Municipalities dataset cleaning

This document is about how the municipality data was obtained and the steps used to clean it.

The dataset was obtained from the website of the province of Ontario (https://www.ontario.ca/page/list-ontario-municipalities#section-3).

The file obtained is a CSV but it needed to be cleaned to remove the innecesary information found in the name of each city so it can be combined with the survey dataset.

One additional step is to remove from the dataset duplicate entries. In the dataset there are some instances that appear as upper tier division (counties, regions) that have the same name of a municipality, thus causing duplicates. The ones to remove are identified below.

```
municipalities_clean[duplicated(municipalities_clean$city),]
```

```
    Municipal.status Geographic.area         city
120       Lower Tier           Essex        Essex
155       Lower Tier  Northumberland     Hamilton
313       Upper Tier           Perth        Perth
318      Single Tier    Peterborough Peterborough
339       Lower Tier         Renfrew      Renfrew
421       Lower Tier        Waterloo     Waterloo
```

R detected the "duplicate" entries because the names of the cities we are interested in sometimes appear after the entry we want to remove. After examining the original municipalities dataset (`municipalities_list.csv`), and comparing it with the list provided above, the duplicate entries in the dataset that need to be removed correspond to:

- County of Essex
- Township of Hamilton (it is an aggregation according to https://en.wikipedia.org/wiki/Hamilton_Townshi
- County of Perth
- County of Peterborough

- County of Renfrew
- Regional Municipality of Waterloo

These entries need to be removed as we are interested in municipalities within larger geographical areas, and these are aggregations that do not fit that criteria. The final step is to remove the duplicates. Because the word that identifies them as aggregations (i.e., "County", "Township") is lost in the process of getting the names of the cities to match them with the survey dataset, we will remove them by rownumber (after comparing with the `municipalities_list.csv`) dataset. The rownumbers are:

- County of Essex: 119
- Township of Hamilton: 155
- County of Perth: 313
- County of Peterborough: 317
- County of Renfrew: 338
- Regional Municipality of Waterloo: 420

```
#| echo: false


municipalities_clean <- municipalities_clean[-c(119,155,313,317,338,420),]


#save the file in csv
write.csv(municipalities_clean,here("data","municipalities_clean.csv"),row.names = FALSE)
```