Opioid Addiction Crisis in Virginia

Xinzuo Wang
xw3xp
xw3xp@virginia.edu

Qian Qu
qq8jn
qq8jn@virginia.edu

Onyi Uche
ocu2t
ocu2t@virginia.edu

I. Abstract

This project examines the opioid addiction crisis in the state of Virginia through related indicators such as Overdose Deaths, ED Visits, Hepatitis C and Diagnosed HIV. By examining the occurrence, we visualized the Virginia Opioid dataset and discovered correlations to opioid addiction in the state of Virginia. These correlations include age group, case count, rate and year for each related opioid addiction indicator. In this project, we use two major strategies: (1) Analyzing and Visualizing the data-set to discover and understand patterns and re-occurrences for the case being studied and (2) Carrying out Machine Learning Techniques in order to find hidden structures and anomalies on data. This projects utilizes machine learning techniques such as Isolation Forest[1], K-Means[2], Mean Shift[3] and Dimensionality reduction techniques such as Principal component analysis(PCA)[4] to better represent and understand the data-set being worked on.

II. Motivation

Virginia is currently under a public health emergency as a result of the opioid addiction crisis. Our state has been severely impacted by opioid abuse, and the situation has been escalating surprisingly fast in recent years. In 1999, the first year for which such data is available, approximately 23 people died from abuse of prescription opioids. However, by 2017, the most recent year for which complete data is available, 1445 people died of the overdose of Fentanyl, Heroin, or Prescription Opioid, an epidemic increase of roughly 6200%. [5] Moreover, the data of 2015-2017 alone, showed us that there was a staggering increase of more than 60% in death attributed to drug overdose. Drug-related deaths have risen unrelentingly, and the drugs kill more people annually in Virginia than either car crashes or gunfire. This situation needs to be controlled, and we should contribute our efforts as community members.

Our project is intended to study the drug abuse problem here in Virginia using machine learning methods, and bring awareness to this widespread public health issue by presenting a thorough analysis of the problem, and possible solutions for different parties to reverse the epidemic of opioid drug overdose deaths and protect the public from overdose and other harms.

III. Data-set

The data-set we are tackling is the Virginia Opioid Dashboard Dataset by Age Group: http://www.vdh.virginia.gov/content/uploads/sites/110/2018/11/Opioid-Dashboard-Dataset-View-Age-Groups.xlsx

This dataset contains the Opioid Overdose records in Virginia from 2011 to 2017 (more than 40,000 rows). The records contains information such as overdose types, geographical features, mortality and ED-visit rates, and age groups.

IV. Related work

For anomaly detection, [1] proposed a tree-based ensemble method that explicitly isolates anomalies instead of profiles normal points. Isolation Forest works well in high dimensional data which has lots of irrelevant attributes.

With regard to feature reduction for high-dimensional data, the essential problem is how to map the high dimensional data into low dimensions without losing important information. A popular method for exploring high-dimensional data is something called PCA[4]. The technique has become widespread in the field of machine learning since it maintains most part of the variance of the original data and thus remains most of the important information. During our study on prior research, such methods are widely used in analyzing problems relating to regional health/crime issue, such as the analysis on U.S. Opiate Prescriptions (https://www.kaggle.com/greenmaverick/exploratory-analysis-on-opioid-prescriptions), and Crime in Chicago (https://www.kaggle.com/fahd09/eda-of-crime-in-chicago-2005-2016).

For clustering problems, what defines a good cluster algorithm depend on applications and various criteria. Among these algorithms, K-Means[2] and Mean Shift[3] are two popular unsupervised algorithms widely used in the machine learning area. Apart from this, there exist many other good algorithms including approaches based on splitting and merging such as ISODATA[6], and methods based on neural nets[7].

V. Method

We used Isolation Forest[1] to do anomaly detection. It works on the natural fact that outliers are 'few and different' in any data set, which is quite different from the typical clustering based or distance based algorithm. We use this method for anomaly detection for two reasons: 1.It is hard to do supervised anomaly detection on our unlabeled data set. 2. As an ensemble method, it is powerful and works well on high dimensional data which distance based method may have problems. We used PCA to do feature reduction so that we can handle the curse of dimensionality. We used K-Means and Mean shift to help find the groups of our data. K-Means is a fairly simple EM algorithms and computes very fast.

The disadvantage of K-Means is that it is highly vulnerable by the initial center positions. The Mean shift, however, are less influenced by initial centers. But it is expensive and time consuming.

VI. Intended Experiments

During the experiments, we will first look into the data to get the basic sense of the data set. Then we will do anomaly detection to see if there exist some red cases in our data, which means these areas appear different from other spaces, we should pay attention to the potential risk of these areas.
Later, we will use coloring to evaluate how well the map preserves the similarity nodes within the same class. Meaning that the nodes with similar property would get close to each other and have the same color. Also, we will use the dimensionality reduction techniques to map the high-dimensional representation to a two-dimensional or three-dimensional space and show as a scatter-plot format. For better visualization, we will deploy our results on the map of Virginia.

VII. Experiments and Results

1. Data Visualization

After preprocessing and cleaning the data, we produced multiple statistical graphics to visualize the data and took a deeper look into how features can influence the results and the pattern of the current opioid addiction situation in Virginia. The tools we used for data visualization includes Matplotlib, Seaborn, and geospatial data visualizing tool Plotly.
The dataset we chose contains very different features. Besides the opioid addiction types and case count/rate of hundreds of countries in Virginia, One major part of the dataset is statistical data of several regional indexes which is formed by multiple counties. To have a clearer view of the opioid addiction crisis in Virginia, we need to visualize the history of the crisis and show how the situation developed in different parts of Virginia in the past seven years. To achieve that, we need to employ multiple visualization methods, including linear regression and grouping categorizing, visualizing univariate and multivariate distributions (e.g., grouped plots, stacked plots), and geospatial data map.

2. Anomaly detection

We first do some feature engineering on the dataset. We noticed that some data is written as '*' meaning any number between 1 to 4. We handle this non-numerical data by add a new column and add a binary number to show whether this column is '*'. For instance, if the data is '*' then the corresponding position in the adding column is 1 and convert the '*' to 0. If not, the corresponding position is 0 and remain the data unchanged.
In the experiment, the number of isolation trees('n_estimators' in sklearn's IsolationForest) was set to 80, the number of random samples it will pick from the original data set for creating Isolation trees ('max_samples' in sklearn's IsolationForest) was set to 'auto' and the ratio of features to draw from training data to train each base estimator ('max_features' in sklearn's IF) was set to 0.6. For prediction, we assume the proportion of outliers('contamination' in sklearn's IF) in the data set is 0.001(about 40 outliers).

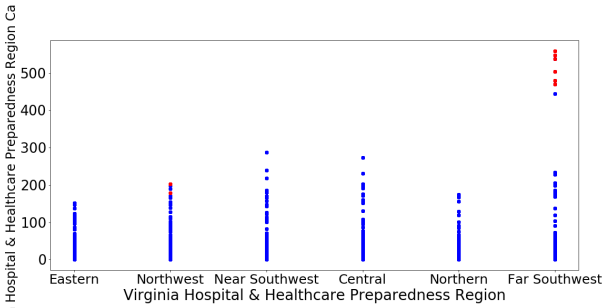

Fig. 1. Virginia Hospital Healthcare Preparedness Region Case Rate for different region(red spots are the outliers)

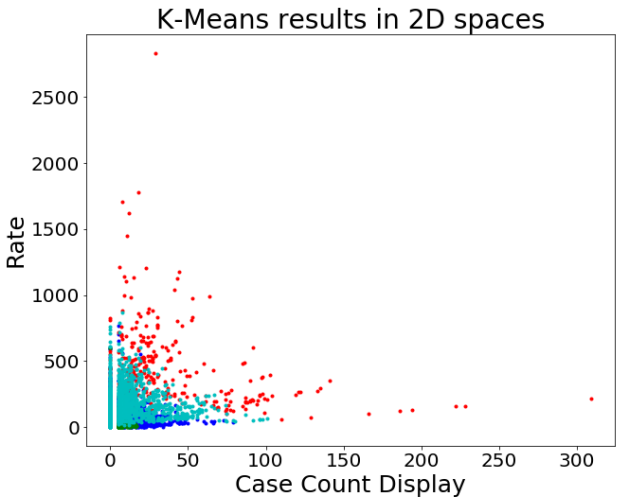

Fig. 2. four clusters of the data regarding Case Count Display vs Rate

Fig. 1 shows the result of the outliers in 2-D space. It is obvious that the top right part of the figure has some outliers indicating Far Southwest region may have some problems for this part apparently have bigger case rate than the other regions. Also, the Northwest part has some anomaly data. We can see the map in Fig. 9 that this area near the West Vriginia State. Later we will look into the outliers and analysis why they are different and what we can do to help this region.

3. Clustering

We first use PCA to reduce the features after data preprocessing. The original feature dimension is 155, and we compress it to 10 dimensions.
Then We use K-Means to do clustering. We chose number of clusters to be 4 and applied the algorithm. Then we chose two features to plot the clusters in 2D space.

Fig. 2 shows the two features in 2D spaces. We can see that the data are automatically clustering to four classes. 1. both Case Count Display and Rate are large(red points), 2. Case

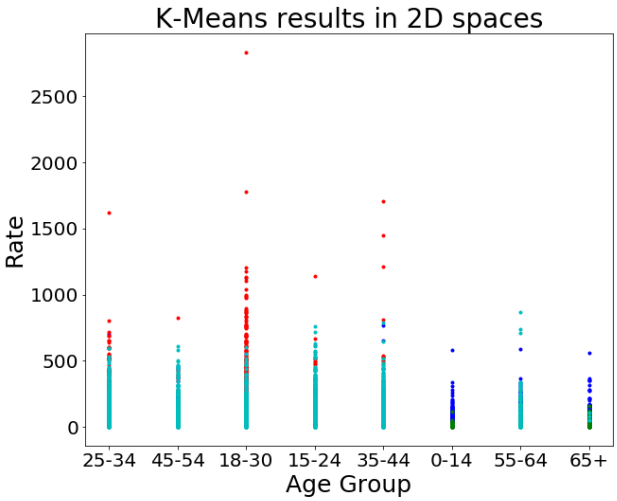

Fig. 3. four clusters of the data regarding Age Group vs Rate

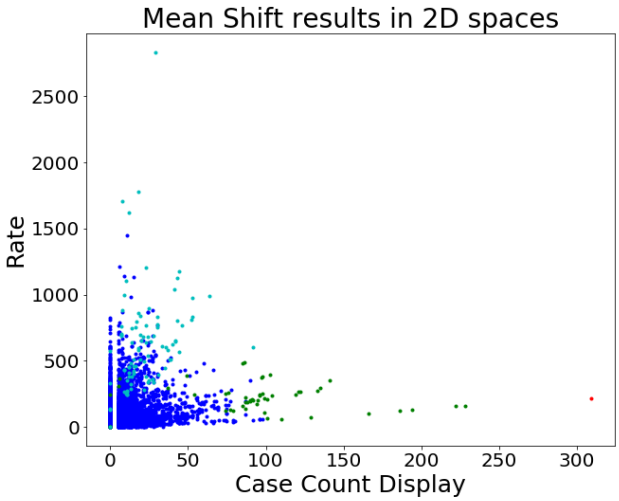

Fig. 4. four clusters of the data regarding to Case Count Display vs Rate

Count Display is large while Rate is small(blue), 3. Rate is at the middle level(light blue), and 4. both Case Count Display and Rate are small(green). These classes are automatically generated by algorithms and the meaning of the classes is clear.

Fig. 3 shows another two features in 2D spaces. We can see that the group at age '0-14' and '65+' performs quite different from the rest for these age groups have little issues about opioid addiction. Also, we can find that the high rates are colored as red which is consistent with Fig. 2.
In addition, we use Mean Shift algorithm to do the same thing. Mean Shift algorithm can find the number of clusters automatically. In our experiment, the hyper parameter 'bandwidth' is 8. And finally it clusters the data into 4 classes.

Fig. 4 shows two features in 2D spaces similar to K-Means algorithm. We can see that the data are automatically clustering to four classes. 1. both Case Count Display and Rate are small or middle(blue points), 2. Case Count Display is large while Rate is small(green), 3. Rate is large while Case Count Display is small(light blue), and 4. Case Count Displays are extremely large (red). The results are different from K-means, however, still meaningful.
4. Insights and Conclusions from Data Visualization
In this part, we mainly used dimensionality reduction to bring down the size to a few factors to describe the data in different aspects. The newly formed factors will be observed to gather insights for a driver analysis. Here we provide several examples of how we use data visualization to further understand how the opioid addiction crisis developed in Virginia.

Fig. 5 is a multi-line graph showing the Hepatitis C new case rate (%) for each VDH Health Region during 2011-2017. To generate this graph, we firstly encoded the labels we need, and

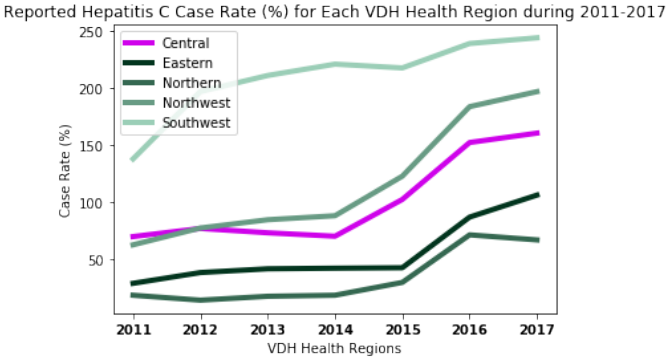

Fig. 5. the Hepatitis C new case rate (%) for each VDH Health Region during 2011-2017.

then grouped the data by Year, Addiction Case Type and VDH region category. After that, we were able to generate the graph above to show the opioid-addiction-related HCV situation in different regions in Virginia. As we can see from the graph, the growth of HCV cases is staggering especially in recent four years, it indicates that all health region of VDH, especially the Southwest health region, have to put more effort in controlling this ongoing increase.

Fig. 6 is a grouped bar graph showing the EMS Narcan Case Rate (%) for Each VA State Police Division(VASPD) during 2011-2016. To generate this graph, we firstly encoded the labels we need, and then grouped the data by Year, Addiction Case Type and VA State Police Division category. After that, we were able to generate the graph above to show the how frequent the Narcan emergency occurs in each VASPD. As we can see from the graph, almost half of all the VASPDs are suffering from an epidemic growth of Narcan in the past four years. Since Narcan emergency rate indicates how much frequent the police are needed to save people from opioid

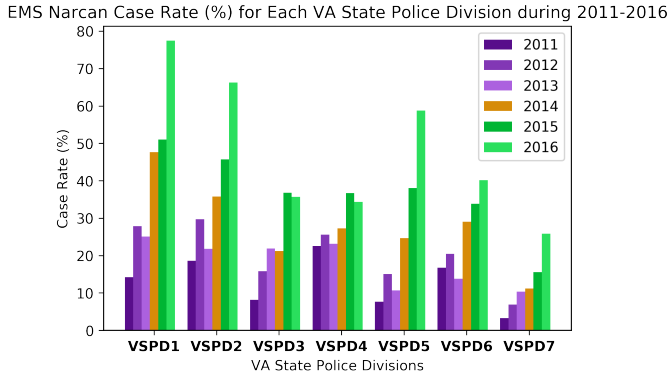

Fig. 6. the EMS Narcan Case Rate (%) for Each VA State Police Division(VASPD) during 2011-2016.

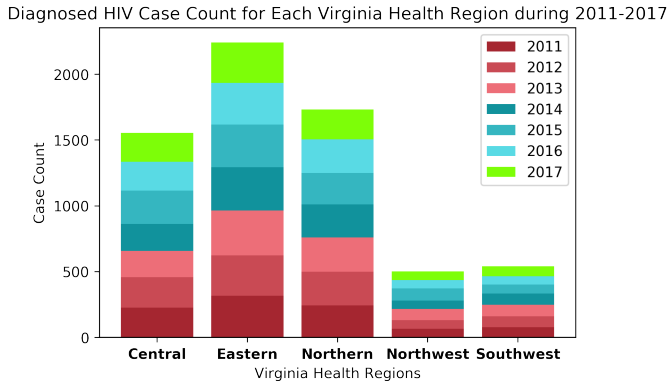

Fig. 7. Diagnosed HIV Case Count for Each Virginia Health Region during 2011-2017

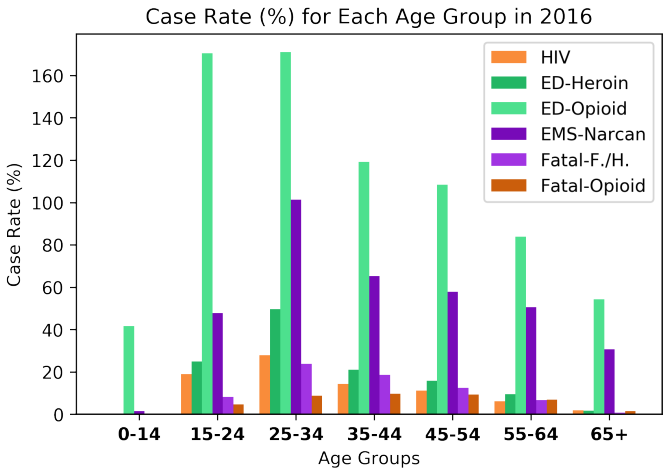

Fig. 8. the Rate of Different Case Types by Age Group in 2016.

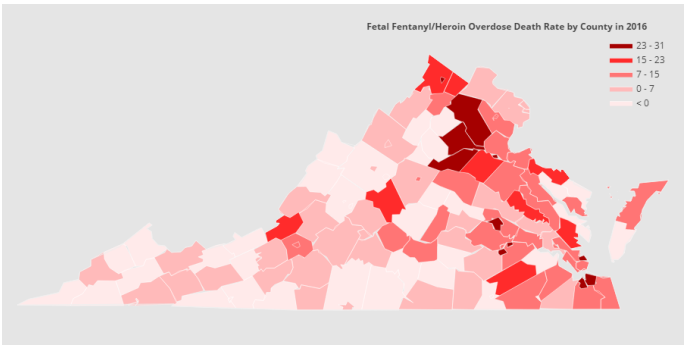

Fig. 9. the Fetal Fentanyl/Heroin Overdose Death Rate by County in 2016.

overdose emergencies, we can conclude that especially in VA State Police Division 1,2 and 5, police officers need more training on how to handle this kind of emergency, and they need more effort in regulating the drug-related activities in each division.

Fig. 7 is a stacked bar graph showing the HIV case count for each Virginia health region during 2011-2017. To generate this graph, we grouped the data by Year, Addiction Case Type and Virginia Health Region category. In this graph, we are able to show the number of HIV cases diagnosed in each Virginia Health Region each year. As we can see from the graph, the overall situation in Virginia seems steady. Based on the case count of each region, we can see the Central, Eastern and Northern regions have higher cases occurring each year. The government can allocate the medical resources for HIV diagnoses and treatment base on this case count statistics. The graph indicates that there is no significant growth of HIV cases in recent years, and the situation seems under control. However, it doesn't mean we don't need further actions, a following analysis of Fig. 8 will raise our attention on the concerns over different case types in different age groups, especially the HIV rates for teenagers and young adults.

In Fig. 8, we can have a look at the situation discussed in previous paragraph from a different perspective. In this graph, we approached the current situation by looking at the occurring rate of different types of cases in different age groups. The reason for choosing the year 2016 instead of 2017 is that, by far statistics on several cases in 2017 is still unavailable. As we can see from this graph, the age group 25-34 have almost all the highest rates of overdose, Narcan emergency, and fatalities. However, in the age group 15-24, we can also notice the opioid overdose rate is among the highest. One of the explanation is, due to their social circle and environment, it is harder for them to encounter more powerful drugs which could lead to serious overdose problems and death. On the other hand, although we have drawn the conclusion in the last part that the overall situation in Virginia seems steady, we can still see from this graph that there are still ways to improve it. The bars indicate that, in age group 15-24 and 25-34, the HIV case rate is much higher than other age groups. Since the opioids are not all injected drugs, it also occurs to us that in some cases, HIV virus is likely being transmitted sexually. Since most people in the age group 15-24 are still in the education system, so providing lectures and consults on HIV-exposed activities are very necessary.

Lastly, to better present the ongoing overdose crisis across

Virginia, we visualized geospatial data and generated the heat map of the fetal Fentanyl/Heroin overdose death rate in each county, as is shown in Fig. 9. As we can see from the graph, the northern and eastern part of the state are having more serious crisis than the other parts. With the broader view of the opioid crisis across Virginia, it is easier for the government to allocate the medical resources for diagnoses and treatment, and police forces for fighting drug-related activities and crimes base on this geospatial statistics. From such map on other case types, we will also gain some insights into the different kind of opioid problems. Based on our findings, we believe the government and other groups need to allocate medical and police resources according to the severeness of the current crisis based on the occurrence of different types of cases. One key prevention strategies include expanding availability and access to emergency treatment and therapies afterward. Also, it is necessary to increase access to prevent the spread of hepatitis C virus infection and HIV infections. To protect our youngsters, it is important to provide more necessary educations on HIV prevention in schools. Public health agencies and law enforcement agencies should work collaboratively to improve detection of and response to outbreaks associated with drug overdoses and protect Virginia against opioid overdose.

VIII. Conclusion

We have shown analysis and visualization of the dataset. We have also described different techniques and experiments carried out on extracted data. From results observed, we learned that counties in the far south-west region have the highest on-going increase in the Hepatitis C case-rate. We also discovered that counties in the northern and southern regions of the state have a serious problem in the Fentanyl/Heroin overdose deaths. We also used clustering algorithms to cluster the data into meaningful classes. With all these observations from experimental results, we see what counties are at risk in the different cases and at what rate incidents in these cases increase. From this, we can determine what counties to be prioritized for each related indicator. In the future, we would like to work more on the data-set by providing predictive analysis by carrying out different prediction techniques on the data. We believe that this will reveal insights on what trends to expect in the coming years and provide ideas on how we can help the counties in dire situations.

IX. Acknowledgements

This work has been a part of the Machine Learning for Virginia project at the University of Virginia in Fall 2018.
We would also like to thank the Virginia Department of Health for providing the Virginia Opioid Dashboard Dataset.

X. Contribution

The link to our project's Github repository: https://github.com/quq99/Opioid_Addiction_Crisis_in_Virginia.git
Xinzuo Wang(xw3xp): 1. Contributed to analyzing and visualizing data. 2. completed data visualization and data analysis.(code and report) 3. complete most preliminary experiments. 4. complete insights and conclusions from data visualization.(code and report) 5. complete most part of report.

Qian Qu(qq8jn): 1. Completed Isolation Forest anomaly detection.(code and report) 2. completed part of data prepossessing.(code) 3. completed clustering part.(code and report) 4. completed the model part and some experiment parts of the report.

Onyi Uche(ocu2t): Contributed to ideas in analyzing, visualizing and consideration of models and experiments on the data as well as organization of content.

References

[1] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, IEEE, 2008, pp. 413–422.

[2] 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 & Machine Intelligence*, no. 7, pp. 881–892, 2002.

[3] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[4] I. Jolliffe, "Principal component analysis," in *International encyclopedia of statistical science*, Springer, 2011, pp. 1094–1096.

[5] *Opioid Overdose and Naloxone Education for Virginia*, Virginia Department of Behavioral Health and Developmental Services, 2017. [Online]. Available: http://dbhds.virginia.gov/behavioral-health/substance-abuse-services/revive.

[6] A. K. Jain and R. C. Dubes, "Algorithms for clustering data," 1988.

[7] T. Kohonen, *Self-organization and associative memory*. Springer Science & Business Media, 2012, vol. 8.