

Homework 2

Deadline: 2019/03/29 (Tuesday) 23:59

Problem 1: Kobe Shot Data Analysis

([hw2_kobe.ipynb](#))

Kobe Bryant marked his retirement from the NBA by scoring 60 points in his final game as a Los Angeles Laker on Wednesday, April 12, 2016. Drafted into the NBA at the age of 17, Kobe earned the sport's highest accolades throughout his long career. Using 20 years of data on Kobe's swishes and misses, many people like to analyze how his shots find the bottom of the net.

In this homework, based on the data we provided you "**kobe.csv**", you can try to be not only a Kobe shot data analyst, but also pandas data analyst. Please answer the following questions based on the dataset. **You are asked to use pandas package at this time.** If you do not use pandas for each question, your score will be zero. In addition, you are asked to write a function for each of the following question.

(1) 計算 Kobe 對戰火箭隊(HOU)的平均兩分球與三分球命中率。

Hint: opponent, shot_type, shot_made_flag

(2) 列出 Kobe 對戰過的球隊中，使得 Kobe 平均得分最低的前五支球隊，並顯示平均得分。

Hint: opponent, shot_made_flag, shot_type

(3) 列出 Kobe 在季後賽最後 3 分鐘內得分最高的前五場球賽，並顯示其得分。

Hint: opponent, playoffs, shot_made_flag, shot_type, minutes_remaining

(4) 列出 Kobe 各球季季後賽中，比賽最後 1 分鐘內的 Jump Shot 命中率，按球季先後排序。

Hint: action_type, playoffs, game_date, seconds_remaining, shot_made_flag, shot_type

(5) 計算 Kobe 「得分命中率 33% 以上的最長連續場數」前 3 名，列出場數以及起訖日期。

Hint: game_date, shot_type, shot_made_flag

(6) 計算 Kobe 「上半場得分多於下半場」且命中率最低的前 3 名場次，列出日期、對手、上半場得分差、該場得分，按命中率由上而下排序。

Hint: period, shot_made_flag, opponent, game_date

(7) 計算 Kobe 「投籃連續失手最多球」之前 3 名場次，列出日期、對手、連續失手球數、及該場得分，按失手次數由上而下排序。

Hint: game_id, shot_made_flag, opponent

題意補充說明

- 所謂“球季”，用 season 這個欄位即可。
- 季後賽“最後 3 分鐘”，只需考慮第 4 節。換言之，延長賽不需考慮，若 Kobe 在第四節沒上場，也不需納入計算。
- “列出球賽”指的是列出球賽(game_id)。
- “上半場得分多於下半場”，“下半場”只需考慮第 3 節與第 4 節，無需考慮延長賽。
- 關於“連續場數”的“連續”，只要將比賽日期按先後順序排序後，前後兩場即為“連續”。

Problem 2: In-Game Purchase Data Analysis

([hw2_kobe.ipynb](#))

In this homework, you are asked to deal with a task of analyzing an “in-game purchase” dataset. Please refer to the dataset “**purchase_data.csv**”. For in-game purchasing, players are able to purchase optional items that enhance their playing experience. Now your task is to generate a report that breaks down the game’s purchasing data into meaningful insights. We provide you basic observation about the dataset, as below. You need to follow the instructions in the ipynb code we provide you (“**hw2_game.ipynb**”), and complete each code block on your own.

- There are 1163 active players. The vast majority are male (84%). There also exists, a smaller, but notable proportion of female players (14%).
- Our peak age demographic falls between 20-24 (44.79%) with secondary groups falling between 15-19 (18.58%) and 25-29 (13.37%).
- The age group that spends the most money is the 20-24 with 1,114.06 dollars as total purchase value and an average purchase of 4.32. In contrast, the demographic group that has the highest average purchase is the 35-39 with 4.76 and a total purchase value of 147.67.

You are forced to use the **pandas** package (and its **data frame** techniques) to generate the data frame that is exactly the same as the table right after each code block of “**hw2_game.ipynb**”. For more details, please refer to “**hw2_game.ipynb**”.

Problem 3: DBSCAN Clustering Implementation

([hw2_dbscan.ipynb](#))

In this problem, your task is to use Python (along with numpy and pandas) to implement the very well-known density-based clustering algorithm, DBSCAN, based on two datasets “**blobs.csv**” and “**varied.csv**”. Remember we provide you a sample code “**dbscan-sample.ipynb**” that uses the **sklearn** package’s **dbscan()** function to perform density-based clustering using these two datasets. Now your task is to implement DBSCAN by yourself, i.e., without relying on any package containing DBSCAN algorithm. What you can use are only the basic packages, including numpy, pandas, math, and matplotlib, as well as the plotting tool we provide you: “**dbscan_lab_helper.py**”.

We know implementing DBSCAN by your own is not easy. Thus, in the ipynb, “**hw2_dbscan.ipynb**”, we are very kind to provide several functions for you to fill in, include **_dist()**, **_eps_neighborhood()**, **_region_query()**, **_expand_cluster()**, and **dbscan()**. The last one is the main function, which is where the algorithm starts executing. The expected results are two pictures shown in “**hw2_dbscan.ipynb**”. Such two pictures are plotted using “**dbscan_lab_helper.py**”, and used to validate the correctness of your code. To complete this task, **you can either follow the functions we provided to fill in each function code block, or write you own ipynb file from scratch** (i.e., not necessary to use the functions we provide, you can do it by yourself). If you choose the latter option, you still need to generate the two pictures as expected.

Important Notes

This is a homework for each **individual**. You are asked to write comments or markdown sections to describe the meaning of each part of your codes in “hw2_kobe.ipynb”, “hw2_game.ipynb”, and “hw2_dbscan.ipynb”.

How to Submit Your Homework?

Submission in NCKU Moodle. Before submitting your homework, please zip your coding files in a zip file, and name the file as “StudentID_hw2.zip”. For example, if your StudentID is H12345678, then name your file name as “H1234f5678_ hw2.zip”. Then submit your zipped file using NCKU Moodle platform <http://moodle.ncku.edu.tw> .

Have Questions about This Homework?

Please feel free to visit TAs, and ask/discuss any questions in their office hours. We will be more than happy to help you.