Logo

Description automatically generated

Week 4 Assignment – MASSACHUSETTS COVID-19 PRELIMINARY ANALYSIS  
  
Group 1

Craig Perkins – [perkins.cr@northeastern.edu](mailto:perkins.cr@northeastern.edu)

Sneha Sharma Satheesha - [satheesha.s@northeastern.edu](mailto:satheesha.s@northeastern.edu)

Pranay Bindela – [bindela.p@northeastern.edu](mailto:bindela.p@northeastern.edu)

Northeastern University

December 2022 – ALY 6150

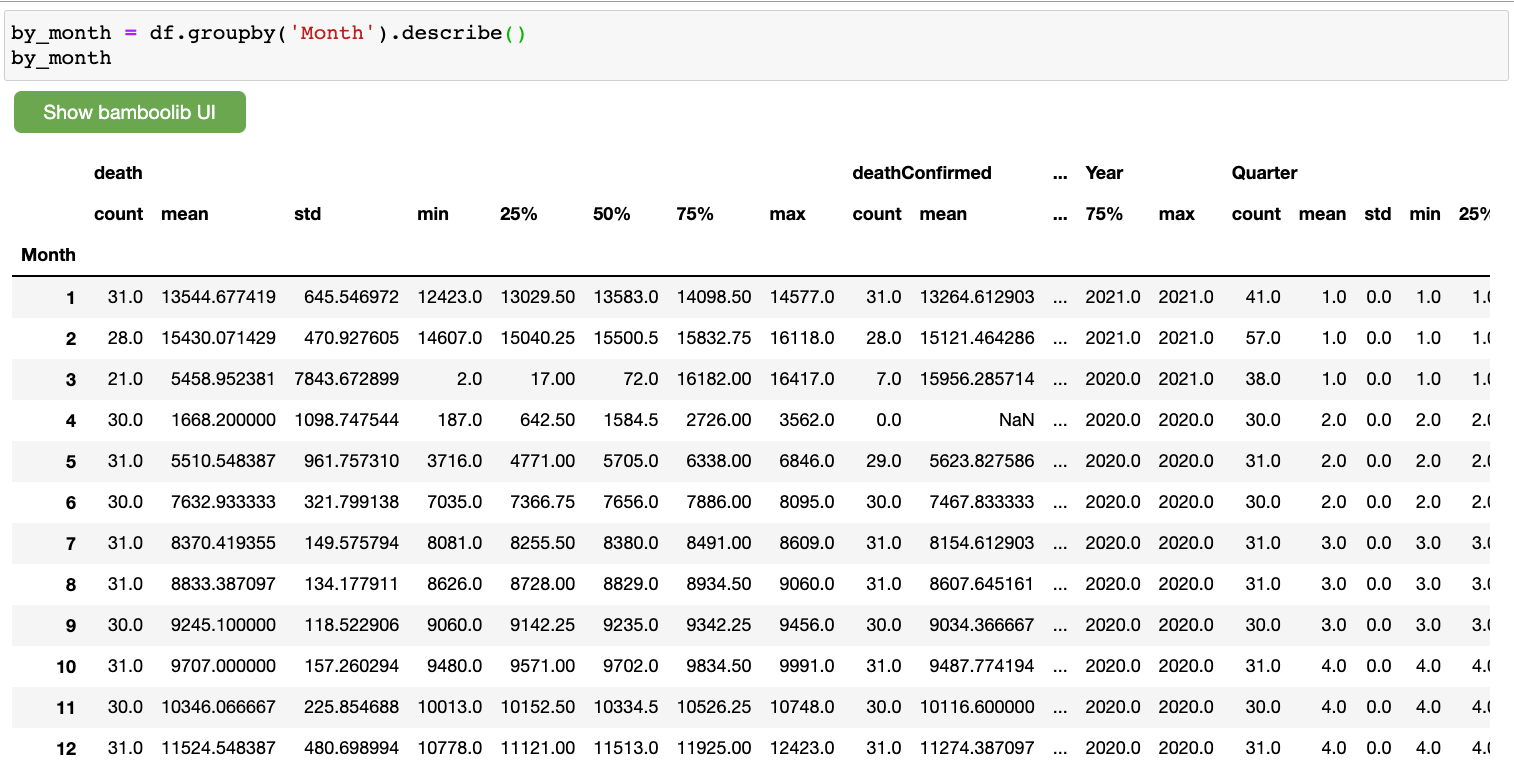
Week 4 Assignment – MASSACHUSETTS COVID-19 PRELIMINARY ANALYSIS  
  
Group 1

For this week’s group assignment, we continued our analysis from week 2 and augmented our dataset with data from mass.gov and the NYTimes. The dataset we chose for week 2’s analysis was the most concise dataset we were able to find that contained features our group was interested in studying, but we also found that the dataset was limiting for subgroup analysis and also lacked recent datapoints that our group is eager to use for analysis. Our group is currently using data from the Covid Tracking Project (<https://covidtracking.com/data/state/massachusetts>) augmented with data from mass.gov (<https://www.mass.gov/info-details/covid-19-response-reporting>) and NYTimes (<https://github.com/nytimes/covid-19-data/blob/master/us-counties.csv>) which provides richer subgroup analysis. The challenge we faced in this analysis is that the mass.gov data is reported in periodic increments and is not collected in a concise time series manner like the dataset from the Covid Tracking Project.

With the predictor’s we built for this week’s analysis, we hope to be able to build tools useful for public health professionals to advise on measures to take to mitigate spread of the virus. If successful, the tool is hyperlocalized and can be used on the county and municipality level in the state of Massachusetts.

**Subgroup Analysis**

Before augmenting the dataset with data from mass.gov and NYTimes, we performed subgroup analysis on the Covid Tracking Project dataset by performing descriptive statistics on the dataset by year, by quarter, by month and by quarters in the first full-year of the pandemic (2021). Below is the output by month. The dataframe is wide since the dataset has a lot of features, so this screenshot focuses on death.



Judging by this output, there is a clear difference in the average amount of deaths between the colder months and warmer months of the year.

The next sub-group analysis performed was by age using data from mass.gov which provides positive testing breakdown by age. The numbers appeared as our group expected with the 20-29 cohort having the highest number of positive tests.

Chart

Description automatically generated with low confidence

The next analysis we performed was using data from NYTimes County data. Our group had to perform filtering on the NYTimes County level data because the original CSV file was 100MB and contained county level data for the whole country.

**Graphical user interface, table

Description automatically generated**

On cursory glance, the case count and death count look to be in line with population counts for the counties. Our group plans to perform a more in-depth analysis to see which counties have disproportionate deaths adjusted for their population

**Conclusion**

From the above graphs, we can conclude the following:

1. The peak covid positive months were January and February 2021
2. The deaths have also occurred in higher numbers during these months.
3. We forecasting model can predict with Root mean square error of 205 at this point in our project.

**References**

1. The COVID Tracking Project – Massachusetts <https://covidtracking.com/data/state/massachusetts>
2. Covid-19 response reporting (Mass.gov) - <https://www.mass.gov/info-details/covid-19-response-reporting>
3. The NYTimes Covid-19 Data - <https://github.com/nytimes/covid-19-data/blob/master/us-counties.csv>