Q 1: Homework Consolidation 200 pts

Write a program to consolidate ALL your homeworks done to date (leave out the LinkedIN one).

In essence, you should have 10 homeworks. The user should be able to enter the homework number, and the homework runs as it is expected to. Once the homework run is complete, we should come back to the original menu, where the user can select another homework (or type 0 to exit).

Note: Do Error Handling in menu inputs, e.g. not allowing any input other than 0-10.

Sample Menu

```
1: BMI

2: While Question String

3: FindRemainder Function

4: Lists with 0

5: First Second File to Third

6: Dictionary List

7: Pandas Basics

8: Stats & Pandas

9: Data Visualization

10: Address Class

    Please Enter Homework Number (and 0 to exit): _
```

2: Train Dataset 250 pts

You have to devise an Object Oriented solution for this problem, i.e. you MUST create a CLASS, and solve the problem calling the class methods.

The train dataset contains data regarding a train crash. It has ~891 records. However, all records are not complete, i.e. some have missing data. In this data set, our outcome of interest is the survived column, on whether an individual survived (1) or died (0) after the Train crash.

1. Clean the data and remove any unwanted records. How many records do you have now? (20 points)
2. The train picked most passengers from which station ? (20 points)
3. Do some basic data exploration (e.g. using commands as head( ), info( ), describe( ), nunique( ), etc). Which variables will you NOT select? (20 points)
4. Are there any outliers in the data? If yes, treat them. (15 points)
5. Partition the data into a training set (with 70% of the observations), and testing set (with 30% of the observations) using the random state of 12345 for cross validation. (15 points)
6. On the partitioned data, build the best KNN model. Show the accuracy numbers. (Hint: What is the best value of k? How do you decide the 'best k'?) (30 points)
7. On the partitioned data, build the best logistic regression model. Show the accuracy numbers. (30 points)
8. On the partitioned data, build the decision tree. Show the accuracy numbers. What tree depth did you choose, i.e. which one is ideal and why? (30 points)
9. Based on the results of k-nearest neighbor, and logistic regression, what is the best model to classify the data? Provide explanation to support your argument. (20 points)
10. Show some interesting graphs of the data, i.e., that can describe the original data. (50 points)

Requirements

- Make all the "best programming practice" decisions, e.g. how to show the output, what prompts to display, how to ask for input etc.
- You are the developer/engineer, it is YOUR decision ... YOUR job. If the program is not presented nicely, you will lose points.
- First, all that is being asked should be done. Second, the displays should be intuitive, self-explanatory and nicely put.
  Do NOT assume that the user knows ANYTHING.
- Write the program such that any new person that sits in front of the terminal, can start playing with it (i.e., the commands, displays etc. are adequate).

Submission and Demo

MULTIPLE file submissions are allowed. Anything you submit is kept.

You will submit the following:

- .ipynb files for all questions, clearly labeled. (Q1_HW.ipynb and Q2_Train.ipynb)
- You are asked to demo the project IN-class or Zoom. Details in D2L/Email.
- The Demo files (i.e. files you submit) should NOT be commented.
