Behind NBA Data

Referee influence | Clustering players | Games predictions

Rubén Martínez Covelo

Data sources:

- **Kaggle:** NBA Enhanced Box Score and Standings Stats.
  - 4 DataSets (Officials, Teams, Players, Standings).
- **Kaggle:** Social Power NBA.
  - 1 DataSets (Player Averages, Social Data, Salaries Data).

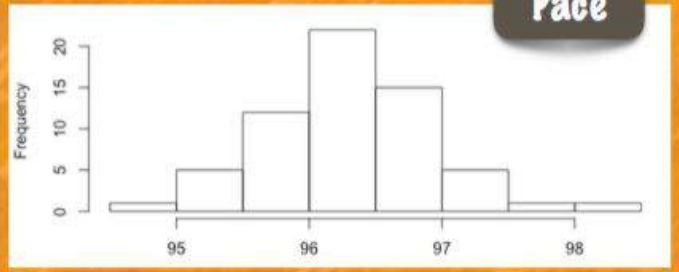
Analysis goal:

- Measure the impact of referees on the game.
- Clustering players.
- Predict season record & single games result.

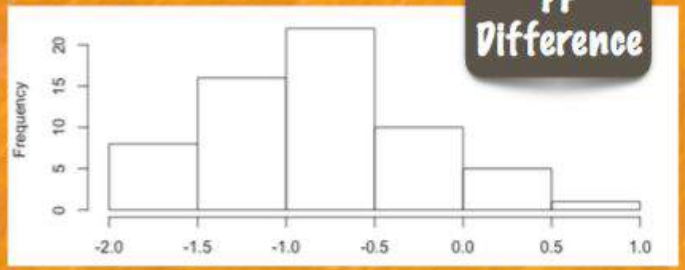
Referees Analysis

Individual analysis: grouped by referee.

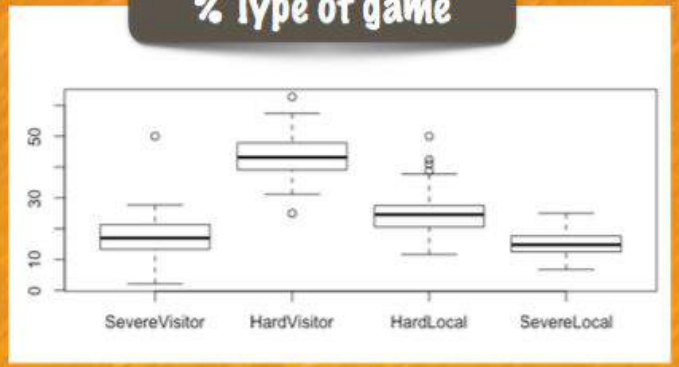
Pace



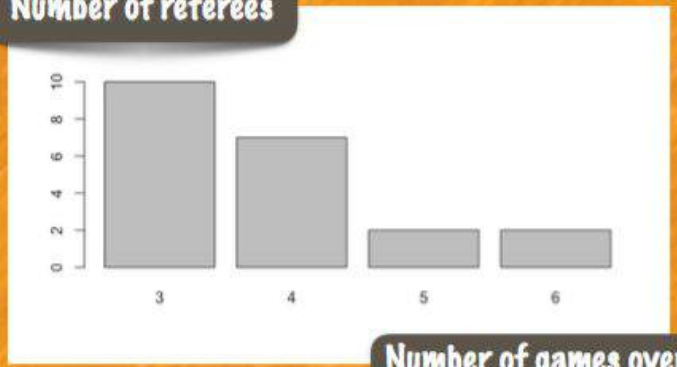
PF Difference



% Type of game



Number of referees



Number of games over -1 OPF difference


Game Hardness Distribution as Local
Away
Home
teamPF
40
30
20
10
BOS GS CLE SA TOR HOU LAC UTA WAS ATL IND MIL OKC MEM CHI MIA POR DEN DET NO CHA DAL NY SAC MIN ORL LAL PHI BKN PHO
Team

Clustering players with 

Step 1

- Remove factor variables.
- Handle NAs.
- Remove outliers with only a few games played.

→ Only 3 variables define the clusters

Step 2

- Log transformation to social metrics due to their huge scale.
- Weight all the stats per 48 minutes.
- Normalize between 0-1.

→ Too many defining variables

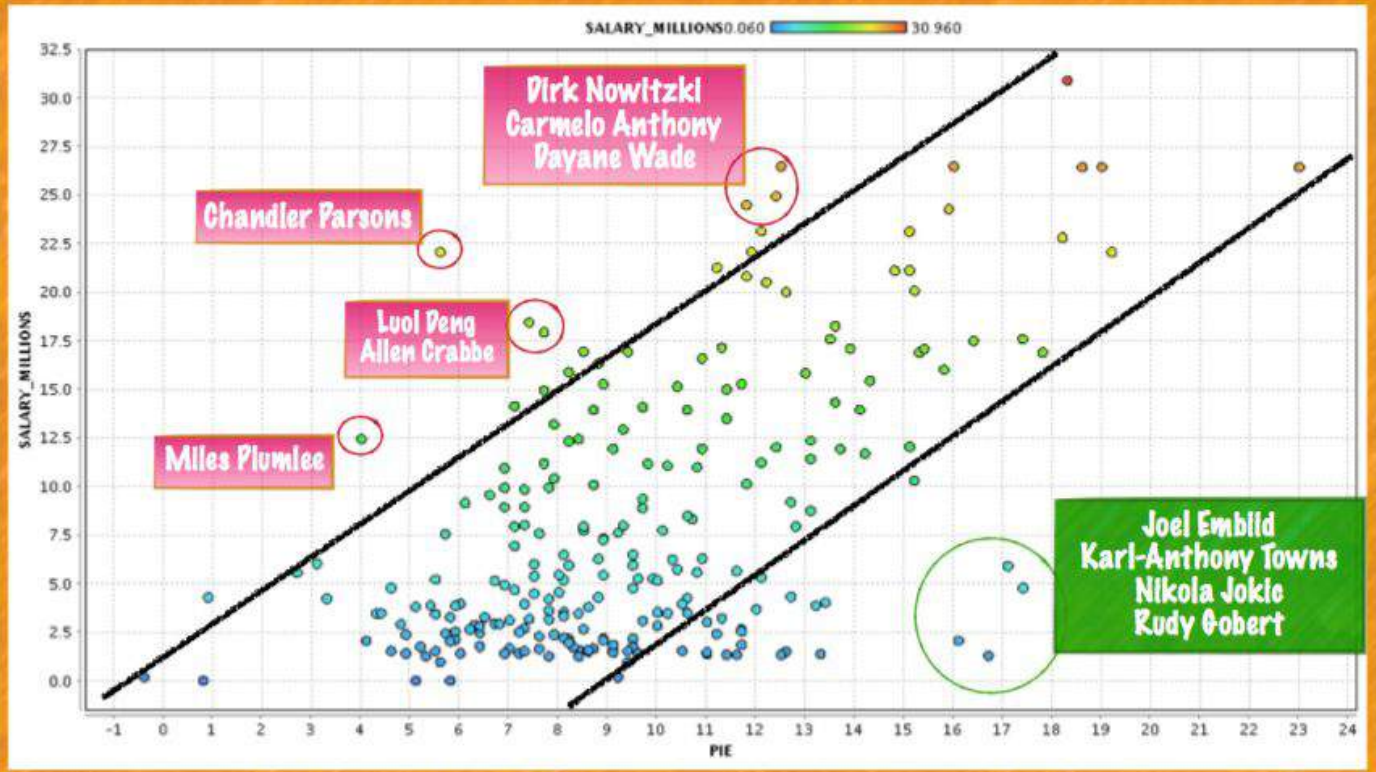
Step 3

- Feature selection (correlations)

→ 5 Identifiable clusters

Clustering players

Salary analysis



Final reflections

Next steps

- Normalization
- Feature transformation
- Feature selection
- Other algorithms
- Use more advanced stats for both teams and players
- Change team insignias due to trades and injuries
- Check records accuracy
- Introduce when players and teams score