Table of Contents

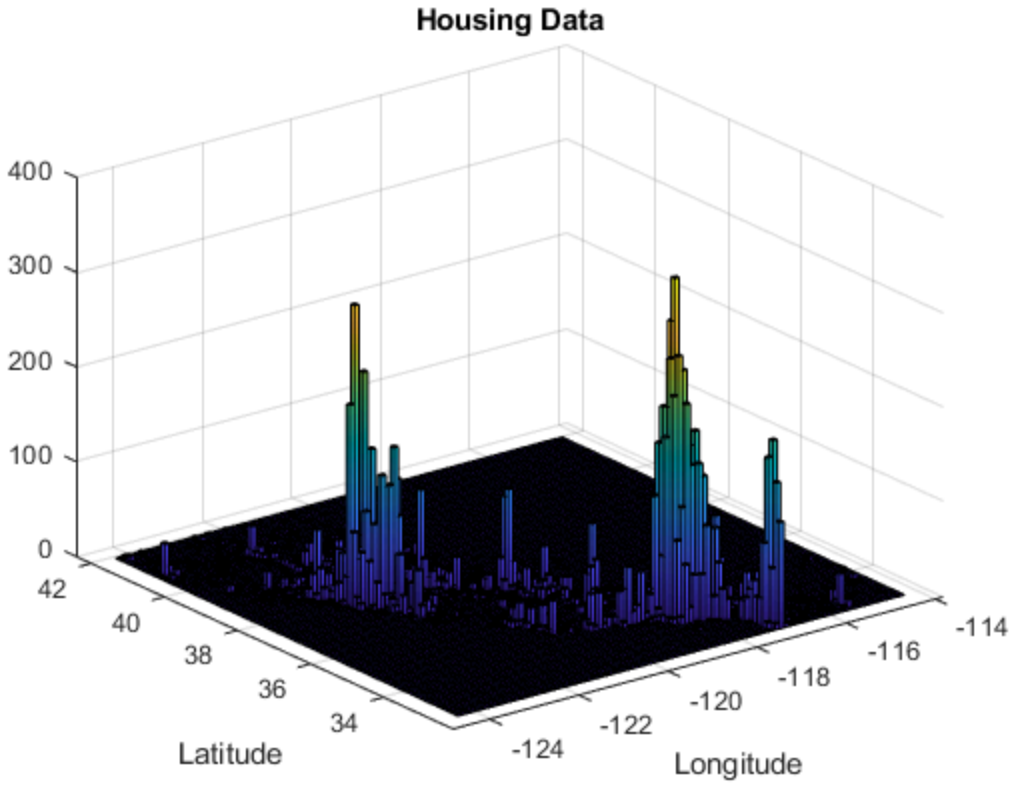
BE606 HW3 Problem 2a

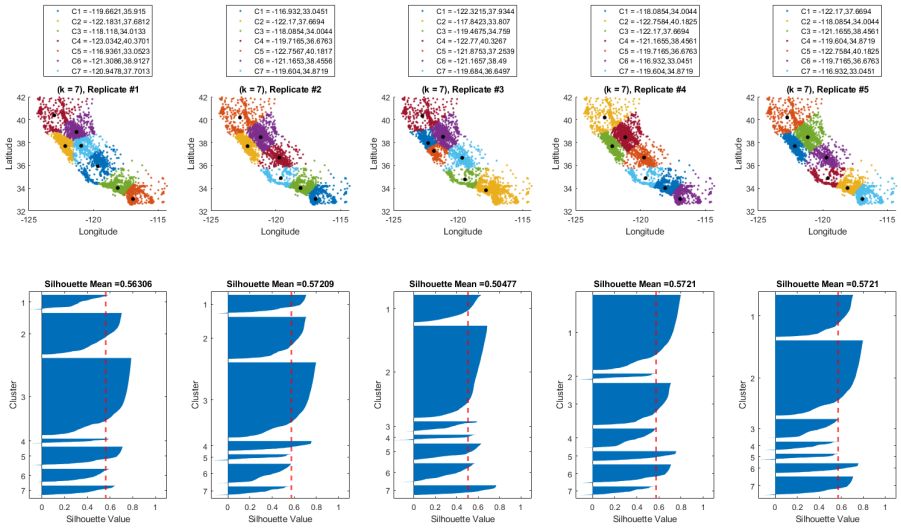
```
clear all
close all
```

Part 2

```
A = readtable('housing.csv');
B = table2array(A(:,1:9));
x1 = B(:,1);
x2 = B(:,2);
X = [x1,x2];


figure;
hist3(X,'CDataMode','auto','FaceColor','interp','Nbins',[100 100])
title('Housing Data')
xlabel('Longitude')
ylabel('Latitude')
% tic
%kmeans
f = figure;

for rr = 1:5
    [class,cent] = kmeans(X,7);
    subplot(2,5,rr)
%     tic
    for kk = 1:7
        hold on
        plot(x1(class==kk),x2(class==kk),'.','DisplayName',...
            ['C',num2str(kk),' =
 ',num2str(cent(kk,1)),',',num2str(cent(kk,2))])
        legend('Location', 'northoutside')


 plot(cent(kk,1),cent(kk,2),'.','MarkerSize',15,'color','k', 'HandleVisibility','o

    end
%     toc

    hold off
    title(['(k = 7), Replicate #', num2str(rr)])
    xlabel('Longitude')
```

```
    ylabel('Latitude')

    subplot(2,5,rr+5)
    sil = silhouette(X,class, 'Euclidean'); %save value for mean
%     toc
    silhouette(X,class, 'Euclidean') %repeat to easily plot
%     toc
    hold on
    xline(mean(sil), 'r--', 'LineWidth', 2);
    hold off

    title(['Silhouette Mean =', num2str(mean(sil))])
end
% toc
f.WindowState = 'maximized'; %use this to maximize plot on screen
%though it does not work so well in the printout
```





Part 3 Questions

```
disp('Q1')
disp('Clusters 2 and 3 of replicate 2 contain the largest population
 centers in California. Cluster 2 contains LA County, the largest
 county in the US')
disp('and cluster 3 contains the bay area.')
disp('Q2')
disp('Yes the 7 clusters make sense, each cluter either contains
 population centers, or the distance is minimized. Centroid 6 is
 Northern California, where there are mainly national forests and some
 smaller cities/towns')
disp('Centroid 2 contains LA County, largest population center in the
 state')
disp('Centroid 3 contains the Bay area and its most likely commuting
 areas')
disp('Centroid 4 is essentially San Diego and its commuting regions,
 as well as reservations.')
disp('Centroid 5 is mostly Yosemite and other forests in the
 interior')
disp('Centroid 1 is similar to centroid 5, with the inclusion of
 Sacramento')
disp('Q3')
disp('Between 2 and 3, there is a decrease in distance and essentially
 a split of LA County, and an increase in the lower middle centroid.')
disp('Q4')
disp('The equation for silhouette coeff, uses max distances in its
 denominator, and since the middle centroid increased in size, and
 pushed two centroids close together, we see a reduction in overall
 silhouette score.')
disp('Q5')
disp('I prefer replicate 2, as it still has a high silhouette score,
 and pretty accurately depicts the Bay Area, LA County, San Diego
```

```
 as well as the national forests/parks. There should be a few large
 clusters, since there are higher population densities in certain
 regions.')
```

```
Q1
Clusters 2 and 3 of replicate 2 contain the largest population centers
 in California. Cluster 2 contains LA County, the largest county in
 the US
and cluster 3 contains the bay area.
Q2
Yes the 7 clusters make sense, each cluter either contains population
 centers, or the distance is minimized. Centroid 6 is Northern
 California, where there are mainly national forests and some smaller
 cities/towns
Centroid 2 contains LA County, largest population center in the state
Centroid 3 contains the Bay area and its most likely commuting areas
Centroid 4 is essentially San Diego and its commuting regions, as well
 as reservations.
Centroid 5 is mostly Yosemite and other forests in the interior
Centroid 1 is similar to centroid 5, with the inclusion of Sacramento
Q3
Between 2 and 3, there is a decrease in distance and essentially a
 split of LA County, and an increase in the lower middle centroid.
Q4
The equation for silhouette coeff, uses max distances in its
 denominator, and since the middle centroid increased in size, and
 pushed two centroids close together, we see a reduction in overall
 silhouette score.
Q5
I prefer replicate 2, as it still has a high silhouette score, and
 pretty accurately depicts the Bay Area, LA County, San Diego as well
 as the national forests/parks. There should be a few large clusters,
 since there are higher population densities in certain regions.
```

Published with MATLAB® R2018b