```
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_range = 4:1:9;
replicates = 100;
figure();


for k = k_range
    lgd_clust = cell(k,1);
    plt_num = find(k_range == k);
    subplot(2,range(k_range)+1,plt_num);
    [idx, cent] = kmeans(X, k,...
        'Replicates', replicates);
    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)',k));
    xlabel('x_1 = Longitude');
    ylabel('x_2 = Latitude');
    legend(lgd_clust, 'Location', 'best');
    hold off;

    subplot(2,range(k_range)+1,plt_num+range(k_range)+1);
    [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;
```

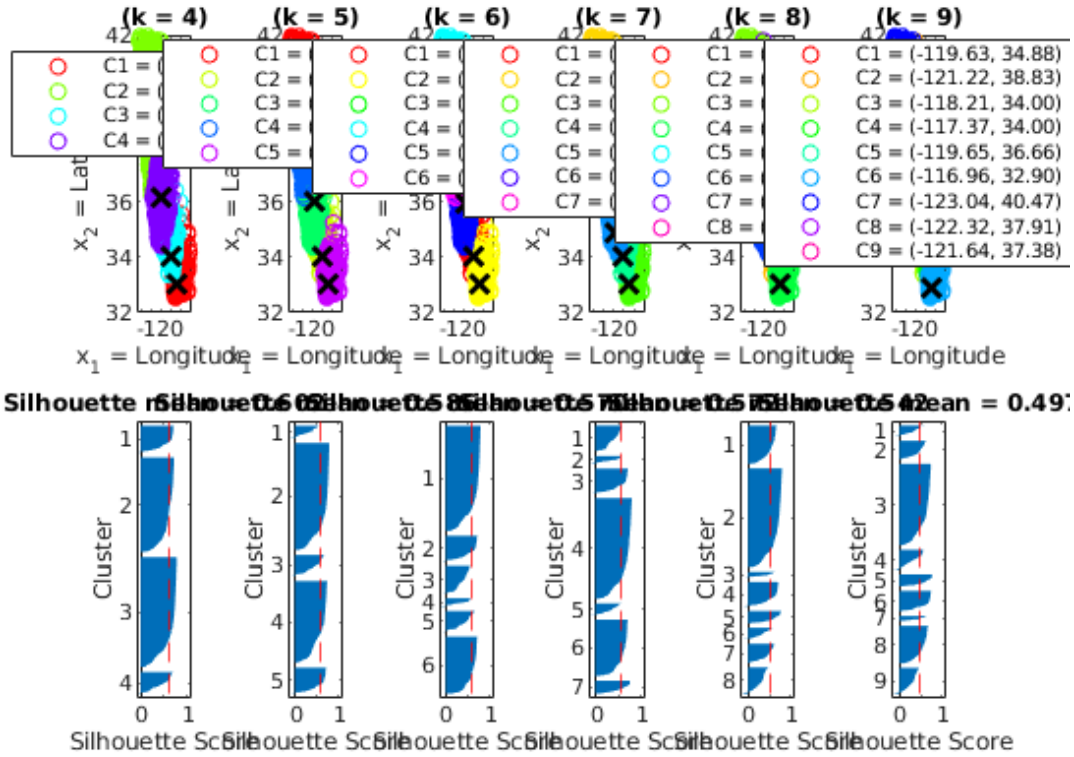
```
end

% 1. Well, the cluster around LA encompasses a ton of houses in the
 city,
% but also the desert to the north where fewer people live. Similarly,
 the
% cluster around San Fran encompasses the many houses in the city and
 those
% in the mountains. Average intracluster distance for these clusters
 is
% slightly higher than expected, and intercluster distances are large.
 For
% these areas, this means a higher silhouette score.

% 2. The improvement we see when increasing clusters from 4 to 5 is
 due to
% an additional cluster in NorCal, decreasing the size of the cluster
% containing San Francisco and allotting specific area to the
 mountainous
% region.

% 3. I think it was worth it to increase clusters from 4 to 7. The 7
% cluster regime makes more sense geographically, showing clusters for
% specific geographic regions around Cali. Yes, we saw a decrease in
% silhouette score, but that is expected. I think it's an affordable
 loss
% given the higher geographic specificity.

% 4. I think 9 clusters is excessive. Clusters in SoCal don't split
 the
% metro regions effectively, and the other clusters are somewhat
 arbitrary
% given the geography they encompass. I think the optimum k given my
 data
% is 7. Seven clusters seem to be a good balance between geographic
% specificity and clustering population densities, and silhouette
 score.
```



Published with MATLAB® R2021a