



PREDICTION OF NBA SALARIES



May 2024

Math 448

Erika Iwule



BACKGROUND

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 uses 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.





GOALS

1

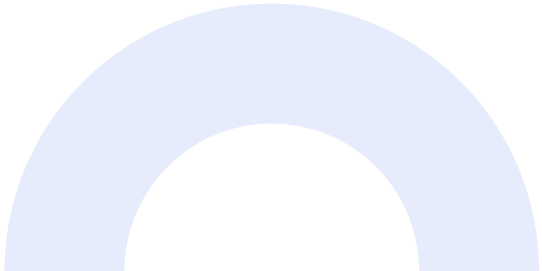

Explore and
analyze dataset
and its
explanatory
variables

2

Test different
regression
models to see
which has the
lowest MSE

3

Select which
predictors best fit
the model



THE DATASET

- 1 The dataset title is “NBA Player Salaries (2022-23 Season)”
- 2 467 observations
- 3 52 Variables
- 4 Response Variable is NBA Salaries

DATA CLEANING

- Used only 14 variables which are; Position of Player, Age, Games Played, Games Started, Minutes Per Game, Field Goals Made Per Game, Field Goal Percentage, Three-Point Field Goals Made Per Game, Three-Point Percentage, Two-Point Field Goals Made Per Game, Two-Point Percentage, Assists Per Game, Points Per Game, and Total Minutes Played
- Eliminated any players who played 20 or fewer games due to injury.
- Dropped any observations with empty data
- Changed the categorical variable “Position of Player” into dummy variables
- Changed Salary to log 10 of Salary to reduce error
- The final sample size is 367 players’ salaries and 18 explanatory variables



NBA SALARIES SUMMARY

Mean Salary: \$10,109,550

Max Salary: \$48,070,010

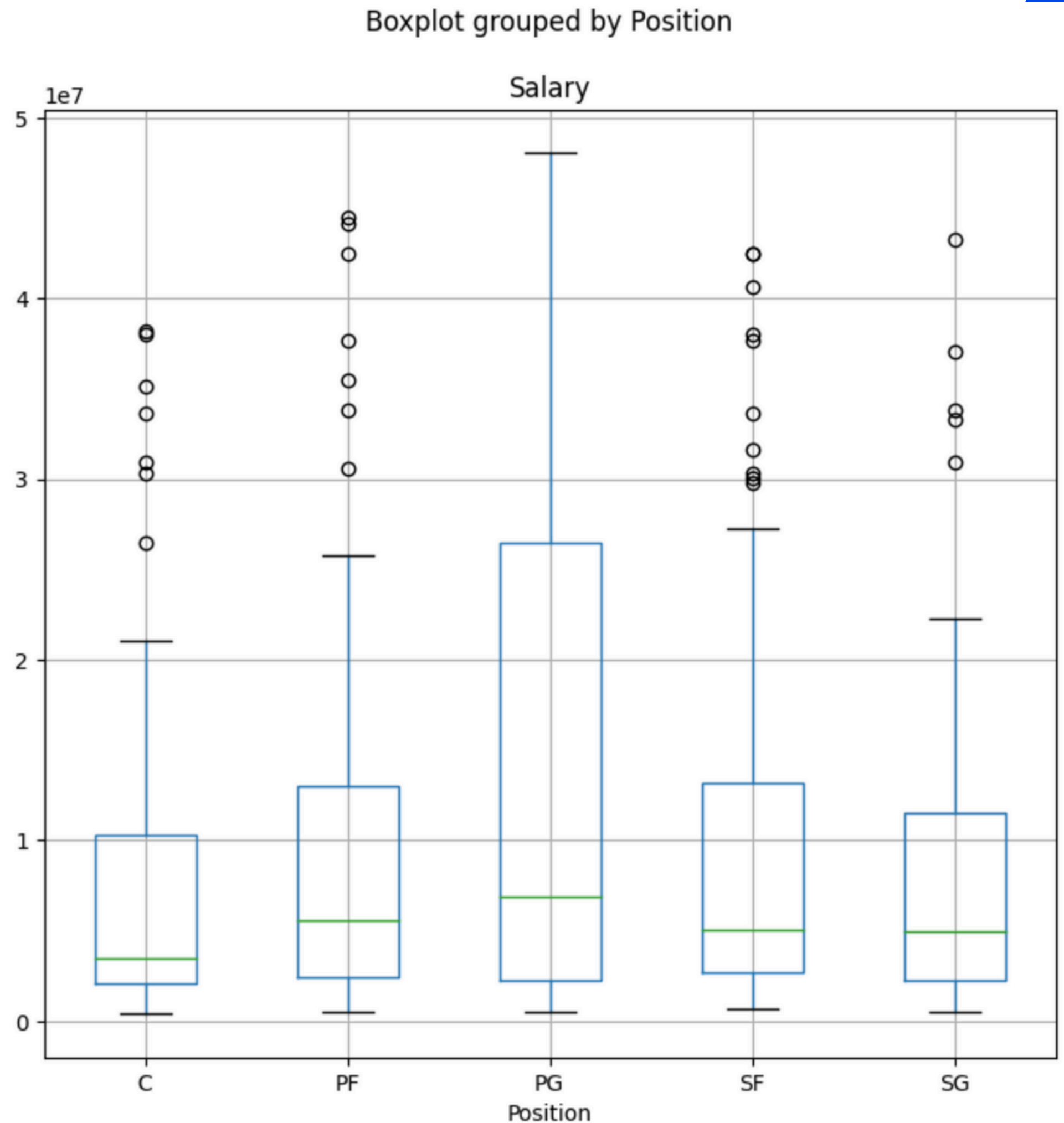
Min Salary: \$386,055

STD Salary: \$11,233,080

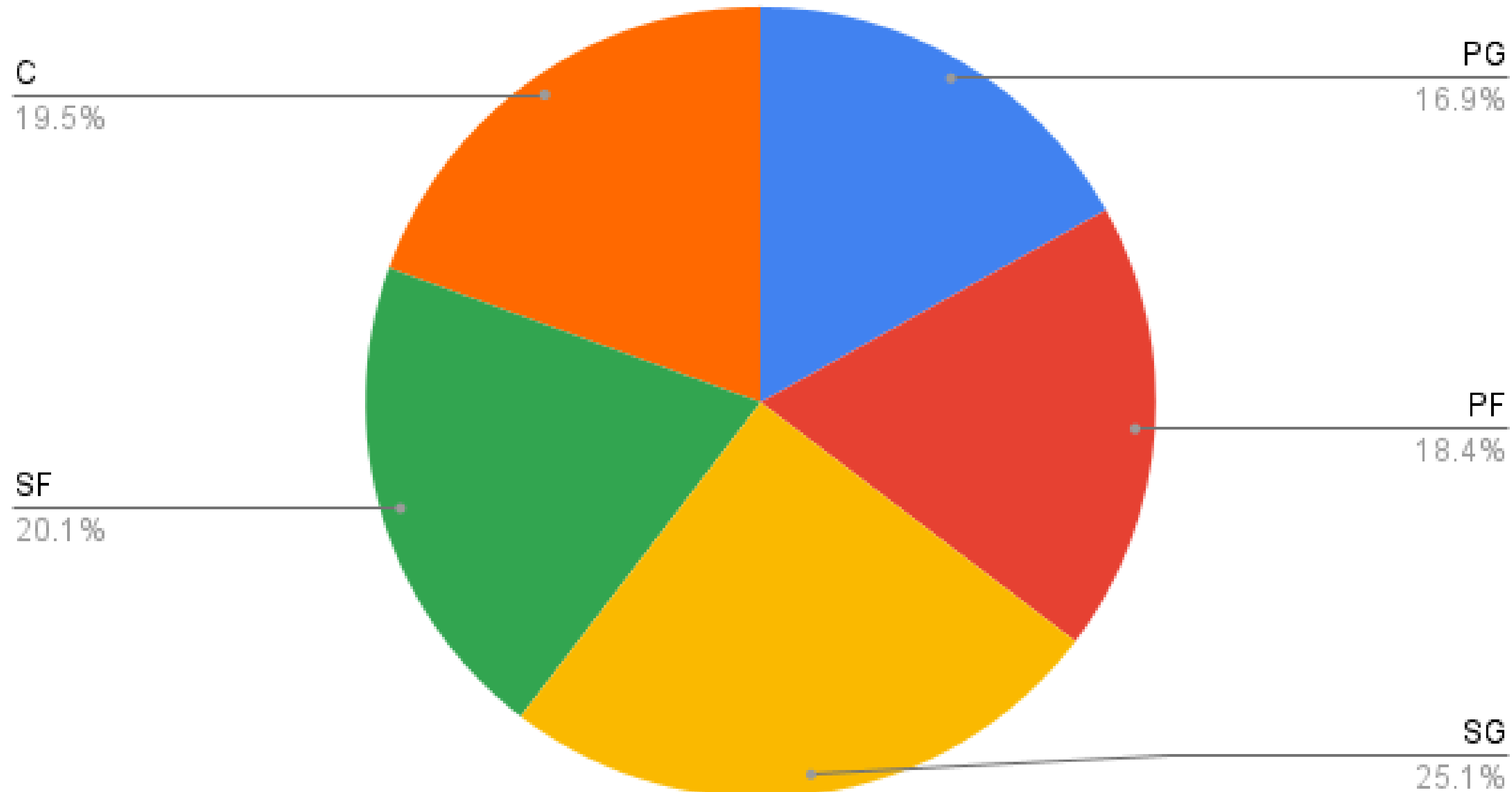


CLOSER LOOK AT THE QUALITATIVE DATA

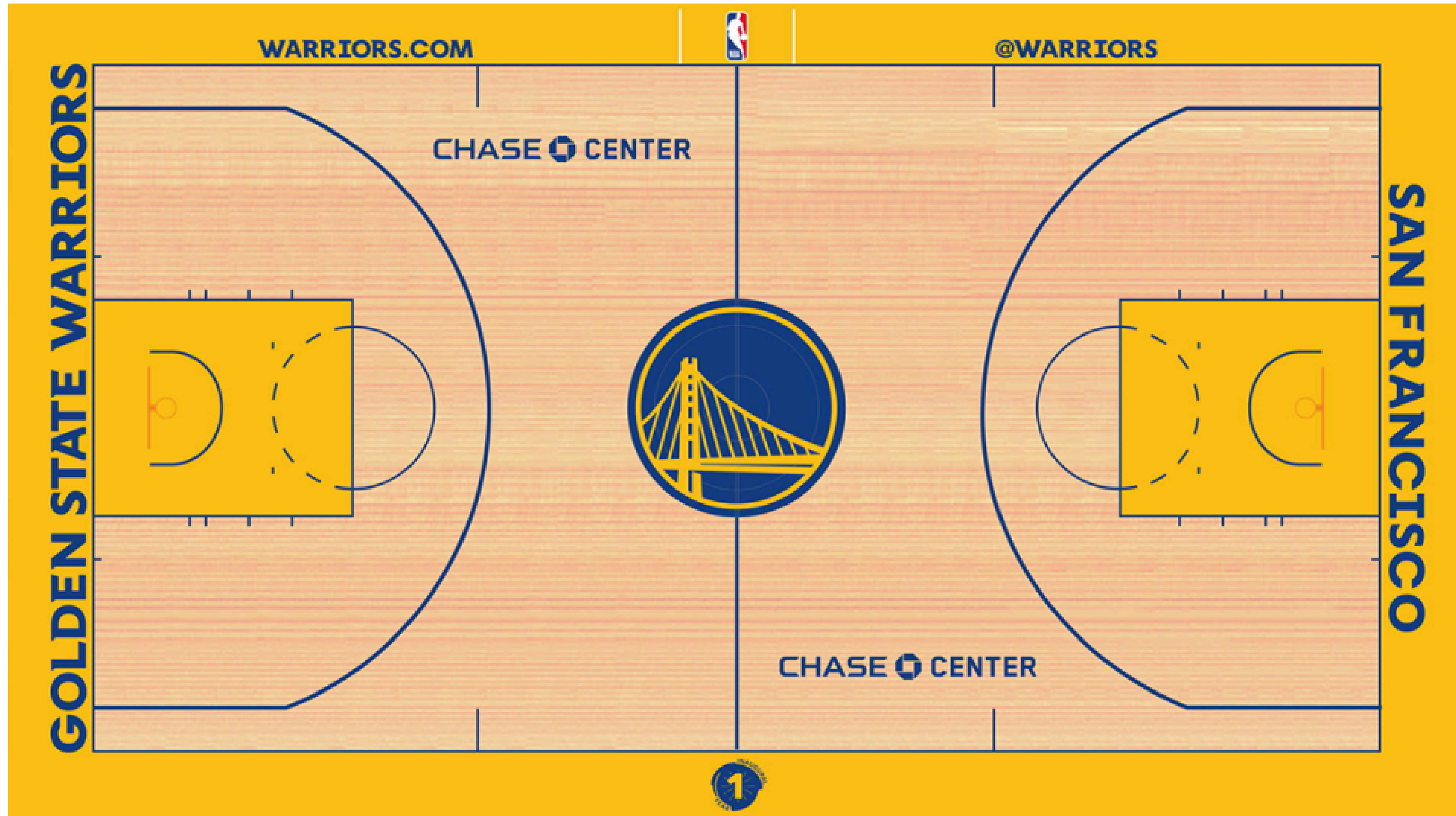
Visualization of Salary vs Position. Based on the chart, PG (Point Guard) has the highest salaries. C (Center) positions make the lowest salaries. C, PF, SF, and SG positions have notable outliers.



Count of Position



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.



REGRESSION SELECTION

LINEAR REGRESSION

- Split into training and validation sets
- Significant predictors for the effect of $\log(\text{Salary})$ are Age, Average of Minutes Per Game(MP), and Positions (C,PF,PG,SF,SG).
- Test MSE = 0.3948

	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

OTHER REGRESSION MODELS

Compared to the Linear Regression MSE = 0.3948, Lasso had the smallest MSE and PCR Model had the highest MSE

RIDGE REGRESSION

Test MSE =
0.3911

LASSO REGRESSION

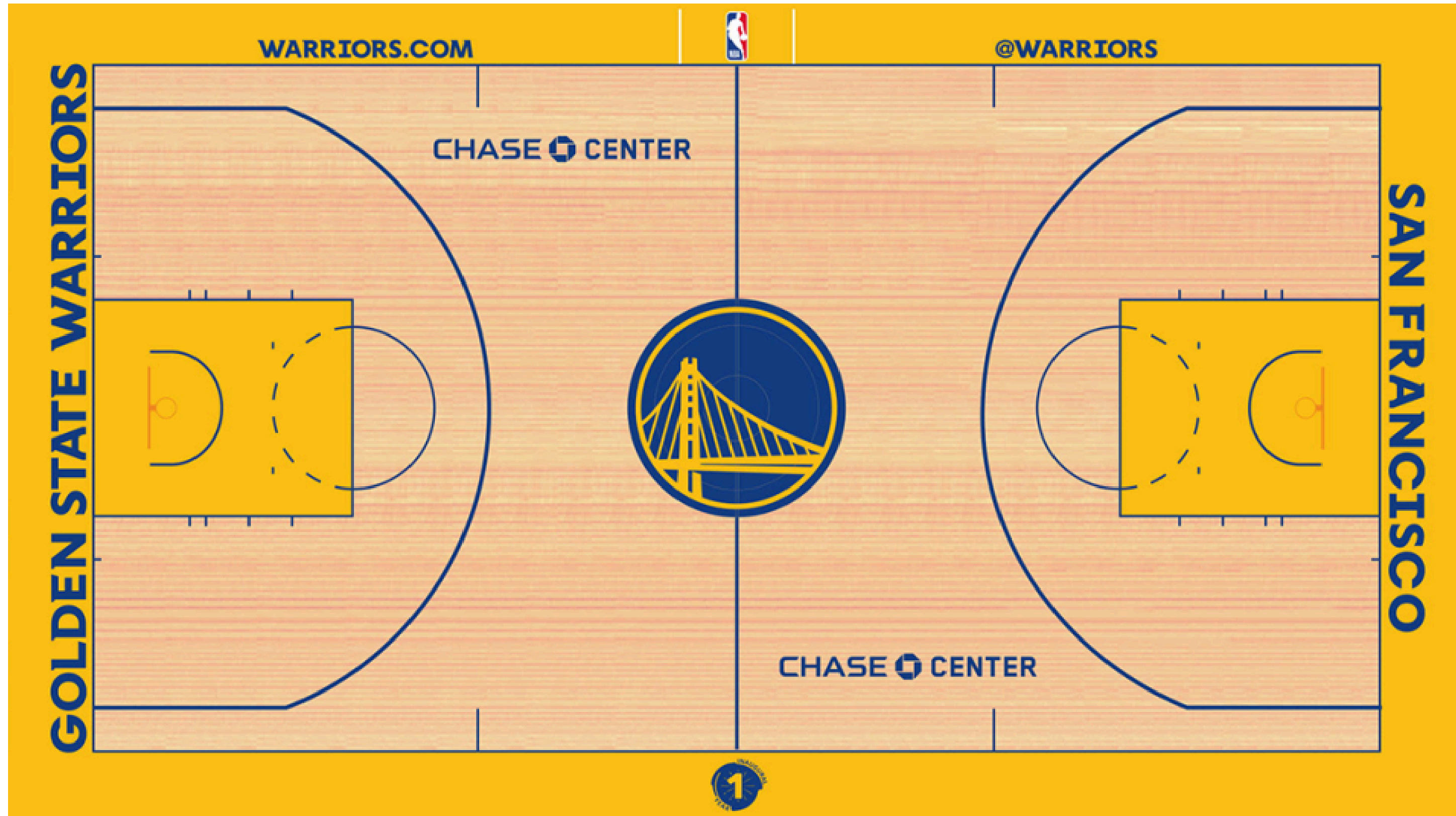
Test MSE =
0.3889

PCR MODEL

Test MSE =
0.5479

PLS MODEL

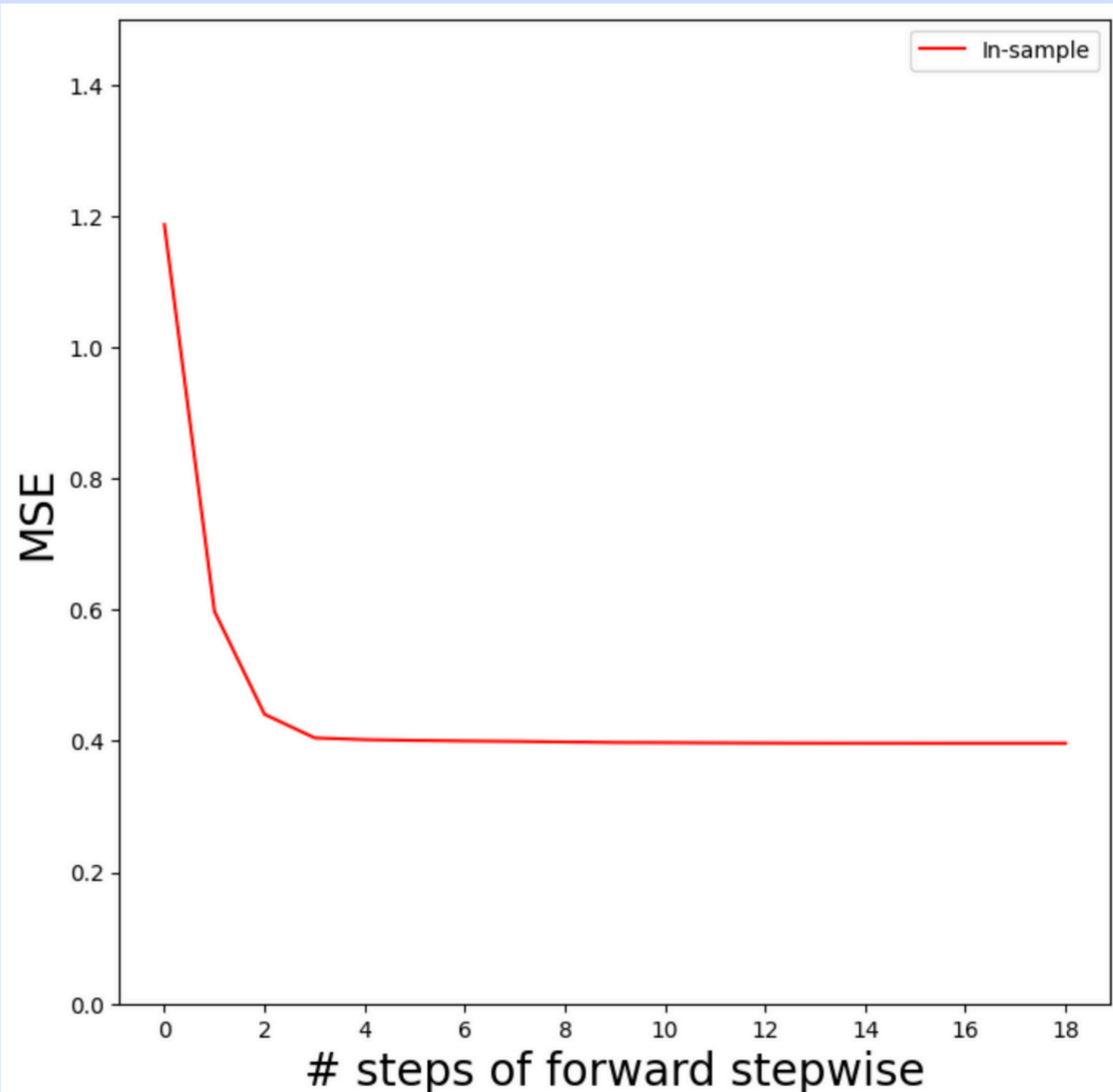
Test MSE =
0.4269



SUBSET SELECTION

FORWARD SELECTION

Using forward selection, the best subset of predictors with the lowest MSE are Age, Average Minutes Played(MP), Average Field Goals Made Per Game(FG), Three-Point Percentage(3P%)



LINEAR REGRESSION

- Split into training and validation sets
- Significant predictors for the effect of $\log(\text{Salary})$ are Age, Average of Minutes Per Game(MP), Average Field Goals Made Per Game(FG),
- Test MSE = 0.3846

	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

OTHER REGRESSION MODELS

Compared to the Linear Regression $\text{MSE} = 0.3846$, PCR had the smallest MSE and Lasso had the highest MSE

RIDGE REGRESSION

Test MSE =
0.3868

LASSO REGRESSION

Test MSE =
0.3908

PCR MODEL

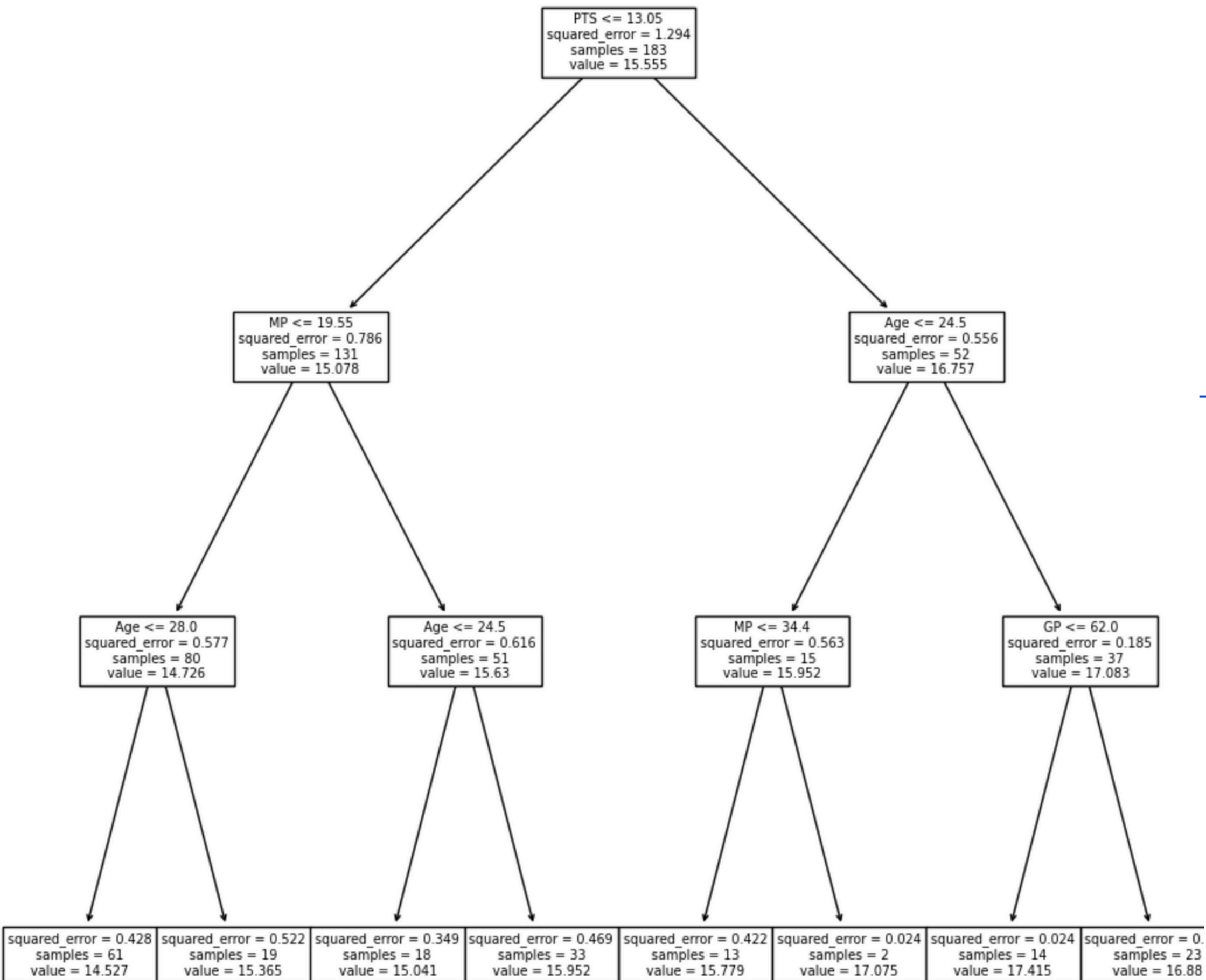
Test MSE =
0.3845

PLS MODEL

Test MSE =
0.3847

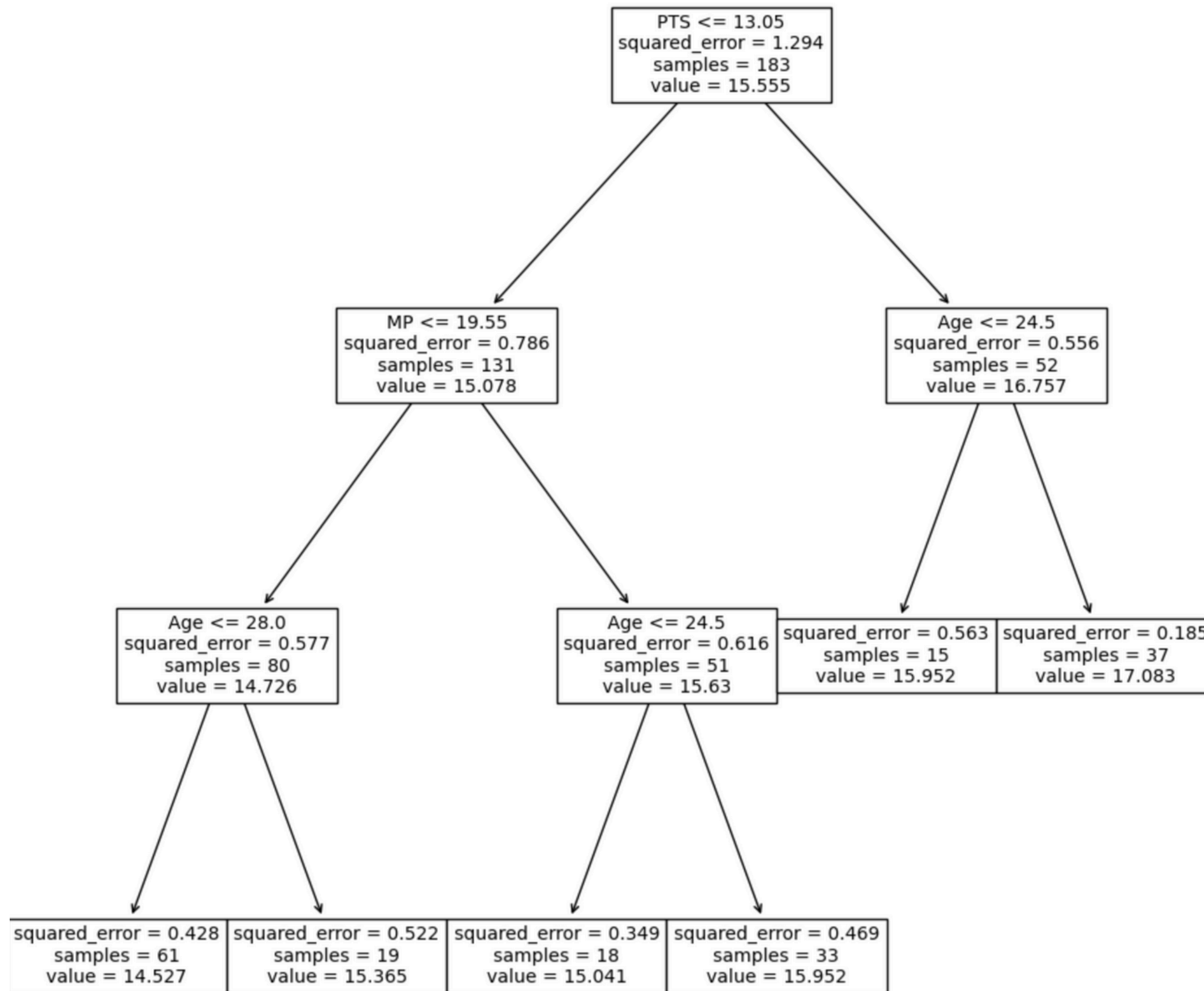
DECISION TREE

Test MSE = 0.5755



PRUNING USING CROSS- VALIDATION

Test MSE = 0.5263



CONCLUSIONS

1. Significant predictors for the effect of $\log(\text{Salary})$ are Age, Average of Minutes Per Game(MP), and Positions (C, PF, PG, SF, SG). For the saturated model.

2. For the saturated model, Lasso is the best regression model with the lowest MSE.

3. By Forward selection, the model with the lowest MSE are Age, Average Minutes Played(MP), Average Field Goals Made Per Game(FG), Three-Point Percentage(3P%)

4. The decision tree, even after pruning was not a good model to represent the data.



THANK YOU!