DS 501 Case Study Assignment 2 (120 Points)

Due: 05:00 p.m. on 03/16/2022

Case Study Objective: Analyzing Data from MovieLens

- Study and analyze the MovieLens 1M data set
- Make conjectures, support or refute those conjectures with data, and tell a story about the data

Case Study Deliverables: Please compress all the below files into a zipped file and submit the zip file (**team#_case_study_2.zip**) to Canvas, where **#** is your team number.

1. **Notebook File**: Save this Jupyter Notebook (see the provided template) for Problem 1 to Problem 4 and find the notebook file in your folder (for example, "filename.ipynb"). This is the file that you need to submit. Please make sure that all the answers and plotted tables and figures are in the notebook.

2. **PDF Report**: Please (1) prepare a report (**At Least 5 pages and Not More Than 10 pages**) to report what you found in the data and (2) include figures and tables in the report, but no source code.
   - Motivation and Background: Why is this case topic interesting and or important to you? Specifically, this case topic should be from a public organization, private company, social community, etc., in any domain.
   - Data Sources: What data did you collected? Please also describe your data characteristics in detail.
   - Methodology: How did you analyze the data? Please explain your step-by-step procedure in detail.
   - Results: What did you find in the data? That it, what insights can you obtain from the data?
   - Conclusions: What conclusions can you make from the results?

3. **PPT Slides**: Please prepare PPT slides (for 30 minutes' talk) to present about the case study.
   - 25 minutes for each presentation
   - 5 minutes for questions

Required Readings: Please read the below chapters of the books, **Python for Data Analysis, 2nd (2017) and 3rd (2022) Edition**, and the corresponding codes for your case study as a reference. You can read these online books free on.

- 2nd Edition of Chapter 14.2 (MovieLens 1M Dataset): https://learning.oreilly.com/library/view/python-for-data/9781491957653/ch14.html#whetting_movielens
- 3rd Edition of Chapter 5, 6, and 9: https://learning.oreilly.com/library/view/python-for-data/9781098104023/

Required Python Libraries:

- **Python for Data Analysis, 2nd Github (**https://github.com/wesm/pydata-book)
- **Pandas** (pandas.pydata.org)
- **Matplotlib** (matplotlib.org)

Note:

(1) The above online book is the latest version that I could find for your reference.
*(2) Please save the given notebook (**.ipynb**) frequently when working in JupyterLab; otherwise, the changes that you made will be lost.*
(3) Assignments are accepted in their assigned Canvas drop box without penalty if they are received by 05:00PM EST on the due date, or 10% of the graded score is deducted for the late submission per day. Work submitted after one week of its original due date will not be accepted.

Grading Criteria:

I. Notebook: Points = 80

Problem 1 (20 Points Total):

(10 points) Importing the MovieLens Data Set and Merging It into a Single Pandas DataFrame

- (2 Points) Download the 1 million ratings data set from the MovieDataSet.zip on Canvas
- (5 Points) Merge all of the data into a single Pandas DataFrame
- (3 Points) Store the data into an HDF5 file.

(10 points) Report Some Basic Details of the Data You Collected. For example:

- (2 Points) How many movies have an average rating over 4.5 overall?
- (2 Points) How many movies have an average rating over 4.5 among men? How about women?
- (2 Points) How many movies have an *median* rating over 4.5 among men over age 30? How about women over age 30?
- (2 Points) What are the ten most popular movies?
  - Choose what you consider to be a reasonable defintion of "popular".
  - Be perpared to defend this choice.
- (2 Points) Make some conjectures about how easy various groups are to please? Support your answers with data!
  - For example, one might conjecture that people between the ages of 1 and 10 are the easiest to please since they are all young children. This conjecture may or may not be true, but how would you support or disprove either conclusion with with data?
  - Be sure to come up with your own conjectures and support them with data!

In [1]:
```
import pandas as pd
#--------------------------------
# Your code goes here.
# Add as many cells as you need
#--------------------------------
```

Problem 2 (20 Points Total): Expand Our Investigation to Histograms

- (2 Points) Plot a histogram of the ratings of all movies.
- (2 Points) Plot a histogram of the *number* of ratings each movie recieved.
- (2 Points) Plot a histogram of the *average rating* for each movie.
- (6 Points) Plot a histogram of the *average rating* for movies which are rated more than 100 times.
  - What do you observe about the tails of the histogram where you use all the movies versus the one where you only use movies rated more than 100 times?
  - Which highly rated movies would you trust are actually good? Those rated more than 100 times or those rated less than 100 times?
- (8 Points) Make some conjectures about the distribution of ratings? Support your answers with data!
  - For example, what age range do you think has more extreme ratings? Do you think children are more or less likely to rate a movie 1 or 5?
  - Be sure to come up with your own conjectures and support them with data!

In [2]:
```
#--------------------------------
# Your code goes here.
# Add as many cells as you need
#--------------------------------
```

Problem 3: (20 Points Total): Correlation - Men versus Women

Let look more closely at the relationship between the pieces of data we have.

- (2 Points) Make a scatter plot of men versus women and their mean rating for every movie.
- (2 Points) Make a scatter plot of men versus women and their mean rating for movies rated more than 200 times.
- (6 Points) Compute the *correlation coefficent* between the ratings of men and women.
  - What do you observe?
  - Are the ratings similiar or not? Support your answer with data!
- (10 Points) Conjecture under what circumstances the rating given by one gender can be used to predict the rating given by the other gender.
  - For example, are men and women more similar when they are younger or older?
  - Be sure to come up with your own conjectures and support them with data!

In [3]:
```
#------------------------------
# Your code goes here.
# Add as many cells as you need
#------------------------------
```

Problem 4: (20 Points Total): Open Ended Question

- (6 Points) Do any of your conjectures in **Problems 1, 2, and 3** provide insights that a movie company might be interested in?
- (6 Points) Propose a business question that you think this data can answer.
- (8 Points) Suppose you are a Data Scientist at a movie company. **Convince your boss that your conjecture is correct!**

In [ ]:
```
```

II. Report (At Least 5 Pages and Not More Than 10 Pages): Points = 20

(1) Motivation and Background: Why is this case topic interesting and or important to you? Specifically, this case topic should be from a public organization, private company, social community, etc., in any domain. **Points: 5**

(2) Data Sources: What data did you collected? Please also describe your data characteristics in detail. **Points: 2**

(3) Methodology: How did you analyze the data? Please explain your step-by-step procedure in detail. **Points: 5**

(4) Results: What did you find in the data? That it, what insights can you obtain from the data? **Points: 5**

(5) Conclusions: What conclusions can you make from the results? **Points: 3**

III. Slides (25 Min of Presentation; 5 Min of Q&A): Points = 20

(1) Motivation and Background: Why is this case topic interesting and or important to you? Specifically, this case topic should be from a public organization, private company, social community, etc., in any domain. **Points: 2**

(2) Data Sources: What data did you collected? Please also describe your data characteristics in detail. **Points: 2**

(3) Methodology: How did you analyze the data? Please explain your step-by-step procedure in detail. **Points: 2**

(4) Results: What did you find in the data? That it, what insights can you obtain from the data? **Points: 2**

(5) Conclusions: What conclusions can you make from the results? **Points: 2**

(6) Overall Quality of Your Presentation: **Points: 10**

- Storytelling, i.e., how all the above items fit together as a story?
- Description of topic motivation/background, data collection, methodology, results, and conclusions
- Explanation of Question of Interest from the audience