COMP1016 Mini Project Part B

Information Semester 2 22/23

# Part B (15% 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.

# 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 15th April, 2023.

# Presentation arrangement: Students will present their findings in the lecture room (FSC901). The presentation date is 17th April, 2023 (Monday) during the lab time. Students may present the works using python notebook, PowerPoint, or any method deemed appropriate.

# Dataset:

# ITSalary.csv

The dataset contains survey results about the salary information conducted in 2019 and 2020 for the IT professionals in Europe.

ITSalary.csv  includes the following columns:

1. Age
2. Gender
3. City
4. Seniority Level
5. Position
6. Total years of experience
7. Yearly salary
8. Number of vacation days
9. Main language at work
10. Contract
11. Company type
12. Year – which year was the survey taken?

# Instruction:

1. Students are required to answer 4 questions in their report (i.e. Pynb or Py format). The first three questions are given as follows:

2. The data itself may not be well suited for plotting graphs because some data fields contain too many distinct values (i.e. over 30 different professionals). Category binning is needed before visualizing the data. Show the steps how you transform the data. For example, transform position data from over 30 different types into six main categories.

Question 1: Describe the big picture about Salary in the IT sector

1. Produce two visualization (tables or charts) to talk about the yearly salary by different positions, and the relationship between years of experience and yearly salary
2. Is there sufficient evidence to show that the average IT salary in Berlin is higher than that in other cities? Show your result using a boxplot.
3. Write a short paragraph to summarize the result.

Question 2: Verify the Central Limit Theorem (the technique learnt in Lab 5)

1. Do data cleaning and transformation if necessary. You can re-use the data cleaning part in Question 1.
2. Verify the central limit theorem using salary data
3. Verify the central limit theorem using age data
4. Discuss your findings

3. For question 3 – 4, students propose and pursue their own research direction using the dataset. The questions should be related to the statistical inference techniques in Lecture 5 to 10.

# 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 (40%) – Question 1 - 2**
   * 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 (40%) – Question 3 - 4**
   * 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.