For my DSC530 Final Project, I wanted to review switch-hitters and whether they would be better served hitting from one side instead of both sides. My initial thought would be that they should hit from their strongest side which would translate into more hits and more runs and more wins for their team. I wanted to do this by analyzing switch-hitter performance against both left and right-handed pitchers. And comparing that performance against left-handed and right-handed hitters against both left and right-handed pitchers. While I was hoping to have a breakthrough that would get my hired on to the Cleveland Indians staff, what I believe I found was what Major League teams have known for a while.

In my initial analysis of the data, I noted that outliers were a bit reversed here. Whereas most of the players had fewer than 200 at-bats and the outliers appeared to be players with more than 200, we wanted to remove the players with fewer at bats as they were not everyday hitters. Beyond that, I did not remove any additional outliers. Though they may have been a mistake. When creating various groups, there were some “everyday” players that rarely played in certain situations. And I think some of these could have been removed. Additionally, there are some very good baseball players that hit very well against everybody. I think these players probably should have been filtered out as well.

For my exploratory data analysis, as I stated, I believe I uncovered what professional baseball knows already. Specifically, that some left-handed hitters really struggle against left-handed pitching. And in general, hitting against the opposite hand is more difficult. And that being a switch-hitter provides some benefit. As noted in my Jupyter Notebook, my recommendation to budding professional players who may be on the cusp, would be to give switch-hitting a try. Especially if they struggle against the opposite handed pitcher.

After about half-way through my analysis, I realized it would have been helpful if I did not have everything broken out by groups from the get-go. All the relevant baseball statistics I used as variables in my analysis were split by pitcher-handedness. I think it may have been helpful to have some overall statistics and then break them down by pitcher handedness. As far as variables that could have been helpful, I am not sure. Major League Baseball provides a wealth of information. One statistic I think may have been nice to use would have been “contact quality”. This indicates how well a hitter contacted the baseball on a pitch in which they swung and hit the baseball. The data I would have liked to have had would have been individual play data for each hitter, especially for swings/whiffs/contact/etc. This data is available but would have taken significant amount of time to pull and generate.

The biggest challenge I faced was trying to determine which baseball statistics to use to help explore my hypothesis. I was a bit all over the place in researching the different exploratory variables and even researching different dependent variables. This probably made my project a bit hard to follow. I also struggled with understanding some of the underlying statistics. Our book provided some nice methods that made generating PMFs, CDFS, histograms, etc. easier. But since his package is not available in the python repository, I probably will not use it for future projects. I wasted time trying to reverse engineer his code to understand it a python level instead of a ThinkStats2 level.

All in all, though, I had fun pulling and researching MLB data. I may have struggled with some aspects. But I also learned a lot. I became significantly more comfortable with Jupyter Notebooks and various Python libraries. While my specific hypothesis may not have been successful, I consider my project a success because of the sheer amount of work and effort I put forth and the knowledge I gained personally.