ICA 11

Derien Weatherspoon

2023-04-12

```
housing <- read.csv("housing.csv")
library(GGally)
```

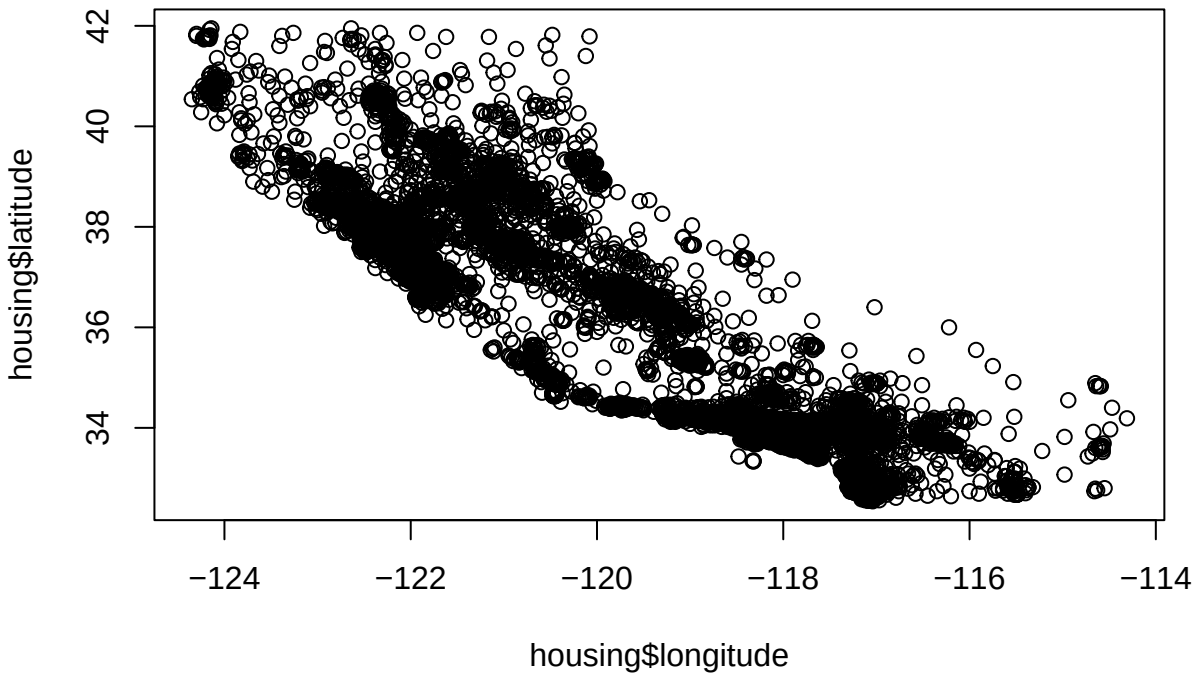
```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggplot2)
```

Question 1: Using "housing.csv", plot the data by longitude and latitude. What data are we looking at?

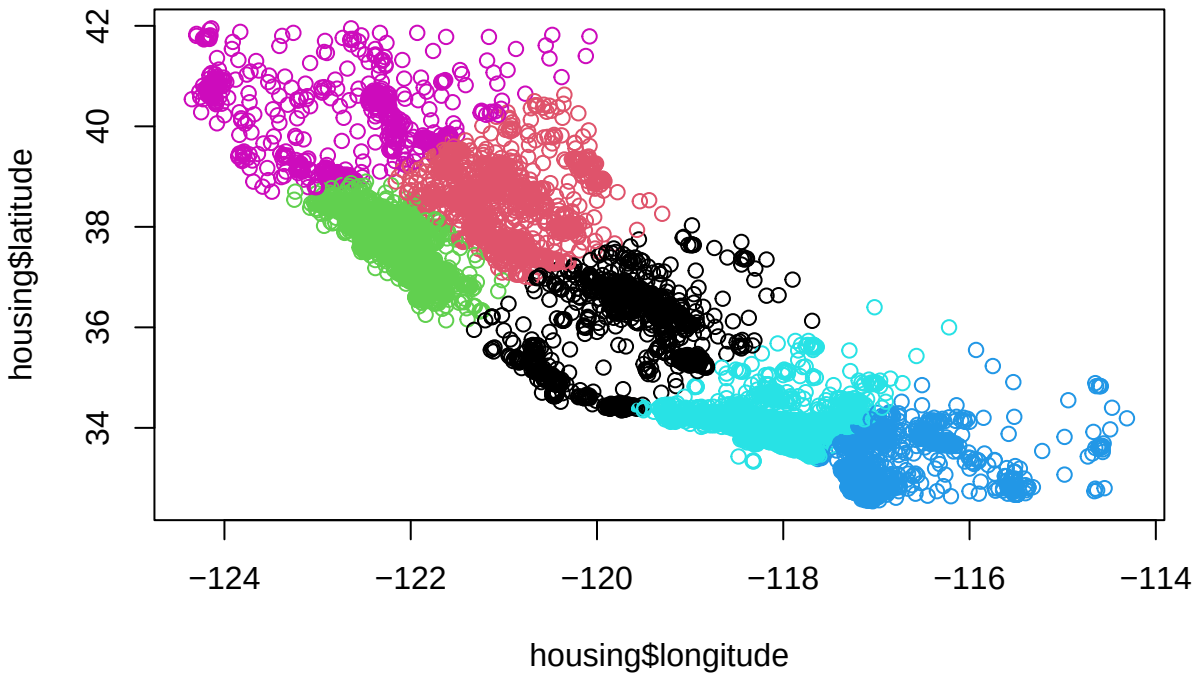
```
plot(housing$longitude, housing$latitude)
```



The graph is pretty much the state of California.

Question 2: Trying different numbers of clusters, cluster the data by these location variables. Visually inspect the result (use cluster id as a color). What seems like an optimal number of clusters

```
housing_scaled <- scale(housing[,1:2])
housing_km <- kmeans(housing_scaled, centers = 6)
plot(housing$longitude, housing$latitude, col = housing_km$cluster)
```



Optimal number of clusters seems to be 6.

Question 2: The rest of the questions are working with the optimal cluster number.

a. How big are the different clusters? The clusters get smaller as you go from top to the bottom. They are not all evenly divided either.

```
housing_cluster <- cbind(housing, housing_km$cluster)
```

b. Filter the data so you just have the largest cluster.

c. Do a second clustering on this largest cluster, using age, population, and income (this is an example of hierarchical clustering). Is there a natural cluster number for this one?

d. For this largest cluster, create a linear regression model to predict median house value.

e. Duplicate the model for the overall dataset. How do the errors compare? Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.