MDA-710 Fall 2022

Homework 2

Intro to Programming II & Pandas & Handling Categorical Data

**Due**: September 29th @ 12AM

Submit on Blackboard

Earnable Points: 100pts

Base Points: 5ts

You can miss 5 points and still get a perfect score on this assignment

# Programming II - Classes

Match the numbered parts in the image with the following terms/phrases. Only use each lettered term/phrase once. Each match is worth 2 points.

1. Keyword used to define a class
2. Parameter
3. Defining a Property
4. Used within a class to reference the current object
5. Object
6. Accessing a property
7. Creating an object/instance
8. Constructor
9. Keyword used to define a function
10. A
11. C
12. H
13. D
14. I
15. B
16. G
17. E
18. F



# Pandas

For each question, answer with the code that would perform the given task. Add a screenshot of the output. Each question is worth 3 points except the last question, which is 5 points. (50 total points)

Using pandas, import the dataset “netflix\_titles” as a dataframe and

1. Display the first 6 rows.

Graphical user interface, text

Description automatically generated

1. Display rows 10 through 20

Graphical user interface, text

Description automatically generated

1. Display rows 10 through 20 and columns 1 through 2



1. Display the names of all the columns

Text

Description automatically generated

1. Display the values only for the column ‘title’ for rows 10 through 20

Text

Description automatically generated

1. Display the standard deviation for the column ‘release\_year’

Graphical user interface, text

Description automatically generated

1. Remove the ‘description’ column and show the results



1. Display all the unique values for ‘type’

Graphical user interface, text, website

Description automatically generated

1. Display all the unique values for ‘rating’

Text

Description automatically generated

1. Display all rows for which the type is ‘Movie’

Graphical user interface, text

Description automatically generated

1. Display all row for which the type is ‘Movie’ and the rating is ‘R’

Graphical user interface, text

Description automatically generated

1. Display all rows for which rating does not start with ‘TV’

Graphical user interface, text

Description automatically generated

1. Display all rows for which the rating is missing (this will be all the ratings that have *nan* values)

A screenshot of a computer

Description automatically generated with medium confidence

1. How many rows have missing ratings?

Graphical user interface, text, application

Description automatically generated

1. Add a new column called ‘isMature’ and initialize all the values to False

Graphical user interface, text

Description automatically generated with medium confidence

1. For each row, set the column value for ‘isMature’ to True if the rating is any of the following: 'PG-13', 'TV-MA', 'TV-14', 'R','NC-17'

Graphical user interface, text

Description automatically generated

# MCAR, MNAR, and MAR

(9 points) Notice that one of the values for ‘rating’ is ‘nan’. This means the value is missing. Match each scenario for why the rating is missing with the correct missing data type (each type can only be used once):

1. Movies that are shorter than an hour and are not released in the US or Europe will not have ratings.

MNAR - Missing not at random

1. We have no idea why the ratings are missing

MCAR - Missing completely at random

1. When constructing the database, entering the ratings were optional

MAR - Missing at random

# Handling Categorical Data

For each question, answer with the code that would perform the given task. Add a screenshot of the output.

Using pandas, import the dataset “netflix\_titles” as a dataframe and

1. (5pts) Cr­eate a reasonable mapping for the ‘ratings’ values replace the column values with the mappings

Text

Description automatically generated

Graphical user interface, text

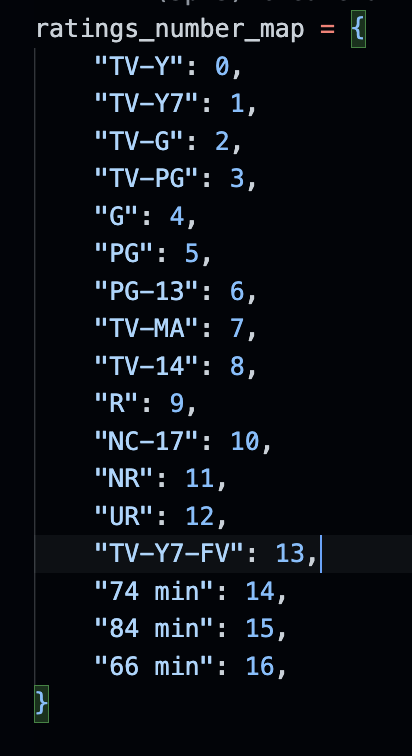
Description automatically generated

1. (5pt) Explain your mapping

I just took a list of all the unique ratings that are present and created a object with serial numbers.

Text

Description automatically generated

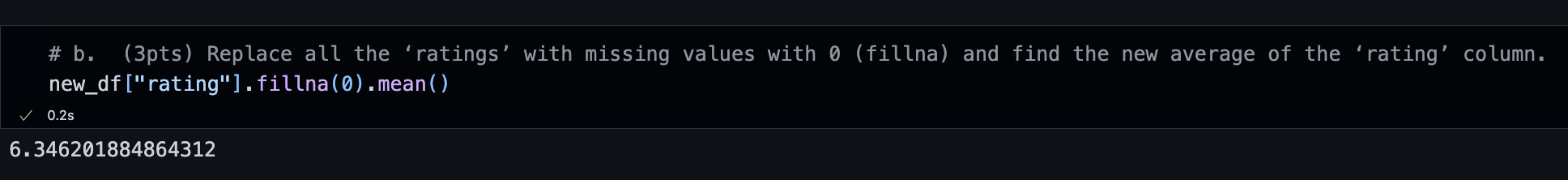


1. After replacing the ‘ratings’ values with your mapping
   1. (3pts) Drop all the ‘ratings’ with missing values (*dropna*) and find the average of the ‘rating’ column.

Graphical user interface, text

Description automatically generated

* 1. (3pts) Replace all the ‘ratings’ with missing values with 0 (*fillna*) and find the new average of the ‘rating’ column.



* 1. (4pts) Replace all the ‘ratings’ with missing values with the average of the ‘rating’ column (*fillna*) and find the new average of the ‘rating’ column

Graphical user interface, text, website

Description automatically generated

1. (3pt) How does the average change based on how you handle it?

**Answer:**

Dropping all the missing value and filling na values with the average gave same output as calculating average using pandas doesn’t consider NA values while doing so.

Filling all na with 0 did reduce the average a bit.