# Monte Carlo Simulation

## 1. Scientific or Statistical Question

How do different initial values in clustering affect clusters of price, score, and location?

## 2. Data

The data-generating model that will be used is the ‘generate_data’ function in R. In create the different clusters, the membership of the data points needs to be identified. Membership is indicated by a “1” and a “0” indicates the absence of that data point in the cluster. The data will be generated by repeating 0 and 1 50-100 times.

Mathematical Notation: $y = b1 + b2 + b3$

The assumptions are that all values are numeric and scaled consistently.

## 3. Estimates

Via simulation, I am estimating what values in the “melted groups” are assigned correctly. “Melted groups” refer to groups that have two “1” values and one “0” value, indicating their double-membership.

## 4. Methods

I am evaluating the effect of differing levels of variance on cluster formation.

The calculations will involve three different types of variance to account for noise, two separate ways. This creates six different treatments.

I hypothesize that higher variance will cause more clusters, because high variance indicates more variability and spread within a dataset.

## 5. Performance Criteria

Assessing the methods will be determined by the clusters that are formed, and if the algorithm got the melted group membership correct.

## 6. Simulation Plan

- Detail how the experiment will be carried out:
  - Number of simulations: Several, so that we get enough repetitions to achieve statistically different numbers
  - Parameter settings: Choosing the type of clustering, selecting the method that calculates distance, numbers of clusters
  - What will be recorded: Clusters that are naturally formed
  - Any changes from Project III's code
    - Add data-generating model
    - Change initial values of clusters

## 7. Anticipated Challenges or Limitations

To have clustering be an efficient method of analysis, an optimal number of clusters must be found. That may be difficult with the added variance and noise. I also anticipate analyzing the clusters for tangible results would be difficult, at least with my current understanding.

## 8. Code & Results

```
# set seed for reproducibility
set.seed(123)

# generate fake data
generate_data <- function(membership) {
  n <- length(membership)

  # identify clusters
  cluster1_price <- rnorm(n, mean = 5, sd = 2)
  cluster2_score <- rnorm(n, mean = 10, sd = 3)
  # create a data frame
  data <- data.frame(
    cluster1_price = cluster1_price,
    cluster2_score = cluster2_score,
    cluster = membership
  )

  return(data)
}

# random list of 0s and 1s to indicate membership
membership_vector <- sample(c(0, 1), size = 100, replace = TRUE)

# generate the data
generated_data_clusters <- generate_data(membership_vector)

head(generated_data_clusters)
```

```
  cluster1_price cluster2_score cluster
1       5.506637      12.363217       0
2       4.942906      12.307127       0
3       4.914259      10.996608       0
4       7.737205       6.974870       1
5       4.548458       9.641642       0
6       8.032941       9.158814       1
```

The code creates a simulation of what cluster membership would look like. The outcome is a lit of 100 randomly generated data points that are grouped into clusters.