# Example of data analysis homework

Average temperature in Corvallis

This is an example of how your response for the data analysis section of the homework should be presented.

Note:

- Use section headings to help the grader navigate (reflect these in your R code too).
- Plots should have English language labels (not R variables).
- You don't have to present every plot you make, just the ones that help tell the story
- You may write in an informal manner, think of this as a conversation with a colleague, not a journal article, but you should still aim for complete, grammatically correct sentences.
- Put all your R code in the Appendix.

## Introduction

```
library(ggplot2)
library(plyr)
library(lubridate)
```

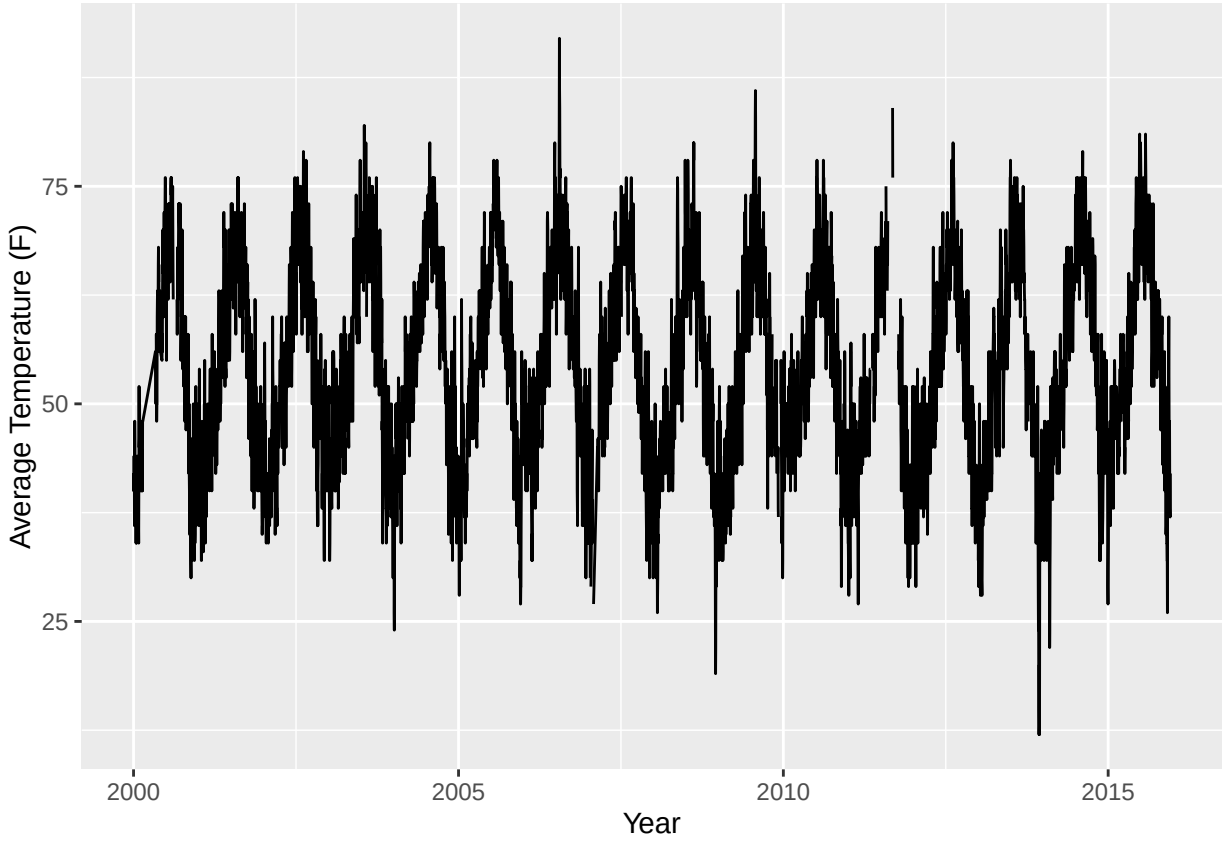
```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:plyr':
##
##     here
```

```
options(stringsAsFactors = FALSE)

# == read in and tidy data == #

# read in the daily weather data
corvallis <- read.csv(url("http://stat565.cwick.co.nz/data/corvallis.csv"))
corv <- corvallis[ , c("PST", "Mean.TemperatureF", "PrecipitationIn" )]
corv <- rename(corv, c("Mean.TemperatureF" = "temp", "PrecipitationIn" = "precip"))

corv$date <- ymd(corv$PST)
corv$year <- year(corv$date)
corv$month <- month(corv$date)
corv$yday <- yday(corv$date)
```

The daily average temperatures in Corvallis from 2000 to 2012 are shown below. The goal of this analysis is describe the seasonality, trend and the properties of the residual variation in this series. Since the plot below is dominated by the annual pattern in temperature I will start by examining this seasonal pattern.

```
# == plot of whole series == #
    qplot(date, temp, data = corv, geom = "line") +
      ylab("Average Temperature (F)") +
      xlab("Year")
```

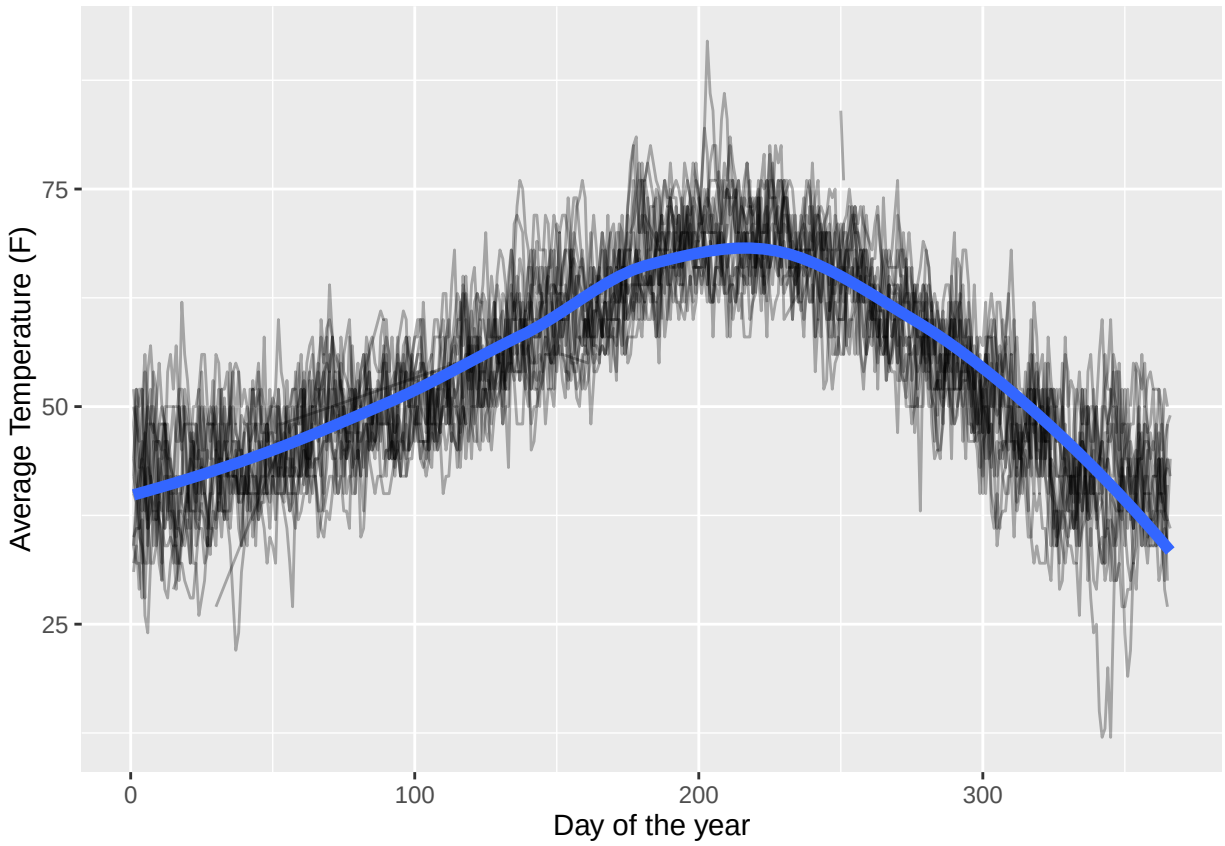



## Seasonality

To examine the seasonal pattern the daily temperature is plotted against the day of the year the temperature was recorded (1 = Jan 1, 32 = Feb 1, etc.). The records that occur in the same calendar year are connected by a line. The blue curve is a smooth through the data using a locally weight regression (loess). Unsurprisingly we see the temperatures are highest in summer and lowest in winter. In Corvallis, the warmest temperatures are generally between mid-July and mid-August. The positive slope from day 1 to 200 seems of a lower magnitude than the negative slope from day 200 to 350. This may be evidence that the temperatures cool in the Fall quicker than they warm in the Spring.

```
# seasonality dominates so start there...

# == seasonality == #
qplot(yday, temp, data = corv, geom = "line", group = year, alpha = I(.3)) +
    geom_smooth(aes(group = 1), method = "loess", se = FALSE, size = 2) +
    ylab("Average Temperature (F)") +
    xlab("Day of the year")
```

```
## Warning: Removed 101 rows containing non-finite values (stat_smooth).
```




```
# fit the seasonal model in preparation for subtraction
lo_fit <- loess(temp ~ yday, data = corv, na.action = na.exclude)
corv$seasonal_smooth <- fitted(lo_fit)

# check it looks ok
qplot(yday, temp, data = corv, geom = "line", group = year, alpha = I(.3)) +
  geom_line(aes(y = seasonal_smooth), colour = "blue",  size = 1)
# subtract off pattern
corv$deseasonalised <- corv$temp - corv$seasonal_smooth
```

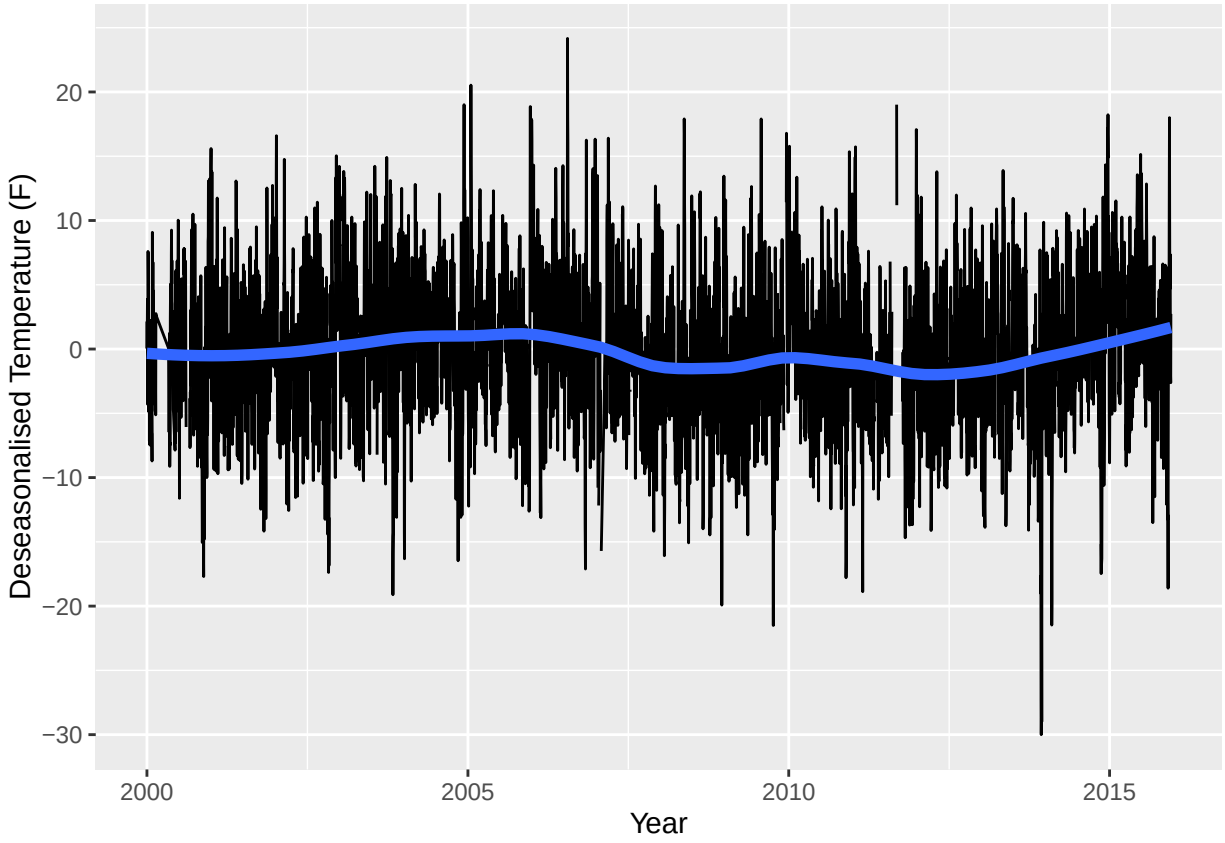
## Trend

To get a better idea of the long term trend in the temperatures in Corvallis, the smooth in the previous plot is used as the model of the seasonal variation and is subtracted from the daily temperature, resulting in seasonally adjusted temperatures. These are plotted below along with another smoother (in blue). Beyond seasonal variation we see there is still daily variation in temperature on the order of 10 degrees. There is some evidence that the years 2008 & 2009 were cooler on average. This smooth line is used as an estimate for the trend in the series and subtracted from the de-seasonalised temperatures. When plotting the resulting residuals there was no obvious non-stationarity left in the mean.

```
# === Trend === #

# play with a few spans
qplot(date, deseasonalised, data = corv, geom = "line") +
```

```
  geom_smooth(se = FALSE, method = "loess", span = 0.4, size = 2)+
  ylab("Deseasonalised Temperature (F)") +
  xlab("Year")
```

```
## Warning: Removed 101 rows containing non-finite values (stat_smooth).
```




```
# fit the trend model in preparation for subtraction
lo_fit_trend <- loess(deseasonalised ~ as.numeric(date), data = corv, na.action = na.exclude,
                      span = 0.4)
corv$trend_smooth <- fitted(lo_fit_trend)
corv$residual <- corv$deseasonalised - corv$trend_smooth
```

```
# === Residuals === #

# check for remaining non-stationarity
qplot(date, residual, data = corv, geom  = "line")
```

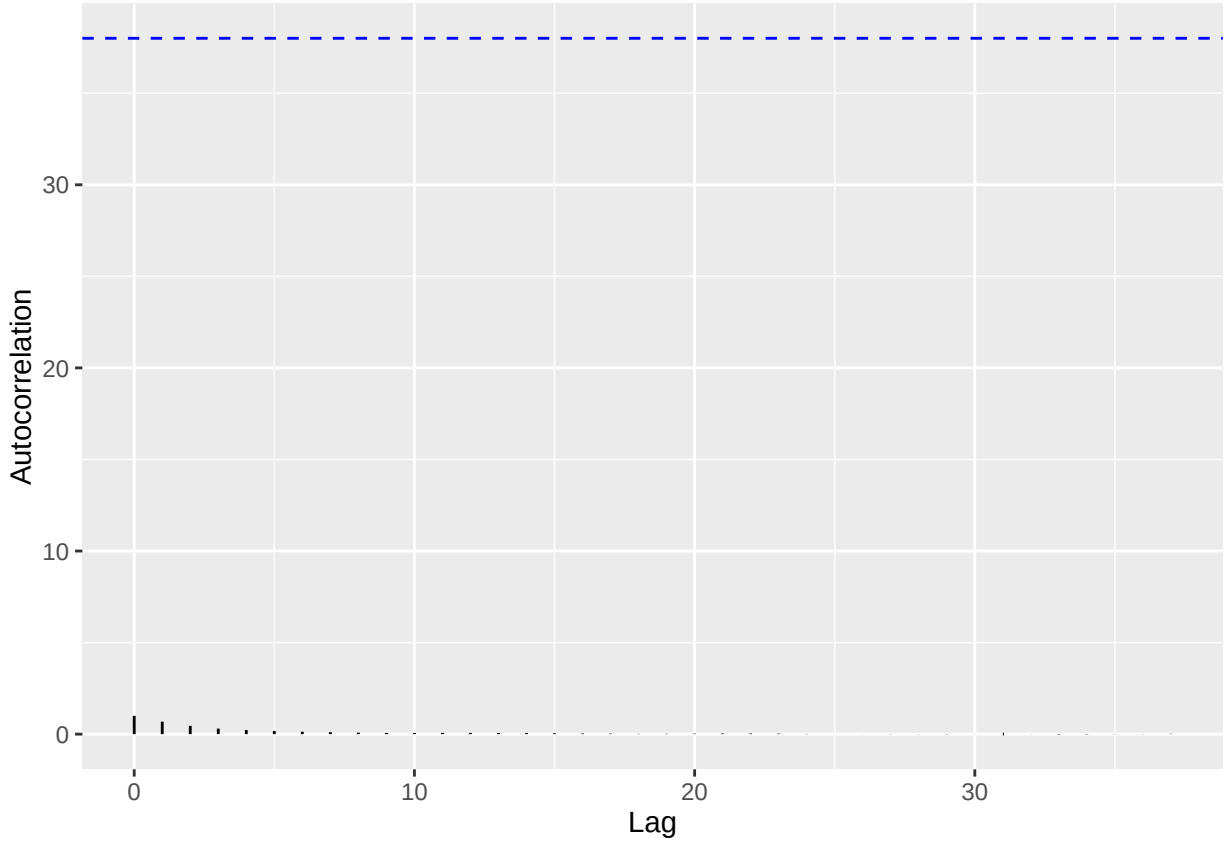
## Residual

```
# force to all be in order still
corv <- corv[order(corv$date), ]
```

```
# correlation of residuals
corv_acf <- acf(corv$residual, na.action = na.pass)

# qplotize it, for uniform look.
acf_df <- data.frame(ACF = corv_acf$acf, Lag = corv_acf$lag)
ci <- qnorm((1 + 0.95)/2)/sqrt(corv_acf$n.used)
```

The autocorrelation function for the estimated residuals is shown below, where a unit of lag is equal to one day. The correlation in average temperature between one day and the next is quite high, 0.68. The correlation rapidly tails off and is negligible beyond two weeks. This implies that, beyond the seasonal norms and long term trend, the temperature today is not useful for predicting the temperature in the future beyond about two weeks.

```
qplot(Lag, ymin = 0, ymax = ACF, data = acf_df, geom = "linerange") +
  geom_hline(aes(yintercept = 38), colour = "blue", linetype = "dashed") +
  ylab("Autocorrelation")
```



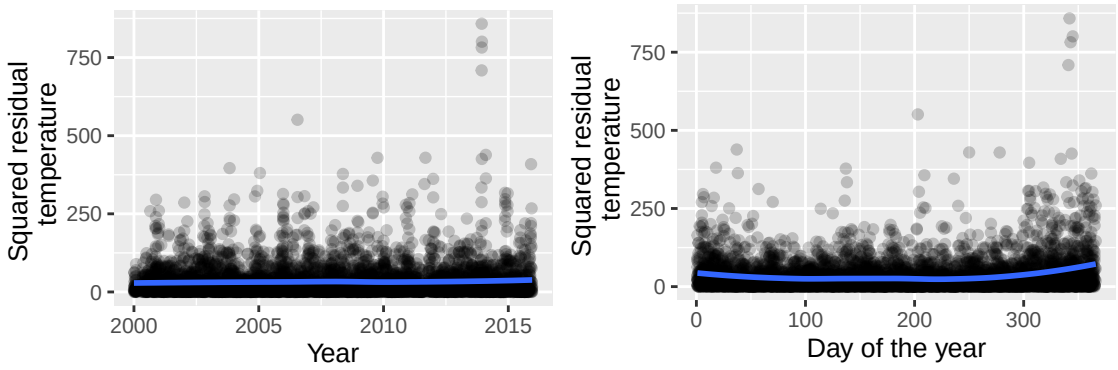

The stationarity of the variance of the residuals is examined by plotting the square of the residuals against time, and day of the year. There is no obvious long term trend but the variability does seem to be higher in the winter months.

```
# stationarity of variance
qplot(date, residual^2, data = corv, alpha = I(0.2)) +
  geom_smooth(method = "loess") +
  ylab("Squared residual \n temperature") +
  xlab("Year") +
  theme_grey(10)
```

```
## Warning: Removed 101 rows containing non-finite values (stat_smooth).

## Warning: Removed 101 rows containing missing values (geom_point).
```

```
qplot(yday, residual^2, data = corv, alpha = I(0.2)) +
  geom_smooth(method = "loess") +
  ylab("Squared residual \n temperature") +
  xlab("Day of the year") +
  theme_grey(10)
```

```
## Warning: Removed 101 rows containing non-finite values (stat_smooth).

## Warning: Removed 101 rows containing missing values (geom_point).
```




In summary,

- The greatest source of variation in daily temperatures in Corvallis is the annual pattern, warm temperatures in the summer, cool in the winter.
- There doesn't appear to be a systematic trend in the temperatures, although 2008 & 2009 appeared to be cooler than average
- The serial correlation of daily temperatures (beyond the seasonal norms and trend) is high but falls rapidly and is negligible on a time scale of two weeks.
- The variation in temperature is higher during the winter but shows no long term changes.

## Appendix

```
library(ggplot2)
library(plyr)
library(lubridate)
options(stringsAsFactors = FALSE)

# == read in and tidy data == #

# read in the daily weather data
corvallis <- read.csv(url("http://stat565.cwick.co.nz/data/corvallis.csv"))
corv <- corvallis[ , c("PST", "Mean.TemperatureF", "PrecipitationIn" )]
corv <- rename(corv, c("Mean.TemperatureF" = "temp", "PrecipitationIn" = "precip"))
```

```r
corv$date <- ymd(corv$PST)
corv$year <- year(corv$date)
corv$month <- month(corv$date)
corv$yday <- yday(corv$date)
# == plot of whole series == #
    qplot(date, temp, data = corv, geom = "line") +
      ylab("Average Temperature (F)") +
      xlab("Year")
# seasonality dominates so start there...

# == seasonality == #
qplot(yday, temp, data = corv, geom = "line", group = year, alpha = I(.3)) +
    geom_smooth(aes(group = 1), method = "loess", se = FALSE, size = 2) +
    ylab("Average Temperature (F)") +
    xlab("Day of the year")
# fit the seasonal model in preparation for subtraction
lo_fit <- loess(temp ~ yday, data = corv, na.action = na.exclude)
corv$seasonal_smooth <- fitted(lo_fit)

# check it looks ok
qplot(yday, temp, data = corv, geom = "line", group = year, alpha = I(.3)) +
  geom_line(aes(y = seasonal_smooth), colour = "blue",  size = 1)
# subtract off pattern
corv$deseasonalised <- corv$temp - corv$seasonal_smooth
# === Trend === #

# play with a few spans
qplot(date, deseasonalised, data = corv, geom = "line") +
  geom_smooth(se = FALSE, method = "loess", span = 0.4, size = 2)+
  ylab("Deseasonalised Temperature (F)") +
  xlab("Year")
# fit the trend model in preparation for subtraction
lo_fit_trend <- loess(deseasonalised ~ as.numeric(date), data = corv, na.action = na.exclude,
                      span = 0.4)
corv$trend_smooth <- fitted(lo_fit_trend)
corv$residual <- corv$deseasonalised - corv$trend_smooth
# === Residuals === #

# check for remaining non-stationarity
qplot(date, residual, data = corv, geom  = "line")
# force to all be in order still
corv <- corv[order(corv$date), ]

# correlation of residuals
corv_acf <- acf(corv$residual, na.action = na.pass)

# qplotize it, for uniform look.
acf_df <- data.frame(ACF = corv_acf$acf, Lag = corv_acf$lag)
ci <- qnorm((1 + 0.95)/2)/sqrt(corv_acf$n.used)
qplot(Lag, ymin = 0, ymax = ACF, data = acf_df, geom = "linerange") +
  geom_hline(aes(yintercept = 38), colour = "blue", linetype = "dashed") +
  ylab("Autocorrelation")
# stationarity of variance
```

```r
qplot(date, residual^2, data = corv, alpha = I(0.2)) +
  geom_smooth(method = "loess") +
  ylab("Squared residual \n temperature") +
  xlab("Year") +
  theme_grey(10)

qplot(yday, residual^2, data = corv, alpha = I(0.2)) +
  geom_smooth(method = "loess") +
  ylab("Squared residual \n temperature") +
  xlab("Day of the year") +
  theme_grey(10)
```