COMP1016 Mini Project Part A

Information Semester 2 22/23

# Part A (10% of COMP1016 Total Score)

**Project aim:** Students apply mathematical methods and programming skills to solve business problems. Emphasis will be placed on the applications of the mathematical concepts in probability and statistics, visualization skills, interpretation of results and insights from data.

**Group Forming:**  2 to 3 students shall form a group to deliver a project using Python programming. **Group enrollment is due by Feb 3rd 2023.**

# Submission: The work should be prepared in pynb format or py format. Please explain your approach and discuss your findings in the markdown section. Submit the zipped file to the site <http://buelearning.hkbu.edu.hk/> no later than Mar 4th 2023.

# Dataset:

# Results.csv

The dataset includes 44,341 results of international football matches starting from the very first official match in 1872 up to 2022. The matches range from FIFA World Cup to FIFI Wild Cup to regular friendly matches.

results.csv includes the following columns:

1. date - date of the match
2. home\_team - the name of the home team
3. away\_team - the name of the away team
4. home\_score - full-time home team score including extra time, not including penalty-shootouts
5. away\_score - full-time away team score including extra time, not including penalty-shootouts
6. tournament - the name of the tournament
7. city - the name of the city/town/administrative unit where the match was played
8. country - the name of the country where the match was played neutral

# Instruction:

1. Each group picks 5 international football teams as the subject in their analysis. **Extract the data from the beginning of 2011 to the end of 2020.**

Iran, Hong Kong, United States

2. Students are required to answer 5 questions in their report (i.e. Pynb or Py format). The first three questions are given as follows:

Question 1: Winning Percentage by team by year

1. Calculate the trend of winning percentage of each selected team by year, from 2011 to 2020. Summarize the result by team in one table. (Table 1a)
   1. make number of winning function and number of matching function -> calculating the number of wins -> calculating the number of matches -> divide
2. Summarize the result by team by year in one table. (Table 1b)
3. Visualize the result of Table 1b in one graph.
4. Write a short paragraph to summarize the result.

Question 2: Correlation between winning percentage and number of goals scored by the selected team.

1. Calculate the average number of goals per match scored by the selected team by year, from 2011 to 2020. Summarize the average number of goals per match by year by team in one table. (Table 2a)
2. Calculate the correlation result between table 1b and 2a. Plot a scatter diagram between the two variables for each selected team.
3. Discuss your findings in Question 2a and 2b.
4. An analyst wanted to discuss the importance of scoring a goal in winning a football match and he came up with the correlation approach (2b). He found the correlation results varied from team and team. Describe two possible reasons for the variations found.

Question 3: Investigate if winning trends exist in football matches. Using the ten years data of **the five selected teams combined\***, calculate the average winning probability in each of the following conditions:

1. The winning percentage, given that one team won the previous two matches.
2. The winning percentage, given that one team won one of the previous two matches.
3. The winning percentage, given that one team lost the previous two matches.
4. Write a short paragraph to summarize the findings in question 3. Discuss if winning trends exist in football matches. Feel free to change the approach or add new metrics to help.

\***Provide the result by each team, and the overall result of the five selected teams combined**.

3. For question 4 – 5, students propose and pursue their own research direction using the same dataset (10 years of 5 selected teams). The two questions should cover the topics in COMP1016, including:

1. Graphical representation of data (such as time series, scatter diagram, side-by-side chart)
2. Descriptive Variability (such as Boxplot, Correlation and Standard Deviation, Standard Score)
3. Probability and Conditional Probability, Probability Distribution

# Assessment:

1. **Presentation (20%)** 
   * Ability to describe the problems, formulate the approach, visualize the results and interpret the findings. Grading will be assessed by presentation skills demonstrated in Question 1 – 5.
2. **Technical Skills (50%) – Question 1 - 3**
   * Students should demonstrate the steps in solving the problems. Put markdown and remarks to assist others understand your approach. Source codes will be examined.
   * Grading will be assessed by the submitted source code, technical capability, visualization, the markdown, and discussion.
3. **Technical Skills (30%) – Question 4 - 5**
   * Students should demonstrate the steps in solving the problems logically. Put markdown and remarks to assist others understand your approach. Source codes will be examined.
   * Grading will be assessed by the submitted source code, technical capability, visualization, the markdown, and discussion.