National University of Computer and Emerging Sciences Islamabad

Department of AI and Data Science

DS5003-Stat. and Math. Methods for Data Science (Fall 2022)

Assignment 1

Total Marks: 60

Due Date: Wednesday, 21st September, 2022 [3:00 pm]

Please read the following instructions

i. All coding assignments are to be implemented in Python
ii. You are required to submit Jupyter notebooks with the solution of each of the tasks given below. You can use Google Colab to create notebooks and submit the link to your implementation. However, you will be required to submit the notebook file (.ipynb) as well.
iii. Please provide sufficient description of your solution using the text blocks.
iv. Label your code blocks clearly describing your implementation.
v. For all questions requiring a theoretical answer, use the text block option in the notebook.
vi. There will be NO extension in deadline

Task-1 [40 marks]:

Descriptive Statistics

Attached herewith is the Boston crimes dataset file (filename: crime.csv). You are required to analyze the data using descriptive statistics in Python. Read the data in a Pandas dataframe and perform the required analysis in Python to answer the following questions. The answers together with the code for each of the questions should be included in the submitted notebook.

i. Present an overall summary of the data. Describe what it is about e.g., total number of data points, how many types of crimes, number of districts, number of years etc. [4]
ii. In which year were the highest number of crimes reported? [2]

iii. Overall, which district had the highest crime rate? [2]

iv. Which district had the highest number of crimes in years 2018, 2016 and 2015? [3]

v. Show using a grouped bar chart how many different types of crimes were reported every year (see figure-1). [5]

vi. Does any district have susceptibility towards a particular crime? Provide statistical evidence for your answer. [5]

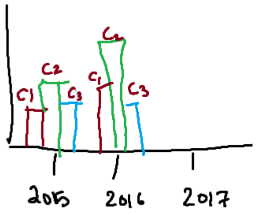

Figure 1- Sample bar chart

vii. Which month are the most and the least number of crimes reported? [2]

viii. Which day of the week are the most and the least number of crimes reported? [2]

ix. Which time of the day (morning, afternoon, evening, night) are the most crimes reported? [5]

x. Which district would you suggest to your friend for moving based on your analysis? Explain with statistics. [2+3]

xi. Mention any interesting insights you have regarding the data. Provide evidence using descriptive statistics. [5]

Task-2[10+10 marks]:

<u>Conditional Probability</u>

This task is related to the Monty Hall Problem.

The *Monty Hall problem* is a famous conundrum in probability which takes the form of a hypothetical game show. The contestant is presented with three doors; behind one is a car and behind each of the other two is a goat. The contestant picks a door and then the gameshow host opens a different door to reveal a goat. The host knows which door conceals the car. The contestant is then invited to switch to the other closed door or stick with their initial choice.

i. Suppose you have *n* doors with a car behind one of those and goats behind the rest. The rest of the conditions stay the same as in the original problem. What would be the increase/decrease in your probability of winning depending on whether you stay with the original choice or switch to one of the other doors after Monty opens a door with a goat behind it? Prove your answer using conditional probability theorems. Upload the scan of your handwritten solution to the Google colab notebook for your solution.

ii. Implement a Python function to simulate a generalized trial of the game. The function should take as input the number of doors, and a binary variable to denote the whether the initial choice has been switched or not. It should return whether you win the car or not. The number of cars would never exceed one. Run the code for 10, 100, 500, 1000, 2000, 4000, 8000, 10000 times with choosing to switch the door and setting the number of doors to 150. Report the results for each of the runs and comment on whether the winning probability converge to your estimated probability in the previous section or not.