CS 7331 Data Mining
Project 1: Data and Visualization

Due Date: See Canvas

Points: out of 100

Please submit your report in **PDF format.** If you want to submit code, then please submit it in an additional file containing sufficient comments to make it understandable.

For this project, we will look at different data sets collected by Google about the spread of the COVID- 19 virus. The data is available on the Google Cloud Platform:

https://console.cloud.google.com/marketplace/browse?filter=solution-%20type:dataset&filter=category:covid19

I have extracted data about the number of cases, demographics, and social distancing and put the data sets on Canvas (under Files/Project). You can always get more and more current data directly from the link above.

Some general questions we are interested in:

- What is the trend in different areas (states, counties) of the US?
- Is social distancing done and is it working?
- Can we identify regions that do particularly well?
- Can we predict the development in a region given the data of other regions?

In this project, we will focus on cleaning and understanding the data.

Follow the CRISP-DM framework

1. Business Understanding [10]

What is COVID-19 and what is social distancing and flattening the curve? Why is it important to look at data about the virus spread, hospitalizations, and available resources? Who is interested in this information? What decisions can be informed using such data? [10 point]

2. Data Understanding [45]

- What data is available? Describe the type of data (scale, values, etc.) of the most important variables in the data. [9 point]
- Verify data quality: Are there missing values? Duplicate Data? Outliers? Are those mistakes? How can these be fixed? [9 Points]

- Give simple appropriate statistics (range, mode, mean, median, variance, etc.) for the most important variables in these files and describe what they mean or if you find something interesting. [9 points]
- Visualize the most important attributes appropriately. Provide an interpretation for each graph. Explain for each attribute type why you chose the visualization. [9 points]
- Explore relationships between attributes: Look at the attributes and then use cross-tabulation, correlation, group-wise averages and box plots, etc. as appropriate. [9 points]

3. Data Preparation [35]

- Create a data set with objects as rows and features as columns. Use as objects counties in the US. [20 points]
  Interesting features may be for example:

  - When was the first case reported?
  - How fast did the virus spread?
  - How (densely) populated is the county?
  - What resources (money, hospital)?
  - What is the social distancing response and how long did it take after the first case?

- Visualize the created attributes. [15 points]

Exceptional Work [10 points]

Examples: Ask your own questions and answer them, insightful visualization of results, in-depth explanation why one method works better than another, developing special preprocessing, and using and explaining a method not covered in class.