Course: *DATA 101: Intro to Data Science* – Professor: *Dr. Osei Tweneboah*

Homework #4 Due on 12/06/2022

Instructions: While discussion with classmates is allowed and encouraged, please try to work on the homework independently and direct your questions to me. Please interpret your analysis results using concise and clear language and focusing on interesting findings.

Problem 1

The Data. The 'maps' package contains a database of world cities i.e. 'world.cities' of population greater than about 40,000. Also included are capital cities of any population size, and many smaller towns.

The database constitutes a list with 6 components, namely "name", "country.etc", "pop", "lat", "long", and "capital", containing the city name, the country name, approximate population (as at January 2006), latitude, longitude and capital status indication (0 for non-capital, 1 for capital, 2 for China Municipalities, and 3 for China Provincial capitals).

Consider the 4,000 biggest cities in the world and their longitudes and latitudes:

```
library(maps)
big_cities <- world.cities %>%
  arrange(desc(pop)) %>%
  head(4000) %>%
  select(long, lat)
glimpse(big_cities)
```

Using *k*-means clustering algorithm separate these 4,000 points each of which is located in a two-dimensional plane into *k* clusters based on their locations alone.

Re-fit the *k*-means algorithm on the BigCities data with a different value of *k*. Experiment with different values of *k* and report on the sensitivity of the algorithm to changes in this parameter.

Problem 2

Baseball players are voted into the Hall of Fame by the members of the Baseball Writers of America Association. Quantitative criteria are used by the voters, but they are also allowed wide discretion.The following code identifies the position players who have been elected to the Hall of Fame and tabulates a few basic statistics,including their number of career hits(H), home runs(HR), and stolen bases(SB).

```
library(Lahman)
hof <- Batting %>%
  group_by(playerID) %>%
  inner_join(HallOfFame, by = c("playerID" = "playerID")) %>%
  filter(inducted == "Y" & votedBy == "BBWAA") %>%
  summarize(tH = sum(H), tHR = sum(HR), tRBI = sum(RBI), tSB = sum(SB)) %>%
filter(tH > 1000)
```

Use the *k*-means algorithm to perform a cluster analysis on these players.Describe the properties that seem common to each cluster.