Applied Data Analytics – Assessment

By Prateek Chauhan (2931465)

Domain: HR Attrition Analytics

Attrition in human resources refers to the gradual loss of employees over time. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture and motivation systems that help the organization retain top employees.

A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits an organization from increasing its collective knowledge base and experience over time. This is especially concerning if the business is customer facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if the organization has constantly new workers.

Dataset: HR Employee Attrition and Performance.

Source: https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/
More information can be found here:
https://community.watsonanalytics.com/discussions/questions/3638/dataset-definition-for-samle-dataset-employee-attr.html

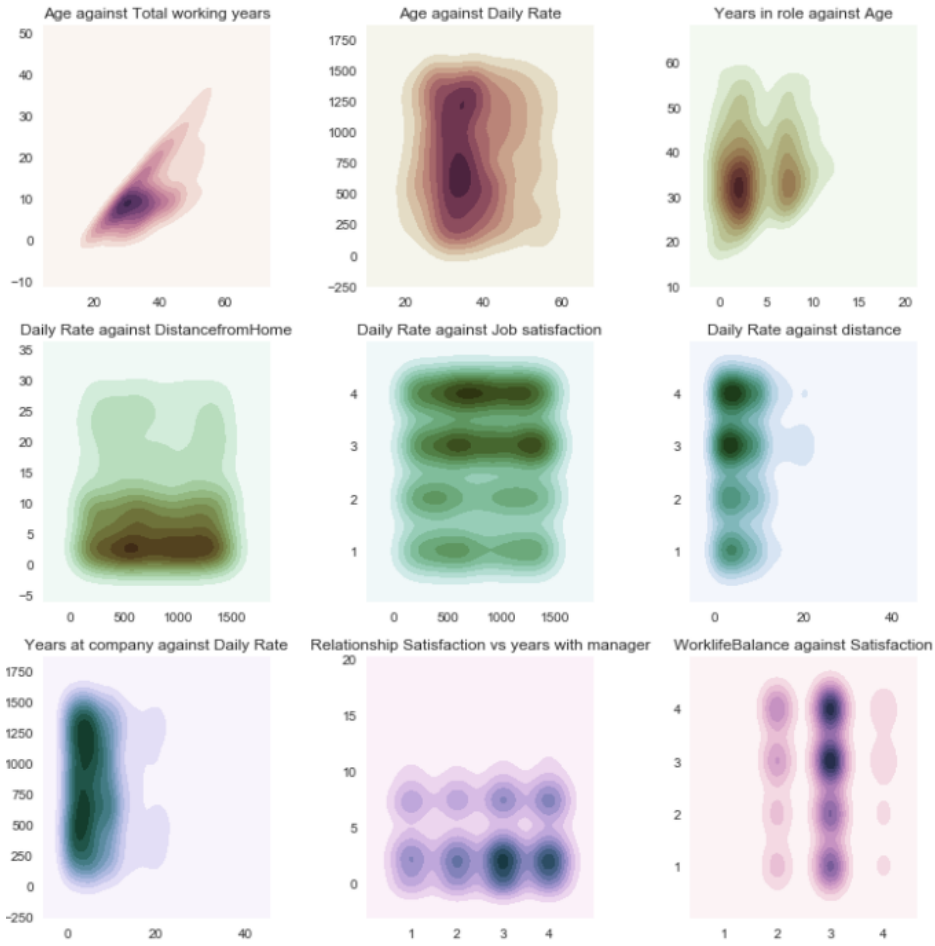
About the data: The dataset has 35 attributes and 1470 records. Some of these attributes are defined as:

Field	Metadata
Attrition	Role: Target
Education	Value Labels: 1. Below College 2. College 3. Bachelor 4. Master 5. Doctor
EnvironmentSatisfaction	Role: Input Value Labels: 1. Low 2. Medium 3. High 4. Very High
JobInvolvement	Value Labels: 1. Low 2. Medium 3. High 4. Very High

Job Satisfaction	Role: Input Value Labels: 1. Low 2. Medium 3. High 4. Very High
NumCompaniesWorked	Measurement level: continuous
PercentSalaryHike	Role: Input Measurement Level: continuous
PerformanceRating	Value Labels: 1. Low 2. Good 3. Excellent 4. Outstanding
RelationshipSatisfaction	Role: Input Value Labels: 1. Low 2. Medium 3. High 4. Very High
TrainingTimesLastYear	Measurement Level: continuous
WorkLifeBalance	Value Labels: 1. Bad 2. Good 3. Better 4. Best
YearsInCurrentRole	Measurement Level: continuous
YearsSinceLastPromotion	Measurement Level: continuous
YearsWithCurrManager	Measurement Level: continuous

Exploratory Data Analysis:

Distribution of the dataset – Generally one of the first few steps in exploring the data would be to have a rough idea of how the features are distributed with one another. To do so, kdeplot function from the Seaborn plotting library in python is used to generate kernel density bivariate plots. A density plot visualises the distribution of data over a continuous interval or time period. An advantage density plots have over histograms is that they are better at determining the distribution shape because they are not affected by the number of bins used. A histogram comprising of only 4 bins wouldn't produce a distinguishable enough shape of distribution as a 20-bin Histogram would. However, with density plots this isn't an issue. These are given as follows:



After observing the results of the kernel density plots, the next logical step would be to figure out that which of the attributes have a statistically significant categorical relationship with Attrition, which is our target attribute. To do this we perform Chi-square test and see the p-value of attributes against Attrition.

	Attrition	p < 0.05
PerformanceRating	0.990075	False
Education	0.545525	False
Gender	0.290572	False
RelationshipSatisfaction	0.154972	False
TrainingTimesLastYear	0.0191477	True
EducationField	0.00677398	True
Department	0.00452561	True
WorkLifeBalance	0.00097257	True
JobSatisfaction	0.0005563	True
EnvironmentSatisfaction	5.12347e-05	True
BusinessTravel	5.60861e-06	True
JobInvolvement	2.86318e-06	True
MaritalStatus	9.45551e-11	True
StockOptionLevel	4.37939e-13	True
JobLevel	6.63468e-15	True
JobRole	2.75248e-15	True
OverTime	8.15842e-21	True

From here, we can divide our attributes into two distinct groups:

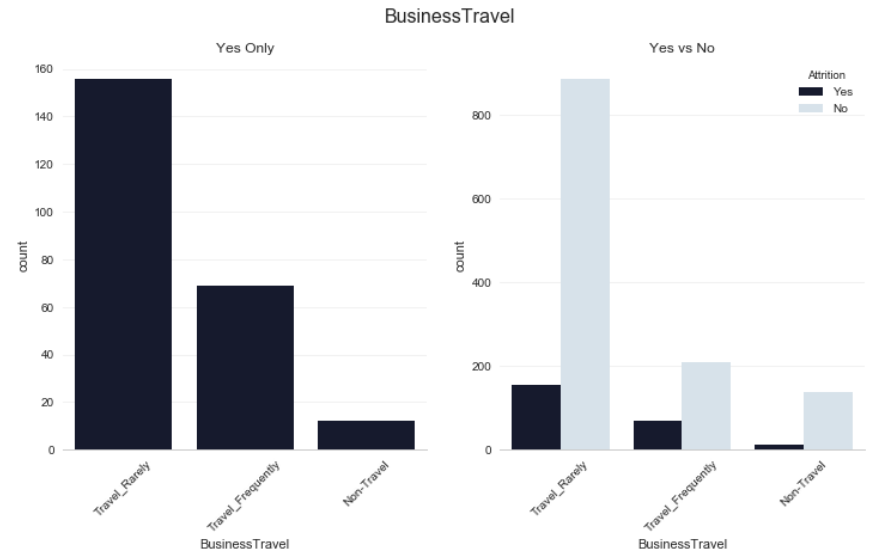
1. Attrition's Non-Significant Categorical Relationships:
   - Gender
   - Education
   - PerformanceRating
   - RelationshipSatisfaction

2. Attrition's Statistically Significant Categorical Relationships:
   - BusinessTravel
   - Department
   - EducationField
   - JobRole
   - MaritalStatus
   - OverTime
   - EnvironmentSatisfaction
   - JobInvolvement
   - JobLevel
   - JobSatisfaction
   - StockOptionLevel
   - TrainingTimesLastYear
   - WorkLifeBalance

Now we can create count plots for each of these categorical attributes, and how they affect attrition.

Majority of employees lost in attrition rarely travel.

```
display_categorical_x_categorical_analysis(data,next(i))
```

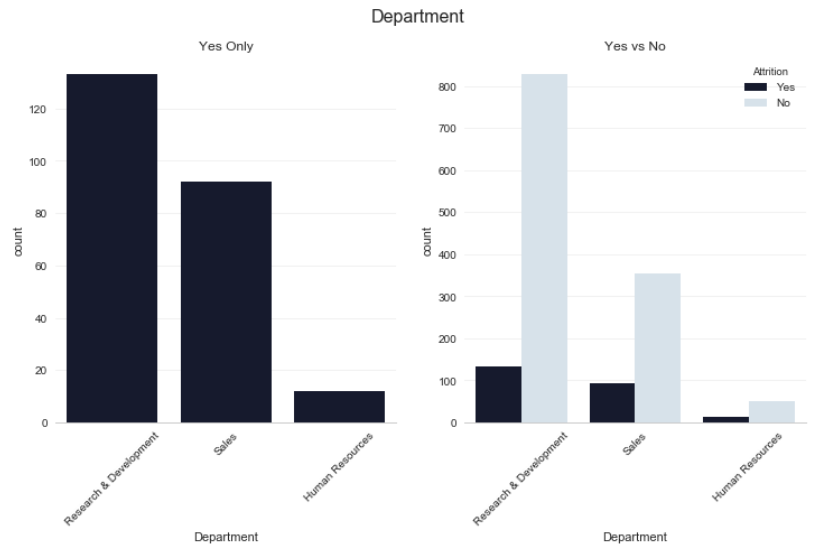



```
Yes Only
Total Count: 237
```

	Percent	Cumulative Percent
Travel_Rarely	65.8%	65.8%
Travel_Frequently	29.1%	94.9%
Non-Travel	5.1%	100.0%

Research & Development and Sales department contribute ~95% of the employees lost in attrition

```
display_categorical_x_categorical_analysis(data,next(i))
```

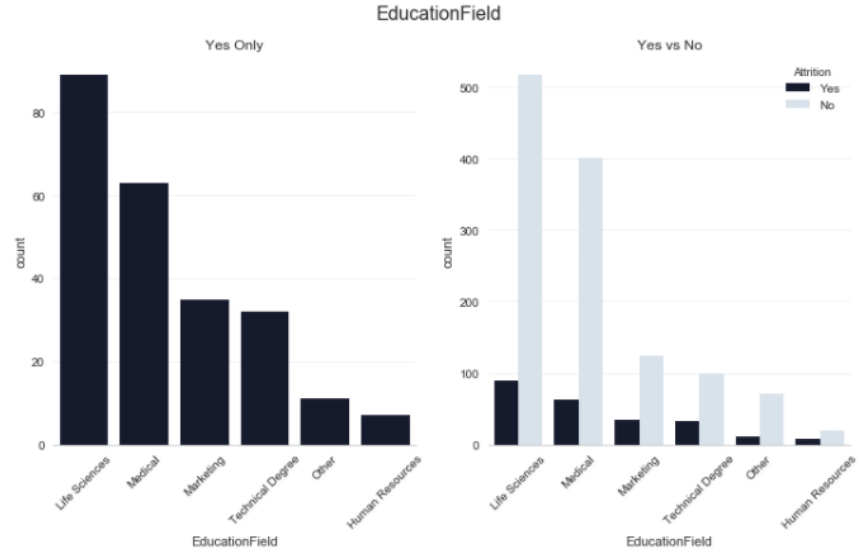



```
Yes Only
Total Count: 237
```

	Percent	Cumulative Percent
Research & Development	56.1%	56.1%
Sales	38.8%	94.9%
Human Resources	5.1%	100.0%

Employees educated in life sciences or medical together make up ~64% of the attrition sample. ¶

```
display_categorical_x_categorical_analysis(data,next(i))
```

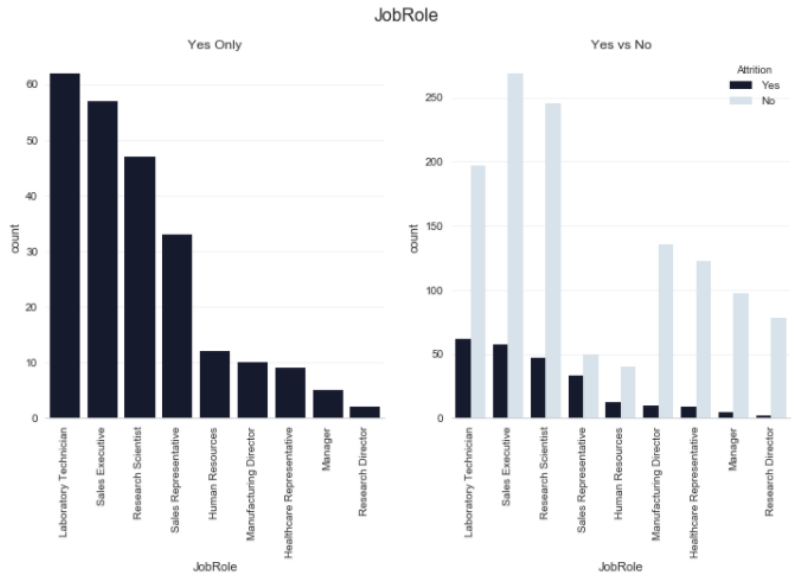



```
Yes Only
Total Count: 237
```

	Percent	Cumulative Percent
Life Sciences	37.6%	37.6%
Medical	26.6%	64.1%
Marketing	14.8%	78.9%
Technical Degree	13.5%	92.4%
Other	4.6%	97.0%
Human Resources	3.0%	100.0%

70% of attrition sample is made up from laboratory technicians, sale executives, and research scientists.

```
display_categorical_x_categorical_analysis(data,next(i))
```

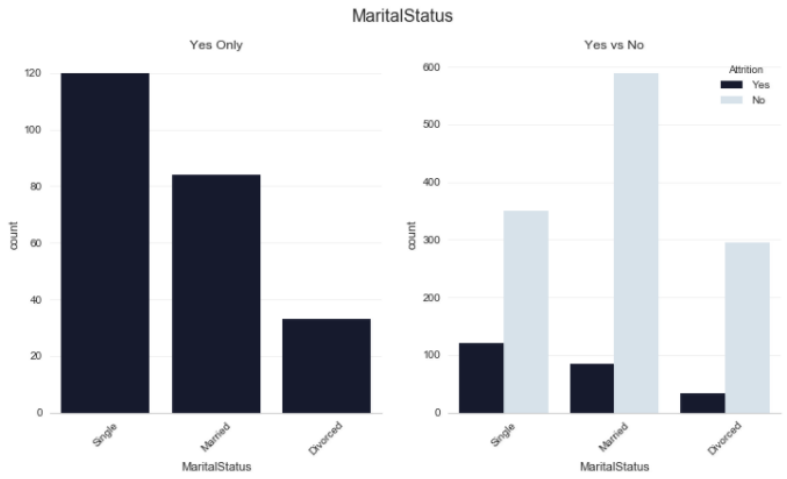



```
Yes Only
Total Count: 237
```

	Percent	Cumulative Percent
Laboratory Technician	26.2%	26.2%
Sales Executive	24.1%	50.2%
Research Scientist	19.8%	70.0%
Sales Representative	13.9%	84.0%
Human Resources	5.1%	89.0%

Employees with the relationship status as single make up over 50% attrition sample.

```
display_categorical_x_categorical_analysis(data,next(i))
```

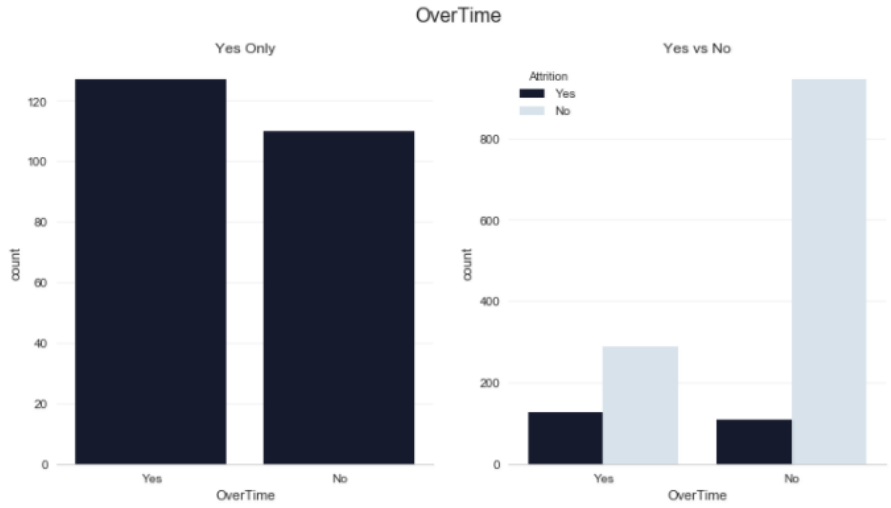



```
Yes Only
Total Count: 237
```

	Percent	Cumulative Percent
Single	50.6%	50.6%
Married	35.4%	86.1%
Divorced	13.9%	100.0%

The ratio of employees working overtime is drastically different across Yes vs No samples.

```
display_categorical_x_categorical_analysis(data,next(i))
```

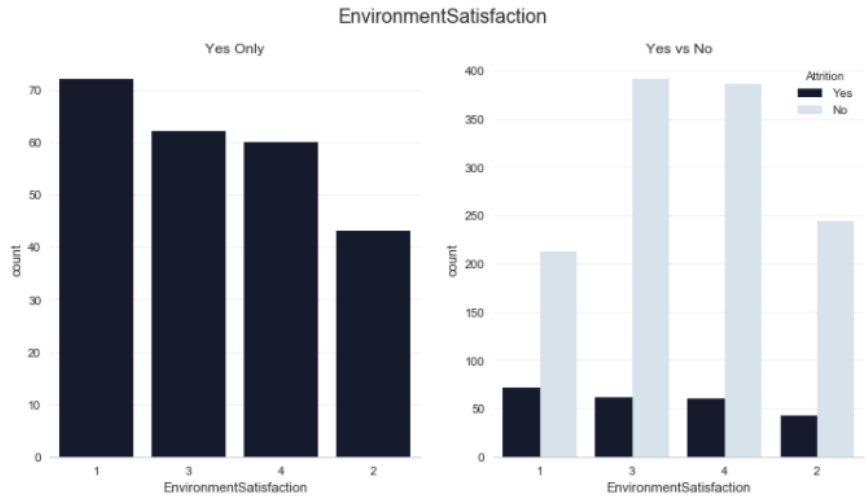



```
Yes Only
Total Count: 237
```

	Percent	Cumulative Percent
Yes	53.6%	53.6%
No	46.4%	100.0%

The distribution of environment satisfaction ratings are different for employees lost in attrition vs those who stayed.

```
display_categorical_x_categorical_analysis(data,next(i))
```

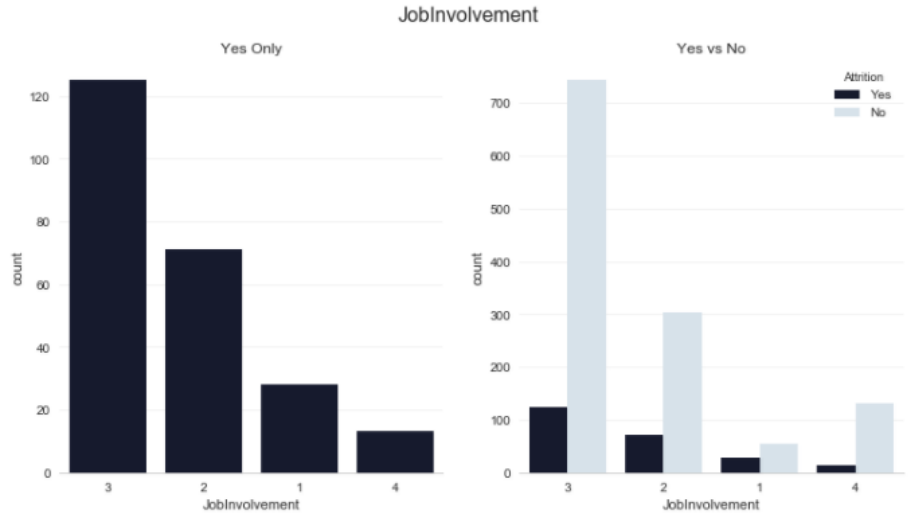



```
Yes Only
Total Count: 237
```

	Percent	Cumulative Percent
1	30.4%	30.4%
3	26.2%	56.5%
4	25.3%	81.9%
2	18.1%	100.0%

~80% of employees lost in attrition rated their level job involvement as moderate to moderately high.

```
display_categorical_x_categorical_analysis(data,next(i))
```

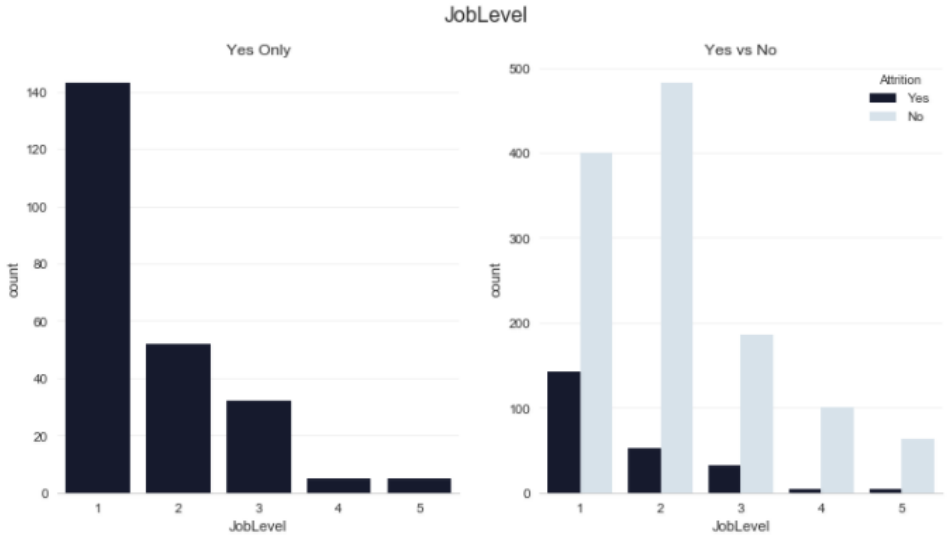



```
Yes Only
Total Count: 237
```

	Percent	Cumulative Percent
3	52.7%	52.7%
2	30.0%	82.7%
1	11.8%	94.5%
4	5.5%	100.0%

Entry level employees make up 60% of the attrition sample

```
display_categorical_x_categorical_analysis(data,next(i))
```

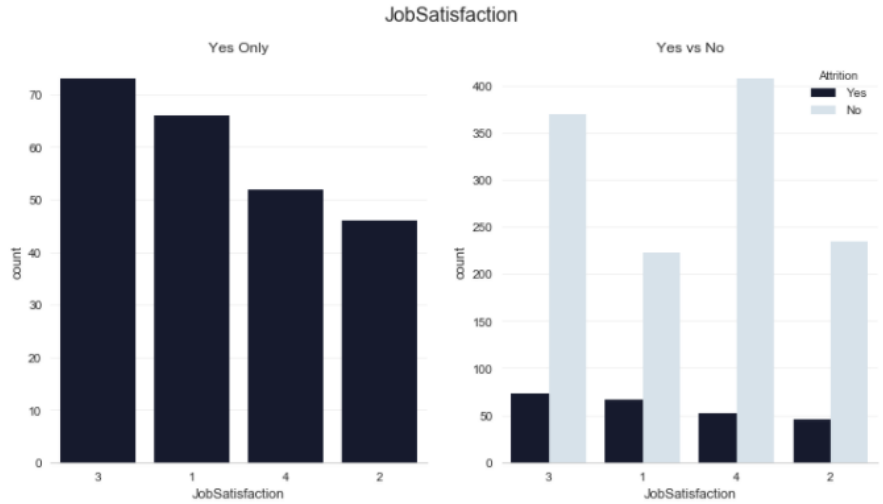



```
Yes Only
Total Count: 237
```

	Percent	Cumulative Percent
1	60.3%	60.3%
2	21.9%	82.3%
3	13.5%	95.8%
4	2.1%	97.9%
5	2.1%	100.0%

Employees who stay had job satisfactions ratings ratio more positively biased than those lost in attrition. ¶

```
display_categorical_x_categorical_analysis(data,next(i))
```

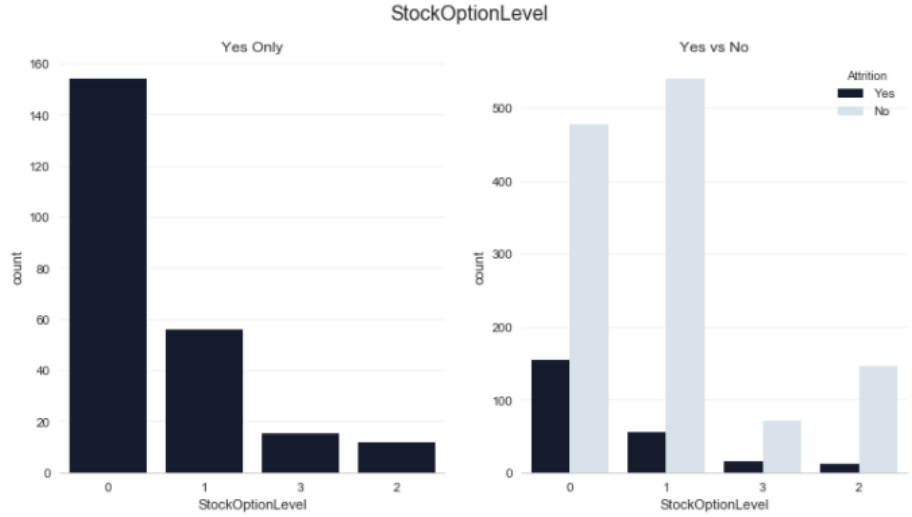



```
Yes Only
Total Count: 237
```

	Percent	Cumulative Percent
3	30.8%	30.8%
1	27.8%	58.6%
4	21.9%	80.6%
2	19.4%	100.0%

Employees lost in attrition had 65% of its members possessing stock option level at 0.

```
display_categorical_x_categorical_analysis(data,next(i))
```

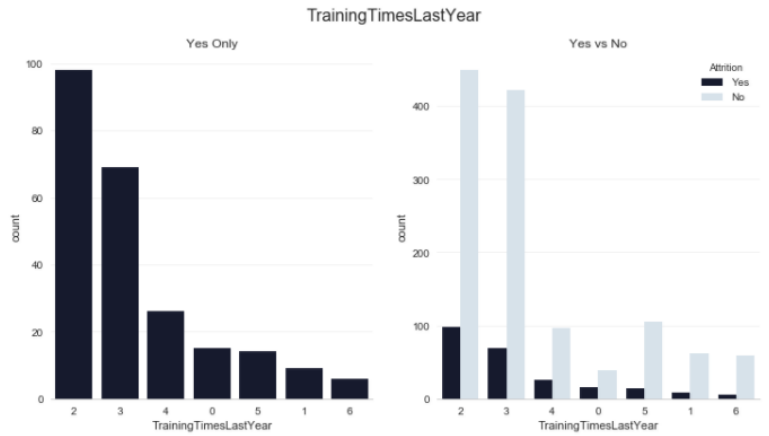



```
Yes Only
Total Count: 237
```

	Percent	Cumulative Percent
0	65.0%	65.0%
1	23.6%	88.6%
3	6.3%	94.9%
2	5.1%	100.0%

Employees lost in attrition had ~70% of members trained 2 - 3 times a year.

```
display_categorical_x_categorical_analysis(data,next(i))
```



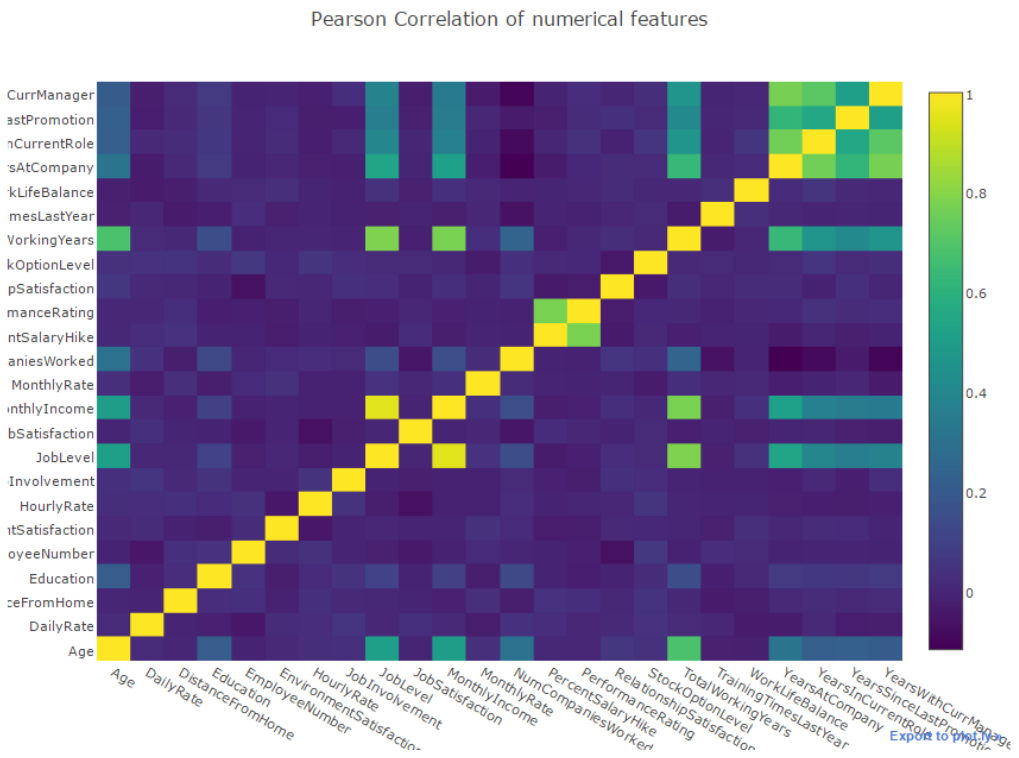

```
Yes Only
Total Count: 237
```

	Percent	Cumulative Percent
2	41.4%	41.4%
3	29.1%	70.5%
4	11.0%	81.4%
0	6.3%	87.8%
5	5.9%	93.7%
1	3.8%	97.5%
6	2.5%	100.0%

Note: refer HR_EDA_category.ipynb

Correlation of features:

The next tool in a data explorer's arsenal is that of a correlation matrix. By plotting a correlation matrix, we have a very nice overview of how the features are related to one another.



From the correlation plot, we can see that quite a lot of our columns seem to be poorly correlated with one another. Generally, when making a predictive model, it would be preferable to train a model with features that are not too correlated with one another so that we do not need to deal with redundant features. In this case, we have quite a lot of correlated features, perhaps a technique such as Principal Component Analysis can be applied to reduce the feature space.

Dashboards:

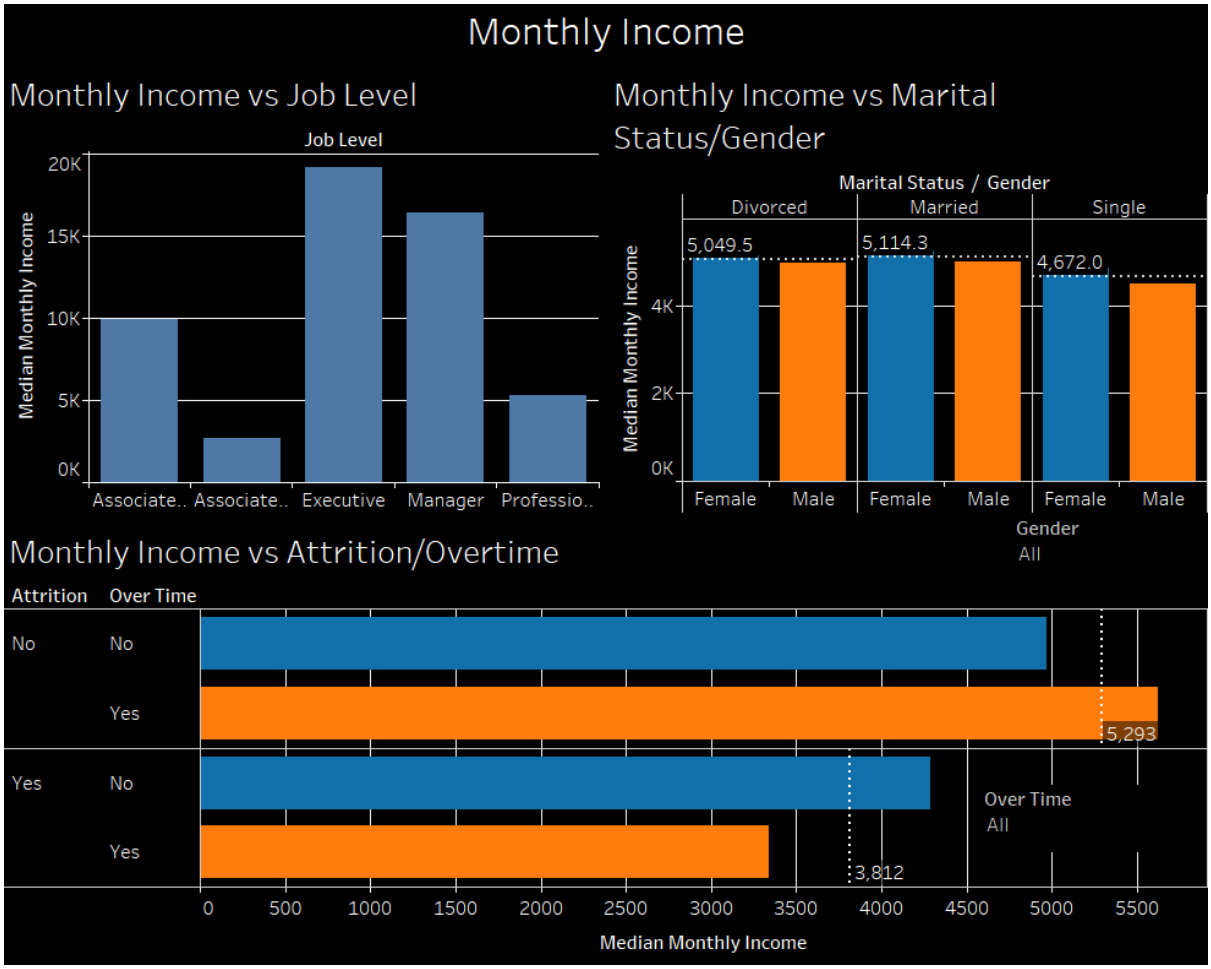
Methodology: First, information gain using ranker's algorithm was implemented on the dataset by taking the target variable as Attrition. A list of ranked attributes was obtained:

```
Attribute Evaluator (supervised, Class (nominal): 2 Attrition):
        Information Gain Ranking Filter

Ranked attributes:
 0.61307458   19 MonthlyRate
 0.58126854   18 MonthlyIncome
 0.41148934    4 DailyRate
 0.05642817    1 Age
 0.05416659   28 TotalWorkingYears
 0.04470628   31 YearsAtCompany
 0.04362863   15 JobRole
 0.03994478   22 OverTime
 0.03864482   12 HourlyRate
 0.0363979    34 YearsWithCurrManager
 0.03603855   14 JobLevel
 0.03403475   32 YearsInCurrentRole
 0.03027354   27 StockOptionLevel
 0.02159121   17 MaritalStatus
 0.01618321    6 DistanceFromHome
 0.0138586    33 YearsSinceLastPromotion
 0.01264528   20 NumCompaniesWorked
 0.01259768   13 JobInvolvement
 0.01165945    3 BusinessTravel
 0.01009658   10 EnvironmentSatisfaction
 0.00851704   16 JobSatisfaction
 0.00731169    8 EducationField
 0.00723988   29 TrainingTimesLastYear
 0.00690596   30 WorkLifeBalance
 0.00669298   23 PercentSalaryHike
 0.00514774    5 Department
 0.00244686   25 RelationshipSatisfaction
 0.00157028    7 Education
 0.00063086   11 Gender
 0.000006     24 PerformanceRating
 0            21 Over18
 0            26 StandardHours
 0             9 EmployeeCount
```

Using the attributes that have the most effect on attrition, dashboards were created. To compare the attributes with one another, the variable which was selected for the dashboard was then evaluated in the correlation matrix and the attributes having maximum correlation with the selected variable were plotted in the dashboard.

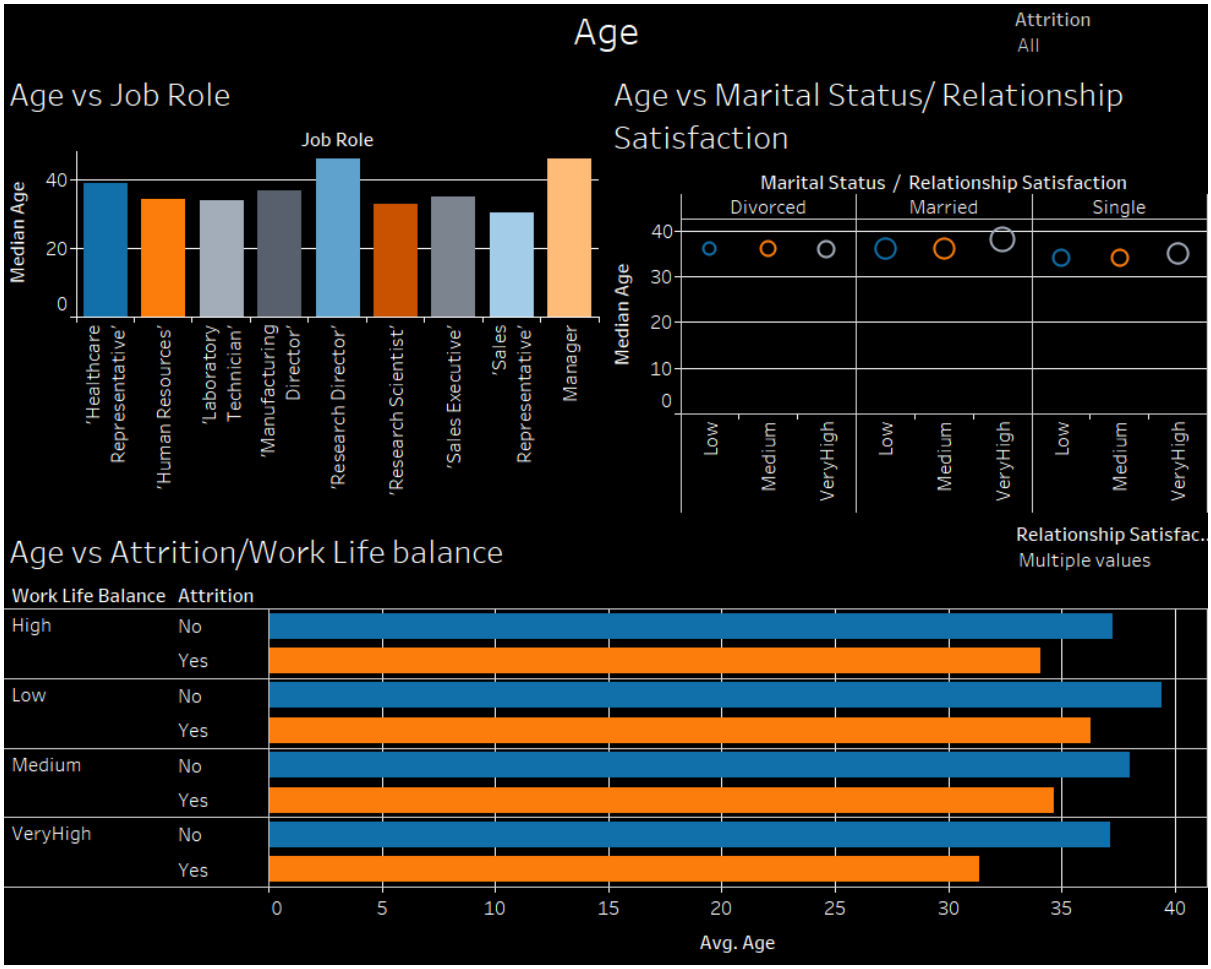
1. **Monthly Income**



Description: dashboard provides information about Monthly Income with various other attributes.

Insights: Monthly income is dependent upon marital status and gender. A female earns more than a male. Also the average monthly income of Married people is highest. Attrition happens when an employee is doing overtime and still receiving monthly income less than the average.

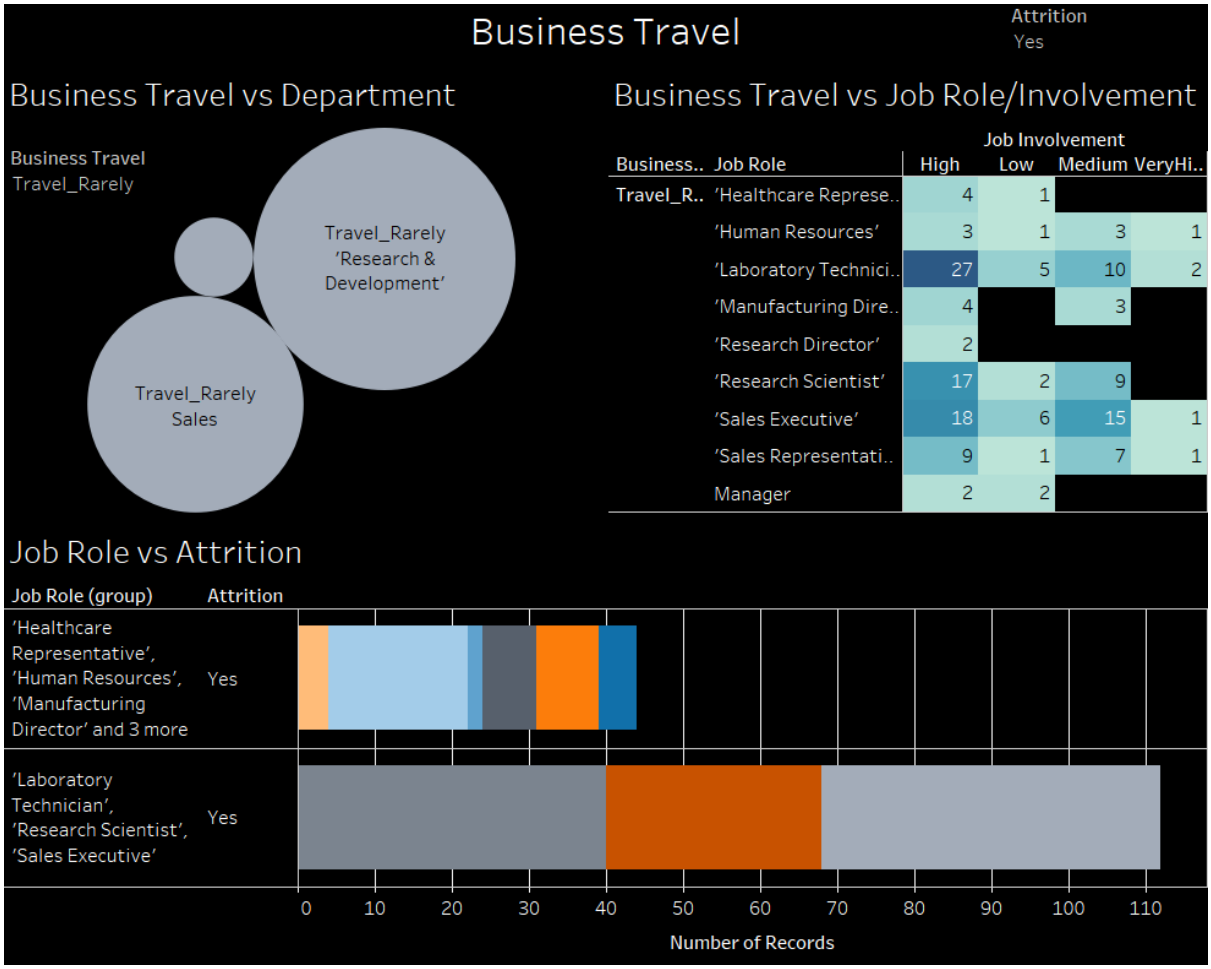
2. Age



Description: dashboard provides information and distribution of age in the organisation and how it is dependent on other variables.

Insights: Employees working as research director and managers have the highest ages. A person having age around 30, even if the work life balance of that employee is very high, there are high chances that the person might want to leave. The reasons can be looking for a new job or for a new opportunity.

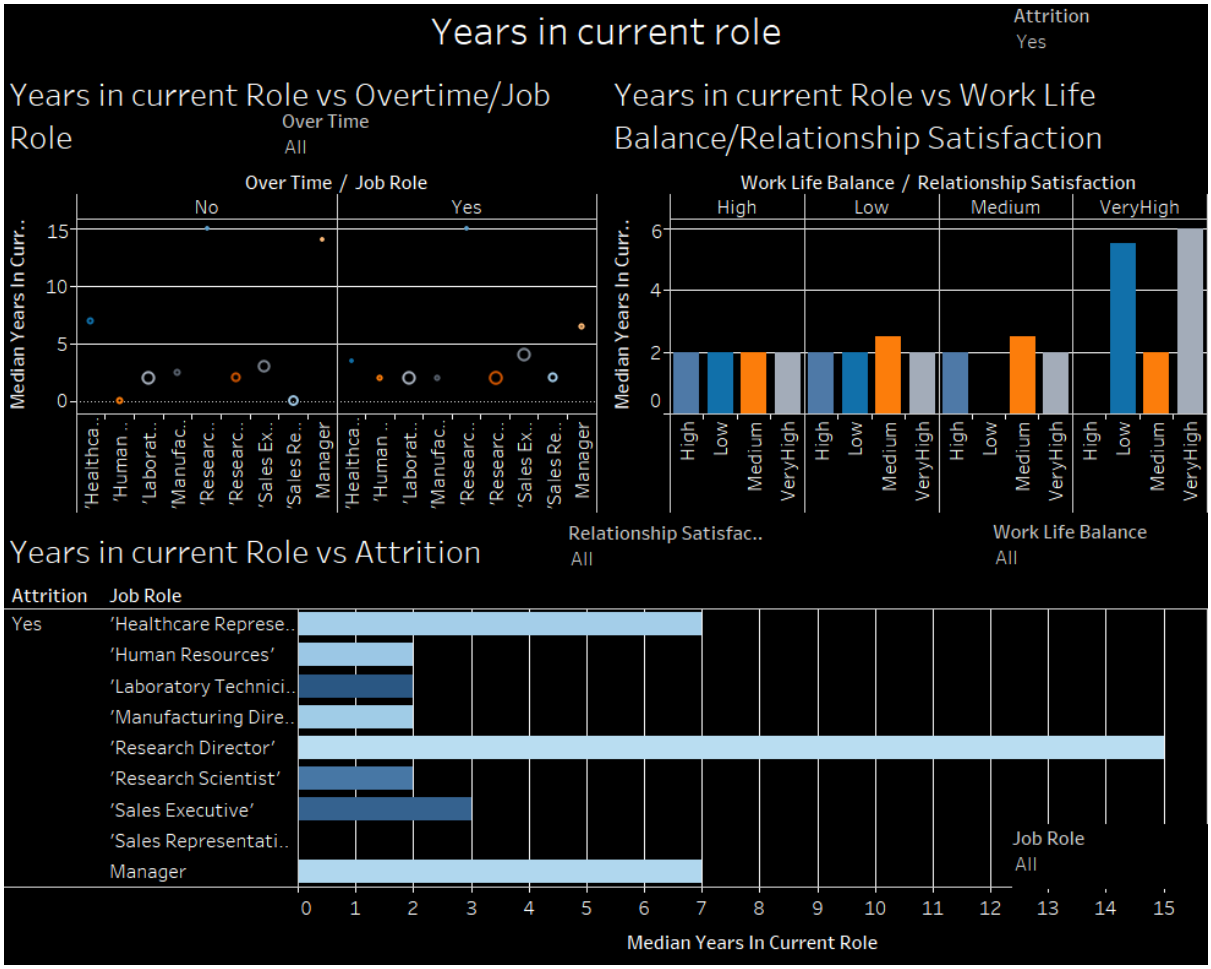
3. Business Travel



Description: Employees in Sales and Research & development travel rarely.

Insights: Going further, employees working in above mentioned departments as Lab technicians, research scientists or sales executive, together are responsible for more than 70% of attrition itself, just because they travel rarely for business trips.

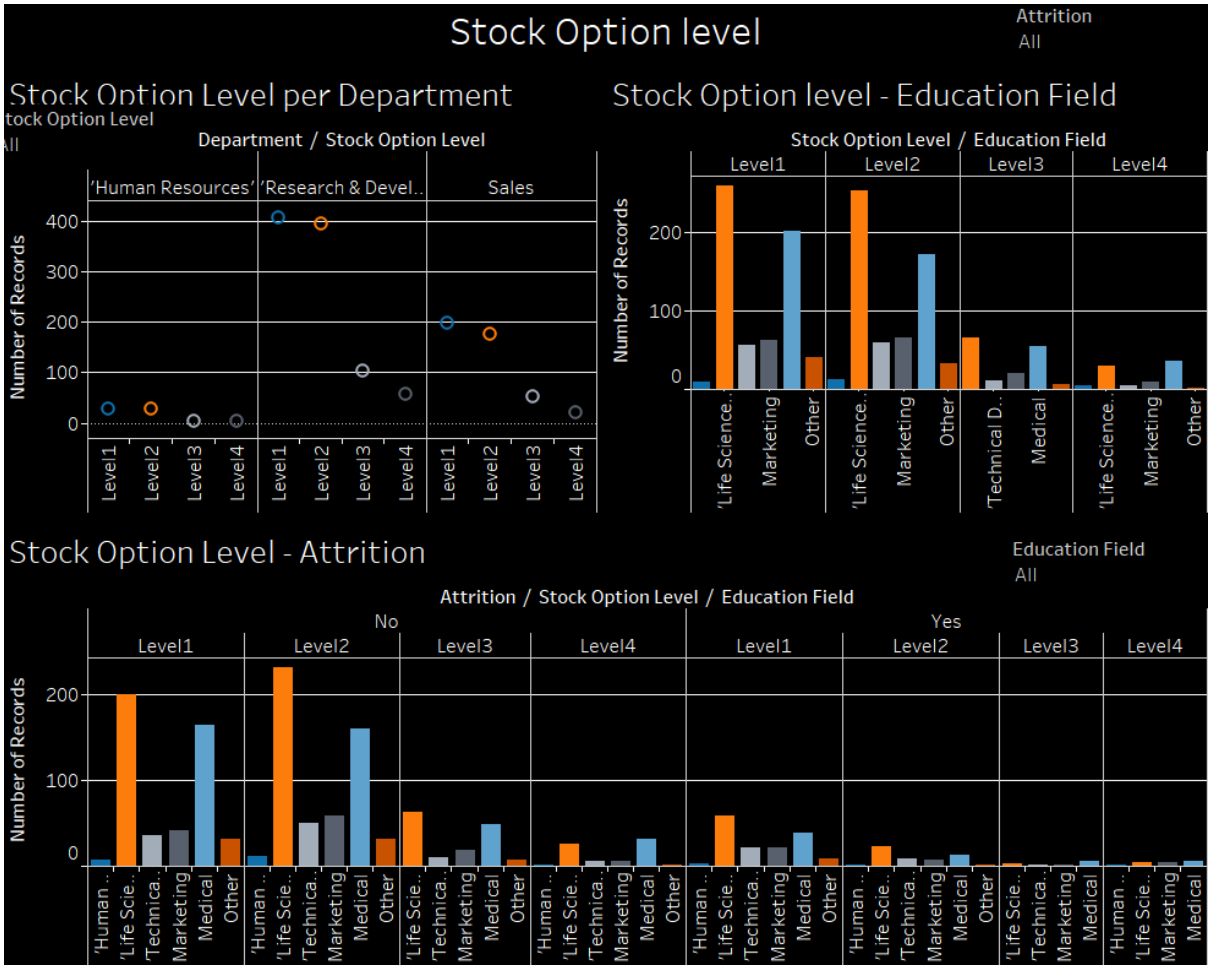
4. **Years in current role**



Description: dashboard provides information about employees and the number of years they have been employed in the same role.

Insights: Most of the attrition is happening from the role of Research Directors because they do most of the overtime in the organisation but there relationship satisfaction is low. That's why they leave.

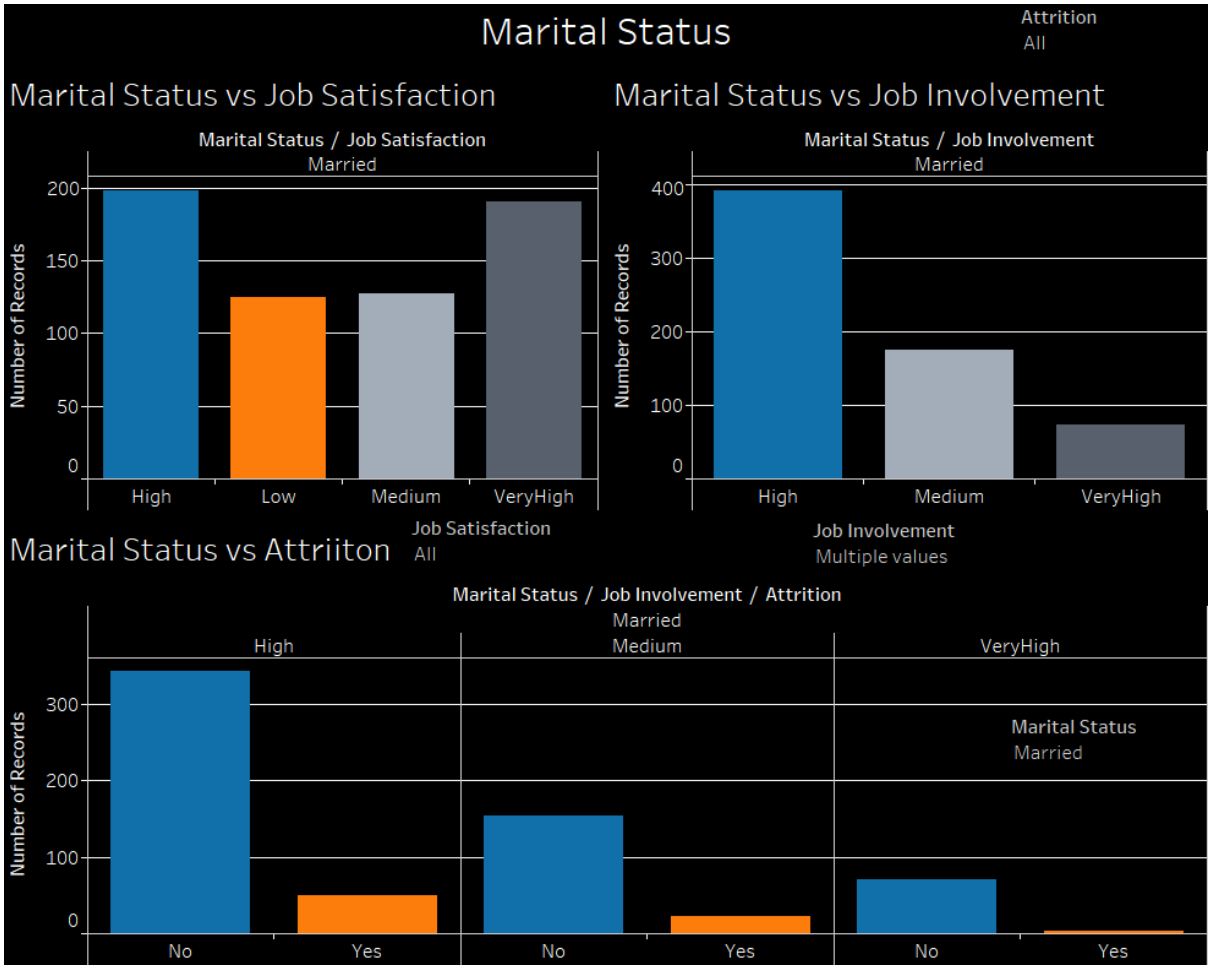
5. **Stock option level**



Description: dashboard provides information about how stock option level can result in attrition.

Insights: An employee can do stock trading on 4 levels within a company. These 4 levels are mentioned here. Employees from Research and Development and Sales do a lot of stock trading than employees in HR. Employees belonging from the field of Life Sciences education tend to do most of the stock trading. Attrition can happen if an employee wants to do trading and is not given stock options as can be seen from the last plot.

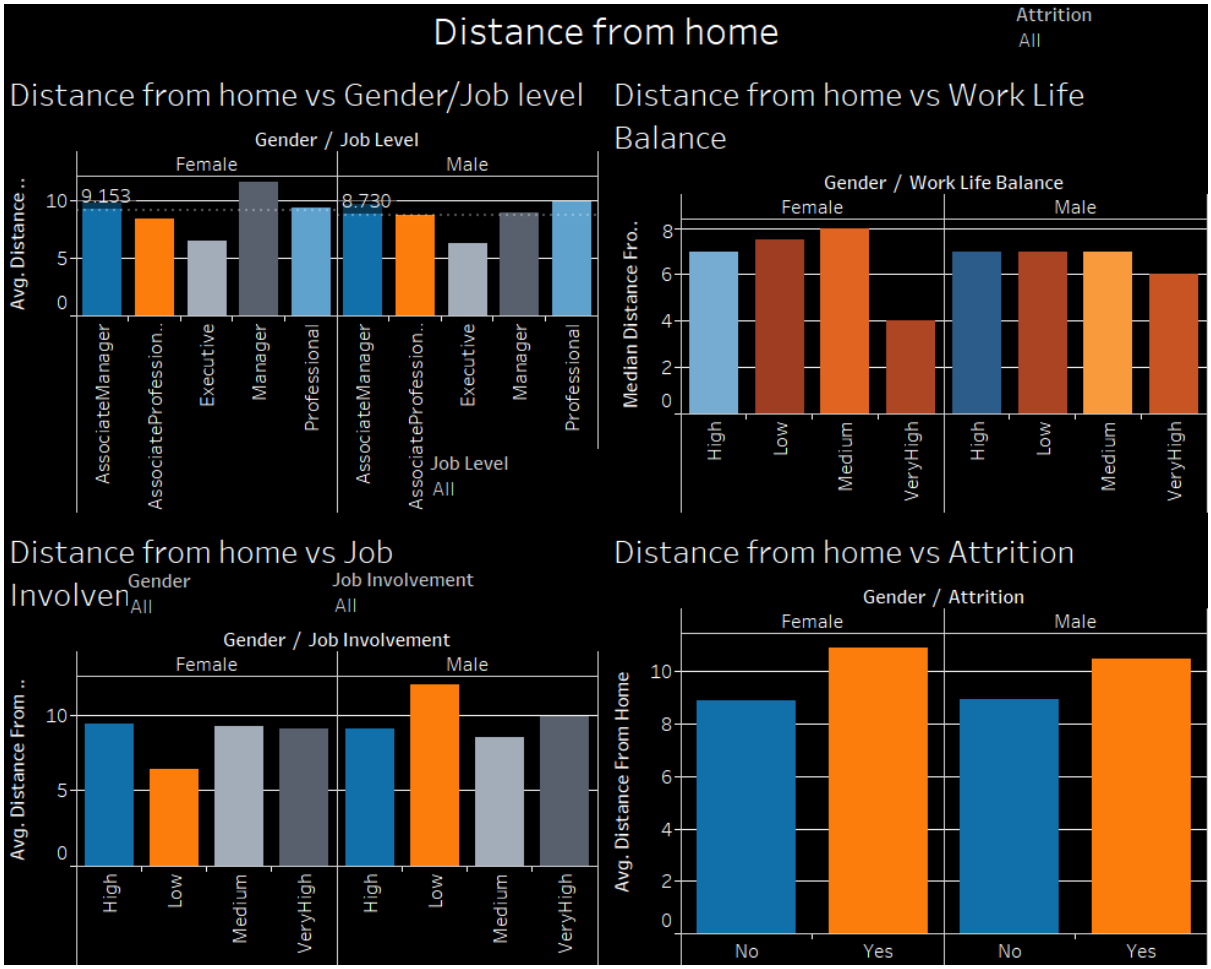
6. **Marital status**



Description: dashboard provides information about marital status and how it affects job satisfaction and job involvement.

Insights: if an employee is married, they are very much satisfied by their jobs. This could also be a result of responsibilities that comes with marriage and therefore the employee has to believe that they are satisfied for securing a good future for the family. The job involvement is also high for married employees. But there are certain employees, who have high job involvement but still are part of the attrition. This could be because of new opportunities somewhere else.

7. **Distance from home**



Description: This dashboard provides information on how an employee's distance from work affects attrition.

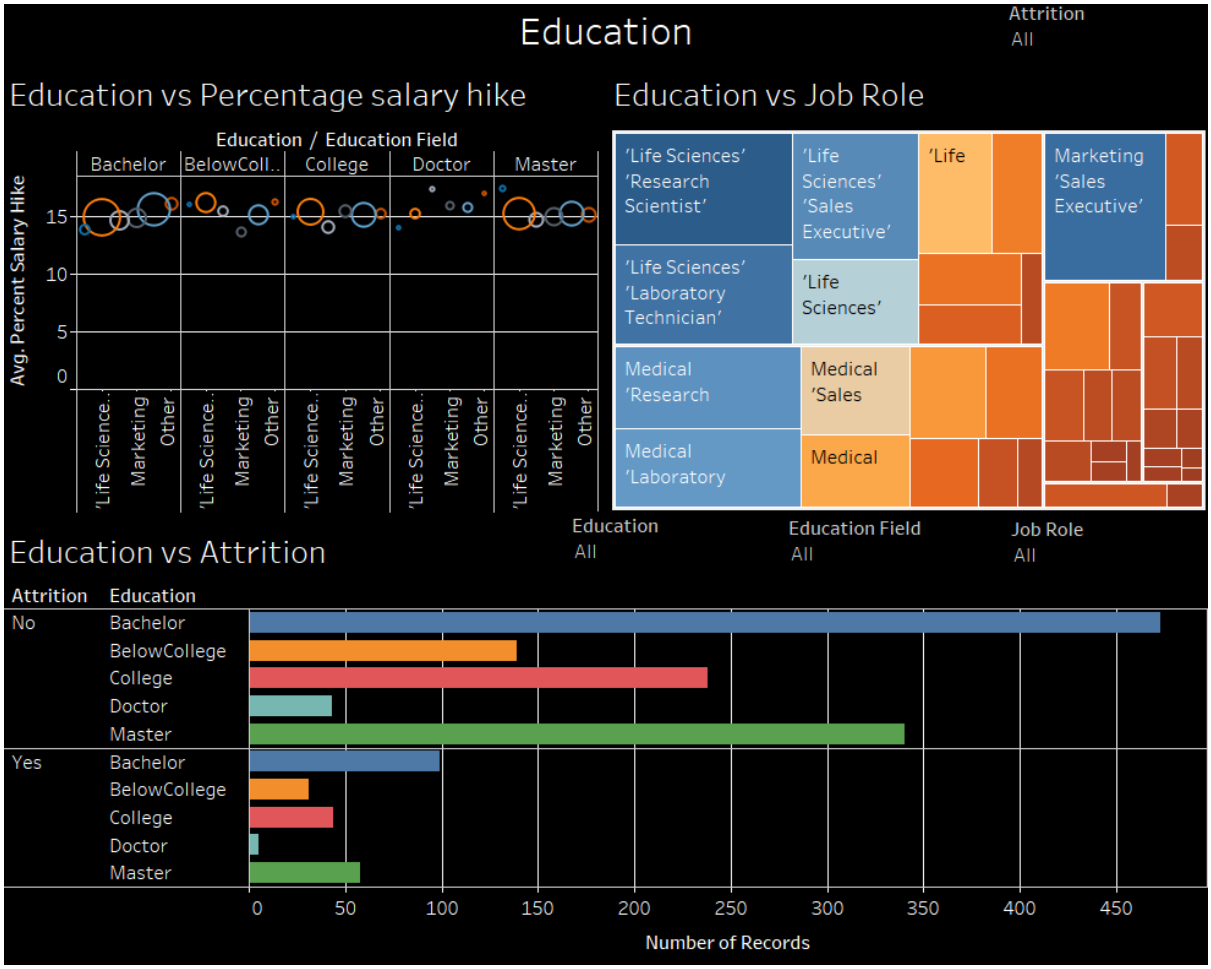
Insights: According to the data, male employees live somewhat closer to their working places as compared to female employees. Executives live the closest to the working place as compared to other Job levels. Male employees have a higher levels of work life balance than female employees. This could be possible due to extra responsibilities on females due to social constructs. Employees living near the working place have higher job involvement and employees living further will be the part of the attrition because they will search for new opportunities which will be closer to their homes.

8. Satisfaction

Description: the dashboard describes satisfaction which includes environment in the work place satisfaction and relationship satisfaction.

Insights: Environment satisfaction and relationship satisfaction are somewhat inversely correlated because if an employee is satisfied with the working environment they are not satisfied with the relationship and vice versa. Satisfaction levels in the working environment either drastically increase or decrease when the number of years in the company increases for employees. Employees who are not satisfied in the relationship tend to leave the company.

9. Education



Description: this dashboard gives information about education levels of employees with respect to education fields and job roles of these employees in the company.

Insights: People from life sciences generally join the research and development team. Also people having marketing background tend to become sales executives. Attrition is maximum seen when the employee just has a bachelor's degree and the least when they have a doctorate. But the company should give attention to why the employee with a doctorate degree want s to leave, because they are invaluable assets to the company.

10. Attrition

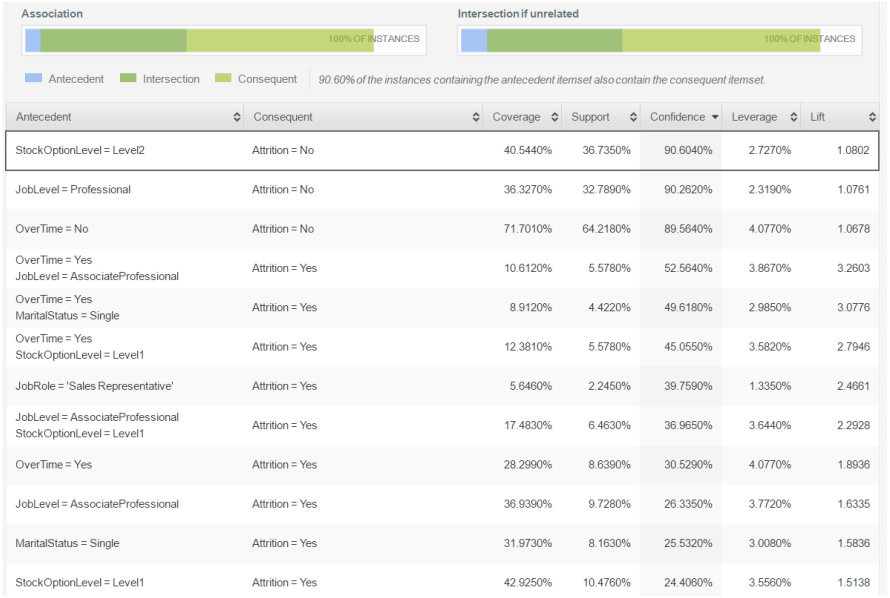
Description: the final dashboard is about attrition itself and tells about how the other major attributes from the ranked list affect attrition.

Insights: employees will leave the organisation when the average monthly income is less than 5k. Also female employees have higher average monthly income than male employees. Most of the attrition is observed in Sales department and finally even if the performance rating of employees is outstanding they can leave the company due to other reasons.

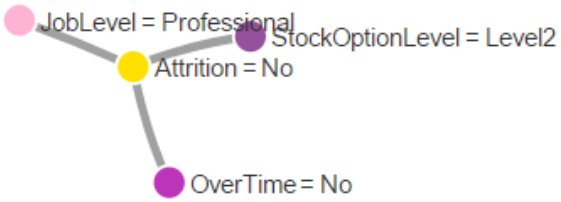
Questions from the first part:

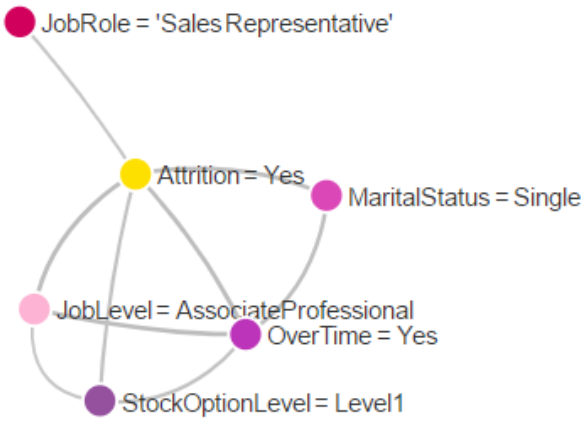
1. Is overtime inspiring an employee to leave the company?
2. Does higher job level results in more overtime?
3. Is overtime killing a marriage?
4. Do stock option levels has to do anything with attrition?
5. Are more stocks given to employees when they reach higher job levels?

Part 2: Identifying three sets of relationships (using Association rules).



Antecedent	Consequent	Coverage	Support	Confidence	Leverage	Lift
StockOptionLevel = Level2	Attrition = No	40.5440%	36.7350%	90.6040%	2.7270%	1.0802
JobLevel = Professional	Attrition = No	36.3270%	32.7890%	90.2620%	2.3190%	1.0761
OverTime = No	Attrition = No	71.7010%	64.2180%	89.5640%	4.0770%	1.0678
OverTime = Yes JobLevel = AssociateProfessional	Attrition = Yes	10.6120%	5.5780%	52.5640%	3.8670%	3.2603
OverTime = Yes MaritalStatus = Single	Attrition = Yes	8.9120%	4.4220%	49.6180%	2.9850%	3.0776
OverTime = Yes StockOptionLevel = Level1	Attrition = Yes	12.3810%	5.5780%	45.0550%	3.5820%	2.7946
JobRole = 'Sales Representative'	Attrition = Yes	5.6460%	2.2450%	39.7590%	1.3350%	2.4661
JobLevel = AssociateProfessional StockOptionLevel = Level1	Attrition = Yes	17.4830%	6.4630%	36.9650%	3.6440%	2.2928
OverTime = Yes	Attrition = Yes	28.2990%	8.6390%	30.5290%	4.0770%	1.8936
JobLevel = AssociateProfessional	Attrition = Yes	36.9390%	9.7280%	26.3350%	3.7720%	1.6335
MaritalStatus = Single	Attrition = Yes	31.9730%	8.1630%	25.5320%	3.0080%	1.5836
StockOptionLevel = Level1	Attrition = Yes	42.9250%	10.4760%	24.4060%	3.5560%	1.5138



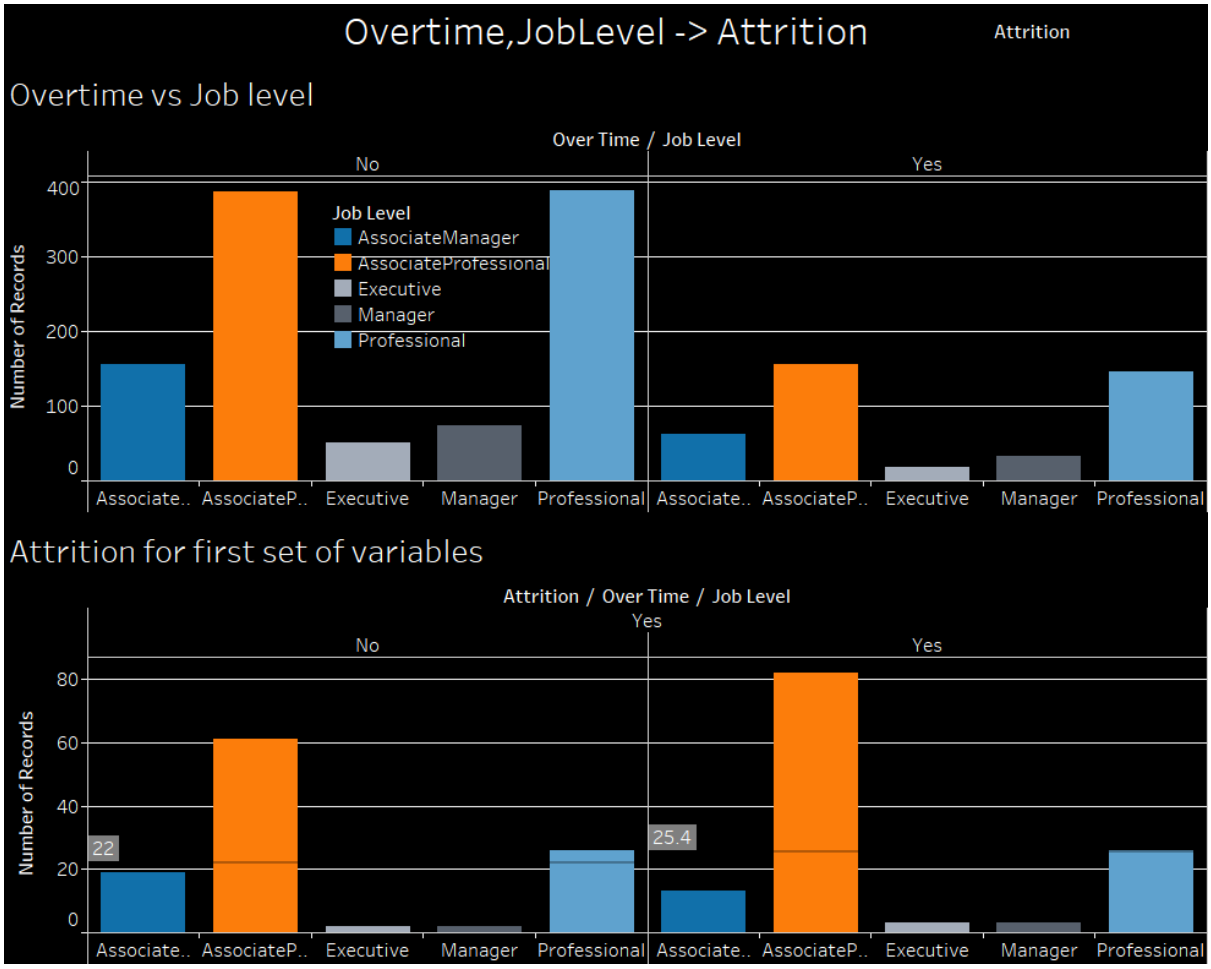


So, the three sets of variables are:

1. Overtime, Job level -> Attrition
2. Overtime, Marital Status -> Attrition
3. Job level, stock option level -> Attrition

Dashboards:

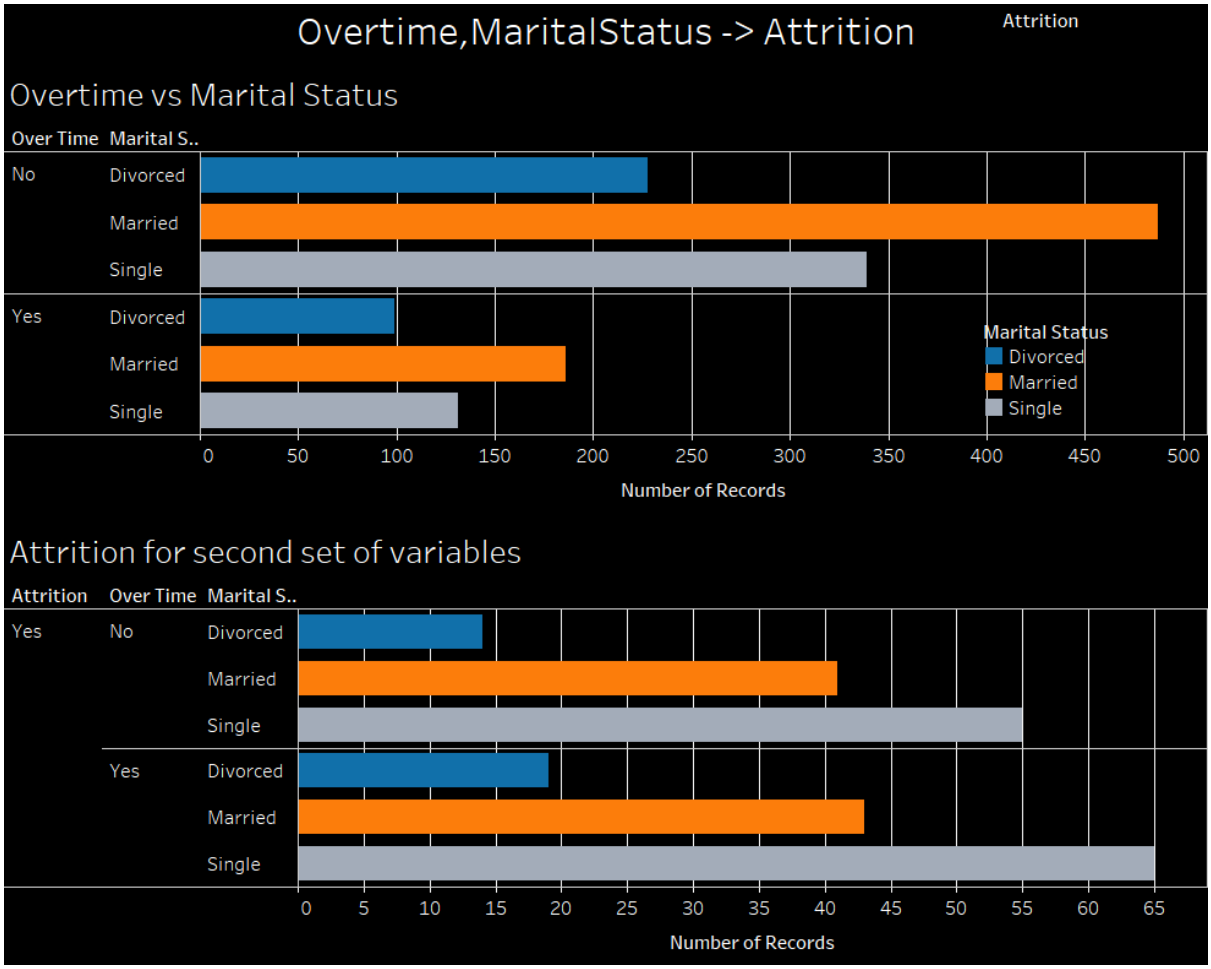
1. Overtime, Job level -> Attrition



Description: the effect of overtime and job level on attrition is shown in this dashboard.

Insights: Most of the overtime is done by employees who are either at Associate Professionals or Professionals level. Also, most of the attrition also happens from these two job levels only. This can mean the manager is not good and is giving more work than intended to the professional who in turn gives it to the associate professional, thus associate professionals are the ones that leave the company in most number followed by professionals.

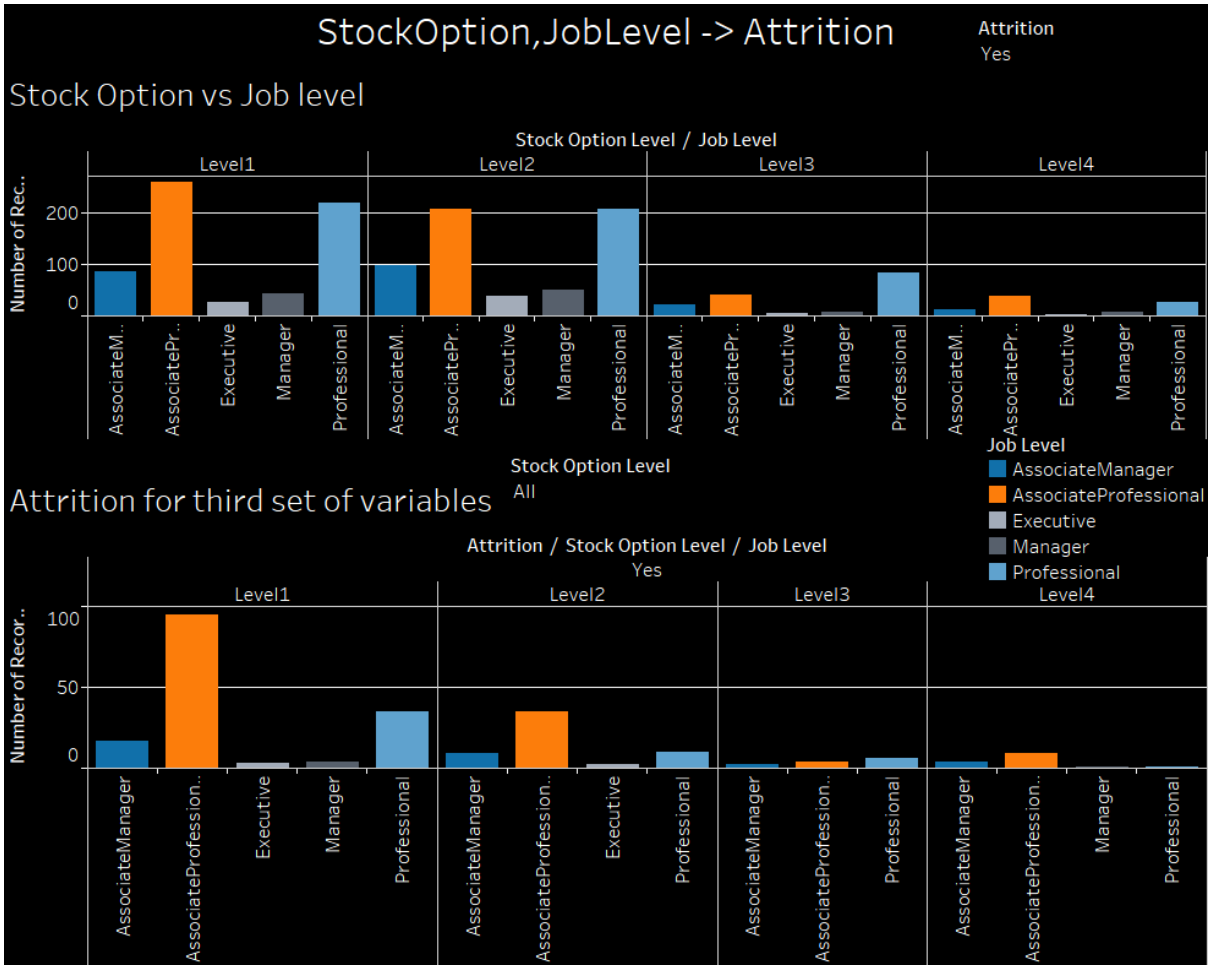
2. Overtime, Marital Status -> Attrition



Description: the effects of overtime and marital status on attrition is shown in this dashboard.

Insights: due to some reasons, one of which can be extra money, married people do the most overtime. But the employees that share the most percentage in attrition are Singles who do overtime but are not paid accordingly that's why leave.

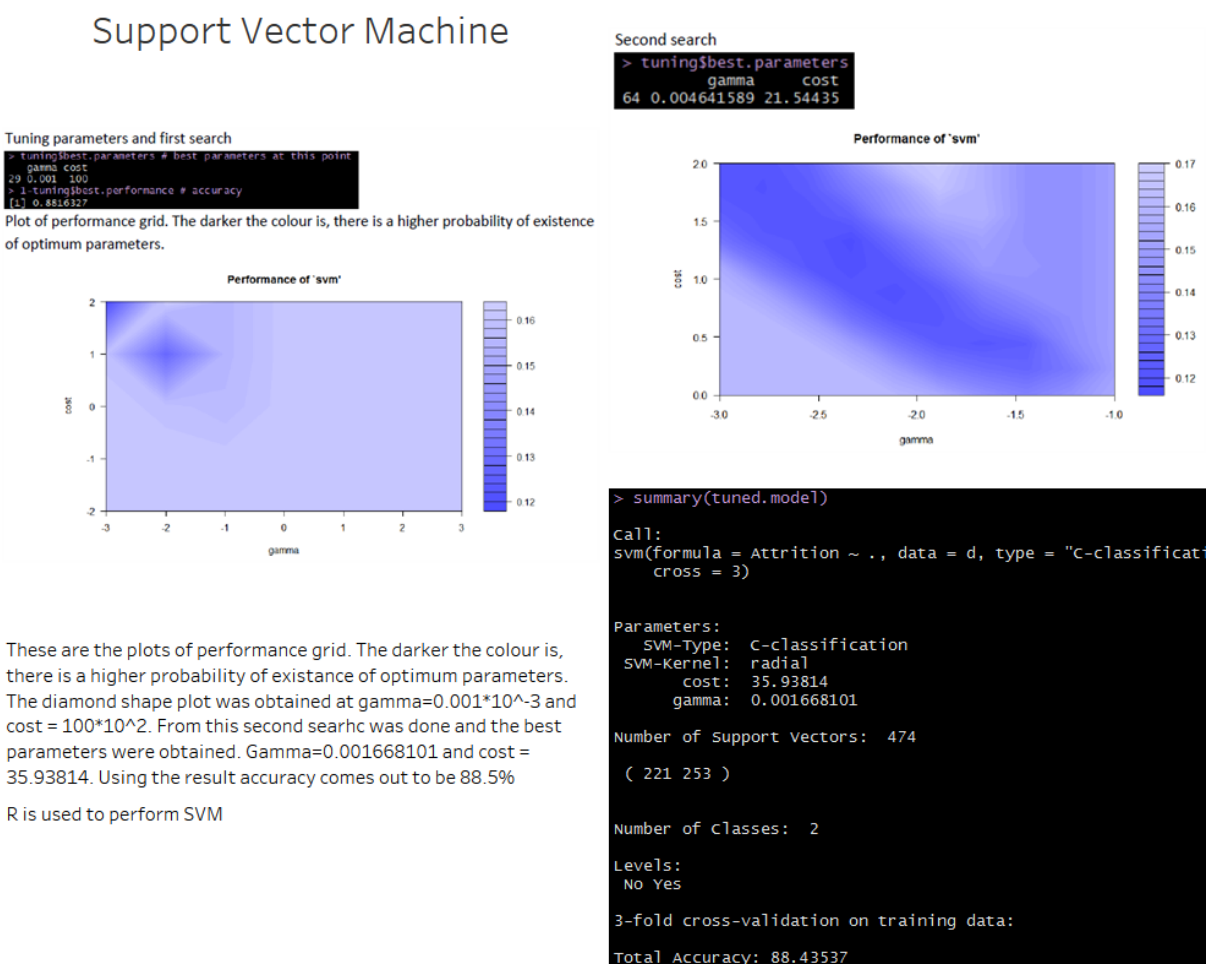
3. Job level, stock option level -> Attrition



Description: the effects of job level and stock option level on attrition is shown in this dashboard.

Insights: Employees who are Associate Professionals work mostly on level 1 and level 2 trading followed by Professionals. Also attrition is seen in these two job levels the most.

Predictive Analysis:



Support Vector Machine was implemented on the dataset using R and the results are being presented on a dashboard that summarises about the prediction model generated by SVM.

Data cleaning included removal of Employee Count, over 18 and Standard Hours attribute because they only contained one value or factor.

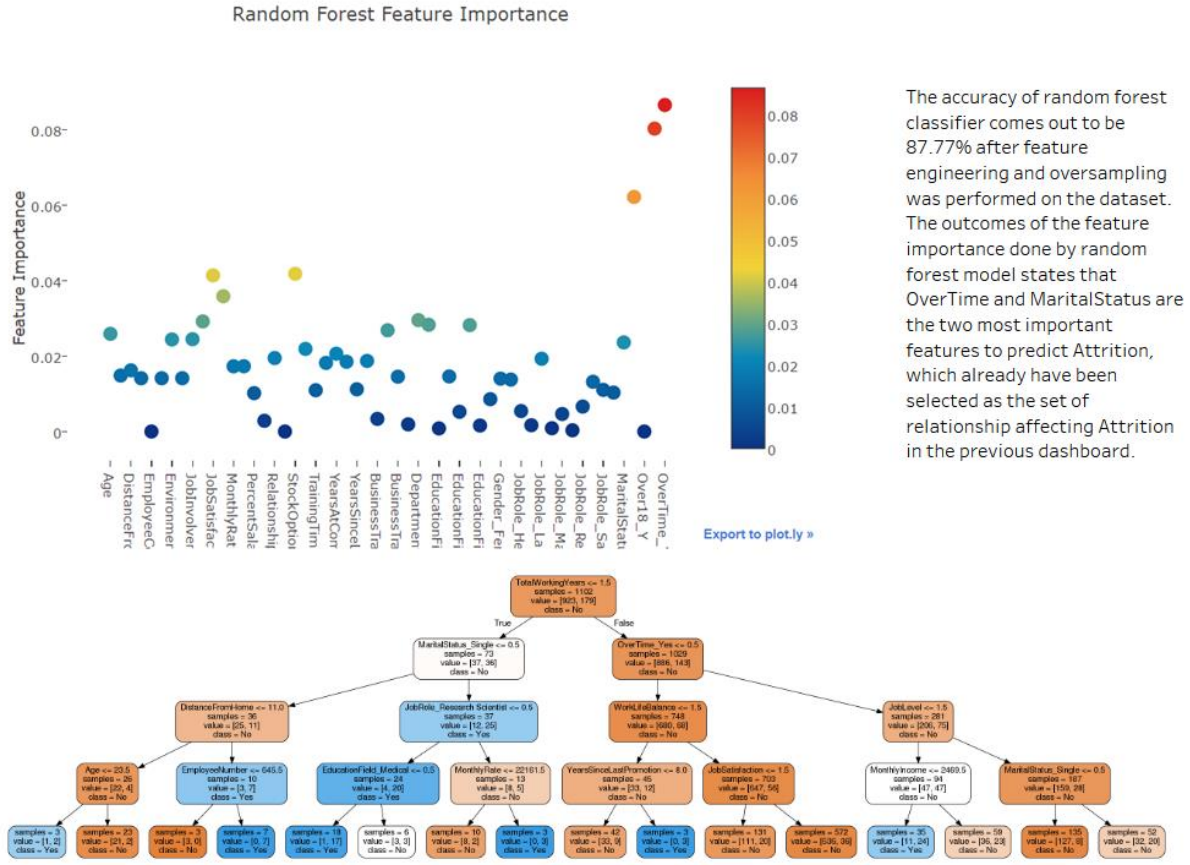
Then the data was split as training and testing data and prediction table was made for training data using 3-fold cross validation.

The tuning parameters obtained after first search were gamma = 0.001*10^-3 and cost = 100*10^2. A plot of performance grid was created. The darker the colour in the performance grid, there is a higher probability of existence of optimum parameters.

The upper left corner was not searched because as cost parameter hikes, it results in a better one-shot prediction but usually this is overfitting.

In second search the best parameters were found to be gamma = 0.001668101 and cost = 35.93814. The accuracy came out to be 88.43%.

Random Forest



The accuracy of random forest classifier comes out to be 87.77% after feature engineering and oversampling was performed on the dataset. The outcomes of the feature importance done by random forest model states that OverTime and MaritalStatus are the two most important features to predict Attrition, which already have been selected as the set of relationship affecting Attrition in the previous dashboard.

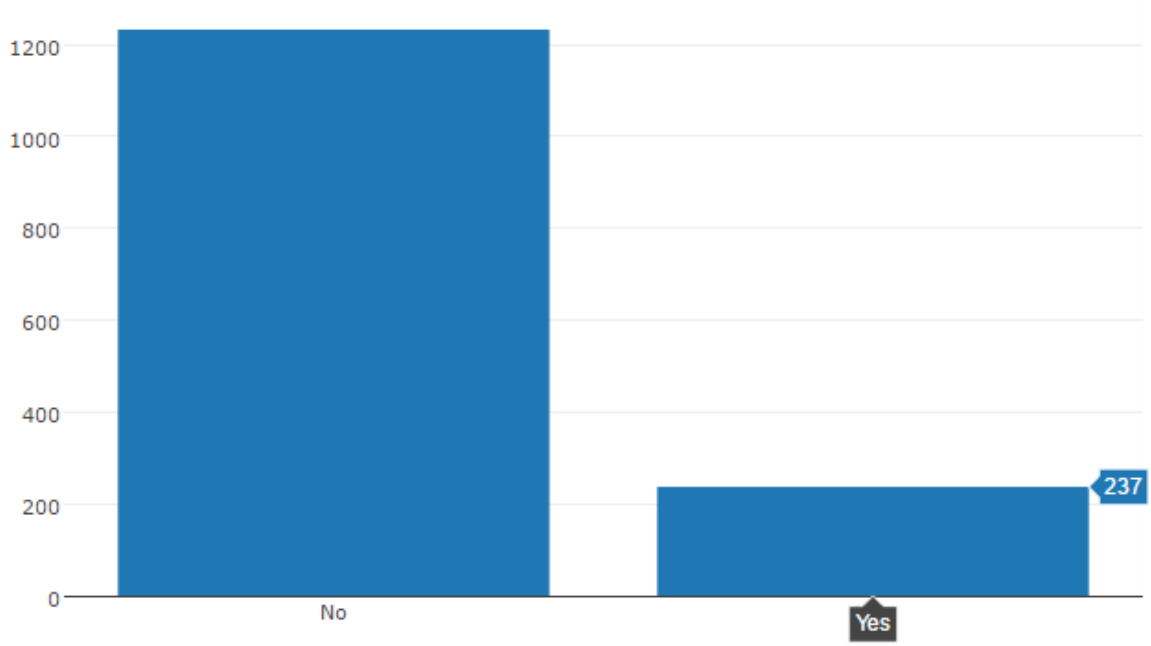
For random forest,

1. **Feature Engineering and Categorical Encoding**

Feature encoding in a nutshell involves creating new features and relationships from the current features that we have. To start off, we can segregate numerical columns from categorical columns and after having identified which of our feature contains categorical data we can encode the values conveniently by applying one line of python code.

Having encoded our categorical columns as well as engineering and created some new features from the numerical data, we can now proceed to merging both data frames into a final set with which we will train and test our models on.

One final step is to generate the target variable. The target in this case is given by the column attrition which contains categorical variables therefore requires numerical encoding. However, there is quite a large skew in target as shown:



There is quite a big imbalance in our target variable. We can use an oversampling technique to treat this imbalance.

2. **Implementing Machine Learning Models:**

Having performed some exploratory data analysis and simple feature engineering as well as having ensured that all categorical values are encoded, we can now proceed to build our model.

The Random Forest method first introduced by Breiman in 2001 can be grouped under the category of ensemble models. Why ensemble? The building block of a Random Forest is the ubiquitous Decision Tree. The decision tree as a standalone model is often considered a "weak learner" as its predictive performance is relatively poor. However, a Random Forest gathers a group (or ensemble) of decision trees and uses their combined predictive capabilities to obtain relatively strong predictive performance - "strong learner".

This principle of using a collection of "weak learners" to come together to create a "strong learner" underpins the basis of ensemble methods which one regularly comes across in Machine learning.

The random forest returns an accuracy of 88% for its prediction. But as our target variable is skewed 84% and 26 %, our model is only predicting slightly better than random guessing.

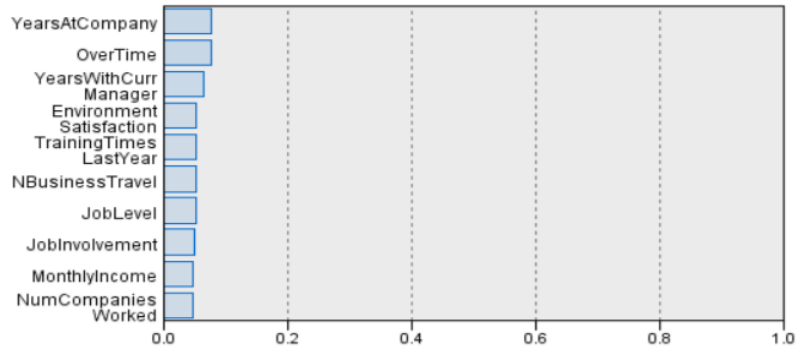
3. **Feature Ranking via the Random Forest**

Feature importance is a very convenient and useful attribute in random forest classifier which tells us which feature within our dataset has been given most importance through the Random Forest algorithm.

Most RF important features: Overtime, Marital Status

Artificial Neural Network

Figure 2. Importance of Variables According ANN analysis



Artificail Neural Network was implemented on the dataset using Python's Keras library and TensorFlow as the backend. Without any data cleaning the important variables and confusion matrix by ANN was given. The accuracy came out to be 85.33%. To improve the accuracy, a number of steps were taken:
1. data cleaning
2. adding derived information
3. Using dummy variables and dummy traps
4. Hyperparameters
5. Training the neural network with K fold cross validation

The accuracy increased and came out out to be 97.5%

Figure 3. ANN Confusion Matrix

Comparing $N-Nattrition with Nattrition

'Partition'	1_Training		2_Testing	
Correct	922	90.39%	384	85.33%
Wrong	98	9.61%	66	14.67%
Total	1,020		450	

Coincidence Matrix for $N-Nattrition (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	852	19
1.000000	79	70
'Partition' = 2_Testing	**0.000000**	**1.000000**
0.000000	352	10
1.000000	56	32

```
print("Best accuracy: ",max)
```

```
Best accuracy:  0.974999979138
```

1. Cleaning the data

Removing the following columns:

- Standard hours: As it was always 80 for all employees this column was useless.
- Over18: Yes, for all
- Employee number: This column was not useful for what we are looking for and could have confused the ANN.

2. Adding some more derived information

Playing with the data we can add some more information to help the network.

We don't need to be expert in the field in Deep Learning, but providing some ratios and extra information can help the network to converge faster.

3. Dummy variable and dummy trap

Here we will create dummy variable for all the categorical data we just encode. (Only if there is more than 2 categories).

We are doing this because if we leave Single as 0, Married as 1 and Divorced as 2, the network would understand that divorced > married, which doesn't make any sense.

Yet we do not want to fall in the dummy variable trap and we will be removing the first column of each of those dummy variables.

Why?

Because if we have 1 0 0 for a Single Person now. We could guess it is single even if we had removed the first column: 0 0 (not divorced, not married --> single). This way we can remove some duplicated features on each OneHotEncoding.

4. **Hyperparameters**

To avoid overfitting such a tiny dataset we will use dropout (randomly putting "off" 10% of the neurons to help them become more independent)

5. **Training the Neural Network, using a K Fold Cross Validation**

For initializing the weight, we will use a truncated normal distribution.

As we want a probability for output we will use a sigmoid activation function on the output layer.

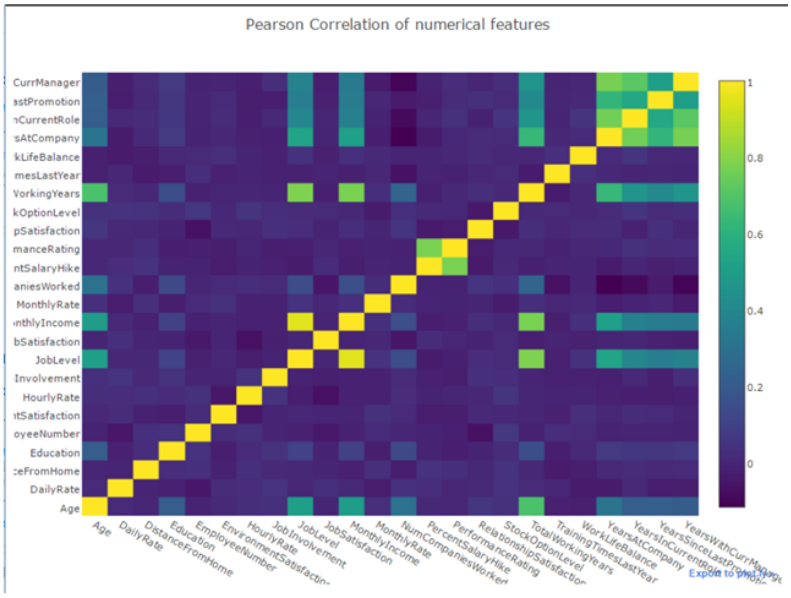
Because we are working on a categorization problem with only 2 categories we will calculate our loss with binary cross entropy.

We will use a 10 Fold Cross validation here: the validation data will be pick randomly K times and we will train the network 10 times on each of those training-validation set.

Accuracy: 97.5%

Statistical Modelling

	Attrition	p < 0.05
PerformanceRating	0.990075	False
Education	0.545525	False
Gender	0.290572	False
RelationshipSatisfaction	0.154972	False
TrainingTimesLastYear	0.0191477	True
EducationField	0.00677398	True
Department	0.00452561	True
WorkLifeBalance	0.00097257	True
JobSatisfaction	0.0005563	True
EnvironmentSatisfaction	5.12347e-05	True
BusinessTravel	5.60861e-06	True
JobInvolvement	2.86318e-06	True
MaritalStatus	9.45551e-11	True
StockOptionLevel	4.37939e-13	True
JobLevel	6.63468e-15	True
JobRole	2.75248e-15	True
OverTime	8.15842e-21	True



After performing EDA, the next logical step was to figure out which of the attributes have a statistically significant categorical relationship with Attrition, which is our target variable. To do this, Chi-square test was performed and the p-values of the attributes were obtained. Attributes having p-values > 0.05 had a non-significant categorical relationship and the others had a statistically significant categorical relationship

From te correlation plot, it was observed that quite a lot of columns were poorly correlated with one another. Generally, when making a predictive model, it would be preferable to train a model with features that are not too correalted with one another so that we do not have to deal with redundant features.

Note: Already discussed previously.