NBA First Non-Rookie Contracts

Summary:

Rookies in the NBA come into the league with a salary that is wholly determined by the position in which they were drafted. These contracts can last up to 4 years, depending on if the incumbent team decides it wants to keep said player. After those 4 years are up, the player is eligible for an entirely new contract. There's a great deal of variance in recently drafted players' skill sets; some quickly become good players, others develop more slowly, and others do not ever become serviceable NBA players. As a result, an NBA player's first non-rookie salary is reflective of how that player performed during those first 4 years in the NBA. The objective of this project is to estimate what this first non-rookie contract will look like for a player based on his body of work up unto that point. This type of model would be of use to NBA front offices, agents, players, and writers among others.

Data Wrangling:

The dataset used for this project was primarily from Basketball-reference.com. The goal was to scrape two main things from BasketballReference:

- A list of players drafted into the NBA
- Relevant statistics for every player in our above list

First, a list of players drafted into the NBA was generated by iterating over the yearly draft pages (e.g. https://www.basketball-reference.com/draft/NBA_2012.html). This became the foundation of our NBA player list, which would later be refined.

Second, each player specific page was scraped for relevant data to be used in the creation of the future salary model (e.g. https://www.basketball-reference.com/players/s/smartma01.html). In each of these pages, tables of several types of data are stored for the player. The tables used for this analysis were:

- Per Game – contains per game statistics, like points, rebounds, assists etc.
- Advanced – contains advanced metrics like TS%, PER, BPM etc.
- Contract – contains salaries by year

The first two tables made up the independent variables used in the eventual models. The final table was used in calculating the dependent variable in our analysis. As players' webpages were scraped, any player who did not play in the NBA long enough to reach his next contract, was removed from the analysis.

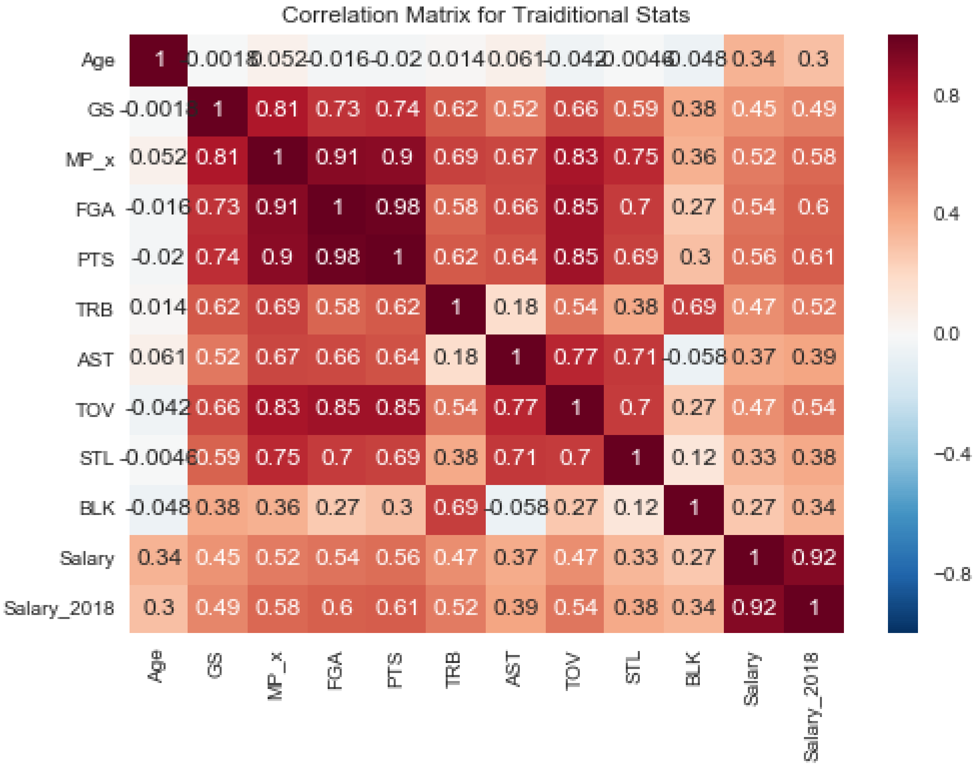
The last piece of information that was brought in was Salary Cap information. This information was brought in from basketball.realgm.com. This information was brought in as a way to scale salaries down to a percentage of the league's cap for that year. This was an important adjustment, as our analysis spanned from 1990 to 2013, over which the NBA's salary cap increased from $12 million to $58 million. Lastly, the numbers were converted to 2018 dollars, where the salary cap sits at $102 million. These 2018-adjusted dollars became the dependent variable in the eventual models.

All of this data was combined and output into a TSV. Some small formatting and null handling was performed to get the data in a suitable format for data analysis.

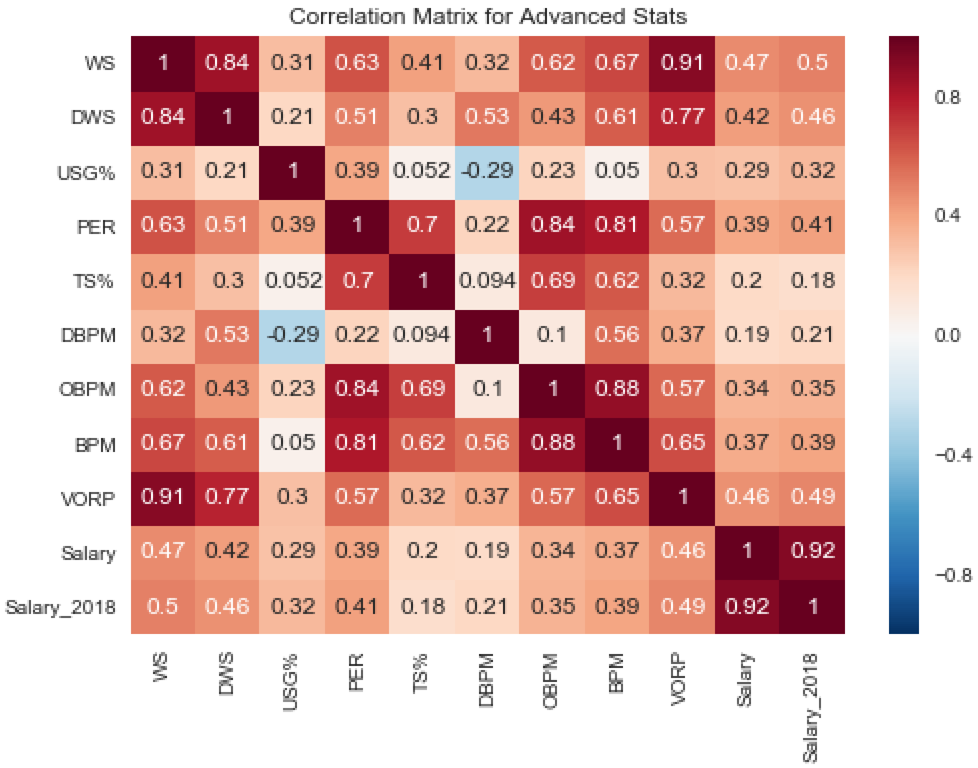
Once the data was neatly stored in a TSV, it was then transformed and condensed so that it could be better handled and analyzed. The TSV of data before transformation had a row of statistics for each year a player played on their rookie contract, 4 years. Rather than having to deal with multiple entries of every variable, a dataframe of weighted averages was created. The weights used were the overall minutes played by the player in that given season. This allowed the data to be condensed into a single row so that inferential statistics and the ultimate model building would be more straightforward.

Inferential Statistics:

Once the data was neatly stored in a TSV file, the data exploration began. First, correlation matrices were created for both the traditional and the advanced statistics. A subset of these are shown below in correlation matrices.



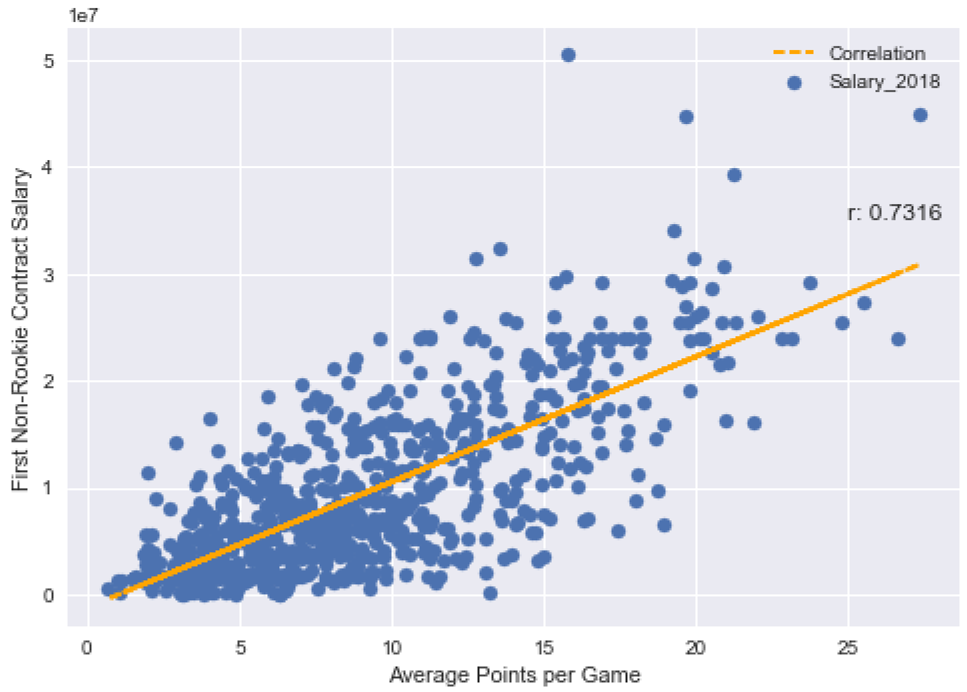
The correlation matrices are used to inform and guide what variables might be useful in the eventual models for first non-rookie salary. Certainly, variables with high correlations with our target variable, Salary_2018, are good candidates for the model. This suggests the use of variables like, PTS (Points), FGA (Field Goal Attempts), and MP_x (Minutes played) among others. Another consideration was the correlation among these independent variables. Things like points and field goal attempts are very highly correlated and would like not provide any additional information if both were included in model. The correlations among independent variables displayed above helped weed out the potential issue of using redundant independent variables.

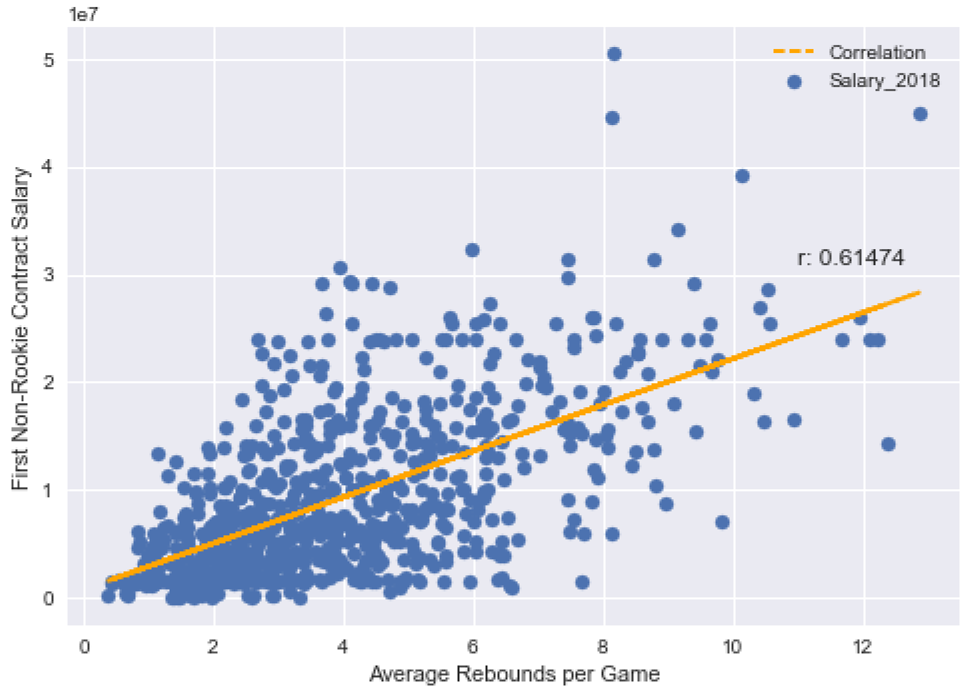


The above correlation matrix for advanced statistics helps accomplish the same goal of guiding the selection of independent variables. Somewhat surprisingly, the correlation levels are lower across the board for the advanced statistics. For some of these metrics, this makes sense as they are rate statistics and not counting statistics. What is meant by that is that statistics like True Shooting % (TS%) only capture how efficient a player is at shooting but does not take into account how *much* a player is shooting. Certainly, a player who shoots 20 times a game at a TS% of 0.60 is more valuable than a player that shoots 1 shot per game at a TS% of 0.70. This is the issue with using rate statistics without incorporating the associated volume in anyway. That said, some metrics like WS and VORP are not rate statistics, and do merge both efficiency and volume, yet these metrics still did not correlate appreciably better than most traditional statistics. This correlation matrix also reveals some interesting findings, particularly with the negative correlation between Defensive Box Plus Minus (DBPM) and Usage (USG%), and the low correlation between Offensive Box Plus Minus (OBPM) and DBPM. Neither of these findings influence the final model but are interesting to note nonetheless.

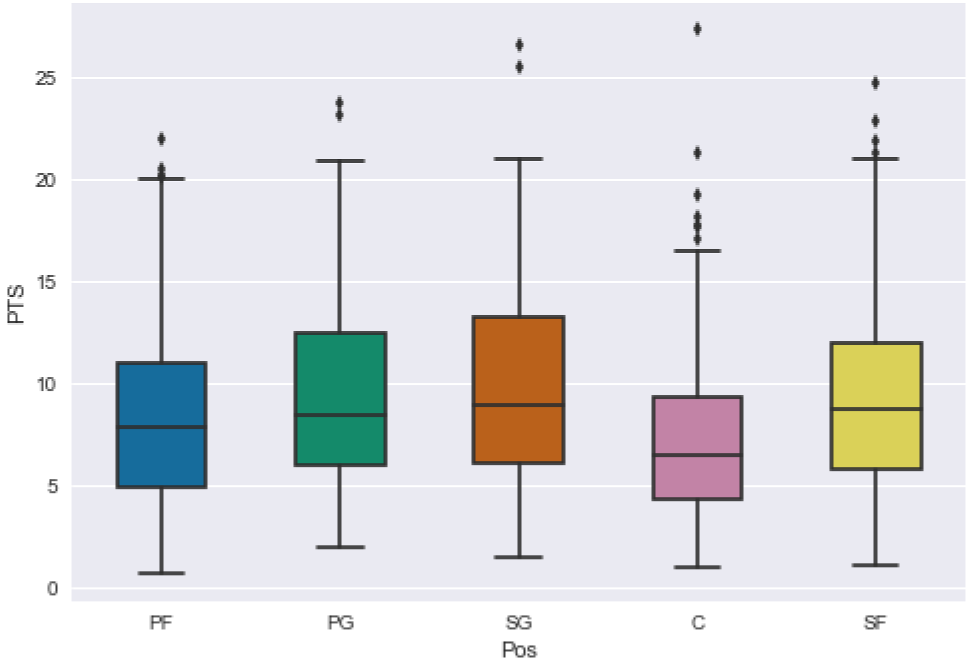
While not pictured above, the Advanced statistics and Traditional statistics were also analyzed so that it could be determined which variables correlate with one another. This was done to ensure that no sets of variables were included in the eventual model that explained the same portion of the target variable’s variance.

With the above correlation matrices informing the variable selection, more in depth variable analysis began. Particularly correlation levels and scatter plots were created for select variables.

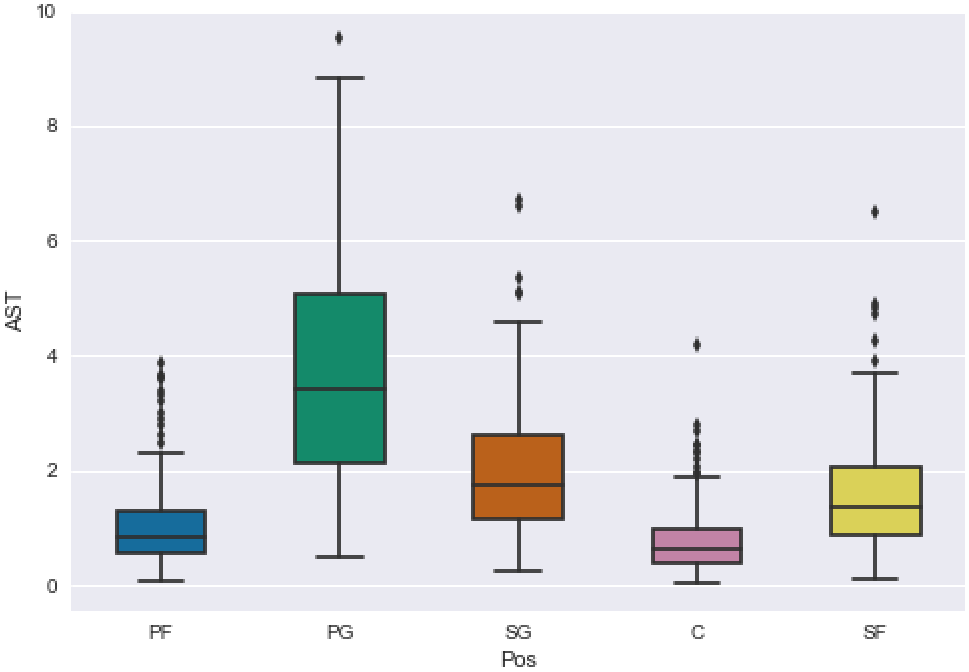




Points and Total Rebounds were among the variables that had the strongest correlation with the Salary % target variable. Scatter plots were generated to visual the correlation level and look for potentially non-linear relationships in the data. In addition to these single variable scatter plots, boxplots were created to see how variables changes across different fields, like Position.

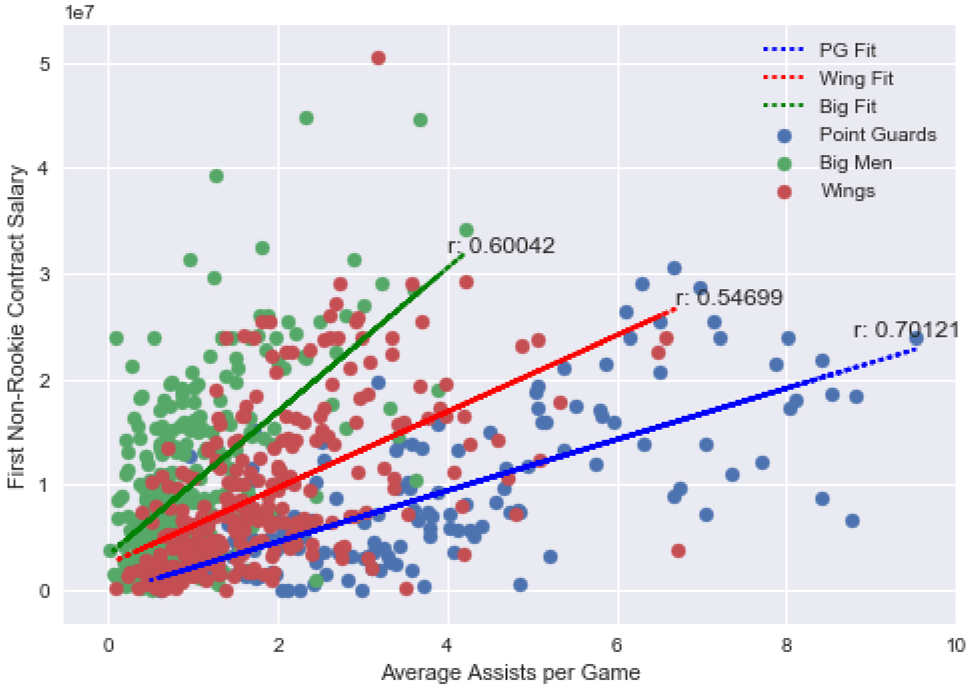


This Points boxplots broken down by position is important in how consistent the boxplots look. The Center (C) position notwithstanding, these box plots look quite similar, especially when juxtaposed to other variables like Total Rebounds (TRB) or Assists (AST) like below. Given the difference in certain statistics by position, it was considered advisable to analyze the effects of those statistics on salary by position as well, potentially using some in the position specific sub-models.



It is clear when looking at this positional breakdown of boxplots for Assists (AST), varies greatly by position. The Shooting Guard (SG) and Small Forward (SF) plots look noticeably wider and higher than the Power Forward (PF) and Center (C) plots, all of which look dramatically different to the Point Guard (PG) plot. The lower quartile of Assist for Point Guards is higher than the median for all other positions and even higher than the upper quartile for Forwards and Centers. This is a good indicator that Assists would be a variable that will differ positionally and thus should be included in an eventual position specific sub-model.

The next portion of the analysis was to see if certain variables have statistically significant differences in their correlation levels with respect to the target variable, Salary%. The previously outlined boxplot analysis informed which variables to test for differing correlation levels.



Researching which variables had differing correlation levels across positions led to the selection of variables for the position-specific portion of the ensemble model. The end result used AST, ORB, 3P, and BLK as position-specific variables in the ensemble model. These variables all make some intuitive sense as they are each more associated with one position than others.

Machine Learning:

Now that the framework of each of the models has been sketched out via the data wrangling and inferential statistics portion of the Capstone, models can be created to explain the current dataset and predict future results.

General Model Approach:

The first model that was created was a generic ordinary least squares regression. The dependent variable, as laid out before, was the Salary % (the Salary of the player as a percentage of the salary cap for that year). The independent variables used were: PTS, AST, TRB, DWS, BLK, and Age. We can see from the below table of summary results that this model has an R^2 of 0.646, meaning this model explains about 65% of the target variables variance. Other information we get from the output is that PTS have considerably the largest impact on the predicted salary of a player, followed by AST and TRB in this model. Unsurprisingly, increased Age has a negative effect on the predicted salary of a player.

OLS Regression Results

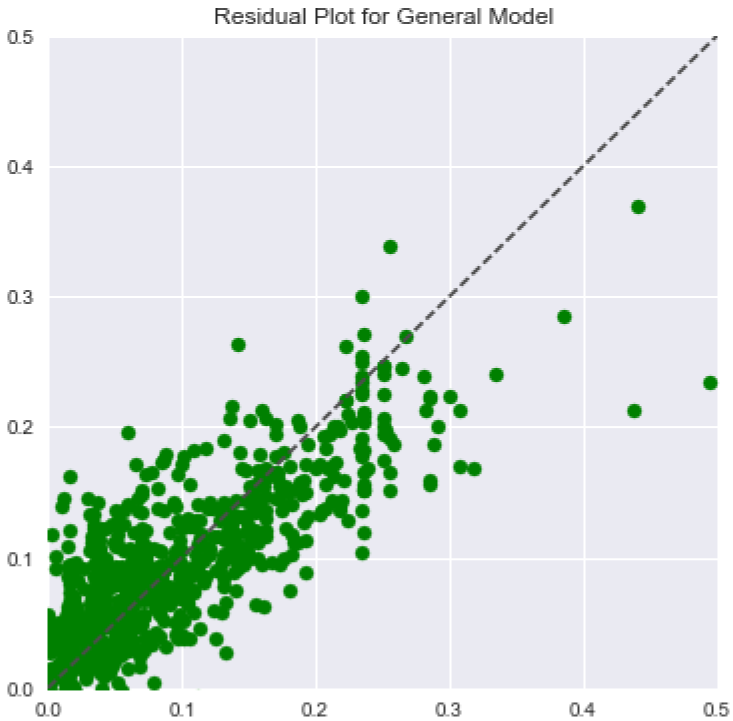
Dep. Variable:	y	**R-squared:**	0.636
Model:	OLS	**Adj. R-squared:**	0.632
Method:	Least Squares	**F-statistic:**	176.2
Date:	Wed, 04 Jul 2018	**Prob (F-statistic):**	3.22e-129
Time:	10:27:43	**Log-Likelihood:**	-10261.
No. Observations:	612	**AIC:**	2.054e+04
Df Residuals:	605	**BIC:**	2.057e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	1.038e+07	2.89e+06	3.588	0.000	4.7e+06 1.61e+07
PTS	6.702e+05	6.58e+04	10.192	0.000	5.41e+05 7.99e+05
AST	4.594e+05	1.68e+05	2.734	0.006	1.29e+05 7.89e+05
TRB	3.649e+05	1.6e+05	2.283	0.023	5.11e+04 6.79e+05
BLK	1.983e+06	5.4e+05	3.674	0.000	9.23e+05 3.04e+06
DWS	1.82e+06	3.01e+05	6.056	0.000	1.23e+06 2.41e+06
Age	-5.754e+05	1.22e+05	-4.703	0.000	-8.16e+05 -3.35e+05

Omnibus:	68.274	**Durbin-Watson:**	2.008
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	228.035
Skew:	0.491	**Prob(JB):**	3.04e-50
Kurtosis:	5.824	**Cond. No.**	386.

The metrics that was used for comparing the quality of the model is the Root Mean Squared Error (RMSE). The RMSE is the standard deviation between the predicted and modeled values. The RMSE for this generic OLS model was: **4,599,389.** We can interpret this as such: 68% of our modeled results fall within $4.6M dollars of the actual 2018-adjusted value.

Looking through the other output of the model, we can gain some additional insights, particularly the incremental effect of various statistics. Averaging a single point per game results in a 2018-adjusted salary increase of $670,000. An increase of a block per game would result in nearly $2M more in 2018-adjusted dollars. It’s important to note that these statistics have different scales. It is much more difficult to average an additional block per game than it is to average an additional point.



The above residual plot for the general model show that the residuals are not biased in any one direction. The residuals fall on either side of the dashed line suggesting minimal bias in the model.

One thing to note about the general model, is that it does not include any train-test split of the data, rather, it uses all the available data in the dataset. The general model was separately built after having trained the data, and tested on the test subset of the data. The general model performed slightly worse with a root mean squared error of $4.68M. It is not surprising that the general model only performed marginally worse. This is because test-train splits are effective ways to combat overfitting of data. This general model was unlikely to be overfitting given the number of data points and few variables used in the model. Had the model been expanded to include many more variables, a train-test split approach would've likely demonstrated the danger of overfitting.

Ensemble Model Approach:

The next model that was created was an ensemble model. The ensemble model had two components, a general component, featuring variables that were not positionally-specific, and a position-specific component.

The general component of the ensemble model used: PTS, DWS, and Age as its independent variables. This general component of the ensemble model still had a R^2 of 0.629, and a RMSE of **4,711,795.**

The position specific sub-models used 3P, AST, ORB, and BLKs as the independent variables. Each of the position specific sub-models had R^2 values around 0.60. To calculate the overall RMSE of the model the following approach was used:

- Iterate over a set of weights (between 0 and 1) which correspond to the weight assigned to the general component, and the position-specific component of the sub-models.
- For each weight, calculate the RMSE for each sub-model
- Calculate the average RMSE for the sub-models.

This analysis yields a weight breakdown of 30/70 where 30% of the weight is assigned to the position-specific portion of the model. The resulting RMSE of this overall model is **4,544,377**.

Conclusion:

There are a variety of takeaways from the entire analysis. The results of the two models demonstrates that for a predictive salary model to be useful, it at least account for some degree of positionality. The position-specific ensemble model saw improvements, albeit slight, over the original general model. The RMSE saw an improvement of about 1.2% in the ensemble approach. While this improvement is not large, it suggests that the position-specific ensemble approach is more appropriate than the general OLS model. With additional data and position-specific variables, its reasonable that the difference would be even larger between the models.

Either of the models have some use as evidenced by the R^2 and RMSE values and could be effectively used as guidelines for NBA general managers, agents, and writers. An RMSE of $4.5M is low enough to be valuable as there are frequently new contracts of sizes that surprise the general NBA audience. Using models such as these could help inform people when a new contract is at least to in tune with the market perception of the player. With that said, this model is still crude enough that NBA teams would likely need much lower RMSE and higher R^2 to make actual decisions off the output.

There are a variety of ways in which this model could be enhanced down the line to further improve its estimates. The most obvious improvement is more data. The dataset could be augmented by simply adding more years' worth of data, though the further back, the less relevant that data likely is to the current NBA. This analysis only focused on Traditional and Advanced statistics from BasketballReference. There are myriad variables that could be considered in the analysis both from BasketballReference and from other sources like NBA.com. Some types of data that come to mind: per-possession data, team data, and teammate data.

In addition to augmenting the dataset, different approaches could be taken to create this model. Lasso or Ridge regression could be implemented which would be particularly helpful if more variables were used in the model. Perhaps players could be first clustered into similar groups (this might differ from positions) and models could be created for those specific clusters. Different amounts of weight could be applied to the years in which a player produced their statistics (more weight to recent years).

Overall the end result was a model that has some real uses, particularly from an educational perspective. For a production quality model, additional work, like that outlined above, would need to be done. This has provided a useful foundation for that work to take place and a more sophisticated and powerful model to be created.

Appendix A

Age - Age
Pos - Position
G – Games played
GS – Games started
MP – Minutes played
FG – Field goals made
FGA – Field goal attempts
FG% - Field goal percentage
3P – Three pointers made
3PA – Three point attempts

3P% - Three point percentage
2P – Two pointers made
2PA – Two pointers attempted
2P% - Two point percentage
eFG% - Effective field goal percentage
FT – Free throws made
FTA – Free throw attempts
FT% - Free throw percentage
ORB – Offensive rebounds
DRB – Defensive rebounds
TRB – Total rebounds
AST - Assists
STL - Steals
BLK - Blocks
TOV - Turnovers
PF – Personal fouls
PTS - Points
PER – Player efficiency rating
TS% - True shooting percentage
3PAr – Three point attempt rate
FTr – Free throws attempt rate
ORB% - Offensive rebound percentage
DRB% - Defensive rebound percentage
TRB% - Total rebound percentage
AST% - Assist percentage
STL% - Steal percentage
BLK% - Block percentage
TOV% - Turnover percentage
USG% - Usage percentage
OWS – Offensive win shares
DWS – Defensive win shares
WS – Win shares
WS/48 – Win shares per 48 minutes
OBPM – Offensive box plus-minus
DBPM – Defensive box plus-minus
BPM – Box plus- minus
VORP – Value over replacement player