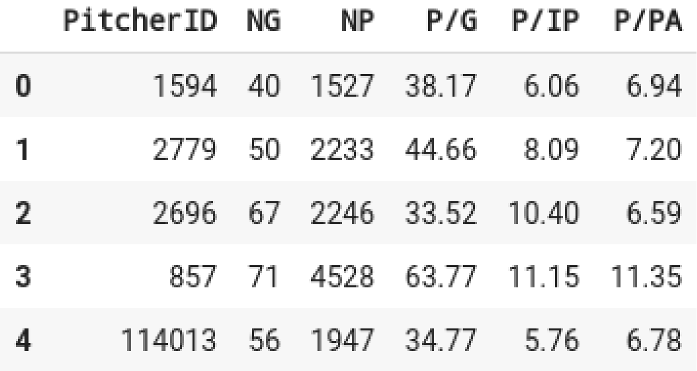
Toronto Blue Jays Task - Hadi Sotudeh

I chose the first question to address here as part of the hiring process.

Pitch Command

1. Which pitcher had the best fastball command? How did you determine this?

Answer: First, I calculate the basic performance statistics[1] of the pitchers, see Table 1.



	PitcherID	NG	NP	P/G	P/IP	P/PA
0	1594	40	1527	38.17	6.06	6.94
1	2779	50	2233	44.66	8.09	7.20
2	2696	67	2246	33.52	10.40	6.59
3	857	71	4528	63.77	11.15	11.35
4	114013	56	1947	34.77	5.76	6.78

Table 1. Basic performance statistics of pitchers

Table 1 shows that pitcher 857 had the highest number of pitches per game.

[1] http://m.mlb.com/glossary/standard-stats, accessed on Jan 19, 2021

Then, I create the pitch velocity boxplot per pitcher, as shown in Fig 1.

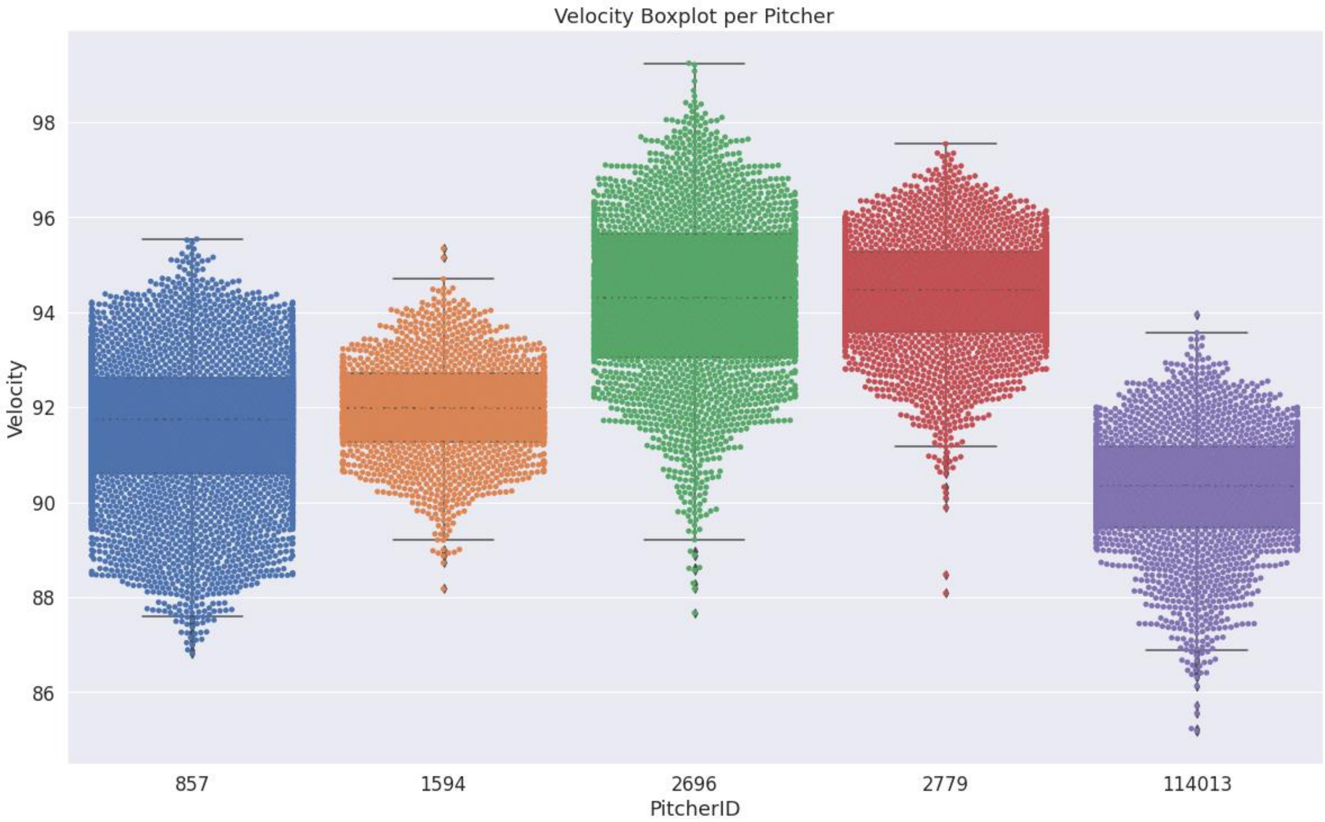


Figure 1. Velocity Boxplot per Pitcher

Defining which pitcher has the *best* fastball command depends on what we mean exactly. Here, I define it as a pitcher with the **highest** and most **consistent** pitch velocities over games. Having a higher velocity makes the batter’s job more difficult and also fewer variations mean the pitcher is more tireless in his performance.

Fig 1 shows that pitchers 2696 and 2779 have the highest pitch velocities on average among the other pitchers, but pitcher 2779 deviates less than pitcher 2696 and is more consistent in his performance. Therefore, I chose pitcher 2779 as the pitcher with the *best* fastball command.

I can also explore the dataset from other dimensions. For example, I create a scatterplot to visualize the horizontal and vertical break and their relation with velocity and pitch type, as shown in Fig 2.

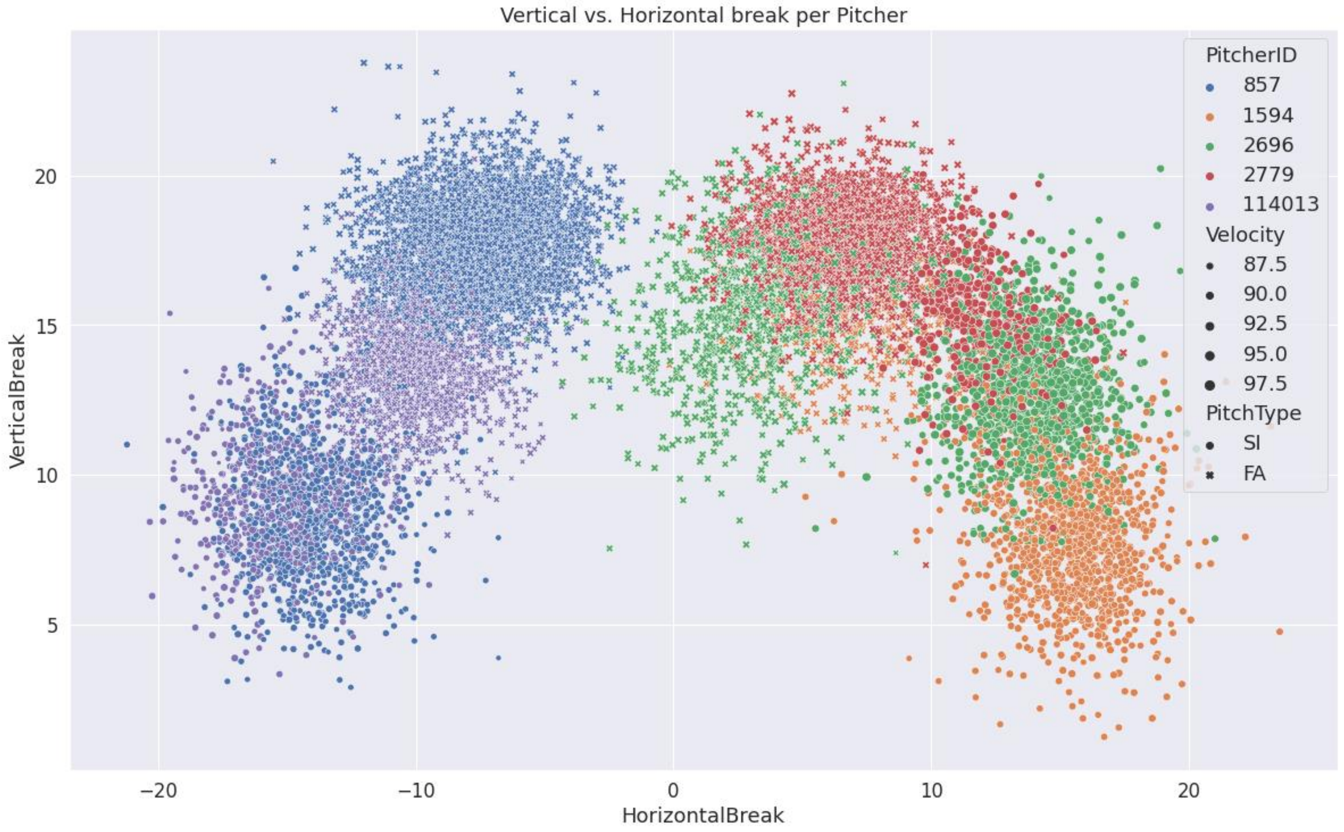


Figure 2. Vertical vs. Horizontal break per Pitcher

In Fig 2, I can clearly see how the pitchers are separated from each other based on their pitch attributes. For instance, the pitch type (FA or SI) separates the pitches pitcher 2696 had only in their horizontal break and not vertically. Although what I did here seems pretty simple, the key point in defining Key Performance Indicators is to keep them simple and meaningful for the coaching staff.

2. If you wanted to create a fastball command metric that could be applied to any pitcher at any level, how might you go about doing so?

Answer:

I will find a large database of all pitchers over a season and take a supervised learning approach by creating a label that determines whether a pitch outcome is successful or not. Then, I can train a classification machine learning model such as Logistic Regression, Random Forest, XGBoost, and Neural Networks that gets pitch features such as velocity, horizontal/vertical breaks, pitch hand and type, and batside as input to calculate the outcome probability.

This probability can act as *Expected Outcomes* in baseball such as what we have as *Expected Goals* in Soccer. Therefore, I can subtract the actual cumulative outcomes from the cumulative expected outcomes over matches to calculate pitch outperformance per pitcher. The pitcher who outperforms his expectation more than others is the one who is more clinical (*better*).
Last but not least, I can also look at the feature importance of the trained model if it is a tree-based model or draw its partial dependence plots[2] regardless of the model type to study which features and to what extent have an impact on the outcome probability.

[2] https://christophm.github.io/interpretable-ml-book/pdp.html, accessed on Jan 19, 2021