**Employee Attrition Prediction and Factors Contributing to it.**

Attrition in business can mean the reduction in staff and employees in a company through normal means, such as [retirement](https://www.investopedia.com/terms/r/retirement.asp) and resignation, the loss of customers or clients to old age or growing out of the company's target demographic.  
  
Attrition can also refer to a business or professional losing customers. Customer attrition generally has a negative effect on the company's profits and growth.

**Step 1 -Define the Goal**

The following questions are among the ones Company would like answered:

1. What proportion of our staff are leaving?
2. Where is it occurring?
3. How does Age and Length of Service affect termination?
4. What, if anything, else contributes to it?
5. Can we predict future terminations?
6. If so, how well can we predict?

**Step 2 – Collect and Manage the Data**

The information contained in the readily available dataset is as follows:

There are 35 Attributes, and 1470 instances.

1. Age
2. Attrition
3. Business Travel
4. Daily Rate
5. Department
6. Distance from Home
7. Education
8. Education Field
9. Employee Count
10. Employee Number
11. Environment Satisfaction
12. Gender
13. Hourly Rate
14. Job Involvement
15. Job Level
16. Job Role
17. Job Satisfaction
18. Marital Status
19. Monthly Income
20. Monthly Rate
21. Number of Companies Worked
22. Over18
23. Over Time
24. Percent Salary Hike
25. Performance Rating
26. Relationship Satisfaction
27. Standard Hours
28. Stock Option Level
29. Total Working Years
30. Training Times Last Year
31. Work Life Balance
32. Years at Company
33. Years in Current Role
34. Years since Last Promotion
35. Years with Current Manager

**Step 3 – Data Exploration**

**Correlation Matrix and Heat Map:**

**Moderate Positively Correlated Features:**

Monthly Income vs Age: 0.497855

**Job level vs Age:** 0.509604

Years at Company vs Monthly Income: 0.514285

Years since Last Promotion vs Years with Current Manager: 0.510224

Years since Last Promotion vs Years in Current Role: 0.548056

Years at Company vs Years since Last Promotion: 0.618409

Total Working Years vs Years At Company: 0.628133

Total Working Years vs Age: 0.680381

Years in Current role vs Years with Current Manager: 0.714365

Years at Company vs Years in Current Role: 0.758754

Years At Company vs Years with Current Manager: 0.769212

Total Working Years vs Monthly Income: 0.772893

Total Working Years vs Job Level: 0.782208

**Moderate Negatively Correlated Feature:**

**Stock Option Level vs Attrition:** -0.137145

Years with Current Manager Vs Attrition: -0.156199

Age Vs Attrition: -0.159205

**Monthly income vs Attrition:** -0.159840

Years in Current role vs Attrition: -0.160545

Job Level vs Attrition: -0.169105

Total Working Years vs attrition: -0.171063

**Stop and Think about the Data:**

* What features affect our target variable the most (Attrition)?
* What features have strong correlations with each other?
* Can we do a more in depth examination of these features?

**Summary:**

From the heatmap, there is strong **positive(+)** correlation between Total working Years, Job Level, Monthly Income, Years with Current Manager, and Years in Current Role. Which could mean that the employees who have worked for more number of Total Years and are at higher Job Level and therefore are having high Monthly Income.

For the **negative(-)** relationships, Attrition and job level + total working years are correlated. I'm assuming that people tend to leave a company more when they are not been promoted and hence are working in current role for larger number of time as well as people at high job level with more number of total working years are expected to leave the company.

**Statistical Test for Correlation:**

**One-Sample T-Test (Measuring Total Working Years):** A one-sample t-test checks whether a sample mean differs from the population mean. Since total working years has the highest correlation with our dependent variable Attrition, let's test to see whether the average ‘total working years’ of employees that had an Attrition differs from those that had no Attrition**.**

**Conducting the T-Test:** Let's conduct a t-test at **95% confidence level** and see if it correctly rejects the null hypothesis that the sample comes from the same distribution as the employee population. To conduct a one sample t-test, we can use the **stats.ttest\_1samp()**function.

**Ttest\_1sampResult(statistic=-7.769571024760564, pvalue=2.407154396299421e-13)**

#### **T-Test Result**

The test result shows the test statistic "t" is equal to -7.7695. This test statistic tells us how much the sample mean deviates from the null hypothesis. If the t-statistic lies outside the quantiles of the t-distribution corresponding to our confidence level and degrees of freedom, we reject the null hypothesis. We can check the quartiles with stats.t.ppf():

#### **T-test Quartile**

If the t-statistic value we calculated above (-7.7695) is outside the quartiles, then we can reject the null hypothesis.

### One Sample T-Test Summary:

#### **T-Test = -7.7695 | P-Value = 2.407154396299421e-13 (approx. P-value = 0) | Reject Null Hypothesis**

Based on the statistical analysis of a one sample t-test, there seems to be some significant difference between the mean Total Working Years of employees who had Attrition and the entire employee population. The super low P-value of approximately equal '0' at a 5% confidence level is a good indicator to reject the null hypothesis.

But this does not necessarily mean that there is practical significance. We would have to conduct more experiments or maybe collect more data about the employees in order to come up with a more accurate finding.

[**Technology used:**](https://www.analyticsinhr.com/blog/blind-hiring-increases-workplace-diversity/)

Python, Anaconda, Jupyter Notebook