Employee EDA

Matt Oehler

January 23, 2018

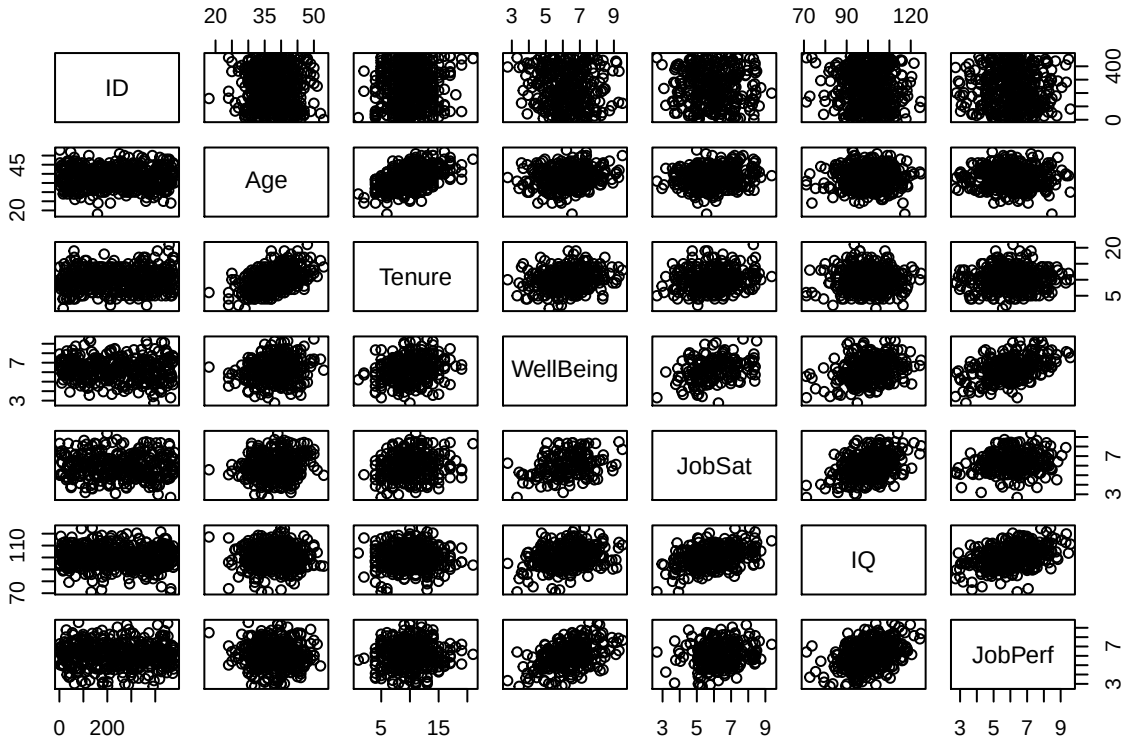
1. Goals of the Analysis

In order for companies to be economically productive, they will want to maximize their profits. One factor that influences a company's profits is its employees. Employees that are active and engaged help to yield higher profits while unmotivated and unhappy employees do not. As a result, companies are interested in how happiness, as well as other possible factors, contribute to an employee's performance. Knowing those relationships will help companies to enact policies or make changes in company culture that will maximize employee performance, and help to increase company profits.

To perform this analysis, researches gathered data on 480 university employees. The data includes measurements on variables including: age, number of years employed, the employees well-being, the employee's satisfaction with their job, the employee's IQ, and the employee's job performance. Using these data, one should be able to model how the other features impact an employee's job performance.

2. Features of the Data

We can use a scatter plot matrix (shown below) to quickly look at all of the data since all of our coviariates are quantitative. At a glance it appears that some of the covariates, such as IQ and well-being, are linearly related with job performance.



Another important note about this data set is that there are many missing values. The summary table below shows this more clearly, but the are several missing data points (NA values) for well-being, job satisfaction, and job performance.

	Age	Tenure	WellBeing	JobSat	IQ	JobPerf
1	Min. :18.00	Min. : 1.00	Min. :2.750	Min. :2.670	Min. : 71.07	Min. :2.730
2	1st Qu.:34.00	1st Qu.: 8.00	1st Qu.:5.455	1st Qu.:5.060	1st Qu.: 94.52	1st Qu.:5.293
3	Median :38.00	Median :10.00	Median :6.320	Median :5.940	Median : 99.92	Median :6.070
4	Mean :37.95	Mean :10.05	Mean :6.270	Mean :5.953	Mean :100.11	Mean :6.074
5	3rd Qu.:42.00	3rd Qu.:12.00	3rd Qu.:7.103	3rd Qu.:6.838	3rd Qu.:105.52	3rd Qu.:6.930
6	Max. :53.00	Max. :21.00	Max. :9.500	Max. :9.370	Max. :124.81	Max. :9.580
7			NA's :160	NA's :160		NA's :64

3. Statistical Method

I think that multiple linear regression would be an appropriate method to analyze this data. The issue of missing data would obviously need to be accounted for, but since we have several covariates and a single quantitative response variable (job performance), it seems that multiple linear regression would fit this scenario well.

4. Things I Don't Know

In the past I have explored methods of imputing missing data such as the EM algorithm, but I don't know how the different imputation methods affect the assumptions and/or results of the model. I am curious to see what other data imputation methods are out there and which of them are the most useful for this kind of situation.