**Project Title: BBC Sports Article Explorer**

**Project Objective:**

To analyze and visualize sports news coverage in the BBC dataset using R. The project will focus on understanding article distribution, identifying text patterns, and building an interactive Shiny dashboard for category-based exploration and keyword insights.

**Project Phases:**

**1. Data Import & Inspection**

* Read the CSV with read\_csv()
* Check structure, missing values, and duplicates

**2. Data Cleaning & Preprocessing**

* Clean text: lowercase, remove stopwords, punctuation, and numbers
* Tokenize and stem if needed
* Handle NAs and blank content

**3. Exploratory Data Analysis (EDA)**

* Count of articles per category
* Distribution of article lengths (word/character count)
* Top keywords per category
* Temporal trends if date exists

**4. Text Analysis**

* Most frequent words across all content
* Word clouds by category
* Term Frequency – Inverse Document Frequency (TF-IDF) ranking
* Sentiment analysis using syuzhet or bing lexicon

**5. Shiny Dashboard Development**

**Tabs & Features:**

* **Overview:** Summary stats, total articles, unique categories
* **Category Explorer:** Bar charts, pie charts by category
* **Word Explorer:** Word cloud + TF-IDF by sport
* **Table View:** Data table with search/filtering using DT
* **Optional:** Sentiment tracker or article similarity visual (e.g., PCA plot)

**6. Dashboard Deployment**

* Deploy to **shinyapps.io** or host locally
* Include interactivity (filter by category, search title, view word cloud)

**Tools & Technologies:**

* **Language:** R
* **Packages:**
  + Data wrangling: dplyr, tidyr, readr
  + Text processing: stringr, tidytext, tm, textclean
  + Visualization: ggplot2, wordcloud, plotly, DT
  + Dashboarding: shiny, shinydashboard, shinyWidgets