ANALYSIS AND FORECASTING OF NBA SALARIES

This report examines NBA salary trends, essential for team management, budgeting, and player recruitment. By analyzing historical salary data alongside player performance metrics, we aim to uncover the factors driving player compensation. Our statistical analysis and forecasting are designed to reveal patterns and predict future trends, offering valuable insights for strategic decision-making in NBA team management.

Major Findings:

Correlation Confirmation:

- There is a strong positive correlation (r = 0.7276) between game impact metrics and player salaries, indicating that players with higher on-court impact scores tend to receive higher salaries.

Significant Predictors:

- Age and Points Scored (PTS) are significant predictors of salary levels, with Importance values of 6.79e+15 and 6.34e+15, respectively. This suggests that younger players with high scoring records are likely to see salary increases.

Factors Influencing Salary:

- The first rotated component (RC1) of factor analysis reveals that factors such as Field Goals (FG), Field Goal Attempts (FGA), 2-Point Field Goals (2P), 2-Point Field Goal Attempts (2PA), Free Throws (FT), Free Throw Attempts (FTA), Turnovers (TOV), Points Scored (PTS), and Usage Percentage (USG%) significantly affect player salaries.

Time Series Trends:

- Time series analysis using ARIMA models predicts a downward salary trajectory for top-performing players, with some exceptions. This indicates

that while there is a general trend, other factors might influence deviations from this pattern.

Model Accuracy:

- The Random Forest regression model explains 72.36% of the variance in salaries, indicating a strong fit. The Mean Squared Residuals of 3.38e+13 further confirm the reliability of these salary predictions.

Additional Insights:

- The model's ability to explain a substantial portion of the salary variance suggests that it can be a valuable tool for teams in salary negotiations and budget planning.
- The deviation in some salary forecasts highlights the potential impact of external factors such as market dynamics, player popularity, and team-specific financial strategies.

RECOMMENDATIONS:

- **Performance Incentives:** Align contracts with player productivity, focusing on Points scored (PTS importance: 6.5e+15). The model explains 72.36% variance in salaries, suggesting scoring significantly impacts earnings.
- **Youth Investment:** Invest in young talent, with Age showing a strong trend in salary predictions (importance: 6.93e+15). Forecasts indicate rising salaries for emerging players.
- **Positional Salary Strategy:** Allocate salary caps strategically by position to maintain financial balance. Regression analysis underscores different salary impacts by position.
- **Smart Contract Negotiations:** Consider peak salary age in contract lengths to optimize value and prevent overcommitment, informed by Age's significant role in salary prediction (importance: 6.93e+15).

ANALYTICAL OVERVIEW

The "Analysis and Forecasting of NBA Salaries" project examines the relationship between NBA players' salaries and performance metrics. Using data from 467 players, it identifies a significant positive correlation between game impact and salary, with age and points scored as notable predictors. The analysis employs linear and Random Forest Regression, yielding high R-squared values, indicating strong model accuracy. Key insights include the importance of aligning salaries with player performance and strategically considering age and role in salary decisions.

APPENDIX

- **Data Cleaning**

The data cleaning process was crucial to ensure the integrity of the dataset used in the analysis. After gathering the data, the first step involved addressing missing values. For this, mean imputation was employed, where missing entries were replaced with the mean of the respective variable. This approach helped maintain the dataset's overall statistical properties. Additionally, standardization procedures were applied to harmonize different formats in player names, team names, and performance metrics. The dataset was thoroughly checked for outliers and inconsistencies, ensuring that it accurately reflected real-world data.

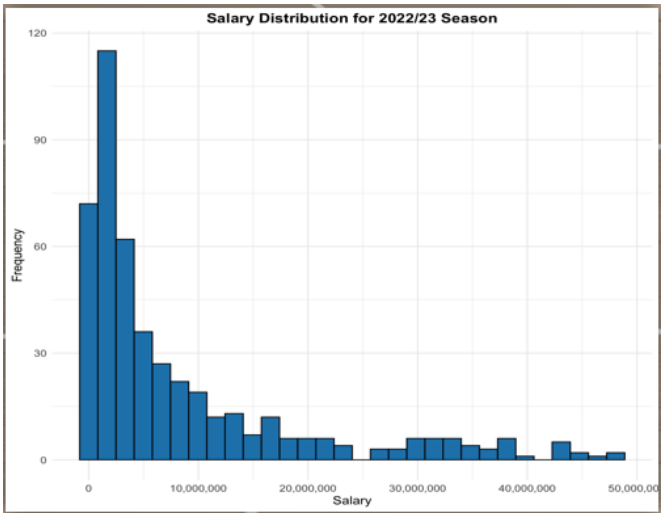
- **Data Scraping and Sources**

The dataset for the 2022-23 NBA season was sourced from Kaggle, providing a structured and comprehensive set of salary and performance data. For the preceding three years (2019-2022), the data was scraped from https://hoopshype.com/salaries/players/. This website offered detailed salary

information for NBA players, which was essential for the longitudinal analysis of salary trends.

Exploratory Data Analysis

Salary Distribution for 2022/23 Season

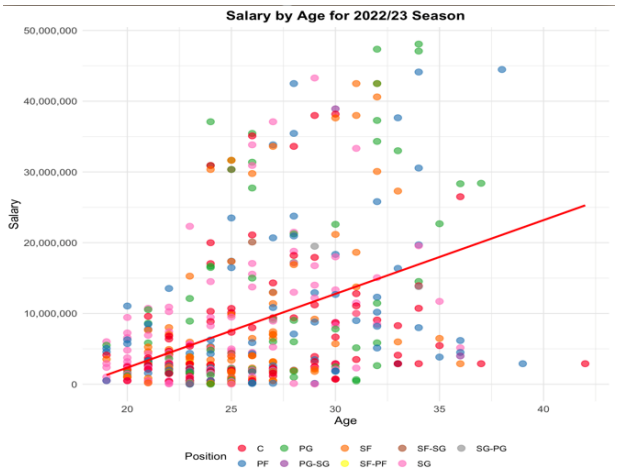


- The majority of the salaries fall between $0 million and $20 million. The highest frequency, indicating the most common salary range, is around $10 million.
- There is a sharp drop in frequency after $20 million, indicating that fewer individuals earn salaries in this range.
- There are very few salaries above $40 million, suggesting that such high salaries are not common.

Boxplot of Salary Distribution for 2022/23 Season

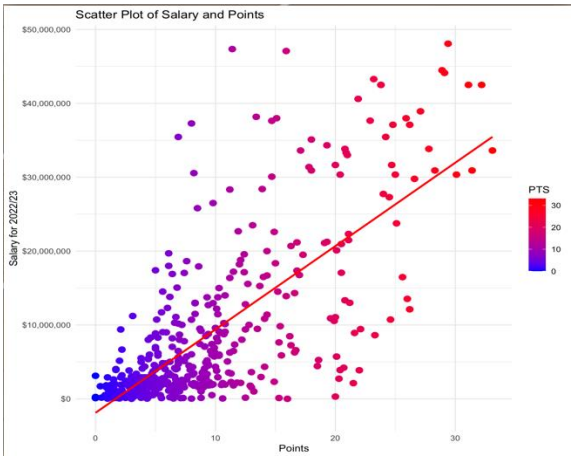
- The line in the middle of the box indicates the median salary, which is around $20 million. This means that half of the salaries are above $20 million and half are below.
- The bottom line of the box represents the lower quartile (25th percentile), which is around $10 million. This means that 25% of the salaries are below $10 million.
- The top line of the box represents the upper quartile (75th percentile), which is around $30 million. This means that 75% of the salaries are below $30 million.

<u>Scatterplot of Salary by age for the 2022/23 season</u>



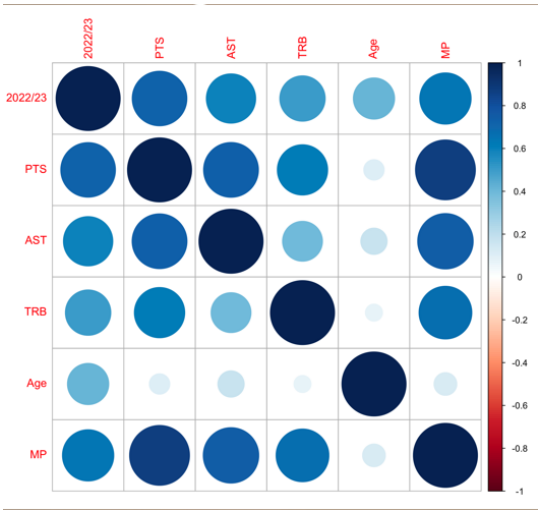
- The scatterplot shows the relationship between age and salary. The red line is a linear regression line, which suggests a positive relationship between age and salary. This means that as age increases, so does the salary.
- The data points are color-coded by position. For example, centers are in blue, power forwards (PF) are in pink, point guards (PG) are in green, small forwards (SF) are in purple, shooting guards (SG) are in orange, and so on. This allows us to see if there's a pattern or difference in salary based on the position.
- The scatterplot allows us to see the distribution of salaries across different ages and positions. For example, we can see if certain positions tend to have higher salaries, or if salaries increase with age.

Scatterplot for Salary by Points for the 2022/23 season



- The scatterplot shows a positive correlation between salary and points. This means that as the salary increases, the points also increase. The red line of best fit indicates this positive trend.
- The majority of individuals have a salary between $0 million and $20 million. This is where the highest concentration of data points is found.
- The points for these individuals mainly range between 0 and 20. This suggests that individuals with salaries in this range tend to score between 0 and 20 points.

Correlation Plot



- The plot shows a strong positive correlation between PTS and AST, TRB and Age, and MP and Age. This means that as one variable increases, the other also increases.
- The plot shows a weak negative correlation between PTS and Age, and AST and MP. This means that as one variable increases, the other decreases.
- The plot shows no correlation between TRB and MP. This means that there is no relationship between these two variables.

Factor Analysis

The factor analysis conducted on the NBA salary data aims to identify underlying patterns in the dataset by reducing the number of variables into a smaller set of factors or components.

In this analysis, Principal Component Analysis (PCA) was used, a common method of factor analysis. PCA seeks to transform the original variables into a new set of uncorrelated variables, called principal components, which are linear combinations of the original variables. These components are ordered so that the first few retain most of the variation present in all the original variables.

```r
str(data)

variables_of_interest <- data[, c('Salary','Age', 'GP', 'GS', 'MP', 'FG', 'FGA', 'FG%', '3P', '3PA', '3P%', '2P', '2PA', '2P%',
                                  'eFG%', 'FT', 'FTA', 'FT%', 'ORB', 'DRB', 'TRB', 'AST', 'STL', 'BLK', 'TOV',
                                  'PF', 'PTS', 'Total Minutes', 'PER', 'TS%', '3PAr', 'FTr', 'ORB%', 'DRB%',
                                  'TRB%', 'AST%', 'STL%', 'BLK%', 'TOV%', 'USG%', 'OWS', 'DWS', 'WS', 'WS/48',
                                  'OBPM', 'DBPM', 'BPM', 'VORP')]

# Remove missing values
variables_of_interest <- na.omit(variables_of_interest)

# Ensure that your data is numeric
variables_of_interest <- as.data.frame(variables_of_interest)

# Ensure that your data is numeric
variables_of_interest <- as.matrix(variables_of_interest)

# Perform factor analysis using Principal Component Analysis (PCA)
factor_analysis_result <- psych::principal(variables_of_interest, nfactors = 5, rotate = "varimax")

# Print factor analysis results
print(factor_analysis_result)
```

```
> print(factor_analysis_result)
Principal Components Analysis
Call: psych::principal(r = variables_of_interest, nfactors = 5, rotate = "var
imax")
Standardized loadings (pattern matrix) based upon correlation matrix
                RC1   RC5   RC2   RC4   RC3    h2    u2 com
Salary         0.66  0.36 -0.04  0.05  0.08 0.584 0.416 1.6
Age            0.06  0.13 -0.14  0.12  0.07 0.059 0.941 3.9
GP             0.17  0.78 -0.08  0.14  0.02 0.657 0.343 1.2
GS             0.51  0.71 -0.03  0.04  0.03 0.765 0.235 1.8
MP             0.62  0.71 -0.13  0.08  0.00 0.916 0.084 2.1
FG             0.85  0.45 -0.04  0.15  0.02 0.952 0.048 1.6
FGA            0.85  0.44 -0.17  0.04 -0.02 0.947 0.053 1.6
FG%            0.08  0.09  0.41  0.86 -0.02 0.926 0.074 1.5
3P             0.46  0.42 -0.61  0.09 -0.06 0.784 0.216 2.8
3PA            0.50  0.42 -0.60  0.00 -0.07 0.788 0.212 2.8
3P%            0.05  0.04 -0.48  0.46 -0.01 0.447 0.553 2.0
2P             0.84  0.37  0.22  0.15  0.05 0.920 0.080 1.6
2PA            0.88  0.36  0.14  0.06  0.02 0.921 0.079 1.4
2P%           -0.01  0.08  0.24  0.74  0.07 0.616 0.384 1.2
eFG%           0.00  0.15  0.10  0.96 -0.03 0.948 0.052 1.1
FT             0.89  0.24  0.03  0.09  0.12 0.866 0.134 1.2
FTA            0.88  0.26  0.12  0.09  0.11 0.875 0.125 1.3
FT%            0.18  0.08 -0.46  0.01  0.10 0.259 0.741 1.5
ORB            0.13  0.48  0.73  0.15  0.10 0.817 0.183 1.9
DRB            0.52  0.64  0.38  0.13  0.07 0.844 0.156 2.7
TRB            0.43  0.64  0.53  0.15  0.09 0.896 0.104 2.9
AST            0.76  0.28 -0.23 -0.03  0.03 0.719 0.281 1.5
STL            0.43  0.50 -0.19 -0.09  0.25 0.550 0.450 2.9
BLK            0.13  0.47  0.54  0.11  0.28 0.619 0.381 2.8
TOV            0.85  0.31 -0.01  0.04 -0.08 0.837 0.163 1.3
PF             0.36  0.64  0.19  0.15 -0.04 0.589 0.411 1.9
PTS            0.86  0.43 -0.10  0.14  0.03 0.966 0.034 1.6
Total Minutes  0.46  0.80 -0.12  0.08  0.02 0.881 0.119 1.7
PER            0.49  0.12  0.15  0.43  0.68 0.926 0.074 2.8
TS%            0.10  0.12  0.05  0.94  0.17 0.935 0.065 1.1
3PAr          -0.25  0.04 -0.75 -0.21  0.04 0.675 0.325 1.4
FTr            0.29 -0.18  0.34  0.18  0.42 0.446 0.554 3.6
ORB%          -0.16  0.03  0.81  0.09  0.04 0.701 0.299 1.1
DRB%           0.05  0.14  0.69  0.16  0.02 0.524 0.476 1.2
TRB%          -0.04  0.11  0.85  0.16  0.04 0.770 0.230 1.1
AST%           0.70  0.00 -0.25 -0.12  0.03 0.565 0.435 1.3
STL%          -0.02 -0.11 -0.12 -0.26  0.82 0.759 0.241 1.3
BLK%          -0.10 -0.04  0.36 -0.06  0.74 0.693 0.307 1.5
```

```
TOV%            0.04 -0.15  0.23  0.04 -0.32 0.177 0.823 2.4
USG%            0.83 -0.12 -0.04 -0.02  0.04 0.715 0.285 1.1
OWS             0.53  0.51  0.12  0.27  0.19 0.662 0.338 2.9
DWS             0.39  0.80  0.12  0.09  0.14 0.830 0.170 1.6
WS              0.52  0.66  0.13  0.22  0.19 0.810 0.190 2.4
WS/48           0.18  0.24  0.07  0.54  0.71 0.882 0.118 2.3
OBPM            0.50  0.27 -0.13  0.44  0.55 0.835 0.165 3.5
DBPM           -0.08  0.07  0.02  0.08  0.90 0.821 0.179 1.0
BPM             0.33  0.23 -0.09  0.35  0.81 0.946 0.054 1.9
VORP            0.67  0.41  0.07  0.16  0.22 0.697 0.303 2.1

                       RC1  RC5  RC2  RC4  RC3
SS loadings          12.35 7.73 5.88 4.72 4.63
Proportion Var        0.26 0.16 0.12 0.10 0.10
Cumulative Var        0.26 0.42 0.54 0.64 0.74
Proportion Explained  0.35 0.22 0.17 0.13 0.13
Cumulative Proportion 0.35 0.57 0.74 0.87 1.00

Mean item complexity =  1.9
Test of the hypothesis that 5 components are sufficient.

The root mean square of the residuals (RMSR) is  0.05
 with the empirical chi square  2937.9  with prob <  1.2e-214

Fit based upon off diagonal values = 0.98
```

Obeservations:

'Salary' has a strong loading on RC1 (0.66) and a moderate loading on RC5 (0.36), suggesting its importance in these components. RC1, with high loadings for 'FG', 'FGA', '2P', '2PA', 'FT', 'FTA', 'PTS', and 'TOV', seems to represent scoring ability and offensive play. RC5, with significant loadings on 'Total Minutes' and 'DWS' might represent overall court presence and defensive wins.

The proportion of variance explained by each component and their cumulative proportions indicate the effectiveness of the PCA in capturing the dataset's variability. The high fit based on off-diagonal values (0.98) and the low RMSR (0.05) suggest a good model fit, indicating that these components effectively summarize the complex relationships in the data.

Regression Testing for Factors:

```r
# Loading the dataset
data <- read_excel("D:/UCD/BANA - Stats/Final Project/Regression_Book2.xlsx")

# Preparing the data
independent_vars <- c('Age', 'GP', 'GS', 'MP', 'FG', 'FGA', 'FG%', '3P', '3PA', '3P%', '2P', '2PA', '2P%',
                      'eFG%', 'FT', 'FTA', 'FT%', 'ORB', 'DRB', 'TRB', 'AST', 'STL', 'BLK', 'TOV',
                      'PF', 'PTS', 'Total Minutes', 'PER', 'TS%', '3PAr', 'FTr', 'ORB%', 'DRB%',
                      'TRB%', 'AST%', 'STL%', 'BLK%', 'TOV%', 'USG%', 'OWS', 'DWS', 'WS', 'WS/48',
                      'OBPM', 'DBPM', 'BPM', 'VORP')
X <- as.matrix(data[independent_vars])
y <- as.numeric(data[['2022/23']])  # Ensuring y is a numeric vector

# Splitting the data into training and testing sets
set.seed(42)  # For reproducibility
train_index <- createDataPartition(y, p = 0.8, list = FALSE)
X_train <- X[train_index, ]
y_train <- y[train_index]


# Linear Regression
linear_model <- lm(Salary ~ ., data = data, subset = train_index)

summary(linear_model)
```

Output:

```r
summary(linear_model)

Call:
lm(formula = Salary ~ ., data = data, subset = train_index)

Residuals:
      Min        1Q    Median        3Q       Max
-17271607  -3780616   -339319   2981066  28347600

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -9943903    8577108  -1.159   0.2472
Age                830406      85255   9.740   <2e-16 ***
GP                  18575      47434   0.392   0.6956
GS                  74307      34322   2.165   0.0311 *
MP                 -52164     271597  -0.192   0.8478
FG                6696598   11249520   0.595   0.5521
FGA              -1765383    6988616  -0.253   0.8007
`FG%`            15421696   54792567   0.281   0.7785
`3P`            -15169808    9325277  -1.627   0.1048
`3PA`             3141834    6975405   0.450   0.6527
`3P%`            -6568710    4356067  -1.508   0.1325
`2P`            -10688414    7512447  -1.423   0.1558
`2PA`             2362820    7069387   0.334   0.7384
`2P%`            -6717433    6005769  -1.118   0.2642
`eFG%`          -26196066   51929686  -0.504   0.6143
FT               -2326901    5108256  -0.456   0.6490
FTA               -182137    1875080  -0.097   0.9227
`FT%`            -1938628    3860145  -0.502   0.6159
ORB              -2343882    6347816  -0.369   0.7122
DRB                226885    6327954   0.036   0.9714
TRB                366521    6299693   0.058   0.9536
AST                 86799     987961   0.088   0.9300
STL                851351    2491032   0.342   0.7327
BLK               4352133    2371275   1.835   0.0674 .
TOV               2879711    2259708   1.274   0.2034
PF                -454629     966512  -0.470   0.6384
PTS               2293112    4760683   0.482   0.6304
```

```
`Total Minutes`     -3407       3133  -1.087   0.2777
PER              -1277824     618956  -2.064   0.0398 *
`TS%`            24649388   32814844   0.751   0.4531
`3PAr`           -2212018    9660044  -0.229   0.8190
FTr              -1174326    3446588  -0.341   0.7335
`ORB%`           -1859105    1497610  -1.241   0.2154
`DRB%`           -2048514    1432573  -1.430   0.1537
`TRB%`            4204628    2887895   1.456   0.1464
`AST%`            -160984     133779  -1.203   0.2297
`STL%`             545708     920439   0.593   0.5537
`BLK%`            -474353     428910  -1.106   0.2696
`TOV%`             -57872     122727  -0.472   0.6376
`USG%`             496388     309047   1.606   0.1092
OWS              15864048    7535271   2.105   0.0360 *
DWS              14490971    7575441   1.913   0.0566 .
WS              -14961439    7570791  -1.976   0.0490 *
`WS/48`           3705912   27783635   0.133   0.8940
OBPM             -5665629    6152290  -0.921   0.3578
DBPM             -6464302    6135833  -1.054   0.2929
BPM               7257082    6164655   1.177   0.2400
VORP               156427    1579380   0.099   0.9212
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6252000 on 327 degrees of freedom
Multiple R-squared:  0.7211,   Adjusted R-squared:  0.681
F-statistic: 17.99 on 47 and 327 DF,  p-value: < 2.2e-16
```

Obserations:

The linear regression model applied to the NBA salaries dataset revealed interesting insights regarding the significance of various factors. Notably, the p-values for most of the factors were quite high, indicating a lack of statistical significance in the model. However, 'Age' emerged as a significant predictor, as indicated by its very low p-value (significantly less than 0.05). This suggests that 'Age' is a strong determinant of a player's salary within this model.

Given the limited number of significant variables in the linear regression model, the decision was made to shift the focus to a model incorporating factors derived from the Factor Analysis.

By focusing on RC1 and RC5, the model aims to capture the most influential aspects of a player's performance that correlate with their salary. This approach is expected to provide a more targeted and relevant analysis, aligning with the goal of developing a robust predictive tool for understanding and forecasting NBA salaries.

Hypothesis Testing:

We started hypothesis testing by initially creating a Impact column in

Statistical Tests and Analysis Techniques

Regression Analysis: