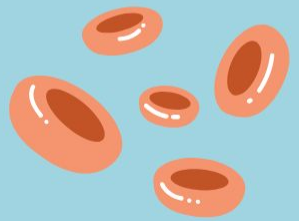
FACTORS THAT CONTRIBUTE TO THE SUSCEPTIBILITY OF INFECTIOUS DISEASE



Group 5
Eddie Almonte
Mark Bishoff
Alanis Perez
Nishi Thapliyal

Which U.S. counties are the most susceptible to infectious disease outbreaks?
What environmental, demographic, & healthcare factors contribute to their vulnerability?

Data Sources

Environmental Factors - Climate (Avg, min, max temps)

- **Source: National Center for Environmental Data**
  - https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/mapping

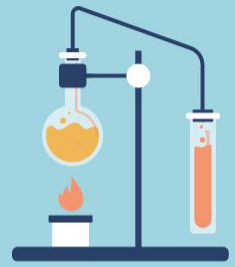
Demographic Factors - Population, Age

- **Source: CENSUS 2020-2023**
  - https://www.census.gov/data/tables/time-series/demo/popest/2020s-counties-detail.html

Healthcare Factors - COVID-19 Vaccination & Death Rates

- **Source: U.S. Department of Health & Human Services, Data.gov**
  - Vaccination Rates: https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh/about_data
  - Death Rates: https://catalog.data.gov/dataset/provisional-covid-19-death-counts-in-the-united-states-by-county

Data Preprocessing



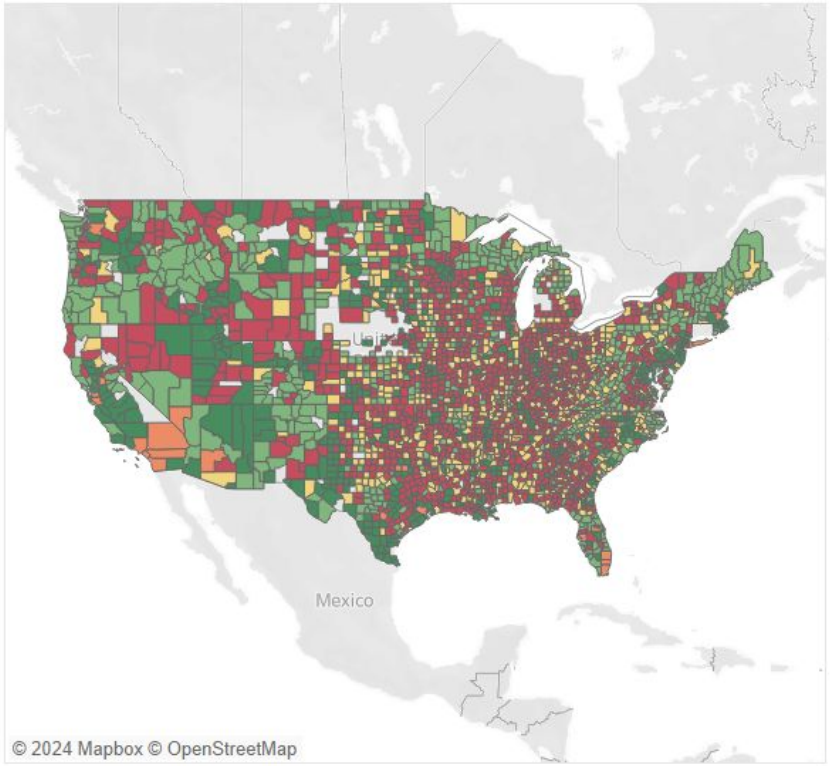
COVID-19 Related Data

- Vaccination rates
  - Administered doses per 100K
- COVID-19 death rates (2020-2023)
  - COVID-19 related deaths per 100K

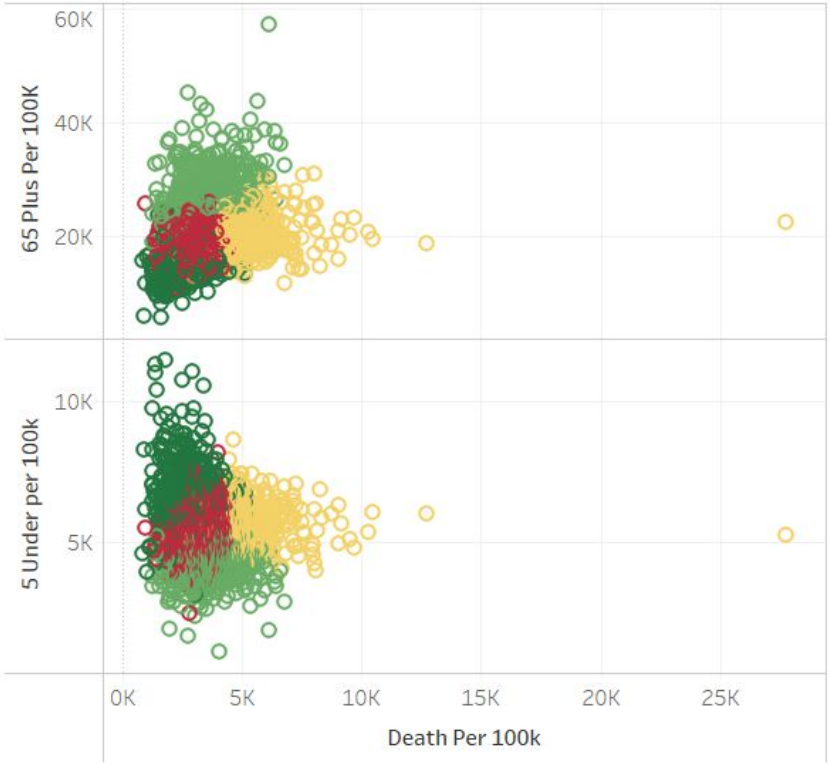
NON-COVID-19 Data

- U.S. Population
  - Seniors per 100k (65+)
  - Children per 100k (5 and Under)
- Temperature (°F)
  - Min yearly temperature
  - Max yearly temperature

Clusters Map



Clusters Scatterplot

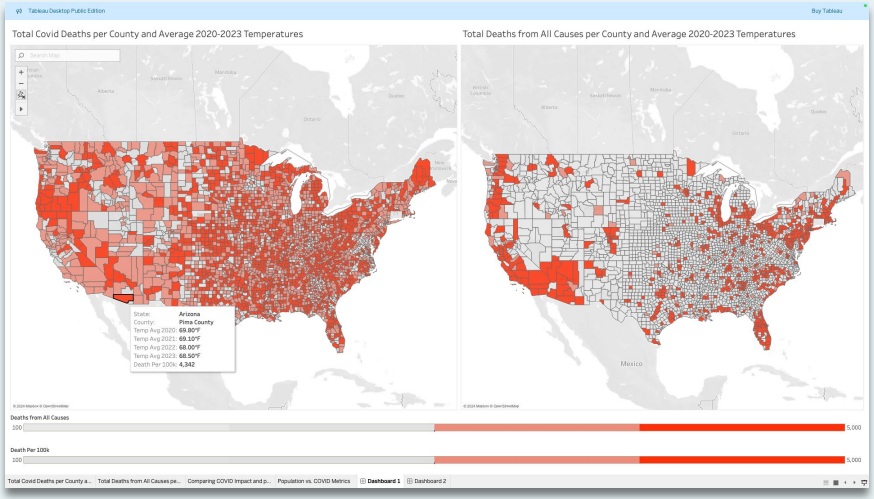


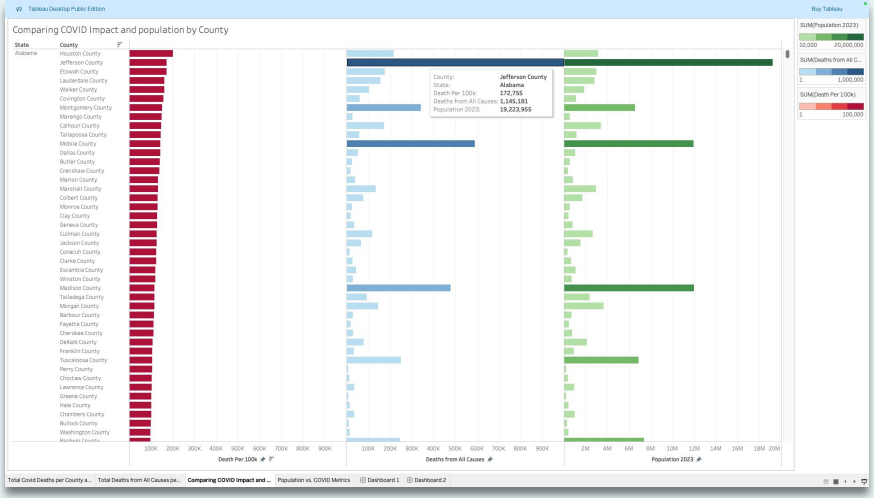
Cluster

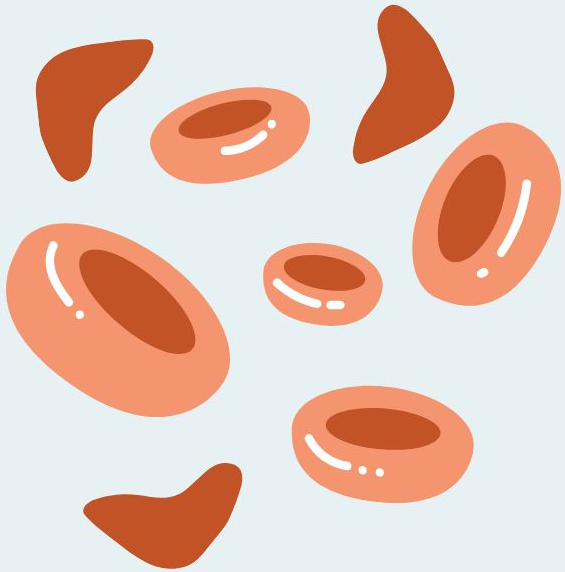
Geographic Distribution of COVID-19 Impact

Comparison of All-Cause Deaths

Temperature Correlation (2020-2023)







Modeling Techniques

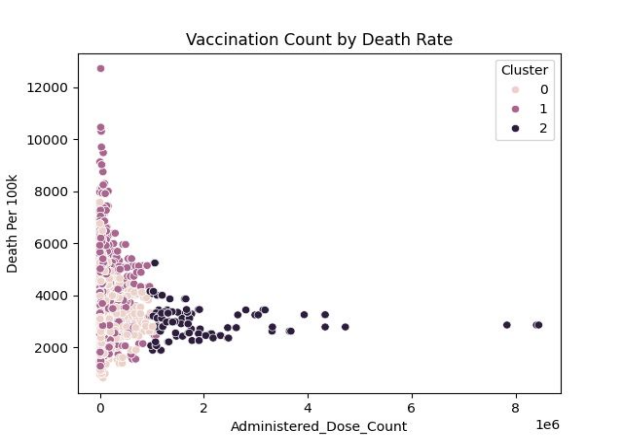
- Visualize the various datasets using Tableau
- Apply machine learning models to predict the COVID-19 death rate
  - Decision Tree Regression
  - Random Forests
  - Neural Networks
  - Linear Regression
  - K-Means Clustering

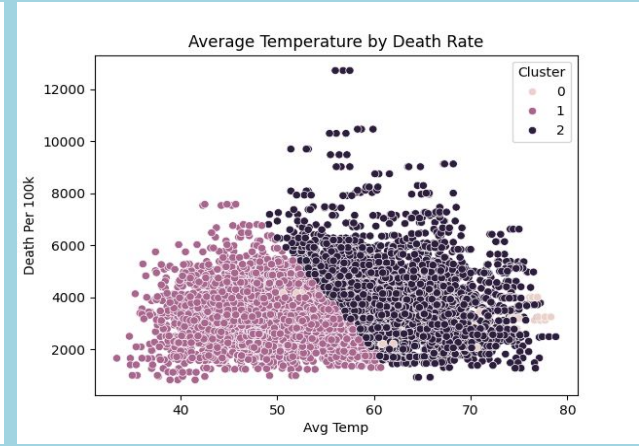
Setting up machine learning models

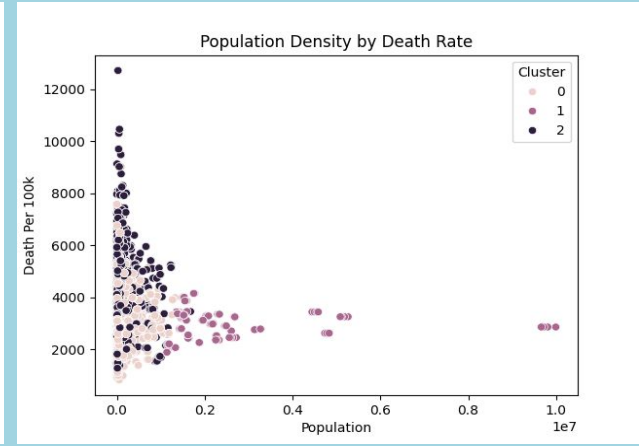
- ❏ Target (‘y’): COVID-19 Death Count
- ❏ Features (‘X’): Average Temperature, Vaccination Count, Population
- ❏ Scaled data
- ❏ Split data into training and testing datasets
- ❏ Apply relevant machine learning models
- ❏ Calculate success metrics: MSE, R^2, MAE, RMSE

KMeans Clustering

- Vaccination Count
  - Lower vaccination counts related to higher death count
- Average Temperature
  - Moderate to high temp vaguely associated with higher death rates
- Population Density
  - Smaller population size associated with higher death rates

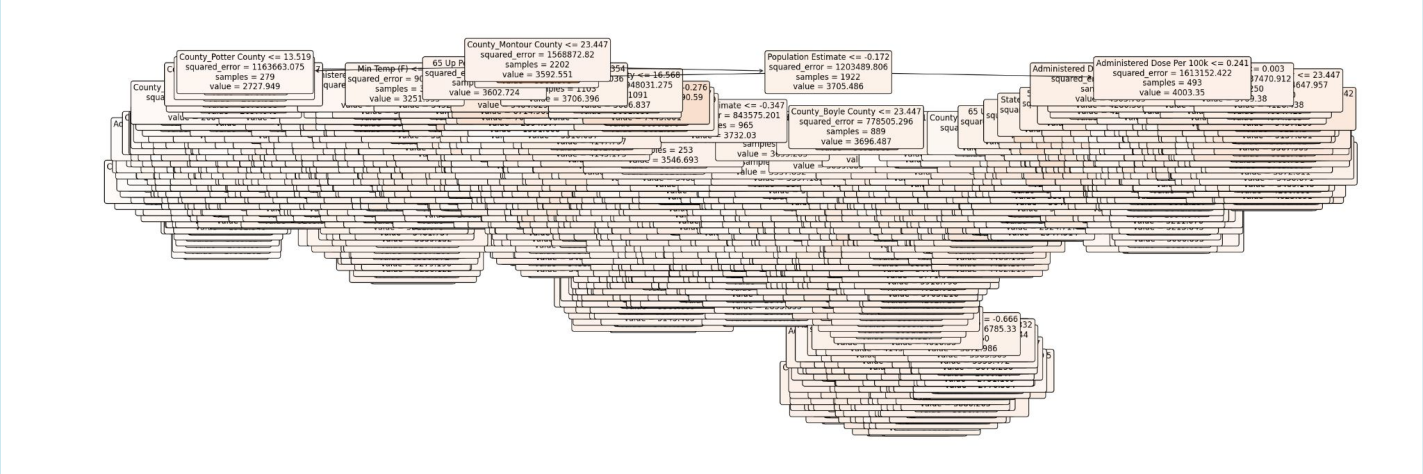






Decision Tree Regression

- Decision tree regression is used for continuous target variables
- High Mean Squared Error: **1,778,541**
- Negative R^2 value: **-0.25689**



Random Forest, Neural Networks, Linear Regression

- Random Forest
  - **MSE**: 1,920,068, **R^2**: 0.1005
- Neural Networks
  - **MSE**: 15,221,333
- Linear Regression
  - **MSE**: 1,961,067, **R^2**: -0.0004

- Poor performance by ML models
  - Large MSE values
  - Low/ negative R^2 values
- Room for improvement and/or different models

Future considerations to build a more accurate model

Environmental

- Flight path
- Heavy traffic flow
- Seasonal patterns

Demographics

- Gender
- Race / Ethnicity
- Poverty

Healthcare

- Vaccination rates
- Illness death counts
- Holistic medicine use
- Infrastructure