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

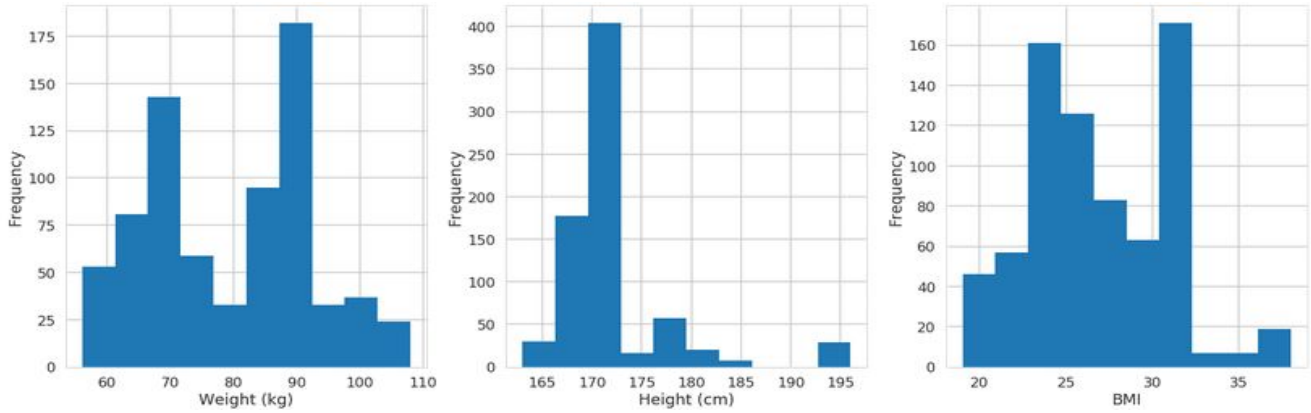
The dataset was provided by the UCI machine learning repo, and was already formatted. No additional steps were needed to clean the data.

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.

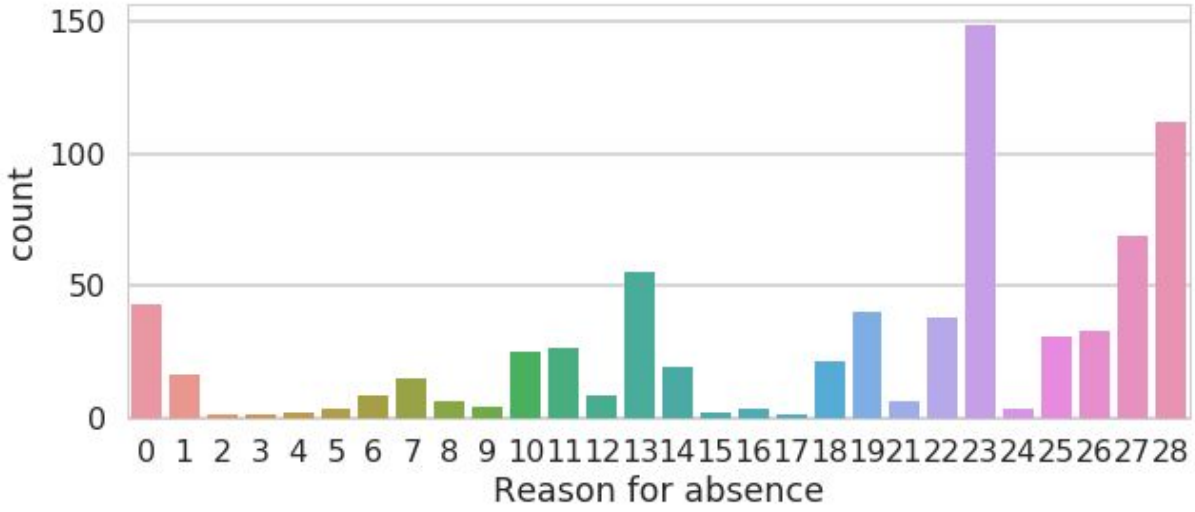
Exploratory data analysis summary

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



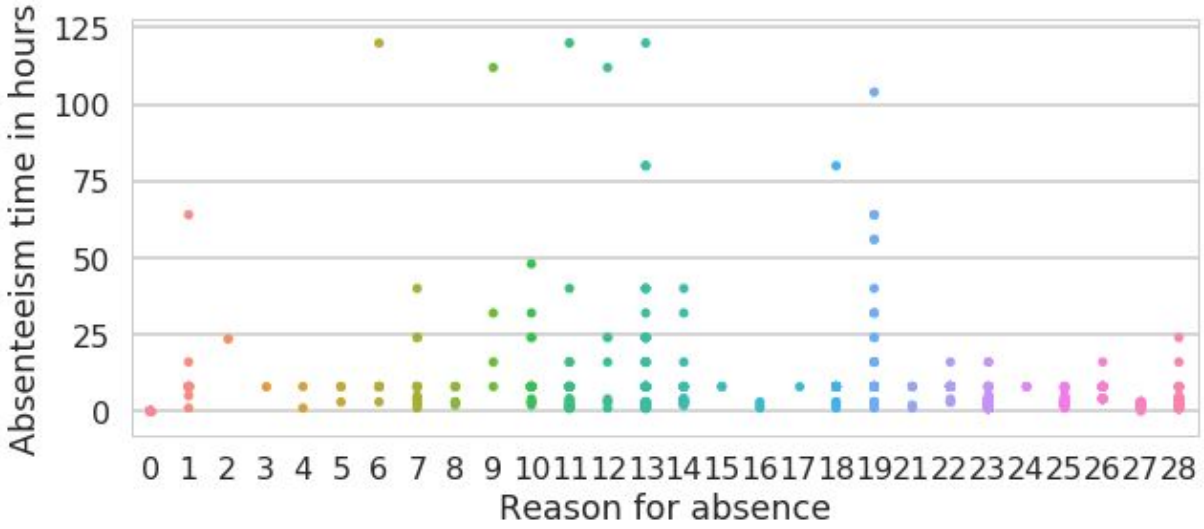
A quick look at the weight, height and BMI reveal that although we expect metrics like weight, height, and BMI to be approximately normal, both weight and BMI appear to have a bimodal curve, and height looks a bit skewed right. However, note that these frequencies are based on frequencies of absences, 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:



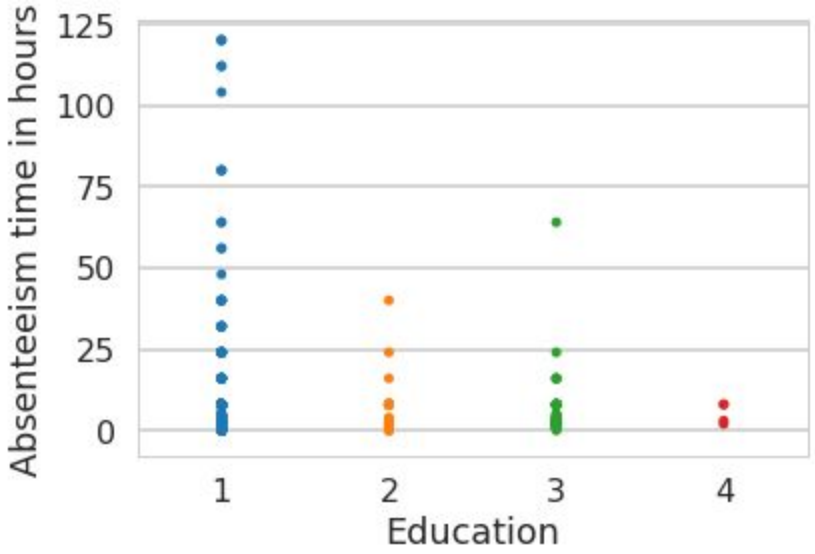
We see that medical consultation(23) and dental consultation(28) are by far the most popular reasons given, followed by physiotherapy(27), diseases of the musculoskeletal system and connective tissue (13), laboratory examinations (24) and unjustified absence(25). Now let's take

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.

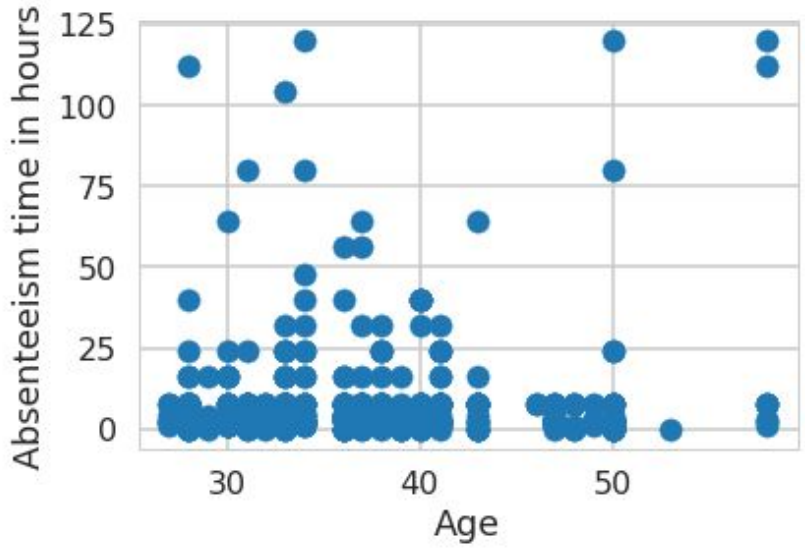
Let's conduct a t-test to see if at least one of the group means is different from the others:

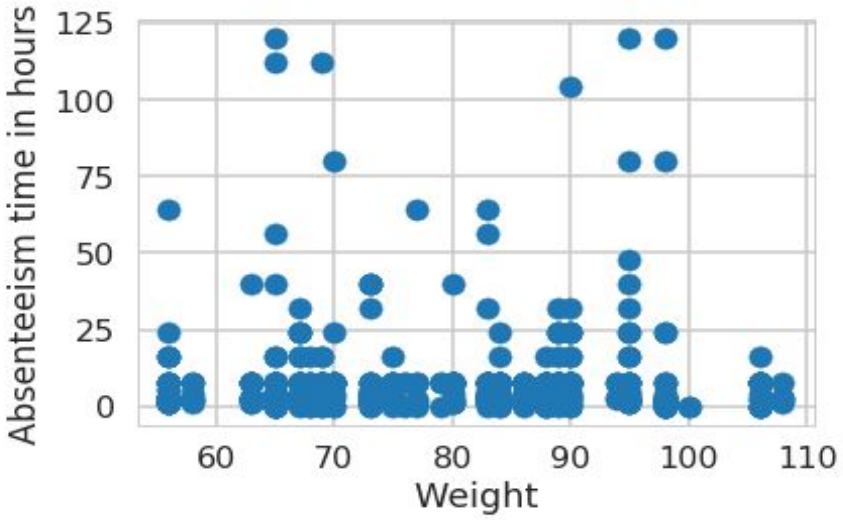
$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

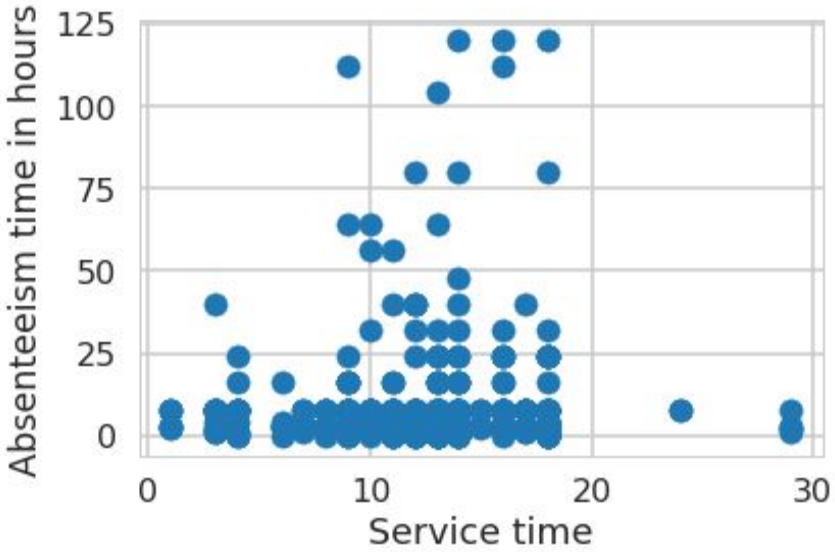
$$H_A : \text{At least one of the means is different from the others}$$

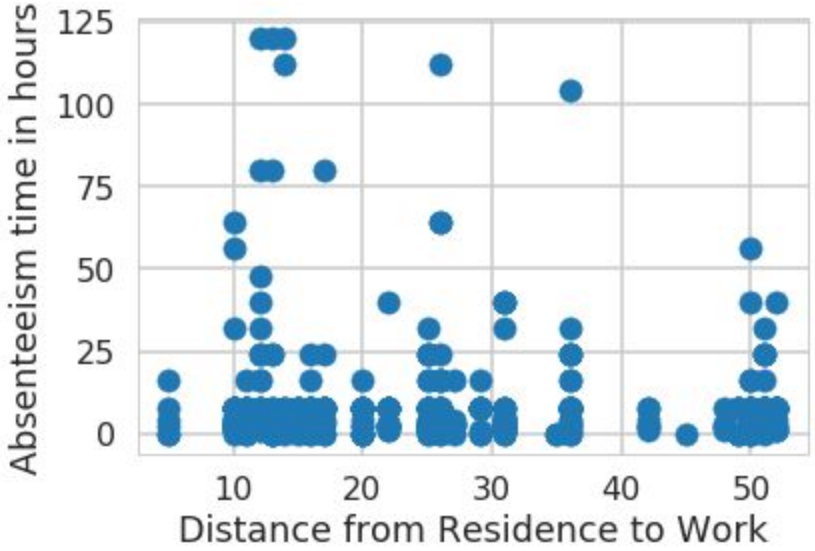
So from the above t-test, we conclude that at least one of the means of the four different education groups is different from the others, as a p-value of $3.43*10^{-28}$, which is less than our alpha of 0.05.

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.