Prediction of NBA Salaries

Math 498 Final Project

Erika Iwule

May 18th, 2024

Introduction of the Data:

The NBA is a multi-billion dollar industry, estimated to be worth around $120 billion. NBA players' salaries come from a collective of TV viewership, sponsorship, and ticket sale revenues. Each team gives each player a salary contract based on guidelines from the NBA. The guidelines use a player's statistics to determine their value to the individual team and the NBA organization as a whole. I have collected a data set from Kaggle that provides each NBA player's statistics from the NBA 2022-2023 season to conduct several analyses.

The first analysis is to predict NBA players' salaries based on numerous parameters from the 2022-2023 NBA season. A few parameters are; the player's position, age, games played, games started, game minutes, points scored, etc. The goal is to create a regression model to predict an NBA player's salary. Additionally, determine which regression model can predict a player's salary more accurately.

The second analysis is to select the best subset of predictors that best fit the model. The goal is to use forward selection to determine the best subset of predictors. Then repeat the first analysis on the new best subset of predictors. As well as, to look at decision trees for the NBA salary data.

Description of the Dataset:

The dataset title is "NBA Player Salaries (2022-23 Season)" from Kaggle.com. The original dimensions are 467 rows/observations and 52 columns/variables. The observations are active players' salaries in the NBA during the NBA 2022-23 season. For the analyses, I cleaned the data by removing unpopular variables. I began only using 14 variables. One variable, Position, is categorical. I turned each position into a dummy variable, which increased my total number of variables to 18. The original sample size is 467 players' salaries. To reduce errors in the models, I have excluded any players who played 20 or fewer games due to injury. I also dropped any empty observations. The final sample size is 367 players' salaries.

Variable Descriptions:

Variables	**Description**
Position	Position of Player
Age	Age of Player
GP	Number of Games Played
GS	Number of Games Started
MP	Average of Minutes Per Game
FG	Average Field Goals Made Per Game
FG%	Field Goal Percentage (The ratio of field goals made to field goals attempted)
3P	Average Three-Point Field Goals Made Per Game
3P%	Three-Point Percentage (The ratio of three-point field goals made to three-point field goals attempted)
2P	Average Two-Point Field Goals Made Per Game
2P%	Two-Point Percentage (The ratio of two-point field goals made to two-point field goals attempted)
ASST	Average Assists Per Game
PTS	Average Points Per Game
Total Minutes	Total Minutes Played in the Season

Position Abbreviation	**Position Description**
C	Center
PG	Point Guard
SG	Shooting Guard
SF	Small Forward
PF	Power Forward

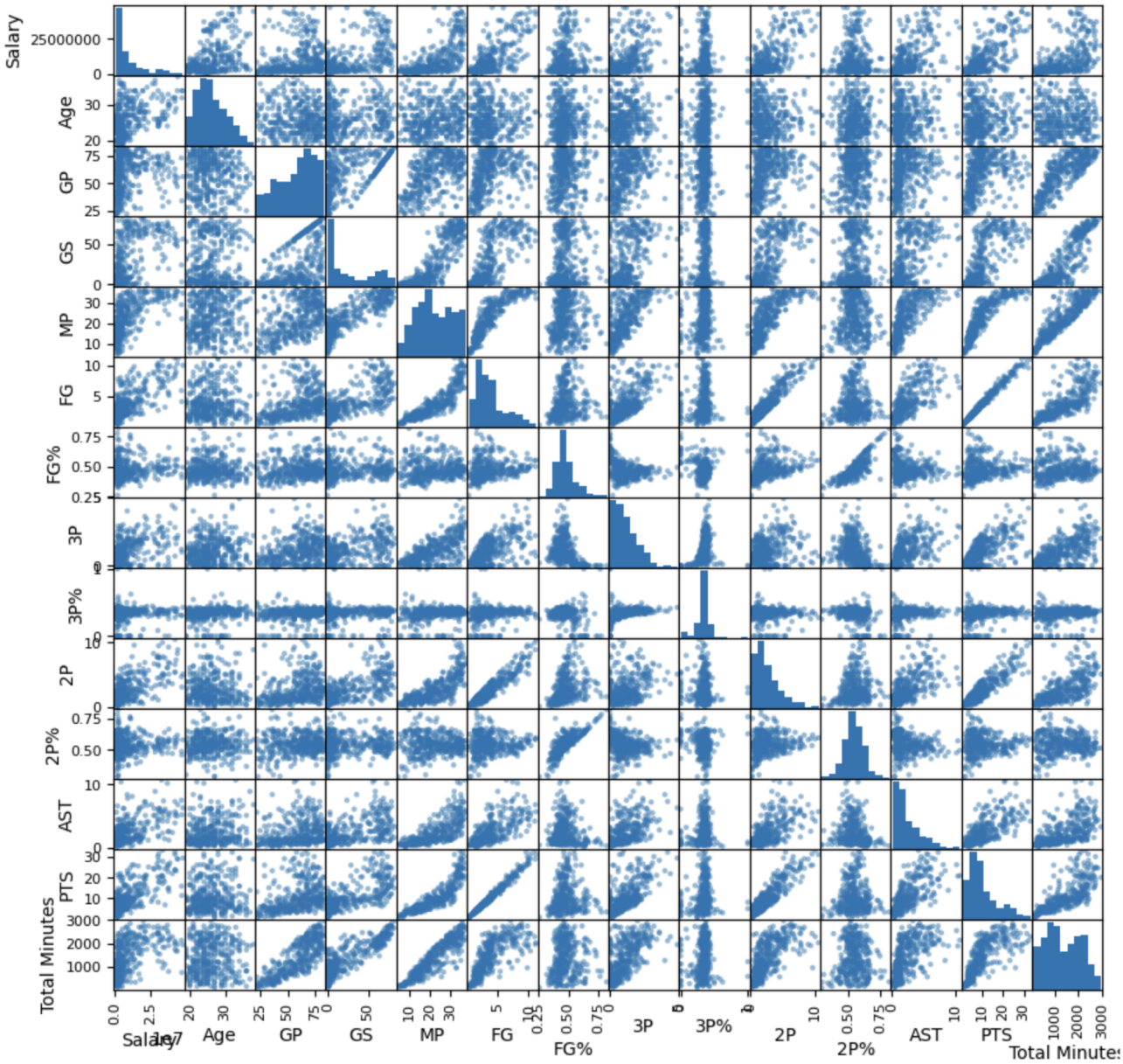
Data Summary:

	Salary	Age	GP	GS	MP	FG	FG%
count	367	367	367	367	367	367	367
mean	$10,109,549	26.02	58.23	28.13	22.25	3.83	0.47
std	$11,233,082	4.30	16.20	27.73	8.54	2.45	0.08
min	$386,055	19.00	22.00	0.00	4.70	0.30	0.26
25%	$2,271,820	23.00	46.00	3.00	15.10	2.00	0.43
50%	$5,155,500	25.00	62.00	16.00	21.70	3.20	0.46
75%	$13,437,409	29.00	72.00	58.00	30.00	5.10	0.51
max	$48,070,014	38.00	83.00	83.00	37.40	11.20	0.78

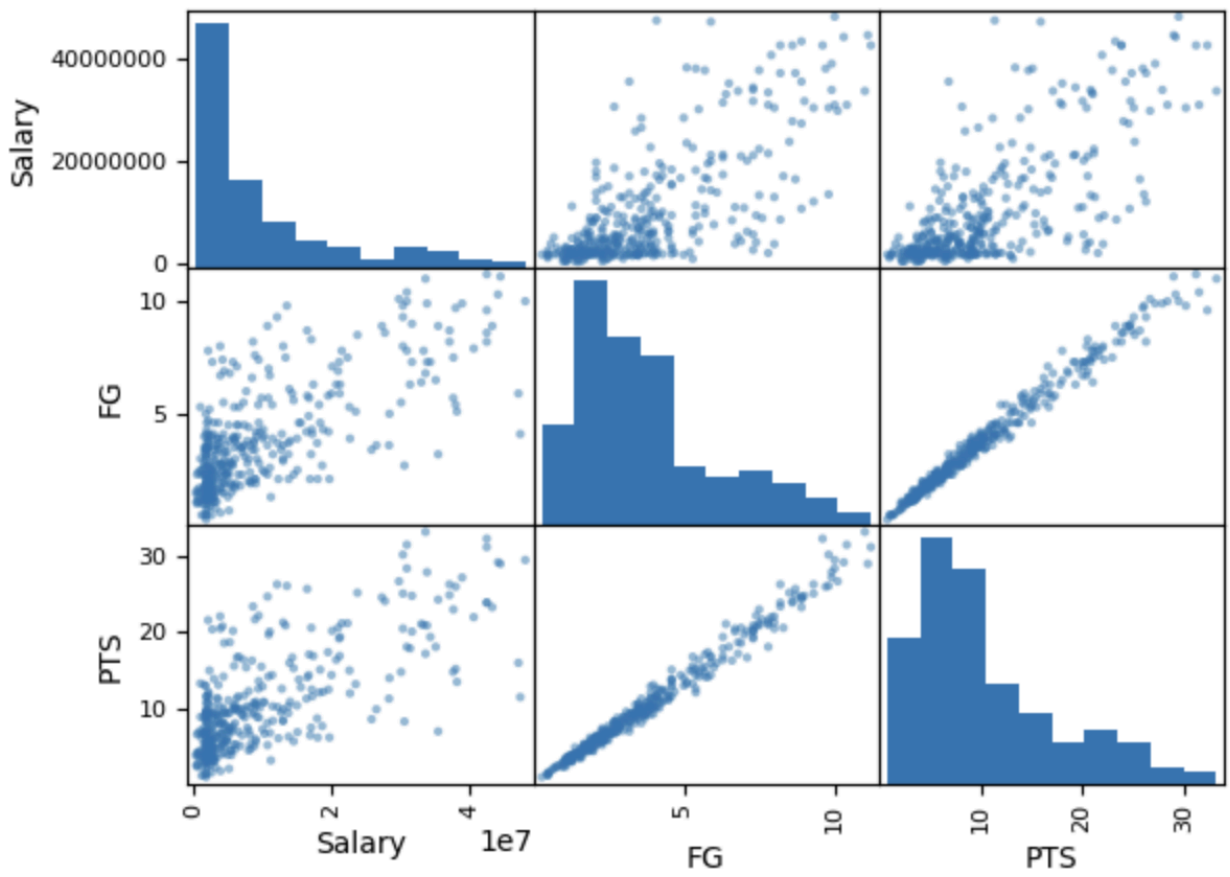
	3P	3P%	2P	2P%	AST	PTS	Total Minutes
count	367	367	367	367	367	367	367
mean	1.14	0.33	2.69	0.54	2.39	10.49	1,369.63
std	0.89	0.11	2.03	0.08	1.97	6.93	726.87
min	0.00	0.00	0.10	0.29	0.20	0.90	107.00
25%	0.50	0.31	1.20	0.49	1.00	5.30	767.50
50%	1.00	0.35	2.00	0.54	1.60	8.70	1,258.00
75%	1.70	0.39	3.75	0.59	3.40	13.75	1,986.00
max	4.90	1.00	10.50	0.78	10.70	33.10	2,963.00

The data summary shows the mean, standard deviation, median, quartiles, min, and max of each variable and the observations. The mean salary of NBA Players is $10,109,549. The salary observations have a maximum value of $48,070,014 and a minimum value of $386,055. The median value for the salary observations is $5,155,500. As well, the upper quartile is $13,437,409 and the lower quartile is $2,271,820. Based on the summary, we can see the salary data is significantly skewed. Variables that are also skewed are GS (Games Started), PTS (Average Points Per Game), AST (Average Assists Per Game), FG (Average Field Goals Made Per Game), 3P (Average Three Point Field Goals Made Per Game), and 2P (Average Two Point Field Goals Made Per Game).

Quantitative Data Visualization and Observations:

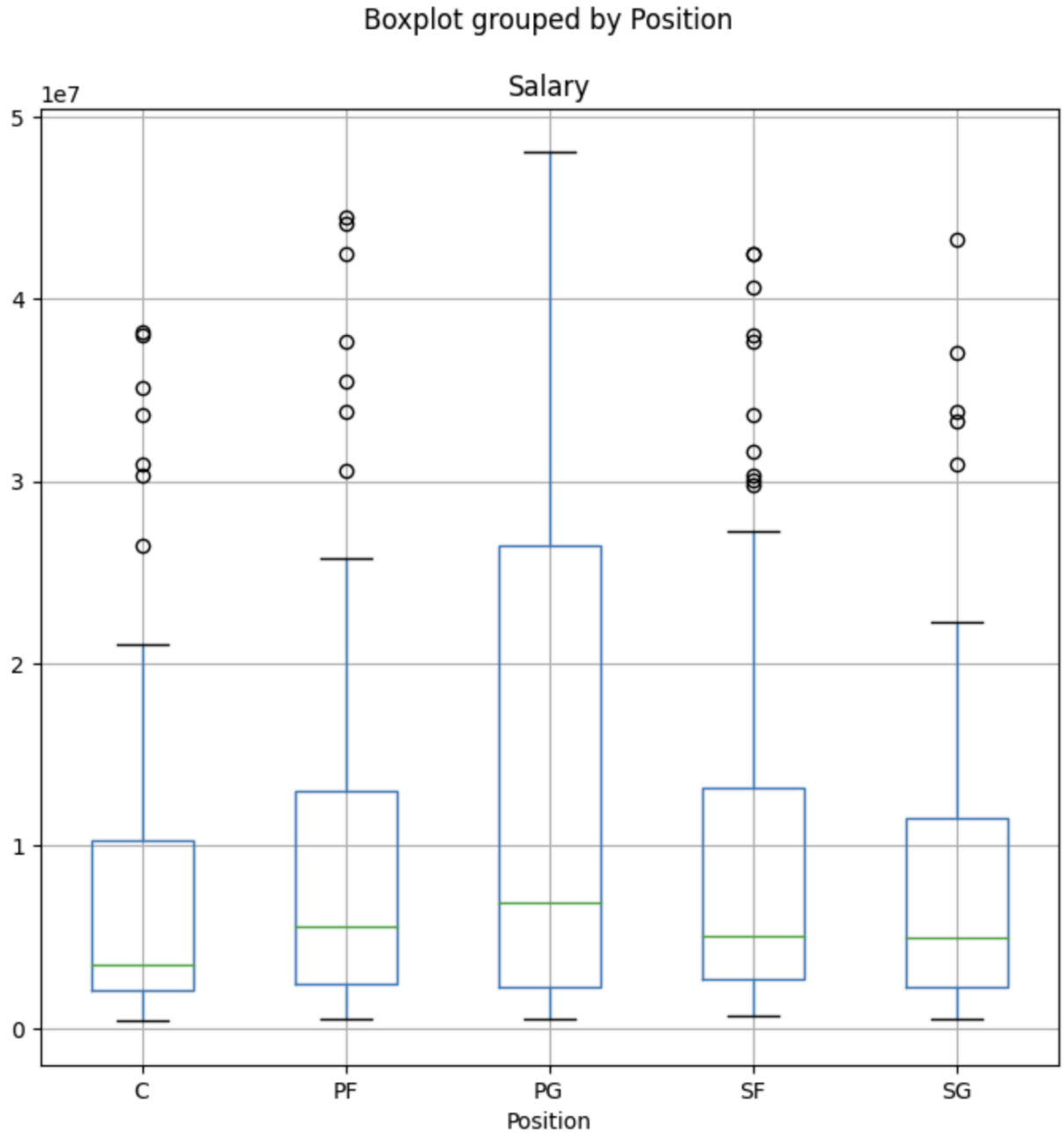


I used a scatter matrix to visualize the data. The matrix depicts the association between Salary and the quantitative variables. Salary vs MG has a positive association. Salary vs FG, 2P, PTS, and Total Minutes all have a positive association as well. The matrix can also depict the association between different variables. Total Minutes vs GP, GS, MP, FG, AST, and PTS has a strong positive association. As well, PTS vs MP, FG, 2P, and AST have a strong positive association. 3P % and 2P % have a normal distribution, while most of the other variables have a skewed distribution.

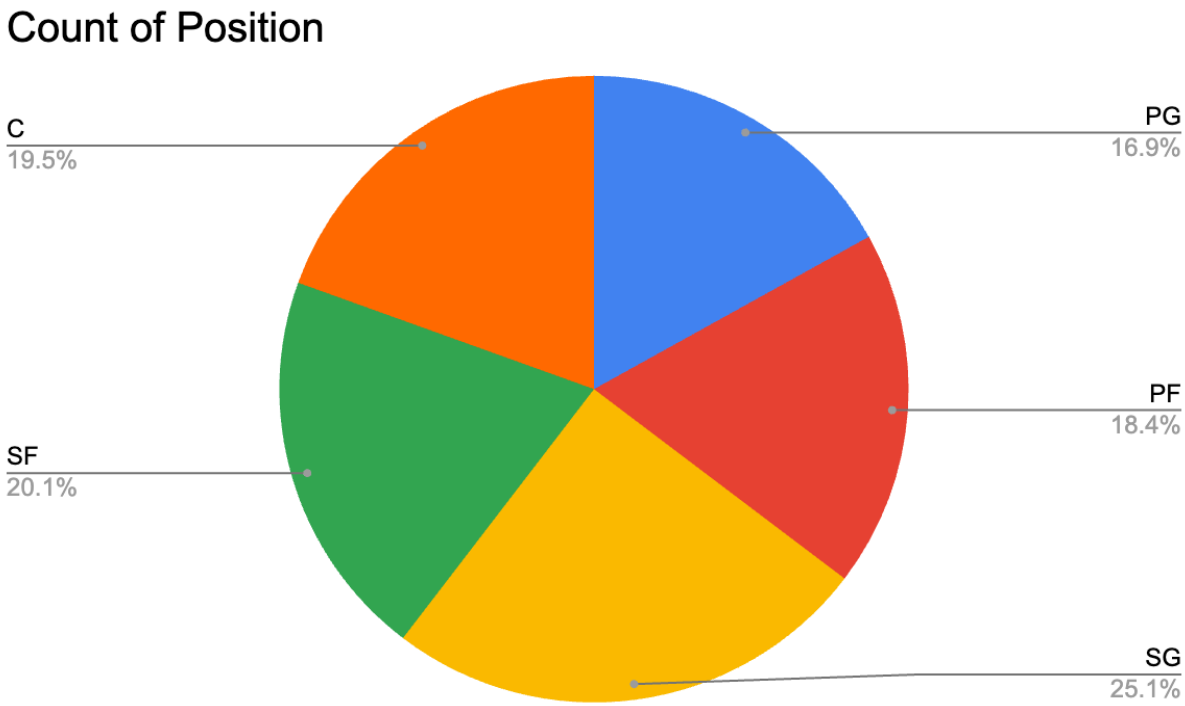


With a closer look, we can see Salary vs FG and Salary vs PTS have a similar positive association. PTS vs FG have a strong positive association. I can draw a conclusion , given FG (Average Field Goals Made Per Game) has a strong linear effect on PTS (Average Points Per Game). FG and PTS data are also right skewed like Salary. It is to be expected for salary data to be skewed.

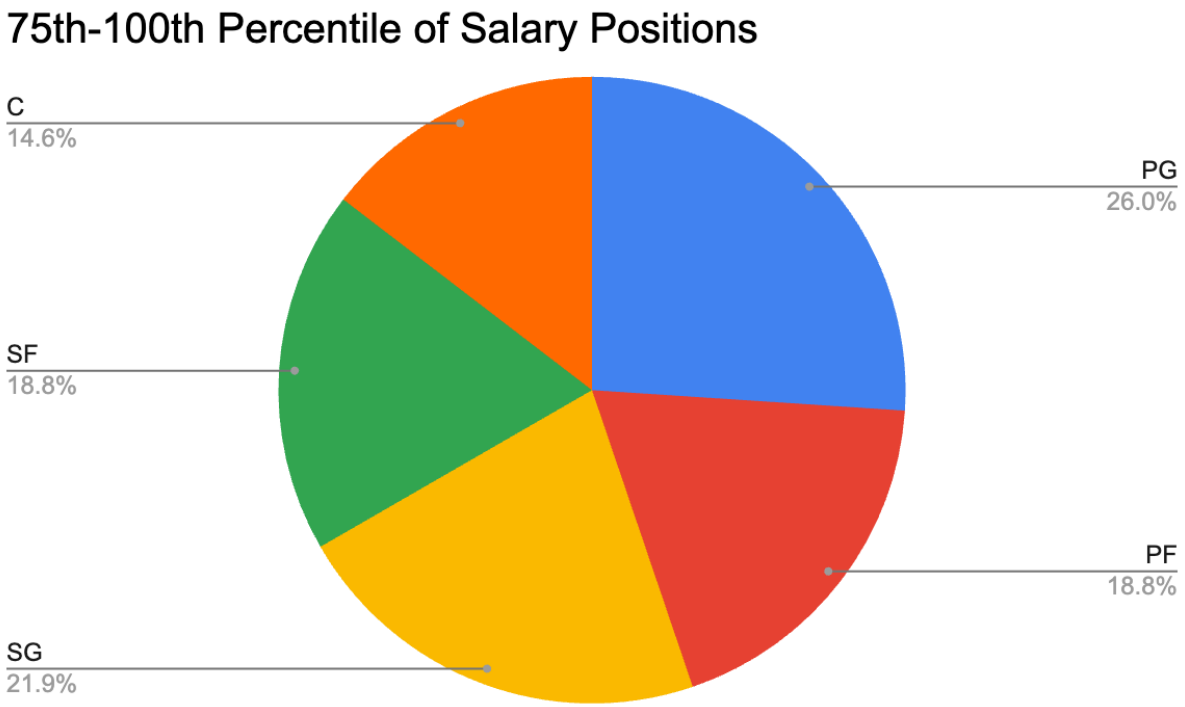
Qualitative Data Visualization and Observations:



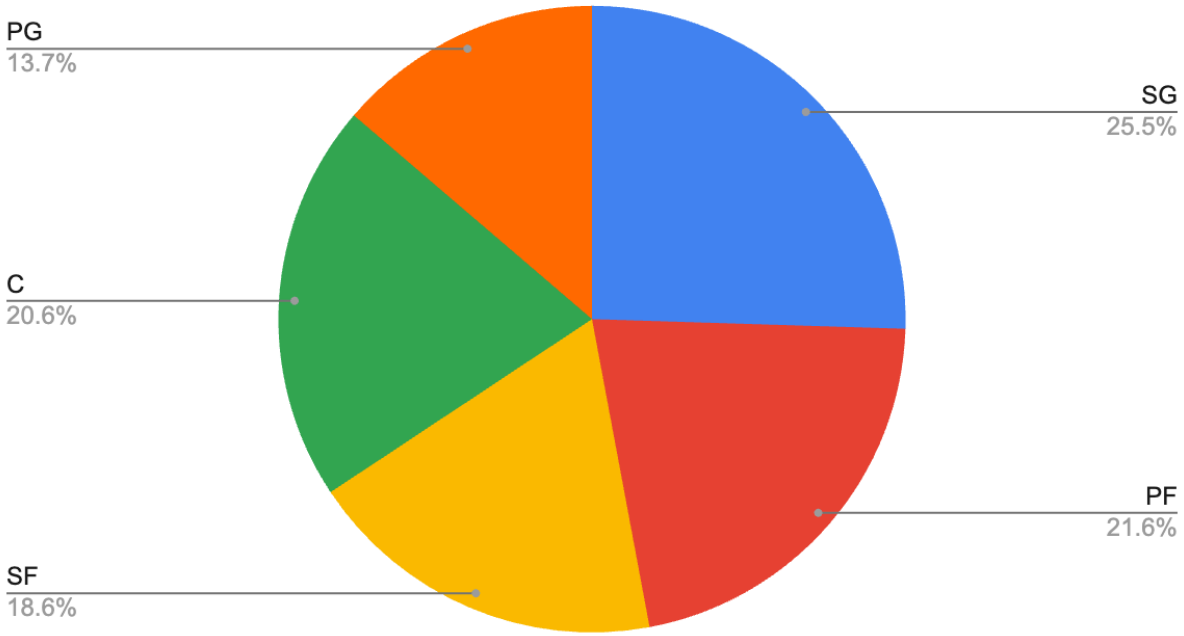
I used a boxplot to visualize the qualitative data of Salary vs Position. Based on the chart, PG (Point Guard) has the highest salaries. C (Center) positions make the lowest salaries. C, PF, SF, and SG positions have notable outliers. I graphed each 25th percentile of Salary based on position as a pie chart below.



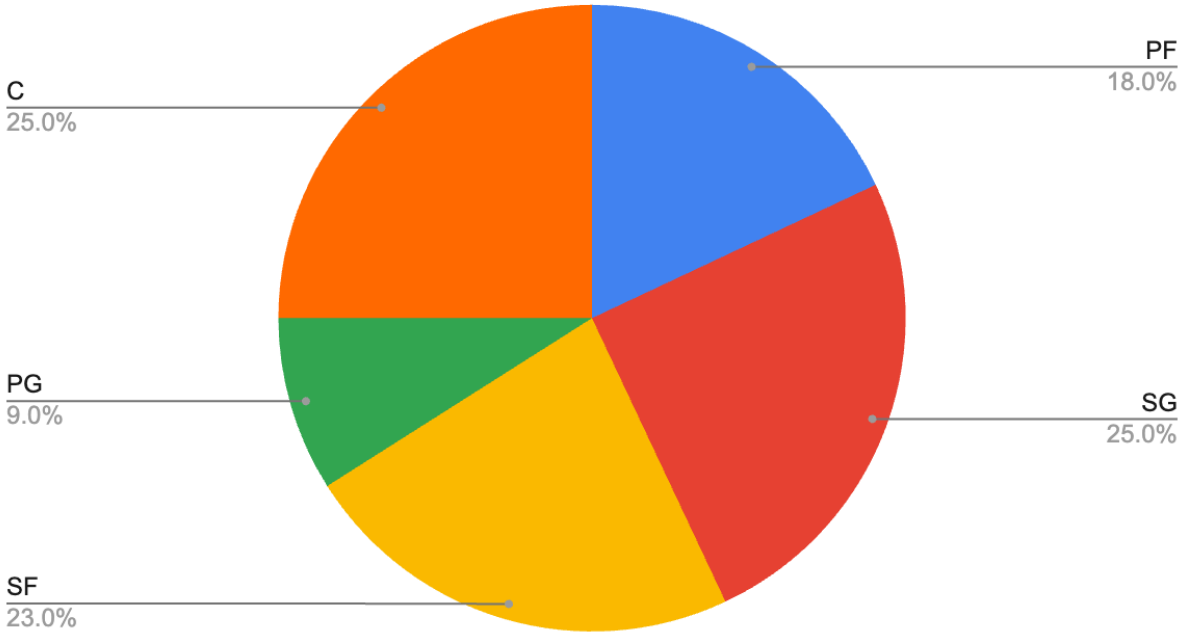
Out of 367 observations, SG is the largest position which makes up 25.1% of all salaries. PG is the smallest position which makes up 16.9.% of all salaries.

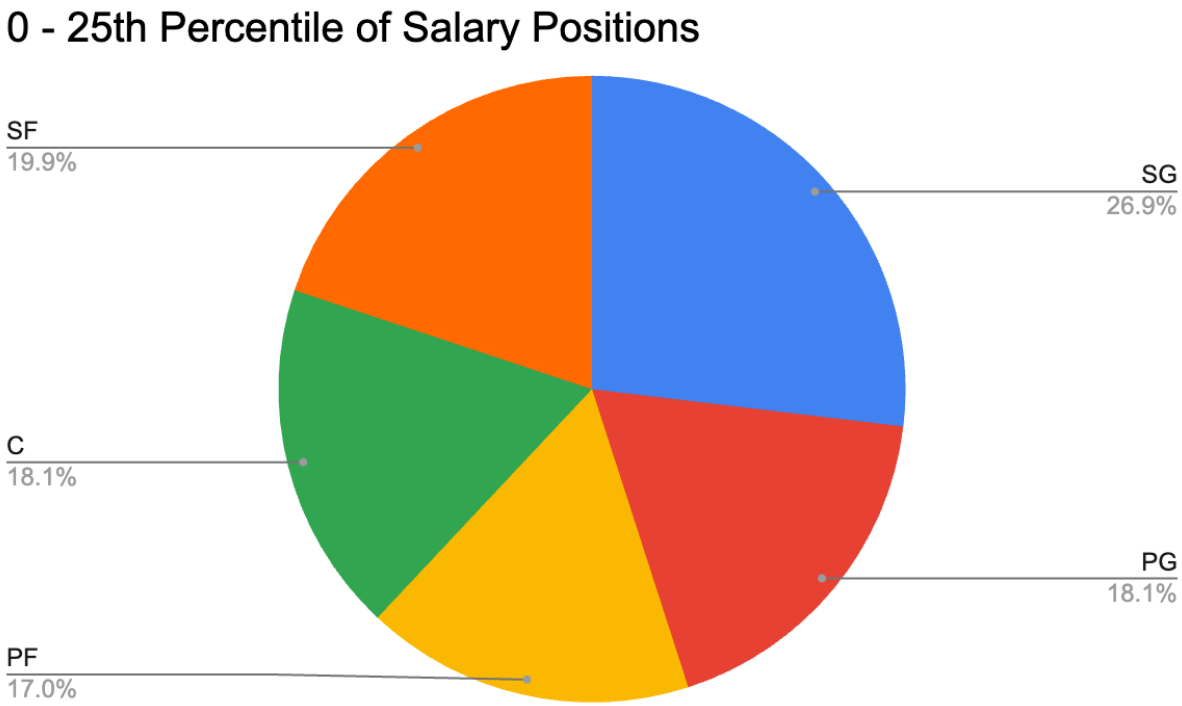


50th -75th Percentile of Salary Positions



25th - 50th Percentile of Salary Positions





The pie charts show PG represents the largest position in the higher percentiles. The Center makes up the least in the higher percentiles. SG represents the largest position in the lower percentiles and PG and PF are the lowest in the lower percentile

Analysis I:
Linear Regression with Least Squares Model:

I first started by performing a linear regression model on my NBA stats dataset. I began by changing the qualitative predictor "Position" into dummy variables. This gave me 5 new predictors; C, PF, PG, SF, and SG. I then had to clean my data again by removing any observations with NA data. The dataset observations decreased from 374 to 367. I then did a log transformation of the salary to reduce the test MSE. Then I split the dataset into a training set and a test set. I split the data in half. Since the number of observations is odd, I let the test size be 0.5 and included a random state of 0. Next, I fitted a linear regression model using least squares on the training set. logSalary is the response variable and; Age, GP, GS, MP, FG, FG%, 3P, 3P%, 2P, 2P%, AST, PTS, Total Minutes, C, PF, PG, SF, and SG are the predictors. After I summarized the results, I saw 7 predictors are significant. The significant predictors are Age, Average of Minutes Per Game (MP), and Positions (C, PF, PG, SF, SG). Next, I obtained the test error of 0.3948. The test error is extremely low, which indicates the model is well-fitted.

	coef	std err	t	P>|t|
intercept	9.6726	0.567	17.060	0.000
Age	0.0928	0.013	7.226	0.000
GP	0.0043	0.011	0.395	0.693
GS	0.0037	0.005	0.756	0.451
MP	0.0527	0.027	1.977	0.050
FG	0.3458	1.173	0.295	0.769
FG%	0.9588	1.568	0.612	0.542
3P	-0.1951	1.158	-0.168	0.866
3P%	-0.2986	0.459	-0.651	0.516
2P	-0.1431	1.159	-0.124	0.902
2P%	-1.1184	1.282	-0.872	0.384
AST	0.0032	0.051	0.062	0.951
PTS	-0.0101	0.067	-0.150	0.881
Total Minutes	-0.0003	0.001	-0.556	0.579
C	1.8668	0.214	8.743	0.000
PF	1.8796	0.160	11.722	0.000
PG	2.0252	0.172	11.772	0.000
SF	1.9718	0.154	12.837	0.000
SG	1.9292	0.134	14.431	0.000

Ridge Regression Model:

Secondly, I fitted a ridge regression model on the training set, with λ chosen by cross-validation. I used the "sklearn.linear_model" class to import RidgeCV and "sklearn.metrics" class to import mean_squared_error to fit this model. I used the default alphas that came with this class, 0.1, 1.0, 10.0. I also used a cross-validation = 5. Then I fitted ridge regression on the training X and Y. I calculated the the ridge error using mean_squared_error. The ridge test error is 0.3912. The ridge test error is less than the linear regression model with least squares. This is to be expected because ridge regression is less flexible than least squares, hence will give improved prediction accuracy.

Lasso Regression Model:

Thirdly, I fitted a lasso regression model on the training set, with λ chosen by cross-validation. I used the "sklearn.linear_model" class to import LassoCV and "sklearn.metrics" class to import mean_squared_error to fit this model. I used the default alphas that came with this class, 0.1, 1.0, 10.0. I also used a cross-validation = 5. Then I fitted ridge regression on the training X and Y. I then calculated the the lasso test error using mean_squared_error. The lasso test error for the l0.3889.The lasso test error is less than the linear regression model with least squares, and less than the ridge regression test error. This is to be expected because lasso regression is less flexible than least squares, hence will give improved prediction accuracy. The lasso regression model also had 6 non-zero coefficients.

PCR Model:

Fourthly, I fitted a PCR model on the training set, with λ chosen by cross-validation. I used the "sklearn.linear_model" class to import Linear Regression, "sklearn.decomposition" class to import PCA, " sklearn.pipeline" class to import Pipeline, "sklearn.model_selection" class to import GridSearchCV, and "sklearn.metrics" class to import mean_squared_error to fit this model. I used the default PCA components that came with this class, 1, 2, 3, 4, and 5. I also used a cross-validation = 5. Then I fitted the PCR model on the training X and Y. I then calculated the the PCR test error using mean_squared_error. The PCR test error is 0.5479 The PCR test error is significantly larger than the linear regression model with least squares. The M chosen by cross-validation is 5.

PLS Model:

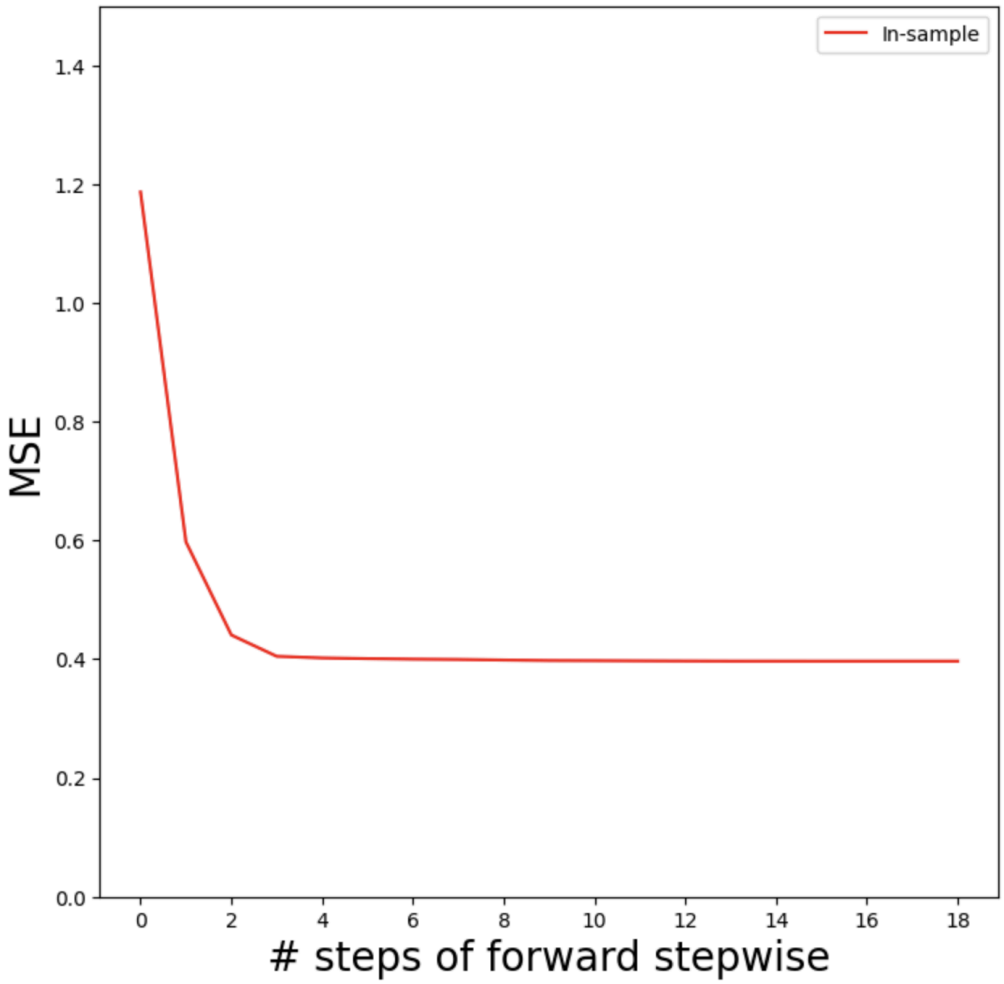
Finally, I fitted a PLS model on the training set, with λ chosen by cross-validation. I used the "sklearn.cross_decomposition" class to import PLSRegression, "sklearn.model_selection" class to import GridSearchCV, and "sklearn.metrics" class to import mean_squared_error to fit this model. I used the default grid components that came with this class, 1, 2, 3, 4, and 5. I also used a cross-validation = 5. Then I fitted the PLS model on the training X and Y. I then calculated the the PLS test error using mean_squared_error. The PLS test error is 0.4269. The PLS test error is larger than the linear regression model with least squares. The M chosen by cross-validation is 2.

Model Comparison Conclusion:

After comparing 4 different models to the least squares model, the lasso regression model has the least test error, 0.3889. The next best model is the ridge regression model with a test error of 0.3912. The test errors for Linear, PLS, and PCR respectively are 0.3948, 0.5479, and 0.4269. PCR had the worst test error. The best model for this dataset is the lasso regression model to predict salary. I am concerned with the test errors in because of how many predictors were used in this analysis. There could be a chance of overfitting. Especially since 7 out of 18 predictors were significant.

Analysis II:
Model selection using Foward Selection:

I wanted to fit the best model by selecting the best subset of variables. I used the forward stepwise selection method on the NBA salary dataset. Forward stepwise selection is a stepwise regression approach that begins with one variable and at each step gradually adds variables from the regression model to find a reduced model that best summarizes the data. Forward stepwise selection uses an information criterion (AIC) to determine which is the best. The model with the lowest AIC which produces the lowest MSE is deemed to be the best model.



After performing the forward selection, the best subset of predictors with the lowest MSE are Age, Average Minutes Played(MP), Average Field Goals Made Per Game(FG), and Three-Point Percentage(3P%). The graph above depicts the lowest MSE at step 4.

Repeat the Regression Models:

Using the new model with Age, Average Minutes Played(MP), Average Field Goals Made Per Game(FG), and Three-Point Percentage(3P%) as the predictors, I repeated the first analysis. I fitted a linear regression of the least squares model. I did a log transformation of the salary to reduce the test MSE. Then I split the dataset into a training set and a test set. I split the data in half. Since the number of observations is odd, I let the test size be 0.5 and included a random state of 0. After I summarized the results, I saw that 3 predictors are significant. The significant predictors are Age, Average Field Goals Made Per Game(FG), and Average of Minutes Per Game (MP). I obtained a test error of 0.3846, which is smaller than the MSE of the full model.

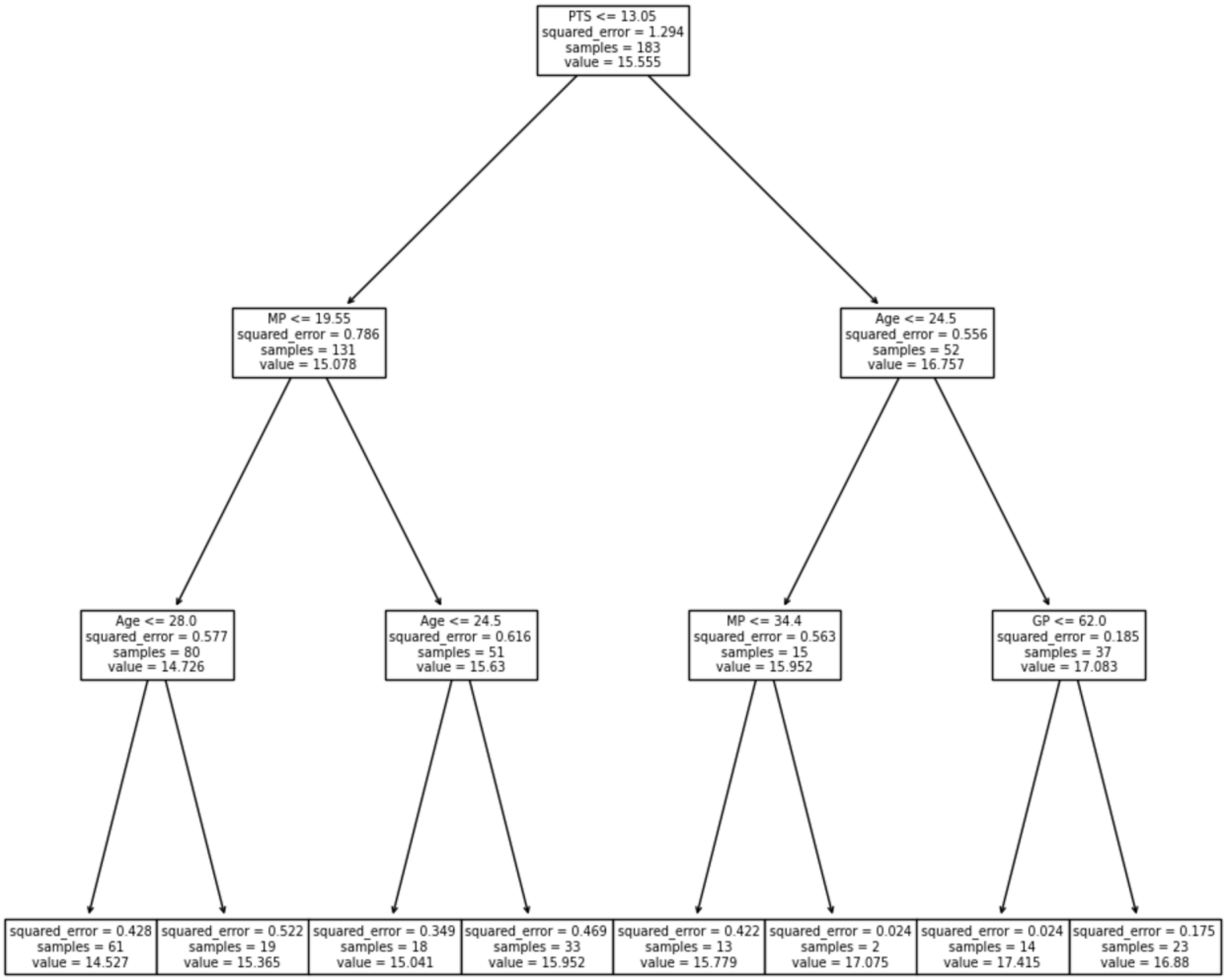
	coef	std err	t	P>|t|
intercept	11.5910	0.318	36.485	0.000
Age	0.0927	0.012	7.908	0.000
FG	0.1684	0.039	4.274	0.000
MP	0.0456	0.012	3.948	0.000
3P%	-0.3968	0.407	-0.976	0.331

For ridge regression using cross-validation, the test MSE is 0.3868. This is larger than linear regression but smaller than the full model ridge regression model. The lasso regression model has a test MSE of 0.3908, which is larger than the full model lasso regression model. The test MSE for the PCR model is 0.3845, which is significantly less than the full model. As well as, less than linear, ridge, and lasso of the reduced model. The M chosen by cross-validation is 3. Lastly, the PLS model test MSE is 0.3847, which is less than the original model. The M chosen by cross-validation is 2.

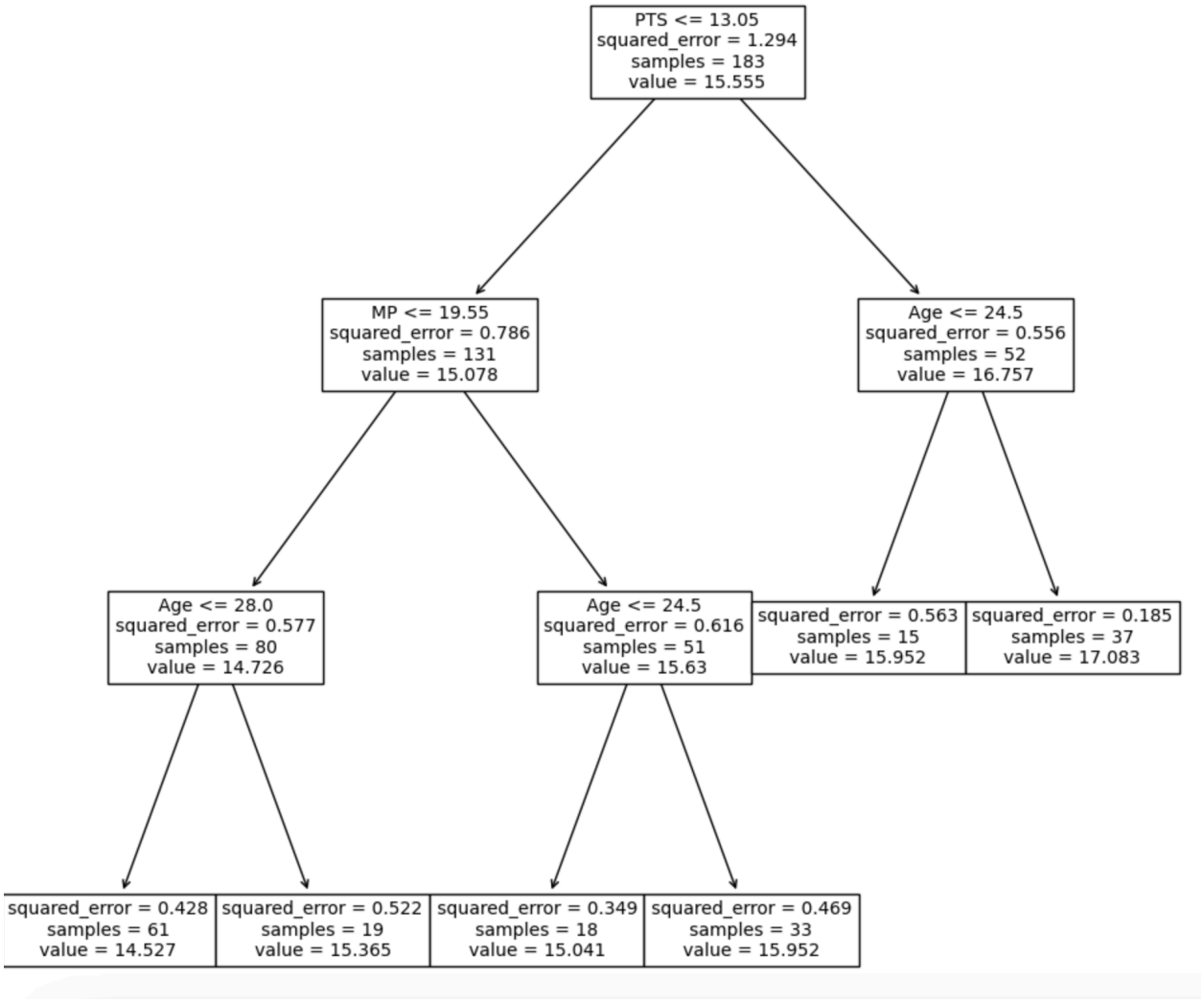
The PCR model had the smallest test MSE for the reduced model. Lasso regression had the highest test MSE. This is to be expected because of lasso performs better with more predictors and worse with fewer predictors. The opposite can be said for the PCR model.

Decision Trees:

A decision tree is a summary of a set of splitting rules used to segment the predictor space. The test MSE for the full model is 0.5648, which is significantly larger. It is even larger than the PCR Model, which is 0.5479. This is not surprising because regregression trees are notorious for not being the best model for fitting data.



To reduce the test MSE, I pruned the regression tree using cross-validation. The new Test MSE is 0.5263. This also is not an optimal model.



Conclusion:

For the NBA, there is a critical need to predict NBA salaries because of the large sum invested into the players. As well as, to maintain fairness amongst players, by the revenue each plate brings to their team. After conducting multiple analyses, the most significant factors for logSalary of the original model are Age, Average of Minutes Per Game(MP), and Positions (C, PF, PG, SF, SG). For the full model, Lasso is the best regression model with the lowest MSE. By Forward selection, the reduced model with the lowest MSE is Age, Average Minutes Played(MP), Average Field Goals Made Per Game(FG), Three-Point Percentage(3P%) The decision tree, even after pruning was not a good model to represent the data. Other methods I would like to try in the future is bagging and random forest and to see which predictors are important to the model.

Citations:

MrSimple. “Injury Prediction Dataset.” *Kaggle*, 24 Feb. 2024, www.kaggle.com/datasets/mrsimple07/injury-prediction-dataset?resource=download.