**Introduction:**

The reality of Covid-19 in today's world is a scary thought. It is everywhere in the world and continues to spread. Therefore, it is important to examine the data around Covid-19 numbers to help us answer tough questions around how this is spreading, and how to slow the spread to help save the lives of some of those most at risk.

Techniques in data science and predictive analytics can help us understand what data we currently have, and how we can use that to inform the future. In this project we will look at data from 87 counties (and 1 "unknown location") in Minnesota (MN) that show Covid-19 rates in the county, and in the nursing facilities in those counties.

**Background:**

Our project will analyze Covid case counts in Minnesota by county to try to identify factors that may contribute to higher rates of cases. We will overlay total Covid case counts by county with multiple potential contributing factors, such as population density, number of long-term care facilities in the county, and whether the county is part of a major metropolitan area. We will do some basic analysis to determine if relationships are present. If relationships are found, we will also analyze the nature of the relationships. Are they linear in nature or do they follow a different pattern?

*We are interested in looking at data to see if any geospatial factors can be identified that correlate with higher infection rates of Covid or increased risk of death from covid. This may indicate if spread and death from the virus is linked to specific behaviors in the community, specific types of institutions within the community, or total density of population.*

**Results:**

This is where we write the results of our question above in yellow and italics.

**Data Overview:**

Our data is Covid-19 case data for the state of Minnesota broken out by individual counties (87 Counties + 1 row for cases for people of unknown counties). For each county we have data points for: “County”, “Total Positive Cases”, “Total Deaths”, “Population, Rate of Infection”, “Death Rate”, “Pop Density”, “Mean Household Size”, “Congregate Living Population”, “No. of Cases in Nursing Homes”, “No. of Nursing Homes in County”, “Nursing Home Cases, % of Total”, and “Cases per number of Nursing Homes”. With our data being mostly numerical we ran linear regression to create an algorithm/equation for predicting Covid-19 numbers across: "Pop Density", "Mean Household Size", "Nursing Home Cases, % of Total". Our goal in working with this data would be to identify any trends in infection rate, or death rate, related either to number of care facilities in a county, the mean household size, or the population density of the county.

**Data Sources**

Our group project looked at Covid case data for at the county level in Minnesota to study if there were any links between the rate of Covid cases in each county and various geospatial factors may be identified which correlate with higher rates of infection or death. We collected data which contained many data points for each county, such as the population, mean household size, number of long-term care facilities in each county, number of cases attributed to long-term care facilities in each county, and total population in congregate living in each county.

The sources for our data were primarily the Minnesota Department of Health and Minnesota State Demographic Center. From the Minnesota Department of Health, we were able to obtain the case counts and total death counts per county (1). We were also able to obtain the number of cases attributed to each individual nursing home facilities in the state (2). This data was aggregated by county to obtain a count of Covid cases attributable to nursing home facilities, as well as a count of long-term care facilities in each county. From the Minnesota State Demographic Center, we were able to obtain the population of each county (3), which was then compiled with the case data to obtain infection rates and death rates for each county. Population density was the one piece of data which we were not able to obtain in a downloadable format, and so this data was added manually in excel using values obtained from usa.com (4).

**Analysis:**

When looking at our model statistics, we will be focusing on the "Score", "MAE", "RSME", and "r2". These statistics represent "The explained variance score", "Mean Absolute Error", "Mean Squared Error", and the "r2". To find the best model, we are looking for the Highest Score and r2, as well as the lowest MAE and RMSE. Both Lasso and Linear Regression have the highest r2 at '.77'. We then look at Score and find Linear Regression is the highest at '.93'. Lastly, we will look at the MAE and RSME to see if there are any striking differences. They are close in value. When looking at all the statistics, there does not to appear to be a substantial difference in models. If we were to select one model, we thought was best (for the data we are using in this example) we would recommend Linear Regression, as it has just a slight edge over the others.

**Conclusion:**

This is where we write up whether our analysis of our original question holds up and predicts our model to be accurate, or not.

**Methods and Observations:**

To help showcase the data the team tried three different Python libraries to select the optimal solution to present the Covid-19 data. Of the three, Matplotlib and Seaborn were chosen to be the best fit. Plotly is not desirable due to its inconvenience of only working in Jupyter notebooks, (and some team members have even had issues getting Plotly to work in notebooks). Plotly (when working) was shown to have its uses, with it being the library best suited for interactive displays. However, we feel that the data in this project may not need to be interactive.

Matplotlib was selected for showcasing most of our data that needs flexibility and creativity. Seaborn was selected to showcase select data that needs to look professional and efficient, while being able to be produced relatively quickly.

The plots that are most useful for explaining our data within our selected libraries of Matplotlib and Seaborn are scatter type plots and line plots. These will show overall trends (or lack thereof) when evaluating one variable based off another. Bar Graph, Histograms, and PairPlots may be useful as well due to our data being more regressive than categorical.

To begin analysis of the features of our data, we first created a correlation matrix in pandas and plotted it to a Seaborn heat map to give a quick visual of correlations within our data.

Analysis of the correlation matrix indicates we may have high multicollinearity between our variables. Population, Total Positive Cases, Total Deaths and No. of Cases in Nursing Homes have an almost perfect relationship, suggesting that they are essentially measuring the same thing. Categories which may be interesting to analyze are the Infection Rate, Death Rate, and total number of deaths. Since Population, Total Deaths, Total Positive Cases and No. of Cases in Nursing Homes all have correlations greater than 98%. Feature analysis will include Total Deaths as a target of interest but will not use the other highly correlated variables as explaining variables.

The initial question for our project was to see if we could identify any geospatial or community level characteristics that would correlate highly with increased spread of Covid or increased risk of death from covid. To begin our analysis, we decided to investigate Rate of Infection, Total Deaths, and Death Rate as potential target variables. Again, we wanted to avoid highly correlated variables that appeared to have extreme multicollinearity. For consistency, we will evaluate each of our target variables of interest against the same explainer variables - 'Population density', 'Mean Household Size', and 'Nursing Home Cases, % of Total'.

To find which algorithm performs best for our data set we first need to create a standard basis on which to compare models. We chose one Target variable "Total Deaths" and three explaining variables: "Pop Density", "Mean Household Size", "Nursing Home Cases, % of Total". This will serve as our basis for analyzing our Linear Regression, Ridge, ElasticNet, and Lasso models.

**References:**

1. Cases, deaths, and no longer needing isolation by County of Residence (County Cases Raw) <https://www.health.state.mn.us/diseases/coronavirus/stats/index.html#wtrmap1>
2. Congregate Care Facilities with Exposures, by County (FacilityDataRaw)

<https://www.health.state.mn.us/diseases/coronavirus/situation.html#ccf1>

1. County Population Data (Countypopulationraw)

<https://mn.gov/admin/demography/data-by-topic/population-data/our-estimates/>

1. Population Density by County

<http://www.usa.com/rank/minnesota-state--population-density--countyrank.htm?hl=&hlst=&wist=&yr=&dis=&sb=DESC&plow=&phigh=&ps=>

Abstract

Background

Results

Discussion

Conclusions

Methods