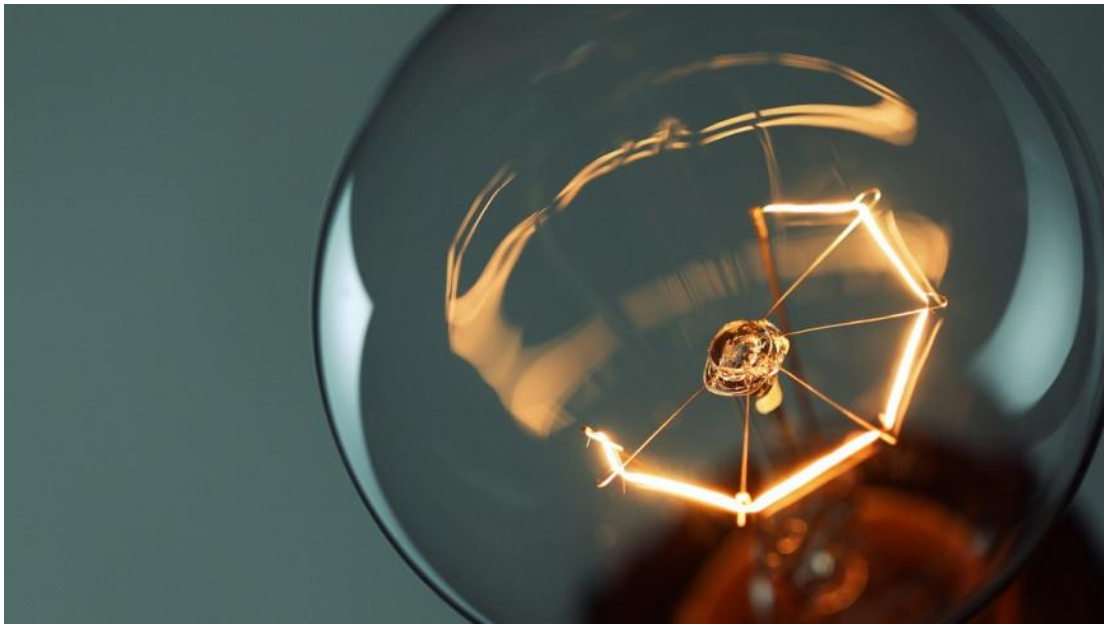
Individual Electricity Cost

Time Series Analysis and Forecasting



Main Objective of Analysis

Household electricity cost comprises a significant portion of a utility bill and depending on the time of the year, fluctuates on factors such as utilizing air conditioners, heaters, and personal electronics, as well as lighting. The objective of this analysis is to track the cost of electricity for an individual living in an apartment in San Jose, California over the course of two years, which was measured by smart meters and shared by Pacific Gas and Electric, a utility company. Various models were created to this end, including single, double, and triple exponential smoothing as well as variations on autoregressive and moving average models such as AR, MA, ARMA, and SARIMA. Forecasts were tested and using standard metrics, and the best model to predict the cost of electricity was determined. The Python library statsmodels was used to conduct the analysis.

Description of Dataset

The dataset was recorded from smart meters, which tracked the electricity usage and cost over time every 15 minutes for a period of 733 days. Several other attributes were included, but not relevant for the purposes of this analysis, so were later dropped during data preprocessing. The following is a summary of the original dataset:

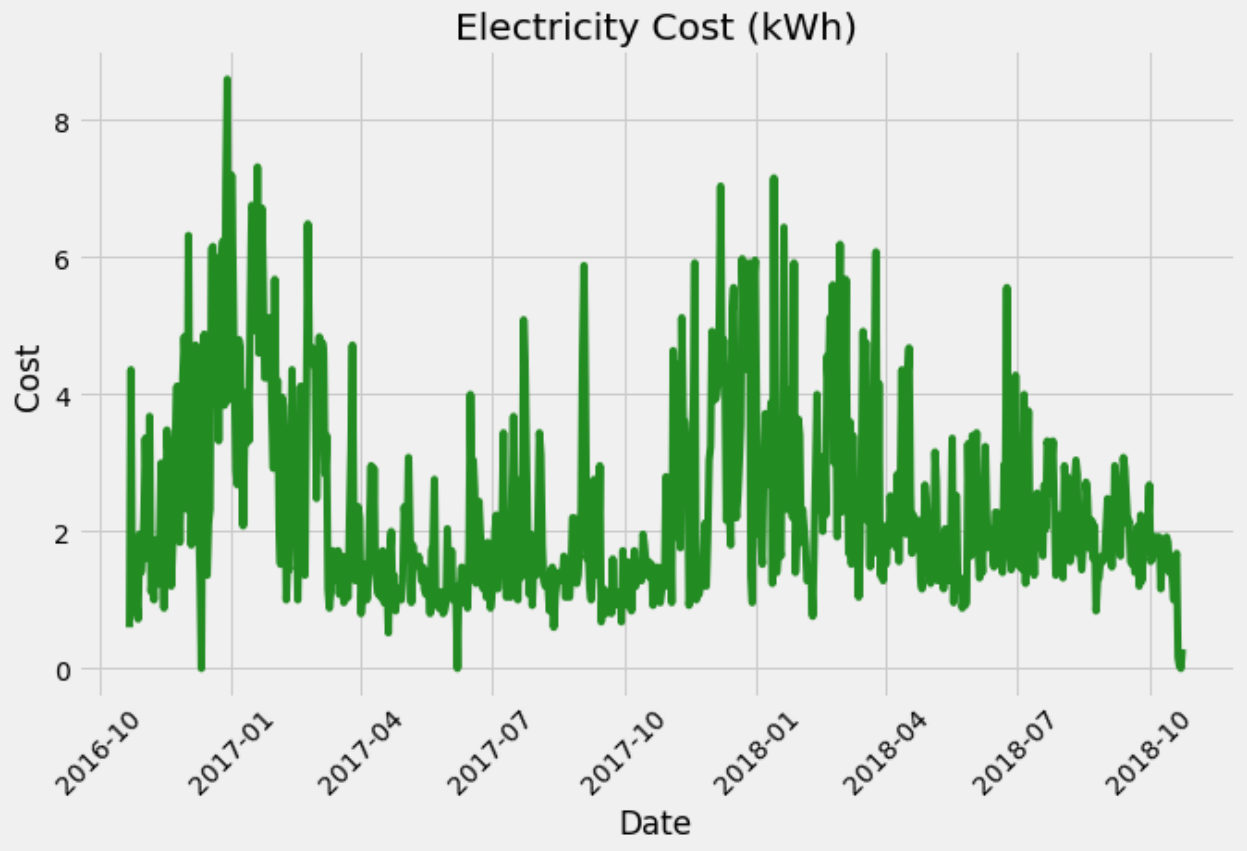
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70368 entries, 0 to 70367
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   TYPE        70368 non-null  object
 1   DATE        70368 non-null  object
 2   START TIME  70368 non-null  object
 3   END TIME    70368 non-null  object
 4   USAGE       70368 non-null  float64
 5   UNITS       70368 non-null  object
 6   COST        70368 non-null  object
 7   NOTES       0 non-null      float64
dtypes: float64(2), object(6)
memory usage: 4.3+ MB
```

Data Preprocessing and EDA

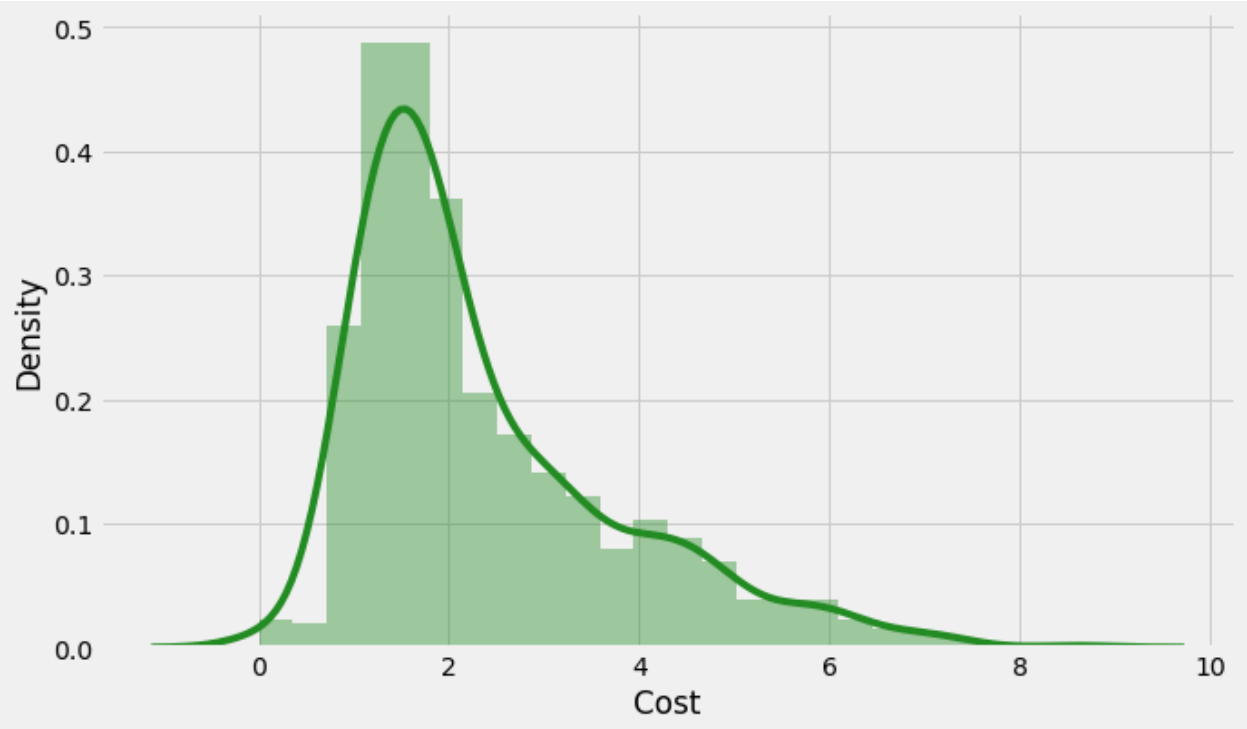
Several attributes were dropped from the original dataset. “TYPE” contained only strings of “Electric usage,” “UNITS” contained only strings of “kWh” and “NOTES” were all null values and because these did not contain any useful information, they were dropped. The analysis used the unit of day for the time series, and because of that, “START TIME” and “END TIME” were dropped as they were in 15 minute intervals. The variable of interest over time is cost and in the original dataset, the values were strings with a dollar sign in front of the number, so the symbol was dropped and then the strings of numbers were cast as numerical float values. “USAGE” was dropped as well, as the objective was to analyze and forecast the cost of electricity. All of the values recorded over the 15 minute intervals were summed per day and the index was set to the date, which was originally a string object, but was converted to a datetime object. The following is a summary of the preprocessed dataset:

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 733 entries, 2016-10-22 to 2018-10-24
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Cost    733 non-null    float64
dtypes: float64(1)
memory usage: 11.5 KB
```

The following is a line graph of the cost of electricity per day over the course of 733 days. It is apparent that seasonal fluctuations are present over the course of time, and that makes intuitive sense given changes in weather causing the individual to use more electricity in winter and summer months for heating and cooling, as well as in the spring months, when the climate is more moderate, electricity costs are lower.

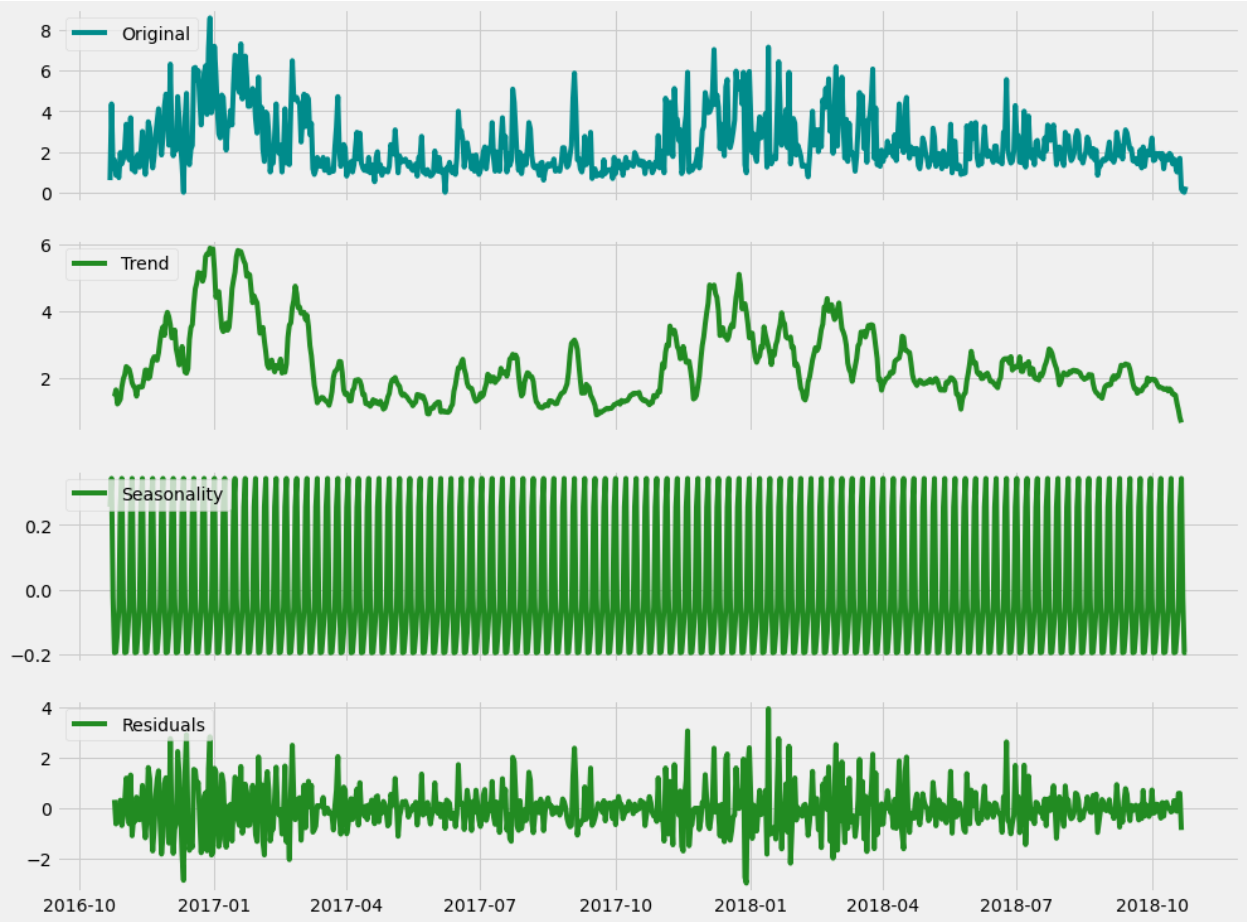


The following is a distribution plot of how frequent different costs in electricity are for the individual per day. The distribution is right skewed and while the individual typically uses less than $2 worth of electricity per day, a significant portion of daily electricity usage is higher than this.

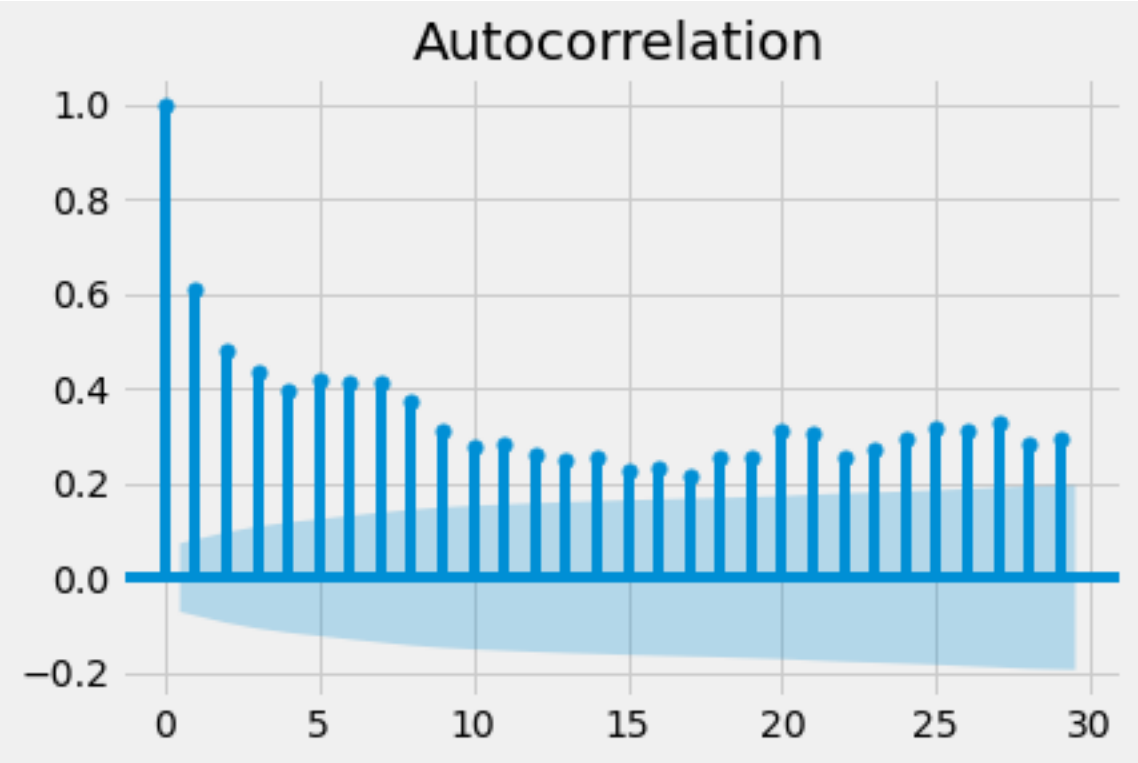


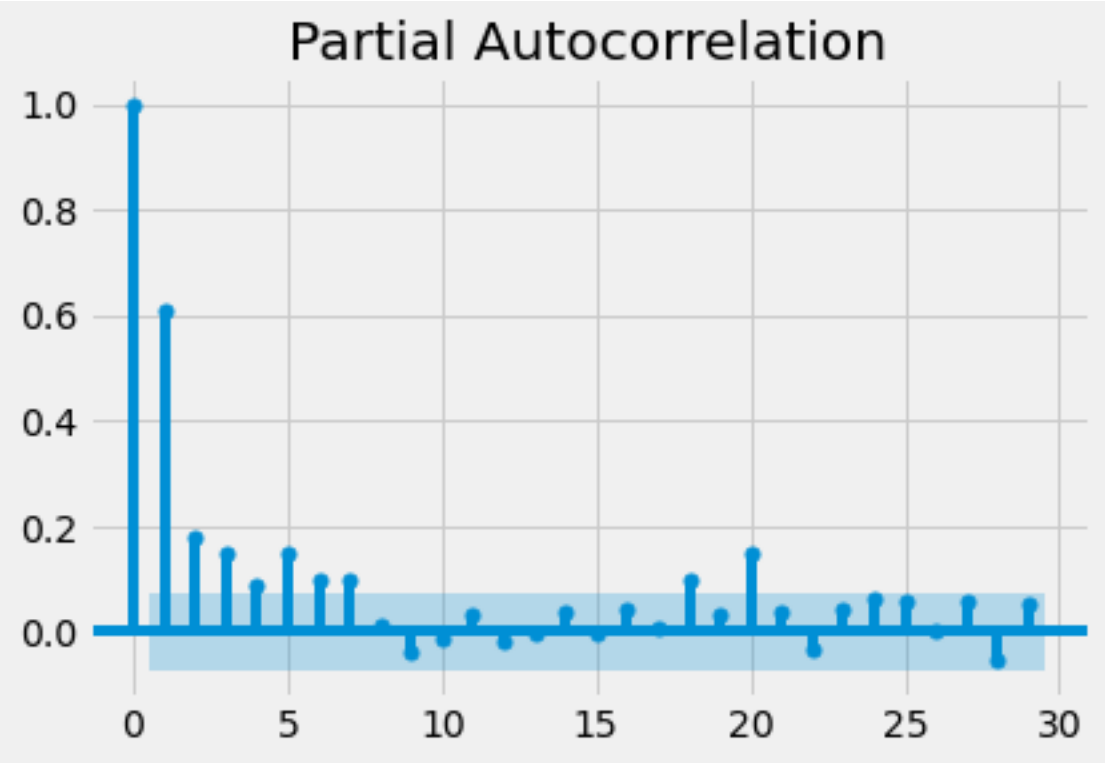
Time Series Models and Forecasting

The original time series was decomposed into trend, seasonality, and residual. The shape of the line plot suggests an additive model with trend and seasonality.



The autocorrelation and partial autocorrelation plots indicate that the time series has significant correlation between lags and because the partial autocorrelation plot starts high and quickly decreases exponentially, an AR model would seem appropriate.



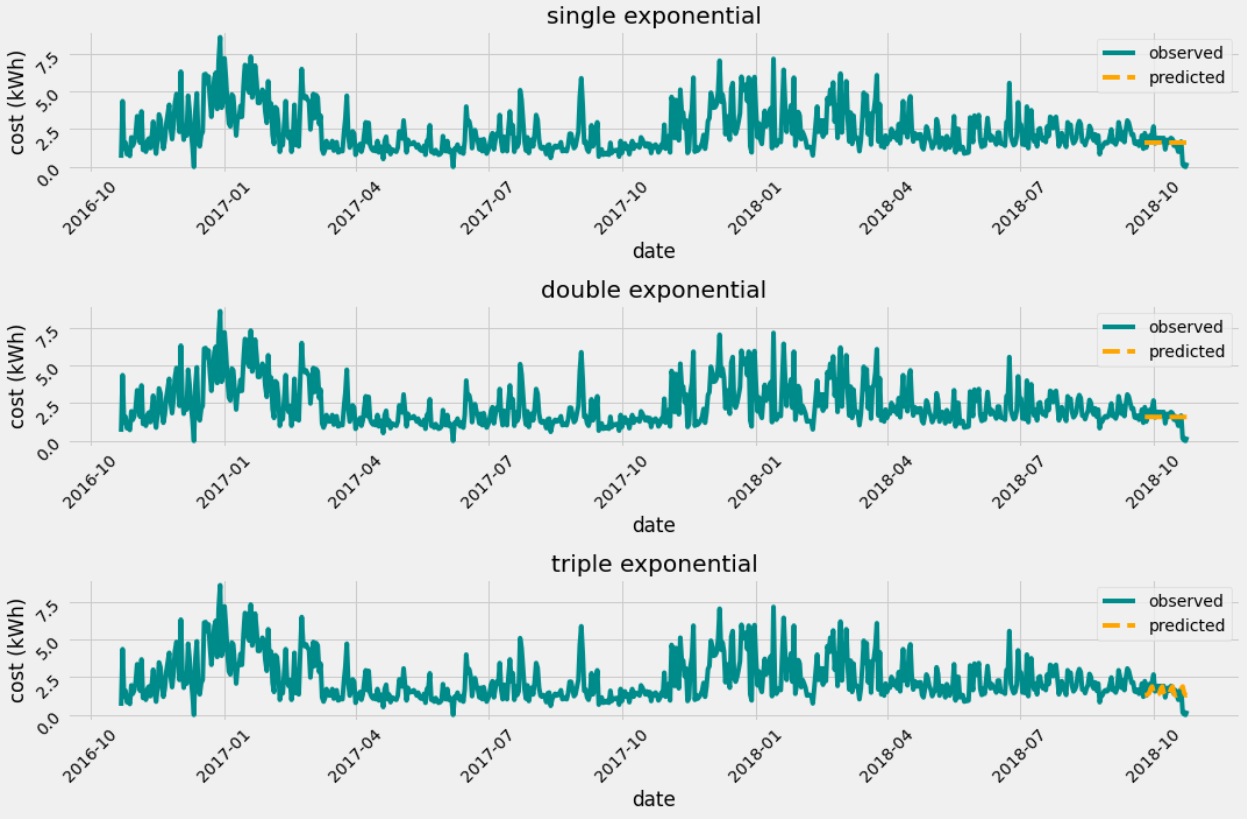


An augmented Dicky-Fuller test was performed to determine the stationarity of the time series. The null hypothesis of this test is that the time series is not stationary. Here, the p-value is about 0.087, which is above the standard 0.05 threshold, so we cannot reject the null hypothesis and conclude that the time series is not stationary.

```
Summary statistics
adf: -2.6315469612542324
p value: 0.08665084537270223
used lag: 19
number of observations: 713
critical values: {'1%': -3.43955476721974, '5%': -2.865602155751202, '10%'
: -2.5689331692727135}
icbest: 2108.3934371297855
```

Exponential Smoothing

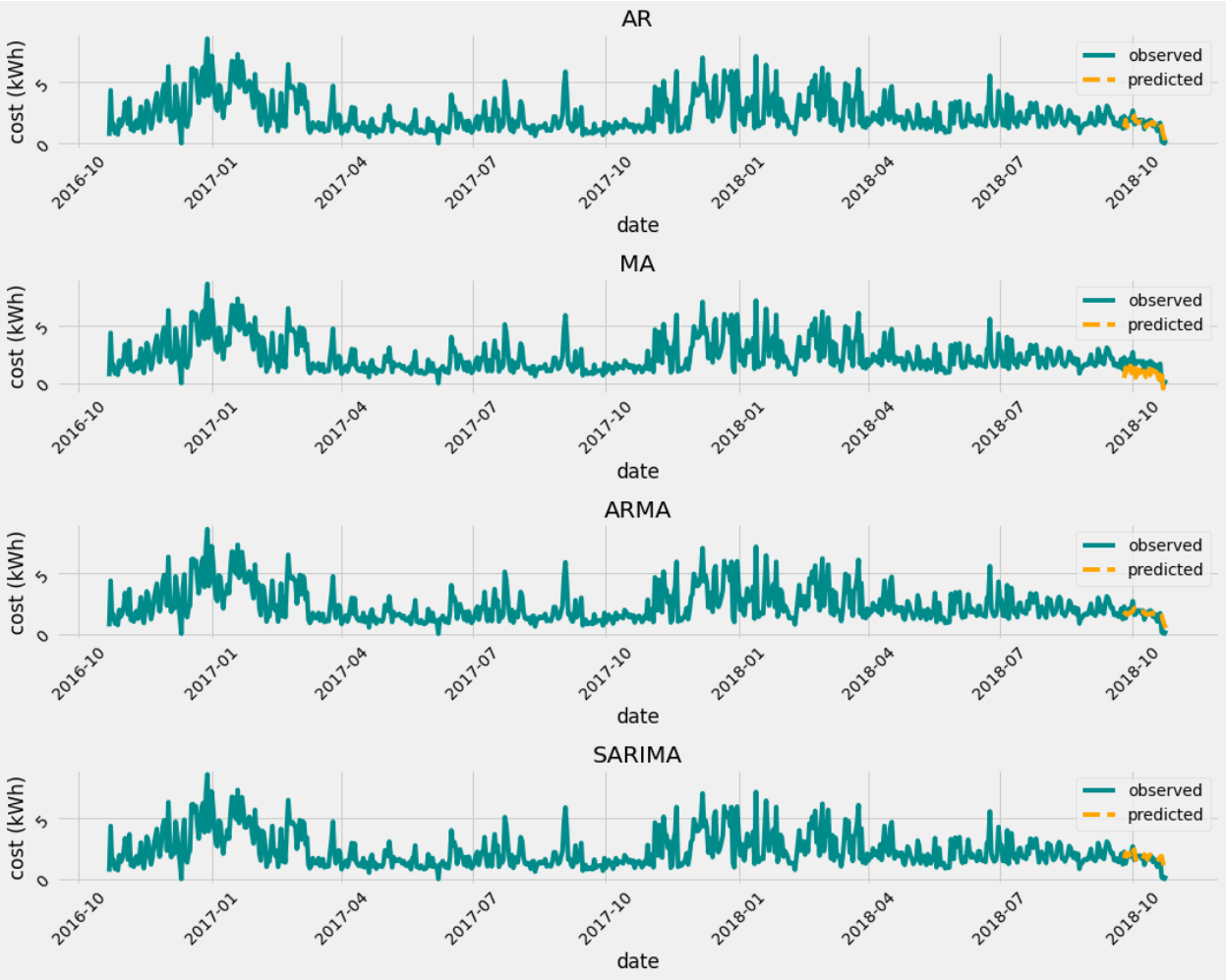
In order to address the stationarity issue in this time series, single, double, and triple exponential smoothing were used to create models and then make forecasts. The test set was the last 30 days in this time series and the train set was everything before. The plots, as well as the scores for mean square error suggest that there were only slight differences between these variations of exponential smoothing.




```
single exponential mse: 12.38
double exponential mse: 12.32
triple exponential mse: 12.35
```

Autoregressive and Moving Average Models

The partial autocorrelation plot indicated that an AR model would be the most appropriate in forecasting electricity cost. To confirm this assumption, various autoregressive and moving average models were created and compared with different hyperparameters, including AR, MA, ARMA, and SARIMA. Models of AR with p=3, MA with q=2, and ARMA with p=1 and q=1 were created. Also, using `pmdarima.auto_arima()`, the best hyperparameters were cross validated for SARIMA, with p=0, d=0, q=2, P=0, D=1, Q=1. As in the previous exponential smoothing models, the test set for the forecast was the last 30 days of the time series and the train set was all of the days before. As expected from the partial autocorrelation plot, the AR model performed the best both visually on the plot, as well as with the lowest mean squared error.




```
AR mse: 6.52
MA mse: 23.19
ARMA mse: 6.69
SARIMA mse: 10.19
```

Key Findings and Insights

Seven different time series models were created to make forecasts in this analysis. The exponential smoothing models all had similar results in the predictions. Double exponential smoothing accounts for trend and triple exponential smoothing accounts for seasonality and because all three models performed similarly in making forecasts of electricity cost, the trend and seasonality of the time series was not as significant as initially assumed. As expected with the partial autocorrelation plot, the AR model performed the best among the other autoregressive moving average models. The MA model had the highest mean squared error and this indicates that the time series had a high dependence on past values rather than past forecast errors. The ARMA model resulted in a good fit as well. The mean squared error of the SARIMA model suggests that differencing did not result in better forecasts and supports the observation that trend and seasonality did not play such a significant role in this time series.

Next Steps

The time series on electricity cost used in this analysis was only from one individual for the course of about two years. The number of observations could be extended back to even more time and perhaps this would result in better forecasts. In addition, a time series of the average electricity cost for San Jose could be used as a point of comparison. Finally, similar data could be collected from different cities and because electricity costs differ depending on locality, the electricity usage could be used as a means of analysis and prediction.