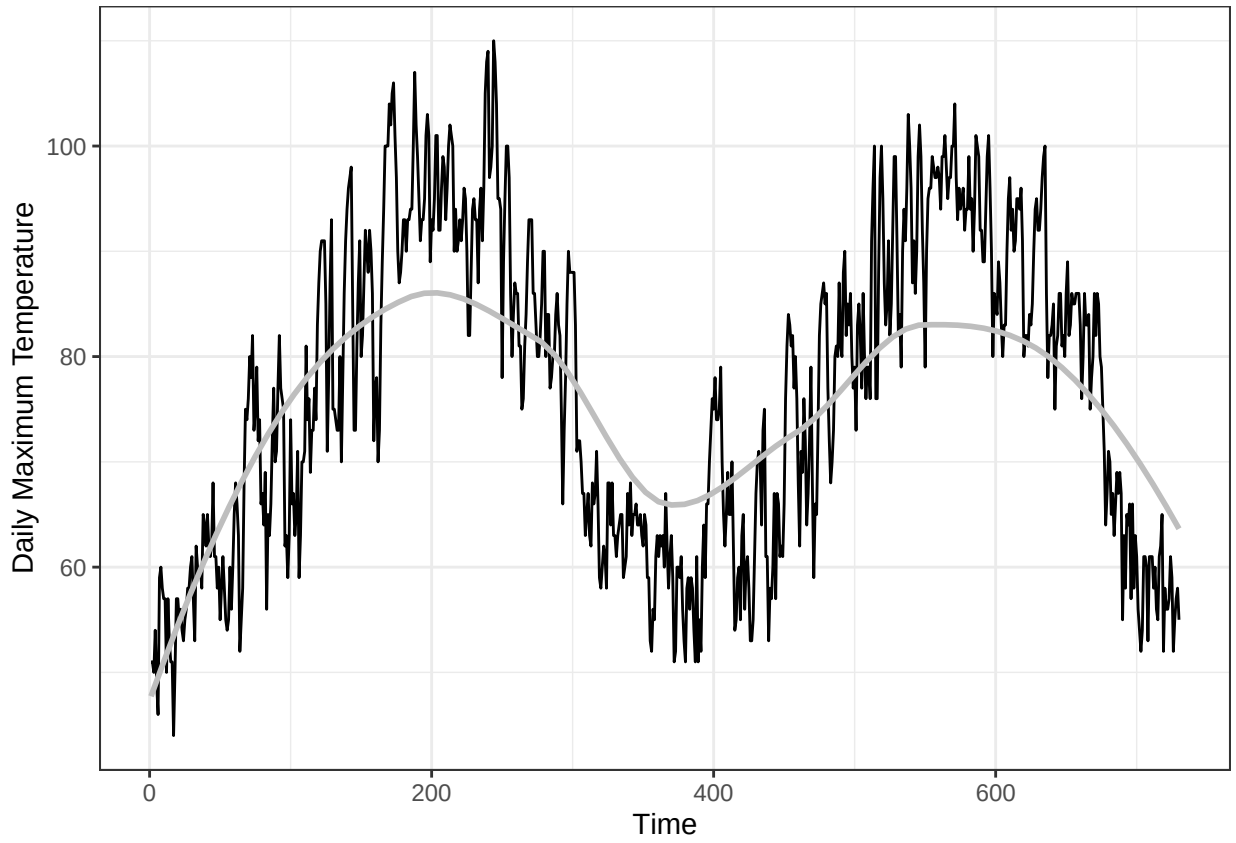
ECN190 Homework 3 Computer Problems

Kevin Chen (914861432) John Mayhew (914807483)

4/3/2020

1. Plot out the daily maximum temperature data of the two years. Explain why this time series is not stationary.



The maxtemp data is not stationary because the expected value of maxtemp changes over time (t); from time 0 to 200, maxtemp increases over time, from 200-400 maxtemp decreases over time, and the pattern repeats from 400-600 and 600-730. The expected value of temp is not constant for all levels of t, meaning the ZCM is not satisfied.

2. The daily maximum temperature data plotted out above has apparent seasonality. Now, regress daily maximum temperature data on monthly dummies.

```
##
## Call:
## lm(formula = maxtemp ~ t + Feb + Mar + Apr + May + Jun + Jul +
##     Aug + Sep + Oct + Nov + Dec, data = DavisWeather)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.7678  -4.0945  -0.1957   3.7832  20.2955
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 55.977784   0.880435  63.580  < 2e-16 ***
## t            0.001331   0.001339   0.994   0.3206
## Feb          6.718808   1.219068   5.511 4.97e-08 ***
## Mar         10.953744   1.189672   9.207  < 2e-16 ***
## Apr         16.255630   1.202881  13.514  < 2e-16 ***
## May         26.549989   1.197869  22.164  < 2e-16 ***
## Jun         34.574455   1.213739  28.486  < 2e-16 ***
## Jul         39.581717   1.211529  32.671  < 2e-16 ***
## Aug         36.346916   1.220508  29.780  < 2e-16 ***
## Sep         33.402027   1.240149  26.934  < 2e-16 ***
## Oct         25.620579   1.242042  20.628  < 2e-16 ***
## Nov         10.070852   1.263989   7.968 6.37e-15 ***
## Dec          3.071662   1.268478   2.422   0.0157 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.609 on 717 degrees of freedom
## Multiple R-squared:  0.8099, Adjusted R-squared:  0.8067
## F-statistic: 254.5 on 12 and 717 DF,  p-value: < 2.2e-16
```

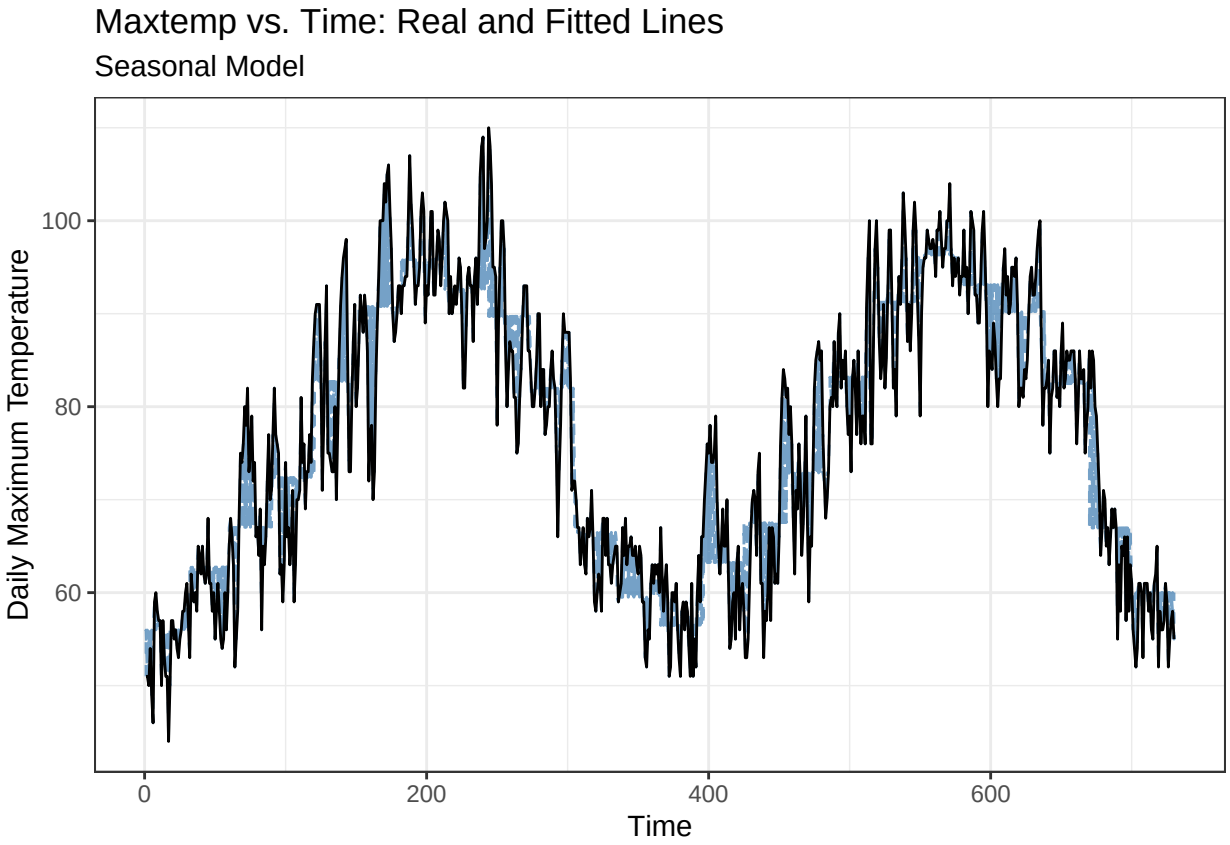
If you want to formally test whether the time series has seasonality, how would you form your null hypothesis?

H_O: $\beta_2 + \beta_3 + \beta_4 + \ldots + \beta_{12} = 0$

H_A: $\beta_2 + \beta_3 + \beta_4 + \ldots + \beta_{12} \neq 0$

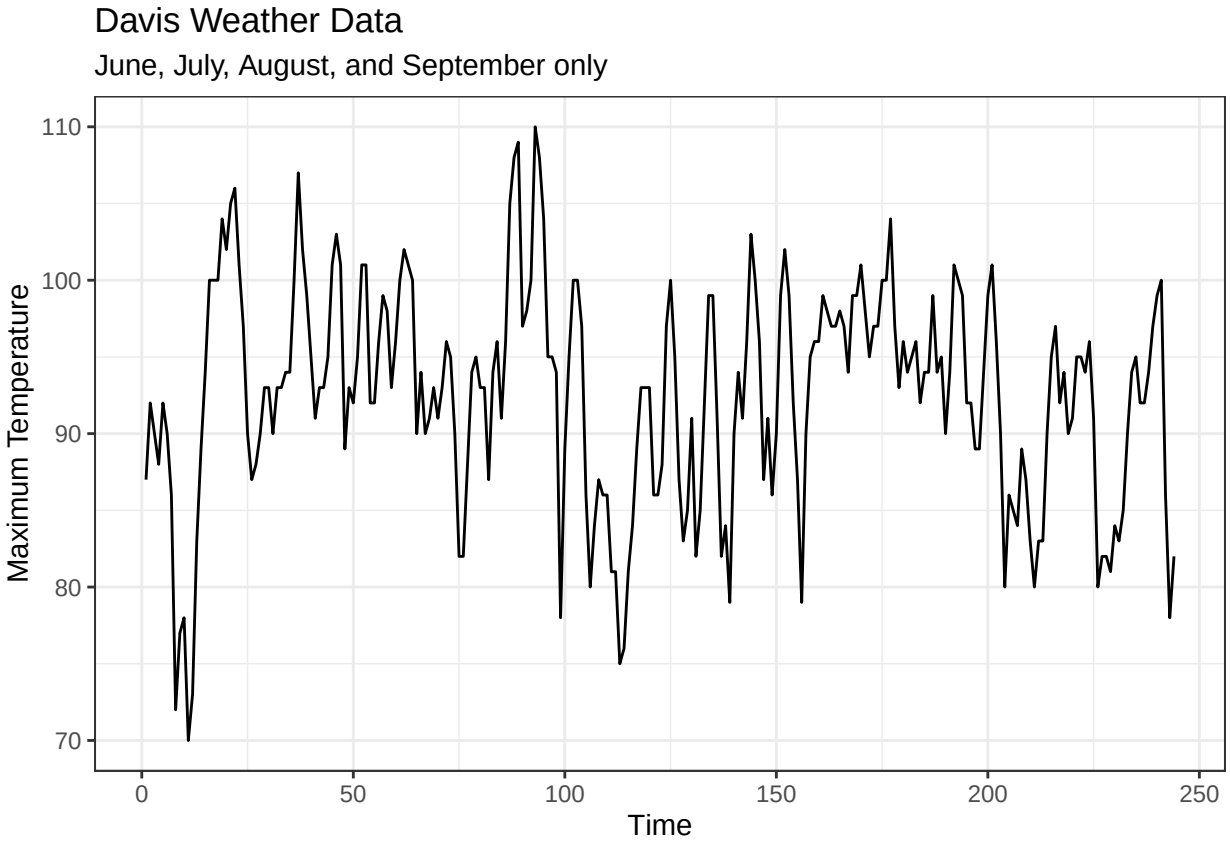
```
## Linear hypothesis test
##
## Hypothesis:
## Feb  + Mar  + Apr  + May  + Jun  + Jul  + Aug  + Sep  + Oct  + Nov  + Dec = 0
##
## Model 1: restricted model
## Model 2: maxtemp ~ t + Feb + Mar + Apr + May + Jun + Jul + Aug + Sep +
##     Oct + Nov + Dec
##
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    718 57067
## 2    717 31320  1     25748 589.44 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

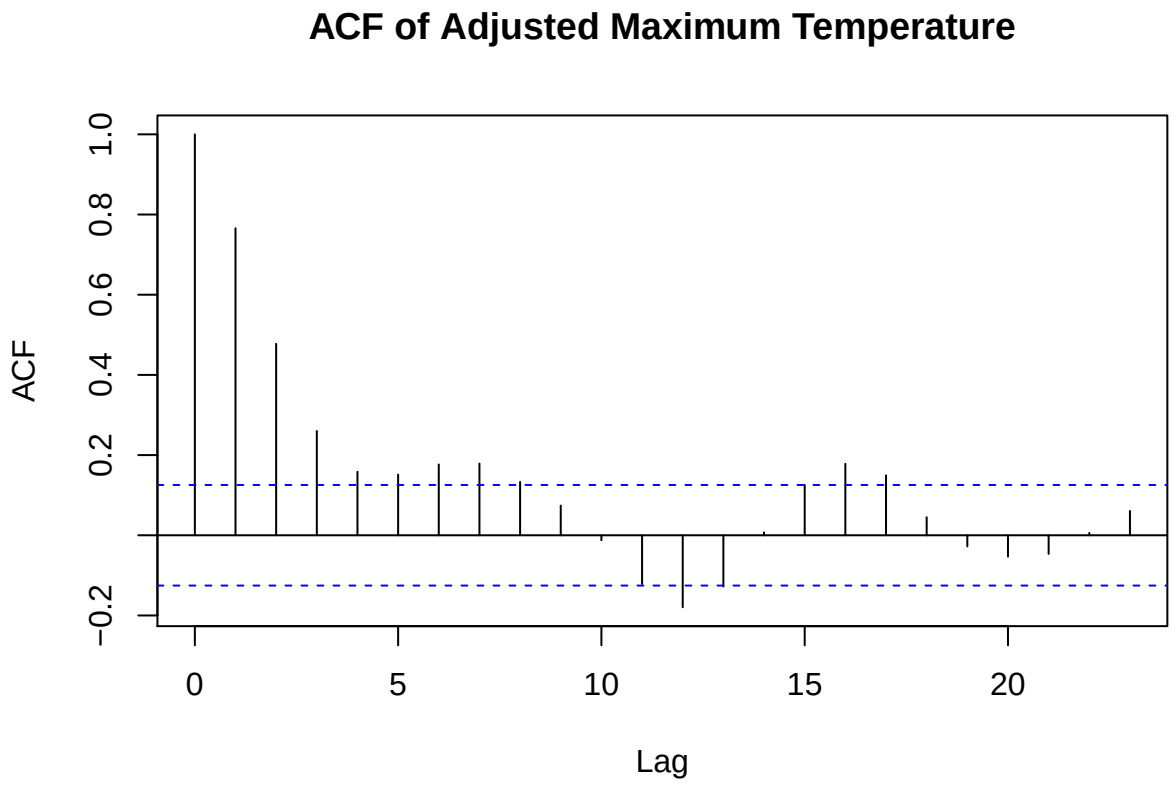
3. Plot the fitted outcome values from your regression in Q2 against the true data. Do you think the model fits the data well?



The model appears to fit the data well. This result is likely because the model has very significant seasonality, so when we added time trend and dummy season variables, the predicted values from the new model fit the data very closely.

4. Now let's further restrict the data sample to daily data from June to Sept. so that the time series would be (roughly) stationary. Do you think the daily maximum temperature data from June to Sept. are weakly dependent? Support your conclusion with graphical evidence.





From both the correlation coefficient and the acf plot, it appears that the daily maxiumum temperature from June to September are weakly dependent. The correlation between maximum temperature with lag 1 and lag 2 are 0.77 and 0.48 respectively, which tells us there is pretty significant correlation between each time period of the maximum temperature.

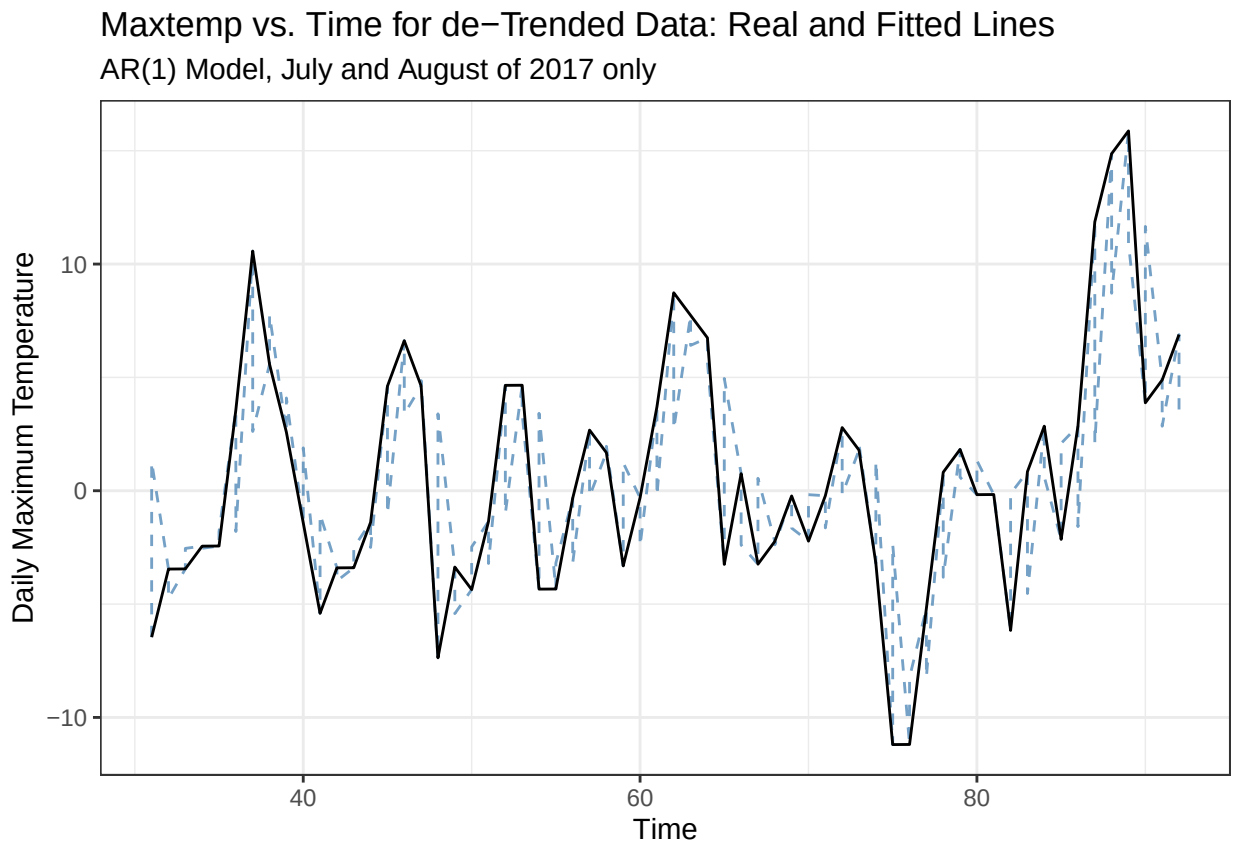
Also, using the ACF plot, we can see that the data is highly correlated and it captures seasonality. As lags increase, the correlation between y_t and y_{t-p} does go to 0 fast enough as p goes to infinity. This is another sign that they are weakly dependent, as there is strong correlation between lags of size 1 and 2 but as the lags increase, the correlations go pretty close to 0.

5. Run an AR(1) model using these two years of summer daily maximum temperature data. How would you interpret the slope coefficient?

```
##
## Call:
## lm(formula = maxtempdet ~ L.maxtempdet, data = DavisSummer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4159  -2.6901   0.5645   3.1513  14.5826
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.004762   0.297499  -0.016    0.987
## L.maxtempdet  0.734856   0.043850  16.759   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.637 on 241 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5382, Adjusted R-squared:  0.5363
## F-statistic: 280.8 on 1 and 241 DF,  p-value: < 2.2e-16
```

The slope coefficient of this regression represents β_1 in an AR(1) detrended model. This coefficient represents the effect the last time period has on the next time period: for every increase in t, daily detrended maximum temperature increases by 0.7348 multiplied by the detrended maximum temperature from the previous time period, in this case the previous day.

6. Plot the fitted outcome values from your regression in Q5 against the true data for July and August of 2017.



7. The fits of regression in Q5 are, of course, only good for the summer months. If we would like to carry out an AR(1) model for the whole time series in 2016 and 2017, how would us modify our AR(1) regression to control for seasonality?

If we would like to carry out an AR(1) model for the whole time series in 2017 and 2018, we can control for seasonality by adding in monthly dummary variables in the model.

8. Use the Breusch-Godfrey test to check whether the error terms in the AR(1) in Q5 is serially correlated. What do you learn from the testing results?

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  ar1.model
## LM test = 21.099, df = 1, p-value = 4.361e-06
```

We would reject the null hypothesis of zero serial correlation and conclude (based on the p-value) that there is serial correlation in the AR(1) model from Q5. The error terms in the AR(1) model are correlated.

Appendix

```
library(tidyverse)
library(sandwich)
library(ggthemes)
library(knitr)
library(car)
library(Hmisc)
library(lmtest)

#1
DavisWeather = read.csv("DavisWeather.csv")[,-1]
DavisWeather$t = 1:nrow(DavisWeather)
DavisWeather = DavisWeather %>% select(t, everything())
DavisWeather %>% ggplot(aes(x = t, y = maxtemp)) + geom_line() + theme_bw() +
  scale_x_continuous("Time") + scale_y_continuous("Daily Maximum Temperature") +
  geom_smooth(method = "loess", formula = "y~x", se = F, color = "grey")

cat("\\newpage")
#2
DavisWeather$Feb = ifelse(DavisWeather$month == 2, 1, 0)
DavisWeather$Mar = ifelse(DavisWeather$month == 3, 1, 0)
DavisWeather$Apr = ifelse(DavisWeather$month == 4, 1, 0)
DavisWeather$May = ifelse(DavisWeather$month == 5, 1, 0)
DavisWeather$Jun = ifelse(DavisWeather$month == 6, 1, 0)
DavisWeather$Jul = ifelse(DavisWeather$month == 7, 1, 0)
DavisWeather$Aug = ifelse(DavisWeather$month == 8, 1, 0)
DavisWeather$Sep = ifelse(DavisWeather$month == 9, 1, 0)
DavisWeather$Oct = ifelse(DavisWeather$month == 10, 1, 0)
DavisWeather$Nov = ifelse(DavisWeather$month == 11, 1, 0)
DavisWeather$Dec = ifelse(DavisWeather$month == 12, 1, 0)

# Regressing MaxTemp on monthly dummies
seasonal.model = lm(maxtemp ~ t + Feb + Mar + Apr + May + Jun + Jul + Aug + Sep +
                      Oct + Nov + Dec, data = DavisWeather)
summary(seasonal.model)

linearHypothesis(seasonal.model, c("Feb + Mar + Apr + May + Jun + Jul + Aug + Sep +
                                   Oct + Nov + Dec = 0"))

DavisWeather$maxtemphat = predict(seasonal.model)

df1 = cbind.data.frame(X = DavisWeather$t, Y = DavisWeather$maxtemp, Type = "Real")
tobind = cbind.data.frame(X = DavisWeather$t, Y = DavisWeather$maxtemphat,
                          Type = "Predicted")
df2 = rbind.data.frame(df1, tobind)
ggplot(data = subset(df2, Type == "Real"), aes(x = X, y = Y)) +
  geom_line(data = subset(df2, Type = "Predicted"),
            linetype = "dashed", color = "steelblue",
            alpha = I(3/4)) + geom_line() +
  theme_bw() + scale_y_continuous("Daily Maximum Temperature") +
  scale_x_continuous("Time") + ggtitle("Maxtemp vs. Time: Real and Fitted Lines",
                                       subtitle = "Seasonal Model")
```

```r
#4
DavisSummer = DavisWeather %>% filter(month > 5, month < 10)
DavisSummer$t = 1:nrow(DavisSummer)
DavisSummer %>% ggplot(aes(x = t, y = maxtemp)) + geom_line() + theme_bw() +
  scale_x_continuous("Time") + scale_y_continuous("Maximum Temperature") +
  ggtitle("Davis Weather Data", subtitle = "June, July, August, and September only")

DavisSummer$L.maxtemp <- Lag(DavisSummer$maxtemp,1)
DavisSummer$L2.maxtemp <- Lag(DavisSummer$maxtemp,2)

#cor(DavisSummer$maxtemp[2:244], DavisSummer$L.maxtemp[2:244])
#cor(DavisSummer$maxtemp[3:244], DavisSummer$L2.maxtemp[3:244])
cat("\\newpage")
acf(DavisSummer$maxtemp, main = "ACF of Adjusted Maximum Temperature")

cat("\\newpage")
#5
detrend.model = lm(maxtemp ~ t + Jul + Aug + Sep, data = DavisSummer)
DavisSummer$maxtempdet = detrend.model$residuals
DavisSummer$L.maxtempdet = Lag(DavisSummer$maxtempdet,1)
ar1.model = lm(maxtempdet ~ L.maxtempdet, data = DavisSummer)
summary(ar1.model)

cat("\\newpage")
DavisSummer$maxtemphatdet[2:nrow(DavisSummer)] = ar1.model$fitted.values
DavisSummer1 = DavisSummer %>% filter(year == 2017 & (Jul == 1 | Aug == 1))
df1 = cbind.data.frame(X = DavisSummer1$t, Y = DavisSummer1$maxtempdet, Type = "Real")
tobind = cbind.data.frame(X = DavisSummer1$t, Y = DavisSummer1$maxtemphatdet,
                          Type = "Predicted")
df2 = rbind.data.frame(df1, tobind)
ggplot(data = subset(df2, Type == "Real"), aes(x = X, y = Y)) +
  geom_line(data = subset(df2,
                          Type = "Predicted"),
            linetype = "dashed", color = "steelblue",
            alpha = I(3/4)) + geom_line() +
  theme_bw() + scale_y_continuous("Daily Maximum Temperature") +
  scale_x_continuous("Time") +
  ggtitle("Maxtemp vs. Time for de-Trended Data: Real and Fitted Lines",
          subtitle = "AR(1) Model, July and August of 2017 only")

bgtest(ar1.model)

cat("\\newpage")
```