Course: Data Wrangling **(4/02/2022)**

Data Wrangling Final Project

Name: Filippo Pacinelli*, Netid:* fpi380

1 Research Question

The (small) discrepancy between measures taken to tackle covid-19 pandemic, people perception of them, and actual consequences on European economies

This small project entails a mainly descriptive analysis where, mostly for European countries, the measures taken to tackle the covid-19 pandemic are analysed and compared to GDP levels among countries. The final goal is to find similarities, differences, and a possible relationship between the type of measures taken on a national level and national GDP. However, in the analysis also another indicator of wealth has been taken into consideration (namely, the gini index) to allow for different 'points of view' so to obtain a broader picture.

2 Data Sources

- Acaps, COVID-19 GOVERNMENT MEASURES DATASET
  https://www.acaps.org/covid-19-government-measures-dataset
- - World Bank Data, GDP of every country from 1960 to 2020
  https://data.worldbank.org/indicator/NY.GDP.MKTP.CD
- - Europa Data, GDP of every european country from 2000 to 2020
  https://ec.europa.eu/eurostat/databrowser/view/sdg$_0$8$_1$0/$default/table?lang = en$
- - Eurostat, Gini coefficient of equivalised disposable income - EU-SILC survey
  https://ec.europa.eu/eurostat/databrowser/view/sdg$_0$8$_1$0/$default/table?lang = en$
- Quarterly growth of GDP in the World from 1970Q1 – 2021Q3 (note: GDP growth, not GDP)
  https://data.oecd.org/gdp/quarterly-gdp.htm

3 Data wrangling methods

The Research Rational:

- Get measures taken to tackle covid for every country
- Analyse those measures
- Get GDP levels for every country, if possible on a quarterly basis
- Analyse GDP
- Get another measure of wealth distribution
- Perform an analysis to grasp eventual relationship among those data

To do so, 4 main stages can be recognised in the project:

1. Retrieving Data (sources showed above)
2. Cleaning Data (prepare and model the dataset to have them ready for subsequent stages)
3. Data Analysis Plotting (the actual analysis of datasets and easily understandable graphs)

4. Drawing Conclusions (drawing conclusions about what stage 3 has suggested)

Main Techniques used in those stages:

1. drive.mount() function to import datasets from Google Drive and Pandas to read the files (csv and excel files)

2. Pandas, Numpy and Datetime have been deployed massively to obtain cleaned, ready-for-use datasets which can actually be used for later analysis. Very often, data were displayed in an unconvienent fashion. For example, dates were displayed as rows, where, for example, for every time and for every country a value was repeated in a row-wise way. An example can be observed in the table below. That required mainly the use of pivoting (pandas) and groupby (pandas). Also, a lot of NaN values where observed. When present in a full row (or column), such row (or column) was deleted from dataset. Afterwards, remaining NaN have been kept just like NaN. That was mainly due to the fact that often figures were very similar and replacing NaNs with, for example, the average value for that entry wouldn't have given any meaningful advantage.

   Secondly, dates were hardly ever saved as datetime object and numbers were sometimes saved as strings. In both cases a transformation was needed to properly deal with the time and numeric nature of those data. Some datasets were pretty well cleaned already but might have a lot of useless (for this analysis) information, thus some columns have been deleted to avoid excessively big datasets. Same apply for too old information (data of more than 30 years ago).

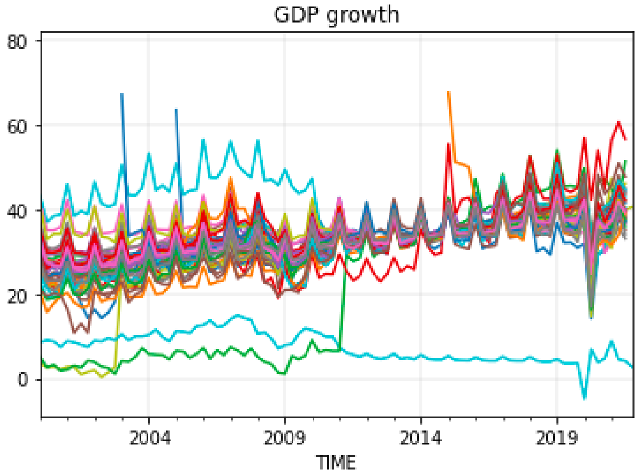
3. As it is a mainly descriptive project, I decided to work mainly on visual analysis. As a consequence, Pyplot and Seaborn have been largely deployed in the analysis section. Obviously, it was often the case that subsets of data and/or groupby/pivoting/transposing or even new-columns creation were needed to properly plot information. In some cases, with time series, a seasonality-removal process has also been applied to obtain less noisy curves. At the end of the analysis, also two statistical test for dependencies have been used: Fisher Exact Test and Chi Squared Test. As it will be showed, the graphical analysis support conclusions which are only in small part in accordance with the statistical tests, since no significance levels have been reached with statistical tests.

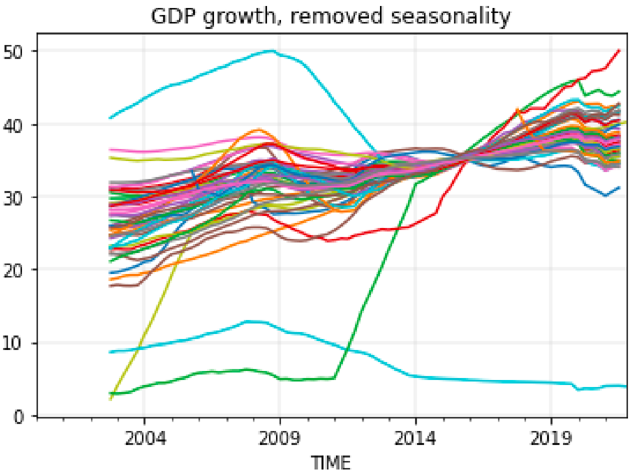
   Thus, this project can be seen as an example of slight discrepancy between what we usually expect, what we observe and what we actually conclude after proper statistical methods are adopted in data analysis.

4. Conclusions are drawn and plots are showed in the next section

4 Conclusion

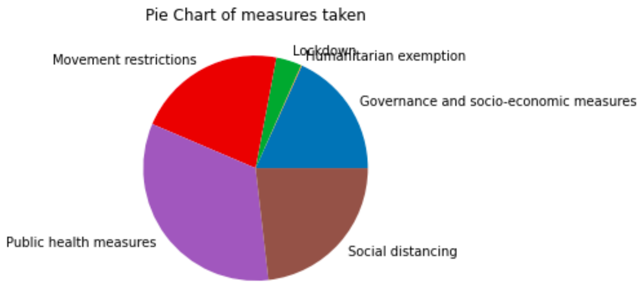
The starting point was to analyse GDP growth in Europe in the last 20 years.

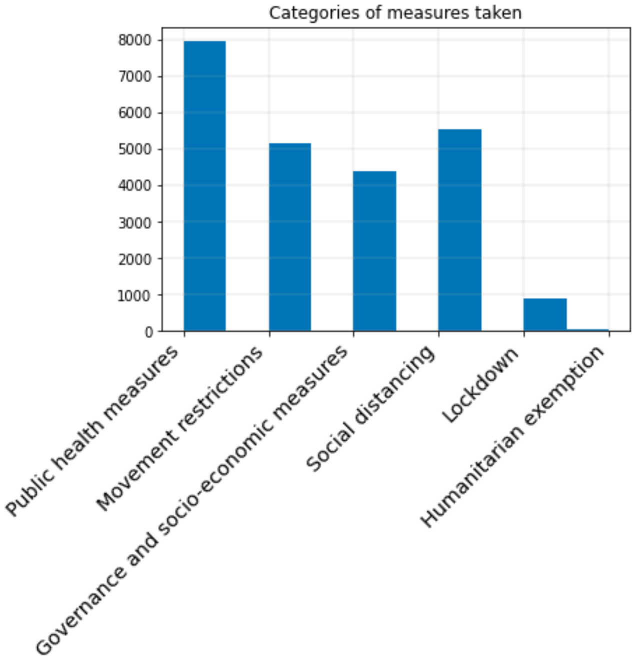




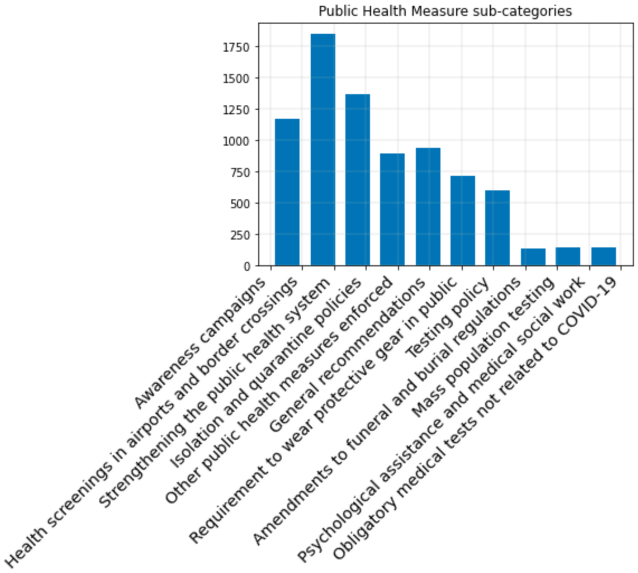
Above, it's possible to observe the GDP growth for all european countries in the time range 2000-2021. On the left, time series are plotted as they are in the dataset. Despite the noise (clear seasonality pattern), a drop in 2020 is clear. Interestingly, removing a little bit of noise (using a rolling average measured over 12 periods, correspoding to 12 quarters, so 3 years), the drop is still present but definetely less drastic (but still bigger than the 2008 Financial Crisis).

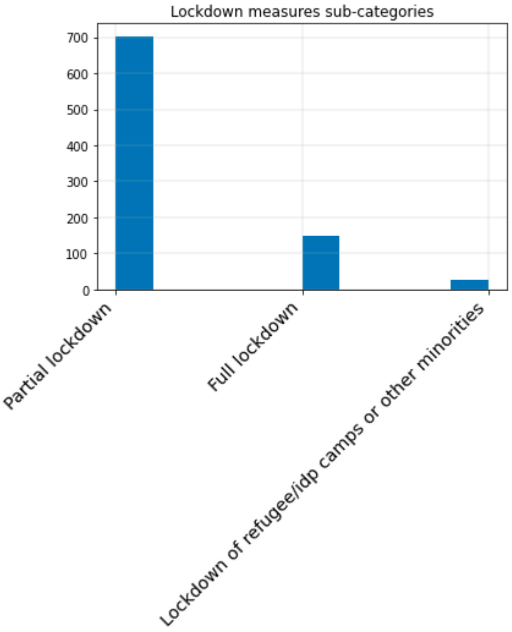
Let's now analyse the most popular measures taken at a governmental level to fight the pandemic.





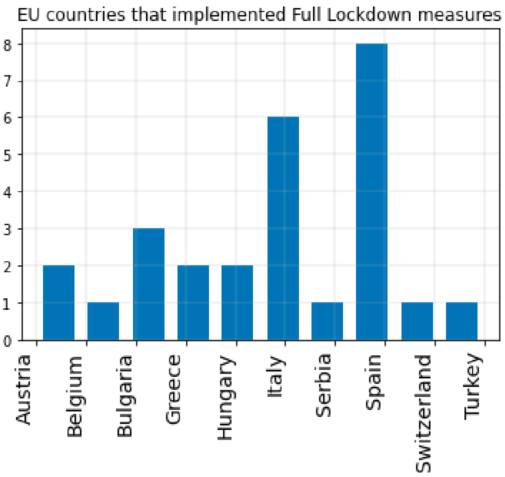
Interestingly, despite the perception of people, lockdown measures were not that popular among decision makers, or at least not that much. As it's possible to observe, the lockdown category is way less common than other type of measures.



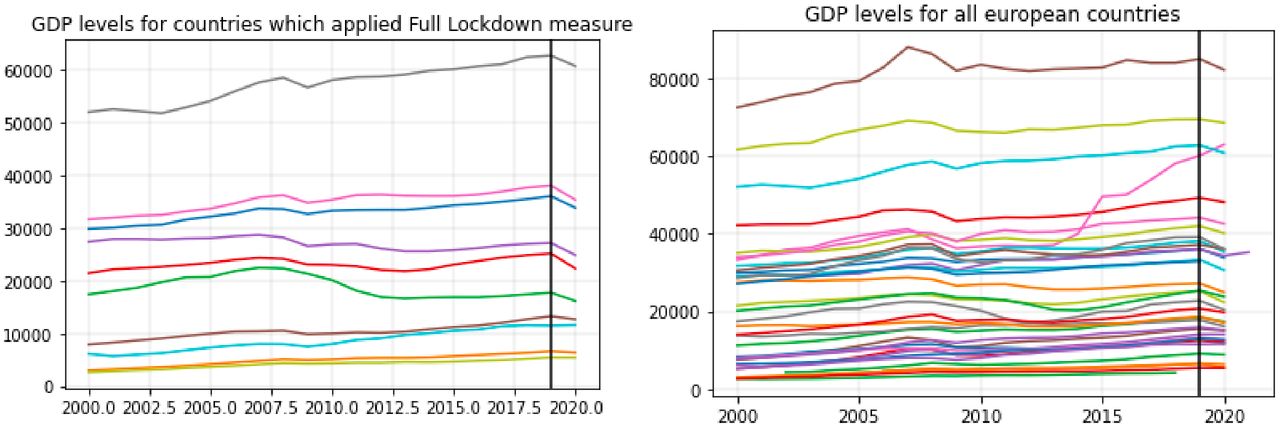


Going more into the details, it's even clearer that the most commonly measures adapted to tackle covid are relatively mild and non-intrusive. For example, the most common subcategories in the *Public Health Measure* category shall be seen as very understandable and, at least from my point of view, not such harmful for freedom. At the same time, going into the details of the *Lockdown* category, it's clear that a *Full Lockdown* regime was very often avoided and a *Partial Lockdown* was usually preferred.

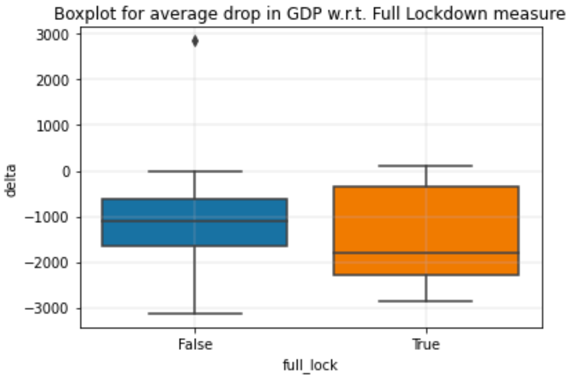
Also, it might be interesting to analyse those european countries where a Full Lockdown was applied.



Let's now see, for those countries, the behavior of their GDP levels (in the last 20 years), represented in the image below on the left and let's compare it with the behavior of all european countries, represented below on the right.



From that graphical analysis, no particularly meaningful differences might be spotted. It can be said that the countries which experienced Full Lockdown measures have a more clear drop between in 2020 (after the black horizontal line, representing the "critical point"). Other techniques of analysis might be needed. To better grasp differences, if any, the difference between 2020 and 2019 levels of GDP have been calculated for all european countries. Then, a box-plot analysis has been deployed. More meaningful results have been actually obtained: countries without full lockdown experienced a smaller drop in GDP levels and in a more consistent way (less



variance of average drop). On the other hand, countries with Full Lockdown measures implemented, despite the larger range of values, have clearly, on average, observed a way larger decrease in their level of GDP.
However, to conclude the analysis, a proper statistical test for dependence should be exploited. A confusion matrix was created and both a Fisher Exact Test and a Chi Squared Test were applied to it to test more rigourously for dependencies between the presence of Full Lockdown measures and higher drop in GDP levels. The output is showed below:

Fisher Exact Test Statistics = 2.25
Fisher Exact Test P-Value = 0.38109337172021146
Chi Squared Test Statistics = 0.21875
Chi Squared Test P-Value = 0.6399940105774465

The statistical tests, however, cannot reject the null hypotheses of independence, meaning that, from a rigorous statistical point of view, it's not possible to talk about strict relation between GDP levels and Full Lockdown measures. However, it should be noted that the data were probably not enough to obtain an extremely reliable conclusion.
Lastly, the level of Gini indexes have been analysed as well to spot some differences. The Gini index is a measure of equality in wealth distribution. A Gini index of 0 represents perfect equality, while an index of 100 implies perfect inequality (more usually, it is represented as a figure between 0-1, but its interpretation is the same). The dataset analysed gives the gini index (0-100) of all European countries.

However, again, the analysis showed mixed results:

Proportion of countries with Full Lockdown measure where Gini index increased from 2019 to 2020: 0.444
Proportion of countries without Full Lockdown measure where Gini index increased from 2019 to 2020: 0.333
Proportion of countries with Full Lockdown measure where Gini index increased from 2018 to 2019: 0.3
Proportion of countries without Full Lockdown measure where Gini index increased from 2018 to 2019: 0.375

So, it can be concluded that from Full Lockdown measure a remarkable increase in wealth distribution occured (almost half of the countries against one third of countries where no Full Lockdown measure were applied).

5 Final Remarks

To conclude, what this project suggests may be summarized in the following points:

- Both statistical and visual analysis are not, by themselves, 100% reliable measures since, as we have seen, they might not be in accordance. However, it's also true that one should not conclude that those are therefore useless, but it's always important to be careful when drawing conclusions.

- That project also shows that it's often the case the perception of people towards public decisions are often not very well supported by data. Lockdown measures are clearly very hard measures and obviously gained more attention, but we have observed that they represent a really small portion of all the measures taken

- Also in terms of wealth distribution equality, which is an highly debated topic, the analysis clearly showed that an increase in equality occured in those countries where harder measures were taken.