Healthcare Analytics: A Clustering Application for Pharmaceutical Data

[1] ***Violeta Maricela Dalgo Flores***
0000-0002-4004-5938
violeta.dalgo@espoch.edu.ec
[2] Grupo de Investigación de Ambiente y Desarrollo. GIADE)
Facultad de Ciencias, Carrera de Bioquímica y Farmacia. Escuela Superior Politécnica de Chimborazo, Panamericana Sur Km $1^{1/2}$, Riobamba, Ecuador.

[1]***Valeria Rodríguez Vinueza***
0000-0001-5515-3539
valeria.rodriguez@espoch.edu.ec
Email: valeria.rodriguez@espoch.edu.ec.
[1]Grupo de Investigación de Tecnología y Atención Farmacéutica de Ecuador (GITAFEC). Facultad de Ciencias, Carrera de Bioquímica y Farmacia. Escuela Superior Politécnica de Chimborazo, Panamericana Sur Km $1^{1/2}$, Riobamba, Ecuador.

[3] ***Bolívar Morales Oñate***
0000-0003-4980-8759
be.morales@uta.edu.ec
[3]Grupo de Investigación Ciencia de Datos (CIDED), Escuela Superior Politécnica de Chimborazo-ESPOCH.
[3]Universidad Técnica de Ambato, Facultad de Ingeniería en Sistemas, Electrónica e Industrial, Carrera de Ingeniería Industrial., Av. Los Chasquis y Río Payamino, Ambato, Ecuador.

[4]**Evelyn Carolina Macias Silva**
0000-0001-7593-6952
[4]Escuela Superior Politécnica de Chimborazo
Grupo de investigacion ESPASCI
Evelyn.macias@espoch.edu.ec

***Abstract*— An unsupervised algorithm is applied on a cleaned dataset to obtain analytical visualizations that depict the trends of the medicine purchase behavior of the Hospital, from which business decisions can be made to minimize medicine wastage, to plan better utilization of resources for medicines. First, three drugs are obtained dated for six months of data. There are three distinct drugs in Antibiotics. The datasets are then cleaned to remove optional parameters that will skew the algorithm. After the cleaning is completed, the medicines under one category are merged into one dataset. Then the K-Means Algorithm has been applied to it in the JupyterLab environment using Python. The output depicts the usage of a particular medicine, which medicine is being bought the most in a specific category and even tells us which gender is using each medicine. Theoretically, these visualizations can help the pharmacy plan on reducing expenditure on buying medication in the first place.**

Keywords—Medicine, Pharmacy, Data, Healthcare Analytics, Clustering Algorithm, K-Means, Resource Planning.

I. INTRODUCTION

Medicines being unused and unsold is a significant loss of revenue to a hospital or a pharmacy[1]. Although this does not affect the manufacturers of the drugs, it affects the hospitals adversely as they can save resources instead of spending it on medicines that will not be sold in the long run. Business Intelligence is required for any corporate establishment to compete in the market. As such, a healthcare analyst must find out viable business metrics on which actions can be taken for the benefit of the organization. Given the enormous amounts of data that are present in the healthcare system, numerous fronts can be explored. In this, pharmaceutical data, in particular sales data, is being investigated using an unsupervised algorithm. Data science in the healthcare industry is still in its embryonic stages. Applying unsupervised or supervised clustering algorithms to health data after identifying key business metrics will help in zeroing in on to the issue at hand and help address it., with the potential to increase profits in this scenario.

II. THE EXISTING PROBLEM

A. Current Scenario

An excess of unused medicines can happen due to overstocking. It's not just hospitals but also households, NGOs, offices stock up drugs past their expiry date, and donate them to local pharmacies. In some instances, repackaging and rebranding of the medicines also take place. Ecopharmacovigilance (defines as science and activities concerning detection, assessment, understanding, and prevention of adverse effects or other problems related to the presence of pharmaceuticals in the environment, which affect humans and other species) is gaining momentum. A study in Ghana showed that 4/5 households dispose of medicines without following due procedure. These drugs – due to their plastic packaging – are not degradable biologically. For example, in India, rules for handling waste apply to solid

waste, hazardous waste, bio-medical waste (BMW), plastic waste, e-waste, and batteries. BMW rules apply to medicines that are expired, contaminated, and should be disposed of carefully by incineration. Another critical factor is that hazardous, radioactive chemicals in certain medicines still do not fall under the BMW category[1]. Therefore, better rules are required for enforcing and accountability to prevent medicine wastage.

Instead of stocking up on too many medicines, is it possible to estimate the number of medications required by clustering previous medicine data for a given organization?

B. Data Mining's Role in Healthcare

Data Mining is gaining momentum in every field as all businesses and enterprises rely on data to drive their business and make effective business decisions that boost their value. However, it has not been thoroughly utilized in the healthcare sector. Data mining is an emerging tool in the field of healthcare. The opportunities where data can be used is immense - be its patient data, medical records, insurance data, safety data. As such, it has a significant load of potential to uncover trends and patterns when data mining techniques such as clustering, regression, and classification are applied. In one instance, people with the genes for diabetes were identified based on certain factors the medical experts observed from patients who had diabetes. As a result, care for the patient who might get diabetes started way in advance, therefore saving his life from grave danger[5]. Vaccines can be planned better; sanitation can be enforced better, diseases can be diagnosed sooner as a result of applying data science to this industry[5, 6]. Because of the gigantic amounts of data in this field, it is key to making business as well as medical decisions, which can be a matter of life and death. Medical data mining produces business intelligence, which is useful for diagnosing the disease.

III. METHODOLOGY

A. Clustering

It is a data science technique that helps us make meaning from a dataset. It is a task to identify smaller groups in the data, which are different from other groups[2]. However, within one subgroup, all the items are very similar. The similarity measure that's used to measure this homogeneity is Euclidian or Hamming Distance[4].

It can be used to find people that are similar in terms of behavior, for example. This logic is even used in advanced imaging where image processing methods are used. It is an entirely unsupervised algorithm as it explores the data by itself and finds the result by itself[2]. In comparison, a supervised algorithm requires a hard target input by the user.

B. Unsupervised Clustering Analysis

Unsupervised algorithms are those algorithms that do not have a hard target. They cluster the dataset and explore the dataset without input from the user as to how to proceed. As a result, previously unknown and unexpected trends can also be found.[2]

In a cluster, similar items are grouped. So one cluster's characteristics and intrinsic properties are inherently different from the other cluster's properties. However, within a cluster, the features are similar for each item. Some of their applications include detecting anomalies, grouping items from a vast dataset, finding behavior patterns of consumers purchasing items on online platforms. Fig.1. Explains the working of an unsupervised algorithm visually.

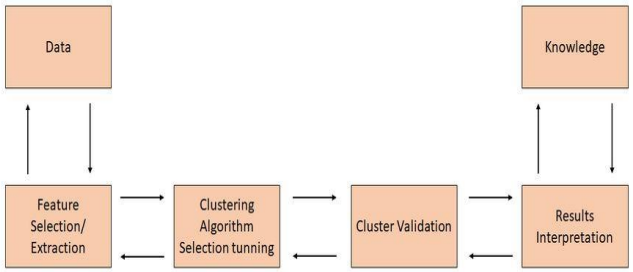

Fig. 1. Process of an Unsupervised Algorithm.

C. K-Means: An Unsupervised Algorithm

This algorithm is iterative and tries to separate the dataset into clusters. These clusters are distinct, but within a cluster, the points are incredibly similar, which is why they are clustered together in the first place[2]. Data points are assigned to a cluster in that the addition of the squared distance between the data points and the cluster's centroid the least. In a cluster, if the scores are similar, they are said to be homogeneous[3]. The procedure is as follows:

1. Define cluster number K.
2. Shuffle the dataset and then choose K points for centroids.
3. Repeat the above step till the centroid does not change.
4. Calculate the squared distance sum(of data points and centroids)
5. Clusters are made by assigning points.
6. Calculate the mean of all data items that belong to a particular cluster to get the centroid.

How K-Means decides on the number of clusters is by something called the Elbow Curve. This is based on the sum of squared distances between data points and their assigned cluster's centroids.[3]

IV. Alternate methodology

A. *DBSCAN*

DBSCAN(Density-Based Spatial Clustering of Applications with Noise) is a method of clustering that separates high-density components from the lower density components.[4]

Advantages of DBSCAN:

- Given enough data, it clusters high vs. low density very effectively.
- It handles outliers also in the clustering.

Disadvantages of DBSCAN:

- Cannot handle variations in the densities.
- Cannot handle similar densities.

B. *Gaussian Mixture Models*

Gaussian Mixture Models(GMM) is an algorithm that clubs data points from a single distribution and makes them into a cluster. These are probabilistic models that use the soft clustering approach to distribute the points in different clusters[4].

Advantages of GMM:

- Most effective clustering.
- Clusters outliers also.

Disadvantages of GMM -

- It needs enormous amounts of data.
- Not suitable for small datasets.

V. Goals

As K-Means is an unsupervised algorithm[2], we do not explicitly tell the algorithm what to find. It finds clusters based on its algorithm and outputs trends and explores the data on its own. These are the goals which were later found -

- To determine which medicine is used the most in each category.
- To determine what is the usual quantity bought by a customer for a particular category.[1]
- To observe if medicines bought vary by gender (male and female).

Theoretically, these visualizations can help the pharmacy plan how many drugs to stock up on in a particular season for a specific category, thereby reducing expenditure on buying medicines in the first place.

VI. Technical and data requirements

The data has been obtained from a Hospital in India on the condition that patients' privacy is respected and the data be anonymized. Due to privacy concerns, the datasets have been anonymized and cleaned with only the required three parameters present in the new dataset on which the algorithm operates. The cleaned dataset contains no information that can be traced back to the original consumer of the medication. The following three medicines were chosen in the 'Antibiotics' category as they were the most prominent medicines sold in the mentioned category for the selected Hospital:

1. Augpen 625.
2. Augpen 325.
3. Cetil 250.

The parameters required for the analysis are:

- Month.
- Gender.
- Quantity of tablets.

The K-Means code is run on the JupyterLab environment, and the language used is Python, specifically the NumPy, pandas, and matplot libraries which are required for the data processing and the visualization.

A. *Steps to be followed*

1. Identify and shortlist the medicines.
2. Select the required parameters(it can be more than the previously mentioned according to insight and business requirement).
3. Obtain the dataset from the Hospital or pharmacy.
4. Clean the dataset of unnecessary and unrequired rows and columns that can skew the data and affect our visualizations.
5. Obtain the K-Means python code from any open-source platform.
6. Convert the datasets into CSV files, as required by K-Means variables.
7. Run K-Means.
8. Infer insights from the visualizations.

VII. Execution of K-Means analysis

The required libraries(numpy, pandas matplotlib, seaborn) are imported into a notebook on JupyterLab. After the importing, the data files are opened in the CSV format. The data frame needs to be allocated as per memory allocation such that the info for the file description is readily available for the algorithm to use. To crosscheck, use describe data to find whether the dataset is well placed.

Next, the covariance is estimated between the three quantities among themselves to see if there is any fluctuation between the two parameters for each. After the estimation, the information is checked to confirm if it has been appropriately uploaded, read, and stored, as shown below.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 994 entries, Augpen 375 to Augpen 375
Data columns (total 3 columns):
Month          994 non-null int64
Gender(1/0)    994 non-null int64
Quantity       994 non-null int64
dtypes: int64(3)
memory usage: 31.1+ KB
```

Fig. 2. Information of the uploaded data frame.

The data frame needs to be allocated as per memory allocation such that the info for the file description is readily available for the algorithm to use. To crosscheck, describe data is used to find whether the dataset is well placed.

Next, the covariance is estimated between the three quantities among themselves to see if there is any fluctuation between the two parameters for each. Gender and month are compared first, shown by fig. 3.

In figure 3, there are no dots plotted between points 4 to 10 on the Y-axis as we had data from November(11) to April(4), and the missing months of data correspond to the missing dots. We see that there is no significant variation in figure 3; purchases have been made in every month. 0 depicts females on the X-axis, and 1 illustrates male on the Y-axis.

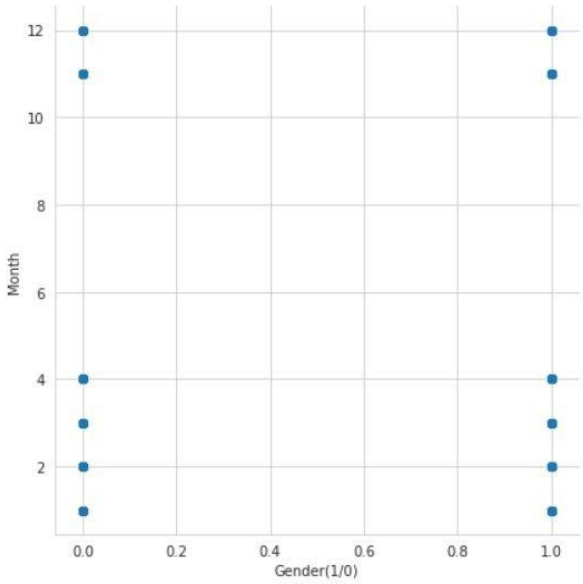

Fig. 3. Covariance of gender and month parameters. Scale: X-axis (Gender) = 0.2 and Y-axis (Month) = 2.

After the first covariance, we check for gender and quantity. In fig. 4, on the Y-axis, dots are plotted, which shows the number of tablets bought. 0 depicts females on the X-axis, and 1 illustrates male on the Y-axis.

After the second covariance, we check for the final covariance between quantity and month. Fig. 5 depicts the number of medicines bought plotted as dots in a particular month.

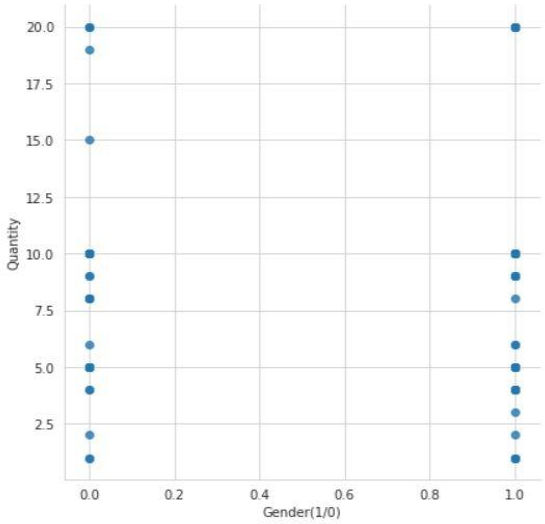

Fig. 4. Covariance of gender and quantity parameters. Scale: X-axis(Gender) = 0.2 and Y-axis(Quantity) = 2.5.

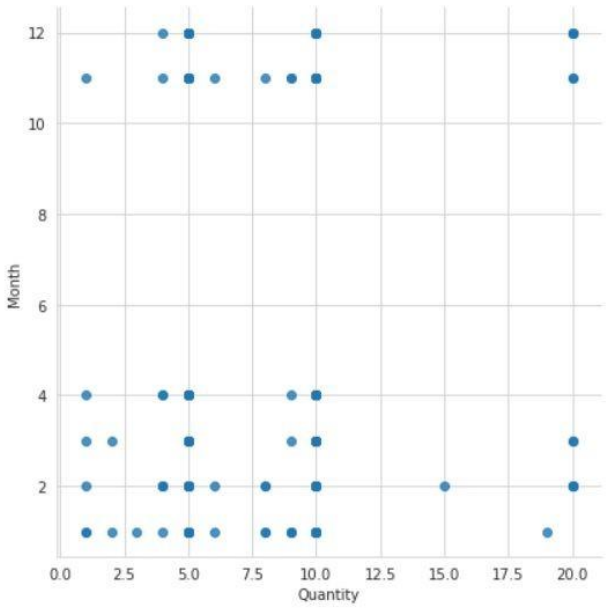

Fig. 5. Covariance of quantity and month parameters. Scale: X-axis(Quantity) = 2.5 and Y-axis(Month) = 2.

Since the covariances are successfully estimated, the quantity of medicines purchased is determined by plotting histograms for quantifiable measures.

In fig. 6, it is observed that 5 and 10 tablets are the most frequently bought number of medicines for a single purchase. This helps estimate the typical illness for which the antibiotics are prescribed. However, this does not convey whether the patients are following through with consuming the medication after the purchase is made.

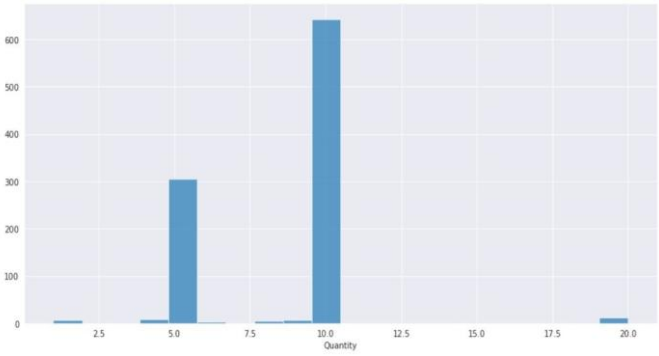

Fig. 6. Antibiotics – the number of tablets bought in a typical purchase. Scale: X-axis (Quantity)=2.5 and Y-axis (Max number of repetitions)=>120.

Since the quantity histogram is visualized, the same needs to be done for the month parameter, as shown below in fig. 7.

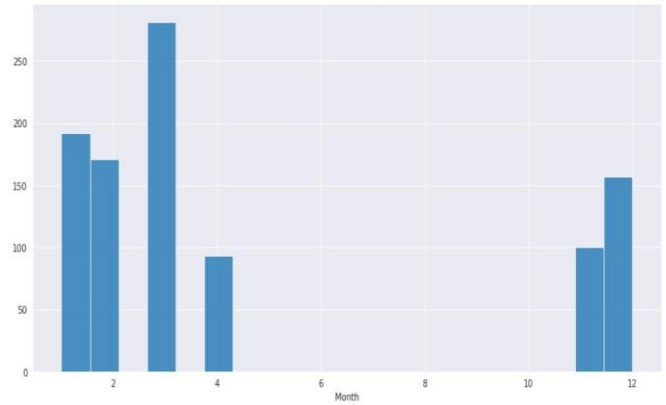

Fig. 7. Antibiotics – a month in which most purchases occurred. Scale: X-axis (Quantity)=2.5 and Y-axis (Max number of repetitions)=>120.

March has witnessed the most purchases out of the given six months of data.
Finally, the third parameter of gender is now plotted, similar to the previous two histograms, as shown below in fig. 8.

For antibiotics, this histogram shows which gender has made more purchases on the whole. The male gender has purchased more of this medication than the female gender.

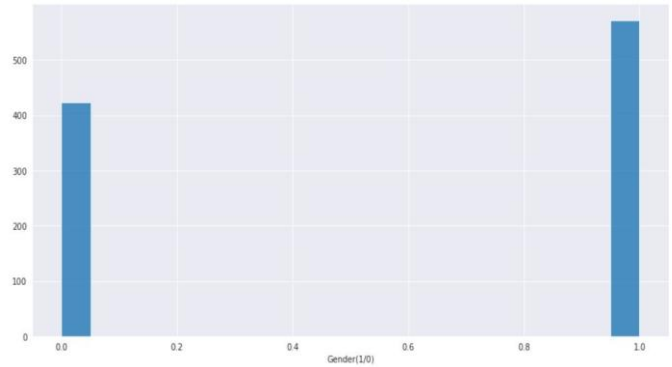

Fig. 8. Antibiotics – the gender which purchased more. 0 denotes female, and 1 denotes male. Scale: X-Axis = 0.2 and Y-axis = Max. items in each dataset.

The covariances and the histograms have been computer and visualized by K-Means Analysis. Further for example to find out a specific quantity range, the number of instances where a particular medicine has been bought in quantity exceeding ten units can be computed.

The following table in fig. 9 lists the instances where a drug has been bought more than ten times. This quantity can be adjusted accordingly for the required number while running the code. Augpen 375 is purchased the most in amounts exceeding ten units.

Name	Month	Gender(1/0)	Quantity
Augpen 375	11	1	20
Augpen 375	11	1	20
Augpen 375	12	1	20
Augpen 375	12	1	20
Augpen 375	12	1	20
Augpen 375	12	1	20
Augpen 375	1	0	19
Augpen 375	2	1	20
Augpen 375	2	1	20
Augpen 375	2	0	20
Augpen 375	2	0	15
Augpen 375	2	0	20
Augpen 375	3	1	20
Zanocin 200	3	0	20

Fig. 9. A table is indicating medicines bought in excess of 10 units.

For the purpose of preprocessing the CSV files to fit into an array, we scale the data from their current values and transform them into a min-max set. This preprocessing is done by using sklearn library and the dataframe is transformed. Fig. 10 displays the output after the preprocessing is finished.

```
array([[0.90909091, 0.        , 0.47368421],
       [0.90909091, 0.        , 0.47368421],
       [0.90909091, 0.        , 0.47368421],
       ...,
       [0.27272727, 0.        , 0.21052632],
       [0.27272727, 1.        , 0.21052632],
       [0.27272727, 1.        , 0.47368421]])
```

Fig. 10. Preprocessing of data into an array.

The array transformation is required for the algorithm to plot the data points in a cluster effectively. Following the array transformation, an elbow curve is planned to find the number of clusters necessary as recommended by the algorithm. An iterative loop from 1-20 takes place as it's a general case of data. This is the optimum range for the given dimension of data. A score is created to find how well our arbitrary assignment of points compares to the original dataset in clustering. The number of clusters are plot on x-axis and the score on y-axis.

As visualized in figure 11, the Elbow Curve breaks at Number of Clusters(NC) = 2, almost equal to 3. The change in points are points of inflection and confirm that these are the critical points by which we can cluster in the most optimal nature.

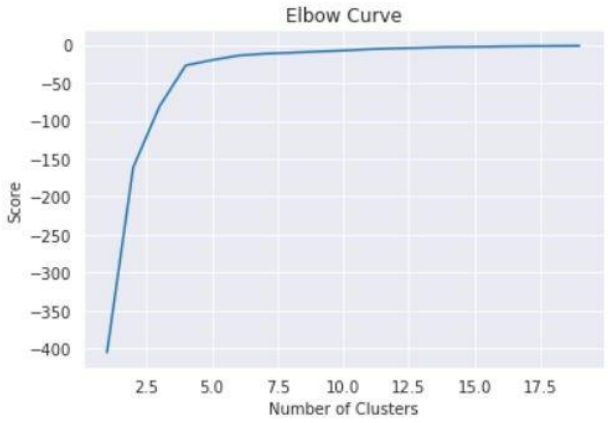

Fig. 11. The number of clusters formed by Elbow Curve. Scale: X-Axis = required clusters and Y-Axis = score of units.

Since the number of clusters has been confirmed by the elbow curve, K-Means clustering algorithm is now used with a maximum iteration set at 300 and then predicted.

In fig.12, Plotting yields useful information about which gender is buying which cluster of medicine in a month. Cluster 1 (green) indicates the drug bought the most while cluster 2 (red) and cluster 3 (blue) indicates the successive clusters. 0 on the X-Axis indicates the months. On the Y-axis, 0 indicates females and 1 shows a male. Therefore, a green dot on the Y-axis' 1' would mean the most bought medicine for males in that month.

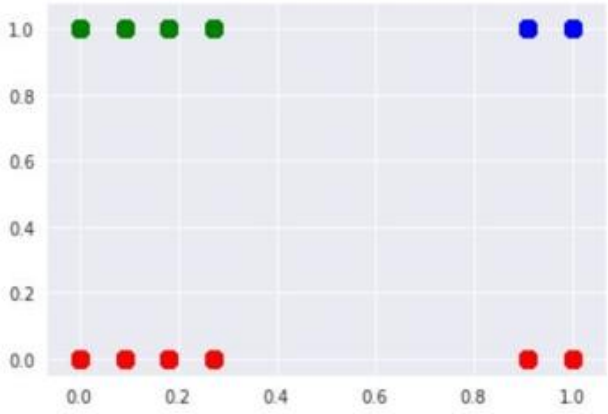

Fig. 12. Scatterplot of three clusters of medicines. Scale: X-axis = month and Y-axis = gender (0 being female and 1 being male)

The clustering is verified by calling upon the head section of the dataframe, and the output is shown in fig. 13.

Name	Month	Gender(1/0)	Quantity	Cluster
Augpen 375	11	0	10	0
Augpen 375	11	0	10	0
Augpen 375	11	0	10	0
Augpen 375	11	0	10	0
Augpen 375	11	1	5	0

Fig. 13. Result of clustering.

The clustering result is shown as '0' denoting the array in respective clusters, which contains the most number of data points after the array transformation. This can be verified by estimating the fit of the cluster. The final result of the cluster is shown as data points of cluster '0', which contains the most points in figure 14.

```
{0: array([ 422,  423,  424,  426,  427,  428,  429,  430,  431,  432,  433,
        434,  435,  436,  437,  438,  439,  440,  442,  443,  444,  445,
        447,  448,  449,  450,  454,  455,  456,  457,  458,  459,  460,
        461,  463,  465,  468,  469,  471,  472,  473,  474,  475,  476,
        477,  478,  479,  480,  482,  483,  484,  485,  486,  488,  490,
        491,  492,  493,  495,  498,  499,  500,  501,  502,  503,  504,
        505,  506,  508,  511,  512,  513,  514,  515,  518,  519,  520,
        521,  522,  523,  524,  525,  526,  527,  529,  530,  532,  533,
        534,  535,  536,  537,  538,  539,  540,  541,  542,  544,  545,
        546,  547,  548,  549,  550,  551,  553,  554,  555,  556,  557,
        558,  559,  560,  561,  562,  567,  568,  569,  570,  571,  572,
        573,  574,  575,  579,  580,  581,  582,  583,  584,  585,  586,
        587,  589,  591,  592,  593,  594,  595,  596,  597,  600,  602,
        603,  605,  607,  609,  610,  611,  612,  613,  615,  616,  617,
        618,  619,  620,  622,  623,  624,  625,  626,  628,  630,  631,
        633,  634,  635,  636,  637,  638,  640,  641,  643,  645,  646,
        647,  649,  650,  651,  652,  654,  656,  657,  658,  662,  665,
        667,  668,  669,  670,  671,  672,  673,  674,  678,  679,  680,
        683,  685,  686,  690,  692,  693,  694,  695,  696,  697,  698,
        699,  700,  701,  702,  704,  705,  706,  708,  710,  712,  714,
        715,  716,  717,  721,  722,  725,  726,  728,  730,  731,  732,
        733,  734,  735,  736,  737,  739,  740,  743,  744,  745,  746,
        747,  748,  750,  751,  752,  754,  755,  756,  757,  758,  759,
        760,  762,  763,  764,  765,  766,  769,  771,  773,  774,  776,
        777,  778,  779,  780,  782,  783,  786,  788,  789,  790,  791,
        792,  793,  795,  796,  797,  798,  799,  800,  801,  802,  804,
        806,  808,  809,  810,  812,  813,  814,  815,  816,  817,  818,
        819,  820,  821,  822,  823,  824,  825,  826,  827,  828,  829,
        830,  831,  832,  833,  834,  835,  836,  837,  838,  840,  841,
```

Fig. 14. The sample output of the cluster '0' showing the data points it contains in an array.

VIII. RESULTS AND DISCUSSION

K-Means analysis has been successfully implemented to the dataset with the prescribed parameters. Clustering is accomplished, and the number of required clusters for all the three categories is 3 clusters.

- The covariance is adequately measured and show that it is plotted as dots on the X, Y-axis depicting the numbers.
- The histograms show us the number of medicines bought, in which month it is purchased, and the gender which is buying more of the drugs. This data can be played around with and be modified according to the need of the organization.
- By selecting a quantity range, it is determined how many medicines are being bought in what quantities, repetitively in each category. This information can be used for the packaging of the drugs efficiently for the required number of medicines.
- The elbow curves are successfully plotted, and they determine the number of clusters to be used, therefore defining the number of clusters required and clustering according to the dataset input by the user.
- The final clusters are spaced far away from each other, showing that similar items are clustered together, telling us that K-Means has been successful.

IX. CONCLUSION AND FUTURE SCOPE

The method used is successful in determining the number of medicines bought, which antibiotics are accessible, which gender is buying more of what dose. Therefore, by applying the K-Means algorithm in the prescribed steps, gleaning useful information in the form of analytics is now possible using data mining. This provides us with crucial information and can now be used for real-world applications on pharmacies to better plan drug expenditure, wastage of medicines by knowing how many medications are usually being sold in a particular season or a month and storage. This model of gleaning business intelligence using data mining

REFERENCES

[1] Pore, S. M., (2014) "Pharmaceutical waste from hospitals and homes: Need for better strategies." *Indian journal of pharmacology*, *46*(4), 459. *(references)*

[2] Victor Roman, "Unsupervised machine learning: clustering analysis," in press.

[3] Imad Dabbura, "K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks," in press.

[4] Cleophas, T. J. (2012), "Machine Learning in Pharmaceutical Research: Data Clustering, Why so? and How so? *Journal of Pharmaceutical Sciences and Research*, *4*(11), 1964.

[5] Doust, D., & Walsh, Z. (2011). "Data mining clustering: A healthcare application."

[6] Sultana, H Parveen; Shrivastava, Nirvishi; Dominic, Dhanapal Durai; Nalini, N; Balajee, J.M. Comparison of Machine Learning Algorithms to Build Optimized Network Intrusion Detection System, Journal of Computational and Theoretical Nanoscience, Volume 16, Numbers 5-6, May 2019, pp. 2541-2549(9).