Corona Virus

Luke Gruszka
Spring 2020
sudozyphr.github.io

Which Domain?

1. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/summary.html
2. https://ourworldindata.org/coronavirus
3. https://www.who.int/emergencies/diseases/novel-coronavirus-2019
4. https://www.worldometers.info/coronavirus/
5. https://informationisbeautiful.net/visualizations/covid-19-coronavirus-infographic-datapack/
6. https://www.ecdc.europa.eu/en/publications-data/rapid-risk-assessment-novel-coronavirus-disease-2019-covid-19-pandemic-increased
7. https://www.barrons.com/articles/latest-coronavirus-data-show-disease-continues-to-spread-even-in-the-u-s-51584224660
8. https://www.theguardian.com/world/2020/mar/13/coronavirus-pandemic-visualising-the-global-crisis
9. https://ourworldindata.org/coronavirus-source-data
10. https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

These domains either contain information or data sets that will be useful to be able to analysis in the project. I will have data that is both in 2019 and 2020 that will be able to see what measures taken have helped the spread. For example, when certain safety measures were taken and how efficient and effective those actions where and what affect those had on the virus spread and outbreak. From this data I will be able to map out and see when the spread will slow down or dissipate.

Which Data?

https://github.com/CSSEGISandData/COVID-19

From this link I will be able to access the updated point in time data set that I could use for the data. Also, I have the CDC that has data sets that will be updated regularly that I will be able to also use if the data set above does not contain all the information that I want. I will be able to perform ETL on the CDC site and get the exact information that I will need.

Research Questions? Benefits? Why analyze these data?

I am proposing to see how long the virus will last. I will examine the data to see if I can find a trend as to when the virus will decimate and leave.

What Method?

I will be using linear regression to analysis the data. I anticipate that my data will be labeled and well defined. If my data is lacking I will look towards undefined data and turn to an unsupervised method such as K-means clustering and then identify the clusters based upon locations.

Potential Issues?

The challenges I will have is keeping the data up to date. The information is constantly being updated and thus my ETL will have to be pulling the data everyday to make sure that I am getting the most up to date information. If I do not have the most up to date information then my prediction and model will be inaccurate.

Concluding Remarks

Today we are faced with a virus that is rapidly spreading and affecting everyone across the world. While the world governments are taking action and implementing safety procedures; we still see that the virus is spreading. This draws the questions if these actions are preventing the spread or simply just stalling the spread. The data being collected will give me the option to be able to see when the virus will be predicted to end and if the safety measurements being taken are actually slowing down the spread for the virus. By using linear regression I will be able to map out the trajectory of the virus. Also, I will be able to use a time series to map out when certain safety precautions have been taken and what effects those actions have on the virus outbreak.