DATA MINING HOMEWORK 1

Deadline: 03/04/2025 早上 9:00 (逾時系統關閉無法受理)，所用程式碼及 outputs 可直接存檔、連同存成 pdf 檔的答案一起放進資料夾，用壓縮檔於 EL 上繳交!

壓縮檔名請標示如以下範例：B1128100 丁小雨 HW1 (學號姓名 HW1)

Please use Python, necessary libraries and the attached "googleplaystore.csv" to complete the following analysis and answer each question. **All the Python codes and outputs, along with your written/typed answers must be submitted**.

1. List the top 20 apps with the largest size. Present the app names and their size. (10%)
2. Check whether each attribute has missingness. For those attributes that have missingness, present the attribute names and their number of missing values. (15%)
3. Let's focus on the attribute "**Rating**".

(1) Calculate its mean, IQR, and standard deviation. (10%)

(2) Identify and report anomalies and/or errors in it. What would you do to make necessary corrections for it? (15%)

(3) Following (2) after corrections being made, re-calculate the mean, IQR, and standard deviation. (15%)

4. Let's focus on the chi-square test.

(1) Check online. What are the assumptions and limitations of the chi-square test? (15%)

(2) Use the chi-square test to investigate the following: whether the ratings $\geqq 4$ or not is associated with whether the price$\geqq$100 or not. Report on your test results. What is your conclusion? (20%)