Homework 2

Deadline: 2024/10/16 (Wed.) 23:59

Problem 1: Movie Data Analysis with Pandas hw2_movie.ipynb

In this homework, you are asked to write a program for answering the following questions based on IMDB Movie data (**IMDB-Movie-Data.csv**). The output format of each question is free. You must use **Pandas** package to answer each question at this time. In addition, you also need to write your code in **Jupyter Notebook** (.ipynb), and use one code block for each question.

	Question
(1)	Top-3 movies with the highest ratings in 2016?
(2)	The actor generating the highest average revenue?
(3)	The average rating of **Emma Watson**'s movies?
(4)	Top-3 directors who collaborate with the most actors?
(5)	Top-2 actors playing in the most genres of movies?
(6)	Top-3 actors whose movies lead to the **largest** *maximum gap of years*? *Example of "maximum gap of years":* Tom Cruise has movies: "Edge of Tomorrow" in 2014, "Mission: Impossible - Rogue Nation" in 2015, "Oblivion" in 2013, "Jack Reacher" in 2012, "Mission: Impossible III" in 2006, "Jack Reacher: Never Go Back" in 2016, "Rock of Ages" in 2012, "Mission: Impossible - Ghost Protocol" in 2011. **The maximum gap of years is 2016-2006 = 10**
(7)	Find all actors who collaborate with **Johnny Depp** in *direct* and *indirect* ways Example: A collaborates with B B collaborates with C and D C collaborates with E and F D collaborates with A and G G collaborates with H → All actors directly and indirectly collaborating with A include: `[B, C, D, E, F, G, H]`

Problem 2: In-Game Purchase Data Analysis hw2_purchase.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_purchase.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_purchase.ipynb**". For more details, please refer to "hw2_purchase.ipynb".

Problem 3: DBSCAN Clustering Implementation hw2_dbscan.ipynb

Problem 4: Data Analysis via Visualization hw2_automobile.ipynb

Problem 5: kNN Graph Plot and Analysis hw2_knn_graph.ipynb

Problem 6: MLB Data Crawling and Analysis hw2_mlb.ipynb

Important Notes

This is a homework for each **individual**. You are asked to **write comments** to describe the meaning of each part of your codes **in either code block or markdown**.

How to Submit Your Homework?

Before submitting your homework, please zip all .ipynb files into a .zip file, and name the file as "StudentID_hw1.zip". For example, if your StudentID is H12345678, then your file name is: "H12345678_hw**2**.zip" or "H12345678_hw**2**.rar". Then submit your 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.