**Lab 4: Data Analysis using Pandas and NumPy**

**Problem 1:** **Warm Up – One-Dimension Random Walk along a line in a 2-D plane**

Imagine a person walking down a road from point ‘A’ to point ‘B’. Ideally one expects that person would walk in a straight line. However, the person walks such that every step she moves forward, she also moves one step to the right or left with equal probability. Write a Python code to simulate this walk and plot it as follows:

1. The user should input the number of steps, and the code should plot the position of the walker v.s. the step number.
2. Run a loop that does the random walk simulation 100 times and for 20 steps for each simulation. Store the position of the walker in a list or an array for each step of each simulation. Plot the average position of the walker for each of the 20 steps taken, along with error bars. As the number of steps taken increases where is the walker most likely to be found at?
3. Redo part(b) for the case with bias, i.e., where the probability of moving one step to the left slightly more – e.g., 5% more – than the probability of moving one step to the right.

**Note:** Put the appropriate label names and legend on the plot. Format and embellish the plot to make it easily understandable.

**Problem 2 – Extracting insights from Tesla Accident data (**source: [Tesla Deaths (Updated 2023) | Kaggle](https://www.kaggle.com/datasets/thedevastator/tesla-accident-fatalities-analysis-and-statistic?resource=download)**)**

1. Load the ‘Tesla Deaths - Deaths.csv’ file into a Pandas dataframe, and print the column headers of the dataframe. You can also refer to https://www.w3schools.com/python/pandas/ for details on syntax.
2. Delete all columns whose headers contain the word “Deceased” and columns that contain weblinks.
3. Using Matplotlib, plot
   1. A histogram of Number of deaths v.s. Year.
   2. A Bar graph of deaths by country
   3. A bar graph of deaths by state in the USA
   4. A pie chart of distribution of fatalities of Tesla driver, occupant, other vehicle, cyclists/peds
4. From the description column, determine a way to describe the cause of highest accidents using commonly occurring words.
5. Determine the number of cyclist related fatalities by country. Does this give a measure of which country is most unsafe for cyclists?
6. What fraction of Autopilot claimed accidents are actually verified? Is there a correlation between autopilot related accidents to any of the other factors based on the provided data?
7. Determine the ‘model’-wise distribution of accidents. Which model results in minimal damage to the driver?
8. Using the ‘Tesla Deaths – Sudden Acceleration.csv’ find the distribution of accidents by model and by State (note, the locations are given as “city, state”. You have to extract state information). Do you find any difference in the distribution of accidents attributed to different ‘models’ for sudden acceleration and accidents?
9. What are the 10 most used words in the ‘Details’ column of Sudden Acceleration cases. Represent them in a plot ranked in the order of citation, giving information on the number of citations?

**Bonus**: Provide your own insights into the data using knowledge from the various TLP courses and plot these insights.