Data Mining Spring 2021

Project 2: Clustering

Due: see Canvas
Points: UG out of 100, Grad out of 110

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.

Source: Businessinsider.com
https://www.businessinsider.com/coronavirus-live-updates-latest-news

We will keep working with data for COVID-19 and try to group similar counties in order to find patterns.

Some general questions we are interested in:

- What counties are similar in terms of makeup of the population, and the response to the virus?
- What is the differences between the found groups of counties.
- Can we identify groups that are more severely hit than others?

Follow the CRISP-DM framework

3. Data Preparation [30 points]

- Describe which features you want to use for clustering and why. Add a table with all the features and basic statistics. [20]
- What is the scale of measurement of the features and what are appropriate distance measures? [10]

4. Modeling [60 points]

- Perform cluster analysis using several methods (at least k-means and hierarchical clustering) using different feature subsets. At least 4 different clusterings. [40]
- How did you determine a suitable number of clusters for each method? [10]
- Use internal validation measures to describe and compare the clusterings and the clusters (some visual methods would be good). [10]
- Can you use a feature as the ground truth and perform external validation? [exceptional work]

5. Evaluation [10 points]

- Describe your results. What findings are the most interesting?

Graduate Students: Exceptional Work [10 points]

- Examples: Use different data, use more clustering algorithms, visualization of results, perform external validation.