Homework 2

Data Collection

Due: July 7, Friday, 5 pm

Instructions

Please carefully read and follow the instructions below.

- Ensure that your responses are clear and presented in an organized manner. Submit the questions in the order given in the homework.
- Additional guidelines can be found on Page 10.
- Almost all questions worth one or two points can be answered in a single line of code. You don't have to answer it in one line, but they can be. The expectation is that you use different library functions and features.
- There are three main questions in total and 6 points for formatting, as mentioned in Q4.
- Create a new folder and name it your FirstName_LastName. Create a new notebook for answering the questions in this folder. Name the notebook HW2_Firstname_Lastname. Store all datasets used in the homework within this folder. Zip the folder and upload it to D2L. This way, we can run everyone's notebooks without a path error.

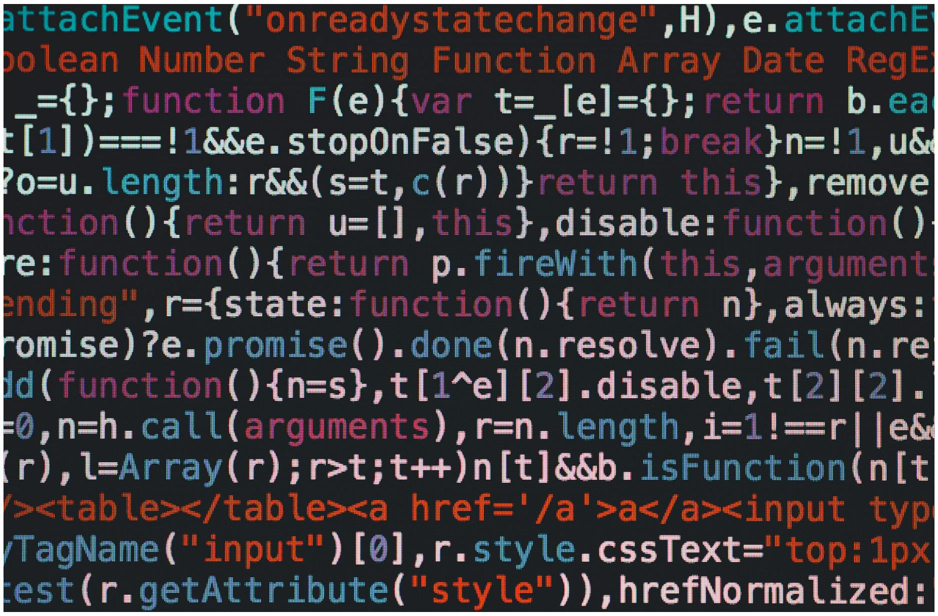
Announcement

Homework one was 15%, and Homework two was 15%. When a question had to be dropped in Homework One, the total went down. So the percentages were adjusted as follows. Now Homework 1 is 10.5% of the total grade, and Homework 2 is 19.5% of the total grade.

Submission	**Total**	**% Contribution to Final Grade**	**Impact of 1 point of Submission in Final Grade**
HW1	14	10.5	0.75
HW2	50	19.5	0.39
HW3	-	20	-
HW4	-	20	-
Final Project	-	20	-
Weekly Checkin	-	5	-
Participation Activities	-	5	-
Total	-	100	-
Bonus Questions	-	5 (tentative)	1

Table 1: Grade Distribution

1 What's the secret code? [4 points]



```
attachEvent("onreadystatechange",H),e.attachE
oolean Number String Function Array Date RegE
_={};function F(e){var t=_[e]={};return b.ea
t[1])===!1&&e.stopOnFalse){r=!1;break}n=!1,u&
?o=u.length:r&&(s=t,c(r))}return this},remove
nction(){return u=[],this},disable:function()
re:function(){return p.fireWith(this,argument
ending",r={state:function(){return n},always:
romise)?e.promise().done(n.resolve).fail(n.re
dd(function(){n=s},t[1^e][2].disable,t[2][2].
=0,n=h.call(arguments),r=n.length,i=1!==r||e&
(r),l=Array(r);r>t;t++)n[t]&&b.isFunction(n[t
/><table></table><a href='/a'>a</a><input typ
/TagName("input")[0],r.style.cssText="top:1px
test(r.getAttribute("style")),hrefNormalized:
```

Credit: Photo by Markus Spiske on Unsplash

The Bonus Question One page appears normal when you load it in a web browser, but there's a hidden secret code embedded within the HTML code.

To find the secret code, you need to inspect the page's HTML. Suppose the secret code is CSC380, Look for the section with the attribute "Secret." Within that section, you will find the code enclosed between two tags: <CSC380></CSC380>.

1. Use the library we learned in class to get the page's HTML. [1 point]
2. Find the section with the secret code by using the Beautiful Soup's find function. [2 points]
3. Clean up the secret code and print it as "The Secret Code is: CSC380 " The secret code should not have the </> symbols on them. [1 point]

You will get no points if you just give the secret code as an answer.

2 Random Facts API [11 points]

Credit: Photo by Agence Olloweb on Unsplash

In a previous class, we explored the TikTok API, where we visited their website, examined the necessary API, and studied the provided examples. Now, it's time to apply our knowledge.

The website Useless Facts - "https://uselessfacts.jsph.pl/" offers a free API, requiring no sign-up, which provides random facts. They have two APIs available. One for random fact and another for the random fact of the day.

2.1 10 Random Facts

Follow instructions as given.

1. **Correct URL** **[1 point]**

   Find the URL for random facts API.

2. **Utilizing the Library for URL Handling** **[1 point]**

   Use the library we discussed in class to call the API 10 times.

3. **Checking Status Code** **[1 point]**

   After making each API call, verify if the status code is 200, indicating a successful response. If the status code is different, it signifies an error or issue that needs to be addressed. You don't have to print the status code for each API call, but the code where you checked if the status code is 200 should be present.

4. **Implementing Time Gaps** **[1 point]**

   To ensure smooth and responsible API usage, introduce time gaps between each API call. You can use the sleep function from the time library. The code where you introduced the time gap should be present.

5. **Creating the Dataframe** **[1 point]**

   Finally, use the collected random facts to create a data frame. Your header should be : (id, text, source, source_url, language, and permalink). Display the data frame.

6. **Display Full Facts** **[2 points]**

   You may notice that by default, pandas trim the facts when displaying them.

   (a) We have learned how to get data from a column. Print the column containing data as a Series. [1 point]

   (b) Convert the Series to a list. Do not use the list() function. We learned of another function to convert a Series to a list. Print the random facts as a list. [1 point]

7. **Show 3 random facts from the data frame** **[1 point]**

   We learned a function that randomly picks a number of rows and displays them. Use it to show 3 rows with random facts.

8. **What is today's random fact?** **[3 points]**

   (a) Utilize the available API to retrieve the **random fact of the day.** [1 point]

   (b) Use the DateTime library to obtain the current date and time when making the API call. [1 point]

   (c) Display the date and time of the API call, along with the corresponding random fact of the day. [1 point]

3 Movies and Shows [29 points]

Credit: Photo by Piotr Cichosz on Unsplash

1. **Download the following datasets. [2 points]**

   You don't have to download using code. You can go to the website, click on download, and get the dataset. But unzip the files using code.

   (a) Hulu: https://www.kaggle.com/datasets/shivamb/hulu-movies-and-tv-shows

   (b) Amazon Prime:https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows

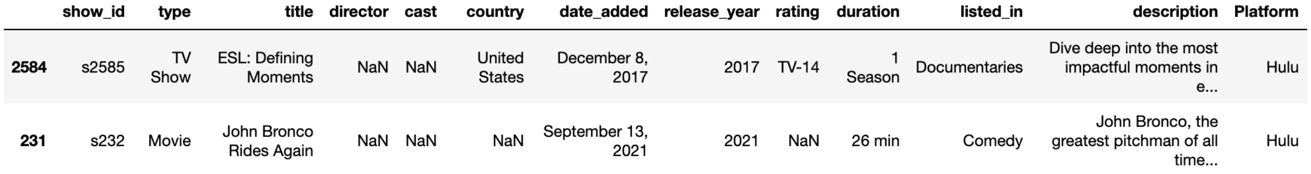
   (c) Netflix : https://www.kaggle.com/datasets/shivamb/netflix-shows

   (d) Disney Plus: https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows

2. **Create one large dataframe. [3 points]**

   (a) Create a separate dataframe for each streaming service. Use names that make it easy to understand which is what. [1 point]

   (b) Create a column named *Platform* with the name of the streaming service for all four dataframes. For example refer to Figure 1. [1 point]



	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	Platform
2584	s2585	TV Show	ESL: Defining Moments	NaN	NaN	United States	December 8, 2017	2017	TV-14	1 Season	Documentaries	Dive deep into the most impactful moments in e...	Hulu
231	s232	Movie	John Bronco Rides Again	NaN	NaN	NaN	September 13, 2021	2021	NaN	26 min	Comedy	John Bronco, the greatest pitchman of all time...	Hulu

Figure 1: Two rows of the dataframe with Hulu information.

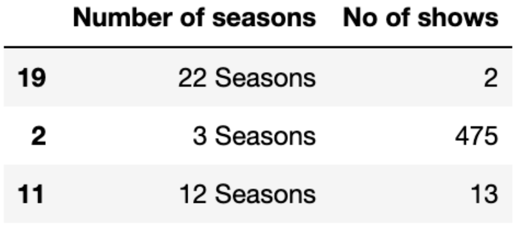
(c) Combine them all to create one large dataframe named *all_ platforms_ df*. Use the pandas functionality to convert a list of dataframe to one. . [1 point]

3. **Longest show and movie** **[6 points]**

(a) Display a dataframe with information on how many shows had how many seasons.

Random rows of the expected dataframe look like Figure 2. Your cell values could be different (if the dataset is updated after this homework is released).

How many seasons were the longest-running show (max number of seasons) ? [3 points].



	Number of seasons	No of shows
19	22 Seasons	2
2	3 Seasons	475
11	12 Seasons	13

Figure 2: Example: A part of the output dataframe Q3.a

(b) ERROR EXPECTED : Repeat Q3.a for Movie duration in minutes. Our goal is to find which is the longest movie and how long it was. Since we have not performed any data cleaning, I am expecting you to get error or incorrect information. Without any preprocessing, follow the instructions.

i. Use the function used to sort a series by the values. What appears to be the longest movie? [1 point]

ii. Use the function used to return a Series containing counts of unique values. What appears to be the longest movie? [1 point]

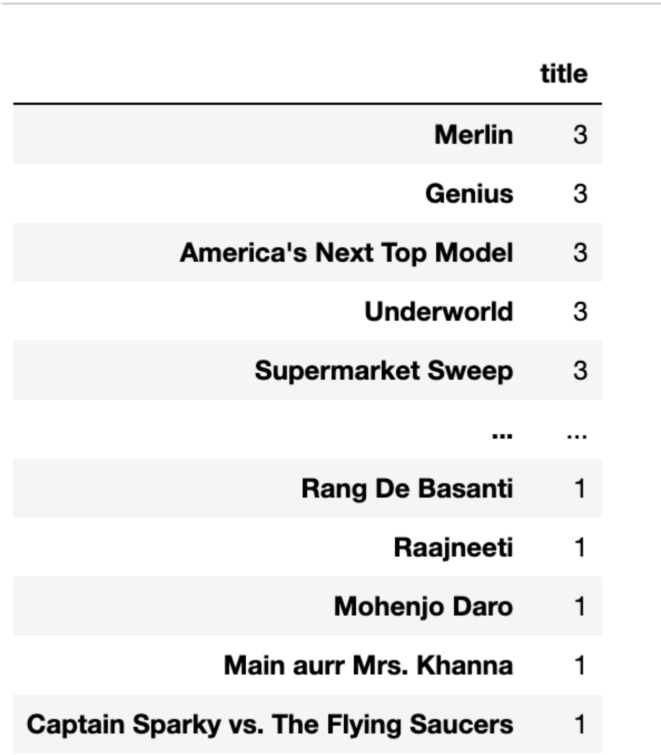
iii. Use the function that returns the maximum value in a series. What appears to be the longest movie? Why do you think you got an error here? [1 point]

4. **Shows streaming on multiple platforms** **[7 points]**

(a) What is the number of rows in *all_platforms_df*? Do not use len(), there is another functionality that will give you the number of rows and columns in a dataframe. Usage of that is the expectation here. [1 point]

(b) How many unique titles are in the dataframe? [1 point]

(c) Count the number of times each title appears in the dataframe *all_platforms_df*. Display this as a dataframe named *titles_count_df*. The expected dataframe has indexes that are shows/movies and a column named title with the number of times the show appeared in the dataframe. Example below. [2 points]



	title
Merlin	3
Genius	3
America's Next Top Model	3
Underworld	3
Supermarket Sweep	3
...	...
Rang De Basanti	1
Raajneeti	1
Mohenjo Daro	1
Main aurr Mrs. Khanna	1
Captain Sparky vs. The Flying Saucers	1

Figure 3: Example of expected answer for Question 4 (c)

(d) Create a new column *name* in *titles_count_df* from the index. Replace the current index with row numbers. [1 point]

(e) Change the column name of *titles_ count_ df* as *Movie or Show Name* in place of *name*. Change the column name *title* with *No of Platforms*. [1 point]

(f) What is the maximum number of streaming platforms a show is on ? The usage of a single function we learned in class is what is expected here. [1 point]

5. **Favorite show or movie** **[2 points]**

   (a) Display the rows with shows that are on at least two platforms. [1 point]

   (b) Picking any show/movie. If your favorite show or movie is in this displayed dataframe in Q4.a, pick that.

   Else pick a random show/movie using the pandas functionality that gives you a random row. Comment the show name in that code line.
   Display the rows in *all_platforms_ df* that has information about the show. [1 point]

6. **Save Data** **[1 points]**
   Save *all_platforms_ df* as a csv without the index.

7. **Name starts with ...** **[8 points]**

   (a) Read the file you saved in the question above. [0 point]

   (b) Display all the rows containing movies starting with your First name's first letter. For instance, my name is Enfa, so I would display all movies that start with E. [2 points]

   (c) How many unique shows start with your name's first letter? [2 points]

   (d) If our TA were to answer Q7b, what would be the output? [1 point]

   (e) If our Ta were to answer Q7c, what would be the output? [1 point]

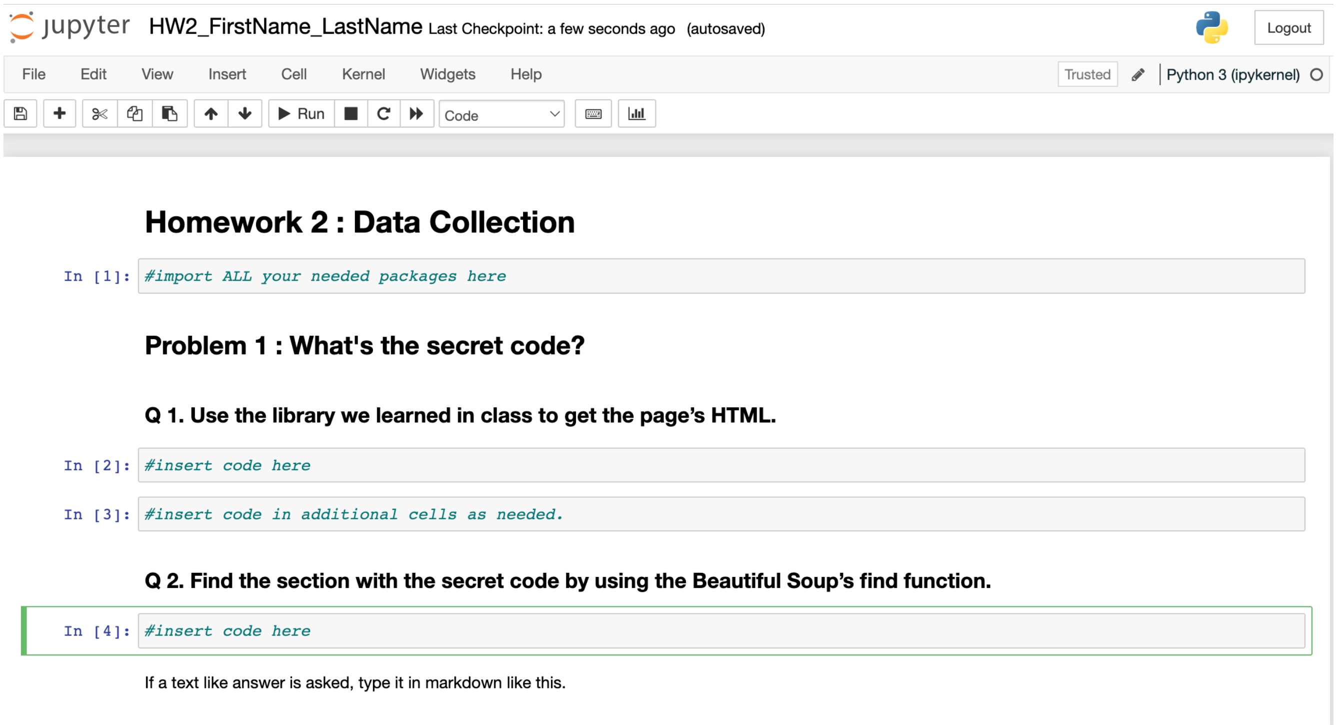
   (f) If our Ta were to answer Q7c, but with his last name instead of first name, what would be the answer? [1 point]

   (g) What is the difference (as in an integer) between Q7.e and Q7.f [1 point]

4 Notebook Presentation [6 points]

To ensure a clear and organized solution, please follow the guidelines below:

1. **Use Markdown:** Utilize markdown cells to create headings for each question or step within the notebook. This will help structure the content and make it easier to read and navigate.
2. **Presentable Formatting:** Format the notebook in a visually appealing manner, with consistent indentation, proper spacing, and clear separation between sections. This will contribute to the overall readability of the notebook.
3. **Display Outputs:** Make sure to include all relevant outputs in the notebook when uploading it. This way, the solution can be reviewed and understood without the need to run the notebook separately. We will run if not all, some notebooks during evaluation.



Homework 2 : Data Collection

```
In [1]: #import ALL your needed packages here
```

Problem 1 : What's the secret code?

Q 1. Use the library we learned in class to get the page's HTML.

```
In [2]: #insert code here
In [3]: #insert code in additional cells as needed.
```

Q 2. Find the section with the secret code by using the Beautiful Soup's find function.

```
In [4]: #insert code here
```

If a text like answer is asked, type it in markdown like this.

Figure 4: A Screenshot of Notebook with guidelines.