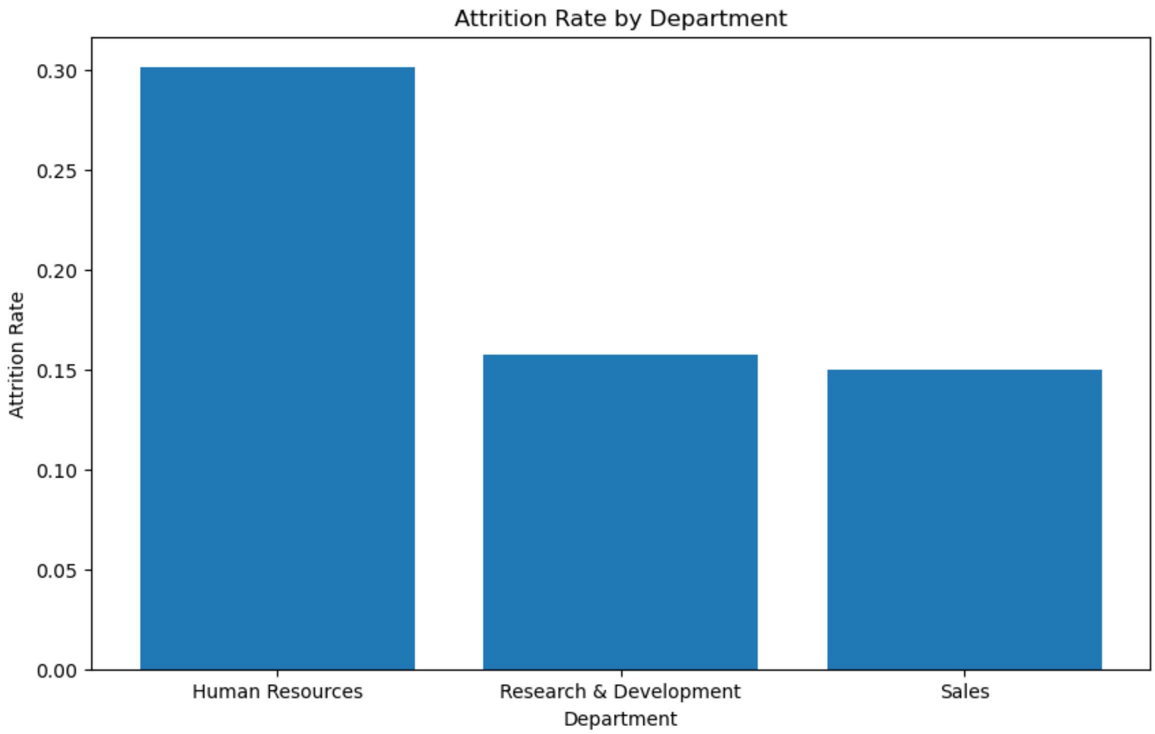
Denise Dodd

Predicting Attrition

10 Questions and Answers

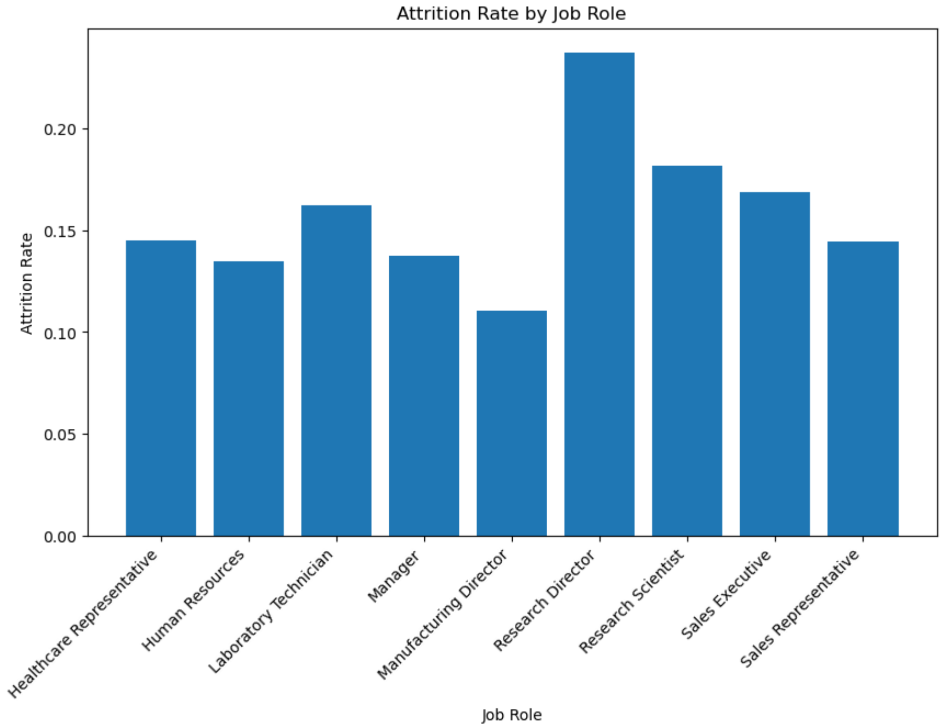
Question #1: What Department has the highest attrition rate?

Human Resources has almost double the rate of attrition as the other two departments in the dataset.



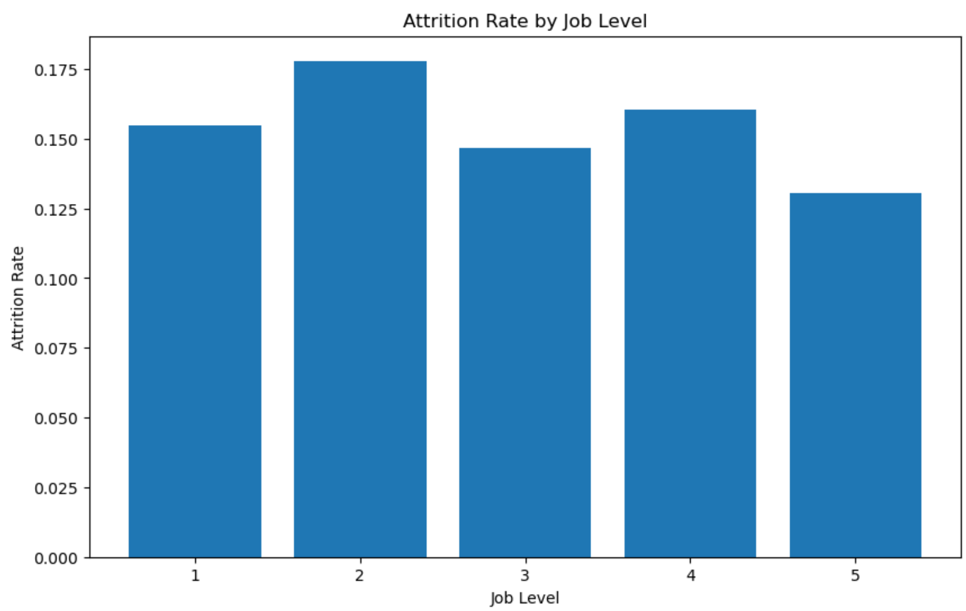
Question #2: What Job Role has the highest attrition rate?

Most Job Roles have between a .13 and a .18 rate of attrition. However, Manufacturing Director has the lowest rate of attrition at .11 and Research Director has the highest rate of attrition at .24.



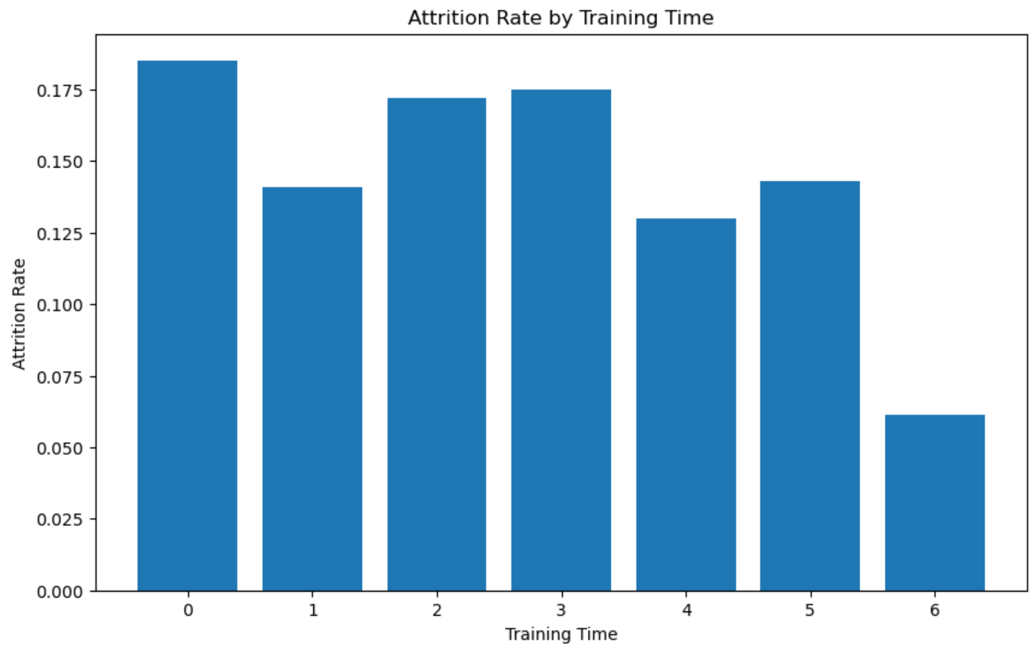
Question #3: Do certain Job Levels attrite more than others?

Most Job Levels have between a .14 and a .16 rate of attrition. However, Job Level 5 has the lowest rate of attrition at .13 and Job Level 2 has the highest rate of attrition at .18.



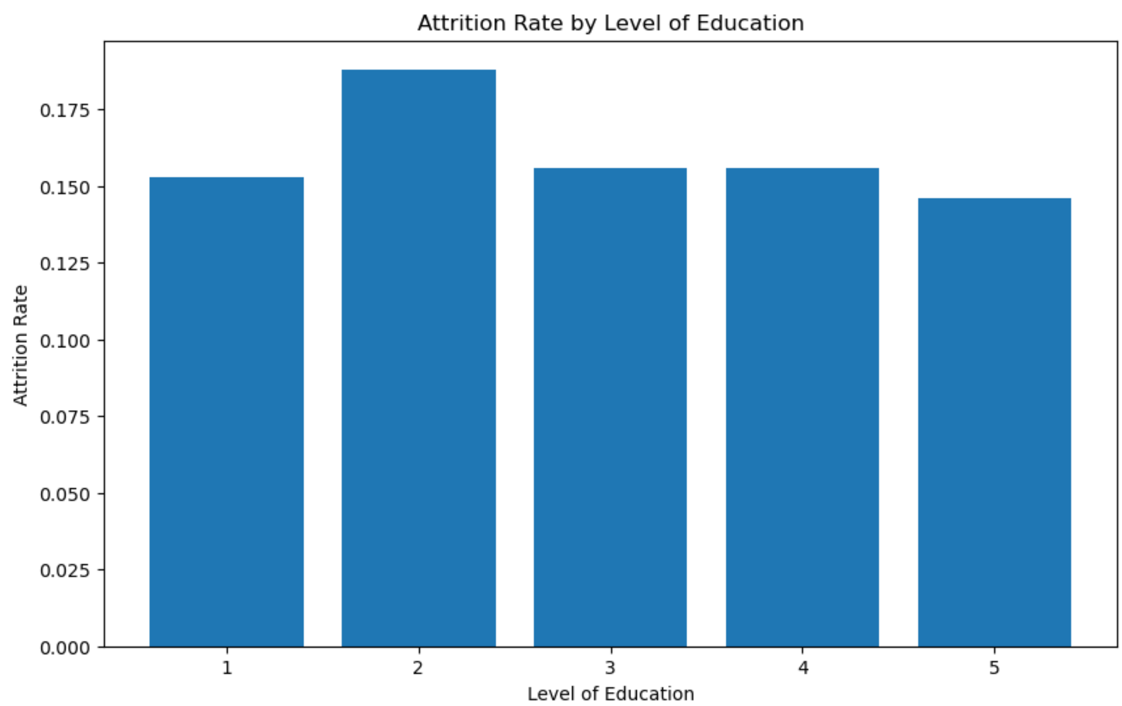
Question #4: Does more training time lead to attrition or less training time?

Less training time is associated with higher attrition. The highest attrition rate has 0 training times, and the lowest attrition rate has 6 training times.



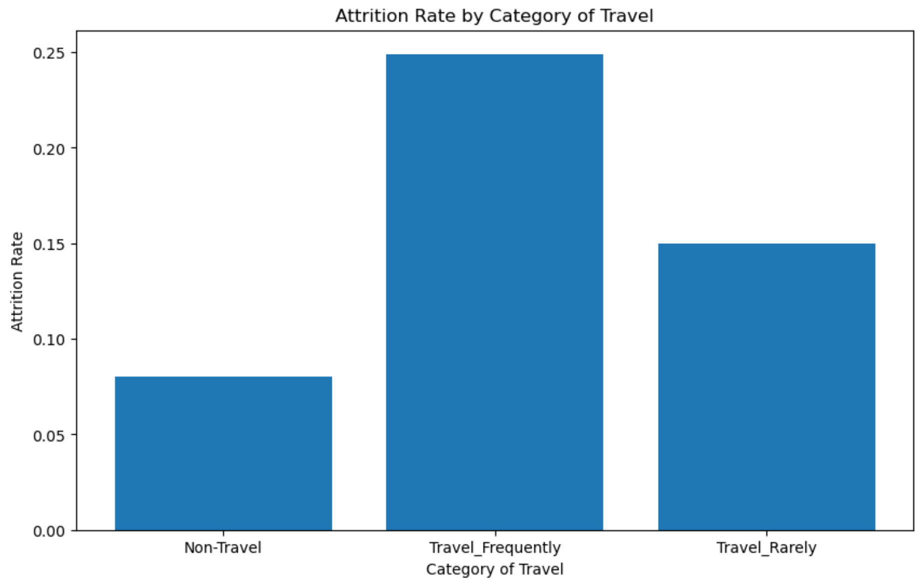
Question #5: Does a higher level of education lead to more or less attrition?

The attrition rates across the different education levels are steady except for a small spike in attrition for the 2nd level of education.



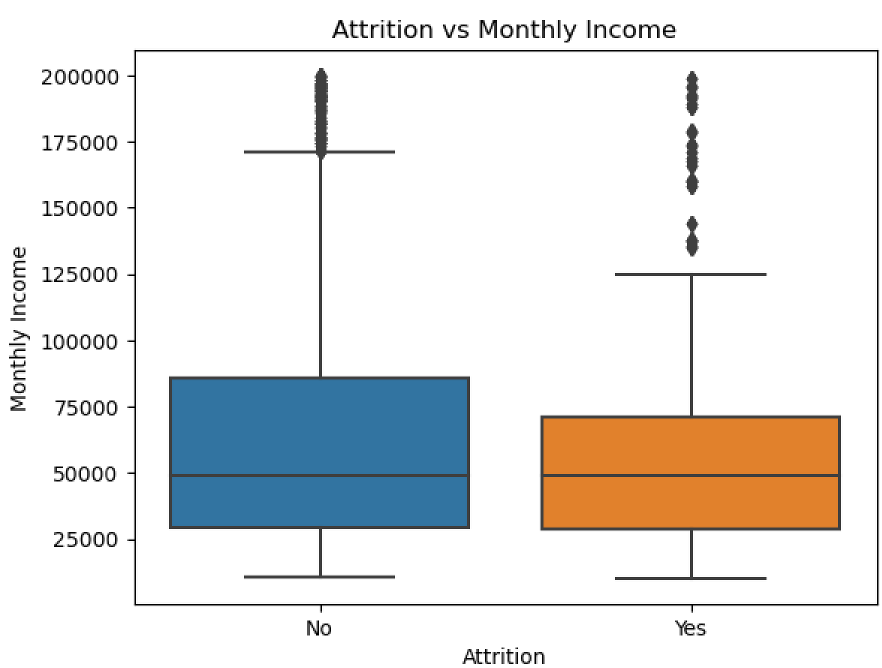
Question #6: Does more travel lead to attrition or less travel?

More travel leads to higher attrition. The attrition level of those who travel frequently is almost double the attrition level of those who do not travel at all.



Question #7: Is there a Monthly Income where attrition plateaus?

Both the "Yes" and "No" groups for attrition have a median income of roughly $50,000. The group that did not attrite has a higher upper bound to the 3rd quartile, a higher range to the whiskers, and a clustering of high range outliers. It is difficult to say if there is a plateau, but it can be concluded that those with a higher monthly income are less likely to attrite.



Question #8: How much has the attrition in the dataset cost the company?

"One estimate places the cost to replace an employee at three to four times the position's salary" (Wallace, 2023).

To determine the cost of attrition in the dataset, I filtered the dataset to only show the entries for employees that had attrited. I then annualized the monthly income and multiplied the yearly income by 3.5, which is the mid-point of the anticipated cost mentioned in the above quote. After summing the attrition cost column, it is determined that the attrition in the dataset had an estimated cost to the company of $1,841,966,280.00.

Question #9: We have assumed that Age and Total Working Years are related. Can this be confirmed?

To find the relationship between Age and Total Working Years, I first found the correlation coefficient of the two variables. The .678 correlation indicates that both variables move together in the same direction. As Age increases, Total Working Years also increases and vice versa.

I then created a scatter plot of Age and Total Working Years and overlayed a regression line to illustrate the lose linear relationship that the variables have.

Question #10: Can this model be run using only variables that the company has control over?

Yes. To accomplish this, you will need to drop the columns that the business has no control over, recreate the dummy variables, and concatenate the categorical dummy variables with the numerical variables in the dataframe. This dataframe can now be used to create features, target, testing, and training variables which can be used to select and fit a predictive model and make predictions. An example of this can be found in the coding notebook.

Question #10.5: Will the updated model have the same metrics?

No, if the data changes the metrics will not be the same. First, you need to determine if the Decision Tree Classifier is still the best fit for the data. In this case, the loop of several classification models shows that the Decision Tree Classifier still has very high classification metrics. However, the hyper-tuned model shows a drop in accuracy from 93% to 84% (as shown in coding notebook). Many of the variables that were dropped because the business had no control over them were ranked in the top 10 most important features for the original Decision Tree Classifier model. Losing this data results in a loss of accuracy when predicting if an employee will attrite.

Original Metrics:

	precision	recall	f1-score	support
0	0.93	0.98	0.96	737
1	0.86	0.65	0.74	145
accuracy			0.93	882
macro avg	0.90	0.81	0.85	882
weighted avg	0.92	0.93	0.92	882

Metrics of dataframe that the business can control:

	precision	recall	f1-score	support
0	0.84	1.00	0.91	737
1	0.00	0.00	0.00	145
accuracy			0.84	882
macro avg	0.42	0.50	0.46	882
weighted avg	0.70	0.84	0.76	882