HR Analytics Case Study

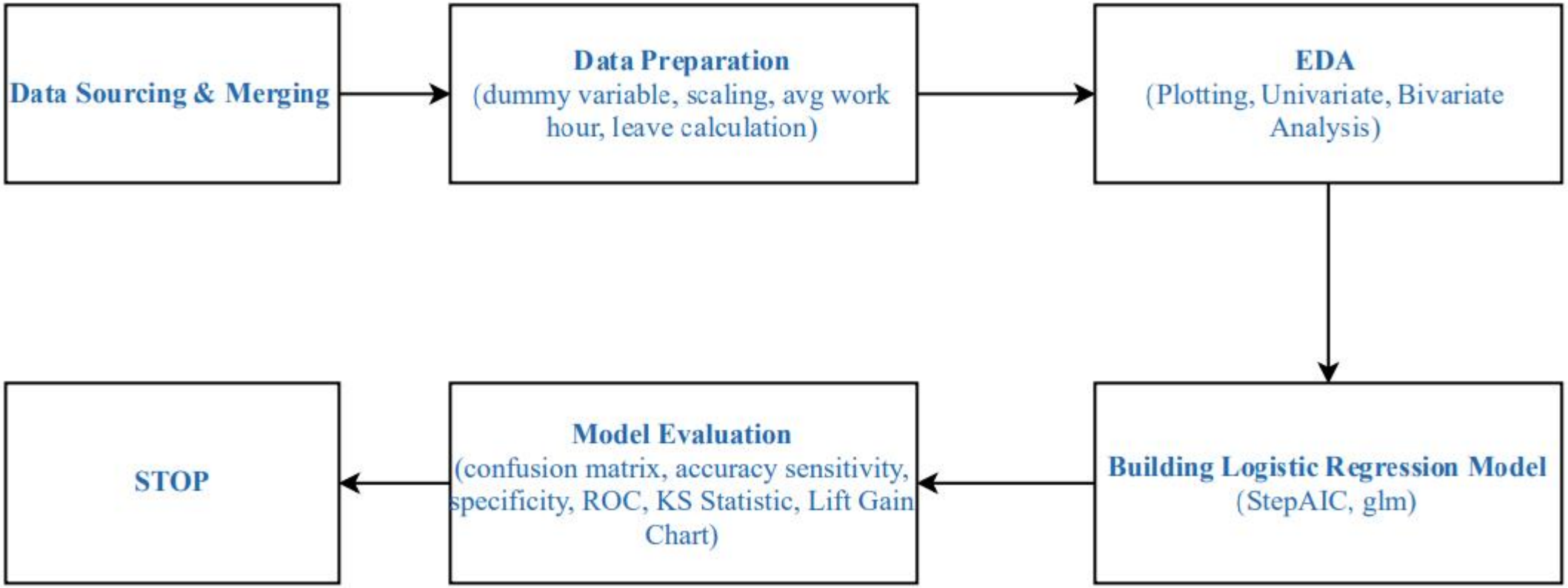
Group Name:

1. Mohit Rathore
2. Mahesh Kulkarni
3. Abdulrahiman Bangara
4. Abhinav Marathe

Abstract

We found factors which are causing attrition of 16% in XYZ, in order to curb attrition. We also suggest what changes they should make to their workplace, in order to get most of their employees to stay. Also, which variables are most important and needs to be addressed right away. To solve this problem we preprocessed data about employees, their performance rating, employee rating & their work hour patterns. From this data we found probability of attrition using logistic regression & prepared a model which also predict churn with Sensitivity of 80%. Model was also evaluated using accuracy, specificity, ROC, KS Statistic, Lift & Gain charts. Using stepwise logistic regression we also found 16 key variables which are highly significant in causing churn.

Problem solving methodology



Analysis: Data Preparation & EDA

1. emp_survey, manager_survey & general_data was combined together in hr dataframe by merging on EmployeeID.
2. We calculated employee average work hours using in_time & out_time data.
3. We also calculated number of leaves taken by employee in past year using this data.
4. Finally leaves & average work hours data was merged with master hr dataset to form a single data frame with all details.
5. Leaves per month are also analysed separately to check absences patterns across the year among 2 kinds of employees.
6. Initial Univariate & Bivariate analysis shows that AverageWorkHours of those who left were higher.
7. EnvironmentSatisfaction, JobSatisfaction was low in employees who left, WorkLifeBalance was almost same for both.
8. Age, NoCompaniesWorked, JobLevel, Percentagehike seems to affect attrition.

Analysis: Building Logistic Regression Model

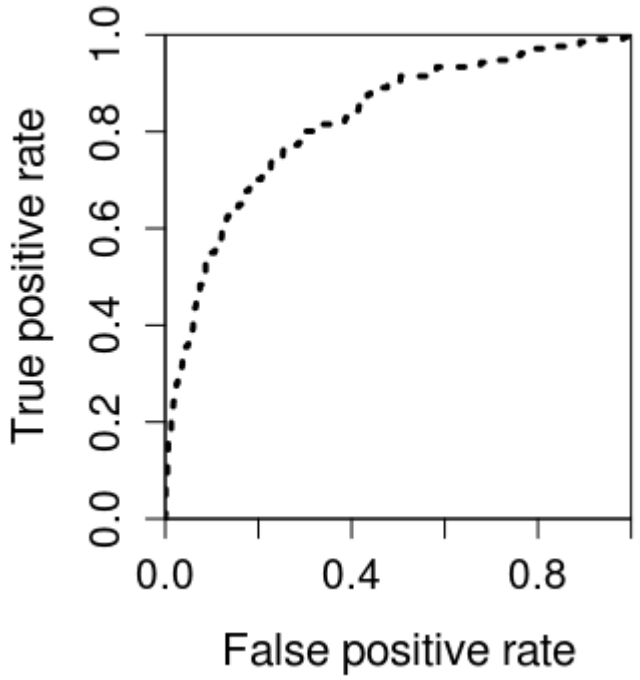
1. Before building logistic regression model dummy variables for categorical variables were created, total 43 variables were generated after this.
2. All continuous variables were normalized using scale function before building model.
3. Features irrelevant for model were removed(Employee ID, Over18).
4. Data was divided in even train test split using split function.
5. An initial logistic regression model was created using glm function which involved all 43 variables.
6. Second model was created using stepAIC method which suggested 28 relevant variables based on automated stepwise regression using p values.
7. Subsequent model were created using manual backward selection method, eliminating variables based on their VIF & p-value.
8. Eventually we stopped 13th Model with 16 highly significant variables remaining. We did further evaluation & prediction using this final model.

Analysis: Model Evaluation

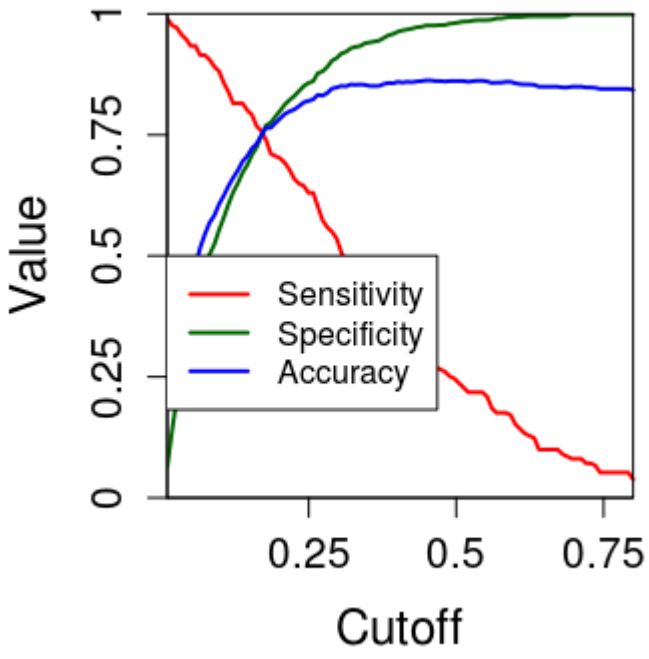
Final model was evaluated using various matrices explained below:

1. **Confusion Matrix:** This matrix helped describe performance of model for both positive & negative cases. Key measures found were Accuracy(72%), Sensitivity(80%) & Specificity(70.2%). Detailed confusion matrix with all statistical results is given on later slide.
2. **Optimal Probability Cut-off:** Initial model was build with default cut-off of 0.5 which gave Accuracy(72%), Sensitivity(24.6%), Specificity(98%) but since we want a model which optimally predict both positive & negative class we derived a probability cut-off chart to get optimal cut-off of ~.15 with focus on Sensitivity.
3. **KS –statistic:** KS statistic is an indicator of how well your model discriminates between the two classes, our model had KS statistic of 50.3% in 3rd decile.
4. **Lift & Gain Chart:** Cumulative gains and lift charts are visual aids for measuring model performance, same was plotted for model along with random & perfect model Gain-Lift chart; our model with decent gain of 80% till 4th Decile & gain of 90% by 6th Decile**.**
5. **ROC Curve:** Receiver Operating Characteristicis a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. Area under ROC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one; For our final model AOC was is 0.822.

Results: ROC & Probability Cut-off Chart



a) ROC curve showing relation between TPR & FPR; Area under curve is **0.82**



b) Probability cutoff chart showing sensitivity specificity tradeoff; Intersecting point after **~.15** was chosen for optimal model

Results: Confusion Matrix and Statistics

Confusion Matrix:

Prediction	No	Yes
No	762	42
Yes	323	169

Accuracy : 0.7184

95% CI : (0.693, 0.7427)

No Information Rate : 0.8372

P-Value [Acc > NIR] : 1

Kappa : 0.3276

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.8009
Specificity : 0.7023

Pos Pred Value : 0.3435
Neg Pred Value : 0.9478

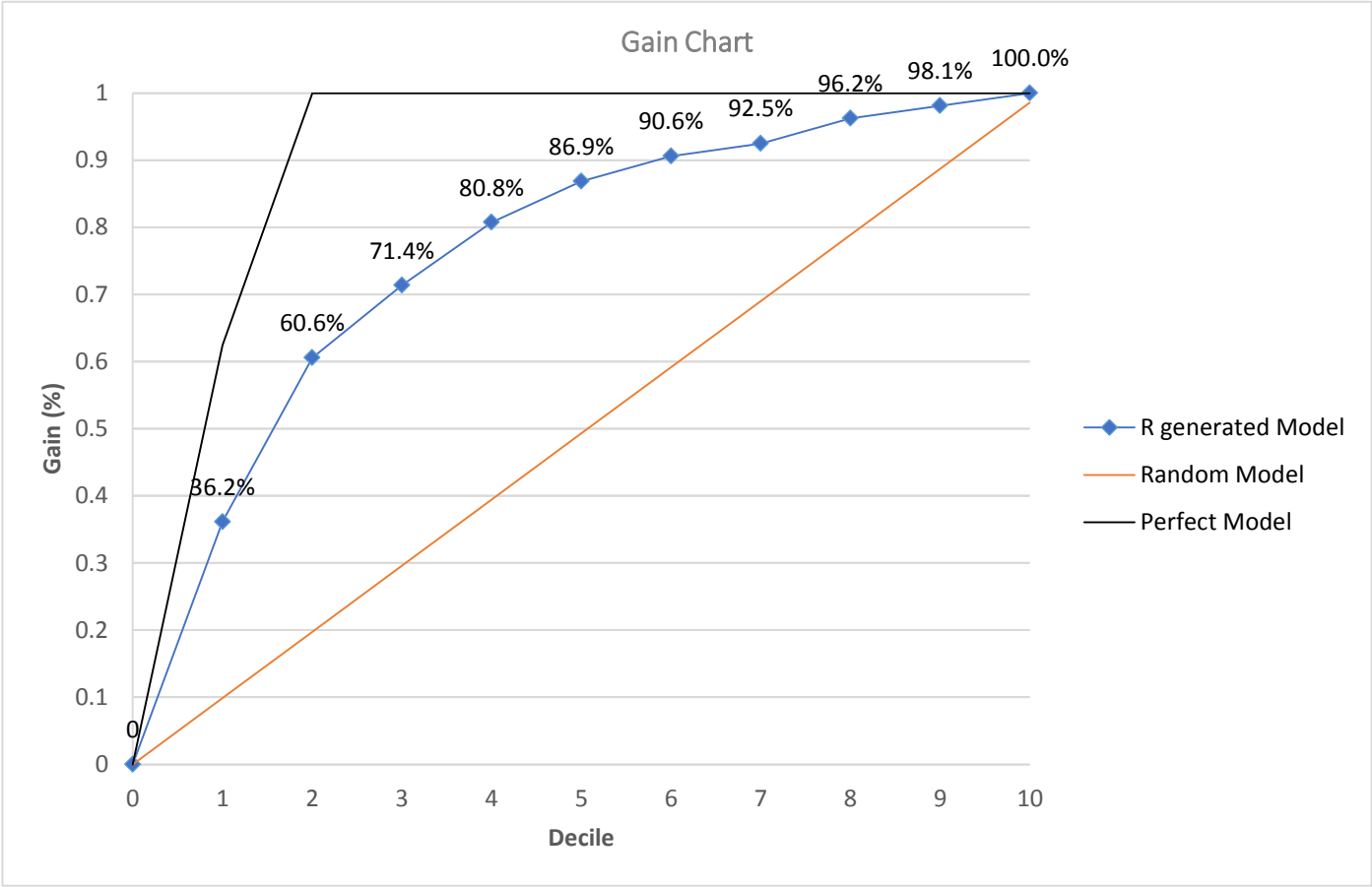
Prevalence : 0.1628
Detection Rate : 0.1304
Detection Prevalence : 0.3796

Balanced Accuracy : 0.7516

'Positive' Class : Yes

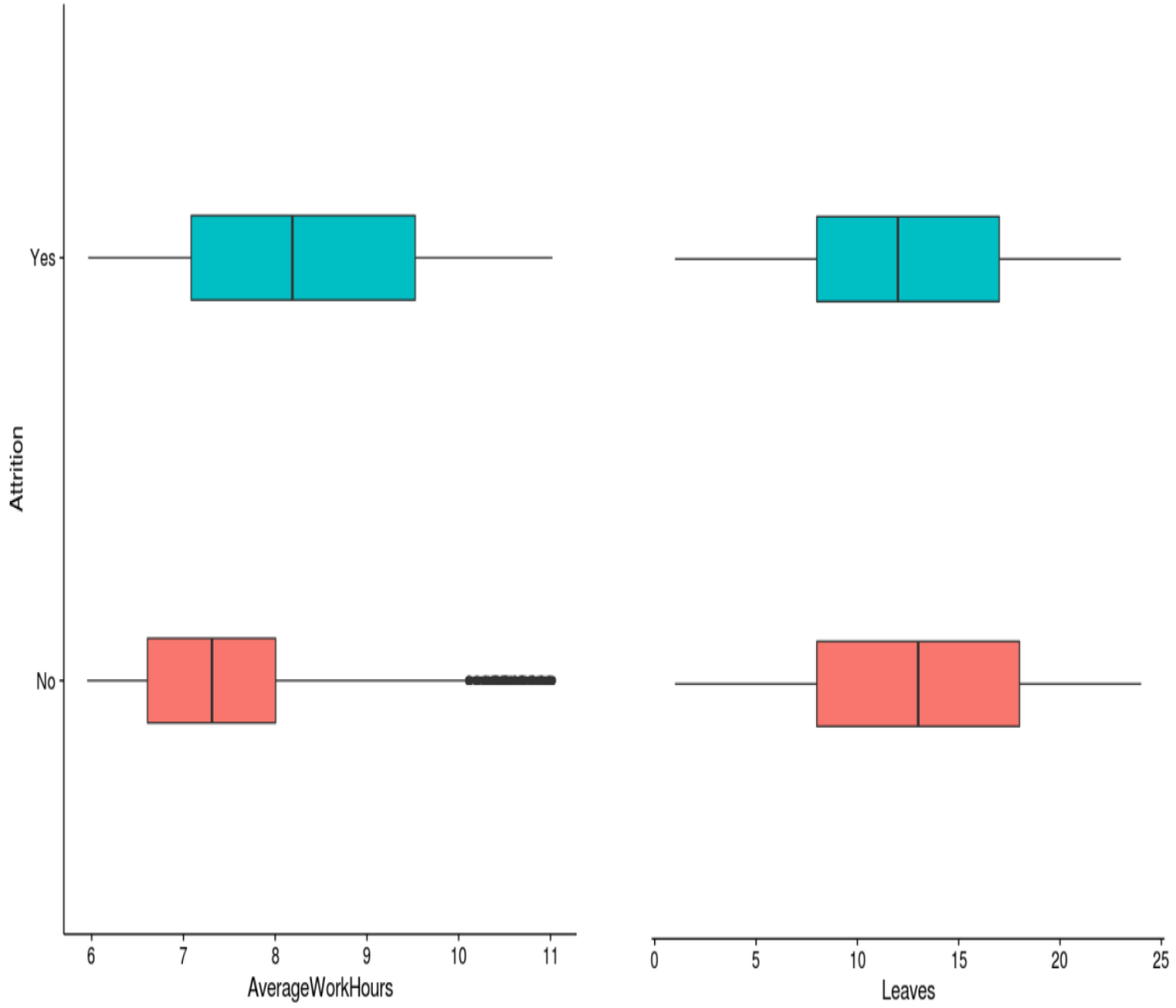
Results: Lift & Gain Chart

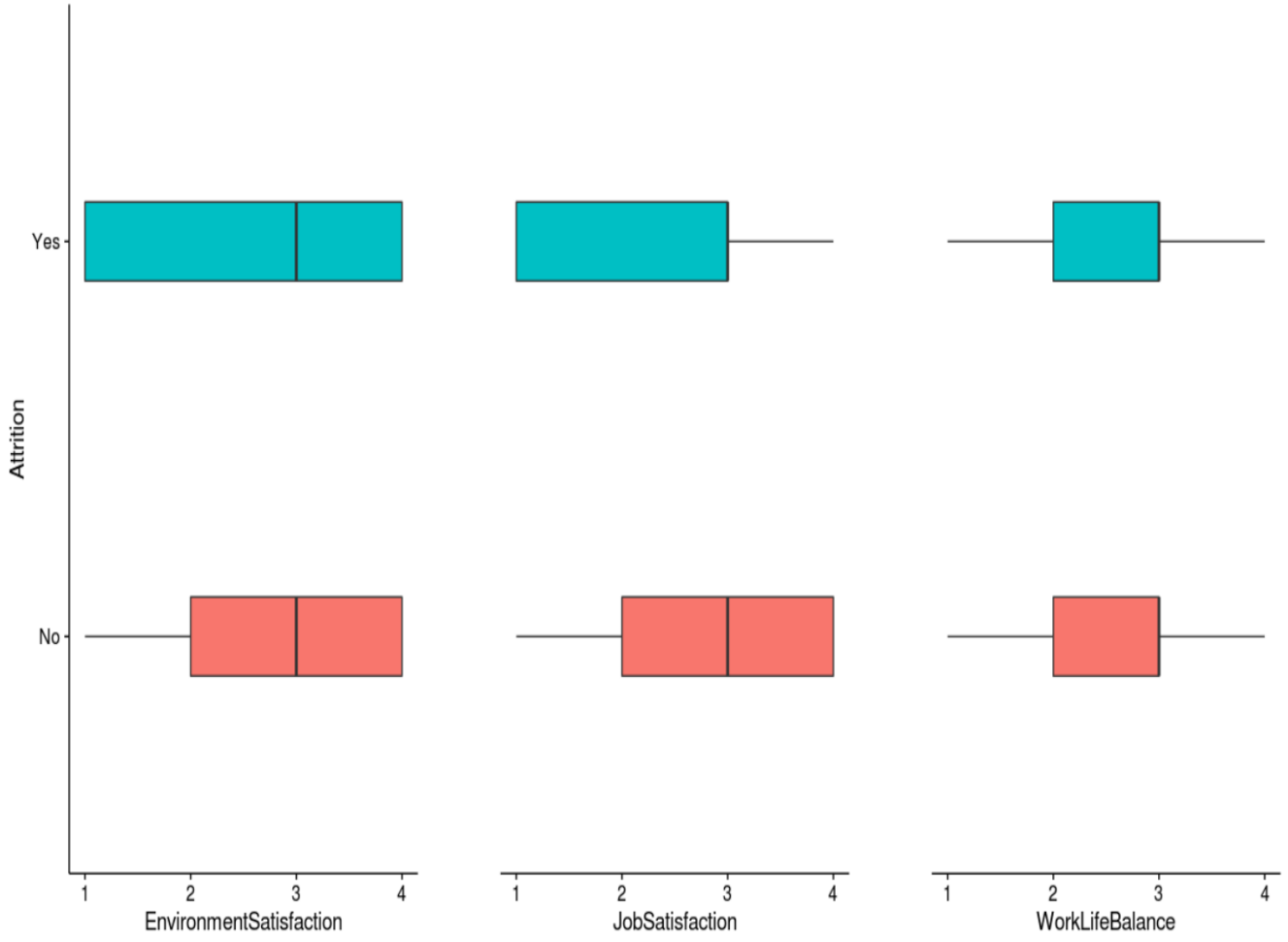
Bucket	Total	Attrition	CumAttrition	Gain	Cum_Lift
1	133	77	77	36.15023	3.6150235
2	132	52	129	60.56338	3.0281690
3	132	23	152	71.36150	2.3787167
4	133	20	172	80.75117	2.0187793
5	132	13	185	86.85446	1.7370892
6	132	8	193	90.61033	1.5101721
7	133	4	197	92.48826	1.3212609
8	132	8	205	96.24413	1.2030516
9	132	4	209	98.12207	1.0902452
10	105	2	211	99.06103	0.9906103
11	27	2	213	100.00000	NA



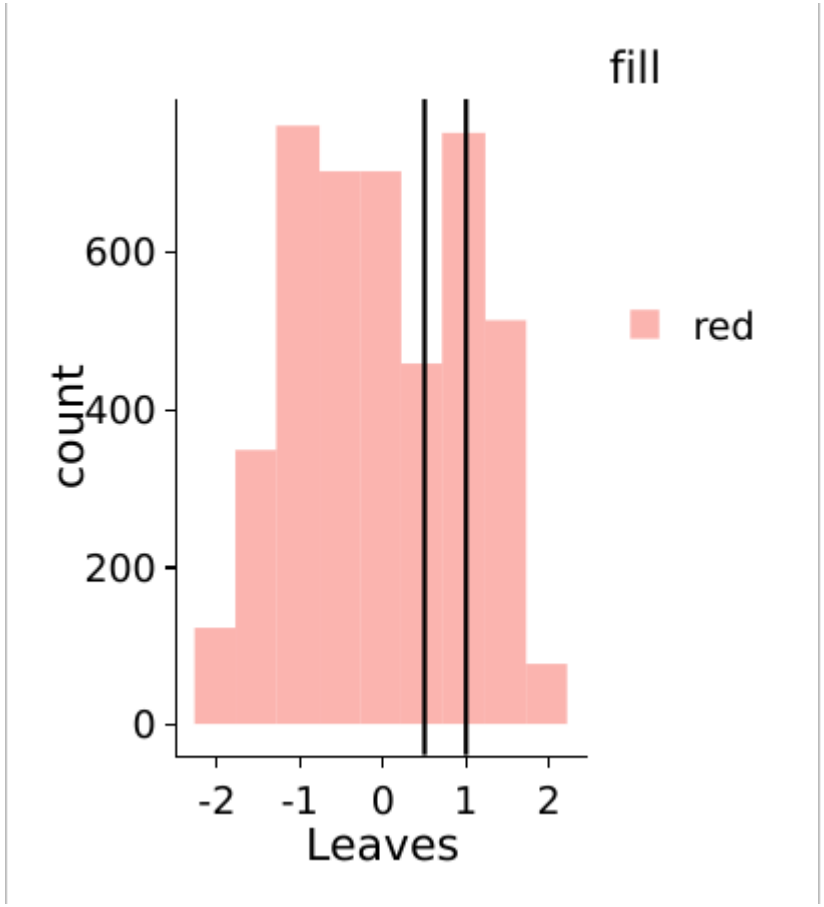
Lift & Gain chart with decent gain of 80% till 4th Decile & gain of 90% by 6th Decile

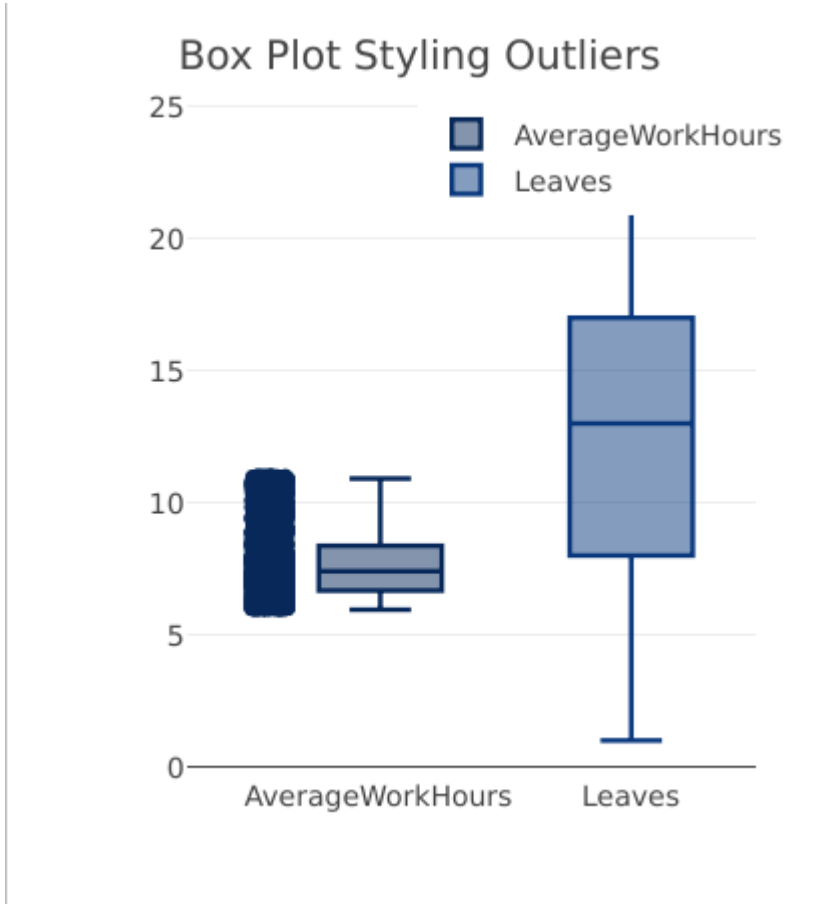
Results: EDA: Boxplots

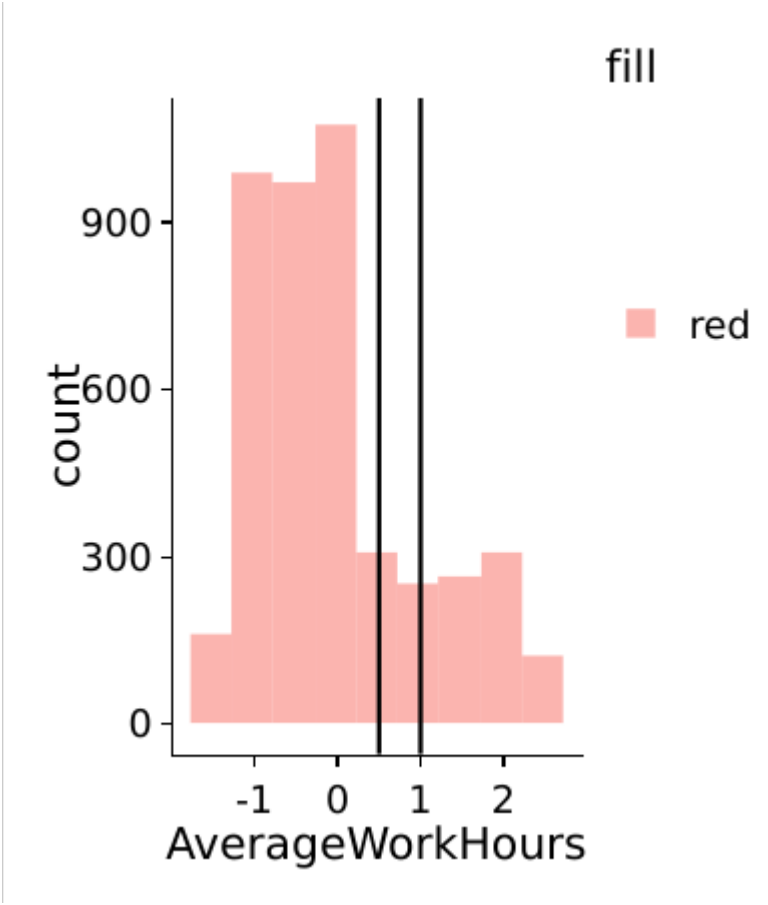




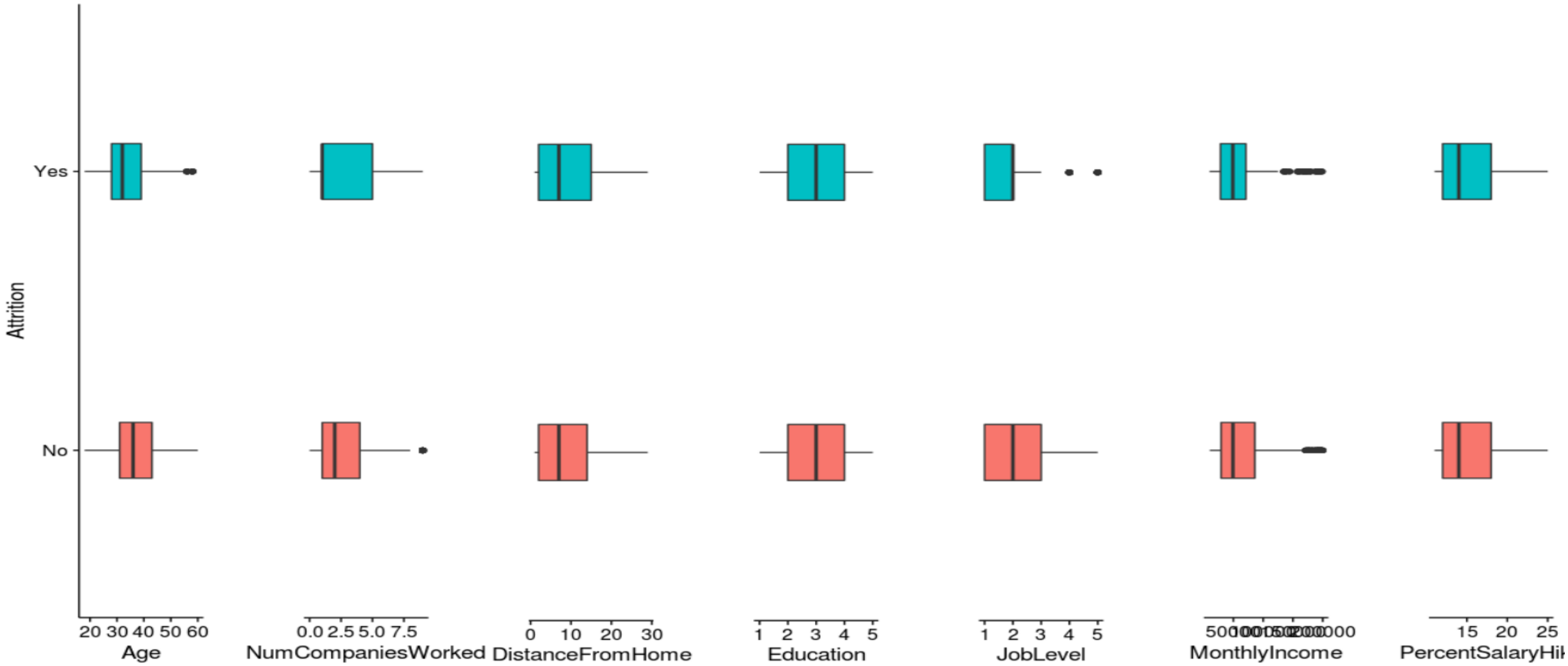
Results: EDA: Interactive Plots for derived variables



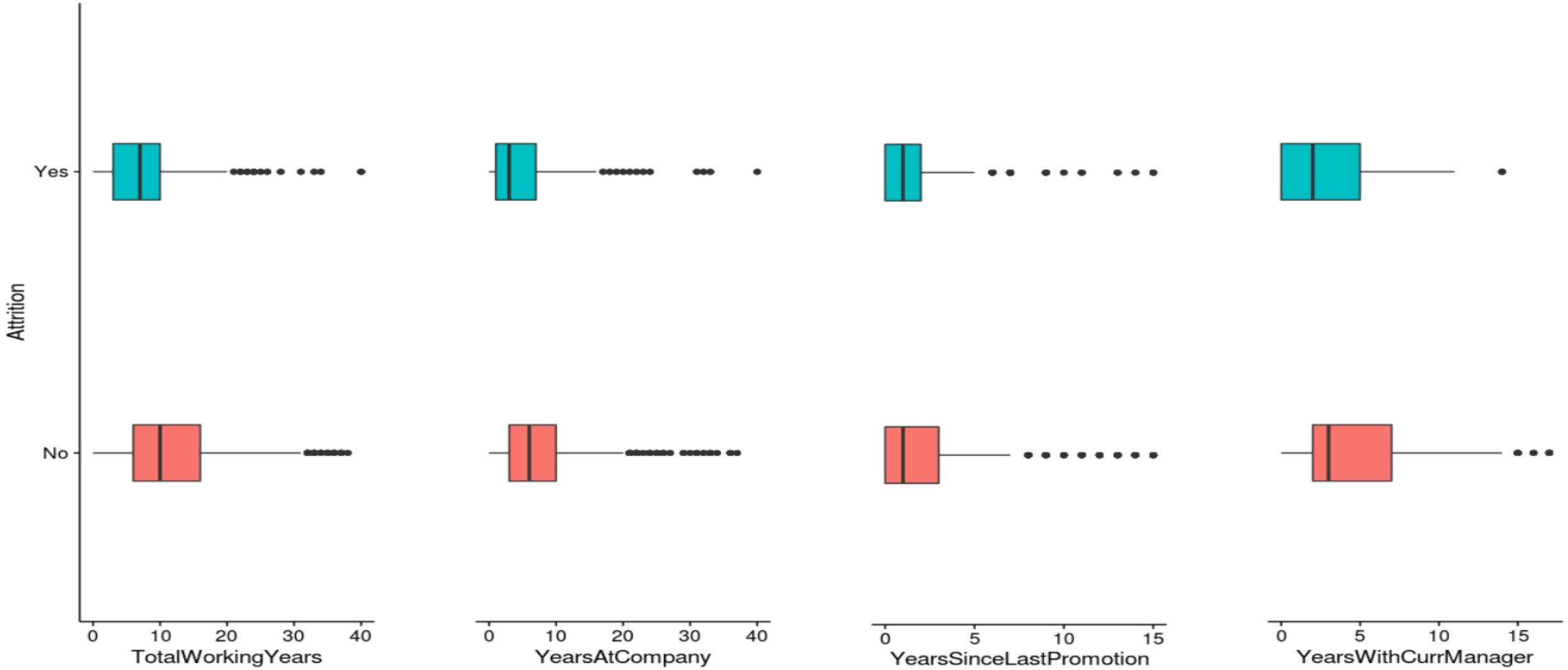




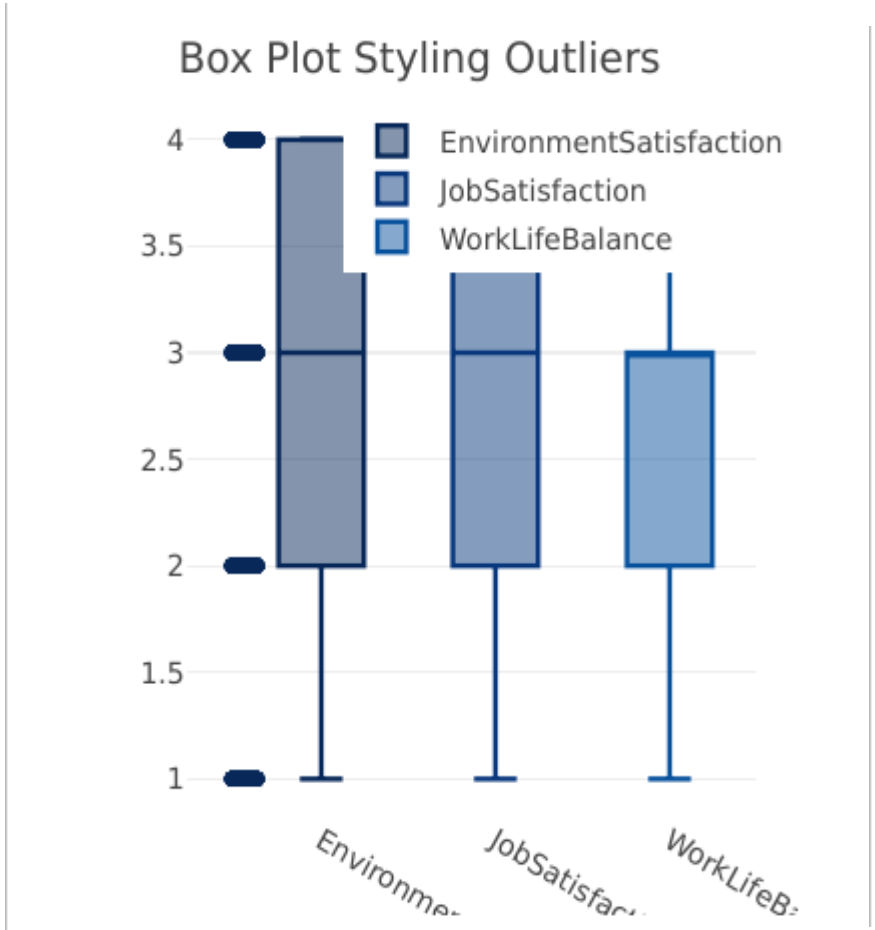
Results: EDA: Boxplots – General attributes vs. Attrition

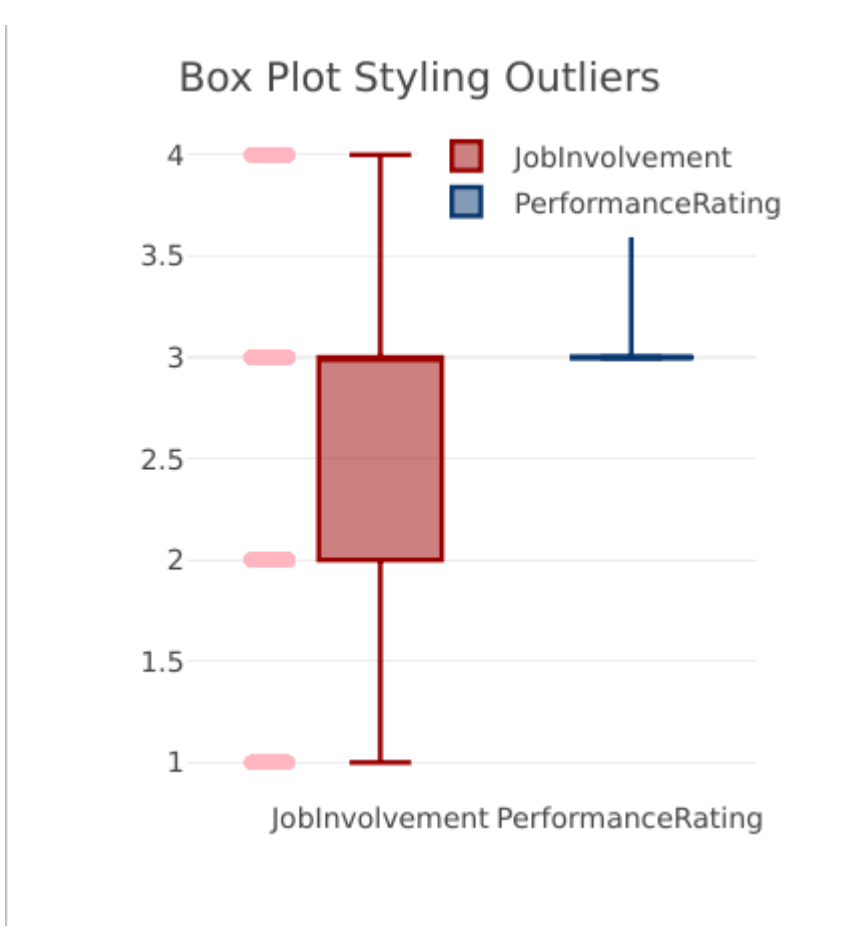


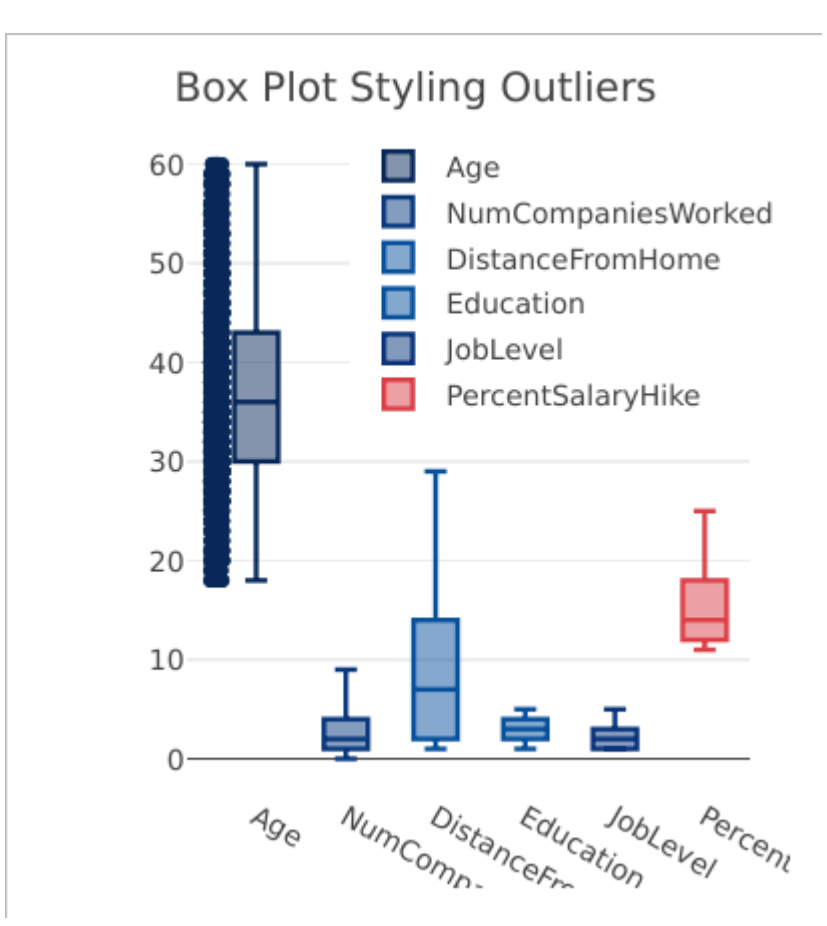
Results: EDA: Boxplots - General attributes vs. Attrition



Results: EDA: Interactive Plots using plotly







Conclusions

1. Based on logistic regression 16 out of 43 variables were found highly significant for attrition, these were: *TrainingTimesLastYear, JobRoleManufacturing.Director, WorkLifeBalance, JobSatisfaction, EnvironmentSatisfaction, MaritalStatusSingle, AverageWorkHours, NumCompaniesWorked, YearsWithCurrManager, Age, YearsSinceLastPromotion, TotalWorkingYears, BusinessTravelTravel_Rarely, BusinessTravelTravel_Frequently, DepartmentSales, DepartmentResearch...Development.*
2. Optimal Logistic Regression model can predict Employees who are likely to leave organization based on above 16 attributes correctly 80 percept of times.
3. As we suspected during EDA factors like *WorkLifeBalance*, *JobSatisfaction*, *EnvironmentSatisfaction*, *AverageWorkHours*, *NumCompaniesWorked*, *Age* does affect attrition as found by the model along with other factors like *BuisnessTravelFrequency*, *DepartmentR&D*, *JobRoleManufacturing etc.* These hidden aspects were only revealed after through analysis & prediction using logistic regression model.
4. XYZ can now focus on these 16 significant factors found by model to control attrition.