HR Analytics Case study

Group Members:

1. Ajay Gupta
2. Shubham Joshi
3. Ashish Bhogawkar

Problem Statement

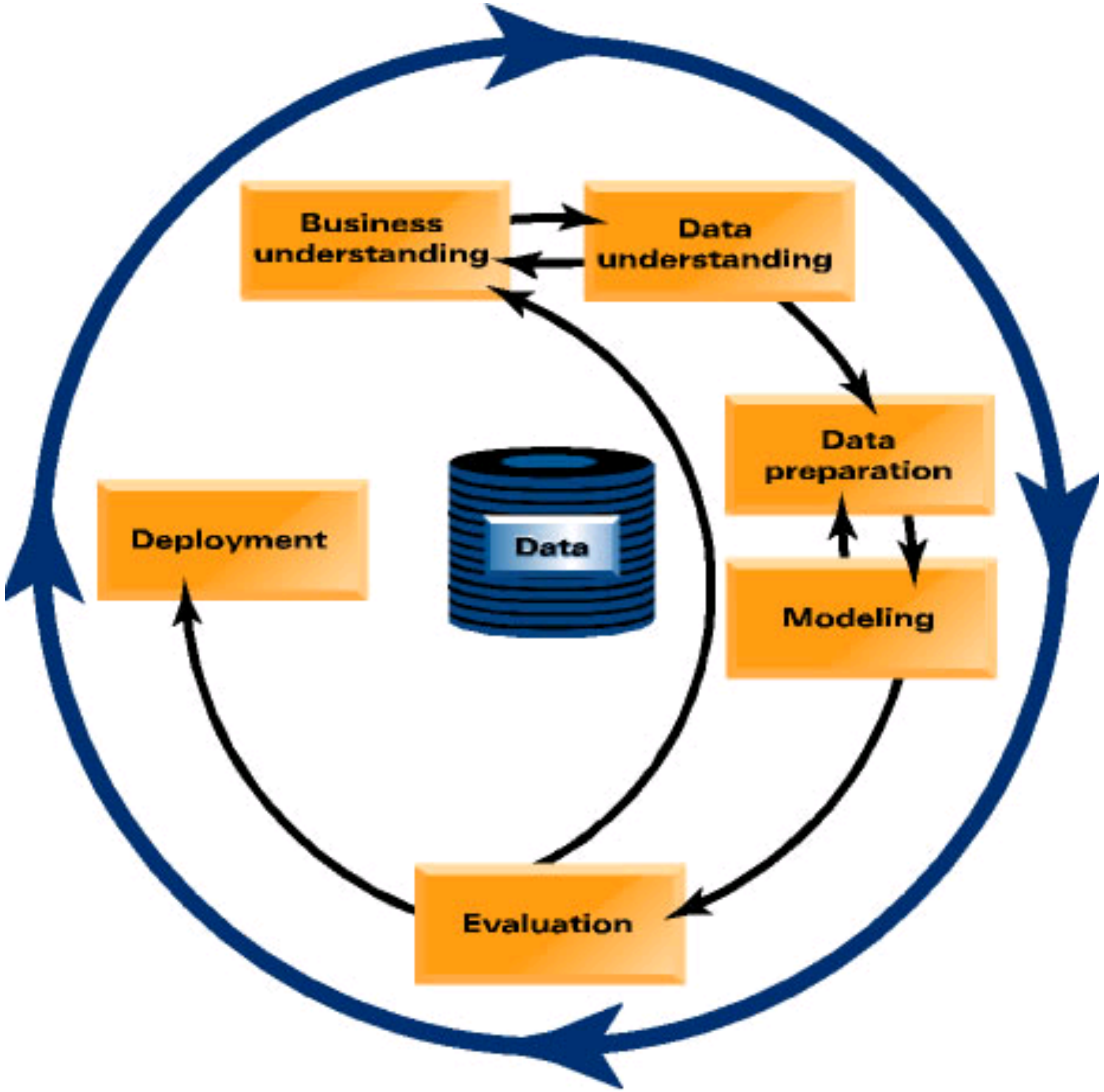
1. The company XYZ is suffering from attrition rate of 15 %.

2. XYZ wants to understand the different factors which are responsible for attrition .

3. By curbing these factors, XYZ wants to reduce attrition.

Analysis Approach

CRISP-dm (Cross Industry Standard Process For Data Mining) Methodology



CRISP-dm (Cross Industry Standard Process For Data Mining) Methodology

UpGrad

1. Business Objectives

- Determine and Evaluate Business Objectives
- Determine Data mining Goals

2. Data Study

- Collect Initial Data and Map Data Needs to sources
- Identify data gaps and mismatch
- Evaluate and suggest data source
- Explore type and quantity

3. Data Preparation

- Data Cleansing
- Filling up the missing values
- Data Formatting
- Labelling the data
- Converting Categorical data into numeric data
- Creating derived fields Analysis & Modeling
- Sampling

4. Modelling

- Select Modelling Techniques
- Generate Test Design
- Build Model
- Access the Model

4. Evaluation

- Select Modelling Techniques
- Build Model
- Generate Test Design
- Test Run on sample data

5.Reporting and Deployment

- Resource Allocation
- Periodic screening & scrutiny of model
- Reporting as based on Client Requirement

Business Understanding:-

Due to 15 % attrition rate, XYZ company is facing below mentioned challenges

1. As the employee’s left XYZ, their projects get delayed, which makes it difficult to meet timelines, resulting in a reputation loss among consumers and partners

2. XYZ has to maintain sizeable department for purpose of recruiting new talent

3. More often than not, the new employees have to be trained for the job and/or given time to acclimatize themselves to the company

Data Understanding:-

The data is given in 5 different tables which are as follows:-

Sr. No	Name	Description
1	general_data	This contains employee specific data regarding various attributes such as Age, Gender, Job Role, Education, Marital Status etc.
2	employee_survey_data	This gives employee specific data such as Worklife balance, Environment satisfaction and Job satisfaction
3	manager_survey_data	This gives employee specific data given by his/her manager. The data covers Job involvement and Performance rating
4	in_time	This covers in time of each employee and the levaes taken by him/her in 2015.
5	out_time	This covers out time of each employee and the levaes taken by him/her in 2015.

Data Preparation:-

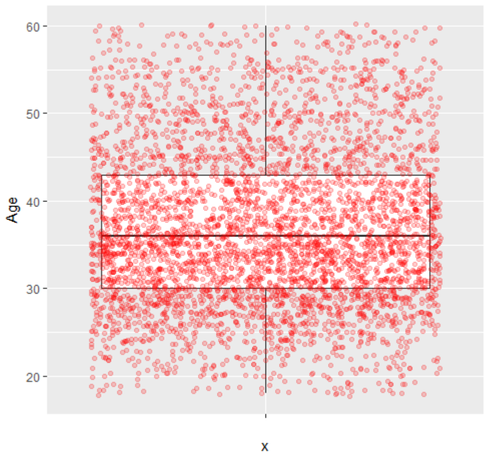
Following are the Data preparation operations Performed:-

Data preparation operations	Description
Data Cleaning:	Imputing the NA values with average values
Outlier Removal:	All the outlier values in numeric continuous variables have been removed
Data Formatting:	Converting dates in string to Date objects.
Data Merging:	All the data sets which are : general_data, employee_survey_data, manager_survey_data, in_time, out_time have been merged to obtain a single data fram HR_data
Exploratory Data Analysis:	Exploratory Data Analysis of Quantitative and Categorical Variables have been done, where **correlation matrix** and **barographs** have been derived to be referred while modelling phase.
Derived Metrics:	From in_time and out_time we have derived 2 variables which are:- 1] **weekly_avg_wrk_hrs :** Average weekly working hours of the employee 2] **leavesTaken_yearly :** Leaves taken by the employee in the year

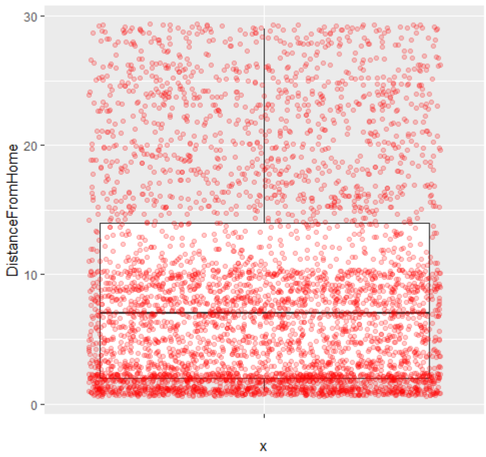
Exploratory Data Analysis – Quantitative Variables:-

The below are the box plot observed after removing the outliers:-

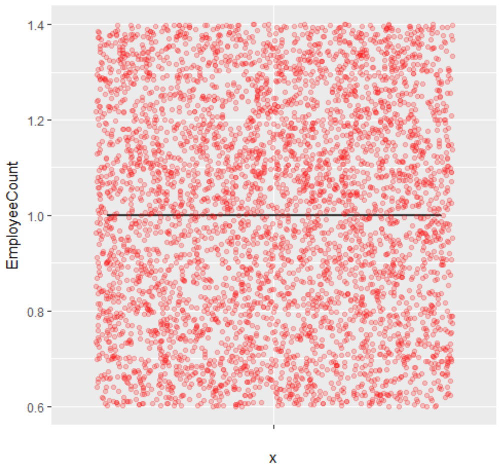
Box plot for Age to check for outliers



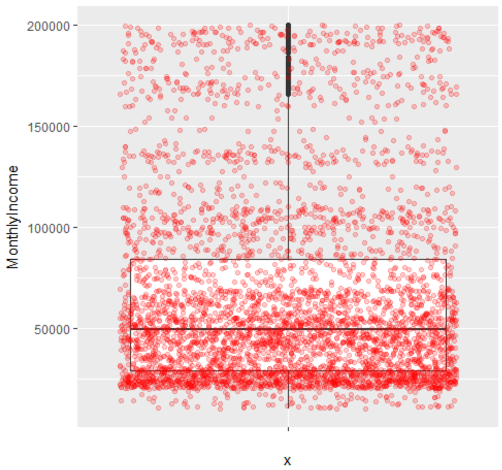
Box plot for DistanceFromHome to check for outliers



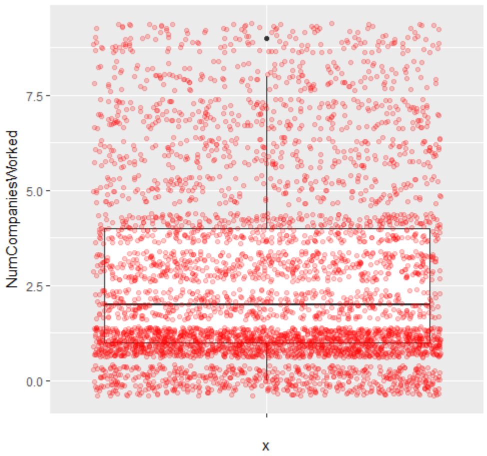
Box plot for EmployeeCount to check for outliers



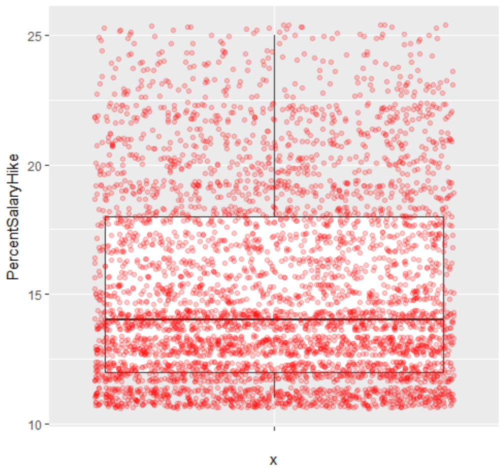
Box plot for MonthlyIncome to check for outliers



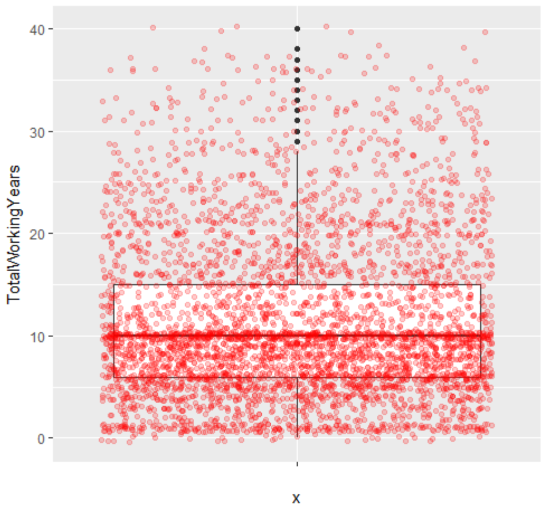
Box plot for NumCompaniesWorked to check for outliers



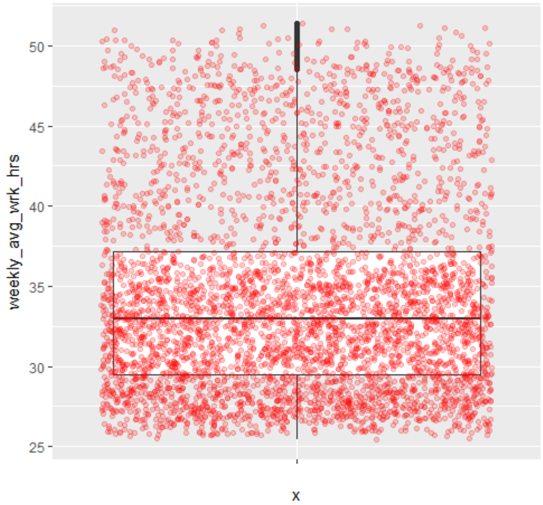
Box plot for PercentSalaryHike to check for outliers



Box plot for TotalWorkingYears to check for outliers



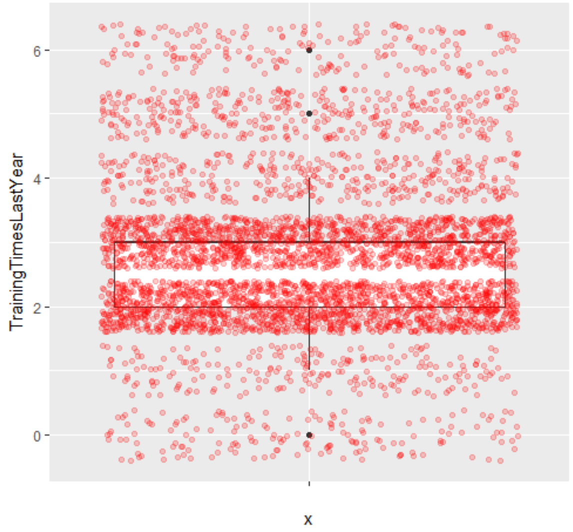
Box plot for weekly_avg_wrk_hrs to check for outliers



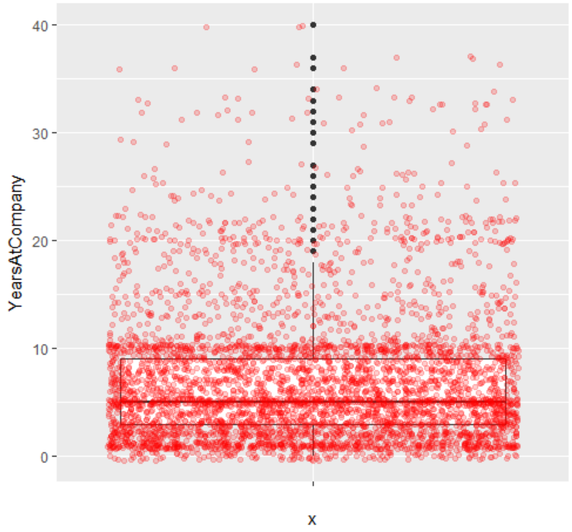
Exploratory Data Analysis – Quantitative Variables:-

The below are the box plot observed after removing the outliers:-

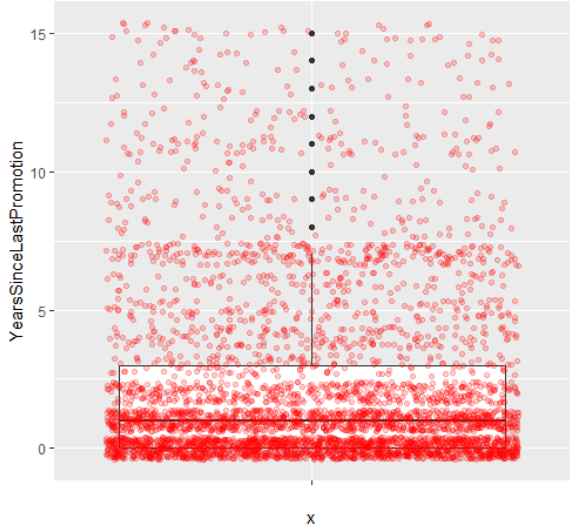
Box plot for TrainingTimesLastYear to check for outliers



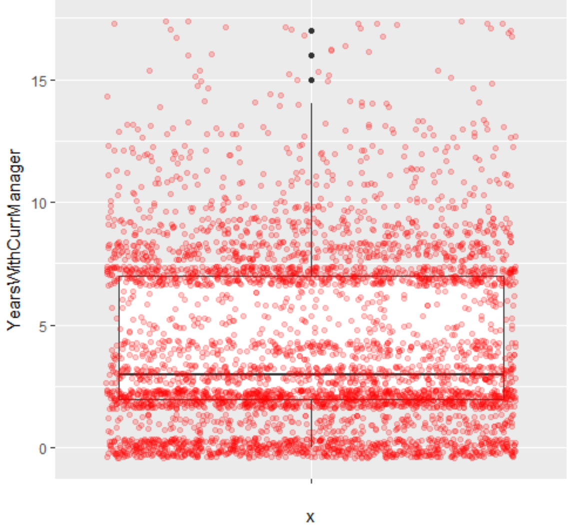
Box plot for YearsAtCompany to check for outliers



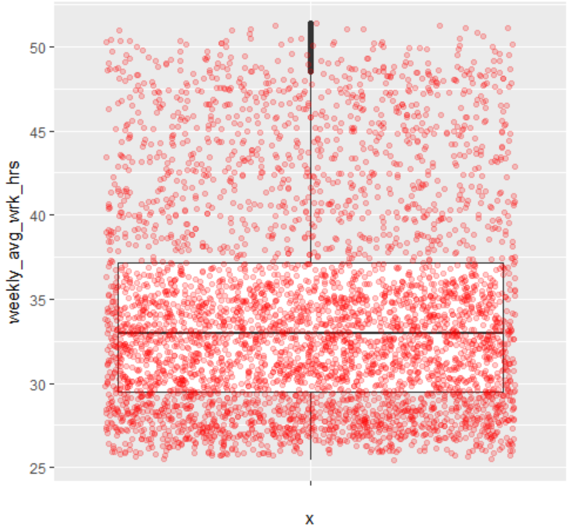
Box plot for YearsSinceLastPromotion to check for outlie



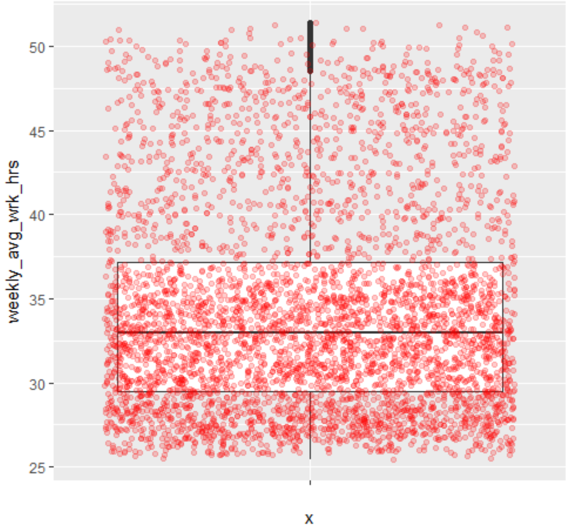
Box plot for YearsWithCurrManager to check for outliers



Box plot for weekly_avg_wrk_hrs to check for outliers

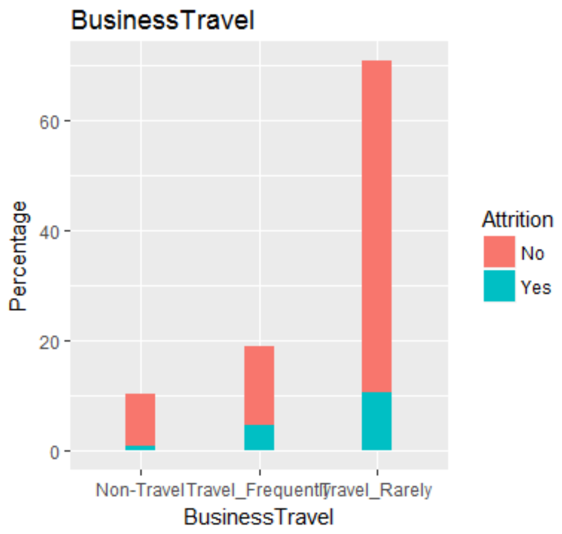


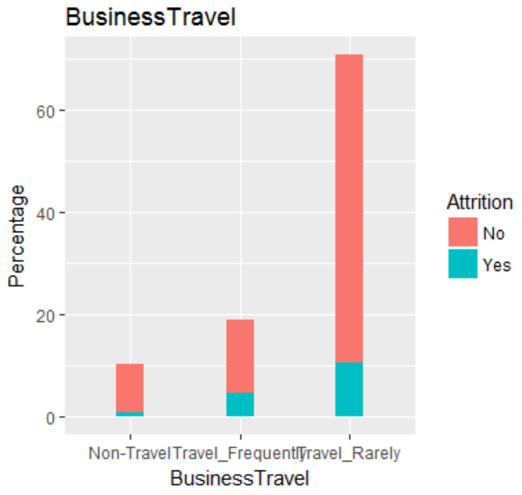
Box plot for weekly_avg_wrk_hrs to check for outliers

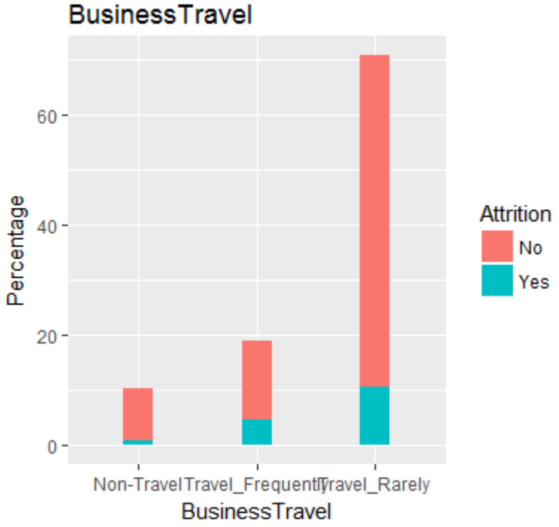


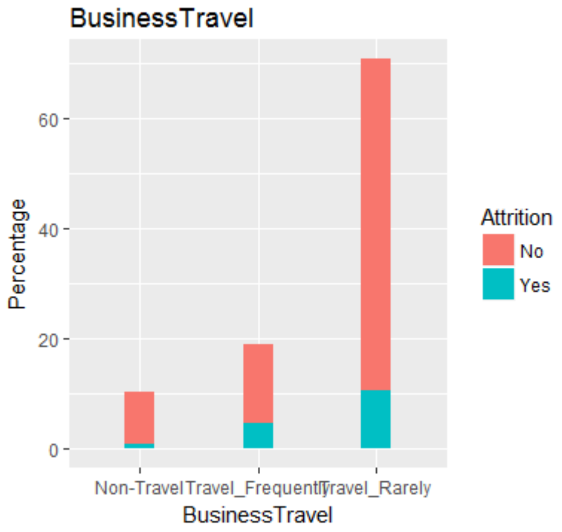
Exploratory Data Analysis – Categorical Variables:-

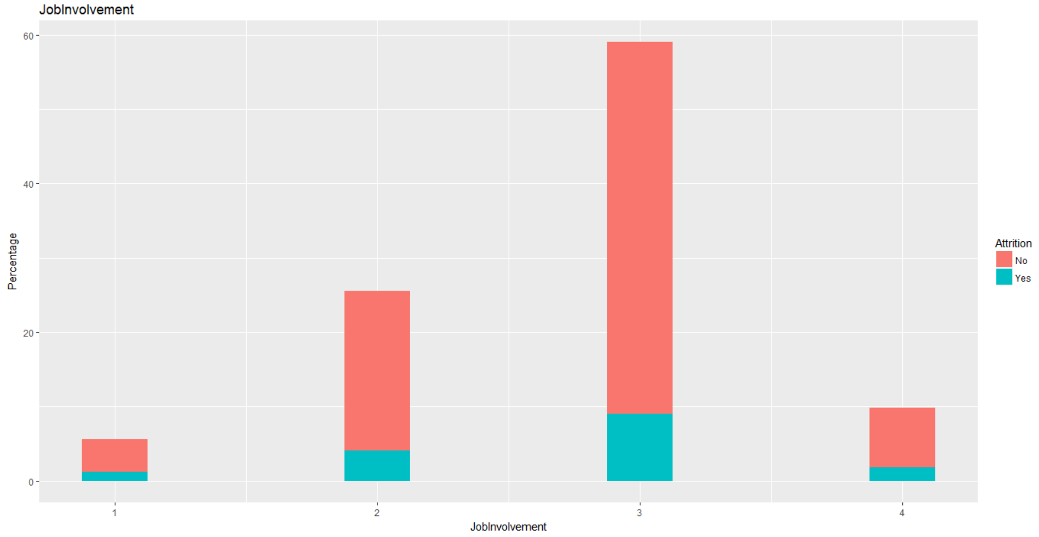
The following are the barographs obtained as a part of EDA of Categorical variables :-

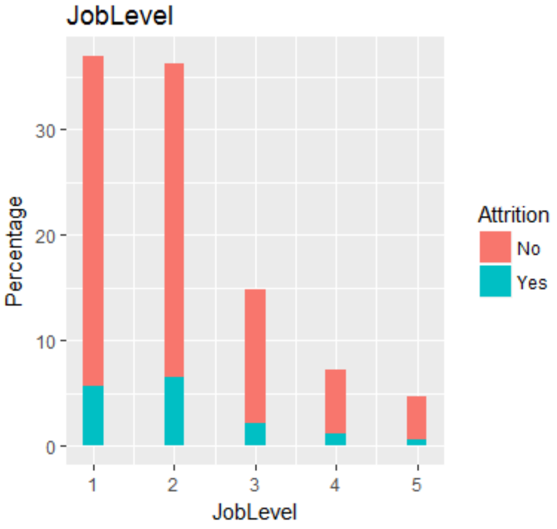




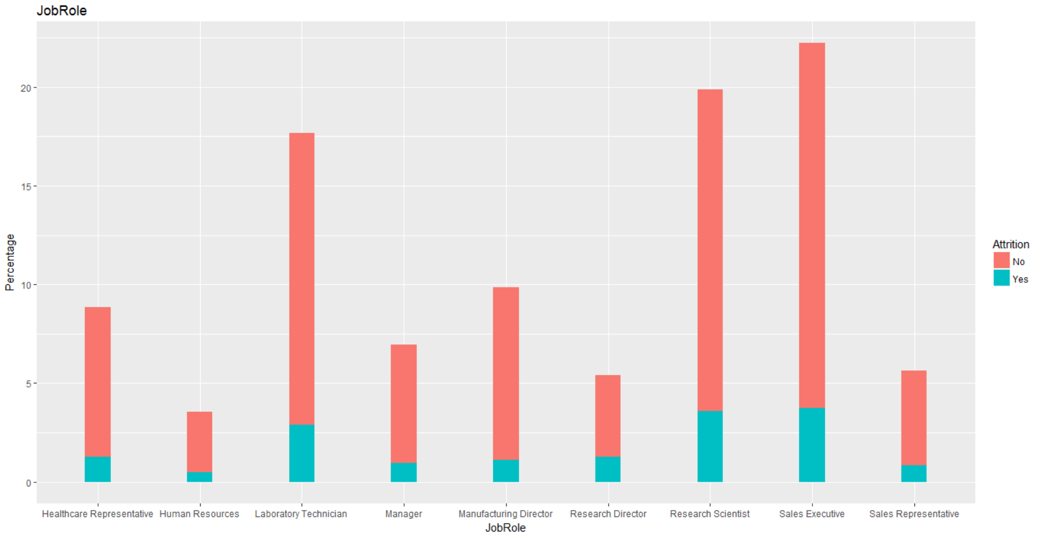


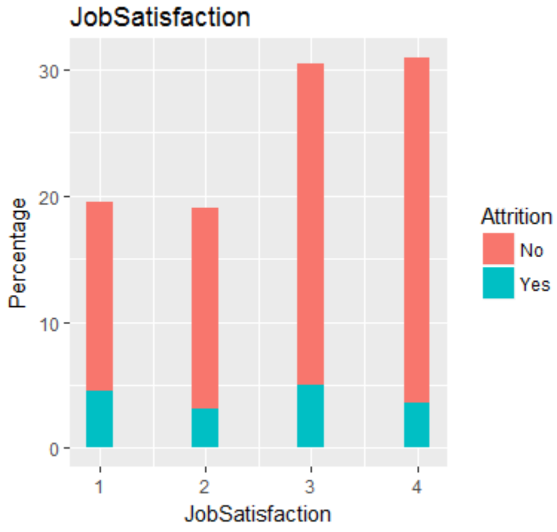


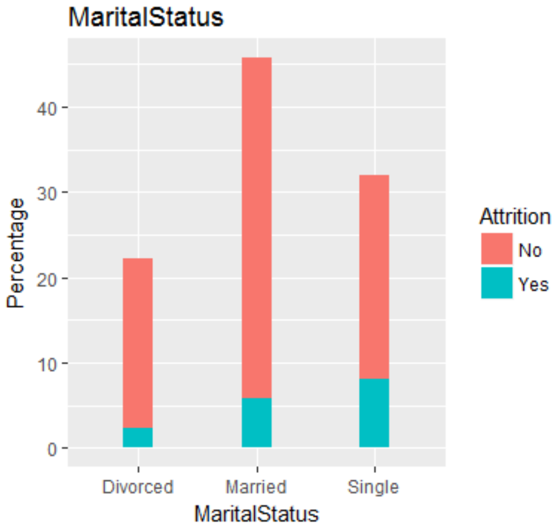


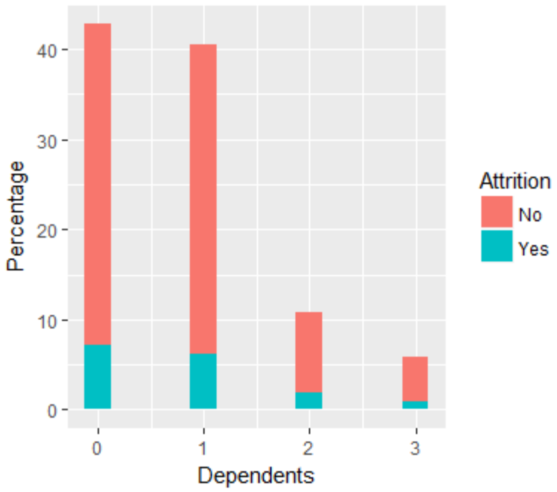


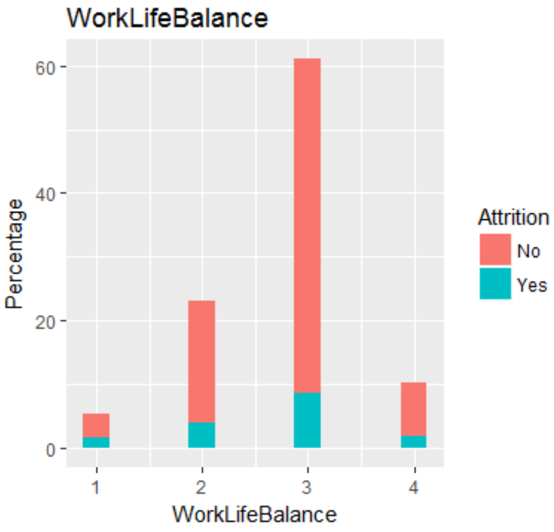
Exploratory Data Analysis – Categorical Variables Continued:-

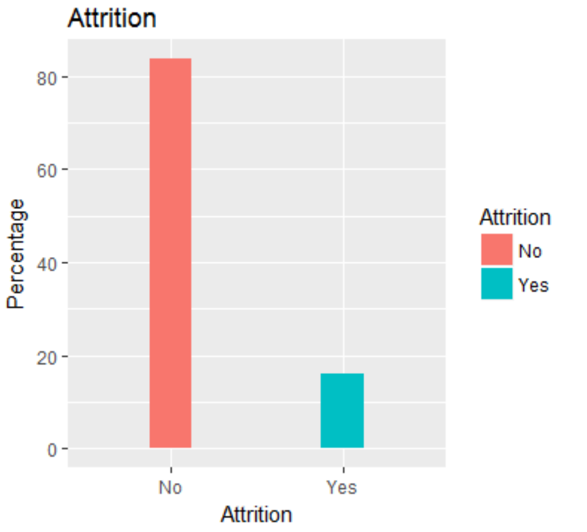




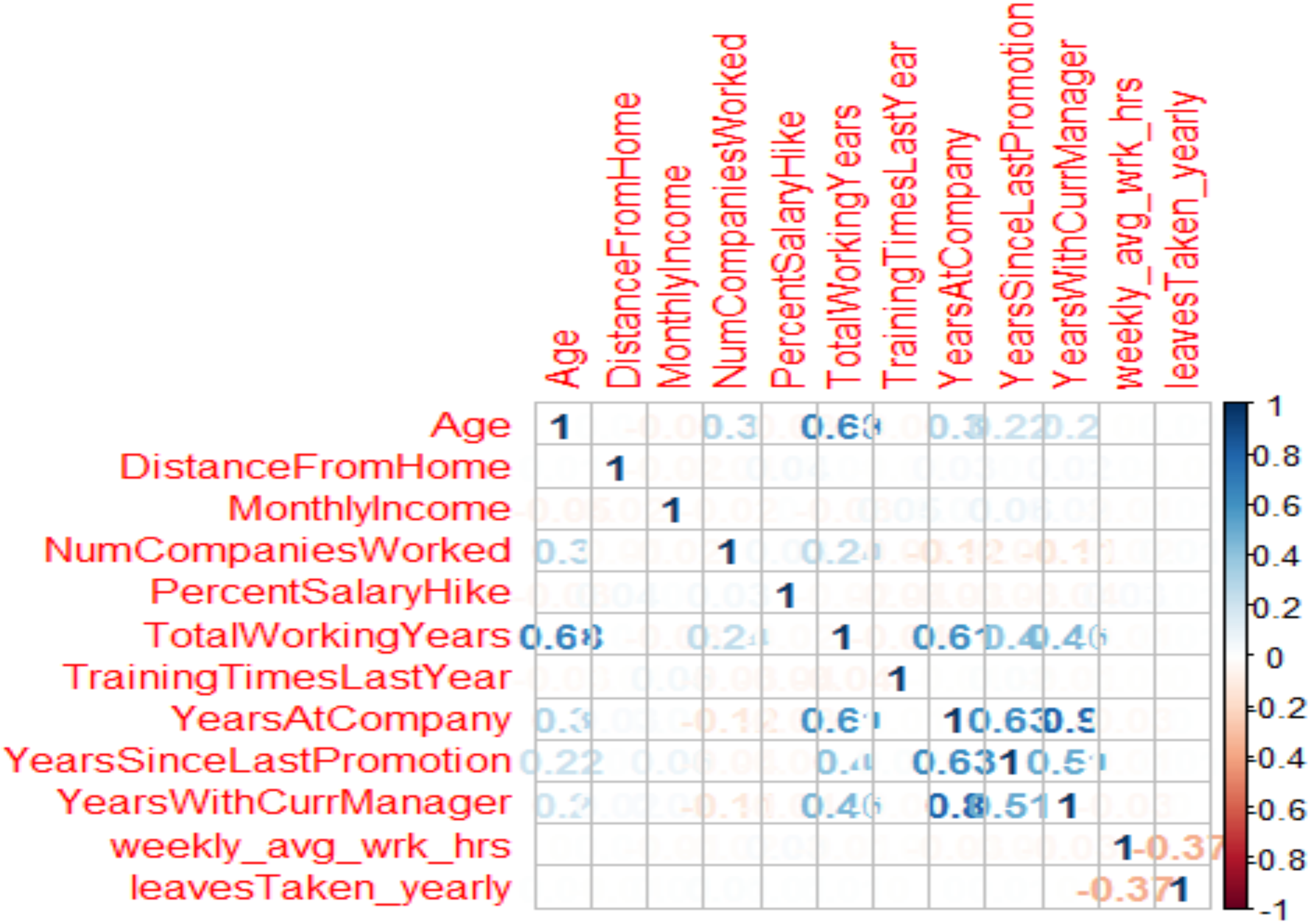








Correlation Matrix between continuous Numeric variables:-



Generating Logistic Regression Model:-

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables. Logistic Regression is a Model using which we can estimate the class probability of values of Dependent variable base on the values of Independent Variables. It determines the presence of a risk factor increases the odds of a given outcome by a specific factor

Logistic regression is used where the dependent variable is dichotomy (i.e. can be divided into 2 categories)

1] For single Independent variable :-

$Ln(p/(p-1))= ax + b$

$p=1/(a+e^{(ax+b)})$

2] For Multiple Independent variable :-

$ln(p/(p-1))= a1x + a2y + a3z \dots . + b$

$p=1/(a+e^{-(a1x + a2y + a3z \dots . + b)})$

Where p is the probability of occurrence of an event.

Logistic Regression Model Details:-

The following are the significant fields and their coefficients obtained as a part of our model:-

Sr. No	Beta	Beta Value	X Value	Description
1	β0	-1.72621	X0	NA
2	β1	-0.26469	X1	Age
3	β2	-0.17815	X2	MonthlyIncome
4	β3	0.31887	X3	NumCompaniesWorked
5	β4	-0.58832	X4	TotalWorkingYears
6	β5	-0.18848	X5	TrainingTimesLastYear
7	β6	0.56757	X6	YearsSinceLastPromotion
8	β7	-0.49543	X7	YearsWithCurrManager
9	β8	-0.37806	X8	EnvironmentSatisfaction
10	β9	-0.40317	X9	JobSatisfaction
11	β10	-0.23053	X10	WorkLifeBalance
12	β11	0.63448	X11	weekly_avg_wrk_hrs

Logistic Regression Model Details continued..:-

Sr. No	Beta	Beta Value	X Value	Description
13	β12	0.92523	X12	BusinessTravelTravel_Frequently
14	β13	-1.07066	X13	DepartmentResearch...Development
15	β14	-1.19188	X14	DepartmentSales
16	β15	0.68412	X15	JobRoleResearch.Director
17	β16	1.04101	X16	MaritalStatusSingle

The above mentiond are the factors which may affect an employee to come to a conclusion to leave the company.

Model Equation = $\Sigma\ \beta iXi$ (i= 0 to 16)

Business Implication - Out of given all variables only mentioned 16 variables have key role in Attrition of any employee in XYZ company. These variable should be controlled to reduce attrition.

Model Evaluation

Model Evaluation – Using Confusion Matrix approach:-

A confusion matrix shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data. The matrix is NxN, where N is the number of target values (classes).

Confusion Matrix		**Actual**			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy = (a+d)/(a+b+c+d)**	
		a/(a+c)	d/(b+d)		

Accuracy : the proportion of the total number of predictions that were correct.
Positive Predictive Value or **Precision** : the proportion of positive cases that were correctly identified.
Negative Predictive Value : the proportion of negative cases that were correctly identified.
Sensitivity or **Recall** : the proportion of actual positive cases which are correctly identified.
Specificity : the proportion of actual negative cases which are correctly identified.

Confusion Matrix approach for our Model:-

Model Evaluation

UpGrad

Confusion Matrix		**Actual**			
		Yes(churn)	No (Non -Churn)		
Prediction	Yes(churn)	51	39	*Positive Predictive Value*	0.56667
	No (Non - Churn)	161	1066	*Negative Predictive Value*	0.86879
		Sensitivity	*Specificity*	**Accuracy =0.8481**	
		0.24057	0.96471		

Model Evaluation using KS-Statistics

The following table is obtained after calculating all the attributes of KS-Statistics :-

deciles	total	Attrition count	Cumulative Attrition	Gain Attrition(% cumulative Attrition)	Lift	Non-Attrition count	Cumulative Non-Attrition	Gain Non-Attrition (% Cumulative Non-Attrition)	(KS statistics) Gain Attrition-Gain NonAttrition
1	132	62	62	29.2453	2.92453	70	70	6.3348416	22.9104
2	132	53	115	54.2453	2.71226	79	149	13.484163	40.7611
3	132	34	149	70.283	2.34277	98	247	22.352941	47.9301
4	131	12	161	75.9434	1.89858	119	366	33.122172	42.8212
5	132	10	171	80.6604	1.61321	122	488	44.162896	36.4975
6	132	15	186	87.7358	1.46226	117	605	54.751131	32.9847
7	131	5	191	90.0943	1.28706	126	731	66.153846	23.9405
8	132	8	199	93.8679	1.17335	124	855	77.375566	16.4924
9	132	8	207	97.6415	1.08491	124	979	88.597285	9.04422
10	131	5	212	100	1	126	1105	100	0

The Ks-Statistics occurs at 47.93 at 3rd decile which covers 70.283% of total population which will churn. The same has been red highlighted in the above diagram.

Thus the HR can predict 70% of the total population accurately in top 3 deciles If he Uses this model.

Thank You