CDS-403-001 Term Project

Start working on your TERM projects. Start by creating a goal for research (what question do you want to answer), do a literature review (who has worked on something similar and what were their results), and find a suitable dataset. Create a Github repostory for your project.

The midterm project is basically your term project, halfway through.

Submit the following:

1. Technical paper draft (in Word or LaTeX):
   1. Introduction - state your goal and hypothesis, why it is original and why nobody solved it before (5p)

***-CDC Background:***

The Centers for Disease Control and Prevention (CDC) is the national public health agency of the United States. It is a United States federal agency, under the Department of Health and Human Services and is headquartered in Atlanta, Georgia. The agency's main goal is the protection of public health and safety through the control and prevention of disease, injury, and disability in the US and worldwide. The CDC focuses national attention on developing and applying disease control and prevention. It especially focuses its attention on infectious disease, food borne pathogens, environmental health, occupational safety and health, health promotion, injury prevention and educational activities designed to improve the health of United States citizens. The CDC also conducts research and provides information on non-infectious diseases, such as obesity and diabetes, and is a founding member of the International Association of National Public Health Institutes.

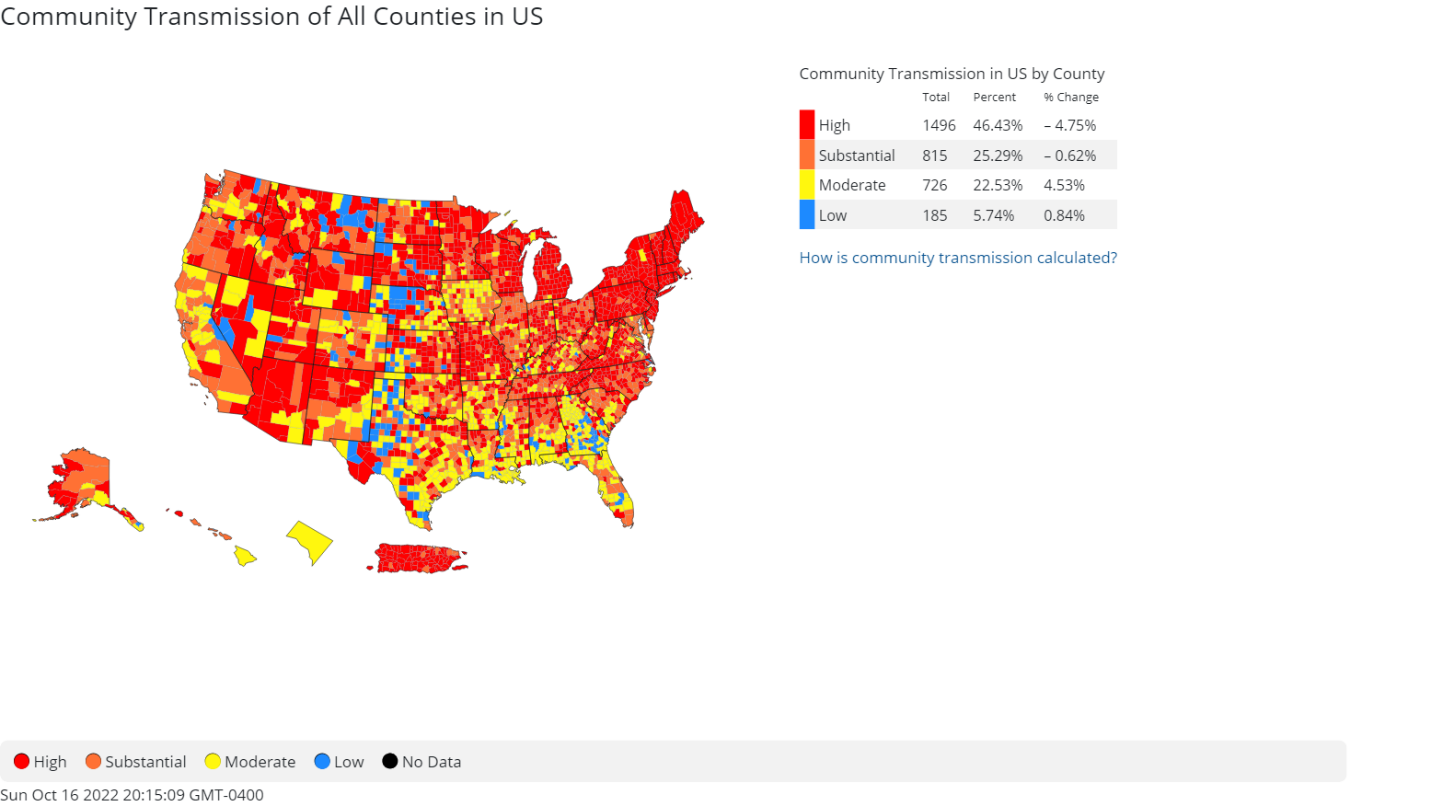
***-Introduction/Goal/Hypothesis/Why It’s Important***

With Covid-19 still being a relatively large problem in the United States despite the severity going down and large segments of the US Population already being partially vaccinated, most of us are taking a big risk due to the fatigue factor in how lightly we take the pandemic and the potential uprise of another wave of the pandemic. The Community Transmission rates for every county in the USA shows a concentration of cases in areas where the vaccination rate is low. The goal of this project is to build on current studies of where cases of Covid-19, and through the methods of machine learning extend the reach of our knowledge on why some places get hit harder than others. One aspect of Covid that was never really talked about after its original outbreak was the effect of temperature on the virus, this is why one of the main goals of this project is to see if on a county level in the United States, does weather have an effect on covid spread? I hypothesis if the initial observation of the virus was correct of it being able to spread faster in certain environments, weather should have a large effect on its spread. The idea of being on the county level scale is to increase the sample size and eliminates potential outliers in the US which tend to be denser than others. The problem with most studies similar to this are they look at the largely populated areas such as NYC and California, which makes sense considering they would need the most resources, however many smaller areas around the US then later lack the resources to treat their patients since they have very high cases as well, but nobody ends up seeing them. The addition of including the weathers impact on top of the current known covid rates and vaccinations, in addition to a county level scale for the project is what makes this an original problem that has a dire need of being solved.

* 1. Literature review - describe similar work that has been done, with references, and where that work failed short of answering your question (5p)

The base data containing the covid rates per county and vaccinations per county that I will be using (linked and cited below in section 1.3) have been used extensively by many organizations and academic institutes to plan for the reentry of workforce back to work place. In a sense the base data itself has done quite a bit to help us get this far into the pandemic and partially subdue it. There was a report on one news channel about how the Internal Revenue Services (IRS) is using this data to plan the re-integration of federal employees back to office buildings in a phased approach where the transmission levels have fallen below a threshold. The problem with this data as it is however, is that it is only showing the percentages of covid throughout the US, but no clear explanation of future indications of where it could possibly further spread in the future. I would like to explore this data and use the data to model future transmissions and positivity rates by meshing the data with the average weather patterns within those counties over the years and then provide prescriptive methods to tackle those situations such as adding more hospital beds, medical personnel and preventive assets like masks and sanitizers to the anticipated locations in advance. At the moment most studies done only show big red areas on a map (usually very populated and dense), and that is where all of our resources go, when in reality we should not only look at a more specific level of geographical locations, but also further extend on finding predictive measures that cause spikes in the pandemic to spread other than human contact to find better preventative measures aside from social distancing.

* 1. Add a link to the dataset/s you plan to work with (must be open source); if the dataset is not very big, you are also welcome to submit it here on Blackboard or on your GitHub repository (2p).



<https://data.cdc.gov/Public-Health-Surveillance/United-States-COVID-19-Community-Levels-by-County/3nnm-4jni> (This will be the main/base dataset which will be used)

Anon. n.d. “United States Covid-19 Community Levels by County.” *Centers for Disease Control and Prevention*. Retrieved October 17, 2022 (https://data.cdc.gov/Public-Health-Surveillance/United-States-COVID-19-Community-Levels-by-County/3nnm-4jni).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

<https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh> (This dataset contains the amount of people and percentages that have received or haven’t received their vaccines yet)

Anon. n.d. “Covid-19 Vaccinations in the United States,County.” *Centers for Disease Control and Prevention*. Retrieved October 17, 2022 (https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

<https://www.weather.gov/dtx/fnttemp1920-1940> (This is a database containing all of the weather collected for the past decades in the US)

US Department of Commerce, N. OA A. 2015. “Temperature Breakdown by Decade.” *National Weather Service*. Retrieved October 17, 2022 (https://www.weather.gov/dtx/fnttemp1920-1940).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Code of the project should include:
   1. first exploration of the data (summary statistics, number of variables, which are the variables of interest, which you will discard, etc.) (3p)

The first dataset which will be the starting point as stated above will the of the Covid-19 levels in the United States by County. There are a total of 12 variables in the data, however the ones I will use are; County, County\_fips (Zip code), State, county\_population, covid\_inpatient\_bed\_utiliztion, covid-19\_community\_level, and date\_updated. The other variables regard cases and admissions in numerical increments which would not correlate to my new data once fully completed with the integration of weather data. The second dataset is very large containing 72 variables, at the moment the only variables of interest are date, FIPS (zip code), Administered\_Dose1\_Recip, Completeness\_pct for those who have received their full vaccines, and Booster\_Doses. These variables will be merged with the original dataset to potentially explain any indications of any variances in covid rates are being affected by the covid vaccination percentages themselves. The weather database will be used once the preliminary results of which counties tend to get the highest covid rates.

* 1. the methods you are using and why you are choosing these methods for the goal you stated. (3p)

The current method of accomplishing the goals of this project is by using 2 main machine learning algorithms. The first is to start with a random forest on the original dataset which will include the base data, as well as the vaccination data merged. The two main features (aka the training data) will consist of the covid rates per county, and the vaccinations per county, and the resulted outcome will be which counties around the US are more likely to get another outbreak of Covid, and which are the least likely. This will then narrow the large number of counties in the US down to a select few which can then be made into a new separate dataset containing those counties as well as the provided weather from the database which will be entered manually. With this then I will be able to apply a Lazy Learning (KNN) algorithm on this dataset, so when future weather is predicted in certain counties, the covid rate within that specific temperature range can be predicted and possibly show if there is any indication of possible preparations of another spread of the virus, which as stated, is the initial goal of this project, to potentially find a correlation between covid rates, vaccination rates, and weather, and then use machine learning to predict future covid rate values with previous vaccination and weather results to provide for those counties which may not receive the help and resources they should.

* 1. discuss the pros and cons for the methodology (2p).

The benefit of this methodology to first narrow the counties in the US to the ones that are the most affected by covid rates and the lowest rates right now is definitely a pro in that it puts much less strain on the algorithms to later predict whether certain ranges of temperatures will increase or decrease the current percentages of covid rates in the counties, however, the con of this methodology revolves around the fact of the ideology itself. Every experiment and every project have to branch from somewhere, and in cases such as these, trying to isolate the weather variable and try to eliminate other factors which have an impact on covid rates is a difficult task, in that there are many aspects that could be attributing to covid rates around the US aside from temperature. The pro side to this however is the major aspects of location and vaccinations within these counties are being limited as much as possible within the datasets themselves which will be merged and created manually, improving the results of the machine learning algorithms, rather than leaving the datasets as they are and trying to assume weather has a large enough influence to be visible through all the other potential variables. That being said, all that’s left to do is merging the datasets properly and the algorithms should have no problems running.

Make sure I have access to any links you are submitting. FYI, I do not have a Kaggle account and don't plan on creating one, therefore if you are working with a dataset from there make sure it can be accessed as open source. Code can be either in R or Python. Make sure you are not copying someone else's project, that your project is not trivial and that you are citing resources (goal, code, dataset).