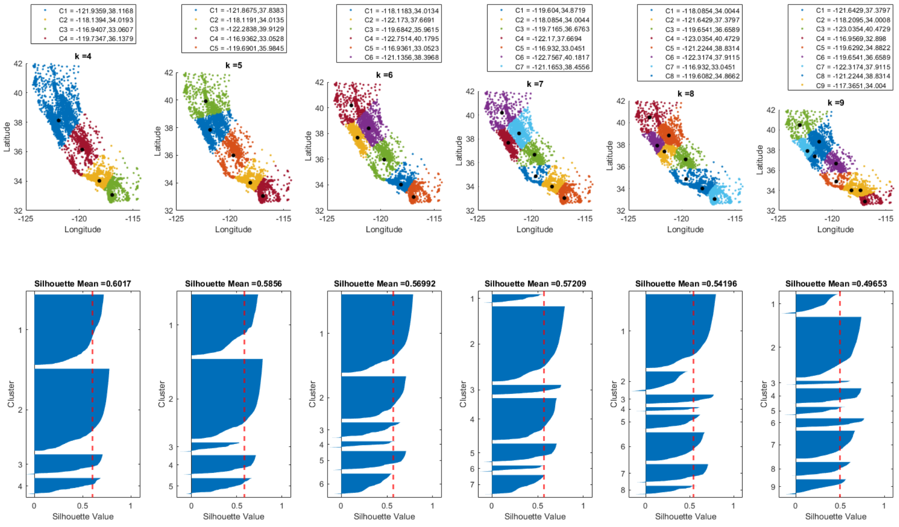
Table of Contents

BE606 HW3 Problem 2b

```
clear all
close all
```

Part 2

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



f = figure;
for kk = 4:1:9
    [class,cent] = kmeans(X,kk,'Replicates',100);
    subplot(2,6,kk-3)

    for jj = 1:kk
        hold on

        plot(x1(class==jj),x2(class==jj),'.','DisplayName',...
            ['C',num2str(jj),' =
 ',num2str(cent(jj,1)),',',num2str(cent(jj,2))])
        legend('Location', 'northoutside')


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

    end
    hold off

    title(['k =', num2str(kk)])
    xlabel('Longitude')
    ylabel('Latitude')


    subplot(2,6,kk+3)
    sil = silhouette(X,class, 'Euclidean'); %save value for mean

    silhouette(X,class, 'Euclidean') %repeat to easily plot

    hold on
    xline(mean(sil), 'r--', 'LineWidth', 2);
    hold off
    title(['Silhouette Mean =', num2str(mean(sil))])
end
f.WindowState = 'maximized';
```



Part 3 Questions

```
disp('Q1')
disp('The two large clusters contain where the majority of the
 population lives. One contains LA County and its surroundings, and
 the other contains nearly a quarter of California.')
disp('Q2')
disp('In k=5, there is a marked improvement as the large northern
 California cluster can be split in two, allowing for more appropriate
 geographical classification. San Francisco is now in its own
 cluster.')
disp('Q3')
disp('Yes, the silhouette score did not decrease much, and now each
 population center in CA is more accurately described. Cluster size is
 also not as drastically large.')
disp('Q4')
disp('After k=7 the silhouette score begins to decrease. Though there
 are more clusters, sectioning off cities and communities possibly
 more effectively, we are beginning to generate too many clusters.
 Realistically the two best were k=6/7 considering the silhouette
 mean is relatively close, and the map intuitively maps the major
 population centers.')
```

```
Q1
The two large clusters contain where the majority of the population
 lives. One contains LA County and its surroundings, and the other
 contains nearly a quarter of California.
```

Q2

In k=5, there is a marked improvement as the large northern California cluster can be split in two, allowing for more appropriate geographical classification. San Francisco is now in its own cluster.

Q3

Yes, the silhouette score did not decrease much, and now each population center in CA is more accurately described. Cluster size is also not as drastically large.

Q4

After k=7 the silhouette score begins to decrease. Though there are more clusters, sectioning off cities and communities possibly more effectively, we are beginning to generate too many clusters. Realistically the two best were k=6/7 considering the silhouette mean is relatively close, and the map intuitively maps the major population centers.

Published with MATLAB® R2018b