DATS 6312 Time Series Project

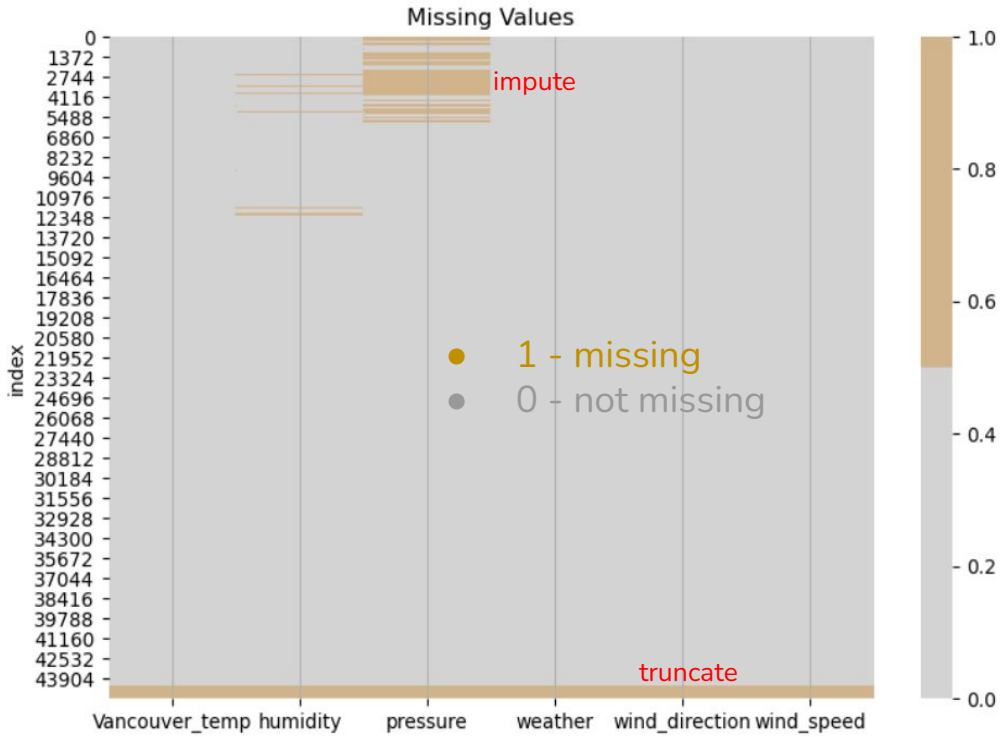
Temperature Forecasting

Author: Yuan Dang

Dataset

- **Source:** Kaggle(https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data/discussion/56293?select=temperature.csv)
- **Timestamp:** Hourly measured from *2012-10-01 12:00* to *2017-11-30 00:00*
- **Size:** 45253 rows x 7 columns
- **Features:**
  - **Temperature(K):** target
  - Humidity(%)
  - Pressure(hPa)
  - Weather Description: 37 categories: [clear, light rain, overcast clouds, mist, ...]
  - Wind Direction(°)
  - Wind Speed(m/s)
  - 36 City: 27 US cities, 3 Canadian cities, 6 Israeli cities.(Vancouver in this case)
- **Goal:** Forecast temperature for future hours

Missing Values



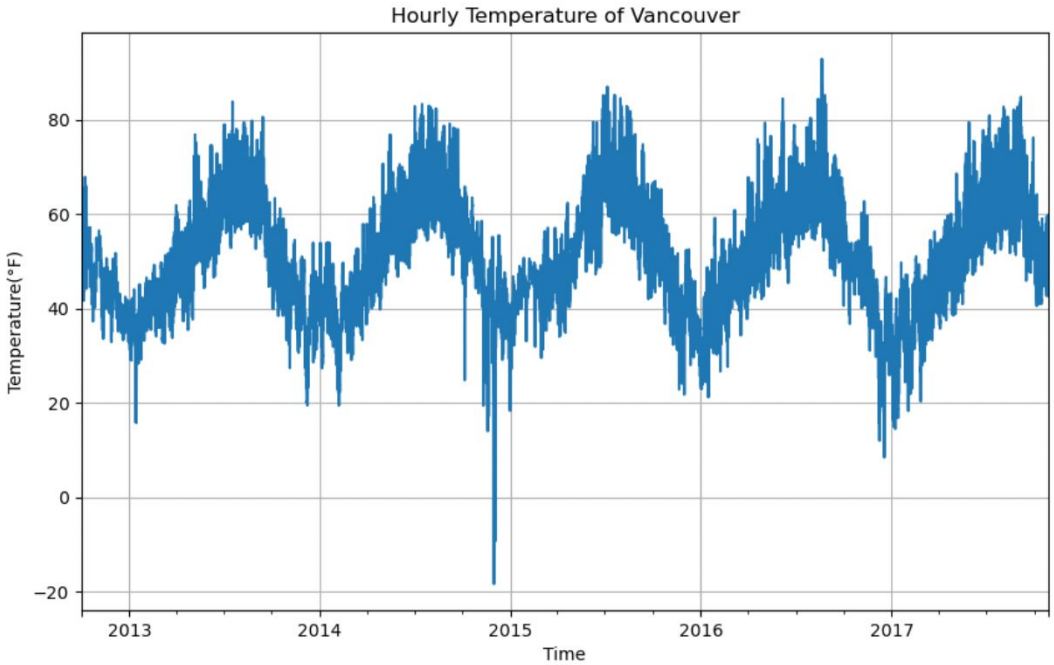
Missing values:

	var	number of missing values
0	Vancouver_temp	795
1	humidity	1826
2	pressure	4234
3	weather	793
4	wind_direction	795
5	wind_speed	795

Impute missing values

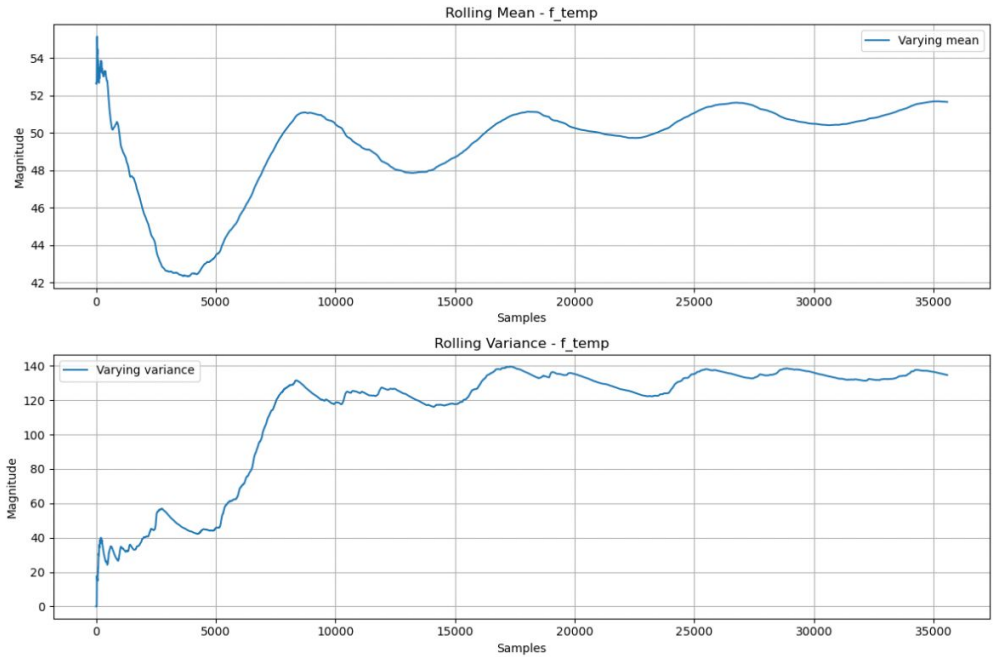
- Backward linear interpolation

Raw Data



Strong Seasonality

Stationarity

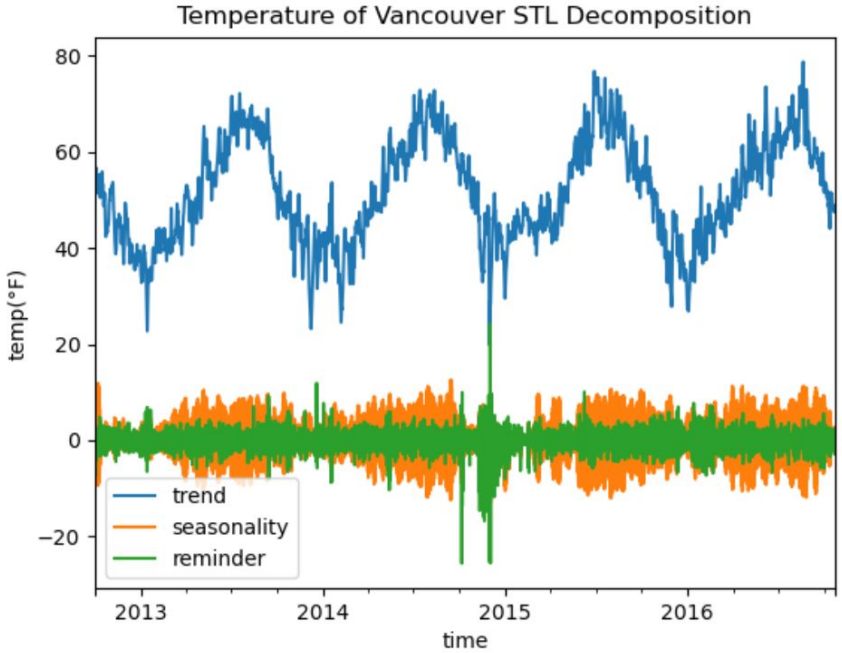



```
Results of KPSS Test:
Test Statistic            1.85279
p-value                   0.01000
LagsUsed                109.00000
Critical Value (10%)      0.34700
Critical Value (5%)       0.46300
Critical Value (2.5%)     0.57400
Critical Value (1%)       0.73900
dtype: float64
```

```
ADF Statistic: -5.510879
p-value: 0.000002
Critical Values:
    1%: -3.431
    5%: -2.862
    10%: -2.567
```

Not stationary

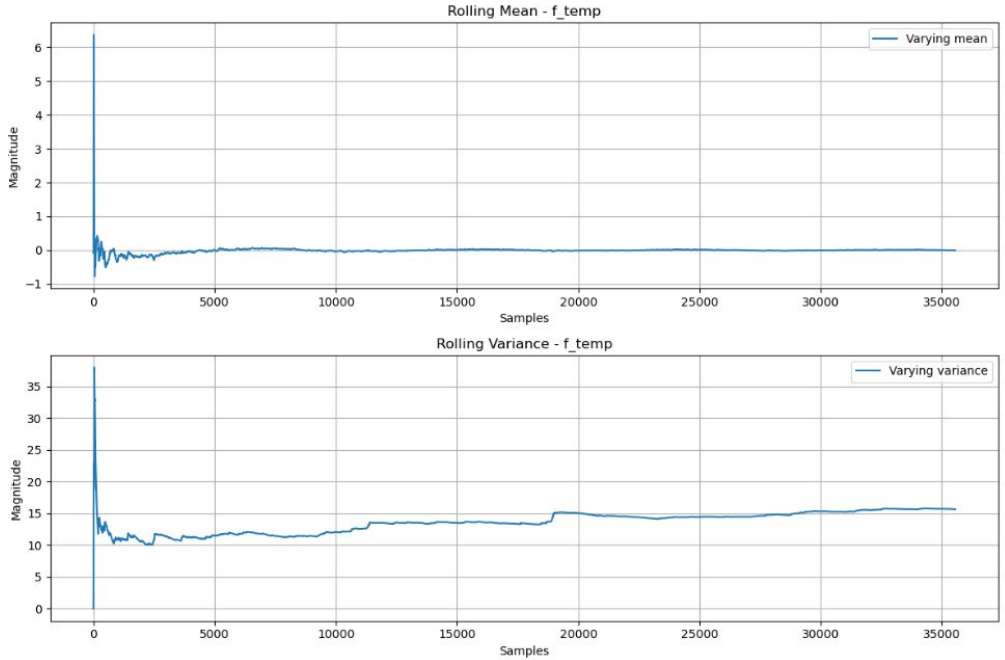
STL Decomposition



Strength of trend: 98.28%
Strength of seasonality: 88.06%

Seasonal Differencing

Seasonal period=24

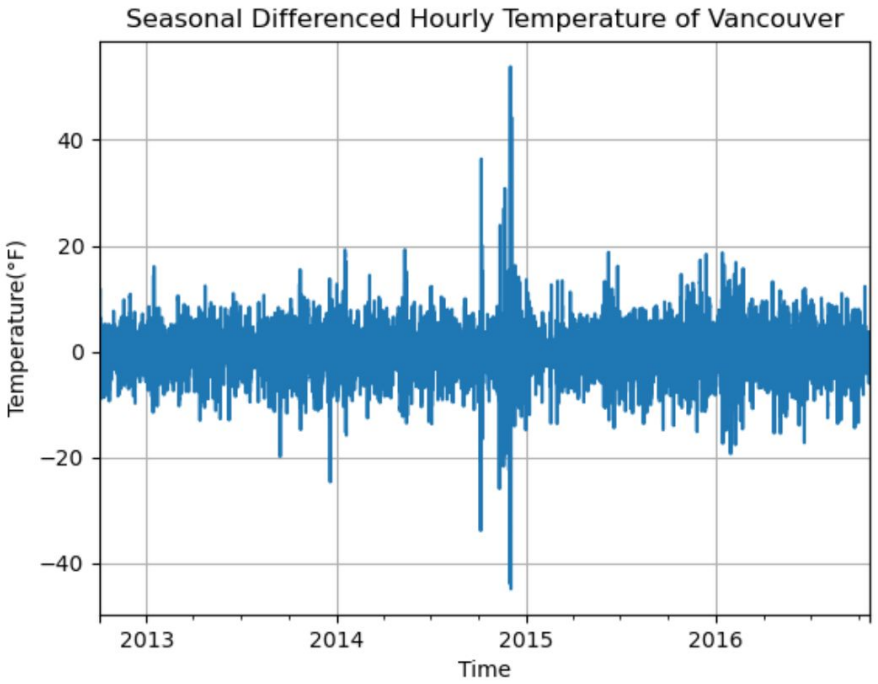



```
Results of KPSS Test:
Test Statistic            0.012732
p-value                   0.100000
LagsUsed                 98.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
dtype: float64
```

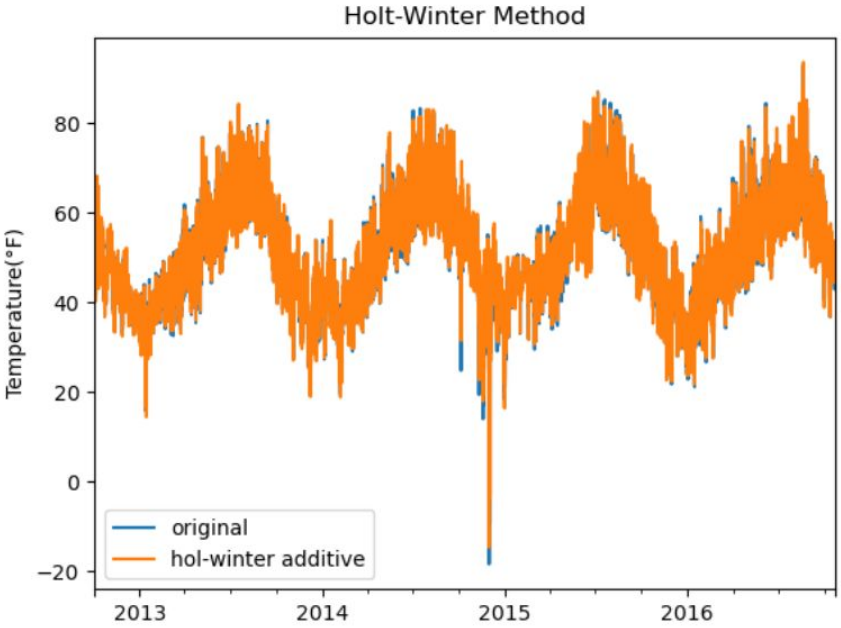
```
ADF Statistic: -26.072101
p-value: 0.000000
Critical Values:
    1%: -3.431
    5%: -2.862
    10%: -2.567
```

Stationary

Seasonal Differencing

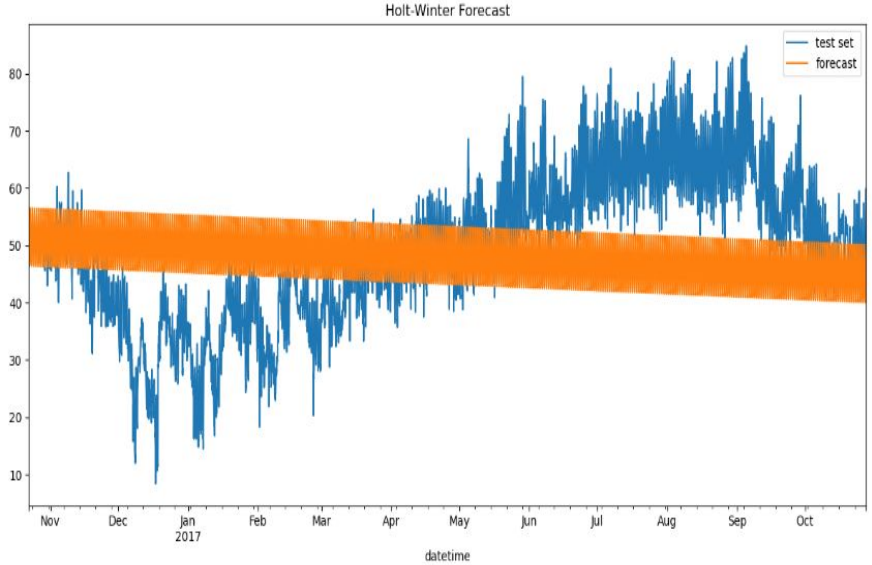


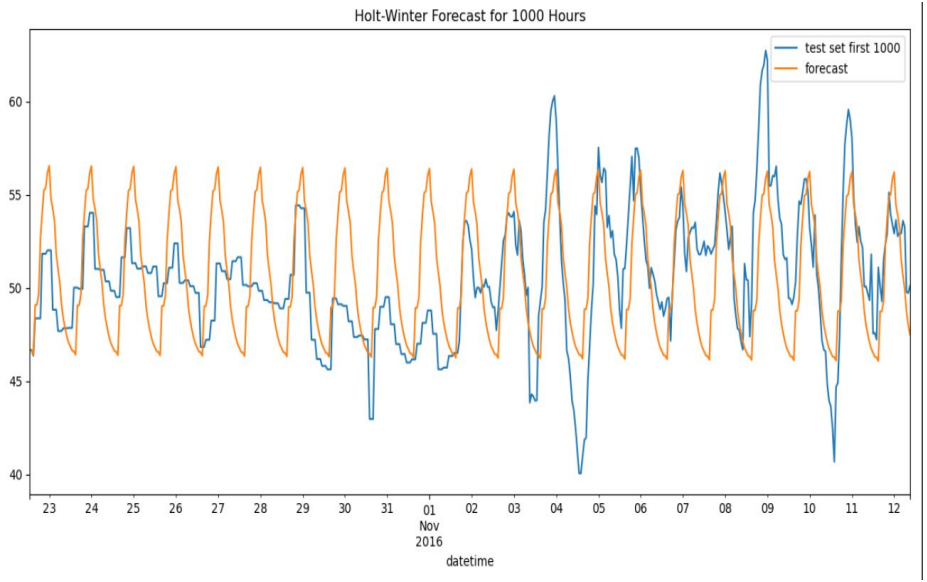
Holt-Winter Method



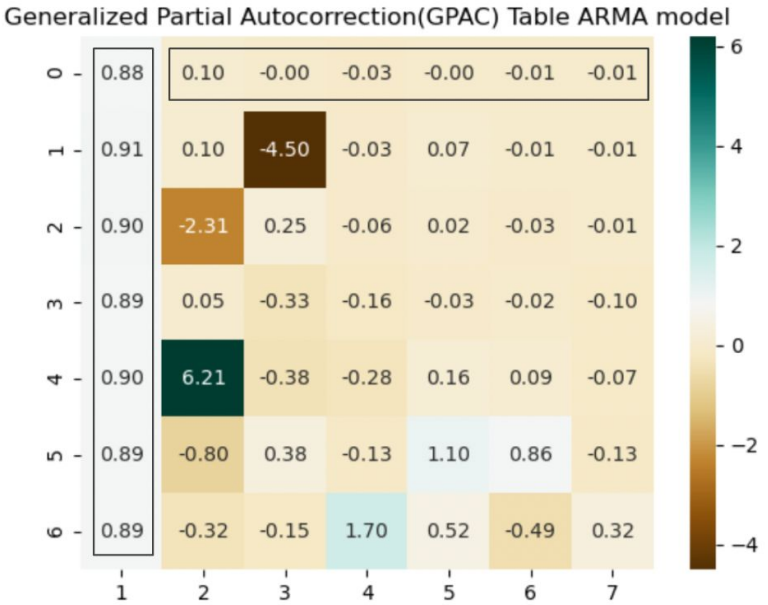
Fitted with additive decomposition

Holt-Winter Method Forecast

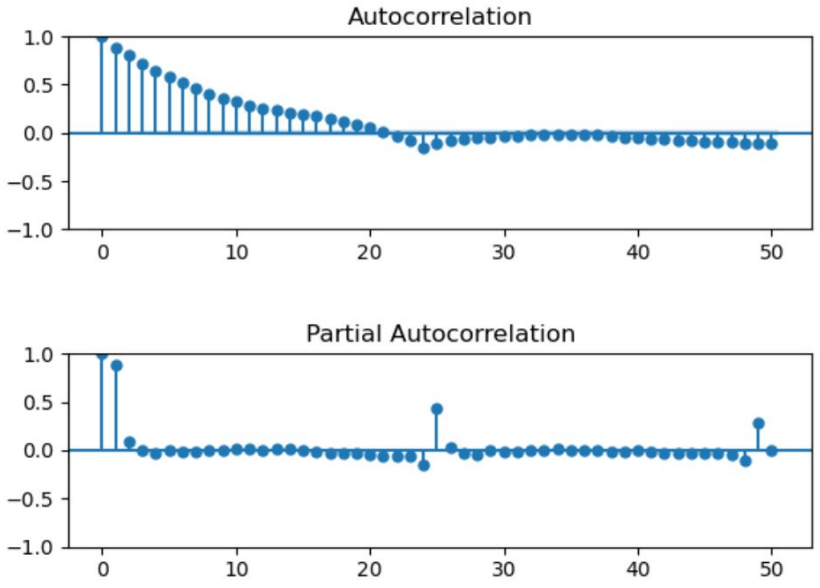




ARMA Model - order determination



na = 1, nb = 0, AR(1)



ARMA Model - parameter estimation

SARIMAX Results

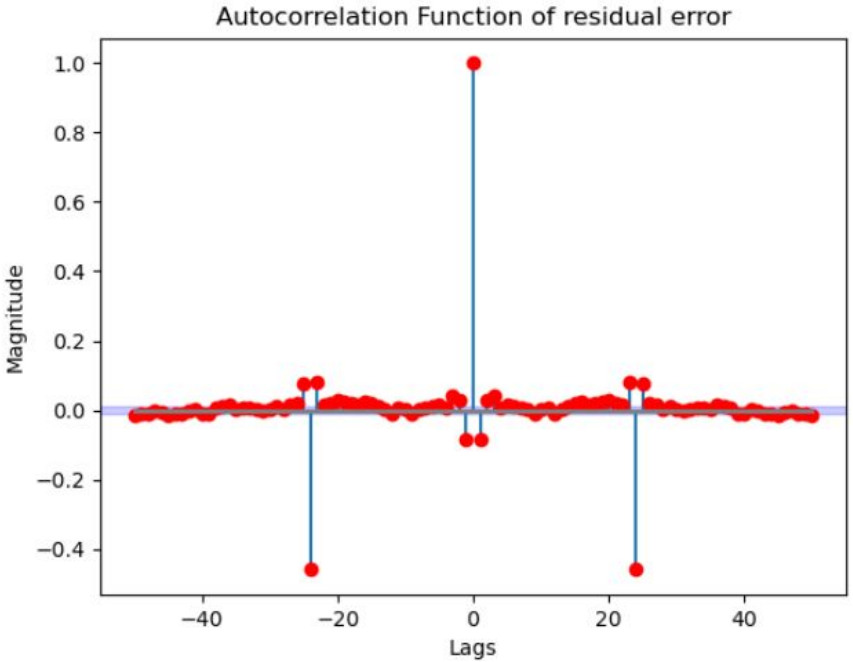
Dep. Variable:	differenced		No. Observations:		35545	
Model:	SARIMAX(1, 0, 0)		Log Likelihood		-72548.739	
Date:	Sat, 17 Dec 2022		AIC		145101.478	
Time:	18:29:27		BIC		145118.435	
Sample:	10-02-2012		HQIC		145106.875	
	- 10-22-2016					
Covariance Type:	opg					

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.8822	0.001	887.462	0.000	0.880	0.884
sigma2	3.4700	0.005	661.057	0.000	3.460	3.480

Ljung-Box (L1) (Q):	251.51	Jarque-Bera (JB):	7647904.54
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	1.01	Skew:	-0.21
Prob(H) (two-sided):	0.45	Kurtosis:	74.86

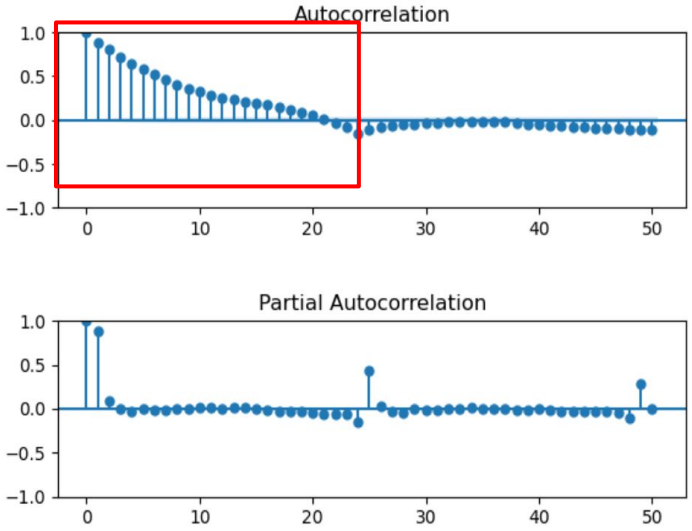
$y(t) - 0.88y(t-1) = e(t)$

ARMA Model - 1 step prediction



Non-seasonal Differencing

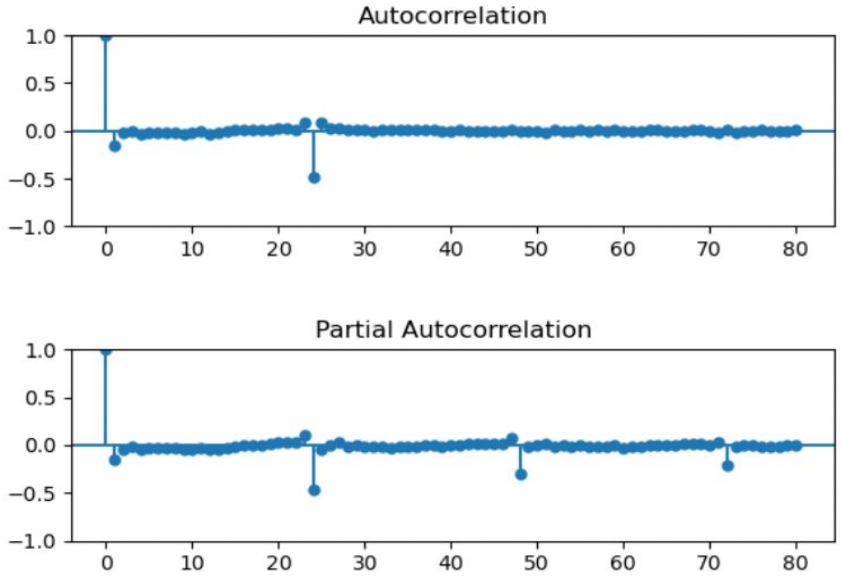
1st order seasonal differenced data(k=24)



May need a further non-seasonal differencing

SARIMA Model - order determination

1st order seasonal & non-seasonal differenced data



Na = 0
Nb = 1

SARIMA(1, 1, 0)xSARIMA(0,1,1)$_{24}$

SARIMA model - parameter estimation

SARIMAX Results

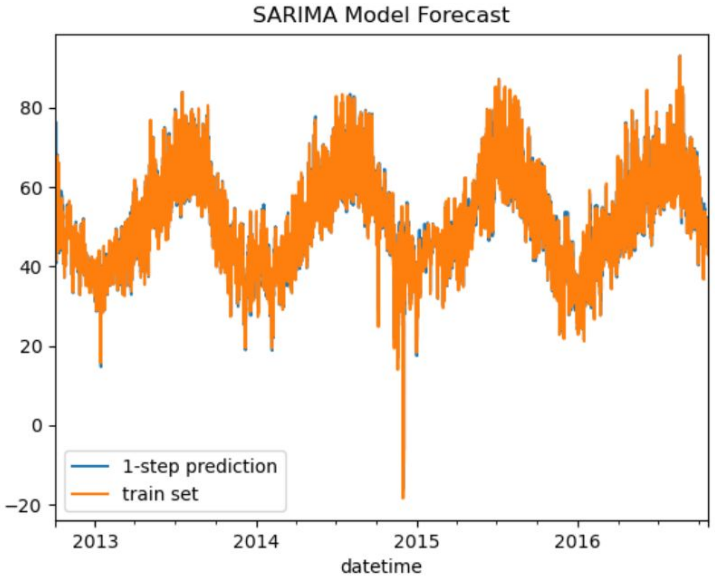
Dep. Variable:	original	No. Observations:	35545
Model:	SARIMAX(1, 1, 0)x(0, 1, [1], 24)	Log Likelihood	-63867.058
Date:	Sat, 17 Dec 2022	AIC	127740.115
Time:	18:49:36	BIC	127765.549
Sample:	10-02-2012	HQIC	127748.211
	- 10-22-2016		
Covariance Type:	opg		

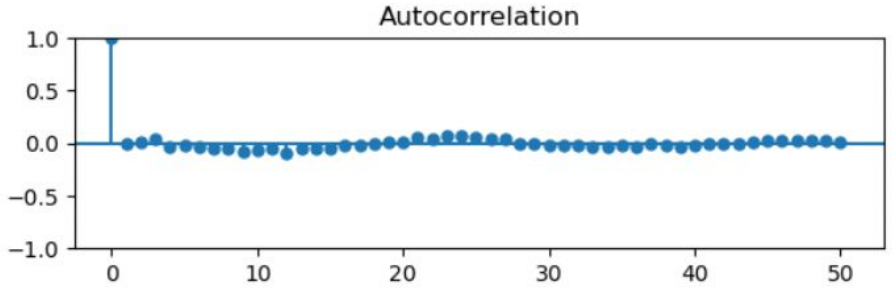
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0855	0.001	-75.378	0.000	-0.088	-0.083
ma.S.L24	-0.9051	0.001	-727.451	0.000	-0.908	-0.903
sigma2	2.1320	0.003	751.346	0.000	2.126	2.138

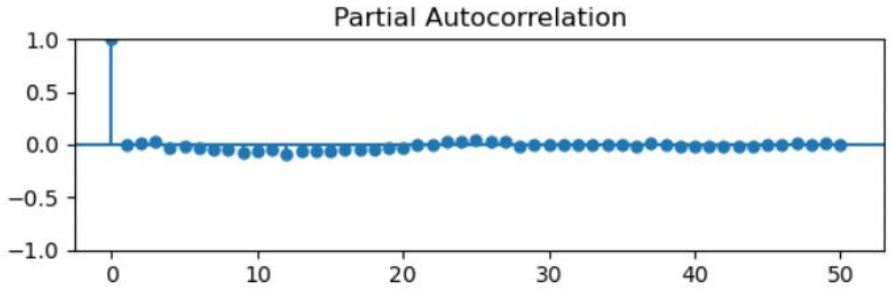
Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	8792801.18
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	1.07	Skew:	0.24
Prob(H) (two-sided):	0.00	Kurtosis:	80.08

$$y(t) + 0.086y(t-1) = e(t) + 0.91e(t-24)$$

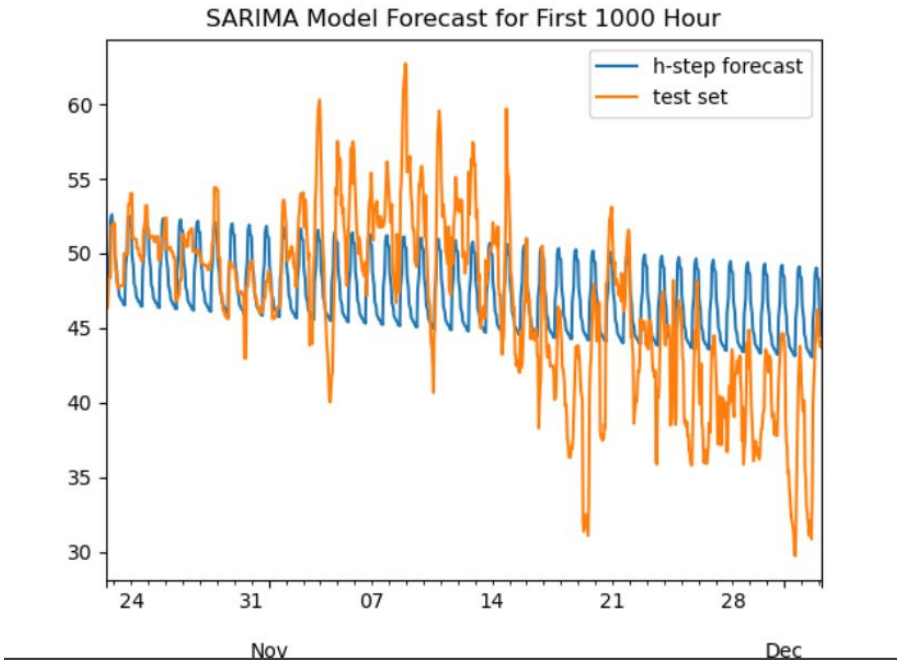
SARIMA Model - 1 step prediction

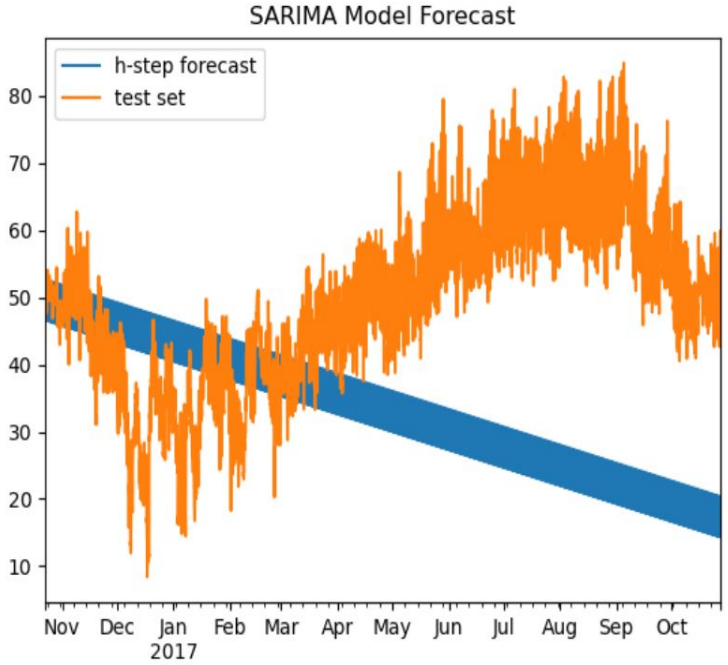




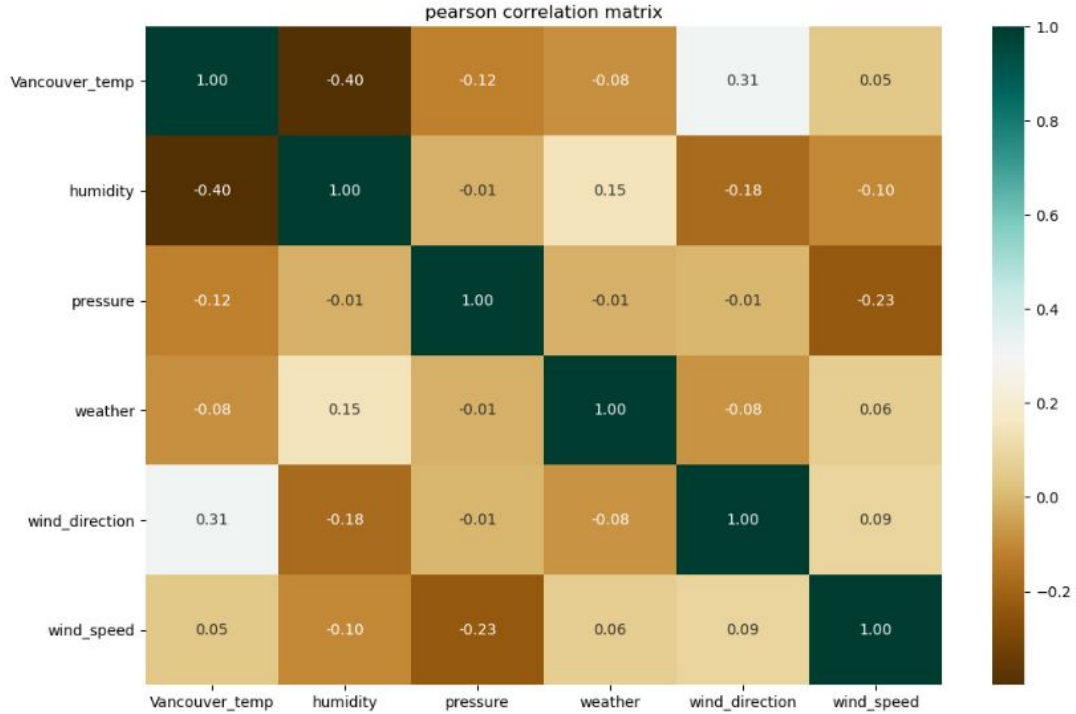


SARIMA Model - h step prediction





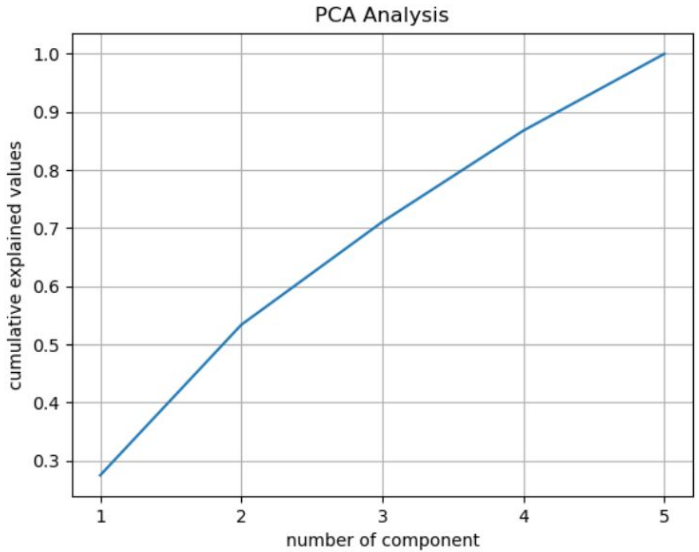
Multiple Linear Regression



Use multiple features

- Humidity
- Pressure
- Weather
- Wind_direction
- Wind_speed

Feature Selection



- Backward Stepwise Selection(p-value<0.05)
- VIF Selection(vif < 3)
- SVD
- PCA

features	humidity	pressure	weather	Wind direction	Wind speed	AIC	BIC	Adjusted R^2
keep or not	1	1	1	1	1	264296.2	264347.1	0.26328

singular values= [48863.87781976, 46010.16482667, 31470.19337718, 27905.09181899, 23475.67215741]

condition number= 1.4427296153944535

Multiple Linear Regression - coefficients

OLS Regression Results

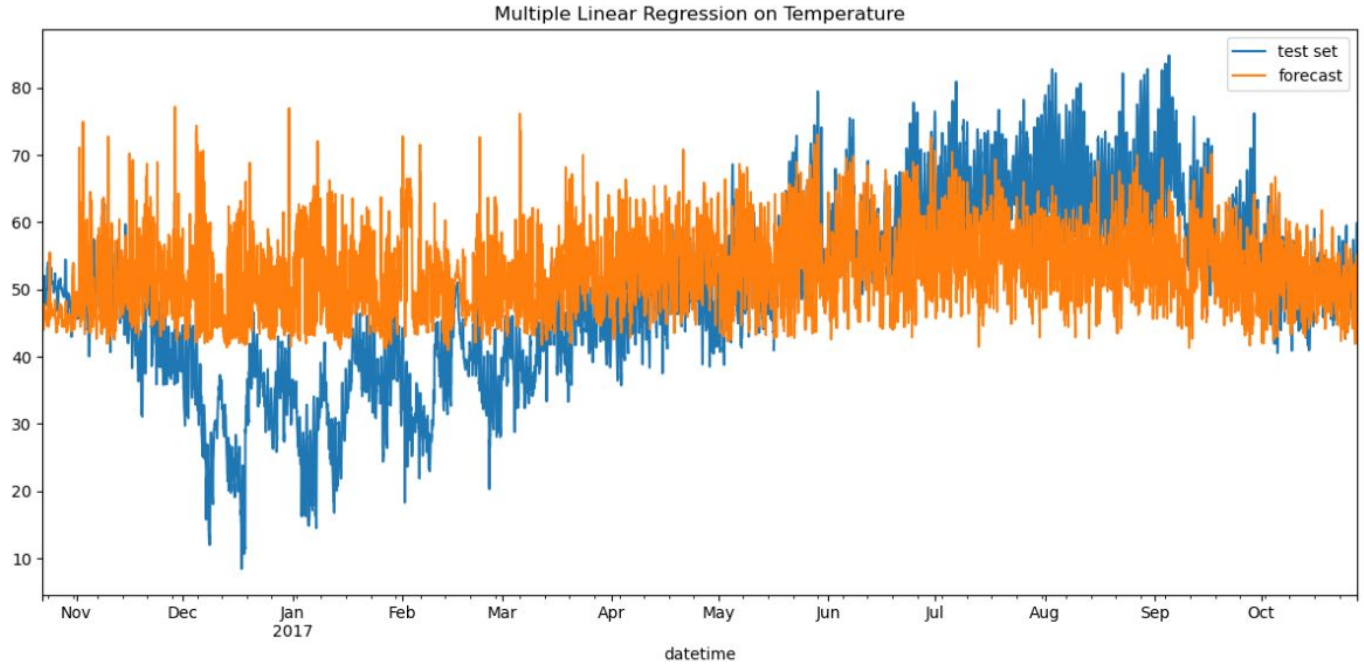
Dep. Variable:	original	R-squared:	0.263
Model:	OLS	Adj. R-squared:	0.263
Method:	Least Squares	F-statistic:	2541.
Date:	Sat, 17 Dec 2022	Prob (F-statistic):	0.00
Time:	18:29:18	Log-Likelihood:	-1.3214e+05
No. Observations:	35545	AIC:	2.643e+05
Df Residuals:	35539	BIC:	2.643e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	51.6602	0.053	977.714	0.000	51.557	51.764
humidity	-4.3742	0.055	-80.235	0.000	-4.481	-4.267
pressure	-1.8783	0.055	-34.345	0.000	-1.985	-1.771
weather	-0.3467	0.054	-6.445	0.000	-0.452	-0.241
wind_direction	2.7896	0.054	51.417	0.000	2.683	2.896
wind_speed	-0.3501	0.055	-6.330	0.000	-0.458	-0.242

Omnibus:	2299.996	Durbin-Watson:	0.137
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3123.428
Skew:	-0.582	Prob(JB):	0.00
Kurtosis:	3.869	Cond. No.	1.39

Temp = 51.66 - 4.37humidity - 1.88pressure - 0.35weather + 2.79windDirection - 0.35windSpeed

Multiple Linear Regression - forecast



Conclusion

- For short term forecasting
  - Holt-winter
  - SARIMA(y(t) + 0.086y(t-1) = e(t) + 0.91e(t-24))
- For long term forecasting
  - Multiple Linear Regression with features

Thank You