# Case Studies in Data Science
# Work Integrated Learning Project

## Identifying Potential Clusters for COVID-19 Vaccination in United States using County-level Data

Final Report

Group 10

Alejandro Rivera Alvarez (S3758098)
Anish Parajuli (S3837878)
Fitrio Pakana (S3778275)
Kanishka Tamang (S3756188)
Nischay Bikram Thapa (S3819491)

October 2020

# Table of Contents

**EXECUTIVE SUMMARY**.......................................................................3

**1. INTRODUCTION**.......................................................................3

PROBLEM STATEMENT.......................................................................3

**2. LITERATURE REVIEW**.......................................................................4

**3. METHODOLOGY**.......................................................................4

3.1. DATA COLLECTION.......................................................................5
3.2. DATA PREPARATION AND MANIPULATION.......................................................................5
3.3. CLUSTERING MODEL SELECTION.......................................................................6
3.4. MODEL VALIDATION.......................................................................7
3.5. SIR MODEL.......................................................................7
3.6. DASHBOARD.......................................................................9

**4. IMPACT AND SIGNIFICANCE**.......................................................................10

**5. PROJECT MANAGEMENT**.......................................................................10

**6. CONCLUSION AND FURTHER DEVELOPMENT**.......................................................................12

**REFERENCES**.......................................................................13

## Executive Summary

As of September 2020, COVID-19 disease has infected around 32.7 million people with close to one million death worldwide where United States currently as the country with highest cumulative number of cases, stands at around 6.7 million cases. This pandemic has adversely impacted live of millions of people around the world with economic impact at its core. Fortunately, to date there are at least 3 companies have advanced to late-stage testing for potential COVID-19 vaccines. However, when a vaccine is finally approved, likely many countries including United States will experience limited supply and therefore an effective, data-driven vaccine deployment strategy is essential. KRAAN Project created prototype of application that provides machine-learning backed information and interactive simulation to support the United States public health officials in developing such strategy. The application groups together counties based on demographic features and epidemiological characteristics such as active cases per thousand population, population density, and 14 days average of new cases. Based on their domain expertise, public health officials and/or epidemiologists makes an interim decision on clusters vaccination prioritisation. The clusters are then fed into an SIR simulator which uses epidemiology mathematical formula to simulate the effects of vaccination over the dynamics of susceptible, infectious, and recovered cases at the country-level population. Based on domain expertise and taking into consideration the information provided by KRAAN, public health officials can develop an effective and well-informed vaccine deployment strategy.

## 1. Introduction

The objective of this Report is to provide detailed information pertaining to a project nicknamed as “KRAAN”, which is a data science Work-Integrated Learning (WIL) Project to build a solution based on a given COVID-19 related challenge. The Report covers brief background of the project, literature review, methodology, and glimpses on the resulting solution prototype.

COVID-19 is a popular term for an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first outbreak happened in Wuhan, Hubei province of China in December 2019 and infected more than 80,000 people with close to 5000 deaths. World Health Organisation declared the outbreak as pandemic in March 2020. As of September 2020, the disease has infected around 30 million people, including 6.7 million in United States, currently the highest number in the world by country, followed by India with 5.3 million cases and Brazil with 4.5 million cases [19].

As a response to the pandemic, many pharmaceutical companies, academic institutions, and government agencies around the world have started developing potential vaccines. In United States as of September 2020, at least 3 companies – Pfizer, Moderna, and AstraZeneca have advanced to late-stage testing for potential coronavirus vaccines. However, most experts think that once available, likelihood the vaccines will have limited supply [1] [2] [3] [4] due to limited production capacity. Government agencies including public health officials will need to start crafting most effective vaccine deployment strategy by answering some critical questions such as who will get the vaccine first and why.

## Problem Statement

Once possible solution to answer this challenge is to look into the cases dynamics and other relevant demographic data to get insights whether there are similar patterns in the population that can be used to construct a set of logical groups or “clusters” of people. For United States where there are more than three thousand counties to be considered, this would not be an easy task. Significant time and effort are needed if such tasks are to be done manually. Once this task is accomplished, based on their domain expertise, public health officials and/or epidemiologists can then make hypothesis of vaccination prioritisation for these clusters e.g. based on some prevailing vaccination objective, which cluster or set of clusters should have higher priority over the others. Further, the hypothesis needs to be validated using a standard disease spread modelling which allows users to observe the dynamics of cases in the population if vaccination to be administered in the selected cluster or set of clusters.

## 2. Literature Review

Vaccine is a product that stimulates immune system in human's body causing it to produce antibodies to fight infections caused by germs [20]. Vaccinated people will then have immunity to a particular disease, where they can be exposed to it without becoming infected. Immunization by administration of a vaccine remains as one the main approaches to protect communities and limit diseases spread.

In a disease outbreak caused by a novel virus, research and development of vaccine will likely be a part of public health response [5]. Within the first few months or even years after a new vaccine is approved for its safety and efficiency, there is a likelihood that it will have limited supply [6] [7], as observed in the 2009 H1N1 (swine flu) pandemic [8] and 2014 Ebola epidemic [9]. One of the main contribution factors of the initial shortage of supply is due to limited production capacities [10].

Driven by the limited supply of vaccines, it is imperative for countries to decide on prioritisation of vaccination target groups because such decision will define the deployment and vaccination activities that works best for them [10]. For influenza A (H1N1) 2009 pandemic, SAGE recommended divided population into four vaccination target groups and noted that the order of priority should be determined by each country based on country-specific conditions [11]. The target groups are health-care workers, pregnant women, person with underlying conditions, and age-specific populations. COVID-19 which is a respiratory illness caused by a large family of viruses called coronavirus, shares similar traits with influenza disease in terms of the disease presentation and transmission [12] and therefore it is likely that the target groups for vaccination will be similar as well.

In the context of disease outbreak, another possible approach that can be used to support in the decision making for vaccine deployment strategy is by grouping communities based on their epidemiological characteristics such number of new cases, deaths, etc. At the moment, there have not been many references that discuss the use of unsupervised machine learning technique to do such grouping that results in "clusters" of people or population subsets. The most recent work is by Zarikas et. al. [21] which clusters countries based on active cases, active cases per population and active cases per population and per area.

Impact of a disease intervention such as vaccination to the dynamics of cases in communities can be projected using some epidemiological modelling technique, such as SIR model. It is a basic compartmental model in epidemiology, considered to have reasonable predictive power for infectious disease that transmits between human [13]. SIR is acronym for Susceptible, Infectious, and Recovered, represent the number of people in each compartment at a particular time [14]. For the normal cases projection (i.e. with no vaccine intervention), the model takes input of number of cases at starting point ($T_0$) then using mathematical equations, it projects the dynamics of cases given epidemiology parameter estimates (i.e. transmission rate and recovery rate) within specified time continuum [15]. If vaccination introduced into the population, vaccinated people would immediately move from Susceptible to Recovered compartment, therefore with slight modification, the basic SIR model can also be used to simulate the effect of vaccination in the dynamics of cases [16].

## 3. Methodology

The project objective was to create a useful tool that helps public health officials in the US to make informed-based decisions about the COVID-19 vaccine distribution across counties in the US.

To accomplish this goal, the project team investigated some reliable sources from COVID-19 historical data and US demographic data. After selecting the proper data sources, a process of data manipulation was performed to clean the data and prepare it for further analysis. Then, an agglomerative clustering technique was applied to group counties based on significant descriptive features. Then, a simulation of the SIR model for spread of disease was implemented based on the predefined clusters and input parameters. Finally, an interactive dashboard was developed to display the functionalities such as summary statistics, available clusters and SIR model simulation.

All data pre-processing, clustering model creation and evaluation, SIR model simulation, and application development was performed using Python version 3.7. The code and files are stored and managed in a central file repository that allows collaborative teamwork.
The following sections describe in detail how each process was performed.

## 3.1. Data Collection

The COVID-19 data was sourced from John Hopkins University available in their GitHub repository (https://github.com/CSSEGISandData/COVID-19) [22]. Additional data related to county-level COVID-19 were referenced from Worldometer website (https://www.worldometers.info/coronavirus/country/us).
The United States population by county was retrieved from the US Census Bureau (https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html) while the land area data were also sourced from US Census bureau (https://www.census.gov/quickfacts/fact/note/US/LND110210). As the John Hopkins data had the cumulative sum of cases and deaths, the latest updated file from the data source was taken into account.

## 3.2. Data Preparation and Manipulation

Several steps were performed during this process and are explained below.

- **Data Retrieval**: The data was retrieved from different sources and then it was structured in a "Tidy format", meaning that each variable had its own column, each observation had its own row and each value in its own cell. In total 2 datasets were created, containing COVID-19 related information and demographic data in the US.
- **Data combination**: Then the structured datasets were combined into one master dataset. In this process, the total confirmed cases, deaths, population, area data were added into one master data.
- **Data cleansing**: This step involved dropping unnecessary columns, formatting, removing duplicates and performing data validation. The main objective was to obtain a structured dataset with an appropriate data type and actual values as reported by trusted sources. Missing values were imputed with the actual values from a trusted source (US census).
- **Feature engineering**: New attributes were created from the existing ones. For example, recovered cases for the county was missing, so this column was calculated as (Confirmed Cases per county / Confirmed cases from the state) * Recovered Cases from state. Similarly, "Mortality rate" and "infection rate" by county were created from confirmed cases, deaths, and total population. "Active per 1000 population" column was created with an Active number of cases divided by the total population for that county.

The final dataset (master dataset) had the required information by each county in the US with x number of rows and y number of columns. The columns information is detailed below.

- Data Attributes
- Active - Active number of cases
- Confirmed - Confirmed Cases
- County - Name of US county
- Date - Date the data was retrieved
- Deaths- Deaths recorded due to coronavirus
- Death per 1000 pop - Deaths per 1000 population in a county
- FIPS- Federation Information Processing Standard
- Infection rate - Total number of confirmed cases / Total population in a county
- Land Area - Total area of a county in square kilometres
- Lat - Latitude
- Long - Longitude
- Mortality rate- Deaths/ Confirmed *100
- Population - Total population in a county
- Recovered- Total number of recovered cases in a county
- State - US state
- Pop density - Population / land area
- Active per 1000 pop - Active number of cases per 1000 population
- Recovered per 1000 pop - Recovered number of cases per 1000 population
- 14-day average - Average number new cases within a 14-day interval

### 3.3. Clustering Model Selection

As the main task was to perform clustering and then simulating the groups to see the dynamics of susceptible, infectious and recovered cases. An Agglomerative Hierarchical Clustering and SIR model were used in this project.

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other [23].

In the Feature Selection for clustering, the following attributes were selected to find similarities between US counties and grouped them together.

- Active per 1000 cases.
- Population density.
- 14 days average of new cases.

These features were standardized to have the same scale by removing the mean and scaling to unit variance. The standard score of sample x is calculated as: z = (xu) / s.

The hyperparameters of clustering algorithms i.e. the ones used to find similarities, were by using Euclidean distance as the affinity measure and ward’s method as the linkage.

Ward's method means calculating the incremental sum of squares. Half square Euclidean distance is the only distance measure that can be used with this clustering method. Therefore, the distance measure is automatically set to Half square Euclidean distance when Ward's method is selected [18]. An attempt of using density-based clusters resulted an undesirable output (were not as per assumptions) and therefore was discarded.

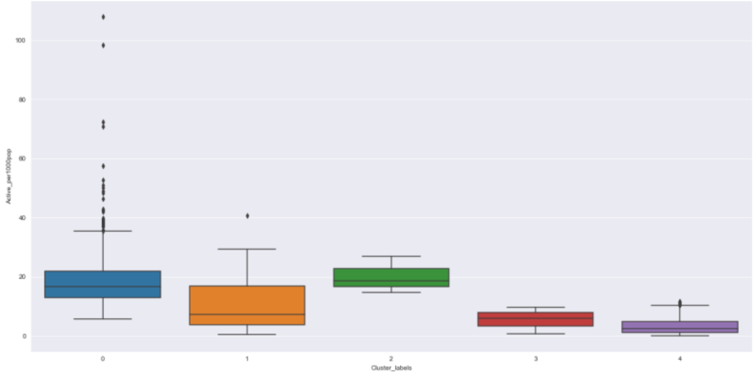


Figure 3.1: Active cases per 1,000 people on different clusters

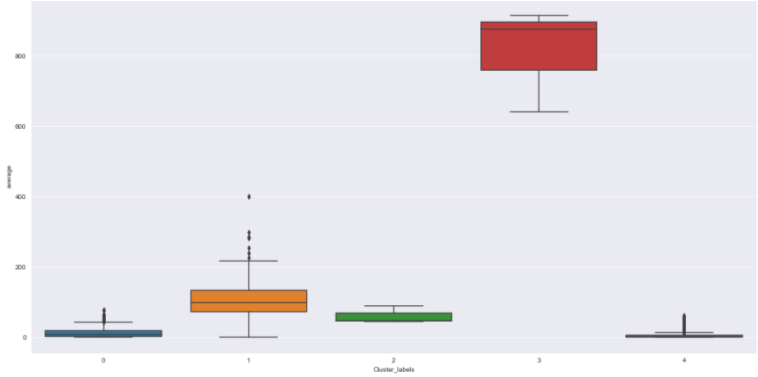


Figure 3.2: 14-day average cases on different clusters

## 3.4. Model Validation

The number of clusters was selected with the help of visual intuition from a dendrogram and with the highest silhouette score.

The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample [24]. The Silhouette Coefficient for a sample is (b - a) / max(a, b). To clarify, b is the distance between a sample and the nearest cluster that the sample is not a part of. We achieved a silhouette score of 0.7 for 5 clusters.

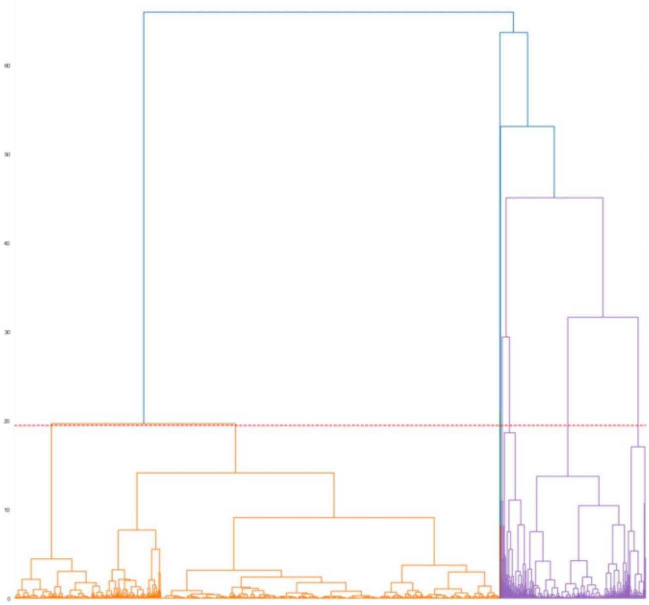


Figure 3.3: Dendrograms of the clusters

## 3.5. SIR Model

The SIR model in its basic form first introduced in 1927 by Kermack and McKendrick in their paper about mathematical theory of epidemics [17]. It is compartmental model because population is assigned to compartment with labels, S – for people susceptible to infection, I – for infectious people, and R – for people that either recovered or died due to the infection.

Movement between compartments is determined by the disease epidemiological characteristics. People from S moves to I compartment determined by the disease transmission rate (beta), and from I to R based on recovery rate (gamma). The model assumes that recovered people will retain the immunity and therefore will never go back to I compartment. The model ignores the vital dynamics (births and deaths), therefore population N is constant [15]. The S, I, and R are modelled as functions of a time variable t, which is measured in days and can be expressed by the following set of differential equations [14]:

$$\frac{dS}{dt} = -\frac{\beta IS}{N}$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

where ($\beta$) is the transmission rate and ($\gamma$) is the recovery rate.

If vaccination intervention being introduced into the population, susceptible people that have been vaccinated will flow directly into the recovered compartment. The number of people considered to be immune after vaccination depends on the vaccine efficacy and the amount of available vaccine stock. Therefore, the differential equations slightly modified as below:

$$\frac{dS}{dt} = -\frac{\beta IS}{N} - (p * v * S)$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I$$

$$\frac{dR}{dt} = \gamma I + (p * v * S)$$

where p is the proportion of population being vaccinated (assumed constant over time) and v is the vaccine efficacy (normally represented in percentage).

The differential equation implementation in KRAAN takes advantage of existing python implementation ordinary differential equation (ODE) function from scipy.integrate package. Taking the input of $t_0$ value of S, I, R together with population N, transmission rate $\beta$, and recovery rate $\gamma$, the function calculates $t_i$ value of S, I, R for given i number of days. The time series line then plotted based on the series of output from this function. This plot is also referred to as 'Baseline' because it the cases projection without vaccine intervention. Figure 3.4 below shows the output of SIR simulation in KRAAN thorough time. The x-axis represents the number of days starting from $t_0$. The y-axis displays the proportion of population. The time series lines show the variation in each compartment.

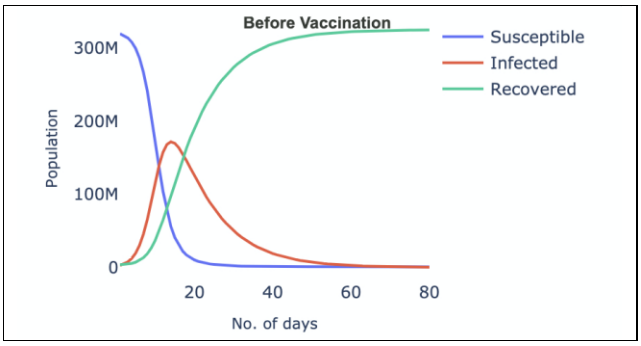


Figure 3.4: SIR Model without vaccination (baseline)

Number of population (N) and $T_0$ data for each of SIR compartment for county in United States are based on the master dataset. External parameters such as number of days to simulate, transmission rate, recovery rate, proportion of population to be vaccinated, and vaccine efficacy are designed as interactive input parameters.

Due to the nature of available master dataset, upon reading the $t_0$ data, some calculations need to be performed as follows:

Susceptible = Population – Confirmed Cases
Recovered = Deaths + Recovered
Infectious = Confirmed Cases - Recovered

Based on clustering output, system allows user to select one or more clusters to simulate the SIR if these clusters being vaccinated. The series for each SIR compartments for these clusters are calculation using differential equations that include vaccination parameters while for the remaining clusters (i.e. those not selected for vaccination), calculation performed using the normal differential equations. The output of these calculations then combined to produce the final dataset that contains the whole population SIR series where the selected clusters being vaccinated. The visualization of SIR with vaccination intervention is based on this dataset with the differential equations taking additional parameters related to the vaccination effort.

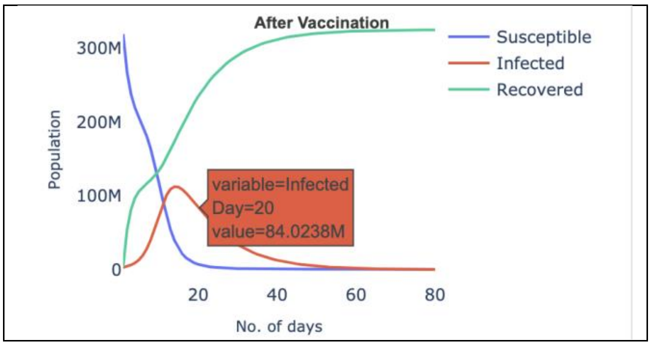
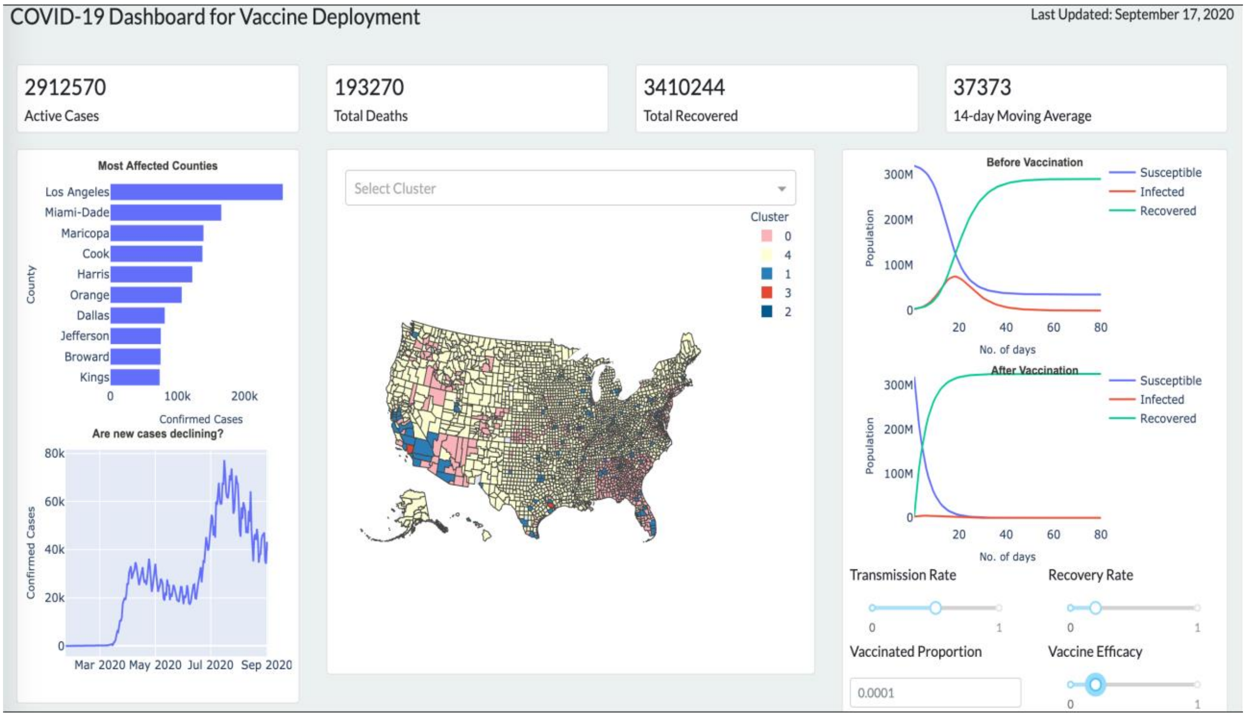


Figure 3.5: SIR Model with vaccination

## 3.6. Dashboard

Python 3.7 is used for building the application. It is chosen because there are numerous frameworks like pandas, plotly, sci-kit-learn, etc that can be used for building data science applications. Also, all the team members were already familiar with this programming language. Pandas is used for data pre-processing and cleaning; Plotly is used for data visualization and scikit-learn is used for building machine-learning model. Similarly, the front-end web application is created using the Dash framework. It is built on top of Plotly.js, React, Flask, and supports the direct integration of visualizations created using Plotly. Although the Dash framework can be used for building all the web UI components like sliders, layout, input field a Dash Bootstrap Components framework is chosen for building the web application as it makes it easier to build styled apps with responsive layouts using modern bootstrap UI components. Figure 3.6 below shows the working prototype of KRAAN dashboard.



COVID-19 Dashboard for Vaccine Deployment

Last Updated: September 17, 2020

2912570
Active Cases

193270
Total Deaths

3410244
Total Recovered

37373
14-day Moving Average

Select Cluster

Vaccinated Proportion: 0.0001

Figure 3.6: KRAAN Dashboard prototype

## 4. Impact and Significance

KRAAN web application provides machine-learning backed information and interactive simulation to support the United States public health officials in developing effective, data-driven vaccine deployment strategy. Using an unsupervised learning technique, KRAAN groups together counties based on demographic features and epidemiological characteristics. The resulting groups are referred to as “clusters”. User can zoom into a cluster to see more detailed information includes additional information such as the top-ten counties from selected clusters based on number of active cases and other important statistics. KRAAN has a built-in SIR simulator that can be used to simulate the dynamics of susceptible (S), infectious (I), and recovered (R) cases at the country-level population within a specified number of days. User can gauge which cluster or combination of clusters that will give the most desirable output on the overall country population from the SIR perspective, if a vaccine intervention introduced to those clusters.

More specifically, KRAAN answers the following needs:

1. Public health official able to identify group of counties (clusters) across all states in the United States based on COVID-19 active cases per thousand population, population density, and 14 days average of new cases.

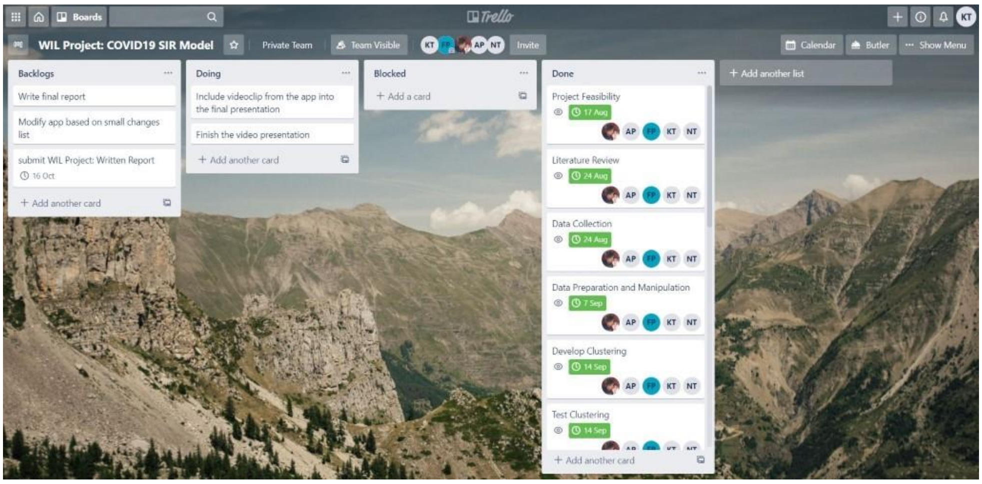
2. Public health official able to simulate country-level COVID-19 cases dynamics in terms of Susceptible, Infectious, and Recovered numbers when vaccine being deployed to one or more clusters. Such information provides valuable insight in crafting of most effective vaccine deployment strategy.

3. General public able to access information related to estimates of COVID-19 cases distribution in their area for necessary adjustment on their personal plan.

## 5. Project Management

The project was built on Agile methodology using the Scrum framework. Agile Scrum framework is chosen as it provides more transparency on the team’s individual effort and helps to track progress accurately. It is mostly suited for projects that have more research involved and are susceptible to change according to the findings. Project progress was made via a series of sprints. The duration of the sprint was set to 2 weeks as it helps to receive feedback faster and provides more opportunities to improve. Each Sprint consist of five activities: sprint planning, daily scrums, development work, sprint review, and sprint retrospective.

During the first meeting, the project requirements was listed out and translated them into product backlog including some R&D tasks. In the Sprint planning, the work was collaboratively planned to be performed during the sprint. Work was allocated to team members according to individual capacity and commitment. At the start of each sprint, a small set of user stories was taken from the product backlog to do R&D, code, and test functionality. Daily scrum was performed on Monday and Thursday of each week at 08:00 PM to get updates from the team members, track the progress of the sprint, to identify the blockers, and to check if any team member requires assistance. Sprint review was performed at the end of the sprint on Saturday at 10:00 am to demonstrate the work done in the sprint and to check if all of the tasks for the sprint are completed as promised. As members of the team were unfamiliar and were working together to develop this project, feedback was taken from all the team members during Sprint Retrospective discussing “What went right in the last sprint?”, “What went wrong in the last sprint?”, and “What could be improved in the next sprint?”. It was observed that sprint retrospective helped to improve team velocity and helped to make the product much better.



Backlogs

- Write final report
- Modify app based on small changes list
- submit WIL Project: Written Report 16 Oct

+ Add another card

Doing

- Include videoclip from the app into the final presentation
- Finish the video presentation

+ Add another card

Blocked

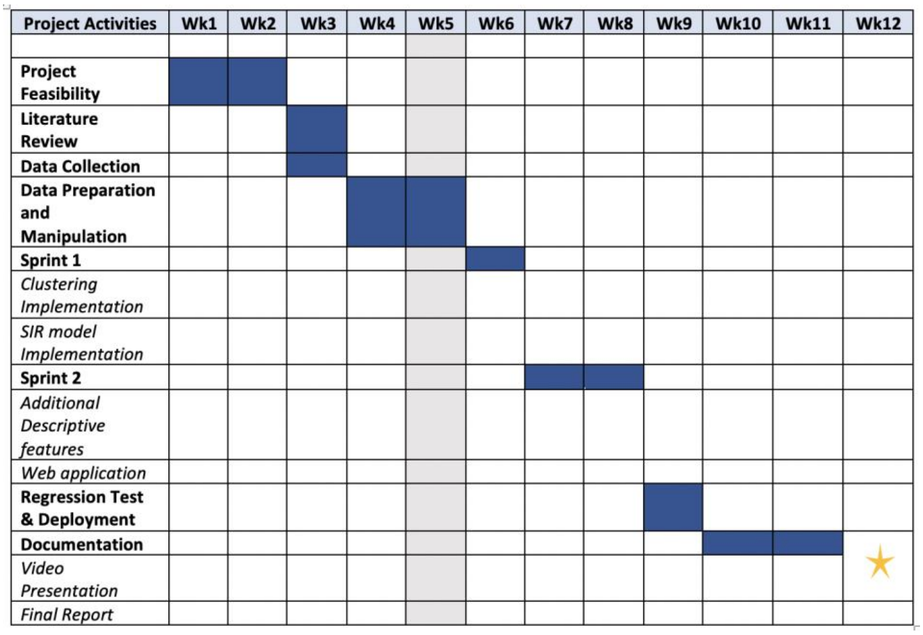
+ Add a card

Done

- Project Feasibility 17 Aug AP FP KT NT
- Literature Review 24 Aug AP FP KT NT
- Data Collection 24 Aug AP FP KT NT
- Data Preparation and Manipulation 7 Sep AP FP KT NT
- Develop Clustering 14 Sep AP FP KT NT
- Test Clustering 14 Sep AP FP KT NT

+ Add another card

Figure 5.1: Task distribution and monitoring using Trello



| **Project Activities** | **Wk1** | **Wk2** | **Wk3** | **Wk4** | **Wk5** | **Wk6** | **Wk7** | **Wk8** | **Wk9** | **Wk10** | **Wk11** | **Wk12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | |
| **Project Feasibility** | ■ | ■ | | | | | | | | | | |
| **Literature Review** | | | ■ | | | | | | | | | |
| **Data Collection** | | | ■ | | | | | | | | | |
| **Data Preparation and Manipulation** | | | | ■ | ■ | | | | | | | |
| **Sprint 1** | | | | | | ■ | | | | | | |
| *Clustering Implementation* | | | | | | | | | | | | |
| *SIR model Implementation* | | | | | | | | | | | | |
| **Sprint 2** | | | | | | | ■ | ■ | | | | |
| *Additional Descriptive features* | | | | | | | | | | | | |
| *Web application* | | | | | | | | | | | | |
| **Regression Test & Deployment** | | | | | | | | | ■ | | | |
| **Documentation** | | | | | | | | | | ■ | ■ | |
| *Video Presentation* | | | | | | | | | | | | ★ |
| *Final Report* | | | | | | | | | | | | |

Figure 5.2: High-level Project Timeline

- Project Feasibility: All the factors with respect to the project which mainly includes the technical, planning and legal considerations to ascertain the likelihood of completing the project successfully was discussed thoroughly.
- Literature Review: All the current substantial findings pertaining to the project was referenced to build the model.
- Data Collection: The data was collected from various sources for the analysis.
- Data Preparation and Manipulation: Feature engineering was carried out to proceed with Data Exploration.
- Sprint 1: Hierarchical Clustering was chosen for clustering the counties in terms of demographic features. These clusters were then the input feature for the SIR model to predict the susceptible-infected-recovered cases.
- Sprint 2: The web application to visualize the findings were built in the Dash Framework and tested.
- Regression Test & Deployment: The entire model and the application deployment was test in this phase.
- Documentation: All the findings as a part of the analysis was written in a form of a report.

## 6. Conclusion and Further Development

The project successfully created a working prototype of KRAAN, a web application that answers some challenges related to COVID-19 pandemic. KRAAN provides machine-learning backed information and interactive simulation to support the United States public health officials in developing an effective, data-driven vaccine deployment strategy in anticipation of imminent limited supply of vaccine when it is approved for safety and efficiency. With the automatically identified clusters, user can observe which cluster or combination of clusters will give the most desirable output on the overall country population from the SIR perspective, if a vaccine intervention introduced to those clusters.

Being a prototype, there are opportunities for further improvements including but not limited to the following:

1. Dataset enrichment to include more county-level population demographic attributes such as age range, employment sector, gender, pregnancy status (for female), existing medical conditions, etc. This additional data allows the clustering algorithm to produce finer granularity of clusters that more closely aligned to the general recommendation of vaccination target groups such as the one issued by Strategy Advisory Group of Experts (SAGE) during Influenza A (H1N1) 2009 pandemic.
2. Integrating other intervention efforts such as places temporary closure, travel restriction, curfew, etc. into the clustering and SIR simulation model.
3. Adoption of epidemiological compartment model that has better performance for COVID-19 pandemic based on latest research, such as the SEIR-QD and SEIR-PO [16]
4. Live data feeds to allow most up-to-date data insights.
5. Expand coverage to not only limited to United States but also for many other countries worldwide.

## References

[1] W. Feuer, "CDC director says there will likely be a limited supply of coronavirus vaccines at first", *CNBC*, 2020. [Online]. Available: https://www.cnbc.com/2020/08/28/cdc-director-says-there-will-likely-be-a-limited-supply-of-coronavirus-vaccines.html. [Accessed: 28 Sep 2020].

[2] T. Brown, "COVID-19 Vaccine Supply Will Be Limited at First, ACIP Says", *Medscape*, 2020. [Online]. Available: https://www.medscape.com/viewarticle/936428. [Accessed: 28 Sep 2020].

[3] X. Chen, M. Li, D. Simchi-Levi and T. Zhao, "Allocation of COVID-19 Vaccines Under Limited Supply", 2020.

[4] S. Scott and E. Clark, "Who will be first in line when a COVID-19 vaccine becomes available, and who will have to wait?", *Abc.net.au*, 2020. [Online]. Available: https://www.abc.net.au/news/2020-09-18/covid-19-vaccine-supply-in-australia-who-gets-it-first/12664888. [Accessed: 28 Sep 2020].

[5] "Vaccines for Pandemic Threats | History of Vaccines", *Historyofvaccines.org*, 2020. [Online]. Available: https://www.historyofvaccines.org/content/articles/vaccines-pandemic-threats. [Accessed: 28 Sep 2020].

[6] D. Fedson, "Pandemic Influenza and the Global Vaccine Supply", *Clinical Infectious Diseases*, vol. 36, no. 12, pp. 1552-1561, 2003. Available: 10.1086/375056 [Accessed 17 October 2020].

[7] *GUIDANCE ON Development and Implementation of a National Deployment and Vaccination Plan for Pandemic Influenza Vaccines*. World Health Organization, 2012. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/75246/9789241503990_eng.pdf [Accessed: 28 Sep 2020].

[8] “2009 swine flu pandemic", *En.wikipedia.org*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/2009_swine_flu_pandemic#Prevention. [Accessed: 28 Sep 2020].

[9] E. Callaway, "‘Make Ebola a thing of the past’: first vaccine against deadly virus approved", *Nature.com*, 2020. [Online]. Available: https://www.nature.com/articles/d41586-019-03490-8. [Accessed: 28 Sep 2020].

[10] *THE COMPLEX JOURNEY OF A VACCINE – PART II*. IFPMA, 2020. [Online]. Available: https://www.ifpma.org/wp-content/uploads/2016/01/IFPMA-ComplexJourney-FINAL-Digital.pdf. [Accessed: 28 Sep 2020].

[11] *Weekly epidemiological record*, 30th ed. World Health Organization, 2020, pp. 301-308. [Online]. Available: https://www.who.int/immunization/sage/SAGE_July_2009.pdf. [Accessed: 28 Sep 2020].

[12] "Q&A: Influenza and COVID-19 - similarities and differences", *Who.int*, 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-similarities-and-differences-covid-19-and-influenza?gclid=CjwKCAjwiaX8BRBZEiwAQQxGxySRoiCRWlOHkEl5I9wVCHCbR0Xpioc1-APxw7wH5PW-2P9c9FgEyBoCWnAQAvD_BwE. [Accessed: 28 Sep 2020].

[13] W. Yang, D. Zhang, L. Peng, C. Zhuge and L. Hong, "Rational evaluation of various epidemic models based on the COVID-19 data of China", 2020. Available: 10.1101/2020.03.12.20034595 [Accessed 3 August 2020].

[14] "Compartmental models in epidemiology", *En.wikipedia.org*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology. [Accessed: 28 Sep 2020].

[15] J. Collins and N. Abdelal, Spread of Disease. Australian Mathematical Science Institute, 2018.

[16] "A Simple Model for Vaccination - Dynamics of Susceptibles | Coursera", *Coursera*, 2020. [Online]. Available: https://www.coursera.org/lecture/developing-the-sir-model/a-simple-model-for-vaccination-rHPrN. [Accessed: 28 Sep 2020].

[17] "Kermack–McKendrick theory", *En.wikipedia.org*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Kermack%E2%80%93McKendrick_theory. [Accessed: 28 Sep 2020].

[18] C. Shalizi, *Distances between Clustering, Hierarchical Clustering*. Carnegie Mellon University, 2009. [Online]. Available: http://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf. [Accessed: 24 Aug 2020].

[19] "Coronavirus (COVID-19) Cases - Statistics and Research", *Our World in Data*, 2020. [Online]. Available: https://ourworldindata.org/covid-cases. [Accessed: 28- Sep- 2020].

[20] Immunization Basics | Vaccines and Immunizations | CDC", *Cdc.gov*, 2020. [Online]. Available: https://www.cdc.gov/vaccines/vac-gen/imz-basics.htm. [Accessed: 28- Sep- 2020].

[21] V. Zarikas, S. Poulopoulos, Z. Gareiou and E. Zervas, "Clustering analysis of countries using the COVID-19 cases dataset", *Data in Brief*, vol. 31, p. 105787, 2020. Available: 10.1016/j.dib.2020.105787 [Accessed 28 September 2020].

[22] E. Dong, H. Du and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time", *The Lancet Infectious Diseases*, vol. 20, no. 5, pp. 533-534, 2020. Available: 10.1016/s1473-3099(20)30120-1 [Accessed 28 September 2020].

[23] W. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods", *Journal of Classification*, vol. 1, no. 1, pp. 7-24, 1984. Available: 10.1007/bf01890115 [Accessed 28 September 2020].

[24] Pedregosa et. all, "Scikit-learn: Machine Learning in Python", *Scikit-learn.org*, 2020. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html. [Accessed: 28- Sep- 2020].