Selected Topics (Data Science and Big Data) Tasks

General Guidelines:

- This is individual task not team task
- Each one will has 1 algorithm to implement and 1 dataset to use

- There are many tasks, everyone will get his task depends on his ID. Your task number is the reminder from (your last digit in your ID modulus 8 (number of tasks)
- Example:
  If your ID is: 20191834, then you will get the last digit (4) modulus 8,
  4 % 8 = 4
  So your task number is 4

- Anyone who will choose any task different from what he supposed to take, will get Zero
- 4 degree on this task

Tasks Main Points:

1. You need to check the dataset if it needs a clean or not. Show me how to check for missing values through code.
2. If dataset needs a clean then clean it
3. Use dataset for training your model
4. After predicting, Display me how many rows are predicted wrong (In case of classification)
5. Choose the best number of clusters through code (In case of Clustering)
6. Visualize the results

8 Tasks

Note: each task has a number. Get task by this number After making calculations mentioned above

1- Simple Linear Regression => dataset: Ground Water Survey
Link for dataset:
https://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/slr/frames/slr09.html#

2- K-Means => dataset: Cambridge Crime Data
Link for dataset:
https://data.world/data-society/cambridge-crime-data-2009-2016
Use username: my-greate-username and password: Aa_12345678

3- Decision Tree (Classification) => dataset: breast-cancer-wisconsin
Link for dataset:
https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

4- Naïve Bayes => dataset: bank-marketing-data
Link for dataset:
https://data.world/data-society/bank-marketing-data
Use username: my-greate-username and password: Aa_12345678

5- Simple Linear Regression => dataset: Fire and Theft in Chicago
Link for dataset:
https://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/slr/frames/slr05.html#

6- Decision Tree (Classification) => dataset: Pima Indians Diabetes
Link for dataset:
https://github.com/npradaschnor/Pima-Indians-Diabetes-Dataset

7- K-Means => dataset: Airplane Crashes
Link for dataset:
https://data.world/data-society/airplane-crashes
Note: Establish the dangerousness in terms of aviation accidents
Use username: my-greate-username and password: Aa_12345678

8- Naïve Bayes => dataset: Glass Identification Data Set
Link for dataset:
https://archive.ics.uci.edu/ml/datasets/glass+identification

Best Wishes