# Tutorial 10 Solution

1. Suppose we have data for five objects on two features:

| **object** | $x_1$ | $x_2$ |
|---|---|---|
| A | 1 | 1 |
| B | 1.5 | 2 |
| C | 3 | 4 |
| D | 3.5 | 5 |
| E | 4.5 | 5 |

We set $k = 2$ to cluster the six data points into two clusters, $\mathcal{P}$ and $\mathcal{Q}$, and initialize the algorithm with the centroids $(x_{1,\mathcal{P}}, x_{2,\mathcal{P}}) = (2, 2)$ and $(x_{1,\mathcal{Q}}, x_{2,\mathcal{Q}}) = (4, 4)$.

Solution:

(a) Fill up the following table to identify the objects in each cluster during the first iteration of the $k$-means algorithm.

| **cluster** | **object(s)** |
|---|---|
| $\mathcal{P}$ | A, B |
| $\mathcal{Q}$ | C, D, E |

For A: the distance from A to centroid of $\mathcal{P}$ is shorter than to the centrod of $\mathcal{Q}$. Hence, A is classified to $\mathcal{P}$.
Similarly, B is classified to $\mathcal{P}$ also.
C, D and E are closer to $\mathcal{Q}$ than to $\mathcal{P}$. Hence, C, D and E are classified to $\mathcal{Q}$.

(b) Compute the new centroids for the two clusters based on cluster assignment in (a).

Answer: $\mathcal{P}$ now has A and B. New centroid for $\mathcal{P}$: $\left(\frac{1+1.5}{2}, \frac{1+2}{2}\right) = (1.25, 1.5)$

$\mathcal{Q}$ now has C, D and E. New centroid for $\mathcal{Q}$: $\left(\frac{3+3.5+4.5}{3}, \frac{4+5+5}{3}\right) \approx (3.67, 4.67)$

(c) Based on the centroids computed in (b), identify the objects in each cluster during the second iteration of the $k$-means algorithm.

Answer: For A, the distance to the new centroid of $\mathcal{P}$, (1.25, 1.5) is shorter than to the centroid of $\mathcal{P}$, (3/67, 4.67). Hence, A is classified to $\mathcal{P}$ again. Similar for B.
For C, D and E, they are closer to the new centroid of $\mathcal{Q}$. Hence, they are classified to $\mathcal{Q}$ again.
The classification is the same as at the end of 1st iteration. Hence, it's converged.

(d) Calculate the Within Sum of Squares (WSS) for the clustering assignment in (c).

Answer: Sum of squares for cluster $\mathcal{P}$: $(1 - 1.25)^2 + (1 - 1.5)^2 + (1.5 - 1.25)^2 + (2 - 1.5)^2 = 0.625$

Sum of squares for cluster $\mathcal{Q}$: $(3-3.67)^2 + (4-4.67)^2 + (3.5-3.67)^2 + (5-4.67)^2 + (4.5-3.67)^2 + (5 - 4.67)^2 \approx 1.833$
Hence, $WSS \approx 0.625 + 1.833 = 2.458$.

2. (K-Means) Consider data set `hdb-2012-to-2014.csv` which was extracted from the published data [1]. The file has information on the HDB resale flats from Jan 2012 to Dec 2014.

   (a) Load data into R. Use $k$ means algorithm to pick an optimal value for $k$ (using $WSS$ as a criterion), based on two variables, `resale_price` and `floor_area_sqm`.

   (b) With the optimal $k$ in part (a), plot the data points in the $k$ clusters determined.

<u>Solution:</u>

(a)
```
> data = read.csv("C:/Data/hdb-2012-to-2014.csv")
> dim(data)
[1] 6047   11
> names(data)
 [1] "X"                   "month"               "town"
 [4] "flat_type"           "street_name"         "storey_range"
 [7] "floor_area_sqft"     "floor_area_sqm"      "flat_model"
[10] "lease_commence_date" "resale_price"
> attach(data)
> # PLOT WSS vs K TO PICK OPTIMAL K:
> K = 15
> wss <- numeric(K)
> for (k in 1:K) {
+    wss[k] <- sum(kmeans(data[,c("floor_area_sqm","resale_price")], centers=k)$withinss)
+ }
> plot(1:K, wss, col = "blue", type="b", xlab="Number of Clusters",
+      ylab="Within Sum of Squares")
> # plot is shown in Figure 1
```

From the plot in Figure 1, $k = 3$ would be a good choice.

(b) With the optimal $k$ in part (a), plot the data points in the $k$ clusters determined.

```
> kout <- kmeans(data[,c("floor_area_sqm","resale_price")],centers=3)
> # visualize the 3 groups:
>
> plot(data$floor_area_sqm,
+      data$resale_price,
+      col=kout$cluster)
> # plot is shown in Figure 2
```

[1] `https://data.gov.sg/dataset/resale-flat-prices`

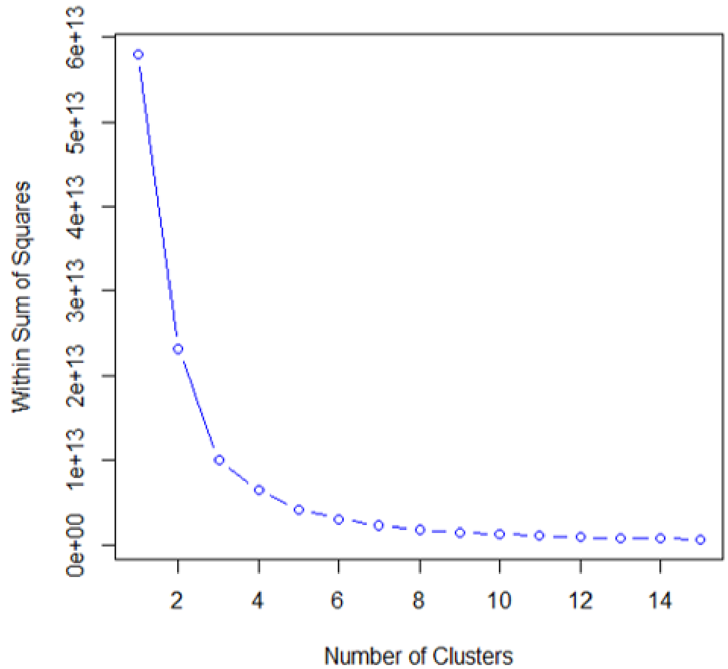


Figure 1: Q2(a)

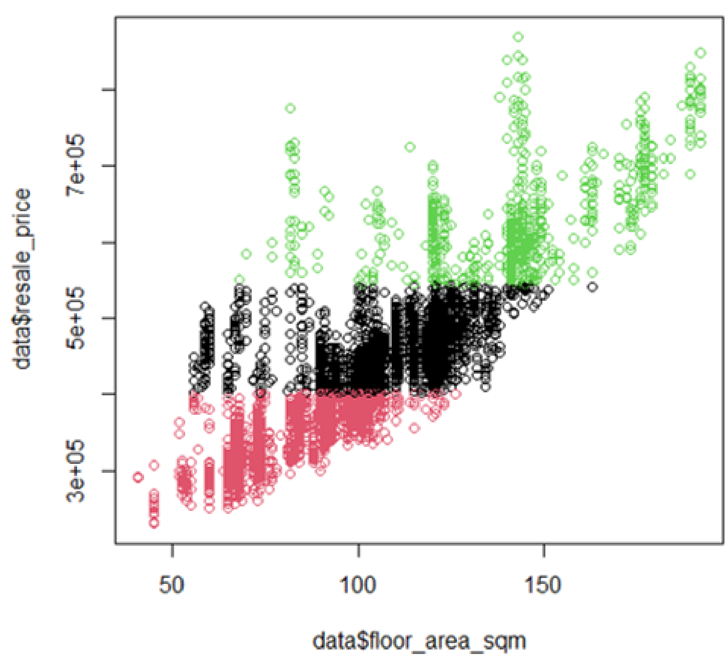


Figure 2: Q2(b)