Week 7 Homework

1. In this question, use the Boston house price market dataset in sklearn to proceed an EDA. The information about the dataset is in https://scikit-learn.org/stable/datasets/index.html#boston-house-prices-dataset

   You will need to start your Python file with

   ```
   from sklearn.datasets import load_boston
   import pandas as pd

   boston = load_boston()
   boston_df = pd.DataFrame(boston.data)
   boston_df.columns =
   ['CRIM','ZN','INDUS','CHAS','NOX','RM','AGE','DIS','RAD',
   'TAX','PTRATIO','B','LSTAT']
   ```

   In the following, explain

   a. What does each column mean?
   b. What are the range of values can be taken in each column?
   c. What would be the method to take each of these column data, and also the method to aggregate the columns together?
   d. Hence, what could be the error sources occurred in this dataset? (You do not have to do an EDA yet)
   e. What could be the value if someone has this dataset to analyse?

2. What are the assumptions of data to use
   a. one sample t-test?
   b. proportion test?
   c. two-sample t-test?
   d. ANOVA?
   e. Wilcoxon test?

3. In this week, we looked at exploratory data analysis (EDA). In this question, use the Boston house price market dataset in sklearn to proceed an EDA.
   a. What are the fields in the dataset?
   b. What is the size of the dataset?

c. Is there any missing data?
d. Are there any duplicated entries in the dataset?
e. Report the statistical summary of the dataset.