

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

Laporan Akhir Revisi ini Disusun untuk Memenuhi Project Akhir Semester
Mata Kuliah Data Wrangling



Disusun oleh:

Kelompok 9 – 2024 C

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

Laili Nurrohmatul Fadhila Zulfa (24031554093 – UNESA)

Dosen Pengampu:

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

Dr. Dian Handayani, M.Si.

**SAINS DATA, FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI SURABAYA
STATISTIKA, FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI JAKARTA**

2025

DAFTAR ISI

DAFTAR ISI	2
BAB I PENDAHULUAN	4
1.1 Latar Belakang	4
1.2 Rumusan Masalah	4
1.3 Tujuan Penelitian	5
1.4 Manfaat Penelitian	5
BAB II TINJAUAN PUSTAKA	6
2.1 Demam Berdarah Dengue (DBD)	6
2.2 Pengaruh Faktor Iklim terhadap DBD	6
2.3 Kepadatan Penduduk dan Kaitannya dengan Penyebaran DBD	6
2.4 Konsep Data Wrangling	7
BAB III METODOLOGI PENELITIAN	8
3.1 Sumber Data	8
3.2 Teknik Pengambilan Data	8
3.2.1 Dataset Kasus Demam Berdarah Dengue (DBD)	8
3.2.2 Dataset Iklim Indonesia	9
3.2.3 Dataset Kepadatan Penduduk menurut Provinsi (jiwa/km²)	9
3.3 Teknik Integrasi Data	9
3.3.1 Penyamaan Nama Provinsi	9
3.3.2 Standarisasi Nama Kolom	9
3.3.3 Penggabungan Dataset	10
3.4 Data Cleaning	10
3.5 Transformasi data	11
3.6 Eksplorasi Data	11
3.6.1 Analisis Univariat	12
3.6.2 Analisis Bivariat	12
3.7 Data Publishing	12
BAB IV HASIL DAN PEMBAHASAN	34
4.1 Hasil Proses Wrangling	34
4.1.1 Pengambilan Data	34

4.1.2	<i>Cleaning Data</i>	36
4.1.3	Integrasi Data	37
4.1.4	Transformasi Data	39
4.2	Hasil Eksplorasi Data (EDA)	40
4.2.1	Statistik Deskriptif	40
4.2.2	Analisis Eksploratif Visual (Visualisasi EDA)	43
BAB V KESIMPULAN DAN SARAN		50
5.1	Kesimpulan	50
5.2	Saran	50
KENDALA DAN RENCANA TINDAK LANJUT		51
DAFTAR PUSTAKA		52
Kontribusi		54

BAB I

PENDAHULUAN

1.1 Latar Belakang

Demam 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/km² 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) menunjukkan bahwa puncak kasus DBD sering terjadi setelah periode peningkatan 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), agregasi, 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 tiga 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/km²) 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/km²)

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/km²) 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 integrasi integrasi data yang digunakan:

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 100.000 Penduduk”, “meninggal”, “CFR (%)” menjadi “provinsi”, “jumlah penduduk 2019”, “jumlah kasus 2019”, “incidence rate per 100.000 penduduk 2019”, “meninggal 2019”, “CFR (%) 2019”, begitupun untuk dataset 2020.

Terakhir, setelah penggabungan dua dataset dbd kolom “jumlah_penduduk”, “jumlah_kasus”, “incidence rate”, meninggal”, “cfr” dilakukan penyederhanaan nama menjadi “JP”, “JK”, “IR”, “M”, “CFR”.

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 `province_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 provinsi sehingga menghasilkan 3 kolom yaitu provinsi, kepadatan penduduk 2019, dan kepadatan penduduk 2020.

3. Penggabungan Dataset DBD

Sama seperti dataset 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 kepadatannya. Cleaning pertama sampai terakhir juga dilakukan untuk dataset kepadatan penduduk 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 Transformasi data

Transformasi data dilakukan untuk menyesuaikan bentuk data agar dapat dianalisis secara bersamaan. Pada dataset iklim bentuk data berupa data harian sedangkan dataset kependudukan dan kasus dbd berbentuk data tahunan. Maka dari itu dataset iklim perlu diseragamkan agar dapat dianalisis. Proses transformasi dimulai dengan mengelompokkan data iklim harian berdasarkan provinsi dan tahun untuk periode 2019-2020. Selanjutnya, menghitung rata-rata tahunan untuk data suhu, kelembapan, dan curah hujan agar dapat dibandingkan dengan variabel lain pada skala yang sama.

3.6 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, seperti heatmap korelasi, bar

chart, ataupun scatterplot untuk melihat pola hubungan antara variabel. Visualisasi ini membantu mengidentifikasi kemungkinan variabel yang memiliki hubungan dengan peningkatan jumlah kasus DBD dan incidence rate per 100.000 penduduk.

3.6.1 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.

3.6.2 Analisis Bivariat

Uji bivariat dilakukan sebagai pendekatan analitik untuk mengevaluasi korelasi antara variabel independen dan variabel dependen. 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 menunjukkan bahwa hubungan memiliki arah positif, menandakan nilai suatu variabel berbanding lurus dengan nilai variabel lain. Sebaliknya nilai r negatif menunjukkan hubungan arah negatif, menandakan hubungan berlawanan, semakin besar nilai suatu variabel, semakin kecil nilai variabel lain.

Analisis bivariat pada penelitian ini menggunakan korelasi Spearman untuk melihat hubungan antara variabel iklim, kepadatan penduduk, dan risiko DBD. Metode ini dipilih karena data tidak sepenuhnya berdistribusi normal dan hubungan antarvariabel cenderung bersifat monotonik. Spearman juga lebih tahan terhadap outlier, sehingga lebih sesuai untuk variasi data antar-provinsi. Hasil korelasi ini menggambarkan kekuatan dan arah hubungan antarvariabel, yang kemudian menjadi dasar dalam menilai faktor-faktor yang berpotensi memengaruhi peningkatan DBD.

3.7 Data Publishing

Pada penelitian 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_REVISI: 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.

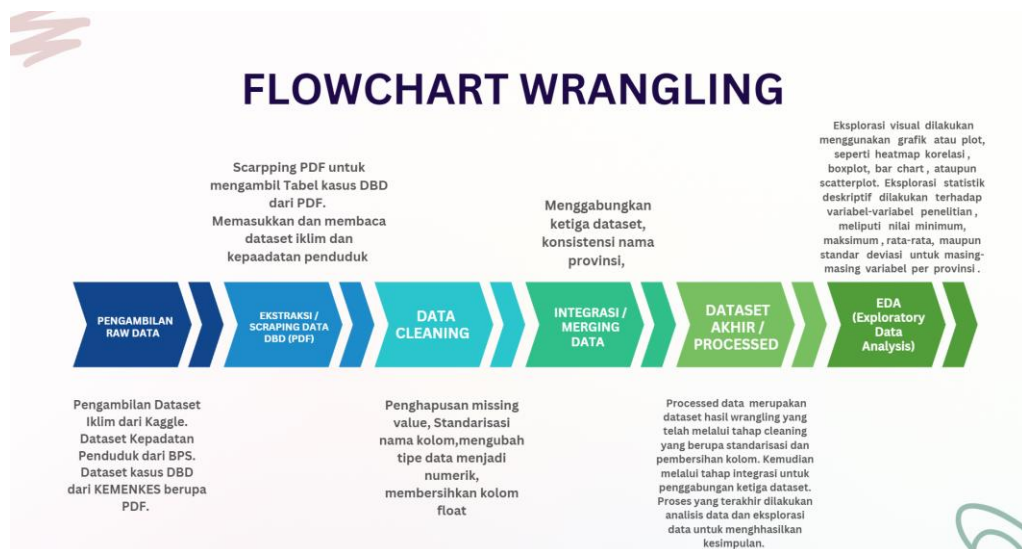
2. Processed Data

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 menghasilkan 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

```
#Unggah File PDF
# Install pustaka yang dibutuhkan
!pip install tabula-py pandas pyreadstat openpyxl PyMuPDF
```

```
#import modul
```

```
import pandas as pd
import os
import pyreadstat
import tabula
import fitz
```

```
#Read File PDF
```

```
# Profil-Kesehatan-Indonesia-2019.pdf
```

```
PDF_2019 = 'Profil-Kesehatan-Indonesia-2019.pdf'
PDF_PATH_2019 = f'/content/{PDF_2019}'
print(f"File PDF telah diunggah ke path")
print(f"Path PDF: {PDF_PATH_2019}")
doc1 = fitz.open(PDF_2019)
doc1
```

```
# Profil-Kesehatan-Indonesia-2020.pdf
```

```
PDF_2020 = 'Profil-Kesehatan-Indonesia-2020.pdf'
PDF_PATH_2020 = f'/content/{PDF_2020}'
print(f"File PDF telah diunggah ke path")
print(f"Path PDF: {PDF_PATH_2020}")
doc2 = fitz.open(PDF_2020)
doc2
```

```
#Dataset Iklim
```

```
import pandas as pd
```

```
# === LOAD DATA ===
```

```
df_climate = pd.read_csv("climate_data.csv")
df_station = pd.read_csv("station_detail.csv")
df_prov = pd.read_csv("province_detail.csv")
```

```
#Dataset Kepadatan Penduduk
```

```
import pandas as pd
```

```
df19 = pd.read_csv("Kepadatan Penduduk menurut Provinsi, 2019.csv")
df20 = pd.read_csv("Kepadatan Penduduk menurut Provinsi, 2020.csv")
```

b. Cleaning Data

○ Pembersihan Data Iklim (df_final)

```
# penggantian nama kolom
df_final = df_final.rename(columns={'province_name':
'provinsi', 'year' : "tahun"})
df_final
```

```
#Mengubah NaN menjadi angka 0
df_final = df_final.fillna(0)
df_final
```

○ Pembersihan Data Kepadatan Penduduk

- 2019

```
#Standarisasi Nama kolom
df19 = df19.rename(columns={
'Unnamed: 1': 'kepadatan 2019',
'38 Provinsi' : 'provinsi'
})
df19
```

```
# drop 2 baris NaN di atas (index 0 dan 1)
df19_1= df19.iloc[2:].reset_index(drop=True)
```

```
# Mengubah tipe kolom menjadi numerik
df19_1['kepadatan 2019'] =
pd.to_numeric(df19_1['kepadatan 2019'],
errors='coerce')
```

```
# hapus baris kosong
df19 = df19_1.dropna(subset=['provinsi', 'kepadatan
2019'])
df19
```

```
df19 = df19[df19['provinsi'] != 'INDONESIA'].copy()
df19
```

- 2020

```
# drop 2 baris NaN kepadatan penduduk 2020 (index 0
dan 1)
df20_1 = df20.iloc[2:].reset_index(drop=True)
```

```
df20_1

# rename kolom
df20_1 = df20_1.rename(columns={
    'Unnamed: 1': 'kepadatan 2020',
    '38 Provinsi': 'provinsi'})
df20_1

# Mengubah tipe kolom menjadi numerik
df20_1['kepadatan 2020'] =
pd.to_numeric(df20_1['kepadatan 2020'],
errors='coerce')

# Menghapus baris kosong
df20 = df20_1.dropna(subset=['provinsi', 'kepadatan
2020'])
df20

df20 = df20[df20['provinsi'] != 'INDONESIA'].copy()
df20
```

- o Pembersihan Data DBD

- 2019

```
# hapus baris index 0 dan 35 karena bukan data yang
digunakan dan bukan provinsi (melainkan Indonesia)
# hapus kolom No (Nomor)
df_dbd_2019 =
df_dbd_2019.drop(columns=["No"]).drop(index=[0,
35]).reset_index(drop=True)

# Informasi sebelum tipe data diubah
df_dbd_2019.info()

# Hapus titik pemisah ribuan & ubah koma ke titik
cols_int = ["Jumlah Penduduk", "Jumlah Kasus",
"Meninggal"]
cols_float = ["Incidence Rate per\n100.000 Penduduk",
"CFR (%)"]

# Bersihkan kolom integer
for col in cols_int:
    df_dbd_2019[col] =
df_dbd_2019[col].str.replace(".", "", regex=False)
    df_dbd_2019[col] =
pd.to_numeric(df_dbd_2019[col], errors="coerce")
```



```
# Bersihkan kolom float
for col in cols_float:
    df_dbd_2019[col] =
df_dbd_2019[col].str.replace(".", "", regex=False)
# hilangkan titik pemisah ribuan
df_dbd_2019[col] =
df_dbd_2019[col].str.replace(",", ".", regex=False) #
ganti koma menjadi titik
df_dbd_2019[col] =
pd.to_numeric(df_dbd_2019[col], errors="coerce")
```

- 2020

```
# hapus baris index 0 dan 35 karena bukan data yang
digunakan dan bukan provinsi
# hapus kolom No (Nomor)
df_dbd_2020 =
df_dbd_2020.drop(columns=["No"]).drop(index=[0,
35]).reset_index(drop=True)
```

```
# Hapus titik pemisah ribuan & ubah koma ke titik
cols_int = ["Jumlah Penduduk", "Jumlah Kasus",
"Meninggal"]
cols_float = ["Incidence Rate per\n100.000 Penduduk",
"CFR (%)"]
```

```
# Bersihkan kolom integer
for col in cols_int:
    df_dbd_2020[col] =
df_dbd_2020[col].str.replace(".", "", regex=False)
df_dbd_2020[col] =
pd.to_numeric(df_dbd_2020[col], errors="coerce")
```

```
# Bersihkan kolom float
for col in cols_float:
    df_dbd_2020[col] =
df_dbd_2020[col].str.replace(".", "", regex=False)
# hilangkan titik pemisah ribuan
df_dbd_2020[col] =
df_dbd_2020[col].str.replace(",", ".", regex=False) #
ganti koma menjadi titik
df_dbd_2020[col] =
pd.to_numeric(df_dbd_2020[col], errors="coerce")
```

c. Integrasi Data

○ Integrasi Menyamakan Nama Provinsi Dataset Iklim

Sebelum melakukan merge dilakukan standarisasi nama provinsi agar dapat di gabungkan.

```
# sesuaikan kolom Provinsi
df_prov["province_name"] =
df_prov["province_name"].str.title()

df_prov["province_name"] =
df_prov["province_name"].replace({"Nanggroe Aceh
Darussalam": "Aceh", "Dki Jakarta" : 'DKI Jakarta', 'Di
Yogyakarta' : 'DI Yogyakarta'})
df_prov['province_name']

df_climate.loc[:, 'station_id'] =
df_climate['station_id'].astype(int)
```

○ Integrasi Dataset Iklim

```
# MERGE climate_data.csv (df_climate) +
station_detail.csv (df_station)
df_merge1 = df_climate.merge(df_station[['station_id',
'province_id']], on='station_id', how='left')
df_merge1

# MERGE dengan province_detail.csv (df_prov)
df_final = df_merge1.merge(df_prov, on='province_id',
how='left')
df_final

# Memfilter tahun 2019-2020
df_final = df_final[df_final['year'].isin([2019,
2020])].copy()
df_final
```

○ 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”.

```
# Menggabungkan dataset 2019 dan 2020
```

```
df_kepduk = df19.merge(df20, on='provinsi', how='inner')
df_kepduk
```

```
# Mengubah nama kolom menjadi title
df_kepduk['provinsi'] =
df_kepduk['provinsi'].str.title()

df_kepduk["provinsi"] =
df_kepduk["provinsi"].replace({"Dki Jakarta" : 'DKI
Jakarta', 'Di Yogyakarta' : 'DI Yogyakarta'})
df_kepduk
```

```
df_kepduk = df_kepduk.melt(
    id_vars='provinsi',
    value_vars=['kepadatan 2019', 'kepadatan 2020'],
    var_name='tahun',
    value_name='KP'
)
df_kepduk
```

```
df_kepduk['tahun'] =
df_kepduk['tahun'].str.extract('(\d+)').astype(int)
df_kepduk
```

- Integrasi Dataset DBD tahun 2019

Sebelum melakukan merge dilakukan pergantian nama kolom dan standarisasi nama provinsi. Kemudian data hasil nya disimpan dengan nama “df_dbd_2019.csv”.

```
df_dbd_2019= df_dbd_2019.rename(columns={
    'Provinsi': 'provinsi',
    'Jumlah Penduduk' : 'jumlah penduduk 2019',
    'Jumlah Kasus' : 'jumlah kasus 2019',
    'Incidence Rate per\n100.000 Penduduk' : 'incidence
rate per 100.000 penduduk 2019',
    'Meninggal' : 'meninggal 2019',
    'CFR (%)' : 'CFR (%) 2019'
})
df_dbd_2019
```

```
df_dbd_2019['provinsi']=
df_dbd_2019["provinsi"].replace({
    "Dki Jakarta": "DKI Jakarta",
    'Di Yogyakarta' : 'DI Yogyakarta',
    'Kepulauan Riau' : 'Kep. Riau',
```

```
'Kepulauan Bangka Belitung' : 'Kep. Bangka
Belitung'})
df_dbd_2019
```

- Integrasi Dataset DBD tahun 2020

Sebelum melakukan merge dilakukan pergantian nama kolom dan standarisasi nama provinsi. Kemudian data hasil nya disimpan dengan nama “df_dbd_2020.csv”.

```
df_dbd_2020= df_dbd_2020.rename(columns={
    'Provinsi': 'provinsi',
    'Jumlah Penduduk' : 'jumlah penduduk 2020',
    'Jumlah Kasus' : 'jumlah kasus 2020',
    'Incidence Rate per\n100.000 Penduduk' : 'incidence
rate per 100.000 penduduk 2020',
    'Meninggal' : 'meninggal 2020',
    'CFR (%)' : 'CFR (%) 2020'
})
df_dbd_2020
```

```
df_dbd_2020['provinsi']=
df_dbd_2020["provinsi"].replace({
    "Dki Jakarta": "DKI Jakarta",
    'Di Yogyakarta' : 'DI Yogyakarta',
    'Kepulauan Riau' : 'Kep. Riau',
    'Kepulauan Bangka Belitung' : 'Kep. Bangka
Belitung'})
df_dbd_2020
```

- Integrasi Dataset DBD tahun 2019 dan Dataset DBD tahun 2020

Data hasil wrangling untuk df_dbd disimpan dengan nama “DBD_2019-2020.csv”.

```
df_dbd = pd.merge(df_dbd_2019, df_dbd_2020,
on='provinsi', how='inner')
df_dbd
```

```
cols_2019 = [
    'jumlah penduduk 2019',
    'jumlah kasus 2019',
    'incidence rate per 100.000 penduduk 2019',
    'meninggal 2019',
    'CFR (%) 2019'
```

```
]

df_dbd_2019 = df_dbd[['provinsi'] + cols_2019].copy()
df_dbd_2019['tahun'] = 2019

df_dbd_2019 = df_dbd_2019.rename(columns={
    'jumlah penduduk 2019': 'jumlah_penduduk',
    'jumlah kasus 2019': 'jumlah_kasus',
    'incidence rate per 100.000 penduduk 2019':
'incidence_rate',
    'meninggal 2019': 'meninggal',
    'CFR (%) 2019': 'cfr'
})
```

```
cols_2020 = [
    'jumlah penduduk 2020',
    'jumlah kasus 2020',
    'incidence rate per 100.000 penduduk 2020',
    'meninggal 2020',
    'CFR (%) 2020'
]

df_dbd_2020 = df_dbd[['provinsi'] + cols_2020].copy()
df_dbd_2020['tahun'] = 2020

df_dbd_2020 = df_dbd_2020.rename(columns={
    'jumlah penduduk 2020': 'jumlah_penduduk',
    'jumlah kasus 2020': 'jumlah_kasus',
    'incidence rate per 100.000 penduduk 2020':
'incidence_rate',
    'meninggal 2020': 'meninggal',
    'CFR (%) 2020': 'cfr'
})
```

```
df_dbd = pd.concat([df_dbd_2019, df_dbd_2020],
ignore_index=True)
df_dbd
```

#Rename kolom

```
df_dbd= df_dbd.rename(columns={
    'jumlah_penduduk': 'JP',
    'jumlah_kasus': 'JK',
    'incidence_rate': 'IR',
    'meninggal': 'M',
    'cfr' : "CFR",
})
df_dbd
```

- Integrasi Dataset Gabungan (3 dataset)

Data hasil wrangling untuk df_gabungan disimpan dengan nama “Gabungan (all dataset).csv”.

Merge Iklim (df_iklim) dan Kepadatan Penduduk (df_kepduk) Tahun 2019 - 2020

```
# dataset iklim
df_iklim['tahun'] = df_iklim['tahun'].astype(int)
# dataset kepadatan penduduk
df_kepduk['tahun'] = df_kepduk['tahun'].astype(int)

# Merge dataset iklim (df_iklim) dan kepadatan penduduk (df_kepduk)
df_IP = df_iklim.merge(
    df_kepduk,
    how='left',
    on=['provinsi', 'tahun']
)
df_IP
```

```
df_IP = df_IP.rename(columns={
    'nilai_kepadatan': 'kepduk',
})
df_IP
```

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

```
df_gabungan = df_IP.merge(
    df_dbd,
    how='left',
    on=['provinsi', 'tahun']
)
df_gabungan
```

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

```
df_gabungan.isnull().sum()
```

d. Transformasi Data

```
#Agregasi dataset iklim
df_iklim = df_final.groupby(['provinsi', 'tahun'],
as_index=False)[['Tn', 'Tx', 'Tavg', 'RH_avg', 'RR', 'ss', 'ff_x', 'f
f_avg']].mean()
df_iklim
```

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

o Statistik Deskriptif

- Rata-rata

```
kolom = [
    "Tavg",    # Suhu rata-rata
    "RH_avg",  # Kelembaban rata-rata
    "RR",      # Curah hujan
    "ss",      # Durasi sinar matahari
    "ff_avg",  # Kecepatan angin rata-rata
    "JK",      # Jumlah Kasus DBD
    "KP"       # Kepadatan Penduduk
]
mean = df_gabungan[kolom].mean()
print(mean)
```

- Standar Deviasi

```
## Standar Deviasi
kolomstd = [
    "Tavg",    # Suhu rata-rata
    "RH_avg",  # Kelembaban rata-rata
    "RR",      # Curah hujan
    "ss",      # Durasi sinar matahari
    "ff_avg",  # Kecepatan angin rata-rata
    "JK",      # Jumlah Kasus DBD
    "KP"       # Kepadatan Penduduk
]
standar_deviasi = df_gabungan[kolomstd].std()
print(standar_deviasi)
```

- Nilai Minimum

```
# Filter tahun 2019
df_2019 = df_gabungan[df_gabungan["tahun"] == 2019]

# Cari indeks nilai minimum jumlah kasus DBD (JK)
idx_minn = df_2019["JK"].idxmin()
idx_minn

# Tampilkan informasi lengkap
df_2019.loc[idx_minn, [
    "provinsi", "tahun", "RR", "Tavg", "RH_avg", "ff_avg",
    "KP",      # kepadatan 2019
    "IR",      # incidence rate per 100.000 penduduk 2019
    "JP",      # jumlah penduduk 2019
    "JK"       # jumlah kasus 2019
]]

# Filter tahun 2020
df_2020 = df_gabungan[df_gabungan["tahun"] == 2020]

# Cari indeks nilai minimum jumlah kasus DBD (JK)
idx_mini = df_2020["JK"].idxmin()

# Tampilkan informasi lengkap
df_2020.loc[idx_mini, [
    "provinsi", "tahun", "RR", "Tavg", "RH_avg", "ff_avg",
    "KP",      # kepadatan 2019
    "IR",      # incidence rate per 100.000 penduduk 2020
    "JP",      # jumlah penduduk 2019
    "JK"       # jumlah kasus 2019
]]
```


- Nilai Maksimum

```
# Filter tahun 2019
df_2019 = df_gabungan[df_gabungan["tahun"] == 2019]

# Cari indeks nilai minimum jumlah kasus DBD (JK)
idx_maks = df_2019["JK"].idxmax()

# Tampilkan informasi lengkap
df_2019.loc[idx_maks, [
    "provinsi", "tahun", "RR", "Tavg", "RH_avg", "ff_avg",
    "KP",      # kepadatan 2019
    "IR",      # incidence rate per 100.000 penduduk 2019
    "JP",      # jumlah penduduk 2019
    "JK"       # jumlah kasus 2019
]]
```

```
# Filter tahun 2020
df_2020 = df_gabungan[df_gabungan["tahun"] == 2020]

# Cari indeks nilai minimum jumlah kasus DBD (JK)
idx_makss = df_2020["JK"].idxmax()

# Tampilkan informasi lengkap
df_2020.loc[idx_makss, [
    "provinsi", "tahun", "RR", "Tavg", "RH_avg", "ff_avg",
    "KP",      # kepadatan 2019
    "IR",      # incidence rate per 100.000 penduduk 2020
    "JP",      # jumlah penduduk 2019
    "JK"       # jumlah kasus 2019
]]
```

○ Analisis Eksploratif Visual (Visualisasi EDA)

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

- `pivot_table()` membuat kolom baru 2019 dan 2020
- `sort_values(by=2020)` mengurutkan provinsi dari kasus terbanyak tahun 2020

```
import matplotlib.pyplot as plt

# Pivot tabel: Provinsi vs Tahun, dengan nilai jumlah kasus (JK)
df_prov_cases = df_gabungan.pivot_table(
    index='provinsi',
    columns='tahun',
    values='JK',
    aggfunc='sum' # bisa sum krn tiap provinsi muncul 1x per tahun
)

# Pastikan kolom tersortir 2020 diurutkan
df_prov_cases = df_prov_cases.sort_values (by = 2020, ascending
= False)

# Plot
df_prov_cases.plot(kind='bar', figsize=(14, 7))
plt.title('Perbandingan Jumlah Kasus DBD per Provinsi (2019 vs
2020)')
plt.ylabel('Jumlah Kasus DBD')
plt.xlabel('Provinsi')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.savefig("bar_chart.png", dpi=300, bbox_inches='tight')
plt.show()
plt.close()
```

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

```
plt.figure(figsize=(10, 6))
colors = {2019: 'blue', 2020: 'red'}
line_styles = {2019: '--', 2020: '-'}
```

```

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

# Menambahkan label provinsi
for i in range(len(group)):
    provinsi_name = group.iloc[i]['provinsi']
    # Offset posisi label
    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]['IR'] + y_offset,
        provinsi_name,
        fontsize=8,
        alpha=0.8,
        ha='center',
        color=colors[year]
    )

# Garis Regresi
X = group['Tavg']
Y = group['IR']
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 (°C)')
plt.ylabel('Incidence Rate per 100.000 Penduduk')
plt.legend(title='Tahun')
plt.grid(True, linestyle='--', alpha=0.5)
plt.tight_layout()
plt.savefig("scatter_tavg_IR.png", dpi=300, bbox_inches='tight')
plt.show()
plt.close()

```

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

```

plt.figure(figsize=(14, 8))
colors = {2019: 'blue', 2020: 'red'}
line_styles = {2019: '--', 2020: '-'}

# Loop per tahun
for year, group in df_gabungan.groupby('tahun'):

    # Scatter plot
    plt.scatter(
        group['RR'],
        group['IR'],
        label=f'Data {year}',
        alpha=0.6,
        s=100,
        color=colors[year]
    )

```

```

)

# Tambahkan label provinsi
for i in range(len(group)):
    provinsi_name = group.iloc[i]['provinsi']
    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]['IR'] + y_offset,
        provinsi_name,
        fontsize=8,
        alpha=0.8,
        ha='center',
        color=colors[year]
    )

# Garis regresi
X = group['RR']
Y = group['IR']
z = np.polyfit(X, Y, 1) # slope & intercept
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 (Curah Hujan) vs Incidence Rate DBD
(dengan Label Provinsi)')
plt.xlabel('Rata-rata Curah Hujan Tahunan (RR)')
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')
plt.show()
plt.close()
```

- Korelasi Spearman digunakan karena tidak mengasumsikan hubungan linear dan tidak mensyaratkan data berdistribusi normal, sehingga lebih sesuai untuk data yang bersifat non-linear, memiliki outlier, serta skala yang tidak selalu interval. Spearman mengukur hubungan monotonik (tidak harus linear, tetapi arah konsisten), sehingga lebih robust terhadap nilai ekstrem dan hubungan yang tidak sepenuhnya linier, yang umum ditemukan pada kasus DBD dan faktor lingkungannya.

- Heatmap Matriks Korelasi Tahun 2019

```
# Filter hanya tahun 2019
df_2019 = df_gabungan[df_gabungan["tahun"] == 2019]
# Pilih semua kolom numerik saja
df_num_2019 = df_2019.select_dtypes(include=['int64', 'float64'])
df_num_2019 = df_num_2019.drop(columns=['tahun'],
errors='ignore')

# Hitung Matriks Korelasi
df_corr_matrix = df_num_2019.corr(method="spearman")
```

```

# Plot Heatmap
plt.figure(figsize=(10, 9))
sns.heatmap(
    df_corr_matrix,
    annot=True,
    cmap='coolwarm',
    fmt=".2f",
    linewidths=.5,
    cbar_kws={'label': 'Koefisien Korelasi'}
)

plt.title('Matriks Korelasi Semua Variabel Numerik (Tahun 2019)')
plt.xticks(rotation=45, ha='right')
plt.yticks(rotation=0)
plt.tight_layout()
plt.savefig("heatmap_2019.png", dpi=300, bbox_inches = 'tight')
plt.show()

```

- Heatmap Matriks Korelasi Tahun 2020

```

# Filter hanya tahun 2020
df_2020 = df_gabungan[df_gabungan["tahun"] == 2020]

# Pilih semua kolom numerik saja
df_num_2020 = df_2020.select_dtypes(include=['int64', 'float64'])
df_num_2020 = df_num_2020.drop(columns=['tahun'],
errors='ignore')

# Hitung Matriks Korelasi
df_corr_matrix = df_num_2020.corr(method="spearman")

# Plot Heatmap
plt.figure(figsize=(10, 9))
sns.heatmap(

```

```

df_corr_matrix,
annot=True,
cmap='coolwarm',
fmt=".2f",
linewidths=.5,
cbar_kws={'label': 'Koefisien Korelasi'}
)

plt.title('Matriks Korelasi Semua Variabel Numerik (Tahun 2020)')
plt.xticks(rotation=45, ha='right')
plt.yticks(rotation=0)
plt.tight_layout()
plt.savefig("heatmap_2020.png", dpi=300, bbox_inches='tight')
plt.show()

```

- Scatter Plot Kepadatan Penduduk vs Incidence Rate per 100.000 Penduduk

➤ 2019

```

# Filter tahun 2019
df_2019 = df_gabungan[df_gabungan["tahun"] == 2019]

plt.figure(figsize=(7,5))
plt.scatter(
    df_2019["KP"],    # Kepadatan Penduduk
    df_2019["IR"],    # Incidence Rate
    alpha=0.5
)
plt.title("Kepadatan Penduduk vs Incidence Rate DBD (2019)")
plt.xlabel("Kepadatan Penduduk (jiwa/km²)")
plt.ylabel("Incidence Rate per 100.000 Penduduk")
plt.grid()

```



```
plt.savefig("scatter_kepduk_IR_2019.png",          dpi=300,  
bbox_inches='tight')  
plt.show()
```

➤ 2020

```
# Filter tahun 2020  
df_2020 = df_gabungan[df_gabungan["tahun"] == 2020]  
  
plt.figure(figsize=(7,5))  
plt.scatter(  
    df_2020["KP"],    # Kepadatan Penduduk  
    df_2020["IR"],    # Incidence Rate  
    alpha=0.5  
)  
plt.title("Kepadatan Penduduk vs Incidence Rate DBD  
(2020)")  
plt.xlabel("Kepadatan Penduduk (jiwa/km²)")  
plt.ylabel("Incidence Rate per 100.000 Penduduk")  
plt.grid()  
plt.savefig("scatter_kepduk_IR_2020.png",          dpi=300,  
bbox_inches = 'tight')  
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 berdasarkan provinsi di Indonesia yang 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` sebagai berikut.
 - 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 untuk tiga dataset yaitu dataset iklim, kepadatan penduduk, 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 nama 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 yogyakarta”. 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 tahun dibuat baru menggunakan dt.year dari kolom tanggal iklim. Perbaikan yang ketiga pada kolom numerik yang dikonversi ke integer. Hasil yang diperoleh yaitu seluruh 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] df_gabungan = pd.merge(df_dbd, df_IP, on='provinsi', how='inner')
[76] 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 x 27 columns

3. Rename kolom

Setelah dataset dbd 2019 dan dbd 2020 di merge menjadi satu, selanjutnya dilakukan perubahan nama kolom yang semula mengandung nama dengan lebih dari satu kata maka akan diganti dengan inisial agar hasil akhir data mudah dibaca

dan dipahami. Hasil akhirnya semua nama kolom menjadi singkat dan dapat dengan mudah dipahami oleh pembaca.

provinsi = 34 Provinsi di Indonesia

tahun = Tahun 2019 dan 2020

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)

ff_avg = kecepatan angin rata-rata (m/s)

KP = Kepadatan Penduduk (jiwa/km²)

JP = Jumlah Penduduk Menurut Provinsi di Indonesia

JK = Jumlah Kasus DBD Menurut Provinsi di Indonesia

IR = Incidence Rate per 100.000 Penduduk / Tingkat Insidensi DBD per 100.000 Penduduk Provinsi

M = Jumlah Penduduk Meninggal Menurut Provinsi di Indonesia

CFR = *Case Fatality Rate* (CFR) / Tingkat Kematian Kasus DBD (%) Menurut Provinsi di Indonesia

4.1.4 Transformasi Data

Pada bagian ini merupakan proses menyamaan bentuk data agar dapat dianalisis bersama. Transformasi data yang dilakukan berupa agregasi pada dataset iklim yang dimana datanya berbentuk data harian. Dengan agregasi perata-rataan data harian menjadi data pertahunan. Dengan begitu bentuk data akan sama dengan dataset lainnya yang juga berbentuk tahunan. Hasil dari transformasi data berupa dataset iklim dengan data suhu, kelembaban, curah hujan, dan sebagainya berbentuk data tahunan.

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

Rata-rata suhu udara (T_{avg}) di seluruh provinsi berada pada kisaran $23,95^{\circ}\text{C}$, mencerminkan karakter iklim tropis Indonesia yang hangat dan relatif stabil sepanjang tahun. Kondisi suhu ini mendukung siklus hidup nyamuk *Aedes aegypti* yang berkembang optimal pada kisaran suhu $24\text{--}30^{\circ}\text{C}$. Kelembaban udara rata-rata (RH_{avg}) mencapai 72,65%, yang merupakan tingkat kelembaban ideal bagi kelangsungan hidup nyamuk, memperpanjang umur nyamuk dewasa serta meningkatkan aktivitas menggigit. Curah hujan rata-rata (RR) sebesar 6,36 mm per hari dan durasi penyinaran matahari (ss) sebesar 5,38 jam menunjukkan kombinasi kondisi lingkungan yang dapat memicu terbentuknya genangan air sebagai tempat perindukan nyamuk. Sementara itu, kecepatan angin rata-rata (ff_{avg}) sebesar 1,73 m/s tergolong rendah sehingga tidak banyak menghambat pergerakan nyamuk. Secara keseluruhan, gambaran kondisi iklim ini menunjukkan lingkungan yang mendukung penyebaran vektor DBD sepanjang tahun.

Dari aspek epidemiologis, jumlah kasus DBD rata-rata antar provinsi mencapai sekitar 3.624 kasus pada periode pengamatan, mencerminkan beban penyakit yang masih cukup tinggi di berbagai wilayah Indonesia. Nilai rata-rata ini

juga menunjukkan adanya variasi signifikan antar provinsi, terutama antara wilayah berpenduduk padat dan wilayah dengan distribusi penduduk yang lebih tersebar. Rata-rata kepadatan penduduk sebesar 740,63 jiwa/km² mengindikasikan bahwa banyak provinsi memiliki lingkungan pemukiman yang padat, kondisi yang dapat mempercepat penularan DBD mengingat jarak jelajah nyamuk *Aedes* relatif pendek. Wilayah padat penduduk cenderung mengalami risiko transmisi lebih tinggi karena peluang kontak antara nyamuk dan manusia menjadi lebih besar. Gambaran rata-rata tersebut memberikan dasar awal dalam memahami bagaimana faktor demografi dan epidemiologi saling berhubungan dengan kondisi iklim dalam memengaruhi dinamika kasus DBD di Indonesia.

4.2.1.2 Standar Deviasi

Variabel iklim, seperti suhu rata-rata ($T_{avg} = \approx 4,69^{\circ}\text{C}$), kelembapan udara ($RH_{avg} = \approx 14,73\%$), curah hujan ($RR = \approx 2,07 \text{ mm}$), durasi penyinaran matahari ($ss = \approx 0,97 \text{ jam}$), dan kecepatan angin rata-rata ($ff_{avg} = \approx 0,45 \text{ m/s}$), memiliki tingkat variasi yang relatif rendah antar provinsi. Hal ini menandakan bahwa kondisi iklim di Indonesia pada periode tersebut cenderung homogen, sesuai karakter iklim tropis yang stabil. Sebaliknya, jumlah kasus DBD menunjukkan penyebaran yang sangat besar dengan standar deviasi mencapai ≈ 4597 kasus, mencerminkan ketimpangan yang besar antara provinsi yang memiliki kasus sangat tinggi dan yang jumlah kasusnya jauh lebih rendah. Variasi serupa terlihat pada kepadatan penduduk ($KP = \approx 2689 \text{ jiwa/km}^2$), yang mengindikasikan bahwa distribusi penduduk antar provinsi sangat tidak merata. Provinsi berpenduduk padat berpotensi mengalami risiko transmisi DBD lebih tinggi, sementara wilayah dengan kepadatan rendah cenderung memiliki risiko yang lebih kecil.

4.2.1.3 Nilai Minimum

- a. Nilai minimum berdasarkan jumlah kasus DBD pada tahun 2019

provinsi	tahun	RR	Tavg	RH_avg	ff_avg	KP	IR	JK
Maluku	2019	5,18	23,05	72,456	2,13	38	13,09	236

Provinsi Maluku memiliki jumlah kasus DBD terendah sebesar 236 kasus dbd pada tahun 2019 dengan nilai incidence rate DBD paling rendah, 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 kepadatan penduduk yang rendah, sehingga peluang kontak antara manusia dan nyamuk

pembawa virus menjadi 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.

- b. Nilai minimum berdasarkan jumlah kasus DBD pada tahun 2020

provinsi	tahun	RR	Tavg	RH_avg	ff_avg	KP	IR	JK
Maluku	2020	8,27	21,08	66,196	2,345	39	4,2	77

Pada tahun 2020, Maluku kembali menjadi provinsi dengan kasus DBD terendah yaitu 77 kasus, mengalami penurunan signifikan dari tahun sebelumnya. Kepadatan penduduk tetap rendah, sebesar 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 DBD semakin terkendali di wilayah Maluku.

4.2.1.4 Nilai Maksimum

- a. Nilai maksimum berdasarkan jumlah kasus DBD pada tahun 2019

provinsi	tahun	RR	Tavg	RH_avg	KP	IR	JK
Jawa Barat	2019	6,5	24,9	75,298	1394	47,6	23483

Pada tahun 2019, provinsi dengan jumlah kasus DBD tertinggi adalah Jawa Barat, yaitu mencapai 23.483 kasus. Tingginya angka kasus ini sejalan dengan 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 (75,3%) dan curah hujan yang cukup (6,5 mm) turut mendukung tumbuhnya habitat perkembangbiakan nyamuk Aedes.

- b. Nilai maksimum berdasarkan jumlah kasus DBD pada tahun 2020

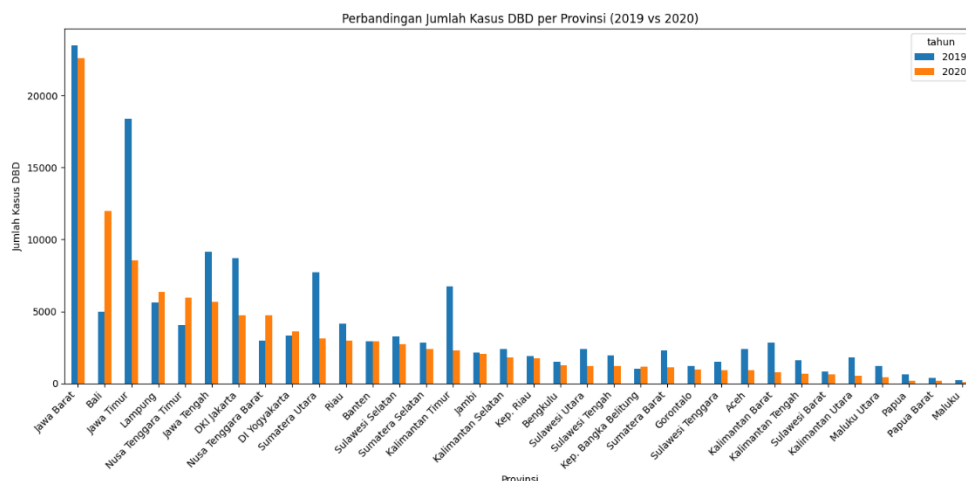
provinsi	tahun	RR	Tavg	RH_avg	KP	IR	JK
Jawa Barat	2020	9,04	24,65	80,11	1365	45,3	22613

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

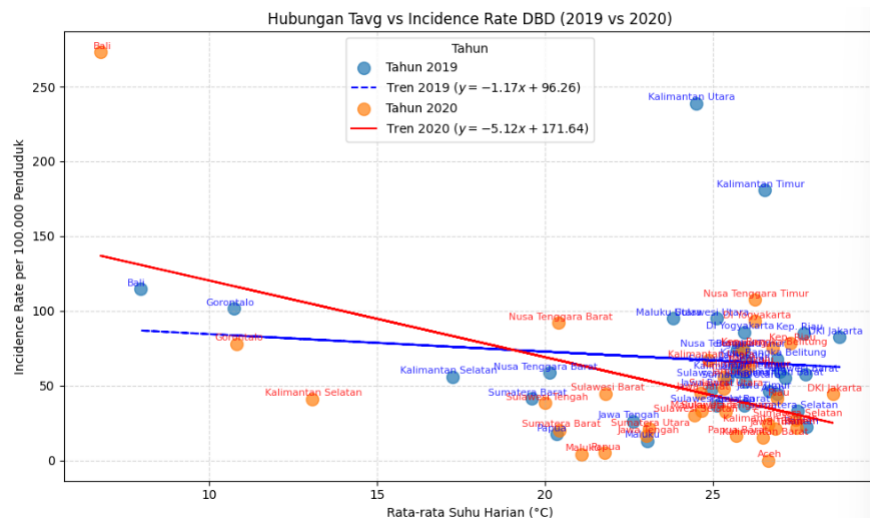


Gambar bar_chart kasus dbd19&20 per provinsi.png

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, Sumatera 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.

4.2.2.2 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.

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

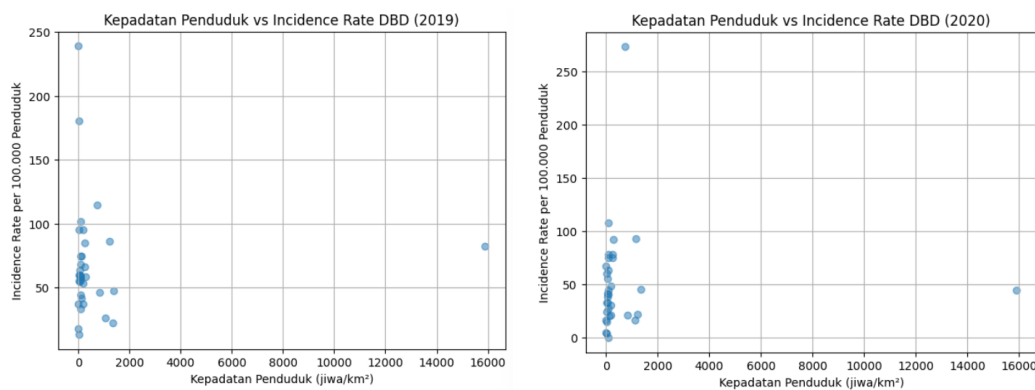


Grafik di atas membandingkan Total Curah Hujan Tahunan (RR - Rainfall) dan Insidence Rate DBD per 100.000 Penduduk (IR) untuk tahun 2019 dan 2020 untuk mengetahui hubungan di antaranya, di mana tahun 2019 (biru) dan tahun 2020 (merah). Kedua tahun menunjukkan pola korelasi yang berlawanan, yang diwakili 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 ($RR \approx 3$ hingga 5 mm), yang ditandai oleh lonjakan ekstrem pada Bali. Hal ini konsisten dengan penelitian Wulandari et al. (2023) yang juga menunjukkan 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.

4.2.2.4 Scatter Plot Kepadatan Penduduk vs Incidence Rate per 100.000 Penduduk



Gambar scatter_kepduk_IR_2019.png Gambar scatter_kepduk_IR_2020.png

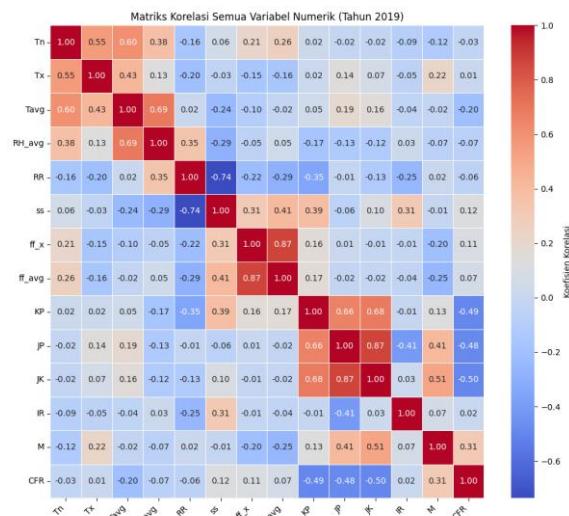
Scatter plot Kepadatan Penduduk vs. *Incidence Rate* (IR) tahun 2019 menunjukkan tidak adanya korelasi linier positif yang signifikan antara kepadatan dan risiko DBD per kapita, yang memperkuat temuan korelasi sebelumnya. Sebagian besar provinsi mengelompok pada kepadatan rendah, tetapi IR mereka sangat bervariasi, mencapai puncaknya di sekitar 250 per 100.000 penduduk di provinsi yang tidak padat. Sebaliknya, provinsi terpadat (seperti DKI Jakarta,

≈ 16.000 jiwa/km²) 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/km²) 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, $\approx 15.907 \approx 16.000$ jiwa/km²) 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 menunjukkan korelasi negatif antara kepadatan penduduk dengan kejadian DBD. Studi menunjukkan 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

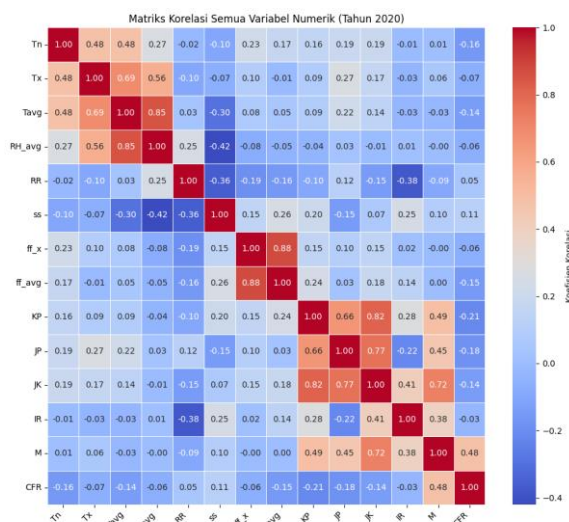
4.2.2.5 Heatmap Matriks Korelasi Tahun 2019



Gambar heatmap_2019.png

Jumlah Kasus (JK) memiliki hubungan yang sangat kuat dan positif dengan Jumlah Penduduk (JP) ($r = 0.87$), serta hubungan yang cukup kuat dan positif dengan Kepadatan Penduduk (KP) ($r = 0.68$). Ini mengindikasikan bahwa semakin banyak dan padat penduduk suatu wilayah, cenderung semakin tinggi pula Jumlah Kasus yang tercatat. Sebaliknya, hubungan antara Jumlah Kasus (JK) dan *Incidence Rate* (IR) adalah sangat lemah dan mendekati nol ($r = 0.03$), menunjukkan bahwa nilai *Incidence Rate* tidak berkorelasi secara linear dengan Jumlah Kasus dalam data ini. Mayoritas variabel iklim, seperti suhu rata-rata (Tavg) dan curah hujan (RR) berkorelasi sangat lemah dengan jumlah kasus DBD dan *incidence rate*, yang menegaskan bahwa faktor demografi, seperti kepadatan penduduk adalah prediktor yang lebih dominan terhadap jumlah kasus DBD dibandingkan faktor iklim.

4.2.2.6 Heatmap Matriks Korelasi Tahun 2020



Gambar heatmap_2020.png

Berdasarkan visualisasi heatmap di atas untuk tahun 2020 menunjukkan bahwa faktor pendorong utama Jumlah Kasus DBD adalah Kepadatan Penduduk, ditunjukkan oleh korelasi positif yang sangat kuat ($\rho \approx 0.82$ pada tahun 2020). Artinya, provinsi dengan populasi lebih padat memiliki jumlah kasus DBD yang jauh lebih tinggi. Namun, ketika kasus diukur sebagai *Incidence Rate* (kasus per 100.000 penduduk), korelasi dengan kepadatan menurun drastis menjadi hanya $\rho \approx 0.41$. 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 (T_{avg} , 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.

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. Terdapat 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.

- 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* maupun jumlah kasus DBD. Metode-metode tersebut 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* pemodelan.

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	<p>Sintaks (Google Colab):</p> <ul style="list-style-type: none"> - 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). <p>Laporan:</p> <ul style="list-style-type: none"> - 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	<p>Sintaks (Google Colab):</p> <ul style="list-style-type: none"> - Mengolah dataset Iklim dan Kepadatan Penduduk tahun 2019 dan 2020 dari download data, read file csv, cleaning data, integrasi dataset iklim dan provinsinya, integrasi kepadatan penduduk 2019 dan 2020, Integrasi ketiga dataset final. <p>Laporan:</p> <ul style="list-style-type: none"> - 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