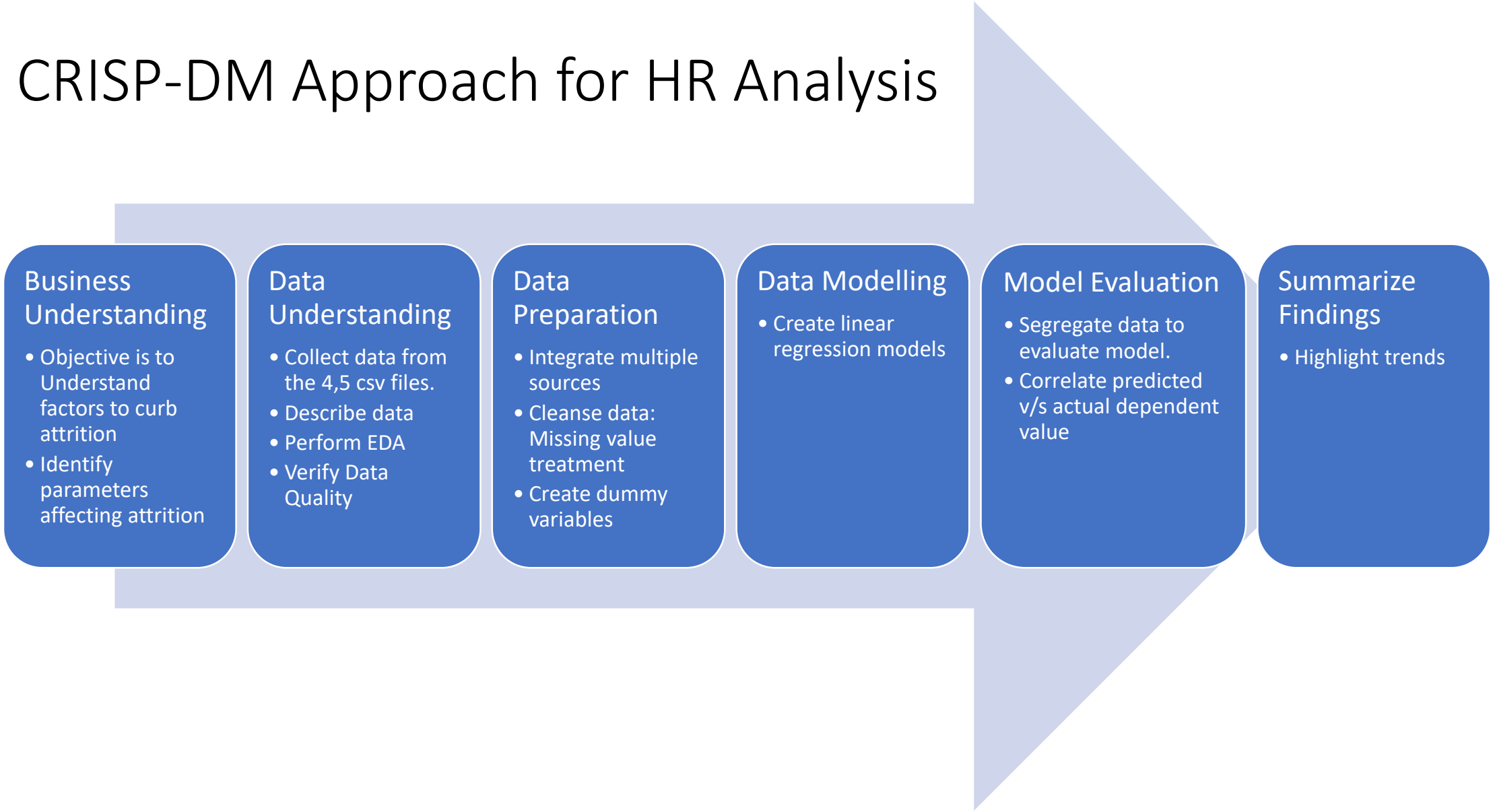
HR Analytics Case Study

1. Varsha Shastri G
2. Guruchetan Prabhakar
3. Samik Bhattacharya

Submission Deadline: 20th Aug 2017 11:59 pm

CRISP-DM Approach for HR Analysis



Business Understanding

- Business Objective
  - Reduce Attrition by understanding factors encourage employees to stay
- Goal of the data analysis
  - Model the probability of attrition using a logistic regression.

The outcome of the data analysis can be used by the management to make improvements to the workplace that reduce attrition.

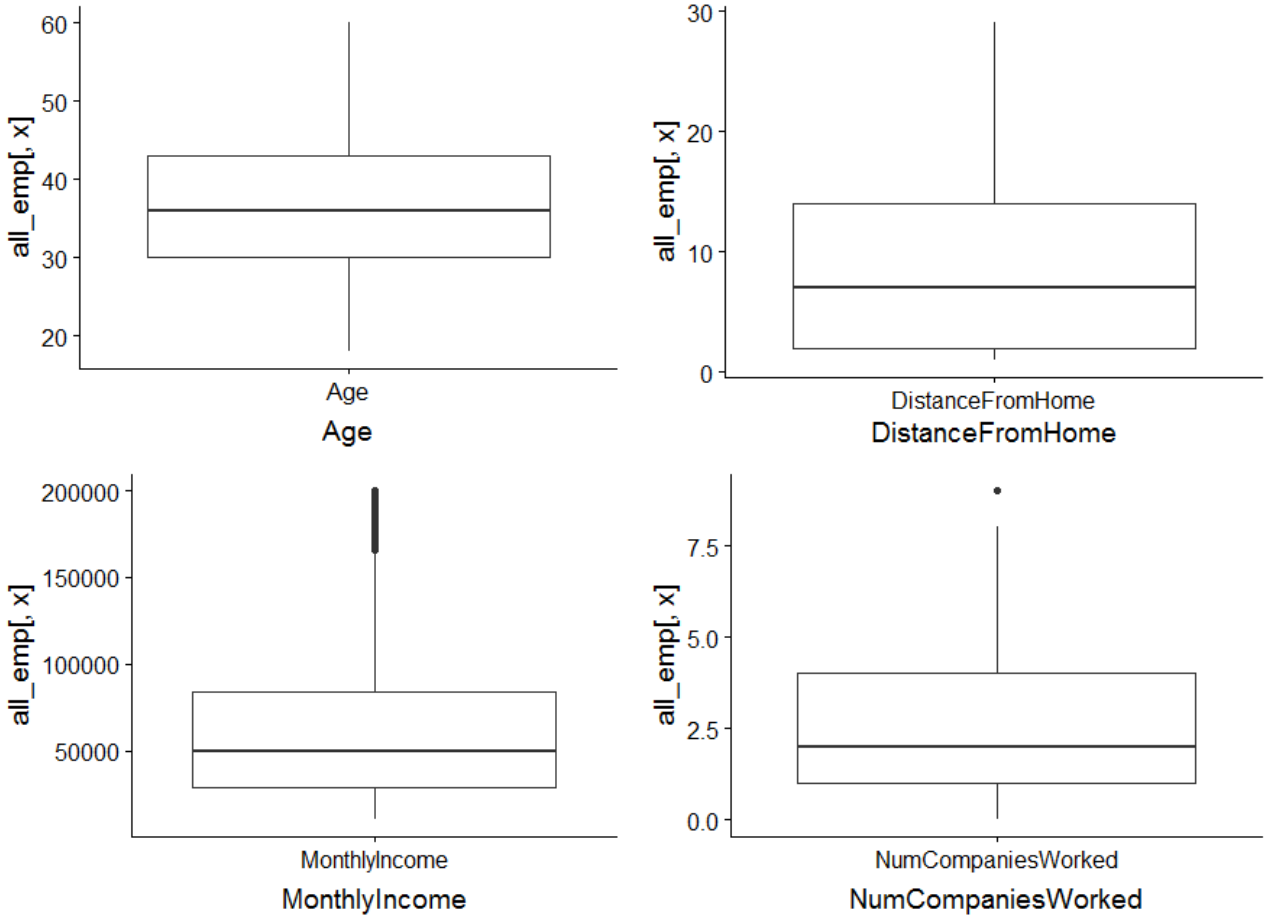
Data Understanding: Data Files

Information for 4410 employees are presented in 5 files.

- 5 files contain data that can be used for analysis.
  - **employee_survey_data.csv**: Job satisfaction parameters captured from employee survey.
  - **general_data.csv**: Attrition (Yes/No) and other attributes collected by the employer
  - **in_time.csv & out_time.csv** : Login in and log out time for each employee
  - **manager_survey_data.csv**: Quantified manager's assessment of employees
  - **data_dictionary.xlsx**: Data field information that includes meanings and levels.
- Information present in in_time and out_time are summarized as the average time spent in office
- Merge information present in various csv files on the basis of "Employee ID".
- The dataset contained 4410 objects with 27 variables

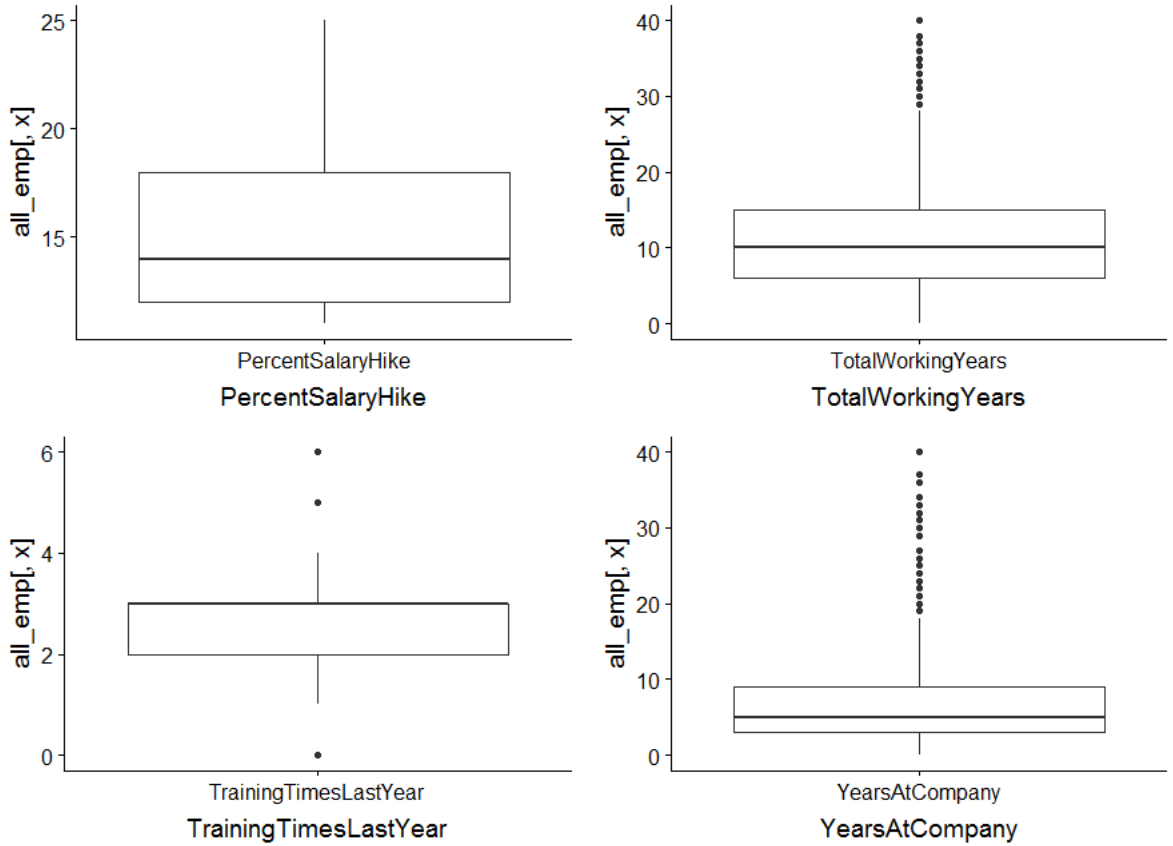
Data Understanding: EDA – Univariate Analysis

- As expected there are no outliers in the Age, Distance from home.
- However there are outliers for Monthly income



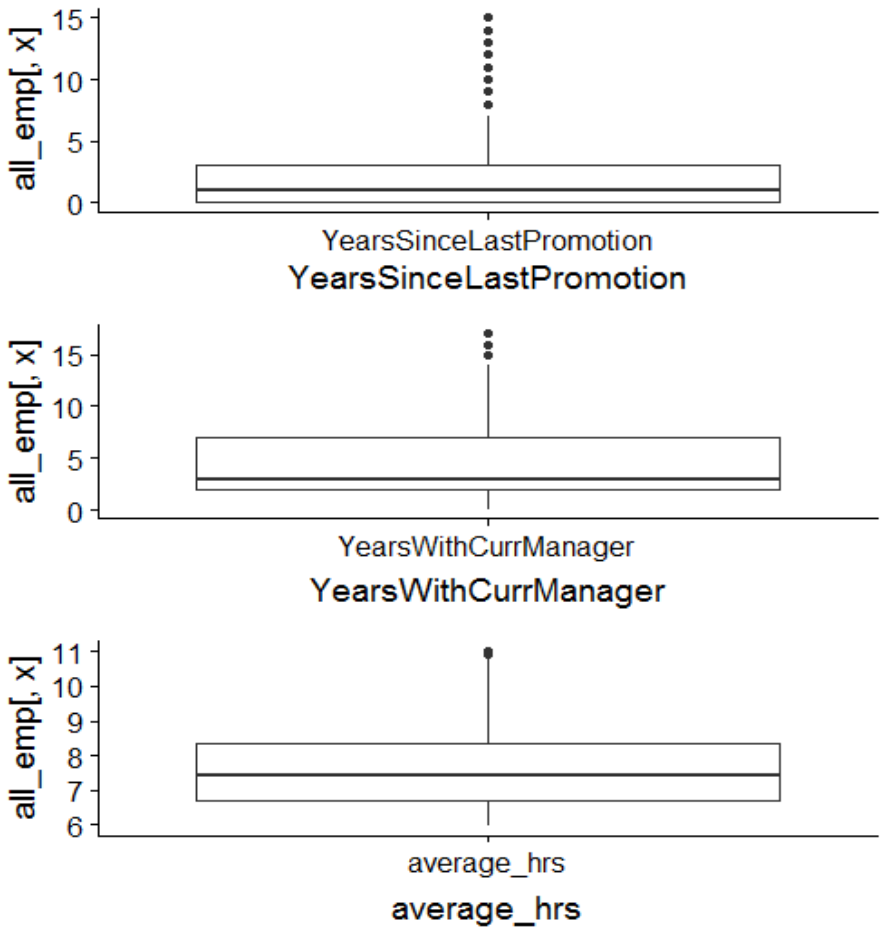
Data Understanding: EDA – Univariate Analysis

- No Outliers exist for Salary Hike in percentage terms.
- However, outlier treatment is required for Total Working Years, Years At Company and Training times Last Year



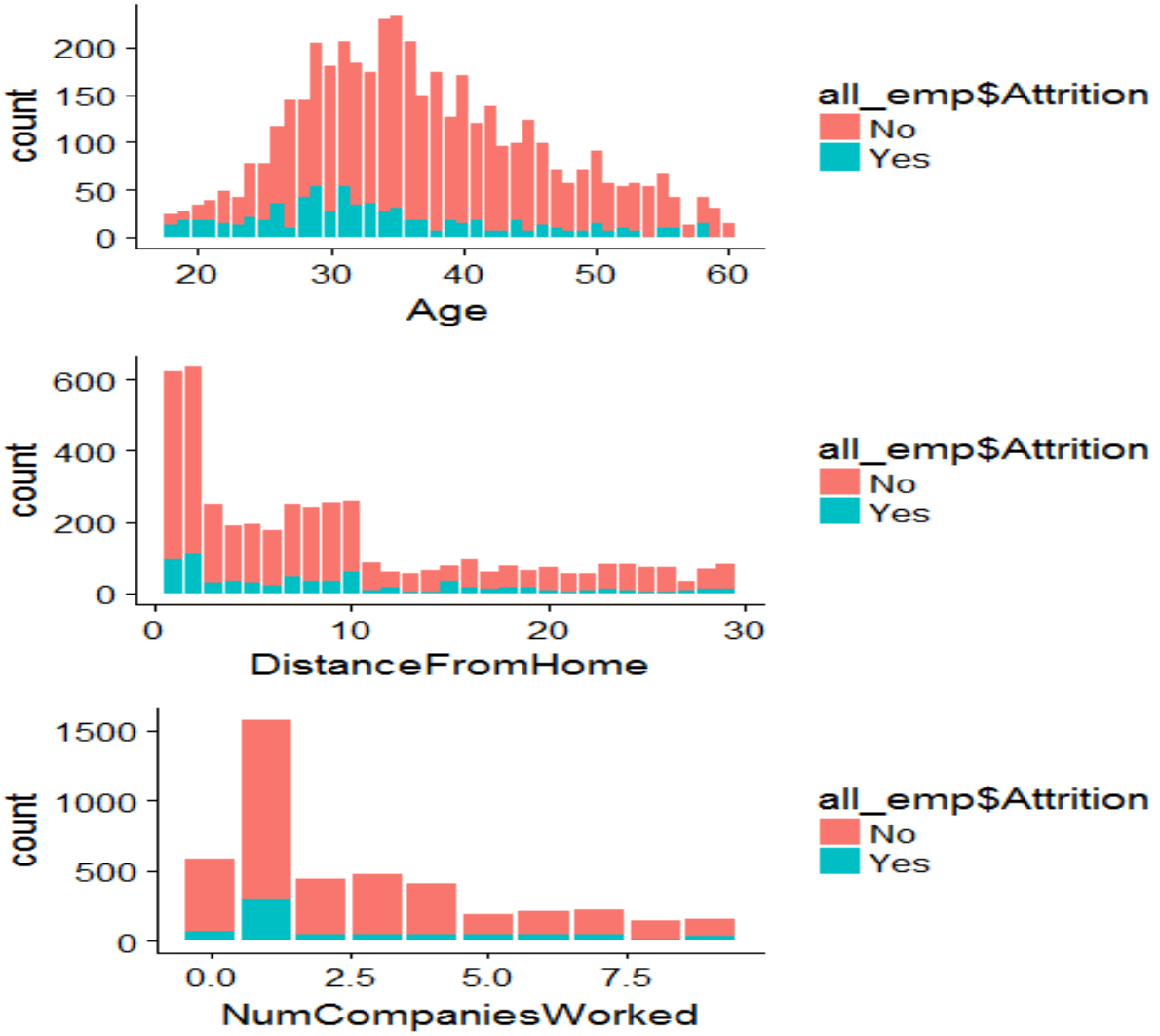
Data Understanding: EDA – Univariate Analysis

- Outliers exist for “Years since Last Promotion” and “Years with Current Manager”
- There is a small proportion of outliers in the average hours spent



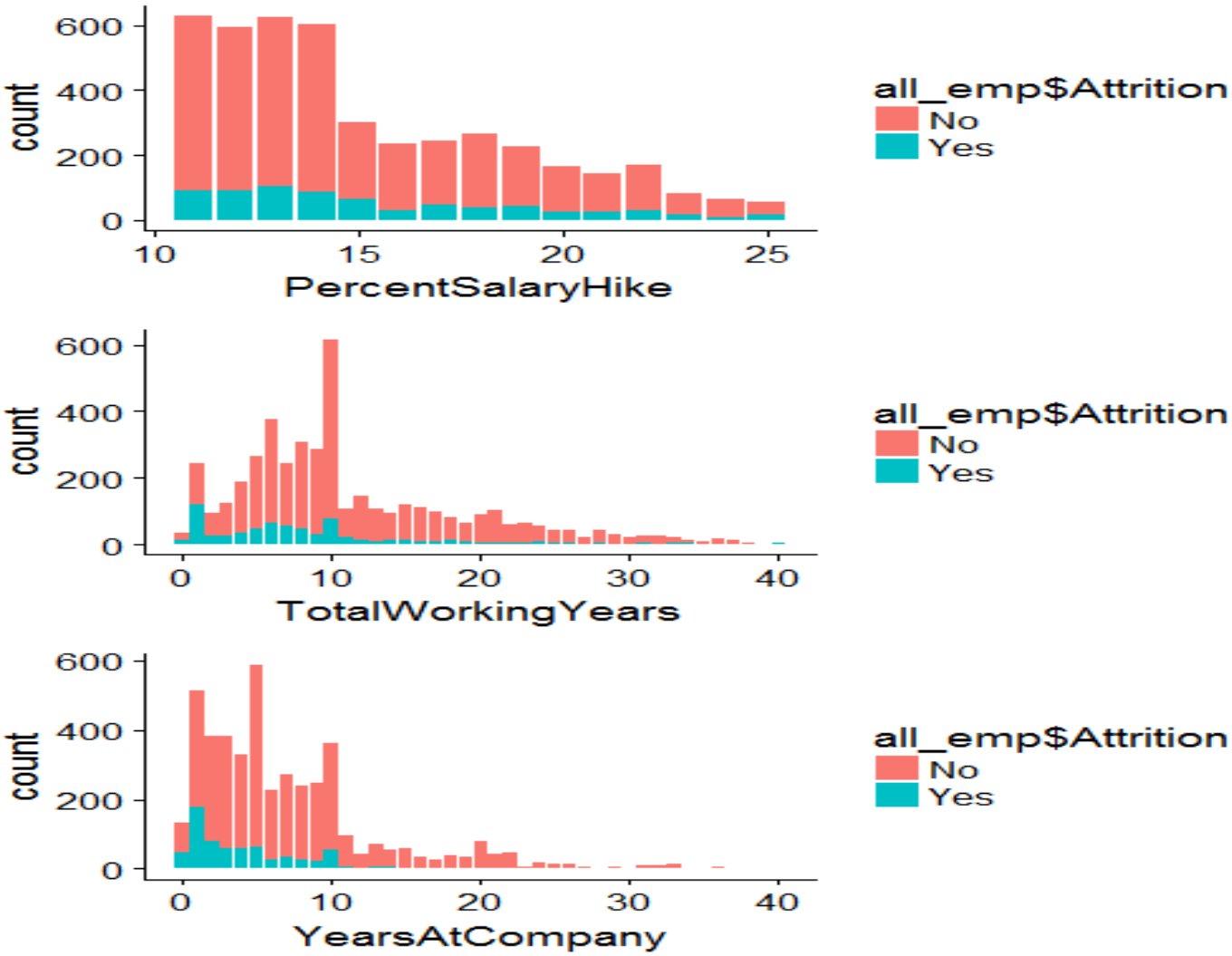
Data Understanding: EDA – Segmented Univariate Analysis

- Higher percentage of Attrition is at the early 20's
- Number of employee attrition is highest around 30 years

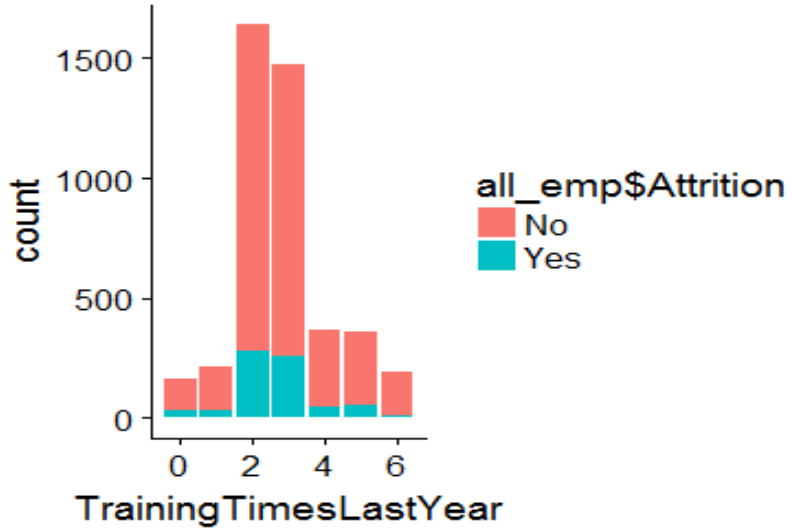


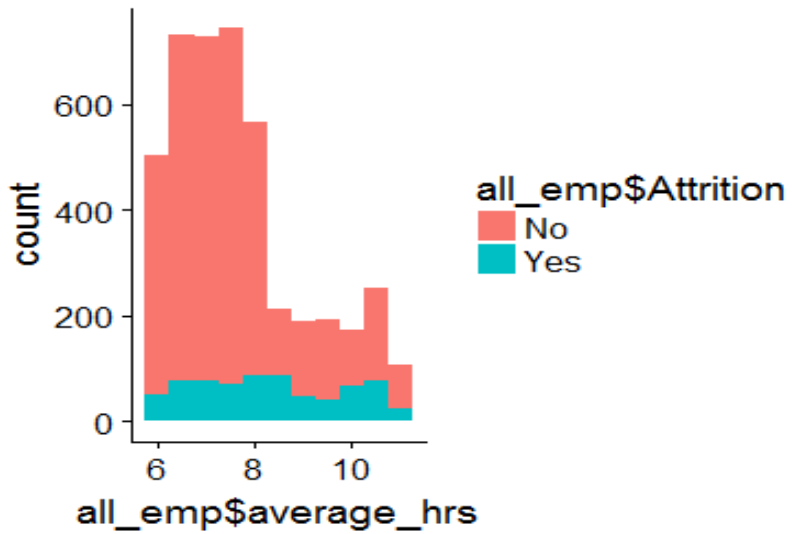
Data Understanding: EDA – Segmented Univariate Analysis

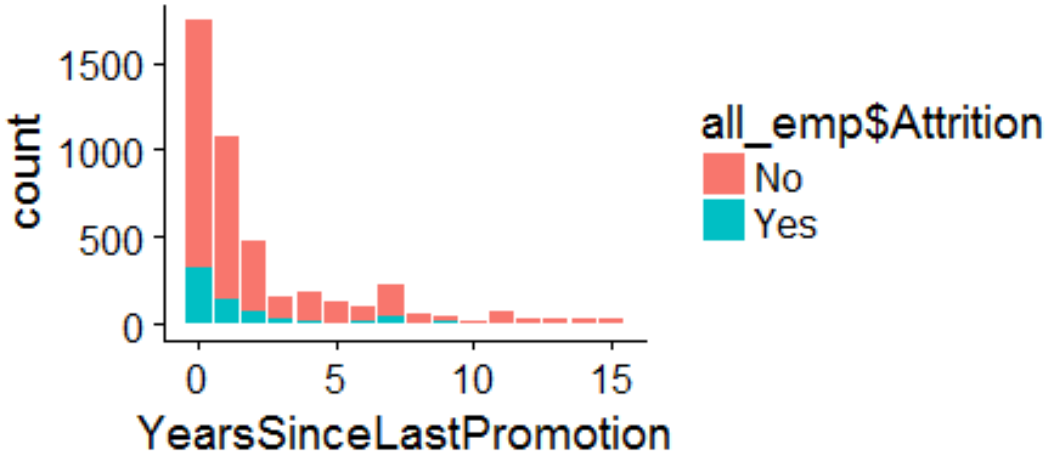
- When percentage hike is lesser there is lesser chance of attrition.
- We see that for people around total working experience of around 10 years, attrition is the least and for 1 year it is the maximum
- Lesser number of years spent in the company, higher the chances of attrition.
- In future slides we can see more graphs showing trends of different variables:

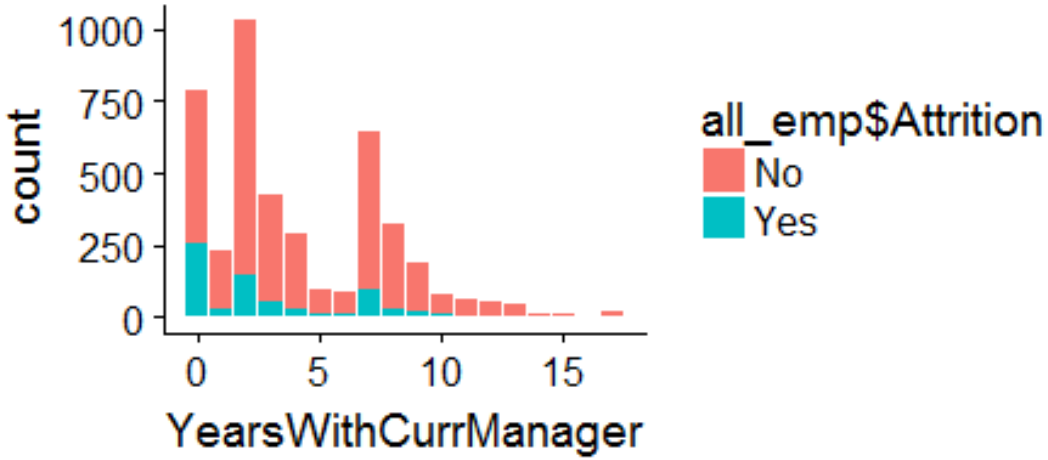


Data Understanding: EDA – Segmented Univariate Analysis

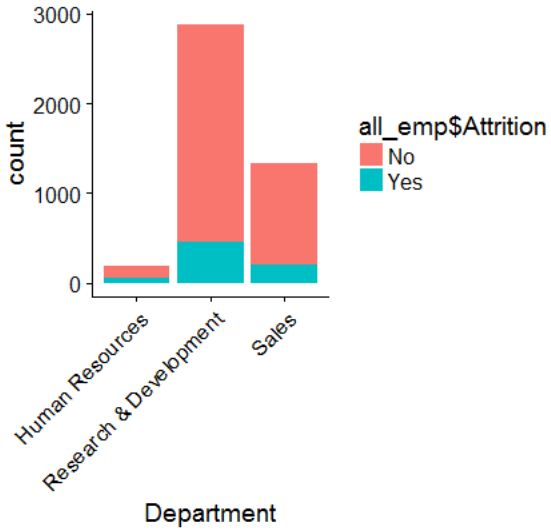


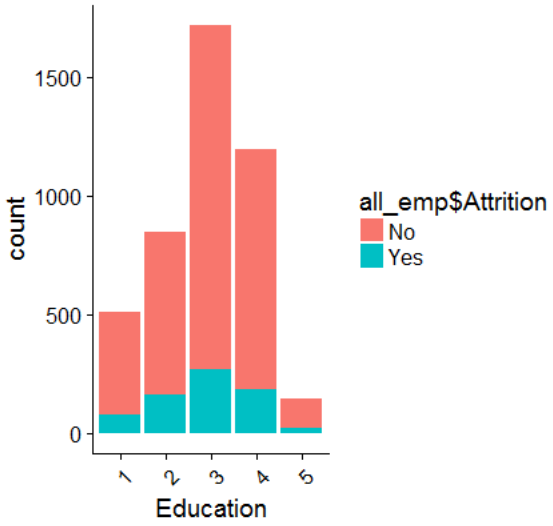


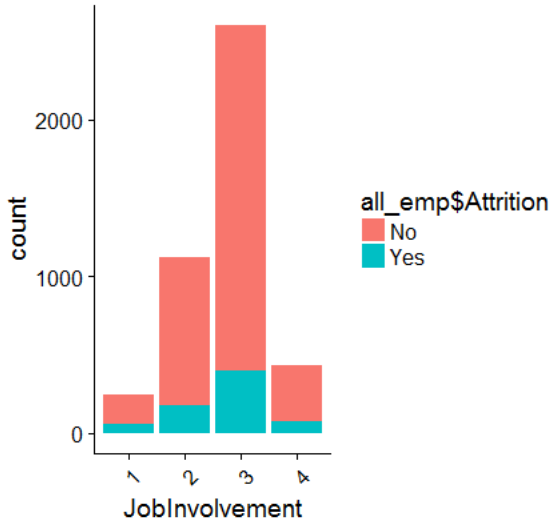


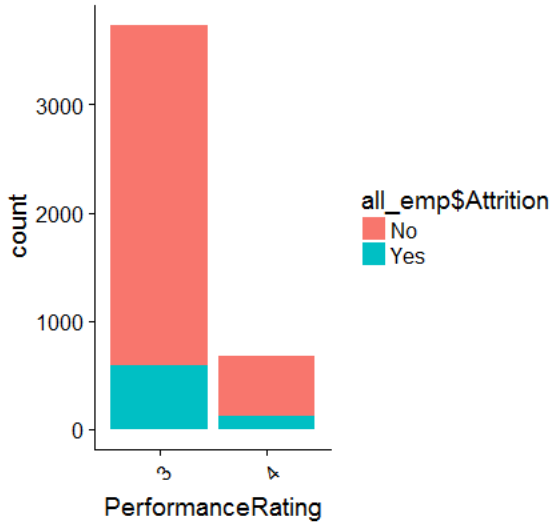


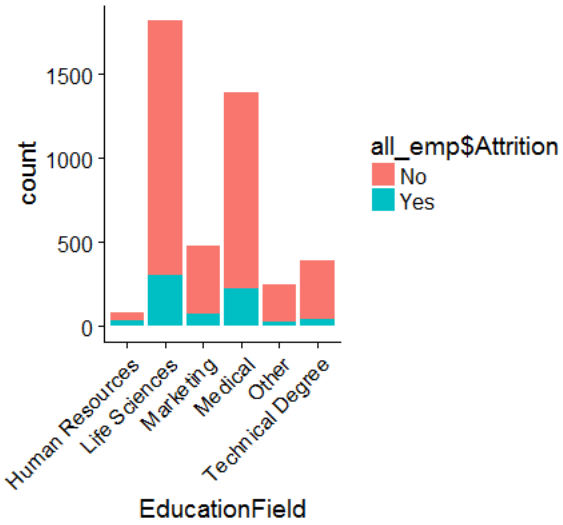
Data Understanding: EDA – Categorical Univariate Analysis

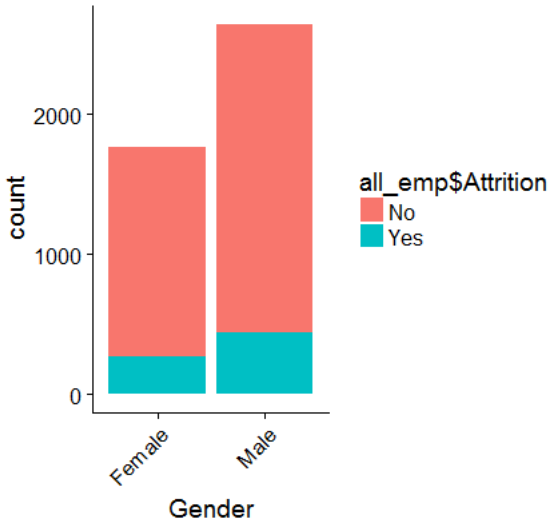


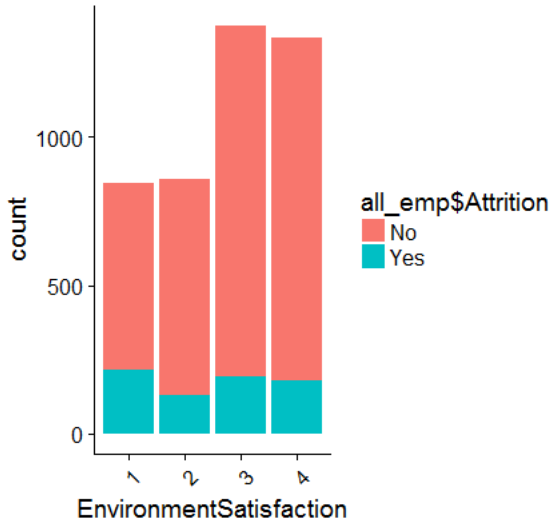


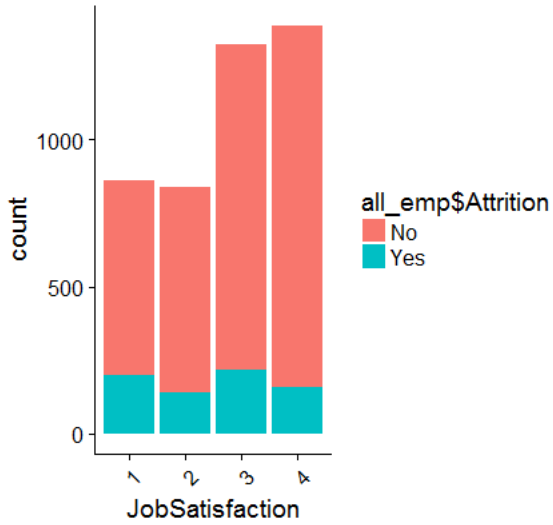




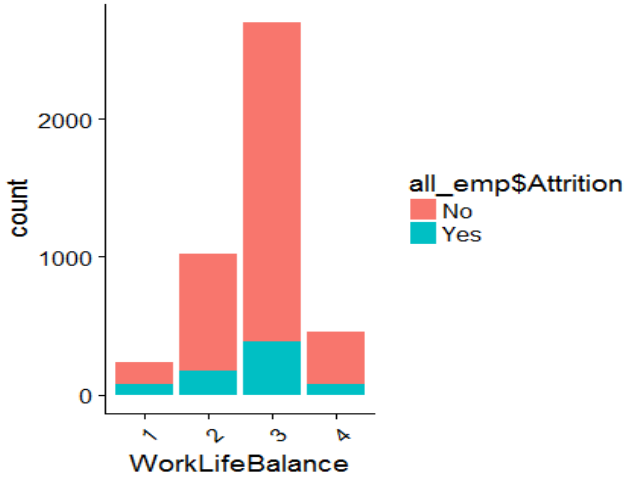


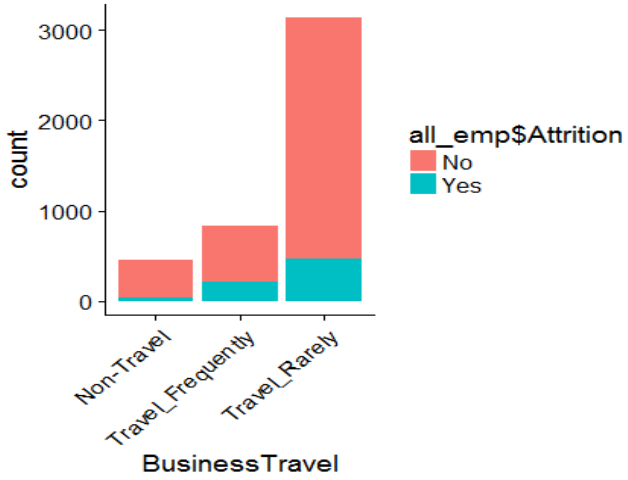


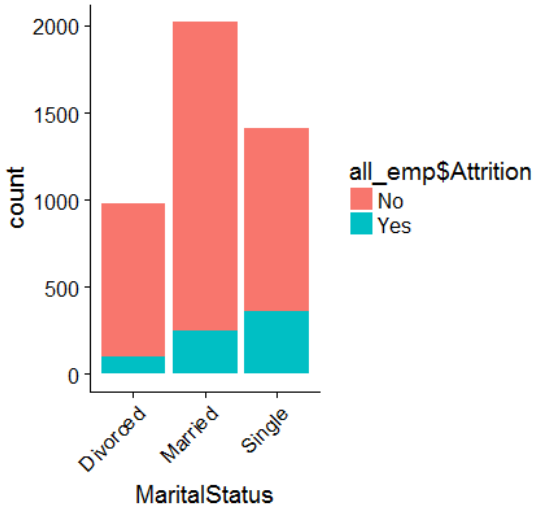


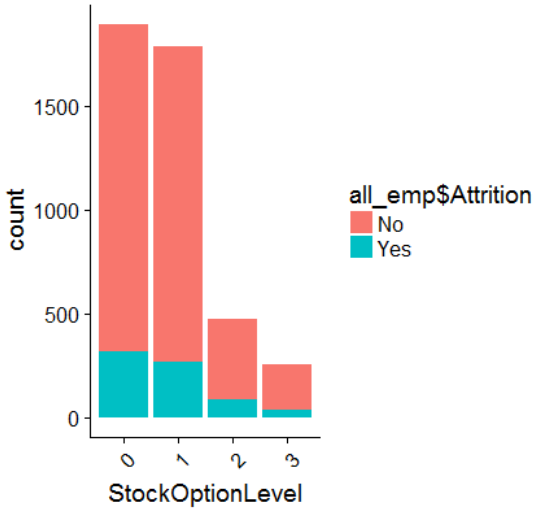


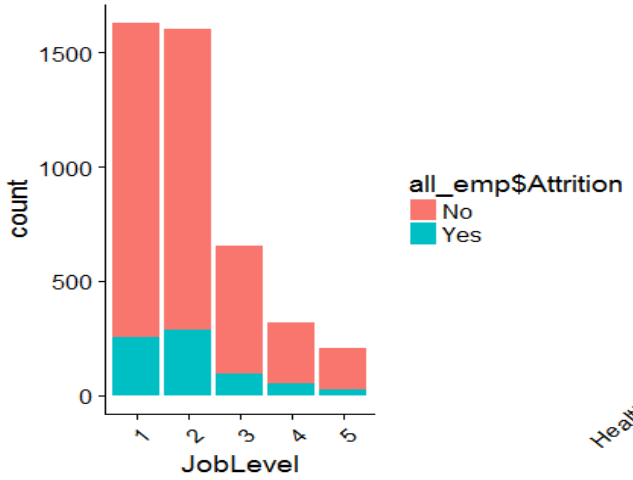
Data Understanding: EDA – Categorical variables

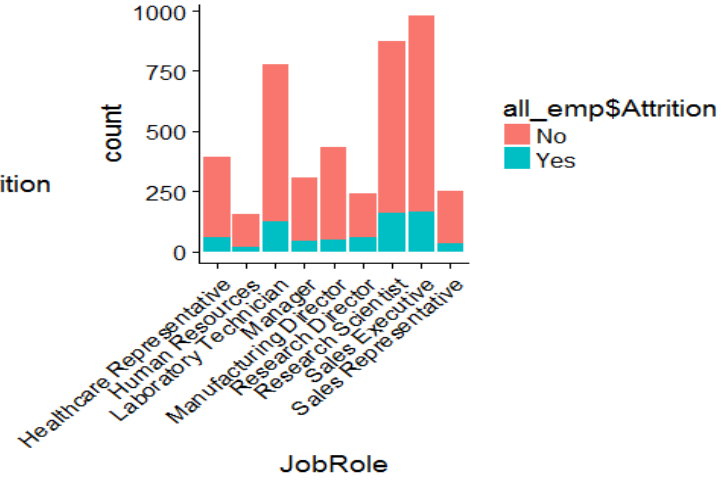


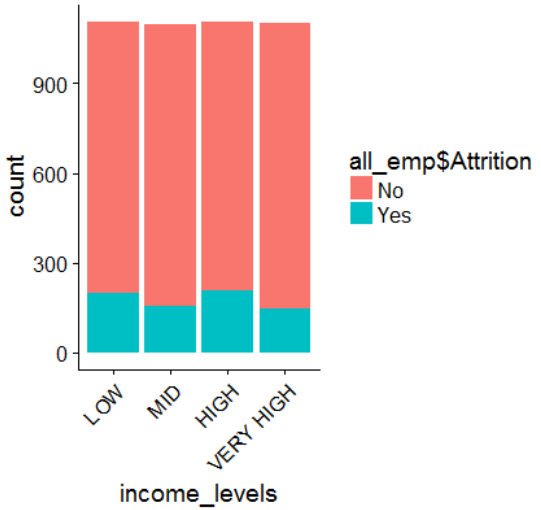












Data Preparation: Outlier treatment, Dummy variables

- Outlier treatment
  - 0.99 percentile of the continuous values will be retained.
- Dummy variable for the following categorical variables that include the following:
  - JobInvolvement, EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance, BusinessTravel, Department, Education, EducationField, JobLevel, JobRole, MaritalStatus, StockOptionLevel
- Scale Continuous variables

Data Modelling: Logistic Regression Model

After several Iterations we ended up a Logistic Regression Model with 11 variables

Null deviance: 2603.2 on 2910 degrees of freedom

Residual deviance: 2064.0 on 2899 degrees of freedom

AIC: 2088

Variable	Coefficient	Meaning
(Intercept)	-1.64549	
NumCompaniesWorked	0.30627	People who worked in more companies have higher attrition rate
TotalWorkingYears	-0.81038	Total number of years negatively affects attrition rate
YearsSinceLastPromotion	0.40471	Longer the time passed since last promotion higher the attrition
YearsWithCurrManager	-0.47331	People who have same manager for longer period of time have lower attrition rate
average_hrs	0.63305	People who spend longer time at work have higher attrition
EnvironmentSatisfaction3	-0.90375	People who voted Env satisfaction as 3 have lower attrition
EnvironmentSatisfaction4	-1.23052	People who voted Env satisfaction as 4 have significantly lower attrition
JobSatisfaction4	-0.70371	People who voted job satisfaction as 4 have lower attrition
EnvironmentSatisfaction2	-0.75958	People who voted Env satisfaction as 2 have lower attrition
BusinessTravelTravel_Frequently	0.82663	People who travel frequently have higher attrition
MaritalStatusSingle	1.00934	Those who are single tend to leave soon.

Model Evaluation

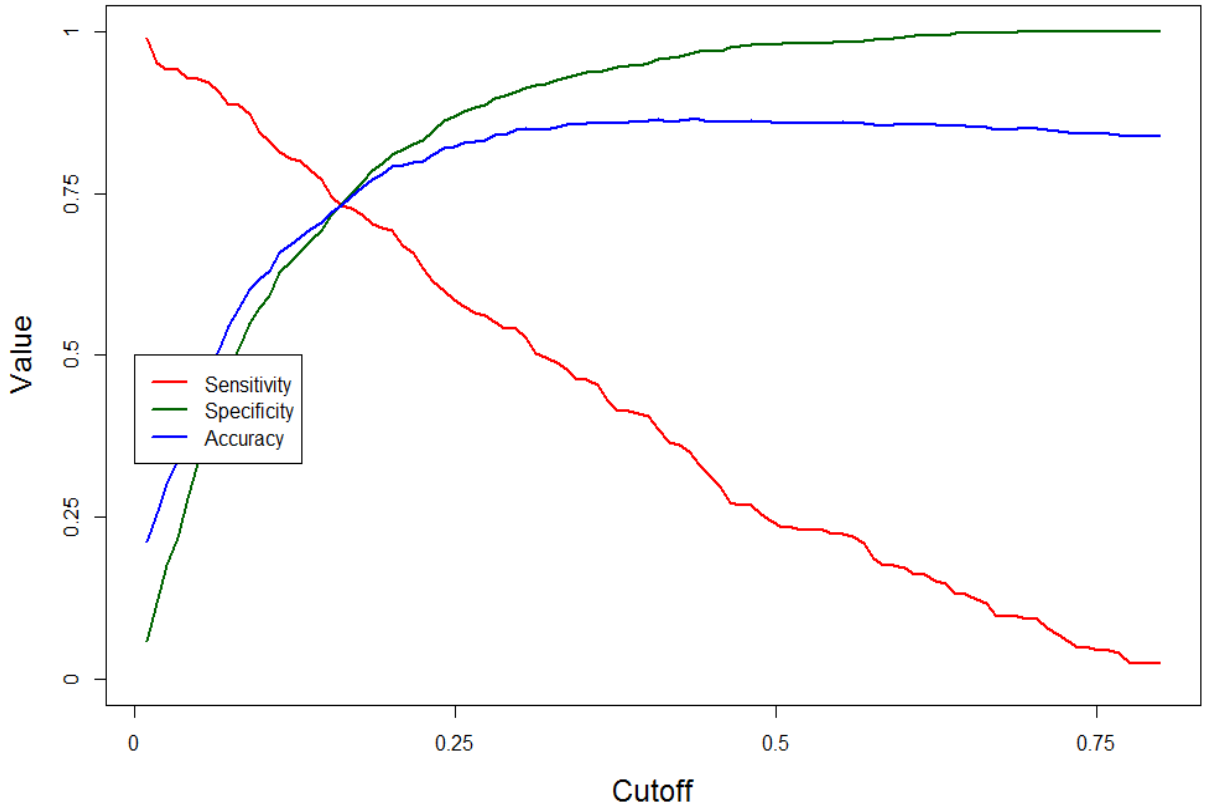
- Data was segregated as Test data (30%) and Training Data (70%)

Final Confusion Matrix at cutoff 0.16162

	Reference	
Prediction	No	Yes
No	763	55
Yes	279	150

Accuracy	**Sensitivity**	**Specificity**
0.7321572	0.7317073	0.7322457

KS Statistic
0.463953



Lift and Gain Charts

From the Decile table, Cumulative Lift and Gain Charts