Homework 4

Christopher Hainzl

2022-12-02

Question 1

```
library(maps)
```

```
## Warning: package 'maps' was built under R version 4.2.2

##
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':
##
##     map
```

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

```
## Rows: 4,000
## Columns: 2
## $ long <dbl> 121.47, 72.82, 67.01, -58.37, 77.21, 120.97, 37.62, 126.99, -46.6~
## $ lat  <dbl> 31.23, 18.96, 24.86, -34.61, 28.67, 14.62, 55.75, 37.56, -23.53, ~
```

```
cities_k_means <- kmeans(big_cities, 400)
cities_k_means$size
```

```
##   [1] 13  4 11  1  3  4 12  2 13  9  3 12  9  7 10 11  5  6  8 12  3  8 21  5 10
##  [26]  1  9 17 15 10  8 13  5 11  6 16 14  5 12 12 14 20  4  6  7  7  6 11  7 12
##  [51] 22 21  9  5 16 27  9  4 14  5 13  6  6 10  6 11  4  7 22  4 30  8  4 20 12
##  [76]  3  5 18  9  5 14  4 12 16 11  1  1 11  2 11  8  6 10  7 12  6 18  7  8 25
## [101]  3 25 30  9  5 16 10 17  7  7  9  4  1  7  4  4  4 11  4  6  5 10 11  9  3
## [126] 16 11  1  2  5 10 12  5  5 11  2  2  7 13  1 20  7  5 11  8  6  4  7  7  8
## [151] 35  7  9  6  7 17 12 13  8 11  6  5 10 25  2 17  8  1  3  7  8 10  3  4  2
## [176]  5 10  8 10 25  5 14  4  9  8  2 10  8  3 11  4  8  8  6 11 13  6 15  9 11
## [201]  7  4 37  6 10 12  8  9 12 25  6 15  4 16 19 15  1  8  7 17 16  5  3 10 13
## [226]  1  9  7  6  9 25 10  9  3  6  6  7  1 27 17  2  9  7 18 10  2 10  4 10 16
## [251]  7  7 12  6 17 11 17 14  6  9  7  4  8 11 12 16 19 15 11 31  9  7  5  7  7
## [276] 12 17  5 24 11  6  9 21 18 12  2 14  9 14 13  7 16  8  8  2 10 10 16  5  8
## [301] 10  9  9 16 15 11 30 15  2 16  7  3 28  3 27 13  6 13  9 15  8 13  9 15  5
## [326]  1  8 45 10  4  9  6 11 13  9 22  7 10  3 12  4  9 25  1 22  5 33  8 38  8
## [351] 12  6  8  4 19 14 10 17  8 14  5  4  9  4  6  5  7  4  1 20 18  7  4  6  5
## [376]  4 17  5  8 13  7 19 11 10  5  9  6  5 10 15 31  8  8  5  7 14  7  1 12  7
```

```
cities_k_means <- kmeans(big_cities, 600)
cities_k_means$size
```

```
##   [1]  8  6  4  6  2  3  1  8  4  6 13  5  1  3  8  3  3 17 13 17 10  9  2  7  5
##  [26]  7  4  7  6  4  9 11  6 14  1  1  5  2  4  5  3  2  6  3 12 11  6  8 22 18
##  [51]  4  6  5  4  6  6  5 13 12  4  7  3  4 14 10  5 11  8  4  3  4  7  3  4  3
##  [76]  4  9  5 10  9  6  5  8  4 10  7 17 18  3  7  4  3  4  5  4  6  6 13  3  6
## [101]  6  3  4  3  4  8 10  4  3  1  5  6  5  1  6  2  5  8 10  3  1  5 12 34  3
## [126] 10  8  3  3 14  2  8  4  6 11  6  4  9 10  3  7  4  7  7  2 11  4  5  6  6
## [151]  3  5  4  9 10  4 12  4  5  6 10  2  8  8  4  8  6  7  7  1  3  2  3 12  3
## [176]  3  5  2 11  5  7 11  8  5  3  5  2 13 13  3  2 12 10 12  5  4  4  8  5  5
## [201]  5  5  6 13  8  8 17  5  4  3  5  7 14  7  5  7  9 11  9  6  9  4  3  3  6
## [226]  5  3  5  7 12  1  6 10  8 13  4  5  4 22  2  5  5  5  3 12  7  6  5  5  9
## [251]  5  7  5  2  4  9  4 12  7  1 13  9  5  8  4  5  2  9  8  5  7  1 10  3  9
## [276] 13  1  9  3  7  7 16 13 11  1  4  4  9  2  4 18 23  4  3  7  9  3  5 10  4
## [301]  7  5  9  6 15 12 10  7  4  6  4  4  5  7  4  8  3  7  2 14  5  2  6  5  7
## [326]  7 12 15  8  6  5 13  7  3  3  4  3  7  7  2  4 10  4  4  5  4  5  5  4  6
## [351]  6 12  8  4  4  4  3  4 12 10  6  4  4  3  9 12 12 10  3 10  4  2  6  7  4
## [376]  1  6  4  5  1 11  4 21  4 10  1 10  3  3  8  5  6  6  3  6 13 11  9  4  7
## [401]  9 10  7 17  7  5  5  6  7  4  9  7  4  3  5  4  2  3  8  4  8  9  4  4 12
## [426]  2  5  6  6  1  5  9  5 15  3  7  3  5  2  8  5  9  6  4  7 14 10  7  9  7
## [451]  3  6 10  5  7  8  4 21  7 15  4  3  7  3  7  6  6  4  3 15  5  2  4  6  5
## [476]  2  8  4  7  8  7  3 13  4  6  1 13  6 14 10  9  4  8  1 11  6  2  7  4  8
## [501] 16  6  3  9  8 12  5  2 10  6 13  4  6  6  5  5 15  7 21  4  7 20  3  3 38
## [526]  1  4  5  9  1 18  2  5 10  3  1  7  5  8  8  3  7  6  4  4  7  5  3  5 12
## [551]  8 12  5  4  5  3  2  1  4  3 12  8 11  9  6  2  2  9  5  7  6 11 10  2  7
## [576] 15 10  6 10  4  9  6 10  8  8  8  5  3  5 16  8  3  3  6  8  6  1  4  6 19
```

- As we increase the value of k, the sizes of some of the clusters appear to decrease, while for others, their sizes appear to increase. This makes sense because as the amount of clusters changes, it affects the amount of variance between the data points.

Question 2

```
library(Lahman)
```

```
## Warning: package 'Lahman' was built under R version 4.2.2
```

```
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)
```

```
hof_filtered <- hof[,c(2:5)]
```

```
hof_clustered <- kmeans(hof_filtered, 20)
hof_clustered
```

```
## K-means clustering with 20 clusters of sizes 5, 6, 5, 3, 2, 3, 1, 5, 5, 3, 1, 4, 7, 5, 3, 6, 6, 5, 2
##
## Cluster means:
##          tH      tHR     tRBI        tSB
## 1  3362.200 116.2000 1493.600  556.00000
## 2  3052.000 370.0000 1686.333  136.66667
## 3  3471.600 569.2000 1982.400  186.80000
## 4  2866.667 575.6667 1836.000  159.00000
## 5  3039.000 223.0000 1007.500 1172.00000
## 6  1649.667 108.0000  785.000  104.00000
## 7  4189.000 117.0000 1944.000  896.00000
## 8  2370.400 451.0000 1454.400  121.80000
## 9  2587.000 530.2000 1760.600   84.20000
## 10 1413.333 314.0000 1049.000   35.00000
## 11 2873.000 714.0000 2217.000  123.00000
## 12 3123.750 197.5000 1257.000  277.00000
## 13 2059.143 334.7143 1298.143   45.42857
## 14 2493.600 123.8000  897.800  420.00000
## 15 2900.667 232.3333 1234.333   59.66667
## 16 2698.833 390.5000 1562.667  149.50000
## 17 2217.500 515.0000 1585.000   47.50000
## 18 2333.800 158.8000 1248.600   95.40000
## 19 2561.000 219.0000 1056.500  748.50000
## 20 2880.200 108.4000 1075.600  423.20000
##
## Clustering vector:
##  [1]  3 20 14  8  9 13 13 12 15  6  2  5 10 20 13  7  6  1 18 16 13 17  8  9 20
## [26] 10  4 16 12 13 16  5  2  9 16  2 20 17 10  1 14  8 14  8  3 17 18  1 19  3
## [51]  3  4 16 13 18 19  8  2 15  4  6 15 11 14 17  2 20 14 13  1 17 18  9 17 18
## [76]  1 12 16  9  2  3 12
##
## Within cluster sum of squares by cluster:
##  [1] 317898.80 183632.67 443086.00  29195.33 144088.50  54514.67      0.00
##  [8]  74885.20 106310.80 210402.67      0.00 124431.75 150130.86 243074.80
## [15] 118412.67 145261.17 106905.00 178618.00  27459.00 208792.00
##  (between_SS / total_SS =  93.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

- The properties that seem common to each cluster are having very large mean values for the ‘tH’ and ‘tRBI’ columns, as well as mean values for the ‘tHR’ and ‘tSB’ columns that are (for the most part) relatively smaller in comparison.