Time Series Forecasting

Done By: Goh Rui Zhuo (2222329)

Objectives

Problem Statement

1. Predict the consumption with forecasting
2. Singapore total electricity, water and gas consumption has been on the rise

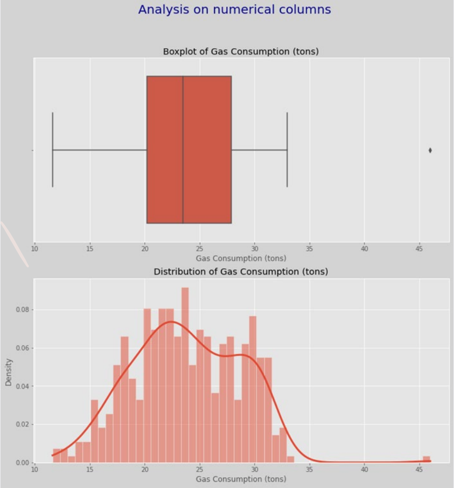
Dataset includes:

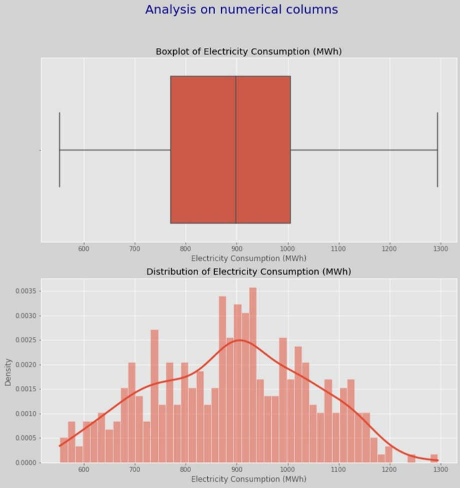
1. 397 Rows x 4 Columns
2. 3 numerical columns and a date column

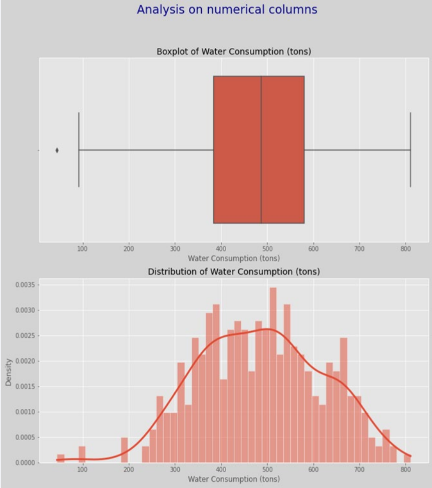
General Info

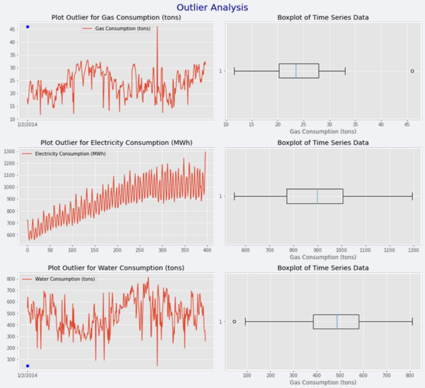
1. No anomaly dtype in the dataset
2. Dataset is clean with no missing values and therefore additional exploratory data analysis can be done
3. No anomaly was observed where in the numerical column

Exploratory data analysis (Part 1)









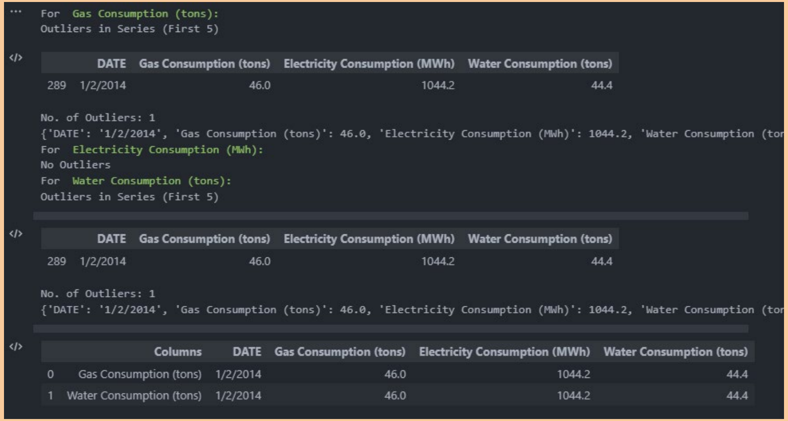
Exploratory data analysis (Part 1)

Numerical Data

1. For Gas Consumption, an analysis of the data reveals an outlier. The distribution of gas consumption is negatively skewed, with a median value of approximately 23 tons over the years. This aligns with the findings from the histogram, which also highlights the outlier value at 45 tons
2. For Electricity Consumption data does not indicate any outlier. The distribution is similarly negatively skewed, yet the median stands around 900 MWh over the years. The histogram supports these observations
3. As for Water Consumption, the data suggests the presence of an outlier. With a distribution that is negatively skewed, the median water consumption is about 500 tons over the years. Correspondingly, the histogram reveals the same trend, including the outlier value, which is less than 100 tons

Outlier Analysis

1. Outlier is when values are above fence or below fence
2. We see that the dataset contains outliers
3. We see that one row of the data is an outlier
4. Forward fill was used to remove the outlier



```
For  Gas Consumption (tons):
Outliers in Series (First 5)
```

	DATE	Gas Consumption (tons)	Electricity Consumption (MWh)	Water Consumption (tons)
289	1/2/2014	46.0	1044.2	44.4

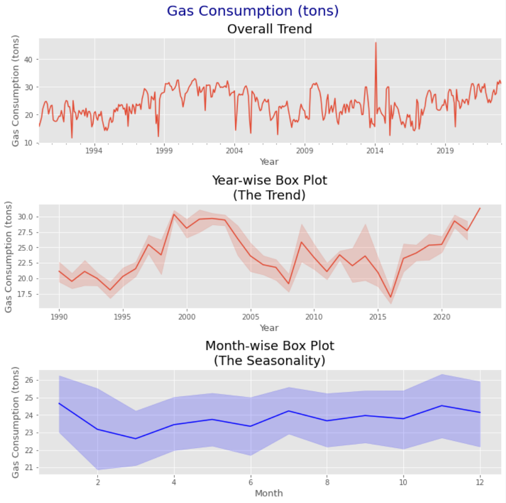
```
No. of Outliers: 1
{'DATE': '1/2/2014', 'Gas Consumption (tons)': 46.0, 'Electricity Consumption (MWh)': 1044.2, 'Water Consumption (tor
For  Electricity Consumption (MWh):
No Outliers
For  Water Consumption (tons):
Outliers in Series (First 5)
```

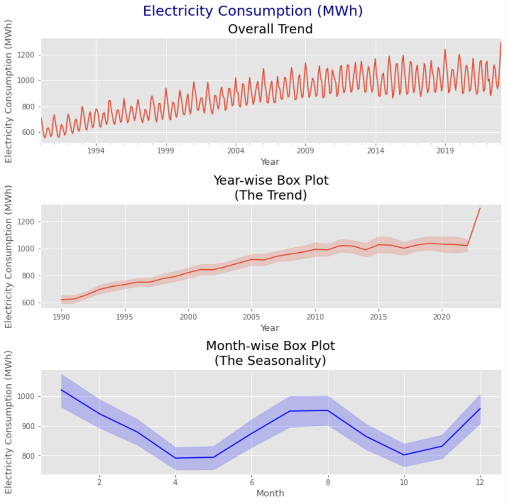
	DATE	Gas Consumption (tons)	Electricity Consumption (MWh)	Water Consumption (tons)
289	1/2/2014	46.0	1044.2	44.4

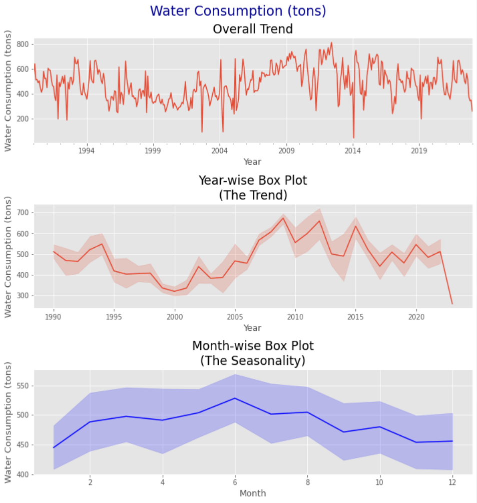
```
No. of Outliers: 1
{'DATE': '1/2/2014', 'Gas Consumption (tons)': 46.0, 'Electricity Consumption (MWh)': 1044.2, 'Water Consumption (tor
```

	Columns	DATE	Gas Consumption (tons)	Electricity Consumption (MWh)	Water Consumption (tons)
0	Gas Consumption (tons)	1/2/2014	46.0	1044.2	44.4
1	Water Consumption (tons)	1/2/2014	46.0	1044.2	44.4

Exploratory data analysis (Part 2)







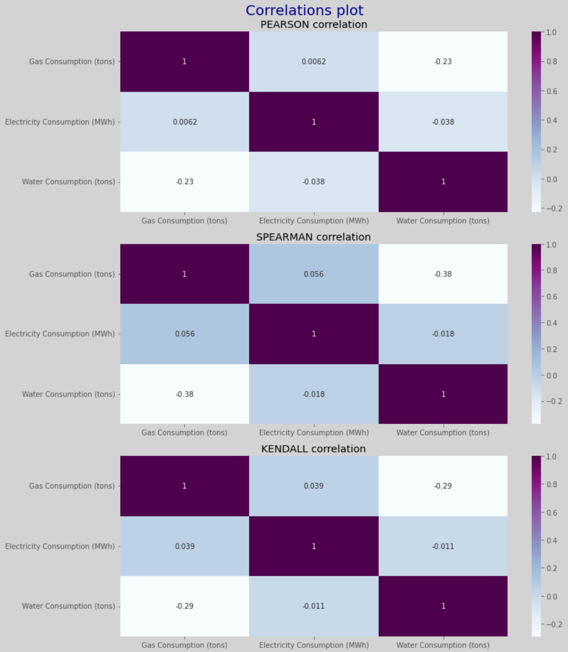
Exploratory data analysis (Part 2)

Analysis on different time frame

1. For Gas Consumption, we observe multiple spikes with a pronounced spike in 2014. On the yearly plot, we note that 1998 was the peak and 2016 the lowest. On the monthly plot, we note that December was the peak and March the lowest.
2. For Electricity Consumption, we observe multiple spikes with an overall increasing trend.On the yearly plot, we note that 2020 was the peak and 1990 the lowest.On the monthly plot, we note that January was the peak and April the nadir.
3. For Water Consumption, we observe multiple spikes with a pronounced spike in 2014.On the yearly plot, we note that 2008 was the peak and 2000 the lowest.On the monthly plot, we note that June was the peak and January the nadir.

Correlation Analysis

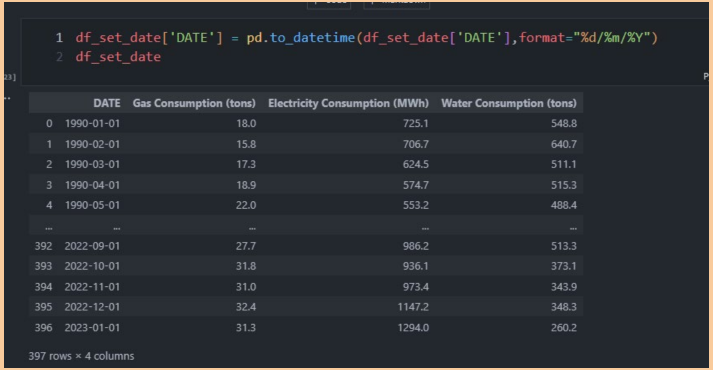
1. Pearson, Spearman and Kendall was used
2. No significant correlation for others for all columns for both normal and interpolated dataframe



Data Preprocessing (1) and Time Series Analysis

Data Preprocessing

1. Change dtype of date in both interpolated and original dataframe



```
df_set_date['DATE'] = pd.to_datetime(df_set_date['DATE'],format="%d/%m/%Y")
df_set_date
```

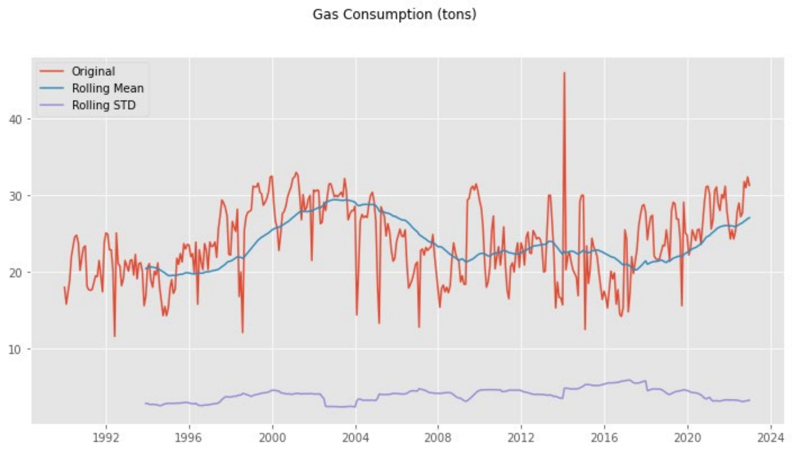
	DATE	Gas Consumption (tons)	Electricity Consumption (MWh)	Water Consumption (tons)
0	1990-01-01	18.0	725.1	548.8
1	1990-02-01	15.8	706.7	640.7
2	1990-03-01	17.3	624.5	511.1
3	1990-04-01	18.9	574.7	515.3
4	1990-05-01	22.0	553.2	488.4
...	...	...	...	...
392	2022-09-01	27.7	986.2	513.3
393	2022-10-01	31.8	936.1	373.1
394	2022-11-01	31.0	973.4	343.9
395	2022-12-01	32.4	1147.2	348.3
396	2023-01-01	31.3	1294.0	260.2

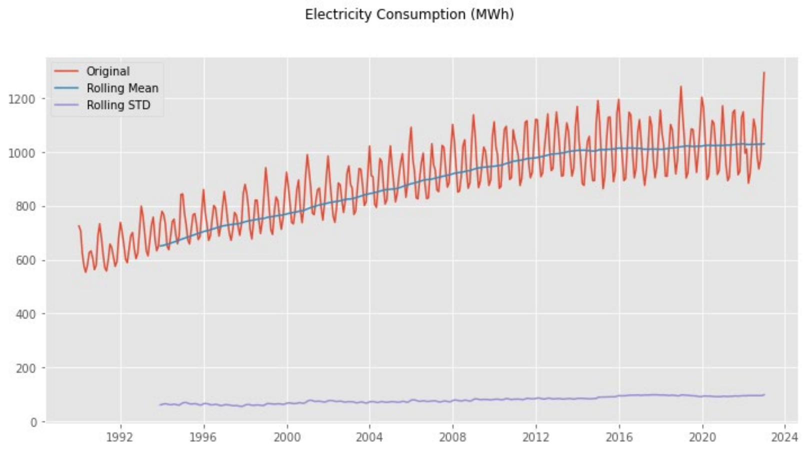
397 rows × 4 columns

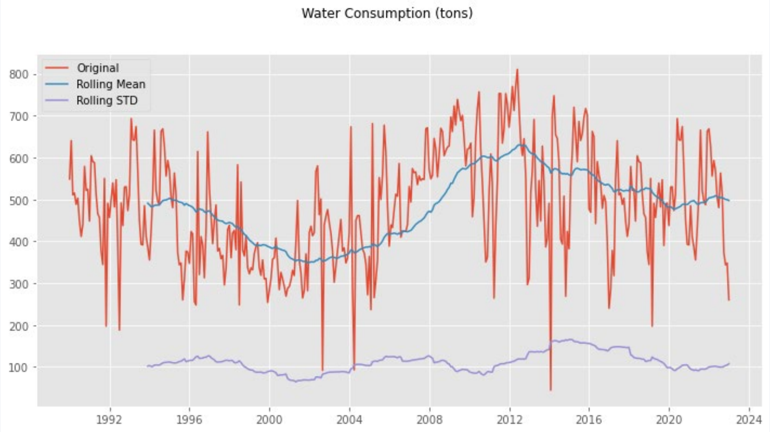
Time Series

1. Seasonal Decomposition
    a. For Gas Consumption, we observe no clear linear trend in the data.
        i. From the seasonality plot, we note that the data exhibits weak seasonality.
        ii. The mean and variance are not consistent throughout the years, with lower values observed during 2000-2004.
    2. For Electricity Consumption, we observe an increasing trend in the data.
        3. From the seasonality plot, we note that the data exhibits strong seasonality.
        4. The mean and variance are not consistent throughout the years.
    5. For Water Consumption, we observe no strong trend in the data.
        6. From the seasonality plot, we note that the data exhibits strong seasonality.
        7. The mean is not consistent, but the variance is consistent throughout the years

Exploratory data analysis (Part 2)







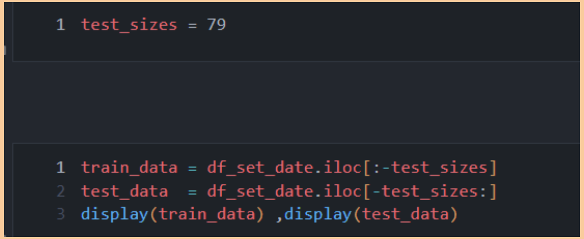
Model Baseline

Time Series (Stationality Test)

1. Visual
    a. For Gas Consumption, mean and variance are consistent.
    b. For Electricity Consumption,t the mean and variance are not consistent.
    c. For Water Consumption, mean and variance are consistent.
2. ADF Test
    a. For Gas Consumption, the data is stationary
    b. For Electricity Consumption, the data is not stationary
    c. For Water Consumption, the data is stationary
3. KPSS test
    a. For Gas Consumption, the data is stationary
    b. For Electricity Consumption, the data is not stationary
    c. For Water Consumption, the data is not stationary

Data Preprocessing

1. Train test split at 80 20



```
test_sizes = 79
```

```
train_data = df_set_date.iloc[:-test_sizes]
test_data  = df_set_date.iloc[-test_sizes:]
display(train_data) ,display(test_data)
```

Evaluation Metrics

1. Lijung Box Test
2. Jarque Bera test
3. HEteroskedasticity test
4. AIC
5. BIC
6. RMSE
7. MAPE

Baseline Model

Model Baseline

1. Average Forecast, Drift Model
   a. For Gas Consumption, all evaluation metrics did not perform well, with a difference of 0 comparing with interpolated data
   b. For Electricity Consumption, all evaluation metrics did not perform well, with a difference of 0, comparing with interpolated data
   c. For Water Consumption, all evaluation metrics did not perform well, with a difference of 0,comparing with interpolated data
2. ARIMA Baseline (Naive)
   a. For Gas Consumption
      i. The data is independently distributed, indicating a white noise time series.
      ii. The data does not have common variance.
      iii. The data is not normally distributed.
      iv. From the residual plot, we see that the majority are close to 0, but with a notable spike in 2014.
      v. The coefficients suggest that the data are independently distributed, residuals show no variance, and the data is not normally distributed.

Baseline Model

Model Baseline

1. ARIMA Baseline (Naive)
    a. For Electricity Consumption:
        i. The data is independently distributed, indicating a white noise time series.
        ii. The data does not have common variance.
        iii. The data is not normally distributed.
        iv. From the residual plot, we see that the majority are close to 0, but with a notable spike in 2014.
        v. The coefficients suggest that the data are independently distributed, residuals show no variance, and the data is not normally distributed.
    b. For Water Consumption:
        i. The data is independently distributed, indicating a white noise time series.
        ii. The data does not have common variance.
        iii. The data is not normally distributed.
        iv. From the residual plot, we see that the majority are close to 0, but with a notable spike in 2014.
        v. The coefficients suggest that the data are independently distributed, residuals show no variance, and the data is not normally distributed

Baseline Model

Model Baseline

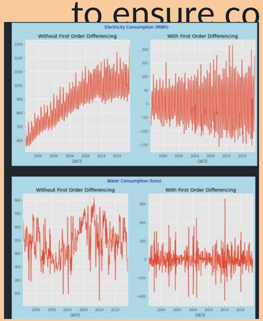
1. Exponential Smoothing
   1. For Gas Consumption:
      1. From the results, we see that the AIC is 836.652 and the BIC is 844.169.
      2. The selected smoothing level is 0.4499155.
      3. Evaluation metrics indicate that all scores are poor.
      4. The forecast does not fit the data well.
   2. For Electricity Consumption:
      1. From the results, we see that the AIC is 2719.964 and the BIC is 2727.481.
      2. The selected smoothing level is 0.995.
      3. Evaluation metrics indicate that all scores are poor.
      4. The forecast does not fit the data well.
   3. For Water Consumption:
      1. From the results, we see that the AIC is 2967.425 and the BIC is 2974.943.
      2. The selected smoothing level is 0.3715462.
      3. Evaluation metrics indicate that all scores are poor.
      4. The forecast does not fit the data wel

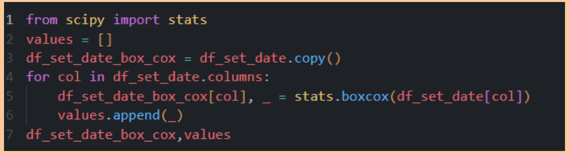
Overall, the baseline model did not perform well and there weren't much difference between interpolate and original data , hence original data will be used the restl.

Advanced Model

Stationarity

1. Differencing
   a. After one time of differencing, data became stationary
2. Box Cox transformation was used to ensure constant mean and

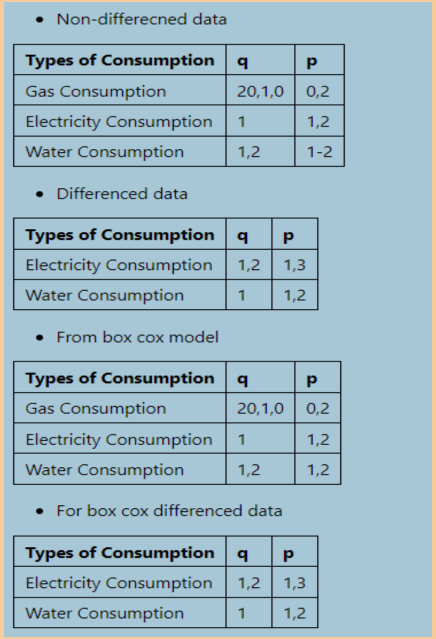




```python
from scipy import stats
values = []
df_set_date_box_cox = df_set_date.copy()
for col in df_set_date.columns:
    df_set_date_box_cox[col], _ = stats.boxcox(df_set_date[col])
    values.append(_)
df_set_date_box_cox,values
```

PACF and ACF

1. To determine the order for arima and sarima model, pacf and acf model is needed



- Non-differecned data

Types of Consumption	q	p
Gas Consumption	20,1,0	0,2
Electricity Consumption	1	1,2
Water Consumption	1,2	1-2

- Differenced data

Types of Consumption	q	p
Electricity Consumption	1,2	1,3
Water Consumption	1	1,2

- From box cox model

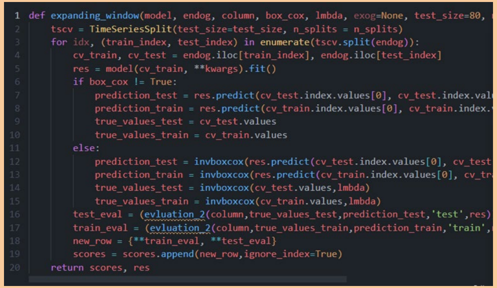
Types of Consumption	q	p
Gas Consumption	20,1,0	0,2
Electricity Consumption	1	1,2
Water Consumption	1,2	1,2

- For box cox differenced data

Types of Consumption	q	p
Electricity Consumption	1,2	1,3
Water Consumption	1	1,2

New Cross Validation Method

1. Expanding window function
2. Provide cross validation among the dataset
3. Evaluation added aic and bic



```python
def expanding_window(model, endog, column, box_cox, lmbda, exog=None, test_size=80, 
    tscv = TimeSeriesSplit(test_size=test_size, n_splits = n_splits)
    for idx, (train_index, test_index) in enumerate(tscv.split(endog)):
        cv_train, cv_test = endog.iloc[train_index], endog.iloc[test_index]
        res = model(cv_train, **kwargs).fit()
        if box_cox != True:
            prediction_test = res.predict(cv_test.index.values[0], cv_test.index.val
            prediction_train = res.predict(cv_train.index.values[0], cv_train.index.
            true_values_test = cv_test.values
            true_values_train = cv_train.values
        else:
            prediction_test = invboxcox(res.predict(cv_test.index.values[0], cv_test
            prediction_train = invboxcox(res.predict(cv_train.index.values[0], cv_tr
            true_values_test = invboxcox(cv_test.values,lmbda)
            true_values_train = invboxcox(cv_train.values,lmbda)
        test_eval = (evluation_2(column,true_values_test,prediction_test,'test',res)
        train_eval = (evluation_2(column,true_values_train,prediction_train,'train',
        new_row = {**train_eval, **test_eval}
        scores = scores.append(new_row,ignore_index=True)
    return scores, res
```

Advanced Model

ARIMA and Sarima (Gas Consumption)

1. ARIMA
   a. From the result of original data frame and box cox transformed data, we can see both performed relatively better for the validation portion on test set with order (1,0,0) and (20,0,0)
   b. box cox transformed data performed better as compared to original dataframe
   c. In addition, we can see that the test AIC is about 500 - 800
   d. We can see that it is independently distributed, hence it is a white noise time series
   e. Residual does not have common variance
   f. Data is not normal
   g. Forecast is not nicely fit
2. SARIMA
   a. From the result of original data frame and box cox transformed data, we can see both performed relatively better for the validation portion on test set with order (20,0,0),(0,0,1,12) and (1,0,2),(0,0,2,12)
   b. Box cox transformed data performed better as compared to original dataframe
   c. In addition, we can see that the test AIC is about 500 - 800
   d. Independently distributed, hence it is a white noise time series
   e. Residual does not have common variance
   f. Data is not normal
   g. Forecast is not nicely fit

Advanced Model

ARIMA and Sarima (Electricity Consumption)

1. ARIMA
    a. From the result of original data frame and box cox transformed data, we can see both performed relatively better for the validation portion on test set with order (3,1,3)
    b. When comparing the difference, box cox transformed data performed better as compared to original dataframe in terms of AIC, BIC but worst in order validation
    c. In addition, we can see that the test AIC is about 1900-2000
    d. We can see that it is independently distributed, hence it is a white noise time series
    e. Residual does not have common variance
    f. Data is not normal
    g. Forecast is pretty nicely fit
2. SARIMA
    a. From the result of original data frame and box cox transformed data, we can see both performed relatively better for the validation portion on test set with order (0,1,2),(0,0,2,12) and (1,1,2),(0,0,2,12)
    b. When comparing the difference, box cox transformed data performed better as compared to original dataframe
    c. In addition, we can see that the test AIC is about 1600-1700
    d. Independently distributed, hence it is a white noise time series
    e. Residual does not have common variance
    f. Data is not normal
    g. Forecast is not nicely fit

Advanced Model

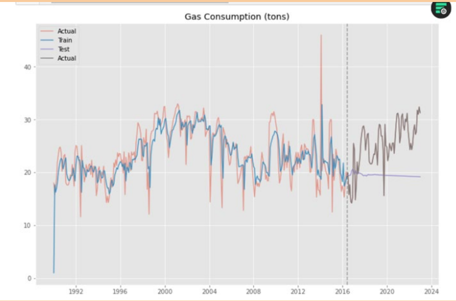
ARIMA and Sarima (Water Consumption)

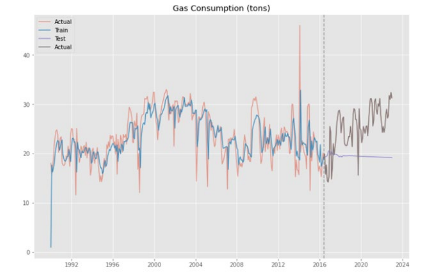
1. ARIMA
    - a. From the result of original data frame and box cox transformed data, we can see both performed relatively better for the validation portion on test set with order (1,1,1) ,(1,0,2)
    - b. When comparing the difference, box cox transformed data performed worst as compared to original dataframe in terms of AIC, BIC but better in other validation
    - c. In addition, we can see that the test AIC is about 1900 - 2000
    - d. We can see that it is independently distributed, hence it is a white noise time series
    - e. Residual does not have common variance
    - f. Data is not normal
    - g. Forecast is not nicely fit
2. SARIMA
    - a. From the result of original data frame and box cox transformed data, we can see both performed relatively better for the validation portion on test set with order (20,0,0),(0,0,1,12) and (1,0,2),(0,0,2,12)
    - b. Box cox transformed data performed better as compared to original dataframe
    - c. In addition, we can see that the test AIC is about 500 - 800
    - d. We can see that it is independently distributed, hence it is a white noise time series
    - e. Residual does have common variance
    - f. Data is not normal
    - g. Forecast is not nicely fit

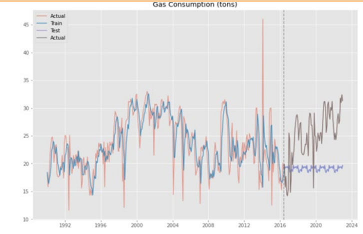
Hyperparameter Tuning

ARIMA and Sarima (Gas Consumption)

1. ARIMA → Tune the order of p,d,q
   a. Best order is 7,0,4
   b. RMSE is 5.79, MAPE is 0.20, AIC is 585 and BIC is 620
   c. Independently distributed, hence it is a white noise time series
   d. Residual does not have common variance
   e. Data is not normal
   f. Forecast is not nicely fit
2. SARIMA
   a. Best order are (3,0,1) (0,0,0,12)
   b. RMSE is 6.19, MAPE is 0.22, AIC is 587 and BIC is 612
   c. Independently distributed, hence it is a white noise time series
   d. Residual does not have common variance
   e. Data is not normal
   f. Forecast is not nicely fit
3. Exponential Smoothing
   a. Best combination are None,None, and 12
   b. RMSE is 8..53, MAPE is 0.33, AIC is 144 and BIC is 186



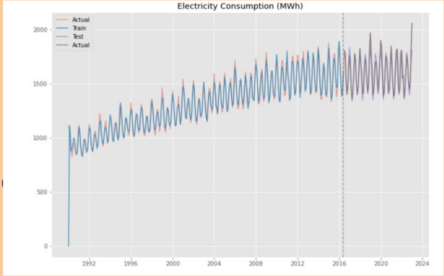


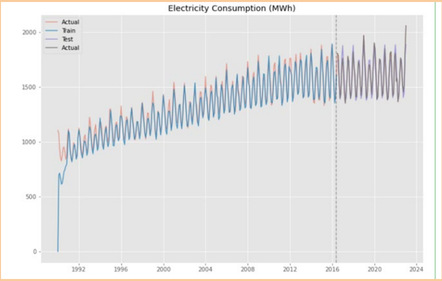


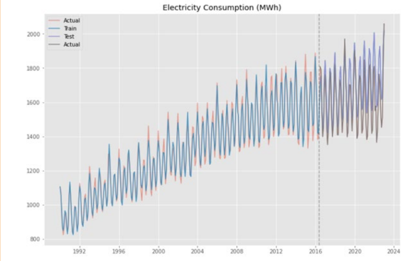
Hyperparameter Tuning

ARIMA and Sarima (Electricity Consumption)

1. ARIMA → Tune the order of p,d,q
   a. Best order is 5,1,7
   b. RMSE 44 MAPE 0.03, AIC 3487, BIC: 3456
   c. We can see that it is independently distributed, hence it is a white nois
   d. Residual does not have common variance
   e. Data is not normal
   f. Forecast is nicely fit
2. SARIMA
   a. Best order are (1,0,4) (1,0,1,12)
   b. RMSE is 36.575, MAPE is 0.032, AIC is 1578 and BIC is 1600
   c. Independently distributed, hence it is a white noise time series
   d. Residual does not have common variance
   e. Data is not normal
   f. Forecast is nicely fit
3. Exponential Smoothing
   a. Best combination are add,,add, and 24
   b. Forcast does much the trend
   c. RMSE is 529, MAPE is 0.06 AIC is 1133 and BIC is 1216



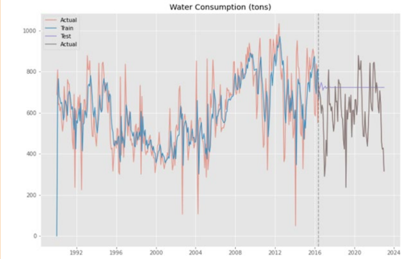


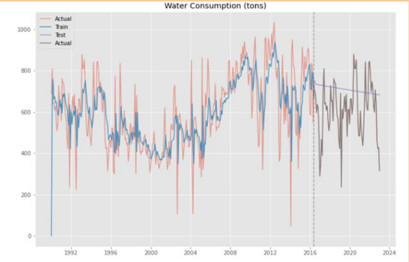


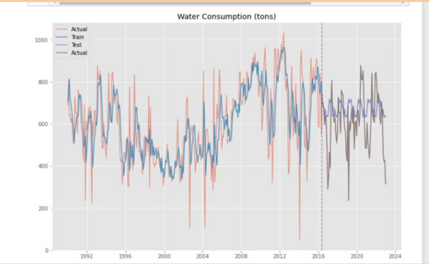
Hyperparameter Tuning

ARIMA and Sarima (Water Consumption)

1. ARIMA → Tune the order of p,d,q
   a. Best order is 9,1,0
   b. RMSE is 144, MAPE is 0.34, AIC is 1970 and BIC is 2000
   c. Independently distributed, hence it is a white noise time series
   d. Residual does not have common variance
   e. Data is not normal
   f. Forecast is not nicely fi
2. SARIMA
   a. Best order are (3,0,1) (0,0,0,12)
   b. RMSE is 121, MAPE is 0.30 , AIC is 1980 and BIC is 1992
   c. Independently distributed, hence it is a white noise time series
   d. Residual does not have common variance
   e. Data is not normal
   f. Forecast is not nicely fit
3. Exponential Smoothing
   a. Best combination are None,add, and 12
   b. RMSE is 146, MAPE is 0.33, AIC is 1542 and BIC is 1584

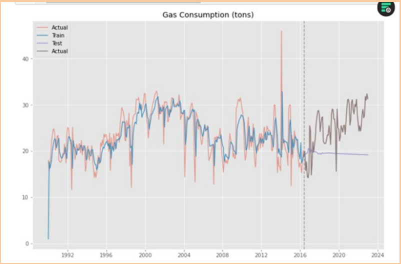


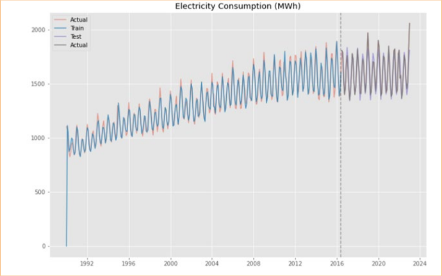


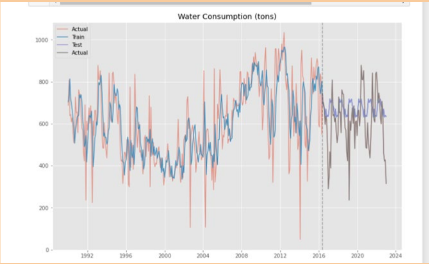


Final Univariate Model

1. Gas Consumption
   a. ARIMA(7,0,4)
   b. RMSE is 5.79, MAPE is 0.20, AIC is 585 and BIC is 620
   c. Independently distributed, hence it is a white noise time series
   d. Residual does not have common variance
   e. Data is not normal
   f. Forecast is not nicely fit
2. Electricity Consumption
   a. ARIMA(5,1,7)
   b. RMSE 44 MAPE 0.03, AIC 3487, BIC: 3456
   c. Independently distributed, hence it is a white noise time series
   d. Residual does not have common variance
   e. Data is not normal
   f. Nicely fit forecast
3. Water Consumption
   a. Best combination are None,add, and 12
   b. RMSE is 146, MAPE is 0.33, AIC is 1542 and BIC is 1584

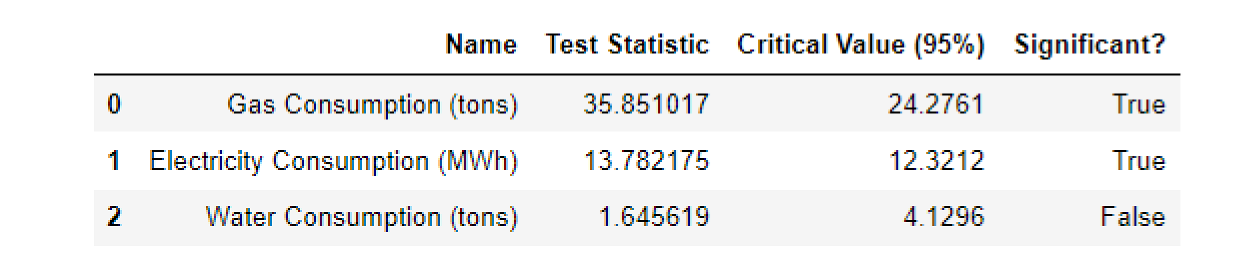




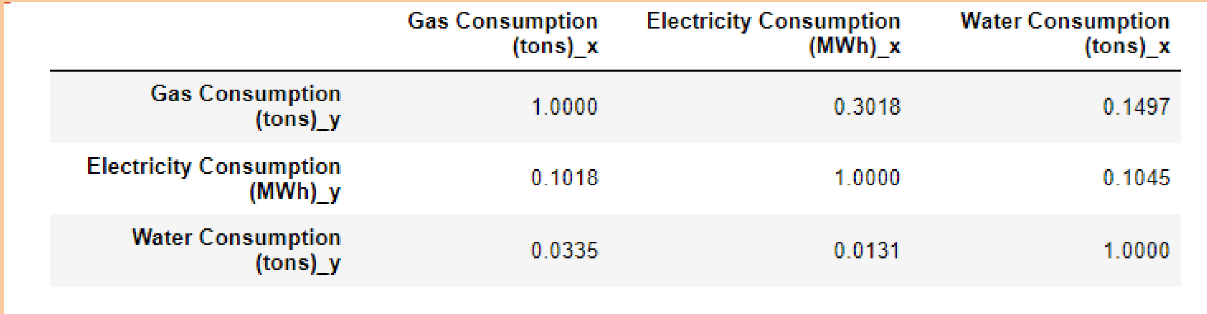


Multivariate Analysis

1. With conintegration test, we can see that time series affect one another
2. With Granger Causality test, we see that whether one time series is useful in forecasting another
3. From both test, all three can be tested for multivariate analysis



	Name	Test Statistic	Critical Value (95%)	Significant?
0	Gas Consumption (tons)	35.851017	24.2761	True
1	Electricity Consumption (MWh)	13.782175	12.3212	True
2	Water Consumption (tons)	1.645619	4.1296	False

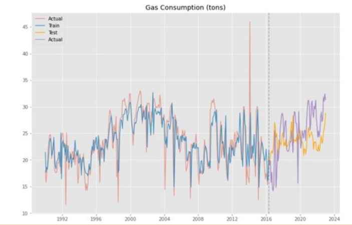


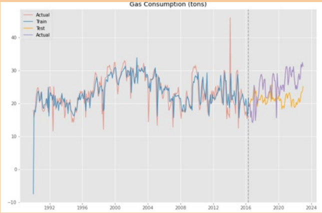
	Gas Consumption (tons)_x	Electricity Consumption (MWh)_x	Water Consumption (tons)_x
Gas Consumption (tons)_y	1.0000	0.3018	0.1497
Electricity Consumption (MWh)_y	0.1018	1.0000	0.1045
Water Consumption (tons)_y	0.0335	0.0131	1.0000

Multivariate Analysis (Hyperparameter tuning)

Gas Consumption

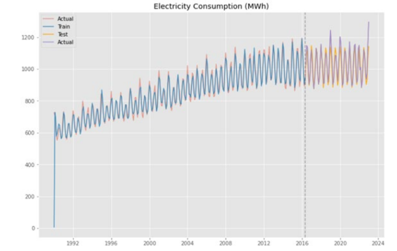
1. Best order for arima is 7,0,2
2. Best order for sarima is (3,1,1) (0,0,0,12)

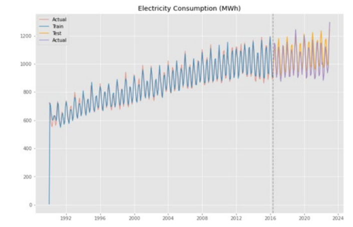




Electricity Consumption

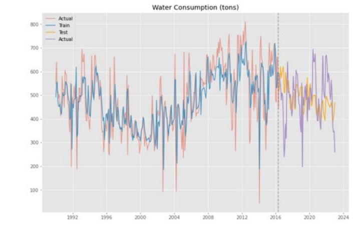
1. Best order for arima is 6,1,9
2. Best order for sarima is (1,1,3) (1,0,1,12)

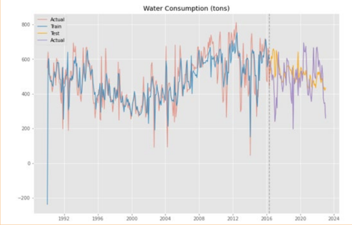




Water Consumption

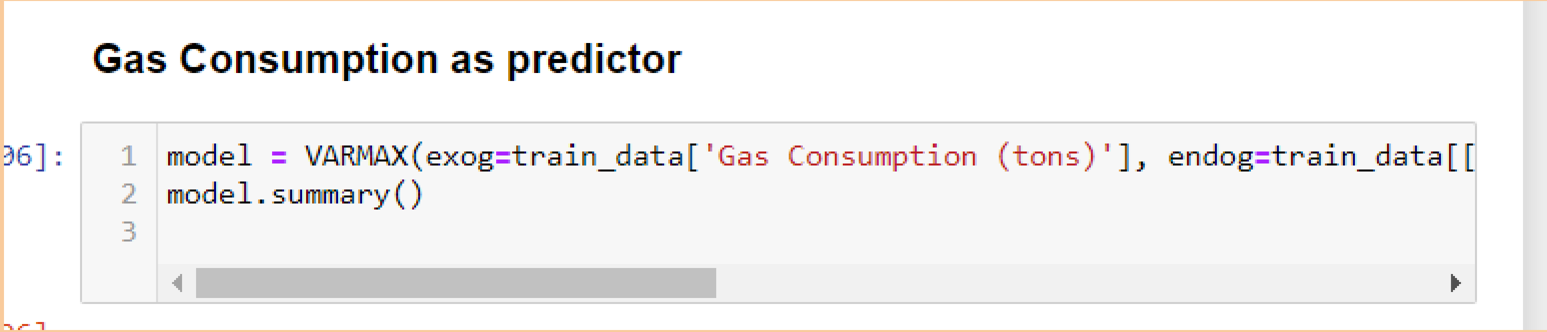
1. Best order for arima is 2,0,3
2. Best order for sarima is (1,1,3) (0,0,1,12)





VARMAX

1. Varmax was also used with each as the predictor
2. However the result was not great
3. Hence, it was not utilise any further



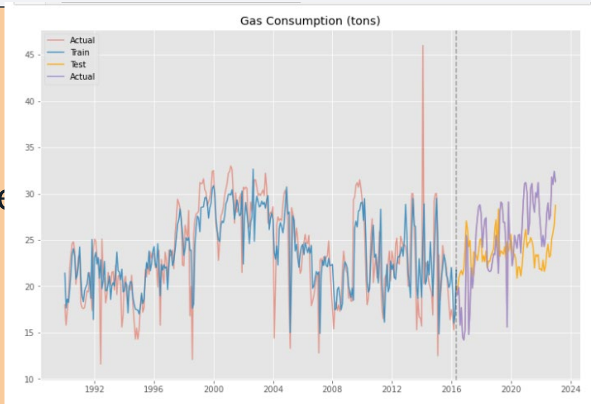
Gas Consumption as predictor

```
06]:  model = VARMAX(exog=train_data['Gas Consumption (tons)'], endog=train_data[[
      model.summary()
```

FInal Multivariate Model

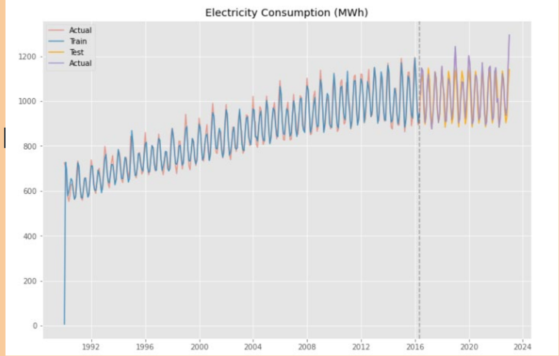
Gas Consumption

1. ARIMA (7,0,2)
2. RMSE 4.37 MAPE 0.15 AIC 1667 BIC 1716
3. Independently distributed, hence it is a white noise time se
4. Residual does not have common variance
5. Data is not normal
6. Forecast is nicely fit



Electricity Consumption

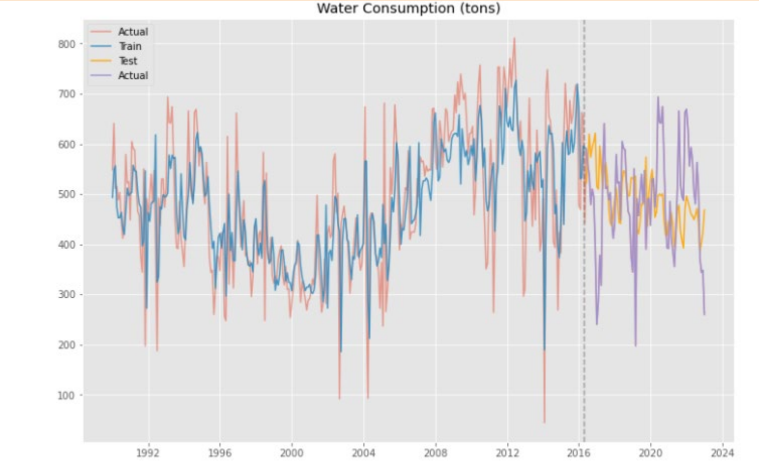
1. ARIMA (7,0,2)
2. RMSE 42.3 MAPE 0.03 AIC 3051 BIC 3118
3. Independently distributed, hence it is a white noise time ser
4. Residual does not have common variance
5. Data is not normal
6. Forecast is nicely fit



FInal Multivariate Model

Water Consumption

1. ARIMA (7,0,2)
2. RMSE 111 MAPE 0.2 AIC 3802 BIC 3836
3. Independently distributed, hence it is a white noise time series
4. Residual does have common variance
5. Data is not normal
6. Forecast is nicely fit



Conclusion, Overall the multivariate model is a Better fit than the univariate one