lab 5

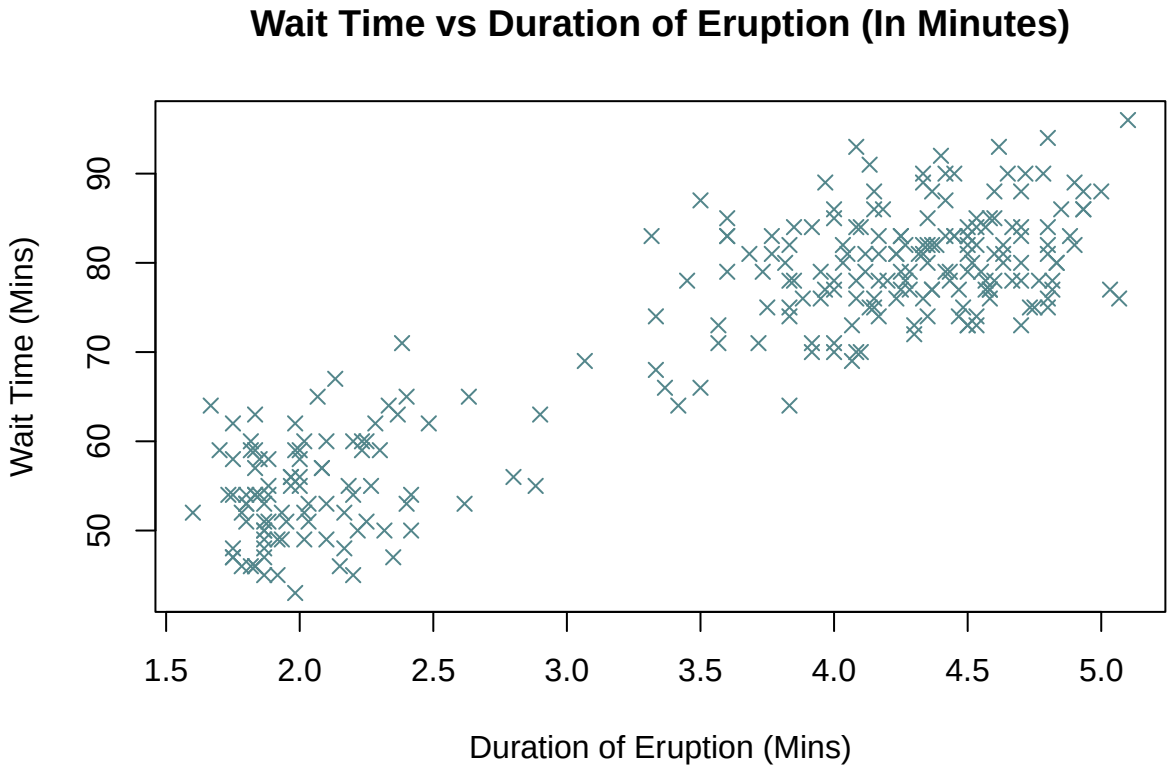
2023-10-05

Ye Olde Faithful

Setting up the preliminaries, (i.e. reading the file)

```
faithful <- read.csv("./data/faithful.csv")
duration <- faithful$duration
wait <- faithful$wait
```

Now, we check if the duration of the eruption and the waiting time are linearly related. Here, I'm assuming that we want to check if a longer duration of the eruption means a longer waiting time. So, waiting time is the "prediction" and duration of eruption is the "predictor". But I think it's possible to try to predict it the other way around too? Maybe a longer waiting time for the next eruption means the duration of the eruption will be longer. I'm not sure what this would change though but it seems like the lab is asking for the one I did so yay.



It does seem possible that the duration of eruption and the wait time are linearly related. This relation is a positive one, since it looks like the wait time is could be increasing as the duration of eruption increases.

Regressing to Before I Was a Geyser

Performing a linear regression and storing it:

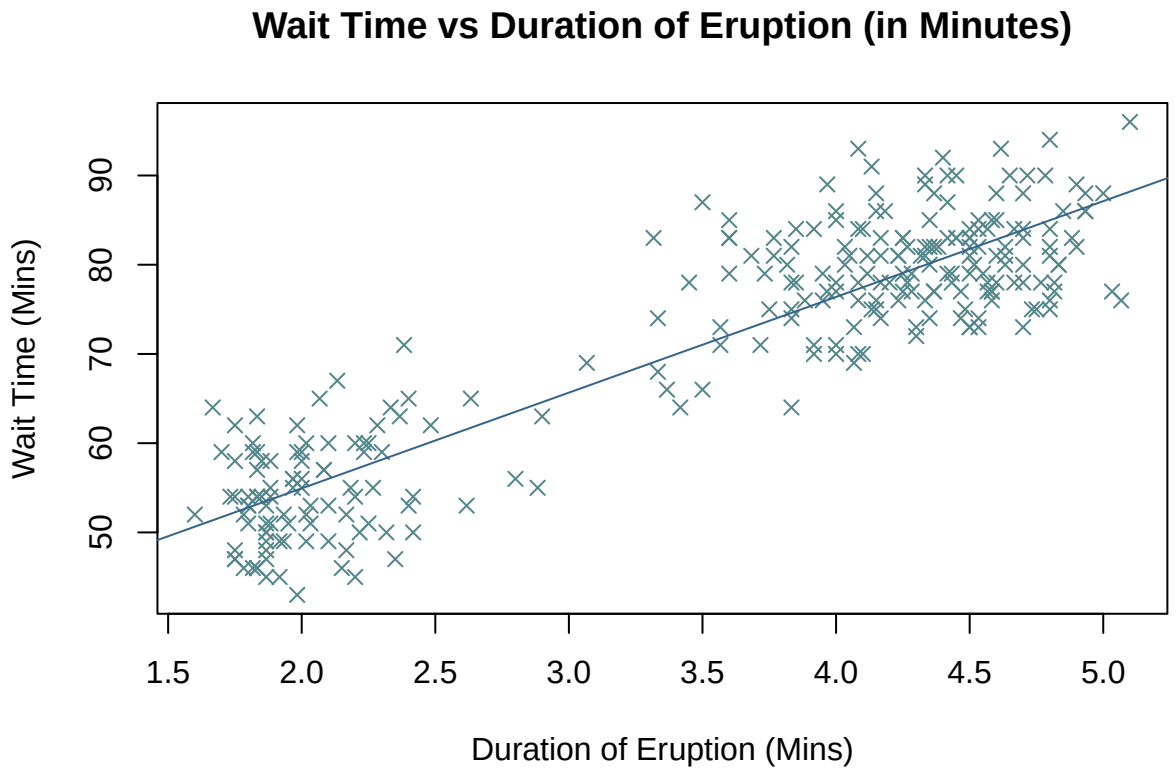
```
linear_regression <- lm(wait~duration)
coefficients <- linear_regression$coefficients
```

The coefficients are:

```
## (Intercept)    duration
##    33.47440    10.72964
```

Rounded up to 3 s.f., the estimated intercept is `33.5` and the slope is `10.7`. The equation of the regression line would thus be `y = 10.7x + 33.5`, where `y` is the wait time and `x` is the duration of eruption.

Re-plotting and adding the line since I couldn't figure out a better way to organise this:



Now we do everything again, but this time in standard units, and we also compute the correlation coefficient:

```
duration_su <- (duration - mean(duration))/sd(duration)
wait_su <- (wait - mean(wait))/sd(wait)

regression_su <- lm(wait_su~duration_su)
coefficients_su <- regression_su$coefficients
correlation_coefficient_su <- cor(duration_su, wait_su)
```

Rounding the slope and the correlation coefficient to 3 s.f., we see that:

```
## [1] The slope is 0.901, and the correlation coefficient is 0.901.
```

```
## [1] It is TRUE that the slope and the correlation coefficient are the same.
```

Prediction

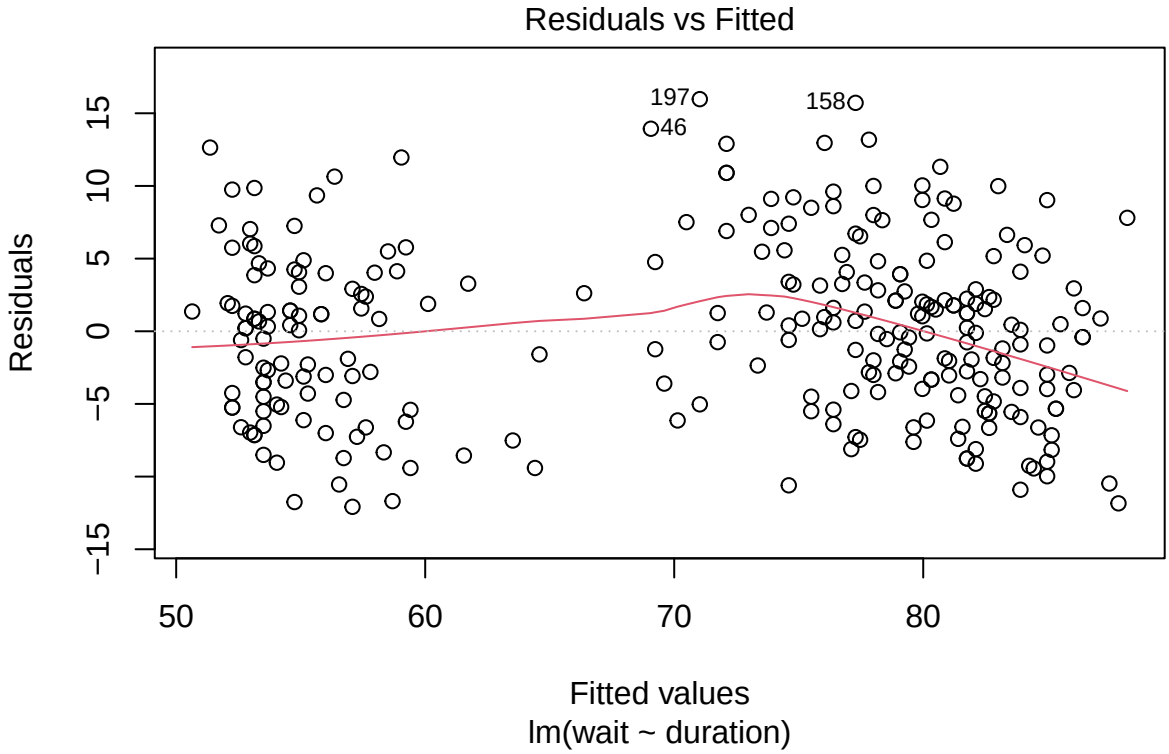
Calculating the prediction for 2 minutes and 5 minutes and printing out the required sentences:

```
## After an eruption lasting 2 minutes,
```

```
##  we predict that you'll wait 54.9 minutes until the next eruption.

## After an eruption lasting 5 minutes,
##  we predict that you'll wait 87 minutes until the next eruption.
```

Now, let's check the residuals:



It seems that the line isn't very straight, so our linear model was not reasonable. Perhaps it would have been better to do a regression for wait times below 70 minutes and wait times above 70 minutes, as it looks like there are two "clouds" of data points before and after the 70 minute point.