**Introduction**

Major League Baseball has been America’s sports pastime for over 100 years and was first founded in 1871 for the National League and 1901 for the American League. Today, there are 15 teams in each league. In the year 2000, the two leagues merged into what is now known as Major League Baseball (MLB). Since the very beginning, statistics in baseball has played a major role in the game. In today’s baseball, advanced metrics are being used by every major league team in order to gain advantage over their competition. MLB organizations employ data science teams to collect this information for executives, general managers and coaches. But, statistics in baseball have always been polarizing. Some managers have lost their jobs recently because they could not adopt and did not believe in advanced metrics.

Quote from Bobby Bragan (baseball manager – 1940’s) – “Say you were standing with one foot in the oven and one foot in the ice bucket. According to the percentage people, you should be perfectly comfortable.”

Quote from Leonard Koppett (A Thinking Mans Guide to Baseball – 1967) – “Statistics are the lifeblood of baseball. In no other sport are so many available and studied so assiduously by participants and fans. Much of the game’s appeal, as a conversation piece, lies in the opportunity the fans get to backup up opinions and arguments with convincing figures, and it is entirely possible that more American boys have mastered long division by dealing with batting averages than in any other way.”

Baseball contracts involve intense negotiations and millions of dollars are at stake. Recently, the St. Louis Cardinals signed Paul Goldschmidt, agreeing to a five-year, $130 million contract. Paul Goldschmidt is 32 years old. Age plays a big role in a baseball players performance. At some point as players get older, their performance on the field inevitably starts to decline. I have wondered whether the Paul Goldschmidt deal was good for the St. Louis Cardinals, and some have said, according to Forbes Magazine, there is reason to believe he could be entering the decline phase of his career. The Cardinals are betting that he will produce through the age of 37 years old. Will Paul Goldschmidt continue to perform through the age of 37? In more general terms, I would like to do analysis on MLB hitters and look at past performance and predict the lifecycle of a major league baseball batter and look at optimal contract terms for the next superstar position players of the game using machine learning models. My customer in this analysis is baseball teams, analysts and fans.

The “steroid era” in baseball was a dark time for MLB baseball. Even today some of the greatest players of all time are not in MLB’s Hall of Fame because of suspected steroid use. Mark McGuire is one such example. There is no exact start and end date to the steroid era. ESPN has defined it as starting in the later 1980’s and ending in the late 2000’s. Others have defined it to be from 1993 to 2003. The latter definition will be used for this project. In 2003, MLB introduced performance enhancing drug testing. Analysis of hitting performance of the pre-steroid era (1982 to 1992), the steroid era and post-steroid era (2004 to 2014) will be performed. I would like to know if modern advanced metrics show a significant advantage to player’s performance using OPS as a metric. My customer is analysts who refuse to let players who took (or suspected of taking) performance enhancing drugs (PEDs) into the hall of fame because of unfair advantage. It is also for fans who want more information on the subject. Did they really have an advantage?

As ESPN Analyst Harold Reynolds said, “All of the sudden, it’s not just BA and Runs Scored, it’s OBA. And what is O-P-S?” Certainly, the “old standard” hitting metrics like batting average and runs scored have given way to more advanced metrics such as OPS (on-base plus slugging) which is a more meaningful metric on how well a player is performing at the plate. There are many other advance metrics today in baseball as well. I will be using OPS for this project. OPS is calculated by adding a player’s on-base percentage with their slugging percentage.

The details of the equations are as follows:

OPS = OBP + SLG

OBP = (H + BB + HBP) / (AB + BB + SF + HBP)

SLG = TB / AB

Where:

H – total number of hits of a player

BB – total number of walks (base on balls) of a player

HBP – total number of times the player was hit by a pitch

AB – total number of plate appearances (times at bat) by the player

SF – total number of sacrifice flies of a player

TB – is the total bases and is a weighted sum ( 1 for single, 2 for double, 3 for triple, 4 for HR).

So, TB is ∑ ( (nSingle \* 1) + (nDoubles \* 2) + (nTriple \* 3) + (nHomeRuns \* 4) ) where nSingles is the number of singles, nDoubles is the number of doubles, nTriple is the number of triples and nHomeRuns is the number of home runs.

OPS – on-base plus slugging

OBP – on-base percentage

SLG – slugging percentage

NOTE: all statistics are taken over a period of time (typically a year)

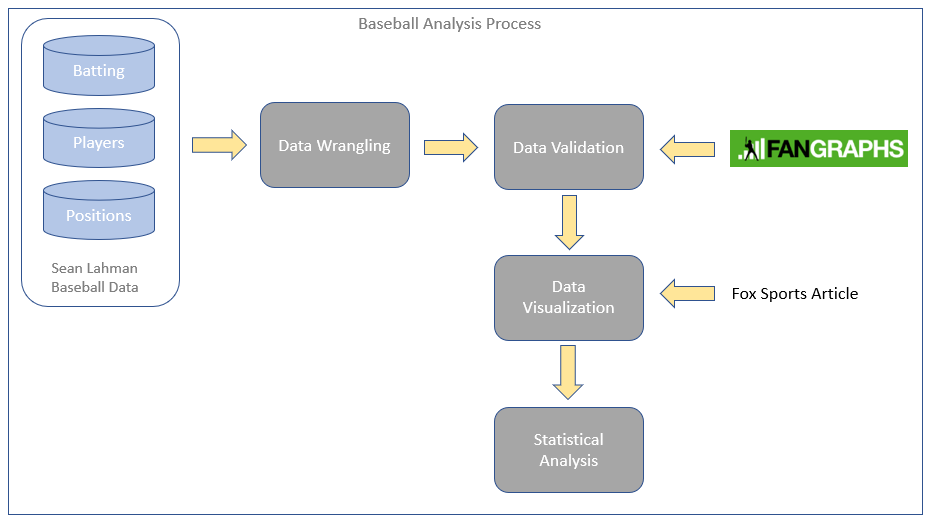
(source : Wikipedia)

I would like to use OPS for both parts (performance/age prediction and PED usage) of analysis. In order to do this analysis, I have done some internet research and found raw baseball data collected from 1871 to 2018 of major league baseball games. All the above atomic data elements such as hits, at bats, etc are available and therefore OPS, OBP and SLG can be computed. Thanks to Sean Lahman and others, they have created a database with yearly baseball statistics from 1871 to 2018. The database has copyright 1996-2018 by Sean Lahman. I have read the license agreement which is licensed under Creative Commons Attribution and will not restrict me from using this data. The raw data needed will be from the 1954 to 2018.

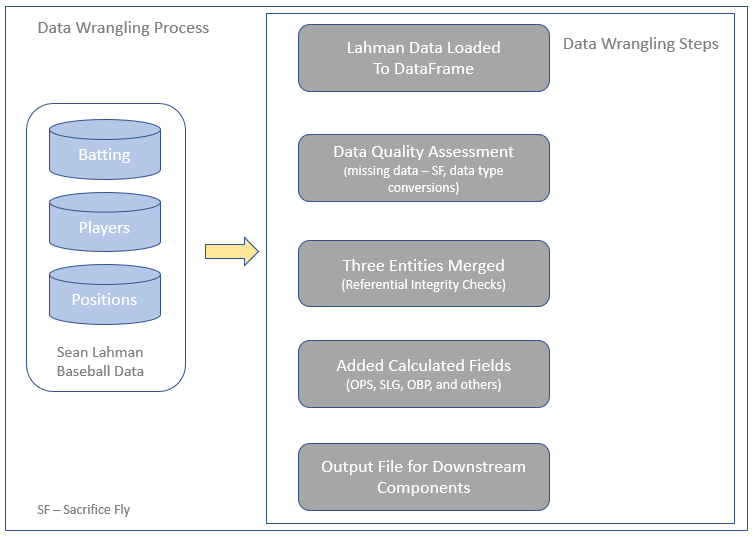
Tony La Russa (ex St. Louis Cardinal Manager) has been quoted as saying (paraphrased) “you may not agree with me, but you don’t have all of the information that I have”. Now we do.

**Stages of Process**

There are four main components of the baseball analysis process:



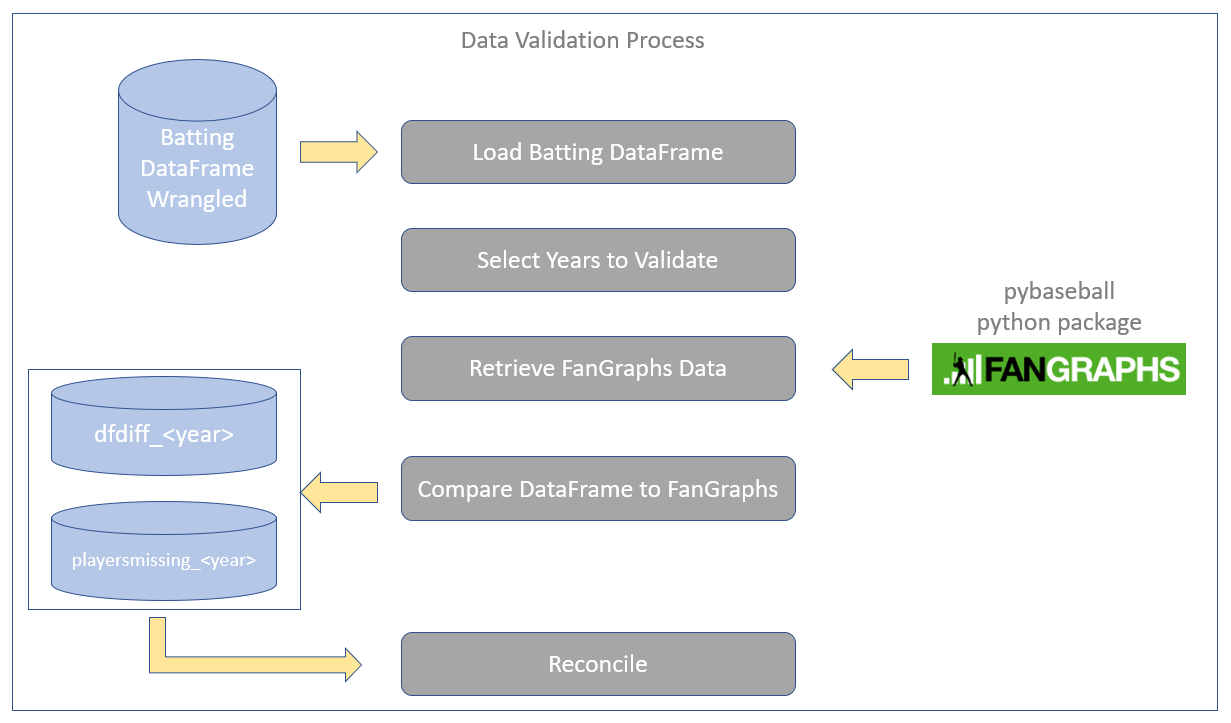
The four steps are Data Wrangling, Data Validation, Data Visualization and Statistical Analysis. The first step is Data Wrangling. Data was downloaded from the Sean Lahman site and staged for loading. There were three main data entities that were loaded: Batting, Player and Position data. Why only use data from 1954 when data was available from 1871 onwards? The sacrifice fly was not tracked consistently until 1954. According to Wikipedia, “batters have not been charged with a time at-bat for a sacrifice hit since 1893, but baseball has changed the sacrifice fly rule multiple times. The sacrifice fly as a statistical category was instituted in 1908, only to be discontinued in 1931. The rule was again adopted in 1939, only to be eliminated again in 1940, before being adopted for the last time in 1954”. Sacrifice flies are required for OPS calculation, and for this reason only data from 1954 onward is used. Here is the process for Data Wrangling:



A data quality assessment was performed, the three entities were joined, additional columns were added and the DataFrame was written to a CSV file to be used in downstream processes. Here is a link to the code:

<https://github.com/paulscheibal/SBDataScienceCert/blob/master/CapstoneP1/DataWrangling/CapstoneBaseballDataWrangling.py>

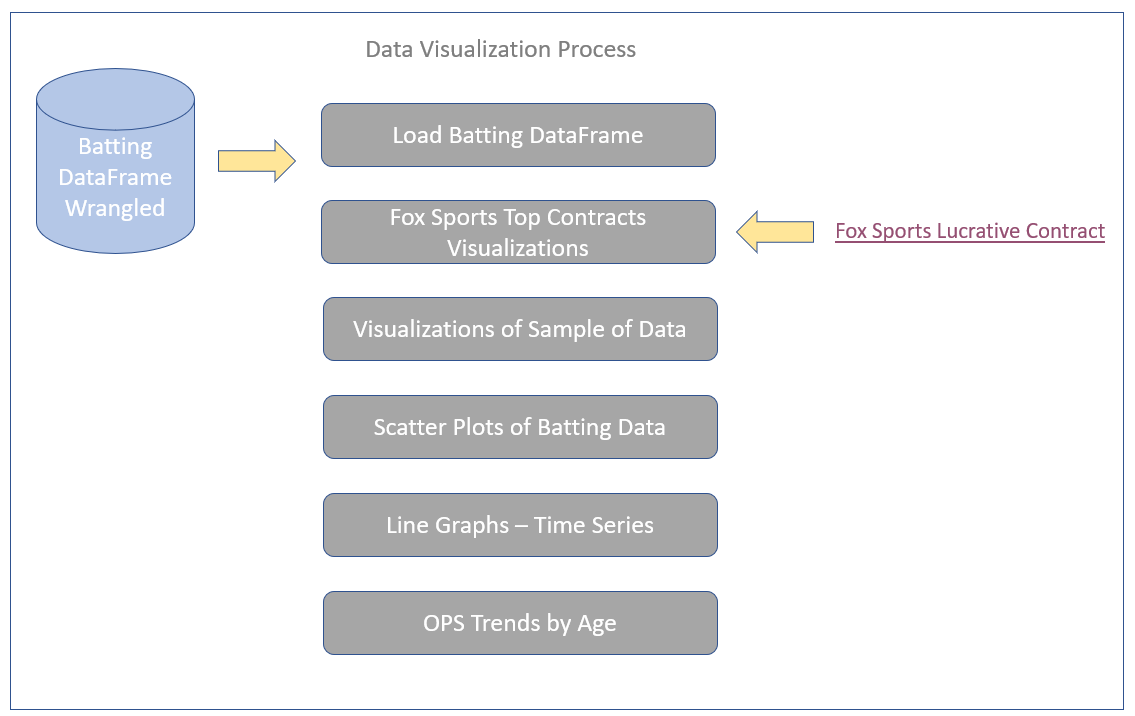
The next step in the process is Data Validation. The following diagram defines the data validation process. In order to independently validate the Lahman data after all the data wrangling was performed, the FanGraphs API was used. The pybaseball package integrated the FanGraph API. All that was needed was a function call which implemented the API which was provided by the pybaseball package.



All the wrangled Lahman data was successfully reconciled using FanGraphs. Here is a link to the Data Validation code:

<https://github.com/paulscheibal/SBDataScienceCert/blob/master/CapstoneP1/Validation/DataValidation.py>

The third step in the process was Data Visualization. During this step, EDA was performed. To make it interesting, a Fox Sports article listed the top contracts in MLB which listed the dollar amount of the contract, the duration of the contract and when it was signed. This data was manually entered into a spreadsheet, loaded and integrated with the wrangled Lahman data. The following flows the overall data visualizations performed:



Here are the links to the Fox Sports article and the link to the data visualization Jupyter Notebook.

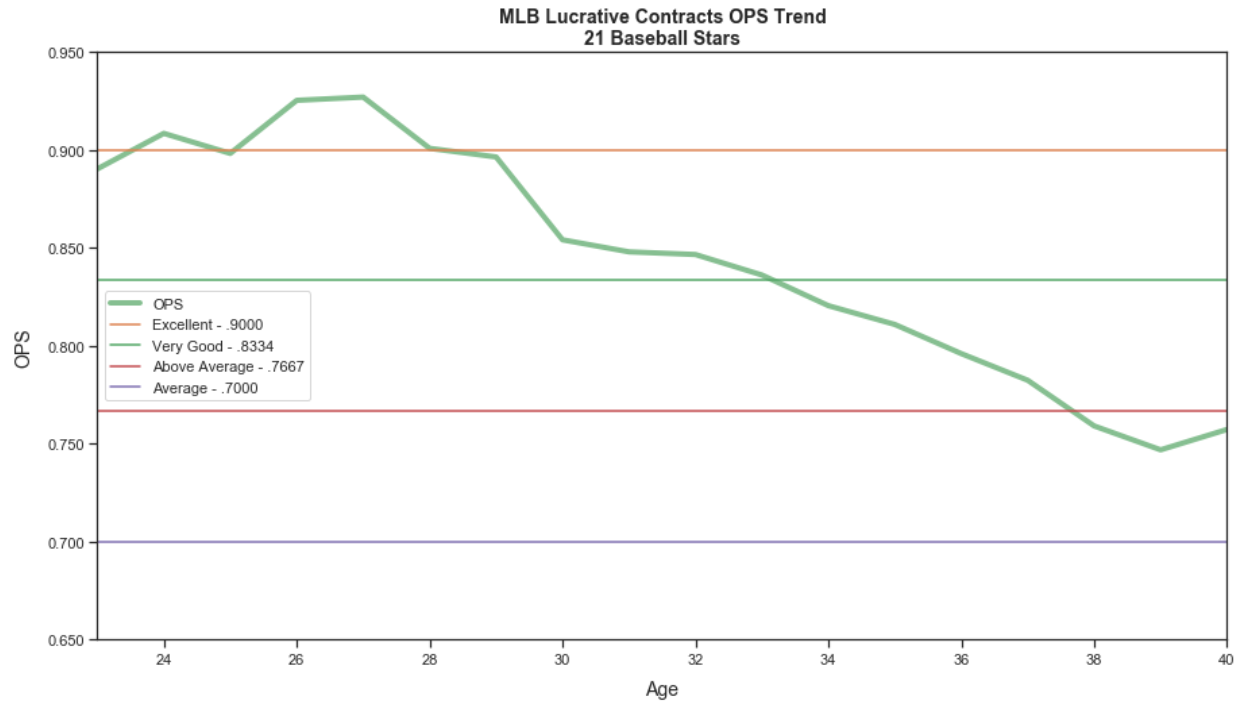
Fox Sports Article:

<https://www.foxsports.com/mlb/gallery/mlb-long-lucrative-contracts-results-120711>

Jupyter Notebook – Data Visualizations

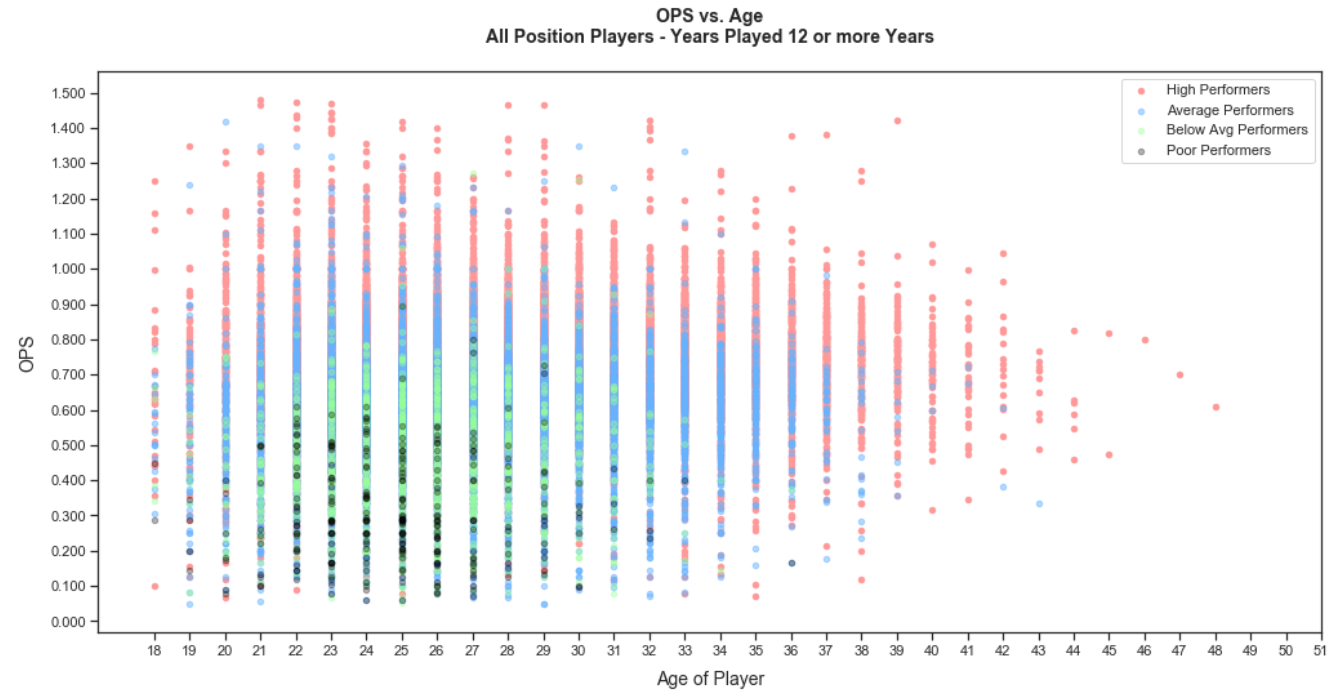
<https://nbviewer.jupyter.org/github/paulscheibal/SBDataScienceCert/blob/master/CapstoneP1/Discovery/DataStory.ipynb>

The two charts that are most interesting to me are the OPS trend of lucrative contracts and the scatter plot of players looking for trends.



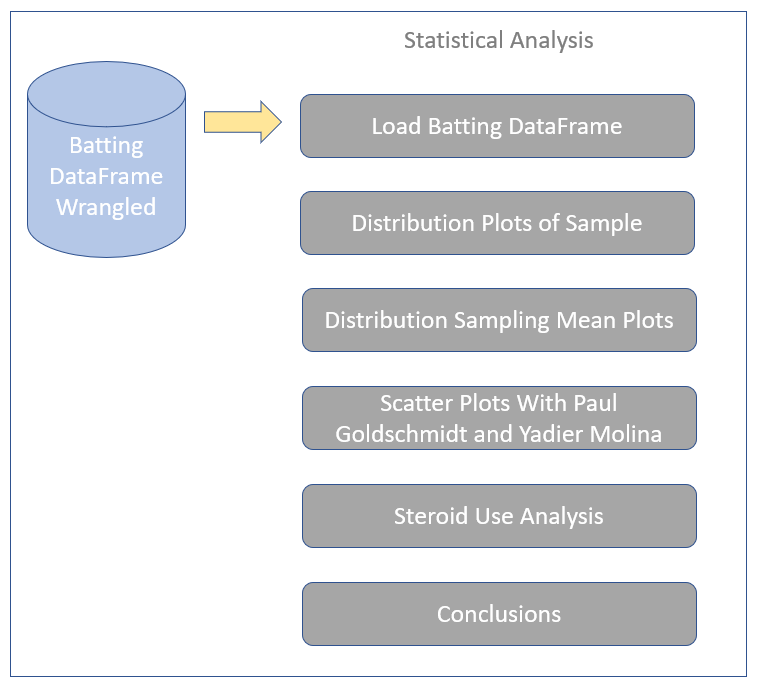
What is interesting to me is the average signing age is 28 years of age. Looking at the above chart, you can see that player performance degrades over time shortly after the contract is signed.

Below is the scatter plot of all MLB players from 1954 to 2018. Note the bands of colors representing different categories of players.



You can already observe in the color bands that there is an upswing in performance of each category of player and then a downswing of performance over their career.

To further my understanding and answer my questions, statistical analysis was performed. The following summarizes the analysis steps as follow:

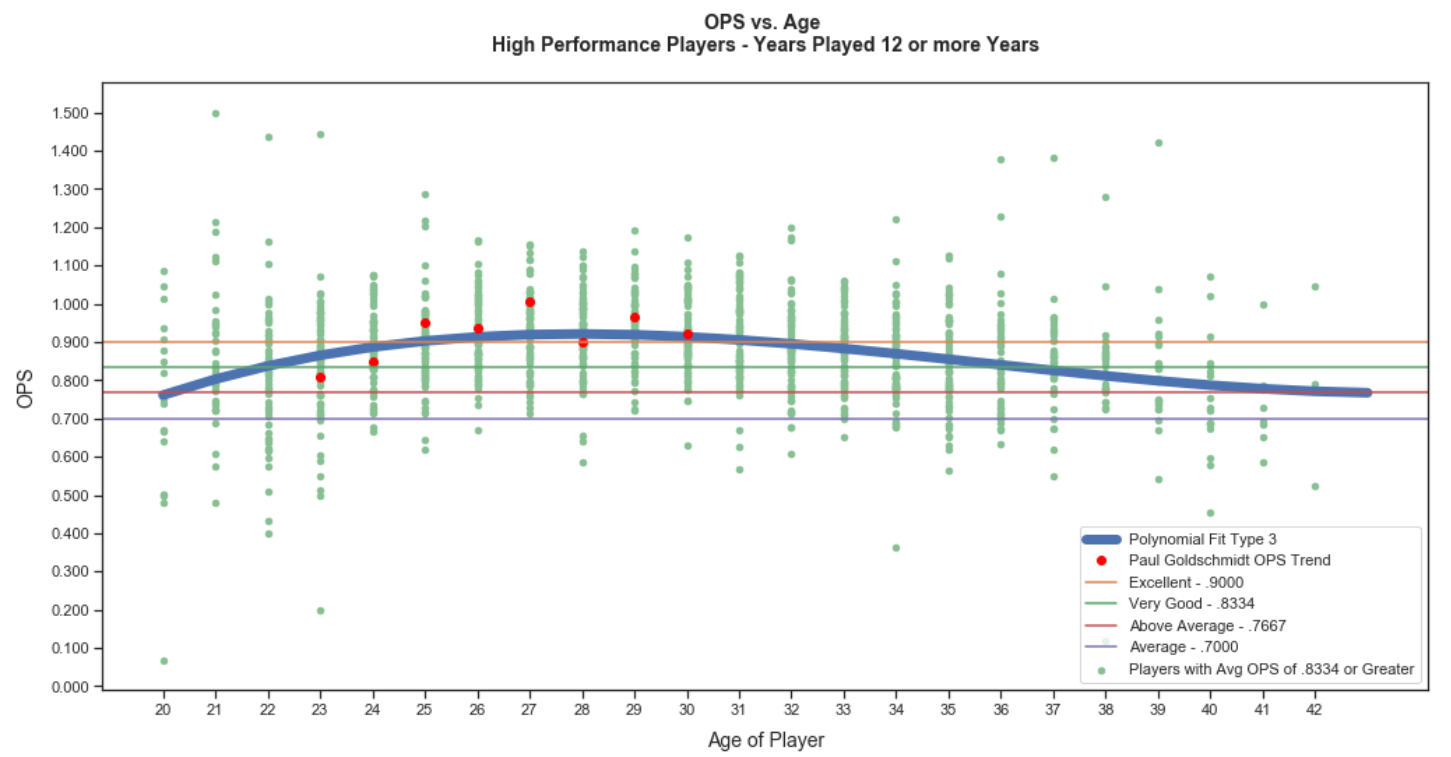


The Jupyter Notebook with the full analysis can be found by clicking on the following link:

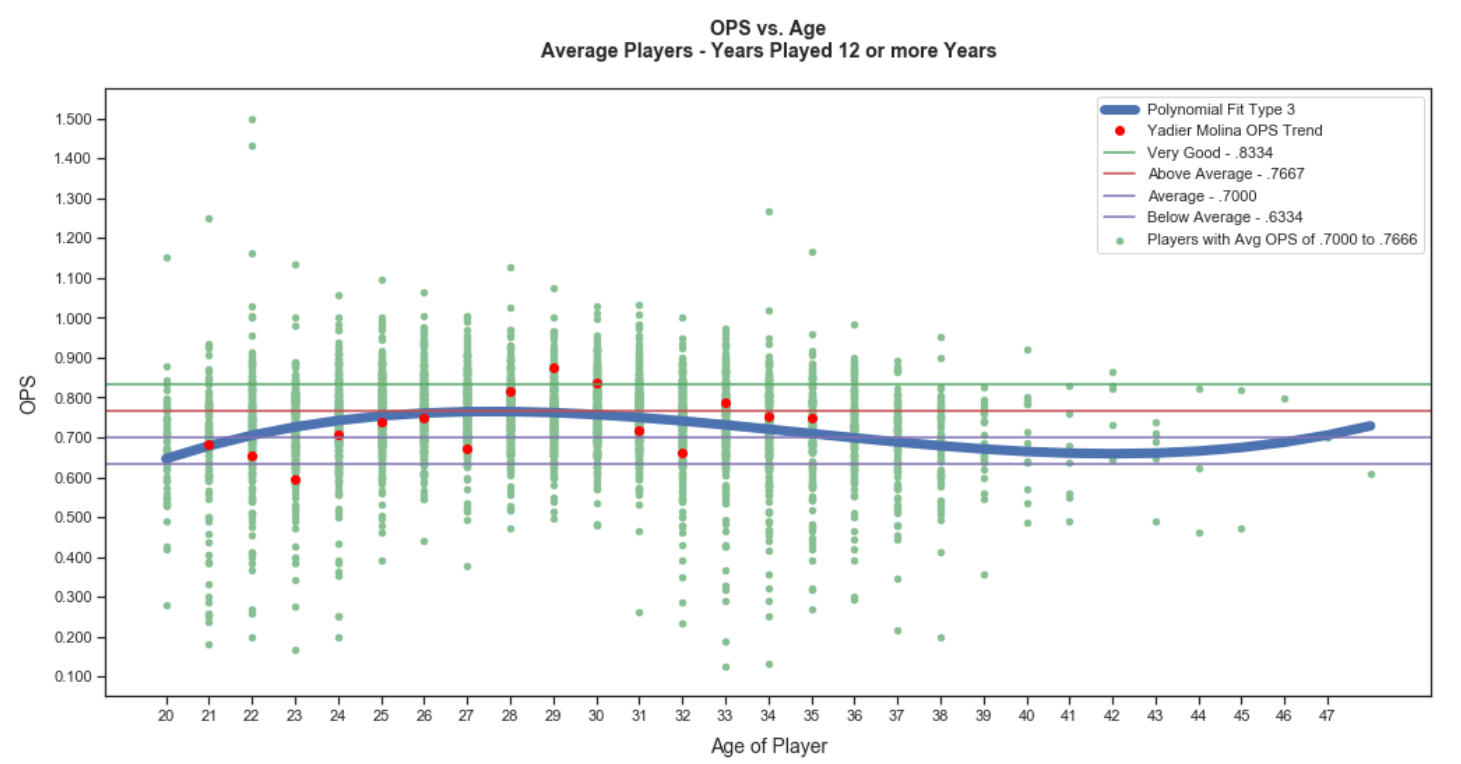
<https://nbviewer.jupyter.org/github/paulscheibal/SBDataScienceCert/blob/master/CapstoneP1/Statistics/StatisticalAnalysisStory.ipynb>

There are four charts that I would like to highlight as part of the statistical analysis: Paul Goldschmidt performance plot, Yadier Molina performance plot and two OPS time series plots regarding steroid usage.

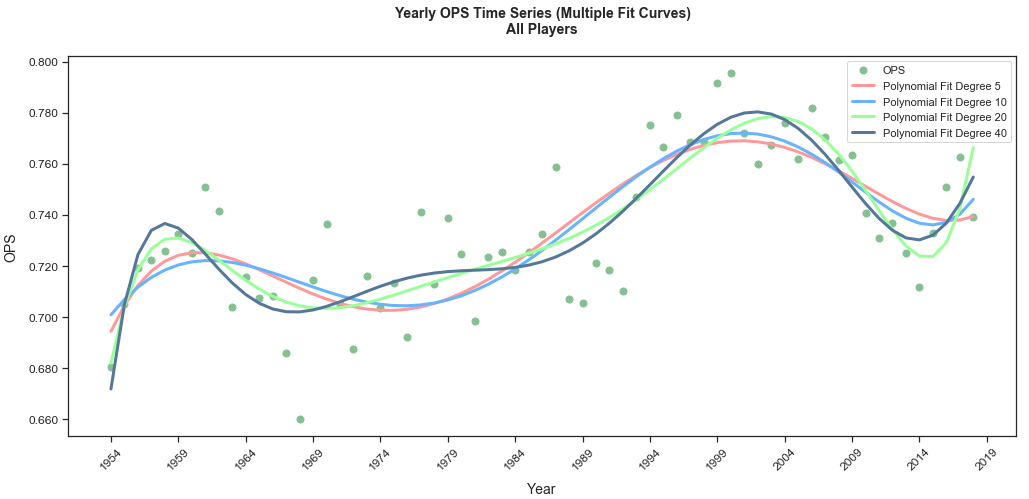
The following is a scatter plot showing high performing players with an average OPS of .8334 or higher. Overlaid on the plot is a predictive curve using the NumPy ployfit function. This function fits a curve over the data that minimizes the squared error. In addition, Paul Goldschmidt’s performance is plotted as well. 2019 was a transition year for Paul as he was traded to the St. Louis Cardinals from the Arizona Diamondbacks and his performance numbers were lower than expected (OPS of .821) which may be due to the transition. According to the predictive curve, his OPS numbers should be between .8334 and .9000 for the rest of his contract with the Cardinals. If his actuals hold true to the predicted curve, the Cardinals will see a good return on their investment. Not superstar level as he once was, but in the “very good” OPS range. I think the St. Louis Cardinals will be quite happy with that.

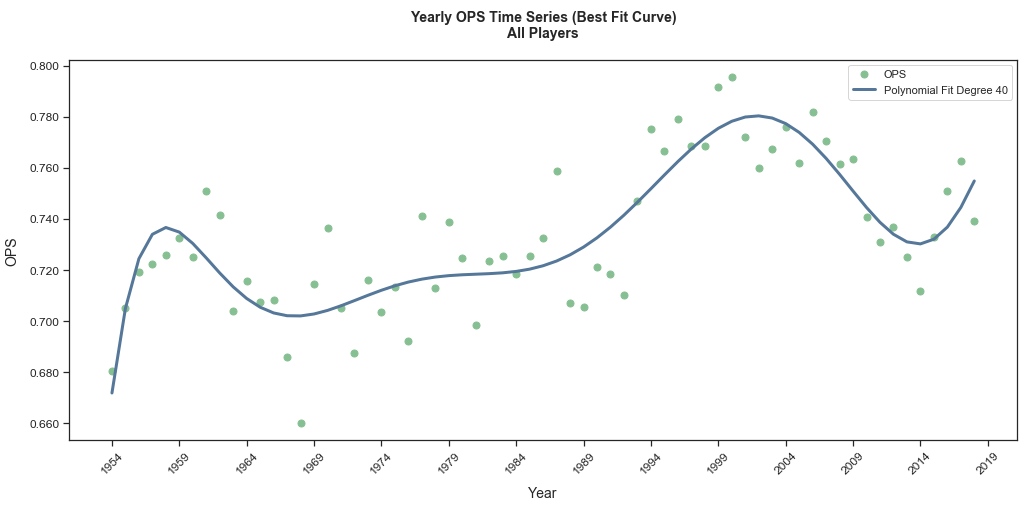


Being a big Yadier Molina fan, I also charted his OPS numbers against the Average OPS predicted curve for fun.



The final two charts show the time series of OPS numbers from 1954 to 2018 with a trend line(s) using NumPy’s polyfit function. The first chart shows a series of polynomial fitting curves overlaid with the time series. The final chart shows the one with the “best fit” and shows an upward trend in the early 1970's and continued to trend upwards until about the mid 2000's. However, at around 1990 the upward trend is much more severe, lending credibility that the steroid era really did benefit position players. This plot does not prove that steroids did give players an advantage, but it does show that there seems to be a "player friendly" correlation between OPS and the steroid era.





For those who vote for MLB Hall of Fame induction, I hope they consider my work. I am not saying steroids didn’t benefited players during this era. There seems to be evidence that it may have. I am just saying I don’t see conclusive evidence that it did.