TSA 2022 Kaggle Competition

Aasha Reddy and Jade Forest

4/12/2022

Introduction

This RMD file showcases 4 TSA models used to participate in the TSA 2022 Kaggle Competition. These models were developed to forecast electricity demand. Exogenous temperature and humidity variables were include in select models. The code for the models is hidden, please see the RMD file for further information.

The project repository is hosted at: https://github.com/aashareddy14/ForestReddy_ENV790_TSA_Competition_S2022.

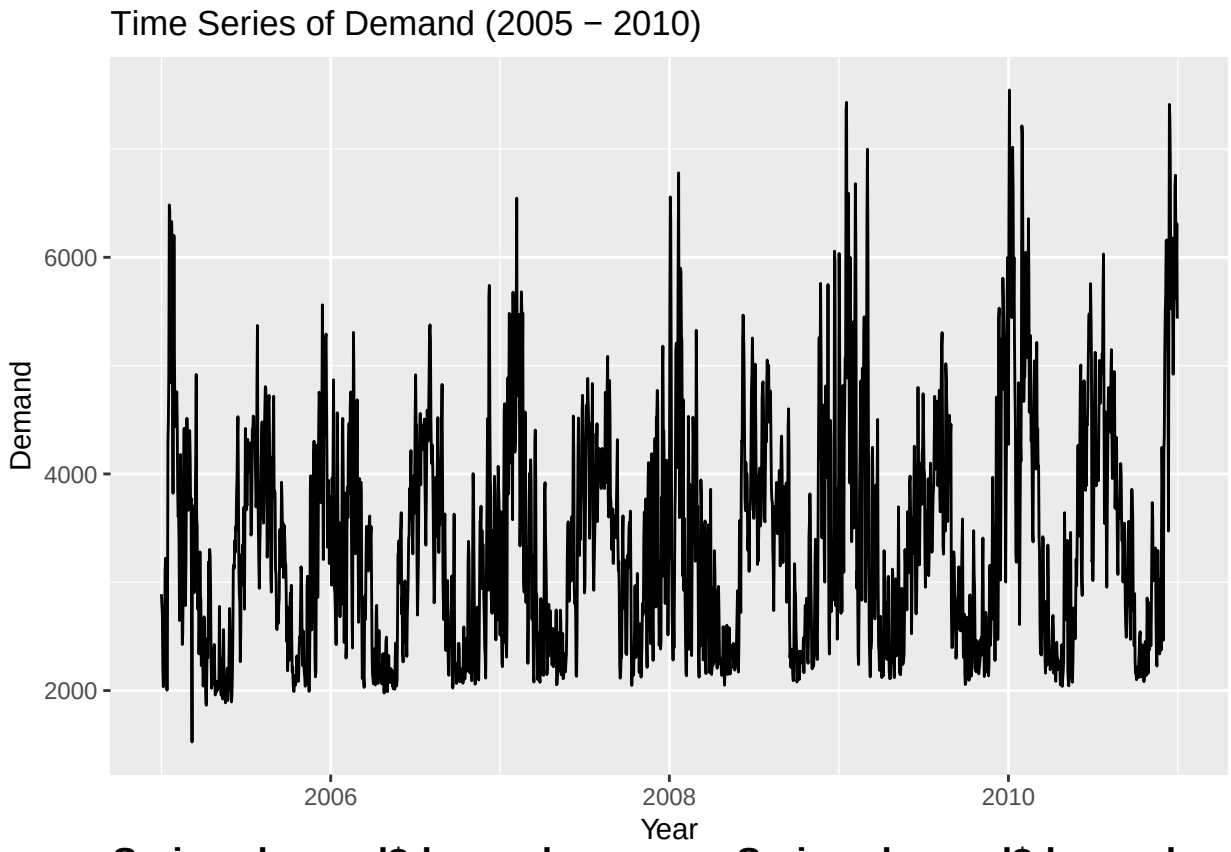
Data Cleaning

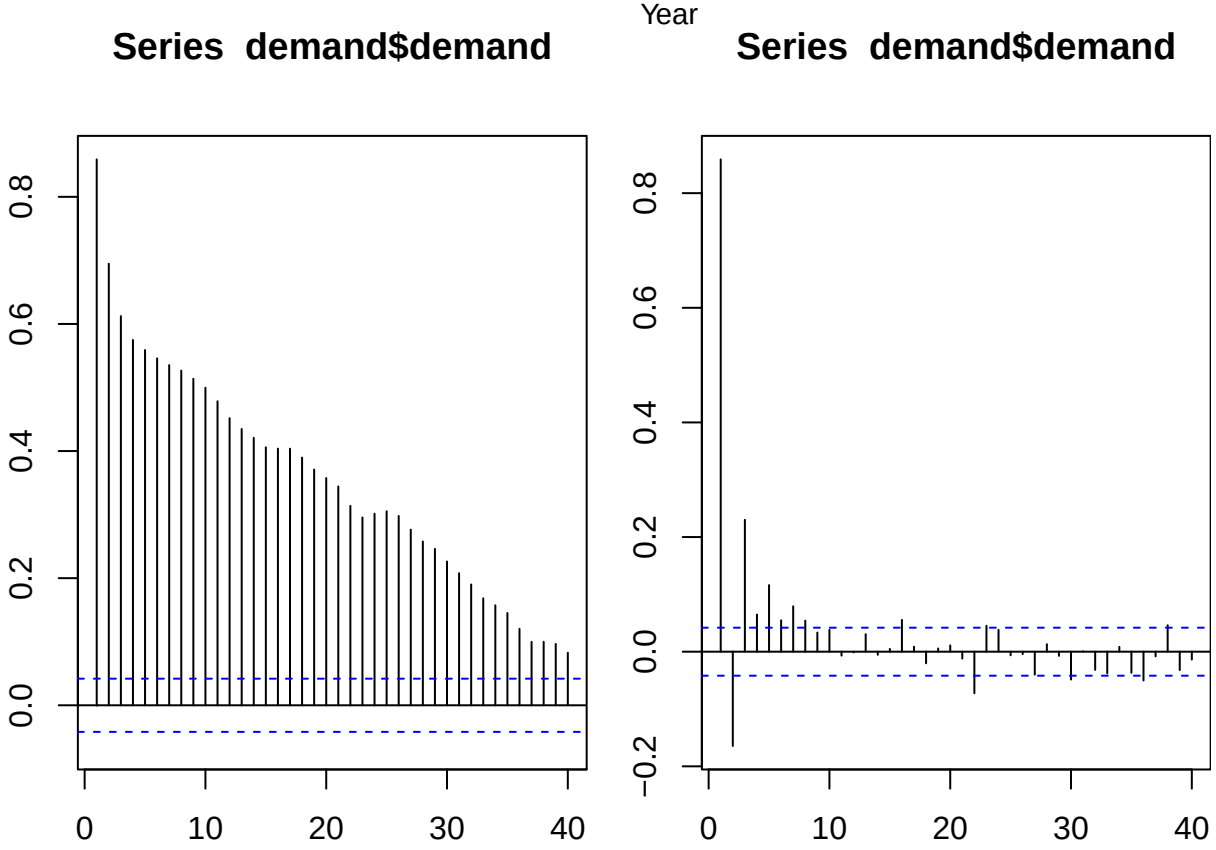
The demand data was examined first. All rows after 2191 were NA and were removed. The meter id column was removed as every row contained a value of 1. The hour columns were averaged to transform hourly data into total daily demand. Lubridate was used to transform the date variable into a datetime object. In the final data set, there are no NA values. In examining the humidity and temperature data sets, it was also noted that there is no missing data. This data was transformed from hourly to daily as well to match the demand data set. The top 6 rows are displayed for reference.

```
## # A tibble: 6 x 4
##   date       demand avg_temp avg_humidity
##   <date>      <dbl>    <dbl>        <dbl>
## 1 2005-01-01  2889.     53.6         76.7
## 2 2005-01-02  2789.     53.8         80.5
## 3 2005-01-03  2708.     55.9         81.2
## 4 2005-01-04  2212.     61.7         74.8
## 5 2005-01-05  2035.     60.4         76.1
## 6 2005-01-06  2110.     62.0         78.0
```

Exploratory Data Analysis

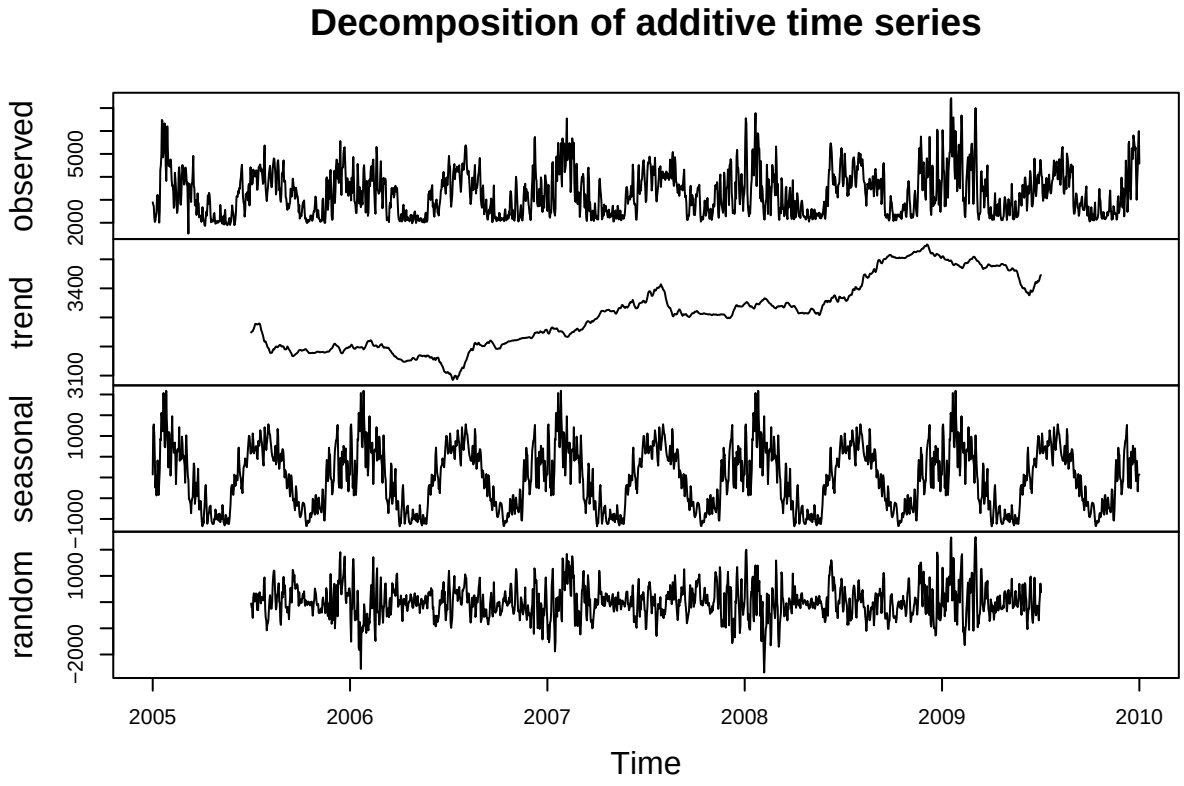
To begin, exploratory analysis was conducted on the demand data. An increasing seasonal trend is visible. A small scalloping pattern is visible in the ACF plot.


Time Series of Demand (2005 − 2010)
Demand
6000
4000
2000
2006
2008
2010
Year

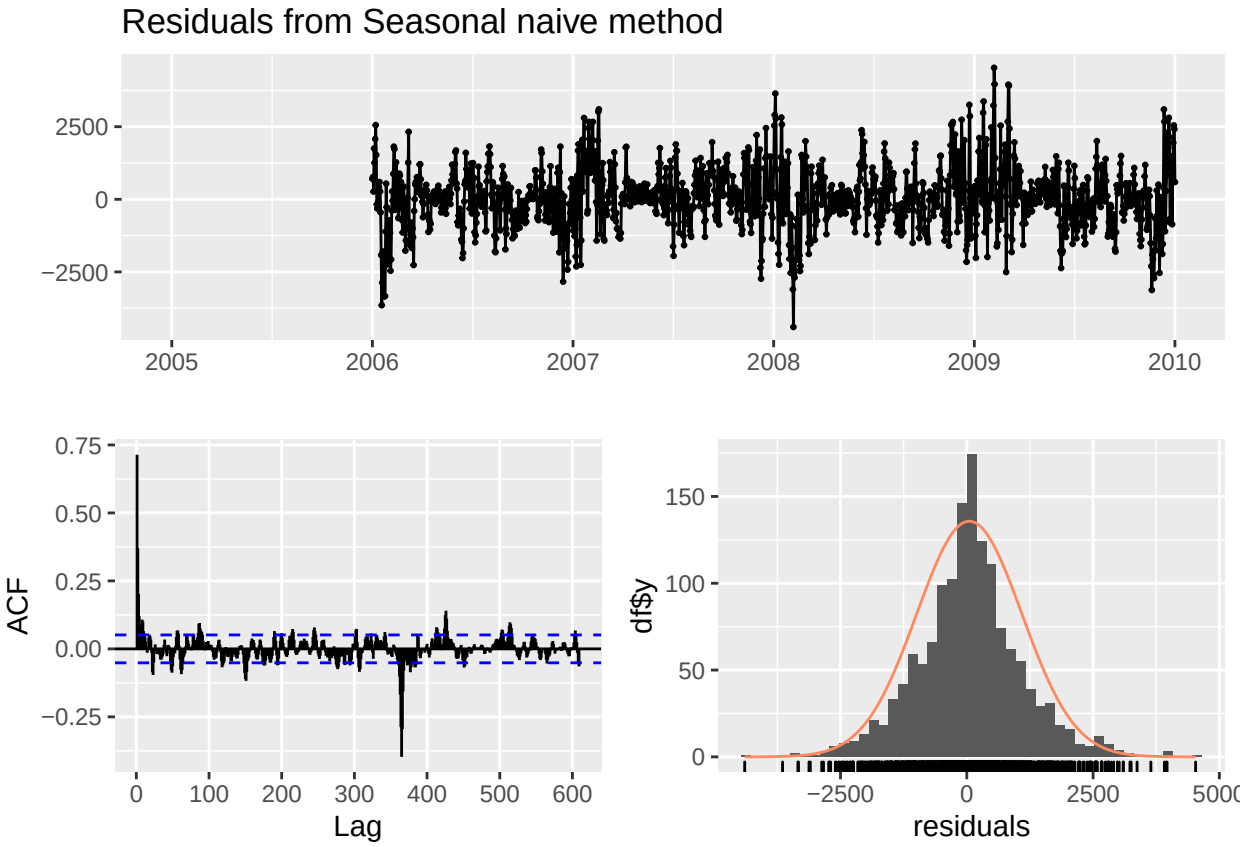

Series demand$demand
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Series demand$demand
−0.2
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40

Methods and Modeling

In order to assess the models, the final year of the data set, 2010, was chosen as the holdout period. Once the data was filtered by time it was transfored into a time series object. The models were trained on data from 2005 - 2009, and then evaluated with the 2010 data. The final model was selected by identifying the highest MAPE. Once the model was identified, the it was retrained with the entire data set. Finally, the best retrained model was used to forecast Jan 2011 demand. The decomposed time series is plotted below.



Model 1 - Seasonal Naive Model

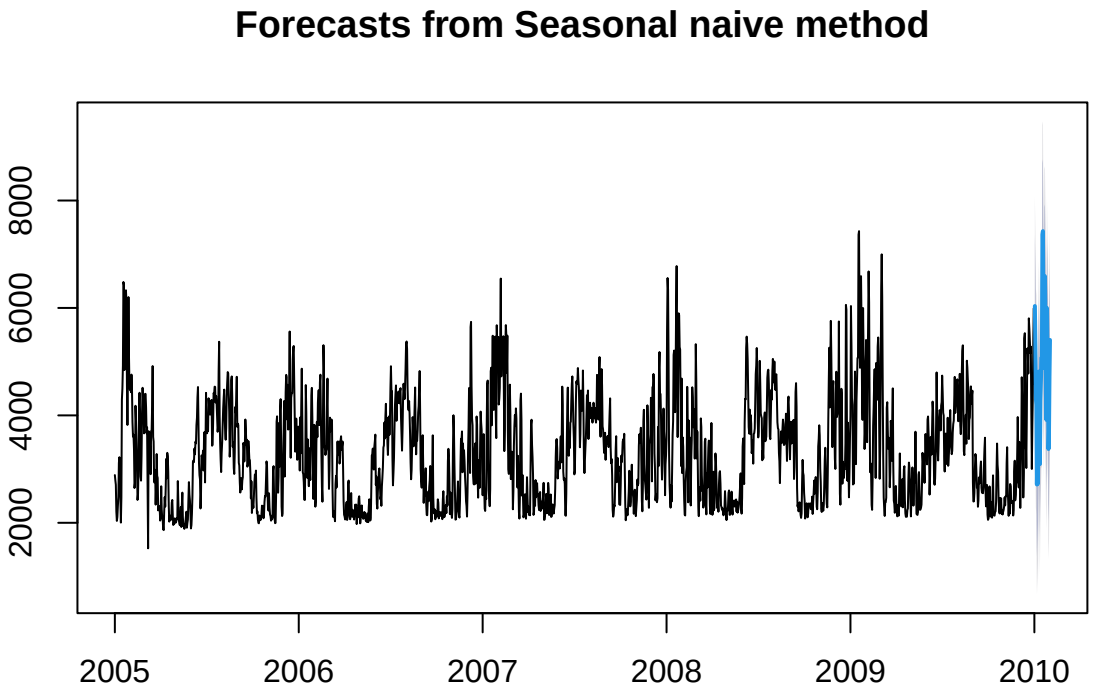



```
##
##  Ljung-Box test
##
## data:  Residuals from Seasonal naive method
## Q* = 2503.7, df = 365, p-value < 2.2e-16
##
## Model df: 0.   Total lags used: 365
```

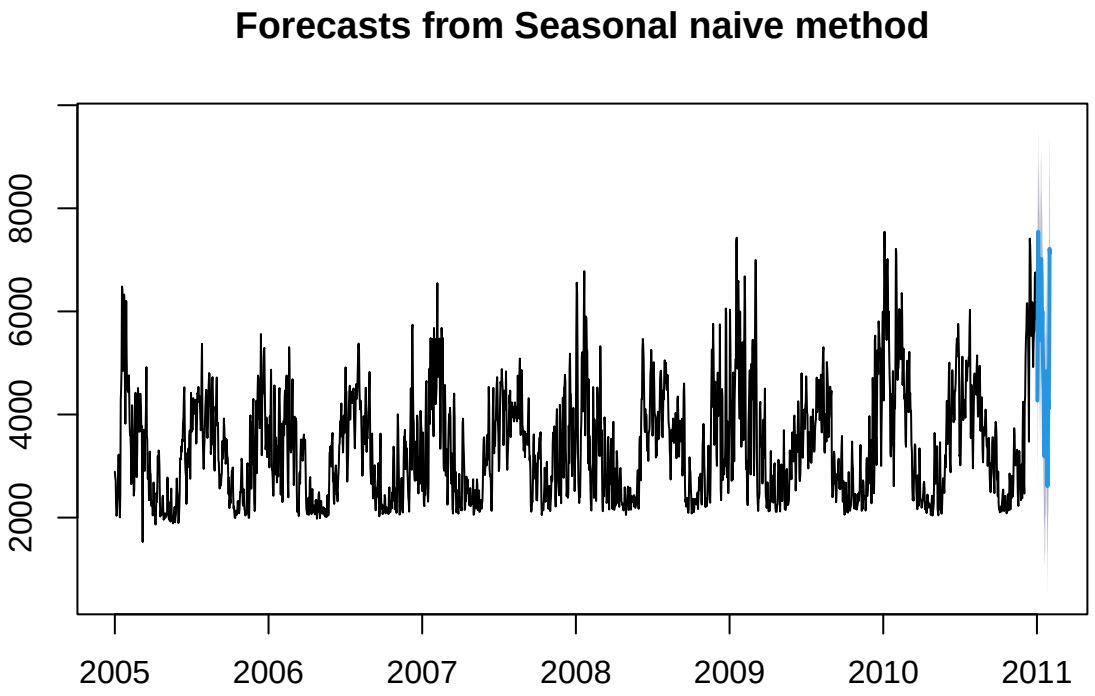
The seasonal naive model residuals are normally distributed. However, there seems to be a pattern and significant correlation in the ACF.

First, demand is forecasted for January 2010, seen in the plot below. The training MAPE is 23.05 while the test MAPE (on January 2010) is 42.36. The MAPE is a lot higher for the test data, meaning the seasonal naive model is likely overfitting to the training data.

```
##                      ME    RMSE       MAE       MPE     MAPE     MASE      ACF1
## Training set  49.31082 1042.46  769.2879 -2.942721 23.05247 1.907120 0.7138548
## Test set     323.78763 2284.72 2003.9946 -5.686023 42.36300 4.968046        NA
```

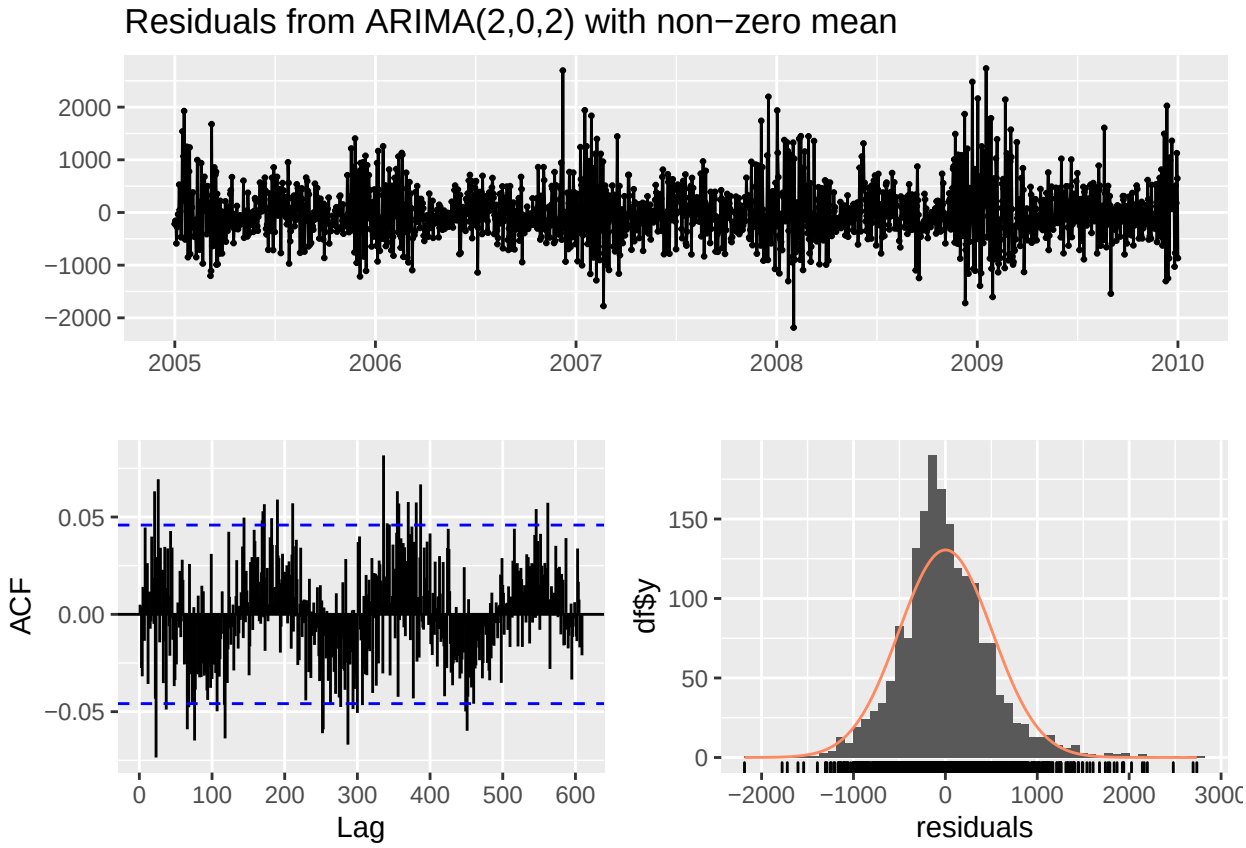


Theseasonal naive model was refit using all data from 2005 - 2010, and then used to forecast for January 2011, shown below



Model 2 - Seasonal ARIMA model (Autofit)

First, the mode was fit using the training data, from 2005 - 2009, and using the auto arima function. The auto arima chooses an ARIMA(2, 0, 2) model.

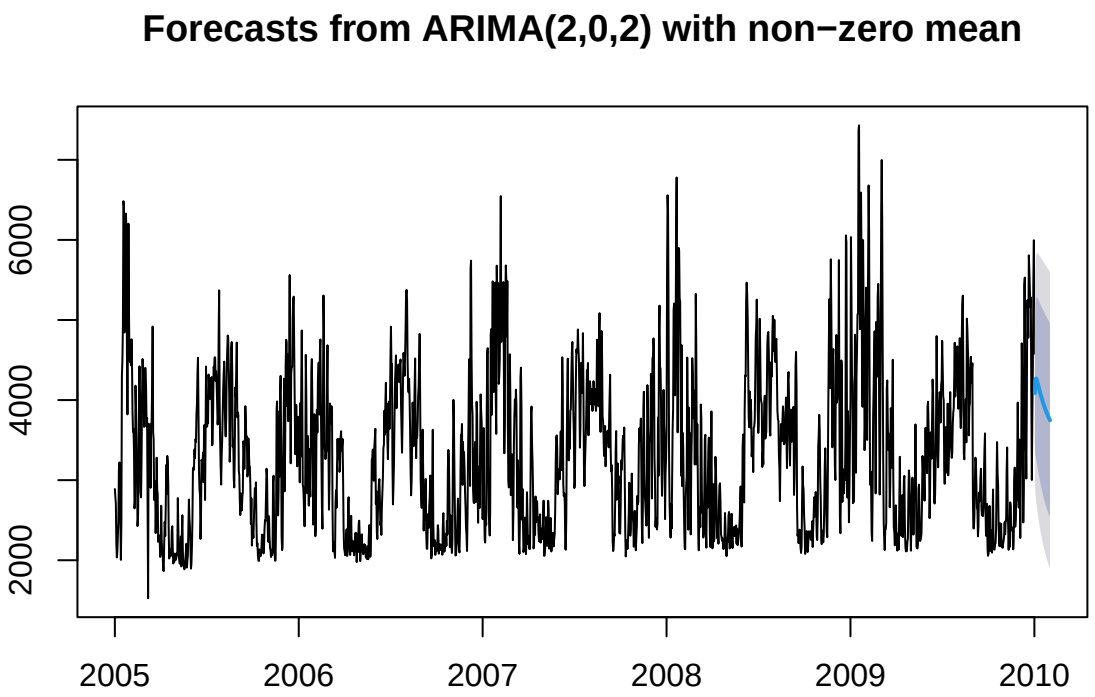



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,0,2) with non-zero mean
## Q* = 515.18, df = 360, p-value = 1.39e-07
##
## Model df: 5.   Total lags used: 365
```

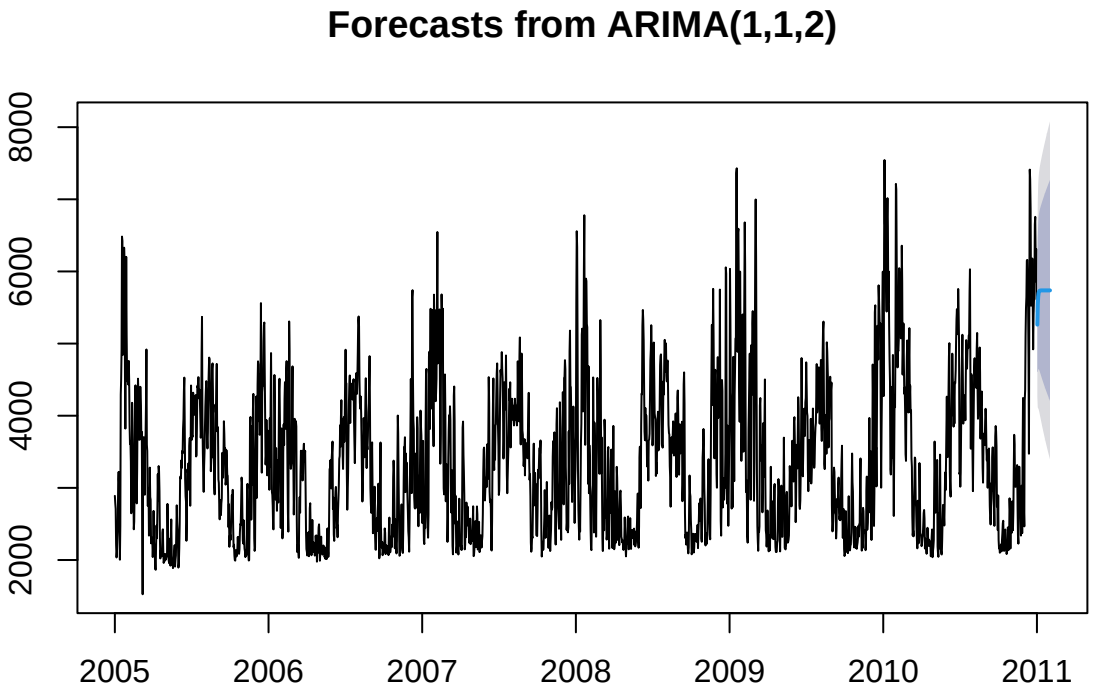
The residuals here are normally distributed but do not seem to be that random. ACF shows significant correlations as well - this means that this likely will not be the best model!

Next, daily demand was forecasted for January 2010, shown in the plot below. The MAPE on the test set (Januray 2010) is 25.1363 here. Note that this is a large improvement from the seasonal naive model.

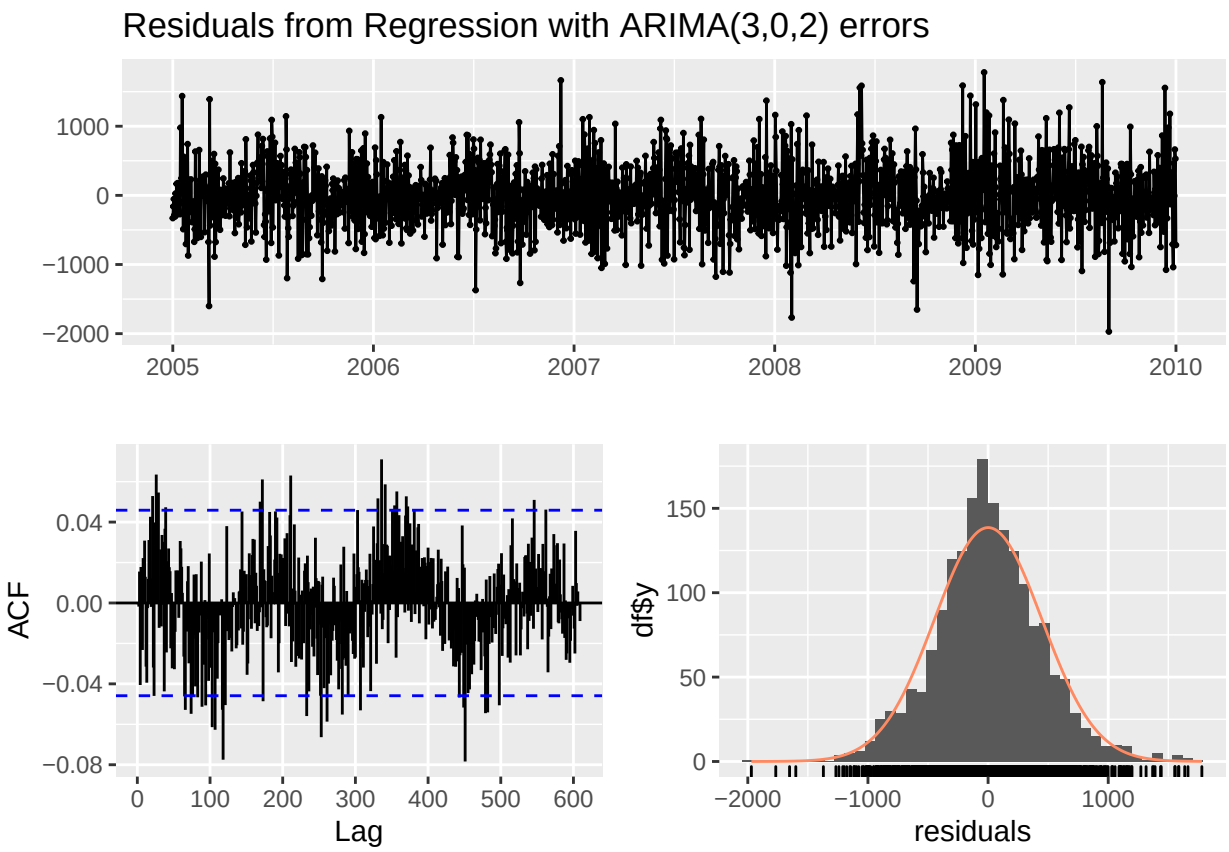
```
##                ME     RMSE      MAE      MPE     MAPE
## Test set 1134.947 1742.689 1419.366 15.95663 25.13631
```



The model was refit using all data from 2005 - 2010, and then used that to forecast for January 2011, shown below.



Model 3 - auto.arima with temperature as exogenous variable

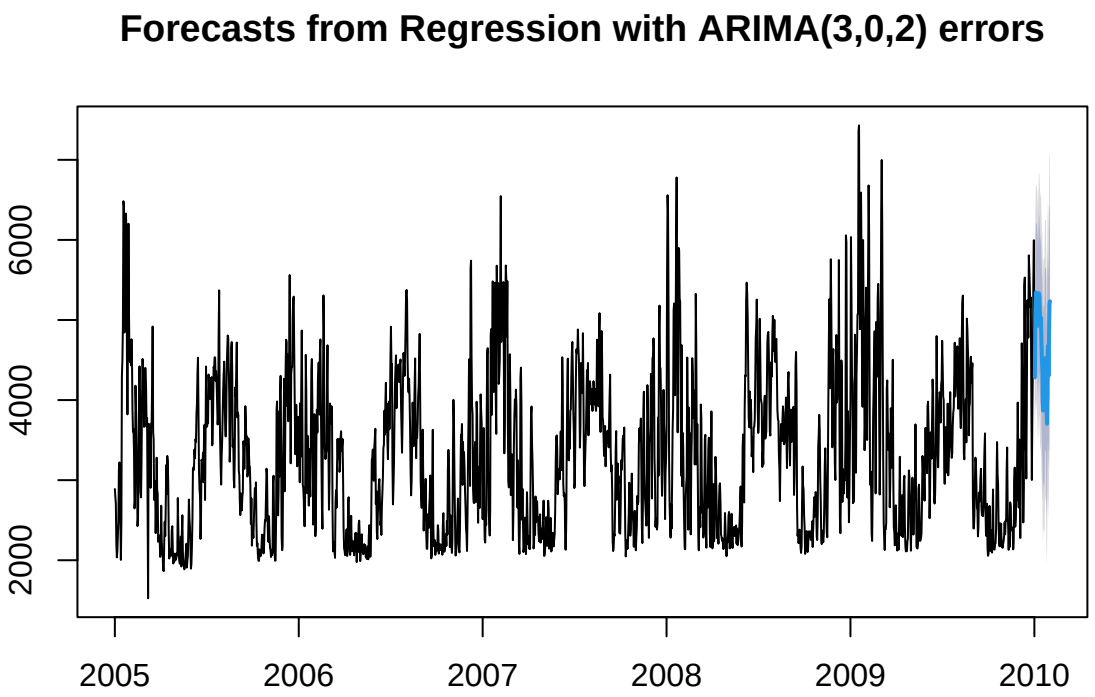



```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(3,0,2) errors
## Q* = 494.24, df = 358, p-value = 2.285e-06
##
## Model df: 7.   Total lags used: 365
```

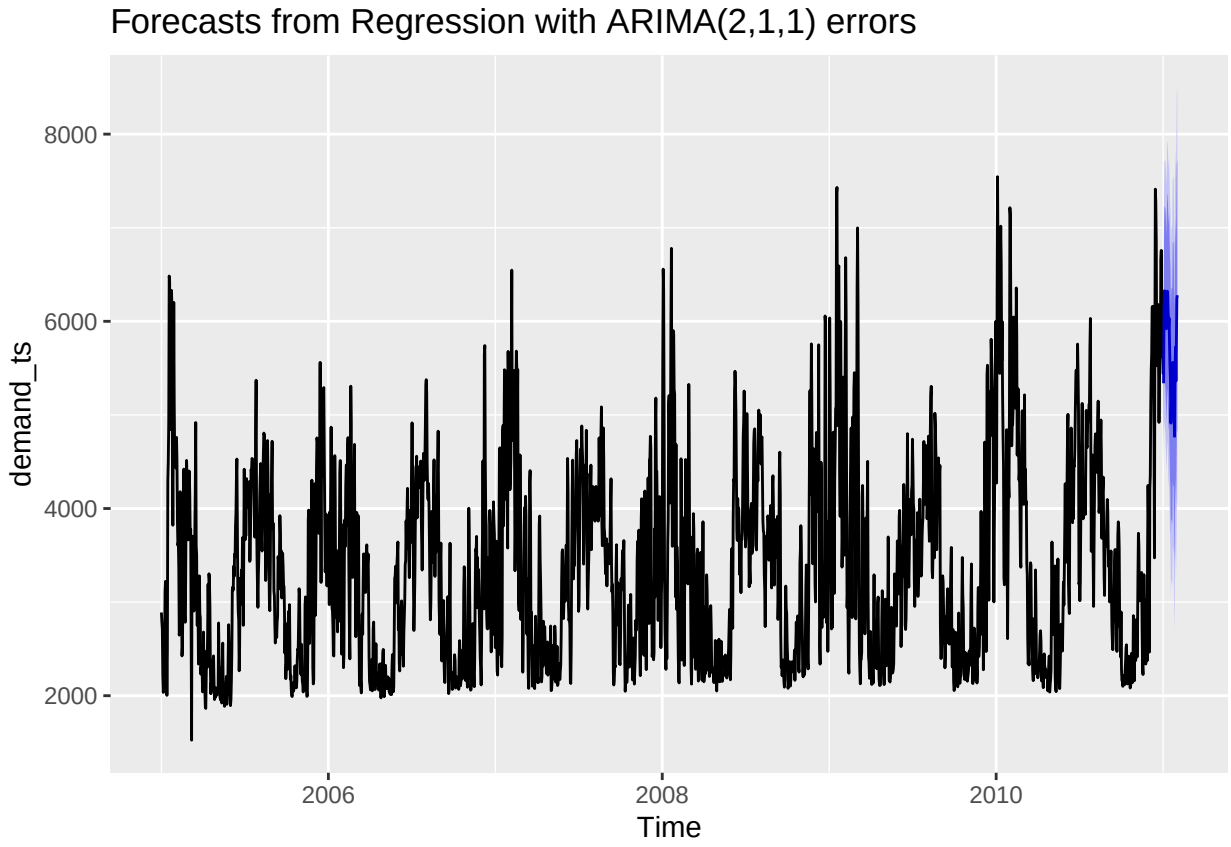
The residuals here are normally distributed. However, the ACF shows significant correlations, similar to the original autofit SARIMA.

Regardless, demand for January 2010 was forecast. The MAPE on the test set (Januray 2010) is 16.26, which is notably smaller that that of the SARIMA(1,2,2) model without temperature as an exogenous variable. The forecast is plotted below.

```
##                ME     RMSE      MAE      MPE     MAPE
## Test set 442.364 1040.276 857.9651 3.742301 16.26123
```

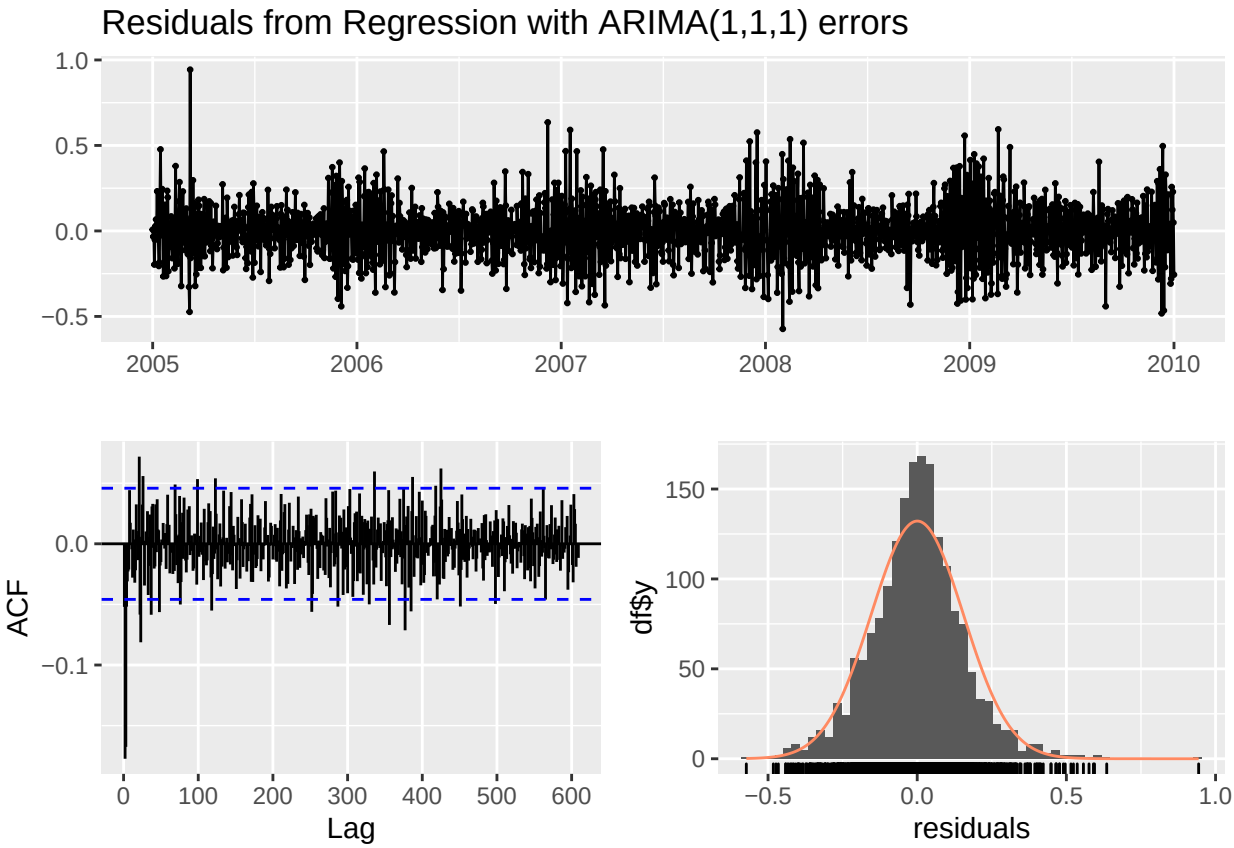


Next, the model was refit using all data from 2005 - 2010, and then use to forecast for January 2011. The January 2011 temperature is assumed to be the same as January 2010 temperature.

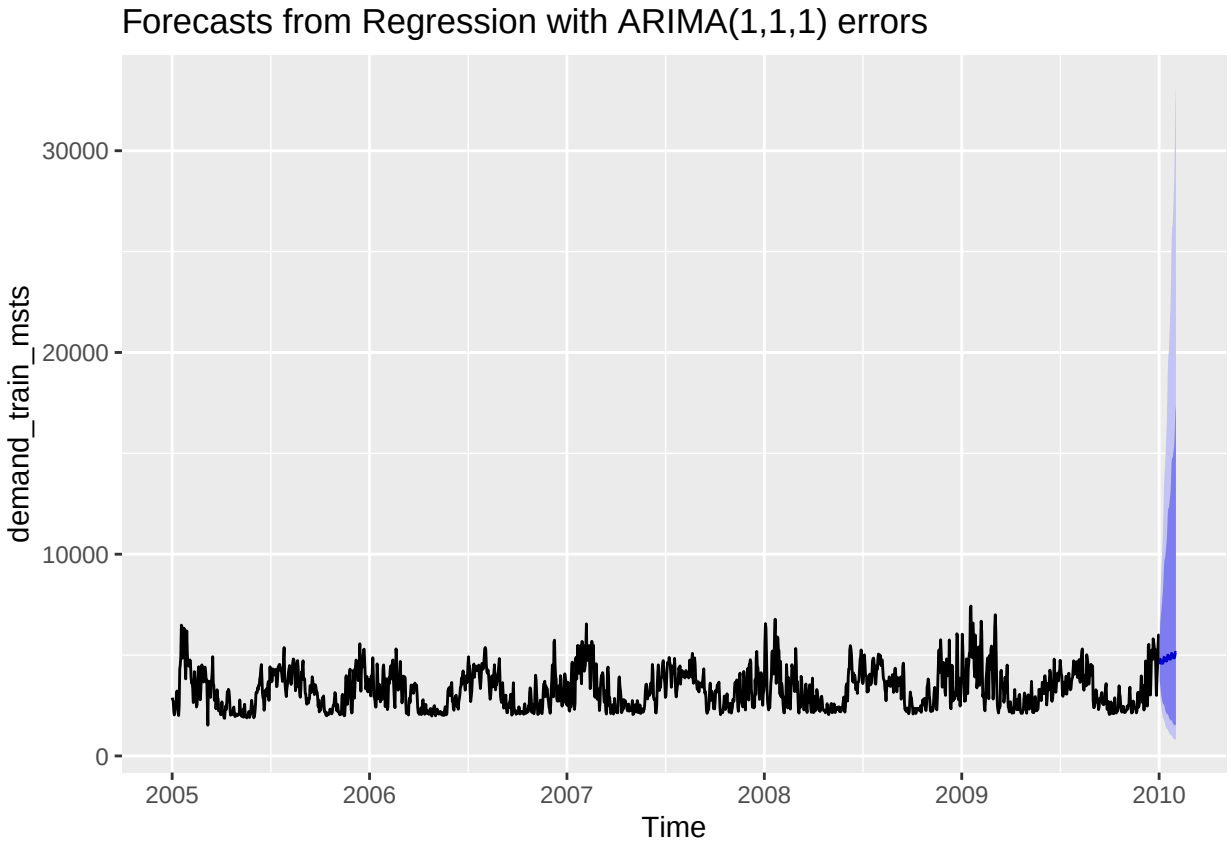


Model 4 - ARIMA + FOURIER

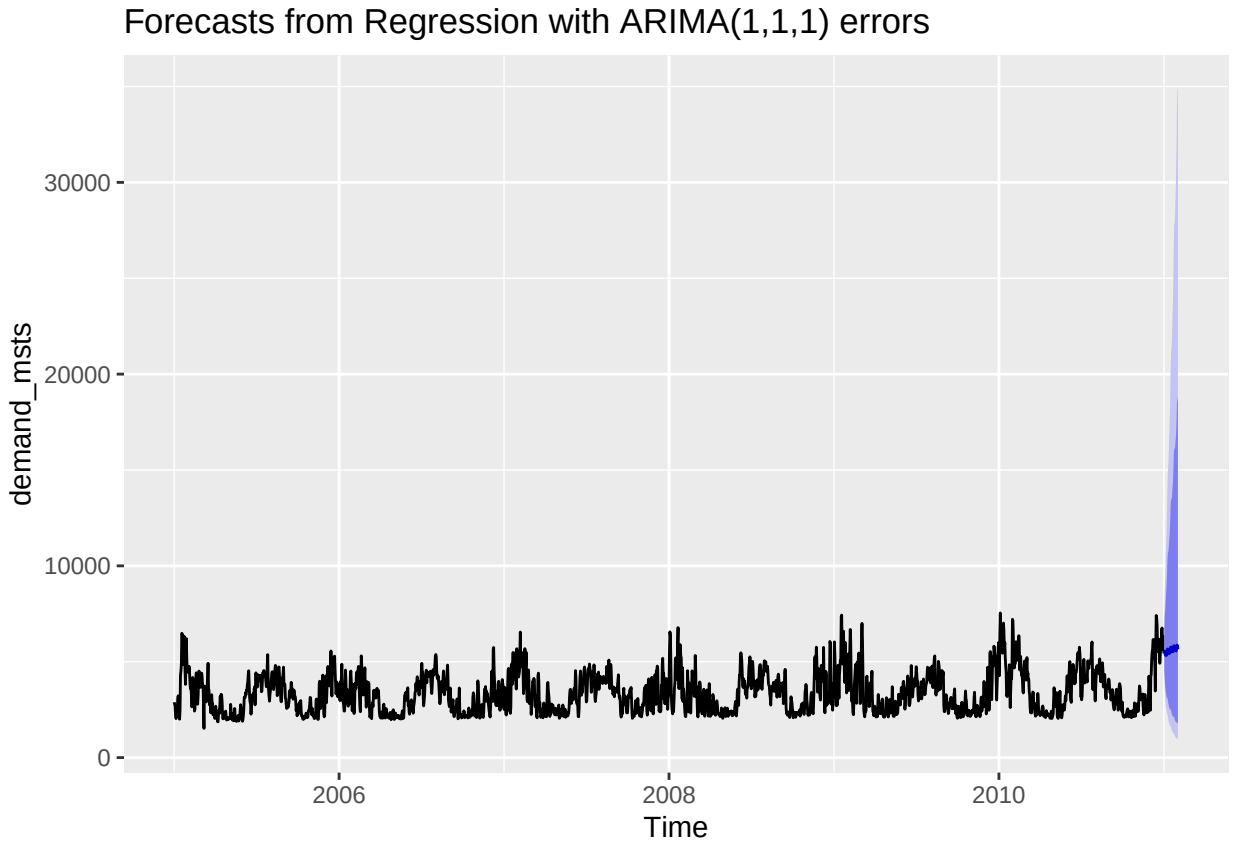
Next an auto fit arima with seasonal fourier terms to account for multiple seasonality present in the data was used. The multiple seasonality comes from the the repeated weekly variation and seasonal monthly variation throughout the year.




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(1,1,1) errors
## Q* = 515.22, df = 351, p-value = 2.434e-08
##
## Model df: 14.   Total lags used: 365
```

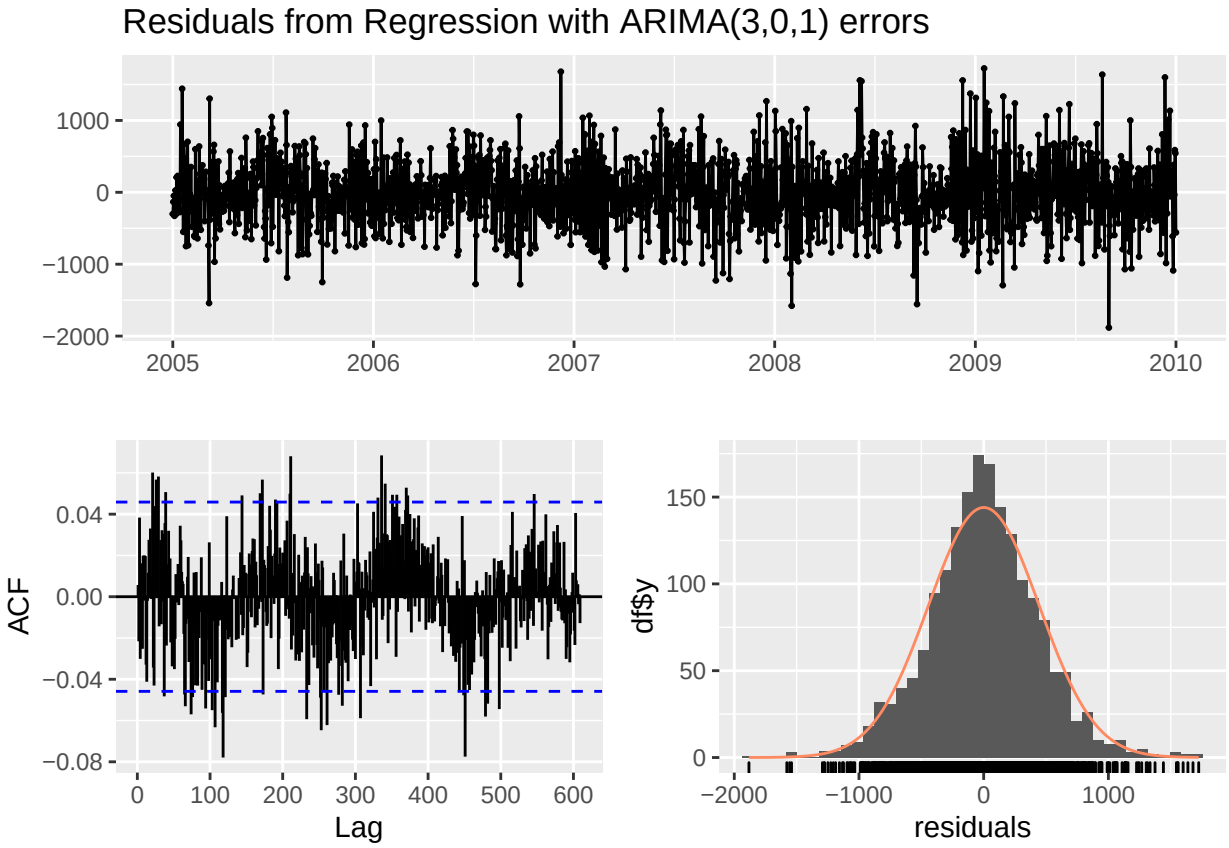


Moving on, the model was refit using all data from 2005 - 2010, and then used to forecast for January 2011.



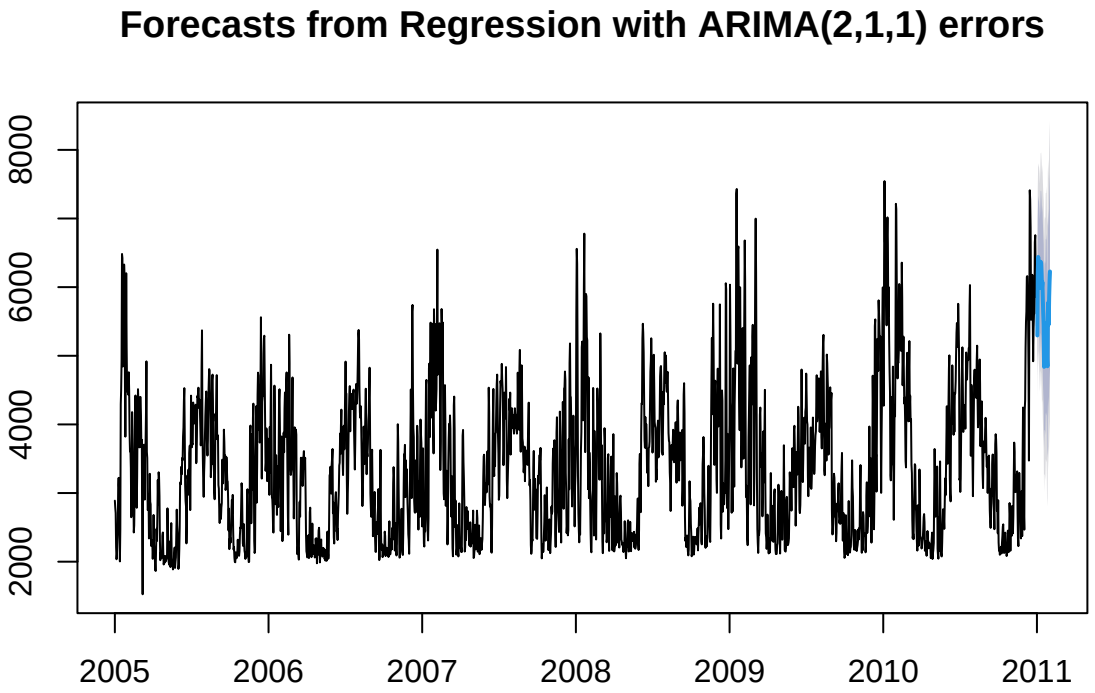
Model 5 - Seasonal ARIMA with temperature and humidity as exogenous

This model implemented a seasonal ARIMA model with temperature and humidity as exogenous regressors. Model order will be identified using the auto.arima function.




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(3,0,1) errors
## Q* = 498.99, df = 358, p-value = 1.14e-06
##
## Model df: 7.   Total lags used: 365
```

```
##                ME     RMSE      MAE      MPE     MAPE
## Test set 403.6135 1013.242 822.6059 3.041937 15.69872
```



Results

Table 1: Performance on Test Set for 5 Models

	ME	RMSE	MAE	MPE	MAPE
SNAIVE	323.8	2284.7	2004.0	-5.7	42.4
SARIMA(1,2,2)	1134.9	1742.7	1419.4	16.0	25.1
SARIMAX(3,0,2)_temp	442.4	1040.3	858.0	3.7	16.3
ARIMA + Fourier	305.2	1468.8	1281.5	-2.3	26.5
SARIMA + temp + humidity	403.6	1013.2	822.6	3.0	15.7

```
## The best model by RMSE is: SARIMA + temp + humidity
```

As shown in clearly in the table above, the best model by RMSE is the SARIMA autofit model with exogenous variables temperature and humidity. On Kaggle, our best performing model was the SARIMA autofit model with temperature alone as an exogenous variable.