populasi lebih padat memiliki jumlah kasus DBD yang jauh lebih tinggi. Namun, ketika kasus diukur sebagai Insidence Rate (kasus per 100.000 penduduk), korelasi dengan kepadatan menurun drastis menjadi hanya $\rho \approx 0.30$. Ini menunjukkan bahwa kepadatan penduduk, meskipun memicu total kasus yang tinggi, bukanlah satu-satunya penentu tingkat risiko penularan per kapita. Sementara itu, korelasi antara semua variabel iklim (Tavg, RH_avg, RR, ss, ff_avg) dengan Incidence Rate DBD terbukti sangat lemah (ρ sebagian besar mendekati nol atau di bawah |0.20|), menguatkan temuan dari analisis *scatter plot* bahwa tidak ada hubungan linier yang signifikan.

4.2.2.8 Heatmap Matriks Korelasi Tahun 2019

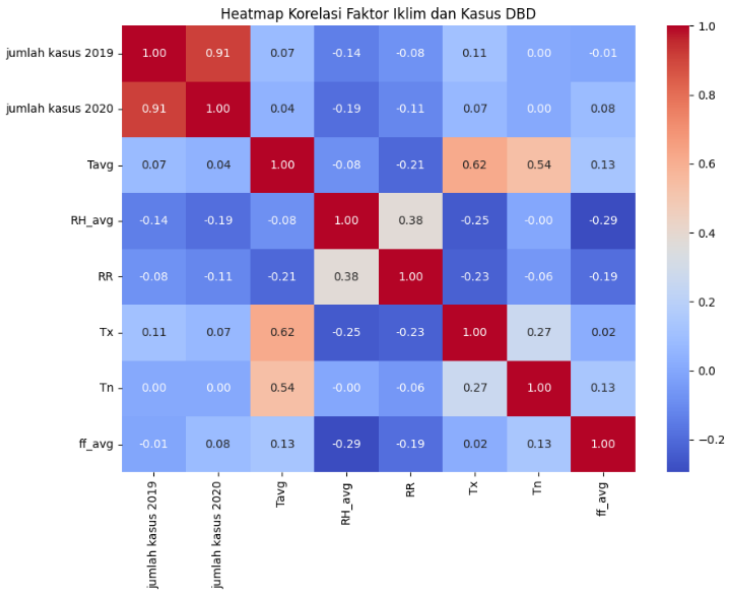


Gambar heatmap_corr_2019.png

Matriks ini menampilkan koefisien korelasi Pearson antara lima variabel kunci: Suhu Rata-rata (Tavg), Kelembaban Rata-rata (RH_avg), Curah Hujan (RR), Kepadatan Penduduk (kepadatan 2019), dan Tingkat Insidensi DBD (incidence rate per 100.000 penduduk 2019).

Analisis korelasi Pearson tahun 2019 menunjukkan bahwa hubungan antar variabel iklim bersifat logis, dengan Suhu Rata-rata (Tavg) dan Kelembaban Rata-rata (RH_avg) memiliki korelasi positif kuat (0.69). Satu-satunya variabel yang menunjukkan hubungan linier sedang dengan Incidence Rate DBD per 100.000 penduduk adalah Curah Hujan Tahunan (RR), dengan korelasi negatif sedang-kuat (-0.43). Korelasi negatif ini mengindikasikan bahwa pada tahun 2019, provinsi dengan curah hujan total yang lebih tinggi cenderung memiliki risiko DBD per kapita yang lebih rendah, mendukung hipotesis efek *flushing* oleh hujan deras. Sebaliknya, Tavg, RH_avg, RR, dan Kepadatan Penduduk memiliki korelasi sangat lemah yang signifikan (semua di bawah 0.05) dengan *incidence rate* DBD, menunjukkan bahwa faktor-faktor ini tidak dapat memprediksi risiko DBD per kapita secara linier.

4.2.2.9 Heatmap Antar Variabel Iklim vs Jumlah Kasus DBD

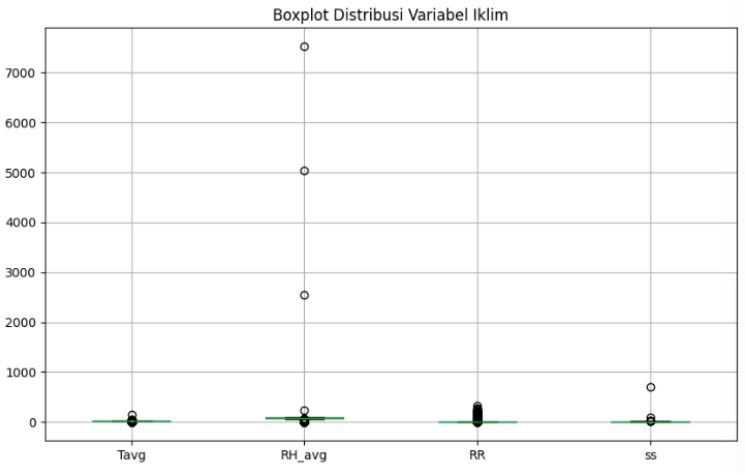


Gambar heatmap_corr_iklim_dbd.png

Heatmap tersebut menggambarkan hubungan korelasi antara berbagai faktor iklim (seperti suhu rata-rata, suhu maksimum, suhu minimum, kelembapan relatif, curah hujan, dan kecepatan angin) dengan jumlah kasus DBD pada tahun 2019 dan 2020. Secara umum, nilai korelasi antara variabel iklim dan kasus DBD cenderung

rendah, yang menunjukkan bahwa tidak ada faktor iklim tunggal yang memiliki hubungan kuat terhadap variasi jumlah kasus DBD antarprovinsi. Korelasi antara jumlah kasus DBD 2019 dan 2020 sangat tinggi (r = 0.91), menandakan bahwa provinsi yang memiliki kasus tinggi pada 2019 cenderung tetap tinggi pada 2020—kemungkinan dipengaruhi oleh faktor populasi, kepadatan penduduk, lingkungan, dan pengendalian vektor, bukan semata faktor iklim. Dari variabel iklim, suhu rata-rata (Tavg) dan suhu maksimum (Tx) menunjukkan korelasi positif sedang dengan kasus DBD menandakan bahwa daerah bersuhu lebih hangat sedikit cenderung memiliki kasus lebih tinggi, meskipun pengaruhnya tetap lemah. Sebaliknya, kelembapan relatif (RH_avg) memiliki korelasi rendah hingga negatif (r = –0.14 dan r = –0.19), dan curah hujan (RR) juga menunjukkan hubungan lemah (r = –0.08 dan r = –0.11). Suhu minimum (Tn) serta kecepatan angin (ff_avg) juga tidak menunjukkan hubungan berarti. Secara keseluruhan, heatmap ini memperlihatkan bahwa faktor iklim hanya memiliki pengaruh kecil terhadap variasi kasus DBD, dan penyebab utama perbedaan antarprovinsi kemungkinan lebih terkait dengan faktor non-iklim seperti demografi, sanitasi, mobilitas penduduk, dan efektivitas program pengendalian nyamuk.

4.2.2.10 Boxplot Faktor Iklim

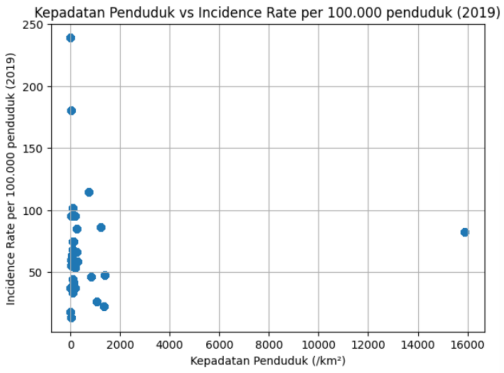


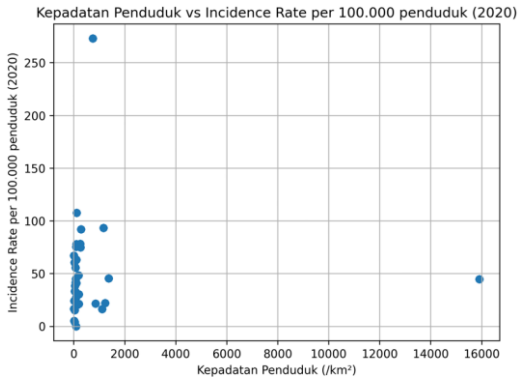
Gambar boxplot_iklim.png

Boxplot distribusi variabel iklim menunjukkan karakteristik penyebaran empat variabel utama, yaitu suhu rata-rata harian (Tavg), kelembapan relatif rata-rata (RH_avg), curah hujan (RR), dan lama penyinaran matahari (ss). Secara umum, grafik ini memperlihatkan bahwa sebagian besar nilai iklim berada dalam rentang yang relatif sempit, tetapi terdapat banyak outlier ekstrem pada variabel tertentu.

Variabel RH_avg dan ss menunjukkan outlier yang sangat tinggi, mengindikasikan bahwa beberapa provinsi memiliki kondisi kelembapan atau penyinaran yang jauh berbeda dari mayoritas wilayah lain. Variabel RR (curah hujan) juga menampilkan outlier besar, yang mencerminkan adanya provinsi dengan curah hujan ekstrem, khas daerah beriklim sangat basah. Sebaliknya, variabel Tavg tampak lebih stabil dengan sedikit outlier dan rentang antar kuartil yang sempit, menunjukkan bahwa suhu rata-rata relatif konsisten antarprovinsi di Indonesia. Pola ini menegaskan bahwa variabel iklim selain suhu cenderung memiliki variasi spasial yang besar, terutama curah hujan, kelembapan, dan durasi penyinaran, sehingga perbedaan geografis antarprovinsi sangat memengaruhi karakteristik iklim yang tercatat.

4.2.2.11 Scatter Plot Kepadatan Penduduk vs Incidence Rate per 100.000 Penduduk





Gambar scatter_kepadatan_IR_2019.png Gambar scatter_kepadatan_IR_2020.png

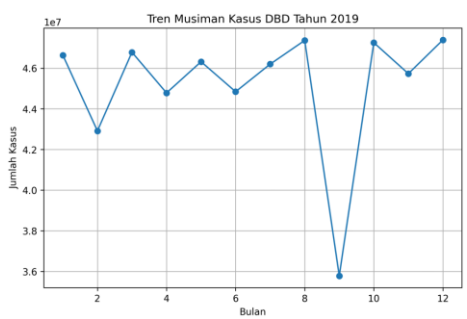
Scatter plot Kepadatan Penduduk vs. *Incidence Rate* (IR) tahun 2019 menunjukkan tidak adanya korelasi linier positif yang signifikan antara kepadatan dan risiko DBD per kapita, yang memperkuat temuan korelasi sebelumnya. Sebagian besar provinsi mengelompok pada kepadatan rendah, tetapi IR mereka sangat bervariasi, mencapai puncaknya di sekitar 250 per 100.000 penduduk di provinsi yang tidak padat. Sebaliknya, provinsi terpadat (seperti DKI Jakarta, ≈16.000 jiwa/km^2) mempertahankan IR yang relatif rendah (<100), sementara penyumbang kasus absolut terbesar (Jawa Barat) berada pada IR yang moderat. Hal ini menyiratkan bahwa pada tahun 2019, risiko penularan DBD yang sesungguhnya di tingkat provinsi lebih dikendalikan oleh efektivitas intervensi kesehatan dan kondisi lingkungan mikro daripada kepadatan populasi secara keseluruhan.

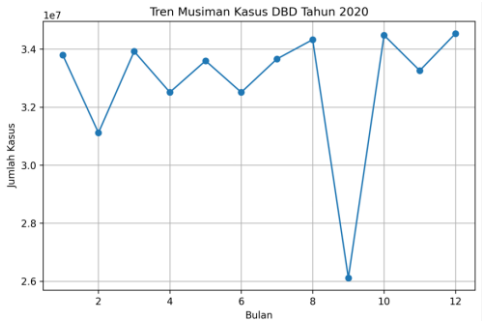
Scatter plot Kepadatan Penduduk vs. *Incidence Rate* (IR) tahun 2020 menunjukkan bahwa risiko DBD per kapita (Incidence Rate) pada tahun 2020 tidak

berkorelasi positif kuat dengan Kepadatan Penduduk. Sebagian besar provinsi mengelompok pada kepadatan rendah (di bawah 2.000 jiwa/km^2) dengan variabilitas IR yang sangat tinggi (mencapai lebih dari 250 per 100.000 penduduk), yang menunjukkan bahwa faktor iklim ekstrem atau mikrolokal adalah pemicu risiko yang lebih dominan. Fenomena ini diperkuat oleh dua anomali: provinsi dengan kepadatan tertinggi (DKI Jakarta, ≈15.907≈16.000 jiwa/km^2) mencatat IR yang relatif rendah (<50) menggarisbawahi efektivitas intervensi kesehatan masyarakat, sementara provinsi dengan IR tertinggi (kemungkinan Bali) berada pada kepadatan rendah, membuktikan bahwa kepadatan hanyalah pemicu Total Kasus Absolut, tetapi bukan penentu utama risiko penularan per kapita.

Hasil yang tidak signifikan antara kepadatan penduduk dengan IR DBD didukung studi oleh Istiqamah et al. (2020) berlawanan dengan hasil di atas, yakni tidak terdapat hubungan signifikan antara kepadatan penduduk dengan kejadian DBD di Kota Kendari pada tahun 2014-2018. Studi yang dilakukan Sajib et al. (2024) pada 11 negara di Asia menunjukan korelasi negatif antara kepadatan penduduk dengan kejadian DBD. Studi menunjukan bahwa wilayah dengan kepadatan penduduk yang lebih tinggi belum tentu mengalami wabah demam berdarah yang lebih parah. Hal ini dapat terjadi jika masyarakat memiliki kepedulian dan kebersihan terhadap lingkungan untuk melakukan pencegahan dan pengendalian vektor nyamuk Aedes. Oleh karena itu, dibutuhkan kolaborasi antara *stakeholder* terkait dan masyarakat untuk menguatkan program pencegahan dan pengendalian Demam Berdarah Dengue yang tetap memperhatikan wilayah yang lebih berisiko. Edukasi kepada masyarakat juga diperlukan secara lebih masif tentang pentingnya praktik pemberantasan sarang nyamuk pada wadah-wadah bekas yang dapat menciptakan genangan air guna mengeliminasi tempat perkembangbiakan nyamuk.

4.2.2.12 Grafik Tren Musiman Kasus DBD per Bulan



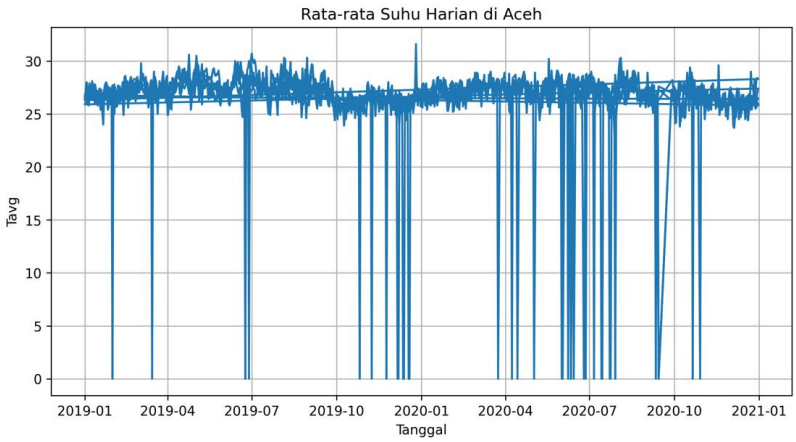


Gambar tren musiman kasus dbd 2019.png Gambar tren musiman kasus dbd 2020.png

Tren musiman kasus DBD pada tahun 2019 dan 2020 menunjukkan pola endemisitas yang konsisten, ditandai oleh puncak kasus yang bertepatan dengan musim hujan. Pada kedua tahun, kasus mencapai puncak utama di awal tahun (Bulan 1 hingga Bulan 3), yang ideal untuk perkembangbiakan nyamuk *Aedes aegypti* karena ketersediaan air dan kelembaban tinggi. Setelah itu, kasus mengalami penurunan signifikan menuju titik terendah tahunan di sekitar Bulan 9 (September), periode yang biasanya menandai puncak musim kemarau di banyak wilayah. Menurut BMKG, puncak musim kemarau di sebagian besar wilayah Indonesia (tergantung Zona Musim (ZOM)), terutama yang dipengaruhi oleh angin monsun Australia, diprediksi terjadi pada periode Agustus hingga September. Periode ini ditandai oleh pergerakan angin yang membawa udara kering dari Benua Australia, mengakibatkan curah hujan minimal dan penurunan kelembaban, yang sesuai dengan data kasus DBD terendah di Bulan 9. Kemudian, kasus kembali menunjukkan kenaikan tajam di akhir tahun (Bulan 10 hingga 12) seiring kembalinya musim hujan. Pola ini menegaskan bahwa kondisi iklim musiman adalah pendorong utama penularan DBD pada skala waktu bulanan.

Meskipun memiliki pola serupa, tahun 2020 memiliki fluktuasi yang lebih tinggi di pertengahan tahun (sekitar Bulan 6 hingga 8) dibandingkan 2019, yang mungkin terkait dengan anomali iklim yang terjadi, meskipun titik terendah di Bulan 9 tetap dominan. Perbedaan yang lebih signifikan terletak pada dampaknya terhadap risiko per kapita di provinsi-provinsi tertentu. Meskipun tren musiman secara nasional konsisten, anomali iklim ekstrem di awal 2020, seperti lonjakan kelembaban, dapat secara dramatis meningkatkan kerentanan di wilayah tertentu, bahkan jika total kasus nasional mengikuti pola musiman yang sama.

4.2.2.13 *Time Series* Rata-Rata Suhu Harian di Provinsi Tertentu



Gambar rata-rata suhu harian di {prov}.png

Grafik time series Rata-rata Suhu Harian (Tavg) di Provinsi Aceh menunjukkan suhu yang sangat stabil dari tahun 2019 hingga 2021, dengan sebagian besar Tavg berfluktuasi ketat antara 26°C hingga 28°C yang merupakan rentang suhu optimal untuk siklus hidup nyamuk *Aedes aegypti*. Meskipun suhunya ideal bagi vektor, Aceh mempertahankan jumlah kasus DBD yang rendah secara absolut pada kedua tahun, yang menguatkan temuan dari analisis korelasi bahwa Tavg bukanlah penentu utama risiko DBD per kapita. Pola visual grafik didominasi oleh garis vertikal tajam yang menurun hingga mendekati nol, yang kemungkinan besar mengindikasikan kesalahan sensor atau data hilang (*noise*) dan bukan anomali iklim yang sesungguhnya. Kurangnya pola musiman yang jelas pada Tavg di Aceh, dikombinasikan dengan kasus DBD yang rendah, menunjukkan bahwa faktor-faktor seperti sanitasi yang baik, program pengendalian vektor yang efektif, atau curah hujan yang sangat tinggi (efek *flushing*) lebih dominan dalam mencegah wabah DBD dibandingkan variabel suhu rata-rata yang stabil.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil analisis data, faktor iklim seperti suhu rata-rata provinsi, curah hujan dan kelembaban tidak menunjukkan hubungan yang kuat dan menunjukkan pengaruh linier yang tidak signifikan terhadap jumlah kasus DBD dan incidence rate selama tahun 2019-2020. Pola perubahan cuaca terlihat bervariasi antar provinsi tetapi hal tersebut tidak terbukti berkaitan dengan naik turunnya kasus DBD. Hal yang sama juga ditemukan pada variabel kepadatan penduduk. Meskipun beberapa provinsi dengan penduduk lebih padat memiliki jumlah kasus yang tinggi, tetapi analisisnya tidak menunjukkan adanya hubungan yang konsisten ataupun cukup kuat untuk menyimpulkan adanya pengaruh yang jelas dari kepadatan penduduk terhadap kasus DBD. kepadatan penduduk hanya berkorelasi kuat dengan jumlah kasus absolut, tetapi tidak terhadap risiko per 100.000 penduduk. Temuan ini mengindikasikan bahwa faktor-faktor makro, seperti iklim dan kepadatan penduduk tidak cukup sensitif untuk menjelaskan variasi angka kejadian DBD pada tingkat provinsi, yang kemungkinan lebih dipengaruhi oleh variabel mikro, seperti kebersihan lingkungan, perilaku masyarakat, keberadaan tempat perkembangbiakan nyamuk, serta efektivitas program intervensi kesehatan.

Secara Umum, hasil ini menunjukkan bahwa penyebaran kasus DBD tingkat provinsi tidak dapat dijelaskan hanya dengan faktor iklim dan kepadatan penduduk. Ada kemungkinan faktor lain seperti perilaku masyarakat, kebersihan lingkungan, program pemberantasan sarang nyamuk (PSN), dan kapasitas kesehatan daerah bisa menjadi penentu yang lebih besar dalam variasi kasus DBD antar provinsi.

5.2 Saran

Hasil penelitian tidak menemukan pengaruh signifikan dari faktor iklim maupun kepadatan penduduk, penelitian selanjutnya disarankan untuk memasukkan variabel lain yang lebih berhubungan dengan aktivitas nyamuk dan perilaku masyarakat, misalnya tingkat kebersihan lingkungan, indeks jentik, atau intensitas program pengendalian vektor (misalnya fogging dan PSN). Selain itu, disarankan menggunakan data yang lebih spesifik seperti tingkat kota/kabupaten atau kecamatan agar variasi lokal yang mungkin tersebar dapat teridentifikasi lebih jelas dan akurat.

KENDALA DAN RENCANA TINDAK LANJUT

Kendala

Terdapat beberapa kendala yang dihadapi dalam proses penyusunan penelitian, antara lain sebagai berikut:

- Terjadi perubahan judul dan topik penelitian karena penyesuaian terhadap ketersediaan dataset yang sesuai.
- Mengalami kesulitan dalam memperoleh dataset yang relevan dengan kajian dan kebutuhan analisis yang kemudian diperoleh berdasarkan hasil diskusi bersama.
- Menghadapi hambatan dalam menjalankan program pada Google Colab akibat waktu pengerjaan yang dilakukan secara bersamaan.
- Mengalami tantangan dalam komunikasi jarak jauh yang berpotensi memperlambat kelancaran koordinasi dan penyelesaian tugas.

Meskipun berbagai kendala tersebut muncul dalam proses pengerjaan, seluruh tantangan dapat diatasi melalui koordinasi yang berkelanjutan, penyesuaian strategi kerja, serta komitmen untuk menyelesaikan penelitian secara optimal.

Rencana Tindak Lanjut

Setelah proses wrangling selesai, langkah tindak lanjut yang direncanakan adalah melakukan eksplorasi data lanjutan serta menyiapkan struktur dataset agar dapat digunakan untuk tahap analisis statistik berikutnya.

- Identifikasi dan penanganan data noise pada Tavg dan RH_avg. Ganti nilai tersebut dengan metode Imputasi Rolling Mean atau Imputasi Mean Bulanan dari provinsi terkait untuk menjaga stabilitas data time series.
- Metode clustering, seperti K-Means untuk mengelompokkan provinsi berdasarkan risiko DBD. Metode klasifikasi menggunakan decision tree atau random forest untuk memprediksi kategori risiko. Analisis regresi untuk memodelkan pengaruh variabel iklim terhadap *incidence rate*. Metode-metode ini relatif mudah diterapkan namun mampu memberikan wawasan lebih mendalam setelah tahap eksplorasi data.
- Melakukan Normalisasi variabel kasus (MinMax Scaling atau Z-Score) pada variabel Jumlah Kasus dan Kepadatan Penduduk untuk membatasi dampak *outlier* pada pemodelan.

- Feature Engineering Lanjutan: Oleh karena korelasi linier dengan Tavg rata-rata terbukti sangat lemah. RH_avg dan Curah Hujan (RR) menunjukkan hubungan non-linier yang lebih kuat, sehingga dapat dihitung variabel baru yang lebih sensitif
    1) Diurnal Temperature Range (DTR): Tmax – Tmin (untuk menangkap variabilitas harian)
    2) RH_avg Kuadratik (RH_avg^2) untuk menangkap hubungan non-linier.
- Validasi Hubungan Non-Linier

  Mengkonfirmasi visualisasi *scatter plot* bahwa hubungan antara iklim dan IR berbentuk kurva ('U' terbalik). Misalnya, pada scatter plot IR vs. Tavg dan IR vs. RR dengan penambahan garis regresi polinomial (orde 2 atau 3).
- Analisis Time-Lag

  Mengatasi kelemahan korelasi instan dan menemukan penundaan waktu antara faktor iklim dan munculnya kasus DBD (siklus hidup nyamuk membutuhkan ≈1-3 bulan. Dapat diterapkan dengan Korelasi Silang (Cross-Correlation Function) antara RH_avg bulanan dengan Kasus DBD bulanan, menguji *lag* (penundaan) 1 hingga 4 bulan.

DAFTAR PUSTAKA

Greegtitan. (2025). Indonesia climate dataset. Kaggle. Diakses pada tanggal 15 November 2025, dari https://www.kaggle.com/datasets/greegtitan/indonesia-climate?select=province_detail.csv

Badan Pusat Statistik Indonesia. (2023). Kepadatan Penduduk menurut Provinsi, 2019. Diakses pada 15 November 2025, dari https://www.bps.go.id/id/statistics-table/2/MTQxIzI=/population-density-by-province.html

Badan Pusat Statistik Indonesia. (2023). Kepadatan Penduduk menurut Provinsi, 2020. Diakses pada 15 November 2025, dari https://www.bps.go.id/id/statistics-table/2/MTQxIzI=/population-density-by-province.html

Kementerian Kesehatan Republik Indonesia. (2020). Profil Kesehatan Indonesia tahun 2019. Kementerian Kesehatan RI. Profil Kesehatan Indonesia 2019

Kementerian Kesehatan Republik Indonesia. (2021). *Profil Kesehatan Indonesia tahun 2020*. Kementerian Kesehatan RI. Profil Kesehatan Indonesia 2020

Sukristi, S. F. (2025). Analisis faktor iklim, kepadatan penduduk, dan angka bebas jentik (ABJ) dengan incidence rate demam berdarah dengue (DBD) di Kota Bogor tahun 2020-2024 = Analysis of climate factor, population density, and larvae-free rate with incidence rate of dengue hemorrhagic fever in Bogor City 2020–2024 [Skripsi, Universitas Indonesia]. Universitas Indonesia. https://lontar.ui.ac.id/detail?id=9999920572203

Wulandari RA, Rahmawati T, Asyary A, Nugraha F. Analysis of Climate and Environmental Risk Factors on Dengue Hemorrhagic Fever Incidence in Bogor District. Kesmas. 2023;18(3):209–14.

Khairinnisa K, Fauzi Y, Nugraheni E, Demam K, Dengue B, Tahun DBD, et al. Analisis Spasio-Temporal Kondisi Iklim dan Jumlah Kejadian Demam Berdarah Dengue (DBD) Tahun 2012-2021 di Bengkulu. 2025;24(November 2024):136 44.

Benedum CM, Seidahmed OME, Eltahir EAB, Markuzon N. Statistical modeling of the effect of rainfall flushing on dengue transmission in Singapore. Reiner RC, editor. PLoS Negl Trop Dis [Internet]. 2018 Dec 6;12(12):e0006935. Available from: https://dx.plos.org/10.1371/journal.pntd.0006935

Rachmatullah F. Analisis Spasial Faktor Iklim dan Kepadatan Penduduk Dengan Kejadian DBD di DKI Jakarta, Kota Bogor, dan Depok Tahun 2015-2018. Skripsi. Universitas Indonesia; 2019.

Rozilawati H, Zairi J, Adanan CR. Seasonal abundance of Aedes albopictus in selected urban and suburban areas in Penang, Malaysia. Trop Biomed. 2007;24(1):83–94.

Istiqamah SNA, Arsin AA, Salmah AU, Mallongi A. Correlation Study between Elevation, Population Density, and Dengue Hemorrhagic Fever in Kendari City in 2014–2018. Open Access Maced J Med Sci [Internet]. 2020 Jul 23;8(T2):63–6. Available from: https://oamjms.eu/index.php/mjms/article/view/5187

Sajib AH, Akter S, Saha G, Hossain Z. Demographic-environmental effect on dengue outbreaks in 11 countries. Colborn J, editor. PLoS One [Internet]. 2024 Sep 11;19(9):e0305854. Available https://dx.plos.org/10.1371/journal.pone.0305854

Kontribusi

Nama Anggota	Kontribusi
Siti Fadilah Nurkhotimah	Sintaks (Google Colab): - Berkontribusi dalam pengolahan dataset kasus DBD berformat PDF yang diambil dari Publikasi tahunan Profil Kesehatan Indonesia tahun 2019 dan 2020 dari Kementerian Kesehatan, yang meliputi proses pengambilan data, pembacaan file PDF, scraping atau ekstraksi tabel dari PDF, pembersihan data, integrasi dataset, serta penyusunan dan interpretasi eksplorasi data (EDA). Laporan: - Mengerjakan bagian BAB II Tinjauan Pustaka. - Mengerjakan bagian BAB III Metodologi Penelitian untuk subbab Sumber Data, Teknik Pengambilan Data, dan Eksplorasi Data. - Berkontribusi dalam penyusunan bagian pengambilan data serta pembuatan analisis eksploratif (EDA) secara keseluruhan untuk ketiga dataset (Iklim, Kepadatan Penduduk, dan Kasus DBD). - Berkontribusi dalam mengelola dan mengunggah data (raw data) dan hasil wrangling (berupa file, gambar, dan Google Colab) di GitHub. - Untuk kesimpulan, saran, kendala, dan rencana tindak lanjut dikerjakan bersama-sama.
Laili Nurrohmatul Fadhila Zulfa	Sintaks (Google Colab): - Mengolah dataset Iklim dan Kepadatan Penduduk tahun 2019 dan 2020 dari read file csv, cleaning data, integrasi dataset iklim dan provinsinya, integrasi kepadatan penduduk 2019 dan 2020, Integrasi ketiga dataset final. Laporan: - BAB 1 PENDAHULUAN - METODOLOGI PENELITIAN Bagian cleaning data, integrasi data, publishing data berupa raw data, cleaning data, integrasi data, dokumen pipeline (flowchart). - HASIL DAN PEMBAHASAN Bagian cleaning data dan integrasi data. - KESIMPULAN DAN SARAN bersama