Department of Statistics, The Chinese University of Hong Kong
STAT 5106, Programming Techniques for Data Science (Term 1, 2024-25)

ASSIGNMENT 3
(Deadline: 14 Nov 2024, 2359)

[Note: Although the assignment is about Python script, we still welcome to use R script for handling those non-Datacamp questions. But please make sure the scripts are runnable and give the correct answers.]

For Q1 and Q2, we continue to investigate nycflights13.

New York John F. Kennedy International Airport

Q1. (40%)
From dataset `flights` in `nycflights13`,

a. find all flights that:
   i. Had a departure delay of two or more hours
   ii. Flew to Houston (IAH or HOU)
   iii. Were operated by United, American, or Delta
   iv. Departure in winter (December, January, and February)
   v. Arrived more than two hours late, but didn’t leave late
   vi. Were departure delayed by at least an hour, but made up over 30 minutes in flight
   vii. Arrived between midnight and 6am (inclusive)
b. Sort flights to find the most arrival delayed flights. Find the flights that landed earliest (smallest delay).

Q2. (20%)
There is another dataset `airport` in `nycflights13`,

a. Compute the average delay by destination, then join on the `airports` data frame so you can show the spatial distribution of delays.
(Hint: Using scatter_geo in plotly.express . Reference is here.)

[The following b-d` are analysis-type questions, which are open-ended. Please provide suitable charts / statistics for supporting your answer.]

b. Is there a relationship between the age of a plane and its delays?

c. What weather conditions make it more likely to see a delay?

d. What happened on June 13 2013? Display the spatial pattern of delays, and then use Google to cross-reference with the weather.
(Again using scatter_geo in plotly.express . Reference is here.)

Q3. (20%)
[Paris 2024 Olympic Data Analysis]

The 2024 Summer Olympics was an international multi-sport event held from 26 July to 11 August 2024 in Paris, France, with some preliminary events that began on 24 July. Paris was the host city, with events (mainly football) held in 16 additional cities spread across metropolitan France, including the sailing center in the second-largest city of France, Marseille, on the Mediterranean Sea, as well as one subsite for surfing in Tahiti, French Polynesia.

Data contains the details of over 11,000 athletes, with 47 disciplines, along with 1698 Teams taking part in the 2024 Paris Olympics. This dataset includes the details of the Athletes, Coaches, Teams participating as well as the Entries by gender. It contains their names, countries represented, discipline, gender of competitors, name of the coaches.

In total we have 58 files. But we concentrate 5 datasets in csv format only:

- athletes.csv: Contains details about the participating Athletes
- coaches.csv: Details about coaches, countries and disciplines along with event
- medals.csv: Contains the Medals and Scoreboard of countries that participated in Paris Olympics
- teams.csv: Details about the Teams, discipline, Name of Country and the event

Please find data.zip in the link.
Also you can find out more information at here:
https://www.kaggle.com/datasets/piterfm/paris-2024-olympic-summer-games

Please complete the following tasks with Python code.

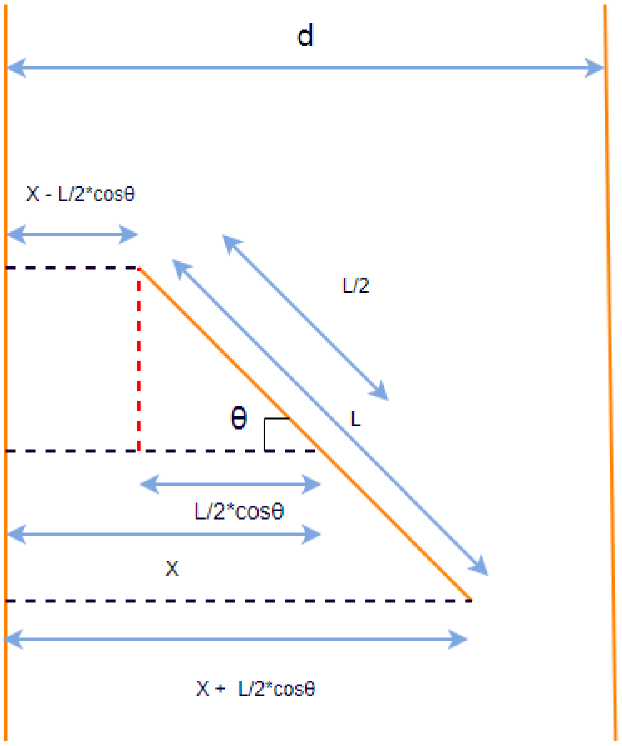
1. Read the 4 csvs to be your pandas data frame into your kernel.
2. In athletes.csv and coaches.xlsx, who are the top 4 coaches leading the most number of Athletes ?
3. In medals.csv, please draw a stacked bar chart, medals no. against countries, having appropriate colors for various medal types, ordering countries by rank in the dataset. Show top 30 countries only.
4. In teams.csv, how many countries where the names are different from the long names of countries? please show results with Python code.

Q4. (20%)
[Buffon’s Needle Problem]

Buffon's needle problem is another way to estimate the value of a with random numbers. The goal in this Monte Carlo estimate of π is to create a ratio that is close to 3.1415926… similar to the example in class - with darts points lying inside/outside a unit circle inside a unit square.

Buffon’s Needle Theorem:
If a short needle, of length l, is dropped on a paper that is ruled with equally spaced lines of distance $d \geq l$, then the probability that the needle comes to lie in a position where it crosses one of the lines is:

$$P(cross\ one\ of\ the\ lines) = \frac{2l}{\pi d}$$



Visual representation.

For simplicity, we set $d = l = 1$.
In this Monte Carlo estimation, you only need to generate two values:

- the distance from left hand bar, $x = [0, 1]$
- the orientation of the needle, $\theta = [0, 2\pi]$

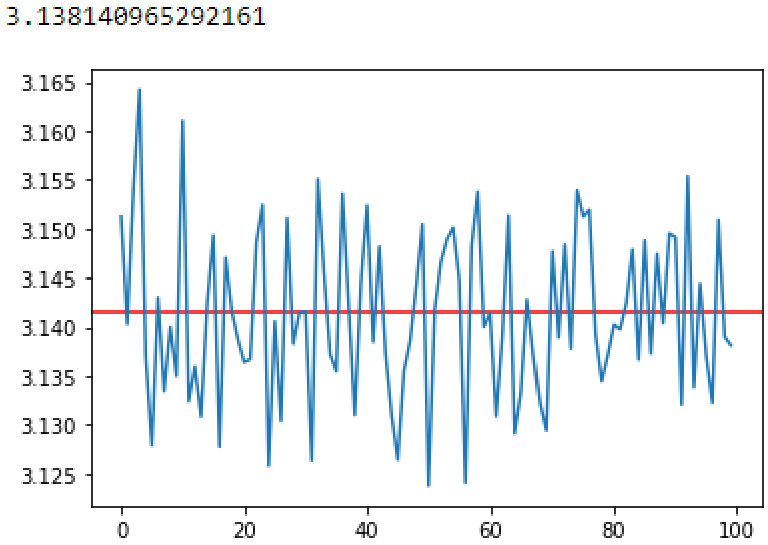
Use python code for the following tasks:

A. Generate 100000 random x and θ values.
B. Calculate the x locations of the 100000 needle ends e.g. $x_{end} = x \pm \frac{\cos\theta}{2}$.

C. Use `np.logical_and` to find the number of needles that have minimum $x_{end\ min} < 0$ and maximum $x_{end\ max} > 0$.

Here $\frac{x_{end\ min} < 0 \ and\ x_{end\ max} > 0}{100000} = \widehat{P}(cross\ one\ of\ the\ lines) = \frac{1}{\pi}$, then we can estimate π.

D. Repeating A, B, C part for generating 100 calculated π.
and plot as the following.
(Hint: using matplotlib, refer to the stack overflow site)



```
3.138140965292161
```