COVID-19 Project

Introduction

The goal of this project is to implement what you studied in this course. The main idea of this project is to research coronavirus cases in the US. You are provided with three data sources that you should apply to your project, and required to add at least one additional data source that adequately answers a meaningful question you will formulate.

The project consists of three parts:

I. Exploration of COVID concern and COVID cases

II. Exploration of Donald Trump's tweets during the concern poll. Can tweets be predictive of COVID concern? Generate a hypothesis based on the exploration. Then, build a machine learning model that uses Trump's tweets as the explanatory variable and concern as response (predicted) variable. Use k-fold cross validation to show model success in terms of pearson's R. **Please note that the models' performance will not be taken into consideration in your grades.**

III. Formulate a meaningful question in the realm of American politics/demography/other and the relationship with COVID cases. For example, can political polarization explain differences in the spread of COVID? Search for a dataset "in the wild" that can be used to ask your questions (e.g., 2016 presidential election results). The data should be the explanatory variable, the response variable must be COVID cases from the wiki dataset. You choose the scope and timeframe, but this should be reasonably motivated **(Do not use the same question as the example provided in this part)**

Each of the task will be assessed with respect to how well you:

1. Explain your way of dealing with the data. Make sure you follow the best practices we covered in class and communicate it clearly.
2. Visualize properly and in a way that makes sense.
3. Write efficient, clear and well-commented code that follows the tidy principles.
4. Make sure you communicate a coherent story that follows the data and not abuse it.

Data Sources

In this project you are required to work with the following data sources:

a. Covid concern (provided, see here for more info)
b. Scrap from wiki covid data (here)
c. Trump tweets (provided)
d. At least one additional meaningful data source (voting behavior, google search volume, other)

Please note that no additional instructions are given on the data, you are required to incorporate what you studied during the course to explore, manipulate, clean and visualize the data.

Instructions

1. The project should be handed in groups of three students!
2. The project due date is 20.06.2021. 23:59.
3. This assignment should be uploaded as a zip file containing your R project. It should contain an RProj file, “data” folder, “figures” folder, 1 rmd files and one knitted html file.
4. Each group is required to present their project, please fill your team in the time slot in the following document: **Link**
5. Each group is required to submit at least one day before the presentation the presentation file and a document which states which student has worked on which part of the project and their contribution.
6. Questions about the project should be asked using the dedicated project forum.
7. Pay attention to possible updates on the project in the announcements forum in Moodle.