Capstone Project 1: In-depth Analysis (Machine Learning)

To objective is to forecast the sales quantities per month of products sold into the renewable energy market.

The dataset is composed of 5 years (2014 – 2018) of sales quantities.

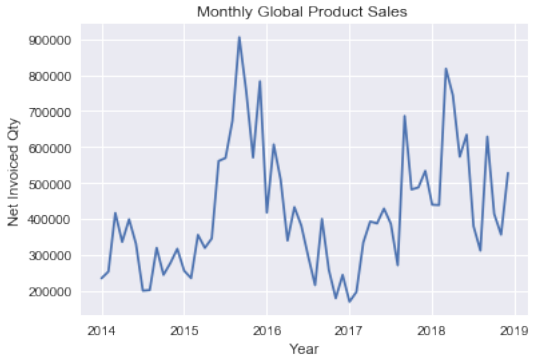
Evaluation of Model:

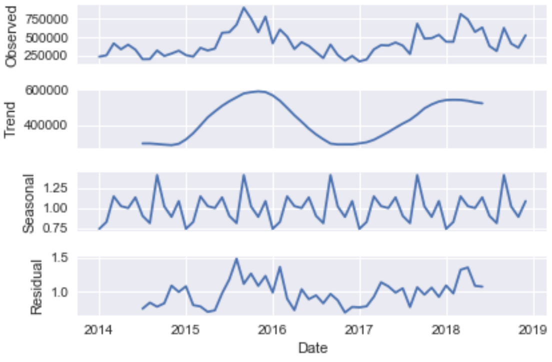
The first 4 years of the data are split into a training and test data. The ARIMA model is used for the prediction of the unit sales

- The performance of the model predictions is evaluated from their RMSE (Root Mean Squared Error) calculated by taking the square root of the output of the mean_squared_error function in the scikit-learn library. This is the criteria used in choosing the optimal hyperparameters of the model.
- To training and test data is split in a ratio of 3:1. 75% of the data is used to train and 25% to test.

Data Analysis:

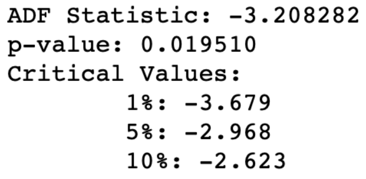
Analysis of any time series assumes that it is stationary (mean, variance, autocorrelation, etc. are all constant over time). Clearly as earlier shown this is not the case. The means vary from year to year. Decomposing the series using seasonal_decompose, we can see that there is seasonality in the series. As shown below



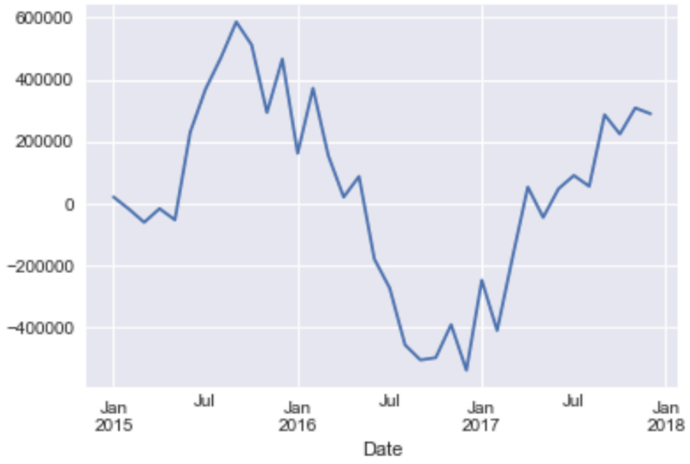


Stationarizing the series:

To stationarize the series we use differencing and then use a statistical test (Augmented Dickey-Fuller test) to confirm that the series is stationary. As shown below the test statistic value -3.208282 is smaller than the critical value at 5% of -2.968. This suggests that we can reject the null hypothesis the data has a unit root and is non-stationary with a significance level of less than 5%. A stationary series has no trend, its variations around its mean have a constant amplitude, and it wiggles in a consistent fashion as the plot shows.

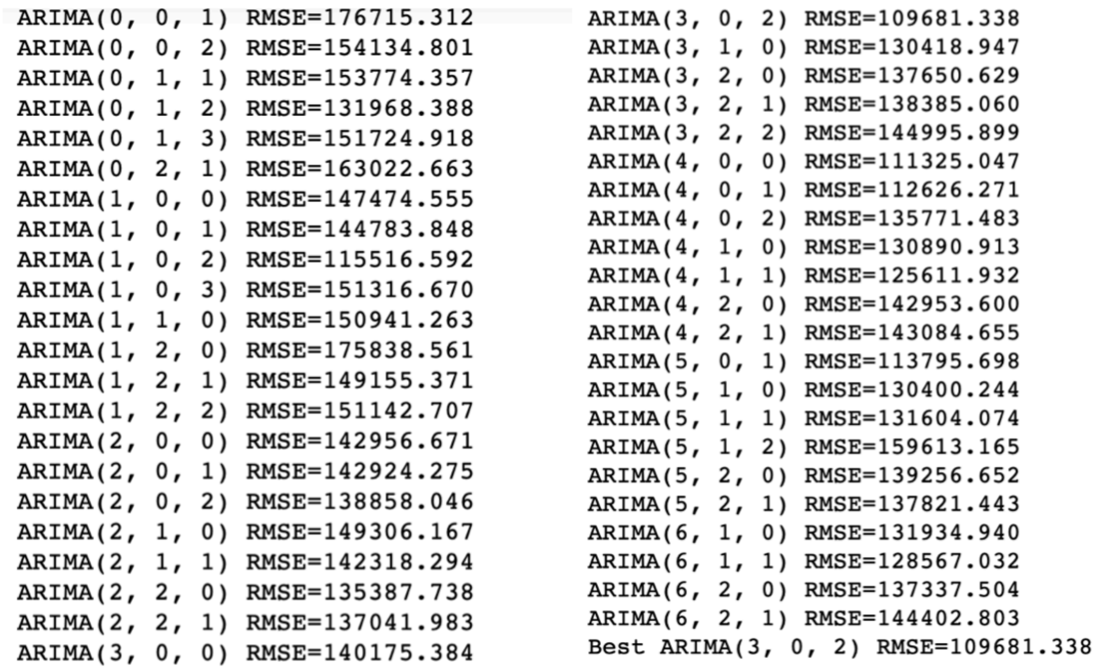


```
ADF Statistic: -3.208282
p-value: 0.019510
Critical Values:
        1%: -3.679
        5%: -2.968
        10%: -2.623
```



ARIMA Model Parameters (p,d,q):

The number of Autoregressive (AR) parameters p (0-6), difference parameters d (0-2) and Moving Average (MA) parameter q (0-6) have 147 combination possibilities. To choose the optimal (p,d,q) hyperparameters, we define a grid search function that iterates over all these combinations and chooses the best based on its RMSE being the smallest as show below.

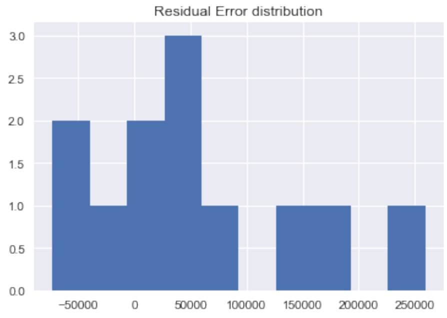


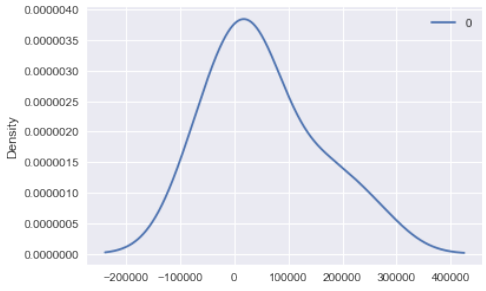
```
ARIMA(0, 0, 1) RMSE=176715.312      ARIMA(3, 0, 2) RMSE=109681.338
ARIMA(0, 0, 2) RMSE=154134.801      ARIMA(3, 1, 0) RMSE=130418.947
ARIMA(0, 1, 1) RMSE=153774.357      ARIMA(3, 2, 0) RMSE=137650.629
ARIMA(0, 1, 2) RMSE=131968.388      ARIMA(3, 2, 1) RMSE=138385.060
ARIMA(0, 1, 3) RMSE=151724.918      ARIMA(3, 2, 2) RMSE=144995.899
ARIMA(0, 2, 1) RMSE=163022.663      ARIMA(4, 0, 0) RMSE=111325.047
ARIMA(1, 0, 0) RMSE=147474.555      ARIMA(4, 0, 1) RMSE=112626.271
ARIMA(1, 0, 1) RMSE=144783.848      ARIMA(4, 0, 2) RMSE=135771.483
ARIMA(1, 0, 2) RMSE=115516.592      ARIMA(4, 1, 0) RMSE=130890.913
ARIMA(1, 0, 3) RMSE=151316.670      ARIMA(4, 1, 1) RMSE=125611.932
ARIMA(1, 1, 0) RMSE=150941.263      ARIMA(4, 2, 0) RMSE=142953.600
ARIMA(1, 2, 0) RMSE=175838.561      ARIMA(4, 2, 1) RMSE=143084.655
ARIMA(1, 2, 1) RMSE=149155.371      ARIMA(5, 0, 1) RMSE=113795.698
ARIMA(1, 2, 2) RMSE=151142.707      ARIMA(5, 1, 0) RMSE=130400.244
ARIMA(2, 0, 0) RMSE=142956.671      ARIMA(5, 1, 1) RMSE=131604.074
ARIMA(2, 0, 1) RMSE=142924.275      ARIMA(5, 1, 2) RMSE=159613.165
ARIMA(2, 0, 2) RMSE=138858.046      ARIMA(5, 2, 0) RMSE=139256.652
ARIMA(2, 1, 0) RMSE=149306.167      ARIMA(5, 2, 1) RMSE=137821.443
ARIMA(2, 1, 1) RMSE=142318.294      ARIMA(6, 1, 0) RMSE=131934.940
ARIMA(2, 2, 0) RMSE=135387.738      ARIMA(6, 1, 1) RMSE=128567.032
ARIMA(2, 2, 1) RMSE=137041.983      ARIMA(6, 2, 0) RMSE=137337.504
ARIMA(3, 0, 0) RMSE=140175.384      ARIMA(6, 2, 1) RMSE=144402.803
                                    Best ARIMA(3, 0, 2) RMSE=109681.338
```

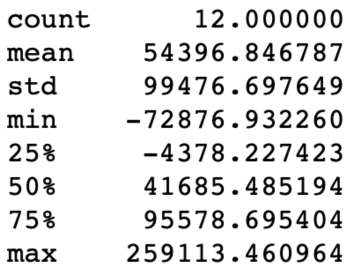
The best (p,d,q) is (3,0,2) chosen by virtue of its RMSE : 109681.3 . This order is used in the ARIMA Models to train the model.

Review of Residual Errors:

Ideally the distribution residual errors of a model should be normal around a mean of zero. The residual error distribution of the data set is approximately normal but not centered around zero as shown by the plot and summary statistics

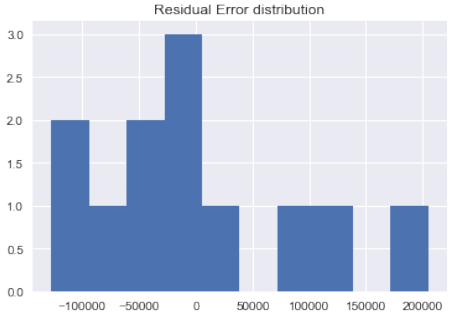


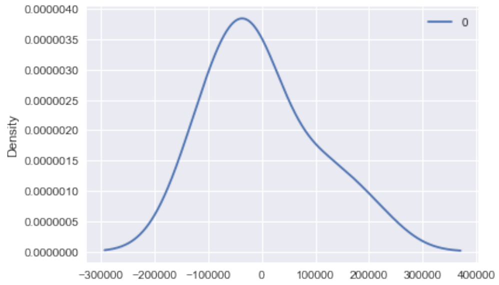


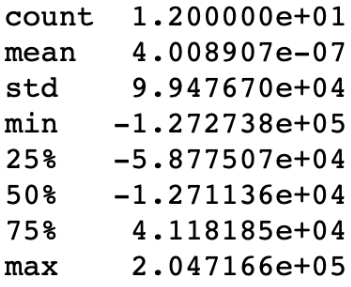


```
count        12.000000
mean      54396.846787
std       99476.697649
min      -72876.932260
25%       -4378.227423
50%       41685.485194
75%       95578.695404
max      259113.460964
```

To bias-correct prediction, the mean of the residual error 54396.85 is added to each prediction. Making this adjustment gives an improved distribution centered very close to zero with an improved RMSE: 95241.687 from RMSE : 109681.3 as shown below:



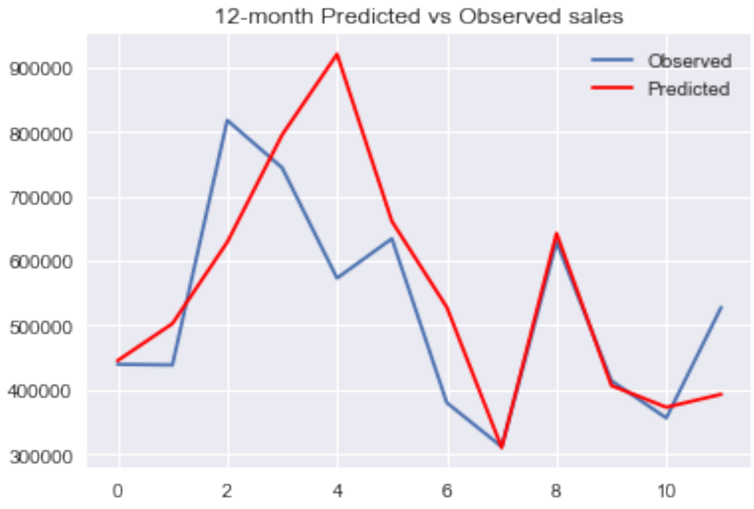


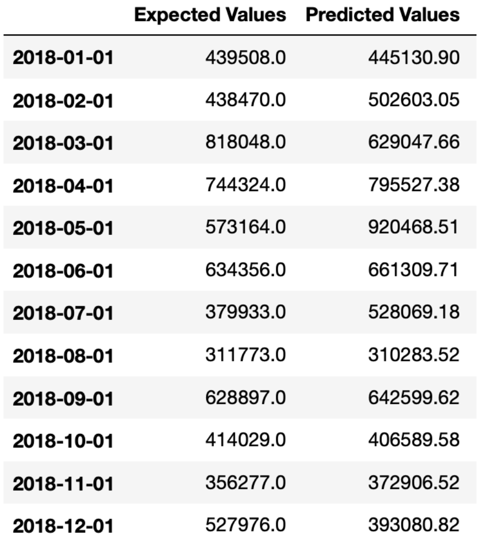


```
count    1.200000e+01
mean     4.008907e-07
std      9.947670e+04
min     -1.272738e+05
25%     -5.877507e+04
50%     -1.271136e+04
75%      4.118185e+04
max      2.047166e+05
```

Model Validation:

The trained data set with optimal hyperparameters, with a bias-correction introduced to the model it predictions compared to the fifth year of sales which was initially held-out as a validation set. The predictions are as shown below.





	Expected Values	Predicted Values
2018-01-01	439508.0	445130.90
2018-02-01	438470.0	502603.05
2018-03-01	818048.0	629047.66
2018-04-01	744324.0	795527.38
2018-05-01	573164.0	920468.51
2018-06-01	634356.0	661309.71
2018-07-01	379933.0	528069.18
2018-08-01	311773.0	310283.52
2018-09-01	628897.0	642599.62
2018-10-01	414029.0	406589.58
2018-11-01	356277.0	372906.52
2018-12-01	527976.0	393080.82

The RMSE is 130542.906 which is higher than on the trained model.