

Assignment-1

Specification

Make Submission

Check Submission

Collect Submission

The assignment data has been extracted from a Movie dataset on Kaggle (<https://www.kaggle.com/rounakbanik/the-movies-dataset>), with some minor modification to make things interesting. The dataset is split into two CSV files **credits** (<https://github.com/mysilver/COMP9321-Data-Services/raw/master/20t1/credits.csv>) and **movies** (<https://github.com/mysilver/COMP9321-Data-Services/raw/master/20t1/movies.csv>). Use the datasets to answer the following questions:

- **Question 1: (based on the both datasets) (0.5 Mark)**

Join the two datasets based on the "id" columns in the datasets, keeping the rows as long as there is a match between the id columns of both dataset (do not concatenate the datasets).

- **Question 2: (based on the dataframe created in Question-1) (0.5 Mark)**

Keep the following columns in the resultant dataframe (remove the rest of columns from the result dataset):

'id', 'title', 'popularity', 'cast', 'crew', 'budget', 'genres', 'original_language', 'production_companies', 'production_countries', 'release_date', 'revenue', 'runtime', 'spoken_languages', 'vote_average', 'vote_count'

- **Question 3: (based on the dataframe created in Question-2) (0.5 Mark)**

Set the index of the resultant dataframe as 'id'.

- **Question 4: (based on the dataframe created in Question-3) (0.5 Mark)**

Drop all rows where the budget is 0

- **Question 5: (based on the dataframe created in Question-4) (1 Mark)**

Assume that there is a ranking scheme for movies defined by " $(revenue - budget)/budget$ ". Add a new column for the dataframe, and name it "success_impact", and calculate it for each movie based on the given formula.

- **Question 6: (based on the dataframe created in Question-5) (1 Mark)**

Normalize the "popularity" column by scaling between 0 to 100. The least popular movie should be 0 and the most popular one must be 100. It is a float number.

- **Question 7: (based on the dataframe created in Question-6) (0.5 Mark)**

Change the data type of the "popularity" column to (int16).

- **Question 8: (based on the dataframe created in Question-7) (1.5 Marks)**

Clean the "cast" column by converting the complex value (JSONs) to a comma separated value. The cleaned "cast"

column should be a comma-separated value of alphabetically sorted characters (e.g., Angela, Athena, Betty, Chester Rush) . NOTE: keep unusual characters e.g., '(uncredited)' as they are; no need for further cleansing.

- **Question 9: (based on the dataframe created in Question-8) (1.5 Marks)**

Return a list, containing the names of the top 10 movies according to the number of movie characters (Harry Potter! is one character! do not count the letters in the title of movies!). The first element in the list should be the movie with the most number of characters.

UPDATE: You can assume that there is no COMMA in the characters.

- **Question 10 : (based on the dataframe created in Question-8) (1 Marks)**

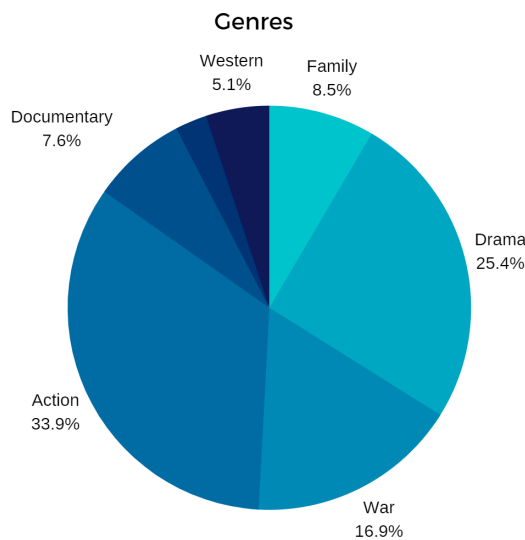
Sort the dataframe by the release date (the most recently released movie should be first row in the dataframe)

- **Question 11: (based on the dataframe created in Question-10) (2 Marks)**

- (1 .5 Mark) Plot a pie chart, showing the distribution of genres in the dataset (e.g., Family, Drama).

- (0.5 Mark) Show the percentage of each genre in the pie chart. Please be noted that the following figure is just a sample and it does not reflect the real values or the list of all genres in the dataset.

UPDATE: You can add a legend to your chart if labels overlap. You can also merge the some of the infrequent labels (up to 4) and name them "other genres".

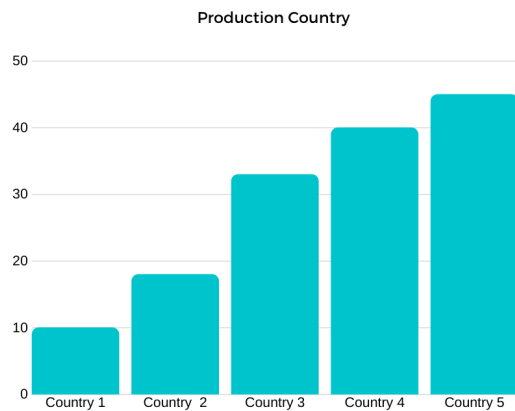


- **Question 12 : (based on the dataframe created in Question-10) (2 Marks)**

- (1.5 Marks) Plot a bar chart of the countries in which movies have been produced. For each county you need to show the count of movies.

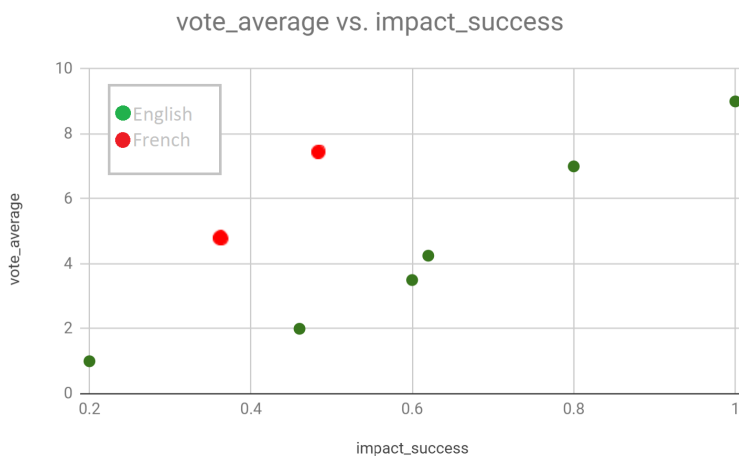
- (0.5 Mark) Countries should be alphabetically sorted according to their names.

Please be noted that the following figure is just a sample and it does not reflect the real values or the list of all countries in the dataset.



• **Question 13: (based on the dataframe created in Question-10) (2.5 Marks)**

- **(1.5 Marks)** Plot a scatter chart with x axis being "vote_average" and y axis being "success_impact".
- **(0.5 Marks)** Ink bubbles based on the movie language (e.g, English, French); In case of having multiple languages for the same movie, you are free to pick any one as you wish.
- **(0.5 Marks)** Add a legend showing the name of languages and their associated colors. **UPDATE: You can use both "original_language" (e.g. "en", "fr") or "spoken_languages" .**



Please be noted that the following figure is just a sample and it does not reflect the real values or the list of all countries in the dataset. (also the x and y axis should be swapped in the figure)

What not to forget!

Due Date: Friday the 13th of March 2020 17:59

Submit your script named " YOUR_ZID .py" (z2123232.py) which contains your code.

You are required to use the following code template (**it is not complete; please download the file**) for your submission:

```

import ast
import json
import matplotlib.pyplot as plt
import pandas as pd
import sys
import os
studentid = os.path.basename(sys.modules[__name__].__file__)
#####
# Your personal methods can be here ...
#####
def log(question, output_df, other):
    print("----- {}-----".format(question))
    if other is not None:
        print(question, other)
    if output_df is not None:
        print(output_df.head(5).to_string())
def question_1(movies, credits):
    """
    :param movies: the path for the movie.csv file
    :param credits: the path for the credits.csv file
    :return: df1
        Data Type: Dataframe
    Please read the assignment specs to know how to create the output dataframe
    """
    #####
    # Your code goes here ...
    #####
    log("QUESTION 1", output_df=df1, other=df1.shape)
    return df1
...

if __name__ == "__main__":
    df1 = question_1("movies.csv", "credits.csv")
    df2 = question_2(df1)
    df3 = question_3(df2)
    df4 = question_4(df3)
    df5 = question_5(df4)
    df6 = question_6(df5)
    df7 = question_7(df6)
    df8 = question_8(df7)
    movies = question_9(df8)
    df10 = question_10(df8)
    question_11(df10)
    question_12(df10)
    question_13(df10)

```

- You can download the code template from : <https://raw.githubusercontent.com/mysilver/COMP9321-Data-Services/master/20t1/z1111111.py> (<https://raw.githubusercontent.com/mysilver/COMP9321-Data-Services/master/20t1/z1111111.py>)
- If you do not follow this structure, you will not be marked.
- You can only add codes in the specified lines (do not edit the rest of the lines):

```

#####
# Your code goes here ...
#####

```

- If your code does not run on CSE machines for any reasons (e.g., hard-coded file path such as C://Users/), you will be penalize at least by 5 marks. We assume that the two csv files are located in the same directory of your script, and the name is the same as the one in the template (movies.csv, and credits.csv)

- **Please look at the documentation for each question method; it describes the inputs (e.g., a dataframe) and output (e.g., dataframe, list of movies) of the method.**

```

"""
:param df7: the dataframe created in question 7
:return: df8
        Data Type: Dataframe
        Please read the assignment specs to know how to create the output dataframe
"""

```

- **Please use the same variable names as mentioned in the comments (e.g., in question 8, you are supposed to create a dataframe and name it df8)**
- **In the last three questions, you need to plot charts; please do not use "plt.show()" function to pop up charts. The code template will automatically save the chart on the disk. What you need to do is to just call the plot functions of the dataframe (e.g., df.plot.pie()). We highly recommend you go through the lab activities to know how to plot charts.**

FAQ:

- **Can I pass extra variables to functions?**
No
- **Can we create our own functions besides the question functions (e.g., question_1)?**
Yes
- **Can I call another function inside the question functions? e.g., calling question_1 inside question_2**
Yes
- **What should I do if my charts are not shown automatically?**
Look at the lab sample codes; if still need a help, ask your tutor during the labs.
- **How should I print my dataframe?**
print(df.to_string())
- **Is it okay that the graph for Q8 does not pop up until the graph for Q7 is closed or should they both pop up at the same time?**
This is fine
- **Do the charts need to look the same (colors, legend position, grid) as the examples shown? or would it be fine to just use the default plotting from pandas?**
The default colours/fonts are fine
- **How are our submissions marked?**
They are marked manually by tutors, by running the following command: python3 z{YOUR_ZID}.py
- **What python packages can I use in my assignment?**
You can only use packages imported in the template file to do the assignment.
- **What version of python should I use?**
Python 3+
- **How I can submit my assignment?**
Go to the assignment page click on the "Make Submission" tab; pick your files which must be named "YOUR_ZID.py". Make sure that the files are not empty, and submit the files together.
- **Can I submit my file after deadline?**
Yes, you can. But 25% of your assignment will be deducted as a late penalty per day. In other words, if you be late for more than 3 days, you will not be marked.

Plagiarism

This is an *individual assignment*. The work you submit must be your own work. Submission of work partially or completely derived from any other person or jointly written with any other person is not permitted. The penalties for such offence may include negative marks, automatic failure of the course and possibly other academic discipline. Assignment submissions will be examined manually.

Do not provide or show your assignment work to any other person - apart from the teaching staff of this course. If you knowingly provide or show your assignment work to another person for any reason, and work derived from it is submitted, you may be penalized, even if the work was submitted without your knowledge or consent. Pay attention that is **also your duty to protect your code artifacts** . if you are using any online solution to store your code artifacts (e.g., GitHub) then make sure to keep the repository private and do not share access to anyone.

Reminder: Plagiarism is defined as (<https://student.unsw.edu.au/plagiarism>) using the words or ideas of others and presenting them as your own. UNSW and CSE treat plagiarism as academic misconduct, which means that it carries penalties as severe as being excluded from further study at UNSW. There are several on-line sources to help you understand what plagiarism is and how it is dealt with at UNSW:

- Plagiarism and Academic Integrity (<https://student.unsw.edu.au/plagiarism>)
- UNSW Plagiarism Procedure (<https://www.gs.unsw.edu.au/policy/documents/plagiarismprocedure.pdf>)


Make sure that you read and understand these. Ignorance is not accepted as an excuse for plagiarism. In particular, you are also responsible for ensuring that your assignment files are not accessible by anyone but you by setting the correct permissions in your CSE directory and code repository, if using one (e.g., Github and similar). Note also that plagiarism includes paying or asking another person to do a piece of work for you and then submitting it as your own work.

UNSW has an ongoing commitment to fostering a culture of learning informed by academic integrity. All UNSW staff and students have a responsibility to adhere to this principle of academic integrity. Plagiarism undermines academic integrity and is not tolerated at UNSW.

Resource created 6 months ago (Thursday 06 February 2020, 07:50:24 AM), last modified 5 months ago (Tuesday 10 March 2020, 10:05:24 AM).

Comments

  (/COMP9321/20T1/forums/resource/41977)

 Add a comment



Jerry Edackatt (/users/z5229867) 4 months ago (Wed Mar 25 2020 10:08:16 GMT+1100 (澳大利亚东部夏令时间)), last modified 4 months ago (Wed Mar 25 2020 10:38:13 GMT+1100 (澳大利亚东部夏令时间))

Where are we meant to check the assignment? The "Collect Submission" tab says that the assignment has not yet been marked.

EDIT: Never mind, it's on give.

Reply



Dominik Tobaben (/users/z5298989) 4 months ago (Wed Mar 25 2020 11:17:19 GMT+1100 (澳大利亚东部夏令时间)), last modified 4 months ago (Wed Mar 25 2020 11:19:39 GMT+1100 (澳大利亚东部夏令时间))

What do you mean with "it's on give"? The "Collect Submission" tab says "not marked yet" for me too (I'm exchange student for this term and I'm not familiar with the general procedure here...)

Reply



Shashank Reddy Boosi (/users/z5222766) 4 months ago (Wed Mar 25 2020 11:25:15 GMT+1100 (澳大利亚东部夏令时间))

Click on the Grades, which is the button next to the your name under Lectures tab. Hope this helps :)

Reply



Dominik Tobaben (/users/z5298989) 4 months ago (Wed Mar 25 2020 13:53:02 GMT+1100 (澳大利亚东部夏令时间))

alright, got it! Thank you!

Reply



Shahryar Babar Bhatti (/users/z5267909) [5 months ago \(Sat Mar 14 2020 01:25:47 GMT+1100 \(澳大利亚东部夏令时间\)\)](#)

hey,

Im sorry i submitted my file and then i slept .I just came here to check my submissions and realized that i uploaded an old file.Now i have uploaded the original how much my assignment would be penalized because there is 25% deduction on 1 day late and im somehow 7 hours late now.

thanks

Reply



Yuao Jiang (/users/z5217274) [5 months ago \(Fri Mar 13 2020 18:14:21 GMT+1100 \(澳大利亚东部夏令时间\)\)](#)

Hey ,I'm sorry I forgot to submit my assignment 1 on time. I thought it was 6.pm but I forgot I'm in China mainland. I submitted 8 mins later after the dead line. Could you please tell me if I'm going to be penalized due to my late submission?

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) [5 months ago \(Fri Mar 13 2020 18:33:26 GMT+1100 \(澳大利亚东部夏令时间\)\)](#), last modified [5 months ago \(Fri Mar 13 2020 18:33:38 GMT+1100 \(澳大利亚东部夏令时间\)\)](#)

Hi,

No worries; I will waive the late submission in such cases.

Reply



Yuao Jiang (/users/z5217274) [5 months ago \(Fri Mar 13 2020 18:47:24 GMT+1100 \(澳大利亚东部夏令时间\)\)](#)

Thanks for your kindness! Really appreciate!

Reply



Shahryar Babar Bhatti (/users/z5267909) [5 months ago \(Fri Mar 13 2020 14:57:12 GMT+1100 \(澳大利亚东部夏令时间\)\)](#)

hey,

im confused with this creating dataframe in assignment submission file.how to i create do this.and should dataframe name be eg df1 ,df2,df3 or they can be as i have them in my code

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) [5 months ago \(Fri Mar 13 2020 15:02:45 GMT+1100 \(澳大利亚东部夏令时间\)\)](#)

Pleaaee keep the same names to avoid any unintentional changes to the template.

You can easily do as following :

df1 = Mydataframe

Reply



Shahryar Babar Bhatti (/users/z5267909) 5 months ago (Fri Mar 13 2020 15:08:37 GMT+1100 (澳大利亚东部夏令时间))

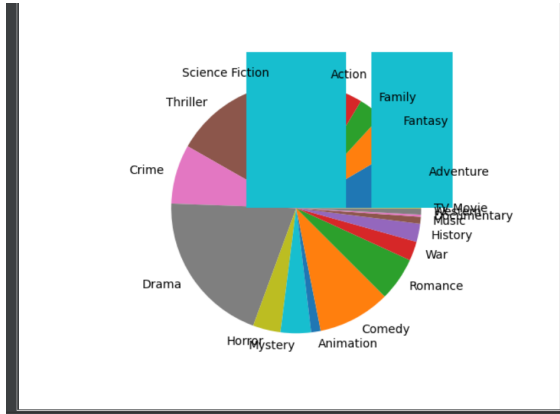
thanks Mohammad.

Reply



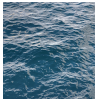
Hao Chen (/users/z5192421) 5 months ago (Fri Mar 13 2020 14:38:23 GMT+1100 (澳大利亚东部夏令时间))

My 'py-Q12.png' showing like this:



It seems overlap with 'py-Q11.png', but it looks normally when i run Q12 individually. Do you know the reason?

Reply



Jiayu Zhou (/users/z5240943) 5 months ago (Fri Mar 13 2020 16:07:12 GMT+1100 (澳大利亚东部夏令时间))

try plt.clf() before you plot the bar chart

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) 5 months ago (Fri Mar 13 2020 14:41:44 GMT+1100 (澳大利亚东部夏令时间))

Clear the plt before drawing a new figure

Reply



Warisara Tuntikornvorakul (/users/z5233249) 5 months ago (Fri Mar 13 2020 13:59:23 GMT+1100 (澳大利亚东部夏令时间))

For Q8, I would like to make sure that how to keep empty strings in characters for updating new cast column. Should it be

„aaa,bb,bb,c

or

“,“,aaa,bb,bb,c

or

None,None,aaa,bb,bb,c

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) 5 months ago (Fri Mar 13 2020 14:09:44 GMT+1100 (澳大利亚东部夏令时间))

„aaa,bb,bb,c

Reply



Adith Kumar Sukumar (/users/z5177910) [5 months ago \(Fri Mar 13 2020 13:20:33 GMT+1100 \(澳大利亚东部夏令时间\)\)](#)

For Q13,

I had programmed in such a way that, using `plt.show()` keeps the legend outside. but removing `plt.show()` pushes the legend back into the scatter plot.

Is this fine?

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) [5 months ago \(Fri Mar 13 2020 13:25:45 GMT+1100 \(澳大利亚东部夏令时间\)\)](#)

The file stored will be marked; do not use `plt.show()`

Reply



Abrar Amin (/users/z5018626) [5 months ago \(Fri Mar 13 2020 10:17:37 GMT+1100 \(澳大利亚东部夏令时间\)\)](#)

Can we use import additional modules from matplotlib? (e.g. colors)

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) [5 months ago \(Fri Mar 13 2020 10:25:07 GMT+1100 \(澳大利亚东部夏令时间\)\)](#)

Yes, you can

Reply



Yifan Ai (/users/z5210859) [5 months ago \(Fri Mar 13 2020 06:00:56 GMT+1100 \(澳大利亚东部夏令时间\)\)](#)

Hi,

In Q5, *'there is a **ranking scheme** for movies defined by " (revenue - budget)/budget ". Add a new column for the dataframe, and name it "success_impact" .*

Do we have to **rank** the movies based on their "success_impact" or we just have to add a column?

Thank you in advance!

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) [5 months ago \(Fri Mar 13 2020 06:53:58 GMT+1100 \(澳大利亚东部夏令时间\)\)](#)

Hi, just add the column

Reply



Yifan Ai (/users/z5210859) [5 months ago \(Fri Mar 13 2020 06:56:51 GMT+1100 \(澳大利亚东部夏令时间\)\)](#), last modified [5 months ago \(Fri Mar 13 2020 15:11:58 GMT+1100 \(澳大利亚东部夏令时间\)\)](#)

Thanks

Reply

Kavitha Narayanan (/users/z5190588) [5 months ago \(Fri Mar 13 2020 00:45:18 GMT+1100 \(澳大利亚东部夏令时间\)\)](#)



For Q11 (pie chart), is it enough if I use the 'autopct' to display the percentage values? Or should I write separate code to calculate percentage and display as part of the labels?

Reply



Mohammadali Yaghoubzadehfard (/users/z5138589) 5 months ago (Fri Mar 13 2020 06:54:57 GMT+1100 (澳大利亚东部夏令时间))

It is fine as long as it shows the percentages

Reply



Kavitha Narayanan (/users/z5190588) 5 months ago (Fri Mar 13 2020 09:15:26 GMT+1100 (澳大利亚东部夏令时间))

Thank you

Reply

Load More Comments