Projet 2

## Introduction

Here we will study open data about politicians in Zürich and see if we can find interesting patterns.

![](https://d7whxh71cqykp.cloudfront.net/uploads/image/data/2221/04-project-zh.png)

The data comes to you in a relational database, a functional `SQLite` database to be precise.

## The data

We placed in your repository a `.db` file:

\* `zh\_politicians.db`

This file is an `SQLite` database and contains data about elected politicians in Zürich over the years.

Some columns like `activity` are in German but this won't affect your analysis.

The different tables have information about:

- The elected persons (name, gender, year of birth, year of death...)

- The mandates people have been elected for

- The addresses of the elected persons

- The affiliations to political organisations

This is a relational database, so you will probably need to do a lot of `JOINS` between tables.

Although we have seen how you can use your normal `{dplyr}` knowledge directly with databases, you will still have to write the code to connect to it and `collect()` the data you need at the right time.

Take the time to study the database structure, either directly with `R` using `{DBI}` and `{dplyr}` or first with the software we used in the course "DB Browser for SQLite".

## Exercises

### Part 1

The `MANDATES` table has information about when people were elected, when their mandate ended and on which assembly they sat.

Using a line plot, show how the number of people with an active mandate changed over the years.

You will have one line per assembly.

![](https://d7whxh71cqykp.cloudfront.net/uploads/image/data/4199/04-project-politics-sketch.png)

You can see above what it should roughly look like.

Don't bother working on exact dates: you can base your "active" years just on years, even if this creates some "dents" in the line (i.e. when looking only at active years, not precise dates, some mandates are double counted in election years since the old mandates end and the new mandates start).

### Part 2

Expand on the plot you just produced.

This time you want to show a facet charts with one chart per assembly.

In each chart, have one line for men and one line for women.

> Consider that the scaling of the axes inside each facet might not be ideal here. By default this scaling is done using the full range of the plotted data. In this case you could consider rescaling each facet by the range of the data just in that facet, using the function argument `scales`. Have a look at `?facet\_wrap()` to see what the options are.

### Part 3

Create a new plot showing the proportion of elected politicians from each party in the year `2000`.

You want to show this by assembly, so use one facet with one pie chart per assembly.

Also show your results in a table.

If you don't know how to do a pie chart, try to search on the internet how to do pie charts with `{ggplot2}`.

We will happily give you some links if you cannot find any, but one of the goals of this course is that you feel confident enough to find your own resources on the internet.

There are several ways to create nice-looking tables in R notebooks.

Have a look at the "Table Suggestions" part in the [cheatsheet](https://github.com/rstudio/cheatsheets/raw/master/rmarkdown-2.0.pdf) and choose your favorite.

### Part 4

Have another look at the composition of the assemblies: this time use a line chart to show how it changed over the years.

### Part 5

For the politicians that have a `YEAR\_OF\_DEATH` in the data, find the average life span.

Does it change if you have a `TITLE` (any title, no need to analyse every possible title separately!).

Use first a plot, and \*\*then a statistical test\*\* to assess if the average life span is different for politicians having a `TITLE` compared to those without any `TITLE`.

Before running the statistical test, write down what you expect the statistical results to look like on the basis of what you see in the plot.

Does your expectation match the statistical test results?

### Part 6

Create a new variable which splits the politicians into 2 subgroups: one subgroup of politicians who were `BORN\_BEFORE\_1918` (i.e. for which `YEAR\_OF\_BIRTH < 1918`), and another subgroup of politicians who were `BORN\_AFTER\_1918` (i.e. for which `YEAR\_OF\_BIRTH >= 1918`).

Reply to the questions asked in the previous part (i.e. graphical comparison \*\*and\*\* statistical test) for each of the 2 subgroups.

I.e. you need to create two visualisations and run two t-tests to answer:

\*is there a difference between politicians with and without title\* in the subgroup born before 1918, and in the subgroup after 1918?

Again, make sure to write down your expectation on the basis of the visualisations.

Is there a difference in results between the subgroup before 1918 and those born on or after 1918?

### Part 7

Which politicians have had the most mandates?

Create a top 10 horizontal bar chart.

### Part 8

Do some politicians have multiple mandates at the same time?

You will need to find a way to check if some mandates started before the previous one of the same politician ended.

### Part 9

Have some politicians been affiliated to different parties over the years?

### Part 10

Take a sample of 20 politicians with a listed address and plot them on a map with `{leaflet}` or `{ggmap}`.

You will need to use an API that converts the addresses to geocoordinates.

> If you decide to use `{leaflet}`, please note that knitting to anything other than a `html\_document` may lead to issues.

### A little reminder

Congratulations,you have now completed the second final project.

At this point, we strongly encourage you to do an informal review of the first two projects, which you can do simply by requesting a 1-1 with an instructor. This will allow you to get some initial feedback as well as an idea if your reports are on the right track.