Table of Contents


```
clear all; close all; clc;
```

Part 1

```
A = readtable('housing.csv', 'NumHeaderLines', 1);
x1 = A.Var1;
x2 = A.Var2;
X = [x1, x2];

figure();
hist3(X, [100, 100], 'CDataMode', 'auto',...
    'FaceColor', 'interp', 'EdgeAlpha', 0.25);
title('Housing Data Density vs. (x_1, x_2)');
xlabel('x_1 = Longitude');
ylabel('x_2 = Latitude');
```

Part 2

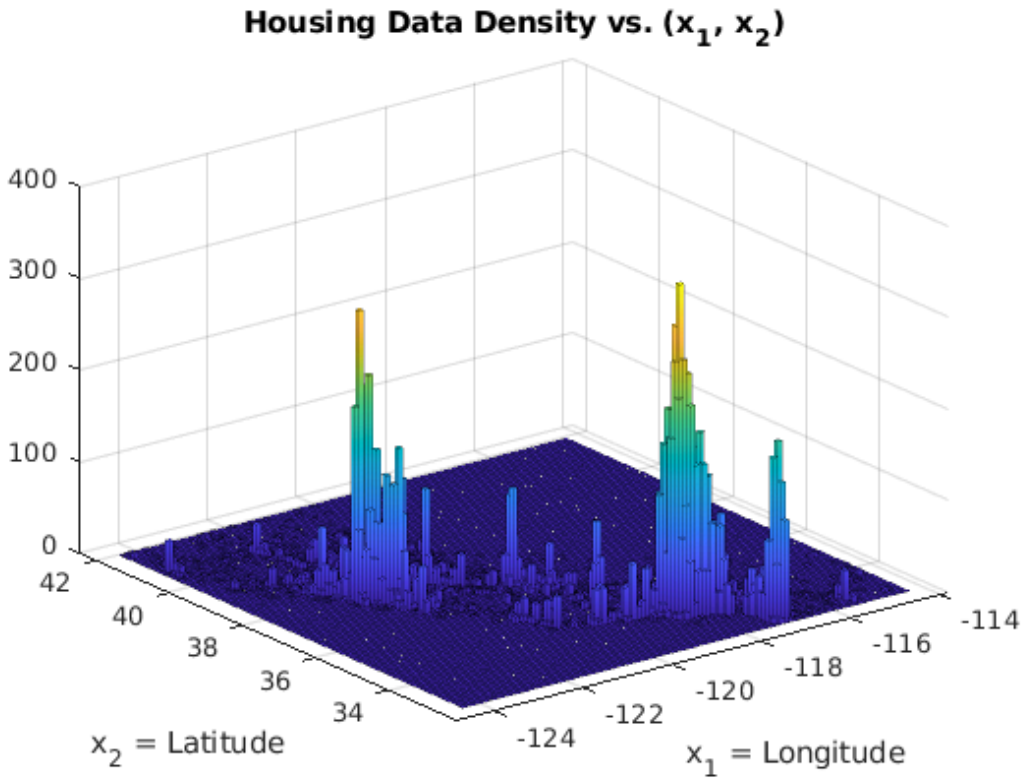
```
k = 7;
replicates = 5;
figure();

lgd_clust = cell(replicates,1);
for rep = 1:replicates
    subplot(2,replicates,rep);
    [idx, cent] = kmeans(X, k);
    gscatter(X(:,1), X(:,2), idx, [], 'o');
    hold on;
    plot(cent(:,1),cent(:,2),'kx','LineWidth', 2, 'MarkerSize', 10);
    for clust = 1:k
        lgd_clust{clust} = sprintf('C%d = (%0.2f, %0.2f)',...
            clust, cent(clust,1), cent(clust,2));
    end
    title(sprintf('(k = %d), rep #%d',k,rep));
    xlabel('x_1 = Longitude');
    ylabel('x_2 = Latitude');
    legend(lgd_clust, 'Location', 'best');
    hold off;

    subplot(2,replicates,rep+replicates);
    [s, h] = silhouette(X, idx, 'Euclidean');
    hold on;
    score = mean(s);
```

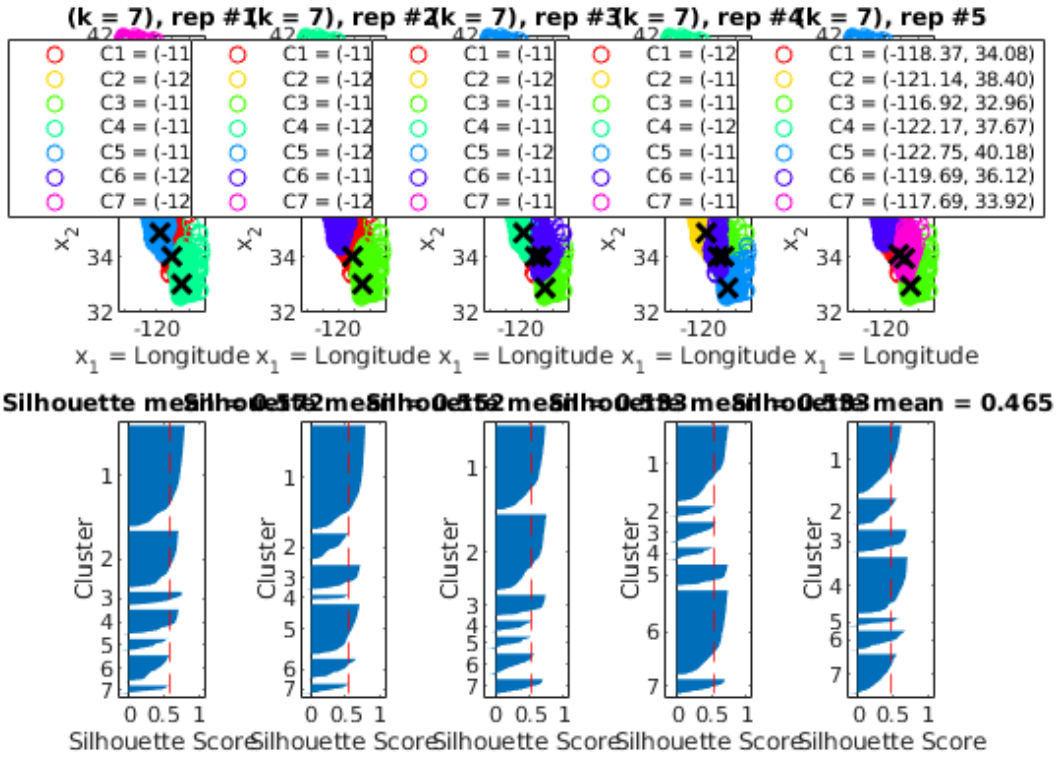
```
    xline(score, '--r');
    title(sprintf('Silhouette mean = %0.3f', score));
    xlabel('Silhouette Score');
    ylabel('Cluster');
    hold off;
end
```

Part 3

```
% 1. Cluster 2 and 3 are much larger than the others because their
% respective centroids are located near large cities (C2 near LA, C3
 near
% San Francisco). There are many houses within these clusters
% (many houses near centroid = smaller mean intracluster distances)
 and
% fewer houses in the next-closest cluster (larger mean intercluster
% distances). The Silhouette score is the ratio of the difference
 between
% these metrics and the largest metric, and thus the scores approach 1
 and
% are comparatively larger than clusters outside of metropolitan
 areas.

% 2. For Replicate 2 in Figure 3 (Andy's Plot), k=7 clusters looks to
 be
% a good choice.
%   Cluster 1 is the Sacramento region.
%   Cluster 2 is the LA region.
%   Cluster 3 is the San Francisco region.
%   Cluster 4 is the San Diego region.
%   Cluster 5 is the Fresno region.
%   Cluster 6 is the Redding/Wine Country region.
%   Cluster 7 is the Santa Barbara region/area between LA and San
 Fran.

% 3. Compared to replicate 2, replicate 3 has 3 clusters whose
 centroids
% are much closer together in SoCal. The cities of San Diego, LA, and
 Santa
% Barbara are almost representative of the centroid locations. The
 clusters
% in replicate 2 were more represenatative of the general regions near
% these cities.

% 4. Replicate 2 has a more homogenous distribution of clusters,
% geographically. Replicate 3 has a higher cluster density in SoCal,
 where
% a large amount of people live. The higher cluster density in this
 region
% will bias the silhouette score to a lower value than replicate 2.

% 5. Personally I like replicate 1. It has a cluster for the SD
 region,
```

```
% LA/Death Valley region, a cluster for Sequoia Park, another for San
 Jose
% and San Fran, and then 2 for the NorCal region. it also has a
 silhouette
% mean of .539, which is not super high (indicating something like
% "overfitting").
```





Published with MATLAB® R2021a