Kmeans Clustering of Jacob deGrom Pitches

Walker Reinfeld

6/8/2019

Executive Summary

As an extension of my 301-1 EDA on Jacob deGrom I was looking for more information about deGrom as a pitcher and maybe why he was more successful in 2018 than he was in 2017. Doing some EDA indicated that there could be some differences between deGrom's pitching with balls in-play vs not in play, strikeouts vs not-strikeouts, and in between years. Kmeans clustering confirmed domain knowledge that deGrom has 5 different pitches, especially that he has two different types of fastballs. Performing Kmeans clustering on the situations listed above yielded results that confirmed mechanical differences from 2017 to 2018. But the clustering did not yield too many results in terms of situational differences.

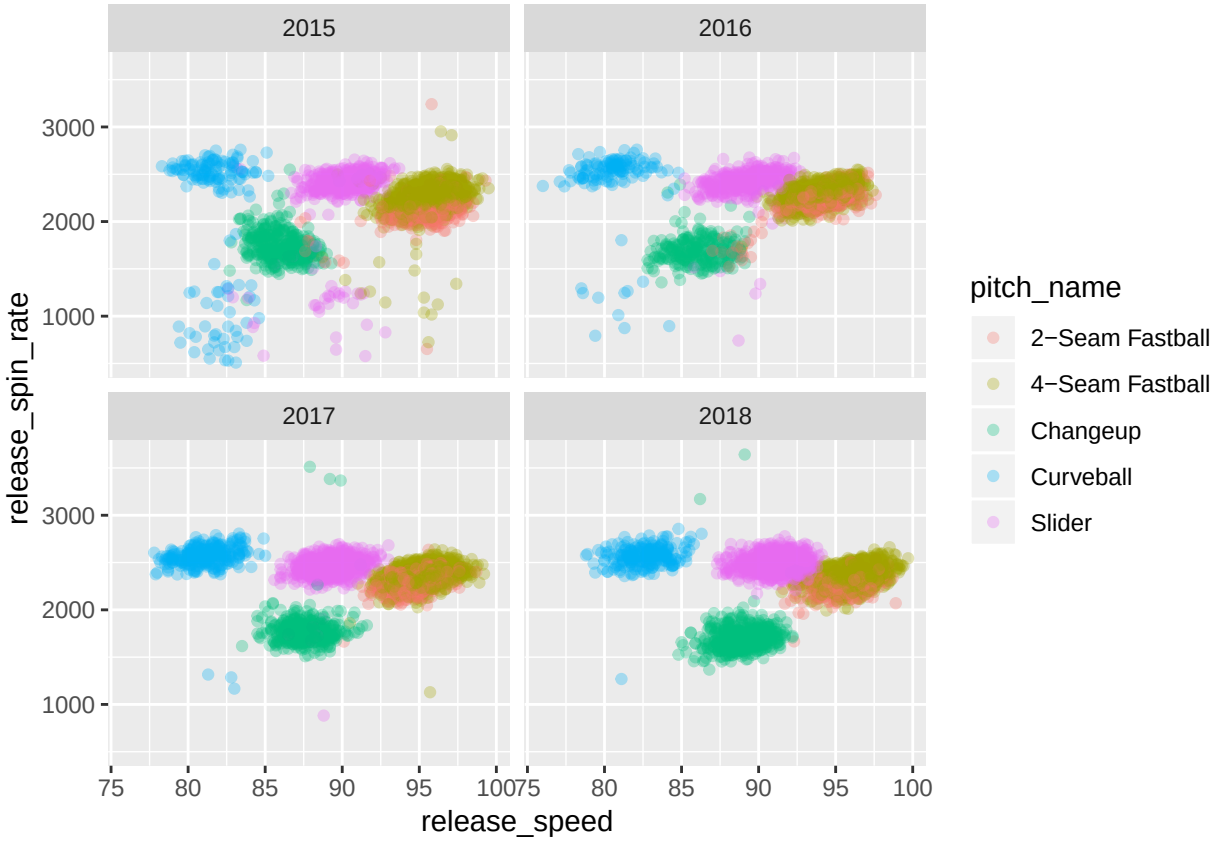
Introduction

As a New York Mets fan I was interested in what made Jacob deGroms 2018 season so much better than his previous ones. During 301-1, I investigated pitch-by-pitch data to try and pinpoint what went right for him. I came to some conclusions regarding his release point becoming more consistent and some of his pitches having a higher spin rate than in previous years. But this analysis I conducted in the fall also raised more questions about deGrom's ability to decieve hitters. Of course with a more conisistent release point it will be harder for hitters to tell what pitch is coming, but I wanted to use clustering techniques to see how similar he was decieving hitters. Kmeans clustering was used in this project to find a number of interesting findings regarding Jacob deGrom and his pitching. The clustering was done for many different scenario's to compare: All pitches in the 2018 season vs the 2017 season, Balls hit in play vs balls not in play, and strikeout pitches vs non-strikeout pitches.

EDA Findings

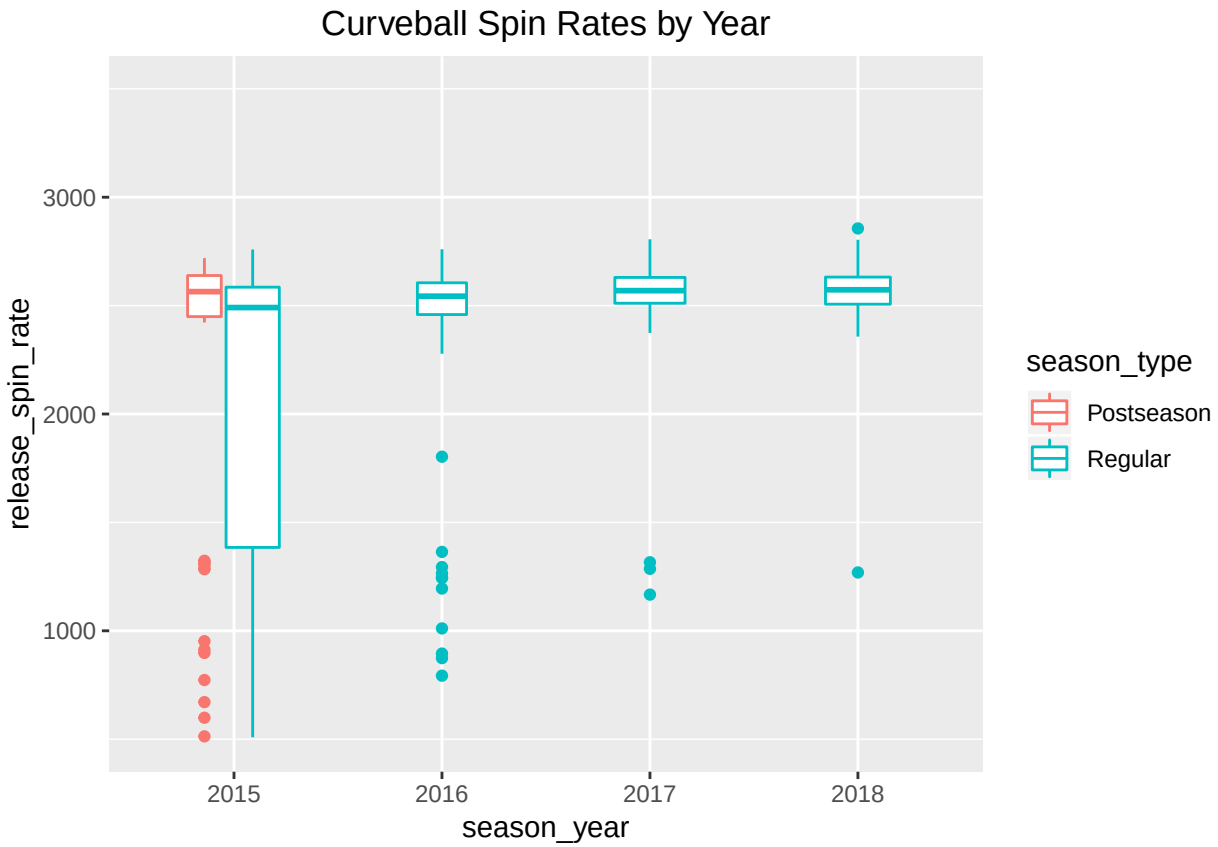
During the EDA for this project there were some helpful findings that helped shaped the idea for this project. The first of which was how deGrom's Spin Rates and speeds progressed over time.

Spin Rate vs Speed



This graphic could be a little hard to interpret, but the imortant thing to note is that in 2017 and 2018 you can see the clusters move right along the x-axis compared to the 2015 and 2016 seasons, which means deGrom's velocity increased over those years. Another important thing to note is that in the 2015 plot you can see a lot of points with decreased spin rates. I beleive this was due to the fact that it was the first year StatCast was being used and they had a lot of trouble reading curveballs as shown below.

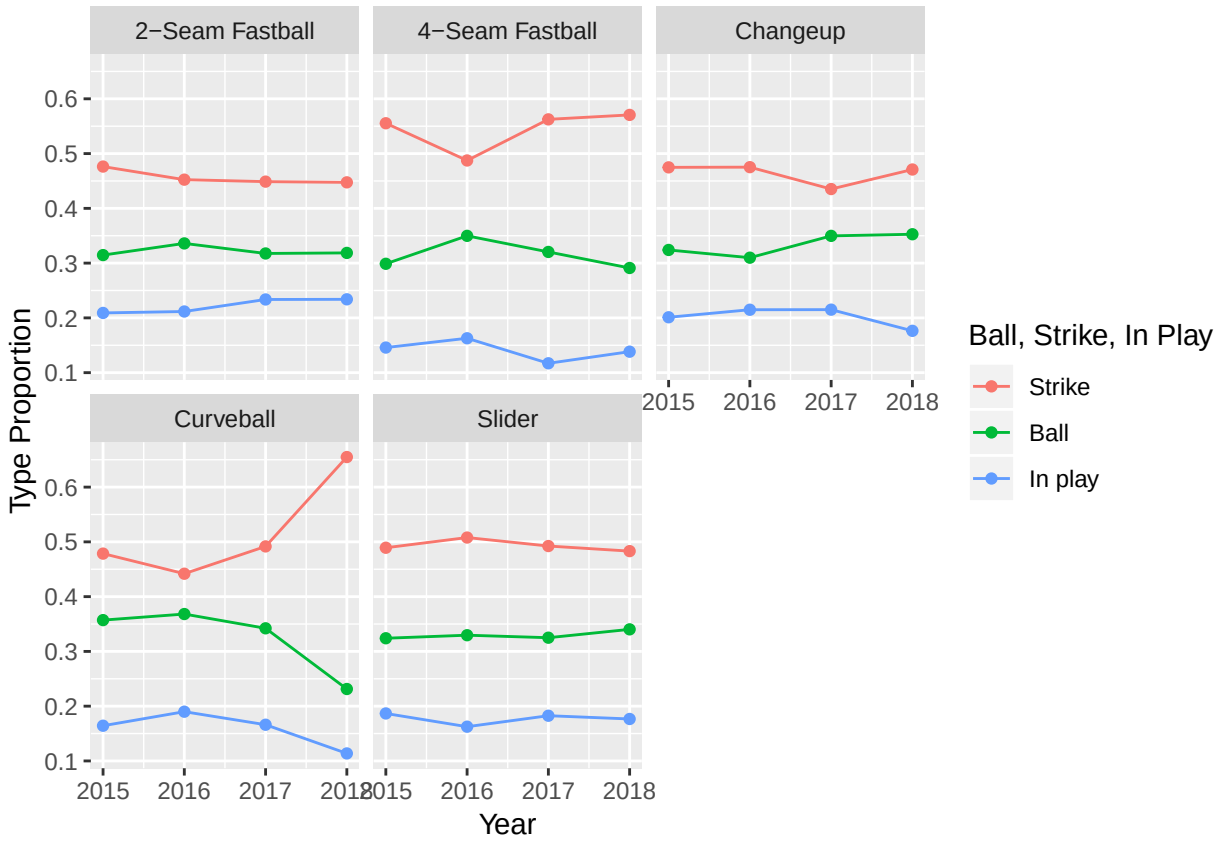
Curveball Spin Rates



The length of the boxplots in 2015 indicate that there was a lot of variability in the recorded spin rates. This will have an effect on clusters that use 2015 data that will be highlighted later.

Balls in-play vs Not in-play

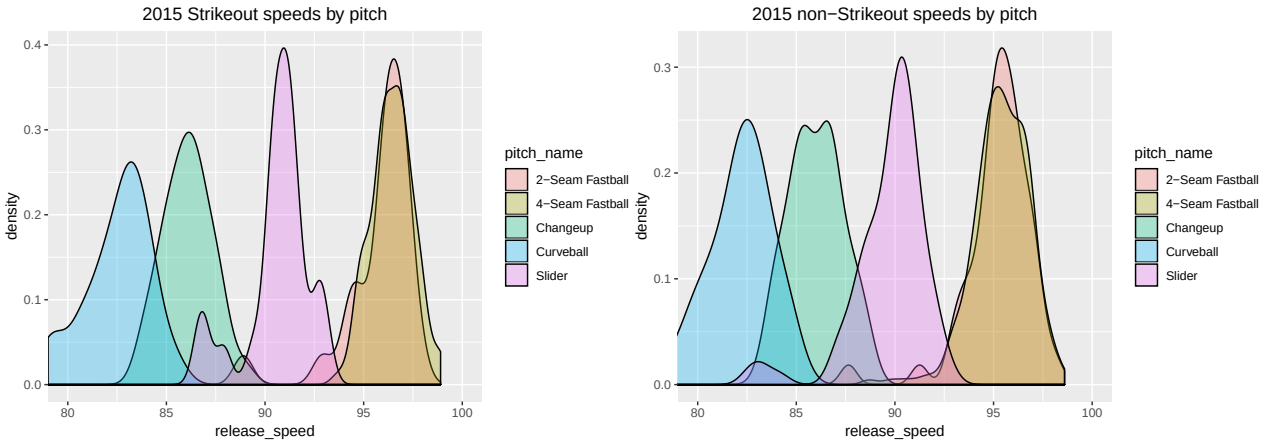
Another interesting finding from the EDA was how deGrom's different pitches were hit in play or thrown for balls and strikes over the years. The graph below highlights these findings.

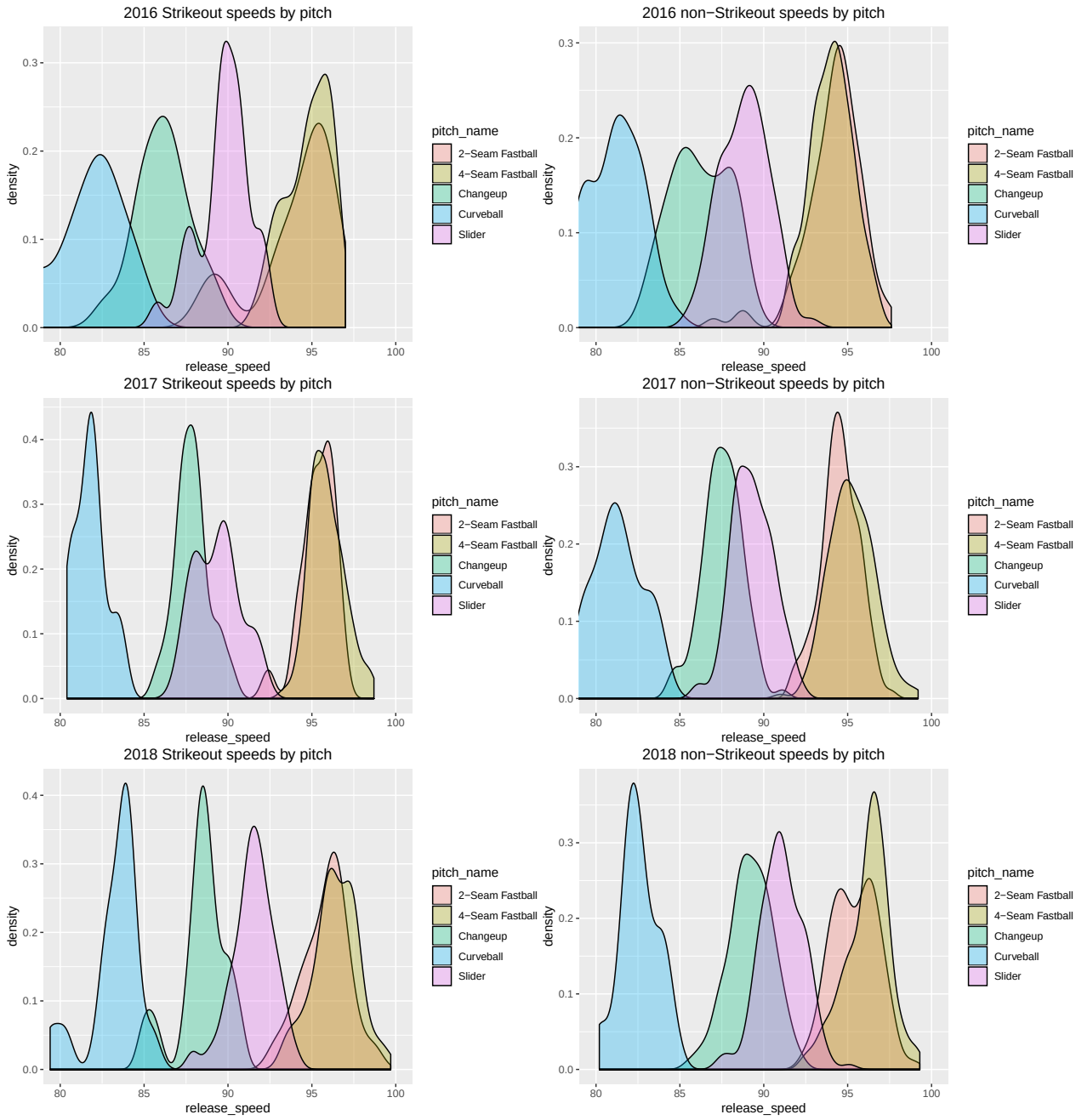


The most drastic finding in this graph is deGrom's increased ability to throw strikes with his curveball. He threw it less than 10% of the time but was much more effective when he did throw it. This graph shows there may be some validity in clustering on balls that were hit-in play vs not in play.

Strikeout speed vs non-strikeout Speed

Another thing I looked at in the EDA was the speed of his pitches on strikeouts vs non-strikeout pitches.

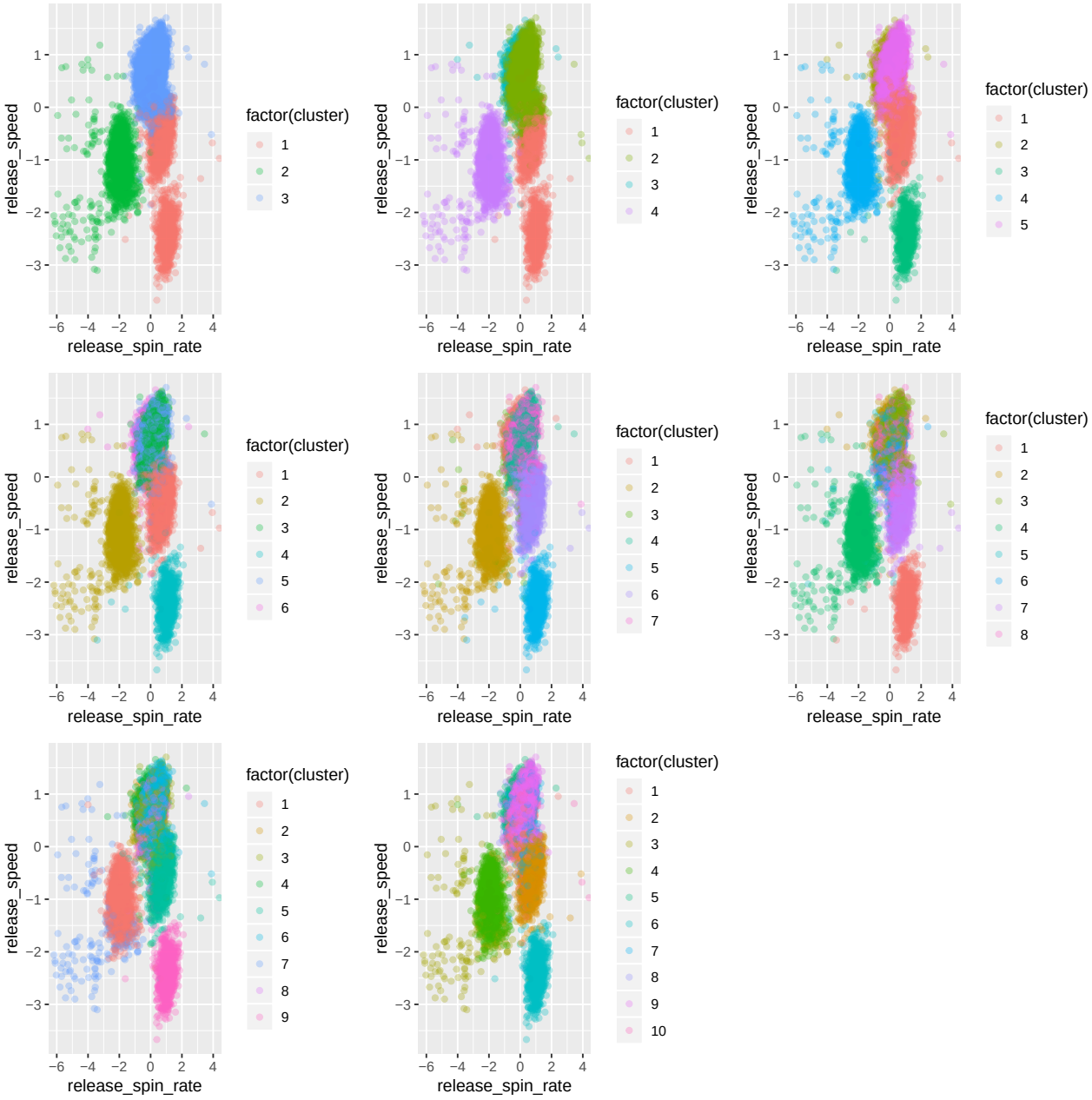




The interesting things to note about these graphs is that it appears deGrom is able to vary his pitch speeds for strikeouts as seen by the multimodal nature of some of the density plots. Compare this to his non-strikeout pitches and the density plots seem to mostly be unimodal.

Clustering

Full data



The most intersting thing to note about this clustering on the full data is that the kmeans method with 5 clusters is able to pick up on the fact that deGrom's four-seam and two-seam fastballs are different pitches. This is highlighted by the fact that there are two different colors in the top right portion of the graph. This is argument enough for using 5 cluster's throughout the remained of this project, as domain knowledge suggests he has 5 pitches and this is backed up by the kmeans clusters. The large cluster of points to the left of the plot with low spin rates are pitches mostly thrown in 2015, when Statcast was brand new and struggled to pick up the spin rate of a curveball. Kmeans has placed this curveballs with the changeup cluster when using 2015 data. For this reason, 2015 data will not excplicitly be looked at in this project.

Looking at the attributes of the 5 clusters from full data.

```
## # A tibble: 5 x 5
##   cluster mean_spin mean_release_spe~ mean_effective_sp~ mean_release_exte~
##     <int>     <dbl>             <dbl>              <dbl>              <dbl>
## 1       1     2441.              90.5               91.4               6.77
## 2       2     2262.              94.8               95.1               6.78
## 3       3     2554.              81.7               82.1               6.46
## 4       4     1677.              87.4               88.7               7.12
## 5       5     2332.              95.3               96.6               7.04
```

We see that the pitches are split up into pretty distinct categories. You have your fastballs with the middle spin rates and highest speeds, the slider with middle speed and higher spin rates, changeup with lowest spin and middle spin, and curveball with highest spin rate and lowest speed. You also may notice that using the data included in 2015 has most likely included any pitch with the really low spin rate in the changeup group as indicated by the slower and faster pitches colored blue.

As we noted the clustering technique was able to pick up on the differences between the two-seam and four-seam fastball. It obviously appears that those differences are not coming from the spin rates or release speeds. So I plucked the centroids and did a subtraction to see what variables were the basis for the difference.

```
##      release_speed  release_extension      release_pos_x
##        0.099400707        0.885453298        0.311305695
##      release_pos_z  release_spin_rate    effective_speed
##        1.894156698        0.243874307        0.316850951
##             type_B             type_S             type_X
##        0.002233346        0.005841826        0.003608480
##            stand_L            stand_R     game_year_2015
##        0.052263749        0.052263749        0.892794084
##     game_year_2016     game_year_2017     game_year_2018
##        0.162378936        0.361899136        0.368516013
##            balls_0            balls_1            balls_2
##        0.002372389        0.001960581        0.002039084
##            balls_3          strikes_0          strikes_1
##        0.002293886        0.005366150        0.003605154
##          strikes_2           inning_1           inning_2
##        0.008971304        0.020009633        0.018867347
##           inning_3           inning_4           inning_5
##        0.006751929        0.017704437        0.016996579
##           inning_6           inning_7           inning_8
##        0.018408969        0.001500207        0.004696213
##           inning_9        runner_on_0        runner_on_1
##        0.004285070        0.060040210        0.060040210
## scoring_position_0 scoring_position_1
##        0.030477818        0.030477818
```

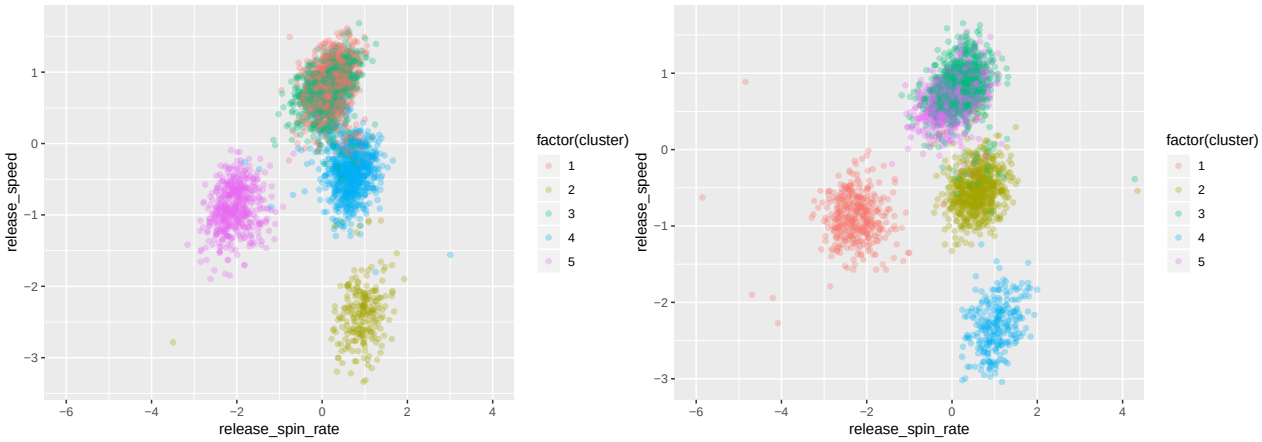
So the variables were scaled, so a tenth of a difference is a pretty big effect, but you can see that the biggest difference in coming from release positions and his release extension. There is also some difference in his spin rates and in future clusters for this project we classified the cluster with the higher of the two spin rates to always be the fourseam fastball for consistency.

Clustering 2017 and 2018 Pitches

```
## # A tibble: 5 x 5
##   cluster mean_spin mean_release_spe~ mean_effective_sp~ mean_release_exte~
##     <int>     <dbl>             <dbl>              <dbl>              <dbl>
```

```
## 1       1     2358.              95.9              97.2              6.92
## 2       2     2563.              82.8              83.2              6.42
## 3       3     2346.              95.8              97.0              6.91
## 4       4     2487.              91.1              92.0              6.65
## 5       5     1693.              89.0              90.4              7.05
```

```
## # A tibble: 5 x 5
##   cluster mean_spin mean_release_spe~ mean_effective_sp~ mean_release_exte~
##     <int>     <dbl>             <dbl>              <dbl>              <dbl>
## 1       1     1760.              87.6               89.4               7.28
## 2       2     2461.              89.3               90.5               6.82
## 3       3     2362.              95.3               96.7               7.10
## 4       4     2570.              81.1               81.9               6.57
## 5       5     2330.              94.7               96.1               7.06
```



The graphs of the clusters alone show something quite interesting about deGrom's pitching in 2018 vs 2017. The 2018 clusters on the left and you see the slider cluster(blue) and fastball clusters (green and red) overlapping a lot more with each other than the 2017 versions (slider: yellow, fastballs: purple and green). This shows that Degrom was able to blend his pitches more in 2018 which made him more successful.

Next I wanted to look closer at differences in those specifics pitches.

Fourseam year difference

```
##      release_speed  release_extension      release_pos_x
##       0.0243434878       0.0025978470       1.7109489371
##      release_pos_z  release_spin_rate    effective_speed
##       0.1731600762       0.0402353052       0.0277140418
##             type_B             type_S             type_X
##       0.0168119569       0.0235714034       0.0067594466
##            stand_L            stand_R            balls_0
##       0.2754474477       0.2754474477       0.1738851382
##            balls_1            balls_2            balls_3
##       0.0856399674       0.0605208782       0.0277242926
##          strikes_0          strikes_1          strikes_2
##       0.1906551193       0.0217813417       0.2124364609
##           inning_1           inning_2           inning_3
##       0.1421948161       0.0160062937       0.0008219119
##           inning_4           inning_5           inning_6
##       0.0619561773       0.0329374074       0.0364796174
##           inning_7           inning_8           inning_9
```

```
##       0.0069923442       0.0098466938       0.0031522419
##        runner_on_0        runner_on_1 scoring_position_0
##       0.3962712011       0.3962712011       0.2290696824
## scoring_position_1
##       0.2290696824
```

The biggest difference on his 4-seam fastball was his release position in the x direction.

Two seam year difference

```
##      release_speed  release_extension      release_pos_x
##        0.064301108        0.071254814        1.353088223
##      release_pos_z  release_spin_rate    effective_speed
##        0.026089137        0.130735715        0.072735705
##             type_B             type_S             type_X
##        0.025005416        0.007696355        0.017309061
##            stand_L            stand_R            balls_0
##        0.205802701        0.205802701        0.147363018
##            balls_1            balls_2            balls_3
##        0.076659204        0.050235224        0.020468590
##          strikes_0          strikes_1          strikes_2
##        0.264399715        0.025472253        0.238927462
##           inning_1           inning_2           inning_3
##        0.134807436        0.025294288        0.018477442
##           inning_4           inning_5           inning_6
##        0.003131158        0.046791982        0.069806250
##           inning_7           inning_8           inning_9
##        0.054818475        0.012228023        0.008196721
##        runner_on_0        runner_on_1 scoring_position_0
##        0.352237205        0.352237205        0.229105840
## scoring_position_1
##        0.229105840
```

Similar to the 4-seam fastball, deGrom altered his release position in the x-direction as well.

Changeup year difference

```
##      release_speed  release_extension      release_pos_x
##        1.661480754        0.525748738        1.397221867
##      release_pos_z  release_spin_rate    effective_speed
##        0.576081016        2.272017458        1.549815269
##             type_B             type_S             type_X
##        0.058867218        0.092501790        0.033634571
##            stand_L            stand_R            balls_0
##        0.009528207        0.009528207        0.171962151
##            balls_1            balls_2            balls_3
##        0.142478458        0.037245517        0.007761824
##          strikes_0          strikes_1          strikes_2
##        0.179887665        0.108196457        0.071691208
##           inning_1           inning_2           inning_3
##        0.206400512        0.089013000        0.047917109
##           inning_4           inning_5           inning_6
##        0.119505706        0.044770358        0.108506734
##           inning_7           inning_8           inning_9
##        0.056951804        0.019129708        0.005533690
```

```
##        runner_on_0        runner_on_1 scoring_position_0
##        0.182125572        0.182125572        0.112117681
## scoring_position_1
##        0.112117681
```

For his changeup degrom altered his speed and spin rate the most between the two seasons. From the EDA plot we can conclude that he was throwing his changeup faster with a slower spin rate.

Slider year difference

```
##      release_speed  release_extension      release_pos_x
##       0.1335517706       0.0988054048       0.2689657638
##      release_pos_z  release_spin_rate    effective_speed
##       0.2335337861       0.0763849065       0.1117872805
##             type_B             type_S             type_X
##       0.0123215784       0.0048213131       0.0075002653
##            stand_L            stand_R            balls_0
##       0.1201764344       0.1201764344       0.0009630613
##            balls_1            balls_2            balls_3
##       0.0246071170       0.0196396429       0.0040044129
##          strikes_0          strikes_1          strikes_2
##       0.0083772317       0.0182741381       0.0266513698
##           inning_1           inning_2           inning_3
##       0.0441726783       0.0335830085       0.0016918644
##           inning_4           inning_5           inning_6
##       0.0008989907       0.0200540996       0.0490140134
##           inning_7           inning_8           inning_9
##       0.0214896816       0.0325458655       0.0044469005
##        runner_on_0        runner_on_1 scoring_position_0
##       0.0627010966       0.0627010966       0.0342617565
## scoring_position_1
##       0.0342617565
```

DeGrom's slider was pretty similar between the two seasons with some variation in his vertical release position.

Curveball year difference

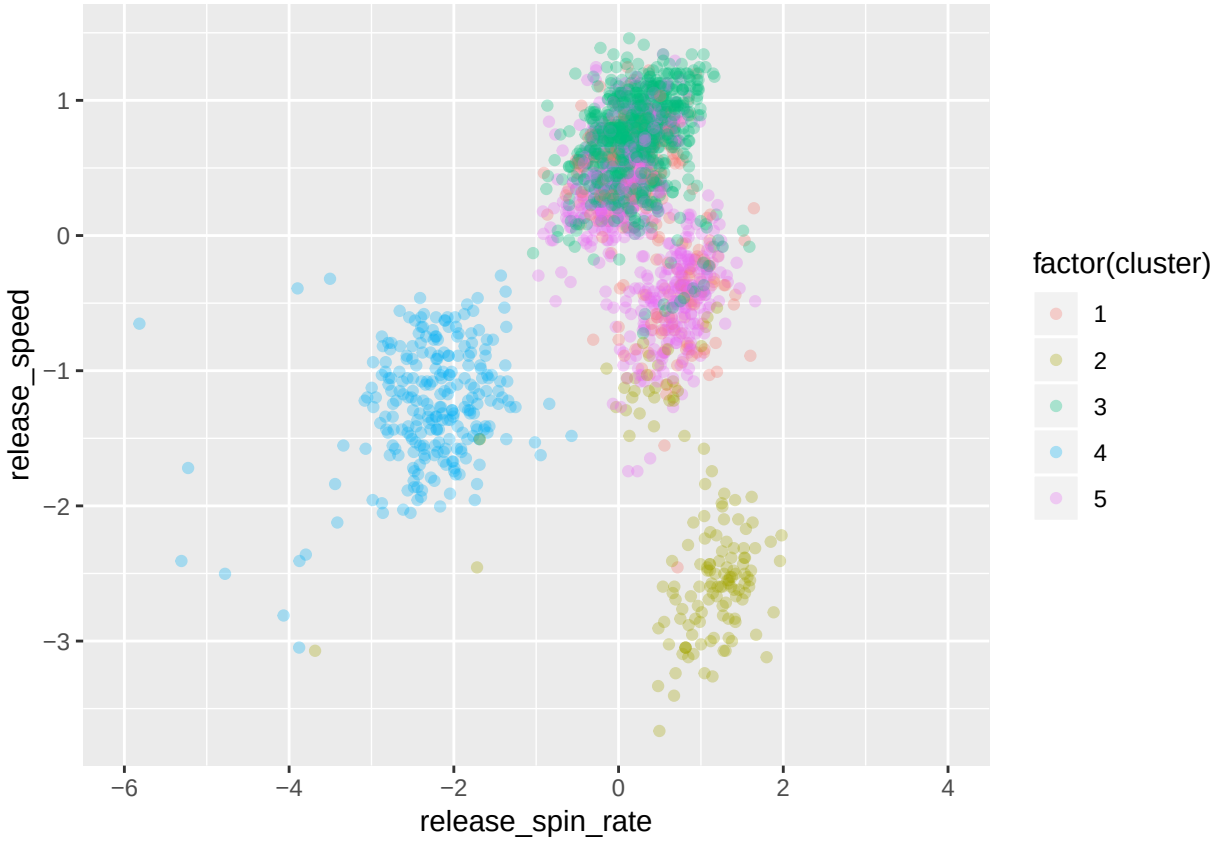
```
##      release_speed  release_extension      release_pos_x
##       0.0695341835       0.1228311693       0.0756832272
##      release_pos_z  release_spin_rate    effective_speed
##       0.5364043696       0.1108113090       0.0728508590
##             type_B             type_S             type_X
##       0.1039604795       0.1608133399       0.0568528603
##            stand_L            stand_R            balls_0
##       0.0368902902       0.0368902902       0.0616580398
##            balls_1            balls_2            balls_3
##       0.0023098582       0.0673905346       0.0080423530
##          strikes_0          strikes_1          strikes_2
##       0.1092377468       0.1306671612       0.0214294144
##           inning_1           inning_2           inning_3
##       0.0739997640       0.0597022475       0.0148539057
##           inning_4           inning_5           inning_6
##       0.0214125542       0.0704253848       0.0806764344
##           inning_7           inning_8           inning_9
##       0.0125440475       0.0011970798       0.0005058084
```

```
##        runner_on_0        runner_on_1 scoring_position_0
##       0.0078737502       0.0078737502       0.0276171368
## scoring_position_1
##       0.0276171368
```

Lastly, similar to his slider deGrom did not have much difference in his curveball besides maybe his vertical release position and release extension.

2016 Clusters

In 2016, deGrom's clusters overlapped more often so it was hard to classify pitches in order to truly compare them to other year's measurements. So I present the clusters below, but do not perform further analysis.



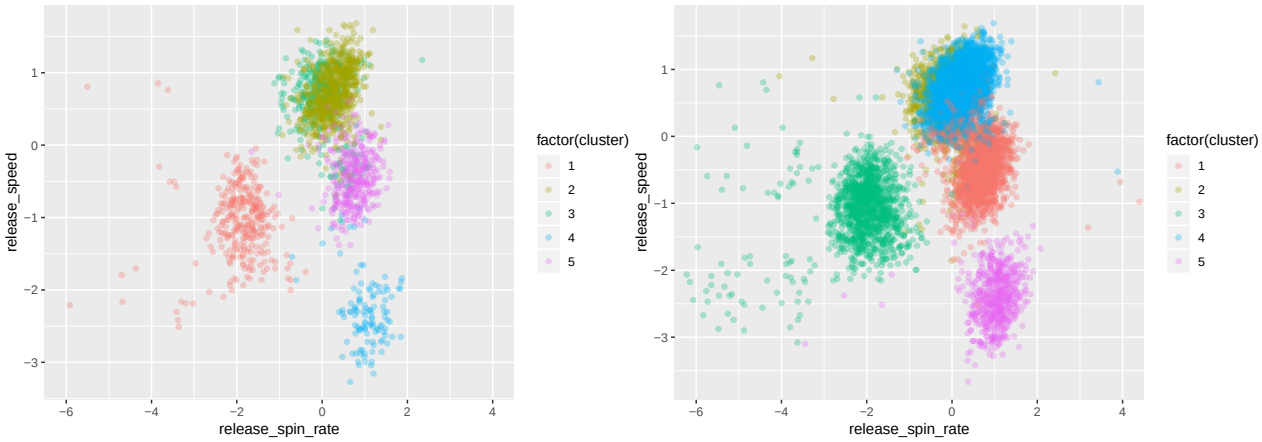

```
## # A tibble: 5 x 5
##   cluster mean_spin mean_release_spe~ mean_effective_sp~ mean_release_exte~
##     <int>     <dbl>             <dbl>              <dbl>              <dbl>
## 1       1     2338.              92.3               93.2               6.82
## 2       2     2499.              81.8               82.2               6.50
## 3       3     2307.              94.4               95.5               7.16
## 4       4     1661.              86.2               87.4               7.27
## 5       5     2316.              92.1               93.0               6.97
```

Balls in play vs Not in Play

```
## # A tibble: 5 x 5
##   cluster mean_spin mean_release_spe~ mean_effective_sp~ mean_release_exte~
##     <int>     <dbl>             <dbl>              <dbl>              <dbl>
```

```
## 1       1     1685.              87.4               88.8               7.13
## 2       2     2311.              95.1               96.4               7.04
## 3       3     2248.              94.6               94.9               6.77
## 4       4     2535               82.0               82.5               6.47
## 5       5     2449.              90.1               91.1               6.74
```

```
## # A tibble: 5 x 5
##   cluster mean_spin mean_release_spe~ mean_effective_sp~ mean_release_exte~
##     <int>     <dbl>             <dbl>              <dbl>              <dbl>
## 1       1     2439.              90.5               91.5               6.77
## 2       2     2266.              94.9               95.2               6.78
## 3       3     1675.              87.4               88.7               7.12
## 4       4     2336.              95.3               96.6               7.04
## 5       5     2556.              81.6               82.0               6.46
```



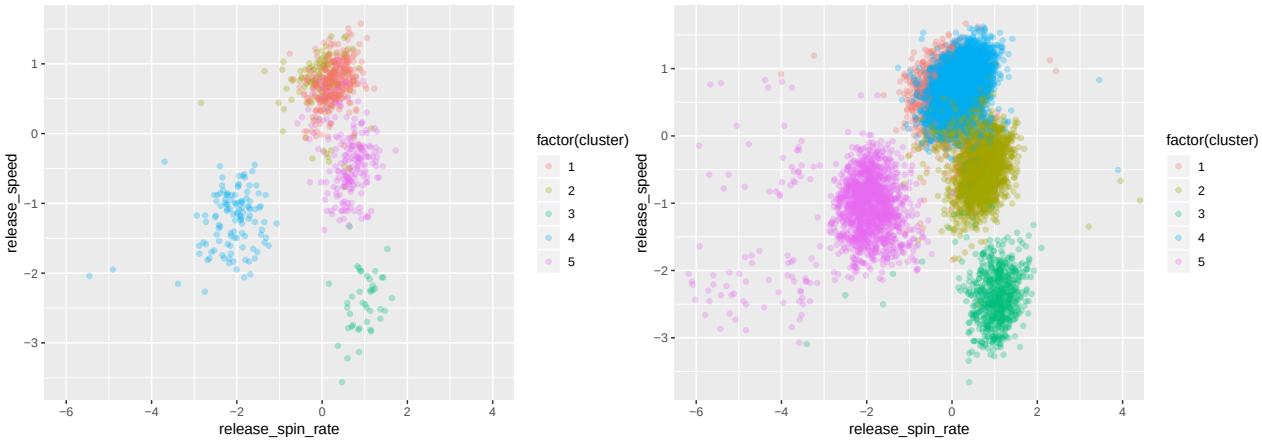
The graphs alone on this clustering does not provide much insight on the differences in his pitches on balls hit in play vs ones that were not. The balls in play cluster is on the left and not in play is on the right.

Performing the centroid subtraction did not yield any large enough differences to warrant any analysis. This suggests that deGrom seems to get unlucky when his pitches are hit in play, because his is not throwing his pitches any worse. I did not include the centroid subtraction to save space on this report.

Strikeout vs Non-Strikeouts

```
## # A tibble: 5 x 5
##   cluster mean_spin mean_release_spe~ mean_effective_sp~ mean_release_exte~
##     <int>     <dbl>             <dbl>              <dbl>              <dbl>
## 1       1     2364.              95.9               97.2               7.02
## 2       2     2292.              95.8               96.1               6.79
## 3       3     2564.              82.1               82.3               6.44
## 4       4     1688.              87.3               88.6               7.14
## 5       5     2464.              91.2               92.3               6.76
```

```
## # A tibble: 5 x 5
##   cluster mean_spin mean_release_spe~ mean_effective_sp~ mean_release_exte~
##     <int>     <dbl>             <dbl>              <dbl>              <dbl>
## 1       1     2260.              94.8               95.1               6.78
## 2       2     2439.              90.4               91.4               6.76
## 3       3     2553.              81.7               82.1               6.46
## 4       4     2329.              95.2               96.5               7.04
## 5       5     1677.              87.4               88.7               7.12
```



Again, it is tough to see any differences between the strikeout pitches clustering on the left against the non-strikeout pitches on the right.

Similar to the clusters when we looked at balls in-play than not in-play. The strikeout vs non-strikeouts did not yield many differences either.

Clustering Results

The fact that we only find large differences from 2017 to 2018 instead of depending on situations suggest that deGrom has done a good job of altering his mechanics to improve in 2018, but does not change what he is doing on the situation. This seems to support my first quarter project when I concluded he made a number of changes between before the 2018 season that allowed him to become an elite pitcher in the Major Leagues.