Analyzing and Predicting COVID-19 Trends in Germany

Derek Yin and Shahrizod Bobojonov

Abstract - **COVID-19 is an infectious disease caused by a particular strain of coronavirus first originating in humans in Wuhan, China. On January 27th, 2020, an employee working at a Bavaria-based company was confirmed to have contracted the first confirmed case of COVID-19 in Germany. On March 11th, 2020, COVID-19 was declared a global pandemic by the World Health Organization (WHO). Over the course of a year, the virus spread rapidly across every German state and county. In this project, we seek to analyze county-level COVID-19 data to observe the differences in infection and death rates among differing gender and age-groups. Additionally, we aim to accurately predict the survival of a COVID patient through several machine learning models. The results are promising and demonstrate the importance of data analysis in handling crisis.**

I. DATASET PARSING, CLEANING, & MERGING

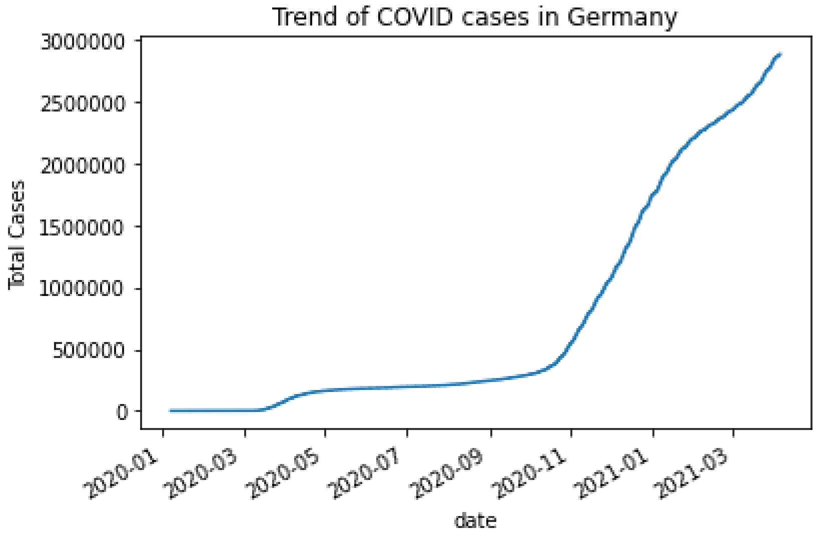
We utilize Google Colaboratory to effectively collaborate on a single Python notebook. The first step involves converting raw csv files into pandas dataframes. We store the two datasets, one named “covid”, and the other named “demographics,” into separate dataframes. Merging these datasets will help us consolidate important information, but the data must be parsed and cleaned before doing so. Firstly, all rows in either dataframe containing null values are dropped. Secondly, we drop any duplicate rows in the demographic dataframe. After some consideration, we leave any duplicate rows in the covid dataframe as they are, because there are no differentiating factors in the data that can confirm if duplicate rows in this dataset were erroneously added or simply describing a different COVID patient.

To prepare these datasets for merging, we parse through the demographic dataframe, (the smaller dataset with consideration for runtime), and replace any instances of ‘male’ or ‘female’ in the gender column with ‘M’ or ‘F’, respectively. This matches the data signature of our larger covid dataset.

Finally, we merge the datasets on state, age group, and gender, using an outer join. After dropping any null/duplicate rows from this new dataset, we are left with a dataset that, for every patient, includes their state, county, age group, gender, date, number of cases, number of recoveries, whether or not the patient died, as well as the total population of patients that shared the same county, gender, and age group.

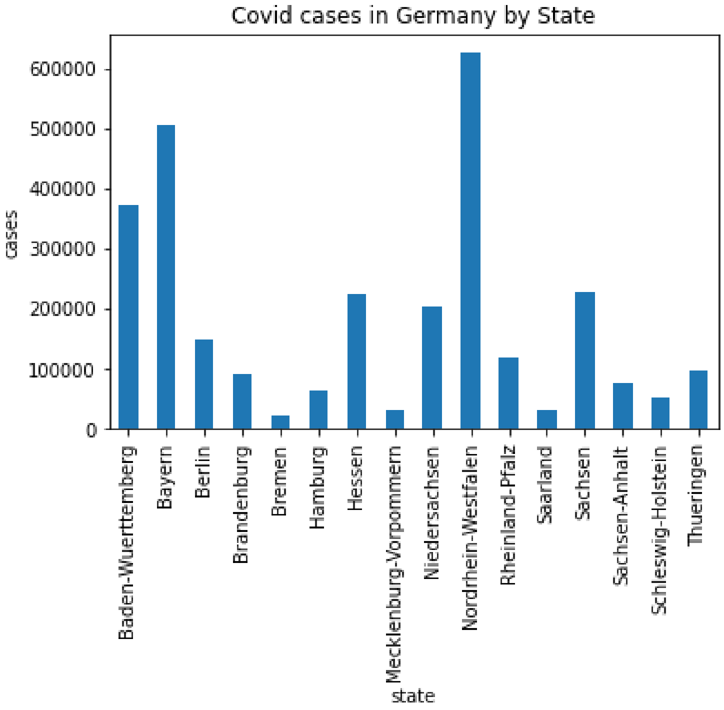
II. ANALYSIS

Now, we want to observe COVID case trends on a state-level basis. To do this, we construct a bar graph that takes the sum of all confirmed cases per state. This is achieved by first grouping the dataframe by state and taking the sum of cases for each state. We store this in a new dataframe which is then used to construct a bar graph.



The x-axis shows us the progression of time, from January of 2020 up until early 2021. The Y-axis shows us the cumulative sum of total COVID-19 cases. The frequency of new COVID cases began to increase drastically in late October to early November of 2020, and continues to rise at the same rate up until early 2021.

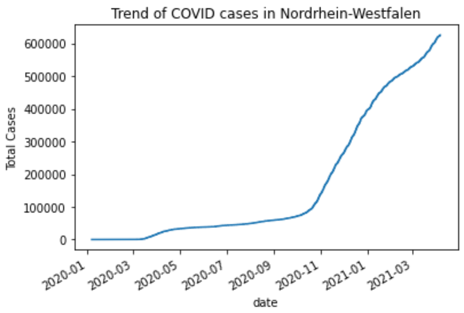
The second baseline analysis we perform is a count of COVID cases for each German state. The corresponding graph is shown below.

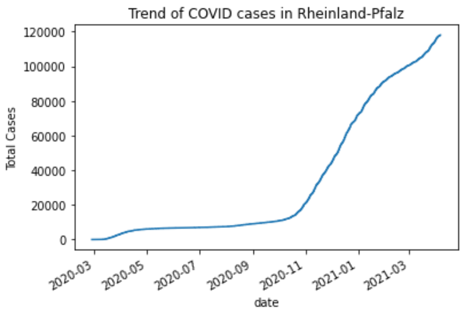


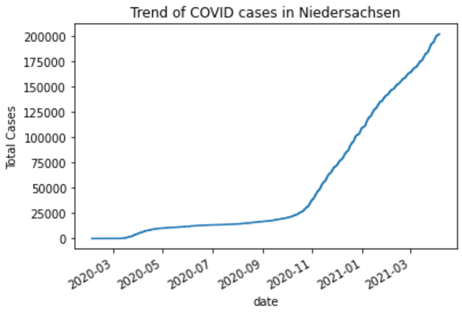
As observed from the above graph, Nordrhein-Westfalen has by far the most Covid cases, followed by Bayern and Baden-Wuerttemberg. The states with the least cases include Bremen, Mecklenburg-Vorpommem, and Saarland.

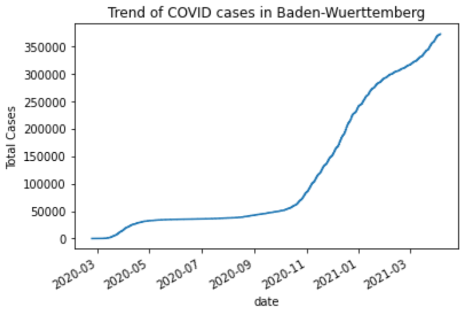
To explain these trends, we can simply observe the total population in each of these states. A lookup of the most and least populous states in Germany reveal that the top three most populous states consist of Nordrhein-Westfaln, Bayern, and Baden-Wuerttemberg, in that order. This reveals the correlation between total population and total confirmed COVID cases in each region. We expect county-level analysis to confirm the same patterns.

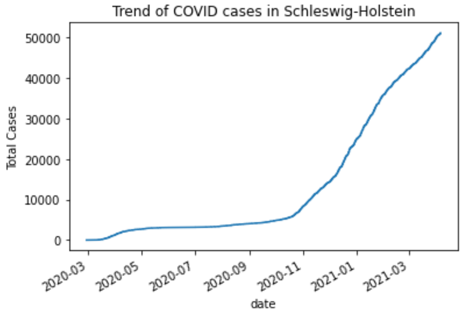
To further build upon these observations, we create confirmed case graphs for each state. These graphs are shown below. We plot simple cumulative sum over time graphs for each state.

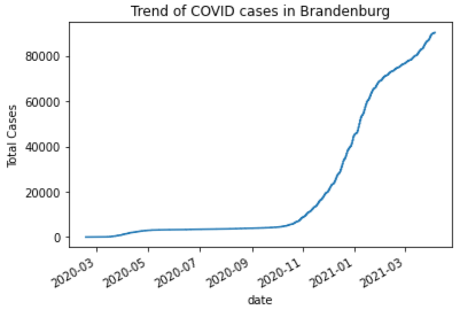


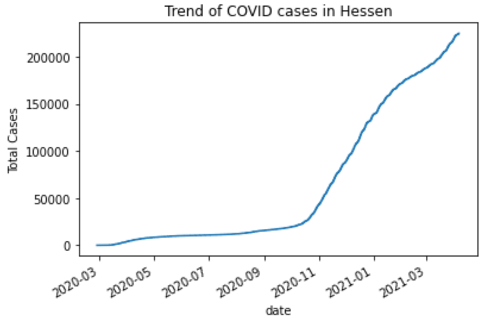


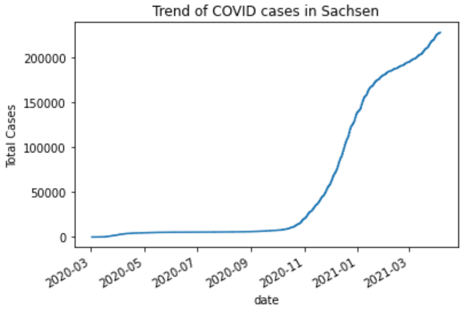


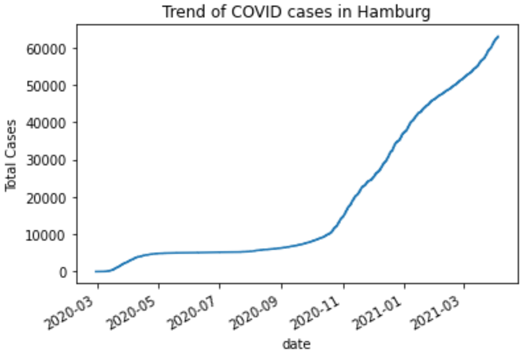




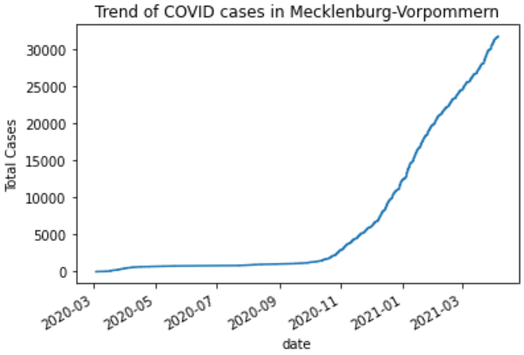




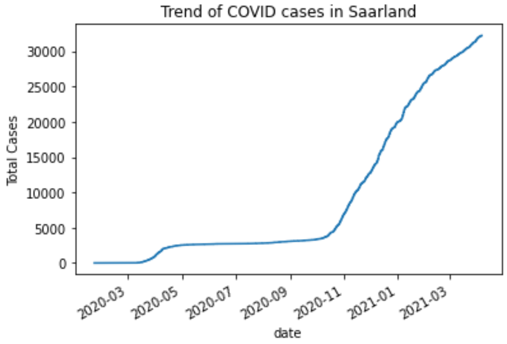




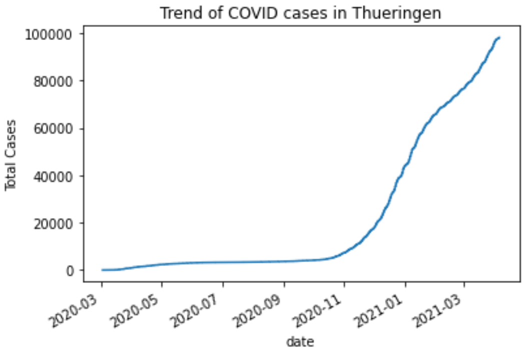
Trend of COVID cases in Hamburg



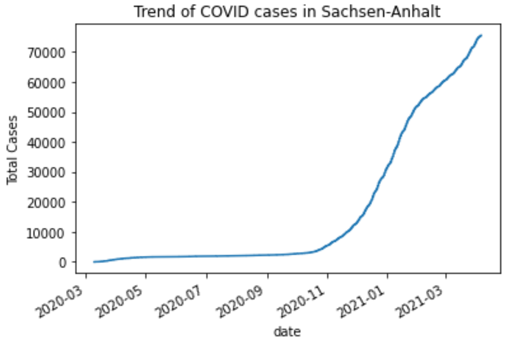
Trend of COVID cases in Mecklenburg-Vorpommern



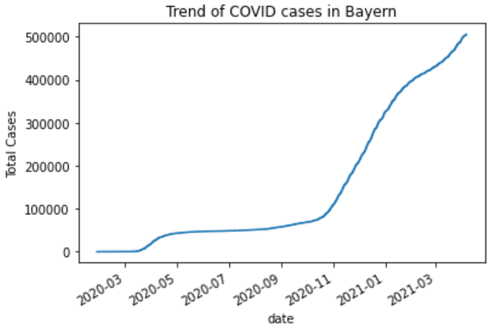
Trend of COVID cases in Saarland



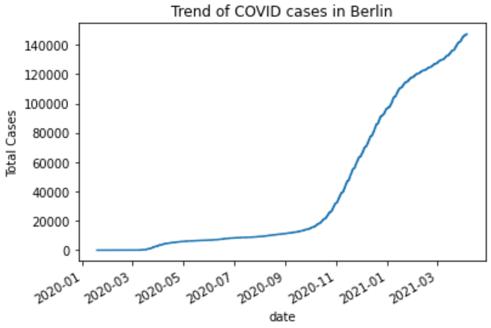
Trend of COVID cases in Thueringen



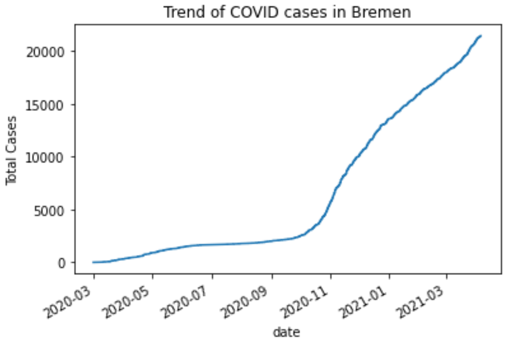
Trend of COVID cases in Sachsen-Anhalt



Trend of COVID cases in Bayern



Trend of COVID cases in Berlin



Trend of COVID cases in Bremen

Evidently, all of the states follow relatively similar trends in terms of COVID cases over time. These graphs are comparable to the country-wide graph, as the lines are very similar.

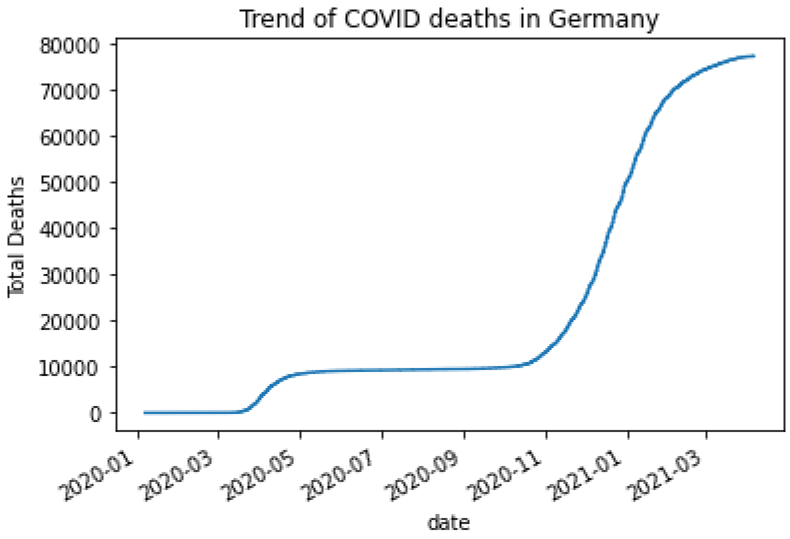
Next, we would like to replicate this analysis for the county level. However, there are simply too many counties to consider all of them in the same way as we did with states. Therefore, we want to study trends in a tabular format and extract information from the top and bottom "performing" counties. To glean more useful county-level information, we can check out the top 5 and bottom 5 counties in regards to number of cases. We achieve this by grouping our merged dataset by county and observing the sum of cases in each county. We observe the results of this analysis in the following table:

county	
SK Hamburg	63007
SK Muenchen	60024
SK Koeln	40244
Region Hannover	38290
SK Frankfurt am Main	30926
...	
LK Ploen	1040
LK Wittmund	980
SK Emden	755
LK Luechow-Dannenberg	588
SK Zweibruecken	545

From this table, we can gather that the 5 counties with the most COVID cases are SK Hamburg, SK Muenchen, SK Koeln, Region Hannover, and SK Frankfurt am Main. The 5 counties with the least cases include LK Ploen, LK Wittmund, SK Emden, LK Luechow-Dannenberg, and finally SK Zweibruecken.

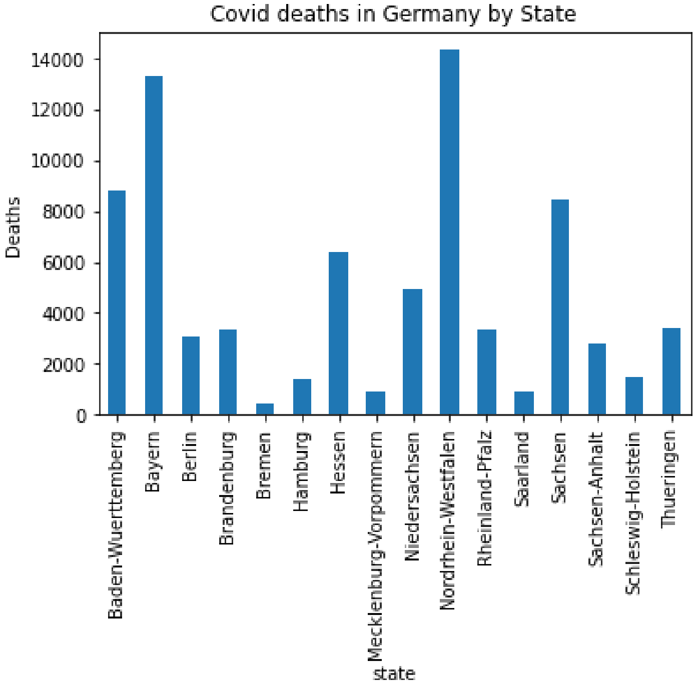
Looking up county-level population data reveals that Berlin, Hamburg, Muenchen, Koeln, and Frankfurt am Main make up the top 5 most populated counties in Germany. Therefore, 4 out of 5 of the counties with the most COVID cases are also the most populated. Our original hypothesis that population is positively correlated with total cases holds, to a reasonable extent.

Next, we wish to repeat this state to county-level analysis, but with COVID deaths in lieu of COVID cases. Below is the country-wide cumulative graph of total COVID deaths in Germany.

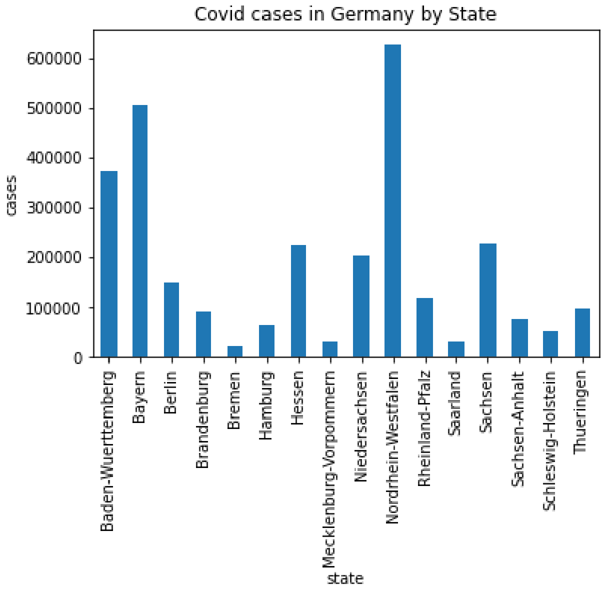


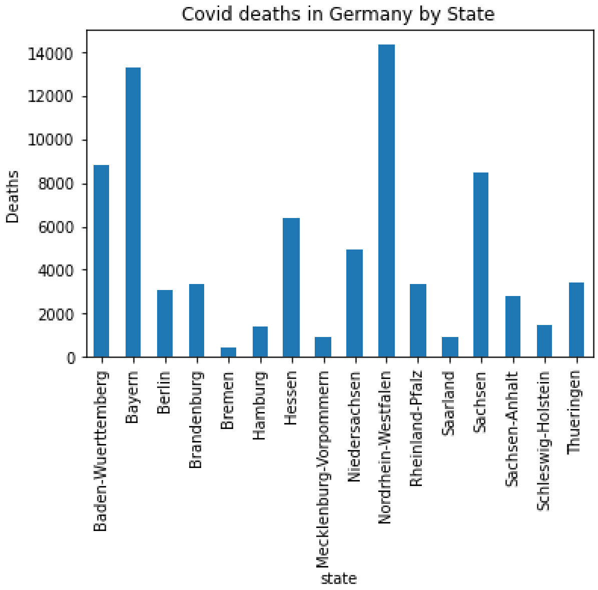
The x-axis shows us the progression of time, from January of 2020 up until early 2021. The Y-axis shows us the cumulative sum of total COVID-19 deaths. Similarly to cases, COVID deaths began to increase drastically in late October to early November of 2020, and continues to rise at the same rate up until early 2021, directly coinciding with the rise in total cases.

We now examine Covid death trends by state level.

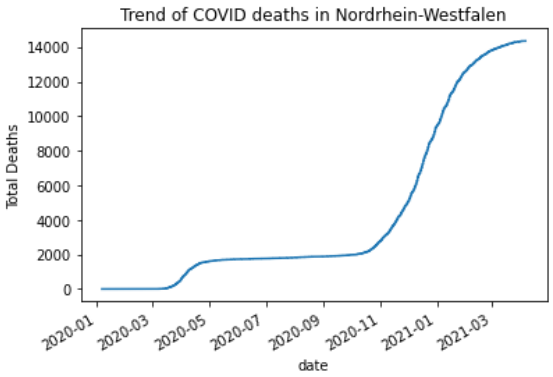


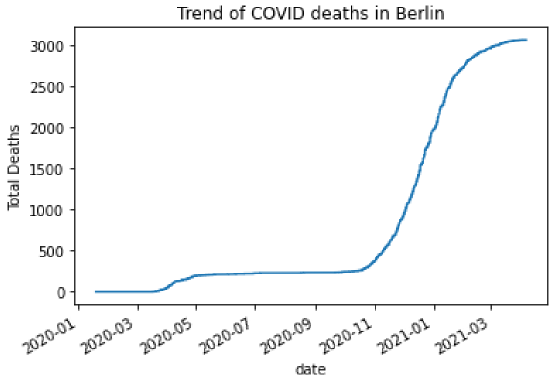
As shown in the above graph, Nordrhein-Westfalen, Bayern, Baden-Wuettermberg, Sachsen, and Hessen form the top 5 states in terms of total COVID deaths. The bottom 5 states include Bremen, Saarland, Mecklenburg-Vorpommern, Hamburg, and Schleswig-Holstein. By observing this graph and the state-level cases bar graph side by side, we can see the direct correlation between deaths and cases on the state level.

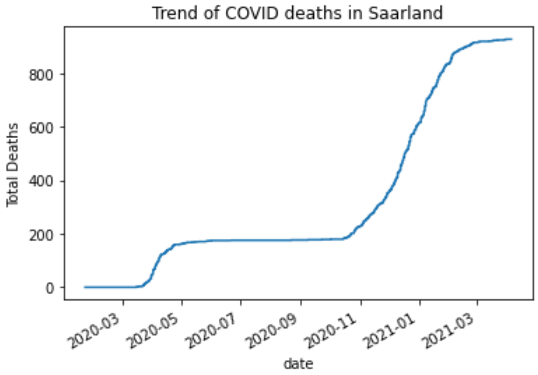


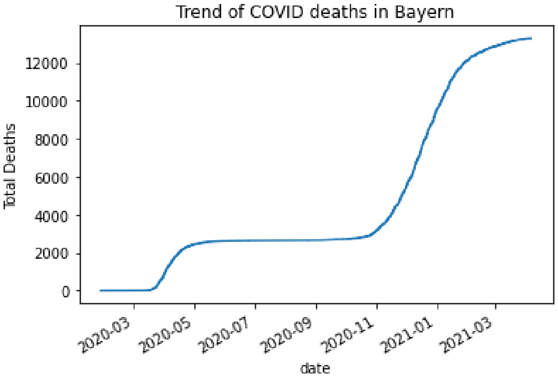


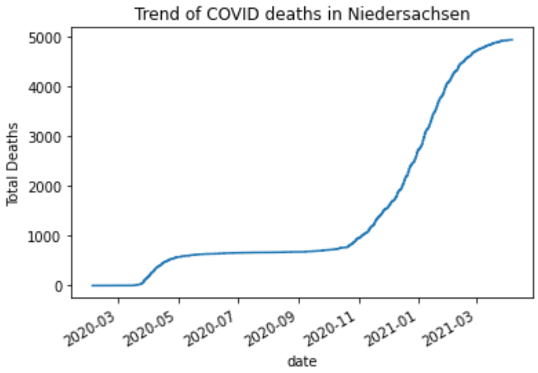
Next, we plot the trend of COVID deaths over time for each state. The cumulative graphs for each state are shown below. Observe how each curve follows a similar shape to the state-wide curve.
We plot a simple cumulative sum over time graph for each state.

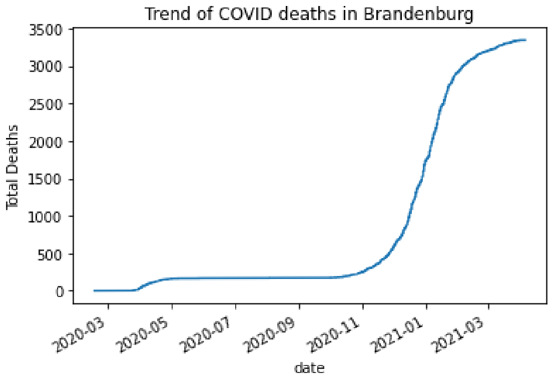


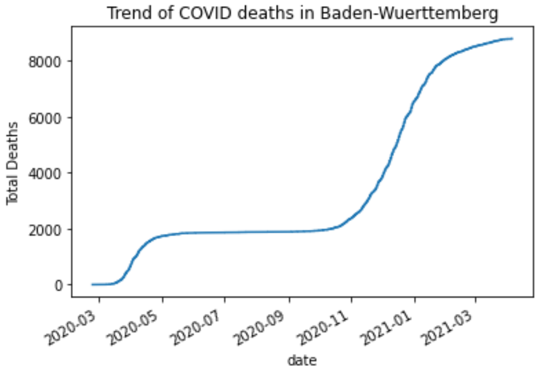


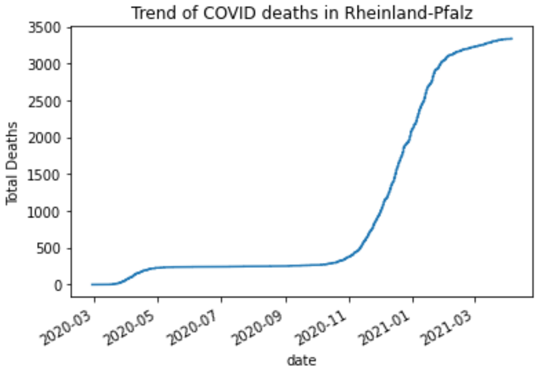


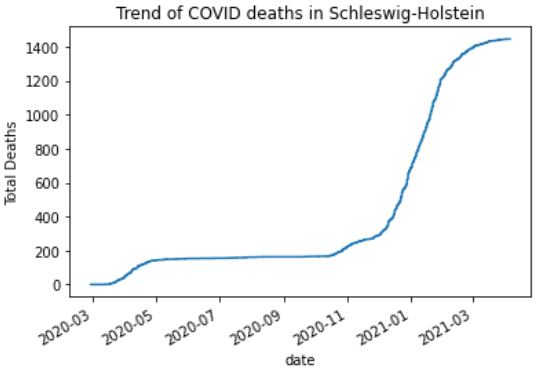


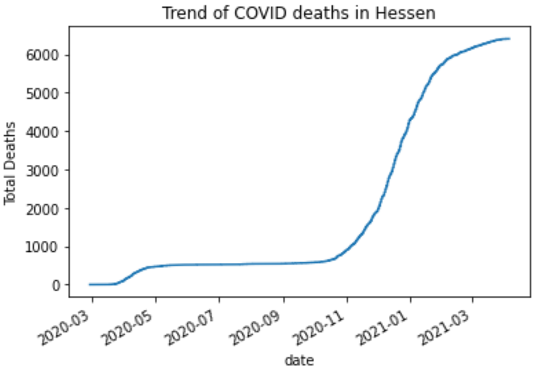


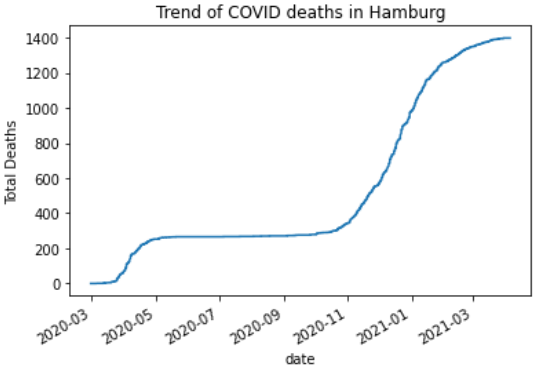


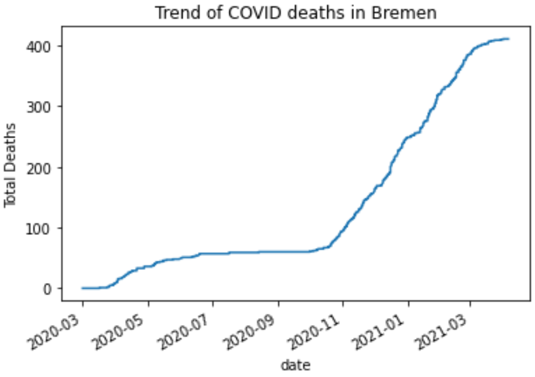


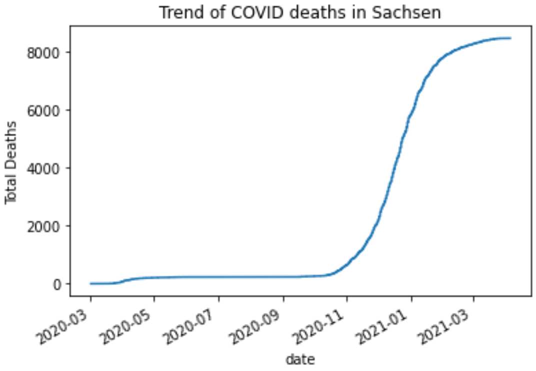


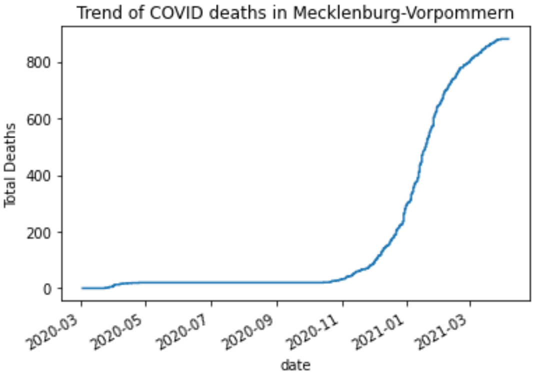

Trend of COVID deaths in Schleswig-Holstein

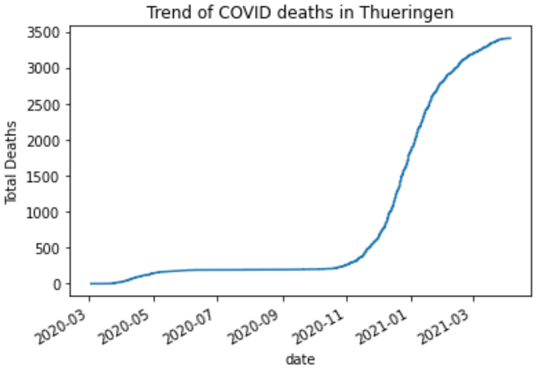

Trend of COVID deaths in Hessen

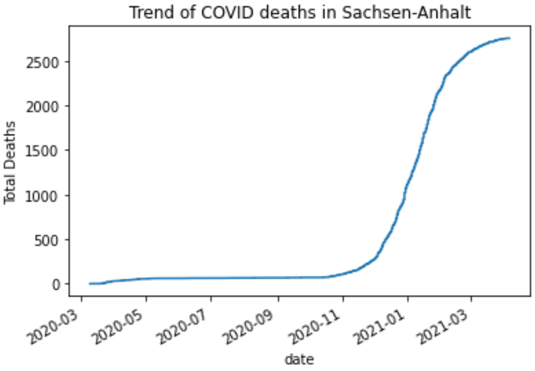

Trend of COVID deaths in Hamburg


Trend of COVID deaths in Bremen


Trend of COVID deaths in Sachsen


Trend of COVID deaths in Mecklenburg-Vorpommern


Trend of COVID deaths in Thueringen


Trend of COVID deaths in Sachsen-Anhalt

The trend of COVID deaths on the state level follows the same curve as the country's death trends. November of 2020 represents an inflection point for an increase in COVID deaths and cases. However, we can also observe a 'tapering off' point for cumulative number of deaths/cases on the state level that occurs around Februrary to March of 2021. This can be attributed to the increase in global vaccination efforts as well as adherence to strict isolation measures implemented around the world.

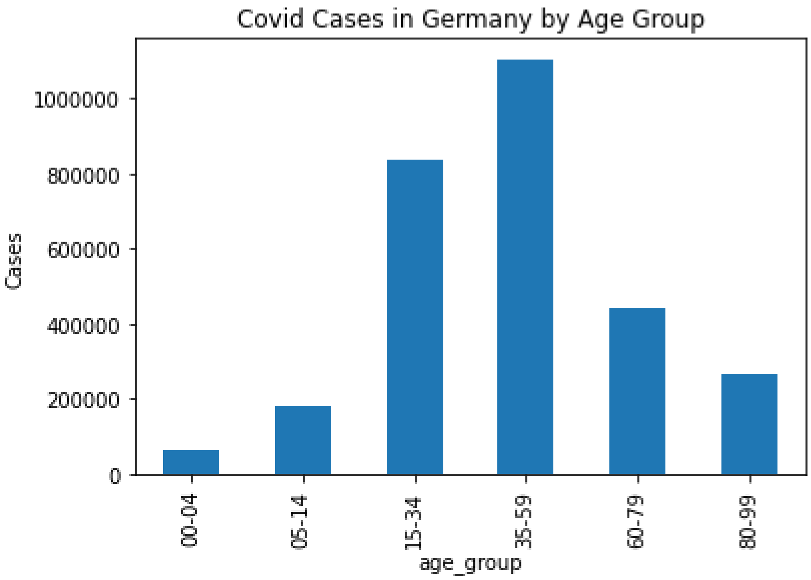
With that being stated, we conclude our state-wide analysis and now want to observe which counties have the highest and lowest death rates. Similarly to the COVID case analysis on the county-level, we cannot observe each county's trends as there are simply too many counties to compile in a readable manner. Instead, we seek to observe the top and bottom 5 counties pertaining to COVID deaths. To do this, we group our merged dataset by county and take the sum of deaths for each county, and sort by top and bottom 5. Below is the table that contains this information.

SK Hamburg	1398
SK Muenchen	1126
LK Goerlitz	1008
SK Dresden	992
Region Hannover	889
...	
SK Neumuenster	22
SK Memmingen	21
SK Amberg	20
SK Emden	7
SK Zweibruecken	4

The counties with the most deaths include SK Hamburg, SK Muenchen, LK Goerlitz, SK Dresden, and Region Hannover. SK Neumuenster, SK Memmingen, SK Amberg, SK Emden, SK Zweibruecken make up the counties with the least COVID deaths. Again, we observe the total populations of these counties and our previous hypothesis holds.

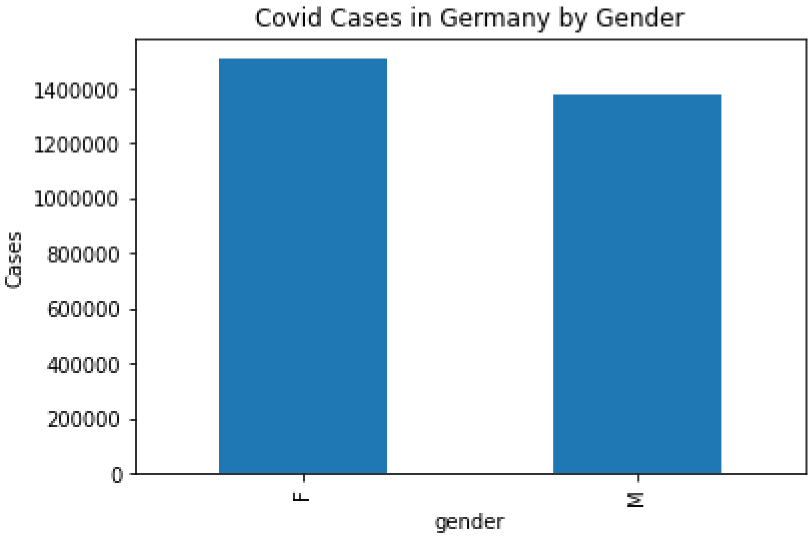
For the final component of our baseline analysis, we want to observe COVID-19 trends based on gender and age group.

We start by plotting the total number of cases for each age group to examine the top and bottom age groups in terms of COVID cases in Germany.



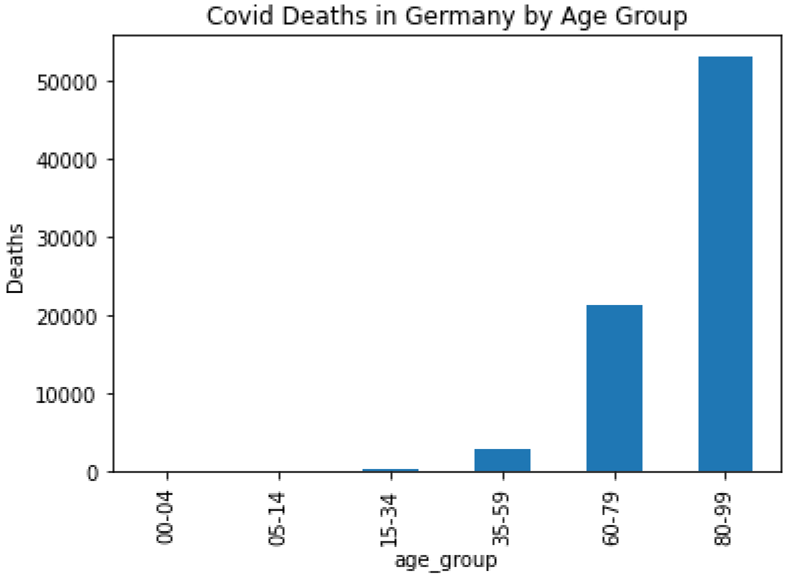
As indicated by the bar graph above, people aged 15-34 through 35-59 comprised the most cases out of any age group. Small children aged 0-4 were the least likely to be affected. These statistics align with the distribution of COVID cases in different age groups in the United States[3], indicating a global trend. Global age distribution is normal. Observe how this distribution is similar to a normal distribution, with the exception of a slight right skew which is attributed to the weaker immune systems of older populations, thereby making them more susceptible to contracting COVID.

Next, we observe gender trends in COVID cases.



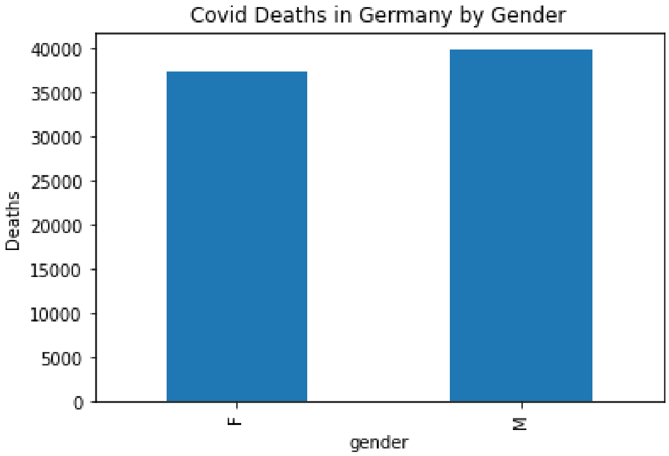
The distribution of cases between women and men were relatively equal. However, females still comprised more of the cases than men.

Next, we examine the trend of COVID deaths by age group.



The graph above shows a clear right skew towards older age groups when concerning COVID deaths. The age group with the highest death rate is ages 80-99. This makes sense - the older population is more susceptible to underlying conditions and weaker immune systems, which may lower their chances of fighting off COVID.

Next, we graph the relationship between gender and COVID deaths.

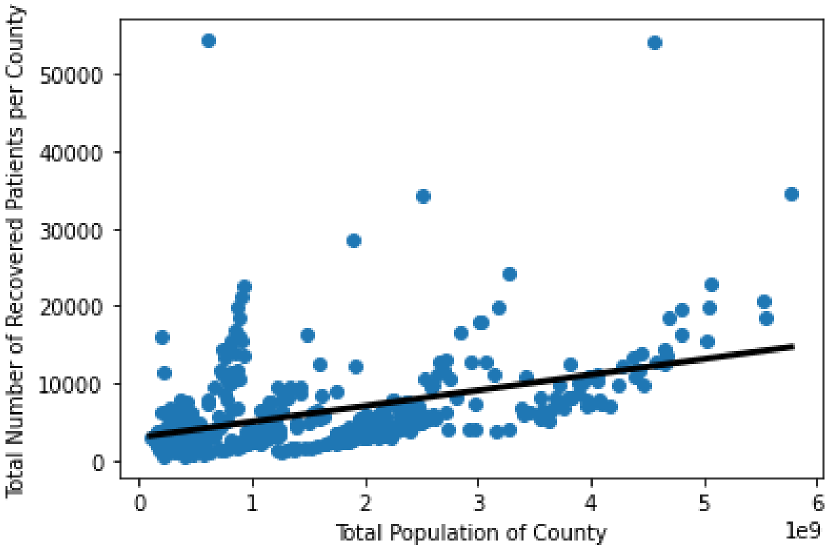


An interesting piece of information to glean from this graph is this: Even though females made up more of the total cases in Germany than males, more men died from COVID than did women.

This concludes our baseline analysis of the COVID-19 data based on different demographics. Next, we delve into machine learning algorithms in an attempt to correctly predict whether or not a COVID patient will die, given their demographics.

III. TRAINING / TESTING MACHINE LEARNING ALGORITHMS

In order to determine if certain attributes would allow us to predict the survivability of individuals with COVID-19 using our dataset, we decided to create three different models, a linear regression model, a logistic regression model, and a K-means model. For our linear regression model, we decided to filter the original data into just the total population by county and the total number of recovered cases per county in order to determine if the two were highly correlated so that we can determine if other factors may be at play. We ended up with a graph as follows with our linear regression in black.



This model has an R^2 value of about 0.18, which is not that great and can likely be explained by the numerous visible outliers which are likely skewing our model. However, given this low R^2 value, this may just also indicate that this model is a poor choice for our data and as such we went further to create a logistic regression model. For this model we decided to also add in gender as a classification variable and converted M to 1 and F to 0 so that we can process its values in a quantitative way. In this instance, our dependent variable is an array of values where each row is equal to 1 if the number of recoveries is equal to the number of cases and 0 otherwise. Upon performing a logistic regression fit on this data, we got an F1 score of 0.91 and an accuracy of 0.84, which is very good and way better than our linear regression model. Our third and final model was a k-means using the same x and y as our logistic regression model and here we got fairly low accuracy but high precision, indicating that there is likely something that can be fixed with the model as a whole, possibly by altering classification variables. When seeking how to improve our model, we also split the data further to create a cross validation set for our kmeans, but this still yielded the same accuracy.

REFERENCES

[1]
https://www.statista.com/statistics/1127686/population-by-federal-state-germany/#:~:text=German%20population%20as%20of%202020%2C%20by%20federal%20state&text=The%20most%20populated%20federal%20state,population%20of%20almost%2018%20million.

[2]
https://worldpopulationreview.com/countries/cities/germany

[3]
https://www.statista.com/statistics/1254271/us-total-number-of-covid-cases-by-age-group/