Sample Baseball Data Report

Matthew Coleman

4/19/2019

```
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("tibble")
library("readxl")
library("knitr")
library("kableExtra")
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

Rank Batted Balls by highest to lowest quality

```
baseball_data <- read_excel("SampleData.xlsx")
```

I began by observing the data and thinking about how I wanted to approach the question. I took a look at the columns and values that were given and determined what I saw as relevant data to my rankings. From here, I would query relevant data and observe the results.

```
#baseball_data

#Omitted result of running baseball_data because of formatting.
```

I am going to rank the bats in terms of impact, while minimizing OutsOnPlay. This way I can maximize impact while minimizing the risk associated with the bat. Therefore, I am going to approach the ranking system in the following hierarchy:

1. Home run

i) Ordered in descending order of runs scored

2. Triple

i) Ordered in ascending order of outs on play and descending runs scored

3. Double

i) Ordered in ascending order of outs on play and descending runs scored

4. Walk
5. Single

i) Ordered in ascending order of outs on play and descending runs scored

6. Sacrifice

i) Ordered in ascending order of outs on play and descending runs scored

Note(1): Descending order means that the highest ranked level has the "largest number", while ascending order means that the highest rank has the lowest number. Ex: 4 runs is the highest rank for the "RunsScored" column, while 0 outs is the highest rank for the "OutsOnPlay" column

Note(2): While a walk is not technically a "batted ball", it fits into the rankings as higher than a single, because there is no risk for an out, where with a single there is risk associated with it.

```
#Below, I took all the hit data that was relevant. This means I took all the acutal bats (including wal
#observations that were not useful for the ranking.)
hit_data <- filter(baseball_data, PlayResult == "HomeRun" | PlayResult == "Triple" |
                     PlayResult == "Double" | PlayResult == "Single" |
                     PlayResult == "Sacrifice"| KorBB == "Walk" )

#Next, I took all the relevant data, and arranged it according to my hierarchy I created.
ranked_hit_data <- hit_data %>% arrange(desc(RunsScored),factor(PlayResult,
            levels =  c("HomeRun", "Triple", "Double","Undefined" ,"Single", "Sacrifice")),
            OutsOnPlay)

#Finally, I am going to remove unnecessary data so that the table can be presented.

clean_ranked_data = select(ranked_hit_data, PitchNo, PitcherId, PitcherThrows,
                           PitcherTeam, BatterId, BatterSide, Inning, KorBB,
                           HitType, PlayResult, OutsOnPlay, RunsScored)
```

Presenting the data:

PitchNo	PitcherId	PitcherThrows	PitcherTeam	BatterId	BatterSide	Inning	KorBB	HitType	PlayResult	OutsOnPlay	RunsScored
111	1000054802	Left	STE_LUM	8883473	Right	4	Undefined	FlyBall	HomeRun	0	1
164	1000054802	Left	STE_LUM	1000023664	Right	5	Walk	Undefined	Undefined	0	1
279	1000025592	Right	STE_LUM	8889133	Right	8	Walk	Undefined	Undefined	0	1
285	8893832	Left	STE_LUM	1000023664	Right	8	Walk	Undefined	Undefined	0	1
118	1000054802	Left	STE_LUM	1000023664	Right	4	Undefined	LineDrive	Single	0	1
158	1000054802	Left	STE_LUM	8889133	Right	5	Undefined	LineDrive	Single	0	1
262	1000025592	Right	STE_LUM	8901602	Right	8	Undefined	Popup	Single	0	1
153	1000054802	Left	STE_LUM	8883473	Right	5	Undefined	FlyBall	Sacrifice	1	1
260	1000025592	Right	STE_LUM	1000053906	Right	8	Undefined	FlyBall	Triple	0	0
49	1000054802	Left	STE_LUM	8882482	Right	2	Undefined	FlyBall	Double	0	0
83	1000054802	Left	STE_LUM	1000023663	Right	3	Undefined	LineDrive	Double	0	0
117	1000054802	Left	STE_LUM	8889133	Right	4	Undefined	FlyBall	Double	0	0
17	1000049209	Right	UCSB_GAU	1000025590	Right	1	Walk	Undefined	Undefined	0	0
275	1000025592	Right	STE_LUM	8882482	Right	8	Walk	Undefined	Undefined	0	0
57	1000054802	Left	STE_LUM	1000023664	Right	2	Undefined	GroundBall	Single	0	0
123	1000054802	Left	STE_LUM	8901601	Right	4	Undefined	GroundBall	Single	0	0
144	1000054802	Left	STE_LUM	1000053906	Right	5	Undefined	LineDrive	Single	0	0
148	1000054802	Left	STE_LUM	8901602	Right	5	Undefined	LineDrive	Single	0	0
184	1000054994	Right	STE_LUM	1000023663	Right	6	Undefined	LineDrive	Single	0	0
239	1000049219	Left	UCSB_GAU	1000025598	Right	8	Undefined	LineDrive	Single	0	0
269	1000025592	Right	STE_LUM	8883473	Right	8	Undefined	LineDrive	Single	0	0

The table above shows the model which involved ranking the batted balls in such a way that maximized impact and minimized risk. The hierarchy can be clearly decoded: rate first by the runs scored on the play, and then rank by the play result (Home Run, Triple, etc.) and the Outs on the play. The same procedure was repeated for the bats where no runs were scored.

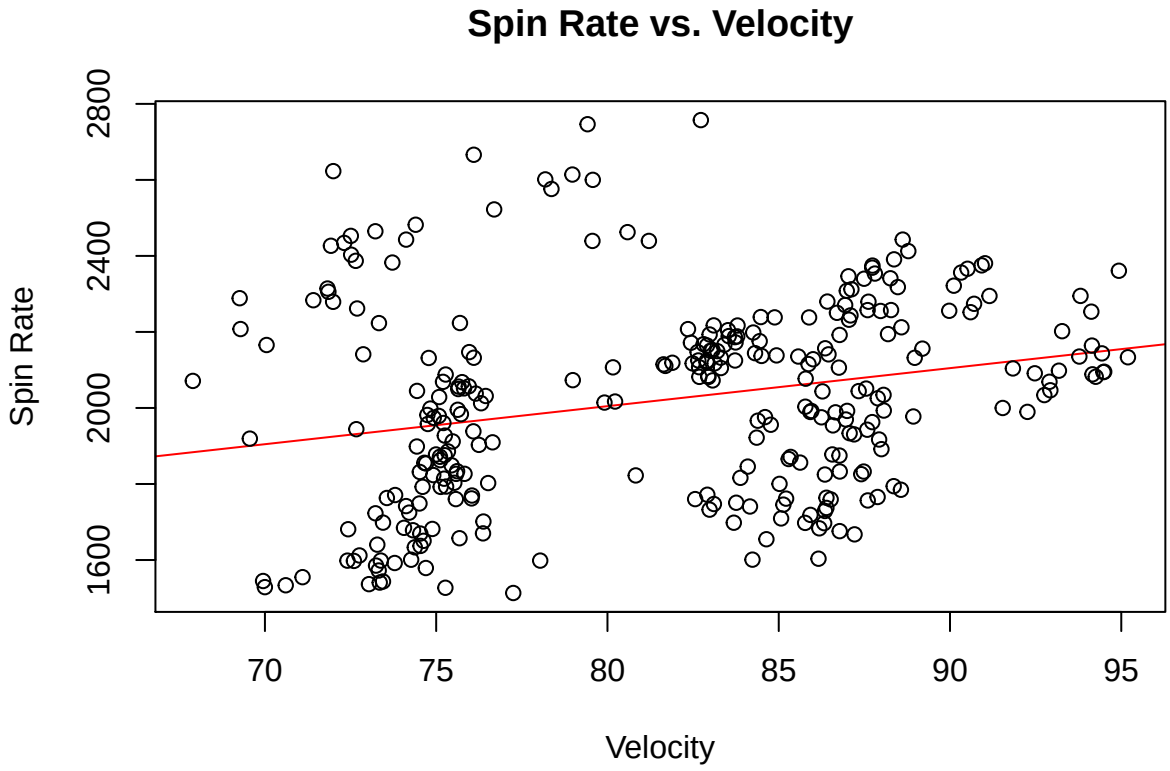
Assess the relationship between Velocity (RelSpeed) and Spin Rate (SpinRate).

To asses the relationship between velocity and spin rate, I am going to first extract the data, and plot velocity against spin rate.

```
velocity <- baseball_data$RelSpeed
spin_rate <- baseball_data$SpinRate

rline <- lm(spin_rate~velocity)
```

```
plot(velocity, spin_rate, main = "Spin Rate vs. Velocity", xlab = "Velocity",
     ylab = "Spin Rate", abline(rline, col = "red"))
```



As observed by the plot, there is a very slight, if any, linear trend between Velocity and Spin Rate. Another observation to be made is that as Velocity of a pitch increases, the variability in Spin Rate decreases. This suggests slower pitches have a greater variability in Spin Rate, meanwhile faster pitches have similar Spin Rates on a pitch-to-pitch basis. Next, I am going to asses R^2, the coefficient of determination.

```
summary(rline)
```

```
##
## Call:
## lm(formula = spin_rate ~ velocity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -463.96 -196.05    7.81  151.37  747.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1204.306    181.652   6.630 1.61e-10 ***
```

```
## velocity      10.005      2.219   4.508 9.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 253.8 on 294 degrees of freedom
## Multiple R-squared:  0.06466,    Adjusted R-squared:  0.06148
## F-statistic: 20.33 on 1 and 294 DF,  p-value: 9.446e-06
```

Here, we can see R^2 is .06466. This means that 6.47% of the variance in spin rate can be explained by a linear relationship with velocity. By this and the plot, I observe there being a weak linear relationship between velocity and spin rate. If I run a test of $H_o : \beta_1 = 0 \, vs \, H_a : \beta_1 \neq 0$ at an $\alpha = .01$ significance level, then at a p-value of 9.45×10^{-6}, I reject the null that $\beta_1 = 0$, and suggest that $\beta_1 \neq 0$. This suggests there is a possible linear relationship between Velocity and Spin Rate, even if it is very small.

I observed the scatter plot is multiple groupings of data points at locations where RelSpeed = 75 mph, 85, and 95. This is consistent with different pitches, where in the 75-85 mph range, there are more "changeup" and "breaking ball" pitches, where the 90-95mph range can have more "fastball pitches". There are also other factors such as release angle and grip which can change the spin rate of the ball. Therefore, there would need to be analysis into separate factors to determine a relationship with Spin Rate.

#####I will take another approach which involves separating the different pitchers to see if the variability in Spin Rate can be reduced.

Because the data contains Spin Rate and Velocity for all pitchers, I am going to group the pitchers and then run an analysis of Spin Rate vs. Velocity for each to remove one possible confounding variable.

```
pitcher_data <- baseball_data %>% group_by(PitcherId) %>% arrange(PitcherId)

#pitcher_data %>% count(PitcherId) (Output below)
```

Pitcher ID	Observations
8893302	18
8893832	8
1000023657	9
1000025592	26
1000049209	48
1000049218	1
1000049219	69
1000054802	92
1000054994	25

Using the above data, I can see that there are 9 separate pitchers. I am going to group 8 of them as following below:

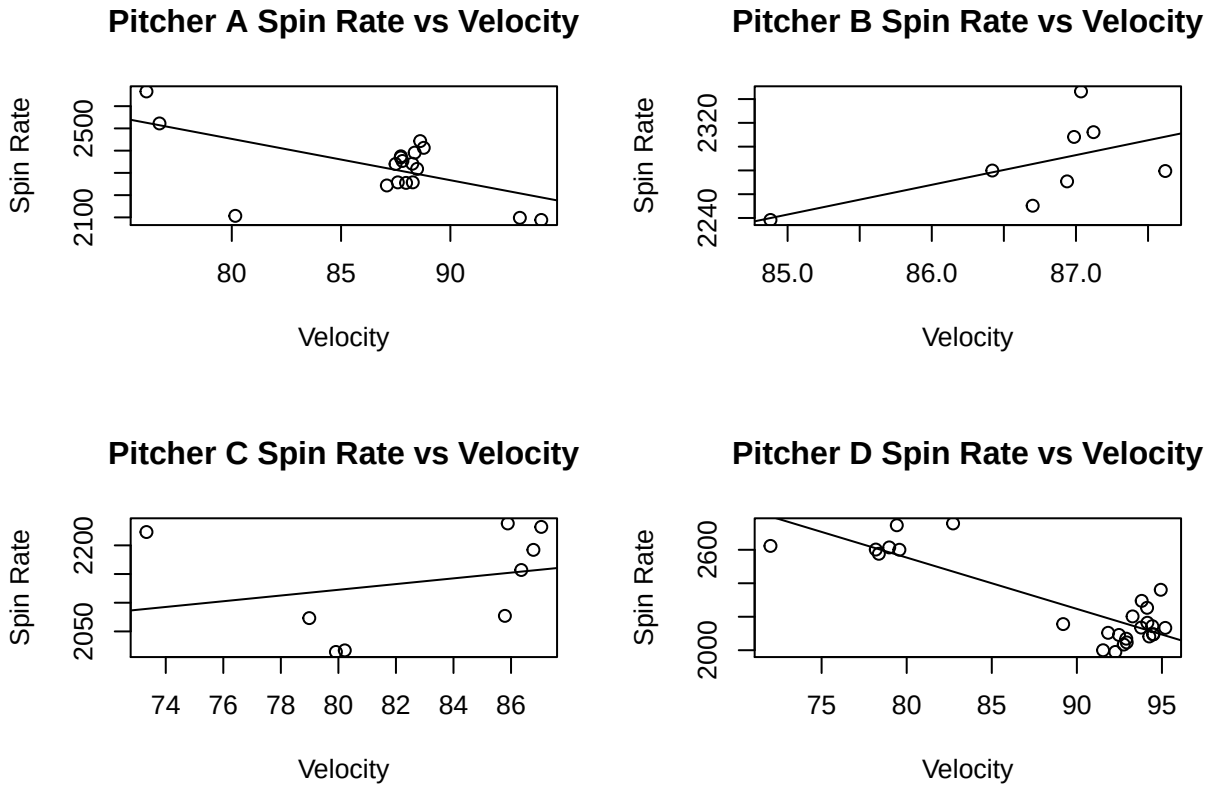
Pitcher ID	Pitcher
8893302	A
8893832	B
1000023657	C
1000025592	D
1000049209	E
1000049219	F
1000054802	G
1000054994	H

Notice that Pitcher #1000049218 has only one observation. There will not be meaningful analysis that can be done on this pitcher, so I will exclude this pitcher. The pitchers are assigned below:

```
pitchera <- filter(pitcher_data, PitcherId == "8893302")
pitcherb <- filter(pitcher_data, PitcherId == "8893832")
pitcherc <- filter(pitcher_data, PitcherId == "1000023657")
pitcherd <- filter(pitcher_data, PitcherId == "1000025592")
pitchere <- filter(pitcher_data, PitcherId == "1000049209")
pitcherf <- filter(pitcher_data, PitcherId == "1000049219")
pitcherg <- filter(pitcher_data, PitcherId == "1000054802")
pitcherh <- filter(pitcher_data, PitcherId == "1000054994")
```

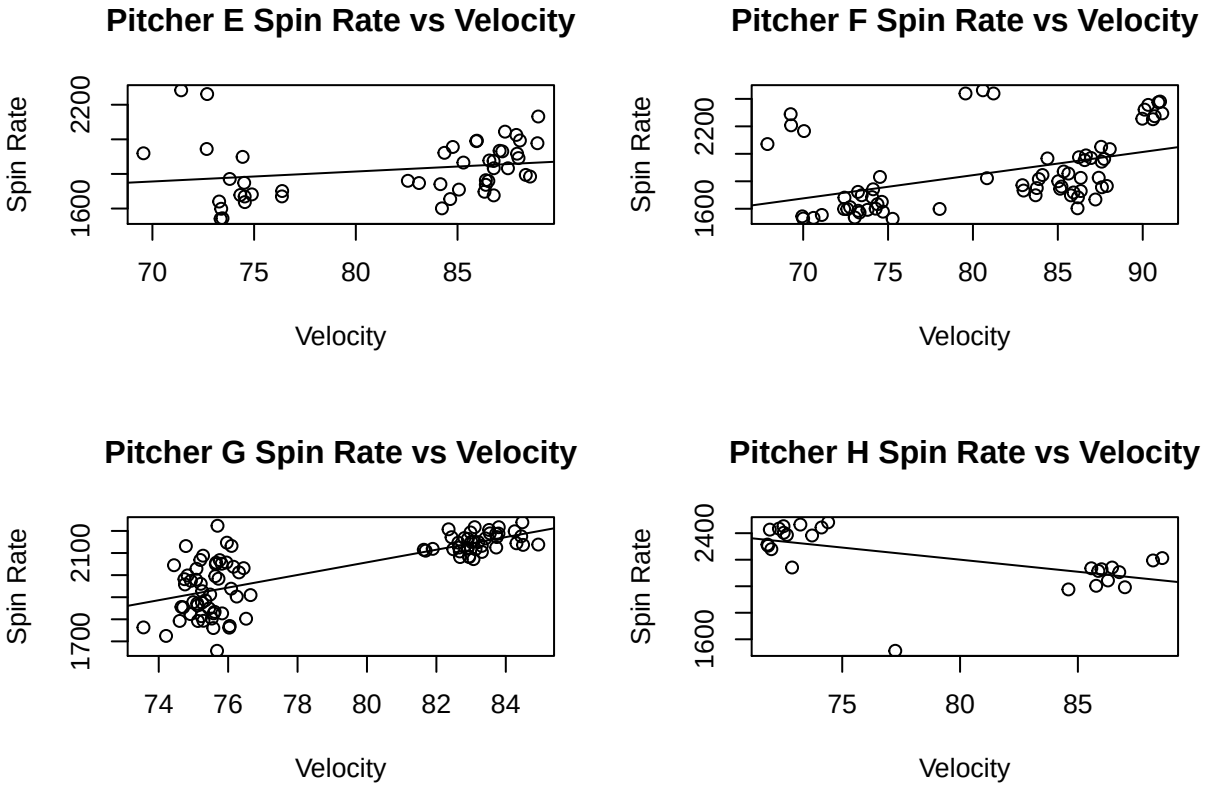
Viewing the data confirms all the group by's have been done correctly. Now, I will create a linear model for each pitcher and plot them simultaneously to view if there are stronger linear relationships across each pitcher.

```
par(mfrow = c(2,2))
plot(pitchera$RelSpeed, pitchera$SpinRate, main = "Pitcher A Spin Rate vs Velocity" ,
     xlab = "Velocity", ylab = "Spin Rate", abline(lm(pitchera$SpinRate~pitchera$RelSpeed)))
plot(pitcherb$RelSpeed, pitcherb$SpinRate, main = "Pitcher B Spin Rate vs Velocity" ,
     xlab = "Velocity", ylab = "Spin Rate", abline(lm(pitcherb$SpinRate~pitcherb$RelSpeed)))
plot(pitcherc$RelSpeed, pitcherc$SpinRate, main = "Pitcher C Spin Rate vs Velocity" ,
     xlab = "Velocity", ylab = "Spin Rate", abline(lm(pitcherc$SpinRate~pitcherc$RelSpeed)))
plot(pitcherd$RelSpeed, pitcherd$SpinRate, main = "Pitcher D Spin Rate vs Velocity" ,
     xlab = "Velocity", ylab = "Spin Rate", abline(lm(pitcherd$SpinRate~pitcherd$RelSpeed)))
```




```
plot(pitchere$RelSpeed, pitchere$SpinRate, main = "Pitcher E Spin Rate vs Velocity" ,
     xlab = "Velocity", ylab = "Spin Rate", abline(lm(pitchere$SpinRate~pitchere$RelSpeed)))
plot(pitcherf$RelSpeed, pitcherf$SpinRate, main = "Pitcher F Spin Rate vs Velocity" ,
     xlab = "Velocity", ylab = "Spin Rate", abline(lm(pitcherf$SpinRate~pitcherf$RelSpeed)))
```

```
plot(pitcherg$RelSpeed, pitcherg$SpinRate, main = "Pitcher G Spin Rate vs Velocity" ,
     xlab = "Velocity", ylab = "Spin Rate", abline(lm(pitcherg$SpinRate~pitcherg$RelSpeed)))
plot(pitcherh$RelSpeed, pitcherh$SpinRate, main = "Pitcher H Spin Rate vs Velocity" ,
     xlab = "Velocity", ylab = "Spin Rate", abline(lm(pitcherh$SpinRate~pitcherh$RelSpeed)))
```



The plots above illustrate how the pitchers have a special selection of pitches. The groupings of pitches suggest many of the pitchers are consistent, throwing with very low variability. This is shown perfectly with pitcher G, who has two tight groupings of pitches in the 75-76 mph and 82-84 mph ranges. While the pitchers may not have linear relationships between Velocity and the spin rate of their pitches, it appears some pitchers have inverse relationships and others have direct relationships with Velocity on Spin rate. As Pitchers A,D, and H throw faster pitches, their spin rates decrease. Meanwhile, pitchers B,C,E, and G have higher spin rates as they throw faster pitches.

To determine a linear relationship with spin rate, I would recommend exploration into different predictors which may show a linear relationship with Spin Rate.