LAPORAN PROJECT DATA WAREHOUSE
KELOMPOK 10

NAMA ANGGOTA :

1. ANDIKA SYUHADA (2355301017)
2. MOHAMAD HAZIQ DAFREN (2355301119)
3. HELFIA GISKA RENATA (2355301082)

KELAS : 2 TI D

DOSEN : Mutia Sari Zulvi, S.S.T., M.M.SI
ILB : Muhammad Anwar, S.Tr.Kom

PROGRAM STUDI TEKNIK INFORMATIKA
POLITEKNIK CALTEX RIAU
TA 2024 / 2025

1. **Memilih Proses :**
   Meningkatkan Prestasi Pendidikan di Indonesia.

2. **Grain:**
   - Total Perolehan Medali berdasarkan Sekolah
   - Total Perolehan Medali berdasarkan Provinsi
   - Total Perolehan Medali berdasarkan Medali
   - Total Perolehan Medali berdasarkan Tahun
   - Total Partisipan berdasarkan Sekolah
   - Total Partisipan berdasarkan Provinsi
   - Total Partisipan berdasarkan Medali
   - Total Partisipan berdasarkan Tahun

3. **Dimensi :**
   - Sekolah
   - Provinsi
   - Medali
   - Waktu (Tahun)

Measurement / Dimensi	**Sekolah**	**Provinsi**	**Medali**	**Tahun**
Total Perolehan Medali	✓	✓	✓	✓
Total Partisipan	✓	✓	✓	✓

4. **Fakta :**
   Tabel Fakta yang akan dibuat adalah Fakta Perolehan Medali dan Partisipasi .
   Measurement:
   1. Total Perolehan Medali
   2. Total Partisipan

5. **Menyimpan Pre-kalkulasi**
   - Medali : count (medali: emas, perak, perunggu)
   - Medali : count(medali: emas,perak,perunggu, partisipan)

6. Melengkapi Tabel Dimensi

Dimensi	Field	Deskripsi
Sekolah	Sekolah	Data dapat dilihat berdasarkan nama sekolah.
Provinsi	Provinsi	Data dapat dilihat berdasarkan nama provinsi.
Medali	Medali	Data dapat dilihat berdasarkan nama medali.
Waktu	Tahun	Data dapat dilihat per tahun.

- Dimensi Sekolah

Atribut	Tipe Data
id_sekolah	varchar(20)
nama_sekolah	varchar(255)

- Dimensi Medali

Atribut	Tipe Data
id_provinsi	varchar(20)
nama_provinsi	varchar(255)

- Dimensi Medali

Atribut	Tipe Data
id_medali	varchar(20)
medali	varchar(255)

- Dimensi Tahun

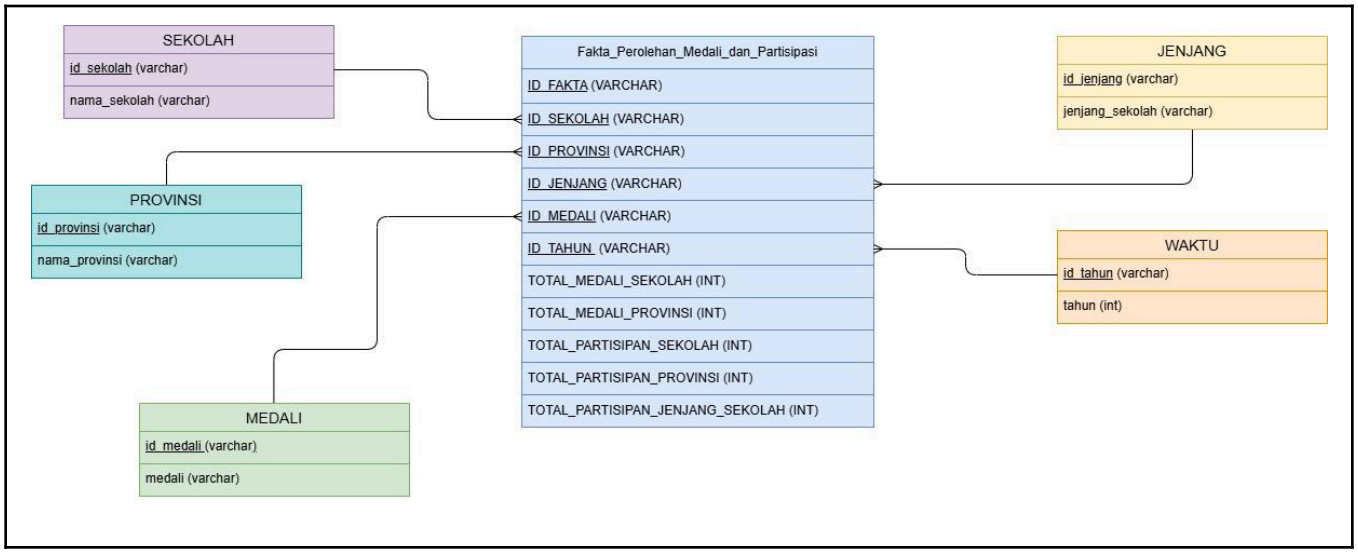
Atribut	Tipe Data
id_tahun	varchar(20)
tahun	int

7. Memilih Durasi dari Database

Durasi dari dataset Olimpiade Sains Nasional yang dimasukkan ke dalam data warehouse adalah sebagai berikut:

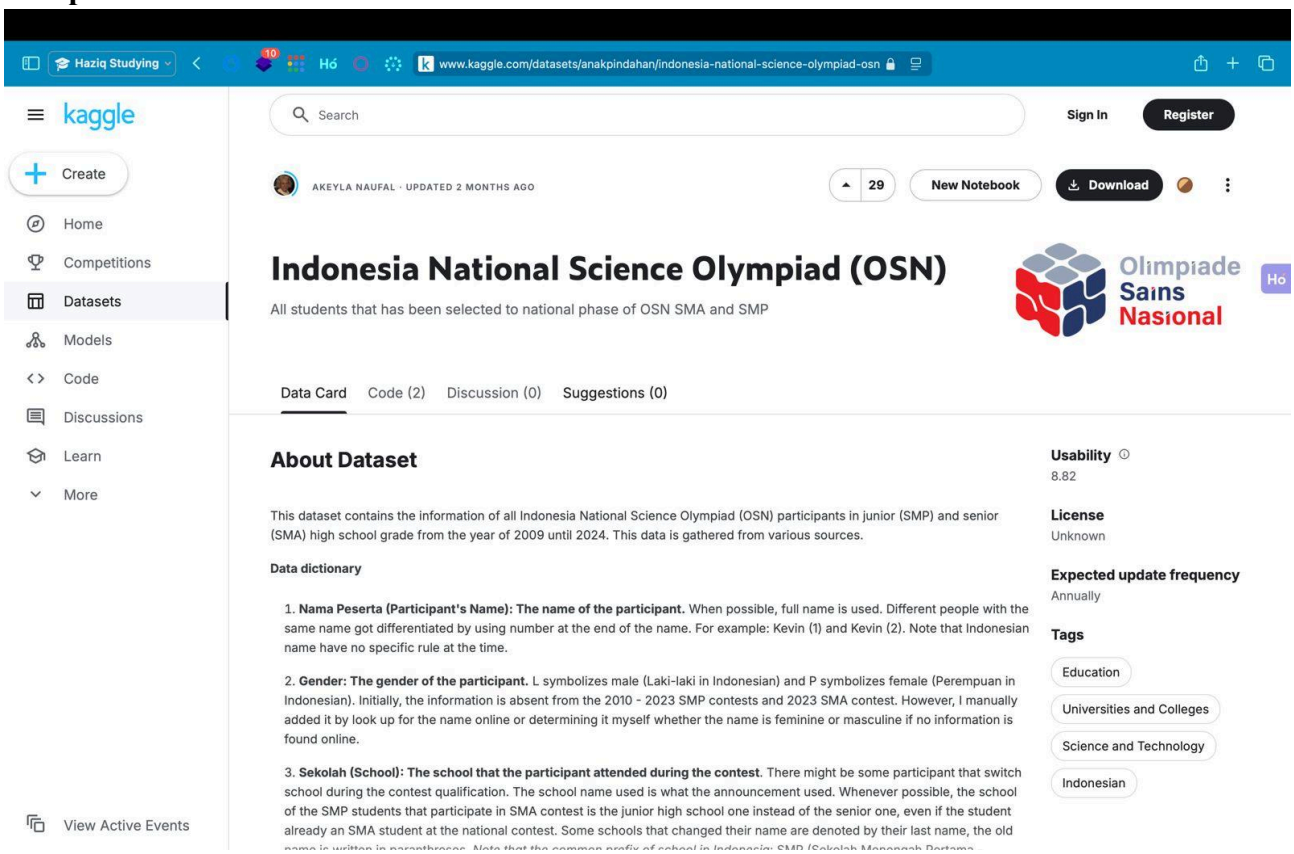
Data ada sejak tahun	Data yang masuk ke dalam Data Warehouse
2009	2009 - 2013 (5 Tahun)

8. Skema (Star)



9. Proses ETL

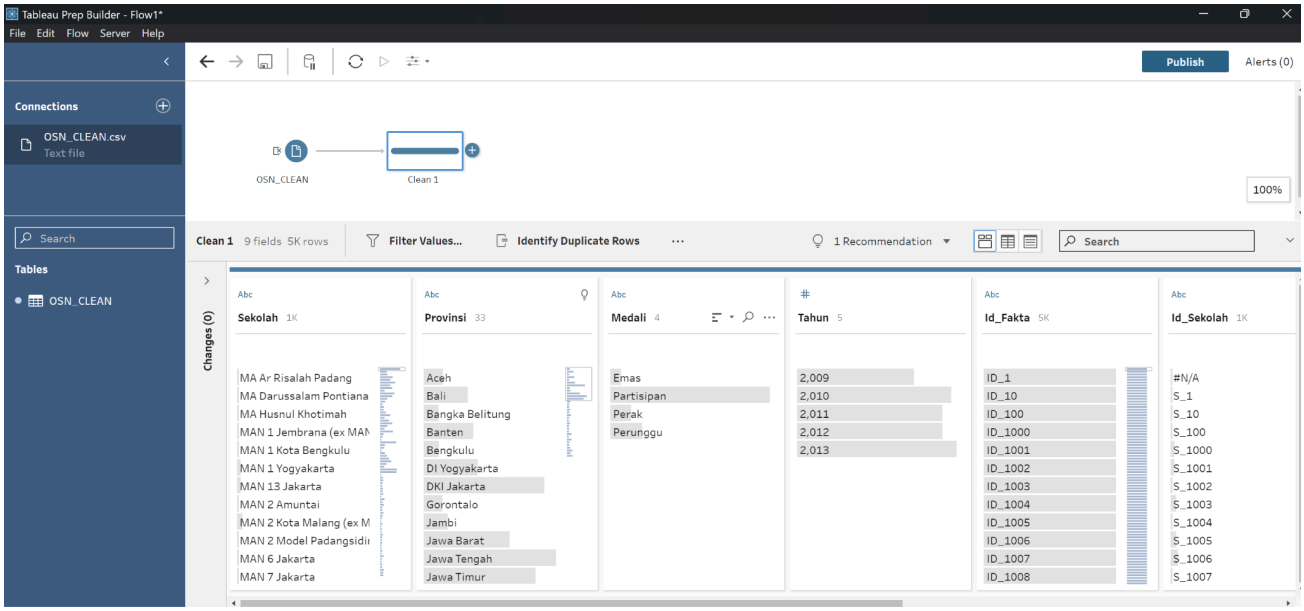
Tahap 1



Penjelasan :

Mengambil sumber data dari source. Dataset ini berisi informasi peserta Olimpiade Sains Nasional (OSN) tingkat SMP dan SMA dari tahun 2009 hingga 2024 di Indonesia. Data mencakup nama peserta, gender, asal sekolah, dll yang digunakan saat tahap kualifikasi OSN. Nama peserta dicatat dengan nama lengkap, namun jika ada kesamaan nama, pembeda diberikan dengan angka di belakang nama, seperti Kevin(1) dan Kevin(2).. Gender peserta juga ditentukan, dengan kode L melambangkan laki - laki dan P untuk perempuan.

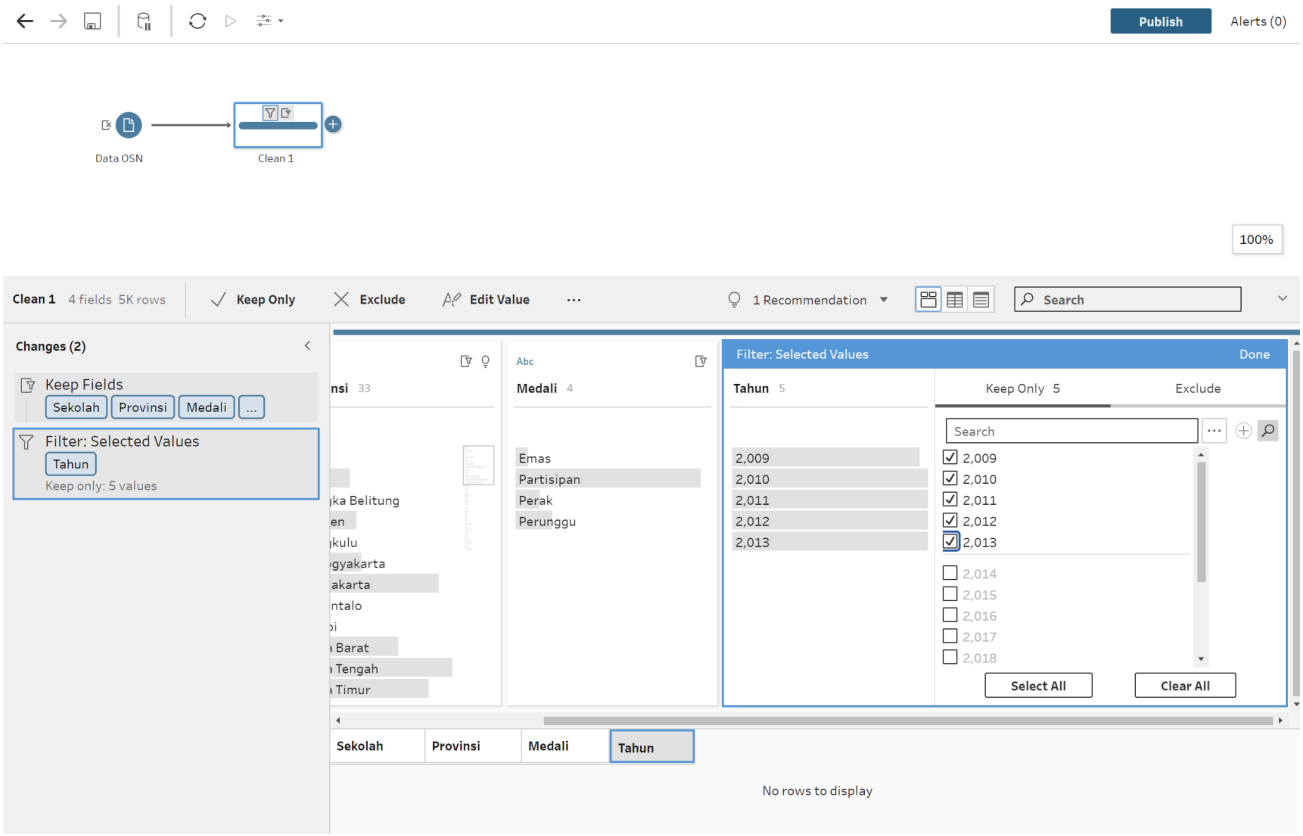
Tahap 2



Penjelasan :

Membersihkan data dengan menghapus kolom yang tidak ingin digunakan dengan menyisakan kolom yang akan dijadikan dimensi. Proses seleksi kolom telah dilakukan untuk menyisakan kolom yang relevan sebagai dimensi utama. Kolom yang dipertahankan antara lain Nama Peserta, Gender, Sekolah, Provinsi, Kab.Kota, Bidang, Jenjang Lomba, Jenjang Sekolah, Kelas, Medali, Prize Tambahan dan Tahun. Dengan menghapus kolom - kolom yang tidak diperlukan, dataset menjadi lebih ringkas dan siap untuk digunakan dalam analisis mendalam. Kolom - kolom tersebut dapat berperan sebagai dimensi dalam proses eksplorasi data, seperti memahami distribusi peserta berdasarkan gender, sekolah, atau provinsi, serta menganalisis perolehan medali dari waktu ke waktu. Langkah ini merupakan tahapan penting untuk memastikan data yang digunakan lebih fokus, bersih dan efisien dalam pengolahan lebih lanjut.

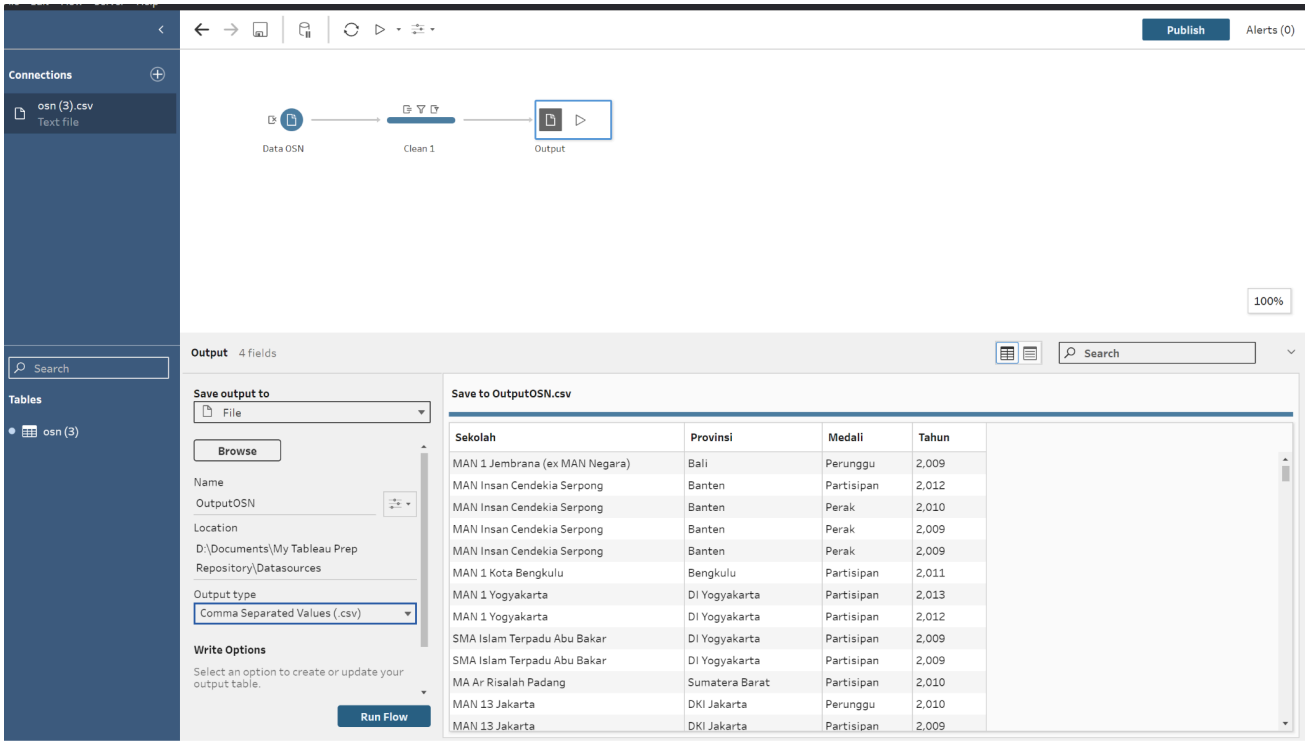
Tahap 3



Penjelasan :

Pada gambar diatas proses yang dilakukan adalah keep only fields. Field yang di keep adalah field sekolah, provinsi, medali, dan tahun. Kemudian dilakukan filter selected values yang digunakan untuk menyaringan data berdasarkan nilai tertentu. Dalam hal ini, penyaringan dilakukan pada kolom "Tahun" sehingga menyisakan tahun 2009 - 2013 saja.

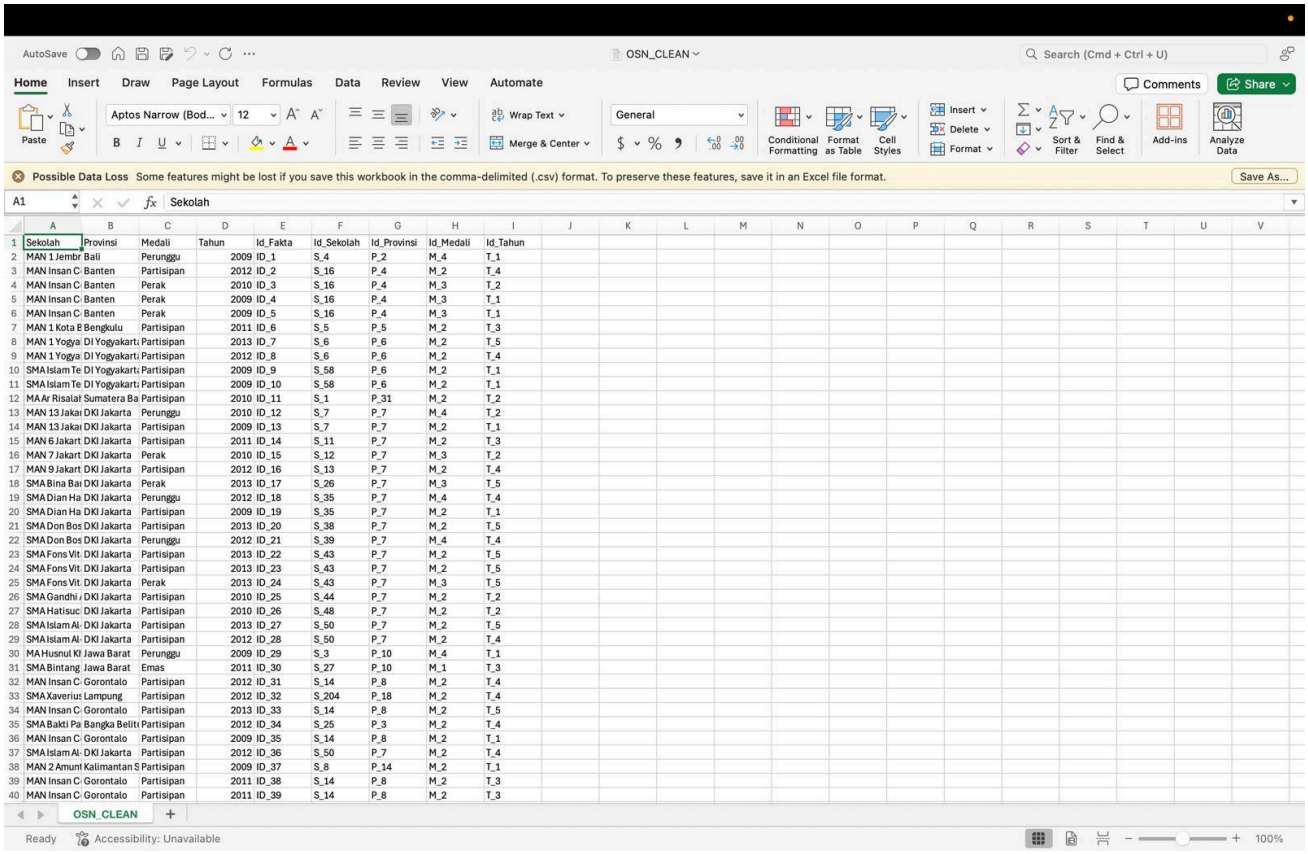
Tahap 4



Penjelasan :

Pada tahap ini kita akan menambahkan step output setelah kita melakukan clean pada field - field kita.

Tahap 5

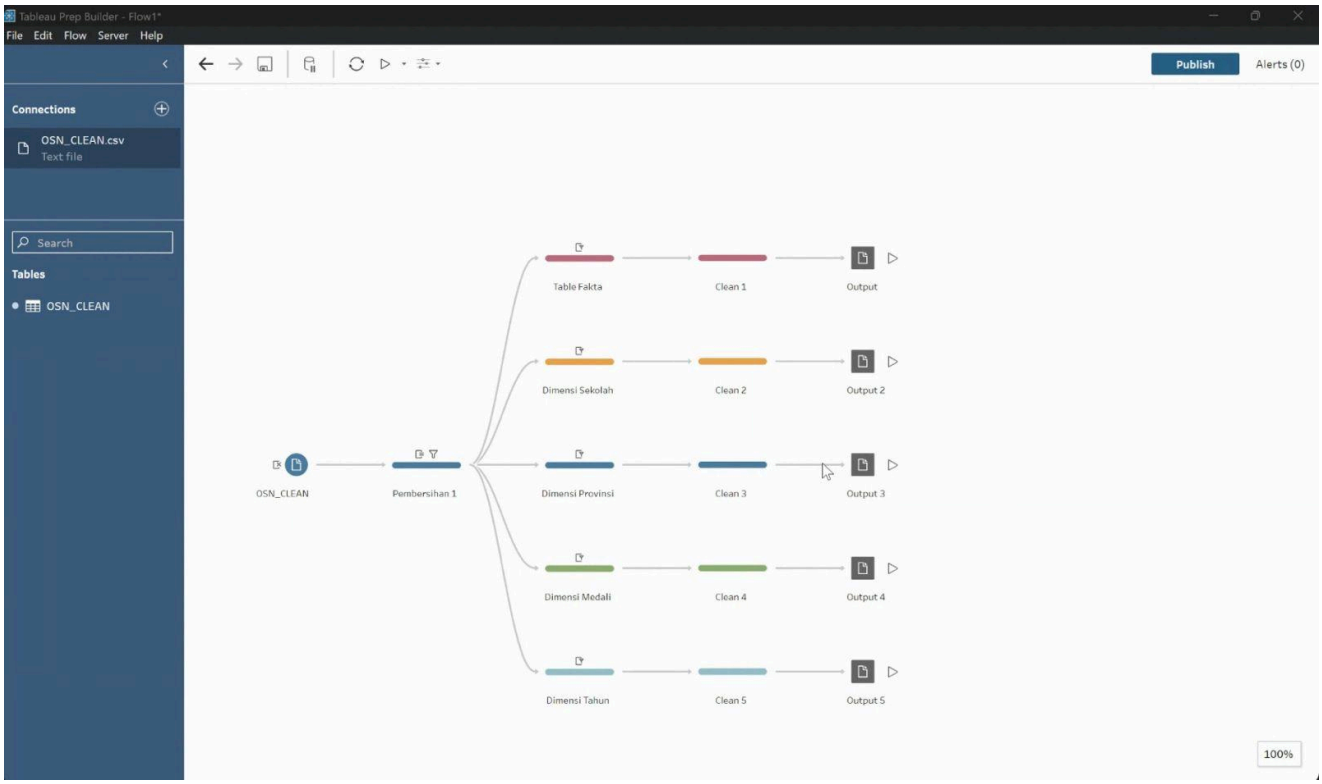


Sekolah	Provinsi	Medali	Tahun	Id_Fakta	Id_Sekolah	Id_Provinsi	Id_Medali	Id_Tahun
MAN 1 Jembr	Bali	Perunggu	2009	ID_1	S_4	P_2	M_4	T_1
MAN Insan C	Banten	Partisipan	2012	ID_2	S_16	P_4	M_2	T_4
MAN Insan C	Banten	Perak	2010	ID_3	S_16	P_4	M_3	T_2
MAN Insan C	Banten	Perak	2009	ID_4	S_16	P_4	M_3	T_1
MAN Insan C	Banten	Perak	2009	ID_5	S_16	P_4	M_3	T_1
MAN 1 Kota B	Bengkulu	Partisipan	2011	ID_6	S_5	P_5	M_2	T_3
MAN 1 Yogya	DI Yogyakarta	Partisipan	2013	ID_7	S_6	P_6	M_2	T_5
MAN 1 Yogya	DI Yogyakarta	Partisipan	2012	ID_8	S_6	P_6	M_2	T_4
SMA Islam Te	DI Yogyakarta	Partisipan	2009	ID_9	S_58	P_6	M_2	T_1
SMA Islam Te	DI Yogyakarta	Partisipan	2009	ID_10	S_58	P_6	M_2	T_1
MA Ar Risalat	Sumatera Ba	Partisipan	2010	ID_11	S_1	P_31	M_2	T_2
MAN 13 Jaka	DKI Jakarta	Perunggu	2010	ID_12	S_7	P_7	M_4	T_2
MAN 13 Jaka	DKI Jakarta	Partisipan	2009	ID_13	S_7	P_7	M_2	T_1
MAN 6 Jakart	DKI Jakarta	Partisipan	2011	ID_14	S_11	P_7	M_2	T_3
MAN 7 Jakart	DKI Jakarta	Perak	2010	ID_15	S_12	P_7	M_3	T_2
MAN 9 Jakart	DKI Jakarta	Partisipan	2012	ID_16	S_13	P_7	M_2	T_4
SMA Bina Ba	DKI Jakarta	Perak	2013	ID_17	S_26	P_7	M_3	T_5
SMA Dian Ha	DKI Jakarta	Perunggu	2012	ID_18	S_35	P_7	M_4	T_4
SMA Dian Ha	DKI Jakarta	Partisipan	2009	ID_19	S_35	P_7	M_2	T_1
SMA Don Bos	DKI Jakarta	Partisipan	2013	ID_20	S_38	P_7	M_2	T_5
SMA Don Bos	DKI Jakarta	Perunggu	2012	ID_21	S_39	P_7	M_4	T_4
SMA Fons Vit	DKI Jakarta	Partisipan	2013	ID_22	S_43	P_7	M_2	T_5
SMA Fons Vit	DKI Jakarta	Partisipan	2013	ID_23	S_43	P_7	M_2	T_5
SMA Fons Vit	DKI Jakarta	Perak	2013	ID_24	S_43	P_7	M_3	T_5
SMA Gandhi	DKI Jakarta	Partisipan	2010	ID_25	S_44	P_7	M_2	T_2
SMA Hatisuc	DKI Jakarta	Partisipan	2010	ID_26	S_48	P_7	M_2	T_2
SMA Islam Al	DKI Jakarta	Partisipan	2013	ID_27	S_50	P_7	M_2	T_5
SMA Islam Al	DKI Jakarta	Partisipan	2012	ID_28	S_50	P_7	M_2	T_4
MA Husnul Kl	Jawa Barat	Perunggu	2009	ID_29	S_3	P_10	M_4	T_1
SMA Bintang	Jawa Barat	Emas	2011	ID_30	S_27	P_10	M_1	T_3
MAN Insan C	Gorontalo	Partisipan	2012	ID_31	S_14	P_8	M_2	T_4
SMA Xaverius	Lampung	Partisipan	2012	ID_32	S_204	P_18	M_2	T_4
MAN Insan C	Gorontalo	Partisipan	2013	ID_33	S_14	P_8	M_2	T_5
SMA Bakti Pa	Bangka Beliti	Partisipan	2012	ID_34	S_25	P_3	M_2	T_4
MAN Insan C	Gorontalo	Partisipan	2009	ID_35	S_14	P_8	M_2	T_1
SMA Islam Al	DKI Jakarta	Partisipan	2012	ID_36	S_50	P_7	M_2	T_4
MAN 2 Amunt	Kalimantan S	Partisipan	2009	ID_37	S_8	P_14	M_2	T_1
MAN Insan C	Gorontalo	Partisipan	2011	ID_38	S_14	P_8	M_2	T_3
MAN Insan C	Gorontalo	Partisipan	2011	ID_39	S_14	P_8	M_2	T_3

Penjelasan :

Pada tahap ini kita akan melakukan penambahan ID berupa id_fakta, id_sekolah, id_provinsi, id_medali, dan id_tahun.

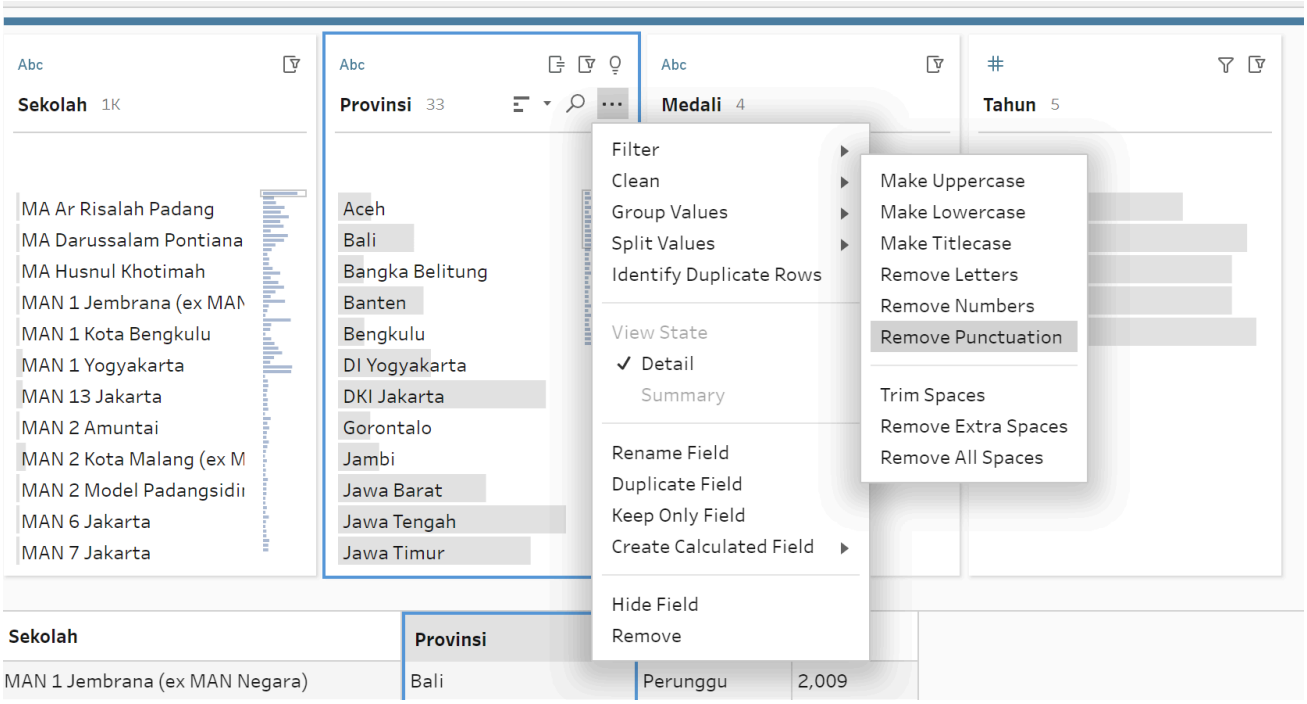
Tahap 6



Penjelasan :

Pada tahap ini menambahkan data yang sudah memiliki id ke dalam aplikasi Tableau Prep untuk melakukan pembersihan data lalu memisahkan data - data tersebut sesuai dengan dimensinya. Kemudian menambahkan step cleaning pada setiap dimensi untuk memastikan bahwa data sudah bersih dengan menggunakan fungsi preview. Pada tahap akhir menambahkan step output untuk mendapatkan file yang akan diproses pada tahap selanjutnya.

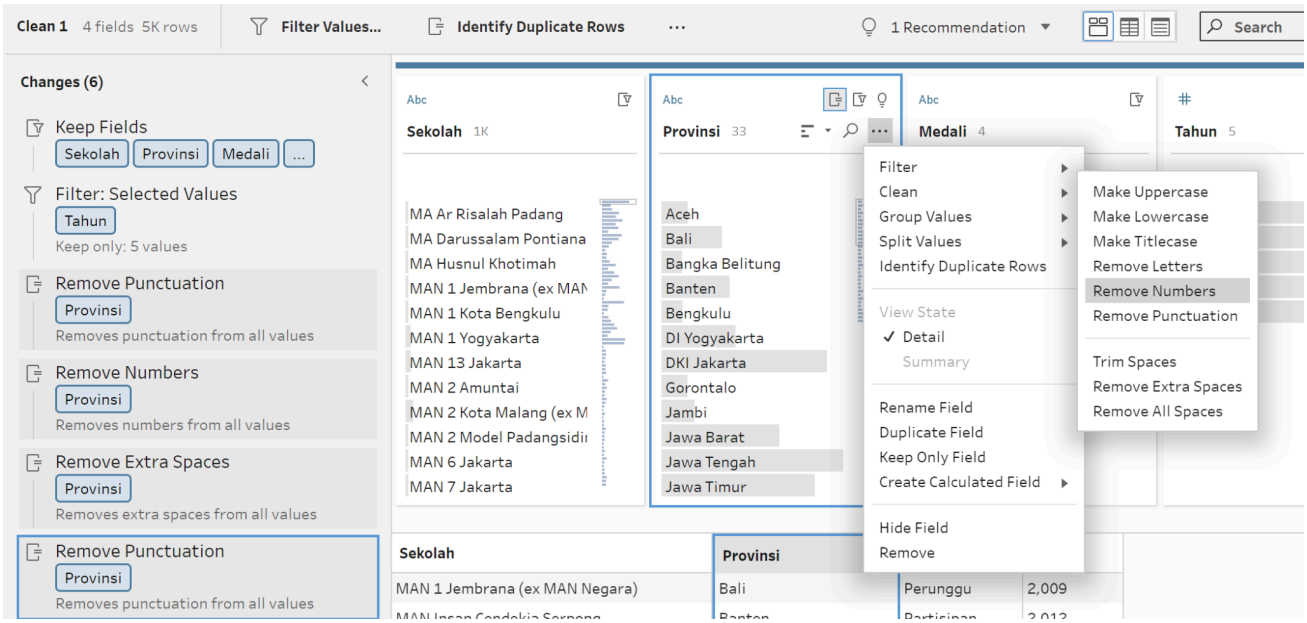
Tahap 7



Penjelasan :

Pada beberapa field, dilakukan pembersihan dengan memilih proses Remove Punctuation. Proses ini dilakukan untuk menghilangkan semua tanda baca pada field provinsi. Tanda baca yang dihilangkan berupa titik, koma, tanda tanya, tanda seru, titik koma, titik dua, tanda kutip, kurung dan lainnya.

Tahap 8

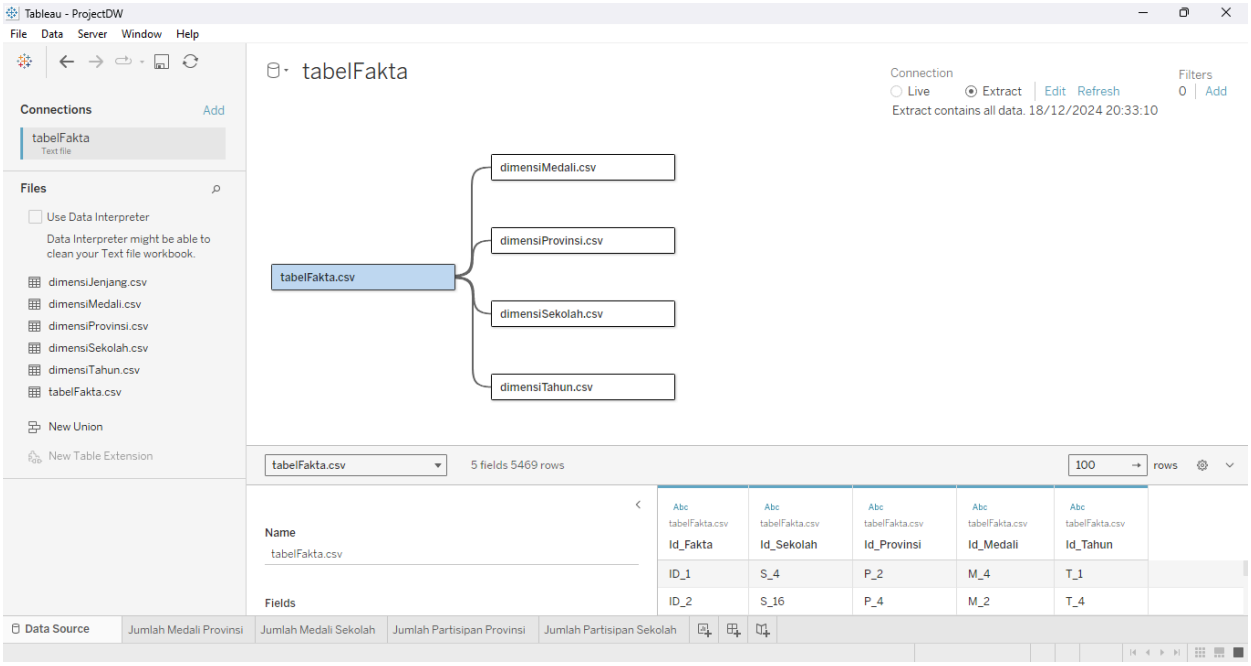


Penjelasan :

Pada tahap ini telah dilakukan remove punctuation yang digunakan untuk menghapus tanda baca berupa tanda titik, koma, tanda tanya, tanda seru, titik dua, titik koma, kurung, dan lainnya pada field . Pada fieldkemudian dilakukan remove numbers yang digunakan untuk menghapus nomor pada field . Kemudian pada field dilakukan remove extra spaces yang digunakan untuk menghapus spasi berlebih pada field.

10. Proses Pembuatan Grafik

a. Data Source



Penjelasan :

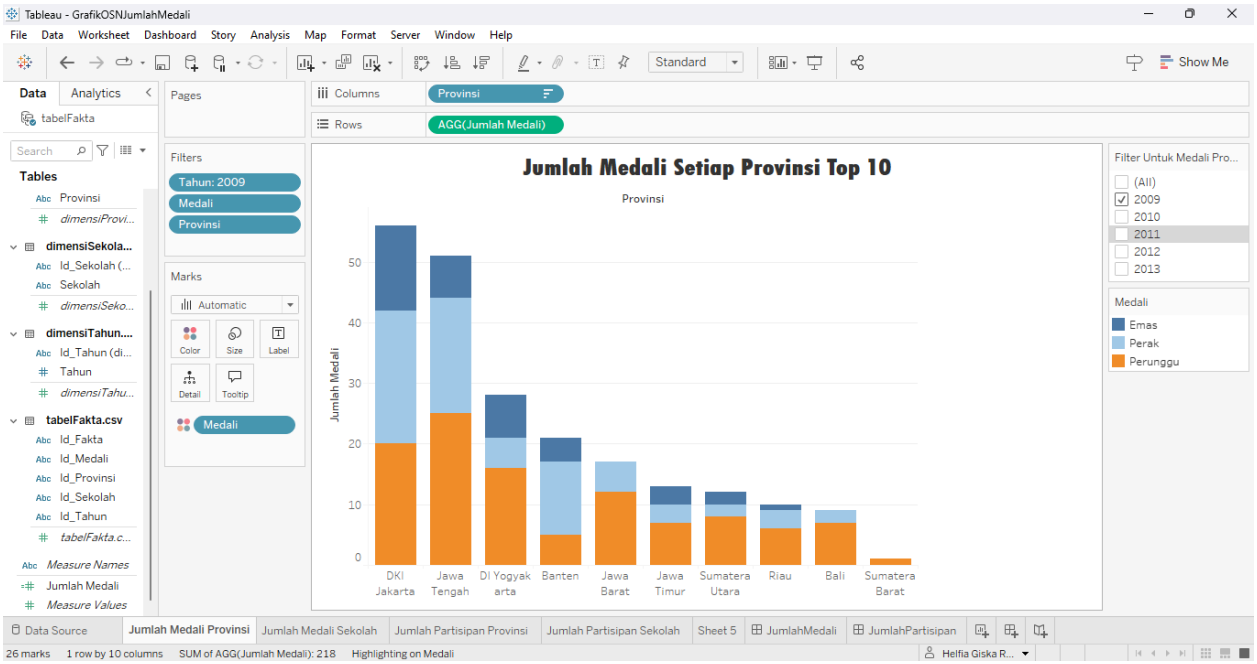
Tabel Fakta (tabelFakta.csv) merupakan pusat dari diagram. Di dalamnya terdapat data yang akan di analisis, seperti jumlah medali dan jumlah partisipan. Tabel fakta memiliki kunci utama yaitu ID yang digunakan untuk menghubungkan dengan dimensi - dimensi lainnya. Tabel fakta dihubungkan dengan setiap dimensi melalui kolom - kolom ID dalam tabel fakta.

Dimensi :

1. dimensiMedali.csv berisi informasi tentang jenis - jenis medali yang ada yaitu emas, perak, dan perunggu). Setiap jenis medali memiliki ID unik yang digunakan untuk menghubungkan dengan tabel fakta.
2. dimensiProvinsi.csv berisi informasi tentang provinsi - provinsi yang menjadi sumber data. Setiap provinsi memiliki ID unik yang digunakan untuk menghubungkan dengan tabel fakta.
3. dimensiSekolah.csv berisi informasi tentang sekolah - sekolah yang terlibat. Setiap sekolah memiliki ID unik yang digunakan untuk menghubungkan dengan tabel fakta.

4. dimensiTahun.csv berisi informasi tentang tahun data yang dikumpulkan. Tahun ini memiliki ID unik yang digunakan untuk menghubungkan dengan tabel fakta.

b. Jumlah Medali Setiap Provinsi Top 10



Penjelasan :

Gambar di atas menampilkan visualisasi data Jumlah Medali Tiap Provinsi Top 10. Visualisasi ini menyajikan informasi mengenai jumlah medali yang diperoleh oleh masing - masing provinsi dalam beberapa tahun terakhir, dengan fokus pada 10 provinsi dengan perolehan medali tertinggi.

Elemen - elemen dalam visualisasi :

- Sumbu X menampilkan nama nama provinsi yang menjadi fokus analisis.
- Sumbu Y menunjukkan jumlah medali yang diperoleh, dengan skala 0 hingga 80.
- Bar Chart atau grafik batang yang digunakan untuk membandingkan jumlah medali antar provinsi dan antar tahun. Setiap batang mewakili jumlah medali tertentu untuk provinsi dan tahun yang bersangkutan.
- Warna yang berbeda digunakan untuk membedakan jenis medali yaitu emas berwarna biru

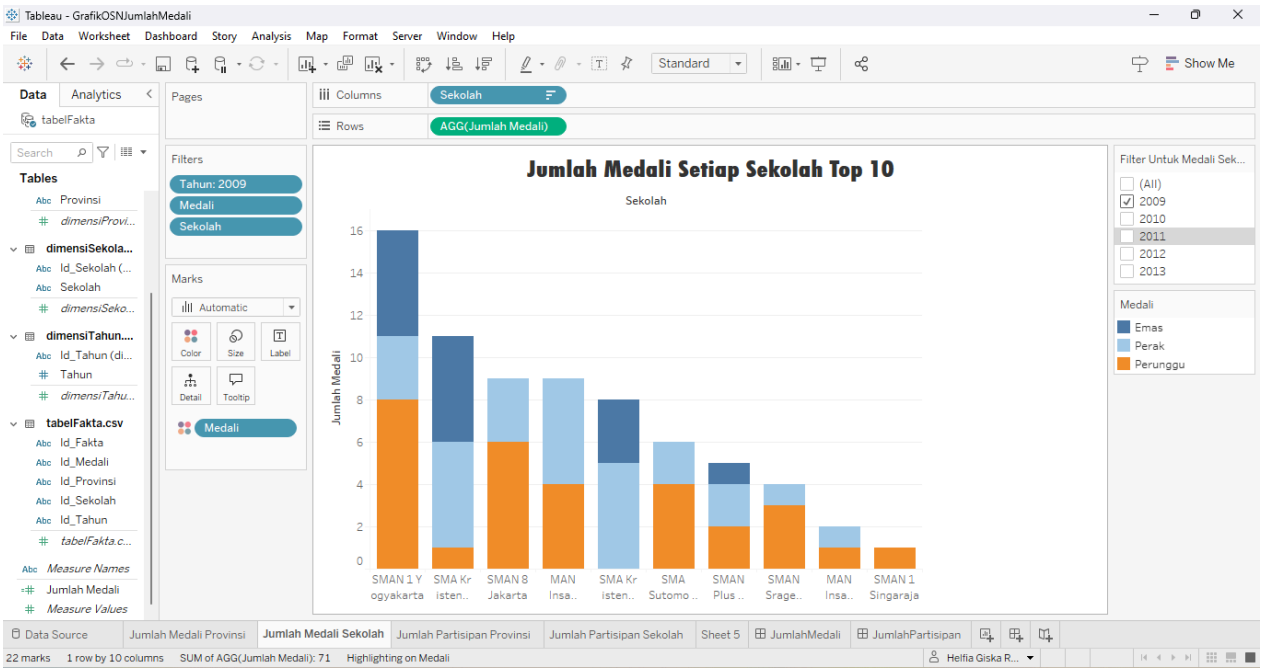
tua, perak berwarna biru muda, dan perunggu berwarna orange. Hal ini memudahkan dalam membandingkan perolehan medali masing - masing jenis.

- Terdapat filter tahun yang memungkinkan pengguna untuk melihat data pada tahun tertentu. Hal ini digunakan untuk menganalisis tren perolehan medali dari waktu ke waktu.

Kesimpulan awal :

- Beberapa provinsi secara konsisten meraih jumlah medali yang lebih tinggi dibandingkan provinsi lainnya. Hal ini mengindikasikan adanya faktor - faktor yang mendukung prestasi akademik di provinsi tersebut, seperti dukungan pemerintah, dan fasilitas belajar yang memadai.
- Dengan menggunakan filter tahun, kita dapat melihat bagaimana jumlah medali yang diperoleh masing - masing provinsi berubah dari waktu ke waktu. Hal ini dapat mengindikasikan adanya peningkatan atau penurunan prestasi di suatu provinsi.
- Visualisasi ini juga menunjukkan distribusi jenis medali yaitu emas, perak, dan perunggu untuk setiap provinsi. Hal ini dapat memberikan gambaran mengenai kekuatan dan kelemahan suatu provinsi dalam Olimpiade Sains Nasional.

c. Jumlah Medali Setiap Sekolah Top 10



Penjelasan :

Gambar di atas menampilkan visualisasi data Jumlah Medali Setiap Sekolah Top 10. Visualisasi ini dirancang untuk menyajikan informasi mengenai jumlah medali yang berhasil diraih oleh 10 sekolah dengan perolehan medali tertinggi dalam beberapa tahun terakhir.

Elemen - elemen dalam visualisasi :

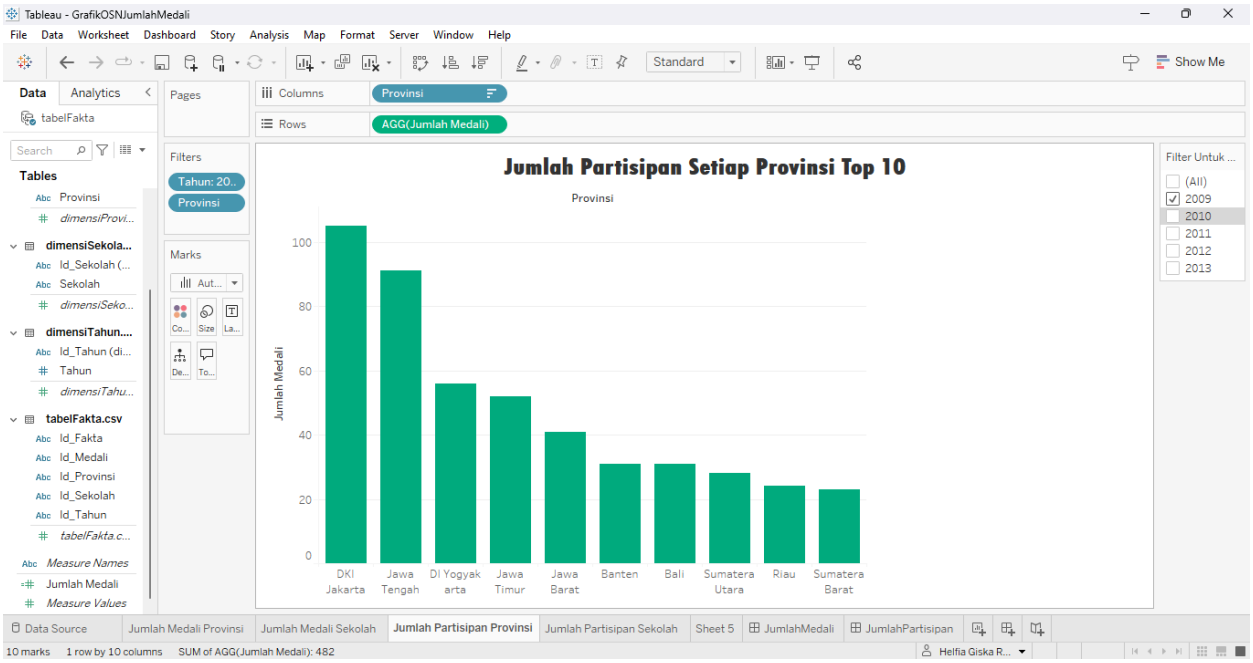
- Sumbu X menampilkan nama - nama 10 sekolah yang memiliki perolehan medali tertinggi.
- Sumbu Y menunjukkan jumlah medali yang diperoleh, dengan skala mulai dari 0 hingga 12.
- Bar Chart atau grafik batang yang mewakili jumlah total medali yang diraih oleh satu sekolah tertentu.
- Grafik batang tersebut dibagi menjadi 3 bagian dengan warna yang berbeda. Warna - warna tersebut mewakili jenis medali yaitu emas berwarna biru tua, perak berwarna biru muda, dan perunggu berwarna orange. Hal ini akan memudahkan user untuk membandingkan perolehan medali tersebut.

- Terdapat filter “Tahun” yang digunakan pengguna untuk melihat data pada tahun tertentu. Fitur ini sangat berguna untuk menganalisis tren perolehan medali dari waktu ke waktu.

Kesimpulan awal :

- Beberapa sekolah secara konsisten meraih jumlah medali yang lebih tinggi dibandingkan sekolah lainnya. Hal ini mengindikasikan adanya faktor - faktor yang mendukung prestasi akademik di sekolah tersebut, seperti fasilitas belajar yang memadai dan dukungan dari berbagai pihak.
- Setiap sekolah memiliki pola perolehan medali yang berbeda - beda. Ada sekolah yang dominan meraih medali emas, semestara sekolah lain mungkin lebih banyak meraih medali perak atau perunggu.

d. Jumlah Partisipan Setiap Provinsi Top 10



Penjelasan :

Gambar di atas menampilkan visualisasi data Jumlah Partisipan Setiap Provinsi Top 10. Visualisasi ini dirancang untuk menyajikan informasi mengenai jumlah partisipan dari masing - masing provinsi dengan jumlah partisipan tertinggi dalam beberapa tahun terakhir.

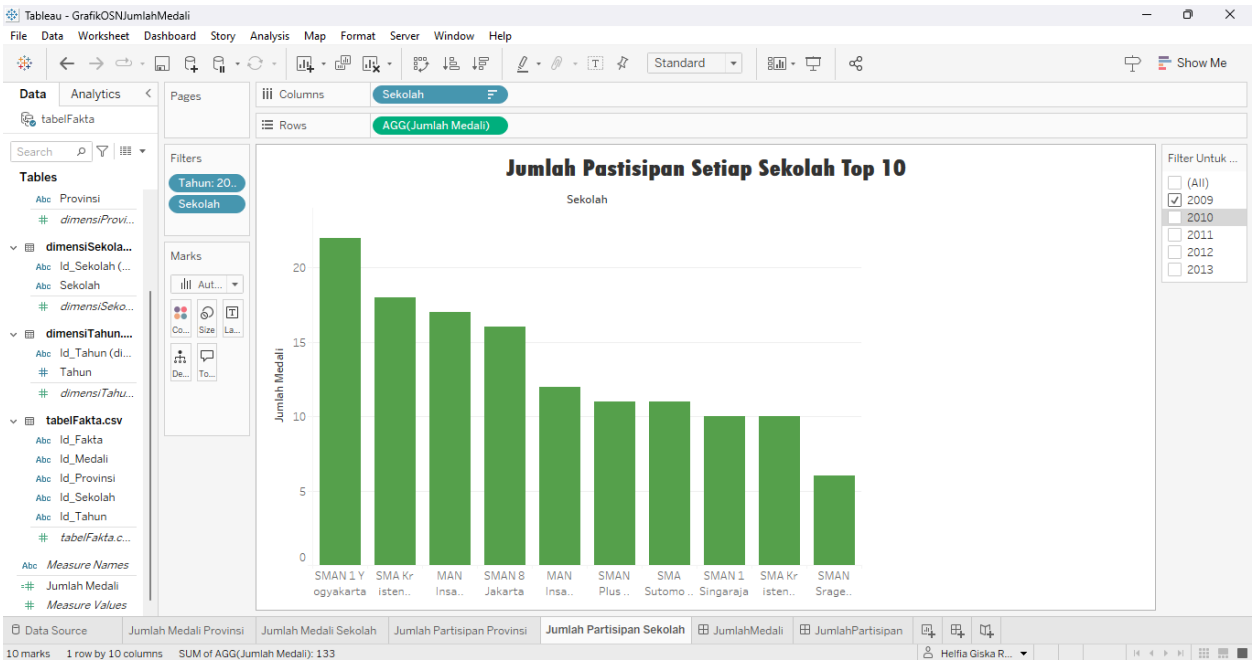
Elemen - elemen dalam visualisasi :

- Sumbu X menampilkan nama nama provinsi yang memiliki jumlah partisipan tertinggi.
- Sumbu Y menunjukkan jumlah total partisipan dari masing - masing provinsi dari skala 0 - 100.
- Bar Chart atau grafik batang yang mewakili jumlah total partisipan dari satu provinsi tertentu. Tinggi batang menunjukkan banyaknya partisipan.
- Semua batang menggunakan warna hijau untuk memberikan tampilan yang konsisten dan mudah dibaca.
- Terdapat filter “Tahun” yang digunakan pengguna untuk melihat data pada tahun tertentu. Fitur ini sangat berguna untuk menganalisis tren jumlah partisipan dari waktu ke waktu.

Kesimpulan awal :

- Beberapa provinsi secara konsisten memiliki jumlah partisipan yang jauh lebih tinggi dibandingkan provinsi lainnya. Ini mengindikasikan adanya faktor - faktor yang mendorong partisipasi aktif di provinsi tersebut, seperti dukungan pemerintah dan fasilitas yang memadai.
- Grafik ini memberikan gambaran jelas tentang peringkat setiap provinsi berdasarkan jumlah partisipannya. Provinsi dengan batang tertinggi memiliki jumlah partisipan terbanyak.

e. Jumlah Partisipan Setiap Sekolah Top 10



Penjelasan :

Gambar di atas menampilkan visualisasi data Jumlah Partisipan Setiap Sekolah Top 10. Visualisasi ini dirancang untuk menyajikan informasi mengenai jumlah partisipan dari masing - masing sekolah dengan jumlah partisipan tertinggi dalam beberapa tahun terakhir.

Elemen - elemen dalam visualisasi :

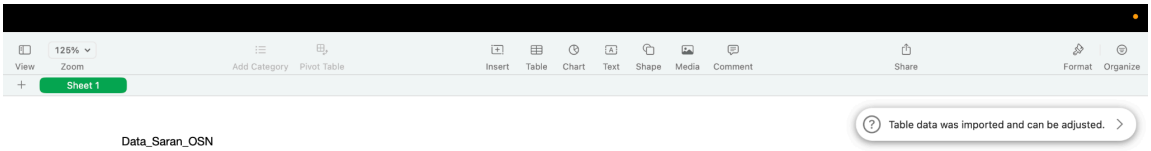
- Sumbu X menampilkan nama - nama 10 sekolah yang memiliki jumlah partisipan tertinggi setiap tahun.
- Sumbu Y menunjukkan jumlah total partisipan dari masing - masing sekolah dari skala 0 - 20.
- Bar Chart atau grafik batang yang mewakili jumlah total partisipan dari satu sekolah tertentu. Tinggi batang menunjukkan banyaknya partisipan.
- Semua diagram batang menggunakan warna hijau untuk memberikan tampilan yang konsisten dan mudah dibaca.
- Terdapat filter “Tahun” yang digunakan pengguna untuk melihat data pada tahun tertentu. Fitur ini sangat berguna untuk menganalisis tren jumlah partisipan dari waktu ke waktu.

Kesimpulan awal :

- Beberapa sekolah secara konsisten memiliki jumlah partisipan yang jauh lebih tinggi dibandingkan sekolah lainnya di tahun 2009. Ini mengindikasikan adanya faktor - faktor yang mendorong partisipasi aktif di sekolah tersebut, seperti dukungan dari sekolah, minat siswa, atau adanya kegiatan ekstrakurikuler yang menarik.
- Grafik ini memberikan gambaran jelas tentang peringkat setiap sekolah berdasarkan jumlah partisipannya setiap tahun. Sekolah dengan batang tertinggi memiliki jumlah partisipan terbanyak.

11. Sentimen Analisis

a. Melihat isi dari data saran



Data_Saran_OSN

Saran
Informasi acara kurang jelas, harap diperbaiki.
Kurangnya sarana teknologi untuk membantu lomba.
Moderator acara sangat profesional, sangat memuaskan.
Jadwal acara terlalu padat, mohon disesuaikan.
Sistem lomba sudah baik, tetapi perlu lebih transparan.
Harap ada waktu istirahat yang cukup di tengah lomba.
Akomodasi untuk peserta dari luar kota perlu diperbaiki.
Sangat luar biasa, semua berjalan lancar.
Kurangnya sarana teknologi untuk membantu lomba.
Panitia sangat ramah dan membantu.
Sistem lomba sudah baik, tetapi perlu lebih transparan.
Moderator acara sangat profesional, sangat memuaskan.
Hadiah perlu ditingkatkan agar lebih menarik.
Acara ini sangat bermanfaat, tetapi kurang promosi.
Panitia sangat ramah dan membantu.
Kurangnya sarana teknologi untuk membantu lomba.
Peserta terlalu banyak dari kota besar, mohon diperhatikan.
Kurangnya sarana teknologi untuk membantu lomba.
Hadiah perlu ditingkatkan agar lebih menarik.
Kurangnya sarana teknologi untuk membantu lomba.
Acara ini sangat bermanfaat, tetapi kurang promosi.
Sistem lomba sudah baik, tetapi perlu lebih transparan.
Acara ini sangat bermanfaat, tetapi kurang promosi.
Semua berjalan dengan lancar, sangat memuaskan.
Semua berjalan dengan lancar, sangat memuaskan.
Acara ini sangat bermanfaat, tetapi kurang promosi.
Jadwal acara terlalu padat, mohon disesuaikan.
Pengumuman hasil lomba perlu dipercepat.
Hadiah perlu ditingkatkan agar lebih menarik.

Penjelasan :

Dataset ini terdiri dari 100 kolom mengenai survei peserta OSN yang akan dianalisis menggunakan Python. Tujuannya adalah untuk mengidentifikasi sentimen positif, negatif, atau netral terhadap berbagai aspek pada Olimpiade Sains Nasional. Proses analisis akan melibatkan pembersihan data, pemprosesan data sebelum dilakukan sentimen analisis, dan pembuatan visualisasi terhadap hasil dari sentimen analisis tersebut menggunakan pustaka dari python.

b. Proses Menambahkan Library

saranCleaning.py > ...

```
1  import pandas as pd
2  from transformers import pipeline
3  import matplotlib.pyplot as plt
4  import seaborn as sns
5  from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
6  import re
```

Penjelasan untuk Library :

- **Pandas (pd):**
  - Digunakan untuk membaca, memproses, dan menyimpan data dalam bentuk tabel seperti file CSV atau Excel.
  - Menyediakan metode untuk manipulasi data, misalnya menambah kolom, filter data, dan lainnya.
- **Transformers:**
  - Library dari Hugging Face yang menyediakan akses ke berbagai model NLP, termasuk analisis sentimen.
  - Memungkinkan kita menggunakan model pra-latih seperti Indonesian RoBERTa untuk klasifikasi teks dalam Bahasa Indonesia.
- **Matplotlib.pyplot (plt):**
  - Digunakan untuk membuat grafik atau visualisasi data.
  - Membantu dalam membuat plot seperti bar chart, line chart, dll.
- **Seaborn (sns):**
  - Library visualisasi data yang dibangun di atas Matplotlib.
  - Memberikan fitur tambahan untuk membuat grafik yang lebih estetis dan informatif.

- **Sastrawi.Stemmer.StemmerFactory:**
  - Library khusus untuk stemming teks dalam Bahasa Indonesia.
  - Digunakan untuk mengubah kata-kata menjadi bentuk dasar, misalnya “berlari” menjadi “lari”.
- **Re:**
  - Library Python untuk manipulasi teks menggunakan pola reguler (regex).
  - Berguna untuk membersihkan teks dari angka, tanda baca, atau karakter tidak diinginkan lainnya.

c. Proses Menambahkan Data

```
# 1. Load Data
data_path = "/Users/haziqdafren/Downloads/Data_Saran_OSN.csv"
df = pd.read_csv(data_path)

# Tampilkan beberapa data
print("Data awal:")
print(df.head())
```

Penjelasan :

- **data_path:** Variabel berisi lokasi file CSV.
- **pd.read_csv(data_path):** Membaca file CSV ke dalam DataFrame df.
- **df.head():** Menampilkan 5 baris pertama data untuk memeriksa isi file.

d. Proses Pembersihan Data

```
21  # Fungsi preprocessing
    Tabnine | Edit | Test | Explain | Document | Ask
22  def preprocess_text(text):
23      text = text.lower()  # Case folding
24      text = re.sub(r'[^\w\s]', '', text)  # Menghapus tanda baca
25      text = re.sub(r'\d+', '', text)  # Menghapus angka
26      return stemmer.stem(text)  # Stemming
27
28  # Tambahkan kolom teks yang sudah diproses
29  df['Saran_Processed'] = df['Saran'].apply(preprocess_text)
30
```

Penjelasan Langkah Preprocessing:

1. **text.lower():** Semua huruf diubah menjadi kecil agar seragam.
2. **re.sub(r'[^\w\s]', '', text):** Menghapus tanda baca seperti titik, koma, tanda seru.
3. **re.sub(r'\d+', '', text):** Menghapus angka dari teks.
4. **stemmer.stem(text):** Mengubah kata ke bentuk dasarnya menggunakan Sastrawi (contoh: "bermain" → "main").
5. **apply(preprocess_text):** Menerapkan fungsi preprocess_text pada setiap baris di kolom Saran.
6. Kolom baru Saran_Processed akan berisi teks yang sudah dibersihkan.

e. Penambahan Model untuk Sentimen Analisis

```
31  # 3. Analisis Sentimen
32  # Load model IndoBERT
33  sentiment_analyzer = pipeline("sentiment-analysis", model="w11wo/indonesian-roberta-base-sentiment-classifier")
34
```

Penjelasan :

w11wo/indonesian-roberta-base-sentiment-classifier: Model NLP yang dilatih untuk analisis sentimen dalam Bahasa Indonesia.

- **Pipeline:** Menggunakan model pra-latih untuk analisis sentimen.
- **Sentiment-analysis:** Tugas NLP untuk menentukan apakah teks bersifat positif, negatif, atau netral.

f. Menjalankan Sentimen Analisis

```
def analyze_sentiment(text):
    try:
        result = sentiment_analyzer(text)[0]
        return result['label'], result['score']
    except Exception as e:
        print(f"Error analyzing sentiment: {e}")
        return "UNKNOWN", 0.0

# Analisis sentimen untuk data
df['Sentimen'], df['Skor'] = zip(*df['Saran_Processed'].apply(analyze_sentiment))

```

Penjelasan :

- **sentiment_analyzer(text)[0]:** Menganalisis teks dan mengembalikan hasil dalam bentuk dictionary ({'label': 'POSITIVE', 'score': 0.98}).
- **result['label']:** Label sentimen (positif, negatif, netral).
- **result['score']:** Skor keyakinan model terhadap prediksi.
- **apply(analyze_sentiment):** Menerapkan analisis sentimen pada setiap teks.
- **zip(*...):** Memisahkan hasil label (sentimen) dan skor ke kolom masing-masing.

g. Proses Menghapus kata berdasarkan stopwords

```
# 4. Menyimpan kata kunci
Tabnine | Edit | Test | Explain | Document | Ask
def extract_keywords(text):
    # Mengambil kata-kata yang berkontribusi pada sentimen
    return [word for word in text.split() if word not in stopwords]

```

Penjelasan :

- **text.split():** Memecah teks menjadi daftar kata.
- **if word not in stopwords:** Hanya memilih kata yang tidak termasuk dalam daftar stopword.

h. Menambahkan Stopword

```
52  # Memuat daftar stopword dari file CSV
53  stopwords_df = pd.read_csv('/Users/haziqdafren/Downloads/stopword.csv')
54  stopwords = set(stopwords_df['stopword'].tolist())
55
56  # Tambahkan kolom kata kunci
57  df['Kata_Kunci'] = df['Saran_Processed'].apply(extract_keywords)
```

Penjelasan :

stopwords: Kumpulan kata-kata umum seperti “dan”, “atau”, “yang”, dll.

i. Menyimpan Hasil ke dalam File

```
59  # Simpan hasil ke file CSV
60  output_path = "/Users/haziqdafren/Downloads/data_saran_sentimen.csv"
61  df.to_csv(output_path, index=False)
62  print(f"Hasil analisis sentimen disimpan di: {output_path}")
63
```

Penjelasan :

df.to_csv(output_path): Menyimpan DataFrame yang sudah diproses ke file CSV.

j. Proses Visualisasi

```
# 5. Visualisasi Hasil
# Hitung distribusi sentimen
sentiment_counts = df['Sentimen'].value_counts()

# Plot bar chart
plt.figure(figsize=(8, 5))
bar_plot = sns.barplot(x=sentiment_counts.index, y=sentiment_counts.values, palette="viridis")
plt.title("Distribusi Sentimen", fontsize=16)
plt.xlabel("Sentimen", fontsize=12)
plt.ylabel("Jumlah", fontsize=12)

# Tambahkan angka di atas setiap batang
for p in bar_plot.patches:
    bar_plot.annotate(f'{int(p.get_height())}',
                      (p.get_x() + p.get_width() / 2., p.get_height()),
                      ha='center', va='bottom',
                      fontsize=12, color='black',
                      xytext=(0, 5),  # Offset untuk angka
                      textcoords='offset points')
```

Penjelasan :

- **value_counts():** Menghitung jumlah teks untuk setiap jenis sentimen (positif, negatif, netral).
- **sns.barplot:** Membuat grafik batang untuk menampilkan jumlah sentimen.
- **p.get_height():** Mendapatkan nilai batang (jumlah sentimen).
- **annotate:** Menambahkan angka di atas setiap batang.

k. Menyimpan Grafik ke dalam File

```
84  # Simpan grafik ke file
85  graph_path = "/Users/haziqdafren/Downloads/sentimen_chart.png"
86  plt.savefig(graph_path)
87  print(f"Grafik sentimen disimpan di: {graph_path}")
88
89  # Tampilkan grafik
90  plt.show()
91
```

Penjelasan :

- **plt.savefig(graph_path)**: Menyimpan grafik ke file gambar.

l. Kesimpulan

Alur kerja ini dirancang untuk menganalisis sentimen dari data teks berbahasa Indonesia dengan pendekatan yang terstruktur dan terintegrasi. Proses dimulai dengan membaca data dari file CSV menggunakan pandas, di mana data diolah untuk menampilkan isi awal dan mempersiapkan kolom yang diperlukan. Selanjutnya, teks diproses melalui serangkaian langkah preprocessing, termasuk case folding, penghapusan tanda baca dan angka, serta stemming menggunakan library Sastrawi untuk mengubah kata menjadi bentuk dasarnya. Setelah teks dibersihkan, model Indonesian RoBERTa dari library transformers digunakan untuk menganalisis sentimen setiap teks, menghasilkan label sentimen (positif, negatif, atau netral) dan skor kepercayaan model terhadap prediksi. Data juga diperkaya dengan kolom kata kunci, yang diekstrak dari teks bersih dengan menghapus stopword menggunakan daftar stopword yang dimuat secara terpisah.

Hasil analisis ini kemudian disimpan ke file CSV untuk dokumentasi dan penggunaan lebih lanjut. Selain itu, data dianalisis secara visual menggunakan grafik distribusi sentimen yang dibuat dengan Matplotlib dan Seaborn, memberikan wawasan kuantitatif tentang jumlah sentimen di dalam dataset. Grafik ini diperkaya dengan anotasi untuk mempermudah interpretasi hasil. Dengan demikian, alur kerja ini tidak hanya membersihkan dan memproses teks, tetapi

juga memberikan analisis mendalam melalui pengelolaan data, analisis sentimen, ekstraksi kata kunci, dan visualisasi data yang informatif. Alur ini sangat fleksibel dan dapat diperluas untuk menangani berbagai jenis data teks dalam Bahasa Indonesia.