# Project One

## Overview of Project One

### Project Description

**This project is designed to test your current ability with using python to read files, loop through the files, perform text manipulation operations and basic analysis, and find and use available help resources online.**

Only concepts already learned to this point are covered in this exercise. This is a more structured project, as such there is a correct **output**that you will be graded against, but there are different ways to achieve the same end goal.

|  |
| --- |
| **Please complete this exercise as independently as possible. Please be reminded of the following excerpt from the class syllabus and schedule.**  *Collaboration between students is a characteristic of a healthy learning environment, however, these guidelines must apply. You are required to do your own thinking, design your own projects, and write your own code/syntax. You are allowed and encouraged to talk to other students about high level concepts and the content of lectures and assigned readings. However, you are not allowed to lead another student to the extent that the solution becomes obvious, and the assignment becomes less challenging for them. You are also not allowed to request such assistance explicitly or implicitly. For instance, never sit down to discuss an assignment with someone else before you have analyzed the problem in depth on your own. I reserve the right to carry out a “wire-pull test” (i.e., ask you to explain specifics of a piece of code or output of work assigned to be done individually). See the link referenced below for detailed examples of acceptable and non-acceptable collaboration between students.*  *Ref: Tom Doeppner, Brown University http://cs.brown.edu/courses/csci1670/collab\_167.pdf* |

Also, remember...

|  |
| --- |
| We will have time set-aside in class for “Project One Check-In” where we can work together on the challenges you are having on the project. |

### What you are to submit

* Write down your answer to the open-ended question
* Paste in your formatted code (I’ll be re-running the code)
* Include any references to resources you consulted online (will be graded also)

### Launching your Python Environment

Use your preferred tool to complete the project.

* [https://www.anaconda.com/download/](https://www.google.com/url?q=https://www.anaconda.com/download/&sa=D&source=editors&ust=1628628476205000&usg=AOvVaw3xzGA6KLQVlPX_QTBtf_MY)) (installation required)
* [https://colab.research.google.com](https://www.google.com/url?q=https://colab.research.google.com&sa=D&source=editors&ust=1628628476205000&usg=AOvVaw389EaasQXREGY3LvG3MsvX) (browser based, login required)
* [https://repl.it/languages/python3](https://www.google.com/url?q=https://repl.it/languages/python3&sa=D&source=editors&ust=1628628476206000&usg=AOvVaw33TGiFT49ZXBBgFXZk32M-)  (browser based, no log in required)

### Project One Data

|  |  |
| --- | --- |
| **Dataset:** | [https://www.dropbox.com/s/hxch4v08bkthvz1/data.csv?dl=1](https://www.google.com/url?q=https://www.dropbox.com/s/hxch4v08bkthvz1/data.csv?dl%3D1&sa=D&source=editors&ust=1628628476208000&usg=AOvVaw3uLVbLUFL4muF1U5J3IWQ8) |
| **About:** | The data contains information about movies and TV shows that have been added to the Netflix platform. The data has been cleaned and is ready for analysis. If there were any errors in the fields that were scraped, such as spelling errors or derivatives like "nite" instead of "night," they haven't been corrected. |
| **Fields:** | The dataset features 7777 observations with the columns (features) show id, show type, title, director, cast, country, rating, release year, year added, duration, seasons, genre, description. |

### Notes on importing the data

Hint: You are importing a csv file. This is different from the kind of imports we did in class, although the code we used in class may also serve. See if you can figure out how to import this file by consulting the documentation of the function here:[https://docs.python.org/2/library/csv.html](https://www.google.com/url?q=https://docs.python.org/2/library/csv.html&sa=D&source=editors&ust=1628628476213000&usg=AOvVaw2aI8rjnIjrZ_KHBNdmG0ta)

Also, consult resources online (via google).

### Exercise One

By searching the country only, answer the following questions:

1.       How many movie titles mention the  United States?

2.       How many movie descriptions mention the United States?

3.       How many movies are from the United States?

### Exercise Two

Write a function to receive any name as input, search the director only and return the number of movies/TV shows that contain that name. Your function should meet the following requirements:

* Your function should receive any name as an input string and return a simple message with the number of movies/TV shows that contain those names. E.g.  There are [5] directors named [supplied name] in this dataset.
* If the name supplied were not found in the database, return the message No directors were found with the name: [supplied name]
* Remember to deal with whatever letter case you are supplied i.e.  all caps or all lowercase, etc.
* Test your function with the word Paul and confirm that your result reads  There are [67] directors named [Paul] in this dataset.

### Exercise Three

Identify the most common words in the description of TOP TEN LONGEST MOVIES ONLY from the dataset.  Please write down the 20 most common words you identify from the description of top ten longest movies.

|  |
| --- |
| Hint: Remember that you are only interested in the top 10 longest movies. |

|  |
| --- |
| You will have to go online and **FIND OUT HOW TO DO THIS!** Find and use the learning resources out there on the internet and modify them for your purpose. Make sure to Include links/references you consulted to your final submission. |

### Exercise Four

Write a function to receive a cast name as input, search the dataset and return information about that actor from the database. Your function should meet the following requirements:

* Your function should receive an actor name as an input string and return
* (1) their average movie duration rounded up to a whole number, and
* (2) a list of their movies from the list.
* If the actor name supplied is not found in the database, return the message: No movie found for the artist: [supplied actor name]. E.g, If I search for an artist named Jennifer Lopez, the function should return: No movie found for the artist: Jennifer Lopez (this is an example!!).
* Remember to deal with whatever letter case you are supplied i.e. all caps or all lowercase, etc.
* Remember to consider the file type of EVERY row of information when writing out your function.

### Exercise Extra (for extra credit)

Using a python graph library select one or two diagrams that will represent in the best way one of the exercises completed above.

Good luck!   
  
Dr. Samayita Guha