ST-534 Project

Titus Dorsey

Jiatao Wang

Wenjin Liu

Data Set Introduction

Temperature affects everyday life. People know approximately what temperature to expect because temperature cycles around the four seasons as the Earth orbits the Sun. In this project, we collected historical temperature data from https://rp5.ru/Weather_in_the_world. The data set contains temperature values measured at 7:00 AM for the Raleigh area between 02/01/2010 and 10/31/2021. This data set is in CSV format so it can be easily imported into SAS for analysis.

Missing Data

Between 02/01/2010 and 10/31/2021, there are a total of 4291 days. Temperature values are available for 4239 days but missing for 52 days. There are many ways to fill in missing data as shown below:

1. Linear Interpolation: Missing data are interpolated by creating a straight line between the two points surrounding the missing data. For example, the value on 01/02/2021 is interpolated by a line between 01/01/2021 and 01/03/2021.

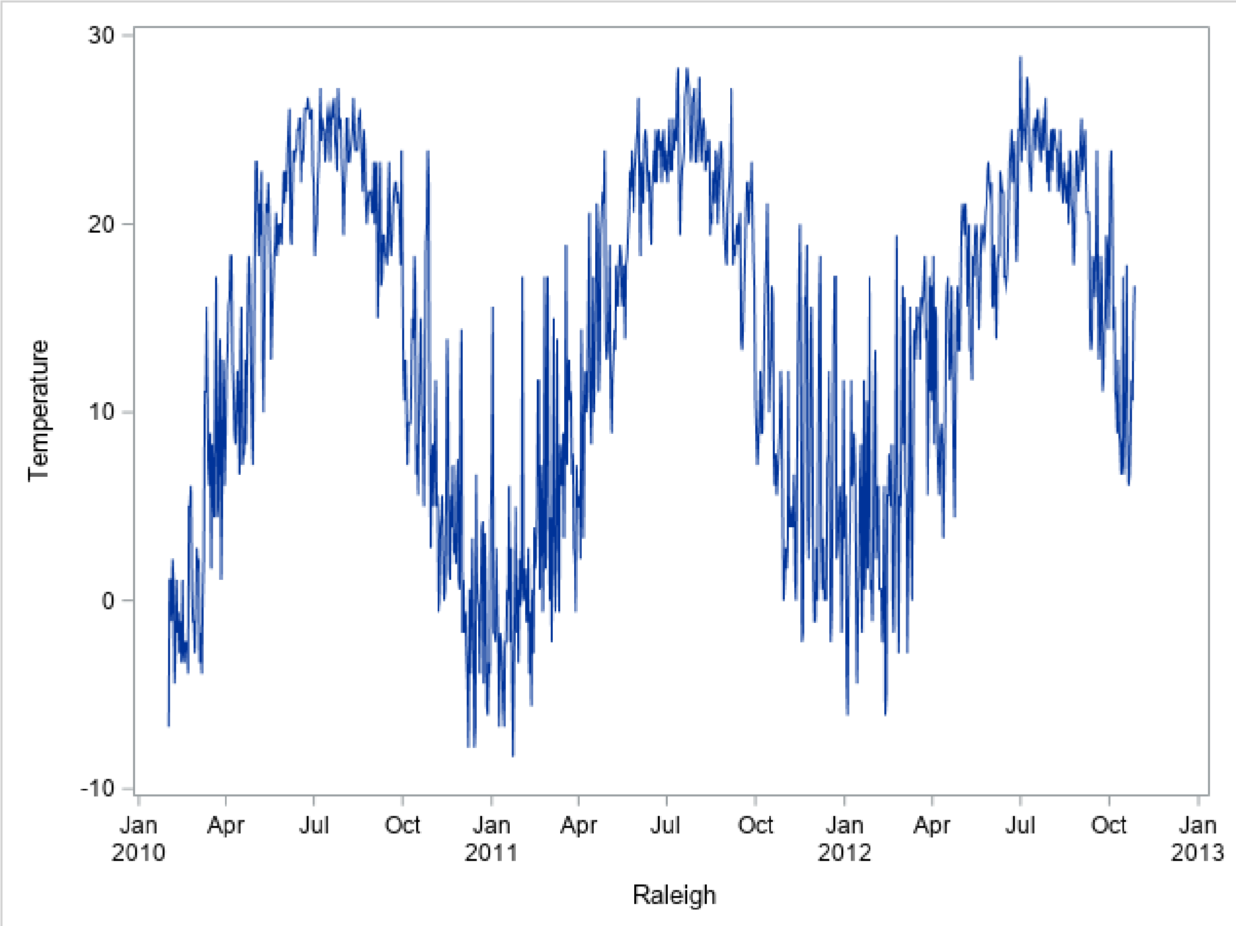
2. Mean Interpolation: Missing data are interpolated by the mean of the values on the same date across different years. For example, the value on 01/02/2021 is interpolated by the mean of the values on 01/02 from 2010 to 2020.
3. Median Interpolation: Missing data are interpolated by the median of the values on the same date across different years. For example, the value on 01/02/2021 is interpolated by the median of the values on 01/02 from 2010 to 2020.

In this data set, missing data are spread over the entire time period instead of focusing on a particular date, so we have enough data to perform mean or median interpolation for each date. To avoid the interpolation being affected by extreme temperature outliers, we used median interpolation to fill in missing data.

Model Fitting

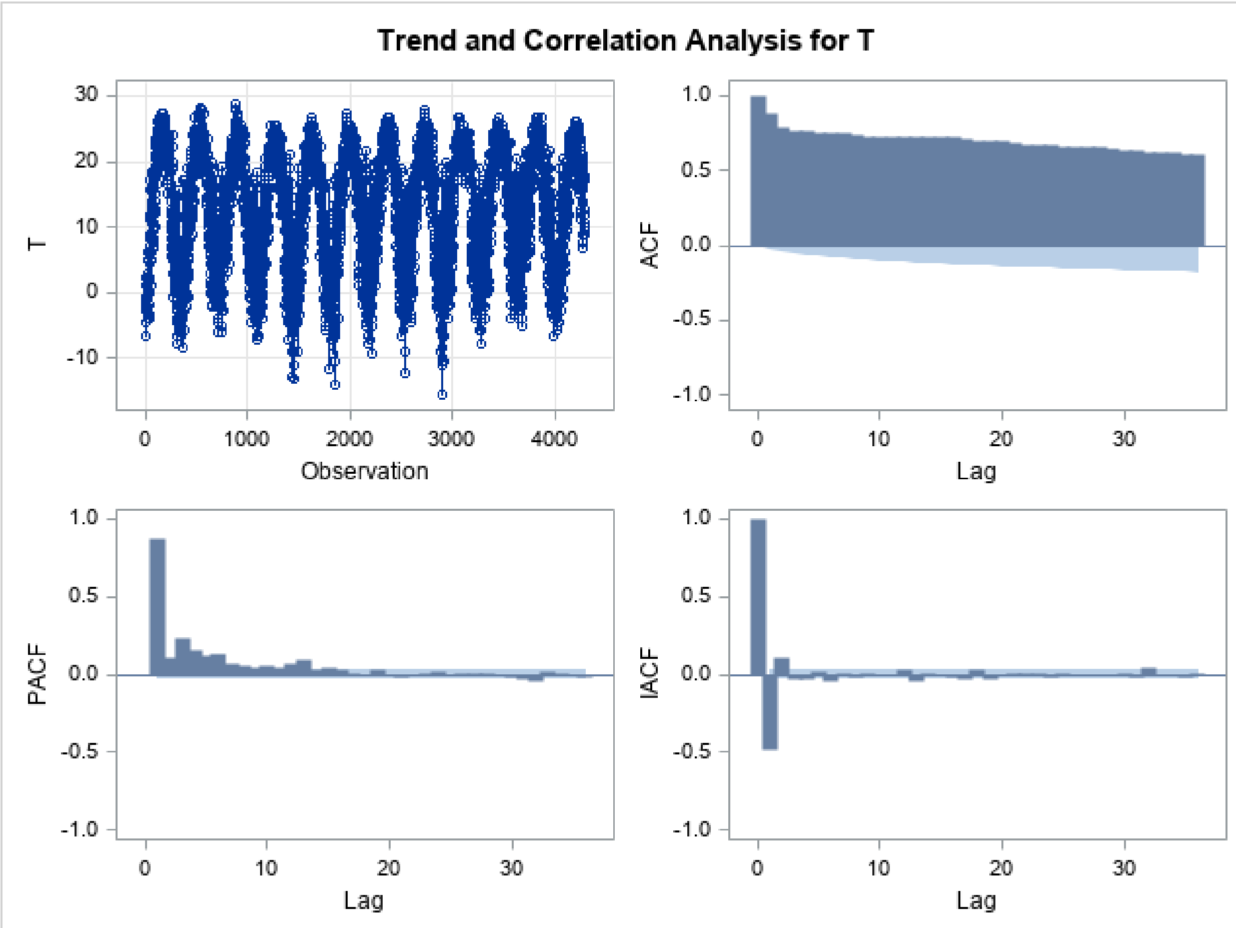
The SAS Code used in this project can be found in the accompanying ST534_project_Read_in_The_Data.docx file.

The first step in the model fitting stage is to plot the data.



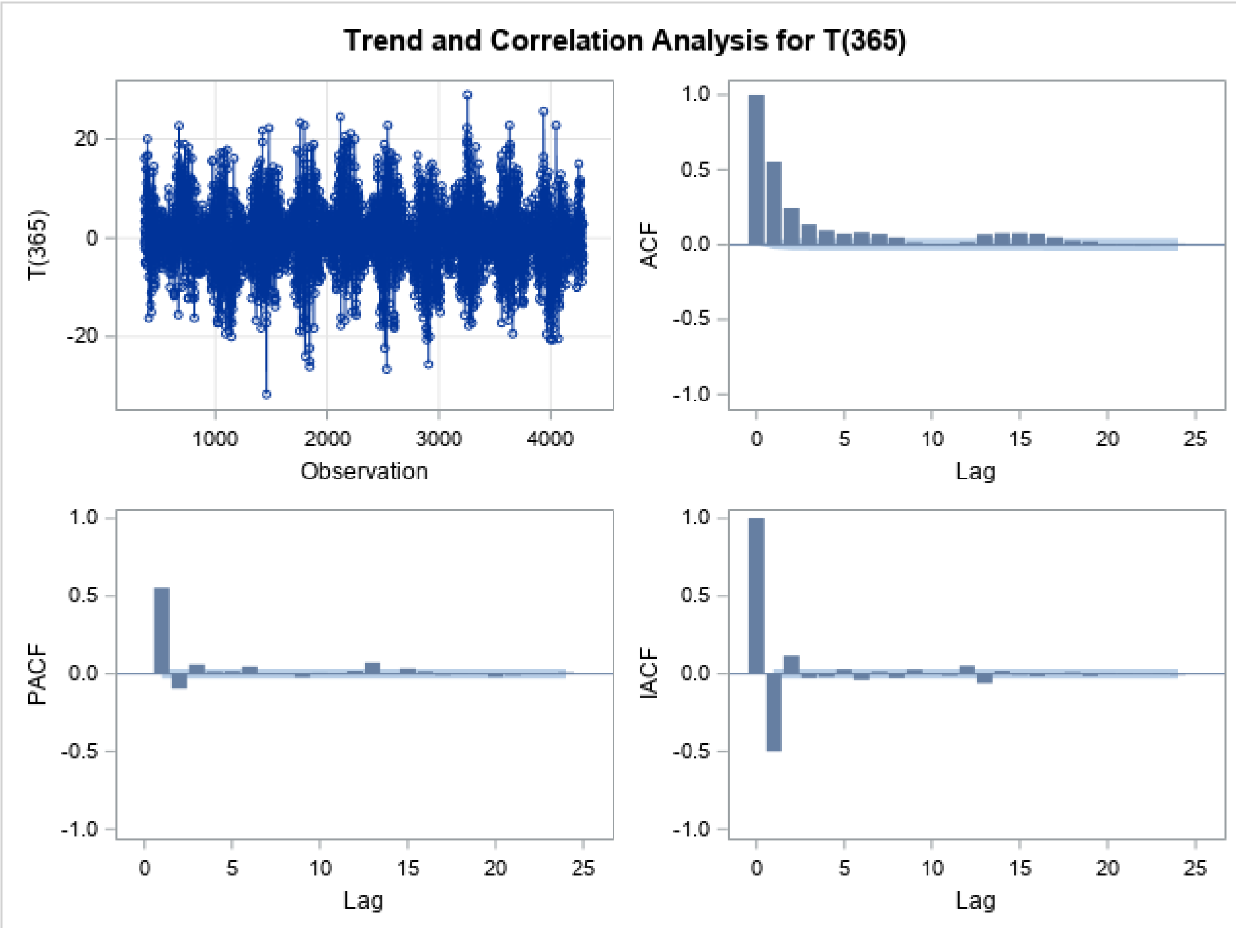
This is a plot of the first three years in the data set. We can easily see the seasonal trend. The temperature data appears to be fluctuating throughout the year as we would expect. The seasonal component in this data is “s = 365”. That is we are comparing the temperature of days that are 365 days apart.

Next, We looked at the ACF and PACF with PROC ARIMA.



We can easily determine that we need to difference this data by looking at the slowly decreasing ACF and the seasonal trend.

We took a seasonal difference and visualized the output.

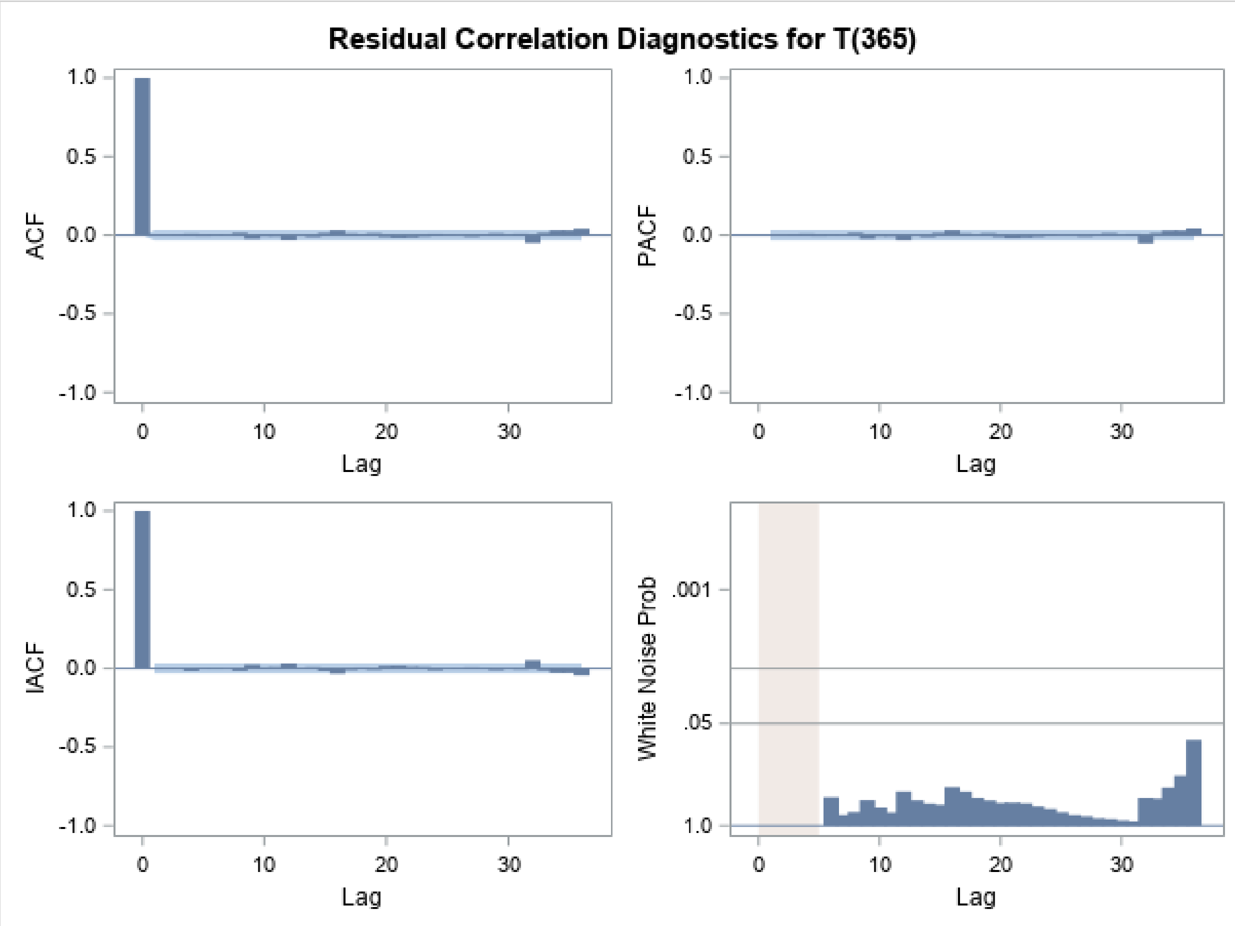


The data looks stationary with mean zero but we performed the Dickey-Fuller test to help guide our decision of not differencing the data again.

Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-1743.37	0.0001	-33.48	<.0001		
Single Mean	0	-1743.50	0.0001	-33.47	<.0001	560.28	0.0010
Trend	0	-1744.28	0.0001	-33.48	<.0001	560.43	0.0010

The test rejects the null hypothesis of there being a unit root in the data. Therefore, the test confirms that we do not need to difference the data again.

Next, we fit an AR(20) model to the data and chose the lags with significant P values. These were 1,2,3,6,and 13. So, we refit the AR(1,2,3,6,13) model to the data.

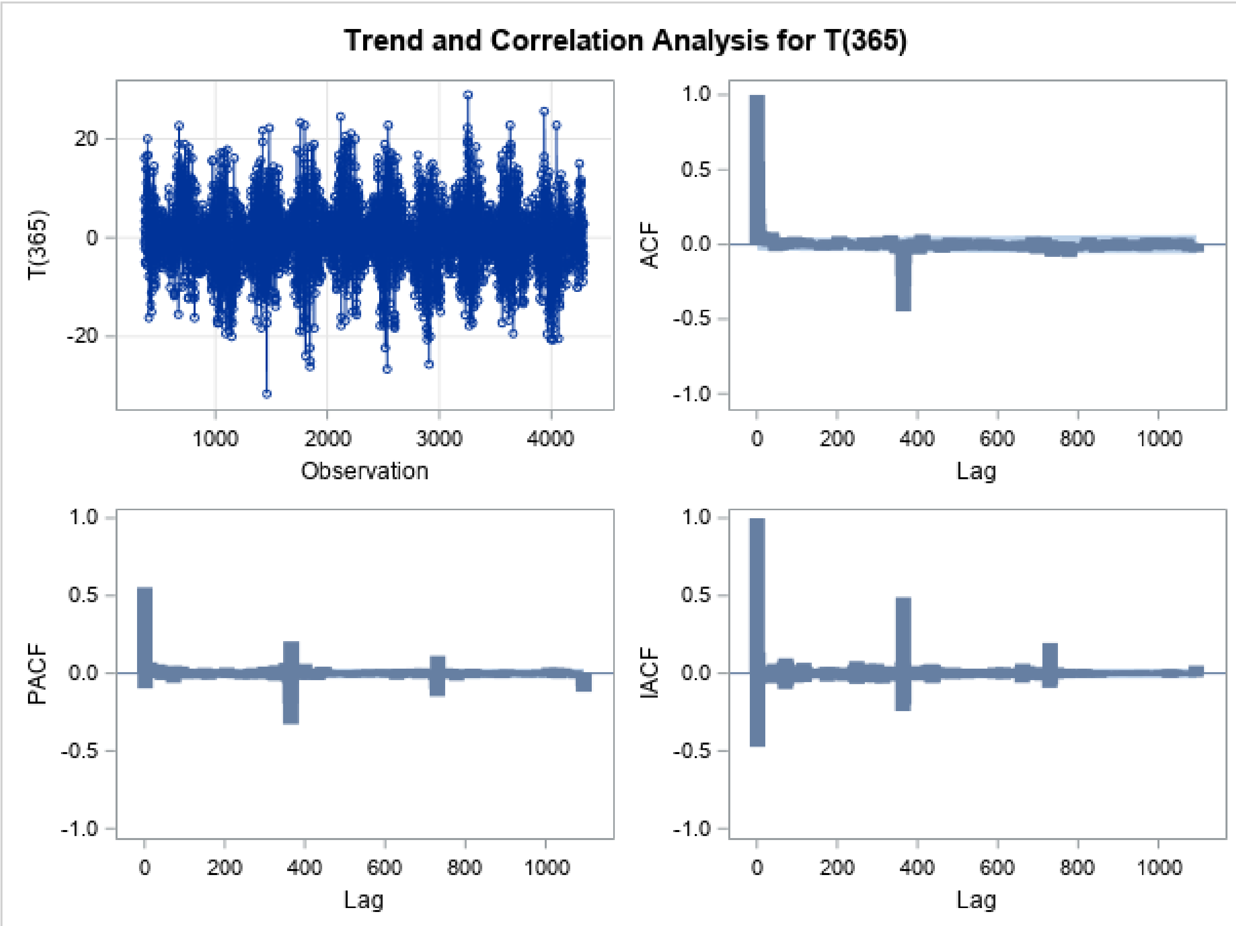


Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	0.61	1	0.4345	-0.000	-0.001	-0.005	0.009	-0.006	0.003
12	7.59	7	0.3699	-0.002	0.015	-0.023	-0.007	-0.008	-0.030
18	13.06	13	0.4430	-0.000	-0.012	0.014	0.030	0.010	-0.005

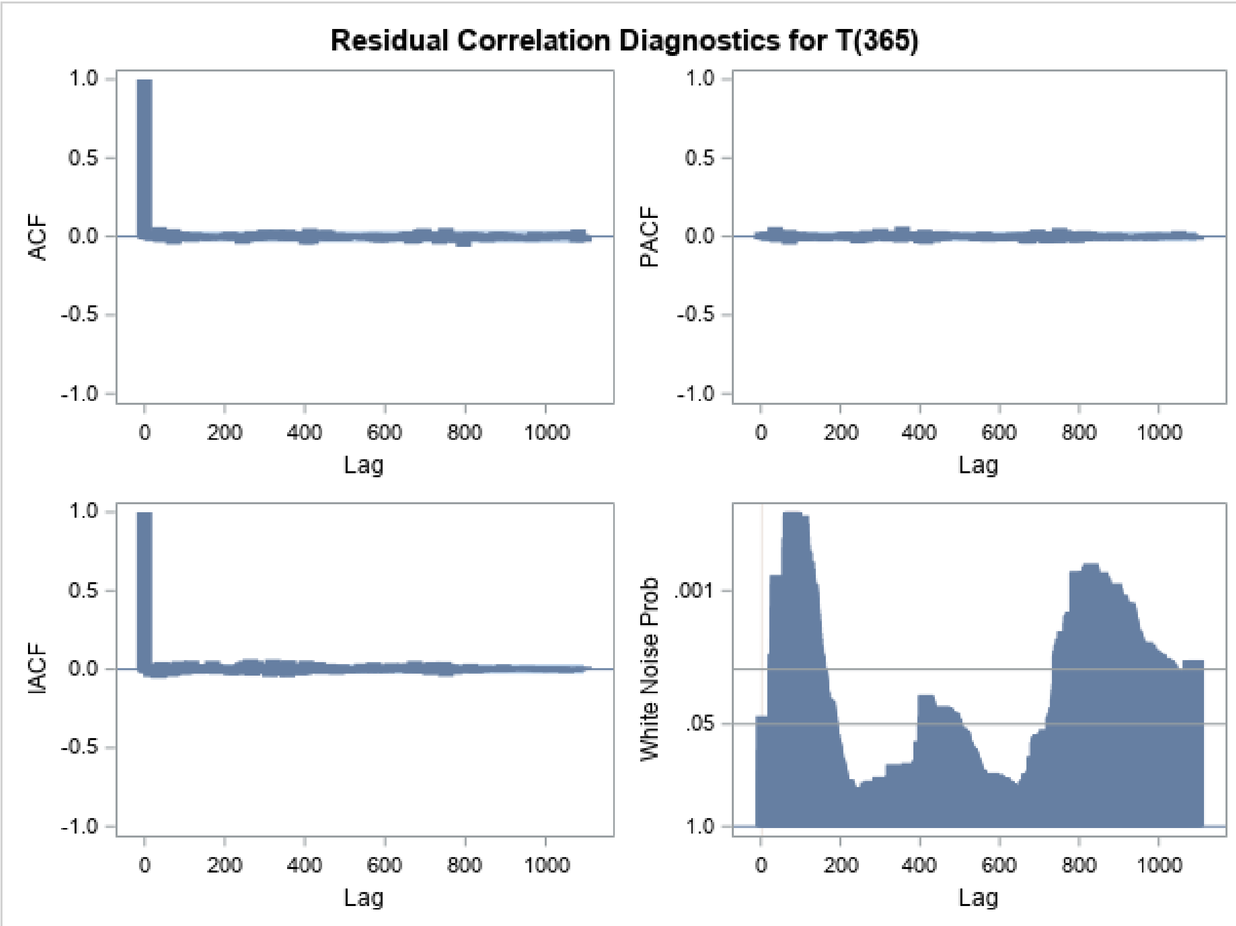
The fit is adequate with AIC: 24547.2 and SBC: 24584.85.

This final model is: [1 - 0.61171 **B**) + 0.13201 **B^2**- 0.05569 **B^3** - 0.03921 **B^6** - 0.05672 **B^13**]Zt = **A**t

Further exploration



Looking at the plots of the data when we extend the number of lags displayed to three times the seasonal component that is equal to 1,095. We can see that there is a spike in the ACF at 365. And recurring spikes in the PACF. This implies that there is an MA component at q = 365 that could be added to the model and potentially improve the previous model.

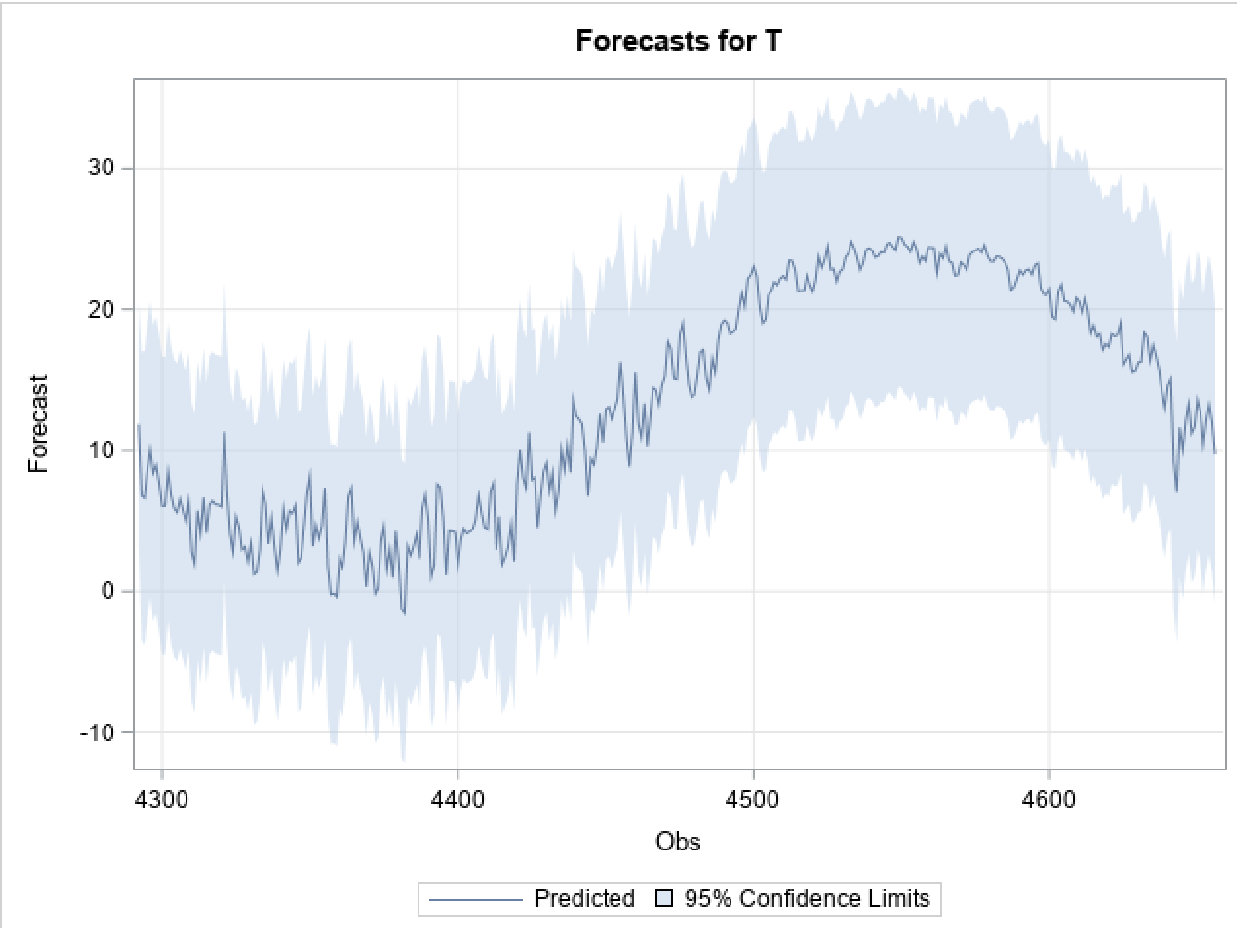


The new model is in fact an improvement upon the previous model with AIC = 22859.93 and SBC = 22903.86. These are significantly smaller than the two previous estimates. Since we are forecasting out a year it makes sense to use this more complicated model in this case.

The final estimated model is:

[1 - 0.61071 **B** + 0.13488 **B^2** - 0.06473 **B^3** - 0.04948 **B^6** - 0.04841 **B^13**]**Z**t = [1 - 0.76802 **B^365**]**A**t

Forecasting



The model is considered a seasonal ARIMA model. $ARIMA(13,0,0) \times (0,1,1)_{365}$

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \phi_6 B^6 - \phi_{13} B^{13})(1 - B^{365})\, Z_t = (1 - \Theta_1 B^{365}) a_t$$

The forecasting is given by expanding the equation above and representing $Z_t$by the past Z_t's and a_t's along with coefficient estimates.

The forecast starts at the date 11/01/2021, forecasting the next one year's daily temperature value at 7:00 AM for the Raleigh area.

We notice that the given graph has standard error that remains unchanged after certain days. probably because the ψ's remained 0 after certain forecasting points. The solid black line connects the daily temperature at 7:am, and the painted light blue area indicates the range of confidence interval(95%) or we can say it indicates the 95% forecast limits under the normal assumption of the residuals from the dataset. After we draw a vertical intersection line for each day, we could observe the black point indicating the forecast temperature and the light blue line indicating the 95% confidence interval. This is to say that we are 95% percent confident that the true temperature values lie within this range of this interval.

Limitations

1. The nlag is set to be default value in the identify statement, if we set the nlag to be 1095, we could see a clear spike at lag 365, which indicates that our model could only be a good fit to a part of the data. (up to the first 30+ lags). After fitting the seasonal ARIMA model by setting the lag to be relatively large, we could see that the AR(1,2,3,6,13) model is not a good fit for the whole data. Actually, the white noise diagnostic test after fitting AR(1,2,3,6,13) shows that the white noise after lag 30+ is completely rejected. The further exploration of the seasonal ARIMA model (residual diagnostics test) shows that the white noise is partially accepted given lag 1000.

The $ARIMA(13,0,0) \times (0,1,1)_{365}$ model is far better than the $ARIMA(13,0,0) \times (0,1,0)_{365}$ model. But the true model could be more complicated.

2. Here we only interpret the temperature based on the past values, there are definitely some factors that could closely relate to the change of temperature

3. Regarding the forecasting part, notice that the standard errors of the forecast values remain unchanged after certain lag because of the simplicity of the model. But the forecasting captures the trend.

Forecasts for variable T

Obs	Forecast	Std Error	95% Confidence Limits	
4292	9.4440	4.8330	-0.0285	18.9166
4293	2.5209	5.6620	-8.5763	13.6182
4294	0.6494	5.7786	-10.6765	11.9752
4295	5.9975	5.8156	-5.4009	17.3960
4296	7.6444	5.8335	-3.7890	19.0778
4297	7.6318	5.8399	-3.8141	19.0777
4298	11.2850	5.8530	-0.1868	22.7567
4299	5.1825	5.8655	-6.3137	16.6787
4300	5.9075	5.8712	-5.5998	17.4149
4301	9.7068	5.8736	-1.8051	21.2188
4302	12.6084	5.8747	1.0942	24.1227
4303	8.2345	5.8753	-3.2809	19.7499
4304	3.4566	5.8757	-8.0595	14.9726
4305	5.9897	5.8829	-5.5405	17.5199
4306	10.1065	5.8922	-1.4419	21.6550
4307	6.6780	5.8967	-4.8794	18.2354
4308	3.4283	5.8988	-8.1331	14.9897
4309	3.6801	5.8999	-7.8835	15.2437
4310	1.1770	5.9005	-10.3877	12.7416
4311	0.3793	5.9010	-11.1864	11.9451
4312	8.7744	5.9015	-2.7924	20.3412
4313	8.1483	5.9019	-3.4192	19.7159
4314	10.5909	5.9022	-0.9771	22.1590
4315	0.9239	5.9023	-10.6444	12.4922
4316	2.5391	5.9024	-9.0294	14.1076
4317	14.1777	5.9024	2.6091	25.7462
4318	11.4452	5.9025	-0.1235	23.0139
4319	5.3403	5.9026	-6.2286	16.9093
4320	6.5194	5.9027	-5.0498	18.0885
4321	13.6309	5.9028	2.0616	25.2002
4322	5.9491	5.9028	-5.6203	17.5184
4323	0.1088	5.9029	-11.4606	11.6782
4324	-0.1228	5.9029	-11.6922	11.4467
4325	4.2548	5.9029	-7.3147	15.8243
4326	4.5080	5.9029	-7.0615	16.0775
4327	3.1866	5.9029	-8.3830	14.7561
4328	1.9983	5.9029	-9.5713	13.5679
4329	2.3945	5.9029	-9.1751	13.9640

4. There is no clear overall trend in this dataset. Due to climate change, there may be some variations of the overall trend of the temperature if we have data that covers decades.
5. Regarding the model fitting: probably modeling the dataset with fewer data points could better capture the seasonal trend.
6. We use the median interpolation (numerical analysis) simple method to fill out some missing values, it is a reasonable method. Even though those filled missing values could barely affect the test methods we used , we still need to take this into consideration of limitations for this simple project.