\section{Association Between Variables}

This final step within the methodology aims to examine the relationship between tweet sentiment and football outcomes. To achieve this, two categorical variables were analyzed using statistical methods, including the Chi-Squared Test and Cramér's V.

Within the framework of our research, a deliberate choice was made to utilize the Chi-Squared Test and Cramér's V instead of correlation analysis methods. The reasoning for this decision is based on the categorical nature of the data under investigation. It is well-established that correlation analysis, such as Pearson or Spearman correlation, is more applicable for continuous or ordinal data. Given that our dataset comprises counts of positive and negative tweets, as well as categorical outcomes, the Chi-Squared Test and Cramér's V are deemed to be more appropriate techniques for examining the associations between these variables.

The Chi-Squared Test is a non-parametric statistical method used to determine if there is a significant relationship between two categorical variables. A contingency table displays the frequency distribution of categorical variables in a matrix format, showing the observed frequencies of different combinations of categories of two or more variables. The Chi-Squared Test employs a mathematical formula to calculate the deviation between the observed and expected data, resulting in a Chi-Squared test statistic that quantifies the discrepancy between these frequencies.

Apart from the Chi-Squared test statistic, the Degrees of Freedom (DF) play a critical role in determining the distribution of the test statistic and the significance of the test. The Degrees of Freedom are calculated using the formula: \begin{equation}

df = (r-1)(c-1)

\end{equation}

Where:

\begin{itemize}

\item \(r\): is the number of rows in the contingency table.

\item \(c\): is the columns in the contingency table.

\end{itemize}

The significance of the test is determined by comparing the p-value to a predetermined level of significance, commonly set at 0.05. While the Chi-Squared Test allows us to determine the existence of an association between the variables, it does not provide information about the strength or direction of the relationship. To assess the strength of the association, Cramers V will be used, which is a measure derived from the Chi-Squared statistic.

Cramér's V is a statistic measure that ranges from 0 to 1, used to determine the strength of the relationship between two variables. Cramér's V can be interpreted to determine the strength and direction of the relationship between the variables. It is derived from the Chi-Squared statistic and calculated using the formula:\begin{equation} \label{eq:kramersv}

V=\sqrt{\frac{\chi^2}{n\min(r-1,c-1)}}

\end{equation}

Where:

\begin{itemize}

\item V: is the value of Cramér's V

\item \(\chi^2\): is the result of the Chi-Squared Test

\item \(n\): is the total number of observations

\item \(r\): is the number of rows in the contingency table

\item \(c\): is the number of columns in the contingency table

\end{itemize}.

To conduct the analysis, two separate contingency tables were created by combining the sentiment counts for both weekly match outcomes and monthly PPG, using the pandas library in Python in which the expected frequencies were calculated assuming independence between variables. The chi2-contingency \footnote{\url{https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2\_contingency.html}} function from the scipy.stats library was used to find degrees of freedom and expected frequencies. The Chi-Squared Test statistic (X²) and its corresponding p-value were calculated, with a manual calculation performed to validate the accuracy of the code.

In order to classify the Points-Per-Game contingency table, a quartile-based stratification method was utilized, wherein the positions of the three quartiles were ascertained using the QUARTILE \footnote{\url{https://support.microsoft.com/en-us/office/quartile-function-93cf8f62-60cd-4fdb-8a92-8451041e1a2a}} function in Microsoft Excel. This process was conducted separately for each month under consideration, as combining the values from the two distinct periods might not yield an appropriate representation, given the potential differences between them. By employing this method, the teams were systematically divided into three distinct categories for each month, ensuring an equitable distribution for inclusion within the contingency table.

\begin{table}[H] %

\centering

\begin{tabular}{lccc}

\toprule

& Match Won & Match Drawn & Match Lost \\

\midrule

Positive Tweets & 45,000 & 30,000 & 25,000 \\

Negative Tweets & 15,000 & 20,000 & 40,000 \\

\bottomrule

\end{tabular}

\caption{A sample contingency table of tweet sentiment and match outcomes}

\label{table:contingency\_table}

\end{table}

\begin{table}[H] %

\centering

\begin{tabular}{lccc}

\toprule

& Low PPG & Medium PPG & High PPG \\

\midrule

Positive Tweets & 139,000 & 21,000 & 45,000 \\

Negative Tweets & 24,000 & 73,000 & 64,000 \\

\bottomrule

\end{tabular}

\caption{A sample contingency table of tweet sentiment and PPG}

\label{table:contingency\_table}

\end{table}

After determining the significance of the association using the Chi-Squared Test, Cramér's V was computed to measure the strength of the association. This allowed us to discern the degree to which tweet sentiment is correlated with match outcomes.