Determining the Security Behavior Intentions of Election Poll Workers: Clustering and Unsupervised Learning of SEBIS Data

Joshitha Mandali
Towson University
jmanda2@students.towson.edu
Advisor: Dr. Josh Dehlinger
COSC 880 Graduate Project
Summer 2021

Abstract— ***In 2017, the Department of Homeland Security (DHS) designated elections infrastructure as critical national infrastructure, and an overlooked component is the nearly 1,000,000 poll workers that are charged with ensuring the integrity of votes and security of voting equipment at the state and local level. Poll workers are the first line of defense in elections security and are trusted insiders to the process. This research analyzes a dataset that comprises of 2,213 poll workers' Security Behavior Intentions Scale (SEBIS) survey responses that were collected during a campaign in spring and summer 2020 targeting poll workers and those who intend to serve during the 2020 General Election and yielded over 2,2000 viable survey responses from 13 states. With the provided non-labelled dataset, this work utilized Jupyter Notebooks to implement an efficient k-means clustering algorithm in the Python language and determined the poll workers security behaviors and their impact on election security.***

KEYWORDS— K-MEANS CLUSTERING, SUPPORT VECTOR MACHINE, SECURITY BEHAVIORS, ELECTION SECURITY

I. INTRODUCTION

Poll workers are one of the primary, visible parts of an election and are the first line of defense in elections security. On every Election Day, hundreds of thousand poll workers in tens of thousands of precincts across the nation are responsible for managing their local polling places [11]. One research [10] concluded that poll workers were not familiar with, nor did they understand election security procedures, though they had a more intuitive understanding of related privacy issues.

When it comes to security, what doesn't happen is frequently as important as what does. How do you effectively train poll workers to look for tampering with a voting machine when they don't understand what a USB port is? [6] Poll workers need to be aware of and vigilant to real-time issues that may occur and threats that may evolve on Election Day.

The main objective of this project is to identify the relationship between the extent of poll worker personal security practices and their impact on election security. Additionally, identifying the personal security practices that might impact polling place security and discuss how those results will impact the supply chain of votes and exploring the approaches outside of the election security The main step is to develop efficient k-means clustering and support-vector machine (SVM) models and to train the model. After training the model, the next step would be to evaluate the model and tune in the parameters to view the accurate results.

II. BACKGROUND AND MOTIVATION

After 2016 United States General Election, the Department of Homeland Security (DHS) revealed that 21 states were susceptible to attacks on their election systems during the election cycle. Several states like Illinois that was subjected to breach that stole registration data for 200,000 voters and Arizona state was a victim of an attack that introduces malware into registration system. Such events, along with the Department of Homeland Security's designation of elections infrastructure as critical infrastructure in 2017, brought a renewed focus on election equipment and its security and integrity.

As discussed in Section I, poll workers are an integral part in the election process. As they have ability to identify and mitigate cyber, physical and insider threats. This paper mainly focuses on identifying the security behaviors and intentions of the poll workers who participated in the survey. The survey questions will be detailed in the description of the dataset section of the paper.

III. DESCRIPTION OF THE DATASET

The dataset used in this work comprises of 2,213 poll workers Security Behavior Intentions Scale (SEBIS) survey responses, which were collected during a campaign in spring and summer 2020 targeting previous poll workers and those who intend to serve in the 2020 General Election and yielded over 2,2000 viable survey responses from 13 states.

For this research, SEBIS responses and demographic question answers provided will be utilized to determine poll workers security behaviors. Because most of the questions are not mandated, especially in the demographic questions, there are missing data. The dataset comprises of 363 rows and 68 columns. The first 18 columns consist of poll workers general details like their IP Address, Email, First name, Last name:

https://github.com/joshitha14/K-means_project880

- ‘StartDate’, ‘EndDate', 'Status', 'IPAddress', 'Progress', 'Duration (in seconds)', 'Finished', 'RecordedDate', 'ResponseId', 'RecipientLastName', RecipientFirstName', 'RecipientEmail', 'ExternalReference', 'LocationLatitude', 'LocationLongitude', 'DistributionChannel', 'UserLanguage’, ‘Personal Computer Security Please enter your participant number, which is the last 4 digits of your telephone number’

This dataset consists of attributes which are categorical variables. A categorical variable is a variable that can take one of limited values, and usually fixed number of possible values assigning each individual or other unit of observation to a particular group or nominal category based on some qualitative property [1]. Example of values that can be represented as categorical variables are blood type of a person: A, B, AB, or O. Gender type: Male, Female. From the 19th column on, the dataset comprises of following categorical variables and their values:

- Q2: Have you previously served as an elections poll worker - Yes, No
- Q3: Do you intend to serve as an elections poll worker during the 2020 Presidential primary or general elections? - Yes, No, Maybe
- Q5: If you have previously served as an elections poll worker, what was your primary role? Please do not include any detailed information that could be used to identify you - Provisional Judge, operations judge etc.
- Q6: What is your age? - 18-24 years old, 25-34 years old, 35-44 years old, 45-54 years old, 55-64 years old, 65-74 years old, 75 years old or older
- Q7: What is the highest level of education that you have completed? - Middle school, High school, Associate degree, bachelor’s degree, master’s degree, Doctoral degree, prefer not to disclose
- Q8: What is your current gender identity? - Male, Female, choose not to mention
- Q9: In what state(s) have you served as a Poll worker? - ‘Maryland’, ‘Florida’, ‘Tennessee’, ‘New York’
- Q10: What were your responsibilities as a poll worker? What equipment or polling stations did you operate? - All of them, as chief judge, ballot etc.
- Q11: For how many years in total have you served as a poll worker? - 0-1 years, 2-3 years, more than 3 years
- Q12: Indicate your level of comfort with using desktop or laptop computers - ‘Extremely comfortable’, ‘Very comfortable’, ‘comfortable’, ‘Somewhat comfortable’, ‘Not comfortable’
- Q31: Please enter your participant number, which is the last 4 digits of your telephone number.
- Q32: Which of the following best describes your current work position? – ‘Professional staff’, ‘Retired’, ‘Student’, ‘Technical’, ‘Student’ etc.
- Q33: Have you previously heard of the potential for cyber, physical, and/or human threats to elections? - Yes, No

The security related SEBIS attributes are classified into the following groups.

Device Securement:

- Q13: I set my computer screen to automatically lock (i.e., sleep) if I don’t use it for a prolonged period
- Q14: I use a password/passcode to unlock my laptop or tablet.
- Q16: I manually lock my computer screen when I step away from it.
- Q17: I use a PIN or passcode to unlock my mobile phone.

Password Generation:

- Q18: I do not change my passwords unless I must.
- Q19: I use different passwords for different accounts that I have.
- Q20: When I create a new online account, I try to use a password that goes beyond the site’s minimum requirements.
- Q22: I do not include special characters in my password if it’s not required.

Proactive Awareness:

- Q23: When someone sends me a link, I open it without verifying where it goes.
- Q24: I know what website I’m visiting based on its look and feel, rather than by looking at the URL bar.
- Q25: I submit information to websites without first verifying that it will be sent securely (e.g., SSL, “https://”, a lock icon).
- Q26: When browsing websites, I mouse over links to see where they go, before clicking them.
- Q27: If I discover a security problem, I continue what I was doing because I assume someone else will fix it.

Updated:

- Q28: When I’m prompted about a software update, I install it right away.
- Q29: I try to make sure that the programs I use are up to date.
- Q30: I verify that my anti-virus software has been regularly updating itself.

Most importantly, from questions Q6 to Q30, every categorial values is assigned a code number to it. For example, Q6 value ‘18-24 years old” is assigned as “1”, value 45-54 years old is assigned as “4”, value 75 years old or older is assigned as “7” code.

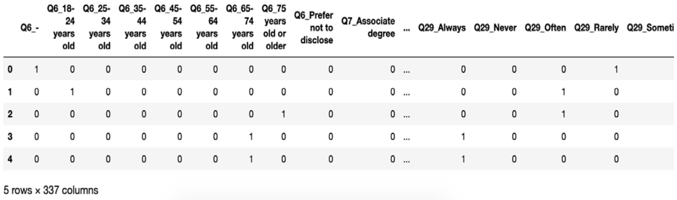
SEBIS questions attributes are assigned codes as well. For instance, from question Q13 to Q18 value ‘Always’ is assigned as code “5”, value ‘Often’ is assigned as code “4”, ‘Sometimes’ is assigned code “3”, value ‘Rarely’ is assigned code “2” and value ‘Never’ is assigned code “1”. Every code column is next to every respective categorical column. ‘Q14 Code’ column is right after ‘Q14’ column.

A. *Feature Selection*

To select features, this work considers only the code columns for visualization, whereas for implementation, the approach followed is one-hot representation because the k-means algorithm isn’t directly applicable to categorical data for various reasons. Converting categorical attributes to binary values and doing k-means clustering was an approach that was implemented.

Categorical data is a problem for most algorithms in machine learning. Suppose, for example, you have some categorical variable called "color" that could take on the values red, blue, or yellow. If we simply encode these numerically as 1,2, and 3 respectively, our algorithm will think that red (1) is closer to blue (2) than it is yellow (3). We need to use a representation that lets the computer understand that these things are all equally different [9].

One simple way is to use one-hot representation. If we take the Q8 gender example, for instance; rather than having one variable like ‘gender’ that can take three values (male, female, choose not to identify), we separate it into three variables. These would be Q8- “male”, Q8- “female”, Q8- “choose not to identify” which only can take on the value 1 or 0 (as shown in the Figure 1).



	Q6_-	Q6_18-24 years old	Q6_25-34 years old	Q6_35-44 years old	Q6_45-54 years old	Q6_55-64 years old	Q6_65-74 years old	Q6_75 years old or older	Q6_Prefer not to disclose	Q7_Associate degree	...	Q29_Always	Q29_Never	Q29_Often	Q29_Rarely	Q29_Someti
0	1	0	0	0	0	0	0	0	0	0	...	0	0	0	1	
1	0	1	0	0	0	0	0	0	0	0	...	0	0	1	0	
2	0	0	0	0	0	0	0	1	0	0	...	0	0	1	0	
3	0	0	0	0	0	0	1	0	0	0	...	1	0	0	0	
4	0	0	0	0	0	0	1	0	0	0	...	1	0	0	0	

5 rows × 337 columns

Figure 1: One-hot representation of the data

B. *Data Cleaning*

Jupyter Notebook and Python were utilized to clean the dataset acquired. In total the dataset comprises 363 rows and 68 columns.

From the dataset, a few columns were dropped that were not helpful to determine the poll worker’s security behavior intentions. On dropping the irrelevant columns, like poll workers general details from column ‘StartDate’ to question “Q5” and questions Q32 and Q33 are also not taken into consideration. So, the remaining columns, along with SEBIS questions, came down to 21columns. There were no duplicate rows to remove.

Additionally, question Q11 consisted of null values. Where the null values have been replaced with data, “choose not to mention” and the code (Q11 code) associated with that has been added as “0” years.

C. *Data Visualization*

Effective data visualization is the important step that will help convey the details in the data. To build some models, this work needed to understand the underlying dataset. To do so, variables were explored in great depth before moving on building a model or working with the data.

Figure 2 shows the pie chart of Q8 (i.e., gender). On the other hand, Figure 3 and Figure 4 depicts the catplot (catplot is a combination of categorical plot with facetgrid, where facetgrid maps a dataset onto multiple axes arrayed in a grid of rows and columns that correspond to the level of variables in the dataset) of the gender against the Q7 (Degree of the poll workers) and Q6 (Age of the poll workers), respectively.

Figure 5 shows the count plot of computer usage vs. gender, from which we can say that female poll workers seem to be extremely comfortable or more comfortable in using the computer or laptop when compared to male poll workers.

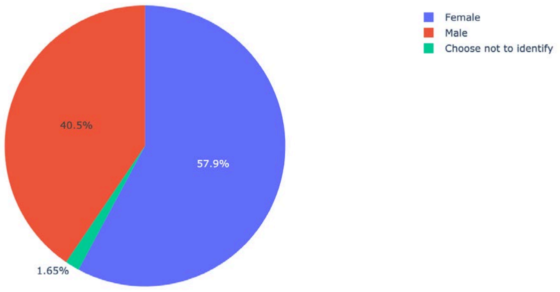


Figure 2: Pie chart represention of Q8 i.e., Gender

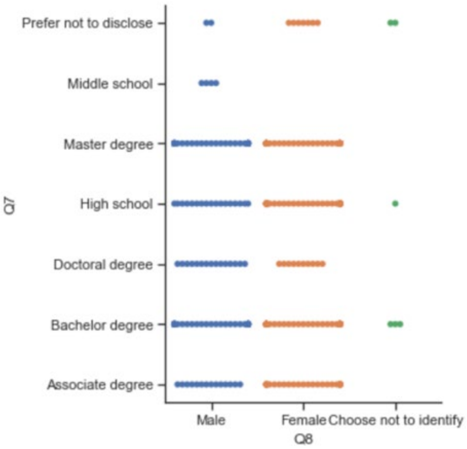


Figure 3: Catplot of Q8 against Q7

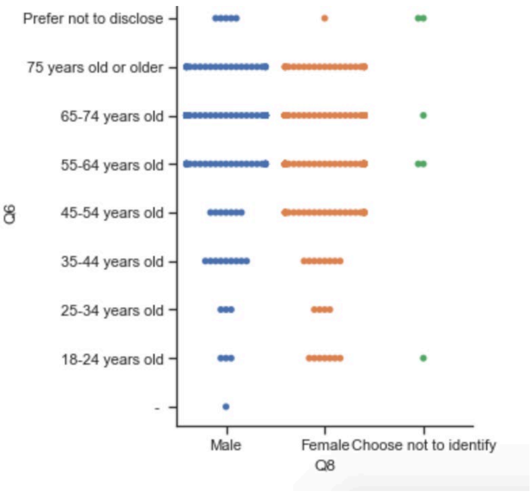


Figure 4: Catplot of Q8 against Q6

Figures 6, 7, 8, 9 show the count plot of the SEBIS questions against the experience. From the bar plots, there are a greater number of poll workers who have chosen “Never” as their option for Proactive Awareness SEBIS questions which includes questions like “When someone sends me link,

https://github.com/joshitha14/K-means_project880

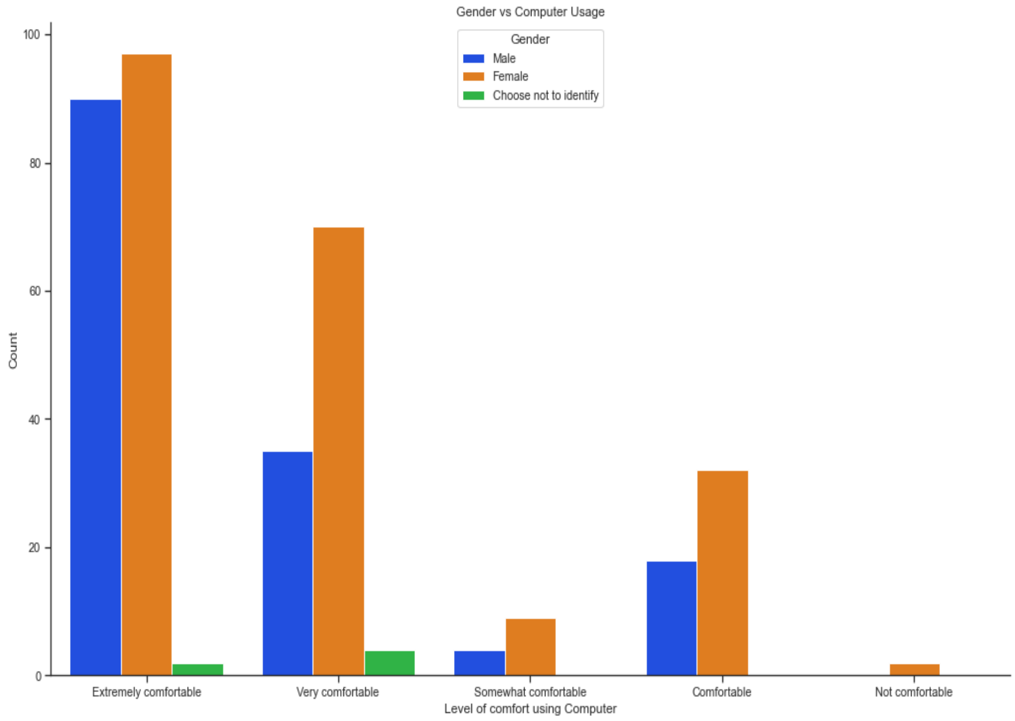


Figure 5: Poll workers gender Vs Poll workers Computer

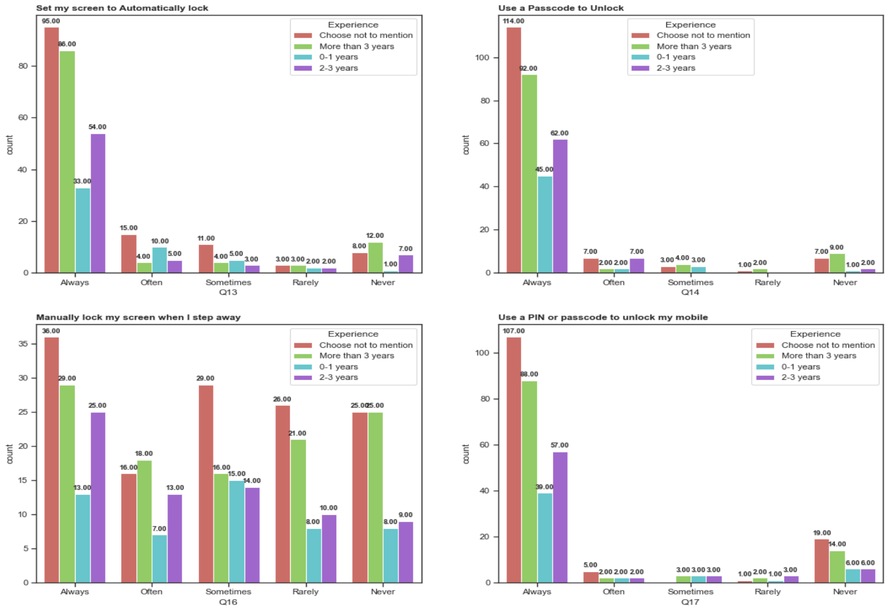


Figure 6: Device Securement questions Vs Poll workers Experience

https://github.com/joshitha14/K-means_project880

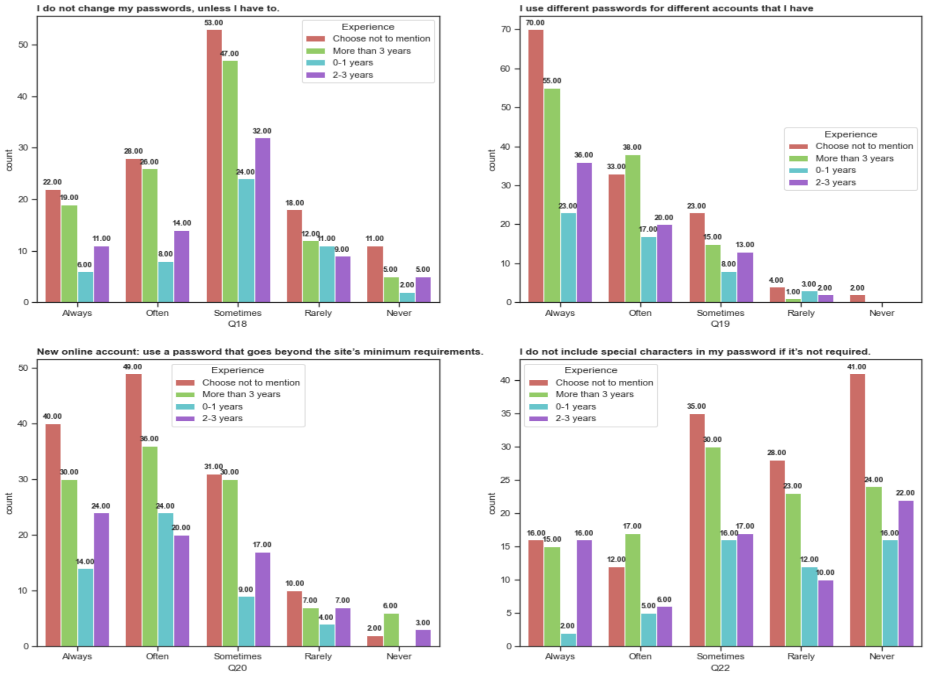


Figure 7:
Password
Generation questions Vs Poll workers experience

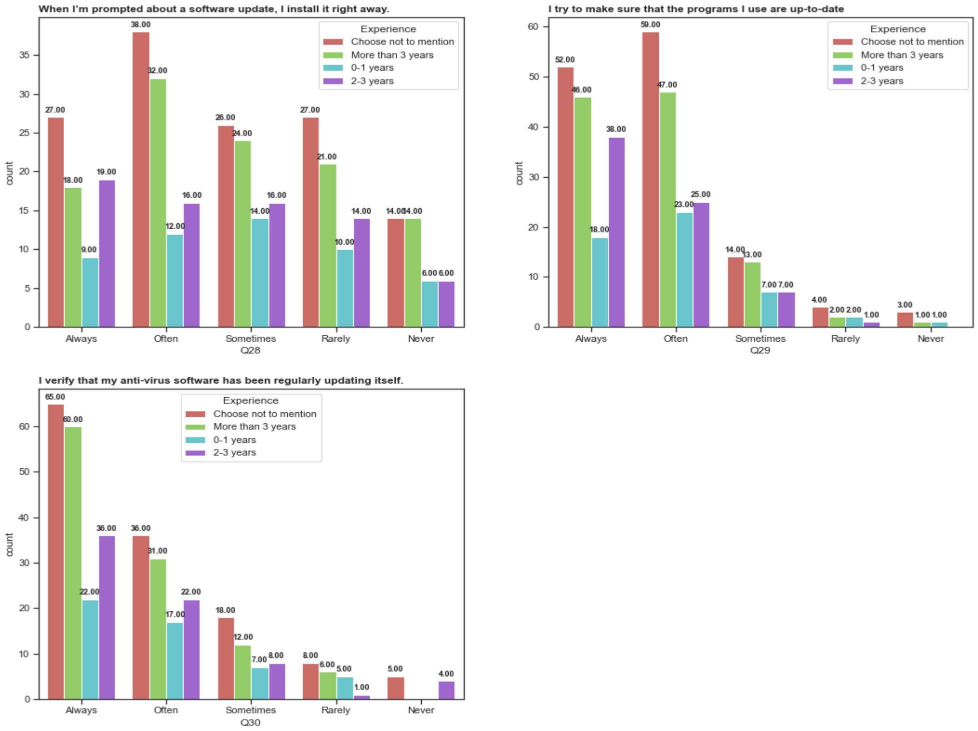
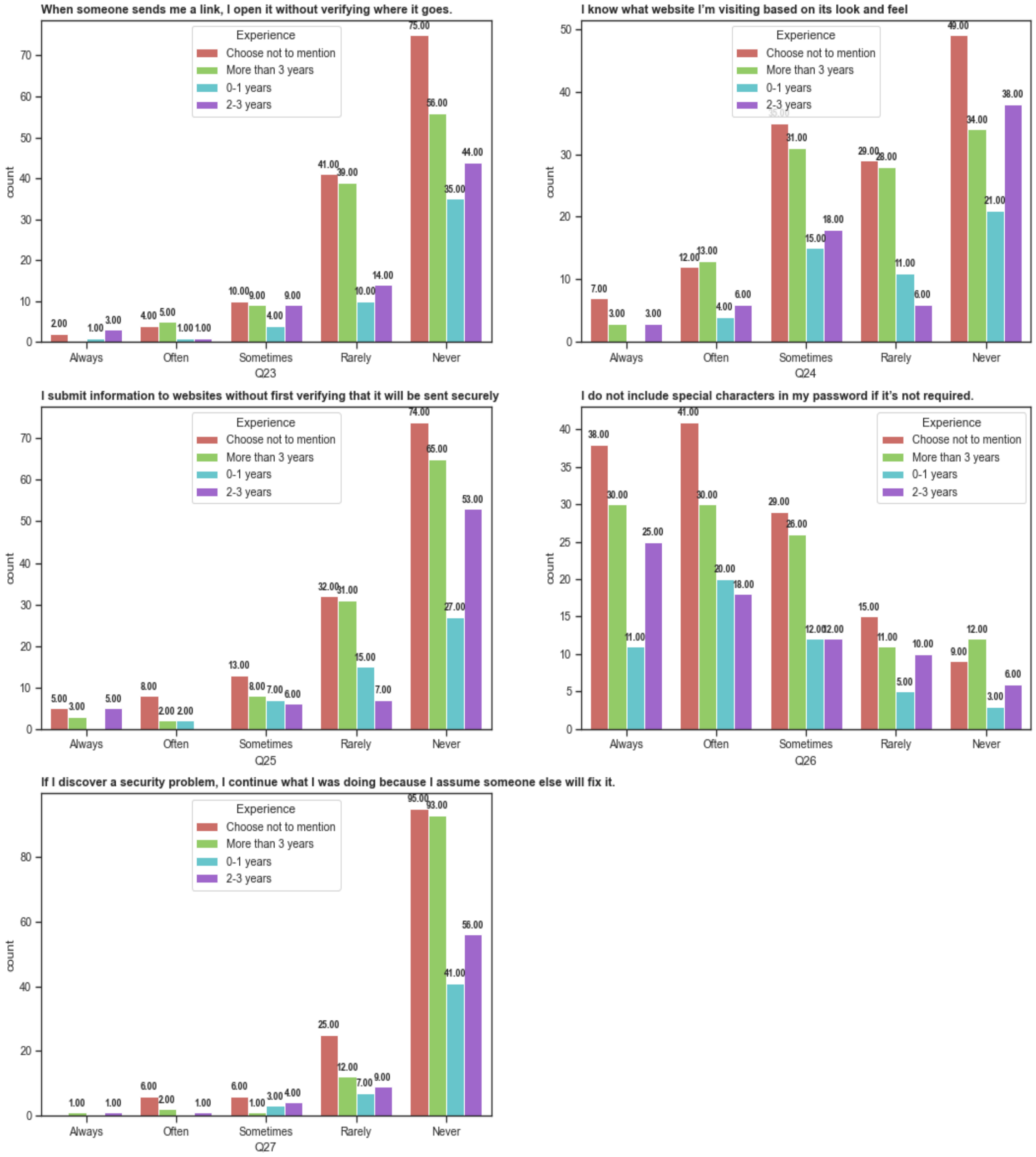


Figure 8:
Proactive Awareness Vs Poll workers experience

https://github.com/joshitha14/K-means_project880



When someone sends me a link, I open it without verifying where it goes.

I know what website I'm visiting based on its look and feel

I submit information to websites without first verifying that it will be sent securely

I do not include special characters in my password if it's not required.

If I discover a security problem, I continue what I was doing because I assume someone else will fix it.

Figure 9: Proactive Awareness questions Vs Poll workers Experience

https://github.com/joshitha14/K-means_project880

I open it without clicking it", "I know what website I am visiting without verifying it" etc.

Whereas a majority number of poll workers have chosen "Always" for Device Securement and Updating SEBIS questions like "I set my computer screen to automatically lock (i.e., sleep) if I don't use it for a prolonged period" and "When I'm prompted about a software update, I install it right away", respectively.

On the contrary, majority number of people have selected the option "Sometimes" for the question "I do not change my passwords unless I must" and have chosen "Never" option for the question "I do not include special characters in my password if it's not required".

IV. Algorithms

A. *Clustering*

Clustering is a process of arranging objects into groups in such a way that the data points in the same groups are alike. A cluster is a collection of objects where these objects are similar and dissimilar to the other cluster [2].

B. *k-means Clustering*

k-means clustering is an unsupervised machine learning. Unsupervised learning is when you only have input data (X) without a corresponding target variable (y) to predict. Its aim is to model the underlying structure of the data to learn from data and identify groups of data (segments / clusters) with similar characteristics / behaviors. It is an iterative process where each data point is assigned to one of the k groups based on similar feature.

Working: It takes "k" as an input which is the number of clusters- and the X variables. It starts by picking the centroids (C1…Ck) to random locations (2 centroids in the example as shown in Figure 10).

It calculates the distance (e.g., Euclidian distance between the centroid (C1) and the point (Xi)). Basically, the centroids are computed by taking mean of all data points which are assigned to a particular cluster.

It then picks the cluster with minimum distance. We must recalculate the centroids by taking all their vectors and averaging them (add them and divide them number of points), which is the new cluster centroid. The same process is repeated until there is no change in the clusters.

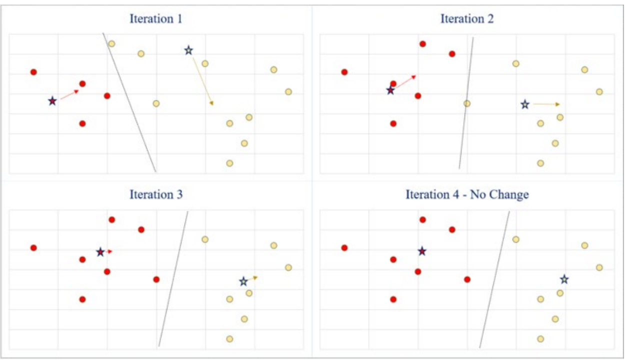


Figure 10: K-means example

https://github.com/joshitha14/K-means_project880

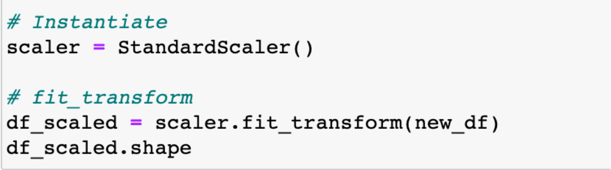
V. Experiment

A. *Data Preprocessing*

To avoid the vast difference between the range of values, this work incorporated an important step in the analysis by standardizing the data.

Standardization is an important part of data preprocessing, and it comes to picture when features of input data set have differences between their ranges, or simply when they are measured in different measurement units.

For this project, a Standard Scaler was used to standardize the data. (Figure 11).



```
# Instantiate
scaler = StandardScaler()

# fit_transform
df_scaled = scaler.fit_transform(new_df)
df_scaled.shape
```

Figure11: Data Standardization

After data standardization, dimensionality reduction is performed for which I have used Principal component Analysis, which is a very popular dimensionality reduction technique used to avoid "the curse of dimensionality".

B. *Principal Component Analysis*

Principal component analysis (PCA) is a method of compressing large data into something that captures the essence of the original data. PCA takes a dataset with lot of dimensions (i.e., lot of cells) and flattens it to 2 or 3 dimensions, so we can look at it. It tries to find a meaningful way to flatten the data by focusing on things that are different between cells.

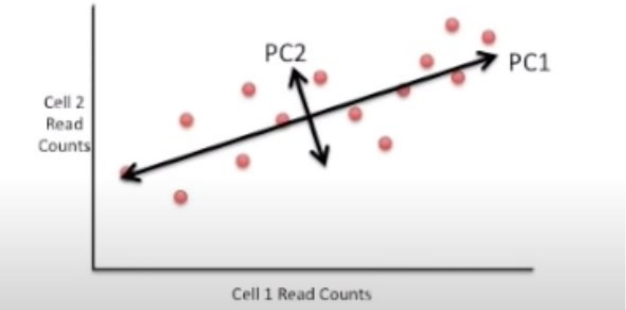


Figure12: PC1 and PC2 plotting

Now, when plotting a point based for each person behavior on how many reads were from each cell (as shown in figure 12). PC1 (the first principal component) captures the direction where most of the variation is, and PC2 captures the direction with the 2nd most variation. PC1 and PC2 now becomes the new x and y axis as shown in Figure 12.

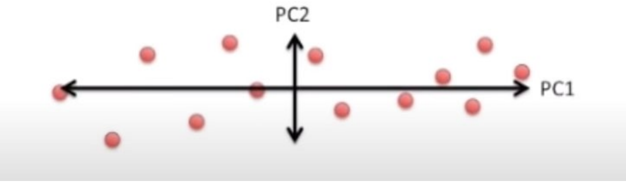
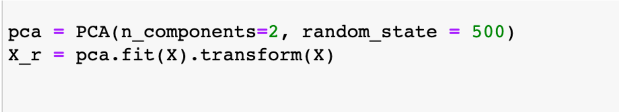


Figure13: PC1 And PC2 as x, y axis

From Figure13, PC1 (the first principal component) is the axis that spans the most variation. PC2 is the axis that spans the second most variation. If we had 4 cells, PC1 would span the direction of the most variation, PC2 would span the direction of the 2nd most variation, PC2 would span the direction of 3rd most variation, PC4 would span the direction of the 4th most variation. There is a principal component for each dimension (cell). If we had 200 cells, we would have 200 principal components. PC200 would span the direction of the 200th most variation.

So, this work had to fit the standardized data (formed in Figure 11) using PCA as shown in Figure 14 [4].



```
pca = PCA(n_components=2, random_state = 500)
X_r = pca.fit(X).transform(X)
```

Figure14: PCA components

Secondly, how many features to keep had to be determined based on the cumulative variance plot, as shown in Figure 15. From the graph, the red line shows the cumulative sum, and we can read off the percentage of the variance in the data explained as we add principal components.

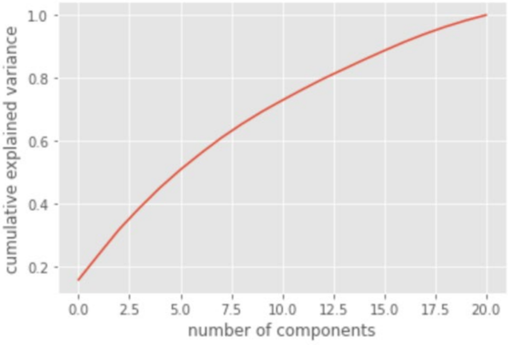
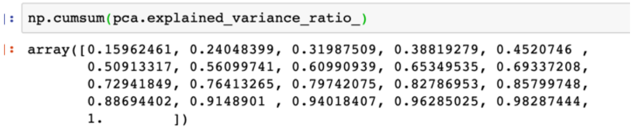


Figure 15: Cumulative Variance plot

From Figure16., we can see that the first principal component explains 15% of the variance of the dataset. The first 2 principal components explain 24% and first 3 explains 31%, first 4 components explain 38% and first 5 principal components explain 45% of the variance of the dataset. As the next step, this work performed PCA with the chosen number of components which is 5 principal components.



```
|: np.cumsum(pca.explained_variance_ratio_)

|: array([0.15962461, 0.24048399, 0.31987509, 0.38819279, 0.4520746 ,
          0.50913317, 0.56099741, 0.60990939, 0.65349535, 0.69337208,
          0.72941849, 0.76413265, 0.79742075, 0.82786953, 0.85799748,
          0.88694402, 0.9148901 , 0.94018407, 0.96285025, 0.98287444,
          1.        ])
```

Figure16: Components for PCA

To perform segmentation based on principal components scores instead of original features is to incorporate newly obtained PCA scores in the K-means algorithm. So, to combine the PCA with K-means, it is necessary to first determine the number of clusters to choose; to do so, the elbow method was utilized to determine the number of clusters [7].

C. Elbow Method

The elbow method is one of the metrics that gives an intuition about the selection of a ‘k’ value. Basically, it gives an idea on what a good ‘k’ number of clusters would be based on the sum of squared distance (SSE) between the data points and their assigned clusters’ centroids. ‘k’ is determined by identifying the spot where SSE starts to flatten out, forming an elbow.

The idea is that we want a small SSE, but that the SSE tends to decrease toward 0 as we increase k (the SSE is 0 when k is equal to the number of data points in the dataset, because then each data point is its own cluster, and there is no error between it and the center of its cluster). So, the primary goal is to choose a small value of k that still has a low SSE, and the elbow usually represents where we start to have diminishing returns by increasing k. [3]

Figure 17 shows that after 4 clusters (i.e., the elbow), the change in the value of inertia is no longer significant and most likely, neither is the variance of the rest of the data after the elbow point. Therefore, we can discard everything after k= 4 and proceed to the next step in the process.

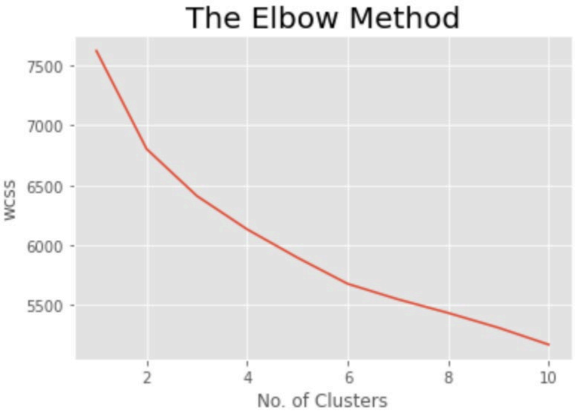


Figure17: The Elbow Method

C. Implementation

Next step was to run K-means with the number of clusters which is equal to 4. Subsequently we are fitting the model with the principal component scores [8] as shown in Figure 18.

https://github.com/joshitha14/K-means_project880

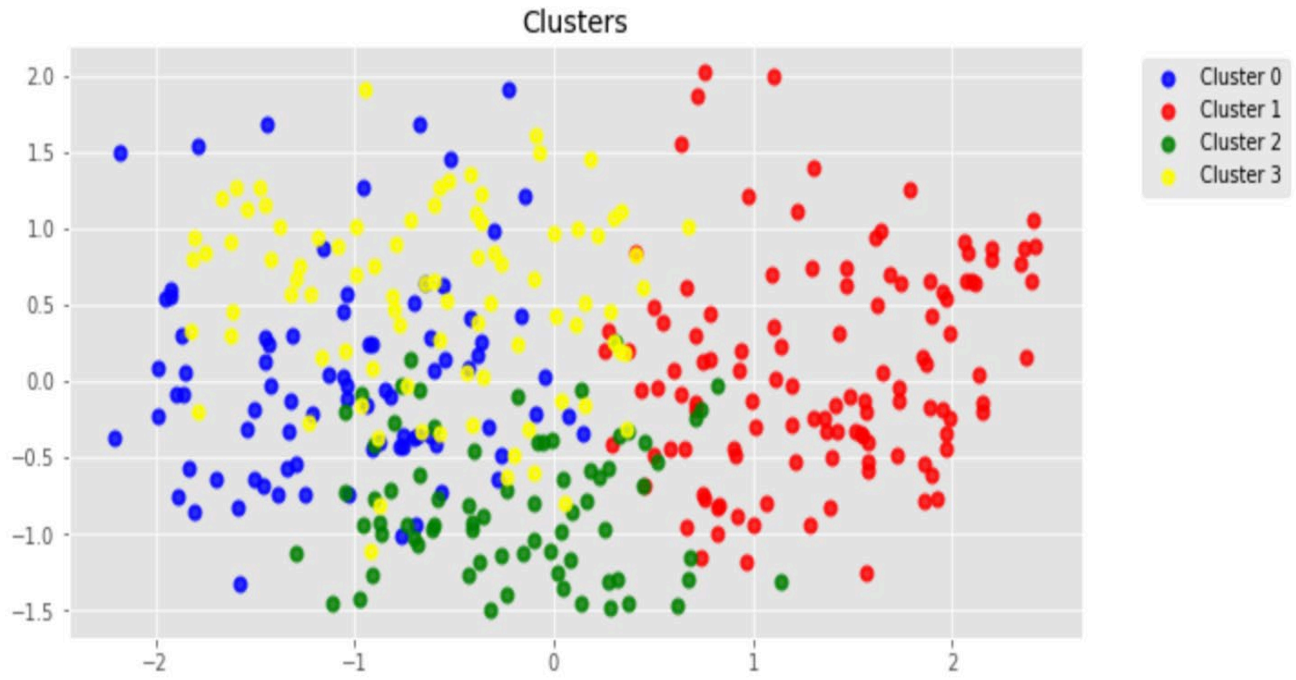
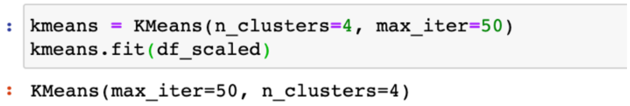


Figure20: Clusters with respect to the components



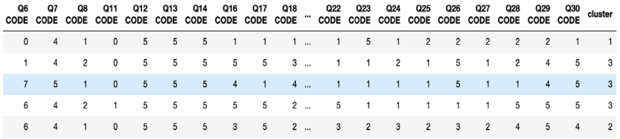
```
: kmeans = KMeans(n_clusters=4, max_iter=50)
  kmeans.fit(df_scaled)

: KMeans(max_iter=50, n_clusters=4)
```

Figure18: Implementing K-means

D. Visualize and Interpret the Clusters

The next step is the most interesting part where I had to analyze the results of the algorithm. Before all else, a new data frame was created where it allowed us to append the clustering labels to the new data frame. Figure 19 shows the snippet of the data frame with the clusters at the end.

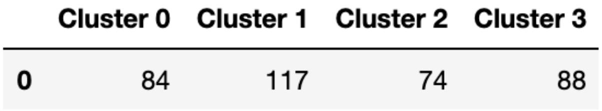


Q6 CODE	Q7 CODE	Q8 CODE	Q11 CODE	Q12 CODE	Q13 CODE	Q14 CODE	Q16 CODE	Q17 CODE	Q18 CODE	...	Q22 CODE	Q23 CODE	Q24 CODE	Q25 CODE	Q26 CODE	Q27 CODE	Q28 CODE	Q29 CODE	Q30 CODE	cluster
0	4	1	0	5	5	5	1	1	1	...	1	5	1	2	2	2	2	2	1	1
1	4	2	0	5	5	5	5	5	3	...	1	1	2	1	5	1	2	4	5	3
7	5	1	0	5	5	5	4	1	4	...	1	1	1	1	5	1	1	4	5	3
6	4	2	1	5	5	5	5	5	2	...	5	1	1	1	1	1	5	5	5	3
6	4	1	0	5	5	5	3	5	2	...	3	2	3	2	3	2	4	5	4	2

Figure19: Cluster column appended to the data frame

To wrap it up by visualizing the clusters on a 2D plane, the first two components were used as axes. The point of PCA was to determine the important components which would explain more variance than the third or fourth one. Figure 20 shows the clusters with respect to the components. Each dot in the figure represents a single poll worker behavior or his/her survey response. The general idea is that the dots with similar responses or poll workers with similar behavior should cluster and we sort of see that in the graph [5].

Overall, Cluster 0 is different from Cluster 1 which is different from Cluster 2 which is different from Cluster 3. But there are few dots club together where we can say that their behavior can be similar.



	Cluster 0	Cluster 1	Cluster 2	Cluster 3
0	84	117	74	88

Figure21: Cluster Count

VI. Analysis of Results

From Figure 20, Figure 21, and based on the analysis from the clusters Excel file in the GitHub link, Cluster 1 and Cluster 2 have highest number of people who are above 50. Additionally, Cluster 1 and Cluster 2 has people with more experienced, who are extremely comfortable in using the laptops. Essentially, pertaining to SEBIS questions Cluster 1 and Cluster 2 have significantly greater number of people (ranking wise) who lock their systems (Device Securement); who use different passwords for different accounts (Password Generation) and has greater number of people who never click a link when someone sends it (Proactive Awareness) and greater number of people who verify their Anti-virus software is regularly updating itself.

On the contrary, Cluster 0 and Cluster 3 have a greater number of people who has degree lower than the associate degree. Also, has the highest number of people who do not lock their systems manually or do not auto lock their systems

https://github.com/joshitha14/K-means_project880

(Device Securement); who do not set their passwords to their mobile phones (Password Generation).

This work concludes that poll workers who have higher education, who are above 40 years of age and who are extremely comfortable in using the computer or laptops has highest number of poll workers (ranking wise) who lock their systems, who use different passwords for different accounts and who never click a link when someone sends it. Moreover, these people update their software or anti-virus up to date.

Precisely based on this analysis, gender or the age may not be accurate because a greater number of females have taken this survey (as shown in Figure 2), and furthermore, there are almost 310 people who are above 40 years of age. So, this work had to just consider the analysis based on qualification and level of comfortability in using the computer.

VII. Lessons Learned

My initial plan was to determine the security behavior of the poll workers using two algorithms, one is k-means clustering and the other one is the support-vector machine algorithm. First, I did not have any prior experience in working with k-means clustering or support vector machine algorithms and learning to use it wasn't happening quickly enough amid the other challenges we were facing this semester. I believe that with the help of the research carried out by Dr. Dehlinger and Dr. Scala aided in understanding the dataset, similarly it helped me in having clear direction in using the k-means clustering algorithm.

While doing some research on the support-vector machine algorithm, we hit a wall as support-vector machine algorithm turned out to be applicable to supervised learning problems. Thus, we have decided to focus on implementing k-means clustering algorithm.

Taking Data Science courses for my masters helped me a lot, especially the Data Mining course which made me understand how to follow the KDD process (the Data cleaning, Data selection and transforming, Data Mining, Pattern Evaluation) and predict the problem in different ways.

Another lesson learnt is that exploring few other unsupervised algorithms would have helped me in achieving accurate results than considering two algorithms.

VIII. Conclusion

In this project, a PCA with k-means clustering algorithm was utilized on a dataset which comprises of 2,213 poll workers Security Behavior Intentions Scale (SEBIS) survey responses which were collected during a campaign in Spring and Summer 2020 targeting previous poll workers and observed that poll workers who have higher education and who are extremely comfortable in using the computer or laptops are the people who lock their system when they are away from the systems. Apparently, these are the people who verify that their anti-virus software has been regularly updating itself and who change their passwords occasionally.

Concerning with the results for the k-means algorithm, we may have achieved better accurate results if the questions like poll worker experience would have been made mandatory. For future work, we can also expect to improve the results by producing some metrics.

https://github.com/joshitha14/K-means_project880

IX. References

[1] https://en.wikipedia.org/wiki/Categorical_variable

[2] https://towardsdatascience.com/k-means-clustering-for-beginners-2dc7b2994a4

[3] https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891

[4] https://365datascience.com/tutorials/python-tutorials/pca-k-means/

[5] https://medium.com/@dmitriy.kavyazin/principal-component-analysis-and-k-means-clustering-to-visualize-a-high-dimensional-dataset-577b2a7a5fe2

[6] https://slate.com/technology/2018/10/elderly-poll-workers-threat-election-security.html

[7] https://365datascience.com/tutorials/python-tutorials/principal-components-analysis/

[8] https://medium.com/@dmitriy.kavyazin/principal-component-analysis-and-k-means-clustering-to-visualize-a-high-dimensional-dataset-577b2a7a5fe2

[9] https://datascience.stackexchange.com/questions/22/k-means-clustering-for-mixed-numeric-and-categorical-data

[10]https://www.usenix.org/system/files/conference/evtwote12/evtwote12-final11-072612.pdf

[11] https://link.springer.com/chapter/10.1007/978-3-319-20376-8_50#CR5