Predicting the Overall Rating of FIFA players with Linear Regression

Jesús Cabezas

Sandra Cunha

Content

Column Selection

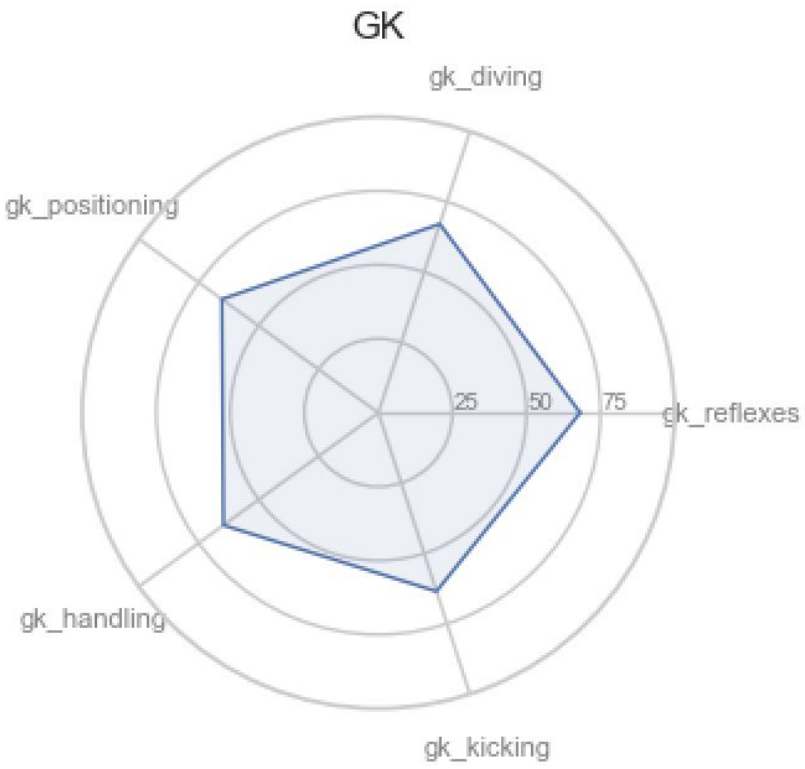
- Age
- OVA
- Nationality
- Club
- BOV
- Position
- POT
- Value
- Crossing
- Finishing
- Heading
- Short Passing
- Volley
- Dribbling
- Curve
- FK Accuracy
- Long Passing
- Ball Control
- Acceleration
- Sprint Speed
- Agility
- Reactions
- Balance
- Stamina
- Strength
- Long Shots
- Aggression
- Interceptions
- Positioning
- Vision
- Shot Power
- Jumping
- Penalties
- Marking
- Standing Tackle
- Sliding Tackle
- GK Diving
- GK Handling
- GK Kicking
- GK Positioning
- GK Reflexes

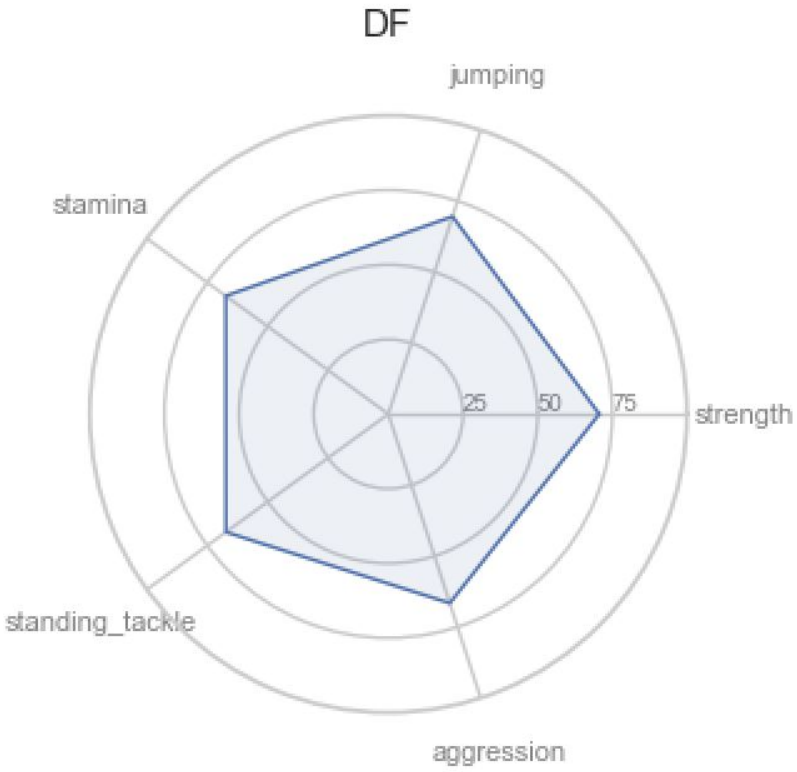
Data Cleaning

- Standardize header names
- Checking for nulls
- Grouping values in “Position” Column to 'DF', 'MF', 'AT', 'GK'

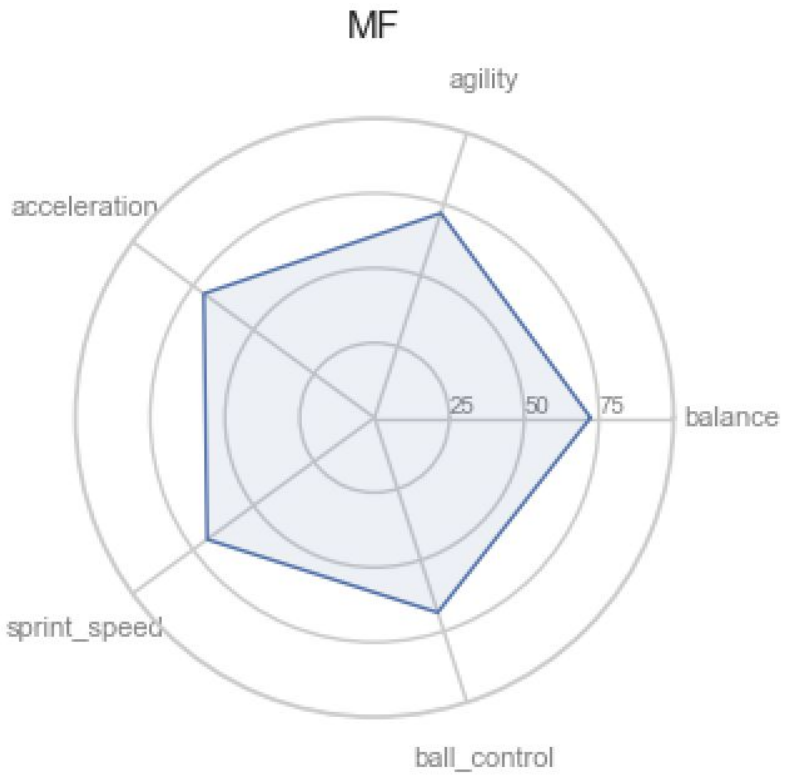
Exploratory Data Analysis

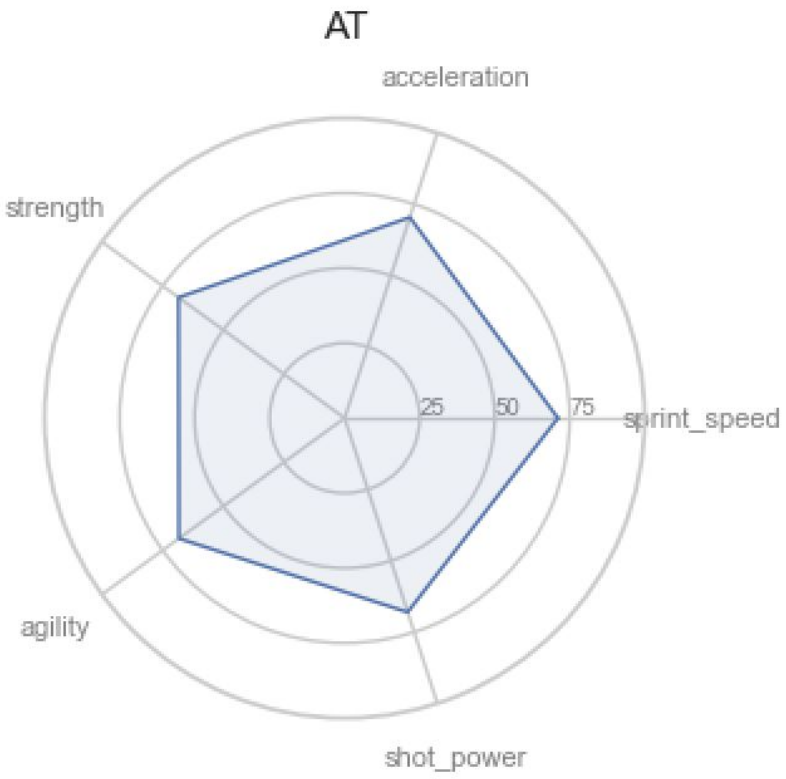
Top 5 Features by Position



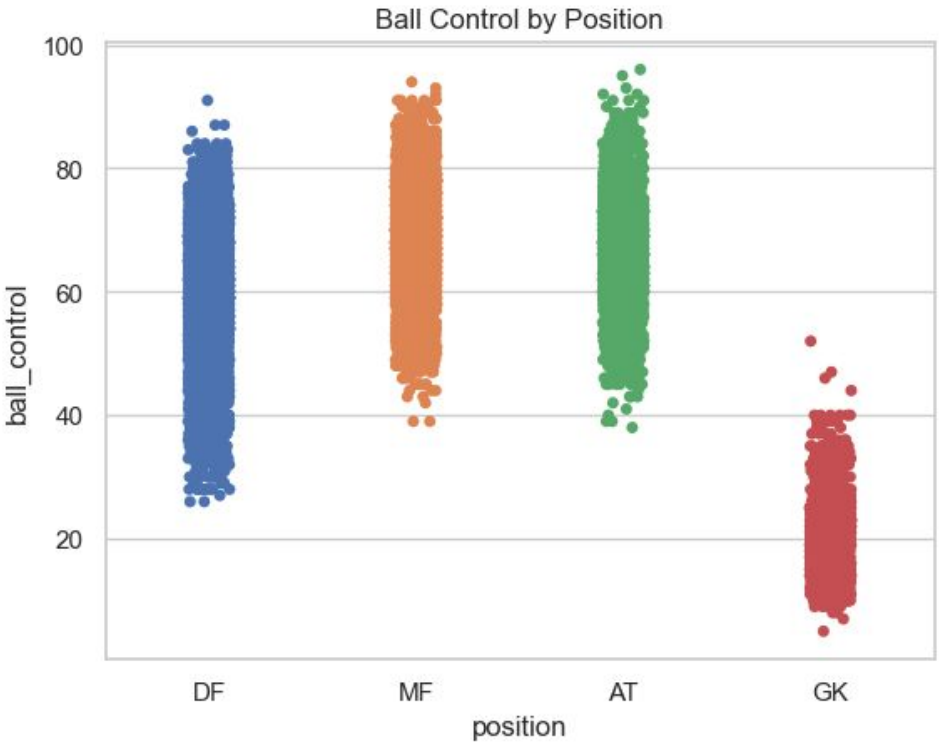


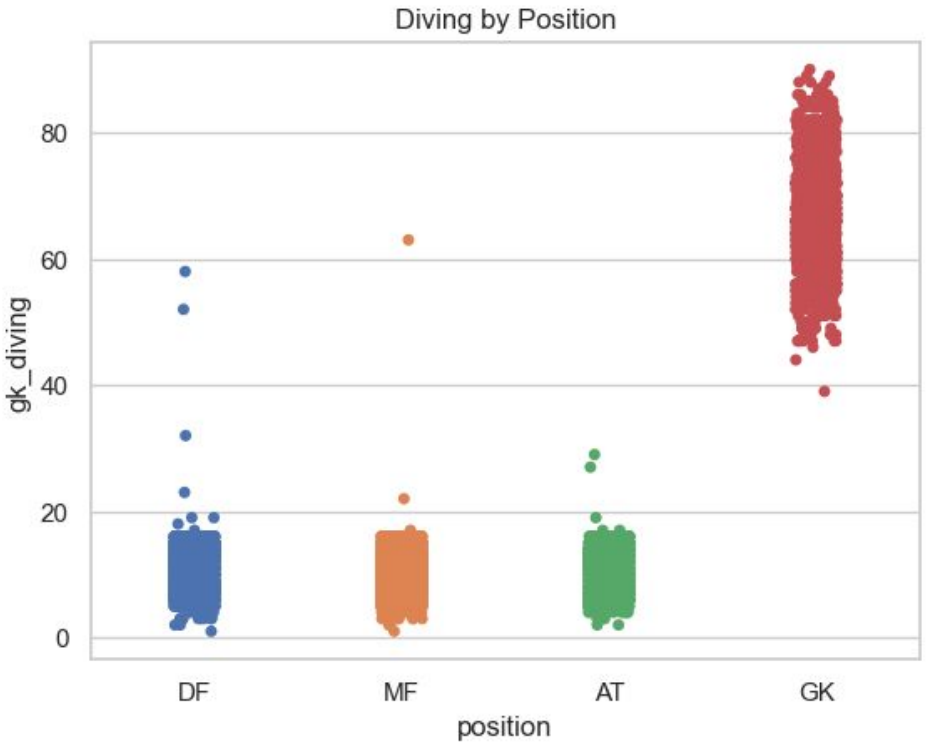
Top 5 Features by Position



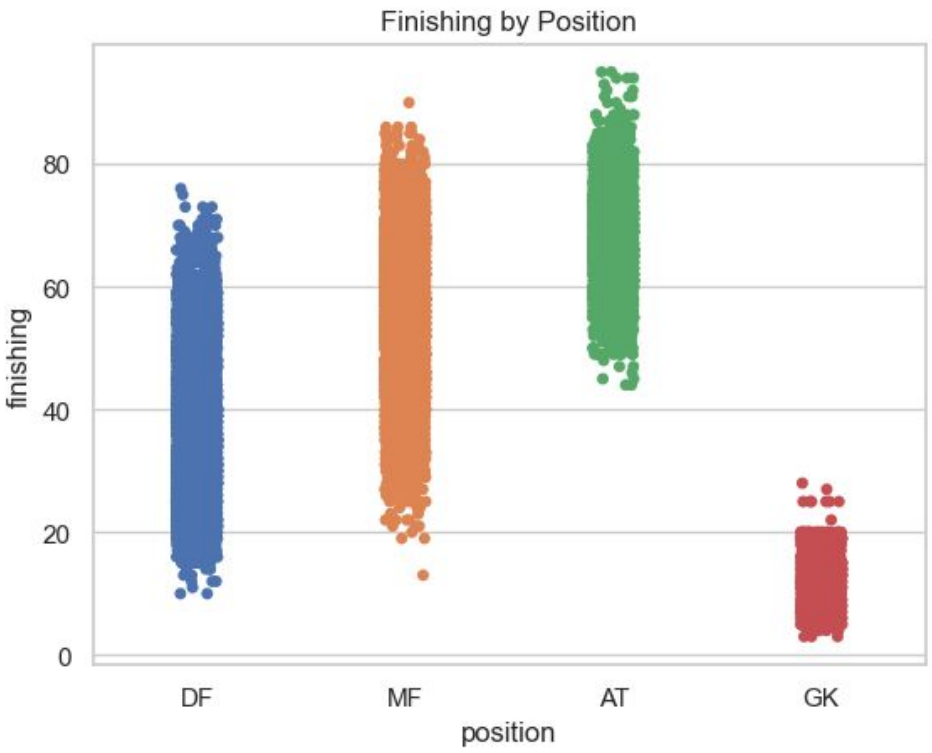


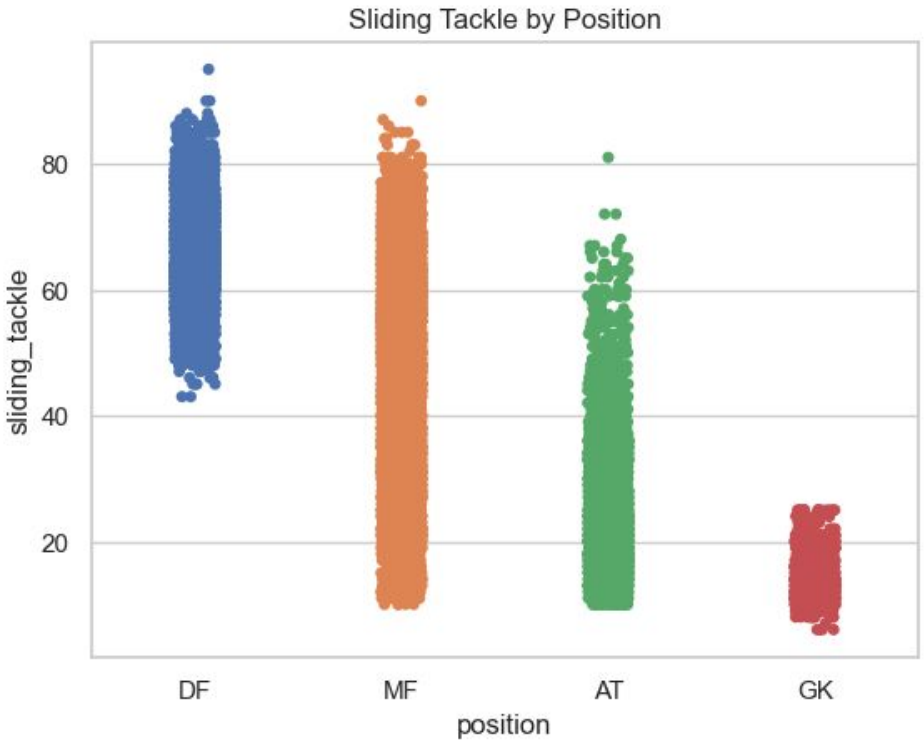
Some differences in features

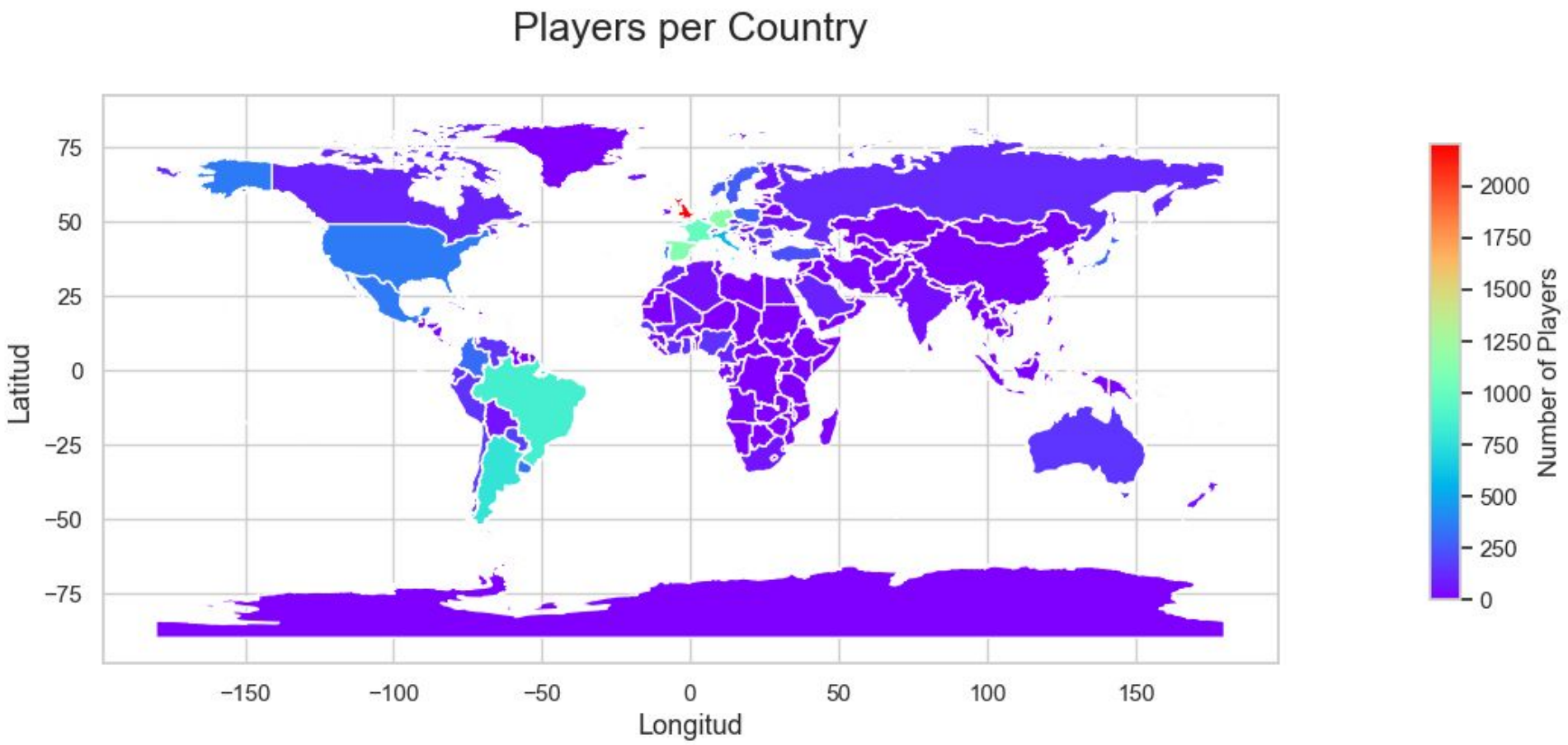




Some differences in features



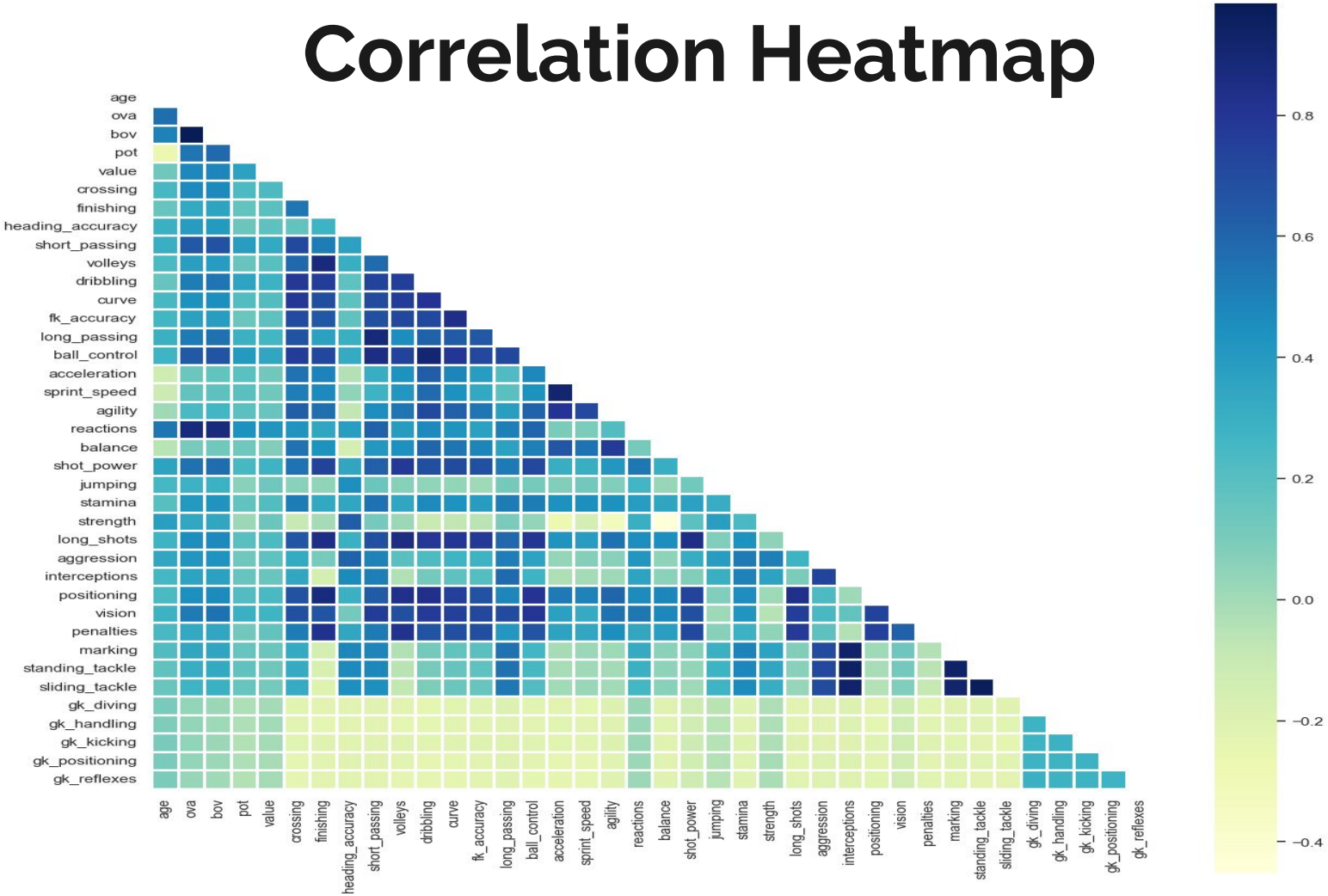



Players per Country
75
50
25
0
−25
−50
−75
Latitud
−150
−100
−50
0
50
100
150
Longitud
2000
1750
1500
1250
1000
750
500
250
0
Number of Players

Correlation Heatmap

- Understand which variables are related to each other;
- The relationship with our target: “ova” - Overall Rating;
- Feature selection: which columns are relevant for our model?

Correlation Heatmap



Feature Selection

Moderate to High Correlated columns (above 0.8) with our target “ova”:

- vision
- shot_power
- reactions
- ball_control
- short_passing
- bov

Preprocessing

Numerical Features	Categorical Features
● Data scaling ● Box-cox	● Dummy Encoding Nationality Position

Modeling

X-y split

y = finaldata["ova"] – Our Target “Overall”

X = finaldata.drop(["ova"], axis=1)

Train-test split

30% --> test , 70% --> train

Model Evaluation

Results with Box-cox:

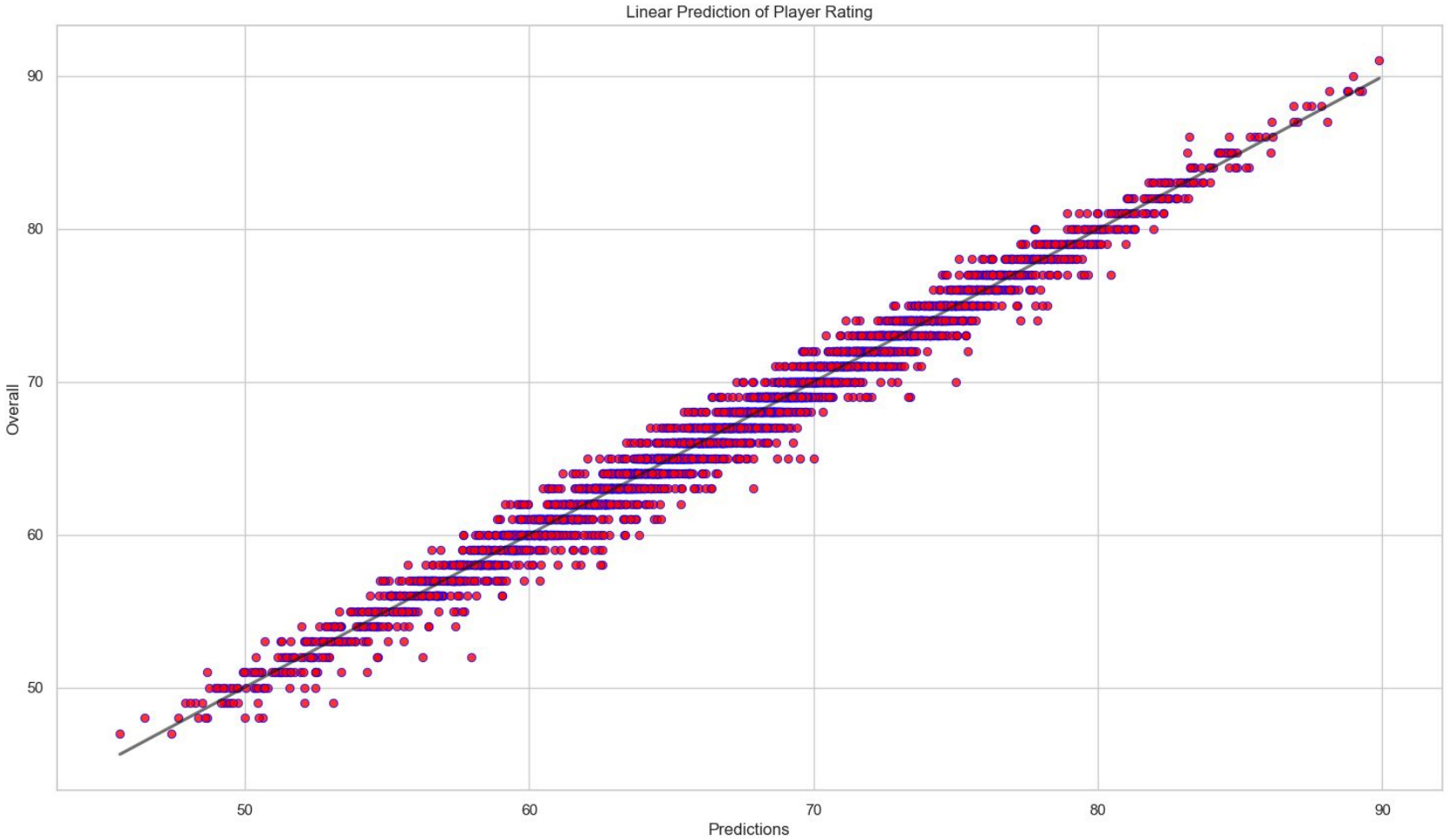
- R2 = 0.98
- RMSE = 0.71
- MSE = 0.50
- MAE = 0.53

Results without Box-cox:

- R2 = 0.98
- RMSE = 0.92
- MSE = 0.85
- MAE = 0.69

98% of the variance of the target is explained by the variance of the features

±0.7/±0.9 difference between values predicted by the model and the actual values


Linear Prediction of Player Rating
Overall
Predictions
90
80
70
60
50
50
60
70
80
90

Conclusions

- Model performs better without categoricals;
- Slightly better prediction using box-cox transformation;

Thank you!!