Cash Flow Forecast Challenge - Solution Development

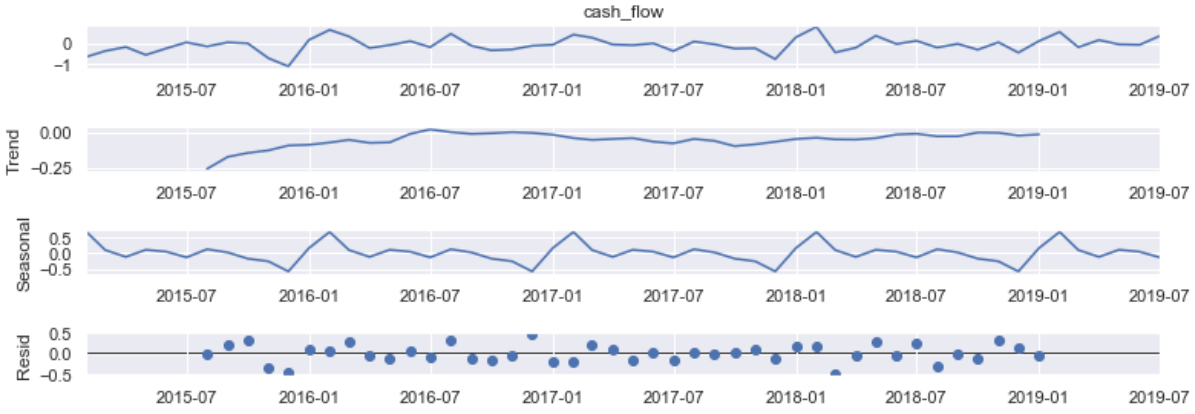
Data Preparation

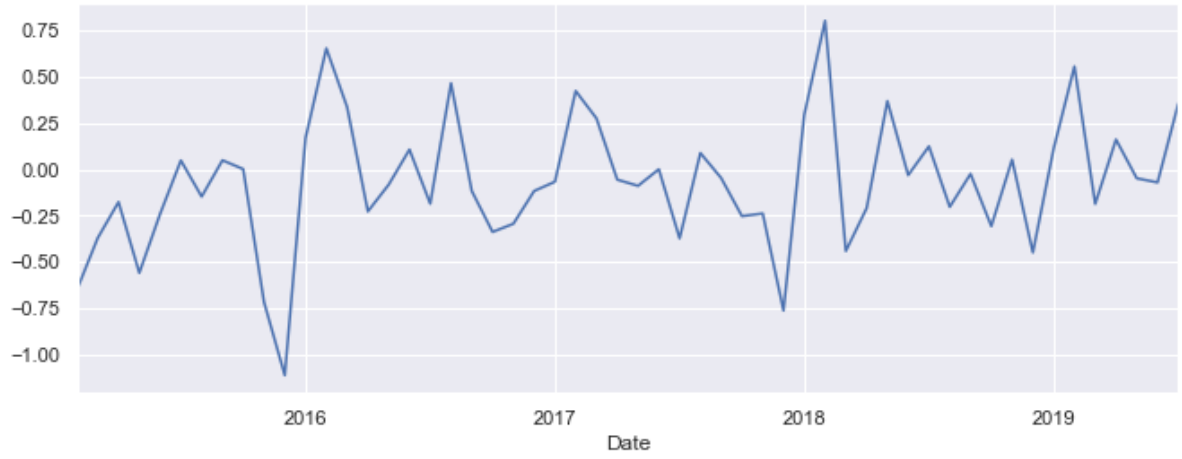
```
count_a_vals = df_train['AP Adj'].values
diffs_a = count_a_vals[:-1] - count_a_vals[1:]
count_b_vals = df_train['Cost'].values
diffs_b = count_b_vals[:-1] - count_b_vals[1:]

df_train = df_train.iloc[1:].copy().reset_index()

df_train['cash_flow'] = np.array(diffs_a)
df_train['net_cost'] = np.array(diffs_b)

df_train = df_train[['Country', 'Date', 'cash_flow', 'net_cost']]

df_train.set_index(pd.DatetimeIndex(df_train['Date']),inplace = True)
```

The target variable cash flow is calculated based on the input provided in the challenge spec.

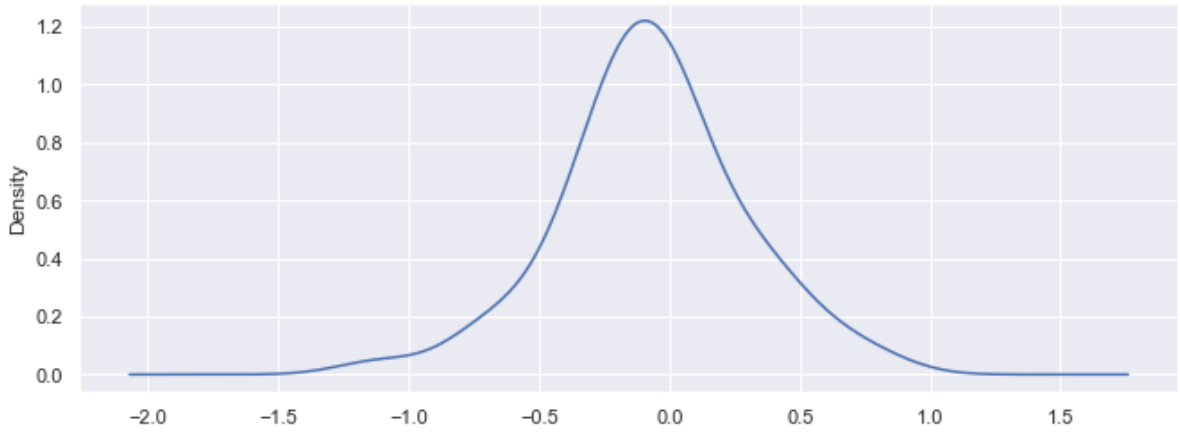
Data Analysis

Checking Trends and Seasonality for each country in the dataset.

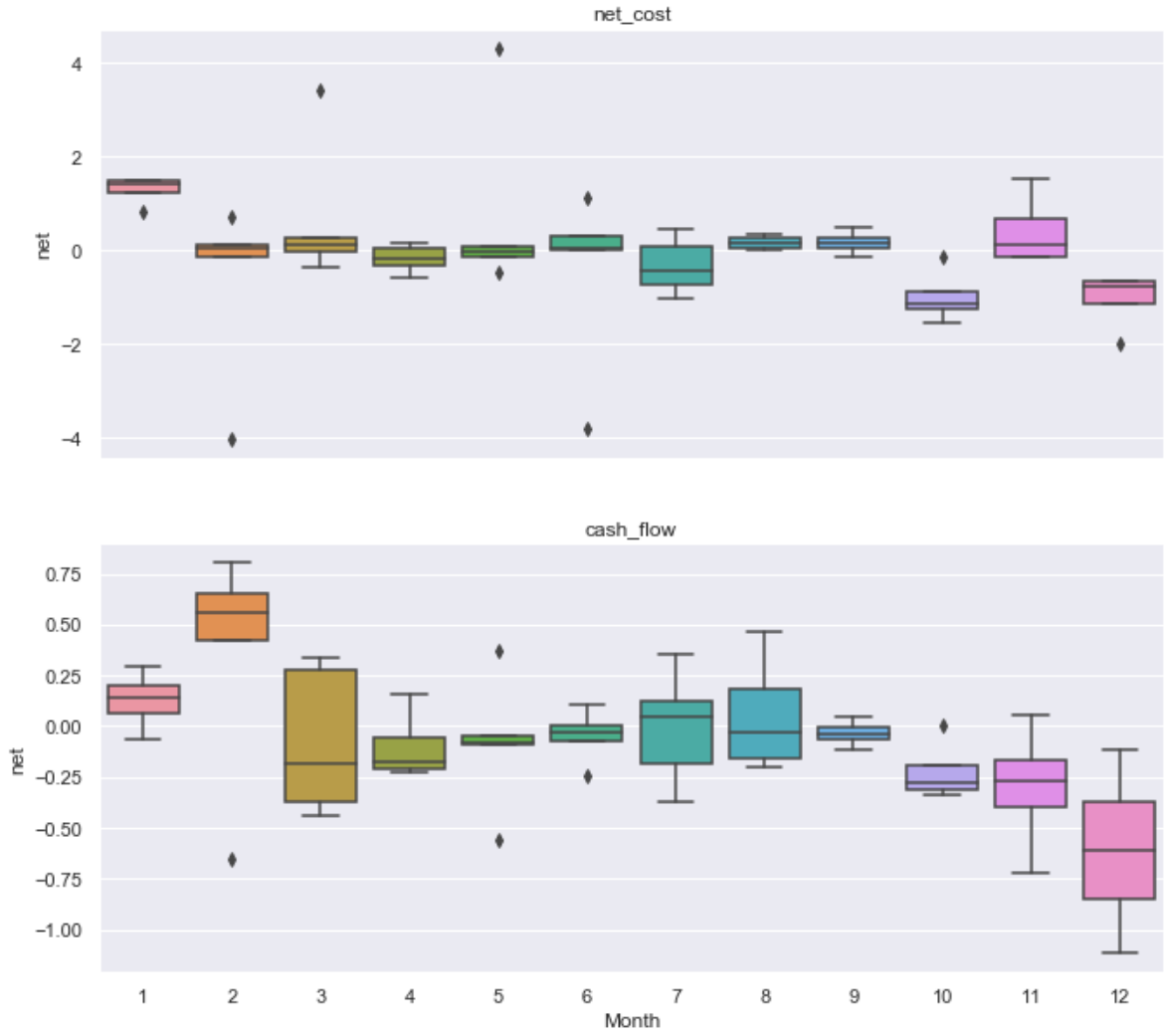




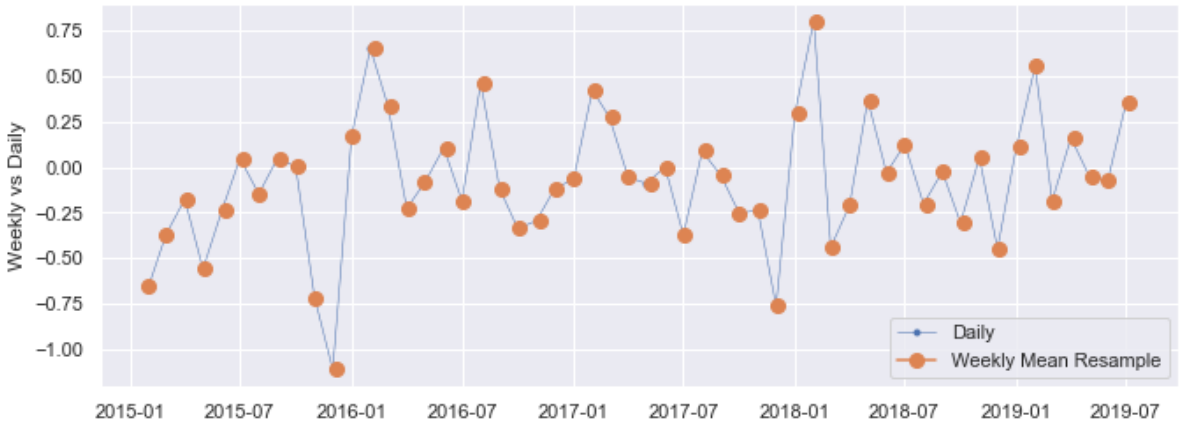
Most of the datapoints for each country was normally distribution with mean 0 and std 1.

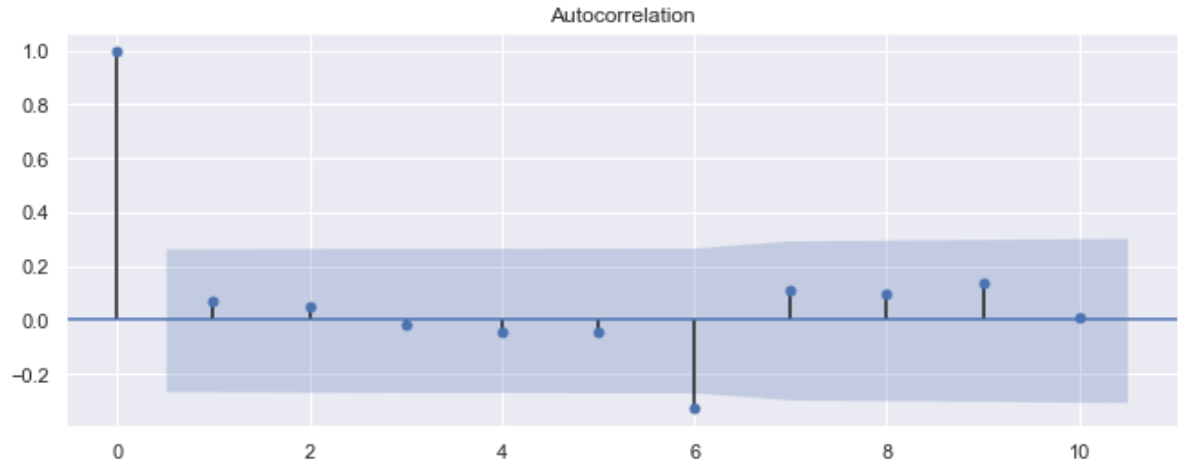


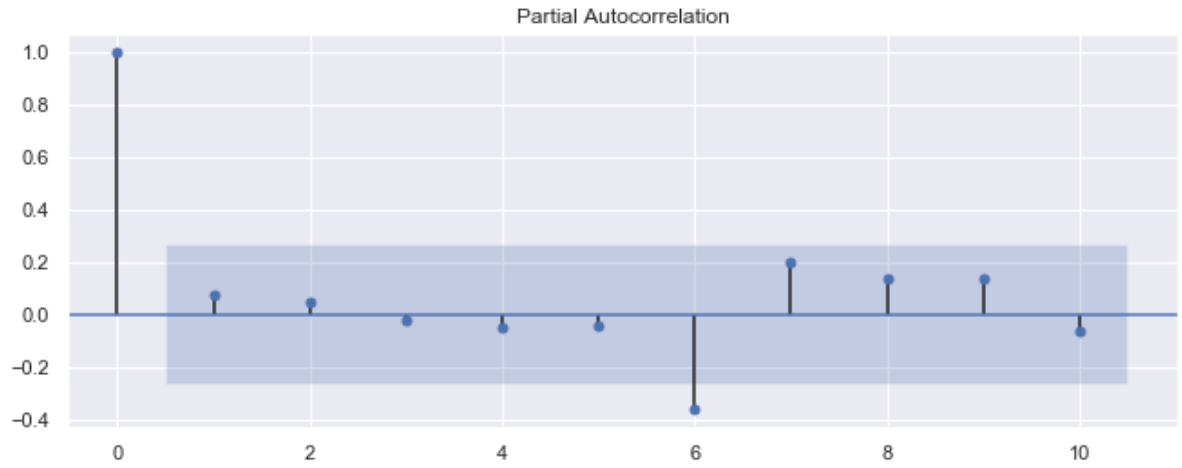
Outliers detected were smoothed based on the country by either rolling mean or double exponential smoothening.



The data generally follows the pattern of steady decrease in the start of the year and the increase in post half of the year.

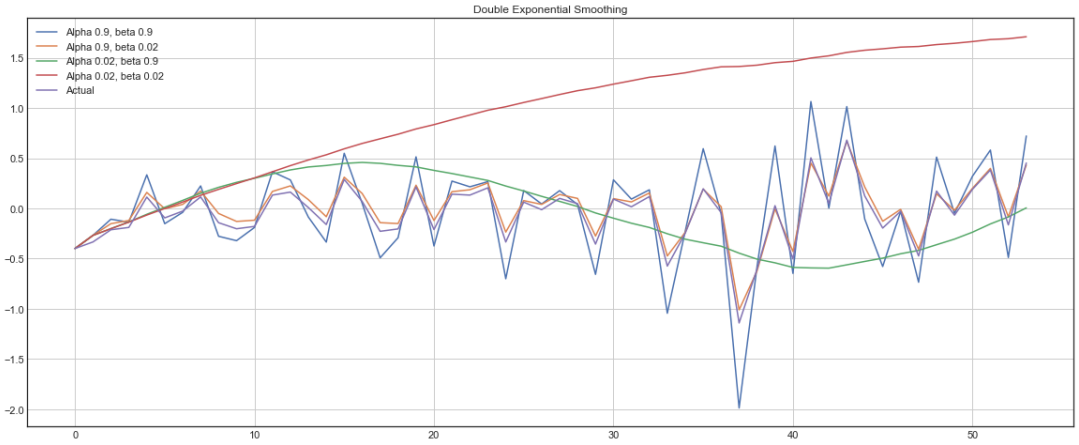




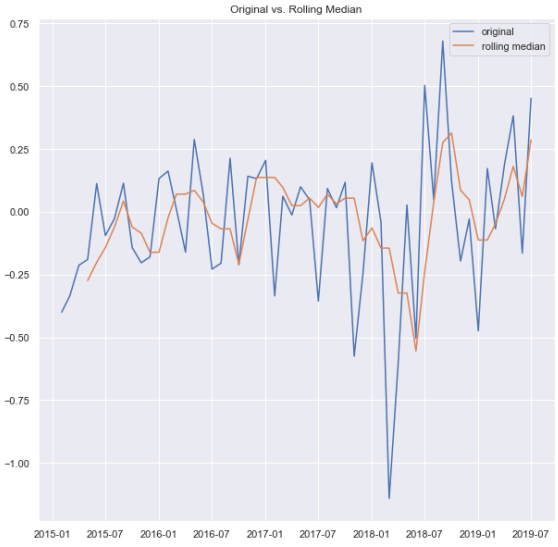


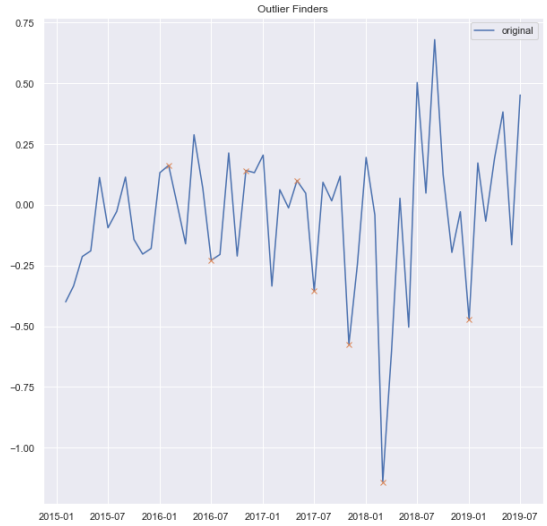
Data Smoothening

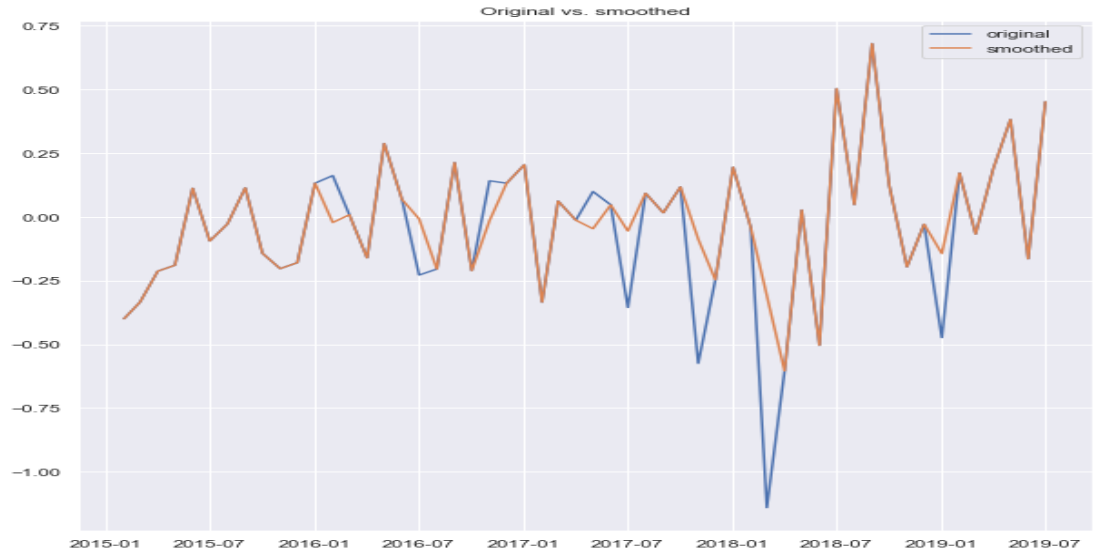
Double exponential smoothening –



Rolling mean/median







Training, Cross Validation & Test Data

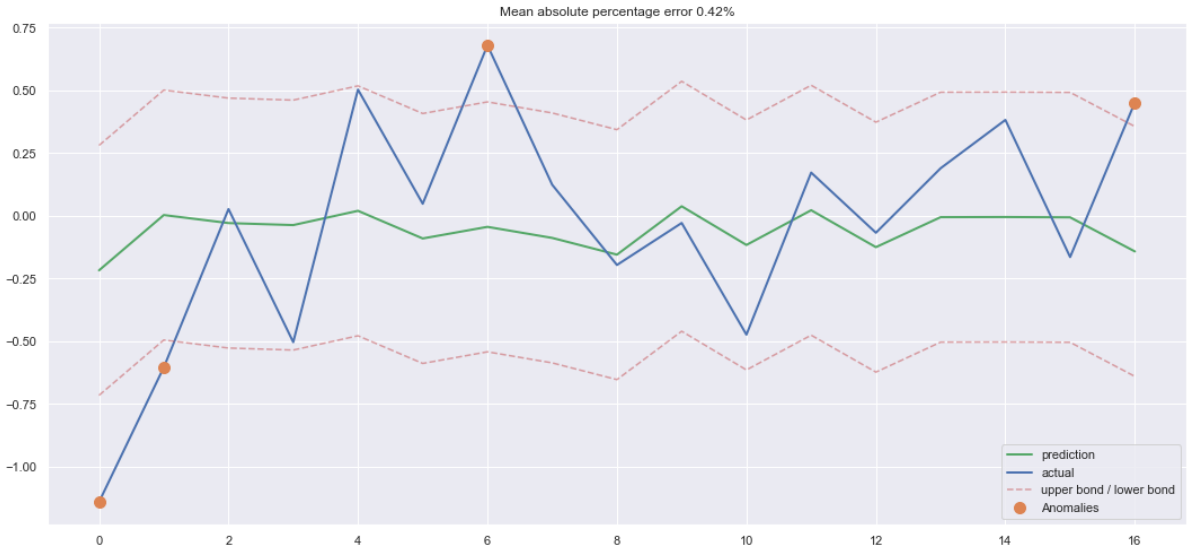
Training Data : 54 data points extracted from anonymized_train_data.csv with one less data point for calculation of cash flow.

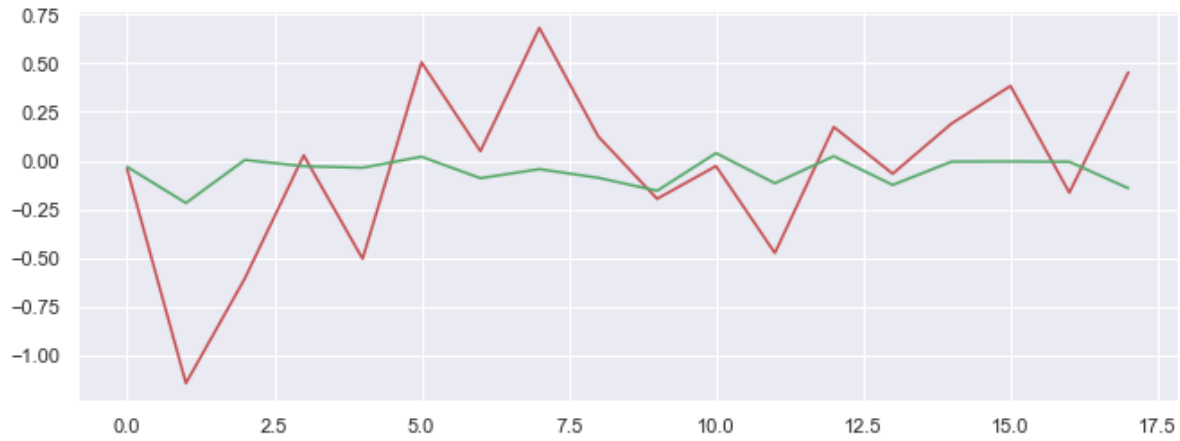
Cross - Validation Strategy/Justification : RMSE was used as a validation justification on different models trained on the last 18 data points for a given country.

Test Data : data was forecasted for next 5 months after 01-07-2019.

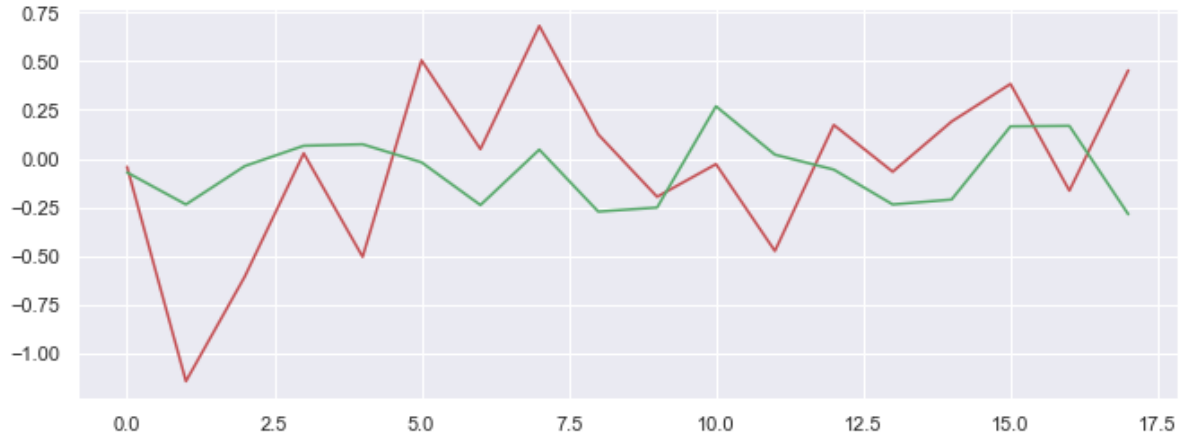
Graphs were generated for each model during model selection such as –

Linear Regression –

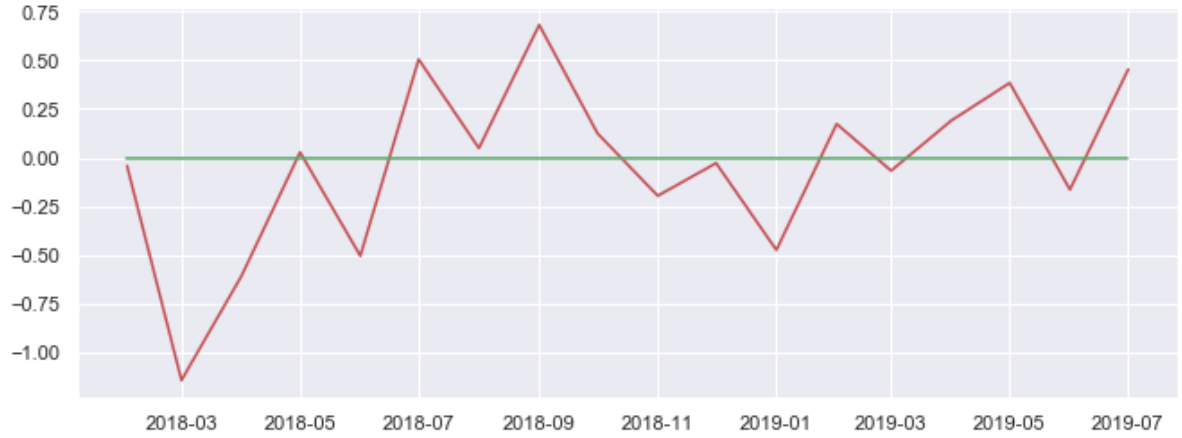




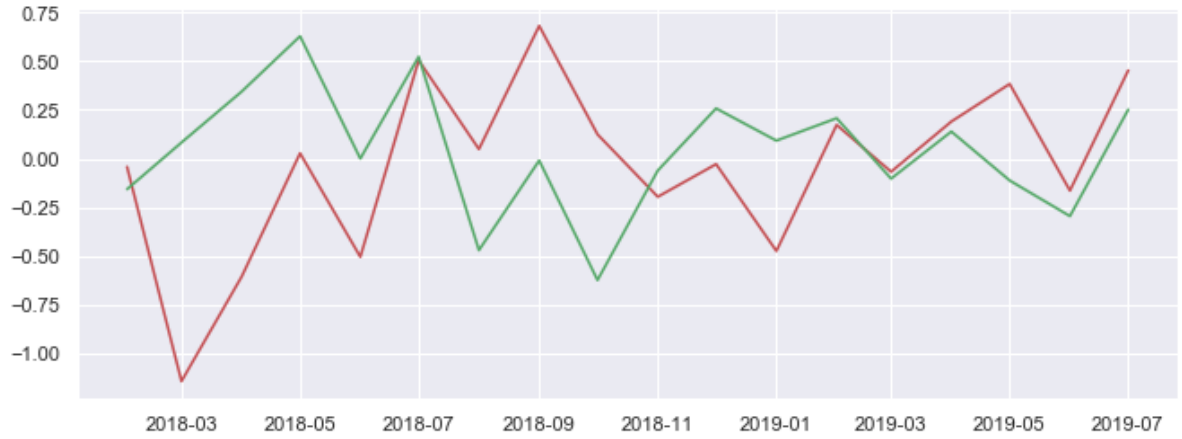
XGB



SARIMAX

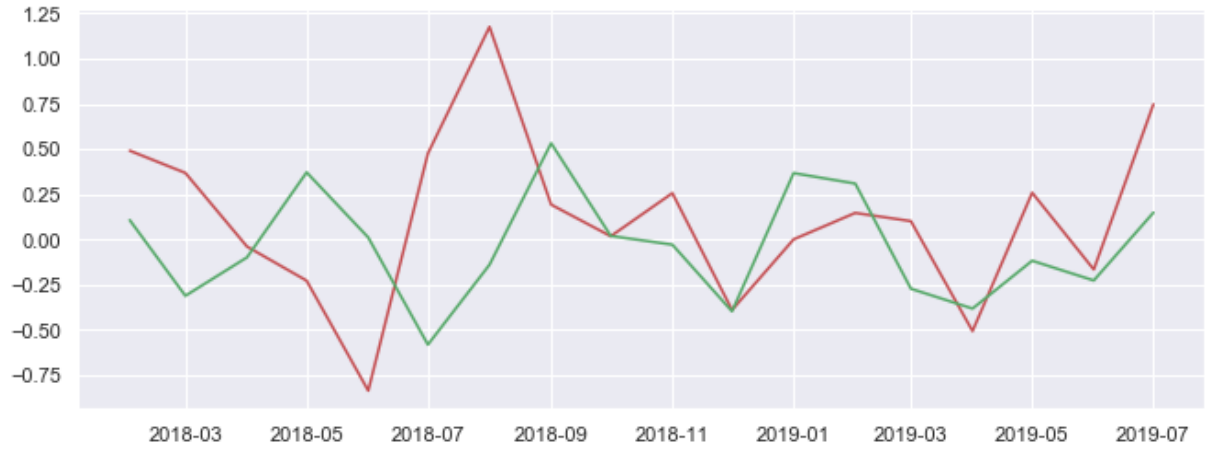


ARIMA

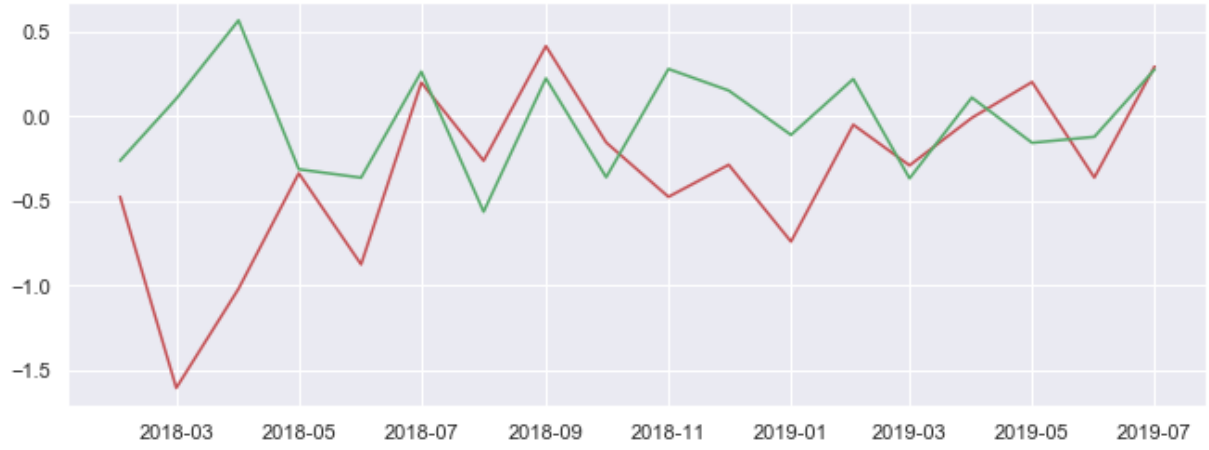


For Each country –

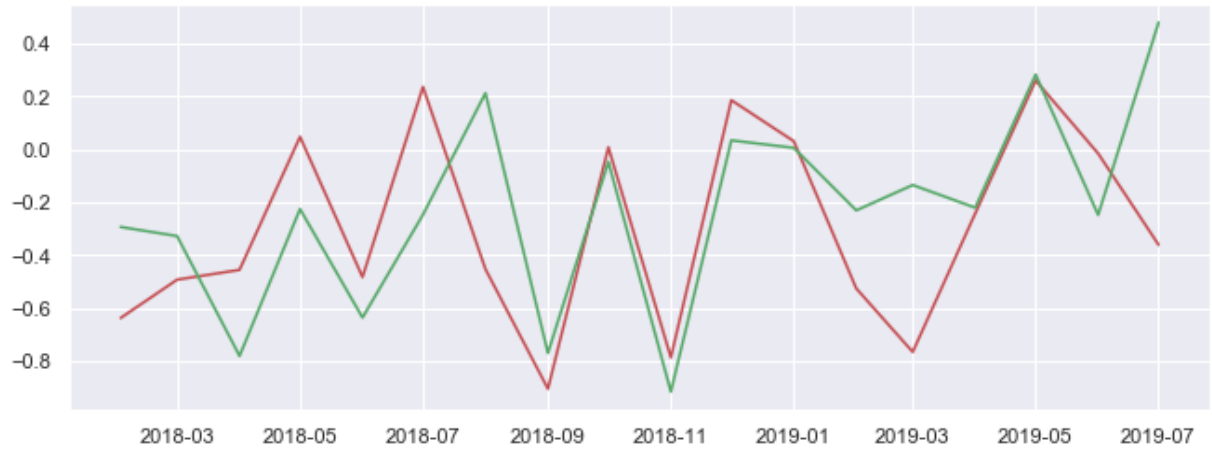
CHINA



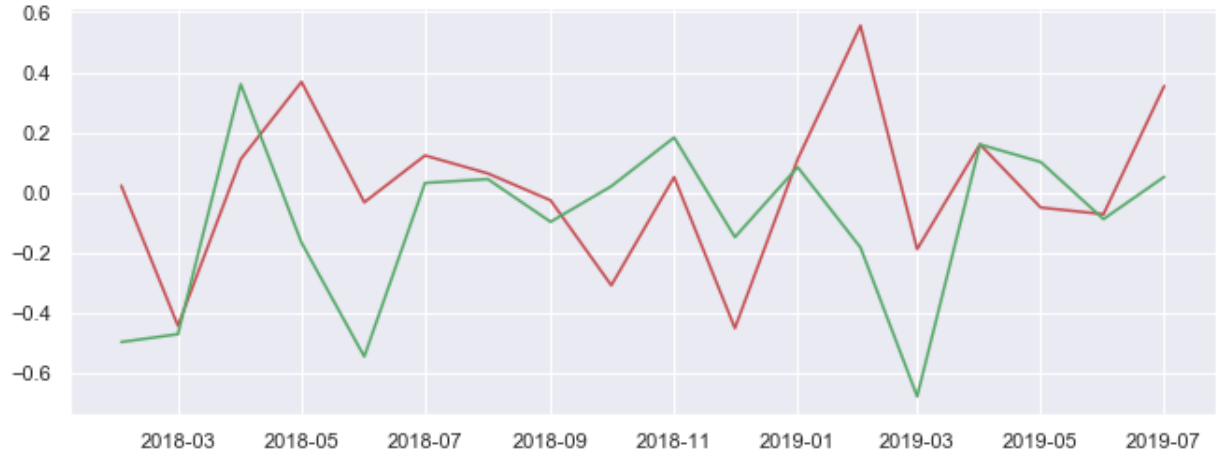
Germany



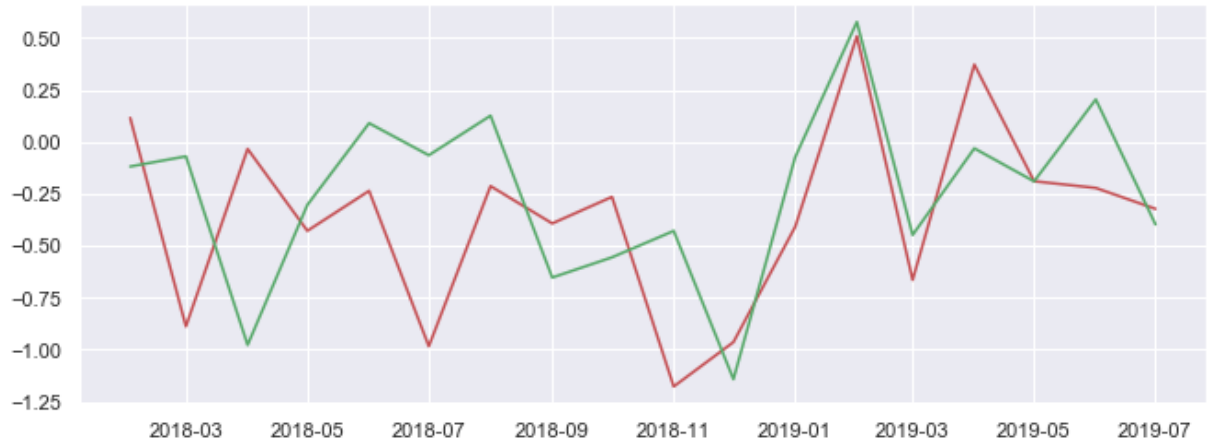
Ireland



Switzerland



USA



Forecasting Model Selection

Automated SARIMAX parameter estimation using Grid Search

```
                           SARIMAX Results
==========================================================================
Dep. Variable:              cash_flow   No. Observations:                   54
Model:                        SARIMAX   Log Likelihood                2.192
Date:                Mon, 23 Nov 2020   AIC                          -2.385
Time:                        17:36:27   BIC                          -0.414
Sample:                    02-01-2015   HQIC                         -1.627
                         - 07-01-2019
Covariance Type:                  opg
==========================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------
sigma2         0.0539      0.009      6.242      0.000       0.037       0.071
===================================================================================
Ljung-Box (Q):                       31.62   Jarque-Bera (JB):                 2.95
Prob(Q):                              0.83   Prob(JB):                         0.23
Heteroskedasticity (H):               3.95   Skew:                             0.30
Prob(H) (two-sided):                  0.01   Kurtosis:                         3.99
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Automated ARIMA parameter estimation using Grid Search

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.cash_flow   No. Observations:                   53
Model:                 ARIMA(0, 1, 1)   Log Likelihood                0.934
Method:                       css-mle   S.D. of innovations             0.229
Date:                Mon, 23 Nov 2020   AIC                            4.131
Time:                        17:36:27   BIC                           10.042
Sample:                  03-01-2015   HQIC                            6.404
                       - 07-01-2019
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              0.0041      0.002      2.069      0.039       0.000       0.008
ma.L1.D.cash_flow -1.0000      0.060    -16.770      0.000      -1.117      -0.883
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
MA.1            1.0000           +0.0000j            1.0000            0.0000
-----------------------------------------------------------------------------
```

Feature Importance

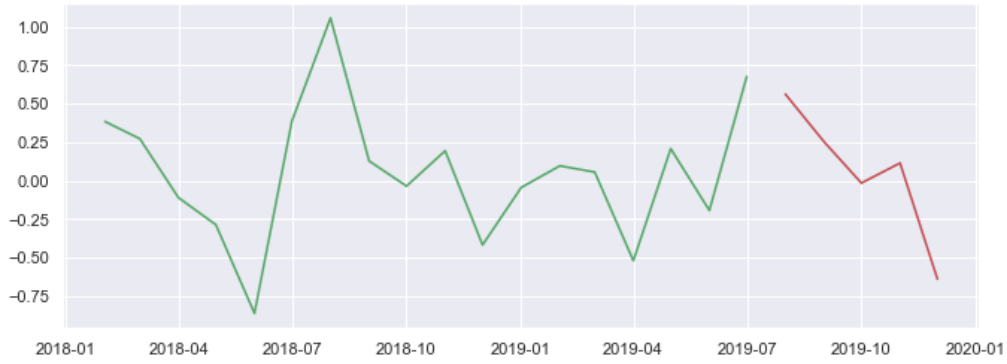
Feature importance : The “coef” column shows the weight (i.e. importance) of each feature and how each one impacts the time series.

Feature Significance : P>|z| column, informs the significance of each feature weight.

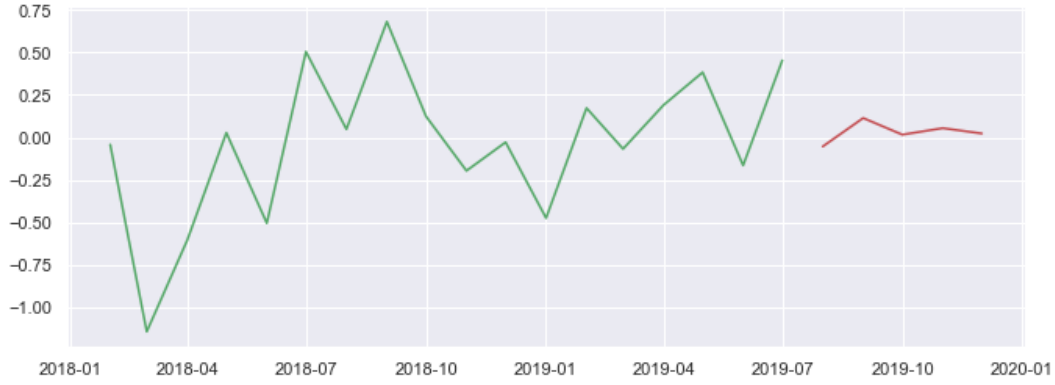
Validation : test_sheet.csv

Test_Prediction : The last five months (Aug-19 to Dec-19) in the output has been forecasted using Models for each country. The test data is never been used during parameter tuning. The forecasted Z-Score has been written in test_sheet.csv.

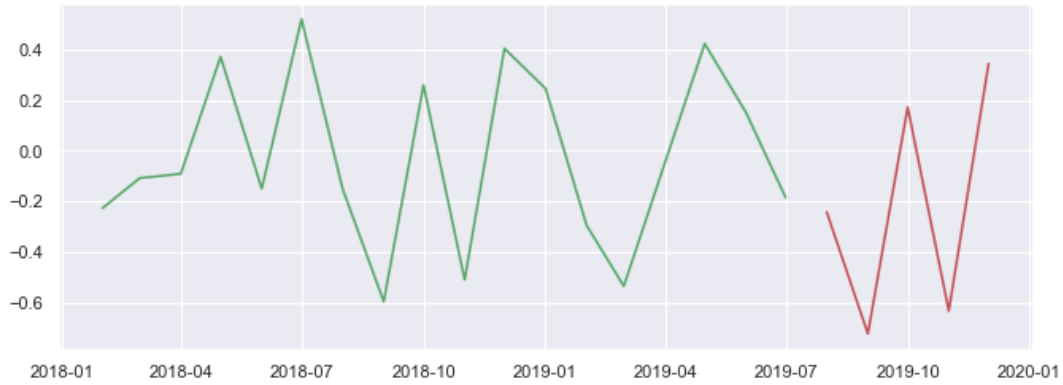
CHINA



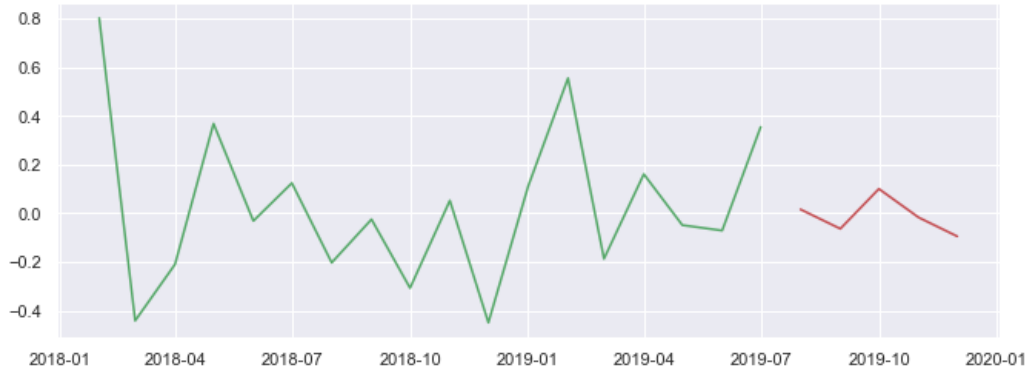
Germany



Ireland



Switzerland



USA