**HR ANALYTICS**

HR analytics is defined as the process of measuring the impact of HR metrics, such as time to hire and retention rate, on business performance. Human resources is a people-oriented function and is so perceived by most people.

Companies rely completely on the human resource at hand and the planning for maintaining existing business and taking up new business is planned based on the resources available with the company. One of the major issues with this planning is attrition of employees.

Attrition is the gradual but deliberate reduction in staff numbers that occurs as employees retire or resign. This is one of the most important metrics for organizations to plan their staff hiring and requirements.

One of the major problems with attrition is the cost and time for the company and if the attrition is unexpected or not planned for, it may be a cause of great business loss to the company. To avoid these, companies are investing in HR analytics to predict the attrition rate so that they can plan their hiring accordingly.

HR Analytics is made up of several components that feed into each other:

-To gain the problem-solving insights that HR Analytics promises, data must first be collected.

-The data then needs to be monitored and measured against other data, such as historical information, norms or averages.

-This helps identify trends or patterns. It is at this point that the results can be analyzed at the analytical stage.

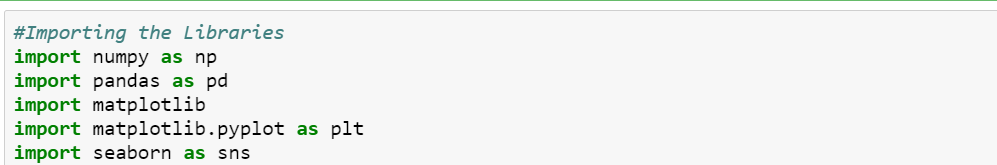
-The final step is to apply insight to organizational decisions.

**Problem Statement**

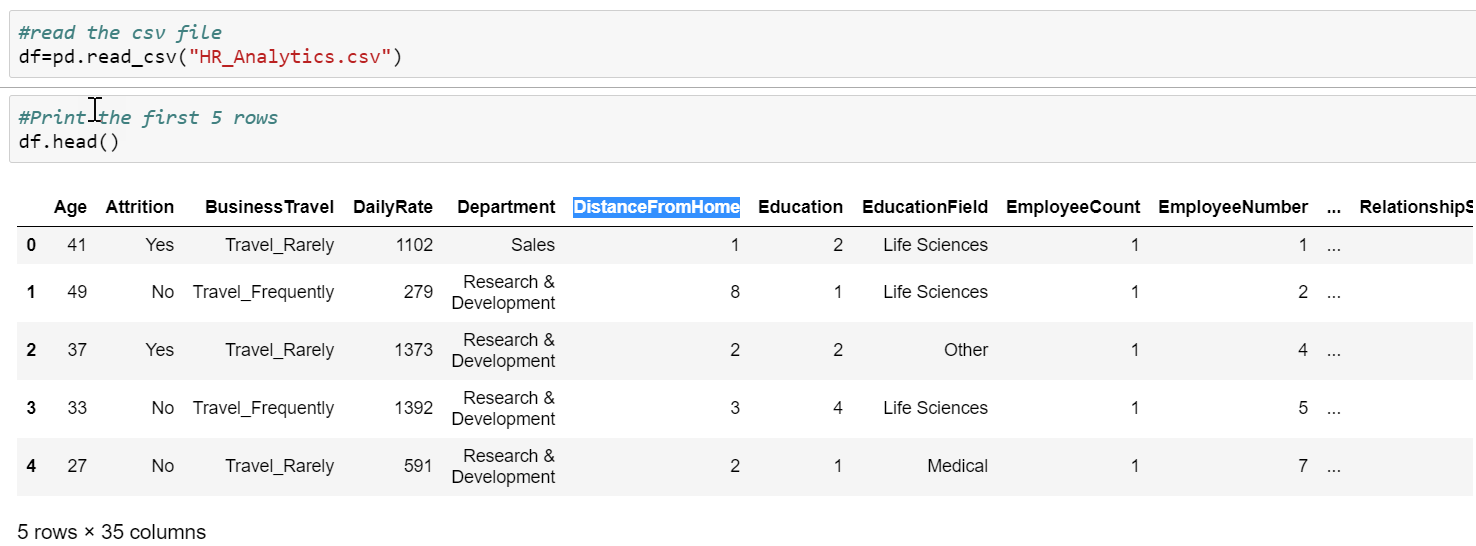
We will be working on IBM HR analytics dataset to predict attrition of employees. The dataset consists of features like Age, Business Travel, Daily Rate, Department, Distance from home, Education, Relationship Satisfaction, Stock Option, Years at company, Years in current role, Years since last promotion, etc. We will study the features and try to analyze the factors that lead to attrition.

Let’s begin with our analysis and create a model for predicting attrition.

Begin with loading libraries which may be utilized at different stages.

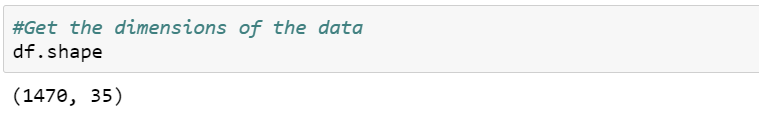


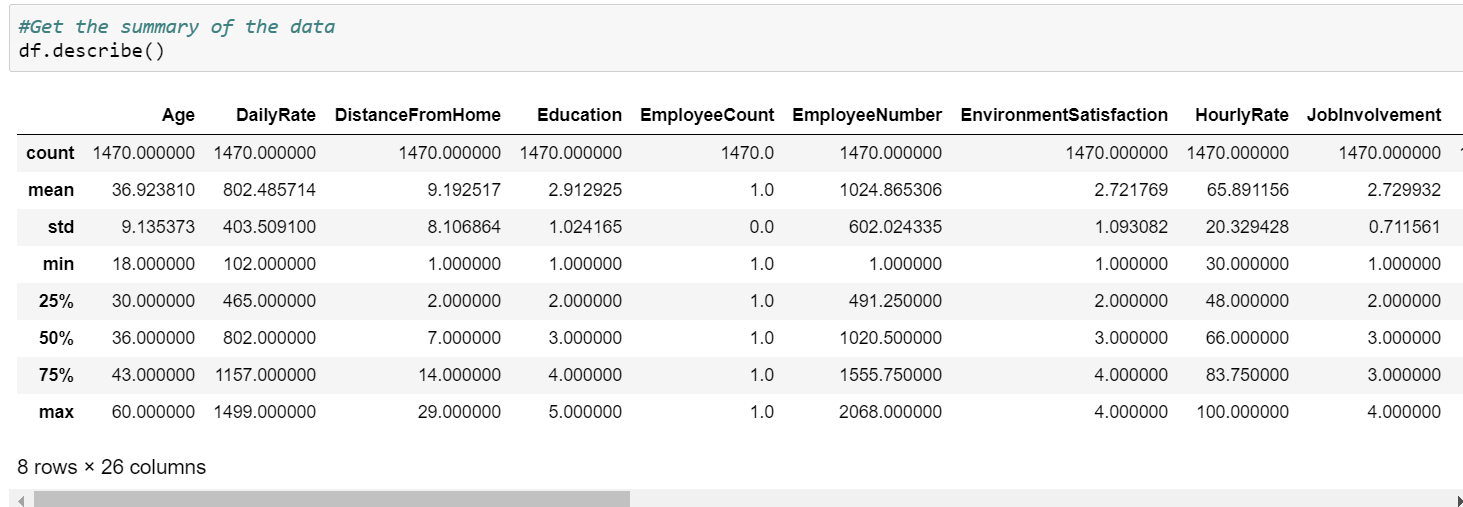
Load the data.



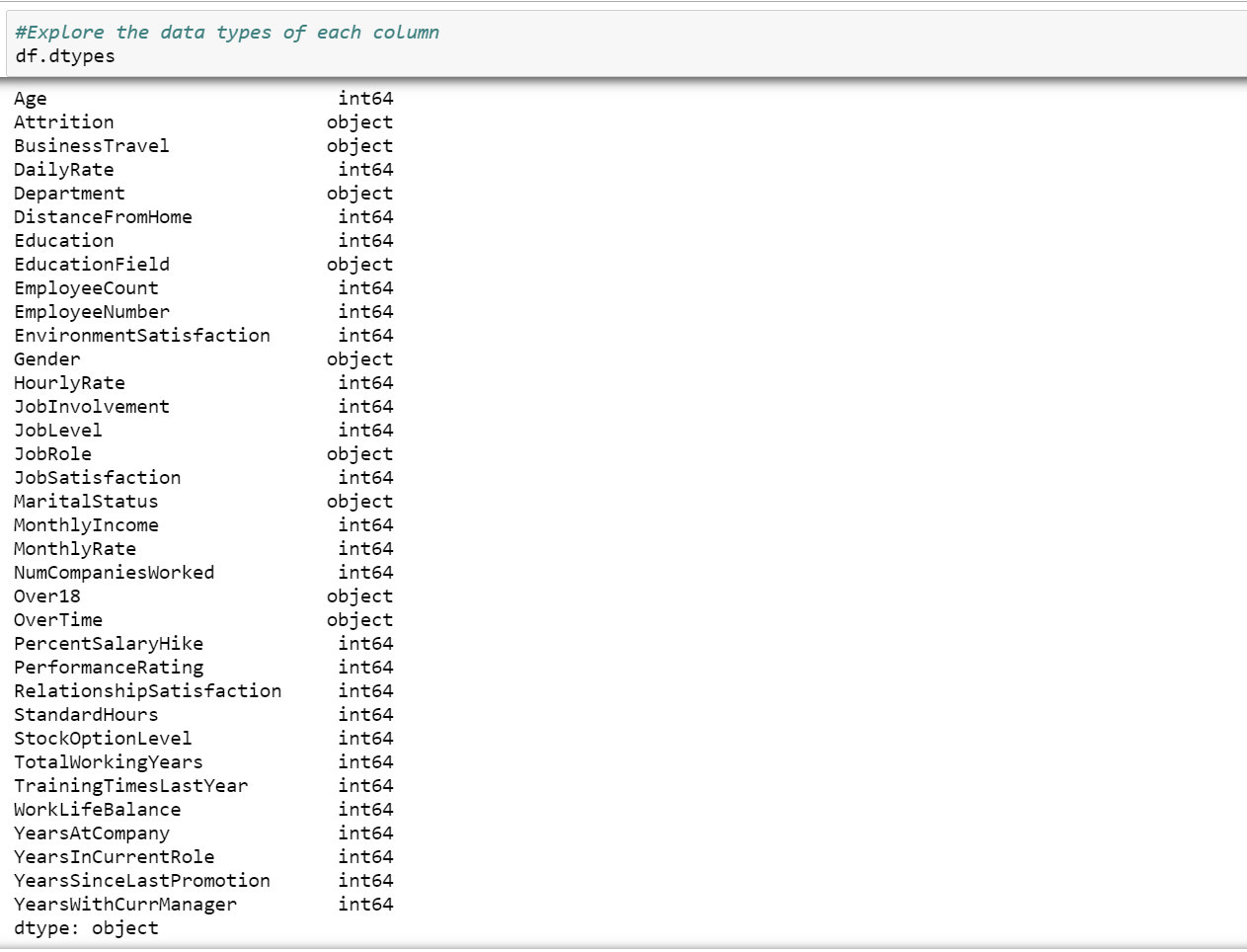
**Data Analysis**

Let’s start with looking into the data and understanding the data we have at our hands.

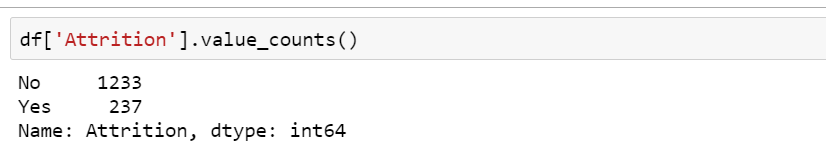
* Checking the dimensions of the data: 
* Check the summary of the data:



* We need to identify the data in each column. Lets the check the data types in each column:



* Let’s identify the different values in attrition column and the count of each value present:

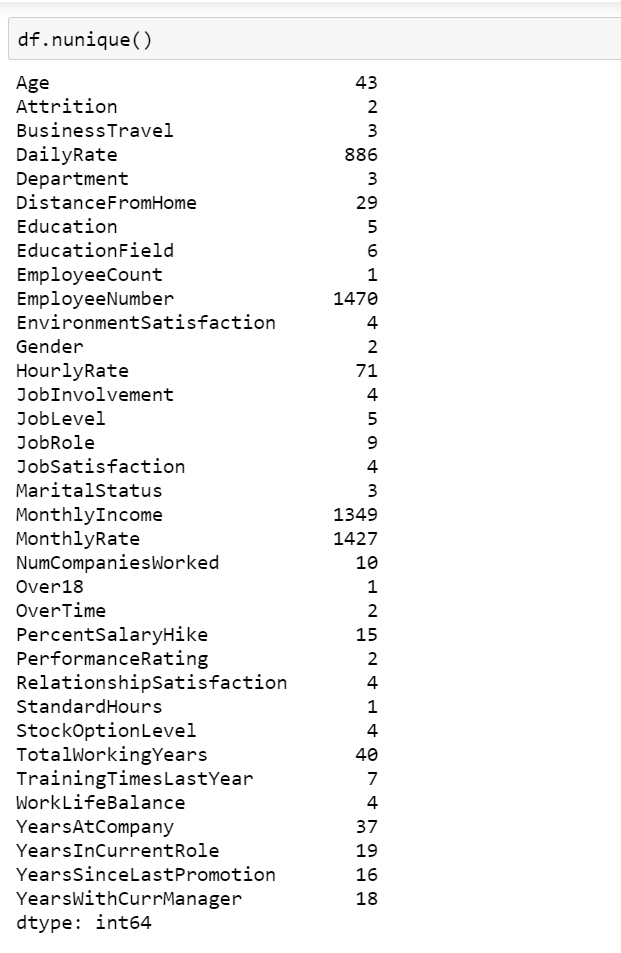


* From first look, doesn’t look like there are NA values in the dataset. Let’s check for the same and drop the NA rows and revalidate the dimensions to see the number of rows dropped.

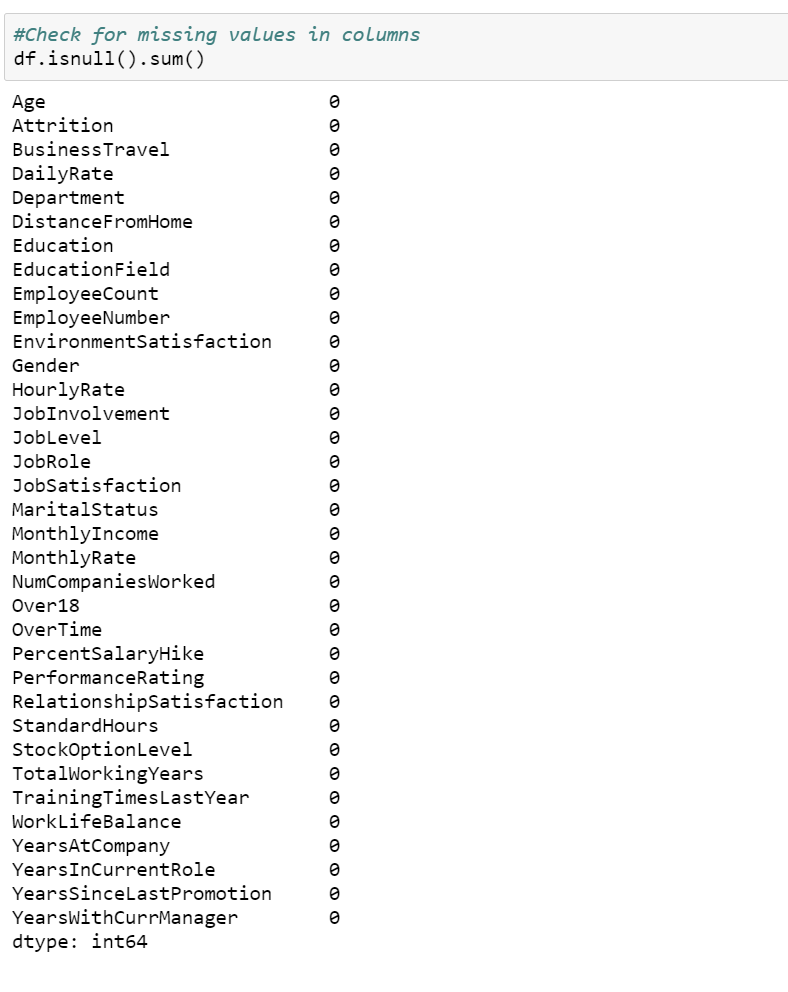


No rows deleted confirms our research that there are no NA values.

* Now we will investigate the unique number of values in each column:

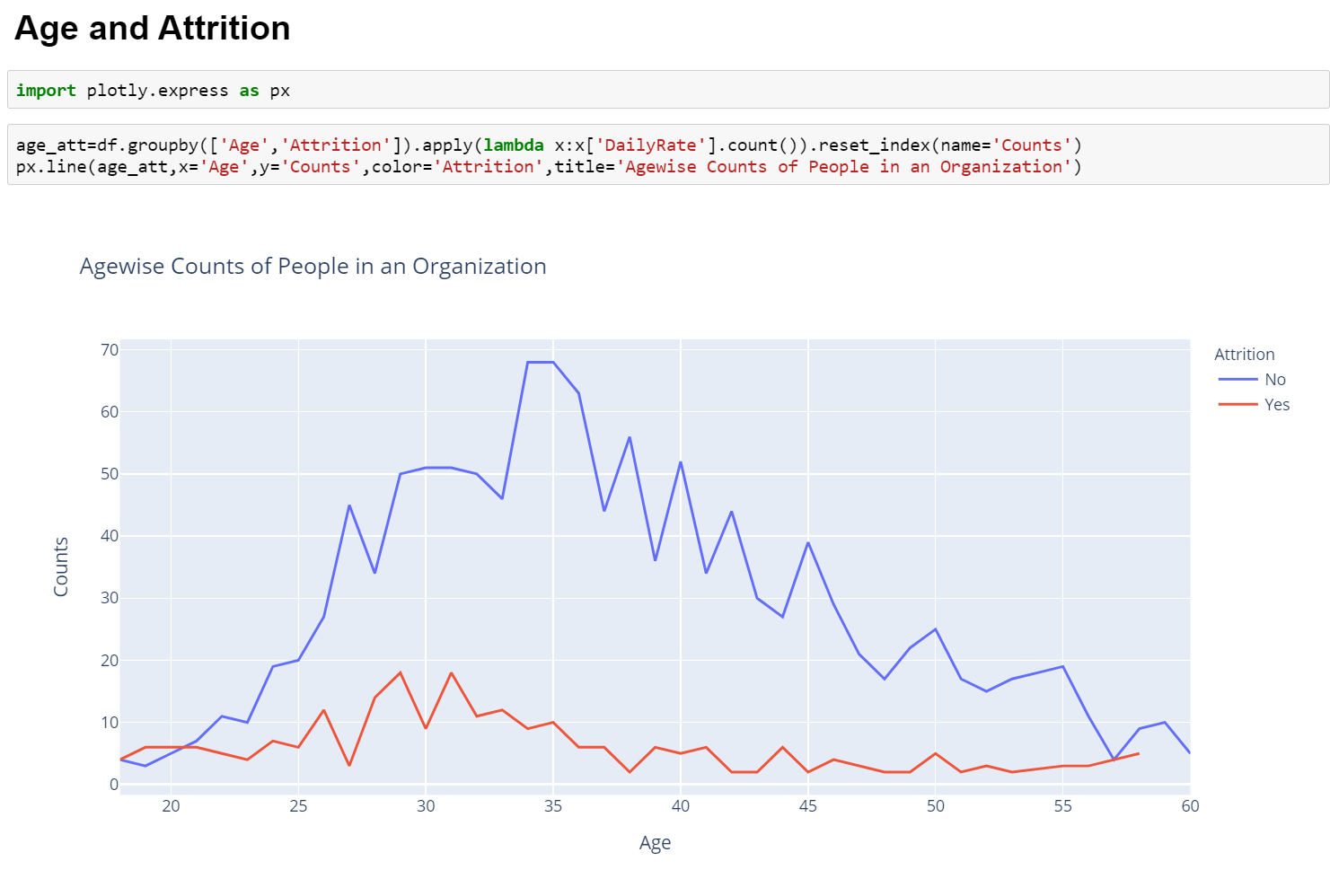


* Before we move ahead with our analysis, lets check for missing values if there are any:



As Attrition is our target variable, let us analyze it with different columns in our dataset.

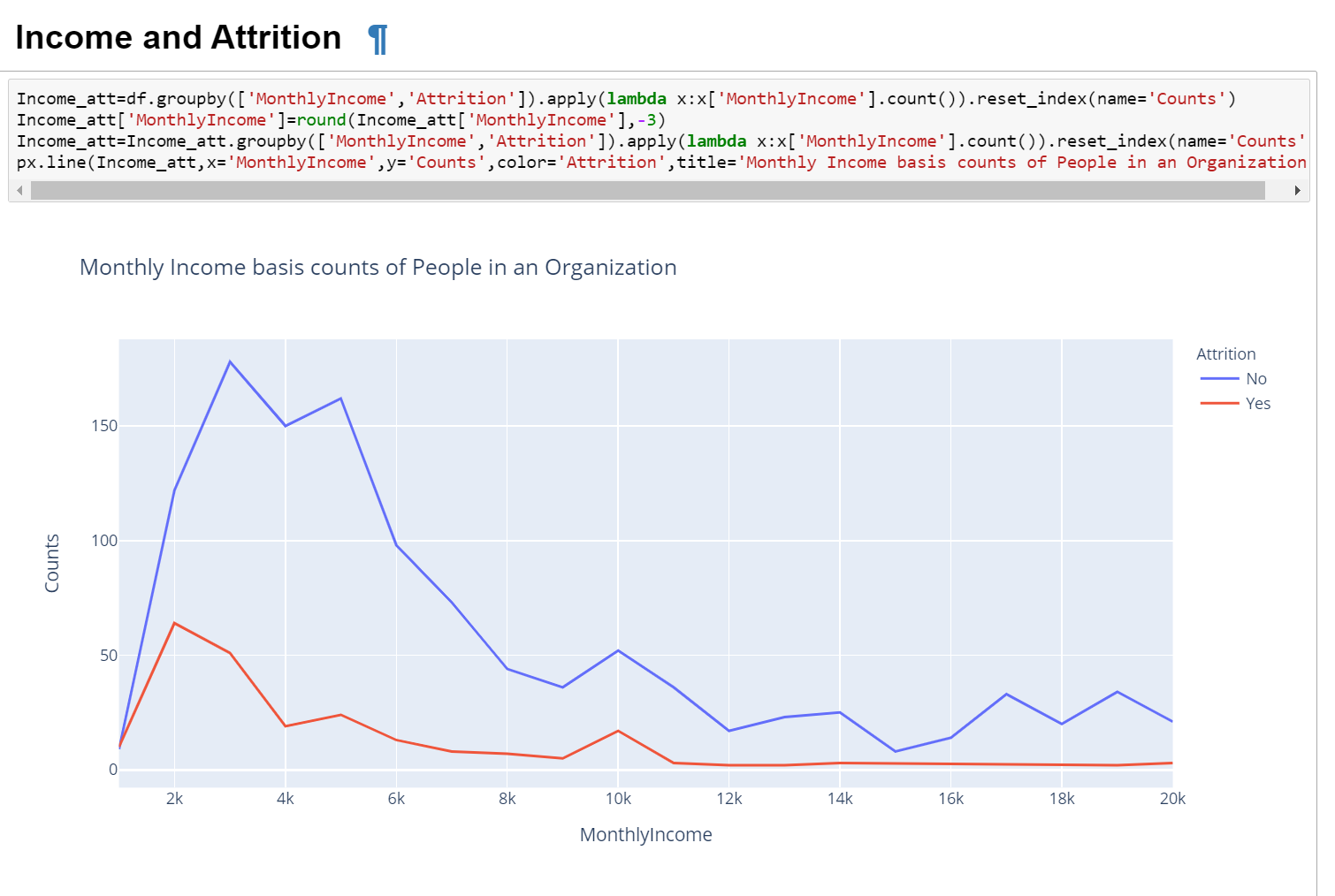
* Age and Attrition – We will plot age and attrition counting the attrition rate for each age.



As we can see in the chart above, the employees aged between 27-32 have maximum attrition rate. Also, another important observation that we can see is that the attrition rate keeps falling with the age.

Another observation from this graph that can be made is at a young age, between 18-21, the chance of employees leaving the organization is higher.

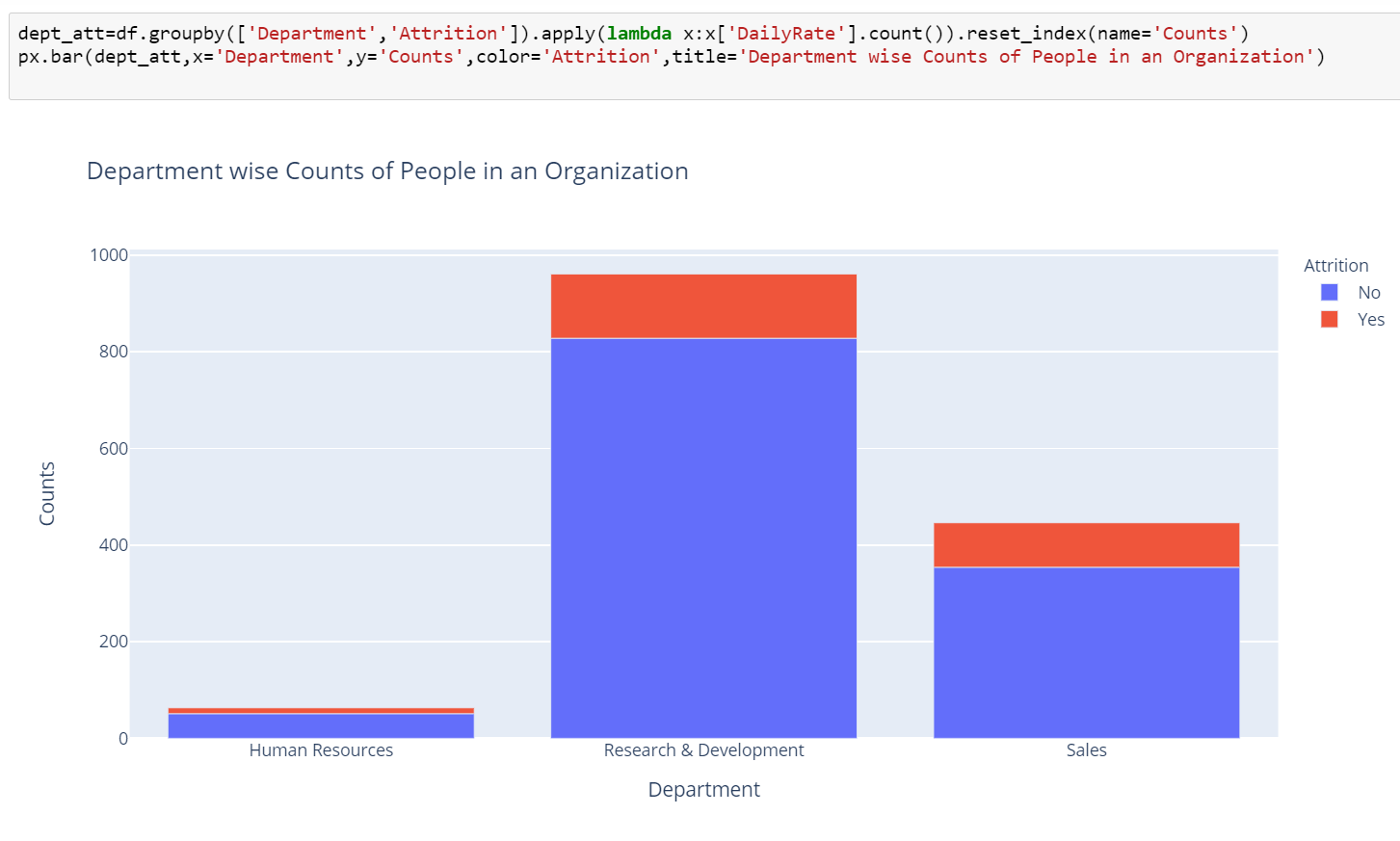
* Income and Attrition



As we can see in the above graph, attrition level is higher with employees with low level monthly income i.e. less than 5k. There is also a small spike at income level 10k.

The graph also depicts that when the income level is higher, the attrition rate is minimal.

* Department and Attrition



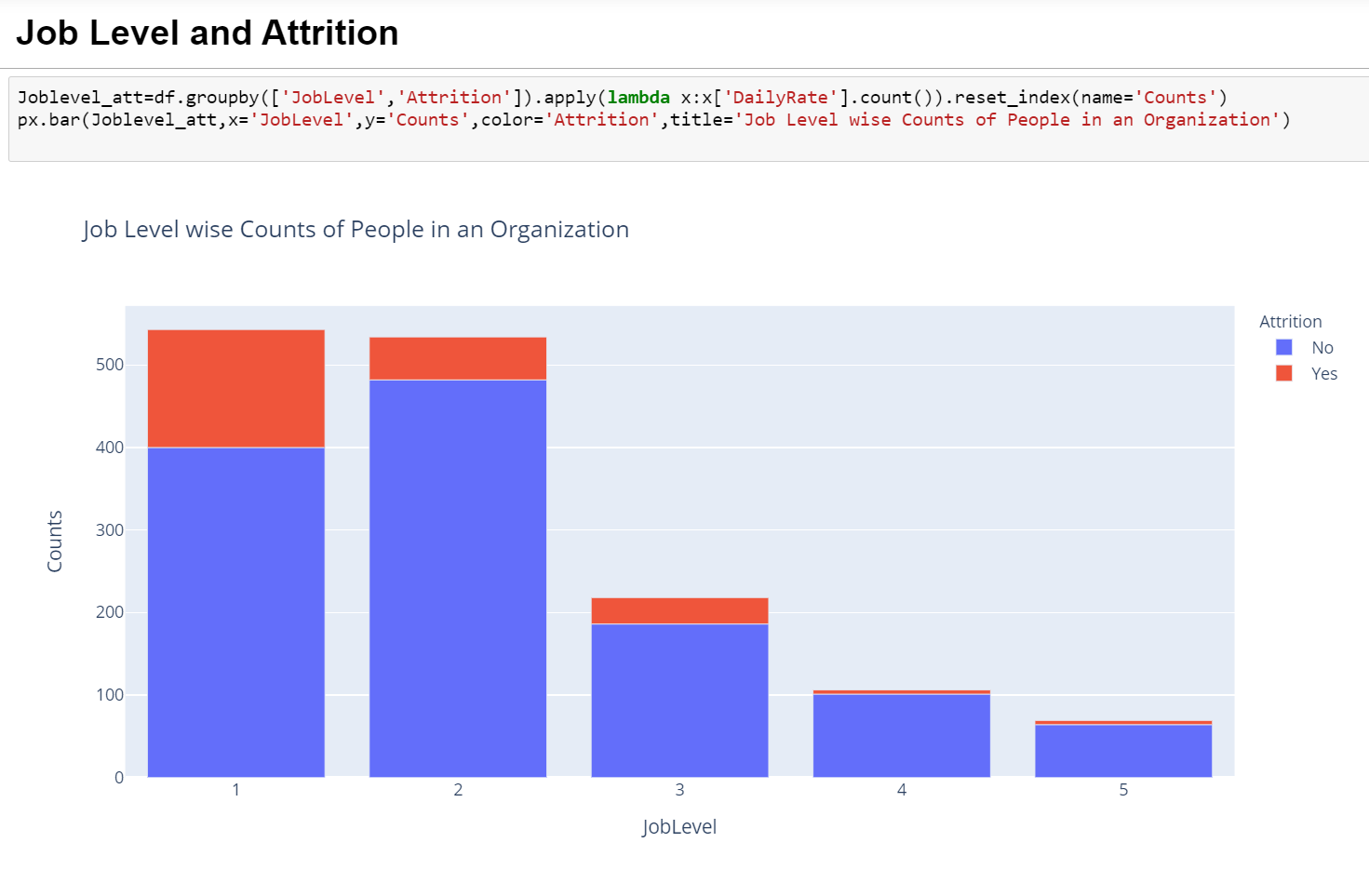
Let’s check the count of people in each department



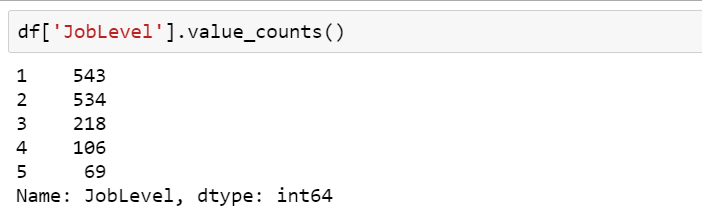
Based on the above analysis, we can see that the attrition rate for sales department is 92/446 ~ 20.63%, for Research and Development department is 133/961 ~ 13.84% and for Human Resources department is 12/63 ~ 19.05%.

The highest attrition is in the Sales Department and the lowest in the Research and Development department.

* Job Level and Attrition



Let’s check the count of people in each Job Level

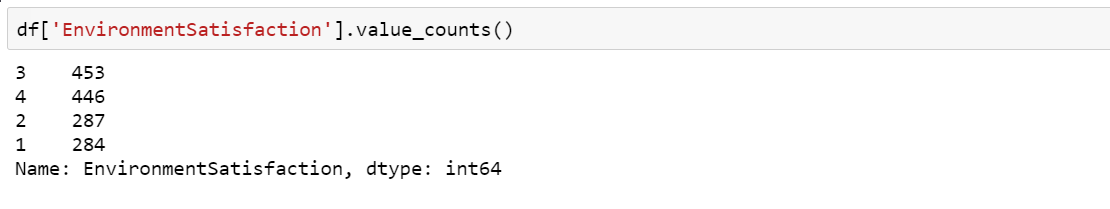


As we can see from the above graph, attrition level is highest at the lowest job level and lowest at the highest job level.

* Environment Satisfaction and Attrition

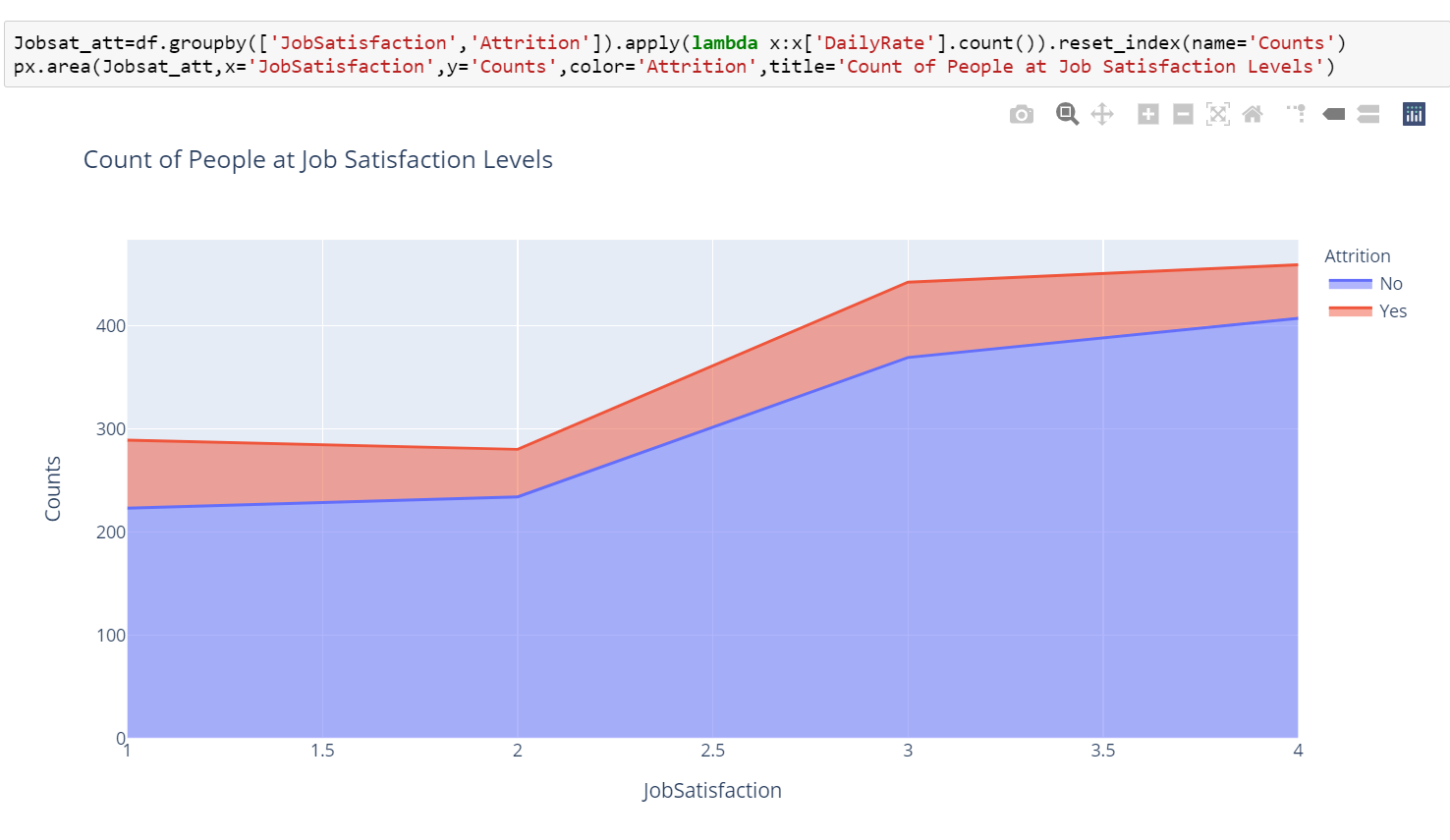


Let’s check the count of people at each Environment Satisfaction Level



We can clearly see that people with least Environment Satisfaction have the highest chances of leaving. Attrition rate is ~ 25.35% at level 1 Environment Satisfaction, ~14.98% at level 2, ~13.69% at level 3, ~13.45% at level 4. This indicates that at level 2, 3 and 4 the attrition rate is almost stagnant.

* Job Satisfaction and Attrition

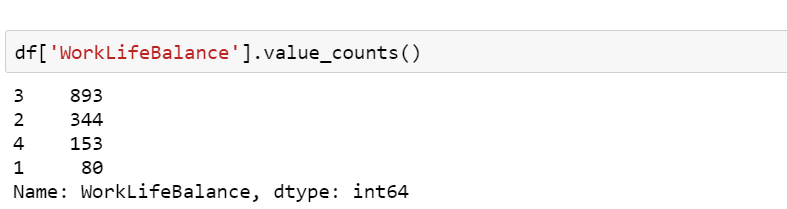


With an increasing job satisfaction, the attrition rate decreases as can be seen in the chart above. Also, from range 1-2, we can infer (as also seen above in the graph for Environment Satisfaction and Attrition), the attrition level falls, but raises from 2-3.

* Work Life balance and Attrition

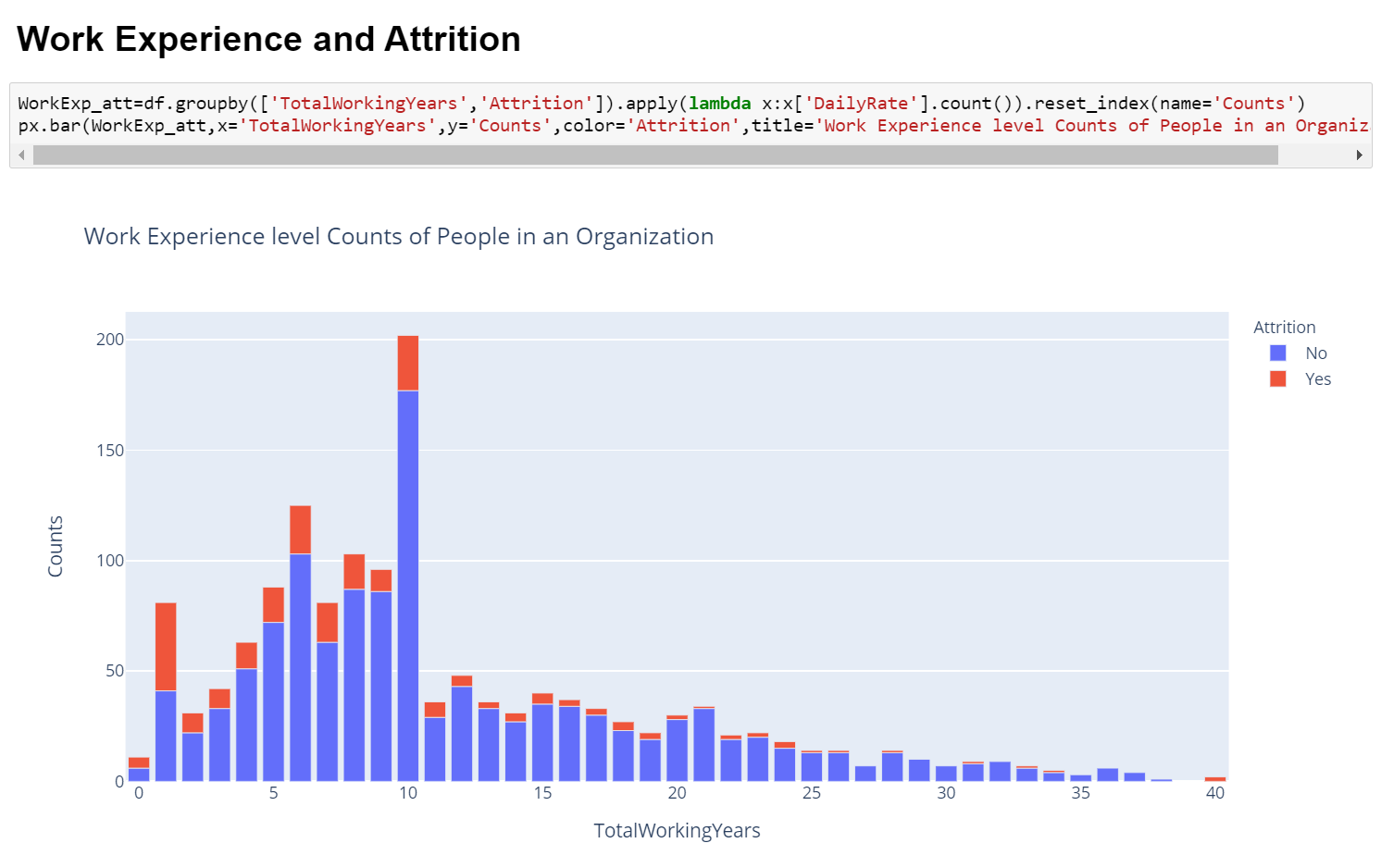


Let’s check the count of people at each Work Life Balance Level



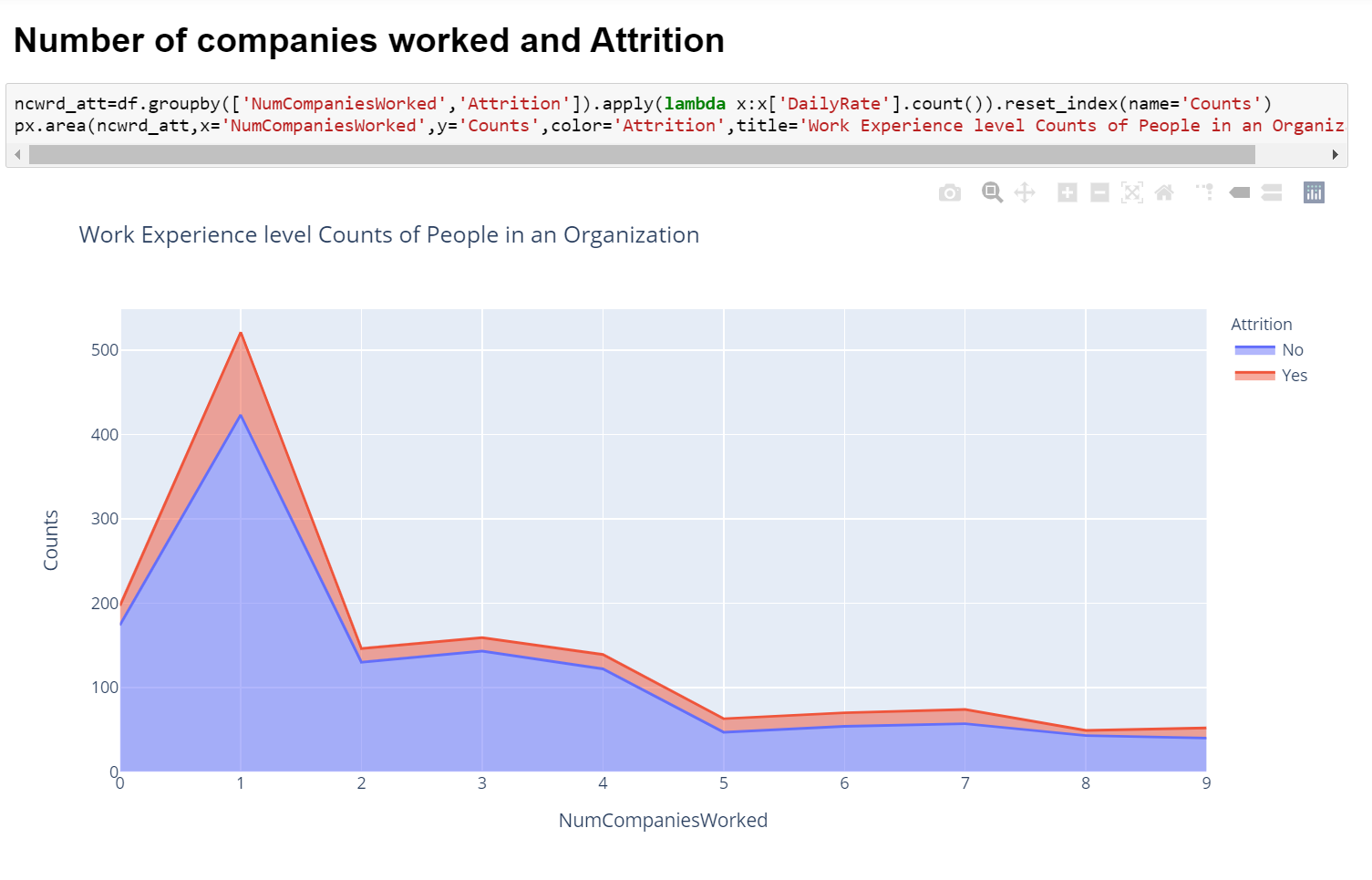
At Work Life balance level 1 attrition rate of employees is 31.25%, at level 2 attrition rate is 16.86%, at level 3 attrition rate is 14.22%, at level 4 attrition rate is 17.64%. As the work life balance increases from level 1, attrition rate is almost stagnant.

* Work Experience and Attrition



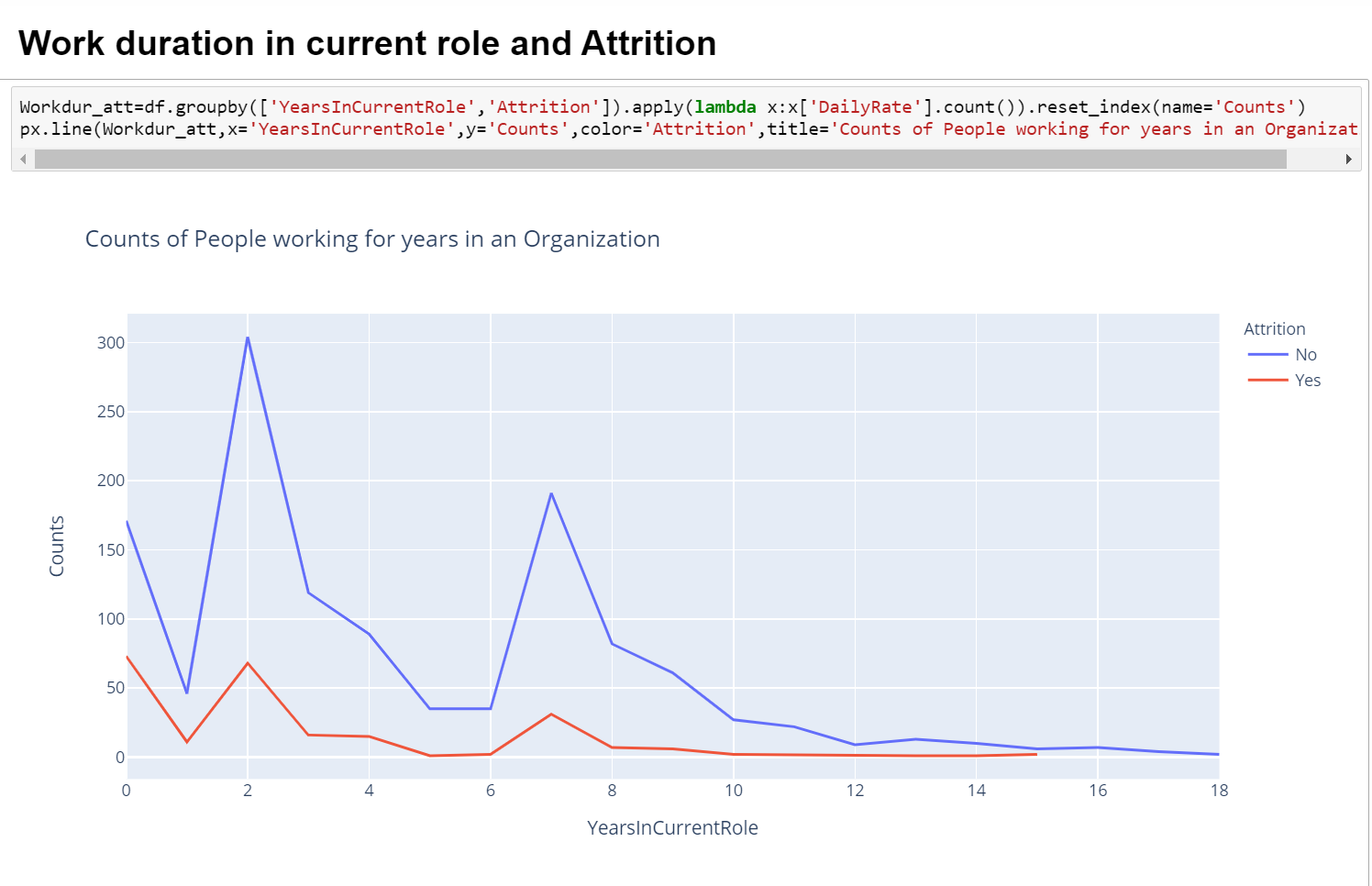
We can clearly see from the plotting that as the work experience of employees increase, the attrition rate tends to fall.

* Number of Companies worked and Attrition



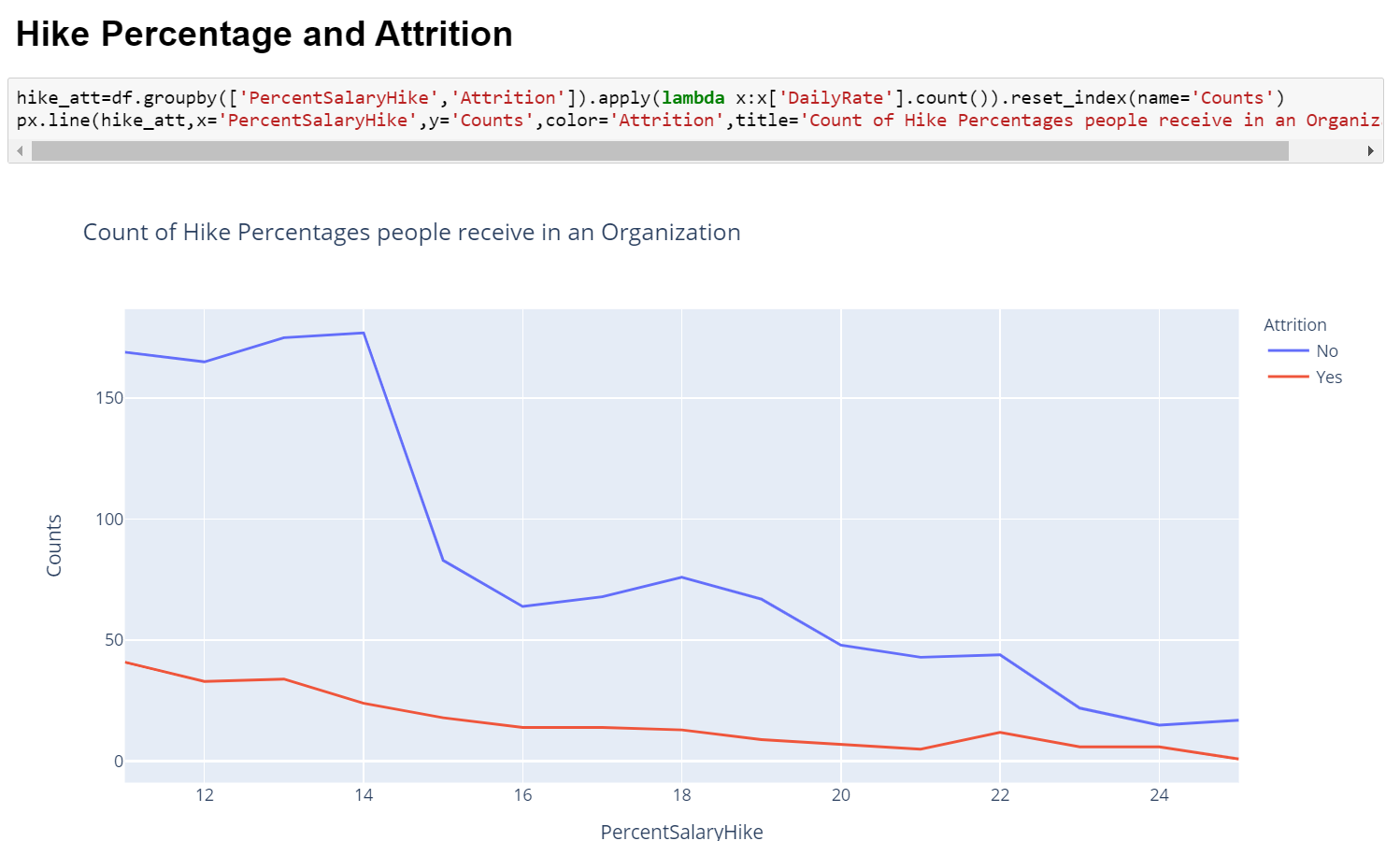
As seen from the chart above, clearly, employees who started their career with the company- or have switched to the company in the initial years of their career, have a higher chance of leaving the organization to a different company. People who have gained much experience- working in multiple companies, tend to stay in the company they join after they gain higher experience.

* Work Duration in Current Role and Attrition



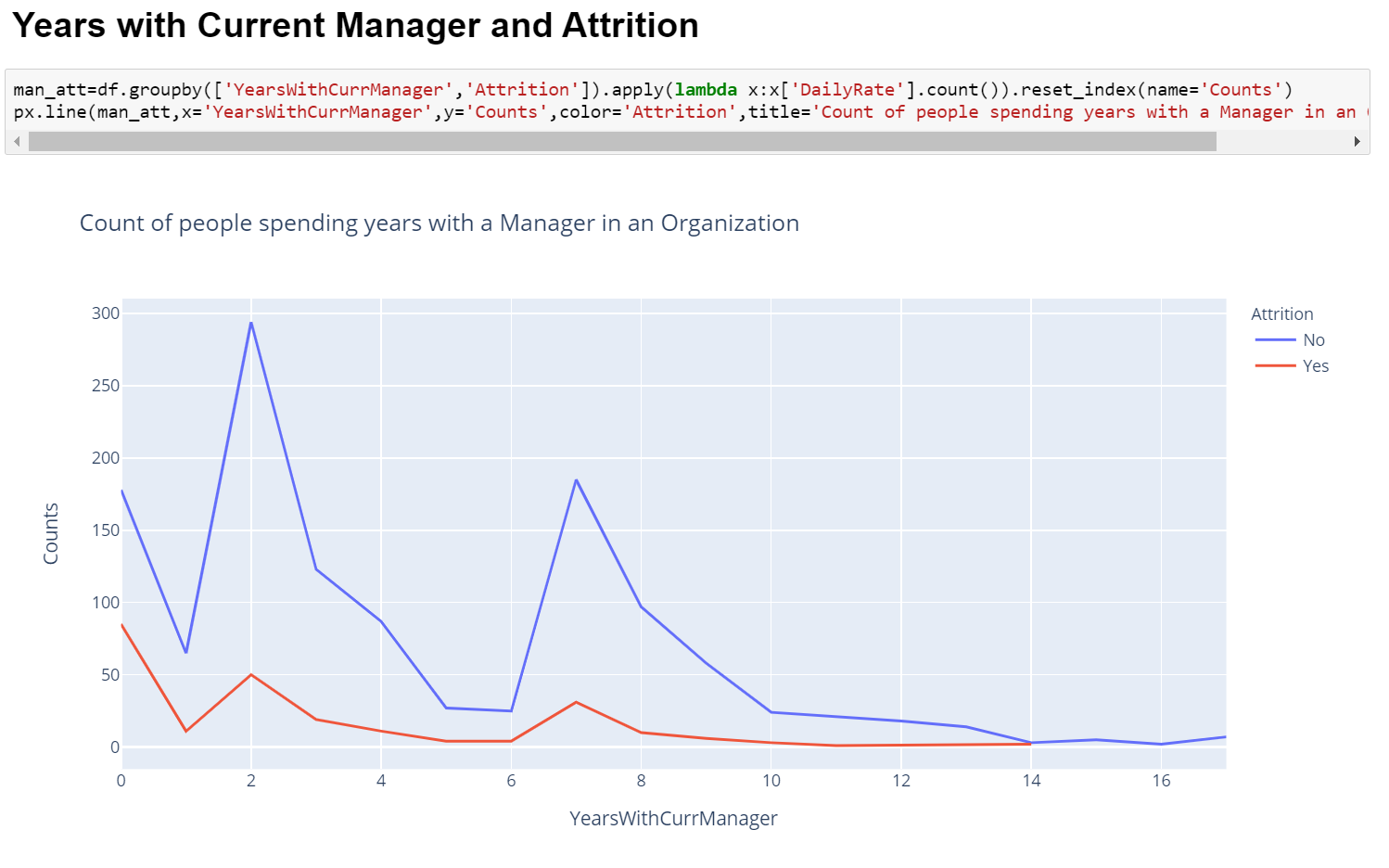
As we can infer from the above graph, attrition rate is higher in the starting years on their role. When people are in the same role for a long period of time, they tend to stay longer.

* Hike Percentage and Attrition



As we can see and is the case usually among employees, Higher the hike, lower the attrition rate and vice versa.

* Years with Current Manager and Attrition



The information that we can infer from the graph is that at the very start, where the time spent with the manager is relatively less- attrition rate is at the highest level. At an average span of 2 years, attrition rate is still higher. When the relative time spend with a manager is very high- attrition rate is very low.

**EDA Concluding Remarks**

As we have seen above in our data analysis, there are many factors which may influence the attrition rate of employees in company. The data we have in each column contributes to our understanding of why an employee may or may not leave the company.

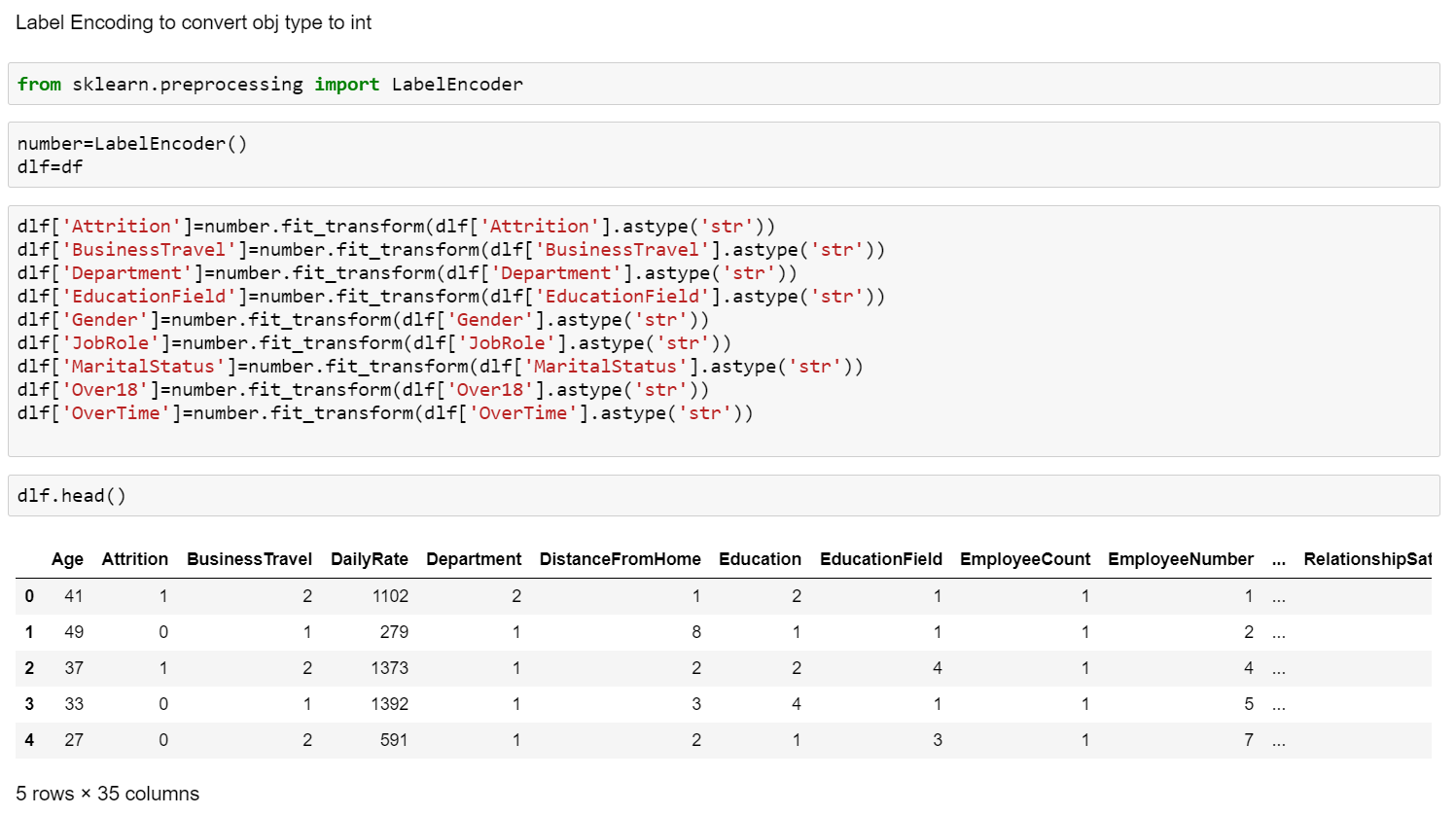
Based on the analysis we have done, and the insights we have on the columns with respect to our target variable, attrition, we will process and prepare the data in the next steps for enabling to us create a machine learning model to predict attrition rate.

**Pre-Processing Pipeline**

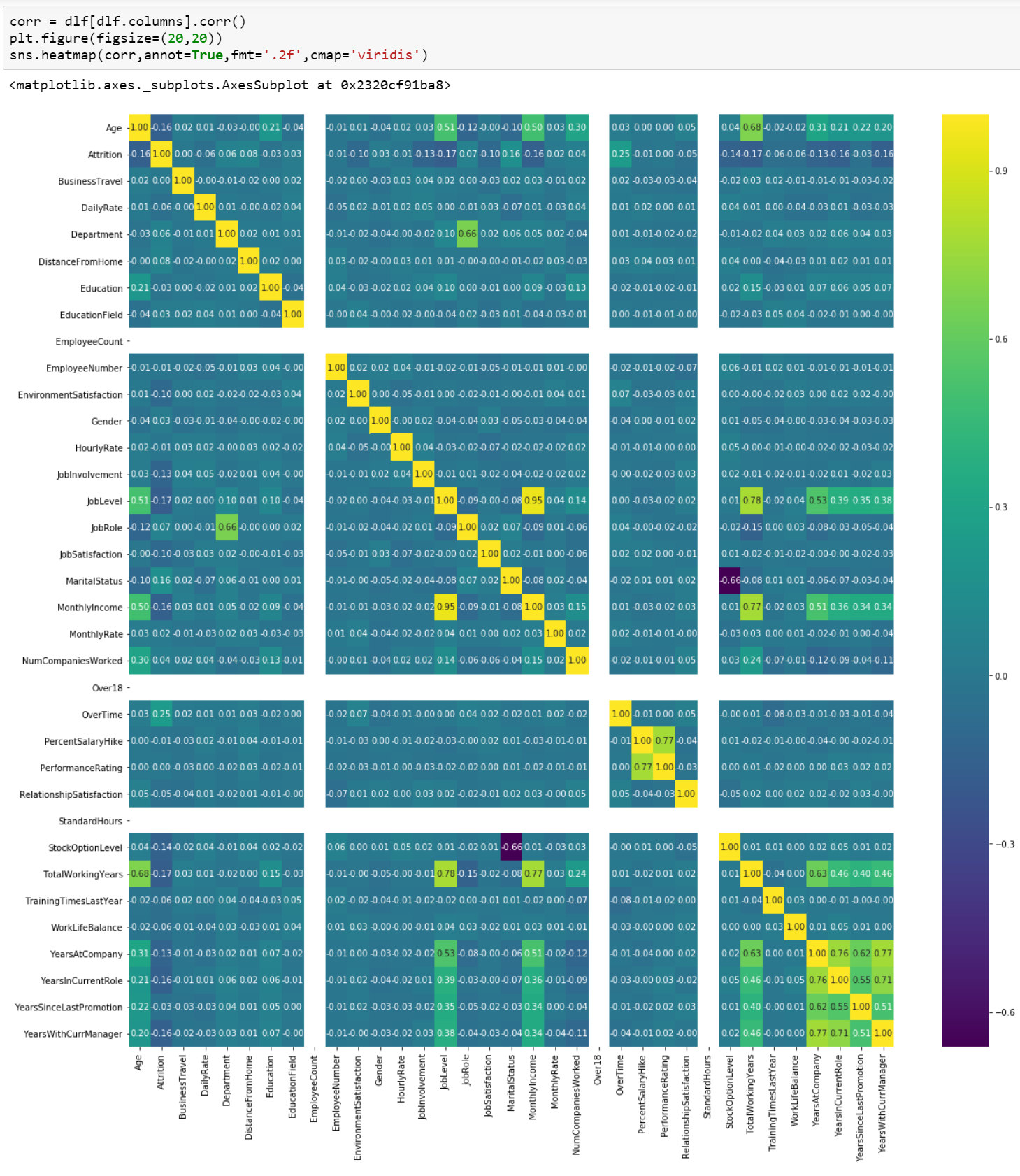
As there are multiple columns with Object datatype, we will be first transforming them to int datatype using Label encoder.

Label encoding is simply converting each value in a column to number.

We will use label encoder to convert the columns and check the dataset after applying the same.



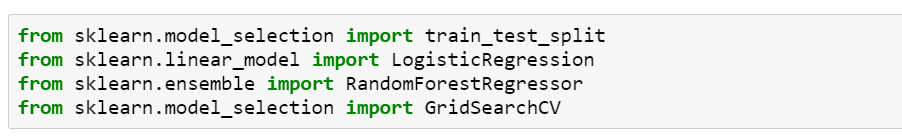
Now, let’s check the correlation of columns with our target variable attrition in order to eliminate any columns which might not be useful in creating our model for prediction.



Based on the correlation of all the columns with Attrition, we can clearly see that the Business travel, performance rating, relationship satisfaction, monthly rate, percent salary hike, hourly rate, employee number columns have a very low co-relation with attrition so this can be removed from the data being used to create our model.

**Building Machine Learning Model**

We will be importing the libraries required for creating train and test data and for creating our machine learning model.



We will be dividing our data into X and Y, where X are the columns used for creating the model and Y is the target column.

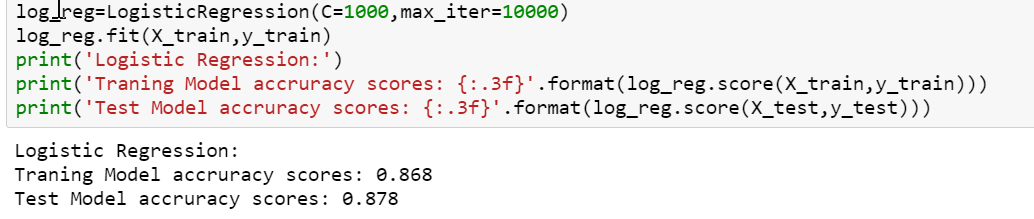


Now we will be splitting the data into Train and Test data for training our data and then testing out the same for verifying the accuracy.



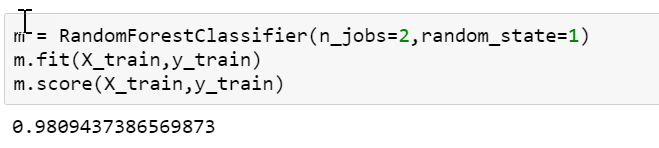
For this problem, we will use Logistic Regression and Random Forest Classifier to create out model. Let’s see which one is best suited in this scenario and for our processed data.

**Logistic Regression**



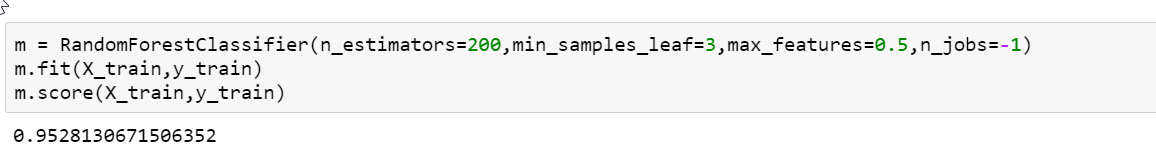
87% accuracy for training and test models. We will see if Random Forest is doing better than this or if Logistic Regression would be the winner for this.

**Random Forest Classifier**



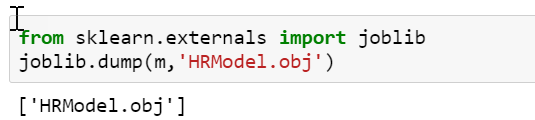
98% accuracy. Random Forest is the clear winner as the accuracy is much higher than that we are getting with the model created by Logistic Regression.

To make sure our model is not over fitting or underfitting, we will add more metrics while creating the model and Hyper tune it.



**Saving the Model**

We will be using joblib library to save our model which can be used later with any other data or production environment.



**Concluding Remarks**

We have successfully created a model to predict attrition rate with 95% accuracy using the data at hand. We were able to do the same as we thoroughly analyzed the model initially and studied the impact of each column/factor contributing to attrition of an employee.

Companies are using these models extensively to predict attrition rate and plan their hiring so that business is not impacted and its business as usual always.