CS287 – Data Science I

Eamon Deffner, Frederick R. “Randy” Carlson  
Subject: Data Science Check-in #5

Working Title: “Use of Data to analyze Base Running Statistics.”  
*Abstract: How much does base running speed affect offensive statistics? This paper mines data from BaseballSavant on sprint speed and batting averages to answer this question. We use the data science capabilities of anaconda python to answer these questions. We found that there is little to no advantage to single base hitting due to speed, but we did find that there is an advantage to speed in slugging. This means that training for longer sprints (from home to 2b) may yield better offensive results than training for the shorter sprint to first. We also found that players who have a better batting average above expected as well as a higher slugging above expected tend to be involved in these plays more often. Another result is that players who have more ground balls tend to have a higher batting average on these ground balls.*   
  
This document serves as a progress report on our project. Our team wishes to provide a condensed overview of this project, its progress to date, and a timeline.

2) Our goal for this week was to analyze the feature importance of our past models, as well as cluster players, first using PCA to find if sprint speed, ground ball percentage, and percent of at bats that result in hits in play. The results were visualized using seaborn plots, showing the clusters. In addition, we created a GitHub file exchange for our results, and created a template for our final presentation and project. The GitHub repository, which has our data, project drafts and templates for the paper and the presentation can be found at:  
<https://github.com/T4rnover/DataScienceProject_Spring>**3) Challenges and Questions**: One factor that we may be missing in our model is the role that stolen bases play. Faster players can steal more bases, giving them the extra base more often on plays where it is not calculated as a part of their slugging percentage. This would be an added stretch goal, as our original focus was to study the importance of speed for getting on base. We are currently looking at the feasibility of adding this question.

**4) Timeline**   
Week 1: Extract data from baseball savant, create a data frame, and do some exploratory data analysis (EDA). - **Complete**  
Week 2: Identify types of data (numeric, categorical, ordinal), determine independent and dependent variables, and create linear regression models; confirm the linearity of relationships. **- Complete**   
Week 3: Find players with the highest offensive profile addition from speed. Visualize this data using tables and graphs to show if speed impacts OPS. - **Complete**  
Week 4: create a model to determine how much of a role *speed*, *batting average above expected*, *slugging above average*, and *percent of at-bats that result in ground balls and hits* determine a player’s OPS. Analyze which features are most important to this model and prune these features in order to not overfit the model. –**Complete**  
Week 5: Use clustering to break players into groups, and then see if speed matters differently for each group of players and how this affects their overall offensive contributions. We plan on using both k-means clustering for this. - **Complete**  
Week 6: Analyze the results of our visualizations, using validation of our results. Begin on creating our final presentation slides, using the results found in previous weeks. - **Complete**

**5) Open Challenges and Questions:** We will still need to incorporate techniques like Bootstrapping and Non-Parametric analysis into our overall paper. There are also opportunities to look at this effect of speed to new rules changes such as the shift rule. We didn’t have the time or resources to investigate bootstrapping, but we did do a PCA analysis in the paper. - **Complete – Had to descope Bootstrapping.**

**6) Revised week-by-week timeline:** We completed the Feature Importance and Non-Parametric portions of the research. We anticipate being able to turn in the paper slightly ahead of our original schedule and have coordinated a 2 May 2023 presentation and turn it date.

**7) Data visualization work to date is shown below**. We are currently using defaults in pandas to degenerate these graphics. We will look at using seaborn and matplotlib to enhance these visualizations.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

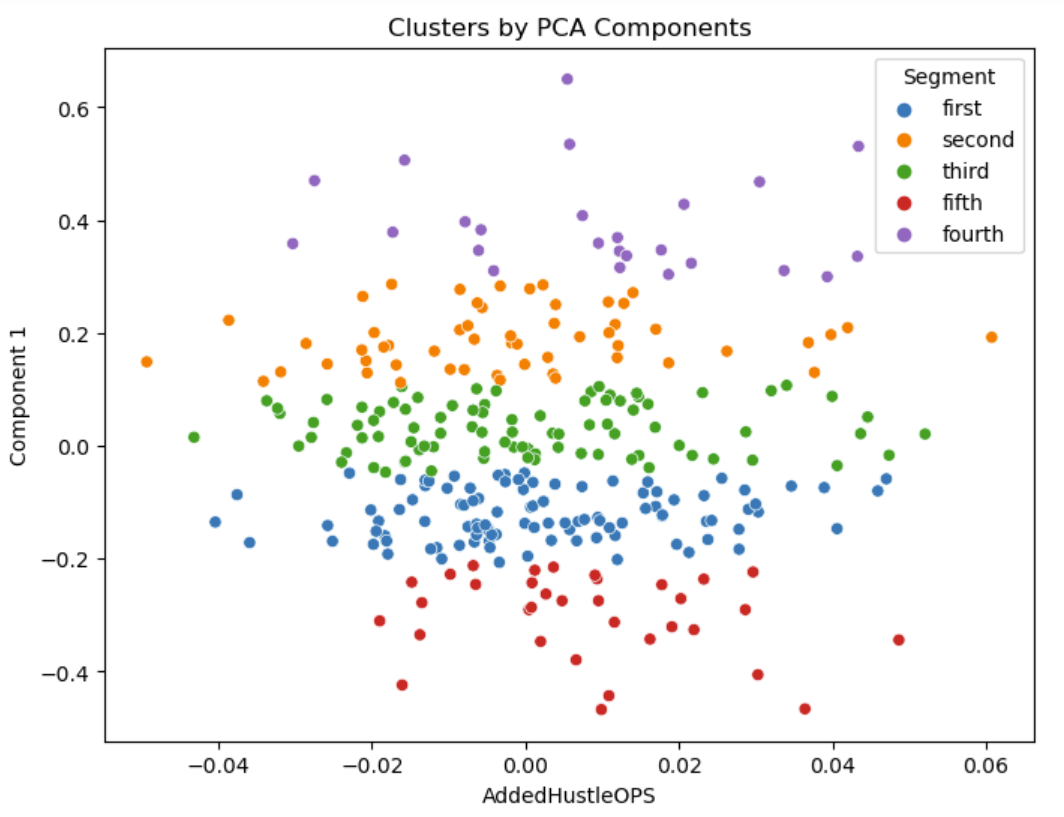
A picture containing text, screenshot, display, diagram

Description automatically generated

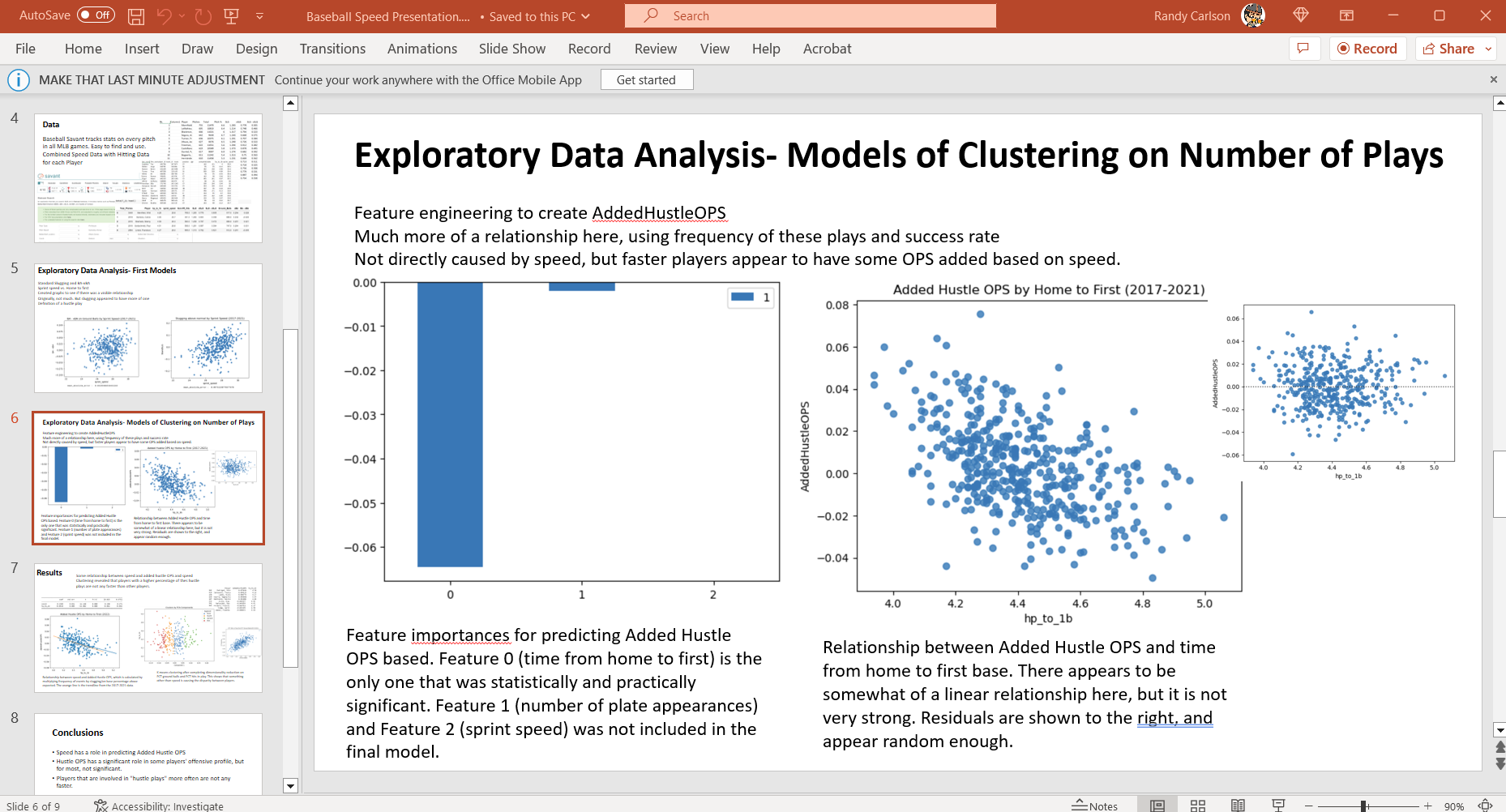
**8) Bootstrap tests.** Bootstrapping was referenced as a valuable deliverable in the class. We are currently investigating how (or if) to integrate this technique into our project. This may have to be abandoned as a technique, although we do want to practice it for pedagogical reasons. Status: **Descoped from paper.**

**9) Non-parametric tests.** This was accomplished by a PCA analysis and a visualization. See “Clustering” below. - **Complete**

**10) Clustering.** We added K-Means clustering to this project. K-Means was valuable to our project and showed that while there is some relationship in speed, percent of ground balls, and slugging on hits in play predicting added OPS on hustle plays, this relationship is not linear. This showed us that there are other factors than speed, possibly random chance.



**Presentation and Paper Status** We have a working PowerPoint presentation, and we have an example slide shown in the graphic below.



**Paper Status.** We have begun construction of the paper in the Overleaf LaTEX editor. We plan to publish a pdf version of the final paper. An example of the Overleaf work is below.  
