CS5010 Final Project Report - Group 7

Github Repo: https://github.com/UVA-CS5010/semester-project
John Hazelton (jch5nb) | Julie Crowe (ykd9ut) | Jake Kolessar (jak4as) | Allie Ridgway (ysx4gm)

Introduction

The purpose of this project is to examine whether there is a significant difference between the salaries of male National Basketball Association (NBA) players and female Women's National Basketball Association (WNBA) players. We wanted to determine the levels of different salaries between the two leagues, and inspect these differences based on their various factors. Such factors include player statistics, age, and league affiliation.

Data Web Scraping, Pre-processing, & Cleaning

With a group interest in both sports and the gender wage-gap, we decided to merge our interests to focus on salary and statistical data for the NBA and WNBA. We found this data significant in not only determining the presence or magnitude of a wage gap in basketball, but also the potential contributors, or differences between leagues that may influence it. The dataset was web scraped from two websites, Spotrac and Basketball Reference, containing player salary information and statistics data, respectively. Web Scraping the data allowed us to tailor the dataframe to the questions being asked and concatenating data from various sources gave us the most robust dataset to analyze.

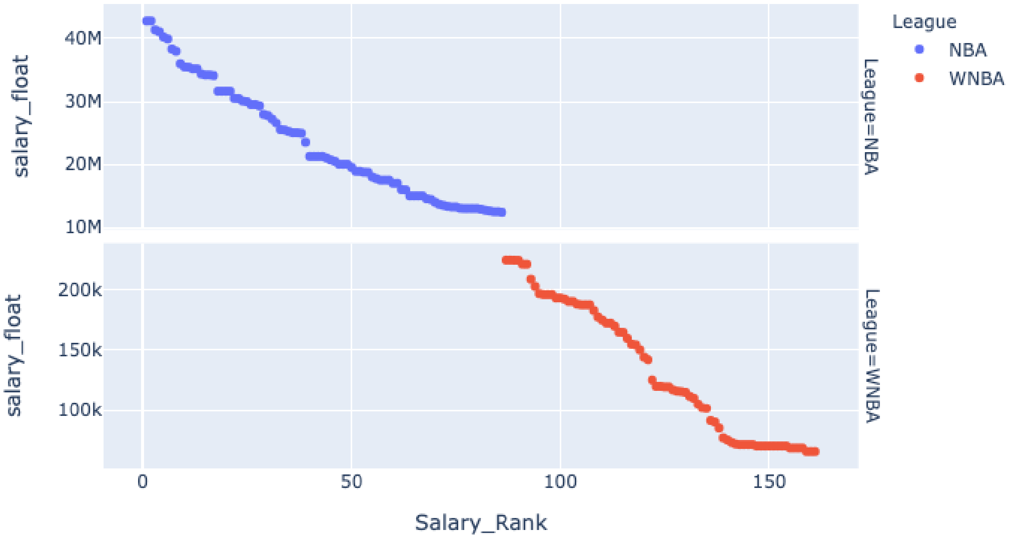
Pre-processing included several steps of aggregating datasets. We first performed a merge or join between salary and statistics dataframes. Upon doing so, we discovered that some international players had been excluded because their names contained accented letters in one dataset and did not in the other. We modified these characters by replacing them with non-accented, English equivalent characters in order for the code to process them correctly. To further clean the data, we utilized dropna(). Additionally, for players that switched teams mid-season, there were multiple records, one for each team as well as a cumulative entry. These duplicates had to be dropped, maintaining only the cumulative records for each player. Once all of the names in the tables were formatted correctly, the statistics and salary data frames were joined via the player name column. The final step involved concatenating the two large NBA and WNBA player datasets. To do so, however, we had to modify some column names and remove any erroneous or useless columns. After this process, the concatenation was essentially stacking one dataframe on top of the other.

Before moving on to our analysis, we decided to bolster our data with additional fields for percentage of games started (GS%) and salary rank. We manually calculated these columns using total games and games started for GS%, and sorting by average salary for the salary rank of each player. Finally we decided to add one final field, salary ratio, as a way of standardizing the widely different salary amounts between the NBA and WNBA. We divided players' average salaries by their respective league's total revenue, as a way of comparing the dollar amounts on an essentially level playing field. The final dataset consisted of 34 columns, some of which included: Player, G, GS, MP, FG, FGA, FG%, 3P, 3PA, 3P%, PTS, League, Team, Position, Age, Avg_Salary, salary_ratio, GS%, and Salary_Rank.

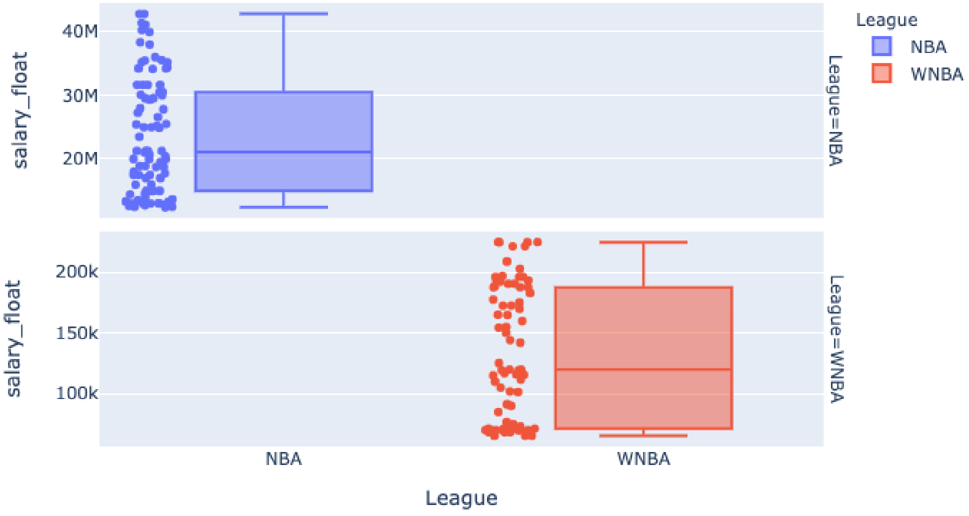
Exploratory Data Analysis

Once cleaning the dataset, we decided to first investigate the data through exploratory data analysis. The beginning plots of NBA vs. WNBA salaries demonstrated the large difference in wages (see below).

Salary Rank vs Salary by League

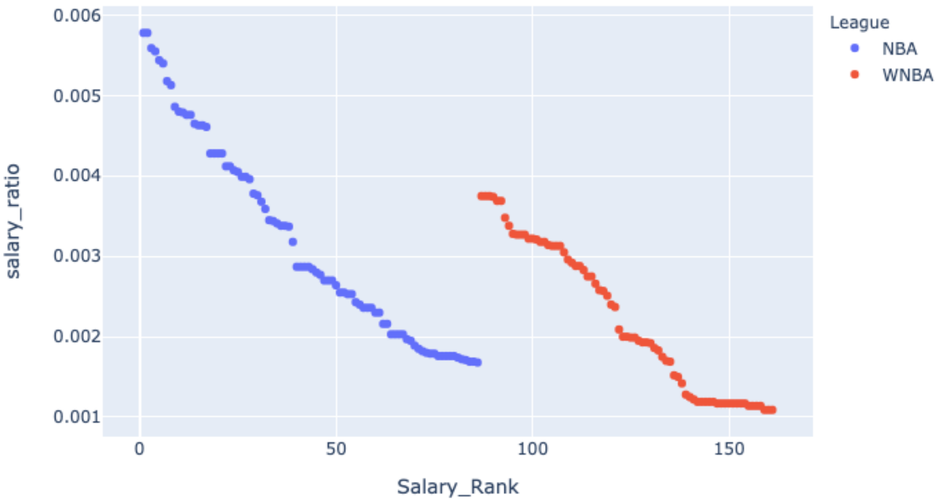


League vs Salary

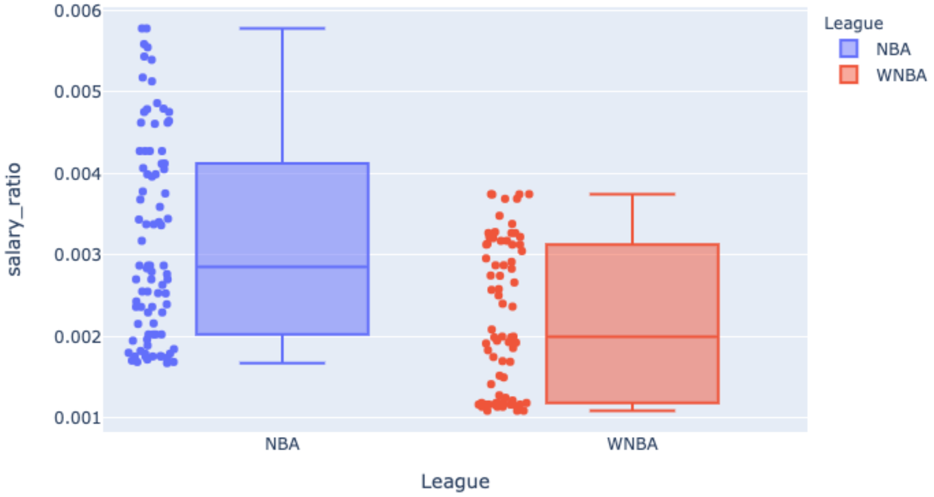


The plots for the two leagues differed to such an extent that the axes were completely unique. Therefore, as a way of more effectively comparing the differences between leagues, we utilized our previously calculated salary ratio statistic (player salary / league revenue). We found this method to be the most efficient way to scale the revenues, though further potential contributing factors are worth noting, such as ticket sales, league profit, or salary caps, which are the maximum amounts teams are allowed to spend on salary, as negotiated between the league and player associations every few years. With an understanding that these scaled salaries are not necessarily a perfect representation of equivalence between leagues, we trusted them in making observations and analyses between the leagues. In the scaled graphs (see below), we discovered that while salaries were now much more comparable between leagues, there was still a rather significant difference. WNBA salary ratios had a median of about 0.2, compared to a value of about 0.3 for NBA players. Furthermore, the spread appeared much greater for NBA players, with those at the top end of the graph making wildly higher portions of league revenue, while WNBA player salaries were grouped much closer together.

Salary Rank vs Salary Ratio by League



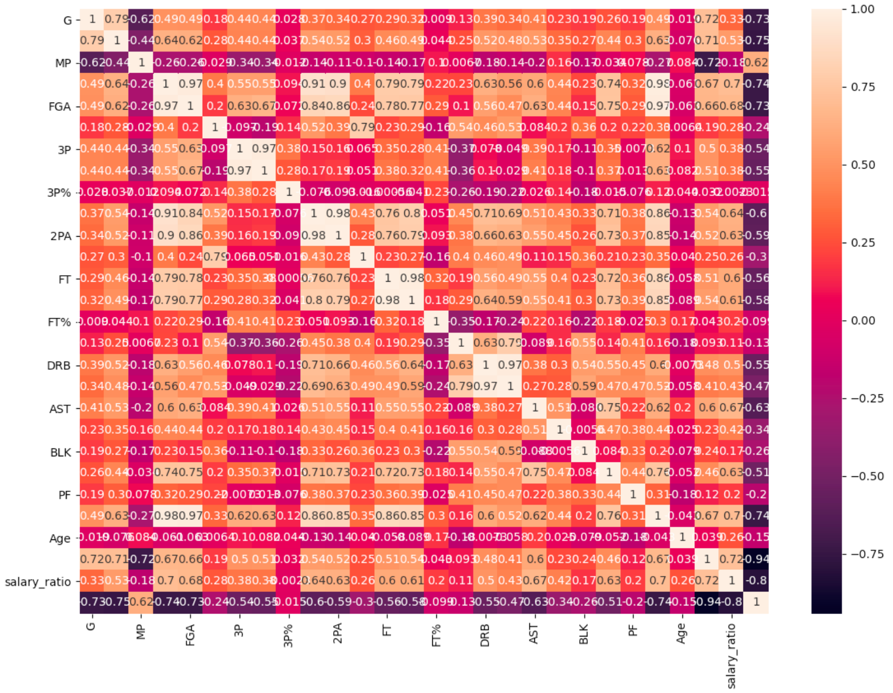
League vs Salary Ratio



Correlation Matrices

For our initial analysis, we chose to create a correlation matrix of each variable within the dataset. The purpose of this analysis was to examine which statistics might be valuable to consider as factors which contributed to salary, as well as statistics which had generally high correlation and might be used as predictors of a successful player. We chose to display the matrix as a heatmap in order to best display the correlations and provide a clear visual representation of the data.

Correlation Matrix of Salary & Statistics

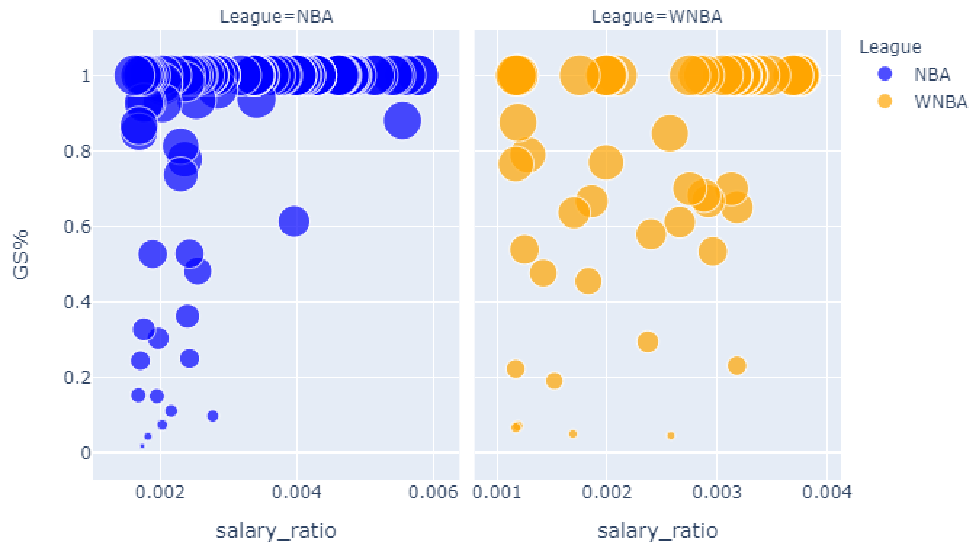


The heat map above shows a lighter color for strong positive correlation, an orange color for low to no correlation, and dark purple for strong negative correlation. Upon viewing the plot, we see that there appears to be an overall light color, meaning that there is generally a strong correlation between the statistics. Specifically, we see light colors throughout the rows of Field Goals, 2-Point Goals and Attempts, Defensive and Total Rebounds, Turnovers, Points, and Salary. We see a strong negative correlation between Minutes Played and Salary Rank. Lastly, there does not appear to be any correlation between 3-Point Goals, Personal Fouls, and Age among all other variables. From the strong positive and negative correlations, we see that there is a relationship between most of the variables within the dataset. Specifically, we may conclude that these game statistics are overall a strong predictor of salary, and the salary of a player is dependent on their performance in games. We may conclude that the ranking of a players salary among the league is directly related to every statistic except for the amount they play per game. We also conclude that age does not have an impact on the statistics of a player, and further assume that players are not evaluated based on age in the league unless their statistics begin to fall.

Salary vs Percentage of Games Started

For our second analysis, we compared salaries with games started, and visualized the differences between the two leagues. In the graph (see below), it is clear that the vast majority of players making higher salaries in the NBA were starting all or nearly all of their games, whereas in the WNBA, there was a greater mix of high salaried players starting only a fraction of their games. This visual difference could simply be indicative of how teams structure their lineups differently between the two leagues, or could suggest one league values a starting spot higher. We later sought out to see if this difference would reveal itself as a key significant predictor in our model building process.

Salary Ratio vs Percentage of Games Started (GS%)

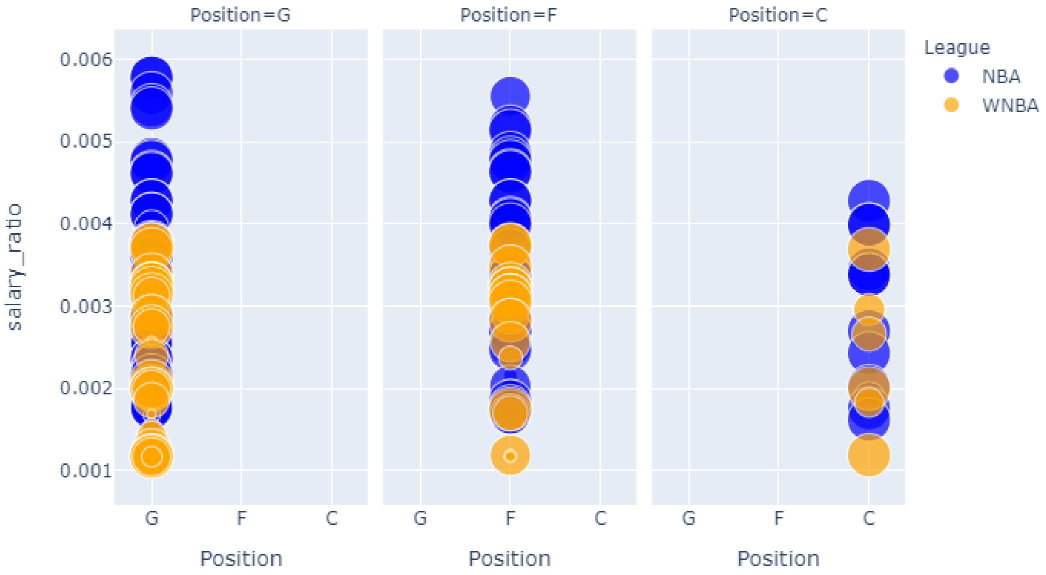


Salary vs Player Position

Our third analysis investigates the effects of player position on their salary. In the first graph (see below), we looked at player position against salary ratio, while still accounting for the percentage of games started (denoted by the size of a data point's bubble). Interestingly, we found a noticeable difference in salaries between positions in the NBA, with guards and forwards reaching much higher salary ratios and than any of their center position counterparts. In the WNBA, however, the data points appeared to be distributed evenly across different positions.

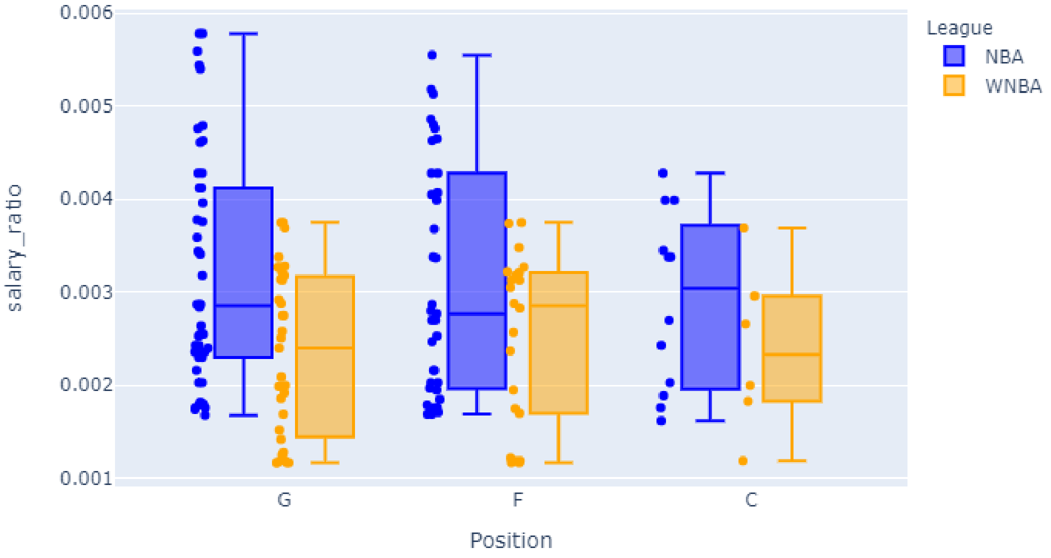
Again, we kept this potential difference in mind as we considered significant predictors later in our model building process.

Position vs Salary by League



When looking at the data in boxplot form (see below), we noticed that the median salaries for centers in the NBA were actually higher than the other positions. It seems only the star guard and forward players, making the highest salaries, were pulling up the averages for those positions, as no centers had salary ratios above 0.005. In the WNBA, median salaries for forwards proved to be much higher than those of guards or centers.

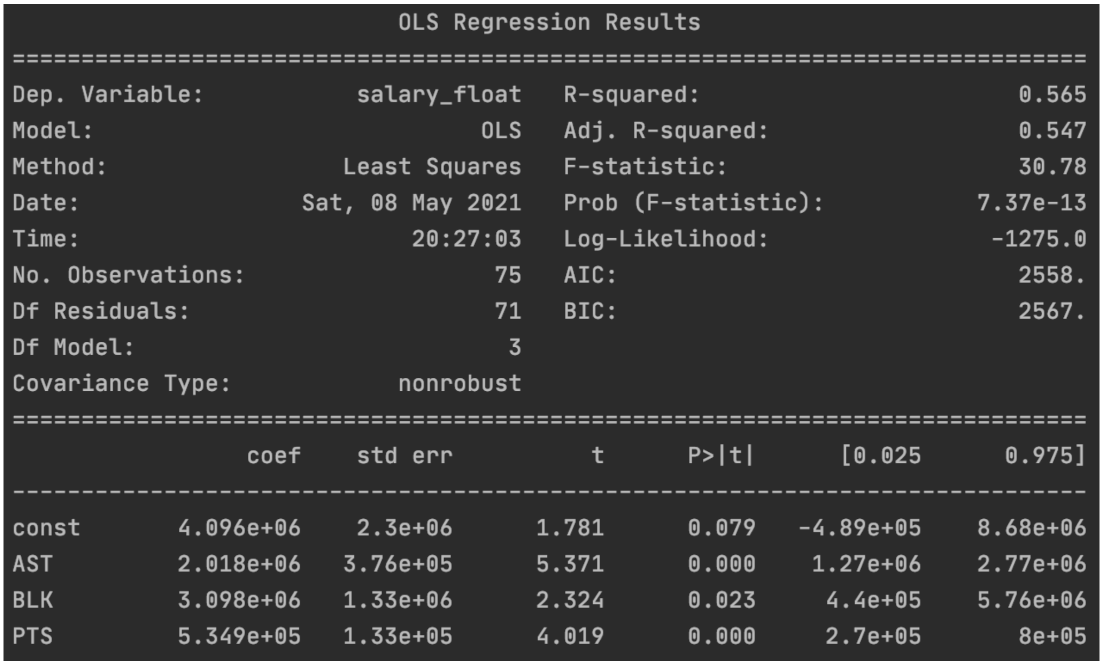
Salary Ratio vs Percentage of Games Started (GS%)



Model Building: Predicting Salaries from Stats

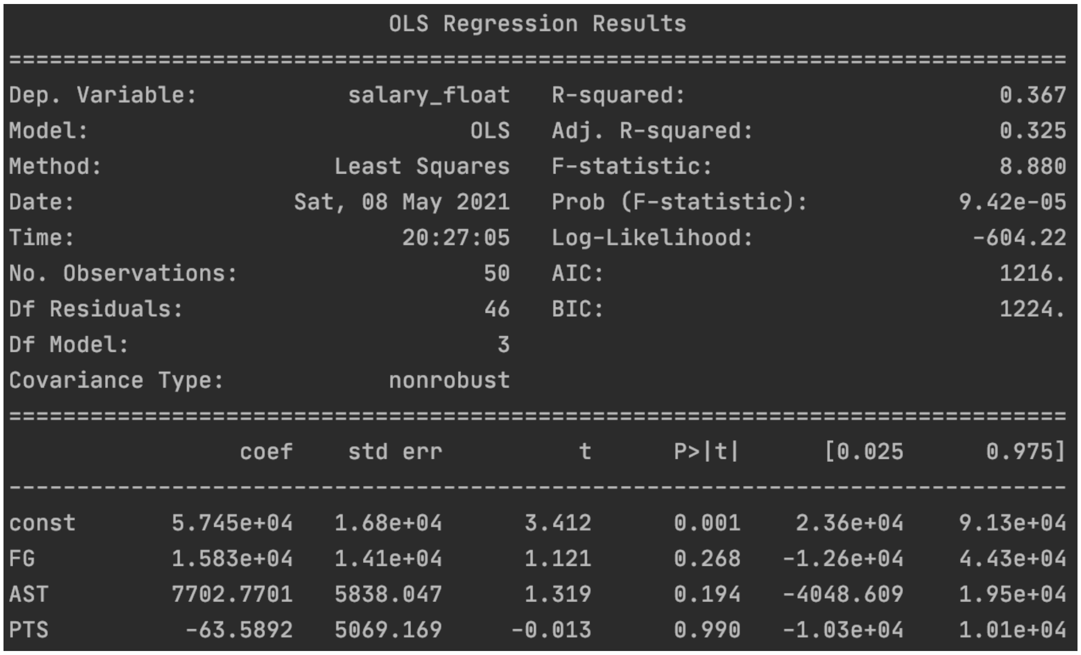
Our fourth and final analysis was to build a statistical model. When setting out to build a predictive model for player salaries based on statistics, we decided to utilize object oriented programming in conjunction with Python's Sklearn module, featuring a powerful tool for predictive data analysis. We first split our data into training and test datasets, and then performed forward and backward selection techniques to arrive at a multiple linear regression model. We repeated this process twice, once for each league. Upon calculating R-squared values for the forward and backward models, we chose the models with the best R-squared, 0.565 for the NBA, and 0.367 for the WNBA. Then, we looked at summary statistics for the chosen models and decided to drop all but three of the predictors, as the others were insignificant. As a result, we arrived at a model with assists (AST), blocks (BLK), and points (PTS) as predictors for NBA salary, and a model with field goals (FG), assists (AST), and points (PTS) as predictors for WNBA salary (see models below). With these results, it seems the WNBA places a higher value on overall defensive output, whereas the NBA values more all rounded offensive players, with a potential favoring of defensive-minded forwards and centers that are more likely to get blocks.

NBA output:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            salary_float   R-squared:                       0.565
Model:                             OLS   Adj. R-squared:                  0.547
Method:                  Least Squares   F-statistic:                     30.78
Date:                 Sat, 08 May 2021   Prob (F-statistic):           7.37e-13
Time:                         20:27:03   Log-Likelihood:                -1275.0
No. Observations:                   75   AIC:                             2558.
Df Residuals:                       71   BIC:                             2567.
Df Model:                            3
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       4.096e+06    2.3e+06      1.781      0.079   -4.89e+05    8.68e+06
AST         2.018e+06   3.76e+05      5.371      0.000    1.27e+06    2.77e+06
BLK         3.098e+06   1.33e+06      2.324      0.023     4.4e+05    5.76e+06
PTS         5.349e+05   1.33e+05      4.019      0.000     2.7e+05       8e+05
```

WNBA Output:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           salary_float   R-squared:                       0.367
Model:                            OLS   Adj. R-squared:                  0.325
Method:                 Least Squares   F-statistic:                     8.880
Date:                Sat, 08 May 2021   Prob (F-statistic):           9.42e-05
Time:                        20:27:05   Log-Likelihood:                -604.22
No. Observations:                  50   AIC:                             1216.
Df Residuals:                      46   BIC:                             1224.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       5.745e+04   1.68e+04      3.412      0.001    2.36e+04    9.13e+04
FG          1.583e+04   1.41e+04      1.121      0.268   -1.26e+04    4.43e+04
AST         7702.7701   5838.047      1.319      0.194   -4048.609    1.95e+04
PTS          -63.5892   5069.169     -0.013      0.990   -1.03e+04    1.01e+04
```

Multiple Linear Regression Model:

NBA: salary = 4,096,000 + 2,018,000*AST + 3,098,000*BLK + 534,900*PTS
WNBA: salary = 57,450 + 15,830*FG + 7703*AST - 63.59*PTS

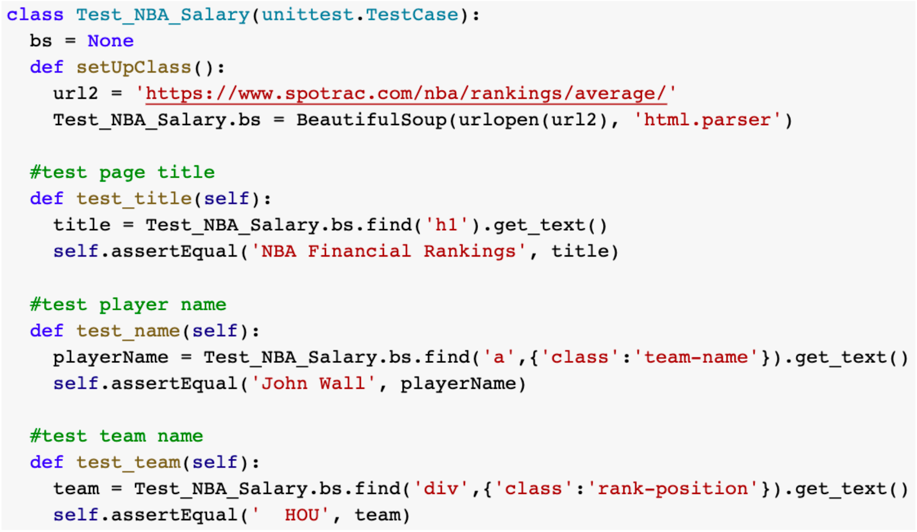
Based on the output, it appears that it may be harder to predict the WNBA salaries based on the statistics of the players. The WNBA model did not have a strong R-squared value and the P-values for the predictors do not meet the criteria for significance. The NBA model produces a stronger R-Squared value and significant P-values. The group concluded that only the NBA model would be useful for predicting salaries. This could be because the WNBA has less variation in salary range from the lowest paid players to the highest paid players. Another reason may be that the league uses different criteria to determine pay rate such as years of experience in the league. From here, a further investigation into what predicts WNBA salaries would be necessary although it was outside the scope of our groups project.

Model Legitimacy: Undervaluing & Overvaluing Players

After producing our models we decided to further apply them, by analyzing the most overvalued and undervalued players based on their predicted salaries. To do so, we added predicted salary and salary residual columns to our dataframe, manually calculated based on our model equations for each league. We then ranked players, ascending and descending by salary residual, to obtain the lists of most undervalued and overvalued players, respectively. With some knowledge and following of NBA players, we decided to test the legitimacy of this ranking by looking at the top three most undervalued players in the NBA based on our model: Giannis Antetokounmpo, Zach LaVine, and Marcus Smart. These results seemed somewhat reasonable, considering Giannis is a multiple-time MVP, LaVine is a young, budding All-Star on a relatively small contract, and Marcus Smart is a quality-level player on one of the smaller contracts for his age and statistics.

Unit Testing

In order to test the accuracy of our dataset, we performed unit testing on the data gathered from the Web Scraping. As previously described, we used separate sources for both salary data and statistical data among the NBA and WNBA. To thoroughly test these sources, we performed a unit test on each column of data using assertEqual() and tested that the data we were pulling was the desired data from the websites. To begin each class of unit tests, we tested the title of each page. For salary data, the unit tests performed were on player name, team name, player position, player age, and salary. The source for the statistics consisted of one large table, so we tested elements in the table such as player name and position. The code below displays a portion of the testing performed on NBA salary data, including the test for page title, player name, and team name for the first player listed on the website: John Wall.



```
class Test_NBA_Salary(unittest.TestCase):
  bs = None
  def setUpClass():
    url2 = 'https://www.spotrac.com/nba/rankings/average/'
    Test_NBA_Salary.bs = BeautifulSoup(urlopen(url2), 'html.parser')

  #test page title
  def test_title(self):
    title = Test_NBA_Salary.bs.find('h1').get_text()
    self.assertEqual('NBA Financial Rankings', title)

  #test player name
  def test_name(self):
    playerName = Test_NBA_Salary.bs.find('a',{'class':'team-name'}).get_text()
    self.assertEqual('John Wall', playerName)

  #test team name
  def test_team(self):
    team = Test_NBA_Salary.bs.find('div',{'class':'rank-position'}).get_text()
    self.assertEqual('  HOU', team)
```

Project Management

After studying agile development and Scrum project management principles, the team decided to follow this approach for the remainder of the project. We developed user stories which aligned with our original goals from the project proposal as this helped to better organize our vision for the project and the steps we would take to fully complete it. Utilizing shared documents, we maintained a backlog of the tasks which needed to be completed and assigned them to team members for the most efficient completion time. The group also met frequently and aimed to structure meetings as agile standups: we first reviewed successes and completed tasks, then reviewed areas of concern or issues we were facing. Not only did this make our development process more efficient, but it bonded the group and increased our meeting effectiveness. We collaborated with many aspects of the project and overall worked together to meet project deadlines.

Extra Credit

For an extra credit opportunity, the group chose to web scrape the dataset. We wanted to challenge ourselves by performing this additional step, and wanted a complete dataset with both salary data and statistical data in order to answer the questions we had about the different leagues. Web scraping the data and combining data from several sources gave us the best dataset we could utilize in order to do so, and allowed us to efficiently analyze the information about basketball leagues and salary relationships.

Peer Feedback

Following the in class presentation, one of the questions asked was whether there was a discrepancy in salary between NBA and WNBA players who had the same statistics otherwise. Unfortunately, there is not a clear answer to this question because the statistics of players between the leagues differs due to differences in game length, physique, and other game rules. Because of this, there are no equivalent NBA and WNBA players with identical statistics that we would be able to hold constant while comparing salaries. This is why we chose to utilize a salary ratio and make comparisons between the leagues with equivalent numbers.

Conclusion

In summary, we can confirm through our analysis that there is a large difference between NBA and WNBA salaries and many contributing factors. By performing exploratory data analysis, we found through examining the salary ratio of players versus the overall league revenue that the WNBA salaries had a lower ratio than the NBA, meaning that the average male player makes a greater percentage of the league's revenue than a female player. Looking at the spread of the salaries in both leagues, this is likely due to the extremely high salaries of top NBA players as compared to the more even small spread of WNBA players. From analysis of the correlation matrix, we saw that nearly every game statistic is highly correlated with the salary of players. In conjunction with the above conclusion regarding salary spread, we can see that top female players should in fact be paid higher amounts for their top performance. After comparing salary with the percentage of games started between both leagues, we determined that top paid NBA players start nearly all of their games while top paid WNBA players do not. Upon examining the salaries between positions of players between the leagues, we again saw that specific guards and forwards, in addition to all centers had high salary ratios, but WNBA players maintained an even spread of a lower ratio except for high paid forwards. Lastly, by utilizing the statistical model we created, gained insights on the statistics of overvalued and undervalued players in both leagues. These results could be used by teams when determining fair salaries for new contracts. They could especially be used by female players looking to pursue the option of negotiating higher salaries. Lastly, they could be viewed by fans interested in learning more about their favorite players and the effects of the players' statistics on salary.

The project has potential for further expansion in many areas. The most significant of these would be continuing to expand other monetary data sources between the leagues. Specifically, ticket prices and sales for different teams, overall league profit, and the salary caps of each team. We found several surprising facts from our analysis, such as the percentage of top paid NBA players starting nearly every one of their games while this was not the case for the WNBA. Given more time, we would like to have further expanded and improved our research and code to make more definitive conclusions about the reasoning for this. Outside of basketball, we would also like to examine how this analysis could be applicable to other gender wage gaps within different industries.