**Project Outline**

**Name –** Temporal and Spatial Trends in US Covid-19 Cases and Deaths

**Introduction**

Covid-19 is novel, highly contagious, acute respiratory virus that was first identified in December 2019 in Wuhan, China. Over the course of the following 14 months, this virus spread rapidly to every corner of the globe, becoming one of the deadliest pandemics in recorded history. In the United States, the first confirmed Covid-19 case was identified in January 2020 and by mid-March there were confirmed cases in every single state and North American territory. In the midst of this rapid pandemic spread, epidemiologists and modelers struggled to accurately forecast the spatial and temporal trends in cases and deaths. However, with regularly updated, publicly-available covid tracking data, a sufficient amount of data now exists to retroactively examine how cases and deaths evolved over the course of this 14 month period. This study utilizes the New York Times Covid Tracking Data to statistically analyze trends in the timeseries of Covid-19 cases and deaths as well as the spatial development of cases at the state level across the united states using cluster analysis.

**METHODOLOGY**

**Data Sources**

Due to the fragmented nature of the US public health system, there is no centralized governmental data repository that is updated daily with Covid-19 case and death data. Instead, this study obtained data from the New York Times (NYTimes) Covid-19 Tracking Project (<https://github.com/nytimes/covid-19-data>). The NYTimes relies on dozens of reporters across multiple time zones to regularly update this tracking database with new information from press conferences, report releases, and local database dumps. Datasets utilized in this analysis reported the daily cumulative case and death counts in the US aggregated at the national, state and county level (US.csv, US-states.csv, US-counties.csv), respectively. Demographic data on state populations were also obtained from the US census bureau to compare per capita cases rates. The *states.area* dataset that is included in R by default was also utilized to derive state population density data.

**Data Formatting**

All data analysis and visualization for this study was conducted in RStudio. The dataset was filtered to only examine cases and deaths reported from the beginning of March 2020 through the middle of February 2021. Raw data was reported as cumulative cases and deaths through time. To examine daily statistics, a filtering function was applied to calculate the finite difference between each consecutive reporting day.

**Exploratory Data Analysis**

**Time Series Visualization**

An initial exploratory data analysis (EDA) was conducted to both elucidate trends and characteristics of the dataset and to guide the model development process. To better understand the temporal evolution of daily new cases and deaths in the US, a timeseries for both of these parameters was first generated (Figure 1). The timeseries of daily US Covid-19 cases depicts four distinct regimes in the change in daily covid 19 cases throughout the course of the pandemic. From March through the end of May 2020, the number of cases grew logistically; growing exponentially in March before asymptoting at a maximum daily new case load of 25,000-30,000 individuals through April and May. Similar but larger magnitude growth trends are evident in June through August, asymptoting around 65,000 daily new cases, and October through December 2020, asymptoting around 250,000 daily new cases. Note that the sharp drop in cases around the end of December is likely a reflection of a decline in reporting around the winter holidays and not a reflection of the actual drop in the real case load. From January 2021 onwards, the case trend differs from the early regimes, with a noticeable linear decline in the reported case numbers through December. Another interesting aspect of this dataset is the seven-day oscillation in the case numbers. New case numbers always tend to be lower on Saturdays and Sundays than during weekdays, reflecting the fact that many labs do not report case numbers on the weekend. The timeseries of daily Covid-19 deaths shows a similar logistic growth rate to the case rate in the early spring 2020. However, the number of deaths, drops significantly from mid-May 2020 and oscillates around 1000 cases a days until November 2020 when the number of daily new deaths rises again, fluctuating around 3000 deaths per days. The seven oscillations, observed in the timeseries of new cases, is even more prominent in the death data, with significant drops in reported deaths during the weekend. This initial visualization and review makes evident that that there are clear similarities and differences in the functional trends between both datasets.

Chart, histogram

Description automatically generated

**Scatter Plot**

To further examine how the relationship between daily new US covid cases and deaths changed through time, a scatter plot of these two variables was generated (Figure 2). Having previously identified four distinct regimes in the case growth rate, these points were colored by the season for each observation. The scatter plot further highlights the different trends in the daily case to death ratios during each of these time periods. In the spring 2020, there is an significantly evident positive correlation in the case to death ratio. The points in this season are all closely clustered with the deaths growing exponentially with cases. In the summer 2020 data, there is no clear positive or negative trend in the correlation between the two parameters with the points scatter roughly in a circle. In the fall season, again a positive correlation is evident, however the case to death ratio is four to five times that observed in the spring. Finally in the winter (December 2020 – February 2021), the ratio of cases to deaths decreases but is still positive.

**Chart, scatter chart

Description automatically generated**

**SPATIAL ANALYSIS**

To examine spatial trends in the spread of Covid-19 across the United States, box plots of the state-level monthly average per capita daily new cases and deaths were visualized (Figure 3). By normalizing the daily state case and death rates by population (i.e. per capita) the virus load between states can be directly evaluated. These visualizations help to determine how new cases and deaths are distributed across the US.

* **Chart, box and whisker chart

  Description automatically generated**

**Chart, scatter chart, box and whisker chart

Description automatically generated**

This data was further subcategorized into regional state groups to examine if there were regional patterns in the Covid-19 case and death rates. The 50 states were separated into four regions that were defined by the default states.region dataframe that is included with R: Northeast, South, North Central (i.e. Midwest), and West. The map in Figure 4 depicts the region that each state was categorized into. Figures 5 and 6 shows comparison of the boxplots for the average per capita case and deaths rates binned by month and aggregated at the regional level. This visualization highlights some key differences in the temporal evolution of Covid-19 in different portions of the US.

**Chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

* New Death rate lagged behind cases by a few weeks
  + Incubation time for virus
  + Takes time for people to get sick, go to hospital, die
* ***Can we develop a model to predict future deaths based on current case numbers?***

**Methodology**

* **Exploratory Data Analysis - Micah**
  + Visualizing timeseries for average daily cases and deaths by state
  + Spatial trends
    - Box plots grouped by states
  + Visuals and statistics for monthly trends in cases by states
    - Box plot of average daily cases per month by state
    - Box plot of average daily deaths per month
  + Animated Map of cases across states
    - Shows when different regions were at their peak
* **Clutering analysis for heriartchy of states**
* **Timeseries lag anayslis**
  + Developing Lag-1 Autogressive model to predict daily deaths as a function of cases
  + What is the lag between the deaths and cases?
  + Is it consistent between states?

**Results**

* EDA
* Clustering Analysis
* Timeseries Analysis
  + How well does AR-1 Model Perform?

**Discussion**

* **Are there relationship between states? - Clustering**
  + What do they have common?
    - Location
    - Poltiics
    - Demographcs
* **Temporal Trends**
  + Does the lag between cases and deaths make sense
  + Can we predict with cases and/or deaths

**Conclusions**

* What are the implications for public policy?

**Appendix –**

* Figures, Tables, Codes