Demographic Characteristics of the Employees and Possible Links to Absenteeism from the Workplace

Executive Summary: This is a report that attempts to set up a connection conceivable segment credits of the staff who will in general be more powerless to miss work. Our report subtleties a thrilling examination concentrated on checking out degrees of truancy and its likely expenses in a few areas (finance, fabricating, utilities/administrations, retail, neighborliness, media, and development). All the more significantly, this concentration additionally saw expected reasons for truancy, and discoveries recommend significant connections that can be utilized to diminish non-attendance and increment efficiency. The report covers the plan, execution, and aftereffects of the review that give benchmark information on levels, expenses, and reasons for non-attendance.

Introduction: Absenteeism alludes to the ongoing non-presence of a representative at their specific employment. Ongoing non-presence stretches out past what is considered to be inside an adequate domain of days from the workplace for genuine causes like booked excursions, infrequent sickness, and family crises. Potential reasons for over-non-attendance incorporate occupation disappointment, progressing private matters, and persistent clinical issues [1]. Notwithstanding the main driver, a specialist who shows a drawn-out example of being missing may discolor their standing, which may therefore undermine their drawn-out employability. Notwithstanding, a few types of absenteeism from work are legitimately secured and can't be the reason for the end. Absenteeism is the ordinary nonattendance from a commitment without reason. In the working environment, truancy alludes to ongoing non-permitted or unannounced nonappearances by representatives. High worker truancy is an indication that a workplace needs changes. It should cause an employer to notice the nature of the board and work conditions inside an association. For instance, River is disappointed with her workspace and occupation obligations. Stream routinely phones are debilitated to work for quite a long time at a time, often missing five days every month, despite the fact that there are no genuine ongoing medical issues[2].

We tracked down the main indicators of absenteeism uncertified were:

1. View of the absence of procedural equity, low degrees of responsibility, low degrees of occupation fulfillment (authoritative issues)
2. Bad weather and individual tasks (primary issues).

We tracked down the main indicators of confirmed truancy were:

1. Actual wellbeing, and

2. Work-family/family-work clashes.

Distinguishing the reasons for truancy permits supervisors and workers, and the exploration organization to think about intercessions to address the causes. Businesses can utilize this data to zero in endeavors on those parts of the work environment that are plainly identified with non-attendance. Consolidating this with an unmistakable comprehension of the levels and expenses of absenteeism permits supervisors, managers, and the exploration organization to survey the expense/advantage compromises of different intercessions. Just as resolving hierarchical issues, the information additionally demonstrates that terrible climate and individual tasks influence uncertified non-appearance levels, and that wellbeing and work/family and family/work clashes influence confirmed levels. These issues additionally must be considered to completely address non-appearance. There might be transportation choices accessible for an awful climate, and the particular idea of the individual tasks might recommend arrangements [3].

Methodology: Nowadays employee absenteeism is the biggest problem & due to this, companies are facing various problems in the present age. Companies are looking for different solutions to get out of this problem. So companies are searching for ways to overcome their employee absenteeism problems very soon. Every company needs some analysis report to get out of this problem because only through the analysis report will they be able to understand why their employees are absent. This study collects the dataset from the SAS official data visual analytics website named ABSENTEEISM [4]. This dataset contains 740 rows and 21 columns or variables, these are representing the demography characteristics of an employee. There is age, body mass index, distance from residence to work, education, height, pets, service time, social drinker, social smoker, son, transportation expense, weight, etc. This study measures the central tendency for an overview of the data. Central tendency is a measure of central values for the probability distribution. This process shows the mean, median, and mode for every variable [5]. Another thing is the variation which is a way of showing how data is spread out. It measures the variability from the mean or the average value.

Variables are visually represented and make correlation and clustering using the SAS platform. The possibility of absenteeism depends on some important attributes of an employee. The relation or dependencies are shown by the correlation among the attributes. Some clustering divides the data into some groups for better prediction for an employee. Education and social drinkers have converted to categorical data for clustering the models created after correlation. Clustering mainly groups the data into some categories. The clustering process was done by three groups, each of the groups carried categorical data. This study converts the attribute education and social smoker to categorical data for clustering operations. There are three groups of clustering used to analyze the data after correlation. Each cluster model has categorical data for prediction analysis. The purpose of this group of clustering is to make a combination of possibilities for absenteeism. To identify which model performs better in analysis. Different data has been added to the groups according to similar activities.

Model 1 Cluster Items	Model 2 Cluster Data	Model 3 Cluster Data
Age	Body mass index	Service time
Distance from residence to work	Distance from residence to work	Distance from residence to work
Transportation expense	Pet	Height
Height	Son	Son
Education(Categorical)	Social drinker(Categorical)	Education(Categorical)

This cluster analysis was done by using a K-means clustering machine learning algorithm. K-means clustering is a distance-based algorithm where the distance has been calculated for assigning the value to a cluster. In K-means, each of the clusters has a centroid and the main goal of this clustering is to minimize the distance between the data point and centroid of a cluster [6]. Firstly, need to identify the value of K in K-means clustering using the Elbow method. Then the distance has been calculated by using the Euclidean Distance method [7]. After the clustering analysis, this study focuses on the relationship between clustering data and Absenteeism time in hours variable. This relation shows the actual dependencies of Absenteeism in the other variables of the dataset.

Result Analysis: Employee Absenteeism is a frequent absence in the workplace for some demographic characteristics. All kinds of characteristics for that absenteeism have been analyzed in this study. This study analyzes the measure of central tendency of the variables. Central tendency is a statistical measure that identifies a single value which represents the whole entire distribution. Mean, Medium, and Mode is the most common and used measure of central tendency[8]. Mean is just the average value of some group data. Median is a center value of some set of data and mode is a number that appears most of the time in a set of data. Santander deviation is the average variability of a set of data and the square of standard deviation called variation[9]. Some measures of this dataset absenteeism is below.

Standard deviation	13.33
Standard error	0.49
Variance	177.72
Coefficient of variation	192.5242

This study also discusses the relationship of the characteristics. Correlation is the most common and useful statistical analysis for describing the degree of relationship between the variables. It also represents the dependencies between the variables[10].

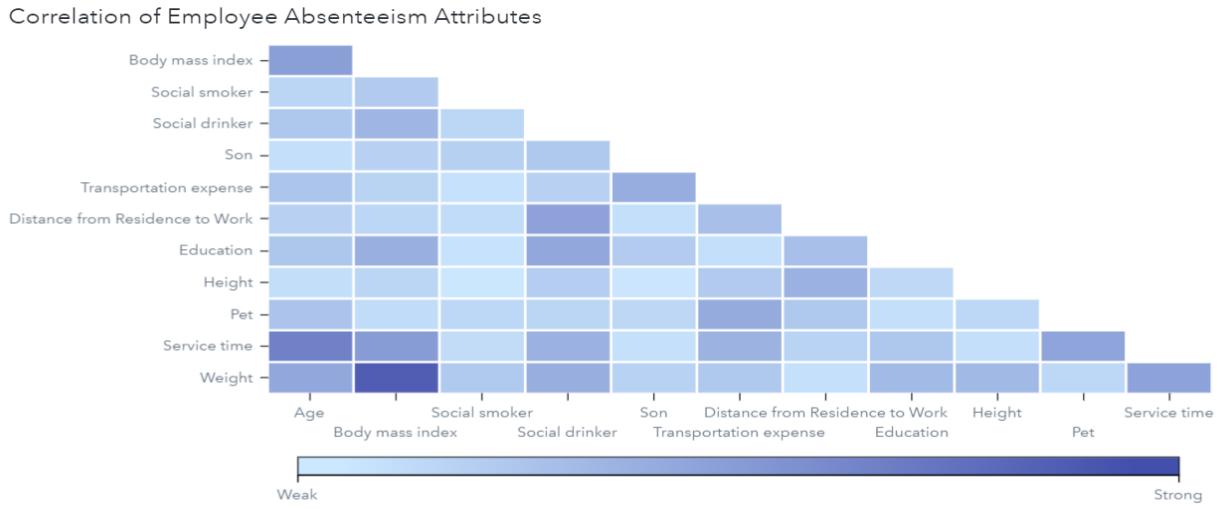


Fig 1: Correlation of Employee Absenteeism variables without the Absenteeism time

This diagram clearly represents the weak and strong relationship between the variables by deep and light color. The relation between the X-axis variables to the Y-axis variables shows in the box color. Correlation analysis becomes useful to explore the associative relationship between independent and dependent variables. All the attributes of that dataset have been added to the correlation diagram without Absenteeism time in an hour.

Absenteeism time in hour represents the number of hours an employee has been absent by employees. This frequency analysis clarifies that most of the employees' absence time is around 0 to 30 hours. A small number of employees have been absent above 30 hours. That means the reason behind absenteeism is not huge. Since most of the employees' absence time is below 20. This analysis is shown below.

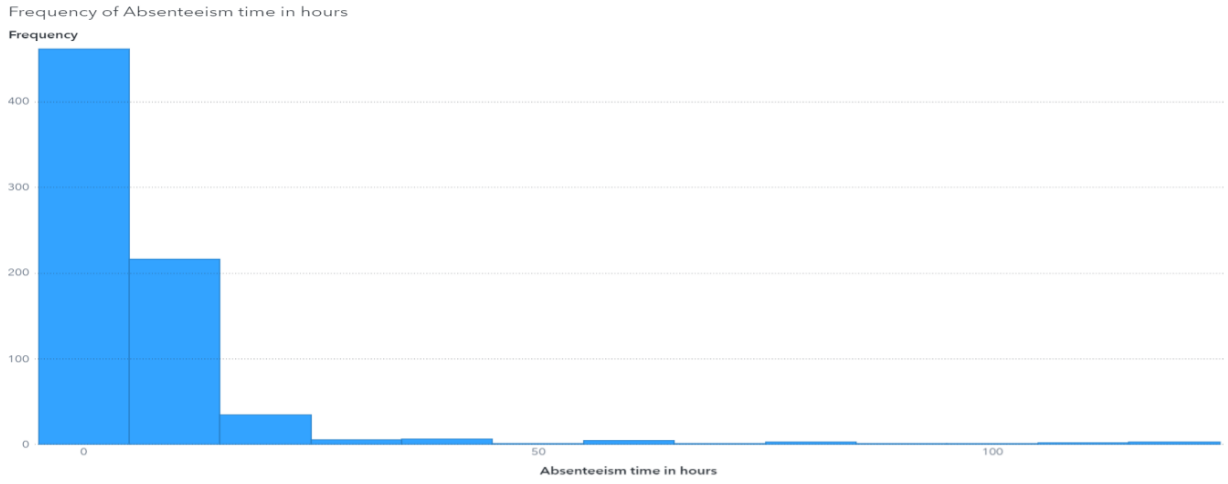


Fig 2: Frequency of Absenteeism time in hours where y-axis indicates the number of employees

The first bar represents the time between 0 to 5, the second represents time 5 to 15, the third represents 15 to 25, and so on. The diagram told us that 461 employees have only below 5 hours of absence. Then the second bar shows 216 employee absence hours between 5 to 15. This analysis indicates the smaller reasons have a bigger impact on Absenteeism.

Clustering is an unsupervised machine learning-based algorithm. Cluster analysis mainly grouping the data that has the same category data in the dataset. Especially based on the similarity between the data, it divides the dataset into many groups[11]. After grouping, this analysis separates the data points in the cluster analysis for better prediction. Clustering helps in the classification when the dataset divides into many groups and makes a prediction with new data according to the cluster analysis report.

Model 1 cluster analysis carries age, Distance from residence to work, Transportation expense, Height, and Education. In this analysis, different data points for every variable are separated using the clustering method K-means clustering. When a model goes for prediction, clustering areas or marked areas shown in the below figure helps a lot by allocating the test data.

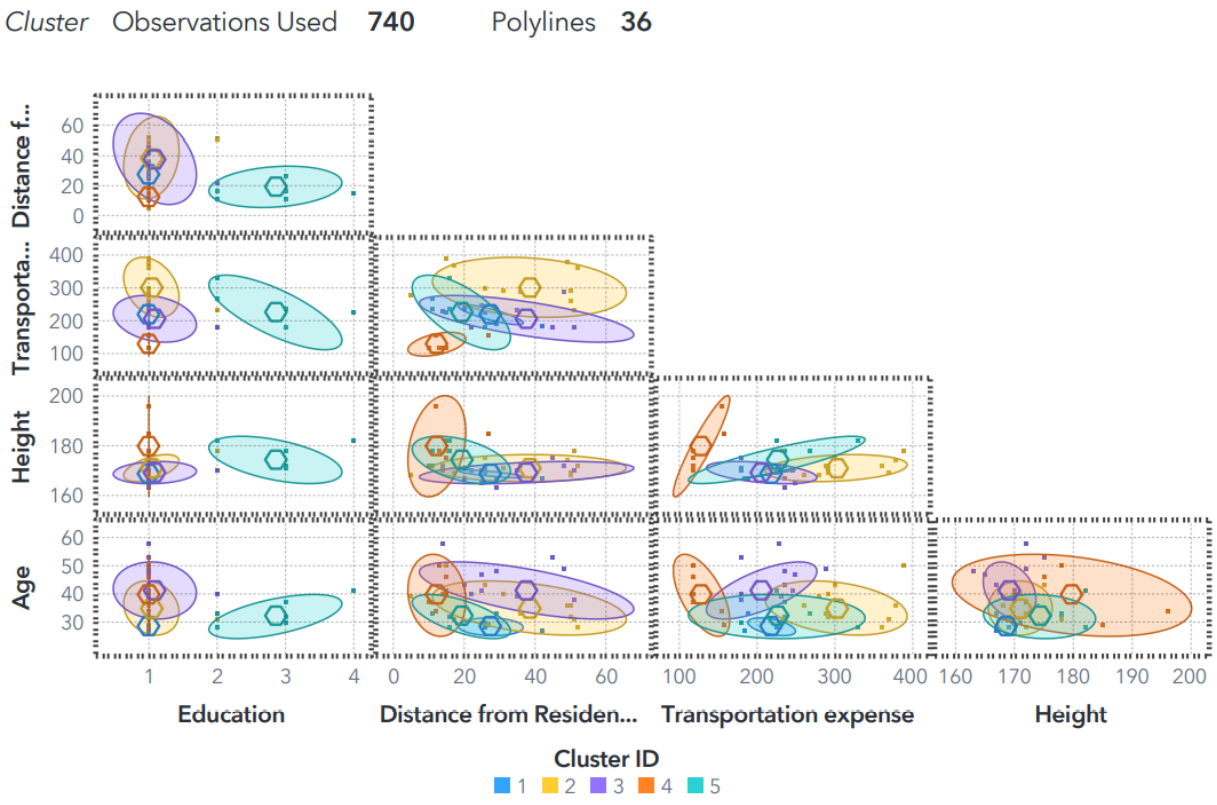


Fig 3: Cluster analysis for model 1 with some similar selected data.

In this figure, different types of data are selected with different color boundaries. That means if any new data comes for the test then easily it will be allocated as its behavior. It really helps when a model has a prediction.

Model 2 cluster analysis carries Body mass index, Distance from residence to work, pet, son, and social drinker. This analysis was also done by using the K means clustering algorithm. Cluster always helps to identify the characteristics of data. When a model can do better clustering then the performance of the model will be so high.

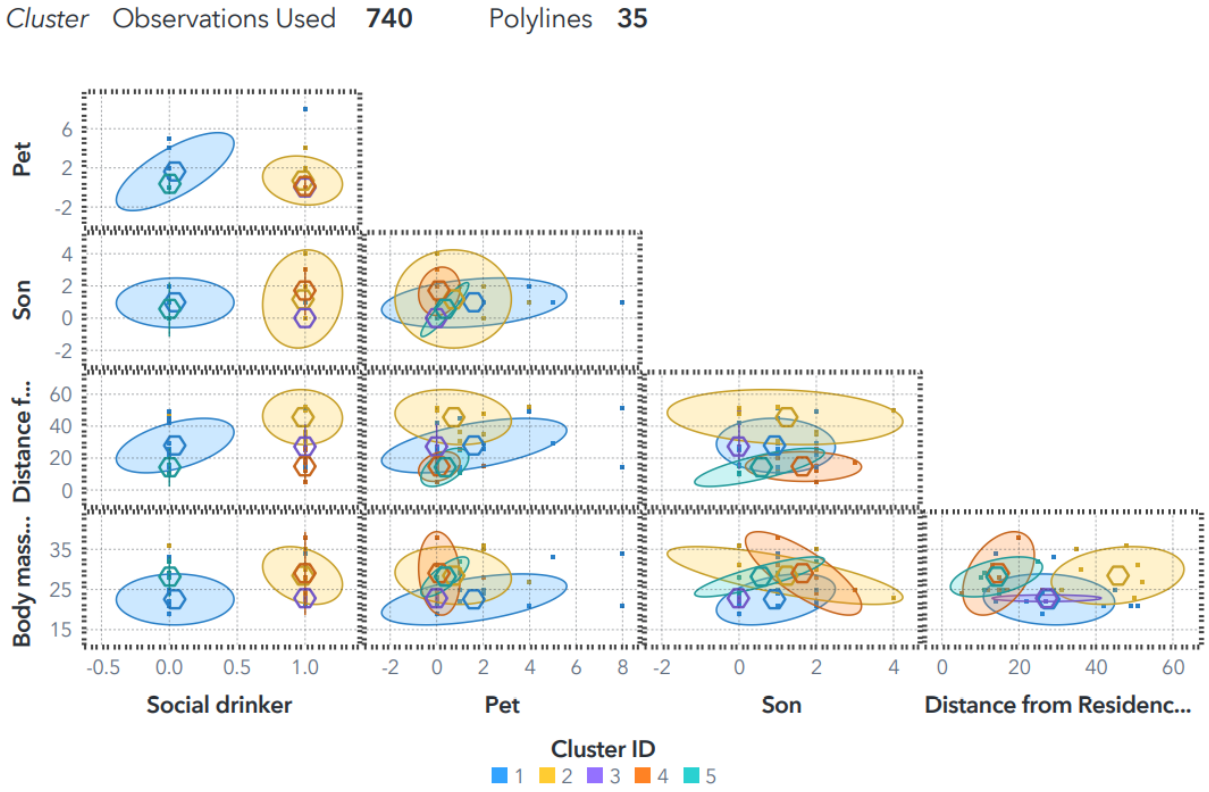


Fig 4: Cluster analysis for model 2 with selected data from the dataset.

In this cluster analysis, social drinker is categorical data and the rest are numerical. Clustering mainly happens with two variables which are shown in fig 4. Which data points are similar characteristics covered with a specific color for visually understanding the grouping.

Model 3 cluster analysis carries Service time, Distance from residence to work, Height, son, Education. It is possible to reduce the complexity of a model by doing clustering. A combination of clusters helps a model for improving prediction results.

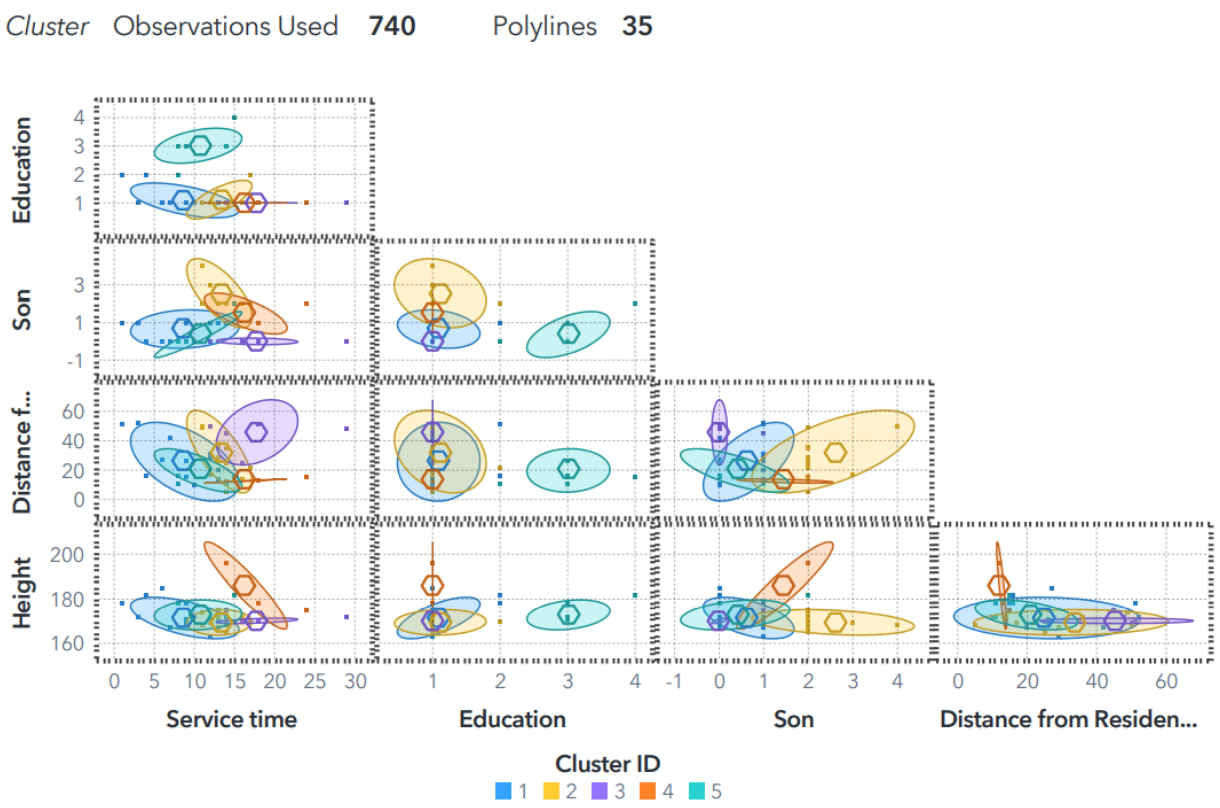


Fig 5: Cluster analysis for model 3 with some selected similar data from the dataset.

Look at this cluster analysis, Education, Son, Distance from residence to work, and Height in the y-axis and Service time, Height, Distance from residence to work, and son in the x-axis. It's a matrix for combined cluster analysis. Every different variable with other variables has a clustering result.

Compare the three models' performance according to the mean and standard deviation of each of the variables.

Model 1		
Variables	Mean	Standerd deviation
Age	12.57	4.39
Distance from residence to work	221.04	66.95
Transportation expense	2.54	1.11
Height	79.06	12.87
Education	0.05	0.05

Model 2		
Variables	Mean	Standerd deviation
Body mass index	172.15	6.08
Distance from residence to work	221.04	66.95
Pet	0.07	0.26
Son	1.3	0.68
Social drinker	1.02	1.09

Model 3		
Variables	Mean	Standerd deviation
Service time	29.67	14.85
Distance from residence to work	221.04	66.95
Height	79.06	12.87
Son	1.3	0.68
Education	0.05	0.05

Standard deviation indicates the farness of the mean. High standard deviation means that the data is more spread out from the mean. Low standard deviation means data are clustered around the value of mean[12]. If the data are clustered around the mean the performance of that model is better. So this study said that the value of standard deviation of model 2 is lower than others. That means the performance of model 2 is better than other models.

After the cluster analysis, this study combines the cluster analysis result with the Absenteeism in hour variables for result prediction. Because Absenteeism in hours is the main result variable of that dataset and this model will show the value of Absenteeism in hours in the prediction phase.

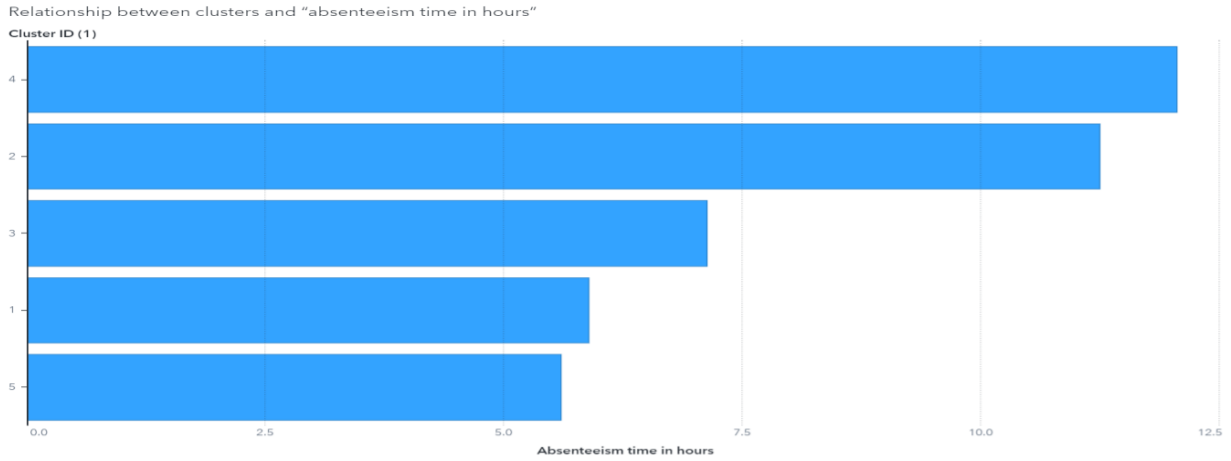
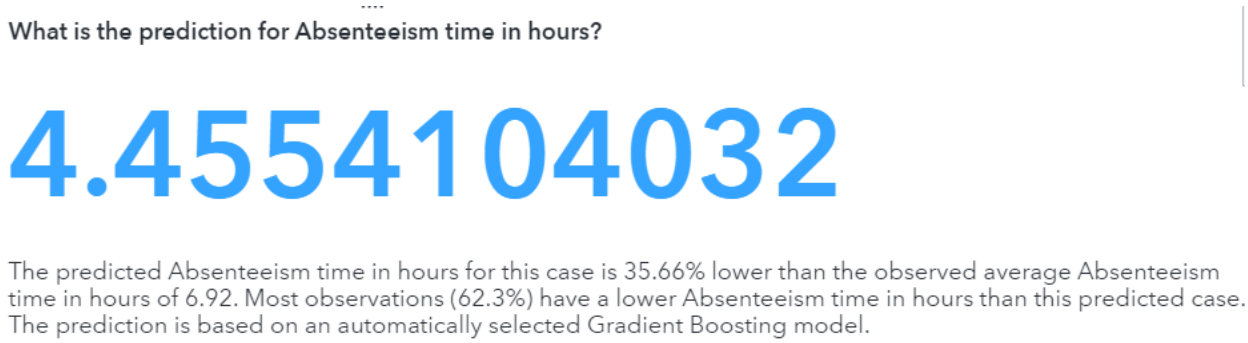


Fig 6: Relation between the clusters and absenteeism in hours variables.

This model will predict according to this diagram. After finishing correlation and clustering analysis then it produces the Absence time. This diagram shows the Absenteeism in hours value according to the number of cluster id. When the cluster id is 4 then the absenteeism time is higher (probably-12.25 hours) and cluster id is 2 then the time is second (probably-11 hours). When the cluster id is 3 then the time is third (probably-7.25 hours) and for cluster-id 1 the time is fourth (probably-6 hours). The last one is 5 cluster id and the time is fifth (probably-5.75 hours).

Now it's time to make predictions of that model using the Gradient boosting model. This prediction process takes one value for each variable. These numbers can predict the possible absenteeism in hours for a specific employee.

Name of the Variable	**value**	**Name of the Variable**	**value**
Reason for absence	23	Day of the week	2
Distance from Residence to Work	26	Body mass index	25
Age	37	Workload Average/day	264249
Weight	83	Hit target	93
Month of absence	3	Seasons	4
Height	170	Pet	0
Son	0	Disciplinary failure	0
Transportation expense	225	Social drinker	1



Conclusion: This review investigates the issue of representative absenteeism and investigates exhaustively precautionary and restorative activities. Truancy contrarily affects an organization's worker resolve. There are various projects that can be executed independently or aggregately to decrease representative non-appearance [13]. Truancy is a genuine and exorbitant issue looked at by organizations all through the world. This issue necessitates that all representatives comprehend the outcomes of such conduct from an organization's viewpoint just as an individual angle. All organizations must approach this issue from a proactive situation with representative counteraction projects and reformist discipline programs. Generally speaking, the outcomes propose that absenteeism levels, for this gathering of organizations, are not as high as some past reports. These evaluations, as noted, are probably going to be traditionalist, and genuine levels for the whole workforce might be higher than those detailed here. In any event, utilizing the moderate numbers, unmistakably the expenses related to absenteeism are high, and that bringing down the degrees of truancy would help all partners [14].

The current outcomes give proof of the main indicators of truancy and propose the issues that should be addressed to bring down absenteeism levels. The full report investigates the manners by which these issues may be tended to. Absenteeism is a difficult issue in the working environment, and companies must be proactive in searching out indications of potential reasons for non-appearance (particularly discouragement) and halting it before it gets an opportunity to begin. Indeed, even things like disappointment with the work environment climate, one of the primary drivers of sorrow in the work environment, can unquestionably somewhat be controlled by the right use of worker directing, work environment motivations, or a blend of the two[15]. The significant thing to recollect is that representatives are people, and it is important to keep them cheerful, whatever the expense.

References:

[1] https://www.investopedia.com/terms/a/absenteeism.asp

[2] https://www.aihr.com/blog/absenteeism/

[3] Journal of Vocational Behavior , Volume 10, Issue 3, June 1977, Pages 316-340

[4] https://v4e044.vfe.sas.com/SASDataExplorer/

[5]https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php

[6]https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

[7] https://sebastianraschka.com/faq/docs/euclidean-distance.html

[8] Manikandan, S., 2011. Measures of central tendency: The mean. *Journal of Pharmacology and Pharmacotherapeutics*, *2*(2), p.140.

[9] https://www.statisticshowto.com/probability-and-statistics/standard-deviation/

[10] https://www.investopedia.com/terms/c/correlation.asp

[11] Gulati, H. and Singh, P.K., 2015, March. Clustering techniques in data mining: A comparison. In *2015 2nd international conference on computing for sustainable global development (INDIACom)* (pp. 410-415). IEEE.

[12]https://www.nlm.nih.gov/nichsr/stats_tutorial/section2/mod8_sd.html#:~:text=Low%20standard%20deviation%20means%20data,above%20or%20below%20the%20mean.

[13] Razmfarsa, A., Othman, R.B., Amidi, A. and Masoomzadeh, A., 2020. An Investigation into The Impact of Absenteeism on the Organizational Workplace in Sepahan Company. *Journal of Energy and Environmental Pollution*, *1*(2), pp.29-34.

[14] Mahajan, P.S., 2015. International Journal of Advance and Innovative Research. *Advance and Innovative Research*, *2*(2), p.7.

[15]http://iaraedu.com/ijair/p5.pdf?fbclid=IwAR03mlr6dO6zl9nLoBVhe0s7yJynZcoO0NpzQS3Mzi0K4YkOEPmjn6wdrNE