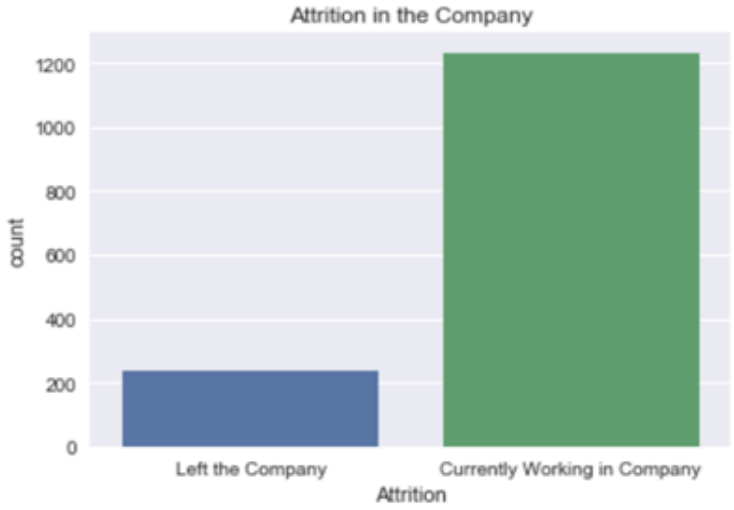
Exploratory Data Analysis

Introduction:

We have 32 features consisting of both categorical as well as the numerical features. Response variable is 'Attrition' of the employees which can 1 and 0 (representing 'Yes' and 'No' respectively). This is what we will predict.

Now, I will try to analyze visually the trends in how and why employees are quitting their jobs. For that, I will deep dive into the details about features and their relationships between each other.

Target Variable (Attrition):



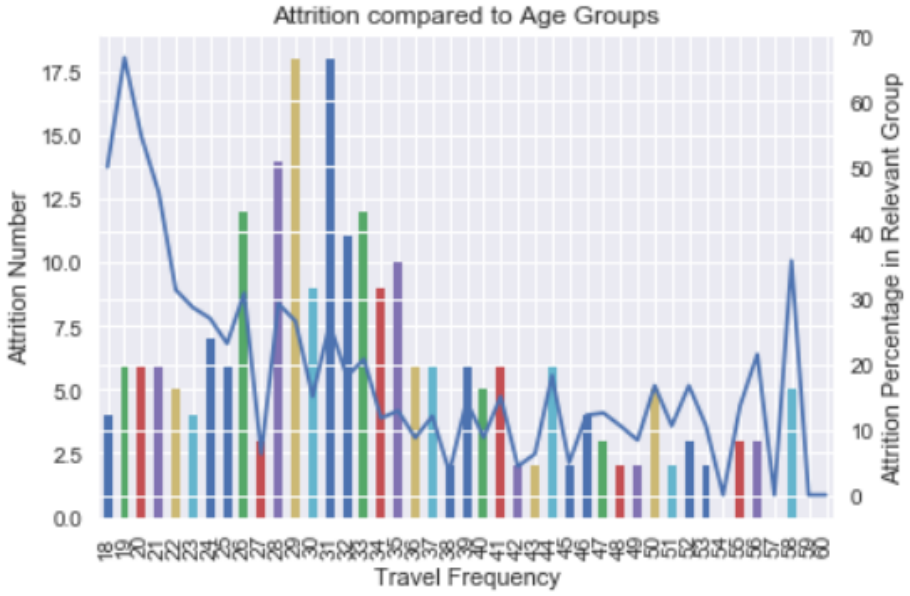
In the company, there are 1470 employees.

237 employees who compose 16% of the total number of employee left the company for some reasons.

Besides that, **1233 employee** is currently continuing to work in the same company.

Features:

Age:

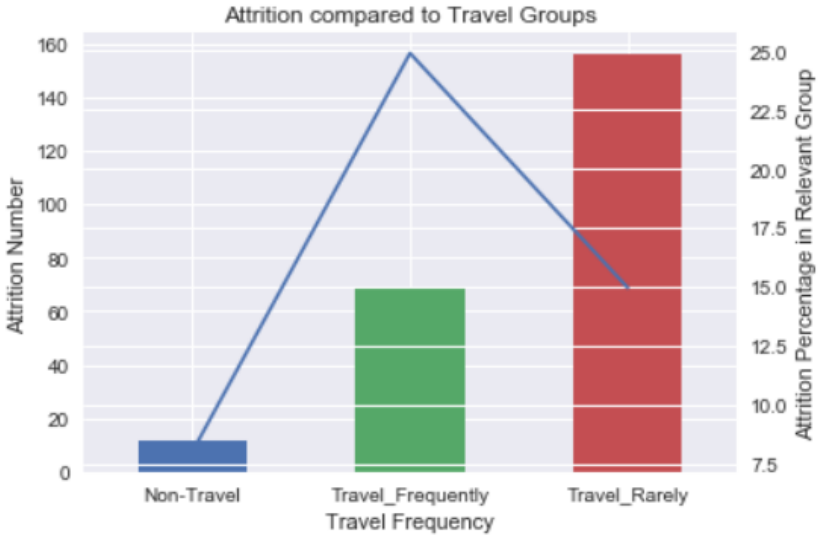


If we evaluate overall attrition number in the company, **26-35 age group's attrition number** is the highest comparing to other age groups. In this group, we have 19% of employee attrition (116 out 606). That makes up 49% of all attrition in the company (22 out of 237 employees).

In **18-21 age group**, young employees are more likely to leave the company. Their attrition proportion to their age group is approximately 54% (22 out of 41 employees) and that makes up 9% of all attrition (22 out of 237 employees).

35-60 age group generally prefers to secure their job in the same company. On the other hand, there are two age groups which come forward to deal with the attrition problem in the company.

Business Travel:



In the company, most of the employee travel rarely or don't travel according to their job description. That group compose the 81.1% of entire company(1193).

The rest of the company employees which is 19.9% of them has to travel frequently (277 out of 1470 employees).

The highest attrition number with 156 belongs to the **employees who travels rarely.** That is approximately 15% of employees in that group (156 out of 1043). But when you put this number overall attrition, it makes up 65.8% of all attrition in the company(156 out of 237).

if we look at the attrition percentage of relevant travel group, the **employees who are traveling frequently** are in the danger zone. Because they have the highest attrition proportion, which is 24.9%, in their individual travel group(69 out of 277). That group's attrition rate composes of the 29.1% of overall attrition in the company (69 out of 237). **Employees who don't travel** in their current role have the lowest attrition rate, which is 8%.

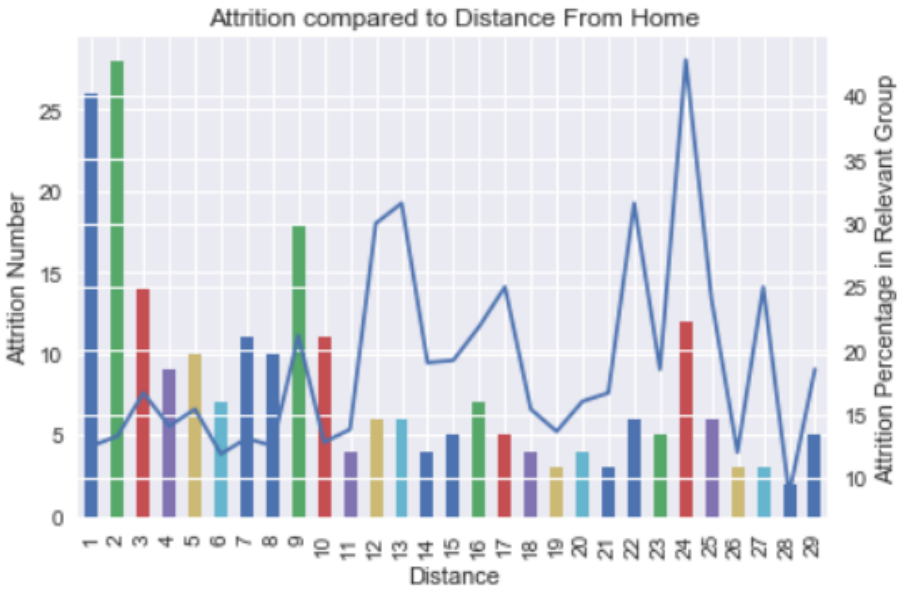
Department:

There are three departments in the company. **Research & Development Department** has the most attrition number in the company. 13.8% of **Research & Development Department** employee left the organization. In numbers, it is equal to 133, which makes us the 56.1% of all attrition in the company. Actually, that attrition is a big number for company, but compared with other departments, **Research & Development Department** has the lowest attrition rate in itself as an individual department.

Sales Department has mostly been affected by the attrition. Because 20.6% of its employees left the organization. This is the highest number compared to the other two departments. That attrition makes up 38.8% of the attrition in the company (92 out of 237).

Human Resources Department follows the **Sales Department** in terms of being affected by attrition itself. 19% of this department employee left the company. But this is not that huge number in terms of whole attrition in company. **Human Resources Department** employee attrition makes up 5% of all attrition in the company (12 out of 237).

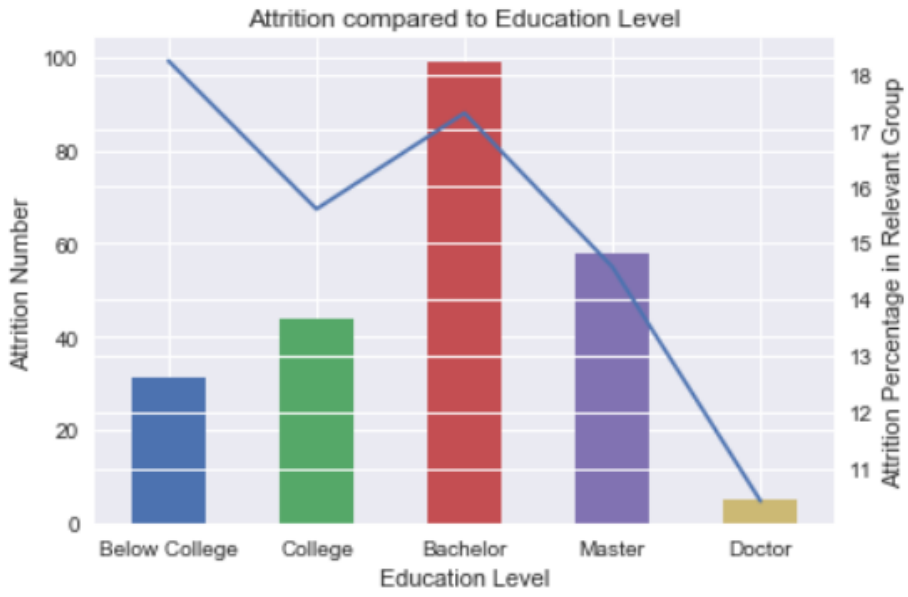
Distance From Home:



Employees whose homes are 1-3 miles far away from the company seem to leave the company more than others. This group's attrition rate is 28.6% of all company's attrition. But there are 502 employees in that distance and attrition rate in their group itself is just 11.5%.

Attrition rate seems to decrease as the distance from home increases. There are some exceptions for that like in 9th miles and 24th miles. This might be outliers or there might be some other reasons for that.

Education:

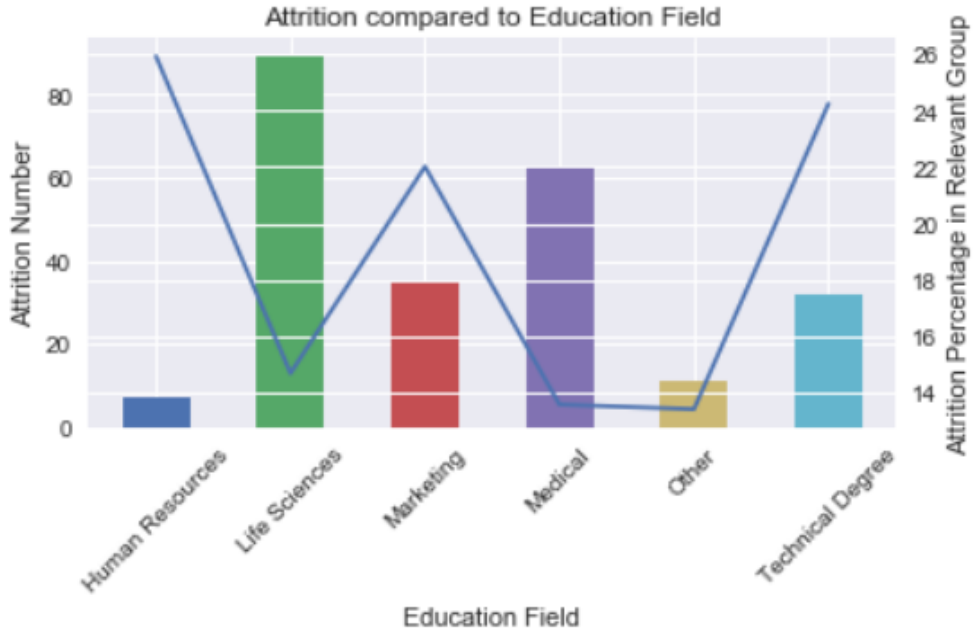


Employees who have bachelor degree have the most attrition number in the company, which makes up 41.8% of all attrition in the company.

Employees who have Ph.D. degree composes the least attrition number in the company.

Employees who have the master, college, and below college degrees are follower of **employees who have bachelor degrees** in terms of the attrition number in the company respectively.

Education Field:

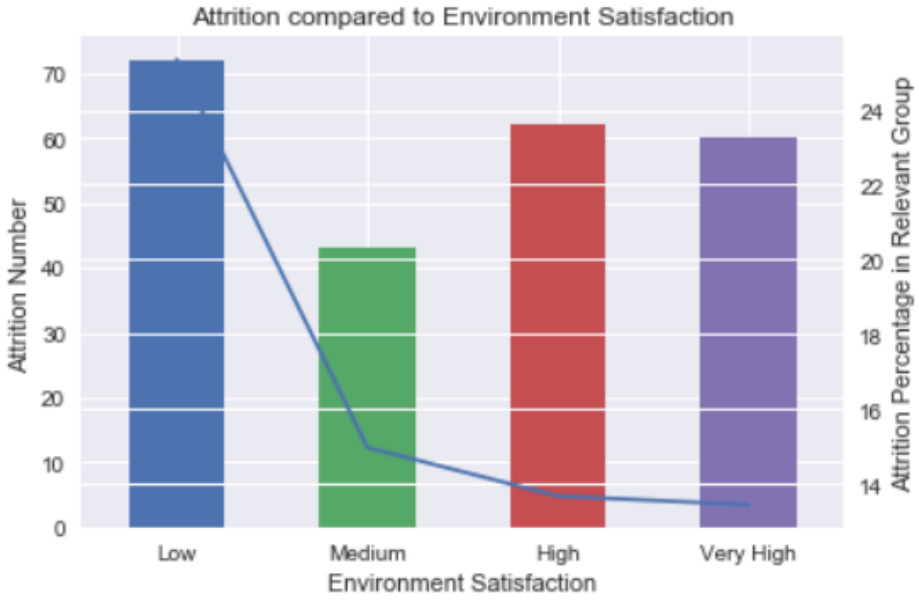


Employees who have Life Science education level have the most attrition number which makes up the 37.5% of all attrition (89 out of 237). But that composes only 14.7% of attrition within Life Sciences field.

Medical education level have the second highest attrition number which makes up the 13.57% of all attrition (63 out of 237). But that composes only 14.7% of attrition within Life Sciences field.

Besides that, **Human Resources, Technical Degree, and Marketing fields** are mostly affected by the attrition respectively. Their approximately 22-26% employees leave the company.

Environment Satisfaction:



As it may be expected, there is a high attrition rate in the **low satisfaction environment**. That composes the 30% of the whole company's attrition number.

Shockingly, in the **high and very high satisfaction environment**, there are still 13.6% of the these each group's employees leave the company. That attrition composes of the 51.5% of the whole company's attrition. This result might tell us that environment satisfaction is not the one of the main reasons for attrition in the company.

Gender:

Male employees are more likely to leave the company than female employees.

Job Involvement:

59% of all employee's job involvement in the company is in the **high** category. The highest attrition number is also observed in high job involvement category. 125 employees in this group, which composes the 52.7% of all attrition, left the company.

Medium job involvement category is following the **high** category group in attrition number with 71 employees.

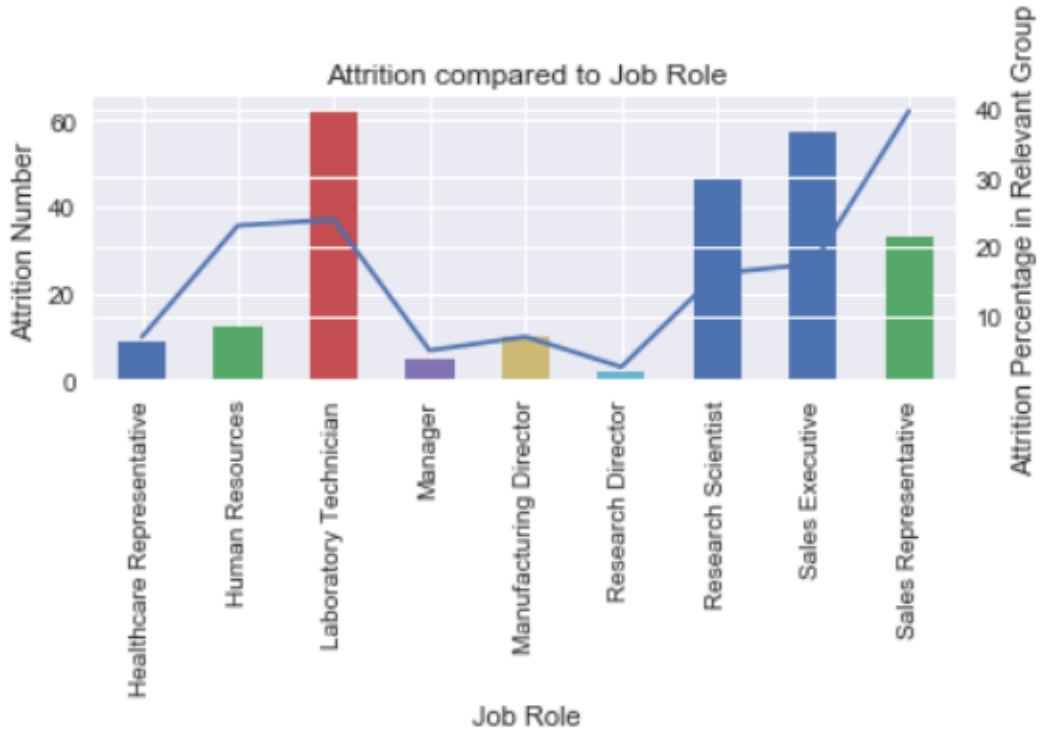
Low job involvement category has the highest employee leaving proportion within individual category when it is compared to the other categories. 33.7% of **Low** Job involvement group left the company.

Job Level:

With an increase in job level, there is a decrease in attrition number throughout the company. The highest attrition is observed in the **job level-1**.

143 employees in the job level-1, who compose the 60.3% of all attrition, left the company.

Job Role:



Laboratory Technician has the most attrition number with the 26.2% of all attrition in the company (62 out of 237 employees). **Sales Executive** and **Research Scientist** are following the **Laboratory Technician** in attrition throughout the company with the 57 and 47 employees respectively. Those both job roles' attrition composes 44% of whole company's attrition.

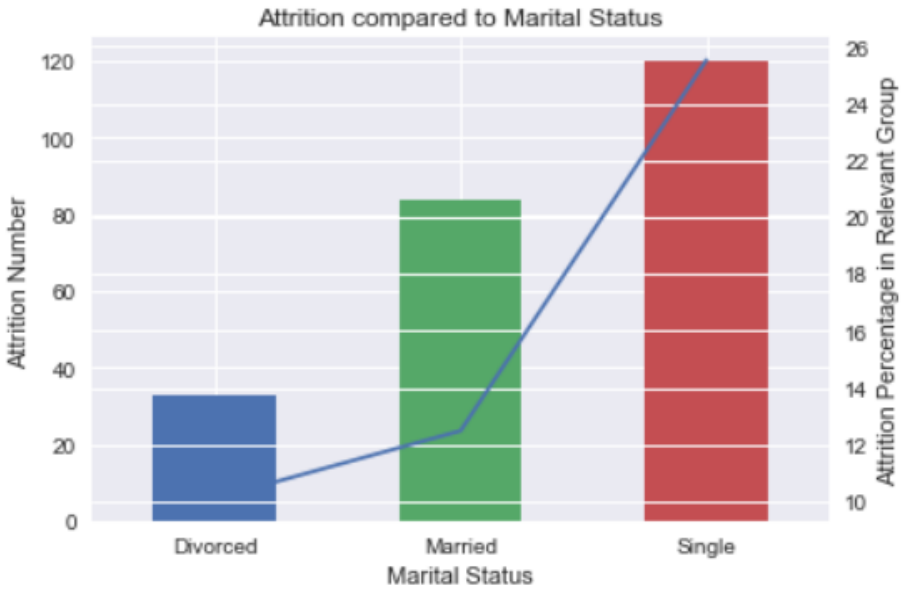
On the other hand, **Research Director** job role has the lowest attrition number not only in the company (2.5%) but only within its own job role(0.8%).

Job Satisfaction:

In **high** job satisfaction, surprisingly employees leaves the company most and their attrition composes 30.8% of company's attrition. From this picture, I assume that job satisfaction should not be the main reason for employees to leave the company.

As it may be expected, in **low** job satisfaction, employees leave the company more than other groups except **high** satisfaction. They compose 27.8% of all attrition in the company.

Marital Status:



Single employees are more likely to leave the company. They have the highest attrition number and compose of the 50.6% employees who left the company.

Married and **Divorced** employees are the followers of **Single** employees in the attrition number of the company.

Monthly Income:

Around 2500 dollar and below monthly income level, there is a high attrition.

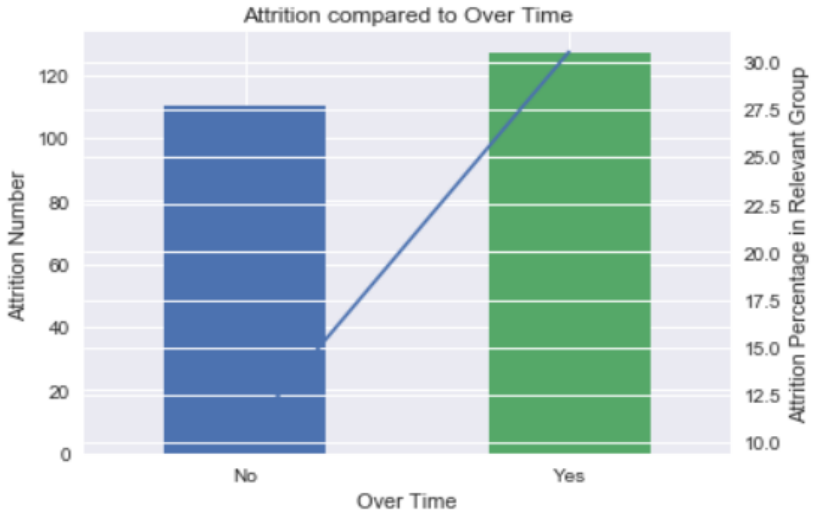
Between 2500-10000-dollar monthly income level, there is an inverse pyramid attrition picture as it may be expected. It is so ordinary to have this graph since the more money employee earn, the more they get satisfied and not willing to leave the company.

Numbers Companies Worked:

If employees have **one company experience** before current company, they are more likely to leave the company. They have the highest attrition number and compose of 41.3% all attrition in the company. Besides, if **employees don't have any experience** in other company, they have the second most attrition number.

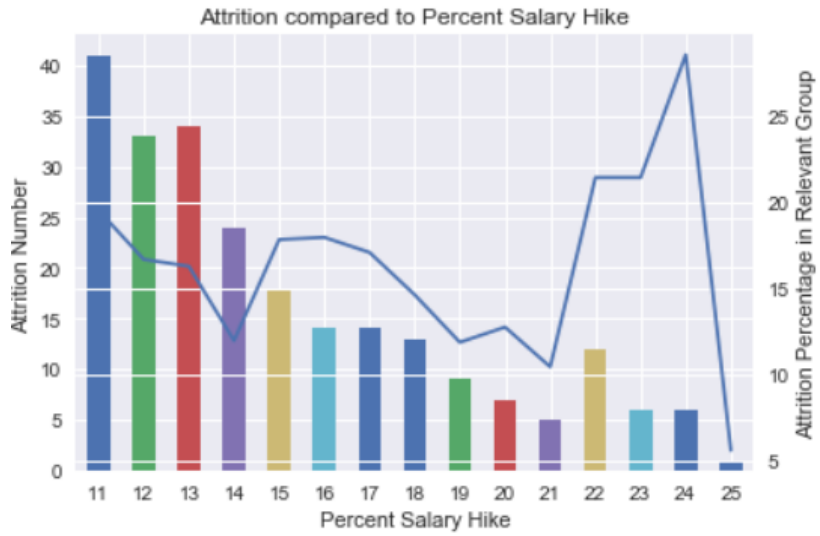
Also, employees, who has more experience such as **working in 5,6,7, and 9 companies** before the current company, have the highest attrition in their individual experienced group.

OverTime:



28.3% of employees have the overtime work in the company and they have higher attrition number than employees who don't have. There is not a significant difference between these two groups' attrition number. But if you compare individually both groups, over time employees are mostly affected by the attrition.

Percent Salary Hike:



As it may be expected, the higher percent salary hike is, the more employees are likely and willingly to stay in the current company.

Performance Rating:

Performance rating has two category such as 3 and 4. Not surprisingly, **performance rating 3 group** has the highest attrition number and compose 84.3% of all attrition in the company (200 out of 237 employees).

Relationship Satisfaction:

High and **very high** relationship satisfaction level have the most attrition number respectively and compose of 52.7% all attrition in the company. I assume that relationship satisfaction is not the one of the reasons for attrition in the company.

Stock Option Level:

If **stock option level is 0**, there occurs a huge attrition in the company and it composes the 65% of the all attrition in the company. Besides, as the stock option level decrease, there is a decrease in attrition number.

Total Working Years:

Employees who have one year working experience are more likely to leave the company and compose the 16.9% of all attrition throughout the company. In addition to that, employees who have 5-10 years' experience have also second highest attrition percentage throughout the company and it compose the 45.1% of all attrition.

Training Times Last Year:

Employees who has **2 and 3 times training last year** has the most attrition number respectively and both of their attrition compose the 70.5% of all attrition in the company.

Work Life Balance:

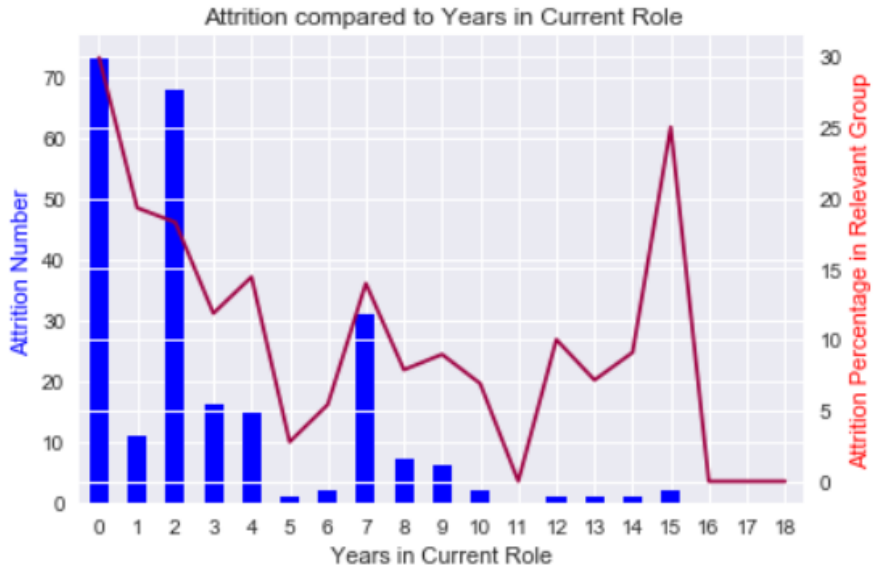
In general, work life balance is satisfactorily good throughout the company. For that reason, it might be misleading to make an inference based on this criterion about attrition.

As it might be seen above, employees whose work life balance are better, have the highest attrition number company. That justifies my assumption.

Years at Company:

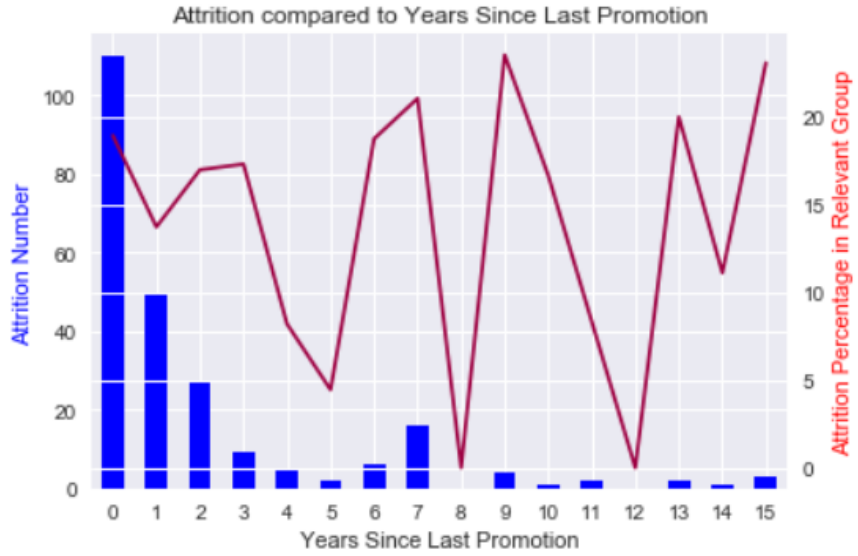
Employees who have up to 5 years' experience in the current company compose 52.8% of all company. They have the highest attrition number (68.4% of all attrition) within these 5 years. Especially, I want to emphasize the one-year experience attrition number. It is the highest attrition number and I assume that first year in the company is challenging or there is something that makes the employees leave the company.

Years in Current Role:



Employees who don't fulfill their first year in their current role are more likely to leave the company. That might be result of challenge or not satisfied with the current role. Besides that, after years in current role, employees are willing to leave the company. That might be result of looking for better opportunities in other companies.

Years Since Last Promotion:



Employees who have up to two years in the company since their last promotion compose the 74.6% of all company's employees.

Employees who don't fulfill his one year since the last promotion in the company are more likely to leave the company(46.4% of all attrition).

And employees who have one- and two-years' experience in the current company since the last promotion have the highest attrition number after the above group in the company respectively.

Years with Current Manager:

Most of the employee quit the company before completing their first year in the company. Other group who leaves the company most is the ones who work the current manager two years.

Other Features:

I also checked the 'Employee Number', 'Daily Rate', 'Hourly Rate' and 'Monthly Rate' features as I did the in previous features of dataset. But there is nothing significant to comment or visualize about these features. For that reason, I didn't include them in my notebook.

Feature/Variable Relationships:¶

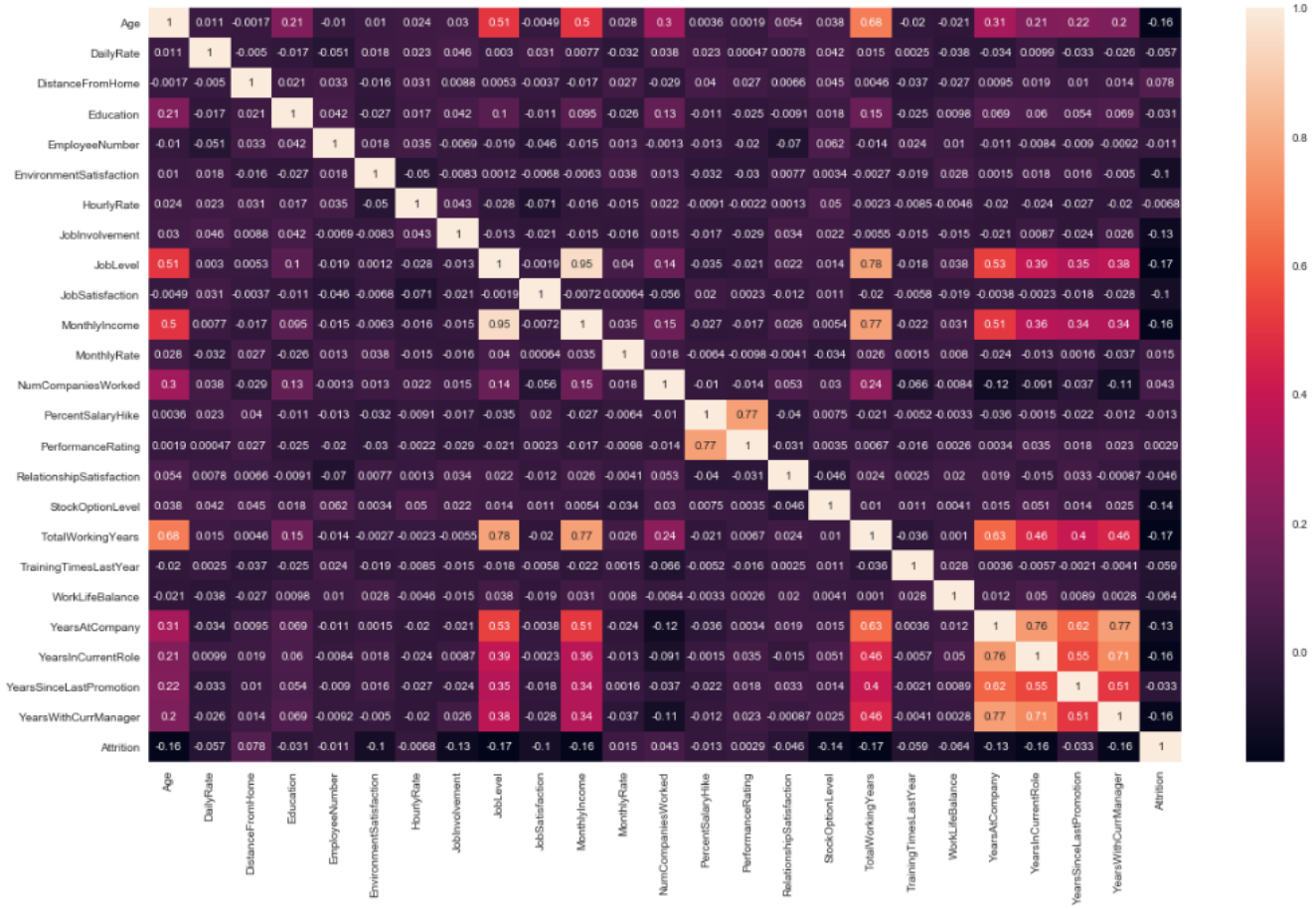
Here we will take a look at how variables related to each other. There are various methods/visualizations for this. I will use correlation matrix (heat map) for this purpose.

Correlation means association - more precisely it is a measure of the extent to which two variables are related. There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation.
A positive correlation is a relationship between two variables in which both variables either increase or decrease at the same time. An example would be height and weight. Taller people tend to be heavier.
A negative correlation is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. An example would be height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature).
A zero correlation exists when there is no relationship between two variables. For example, there is no relationship between the amount of tea drunk and level of intelligence.

Strength of Correlation:

Perfect: +1 -1
Strong: +0.9, +0.8, +0.7 -0.9, -0.8, -0.7
Moderate: +0.6, +0.5, +0.4 -0.6, -0.5, -0.4
Weak: +0.3, +0.2, +0.1 -0.3, -0.2, -0.1
Zero: 0

Correlation Matrix (Heat Map):



Based on the fact which is given strength of correlation chart, we can identify the features which have strong, moderate, weak and zero correlations between each other. I will just outline the strong and moderate correlations here.

Features which have strong correlations:

Percent Salary Hike and Performance Rating,

Total Working Years, Monthly Income and Job Level,

Years at Company, Years with Current Manager, and Years in Current Role,

Features which have moderate correlations:

Age has moderate correlation with Total Working Years, Monthly Income, and Job Level,

Job Level has moderate correlation with Years at Company and Age,

Total Working Years has moderate correlation with Years with Current Manager, Years Since Last Promotion, Years in Current Role, Years at Company, and Age,

Years at Company has moderate correlation with Years Since Last Promotion, Total Working Years, Monthly Income, Job Level,

Years in Current Role has moderate correlation with Years Since Last Promotion, Total Working Years,

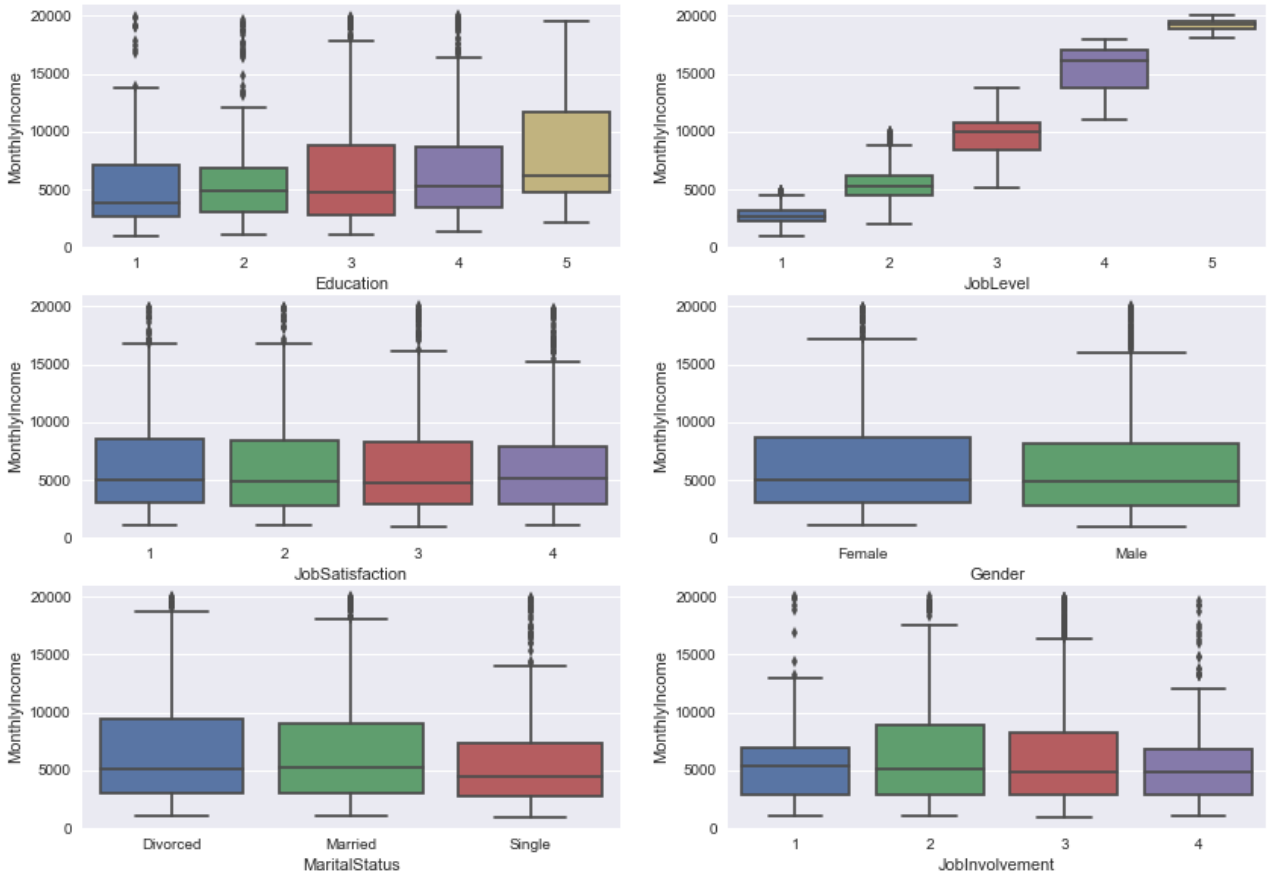
Years Since Last Promotion has moderate correlation with Years with Current Manager, Years in Current Role, Years at Company, Total Working Years,

Years with Current Manager has moderate correlation with Years Since Last Promotion, Total Working Years.

Generally, for the training model, we don't select features that have a strong correlation because it will have multicollinearity problem. Heatmap is a good way to detect this kind of situation. In this case, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion and YearWithCurrManager have strong correlations with each other which we should keep in our mind in further steps. Besides, that does not mean that is always case. We should try each variable and select the features which will give the best results in the model.

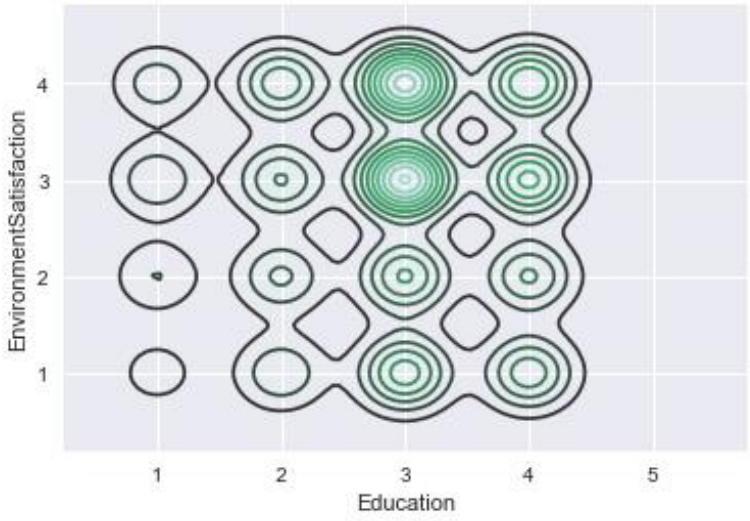
Box Plot:

I can make an assumption to see the relationship between features clearly. I don't need to show all of them right now but some that I think maybe matter a lot, such as Education, Job Level, Job Satisfaction, Job Involvement. I also would like to check other interesting features like Gender and Marital Status. I will use boxplot to show the trend.

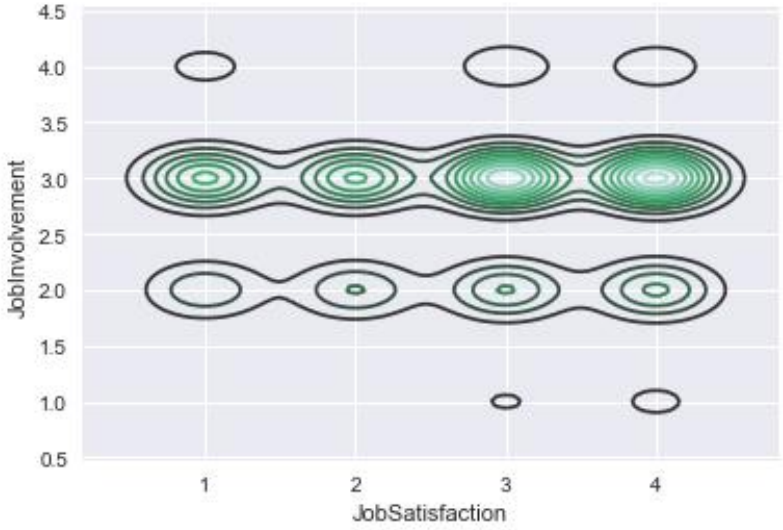


Job Level has an extremely effect on the income, apparently higher job level means higher income.

Density Plot:

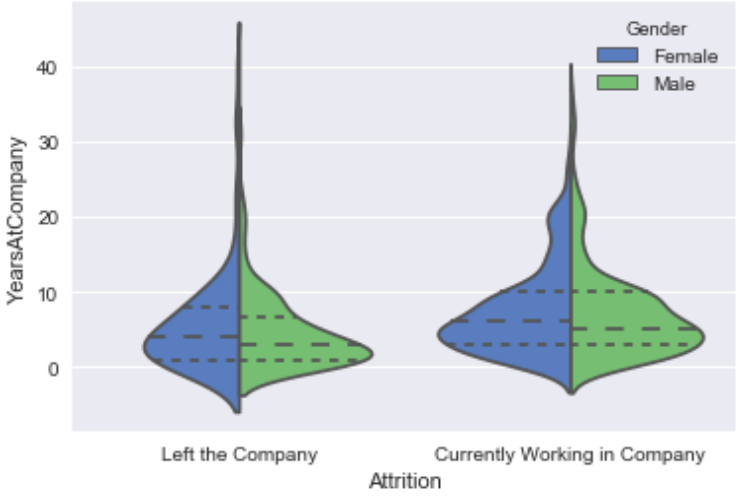


Look for job satisfaction at various job level. Job involvement is mostly high with high job satisfaction.



Look for satisfaction with environment with education levels of the employee. Employees with Bachelor and master's degree are mostly likely satisfied with the work environment.

Violin Plot:



Violin plots are like box plots, but they have the capability to explain the data better. The distribution of data is measured by the width of the violin plot. Here, I have plotted the number of years spent in an organization based on gender. The middle-dashed line shows the median. The lines above and below the median show the interquartile range. The denser part shows the maximum population falls under that range and thinner part shows the lesser population.