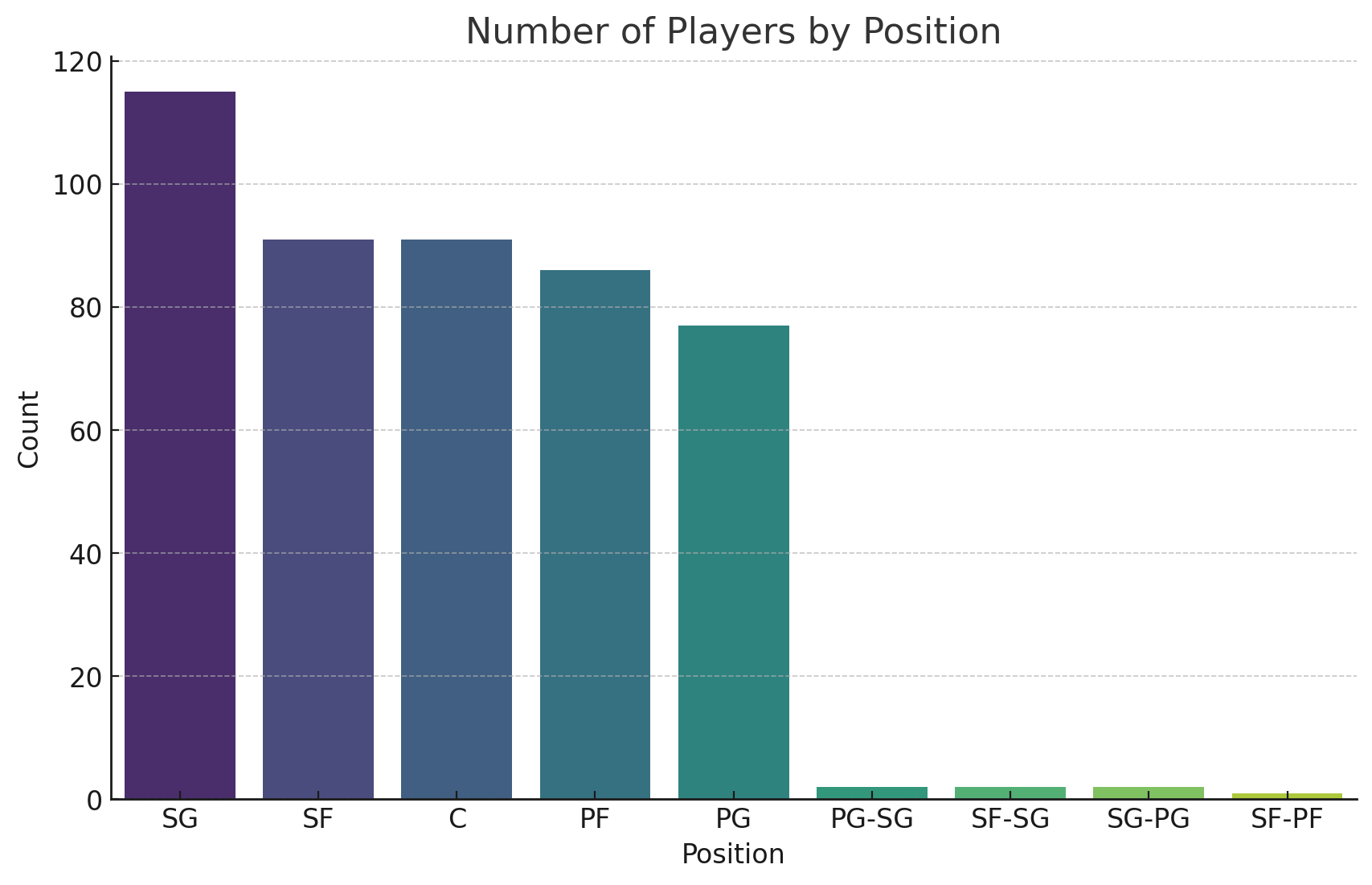
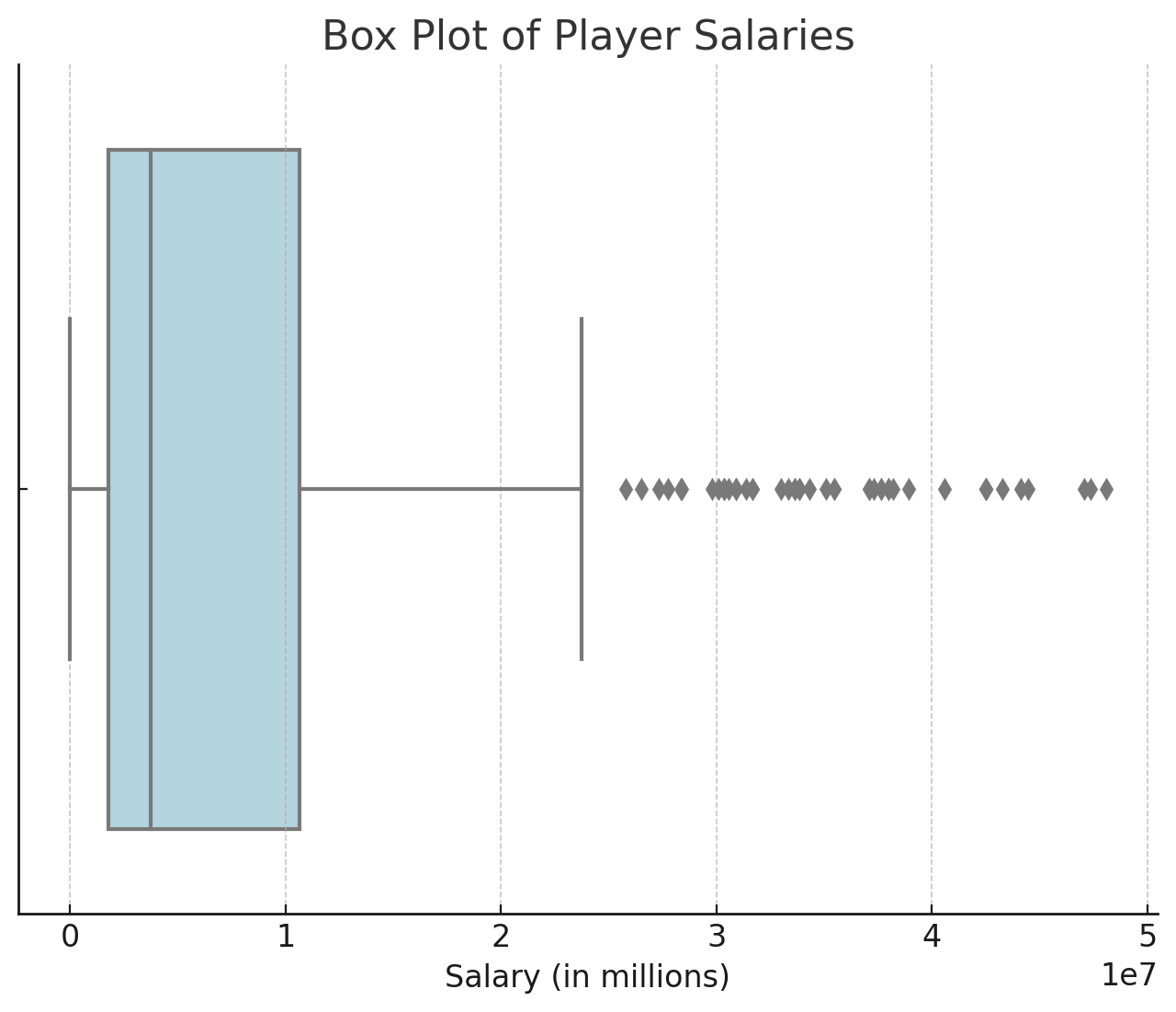
### **Dataset Overview**

The dataset contains 467 rows representing players and 52 columns capturing attributes. It includes 43 numerical columns (e.g., PTS, RPG, WS, USG%), three categorical columns (Player Name, Position, Team), and six integer columns (e.g., GP, FG). Key statistics indicate a mean of 9.13 points per game (PTS) with a maximum of 33.1, an average salary of $8.42M with a maximum of $48.07M, and average minutes played (MP) of 19.87 with a standard deviation of 9.55.

### **Preprocessing and Data Cleaning**

The dataset was trimmed to retain critical statistics such as PTS, TRB, AST, FG, FT, TOV, TS%, and PER, while irrelevant or redundant columns were removed. Missing values in columns like FG%, FT%, and TS% were addressed using imputation. Notable outliers were observed in Salary, PTS, AST, and BLK; these were retained for richness but monitored during analysis. Numerical features were normalized using StandardScaler for consistent scaling.

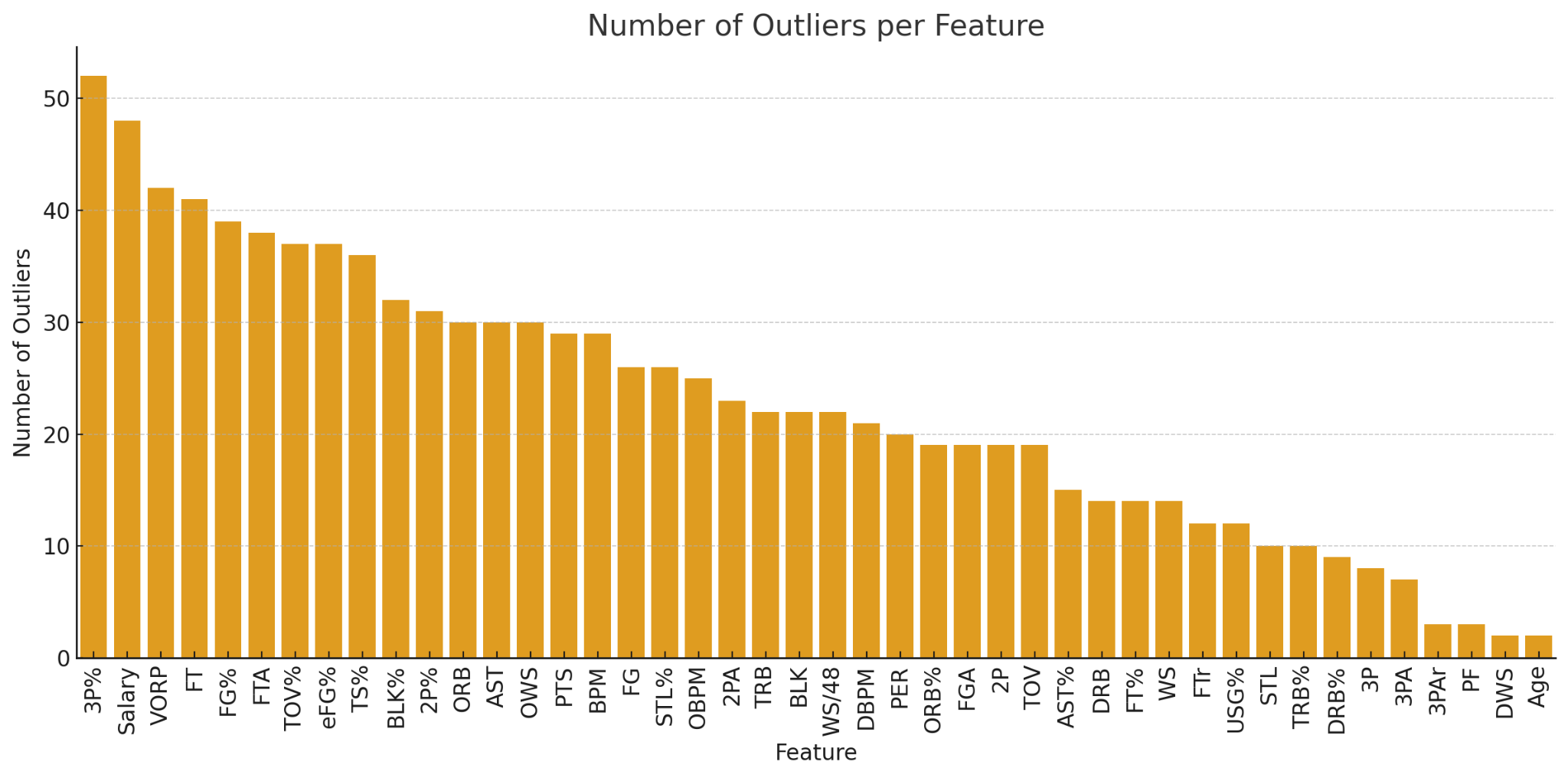
### **Exploratory Data Analysis (EDA)**

The visualizations included a correlation heatmap showing a strong positive correlation between PTS, MP, and Salary. A box plot highlighted salary outliers exceeding $40M. A bar chart showed guards had the highest positional variability. Salary, FT, and BLK were positively skewed, FG% and TS% slightly negatively skewed, and GP and PF symmetrically distributed.

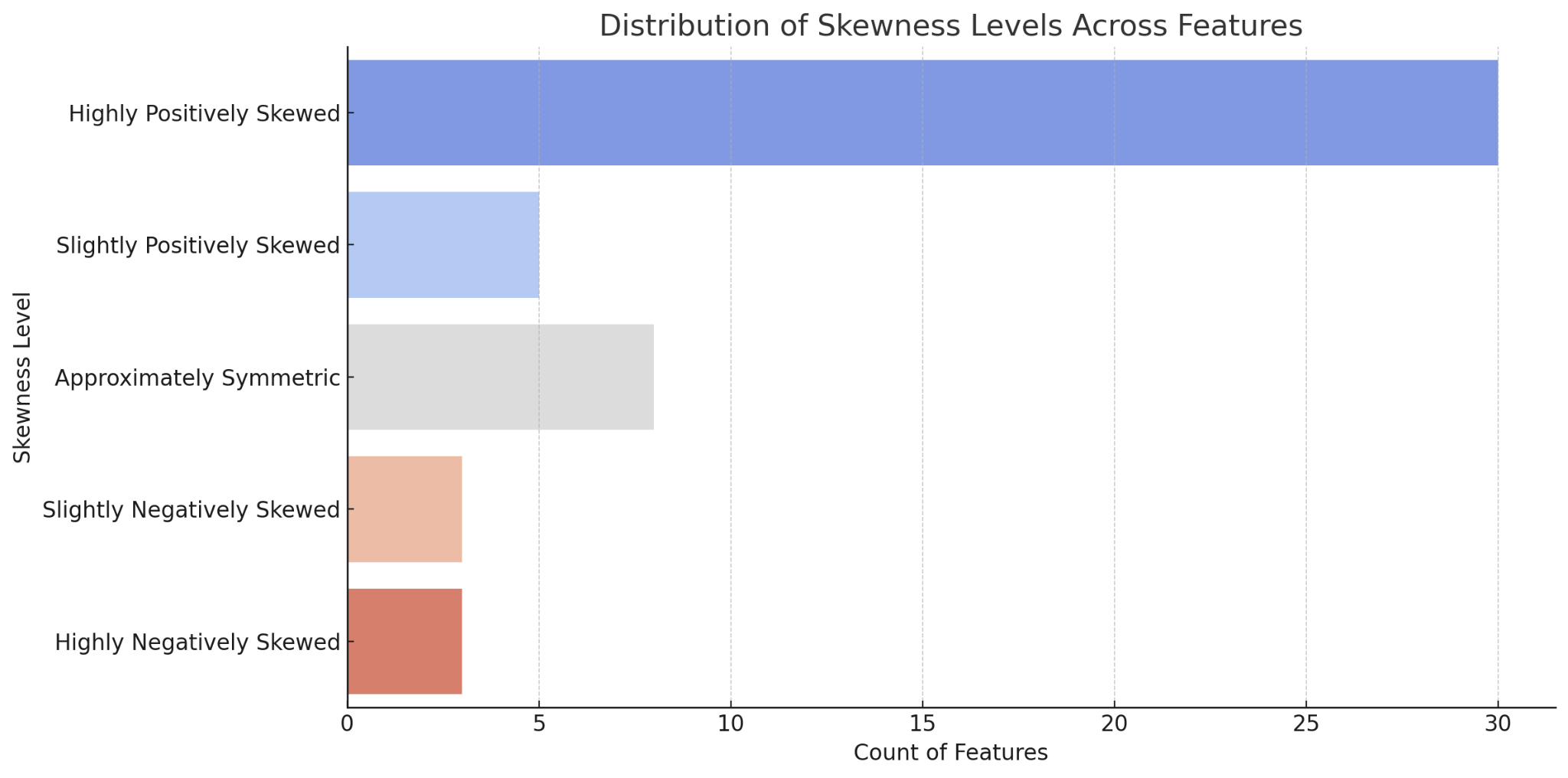


**Dataset Outliers and Feature Distribution**

The dataset showed 48 salary outliers (e.g., $48.07M), 41 and 38 outliers in Free Throws (FT) and Attempts (FTA), and notable outliers in Offensive Rebounds, Assists, and Blocks (30, 30, and 22, respectively). Salary, FT, and BLK were positively skewed, GP and PF were symmetrical, and FG% and TS% were slightly negatively skewed toward higher values.



The dataset showed highly positive skewness in Salary, Free Throws, and Blocks, with most values concentrated at the lower end and a few extremes. GP and PF had balanced distributions, while FG% and TS% were slightly negatively skewed, favoring higher values.



**Model Implementation**

Two models were implemented: Decision Tree for simplicity and non-linear relationships and SVM for accuracy and high-dimensional data. The dataset was split into 80-20 (373 training rows, 94 testing rows) and 70-30 (326 training rows, 141 testing rows).

In the 80-20 split, Decision Tree achieved an MSE of 0.0924, R² of 0.9362, and Precision/Recall/F1 of 0.9286, while SVM outperformed it with an MSE of 0.0516, R² of 0.9644, and Precision/Recall/F1 of 0.9816. Similarly, in the 70-30 split, Decision Tree recorded an MSE of 0.1104, R² of 0.9159, and Precision/Recall/F1 of 0.8738, while SVM achieved an MSE of 0.0498, R² of 0.9621, and Precision/Recall/F1 of 0.9531.

While SVM consistently outperformed Decision Trees in accuracy and robustness, Decision Trees offered speed, simplicity, and better interpretability, making them useful for explainable models.

**Interpretation of Results**

The models, Decision Tree, and SVM were evaluated on MSE, R², Precision, Recall, and F1 Score. SVM consistently outperformed Decision Tree, achieving higher R² scores (up to 0.9644) and Precision/Recall/F1 scores (e.g., 0.9816 in the 80-20 split). Decision Trees, while slightly less accurate, offered simplicity and interpretability. Critical factors included feature normalization, which improved SVM's handling of high-dimensional data, and robust performance across data splits (80-20 and 70-30). SVM excelled with complex relationships, while Decision Trees risked overfitting without pruning. Preprocessing ensured unbiased results.

**Insights and Real-world Applications**

Salary is strongly correlated with performance metrics like PPG and MP, as shown in the correlation heatmaps. Guards (PG, SG) exhibited higher variability due to their diverse roles. SVM is ideal for tasks needing high accuracy, like predicting player value or identifying star performers. At the same time, Decision Trees are better for interpretability, such as explaining salary decisions or player rankings. Applications include optimizing player salaries, predicting draft success, and engaging fans with interpretable analytics like decision tree outputs. These insights highlight the importance of aligning model choice with analysis goals.

**Questions & Answers**

1. A key challenge was distinguishing between missing values and valid zeroes, such as zero points scored or zero assists, which required careful inspection to avoid misinterpretation. The dataset also contained extreme outliers, like a $48.07M salary or 52 points in a single game, which were retained for richness but monitored closely to minimize their impact on analysis.
2. SVM was chosen for its accuracy and ability to handle high-dimensional data, making it ideal for predicting performance and salary based on multiple metrics. Decision Trees were selected for their simplicity, interpretability, and ability to model non-linear relationships, providing clear rules to explain how metrics like PPG or position influence salary.
3. Key factors included data quality, filling missing values and normalization, improved consistency, and feature selection, with metrics like PPG and MP enhancing correlations with salary. Player variability, particularly among guards with diverse roles, also impacted model performance, requiring SVM to generalize effectively while tuning Decision Trees to prevent overfitting.
4. SVM outperformed Decision Trees, delivering higher R², Precision, Recall, and F1 scores. It excelled in capturing relationships between metrics like PPG and salary. While less accurate, decision trees provided clear, interpretable rules, such as linking high points and minutes played to higher salary brackets, making them valuable for explanation.
5. The analysis showed a strong correlation between performance metrics like PPG and MP and salary, with better stats leading to higher contracts. Guards displayed higher variability due to their diverse roles in scoring and playmaking. These insights can help optimize salaries, predict draft success, and enhance fan engagement through interpretable analytics.