# Analyzing Absenteeism at Work using Clustering Method

**Rachita Mehta**

**California State University – East Bay**

**Class – Data Mining (BAN 620), 2018**

## I. INTRODUCTION

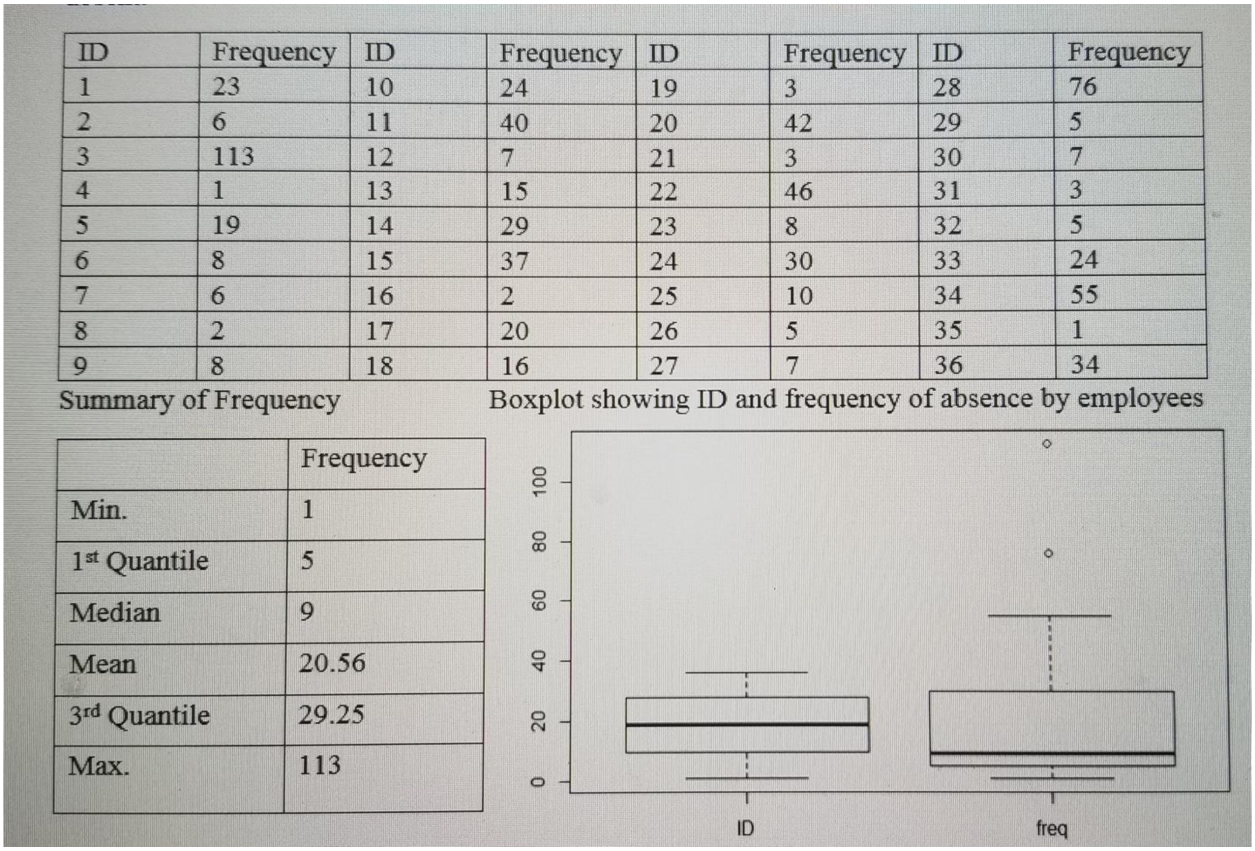
"In 2017, the U.S. Department of Labor (DOL) estimated that almost 3% percent of an employer's workforce was absent on any given day." (Society for Human Resource Management). Employee absenteeism has always been a concern to the employers. Absenteeism is the practice or the act of being away from work. There are a lot of reasons why employees choose to not show up to work. Some of the reasons are illness, unexpected emergency, medical consultation, unpleasant workplace environment, stressful work, family reasons, unjustified absences etc. High absenteeism can lead to decrease in company profits, delay in completion of projects, increase in cost to employers due to loss of productivity, reduced performance of the employee etc. Presence of each employee affects the productivity in some way. To reduce absenteeism, patterns in employees' absenteeism behavior should be recognized in order to take appropriate measures for groups of employees who share similar characteristics.

## II. LITERATURE REVIEW

Many researches has been conducted by labor economists, psychologists, data scientists etc. on employee absenteeism which go in depth explaining the reasons, consequences and solutions to reduce absenteeism at work. One of the researches on employee absenteeism, Artificial Neural Network and Their Application in the Prediction of Absenteeism by Ricardo Pinto Ferreira., Andréa Martiniano., Domingos Napolitano., Edquel Bueno Prado Farias and Renato José Sassi, makes use of machine learning concept to predict employee. This research paper makes use of Artificial Neural Network to predict absenteeism at work. This paper has been a guiding paper for this report as it helped in understanding how a data mining technique in such kind of a research.

## III. DATA

The dataset used to recognize patterns in employees' absenteeism is the same dataset used in Artificial Neural Network and Their Application in the Prediction of Absenteeism research paper. The dataset contains 740 observations and 21 attributes. The attributes are ID (employee or individual identification), reason for absence, month of absence, day of the week, seasons, transportation expense, distance from residence to work, service time, age, work load average per day, hit target, disciplinary failure, education level, number of children, whether the employee is social drinker, social smoker, number of pets, weight, height, body mass index and absenteeism time in hours. There are a total of 28 reasons for absence, out of which 21 are the diseases identifies by International Code of Diseases, and the remaining 7 some common reasons identified at work place. The entire dataset contains integer values. All categorical attributes such as reason for absence, day of the week, season, disciplinary action, social drinker and social smoker were already converted into integers. For this report all the integer values were converted into numeric values in order to run clustering algorithm. By analyzing the dataset, there were employee IDs that repeated many times, which indicated the number of times that employee was absent from work. Frequency table was used to check the number of times an employee was absent.



| ID | Frequency | ID | Frequency | ID | Frequency | ID | Frequency |
|---|---|---|---|---|---|---|---|
| 1 | 23 | 10 | 24 | 19 | 3 | 28 | 76 |
| 2 | 6 | 11 | 40 | 20 | 42 | 29 | 5 |
| 3 | 113 | 12 | 7 | 21 | 3 | 30 | 7 |
| 4 | 1 | 13 | 15 | 22 | 46 | 31 | 3 |
| 5 | 19 | 14 | 29 | 23 | 8 | 32 | 5 |
| 6 | 8 | 15 | 37 | 24 | 30 | 33 | 24 |
| 7 | 6 | 16 | 2 | 25 | 10 | 34 | 55 |
| 8 | 2 | 17 | 20 | 26 | 5 | 35 | 1 |
| 9 | 8 | 18 | 16 | 27 | 7 | 36 | 34 |

Summary of Frequency

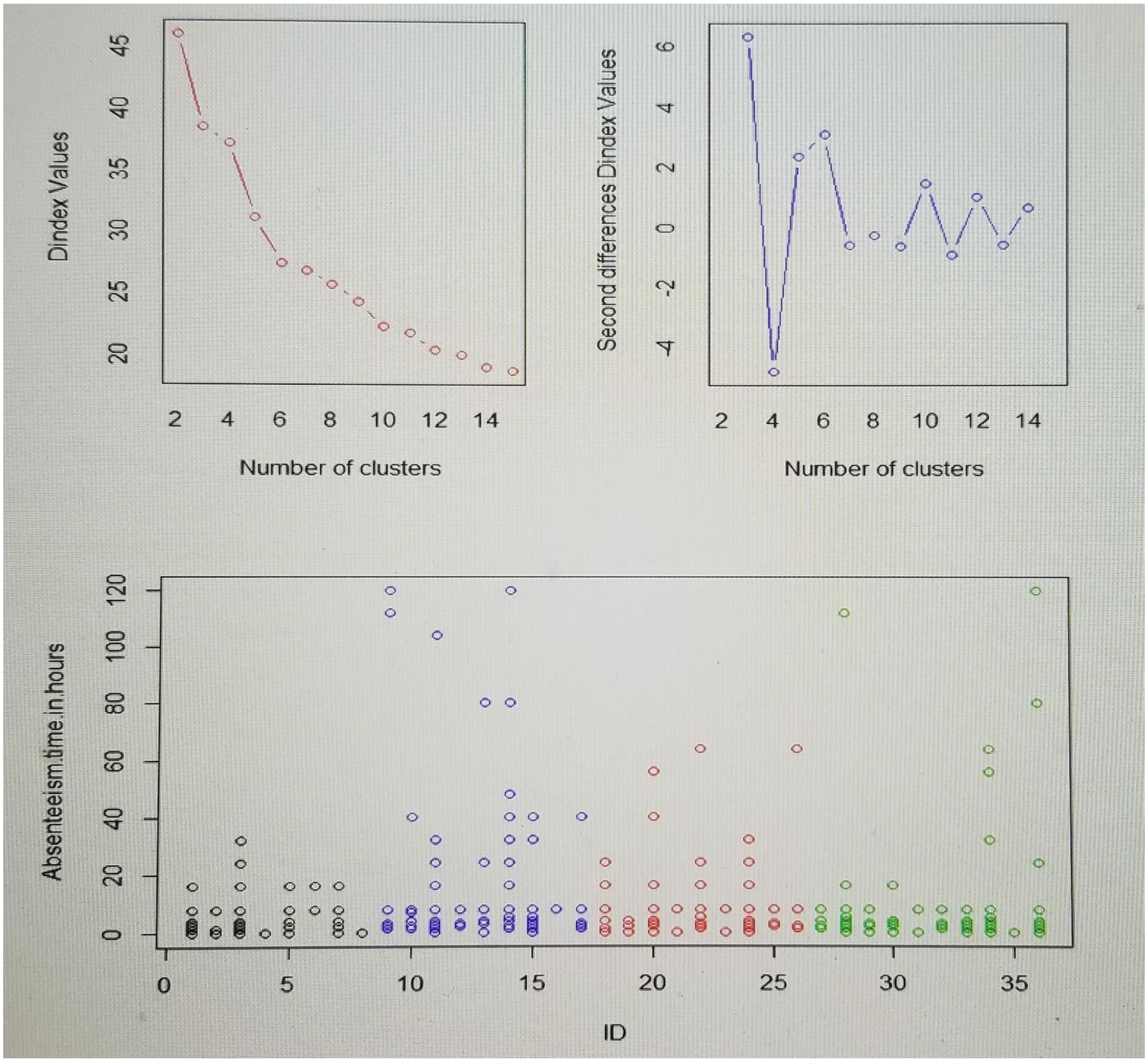
|  | Frequency |
|---|---|
| Min. | 1 |
| 1st Quantile | 5 |
| Median | 9 |
| Mean | 20.56 |
| 3rd Quantile | 29.25 |
| Max. | 113 |

Boxplot showing ID and frequency of absence by employees

0
20
40
60
80
100
ID
freq

## IV. METHOD

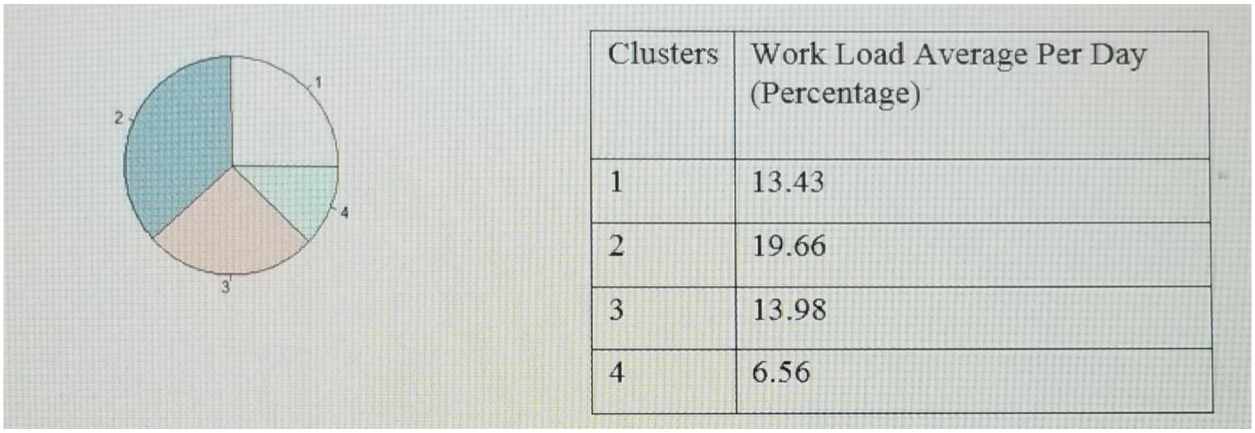
To identify patterns in absenteeism in employees, clustering method is used. Clustering is a data segmentation technique. Clustering is an unsupervised method of segmenting the data into similar characteristics. It is used when class labels are unknown. In this case the class label of employee ID is unknown. The task is to find cluster of employees who have similar absenteeism characteristics. There are many clustering methods such as hierarchical clustering, density based clustering, self-organizing maps, k-means etc. For this report k-means clustering is used. The dataset is not a large, it is small. K-means clustering is relatively easier and faster to implement than the other clustering techniques. Since k-means makes use of mean, it is assumed that cluster will have very similar characteristics. Since, k-means does not guarantee to converge to global optimum, the process was repeated 25 times to give the best results. Using k-means clustering, the best number of clusters for this dataset is 4 as depicted by the elbow method.



Using k-means clustering, 4 clusters of employee ID based on their absenteeism time in hours is plotted above. The size of Cluster 1 is 178, Cluster 2 is 163, Cluster 3 is 217, and Cluster 4 is 182. The sum of squares with the Cluster 1 is 381.62, Cluster 2 is 741.18, Cluster 3 is 2219.74, and Cluster 4 is 1025.98. The total sum of squares within the clusters is 4368.52 and between the clusters is 85396.25. Although cluster 3 is the largest cluster, cluster 2 and cluster 4 have the maximum number of absent hours. Cluster 2 and cluster 4 should be analyzed in order to find absenteeism patterns.

## V. ANALYSIS

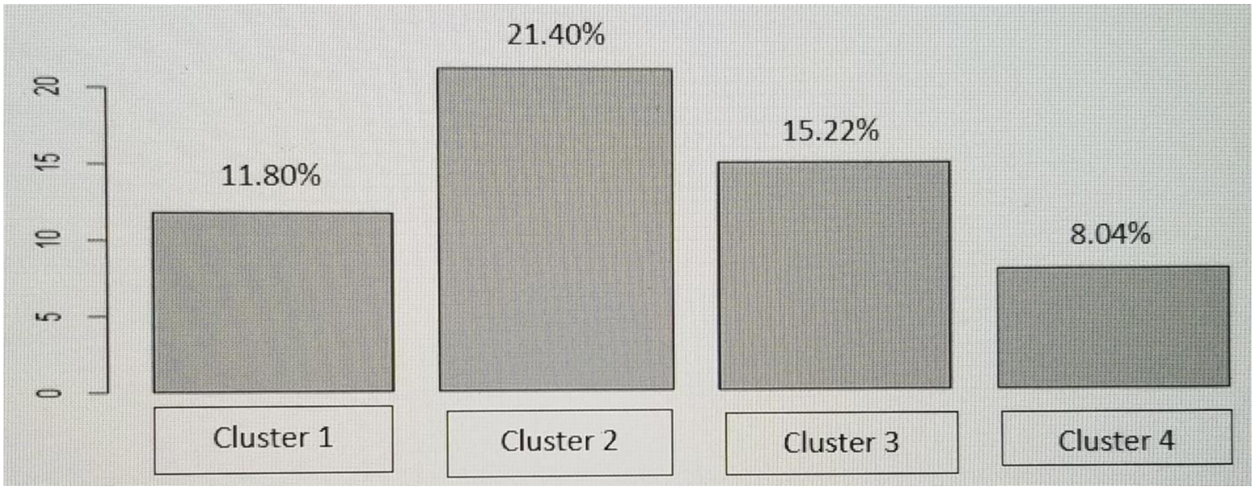
Heavy work load can stress out employees and cause them to remain absent from work. In order to analyze the absenteeism in employees, the work load percentage of each cluster group is used.



| Clusters | Work Load Average Per Day (Percentage) |
|---|---|
| 1 | 13.43 |
| 2 | 19.66 |
| 3 | 13.98 |
| 4 | 6.56 |

Employees in Cluster 2 have the maximum work load average per day. The employees in Cluster 2 can be burnt out by the amount of work load they have and therefore are absent from work. Cluster 4 has the least amount of work load and still has a lot of absenteeism. There can be something else that could be affecting it.

Absenteeism affects the performance of the employees. The below bar graph shows the percentage of targets achieved by each clusters. Despite having the maximum amount of workload, employees in cluster 2 have the highest percentage of target received.



Cluster 4 is at the least in the percentage of target hit despite having the least amount of work load average per day.

Employees may also remain absent from work if they have to incur travel expenses. The correlation of transportation expense and absenteeism time in hours is positive of 0.027584631. The percentage of transportation expenses in cluster 1 is 11.19%, in cluster 2 is 19.01%, in cluster 3 is 13.71% and in cluster 4 is 8.44%. Employees in cluster 1, 2 and 3 have more transportation expenses than in cluster 4.

In the data set, there are 28 reasons for absence. One of the reasons for absence is unjustified absence. It was very interesting to observe the results of this analysis. Clusters 1, 2, and 3 had zero percentage of unjustified absences. Only cluster 4 had 24.24% of unjustified absences. In order to find the main reason for absence in employees in cluster 2, the frequencies for reasons for absences was calculated. Medical consultation and dental consultation had the highest number of frequencies. Cluster 2 was analyzed with medical consultation and dental consultation as a reason for absence. The percentage of absence with respect to medical consultation is 34.23% and dental consultation is 29.46%.

## VI. CONCLUSION

One of the drawbacks of using k-means clustering technique is that it is sensitive of outliers. By looking at the box plot above, outliers were detected. While clustering the IDs, the outlier point would have shifted the mean. Apart from this, k-means did aid in identifying patterns in employees' absenteeism from which measures can be recommended.

Employees in cluster 2 have a lot of work load, which can burn them out. They frequently remain absent to seek medical consultation. Work load for these employees can be reduced a little bit to avoid burn out situations. They have the maximum percentage of target hit. They can hit higher targets if they are more present at the work place.

Employees in cluster 4 need more attention. These are the employees who are frequently absent without any justified reason. They have least amount of work load and have least percentage of target hit. They might be remaining absent for other reasons like work place issues, family reasons etc. Managers can improve the attendance of this group by setting clear attendance expectation, rewarding for performances, providing work life balance etc.

## VII. CITATION

Managing Employee Attendance. (n.d). Retrieved December 9, 2018, from https://www.shrm.org/resourcesandtools/tools-and-samples/toolkits/pages/managingemployeeattendance.aspx

Investopedia. (2013, July10). The Causes And Costs of Absenteeism in The Workplace. Retrieved from https://www.forbes.com/sites/investopedia/2013/07/10/the-causes-and-costs-of-absenteeism-in-the-workplace/#12e7e7733eb6

Lotich, P.(2018, September 25). 4 Tips For Reducing Absenteeism in the Workplace. Retrieved from https://thethrivingsmallbusiness.com/4-ways-to-reduce-employee-absenteeism

Cluster Analysis Using K-means Explained. (2017, February 19). Retrieved from https://codeahoy.com/2017/02/19/cluster-analysis-using-k-means-explained/

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques*. Amsterdam: Elsevier/Morgan Kaufmann.

Data – (n.d). Retrieved from https://archive.ics.uci.edu/ml/datasets/Absenteeism-at-work#

Ricardo Pinto Ferreira et al. 2018, ARTIFICIAL NEURAL NETWORK AND THEIR APPLICATION IN THE PREDICTION OF ABSENTEEISM AT WORK(Unpublished master's thesis). Nove de Julho University, Brazil. http://dx.doi.org/10.24327/ijrsr.2018.0901.1447