
NFL
EDA PROJECT

CONTENT

- Business Case
- Dataset Analysis
- Interesting Finding
- Feature Selection
- Conclusion

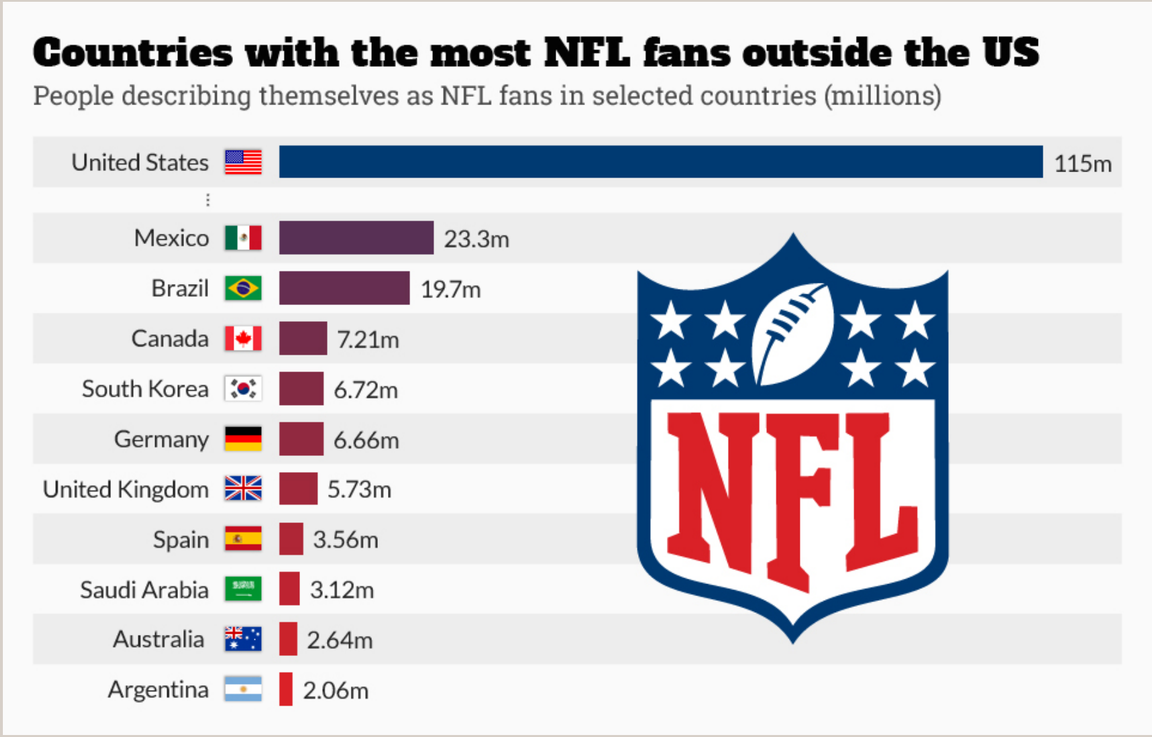
BUSINESS CASE

NFL is one of the **biggest sporting organizations** in the USA. With the **Superbowl** being the **most livestreamed** sporting event in the USA.

Identify the key factors contributing **to game ratings**. What **features drive the ratings** of a game and how can we leverage these to hone the maximum potential of the NFL?

Rating: Ranges from 0-100

Side Objective: Does the **Superbowl affect** the **rating** of the game, if so, in what way?



DATA CLEANING

Datatypes: 24 Numeric, 1 Category, 4 Objects

Missing Data: 3 columns substantial -> deleted

Duplicate Data: None

Shape Data: 15217 Rows, 30 Columns

BEFORE

	date	season	neutral	playoff	team1	team2	elo1_pre	elo2_pre	\
0	1920-09-26	1920	0	NaN	RII	STP	1503.947	1300.000	
1	1920-10-03	1920	0	NaN	AKR	WHE	1503.420	1300.000	
2	1920-10-03	1920	0	NaN	BFF	WBU	1478.004	1300.000	
3	1920-10-03	1920	0	NaN	DAY	COL	1493.002	1504.908	
4	1920-10-03	1920	0	NaN	RII	MUN	1516.108	1478.004	

	elo_prob1	elo_prob2	...	qb2_game_value	qb1_value_post	qb2_value_post	\
0	0.824651	0.175349	...	NaN	NaN	NaN	
1	0.824212	0.175788	...	NaN	NaN	NaN	
2	0.802000	0.198000	...	NaN	NaN	NaN	
3	0.575819	0.424181	...	NaN	NaN	NaN	
4	0.644171	0.355829	...	NaN	NaN	NaN	

	qbelo1_post	qbelo2_post	score1	score2	quality	importance	total_rating
0	NaN	NaN	48	0	NaN	NaN	NaN
1	NaN	NaN	43	0	NaN	NaN	NaN
2	NaN	NaN	32	6	NaN	NaN	NaN
3	NaN	NaN	14	0	NaN	NaN	NaN
4	NaN	NaN	45	0	NaN	NaN	NaN

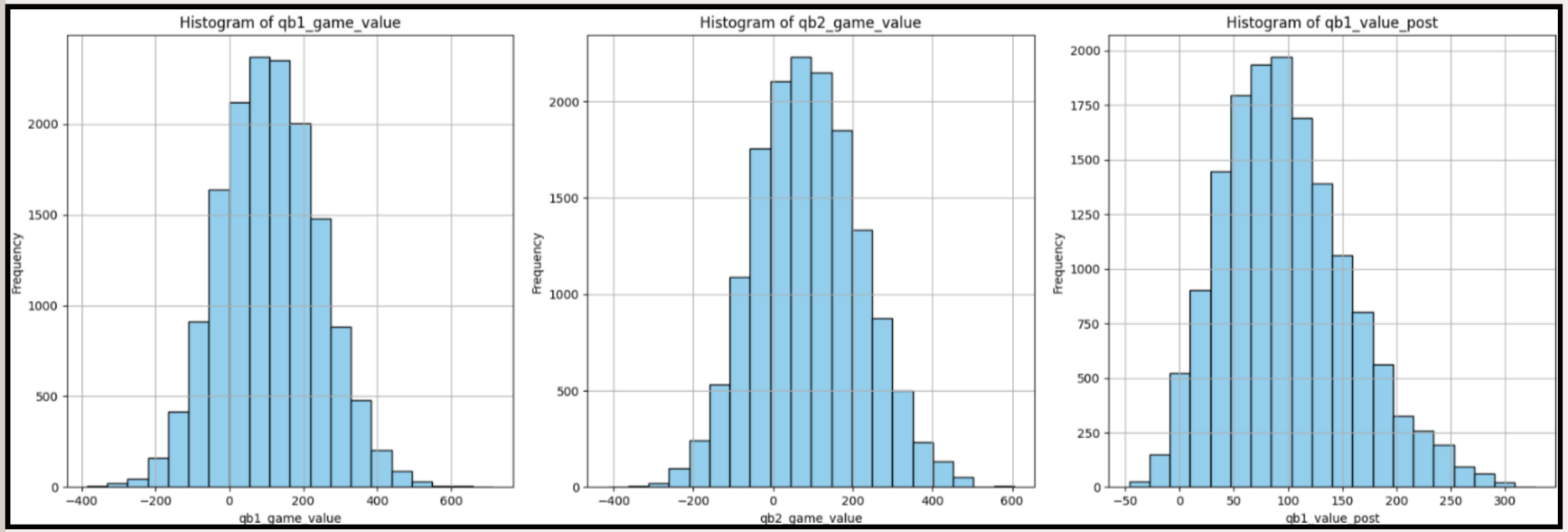
A lot of missing values and unnecessary columns

AFTER

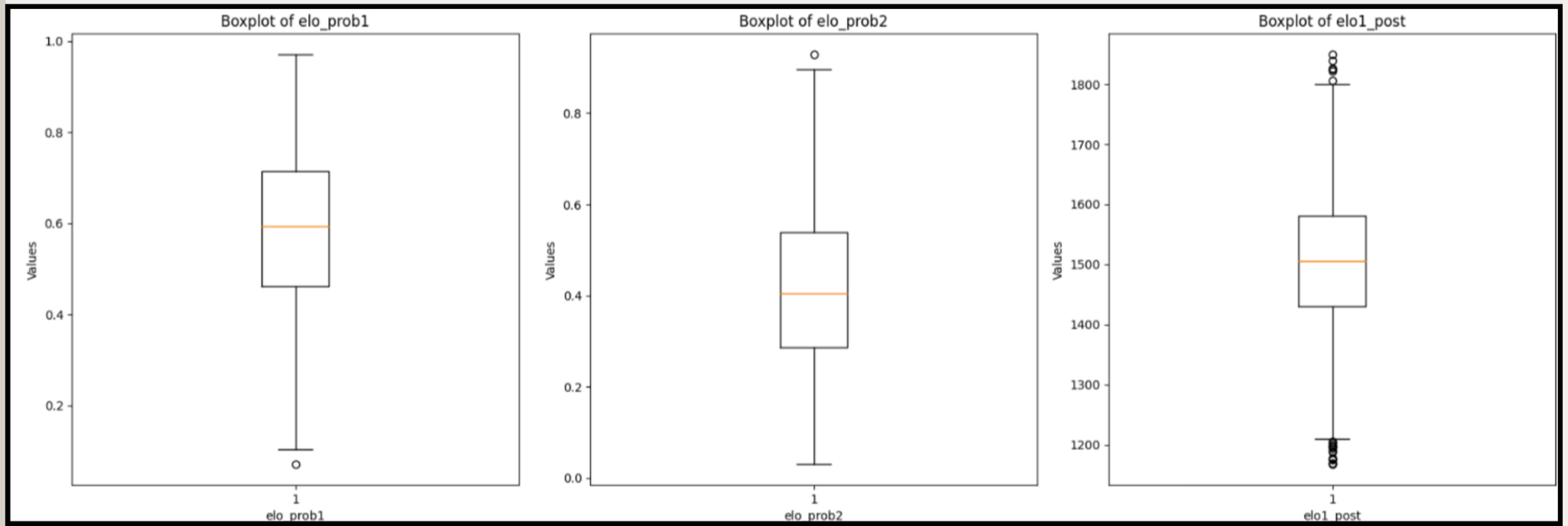
	date	season	neutral	team1	team2	elo1_pre	elo2_pre	elo_prob1	elo_prob2	elo1_post	...	qbelo_prob2	qb1_game_value	qb2_game_value	qb1_value_post	qb2_value_post	qbelo1_post	qbelo2_post	score1	score2	rating
2162	1950-09-16	1950	0	PHI	CLE	1674.314	1647.304	0.629402	0.370598	1631.511	...	0.379347	-123.09	275.22	-12.309	27.522	1630.613026	1685.827280	10	35	98.0
2163	1950-09-17	1950	0	BCL	WSH	1337.541	1454.448	0.425851	0.574149	1310.758	...	0.574643	3.63	251.79	0.363	25.179	1310.867671	1481.487455	14	38	6.0
2164	1950-09-17	1950	0	PIT	NYG	1485.849	1461.717	0.625529	0.374471	1453.448	...	0.387413	-34.98	-54.12	-3.498	-5.412	1454.118729	1493.988602	7	18	35.0
2165	1950-09-17	1950	0	GB	DET	1353.646	1449.128	0.456245	0.543755	1320.673	...	0.556100	-37.62	115.17	-3.762	11.517	1322.441687	1481.661757	7	45	7.0
2166	1950-09-17	1950	0	LAR	CHI	1564.606	1628.688	0.501321	0.498679	1548.463	...	0.501482	-12.21	-30.03	-1.221	-3.003	1549.226817	1643.798965	20	24	88.0

No missing data, correct data types, no duplicates.

UNIVARIATE ANALYSIS

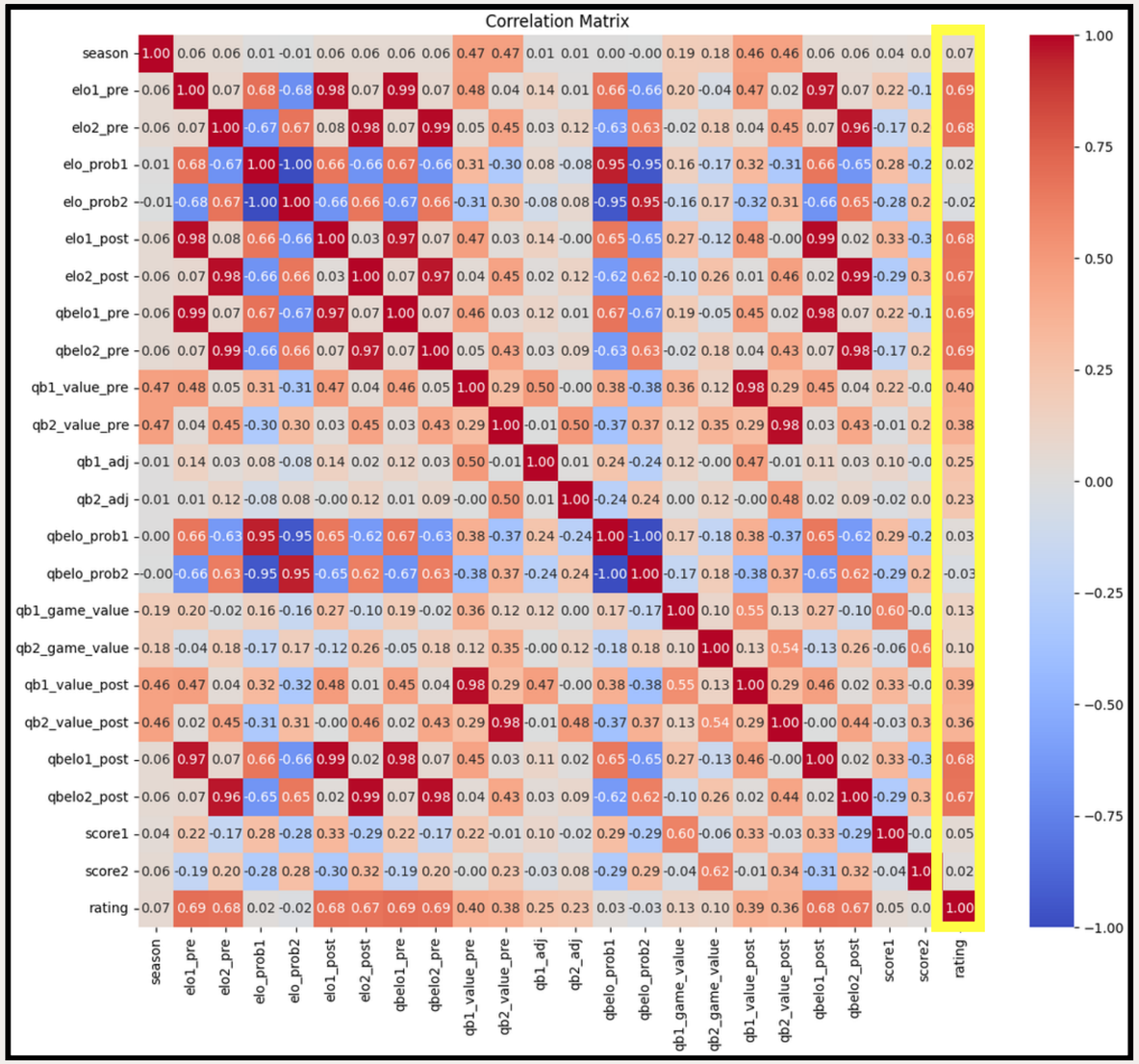


Gaussian Distribution

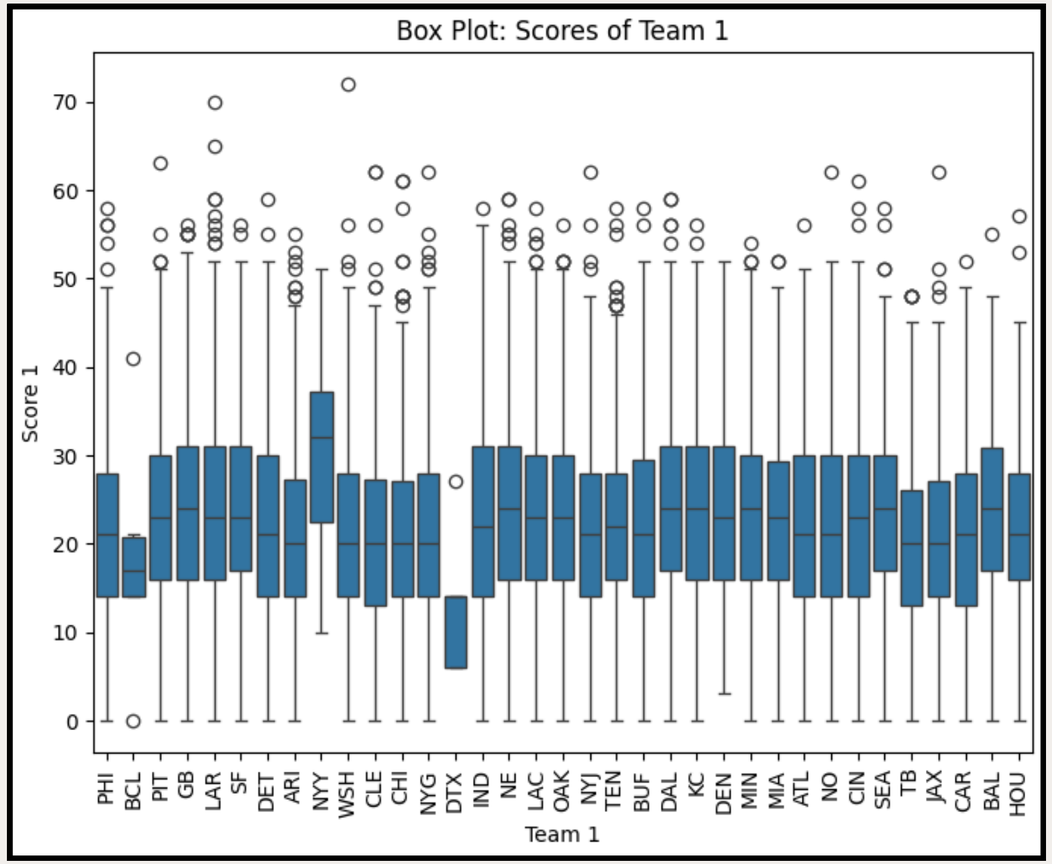


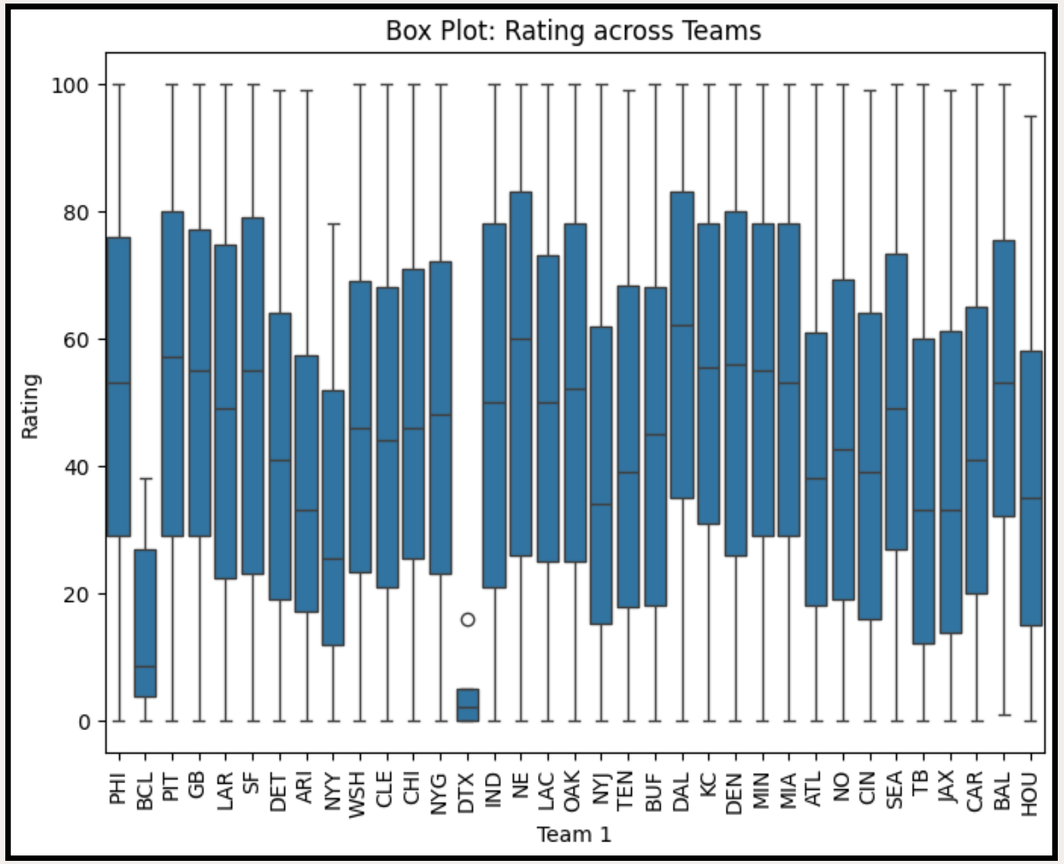
Few Outliers

BIVARIATE ANALYSIS

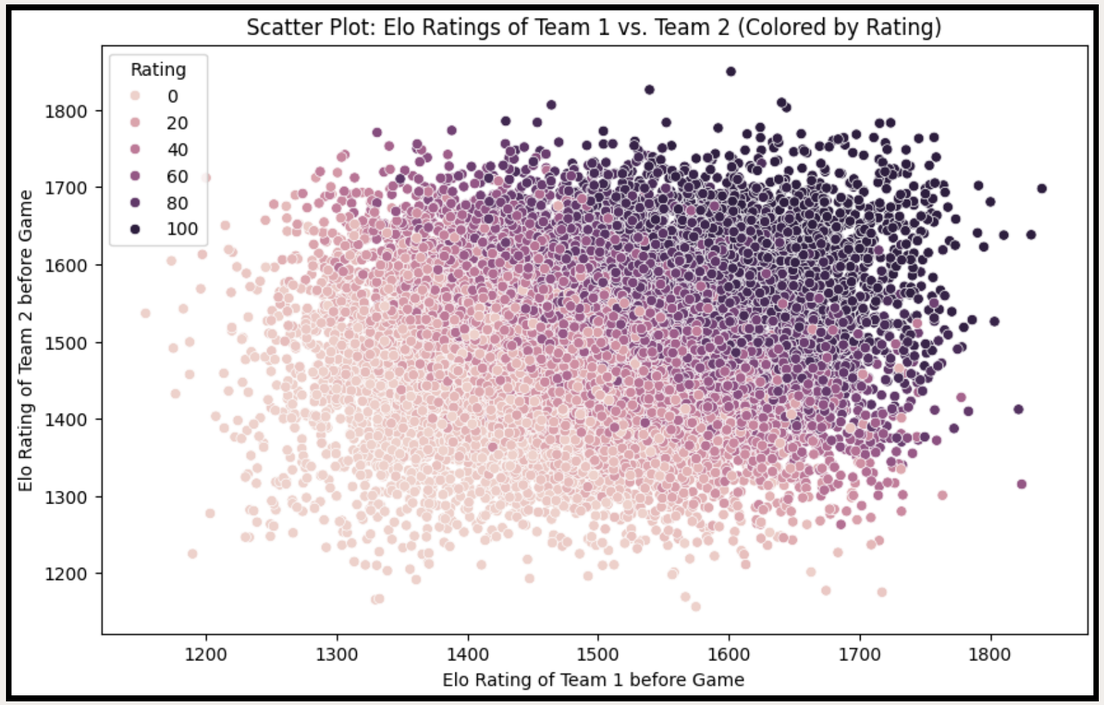


BIVARIATE ANALYSIS



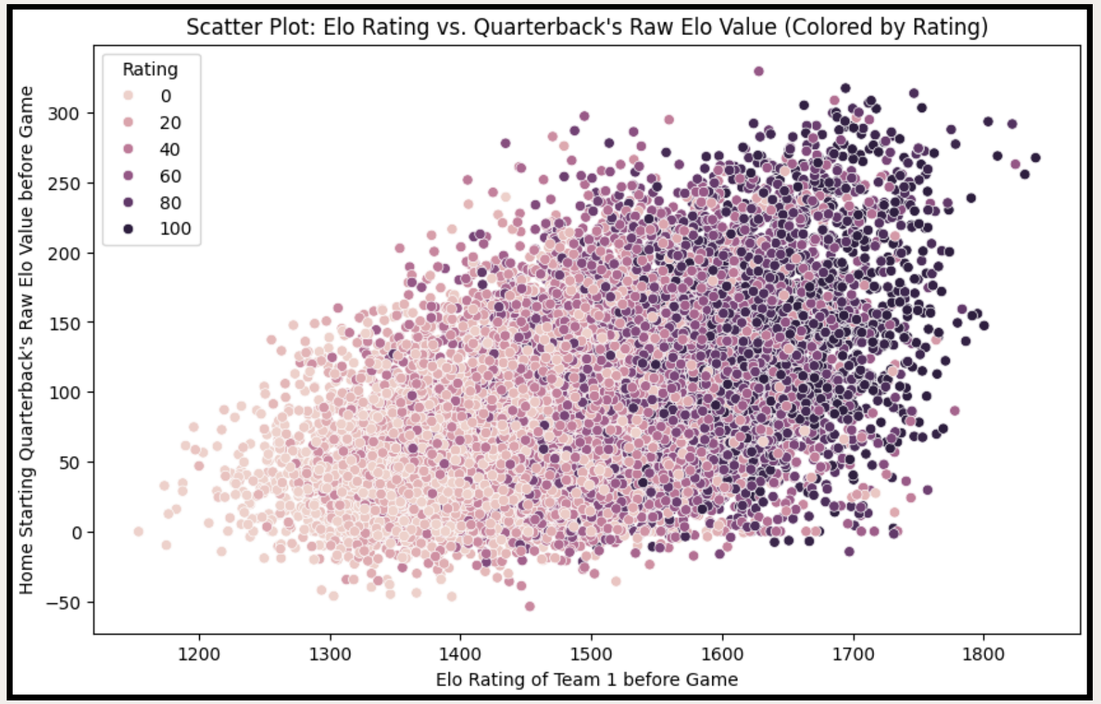


MULTIVARIATE ANALYSIS

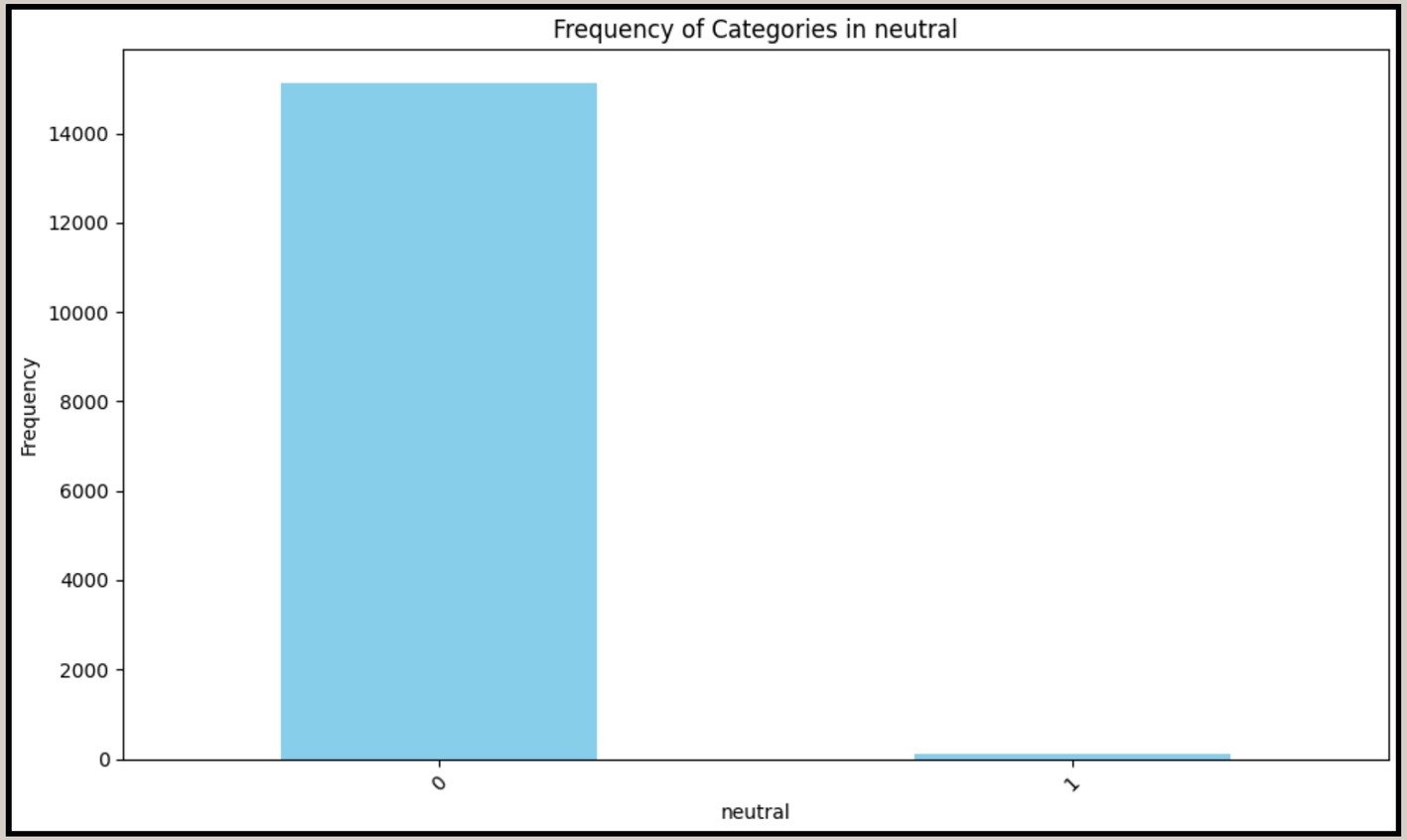


Great Pointers for Rating

Good gradient



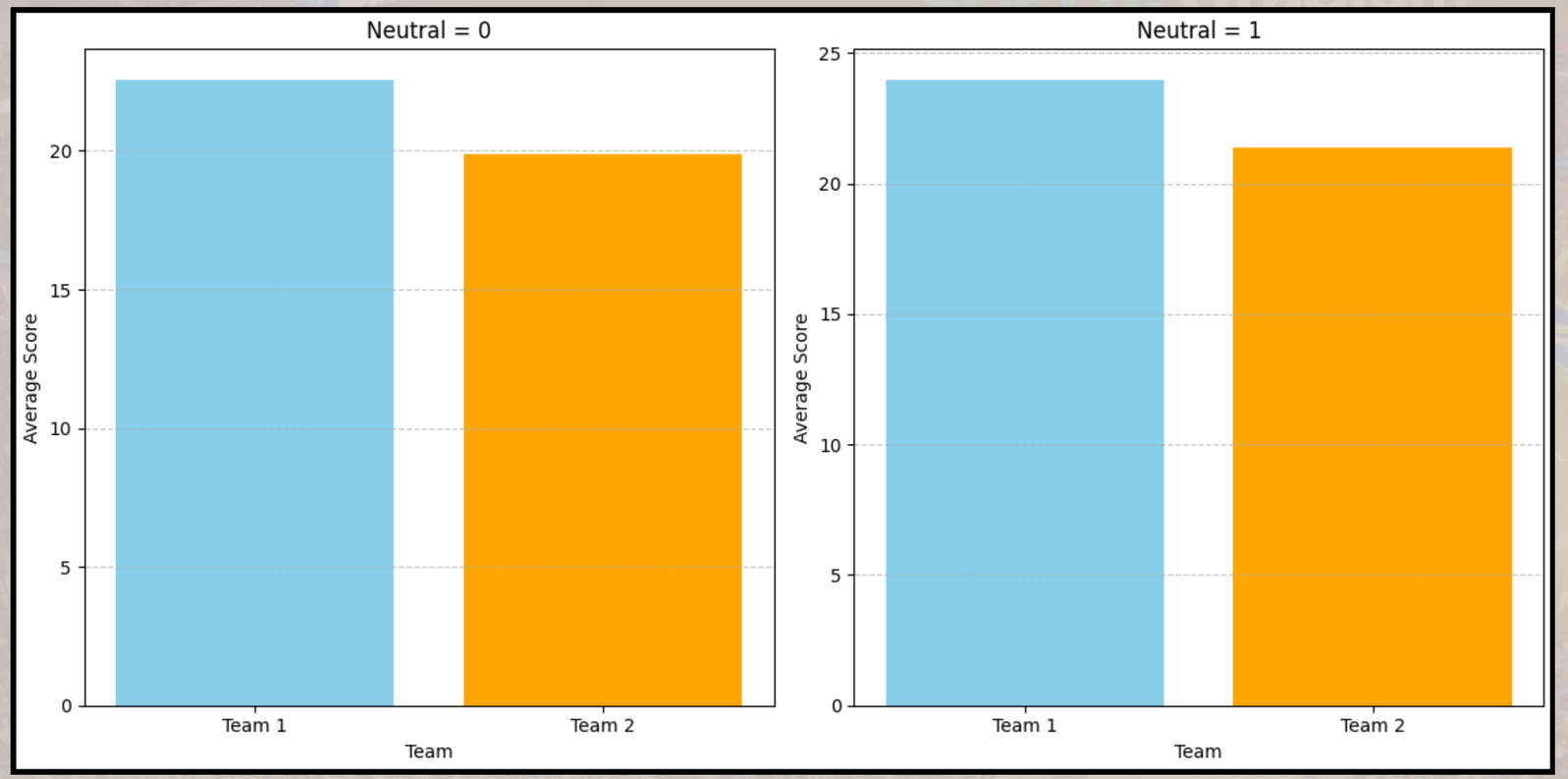
INTERESTING FINDING



Class Disproportion

Why?

SUPERBOWL



<u>Higher</u> AVG <u>score count</u> during Superbowl

Data transformation

Removal of outliers.

- Based on the z-scores
- Sorted them from worst to best
- Deleted 5%

Scaled the data

- Standardized, because retention of integrity of the features.

FEATURE SELECTION

KBest

- Rating **Team 1**
- Rating **Team 2**
- **Upd** Rating **Team 1**
- **QB** Rating Team 1
- **QB** Rating Team 2
- **Upd QB** Rating Team 1

LASSO

- **QB** Rating Team 1
- **QB** Rating Team 2
- **Adj QB** Rating Team 1
- **AdjQB** Rating Team 2
- Rating **Team** 2
- Rating **Team** 1

Decision Tree

- **QB** Rating Team 1
- **QB** Rating Team 2
- Rating **Team** 2
- **Adj QB** Rating Team 1
- **AdjQB** Rating Team 2
- **Upd QB** Rating Team 1

Interaction Terms

Added:

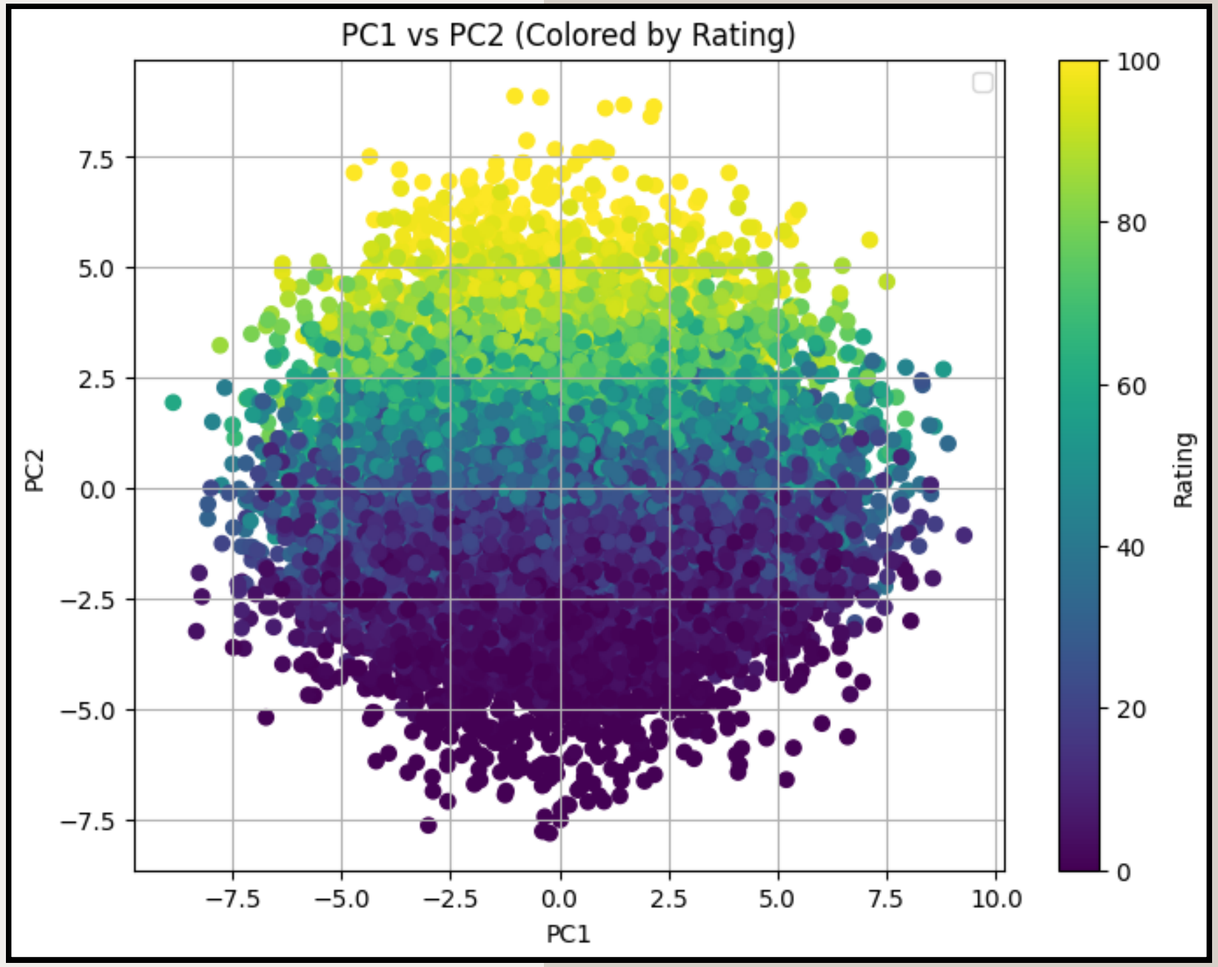
1. Quarterback Performance and Game Outcome
2. Seasonal Influence on Elo Probability
3. Elo Probability and Score Differential

!Same Results!

PCA

PC1 and PC2 mainly consists of

- **Elo ratings** (pre and post)
- **Probability values**
- **Game values** related to Elo **ratings**



CONCLUSIONS

- **Neutral Feature is significant**
  - Superbowl **DOES** affect the rating
- **Most important features:**
  - team ratings, quarterback performance, and game probabilities
  - supported by KBest, LASSO, DT, PCA
- **NFL could <u>adjust schedules</u> accordingly**

THANK YOU.

Vincent Mostert