First year project 2, Spring 2021
COVID-19 and the Weather

Michele Coscia

February 23, 2021

1 Overview: Spatial Data Science

In this project, you will complete tasks similar to data scientists working for a public health agency, to inform central governments about possible correlates between COVID-19 and environmental factors facilitating or hampering its spread. You will explore the latest data set of all infections in the year 2020 provided by the government, and weather data from the IBM Pairs system. IBM will give an introductory guest lecture about their study on the same topic.

The major parts of the project are:

- Exploring and transforming the data, making numerical and visual reports;
- Connecting data tables (weather, infections);
- Investigating possible statistical associations by filtering for a variety of attributes;
- Visualizing the data on a map;
- Involving self-obtained external data sets in the analysis.

2 Requirements

In addition to the requirements in the course description, you must work on github. All Python packages are allowed, but comment your code assuming the reader has no knowledge of extra packages that were not covered in the first semester, such as pandas.

3 Assignment

You are the data scientist team working for the public health agency of a country:

Country	Denmark	Sweden	Germany	Netherlands
Group	1,2,3,4	5,6,7,8	9,10,11,12	13,14,15,16

Your job is to create a report for the government planner, informing politicians about the relationship between the intensity of the COVID-19 outbreak and the weather.

3.1 Task 0: Data filtering and cleaning

Out of the raw dataset, create a processed data set that contains only fields and records that are relevant to your analysis. You are going to use this processed data set for all the tasks below. Briefly describe your data set in a numerical summary (e.g. number/meaning of fields and records, statistical key metrics).

3.2 Task 1: Single variable analysis

Report key statistics for the relevant variables you selected for your analysis (e.g. number of infected and temperature) for 1) the different country regions, and 2) in different levels of aggregation (daily, weekly, monthly).

3.3 Task 2: Associations

Research whether there is a significant statistical association in your country between weather data and infection rates. Report whether there is a statistically significant association between your chosen variables or not, together with the appropriate statistical metric(s). Discuss why this association, or the lack of this association, is relevant for policies.

3.4 Task 3: Map visualization

Visualize a selection of important variables on a map of your country.

3.5 Task 4: Open question

Use the data to formulate, motivate, answer, and discuss another research question of your choice. For example, compare your country to another one in the dataset; or investigate additional variables that you could find elsewhere (e.g. total population, population density, population demographics such as age distribution, number of elders, etc); or identify if there are temporal patterns (is one variable more important in some months? Are weekends different from weekdays? Etc), ...

4 Hand-in

You must hand in:

- gitlog.txt: Your repo's git log, e.g. by running: `git log > gitlog.txt`.
- code.zip: One zip file containing one Jupyter notebook (.ipynb) of your commented code that runs fully without errors using the three raw data files, and reproduces your findings. Do not include the raw data files here. If your code is making use of external data sets or .py scripts, include them here.
- report.pdf: A project report.

The project report must be between 3 and 5 pages long including figures (with 11 pt font size and about 1.5 cm margins, like in the first project), and should consist of precisely the following sections:

1. Introduction: Here you provide the context and motivation for the problem. What are your research questions, and why does your research provide value to the country?
2. Data: Here you describe all your data tables/sets, and briefly summarize how you obtained and cleaned/transformed them without referring to any code. You should also describe here how you dealt with data issues such as missing data.
3. Results and discussion: Here you provide the technical results and a discussion of your findings over the data.
4. Limitations: Here you give an account on the major short-coming(s) of your methodology/data.
5. Concluding remarks and future work: Here you provide a couple of sentences summarizing the results of the project, why your report is relevant for your country's health, and indicate how the methods, data, and analysis could be improved or extended.
6. Disclosure statement (optional): Here you may state if there were any serious unequal workloads among group members.

For each of the tasks 1,2,3,4 you must provide both a textual description and exactly one figure, so the report will contain exactly 4 figures (figures may have subplots). Your whole report can contain maximum one table with numerical results. In your report, cite at least one reference, for example in the introduction and/or conclusions. The reference can be academic or non-academic (e.g. a news piece talking about something related to your analysis, to show real-world relevance). Your hand-in should be self-contained. You are required to master your code for the oral exam. Your 10 minute oral presentation should correspond to the structure of your report. However, you are encouraged to have slide headings that are more communicative.