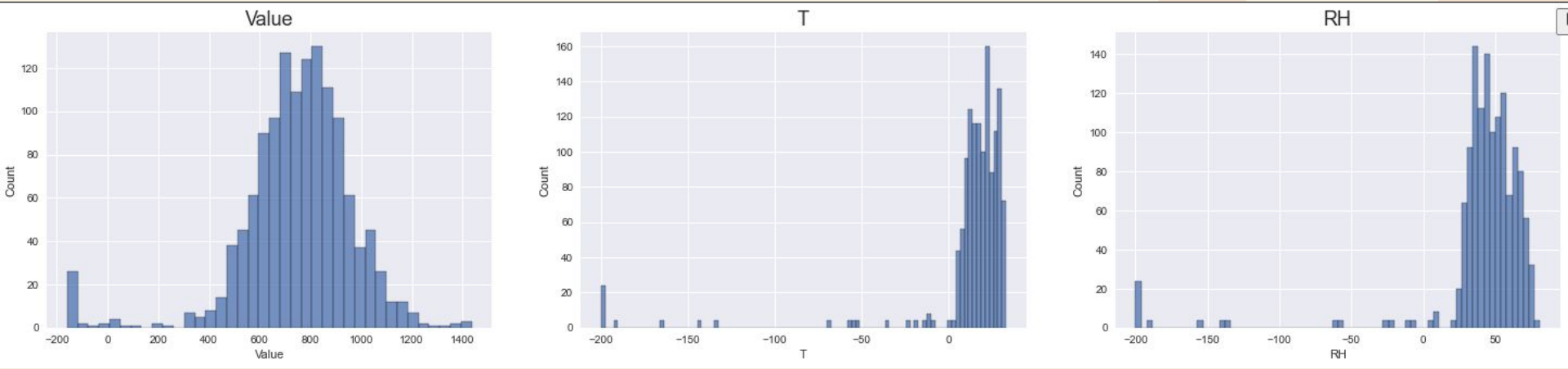
Air Pollution Forecasting

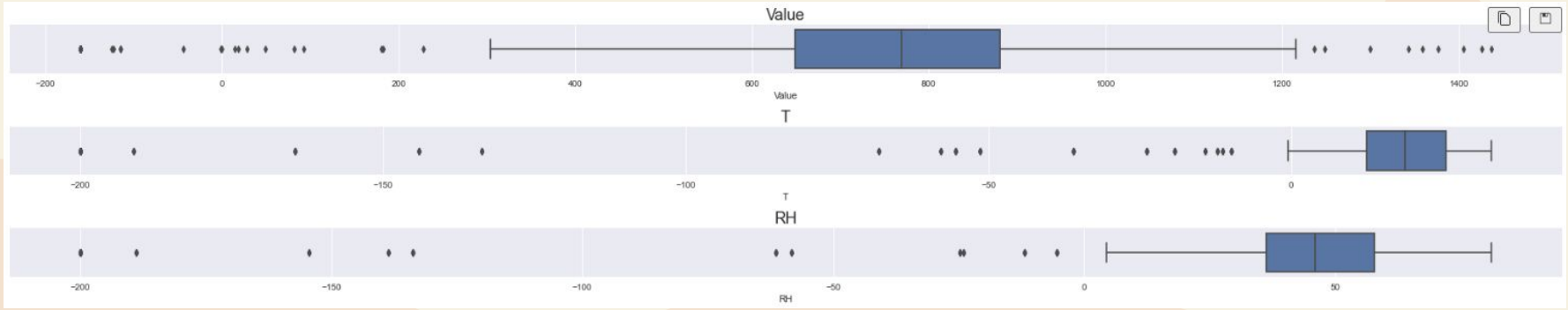
Using Time Series Modelling

Lim Jun Jie
DAAA/FT/2A/02

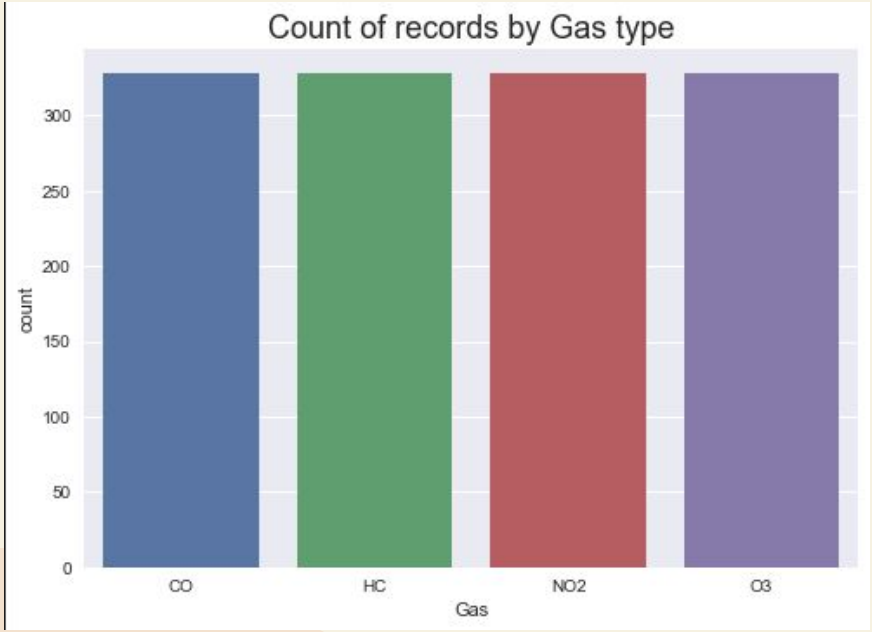
TABLE OF CONTENTS

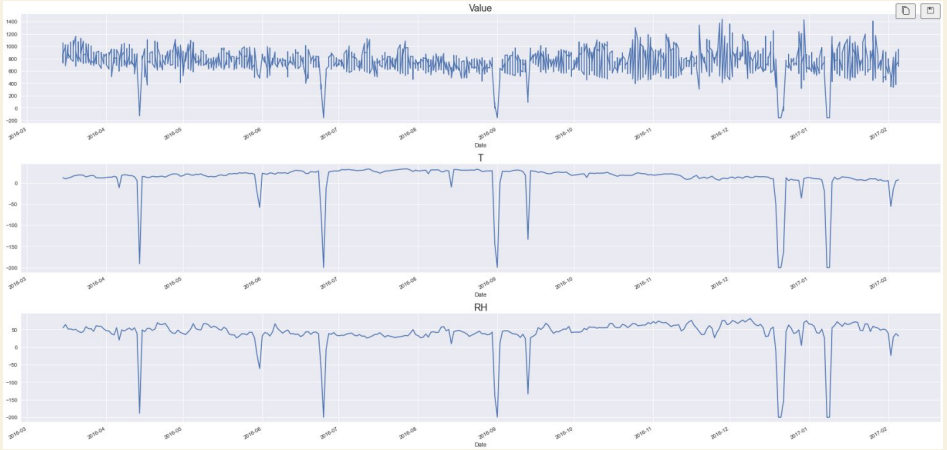
EDA - Histograms and Boxplots



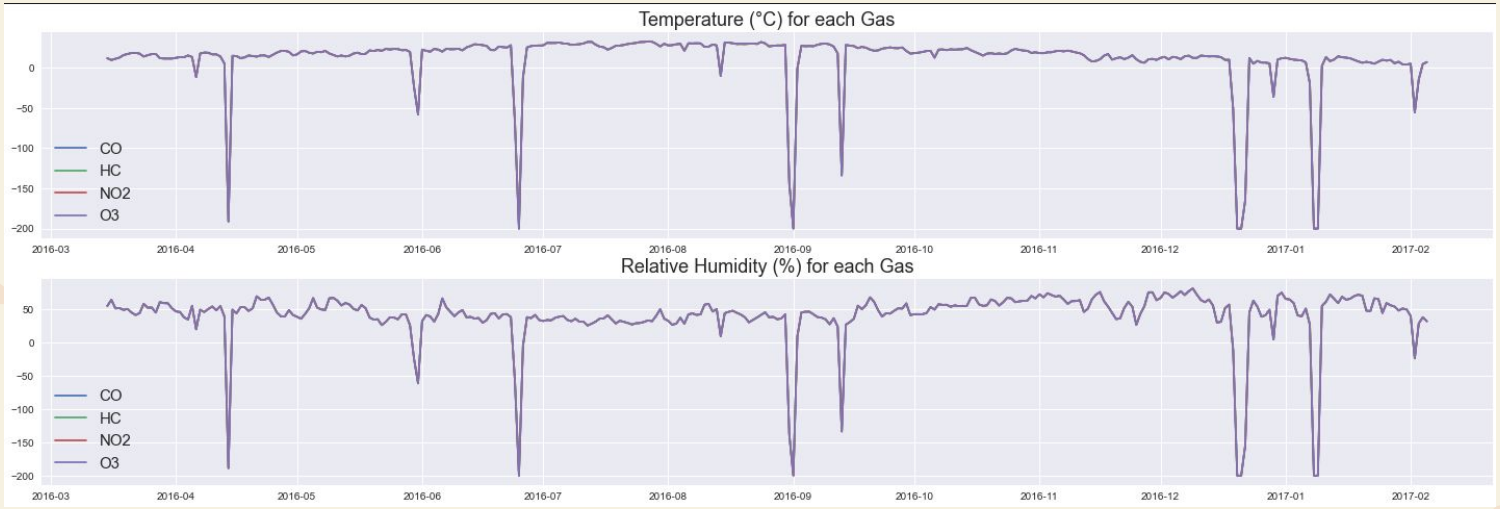


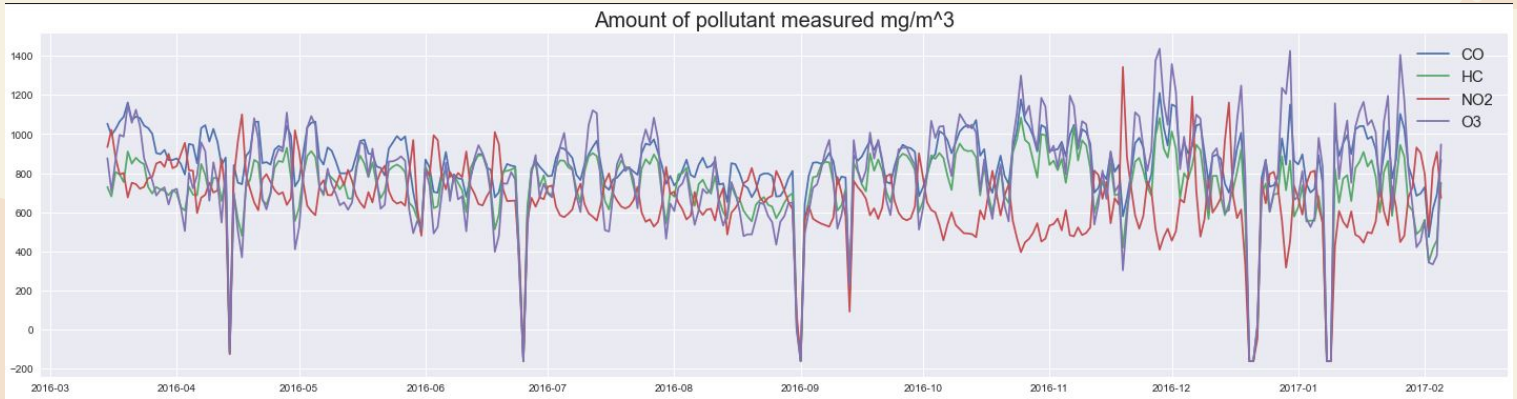
EDA - Barplot and Time Series plot





EDA - Barplot and Time series plot

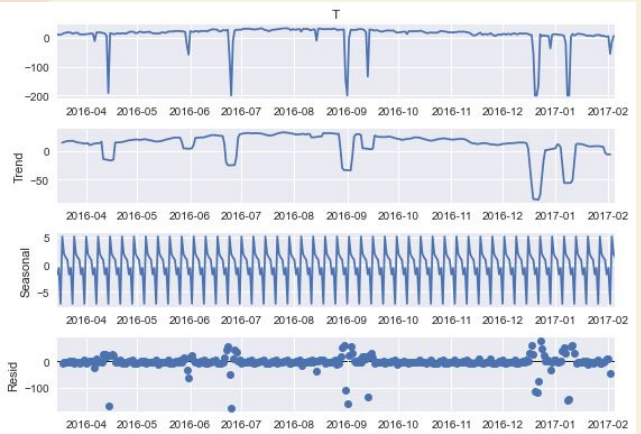


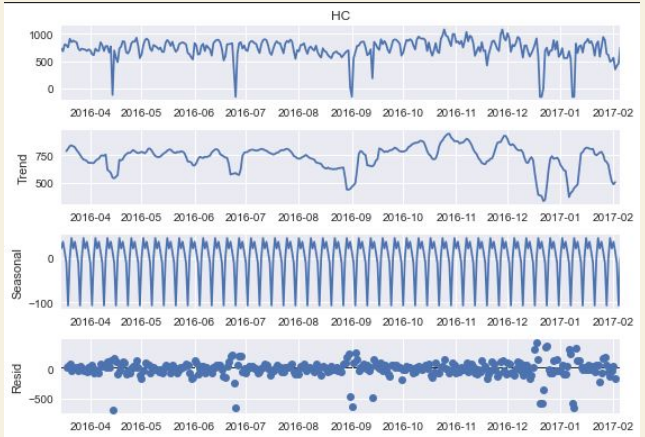


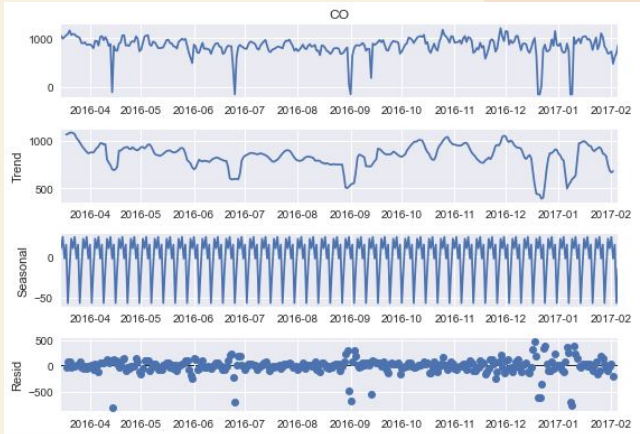
Pivot dataframe

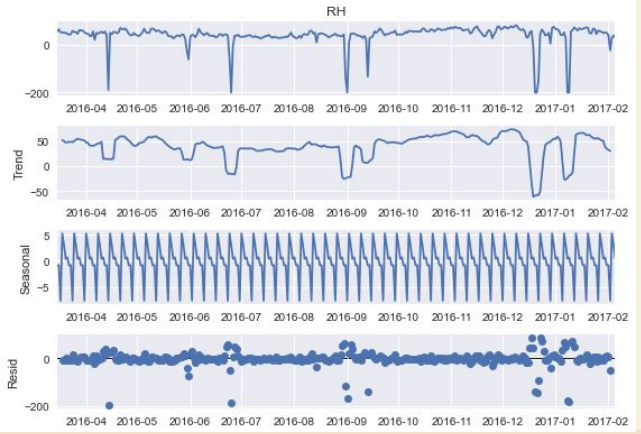
Date	CO	HC	NO2	O3	T	RH
2016-03-15	1053.200000	729.800000	933.800000	876.833333	12.020833	54.883334
2016-03-16	995.250000	681.441667	1021.750000	708.025000	9.833333	64.069791
2016-03-17	1025.250000	806.583333	881.375000	867.375000	11.292708	51.107292
2016-03-18	1064.444444	794.258333	794.527778	996.625000	12.866319	51.530903
2016-03-19	1088.741667	755.083333	800.883333	987.341667	16.016667	48.843750
...	...	...	...	...	...	...
2017-02-01	729.422222	562.650000	797.647222	553.180556	5.267708	39.614930
2017-02-02	474.291667	347.480556	508.180556	343.500000	-55.515972	-24.010417
2017-02-03	615.700000	414.475000	819.733333	334.458333	-14.272917	28.563542
2017-02-04	691.713889	458.947222	909.375000	379.513889	4.848611	37.832986
2017-02-05	867.600000	751.833333	673.741667	947.333333	7.273958	31.809375

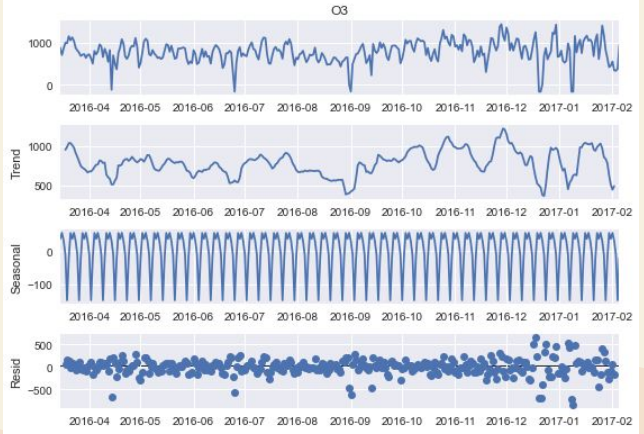
Seasonality decomposition

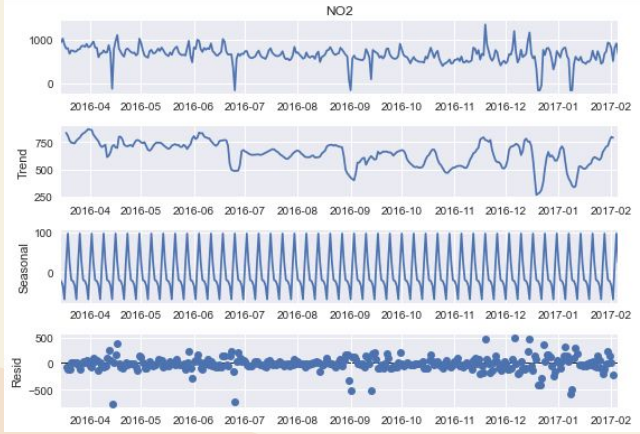












Adfuller test

```
ADF Results for T feature
----------------------------------------
ADF Statistic: -10.250990377696628
p-value: 0.0
Level of significance: 5%
Null hypothesis rejected, data is stationary


ADF Results for RH feature
----------------------------------------
ADF Statistic: -10.245827648849437
p-value: 0.0
Level of significance: 5%
Null hypothesis rejected, data is stationary


ADF Results for CO feature
----------------------------------------
ADF Statistic: -9.54352153969917
p-value: 0.0
Level of significance: 5%
Null hypothesis rejected, data is stationary
```

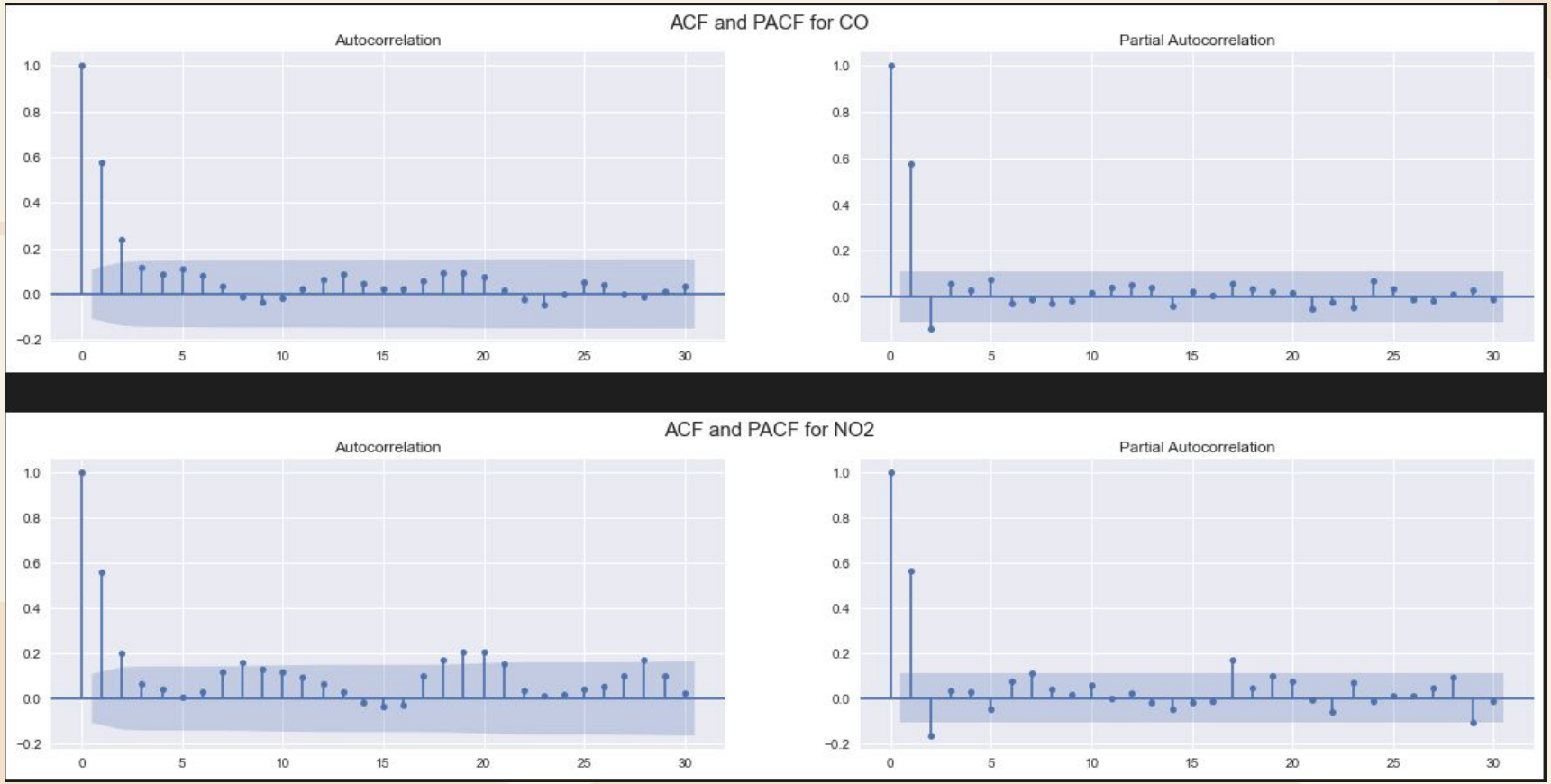
```
ADF Results for NO2 feature
----------------------------------------
ADF Statistic: -10.055200358410564
p-value: 0.0
Level of significance: 5%
Null hypothesis rejected, data is stationary


ADF Results for HC feature
----------------------------------------
ADF Statistic: -9.668344847083937
p-value: 0.0
Level of significance: 5%
Null hypothesis rejected, data is stationary


ADF Results for O3 feature
----------------------------------------
ADF Statistic: -9.569086642802421
p-value: 0.0
Level of significance: 5%
Null hypothesis rejected, data is stationary
```

Granger Causality test

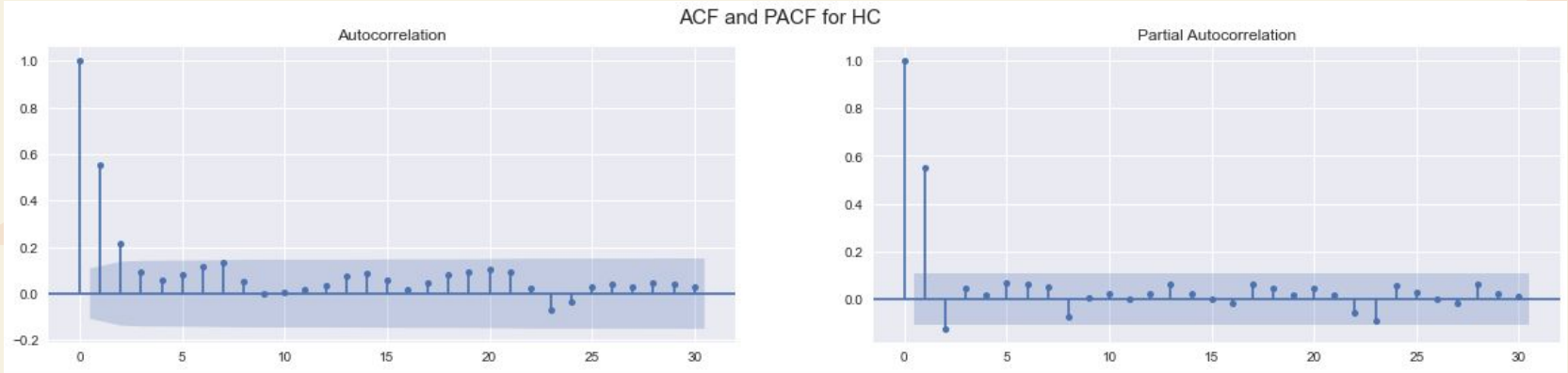
	T	RH	CO	NO2	HC	O3
CO	0.036600	0.026400	1.000000	0.101600	0.145800	0.227200
NO2	0.173200	0.317200	0.472300	1.000000	0.155200	0.085600
HC	0.008200	0.004100	0.000500	0.003000	1.000000	0.239400
O3	0.021300	0.038100	0.008400	0.005800	0.410500	1.000000

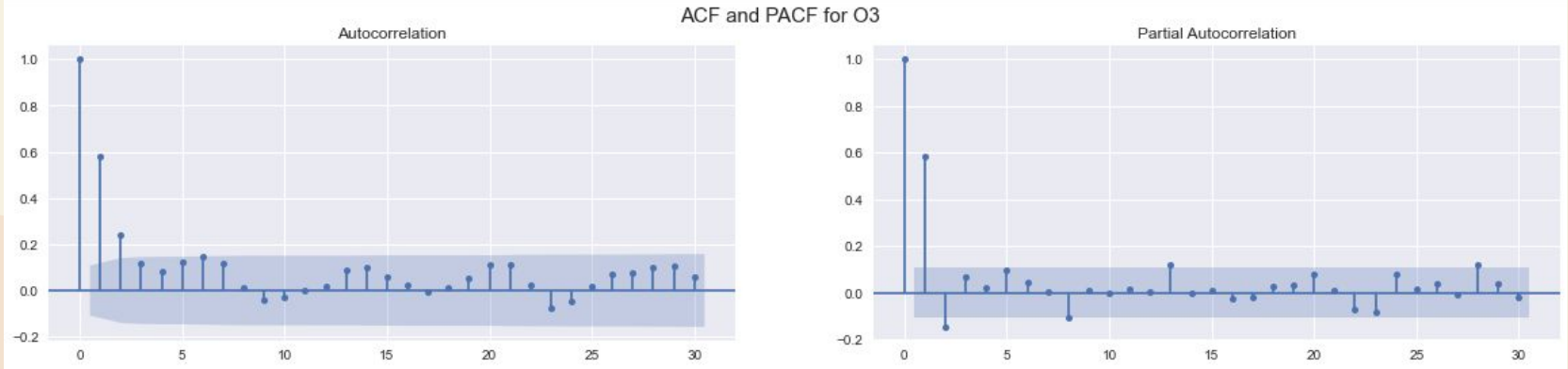
Null Hypothesis H0: Time Series X does not Granger-cause Time Series Y
Alternative Hypothesis H1: Time Series X Granger-causes Time Series Y

ACF and PACF plots (1)

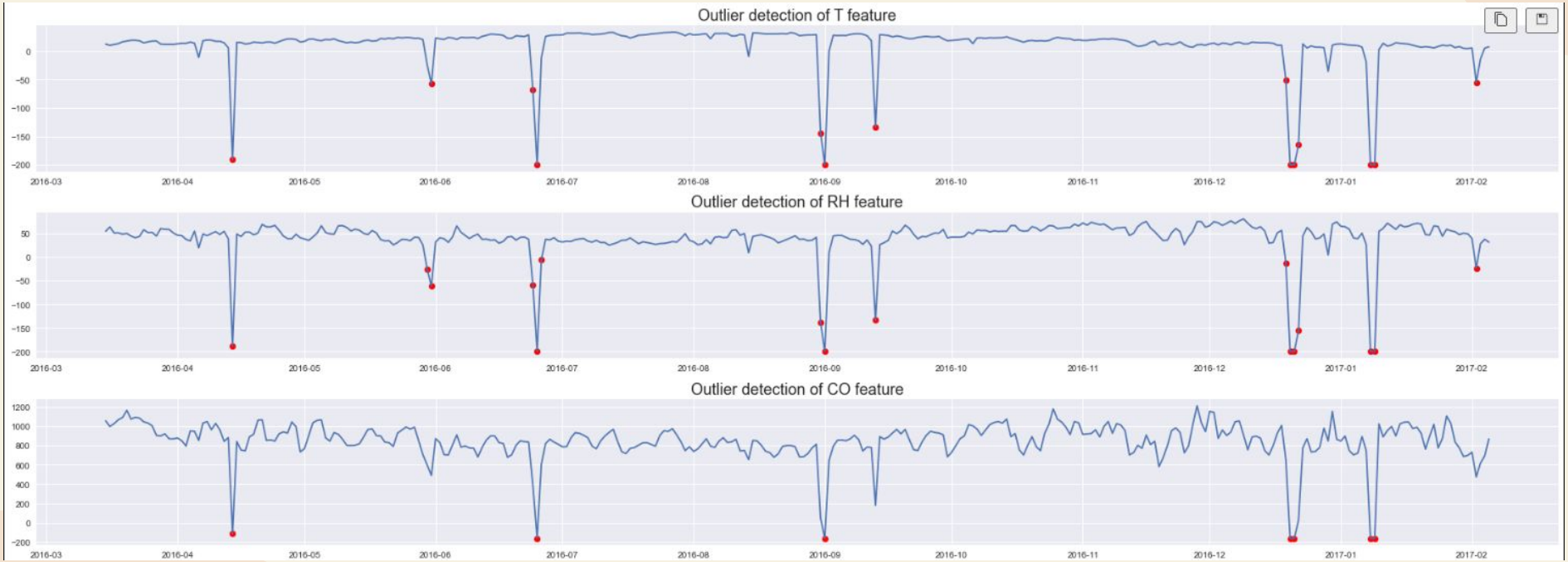


ACF and PACF plots (2)



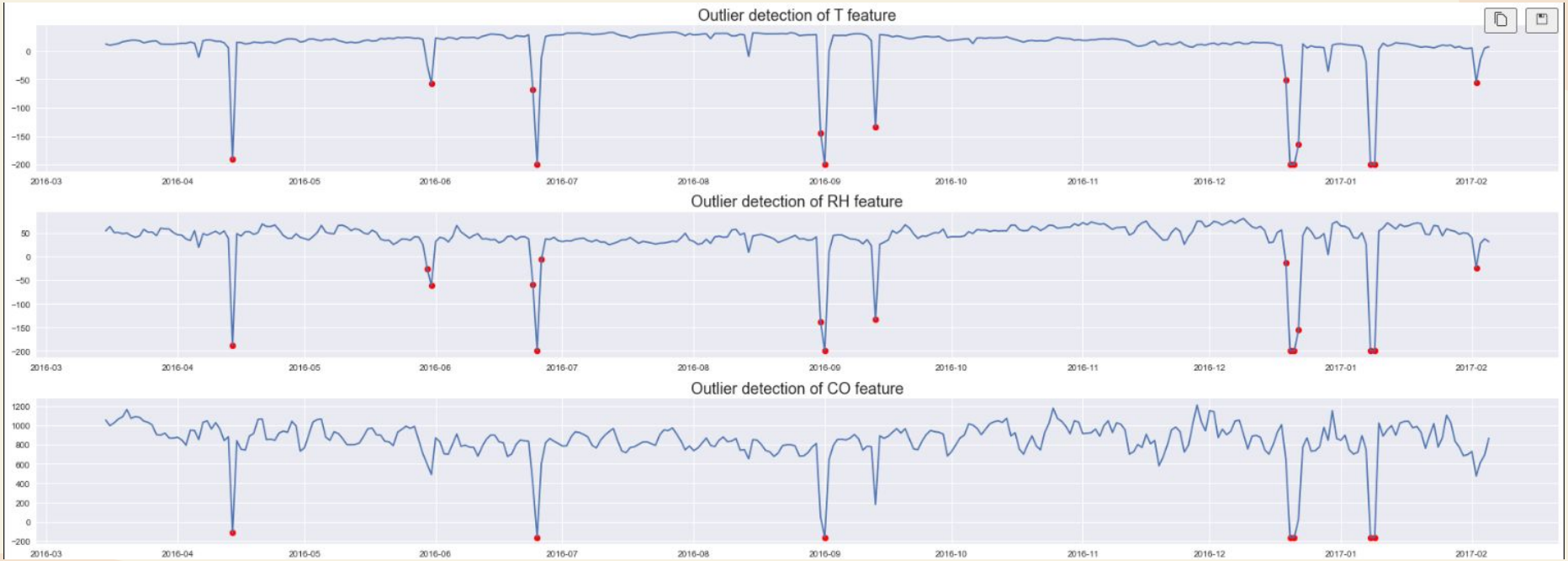


Outlier Analysis



Outlier Analysis

Outlier Analysis



Outlier Analysis

WHY VARMAX?

- Some variables Granger cause other variables in this dataset
- In theory, gases react with each other to form even more gases, thus there is some collinearity between gas concentrations

Hyperparameter Tuning Varmax

	Train MAPE (%)	Train RMSE	Test MAPE (%)	Test RMSE	AIC	BIC
0, 1	12.616746	93.001940	44.588911	149.358520	7300.499644	7413.722052
0, 2	12.134162	90.063254	30.866924	149.670481	7221.011267	7381.906269
0, 3	11.723678	88.209959	29.514559	148.052540	7218.163153	7426.730747
0, 4	10.334215	86.282657	32.109940	147.940921	7224.429956	7480.670144
1, 0	29.922634	145.915025	349.825161	268.434243	8820.895426	8934.117835
1, 1	36.775266	128.271963	346.439069	208.658522	7934.261230	8095.156231
1, 2	38.348375	126.414196	516.549668	227.083672	7716.589685	7925.157280
1, 3	40.132254	127.908473	399.196026	197.458545	7726.219458	7982.459645
1, 4	40.757700	128.191370	530.824332	254.531910	7951.227365	8255.140145
2, 0	23.783968	115.379684	122.731301	240.036457	8624.069008	8784.964009
2, 1	34.721010	111.597486	367.939592	255.157119	7708.974055	7917.541650
2, 2	31.607568	107.610008	261.515471	207.138097	7663.187635	7919.427823
2, 3	36.282750	109.804035	350.714276	190.448747	7772.486387	8076.399167

Hyperparameter Tuning Varmax (2)

	Train MAPE (%)	Train RMSE	Test MAPE (%)	Test RMSE	AIC	BIC
0, 1	12.616746	93.001940	44.588911	149.358520	7300.499644	7413.722052
0, 2	12.134162	90.063254	30.866924	149.670481	7221.011267	7381.906269
0, 3	11.723678	88.209959	29.514559	148.052540	7218.163153	7426.730747
0, 4	10.334215	86.282657	32.109940	147.940921	7224.429956	7480.670144
0, 5	11.387256	87.119676	39.309186	148.619320	7240.297538	7544.210318
0, 6	10.323629	86.071937	28.784010	148.577293	7259.309460	7610.894834
0, 7	9.954509	82.917908	30.765395	147.992387	7248.174320	7647.432286
0, 8	9.963268	81.753080	27.781674	149.502292	7271.113733	7718.044293
0, 9	12.609518	82.049876	37.713643	149.315980	7292.912331	7787.515484
0, 10	12.024751	81.936354	39.922922	148.755448	7322.438829	7864.714575
0, 11	9.981161	79.333876	39.168134	149.944919	7336.872909	7926.821247
0, 12	17.482444	119.829914	43.720534	148.734658	7957.203947	8594.824879
0, 13	9.046800	78.262280	23.027477	149.696568	7389.371806	8074.665331
0, 14	9.030894	77.974183	36.125748	150.056886	7423.246028	8156.212146
0, 15	9.558801	76.866726	59.712062	149.121846	7449.833769	8230.472479

Conclusion

10	2A02-Jun Jie		151.62686	7	5d

Using my tuned hyper parameters, i managed to clinch 10th place in this Kaggle competition.