Capstone Project 1 Absentee Profiling

Andrew Liu

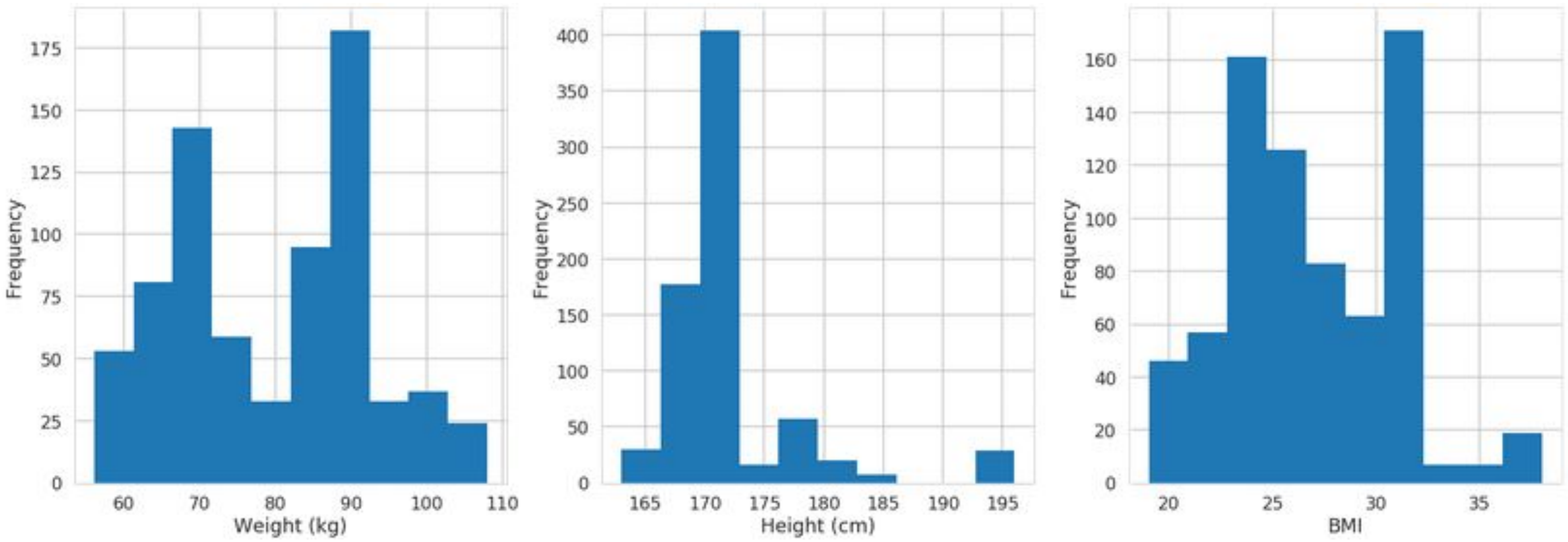
Proposal with Problem Statement

- 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.
- Provided by 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
- Once clustered, 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.
- 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

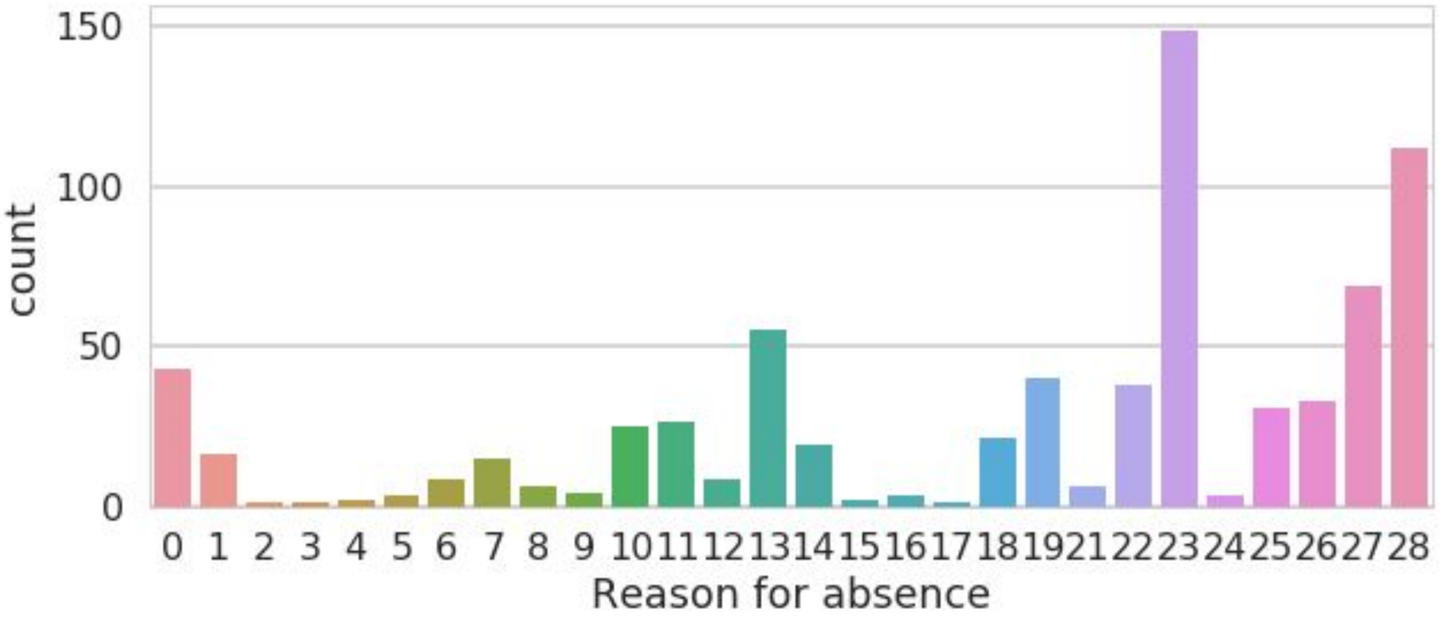
- The dataset was provided by the UCI machine learning repo, and was already formatted. No additional steps were needed to clean the data.
- A few of the entries for “month” were 0, so while the data wasn’t explicitly missing, the entry didn’t make much sense. The entries were left as 0, and were properly 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, as there were only 740 entries in this dataset..

Distritbution of Height, Weight, BMI



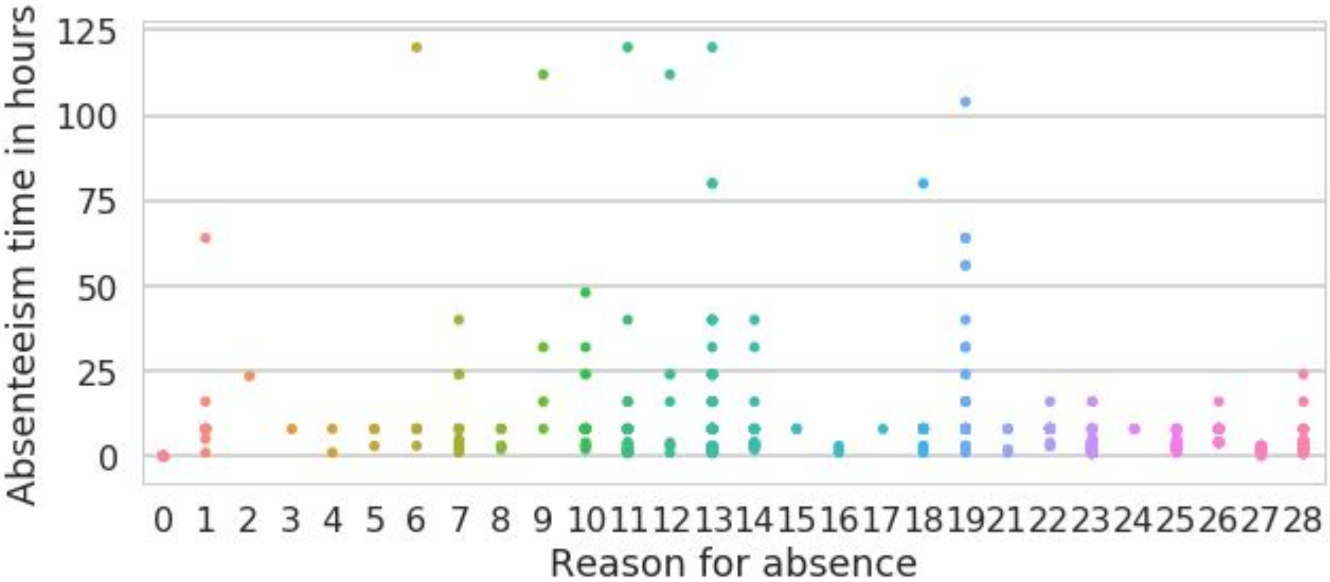
- Some of the distributions from the previous slide look a bit weird! (bimodal, skewed, etc)
- We expect things like height, weight, and BMI to be approximately normally distributed
- However, these histograms are a histogram of absences, not people, and although we believe the distribution of weights across all people to be normal, we shouldn’t expect the same from absences

Distribution for Reason for absence



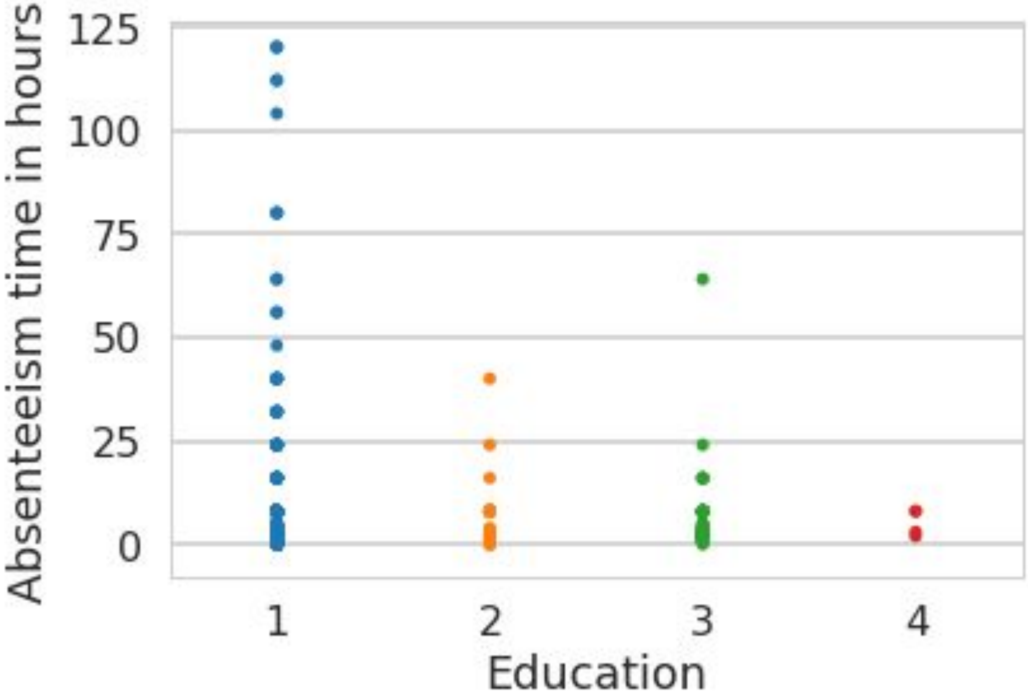
- Medical consultation(23) and dental consultation(28) are by far the most popular reasons given
- Physiotherapy(27), diseases of the musculoskeletal system and connective tissue (13), laboratory examinations (24) and unjustified absence(25)

Absence duration based on Reason



- The last seven reasons for absence (22 through 28) were not attested by the International Code of Diseases (ICD)
- Reasons 1 through 21 were attested by the ICD
- These last seven reasons had all absences take less than 24 hours, while the first 21 types of absences have lengthier occurrences

Relationship between education level and Absence duation?



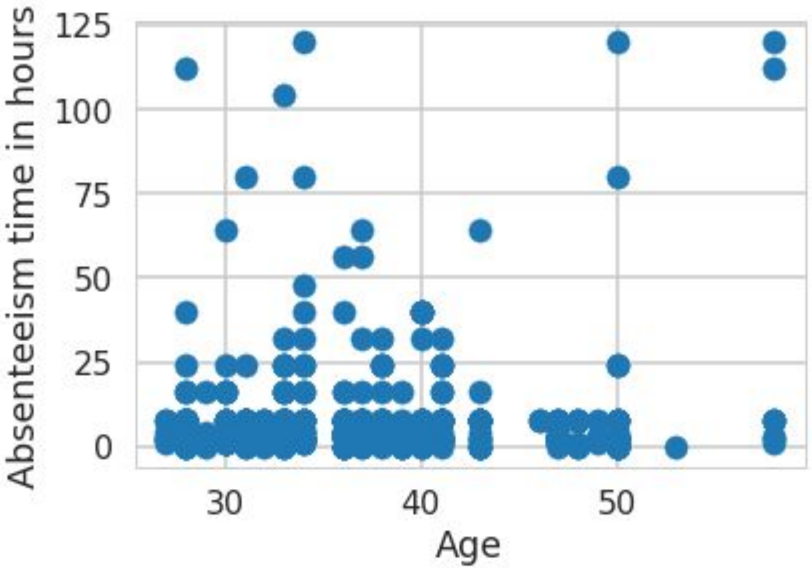
- Conduct a t-test to see if at least one of the group means is different from the others:

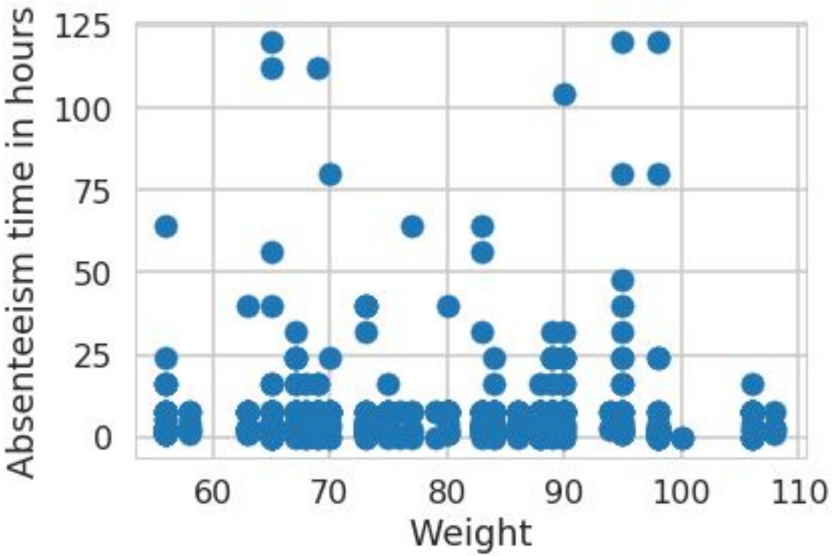
$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

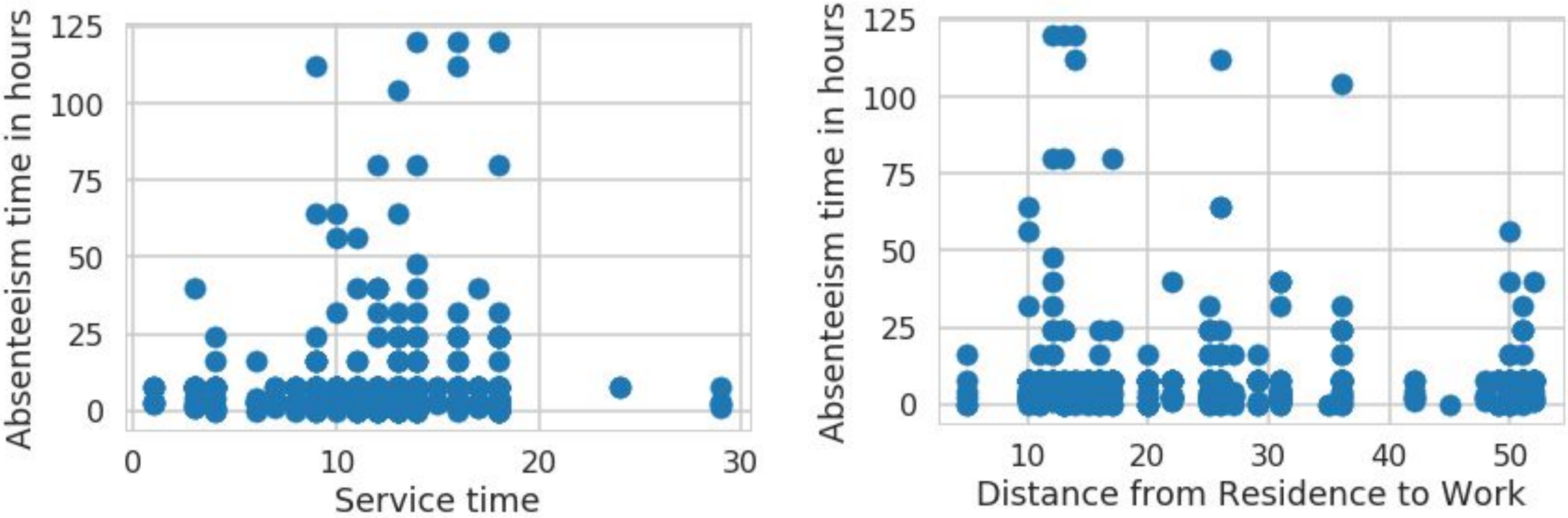
H_A : At least one of the means is different from the others

- T-test results: p-value of $3.43*10^{-28}$, which is less than our alpha of 0.05.
- Conclude that at least one of the means of the four different education groups is different from the others

How are other feature related to absence duration?






Absenteeism time in hours
125
100
75
50
25
0
0
10
20
30
Service time
Absenteeism time in hours
125
100
75
50
25
0
10
20
30
40
50
Distance from Residence to Work

Linear correlation: Pearson correlation coefficients

	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

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