HR-ANALYTICS CASE STUDY

PRESENTATION

Group Name:

1. Shahana Abdul karim
2. Karthik Ganapavarapu
3. Nikhil Gupta
4. Sai Manideep Allu

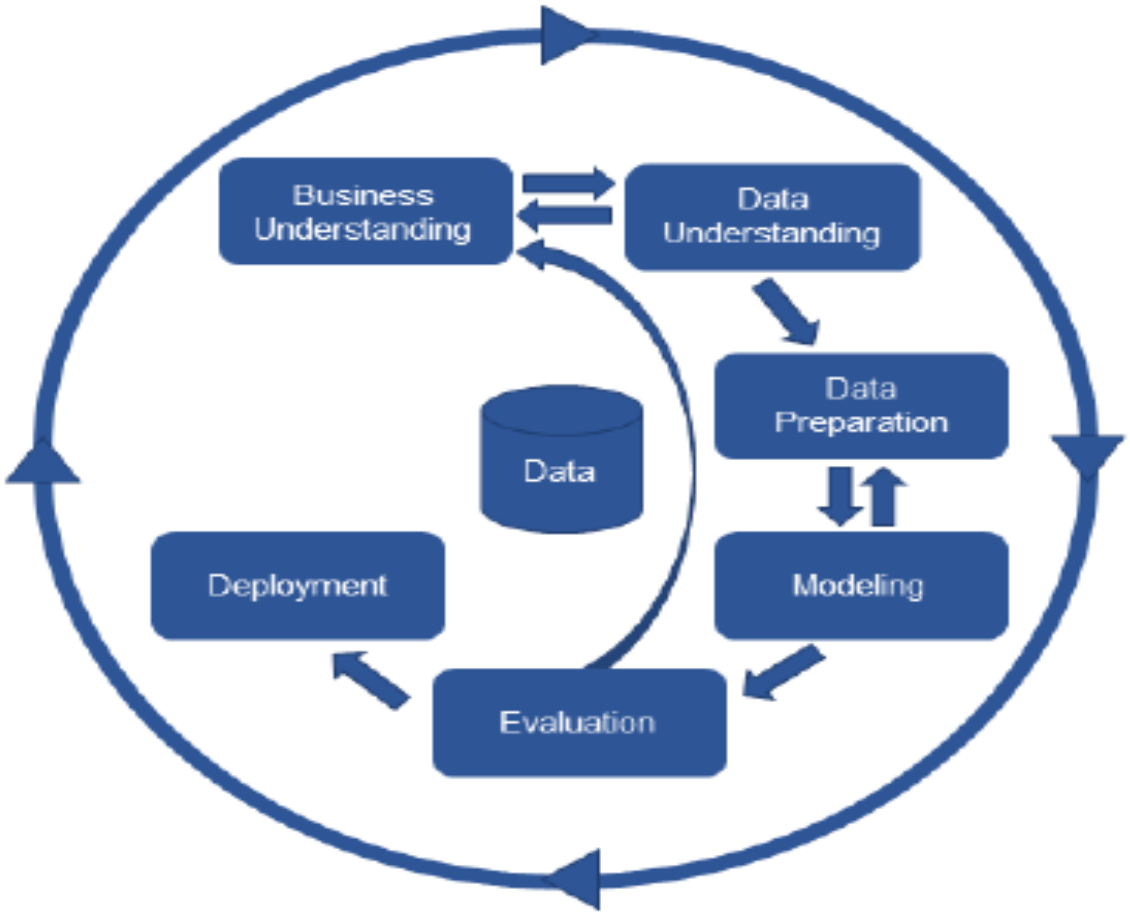
Business Objective

XYZ company is a large company with 4000 employees at any point of time, However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company.

Hence, the management has contracted an HR analytics firm to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay. Also, they want to know which of these variables is most important and needs to be addressed right away.

Goal of this case study is to find the factors affecting the attrition rate of the employee using Logistic Regression and present them in business form so that the company can take appropriate steps to retain the employees

Problem Solving Methodology



- **Data Understanding**: Data has been provided in 5 categories – General data of employee, Employee survey data, Manager survey data, In and out time details for year 2015.
- **Data Preparation**: It includes handling missing values, outliers, duplicates, creation of dummy variable for categorical variables, deriving new metrics (wherever applicable) and formatting data before proceeding with modeling.
- **EDA** – To understand the trends within data against target variable (Attrition).
- **Modeling:** It involves analyzing the nature of predicted variable. As attrition is binomial in nature, so we use glm function available in R to build logistic regression model.
- **Evaluation**: Steps involve to validate the model for the test dataset using Confusion Matrix(determining accuracy, specificity and sensitivity), also comparing outputs from the model with a random model to check the performance and summarizing the results keeping the business success constraints in mind.

Selected attributes through initial visual analysis

S.No	Attribute Name	Description
1.	Age	Age of employee.
2.	Attrition	Whether the employee left in the previous year or not
3.	Business Travel	How frequently the employees travelled for business purposes in the last year
4.	Department	Department in company
5.	DistanceFromHome	Distance from home in kms
6.	Education	Education Level
7.	Education Field	Field of education
8.	Gender	Gender of employee
9.	JobLevel	Job level at company on a scale of 1 to 5
10	JobRole	Name of job role in company
11.	MaritalStatus	Marital status of the employee

Selected attributes through initial visual analysis Contd….

S.No	Attribute Name	Description
12.	MonthlyIncome	Monthly income in rupees per month
13.	NumCompaniesWorked	Total number of companies the employee has worked for
14.	PercentSalaryHike	Percent salary hike for last year
15.	StockOptionLevel	Stock option level of the employee
16.	TotalWorkingYears	Total number of years the employee has worked so far
17.	TrainingTimesLastYear	Number of times training was conducted for this employee last year
18.	YearsAtCompany	Total number of years spent at the company by the employee
19.	YearsSinceLastPromotion	Number of years since last promotion
20.	YearsWithCurrManager	Number of years under current manager
21.	EnvironmentSatisfaction	Work Environment Satisfaction Level
22.	JobSatisfaction	Job Satisfaction Level

Selected attributes through initial visual analysis Contd….

S.No	Attribute Name	Description
23.	WorkLifeBalance	Work life balance level
24.	JobInvolvement	Job Involvement Level
25.	PerformanceRating	Performance rating for last year
26.	averagetime	Average time spent by the employee
27.	worktime	Total time worked by the employee

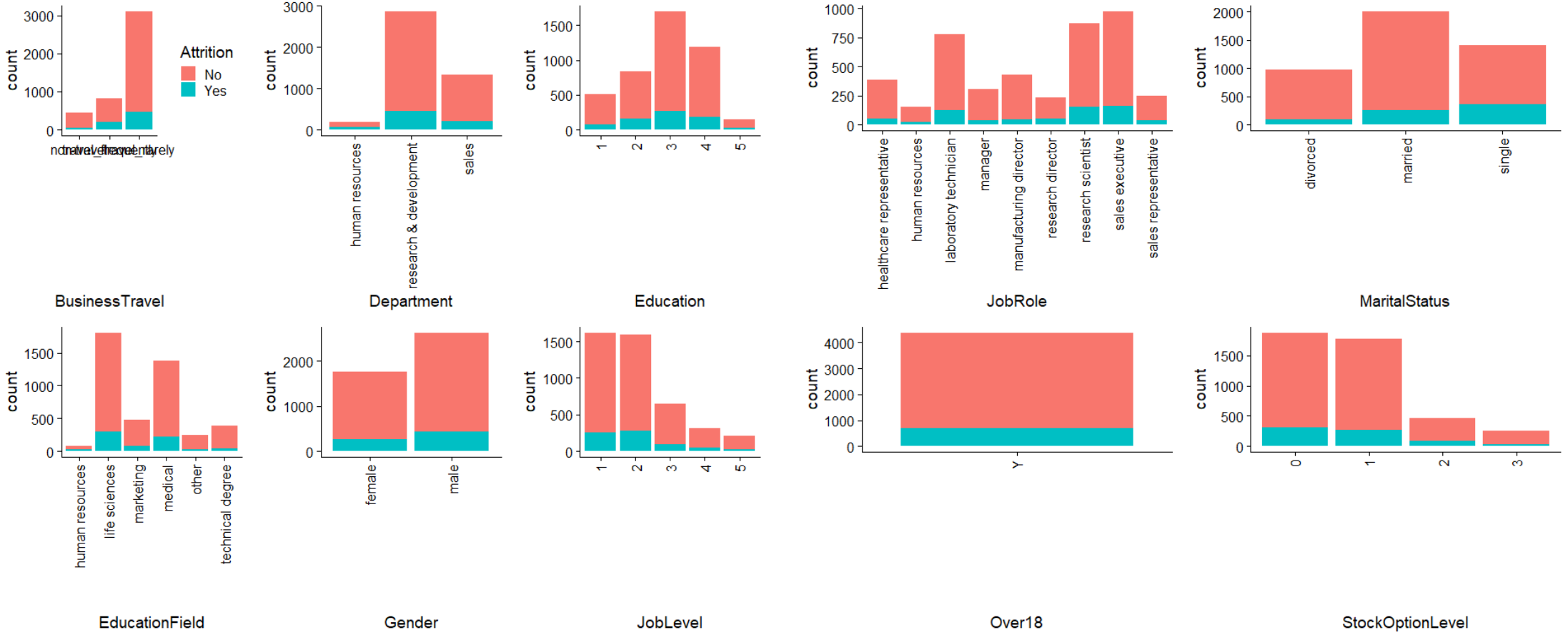
Data Cleaning And Preparation

- Reading Excel Data available in R.
- Merging of the Data in a single master frame for further analysis.
- Converting the time data into R compatible date format and finding the time difference for each day for every employee in a separate column
- Removing the rows having NA values for column Total working years and Number of companies worked as the NA values have very less percentage
- Removing those columns which are having only one kind of values in all rows.
- Imputation of the remaining NA values with the mode of the corresponding columns, Columns are EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance
- Assumption taken for Number of Companies worked column that Current Company is not counted in this column and the data is cleaned accordingly for this column

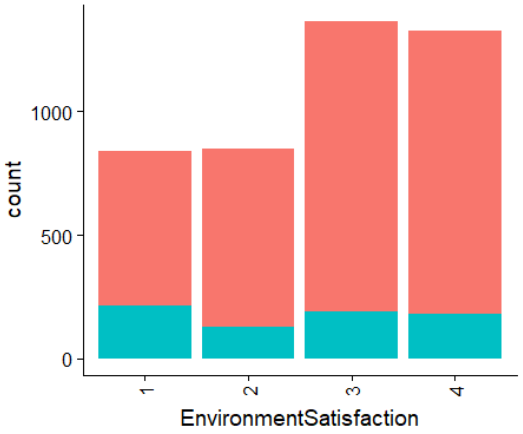
Data Cleaning And Preparation

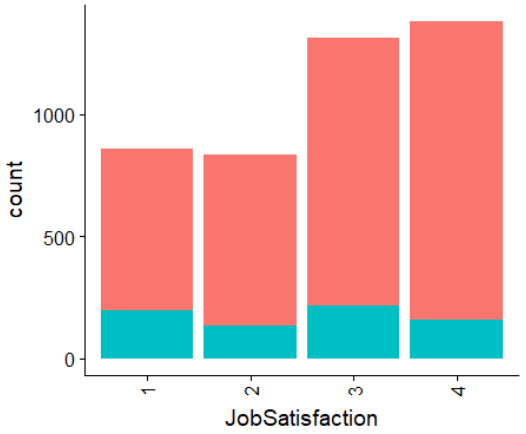
- Making the box plot for every continuous variable and checking for the outliers.
- Imputing the outliers using the Interquartile function.
- Outliers found in the columns Monthly Income, Total Working Years, Years At Company, Years since Last Promotion ,Years with Current Manager.
- Converting discrete variables into categorical variables using as.factor command in R.

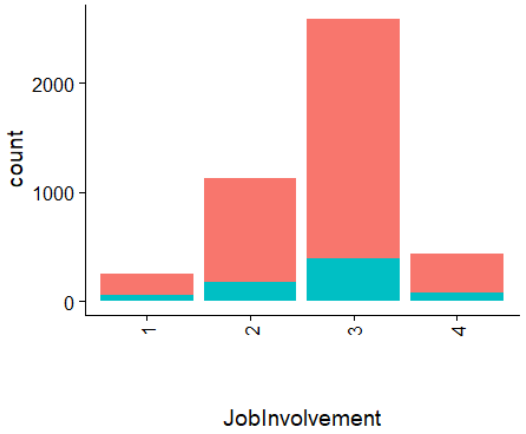
EDA - Categorical Variables

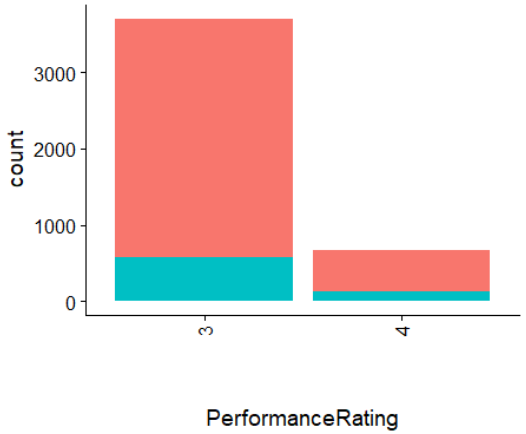


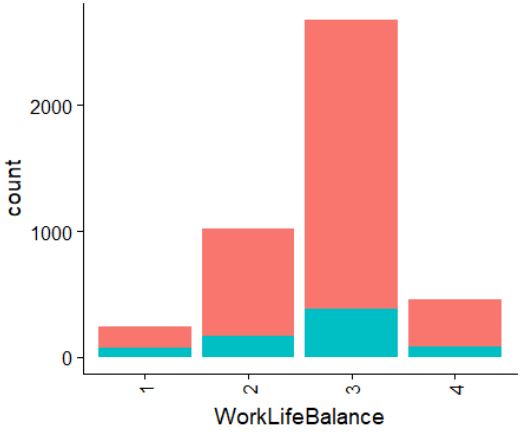
EDA- Categorical Variables

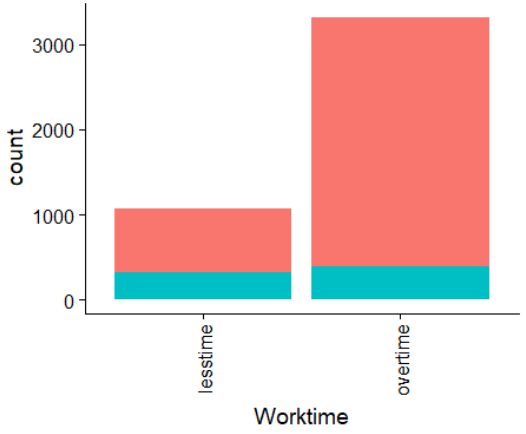




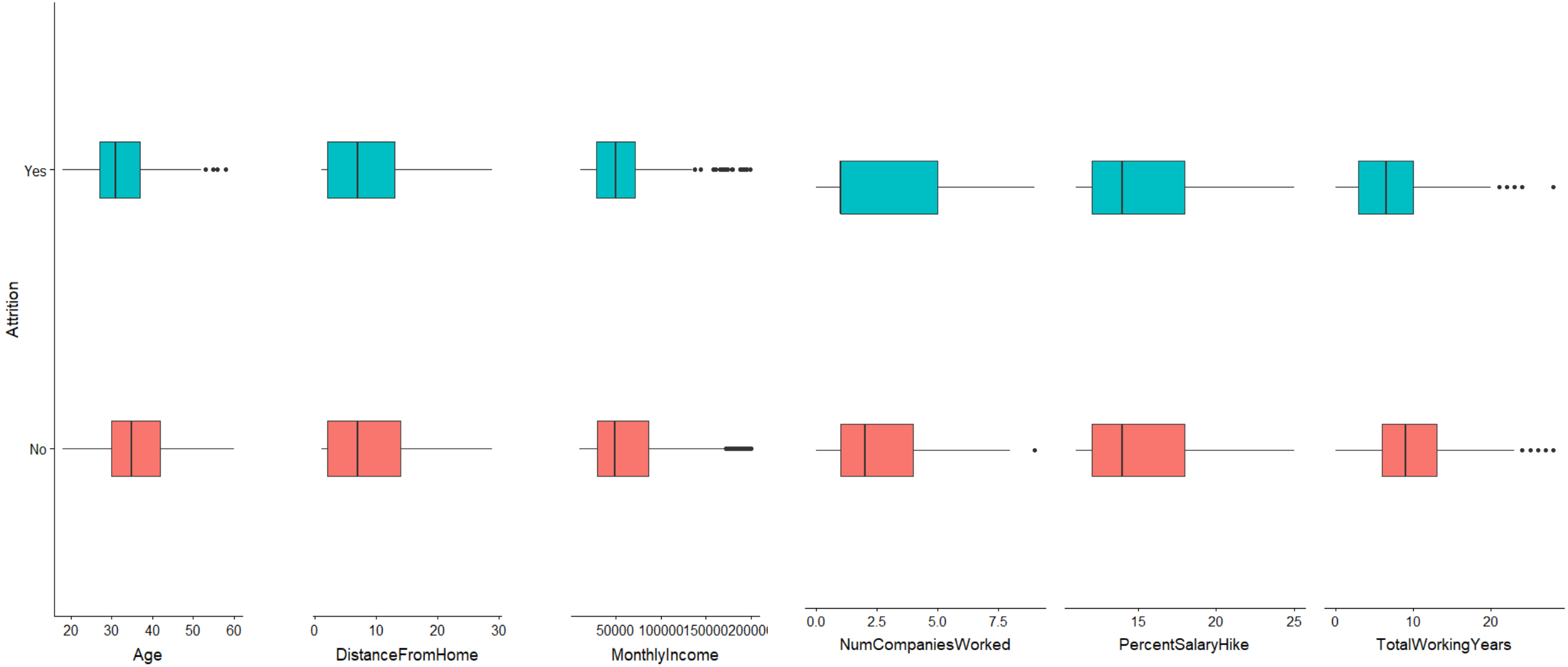




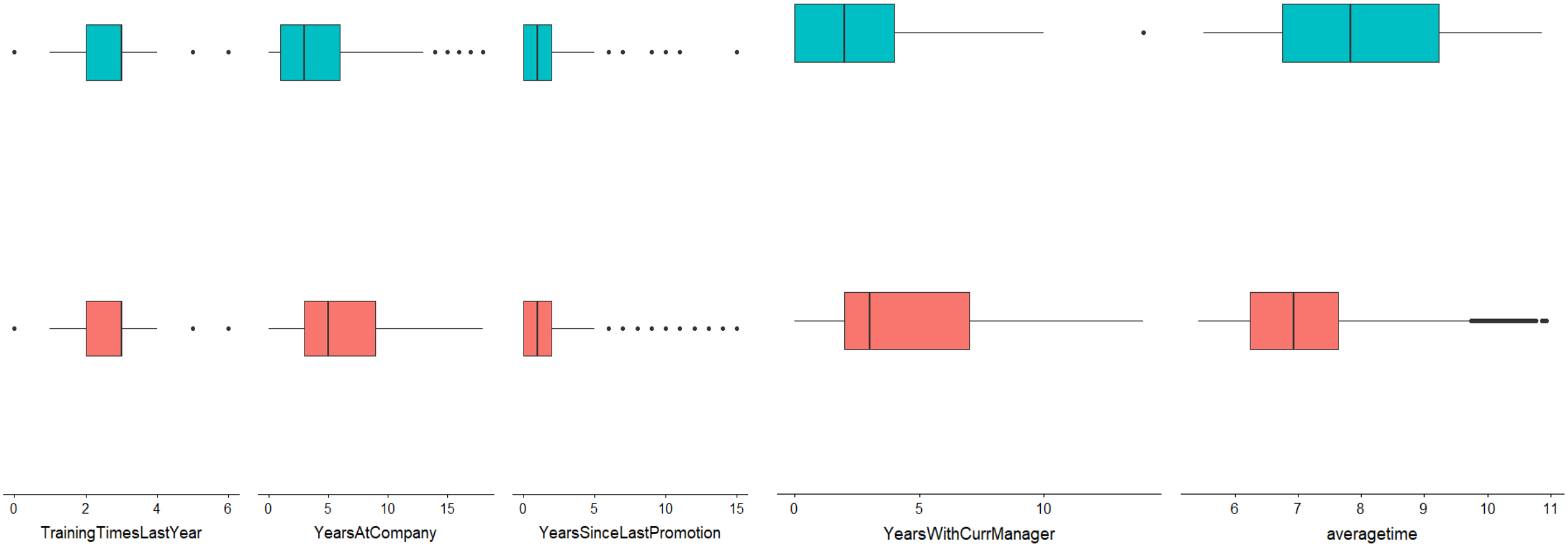




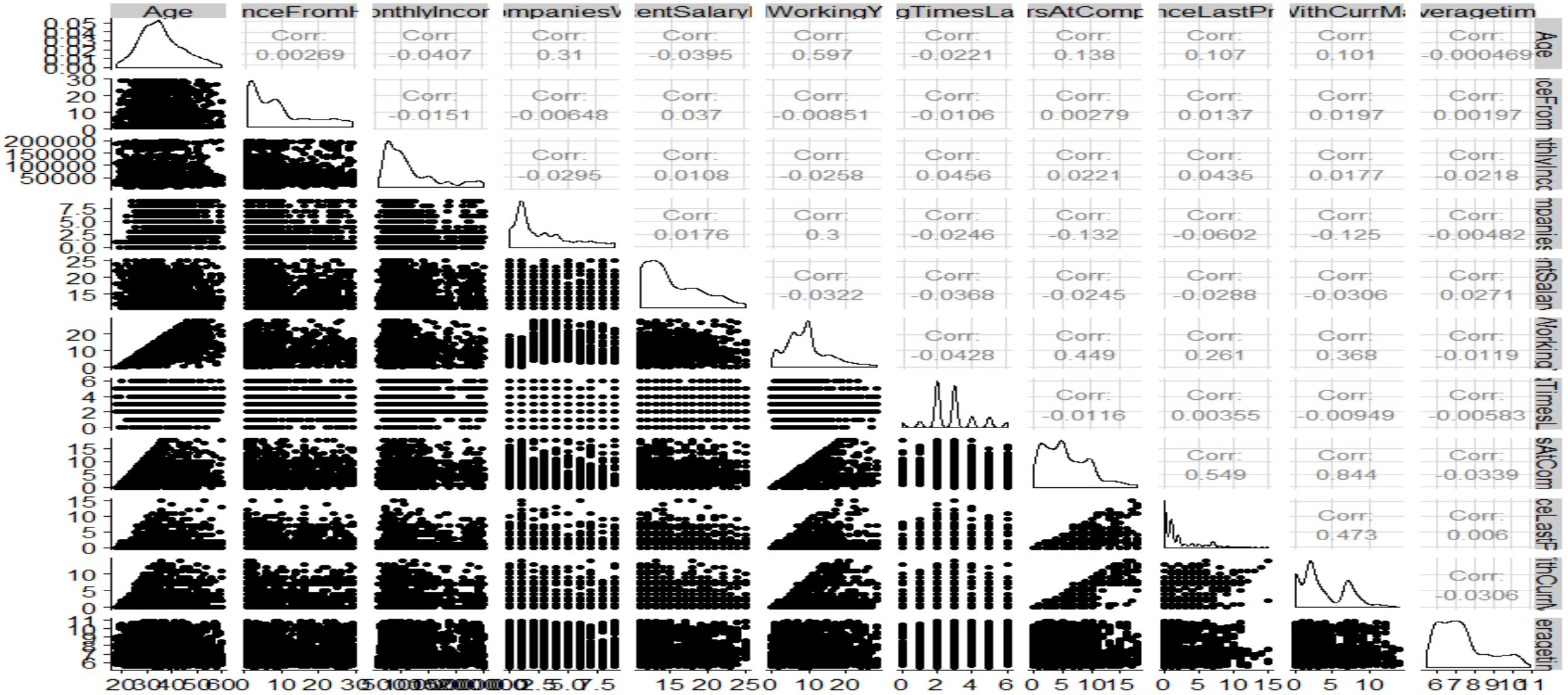
EDA- Numerical Variables



EDA- Numerical Variables

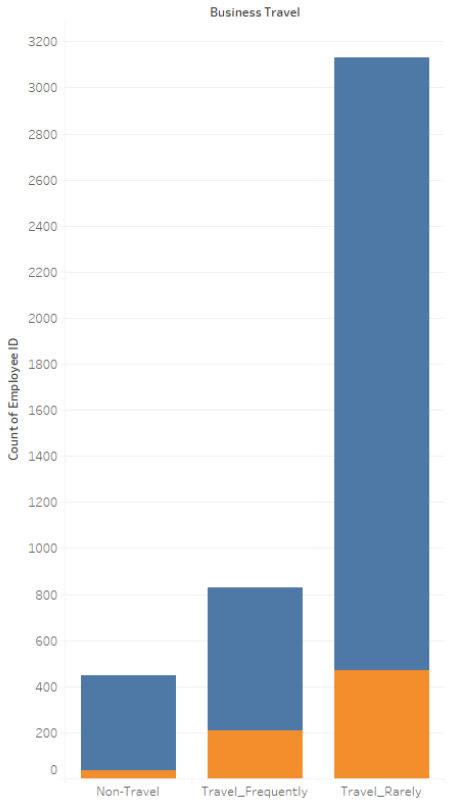


Corelation Between Numerical Variables

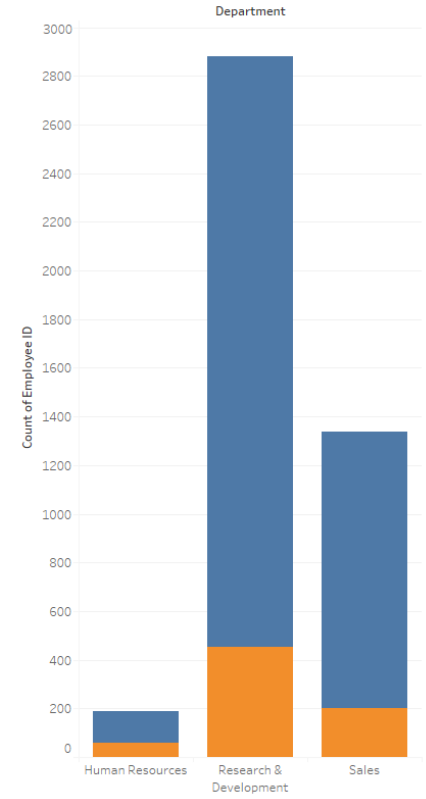


Attrition VS Categorical Variables

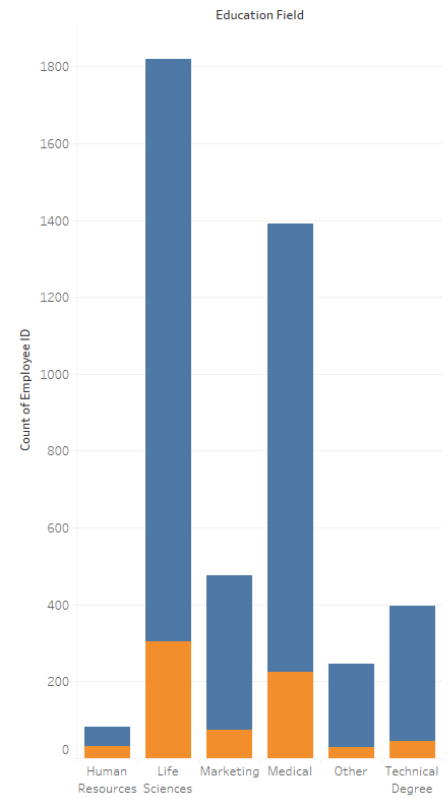
Attrition Vs Business Travel



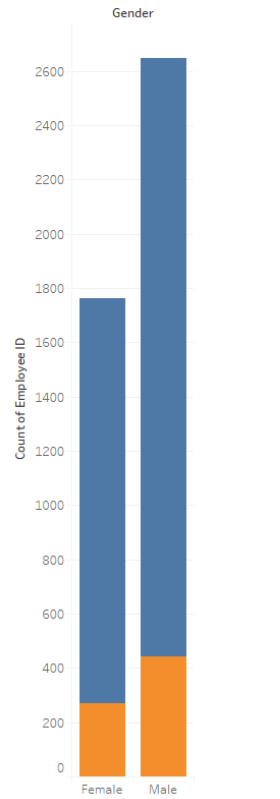
Attrition Vs Department



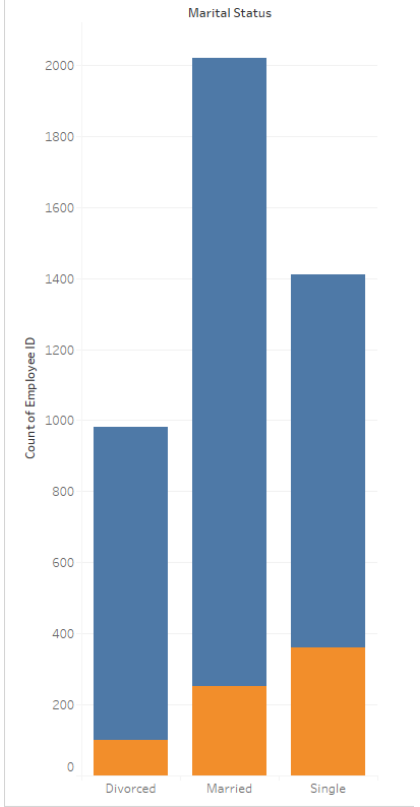
Attrition Vs Education Field

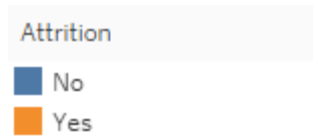


Attrition Vs Gender



Attrition Vs Maritial Status





Logistic Regression-Modelling

- Our Response variable is “Attrition” (1 == Yes, & 0 == No)
- Rest all non constant numeric variables are scaled to aid in regression modelling.
- Splitting data into training and test data set.
- For creating Train and test datasets from final data set:
- We fixed seed to 100 and Used split ratio of 0.7 for training dataset and remaining data has been assigned to test dataset
- Initial model has been conceived with glm function, then StepAIC has been applied to arrive at standard model which yielded on iterative predictor selection without major reduction in AIC Score.
- Removed the variables having high VIF value and low significance i.e. if p-value > 0.05
- Checked the correlation among variables appropriately and removed from the model accordingly.
- Then based on VIF (variance inflation factor) and P - value (with significance) predictors have been filtered and after another 28 iterations we could achieve our final model. with almost all predictors being significant with lowest VIF are present.

Significant Variables and there relation

- Overall there are 10 significant variables with positive and negative coefficients, positive coefficient means more is the chance of attrition with that variable whereas, negative coefficient means they will be affecting attrition rate but at a lesser rate.
- In case of categorical variables like EnvironmentSatisfaction the coefficients are negative as the level of satisfaction is high (with highest rating =3,4), attrition rate is thus lower compared to positive coefficients.
- YearsSinceLastPromotion 0.31717
- YearsWithCurrManager -0.66140
- Worktime -1.39222
- BusinessTravel.xtravel_frequently 0.73951
- MaritalStatus.xsingle 1.12050
- EnvironmentSatisfaction.x2 -0.82388
- EnvironmentSatisfaction.x3 -0.93616
- EnvironmentSatisfaction.x4 -0.92667
- JobSatisfaction.x4 -0.97198
- WorkLifeBalance.x3 -0.46299
- Therefore, Years Since Last Promotion , Frequent Business Travel, Single Marital Status affect attrition rate of employees at a high rate compared to the other significant variables.

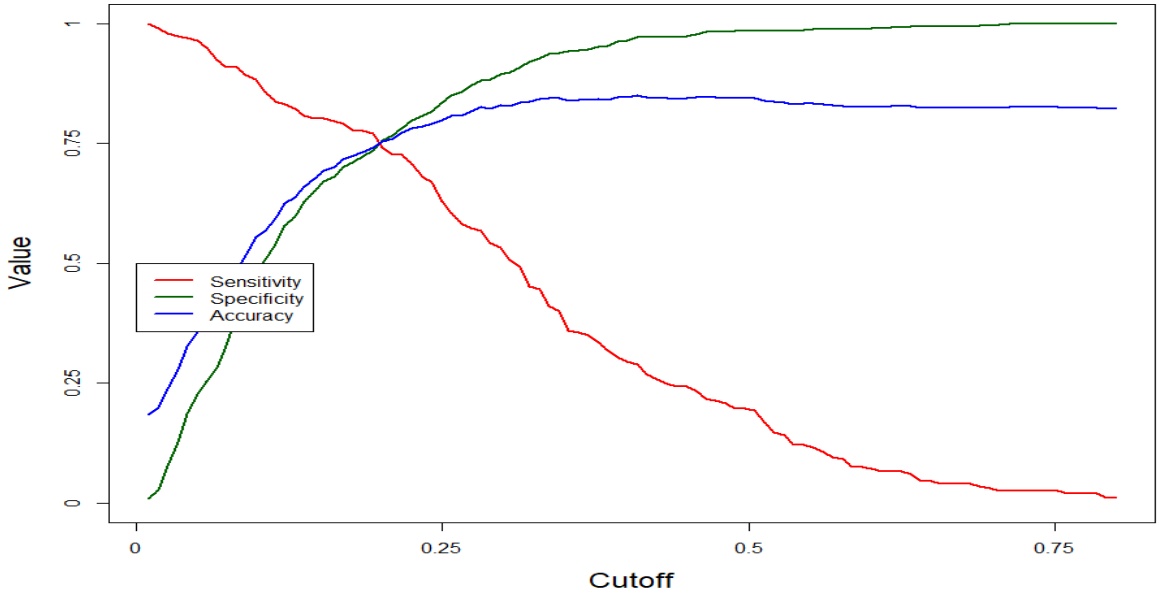
Model Evaluation

Generating confusion matrix with different probability to attrition cutoff levels and then comparing it's accuracy, sensitivity, specificity

Finding the optimal cutoff value :

Cut-off Probability	Accuracy	Sensitivity	Specificity
50%	84.5%	20.2%	97.8%
40%	83.3%	26.23%	95%
30%	82.1%	48.5%	89%
18% (optimal)	70.6%	72.7%	70.2%

Model Evaluation



The model catches 2.1 times more attritions than a random model would have caught and is optimal at 18% probability Of attrition.

It is a good model as it's KS statistic is above 42% falling within the 4th decile.

72.7% employee will churn accuracy model
+ 70.2 % employee won't churn accuracy model

bucket	total	totalresp	Cumresp	Gain	Cumlift
1	119	63	63	31.2	3.12
2	119	37	100	49.5	2.48
3	118	28	128	63.4	2.11
4	119	22	150	74.3	1.86
5	118	15	165	81.7	1.63
6	119	10	175	86.6	1.44
7	119	10	185	91.6	1.31
8	118	6	191	94.6	1.18
9	119	8	199	98.5	1.09
10	118	3	202	100	1

Gain/Lift table

- by the 4th decile,then among the top 40% employees who are sorted according to the probability in decreasing order,
- 74.3% of those employees are likely to leave/get fired.
- by the model's gain by the end of the 3rd decile is 2.1 times that of a random model's gain at the end of 3 deciles.
- In other words, the model catches 2.1 times more attritions than a random model would have caught.

Conclusion

- Model and EDA lay out the following factors for the company:
- 1. Company has to take the factors like number of companies worked for before, Years Since Last Promotion , Frequent Business Travel, Single Marital Status into consideration .
- 2. Company must focus on improving improve the Working conditions in order to improve employee efficiency and productivity.
- 3. Company must also cater to requirements or conditions of Research & Development , Sales departments.
- 4. Then there are some factors such as Job role, Work Life balance which are interlinked with the working conditions.
- 5. Number years worked under a particular manager also show attrition. Constant reviewing of managers- employee performance may also help fix problems of attrition.