Project Definition: The main objective of this project is to perform hierarchical clustering to find new knowledge from the dataset absenteeism at work. The dataset absenteeism at work contains 21 attributes and 740 instances. The data set allows for several new combinations of attributes and attribute exclusions, or the modification of the attribute type (categorical, integer, or real) depending on the purpose of the research. Which is to be put through hierarchical clustering to discover new knowledge. Used Machines:

- IBM SPSS Statistics Data Editor
- Microsoft Excel

Literature Survey: The idea of Hierarchical Clustering is to assign every object into a cluster, then repeatedly merging the closest pair of those clusters till its just one single cluster which will have multiple sub cluster. It forms like a tree with root, node and leaf clusters. This entire tree is called dendrogram.

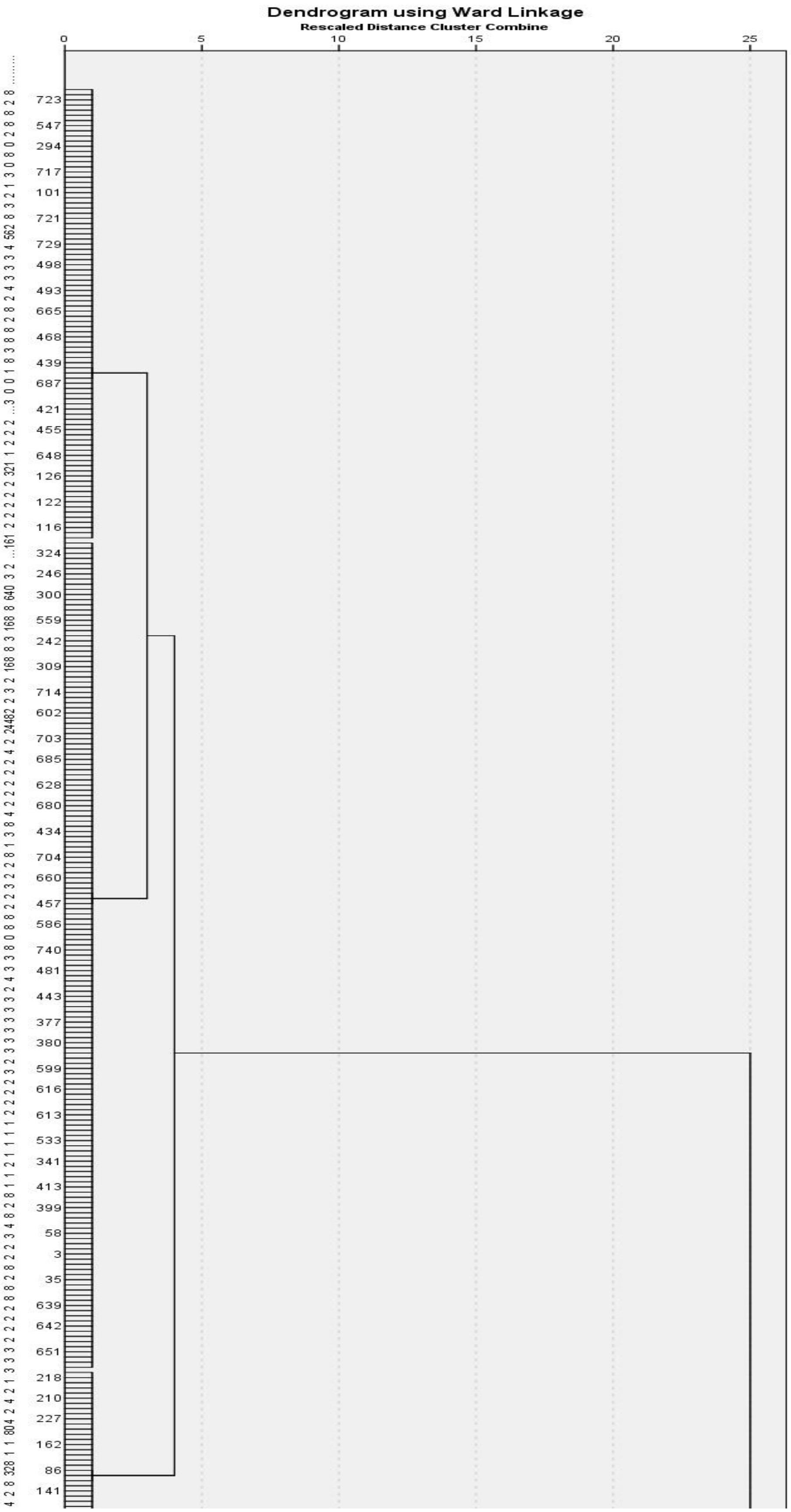
To gather data from the dendrogram, it would require to choose a sub cluster from it and take the instances under the sub cluster apart to look onto the values of those instances to find relation among them as to why they were under this single sub cluster of the entire dendrogram hoping to find some insights.

Since the data set contains its data in (.arff) format alongside with other common formats. Weka offers to create a dendrogram from the given data set via .arff file. But the dendrogram viewfinder tool is quite backdated to represent a quite large sized dendrogram that has been produced by this data set.

Not going with the path of WEKA anymore, instead a tool from IBM named IBM SPSS Statistics. Which provided dendrogram of the data set with label that could be interpreted easily. Using Ward Linkage to create the dendrogram, it was decided to analyze the dendrogram on the sub cluster group that had 8 clusters on it.

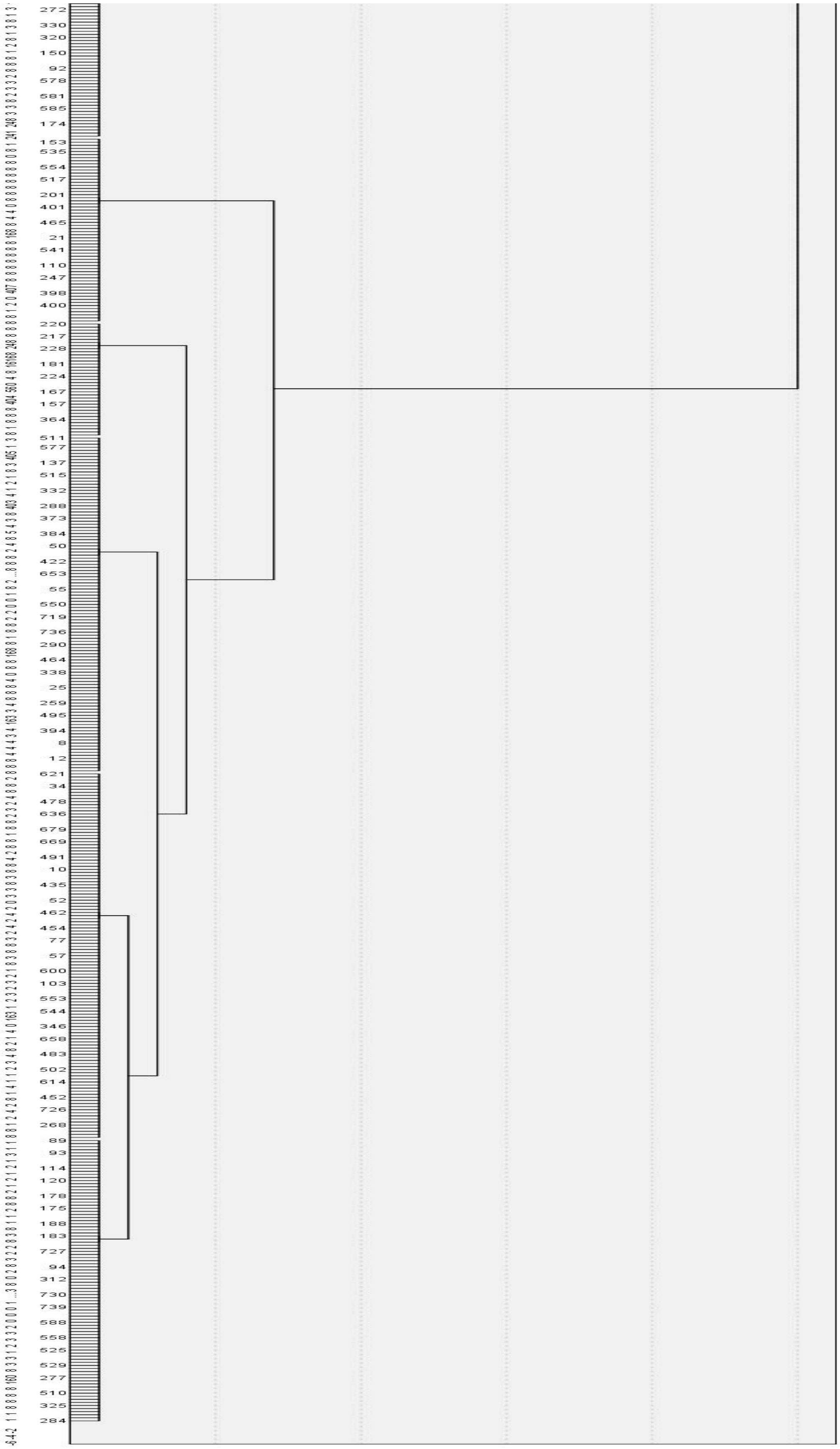
Method: The hierarchical clustering was done using the software IBM SPSS statistics data editor. Where the attribute Absenteeism at work was used to label cases by and all of the other attributes were put into variables section.

Clustering was done by cases and labeled by Absenteeism time in hours. Ward's method was used for clustering and for distance squared Euclidean distance formula was used. Weight height was excluded as there exists another instance Body mass index. Which lead to the dendrogram given below:



Dendrogram using Ward Linkage

Rescaled Distance Cluster Combine

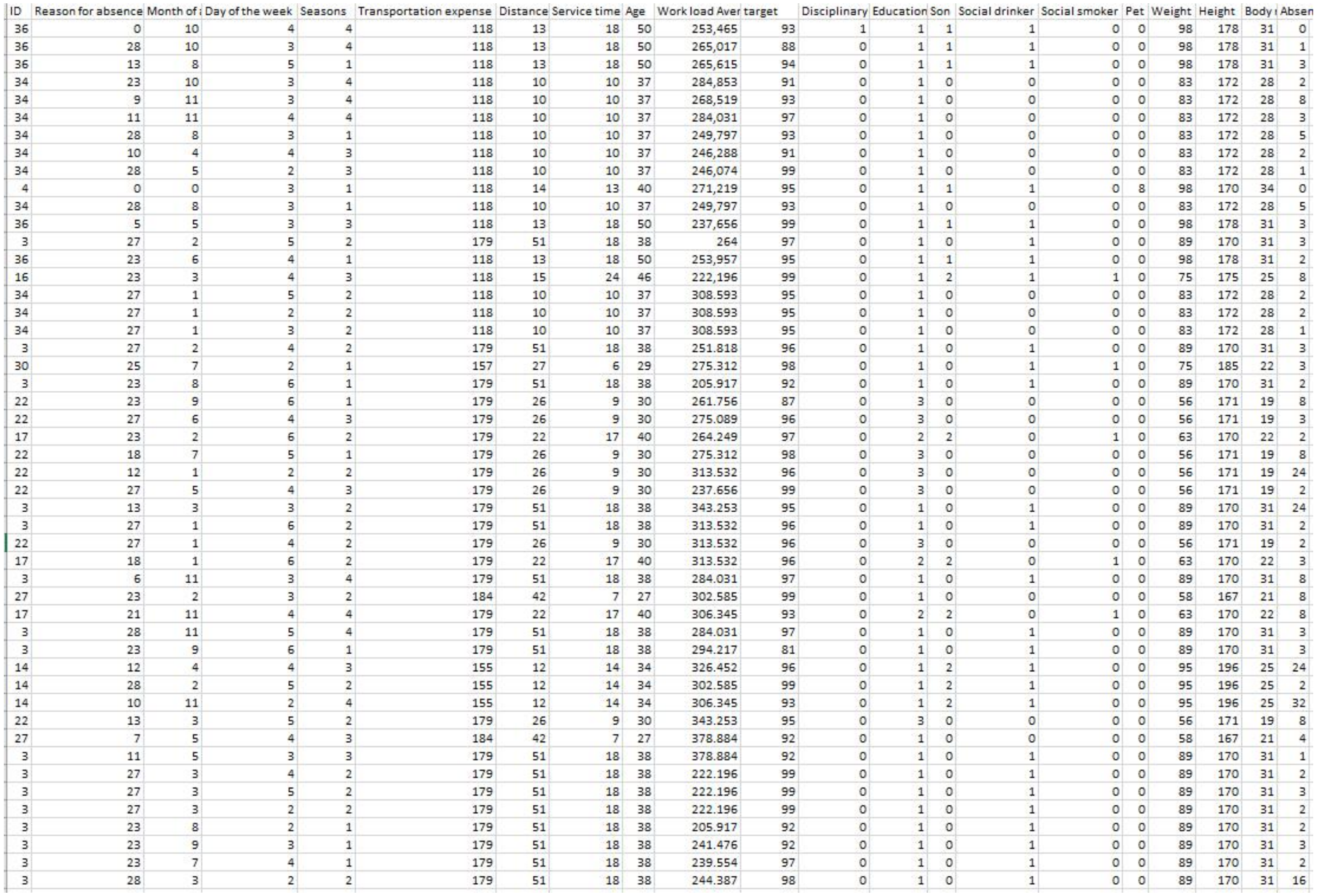

Y
272
330
320
150
92
578
581
585
174
153
535
554
517
201
401
465
21
541
110
247
398
400
220
217
228
181
224
167
157
364
511
577
137
515
332
288
373
384
50
422
653
55
550
719
736
290
464
338
25
259
495
394
8
12
621
34
478
636
679
669
491
10
435
52
462
454
77
57
600
103
553
544
346
658
483
502
614
452
726
268
89
93
114
120
178
175
188
183
727
94
312
730
739
588
558
525
529
277
510
325
284

Knowledge finding Analyzing the Dendrogram:

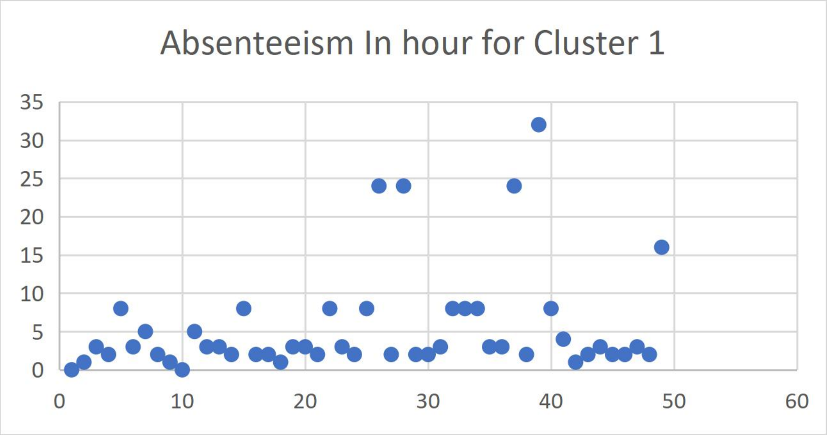
To get more precise knowledge from the dendrogram, from rescaled combined distance cluster we have cut down point at 10 where 8 clusters are combined into 2 clusters.

From the recombined cluster 1, 50 instances are taken randomly to get an optimal view labeled by Absenteeism in hours.

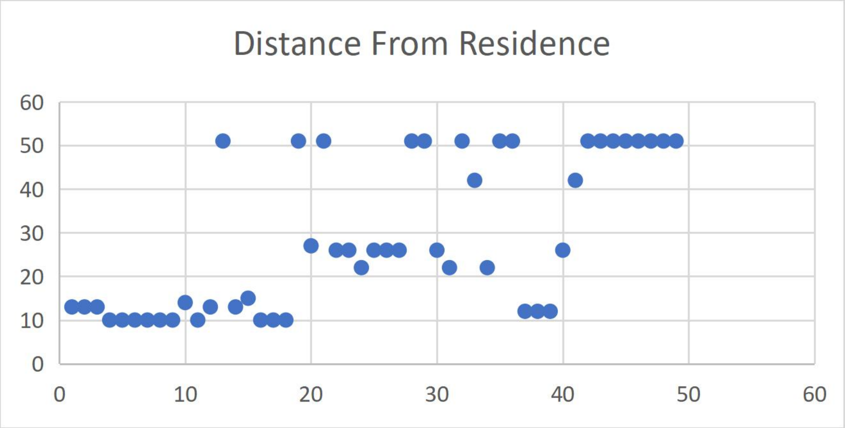
cluster1:



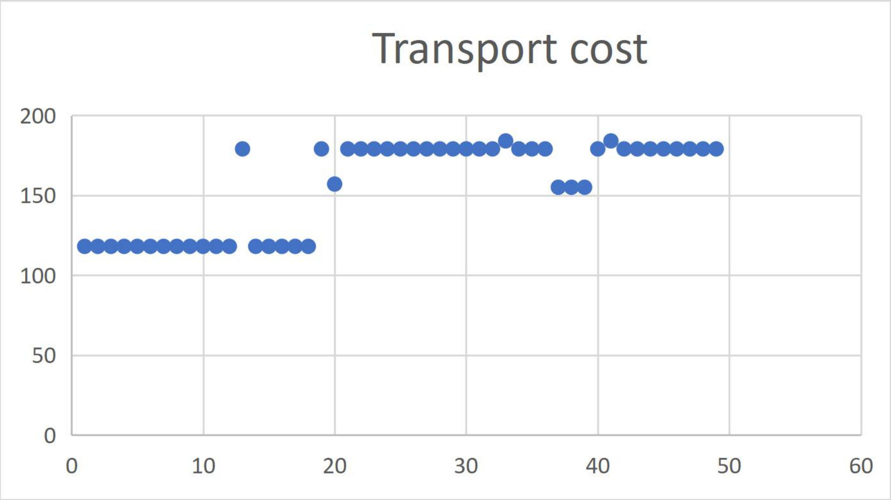
ID	Reason for absence	Month of	Day of the week	Seasons	Transportation expense	Distance	Service time	Age	Work load Aver	target	Disciplinary	Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body	Absen
36	0	10	4	4	118	13	18	50	253,465	93	1	1	1	1	0	0	98	178	31	0
36	28	10	3	4	118	13	18	50	265,017	88	0	1	1	1	0	0	98	178	31	1
36	13	8	5	1	118	13	18	50	265,615	94	0	1	1	1	0	0	98	178	31	3
34	23	10	3	4	118	10	10	37	284,853	91	0	1	0	0	0	0	83	172	28	2
34	9	11	3	4	118	10	10	37	268,519	93	0	1	0	0	0	0	83	172	28	8
34	11	11	4	4	118	10	10	37	284,031	97	0	1	0	0	0	0	83	172	28	3
34	28	8	3	1	118	10	10	37	249,797	93	0	1	0	0	0	0	83	172	28	5
34	10	4	4	3	118	10	10	37	246,288	91	0	1	0	0	0	0	83	172	28	2
34	28	5	2	3	118	10	10	37	246,074	99	0	1	0	0	0	0	83	172	28	1
4	0	0	3	1	118	14	13	40	271,219	95	0	1	1	1	0	8	98	170	34	0
34	28	8	3	1	118	10	10	37	249,797	93	0	1	0	0	0	0	83	172	28	5
36	5	5	3	3	118	13	18	50	237,656	99	0	1	1	1	0	0	98	178	31	3
3	27	2	5	2	179	51	18	38	264	97	0	1	0	1	0	0	89	170	31	3
36	23	6	4	1	118	13	18	50	253,957	95	0	1	1	1	0	0	98	178	31	2
16	23	3	4	3	118	15	24	46	222,196	99	0	1	2	1	1	0	75	175	25	8
34	27	1	5	2	118	10	10	37	308.593	95	0	1	0	0	0	0	83	172	28	2
34	27	1	2	2	118	10	10	37	308.593	95	0	1	0	0	0	0	83	172	28	2
34	27	1	3	2	118	10	10	37	308.593	95	0	1	0	0	0	0	83	172	28	1
3	27	2	4	2	179	51	18	38	251.818	96	0	1	0	1	0	0	89	170	31	3
30	25	7	2	1	157	27	6	29	275.312	98	0	1	0	1	1	0	75	185	22	3
3	23	8	6	1	179	51	18	38	205.917	92	0	1	0	1	0	0	89	170	31	2
22	23	9	6	1	179	26	9	30	261.756	87	0	3	0	0	0	0	56	171	19	8
22	27	6	4	3	179	26	9	30	275.089	96	0	3	0	0	0	0	56	171	19	3
17	23	2	6	2	179	22	17	40	264.249	97	0	2	2	0	1	0	63	170	22	2
22	18	7	5	1	179	26	9	30	275.312	98	0	3	0	0	0	0	56	171	19	8
22	12	1	2	2	179	26	9	30	313.532	96	0	3	0	0	0	0	56	171	19	24
22	27	5	4	3	179	26	9	30	237.656	99	0	3	0	0	0	0	56	171	19	2
3	13	3	3	2	179	51	18	38	343.253	95	0	1	0	1	0	0	89	170	31	24
3	27	1	6	2	179	51	18	38	313.532	96	0	1	0	1	0	0	89	170	31	2
22	27	1	4	2	179	26	9	30	313.532	96	0	3	0	0	0	0	56	171	19	2
17	18	1	6	2	179	22	17	40	313.532	96	0	2	2	0	1	0	63	170	22	3
3	6	11	3	4	179	51	18	38	284.031	97	0	1	0	1	0	0	89	170	31	8
27	23	2	3	2	184	42	7	27	302.585	99	0	1	0	0	0	0	58	167	21	8
17	21	11	4	4	179	22	17	40	306.345	93	0	2	2	0	1	0	63	170	22	8
3	28	11	5	4	179	51	18	38	284.031	97	0	1	0	1	0	0	89	170	31	3
3	23	9	6	1	179	51	18	38	294.217	81	0	1	0	1	0	0	89	170	31	3
14	12	4	4	3	155	12	14	34	326.452	96	0	1	2	1	0	0	95	196	25	24
14	28	2	5	2	155	12	14	34	302.585	99	0	1	2	1	0	0	95	196	25	2
14	10	11	2	4	155	12	14	34	306.345	93	0	1	2	1	0	0	95	196	25	32
22	13	3	5	2	179	26	9	30	343.253	95	0	3	0	0	0	0	56	171	19	8
27	7	5	4	3	184	42	7	27	378.884	92	0	1	0	0	0	0	58	167	21	4
3	11	5	3	3	179	51	18	38	378.884	92	0	1	0	1	0	0	89	170	31	1
3	27	3	4	2	179	51	18	38	222.196	99	0	1	0	1	0	0	89	170	31	2
3	27	3	5	2	179	51	18	38	222.196	99	0	1	0	1	0	0	89	170	31	3
3	27	3	2	2	179	51	18	38	222.196	99	0	1	0	1	0	0	89	170	31	2
3	23	8	2	1	179	51	18	38	205.917	92	0	1	0	1	0	0	89	170	31	2
3	23	9	3	1	179	51	18	38	241.476	92	0	1	0	1	0	0	89	170	31	3
3	23	7	4	1	179	51	18	38	239.554	97	0	1	0	1	0	0	89	170	31	2
3	28	3	2	2	179	51	18	38	244.387	98	0	1	0	1	0	0	89	170	31	16



Where we can see that, most of the workers having low absenteeism which is around 5 hours. The reason can be seen observing scatter points of Distance to the workplace from residence,

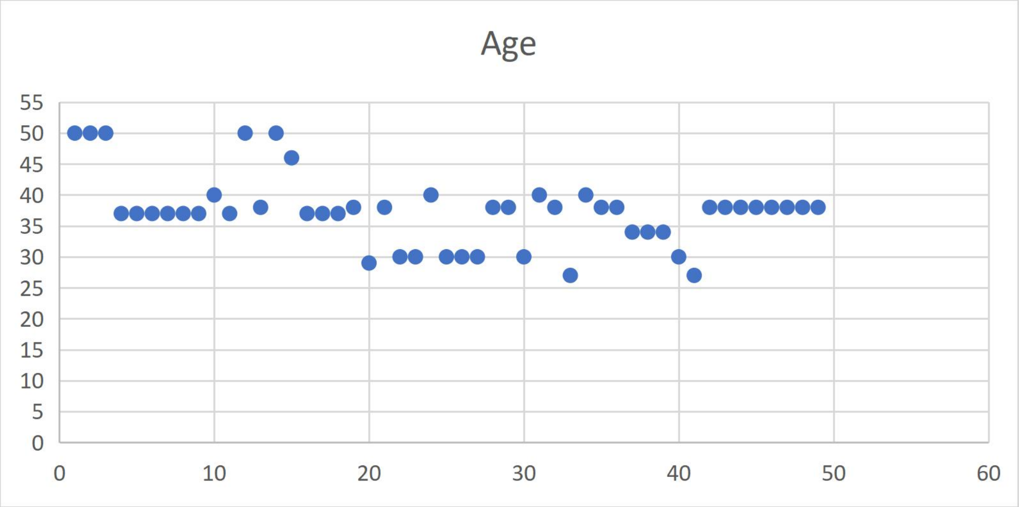


Where we can see that more than half of the workers' residence value is less than 30, though there are almost half of the workers are at a distance of about 50. To get more precise knowledge, by observing transport cost,



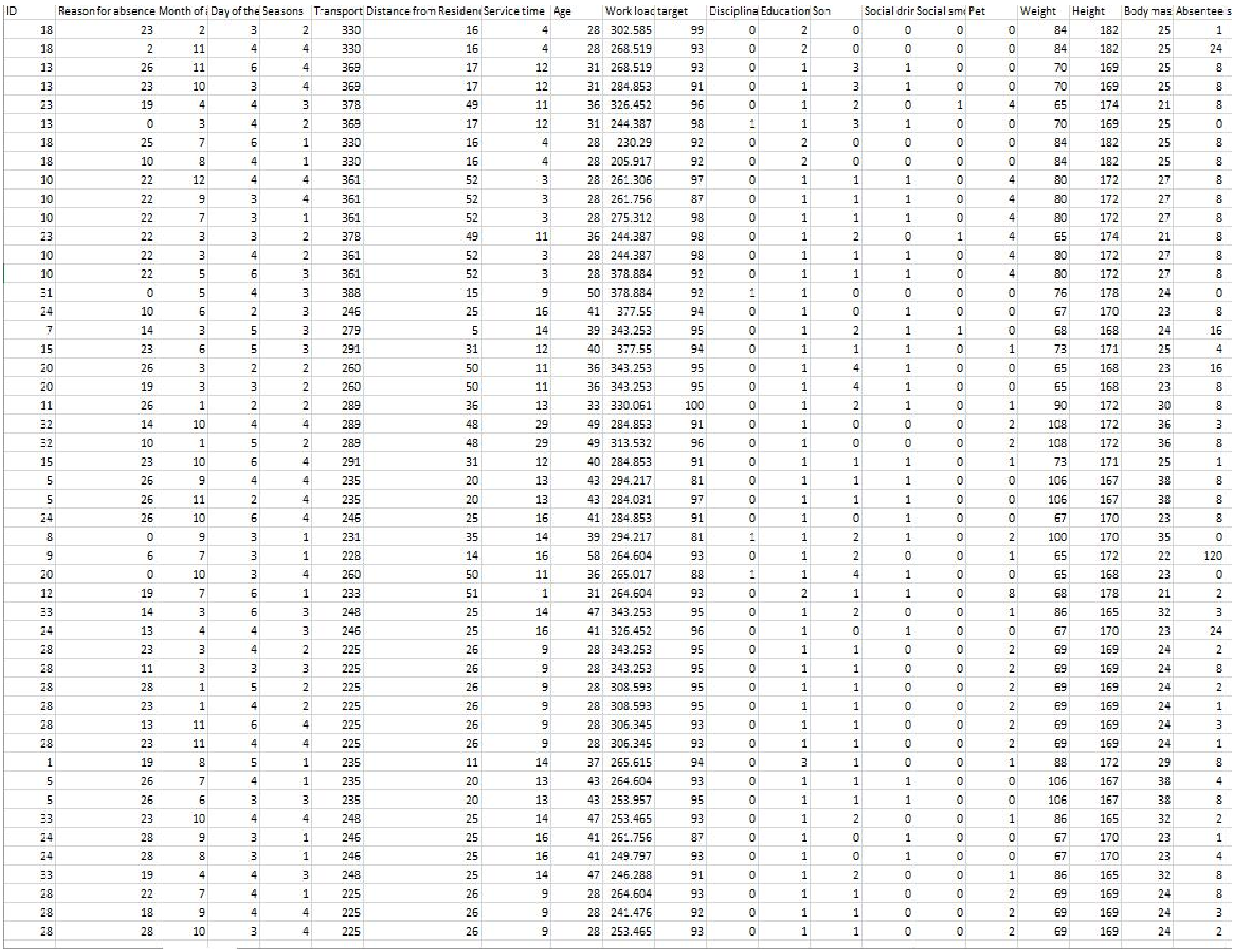
It is seen that despite of being almost half of the workers' residence distance about 50, all the workers'

Transport cost is seen below 190.which seems to be the major reason for the workers' absenteeism being such a low value. Looking from another angle we can see that,



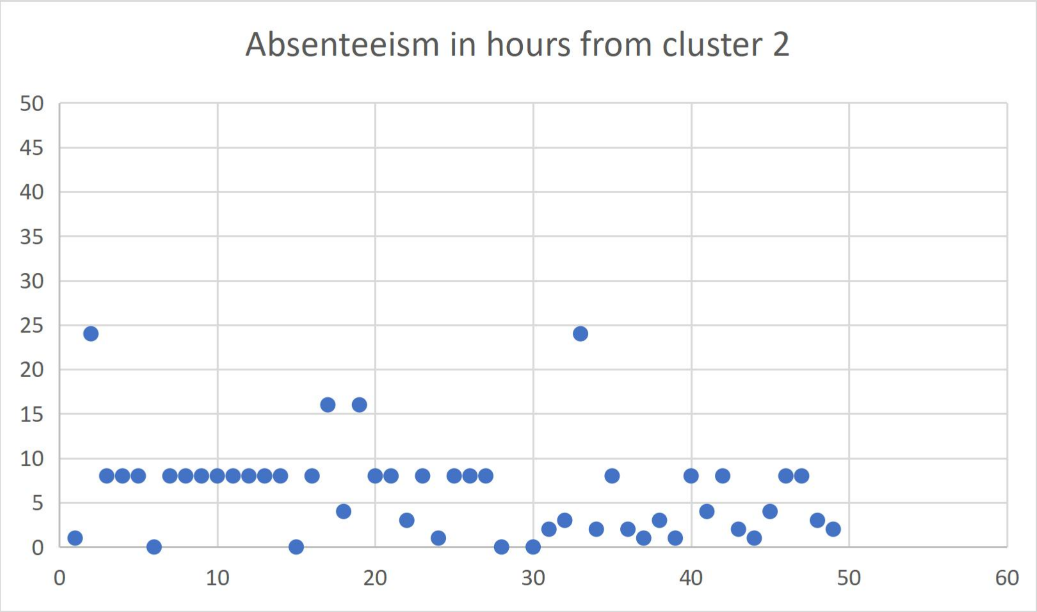
The age scatter graph shows that the workers' age of cluster 1 majority being in the range of 35-40.

cluster 2:

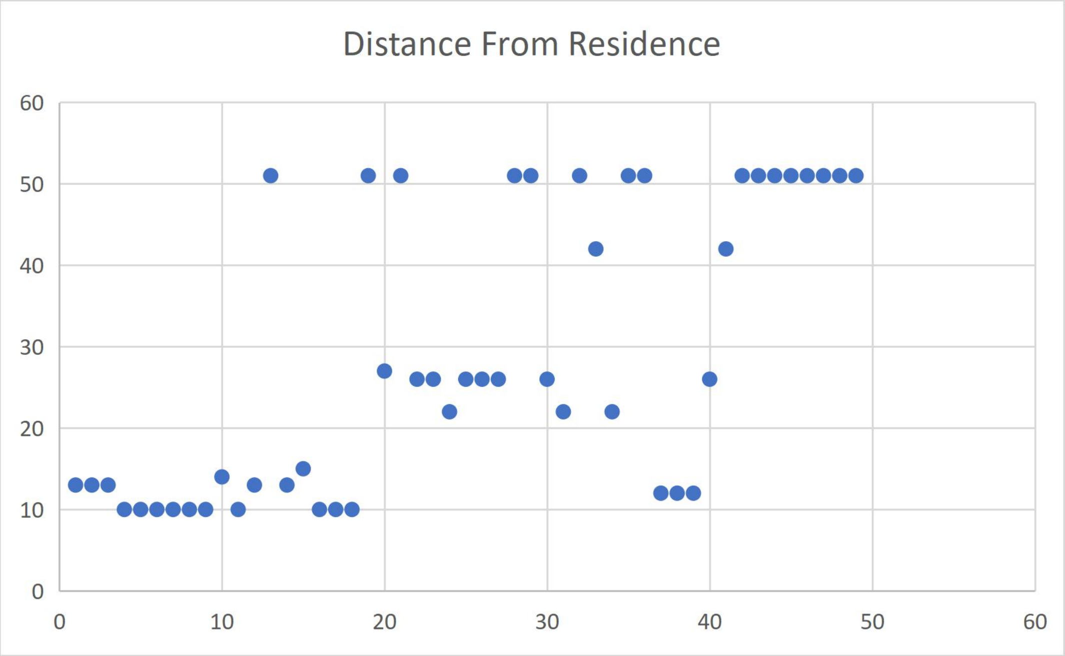


ID	Reason for absence	Month of	Day of the	Seasons	Transport	Distance from Residen	Service time	Age	Work loac	target	Disciplina	Education	Son	Social drir	Social sm	Pet	Weight	Height	Body mas	Absenteeis
18	23	2	3	2	330	16	4	28	302.585	99	0	2	0	0	0	0	84	182	25	1
18	2	11	4	4	330	16	4	28	268.519	93	0	2	0	0	0	0	84	182	25	24
13	26	11	6	4	369	17	12	31	268.519	93	0	1	3	1	0	0	70	169	25	8
13	23	10	3	4	369	17	12	31	284.853	91	0	1	3	1	0	0	70	169	25	8
23	19	4	4	3	378	49	11	36	326.452	96	0	1	2	0	1	4	65	174	21	8
13	0	3	4	2	369	17	12	31	244.387	98	1	1	3	1	0	0	70	169	25	0
18	25	7	6	1	330	16	4	28	230.29	92	0	2	0	0	0	0	84	182	25	8
18	10	8	4	1	330	16	4	28	205.917	92	0	2	0	0	0	0	84	182	25	8
10	22	12	4	4	361	52	3	28	261.306	97	0	1	1	1	0	4	80	172	27	8
10	22	9	3	4	361	52	3	28	261.756	87	0	1	1	1	0	4	80	172	27	8
10	22	7	3	1	361	52	3	28	275.312	98	0	1	1	1	0	4	80	172	27	8
23	22	3	3	2	378	49	11	36	244.387	98	0	1	2	0	1	4	65	174	21	8
10	22	3	4	2	361	52	3	28	244.387	98	0	1	1	1	0	4	80	172	27	8
10	22	5	6	3	361	52	3	28	378.884	92	0	1	1	1	0	4	80	172	27	8
31	0	5	4	3	388	15	9	50	378.884	92	1	1	0	0	0	0	76	178	24	0
24	10	6	2	3	246	25	16	41	377.55	94	0	1	0	1	0	0	67	170	23	8
7	14	3	5	3	279	5	14	39	343.253	95	0	1	2	1	1	0	68	168	24	16
15	23	6	5	3	291	31	12	40	377.55	94	0	1	1	1	0	1	73	171	25	4
20	26	3	2	2	260	50	11	36	343.253	95	0	1	4	1	0	0	65	168	23	16
20	19	3	3	2	260	50	11	36	343.253	95	0	1	4	1	0	0	65	168	23	8
11	26	1	2	2	289	36	13	33	330.061	100	0	1	2	1	0	1	90	172	30	8
32	14	10	4	4	289	48	29	49	284.853	91	0	1	0	0	0	2	108	172	36	3
32	10	1	5	2	289	48	29	49	313.532	96	0	1	0	0	0	2	108	172	36	8
15	23	10	6	4	291	31	12	40	284.853	91	0	1	1	1	0	1	73	171	25	1
5	26	9	4	4	235	20	13	43	294.217	81	0	1	1	1	0	0	106	167	38	8
5	26	11	2	4	235	20	13	43	284.031	97	0	1	1	1	0	0	106	167	38	8
24	26	10	6	4	246	25	16	41	284.853	91	0	1	0	1	0	0	67	170	23	8
8	0	9	3	1	231	35	14	39	294.217	81	1	1	2	1	0	2	100	170	35	0
9	6	7	3	1	228	14	16	58	264.604	93	0	1	2	0	0	1	65	172	22	120
20	0	10	3	4	260	50	11	36	265.017	88	1	1	4	1	0	0	65	168	23	0
12	19	7	6	1	233	51	1	31	264.604	93	0	2	1	1	0	8	68	178	21	2
33	14	3	6	3	248	25	14	47	343.253	95	0	1	2	0	0	1	86	165	32	3
24	13	4	4	3	246	25	16	41	326.452	96	0	1	0	1	0	0	67	170	23	24
28	23	3	4	2	225	26	9	28	343.253	95	0	1	1	0	0	2	69	169	24	2
28	11	3	3	3	225	26	9	28	343.253	95	0	1	1	0	0	2	69	169	24	8
28	28	1	5	2	225	26	9	28	308.593	95	0	1	1	0	0	2	69	169	24	2
28	23	1	4	2	225	26	9	28	308.593	95	0	1	1	0	0	2	69	169	24	1
28	13	11	6	4	225	26	9	28	306.345	93	0	1	1	0	0	2	69	169	24	3
28	23	11	4	4	225	26	9	28	306.345	93	0	1	1	0	0	2	69	169	24	1
1	19	8	5	1	235	11	14	37	265.615	94	0	3	1	0	0	1	88	172	29	8
5	26	7	4	1	235	20	13	43	264.604	93	0	1	1	1	0	0	106	167	38	4
5	26	6	3	3	235	20	13	43	253.957	95	0	1	1	1	0	0	106	167	38	8
33	23	10	4	4	248	25	14	47	253.465	93	0	1	2	0	0	1	86	165	32	2
24	28	9	3	1	246	25	16	41	261.756	87	0	1	0	1	0	0	67	170	23	1
24	28	8	3	1	246	25	16	41	249.797	93	0	1	0	1	0	0	67	170	23	4
33	19	4	4	3	248	25	14	47	246.288	91	0	1	2	0	0	1	86	165	32	8
28	22	7	4	1	225	26	9	28	264.604	93	0	1	1	0	0	2	69	169	24	8
28	18	9	4	4	225	26	9	28	241.476	92	0	1	1	0	0	2	69	169	24	3
28	28	10	3	4	225	26	9	28	253.465	93	0	1	1	0	0	2	69	169	24	2

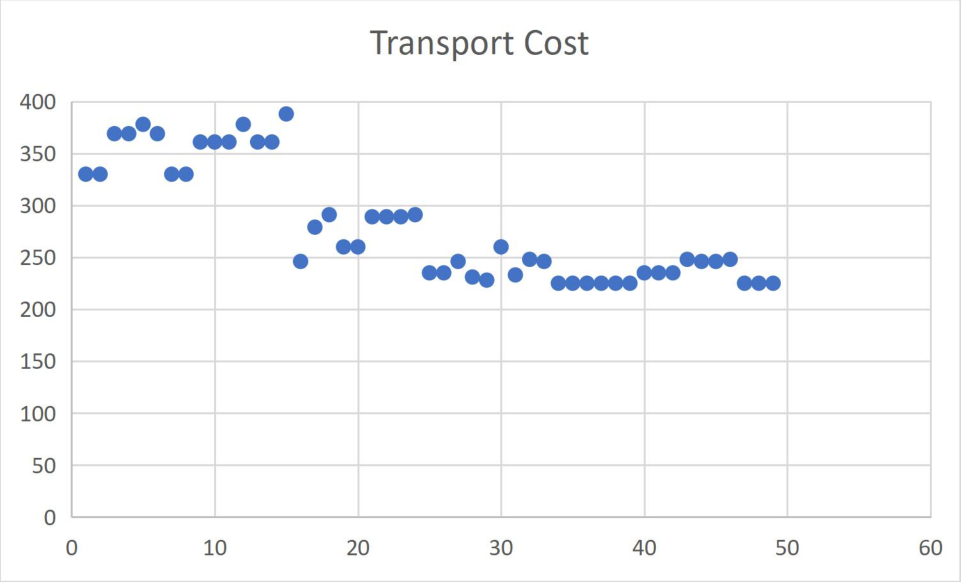
Observing the recombined Cluster 2 labelled by absenteeism in hours, from the randomly taken 50 instances from cluster 2, absenteeism values are found visualized in the scatter graph,



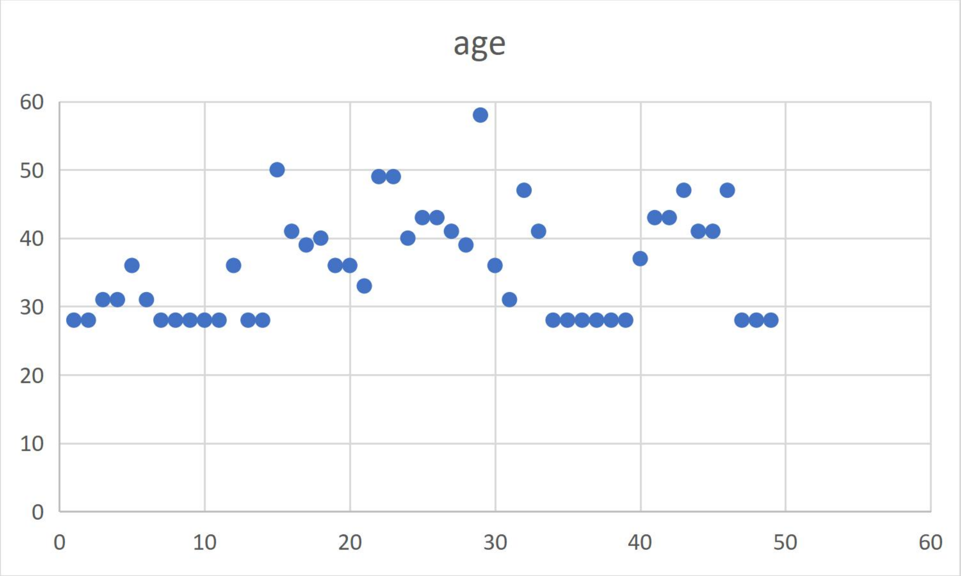
Where it seen that majority of the workers' absenteeism being above 5 hours. for the reason Same for the cluster 2 as cluster 1, by observing the attribute Distance from residence



Where above 60% of the workers' distance from workplace is above 20. though there are workers living near the workplace have absenteeism of higher value. Because from the scatter graph generated from the attribute,



Here we can clearly see that despite of being some residences close to work place, transport cost is above 200. Which seems to be the major reason for a higher absenteeism value in recombined cluster 2. following cluster 1, looking at the scatter graph generated from Age,



Workers with the higher Absenteeism value seems to be aged under 30 mostly.

Analyzing the above observations, it is clear that workers who have less absenteeism are mostly middle aged, have a service time of about 10 and have residence near to the workplaces resulting low transport expense where workers having higher absenteeism are mostly younger comparing age and service time with the workers having low absenteeism. Also, its noticeable that they have residence comparatively far from the workplace resulting higher transport expense. Coincidently, they seem to have more or less pets where workers with lower absenteeism has no pets in the records from the taken instances given in the dataset.