Personal Data Set – IBM\_HR data set

1. The subjects of my data are fictional IBM employees. This HR data set was created by data scientists at IBM. It is available at the following URL:

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

I further processed the data in SPSS, removing all but the following 4 variables, and ordering by the Attrition variable.

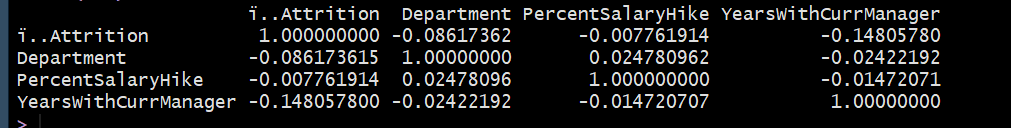
1. My two dichotomous variables are:
   1. Attrition – Yes as 1, and No as 0
   2. Department – 1 as Research and Development, and 0 as Sales
      1. The original data set had 4.3% Human Resources employees, but this portion of the data was discarded.
2. My two continuous variables are:
   1. Percent Salary Hike, Measured as a percentage of salary in dollars
   2. Years with current manager, measured in years
3. This data set does have some extreme values, but they do not appear to be errors. I think these values should remain in the data set for analysis.
4. **Do your two groups (defined by the dichotomous variable on which you sorted) appear to differ considerably on either of the continuous variables, or do they appear to have quite similar distributions on the continuous variables? If they differ, describe the differences.**
5. For the variable attrition, I found skewness and kurtosis considerably different from a normal distribution. For the variable Department, I found kurtosis significantly different from a normal distribution.

The strongest relationship I found in the correlation matrix was – Inverse correlation of -.15 between Attrition and Years with current manager. This makes sense because these variables are actually measuring the same thing – How long an employee stays with the company.

.

* At the very end of the document, paste in the SAS (or SPSS or R) syntax you used to conduct the analysis.

Correlation matrix:



Descriptive Statistics:

Sources:

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