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.

A. Keyword used to define a class
B. Parameter
C. Defining a Property
D. Used within a class to reference the current object
E. Object
F. Accessing a property
G. Creating an object/instance
H. Constructor
I. Keyword used to define a function

1. __
2. __
3. __
4. __
5. __
6. __
7. __
8. __
9. __


```python
import numpy as np

class Person:
    name = None
    age = np.nan
    favorite_color = None

    def __init__(self, name, age, favorite_color=None):
        self.name = name
        self.age = age
        self.favorite_color = favorite_color

    def __str__(self):
        return "Person\t" + self.name + "\t" + str(self.age) + "\t" + self.favorite_color

    def is_older_than(self, other_person):
        return self.age > other_person.age

danielle = Person("Danielle", 30, "pink")
sister = Person("Elise", 18)

print(danielle)
print(danielle.name + " " + ("is" if danielle.is_older_than(sister) else "is not") + " older than " + sister.name)
```

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.
2. Display rows 10 through 20
3. Display rows 10 through 20 and columns 1 through 2
4. Display the names of all the columns
5. Display the values only for the column ‘title’ for rows 10 through 20
6. Display the standard deviation for the column ‘release_year’
7. Remove the ‘description’ column and show the results
8. Display all the unique values for ‘type’
9. Display all the unique values for ‘rating’
10. Display all rows for which the type is ‘Movie’
11. Display all row for which the type is ‘Movie’ and the rating is ‘R’
12. Display all rows for which rating does not start with ‘TV’
13. Display all rows for which the rating is missing (this will be all the ratings that have *nan* values)
14. How many rows have missing ratings?
15. Add a new column called ‘isMature’ and initialize all the values to False
16. 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'

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.
2. We have no idea why the ratings are missing
3. When constructing the database, entering the ratings were optional

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
2. (5pt) Explain your mapping
3. After replacing the ‘ratings’ values with your mapping
    a. (3pts) Drop all the ‘ratings’ with missing values (*dropna*) and find the average of the ‘rating’ column.
    b. (3pts) Replace all the ‘ratings’ with missing values with 0 (*fillna*) and find the new average of the ‘rating’ column.
    c. (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
4. (3pt) How does the average change based on how you handle it?