Project Report

Aasha Reddy

3/20/2022

Introduction

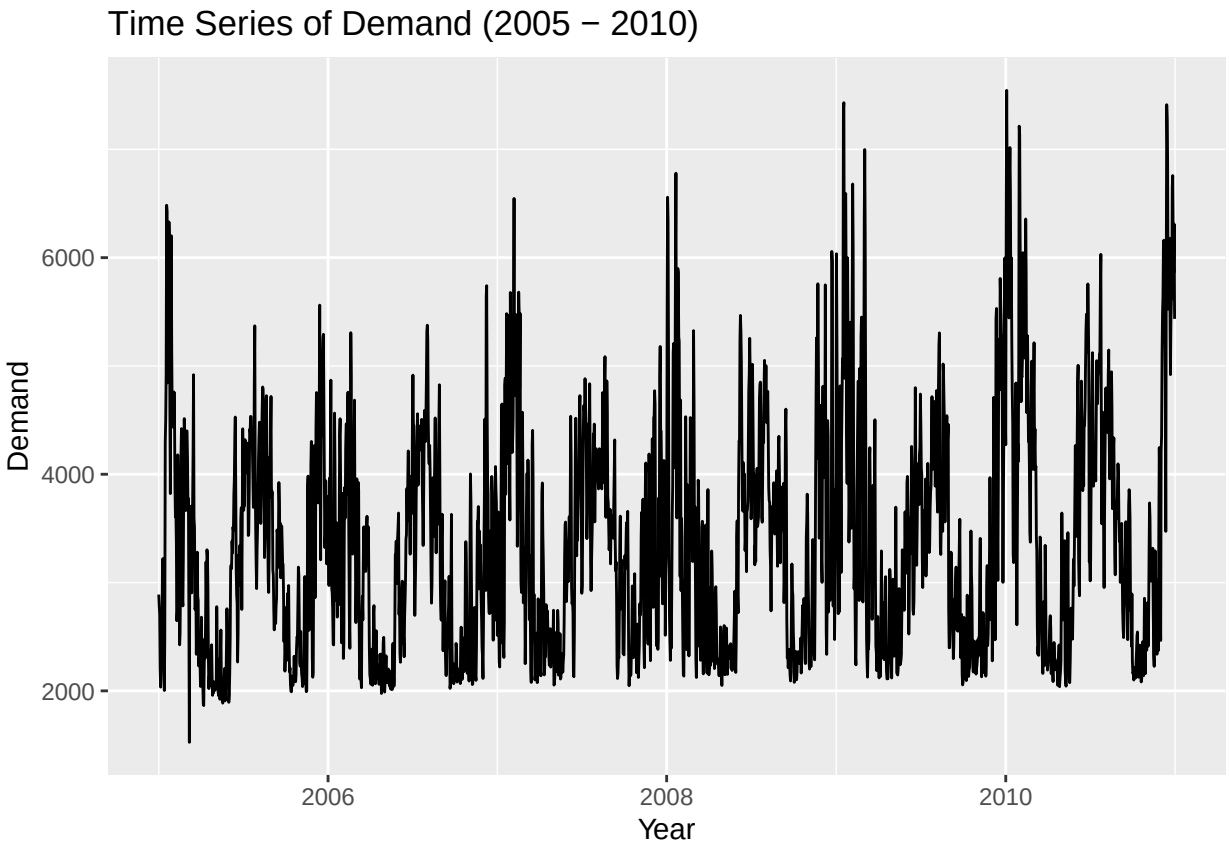
In this assignment, we participate in a Kaggle competition

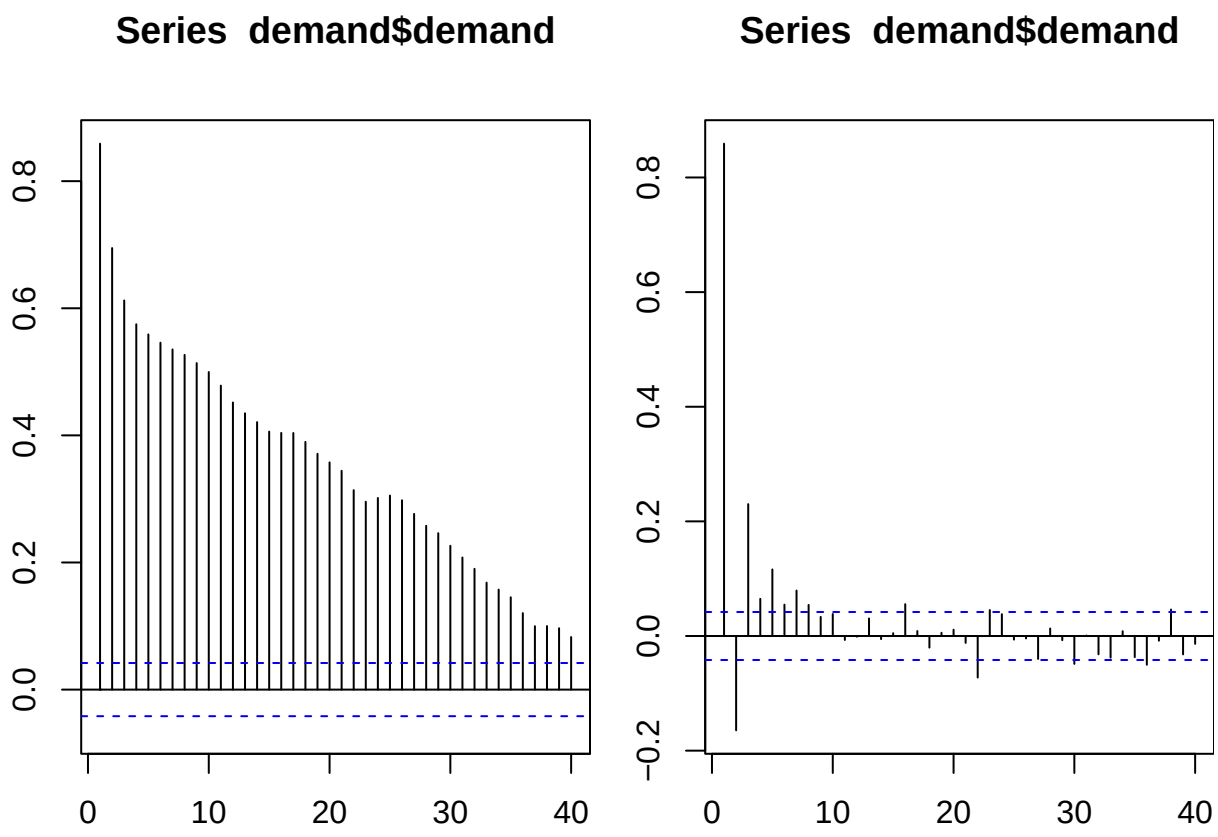
Data Cleaning

First we examine the demand dataset. We note that all rows after 2191 are NA, so we remove those rows. We also note that all meter_id's are 1, so we can remove that column from the dataset. We then average across all the hour columns to transform hourly data into total daily demand. We also transform the date variable into a datetime object using lubridate. In the final dataset, there are no NA values. In examining the humidity and temperature datasets, we also note that there is no missing data. We transform this data from hourly to daily as well to match the demand dataset.

Exploratory Data Analysis

First we will start by examining the demand data. We see that there is definitely a seasonal trend, and it looks like an increasing trend as well. We also see that the ACF plot has maybe a small scalloping pattern.


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

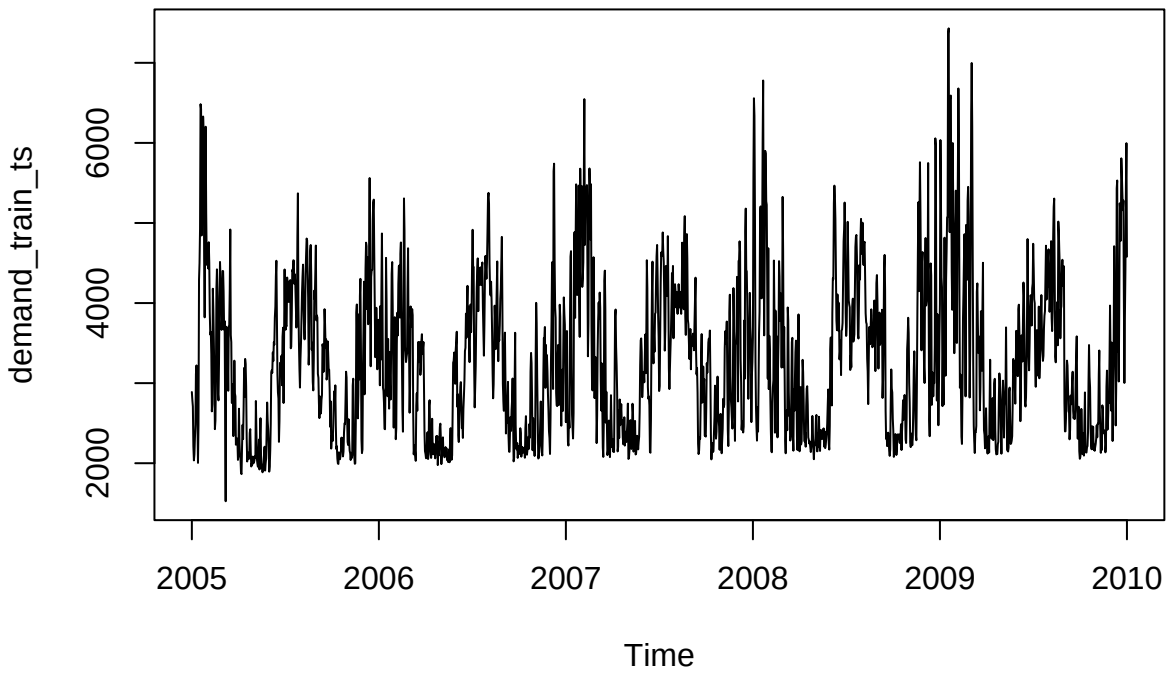


Methods and Modeling

In order to assess our models, we will choose the holdout period to be the final year in the dataset, 2010. We will thus train our models on data from 2005 - 2009, and then evaluate how well our models perform on the 2010 data, choosing the final model to be the one that acheives the highest MAPE. Finally, we will forecast demand for 2011 using our best model.

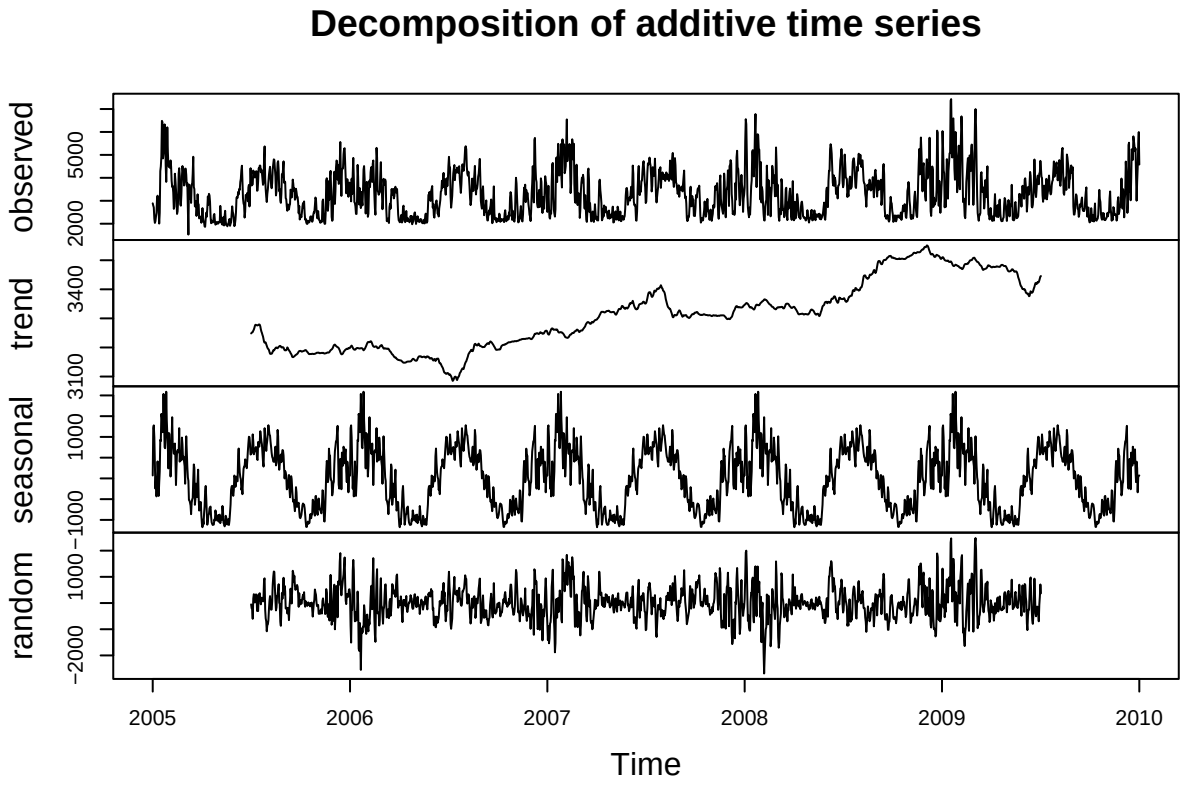
We also transform the demand dataframe into a time series object.

```
## Time Series:
## Start = c(2005, 1)
## End = c(2005, 15)
## Frequency = 365
##  [1] 2889.125 2788.958 2708.458 2211.583 2035.125 2109.625 2313.375 2479.708
##  [9] 3115.542 3221.583 3055.542 2498.833 2003.917 2475.667 4293.500
```

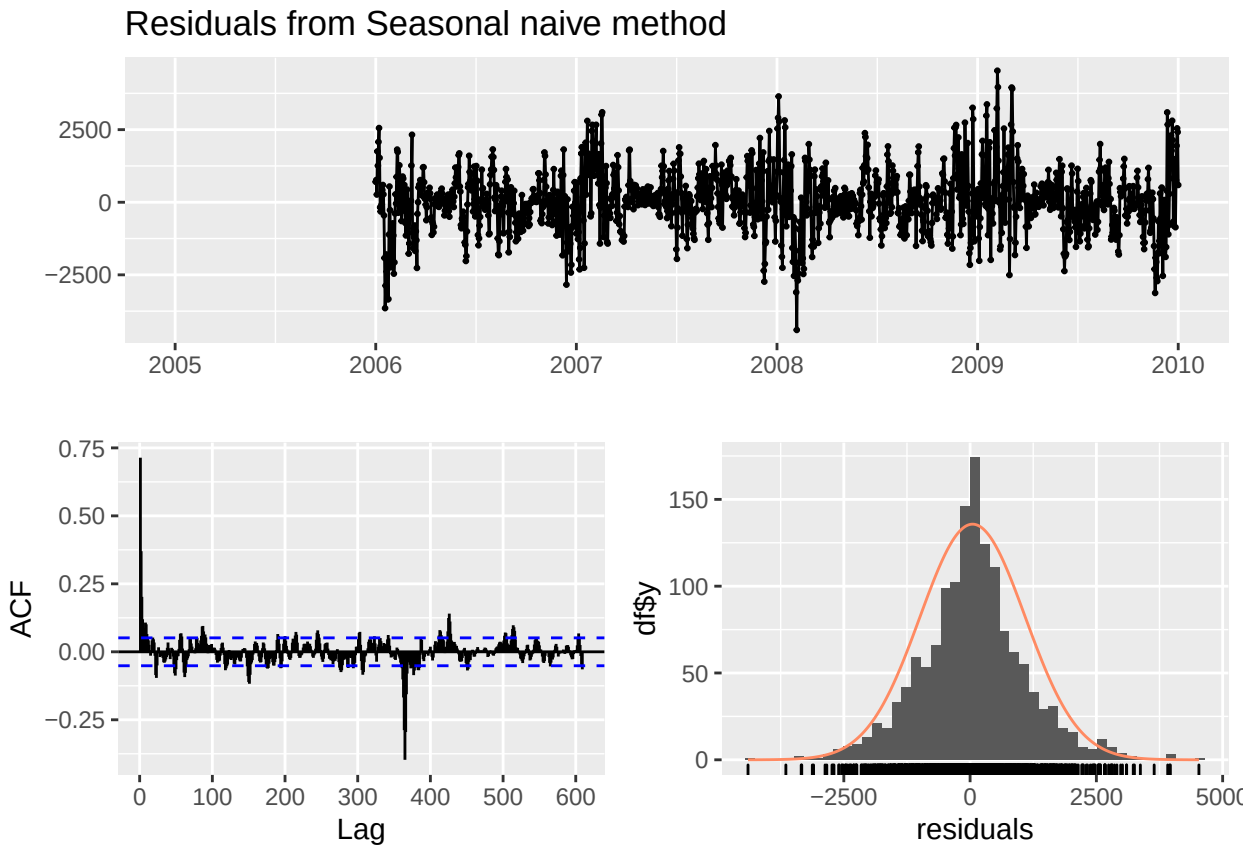


Model 1 - Seasonal Naive Model

The first model we try is a seasonal naive model. To do this we first decompose and deseason the demand series.



Next we fit the model using the training data, from 2005 - 2009.

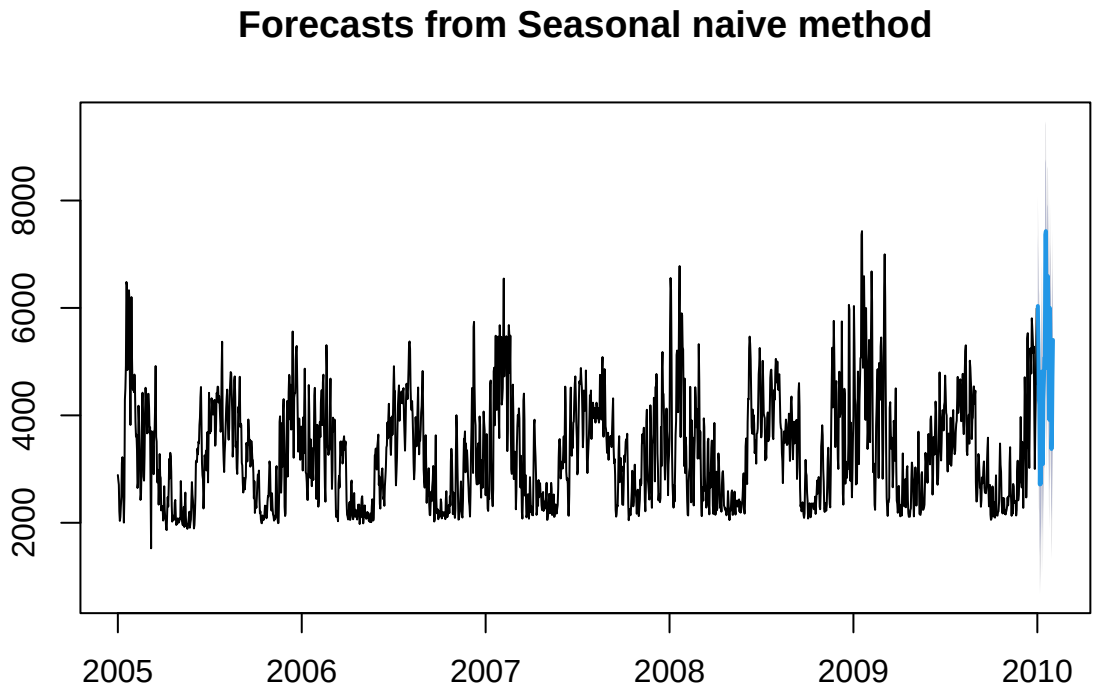



```
##
##  Ljung-Box test
##
## data:  Residuals from Seasonal naive method
## Q* = 2503.7, df = 365, p-value < 2.2e-16
##
## Model df: 0.   Total lags used: 365
```

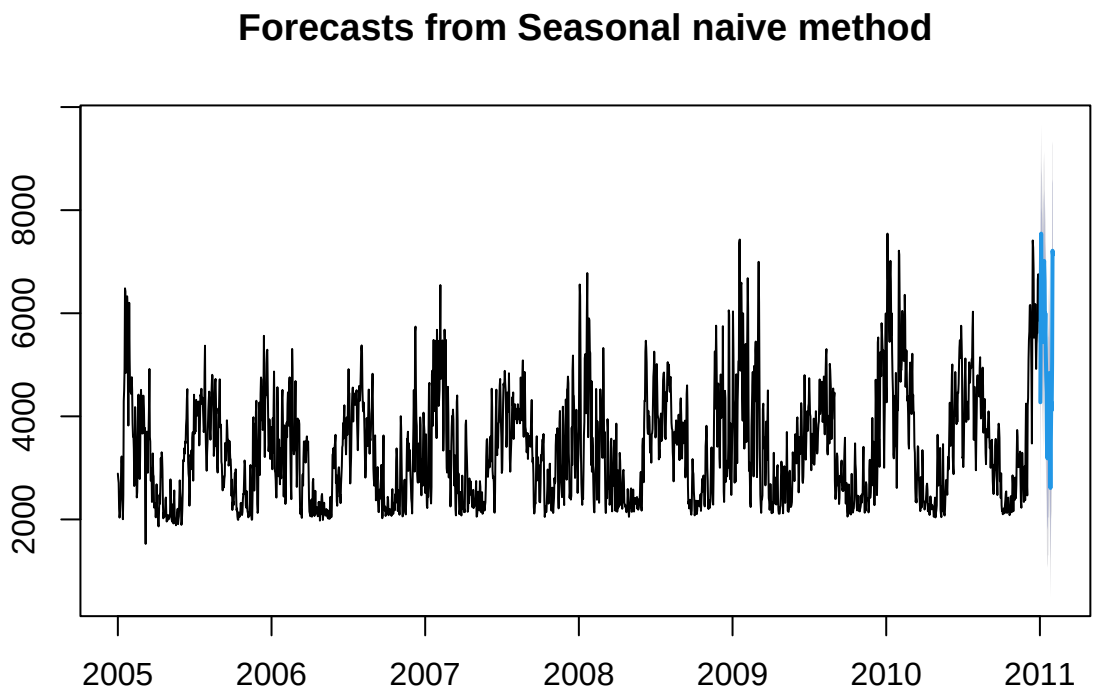
We see that the residuals are normally distributed. However, there seems to be a pattern and significant correlation in the ACF. We hope to improve on this in our next model.

We then forecast daily demand 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 our seasonal naive model is 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
```



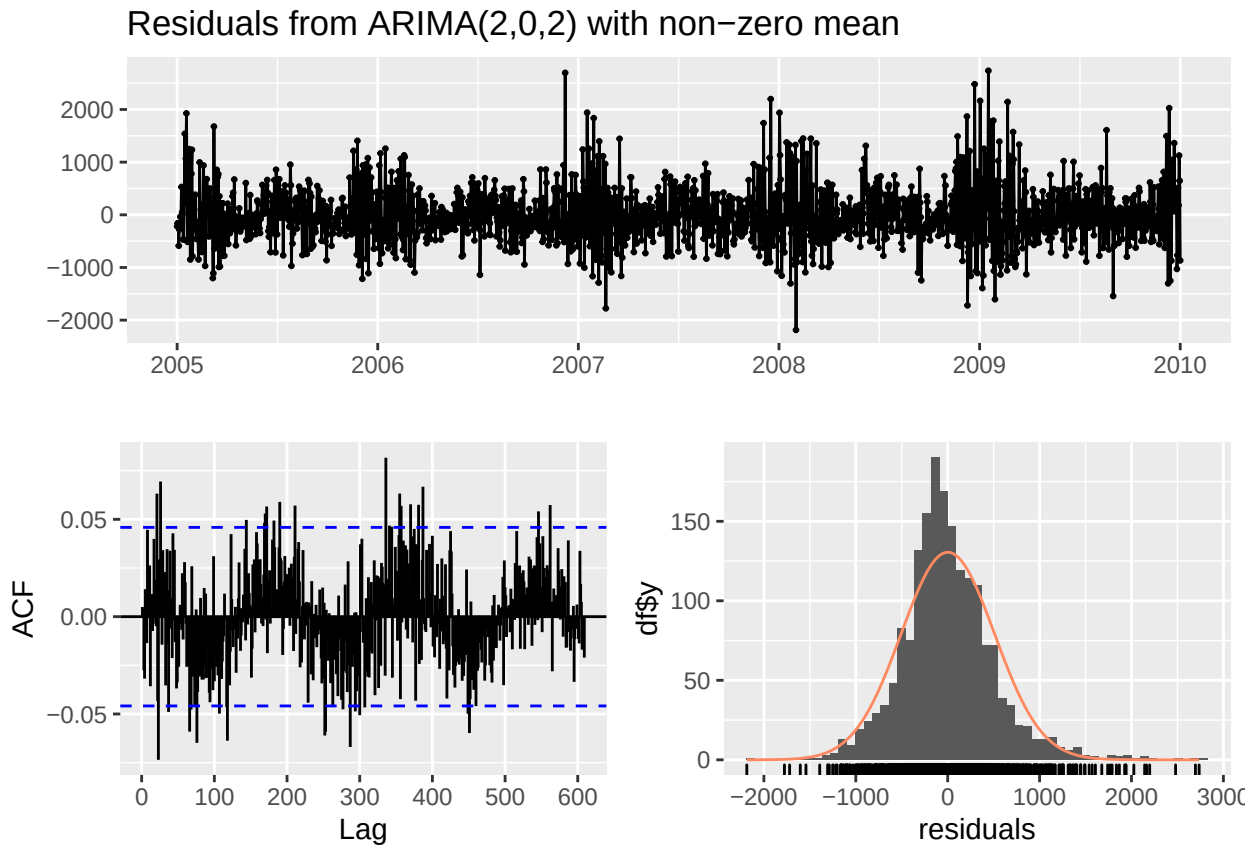
Now we refit the seasonal naive model using all data from 2005 - 2010, and then use that to forecast for January 2011.



Model 2 - Seasonal ARIMA model (Autofit)

Our second model will be a seasonal ARIMA model.

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

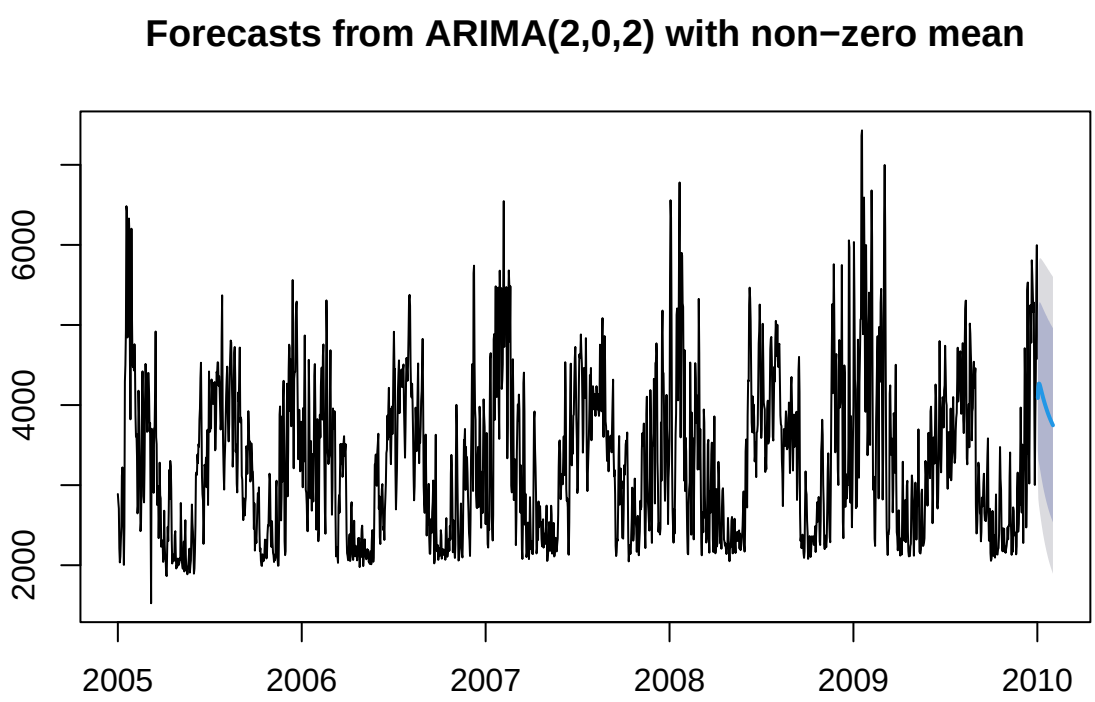



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,0,2) with non-zero mean
## Q* = 515.18, df = 360, p-value = 1.39e-07
##
## Model df: 5.   Total lags used: 365
```

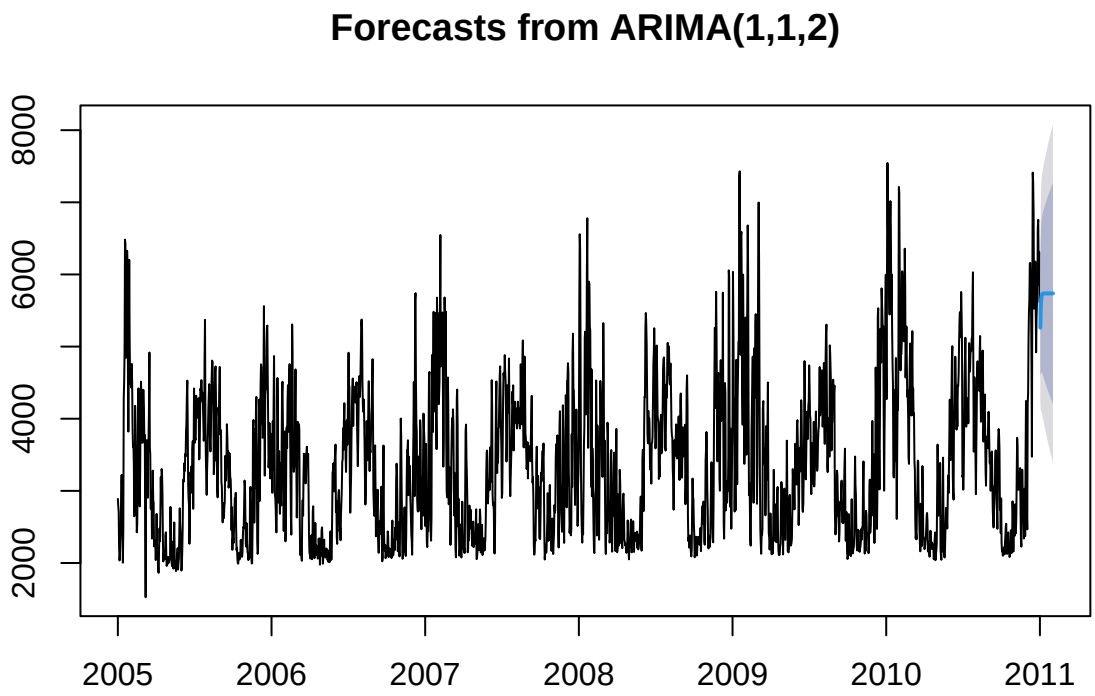
We see that the residuals here are normally distributed but do not seem to be that random. ACF shows significant correlations as well - this means that this likely will not be our best model!

Still, we move forward and forecast daily demand for January 2010, seen in the plot below. The MAPE on the test set (Januray 2010) is 25.1363 here, and we note that this is a large improvement from the seasonal naive model. We also see the plot forecast for Jan 2010 below.

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

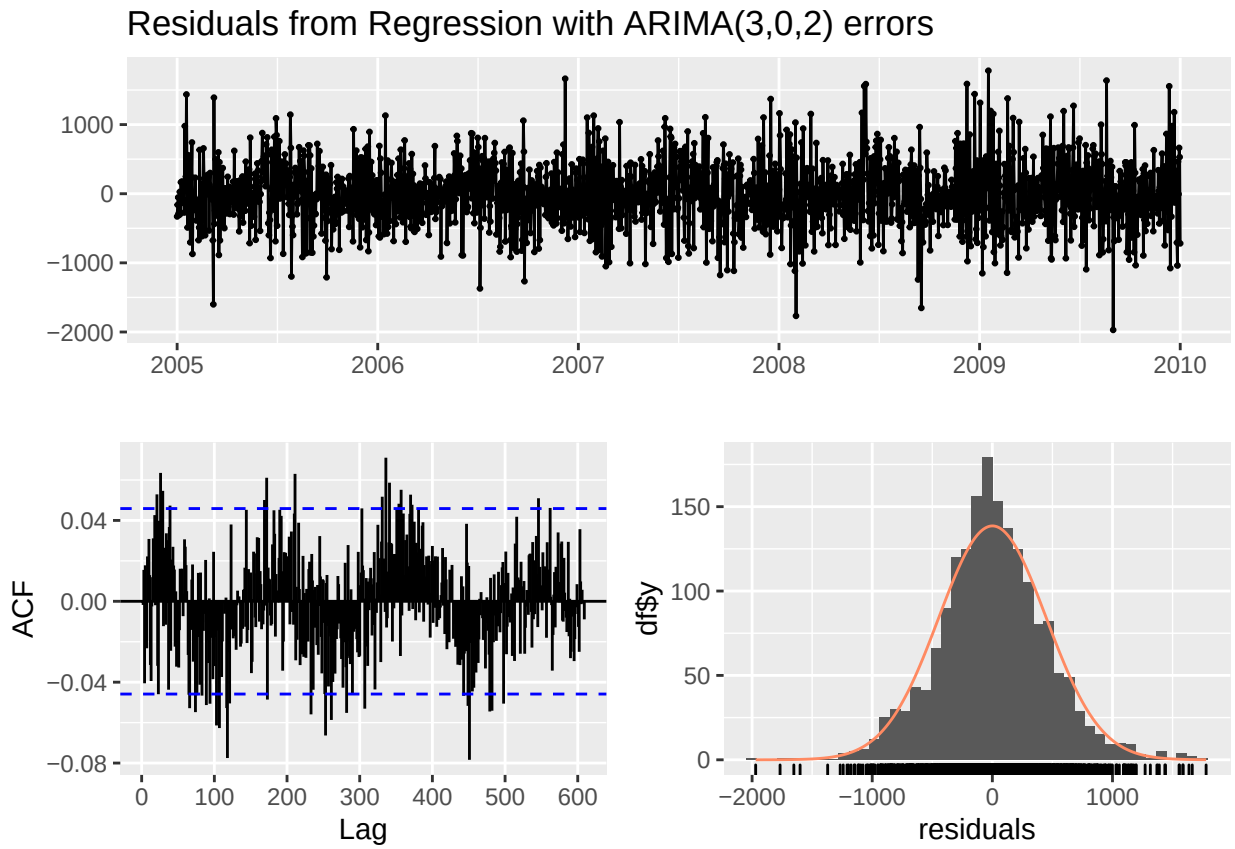


Now we refit the seasonal naive model using all data from 2005 - 2010, and then use that to forecast for January 2011.



Model 3 - auto.arima with temperature as exogenous variable

Next we will fit a use auto.arima to fit a seasonal arima model that includes temperature as an exogenous variable.

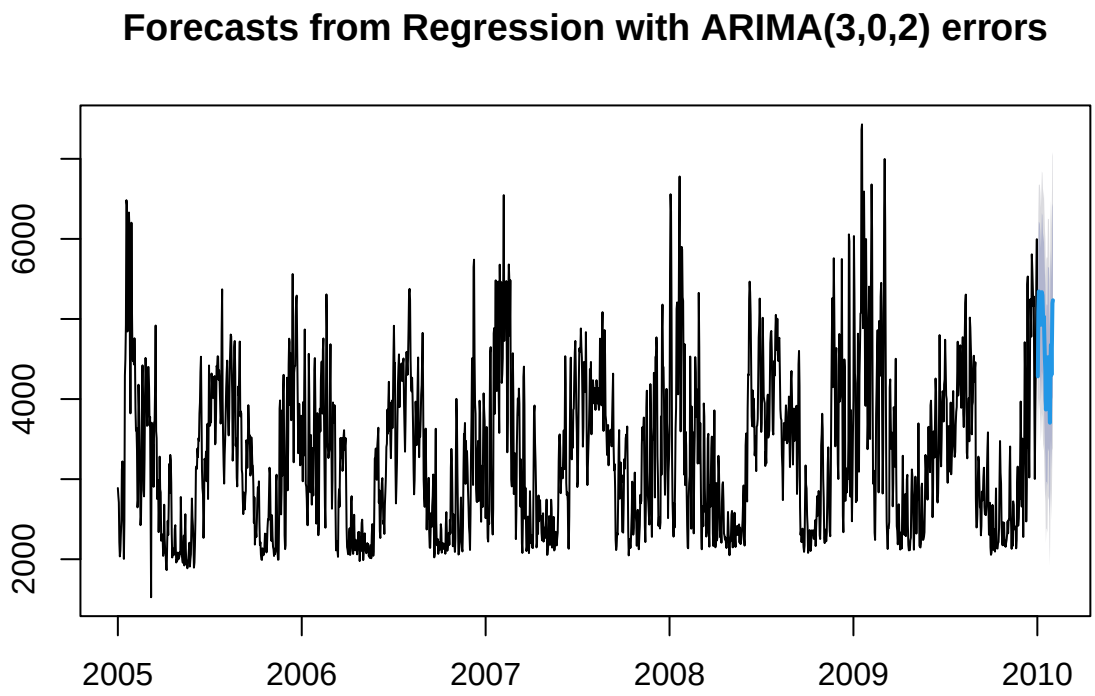



```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(3,0,2) errors
## Q* = 494.24, df = 358, p-value = 2.285e-06
##
## Model df: 7.   Total lags used: 365
```

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

Regardless, we will forecast demand for January 2010. 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.

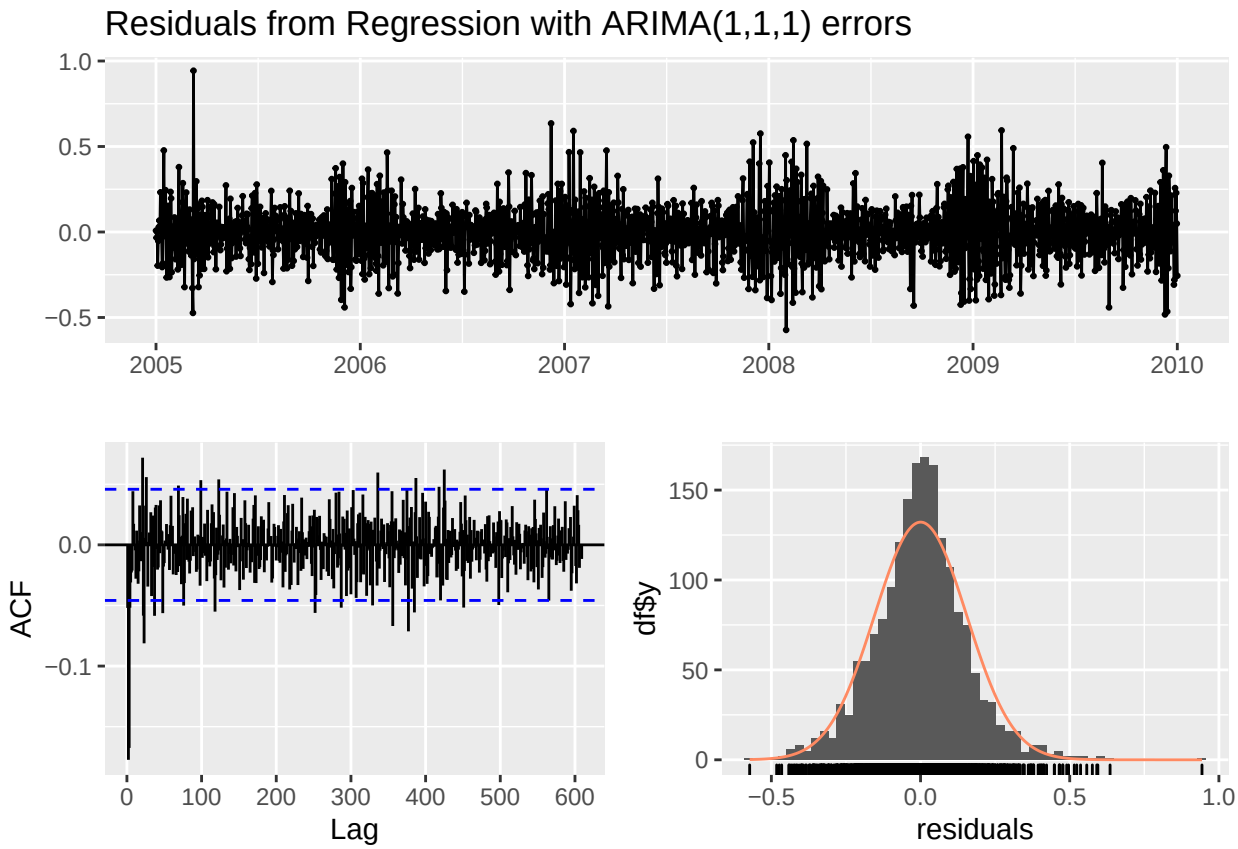
```
##                ME     RMSE      MAE      MPE     MAPE
## Test set 442.364 1040.276 857.9651 3.742301 16.26123
```



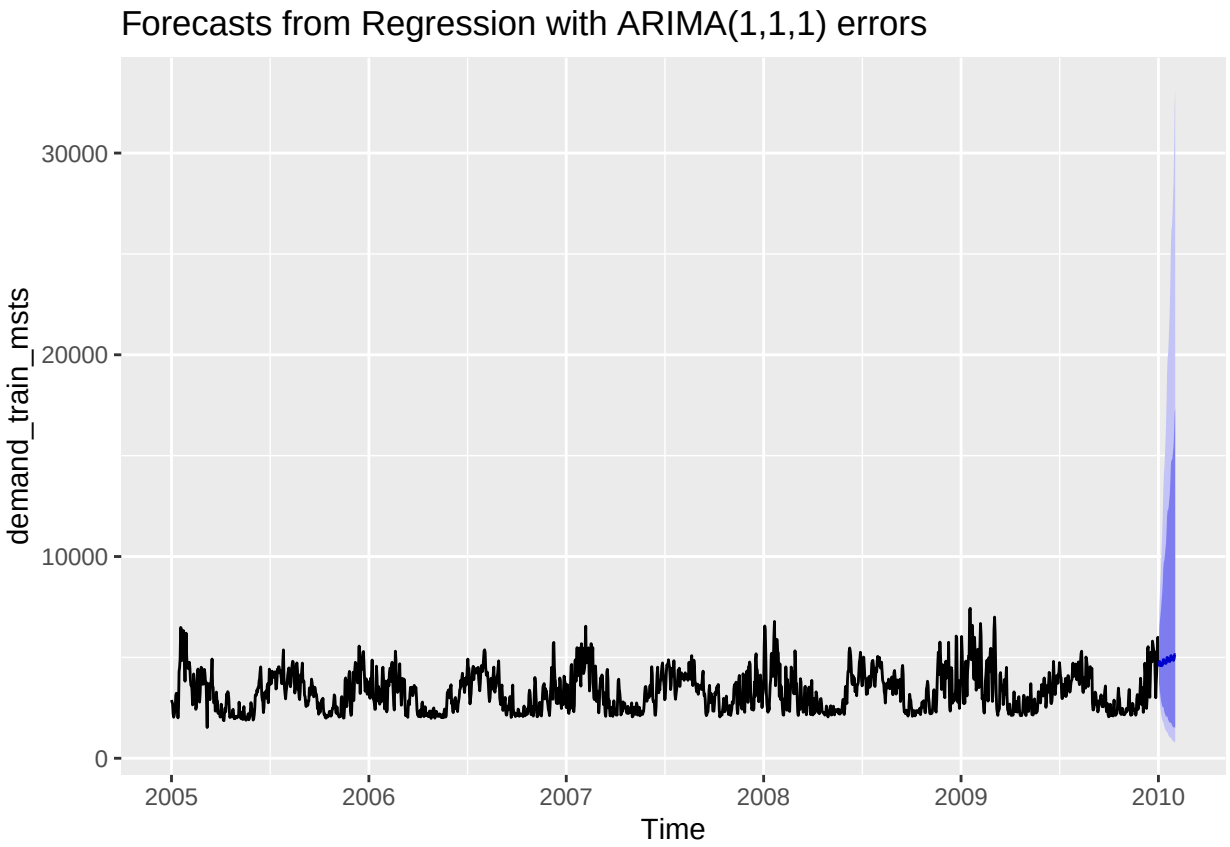
Now we refit the seasonal naive model using all data from 2005 - 2010, and then use that to forecast for January 2011. The January 2011 temperature is assumed to be the same as January 2010 temperature.

Model 4 - ARIMA + FOURIER

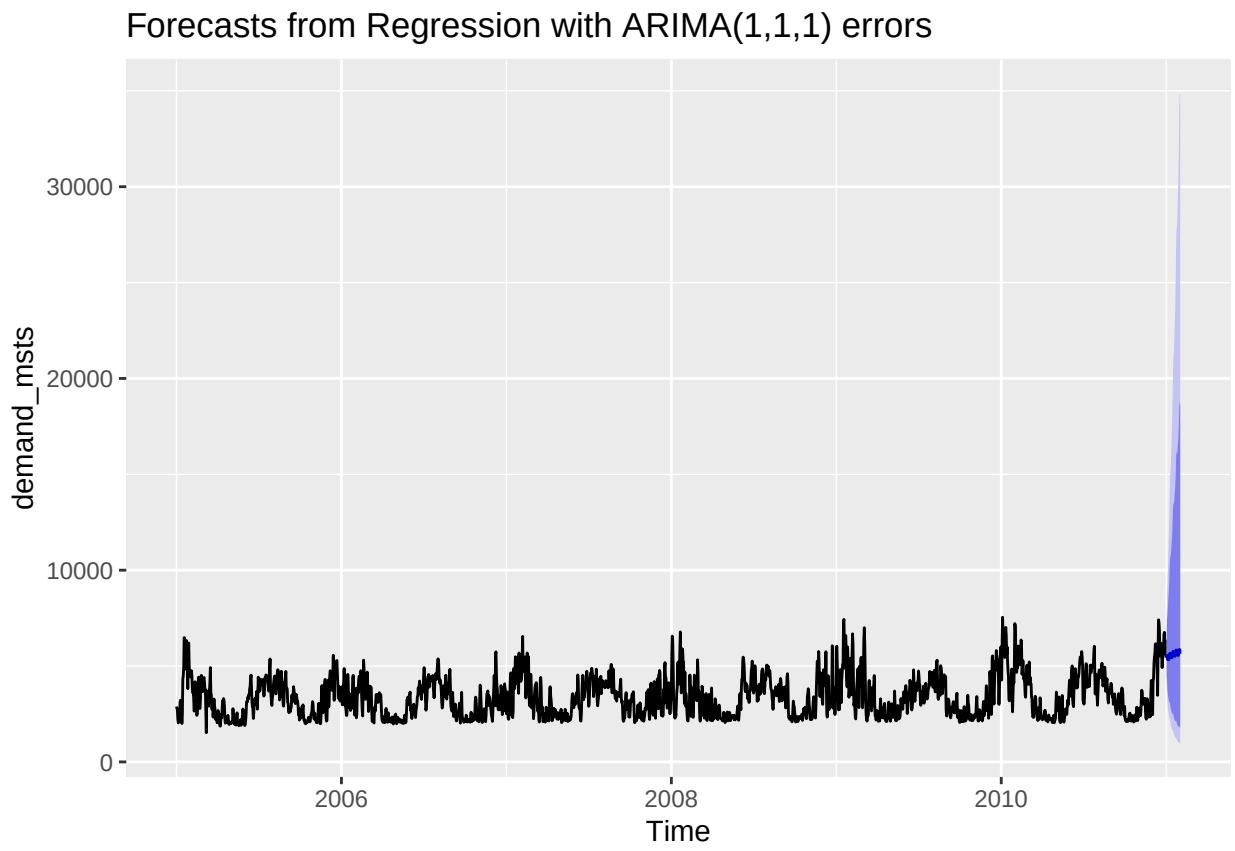
Next we will use a autofit arima with seasonal fourier terms to account for multiple seasonalilty present in the data. The multiple seasonality comes from the the repeated weekly variation and seasonal monthly variation throughout the year.




```
##
##  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
```

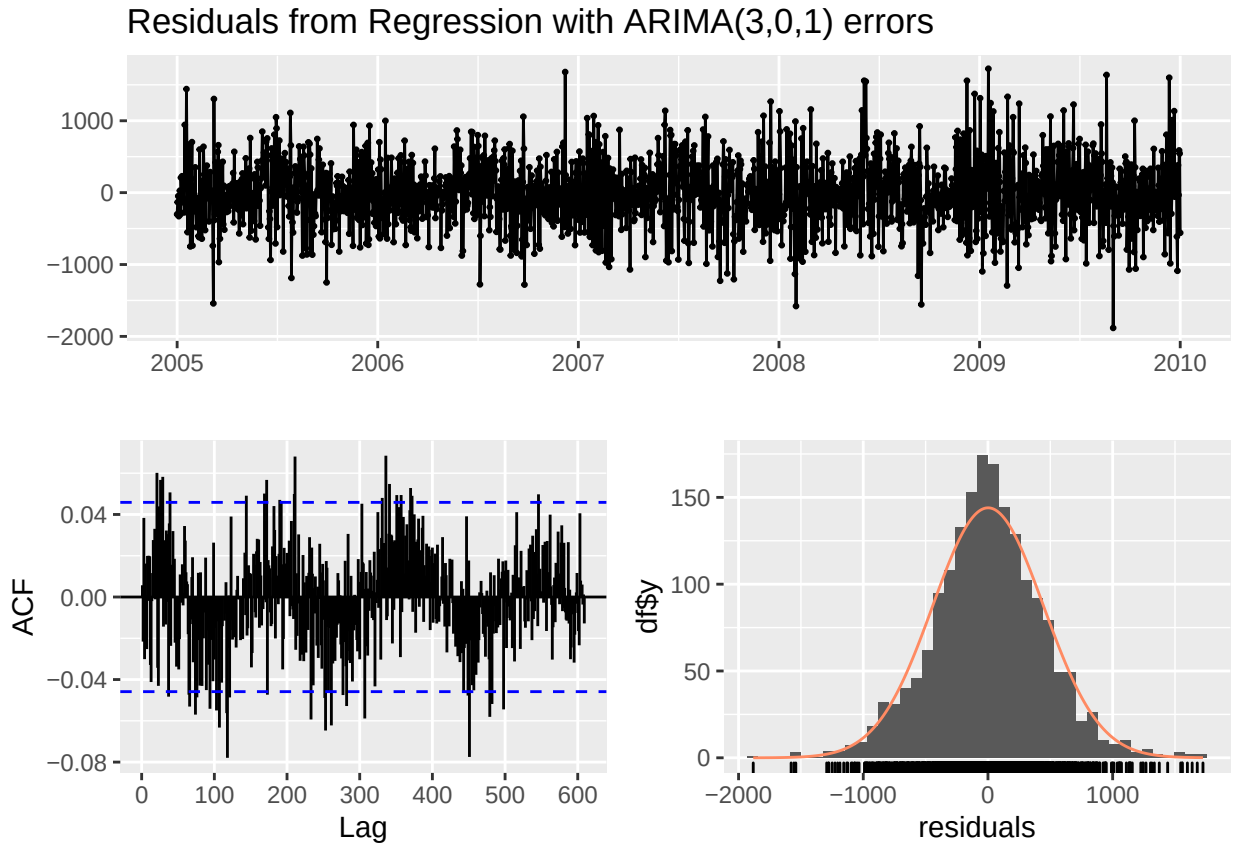


Next, we refit the model using all demand data.



Model 5 - Seasonal ARIMA with temperature and humidity as exogenous

This model will implement 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
```

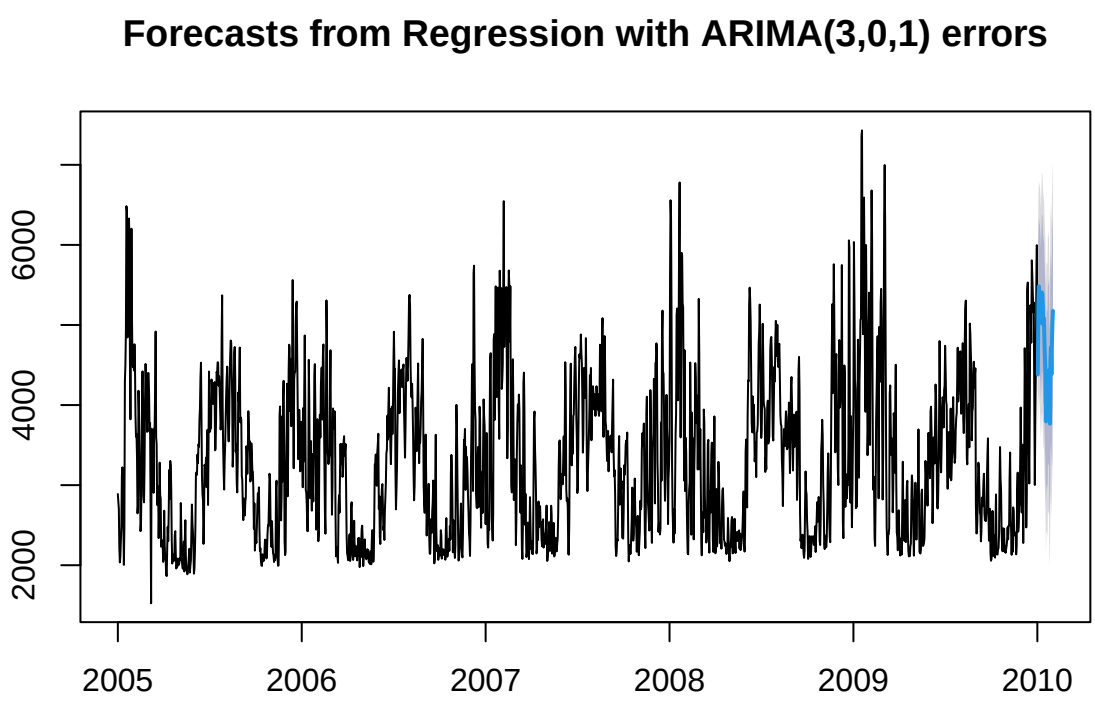
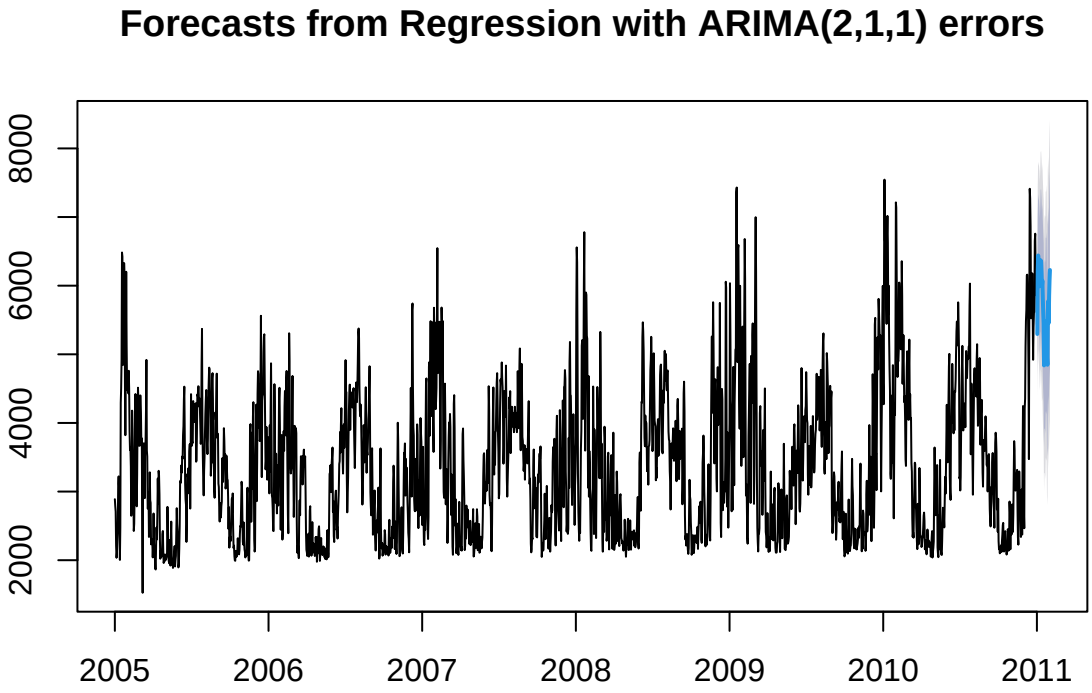

Forecasts from Regression with ARIMA(3,0,1) errors
2000
4000
6000
2005
2006
2007
2008
2009
2010

Table 1: Performance on Test Set for 5 Models

	ME	RMSE	MAE	MPE	MAPE
SNAIVE	323.79	2284.72	2003.99	-5.69	42.36
SARIMA(1,2,2)	1134.95	1742.69	1419.37	15.96	25.14
SARIMAX(3,0,2)_temp	442.36	1040.28	857.97	3.74	16.26
ARIMA + Fourier	305.23	1468.77	1281.52	-2.35	26.47
SARIMA + temp + humidity	403.61	1013.24	822.61	3.04	15.70



Results

As we can see from the below table, 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.

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