CITADEL | CITADEL | Securities

West Coast Regional Datathon
Team 6 Report

Focus

We focused our analysis on determining geographical trends for COVID Spread in Europe, analyzing the spread of the disease and the mortality rate through the prism of social mobility, and quantifying the impact of political affiliation with respect to the spread of COVID in the United States.

Death toll findings in Europe

When discussing COVID-19 impact on human life in Europe, we observe several types of patterns emerging amongst different countries. We observe these patterns correlate most strongly with geographical proximity, and see a clear divide at different granularity levels between Western and Eastern Europe.

Methodology

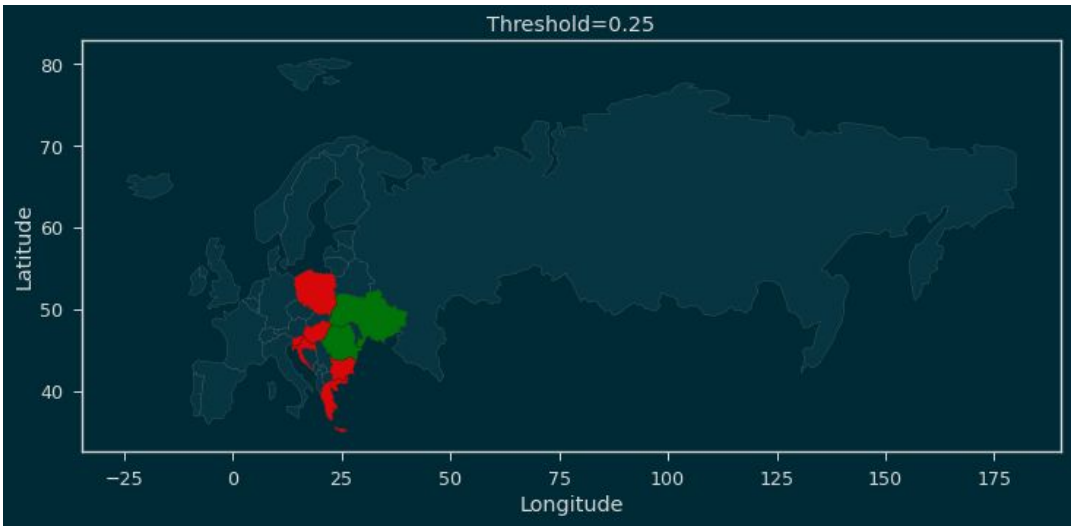
We consider the OWID covid dataset, and focus on *new_deaths_smoothed* to quantify direct COVID impact on human life. We prune outlier countries from our analysis, for which data is not reported in at least 20% of considered dates. For the remaining countries, we clip the reported deaths to 0, and fill any nans with 0. We normalize the number of deaths over the reported period, to be able to directly analyze the trend of deaths.

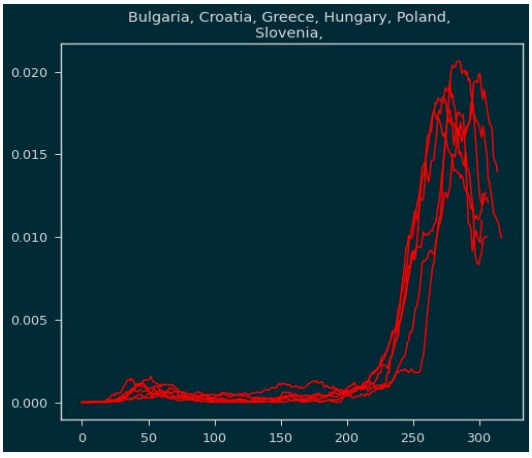
When comparing the similarity between two countries, we compute the edit distance between the normalized number of deaths. We then set different similarity thresholds for determining which countries are similar, and report the connected components as clusters of similar countries. For different thresholds we observe different clusterings of neighboring countries.

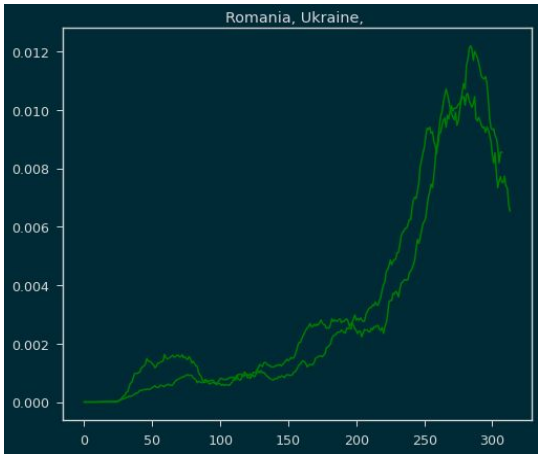
We also investigate how / if different measures deployed across clusters affect the outcome on the number of deaths. We perform a qualitative analysis focused on clusters formed by the .3 threshold.

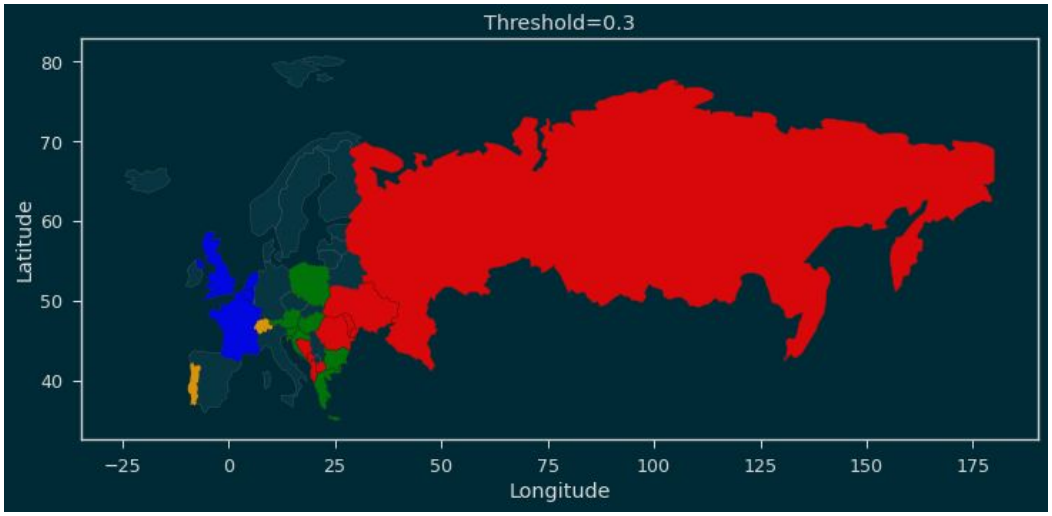
Results

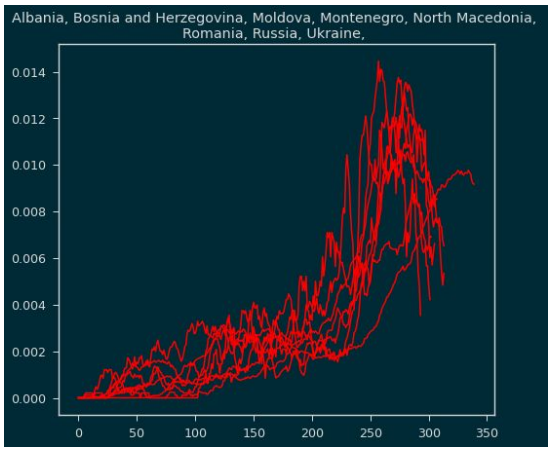
We first report the cluster formations at different levels of abstraction. We selected a threshold between .1 (every country is independent) to .5 (all countries are part of one cluster). Recall these values are only representative with respect to the distance function we chose, so a different distance metric may require different cut-off points.

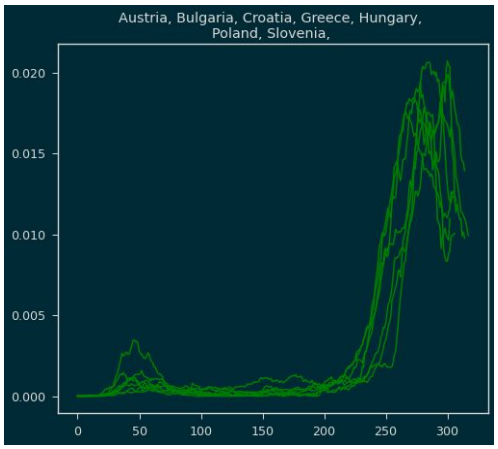


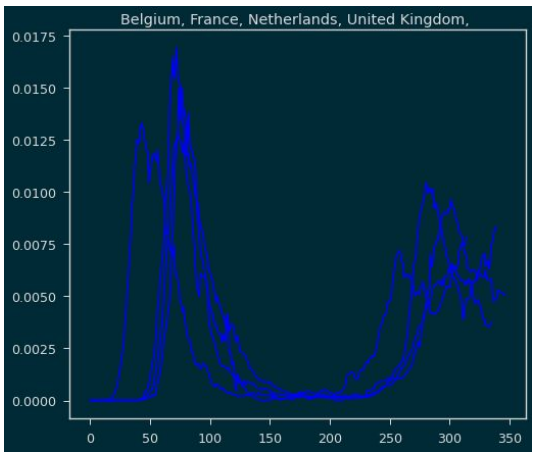


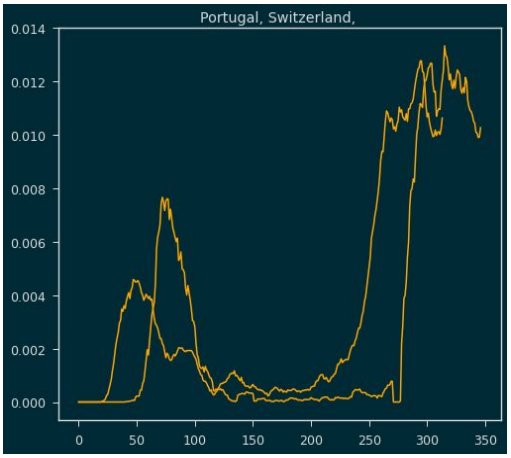


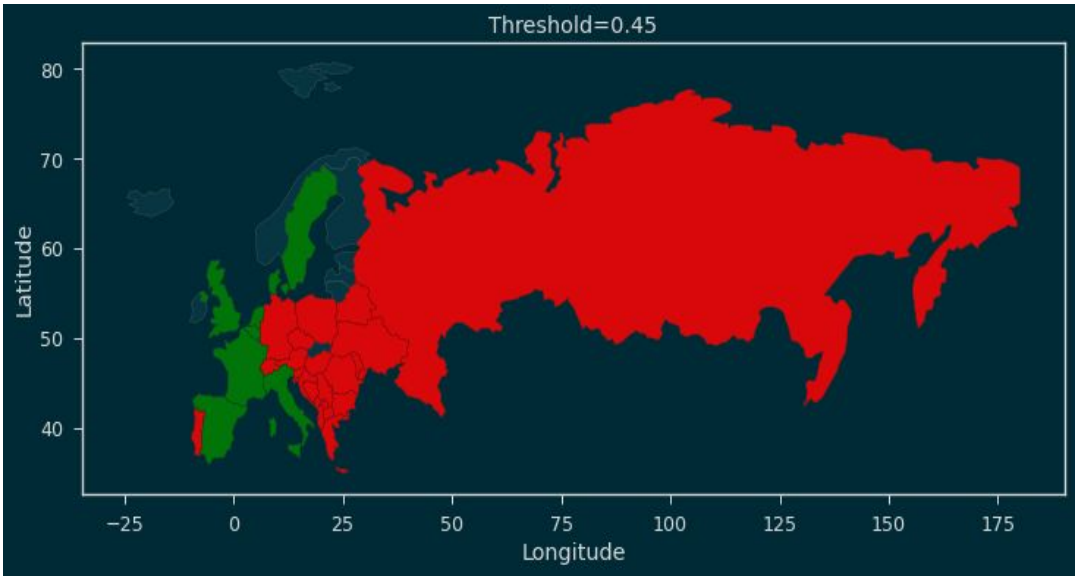

Threshold=0.3
Latitude
Longitude
80
70
60
50
40
−25
0
25
50
75
100
125
150
175

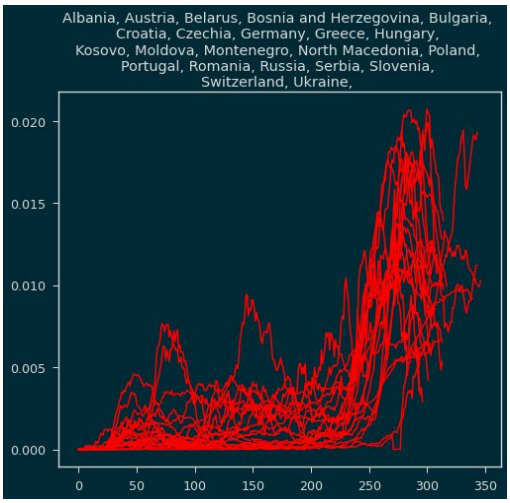

Albania, Bosnia and Herzegovina, Moldova, Montenegro, North Macedonia,
Romania, Russia, Ukraine,
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
50
100
150
200
250
300
350

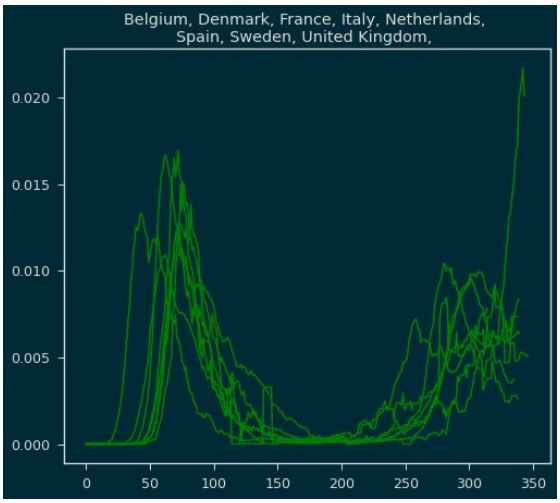

Austria, Bulgaria, Croatia, Greece, Hungary,
Poland, Slovenia,
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300


Belgium, France, Netherlands, United Kingdom,
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
50
100
150
200
250
300
350


Portugal, Switzerland,
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
50
100
150
200
250
300
350

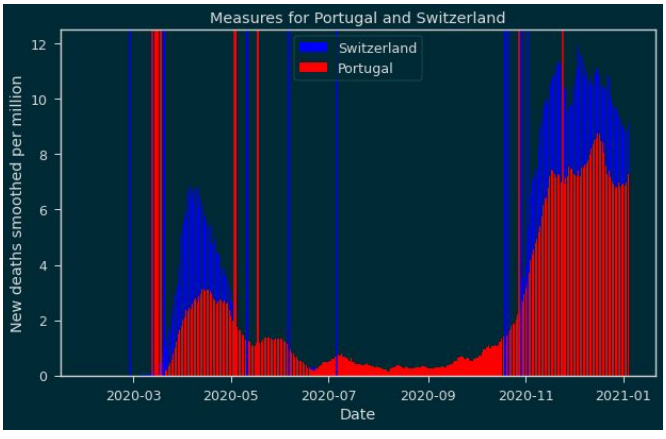




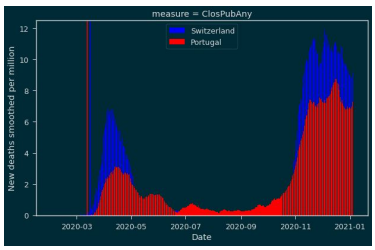


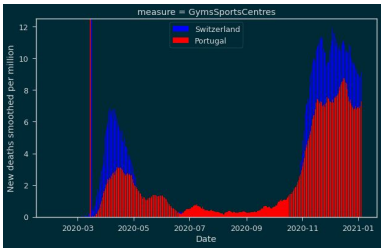
We observe the most detailed clustering forms for the .3 threshold, where Russia and Eastern European countries share similar trends in deaths. We also observe Central and Southern European countries sharing a smaller relative number of deaths after the ‘first wave’ and the larger number of deaths after the ‘second wave’ of COVID.

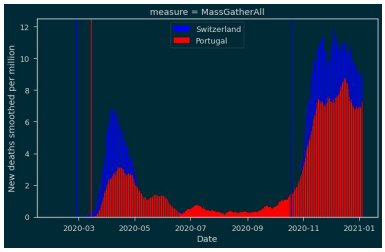
In the case of countries with higher impact from the second wave, we consider the cluster formed by Portugal And Switzerland. We investigate which measures - if any, had a strong or rather lack of impact on the final death toll. To the left we plot the measures in relation to the COVID fatalities, vertical lines representing the measure implementation date, and bar plots representing *total_deaths_smoother_per_1_million*. This indicates that the measures in question

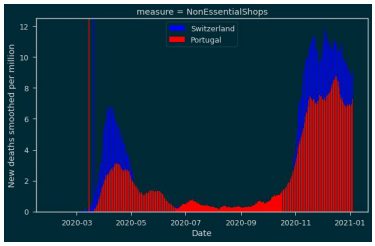


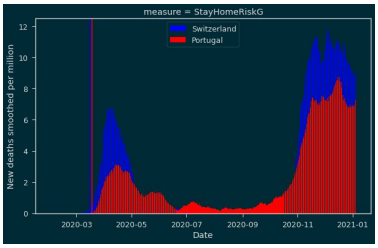
had a limited effect on the spread and ultimately fatality rate due to COVID, assuming the data obtained for the first wave is accurate. This conclusion can be further supported if we look at individual measures with respect to the fatality rate.



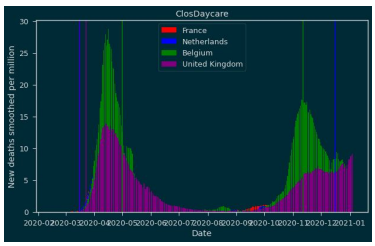


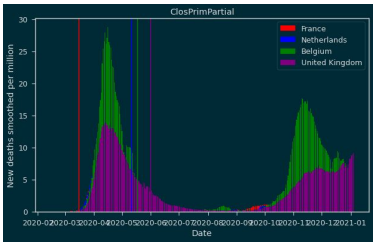


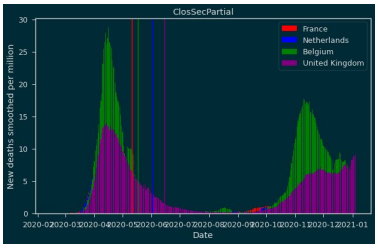


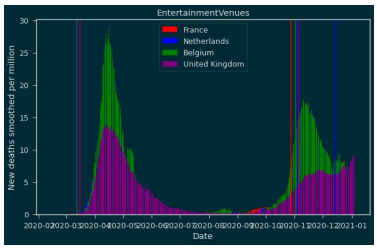


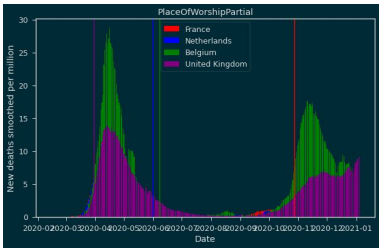
We also observe that different clusters of countries were able to mitigate the second wave of covid. These are France, Belgium, The Netherlands and The United Kingdom. Observing only measures implemented in this set of countries that were **not** implemented in Switzerland or Portugal, we find the following implementation timeline.









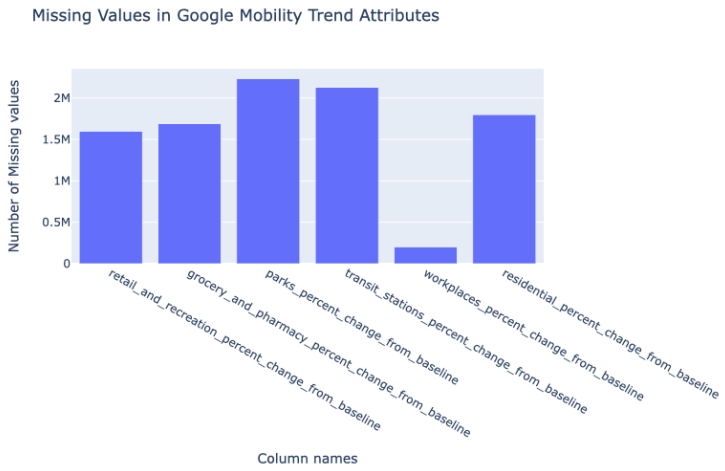


In addition to mandatory mask policies implemented by a subset of these countries, the above measures were the only common amongst all. We observe all these measures are focused on limiting large gatherings - but in a more strictly controlled manner. This entails that while most countries implemented a ban on gatherings above a certain threshold, such measures are hard to enforce. Moreover, daycare centers and places of worship offer situations

where the contact between strangers is closer, compared to a bar or gym where even if the virus spreads, it is less likely to affect the whole population present.

Mortality and Google Mobility Reports

The Global Google Mobility Reports dataset provides lots of insight into what has changed in response to policies aimed at combating COVID-19. The reports chart movement trends over time by geography, across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential locations. Location accuracy and the understanding of types of categorized places varies from region to region, so we didn't use this data to compare changes between countries but used it as a good metric for changes across the continents as a whole. When preprocessing the dataset, we also noticed that there were many missing values in the trend attributes, as indicated in the figure below. We also noticed that much of the missing data came from Britain, Colombia and Turkey. However, the overall quality of the data was satisfactory and provided interesting results.

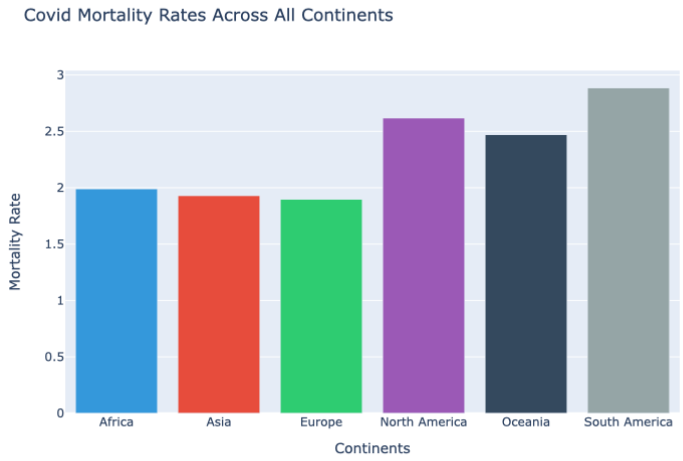


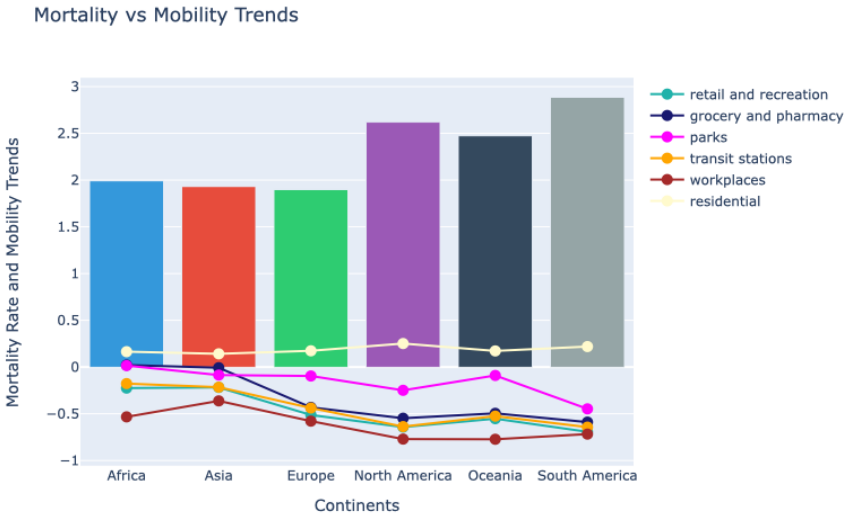
In particular, the heatmap below shows several salient relations between the mobility trends. The *residential_percent_change_from_baseline* feature has negative correlations with the rest of the attributes, indicating generic global lockdown and quarantine measures in the majority of the countries forced people to stay in residential areas. Additionally, the *parks_percent_change_from_baseline*, surprisingly, does not show any strong correlations with other features; this indicates an increase in outdoor involvement during the pandemic as many countries advised against indoor activities. However, there is a subset of countries where park visits dropped dramatically, as many countries enforced bans on park visits. Finally, the rest of the attributes displayed strong pair correlations: the lockdown restrictions resulted in a drop in visits around all of the attributes except residential.

Methodology

We wanted to determine if the mobility of people during COVID-19 had any effect on the mortality rates across the continents. In order to do this, we utilized the *wikipedia-iso-country-codes.csv*, as well as the ACAPS Government Measures dataset, the provided *owid-covid-data.csv*, and the Global Mobility Report. We merged and wrangled the datasets, including adding a mortality feature that was calculated as the percentage $total\_deaths/total\_cases * 100$ and replacing any numerical NaNs with 0's. We then grouped the data by calculating the mean mortality rate across all countries in each continent. The resulting data is shown below.

continent	Africa	Asia	Europe	North America	Oceania	South America
census_fips_code	NaN	NaN	NaN	NaN	NaN	NaN
retail_and_recreation	-0.225000	-0.217656	-5.129630e-01	-6.420000e-01	-0.552500	-6.927273e-01
grocery_and_pharmacy	0.021500	-0.007187	-4.316667e-01	-5.475000e-01	-0.495000	-5.890909e-01
parks	0.015610	-0.085781	-9.555556e-02	-2.490000e-01	-0.090000	-4.472727e-01
transit_stations	-0.177000	-0.215469	-4.381818e-01	-6.370000e-01	-0.525000	-6.427273e-01
workplaces	-0.532927	-0.362031	-5.785455e-01	-7.705000e-01	-0.772500	-7.172727e-01
residential	0.164500	0.141563	1.738889e-01	2.515000e-01	0.172500	2.190909e-01
total_cases	101990.634146	451454.437500	1.293361e+06	1.091424e+06	7812.750000	1.164294e+06
total_deaths	2262.024390	7678.868852	2.514582e+04	2.448050e+04	236.000000	3.225382e+04
mortality	1.989693	1.928943	1.897117e+00	2.617202e+00	2.471326	2.886785e+00



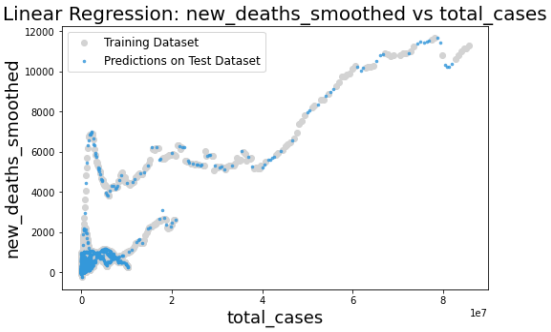


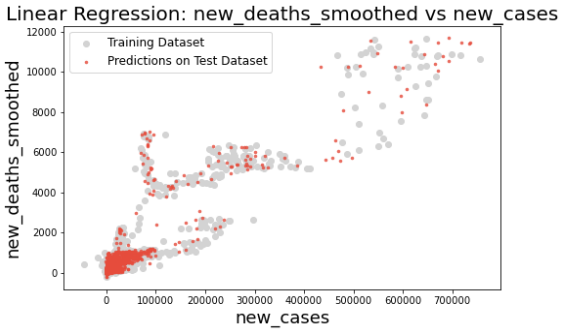
Results

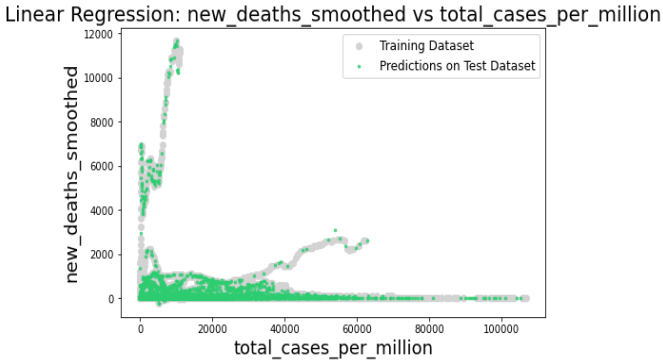
From the bar graphs above, it is clear Europe has the lowest mortality rate - however, in comparison to North America, it seems that most people were still mobile for all attributes (except residential, indicating high mobility everywhere else). In addition, South America maintained some of the most restrictions and saw a large decrease in mobility yet has the highest mortality rate. There is no clear trend between mortality and mobility: Europe as a whole generally had less of a decrease from the baseline in mobility yet has the lowest mortality, while North America had more of a decrease from the baseline in mobility and has the second highest mortality rate.

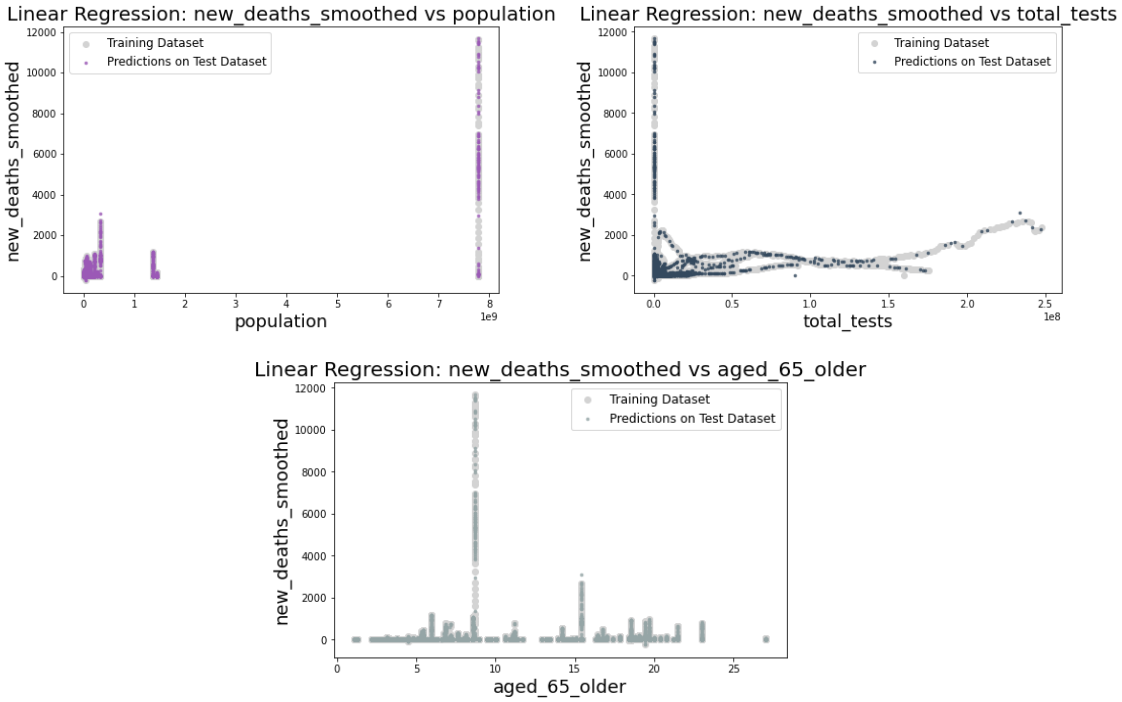
Linear Regression Modeling Results

After concluding that there is no suggestive trend between mortality and mobility rates, we wanted to look closer into modeling new deaths. The provided OWID covid dataset which includes data on confirmed cases, deaths, testing, and more, contains many variables of interest for linear regression analysis to predict the *new_deaths_smoothed* feature. In this case, we used Multiple Linear Regression (MLR) with explanatory variables ['*total_cases*', '*new_cases*','*total_cases_per_million*','*population*', '*total_tests*', '*aged_65_older*'] to predict the response variable *new_deaths_smoothed*. We split the data into 67% training and 33% testing. A mean absolute error of 0.26 and an R-squared of 1.000 was recorded. The prediction plots are included below:









American Findings

Looking at the spread of COVID-19 throughout the United States, we analyzed a variety of different patterns potentially connecting how the disease transformed across the country. One pattern that we looked at more closely was how the political affiliation of states may have affected how the states responded to the COVID-19 outbreak.

Methodology

Initially, we utilized a presidential dataset published by Harvard Dataverse in order to break down how each state was categorized. A state was labeled “DEMOCRAT” or “REPUBLICAN” based on the majority popular vote during the 2020 presidential election. This allowed us to have a general understanding of a divide between “blue” or “red” states, which would be used to compare a variety of different factors.

In the current political climate, the concept of quarantine is heavily argued between the different parties. As seen in multiple polls over the past year, like in this poll conducted by Gallup, Democrats were more likely to support and follow concepts of social distancing when compared to Republicans. However, would this data translate to real life actions? We analyzed the Google mobility reports dataset mentioned previously to access this information. Two key features we looked at were *residential_percent_change_from_baseline* and *retail_and_recreation_change_from_baseline*. These two features would allow us to determine how different political states followed strongly recommended CDC guidelines.

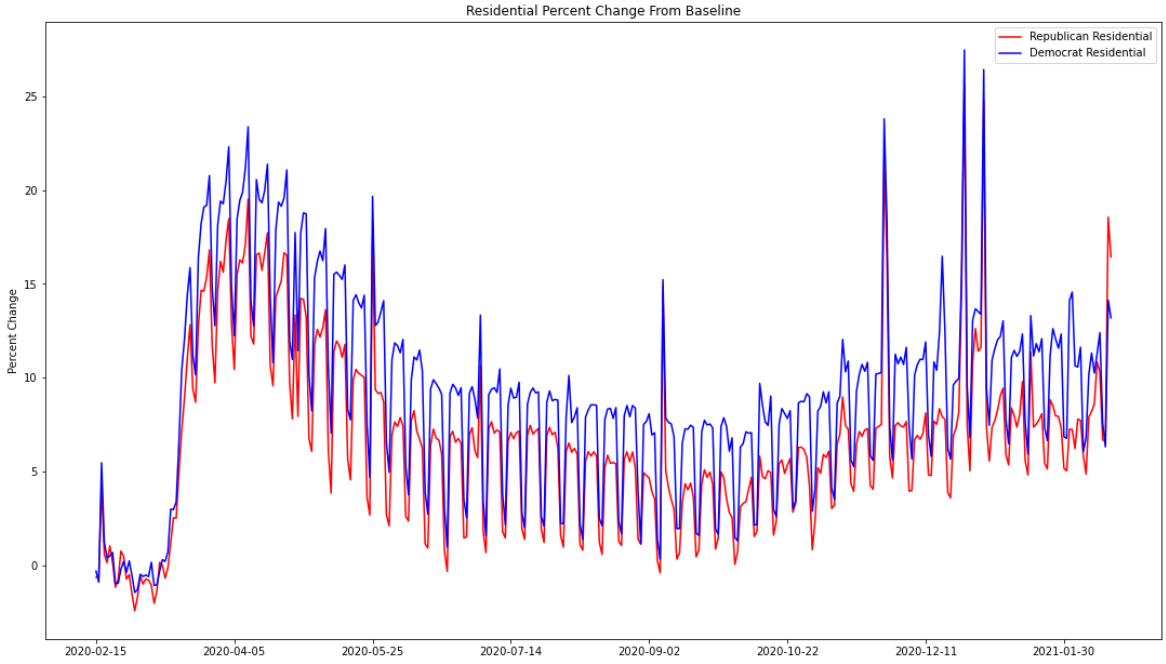
To construct these graphs based on these features, we merged the following datasets:

- Presidential dataset (to determine Democrat/Republican status)
- Google regional mobility dataset
- Covidtracking state-level datasets, given by the contest organizers

These datasets were checked for null values or NaNs, which would then be cleaned and discarded. We also noticed several 0 values in the state-level datasets when looking at the total number of tests done. These values were discarded and nulled.

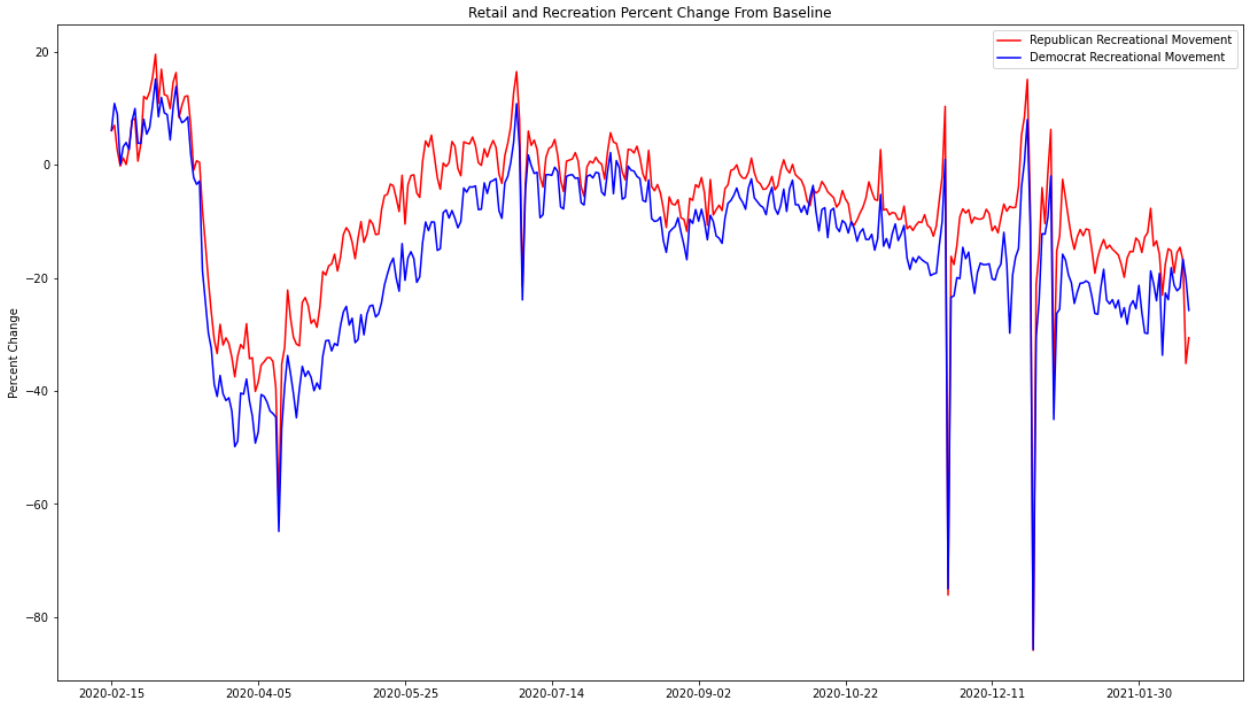
We grouped entries within this dataset based on date and party affiliation. These entries were then averaged across the different states to obtain an average degree of social distancing based on the features selected above. Other features either did not contain as much relevant data or would not support evidence to social distancing (eg going to parks is a safe social activity, so would not be as relevant to this use case).

Afterwards, we compared the percentage of positive tests in Democratic and Republican states across the United States. This allowed us to consider factors that may have been overlooked in the previous social distancing analysis, such as some Democratic states may have stricter guidelines (like California, New York) because of their location as border to the outside world, especially in comparison to states like Oklahoma, landlocked states which are less likely to have the virus enter through international means. However, we will discuss this further in the analysis below.

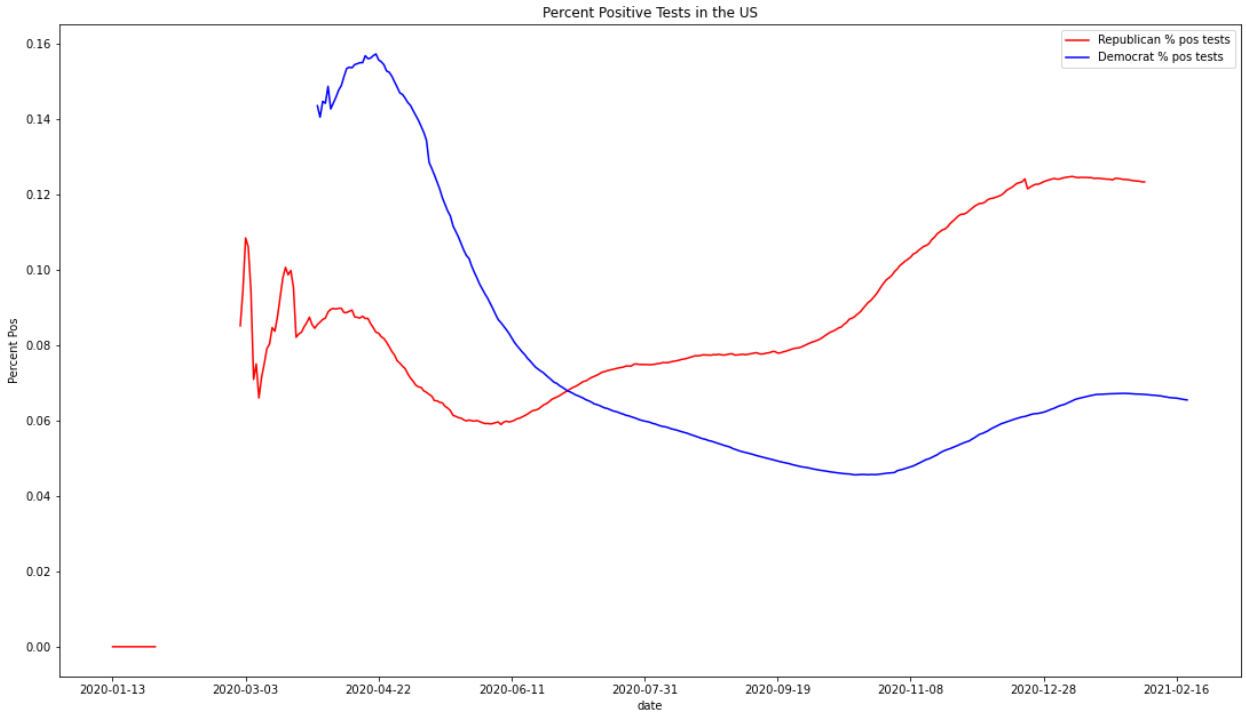


The above graph shows how citizens of Democrats and Republican states spent time in their respective private residences when compared to normal baselines. As seen in the visualization, we can tell that Democratic states spent a higher percentage of time in their

homes, following the shelter-in-place orders given by either state or national governments. This is supported by more data shown in the graph below, where it is evident that Republican states are more likely to go out and shop or conduct recreational activities. This supports the notion that Republican states are less likely to care and follow social distancing guidelines, inferred from the polls taken across the US.



But how does this difference of attitudes towards social distancing affect the spread of COVID-19 across the United States? We explored this in terms of positive tests/number of tests done and compared the results between Republican and Democratic states in the graph below. As can be evidenced, Democratic states had a much larger spike of positive cases, likely due to the nature of Democratic states and how international travelers/immigrants are likely to arrive at these states first, bringing with them the threat of the COVID-19 virus. However, as time passed on, Democratic states were evidently able to bring the rampant spread of the virus a bit more under control, with heavier testing and lower positive rates. We can contribute this in part to the more diligent social distancing practiced by Democratic citizens. Without it, this sort of stop would be impossible. Republican states, in contrast, have been heavily affected by the lack of government regulation, with the rate of positive cases continually increasing, even with more accurate tests and ease of testing.


Percent Positive Tests in the US
Republican % pos tests
Democrat % pos tests
Percent Pos
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
2020-01-13
2020-03-03
2020-04-22
2020-06-11
2020-07-31
2020-09-19
2020-11-08
2020-12-28
2021-02-16
date