To Do’s:

1. Data description and cleaning
2. Interpretation and results
   1. Q1: EDA of binary variables (player, team, etc.) of all tweets
   2. Q2: What are the most common topics discussed by Supercar fans on Twitter during live racing events, and how do these topics relate to online user engagement and live television viewership? (topic)
   3. Q3: How does the sentiment of live tweets during Supercar racing events impact online user engagement and live television viewership? (sentiment)
   4. Q4: Can we predict online user engagement and live television viewership for Supercar racing events based on the content and sentiment of live tweets? (regression) -
   5. Q5: What is the relationship between the time of day when tweets are posted and the online user engagement and live television viewership for Supercar racing events? (regression)
3. Overall recommendations
4. Limitations

Question 3:

we had performed data aggregation and correlation analysis on two data sources: tweet\_sport and Tv\_rating. We aggregates tweet data by minute and calculates the mean sentiment and engagement, then calculates the correlation between sentiment and engagement, and sentiment and TV ratings, which shows us the relation between the tweets and tv rating.

Important steps followed:

* The sentiment of each tweet is calculated using the sentiment analysis tool.
* The polarity\_scores function is used to calculate the sentiment score for each tweet.
* The time column in the tweet data is converted to a pandas datetime format and set as the index for the data frame.
* The sentiment column is converted to a numeric type so that it can be used for calculations later.
* The tweet data is then resampled by minute and aggregated to calculate the mean sentiment and engagement (sum of retweets and likes) for each minute.
* The TV rating data is similarly formatted by converting the start and end times to datetime format and setting them as the index for the data frame.
* The tweet data and TV rating data are merged together using an outer join, so that all data is included even if there are missing values.
* Any rows with missing values are dropped from the merged data frame.
* The correlation between sentiment and engagement, as well as between sentiment and TV ratings, is calculated using the corr() function.

The output shows the correlation coefficients between the sentiment of tweets and two other variables: engagement (measured as the sum of retweets and likes) and TV ratings.

* The correlation coefficient between sentiment and engagement is -0.39, which indicates a moderate negative correlation. This means that as the sentiment of tweets becomes more negative, the level of engagement (retweets and likes) tends to decrease.
* The correlation coefficient between sentiment and TV ratings is 0.01, which indicates a very weak positive correlation. This means that there is little to no relationship between the sentiment of tweets and TV ratings.

**Limitation :**

This analysis has some limitations that should be considered when interpreting the results.  
**Firstly**, the sentiment analysis algorithm used in this analysis, the VADER algorithm, may not be entirely accurate in assessing the sentiment of tweets. The VADER algorithm is based on a lexicon of words that have been manually labelled with their sentiment score, which may not always capture the nuances and context of language used in tweets. Additionally, the sentiment score assigned to a tweet is based on a set of predetermined rules and heuristics, which may not always be applicable to different types of tweets or topics.  
**Secondly**, the sample of tweets used in this analysis may not be representative of the entire population of tweets related to sports. The tweets were obtained from a single source, and only include tweets that contain certain keywords related to sports. This may lead to a biased sample that does not accurately reflect the sentiment or engagement of the broader population of tweets related to sports.  
**Thirdly**, the analysis assumes that there is a causal relationship between sentiment and engagement or TV ratings. While there may be a correlation between these variables, it is not necessarily the case that changes in sentiment directly cause changes in engagement or TV ratings. Other factors, such as the timing of events or external news events, may also influence engagement or TV ratings and need to be accounted for in the analysis.

The analysis in question 3 only considers correlations between variables and does not establish any causal relationships.