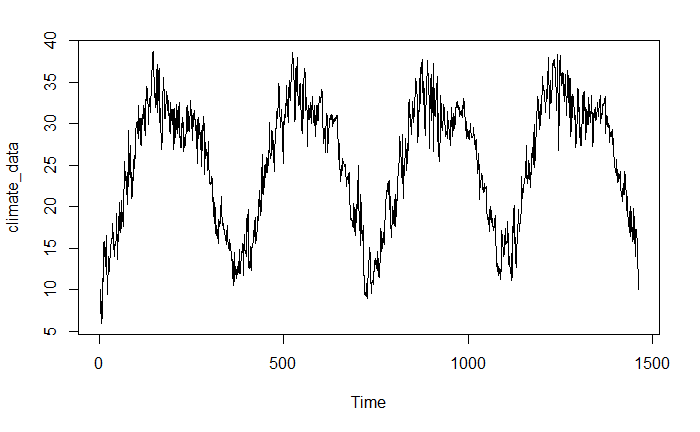
**Exploration**

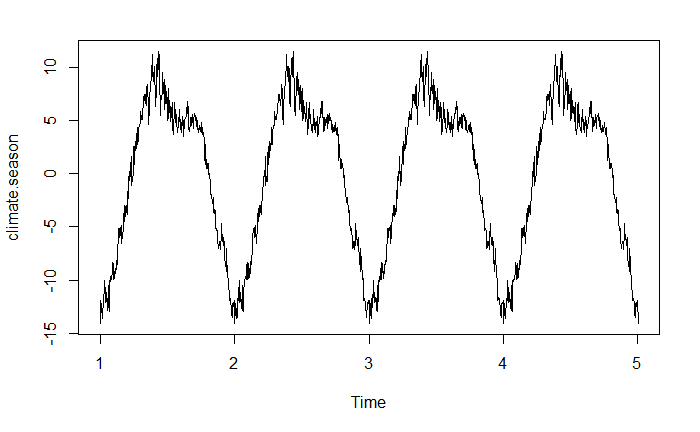
The first step in analyzing the climate data is plotting the raw time series, viewing temperature as the response variable against time. This time series is not stationary, as the mean changes over time and there is clearly seasonality present. Since there is no change in variance over time, no log transformation is required to stabilize the data in this scenario. The time series plot shows a clear seasonality pattern, with a period every ~365 days. For the 4 years of data, there are 4 periods, which makes sense because the temperature changes are likely caused by seasons throughout the year. Therefore, we will estimate the seasonal component as 365 days in the process of making the time series stationary. There is also a slight upward trend over time (Figure 1). We need to remove seasonality and trend to achieve stationarity by fitting a linear regression of the time series and analyzing residuals.



**Figure 1.** Climate time series data plotted before removing trend and seasonality.

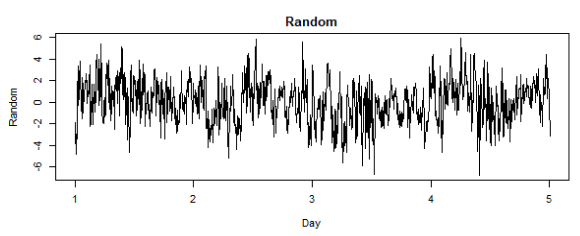
**Removing Trend and Seasonality**

We remove the trend from the time series by extracting the residuals from the linear model. The time series no longer shows an upward trend as seen in Figure 2. The mean is constant over time, but seasonality remains.



**Figure 2.** Climate time series data plotted after removing the positive trend.

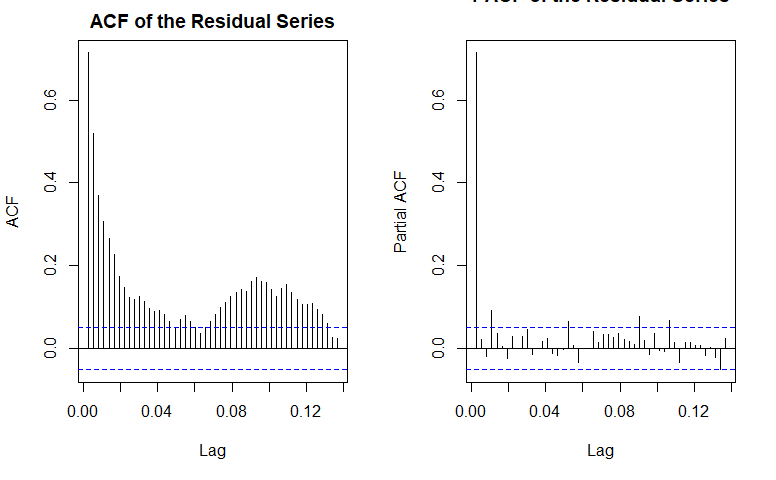
We estimate the seasonality from the residuals of the linear model, and then remove the seasonality. We observe that the ‘Random’ plot after removing seasonality behaves like white noise. Therefore, we have successfully removed trend and seasonality from the time series to obtain a stationary series (Figure 3).



**Figure 3**. Stationary times series after removing trend and seasonality resembles a white noise process.

**Fitting ARMA Model**

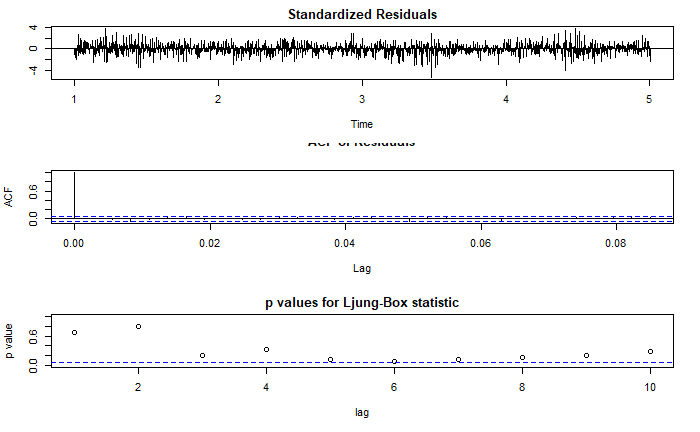
We will fit an ARMA model to the stationary time series by making conclusions from the ACF and PACF plots. These plots show where autocorrelation may occur in the time series data. We observe that the ACF plot slowly decreases to zero and the PACF plot has a non-zero first lag. There are many models that may work for this data, so several models will be compared. There are also non-zero PACF values at the period lags, so we will fit a SARIMA model just in case. Based on the ACF and PACF plots, we chose to compare a few model candidates: AR(1), AR(2), AR(3), ARMA(1,1), ARMA(2,2), ARMA(2,1), and ARIMA(1,1,0)x(0,1,0)12. AIC is used to compare these models, where the ARMA(2,1) model returns the smallest value of all the other models (4990, see appendix for other model results).



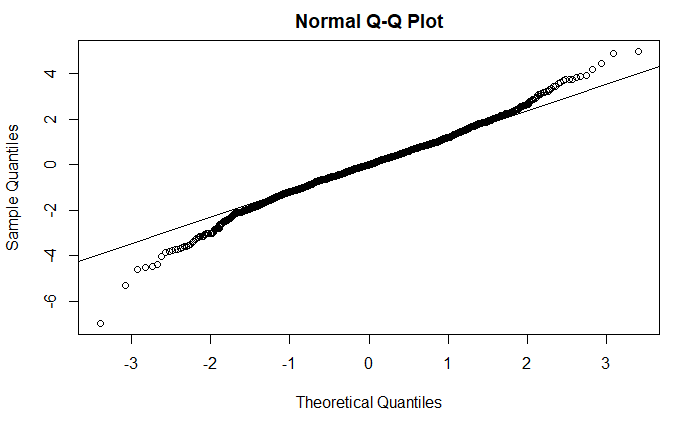
**Figure 4**. ACF and PACF plots for the stationary climate time series.

**Model Diagnosis and Analysis**

We can check the residuals of our selected ARMA(2,1) model to evaluate the fit. The model is evaluated with diagnostics using R function tsdiag(). The residuals look reasonable - the process is centered at zero and does not have a trend (Figure 5, top). The residual ACF and PACF plots resemble white noise - there is no erroneous autocorrelation in the residuals (Figure 5**,** middle. PACF plot can be seen in appendix). The p-values are larger than 0.05, which suggests residuals come from white noise (Figure 5, bottom). The ‘qqplot’ supports the assumption of normal distribution of the residuals because the data points fall along the straight line. The model fitting performs reasonably well.



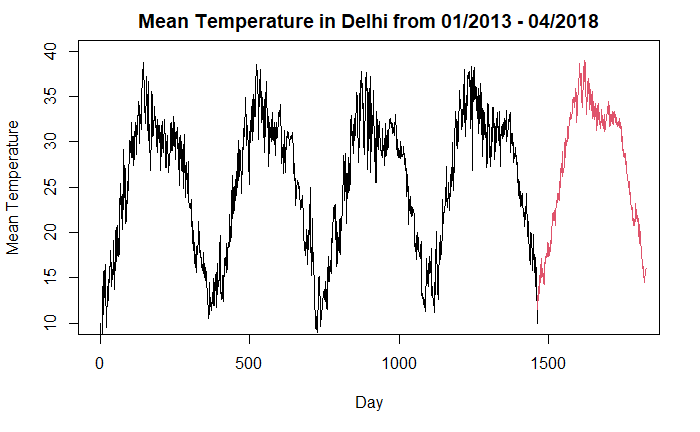
**Figure 5**. Results from tsdiags() show a well-fitted model.

****

**Figure 6**. ‘qqplot’ of the residuals support assumption of normal distribution.

**Model Prediction and Forecasting**

To forecast the next year of temperature data, we can use the predict() R function to predict the next 365 data points. We add the trend and seasonality back into the time series model for the prediction to obtain the plot of the new predicted data below.



**Figure 7.** Predicted values from the fitted ARMA(2,1) model for the next year of temperature.

Notes for Connor’s results/discussion:

* Better results may have been obtained by comparing more ARMA or SARIMA models
* Alternative approach would have been to using function diff() to difference the time series data
* If it were provided, we could have also used a ‘test’ data set to evaluate the performance of our data by comparing our ARMA(2,1) predictions with real results
* Overall the ARMA(2,1) model performs fairly well and the forecast reflects the behavior of the time series very well.
* The forecast shows increasing temperature in the following year, which is what we expected