Interview case study Xaringan Lina Berbesi

Lina Berbesi

October, 2023

Unsplash images Taken from: https://unsplash.com/photos/

Case Study

Situation

Airbnb has allowed regular people take the space of formal hospitality businesses to offer both short and long term accomodation. [1]

For government purposes temporal accomodations hosted in residential properties represents a challange since they can not be easily identified as commercial acommodation does. For research purposes while commercial properties can be recognized by using the building consents or business registers hold by the government Airbnb data is out of reach and has a cost since it is owned by a private company.

[1] Airbnb https://www.airbnb.co.nz/

Action

It is difficult to access Airbnb data without paying for their API services. Nonetheless, there are certain projects that advocates for Airbnb data to be public such as Inside Airbnb. [2]

Using Inside Airbnb publicly available data we can identify if Airbnbs are being randomly distributed across New Zealand or if there was a particular concentration in certain areas. There is the potential of also looking into the characteristics of the dwellings that were being clustered to see if these display any particular pattern but this is going to be taken out of scope for this exercise.

Given that the number of clusters was unknown Hierarchical clustering was chosen as the preferred methodology. Hierarchical clustering is a form of unsupervised learning that helps to draw inferences from unlabeled data. For this particularly agglomerative clustering based on the distance between the Airbnbs it is going to be used to select which dwellings belong to each cluster.

[2] Inside Airbnb http://insideairbnb.com

Result

Data for New Zealand was taken from the regional files shared in the Inside Airbnb project.[3]

```
con <- gzcon(url(paste("http://data.insideairbnb.com/new-zealand/2023
                       "listings.csv.gz", sep="")))
txt <- readLines(con)
airbnb_listings_tmp  <- read.csv(textConnection(txt))
```

[3] Open source Airbnb data http://insideairbnb.com/get-the-data

Selecting only the North Island and sampling 1000 dwellings.

```
airbnb_listings<-airbnb_listings_tmp %>% filter(latitude>-45 & longit

set.seed(123)
index <- sample(1:nrow(airbnb_listings), 1000)
airbnb_listings_north<-airbnb_listings[index, ]
```

Keeping only the latitude and longitude while converting the data to a Spatial Points Data Frame object.

```
x<-airbnb_listings_north$longitude
y<-airbnb_listings_north$latitude

xy <- SpatialPointsDataFrame(
  matrix(c(x,y), ncol=2), data.frame(ID=seq(1:length(x))),
  proj4string=CRS("+proj=longlat +ellps=WGS84 +datum=WGS84"))
```

Using the distm function to generate a geodesic distance matrix in meters.

```
mdist <- distm(xy)
```

Clustering all dwellings using a hierarchical clustering approach.

```
hc <- hclust(as.dist(mdist), method="complete")
```

Defining the distance threshold, in this case 5km - 5000m.

```
d=5000
```

Defining the clusters based on a tree "height" cutoff "d" and adding them to the Spatial Data Frame.

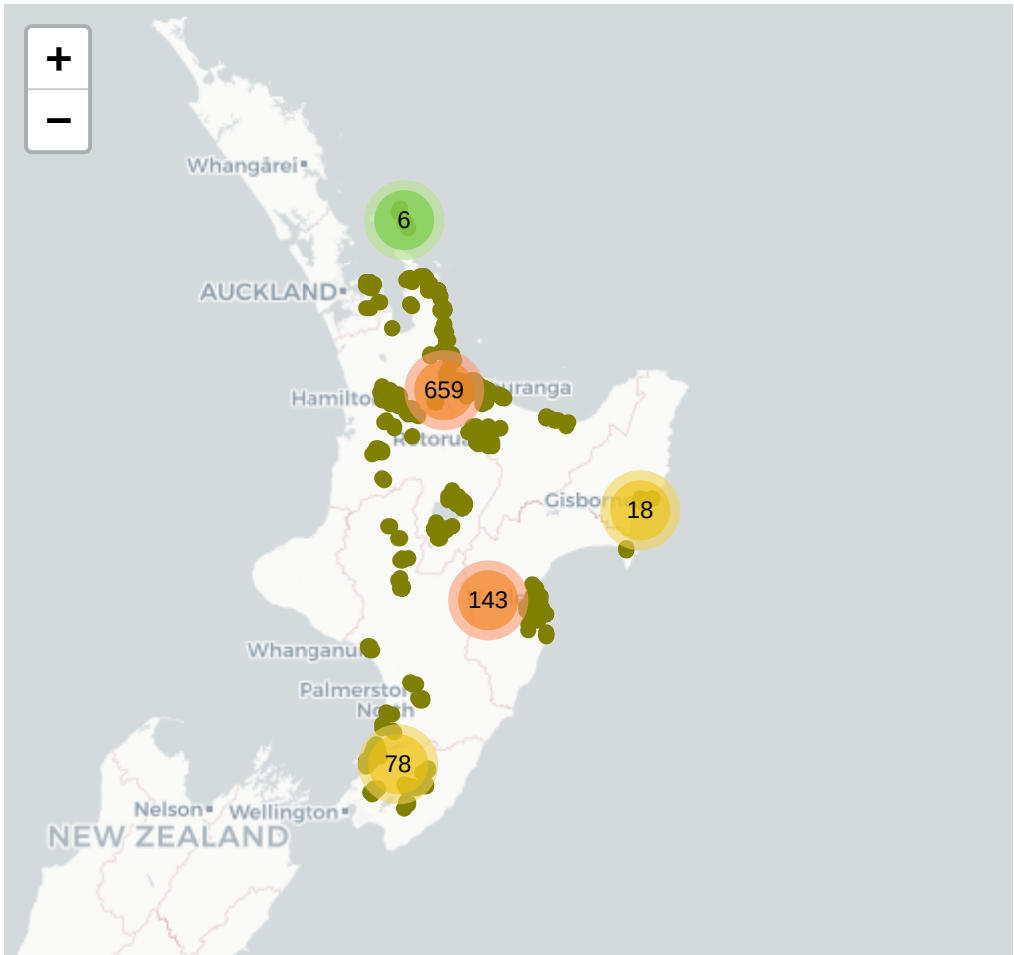
```
xy$clust <- cutree(hc, h=d)
```

Selecting the Id variable to name the clusters.

```
xy_sp<- st_as_sf(xy) %>% filter(duplicated(clust) | duplicated(clust,
  group_by(clust) %>%
  mutate(clust_fnl= cur_group_id()) %>%
  ungroup() %>%
  arrange(clust_fnl) %>%
  dplyr::select(-c("clust"))
```

Identified Clusters In leaflet.

```
leaflet()%>% addProviderTiles(providers$CartoDB.Positron) %>%
  addCircleMarkers(data=xy_sp ,color='olive',radius=4,stroke=FALSE,f
  addMarkers(data=xy_sp ,clusterOptions = markerClusterOptions())
```



Unsplash images Taken from: https://unsplash.com/photos/