Absentee Profiling : Milestone Report

Proposal with problem statement

The goal is to create a profiling scheme using predictor variables detailing absences from work, such as the type of illness, month of absence, and worker age. This type of analysis would be useful to help promote a healthier workers, and help health insurance companies to create better health insurance policies, especially if health insurance companies are contracted to supply healthcare benefits to another organization's employees.

The absentee data is provided on the UCI Machine Learning Repository and has already been acquired.

We can use any clustering method (KMeans) to create clusters on our data. To select k we can use a variety of performance metrics like SS and silhouette scoring. Once we've determined k and formed the clusters, we can individually investigate each of the clusters to see if they intuitively make sense, and create labeling characteristics that define each of the clusters.

Furthermore, the ID of the worker for each absence has been recorded, so we can also investigate each worker individually (36 workers) to see if there is a pattern based on worker (as opposed to each instance of an absence).

Data collection and wrangling summary

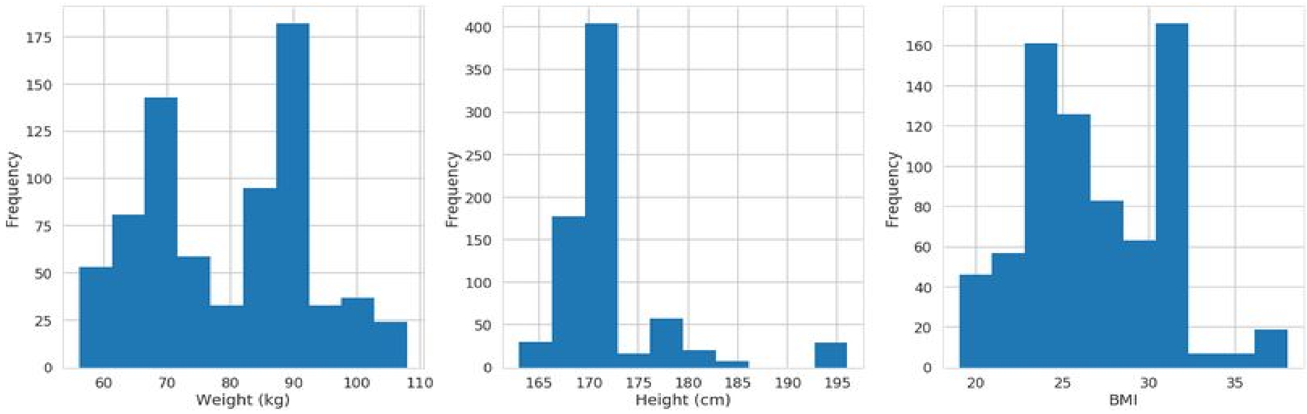
The dataset was provided by the UCI machine learning repo, and was already formatted. No additional steps were needed to clean the data.

A few of the entries for "month" were 0, so while the data wasn't explicitly missing, the entry didn't make much sense. The entries were left as 0, and were properly encoded as a row of 0's when one-hot encoding was used.

There were a few data entries that had uncommon values for a few of the features, but these entries were left as is to capture the variability in profiling.

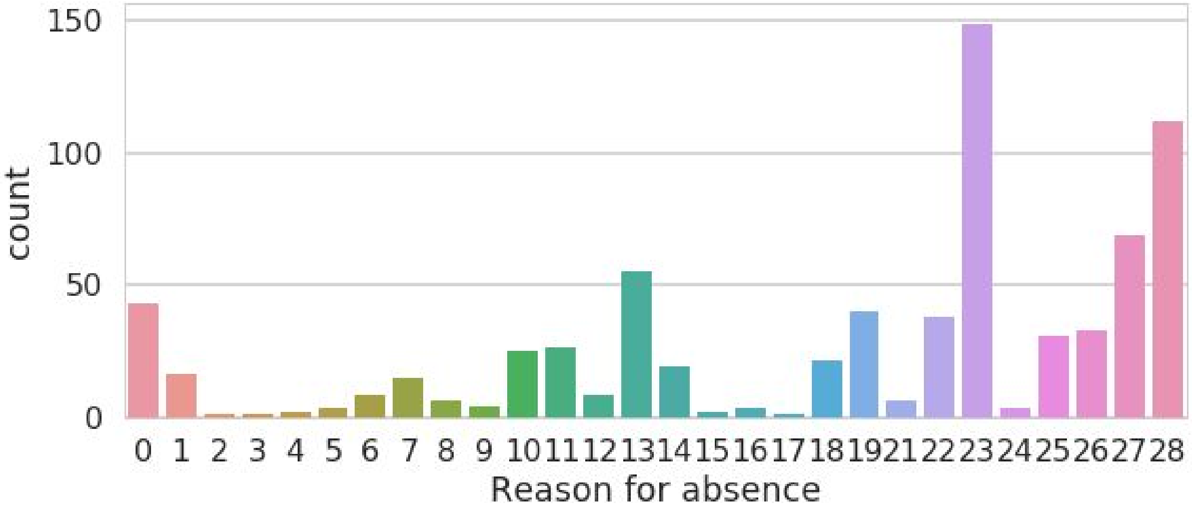
Exploratory data analysis summary

First we examine the weight, height, and BMI columns in the dataset:

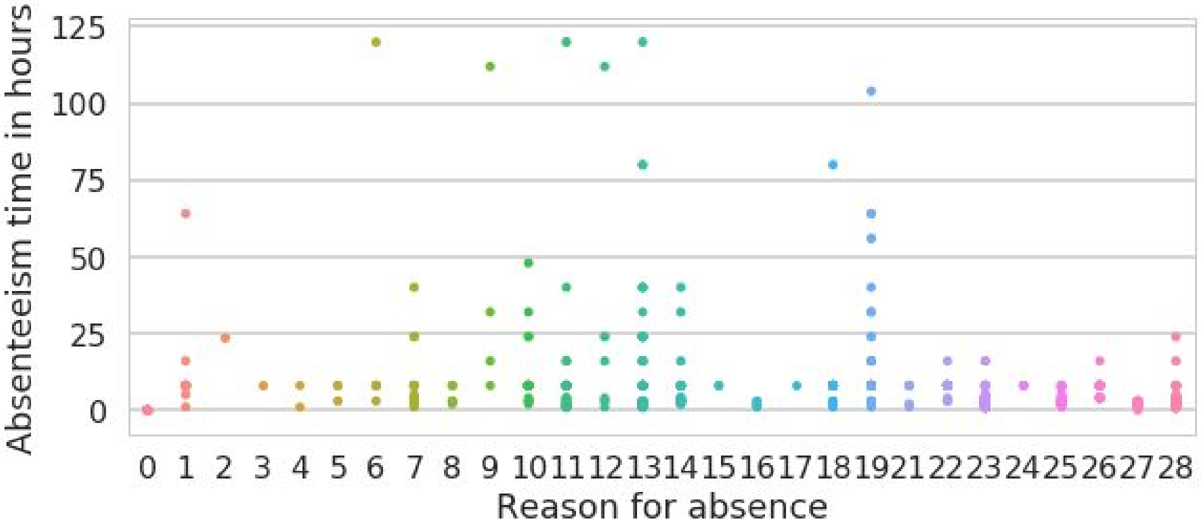


A quick look at the weight, height and BMI reveal that although we expect metrics like weight, height, and BMI to be approximately normal, both weight and BMI appear to have a bimodal curve, and height looks a bit skewed right. However, note that these frequencies are based on frequencies of absences, so a particularly lighter person may have more absences than a medium-weight person, which will cause the histogram to have more counts for the lighter weight person. In other words, these histograms are a histogram of absences, not people, and although we believe the distribution of weights across all people to be normal, we shouldn't expect the same from absences.

Let's also check out the distribution of the reason for absences:

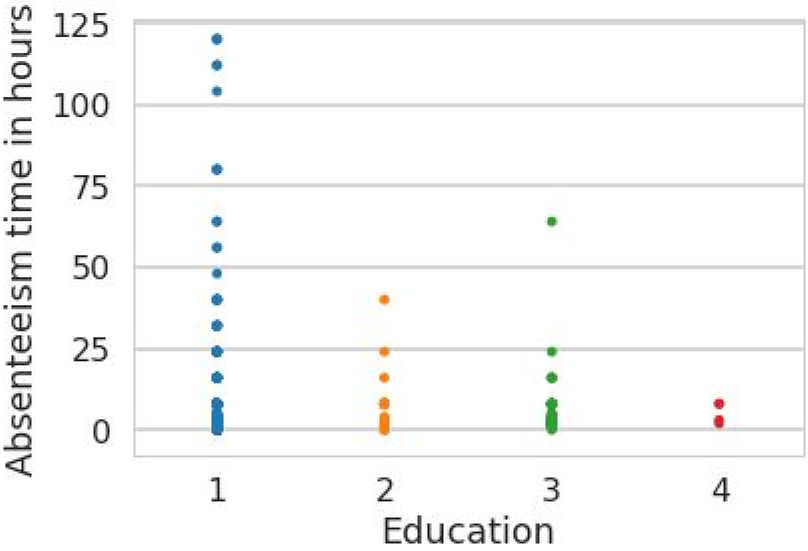


We see that medical consultation(23) and dental consultation(28) are by far the most popular reasons given, followed by physiotherapy(27), diseases of the musculoskeletal system and connective tissue (13), laboratory examinations (24) and unjustified absence(25). Now let's take a look at how long the absences were for each of these illnesses:



Referring back to the data info in the repo: the last seven reasons for absence (22 through 28) were not attested by the International Code of Diseases (ICD), while reasons 1 through 21 were. Note these last seven reasons had all absences take less than 24 hours, while the first 21 types of absences have lengthier occurrences.

Given the varying amounts of absence durations, we now ask if there is a relationship between absence length and education level. We initially investigate with this plot:



It appears that as the level of education gets higher, the length of the absence tends to go down. However, we note that this also might be due to there simply being more instances of education level 1 than 2, 2 than 3, and so on. If we were to test specific education level means, we could again use the t-test, or we could use a multiple comparison test if we wanted to simultaneously compare groups.

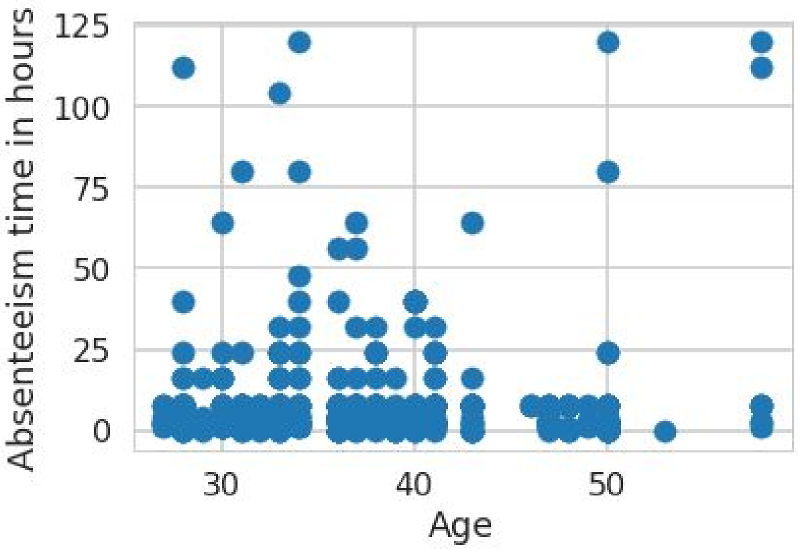
Let's conduct a t-test to see if at least one of the group means is different from the others:

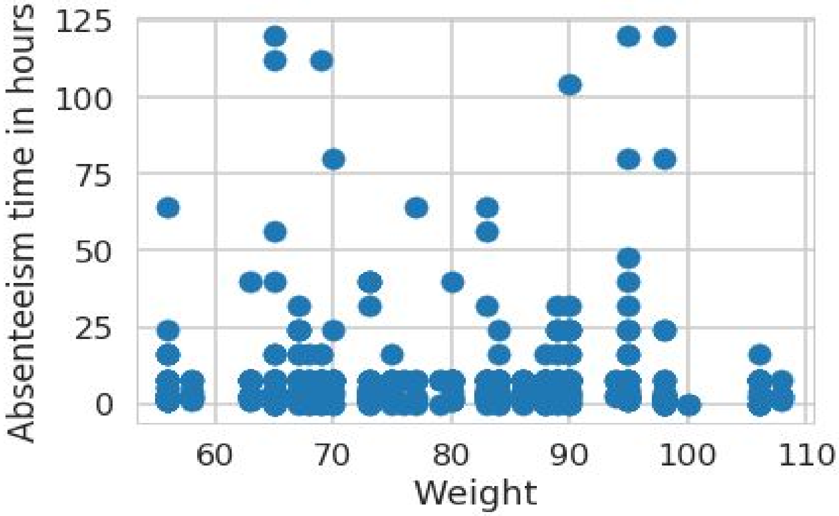
$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

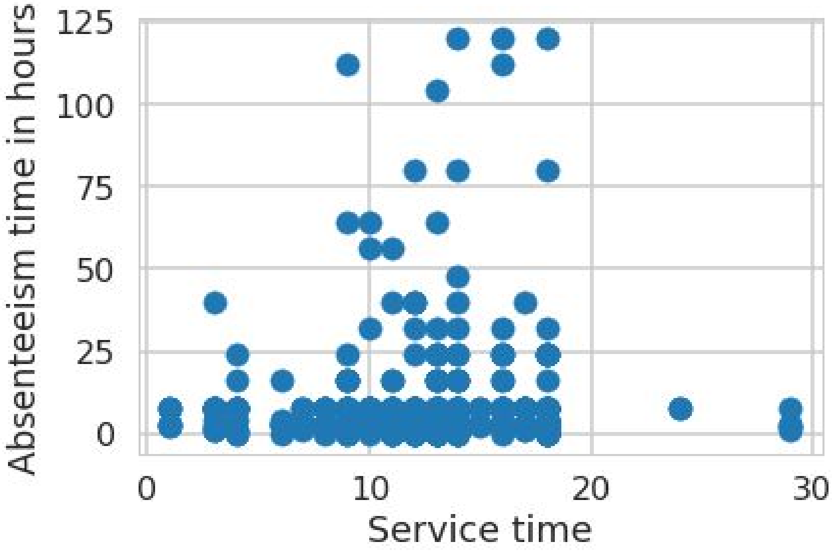
$$H_A : \text{At least one of the means is different from the others}$$

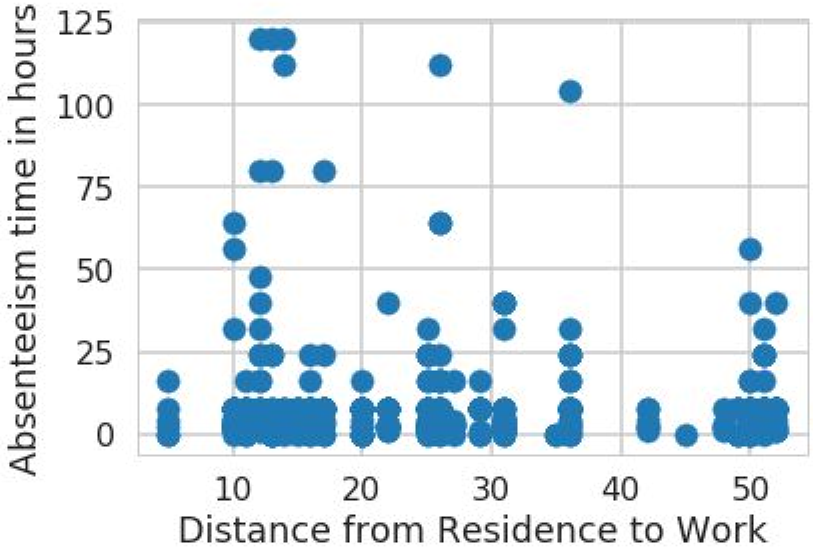
So from the above t-test, we conclude that at least one of the means of the four different education groups is different from the others, as a p-value of 3.43*10^{-28}, which is less than our alpha of 0.05.

As absence duration is perhaps one of the more interesting variables (at least from an employer's perspective), so now we want to look at how absence duration is correlated with the other variables:

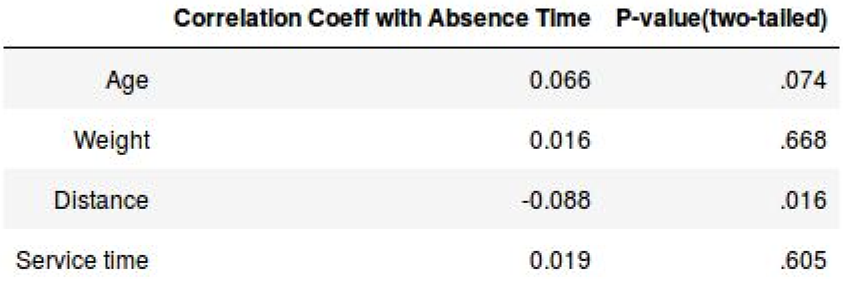








Below is a table showing the Pearson correlation coefficients between Absence time and various predictors:



	Correlation Coeff with Absence Time	P-value(two-tailed)
Age	0.066	.074
Weight	0.016	.668
Distance	-0.088	.016
Service time	0.019	.605

It would appear that none of our proposed variables (age, weight, distance between residence and work, and service time) have a strong linear correlation to the absence time. However, we should note that there may be other types of correlations (e.g. quadratic) that would better capture the relationships.

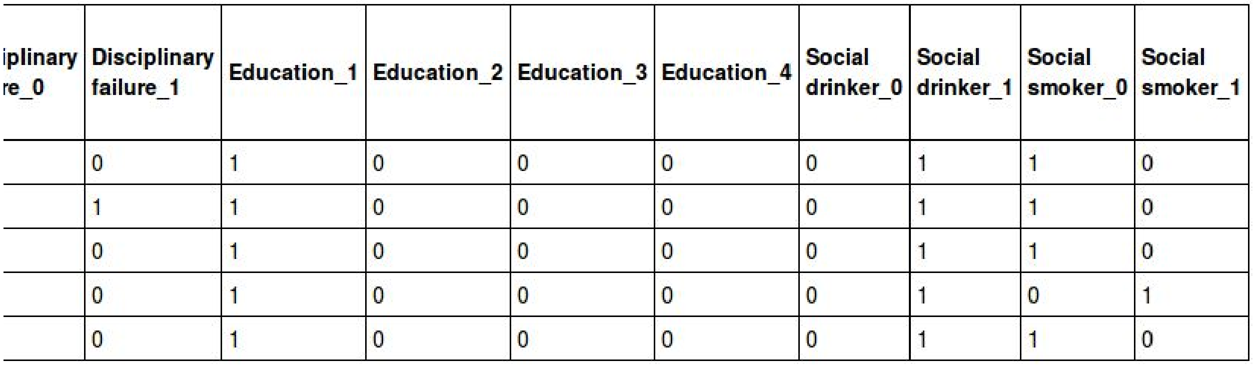
Results and in-depth analysis

We create two sets of profiles: one for the actual absences, and one for the individual workers (who have unique IDs labeled 1 through 36).

To form our clusters, we will employ K-means clustering. To do so,, we first must select k, run the k-means algorithm on our data, then compare the features of our clusters.

Profiling: Absences

Before running k-means, we first want to handle our categorical data by using one-hot encoding. Below is an example of one-hot encoding applied to one of our categorical variables:

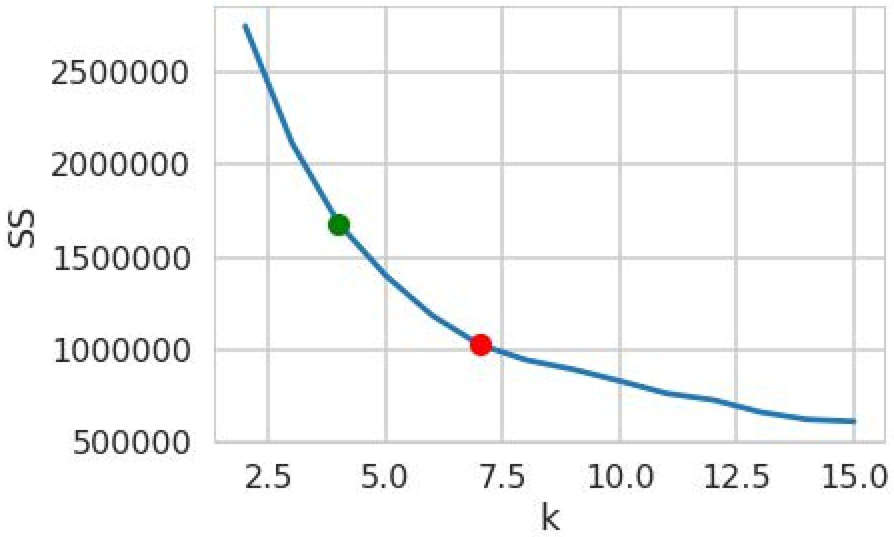


iplinary re_0	Disciplinary failure_1	Education_1	Education_2	Education_3	Education_4	Social drinker_0	Social drinker_1	Social smoker_0	Social smoker_1
	0	1	0	0	0	0	1	1	0
	1	1	0	0	0	0	1	1	0
	0	1	0	0	0	0	1	1	0
	0	1	0	0	0	0	1	0	1
	0	1	0	0	0	0	1	1	0

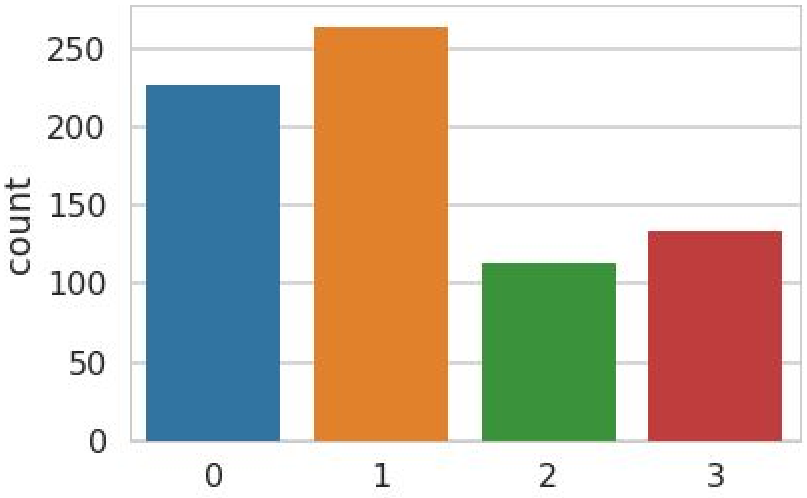
Note how the four levels of the education variable are now all represented as their own variable, with the ones and zeroes acting as indicator variables for the level (ie a 1 in education_3 means that the education level of the worker was level 3).

Now that we've adjusted for our categorical variables, let's try to select k.

Below is a graph of the sum of squared residuals for various k:

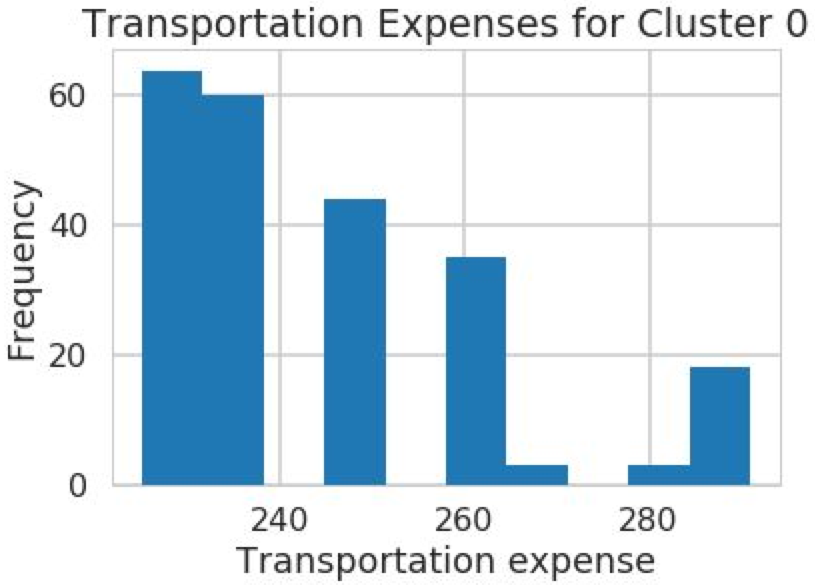


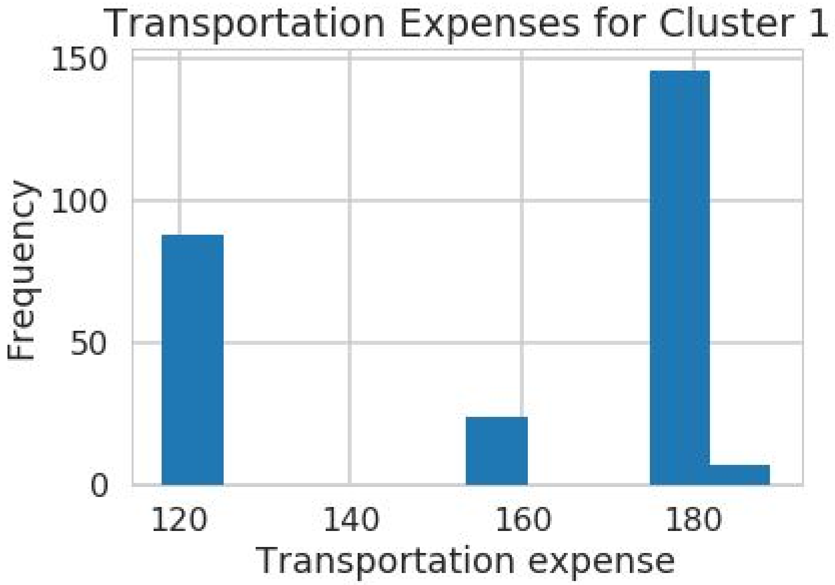
Using the elbow method, we see that there is no apparent 'elbow', so we might initially select 4 or 7 clusters. In this case, we select 4 clusters as generally speaking, it's easier to find differences between groups(clusters) if there are fewer groups to compare.

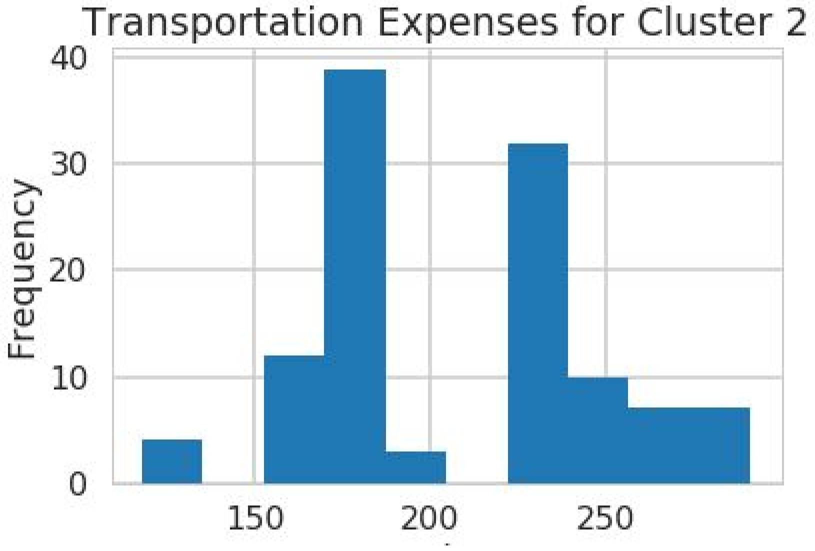


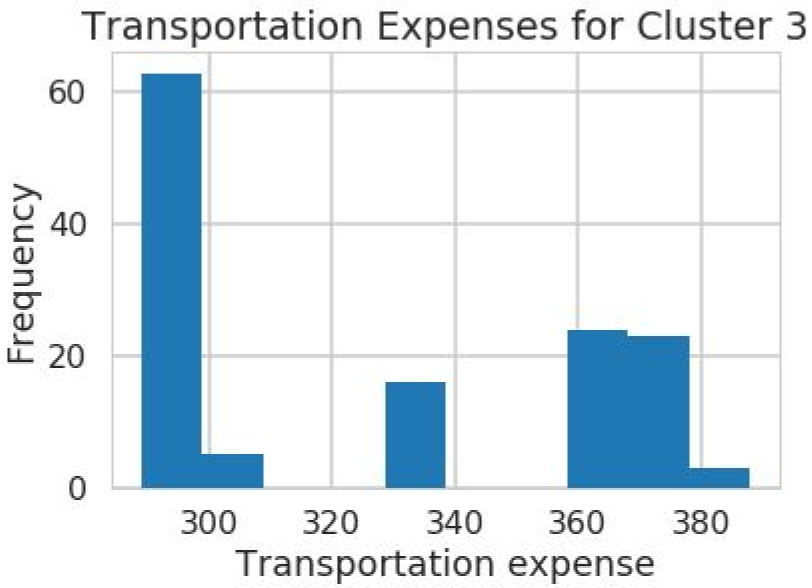
Once we've made our four groups, we first check the size of each group. While groups 2 and 3 are noticeably smaller than 0 and 1, they're not too small to the point where we'd suspect overfitting.

Next we take a look at some of the variables and how they're distributed among clusters. For this example, we examine the transportation expense:


Transportation Expenses for Cluster 0
Frequency
0
20
40
60
240
260
280
Transportation expense


Transportation Expenses for Cluster 1
Frequency
0
50
100
150
120
140
160
180
Transportation expense


Transportation Expenses for Cluster 2
Frequency
0
10
20
30
40
150
200
250



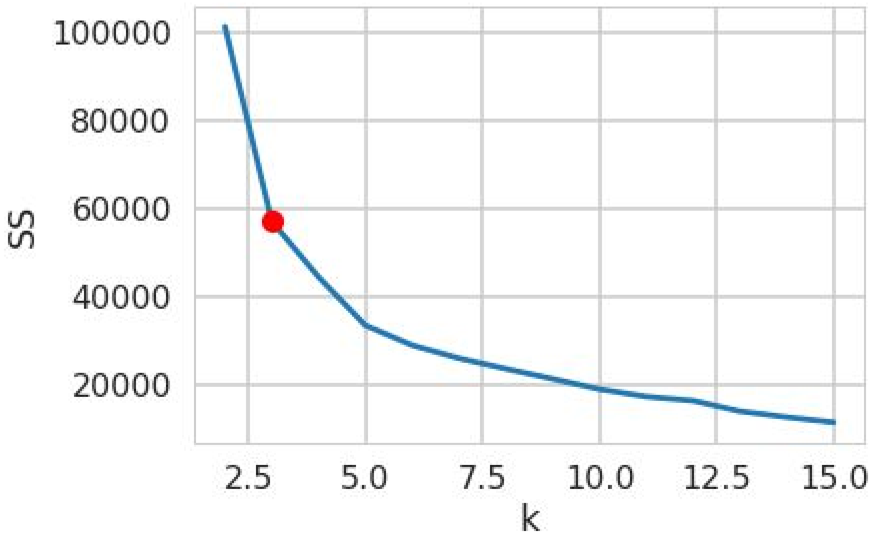
Based on the transportation expenses, we can see a clear difference in the distributions among the groups: group 1 has mostly sub-200 expenses, group 3 has mostly expenses over 300, group 0 has mostly entries in the 200 to 260 range, while group 2 has expenses (1) around 175 and (2) over 225.

We can repeat this process for other variables across clusters, and we report our findings in the following table:

Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**
	Under 200 Transportation expense	Between 150 and 200 Transportation expense	At least 300 transportation expense
Late 20s	Late 30s to early 40s	Under 30 y.o	
Weighs right under 70kg	Between 80 and 90kg		
Has 1 Child	No Children		
Between 0 and 10 hours absent			At least 10 hours absent

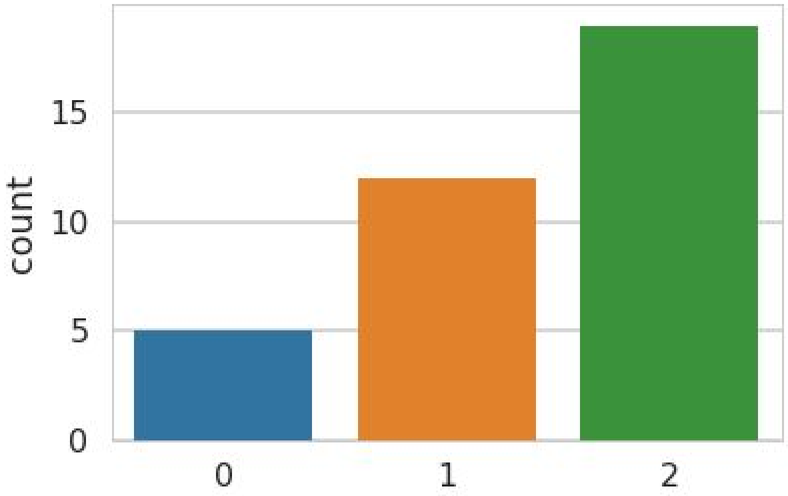
Profiling: Individuals

Now we apply the same steps for profiling absences to individuals. After we've one-hot encoded the categorical variables, we select k by comparing sum of squared residuals:



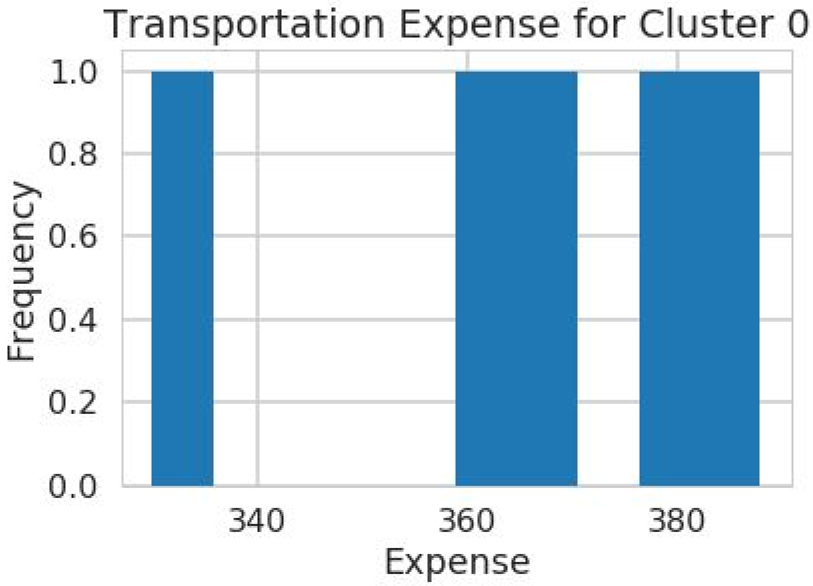
From the above SS, we choose k to be 3.

After we've formed our clusters, we can check the size of each cluster:

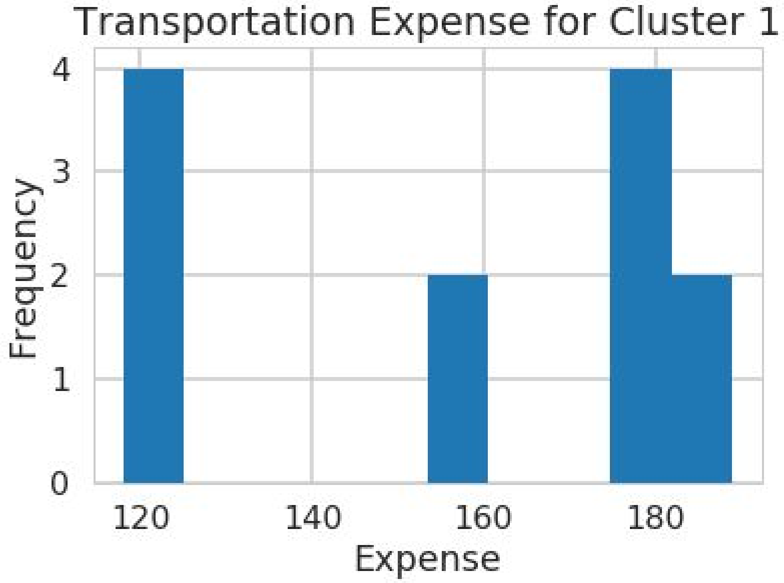


While cluster 0 has only 5 workers in it, we conclude that this is probably just due to variance in cluster size.

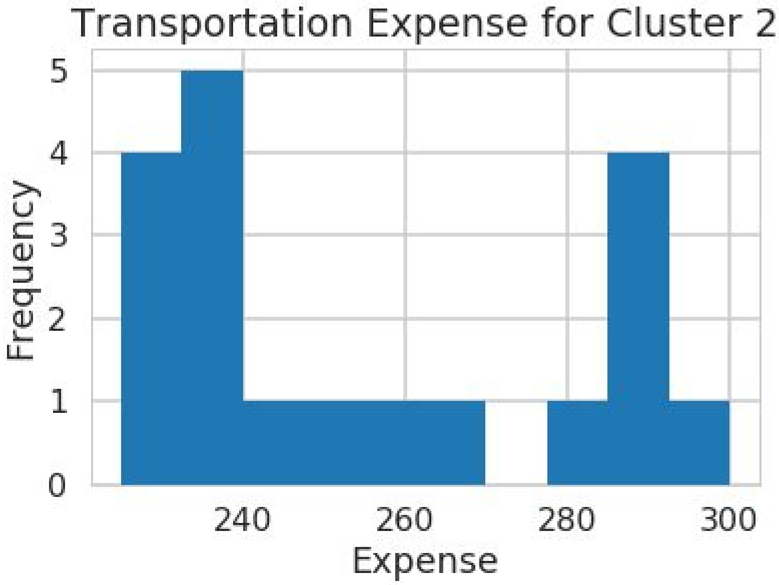
Next, we examine the transportation expense for each cluster:



Transportation Expense for Cluster 0



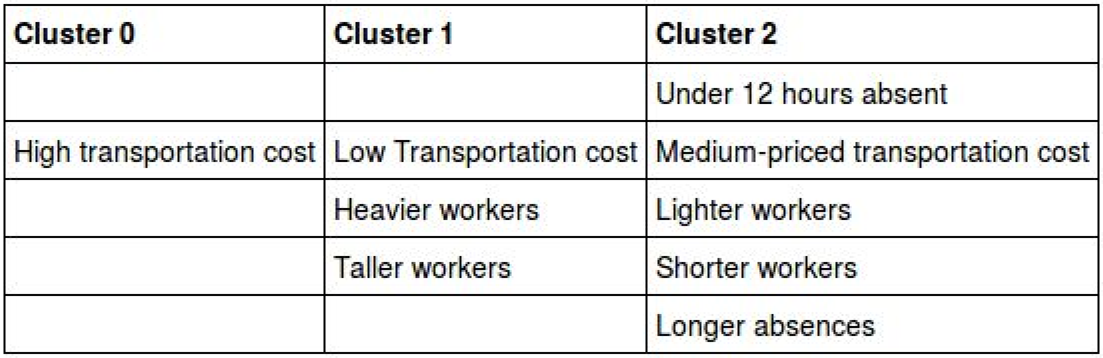
Transportation Expense for Cluster 1



Transportation Expense for Cluster 2

From the above histograms, we see that cluster 0 has the highest expenses, cluster 1 has the cheapest expenses, and cluster 2 has medium-priced expenses.

We apply this methodology to other variables across all clusters, and conclude the following:



Cluster 0	**Cluster 1**	**Cluster 2**
		Under 12 hours absent
High transportation cost	Low Transportation cost	Medium-priced transportation cost
	Heavier workers	Lighter workers
	Taller workers	Shorter workers
		Longer absences