LAPORAN TUGAS BESAR KOMPUTASI STATISTIK

"Analisis dan Optimasi Kualitas Dataset Produksi Durian Provinsi Lampung (2021–2024) Menggunakan Teknik Munging dan Wrangling dalam Komputasi Statistik"

Disusun Oleh:
Kelompok 9 RB

Devyna Sonya Palupi Sanjaya	(123450007)
Ken Gracya Waoma	(123450045)
Ali Aristo Muthhahari Parisi	(123450088)
Hanifah Inaya Sani	(123450123)

Dosen Pengampu:
Mika Alvionita S, S.Si., M.Si., Fitri Nurjanah, S.Si., M.Mat.

PROGRAM STUDI SAINS DATA
FAKULTAS SAINS
INSTITUT TEKNOLOGI SUMATERA

2025

ABSTRAK

Penelitian ini bertujuan untuk meningkatkan kualitas dataset produksi durian Provinsi Lampung tahun 2021–2024 melalui penerapan teknik data munging dan data wrangling. Dataset awal menunjukkan berbagai permasalahan, seperti nilai hilang, ketidakkonsistenan format, perbedaan struktur antar sumber data yang dapat mengganggu akurasi analisis komputasi statistik. Proses penelitian dimulai dengan inspeksi struktur data untuk mengidentifikasi pola, tipe variabel, serta kelengkapan nilai. Selanjutnya dilakukan tahapan munging, meliputi standarisasi tipe data, penyesuaian nama kolom, serta penggabungan beberapa dataset berdasarkan konteks waktu. Tahapan wrangling kemudian diterapkan melalui pembersihan nilai hilang dan validasi format. Perbandingan kondisi data sebelum dan sesudah preprocessing menunjukkan peningkatan signifikan pada keterbacaan, konsistensi, dan integritas dataset. Dataset hasil preprocessing dinilai layak untuk digunakan dalam analisis lanjutan, seperti pemodelan statistik, analisis tren, maupun peramalan produksi durian. Penelitian ini juga memberikan gambaran mengenai pentingnya preprocessing yang terstruktur dalam pengolahan data pertanian agar informasi yang dihasilkan lebih akurat dan dapat diandalkan.

Kata Kunci: data munging, data wrangling, preprocessing, komputasi statistik, produksi durian, Lampung.

DAFTAR ISI

BAB I

PENDAHULUAN

I.1 Latar Belakang

Produksi durian di Lampung terus menunjukkan peningkatan dari tahun ke tahun dan menjadi salah satu komoditas yang memberikan kontribusi penting bagi sektor hortikultura (cabang ilmu pertanian yang berfokus pada budidaya tanaman buah-buahan, sayuran, tanaman hias, dan tanaman obat) daerah. Informasi mengenai perkembangan produksi ini sangat dibutuhkan, baik untuk mendukung perencanaan pertanian, memetakan potensi wilayah, hingga menyusun strategi pemasaran. Namun, ketersediaan data yang memadai tidak selalu diikuti dengan kualitas data yang baik, sehingga sering kali menimbulkan kendala pada tahap analisis. Di lapangan atau di perkebunannya sendiri masih banyak data pertanian yang ditemukan dalam kondisi tidak rapi. Beberapa masalah yang sering sekali muncul yaitu nilai yang hilang, pencatatan yang tidak konsisten antar periode, dan format yang berbeda-beda. Kondisi seperti ini dapat menghambat proses analisis karena perhitungan statistik hanya dapat menghasilkan informasi yang akurat apabila data yang digunakan memiliki kualitas yang baik dan terstruktur dengan jelas.

Permasalahan semakin menjadi ketika data diperoleh dari berbagai sumber, seperti survei petani, laporan instansi pemerintah, atau file yang disusun dalam format digital yang berbeda beda. Ketidaksamaan struktur data, perbedaan nama kolom, atau variasi skema pencatatan sering membuat proses penggabungan data menjadi sulit dilakukan. Apabila data dari berbagai sumber ini tidak diselesaikan terlebih dahulu, hasil analisis yang dihasilkan bisa tidak akurat dan menimbulkan interpretasi yang keliru mengenai kondisi produksi durian di Lampung. Untuk mengatasi tantangan tersebut, dibutuhkan proses pembersihan dan penataan data sebelum data dianalisis lebih lanjut. Pada tahap inilah teknik data munging dan data wrangling berperan penting. Melalui teknik ini, data dapat diperbaiki dengan cara mengisi nilai yang hilang, menghapus data yang tidak relevan, menstandarkan format, serta menyesuaikan struktur agar setiap sumber data dapat digabungkan dengan benar. Proses ini membantu memastikan bahwa data memiliki kualitas yang cukup baik untuk digunakan dalam analisis komputasi maupun pengolahan statistik.

Dengan menerapkan data munging dan wrangling, hasil analisis yang dihasilkan menjadi lebih akurat dan dapat dipercaya. Hal ini sangat penting mengingat data produksi durian digunakan sebagai dasar dalam pengambilan keputusan, seperti memprediksi potensi panen, menentukan strategi pengembangan wilayah, hingga menilai kebutuhan distribusi. Kualitas data yang baik akan memberikan gambaran yang lebih jelas mengenai kondisi produksi, sehingga dapat mendukung upaya peningkatan sektor pertanian di Lampung secara lebih terarah dan efektif.

I.2 Rumusan Masalah

Rumusan masalah dalam penelitian ini dapat dirumuskan sebagai berikut::

1. Bagaimana kondisi awal dataset produksi durian Lampung tahun 2021–2024 dilihat dari aspek struktur, kelengkapan, serta kualitas datanya?
2. Teknik munging dan wrangling apa saja yang diterapkan untuk membersihkan, menata, serta meningkatkan keandalan dataset tersebut?
3. Bagaimana perubahan kualitas dataset setelah proses preprocessing dilakukan, dan sejauh mana perbaikan tersebut mendukung kelayakan dataset untuk analisis lebih lanjut?

I. 3 Tujuan Penelitian

Tujuan dari penelitian ini dapat dijelaskan sebagai berikut:

1. Mendeskripsikan kondisi awal dataset berdasarkan struktur atribut, tingkat kelengkapan data, distribusi nilai, serta potensi masalah kualitas data lainnya.
2. Menerapkan teknik data munging dan data wrangling guna meningkatkan konsistensi, integritas, dan kesiapan dataset sebelum memasuki tahap analisis statistik atau pemodelan.
3. Membandingkan dataset sebelum dan sesudah preprocessing untuk menilai efektivitas proses perbaikan data serta dampaknya terhadap kualitas dataset secara keseluruhan.

I.4 Manfaat Penelitian

Manfaat yang dihasilkan dari penelitian ini meliputi:

1. Tersedianya dataset produksi durian yang lebih bersih, terstruktur, dan siap digunakan dalam berbagai analisis komputasional maupun pengembangan model prediktif.
2. Menjadi contoh penerapan preprocessing data pada data pertanian yang dapat digunakan sebagai referensi atau studi kasus bagi mahasiswa, peneliti, maupun instansi terkait.
3. Memberikan pemahaman yang lebih mendalam mengenai pentingnya preprocessing dalam komputasi statistik, khususnya dalam meningkatkan akurasi, reliabilitas, dan validitas hasil analisis.

BAB II

TINJAUAN PUSTAKA

II.1 Produksi Durian dan Konteks Pertanian Hortikultura

Durian merupakan salah satu komoditas hortikultura yang memiliki peranan penting dalam mendukung perekonomian masyarakat, terutama di wilayah pedesaan. Keunggulan nilai ekonomi durian tidak hanya berasal dari konsumsi segar, tetapi juga dari potensi pengembangan produk turunan dan peluang ekspor. Dalam konteks pertanian hortikultura, durian termasuk dalam kelompok buah-buahan yang memberikan kontribusi besar terhadap pendapatan daerah serta penyerapan tenaga kerja.

Produksi durian sendiri dipengaruhi oleh berbagai faktor internal maupun eksternal, yang mencakup kapasitas sumber daya alam, kemampuan petani, infrastruktur usaha tani, serta dinamika pasar.Meski demikian, sejumlah hambatan masih ditemukan, seperti penggunaan bibit non-unggul, pola pemasaran tradisional, dan keterbatasan modal. Dari sisi eksternal, peluang muncul dari meningkatnya permintaan pasar dan potensi pengembangan kebun durian secara lebih luas. Namun, ancaman seperti perubahan cuaca, gangguan hama, dan kurang optimalnya sistem irigasi tetap menjadi tantangan yang perlu diatasi. Oleh karena itu, pengembangan agribisnis durian memerlukan strategi terpadu yang meliputi peningkatan kapasitas petani, penerapan teknologi budidaya yang lebih baik, serta penguatan kelembagaan kelompok tani agar keberlanjutan produksi durian dapat terjaga dalam jangka panjang[1].

II.2 Data Munging/Data Wrangling

Data munging merupakan proses mengolah dan merapikan data mentah sehingga menjadi lebih terstruktur serta mudah digunakan dalam berbagai analisis. Data yang awalnya tidak teratur atau masih dalam bentuk asli diubah menjadi bentuk yang rapi dan konsisten. Proses ini sangat penting terutama ketika data berasal dari banyak sumber dengan format yang tidak seragam, sehingga diperlukan penyesuaian sebelum data dapat dianalisis lebih lanjut.[2] Metode data munging mencakup berbagai aktivitas, seperti menyesuaikan tipe data, mengganti nama kolom agar konsisten, serta menggabungkan beberapa sumber data menjadi satu kesatuan yang terorganisir. Penataan ulang struktur ataupun urutan data juga dapat dilakukan agar sesuai dengan kebutuhan analisis. Dengan penataan yang tepat, peneliti atau analis dapat lebih mudah mengenali pola data tanpa terganggu oleh format yang berantakan.[3] Data wrangling, yang sering disebut juga sebagai data cleaning, merupakan proses yang lebih luas dan kompleks dibandingkan sekadar merapikan data. Proses ini melibatkan serangkaian tahapan yang sering kali perlu dilakukan berulang agar kualitas data meningkat dan memenuhi standar analisis. Ketika ditemukan ketidaktepatan seperti anomali, kesalahan input, atau ketidaksesuaian format, langkah-langkah

dalam data wrangling harus diterapkan kembali hingga diperoleh data yang benar-benar layak diproses[4].

Dalam praktiknya, data wrangling mencakup teknik-teknik seperti *filtering* untuk menyaring data yang tidak relevan, imputasi untuk menangani nilai hilang (*missing values*), serta standarisasi format agar setiap variabel memiliki konsistensi yang baik. Selain itu, proses validasi nilai juga diperlukan untuk memastikan bahwa data tidak mengandung kesalahan seperti duplikasi atau ketidaksesuaian semantik. Semakin bersih data yang dihasilkan, semakin kredibel hasil analisis yang diperoleh.[5] Secara keseluruhan, tujuan data munging dan wrangling adalah menghasilkan dataset yang akurat, konsisten, dan bebas dari kesalahan sehingga dapat digunakan tanpa menimbulkan kendala pada tahap analisis maupun visualisasi. Data yang telah melalui proses ini akan lebih mudah diolah menggunakan berbagai teknik statistik atau machine learning. Dengan kualitas data yang terjaga, hasil analisis menjadi lebih dapat dipercaya dan mampu memberikan insight yang lebih tepat sasaran.[6]

II.3 Missing Value

Missing value adalah kondisi ketika suatu data tidak tercatat atau tidak muncul pada suatu variabel dalam sebuah dataset[7]. Ketidakhadiran nilai ini dapat dipengaruhi oleh berbagai faktor, seperti kesalahan input data, instrumen pengukuran yang tidak berfungsi, atau respon yang tidak diberikan oleh responden dalam survei. Jika tidak ditangani dengan baik, missing value dapat menyebabkan distorsi pola data, menghasilkan bias dalam perhitungan statistik, serta menurunkan kualitas kesimpulan analisis. Oleh karena itu, diperlukan teknik penanganan khusus seperti imputasi, penghapusan baris tertentu, atau transformasi data agar hasil analisis tetap akurat dan dapat diandalkan.

BAB III

METODOLOGI

III.1 Jenis Data

Penelitian ini memanfaatkan data sekunder, yaitu data yang sudah dikumpulkan dan dipublikasikan oleh lembaga resmi. Dalam konteks penelitian ini, data diperoleh dari catatan produksi durian pada tingkat kabupaten/kota di Provinsi Lampung untuk periode 2021–2024. Penggunaan data sekunder dipilih karena informasi yang dibutuhkan telah tersedia, memiliki cakupan wilayah yang luas, serta disajikan dalam format yang dapat langsung diolah untuk keperluan analisis.

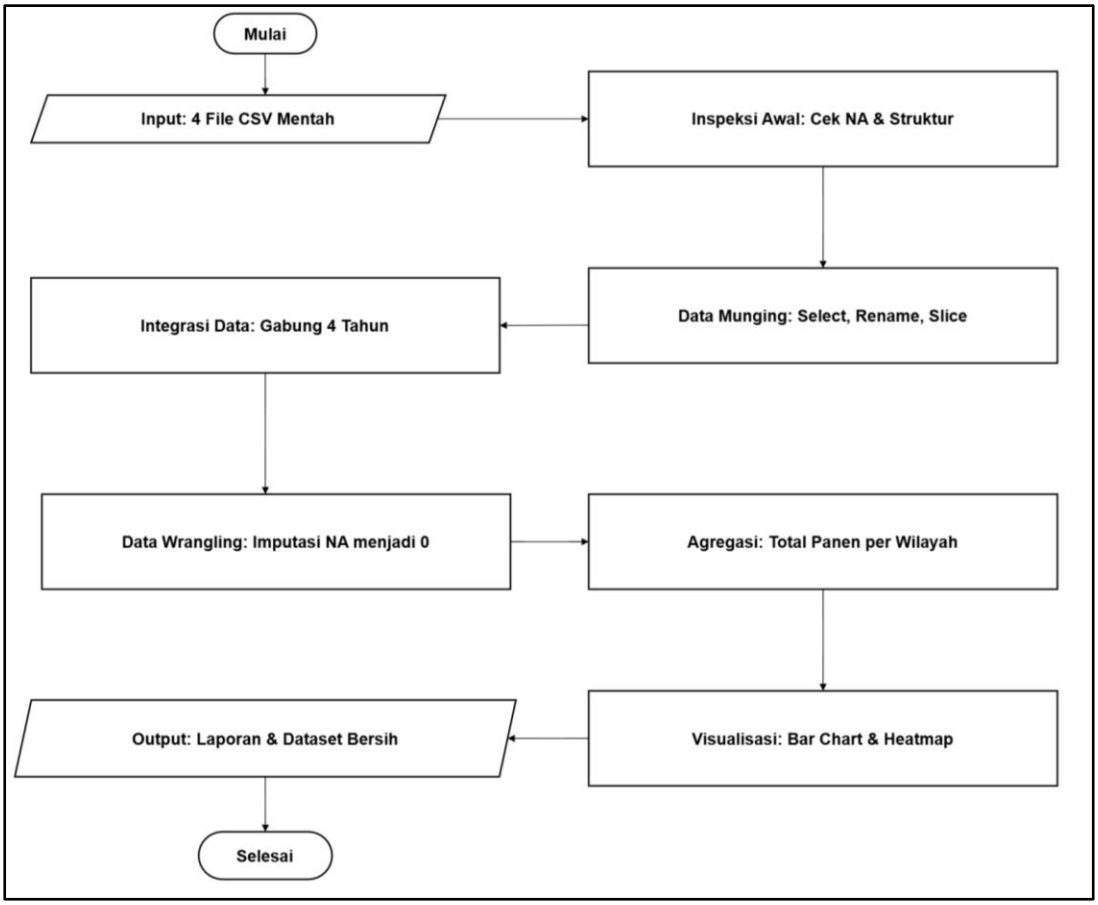
III.2 Teknik Pengumpulan Data

Pengumpulan data dilakukan melalui beberapa langkah sistematis. Langkah pertama adalah mengimpor dataset dalam format CSV atau Excel ke dalam perangkat lunak R. Dataset tersebut diperoleh dari Badan Pusat Statistik (BPS) Provinsi Lampung sebagai sumber resmi yang menyediakan data produksi durian per kabupaten/kota. Setelah data berhasil dimuat, dilakukan pemeriksaan awal untuk memastikan struktur, tipe variabel, serta kelengkapan data. Tahapan ini penting agar setiap variabel dapat diolah dengan tepat dan untuk mengidentifikasi jika terdapat kesalahan format atau ketidaksesuaian sebelum masuk ke tahap analisis lebih lanjut.

III.3 Variabel yang Diamati

1. Nama Kabupaten/Kota
2. Produksi Durian 2021
3. Produksi Durian 2022
4. Produksi Durian 2023
5. Produksi Durian 2024

III.4 Diagram Alir



Gambar 3.4 Diagram Alir

BAB IV

HASIL DAN PEMBAHASAN

IV.1 Deskripsi Data Awal

Tabel 1. Data Mentah Tahun 2021

Kabupaten/Kota	Produksi Durian (kuintal) (Kw)
Lampung Barat	10081.00
Tanggamus	20393.16
Lampung Selatan	19238.40
Lampung Timur	66375.75
Lampung Tengah	20863.79
Lampung Utara	9668.00
Way Kanan	8536.00
Tulangbawang	1473.00
Pesawaran	19222.75
Pringsewu	4925.00
Mesuji	NA
Tulang Bawang Barat	106.00
Pesisir Barat	18623.33
Kota Bandar Lampung	4422.65
Kota Metro	8.00
Lampung	203936.83

Pada tabel data mentah tahun 2021, baris Provinsi Lampung masih termuat di dalam dataset. Baris ini merupakan nilai total agregat yang seharusnya tidak disatukan dengan data per kabupaten/kota. Selain itu, terdapat nilai NA pada wilayah Mesuji yang perlu ditangani sebelum analisis dilakukan.

Tabel 2. Data Mentah Tahun 2022

Kabupaten/Kota	Produksi Durian (kuintal) (Kw)
Lampung Barat	6346.00
Tanggamus	32054.64
Lampung Selatan	28168.00
Lampung Timur	140808.31
Lampung Tengah	11038.44
Lampung Utara	11203.87
Way Kanan	2852.40
Tulangbawang	1431.00
Pesawaran	21092.37
Pringsewu	1119.00
Mesuji	NA
Tulang Bawang Barat	155.00
Pesisir Barat	17005.27
Kota Bandar Lampung	2622.90
Kota Metro	NA
Lampung	275897.20
NA	NA
Catatan	NA
Angka tetap	NA

Pada tabel data mentah tahun 2022, baris Provinsi Lampung masih termuat di dalam dataset, yang seharusnya dipisahkan dari data kabupaten/kota. Selain itu, struktur data terganggu oleh adanya baris NA, Catatan, dan Angka tetap yang sama sekali tidak memiliki pengaruh ke produksi durian. Masalah kelengkapan data juga terlihat karena masih terdapat nilai NA pada wilayah Mesuji dan Kota Metro yang harus ditangani.

Tabel 3. Data Mentah Tahun 2023

Kabupaten/Kota	Produksi Durian (kuintal) (Kw)
Lampung Barat	13471.78
Tanggamus	29123.20
Lampung Selatan	49039.50
Lampung Timur	258452.12
Lampung Tengah	13241.18
Lampung Utara	2887.40
Way Kanan	7727.82
Tulangbawang	1284.00
Pesawaran	59294.49
Pringsewu	1978.32
Mesuji	NA
Tulang Bawang Barat	497.60
Pesisir Barat	11649.06
Kota Bandar Lampung	1915.03
Kota Metro	461.60
Lampung	451023.10
NA	NA
Catatan	NA
Penambahan komoditas melinjo dan petai	NA

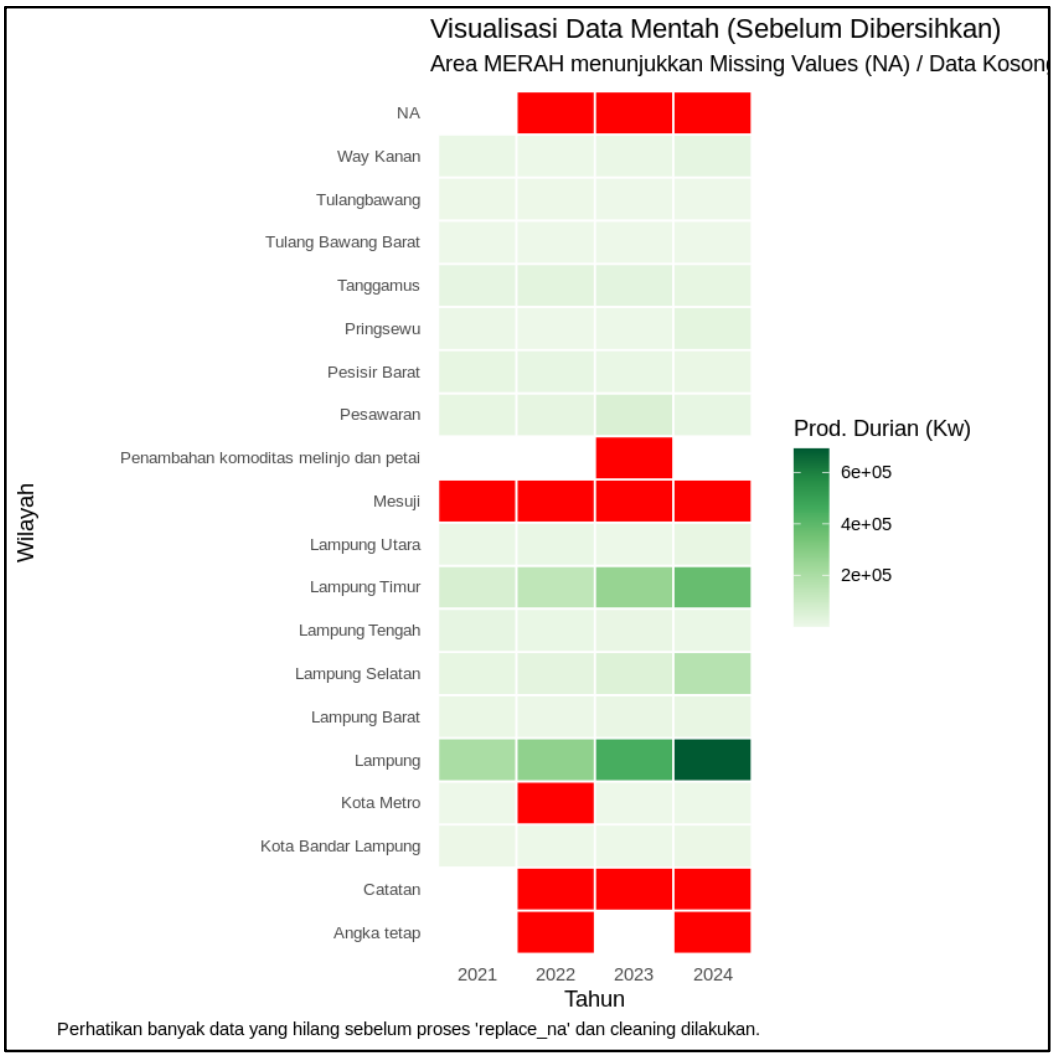
Pada tabel data mentah tahun 2023, baris Provinsi Lampung masih termuat di bagian bawah tabel. Selain itu, struktur data terganggu oleh adanya baris NA, Catatan, dan Penambahan komoditas melinjo dan petai yang tidak memiliki pengaruh ke produksi durian. Keberadaan nilai NA juga masih ditemukan pada wilayah tertentu, sehingga data belum siap untuk dianalisis secara langsung.

Tabel 4. Data Mentah Tahun 2024

Kabupaten/Kota	Produksi Durian (kuintal) (Kw)
Lampung Barat	16120.64
Tanggamus	18280.10
Lampung Selatan	166817.87
Lampung Timur	379198.86
Lampung Tengah	7927.01
Lampung Utara	15721.00
Way Kanan	23198.60
Tulangbawang	984.00
Pesawaran	17088.83
Pringsewu	27143.50
Mesuji	NA
Tulang Bawang Barat	1151.74
Pesisir Barat	10392.50
Kota Bandar Lampung	7046.20
Kota Metro	2267.20
Lampung	693338.05
NA	NA
Catatan	NA
Angka tetap	NA

Pada tabel data mentah tahun 2024, baris Provinsi Lampung masih termuat, mengulang pola ketidakkonsistenan dari tahun sebelumnya. Selain itu, struktur data terganggu oleh adanya baris NA, Catatan, dan Angka tetap yang tidak memiliki pengaruh ke produksi durian dan harus dibersihkan. Selain itu, masih terdapat nilai NA yang perlu diimputasi agar tidak menghambat proses komputasi statistik.

IV.2 Visualisasi Sebelum Munging/Wrangling



Gambar 4.2 Visualisasi Data Mentah

IV.3 Proses Munging/Wrangling

Dalam proses munging/wrangling data produksi durian, terdapat empat dataset terpisah yang mewakili periode tahun 2021 hingga 2024. Melakukan proses munging/wrangling secara manual dan rentan akan terjadinya human error. Oleh karena itu, dibuat sebuah fungsi otomatis untuk proses munging/wrangling bernama bersihkan_data().

```r
# Fungsi Pembersih (Wrangling Function)
bersihkan_data <- function(df, tahun) {
  df %>%
    # Memilih hanya kolom Kabupaten dan Durian
    select(`Kabupaten/Kota`, contains("Durian")) %>%

    # Mengganti nama kolom jadi simpel
    rename(Wilayah = `Kabupaten/Kota`,
           Produksi = 2) %>%

    # Mengambil hanya baris 1 sampai 15
    slice(1:15) %>%

    # Menambahkan penanda Tahun
    mutate(Tahun = as.character(tahun)) %>%

    # Mengatasi NA: Mengisi data kosong dengan 0
    mutate(Produksi = replace_na(Produksi, 0))
}
```

Gambar 4.3 Fungsi Pembersih

IV.3.1 Pendefinisian fungsi

→ bersihkan_data <- function(df, tahun) {

df %>%

Proses awal dilakukan dengan mendefinisikan fungsi bernama bersihkan_data() yang menerima dua parameter input yaitu df (data mentah) dan tahun (label waktu). Dilanjutkan dengan tanda %>% (pipe operator) yang langsung meneruskan data df ke proses selanjutnya.

IV.3.2 Seleksi variabel

→ select(`Kabupaten/Kota`, contains("Durian")) %>%

Seleksi kolom dilakukan menggunakan fungsi select dengan argumen `Kabupaten/Kota` untuk memilih kolom Kabupaten/Kota. Sedangkan fungsi contains("Durian") akan otomatis mendeteksi dan mengambil kolom yang mengandung kata durian.

IV.3.3 Standardisasi Nama Kolom

→ rename(Wilayah = `Kabupaten/Kota`,

Produksi = 2) %>%

Standardisasi menggunakan fungsi rename membuat nama variabel menjadi seragam. Mengubah kolom `Kabupaten/Kota` menjadi Wilayah. Argumen Produksi = 2 mengubah nama kolom durian hasil seleksi sebelumnya menjadi Produksi.

IV.3.4 Pemotongan Baris

→ slice(1:15) %>%

Fungsi slice dipakai untuk mengambil baris data yang relevan untuk digunakan.

IV.3.5 Penambahan Atribut Waktu Dan Perbaikan Tipe Data

→ mutate(Tahun = as.character(tahun)) %>%

Fungsi Mutate digunakan untuk menambah kolom baru bernama Tahun. Nilai tahun yang diinputkan diubah menjadi tipe data teks menggunakan as.character(). Pengubahan ini dilakukan agar variabel tahun dianggap sebagai kategori diskrit dalam visualisasi.

IV.3.6 Penanganan Missing Values

→ mutate(Produksi = replace_na(Produksi, 0))}

Menggunakan fungsi replace_na(Produksi, 0) untuk menangani missing value pada kolom produksi dengan angka 0. Penggunaan imputasi 0 dilakukan karena di wilayah tersebut tidak ada volume produksi di tahun tersebut.

IV.3.7 Penerapan Fungsi pada Setiap Tahun

```
# Menerapkan fungsi ke semua data
clean_21 <- bersihkan_data(df_2021, 2021)
clean_22 <- bersihkan_data(df_2022, 2022)
clean_23 <- bersihkan_data(df_2023, 2023)
clean_24 <- bersihkan_data(df_2024, 2024)
```

Gambar 4.3.7 Penerapan fungsi pada dataset

Hasil dari proses kode diatas adalah empat objek data baru yaitu clean_21, clean_22, clean_23, dan clean_24. Keempat objek data memiliki struktur seragam yang terdiri dari kolom Wilayah, Produksi, dan Tahun, serta bebas dari nilai NA.

Tabel 5. Data Bersih Tahun 2021

Wilayah	Produksi	Tahun
Lampung Barat	10081.00	2021
Tanggamus	20393.16	2021
Lampung Selatan	19238.40	2021
Lampung Timur	66375.75	2021
Lampung Tengah	20863.79	2021
Lampung Utara	9668.00	2021
Way Kanan	8536.00	2021
Tulangbawang	1473.00	2021
Pesawaran	19222.75	2021
Pringsewu	4925.00	2021
Mesuji	0.00	2021
Tulang Bawang Barat	106.00	2021
Pesisir Barat	18623.33	2021
Kota Bandar Lampung	4422.65	2021
Kota Metro	8.00	2021

Data diatas menampilkan produksi (dalam satuan tertentu) untuk 15 wilayah di Provinsi Lampung. Nilai produksi bervariasi, dengan beberapa wilayah memiliki output tinggi dan beberapa sangat rendah atau nol.

Tabel 6. Data Tahun 2022

Wilayah	Produksi	Tahun
Lampung Barat	6346.00	2022
Tanggamus	32054.64	2022
Lampung Selatan	28168.00	2022
Lampung Timur	140808.31	2022
Lampung Tengah	11038.44	2022
Lampung Utara	11203.87	2022
Way Kanan	2852.40	2022
Tulangbawang	1431.00	2022
Pesawaran	21092.37	2022
Pringsewu	1119.00	2022
Mesuji		2022
Tulang Bawang Barat	155.00	2022
Pesisir Barat	17005.27	2022
Kota Bandar Lampung	2622.90	2022
Kota Metro	0.00	2022

Data ini menunjukkan produksi tiap wilayah di Lampung dengan total 15 entri. Produksi tertinggi tercatat di Lampung Timur, sedangkan beberapa wilayah seperti Mesuji memiliki nilai nol.

Tabel 7. Data Tahun 2023

Wilayah	Produksi	Tahun
Lampung Barat	13471.78	2023
Tanggamus	29123.20	2023
Lampung Selatan	49039.50	2023
Lampung Timur	258452.12	2023
Lampung Tengah	13241.18	2023
Lampung Utara	2887.40	2023
Way Kanan	7727.82	2023
Tulangbawang	1284.00	2023
Pesawaran	59294.49	2023
Pringsewu	1978.32	2023
Mesuji	0.00	2023
Tulang Bawang Barat	497.60	2023
Pesisir Barat	11649.06	2023
Kota Bandar Lampung	1915.03	2023
Kota Metro	461.60	2023

Tabel ini memuat data produksi untuk 15 wilayah di Lampung. Lampung Timur kembali menjadi wilayah dengan produksi terbesar, jauh melampaui wilayah lainnya. Beberapa daerah seperti Mesuji tetap menunjukkan produksi nol.

Tabel 8. Data Tahun 2024

Wilayah	Produksi	Tahun
Lampung Barat	16120.64	2024
Tanggamus	18280.10	2024
Lampung Selatan	166817.87	2024
Lampung Timur	379198.86	2024
Lampung Tengah	7927.01	2024
Lampung Utara	15721.00	2024
Way Kanan	23198.60	2024
Tulangbawang	984.00	2024
Pesawaran	17088.83	2024
Pringsewu	27143.50	2024
Mesuji	0.00	2024
Tulang Bawang Barat	1151.74	2024
Pesisir Barat	10392.50	2024
Kota Bandar Lampung	7046.20	2024
Kota Metro	2267.20	2024

Data diatas menampilkan jumlah produksi 15 wilayah Lampung dengan selisih yang cukup besar antar daerah. Lampung Timur tetap mendominasi dengan produksi tertinggi, sementara Mesuji kembali mencatatkan produksi nol.

IV.3.8 Agregasi Data untuk Analisis Wilayah

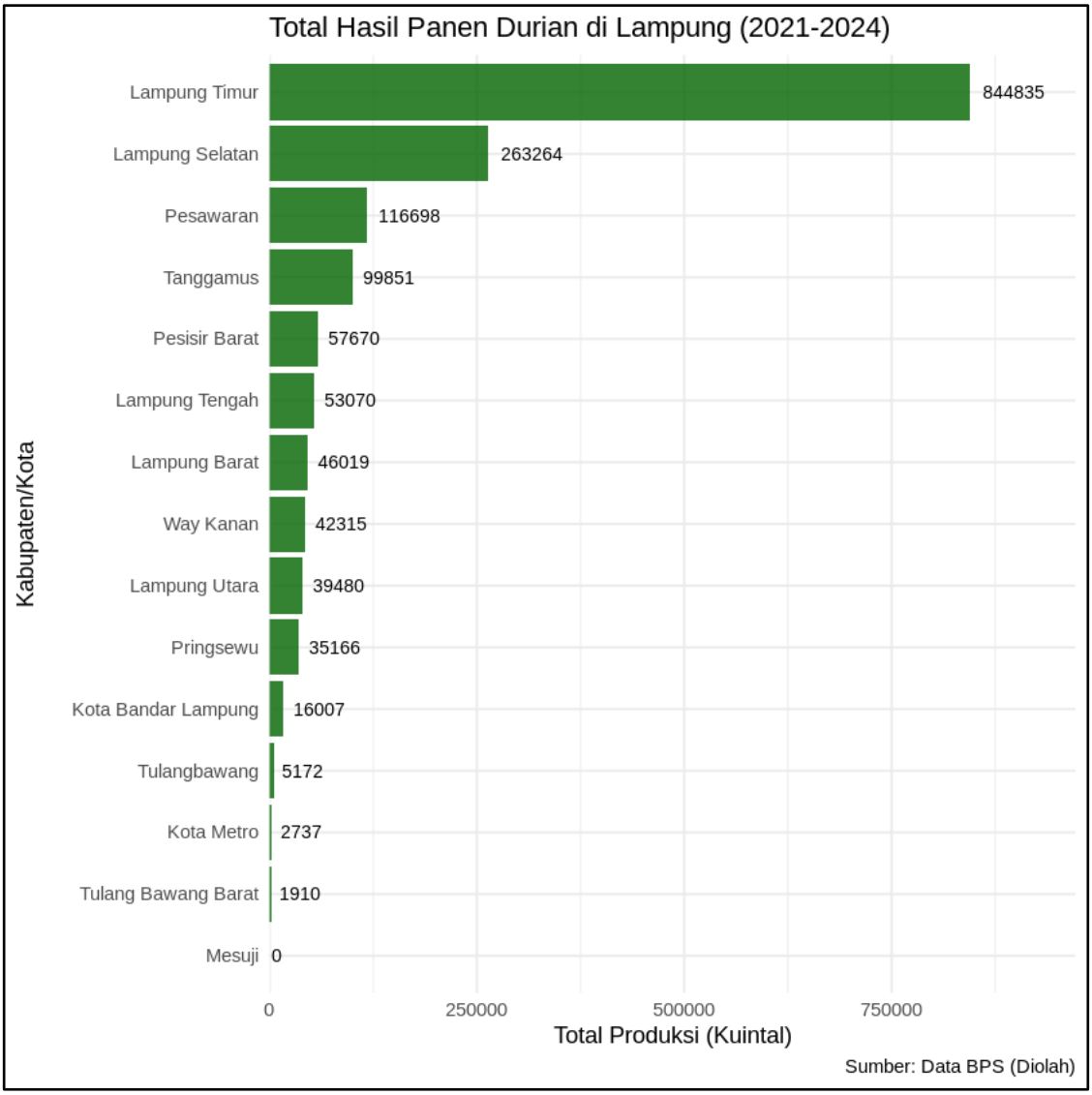
Data gabungan diagregasi untuk melihat total produksi durian per wilayah selama periode 2021-2024. Hal ini bertujuan mengidentifikasi pusat produksi durian

```r
# Mengatasi Agregasi (Total Panen per Wilayah)
durian_final <- durian_gabungan %>%

  # Menjumlahkan total produksi dari 2021-2024 per wilayah
  group_by(Wilayah) %>%
  summarise(Total_Panen = sum(Produksi)) %>%

  # Mengurutkan data dari yang terbanyak
  arrange(desc(Total_Panen))

# Tampilkan Data Hasil Wrangling
print(durian_final)
```

Gambar 4.3.8.1 Kode agregasi data 2021-2024

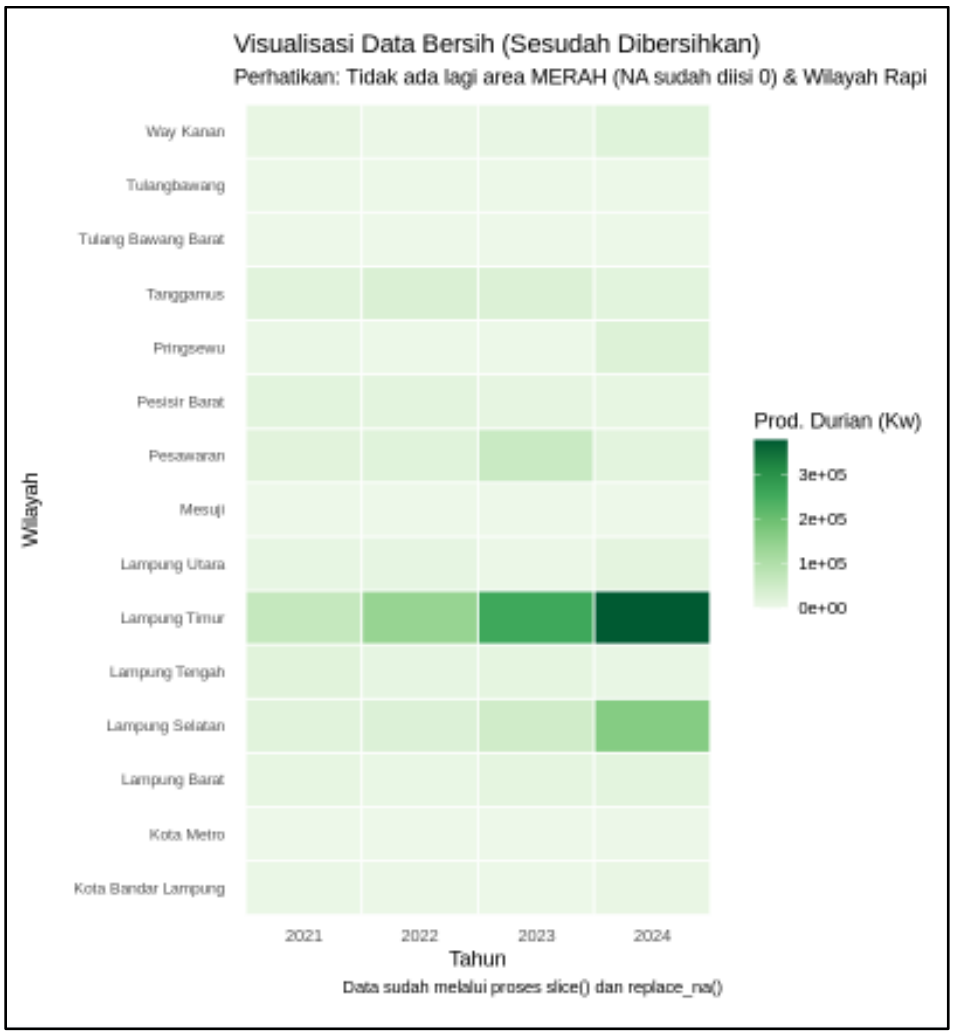
Tabel 9. Data total produksi durian per wilayah selama periode 2021-2024

Wilayah	Total Panen
Lampung Timur	844835
Lampung Selatan	263264
Pesawaran	116698
Tanggamus	99851
Pesisir Barat	57670
Lampung Tengah	53070
Lampung Barat	46019
Way Kanan	42315
Lampung Utara	39480
Pringsewu	35166
Kota Bandar Lampung	16007
Tulangbawang	5172
Kota Metro	2737
Tulang Bawang Barat	1910
Mesuji	0



Gambar 4.3.8.2 Total hasil panen durian di Lampung periode 2021-2024

IV.4 Visualisasi Setelah Munging/Wrangling



Gambar 4. 4Visualisasi data bersih

IV.5 Pembahasan

Analisis terhadap proses *data munging* dan *wrangling* yang telah dilakukan menunjukkan adanya transformasi signifikan pada kualitas dataset produksi durian Provinsi Lampung (2021-2024). Pembahasan ini menyoroti tiga aspek utama: penanganan inkonsistensi struktural, validasi penanganan nilai hilang (*missing values*), dan relevansi hasil terhadap prinsip komputasi statistik.

IV.5.1 Penanganan Inkonsistensi Struktural dan Dirty Data

Berdasarkan deskripsi data awal, dataset mentah mengandung banyak elemen non-observasi yang dapat mendistorsi hasil analisis. Pada dataset tahun 2021 hingga 2024, ditemukan baris-baris "sampah" di bagian bawah tabel seperti baris agregat "Lampung", baris "Catatan",

"Angka tetap", hingga "Penambahan komoditas melinjo dan petai". Jika baris-baris ini tidak dibersihkan, fungsi statistik agregat (seperti *mean* atau *sum*) akan menghitung nilai ganda (karena baris total provinsi ikut terhitung) atau mengalami kegagalan proses akibat tipe data yang tidak seragam. Penerapan teknik *slicing* (slice(1:15)) terbukti efektif untuk mengisolasi 15 baris observasi valid yang merepresentasikan kabupaten/kota, sehingga secara otomatis membuang baris footer yang tidak relevan. Selain itu, standarisasi nama kolom menjadi "Wilayah" dan "Produksi" mengatasi masalah perbedaan format penamaan antar tahun, memungkinkan penggabungan (*merging*) dataset multi-tahun menjadi satu kesatuan yang koheren.

IV.5.2 Justifikasi Metodologis Penanganan *Missing Values*

Salah satu temuan krusial dalam data awal adalah keberadaan nilai NA pada wilayah tertentu, seperti Kabupaten Mesuji dan Kota Metro pada tahun-tahun tertentu. Dalam penelitian ini, keputusan untuk melakukan imputasi nilai NA menjadi 0 menggunakan fungsi replace_na() didasarkan pada logika domain pengetahuan (*domain knowledge*) bahwa ketidakadaan data pencatatan pada konteks ini mengindikasikan tidak adanya volume produksi durian di wilayah tersebut pada tahun yang bersangkutan. Langkah ini krusial karena membiarkan nilai NA tetap ada akan menyebabkan hasil analisis statistik menjadi NaN atau *error* saat dilakukan operasi aritmatika, yang pada akhirnya menghambat proses pemodelan atau visualisasi.

IV.5.3 Peningkatan Integritas Data dan Prinsip GIGO

Proses *preprocessing* yang dilakukan telah berhasil membuktikan prinsip *Garbage In, Garbage Out* (GIGO). Sebelum *wrangling*, data mentah dikategorikan sebagai "Garbage" karena mengandung ketidakkonsistenan yang berisiko tinggi menghasilkan kesimpulan yang bias atau salah. Dengan mentransformasi tipe data kolom "Tahun" menjadi karakter (as.character), data yang semula mungkin dianggap sebagai nilai numerik kontinu kini diperlakukan sebagai kategori diskrit yang tepat untuk analisis tren waktu (*time-series*). Hasil akhirnya adalah dataset yang memenuhi prinsip *Tidy Data*, di mana setiap variabel membentuk kolom, setiap observasi membentuk baris, dan setiap nilai memiliki sel sendiri. Hal ini menjamin bahwa output visualisasi dan perhitungan statistik yang dihasilkan di bagian selanjutnya adalah valid dan dapat dipertanggungjawabkan secara ilmiah.

BAB V

KESIMPULAN DAN SARAN

V.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa dataset produksi durian Provinsi Lampung tahun 2021–2024 memiliki berbagai permasalahan kualitas data sebelum dilakukan preprocessing. Permasalahan tersebut meliputi keberadaan baris yang tidak relevan seperti baris total provinsi dan catatan, ketidakkonsistenan format antar tahun, nilai hilang (missing values) yang mengganggu proses analisis statistik. Melalui penerapan teknik data munging dan data wrangling, seluruh masalah tersebut dapat ditangani secara sistematis. Proses standarisasi nama kolom, pemilihan baris relevan, penghapusan bagian yang tidak diperlukan, imputasi nilai hilang, hingga penambahan atribut tahun berhasil menghasilkan dataset yang lebih rapi, bersih, dan konsisten. Penggabungan keempat dataset juga berjalan optimal karena struktur data telah distandarisasi.

Dataset akhir yang telah melalui proses preprocessing terbukti memiliki integritas yang lebih baik dan siap digunakan untuk analisis statistik lanjutan. Hasil ini menunjukkan bahwa penerapan munging dan wrangling berperan penting dalam meningkatkan kualitas data, sehingga mendukung analisis produksi durian secara lebih akurat, terpercaya, dan dapat dipertanggungjawabkan.

V.2 Saran

1. Penelitian selanjutnya disarankan menambah cakupan data dari tahun-tahun sebelum 2021 atau setelah 2024 untuk memperoleh gambaran tren produksi durian yang lebih panjang dan stabil. Rentang waktu yang lebih luas akan memudahkan analisis pola jangka panjang serta peramalan produksi.
2. Perlu mempertimbangkan penggunaan metode imputasi yang lebih canggih, seperti imputasi berbasis regresi atau metode berbasis model machine learning, terutama untuk dataset yang memiliki pola missing values yang kompleks.
3. Pengelolaan data dari instansi resmi seperti BPS atau dinas pertanian sebaiknya distandarisasi agar format pencatatan antar tahun konsisten. Hal ini akan memudahkan proses integrasi data pada penelitian serupa di masa mendatang.
4. Perlu dilakukan analisis lanjutan seperti visualisasi tren, analisis korelasi, atau model prediksi produksi untuk memanfaatkan dataset hasil preprocessing secara optimal, sehingga hasil penelitian dapat memberikan nilai tambah bagi perencanaan sektor pertanian daerah.
5. Penggunaan fungsi otomatis dalam munging/wrangling (seperti bersihkan_data()) perlu dikembangkan lebih lanjut, khususnya dengan menambahkan mekanisme deteksi baris observasi secara dinamis. Hal ini diperlukan untuk menggantikan metode pemotongan

baris manual (*hard-coded slicing*) yang saat ini digunakan, agar fungsi menjadi lebih adaptif dan tetap akurat meskipun terjadi perubahan struktur tabel atau penambahan jumlah wilayah administrasi di masa depan.

DAFTAR PUSTAKA

[1] Z. Arifin, *Potensi Pengembangan dan Strategi Usaha Agribisnis Buah Durian di Desa Tebul Timur Kecamatan Pegantenan Kabupaten Pamekasan*. Fakultas Pertanian, Universitas Islam Madura.

[2] F. Endel and H. Piringer, "Data Wrangling: Making data useful again," *IFAC-PapersOnLine*, vol. 48, no. 1, pp. 111–112, Jan. 2015, doi: 10.1016/j.ifacol.2015.05.197.

[3] I. Setiawan, A. M. Dawis, and Program Studi Sistem dan Teknologi Informasi Fakultas Sains dan Teknologi, *Data Science: Pendekatan dan Langkah Praktis dengan Excel*. 2023.

[4] F. Ridzuan and W. M. N. Wan Zainon, "A Review on Data Cleansing Methods for Big Data," *Procedia Computer Science*.

[5] W. Widyanto, V. L. Mahendra, F. A. R. Saputra, and R. R. Muhima, "Perancangan Data Infrastruktur dengan Menerapkan Teknik Data Wrangling Studi Kasus: Data Users di Narasio Data," Institut Teknologi Adhi Tama Surabaya, 2023.

[6] O. Azeroual, "Data Wrangling in Database Systems: Purging of Dirty Data," *Data*, vol. 5, no. 2, p. 50, Jun. 2020, doi: 10.3390/data5020050.

[7] A. S. Arifianto, K. D. Sawitri, K. Agustianto, and I. G. Wiryawan, "Pengaruh Prediksi Missing Value pada Klasifikasi Decision Tree C4.5," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 9, no. 4, p. 780, Agt. 2022, doi: 10.25126/jtiik.2022944778.

LAMPIRAN

Folder Codingan, Poster, PPT, Video, Dataset:

https://s.itera.id/Komstat_9_RB