AIML
CA2

Air Quality Forecasting

Wong Zhao Wu

Problem Overview & Modelling Objectives

The training dataset consist of 328 observations from March 2004 to January 2005 with hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Total Nitrogen Oxides (NOx) and Ozone. Test set, on the other hand, consist of 63 observations from February 2005 to April 2005.

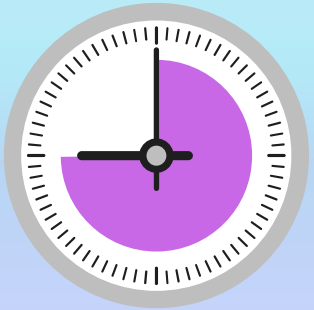
Modelling Objectives

To forecast the Air Quality of Carbon Monoxide CO, Nonmethane Hydrocarbon NMHC, Nitrous Oxide NOx and Ozone O3 for the next 63 Days(test size) using Univariate and Multivariate Statistical Model.

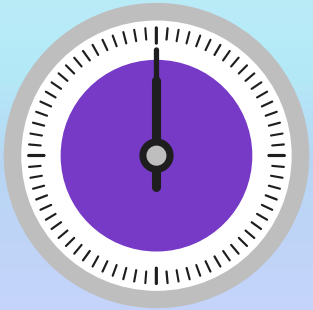
CO

NMHC

NOx



O3



Data Preprocessing

Date	T	RH	CO	NMHC	NOx	O3	_merge
2004-03-10	12.020833	54.883334	1316.500000	912.250000	1167.250000	1096.041667	both
2004-03-11	9.833333	64.069791	1244.062500	851.802083	1277.187500	885.031250	both
2004-03-12	11.292708	51.107292	1281.562500	1008.229167	1101.718750	1084.218750	both
2004-03-13	12.866319	51.530903	1330.555556	992.822917	993.159722	1245.781250	both
2004-03-14	16.016667	48.843750	1360.927083	943.854167	1001.104167	1234.177083	both
...	...	...	...	...	...	...	...
2004-08-27	NaN	NaN	-200.000000	-200.000000	-200.000000	-200.000000	right_only
2004-12-15	NaN	NaN	-200.000000	-200.000000	-200.000000	-200.000000	right_only
2004-12-16	NaN	NaN	-200.000000	-200.000000	-200.000000	-200.000000	right_only
2005-01-03	NaN	NaN	-200.000000	-200.000000	-200.000000	-200.000000	right_only
2005-01-04	NaN	NaN	-200.000000	-200.000000	-200.000000	-200.000000	right_only

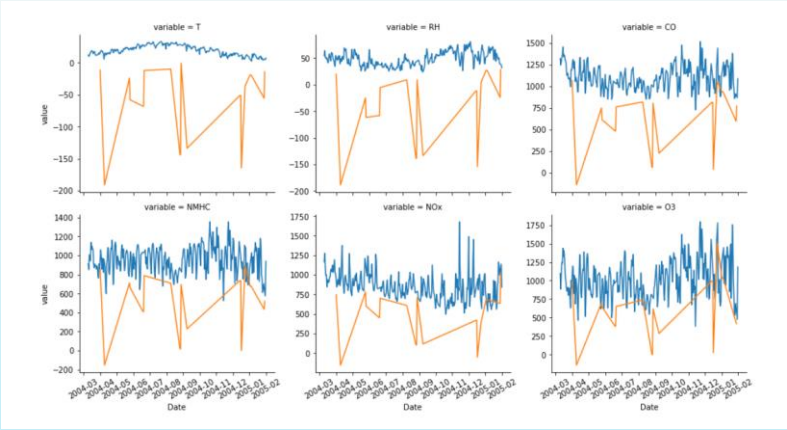
328 rows × 7 columns

Pivoting Table

Pivot the dataframe from **stacked format to unstacked format**.

E-Nose Drifting

By using **Temperature< 0 as the indicator**, I have noticed the **randomly scattered anomaly data across all gasses.**



Missing Values

-200 is observed as for the **missing values.**
When either one of the gasses is recorded -200, the rest of the gases will also be -200.

Front-Fill Anomalies

Front Filling the anomaly data for the **"-200" missing value** and **"drifted" value**.

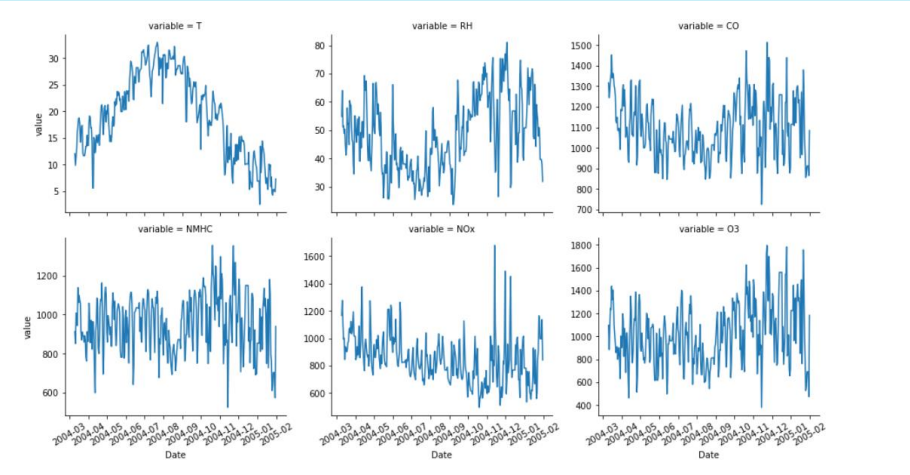
The same preprocessing process is done for both Train set and Test set.

Time Series Analysis : Stationarity Series

To model a time-series using ARMA model, we need to ensure the time-series is stationary, that is the the **mean, variance and covariance does not vary with time.**

Data Visualisation

The series does not have a specific trend expect for T with a seemingly inverse “U” shape. However, let us utilize the Dickey Fuller Statistical Test before jumping into the conclusion of stationarity.



Augmented Dickey Test

at Significant level, p=0.05 with,
H_0: Time-Series is Non-Stationary
H_1: Time-Series is Stationary

	Variable	P_Value	Stationary
0	T	7.839651e-01	False
1	RH	5.004329e-04	True
2	CO	3.744836e-13	True
3	NMHC	4.124349e-03	True
4	NOx	1.459045e-03	True
5	O3	1.514280e-02	True

Time Series Analysis : ACF & PACF

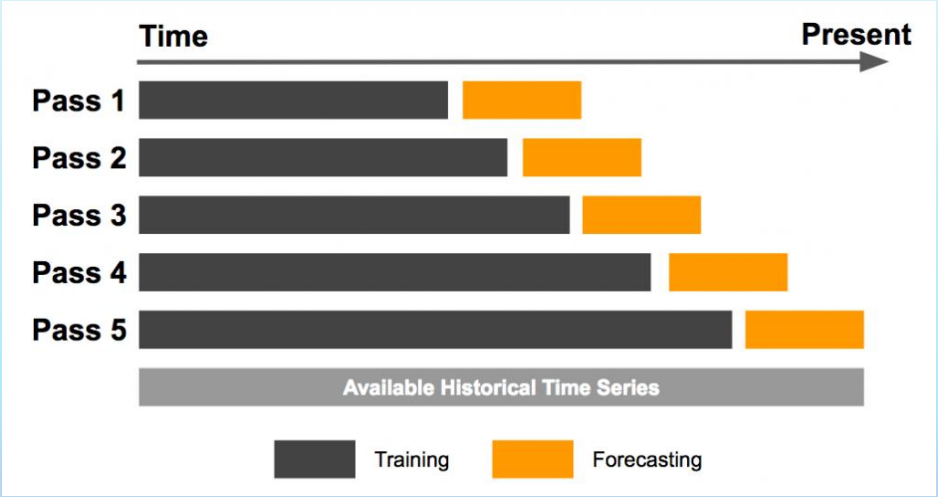
After ensuring our time-series is stationary, ACF and PACF can be helpful for us to **visualising potential trend** as well as **interpret the initial orders** for my time series model.

Observation Summary

Gasses	AR (p lags)	MA (q lags)	Seasonal Trend (m)	Potential Initial Model
CO	2	3	-	ARMA(2,3)
NMHC	2	2	Weak Trend every 7 Iteration	ARMA(2,2)/ SARIMA(2,0,2)(1,0, 4,7)
NOx	2	5	Weak Trend every 7 Iteration	ARMA(2,2)/ SARIMA(2,0,5)(1,0, 4,7)
O3	2	3	Weak Trend every 7 Iteration	ARMA(2,3)/ SARIMA(2,0,3)(0,0, 4,7)
CO (d=1)	5	3	Weak Seasonal Trend every 28 Iteration	ARIMA(5,1,3)/SARI MA(5,1,3)(1,1,2,28)
NMHC (d=1)	6	5	Weak Seasonal Trend every 7 Iteration	ARIMA(6,1,5)/ SARIMA(6,1,5)(3,1, 4,7)
NOx (d=1)	6	3	Weak Seasonal Trend every 21 Iteration	ARIMA(6,1,3)/ SARIMA(6,1,5)(2,1, 2,21)
O3 (d=1)	6	5	Weak Seasonal Trend every 21 Iteration	ARIMA(6,1,5)/SARI MA(6,1,5)(1,1,2,21)

Time Series Analysis :
Sampling Techniques & Baseline Estimation

An unbiased model evaluation technique is crucial for us to understand the general potential of the model. One way to achieve it is through **training and evaluating the model multiple time each with different window of dataset**. This is where Expanding Window technique comes into play in addition to the **train set split** with test size of 63 observations.



Original article: https://eng.uber.com/omphalos/

Average Baseline Forecast

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

I will be using an Average Forecast that **Predicts Current Observation based on the Mean** of the historical data.
The performance will be **Evaluated Using the Same Expanding Window** techniques as mentioned above.

Time Series Analysis : ARIMA Forecast

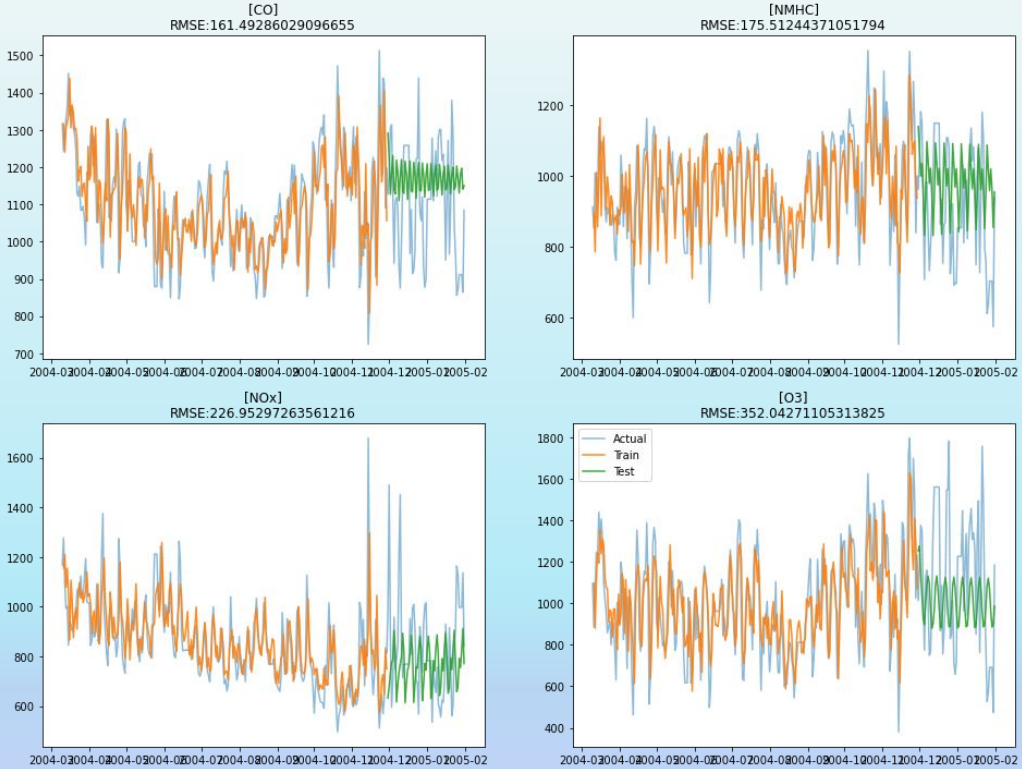
Based on the interpreted model from PACF and ACF, I've build the initial ARMA model. However, I realise that it is just **performing slightly better than the baseline** average forecast for d=0.

Gasses	AR(p significant lags)	I(d integrate lags)	MA(q significant lags)	RMSE	AIC	Baseline
CO	3	1	4	130.167628	1603.053095	153.366220
NMHC	6	1	6	134.313186	1621.743978	146.519079
NOx	8	1	8	146.977765	1664.471816	183.332560
O3	7	1	7	242.780002	1795.741758	257.749546

	RMSE	AIC
CO (2, 0, 3)	150.163563	1622.959167
NMHC (2, 0, 2)	145.953269	1661.373538
NOx (2, 0, 5)	172.189272	1689.601666
O3 (2, 0, 3)	253.258412	1805.618220
CO (5, 1, 3)	130.520877	1609.662279
NMHC (6, 1, 5)	140.630756	1627.391221
NOx (6, 1, 3)	153.317551	1673.488005
O3 (6, 1, 5)	248.079196	1799.304067

Then, I explore ARIMA model further by Grid Searching the orders for my model. I will be **fitting different model for each of the gasses** and perform **hyperparameter tuning independently**, using the **Expanding Window technique**. I will be minimizing the **Root Mean Squared Error(RMSE) and Aikeike Information Criteria(AIC)** to ensures maximum likelihood and simpleness of my model.

Time Series Analysis : ARIMA Forecast



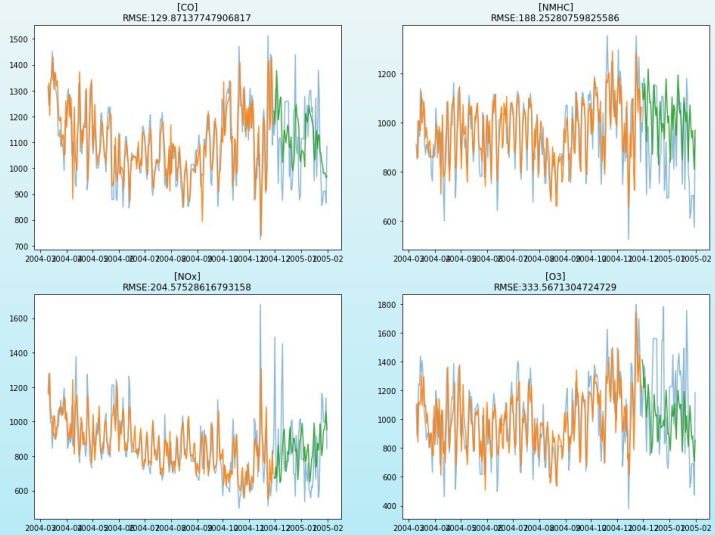
Based on the cross-validated result, it seems that ARIMA is able to **Perform better than the Interpreted Value**. Let us visualise the prediction by plotting it out.

By visualising the test set, I realise that the Univariate ARIMA Model is just **oscillating with a specific trend** around the average. Besides, I also realised that ARIMA Model works fine for NMHC and NOx however it seems to be underfitting for CO and O3.

Time Series Analysis : ARIMAX Forecast

Correlation Matrix

It seems that the Exogenous Variable (i.e. T, RH) has **moderate linear relationship** with our target variables which corroborate the notion of Multivariate Model.



From the cross-validated result as well as through visualizing the final test result, I have noticed that ARIMAX is much **better at capturing the general trend of movement** for the data than ARIMA model.

Besides, I also printed out the model summary and realise that the **coefficients are significantly different from zero**, implying that eXogenous variables is useful for our forecasting.

Gasses	AR(p significant lags)	I(d integrate lags)	MA(q significant lags)	RMSE	AIC
CO	2	1	4	169.775843	1571.226480
NMHC	5	1	8	142.601644	1602.235755
NOx	7	1	7	155.411699	1649.813866
O3	7	0	8	262.8634024	1811.5783

T	16.0404
RH	3.9266

T	12.5272
RH	1.3188

T	-12.2610
RH	-3.6351

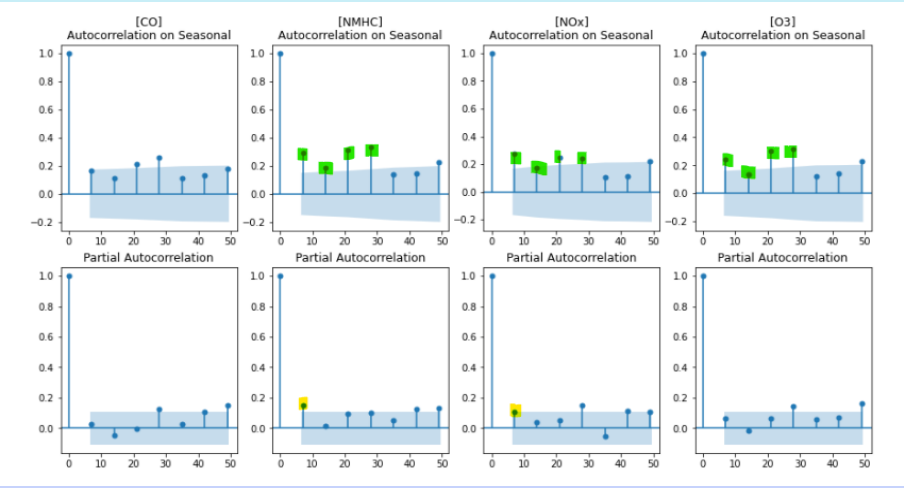
T	3.0492
RH	6.6162

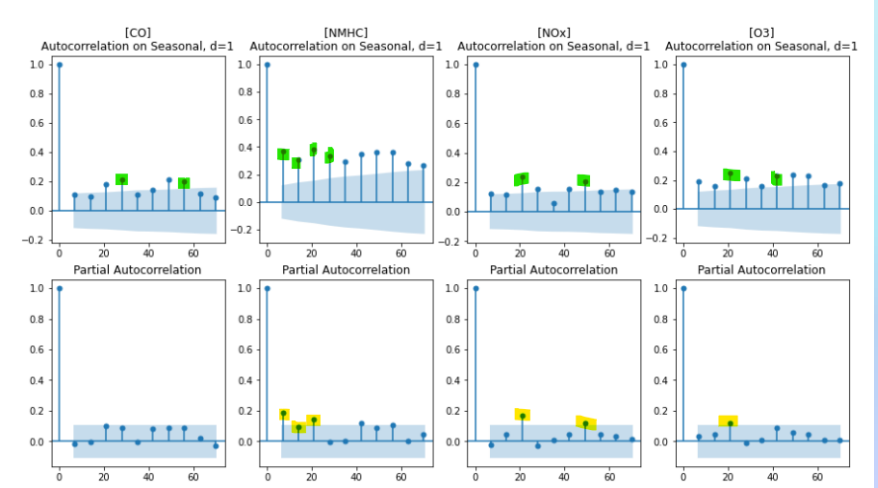
Time Series Analysis : Seasonal ARIMAX

Seasonal ACF and PACF

By analysing the ACF and PACF of seasonal trend for multiples of 7, I can interpret the P and Q order for the Seasonal Component of ARIMA as documented below:

Gasses	d	m	P	Q
CO	0	7	None	None
NMHC	0	7	1	4
NOx	0	7	1	4
O3	0	7	0	4
CO (d=1)	1	28	0	2
NMHC (d=1)	1	7	3	4
NOx (d=1)	1	21	2	2
O3 (d=1)	1	21	1	2



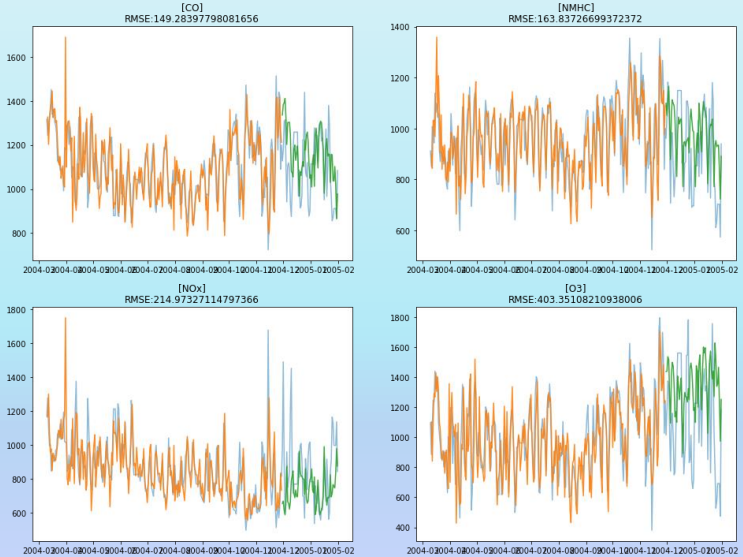


Time Series Analysis : SARIMAX Forecast

Based on the interpreted model from seasonal PACF and ACF, I have build the model as a comparison towards our cross-validated result.

	RMSE	AIC
NMHC : (2,0,2),(1,0,4,7)	173.150673	1625.627387
NOx : (2,0,5),(1,0,4,7)	235.109636	1709.281055
O3 : (2,0,3),(0,0,4,7)	351.066904	1813.067125
CO : (5,1,3),(1,1,2,28)	186.473693	1292.835844
NMHC : (6,1,5),(3,1,4,7)	134.082659	1546.460306
NOx : (6,1,5),(2,1,2,21)	135.550182	1436.865677
O3 : (6,1,5),(1,1,2,21)	239.958456	1542.921261

Gas	ARIMA Order	Seasonal Order	RMSE	AIC	Baseline
CO	(4,1,1)	(1,1,3,21)	125.195855	1358.510975	153.366220
NMHC	(4,1,3)	(3,1,1,7)	125.161663	1525.092355	146.519079
NOx	(2,1,4)	(1,1,2,21)	130.787961	1427.906009	183.332560
O3	(4,1,4)	(1,1,1,21)	234.168264	1536.897802	257.749546



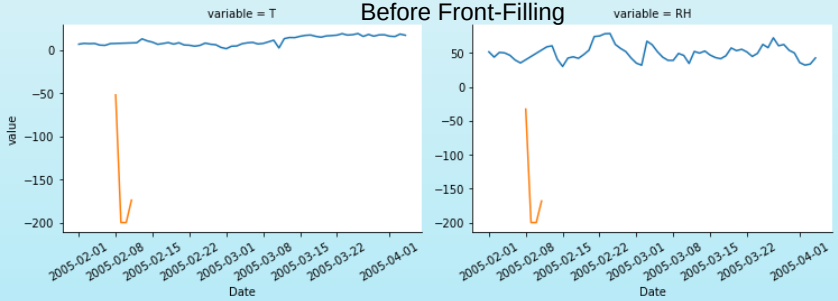
From tobservation, it seems that all SARIMAX model is able to perform better than pure ARIMAX except for O3 gas.
he SARIMAX
Hence for the final submission, I will be using SARIMAX for all 3 other gases but ARIMAX with O3.

Kaggle Test Set Prediction

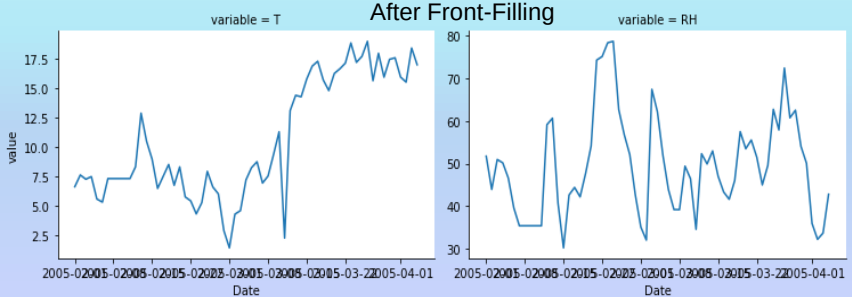
Visualise the Drifting

By visualizing the endogenous variable in my test set, I realise the same drifting event has occurred. Has I have cleaned the test set by the front-filling method.

Before Front-Filling



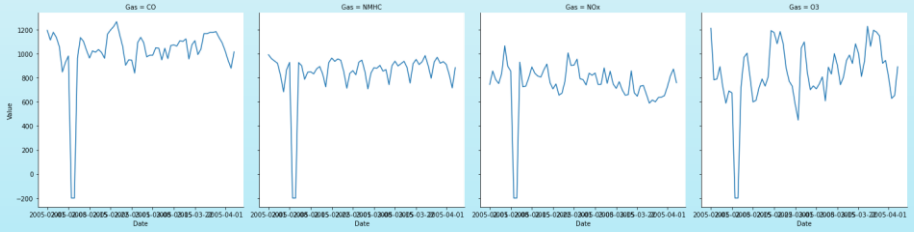
After Front-Filling



Generate Final Prediction

Finallly,I have trained my final SARIMAX model based on the orders as shown in the table below for the final Kaggle Submission via Kaggle API. Before I submit the forecasted value, I have **padded the missing values with -200**.

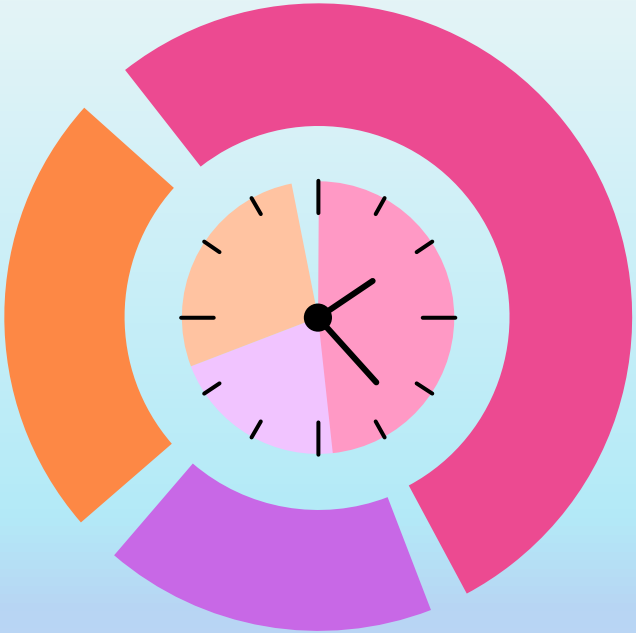
Forecasted Value Visualisation



Final SARIMAX orders

Gasses	p	d	q	P	D	Q	M
CO	4	1	1	1	1	3	21
NMHC	4	1	3	3	1	1	7
NOx	2	1	4	1	1	2	21
O3	7	0	9	0	0	0	0

Thank You



Personal Learning Journey

This is the first time that I embarked on solving a time-series related project and given the limited time I have, the best I can do is just to **interpret the ACF and PACF Plots**, **Tune a Simple ARIMA, ARIMAX and SARIMAX model** and implement some cross-validation techniques like **Expanding Window selection**. I hope that I build a more robust model by utilising **LSTM Networks** (to maybe to predict some stock price next time).