CS37300 Homework	— Homework 5—	Last update: December 9, 2025
Due: Dec 8, 2025 (11:59 pm)		

Name: _____ PUID: _____

Instructions and Policy: Each student should write up their own solutions independently. You need to indicate the names of the people you discussed a problem with; ideally you should discuss with no more than two other people.

- YOU MUST INCLUDE YOUR NAME IN THE HOMEWORK
- The answers (without the python scripts) MUST be in submitted via Gradescope.
- Please write clearly and concisely - clarity and brevity will be rewarded. Refer to known facts as necessary.
- Theoretical questions MUST include the intermediate steps to the final answer.
- Zero points in any question where the python code answer doesn't match the answer on Gradescope.
- If the answer is a plot, it should be added to the PDF and, in the code, it should always be saved as an file (image or PDF), and not using plt.show().

Your code is REQUIRED to run on Python 3 at scholar.rcac.purdue.edu. The TA's will help you with the use of the scholar cluster. If the name of the executable is incorrect, it won't be graded. Please make sure you didn't use any library/source explicitly forbidden to use. If such library/source code is used, you will get 0 pt for the coding part of the assignment. If your code doesn't run on scholar.rcac.purdue.edu, then even if it compiles in another computer, your code will still be considered not-running and the respective part of the assignment will receive 0 pt.

Q0 (0 pts pts): A correct answer to the following questions is worth 0pts. An incorrect answer can result in an F grade in the course.

(1) Student interaction with other students / individuals:

   (a) I have copied part of my homework from another student or another person (plagiarism).

   (b) Yes, I discussed the homework with another person but came up with my own answers. Their name(s) is (are) _____

   (c) No, I did not discuss the homework with anyone

(2) On using online resources:

   (a) I have copied one of my answers directly from a website (plagiarism).

   (b) I have used online resources to help me answer this question, but I came up with my own answers (you are allowed to use online resources as long as the answer is your own). Here is a list of the websites I have used in this homework:

   _____

   (c) I have not used any online resources except the ones provided in the course website.

1 Theoretical Questions (10 + 30 = 40 pts)

Please submit your answers on Gradescope.

Q1 (10 pts): True or False Questions

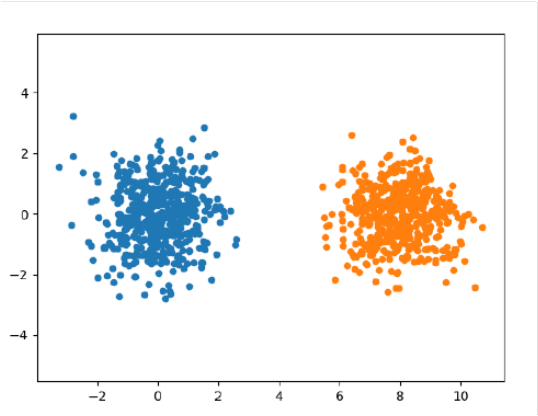
Answer the following as True or False with *a justification or an example*. Points are uniformly distributed wihtin the questions.

1. (True or False) For each question below, answer True or False and justify your answer.

   (a) **(5 pts)** k-means clustering is guaranteed to always converge to the optimal cluster assignment (smallest score), for a 2-dimensional dataset when using the euclidean distance.

   False, K-means is not guaranteed to always converge to the optimal cluster assignment. The convergence depend on the initial condition of the centroids and is possible to get stuck there.

   (b) **(5 pts)** The k-means score is minimum at k = 2 for the following dataset:



   True, the data shows two well separated clusters. This implies that the k-score should be minimized at k=2. For k greater than 2 it would yield only small improvements, so the curve flattens.

Q2 (30 pts): Clustering Algorithms

(a) (5 pts) Consider the following figure of 2D points. There are two true clusters (k = 2) depicted as yellow and blue. Which of the following clustering method(s) is (are) most likely to produce the true yellow (cluster 1) and blue (cluster 2) clustering? Choose the most likely methods (or method) and explain why they (it) will work better than others briefly. Use several sentences to explain your answer. Note that the answer may include more than one clustering method.

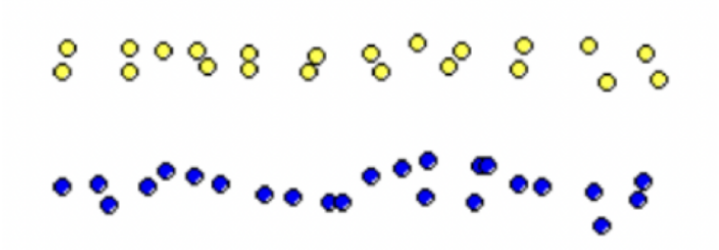
i. Hierarchical clustering with a single link (nearest neighbor) as a distance measure between clusters.

ii. Hierarchical clustering with complete link (furthest neighbor) as a distance measure between clusters.

iii. Hierarchical clustering with average link as a distance measure between clusters.

iv. K-means.

v. GMM (Gaussian mixture model with learnable means, covariance matrices, and mixture coefficients).



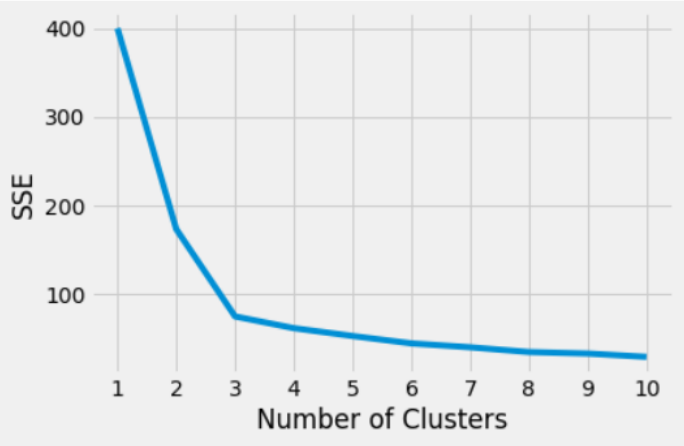
The methods most likely to recover the yellow and blue horizantal clusters are (2) and (3). Unlike k-means which assumes roughly spherical clusters and can cut them vertically, these linkage criteria respect the long band structure and maintain separation across the full horizontal extent.

(b) (15 pts) Finding the right k number in k-means clustering is important and fundamental. There are two popular methods to determine the optimal k number:

i. (5 pts) The elbow method:

- Step 1: Run the algorithm for different choices of k and record the Sum of Squared Errors (SSE).
- Step 2: Plot the k-SSE curve, then choose the k value at which an increase in k will cause a very small decrease in the error sum, while a decrease will sharply increase the error sum (the elbow of the curve).

Based on the definition, choose an appropriate k value in the following plot. Justify your answer.



To choose an opimal k, we need to look for the "elbow" in the graph. This is where the SSE stars becoming small. In this particular case the optimal k is 3.

ii. (5 pts) The silhouette coefficient: The silhouette coefficient quantifies how well a data point fits into its assigned cluster based on the following criteria: (C1) how far away the data point is from points in other clusters, and (C2) how close the data point is to other points in the cluster To compute the silhouette coefficient for data point i, we first compute the following two values:

$$a(i) = \frac{1}{|C_{c_i}| - 1} \sum_{j \in C_{c_i}, i \neq j} d(i,j),$$

$$b(i) = \min_{c_j \neq c_i} \frac{1}{|C_{c_j}|} \sum_{j \in C_{c_j}} d(i,j),$$

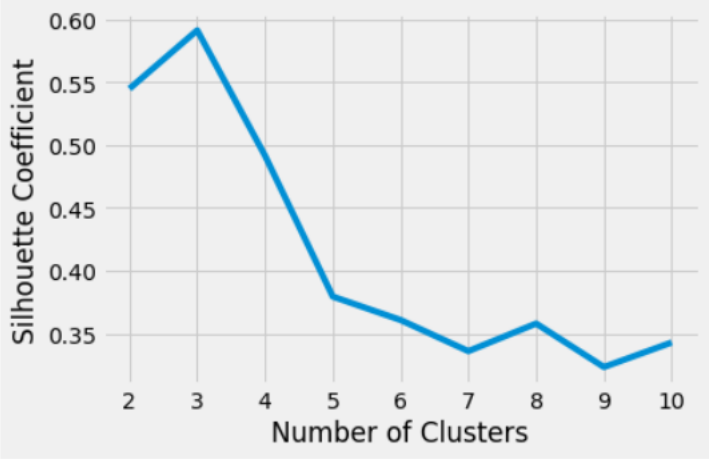
where i and j are data points, c_i is the cluster assigned to data point i, C_m is the set of points that belong to cluster $m \in 1, \ldots, k$. Then, the average silhouette coefficient (whose value ranges between -1 and 1) is given by:

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{b(i) - a(i)}{\max(a(i), b(i))}.$$

Larger numbers indicate that the data points are closer to points in their own clusters than to points in other clusters. The value of S is then used to choose the number of clusters by plotting it against different choices of k, the number of clusters, and choosing the k with the highest score. Is a(i) computing a quantity related to criteria (C1) or (C2)? Is b(i) computing a quantity related to criteria (C1) or (C2)? Justify your answer.

a(i) is related to C2 because it averages the distances from point i to all other points in its own cluster, measuring how close it is to the assigned cluster. b(i) is related to C1 because it takes over all the other clusters, the minimum average distance from i to those clusters, hence quantifying how far i is from points in other clusters.

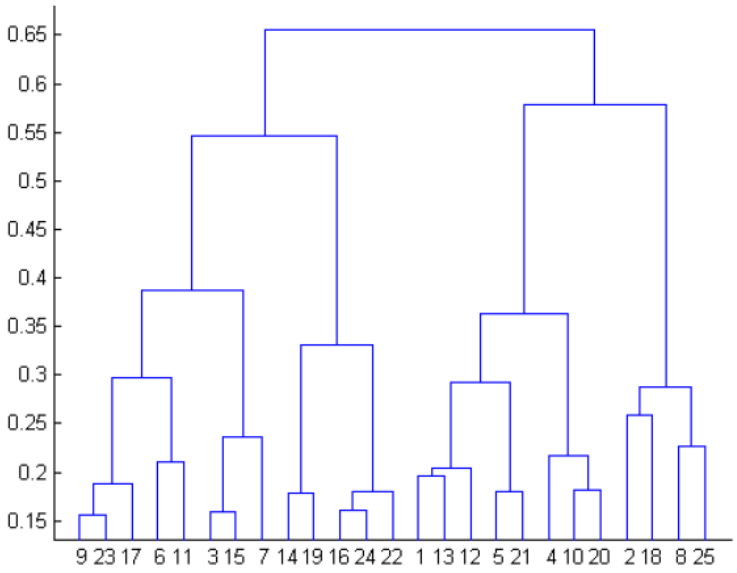
iii. (5 pts) For the following plot of the silhouette coefficient method, which is the best value of k according to the silhouette coefficient criteria?



The silhoutte coefficient measures quality of clustering where a higher score implies better defined clusters. So in this particular case the highest coefficient is k = 3.

(c) (10 pts) The dendrogram of a hierarchical clustering method shows the data points in the x-axis with the y-axis showing the average (or single or complete) link distance threshold when a given cluster splits into two clusters (going from the top to bottom). One way to choose the optimal number of clusters with a dendrogram is to find the y-axis region with maximum distance (vertically) which does not include a new branching (split). Then the number of clusters can be determined by the number of branches that intercept a horizontal line within that region.

i. (5 pts) Based on the above criteria, find the appropriate number of clusters (and the region with maximum distance (vertically) which does not include a new branching (split)) in the following dendrogram (you may directly draw on the figure below for your answer).



ii. (5 pts) Explain why the above method is a reasonable way to choose the number of clusters. (Hint: use your y-axis region to justify your choice).

This method looks for the largest vertical gap on the y-axis with no merges, which is where the biggest jump in distance when clusters are merged. Cutting the dendogram in that region means we stop joining right before dissimilar groups are forced to be grouped together, so points within each cluster are somewhat close, small linkage distance, while different clusters are well separated, large linkage distance.

2 Programming Part (60 Pts)

The objective of this programming task is to utilize k-means clustering to analyze the California Housing Prices dataset, segmenting the housing data based on various features. This dataset is a comprehensive resource for predicting housing prices in California, encompassing o distinct features as well as the class label, which is the median house value, as detailed in Table 1. These features offer valuable insights into multiple aspects of housing, including geographic location, housing age, number of rooms and bedrooms, population, and the number of households in the area. Additionally, it covers economic factors such as median income. The dataset's rich geographic and demographic diversity makes it ideal for clustering analysis, helping to uncover patterns and trends in the California housing market.

In this programming assignment, no skeleton code is provided, and we are not seeking exact responses. Instead, the emphasis is on assessing your analytical skills while navigating the dataset. Your code will be tested using the following command: `python yourusername-kmean.py housing.csv outputfolder`. You should store your results (execution log) and analytical figures in the specified output folder.

Variable	**Description**
longitude	Longitude of the house location
latitude	Latitude of the house location
housing_median_age	Median age of the house
total_rooms	Total number of rooms in the house
total_bedrooms	Total number of bedrooms in the house
population	Population in the house block
households	Number of households in the house block
median_income	Median income of the household
median_house_value	Median house value
ocean_proximity	Proximity to the ocean

Table 1: House Data Description

Part 1. Data Exploration and Processing (10 pts)

In the first part, concentrate on importing the dataset into your Python environment with the pandas package and obtaining some preliminary insights into the data.

1. **(5 pts)** First, perform basic exploratory data analysis to understand the structure and characteristics of the data. For each of the 9 feature, report the summary statistics, data types, and any missing values.

```
\          longitude     latitude  housing_median_age   total_rooms  total_bedrooms    population    households  median_income  median_house_value ocean_proximity
count   20640.000000  20640.000000        20640.000000  20640.000000    20433.000000  20640.000000  20640.000000   20640.000000        20640.000000           20640
unique           NaN           NaN                 NaN           NaN             NaN           NaN           NaN            NaN                 NaN               5
top              NaN           NaN                 NaN           NaN             NaN           NaN           NaN            NaN                 NaN       <1H OCEAN
freq             NaN           NaN                 NaN           NaN             NaN           NaN           NaN            NaN                 NaN            9136
mean     -119.569704     35.631861           28.639486   2635.763081      537.870553   1425.476744    499.539680       3.870671       206855.816909             NaN
std         2.003532      2.135952           12.585558   2181.615252      421.385070   1132.462122    382.329753       1.899822       115395.615874             NaN
min      -124.350000     32.540000            1.000000      2.000000        1.000000      3.000000      1.000000       0.499900        14999.000000             NaN
25%      -121.800000     33.930000           18.000000   1447.750000      296.000000    787.000000    280.000000       2.563400       119600.000000             NaN
50%      -118.490000     34.260000           29.000000   2127.000000      435.000000   1166.000000    409.000000       3.534800       179700.000000             NaN
75%      -118.010000     37.710000           37.000000   3148.000000      647.000000   1725.000000    605.000000       4.743250       264725.000000             NaN
max      -114.310000     41.950000           52.000000  39320.000000     6445.000000  35682.000000   6082.000000      15.000100       500001.000000             NaN
longitude             float64
latitude              float64
housing_median_age      int64
total_rooms             int64
total_bedrooms        float64
population              int64
households              int64
median_income         float64
median_house_value      int64
ocean_proximity        object
dtype: object
longitude               0
latitude                0
housing_median_age      0
total_rooms             0
total_bedrooms        207
population              0
households              0
median_income           0
median_house_value      0
ocean_proximity         0
dtype: int64
```

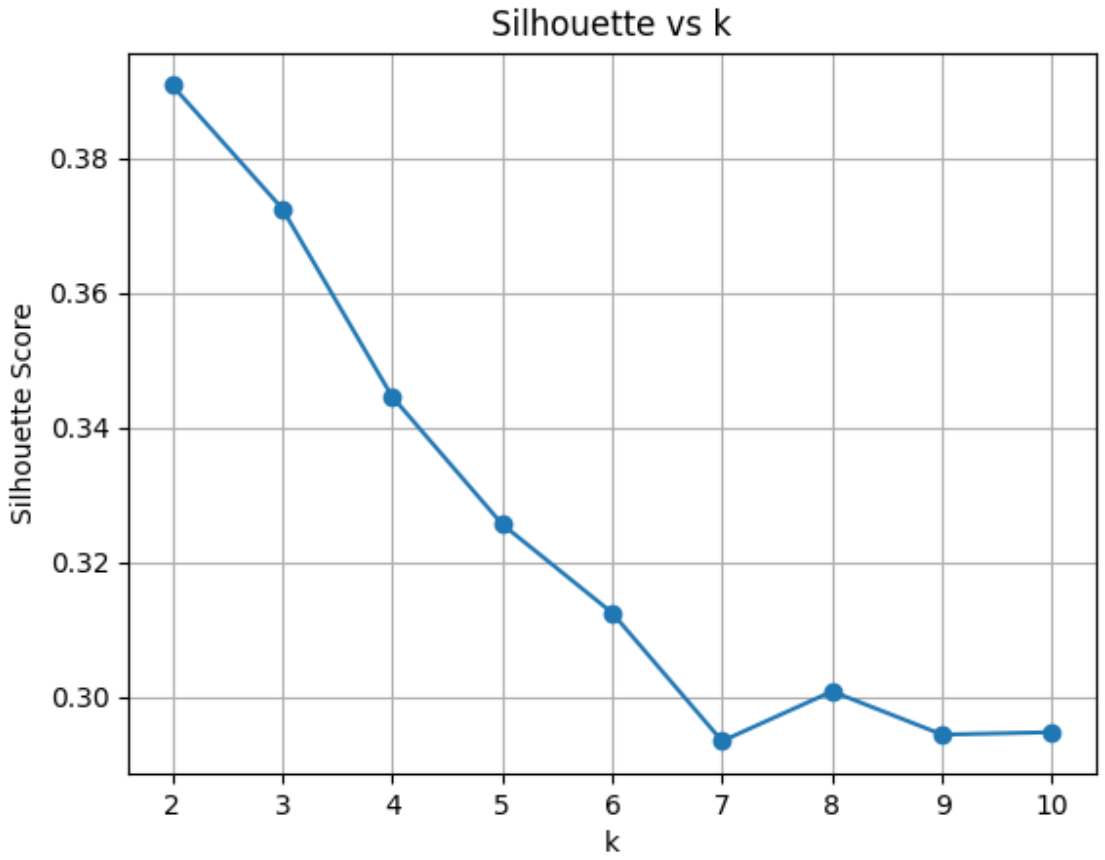
2. **(5 pts)** Next, use pandas to handle any missing values and normalize the features to prepare the data for clustering. Provide a brief summary of the data modifications.

```
          longitude      latitude  housing_median_age   total_rooms  total_bedrooms    population    households  median_income  median_house_value  ocean_proximity
count  20640.000000  20640.000000        20640.000000  20640.000000    20640.000000  20640.000000  20640.000000   20640.000000        20640.000000     20640.000000
mean    -119.569704     35.631861           28.639486   2635.763081      537.870553   1425.476744    499.539680       3.870671       206855.816909         1.165843
std        2.003532      2.135952           12.585558   2181.615252      419.266592   1132.462122    382.329753       1.899822       115395.615874         1.420662
min     -124.350000     32.540000            1.000000      2.000000        1.000000      3.000000      1.000000       0.499900        14999.000000         0.000000
25%     -121.800000     33.930000           18.000000   1447.750000      297.000000    787.000000    280.000000       2.563400       119600.000000         0.000000
50%     -118.490000     34.260000           29.000000   2127.000000      438.000000   1166.000000    409.000000       3.534800       179700.000000         1.000000
75%     -118.010000     37.710000           37.000000   3148.000000      643.250000   1725.000000    605.000000       4.743250       264725.000000         1.000000
max     -114.310000     41.950000           52.000000  39320.000000     6445.000000  35682.000000   6082.000000      15.000100       500001.000000         4.000000
longitude           float64
latitude            float64
housing_median_age  float64
total_rooms         float64
total_bedrooms      float64
population          float64
households          float64
median_income       float64
median_house_value  float64
ocean_proximity     float64
dtype: object
```

Part 2. Implement and Apply K-Means Clustering (35 pts)

1. **(15 pts)** Implement the k-means clustering algorithm and apply it to the dataset. Using the existing implementations (e.g. from the sklearn library) is not allowed.

2. **(10 pts)** Use your implementation to conduct Silhouette analysis to determine the optimal number of clusters. Plot the silhouette scores for various cluster counts and identify the best value(s) if the algorithm can pinpoint them.

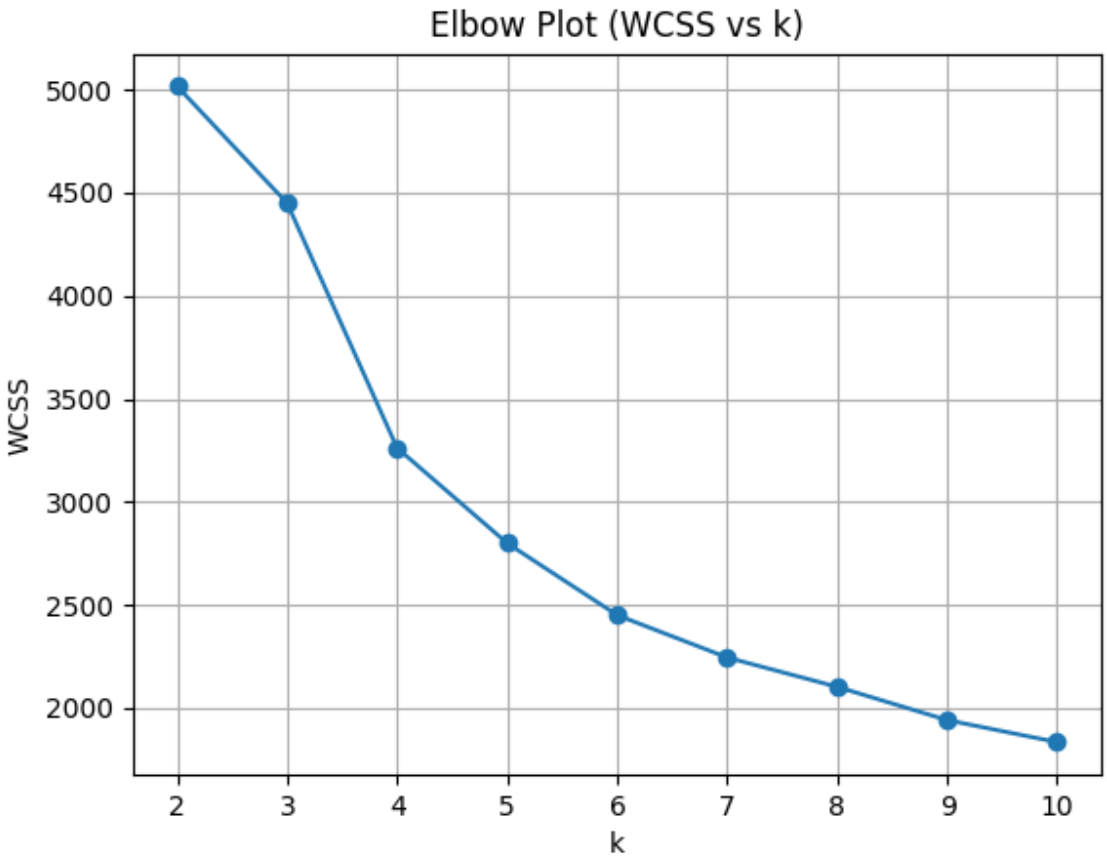


3. **(10 pts)** Use the Elbow method to find the optimal number of clusters by plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the "elbow" point. If the elbow point is unclear, use the `kneed` Python package to identify it programmatically: `KneeLocator(k-range, wcss, curve=''convex'', direction=''decreasing'')`. Compare this result with the Silhouette analysis. If there's a discrepancy, explain why.

Part 3: Analyze Cluster Characteristics (15 pts)

1. **(5 pts)** Apply k-means with the optimal numbers of clusters identified by these two methods on the data (if either or both algorithms provide an efficient value). If your analysis does not yield an optimal (k), document the (k) value you selected.

2. **(10 pts)** Examine the characteristics of each cluster by analyzing the mean and standard deviation of the features within each cluster as well as the class label. Describe the typical profile for each cluster.




```
Best k according to Silhouette: 2
Elbow k according to KneeLocator: 5
2 for final clustering.
```

```
Cluster sizes:
cluster
0    15688
1     4952
Name: count, dtype: int64

Cluster means (numeric columns only):
         longitude   latitude  housing_median_age  total_rooms  total_bedrooms   population  households  median_income  median_house_value  ocean_proximity  cluster
cluster
0       -119.217395  35.467177           27.188233  2665.553034      541.190061  1466.250637  500.895589       3.803942       191949.111104         0.417708      0.0
1       -120.685826  36.153584           33.237076  2541.388126      527.354307  1296.304523  495.244144       4.082068       254080.453554         3.535945      1.0

Cluster standard deviations (numeric columns only):
         longitude  latitude  housing_median_age  total_rooms  total_bedrooms   population  households  median_income  median_house_value  ocean_proximity  cluster
cluster
0         1.780333  2.066385           12.056752  2257.654445      433.458527  1180.381732  392.755606       1.857123       108800.284560         0.493326      0.0
1         2.247576  2.265240           13.111085  1918.110667      370.581208   953.735555  347.242373       2.014824       122756.396749         0.500374      0.0
```

General Instructions

For this assignment, please download the California housing dataset from Bright Space. Note that you cannot use sklearn for your k-means clustering implementation. You should upload a single Python file for the programming part and store the artifacts of each part with meaningful names in the output folder.

3 Submission Instructions

Upload your PDF report, containing the answers of both the theoretical questions and any question from the programming part that instructs so, must be submitted in Gradescope. When submitting your PDF, you have to use Gradescope's interface to assign each question to its corresponding page.

Gradescope Help - Submitting an Assignment:

`https://help.gradescope.com/article/ccbpppziu9-student-submit-work`