Machine Learning – Forecasting Energy (Electricity) Consumption

Spencer Rubin

March 18, 2021

Project Summary

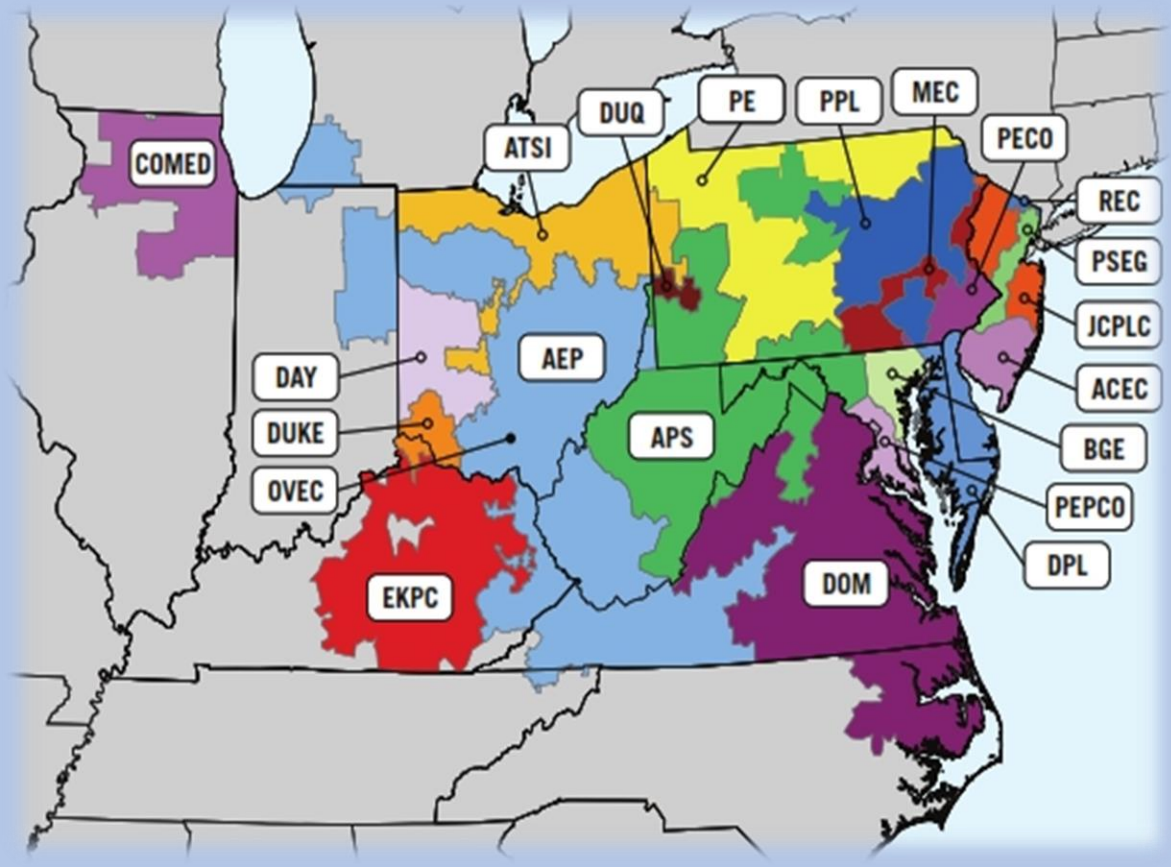
- Energy consumption fluctuates by season, period of the day, and climate
- Extreme weather events may impact peaks in energy demand
- Projecting demand for electricity can reduce power outages, improve customer satisfaction
- *Can we utilize time series data, machine learning on electricity consumption to improve forecasting?*

Data

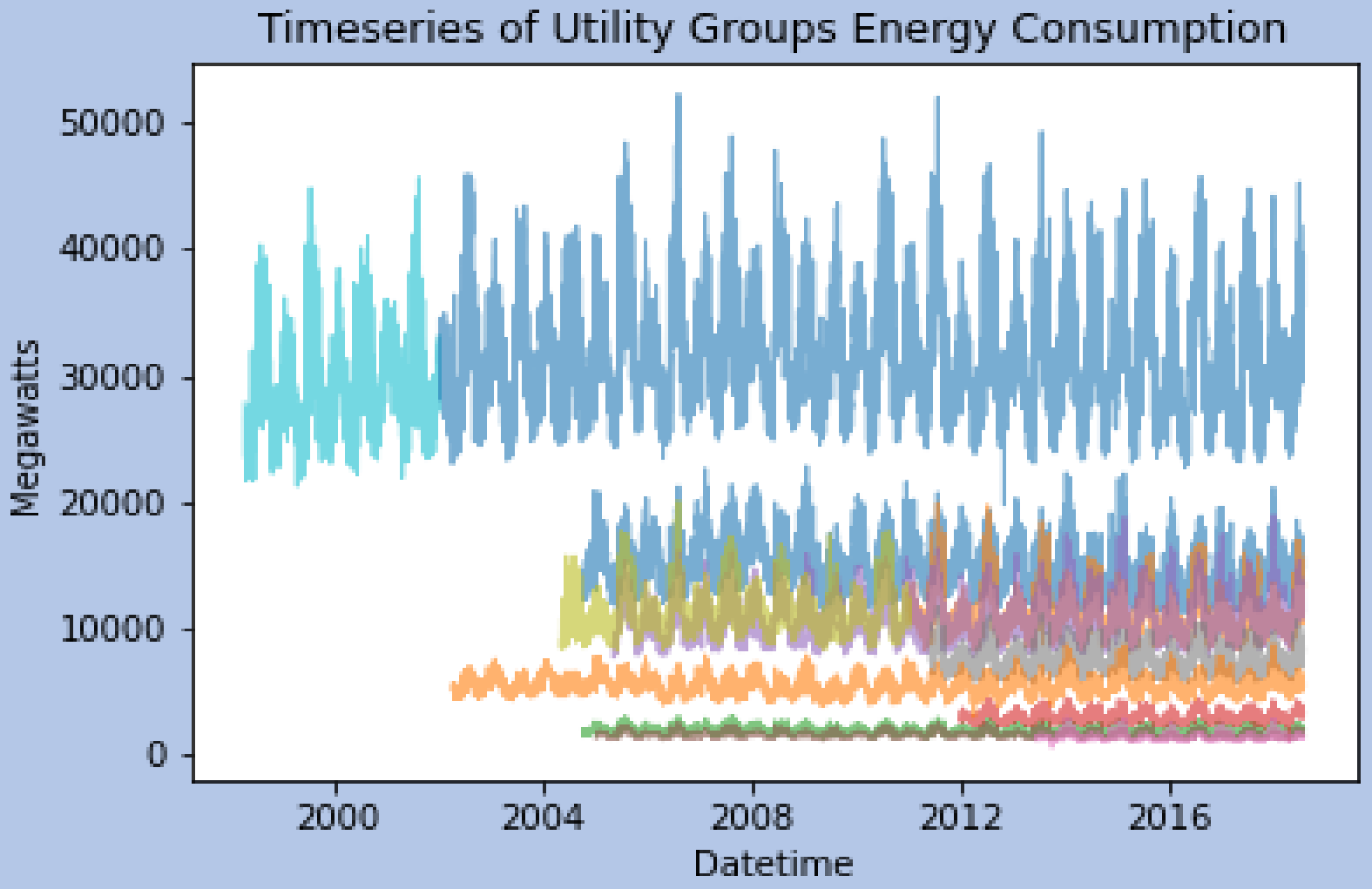
- PJM Interconnection, Inc. – transmission line serving multi-state, multi-utility region in eastern U.S.

- Hourly data of electricity consumption by utility company (megawatts)

- Data freely available via Kaggle: https://www.kaggle.com/robikscube/hourly-energy-consumption
  - Multiple .csv files per utility company under PJM

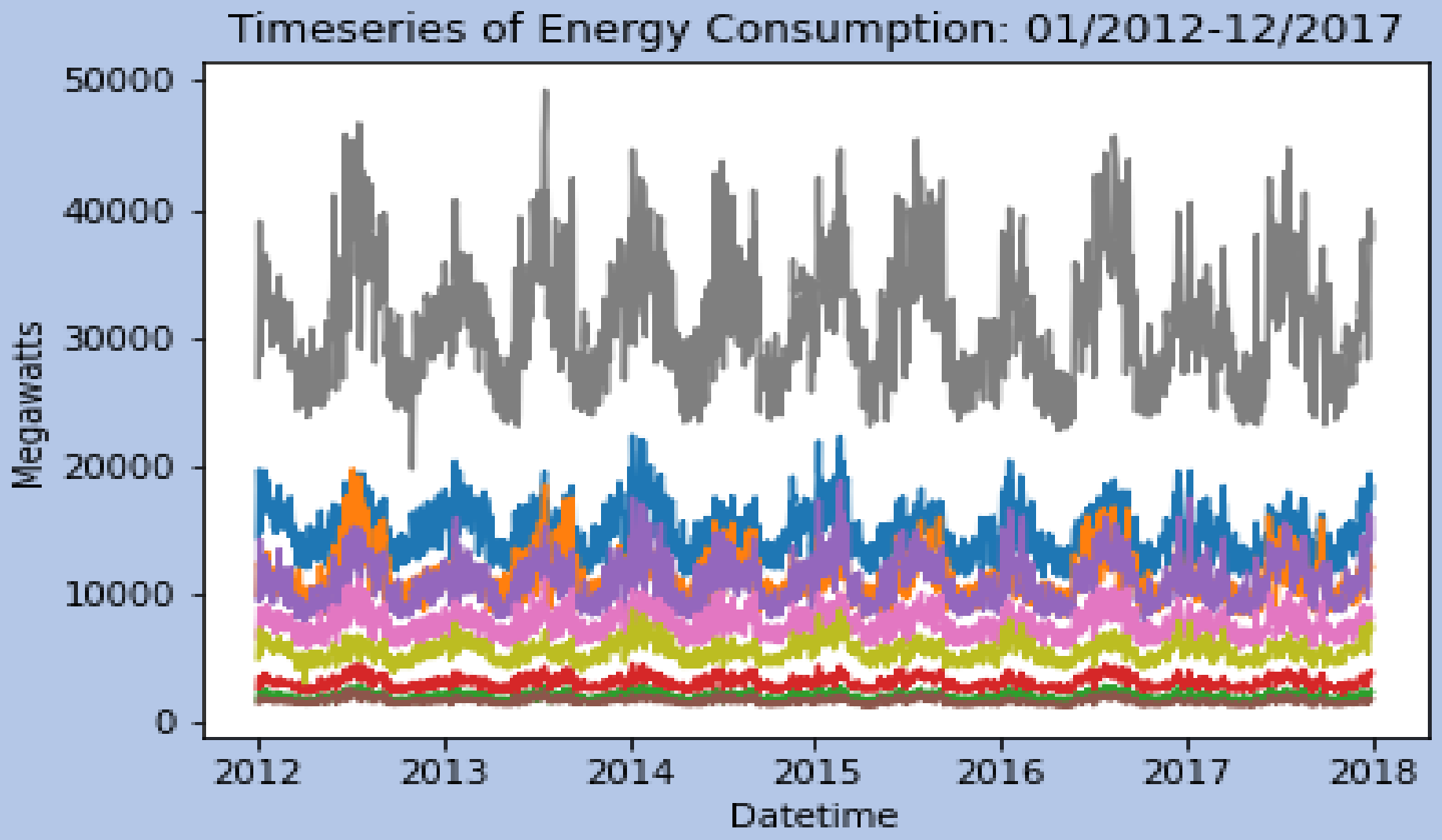


Data Wrangling



- Plotting all utility power as individual features
- Inconsistent time frames
- Utility changes w/in PJM

Data Wrangling



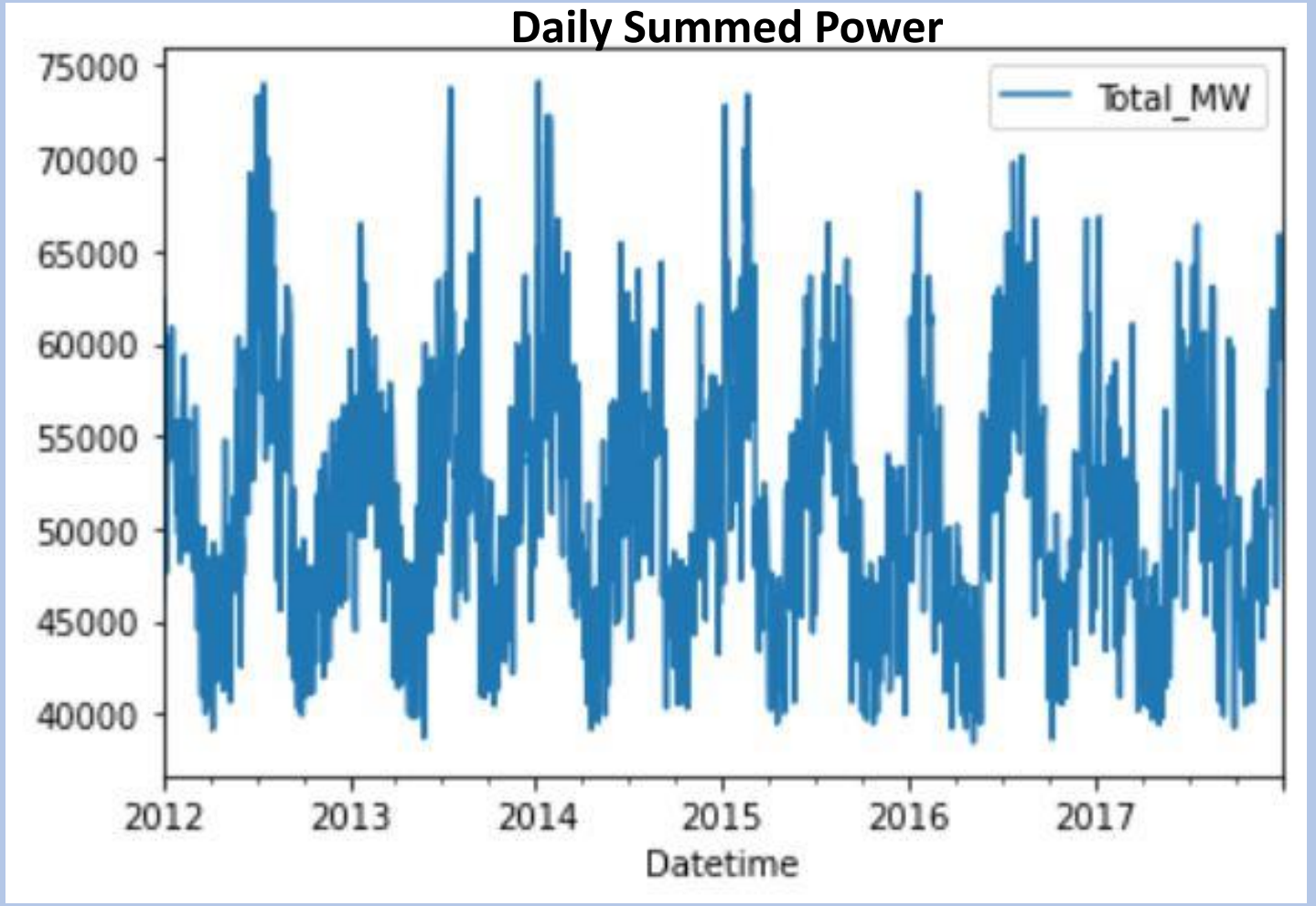
Data Wrangling

Datetime	AEP_MW	COMED_MW	DAYTON_MW	DEOK_MW	DOM_MW	DUQ_MW	FE_MW	PJME_MW	PJMW_MW
2012-01-01	14397.375000	9894.000000	1715.125000	2629.913043	9475.250000	1504.916667	6571.583333	26920.916667	4902.875000
2012-01-02	16882.958333	11217.833333	2061.166667	3179.875000	10587.333333	1686.375000	7533.041667	30463.166667	5857.791667
2012-01-03	19687.333333	12487.833333	2352.083333	3503.375000	13584.708333	1883.500000	8889.125000	36219.208333	6965.708333
2012-01-04	19357.458333	12119.541667	2323.208333	3470.541667	14265.166667	1947.750000	8903.041667	39069.166667	7296.583333
2012-01-05	17448.541667	11694.000000	2155.833333	3210.125000	11898.708333	1794.750000	8425.541667	35579.583333	6425.333333

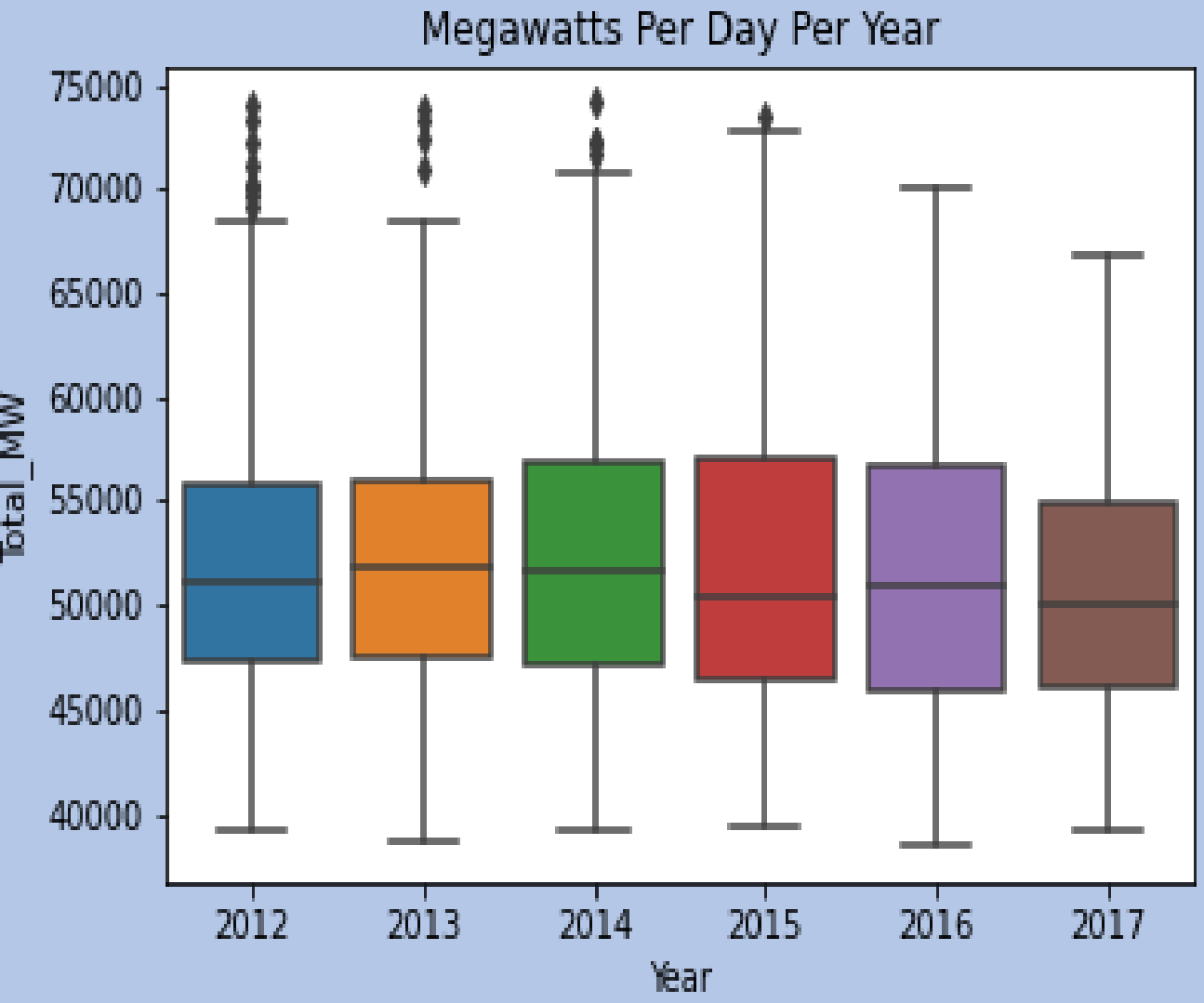
- Join individual features (columns):
  - Index – Timeseries
  - Column = daily sum of utility power consumption

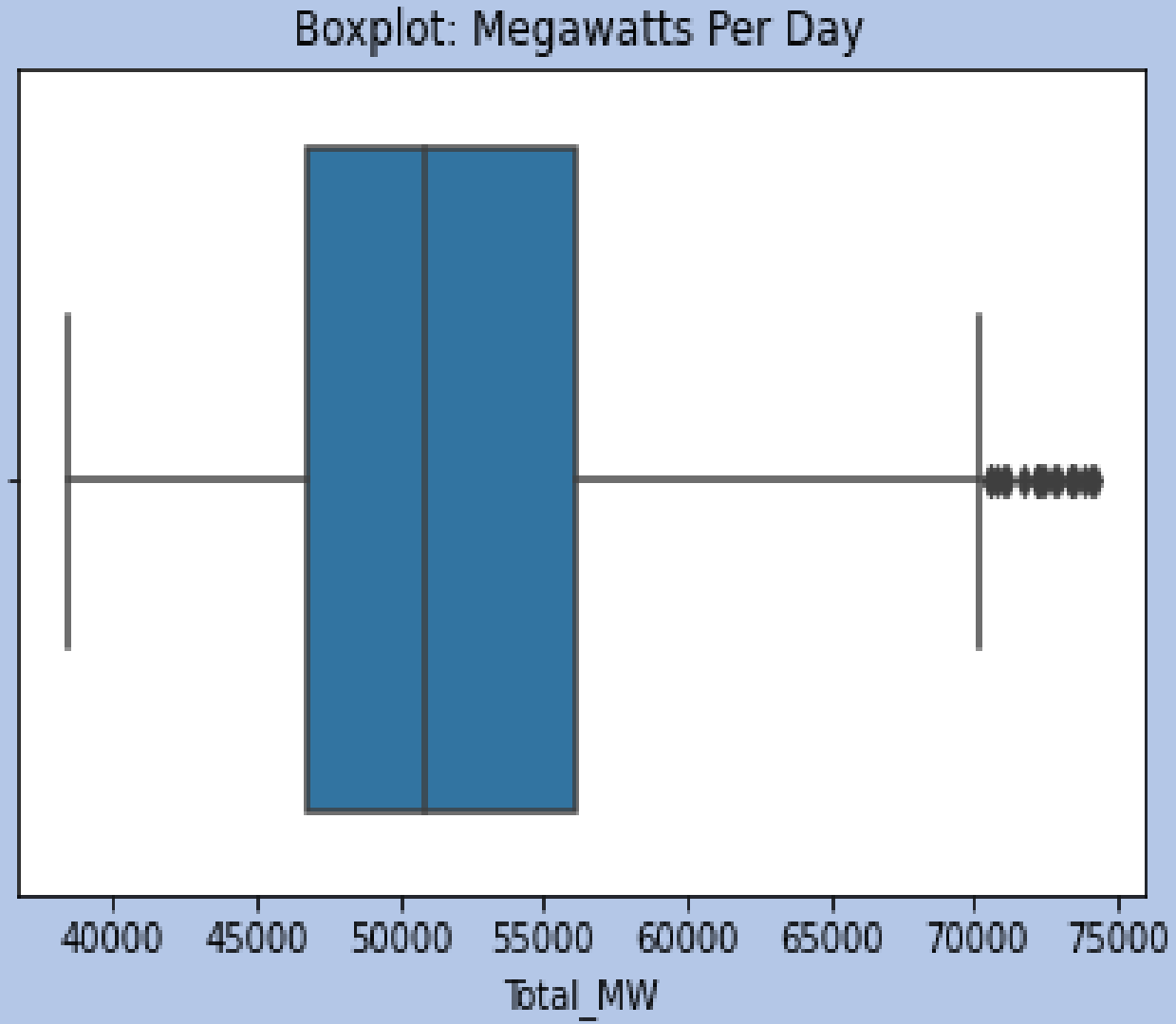
Data Wrangling

Datetime	Total_MW
2012-01-01	46188.163043
2012-01-02	53148.583333
2012-01-03	62387.958333
2012-01-04	62386.708333
2012-01-05	56627.500000



Exploring The Data

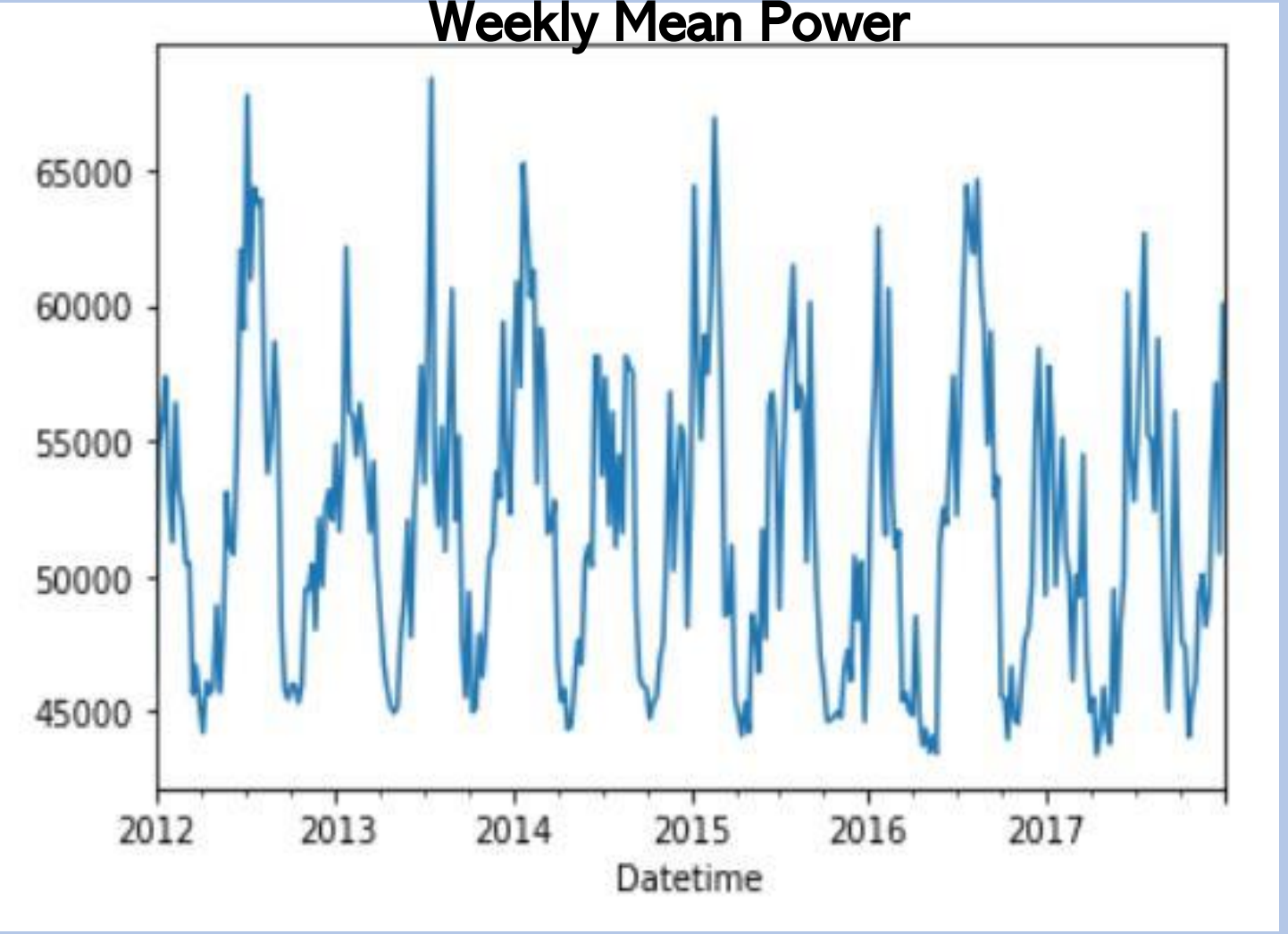




Exploring The Data

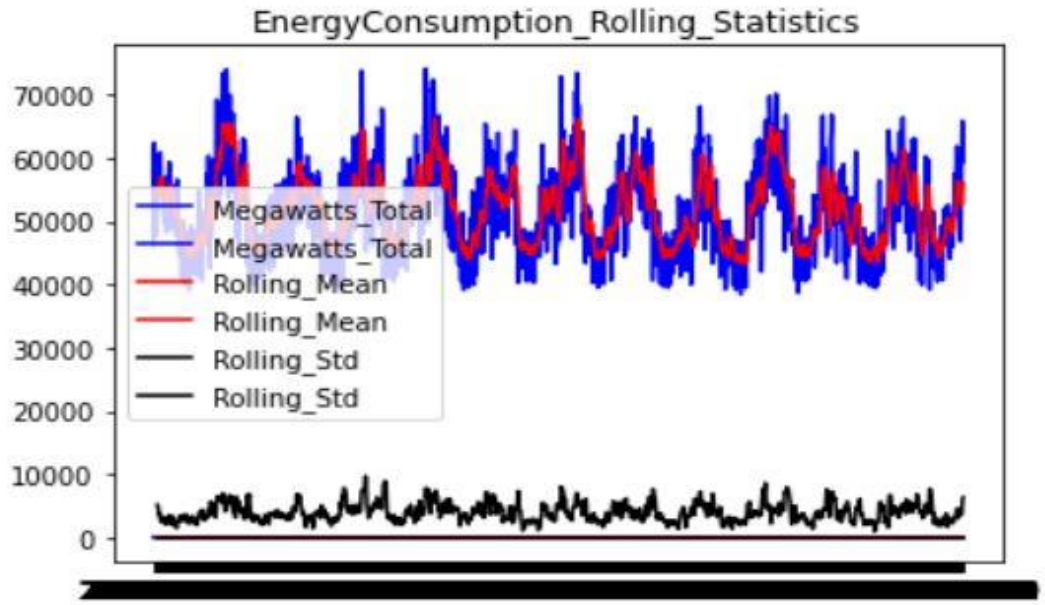
Resampled for weekly average

Datetime	Total_MW	Weeks
2012-01-01	46188.163043	1
2012-01-08	54688.875000	2
2012-01-15	55755.767857	3
2012-01-22	57336.458333	4
2012-01-29	53328.892857	5



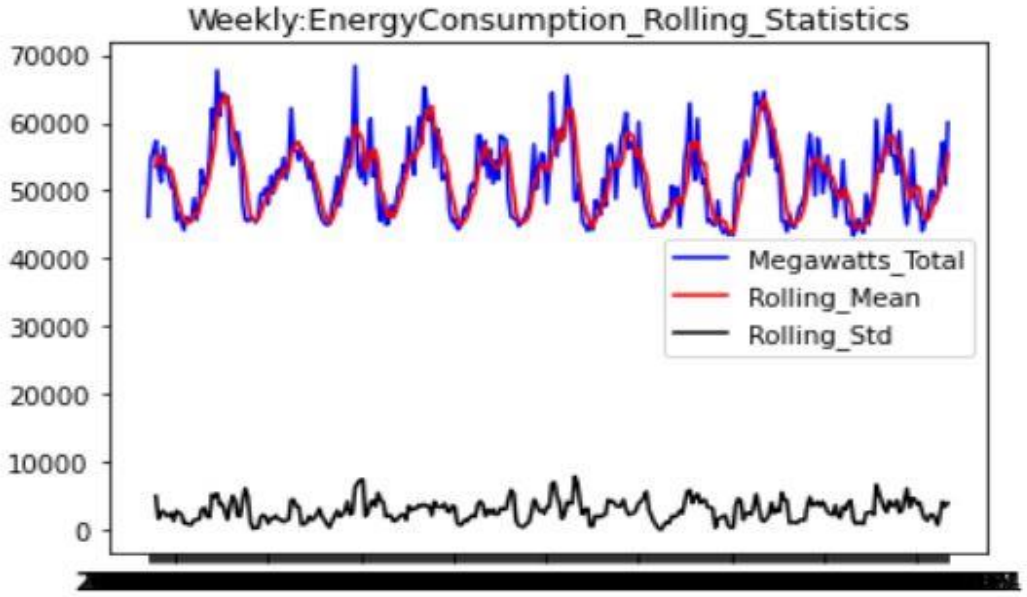
Augmented Dicky-Fuller Test: Stationarity

Daily Summed




```
Result: Dicky-Fuller Test
Test Statistics                  -4.372127
p-value                           0.000332
No. of lags used                 26.000000
Number of observations used    2165.000000
critical value (1%)              -3.433374
critical value (5%)              -2.862876
critical value (10%)             -2.567481
dtype: float64
```

Weekly Summed




```
Result: Dicky-Fuller Test
Test Statistics                -8.085169e+00
p-value                         1.426286e-12
No. of lags used                1.100000e+01
Number of observations used     3.020000e+02
critical value (1%)            -3.452190e+00
critical value (5%)            -2.871158e+00
critical value (10%)           -2.571895e+00
dtype: float64
```

Train – Test Sets

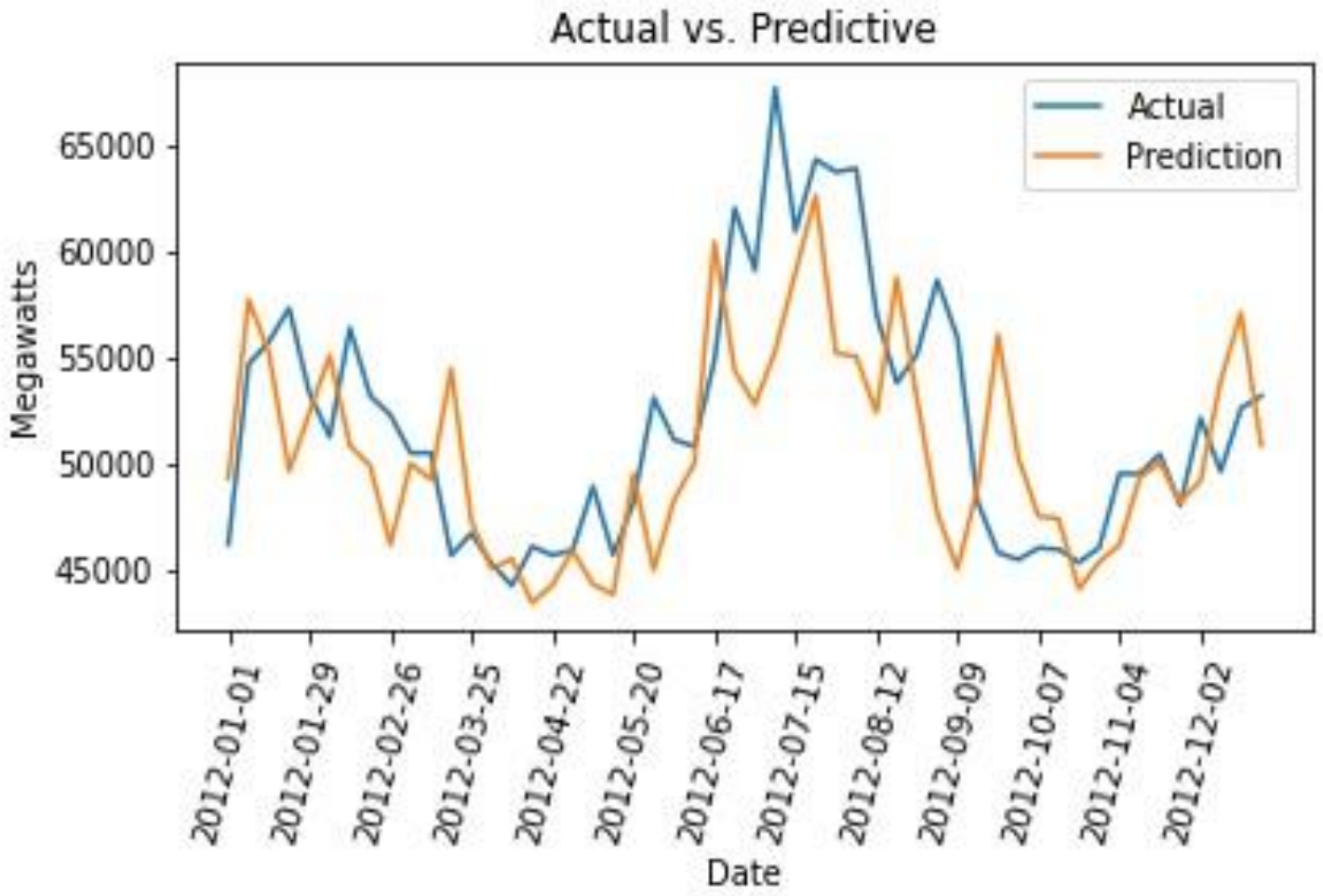
- 6 years ~ 314 weeks
- Train set: 262 weeks (5-yr)
- Test set: 52 (1yr)

```
X = df_W.Total_MW
size = int(len(X) * 0.837)
train, test = X[0:size], X[size:len(X)]

print("Train set: " + str(len(train)) + " weeks")
print("Test set: " + str(len(test)) + " weeks")
```

```
Train set: 262 weeks
Test set: 52 weeks
```

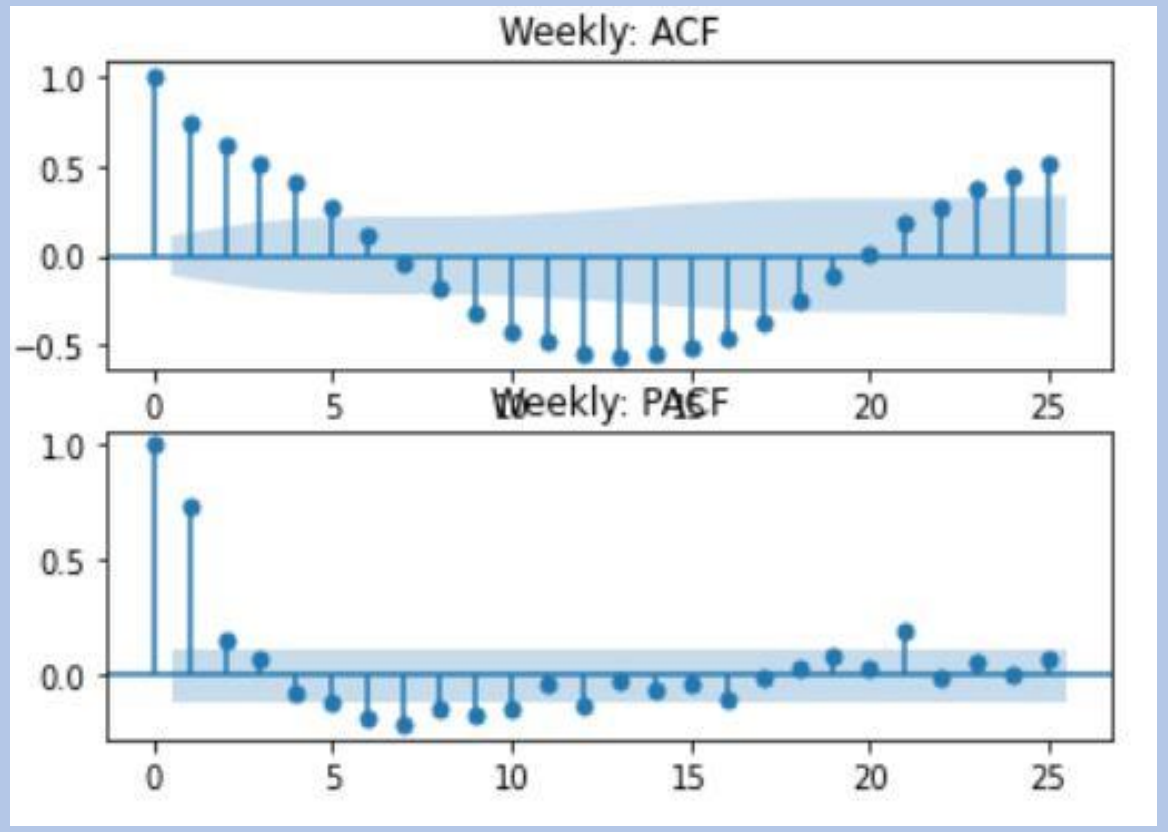
Persistence Model

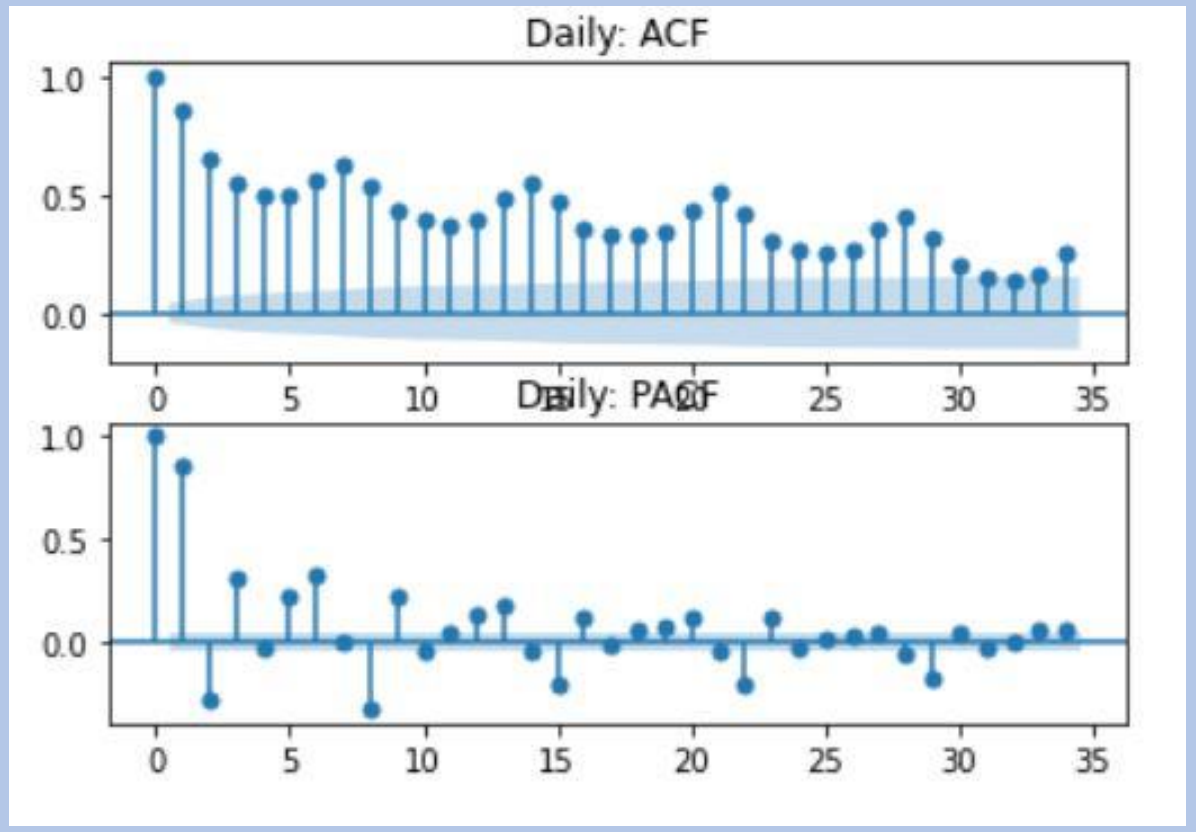


Baseline model using the previous step (t-1) to predict the following step (t+1)

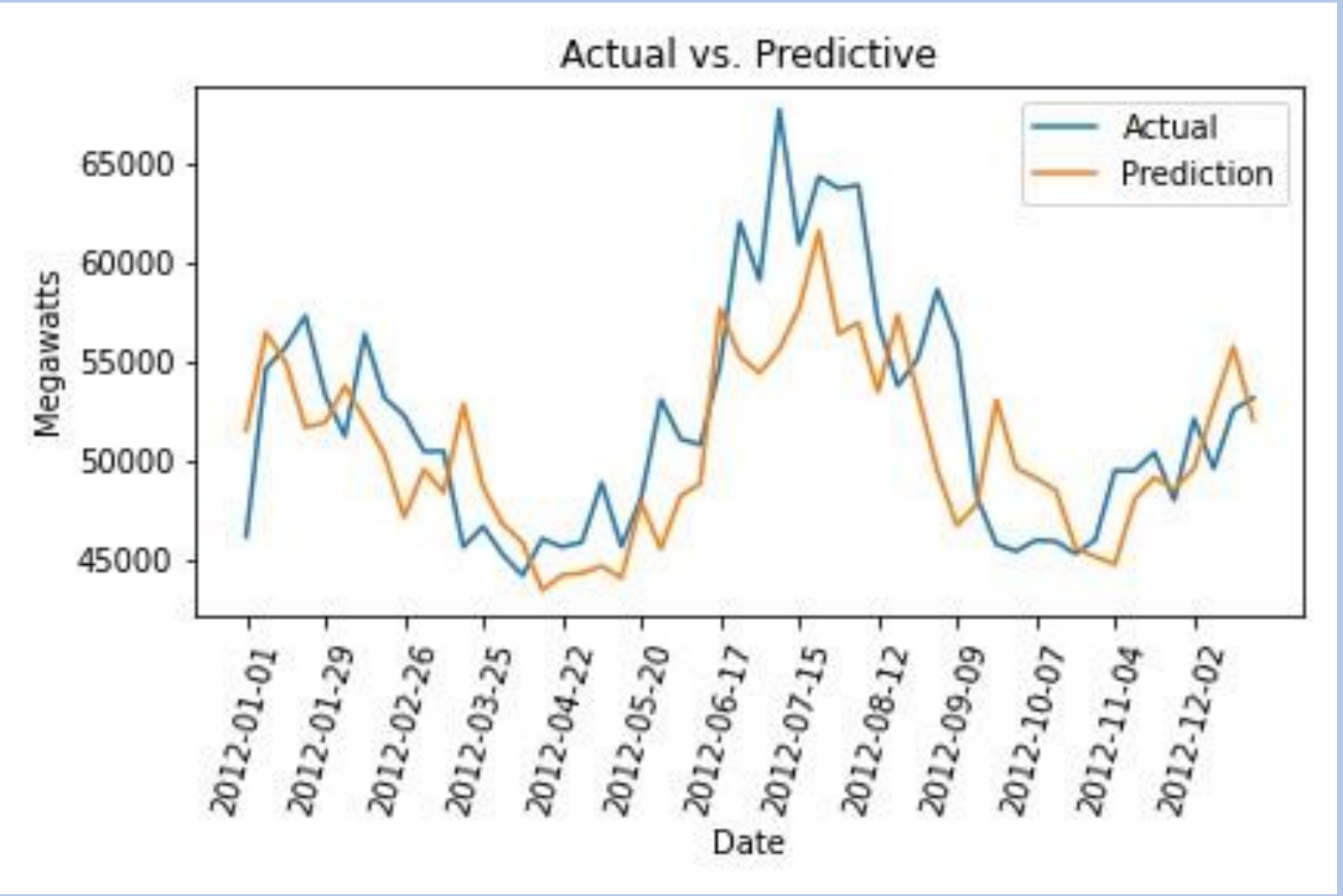
- *RMSE: 4344*
- *Can we improve on the accuracy of this model?*

Autocorrelation – Partial Autocorrelation





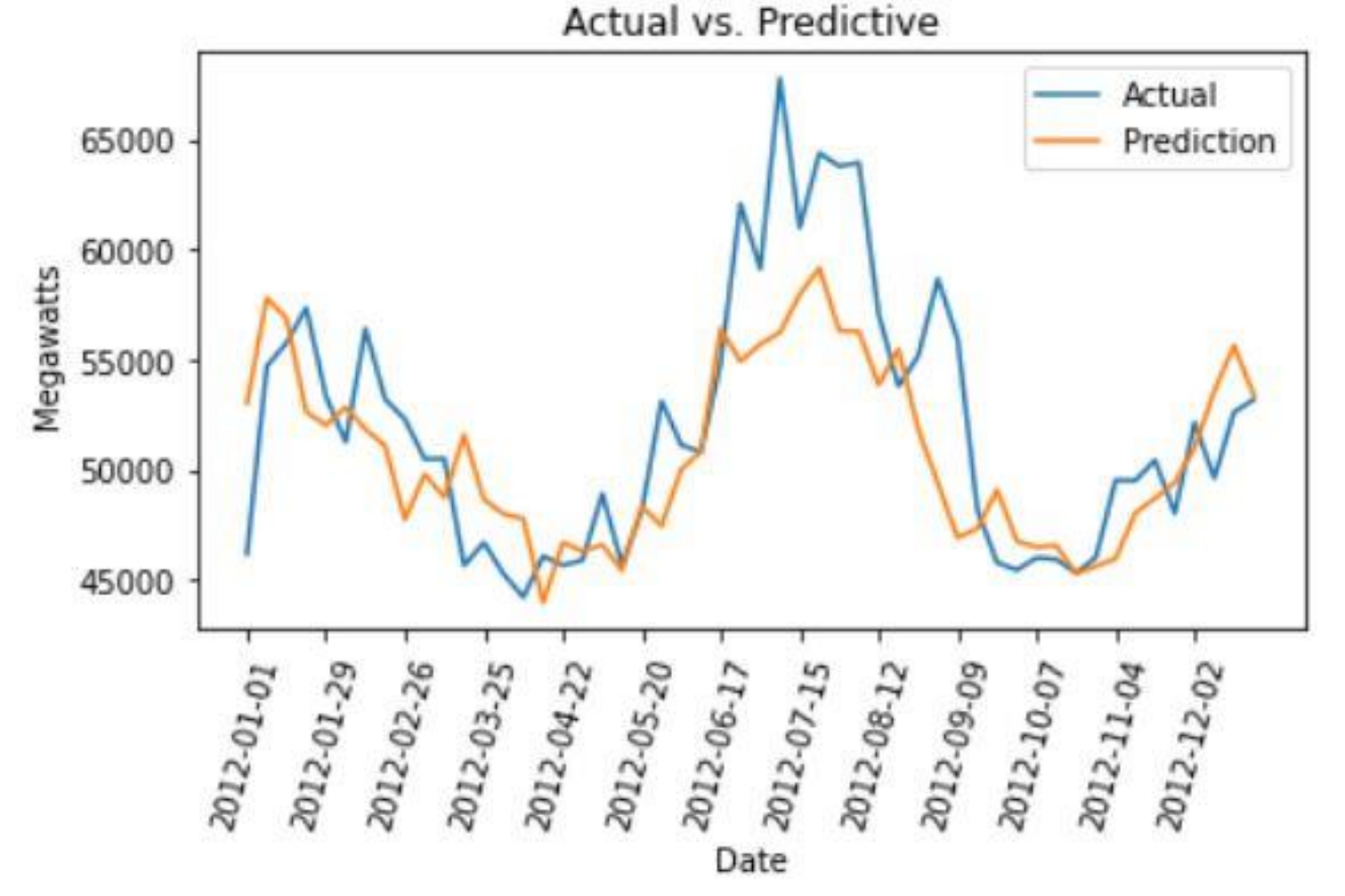
Autoregressive Integrated Moving Average (ARIMA)



ARIMA(6,1,1)

- *RMSE: 4093*
- *Slightly reduced RMSE compared to persistence*

ARIMA: Hyperparameter Tuning

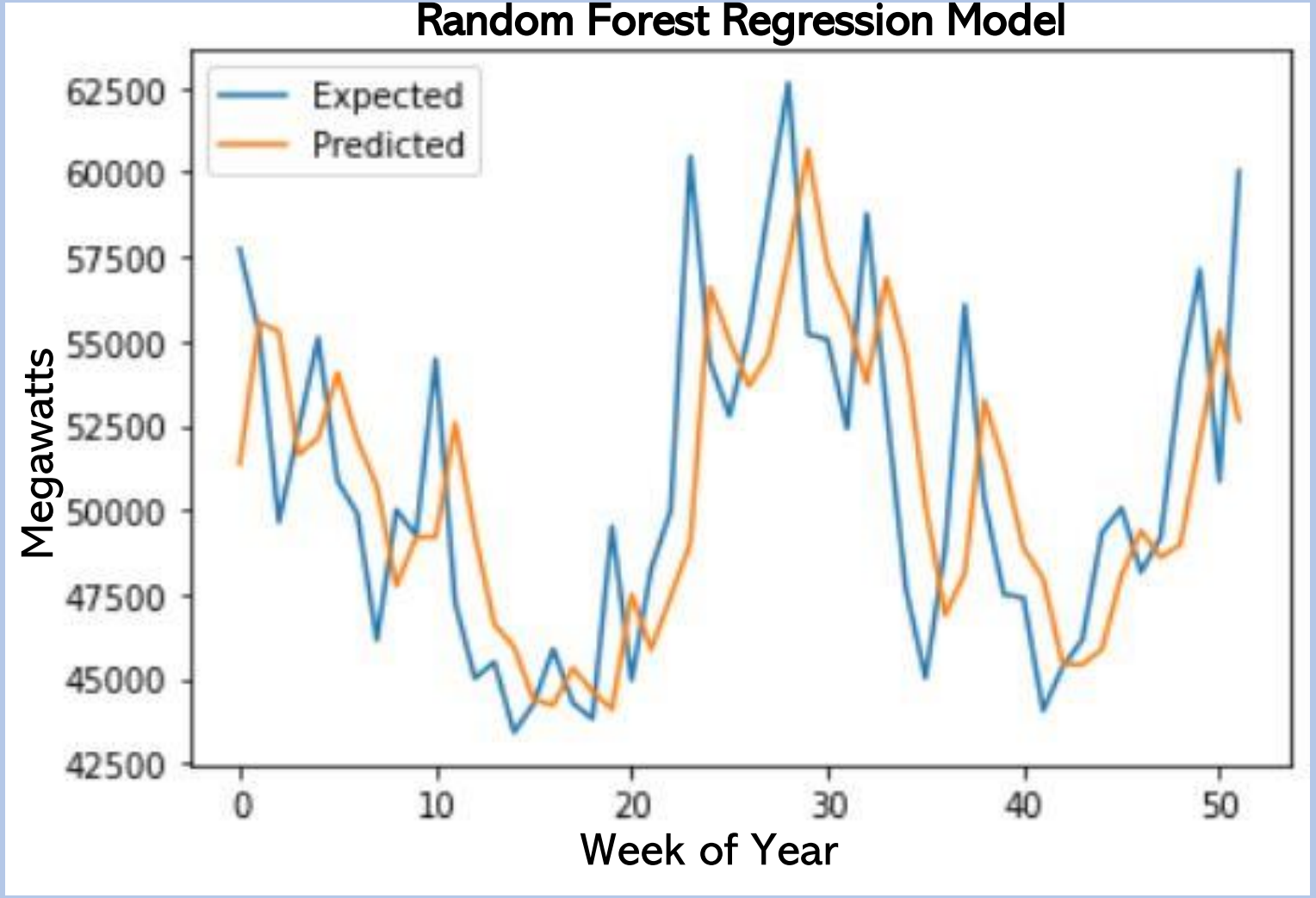


Search over (p,d,q) values…

ARIMA(7,0,2) smaller RMSE

- *RMSE: 3684*
- *Slightly reduced RMSE compared to persistence*

Random Forest Regression Model

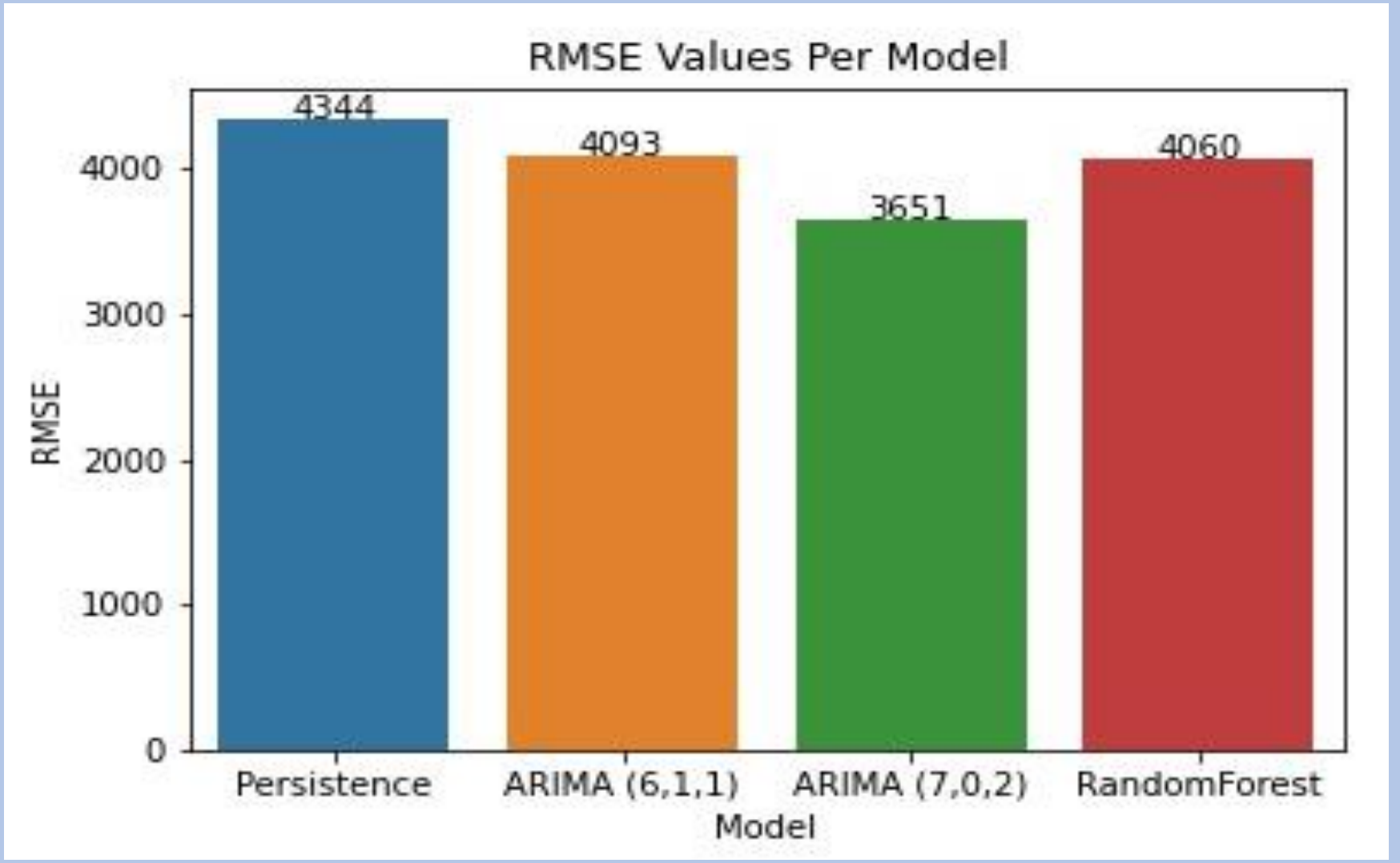


- Add column “Weeks”
- Create “supervised learning”

	Weeks	Total_MW
0	1	46188.163043
1	2	54688.875000
2	3	55755.767857
3	4	57336.458333

- *RMSE: 4060*
- *Better than “persistence”*

Model Performance



Conclusions

- All models performed well in predicting trends of power consumption
- Predicting peaks from unseasonal weather, extreme events may require further feature engineering
- Hyperparameter tuning can provide additional error reduction, improved accuracy
- Generalized trends can be well understood using ARIMA, Persistence Modeling, and Random Forest Regression

Future Considerations

- Adding additional features (e.g. weather, population of utility regions) can further support supervised learning models

- Correlation of features relating to predicting spikes in power may allow the model to more accurately predict summer and winter power consumption extremes
  - This can be critical for transmission line planning and performance

- Additional hyperparameter tuning of well performing models may increase accuracy
  - ARIMA(p,d,q) grid search

- Testing additional model types (e.g. Prophet, LSTM) can provide more detailed comparisons