VIRGINIA TECH™

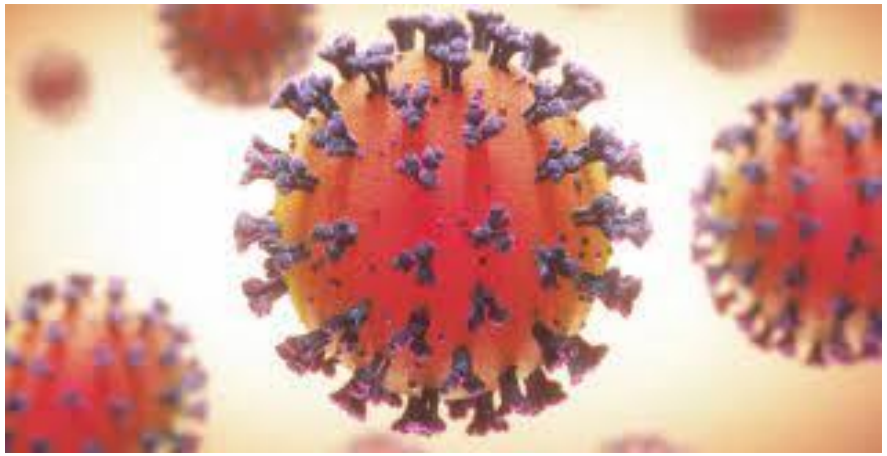
# Ability Evaluation of Counties against COVID-19 in United States

WENJUN HAN

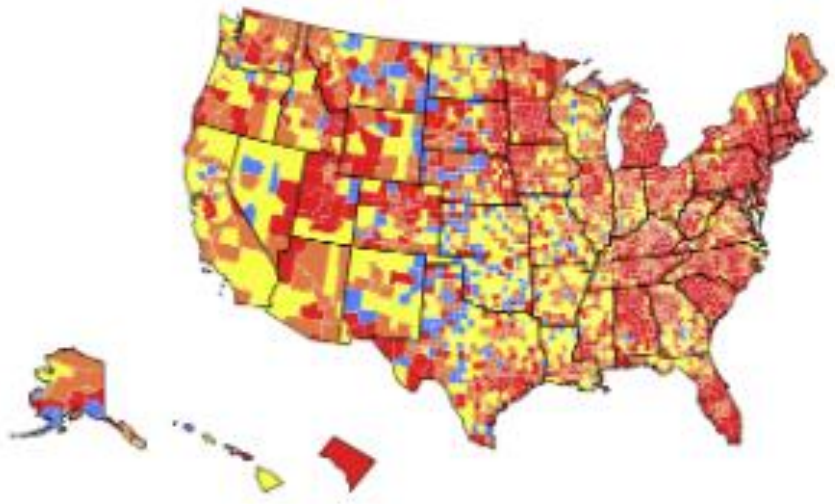
01

# Introduction

# *Covid-19 and Pandemic*



- March 11 WHO declared pandemic
- Caused nearly 3 million deaths around the world
- Total identified cases in US is nearly 30 million
- Health needs is estimated go beyond the capacity of US hospitals



## *Research Goals*

Evaluate the ability of counties against COVID-19 and identify vulnerable counties

02

# Data

# Data Description

| Data Title | Description |
|---|---|
| Total Cases | Data on cumulative coronavirus cases by county and by day |
| Total Deaths | Data on cumulative coronavirus deaths by county and by day |
| Underlying medical conditions | Percentage estimates of the prevalence for any of five underlying medical conditions that increase the risk for severe COVID-19-associated illness (obesity, diabetes, heart disease, Chronic obstructive pulmonary disease (COPD), Chronic kidney disease(CKD)) |
| County Population | Population estimation in 2019 by county |
| County Population of age 60+ | Total population of elder people who aged 60+ |
| Percent of population aged 60+ | Percentage of population of elder aged 60+ |
| Total ICU beds number | Number of ICU beds in the county |
| Residents aged 60+ per ICU bed | County residents Aged 60+ Per Each ICU Bed |
| GDP by dollars | Real gross domestic product by thousands of chained dollars(2019) |
| Poverty Condition | Individuals (all ages) in the county classified as living in poverty (2019)(log-transformed) |

# Data Exploration

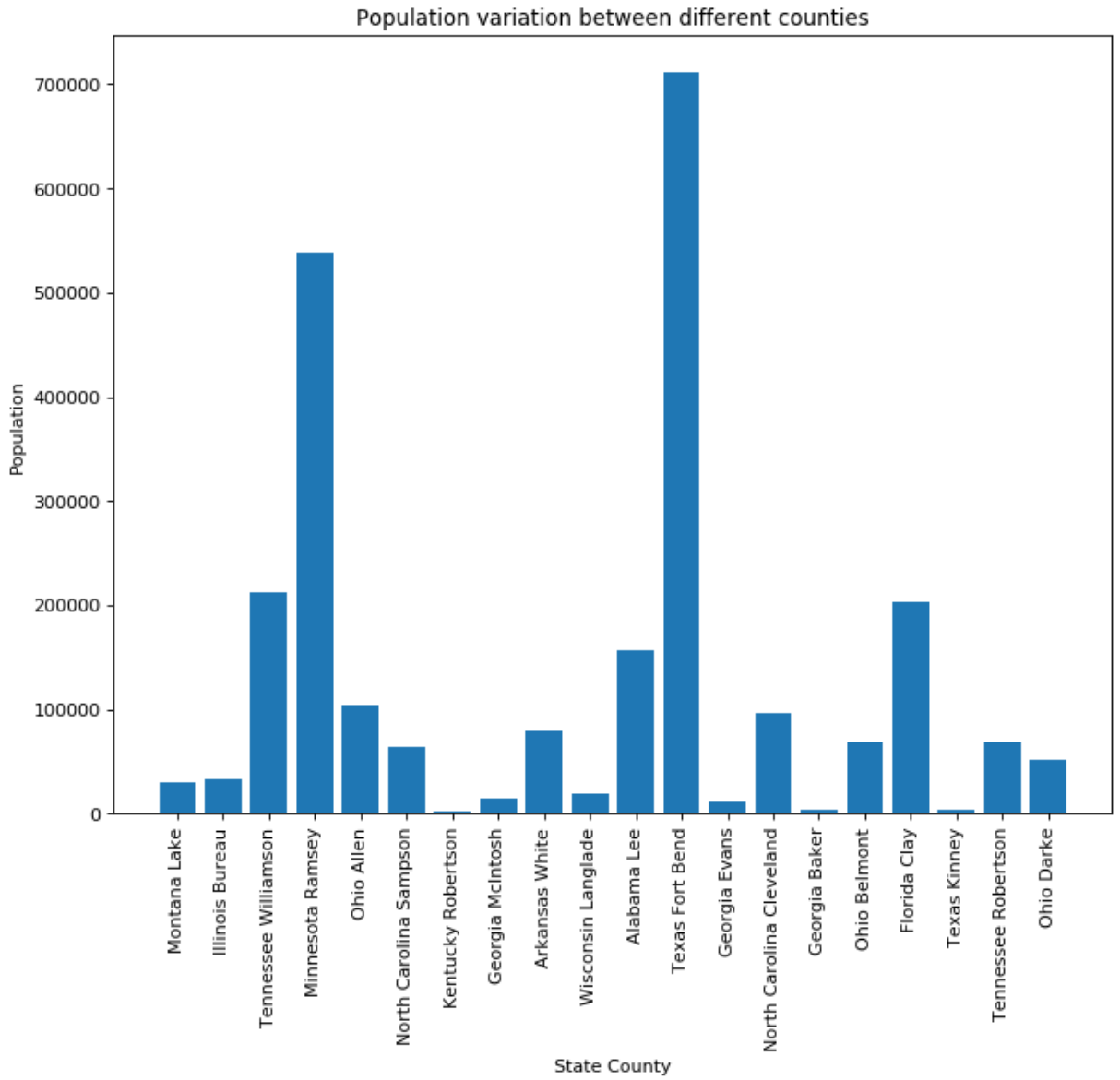
## Original Data

| state | county | icu beds | total population | population aged 60+ | percent of population aged 60+ | residents aged 60+ per each icu bed | gdp _dollars | cases | deaths | obesity | heart_disease | COPD | diabetes | CKD | poverty |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Alabama | Autauga | 6 | 55036 | 10523 | 19.1 | 1754.0 | 1501769 | 5683 | 69.0 | 35.8 | 7.9 | 8.6 | 12.9 | 3.1 | 10.916415 |
| Alabama | Baldwin | 51 | 203360 | 53519 | 26.3 | 1049.0 | 6140514 | 18211 | 224.0 | 29.7 | 7.8 | 8.6 | 12.0 | 3.2 | 12.279579 |
| Alabama | Barbour | 5 | 26201 | 6150 | 23.5 | 1230.0 | 762856 | 1956 | 40.0 | 40.7 | 11.0 | 12.1 | 19.7 | 4.5 | 9.997843 |
| Alabama | Bibb | 0 | 22580 | 4773 | 21.1 | 8470.0 | 389547 | 2309 | 52.0 | 38.7 | 8.6 | 10.0 | 14.1 | 3.3 | 9.914032 |
| Alabama | Blount | 6 | 57667 | 13600 | 23.6 | 2267.0 | 869049 | 5720 | 100.0 | 34.0 | 9.2 | 10.5 | 13.5 | 3.4 | 10.954973 |

## Data Summary

| | icu beds | population | people aged 60+ | %people aged 60+ | residents aged 60+ per each icu bed | GDP | cases | deaths | obesity | Heart Disease | COPD | diabetes | CKD | poverty |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 2,967.00 | 2,967.00 | 2,967.00 | 2,967.00 | 2,967.00 | 2,967.00 | 2,967.00 | 2,967.00 | 2,967.00 | 2,967.00 | 2,967.00 | 2,967.00 | 2,967.00 | 2,967.00 |
| mean | 23.10 | 100,073.07 | 20,980.69 | 24.94 | 5,132.19 | 5,743,794.47 | 8,227.42 | 132.55 | 35.03 | 8.64 | 9.11 | 13.07 | 3.45 | 10.23 |
| std | 84.00 | 327,317.68 | 61,097.24 | 5.53 | 3,619.17 | 24,207,952.06 | 31,440.97 | 495.70 | 4.47 | 1.75 | 2.34 | 2.70 | 0.56 | 1.49 |
| min | 0.00 | 74.00 | 29.00 | 5.80 | 10.00 | 22,870.00 | 1.00 | 0.00 | 15.20 | 3.50 | 3.50 | 6.10 | 1.80 | 5.02 |
| 25% | 0.00 | 10,919.50 | 2,820.50 | 21.40 | 1,176.50 | 376,257.50 | 864.50 | 14.00 | 32.50 | 7.50 | 7.35 | 11.20 | 3.10 | 9.27 |
| 50% | 0.00 | 25,574.00 | 6,262.00 | 24.60 | 8,470.00 | 963,407.00 | 2,074.00 | 36.00 | 35.40 | 8.60 | 8.90 | 12.80 | 3.40 | 10.11 |
| 75% | 12.00 | 66,275.00 | 15,814.00 | 27.90 | 8,470.00 | 2,734,368.50 | 5,399.00 | 87.00 | 37.90 | 9.80 | 10.60 | 14.80 | 3.80 | 11.07 |
| max | 2,126.00 | 10,105,722.00 | 1,800,341.00 | 64.20 | 8,470.00 | 726,943,301.00 | 1,121,349.00 | 16,854.00 | 49.90 | 15.10 | 19.90 | 25.60 | 5.90 | 16.11 |

# *Large Difference between County*



# *Data Pre-possessing*

- Directly use cases, deaths, population, GDP is not suitable due to large variation between counties
- Death ratio instead of Deaths
  - Death ratio = Deaths/Population
- Cases ratio instead of Cases
  - Cases ratio = Cases/Population
- Z-Score Normalization is applied

| state | county | icu beds | total population | population aged 60+ | percent of population aged 60+ | residents aged 60+ per each icu bed | gdp _dollars | obesity | heart_disease | COPD | diabetes | CKD | poverty | cases ratio | deaths ratio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Alabama | Autauga | -0.204 | -0.138 | -0.171 | -1.056 | -0.934 | -0.175 | 0.173 | -0.425 | -0.217 | -0.061 | -0.622 | 0.463 | 0.647 | -0.266 |
| Alabama | Baldwin | 0.332 | 0.316 | 0.533 | 0.246 | -1.128 | 0.016 | -1.193 | -0.483 | -0.217 | -0.394 | -0.444 | 1.378 | 0.174 | -0.420 |
| Alabama | Barbour | -0.215 | -0.226 | -0.243 | -0.261 | -1.078 | -0.206 | 1.271 | 1.351 | 1.278 | 2.455 | 1.879 | -0.154 | -0.339 | 0.012 |
| Alabama | Bibb | -0.275 | -0.237 | -0.265 | -0.695 | 0.922 | -0.221 | 0.823 | -0.024 | 0.381 | 0.383 | -0.265 | -0.210 | 0.613 | 0.800 |
| Alabama | Blount | -0.204 | -0.130 | -0.121 | -0.242 | -0.792 | -0.201 | -0.230 | 0.319 | 0.594 | 0.161 | -0.086 | 0.489 | 0.507 | 0.222 |

# Pearson Correlation Analysis

| | icu beds | population | people aged 60+ | %people aged 60+ | residents aged 60+ per each icu bed | gdp _dollars | obesity | heart_disease | COPD | diabetes | CKD | poverty | cases ratio | deaths ratio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| icu beds | 1.000 | 0.924 | 0.922 | -0.209 | -0.328 | 0.887 | -0.214 | -0.267 | -0.216 | -0.146 | -0.188 | 0.527 | -0.003 | -0.043 |
| population | 0.924 | 1.000 | 0.988 | -0.220 | -0.284 | 0.954 | -0.245 | -0.296 | -0.245 | -0.173 | -0.218 | 0.541 | -0.024 | -0.059 |
| people aged 60+ | 0.922 | 0.988 | 1.000 | -0.181 | -0.304 | 0.931 | -0.267 | -0.281 | -0.236 | -0.174 | -0.208 | 0.568 | -0.044 | -0.058 |
| %people aged 60+ | -0.209 | -0.220 | -0.181 | 1.000 | 0.315 | -0.201 | -0.026 | 0.594 | 0.320 | 0.259 | 0.499 | -0.453 | -0.285 | 0.039 |
| residents aged 60+ per each icu bed | -0.328 | -0.284 | -0.304 | 0.315 | 1.000 | -0.237 | 0.197 | 0.336 | 0.190 | 0.220 | 0.303 | -0.680 | 0.017 | 0.089 |
| gdp _dollars | 0.887 | 0.954 | 0.931 | -0.201 | -0.237 | 1.000 | -0.251 | -0.283 | -0.243 | -0.169 | -0.211 | 0.469 | -0.032 | -0.055 |
| obesity | -0.214 | -0.245 | -0.267 | -0.026 | 0.197 | -0.251 | 1.000 | 0.549 | 0.575 | 0.690 | 0.540 | -0.308 | 0.239 | 0.256 |
| heart_disease | -0.267 | -0.296 | -0.281 | 0.594 | 0.336 | -0.283 | 0.549 | 1.000 | 0.884 | 0.831 | 0.892 | -0.501 | -0.050 | 0.217 |
| COPD | -0.216 | -0.245 | -0.236 | 0.320 | 0.190 | -0.243 | 0.575 | 0.884 | 1.000 | 0.784 | 0.727 | -0.294 | -0.058 | 0.121 |
| diabetes | -0.146 | -0.173 | -0.174 | 0.259 | 0.220 | -0.169 | 0.690 | 0.831 | 0.784 | 1.000 | 0.916 | -0.328 | 0.045 | 0.292 |
| CKD | -0.188 | -0.218 | -0.208 | 0.499 | 0.303 | -0.211 | 0.540 | 0.892 | 0.727 | 0.916 | 1.000 | -0.448 | -0.015 | 0.298 |
| poverty | 0.527 | 0.541 | 0.568 | -0.453 | -0.680 | 0.469 | -0.308 | -0.501 | -0.294 | -0.328 | -0.448 | 1.000 | -0.080 | -0.182 |
| cases ratio | -0.003 | -0.024 | -0.044 | -0.285 | 0.017 | -0.032 | 0.239 | -0.050 | -0.058 | 0.045 | -0.015 | -0.080 | 1.000 | 0.467 |
| deaths ratio | -0.043 | -0.059 | -0.058 | 0.039 | 0.089 | -0.055 | 0.256 | 0.217 | 0.121 | 0.292 | 0.298 | -0.182 | 0.467 | 1.000 |

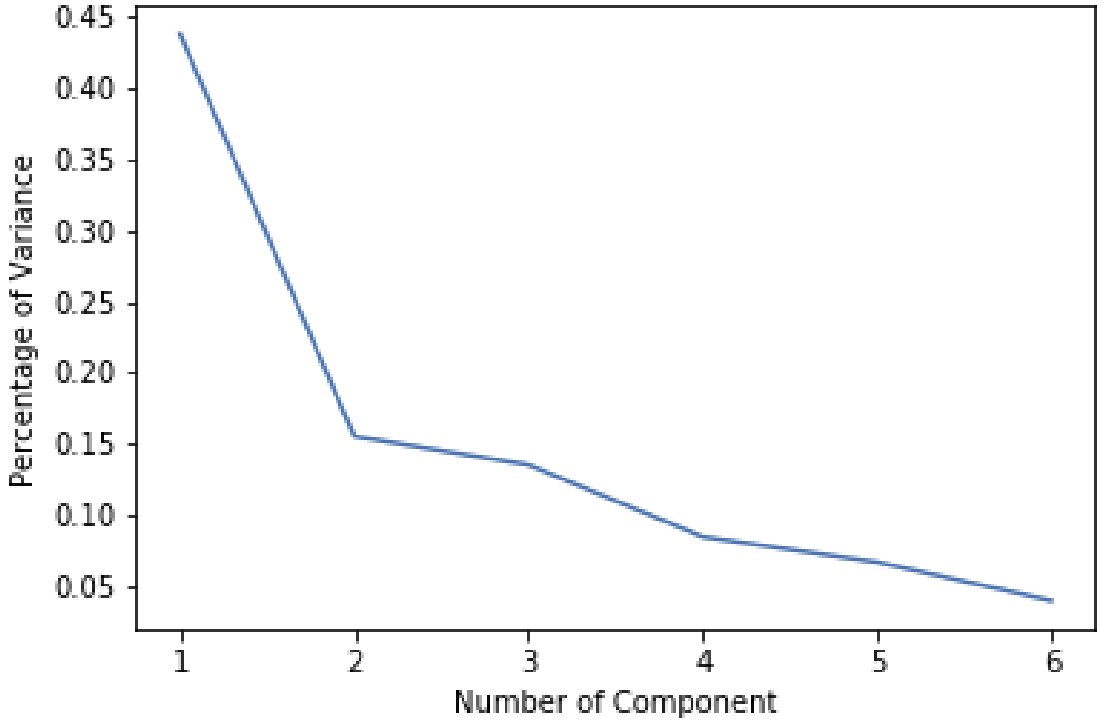


Delete ICU Beds, population, people aged 60+

# PCA

- First 6 components are selected
- Variance ratio: (total ~91.8%)

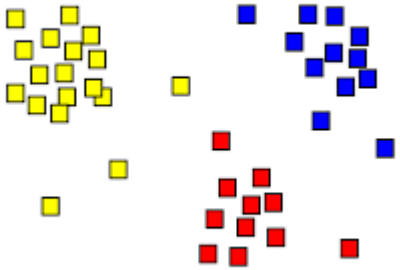
  [0.43765049 0.15518187 0.13522841 0.08474972 0.06695598 0.03996605]
- Percentage of Variance vs. Number of Component



03

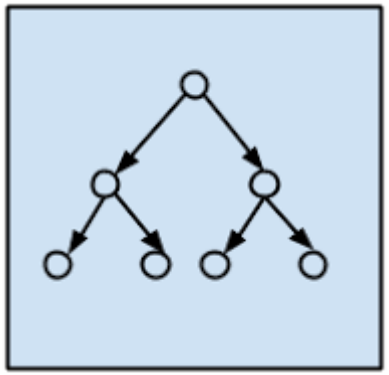
# Modeling

# Method



## Cluster Algorithm

Separate counties into groups



## Supervised Learning

Train and test model

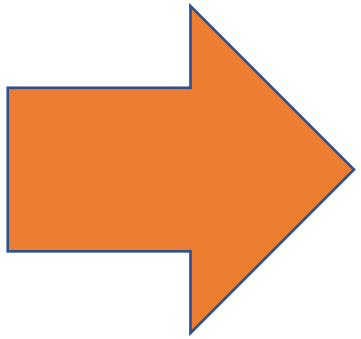
## Resource Allocation

Provide reference for decision-making

# Method Flow Chart

Cluster Algorithm

- K-Means
- Hierarchical Clustering
- DBSCAN

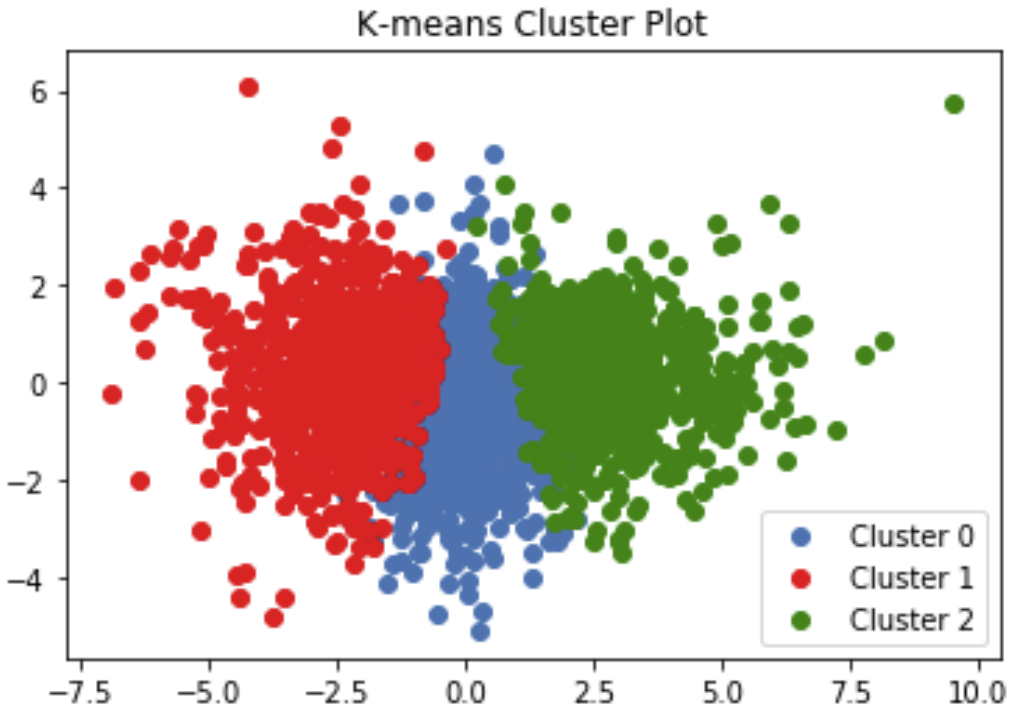


Classification Algorithm

- KNN
- Decision Tree
- Random Forest
- Logistic Regression

# *Clustering: K-means*

- Plan to have 3 clusters: Low risk, Moderate risk, High risk
- K-means algorithm assigns labeling as 0,1,2
- PCA data component 1 and component 2 for figure



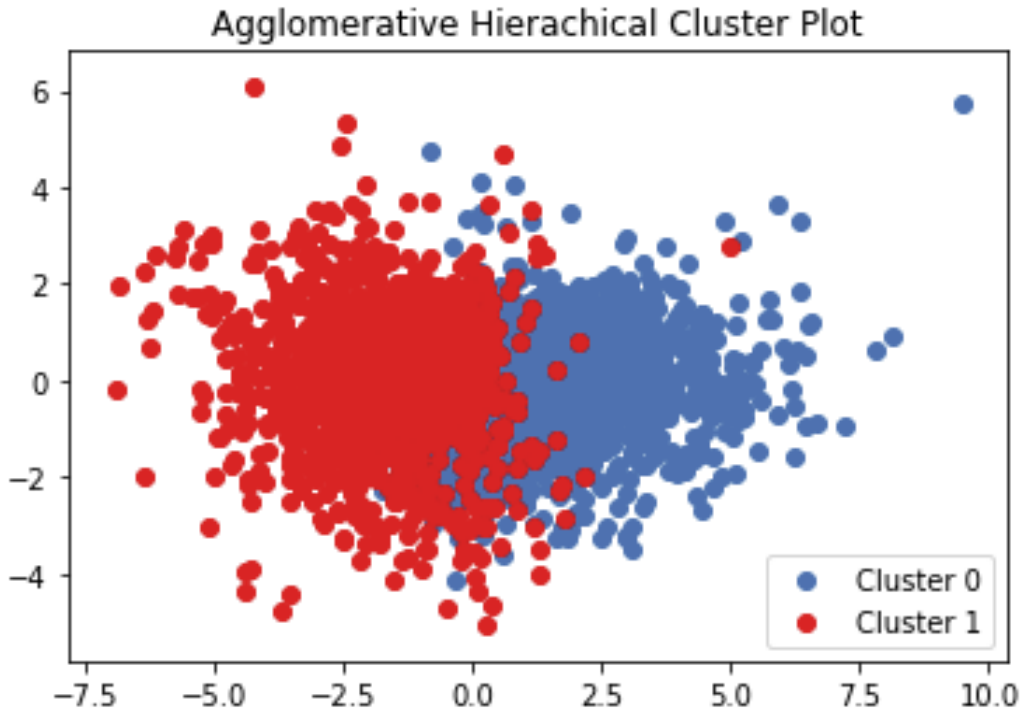

## *Data Details*

| GROUP | %aged 60+ | Elderly per ICU | %obesity | %heart disease | %diabetes | poverty | Cases ratio | Deaths ratio |
|---|---|---|---|---|---|---|---|---|
| 0 | 26.1 | 5996 | 34.8 | 8.5 | 12.4 | 9.7 | 0.086 | 0.0015 |
| 1 | 27 | 6384 | 38.4 | 10.6 | 16.2 | 9.58 | 0.084 | 0.0019 |
| 2 | 20 | 2002 | 31.4 | 6.6 | 10.5 | 11.9 | 0.081 | 0.0011 |

Group 1: High risk, Group 0: Moderate risk, Group 2: Low risk

# *Clustering: Hierarchical Clustering*

- Hierarchical clustering algorithm classify them to two clusters
- Assigned labels: 0 and 1
- PCA data component 1 and component 2 for figure



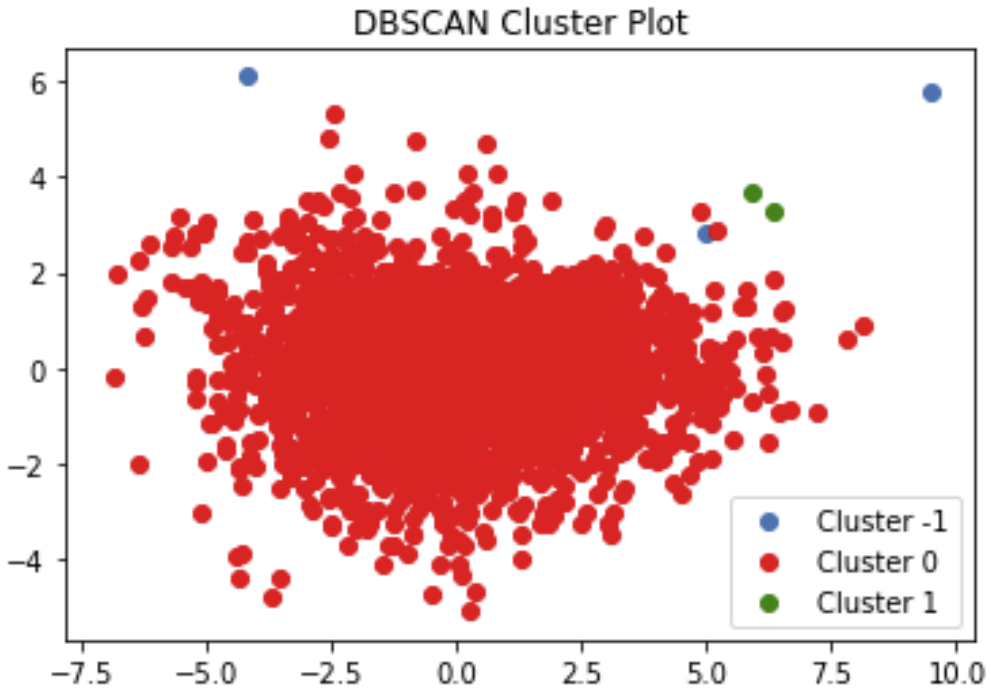

## *Data Details*

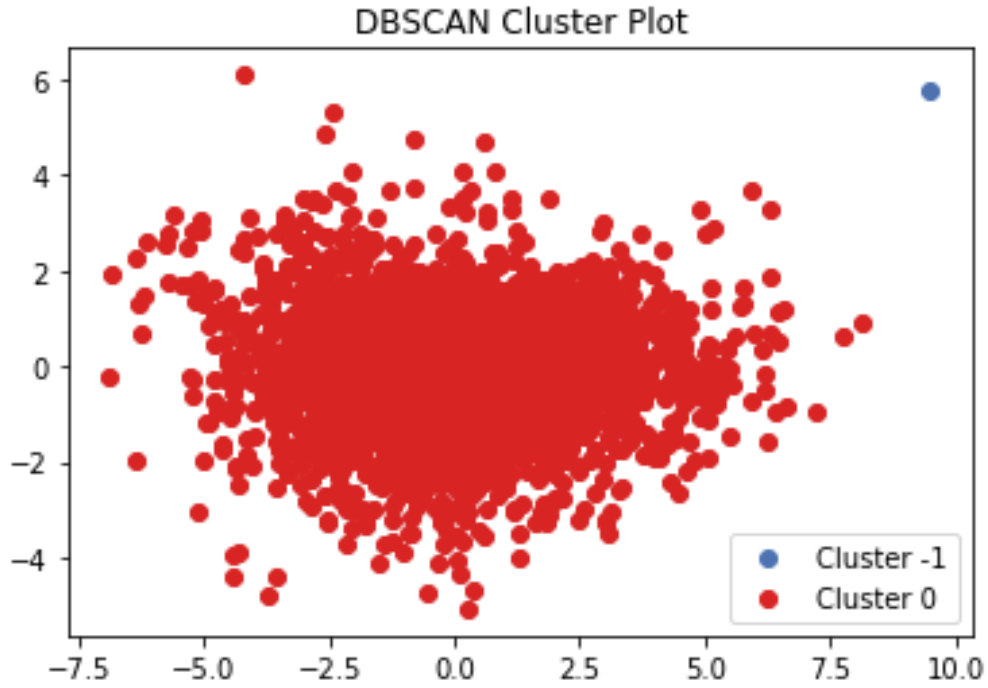
| GROUP | %aged 60+ | Elderly per ICU | %obesity | %heart disease | %diabetes | poverty | Cases ratio | Deaths ratio |
|---|---|---|---|---|---|---|---|---|
| 0 | 23 | 3909 | 33 | 7.43 | 11.3 | 10.8 | 0.082 | 0.0012 |
| 1 | 27 | 6429 | 37 | 9.92 | 15 | 9.6 | 0.088 | 0.0018 |

Group 0: Low risk, Group 1: High risk

# *Clustering: DBSCAN*

- Assigned labels: [0 ,1], [-1,0,1]
- DBSCAN is not suitable for the data since no pattern/shape



# *Classification: KNN*

- 5-fold cross validation
- Evaluation through classification report

## *K-means + KNN*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.92 | 0.94 | 0.93 | 1354 |
| 1 | 0.93 | 0.93 | 0.93 | 885 |
| 2 | 0.96 | 0.93 | 0.95 | 728 |
| accuracy | | | 0.93 | 2967 |
| macro avg | 0.94 | 0.93 | 0.93 | 2967 |
| weighted avg | 0.93 | 0.93 | 0.93 | 2967 |

## *Hierarchical + KNN*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.91 | 0.99 | 0.95 | 1527 |
| 1 | 0.99 | 0.90 | 0.94 | 1440 |
| accuracy | | | 0.94 | 2967 |
| macro avg | 0.95 | 0.94 | 0.94 | 2967 |
| weighted avg | 0.95 | 0.94 | 0.94 | 2967 |

# *Classification: Decision Tree*

- 5-fold cross validation
- Evaluation through classification report

## *K-means + DT*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.92 | 0.90 | 0.91 | 1354 |
| 1 | 0.92 | 0.94 | 0.93 | 885 |
| 2 | 0.91 | 0.92 | 0.92 | 728 |
| accuracy | | | 0.92 | 2967 |
| macro avg | 0.92 | 0.92 | 0.92 | 2967 |
| weighted avg | 0.92 | 0.92 | 0.92 | 2967 |

## *Hierarchical + DT*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.89 | 0.89 | 1527 |
| 1 | 0.88 | 0.90 | 0.89 | 1440 |
| accuracy | | | 0.89 | 2967 |
| macro avg | 0.89 | 0.89 | 0.89 | 2967 |
| weighted avg | 0.89 | 0.89 | 0.89 | 2967 |

# *Classification: Random Forest*

- 5-fold cross validation
- Evaluation through classification report

## *K-means + RF*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.92 | 0.95 | 0.94 | 1354 |
| 1 | 0.97 | 0.94 | 0.95 | 885 |
| 2 | 0.94 | 0.93 | 0.93 | 728 |
| accuracy | | | 0.94 | 2967 |
| macro avg | 0.94 | 0.94 | 0.94 | 2967 |
| weighted avg | 0.94 | 0.94 | 0.94 | 2967 |

## *Hierarchical + RF*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.95 | 0.92 | 1527 |
| 1 | 0.95 | 0.89 | 0.91 | 1440 |
| accuracy | | | 0.92 | 2967 |
| macro avg | 0.92 | 0.92 | 0.92 | 2967 |
| weighted avg | 0.92 | 0.92 | 0.92 | 2967 |

# *Classification: Logistic Regression*

- 5-fold cross validation
- Evaluation through classification report

## *K-means + LR*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.96 | 0.93 | 0.95 | 1354 |
| 1 | 0.94 | 0.98 | 0.96 | 885 |
| 2 | 0.95 | 0.96 | 0.95 | 728 |
| accuracy | | | 0.95 | 2967 |
| macro avg | 0.95 | 0.96 | 0.95 | 2967 |
| weighted avg | 0.95 | 0.95 | 0.95 | 2967 |

## *Hierarchical + LR*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.91 | 0.90 | 1527 |
| 1 | 0.90 | 0.88 | 0.89 | 1440 |
| accuracy | | | 0.89 | 2967 |
| macro avg | 0.89 | 0.89 | 0.89 | 2967 |
| weighted avg | 0.89 | 0.89 | 0.89 | 2967 |

# ***Classification***: *Logistic Regression*

*K-means + LR*

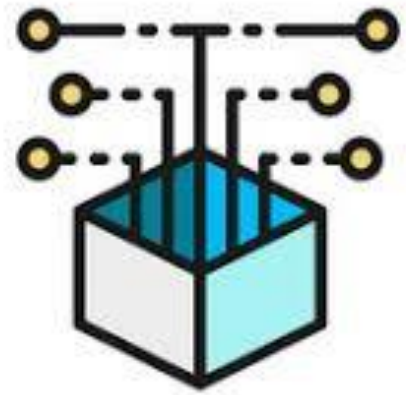
## *Confusion Matrix*

| | | | |
|---|---|---|---|
| **Moderate risk** | 1263 | 54 | 37 |
| **High risk** | 15 | 870 | 0 |
| **Low risk** | 31 | 0 | 697 |
| | **Moderate risk** | **High risk** | **Low risk** |

03

# Conclusion

# Conclusion



- Goal: Evaluate the ability of counties against COVID-19, identify vulnerable counties
- A clustering + supervised learning model is selected for county classification(K-means + Logistic Regression)
- Successfully identify counties that are vulnerable to COVID with high precision
- Can be a powerful tool for decision makers to allocate the priority of medical resources

# Reference & Resource


https://covid.cdc.gov/covid-data-tracker/#underlying-med-conditions
https://covid.cdc.gov/covid-data-tracker/#county-view
https://www.kaggle.com/fireballbyedimyrnmom/us-counties-covid-19-dataset
https://www.kaggle.com/jaimeblasco/icu-beds-by-county-in-the-us
https://www.kaggle.com/laurindogarcia/covid-19-race-gender-poverty-risk-us-county
https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas
https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html
https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/

# Reference & Resource

Ezekiel J Emanuel, Govind Persad, Ross Upshur, Beatriz Thome, MichaelParker, Aaron Glickman, Cathy Zhang, Connor Boyle, Maxwell Smith,and James P Phillips. 2020. Fair allocation of scarce medical resourcesin the time of Covid-19.

Nezir Aydin and Gökhan Yurdakul. 2020. Assessing countries' performances against COVID-19 via WSIDEA and machine learning algorithms. Applied Soft Computing 97 (2020), 106792.

*Thank you!*