Homework 6

Overview

In this assignment, you will apply K-means clustering to a dataset and interpet the results.

For this assignment, use this file as the template for your homework assignment. You should add new code chunks as needed.

Objectives

- Apply K-means clustering using R
- Interpret the resulting clusters
- Apply a different clustering algorithm of your choice to the dataset and compare its results to the K-means clustering results.

Grading

- Uploaded requested files, 5%
- File is properly formatted, using the template provided, 5%
- Writing is clear and appropriately formal, 5%
- Random number seed set to 1 before clustering (done for you in this file): 5%
- Questions 1-8: 10% each

Deliverables

- .pdf file (generated by knitting your .Rmd file)
- .Rmd file (used to generate your pdf file)

Setup

Load all of the packages you need in the code chunk below (you may need to install some packages that you do not already have installed). Feel free to load any extra packages that you want to use for your homework.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(cluster)
```

Set the random number seed to `1`:

```r
# Set the random number seed:
set.seed(1)
```

Part A. Running K-means

1. Load the dataset

Download and load the dataset provided on blackboard (data.csv) as a dataframe. There are no missing values, so you don't need to remove any missing values. Normally, you would standardize each attribute you plan to cluster with, but for this assignment, don't do any scaling/standardizing for the clustering.

```r
# Your code here.
```

These data represent people: their age and how they spend their money (media, food, transportation, housing, and pets). In the questions below, You will use the K-means clustering algorithm to cluster these data.

2. Apply K-means clustering to the data

Apply K-means clustering with `K=5`.

```r
# Your code here.
```

3. What are the sizes of each cluster?

```r
# You code here.
```

4. What are the centroids for each cluster?

```r
# Your code here.
```

5. In your own words, describe each cluster.

Look at the centroids for each cluster. Based on the centroids, describe each of the five clusters. Are there any obvious details about the people in each cluster that you can see based on the cluster centroids?

- Cluster 1:
- Cluster 2:
- Cluster 3:
- Cluster 4:
- Cluster 5:

6. Which cluster contains the most within-cluster variation (i.e., is the least compact)?

Look at each cluster's within-cluster sum of squares. Use this information to identify which cluster contains the most within-cluster variation.

```
# Your code here
```

Part B. Clustering continued

7. Apply another clustering algorithm of your choice (not K-means) to the dataset provided for this homework assignment (`data`).

```
# Your code here
```

8. Describe how the clustering results from the clustering algorithm you used in #7 are different from (or similar to) the results from K-means.