# Analysis of Manchester City Games

Gaurab Baral and Joshua Osakwe
Department of Mathematics and Statistics, Northern Kentucky University

## Abstract

This project, which seeks to analyze soccer games of Manchester City from 2018 to May 2023, has 2 aims. This research project examines different soccer games of Manchester City and seeks to answer the question if there is a relation between the formation played by Manchester City and the outcome of the game. The study also addresses if player statistics such as shots, shots on target, possession percentage, fouls, corners, and pass completion rate relates with the total goals scored by Manchester City. Utilizing a sample of 200 randomly selected games, the analysis employs statistical tests and linear modeling techniques to answer these questions. Results showed that Manchester City generally win regardless of any formation they play in and "shots on target" is the most related factor with the total goals scored for Manchester City. We can predict total goals scored by Manchester City only by using Shots on Target, Yellow Cards, and Corners.

## Introduction

This research paper explores Manchester City's soccer games from 2018 to May 2023, analyzing factors like playing formations and player statistics to understand their impact on match outcomes. Extracted from Google's reputable soccer database, our goal is to identify the significant elements affecting Manchester City's performance and answer two key questions: the influence of playing formations on win/loss outcomes and how player statistics contribute to total goals scored.

## Variables

- **Outcome:** Categorical
- **Formation:** Categorical
- **Shots:** Quantitative
- **Shots on Target:** Quantitative
- **Possession:** Quantitative
- **Passes:** Quantitative
- **Fouls:** Quantitative
- **Yellow Cards:** Quantitative
- **Corners:** Quantitative

## Data Collection

- Initially, we gathered data from the Google Database of Manchester City. This data had 331 rows and 15 columns.
- We sampled 200 random rows from our initial population dataset.

### Data Pre-Processing

- Out of the initial 200 data rows, we chose 182 rows by excluding games where the formation used was unique to the sample
- We modified our data, categorizing "win" as "win", "lost" and "draw" as "no win".

## Research Questions

- To address the first question, we utilized a Test for Independence, assessing the relationship between game outcomes and playing formations.
- For the second question, we conducted a stepwise analysis, correlating explanatory variables and iteratively building linear models while tracking adjusted R^2 values. Additionally, we explored the relationship between "Shots on Target," "Outcome," and total goals through regression, eliminating non-significant variables and predicting scores in various scenarios.

## Results:

Table1: Outcome based on Formation

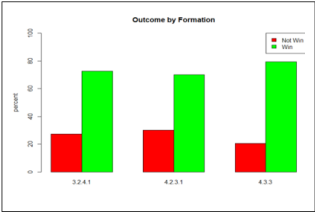| Formation | Not Win | Win | Total |
|---|---|---|---|
| 3.2.4.1 | 3(27.3%) | 8(72.7%) | 100% |
| 4.2.3.1 | 9(30%) | 21(70%) | 100% |
| 4.3.3 | 29(20.6%) | 112(79.4%) | 100% |


Figure1: Outcome based on Formation.
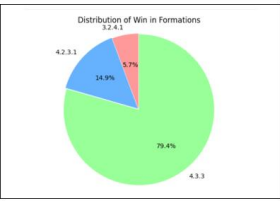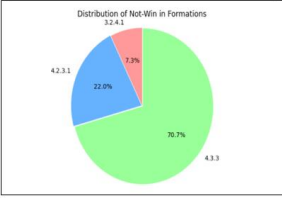

Figure 2: Wins distributed to Formations.


Figure 3: Not-Wins distributed to Formations.

Looking at the numerical and graphical summaries *(Table 1 and Figure 1,2 and 3)*, we can say that the likelihood of Manchester City winning in their games is greater than the likelihood of not winning. We found that the chi-squared tests yielded p-values of 0.5277 indicating that there is no significant relationship between the outcome of the game alone and the formation played by Manchester City.

At the 95% significance level, the proportion of games that Manchester City wins is estimated to be around 70.58% to 83.20% and the proportion of games that Manchester City did not win is estimated to be around 16.82% to 29.42%.

*Table2: Correlation Coefficient of Goal Scored with different factors*

| Goal Scored vs Shots | Goal Scored vs Shot on Target | Goal Scored vs Possession | Goal Scored vs Passes | Goal Scored vs Fouls | Goal Scored vs Yellow Cards | Goal Scored vs Corners |
|---|---|---|---|---|---|---|
| 0.3331 | 0.6109 | 0.0964 | 0.1645 | -0.0557 | -0.1798 | -0.1224 |


Figure 4: Scatter Plots of Total Goal Scored with explanatory variables
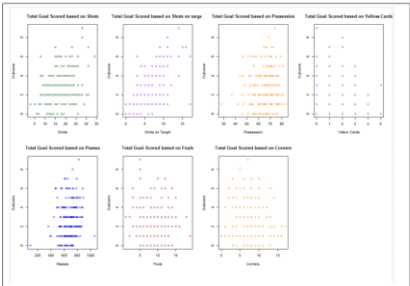
Looking at the graphical and numerical summaries (Table 2 and Figure 4) of:

- Total goals vs shots: relation is linear and positive. (Correlation: 0.3331)
- Total goals vs shot on target: relation is linear and positive. (Correlation: 0.6109)
- Total goals vs possession: relation is linear and neither positive nor negative. (Correlation: 0.0964)
- Total goals vs Passes: relation is linear and neither positive nor negative (Correlation: 0.1645)
- Total goals vs Fouls: relation is linear and neither positive nor negative (Correlation: -0.0557)
- Total goals vs yellow cards: relation is linear and neither positive nor negative. (Correlation: -0.1798)
- Total goals vs Corners: relation is linear and neither positive nor negative (Correlation: -0.1224)

The linear model with all 7 variables (Shots, Shots on Target, Possession, Passes, Fouls, Yellow: Cards and Corners) of Manchester City had a p-value of 2.2e-16. This signifies that there is a relation between the explanatory variables. The final model after backward elimination can be represented by the equation below:

$$\hat{y} = 1.58224 + 0.39679(Shots\ on\ Target) - 0.21107(Yellow\ Cards) - 0.16494(Corners)$$
where $\hat{y}$ is the total goal scored by Manchester city.

For the final model, 46.99% of the variation in total goal scored by Manchester City can be explained by Shots on Target, Yellow Cards and Corners received by Manchester City.

| Scenario | Shots on Target | Yellow Cards | Corners | Total Goal Scored ($\hat{y}$) | Lower Prediction Interval | Higher Prediction Interval |
|---|---|---|---|---|---|---|
| 1 | 5 | 1 | 2 | 3.025469 | 0.2805848 | 5.770354 |

In a game where Manchester City had 5 shots on target, 1 yellow card and 2 corner kicks, it can be predicted that the total goal scored will be between 0 and 6. (95% confidence)

## Discussions/Conclusions:

- Strategic Adjustments
- Key Performance Indicators
- Continuous Monitoring

## Limitations:

- Data Quality
- Additional Variables
- Machine Learning Approaches
- Qualitative Analysis
- External Validity

## Acknowledgement: