More on Multivariate Regressions

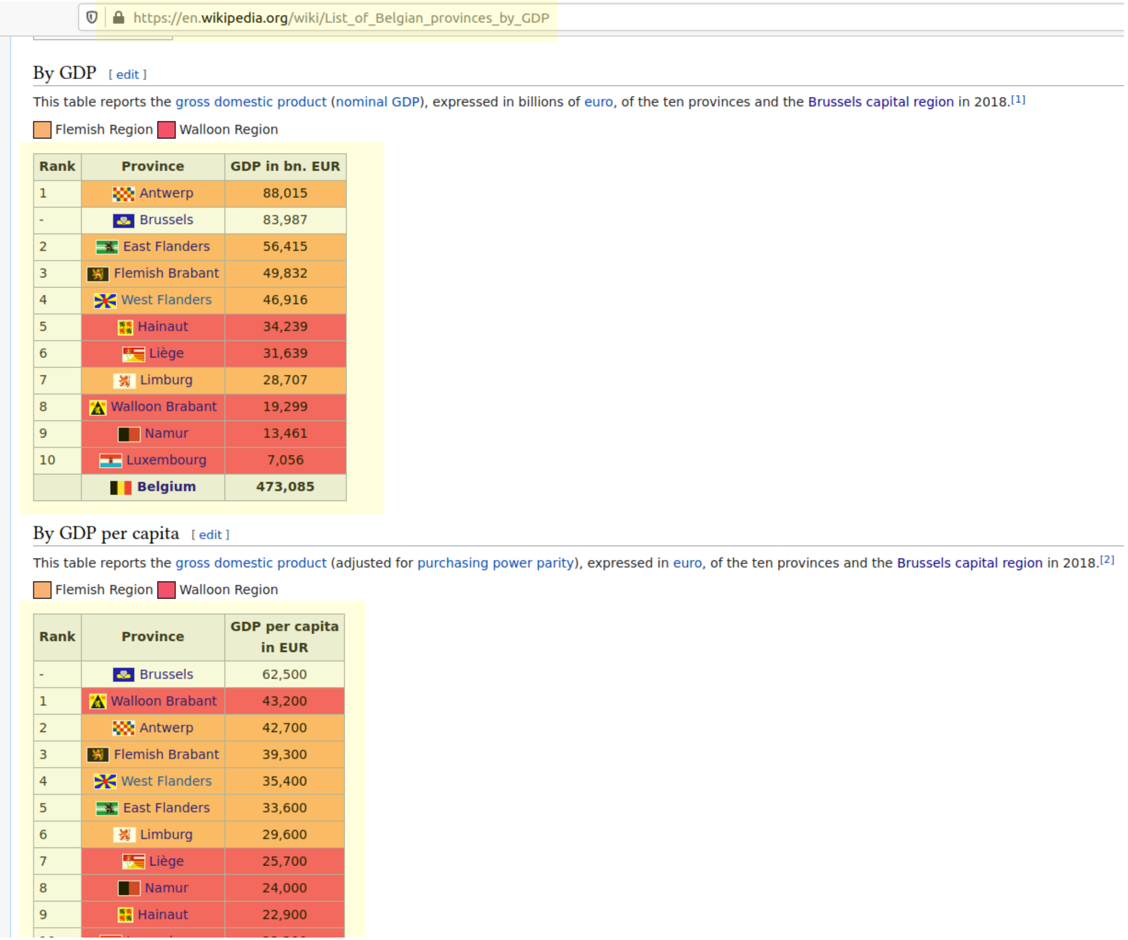
Michele Coscia

First Year Project #2

March 12th, 2021

Looking for data

- Wikipedia is always a great bet...



https://en.wikipedia.org/wiki/List_of_Belgian_provinces_by_GDP

By GDP [edit]

This table reports the gross domestic product (nominal GDP), expressed in billions of euro, of the ten provinces and the Brussels capital region in 2018.[1]

🟧 Flemish Region 🟥 Walloon Region

Rank	Province	GDP in bn. EUR
1	Antwerp	88,015
-	Brussels	83,987
2	East Flanders	56,415
3	Flemish Brabant	49,832
4	West Flanders	46,916
5	Hainaut	34,239
6	Liège	31,639
7	Limburg	28,707
8	Walloon Brabant	19,299
9	Namur	13,461
10	Luxembourg	7,056
	Belgium	**473,085**

By GDP per capita [edit]

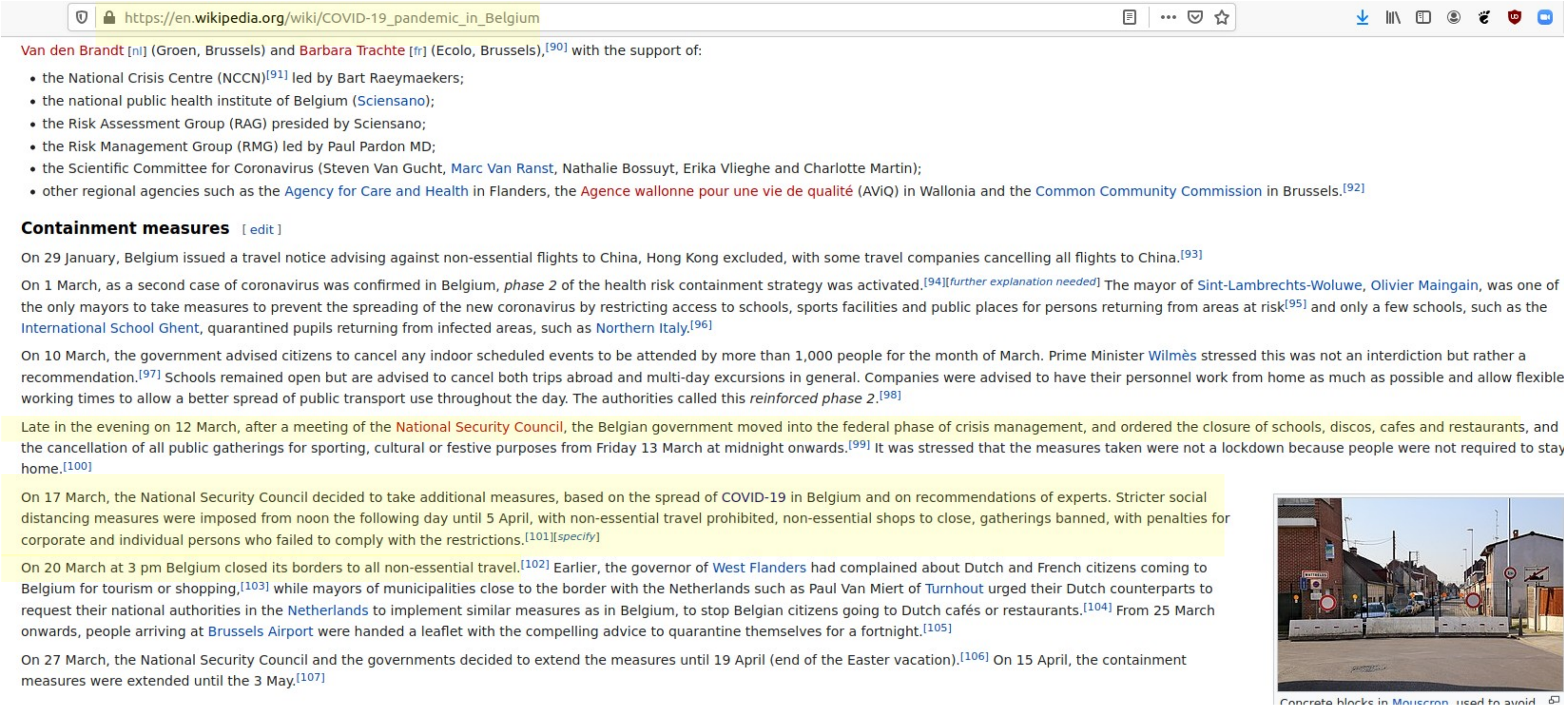
This table reports the gross domestic product (adjusted for purchasing power parity), expressed in euro, of the ten provinces and the Brussels capital region in 2018.[2]

🟧 Flemish Region 🟥 Walloon Region

Rank	Province	GDP per capita in EUR
-	Brussels	62,500
1	Walloon Brabant	43,200
2	Antwerp	42,700
3	Flemish Brabant	39,300
4	West Flanders	35,400
5	East Flanders	33,600
6	Limburg	29,600
7	Liège	25,700
8	Namur	24,000
9	Hainaut	22,900

Looking for data

- ...sometimes in creative ways!



https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Belgium

Van den Brandt [nl] (Groen, Brussels) and Barbara Trachte [fr] (Ecolo, Brussels),[90] with the support of:

- the National Crisis Centre (NCCN)[91] led by Bart Raeymaekers;
- the national public health institute of Belgium (Sciensano);
- the Risk Assessment Group (RAG) presided by Sciensano;
- the Risk Management Group (RMG) led by Paul Pardon MD;
- the Scientific Committee for Coronavirus (Steven Van Gucht, Marc Van Ranst, Nathalie Bossuyt, Erika Vlieghe and Charlotte Martin);
- other regional agencies such as the Agency for Care and Health in Flanders, the Agence wallonne pour une vie de qualité (AViQ) in Wallonia and the Common Community Commission in Brussels.[92]

Containment measures [edit]

On 29 January, Belgium issued a travel notice advising against non-essential flights to China, Hong Kong excluded, with some travel companies cancelling all flights to China.[93]

On 1 March, as a second case of coronavirus was confirmed in Belgium, *phase 2* of the health risk containment strategy was activated.[94][*further explanation needed*] The mayor of Sint-Lambrechts-Woluwe, Olivier Maingain, was one of the only mayors to take measures to prevent the spreading of the new coronavirus by restricting access to schools, sports facilities and public places for persons returning from areas at risk[95] and only a few schools, such as the International School Ghent, quarantined pupils returning from infected areas, such as Northern Italy.[96]

On 10 March, the government advised citizens to cancel any indoor scheduled events to be attended by more than 1,000 people for the month of March. Prime Minister Wilmès stressed this was not an interdiction but rather a recommendation.[97] Schools remained open but are advised to cancel both trips abroad and multi-day excursions in general. Companies were advised to have their personnel work from home as much as possible and allow flexible working times to allow a better spread of public transport use throughout the day. The authorities called this *reinforced phase 2*.[98]

Late in the evening on 12 March, after a meeting of the National Security Council, the Belgian government moved into the federal phase of crisis management, and ordered the closure of schools, discos, cafes and restaurants, and the cancellation of all public gatherings for sporting, cultural or festive purposes from Friday 13 March at midnight onwards.[99] It was stressed that the measures taken were not a lockdown because people were not required to stay home.[100]

On 17 March, the National Security Council decided to take additional measures, based on the spread of COVID-19 in Belgium and on recommendations of experts. Stricter social distancing measures were imposed from noon the following day until 5 April, with non-essential travel prohibited, non-essential shops to close, gatherings banned, with penalties for corporate and individual persons who failed to comply with the restrictions.[101][*specify*]

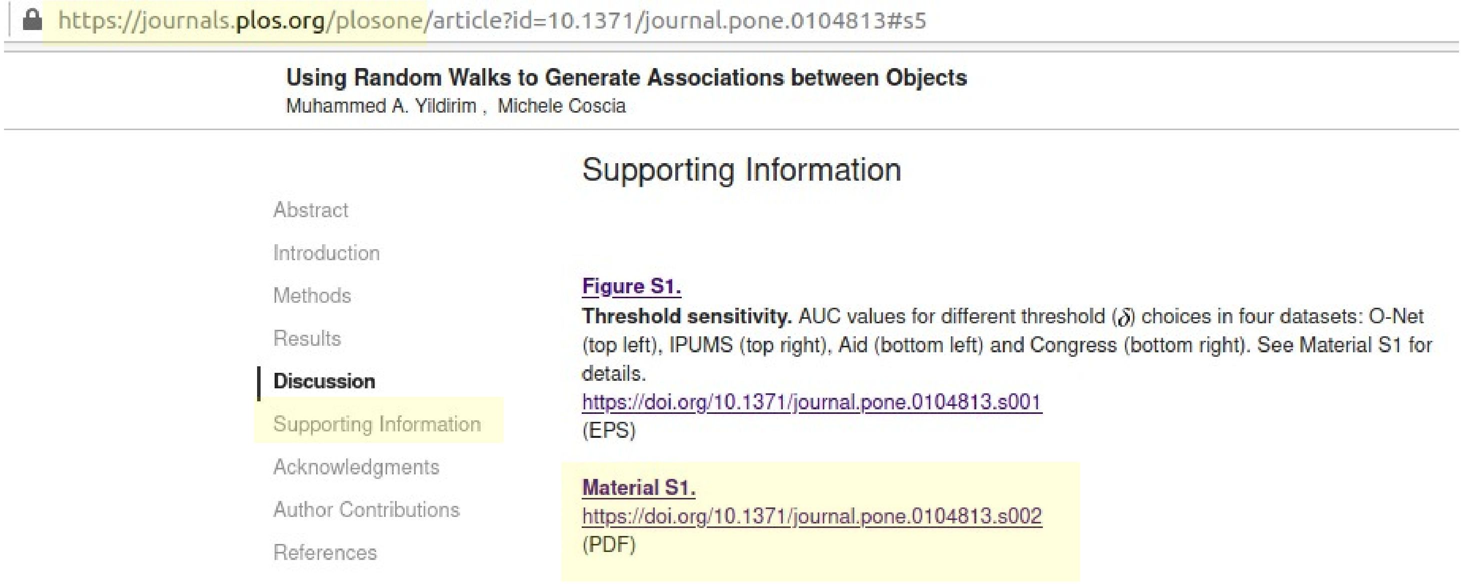
On 20 March at 3 pm Belgium closed its borders to all non-essential travel.[102] Earlier, the governor of West Flanders had complained about Dutch and French citizens coming to Belgium for tourism or shopping,[103] while mayors of municipalities close to the border with the Netherlands such as Paul Van Miert of Turnhout urged their Dutch counterparts to request their national authorities in the Netherlands to implement similar measures as in Belgium, to stop Belgian citizens going to Dutch cafés or restaurants.[104] From 25 March onwards, people arriving at Brussels Airport were handed a leaflet with the compelling advice to quarantine themselves for a fortnight.[105]

On 27 March, the National Security Council and the governments decided to extend the measures until 19 April (end of the Easter vacation).[106] On 15 April, the containment measures were extended until the 3 May.[107]

Concrete blocks in Mouscron, used to avoid

Looking for data

- Some journals mandate authors to share their data publicly: Plos One, Nature Human Behaviour, …



https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0104813#s5

Using Random Walks to Generate Associations between Objects
Muhammed A. Yildirim , Michele Coscia

Supporting Information

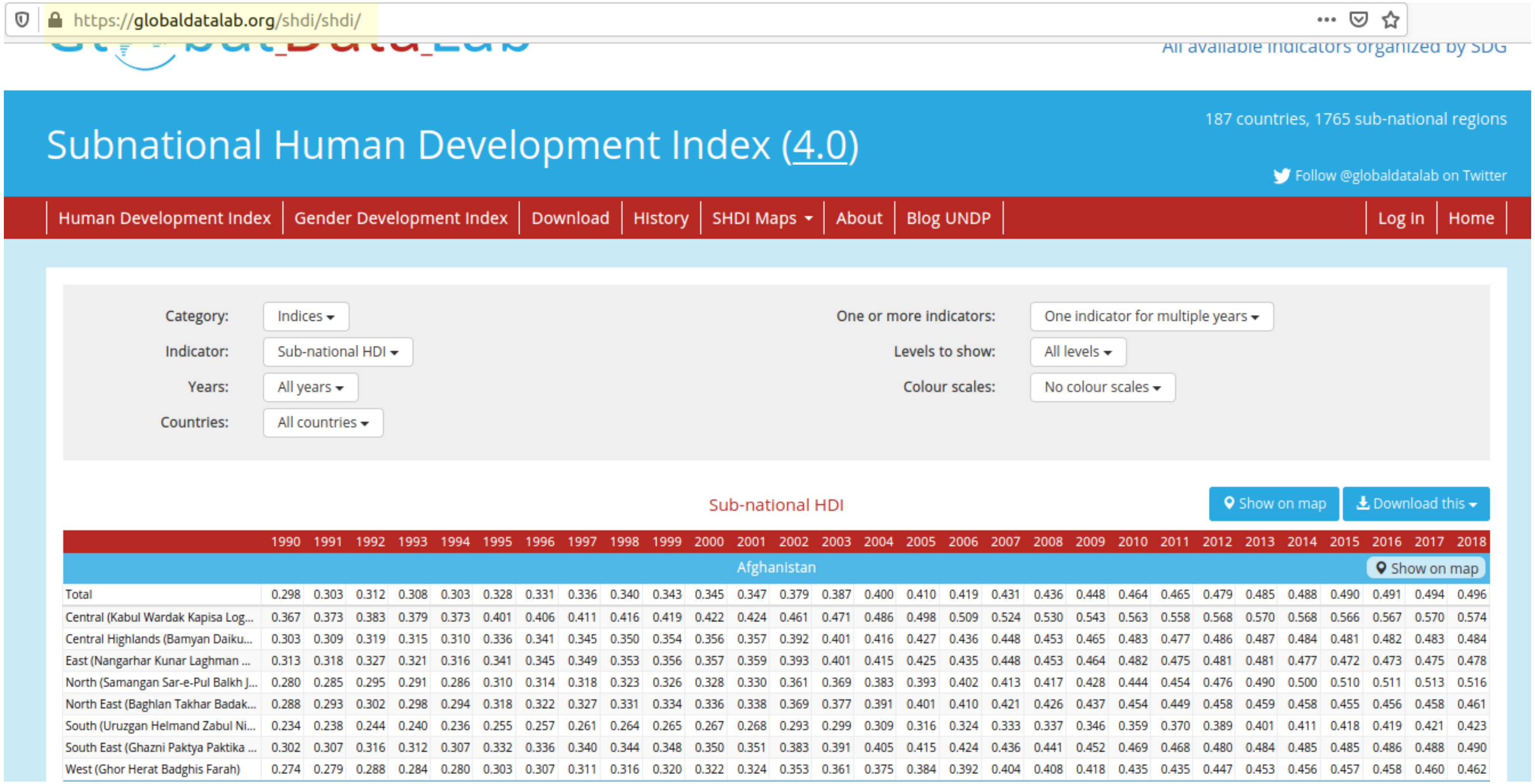
Figure S1.

Threshold sensitivity. AUC values for different threshold (δ) choices in four datasets: O-Net (top left), IPUMS (top right), Aid (bottom left) and Congress (bottom right). See Material S1 for details.
https://doi.org/10.1371/journal.pone.0104813.s001
(EPS)

Material S1.
https://doi.org/10.1371/journal.pone.0104813.s002
(PDF)

Looking for data

- International institutions are waking up!
- Hit the World Bank, European agencies, …
- Example, HDI by region:



https://globaldatalab.org/shdi/shdi/
All available indicators organized by SDG

Subnational Human Development Index (4.0)

187 countries, 1765 sub-national regions

Category: Indices

Indicator: Sub-national HDI

Years: All years

Countries: All countries

One or more indicators: One indicator for multiple years

Levels to show: All levels

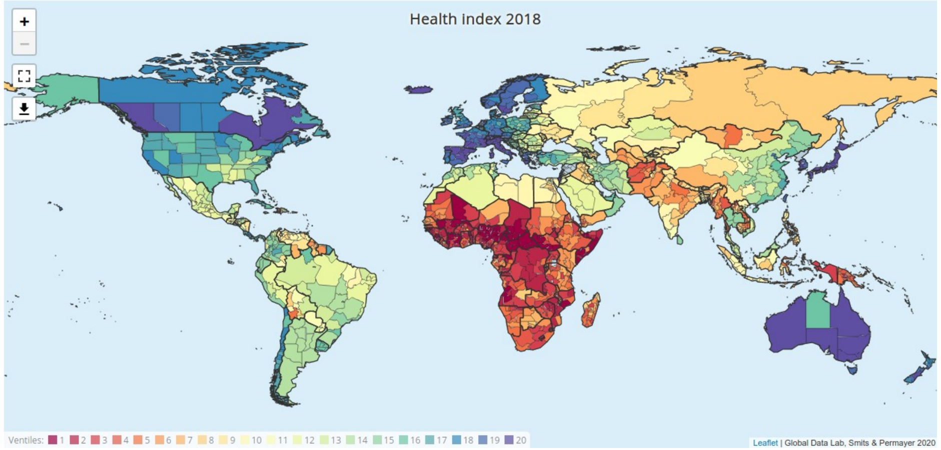
Colour scales: No colour scales

Sub-national HDI

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Afghanistan																													Show on map
Total	0.298	0.303	0.312	0.308	0.303	0.328	0.331	0.336	0.340	0.343	0.345	0.347	0.379	0.387	0.400	0.410	0.419	0.431	0.436	0.448	0.464	0.465	0.479	0.485	0.488	0.490	0.491	0.494	0.496
Central (Kabul Wardak Kapisa Log...	0.367	0.373	0.383	0.379	0.373	0.401	0.406	0.411	0.416	0.419	0.422	0.424	0.461	0.471	0.486	0.498	0.509	0.524	0.530	0.543	0.563	0.558	0.568	0.570	0.568	0.566	0.567	0.570	0.574
Central Highlands (Bamyan Daiku...	0.303	0.309	0.319	0.315	0.310	0.336	0.341	0.345	0.350	0.354	0.356	0.357	0.392	0.401	0.416	0.427	0.436	0.448	0.453	0.465	0.483	0.477	0.486	0.487	0.484	0.481	0.482	0.483	0.484
East (Nangarhar Kunar Laghman ...	0.313	0.318	0.327	0.321	0.316	0.341	0.345	0.349	0.353	0.356	0.357	0.359	0.393	0.401	0.415	0.425	0.435	0.448	0.453	0.464	0.482	0.475	0.481	0.481	0.477	0.472	0.473	0.475	0.478
North (Samangan Sar-e-Pul Balkh J...	0.280	0.285	0.295	0.291	0.286	0.310	0.314	0.318	0.323	0.326	0.328	0.330	0.361	0.369	0.383	0.393	0.402	0.413	0.417	0.428	0.444	0.454	0.476	0.490	0.500	0.510	0.511	0.513	0.516
North East (Baghlan Takhar Badak...	0.288	0.293	0.302	0.298	0.294	0.318	0.322	0.327	0.331	0.334	0.336	0.338	0.369	0.377	0.391	0.401	0.410	0.421	0.426	0.437	0.454	0.449	0.458	0.459	0.458	0.455	0.456	0.458	0.461
South (Uruzgan Helmand Zabul Ni...	0.234	0.238	0.244	0.240	0.236	0.255	0.257	0.261	0.264	0.265	0.267	0.268	0.293	0.299	0.309	0.316	0.324	0.333	0.337	0.346	0.359	0.370	0.389	0.401	0.411	0.418	0.419	0.421	0.423
South East (Ghazni Paktya Paktika ...	0.302	0.307	0.316	0.312	0.307	0.332	0.336	0.340	0.344	0.348	0.350	0.351	0.383	0.391	0.405	0.415	0.424	0.436	0.441	0.452	0.469	0.468	0.480	0.484	0.485	0.485	0.486	0.488	0.490
West (Ghor Herat Badghis Farah)	0.274	0.279	0.288	0.284	0.280	0.303	0.307	0.311	0.316	0.320	0.322	0.324	0.353	0.361	0.375	0.384	0.392	0.404	0.408	0.418	0.435	0.435	0.447	0.453	0.456	0.457	0.458	0.460	0.462

Global Data Lab

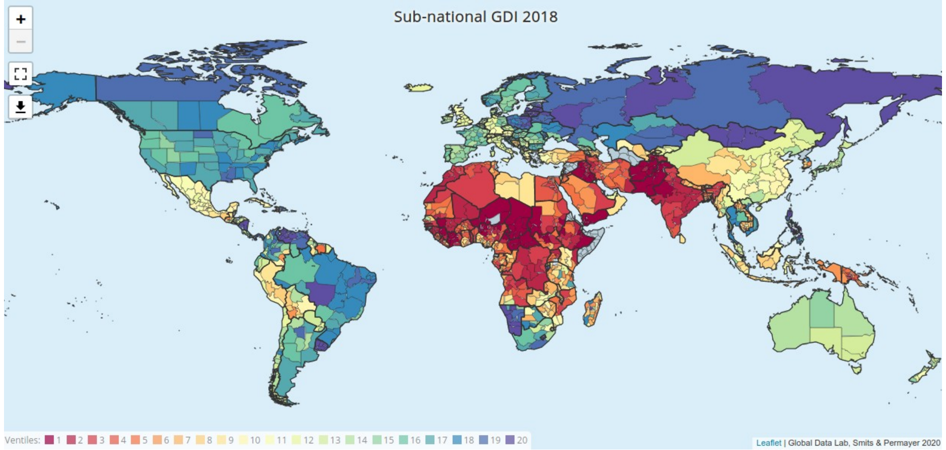
- Health index
- Gender equality
- And more!



Health index 2018

+
–
Ventiles: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Leaflet | Global Data Lab, Smits & Permayer 2020



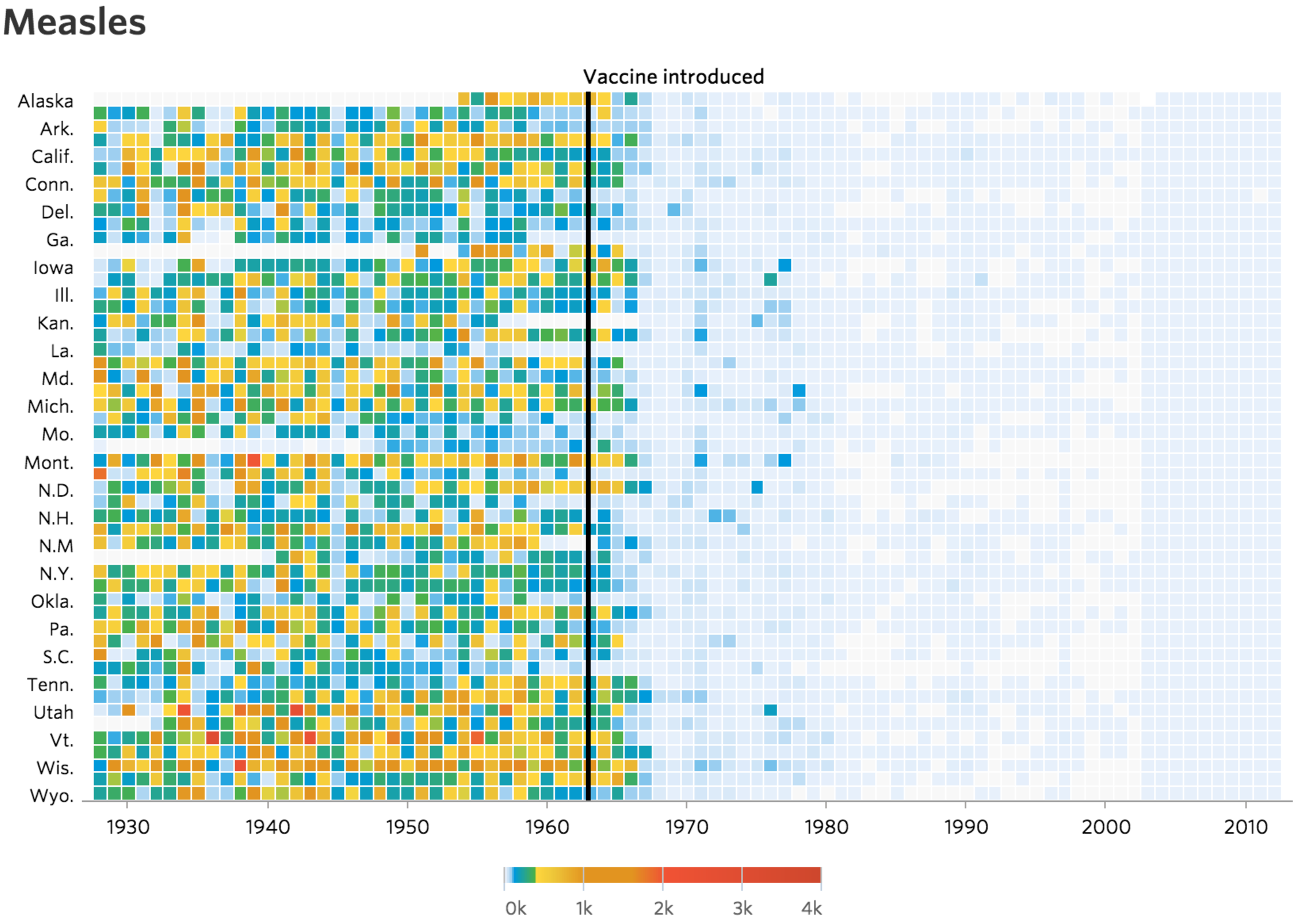
Sub-national GDI 2018

+
–
Ventiles: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Leaflet | Global Data Lab, Smits & Permayer 2020

Intervention Effects

The Ideal Scenario



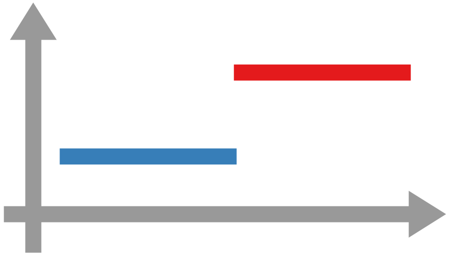
It’s not that easy...

- Did the lockdown work?
- Too many things happen at the same time:
  - New variants
  - Ebb & flow of epidemics
  - Individual & collective behavior
  - Influence of weather
    - Which interacts with collective behavior
  - Vaccine development

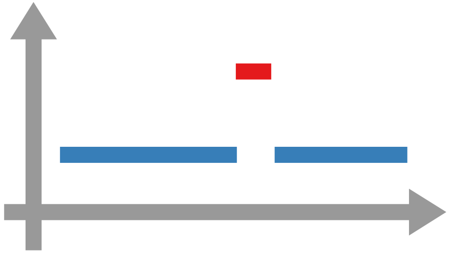
It’s not that easy...

- Also: what type of effect did the intervention cause?

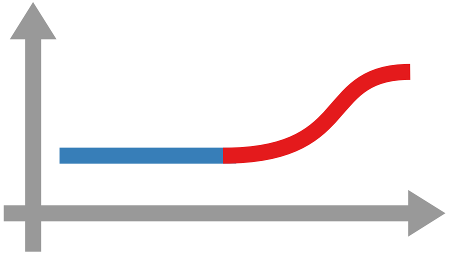
Before intervention After intervention



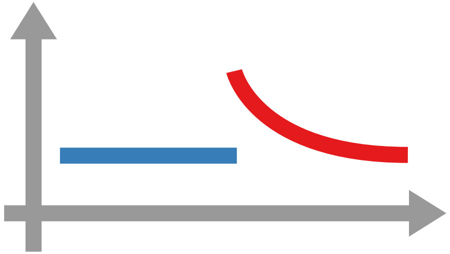
Immediate & sustained



Immediate & temporary



Warm up & sustained

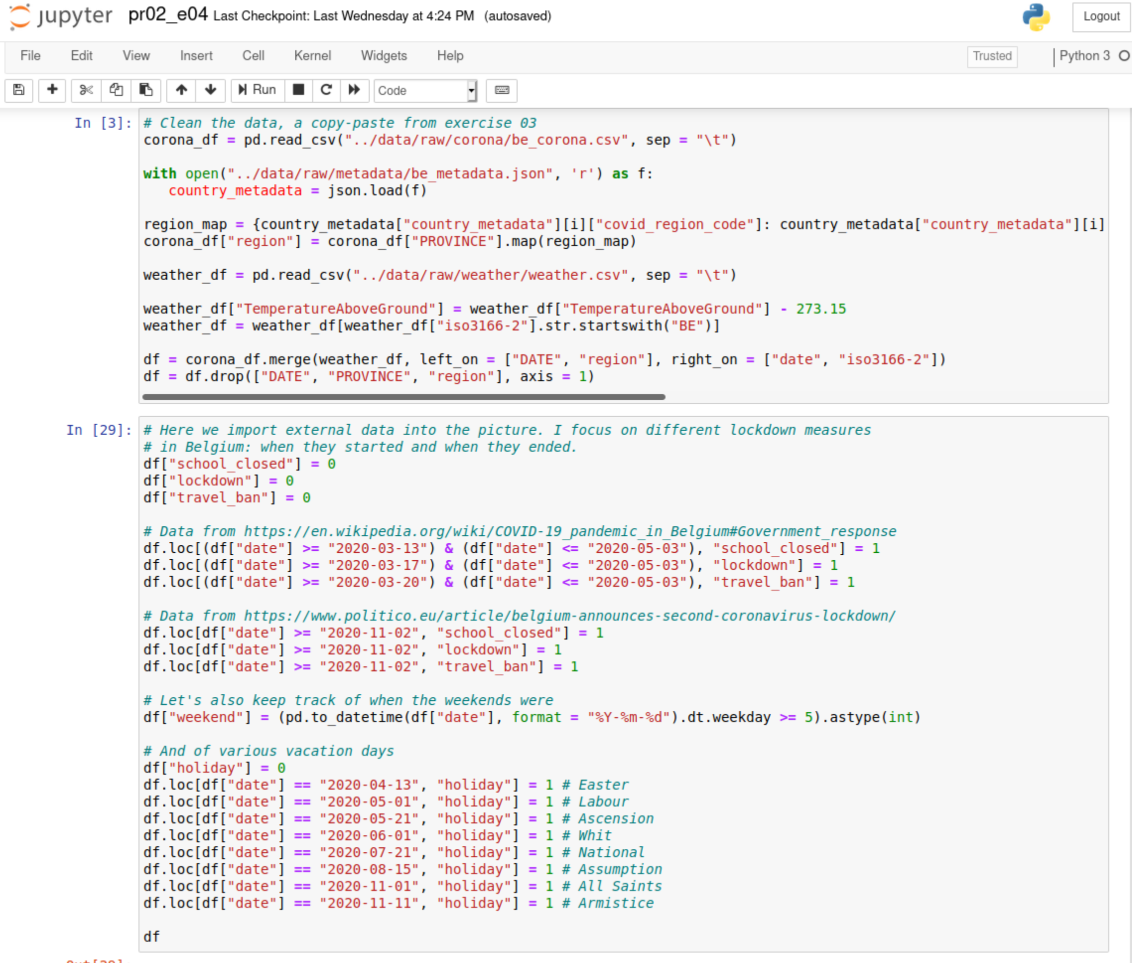


Immediate & cool down

Today

- Learning not how to do it properly...
- … but why it’s so tricky!
- Get data about lockdown
- Show how it correlates with weather

Practicalities



In [3]:

```
# Clean the data, a copy-paste from exercise 03
corona_df = pd.read_csv("../data/raw/corona/be_corona.csv", sep = "\t")

with open("../data/raw/metadata/be_metadata.json", 'r') as f:
   country_metadata = json.load(f)

region_map = {country_metadata["country_metadata"][i]["covid_region_code"]: country_metadata["country_metadata"][i]
corona_df["region"] = corona_df["PROVINCE"].map(region_map)

weather_df = pd.read_csv("../data/raw/weather/weather.csv", sep = "\t")

weather_df["TemperatureAboveGround"] = weather_df["TemperatureAboveGround"] - 273.15
weather_df = weather_df[weather_df["iso3166-2"].str.startswith("BE")]

df = corona_df.merge(weather_df, left_on = ["DATE", "region"], right_on = ["date", "iso3166-2"])
df = df.drop(["DATE", "PROVINCE", "region"], axis = 1)
```

In [29]:

```
# Here we import external data into the picture. I focus on different lockdown measures
# in Belgium: when they started and when they ended.
df["school_closed"] = 0
df["lockdown"] = 0
df["travel_ban"] = 0

# Data from https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Belgium#Government_response
df.loc[(df["date"] >= "2020-03-13") & (df["date"] <= "2020-05-03"), "school_closed"] = 1
df.loc[(df["date"] >= "2020-03-17") & (df["date"] <= "2020-05-03"), "lockdown"] = 1
df.loc[(df["date"] >= "2020-03-20") & (df["date"] <= "2020-05-03"), "travel_ban"] = 1

# Data from https://www.politico.eu/article/belgium-announces-second-coronavirus-lockdown/
df.loc[df["date"] >= "2020-11-02", "school_closed"] = 1
df.loc[df["date"] >= "2020-11-02", "lockdown"] = 1
df.loc[df["date"] >= "2020-11-02", "travel_ban"] = 1

# Let's also keep track of when the weekends were
df["weekend"] = (pd.to_datetime(df["date"], format = "%Y-%m-%d").dt.weekday >= 5).astype(int)

# And of various vacation days
df["holiday"] = 0
df.loc[df["date"] == "2020-04-13", "holiday"] = 1 # Easter
df.loc[df["date"] == "2020-05-01", "holiday"] = 1 # Labour
df.loc[df["date"] == "2020-05-21", "holiday"] = 1 # Ascension
df.loc[df["date"] == "2020-06-01", "holiday"] = 1 # Whit
df.loc[df["date"] == "2020-07-21", "holiday"] = 1 # National
df.loc[df["date"] == "2020-08-15", "holiday"] = 1 # Assumption
df.loc[df["date"] == "2020-11-01", "holiday"] = 1 # All Saints
df.loc[df["date"] == "2020-11-11", "holiday"] = 1 # Armistice

df
```

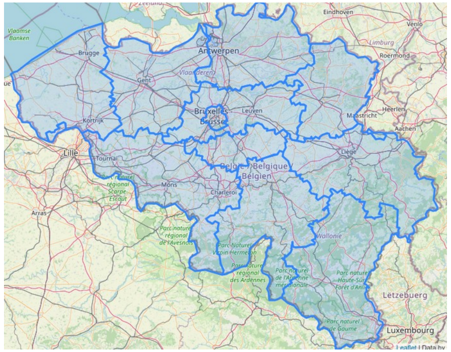
Fixed Effects
(Dummy Variables)

Problem

- Sometimes you **know** something affected your outcome
- You just don’t have any measure for it
- In our case: different local governments work differently

Fixed Effects

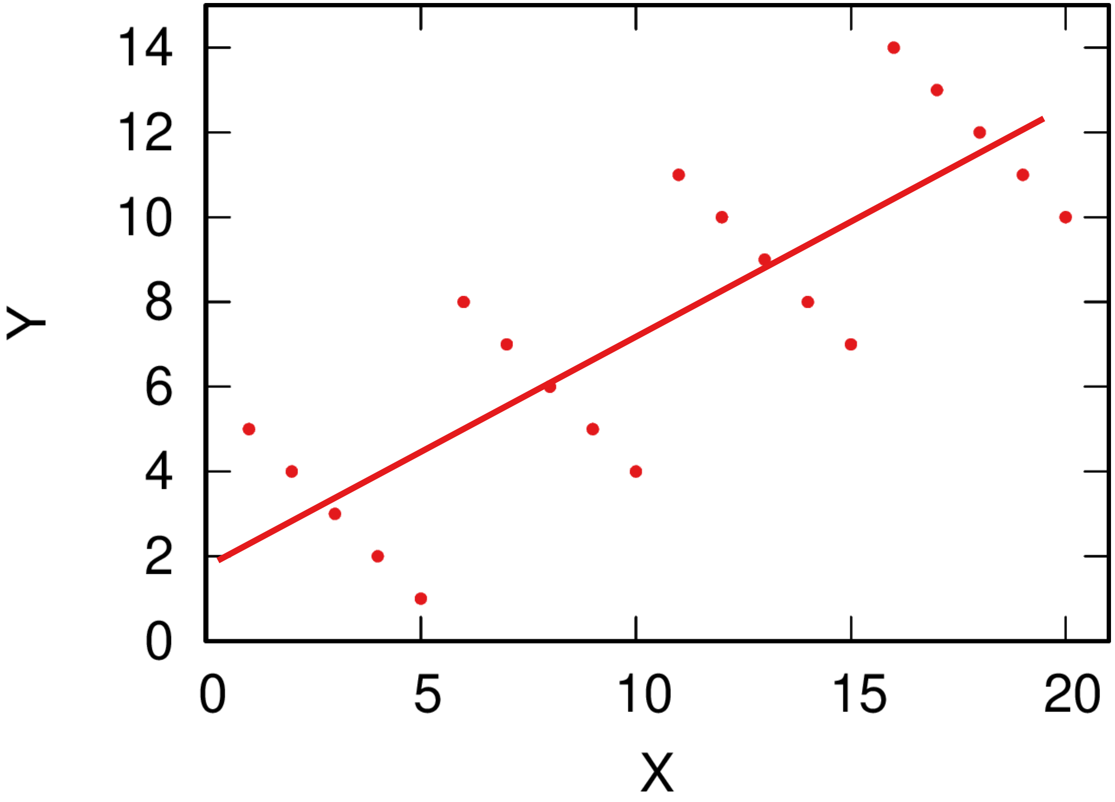
- You know your observations belong to specific groups
  - In our case, Belgian regions


Vlaamse Banken
Eindhoven
Venlo
Limburg
Roermond
Brugge
Antwerpen
Gent
Vlaanderen
Bruxelles
Brussel
Leuven
Kortrijk
Heerlen
Maastricht
Aachen
Lille
Tournai
Liège
België / Belgique / Belgien
Parc naturel régional Scarpe-Escaut
Mons
Charleroi
Arras
Parc naturel régional de l'Avesnois
Wallonie
Parc Naturel Viroin Hermeton
Parc naturel régional des Ardennes
Parc naturel de l'Ardenne méridionale
Parc naturel Haute-Sûre Forêt d'Anlier
Lëtzebuerg
Parc naturel de Gaume
Luxembourg
Leaflet | Data by

- The avg of each group is fixed
- Everything that group does differently from the other groups is captured here

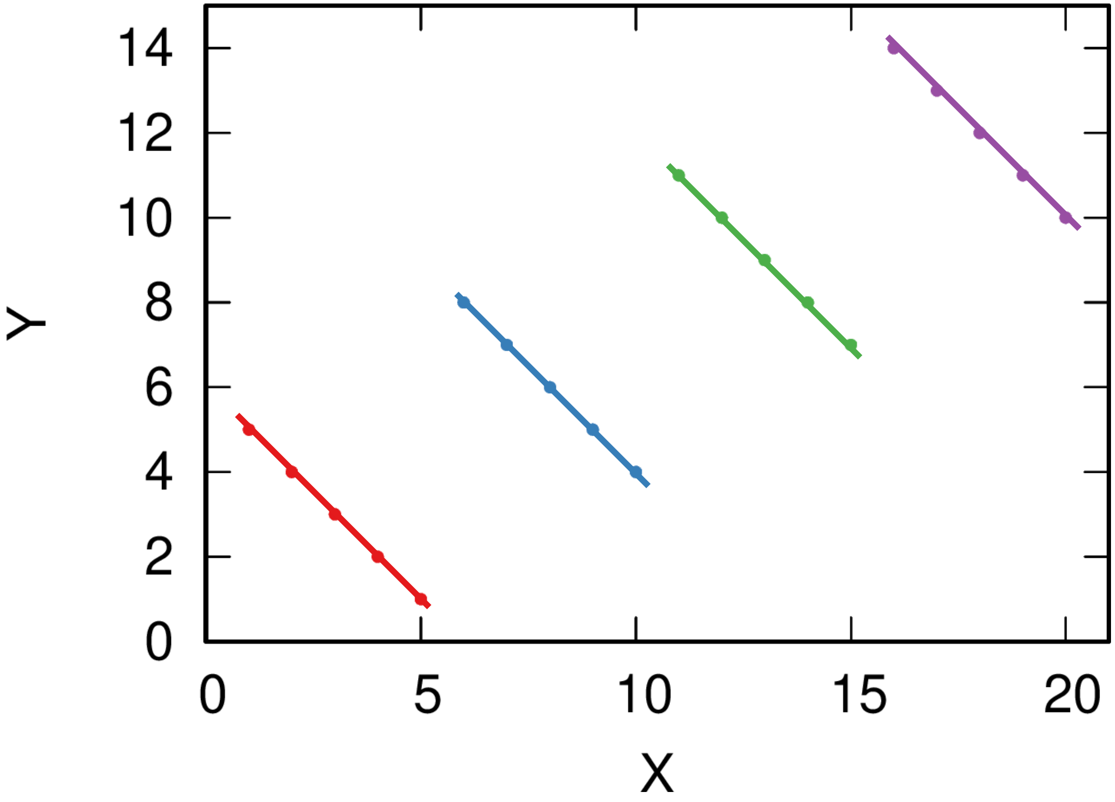
Why would we do this?

- Corr ~ 0.8!!
- Best fit
- Something Fishy...



Why would we do this?

- Groups!
- Related with X
- The true relationship is actually negative!



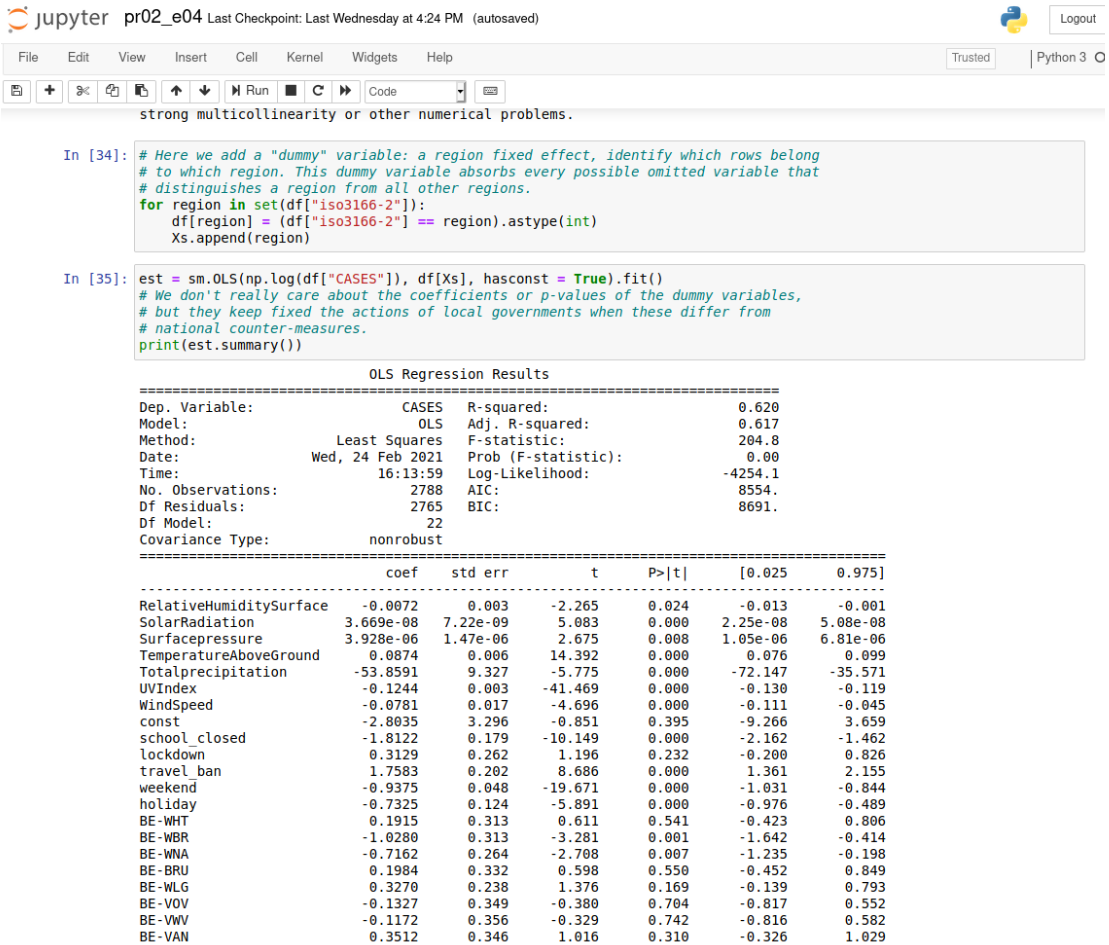
How to do it practically

- In R, it’s automatic
  - Just pass a categorical variable to your regression function
- In general, you can add a “dummy variable”
  - One variable per group
  - 1 if observation belongs to the group, 0 otherwise
  - You need to omit one group (the reference)

Interpretation

- Coefficient tells you the effect of being part of the group
  - Specifically: the difference between your group and the reference one
- If group membership is important for your question, you can interpret it
  - But careful, because you’re absorbing everything!
- Most often, it’s just a control → Ignore

Practicalities



```
strong multicollinearity or other numerical problems.
```

In [34]:

```python
# Here we add a "dummy" variable: a region fixed effect, identify which rows belong
# to which region. This dummy variable absorbs every possible omitted variable that
# distinguishes a region from all other regions.
for region in set(df["iso3166-2"]):
    df[region] = (df["iso3166-2"] == region).astype(int)
    Xs.append(region)
```

In [35]:

```python
est = sm.OLS(np.log(df["CASES"]), df[Xs], hasconst = True).fit()
# We don't really care about the coefficients or p-values of the dummy variables,
# but they keep fixed the actions of local governments when these differ from
# national counter-measures.
print(est.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  CASES   R-squared:                       0.620
Model:                            OLS   Adj. R-squared:                  0.617
Method:                 Least Squares   F-statistic:                     204.8
Date:                Wed, 24 Feb 2021   Prob (F-statistic):               0.00
Time:                        16:13:59   Log-Likelihood:                -4254.1
No. Observations:                2788   AIC:                             8554.
Df Residuals:                    2765   BIC:                             8691.
Df Model:                          22
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
RelativeHumiditySurface    -0.0072      0.003     -2.265      0.024      -0.013      -0.001
SolarRadiation           3.669e-08   7.22e-09      5.083      0.000    2.25e-08    5.08e-08
Surfacepressure          3.928e-06   1.47e-06      2.675      0.008    1.05e-06    6.81e-06
TemperatureAboveGround      0.0874      0.006     14.392      0.000       0.076       0.099
Totalprecipitation        -53.8591      9.327     -5.775      0.000     -72.147     -35.571
UVIndex                    -0.1244      0.003    -41.469      0.000      -0.130      -0.119
WindSpeed                  -0.0781      0.017     -4.696      0.000      -0.111      -0.045
const                      -2.8035      3.296     -0.851      0.395      -9.266       3.659
school_closed              -1.8122      0.179    -10.149      0.000      -2.162      -1.462
lockdown                    0.3129      0.262      1.196      0.232      -0.200       0.826
travel_ban                  1.7583      0.202      8.686      0.000       1.361       2.155
weekend                    -0.9375      0.048    -19.671      0.000      -1.031      -0.844
holiday                    -0.7325      0.124     -5.891      0.000      -0.976      -0.489
BE-WHT                      0.1915      0.313      0.611      0.541      -0.423       0.806
BE-WBR                     -1.0280      0.313     -3.281      0.001      -1.642      -0.414
BE-WNA                     -0.7162      0.264     -2.708      0.007      -1.235      -0.198
BE-BRU                      0.1984      0.332      0.598      0.550      -0.452       0.849
BE-WLG                      0.3270      0.238      1.376      0.169      -0.139       0.793
BE-VOV                     -0.1327      0.349     -0.380      0.704      -0.817       0.552
BE-VWV                     -0.1172      0.356     -0.329      0.742      -0.816       0.582
BE-VAN                      0.3512      0.346      1.016      0.310      -0.326       1.029
```