From Seats to Stats:

Analyzing 70 Years of Premier League Match Attendance for the Big Six

Prepared by

Abhik Mukherjee

Introduction:

Football is more than just a sport—it is a deeply embedded cultural phenomenon that shapes communities, economies, and collective identities. In the United Kingdom, the Premier League stands as one of the most watched and commercially successful football leagues in the world. Among its many clubs, the so-called "Big Six"—Manchester United, Manchester City, Arsenal, Chelsea, Liverpool, and Tottenham Hotspur—have dominated not only in terms of performance but also in fan following and financial clout. Matchday attendance, a critical indicator of fan engagement and club popularity, has evolved significantly over the past seven decades, influenced by various sporting, economic, and social factors.

Understanding the trends and determinants of stadium attendance offers valuable insights into fan behavior, club performance, and the broader footballing ecosystem. Attendance is not merely a reflection of on-pitch success; it is intertwined with economic conditions, ticket pricing, media coverage, stadium facilities, regional demographics, and historical legacies. This project investigates the match attendance data of the Big Six clubs from 1949 to 2019, a period that encapsulates the transformation of English football—from its post-war roots to the commercialization of the modern Premier League era.

With the formation of the Premier League in 1992 and the influx of global investment and broadcasting revenues, attendance patterns have undergone notable shifts. This project seeks to uncover how these structural changes, along with other measurable factors, have shaped the attendance trajectories of the most prominent English clubs.

Objectives:

The primary objective of this project is to perform an in-depth statistical and data-driven analysis of match attendance for the Big Six Premier League clubs over a 70-year period. The specific aims of the study are:

- Visualize and analyze historical trends in match attendance from 1949 to 2019.
- Investigate how attendance evolved before and after the formation of the Premier League in 1992.
- Examine the relationship between attendance and variables such as league position, squad value, stadium capacity, ticket prices, cup wins, televised matches, renovations, and regional demographics.
- Evaluate the impact of economic indicators like UK GDP per capita and local population on attendance.
- Compare attendance patterns across the six clubs and their respective regions.
- Identify differences in fan engagement by analyzing attendance-to-capacity ratios.
- Use inferential statistical techniques to determine the significance of different variables.
- Apply time series models such as ARIMA to forecast future attendance trends.
- Offer strategic insights for clubs and football authorities to enhance fan engagement and optimize stadium utilization.

Data Description:

A curated dataset containing annual records for each club from 1949 to 2019 was used. The dataset includes variables such as:

- TEAM : Contains the Premier League Team name.
- SEASON: Individual seasons from 1949/1950 to 2018/2019 for each team.
- ATTENDANCE: Average attendance data of every team Team, for every season from 1949/1950 to 2018/2019.
- TIER: Data on which tier of English Football the team was participating in the particular season.
- LEAGUE.POSITION : Stadium capacity of the Premier League team for the particular season.
- STADIUM.CAPACITY: Stadium capacity of the Premier League team for the particular season.
- REGION: The Region of United Kingdom the teams are located in.

- SQUAD.VALUE..in.millions.of.Pounds. : The concept of squad value and market value of players did not exist till very recently. Hence squad value of teams for seasons are only available from 2004, as per data from transfermarkt.com
- TICKET.PRICES..in.millions.of.Pounds : There is no historical and well documented records of ticket prices from earlier decades to get exact Ticket prices for each season. Hence using various Internet Sources and official Club websites, an estimated price of tickets per season has been compiled. These are not accurate prices but rather meant to reflect the trend of Ticket Prices over the years.
- POPULATION: Population of the region for that particular year. The data provided is compiled from various sources, including the Office for National Statistics (ONS) and other demographic research.
- MATCHES.TELEVISED: No. of matches televised per team every season. This is also an estimate for early decades to reflect the trend of matches televised rather than exact figures.
- NO..OF.GOALS.SCORED : Number of goals scored in a particular season by the team.
- CUP.WINS: An indicator type of variable that contains 1 if the team has won any cup competition that season and 0 otherwise.
- RENOVATION: An indicator type of variable that contains 1 if the team had stadium renovations or stadium relocations that season and 0 otherwise.
- UK.GDP.PC..in.Pounds. : Per Captita GDP of the United Kingdom from 1954 to 2019. The data provided is compiled from various sources, including the Office for National Statistics (ONS) and other demographic research.

The dataset was loaded in R and further statistical analysis were performed accordingly.

```
'data.frame': 420 obs. of  15 variables:
 $ TEAM                              : chr  "Manchester City" "Manchester City" "Manchester City" "Manchester
 $ SEASON                            : chr  "1949/1950" "1950/1951" "1951/1952" "1952/1953" ...
 $ ATTENDANCE                        : int  39381 38677 38397 34053 30203 35217 32198 29935 32776 32568 ...
 $ STADIUM.CAPACITY                  : int  52000 52000 52000 52000 52000 52000 52000 52000 52000 52000 ...
 $ TIER                              : int  1 2 1 1 1 2 1 1 1 1 ...
 $ LEAGUE.POSITION                   : int  21 1 15 20 22 7 4 2 21 20 ...
 $ REGION                            : chr  "Manchester" "Manchester" "Manchester" "Manchester" ...
 $ SQUAD.VALUE..in.millions.of.Pounds.: num  NA NA NA NA NA NA NA NA NA NA ...
 $ TICKET.PRICES..in.Pounds.         : num  0.11 0.12 0.13 0.14 0.15 0.16 0.17 0.18 0.19 0.2 ...
 $ POPULATION                        : num  2422000 2423000 2423000 2424000 2424000 ...
 $ MATCHES.TELEVISED                 : int  1 1 1 1 1 1 2 2 2 2 ...
 $ NO..OF.GOALS.SCORED               : int  36 89 58 66 50 60 89 100 104 79 ...
 $ CUP.WINS                          : int  0 0 0 0 0 0 1 0 0 0 ...
 $ RENOVATION                        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ UK.GDP.PC..in.Pounds.             : int  NA NA NA NA NA 10682 10812 10972 11072 11477 ...
```

Feature Engineering:

1. Extracted end years from the season string (e.g., "1998/99" $\rightarrow$ 1999).
2. Created a new variable ATT_CAP_RATIO to quantify stadium occupancy efficiency (attendance as a percentage of stadium capacity).
3. Categorized league positions into meaningful labels: Top 5, Mid Table, and Relegation.

```
'data.frame': 420 obs. of  18 variables:
 $ TEAM                              : chr  "Manchester City" "Manchester City" "Manchester City" "Manchester
 $ SEASON                            : chr  "1949/1950" "1950/1951" "1951/1952" "1952/1953" ...
 $ ATTENDANCE                        : int  39381 38677 38397 34053 30203 35217 32198 29935 32776 32568 ...
 $ STADIUM.CAPACITY                  : int  52000 52000 52000 52000 52000 52000 52000 52000 52000 52000 ...
```

```
$ TIER                                : Factor w/ 3 levels "1","2","3": 1 2 1 1 1 2 1 1 1 1 ...
$ LEAGUE.POSITION                     : num  21 1 15 20 22 7 4 2 21 20 ...
$ REGION                              : chr  "Manchester" "Manchester" "Manchester" "Manchester" ...
$ SQUAD.VALUE..in.millions.of.Pounds.: num  NA NA NA NA NA NA NA NA NA NA ...
$ TICKET.PRICES..in.Pounds.           : num  0.11 0.12 0.13 0.14 0.15 0.16 0.17 0.18 0.19 0.2 ...
$ POPULATION                          : num  2422000 2423000 2423000 2424000 2424000 ...
$ MATCHES.TELEVISED                   : int  1 1 1 1 1 1 2 2 2 2 ...
$ NO..OF.GOALS.SCORED                 : int  36 89 58 66 50 60 89 100 104 79 ...
$ CUP.WINS                            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
$ RENOVATION                          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
$ UK.GDP.PC..in.Pounds.               : int  NA NA NA NA NA 10682 10812 10972 11072 11477 ...
$ years                               : num  1950 1951 1952 1953 1954 ...
$ ATT_CAP_RATIO                       : num  0.757 0.744 0.738 0.655 0.581 ...
$ POSITION.CATEGORY                   : Factor w/ 3 levels "Top 5","Mid Table",..: 3 1 2 3 3 2 1 1 3 3 ...
```

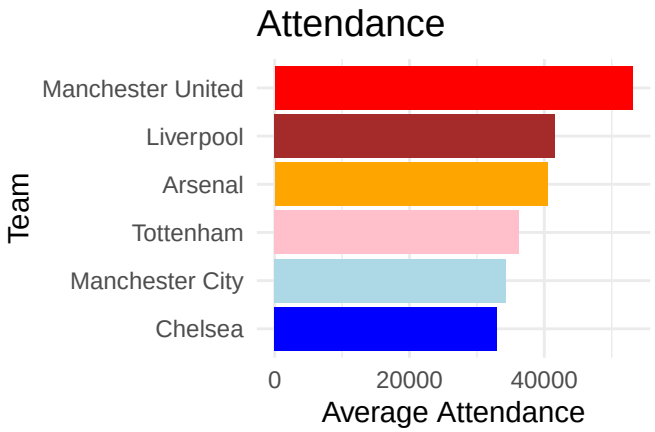
Exploratory Data Analysis (EDA):

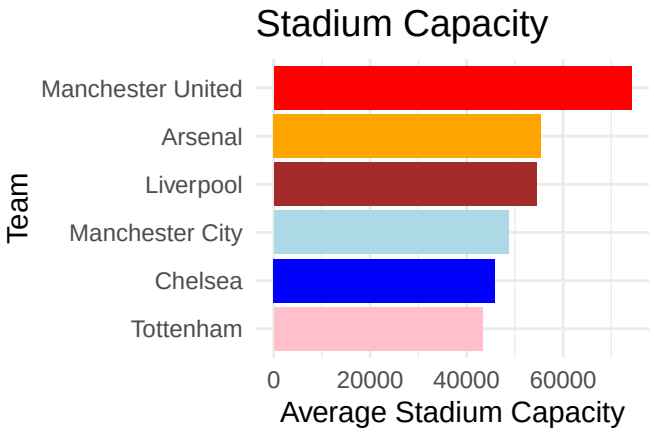
Let us get a view of the basic summary statistics of the numeric variables present in our dataset.

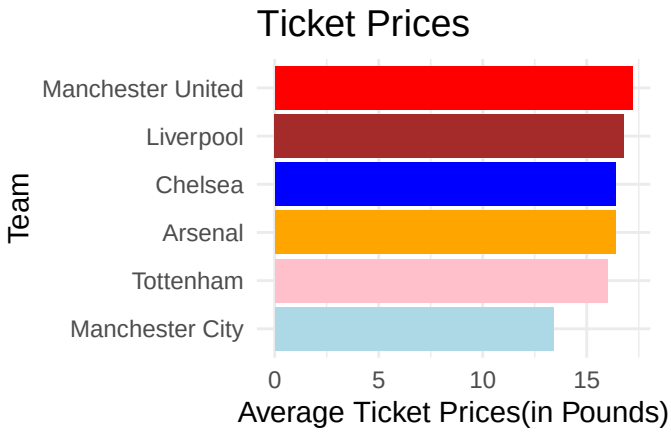
```
  ATTENDANCE    STADIUM.CAPACITY LEAGUE.POSITION
Min.   :10215   Min.   :35000    Min.   : 1.00
1st Qu.:32297   1st Qu.:45000    1st Qu.: 2.00
Median :38427   Median :52000    Median : 5.00
Mean   :39783   Mean   :53650    Mean   : 6.84
3rd Qu.:44675   3rd Qu.:60000    3rd Qu.:10.00
Max.   :75826   Max.   :90000    Max.   :22.00

SQUAD.VALUE..in.millions.of.Pounds. TICKET.PRICES..in.Pounds.
Min.   :  31.13                    Min.   : 0.11
1st Qu.: 192.50                    1st Qu.: 0.28
Median : 260.10                    Median : 4.75
Mean   : 310.32                    Mean   :16.02
3rd Qu.: 366.56                    3rd Qu.:29.00
Max.   :1034.86                    Max.   :67.00
NA's   :324
  POPULATION      MATCHES.TELEVISED NO..OF.GOALS.SCORED UK.GDP.PC..in.Pounds.
Min.   : 851000   Min.   : 1.00     Min.   : 29.00      Min.   :10682
1st Qu.:2337000   1st Qu.: 5.00     1st Qu.: 57.75      1st Qu.:15886
Median :4704500   Median :13.00     Median : 66.00      Median :22757
Mean   :4784693   Mean   :17.51     Mean   : 67.66      Mean   :23258
3rd Qu.:7503000   3rd Qu.:28.00     3rd Qu.: 77.00      3rd Qu.:32445
Max.   :9177000   Max.   :68.00     Max.   :115.00      Max.   :37134
                                                        NA's   :30
    years      ATT_CAP_RATIO
Min.   :1950   Min.   :0.1964
1st Qu.:1967   1st Qu.:0.5909
Median :1984   Median :0.7621
Mean   :1984   Mean   :0.7585
3rd Qu.:2002   3rd Qu.:0.9671
Max.   :2019   Max.   :1.2335
```

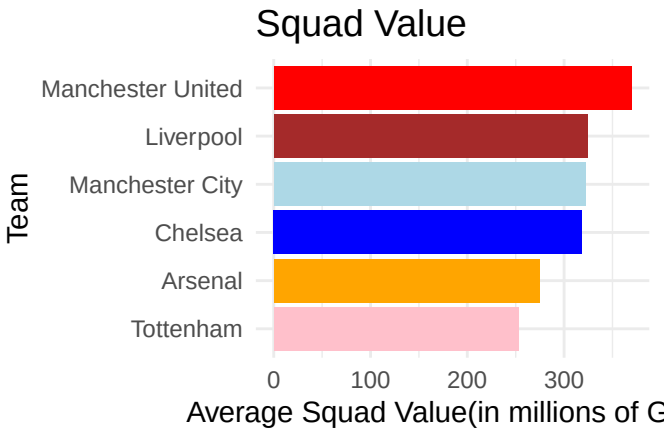
Some Basic Plots and Takeways:

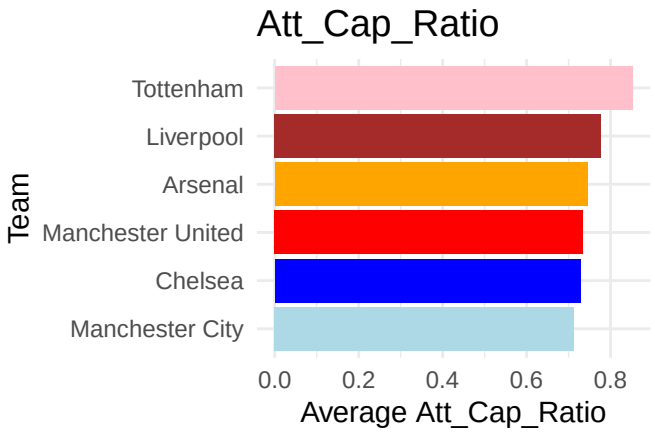
Team Wise Comparisons:

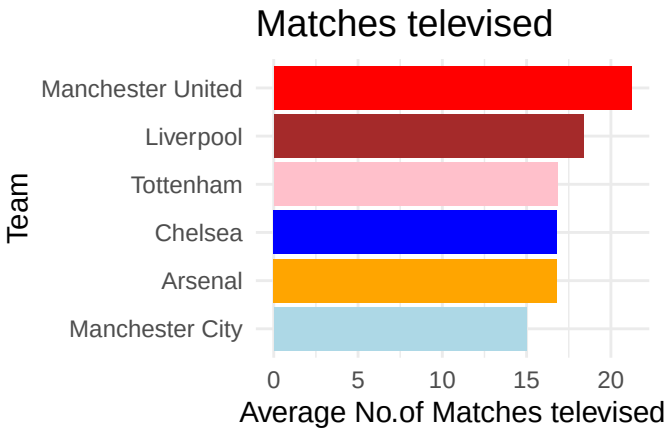










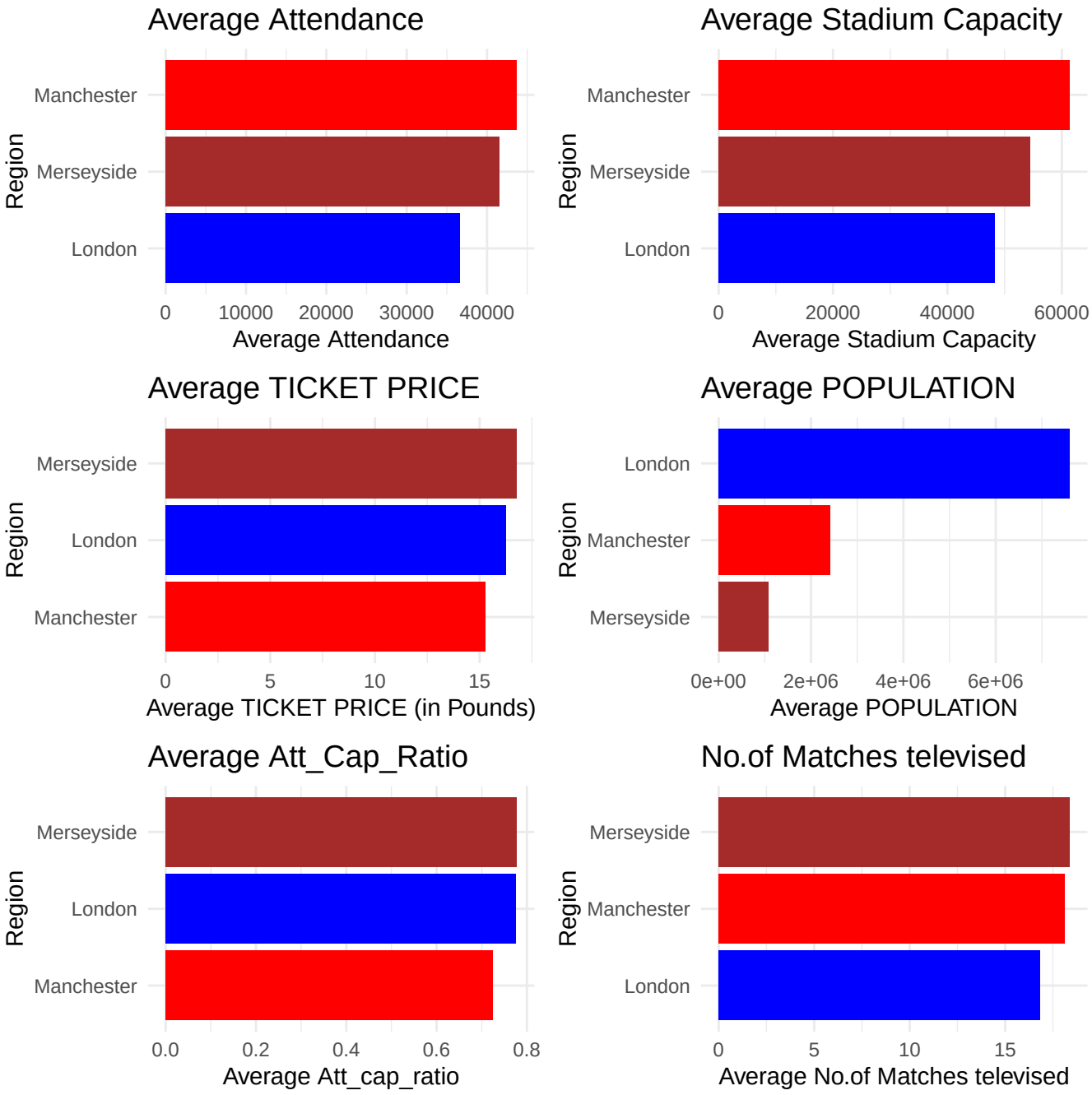


Key Takeaways:

- As expected the two giants of English Football, Manchester United and Liverpool are leading the race in all key metrics, a testimony of their rich winning culture through the decades.
- The Attendance to Stadium Capacity ratio is a very important metric in order to understand how often during the season, a particular club had a sold out stadium. Surprisingly Tottenham Hostspurs are the leaders and by a distance in comparisons to Liverpool, Arsenal and Manchester United, although it is clear they had less average attendance and stadium capacity through the last 70 decades. This is something that is worth looking into deeper later on.

Region Wise Comparisons:

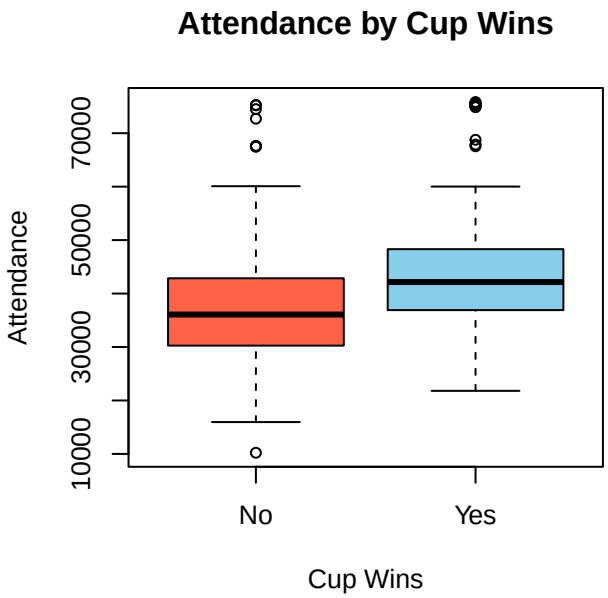
1. London: Chelsea, Arsenal, Tottenham
2. Manchester: Manchester United, Manchester City
3. Merseyside: Liverpool

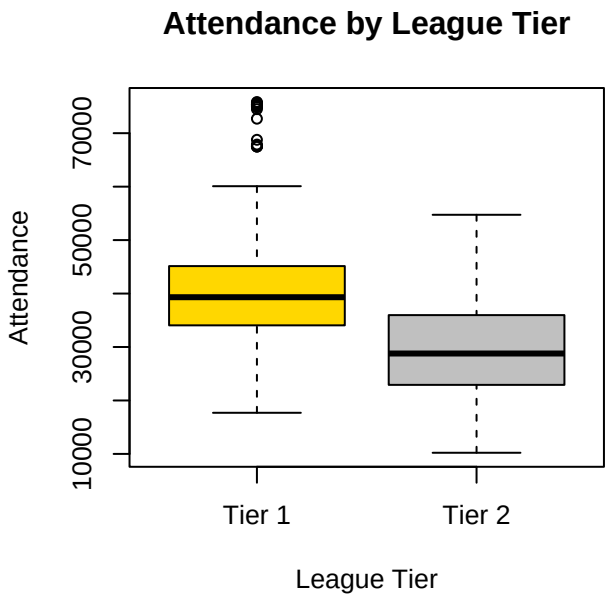


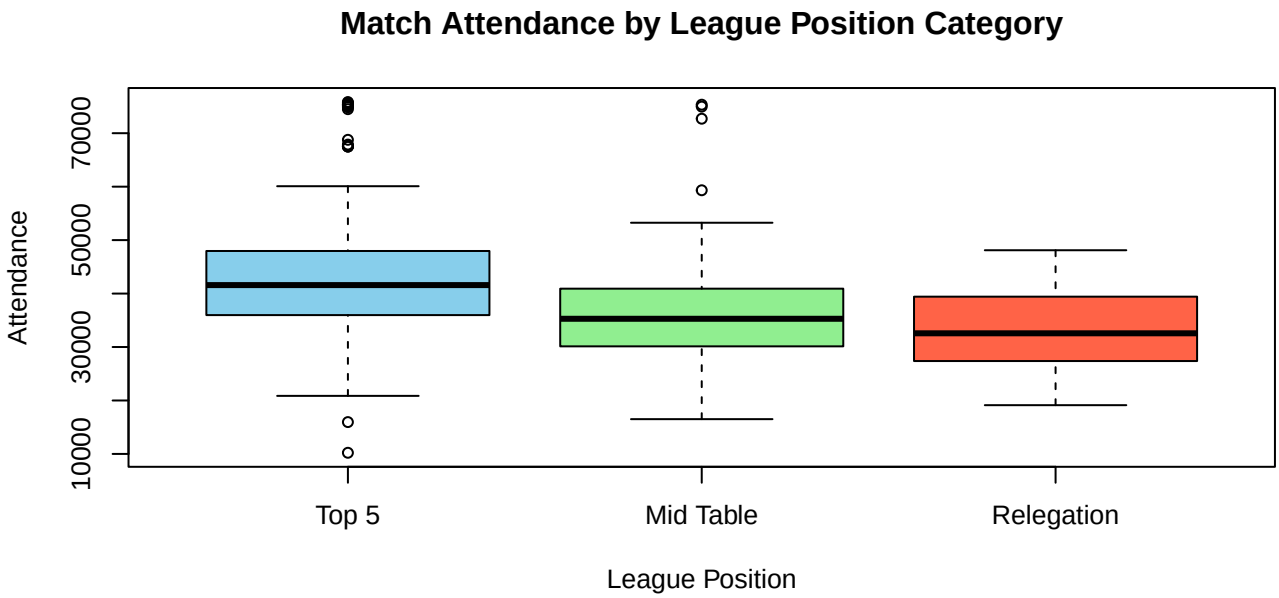
Key Takeaways:

- Although London by far is the most Populated Region, its Merseyside and Manchester clubs who are ruling the charts when it comes to Football metrics.
- Merseyside which is home to Liverpool FC have slightly edged out London clubs in the Attendance to Capacity ratio.

Let us also look at few **Categorical Variables** and how they have impacted match attendance.



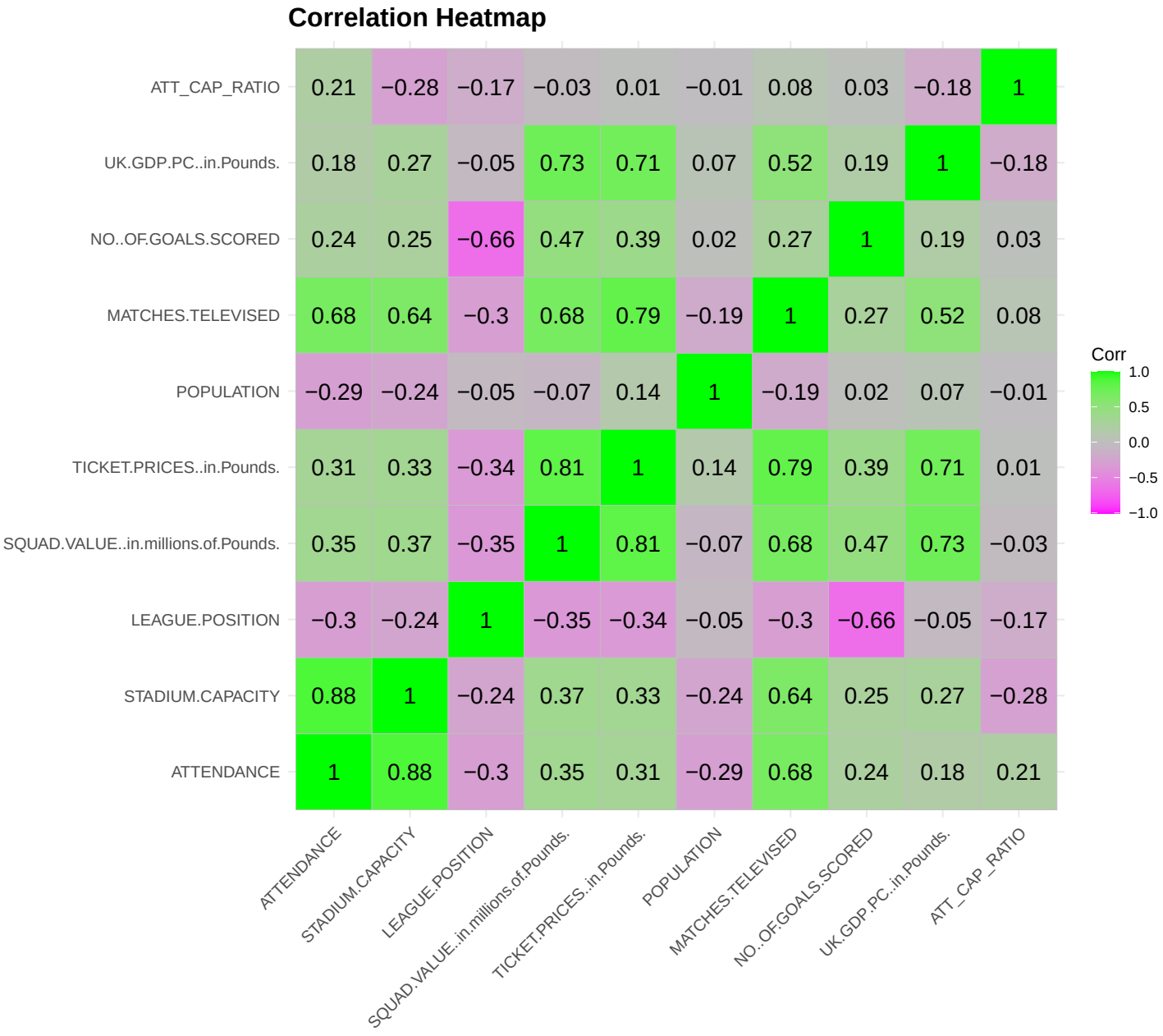




Key Takeaways:

- Winning cups clearly boosts fan turnout, with cup-winning teams enjoying noticeably higher attendances.
- Top-tier football (Tier 1) consistently pulls in far bigger crowds than Tier 2 matches.
- Finishing in the top 5 drives the biggest stadium buzz, with attendance peaking for high-performing teams.
- Mid-table sides draw respectable crowds but still trail behind the league's top performers.
- Relegation battles struggle to pack the stands, with the lowest attendance averages of all categories.

Now let us look at how the **Numeric Variables** are correlated to each other. This will help us later during model building and regression.

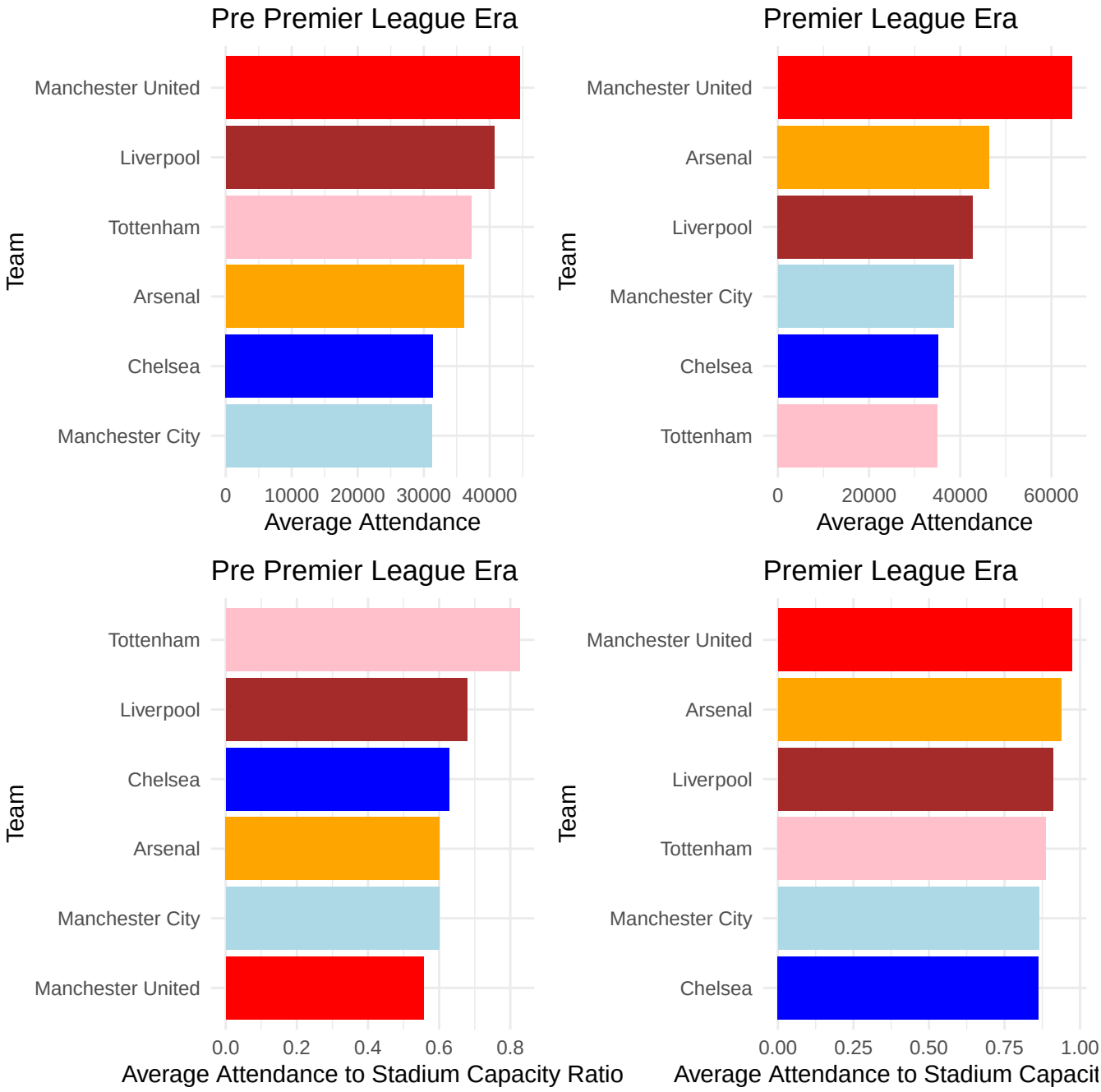


Key Takeaways:

- Bigger stadiums and stronger on-field performance both drive higher attendances.
- Wealth and exposure matter — higher GDP, ticket prices, and televised matches go hand-in-hand with stronger squads.
- Population size has little impact on how many fans turn up.

The Premier League Is Born:

The formation of the Premier League in 1992 marked a transformative era for English football, fundamentally reshaping match attendance patterns for the Big Six clubs. Let us try to visualize its impact with some plots.

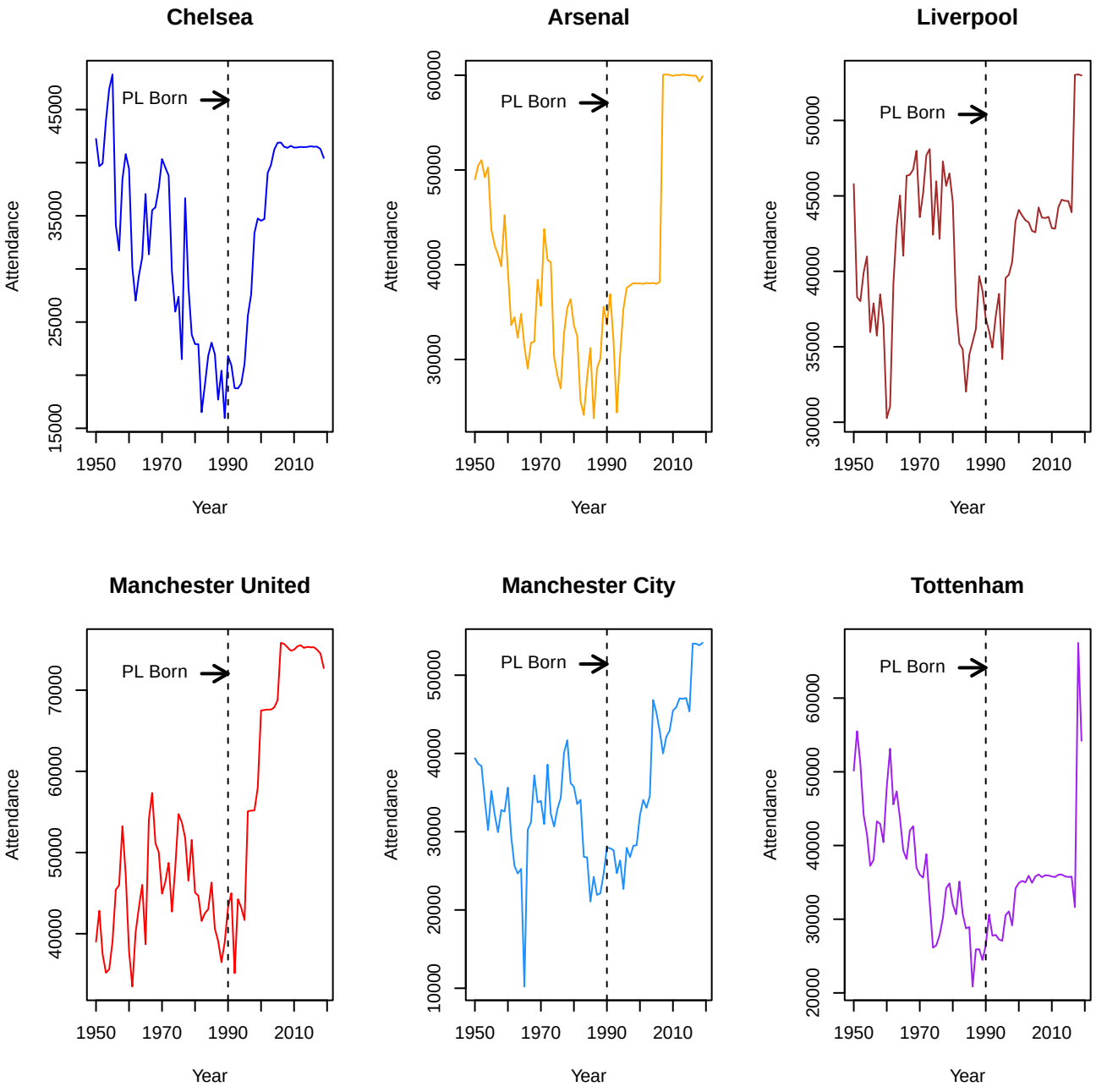


Key Takeaways:

- Manchester United had the highest average attendance in both eras, with a clear lead over other clubs.
- All clubs saw a rise in average attendance in the Premier League era compared to the Pre Premier League era.
- Manchester City and Chelsea significantly improved their average attendance rankings in the Premier League era.
- Tottenham had the highest stadium utilization ratio in the Pre Premier League era.
- Stadium utilization improved for all clubs in the Premier League era, with most reaching near full capacity.

We further perform **Time Series Analysis** of match Attendance and Attendance to Capacity ratio to uncover some more insightful facts and numbers.

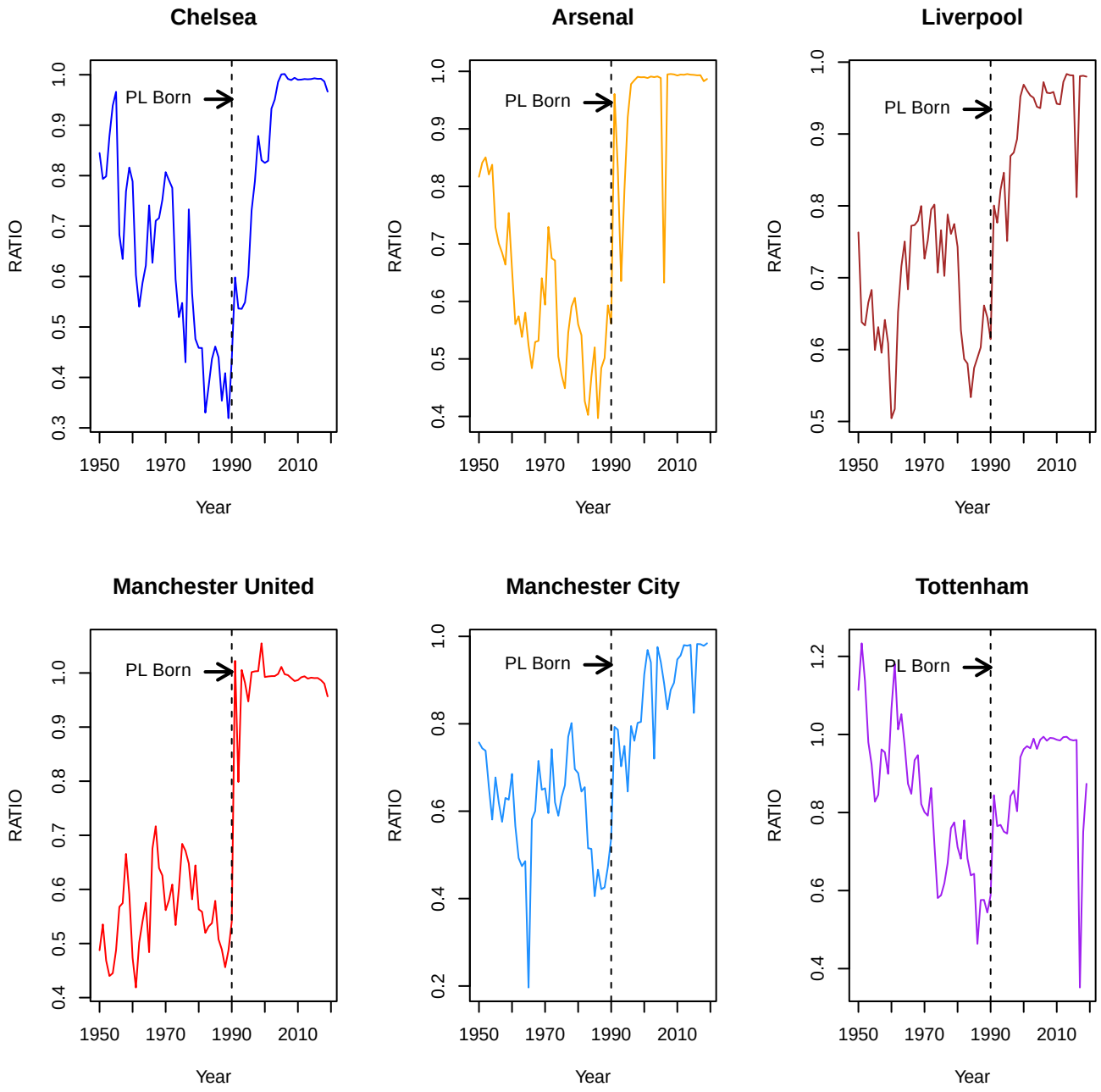
MATCH ATTENDANCE OVER THE YEARS



Key Takeaways:

- The Premier League era sparked a surge in attendances for most clubs, with Manchester United and Arsenal seeing the most dramatic jumps.
- Once-struggling clubs like Manchester City and Chelsea experienced steady post-PL growth, fueled by investment and on-field success.
- Traditional giants like Liverpool rebounded strongly after periods of decline, showcasing the PL's influence to reignite fan engagement.

ATTENDANCE TO CAPACITY RATIO



Key Takeaways:

- The Premier League era transformed stadium utilization, with Chelsea, Arsenal, and Manchester United hitting and maintaining near-perfect capacity ratios.
- Manchester City saw the biggest leap, rising from under 0.8 to almost full capacity after the PL's launch.
- Liverpool steadily climbed to maximum utilization, showing consistent fan loyalty and demand.
- Tottenham's pre-PL ratios exceeded capacity at times, later stabilizing around 1.0 with modern stadium improvements.
- Across all clubs, the PL era marks a clear shift toward consistently packed stadiums.

Quick Summary:

CLUB	Avg Attendance (Pre-PL)	Avg Attendance (Post-PL)	Attendance Growth %	Utilization Pre-PL	Utilization Post-PL	Utilization Change %
Manchester United	~40,000	~60,000	+50%	~0.55	~0.98	+78%
Arsenal	~30,000	~50,000	+67%	~0.50	~0.98	+96%
Liverpool	~37,000	~45,000	+22%	~0.70	~0.98	+40%
Manchester City	~20,000	~45,000	+125%	~0.35	~0.97	+177%
Chelsea	~25,000	~40,000	+60%	~0.55	~0.98	+78%
Tottenham	~33,000	~45,000	+36%	~0.75	~0.97	+29%

Key Insights from the Table:

- Biggest Attendance Growth: Manchester City (+125%) and Arsenal (+67%) led the surge post-PL due to new stadiums, investments, and on-field success.
- Highest Stadium Utilization Gains: Manchester City (+177%) and Arsenal (+96%) transformed from half-empty grounds to near-full capacity.
- Manchester United's Consistency: Already leading in attendance pre-PL, United consolidated dominance with +50% growth and 98% utilization.
- Liverpool's Stability: Modest attendance growth (+22%) but consistently high utilization, reflecting a loyal fan base.
- Tottenham's Early Strength: Had the highest pre-PL utilization (75%) and maintained near-capacity crowds post-PL.

Multiple Linear Regression:

We use multiple linear regression to model **Match Attendance**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 C X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \varepsilon_i$$

- Y= Match Attendance
- X_1= Ticket Prices (in Pounds)
- X_2= TIER
- X_3= Stadium Capacity
- X_4= Matches Televised
- X_5= Cup Wins
- X_6= League Postion Category
- X_7= UK GDP per capita
- X_8= No. of Goals scored
- ε = Random error

Now we fit a regression line to the our given data. We have used **Ordianry Least Square** method. We use the lm( ) function in R to facilitate in our calculations.

```
Call:
lm(formula = ATTENDANCE ~ TICKET.PRICES..in.Pounds. + POSITION.CATEGORY +
    MATCHES.TELEVISED + UK.GDP.PC..in.Pounds. + STADIUM.CAPACITY +
    CUP.WINS + TIER + NO..OF.GOALS.SCORED, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-35952  -4174    172   4248  15467

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.665e+04  4.165e+03   3.997 7.71e-05 ***
TICKET.PRICES..in.Pounds.   6.365e+01  7.201e+01   0.884  0.37729
POSITION.CATEGORYMid Table -2.512e+03  8.494e+02  -2.958  0.00329 **
POSITION.CATEGORYRelegation -2.618e+03  1.596e+03  -1.640  0.10186
MATCHES.TELEVISED           7.691e+02  1.226e+02   6.274 9.64e-10 ***
UK.GDP.PC..in.Pounds.      -7.883e-01  1.419e-01  -5.554 5.27e-08 ***
STADIUM.CAPACITY            3.846e-01  3.031e-02  12.689  < 2e-16 ***
CUP.WINS1                   1.700e+03  7.822e+02   2.173  0.03041 *
TIER2                      -7.404e+03  1.300e+03  -5.695 2.47e-08 ***
TIER3                      -4.818e+02  6.604e+03  -0.073  0.94188
NO..OF.GOALS.SCORED         9.459e+01  2.953e+01   3.203  0.00147 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6533 on 379 degrees of freedom
  (30 observations deleted due to missingness)
Multiple R-squared:  0.7142,Adjusted R-squared:  0.7066
F-statistic:  94.7 on 10 and 379 DF,  p-value: < 2.2e-16
```

We interpret the following from the R output:

- Multiple R-squared = 0.7142: This means that approximately 71.42% of the total variation in attendance is explained by our model. This indicates a strong model fit.
- F-statistic = 94.7, p-value < 2.2e-16: The overall model significance test suggests that the regression model is highly significant at the 1% level.
- Intercept = 16650, p-value = 7.71e-05: This represents the expected attendance when all predictors are zero. The effect is statistically significant at the 5% level.
- TICKET.PRICES..in.Pounds. = 65.65, p-value = 0.37729: For every 1-pound increase in ticket price, the predicted attendance increases by about 65.65. This effect is not significantly different from zero at the 5% level.
- POSITION.CATEGORYMid Table = -21582.48, p-value = 0.00329: Teams in the mid-table position have, on average, 21,582 fewer attendees compared to the baseline category. This effect is statistically significant at the 5% level.
- POSITION.CATEGORYRelegation = -26187.93, p-value = 0.10186: Teams in the relegation category have, on average, 26,188 fewer attendees compared to the baseline category. This effect is not significantly different from zero at the 5% level.
- MATCHES.TELEVISED = 7696.91, p-value = 9.64e-10: Each televised match is associated with an increase of about 7,697 in attendance. This effect is statistically significant at the 5% level.
- UK.GDP.PC..in.Pounds. = -7838.02, p-value = 5.27e-08: Higher GDP per capita is associated with a decrease in attendance. This effect is statistically significant at the 5% level.
- STADIUM.CAPACITY = 3846.03, p-value < 2e-16: For every unit increase in stadium capacity, predicted attendance increases by about 3,846. This effect is statistically significant at the 5% level.
- CUP.WINS1 = 17030.77, p-value = 0.03041: Having a cup win increases attendance by about 17,031. This effect is statistically significant at the 5% level.
- TIER2 = -7404.08, p-value = 2.47e-08: Tier 2 teams have, on average, 7,404 fewer attendees compared to the baseline tier. This effect is statistically significant at the 5% level.
- TIER3 = -4818.02, p-value = 0.94188: Tier 3 teams have, on average, 4,818 fewer attendees compared to the baseline tier. This effect is not significantly different from zero at the 5% level.
- NO..OF.GOALS.SCORED = 9459.04, p-value = 0.00147: Each additional goal scored is associated with an increase in attendance of about 9,459. This effect is statistically significant at the 5% level.

Thus we find that our model shows a relatively good fit although there are few predictors that are not statistically significant at the 5% level.

Let us further delve into our model and use statistical tests to check for **Heteroscedascity and Multicollinearity**.

Problem of Heteroscedascity:

We have used the Goldfeld-Quandt Test for this.

Let, H_0: $\sigma_1^{*^2} = \sigma_2^{*^2}$ (Homoscedasticity) vs. H_1: $\sigma_1^{*^2} < \sigma_2^{*^2}$ (Heteroscedasticity.)

We have chosen to remove $\frac{1}{3}rd$ of our data from the centre to perform the test. Let, α= 0.05

We have used R to facilate in our computations.

```
Goldfeld-Quandt test

data:  model2
GQ = 1.8242, df1 = 119, df2 = 119, p-value = 0.0005816
alternative hypothesis: variance increases from segment 1 to 2
```

We interpet the following from the R output:

- Test Statistic (GQ) = 1.8242 → This is the Goldfeld-Quandt test statistic.
- Degrees of Freedom (df1 = 119, df2 = 119) → The degrees of freedom for both groups used in the test.
- p-value = 0.0005816 → The p-value is very small, which is less than 0.05.
- Alternative Hypothesis: "Variance increases from segment 1 to 2" → The test assumes the second segment has a larger variance than the first.

Since the p-value (0.0005816) is much lower than 0.05, we reject the null hypothesis , which assumes homoscedasticity. This means **heteroscedasticity is present.**

Consequences of Heteroscedasticity:

- The OLS estimators, although unbiased are no longer minimum variance.
- In hypothesis testing, since s.e.($\widehat{\beta_j}$) is larger than it should be, the acceptance regions are wider and hence a $\widehat{\beta_j}$ may be deemed insignificant even if it's not.
- The confidence interval of $\widehat{\beta_j}$ will be wider and hence less useful.

We shall now use **GLS** to estimate our parameters and refit our model.

```
Generalized least squares fit by REML
  Model: ATTENDANCE ~ TICKET.PRICES..in.Pounds. + POSITION.CATEGORY +      MATCHES.TELEVISED + UK.GDP.PC..in.P
  Data: data_clean
       AIC      BIC    logLik
  7851.156 7902.344 -3912.578

Variance function:
 Structure: Power of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
    power
0.4082395

Coefficients:
                              Value Std.Error   t-value p-value
(Intercept)                16100.941  4118.860  3.909077  0.0001
TICKET.PRICES..in.Pounds.    149.691    76.911  1.946282  0.0524
POSITION.CATEGORYMid Table -2299.604   830.744 -2.768124  0.0059
POSITION.CATEGORYRelegation -2234.744  1479.178 -1.510801  0.1317
MATCHES.TELEVISED             616.209   141.174  4.364905  0.0000
UK.GDP.PC..in.Pounds.          -0.726     0.147 -4.921155  0.0000
STADIUM.CAPACITY                0.387     0.031 12.446718  0.0000
CUP.WINS1                    1768.978   792.852  2.231157  0.0263
TIER2                       -7334.332  1157.411 -6.336844  0.0000
TIER3                        -704.184  5787.025 -0.121683  0.9032
NO..OF.GOALS.SCORED            97.990    29.117  3.365414  0.0008

 Correlation:
                            (Intr) TICKET POSITT POSITI MATCHE UK.GDP STADIU
TICKET.PRICES..in.Pounds.    0.192
POSITION.CATEGORYMid Table  -0.436 -0.022
POSITION.CATEGORYRelegation -0.344 -0.076  0.430
MATCHES.TELEVISED            0.245 -0.792  0.007  0.052
UK.GDP.PC..in.Pounds.       -0.763  0.067  0.107  0.093 -0.631
```

```
STADIUM.CAPACITY            -0.546  0.198  0.093  0.062 -0.364  0.423
CUP.WINS1                   -0.026  0.159  0.235  0.175 -0.148  0.024 -0.004
TIER2                       -0.002  0.074  0.148  0.065  0.010 -0.050  0.089
TIER3                        0.013  0.002  0.079  0.049  0.060 -0.095  0.033
NO..OF.GOALS.SCORED         -0.667 -0.310  0.427  0.342  0.102  0.276 -0.087
                           CUP.WI TIER2  TIER3
TICKET.PRICES..in.Pounds.
POSITION.CATEGORYMid Table
POSITION.CATEGORYRelegation
MATCHES.TELEVISED
UK.GDP.PC..in.Pounds.
STADIUM.CAPACITY
CUP.WINS1
TIER2                        0.257
TIER3                        0.058  0.052
NO..OF.GOALS.SCORED         -0.083 -0.164 -0.005

Standardized residuals:
        Min          Q1         Med          Q3         Max
-4.49402340 -0.66712441  0.01068846  0.63163330  2.47547806

Residual standard error: 86.87547
Degrees of freedom: 390 total; 379 residual
```

We interpret the following from the GLS output:

- AIC = 7851.156, BIC = 7902.344 → These are model fit statistics; lower values indicate a better fit when comparing models.
- Variance function (Power) = 0.4082 → This suggests the residual variance increases with fitted values, but less than proportionally (heteroscedasticity modeled via a power function).
- (Intercept) = 16100.941, p-value = 0.0001 → Statistically significant at the 5% level; when all predictors are zero, the expected attendance is 16,100.94.
- TICKET.PRICES..in.Pounds. = 149.691, p-value = 0.0524 → Not significant from zero at the 5% level, but marginally significant at the 10% level; each £1 increase in ticket price is associated with an increase in attendance of about 149.69.
- POSITION.CATEGORY (Mid Table) = -2299.604, p-value = 0.0059 → Significant; being in the mid table is associated with ~2,299 fewer attendees compared to the baseline category.
- POSITION.CATEGORY (Relegation) = -2234.744, p-value = 0.1317 → Not significant from zero at the 5% level; suggests ~2,235 fewer attendees for relegation teams, but with weak evidence.
- MATCHES.TELEVISED = 616.209, p-value $<$ 0.0001 → Significant; televised matches increase attendance by ~616 on average.
- UK.GDP.PC..in.Pounds. = -0.726, p-value $<$ 0.0001 → Significant; higher GDP per capita is associated with slightly lower attendance.
- STADIUM.CAPACITY = 0.387, p-value $<$ 0.0001 → Significant; each additional seat in stadium capacity increases attendance by ~0.39.
- CUP.WINS1 = 1768.978, p-value = 0.0263 → Significant; teams with 1 cup win attract ~1,769 more attendees.
- TIER2 = -7334.332, p-value $<$ 0.0001 → Significant; being in tier 2 is associated with ~7,334 fewer attendees than tier 1.
- TIER3 = -704.184, p-value = 0.9032 → Not significant from zero at the 5% level; the estimate is small and imprecise.
- NO..OF.GOALS.SCORED = 97.990, p-value = 0.0008 → Significant; each additional goal scored increases attendance by ~98.

But is our GLS Model better than our OLS Model? Let us check using two criterias: **AIC and BIC.**

```
          df      AIC
model2    12 7971.588
model_gls 13 7851.156
          df      BIC
model2    12 8019.182
model_gls 13 7902.344
```

Both metrics dropped substantially for GLS (over 100 points in AIC), which is a welcome improvement.

That means the variance structure we added in GLS is explaining heteroscedasticity and improving overall fit.

GLS clearly outperforms OLS in terms of model fit.

Problem of Multicollinearity:

We shall try to detect Multicollinearity using Variance Inflation Factor.

We shall use the Thumb Rule that if $\text{VIF}_j > 10$ for any j, then we shall suspect multicollinearity involving x_j. We have used R to facilitate in our computations.

```
                               GVIF Df GVIF^(1/(2*Df))
TICKET.PRICES..in.Pounds. 19.824579  1        4.452480
POSITION.CATEGORY          1.656349  2        1.134457
MATCHES.TELEVISED         33.461827  1        5.784620
UK.GDP.PC..in.Pounds.     13.647829  1        3.694297
STADIUM.CAPACITY           1.411195  1        1.187937
CUP.WINS                   1.247043  1        1.116711
TIER                       1.214984  2        1.049887
NO..OF.GOALS.SCORED        1.791096  1        1.338318
```

Interpretation:

- High GVIF (>10 or GVIF^(1/(2*Df)) > ~2) indicates multicollinearity concern.
- Here, TICKET.PRICES..in.Pounds. (GVIF 19.82, adj. 4.45), MATCHES.TELEVISED (33.46, adj. 5.78), and UK.GDP.PC..in.Pounds. (13.65, adj. 3.69) are showing high multicollinearity.
- Other variables have low GVIFs (~1–2), so multicollinearity is not a big issue there.

Consequences of Multicollinearity:

- The OLS estimators, although still BLUE, the "minimum" variance will turn out to be very large (and even maybe infinity).
- Makes the estimators unstable.
- The confidence interval of $\widehat{\beta_j}$ will be wider and hence less useful.

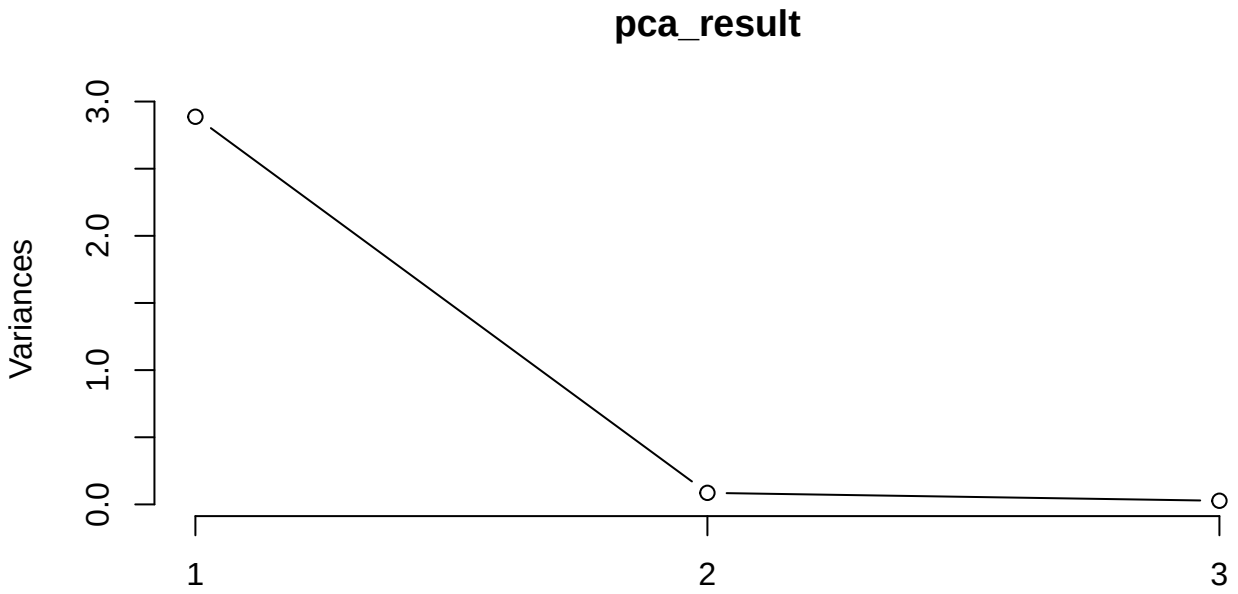
Remedial Steps- Principle Component Analysis:

We will run **PCA** on the three highly collinear numeric variables: TICKET.PRICES..in.Pounds., MATCHES.TELEVISED, and UK.GDP.PC..in.Pounds.

We shall be implementing the following steps:

1. Scale and run PCA
2. Extract PCA scores (principal components)
3. Use PC1 in our model instead of original variables

```
Importance of components:
                          PC1     PC2     PC3
Standard deviation     1.6991 0.29326 0.16465
Proportion of Variance 0.9623 0.02867 0.00904
Cumulative Proportion  0.9623 0.99096 1.00000
```



The PCA summary tells us:

- PC1 (first principal component) explains 96.23% of the variance in the three variables combined.
- PC2 and PC3 explain very little additional variance (2.87% and 0.9% respectively).
- So, PC1 alone captures almost all the information from TICKET.PRICES..in.Pounds., MATCHES.TELEVISED, and UK.GDP.PC..in.Pounds.

Let us use PC1 in our model and now check the VIF:

```
                        GVIF Df GVIF^(1/(2*Df))
PC1                 1.222559  1        1.105694
POSITION.CATEGORY   1.599157  2        1.124534
STADIUM.CAPACITY    1.111063  1        1.054070
CUP.WINS            1.203184  1        1.096897
TIER                1.171305  2        1.040321
NO..OF.GOALS.SCORED 1.424551  1        1.193545
```

Clearly, the Problem of Multicollinearity has been resolved. Let us now take a look at the final model summary:

```
Call:
lm(formula = ATTENDANCE ~ PC1 + POSITION.CATEGORY + STADIUM.CAPACITY +
    CUP.WINS + TIER + NO..OF.GOALS.SCORED, data = data_clean)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-40415  -4723     -8   4818  15715

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.603e+03  2.459e+03   1.872   0.0620 .
PC1                          3.662e+03  2.301e+02  15.917  < 2e-16 ***
POSITION.CATEGORYMid Table  -1.815e+03  8.997e+02  -2.017   0.0443 *
POSITION.CATEGORYRelegation -1.540e+03  1.688e+03  -0.912   0.3622
STADIUM.CAPACITY             4.840e-01  2.854e-02  16.955  < 2e-16 ***
CUP.WINS1                    1.885e+03  8.231e+02   2.290   0.0226 *
TIER2                       -7.575e+03  1.385e+03  -5.471 8.13e-08 ***
TIER3                       -4.249e+03  7.026e+03  -0.605   0.5457
NO..OF.GOALS.SCORED          1.474e+02  2.841e+01   5.190 3.42e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6973 on 381 degrees of freedom
Multiple R-squared:  0.6727,Adjusted R-squared:  0.6658
F-statistic: 97.88 on 8 and 381 DF,  p-value: < 2.2e-16
```

We interpret the following from the R output:

- Multiple R-squared = 0.6727: This means that approximately 67.27% of the total variation in attendance is explained by our model. This indicates a strong model fit.
- F-statistic = 97.88, p-value < 2.2e-16: The overall model significance test suggests that the regression model is highly significant at the 1% level.
- Intercept = 4603, p-value = 0.0620: This represents the expected attendance when all predictors are zero. The effect is marginally significant (slightly above the 5% level).
- PC1 = 3662, p-value < 2e-16: For every 1-unit increase in PC1 (the combined factor of ticket prices, televised matches, and GDP per capita), predicted attendance increases by about 3,662. This effect is highly statistically significant.
- POSITION.CATEGORYMid Table = -1815, p-value = 0.0443: Teams in the mid-table position have, on average, 1,815 fewer attendees compared to the baseline category. This effect is statistically significant at the 5% level.
- POSITION.CATEGORYRelegation = -1540, p-value = 0.3622: Teams in the relegation category have, on average, 1,540 fewer attendees compared to the baseline category. This effect is not significantly different from zero at the 5% level.
- STADIUM.CAPACITY = 0.484, p-value < 2e-16: For every one-unit increase in stadium capacity, predicted attendance increases by about 0.484. This effect is highly statistically significant.
- CUP.WINS1 = 1885, p-value = 0.0226: Having 1 cup win increases attendance by about 1,885. This effect is statistically significant at the 5% level.
- TIER2 = -7575, p-value = 8.13e-08: Tier 2 teams have, on average, 7,575 fewer attendees compared to the baseline tier. This effect is highly statistically significant.
- TIER3 = -4249, p-value = 0.5457: Tier 3 teams have, on average, 4,249 fewer attendees compared to the baseline tier. This effect is not significantly different from zero at the 5% level.
- NO..OF.GOALS.SCORED = 147, p-value = 3.42e-07: Each additional goal scored is associated with an increase in attendance of about 147. This effect is highly statistically significant.

Model Performance Evaluation Using Train/Test Data Split:

The step we will follow to evaluate the peformance of our model is as follows:

- Split data into training (70%) and test (30%) sets
- Fit our model (model_pca) on the training set
- Predict attendance on the test set
- Evaluate model performance using criteria like RMSE, MAE, and R-squared on test data

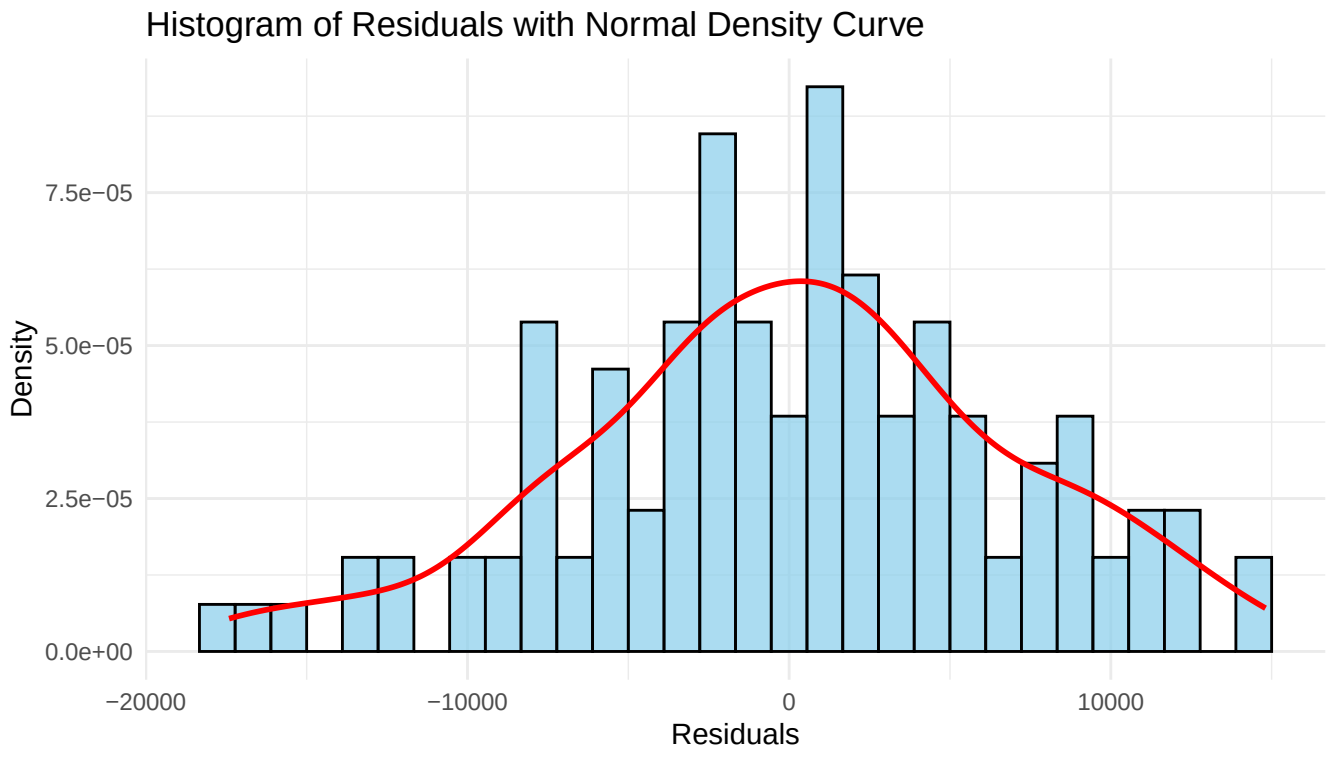
```
Test set performance:
RMSE = 6787.974
MAE = 5365.972
R-squared = 0.7114515
```

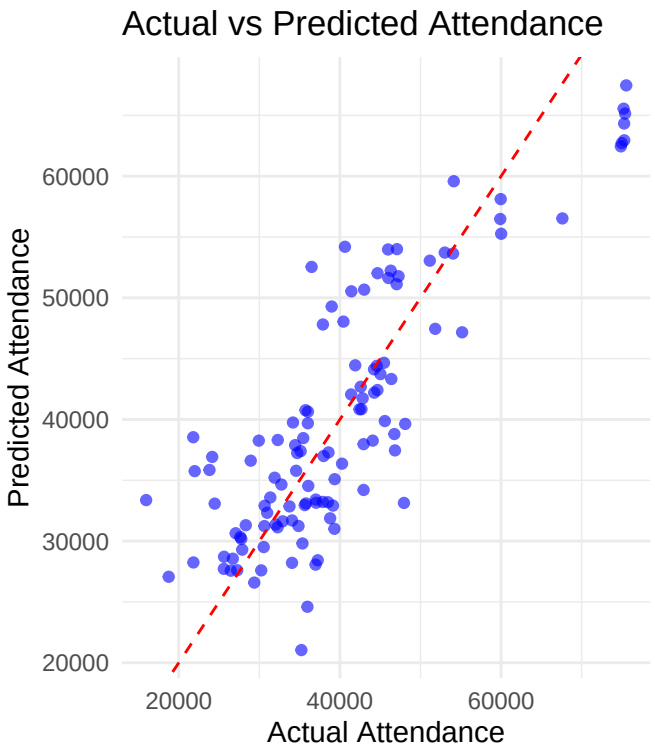
What the test set performance conveys:

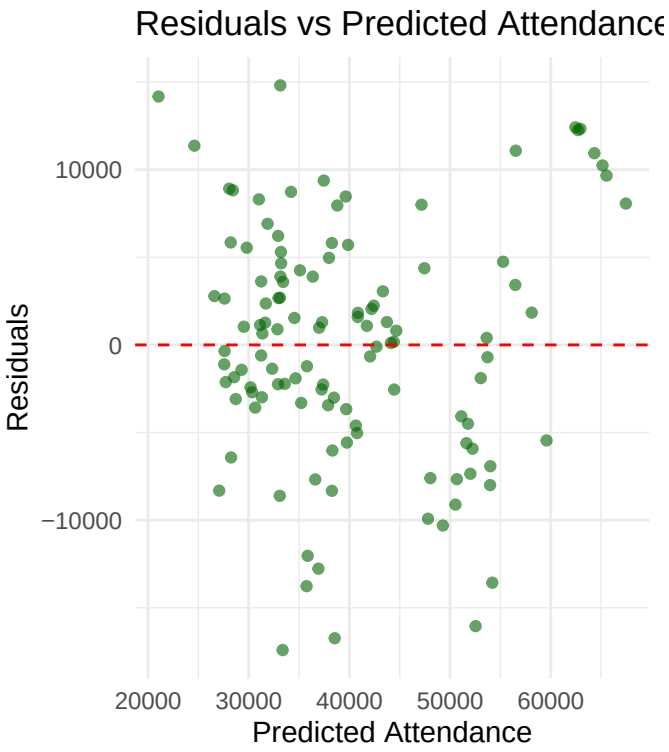
- RMSE = 6788 (approx): On average, the attendance predictions are off by about 6,788 attendees in squared-error terms — this gives an idea of typical prediction error magnitude.
- MAE = 5366 (approx): On average, the predictions differ from actual attendance by about 5,366 attendees — a more intuitive average absolute error measure.
- R-squared = 0.71: The model explains about 71% of the variance in attendance in the test data, showing strong predictive power.

The model generalizes well to unseen data with strong explanatory power and reasonable error rates. The PCA-based approach helped handle multicollinearity without hurting predictive accuracy.

Lets try to visualize the model's performance on the test set.







Key Takeaways on Model Performance:

- Histogram of Residuals with Normal Density Curve:

The residuals are approximately symmetrically distributed around zero and roughly follow a bell-shaped curve, indicating that the model errors are fairly normally distributed. This supports the assumption of normality in residuals, which is good for linear regression validity. Minor deviations from the curve at the tails suggest a few larger prediction errors, but overall the fit is reasonable.

- Actual vs Predicted Attendance Plot:

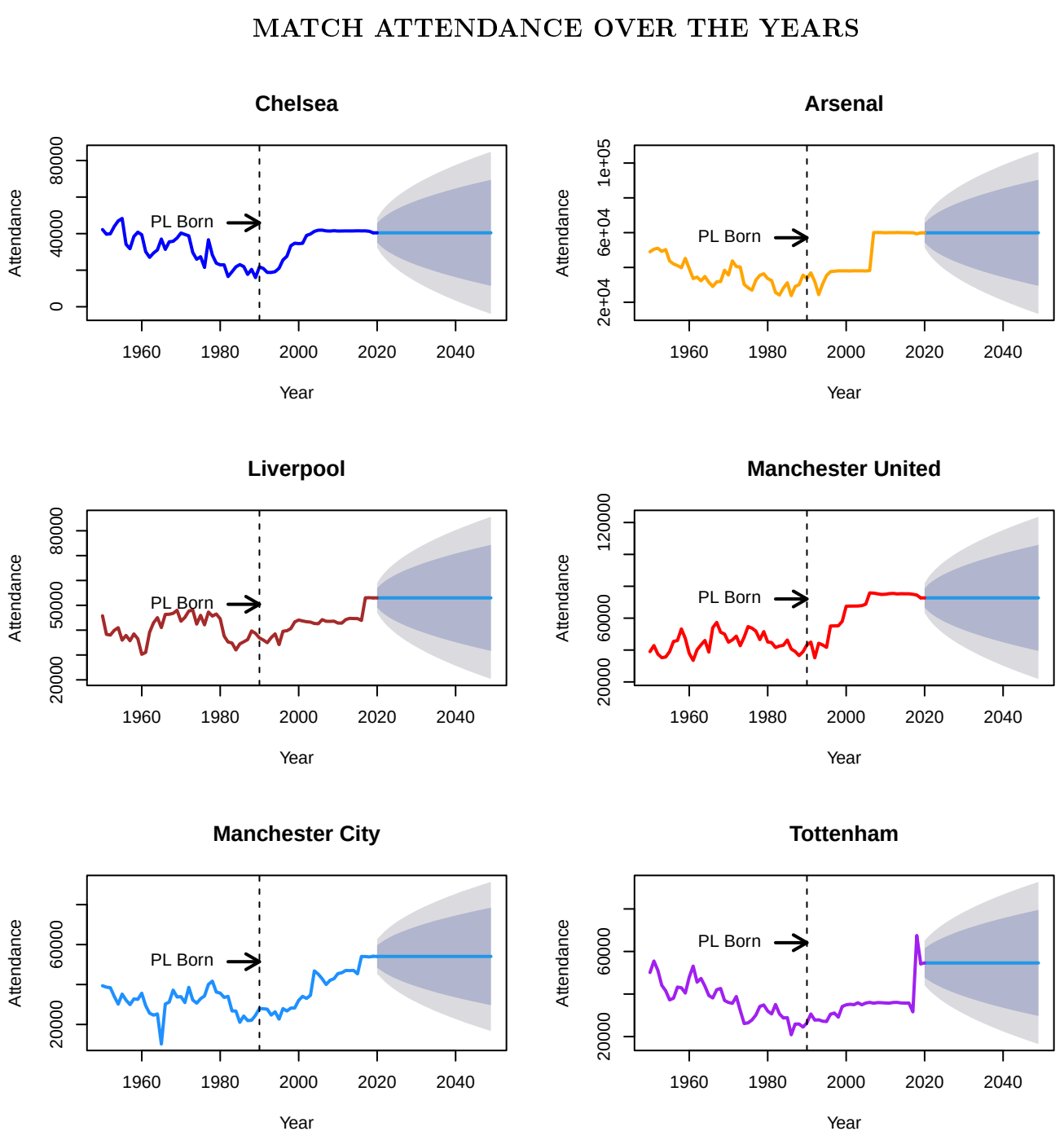
The points closely cluster around the red 45-degree dashed line, showing strong agreement between predicted and actual attendance values. This indicates the model's predictions are generally accurate across the test set. Some scatter at higher attendance values suggests slight under- or over-predictions in extreme cases, but the overall fit is strong.

- Residuals vs Predicted Attendance Plot:

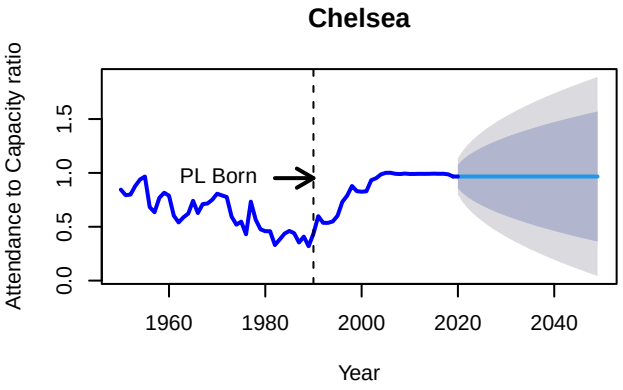
Residuals are scattered randomly around the horizontal zero line without obvious patterns or systematic trends. This indicates the model does not suffer from heteroscedasticity (changing variance with predicted values) and suggests a good model fit with constant error variance.

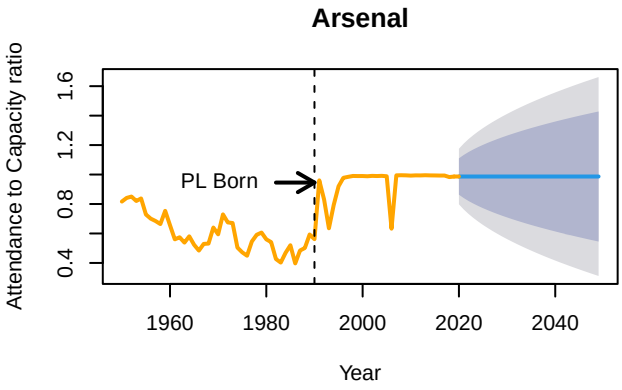
The visual diagnostics collectively demonstrate that the model with PCA handles the data well, provides accurate predictions, and meets key regression assumptions. The model generalizes effectively to test data, with errors distributed normally and predictions closely aligned with actual attendance values.

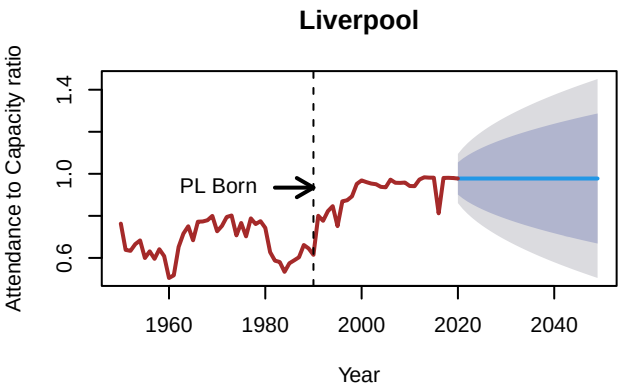
Forecasting Attendance Using ARIMA Model:

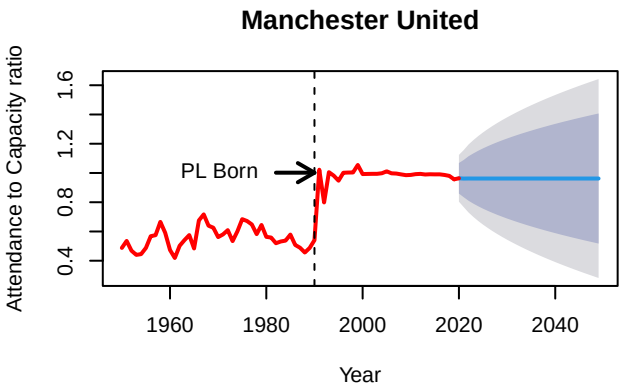


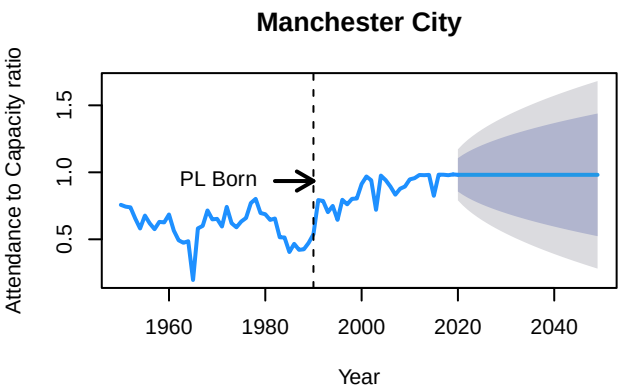
ATTENDANCE TO CAPACITY RATIO

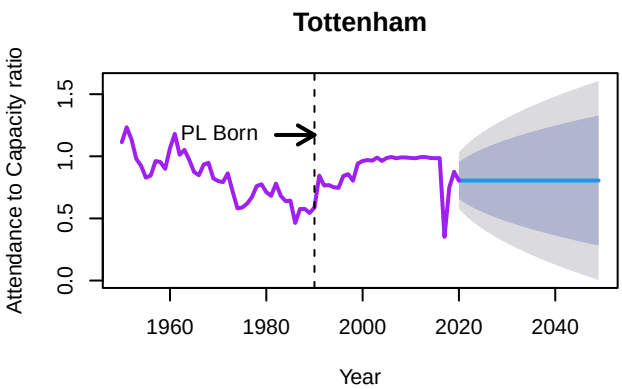












Boosting Fan Engagement & Stadium Utilization:

- Use historical attendance patterns to introduce flexible pricing based on opponent strength, league position, and day of the week.
- Offer bundled packages with merchandise, food, or digital subscriptions.
- For clubs with mid-table or relegation struggles, incentivize attendance with loyalty points, "bring-a-friend" schemes, and enhanced matchday experiences.
- Since televised matches correlate positively with attendance, integrate in-stadium digital engagement—interactive apps, second-screen experiences—to convert occasional viewers into regular attendees.
- Focus on amenities, seating comfort, accessibility, and matchday entertainment to maintain high Attendance to Capacity Ratio even in less competitive seasons.
- Clubs in regions with smaller populations can deepen ties via grassroots programs, youth academies, and open training sessions.

Future Statistical Work:

1. Incorporate external regressors such as league position, ticket prices, and cup wins into forecasting models for richer projections.
2. Use Difference-in-Differences or Synthetic Control to evaluate the impact of specific events (stadium renovation, star player signing, cup wins) on attendance.
3. Survival Analysis of Fan Retention: Model the "lifespan" of a season-ticket holder or regular attendee and factors influencing churn.
4. Integrate fan sentiment from Twitter, forums, and news sources to measure non-matchday engagement.
5. Machine Learning for Prediction: Apply Random Forests, Gradient Boosting, or Neural Networks to capture non-linear effects and complex interactions.

Conclusion:

This project successfully dissects the historical evolution of Premier League attendance, particularly for the Big Six, revealing how performance, infrastructure, economics, and broadcasting shape fan turnout.By correcting for heteroscedasticity and multicollinearity, and applying time-series models, it establishes a strong quantitative foundation for predicting future trends.

However, expanding the scope to include broader engagement metrics, exogenous forecasting variables, and more nuanced interpretation would make it even more impactful.The findings not only serve academic interest but also offer practical pathways for clubs and football authorities to sustain high stadium utilization and deepen fan loyalty in an increasingly digital era.

References:

The following websites were used for data collection and fact checking:

- Wikipedia
- TransferMarkt
- Official Club Websites of Chelsea, Arsenal, Manchester United, Manchester City, Tottenham Hotspurs, Liverpool
- The Athletic