Let’s Talk About Baseball

Isabel Lopez
Department of Human-Computer Interaction
School of Informatics & Computing
lopezi@iu.edu

Wyatt Van Dyke
Department of Human-Computer Interaction
School of Informatics & Computing
wyvandyk@iu.edu

Introduction

Baseball has grown into an extremely popular sport over the years. Its origins are officially unknown, but histories say it was inspired by the England sports cricket and rounders. The sport of baseball started from just a small league based out of New York, to now a multi-million dollar sport. ((*Who really invented baseball?,* 2021))Baseball has become a sport rich in statistics, which makes it a well-suited subject for statistical analysis. However, the sheer volume of statistics means that while there are a wide variety of questions that can be asked, it can be difficult to determine, and best be able to leverage so much data. This project is an attempt to remedy that by providing methods to better understand baseball statistics, describe challenges that you may encounter while handling baseball data, and show some data analysis that can directly influence players in the current sport of baseball.

Problem Description

As mentioned before, baseball’s long history and depth and breadth of information and statistics make it a valuable source of statistics, but the volume of data can be difficult to make sense of. Our objective is to create models to make predictions and aid understanding of baseball statistics. The question we want to answer specifically is what impact does age have on performance of baseball player on the field, and what model is best for predicting the relation between age and performance if there is any.

Data Description

The official dataset which can be found on Kaggle provided us with 20 different files ranging from Manger information, to Awards to, player stats. In total the data contained a total of 291 attributes ranging from continuous to discrete, and containing over 100,000 different lines of data. From this dataset, we used an altered version, which we found would better get at the question we were trying to answer. We used only the MASTER, Batting, Pitching, FieldingOF, and Salaries tables found in the dataset in these links. Our specific files that we called

PeopleBatting, PeoplePitching, PeopleFieldingOF, and PeopleSalaries, were created by an inner join with the MASTER table and the respective table, with an age column calculated by the year the player played in minus the player's birth year. We added the MASTER table specifically to join the tables by PlayerID, the MASTER table contained the player's birthday, and other relative information that could help us while performing analysis on our data.

We joined and cleaned our data in R to be able easily use our data in Pyspark. To clean the data we used an onmit function that took out data that was missing data from several columns. We kept rows that had less than 4 columns of data missing, to not only have as many data rows as possible, but to be able to have a wide range when it came to running our analysis.

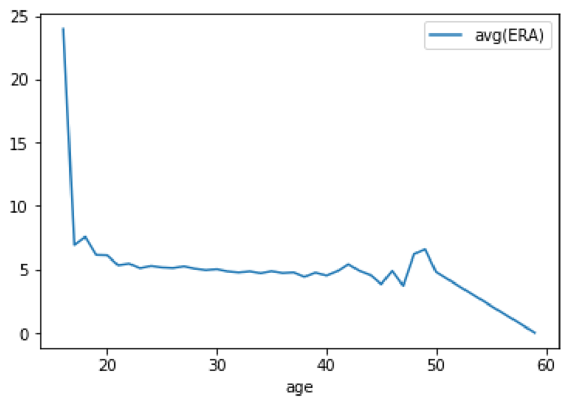
Methodology

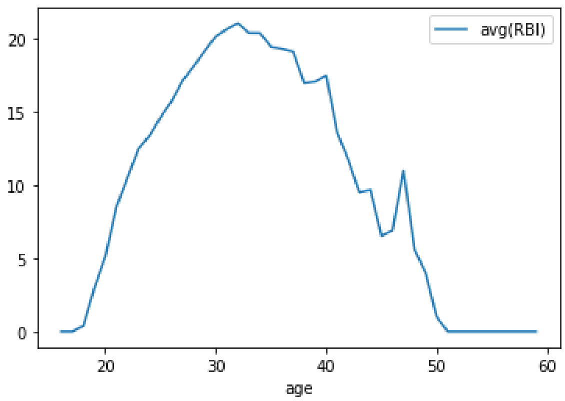
We initially did this project in R where we did some exploratory analysis and even created a few linear regressions, but we found out later that we were required to do it in Pyspark, so we recreated the work we did in R up to that point in Python. We faced a major roadblock in the form of troubles with the LinearRegression function that caused us about a week's worth of delay, but we managed to get it to work.

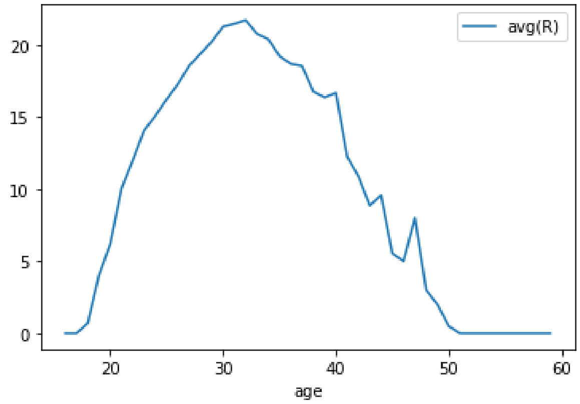
To prepare for the models we used a 70%, 30% split for each of the models we ran. We then ran a linear regression for the PeopleBatting, PeoplePitching, and PeopleSalaries datasets to predict "age" based on a number of predictors that we inferred that age would logically impact. In addition, we took the mean by age of some significant statistics and plotted them against age to determine the relationship of age on that specific statistic directly.

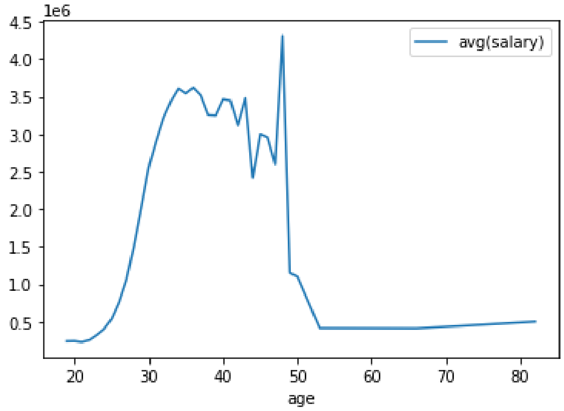
Results

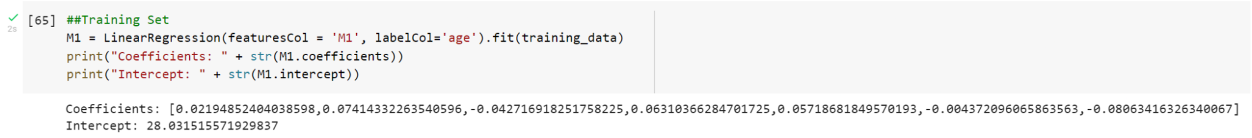
The graphs below plots we created that we believe would be significant to the sport of baseball. In addition to running linear regressions, we divided the datasets into their different age values and took the average value of a relevant statistic. Each graph is related to a player's performance on the field, and how age can influence the player directly. You can see that we also graphed a players salary, which can be directly correlated to a players age as well. We see that Age has a positive correlation on several variables, up until a certain age, the graph then drops off significantly. We found the graph for the mean Earned Run Average by age had a very high point at first before rapidly going low and plateauing, then spiking again and lowering again. Means by age for Runs Batted In and Runs both had a mostly bell-curve like form, with a slight skew to the left. Salary was unusual in that it starts out like a bell curve then gets highly irregular before making a big spike upwards then going down rapidly to plateau, then rising slightly. In the sport of baseball this spike is common when a player is near retirement, and has signed to the team they will be retiring with.



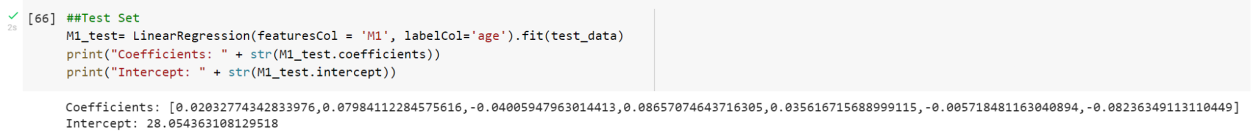






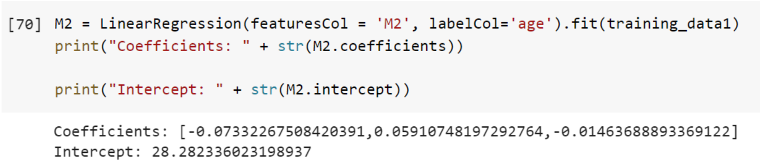


```
[65] ##Training Set
     M1 = LinearRegression(featuresCol = 'M1', labelCol='age').fit(training_data)
     print("Coefficients: " + str(M1.coefficients))
     print("Intercept: " + str(M1.intercept))

     Coefficients: [0.02194852404038598,0.07414332263540596,-0.042716918251758225,0.06310366284701725,0.05718681849570193,-0.004372096065863563,-0.08063416326340067]
     Intercept: 28.031515571929837
```

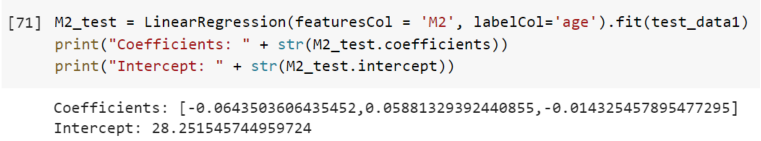


```
[66] ##Test Set
     M1_test= LinearRegression(featuresCol = 'M1', labelCol='age').fit(test_data)
     print("Coefficients: " + str(M1_test.coefficients))
     print("Intercept: " + str(M1_test.intercept))

     Coefficients: [0.02032774342833976,0.07984112284575616,-0.04005947963014413,0.08657074643716305,0.035616715688999115,-0.005718481163040894,-0.08236349113110449]
     Intercept: 28.054363108129518
```

The linear regression for the pitching data had a larger set of predictors than the other linear regressions: hits, earned runs, walks, wins, losses, strikeouts, and runs allowed. From the coefficients, we can see that hits, earned runs, and wins and losses are positively correlated with age, while walks, strikeouts, and runs allowed are negatively correlated with age.

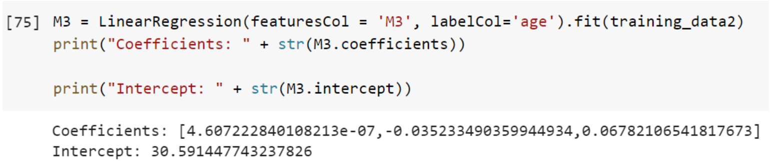


```
[70] M2 = LinearRegression(featuresCol = 'M2', labelCol='age').fit(training_data1)
     print("Coefficients: " + str(M2.coefficients))

     print("Intercept: " + str(M2.intercept))

     Coefficients: [-0.07332267508420391,0.05910748197292764,-0.01463688893369122]
     Intercept: 28.282336023198937
```

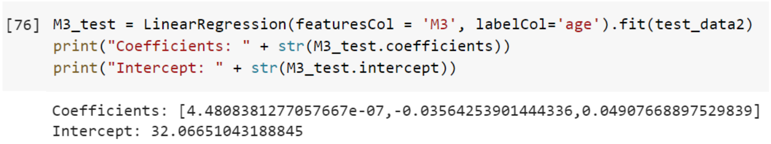


```
[71] M2_test = LinearRegression(featuresCol = 'M2', labelCol='age').fit(test_data1)
     print("Coefficients: " + str(M2_test.coefficients))
     print("Intercept: " + str(M2_test.intercept))

     Coefficients: [-0.0643503606435452,0.05881329392440855,-0.014325457895477295]
     Intercept: 28.251545744959724
```

The predictors we used for the second linear regression model for our batting dataset were homeruns, runs batted in, and hits. From the coefficients, runs batted in had a positive correlation and homeruns and hits had a negative correlation.

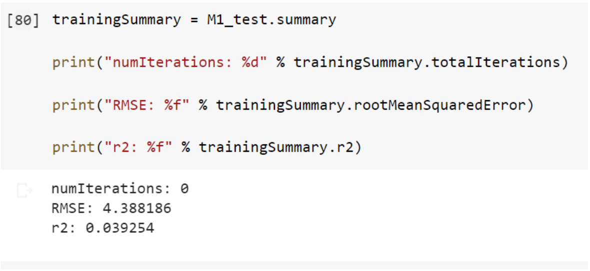


```
[75] M3 = LinearRegression(featuresCol = 'M3', labelCol='age').fit(training_data2)
     print("Coefficients: " + str(M3.coefficients))

     print("Intercept: " + str(M3.intercept))

     Coefficients: [4.607222840108213e-07,-0.035233490359944934,0.06782106541817673]
     Intercept: 30.591447743237826
```

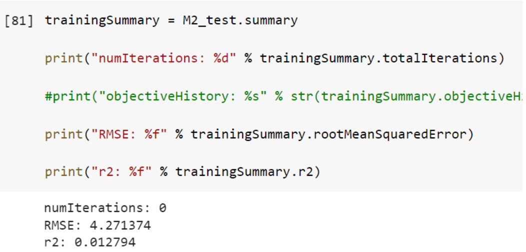


```
[76] M3_test = LinearRegression(featuresCol = 'M3', labelCol='age').fit(test_data2)
     print("Coefficients: " + str(M3_test.coefficients))
     print("Intercept: " + str(M3_test.intercept))

     Coefficients: [4.4808381277057667e-07,-0.03564253901444336,0.04907668897529839]
     Intercept: 32.06651043188845
```

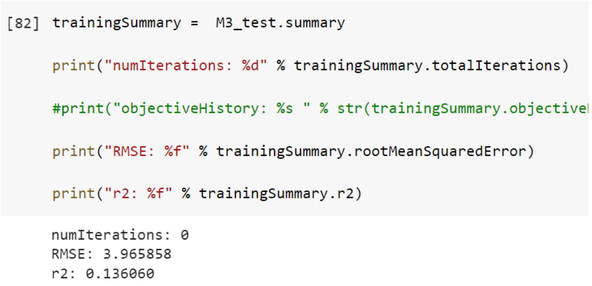
For our linear regression model for our salary dataset, we used height, weight, and salary as predictors. Height has no correlation with age, but weight had a negative correlation with age while salary had a positive correlation.



```
[80] trainingSummary = M1_test.summary

     print("numIterations: %d" % trainingSummary.totalIterations)

     print("RMSE: %f" % trainingSummary.rootMeanSquaredError)

     print("r2: %f" % trainingSummary.r2)

     numIterations: 0
     RMSE: 4.388186
     r2: 0.039254
```



```
[81] trainingSummary = M2_test.summary

     print("numIterations: %d" % trainingSummary.totalIterations)

     #print("objectiveHistory: %s" % str(trainingSummary.objectiveH

     print("RMSE: %f" % trainingSummary.rootMeanSquaredError)

     print("r2: %f" % trainingSummary.r2)

     numIterations: 0
     RMSE: 4.271374
     r2: 0.012794
```



```
[82] trainingSummary =  M3_test.summary

     print("numIterations: %d" % trainingSummary.totalIterations)

     #print("objectiveHistory: %s " % str(trainingSummary.objective

     print("RMSE: %f" % trainingSummary.rootMeanSquaredError)

     print("r2: %f" % trainingSummary.r2)

     numIterations: 0
     RMSE: 3.965858
     r2: 0.136060
```

Every linear regression we ran had a very low R-squared value and a very high RMSE value, though our salary linear regression model had the lowest RMSE and highest R-squared value. This may indicate that more models need to be tested for the optimal method of modeling the different datasets. This would most likely be a non-linear model such as a polynomial regression model, judging from the curve the Runs and Runs Batted In versus mean by age charts had. That chart suggests that for batters, skill develops as players get older following a bell-curve, and have a significant drop off when they reach a certain age judging from the graphs with tends to be between 30-40.

Conclusion

When it came to answering our question we came into a lot of challenges that prevented us from answering the question we would have liked. With our models R-squared being so low, we weren't able to come up with a model with a statistically significant predictor. What we were able to come to understand was that the way your data is structured and the way you analyze are extremely important to building linear regression models. We have come to the conclusion that the sport of baseball itself is structured in such a way that you would more than likely need to use multiple types of models to better understand a single predictor such as age. In the future we hope to take what we have learned and apply it to real world experience, and be able to solve problems like these quickly, and efficiently.

References

https://www.kaggle.com/open-source-sports/baseball-databank
http://www.seanlahman.com/baseball-archive/statistics/

Encyclopædia Britannica, inc. (n.d.). *Who really invented baseball?* Encyclopædia Britannica. A Retrieved December 13, 2021, from a https://www.britannica.com/story/who-really-invented-baseball.

Appendix 1: Contributions from each member
Project Idea: Wyatt Van Dyke
Data Source Finding: Wyatt Van Dyke
Data Cleaning & Concatenation: Isabel Lopez
Project Proposal: Wyatt Van Dyke, Isabel Lopez
Mean statistic charts: Wyatt Van Dyke
Linear Regressions: Isabel Lopez
Code Corrections: Wyatt Van Dyke

Final Report: Wyatt Van Dyke (outline, introduction, problem description, data description, methodology, pictures and text for results, references), Isabel Lopez (fleshing out, editing, making it look pretty, references)