Analyzing Employee Attrition at XYZ

Vikas Jangra
Anjali Nair
Sudha Choudhary
Vernon Fernandes

Attrition a huge concern at XYZ

- 15% annual attrition
- Leading to project delays & missing deadlines
- Loss of name / reputation in competitive business environment
- In consistency in maintaining minimum headcount
- Time to train new employees to get on the active floor hence impacting both productivity & billing

Objective

- To understand the factors for attrition at XYZ
- To highlight factors adding to the attrition problem
- Call attention to key focus areas that needs immediate responsiveness
- Recommendation to reduce attrition at XYZ

What data do we have

- 5 Data Sets

File Name	Description	Number of Records	Remarks
General_Data	Employee details with status of attrition	4410	24 fields with EmployeeID as the primary key
Employee_Survey_Data	Gives us an indication of the work environment index	4410	3 fields with EmployeeID as the primary key
Manager_Survey_Data	Talk about Job involvement & performance ratings	4410	2 fields with EmployeeID as the primary key
In_Time	In timestamp of all employees	4410	Data made available from 1st Jan to 31st Dec 2015
Out_Time	Out timestamp of all employees	4410	Data made available from 1st Jan to 31st Dec 2015

Our Approach

Data Preparation

Data Cleaning

Data Standardization

Data OneView

Exploratory Data Analysis

Univariate Analysis to identify lead indicators if any

Modelling

Identifying Predictor and Response Variables

Splitting Data in to Training and Test Data set randomly

Developing the model based on training set

Fine Tuning

Fine Tuning the model

Selection of most appropriate model based on algorithm and Statistical criterion

Prediction from test set

Recommendation

Assessment of Quality of Prediction of Response Variable

Final Recommendations

Data Preparation

- Total Employee Data: 4410
- Date Period: 1st Jan to 31st December
- Count of Attrition: 771 (16.12%)

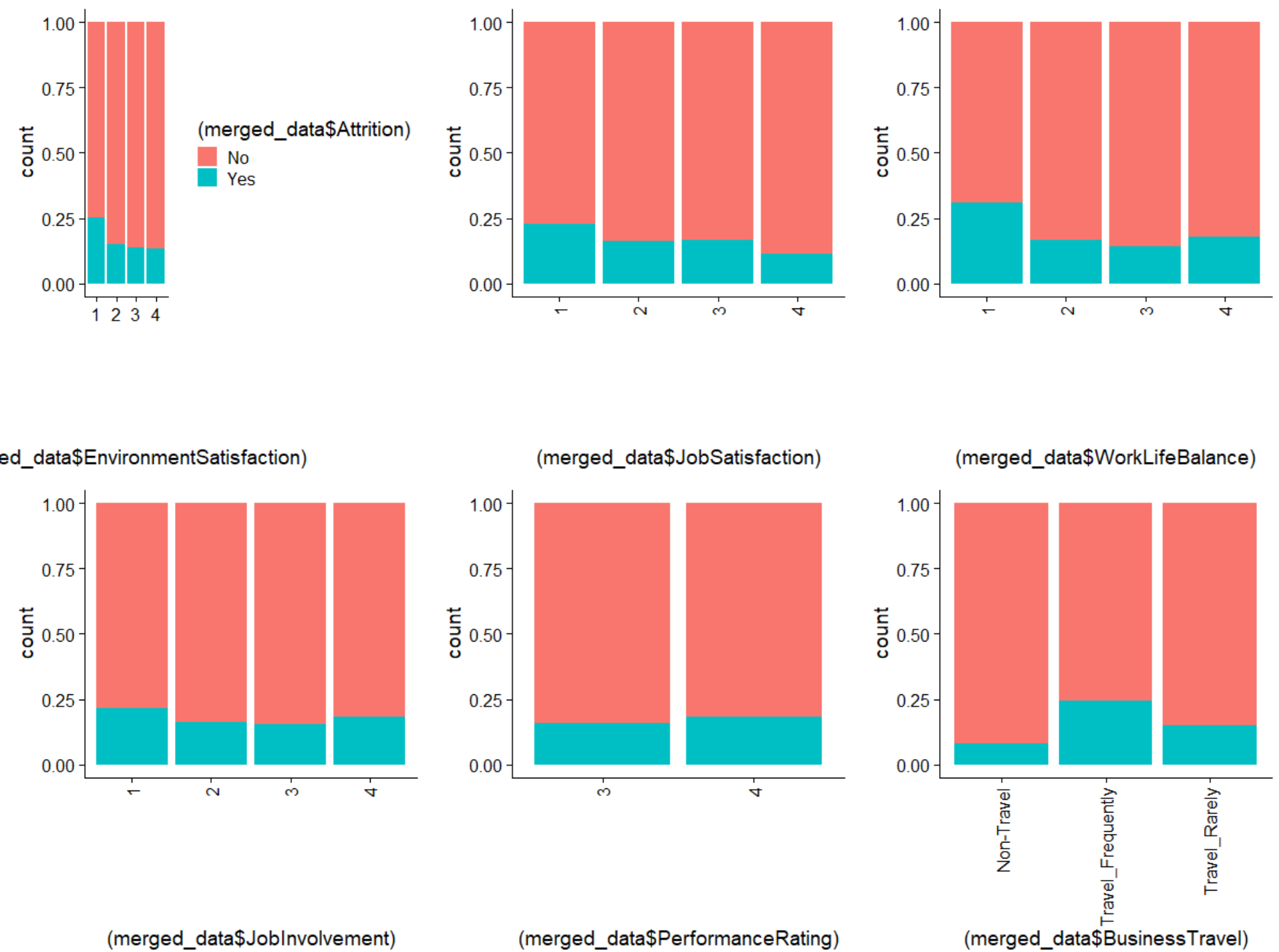
In_Time & Out_Time

- Formatting the date and the time variable
- Calculating average number of working hours for each employee
- Calculating number of leave taken during the 12 month period
- Columns where all NA are present indicates a national holiday

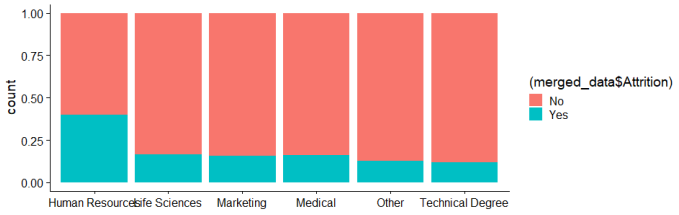
Treatment to various datasets - Summary

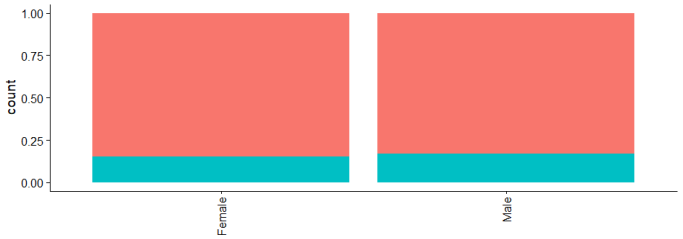
File Name	Checking for Dupli & NA employee ID	NA Values	Treatment of NA	Converting category as factors	Delating Unwanted (Rows or Columns)	Checking & Treating Outliers	Dummy Variable Creation
General_Data	No	#NumCompaniesWorked-19 #TotalWorkingYears-9	Removed rows having NA	Attrition, business travel, department, education field, education, gender, job level, job role, martial status, stock option, job level all converted	Over18, Standard hours & Employee counts both have only one value	present in 7 fields. All outliers have been removed	Yes for categorical variables with levels
Employee_Survey_Data	No	#Environment Satisfaction -25 #$JobSatisfaction-20 #Work Life Balance -38	Removed rows having NA	All 3 variable in the database have been converted			
Manager_Survey_Data	No	No NA		Both variable in the database have been converted			
In_Time	No		NA against each employee for the particular days was marked as leave		Removed all columns where entire column is marked as N.A		
Out_Time	NO				Removed all columns where entire column is marked as N.A		

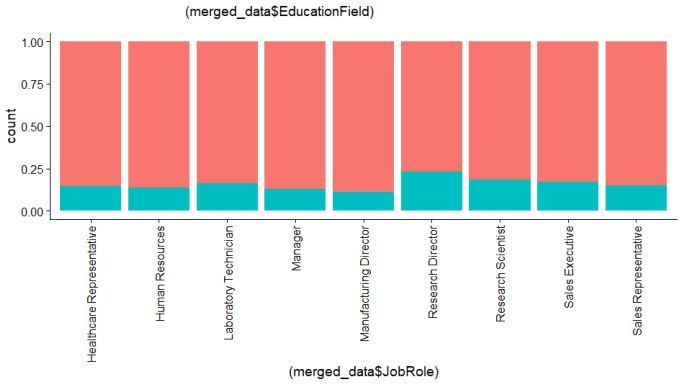
Exploratory Data Analysis



- People with very low environment and job satisfaction tend to leave company
- People with low work life balance tend to leave company which is understandable.
- Performance rating -3 have high attrition
- Travel - people with frequent travel have high attrition that can be because of travel stress and more independency to look other option as on travel

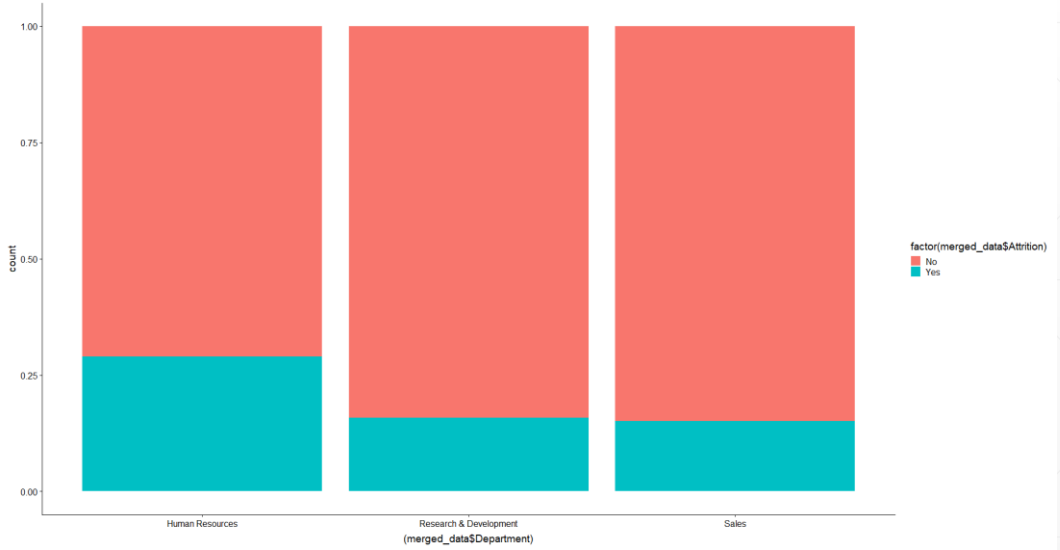


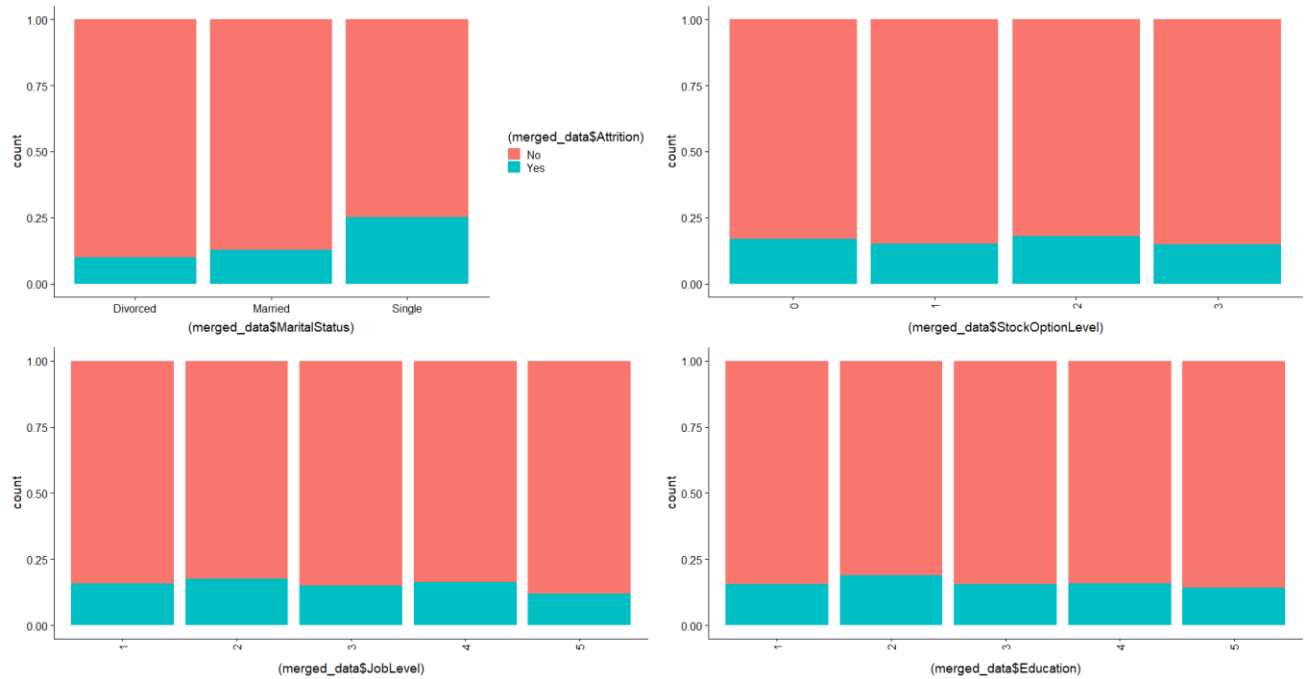




- No clear attrition based on gender. although it looks like people from human resources have high attrition.
- Reason is unclear at this point of time though

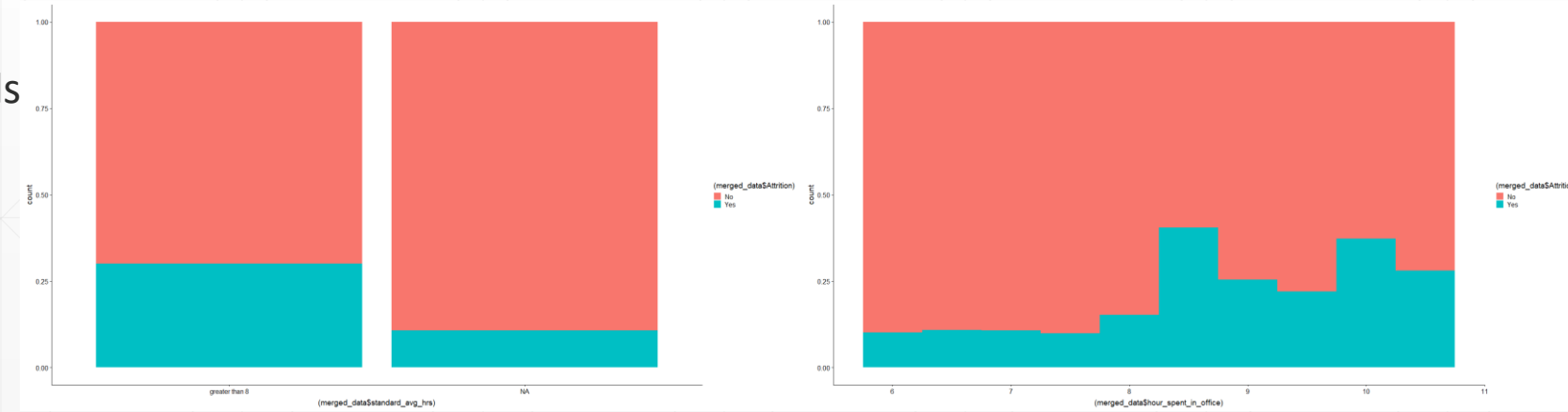
- High Attrition is found in Human and resources, which is in line our finding from job role.
- Looks like job role and department can be correlated all together.

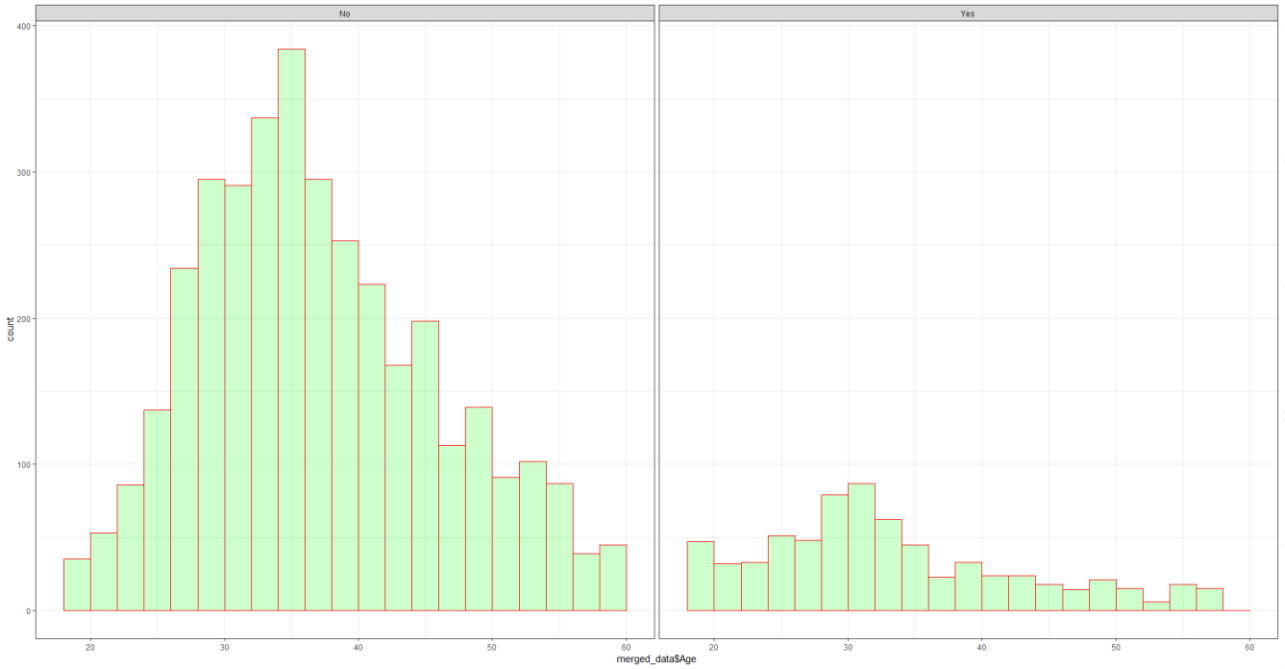




- Marital status seems to be strong factor behind high attrition. People having single status tends to leave job more.
- There is a reason for it because Single people are more risk taker, mostly independent. if they don't like the job they don’t
- Hesitate to look for other option whereas married and divorce people have more responsibilities.
- Stock option level 0,1 ,Education - Level 3 ('Bachelor') & job level does not reveal any noticeable difference in attrition.

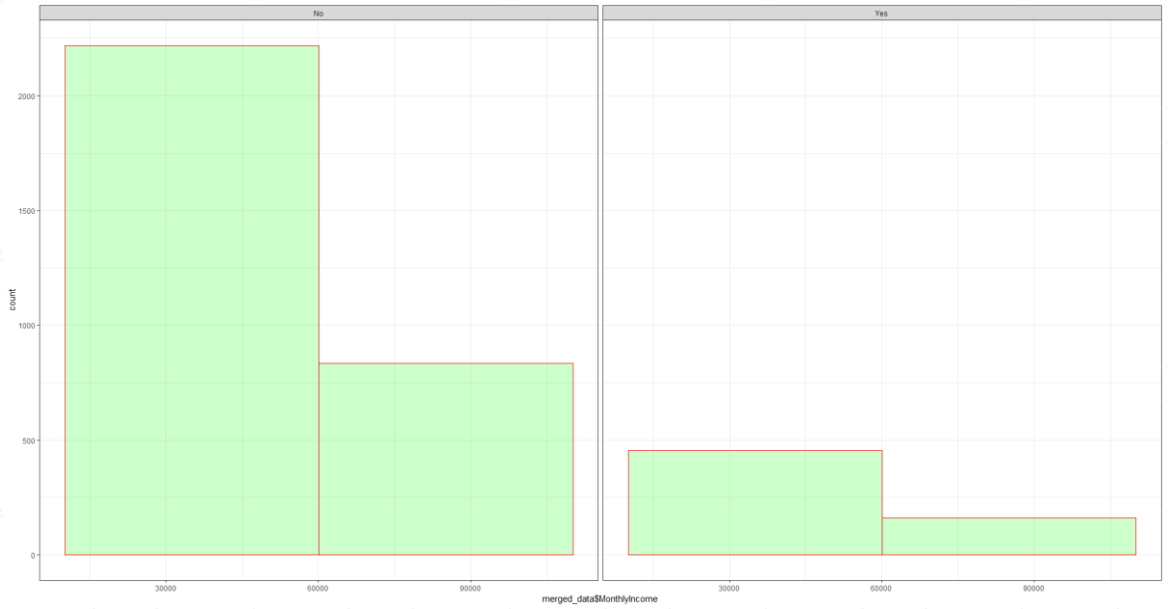
- Attrition is higher in when employee spends greater than 8 hrs
- From which it seems that overworked employees tend to leave company faster.





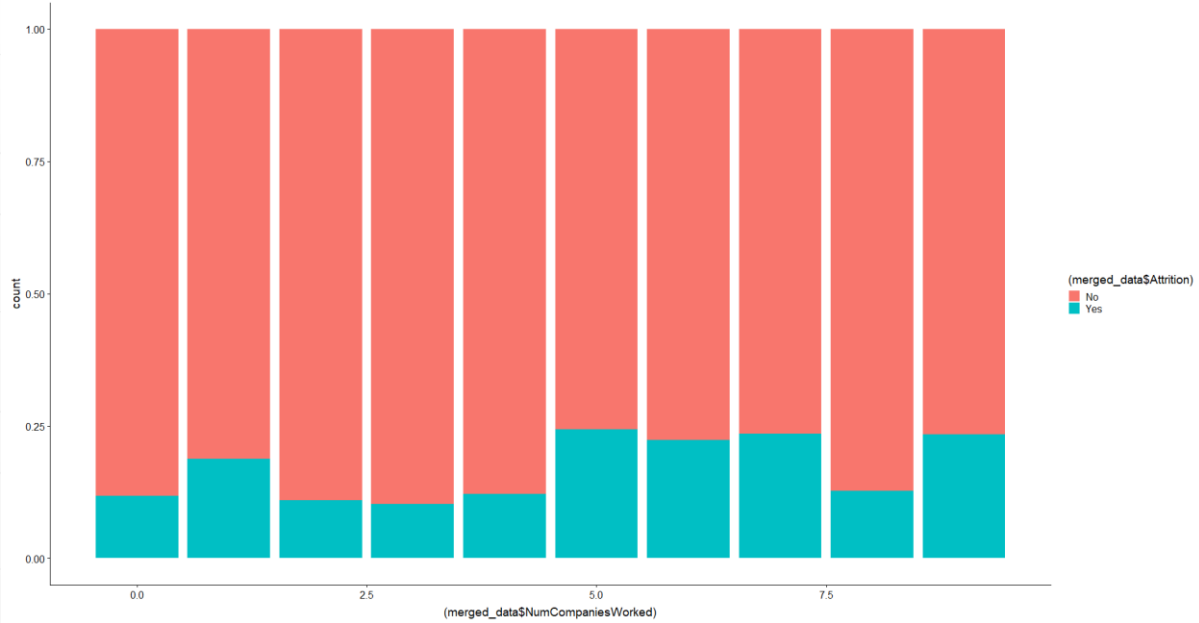
- most of the attrition happens are between 30 and 40
- note -0 indicates no attrition

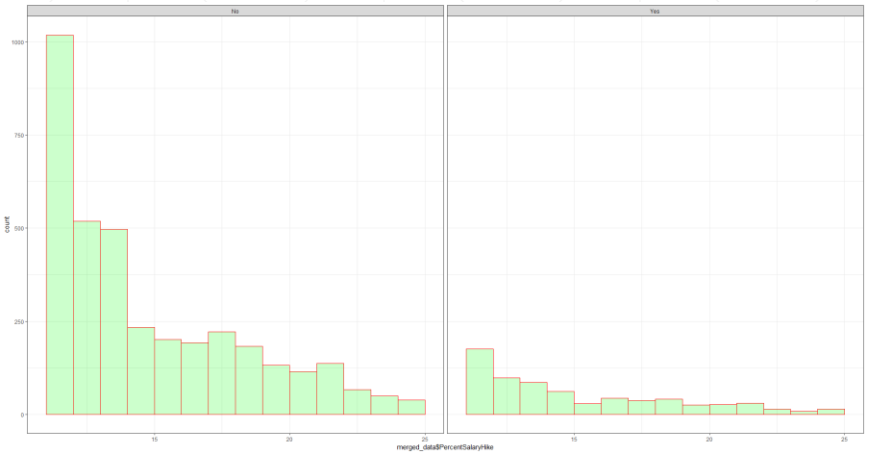
- Attrition rate more for the people living nearer to office

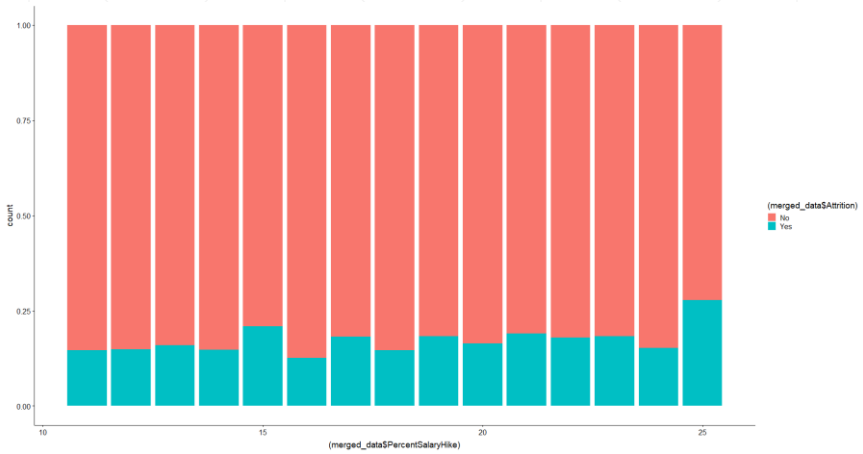


- #Lower income people tend to leave company more. Which is expected.

- #although at first look it looks like people who have not worked in other companies are leaving more.
- #attrition rate is almost similar in other categories too.
- # attrition is high due to the fact that there are more number of employees in 0 num companies worked.

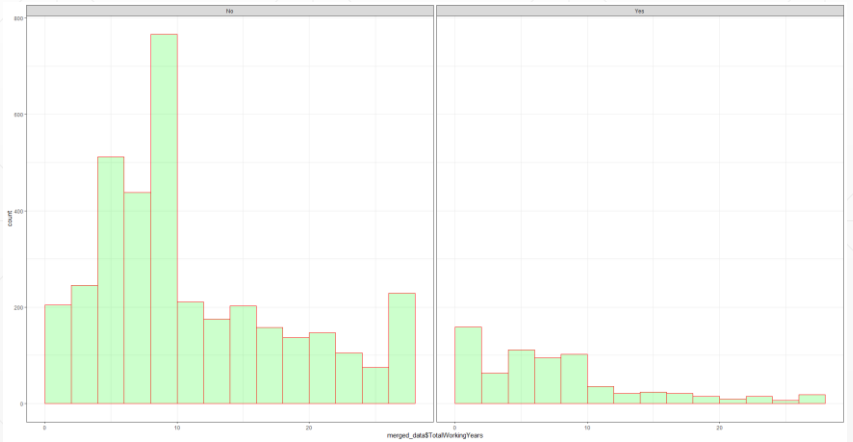


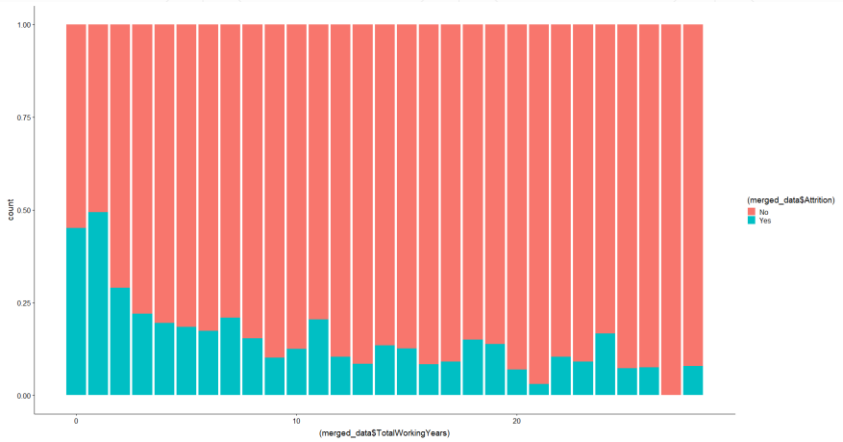


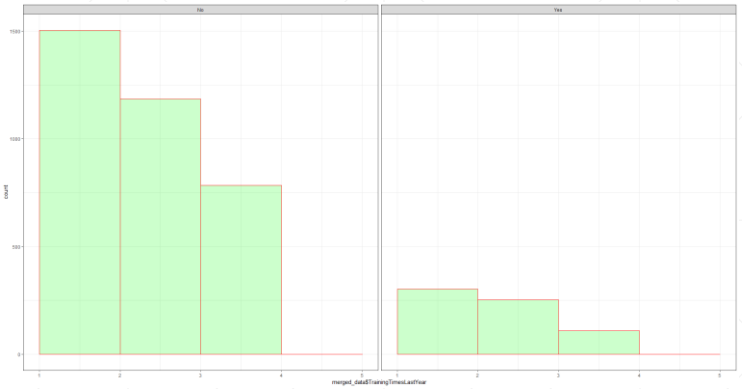


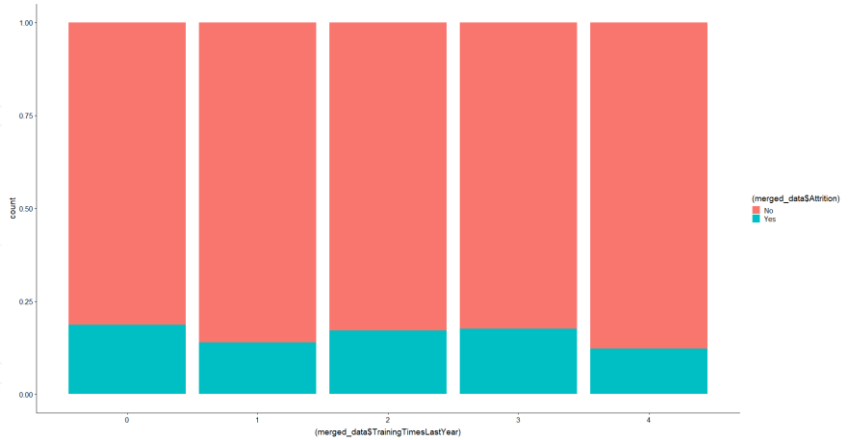
- there also attrition percentage is not high however total attritions are high in lower percent salary hike.

- people with less number of years have high attrition.



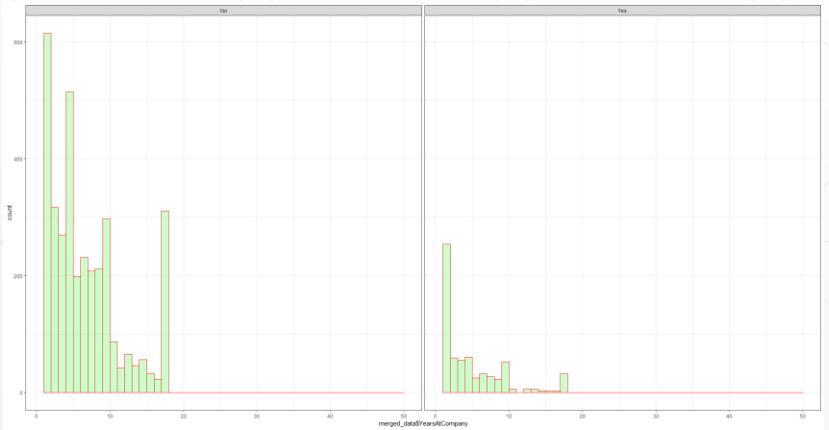


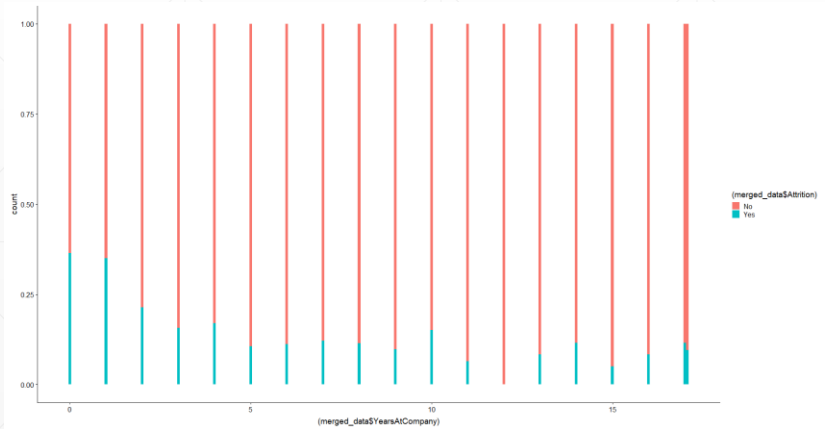


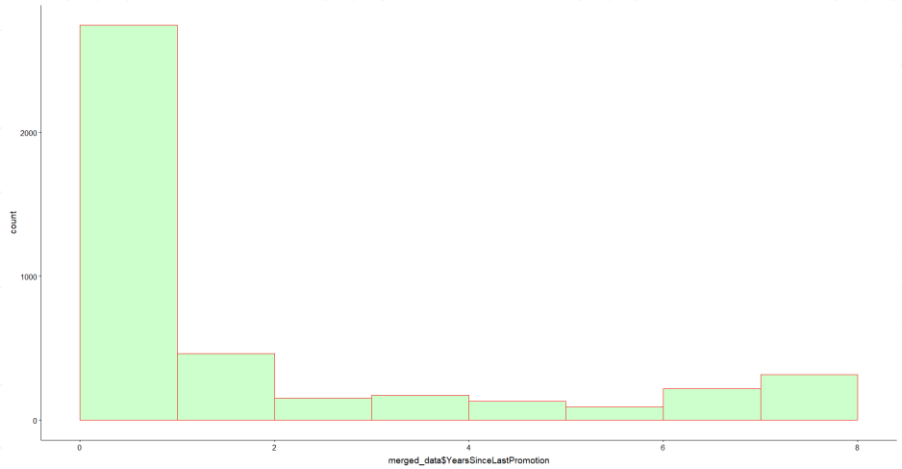


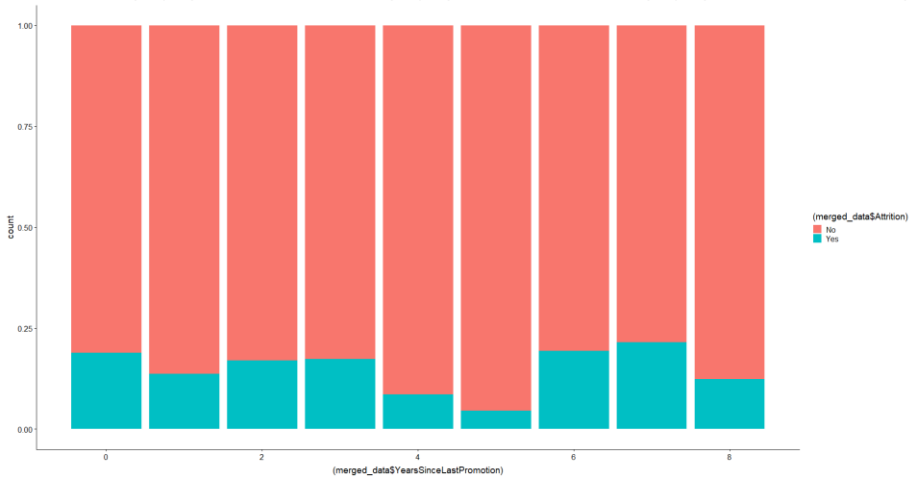
- Attrition of people who have no training last year has highest attrition rate amongst others with 6 times training having lowest.

- # people who have spent less number of years at the company tends to leave company more not in terms of number
- #but also in terms of percentage, This is evident from graph.



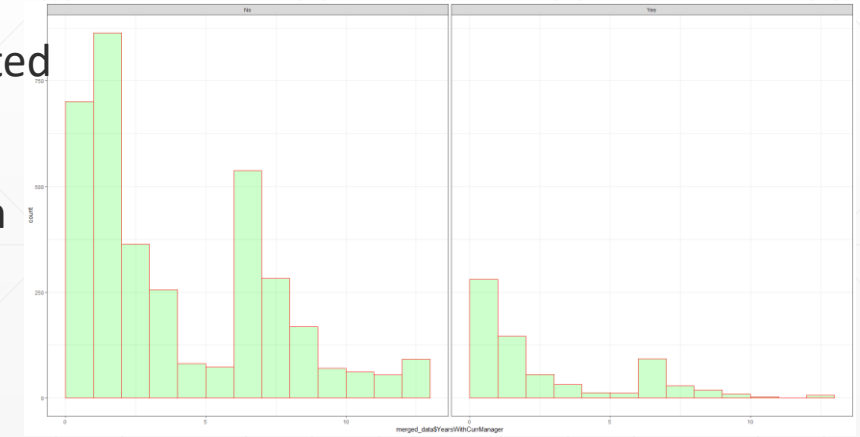


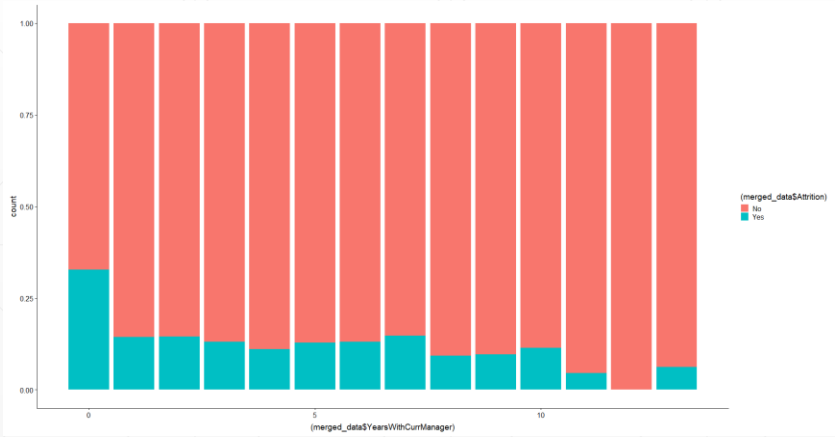


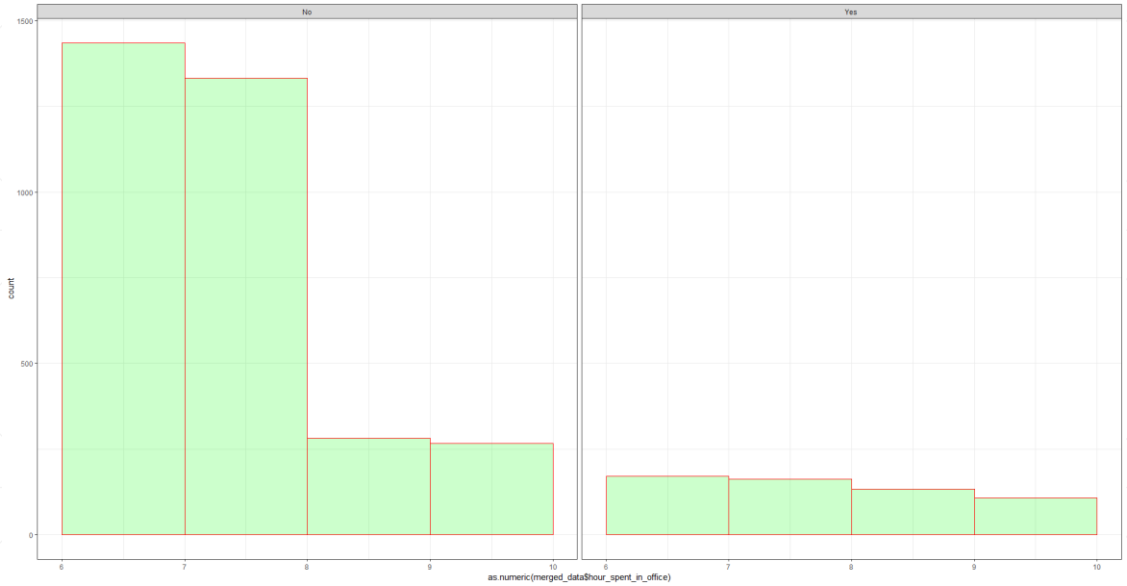


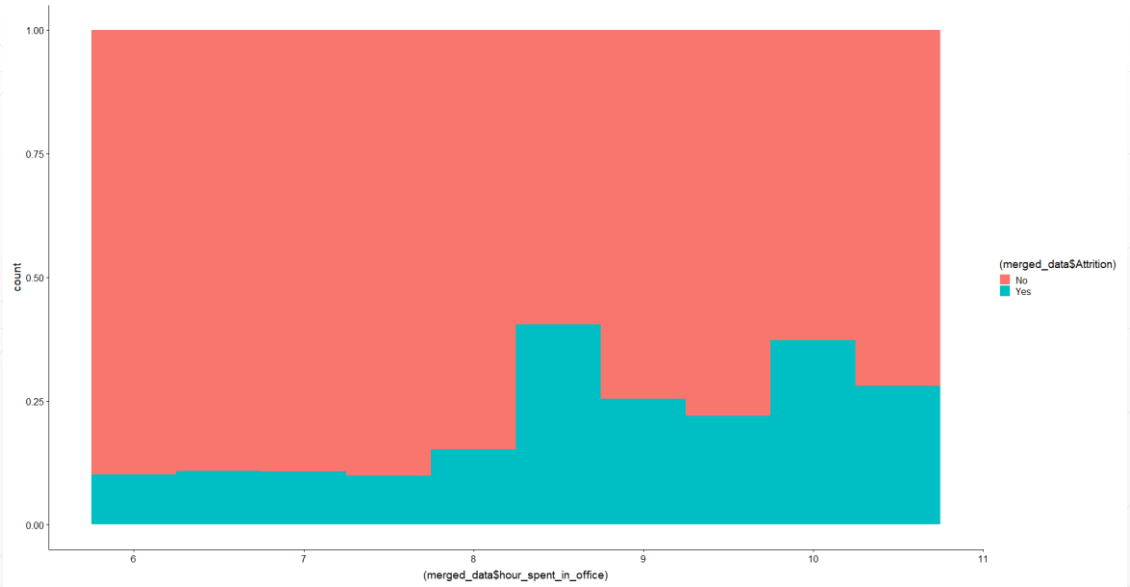
- #no clear attrition rate trend as we increase the yearssince last promotion

- # people whose manager is just appointed recently, attrition rate is high,
- # first year is very critical to the attrition once employees cross that 1 year threshold with a manager, attrition rate lower down by almost half









- # this is also very evident from the second plot that as soon as average number of hour increase from 8 hours,
- #people feel overworked and tend to leave company more.

Model Creation

- Final Dataset: attrition_final
- Record Count: 4099 obs. of 59 variables
- Split Ratio: 70:30
- Seed Set: 100

Procedure

- Model 1 & Model 2
  - Family = Binomial
  - Stepwise selection (automatically selecting a reduced number of predictor variables for building the best performing logistic regression model)
  - Removing multicollinearity through VIF check

Model No.	Field Eliminated	Reason
2	EducationField.x Life.Sciences	because of high VIF. It has been high VIF because of since model_1
3	hour_spent_in_office	high VIF and high insignificance
4	▪ hours_spent_office & ▪ BusinessTravel.xTravel_Rarely	▪ AIC did not hcange much ▪ high VIF of 3.557704 and high insignificance
5	worklife.balance.x2	▪ both WorkLifeBalance.x3 and WorkLifeBalance.x2 but WorkLifeBalance.x2 has lower significance than WorkLifeBalance.x3 ▪ seems both variable is highly correlated. ▪ We can check that ▪ correlation is -0.68 which shows high collinearity ▪ we can remove worklife.balance.x2 from the data because of collinearity ▪ removing WorkLifeBalance.x2 because of high VIF of 3.035984, it has been high VIF since model_1
6	EducationField.xMarketing	▪ All variable have VIF below 2.5 ▪ EducationField.xMarketing high insignificance of 0.46
7	EducationField.xMedical	high insignificance of 0.453
8	WorkLifeBalance.x4	high insignificance of 0.171
9	EducationField.xTechnical.Degree	high insignificance of 0.157
10	EducationField.xOther	high insignificance of 0.184
11	remove JobLevel.x5	high insignificance of 0.121
12	JobRole.xLaboratory.Technician	high insignificance of 0.083895

Model No.	Field Eliminated	Reason
13	Education.x5	high insignificance of 0.071647
14	JobRole.xResearch.Scientist	high insignificance of 0.0866
15	StockOptionLevel.x1	high insignificance of 0.045264
16	JobRole.xResearch.Director	high insignificance of 0.011240
17	TrainingTimesLastYear	high insignificance of 0.017454
18	JobRole.xSales.Executive	high insignificance of 0.015578
19	JobSatisfaction.x2	high insignificance of 0.002215
20	JobSatisfaction.x3	high insignificance of 0.052719
21	worklifebalance.x3	▪ Although all the variables are looking decently significant. however worklifebalance.x3 is on the verge of insignificance (p = 0.000826) lets check removing that wel will keep removing it until we get P value very low (~10e-4) and use all combinations to see which gives us better result.
22	JobRole.x Manufacturing.Director	▪ insignificance of 0.000186
23		▪ removing the variable increased AIC drastically . it not advisable to remove JobRole.xManufacturing. Director our final model is model_22

model_22

glm(formula = Attrition ~ Age + NumCompaniesWorked + TotalWorkingYears + YearsSinceLastPromotion + YearsWithCurrManager + EnvironmentSatisfaction.x2 + EnvironmentSatisfaction.x3 + EnvironmentSatisfaction.x4 + JobSatisfaction.x4 + BusinessTravel.xTravel_Frequently + JobRole.xManufacturing.Director + MaritalStatus.xSingle + standard_avg_hrs.xgreater.than.8, family = "binomial", data = train)

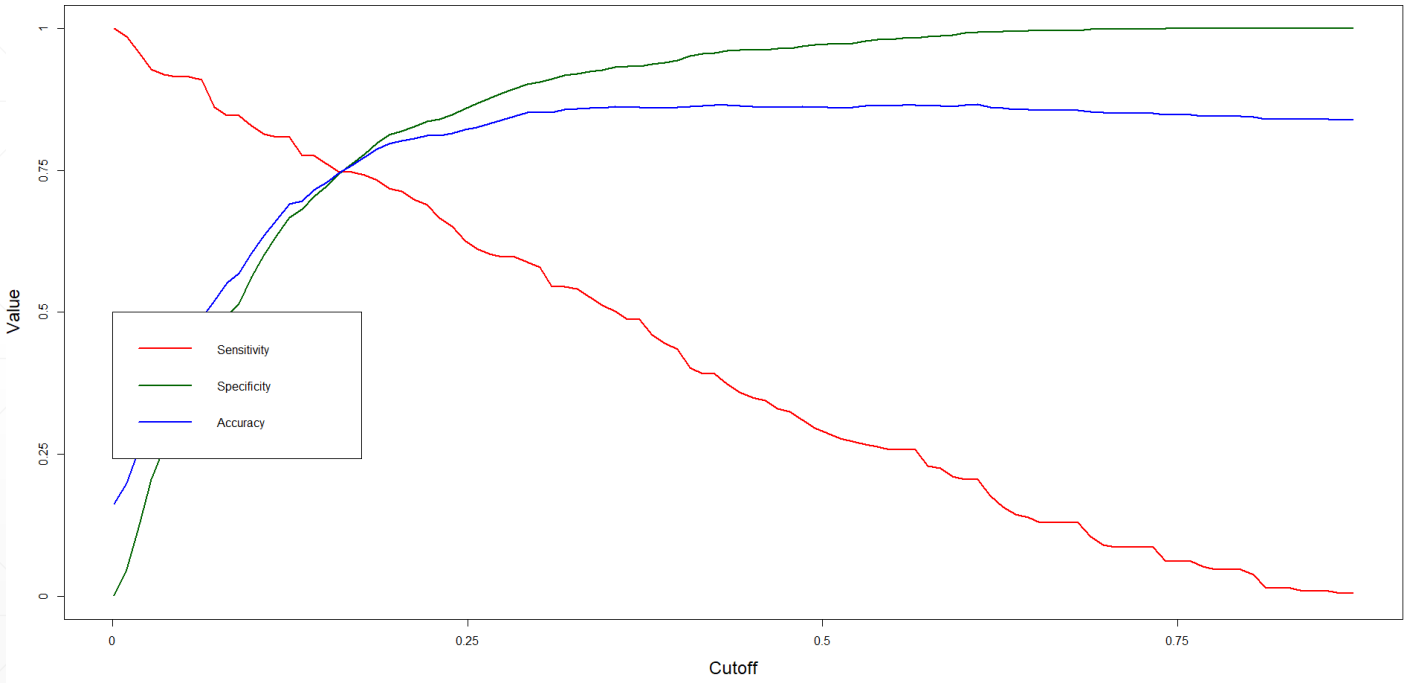
Model Assessment & Recommendation

Prediction Based on Final Model

Confusion Matrix

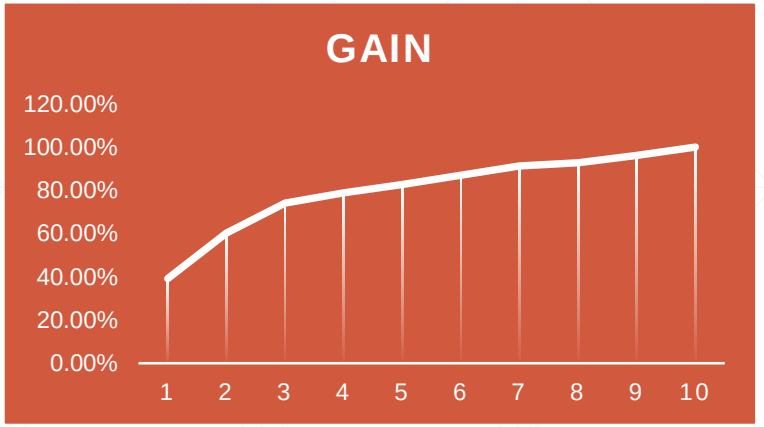
Prediction	Reference		
	No	Yes	Total
No	1025	122	1147
Yes	56	87	143
Total	1081	209	1323

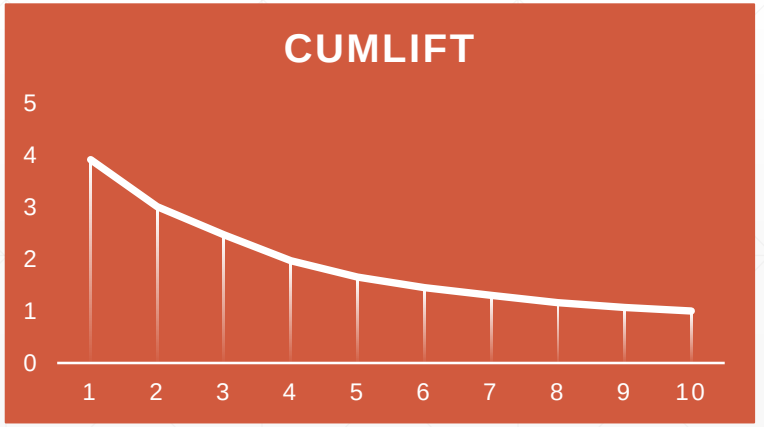
Accuracy : 74%
Sensitivity : 75%
Specificity : 74%



Model Assessment (GAIN & LIFT charts)

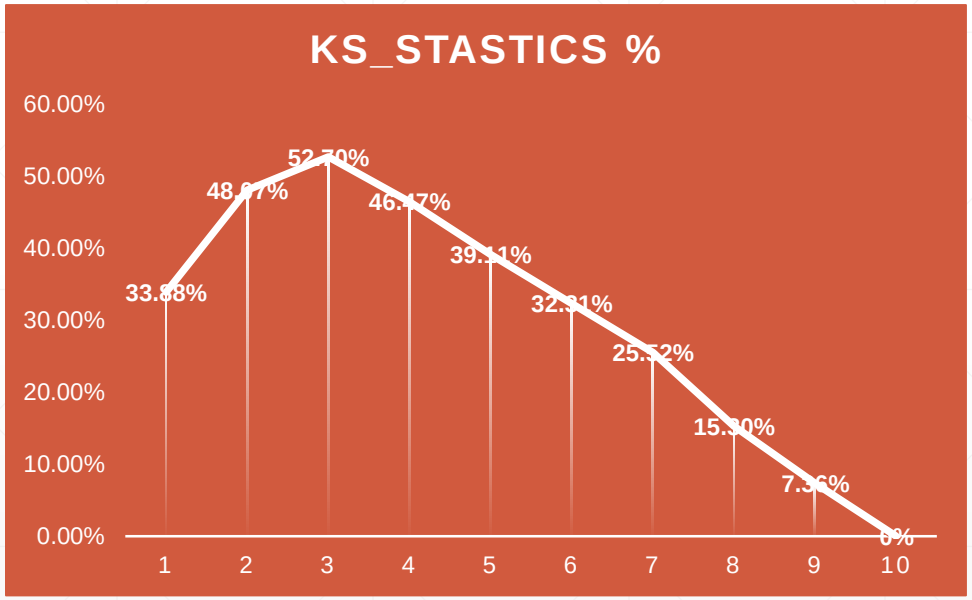
Bucket	Total	Totalresp	Cum Resp	Gain	Cumlift
1	129	82	82	39.23%	3.923445
2	129	44	126	60.29%	3.014354
3	129	29	155	74.16%	2.472089
4	129	10	165	78.95%	1.973684
5	129	8	173	82.78%	1.655502
6	129	9	182	87.08%	1.451356
7	129	9	191	91.39%	1.305537
8	129	3	194	92.82%	1.160287
9	129	7	201	96.17%	1.068581
10	129	8	209	100%	1
Total	**1290**	**209**			





Model Assessment (KSS)

Bucket	Total	Totalresp	Cum Resp	Gain	Cumlift	non_attrition	cum_non_attrition	cum_non attrition_%	KS_Stastics %
1	129	82	82	39.23%	3.923445	47	47	4.34%	33.88%
2	129	44	126	60.29%	3.014354	85	132	12.21%	48.07%
3	129	29	155	74.16%	2.472089	100	232	21.46%	52.70%
4	129	10	165	78.95%	1.973684	119	351	32.46%	46.47%
5	129	8	173	82.78%	1.655502	121	472	43.66%	39.11%
6	129	9	182	87.08%	1.451356	120	592	54.74%	32.31%
7	129	9	191	91.39%	1.305537	120	712	65.86%	25.52%
8	129	3	194	92.82%	1.160287	126	838	77.52%	15.30%
9	129	7	201	96.17%	1.068581	122	960	88.80%	7.36%
10	129	8	209	100%	1	121	1081	100%	0%
Total	**1323**	**229**			**1094**				



Model Summary

- The model has an increasing Gain and a decreasing Lift.
- The Model predicts close to 79% of the attritions within the 4th decile with 74% accuracy
- Max KSS shows at 3rd decile at 52.70%
- The KS statistic shows that the model is very good in distinguishing between employees who will leave the company and employees who won’t

Conclusion

- The model results show that the company should focus on the following attributes to curb attrition
  - Lesser Age
  - NumCompaniesWorked
  - TotalWorkingYears
  - YearsSinceLastPromotion
  - YearsWithCurrManager
  - EnvironmentSatisfaction.x2 / x3 / x4 (Medium / High / Very High)
  - JobSatisfaction.x4
  - BusinessTravel.xTravel_Frequently
  - JobRole.xManufacturing.Director
  - MaritalStatus.xSingle
  - standard_avg_hrs.xgreater.than.8