SCMA248 Introduction to Data Science

. Course Description

(In Thai) การแนะนำสู่แง่มุมสำคัญของวิทยาการข้อมูล การดึงข้อมูล และการจัดการข้อมูล การแสดงข้อมูล การคำนวณเชิงสถิติ การเรียนรู้ด้วยเครื่อง การนำเสนอและการสื่อสาร การคำนวณร่วมสมัย สิ่งแวดล้อมด้านฐานข้อมูล เช่น อาร์ และ เอสคิวแอล กรณีศึกษาจากนอกห้องเรียน ทักษะพื้นฐาน สำคัญสำหรับ การเปลี่ยนข้อมูล เป็นความรู้ การฝึกทักษะการสืบค้นข้อมูล เพื่อทำงานกลุ่มและ นำเสนอ ในห้องเรียน

(In English) An introduction to key aspects of data science: data retrieval and manipulation, data visualization, statistical computation and machine learning, presentation and communication; an introduction to contemporary computing and database environments such as R and SQL; case studies from outside the classroom; foundational skills necessary to turn data into information; practicing of information searching skill for working on group assignments and doing presentation in the classroom

Midterm Examination: Part 1

Due date: 15 March 2022 before 10 am.

Points: 15 points for Part 1 (from 21 questions)

Do not alter this file.

Duplicate this file and move it into the Personal folder. Rename the file as Midterm_id (id is your student ID number, e.g. Midterm_Part1_6305001).

Write Python commands to answer each of the questions. For each question that requires numerical values (not list or dataframe), you also need to assign the variable e.g. ans1 to store the numerical answer for question 1. If there is more than 1 answer required, you must create more variables e.g. ans1_1, ans1_2 to store the values of the answers.

When you want to submit your file, you simply share access with me using my email pairote.sat@mahidol.edu and my TA p.pooy.pui.i@gmail.com. Do not move your file into the DS@MathMahidol team.

Data sets

There are 3 data sets that we will work with:

You can download the files from my github page:

- Population dataset:

https://raw.githubusercontent.com/pairote-sat/SCMA248/main/Data/US_population.csv

- US state abbreviations

https://raw.githubusercontent.com/pairote-sat/SCMA248/main/Data/us_state_abbreviations.csv

- Covid-19 dataset:

https://raw.githubusercontent.com/pairote-sat/SCMA470/master/us-counties.csv

```
import pandas as pd
import numpy as np

# https://stackoverflow.com/questions/20625582/how-to-deal-with-settingwithcopywarning-in-panda
pd.options.mode.chained_assignment = None
```

```
from plotnine import *
```

```
ModuleNotFoundError: No module named 'plotnine'
```

Show error details Search on Stack Overflow

1. Enter student IDs of your group.

key aspects of data science: data retrieval and manipulation, data

```
id =
```

Population Dataset: 2020 Census Demographic Data by County.

We begin with the first dataset, which we obtained from

https://www.dataemporium.com/dataset/254/?gclid=CjwKCAiAg6yRBhBNEiwAeVyL0JI9xZg-nt9evBLB04fAZPc-TPTEmrW9kMfoIqMhBJvHjXQ-GV5fPBoChYIQAvD_BwE

This dataset was created from the 2020 Census. It contains one row per county with the total population and a breakdown by race.

```
path = '/Users/Kaemyuijang/SCMA248/Data/US_population.csv'

#df = pd.read_csv(path, parse_dates=True, index_col = 'date')

population = pd.read_csv(path)
```

Detailed description

This table breaks down the total population and population by race for each county in the United States, based on the 2020 Census.

Some Terminology.

- A city is created by any population that has its own system of government and a semblance of a legal system. Cities are located within a county, within a state.
- A county is a geographic unit created for political purposes within a state.

Read more : Difference between city and county
http://www.differencebetween.net/miscellaneous/difference-between-city-and-county/#ixzz7NKcARPB5

This table only includes numbers for people who checked off a single race in the census. This is the majority of people. You can see the number of people who belong to two or more races by subtracting POP_ONE_RACE from TOTAL_POPULATION.

Before proceeding to the next step, make sure you understand the difference between POP_ONE_RACE from TOTAL_POPULATION.

Q1 : Verify that the sum of counts in the columns ['white','black','amarican_indian','asian','hawaiian','other'] is equal to the count in the column 'pop_one_race'.

key aspects of data science: data retrieval and manipulation, data

Q2: How many counties are included in this dataset?

key aspects of data science: data retrieval and manipulation, data

Q3: How many unique county names are there?

key aspects of data science: data retrieval and manipulation, data

Q4: What can be concluded from the difference between the number of counties and the number of unique county names?

key aspects of data science: data retrieval and manipulation, data

Q5: Create a table with the number of counties in each U.S. state.

key aspects of data science: presentation and communication

Q6: The following Python command uses `np.random.seed(id)` to set the seed number of your student ID. Complete the following command to create the variable 'state_given', which stores a randomly selected US state.

```
np.random.seed(id)
states_given = np.random.choice(..., 1).tolist()
print(states_given)
```

Q7: List all counties in the randomly selected US state defined by `states_given`.

key aspects of data science: data retrieval and manipulation, data

Q8: Write Python code to select two random US states and include them in a list named **states_list**. Do not forget to use np.random.seed to set the random value to your ID.

```
np.random.seed(id)
states_list = np.random.choice(..., 1, replace = True).tolist()
```

Q9: List all counties for each state in the **states_list**.

key aspects of data science: data retrieval and manipulation, data

Population by US states

Q10 : Write Python code to create a table including the population for each U.S. state. The table should include total population, pop_one_race, white, black, etc. for each state.

key aspects of data science: statistical computation, presentation and communication
foundational skills necessary to turn data
into information; practicing of information

Q11: From the table of population by race for each US state created above, write Python code to calculate the (row) percentages (for each US state) of the following variables
'white','black','amarican_indian','asian','hawaiian','other'.

```
races_list = ['white','black','amarican_indian','asian','hawaiian','other']
```

Q12: List the first five states with the highest percentage of white Americans.

key aspects of data science: statistical computation
foundational skills necessary to turn data
into information; practicing of information

Q13: List the first five states with the highest percentage of black Americans.

key aspects of data science: statistical computation
foundational skills necessary to turn data
into information; practicing of information

Visualization of U.S. population by race and U.S. states

Q14: Write Python to graph the US population by race broken down by state.

statistical computation, presentation and communication;
foundational skills necessary to turn data
into information; practicing of information

US Covid-19 Dataset

The second dataset can be downloaded from

https://www.kaggle.com/fireballbyedimyrnmom/us-counties-covid-19-dataset.

Each data row contains data on cumulative coronavirus cases and deaths.

The specific data here, are the data **PER US COUNTY**.

We will work with this Covid-19 dataset by preprocessing data, performing statistical data analysis and presenting the results.

```
path = '/Users/Kaemyuijang/SCMA248/Data/us-counties.csv'

#df = pd.read_csv(path, parse_dates=True, index_col = 'date')

df = pd.read_csv(path, parse_dates=['date'], index_col = 'date')
```

Data Cleaning and Preparation: Handling Missing Values

Q15: List all variables (or columns) in this Covid-19 dataset with missing data.

key aspects of data science: data retrieval and manipulation, data

Q16: Calculate the number of missing data for each variable (i.e. in each column).

key aspects of data science: data retrieval and manipulation, data

Filtering out Missing Data

Q17: Write Python code to drop rows only when the column 'fips' has NaN in it. How many rows are there in the resutling DataFrame.

key aspects of data science: data retrieval and manipulation, data

Q18: How many NaN values remain in the **Deaths** column in the resutling DataFrame after deleting only rows where the 'fips' column contains NaN?

key aspects of data science: data retrieval and manipulation, data

Q19: Drop all rows with missing values in the deaths column

key aspects of data science: data retrieval and manipulation, data

Merging multiple DataFrames

The Covid-19 dataset contains only data on cumulative coronavirus cases and deaths. To perform our analysis, e.g., to compare infection rates among different U.S. states, we need to add more information on the population size of each county.

We will need to download another data frame that contains the list of state abbreviations. We will then combine multiple data frames by adding the 'total_population' to the Covid-19 dataset.

- List of State Abbreviations:

https://worldpopulationreview.com/states/state-abbreviations

```
path = '/Users/Kaemyuijang/SCMA248/Data/us_state_abbreviations.csv'

abbr = pd.read_csv(path)
```

Q20: Using the population and us_state_abbreviations datasets, write Python code to add the 'total_population' to the Covid-19 dataset.

key aspects of data science: data retrieval and manipulation, data

Q21: Write Python to check for NaN values after adding 'total_population' to the Covid-19 dataset.

key aspects of data science: data retrieval and manipulation, data