


LAPORAN TUGAS

Data Acquisition dan Cleaning



Matakuliah	TI0263 – Kecerdasan Buatan (Grup A) - Genap 2022/2023
Dosen Pengampu	Matahari Bhakti Nendya, S.Kom., M.T
Nama Kelompok	Solo Player
Anggota Kelompok	1. <i>Revyanto Tangiloang Sukardi (71200610)</i> 
Deklarasi	Dengan ini saya menyatakan bahwa tugas ini merupakan hasil karya kelompok saya, tidak ada manipulasi data serta bukan merupakan plagiasi dari karya orang lain.



UNIVERSITAS KRISTEN DUTA WACANA
Fakultas Teknologi Informasi
Program Studi Informatika



BAB 1

Latar Belakang

Pengantar

Bagian pertama dalam pengembangan aplikasi berbasis AI adalah data acquisition dan cleaning. Tahapan data acquisition dan cleaning perlu dilakukan untuk mendapatkan kualitas dan konsistensi data yang akan digunakan dalam pengembangan proses inferensi.

BAB 2

Pembahasan

Jawab

1. Model data

Pemodelan yang digunakan :

- SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors)
- LSTM (Long Short-Term Memory)

2. Sumber Data

- Sumber
Data yang digunakan bersumber dari Kaggle dengan rentang waktu 12 tahun yaitu 2010 - 2021
- Lokasi
DKI Jakarta
- Jenis Polutan
Jenis polutannya antara lain : PM10 (Particulate Matter 10), SO2 (Sulfur Dioxide), CO (Carbon Monoxide), NO2 (Nitrogen Dioxide), dan O3 (Ozone)

3. Data Acquisition

- Load Data

	tanggal	pm10	so2	co	o3	no2	max	critical	categori
0	01/01/2010	60.0	4.0	73.0	27.0	14.0	73.0	CO	SEDANG
1	02/01/2010	32.0	2.0	16.0	33.0	9.0	33.0	O3	BAIK
2	03/01/2010	27.0	2.0	19.0	20.0	9.0	27.0	PM10	BAIK
3	04/01/2010	22.0	2.0	16.0	15.0	6.0	22.0	PM10	BAIK
4	05/01/2010	25.0	2.0	17.0	15.0	8.0	25.0	PM10	BAIK
...
4378	27/12/2021	52.0	61.0	15.0	20.0	15.0	71.0	PM25	SEDANG
4379	28/12/2021	51.0	53.0	15.0	18.0	13.0	65.0	PM25	SEDANG
4380	29/12/2021	31.0	54.0	10.0	24.0	11.0	54.0	SO2	SEDANG
4381	30/12/2021	55.0	53.0	16.0	23.0	14.0	71.0	PM25	SEDANG
4382	31/12/2021	62.0	52.0	23.0	20.0	14.0	85.0	PM25	SEDANG

4383 rows × 9 columns

- Tipe Data

```
Data columns (total 9 columns):
#   Column      Non-Null Count  Dtype
---  -
0   tanggal      4383 non-null     object
1   pm10          4171 non-null     float64
2   so2           4208 non-null     float64
3   co            4233 non-null     float64
4   o3            4180 non-null     float64
5   no2           4190 non-null     float64
6   max           4374 non-null     float64
7   critical      4273 non-null     object
8   kategori      4383 non-null     object
dtypes: float64(6), object(3)
memory usage: 308.3+ KB
```

4. Data Cleaning

- Melihat Data yang valuenya null atau kosong

```
# melihat null value
df.isnull().sum()

tanggal      0
pm10         212
so2          175
co           150
o3           203
no2          193
max           9
critical     110
kategori      0
dtype: int64
```

- Mengisi Data yang valuenya null atau kosong

```
# mengisi nilai null dari setiap kolom

data["pm10"] = data["pm10"].interpolate(method='linear', limit_direction='forward', axis=0)
data["so2"] = data["so2"].interpolate(method='linear', limit_direction='forward', axis=0)
data["co"] = data["co"].interpolate(method='linear', limit_direction='forward', axis=0)
data["o3"] = data["o3"].interpolate(method='linear', limit_direction='forward', axis=0)
data["no2"] = data["no2"].interpolate(method='linear', limit_direction='forward', axis=0)
```

- Menisi nilai null pada kolom kategori

	index	tanggal	pm10	so2	co	o3	no2	max	critical	kategori
0	175	25/06/2010	58.666667	18.333333	30.000000	43.0	22.333333	0.0	NaN	TIDAK ADA DATA
1	215	04/08/2010	68.500000	14.500000	33.333333	52.0	21.500000	0.0	NaN	TIDAK ADA DATA
2	304	01/11/2010	36.000000	13.000000	24.000000	30.5	11.000000	0.0	NaN	TIDAK ADA DATA
3	419	24/02/2011	44.333333	14.666667	25.333333	37.0	11.333333	0.0	NaN	TIDAK ADA DATA
4	420	25/02/2011	41.666667	14.333333	24.666667	33.0	11.666667	0.0	NaN	TIDAK ADA DATA
...
105	3621	01/12/2019	54.500000	19.500000	16.000000	60.5	11.500000	0.0	NaN	TIDAK ADA DATA
106	3631	11/12/2019	57.666667	21.500000	19.333333	49.0	13.000000	0.0	NaN	TIDAK ADA DATA
107	3642	22/12/2019	55.500000	22.500000	24.000000	61.0	14.000000	0.0	NaN	TIDAK ADA DATA
108	3648	28/12/2019	52.000000	26.000000	23.333333	43.0	10.333333	0.0	NaN	TIDAK ADA DATA
109	3649	29/12/2019	52.000000	24.000000	25.666667	45.0	9.666667	0.0	NaN	TIDAK ADA DATA

110 rows x 10 columns

5. Evaluasi Data

	precision	recall	f1-score	support
BAIK	1.00	1.00	1.00	316
SEDANG	1.00	1.00	1.00	921
TIDAK SEHAT	1.00	0.98	0.99	45
accuracy			1.00	1282
macro avg	1.00	0.99	1.00	1282
weighted avg	1.00	1.00	1.00	1282

6. Mengubah Data harian menjadi mingguan

	pm10	so2	co	o3	no2
tanggal					
2010-01-01	60.0	4.0	73.0	27.00	14.0
2010-01-02	41.0	7.0	30.0	17.00	13.0
2010-01-03	60.0	5.0	39.0	44.00	19.0
2010-01-04	59.0	5.0	46.0	72.75	21.0
2010-01-05	60.0	15.5	24.0	28.00	10.0
...
2021-12-27	52.0	61.0	15.0	20.00	15.0
2021-12-28	51.0	53.0	15.0	18.00	13.0
2021-12-29	31.0	54.0	10.0	24.00	11.0
2021-12-30	55.0	53.0	16.0	23.00	14.0
2021-12-31	62.0	52.0	23.0	20.00	14.0

4383 rows × 5 columns

	pm10	so2	co	o3	no2
tanggal					
2010-01-04	55.000000	5.250000	47.000000	40.187500	16.750000
2010-01-11	51.571429	16.357143	26.714286	43.785714	16.714286
2010-01-18	37.857143	6.571429	30.142857	19.142857	14.000000
2010-01-25	48.285714	4.142857	41.714286	20.571429	15.571429
2010-02-01	47.285714	6.285714	33.142857	25.142857	14.142857
...
2021-12-06	44.857143	24.571429	13.714286	19.142857	21.428571
2021-12-13	46.428571	31.000000	9.857143	19.000000	16.571429
2021-12-20	53.428571	40.857143	18.142857	23.142857	13.000000
2021-12-27	49.285714	53.571429	17.571429	20.428571	14.142857
2022-01-03	49.750000	53.000000	16.000000	21.250000	13.000000

627 rows × 5 columns

7. Membagi data menjadi data train dan data test

	pm10	so2	co	o3	no2
tanggal					
2010-01-04	55.000000	5.250000	47.000000	40.187500	16.750000
2010-01-11	51.571429	16.357143	26.714286	43.785714	16.714286
2010-01-18	37.857143	6.571429	30.142857	19.142857	14.000000
2010-01-25	48.285714	4.142857	41.714286	20.571429	15.571429
2010-02-01	47.285714	6.285714	33.142857	25.142857	14.142857
...
2019-07-15	56.371429	23.238095	18.904762	66.428571	15.619048
2019-07-22	60.714286	21.571429	23.428571	58.714286	18.142857
2019-07-29	63.857143	19.285714	21.142857	54.285714	17.571429
2019-08-05	52.285714	16.857143	27.285714	75.142857	17.500000
2019-08-12	54.914286	20.000000	18.428571	62.000000	13.000000

502 rows × 5 columns

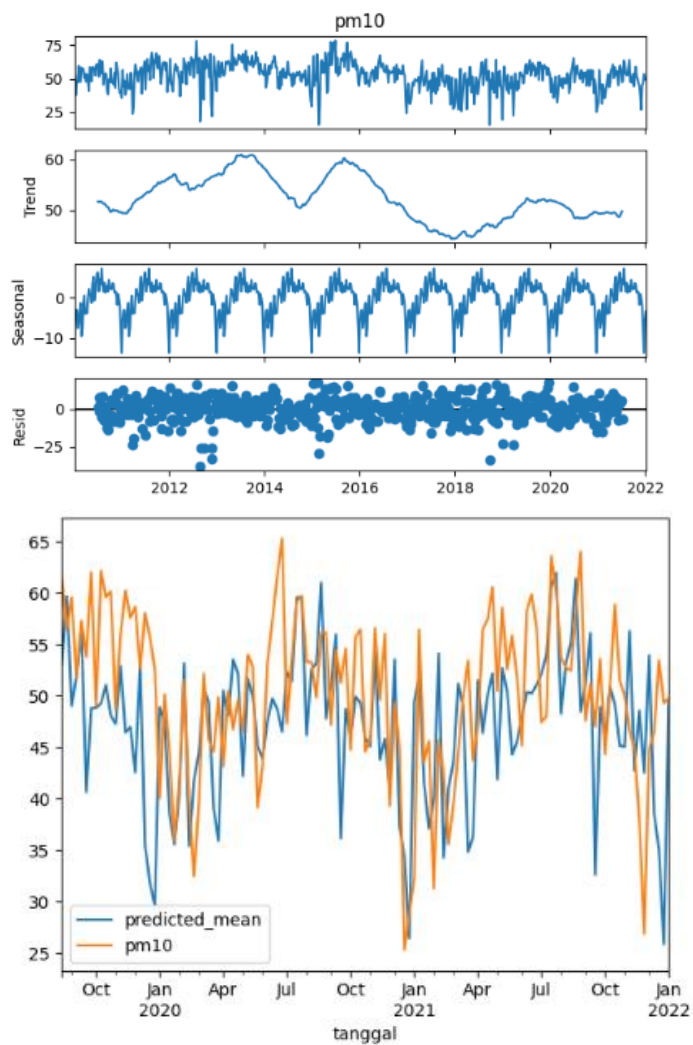
	pm10	so2	co	o3	no2
tanggal					
2019-08-19	61.500000	17.714286	20.285714	64.142857	15.285714
2019-08-26	56.428571	19.142857	21.000000	60.000000	15.285714
2019-09-02	59.571429	17.000000	22.571429	86.714286	13.142857
2019-09-09	51.714286	14.285714	22.857143	71.000000	16.571429
2019-09-16	57.314286	21.714286	18.142857	72.714286	15.285714
...
2021-12-06	44.857143	24.571429	13.714286	19.142857	21.428571
2021-12-13	46.428571	31.000000	9.857143	19.000000	16.571429
2021-12-20	53.428571	40.857143	18.142857	23.142857	13.000000
2021-12-27	49.285714	53.571429	17.571429	20.428571	14.142857
2022-01-03	49.750000	53.000000	16.000000	21.250000	13.000000

125 rows × 5 columns

8. Modeling Build

- Modelling PM10

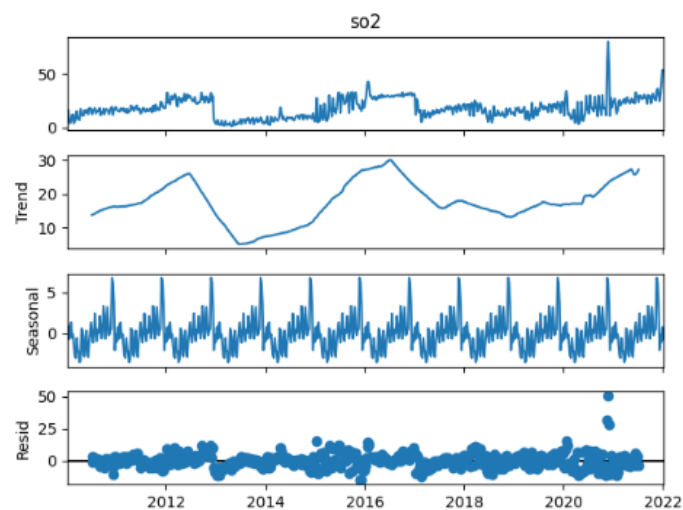
SARIMAX

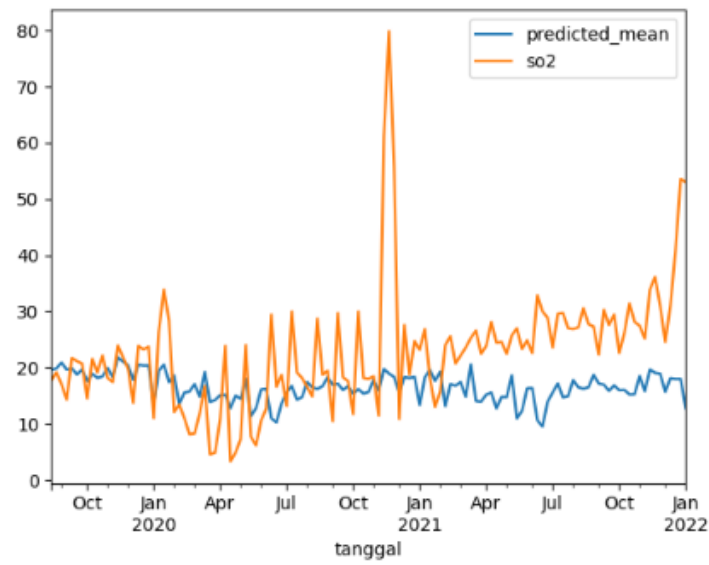


Nilai RMSE : 8,6

- Modelling SO2

SARIMAX

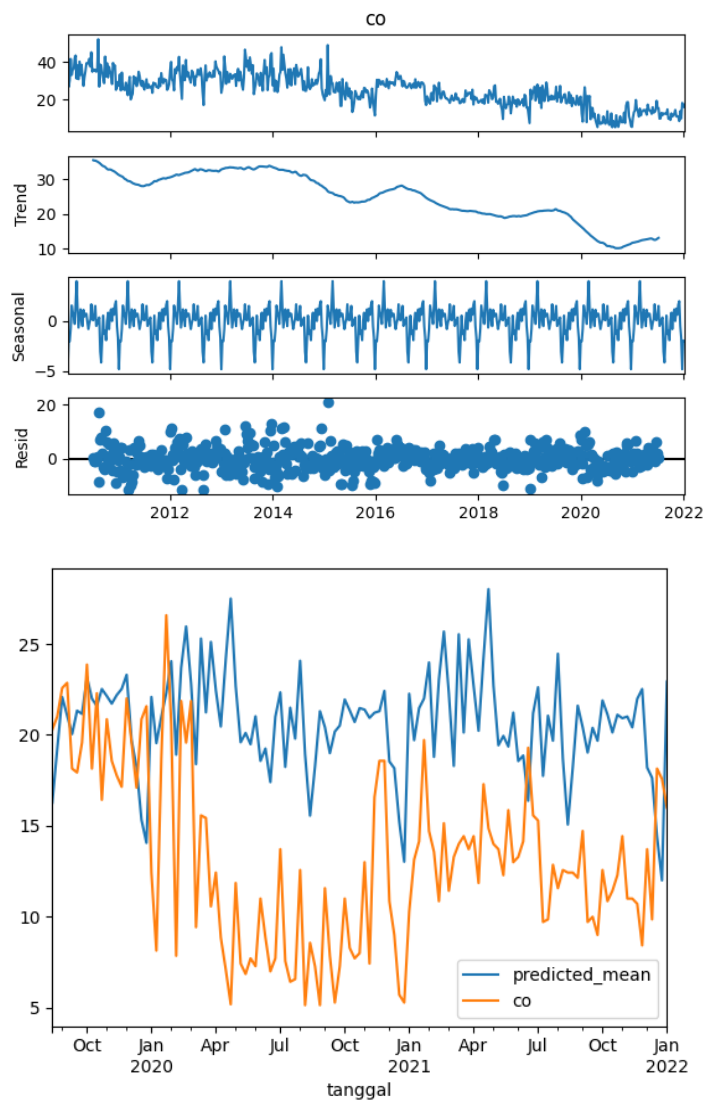




Nilai RMSE : 12,4

- Modelling Sarimax CO

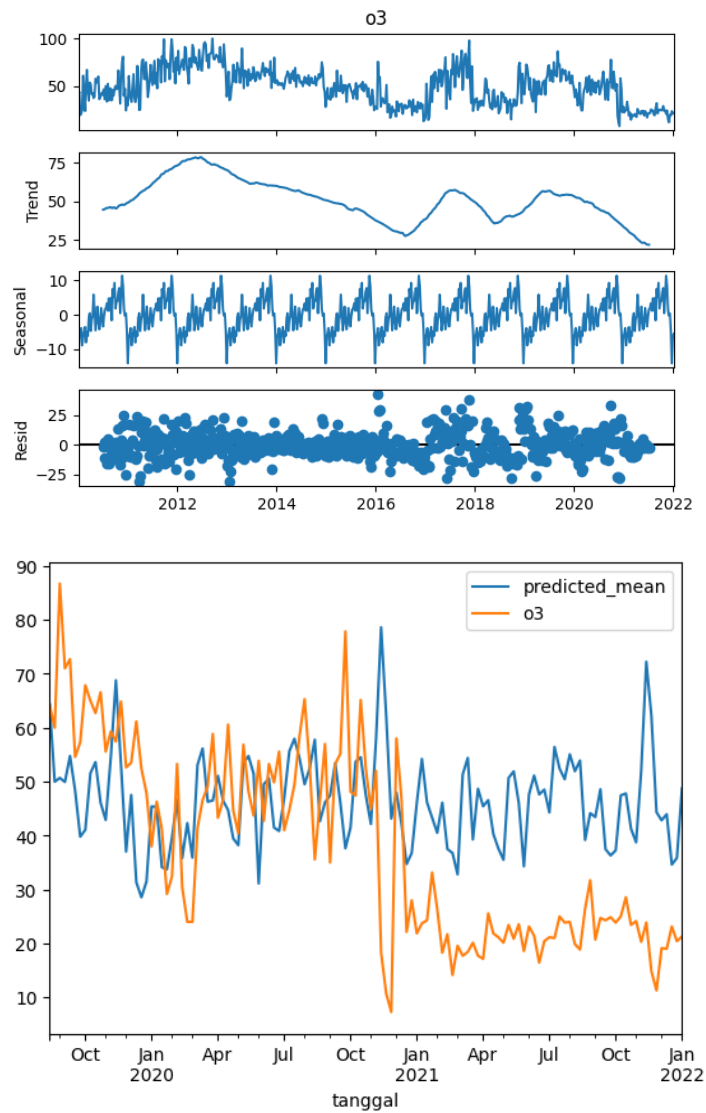
SARIMAX



Nilai RMSE : 9,07

- Modelling Sarimax O3

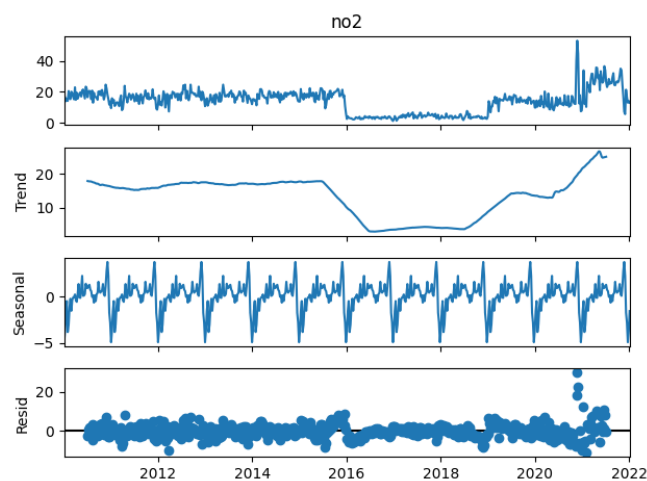
SARIMAX

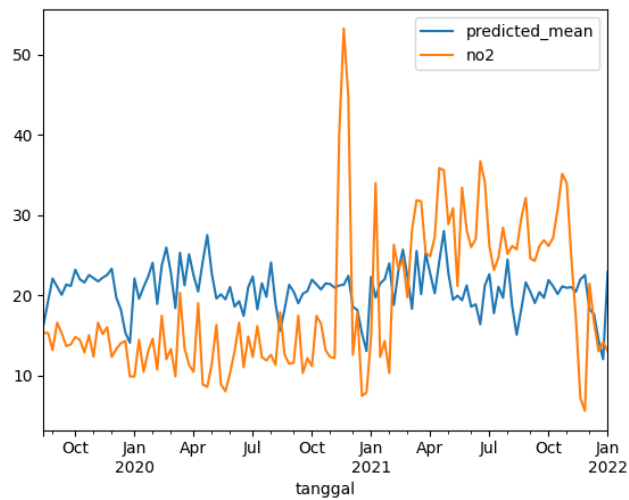


Nilai RMSE : 20,8

- Modelling NO2

SARIMAX





Nilai RMSE : 9,3

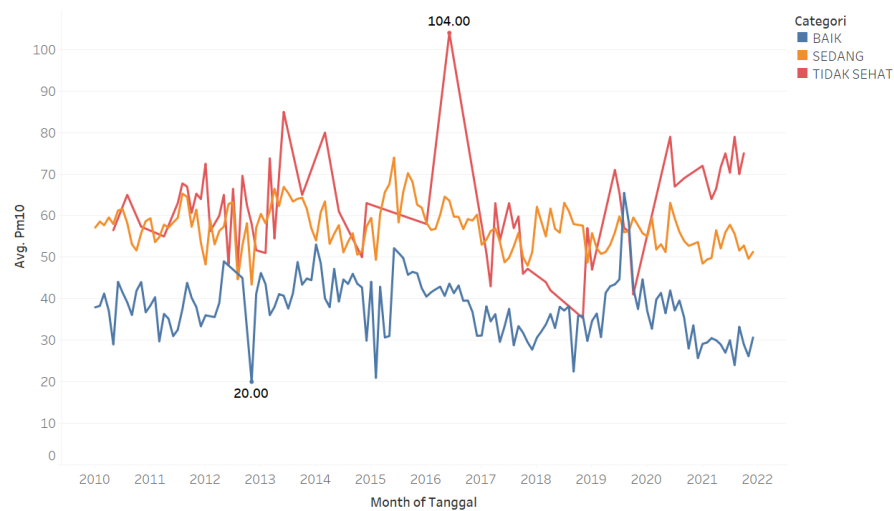
9. Visualisasi Data

9.1. PM10

- Actual

Partikel PM10 mencapai tingkat kesehatan paling buruk pada bulan Juni 2016, sedangkan mencapai tingkat kesehatan paling baik pada bulan November 2012.

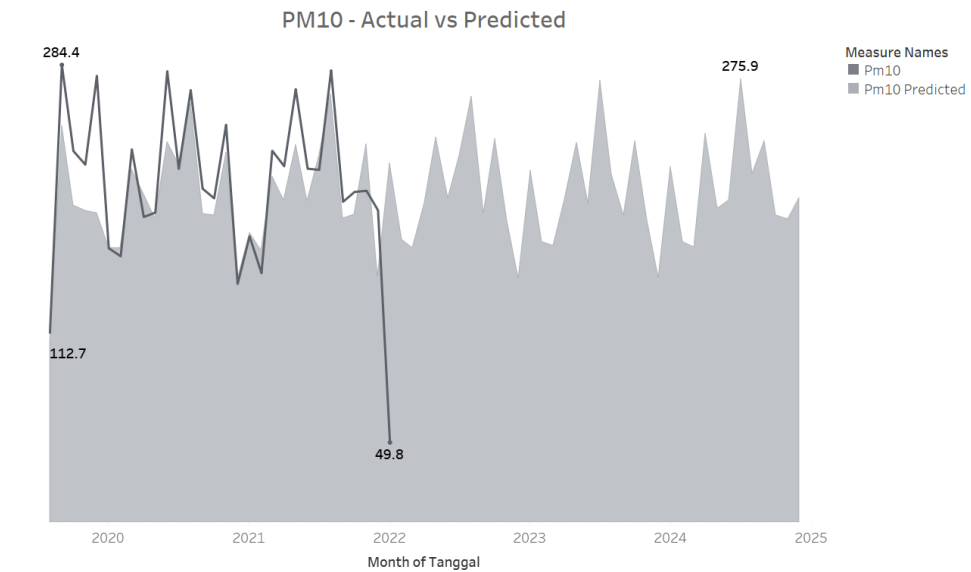
PM10 Actual



The trend of average of Pm10 for Tanggal Month. Color shows details about Kategori. The marks are labeled by average of Pm10. The view is filtered on Kategori, which keeps BAIK, SEDANG and TIDAK SEHAT.

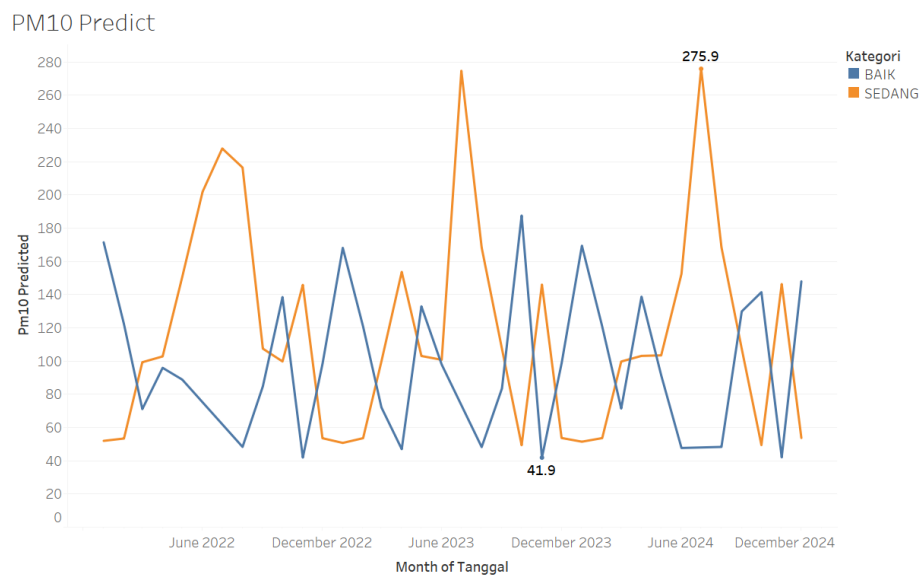
- Actual vs Predict

Perbandingan data actual dengan hasil prediksi.



- Predicted

Partikel PM10 mencapai tingkat kesehatan paling buruk pada bulan Juli 2024, sedangkan mencapai tingkat kesehatan paling baik pada bulan November 2023.

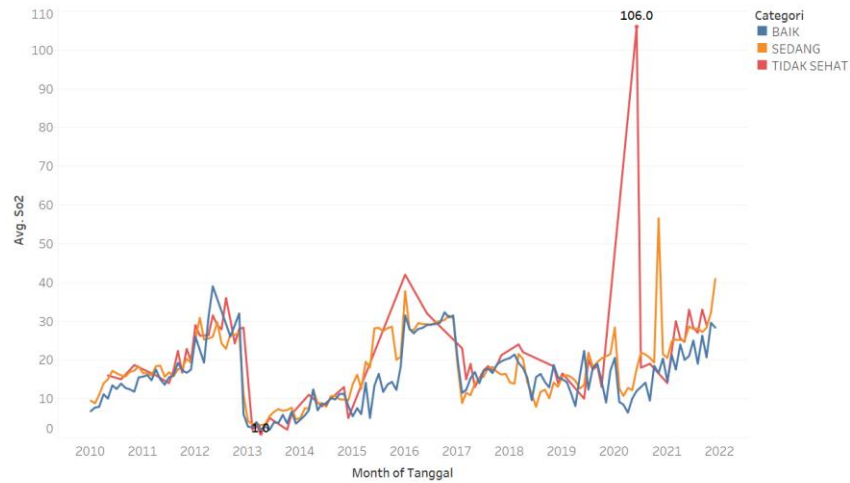


9.2. SO2

- Actual

Partikel SO2 mencapai tingkat kesehatan paling buruk pada bulan Juni 2020, sedangkan mencapai tingkat kesehatan paling baik pada bulan Juni 2013.

SO2 Actual

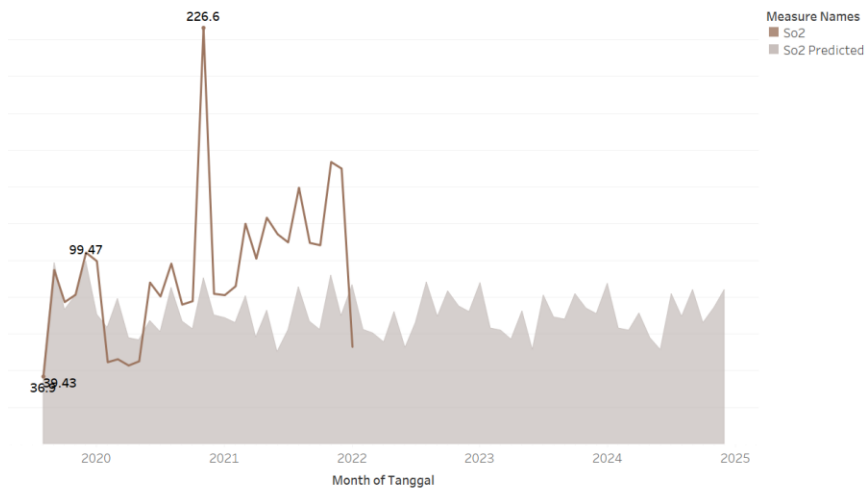


The trend of average of So2 for Tanggal Month. Color shows details about Kategori. The marks are labeled by average of So2. The view is filtered on Kategori, which keeps BAIK, SEDANG and TIDAK SEHAT.

- Actual vs Predict

Perbandingan data actual dengan hasil prediksi.

SO2 - Actual vs Predicted

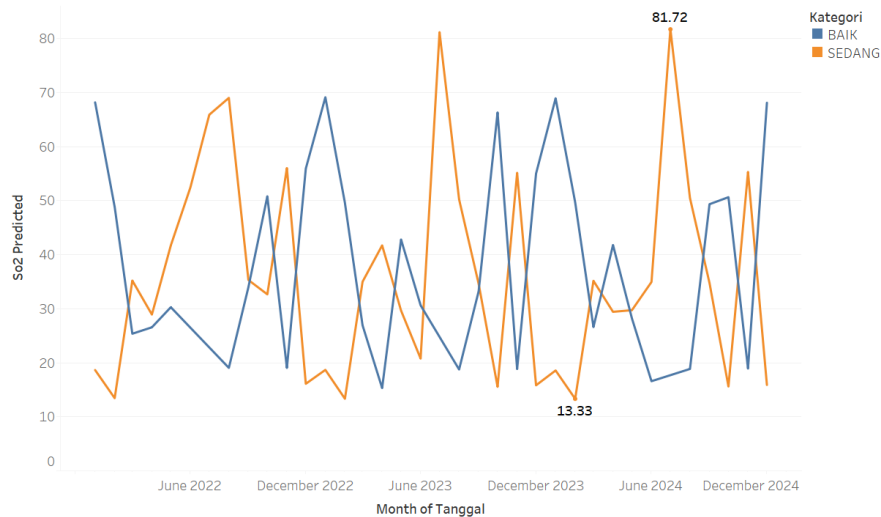


The trends of So2 Predicted and So2 for Tanggal Month. Color shows details about So2 Predicted and So2. For pane Sum of So2 Predicted: The marks are labeled by So2 Predicted. For pane Sum of So2: The marks are labeled by So2.

- Predicted

Partikel SO2 mencapai tingkat kesehatan paling buruk pada bulan Juli 2024, sedangkan mencapai tingkat kesehatan paling baik pada bulan April 2023.

SO2 Predict



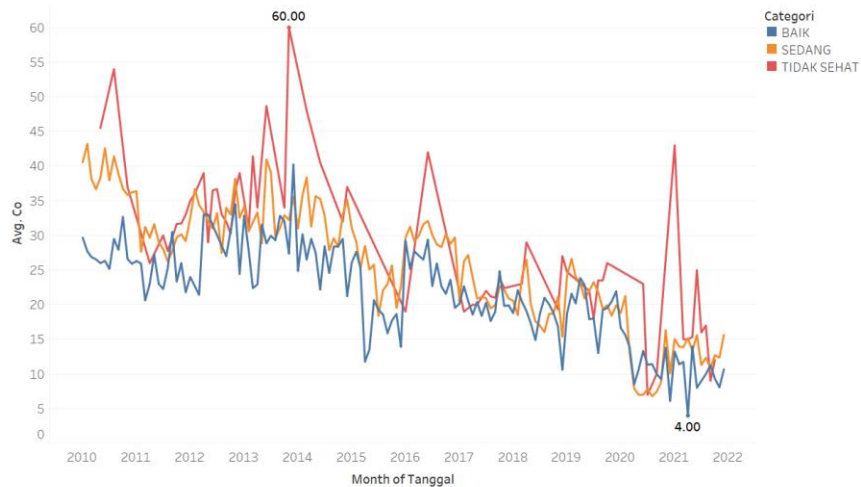
The trend of sum of So2 Predicted for Tanggal Month. Color shows details about Kategori. The marks are labeled by sum of So2 Predicted. The view is filtered on Kategori, which keeps BAIK and SEDANG.

9.3. CO

- Actual

Partikel CO mencapai tingkat kesehatan paling buruk pada bulan November 2013, sedangkan mencapai tingkat kesehatan paling baik pada bulan April 2021.

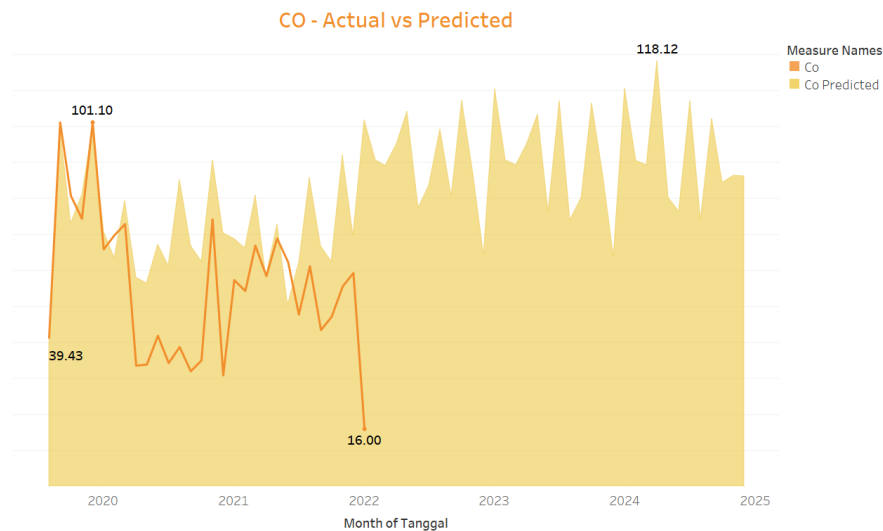
CO Actual



The trend of average of Co for Tanggal Month. Color shows details about Kategori. The marks are labeled by average of Co. The view is filtered on Kategori, which keeps BAIK, SEDANG and TIDAK SEHAT.

- Actual vs Predict

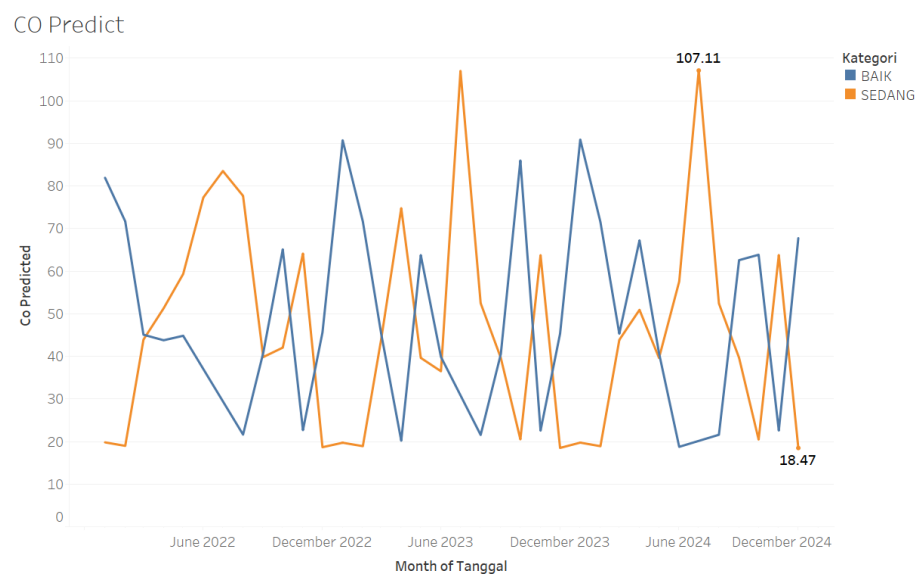
Perbandingan data actual dengan hasil prediksi.



The trends of Co Predicted and Co for Tanggal Month. Color shows details about Co Predicted and Co. For pane Sum of Co Predicted: The marks are labeled by Co Predicted. For pane Sum of Co: The marks are labeled by Co.

- Predicted**

Partikel CO mencapai tingkat kesehatan paling buruk pada bulan Juli 2024, sedangkan mencapai tingkat kesehatan paling baik pada bulan Juni 2024.



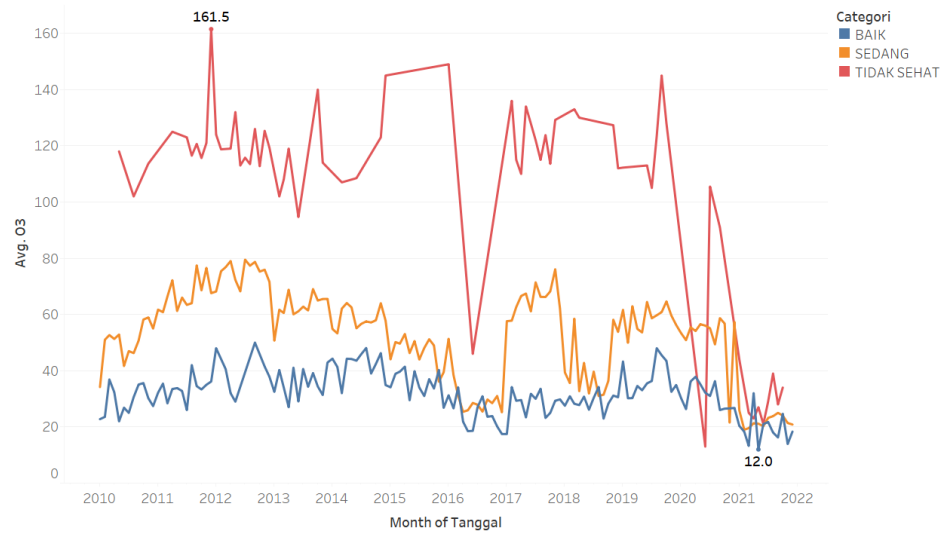
The trend of sum of Co Predicted for Tanggal Month. Color shows details about Kategori. The marks are labeled by sum of Co Predicted. The view is filtered on Kategori, which keeps BAIK and SEDANG.

9.4 O3

- Actual**

Partikel O3 mencapai tingkat kesehatan paling buruk pada bulan Desember 2011, sedangkan mencapai tingkat kesehatan paling baik pada bulan Mei 2021.

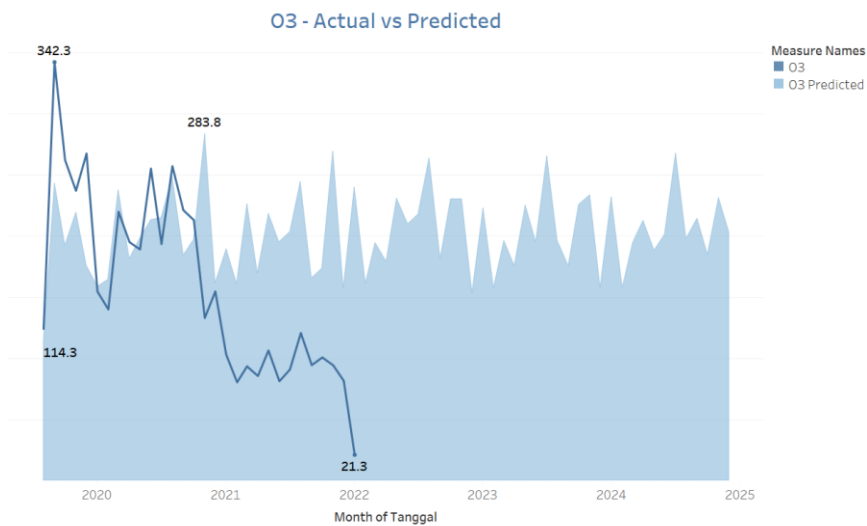
O3 Actual



The trend of average of O3 for Tanggal Month. Color shows details about Kategori. The marks are labeled by average of O3. The view is filtered on Kategori, which keeps BAIK, SEDANG and TIDAK SEHAT.

Actual vs Predict

Perbandingan data actual dengan hasil prediksi.

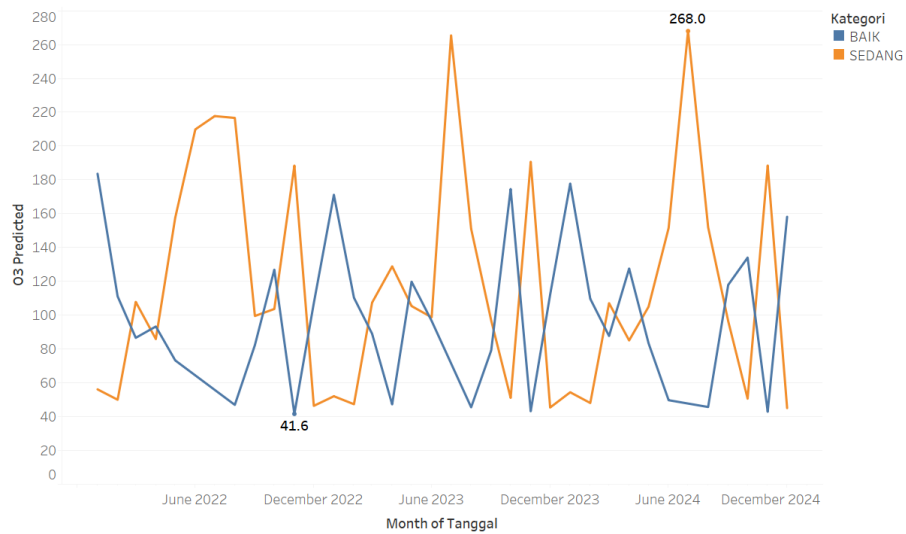


The trends of O3 Predicted and O3 for Tanggal Month. Color shows details about O3 Predicted and O3. For pane Sum of O3 Predicted: The marks are labeled by O3 Predicted. For pane Sum of O3: The marks are labeled by O3.

Predicted

Partikel O3 mencapai tingkat kesehatan paling buruk pada bulan Juli 2024, sedangkan mencapai tingkat kesehatan paling baik pada bulan November 2022.

O3 Predict



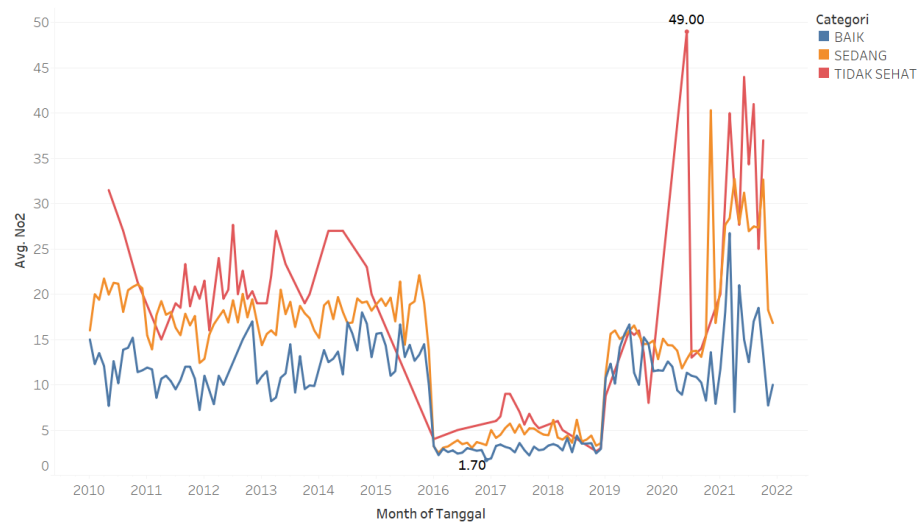
The trend of sum of O3 Predicted for Tanggal Month. Color shows details about Kategori. The marks are labeled by sum of O3 Predicted. The view is filtered on Kategori, which keeps BAIK and SEDANG.

9.5 NO2

- Actual

Partikel NO2 mencapai tingkat kesehatan paling buruk pada bulan Juni 2020, sedangkan mencapai tingkat kesehatan paling baik pada bulan Desember 2016.

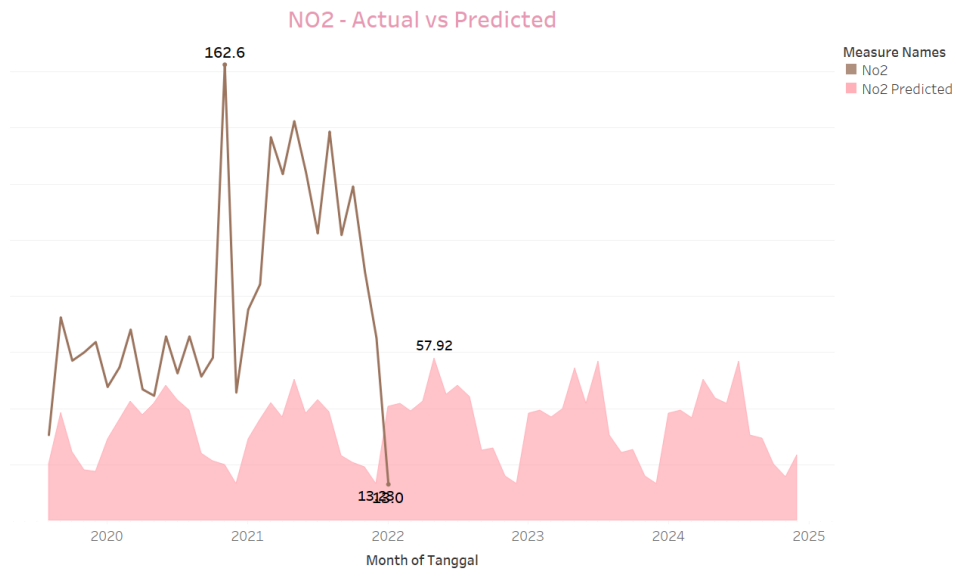
NO2 Actual



The trend of average of No2 for Tanggal Month. Color shows details about Kategori. The marks are labeled by average of No2. The view is filtered on Kategori, which keeps BAIK, SEDANG and TIDAK SEHAT.

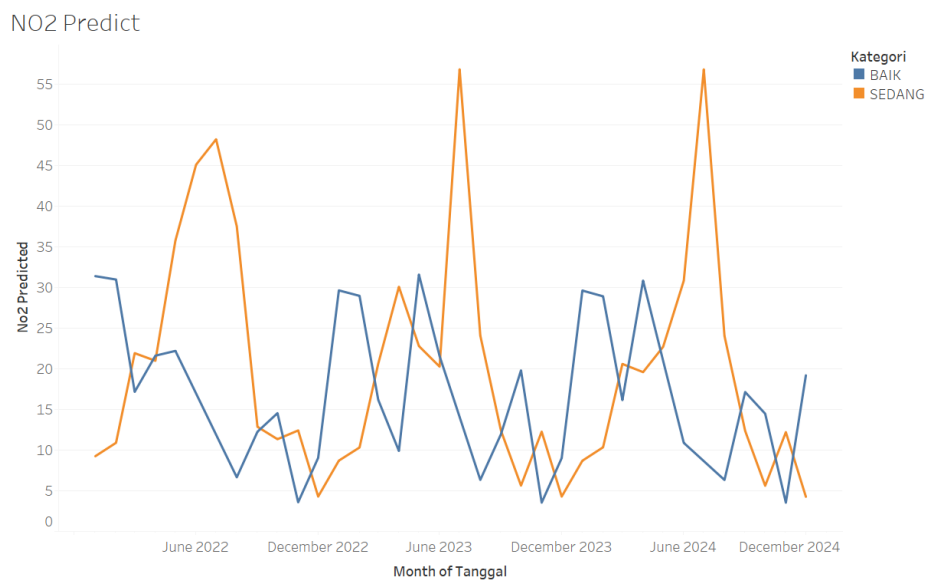
- Actual vs Predict

Perbandingan data actual dengan hasil prediksi.

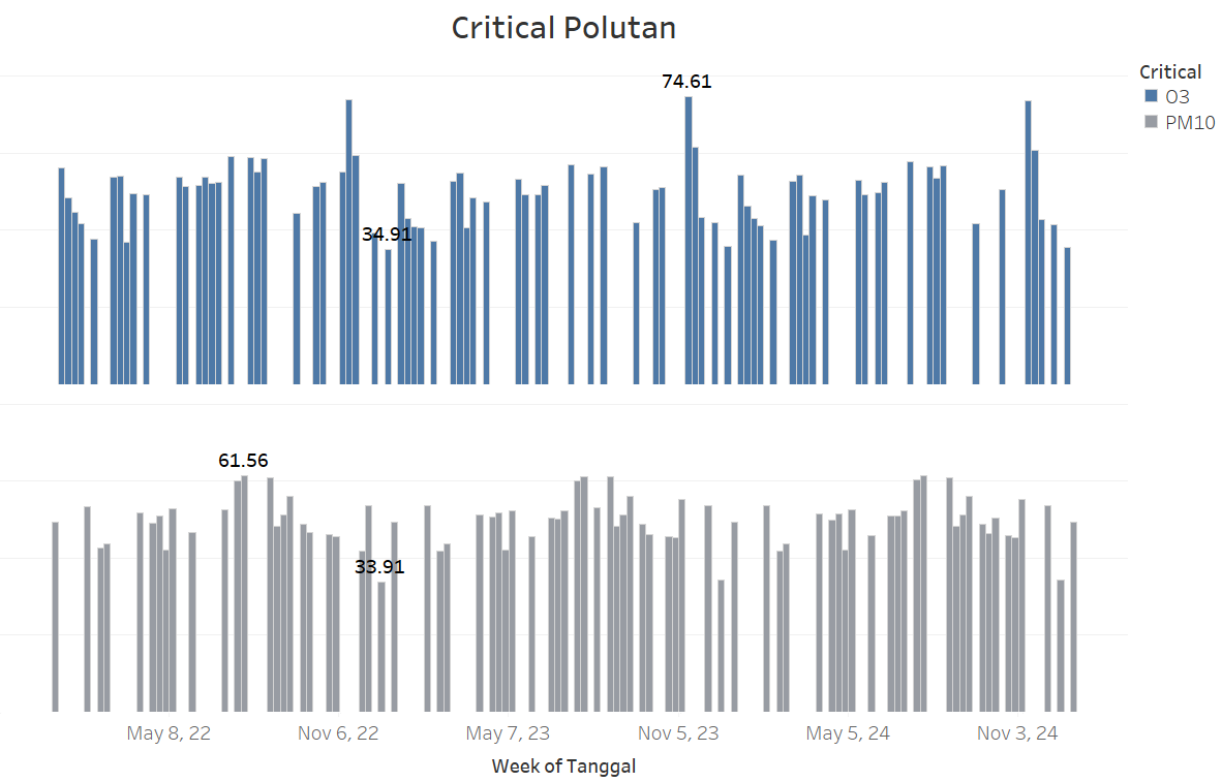


- **Predicted**

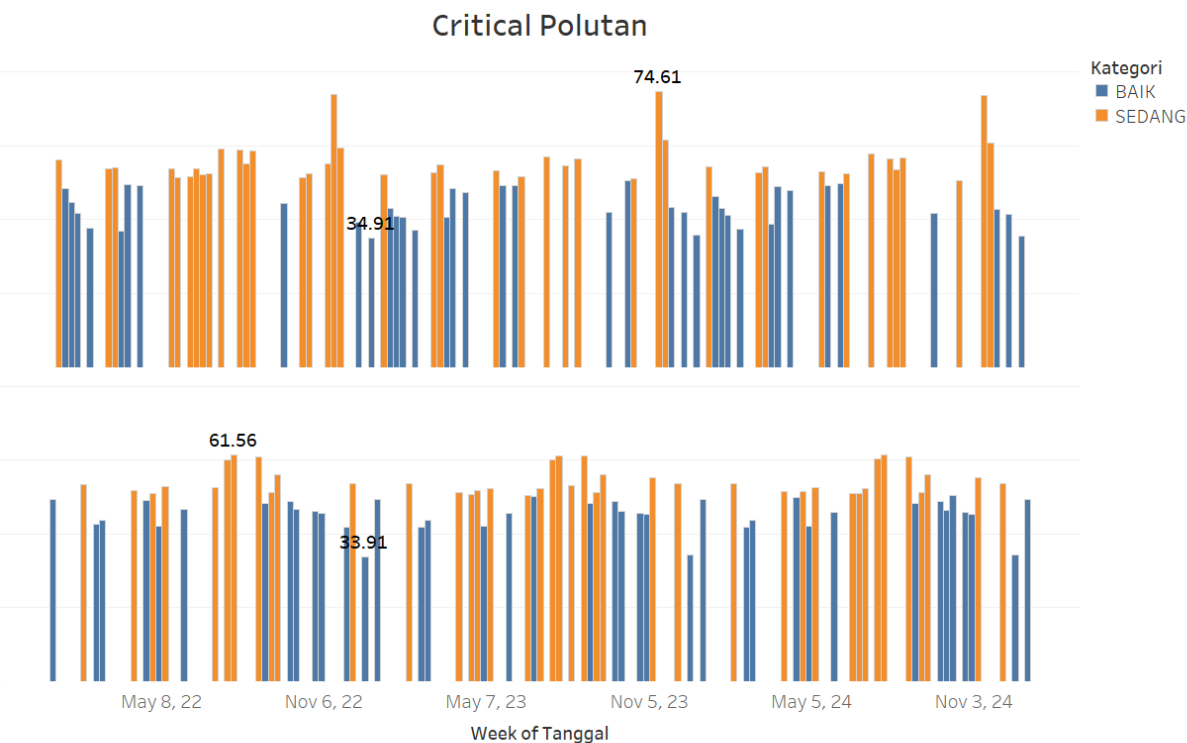
Partikel NO2 mencapai tingkat kesehatan paling buruk pada bulan Juli 2023, sedangkan mencapai tingkat kesehatan paling baik pada bulan November 2024.



Critical Polutan dari prediksi adalah O3 dan PM10



The plot of sum of Max for Tanggal Week broken down by Critical. Color shows details about Critical. The marks are labeled by sum of Max. The view is filtered on Critical, which keeps O3 and PM10.



The plot of sum of Max for Tanggal Week broken down by Critical. Color shows details about Kategori. The marks are labeled by sum of Max. The view is filtered on Critical, which keeps O3 and PM10.