

Exploring Prejudice and Salary Prediction in the NBA Through a Statistical Model - Final Report

Author: Berkeley Reynolds

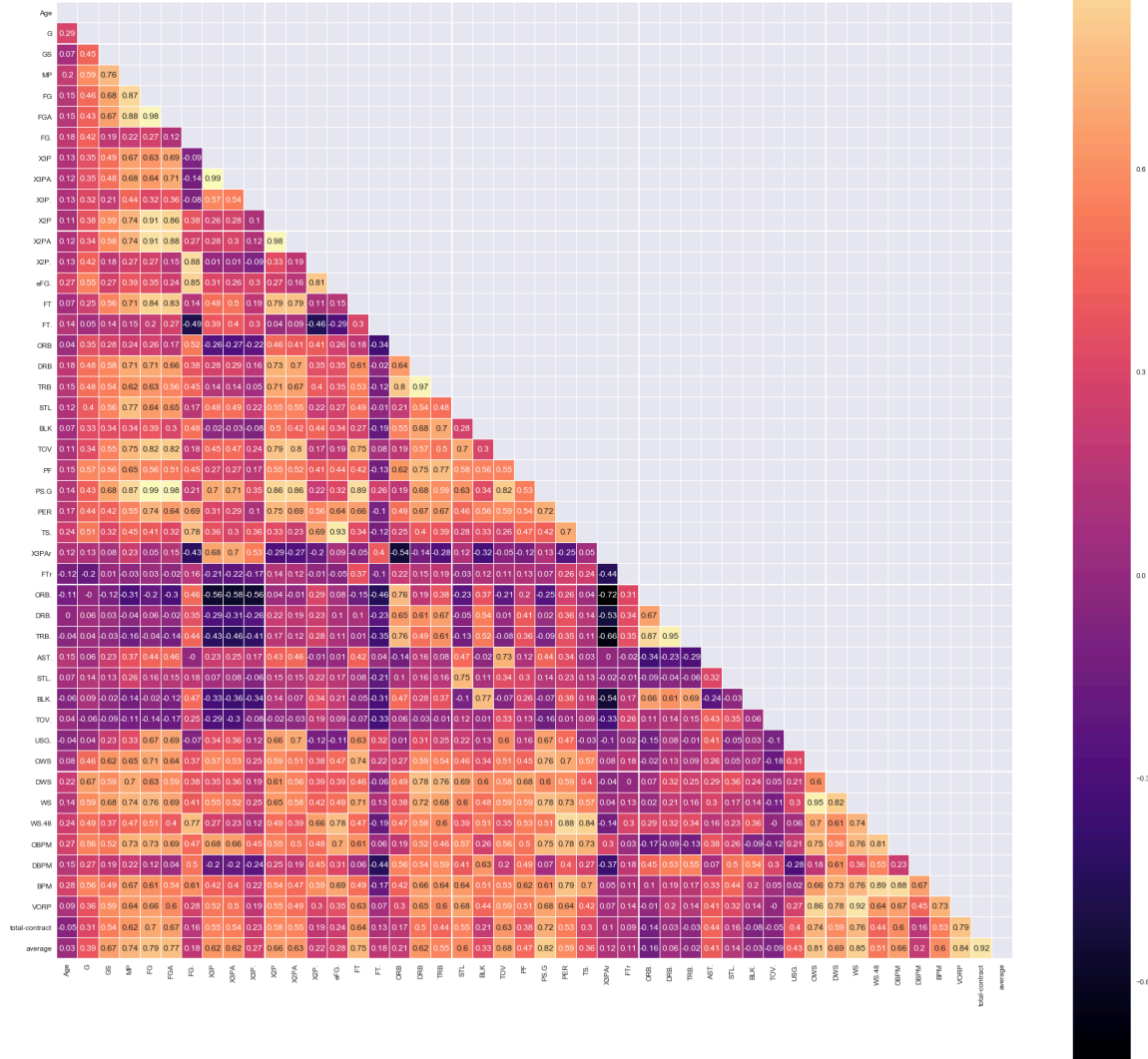
Introduction and Related Work:

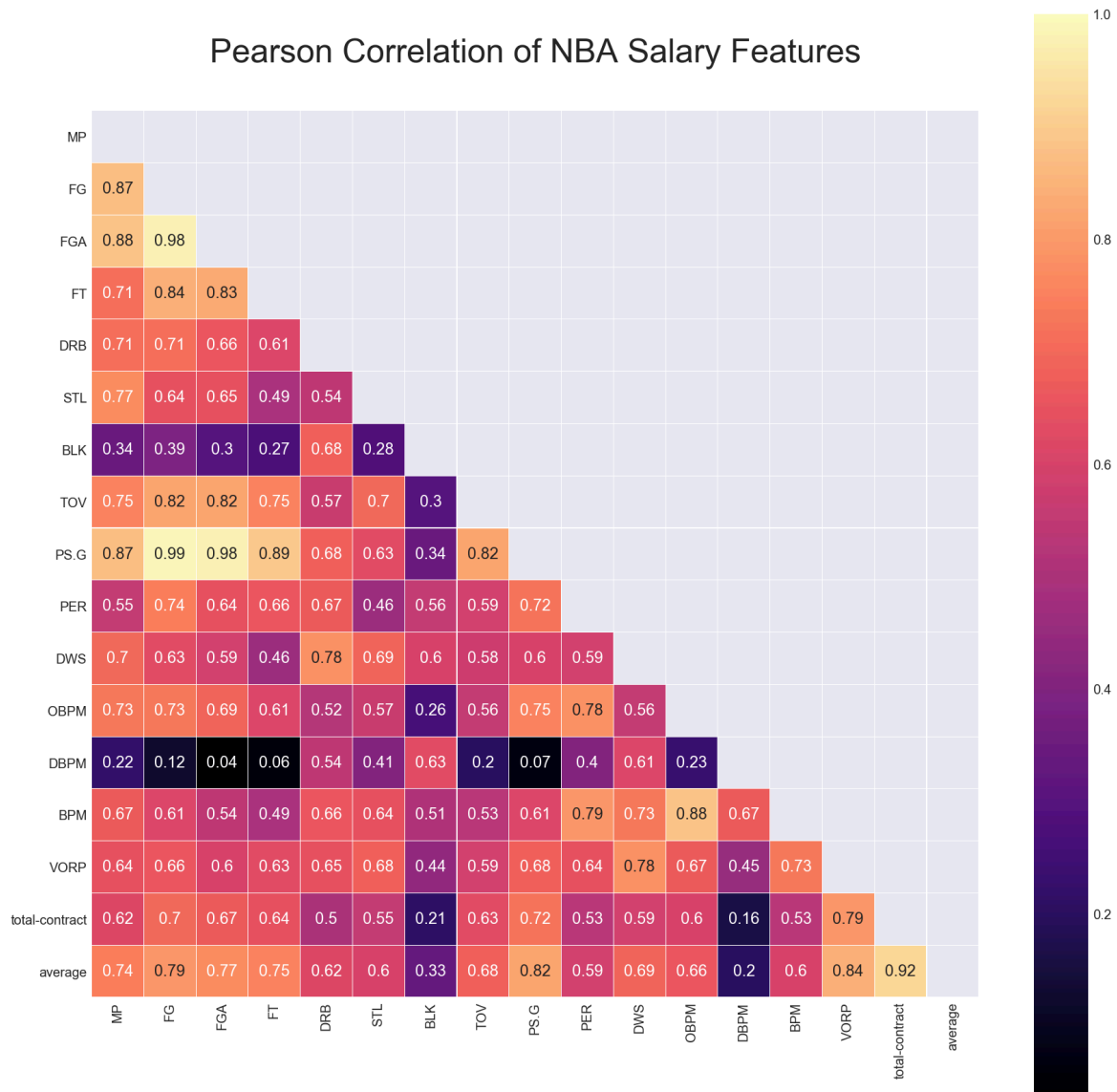
Introduction of New Techniques:

In this phase of the project, I plan to build upon the work I completed in the first quarter of senior project (CSC 491). I plan to explore some new techniques, methodologies, and algorithms in this phase. One new technique I will implement is the scraping of data from a website. I plan to write XPath queries to scrape data about players' nationalities and 2018 salaries from basketballreference.com. I learned a bit about XPath queries in the database courses I have taken at Cal Poly, but it is still a relatively new technique for me, so I plan to research this topic more to figure out how to write the correct queries. I also plan to write code in R or Python to utilize join operations to add this new data to the existing data from the first phase of the project. I have also gained a bit of familiarity with this topic in the database courses I have taken, but it is still a new technology that I will have to learn more about as well in order to execute the task. Furthermore, I plan to utilize new technologies in ggplot or matplotlib to create visualizations to display the new discoveries that are found in the analysis conducted in the second phase of this project.

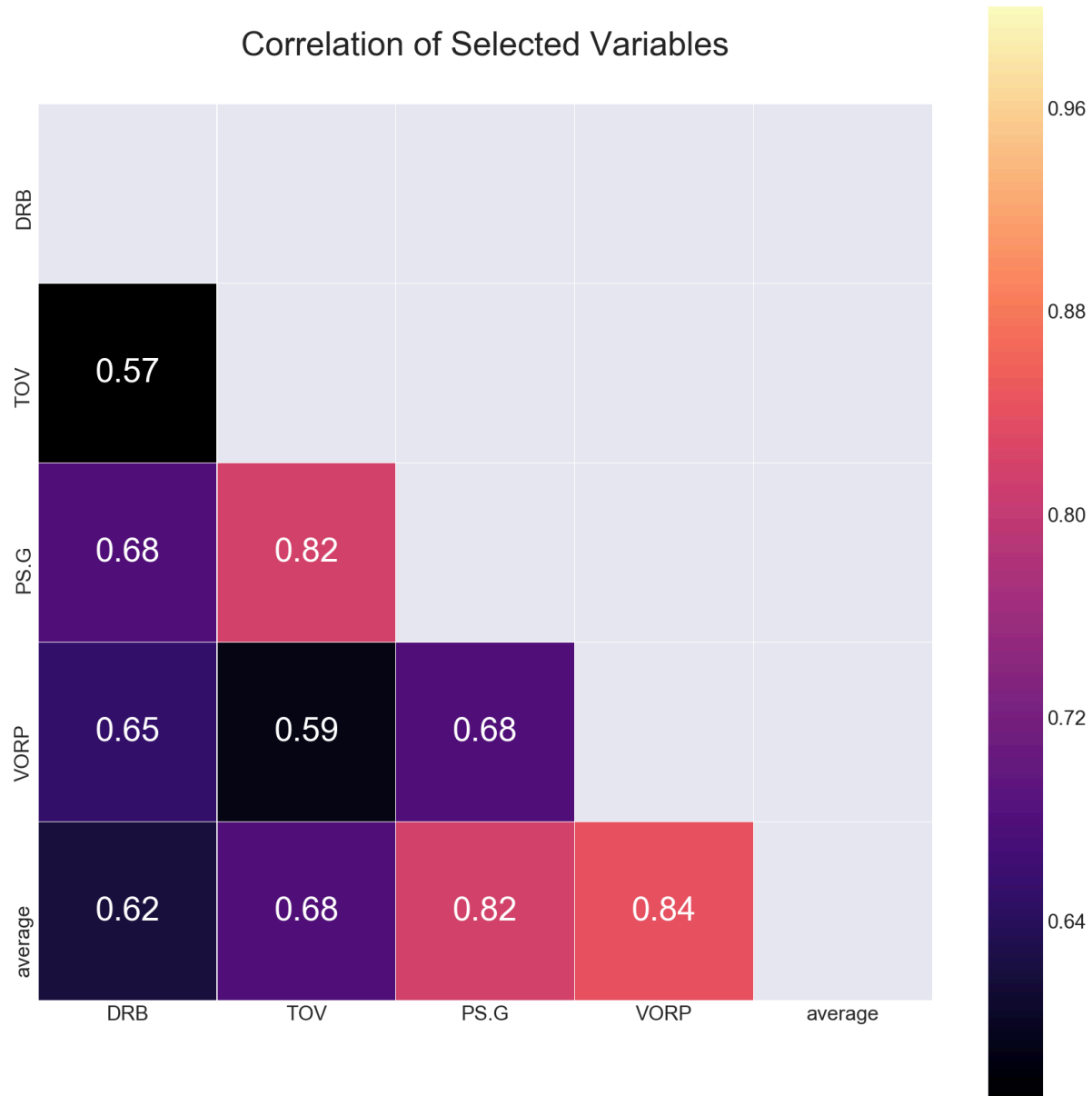
Differences from Previous Study:

In the second phase of this project, I plan to build upon the work completed in the first phase. So, I am not completing a brand new project, but rather extending the lengths of the previously done work. In the previous phase, I explored a model that essentially predicts the salary of upcoming NBA free agents using data about players' statistics in previous NBA seasons. The original model used Spotrac's online database to both generate yearly lists of NBA free agents and collect data pertaining to contract signings within each year. For player statistics, the model used seasonal data collected by Basketball Reference, evaluating all metrics listed in the Per-Game, Per 36 Minutes, Per 100 Possessions, and Advanced data tables under each relevant season. Furthermore, the model also referenced Spotrac's

[illegible]

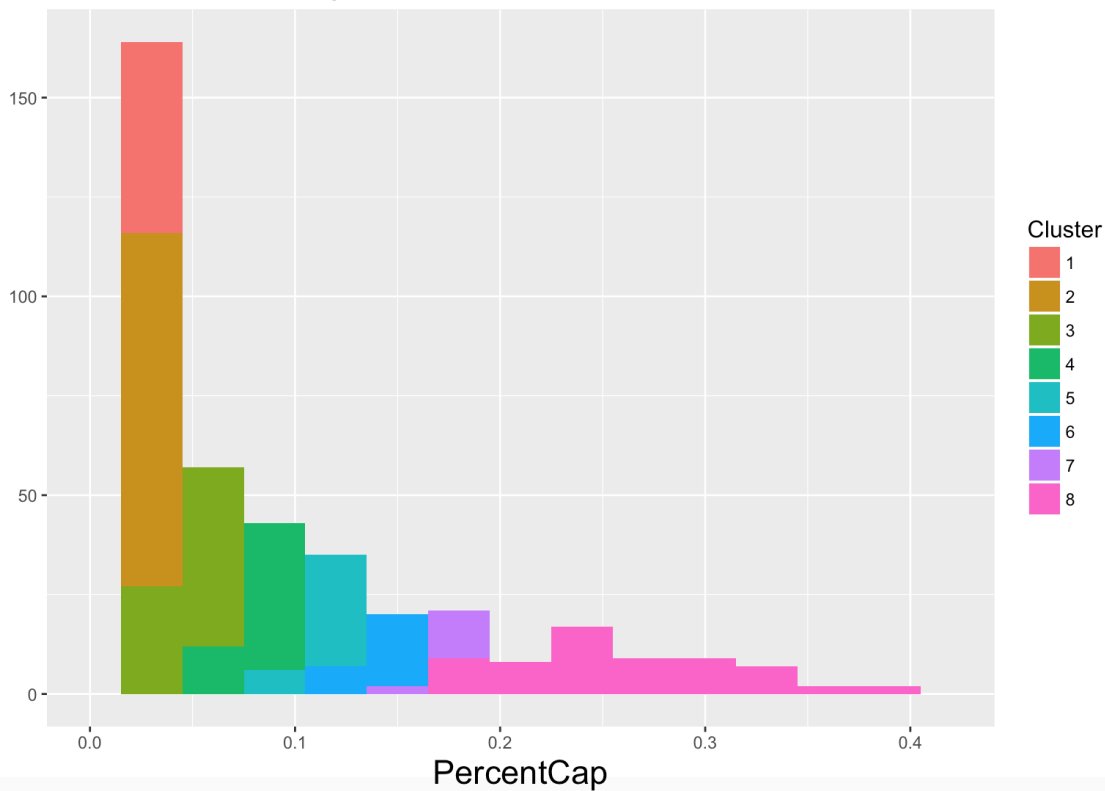


After doing one more analysis of this correlation table, the statistics were narrowed down once again, resulting in this final correlation table below:



After this, each of the four variables were standardized using the respective means and standard deviations among the free agent class of each same year. Afterwards, the salaries were grouped into clusters that were formed to encapsulate several of the salary standards within the NBA. The clusters' bounds are based off of the distribution of percent-cap shown below:

How Percent-cap is Divided



The bounds are narrower at lower percentages, distinguishing between several different contract exceptions and minimum salary rates. They are wider among higher salaries, to account for their higher variability.

After the standardization and clustering process, the dataset was used to train a K-Nearest-Neighbors classification model. An 80-20 training-test split, done through simple random selection, was used. After training, the model was used to predict the salaries of upcoming NBA free agents in 2018, based off of player statistics as of January 11, 2018. Classification was performed on the test set using the 17 nearest-neighbors in unweighted Euclidean distance between the four variables, allowing for a conservative estimate at a reasonable error size.

One of the biggest differences I plan to carry out in the second phase of this project is the addition of nationality into the predictive model. I plan to collect data about the nationality of each player in the existing dataset. I then want to use data analyzation and visualization techniques to explore the discrepancy in how players of different nationalities are paid. I want to explore if players of certain

nationalities are being under or over paid by comparing their true salary with the salary that the model predicts. Another addition I want to explore in the second phase of this project is how accurate the model is in its ability to predict players' salaries in hindsight. The model from the first phase of the project used data from the NBA seasons leading up to 2018 to predict the salaries of upcoming 2018 NBA free agents. Now that we are in 2024, I wish to collect data about how much each 2018 NBA free agent actually ended up getting paid during the 2018 offseason. I then want to compare these real salaries with the predicted salaries to analyze the accuracy of the predictive model. This will give great insight to how well the model performs and whether the model could be improved by tweaking different metrics in the model. If time permits, I also wish to explore the tuning of this model to create the most accurate predictive model possible. This will give insight to what statistics and metrics are most effective in predicting player salaries. Another addition that I would love to add if time permits is adding the data from all the NBA seasons since 2018 to evaluate how the model performs with the current landscape of the NBA. The game of basketball and the NBA have been changing rapidly, so it would be very interesting to see how the model performs with current NBA data.

Purpose:

The purpose for the additions to this project are to further explore the capabilities of the model that predicts the salaries of NBA players. First of all, I want to incorporate the nationalities of players into the predictive model. The purpose of this is to explore the discrepancy between how NBA players of different nationalities are paid. The history of humanity has been plagued with discrimination. Even though we have made great progress and big strides in reducing discrimination, it still exists in many places around the world. People all over the world are over or underpaid based on their ethnicity or nationality. Particularly in the workplace, people are discriminated against based on these traits. People often think about NBA players as people who are simply paid to put a basketball through the hoop. However, these players are people just like the rest of us. So, I believe it is possible that this discrimination takes place in the NBA as well. Thus, I plan to investigate this. I want to see if there is a discrepancy between nationalities, and if there is, I want to explore which nationalities are being

discriminated against the most. I think this is a very important topic, and if there truly is a discrepancy, then I believe the world needs to know about this, so we can collectively fix the problem. I also want to analyze how well the predictive model performs by comparing the results of the model with the true salaries that players ended up receiving in the 2018 NBA offseason. The purpose of this is to, first of all, determine how accurate the model is. Furthermore, it could provide insight as to which players were under or over paid and why this was the case. I wish to look for certain organizations that are consistently over or under paying players. I hypothesize that it may be the case that teams in cities that players are more likely to want to live in, such as Los Angeles, Miami, and New York City, are able to underpay players and get away with it. On the contrary, teams in cities that may be less desirable, such as Oklahoma City and Indianapolis, may have to overpay players to get them to join their team. I think this would be beneficial to explore because it may give extra insight as to why certain players are being under or over paid. With this knowledge, the model could be tuned more to make even more accurate predictions.

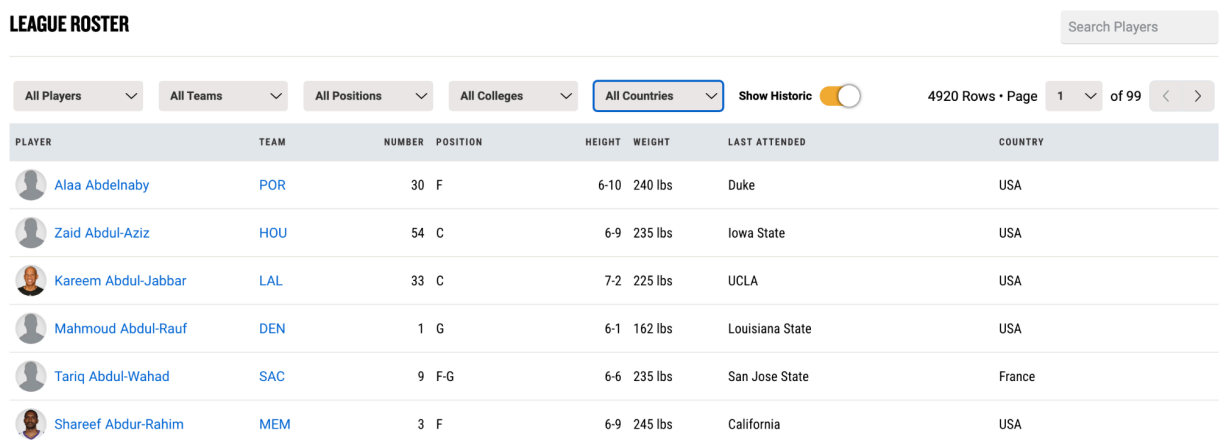
Methodology:

The approach I will take to tackle this problem will begin with building upon the work I completed in the first part of my senior project. I plan to explore and use some new tools, technologies, algorithms, and methodologies in this part of the project. I will use XPath queries to extract data regarding players' nationalities and salaries from basketballreference.com. This is an example of the use of a new technology or tool. I was exposed to this topic in a database design and implementation course that I took at Cal Poly, and I am excited to continue my exploration of this technology in this phase of the project. I will then use R code to execute the joining of this new data into the existing datasets. I have been exposed to some data science topics within my curriculum at Cal Poly, but I had not previously used R until now. I have been gaining some familiarity with R and its plethora of available packages through a statistics class I am currently taking at Cal Poly. I will continue to grow and exercise my skills in R through this phase of the project. Another new tool I will use is the dplyr package of R. This package is useful for all kinds of data transformation, manipulation, and summarization. I will use this to make various changes to the dataset that will allow for more useful and

purposeful analysis. I will also be utilizing new technologies through the use of the ggplot2 package of R. This package arms you with an array of different methods that are useful for creating data visualizations. It provides the necessary functions to construct any graph, plot, or visualization to explore and present findings in a dataset.

Implementation:

In order to explore the difference in how NBA players coming from different countries are compensated, the first step I had to take was adding into the dataset the nationality of each player in the dataset. The NBA has put together a great resource that provides, among other things, the birth country of every player who has ever played in the league. This data can be accessed on the NBA's main website. Below is an image from their website, showcasing that the country of each player is included.



The screenshot shows the NBA League Roster page. At the top, there's a search bar and filters for All Players, All Teams, All Positions, All Colleges, and All Countries (which is highlighted). There's also a 'Show Historic' toggle and a pagination indicator showing 4920 Rows, Page 1 of 99. Below the filters is a table with the following columns: PLAYER, TEAM, NUMBER, POSITION, HEIGHT, WEIGHT, LAST ATTENDED, and COUNTRY. The table lists six players: Alaa Abdelnaby (POR, 30, F, 6-10, 240 lbs, Duke, USA), Zaid Abdul-Aziz (HOU, 54, C, 6-9, 235 lbs, Iowa State, USA), Kareem Abdul-Jabbar (LAL, 33, C, 7-2, 225 lbs, UCLA, USA), Mahmoud Abdul-Rauf (DEN, 1, G, 6-1, 162 lbs, Louisiana State, USA), Tariq Abdul-Wahad (SAC, 9, F-G, 6-6, 235 lbs, San Jose State, France), and Shareef Abdur-Rahim (MEM, 3, F, 6-9, 245 lbs, California, USA).

PLAYER	TEAM	NUMBER	POSITION	HEIGHT	WEIGHT	LAST ATTENDED	COUNTRY
Alaa Abdelnaby	POR	30	F	6-10	240 lbs	Duke	USA
Zaid Abdul-Aziz	HOU	54	C	6-9	235 lbs	Iowa State	USA
Kareem Abdul-Jabbar	LAL	33	C	7-2	225 lbs	UCLA	USA
Mahmoud Abdul-Rauf	DEN	1	G	6-1	162 lbs	Louisiana State	USA
Tariq Abdul-Wahad	SAC	9	F-G	6-6	235 lbs	San Jose State	France
Shareef Abdur-Rahim	MEM	3	F	6-9	245 lbs	California	USA

So, by using this resource provided by the NBA, I was able to obtain the nationality of each player in the dataset of 2018 NBA free agents. I then created and filled out a new column in the dataset to hold this data regarding the nationality of each player. Besides a possible discrepancy between the compensation of players that come from different nations, I also wanted to explore the accuracy of the model in hindsight. The original model was used to make predictions for how much money each upcoming 2018 NBA free agent would earn. So, being that it is now 2024, I was able to obtain data about how much money each of those free agents really ended up obtaining. For this data, I utilized Spotrac's databases. Spotrac provides the full salary history for every player that

has played in the NBA. So, I used this resource to obtain the actual 2018 salary obtained for each of the 2018 upcoming free agents in the dataset. I then added another new column to hold this data regarding the actual 2018 salary obtained from these free agents. Below, I have included a look at the dataset after adding these two new columns of data.

...	1	Player	Pos	Age	Tm	Cluster	Predicted Salary	Current Salary	Nationality	Actual 2018 Salary
1	1	Arron Afflalo	SG	32	ORL	1	< \$2,020,000	\$2,328,652.00	USA	LEFT NBA
2	2	Will Barton	SG	27	DEN	2	\$2,020,000 – \$4,040,000	\$3,533,333.00	USA	\$13,500,000.00
3	3	Marco Belinelli	SG	31	ATL	2	\$2,020,000 – \$4,040,000	\$6,333,333.00	Italy	\$6,000,000.00
4	4	Avery Bradley	SG	27	DET	5	\$10,100,000 – \$13,130,000	\$8,000,000.00	USA	\$12,500,000.00
5	5	Aaron Brooks	PG	33	MIN	1	< \$2,020,000	\$2,116,955.00	USA	LEFT NBA
6	6	Jose Calderon	PG	36	CLE	1	< \$2,020,000	\$2,328,652.00	Spain	\$2,393,887.00
7	7	Kentavious Caldwell-Pope	SG	24	LAL	3	\$4,040,000 – \$7,070,000	\$17,745,894.00	USA	\$12,000,000.00
8	8	Michael Carter-Williams	PG	26	CHO	2	\$2,020,000 – \$4,040,000	\$2,700,000.00	USA	\$1,800,000.00
9	9	Mario Chalmers	PG	31	MEM	3	\$4,040,000 – \$7,070,000	\$2,106,470.00	USA	LEFT NBA
10	10	Ian Clark	SG	26	NOP	1	< \$2,020,000	\$1,577,230.00	USA	\$1,757,429.00
11	11	Pat Connaughton	SG	25	POR	1	< \$2,020,000	\$838,158.00	USA	\$1,682,025.00
12	12	Wayne Ellington	SG	30	MIA	4	\$7,070,000 – \$10,100,000	\$6,135,000.00	USA	\$6,270,000.00
13	13	Tyreke Evans	SG	28	MEM	8	> \$19,190,000	\$3,290,000.00	USA	\$12,000,000.00
14	14	Raymond Felton	PG	33	OKC	3	\$4,040,000 – \$7,070,000	\$2,328,652.00	USA	\$2,393,887.00
15	15	Yogi Ferrell	SG	24	DAL	5	\$10,100,000 – \$13,130,000	\$207,798.00	USA	\$3,075,000.00
16	16	Bryn Forbes	SG	24	SAS	1	< \$2,020,000	\$724,360.00	USA	\$3,000,000.00
17	17	Tim Frazier	PG	27	WAS	1	< \$2,020,000	\$2,045,000.00	USA	DIDNT RECEIVE CONTRACT
18	18	Treveon Graham	SG	24	CHO	2	\$2,020,000 – \$4,040,000	\$543,471.00	USA	\$1,578,979.00
19	19	Gerald Green	SG	32	HOU	3	\$4,040,000 – \$7,070,000	\$2,328,652.00	USA	\$2,400,000.00
20	20	Devin Harris	PG	34	DAL	4	\$7,070,000 – \$10,100,000	\$4,140,721.00	USA	\$2,393,887.00
21	21	Joe Harris	SG	26	BRK	5	\$10,100,000 – \$13,130,000	\$1,015,838.00	USA	\$8,000,000.00
22	22	Rodney Hood	SG	25	UTA	1	< \$2,020,000	\$1,608,046.00	USA	\$3,472,887.00
23	23	Jarrett Jack	PG	34	NYK	2	\$2,020,000 – \$4,040,000	\$2,328,652.00	USA	\$2,393,887.00
24	24	Shane Larkin	PG	25	BOS	1	< \$2,020,000	\$1,524,305.00	Turkey	LEFT NBA

After obtaining and adding into the dataset all of the relevant data to carry out the analysis, I began cleaning and transforming the data. First, I renamed all of the columns, removing spaces in the names. This allowed for easier and cleaner data processing down the line. I also converted all unknown salary values to NA (or, in other words, a null value). This allows there to be a uniform way to indicate to the data processing functions I utilized to ignore the data that was not available. Next, I removed the '\$' token, as well as all commas, from all salaries. This allowed the conversion of the type of all salary values from a character type to a numeric type. This was necessary to do in order to evaluate the salaries using data manipulation and processing functions down the line. Then, I converted the variable that represents the cluster number from a numeric type into a factor. A factor is a data type used in R in situations of categorical variables. Converting the cluster number to a factor makes sense because there is a set number of clusters. The conversion to a factor allows for functions in the forcat package of R to be applied to this data. Next, I created a new column to hold the midpoint salary value of each cluster

number. Since the clusters represent salary ranges, I took the midpoint of each salary range and included that in the new column. This was done in order to enable comparison between the true and predicted salaries. Furthermore, I created another column to hold the ratio difference between the actual 2018 salary of each player and the predicted amount, represented as the midpoint of the salary cluster. Similarly, another new column was created to hold the ratio difference between the 2017 salary of each player and the midpoint of the salary cluster. Lastly, I made a new column to hold the exact dollar difference between the actual 2018 salary and the midpoint of the salary cluster. It was useful to look into both the ratio difference as well as the exact dollar difference between the actual and predicted salary values. This is because a difference in, for example 1 million dollars, can be much different when considering it from a 40 million dollar contract versus a 500,000 dollar contract. A change in 1 million dollars is not that big when it comes from a 40 million dollar contract, but it is a massive change when it comes from a 500,000 dollar contract. After the transformation and cleaning of the data was completed, it was time to dive into the analysis. I created a number of different tables and visualizations that display different types of data within the dataset. I utilized an array of functions from the tidyverse package of R to accomplish this.

Results:

The results of this investigation and analysis can be best seen through the tables and visualizations that have been created. Meaningful insights and ideas regarding how NBA players are paid can be drawn by taking a dive into the results of the data analysis.

NationalityRegion	Average Ratio Predicted To 2017 Salary	Max Ratio Predicted To 2017 Salary	Min Ratio Predicted To 2017 Salary
Other	0.6259613	3.741465	-0.9062522
USA	0.8497692	15.976077	-0.9821095

The first table we can look at is exploring the relationship between the predicted salaries obtained from the model and the, at the time, current 2017 salaries the players were receiving. In this table, we can see that players from the USA, on average, have a larger ratio difference between their 2017 salaries and the predicted

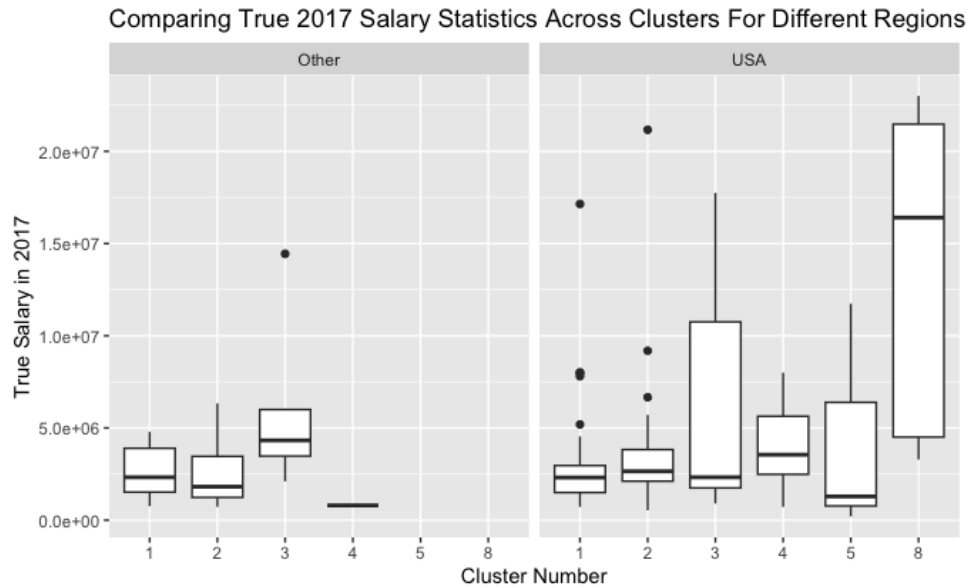
salaries provided by the model. This points to the fact that, at the time, players from the USA were normally getting paid higher than what their value is, based on the statistical model, to more of an extent than players not from the USA. We can also see that the maximum ratio difference is much higher for the players from the USA. This means that, in the most extreme cases, overpaid players from the USA were being overpaid to a much higher extent than overpaid foreign players were. However, we can also see that the minimum ratio difference is a bit smaller for the USA, but it is very comparable between players from the USA and foreign players. This means that in the most extreme cases of underpaid players, the American and foreign players were being underpaid by a similar amount.

Average Ratio Predicted To Future Salary	Max Ratio Predicted To Future Salary	Min Ratio Predicted To Future Salary	Standard Dev Ratio Predicted To Future Salary
0.8101248	7.910891	-0.8697718	1.558917

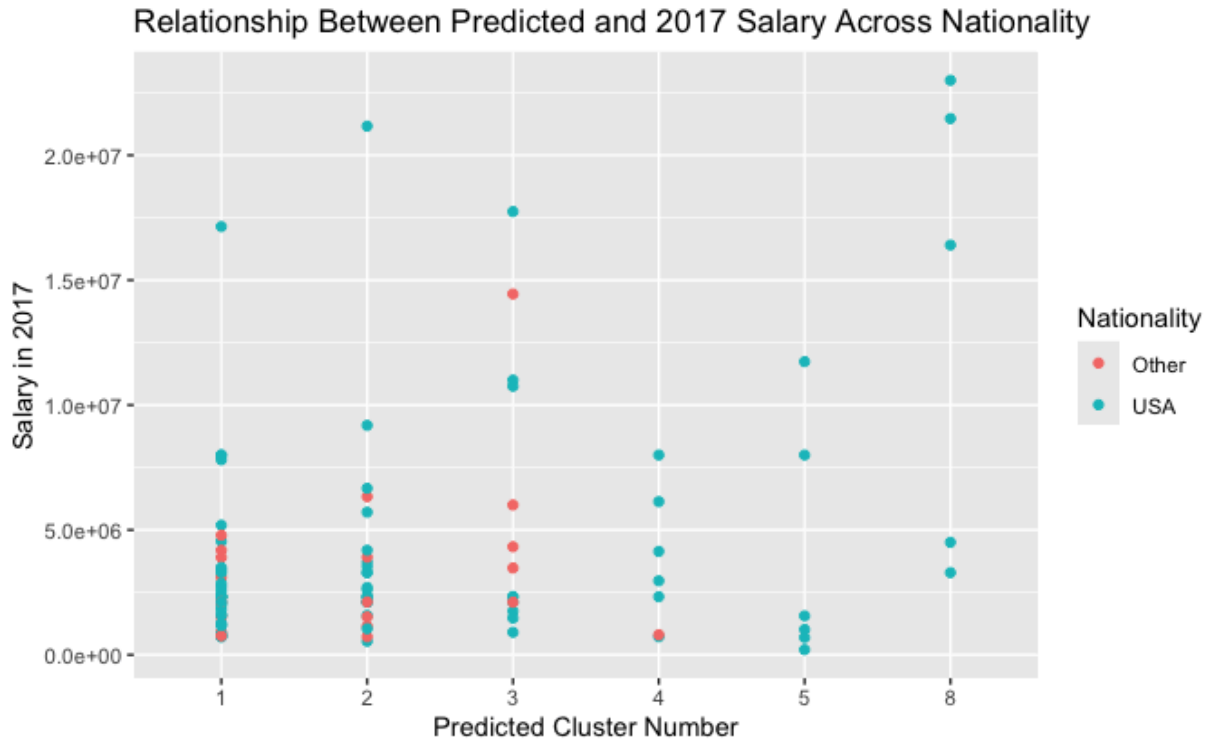
The next table is exploring the ratio difference between the salaries that were predicted by the model and the true salaries that the 2018 NBA free agents ended up receiving during the offseason. We can see that, on average, the actual salaries received were about 81% higher than the salaries predicted by the model. First, this could indicate that the model is consistently under evaluating the worth of players. But, this could also indicate that there are certain traits that players hold that do not show up in the stat sheet. The model uses strictly in-game statistics to evaluate players' worth. However, there are many intangible parts to the game of basketball, so players have worth that is not accounted for by the model. So, intuitively, it makes sense that the model would report a lower value for players than what their true value is. We can see that in the most extreme case of under evaluation by the model, the true salary was 791% higher than the predicted salary. On the other hand, in the most extreme case of over evaluation by the model, the true salary was 87% lower than the predicted salary. This again reinforces the idea that the model is more likely to undervalue players than it is to overvalue players. Lastly, we see that the standard deviation in this case is about 156%. So, players' true salaries are likely to be within 81 plus/minus 156% of the predicted value, showing that the results of the model do range by a little bit.

Average Difference Predicted To Future Salary	Max Difference Predicted To Future Salary	Min Difference Predicted To Future Salary	Standard Dev Difference Predicted To Future Salary
628861.4	14970000	-10190000	4703439

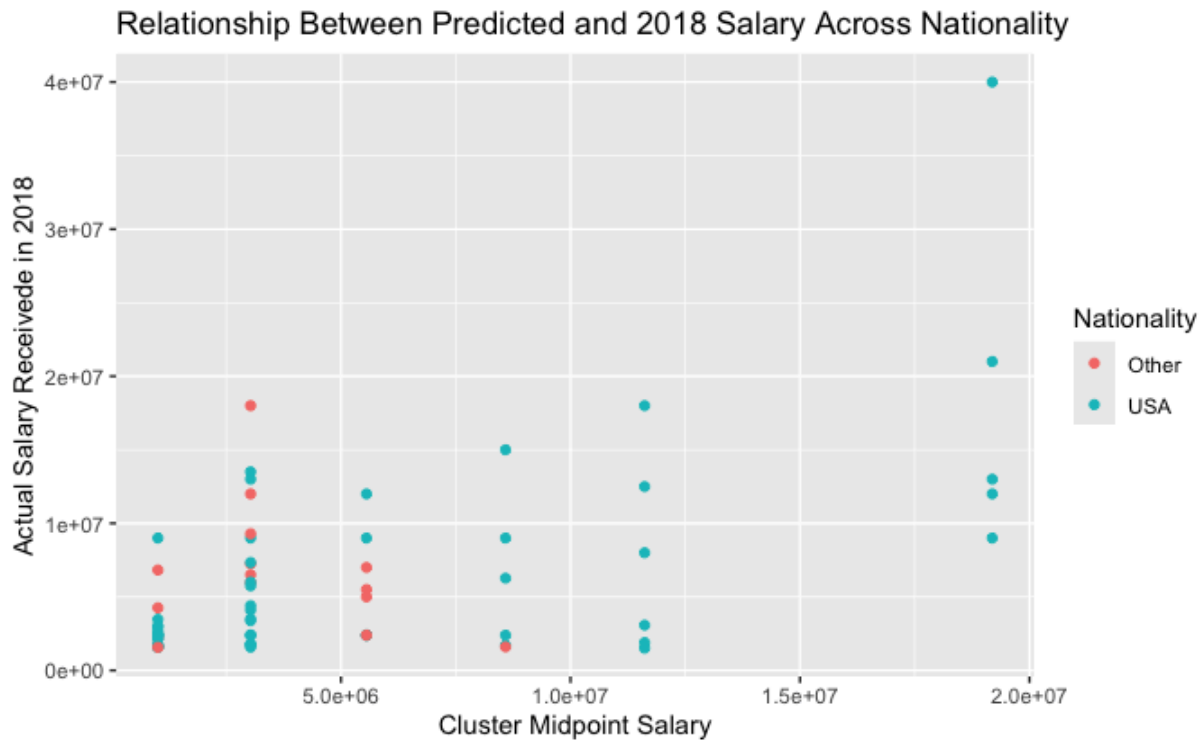
The last table is looking into the exact difference in dollars between the salaries that were predicted by the model and the true salaries that the players ended up receiving in the 2018 offseason. So, this table represents similar ideas to the last one, but gives us a better idea of the real difference in dollars, giving us a unit that we better wrap our heads around. We can see that, on average, the actual salaries received were more than the predicted salaries by about \$628,861. So, when considering that some players are making salaries upward of \$20 million, the model was pretty accurate in its predictions. However, it is important to also remember that many players in the NBA make salaries that are less than \$1 million. So, in these cases, a \$628,861 difference in predicted versus true salary can be quite a big difference. We can see that the most overpaid player received a salary that was \$14,970,000 higher than the predicted salary. On the contrary, we can see that the most underpaid player received a salary that was \$10,190,000 less than the predicted salary. This is a very large amount to be underpaid, so this represents a good example of a player that filled the stat sheet up pretty well, but did not bring that much value to his team. We see that the standard deviation in the difference between predicted and true salaries was \$4,703,439, showing that the model has quite a big range of prediction accuracy. This is a great example of where seeing the true dollar amount provides more context than the ratio amount, but it is important to keep in mind the caveats of this that were described earlier in the report.



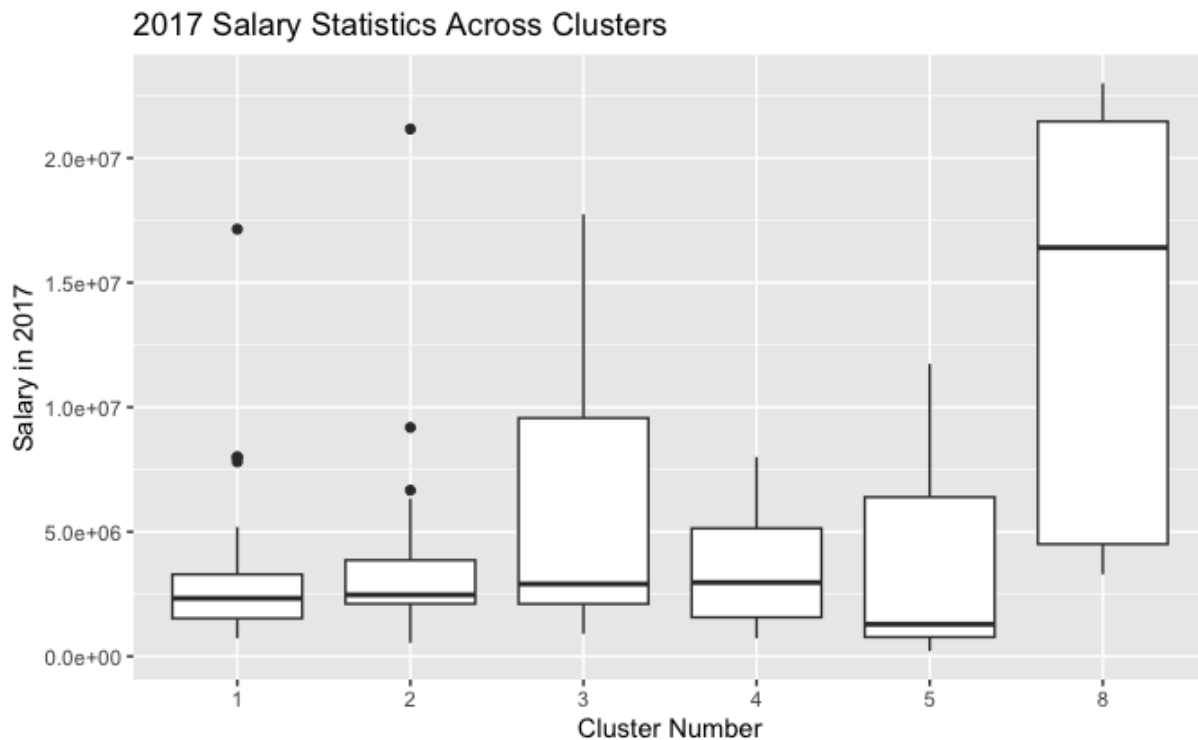
In this visualization, we can observe the difference in 2017 salaries between two different groups of players - those from the USA and those from other regions of the world. I would like to look into the difference between each of the other regions, but due to a lack of available data, I have grouped all other nations into one group. However, we can see that players from the USA, on average, were much higher paid than foreign players. There were more American players in higher clusters, which represents players receiving higher salaries. We can see that there were no foreign players in clusters 5-8. Within the lower clusters, the American players received similar salaries to those from other regions. However, in cluster 3, the mean salary for foreign players was actually higher than that of American players. On the contrary, within cluster 3, there were more American outliers than foreign outliers, pointing to the fact that American players were being overpaid more frequently than foreign players.



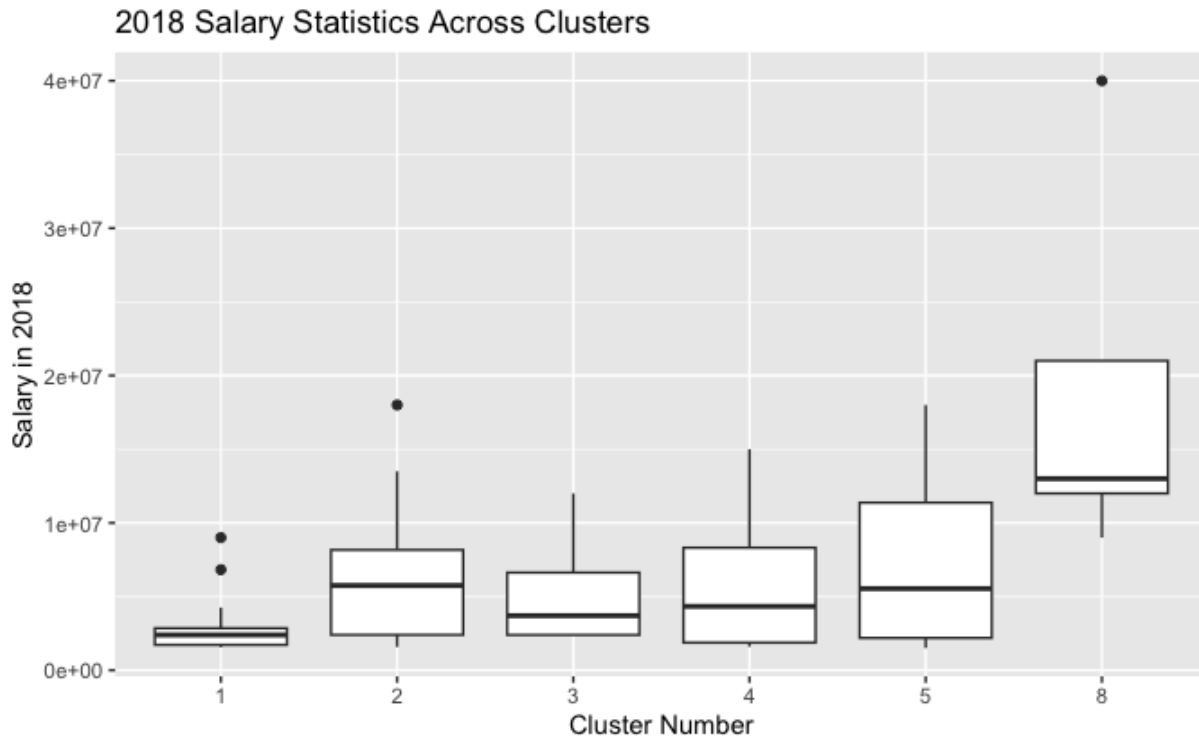
In this visualization, we can see the relationship between the predicted cluster of each player and their 2017 salaries. We can see that the higher clusters contain almost exclusively American players. If the model were perfect, we would expect to see a linear relationship between these two variables. However, this is not exactly the case here. If a player's salary in 2017 was in perfect accordance to the predicted cluster, the dot representing that player would fall exactly at the value on the y-axis that is equal to the midpoint of the cluster. So, players that are unproportionately high on the graph are being overpaid and the players that are unproportionately low on the graph are being underpaid. It can be noticed that most of the players that are being overpaid are American, whereas both American and foreign players can be seen as being underpaid.



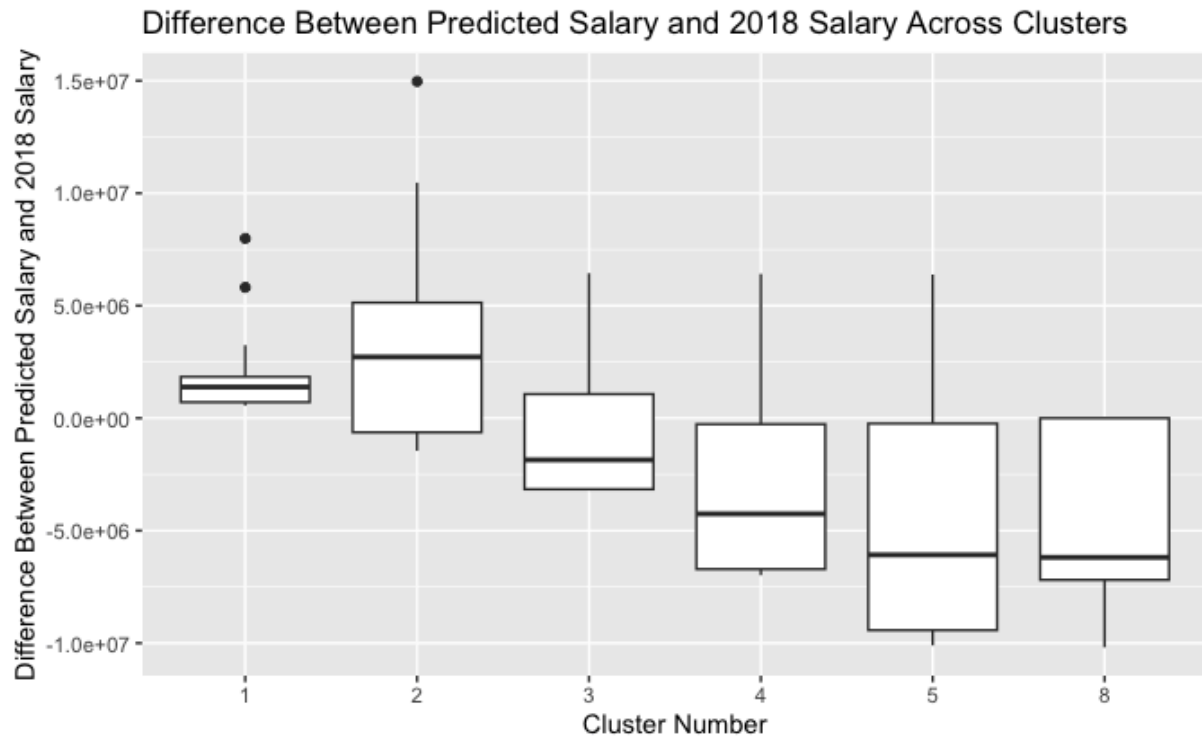
This visualization shows the relationship between the predicted clusters and the true salaries that the NBA free agents ended up receiving in the 2018 offseason. This visualization offers similar insights to the last visualization, but gives a perspective of how the model really performed on new data. Many of the players seem to be lower on the graph, proportionately to their cluster number, than they were in the previous visualization. However, this is mainly due to an extremely high outlier that was introduced in cluster 8, which shifted the scale of the y-axis. Another thing we can draw from this visualization is that foreign players in cluster 2 seem to be getting paid consistently with their statistics, but the same cannot be said about foreign players in cluster 1. This makes sense because new players that have not been in the NBA for long are likely to be found in the lower clusters. This is because teams are usually not willing to invest heavily in a player until the player proves they are worth it. So, the fact that this change between clusters 1 and 2 is more prominent in the foreign players points to the fact that this phenomena gets amplified for foreign players. This points to the fact that NBA managers are especially hesitant to invest in a foreign player before the player proves their worth. However, the big jump in salaries for foreign players in cluster 2 points to the fact that once a foreign player has proven himself, the NBA teams are willing to give them the contract they deserve.



This visualization shows the relationship between the predicted cluster number and the, at the time, current 2017 salaries. However, this visualization does not separate players based on their nationality region. You would expect the boxplots to get higher across the clusters, because higher cluster numbers represent higher predicted salaries. The visualization depicts this somewhat, but with the exception of the 8th cluster, the boxplots are at similar levels. This is especially true when focusing on the mean salary in each cluster. For example, the mean salary in cluster 5 is actually the lowest out of all clusters. This is interesting because we would expect it to be one of the highest. Another takeaway from this visualization is that the variability in salaries among clusters seems to increase as the cluster number increases. This tells us that the predictions become more loose and inaccurate as the cluster number, or predicted salary, increases.



This visualization shows the relationship between the predicted cluster number and the actual salaries the NBA free agents ended up receiving in the 2018 offseason. If the predictions were perfect, we would expect to see the boxplots get higher as the cluster number increases. This is the case here, with the exception of cluster 2. The players in cluster 2 have a higher mean salary value than all other clusters besides cluster 8. This is interesting, and it can most likely just be attributed to random error. Other than cluster 2, the upwards trend is followed quite nicely. Each cluster has a higher mean value than the previous clusters (excluding cluster 2), which is what we would expect. So, this is evidence that the model has done a good job in accurately predicting the salaries of the 2018 NBA free agents.



This visualization displays the difference between predicted salary and the actual salary the 2018 NBA free agents ended up receiving, in an exact dollar amount. We can see that the difference seems to get smaller as the cluster number increases. However, this is in regards to total difference, not absolute difference. For example, a difference of -5,000,000 indicates that a player received \$5,000,000 dollars less than the model predicted. So, in this case, a value close to zero indicates that the model has made an accurate prediction. By this logic, we can see that the predictions were most accurate for players in cluster 1, and the predictions become less accurate as the cluster number increases. This makes sense because the difference is being presented in an exact dollar amount. So, since the overall value of contracts increases and the cluster number increases, we would expect this difference in total dollar amount to increase as the cluster number increases as well.

Discussion:

I came across a statistical model that was created back in 2017 and the purpose of it was to use in-game statistics to predict the salary of each 2018 NBA free agent. When I came across this, my mind started buzzing with ideas about how I could take the previous work and extend it to provide more insight and discover new ideas. After considering all these thoughts that were flying through my head, I was able to come to a couple of specific ideas or areas that I wanted to explore the most.

The first notion I wanted to look into was the idea of nationality in the NBA. In particular, I wanted to discover whether there was a discrepancy between how players of different nationalities get paid in the NBA. People experience prejudice and unjust pay in workplaces all over the world, so I figured it could be taking place in the NBA as well, even though you do not hear about it often. If the data were to point to the idea of a prejudice against certain nationalities, this could provide major breakthroughs in the world of sports. In this day and age, people are trying their best to remove and reduce prejudice wherever possible. But this goes unnoticed in sports sometimes, so being able to analyze the data and come to a conclusion of prejudice in this field would give the opportunity to share this problem with the broader sports community. And upon informing the community on this issue, hopefully, as a collective, we could take the necessary steps to eliminate the prejudice.

The second idea that I wanted to explore was the accuracy of the model that I was using. The model was created in 2017, with hopes to make predictions about the NBA free agents of 2018. So, now that the real salaries that the NBA free agents of 2018 received are known and placed in databases online, I thought this would be an amazing chance to revisit the model. I wanted to explore exactly how well the predictions really were. Depending on how well the model's predictions were, this could give insight into how we can improve models in the future. We could explore which aspects of the model worked well, and which ones did not. From there, we could use that knowledge to tweak the model and make one that is even better at predicting the salaries or worth of NBA players by using their in-game statistics.

After taking a deep dive into this data regarding 2018 NBA free agents, I was able to come to some partial conclusions and gain insight on new ideas. In terms of the idea of discrimination or prejudice against players of different nationalities, I was able to draw some insight, but it would require some more analysis on a larger set of data to be able to decisively conclude that there is serious prejudice in the NBA. The best insight I got regarding this idea can be seen through the visualizations in this study. First of all, it was apparent right away that, on average, American players get paid a substantial amount more than foreign players do. This was evident because, for almost every single cluster of players, the American players averaged a higher salary than the foreign players did. Furthermore, one of the biggest indicators was the number of outliers in these visualizations. In almost all cases, there were more American players that were extremely high outliers than there were foreign players. In this case, extremely high outliers is referring to players that make a salary that is much higher than the salary that the model suggests. Since the model only uses concrete numbers from in-game statistics, it is a completely unbiased predictor. So, since American players get paid a lot more than the predicted value much more frequently than foreign players, this could be a sign that American players are getting overpaid and foreign players are getting underpaid. So, all in all, this data does not give decisive enough reason to come to a complete conclusion of prejudice in the NBA, but it does give us enough reason to warrant further investigation into the issue of discrimination in the NBA.

With regards to the accuracy of the model, I was also able to gain some insight and reach conclusions through the data analysis. First, by comparing the predicted salaries to the 2017 salaries, we could see the beginning of a trend that was then enforced by comparing the predicted salaries to the 2018 salaries. Comparing the predicted salaries to the 2017 salaries provided a good baseline, but it was not a good way to come to a final conclusion. The 2017 salary data was used in the building and training of the model, so it is never a good way to come to a decisive conclusion, but it did set the stage for the comparison with the 2018 salary data. When comparing the predicted salaries with the 2018 salaries, it became clear that the model performed well. This was apparent in the boxplot visualization titled “2018 Salary Statistics Across Clusters.” If the model was perfect, we would expect to see the 2018 salaries increase as the prediction cluster number increased. This was indeed exactly what was observed. With the exception of one cluster,

every prediction cluster correctly had a higher average 2018 salary than the previous cluster. This showed us that the model did a good job of identifying which players should be in which cluster. So, all in all, the exploration of the model accuracy proved to be worthwhile and indicates that this model does a good job in predicting NBA player salaries given data regarding the players' in-game statistics.

Conclusion:

After embarking on a journey through the data to explore the ideas of prejudice in the NBA, as well as investigate the accuracy of the predictive model at hand, I was able to gain some useful insight and draw useful conclusions. Some of them turned out to be sound conclusions, and others did not quite make it there, but provided enough insight to warrant a further investigation of the issue. The investigation into the accuracy of the predictive model reached a decisive conclusion that the model does a good job in predicting the salaries or worth of NBA players, based on in-game statistics. We saw this through the comparison of the predicted salaries of the 2018 NBA free agents to the true salaries those free agents ended up receiving in hindsight. The investigation of prejudice in the NBA was not able to produce a conclusion as decisive as that of the investigation into the accuracy of the model. It gave us reasons to believe that American players, on average, get paid in a more just way than foreign players. However, the sample size proved to be too small to draw a final conclusion that foreign players experience prejudice in the NBA. The insight it did give us, though, is enough evidence to warrant further investigation of this issue. These conclusions could provide great importance and impact within the fields of computer science and data science in the greater sports community. Prejudice is an issue that we, as humans, should aim to reduce every day. This project gives us a reason to believe that there may be prejudice present in the payment that foreign players receive in the NBA. That, in itself, warrants a further investigation into this issue with a broader set of data to look into. This is a great discovery because it gives us reason to not turn the other way when hearing about prejudice in sports, but rather to challenge it at face-value and determine if there truly is discrimination occurring. If there is, we must get to the bottom of it and abolish it, to continue bettering sports and humanity as a whole. This project also provided great insight into how to make an accurate predictive model in the world of sports data. This project verified that the model at hand performed well, and so

we can now use this model as a great example of a predictive model in the realm of sports. This, again, is an important topic. By introducing and using predictive statistical models to explain sports phenomena, we can improve aspects of sports, such as fairness, equality, revenue, and quality of players. Sports are a great way to foster community, so improving the outlook of sports has the power to bring more people together and help individuals attain happiness in their community. At the end of the day, we all aim to live happy lives within our communities and to eliminate prejudice and discrimination, and this project helps us take some initial steps in that direction.