|  |
| --- |
| Here’s my belated Statistics Sunday post, using a cool technique I just learned about: [gghighlight](https://www.littlemissdata.com/blog/highlight). This R package works with ggplot2 to highlight a subset of data. To demonstrate, I’ll use a dataset I analyzed for a previous post about my [2017 reading habits](http://www.deeplytrivial.com/2018/01/data-analysis-of-my-2017-reading.html). [Side note: My [reading goal for this year](https://www.goodreads.com/challenges/7501-2018-reading-challenge) is 60 books, and I’m already at 43! I may have to increase my goal at some point.]  setwd("~/R") library(tidyverse)  books<-read\_csv("2017\_books.csv", col\_names = TRUE)  ## Warning: Duplicated column names deduplicated: 'Author' => 'Author\_1' [13]  ## Parsed with column specification: ## cols( ## .default = col\_integer(), ## Title = col\_character(), ## Author = col\_character(), ## G\_Rating = col\_double(), ## Started = col\_character(), ## Finished = col\_character() ## )  ## See spec(...) for full column specifications.  One analysis I conducted with this dataset was to look at the correlation between book length (number of pages) and read time (number of days it took to read the book). We can also generate a scatterplot to visualize this relationship.  cor.test(books$Pages, books$Read\_Time)  ##  ## Pearson's product-moment correlation ##  ## data: books$Pages and books$Read\_Time ## t = 3.1396, df = 51, p-value = 0.002812 ## alternative hypothesis: true correlation is not equal to 0 ## 95 percent confidence interval: ## 0.1482981 0.6067498 ## sample estimates: ## cor  ## 0.4024597  scatter <- ggplot(books, aes(Pages, Read\_Time)) +  geom\_point(size = 3) +  theme\_classic() +  labs(title = "Relationship Between Reading Time and Page Length") +  ylab("Read Time (in days)") +  xlab("Number of Pages") +  theme(legend.position="none",plot.title=element\_text(hjust=0.5))    There’s a significant [positive](http://www.deeplytrivial.com/2018/01/statistical-sins-types-of-statistical.html) correlation here, meaning the longer books take more days to read. It’s a moderate correlation, and there are certainly other variables that may explain why a book took longer to read. For instance, nonfiction books may take longer. Books read in October or November (while I was gearing up for and participating in NaNoWriMo, respectively) may also take longer, since I had less spare time to read. I can conduct regressions and other analyses to examine which variables impact read time, but one of the most important parts of sharing results is creating good data visualizations. How can I show the impact these other variables have on read time in an understandable and visually appealing way?  gghighlight will let me draw attention to different parts of the plot. For example, I can ask gghighlight to draw attention to books that took longer than a certain amount of time to read, and I can even ask it to label those books.  library(gghighlight)  scatter + gghighlight(Read\_Time > 14) +  geom\_label(aes(label = Title),  hjust = 1,  vjust = 1,  fill = "blue",  color = "white",  alpha = 0.5) |

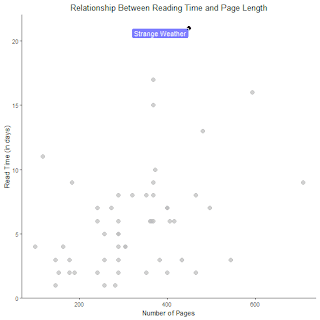
Here, the gghighlight function identifies the subset (books that took more than 2 weeks to read) and labels those books with the Title variable. Three of the four books with long read time values are non-fiction, and one was read for a course I took, so reading followed a set schedule. But the fourth is a fiction book, which took over 20 days to read. Let’s see how month impacts reading time, by highlighting books read in November. To do that, I’ll need to alter my dataset somewhat. The dataset contains a starting date and finish date, which were read in as characters. I need to convert those to dates and pull out the month variable to create my indicator.

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

books$Started <- mdy(books$Started)  
books$Start\_Month <- month(books$Started)  
books$Month <- ifelse(books$Start\_Month > 10 & books$Start\_Month < 12, books$Month <- 1,  
 books$Month <- 0)  
scatter + gghighlight(books$Month == 1) +  
 geom\_label(aes(label = Title), hjust = 1, vjust = 1, fill = "blue", color = "white", alpha = 0.5)

[](https://i0.wp.com/2.bp.blogspot.com/-CHyZ17fDkWo/W2ncu6eYNbI/AAAAAAAANYs/mPH6cmNgLu88tYOTttaK-yGNEl9xxlnOgCLcBGAs/s1600/unnamed-chunk-4-1.png?ssl=1)

The book with the longest read time was, in fact, read during November, when I was spending most of my time writing.