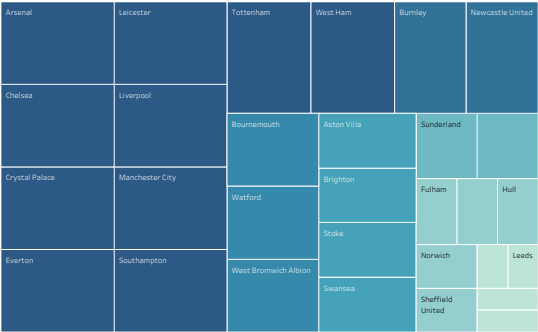


Using Expected Goals and Match Statistics to Explain Manchester United’s Premier League Match Results

Executive Summary

In this project, I used match-level Premier League data from 266 Manchester United matches between the 2014/15 and 2020/21 seasons to test which performance stats best explain points earned per match. I used R for data cleanin...

Match Distribution by Opponent

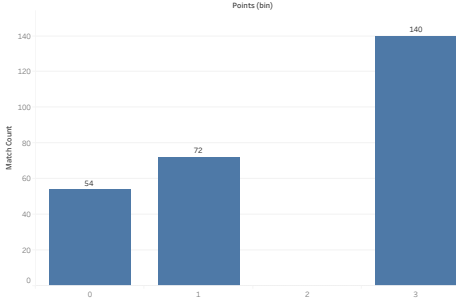


Description of the Issue/Problem

Match results are usually judged by the final score, but the score alone does not always show how well a team actually played. In soccer analytics, expected goals, or xG, is useful because it focuses on chance quality instead of only the final outcome (Mead et al., 2023). That matters for a team like Manchester United, where a win or loss can sometimes hide whether the team created the better chances in the match. The problem is that basic match stats do not always make it clear which parts of performance are most connected to league points. This project looks at that gap by testing whether xG-based match metrics explain Manchester United’s Premier League points better than simpler attacking numbers.

×

Distribution of Match Points



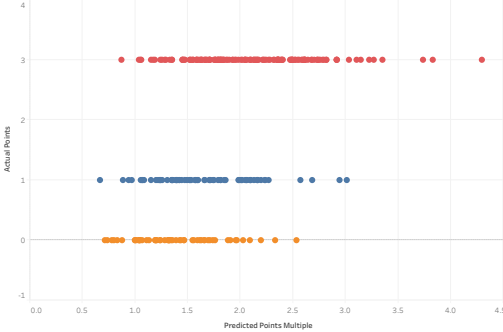
Description and Interpretation of Results

The descriptive results already showed a clear pattern. Manchester United averaged 2.03 points at home compared with 1.67 away, and matches they won had a much stronger average xG difference than matches they lost. The correlation between xG difference and points was 0.460, which suggested a moderate positive relationship before running the regression.

In the simple regression, the intercept was 1.60 and the coefficient for xG difference was 0.509, with  $p < .001$ . This means that a 1.0 increase in xG difference was associated with about 0.51 more points in a match on average. The model explained 21.2% of the variation in points, so it did not explain everything, but it did explain a meaningful part of the match-to-match results. Because the xG difference coefficient was statistically significant, I reject the null hypothesis for the simple regression.

In the multiple regression, xG for was positive and significant ( $b = 0.568, p < .001$ ), and xG against was negative and signifc...

Actual vs Predicted



Description and Exploration of the Data

The dataset comes from public Understat Premier League shot-level data, accessed through a public GitHub mirror of the files (Douglas BC, n.d.; Understat, n.d.). I cleaned and aggregated the shot events in R so that each row represents one Manchester United Premier League match. The final dataset includes 266 complete league matches from 2014/15 through 2020/21. I excluded the 2021/22 mirror file because it was incomplete, so keeping it would have weakened the analysis. The main variables used in the analysis are match date, opponent, home or away status, goals for, goals against, points earned, xG for, xG against, xG difference, shots, and shots on target. On average, Manchester United earned 1.85 points per match, created 1.58 xG, allowed 1.09 xG, and had an average xG difference of 0.49.

Specific Analytics

I used the following analysis steps in order:

1. I imported the raw shot-level data in R and cleaned the variables.
2. I combined the shot events into one row per Manchester United match.
3. I created descriptive statistics for points, goals, xG, shots, and home versus away matches.
4. I calculated correlations to see which variables moved most closely with points.
5. I ran a simple linear regression with points as the dependent variable and xG difference as the predictor.
6. I ran a multiple regression with xG for, xG against, shots on target, and home status as predictors.
7. I checked diagnostics, including residual plots, normality, homoscedasticity, independence of errors, and multicollinearity.
8. I used Tableau to build the final brochure-style dashboard and visualize the main findings.

For the simple regression, the hypotheses were:

H0: There is no linear relationship between xG difference and Manchester United’s points per match. In the model, this means the slope for xG difference equals 0.

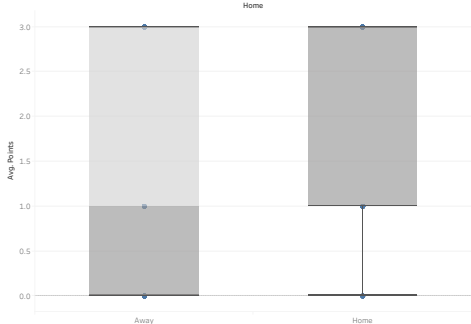
H1: There is a linear relationship between xG difference and Manchester United’s points per match. In the model, this means the slope for xG difference does not equal 0.

For the multiple regression, the hypotheses were:

H0: xG for, xG against, shots on target, and home status do not explain points once they are included together. In the model, all slope coefficients equal 0.

...

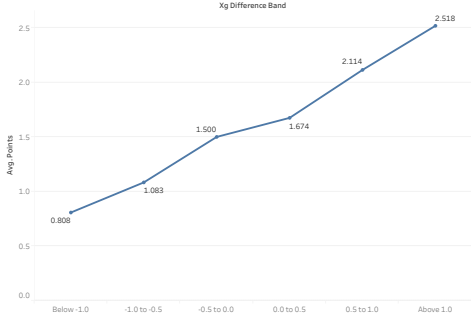
Home vs Away



Implications for Operational Practices

The biggest takeaway from this project is that chance quality mattered more than basic shot volume in explaining Manchester United’s league results. One reason is that not every shot is equally dangerous, so xG gives more information than just counting attempts. From an operational point of view, the results suggest that Manchester United should focus on improving the gap between xG created and xG allowed. In simple terms, that means creating better chances, not just more shots, while also limiting the quality of chances given up to opponents. For coaches and analysts, this makes xG difference a useful match-level performance measure to track.

Average Points by xG Band



References

Douglas BC. (n.d.). "scraping-understat-dataset" [Data repository]. GitHub. <https://github.com/douglasbc/scraping-understat-dataset>

Mead, J., O'Hare, A., & McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. \*PLOS ONE, 18\*(4), e0282295. <https://doi.org/10.1371/journal.pone.0282295>

Understat. (n.d.). "Premier League match and shot data". <https://understat.com/league/EPL>