Project Data thief

Team: Pandas (Bita Najd, Thomas Bentler, Jens Kleiber)
Date: 10.04.2020

Disclaimer

As students of the Data Analytics Bootcamp of the cohort 03/2020 the project “Data Thief” has the goal to apply all acquired skills of the first module. The task consists of choosing a topic and find all the relevant data by connecting to an API, finding a dataset or scraping data from the web. These data then must organized, cleaned, analyzed and presented.

Following presentations shows the result of the team Pandas.

Question

Are there correlations between socio-economic topics and deaths caused by Corona virus in Europe?

Workflow

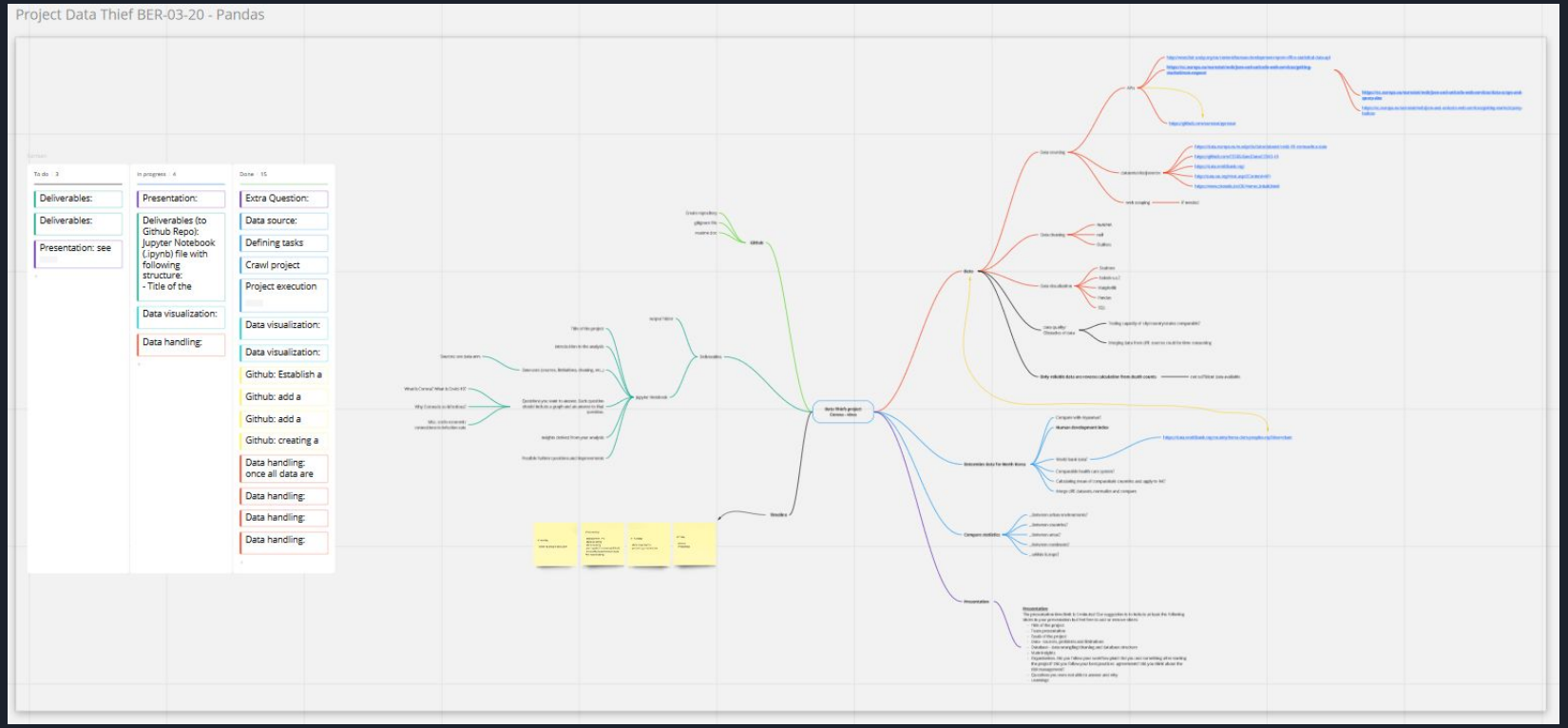
1. Project management
2. Data sourcing - API, file download
3. Data handling - MySQL, Github, Python
4. Re-evaluating questions - focusing
5. Preparing presentation

1. Project management

We used following tools for online collaboration:

- Zoom: Breakout room
- Slack: Sharing misc.
- Miro: MindMap, Kanban-board, Time schedule
- G-sheets: Visualization of tables during discussions
- G-Presentations: Preparing presentation
- Github: Collecting data, Jupyter Notebooks, sharing code
- MySQL database: Visualizing data for inspection
- Jupyter Notebook: Python, Matplotlib, Pandas

1. Project management - Miro



2. Data sourcing

API - connection:

→ EUROSTAT (Statistical Office of the European Union)

API - download of json-files:

→ WHO (World Health Organization)

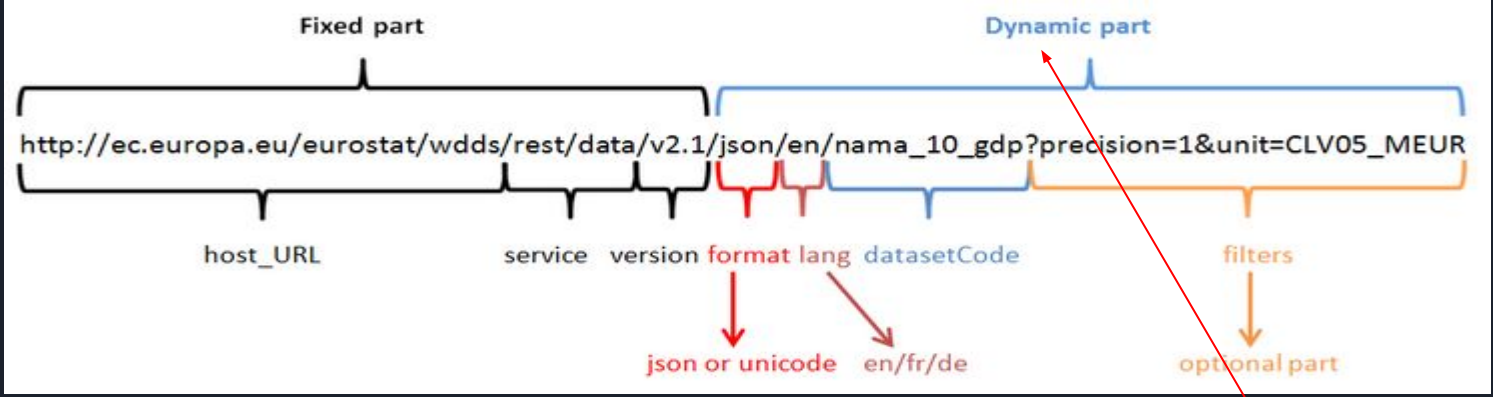
Direct download of csv-files:

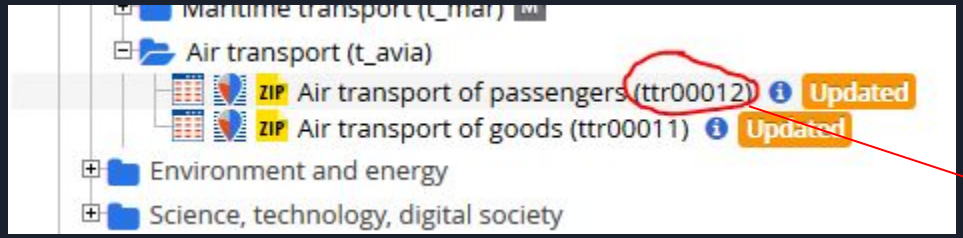
→ Kaggle, Robert-Koch-Institut

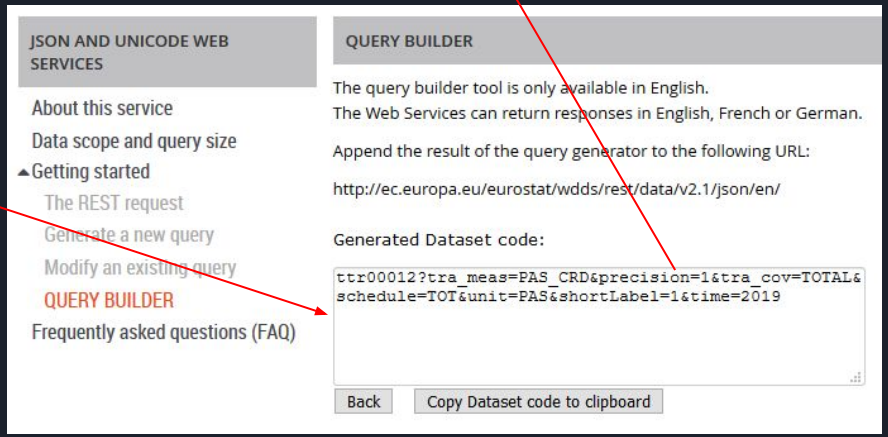
Handmade table:

→ Human development index (HDI)

2. Data sourcing - API - EUROSTAT







2. Data sourcing - API - EUROSTAT

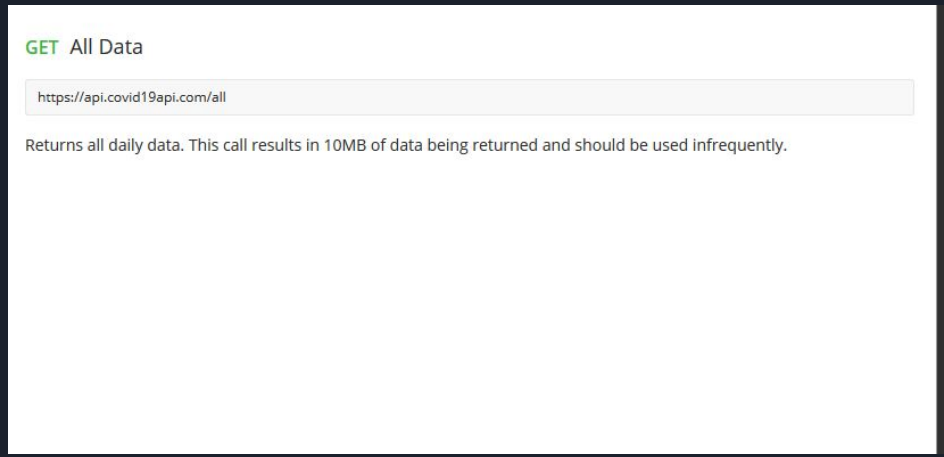
```
Eurostat_title =  "Tourism"
json_RQST = 'tin00173?precision=1&unit=PC&accommod=BEDPL&shortLabel=1'

#Eurostat_title =  "Eurostat population density 2018"
#json_RQST = "tps00003?unit=PER_KM2&precision=1&time=2018"

#Eurostat_title =  "Eurostat air transport passengers 2018"
#json_RQST = "ttr00012?tra_meas=PAS_CRD&precision=1&tra_cov=TOTAL&schedule=TOT&unit=PAS&time=2018"

#Eurostat_title =  "Eurostat gdp/head 2018"
#json_RQST = "nama_10_pc?na_item=B1GQ&precision=1&unit=CLV10_EUR_HAB&time=2018"

#Eurostat_title =  "Eurostat health euro expenditure per capita and totals 2017"
#json_RQST = "tps00207?precision=1&unit=EUR_HAB&unit=MIO_EUR&unit=PC_GDP&icha11_hc=TOT_HC&time=2017"

#Eurostat_title = "Eurostat total nights spent by residents/non residents 2018"
#json_RQST = "tin00175?c_resid=FOR&c_resid=NAT&c_resid=TOTAL&precision=1&unit=NR&time=2018&nace_r2=I551-I553"

#Eurostat_title = "Eurostat total deaths by pnomonia 2016"
#json_RQST = "tps00128?precision=1&sex=F&sex=M&sex=T&unit=RT&time=2016&age=TOTAL&icd10=J12-J18"

#Eurostat_title = "Eurostat Standardised death rate due to chronic diseases by sex"
#json_RQST = "sdg_03_40?sex=F&sex=M&sex=T&precision=1&time=2016"

#Eurostat_title = "Eurostat Total Population in January1 st 2018"
#json_RQST = "proj_18np?precision=1&age=TOTAL"

#Eurostat_title = "Eurostat # of practising physicians per 100k inhabitants 2017"
#json_RQST = "tps00044?precision=1&isco08=OC221&unit=P_HTHAB&wstatus=PRACT&time=2017"

#Eurostat_title =  "Eurostat Share of people with good or very good perceived health by sex above 16 2019"
#json_RQST = "sdg_03_20?precision=1&sex=F&sex=M&sex=T&unit=PC&quantile=TOTAL&time=2019&age=Y_GE16&levels=VG_G"

base = "http://ec.europa.eu/eurostat/wdds/rest/data/v2.1/json/en/"
url = base + json_RQST
```

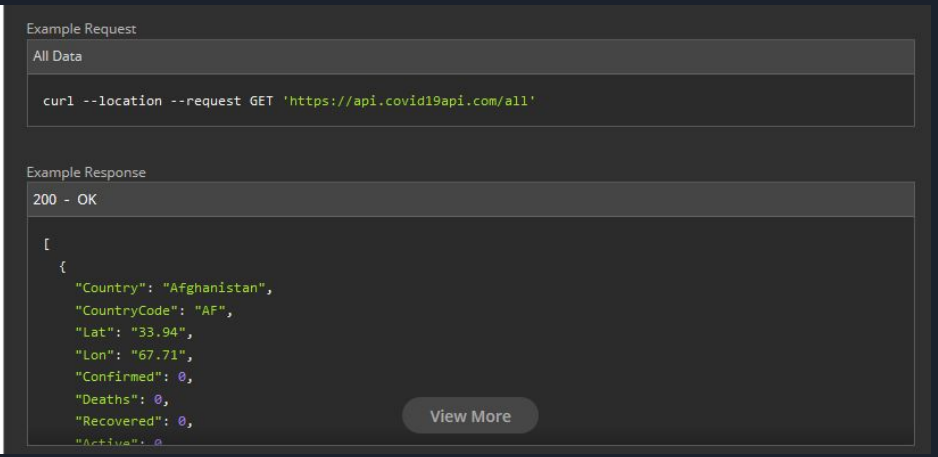
In [180]:

```
def eurostat(url):
    from pyjstat import pyjstat
    from collections import OrderedDict
    # read from json-stat
    dataset = pyjstat.Dataset.read(url)
    # write to dataframe
    df = dataset.write('dataframe')
    return df

df = eurostat(url)

#### Display dataset ##########
print ("Dataset for" , Eurostat_title)
df
```

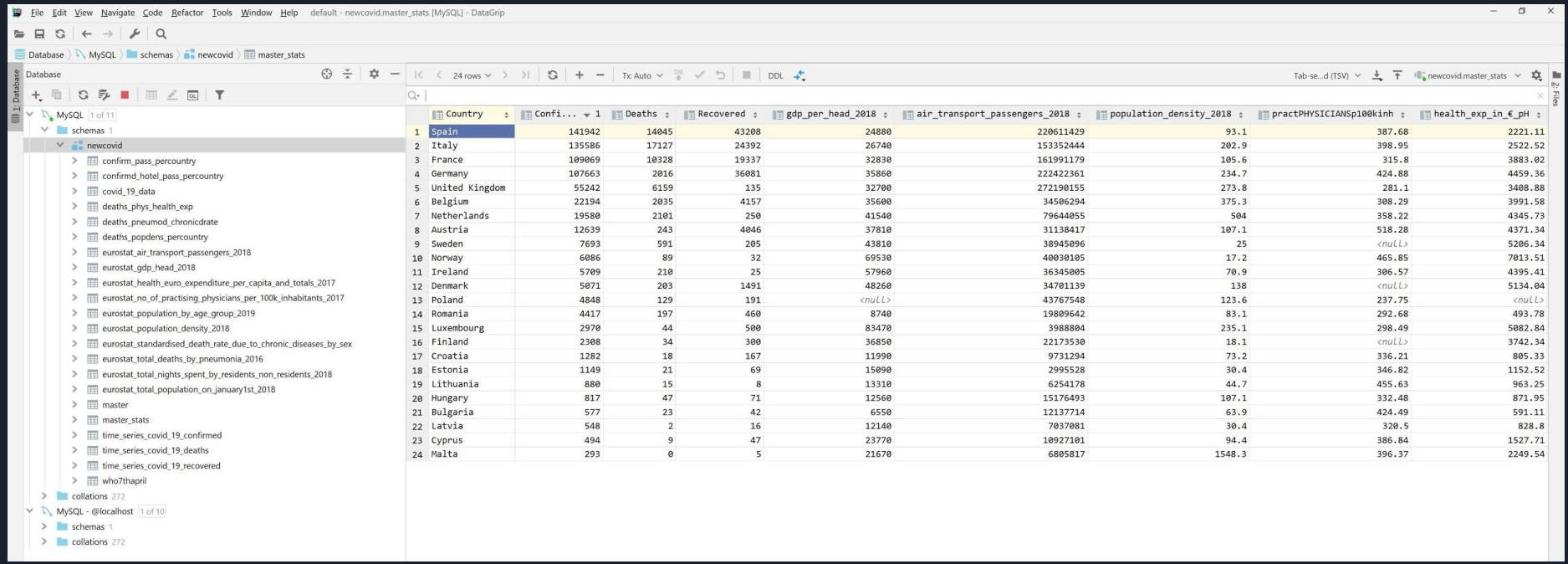
```
Dataset for Tourism
```

2. Data sourcing - API - WHO





3. Data handling - MySQL

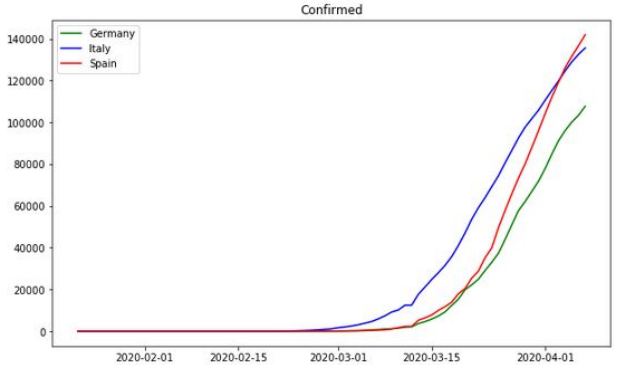


	Country	Confi... 1	Deaths	Recovered	gdp_per_head_2018	air_transport_passengers_2018	population_density_2018	practPHYSICIANSp100kinh	health_exp_in_€_pH
1	Spain	141942	14045	43208	24880	220611429	93.1	387.68	2221.11
2	Italy	135586	17127	24392	26740	153352444	202.9	398.95	2522.52
3	France	109069	10328	19337	32830	161991179	105.6	315.8	3883.02
4	Germany	107663	2016	36081	35860	222422361	234.7	424.88	4459.36
5	United Kingdom	55242	6159	135	32700	272190155	273.8	281.1	3408.88
6	Belgium	22194	2035	4157	35600	34506294	375.3	308.29	3991.58
7	Netherlands	19580	2101	250	41540	79644055	504	358.22	4345.73
8	Austria	12639	243	4046	37810	31138417	107.1	518.28	4371.34
9	Sweden	7693	591	205	43810	38945096	25	<null>	5206.34
10	Norway	6086	89	32	69530	40030105	17.2	465.85	7013.51
11	Ireland	5709	210	25	57960	36345005	70.9	306.57	4395.41
12	Denmark	5071	203	1491	48260	34701139	138	<null>	5134.04
13	Poland	4848	129	191	<null>	43767548	123.6	237.75	<null>
14	Romania	4417	197	460	8740	19809642	83.1	292.68	493.78
15	Luxembourg	2970	44	500	83470	3988804	235.1	298.49	5082.84
16	Finland	2308	34	300	36850	22173530	18.1	<null>	3742.34
17	Croatia	1282	18	167	11990	9731294	73.2	336.21	805.33
18	Estonia	1149	21	69	15090	2995528	30.4	346.82	1152.52
19	Lithuania	880	15	8	13310	6254178	44.7	455.63	963.25
20	Hungary	817	47	71	12560	15176493	107.1	332.48	871.95
21	Bulgaria	577	23	42	6550	12137714	63.9	424.49	591.11
22	Latvia	548	2	16	12140	7037081	30.4	320.5	828.8
23	Cyprus	494	9	47	23770	10927101	94.4	386.84	1527.71
24	Malta	293	0	5	21670	6805817	1548.3	396.37	2249.54

3. Data handling - Jupyter Notebook

In [322]:

```
fig, ax = plt.subplots(figsize=(10,6))

plt.plot(df_who_Germany['Date'], df_who_Germany['Confirmed'], zorder=1 , color="g",label = "Germany")
plt.plot(df_who_Italy['Date'], df_who_Italy['Confirmed'], zorder=2 , color="b" , label = "Italy")
plt.plot(df_who_Spain['Date'], df_who_Spain['Confirmed'], zorder=3 , color="r", label = "Spain")
ax.set_title("Confirmed")

plt.legend(loc="upper left");
```



In [323]:

```
fig, ax = plt.subplots(figsize=(10,6))

plt.plot(df_who_Germany['Date'], df_who_Germany['Deaths'], zorder=1 , color="g",label = "Germany")
plt.plot(df_who_Italy['Date'], df_who_Italy['Deaths'], zorder=2 , color="b", label = "Italy")
plt.plot(df_who_Spain['Date'], df_who_Spain['Deaths'], zorder=3 , color="r", label = "Spain")
ax.set_title("Deaths")

plt.legend(loc="upper left");
```

Data - problems and limitations

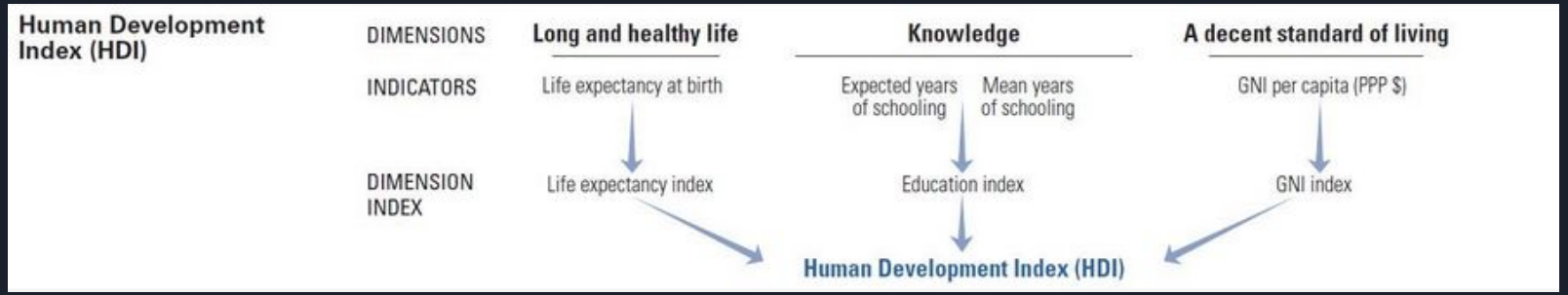
- EUROSTAT API: didn’t read the description properly
- HDI - API: registration failed several times, no response from support
- using data from different sources
  - formatting data properly to have master table
- Data interpretation, i.e. France had several rows, but former colonies included
- Some ideas discarded due to lack of data
  - back-calc. from dead to infected via lethality rate
  - determine dead/infected for North Korea

Question

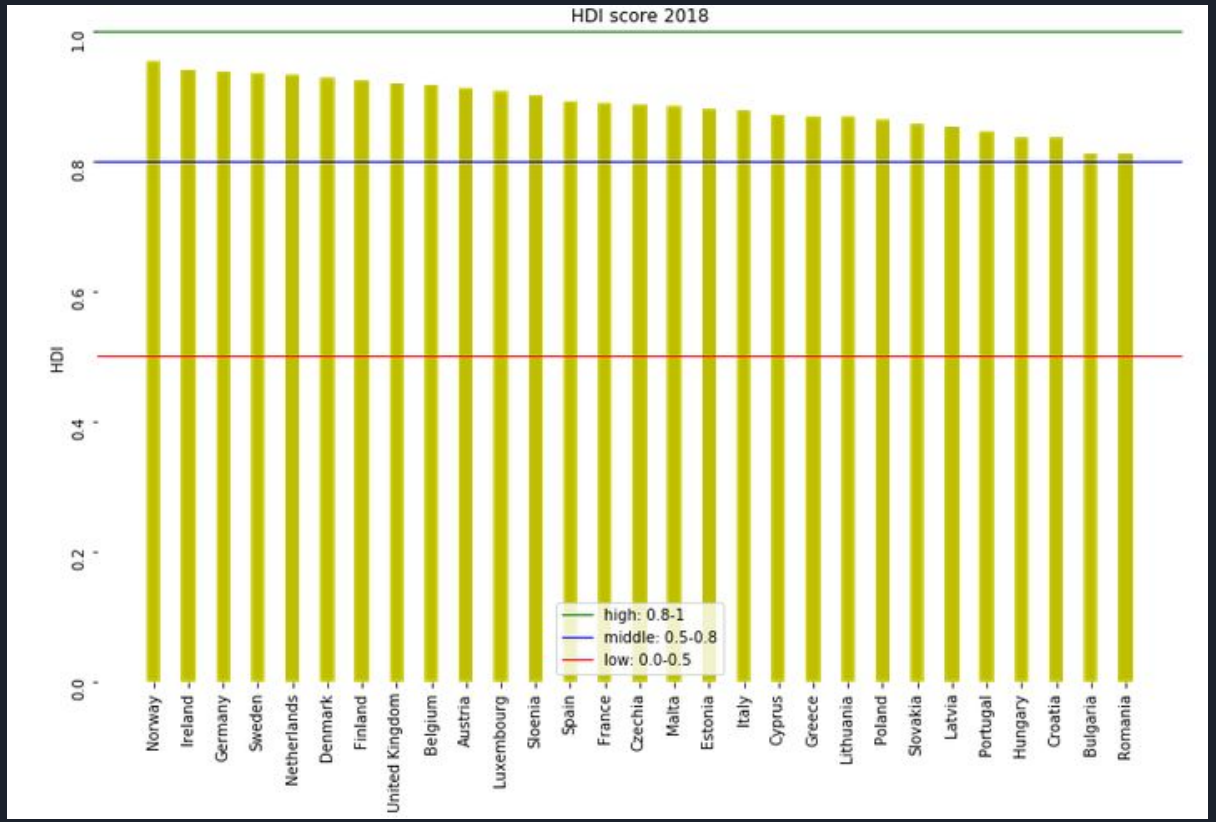
Are there correlations between socio-economic topics and deaths caused by Corona virus in Europe?

4. Does a low HDI lead to more deaths?

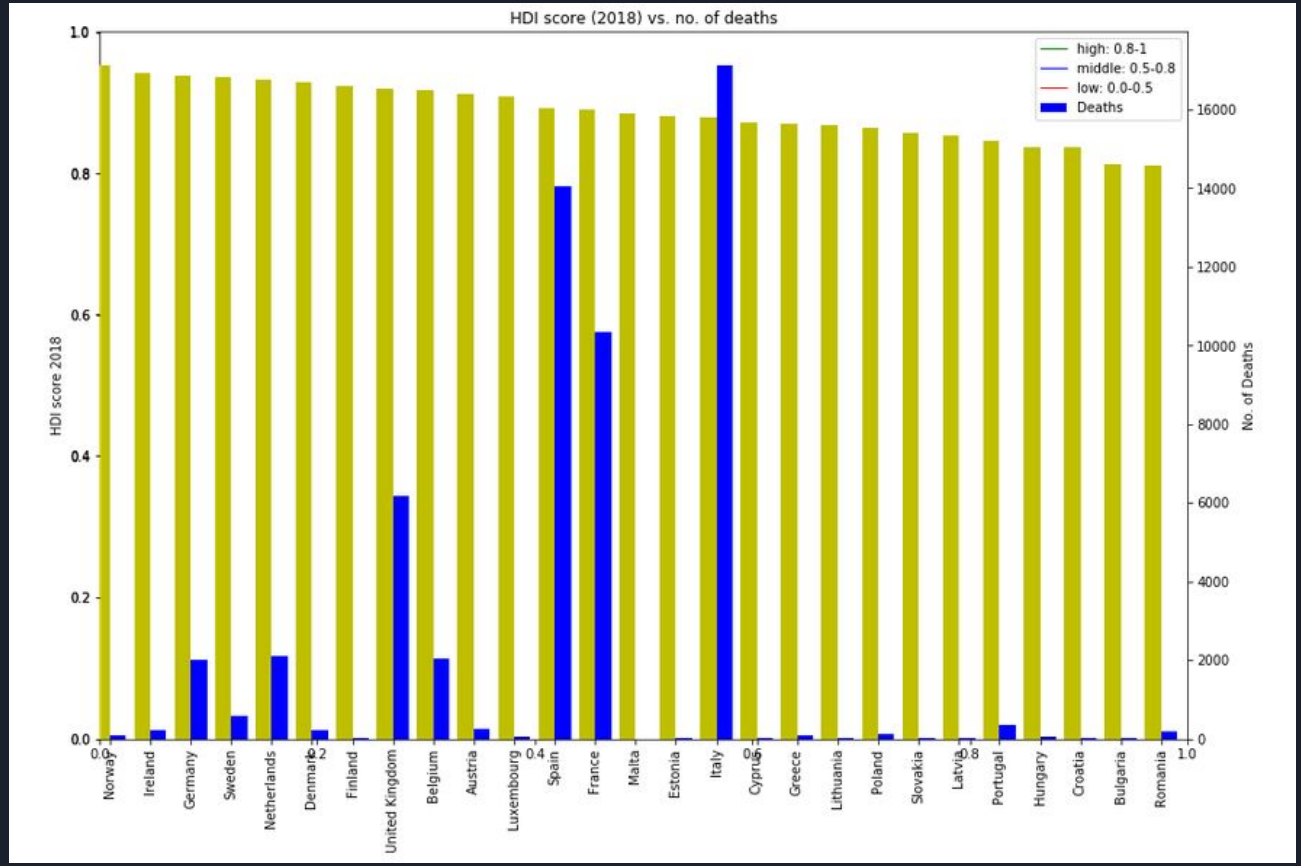
HDI is representing socio-economic topics, which can also be defined as a prosperity factor of states.



Does a low HDI lead to more deaths in Europe?



Does a low HDI lead to more deaths in Europe?

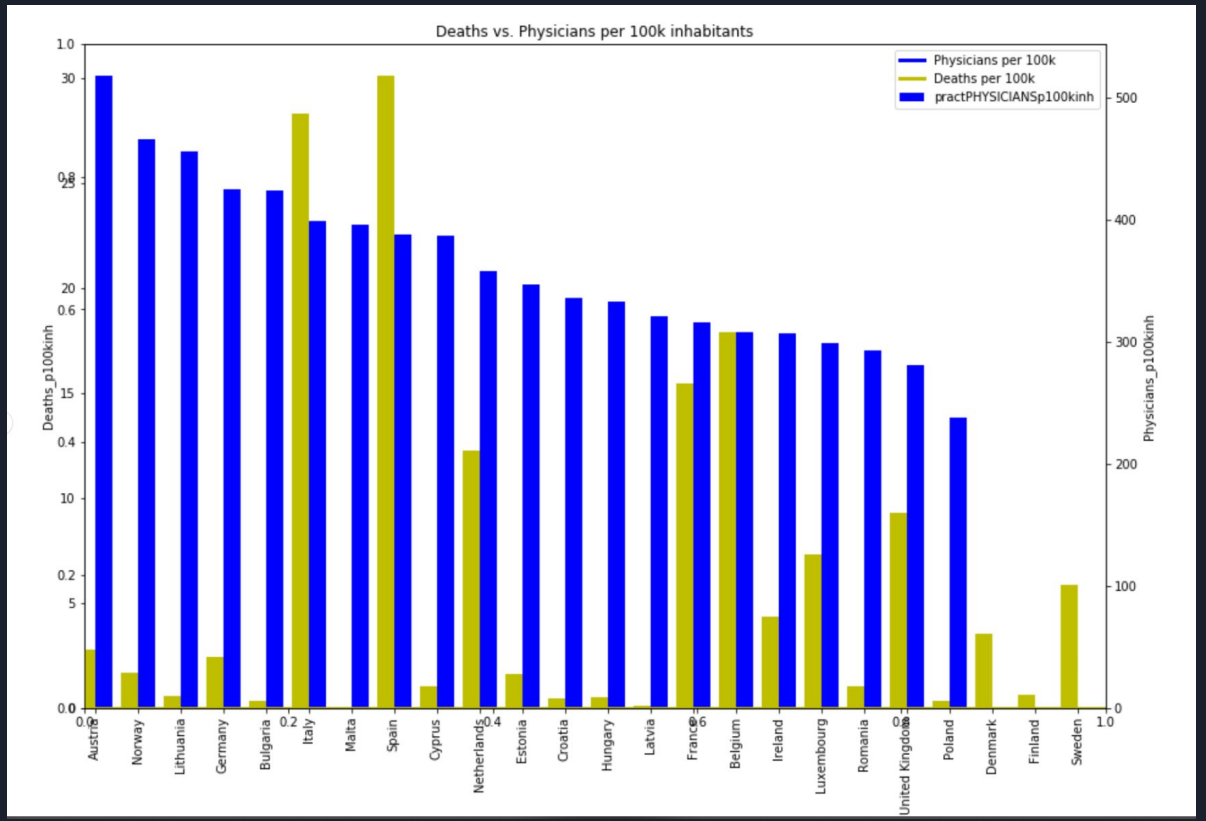


No correlations with given data can be found.

...hence we need to investigate other correlations of which two examples are shown in following slides.

Does a low no. of physicians lead to more deaths in Europe?

Further investigations f.i. no. of hospital beds vs. no. of deaths can be done.



Correlation between infections and tourism exists.

There is a clear correlation between the spent nights and no. of infections. A further investigation should be the integration of f.i. Airbnb stays.

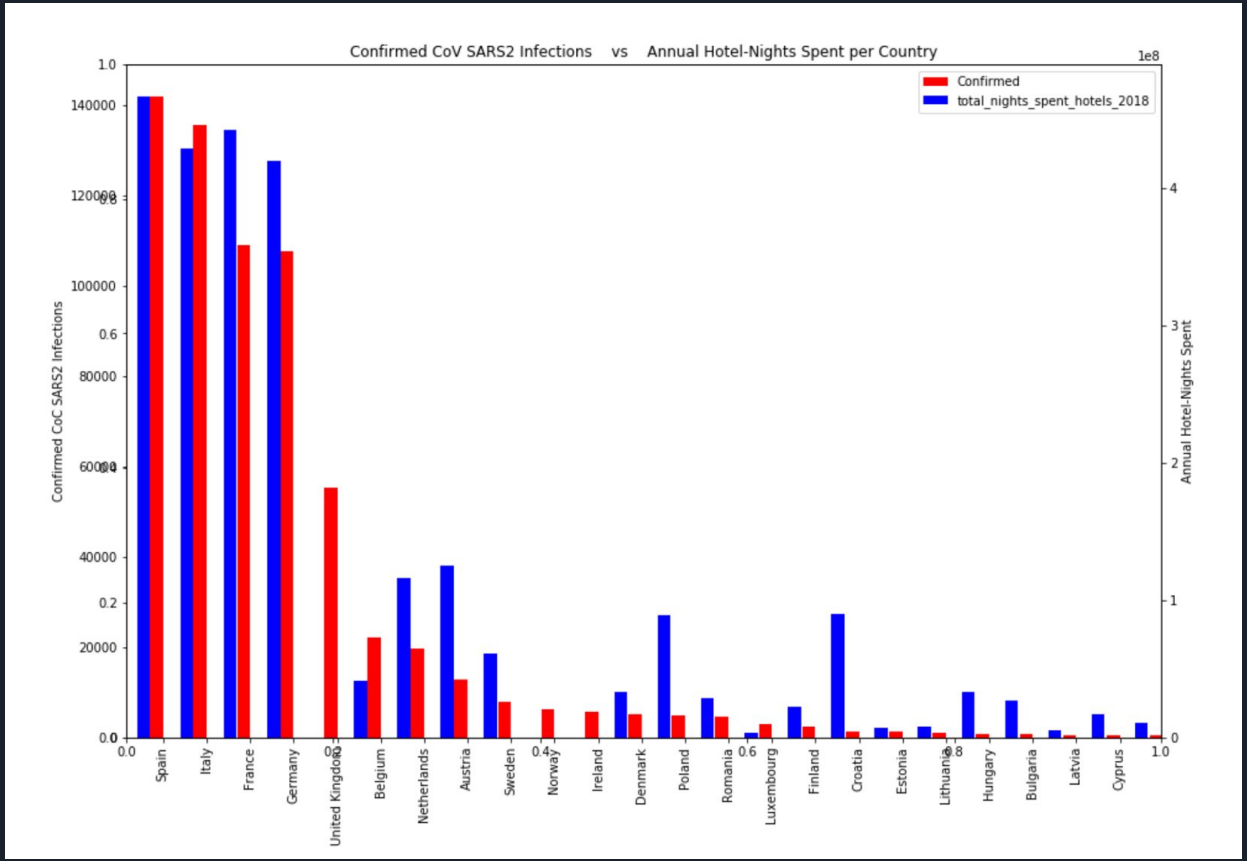


Table of links

EUROSTAT

https://ec.europa.eu/eurostat/web/json-and-unicode-web-services/getting-started/rest-request

WHO

https://documenter.getpostman.com/view/10808728/SzS8rjbc?version=latest#81415d42-eb53-4a85-8484-42d2349debfe

HDI

http://hdr.undp.org/en/content/human-development-index-hdi

RKI

https://npgeo-corona-npgeo-de.hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0

GitHub

https://github.com/ThomasBentler/DATA-THIEVES-COVID19.git