Will Eddy

Psych 6430

9/12/2018

**Personal Data Set – IBM HR Data: Attrition and relationship to other factors**

The subjects of my data are fictional IBM employees. This dataset was created by data scientists at IBM.

<https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>

I used SPSS to recode the alphanumeric dichotomous variable values to numeric.

My two dichotomous variables are:

**Attrition** – Yes as 1, and No as 0. This shows whether an employee left IBM by their own choosing.

**Department** –Research and Development as 1, and Sales as 0. This shows where in the company the employee works. The original data set also shows a person’s position within the department.

The original data set included 4.3% Human Resources employees, but I discarded this portion of the data using SPSS.

My two continuous variables are:

**Percent Salary Hike,** Measured as a percentage of salary in dollars. I could not find clarification about the period of time the salary hike represents.

**Years with current manager,** measured in years.

In observing the sample as a whole, the strongest relationship I found in the correlation matrix was a weak negative correlation of -.15 between Attrition and Years with current manager. This would make sense because these variables are measuring something that is related. This data set does have some extreme values, but they do not appear to be errors. These values should remain in the data set for analysis.

I created two subset groups: Attrition = Yes and Attrition = No.

In the Attrition = Yes group, there appears to be more of a relationship between PercentSalaryHike and YearsWtihCurrManager when compared with the Attrition = No group.

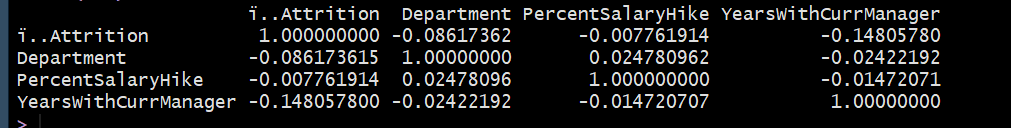
For the variable attrition, I found skewness and kurtosis considerably different from a normal distribution. For the variable Department, I found kurtosis considerably different from a normal distribution. Negative kurtosis is more extreme in the Department field in the Attrition = Yes group. Skewness is more extreme in the Department field in the Attrition = No group.

Overall, it appears that no significant relationships were found.

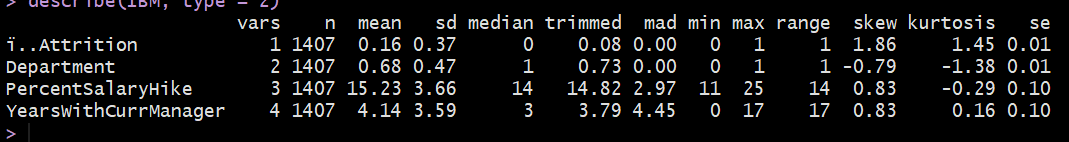
Going forward, I think it might be wise to create a new variable – We know salary information and percentage raise. These variables can be combined to create a raw raise in dollars variable.

Additionally, I may need to consider performing a transformation to nudge the data set towards normality, to make way for additional analysis.

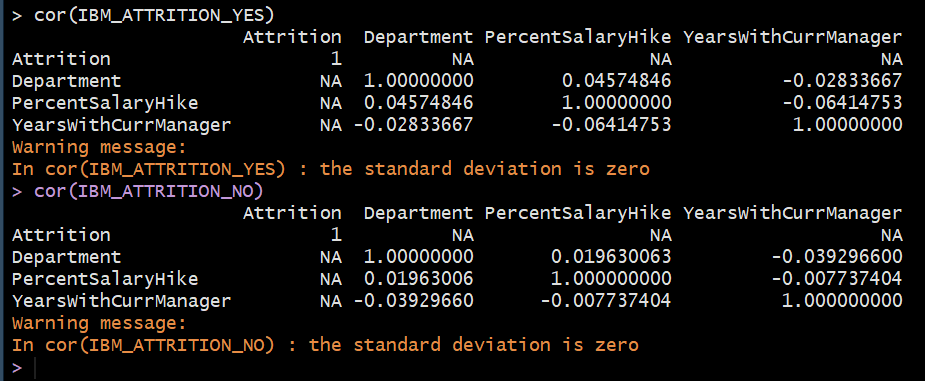
Correlation matrix for whole sample:



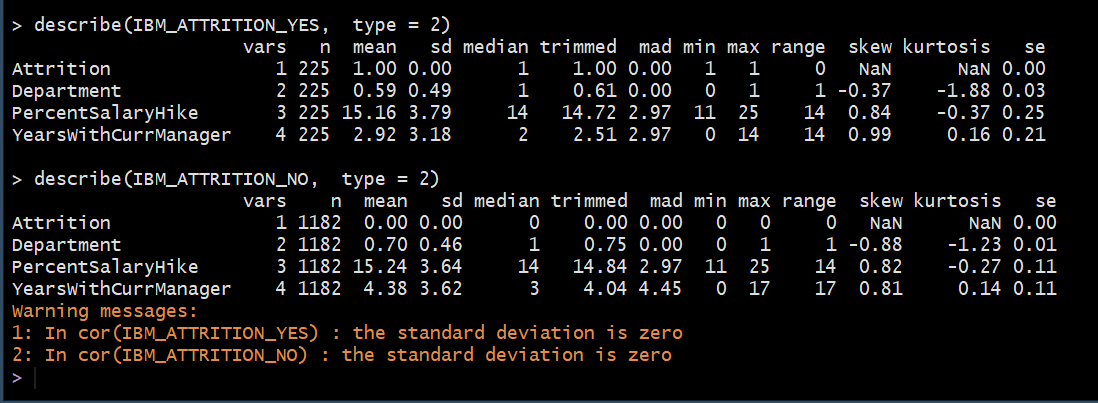
Descriptive Statistics for whole sample:



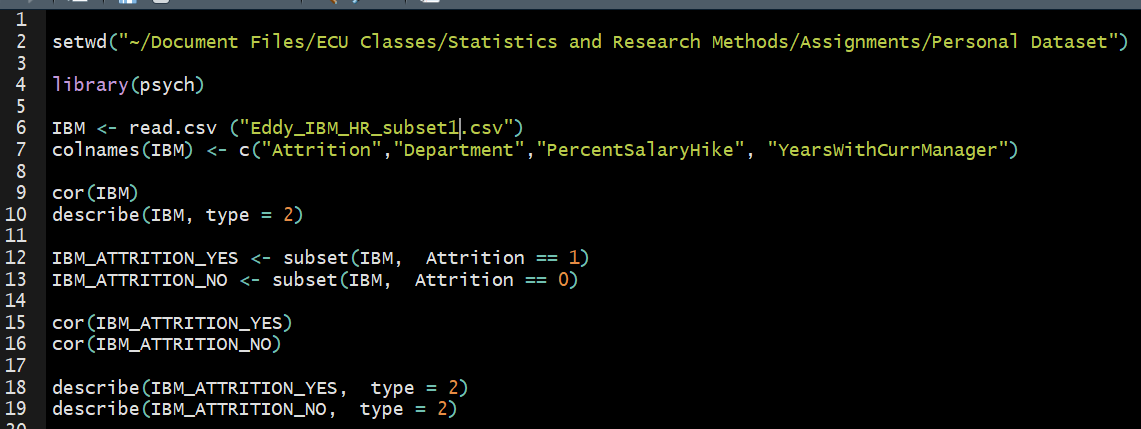
Correlation matrix for Attrition = YES and Attrition = NO



Descriptive statistics for Attrition = YES and Attrition = NO



Code used in R:



Sources:

http://core.ecu.edu/psyc/wuenschk/R-Lessons/R-Lessons.htm

<https://www.kaggle.com/ganeshn88/exploratory-data-analysis-of-hr-attrition>

<https://www.r-bloggers.com/5-ways-to-subset-a-data-frame-in-r/>