Air Quality (PM 2.5) Forecasting in NYC

Image souce from freepik

Background

Air pollution is the **2nd highest risk factor for non-communicable diseases**. It is estimated to cause 4.2 million premature deaths worldwide per year in 2019; this mortality is due to exposure to fine particulate matter, which causes cardiovascular and respiratory disease, and cancers.[1]

[1]https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health#:~:text=Ambient%20(outdoor)%20air%20pollution%20is,Asia%20and%20Western%20Pacific%20Regions.

Image souce from freepik

Background

Particulate Matter 2.5 (PM2.5) refers to fine inhalable particles with a diameter of 2.5 micrometers or smaller.

Despite their small size, PM2.5 particles can deeply penetrate the respiratory system, bypassing natural defenses, and enter the bloodstream, potentially leading to respiratory and cardiovascular diseases, including aggravated asthma, lung infections, and even heart attacks.

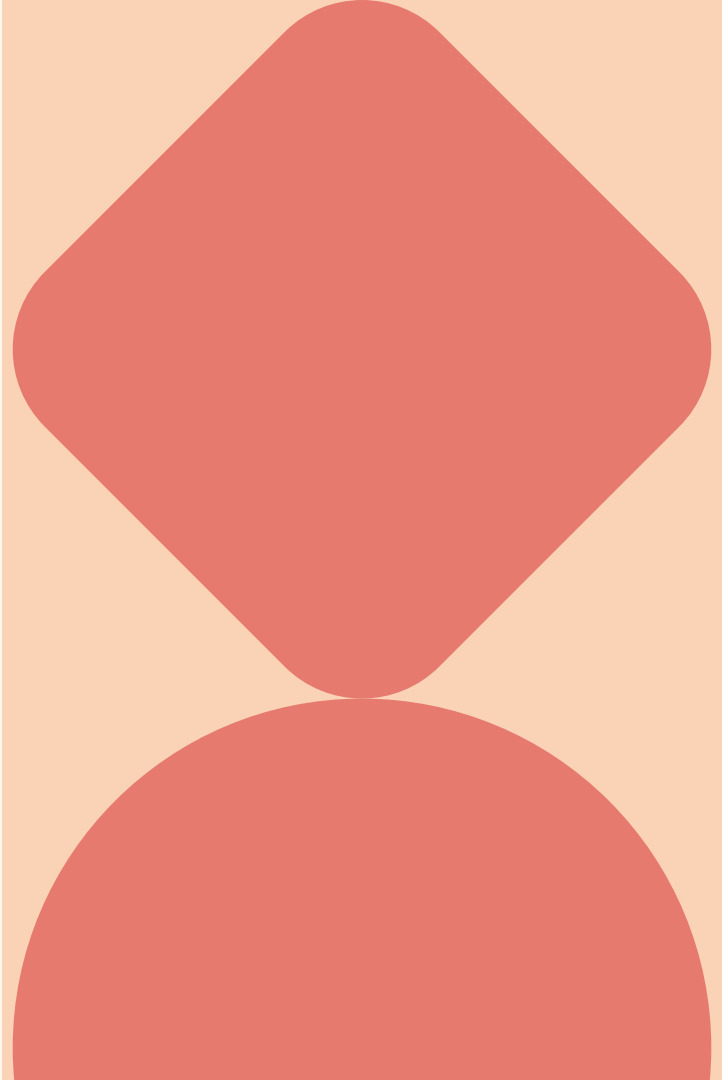
Image souce from freepik

Background

As a challenge for myself, I tried to apply my forecasting skill on air quality data in NYC, specifically to make a **one-step ahead forecast model for PM 2.5.**

I don't make Root Cause Analysis or hypothesis. **The goal is simply to make a working forecasting model.**

Data Analysis and Forecasting



Step-by-step analysis

Data cleaning.
The data is quite clean, some dates are missing but no big deal. I only change data type of ‘date’ to datetime.

Exploratory Data Analysis.
Simple visualization of the concentrations of PM2.5 over time in NYC.

Forecasting using ARIMA.

Data Overview

There are 6317 row, no null value. PM2.5 is in μg/m^3

	date	pm25
0	1999-07-01	20.000000
1	1999-07-02	23.900000
2	1999-07-03	36.700000
3	1999-07-04	39.000000
4	1999-07-05	28.171429

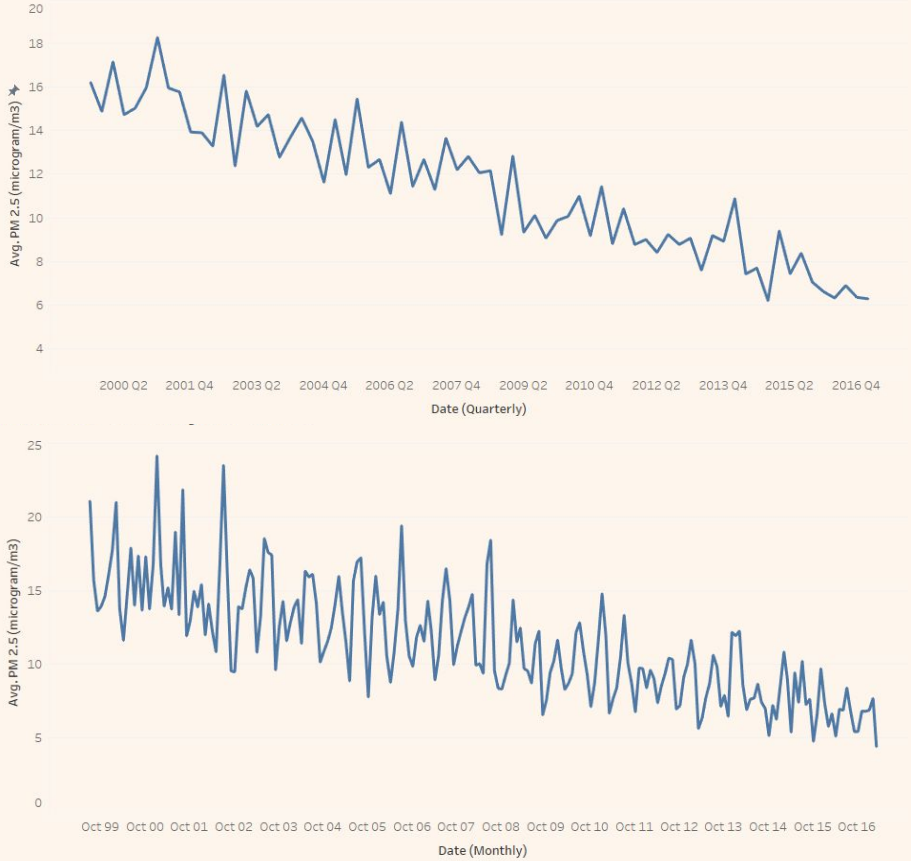
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6317 entries, 0 to 6316
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   date    6317 non-null   object
 1   pm25    6317 non-null   float64
dtypes: float64(1), object(1)
memory usage: 98.8+ KB
```

PM2.5 yearly in NYC

The air concentration of PM2.5 in New York City is **trending down** over year.

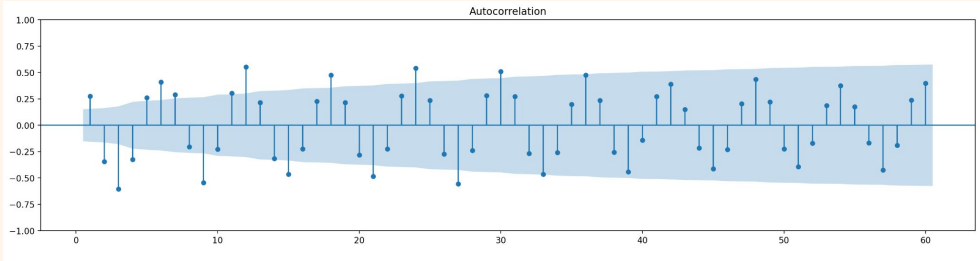
Seasonality Analysis

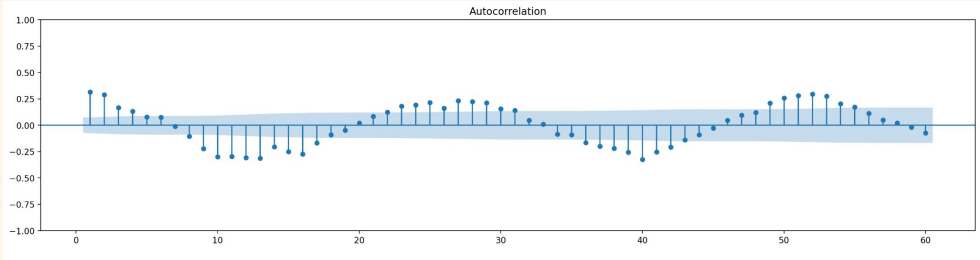
There are no apparent seasonality in quarterly or monthly. To help ascertain the seasonality, I will use statistical tools.



Seasonality Analysis

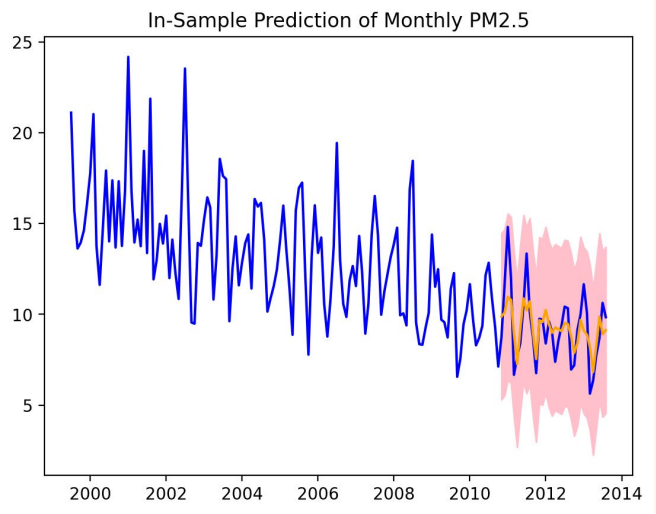
After detrending the data, I made autocorrelation plot for weekly and monthly data. I found that there are **3 month seasonality** and around **12-13 weekly seasonality**.





Building ARIMA Model (🚨technical alert🚨)

With the help of auto_arima, I built ARIMA model from daily data, while from weekly and monthly data I use SARIMA model. I let auto_arima chose the order of AR, MA, the differencing as well as the seasonal component. The selection is based on the lowest AIC (Akaike Information Criterion)



Timeframe	Test set RMSE
Daily	4.62
Weekly	3.274
Monthly	1.644

So, what is the forecast of PM2.5 next month in NYC?

It is predicted that at April 2017, the average PM2.5 concentration is **5.89 μg/m³** (95% confidence interval within 1.6 - 10.18)

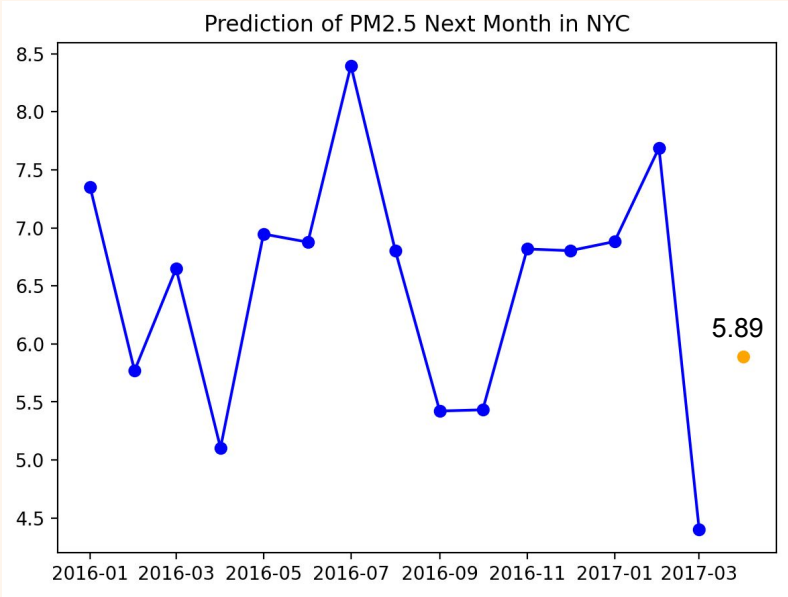


Image souce from freepik

How about Indonesia?

Based on BMKG report[2], the average concentration of PM2.5 in **Jakarta 2017** is **25.7µg/m^3**, far above the WHO standard of 10 µg/m^3 and now become 5 µg/m^3.

And it doesn't stop there. Today (28 August 2023) the PM2.5 concentration is **41 µg/m^3** 😔

[2]https://gawpalu.id/bgb/index.php/bgb/article/download/7/3/5

Let’s make a better future!

Thanks for reading my small project!