Historical Weather in Santa Barbara Analysis

Alaina Liu

2023-06-14

Introduction

This project takes a look at historical weather in Santa Barbara. I am using a dataset called "Historical Weather in Santa Barbara (2015-2022)" from kaggle, with observations collected daily from May 1, 2015 to May 1, 2022. The variables contained are: Date, Average Temperature (ºF), Average Dew Point (ºF), Average Humidity (%), Maximum Wind Speed (mph), Average Wind Speed (mph), Average Pressure (inHg), Precipitation (in).

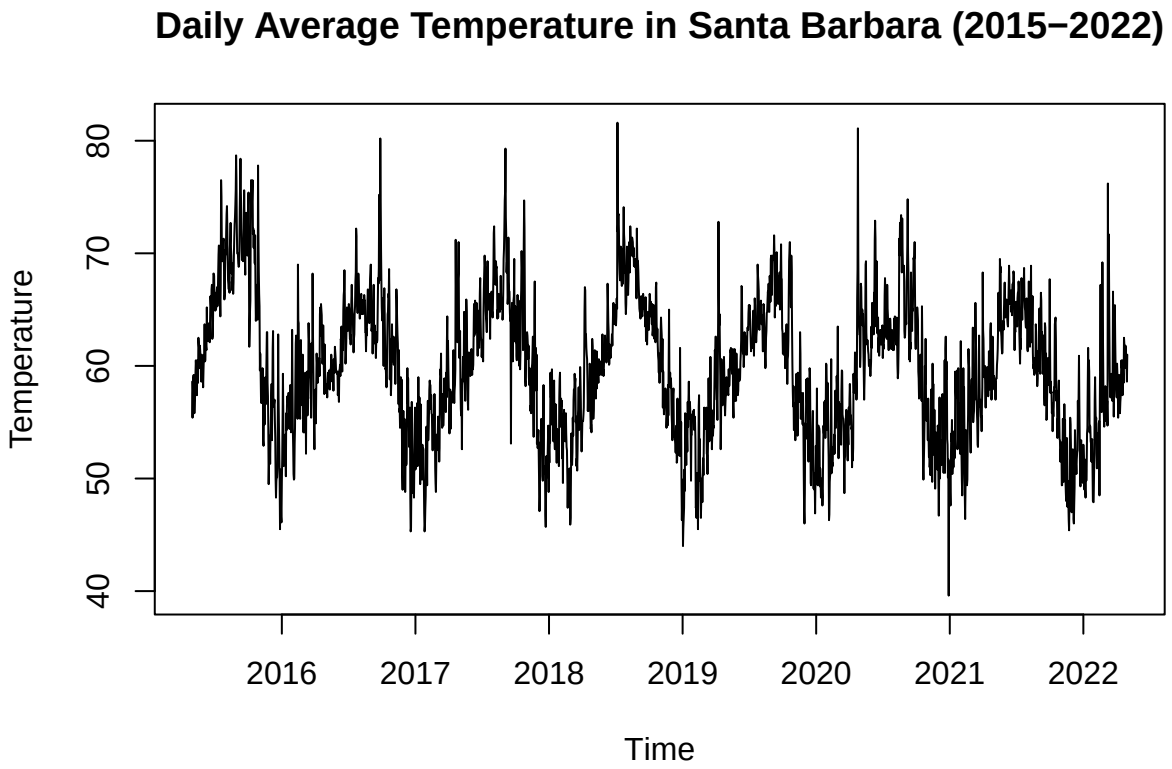
I chose this topic because it is really relevant to me as a student at UCSB. The weather this past year especially has been extremely irregular and has been a topic widely talked about in the area. At the moment, it is already June, which is a time we would expect the weather to be hot and sunny here, yet most days the temperature doesn't even hit 70 degrees Fahrenheit on most days. I wanted to explore temperature changes over the past few years so observe any trends that may explain these recent patterns.

The two models I will be applying to our data are a SARIMA (p,d,q) x (P,D,Q) model and an ARMAX model. I will be using the SARIMA model to look at the average temperature.

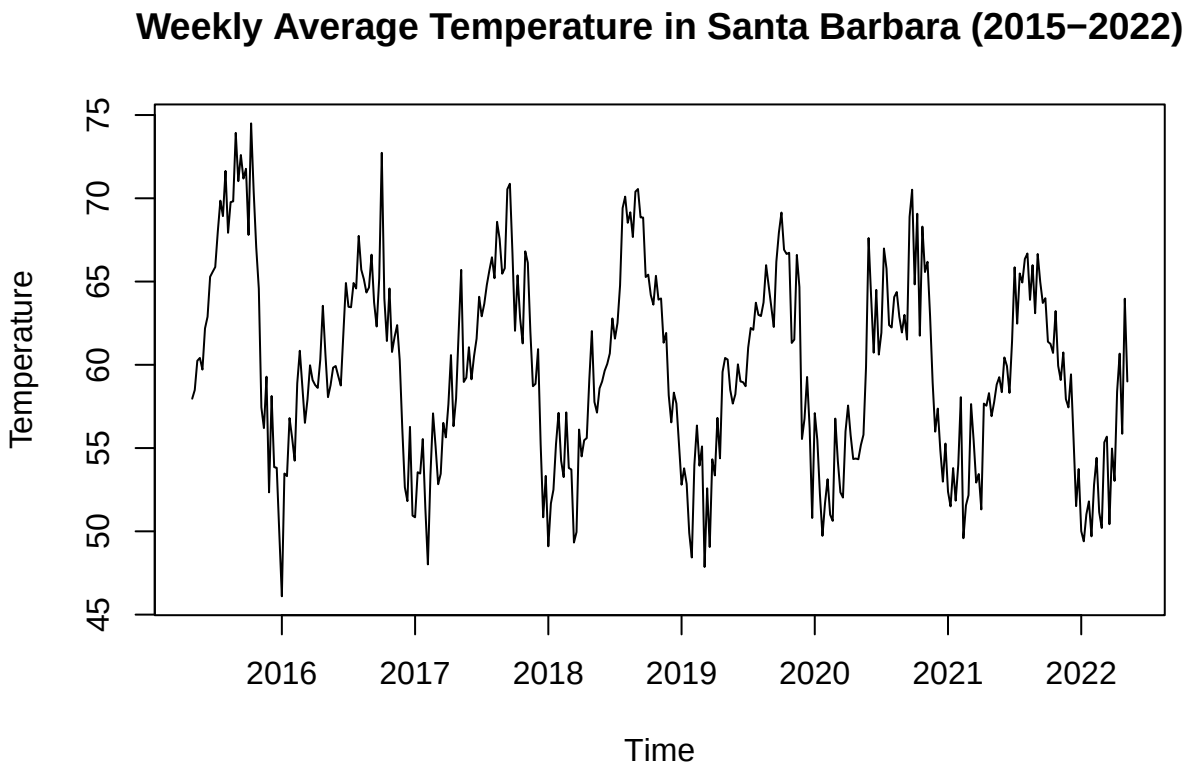
Exploring the Data

As mentioned in the previous section, the dataset contains information about the daily average temperature, average dew point, average humidity, maximum wind speed, average wind speed, average pressure, and precipitation. There are 2529 total obserations. Let's take a quick look at our dataset, and a plot of the temperature variable:

Date	Temperature	DewPoint	Humidity	MaxWind	WindSpeed	Pressure	Precipitation
5/1/15	58.6	52.7	81.6	16	6.4	29.9	0
5/2/15	55.4	49.7	82.2	18	5.5	29.9	0
5/3/15	57.5	49.4	75.1	17	7.7	29.9	0
5/4/15	59.0	49.0	70.2	14	7.2	29.9	0
5/5/15	59.2	48.3	67.7	12	6.3	29.9	0
5/6/15	58.1	45.5	64.3	17	6.4	29.9	0



Since the data was collected daily, there are a lot of data points across seven years. I will choose to take the weekly averages to make the data easier to work with.

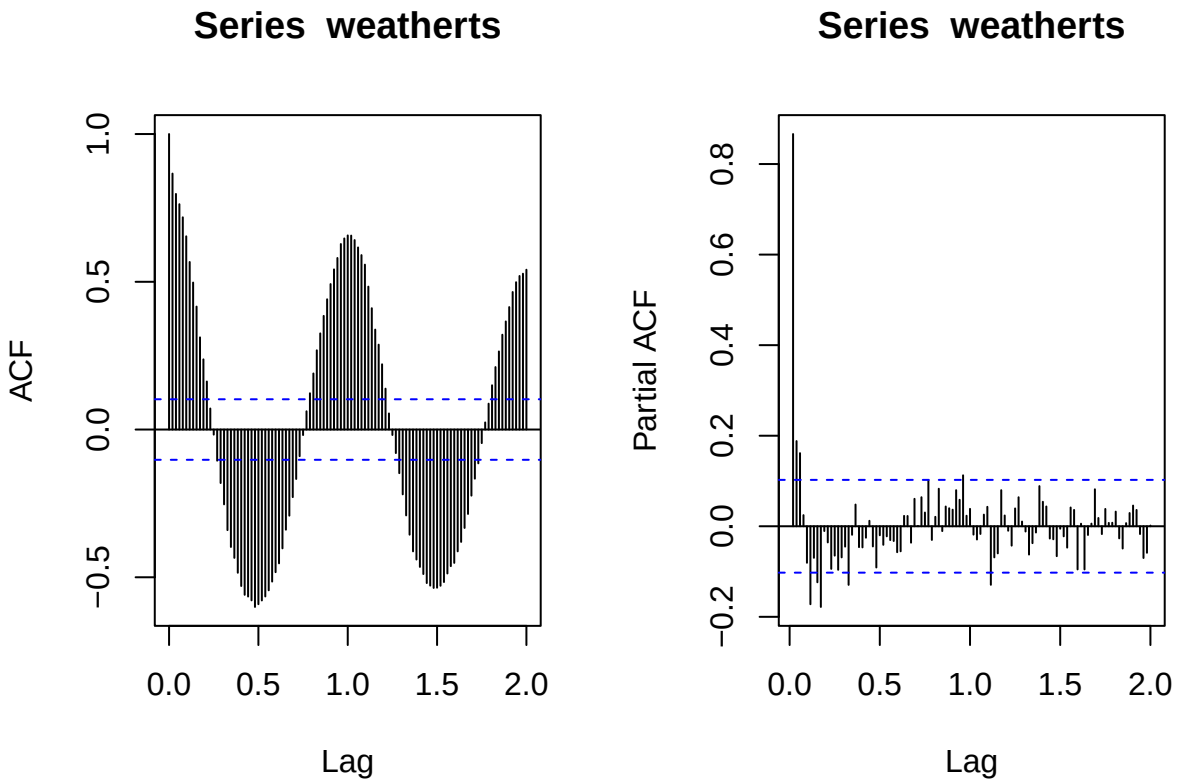


The plot above is our time series of temperature from 2015 to 2022, with weekly observations. There is clearly a seasonal trend with each cycle being one year. The temperature peaks after the midway point of each year, and dips in the beginning of each year. This is within our expectations of weather patterns. It looks like there is a slight downward drift, which is something worth looking into in our models.

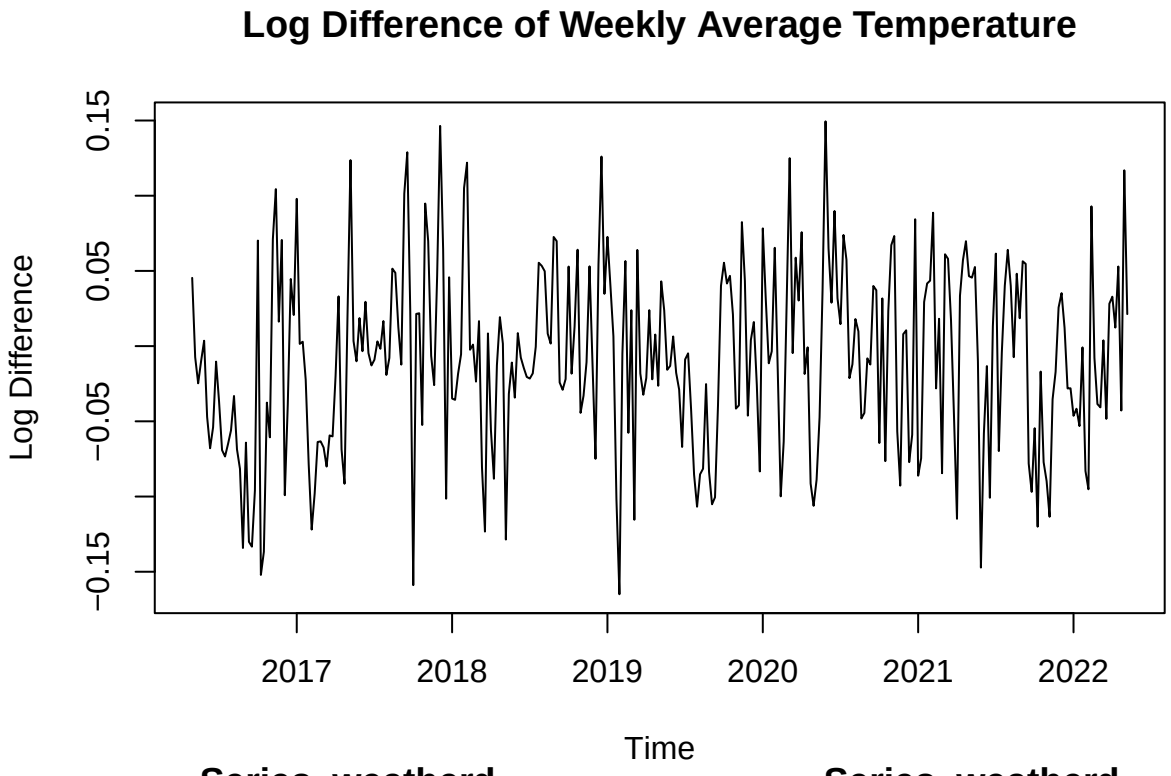
SARIMA Model

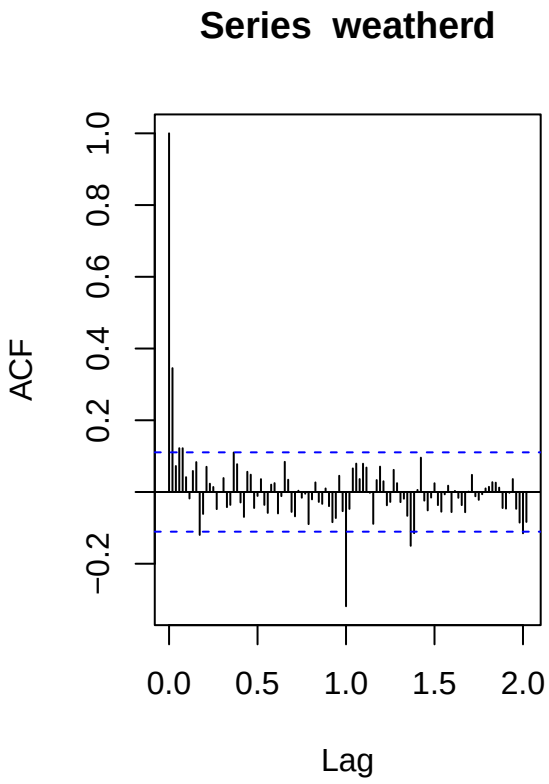
The first model I will be using is the SARIMA (p,d,q) x (P,D,Q), or a Seasonal AutoRegressive Integrated Moving Average model. This model is an extension of the ARIMA model that includes seasonal patterns, which is fitting with how seasonal the temperature data is. I will use a time series of the average temperatures as my main variable of interest for this model. The previous plots show that temperature follows a yearly pattern, therefore it is clearly not stationary. The ACF and PACF plots also depend on time. A transformation of the data is needed.

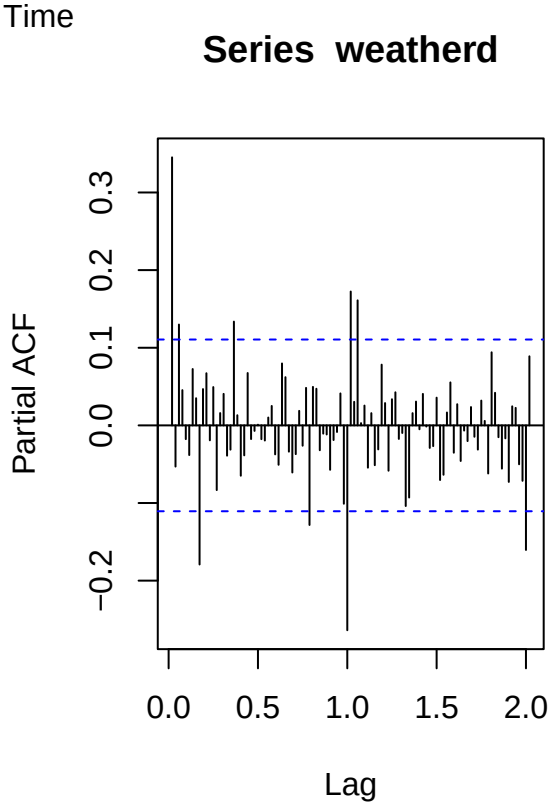
Model Fitting



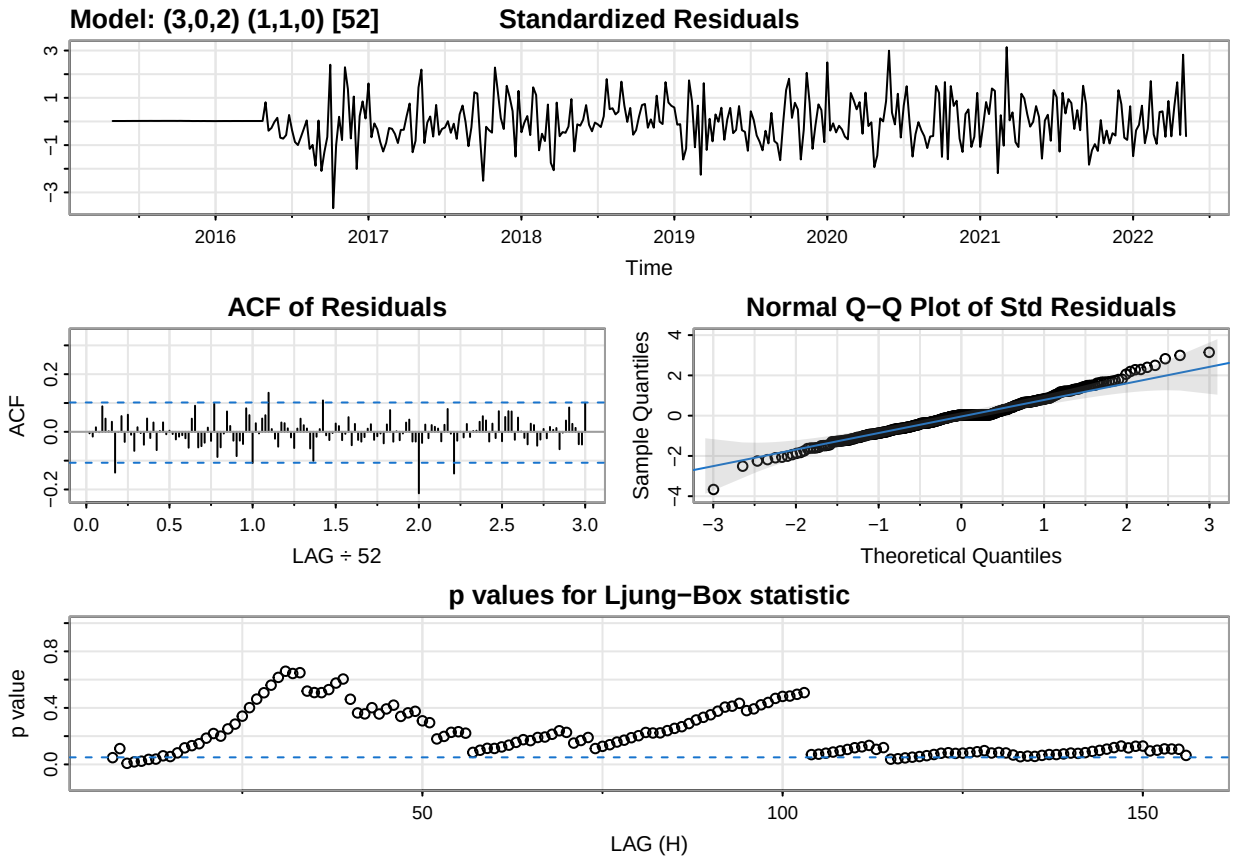
I will perform seasonal differencing to remove seasonality of the data. After looking at both the log difference and second difference, the log differenced data has lower variance, and the ACF and PACF plots show it is stationary as they do not depend on time. There are a few values of the ACF and PACF above the blue dashed lines but they can be thought of as exceptions and do not affect the stationarity of the series.







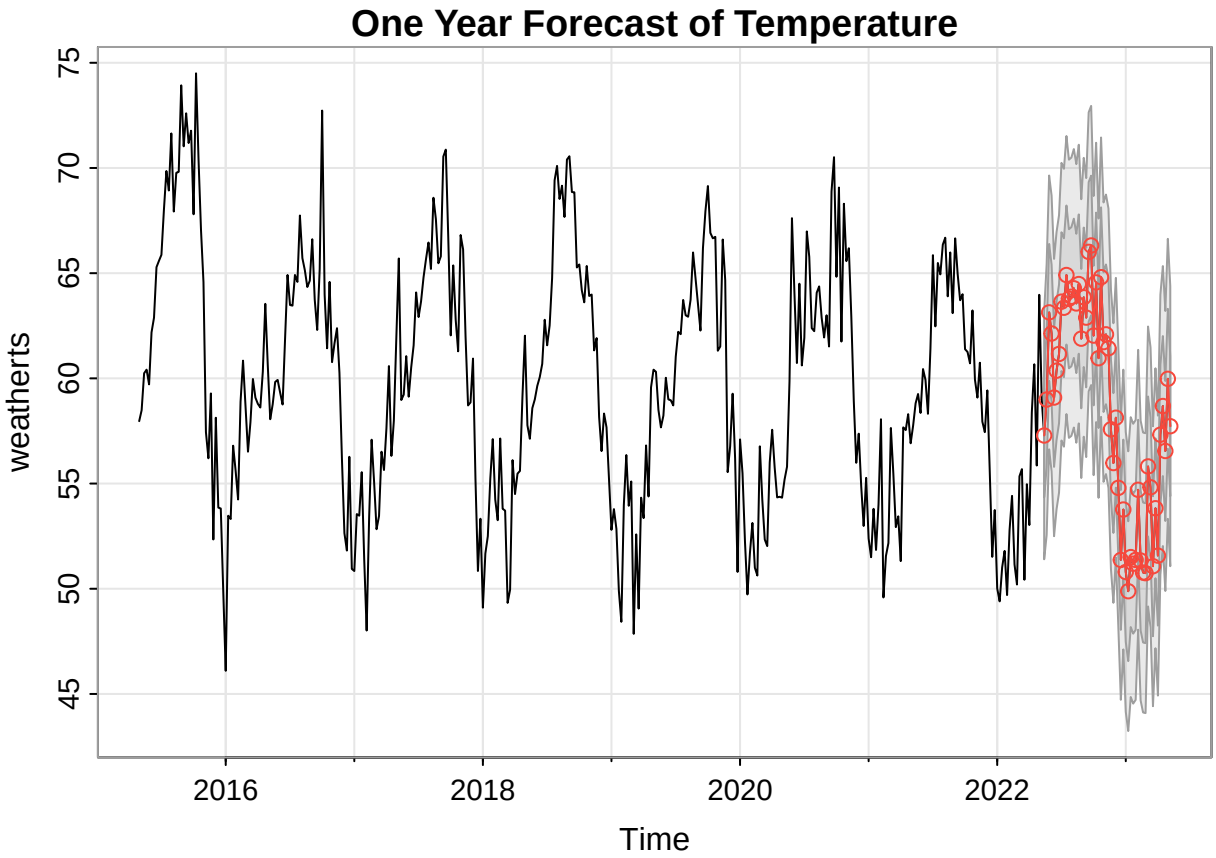
Next, I will fit a SARIMA model to find the best model by trying out different values of p, d, q, P, D, Q. I will use D = 1 since there is a seasonal difference. I will experiment of values of the other parameters by looking at the peaks of the ACF and PACF plots.



The best model here is ARIMA(3,0,2)(1,1,0)[52]. Looking at the standardized residuals plot, the data looks randomly scattered around zero and there is no pattern or trend. The ACF of residuals plot shows that the ACF also does not have any patterns or depends on time. The Normal Q-Q plot of the residuals shows that the residuals are close to Normal as most of the points lie on the diagonal line. The p-values for the Ljung-Box statistic are above 0.5, so we fail to reject the null that the model is adequate. These all suggest that the residual is white noise

SARIMA Forecasting

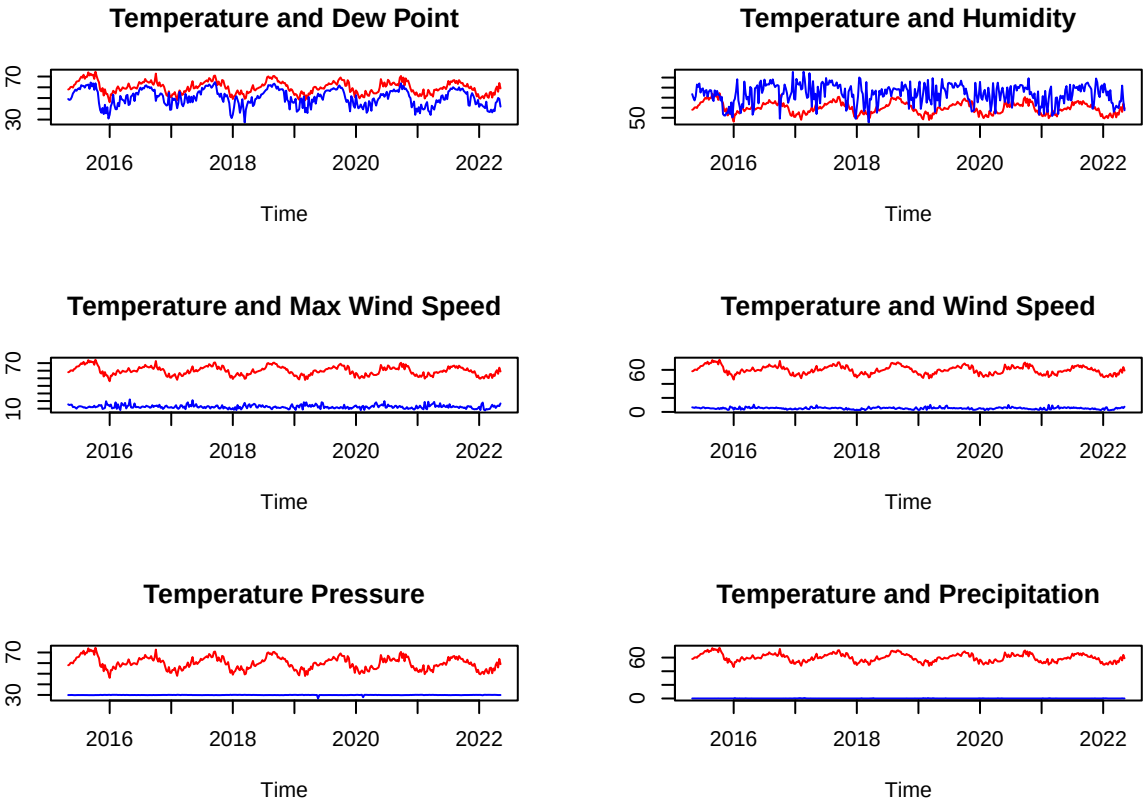
I am interested in forecasting the temperature of the next year. To do this, I will use the sarima.for function and plot the results.



Just from lookg at the visualization, the SARIMA model seems to do a good job, as the predicted values follow the yearly trend of the previous years.

ARMAX Model

Because the original dataset contains not just temperature data but also data of other variables, I am interested in looking at any possible interactions between them. An ARMAX model is based on the concept of ARMA but incorporates exogenous variables. The aim of fitting this model to explore how the other variables affect temperature. Let's take a look of plots of temperature and all of the other variables.



The plots show that certain variables like dew point seem to have a correlation with temperature. It looks like dew point and temperature follow a similar trend. The relationships of the other variables aren't as visible from the plots, but should not be disregarded.

Selection Criteria

Before fitting an ARMAX model, I want to take a look at what lag value to use for the predictors.

```
## $selection
## AIC(n)  HQ(n)  SC(n) FPE(n)
##      4      1      1      4
##
## $criteria
##                 1          2           3           4          5          6
## AIC(n) -2.2333931 -2.2643535 -2.41502143 -2.44920601 -2.3889047 -2.3003945
## HQ(n)  -1.9606193 -1.7794223 -1.71793279 -1.53995996 -1.2675013 -0.9668336
## SC(n)  -1.5476609 -1.0452742 -0.66259489 -0.16343227  0.4302162  1.0520737
## FPE(n)  0.1071723  0.1039413  0.08947824  0.08660468  0.0922151  0.1011145
##                 7          8           9          10
## AIC(n) -2.2316834 -2.1639080 -2.1119570 -2.0579997
## HQ(n)  -0.6859652 -0.4060323 -0.1419239  0.1241908
```

```
## SC(n)   1.6541319  2.2552545  2.8405528  3.4278572
## FPE(n)  0.1088492  0.1172556  0.1245597  0.1328655
```

I used the VARselect function to choose the best lag to use for the model. AIC and FPE both pick p=4, and HQ and BIC both pick p=1. I will go with p=1 and fit my ARMAX model, starting with using all other variables.

Model Fitting

```
##
## Call:
## lm(formula = y ~ -1 + ., data = datamat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4217 -1.5163 -0.2513  1.5033  8.4396
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## weatherts.l1       -0.088601   0.211799  -0.418 0.675961
## dewpointts.l1       0.900338   0.221645   4.062 5.99e-05 ***
## windspeedts.l1      0.131781   0.224082   0.588 0.556844
## humidityts.l1      -0.294129   0.102289  -2.875 0.004277 **
## maxwindts.l1       -0.058793   0.134726  -0.436 0.662821
## pressurets.l1       0.572197   0.643080   0.890 0.374187
## precipitationts.l1 -6.170098   1.737422  -3.551 0.000435 ***
## const              25.198586  21.671781   1.163 0.245715
## trend              -0.002489   0.001349  -1.845 0.065854 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.624 on 356 degrees of freedom
## Multiple R-squared:  0.7987, Adjusted R-squared:  0.7942
## F-statistic: 176.5 on 8 and 356 DF,  p-value: < 2.2e-16
```

First, I fit the model using all the other variables. From the coefficients, only dew point, huidity, and precipitation with lag 1 are significant predictors for temperature, as they have p-values less than 0.05 while the other variables have higher p-values. Therefore, for the next model I will only use those three as predictors.

```
##
## Call:
## lm(formula = y ~ -1 + ., data = datamat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4032 -1.4921 -0.2588  1.5377  8.4237
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## weatherts.l1       -0.111550   0.201093  -0.555  0.57943
## dewpointts.l1       0.926608   0.207440   4.467 1.06e-05 ***
```

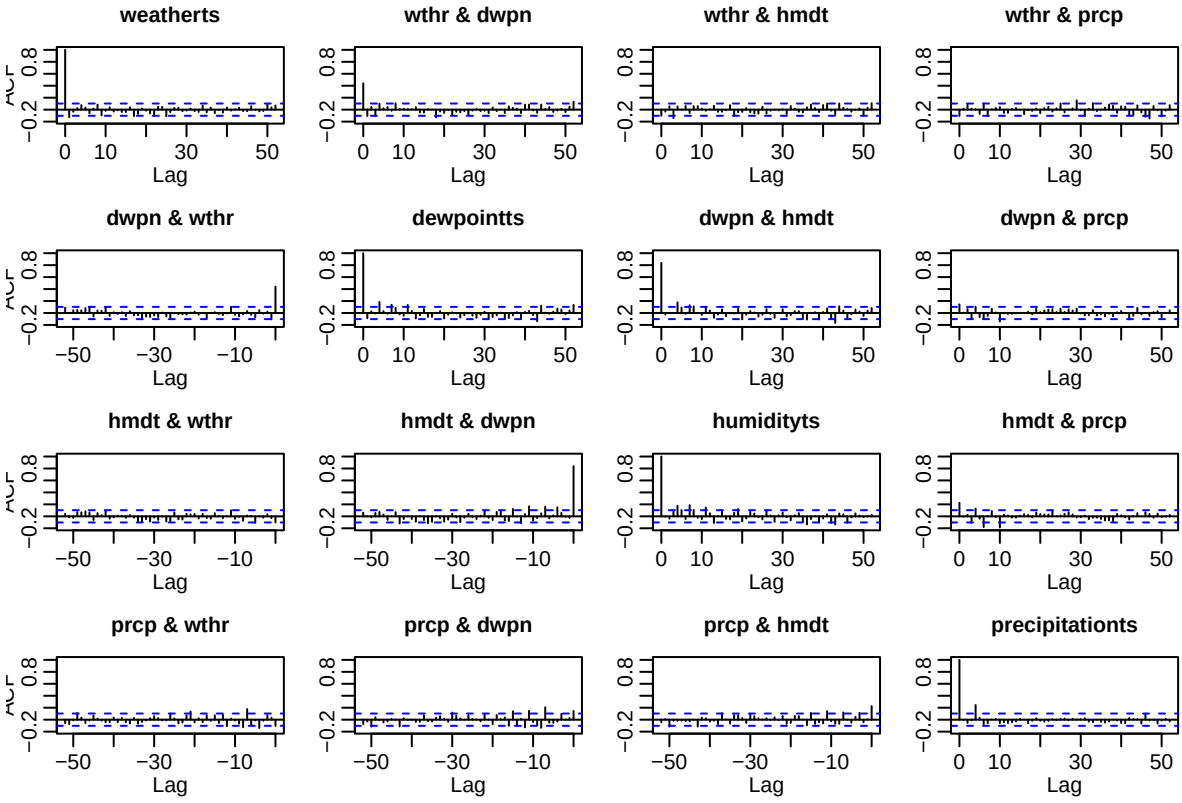
```
## humidityts.l1      -0.305826   0.095026  -3.218  0.00141 **
## precipitationts.l1 -6.094675   1.632722  -3.733  0.00022 ***
## const              43.194961   8.611787   5.016 8.32e-07 ***
## trend              -0.002521   0.001340  -1.882  0.06069 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.617 on 359 degrees of freedom
## Multiple R-squared:  0.798,  Adjusted R-squared:  0.7952
## F-statistic: 283.7 on 5 and 359 DF,  p-value: < 2.2e-16
```

ARMAX Model Results

The fitted model gives the following prediction equation:

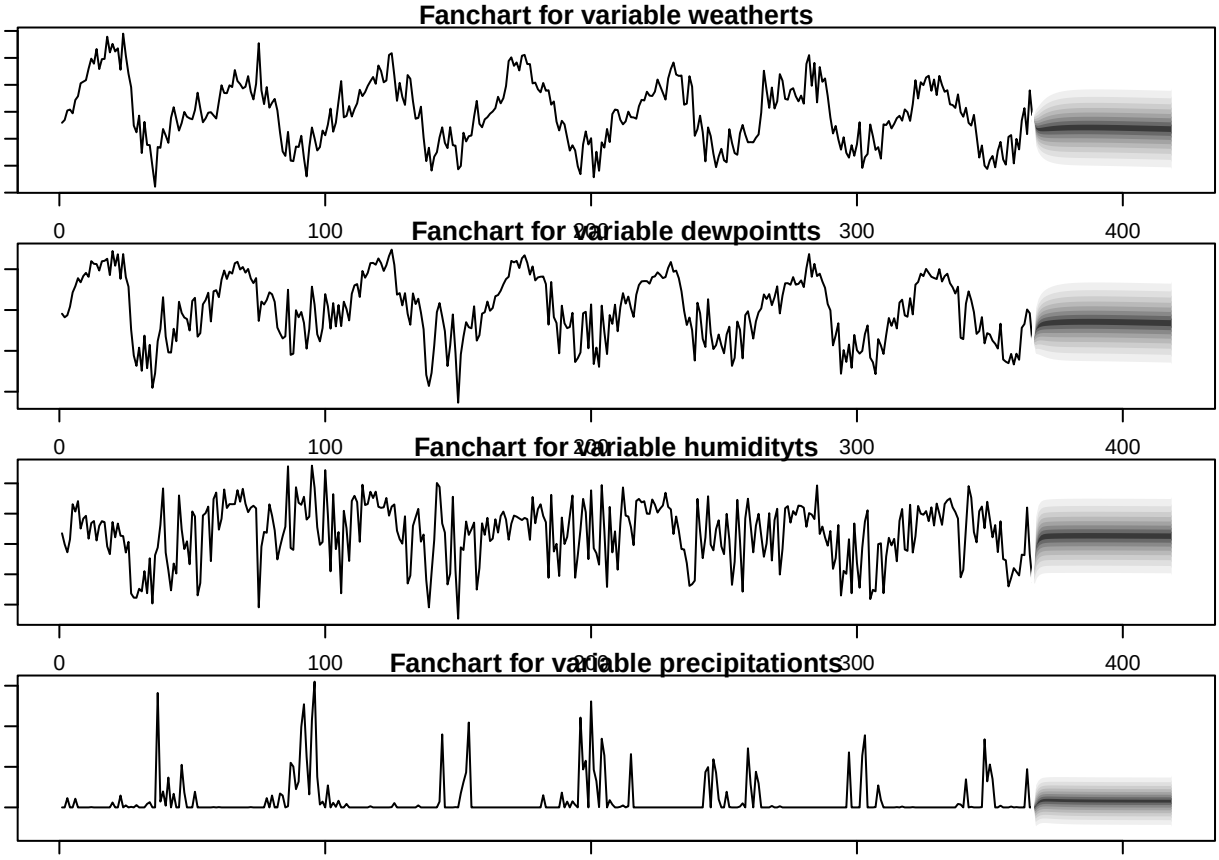
$$\hat{T}_t = 43.19 - 0.0025t + 0.92D_{t-1} - 0.31H_{t-1} - 6.09P_{t-1}$$

Where T_t, D_t, H_t, and P_t denote the temperature, dew point, humidity, and precipitation, respectively. The R^2 goodness of fit value of 0.798 means that approximately 80% of the variation in temperature can be explained by dew point, humidity, and precipitation. This is a pretty high value, suggesting that the model does well capturing the relationship between the predictors and temperature.




```
##
##  Portmanteau Test (adjusted)
##
## data:  Residuals of VAR object var_model
## Chi-squared = 390.48, df = 176, p-value < 2.2e-16
```

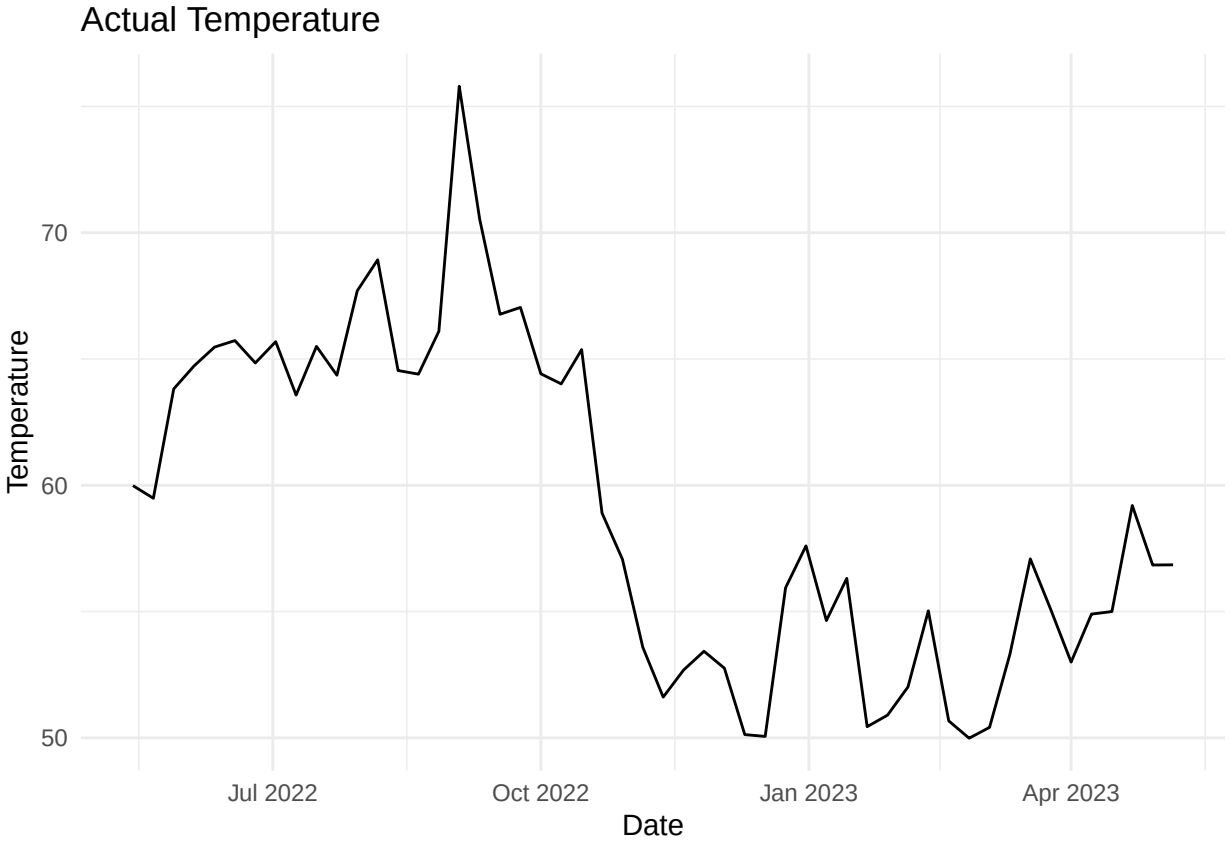
Upon examining the ACF and CCF of the residuals of the ARMAX model, most do not have correlation. However, some plots, like the one between humidity and dewpoint, appear to have non-zero correlation. So it is not plausible say that the noise is white. The adjusted Postmanteau test results in a small p-value, so I reject the null that the autocorrelations in the residuals are not different from zero. My ARMAX model still has some autocorrelation and may need to be further improved upon to accurately predict temperature.



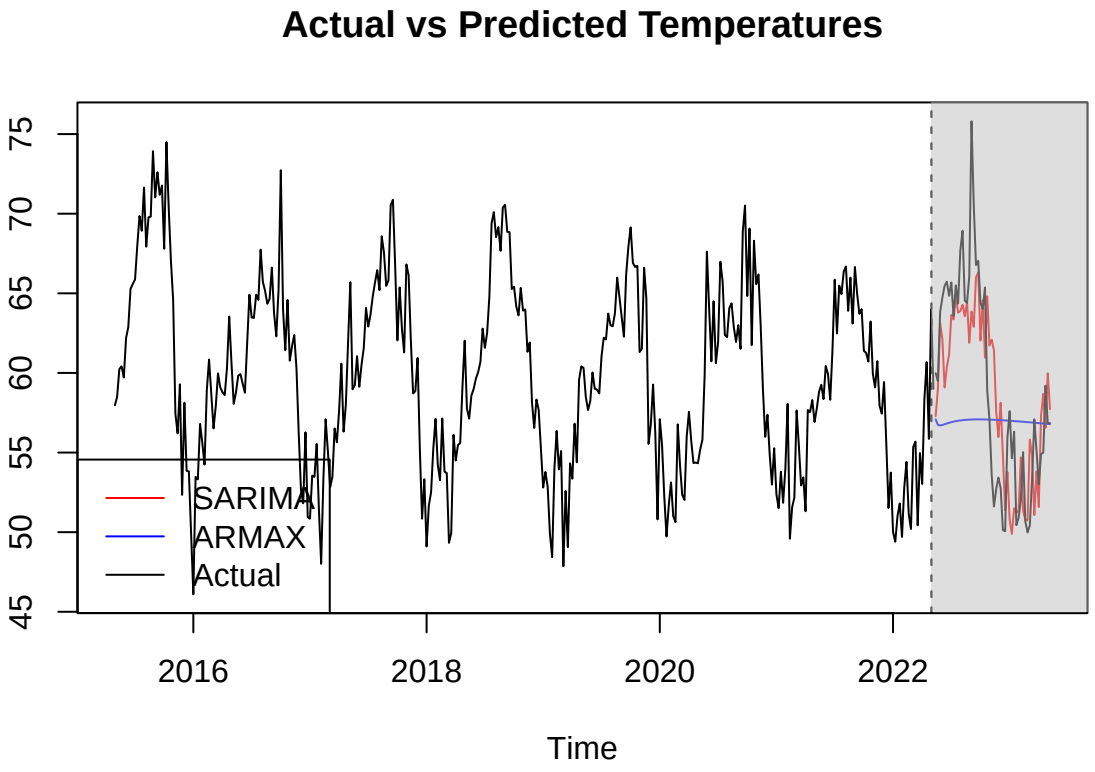
The plots above show the forecasted values for all the variables used in my ARMAX model. In particular, looking at the temperature graph, the model's predictions are a constant flat trend, unlike the SARIMA model which its forecasted values followed the same cyclical trend as the data. Therefore, the ARMAX model did not produce the best results.

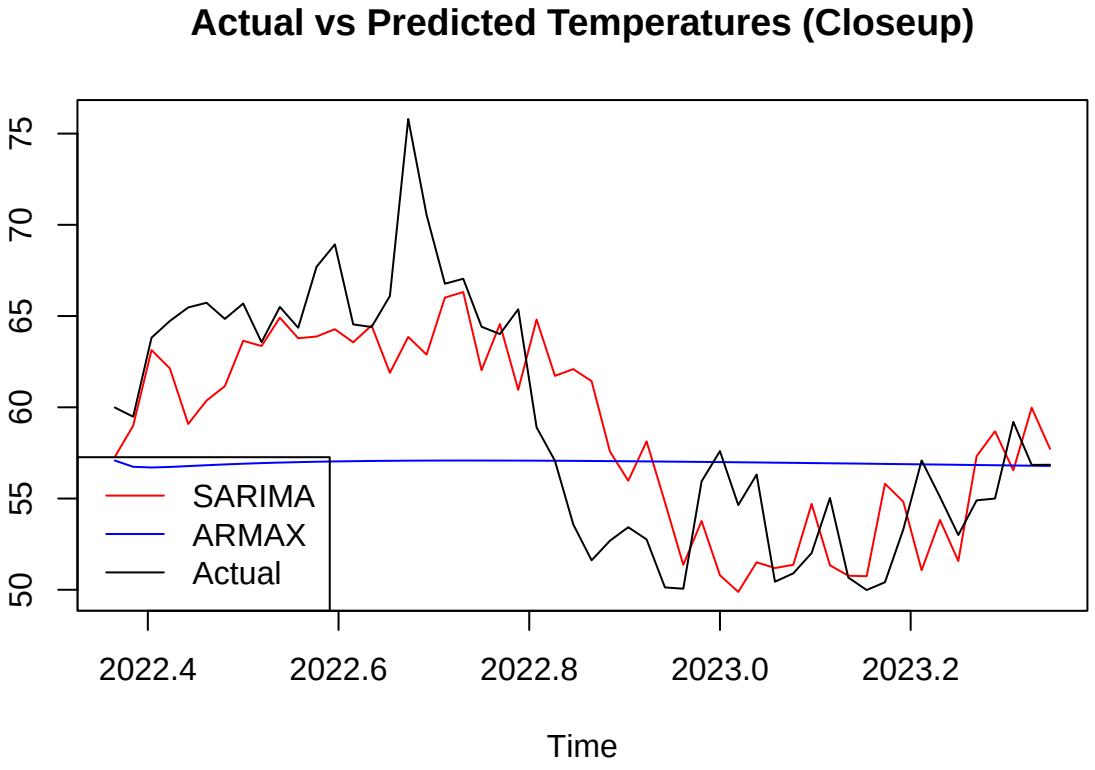
Comparing results

After forecasting temperature from one year after May 1, 2022, I am interested in comparing the forecasted values to the true temperatures of the past year. I collected temperature data from Visual Crossing, which contained daily temperature values as well as many other variables, and tidied it up the same way as before.



Let's look at a plot of our predicted values from both models and how they stack up against the actual data:


Actual vs Predicted Temperatures
75
70
65
60
55
50
45
SARIMA
ARMAX
Actual
2016
2018
2020
2022
Time


Actual vs Predicted Temperatures (Closeup)
75
70
65
60
55
50
SARIMA
ARMAX
Actual
2022.4
2022.6
2022.8
2023.0
2023.2
Time

Conclusion

The results show that my SARIMA model did a great job with forecasting temperatures values. The black (actual temperatures data) and red (SARIMA forecasted temperatures) lines follow the same pattern, with the temperature peaking in after the halfway point of 2022 and dipping in the end and beginning of 2023. On the other hand, my ARMAX model did not produce a good forecast. It produced a roughly flat line around the mean, but did not take into consideration any of the seasonal trends present in the actual data.

These results are quite interesting. I learned that the SARIMA model I used is able to predict my seasonal temperature data really well, but the ARMAX model still needed some work. For future study, I might want to consider changing up the selection criteria and/or the variables I used in the model. I used a lag of 1, but a different lag may have worked better. As there was a noticeable relationship between dew point and temperature, I had wanted to try using a model that involves another variable, but the result was not satisfactory. It may be worth experimenting with different models to improve forecasting.

References

My dataset “Historical Weather in Santa Barbara (2015-2022)” is from kaggle.

Time Series Analysis and its Applications with R Examples by R.Links to an external site. H. Shumway and D. S. Stoffer

My dataset of observations from May 2022 to May 2023 if from Visual Crossing. https://www.visualcrossing.com/weather/weather-data-services#

Appendix

```
weather <- read.csv("/Users/alainaliu/Downloads/PSTAT 174/SB Weather Data/weatherdatanew.csv")

#changing column names for better readability
colnames(weather) <- c("Date", "Temperature", "DewPoint", "Humidity", "MaxWind", "WindSpeed", "Pressure
# sorting date
weather <- weather %>%
  arrange(mdy(weather$Date))
head(weather) %>% kable()

weatherts1 <- ts(weather$Temperature, start=c(2015, 121), end=c(2022, 121), frequency=365)
plot(weatherts1)

weather2 <- weather
weather2$Year <- year(mdy(weather2$Date))
weather2$Week <- week(mdy(weather2$Date))

weatherweek <- as.data.frame(weather2 %>%
  group_by(Year, Week) %>%
  summarize(AvgTemp=mean(Temperature),
            AvgDewPoint=mean(DewPoint),
            AvgHumidity=mean(Humidity),
            AvgMaxWind=mean(MaxWind),
            AvgWindSpeed=mean(WindSpeed),
            AvgPressure=mean(Pressure),
            AvgPrecipitation=mean(Precipitation)))
weatherweek %>% head() %>% kable()

weatherts <- ts(weatherweek$AvgTemp, start=c(2015, 18), end=c(2022, 19), frequency=52)
dewpointts <- ts(weatherweek$AvgDewPoint, start=c(2015, 18), end=c(2022, 19), frequency=52)
humidityts <- ts(weatherweek$AvgHumidity, start=c(2015, 18), end=c(2022, 19), frequency=52)
maxwindts <- ts(weatherweek$AvgMaxWind, start=c(2015, 18), end=c(2022, 19), frequency=52)
windspeedts <- ts(weatherweek$AvgWindSpeed, start=c(2015, 18), end=c(2022, 19), frequency=52)
pressurets <- ts(weatherweek$AvgPressure, start=c(2015, 18), end=c(2022, 19), frequency=52)
precipitationts <- ts(weatherweek$AvgPrecipitation, start=c(2015, 18), end=c(2022, 19), frequency=52)

plot(weatherts)

par(mfrow=c(1,2))
acf(weatherts, 104); pacf(weatherts, 104)

# taking the log difference of the data to make it more stationary
weatherd <- diff(log(weatherts), 52) # lag of one year
weatherd2 <- diff(weatherd, 52)
var(weatherd); var(weatherd2)
# weatherd has lower variance

plot(weatherd)
par(mfrow=c(1,2))
acf(weatherd, 105); pacf(weatherd, 105) # acf and pacf with lag of two years

#auto.arima(weatherts)
```

```r
# model predicted from auto.arima()
sarima(weatherts, 2,0,2,1,1,0, S=52, details=F)
# testing other models
sarima(weatherts, 3,0,1,0,1,0, S=52, details=F)
sarima(weatherts, 2,0,1,1,1,0, S=52, details=F)
sarima(weatherts, 3,0,2,1,1,0, S=52, details=F)

sarima_pred <- sarima.for(weatherts, n.ahead=52, 3, 0, 2, 1, 1, 0, 52, plot.all=T)

par(mfrow=c(2,1))
ts.plot(weatherts, dewpointts, gpars=list(col=c("red", "blue")))
title("Temperature and Dew Point")
ts.plot(weatherts, humidityts, gpars=list(col=c("red", "blue")))
title("Temperature and Humidity")
ts.plot(weatherts, maxwindts, gpars=list(col=c("red", "blue")))
title("Temperature and Max Wind Speed")
ts.plot(weatherts, windspeedts, gpars=list(col=c("red", "blue")))
title("Temperature and Wind Speed")
ts.plot(weatherts, pressurets, gpars=list(col=c("red", "blue")))
title("Temperature Pressure")
ts.plot(weatherts, precipitationts, gpars=list(col=c("red", "blue")))
title("Temperature and Precipitation")

x = cbind(weatherts, dewpointts, windspeedts, humidityts, maxwindts, pressurets, precipitationts)
VARselect(x, lag.max=10, type='both')

var_model1 <- VAR(x, p=1, type='both')
summary(var_model1)$varresult$weatherts

x2 = cbind(weatherts, dewpointts, humidityts, precipitationts)
var_model <- VAR(x2, p=1, type='both')
summary(var_model)$varresult$weatherts

acf(resid(var_model), 52)
serial.test(var_model, lags.pt=12, type="PT.adjusted")

armax_predict <- predict(var_model, n.ahead=52, ci=0.95)
par(mar = c(1, 1, 1, 1))
fanchart(armax_predict)
armax_pred <- ts(armax_predict$fcst$weatherts[,1], start=c(2022, 20), end=c(2023, 19), frequency=52)

weather2023 <- read.csv("/Users/alainaliu/Downloads/PSTAT 174/SB Weather Data/Santa Barbara Weather 202

weather2023$Year <- year(ymd(weather2023$datetime))
weather2023$Week <- week(ymd(weather2023$datetime))

weather2023week <- as.data.frame(weather2023 %>%
  group_by(Year, Week) %>%
  summarize(AvgTemp=mean(temp)))[-c(1,2),]

weather2023ts <- ts(weather2023week$AvgTemp, start=c(2022, 20), end=c(2023, 19), frequency=52)
plot(weather2023ts)
weather2023ts %>%
```

```
  ggplot(aes(x=seq(from=as.Date("2022-05-14"), by="week", length.out=52), y=weather2023ts)) +
  geom_line(color="black") +
  labs(x="Date", y="Temperature", title="Actual Temperature") +
  theme_minimal()

armax_predict1 <- predict(var_model1, n.ahead=52, ci=0.95)
par(mar = c(1, 1, 1, 1))
fanchart(armax_predict1)
armax_pred1 <- ts(armax_predict1$fcst$weatherts[,1], start=c(2022, 20), end=c(2023, 19), frequency=52)

ts.plot(sarima_pred$pred, armax_pred, weather2023ts, weatherts, col=c("red", "blue", "black", "black"))
abline(v=2022.33, lty=2)
rect(xleft=2022.33, xright=2024, ybottom=45, ytop=77, border=NA, col=adjustcolor("grey", alpha=0.5))
title("Actual vs Predicted Temperatures")
legend("bottomleft", legend=c("SARIMA", "ARMAX", "Actual"), col=c("red", "blue", "black"), lty=1)

ts.plot(sarima_pred$pred, armax_pred, weather2023ts, col=c("red", "blue", "black"))
title("Actual vs Predicted Temperatures (Closeup)")
legend("bottomleft", legend=c("SARIMA", "ARMAX", "Actual"), col=c("red", "blue", "black"), lty=1)
```