Numpy, Pandas, Matplotlib, Data Cleaning and Pandas SQL

Assignment 1

Assignment 1

Table of Contents

1.Introduction

This assignment will help you to consolidate the concepts learnt in the session.

2.1. Problem Statement: Numpy

Problem Statement 1:

Write a function so that the columns of the output matrix are powers of the input vector.

The order of the powers is determined by the increasing boolean argument. Specifically, when increasing is False, the i-th output column is the input vector raised element-wise to the power of N - i - 1.

HINT: Such a matrix with a geometric progression in each row is named for Alexandre-Theophile Vandermonde.

Problem Statement 2:

Given a sequence of n values x1, x2, ..., xn and a window size k>0, the k-th moving average of the given sequence is defined as follows:

The moving average sequence has n-k+1 elements as shown below.

The moving averages with k=4 of a ten-value sequence (n=10) is shown below

```
  i    1 2 3 4 5 6 7 8 9 10
===== == == == == == == == == == ==
Input 10 20 30 40 50 60 70 80 90 100
y1     25 = (10+20+30+40)/4
y2     35 = (20+30+40+50)/4
y3     45 = (30+40+50+60)/4
y4     55 = (40+50+60+70)/4
y5     65 = (50+60+70+80)/4
y6     75 = (60+70+80+90)/4
y7     85 = (70+80+90+100)/4
```

Thus, the moving average sequence has n-k+1=10-4+1=7 values.

Question: Write a function to find moving average in an array over a window:
Test it over [3, 5, 7, 2, 8, 10, 11, 65, 72, 81, 99, 100, 150] and window of 3.

2.2. Problem Statement: Pandas

Problem Statement 1:

1) How-to-count-distance-to-the-previous-zero

For each value, count the difference of the distance from the previous zero (or the start of the Series, whichever is closer) and if there are no previous zeros, print the position

Consider a DataFrame df where there is an integer column {'X'[7, 2, 0, 3, 4, 2, 5, 0, 3, 4]}

The values should therefore be [1, 2, 0, 1, 2, 3, 4, 0, 1, 2]. Make this a new column 'Y'.

```
import pandas as pd

df = pd.DataFrame({'X': [7, 2, 0, 3, 4, 2, 5, 0, 3, 4]})
```

2) Create a DatetimeIndex that contains each business day of 2015 and use it to index a Series of random numbers.

3) Find the sum of the values in s for every Wednesday

4) Average For each calendar month

5) For each group of four consecutive calendar months in s, find the date on which the highest value occurred.

Problem Statement 2:

Read the dataset from the below link

https://raw.githubusercontent.com/guipsamora/pandas_exercises/master/06_Stats/US_Baby_Names/US_Baby_Names_right.csv

Questions:

1) Delete unnamed columns

2) Show the distribution of male and female

3) Show the top 5 most preferred names

4) What is the median name occurrence in the dataset

5) Distribution of male and female born count by states

2.3. Problem Statement: Use Cases on Numpy and Pandas

1.) Write a Python program which accepts a list named : randomList = ['a', 0, 2]. Use exception handling using try-catch which gives the output as:

Output:

1) If the List element is a alphabet or string, the output will be

The entry is a
Oops! <class 'ValueError'> occured.
Next entry.

2) If the List element is “0”,the output will be

The entry is 0
Oops! <class 'ZeroDivisionError'> occured.
Next entry.

3) If the List element is and integer except 0,then output will be:

The entry is 2
The reciprocal of 2 is 0.5 // reciprocal of an integer

2) Array out of Bound Exception

Write a Python program to give exception “Array Out of Bound” if the user wants to access the elements beyond the list size (use try and except)

3) Write a python module script that contains fib2() method to calculate the Fibonacci series till 1000 and save it as fibo.py.

Note : The module created as fibo.py has to be placed in lib folder

For linux/ubuntu path = /home/anaconda/lib/python3
For Windows path = C:\Users\Ajit\Anaconda3\Lib

4) Write a python module script that contains ispalindrome() method to calculate the input string as palindrome string or not and save it as palindrome.py.

5) Write a program in Python with one class called Cipher. Within the constructor of this class, ask user for a string and store it. Use a static variable, key to store a randomly generated integer between 1 and 50 inclusive. Implement two methods, encrypt and decrypt within this class. Encrypt generates and prints a cipher text using the user-entered string and the key and ecrypt generates decrypted string from ciphertext. The cipher only encrypts alpha and numeric (A-Z, a-z, 0-9). All Symbols, such as - , ; %, remain unencrypted. The cipher text can have special characters. Use generator expression to filter out alpha and numeric characters of the input string and to generate cipher text. Create an instance of this class, encrypt and decrypt back the user entered string.

6) Get Data from the following link:

http://files.grouplens.org/datasets/movielens/ml-20m.zip

We will be using the following files for this exercise:

ratings.csv : userId,movieId,rating, timestamp
tags.csv : userId,movieId, tag, timestamp
movies.csv : movieId, title, genres

I. Read the dataset using pandas.

II. Extract the first row from tags and print its type.

III. Extract row 0, 11, 2000 from tags DataFrame.

IV. Print index, columns of the DataFrame.

V. Calculate descriptive statistics for the 'ratings' column of the ratings DataFrame. Verify using describe().

VI. Filter out ratings with rating > 5

VII. Find how many null values, missing values are present. Deal with them. Print out how many rows have been modified.

VIII. Filter out movies from the movies DataFrame that are of type 'Animation'.

IX. Find the average rating of movies.

X. Perform an inner join of movies and tags based on movieId.

XI. Print out the 5 movies that belong to the Comedy genre and have rating greater than 4.

XII. Split 'genres' into multiple columns.

XIII. Extract year from title e.g. (1995).

XIV. Select rows based on timestamps later than 2015-02-01.

XV. Sort the tags DataFrame based on timestamp.

2.4. Problem Statement: Matplotlib

Matplotlib:

This assignment is for visualization using matplotlib:

data to use:

url=https://raw.githubusercontent.com/Geoyi/Cleaning-Titanic-Data/master/titanic_original.csv

titanic = pd.read_csv(url)

Charts to plot:

1. Create a pie chart presenting the male/female proportion

2. Create a scatterplot with the Fare paid and the Age, differ the plot color by gender

2.5. Problem Statement: Data Cleaning

It happens all the time: someone gives you data containing malformed strings, Python, lists and missing data. How do you tidy it up so you can get on with the analysis?

Take this monstrosity as the DataFrame to use in the following puzzles:

```
df = pd.DataFrame({'From_To': ['LoNDon_paris', 'MAdrid_miLAN', 'londON_StockhOlm',
                               'Budapest_PaRis', 'Brussels_londOn'],
                   'FlightNumber': [10045, np.nan, 10065, np.nan, 10085],
```

'RecentDelays': [[23, 47], [], [24, 43, 87], [13], [67, 32]],

'Airline': ['KLM(!)', '<Air France> (12)', '(British Airways.)',

'12. Air France', '"Swiss Air"']})

1. Some values in the the FlightNumber column are missing. These numbers are meant to increase by 10 with each row so 10055 and 10075 need to be put in place. Fill in these missing numbers and make the column an integer column (instead of a float column).

2. The From_To column would be better as two separate columns! Split each string on the underscore delimiter _ to give a new temporary DataFrame with the correct values. Assign the correct column names to this temporary DataFrame.

3. Notice how the capitalisation of the city names is all mixed up in this temporary DataFrame. Standardise the strings so that only the first letter is uppercase (e.g. "londON" should become "London".)

4. Delete the From_To column from df and attach the temporary DataFrame from the previous questions.

5. In the RecentDelays column, the values have been entered into the DataFrame as a list. We would like each first value in its own column, each second value in its own column, and so on. If there isn't an Nth value, the value should be NaN.

Expand the Series of lists into a DataFrame named delays, rename the columns delay_1, delay_2, etc. and replace the unwanted RecentDelays column in df with delays.

2.6. Problem Statement: Pandas SQL

Problem statement 1:

Read the following data set:

https://archive.ics.uci.edu/ml/machine-learning-databases/adult/

Rename the columns as per the description from this file:

https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names

Task:

Create a sql db from adult dataset and name it sqladb

1. Select 10 records from the adult sqladb

2. Show me the average hours per week of all men who are working in private sector

3. Show me the frequency table for education, occupation and relationship, separately

4. Are there any people who are married, working in private sector and having a master's degree

5. What is the average, minimum and maximum age group for people working in different sectors

6. Calculate age distribution by country

7. Compute a new column as 'Net-Capital-Gain' from the two columns 'capitalgain' and 'capital-loss'

Problem statement 2:

Read the following data set:

https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

Task:

1. Create an sqlalchemy engine using a sample from the data set

2. Write two basic update queries

3. Write two delete queries

4. Write two filter queries

5. Write two function queries

Note: Solution submitted via github must contain all the detailed steps.

3.Output

This assignment consists of 1500 marks and needs to be submitted in Github. You can follow Github submission guide provided to do the same.