Marissa, Moe, Star, Matt

COVID-19

Data Engineering Bootcamp

Project 1

CONTENTS

BACKGROUND

Data Source

- Kaggle Novel Coronavirus 2019 Dataset
- Total Covid-19 data conducted worldwide

COVID-19

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus. The disease was first identified in 2019 in Wuhan, the capital of China's Hubei province, and has since spread globally, resulting in the ongoing 2019–20 coronavirus pandemic.

RESEARCH QUESTIONS

DEMOGRAPHICS

Which age group has a higher recovery rate?

Which gender has a higher chance of recovery?

HYPOTHESIS: Females that are under the age of 60 have a higher rate of recovery.

TIME SERIES

Which countries have begun to “flatten the curve?”

Hypothesis: Countries that have had the disease longer are more likely to see flattening.

TEMPERATURE

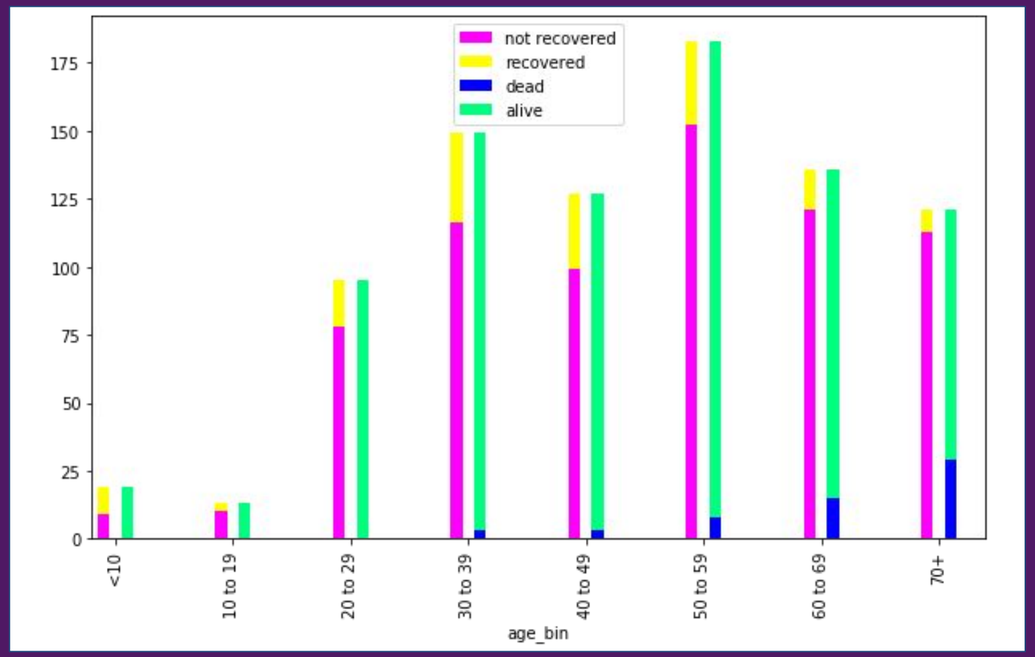
Does a country’s temperature impact the amount of infections is has?

Hypothesis: Temperature and infection cases are related. Countries with warmer temperatures will have less infection rates.

DEMOGRAPHICS

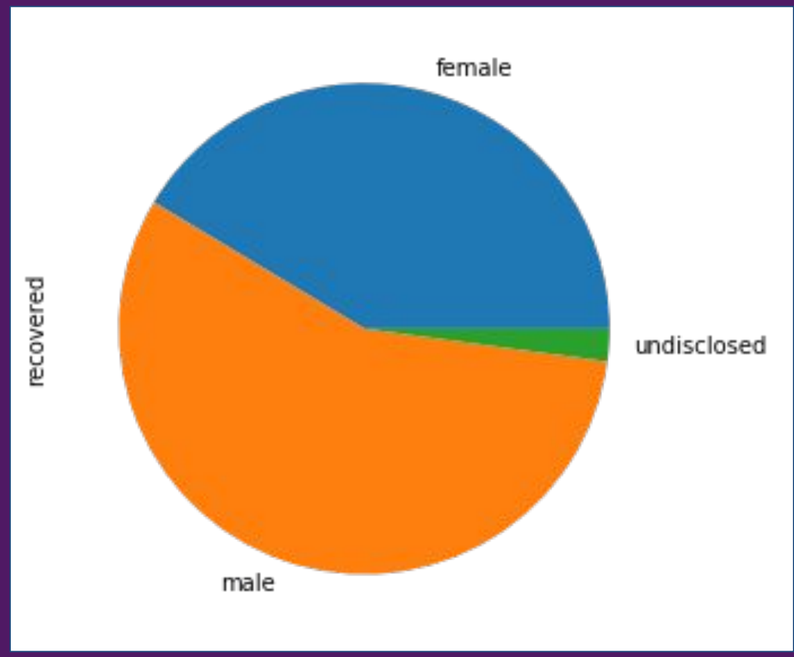
AGE

- 50-59: largest number of cases
- 70+: highest fatalities



GENDER

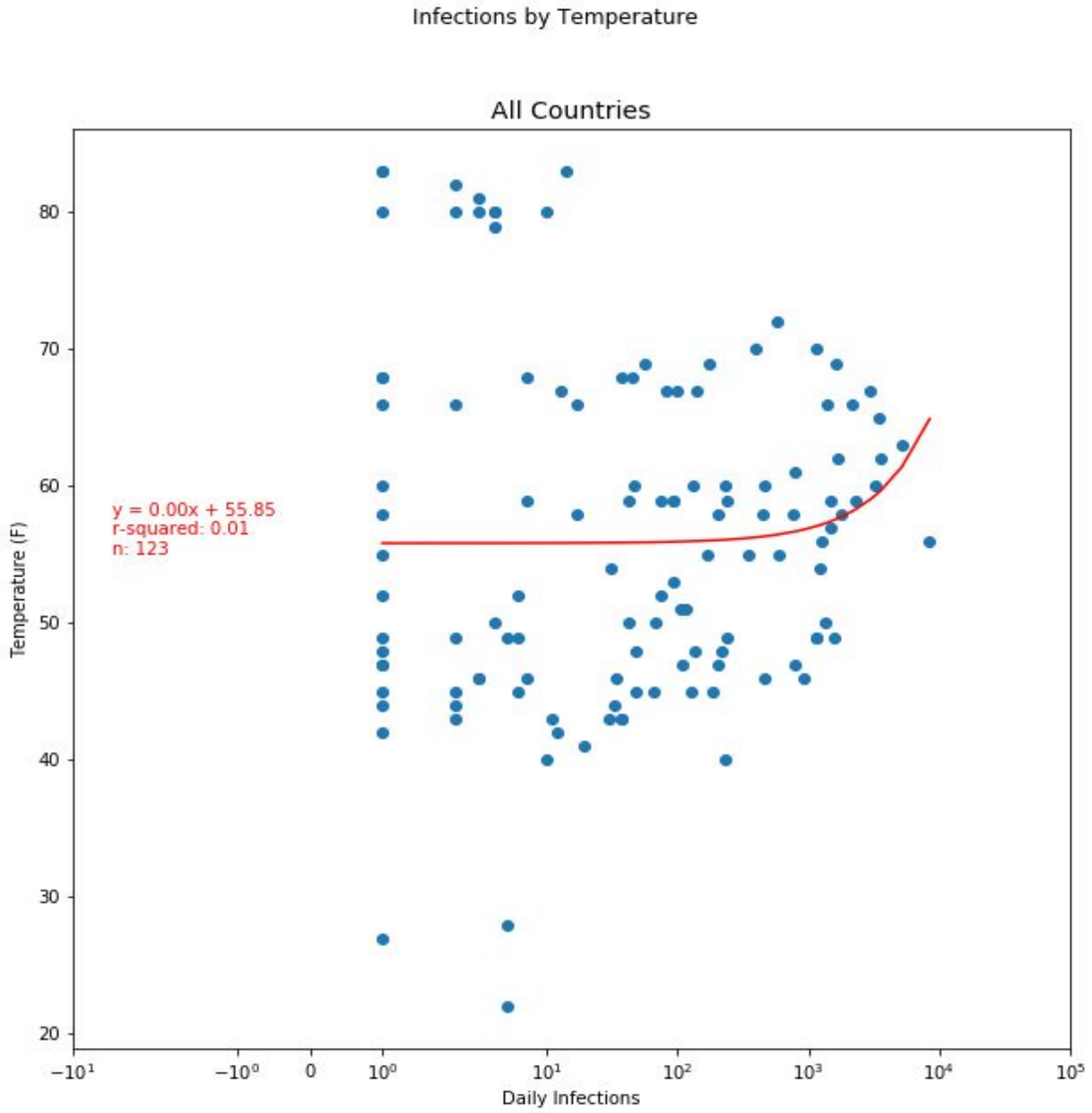
- Males have higher reported recovered cases



TEMPERATURE & INFECTIONS

Daily Average Temperatures (F) from 2/1/20-3/15/20

Countries (7):
Sweden, Italy, Spain, Germany, Malaysia, UK*, NY*

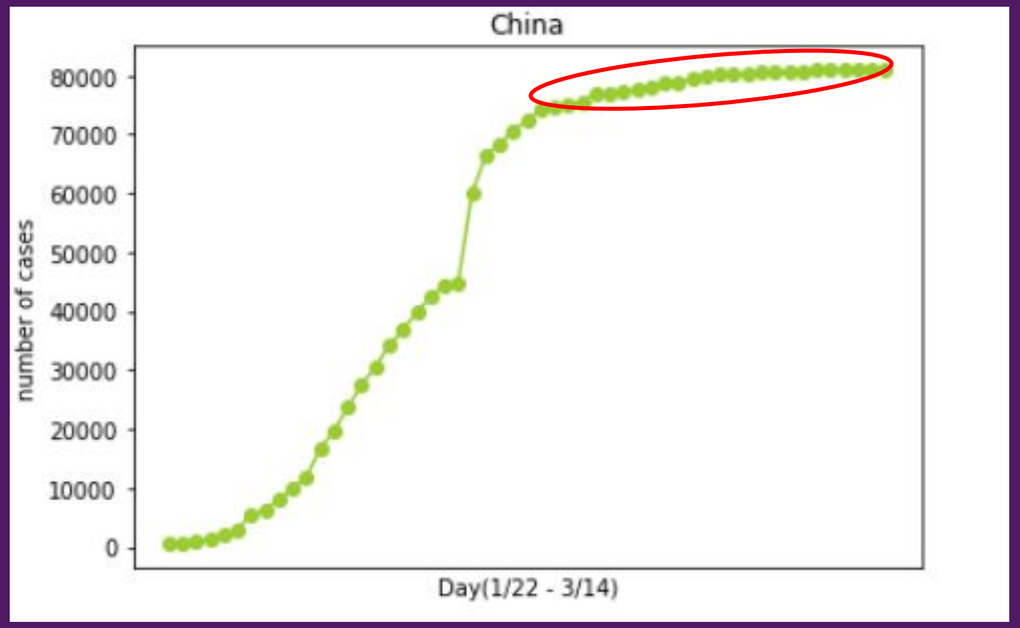


- Temperature and daily infections (confirmed cases of Coronavirus) are not correlated (r-squared = 0.01)

Temperature Data Source:
https://www.ncdc.noaa.gov/sotc/global/201913

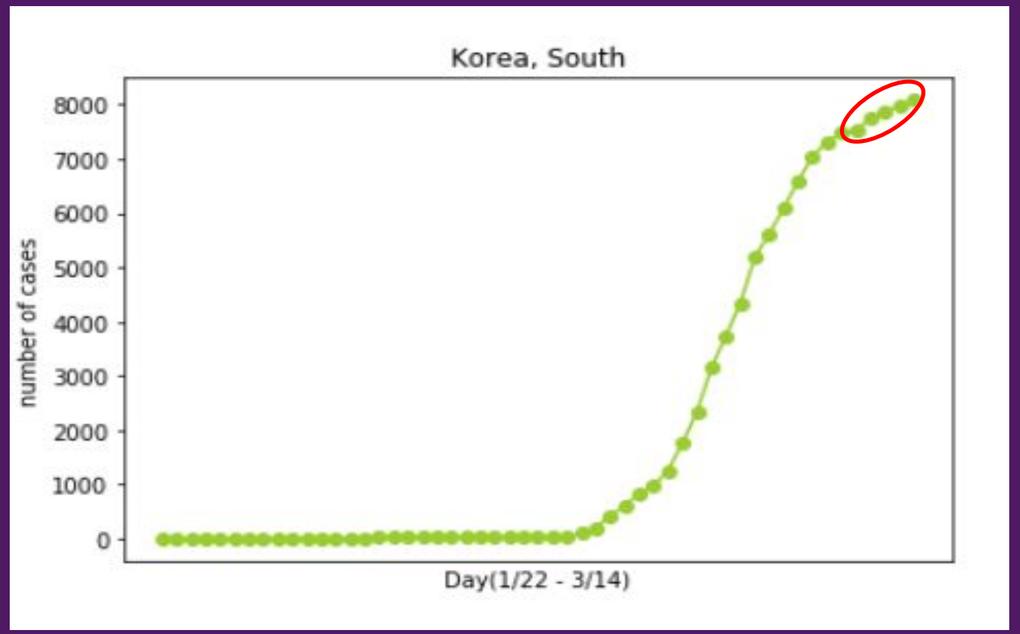
Google Heat Map

Number of Confirmed Cases

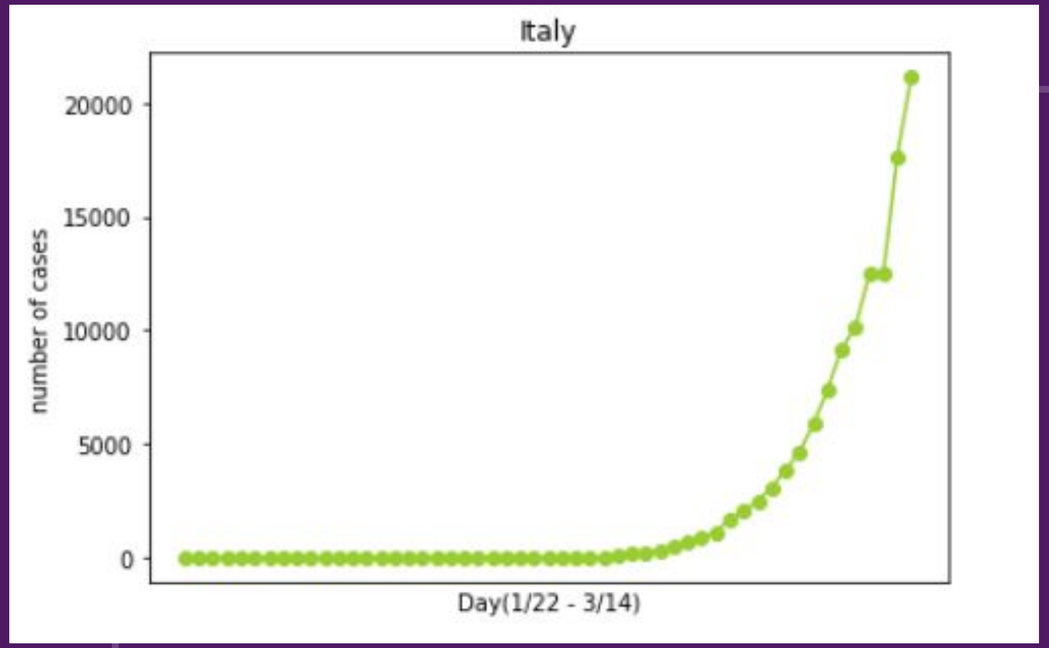


Time Series

Confirmed Cases by Day

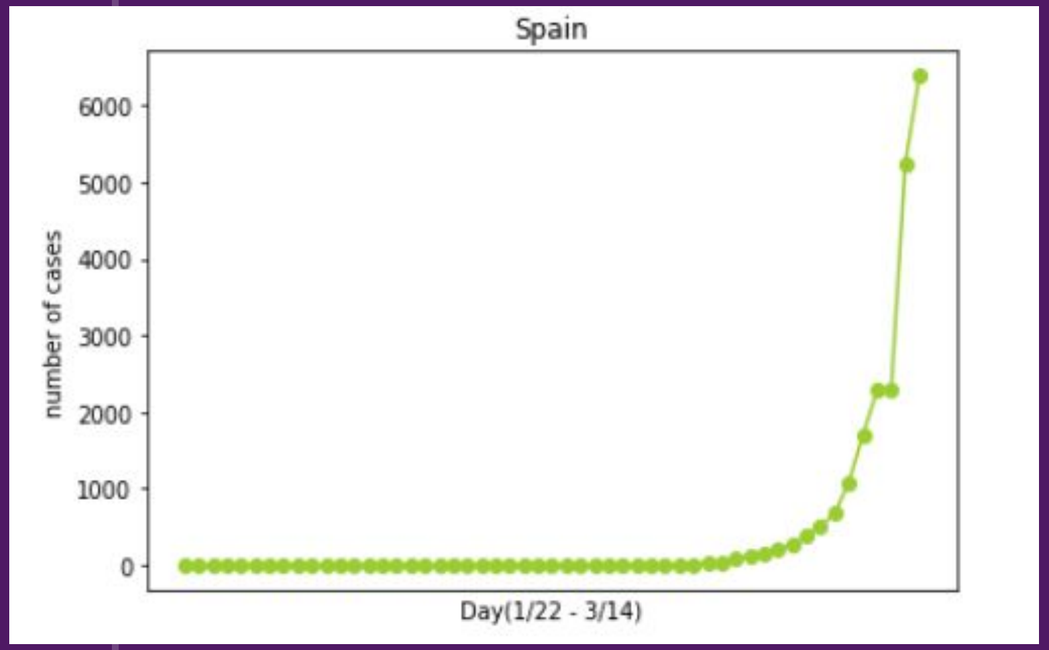


- China & South Korea already experienced a rapid growth period and reached a point the curve started begins to taper & the growth speed slows.

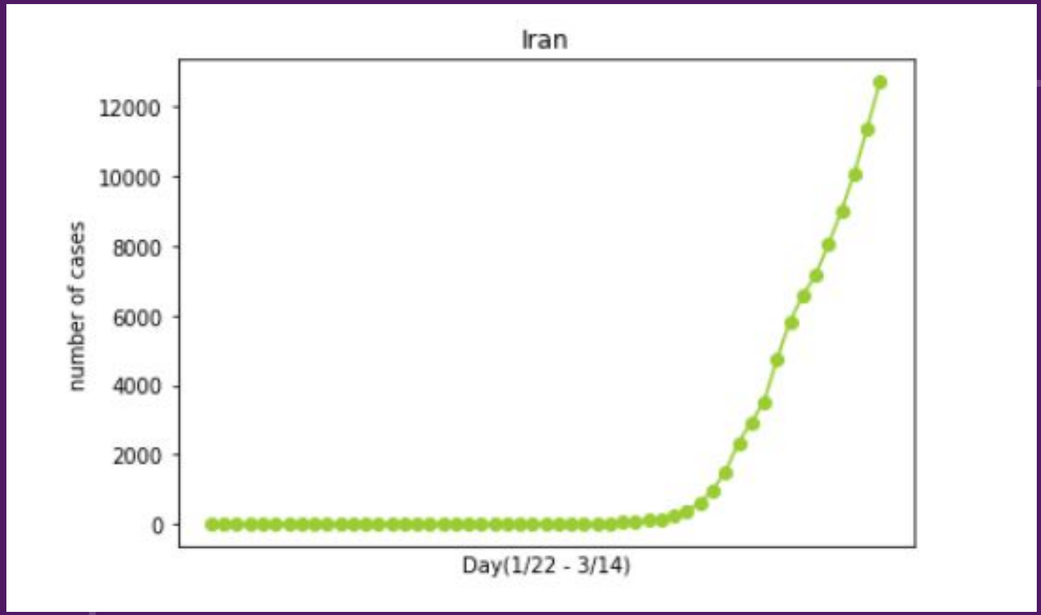


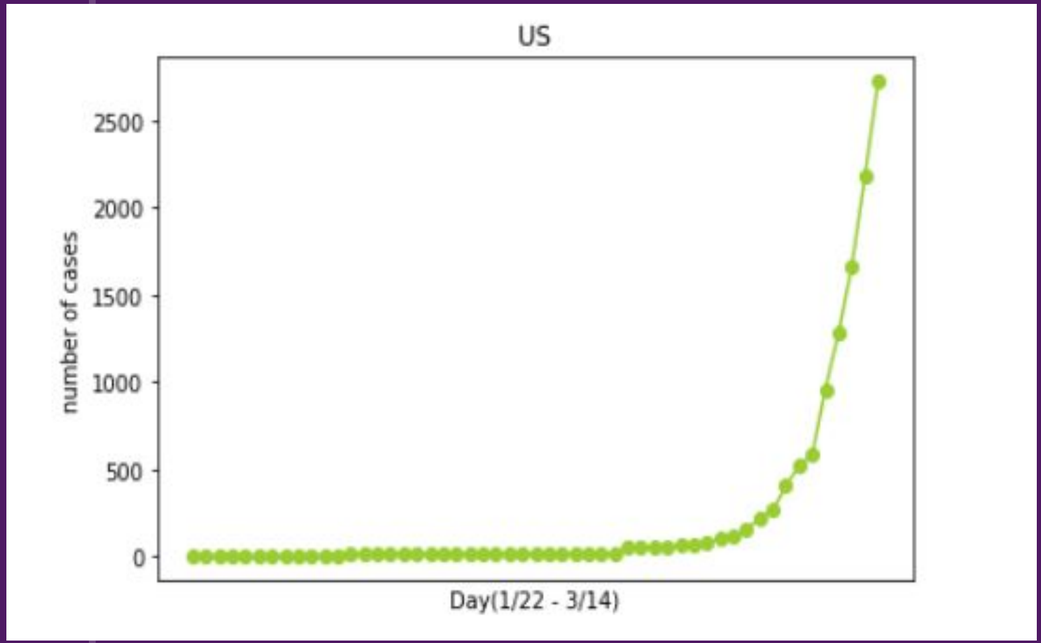
Time Series

Confirmed cases by Day



- Spain & Italy experienced some sharp increases w/cases increasing significantly over the course of day.

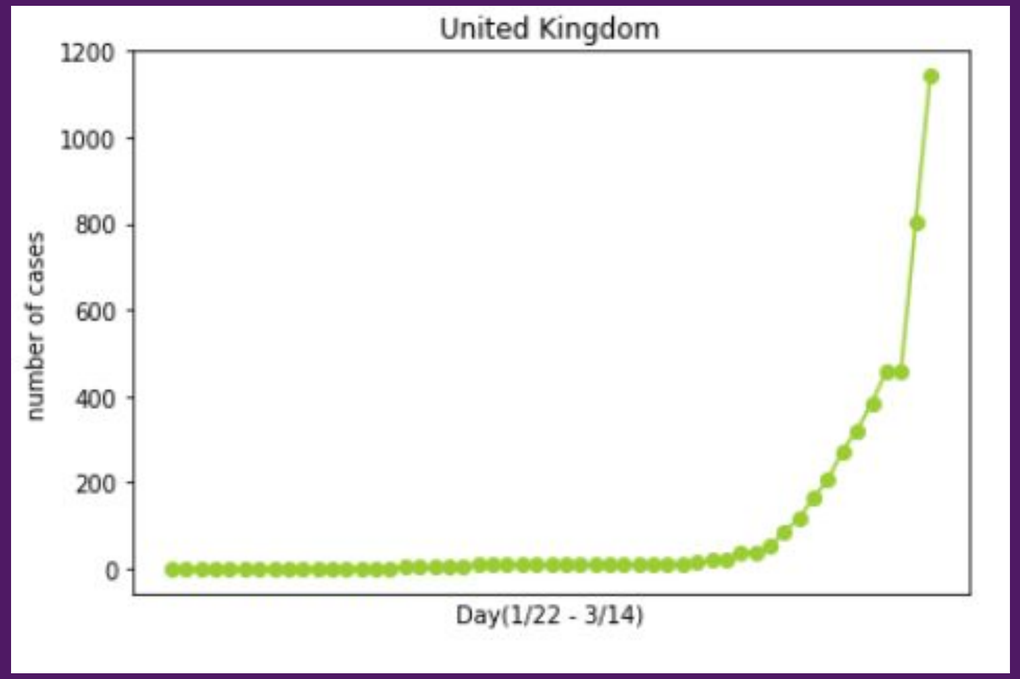




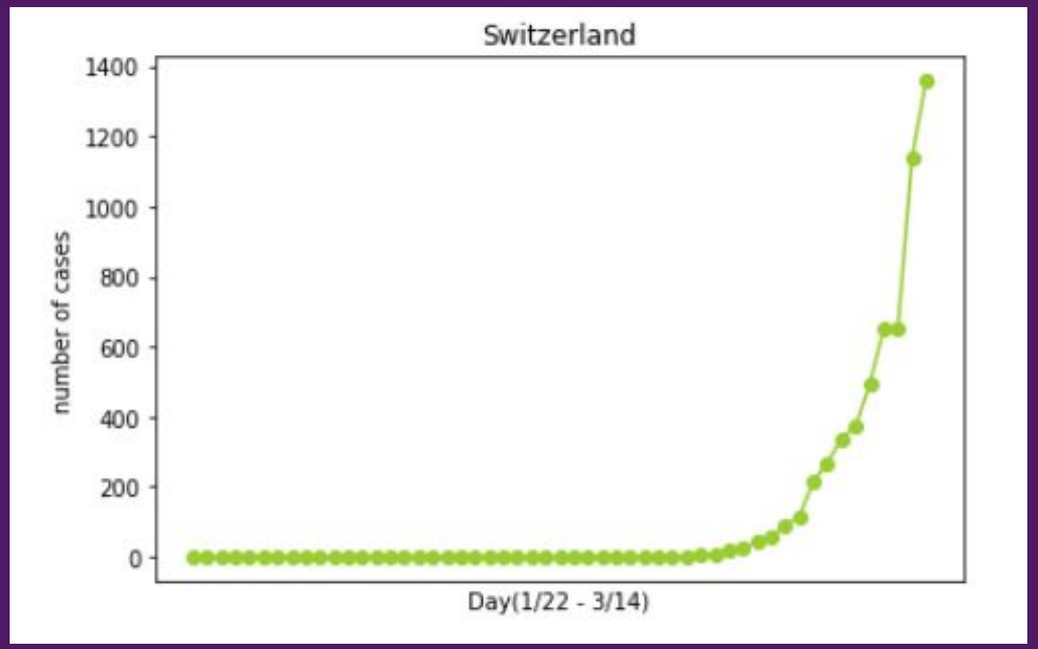
Time Series

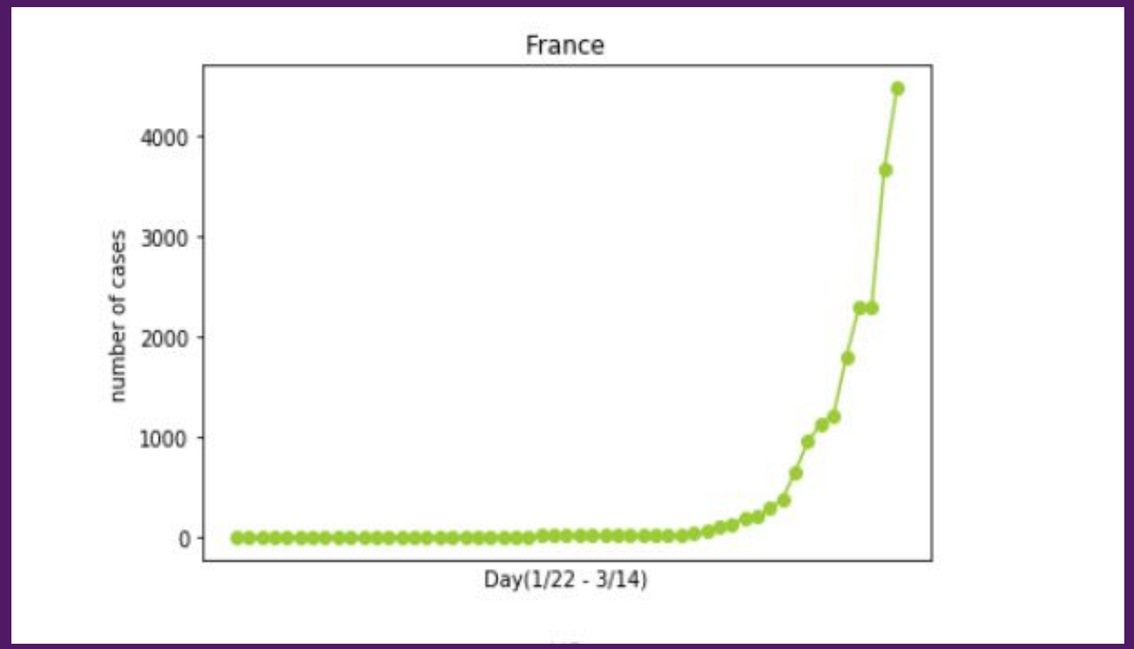
Confirmed cases by Day

- Certain countries seem to increase at a more stable and consistent rate.



Time Series

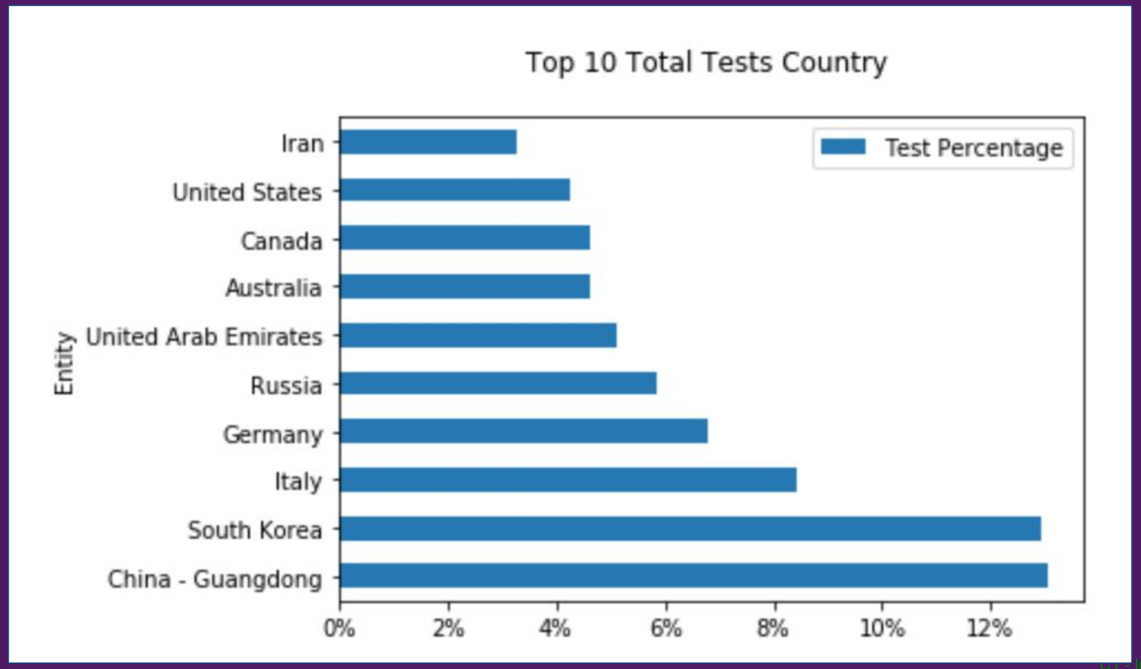


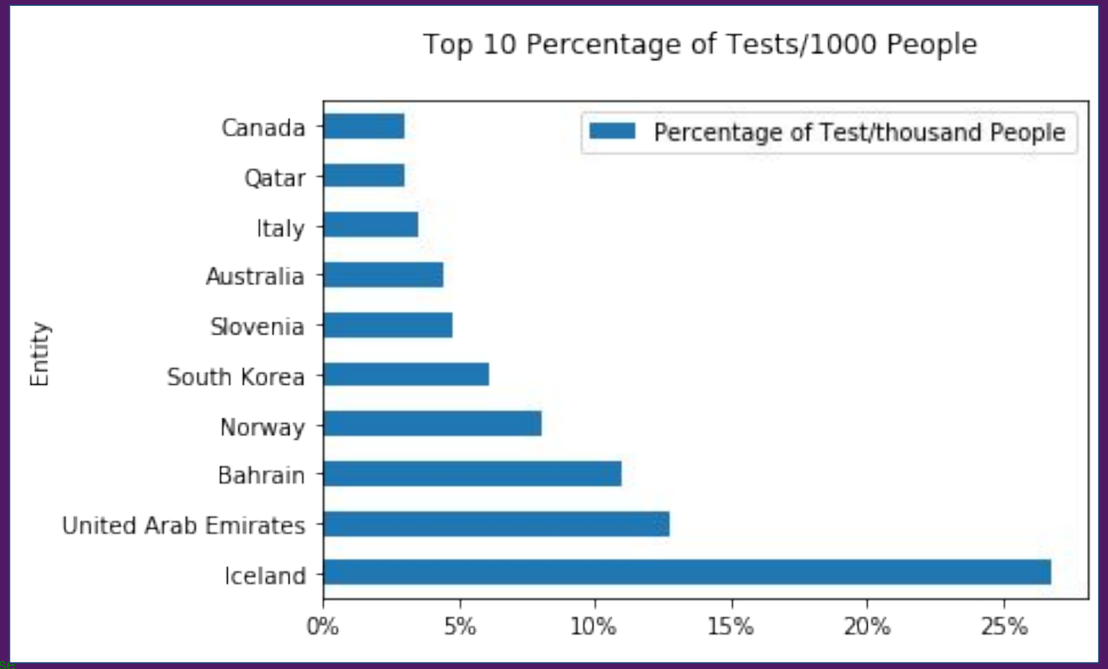


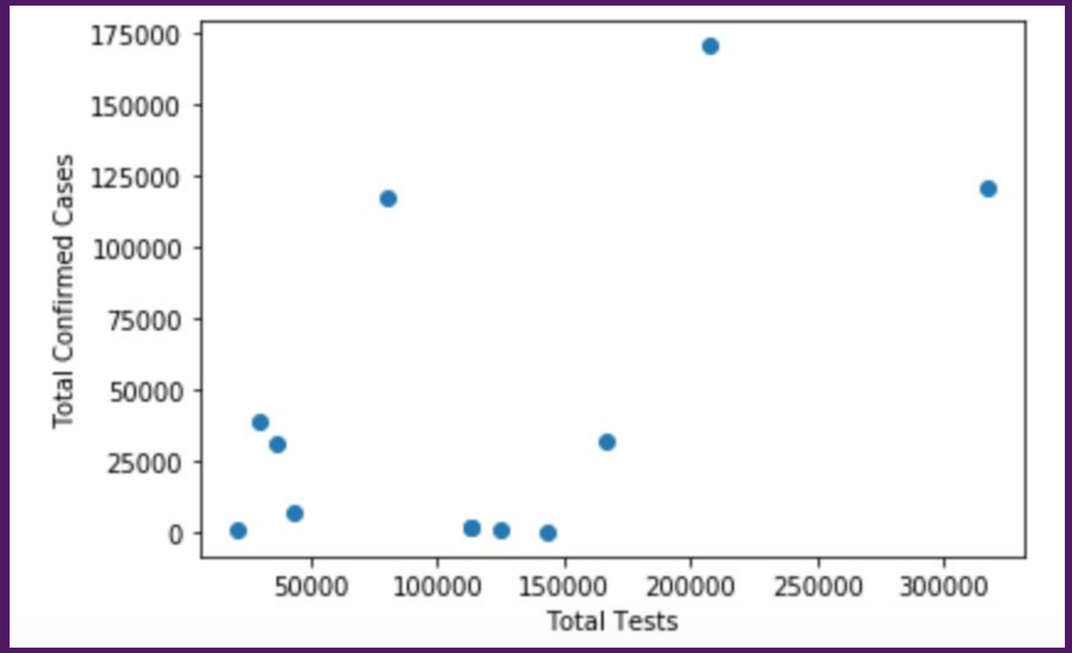
- Top 10 countries of confirmed cases: China, South Korean, Italy, Iran, Spain, Germany, France, US, Switzerland, UK
- Testing Capacity
- Data Limitation
- % of population been tested

Test Rate

Whether # of tests have an impact on confirmed cases?



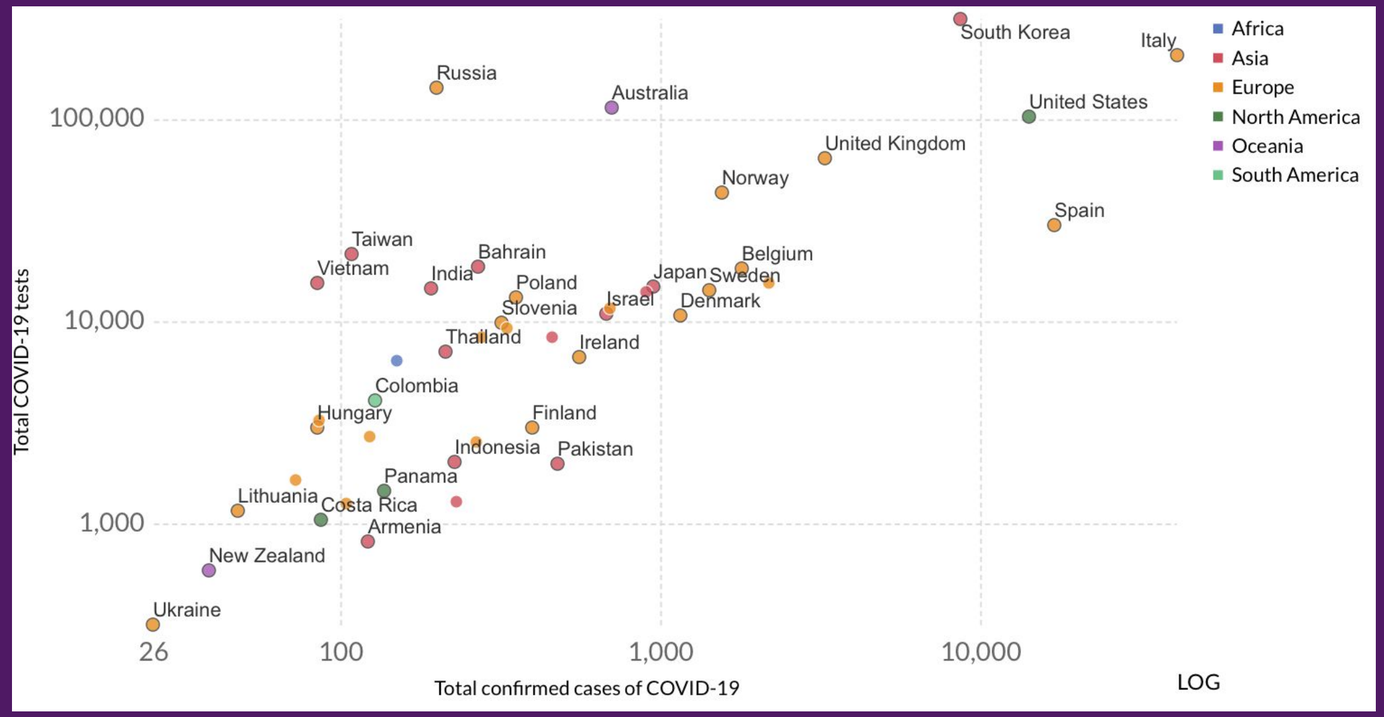




Test Rate

Correlation between total tests VS Confirmed cases of top 20 countries

- No strong correlation between total tests VS confirmed cases



RESEARCH CONCLUSIONS

DEMOGRAPHIC

Which age group has a higher recovery rate?

Age groups less than 60 yrs old have a higher chance of recovery.

Which gender has a higher chance of recovery?

Our data shows that males have more reported cases of recovery.

Time Series

Which countries have been successful in terms of "flattening the curve?"

China and South Korea have experienced flattening of the curve.

Italy, Spain and China experienced some of largest one day spikes.

TEMPERATURE

Does a country's temperature impact the amount of infections is has?

The results are inconclusive, although the available data suggest weather and infection rates aren't related.

To strengthen this analysis, it would be helpful to look at more data points for each country, a greater variety and amount of countries (representing all continents), multiple measures of weather (humidity, precipitation) and extend the time frame beyond one month to detect seasonality.