K-Means Project Midterm Report

CSCI 3302 - Machine Learning & Knowledge Discovery

Joshua Kyle Aceret, Alyssa Lawton
Computer Science
Hawaii Pacific University

I. INTRODUCTION

The K-Means Clustering algorithm is a type of unsupervised learning which can be applied to the study of viruses.

[1] proposed an algorithm that is based on the basic k-mean clustering algorithm. The k-mean clustering algorithm is traditionally used on numeric data. The authors have tweaked the algorithm to allow for mixed-numeric data and categorical features. These include new cost functions and distances measures based on co-occurrence of values. The measures also look at the significance of an attribute in the clustering process. They have tested this improved algorithm against other clustering algorithms to measure its effectiveness. Some of the algorithms the authors looked at included CACTUS, BIRCH, and CURE. One data set that was used was a film data set that included a film's director, genre, and famous actor. [1] In our implementation, we may want to adapt this algorithm as a comparison to the standard K-Means Clustering algorithm and determine if it suits our data better.

Due to the frequent use and wide access of social networking platforms, [2] detected diseases such as dengue or flu based on social media posts. Overall, responses and updates are received much sooner due to Online Social Network Sites (OSNSs). Naive Bayes and Support Vector Machine (SVM) were the two supervised learning classification algorithms used to detect flu outbreaks in tweets. Unsupervised learning techniques were effective since the data was unlabeled and the predicted outcome was unknown at the time. The K-Means Clustering algorithm was used for spatial analysis which proved to be more efficient because it utilizes distance-based measure to assess similarity between data sets. An elbow test was implemented to validate the number of clusters created. In the end, regions infected with Dengue were graphically represented based on the K-Means Clustering. In our project, K-Means Clustering can possibly serve to find trends within separate data-sets, linking them to find similarities. We might want to look at different figures that are available and compare them.

Although the K-Means Clustering algorithm has not been directly applied to the number of cases, it has been implemented to determine what caused patients to suffer severe COVID-19, single-cell RNA-seq using peripheral blood mononuclear cells (PBMCs) were obtained from donors who were diagnosed with mild or severe COVID-19 and severe influenza. This study concluded that type I IFN-driven inflammatory response is significant to the severity of COVID-19. Clustering was based on relative gene expression changes against a healthy donor group. The k-mean clustering algorithm followed the specific regulated gene expression patterns across all cell types among PBMCs. Besides this application, k-mean was used in the analysis of monocytes (type of white blood cell) in a comparison between mild COVID-19 and severe COVID 19 and influenza. [3]

Preparation for any upcoming pandemics on a national-level is critical in the case of a global epidemic and pandemic. This study wanted to determine what other essential capacities affect the current assessments and improve overall preparedness. An Epidemic Preparedness Index (EPI) was created to measure the economic resources of countries. K-Means Clustering algorithm helped evaluate the EPI by grouping the 188 countries into five clusters. These clusters were then validated by testing the index against detection and response outcomes during past pandemics and epidemics. Specifically, the 2009 H1N1 influenza pandemic was used to test the outbreak response. Other topics such as, the timeliness of outbreak detection and reporting along with the vaccination rate were events tested as well. Similarly to how the K-Means Clustering algorithm was applied to the 2009 H1N1 influenza pandemic in [4], we may be interested in determining the effect that the COVID-19 pandemic has had an effect on the Epidemic Preparedness Index (EPI).

To predict the hygiene rate of hospitals during COVID-19 on Nursing home data sets, K-Means Clustering algorithm was paired with three classification algorithms. K-Means Clustering performed better with the Naive Bayes algorithm than the random forest and decision tree algorithms. The Elbow method and silhouette score method were applied to validate the K-value selected that was dependent on the attributes, Passed Quality Assurance Test (PQAT) and Weekly Personal Protection Equipment (WTPPE). In conclusion, it was evident that the accuracy value of 98.1% from the K-mean and Naive Bayes algorithm occurred due to the decrease in cases amongst the hospital staff when personal protective equipment (PPE) was supplied to the hospitals and passed the personal quality test. In our project, we could possibly pair the K-Means Clustering algorithm with either the random forest, decision tree, or Naive Bayes algorithm to see if we can obtain an improved result as done in [5].

In [6], the goal was to find the closest relatives of the new H1N1 virus from public data available to them from databases. Influenza A virus comes in many forms. However, when two viruses were to infect a cell, a child-virus can be made with a reassortment of genes from the two viruses. It is assumed that reassortment has become more common due to swine and poultry farms with close contact. They authors use K-Means Clustering as a way to cluster the strains into sub groups. The unsupervised methods provide a way of identifying clusters without relying on other information. They also provide some extra material at the end of the article. We looked at this article as it was a possible idea to attempt. It clustered substrains of the flu, which could possibly be applied to Corona Virus and its possible strains.

In order to find a relation between food-group and prediction of Coronary Heart Disease and Stroke, [7] conducted a study with 40,011 men and women. Food items were consolidated into 31 food-groups. After 13 years, 1,843 CHD and 588 stroke cases were documented. Consumption was also split between prudent and western style consumption. They used both Principal Component Analysis and K-means Cluster Analysis. They found that a more prudent style of eating (high intakes of fish, high-fiber products, raw vegetables, and wine and low consumptions of potatoes) correlated with a decreased risk of CHD and stroke compared to a western style of eating. [7] showed that K-Means Clustering, could be applied to a health related dataset. Our implementation could look into incorporating PCA, which might be difficult, regarding the recency COVID-19 and data that is publically available.

Another application of implementing the K-Means Clustering algorithm was in a new screening method for influenza patients that is non-contact. [8] describes that the screening can occur within tens of seconds using non-contact monitored parameters. They developed a non-linear screening method which uses a neural network and K-Means Clustering. The authors chose to use SOM or Kohonen's self-organizing map combined with K-Means Clustering to discriminate between those with and those without the influenza virus. This is all done without prior medical and personal background (unsupervised). 92 test subjects were tested, 57 subjects were hospitalized with influenza and 35 were normal control subjects. They claim that this appears promising for future screening techniques for unknown/unexpected infectious diseases. [8] concludes that K-Means Clustering performed better to determine if a person had influenza. This is a plus for our implementation is it could mean that K-Means Clustering can be used for screening.

[9] is a tutorial and introduction to the K-Means Clustering Algorithm. It gives an explanation as to the use of K-Means Clustering and provides a sample data set to use for explanation. There is a flowchart that helps to visualize the process of K-Means Clustering. The tutorial provides an explanation for each step of the algorithm and provides example calculations on the data set. Towards the end of the tutorial the authors provide some answers to questions regarding K-Means Algorithm. They also included some code for the algorithm.

Influenza forecasting is conducted in the US to improve planning for our current health systems and better the behavior of its citizens who may be at risk. Seasonal influenza are deemed as ideal cases studies for a pandemic influenza and plan the course of action for the future. Data was collected from the point-of-care (POC) diagnostic machines over three seasons. The K-Means Clustering algorithm was applied to 136 qualifying counties that were filtered out by a certain criteria. As a result, the forecasting of influenza A positive POC data was more accurate when grouped in clusters rather than applying the forecast to individual counties. As it was determined in [10] instead of applying the K-Means clustering algorithm to individual islands, this helped us determine that our data would be more accurate by applying the algorithm to the entire state's population.

[11] uses the K-Means Clustering to determine the likelihood that a patient with symptoms similar to hepatitis B and hepatitis C, actually has it. These strains of hepatitis were selected as they were likely to cause liver cancer. Liver cancer is the seventh most common cancer and third most deadly. They used data points with 7 features including: sex, SGOT, SGPT, HbsAg, Anti-HCV, Ureum and Creatinin. They showcase some code and equations they used to make the program. The highest success rate of the program was roughly 84.85%. The program also was pretty fast with around 0.08 seconds. The authors note that K-Means Clustering could be used to predict other health conditions. [11] is an article in which its idea is one that could be implemented. Our implementation could do something similar in an effort to determine the likelihood of someone having COVID-19 based on features like symptoms.

In efforts to implement a simplistic version of Lloyd's K-Means Clustering algorithm which is referred to as the filtering algorithm which is a popular heuristic for K-Means Clustering. Lloyd's algorithm improves the final distortion at the post processing stage in statistical analysis. The authors' filtering algorithm differs due to the values selected to bisect the hyperplane which helped select candidates closer to a cell's midpoint. This overall became efficient because data points did not vary throughout computations so the data structures were not required to be recalculated at each step. In the theoretical data sensitive analysis, the results concluded that the more separated a cluster was, the running time of the algorithm would shorten. The empirical analysis through experiments done on both synthetically generated and real data sets proved the theoretical conclusion. The implementation in [12] of Lloyd's K-Means Clustering might be an algorithm we may be interested in implementing as a comparison to the generic K-Means Clustering algorithm.

II. PROBLEM

Coronavirus Pandemic (COVID-19) results in an unknown outcome of cases. In light of any pandemic, it is optimal to determine what is the leading cause to the spread of a virus. Hawaii's economy heavily relies on tourism which was highly impacted by the statewide lockdown due to COVID-19 in March of 2020. Provided statistics on the number of COVID-19 cases and the number of visitors, we hope to determine whether or not tourism has an impact on the number of COVID-19 cases in Hawaii. By clustering the groups of data based on external information collected by the State of Hawaii, we plan to use the K-Means Clustering Algorithm to extract any possible information in regards to the Hawaii COVID-19 dataset to help take preventive measures in the future.

III. DATA SETS

- Hawaii Data Collaborative COVID Dashboard Data:
  - https://public.tableau.com/views/RiskFactorCountyTravelprod11242020/BarsDash?%3Adisplay_count=y&%3Aorigin=viz_share_link&%3Aembed=y&%3AshowVizHome=no
- Hawaii Daily Visitor Statistics:
  - https://www.hawaiitourismauthority.org/covid-19-updates/trans-pacific-passenger-arrivals/

IV. ALGORITHM OVERVIEW

In our machine learning course, we learned about the K-Nearest Neighbors algorithm that gives us new data points in respect to the k number of closest neighbors points. The K-Means Clustering algorithm varies because it is unsupervised learning which is applied to unlabeled data points. The anticipated result of the algorithm is grouping the data into K number of clusters which share a similarity in features. As new data is added to the set, this data point will be grouped in one of the generated clusters based on its respective features. In general, the K-Means Clustering algorithm can be used to find trends in data and classify new data points accordingly. This is why we were inclined to apply this algorithm on the effect of travel on the number of COVID cases in Hawaii.

V. FINAL RESULT ANALYSIS

We implemented our data using MATLAB's K-Means Clustering algorithm to produce cluster centroids. From our data, we are able to predict the number of expected COVID-19 cases that may occur due to the number of incoming trans-pacific passenger arrivals. Evidently, there is no linear relationship amongst the number of passenger arrivals and the number of travel-associated COVID-19 cases.

To determine the number of clusters we wanted to implement, MATLAB provided a suggested comparison of silhouette values. By plotting potential cluster values ranging from 2 to 30, we were able to determine the optimal number of clusters that would represent each data set. The silhouette values determine how close each point is in respect to distant neighboring clusters.

TABLE I: STATE DATE RANGES

Cluster	Ranges	
1	3/27/20 - 10/14/20	10/30/20
	10/20/20	1/26/21
	10/25/20 - 10/28/20	---
2	12/19/20	3/4/21
	12/26/20	3/10/21 - 3/15/21
	2/11/21 - 2/13/21	3/17/21 - 4/30/21
3	10/15/20 - 10/19/20	12/27/20 - 1/25/21
	10/21/20 - 10/24/20	1/27/21 - 2/10/21
	10/29/20	2/14/21 - 3/3/21
	10/31/20 - 12/18/20	3/5/21 - 3/9/21
	12/20/20 - 12/25/20	3/16/21

TABLE II. STATE CENTROID VALUES

Cluster	Cluster Centroid		
	Passenger Arrivals	**COVID-19 Cases**	**% Travel COVID-19 Cases**
1	1,725	2	10.74%

2	18,971	4	7.85%
3	8,997	9	12.16%

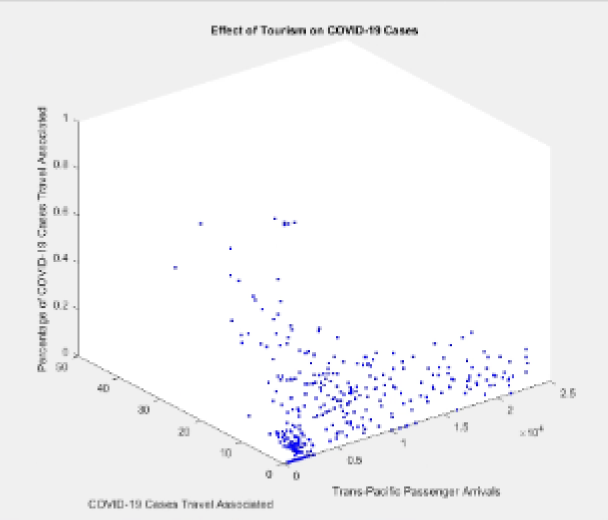

Fig. 1. Plot of all data points, presenting the relationship amongst the number of trans-Pacific passenger arrivals and number of new covid cases in the State of Hawaii.

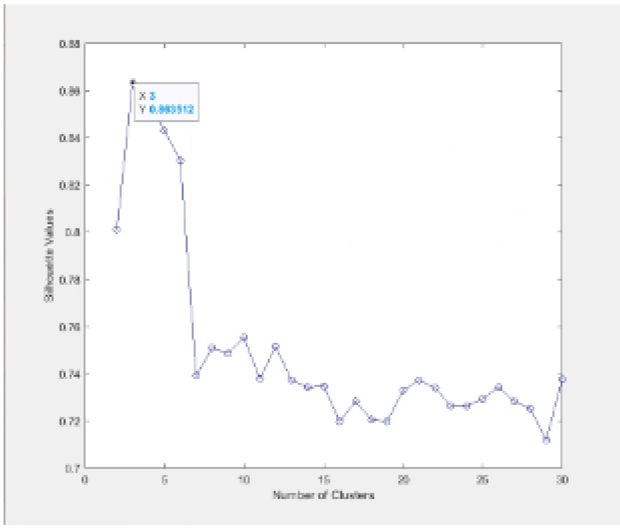

Fig. 2. Plot of silhouette values based on clusters ranging from 2 to 30 for the State of Hawaii.

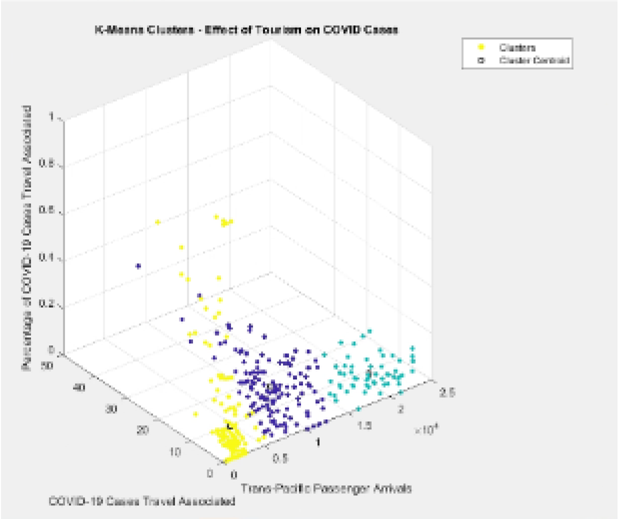

Fig. 3. Plot of K-Means Clustering produced by MATLAB's k-means algorithm for the State of Hawaii

The first cluster ranges from March 2020 to October 2020 which includes the Stay-At-Home period from late March 2020 to the end of May in 2020. According to the cluster centroid value, the passenger arrivals have been lower as the number of travel cases have lessened but the percentage of travel-associated COVID-19 cases was the second highest. In comparison, the second cluster represents when the State of Hawaii decided to open up vaccines to the public where anyone 16 years and older, had the option to take the COVID-19 vaccine. The second cluster centroid accurately represents this event because the passenger arrivals are the highest while the number of travel-associated COVID-19 cases are neither the highest or lowest of the centroid values due to the accessibility of vaccinations. Although the number of travel-associated COVID-19 cases have been reduced, the percentage is the lowest of the three centroids which was unexpected. The percentage could have decreased due to external factors such as vaccinations occurring outside of the state of Hawaii. Lastly, our third centroid represents the period in which vaccines were beginning to be distributed to frontline workers, essential workers, and elders. The cluster centroid shows an increase in the number of passenger arrivals along with the highest number and percentage of travel-associated COVID-19 cases. This occurred because the number of flights increased during this time period and only people that met a certain criteria had access to the COVID-19 vaccine.

TABLE III. OAHU DATE RANGES

Cluster	Ranges	
1	3/27/20 - 8/1/20	10/27/20 - 10/28/20
	8/3/20	10/30/20
	8/5/20 - 10/14/20	1/26/21
	10/25/20	---
2	3/11/21 - 3/13/21	4/21/21 - 4/24/21
	3/17/21 - 4/12/21	4/26/21
	4/14/21 - 4/17/21	4/28/21 - 4/30/21
	4/19/21	---
3	11/21/20	3/1/21
	12/17/20 - 12/21/20	3/3/21 - 3/4/21
	12/23/20 - 12/24/20	3/8/21
	12/26/20 - 12/31/20	3/10/21
	1/2/21 - 1/4/21	3/14/21 - 3/16/21
	1/7/21	4/13/21
	2/10/21 - 2/13/21	4/18/21
	2/15/21	4/20/21
	2/18/21 - 2/20/21	4/25/21
	2/24/21 - 2/25/21	4/27/21
	2/27/21	---
4	8/2/20	1/8/21 - 1/25/21
	8/4/20	1/27/21 - 2/9/21
	10/15/20 - 10/24/20	2/14/21
	10/26/20	2/16/21 - 2/17/21
	10/29/20	2/21/21 - 2/23/21
	10/31/20 - 11/20/20	2/26/21
	11/22/20 - 12/16/20	2/28/21
	12/22/20	3/2/21
	12/25/20	3/5/21 - 3/7/21
	1/1/21	3/9/21
	1/5/21 - 1/6/21	---

TABLE IV. OAHU CENTROID VALUES

Cluster	Cluster Centroid		
	Passenger Arrivals	COVID-19 Cases	% Travel COVID-19 Cases
1	109	0	15.14%
2	1,847	2	31.12%
3	2,807	1	18.10%
4	1,031	3	29.57%

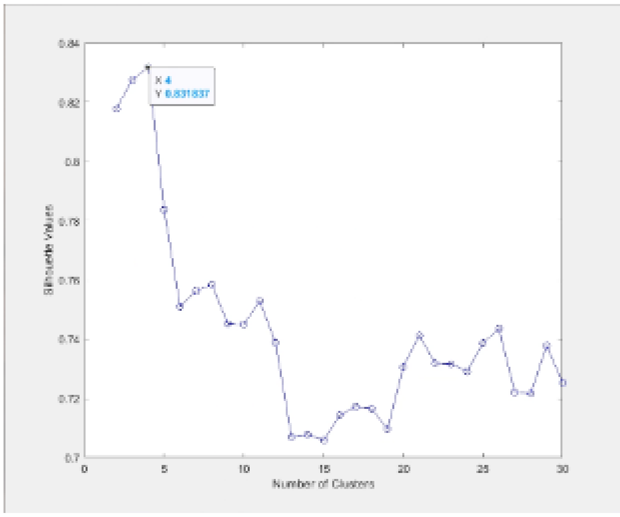

Fig. 4. Plot of silhouette values based on clusters ranging from 2 to 30 for the island of Oahu.

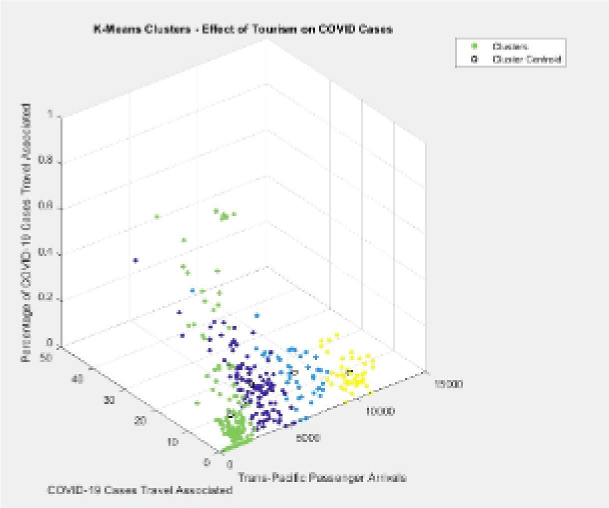

Fig. 5. Plot of K-Means Clustering produced by MATLAB's k-means algorithm for the island of Oahu

For the island of Oahu, the stay-at-home mandate was respresnted by the cluster centroid 1. During this time period, the numbers and percentage of arrivals and travel-associated COVID-19 cases were at their lowest. This was due to the restrictions and mandates enforced at the time. When vaccinations were starting to be administered, this time period was represented by cluster centroid 4. Similar to the State of Hawaii, this is when the most travel-associated COVID-19 cases occurred. Cluster centroid 2, represents the most recent phases in which vaccinations are more open to the general public. The centroids value shows an increase in the number of passenger arrivals along with an increase in the number and percentage of travel-associated COVID-19 cases. This is anticipated because the rate at which the community is being vaccinated is increasing.

TABLE V. HAWAII DATE RANGES

Cluster	**Ranges**	
1	3/27/20 - 10/14/20	10/25/20 - 10/28/20
	10/19/20 - 10/21/20	---
2	3/11/21 - 3/13/21	4/21/21 - 4/24/21
	3/17/21 - 4/12/21	4/26/21
	4/14/21 - 4/17/21	4/28/21 - 4/30/21
	4/19/21	---
3	11/21/20	3/1/21
	12/17/20 - 12/21/20	3/3/21 - 3/4/21
	12/23/20 - 12/24/20	3/8/21
	12/26/20 - 12/31/20	3/10/21
	1/2/21 - 1/4/21	3/14/21 - 3/16/21
	1/7/21	4/13/21
	2/10/21 - 2/13/21	4/18/21
	2/15/21	4/20/21
	2/18/21 - 2/20/21	4/25/21
	2/24/21 - 2/25/21	4/27/21
	2/27/21	---
4	8/2/20	1/8/21 - 1/25/21
	8/4/20	1/27/21 - 2/9/21
	10/15/20 - 10/24/20	2/14/21
	10/26/20	2/16/21 - 2/17/21
	10/29/20	2/21/21 - 2/23/21
	10/31/20 - 11/20/20	2/26/21
	11/22/20 - 12/16/20	2/28/21
	12/22/20	3/2/21
	12/25/20	3/5/21 - 3/7/21
	1/1/21	3/9/21
	1/5/21 - 1/6/21	---

TABLE VI. HAWAII CENTROID VALUES

Cluster	Cluster Centroid		
	Passenger Arrivals	COVID-19 Cases	% Travel COVID-19 Cases
1	1,390	2	10.91%
2	10,322	4	7.15%
3	7,127	7	11.62%
4	4,315	8	11.73%

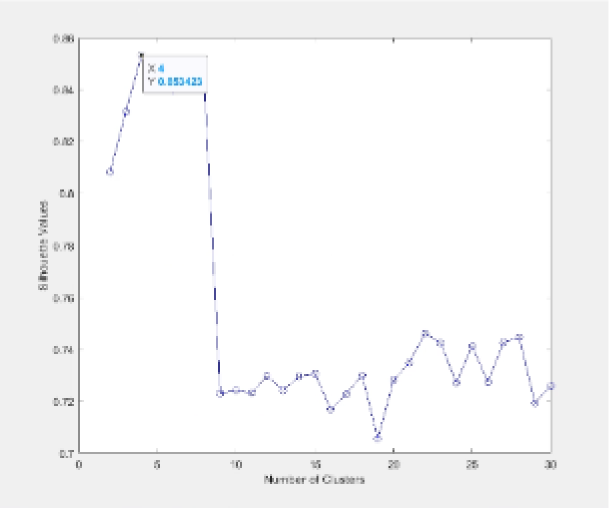

Fig. 6. Plot of silhouette values based on clusters ranging from 2 to 30 for the island of Hawaii.

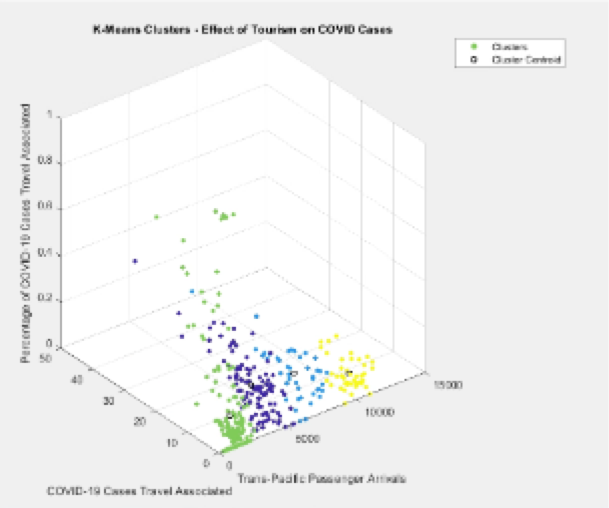

Fig. 7. Plot of K-Means Clustering produced by MATLAB's k-means algorithm for the island of Hawaii.

For the island of Hawaii, cluster centroid 1 represented the stay-at-home mandate. In comparison to the overall state, the percentage was not the lowest of all the cluster centroid but the number of passenger arrivals and travel-associated COVID-19 cases were the smallest of their features. This could have occurred due to the ratio of people diagnosed with COVID-19 were more likely to have travel history. Since the outer islands have less international and mainland flights, which meant less community spread. Cluster centroid 2 represents phase 1a and phase 2 of the vaccination release plan. This represents the largest influx of passenger arrivals but the second lowest amount of travel-associated COVID-19 cases and the lowest percentage of COVID-19 cases. This is evident since the vaccines were distributed. We are perplexed unto why the second cluster centroid represents the initial release of vaccines. Cluster centroids 4 and 3 represent the intermediate phases of the vaccination release plan. This is when the number of COVID-19 cases are the highest probably due to the limited availability of vaccines to the general population. Since the number of passenger arrivals increased, this affected the number and percentage of travel-associated COVID-19 cases.

TABLE VII. MAUI DATE RANGES

Cluster	Ranges	
1	3/27/20 - 10/14/20	---
2	12/26/20	4/8/21 - 4/9/21
	3/12/21	4/15/21 - 4/16/21
	3/24/21 - 3/25/21	4/22/21
	3/29/21	4/29/21
	3/31/21	---
3	12/28/20 - 12/30/20	2/21/21 - 2/22/21
	1/2/21	2/24/21 - 2/26/21
	1/16/21	2/28/21
	2/5/21	3/1/21
	2/15/21	3/5/21
	2/17/21 - 2/18/21	---
4	11/21/20	3/3/21 - 3/4/21
	12/18/20	3/6/21

	12/20/20	3/8/21
	12/24/20 - 12/25/20	3/10/21
	12/27/20	3/15/21 - 3/16/21
	2/6/21	3/23/21
	2/10/21 - 2/11/21	4/13/21
	2/14/21	4/20/21
	2/19/21 - 2/20/21	4/25/21 - 4/27/21
	2/27/21	---
5	11/6/20 - 11/7/20	1/13/21
	11/13/20 - 11/14/20	1/17/21
	11/18/20 - 11/19/20	1/20/21 - 1/22/21
	11/25/20 - 12/4/20	1/27/21 - 1/29/21
	12/10/20 - 12/11/20	2/7/21
	12/13/20	2/23/21
	1/4/21 - 1/6/21	3/7/21
	1/10/21	---
6	12/19/20	4/11/21 - 4/12/21
	2/12/21 - 2/13/21	4/14/21
	3/11/21	4/18/21 - 4/19/21
	3/14/21	4/21/21
	3/17/21 - 3/18/21	4/23/21
	3/22/21	4/28/21
	3/30/21	4/30/21
	4/6/21 - 4/7/21	---
7	10/18/20 - 10/23/20	12/8/20
	10/25/20 - 10/31/20	1/19/21
	11/2/20 - 11/3/20	1/26/21
8	3/20/21	4/2/21 - 4/3/21
	3/27/21	4/10/21
9	3/13/21	4/1/21
	3/19/21	4/4/21 - 4/5/21
	3/21/21	4/17/21
	3/26/21	4/24/21
	3/28/21	---
10	10/15/20 - 10/17/20	12/6/20 - 12/7/20
	10/24/20	12/9/20
	11/1/20	12/14/20 - 12/15/20
	11/4/20 - 11/5/20	1/11/21 - 1/12/21
	11/8/20 - 11/12/20	1/18/21
	11/15/20 - 11/17/20	1/24/21 - 1/25/21
	11/24/20	1/31/21
	11/29/20	2/2/21
11	11/23/20	1/23/21
	12/5/20	1/30/21
	12/16/20	2/1/21
	12/22/20	2/3/21 - 2/4/21
	12/31/20 - 1/1/21	2/8/21 - 2/9/21
	1/3/21	2/16/21
	1/7/21 - 1/9/21	3/2/21
	1/14/21 - 1/15/21	3/9/21

TABLE VIII. MAUI CENTROID VALUES

Cluster	Cluster Centroid		
	Passenger Arrivals	COVID-19 Cases	% Travel COVID-19 Cases
1	115	0	12.47%
2	6,374	1	5.41%
3	3,379	2	13.56%
4	4,241	1	8.17%
5	2,293	2	13.41%
6	5,577	1	9.57%
7	1,210	1	8.65%
8	8,122	3	11.78%
9	7,173	1	3.52%
10	1,820	2	28.18%
11	2,835	2	9.68%

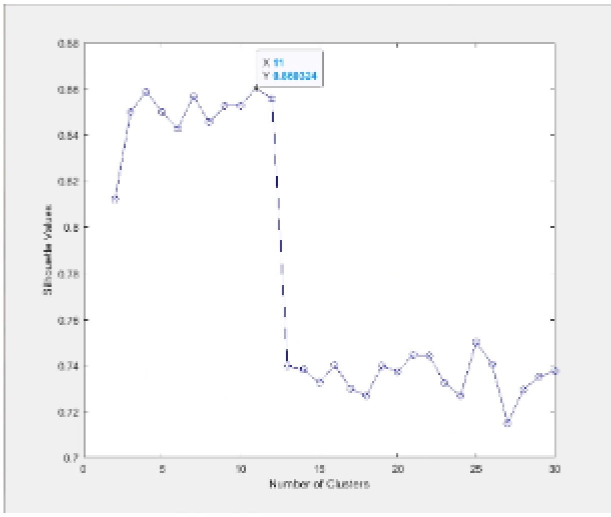

Fig. 8. Plot of silhouette values based on clusters ranging from 2 to 30 for the island of Maui.

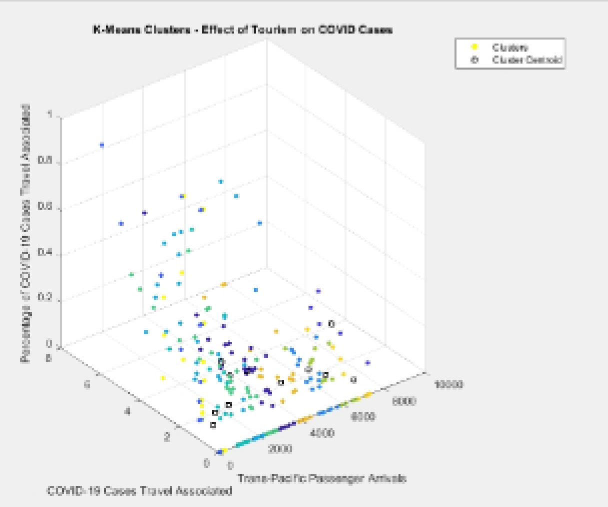

Fig. 9. Plot of K-Means Clustering produced by MATLAB's k-means algorithm for the island of Maui.

In comparison to the other sets of data, Maui had the most number of suggested clusters. This number of clusters may be overrepresenting the entire data set which in return makes the values too specific. Cluster centroid 1 has the lowest amount of passenger arrivals and travel-associated COVID-19 cases. Similarly, to the island of Hawaii, the percentage of COVID-19 cases was not the lowest since it is one of the outer islands not receiving out of state flights. Cluster centroid 3 represents the initial release of early phase of the vaccines. Overall, the COVID-19 cases for the island of Maui are on the lower end in comparison to neighboring islands. The number of passenger arrivals along with the percentage in travel related COVID-19 cases have both increased. This is probably due to the lifting of restrictions. In the next phase 1b of the vaccination release plan, the cluster centroid 11 represented this date range. Interestingly enough, there was a reduction in the number of passenger arrivals and percentage of COVID-19 cases. In the more recent phases of the vaccination release, the sixth cluster centroid shows an increase in passenger arrivals and slight decrease in percentage of travel related COVID-19 cases. Once again, external factors may have contributed to this change in percentage.

TABLE IX. KAUAI DATE RANGES

Cluster	Ranges	
1	3/27/20 - 6/30/20	10/25/20 - 10/28/20
	7/2/20 - 10/14/20	11/29/20
	10/20/20 - 10/21/20	12/2/20 - 4/4/21
2	7/1/20	10/29/20 - 11/28/20
	10/15/20 - 10/19/20	11/30/20 - 12/1/20
	10/22/20 - 10/24/20	4/5/21 - 4/30/21

TABLE X. KAUAI CENTROID VALUES

Cluster	Cluster Centroid		
	Passenger Arrivals	COVID-19 Cases	% Travel COVID-19 Cases
1	73	0	15.31%
2	939	1	40.11%

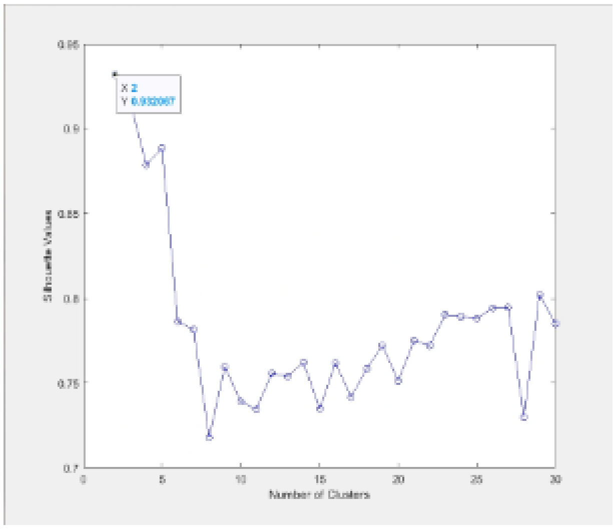

Fig. 10. Plot of silhouette values based on clusters ranging from 2 to 30 for the island of Kauai.

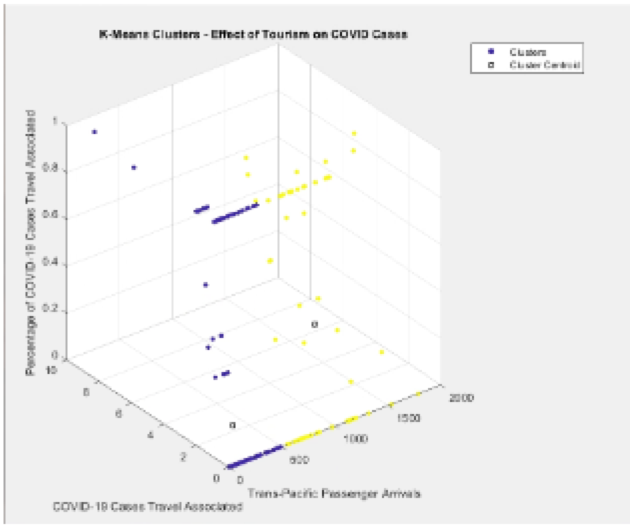

Fig. 11. Plot of K-Means Clustering produced by MATLAB's k-means algorithm for the island of Kauai.

Kauai's data set varied the most from its neighboring islands. When a new COVID-19 case occurred, it was almost always travel related. This sets the data points on the limits of the percentage of travel-associated COVID-19 cases. Up till late March of 2021, the data was represented with cluster centroid 1. The number of passenger arrivals and percentage in COVID-19 cases were the lowest and the number of COVID-19 cases is 0. This is most likely due to the additional restrictions enforced by the island itself. On the other hand, all three features increased in value for the last two phases in the vaccine distribution. This is expected because the number of vaccinations released to the public have increased, hence the community related COVID-19 cases have decreased. Since there are more passenger arrivals, it is more likely that there will be more travel related COVID-19 cases.

TABLE XI. MAJOR EVENT CLUSTERING

Events	STATE	OAHU	HAWAII	MAUI	KAUAI
Stay-At-Home Order (3/23-5/31)	1	1	1	1	1
1a - "Health Care Personnel LTC Facility Residents" (12/15)	3	4	2	3	1
1b - "75+ Frontline	3	4	4	11	1

Essential Workers" (12/22)					
1c - "65-74 16-64 with high-risk medical condition, essential workers not recommended for vaccination in 1b" (3/29)	2	2	3	6	2
2 - "Individuals 16+" (4/19)	2	2	2	6	2

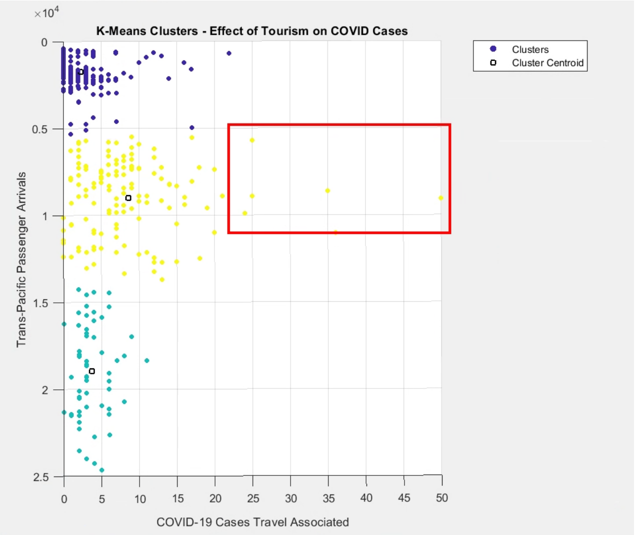

Fig. 12. Relationship between Passenger Arrivals and Travel Related COVID-19 Cases. Outliers depicted in red rectangle.

These outlier points outlined in the red rectangle occurred during the 7 to 14 day period after major holidays such as Christmas and New Year's which tells us that the COVID-19 cases associated with travel history are higher during the holiday season.

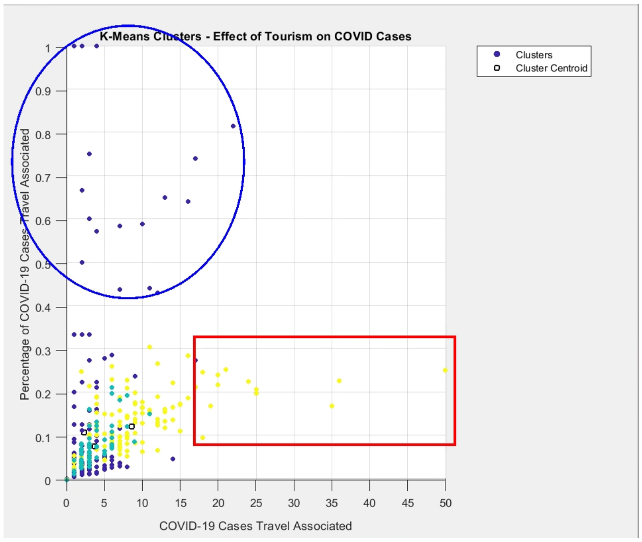

Fig. 13. Relationship between Percentage and Number of Travel Related COVID-19 Cases. Two sets of outliers depicted in blue ellipse and red rectangle.

When looking at the relationship between the percentage and number of COVID-19 cases associated with travel, we have identified two regions of outliers. The outliers that are located in the blue ellipse occur at the start of the COVID-19 pandemic, from March to May of 2020. This makes sense because COVID-19 did not originate from the State of Hawaii and was brought in through trans-pacific passenger arrivals.

The second set of outliers outlined in the red rectangle represent the cases that occurred during the 7 to 14 day period after major holidays. This includes Christmas and New Years, along with Veterans Day and Thanksgiving. Although there were a fair amount of COVID-19 cases associated with travel, all of these data points' percentages were below 30%. This is true for the majority of the overall data which means as a state, tourism does not have as much of an impact on the number of COVID-19 cases that occur in Hawaii.

VI. CONCLUSION

Our implementation of the K-Means Clustering algorithm helped us determine the relationship between tourism and COVID-19 cases in Hawaii. We were able to obtain an accurate representation of the effect of tourism on the state of Hawaii. Based on the clusters generated, we could conclude that tourism is not a major contributor to the overall number of COVID-19 cases that occur in the state. In contrast to this overall trend, tourism does have a large impact on the island of Kauai since the majority of cluster centroids had higher values in comparison. The K-Means Clustering algorithm had a tendency to group the clusters in relation to the number of Trans-Pacific passenger arrivals due to the shorter distance between points amongst this feature. If given the opportunity to continue this analysis, we would like to see the relationship between March to May of 2020 and March to May of 2021 to see if the vaccine creates any outlier points similar to the early dates during the pandemic.

REFERENCES

[1] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data & Knowledge Engineering*, vol. 63, no. 2, pp. 503–527, Nov. 2007.

[2] S. Amin, M. I. Uddin, M. A. Zeb, A. A. Alarood, M. Mahmoud, and M. H. Alkinani, "Detecting Dengue/Flu Infections Based on Tweets Using LSTM and Word Embedding," *IEEE Access*, vol. 8, pp. 189054–189068, Oct. 2020.

[3] J. S. Lee, S. Park, H. W. Jeong, J. Y. Ahn, S. J. Choi, H. Lee, B. Choi, S. K. Nam, M. Sa, J.-S. Kwon, S. J. Jeong, H. K. Lee, S. H. Park, S.-H. Park, J. Y. Choi, S.-H. Kim, I. Jung, and E.-C. Shin, "Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19," *Science Immunology*, vol. 5, no. 49, pp. 1–16, Jul. 2020.

[4] B. Oppenheim, M. Gallivan, N. K. Madhav, N. Brown, V. Serhiyenko, N. D. Wolfe, and P. Ayscue, "Assessing global preparedness for the next pandemic: development and application of an Epidemic Preparedness Index," *BMJ Global Health*, vol. 4, no. 1, pp. 1–9, Jan. 2019.

[5] A. M. Qahtani, B. M., H. Alhakami, S. Abuayeid, and A. Baz, "Predicting Hospitals Hygiene Rate during COVID-19 Pandemic," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, pp. 815–823, 2020.

[6] A. Solovyov, G. Palacios, T. Briese, W. I. Lipkin, and R. Rabadan, "Cluster analysis of the origins of the new influenza A(H1N1) virus," *Eurosurveillance*, vol. 14, no. 21, May 2009.

[7] M. Stricker, N. Onland-Moret, J. Boer, Y. V. D. Schouw, W. Verschuren, A. May, P. Peeters, and J. Beulens, "Dietary patterns derived from principal component- and k-means cluster analysis: Long-term association with coronary heart disease and stroke," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 23, no. 3, pp. 250–256, Mar. 2013.

[8] G. Sun, Y. Hakozaki, S. Abe, N. Q. Vinh, and T. Matsui, "A novel infection screening method using a neural network and k-means clustering algorithm which can be applied for screening of unknown or unexpected infectious diseases," *Journal of Infection*, vol. 65, no. 6, pp. 591–592, Dec. 2012.

[9] K. Teknomo, "K-Mean Clustering Tutorials," *K-Mean Clustering Tutorial*, 2019. [Online]. Available: https://people.revoledu.com/kardi/tutorial/kMean/index.html. [Accessed: 12-Mar-2021].

[10] J. Turtle, P. Riley, M. Ben-Nun, and S. Riley, "Accurate influenza forecasts using type-specific incidence data for small geographical units," *medRxiv*, Nov. 2019.

[11] G. Kurniawan and Z. Rustam, "Enhancement of hepatitis virus outcome predictions with application of K-means clustering," *Proceedings of the 4th International Symposium on Current Progress in Mathematics and Sciences (ISCPMS2018)*, vol. 2168, no. 1, pp. 1–5, Nov. 2019.

[12] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, Jul. 2002.