

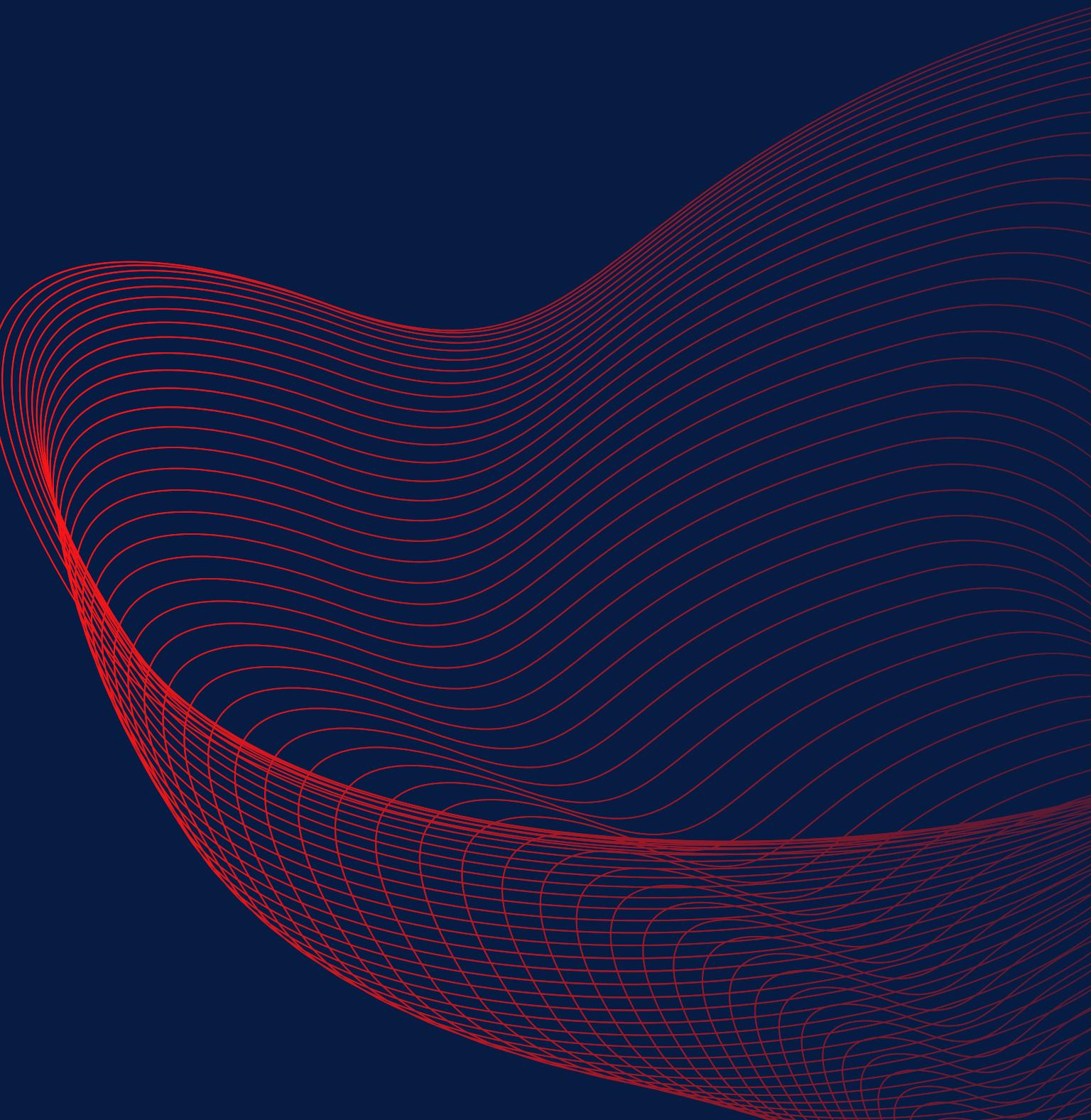


FORECASTING KONSENTRASI PARTIKULAT PM 2.5 DI JAKARTA PUSAT

PACMANN STUDENT HACKATHON, KELOMPOK 3

NOVEMBER 2022

- Alvin N
- Fadhiba Annisa Maksum
- Muhammad Farid Zaki



Introduction

Jakarta merupakan salah satu kota di dunia dengan konsentrasi PM 2.5 yang cukup tinggi. Bahkan pada Juni 2022 lalu, sebagaimana yang kita ingat, Jakarta menjadi kota dengan kualitas udara terburuk di dunia karena konsentrasi PM 2.5 yang terdeteksi sangat tinggi.

PM 2.5 dengan komponen penyusunnya yang berasal dari sulfat, nitrat, amonia, natrium klorida, karbon hitam, debu mineral, dan air, dinilai berbahaya karena dapat lebih merusak kesehatan tubuh dibandingkan dengan PM 10 atau partikel polusi udara lainnya. Eksposur berlebih secara terus-menerus terhadap manusia dapat menyebabkan berbagai macam penyakit pernapasan hingga bisa memicu kanker paru-paru dan penyakit jantung, yang tentu saja dapat menyebabkan kematian. Oleh karena itu, penting bagi manusia untuk menggunakan masker saat konsentrasi PM 2.5 dan kualitas udara sedang memburuk sehingga dapat mencegah penyakit-penyakit di masa mendatang.

Proyek ini bertujuan untuk membuat sebuah sistem yang dapat memprediksi konsentrasi dari PM 2.5 sehingga nantinya diharapkan pengguna dapat mempersiapkan diri (masker dan obat-obatan pendukung) ketika kondisi konsentrasi PM 2.5 diprediksi akan memburuk namun disaat yang bersamaan pengguna juga harus beraktifitas di luar rumah.

Proyek ini menggunakan data konsentrasi PM 2.5 yang sensornya berada di Kedutaan Besar Amerika Serikat di Jakarta Pusat, dan data klimatologi pendukung dari Jakarta Observatory. Algoritma Long-Short Term Memory (LSTM) dipilih untuk melakukan training model pada proyek ini.

Related work

- **Short-Term Prediction of PM2.5 Using LSTM Deep Learning Methods**
- **Prediction of PM2.5 Concentration Based on the LSTM-TSLightGBM Variable Weight Combination Model**

Data Preparation

Data cleansing, transformation, dan aggregation dilakukan pada tahapan ini sehingga dapat memudahkan tahapan selanjutnya. Pada dataset PM 2.5, observasi dilakukan setiap satu jam selama 24 jam dalam sehari, dari 1 Januari 2016 hingga 25 November 2022. Upsampling dilakukan pada dataset ini dan diambil median dari konsentrasi observasinya sehingga hasil akan lebih robust meskipun terdapat error yang dihasilkan oleh sensor pengukur. Dari median konsentrasi yang didapatkan dalam satu hari, dilakukan perhitungan skor Air Quality Index dan juga kategori yang dihasilkan dari kondisi udara menggunakan equation dari EPA.

Kedua dataset kemudian digabungkan menjadi satu berdasarkan datetimeindex. Interpolation digunakan untuk melakukan handling pada missing values di dataset final.

Jakarta PM 2.5 Dataset, 1 Jan 2016 – 25 Nov 2022, Hourly Observation, 118182 rows

Site	Parameter	Date (LT)	Year	Month	Day	Hour	NowCast Conc.	AQI	AQI Category	Raw Conc.	Conc. Unit	Duration	QC Name	
0	Jakarta Central	PM2.5 - Principal	2016-01-01 01:00 AM	2016	1	1	1	256.6	307	Hazardous	412.0	UG/M3	1 Hr	Valid
1	Jakarta Central	PM2.5 - Principal	2016-01-01 02:00 AM	2016	1	1	2	203.3	253	Very Unhealthy	150.0	UG/M3	1 Hr	Valid
2	Jakarta Central	PM2.5 - Principal	2016-01-01 03:00 AM	2016	1	1	3	111.1	180	Unhealthy	19.0	UG/M3	1 Hr	Valid
3	Jakarta Central	PM2.5 - Principal	2016-01-01 04:00 AM	2016	1	1	4	60.5	154	Unhealthy	10.0	UG/M3	1 Hr	Valid
4	Jakarta Central	PM2.5 - Principal	2016-01-01 05:00 AM	2016	1	1	5	40.6	114	Unhealthy for Sensitive Groups	21.0	UG/M3	1 Hr	Valid

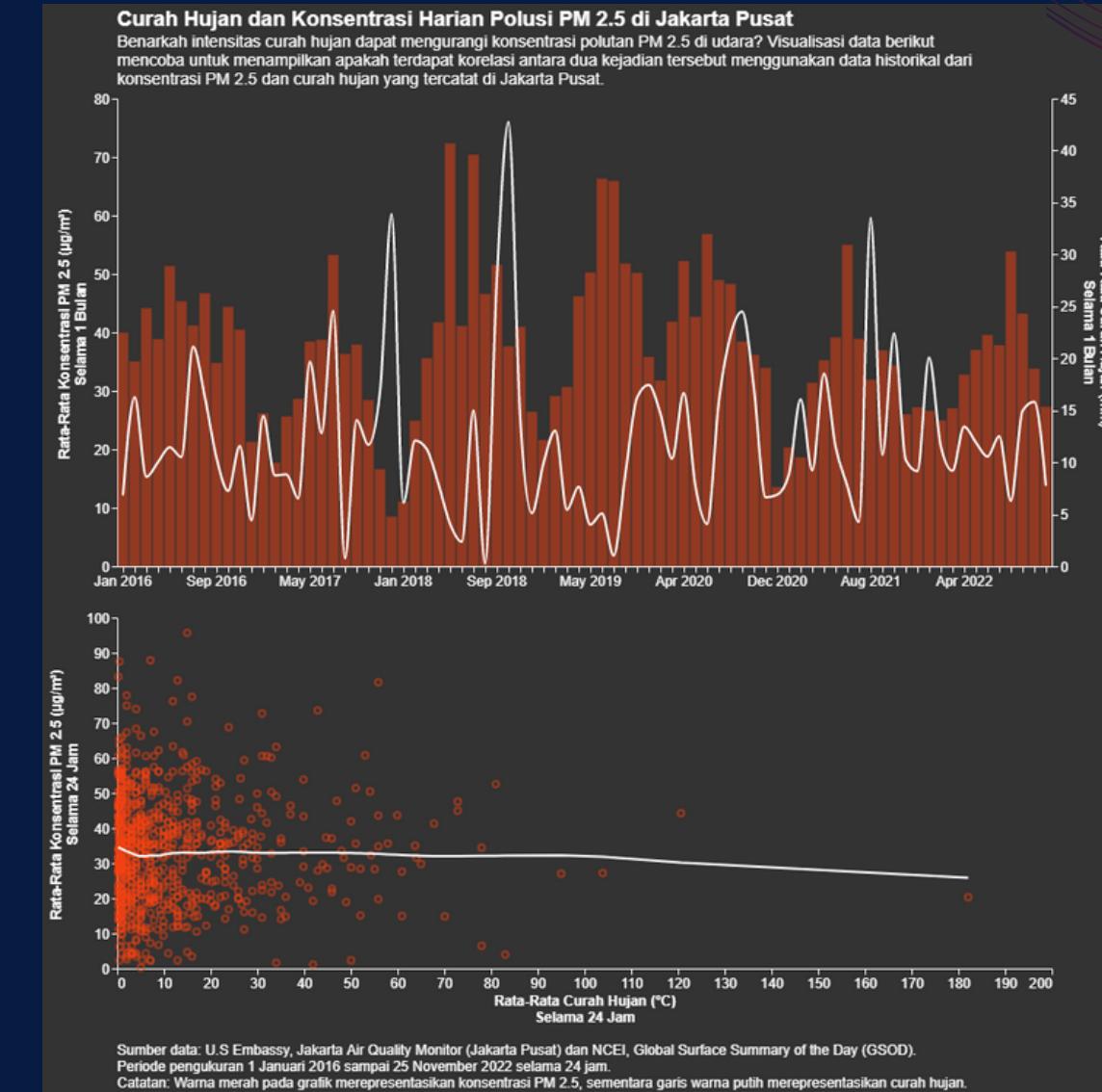
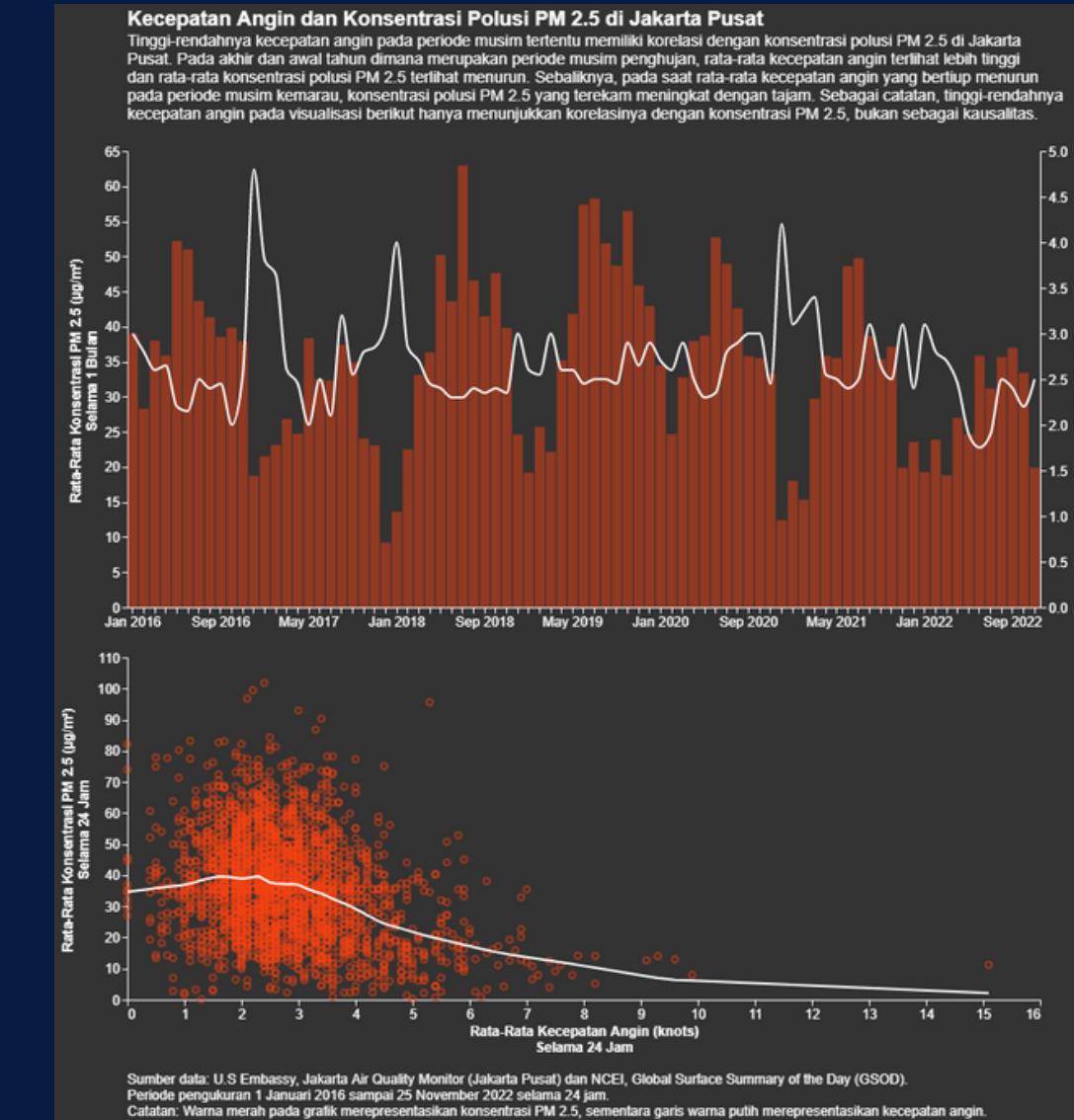
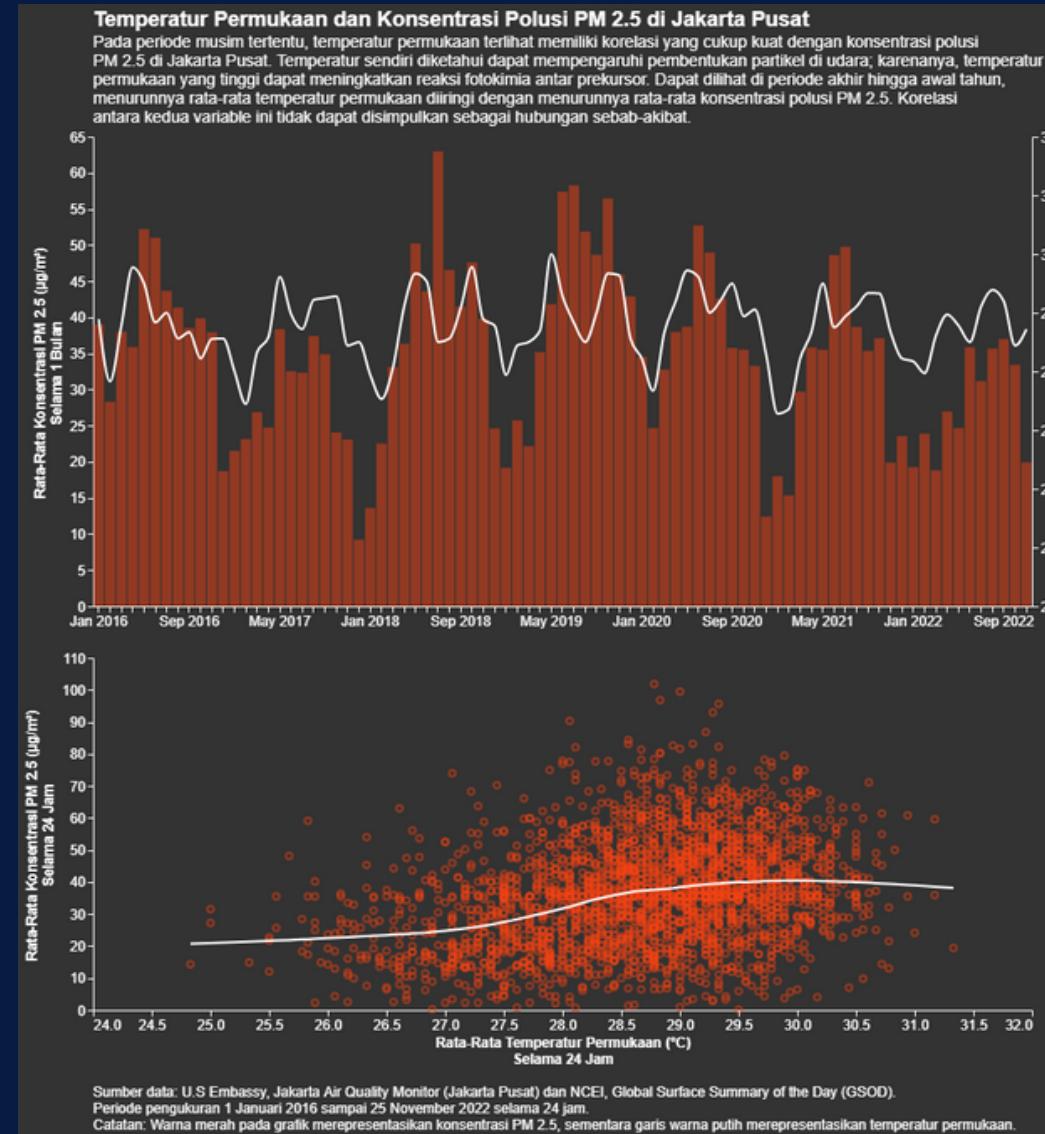
Jakarta Climate Dataset, 1 Jan 2016 – 19 Nov 2022, Daily Observation, 2507 rows

DATE	STATION	NAME	LATITUDE	LONGITUDE	ELEVATION	GUST	PRCP	TEMP	VISIB	WDSP	
2016-01-01	96745099999	JAKARTA OBSERVATORY, ID	-6.183333	106.833333		8.0	999.9	0.00	82.6	4.1	3.3
2016-01-02	96745099999	JAKARTA OBSERVATORY, ID	-6.183333	106.833333		8.0	999.9	0.60	81.4	3.9	3.2
2016-01-03	96745099999	JAKARTA OBSERVATORY, ID	-6.183333	106.833333		8.0	999.9	0.00	84.2	4.3	3.6
2016-01-04	96745099999	JAKARTA OBSERVATORY, ID	-6.183333	106.833333		8.0	999.9	0.00	84.5	4.1	3.3
2016-01-05	96745099999	JAKARTA OBSERVATORY, ID	-6.183333	106.833333		8.0	999.9	0.47	83.2	4.0	2.3

Jakarta PM 2.5 & Climate Processed Dataset, 1 Jan 2016 – 25 Nov 2022, Daily Observation, 2521 rows

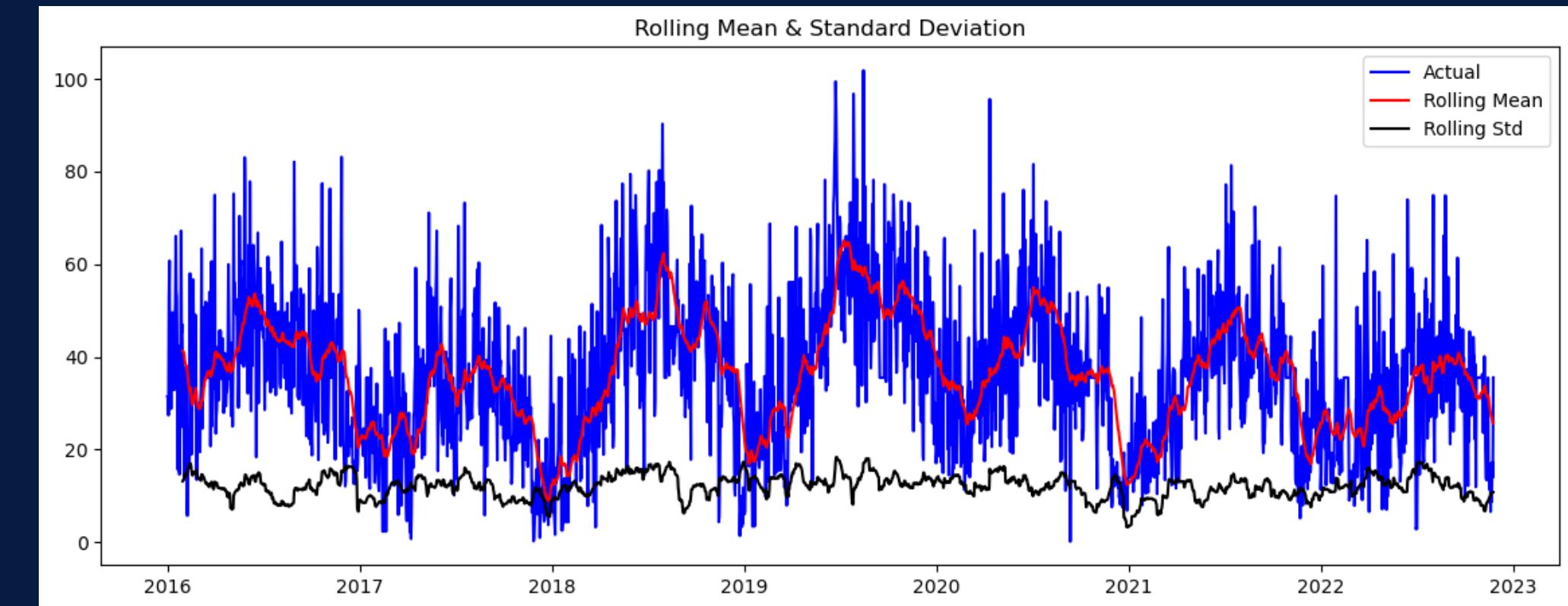
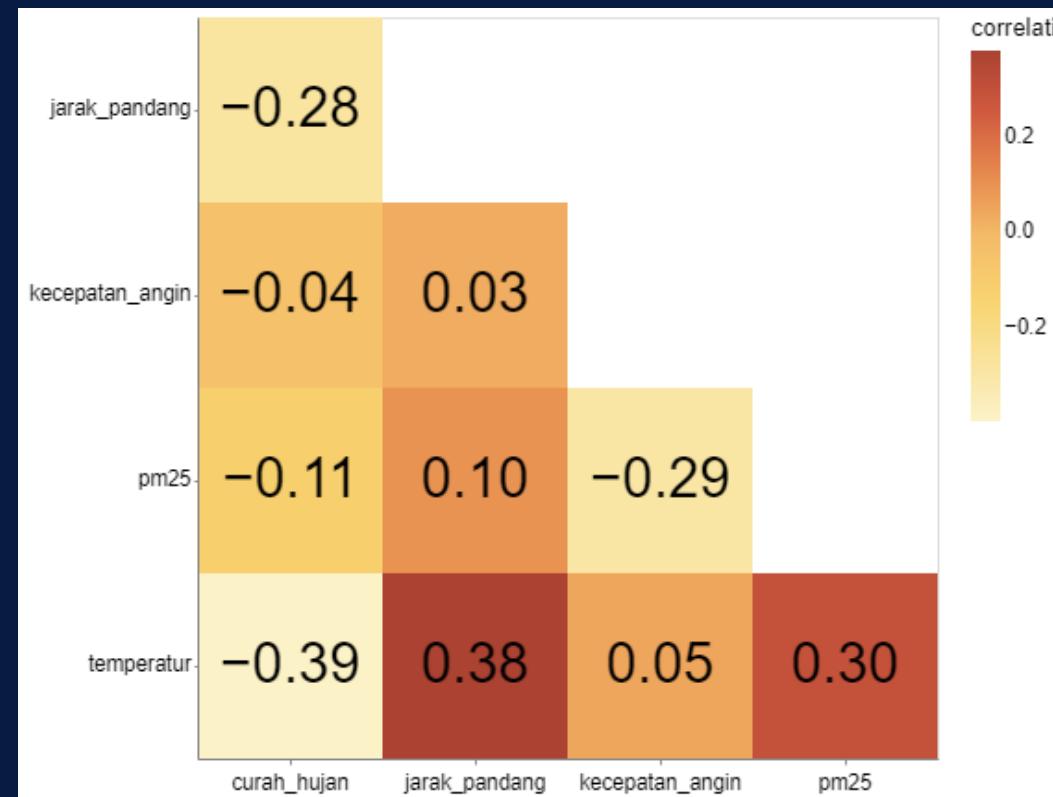
tanggal	lokasi	pm25	aqi	kategori	temperatur	curah_hujan	jarak_pandang	kecepatan_angin	
1	2016-01-01	Jakarta Pusat	31.40	91.6	Sedang	28.11	0.00	4.1	3.3
2	2016-01-02	Jakarta Pusat	31.00	90.7	Sedang	27.44	15.24	3.9	3.2
5	2016-01-03	Jakarta Pusat	27.40	83.2	Sedang	29.00	0.00	4.3	3.6
7	2016-01-04	Jakarta Pusat	52.80	143.6	Tidak Sehat Untuk Kelompok Rentan	29.17	0.00	4.1	3.3
8	2016-01-05	Jakarta Pusat	60.75	153.7	Tidak Sehat	28.44	11.94	4.0	2.3
10	2016-01-06	Jakarta Pusat	36.10	102.5	Tidak Sehat Untuk Kelompok Rentan	28.89	0.25	4.3	2.3
12	2016-01-07	Jakarta Pusat	37.90	106.9	Tidak Sehat Untuk Kelompok Rentan	30.17	0.00	4.3	3.9
14	2016-01-08	Jakarta Pusat	28.80	86.1	Sedang	29.83	0.00	4.3	3.0
16	2016-01-09	Jakarta Pusat	44.90	124.1	Tidak Sehat Untuk Kelompok Rentan	29.28	0.00	4.2	4.2
18	2016-01-10	Jakarta Pusat	49.55	135.6	Tidak Sehat Untuk Kelompok Rentan	29.50	0.00	4.2	2.7

Exploratory Data Analysis



Pada tahapan Exploratory Data Analysis (EDA), kami mencoba untuk melihat korelasi dari temperatur permukaan, curah hujan, dan kecepatan angin yang tercatat di Jakarta Pusat dengan konsentrasi PM 2.5 selama 2016 hingga 2022. Selain menampilkan data historikal secara visual, kami juga coba melakukan hypothesis testing ketiga fitur ini dengan konsentrasi PM 2.5 menggunakan Spearman Correlation. Hasilnya, ketiga null hypothesis (H_0) dapat ditolak sehingga dapat dikatakan jika ketiga fitur ini memiliki korelasi dengan konsentrasi PM 2.5.

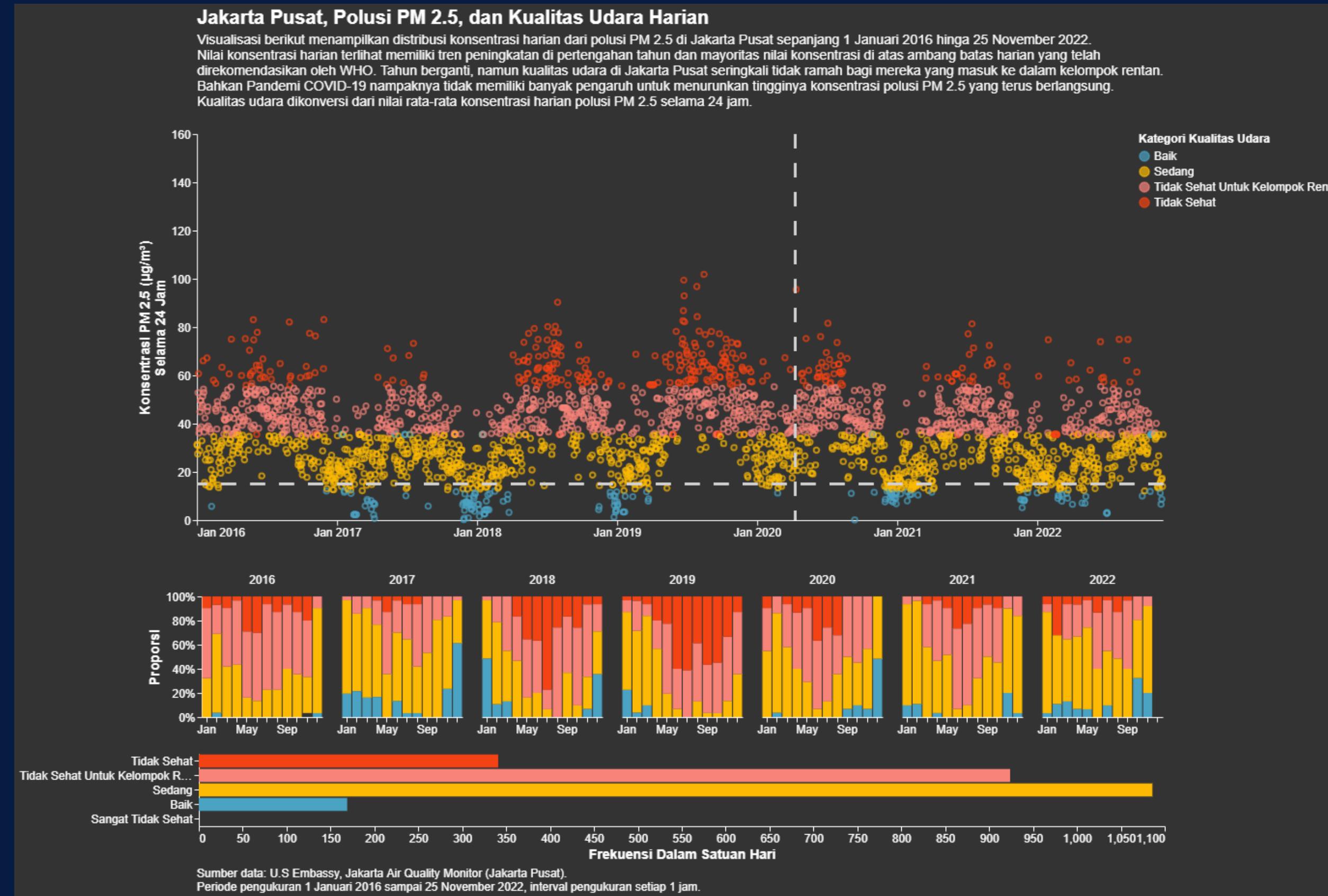
Exploratory Data Analysis



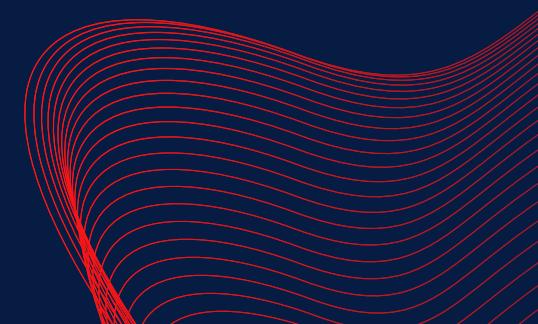
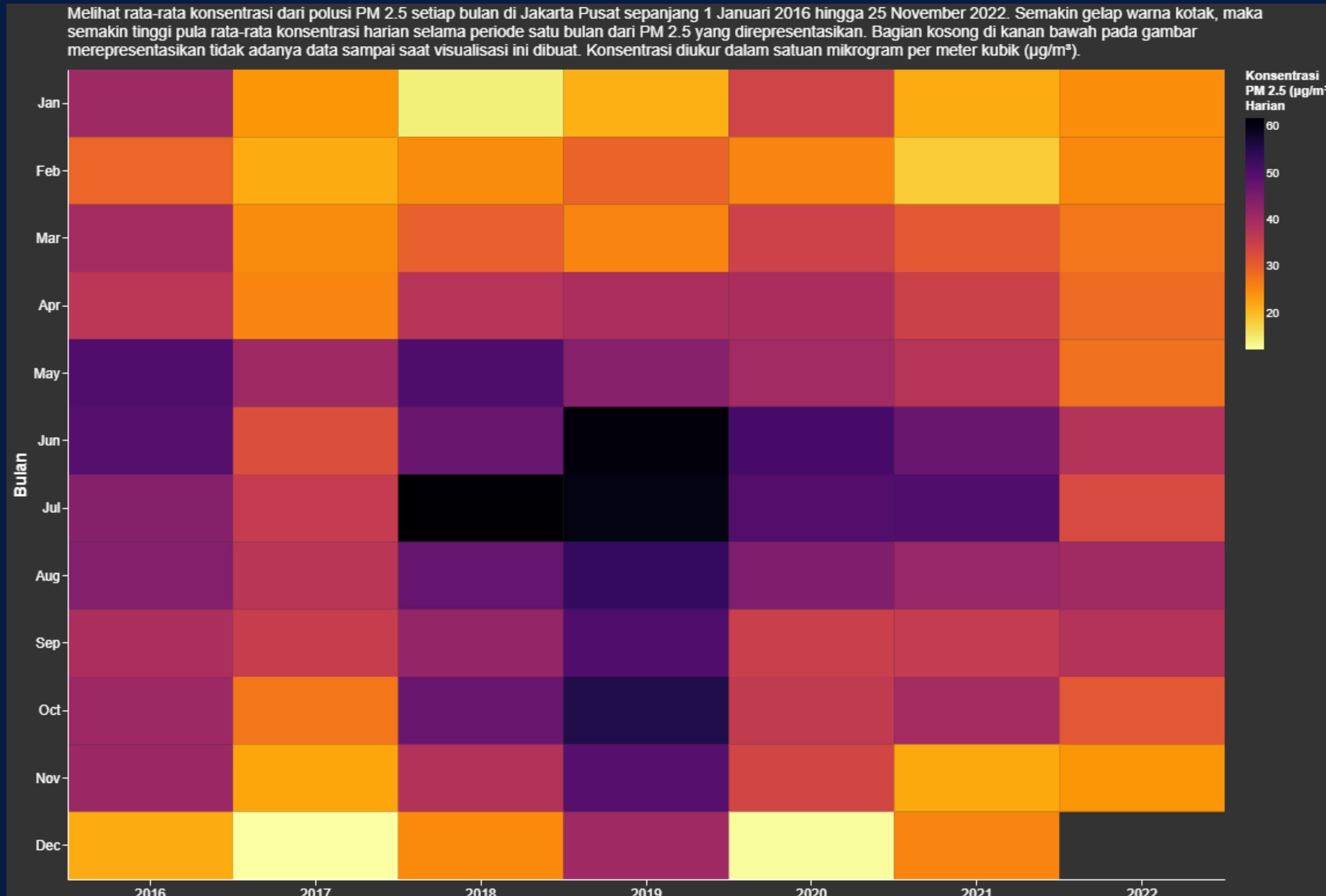
Gambar di sebelah kiri menunjukkan korelasi dari temperatur, jarak pandang, kecepatan angin, dan curah hujan dengan konsentrasi PM 2.5. Jarak pandang dan curah hujan memiliki korelasi yang lemah dengan konsentrasi PM 2.5,. Sementara itu, temperatur dan kecepatan angin terlihat memiliki korelasi yang cukup moderat dengan konsentrasi PM 2.5 di Jakarta Pusat.

Gambar di sebelah kanan merupakan hasil pengujian stasioneritas menggunakan Dickey-Fuller Test pada dataset kami yang merupakan timeseries dataset. Hasil pengujian menunjukkan jika dataset yang kami gunakan adalah data stasioner, yang mana berarti nilai rata-rata dan variansnya tidak mengalami perubahan yang secara sistematis sepanjang waktu atau dengan kata lain, rata-rata dan variansnya konstan.

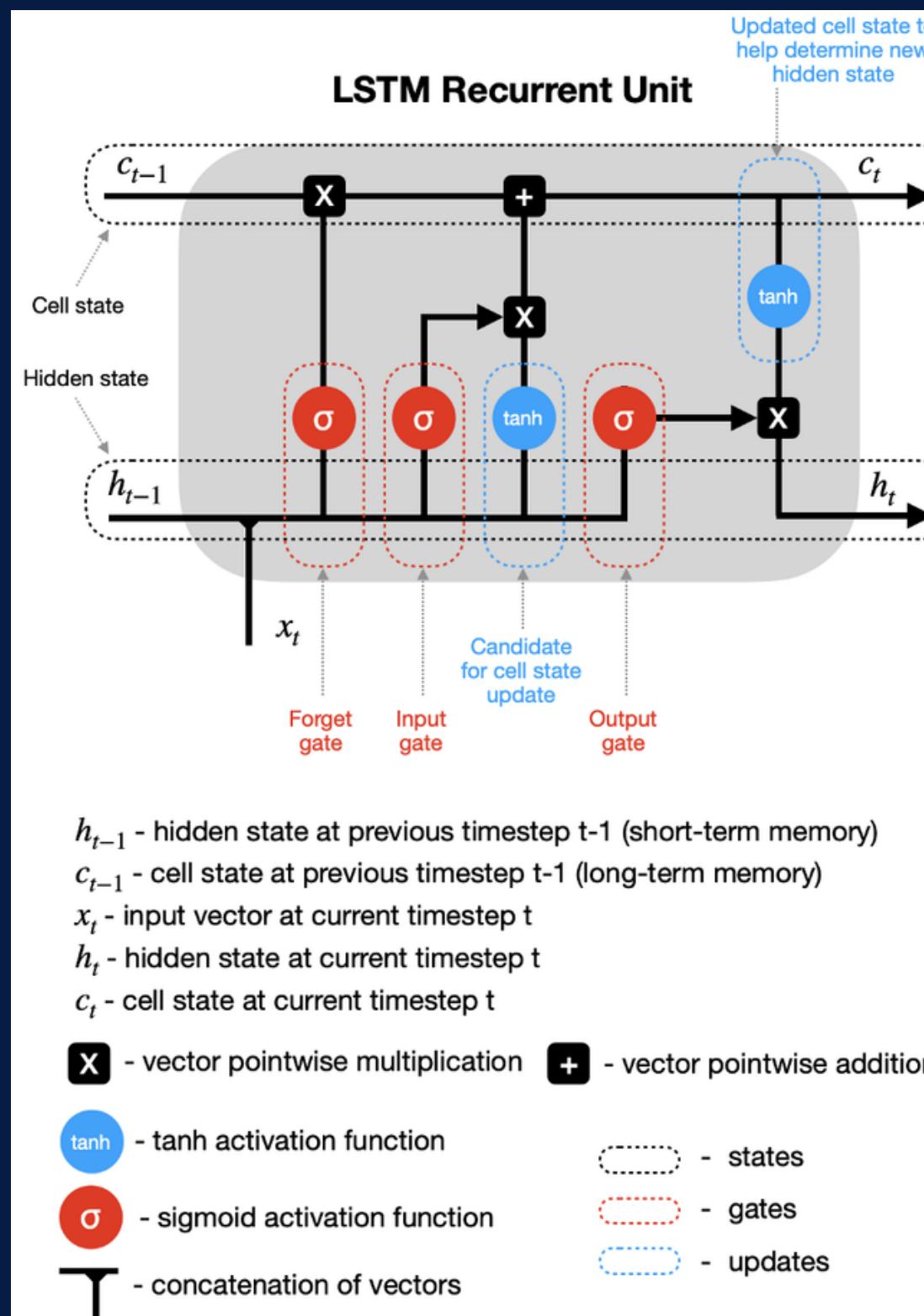
Exploratory Data Analysis



Exploratory Data Analysis



Long-Short Term Memory



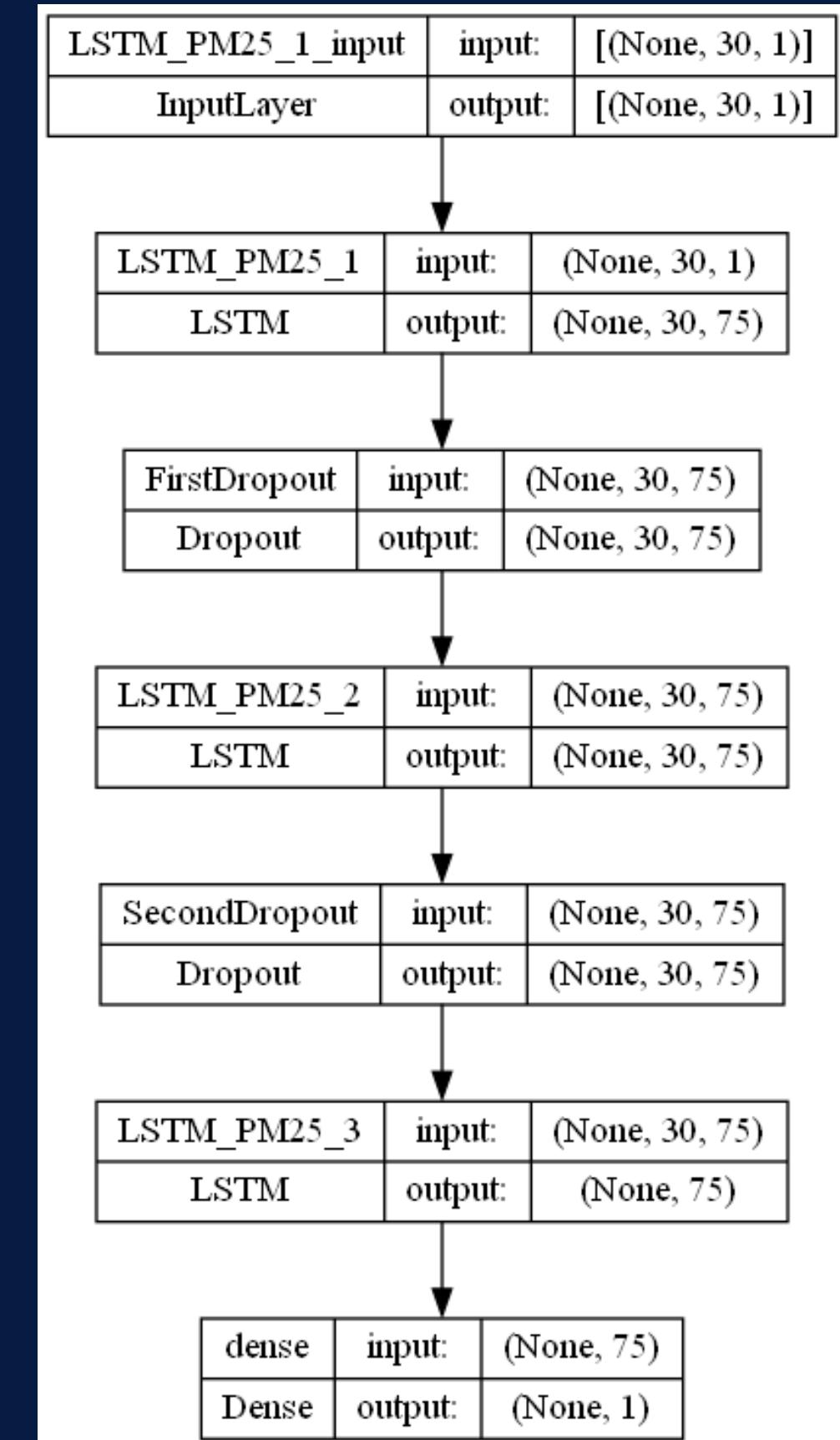
Long-short Term Memory (LSTM) adalah salah satu algoritma dari deep learning yang merupakan hasil pengembangan dari algoritma Recurrent Neural Network (RNN). Kami memilih menggunakan LSTM dalam pengembangan proyek ini dikarenakan tipe dataset yang kami gunakan dan juga LSTM memiliki kemampuan untuk mengingat kumpulan informasi yang telah disimpan dalam jangka waktu panjang, sekaligus menghapus informasi yang tidak lagi relevan. Kami menilai LSTM memiliki efisiensi dalam memproses, memprediksi, sekaligus mengklasifikasikan data berdasarkan urutan waktu tertentu.

Gambar di samping merupakan diagram cara kerja dari LSTM. LSTM memutuskan informasi apa yang harus tetap utuh dan apa yang harus dibuang dari cell state, kemudian menentukan informasi baru apa yang harus disimpan dan menggantikan yang tidak relevan yang berhasil diidentifikasi pada langkah sebelumnya. Untuk lebih lengkapnya tentang LSTM, dapat membaca pada tautan bertikut: <https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9>

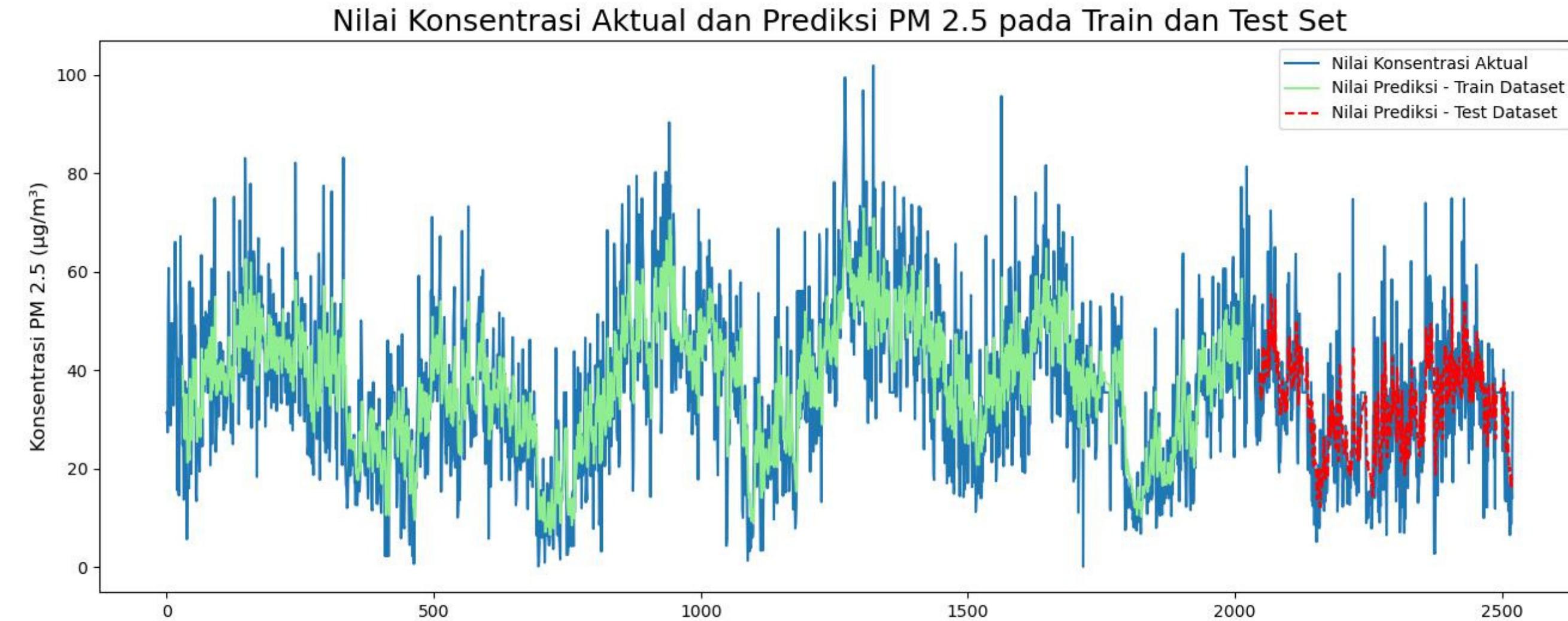
BUILD LSTM - MODEL ARCHITECTURE

Sebelum mengkonstruksi arsitektur dari model LSTM, kami melakukan preprocessing dan feature engineering lebih dulu pada dataset kami. Dataset dinormalisasi kemudian kami bagi menjadi 80% untuk training (2016 baris) dan 20% untuk validasi (505 baris). Input kemudian kami reshape menjadi 3D sehingga dapat diterima oleh model.

Setelah melakukan beberapa kali uji coba, arsitektur dari model LSTM yang cocok untuk kami memiliki 3 hidden layers dengan masing-masing mempunyai 75 neurons, Dropout sebesar 20% dan input shape yang terdiri dari 1 time step dengan 30 time windows. Kami menggunakan Mean Square Error sebagai loss functionnya dan adam optimization. Untuk menghindari overfitting, kami menggunakan mekanisme EarlyStopping di dalam model ini. Model kemudian ditraining di dalam 100 training epochs dengan batch size sebesar 64.



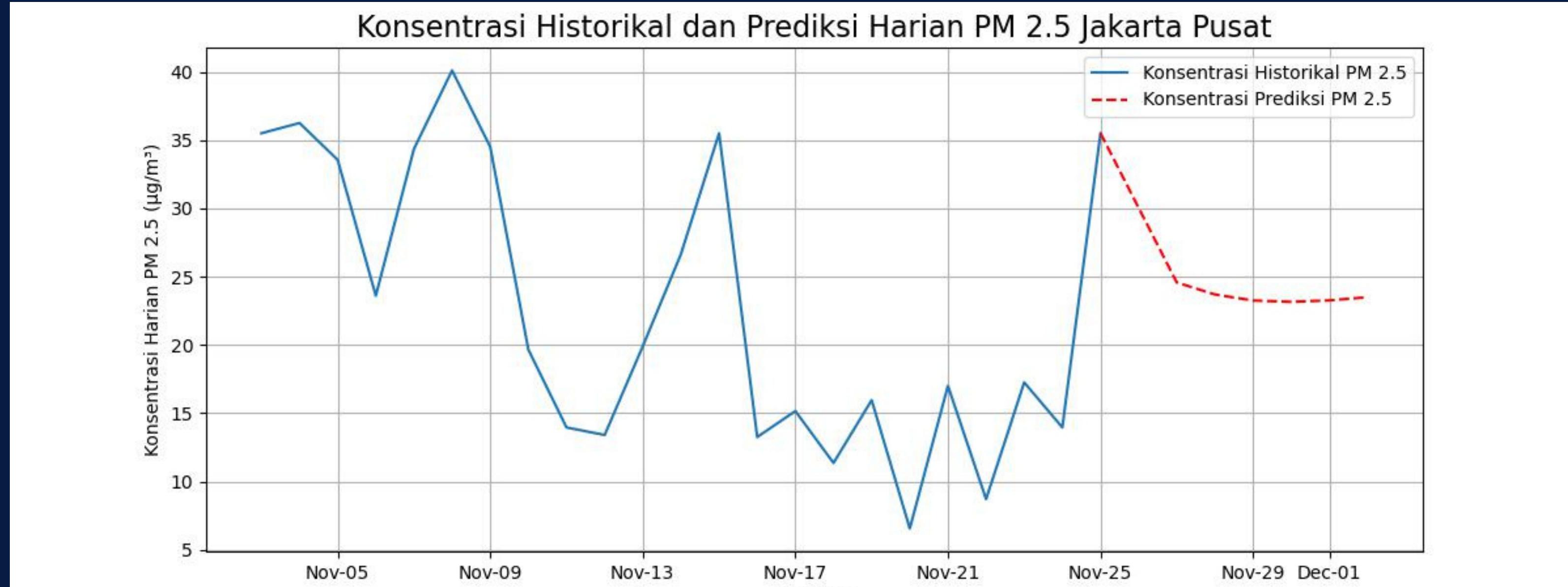
MODEL PERFORMANCE



Gambar di atas adalah visualisasi bagaimana performa dari model LSTM kami terhadap prediksi di training dan test dataset. Model terlihat cukup mumpuni untuk melakukan prediksi. RMSE pada training dataset sebesar 38.8133, sedangkan RMSE di test dataset sebesar 33.3895.

Kedepannya, performa model akan terus dikembangkan sehingga RMSE dapat ditekan hingga lebih rendah dari RMSE yang ada saat ini.

FORECASTING KONSENTRASI PM 2.5



Sesuai dengan judul proyek ini, kami melakukan prediksi seberapa besar konsentrasi PM 2.5 dalam beberapa hari mendatang. Prediksi dilakukan pada data harian terbaru yang kemudian dipelajari oleh trained model. Pada visualisasi di atas, model melakukan prediksi konsentrasi PM 2.5 di Jakarta Pusat dalam 7 hari kedepan, ditandai menggunakan garis putus berwarna merah. Sementara itu untuk garis berwarna biru, adalah konsentrasi historikal dari PM 2.5.

Conclusion and Future Work

Dari proyek ini, kami menyimpulkan jika konsentrasi dari PM 2.5 memiliki pola pada periode waktu tertentu, dan faktor klimatologi terlihat memiliki korelasi dan juga impact meskipun terbilang cukup kecil. Lalu dikarenakan konsentrasi PM 2.5 ini terpengaruh oleh historikal data sehingga menggunakan LSTM adalah pilihan yang cukup tepat.

Kami akan coba untuk terus mengembangkan model yang ada saat ini menjadi lebih baik lagi. Selain itu, kami juga berkeinginan untuk dapat membangun sebuah web app yang dapat menampilkan informasi persebaran dari konsentrasi PM 2.5 di Jakarta dan sekitarnya secara rea-timel beserta dengan prediksinya sehingga dapat memiliki manfaat bagi masyarakat yang membutuhkannya.

Anda dapat menngunjungi halaman <https://github.com/datawithalvin/pm25-prediction-with-lstm> untuk melihat sumber kode yang kami gunakan di dalam proyek ini.

Terima kasih.

