**Quiz 8**

Content: kmeans()

Data set: <hdbresale\_cluster.csv>

We would cluster all the flats in the data set into groups.

**Q1**. In order to use K-Means algorithm for a data set, we need to identify the response column in that data set first, then apply function kmeans() for the response column. True or False?

Ans: False

Explanation: K-Means is an algorithm for unsupervised learning, where there is no response for the ‘model’ to learn.

**Q2**. The number of clusters (or groups) for all the flats in the data set given could be chosen differently from user to user. True or False?

Ans: True

Explanation: One might choose 2, other might choose 3 or 4, etc.

**Q3**. Clustering by K-Means algorithm should only be used for quantitative variables. True or False?

Ans: True

Explanation: the algorithm calculates the Euclidian distance between the data points, hence, it should only be used for quantitative variable, not for categorical variable, especially nominal ones.

**Q4**. We could check the goodness of fit of the K-Means algorithm for the data set given above, for example, by calculating the accuracy of it. True or False?

Ans: False.   
Explanation: There is no response in that data set, and there is no real cluster for those flats. Hence, the clustering by the algorithm is just to separate those flats into different groups, and there is no confirmation for the clustering be right or wrong.

**Q5**. In order to decide on the number of clusters for a data set, the plot below could help

1. Box plot of each column
2. Histogram of each column
3. Scatter plot of every two quantitative columns
4. Pie chart of every column

Ans: Option C

**Q6**. If the quantitative variables in the data set are of very different magnitudes, we should standardize each of them before applying the K-Means algorithm. True or False?

Ans: True.

Explanation: This is because the algorithm calculates the Euclidian distance between data points. Hence, the variable with large magnitude will dominate the distance. That means, the clustering will mainly be affected by that variable and the contribution of variable (s) with small magnitude will become minor in clustering step.