Technische Universität Dortmund

Department of Statistics

Application Report Master Data Science

Wintersemester 2024/2025

Student

Stasevski Emil

Contents

1 Introduction

Research questions

1) **Does the winner of a match have a higher passing rate than the loser?** 2) **Is the difference in the passing rate in games with a winner higher than the difference in games that ended in a draw?**

The predictor "passing_quote" (passing rate) may reflect various aspects of a team's dynamics, including player cooperation, the distribution of responsibilities (theoretically, teams where a few players handle most of the workload can still be successful), and the team's ability to maintain control of the ball. Therefore, it is important to consider that a low "passing_quote" does not necessarily lead to defeat in a game.

2 Data integrity & quality

Before drawing any significant statistical conclusions, researchers must evaluate the quality of the data. The dataset includes information on 153 games, with variables comprising passing rates (the ratio of passes made by a team to passes received by a player from the same team) and the outcome for each team (victory or defeat). Fortunately, there are only two missing values associated with game number 139: the passing rates for both the losing and winning teams. Thus, the dataset contains a total of 306 observations, with two values absent. In quantitative research, while it is crucial to acquire as much data as possible, the available dataset should be considered adequate for conducting statistical analysis to address the research questions posed (according to Cohen (Cohen, 2013)).

Transitioning from a general discussion of data, I now intend to delve into the specific variables under consideration. The variable *winner* should be classified as a categorical variable. A primary concern in this context is the balance of the classes, particularly because *winner* serves as the target variable. The dataset comprises 190 defeats and 114 victories , as it could be seen on Figure 1. From a statistical standpoint, unbalanced data can adversely affect statistical inference. Nevertheless, this proportion appears reasonable for the given data, as it may reflect a natural selection process wherein only the strongest teams succeed (Carpita et al., 2019; Gomez et al., 2016).

Numerical variable *"passing_quote"* requires deeper study. First, lets study it generally, then down by winning status. In general, this variable distribution is right-skewed (Figure 2) - players tend to

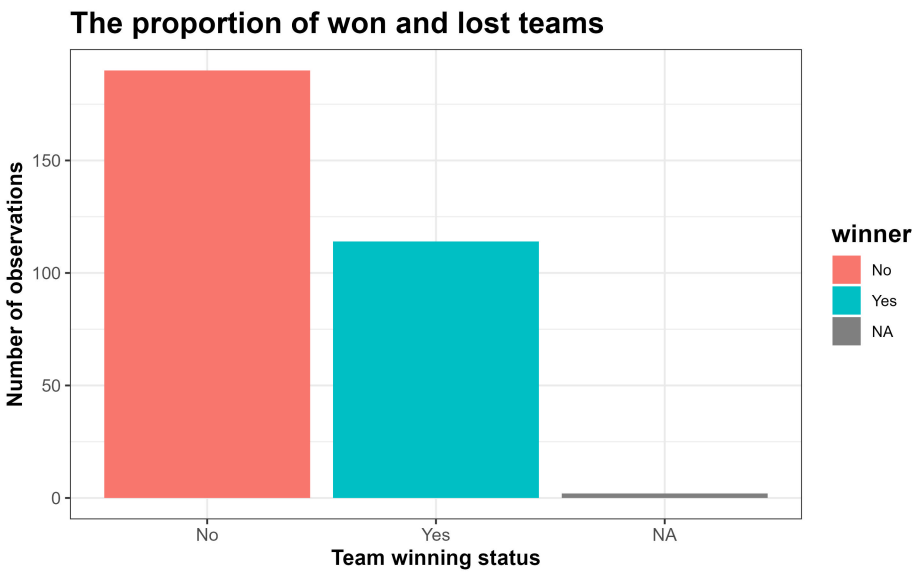

Figure 1:

receive more than 60 % of all the passes played by their team in 98% of cases. Whereas, the minimal ratio of passes is equal to 53 % (these are outliers - Figure 3. So we could see that in all the games at least 53% of the passes were delivered from one member of the team to another. The right-skewed distribution indicates the lack of normal distribution, so that later we should use non-parametric methods, so here we apply *Shapiro* and *Levene tests* in order to check that properly (Schultz, 1985; Shapiro & Wilk, 1965).

Both the Shapiro-Wilk test and Levene's test yield p-values significantly less than 0.05, it indicated that the assumptions of normality and homogeneity of variances are violated. So further we apply *non-parametric methods*.

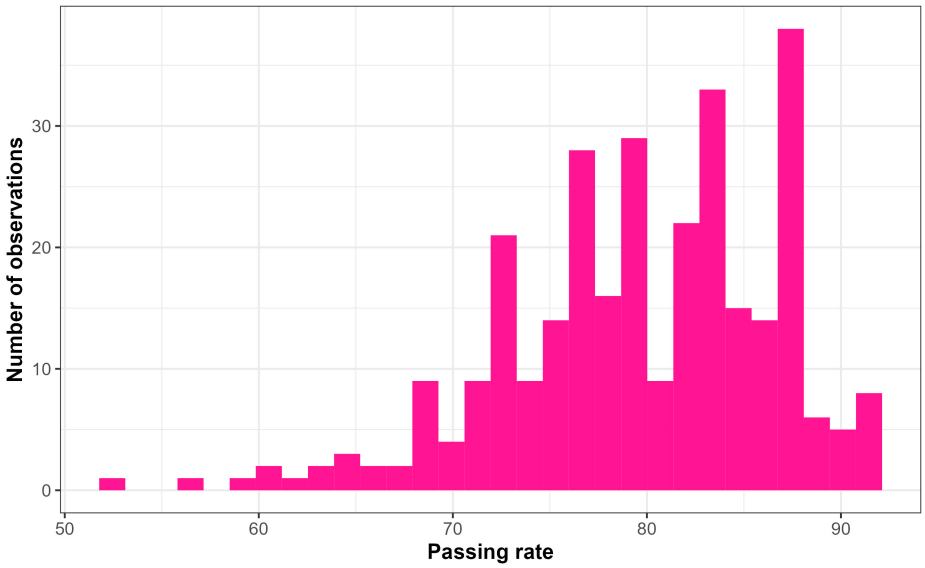

Figure 2: The general distribution of "passing_quote" teams passing rate (historgram)

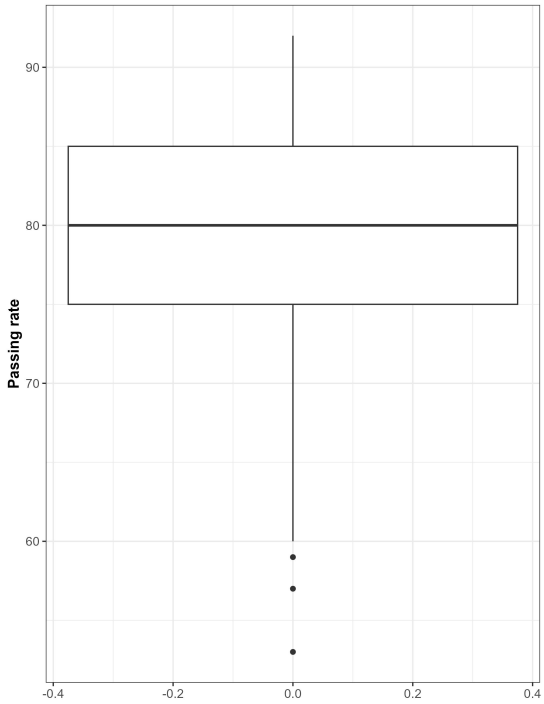

Figure 3: The general distribution of "passing_quote" teams passing rate (boxplot with outliers)

Passing rate and winning

As we can see on Figures 4 , 5 and Table 1 0.75 quartile, median and mean values for won teams are higher than for the lost teams. Yet, these statistics for these groups. Moreover, minimal values (outliers) of passing rate for winning teams are lower than for loosing teams. As I underlined before, there could successful and victorious teams with few strong players who do all the job.

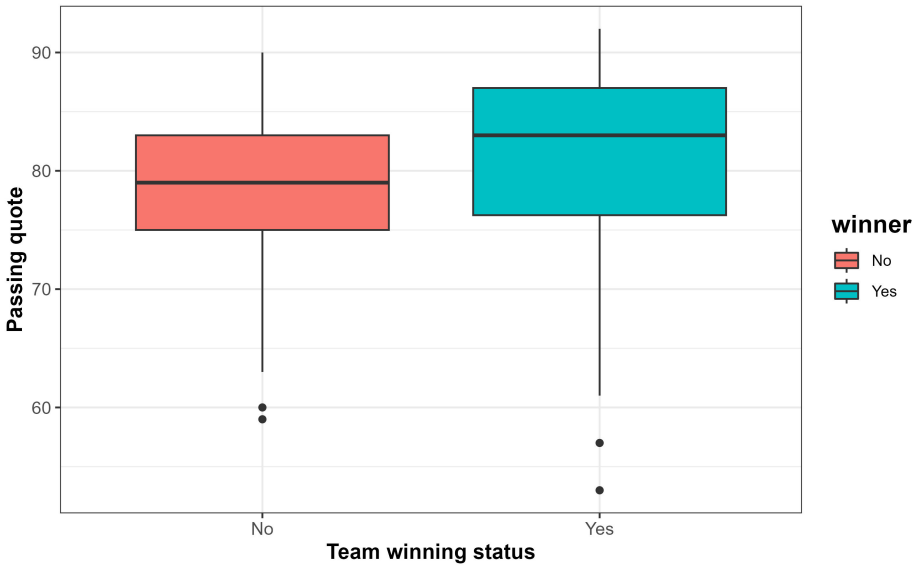

Figure 4: The distribution of "passing_quote" teams passing rate down by winning status

So, apart from graphics we have non-parametric methods - *Mann-Whitney U Test (Wilcoxon*

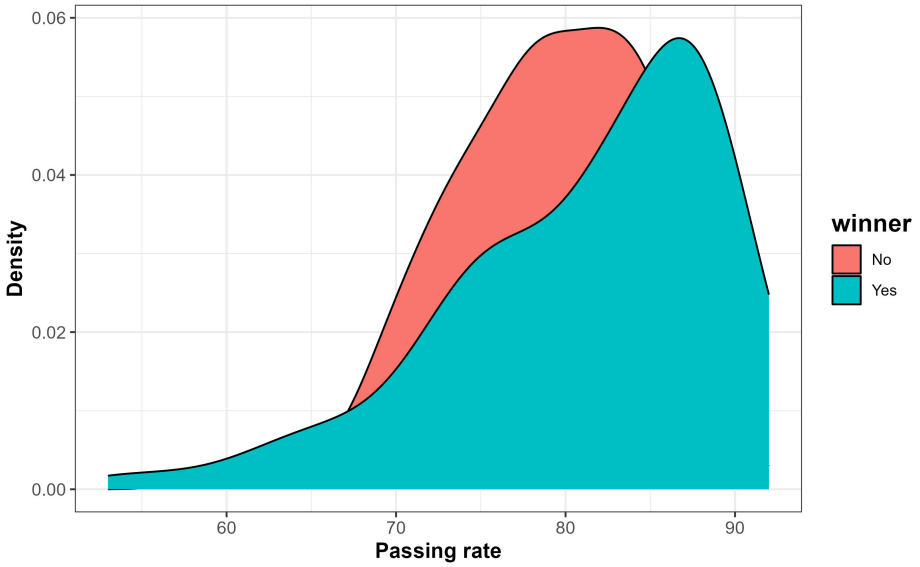

Figure 5: The distribution of "passing_quote" teams passing rate down by winning status

Table 1: Summary Statistics by Winning Status

	winner	statistic	value
1	No	mean_passing_quote	78.84
2	No	sd_passing_quote	6.07
3	No	min_passing_quote	59
4	No	max_passing_quote	90
5	No	n	190
6	Yes	mean_passing_quote	81.08
7	Yes	sd_passing_quote	8.06
8	Yes	min_passing_quote	53
9	Yes	max_passing_quote	92
10	Yes	n	114

Rank-Sum Test) and *Kruskal-Wallis Test.* The first compares the distribution of passing rates between the two groups (win/loss) without assuming normality (t-test alternative), whereas the second compares the distributions across more than two groups, also without assuming normality or equal variances (ANOVA alternative) (Kruskal, 1952; Wilcox, 2012).

For both tests, we obtained p-values significantly less than 1%, indicating a statistically significant difference in passing rates between winning and losing teams at the 1% significance level. Specifically, we reject the null hypothesis H_0, concluding that the passing rate distributions for these groups are different. Additionally, I performed parametric tests, including the t-test and ANOVA, despite the data not being normally distributed. These tests, which generally have low statistical power under such conditions, yielded consistent results.

3 Passing rate and tied score in the game

Applying basic data engineering, I select rows where teams lost and further filter duplicating "game_id" - games with raw results. Here again we have disbalance of classes, there only 76 draws, whereas cases one of the teams won are four times more - 230 (Figure 6).

In terms of passing rate 0.75 quartile for draw and no-draw games coincides. Median and mean values are slightly higher for cases one team won compared to draw games. Yet again lowest minimal passing rate values (outliers) belong to the games where one teams won - Figures 6, 7 and table 2. So, both draw and no-draw game supposes a fierce competition, however, because of specific teams with low passing rate yet with capability to win, the passing rate itself is not the most appropriate measure to predict game's outcomes. The Kruskal-Wallis and Mann-Whitney U tests' results confirm this assumption. The p-value is quite close to 5% significance threshold. We cannot assert with 100% that there is statistical significance in passing rate between these two types of games (also see Figure 8).

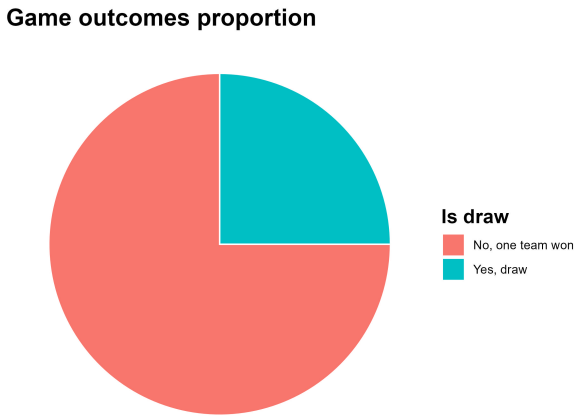

Figure 6: The proportion of draw and won games

Table 2: Summary Statistics by Game outcome

	is_draw	statistic	value
1	No, one team won	mean_passing_quote	80.17
2	No, one team won	sd_passing_quote	6.81
3	No, one team won	min_passing_quote	53
4	No, one team won	max_passing_quote	92
5	No, one team won	n	228
6	Yes, draw	mean_passing_quote	78.21
7	Yes, draw	sd_passing_quote	7.25
8	Yes, draw	min_passing_quote	59
9	Yes, draw	max_passing_quote	89
10	Yes, draw	n	76

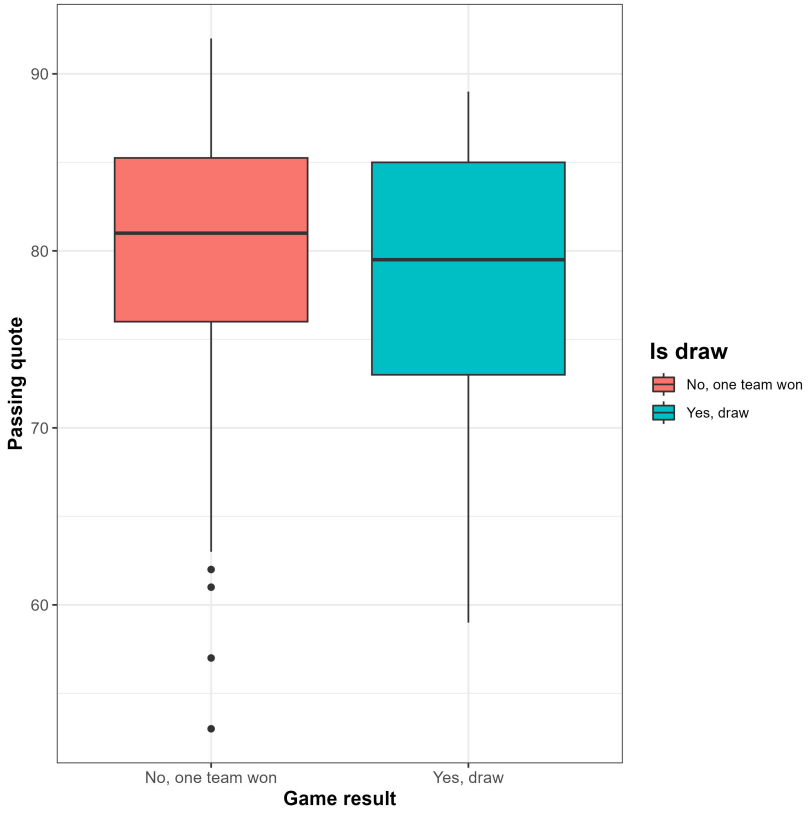

Figure 7: The game outcome and passing rate

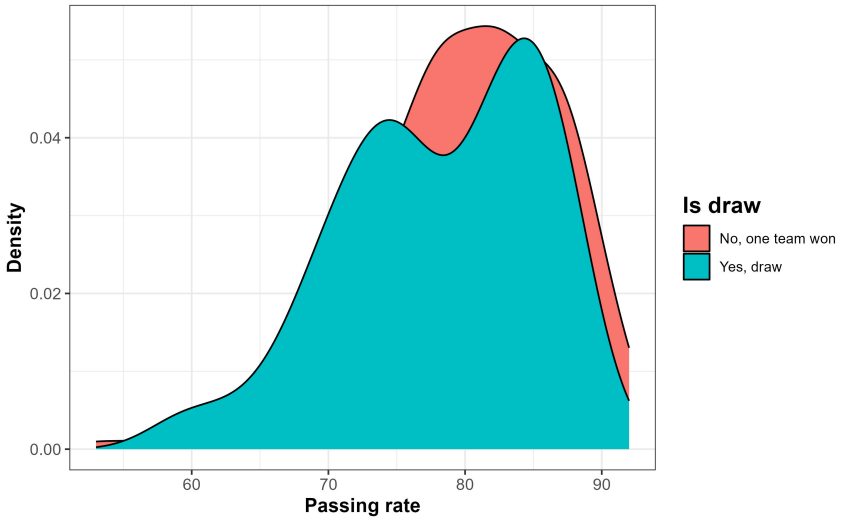

Figure 8: The game outcome and passing rate

4 Conclusions

With a few exceptions, winning teams generally exhibit higher passing rates, although the differences in means and medians are not substantial. This suggests that teamwork and cooperation among players are important factors in achieving victory. However, there are some teams, such as those in games with IDs 26 and 152, where the passing rate is low. This anomaly may be attributed to the presence of strong individual players who carried the team to victory despite a lack of cooperation. This observation

highlights a potential issue of endogeneity due to the absence of other explanatory variables that could account for the reasons behind a team's victory. Consequently, passing rate alone may not be the most reliable predictor of game outcomes. Moreover, the difference in passing rates between winning and drawing games is almost statistically insignificant, indicating that further investigation and additional data are required to better understand the factors that contribute to game outcomes.

References

Carpita, M., Ciavolino, E., & Pasca, P. (2019). Exploring and modelling team performances of the kaggle european soccer database. *Statistical Modelling*, *19*(1), 74–101.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.

Gomez, M.-A., Lago-Peñas, C., & Owen, L. A. (2016). The influence of substitutions on elite soccer teams' performance. *International Journal of Performance Analysis in Sport*, *16*(2), 553–568.

Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics*, 525–540.

Schultz, B. B. (1985). Levene's test for relative variation. *Systematic Zoology*, *34*(4), 449–456.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Academic press.

5 Applications

All code utilized in the research (data preparation, descriptive analysis and stochastic actor oriented model) is accessible via a provided hyperlink.

The Shapiro-Wilk test statistic W

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Where:

- $x_{(i)}$ is the i-th order statistic (i.e., the i-th smallest number in the sample).
- $\bar{x}$ is the sample mean.

- a_i are constants generated from the means, variances, and covariances of the order statistics of a sample of size n from a standard normal distribution.

The Kruskal-Wallis test statistic H i:

$$H = \left(\frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \bar{R} \right)^2 \right)$$

Where:

- N is the total number of observations across all groups.
- k is the number of groups.
- n_i is the number of observations in group i.
- $\bar{R}_i$ is the average rank of observations in group i.
- $\bar{R}$ is the overall average rank of all observations.

The Mann-Whitney U test statistic U :

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

Where:

- n_1 and n_2 are the sample sizes of the two groups.
- R_1 is the sum of the ranks for the first group.
- R_2 is the sum of the ranks for the second group.