MLDS 422 – Fall 2023
Project 1
Due Friday, 10/27/23 at 11:59pm

Exercise 1: OOP & Pandas Practice

Combine the **baby names by state** data set into a single file using shell commands.

Write Python code to do the following:

1. Create a **class** called **BabyNames**: The class should offer the following interfaces:
    a. **Constructor** where you pass the file location to create a Pandas DataFrame
    b. **Count** (state='', year=''): returns the total number of births. A blank state or year should return all births of the empty input.
    c. **Top10BabyNames** (state='IL', year=2015): your output should look like (empty state means all, same for year):

Rank	Male	Female
1	Noah	Emma
…	…	…

    d. **ChangeOfPopularity** (fromYear=2014, toYear=2015, top=10): This function should list baby names (male or female) that showed change in popularity as follows:
        - Names that increased in popularity
        - Names that decreased in popularity
        - Names having the same popularity
    e. **Top5NamesPerYear** (year=2015, sex=''): Returns a table that shows the five most frequent given names, by State, for male, female, or both in a given year. The number to the right of each name is the number of occurrences in the data. (see below for format)

Top Five Female Names for Births in 2015										
State	**Rank 1**	**Num**	**Rank 2**	**Num**	**Rank 3**	**Num**	**Rank 4**	**Num**	**Rank 5**	**Num**
Alabama	Ava	297	Emma	285	Olivia	258	Harper	213	Elizabeth	186
Alaska	Olivia	56	Emma	49	Aurora	46	Amelia	39	Ava	39

    f. **NamePopularityPlot** (name='Jim', yearRange=(2000,2015), state='IL', sex='M'): This function will create a plot that shows the name popularity changes over the years. (popularity is based on the proportional use of the name within a state and year)

g. **NameFlip(n=10)**: List top n names that flipped over the years. (i.e. from boy name to girl or the reverse). Provide a plot of the names showing the year.

2. Make sure to document your class and follow the Python standards.

3. Tell another story from this baby names data set. Support your story with plots.

Exercise 2: Statistics & Data Visualization Practice

In a given course the following applies (the data set **exams.csv** file is attached):

- The system tracks students by student name and unique ID.
- Grades are based on:
    - Exams Score (40% of the final grade)
    - Projects Score (30% of the final grade)
    - Quizzes Score (30% of the final grade)
- Final Grades are based on the final score (out of 100) as follows:
    - [90-100]: **A**
    - [80-90): **B**
    - [20-80): **C**
    - [10-20): **D**
    - [0-10): **F**

1. Load the **exams.csv** file into a DataFrame.

2. Identify outlier students, or those who have final scores that are outside of the mean +/- two standard deviations of the final scores.

3. Create box plot parameters (not drawing them, but just computing the numbers, min, max, median, Q1 and Q3 for a box plot).

4. Create a seaborn visualization that shows the final letter grades distribution. Choose the visualization that you think best represents the data. Explain your reasoning.

5. Discover two more insights from the data. Support your insights with calculations and/or seaborn plots.