Objective

This final report outlines our final project. It includes generating random test data, applying quality filters, performing a model fit with error analysis, and providing explanations for the model fit to data.

Data Source

We obtained our primary dataset, which comprises 1,035 MLB player records, from the UCLA statistics department. These records contain information about each player's team, game position, height, weight, and age. This dataset offers a rich source of information that can be used to explore various relationships within Major League Baseball player characteristics. We also created our own fake data in order to run similar regressions on MLB players and highlight its relation, as well as its reasonability in relation to our actual data.

Research Execution

1. Data Selection and Filtering

We began by selecting relevant variables from the dataset, including height and weight, which are the focus of our analysis. We made sure there were no outliers or missing data points.

2. Model Selection

Our analysis consists of constructing a polynomial regression model to investigate the relationship between height and weight in MLB players. We chose the polynomial regression model due to its flexibility in capturing nonlinear relationships in data. We expected to find an equation like this, where W is weight, H is height, and a,b, and c are constants.

$$W = aH^2 + bH + c$$

3. Model Fitting

We performed a regression analysis on the filtered dataset as well as the faked dataset, using two different Python packages. This analysis resulted in a line of best fit, which was accurate for both real and fake data. This went well because the regression analysis was able to accurately impose a line of best fit on both our real and fake data.

4. Random Test Data

To verify the reliability of our model, we randomly generated 1,000 test records (as mentioned above). These test records are used to evaluate the model's predictive capabilities and determine whether it produces reasonable results. The reasonability of our results was verified by this data after performing a regression analysis, and this is seen in the fact that our fake data is linearized similarly to our real data, and can be analyzed in a similar fashion (as seen in the line of best fit).

Outcomes

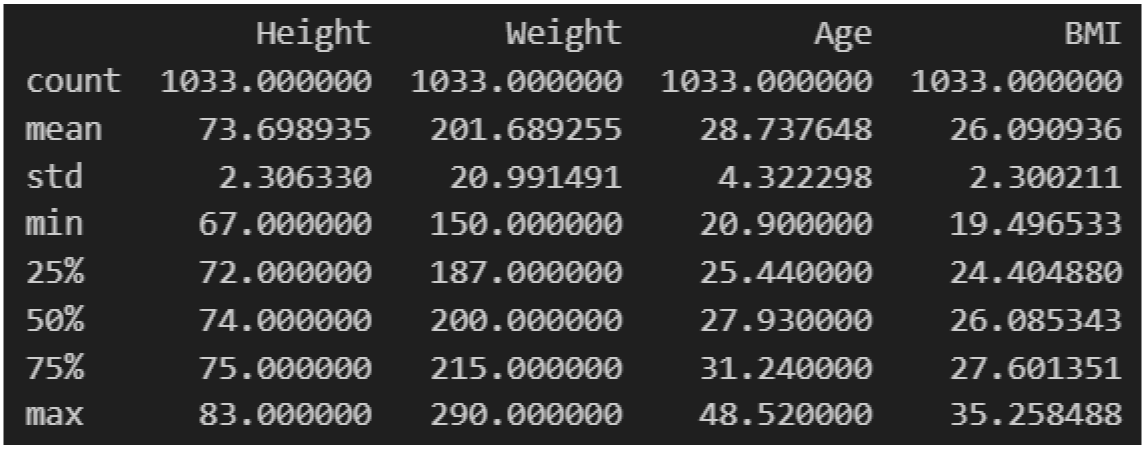
The analysis was expected to reveal insights into the relationship between height and weight among MLB players, and it was a success. Furthermore, the random test data served as a validation step to ensure the model's generalizability.

Conclusion

By utilizing the provided dataset, filtering it, selecting a suitable model, and conducting a model fit with error analysis, we were able to provide insights into the height-weight relationship among MLB players. In addition, we were able to detect trends on a baseball team, player position, and age basis. The graphs and their associated meanings are in the appendix.

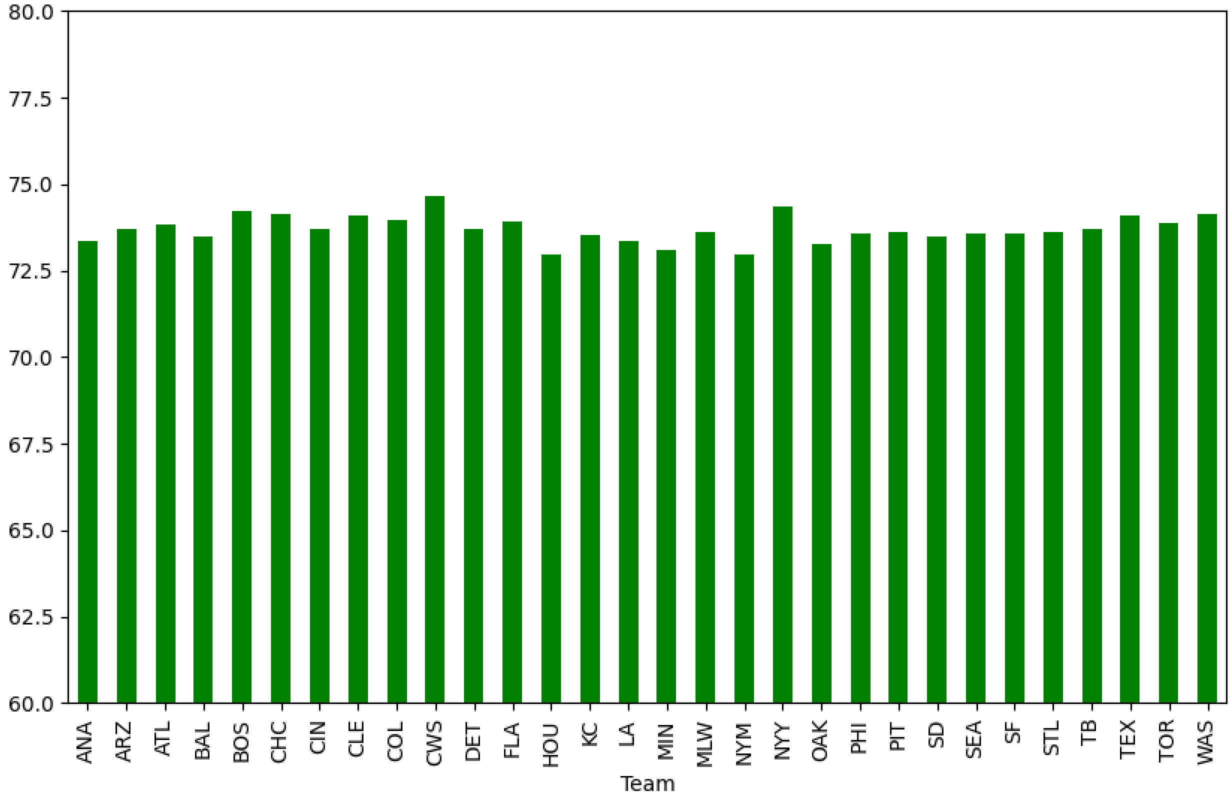
Appendix

We found the mean and standard deviation for Weight, Age, and Height in the data set.

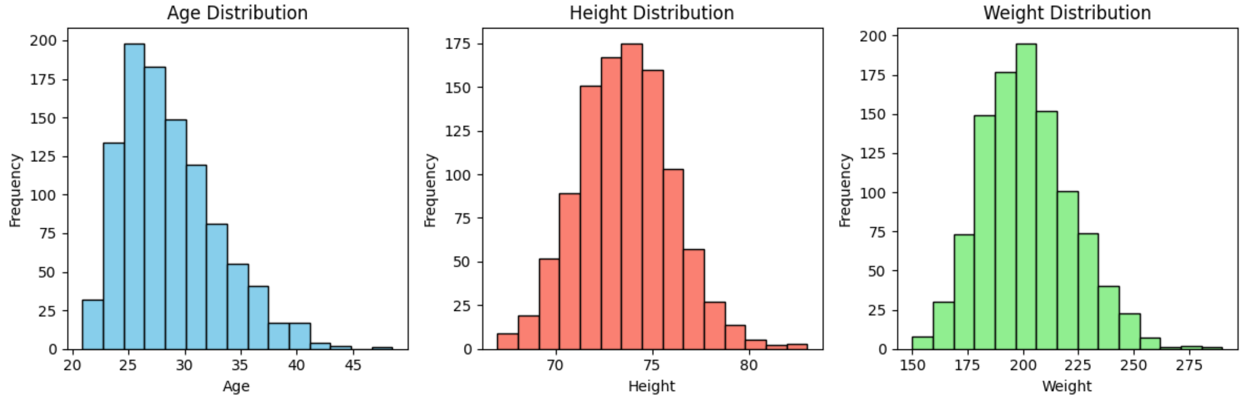


```
            Height       Weight          Age          BMI
count  1033.000000  1033.000000  1033.000000  1033.000000
mean     73.698935   201.689255    28.737648    26.090936
std       2.306330    20.991491     4.322298     2.300211
min      67.000000   150.000000    20.900000    19.496533
25%      72.000000   187.000000    25.440000    24.404880
50%      74.000000   200.000000    27.930000    26.085343
75%      75.000000   215.000000    31.240000    27.601351
max      83.000000   290.000000    48.520000    35.258488
```

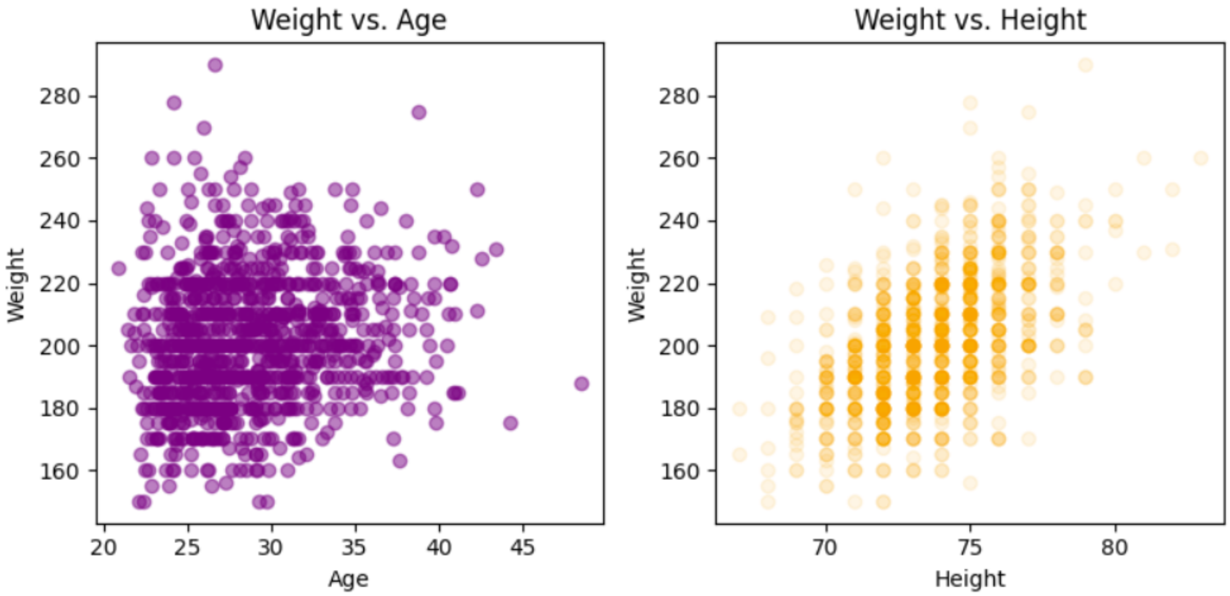
We also analyzed this on a per-team basis and did not notice any significant trends - the teams appear uniform. Here's the average height by team, for instance. We have to zoom into the top to see differences.



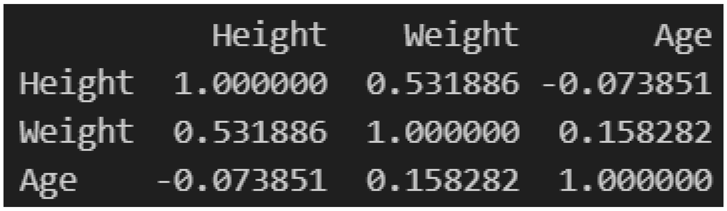
Height and weight appear to be normally distributed. Age leans towards the younger side.



There does not appear to be a correlation of weight with age, but there may be one with height.



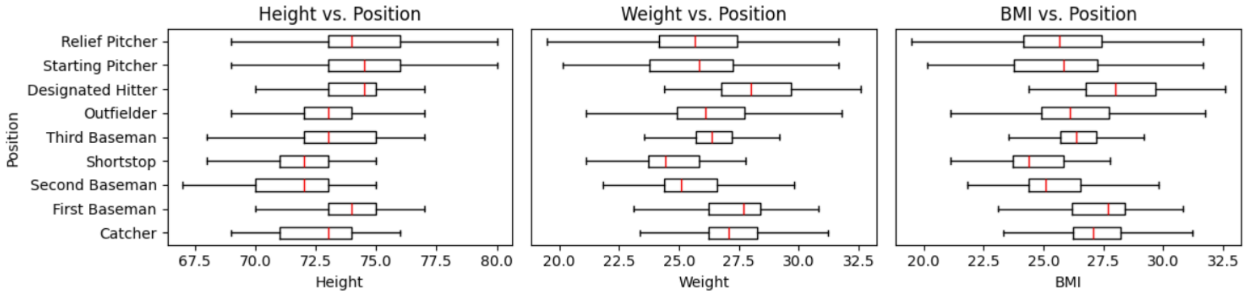
This is confirmed by a correlation matrix, only height and weight appear to be correlated, and not by much.



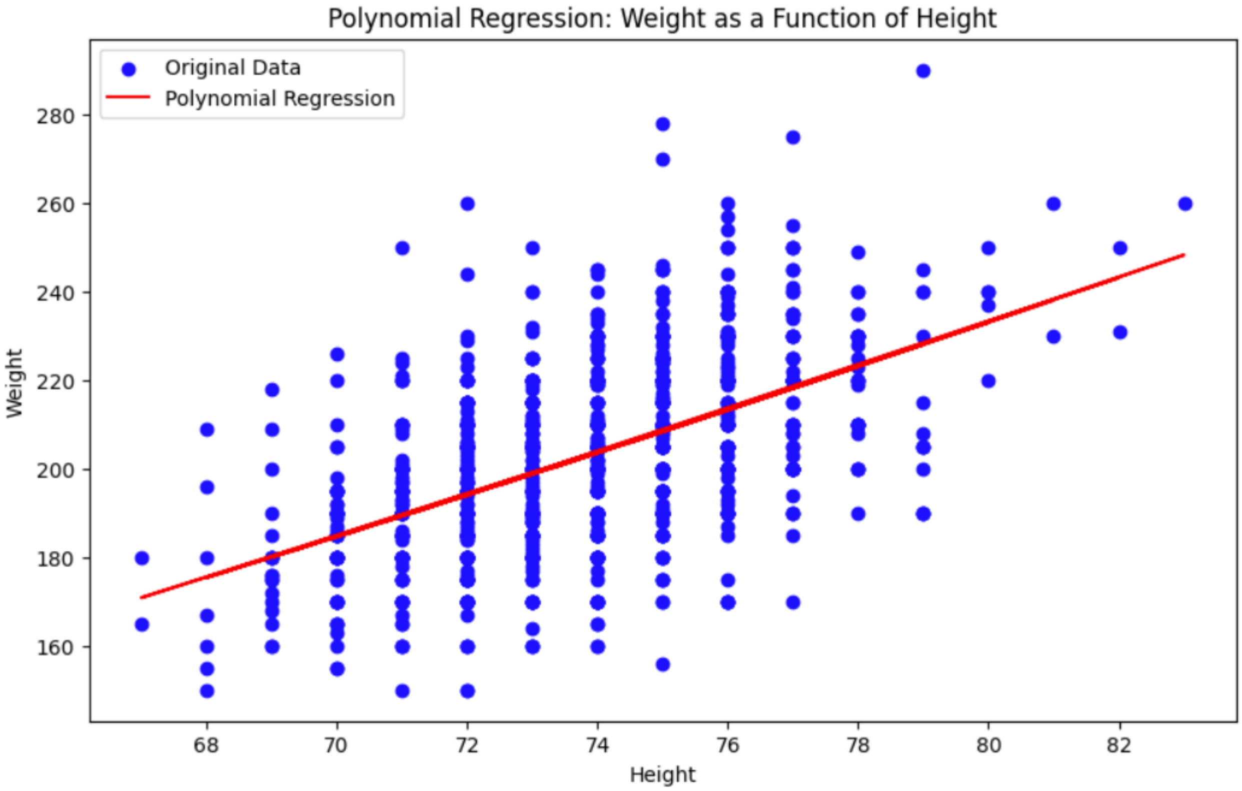
```
          Height    Weight       Age
Height  1.000000  0.531886 -0.073851
Weight  0.531886  1.000000  0.158282
Age    -0.073851  0.158282  1.000000
```

We notice a few things from these graphs below -

1. Shortstops and second basemen are shorter and weigh less than average.
2. Designated hitters have a relatively high BMI and also weigh the most.



The result of a polynomial regression:


Polynomial Regression: Weight as a Function of Height
Original Data
Polynomial Regression
Weight
Height