ISYE 6402 Homework 2 Q2 Solutions

Background

In this problem, we will analyze aggregated temperature data.

Data *LA Temp Monthly.csv* contains the monthly average temperature of Los Angeles from January 1950 through December 2018. Run the following code to prepare the data for analysis:

Instructions on reading the data

To read the data in `R`, save the file in your working directory (make sure you have changed the directory if different from the R working directory) and read the data using the `R` function `read.csv()`

You will perform the analysis and modelling on the `Temp` data column.

```
fpath <- "LA Temp Monthly.csv"
df <- read.csv(fpath, head = TRUE)
```

Here are the libraries you will need:

```
library(mgcv)
library(TSA)
library(dynlm)
```

Run the following code to prepare the data for analysis:

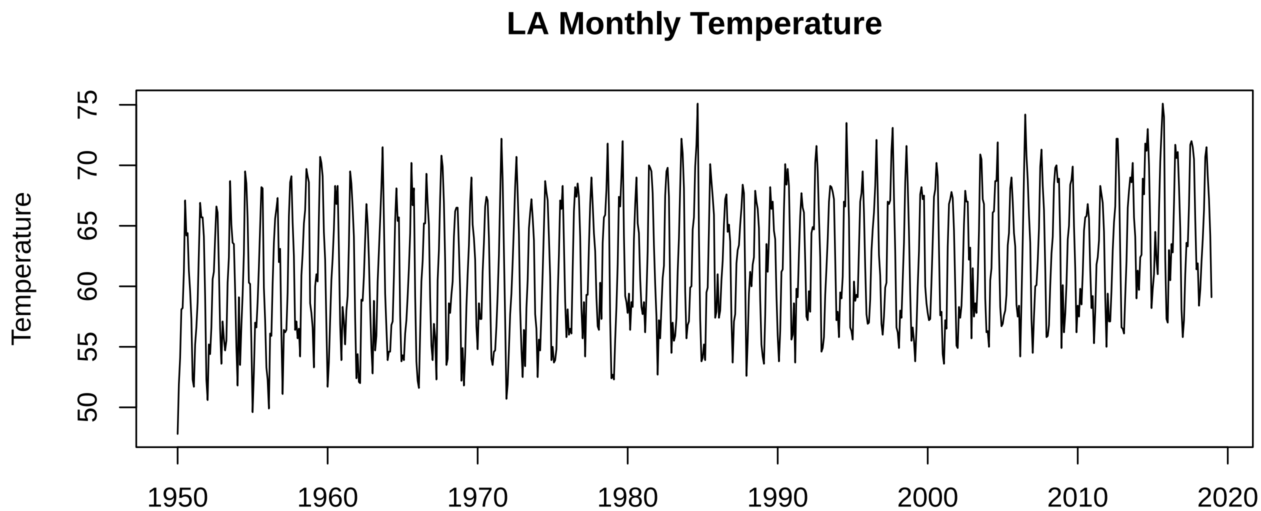
```
df$Date <- as.Date(paste0(df$Date, "01"), format = "%Y%m%d")
temp <- ts(df$Temp, start = 1950, freq = 12)

datenum <- ts(df$Date)
```

Question 2a: Exploratory Data Analysis

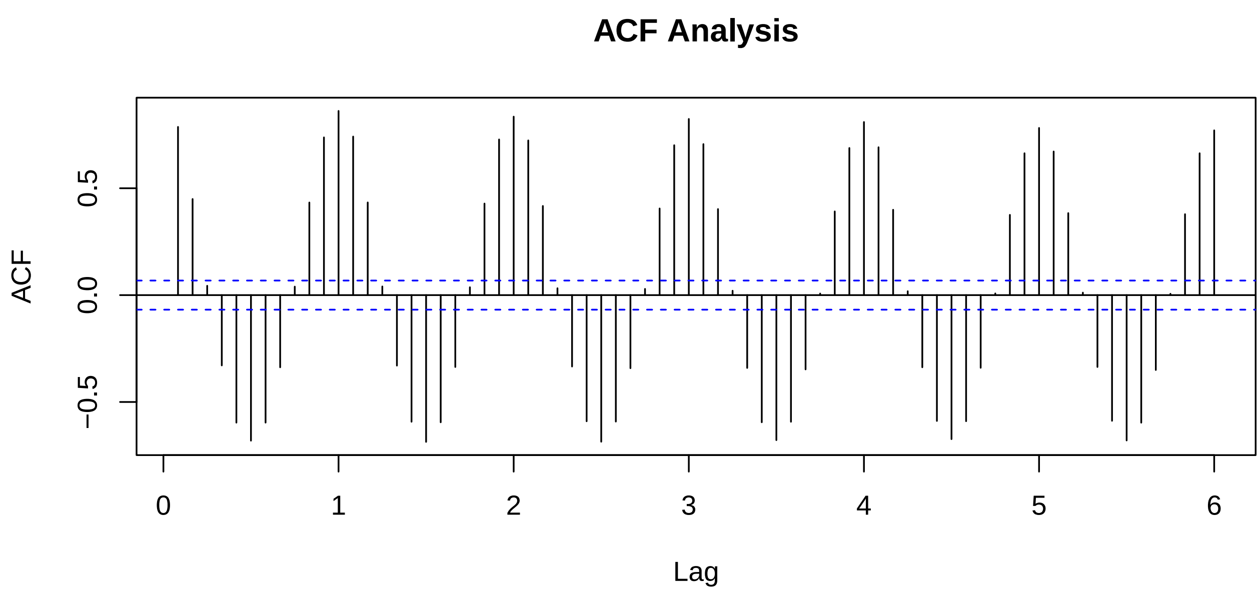
Plot both the Time Series and ACF plots. Comment on the main features, and identify what (if any) assumptions of stationarity are violated. Additionally, comment if you believe the differenced data is more appropriate for use in fitting the data. Support your response with a graphical analysis.

Hint: Make sure to use the appropriate differenced data.

```
plot(temp, xlab = "", ylab = "Temperature", main = "LA Monthly Temperature")
```



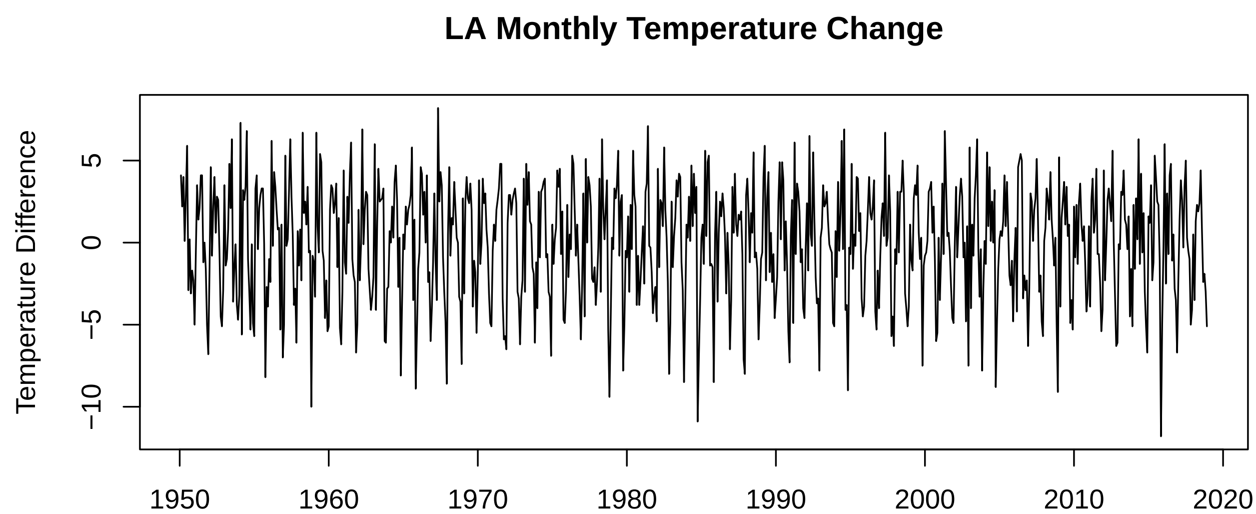

```r
acf(temp, lag.max = 12*6, main = "ACF Analysis")
```



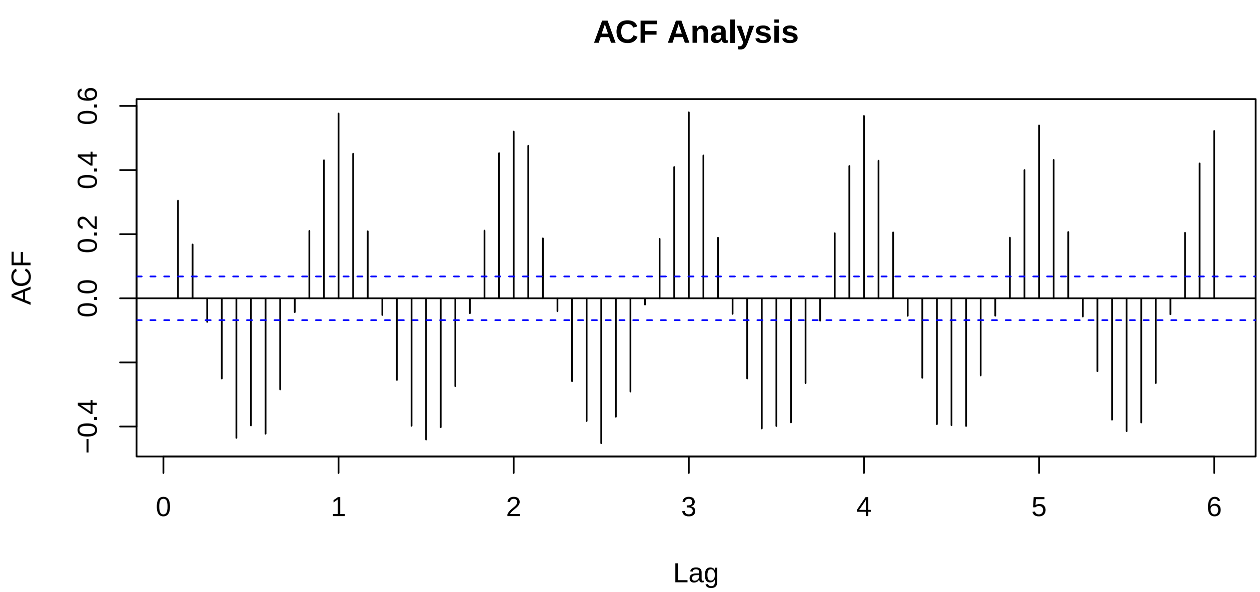
Response: Comments about the time series and ACF plots of the original time series

From the two plots, we can clearly see that the values stay within the confidence band and are following a cyclical pattern. A general increasing trend can also be observed in the graph. From the ACF plots, there is a clear seasonality pattern being observed.

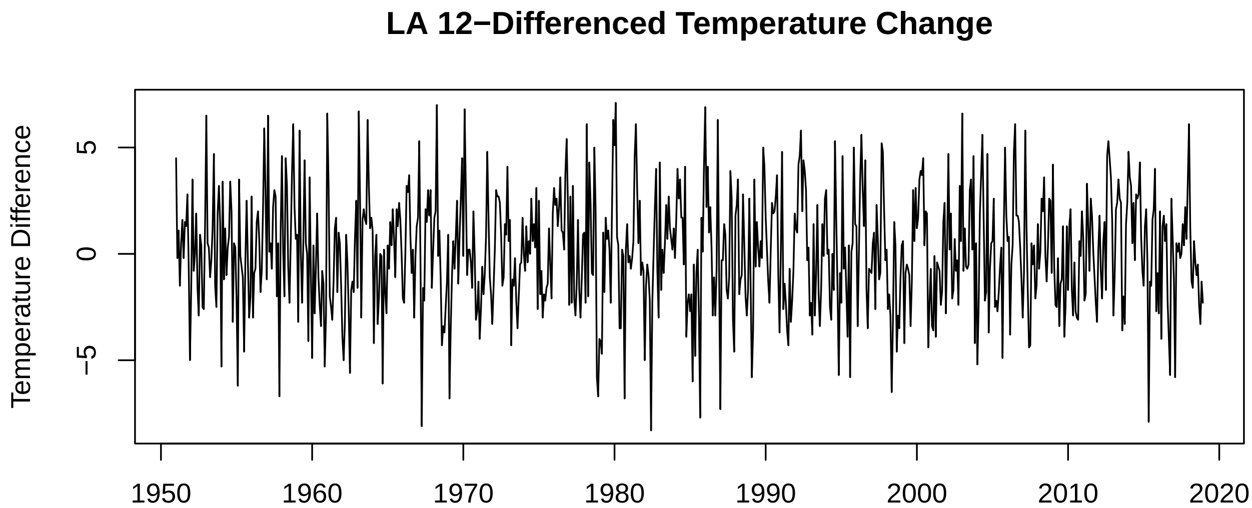
```r
plot(diff(temp), xlab = "", ylab = "Temperature Difference",
     main = "LA Monthly Temperature Change")
```



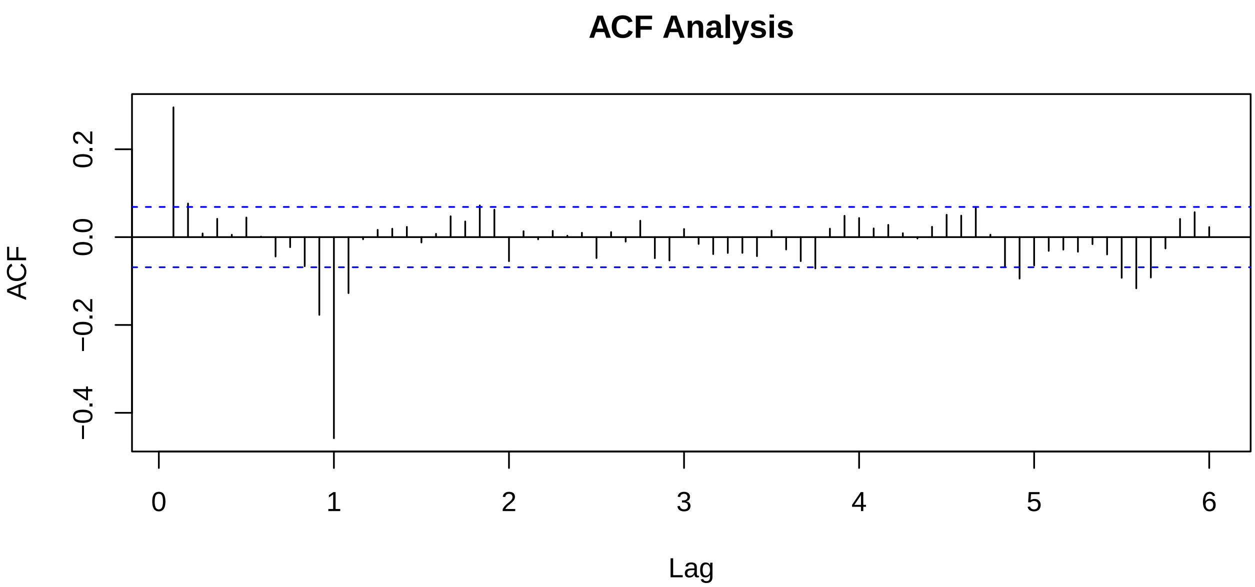

```r
acf(diff(temp), lag.max = 12*6, main = "ACF Analysis")
```




```r
plot(diff(temp, 12), xlab = "", ylab = "Temperature Difference",
     main = "LA 12-Differenced Temperature Change")
```




```
acf(diff(temp, 12), lag.max = 12*6, main = "ACF Analysis")
```



Response: Comments about the time series and ACF plots of the difference time series

The plot of the differenced data shows that trend has been removed. The seasonality effect, however, still seems to be present. For the differenced data, the first seasonal lag in the ACF is close to 1 and decays slowly over multiples of the lag. Clearly, the 1st order differenced data is not appropriate for use in fitting the seasonality of the data.

Since we know that the 1st order difference doesn't appropriately address seasonality, we can apply a 12 lag difference as provided above. The ACF plot still shows that some acf values are statistically significant for some small lags but seasonality has been removed to a great extent.

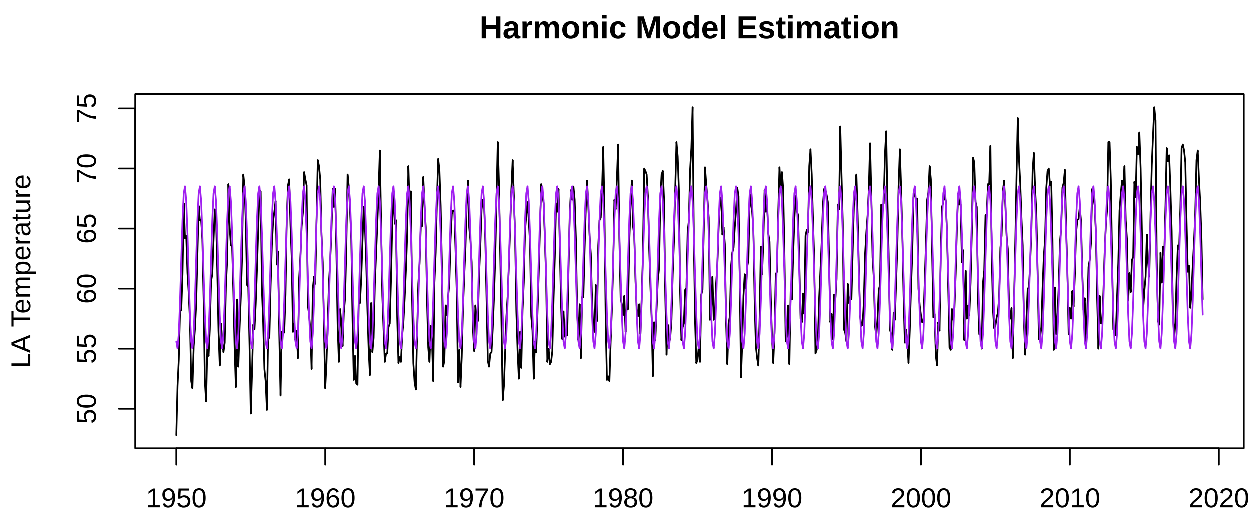
Question 2b: Seasonality Estimation

Separately fit a seasonality harmonic model and the ANOVA seasonality model to the temperature data. Evaluate the quality of each fit with residual analysis. Does one model perform better than the other? Which model would you select to fit the seasonality in the data?

```r
# 1. Estimate seasonality using harmonic model

model.harmonic <- dynlm(temp ~ harmonic(temp))
summary(model.harmonic)
```

```
##
## Time series regression with "ts" data:
## Start = 1950(1), End = 2018(12)
##
## Call:
## dynlm(formula = temp ~ harmonic(temp))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8022 -1.7405 -0.0916  1.5677  9.4140
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 61.76836    0.08714  708.81   <2e-16 ***
## harmonic(temp)cos(2*pi*t) -6.16619    0.12324  -50.03   <2e-16 ***
## harmonic(temp)sin(2*pi*t) -2.81769    0.12324  -22.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.508 on 825 degrees of freedom
## Multiple R-squared:  0.7858, Adjusted R-squared:  0.7853
## F-statistic:  1513 on 2 and 825 DF,  p-value: < 2.2e-16
```

```r
plot(temp, type = "l", xlab = "", ylab = "LA Temperature", main = "Harmonic Model Estimation")
lines(fitted(model.harmonic), col = "purple")
```

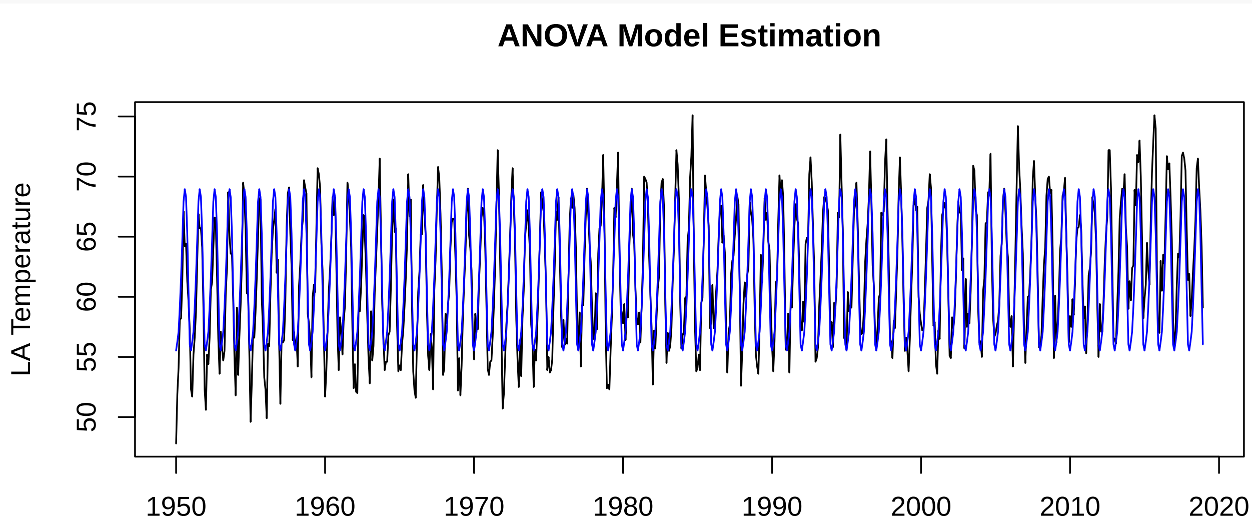



```r
# 2. Estimate seasonality using ANOVA model

model.anova <- dynlm(temp ~ season(temp))
summary(model.anova)
```

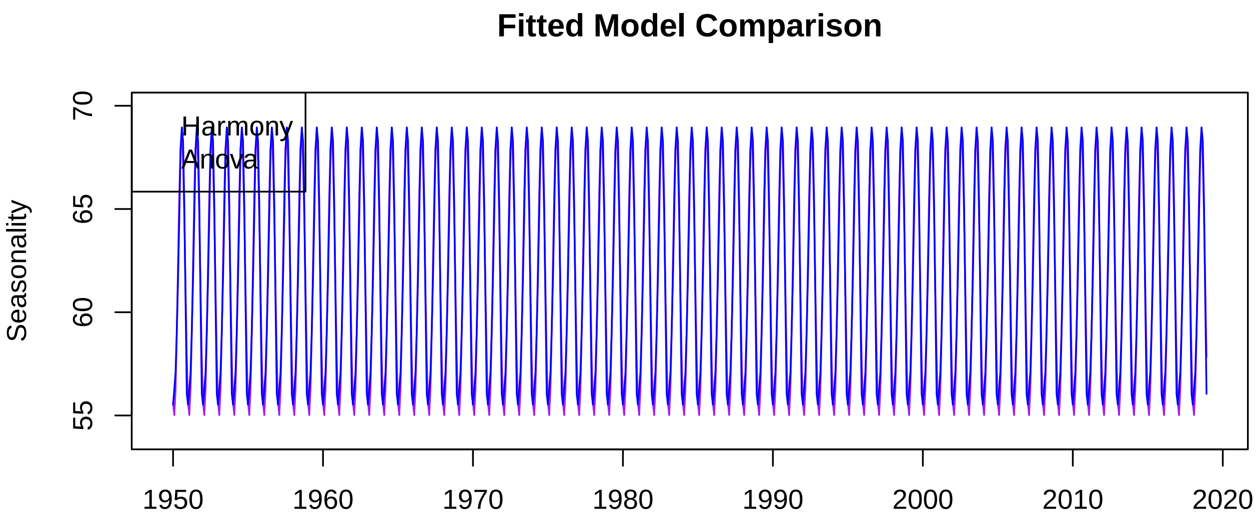
```
##
## Time series regression with "ts" data:
## Start = 1950(1), End = 2018(12)
```

```
##
## Call:
## dynlm(formula = temp ~ season(temp))
##
## Residuals:
##     Min      1Q Median      3Q     Max
## -7.729 -1.581 -0.037  1.625  8.735
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      55.5290     0.2841 195.476  < 2e-16 ***
## season(temp)Feb   0.7275     0.4017   1.811 0.070511 .
## season(temp)Mar   1.5623     0.4017   3.889 0.000109 ***
## season(temp)Apr   3.6043     0.4017   8.972  < 2e-16 ***
## season(temp)May   6.0812     0.4017  15.137  < 2e-16 ***
## season(temp)Jun   9.1159     0.4017  22.691  < 2e-16 ***
## season(temp)Jul  12.3841     0.4017  30.826  < 2e-16 ***
## season(temp)Aug  13.4246     0.4017  33.417  < 2e-16 ***
## season(temp)Sep  12.7710     0.4017  31.790  < 2e-16 ***
## season(temp)Oct   9.7362     0.4017  24.235  < 2e-16 ***
## season(temp)Nov   4.9464     0.4017  12.313  < 2e-16 ***
## season(temp)Dec   0.5188     0.4017   1.291 0.196897
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.36 on 816 degrees of freedom
## Multiple R-squared:  0.8124, Adjusted R-squared:  0.8098
## F-statistic: 321.2 on 11 and 816 DF,  p-value: < 2.2e-16
```

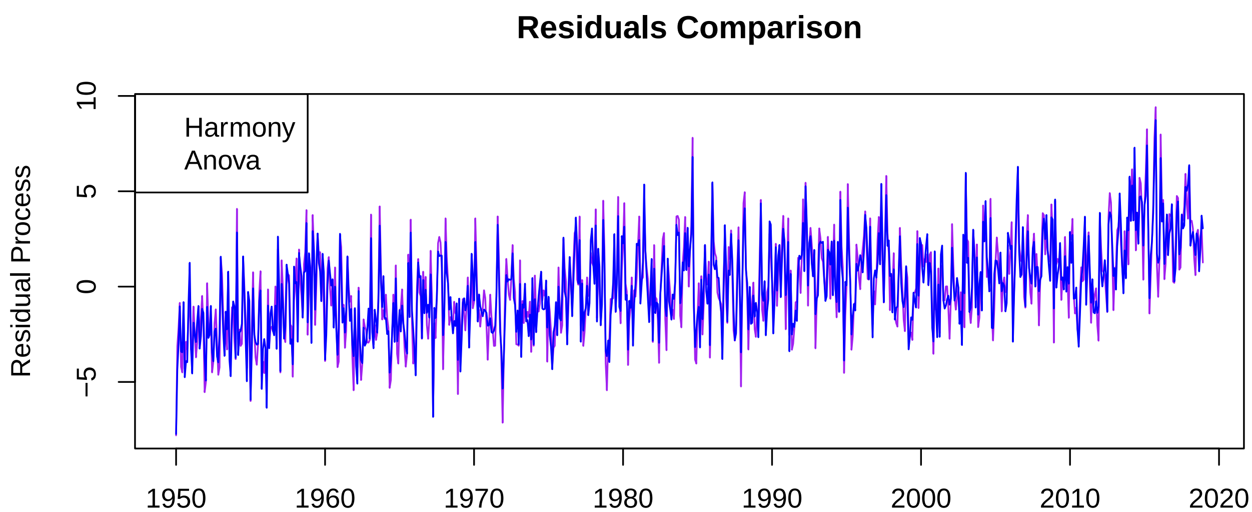
```r
plot(temp, type = "l", xlab = "", ylab = "LA Temperature",
     main = "ANOVA Model Estimation")
lines(fitted(model.anova), col = "blue")
```



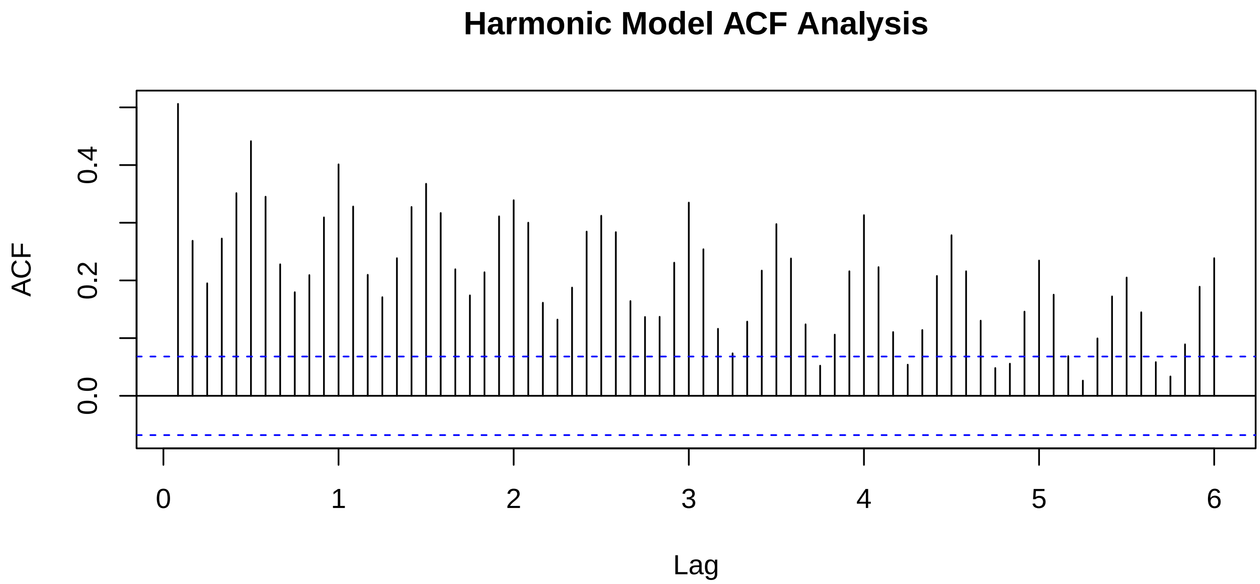

```r
# Compare two models

plot(fitted(model.harmonic), xlab = "", ylab = "Seasonality", col = "purple",
     ylim = c(54, 70), main = "Fitted Model Comparison")
lines(fitted(model.anova), col= "blue")
legend("topleft", legend = c("Harmony", "Anova"), col = c("purple", "blue"))
```



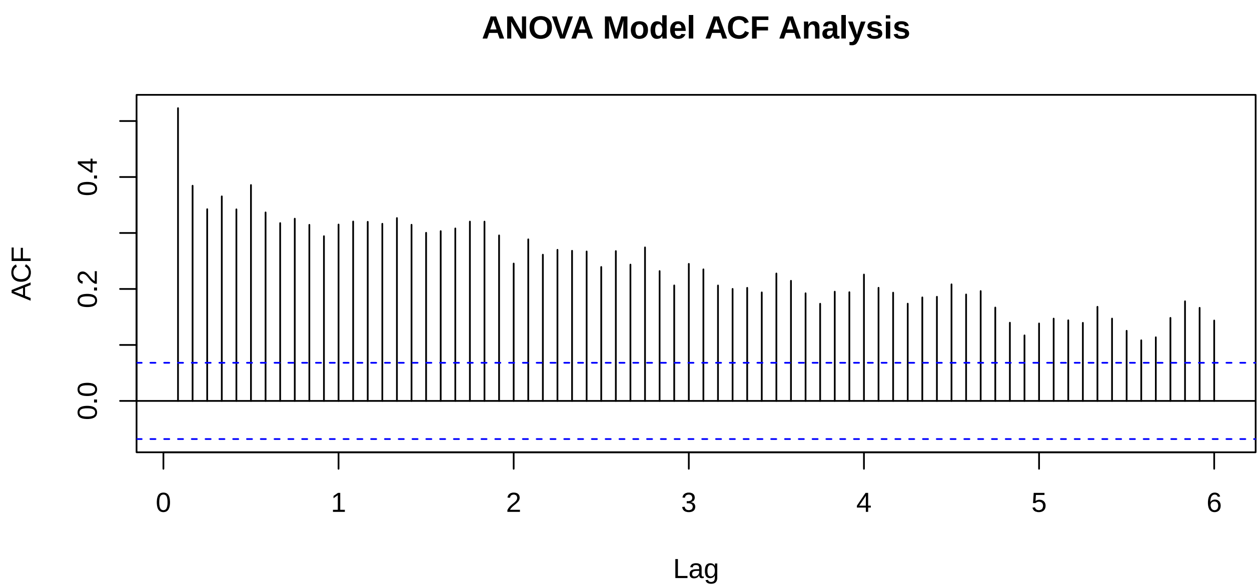

```r
# Compare the residuals of two models

resid.harmonic <- residuals(model.harmonic)
resid.anova <- residuals(model.anova)
ylim <- c(min(resid.harmonic, resid.anova),
          max(resid.harmonic, resid.anova))

ts.plot(resid.harmonic, xlab = "", ylab = "Residual Process",
        col="purple", ylim = ylim, main = "Residuals Comparison")
lines(resid.anova, col = "blue")
legend("topleft", legend = c("Harmony", "Anova"), col = c("purple", "blue"))
```




```r
# ACF analysis of two models

acf(resid.harmonic, lag.max = 12 * 6, main = "Harmonic Model ACF Analysis")
```




```
acf(resid.anova, lag.max = 12 * 6, main = "ANOVA Model ACF Analysis")
```



Response: Compare Seasonality Models

The regression coefficients for both models are statistically significant, indicating that both models capture a seasonal pattern. The two models perform similarly, except that the ANOVA model overestimates, and the harmonics models underestimates seasonality based on the comparison of the fitted values.

The residuals time series plots for both models show an increasing trend; this suggests that we will need to jointly fit both trend and seasonality. The acf plots show also that the residuals are not stationary, with acf values slowly decreasing, again suggesting the presence of a trend. The ANOVA model seems to capture seasonality better since the acf values are not maintaining a seasonality pattern as for the harmonics model.

Question 2c: Trend-Seasonality Estimation

Using the time series data, fit the following models to estimate the trend with seasonality fitted using ANOVA:

- Parametric Polynomial Regression

- Non-parametric model

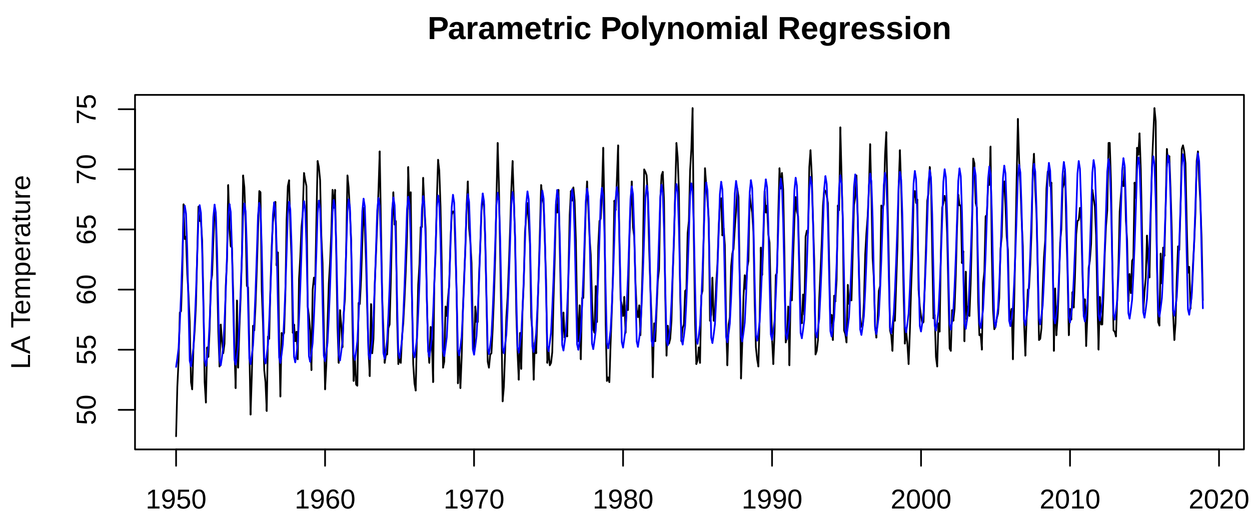
Overlay the fitted values on the original time series. Plot the residuals with respect to time. Plot the ACF of the residuals. Comment on how the two models fit and on the appropriateness of the stationarity assumption of the residuals.

What form of modelling seems most appropriate and what implications might this have for how one might expect long term temperature data to behave? Provide explicit conclusions based on the data analysis.

```r
points <- 1:length(temp)
points <- c(points - min(points)) / max(points)
x1 <- points
x2 <- points ^ 2

# Parametric Polynomial Regression

model.para <- dynlm(temp ~ x1 + x2 + season(temp))
summary(model.para)
```

```
##
## Time series regression with "ts" data:
## Start = 1950(1), End = 2018(12)
##
## Call:
## dynlm(formula = temp ~ x1 + x2 + season(temp))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.767 -1.377 -0.177  1.307  6.908
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       53.5666     0.3066 174.685  < 2e-16 ***
## x1                 3.1245     0.9520   3.282  0.00107 **
## x2                 1.2963     0.9228   1.405  0.16048
## season(temp)Feb    0.7222     0.3370   2.143  0.03238 *
## season(temp)Mar    1.5517     0.3370   4.605 4.78e-06 ***
## season(temp)Apr    3.5884     0.3370  10.649  < 2e-16 ***
## season(temp)May    6.0599     0.3370  17.984  < 2e-16 ***
## season(temp)Jun    9.0893     0.3370  26.975  < 2e-16 ***
## season(temp)Jul   12.3521     0.3370  36.657  < 2e-16 ***
## season(temp)Aug   13.3873     0.3370  39.729  < 2e-16 ***
## season(temp)Sep   12.7284     0.3370  37.774  < 2e-16 ***
## season(temp)Oct    9.6882     0.3370  28.751  < 2e-16 ***
## season(temp)Nov    4.8930     0.3370  14.521  < 2e-16 ***
## season(temp)Dec    0.4601     0.3370   1.365  0.17248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.979 on 814 degrees of freedom
## Multiple R-squared:  0.8683, Adjusted R-squared:  0.8662
## F-statistic: 412.9 on 13 and 814 DF,  p-value: < 2.2e-16
```

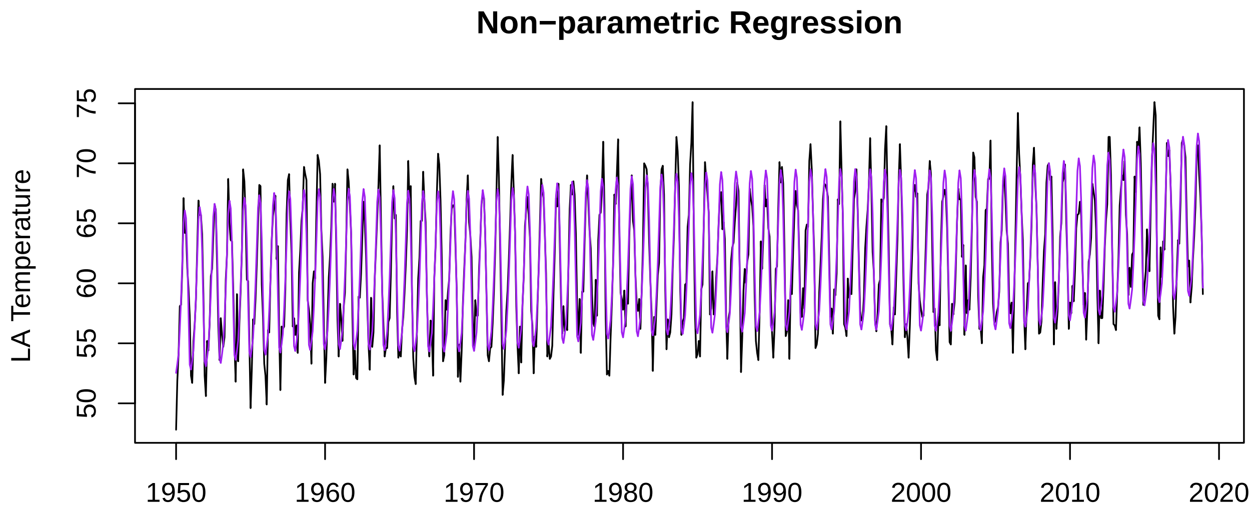
```r
plot(temp, type = "l", xlab = "", ylab = "LA Temperature",
     main = "Parametric Polynomial Regression")
lines(fitted(model.para), col = "blue")
```




```r
# Non-parametric model

month <- as.factor(format(df$Date, "%b"))
model.gam <- gam(temp ~ s(points) + month)
summary(model.gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## temp ~ s(points) + month
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.1535     0.2323 254.635  < 2e-16 ***
## monthAug      9.7879     0.3285  29.793  < 2e-16 ***
## monthDec     -3.1502     0.3286  -9.588  < 2e-16 ***
## monthFeb     -2.8606     0.3285  -8.707  < 2e-16 ***
## monthJan     -3.5801     0.3285 -10.897  < 2e-16 ***
## monthJul      8.7554     0.3285  26.650  < 2e-16 ***
## monthJun      5.4954     0.3285  16.727  < 2e-16 ***
## monthMar     -2.0339     0.3285  -6.191 9.49e-10 ***
## monthMay      2.4687     0.3285   7.515 1.51e-13 ***
## monthNov      1.2854     0.3286   3.912 9.91e-05 ***
## monthOct      6.0834     0.3285  18.516  < 2e-16 ***
## monthSep      9.1262     0.3285  27.778  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df     F p-value
## s(points) 7.271  8.282 49.16  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.873   Deviance explained = 87.6%
## GCV = 3.8122  Scale est. = 3.7235    n = 828
```

```r
plot(temp, type = "l", xlab = "", ylab = "LA Temperature",
     main = "Non-parametric Regression")
lines(ts(fitted(model.gam), start = 1950, freq = 12), col = "purple")
```

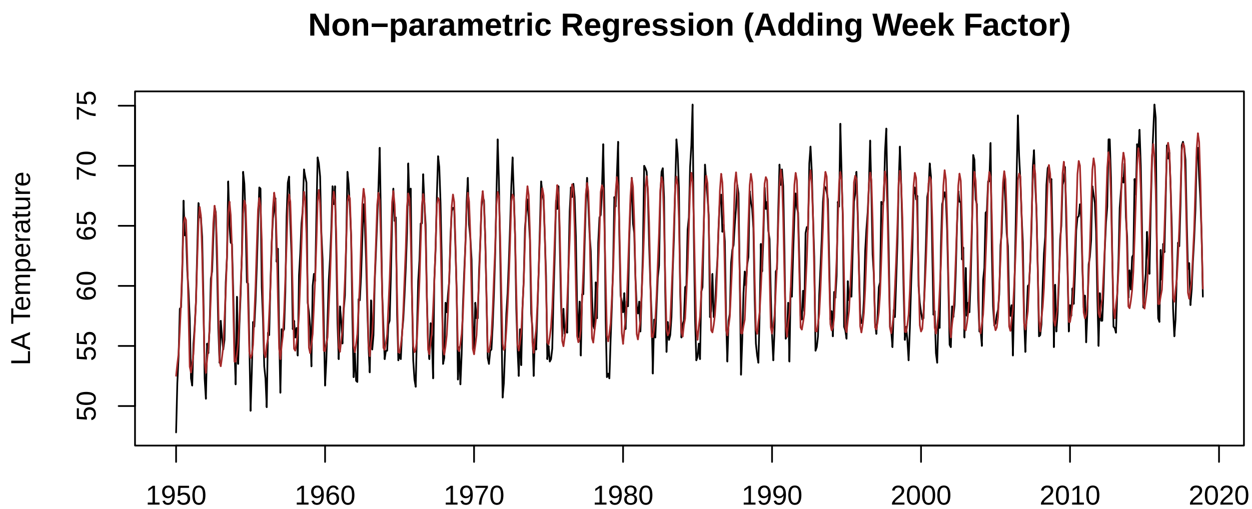



```r
# Adding week factor

week <- as.factor(weekdays(df$Date))
model.gam2 <- gam(temp ~ s(points) + month + week)
summary(model.gam2)
```

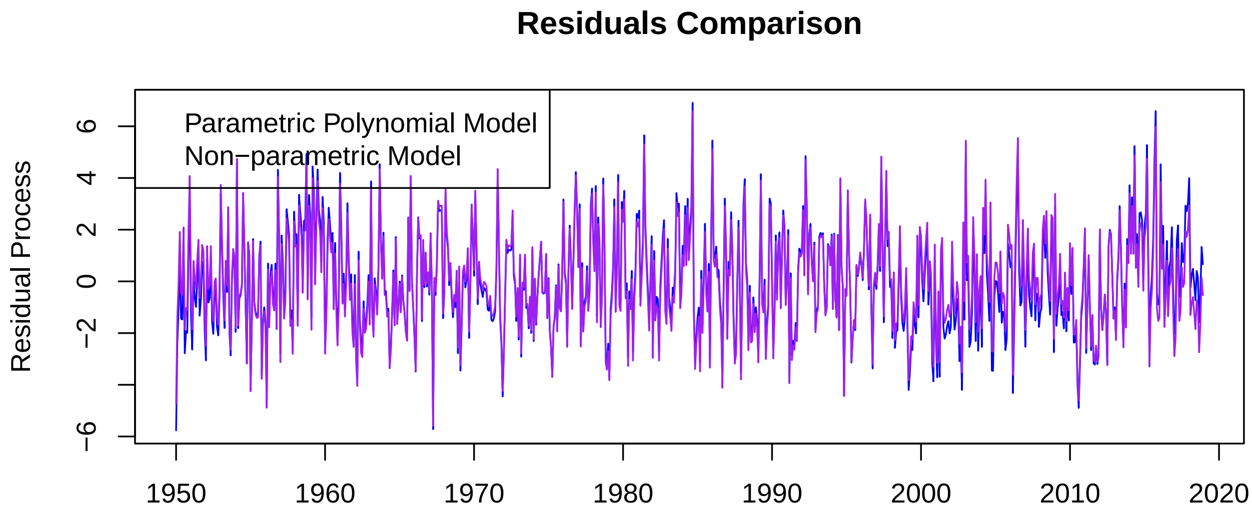
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## temp ~ s(points) + month + week
##
## Parametric coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     59.21400    0.28439 208.217  < 2e-16 ***
## monthAug         9.78448    0.32852  29.783  < 2e-16 ***
## monthDec        -3.14798    0.32857  -9.581  < 2e-16 ***
## monthFeb        -2.86591    0.32852  -8.724  < 2e-16 ***
## monthJan        -3.57338    0.32850 -10.878  < 2e-16 ***
## monthJul         8.75544    0.32846  26.656  < 2e-16 ***
## monthJun         5.49266    0.32852  16.720  < 2e-16 ***
## monthMar        -2.03157    0.32852  -6.184 9.94e-10 ***
## monthMay         2.48010    0.32852   7.549 1.19e-13 ***
## monthNov         1.28778    0.32854   3.920 9.62e-05 ***
## monthOct         6.08076    0.32856  18.507  < 2e-16 ***
## monthSep         9.12844    0.32855  27.784  < 2e-16 ***
## weekMonday      -0.08857    0.25180  -0.352    0.725
## weekSaturday     0.07821    0.25133   0.311    0.756
## weekSunday      -0.07572    0.25022  -0.303    0.762
## weekThursday    -0.11926    0.25081  -0.475    0.635
## weekTuesday     -0.39999    0.25128  -1.592    0.112
## weekWednesday    0.17502    0.25127   0.697    0.486
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df     F p-value
## s(points) 7.296  8.301 49.11  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.873   Deviance explained = 87.7%
## GCV = 3.8392  Scale est. = 3.7219    n = 828
```

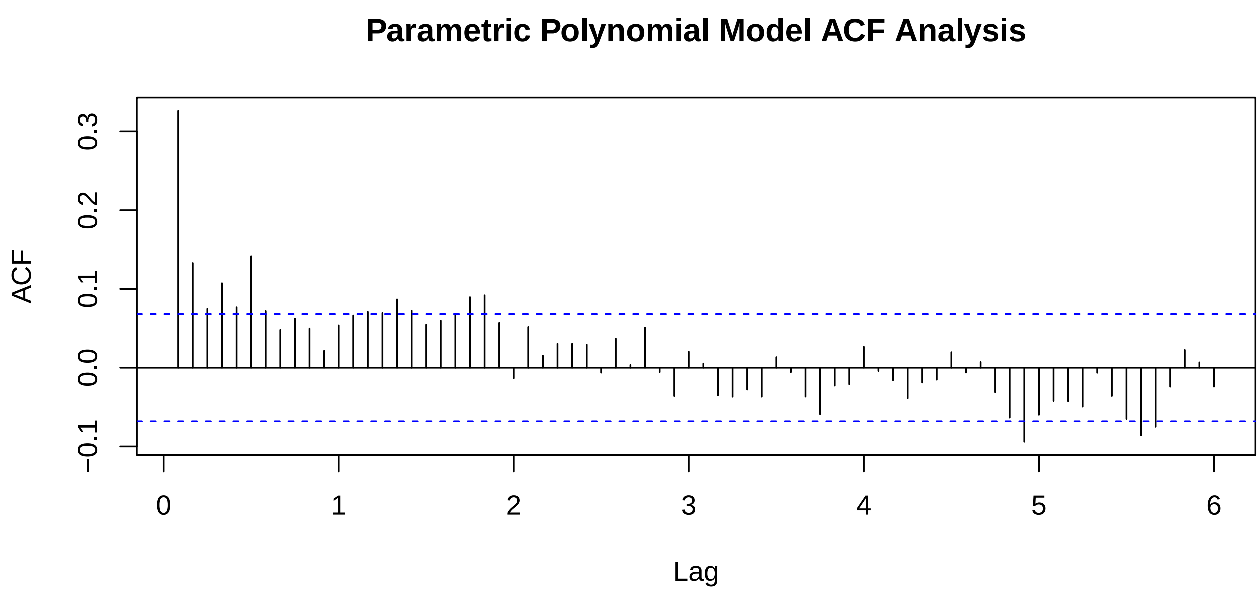
```
plot(temp, type = "l", xlab = "", ylab = "LA Temperature",
     main = "Non-parametric Regression (Adding Week Factor)")
lines(ts(fitted(model.gam2), start = 1950, freq = 12), col = "brown")
```



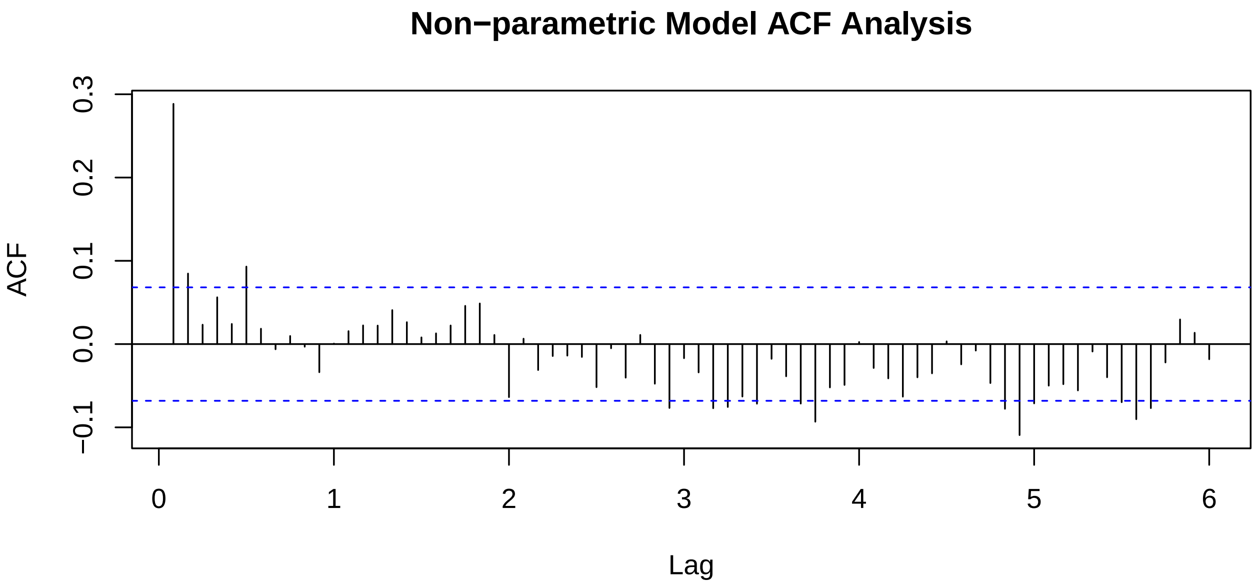

```
# Compare the residuals of two models

resid.para <- residuals(model.para)
resid.gam <- ts(residuals(model.gam), start = 1950, freq = 12)
ylim <- c(min(resid.para, resid.gam),
          max(resid.para, resid.gam))

ts.plot(resid.para, xlab = "", ylab = "Residual Process",
        col = "blue", ylim = ylim, main = "Residuals Comparison")
lines(resid.gam, col = "purple")
legend("topleft",legend = c("Parametric Polynomial Model",
                            "Non-parametric Model"),
       col = c("blue", "purple"))
```




```r
# ACF analysis of two models
acf(resid.para, lag.max = 12 * 6,
    main = "Parametric Polynomial Model ACF Analysis")
```




```r
acf(resid.gam, lag.max = 12 * 6,
    main = "Non-parametric Model ACF Analysis")
```



Response: Model Comparison

From the fitted models, we see that the parametric polynomial regression shows a linear trend fitting on the original data, while the seasonality is quite effectively captured. In case of the non parametric model, while the seasonality is effectively captured, the trend is also fitted better that the polynomial model. Adding the week data does not increase the predictive power of the non-parametric model, with all the regression coefficients corresponding to the week seasonality being not statistically significant.

From residual analysis of the two models, we see that the residuals of the parametric polynomial regression models show somewhere larger variability. From the ACF of the residuals, we see that the residuals from the non-parametric model fit are stationary treats whereas those from the parametric model show some serial correlation.

We can clearly see here that the non parametric model of the trend seems to work for the temperature data. We can predict that the temperature will follow a general(but not linear) rising trend, with seasonality on an annual basis. Hence this model can be used towards predicting temperature.

Overall, we find that temperature has shown an increase over the past 70 years hence seasonality is not sufficient to capture the variability in the data. We have seen a similar finding in the data example provided in the class. Hence this is phenomena may be consistent over other geographic areas.