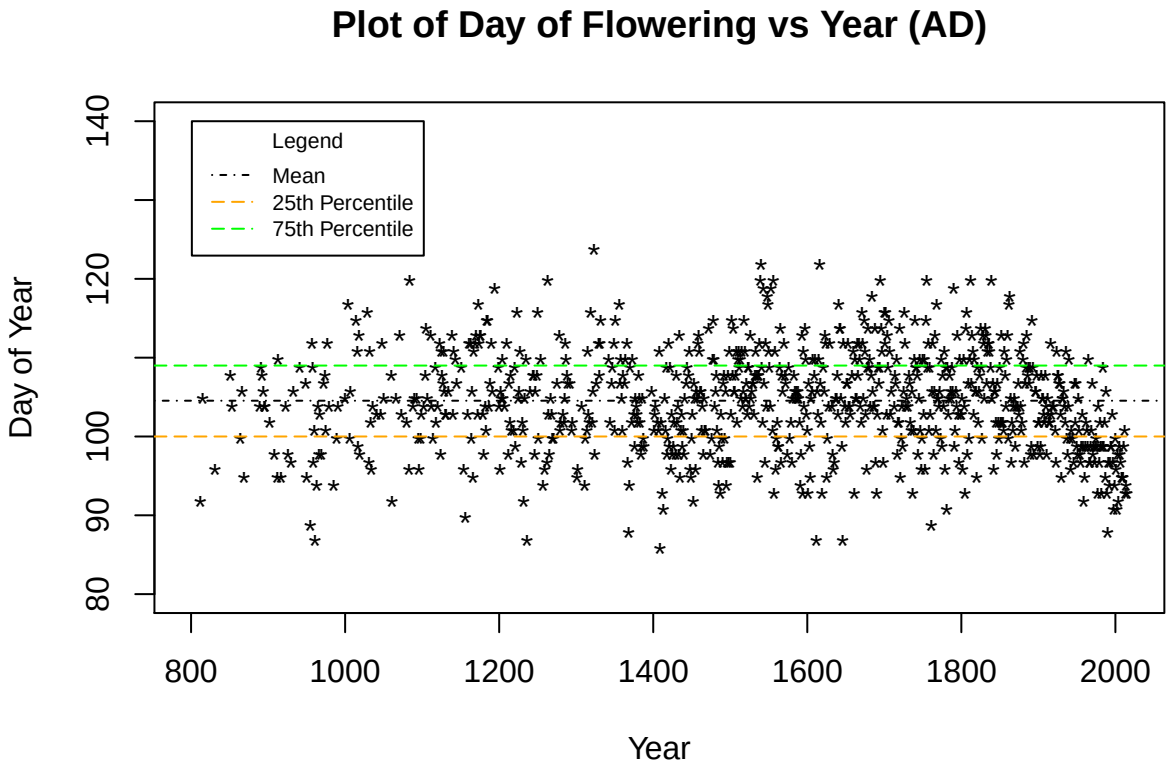
36-467 Homework 1

Name: Eu Jing Chua
Andrew ID: eujingc

Q1 It makes less sense to calculate summary statistics about the date column as it is not "numerical", unlike the day of the year.

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
##  86.000000 100.000000 105.000000 104.540508 109.000000 124.000000
##       NA's
##        388
```

Q2

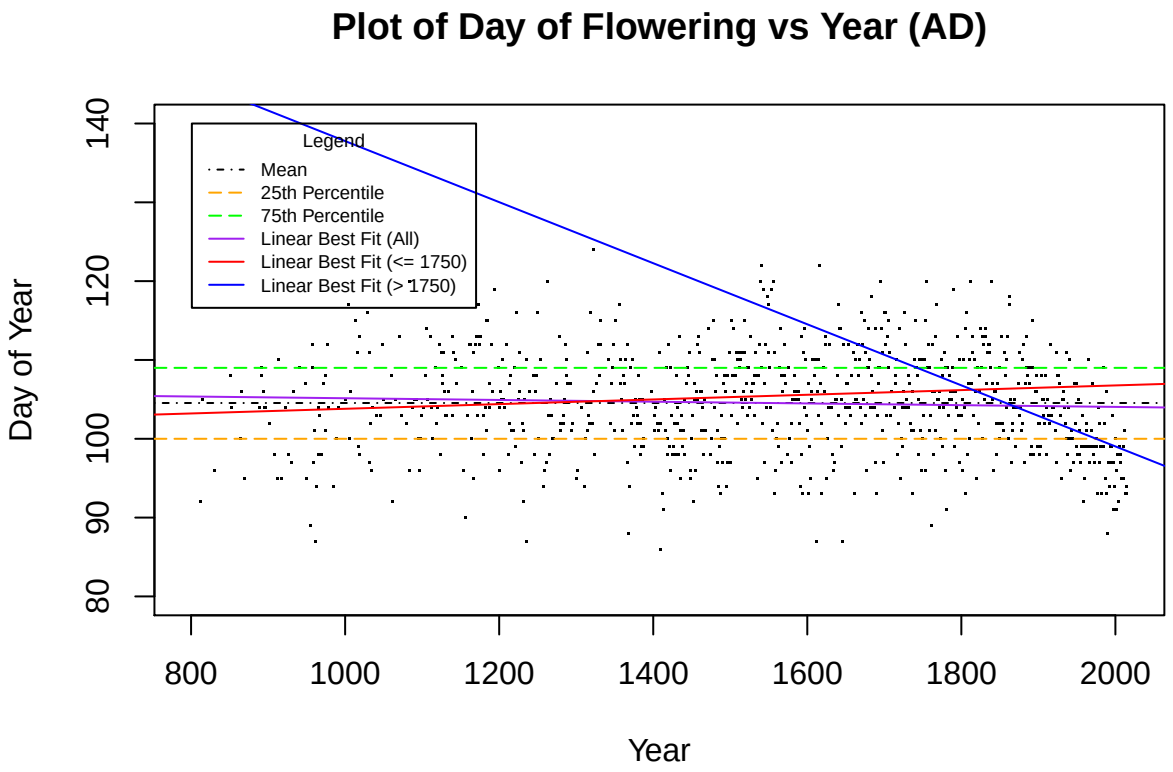


Q2 d)

- The density of the scatter seems to increase across the years, which follows how less data was available in the early years.
- The interquatile range does not seem to capture the majority of the scatter, with alot of outliers in the later years.
- There seems to be a downward trend in the denser scatter within the years 1800 to 2000.
- The scatter does not seem to be centered around the mean in the years 1800 to 2000.

- The variance of the scatter does not seem to be constant.

Q3



Q3 a) The slope is -0.001098, which means for each increase in year, the predicted day of year of cherry blossom flowering decreases by 0.001098.

Q3 b) The slope is 2.965e-03, which means for each increase in year, the predicted day of year of cherry blossom flowering increases by 2.965e-03.

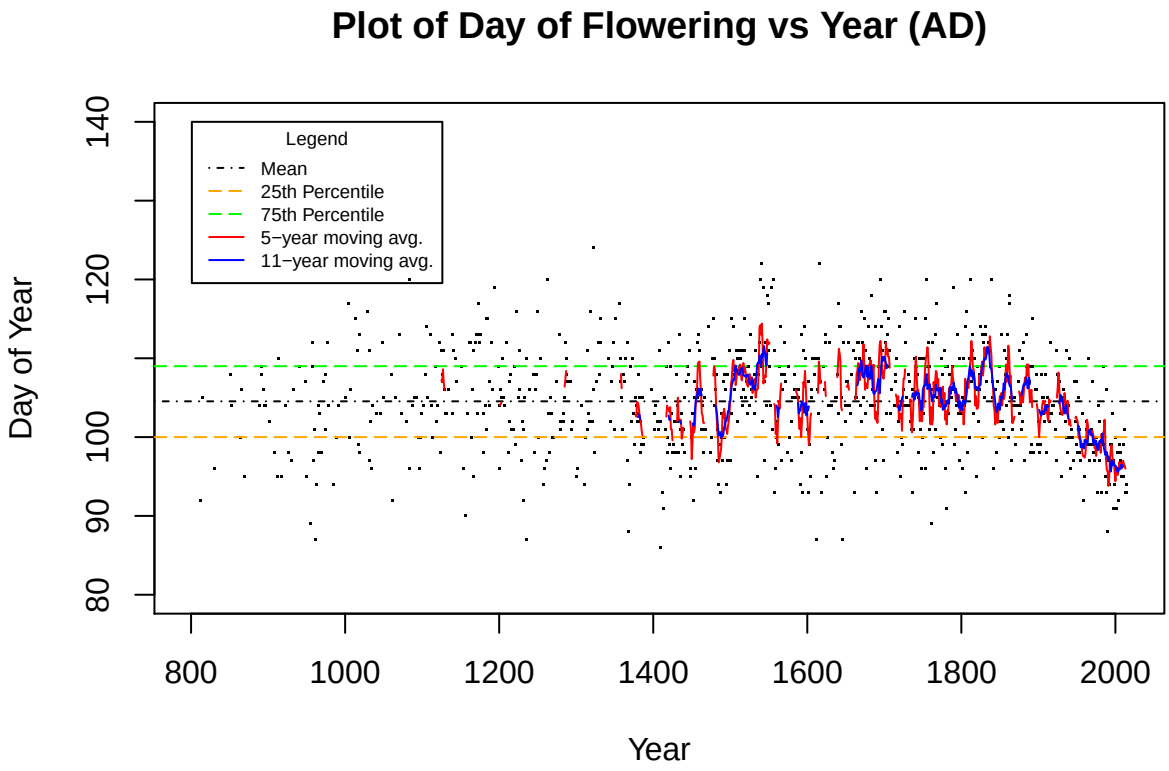
Q3 c) The slope is -0.03874, which means for each increase in year, the predicted day of year of cherry blossom flowering decreases by 0.03874.

Q3 d) The slope of the overall regression (purple line) is small and negative. However, the slope of the regression of data points before 1750 (red line) is a positive one that is larger in magnitude compared to the overall. The slope of the regression of data points after 1750 (blue line) is a negative one that is larger in magnitude compared to the previous two.

Q4

```
plot(Flowering.DOY ~ Year.AD, data = df, ylim = c(80, 140), pch = ".",
     xlab = "Year", ylab = "Day of Year")
abline(h = mean(df$Flowering.DOY, na.rm = TRUE), lty = 4)
```

```r
abline(h = flowering.DOY.quantiles[["25%"]], lty = 5, col = "orange")
abline(h = flowering.DOY.quantiles[["75%"]], lty = 5, col = "green")

# Q4 a)
library(zoo)
moving.avg.5 <- rollmean(df$Flowering.DOY, 5, na.pad=TRUE)
index.1945 <- which(df$Year.AD == 1945)

# Q4 c)
moving.avg.11 <- rollmean(df$Flowering.DOY, 11, na.pad=TRUE)

# Q4 d)
lines(df$Year.AD, moving.avg.5, col="red")
lines(df$Year.AD, moving.avg.11, col="blue")

legend(df$Year.AD[1], 140,
       c("Mean",
         "25th Percentile",
         "75th Percentile",
         "5-year moving avg.", "11-year moving avg."),
       cex=0.6, col=c("black", "orange", "green", "red", "blue"),
       lty=c(4, 5, 5, 1, 1, 1), title="Legend")

title("Plot of Day of Flowering vs Year (AD)")
```

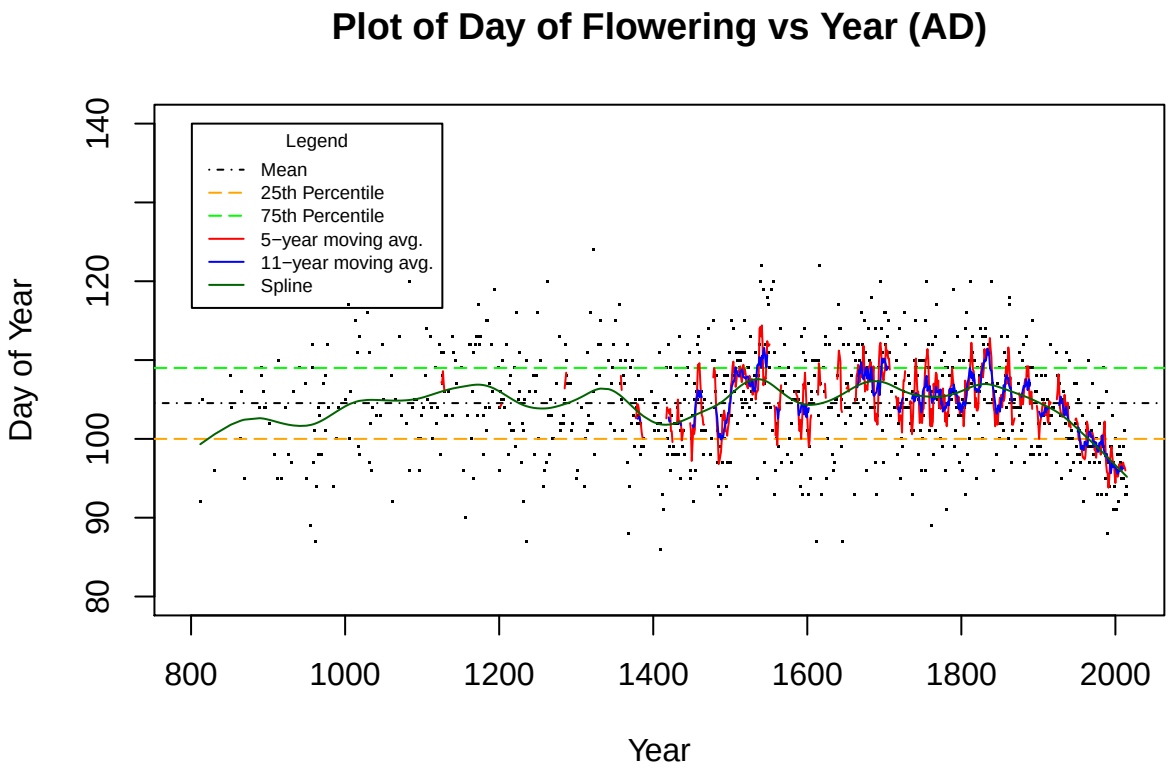


Q4 b) The code handles NA values by propogating them in calculations. The value of the flowering day-of-

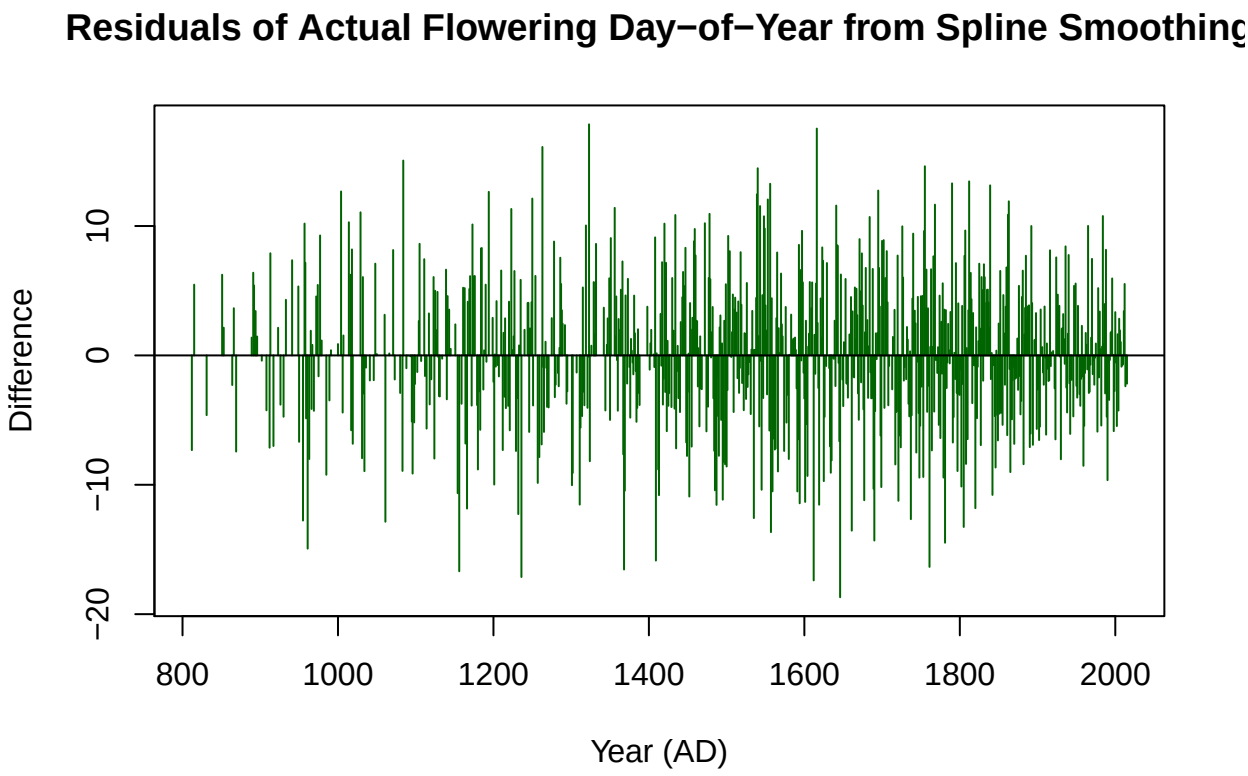
year at 1945 is NA, and the nearby values including itself are 101, 100, NA, 97, 107. The calculated moving averages with k=5 around 1945 is also then follows: NA, NA, NA, NA, NA. As the calculation for each of these moving averages includes the value at 1945, the NA has been propogated and thus their results are also NA, not just the value of the moving average at 1945.

Q4 e) The 11-year moving average is has slightly smoother than the 5-year moving average, but both are still quite noisy to the eye. However, both do seem to show more obviously a downward trend between the years 1800 and 2000. The 11-year moving average has more gaps in it compared to the 5-year moving average as there is much more propogation of NA values from the wider window being used.

Q5



Q5 b)



The residuals seem to be centered around 0, with more residuals in the later years compared to the earlier years, an artifact of the missing data in the earlier years. However, they are not very evenly distributed, hinting at unequal variances.

Q5 c)
When setting $df = 2$, the resulting curve is very smooth and almost seems linear.
When setting $df = 100$, the resulting curve follows the data points much more closely but also does not seem to generalize the trend well, most likely an overfit.
I would pick the curve from 5a), as according to R's helpfile, a systematic approach is used to estimate the optimal degrees-of-freedom (leave-one-out cross-validation), which according to results was around 21. The curve that resulted from this seems to avoid overfitting too much while still smoothing out the data, striking a balance compared to the arbitrarily picked values of df (2 or 100). Using cross-validation allows for a model that generalizes well to unseen data as it simulated "unseen" data with the test set of data.

Q6 a)

```
before.1750.quantiles <- quantile(before.1750.df$Flowering.DOY, na.rm = TRUE)
before.1750.quantiles[["75%"]]
```

```
## [1] 109
```

The 75^{th} percentile is 109.

Q6 b)

```
approx.under.1750.75.perc <- 0.75 * length(after.1750.df$Flowering.DOY)
```

If the distribution of flowering-days were unchanged over time, then approximately 198.75 years post-1750 should be below the pre-2750 75^{th} percentile.

Q6 c)

```
actual.under.1750.75.perc = length(which(
  after.1750.df$Flowering.DOY < before.1750.quantiles[["75%"]]))
```

The number of years post-1750 that is actually below the pre-1750 75^{th} percentile is 200.

Q6 d)

There are 0 intervals pre-1750 that satisfy the conditions.

Q6 e)
The ratio of number of intervals found above to number of possible intervals is a not p-value. It would seem like it is a p-value of the following test:

Let $p_{75,pre}$ be the 75^{th} percentile of pre-1750 flowering day-of-year
Let $p_{75,post}$ be the 75^{th} percentile of pre-1750 flowering day-of-year

$H_0 : p_{75,pre} = p_{75,post}$
$H_a : p_{75,pre} \geq p_{75,post}$

But due to the overlapping of intervals, where there are 264 common samples between each neighbouring interval, there is implicitly covariance between nearby intervals. Since this proportion is the result of all these non-independent samples, it is not an accurate estimate of the probability of more extreme values than observed assuming non-changing distributions over time. In significance testing with p-values, it is usually assumed that observations and samples are independent of each other.
