



REPORT ON FIFA MEN'S WORLD RANKINGS

INTRODUCTION

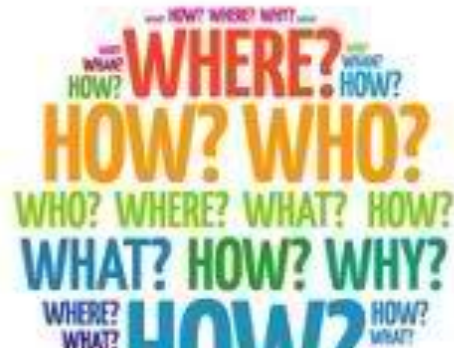
- The beginnings of contemporary football can be tracked back to 19th century Britain. Prior to this, folk football was played with numerous regulations but it was at the public schools that a standardized form of the game began to develop.
- You have been recruited as a football analyst in a company - Mchezopesa Ltd and tasked to accomplish the task below;
- A prediction result of a game between team 1 and team 2, based on who's home and who's away, and on whether or not the game is friendly (include rank in your training



DEFINE THE QUESTION



- Predict results of a game between team 1 and team 2, based on who's home and who's away, and on whether or not the game is friendly (include rank in your training).



METRICS FOR SUCCESS



- Optimal and reliable prediction Model:
- About 80% Accuracy score and above
- With the least RMSE score achievable

1 **+1** **BRAZIL**

2 **-1** **BELGIUM**

3 **—** **FRANCE**


4 **—** **ARGENTINA**

5 **—** **ENGLAND**

CONTEXT

- The FIFA Men's World Ranking ranks teams of FIFA's member nations on the basis of their game results from full international matches, awarding points to teams depending on the outcome.
- Brazil is presently ranked highest, followed by eight other countries who have held the top spot since its introduction in December 1992.
- In response to criticism, the ranking system has been revised multiple times, most recently transitioning to the Elo rating system found in chess in August 2018.
- A more detailed explanation and history of the rankings is available here: [\[Link\]](#)

Experimental design taken

- 
- Perform your EDA
 - Perform any necessary feature engineering
 - Check of multicollinearity
 - Building a model
 - Approach 1: Polynomial regression model
 - Model 1: Predict how many goals the home team scores
 - Model 2: Predict how many goals the away team scores
 - Approach 2: Logistic regression model
 - Figure out from the home team's perspective if the game is a Win, Lose or Draw (W, L, D)
 - Cross-validate the model
 - Compute RMSE
 - Create residual plots for the model
 - Assess Heteroscedasticity using Bartlett's test
 - Challenge the solution.
 - Create a dashboard that communicates the findings.

Appropriateness of the available Data

- Two datasets are available:
- Ranking dataset: contains the team ranks from 1993 to 2018
- Results dataset: contains matches and the team scores since 1892 to 2019

Resources

- Datasets: [Link](#)
- Tools used
 - Google Colab
 - Github
 - Moringa school canvas access for data and problem statement access

Assumptions

- Data provided is sufficient to fulfil the research objectives

Constraints

- Data is from varied times.

Cost

- Analysis time.



DATA UNDERSTANDING

Data Structure and Exploration

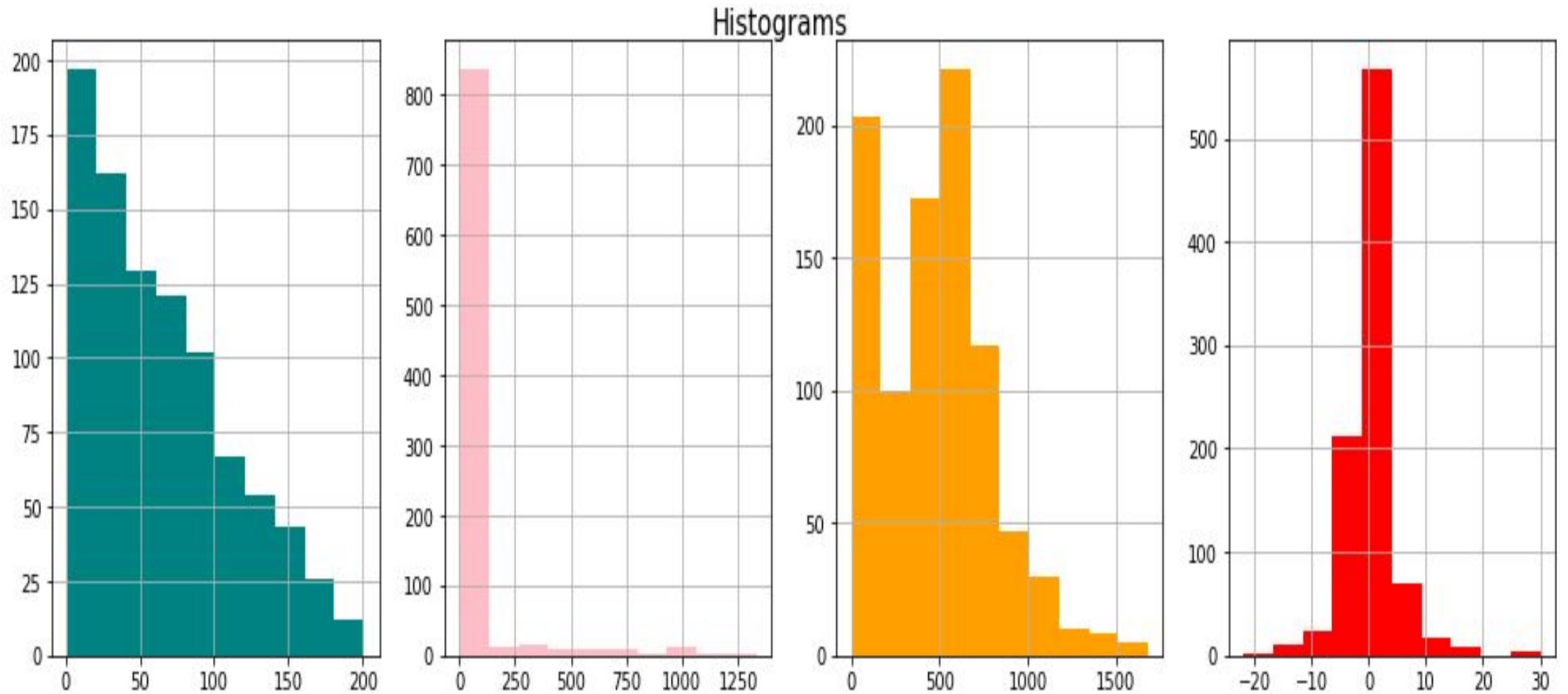
results data frame

	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
	1872-11-30	Scotland	England	0	0	Friendly	Glasgow	Scotland	False
	1873-03-08	England	Scotland	4	2	Friendly	London	England	False
	1874-03-07	Scotland	England	2	1	Friendly	Glasgow	Scotland	False
	1875-03-06	England	Scotland	2	2	Friendly	London	England	False
	1876-03-04	Scotland	England	3	0	Friendly	Glasgow	Scotland	False

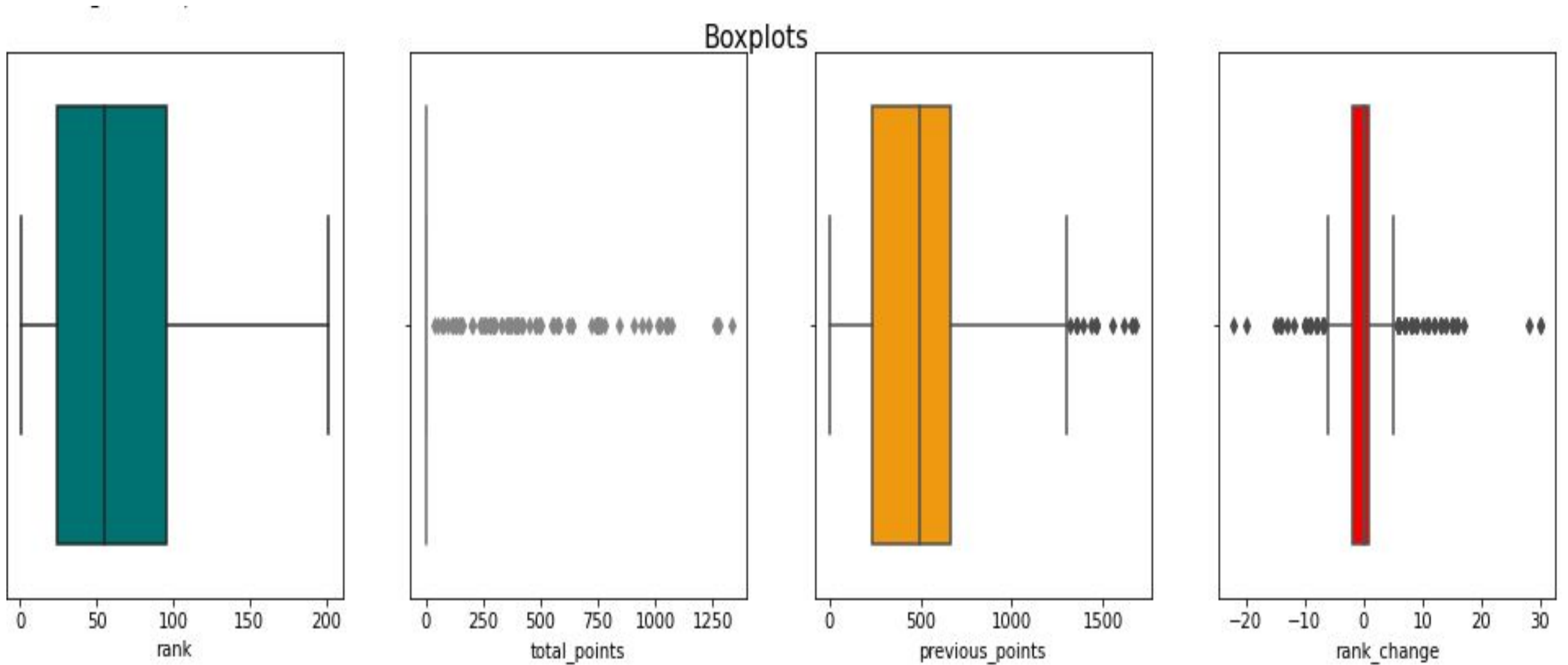
ranking data frame

	rank	country_full	country_abrv	total_points	previous_points	rank_change	cur_year_avg	cur_year_avg_weighted	last_year_avg
0	1	Germany	GER	0.0	57	0	0.0		0.0
1	2	Italy	ITA	0.0	57	0	0.0		0.0
2	3	Switzerland	SUI	0.0	50	9	0.0		0.0
3	4	Sweden	SWE	0.0	55	0	0.0		0.0
4	5	Argentina	ARG	0.0	51	5	0.0		0.0

Univariate distribution checked using histograms, the below diagrams show how the numeric variables are distributed.

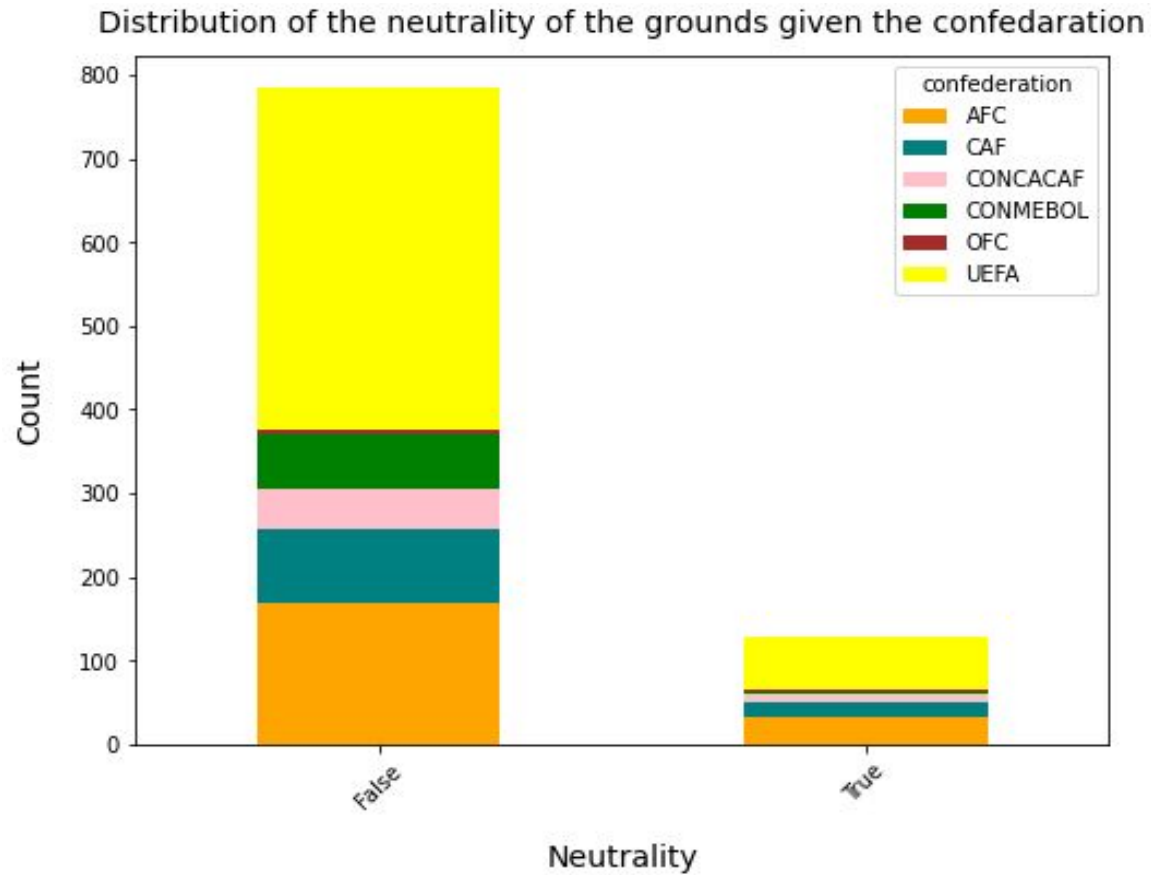


Outliers - from initial observations, the most of the numeric variables data is heavily dispersed. Box plots used to check the presence of outliers.



BIVARIATE ANALYSIS

In this section we explored the relationships between our variables in pairs using charts and testing for correlation which we visualized using a heat map..



BUILDING A MODEL

We build the model using appropriate regression techniques, and cross-validated the model to ensure its accuracy. We then computed the RMSE to measure the model's performance

Conclusion

. 1. Predicting whether home team wins, loses or draws in a match

The best accuracy score using two hyperparameters is 57% Using XGboost algorithm the accuracy score was 75% Therefore, another algorithm can be considered to improve the accuracy score or more hyperparameters should be tuned

2. Predicting the rank of a team

None of the regularised regression models is a good fit to predict the rank; they all have RMSE scores greater than the mean of the actual. Since this is also more of a classification problem, using regressions is not the best approach though this project was restrictive.