pitches Dataset

The Dataset

As someone who cut his statistical teeth on sports data, I couldn't help but include a baseball dataset in these labs. This dataset—*pitches* in the *pitchRx* library—is a subset of one of the more consequential datasets in baseball. The *pitches* dataset contains 70 variables on over 1,800 pitches thrown by Mariano Rivera and Phil Hughes in 2011. I won't list all the variables, as many of them aren't the most useful, but here are a listing of some of the more interesting ones.

- *type*: A categorical variable showing the result of the pitch
- *start_speed*: The release speed of the pitch.
- *end_speed*: The speed of the pitch when it crosses the plane of the plate.
- *sz_top*: The top of the strike zone for a batter.
- *sz_bot*: The bottom of the strike zone for a batter.
- *pfx_x*: The amount of movement in the horizontal plane (From the catcher's point of view) compared to a pitch thrown from the same location and same speed with no spin.
- *pfx_x*: The amount of movement in the vertical plane (From the catcher's point of view) compared to a pitch thrown from the same location and same speed with no spin.
- *px*: Where a pitch crosses the front edge of the plate in the horizontal plane.
- *pz*: Where a pitch crosses the front edge of the plate in the vertical plane.
- *pitch_type*: The best guess at the type of pitch.
- *spin_rate*: The estimated spin (in RPM) of the pitch
- *on_1b*: Player ID for a runner on first base, NA otherwise.
- *on_2b*: Player ID for a runner on second base, NA otherwise.

- *on_3b*: Player ID for a runner on third base, NA otherwise.
- *stand*: The side of the plate the batter is standing on.
- *pitcher_name*: Pitcher Name. For this dataset, either Mariano Rivera or Phil Hughes.

Possible Questions

The number of possible questions in this dataset is huge. Many of them have been covered by sabermetricians over the years, some that you could think of are still unsolved. However, here's a short list, with the ones we'll go through highlighted.

1. **Does Phil Hughes throw more strikes than the average pitcher? (One-sample proportion)**
2. **Does Phil Hughes or Mariano Rivera get more strikes called on pitches outside the strike zone? (Two-sample proportion)**
3. Does Mariano Rivera pitch inside to right-handed batters more often than left-handed batters? (Two-sample proportion)
4. Do pitchers throw more fourseam fastballs when behind in the count? (Two-sample proportion)
5. Does Phil Hughes or Mariano Rivera get more horizontal movement on their cutters (Cut Fastballs)? (Two-sample mean)
6. Does Mariano Rivera keep pitches lower in the zone with runners on base? (Two-sample mean)
7. Is spin rate correlated with movement? (Correlation)
8. **How does initial speed affect speed at the plate? (Regression)**
9. Who has a better cutter: Phil Hughes or Mariano Rivera?*
10. **What attributes makes a good fastball?***

** I would denote this a challenge question. Challenge questions are ones that, due to the open-ended nature of the question or coding that can stretch the student or may not even be in book, challenges students to a certain degree.*

Question 1

Phil Hughes has always been known as a command pitcher. In 2014, he set the record for the highest strikeout per walk ratio with 11.625. Thus, it would stand to reason that he would throw a higher rate of strikes than the average, which was 63.2% (Courtesy of a little online research) in 2011. We can test to see if this assumption holds up, with the hypotheses

$$H_0 : p = 0.632$$
$$H_A : p > 0.632$$

In order to investigate this, we can look at the *type* variable, which has three levels: "B" for a ball, "S" for a strike, and "X" for contact. Generally speaking, contact is included in strike counts, so we'll add the "S" and "X" counts together to get our proportion of $\hat{p} = 0.702$.

Ball	Strike	Contact
282	481	183

Using the *prop.test* function, we can see that our test statistic will be $t = \sqrt{19.875} = 4.458$ with a p-value of $P\big(N(0,1) > 4.458\big) = 4.133^{-6}$. Importantly, our assumptions in terms of sample size and observed successes and failures appear to be met. Thus, we would confirm that Phil Hughes does seem to throw strikes at a higher rate than league average.

```
library(pitchRx)
table(pitches$type[pitches$pitcher_name==''Phil Hughes'',]
prop.test(x=481+183,n=481+183+282, p=0.632,
alternative=''greater'', correct=F)
```

Question 2

There is long a precedence in baseball that respected pitchers get more calls to go their way. They'll get a few extra strike calls outside the zone, much to the frustration of batters. While Phil Hughes has had a fine career, few players were as respected as Mariano Rivera—to the point that he would go on to be the first player elected to the Hall of Fame unanimously. It would be reasonable to think that Rivera would get a few extra calls.

In order to investigate this, we need to first see if each pitch's location should have been called a ball. A ball is any pitch that has one of the following conditions...

- $pz > sz\_top$
- $pz < sz\_bot$
- $|px| > \dfrac{8.5}{12}$

This last condition can be derived from the fact that baseball's home plate is 17 inches wide, with the center of the plate resulting in a *px* value of 0. From these conditions, we can define a variable *ooz*—short for "out of zone."

Now we need to cut down out dataset to only consider pitches where the batter did not swing at the pitch (And the batter wasn't hit by the pitch) that were out of the zone. This can be done using the *des* variable, which classifies the results of the pitch into a number of categories. These resulting pitches are the ones where there's a chance that the pitcher was given the benefit of the doubt. This results in 676 pitches—332 for Mariano Rivera and 344 for Phil Hughes. Using the *table* function, we can see the balls and strikes on these out of zone pitches.

	Balls	Strikes
Rivera	257	75
Hughes	266	78

In order to see if Rivera gets more respect than Hughes on these out of zone pitches—which would mean he gets a higher percentage of these out of zone pitches called as strikes—we would want to test the following hypotheses at the $\alpha = 0.05$ level.

$$H_0 : p_{Rivera} - p_{Hughes} = 0$$
$$H_A : p_{Rivera} - p_{Hughes} > 0$$

The *prop.test* function gives us our answer. With a p-value of 0.5104, we would fail to reject the null hypothesis and conclude that it's plausible that Rivera and Hughes have the same proportion of out of zone strikes called. Additionally, the assumptions related to sample size, successes, and failures seem to be met.

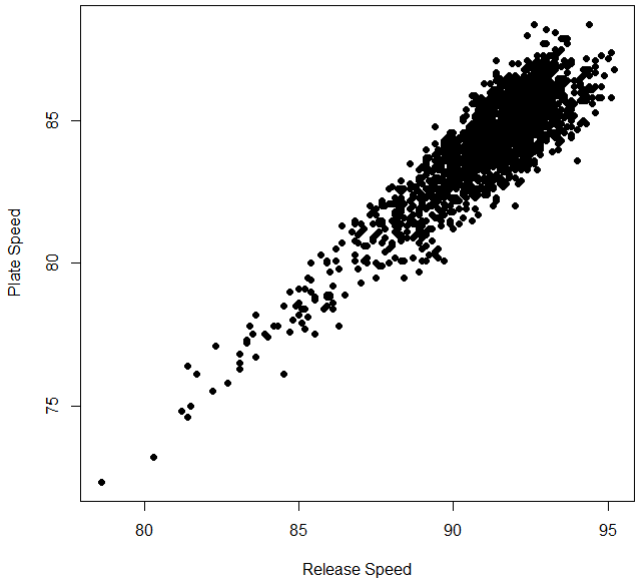
I should note that there are better ways to answer this question. This test does not account for the probability that the pitch would be called a strike normally, as pitches closer to the strike zone should be more likely to be called a strike. Fully answering the question would likely involve using this probability for each pitch and seeing what proportion of strikes we'd expect to see given this information.

```
library(pitchRx)
pitches$ooz=ifelse(pitches$pz>pitches$sz_top|pitches$pz<pitches$sz_bot|
abs(pitches$px)>8.5/12,''yes'',''no'')
balls=pitches[pitches$ooz==''yes'',]
table(balls$ooz,balls$pitcher_name)
prop.test(x=c(75,78), n=c(75+257,89+266),
alternative=''greater'', correct=F)
```

Question 8

It makes intuitive sense that release speed and at plate speed would be correlated. Obviously, the harder the pitch is thrown, the harder it will be at the plate. However, it's reasonable

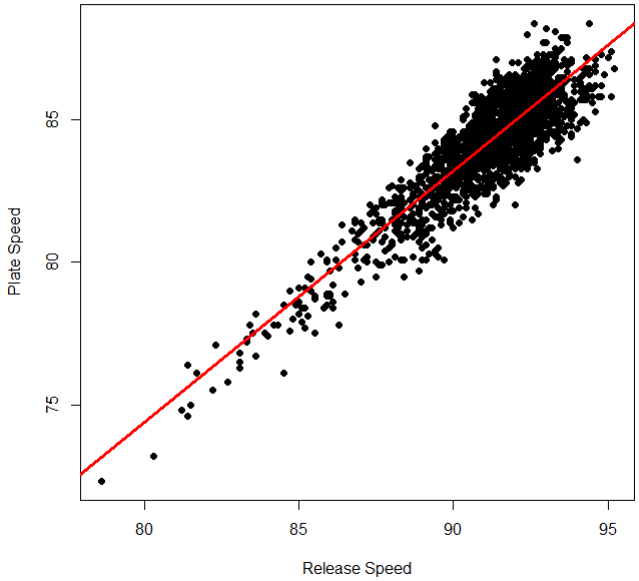
to ask how does it affect release speed affect plate speed. Essentially, what is the sort of deceleration factor? A physicist could calculate a theoretical factor based on air density, spin, etc., but here we'll look at the observed data.



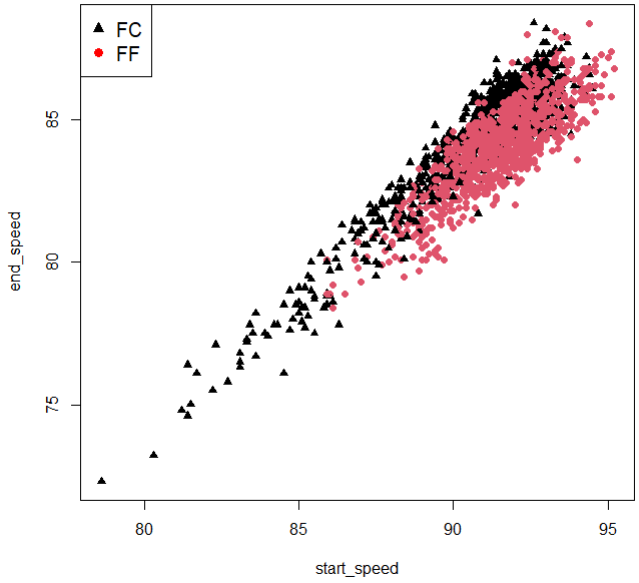
There's a clear linear relationship between the variables, with some error around the line. It seems like—in this range—pitches lose around 7.5 MPH on their pitches from release to plate. Linear regression is the clear answer to investigating the relationship between the variables, and is simple to implement.

```
lm(formula = end_speed ~ start_speed, data = pitches)

Residuals:
     Min       1Q   Median       3Q      Max
-3.13118 -0.63432  0.07318  0.68425  2.90282

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.87675    0.98598   3.932 8.74e-05 ***
start_speed  0.88143    0.01082  81.441  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9232 on 1856 degrees of freedom
Multiple R-squared:  0.7814,    Adjusted R-squared:  0.7812
F-statistic:  6633 on 1 and 1856 DF,  p-value: < 2.2e-16
```

There's a lot we can get from the regression results. First off, unsurprisingly release speed has an effect on plate speed, seen through the hypothesis test results for *start_speed*. Further, this is a fairly good fit for our data, evidenced by the R^2 value of 0.7814 and the residual standard error of $\hat{\sigma}^2 = 0.92$. We can see that the regression line splits the data well, and a little calculation of predicted values show that on the average, pitches lose 6-8 MPH from release to the plate.



It's interesting to note that if students want to continue investigating this question, there's a little more that can be extracted from this data. If we split up pitches by their pitch type, it seems like fourseam fastballs and cutters exhibit slightly different behavior. It would be interesting to split up the data and analyze it separately, or for those familiar with multiple regression including pitch type as another variable in the model.



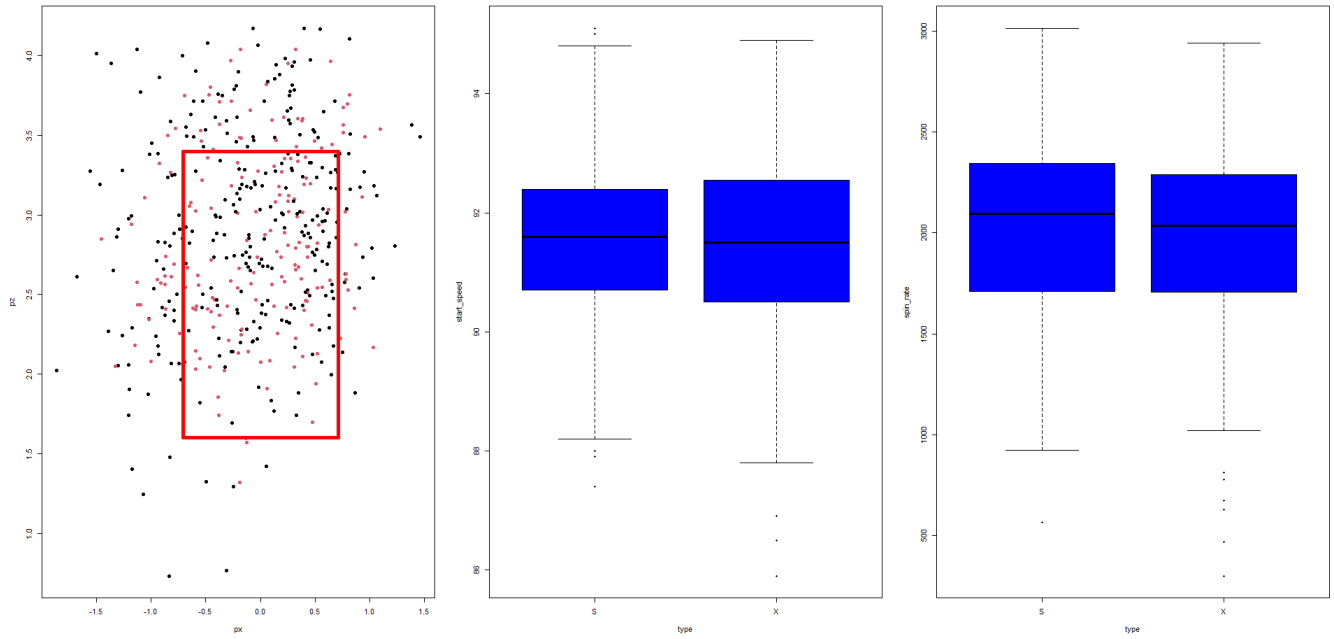

```
plot(pitches$start_speed, pitches$end_speed, pch=16,
xlab=''Release Speed'', ylab=''Plate Speed'')
lm=lm(end_speed~start_speed, data=pitches)
summary(lm)
abline(lm, col=''red'', lwd=3)
plot(pitches$start_speed, pitches$end_speed,
xlab=''Release Speed'', ylab=''Plate Speed'',
pch=ifelse(pitches$pitch_type==''FF'',16,17),
col=as.factor(pitches$pitch_type))
predict(lm,newdata=data.frame(start_speed=c(80, 85, 90, 95))
```

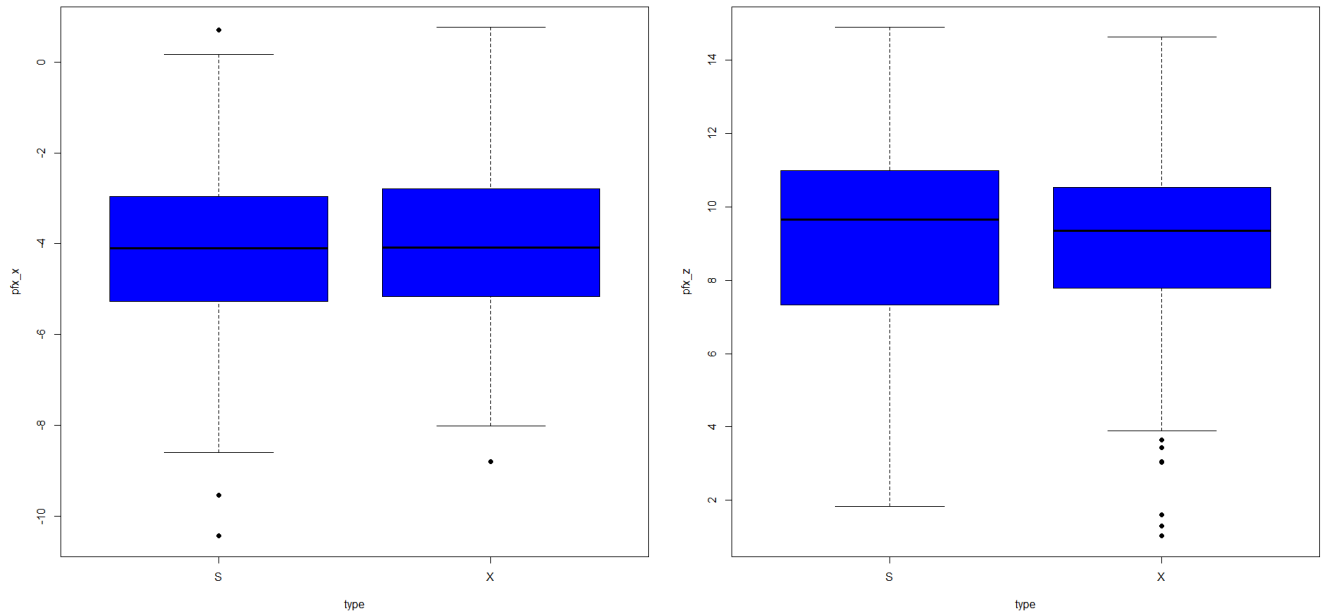
Question 10

This is a wonderful question because there are so many routes you can go with it. The biggest challenge within the question, in my mind, is how do you define a good fastball? Is it a results-based criterion, for example dealing with generating swinging strikes from batters or inducing weak contact? Is it attribute based, looking at speed, spin, or movement? Is it independent of location? There are so many ways you can go, which is what I believe makes this open-ended question so great!

If I'm fully answering the question, I tend to go somewhere in between. However, for the sake of keeping this fairly simple I'm going to define a good fastball as being one that generates little contact given that the batter is swinging. So, we'll need to look at what attributes induce a batter to swing and miss. I'll look primarily at five variables for this:

plate location, speed (*start_speed*), horizontal movement, vertical movement, and spin rate. Further, in order to simplify this question I'll only be looking at fourseam fastballs, and only on pitches where the batter swung (Given by the *des* variable and little coding).





Looking at this, it's hard to see many differences, but there are a few things. First off, it seems like there are potential tiny differences between contact and whiffs in speed,

spin rate, and vertical movement. As the majority of a fourseam fastball's spin goes to vertical movement, I'm going to eliminate vertical movement as a potential variable. A quick hypothesis test to see if the speed or spin rate differs between whiffs and contacts fails to reject either null hypothesis ($H_0 : \mu_{Whiff} - \mu_{Contact} = 0$) at the $\alpha = 0.05$ level.

Interpreting the result of plate location seems to be a bit of a challenge, but with a little more work we begin to see some results. Let's create a variable that simplifies the height of the pitch to either high ($pz > 3.25$), low ($pz < 1.8$), or middle ($1.8 \leq pz \leq 3.25$). These numbers are somewhat arbitrarily chosen, though both are about a baseball's width from the average upper or lower edges of the zone. We can then look at if the contact rates differ for these regions.

	High	Low	Mid
Whiff	92	21	172
Contact	48	8	123

Keeping the pitch out of the middle heights seems to be imporant, which a quick hypothesis test ($H_A : p_{High/Low} - p_{Mid} > 0$, where p is the proportion of strikes) confirms. However, which is more important: keeping the ball up or down? For this, we actually have to go back a step; considering how often high pitches versus low pitches are swung at. When looking at all fourseam fastballs—regardless of if the batter swung—we can see that low pitches are much less likely to be swung at, confirmed via a quick hypothesis test ($H_A : p_{High} - p_{Low} > 0$, where p is the proportion of swings).

	High	Low	Mid
No Swing	129	90	238
Swing	140	29	295

Thus, it seems like high fastballs are key to success, when it comes to defining success as limiting contact while inducing swings. I'll note that this is probably not the best way to try to answer this question, even as I've defined success. A better way would be to use logistic regression to try and see what variables help us model the probability of a whiff using logistic regression. Alas, that's outside the scope of this material. Additionally, more in-depth answers would likely find that there is interactions with variables. One I particularly have in mind is that high speed, high spin fastballs are likely extremely effective when high in the zone. However, students have a wide variety of options when trying to figure out this question.

```
library(pitchRx)
pitches$swing=ifelse(pitches$des==''Ball''|pitches$des==''Ball In Dirt''|
pitches$des==''Called Strike''|pitches$des==''Hit By Pitch'',
``yes'',''no'')
ff=pitches[pitches$pitch_type==''FF'',]
ff_swing=ff[ff$swing==''yes'',]
```

```
plot(ff_swing$px, ff_swing$pz, pch=16, col=as.factor(ff_swing$type))
boxplot(start_speed~type, data=ff_swing)
boxplot(spin_rate~type, data=ff_swing)
boxplot(pfx_x~type, data=ff_swing)
boxplot(pfx_z~type, data=ff_swing)
t.test(start_speed~type, data=ff_swing,
alternative=''greater'', var.equal=T)
t.test(spin_rate~type, data=ff_swing,
alternative=''greater'', var.equal=T)
table(ff_swing$swing, ifelse(ff_swing$pz>3.25,''high'',
ifelse(ff_swing$pz<1.8,''low'',''mid'')))
prop.test(x=c(92+21,172), n=c(92+21+48+8, 172+123),
alternative=''greater'', correct=F)
table(ff$type, ifelse(ff_swing$pz>3.25,''high'',
ifelse(ff_swing$pz<1.8,''low'',''mid'')))
```