Covid dataset

* 2 different sets: till 2021, from 2021
* Both load in 2 lists 🡪 all the data is made in safety/security regions, but we have to convert them to province / province group

Formatting data

* merge the datasets together
* group the data by year and month so that we can later compare this to the other data set on the same time scale
* we filter all the relevant columns out of the data 🡪 (speaks for itself) 🡪 security region code (year, month, security region code (to see what province it actually happens)
* drop irrelevant data to make the code more efficient

1st real process:

* making the provinces and provincegroup 🡪 foud through looping through the whole dataset, i2 & i4 are the columns we want to change we put the new data there which we will use. We take the code from the certain row, region code, we strip the text VR from it (so it only contains numbers, then we change that into an integer and do that number -1 so we get an index number for a list). (we could have also created a dictionary but we realized that later, in that way we didn t need to strip the data).
* We then rename the column for name clarity
* At last we grouped the values by province so that we can determine the
* The outcome is an overview of the hospitalized covid cases per month per province.

2nd process – travel dataset

* The dataset is different because it contains metadata, code that stands for outcomes so we first have to convert that into usable data.
* 1st scrap all the irrelevant data: data that is not about provincegroups
* We do that by checking if a certain region cchteriatic has LD in it.
  + If LD is not in the unit we append a new array: true, if it is: false. By that we filter the data.
* We have the data frame with true/false as outcome, of which we deleted the true ones.
* The same is applied for every trip characteristic with the ID 2031, which means that it is talking about months/days: we want to keep the monthly data (so days will be removed).
* After that we also remove the data of which after 2031 the number lower than 160 follows: because that stands for days (instead of months).
* Remove all the data from margins that have no code: MW00000, because that means that this data doesn’t include all the data we need (upper/lowerbound etc.).
* We also remove populationkey: A048709, because this
* Lastly we also remove all travel keys that do not include T001080 \*\* Incorrect
* And we rename values in region characteristics and trip char. So they state the province group & month instead of values.
* Lastly we remove the columns of ID, population, and margins.