**Lab #3 - Pandas**

The notebook should be well organized. Each section should be **clearly labeled with the exercise (and part) that it addresses** (e.g., Exercise #1a, #1b, #2) in a Markdown cell block. Use (clear and concise) comments as needed to help describe each step of your process. All notebook cells that contain essential steps should be executed and the output should be visible, so as to demonstrate your successful completion of the exercise. If you cannot complete an exercise in its entirety, you should make an effort to demonstrate your intermediate progress in order to maximize partial credit, and move forward as best as possible. You may submit any written answers to the exercises in the notebook as text cells.

**Background**

The [attached data filePreview the document](https://umd.instructure.com/courses/1266059/files/54045060/download?wrap=1) (acquired via Kaggle) contains information about historical video game sales, including the rank (in sales), title, platform (system), release year, genre, publisher, and sales (in millions) in North America, Europe, Japan, other countries, and in total. Complete the following exercises using the data.

**Exercise #1 - 1 point**

Step 1.1: Load the data file into a data frame, setting the Rank column as the index.

Step 1.2: Preview the first ten titles in the data set.

Step 1.3: Report the total number of titles.

Step 1.4: Report some summary statistics about the distribution of global sales.

Step 1.5: What proportion (or percentage) of titles earned at least $1 million in global sales? Hint: Use a filter and .count() then divide by your result from 1.3.

**Exercise #2 - 2 points**

Step 2.1: Filter the DataFrame to show  the top 3 sellers in the Action genre.  Hint:  Use sort\_values() and .head().

Step 2.2: Generalize the filter you created above. Write a function **top\_sellers** that takes in the DataFrame of video game sales (*df*), a category\* (*cat*, default=None), value (*val*, default=None), and a number (*n*, default=5) that specifies the number of rows to display in the result. The function should return a DataFrame containing the top *n* video games (in terms of global sales) for a specific value in the category (if specified) or the top *n* video games overall (if cat and val are not specified).

\*The category can be any column in the DataFrame, for example Platform or Publisher. The value is a value that exists in that column. You DO NOT have to implement error handling to account for function inputs that do not exist in the DataFrame.

Step 2.3: Use the function to display the following top selling games:

* Top 5 overall sellers
* Top 10 NES games
* Top 10 selling games in the year 2000
* Top 10 selling sports games

**Exercise #3 - 1 point**

Step 3.1: Generate a (linear) correlation matrix between sales in specific markets (North America, Europe, Japan, Other) and global sales. (Hint: How do we select multiple columns (all the sales columns) using pandas? Then use .corr() on the columns we are interested in.) What do you observe about the relationship between sales in these various markets?

Step 3.2: Repeat this calculation for the subset of video games released in the year 2000, and compare these results to those for the subset of video games released in the year 2015. What do you observe about how the relationship between the sales figures has changed over the 15-year period?