HR Analytics Study

SUBMISSION

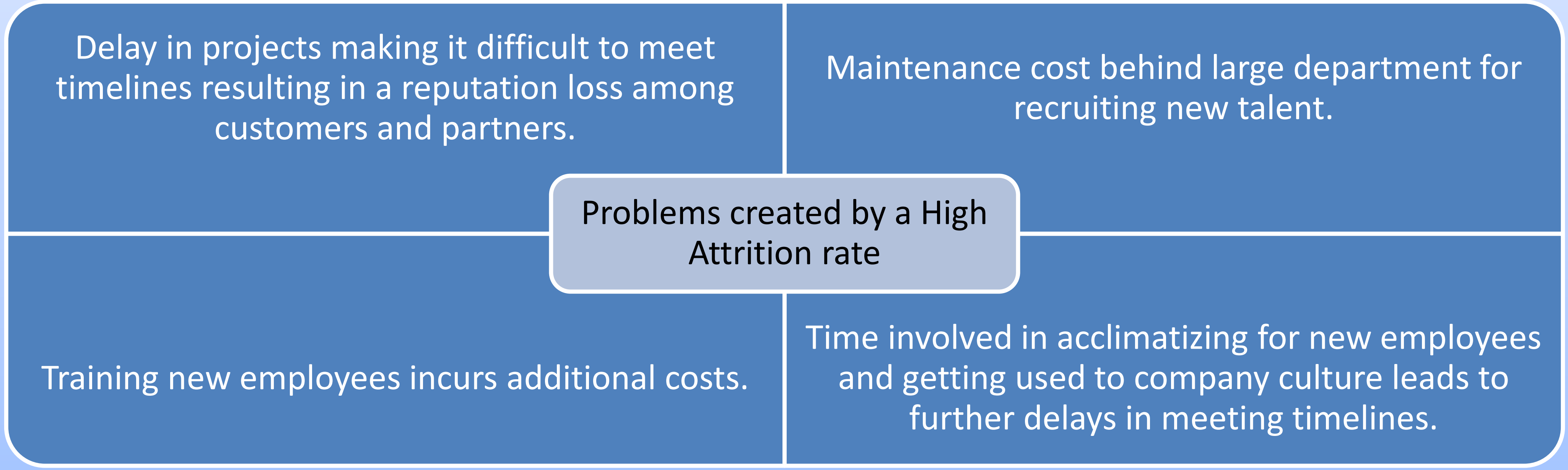
Group Name:

1. Subhanshu Rathi
2. Pavan ML
3. Ravi Shekhar Rai
4. Srikanth K

The Problem Statement

Some statistics on XYZ company

- Current workforce strength ~ 4000
- Each year nearly 15% of employee leave the company.
- The attrition rate is much higher then the ideal turnover rate of 10 percent.



Problems created by a High Attrition rate

Delay in projects making it difficult to meet timelines resulting in a reputation loss among customers and partners.

Maintenance cost behind large department for recruiting new talent.

Training new employees incurs additional costs.

Time involved in acclimatizing for new employees and getting used to company culture leads to further delays in meeting timelines.

Goals and Methodology

Management of xyz company have contracted our firm to figure out

•Most important factor responsible for attrition

•Provide Suggestion to reduce attrition rate

•Pinpoint and make changes to their workplace to get most employees to stay.

Our problem solving methodology

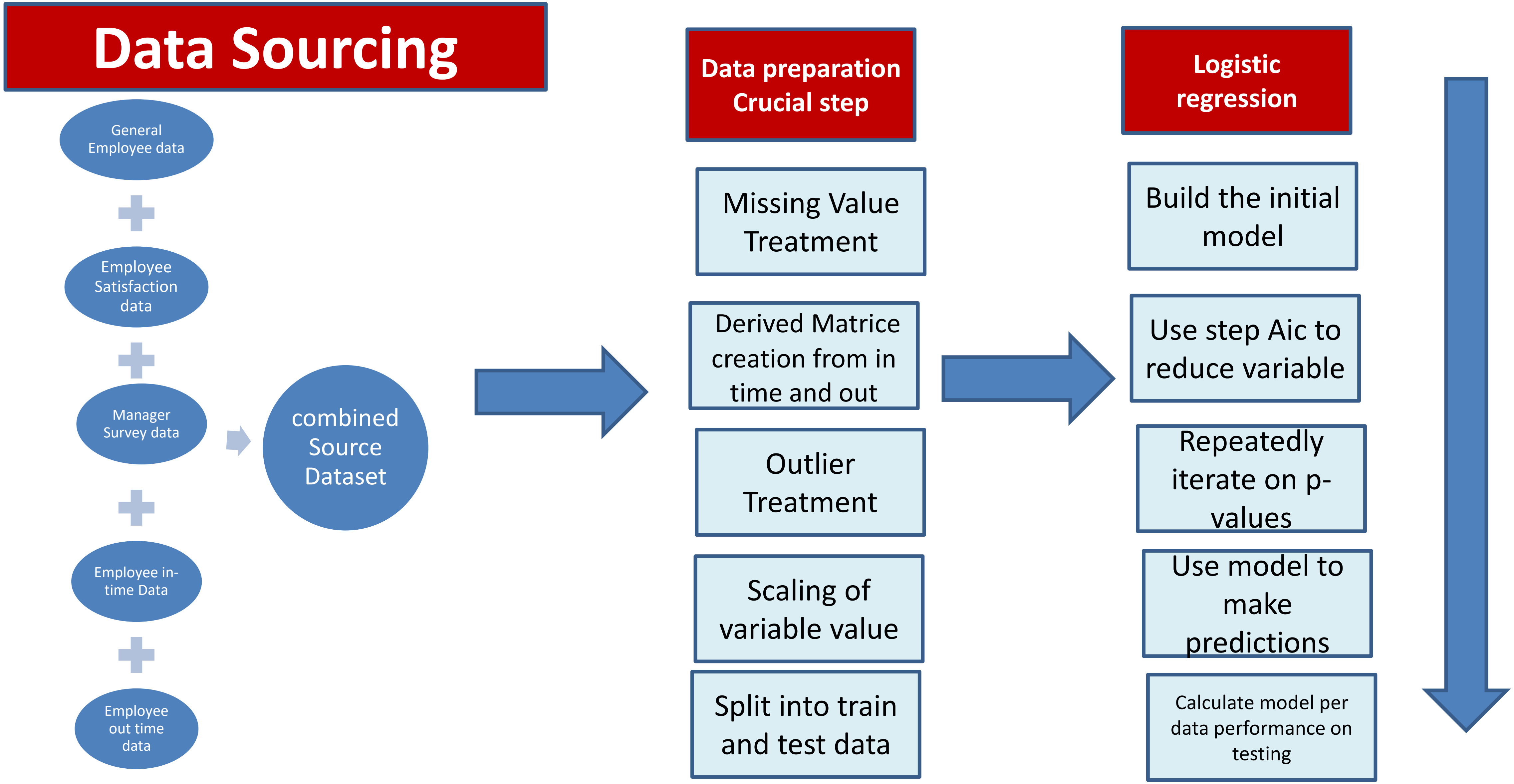
•Model probability of attrition using a Logistic regression model.

•Figure out the most important variable from the model.

•Use the variable and their co- efficients to infer how they are related to Attrition rate.

•Suggest how to curb the attrition rate based on the finding.

Data Analysis Methodology



Data Sourcing

General Employee data

+

Employee Satisfaction data

+

Manager Survey data

+

Employee in-time Data

+

Employee out time data

combined Source Dataset

Data preparation Crucial step

Missing Value Treatment

Derived Matrice creation from in time and out

Outlier Treatment

Scaling of variable value

Split into train and test data

Logistic regression

Build the initial model

Use step Aic to reduce variable

Repeatedly iterate on p-values

Use model to make predictions

Calculate model per data performance on testing

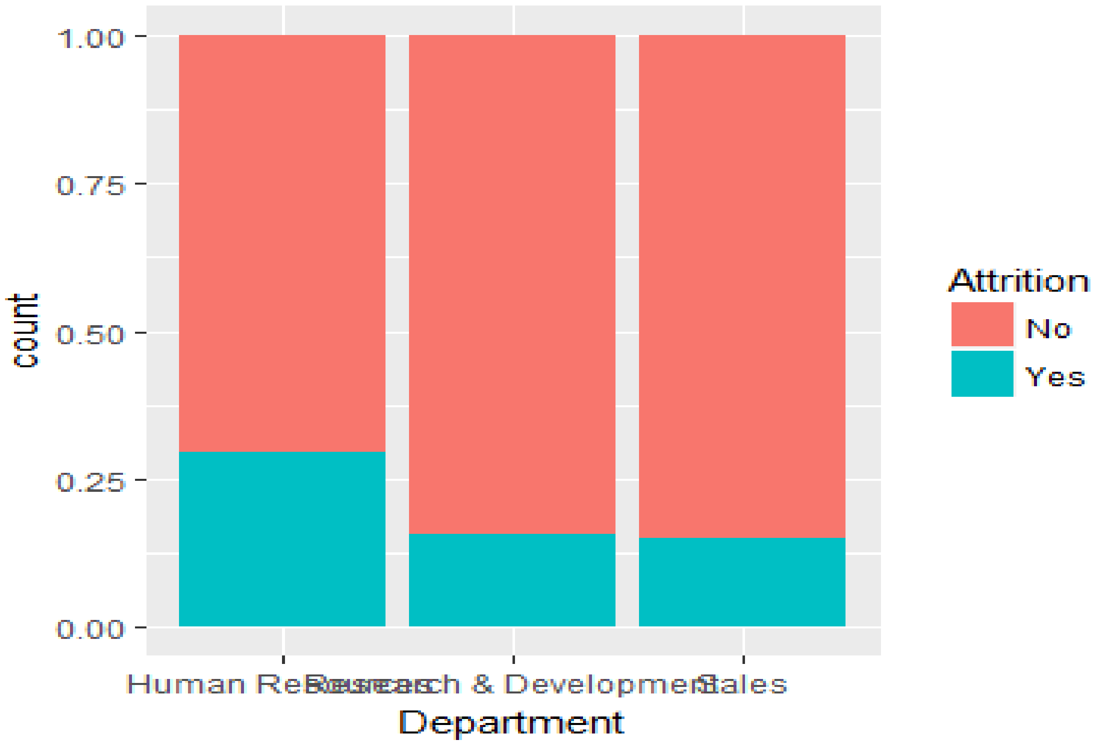
DATA CLEANSING

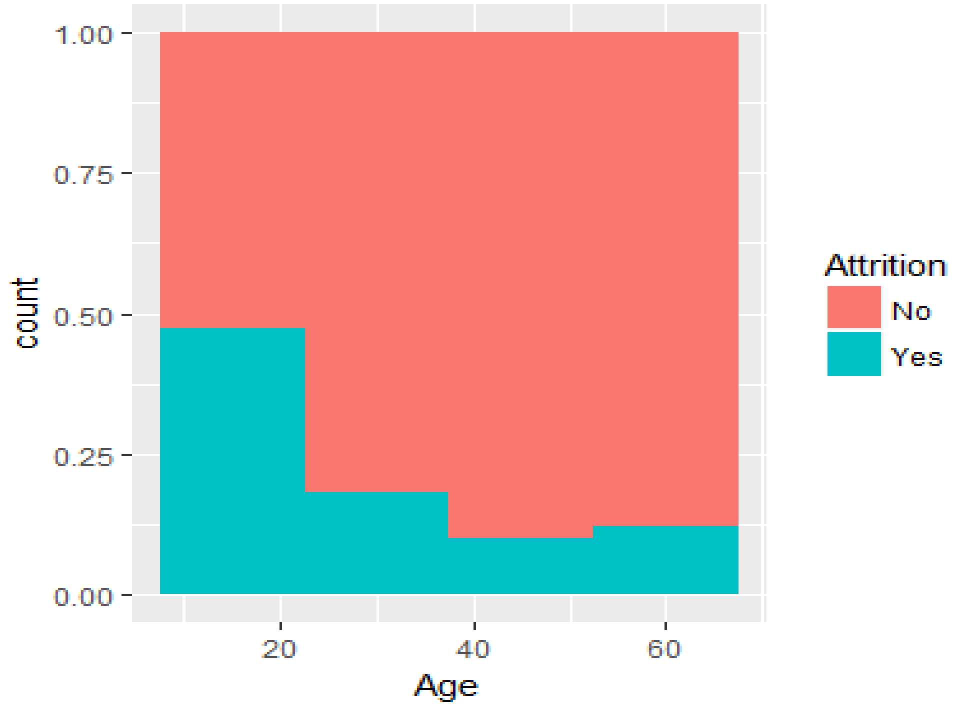
S.No	Category	Activity
1	Duplicate Value handling	- Checked all data sets present and no duplicate value found
2	NA Column Handling	- manager_survey_data - No NA values Present - empl_survey_data - returns 83, which is only 1.8% of total observations(we will check for the columns and replace NA's with mode of the values as mode gives the maximum occurrence of feedback points and also results in integral value unlike mean) - general_data - returns 28, which is a fairly small number considering 4410 (only two columns are having NA values and they are Num Companies Worked (19 NAs) & Total Working Years(9 NAs) will remove the rows with NA values after merging) - in_time, out_time- too many NA values same number of NA values as in_time, which might be there because of the absence of person from office on particular dates we will look at them after merging the data frames.
3	Blank Values	- Checked all data sets no Blank Values present.
4	Primary Key Detection	- Employee id is the common primary key in all given data sets.
5	Date time format	- Date format is changed to Y-M-D format. - Time is changed to HH:mm:YY Format.
6	Unnecessary column handling	- Removed all the unnecessary columns which are not relevant for our analysis like Employee Count, Over18 & Standard hours which have static value and the remaining NA's in the merged final data set

DATA PREPARATION

S.No	Category	Activity
1	Derived Columns	• avg_working_time • working_overtime(Yes or NO) • overtime_count • leave_counts
2	Working Hour Calculation	• Checking number of leaves • The leaves are indentified from in_time and out_time and is used for data analysis. • Merging in_time and out_time and all the NA values are in same row in In_time and out_time records which signal that the employee is absent on respective date.
3	Column name addition	• Changing the name of first column in both data frames , "in_time" & "out_time“ to "EmployeeID"

Univariate/Bivariate Analysis of Data

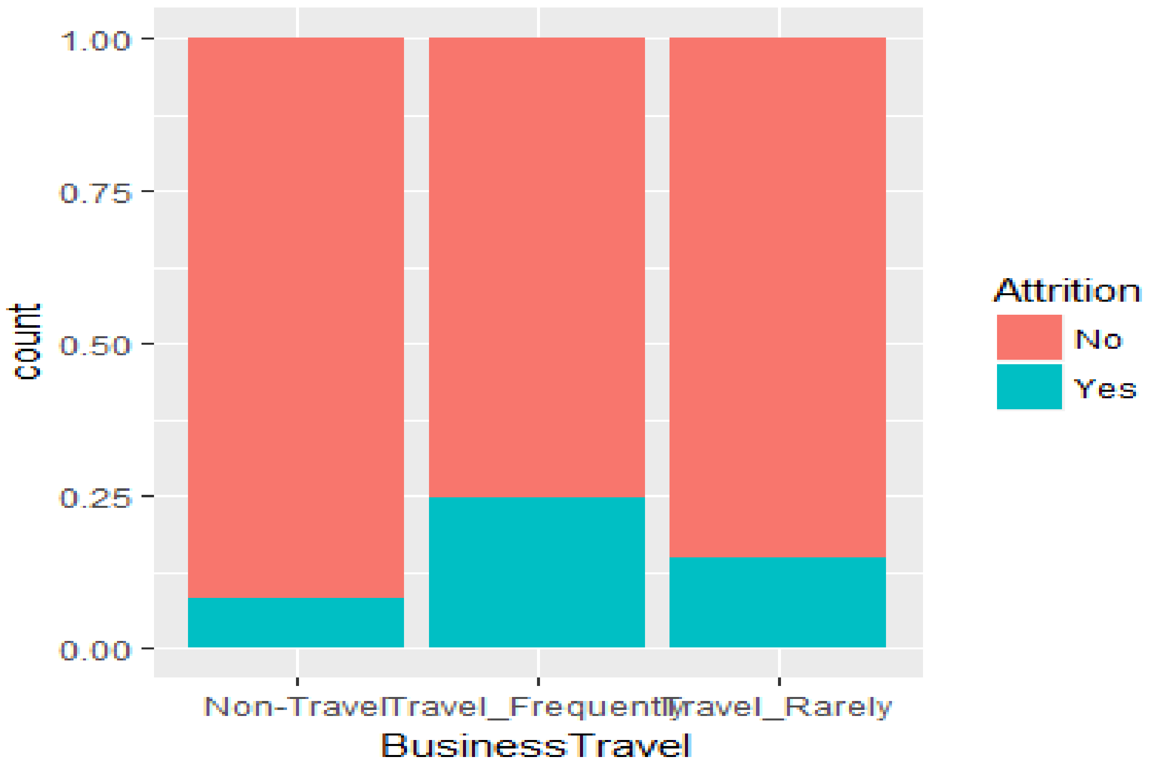




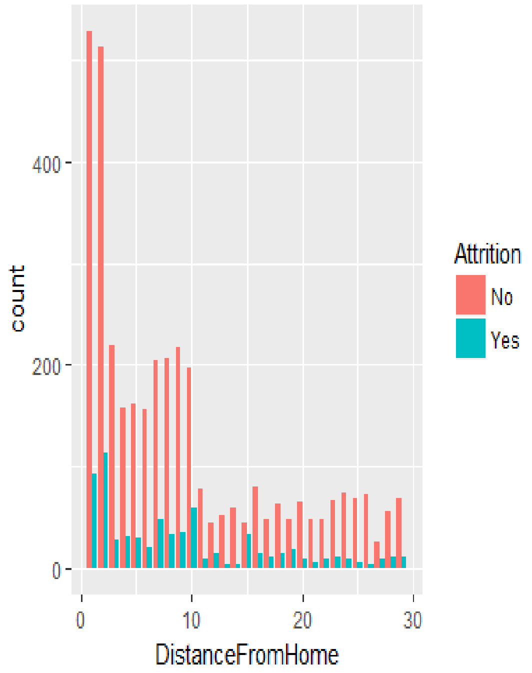
- ✓ HR department has high attrition rate but not significant enough to conclude anything

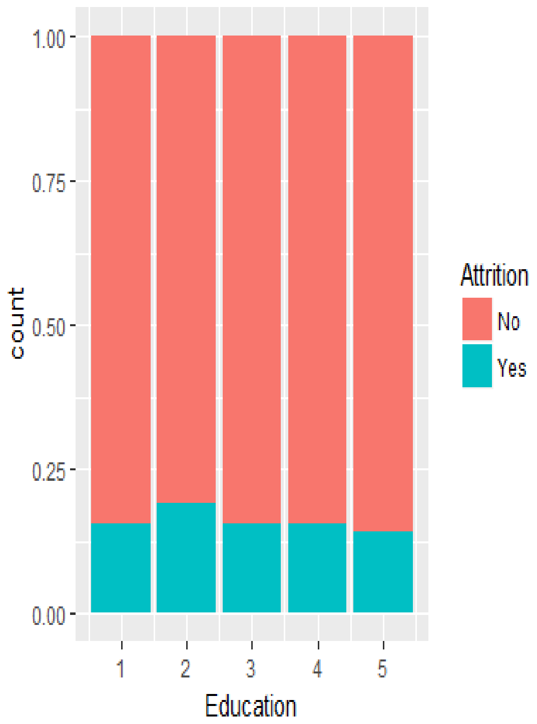
- ✓ Lower age groups are showing high attrition rates

Univariate/Bivariate Analysis of Data



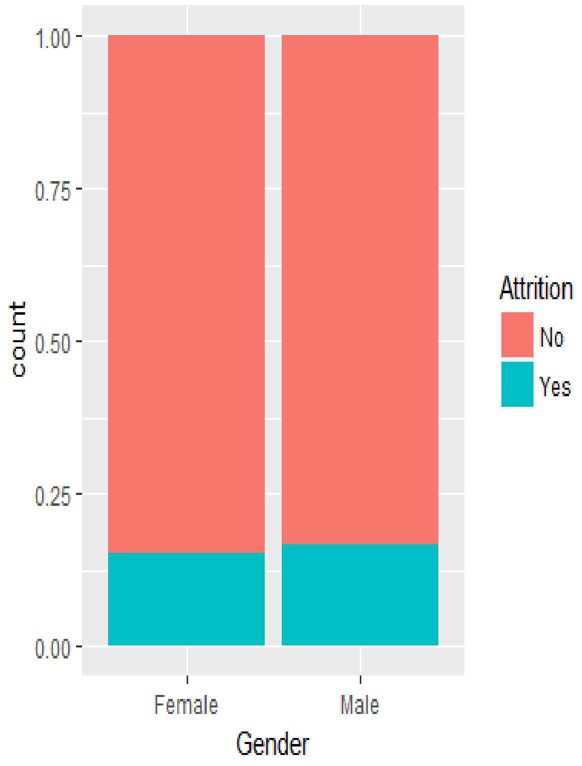
- ✓ Those who travel frequently have higher attrition rates

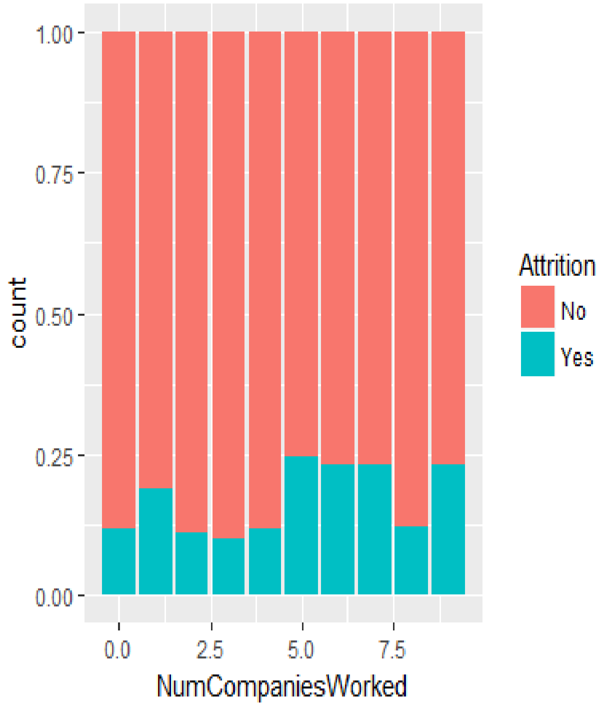




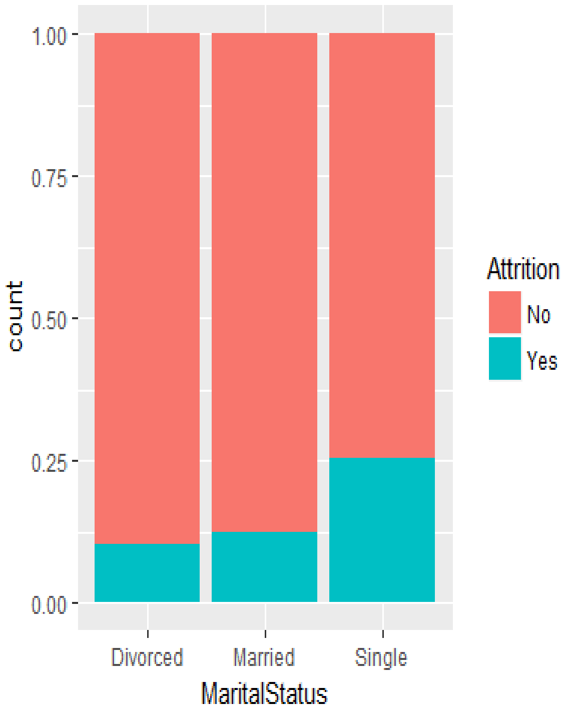
- ✓ Education doesn’t seem to be a factor effecting attrition.
- ✓ Distance From Home doesn't seem to be a factor effecting attrition.

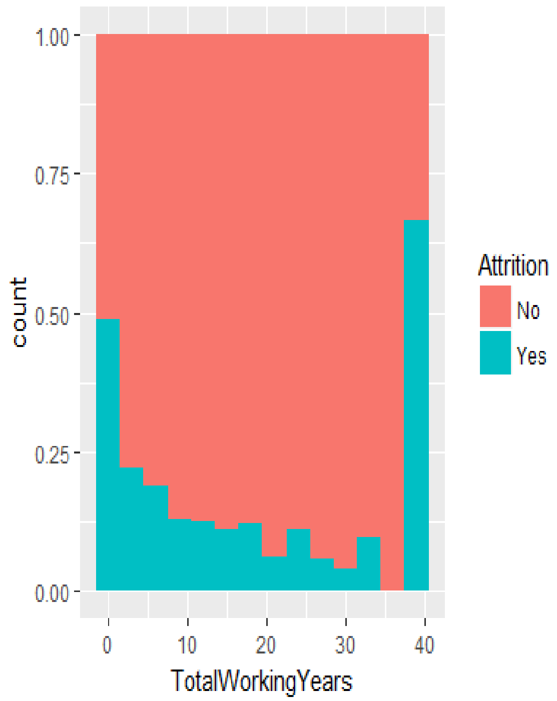
Univariate/Bivariate Analysis of Data





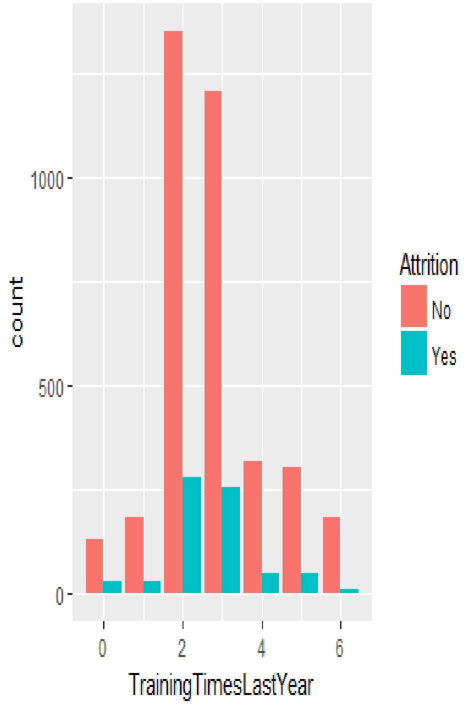
- ✓ No apparent effect of gender & Job level on attrition
- ✓ Number of companies switched is not affecting attrition

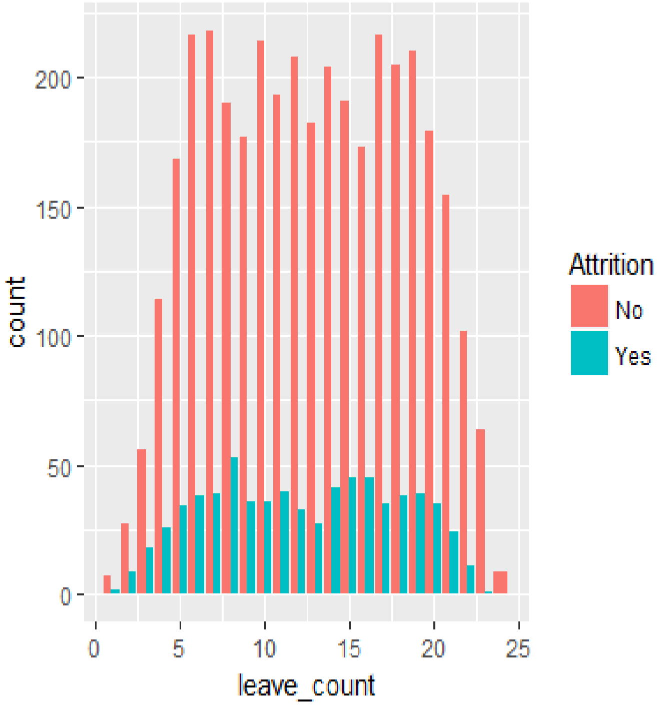


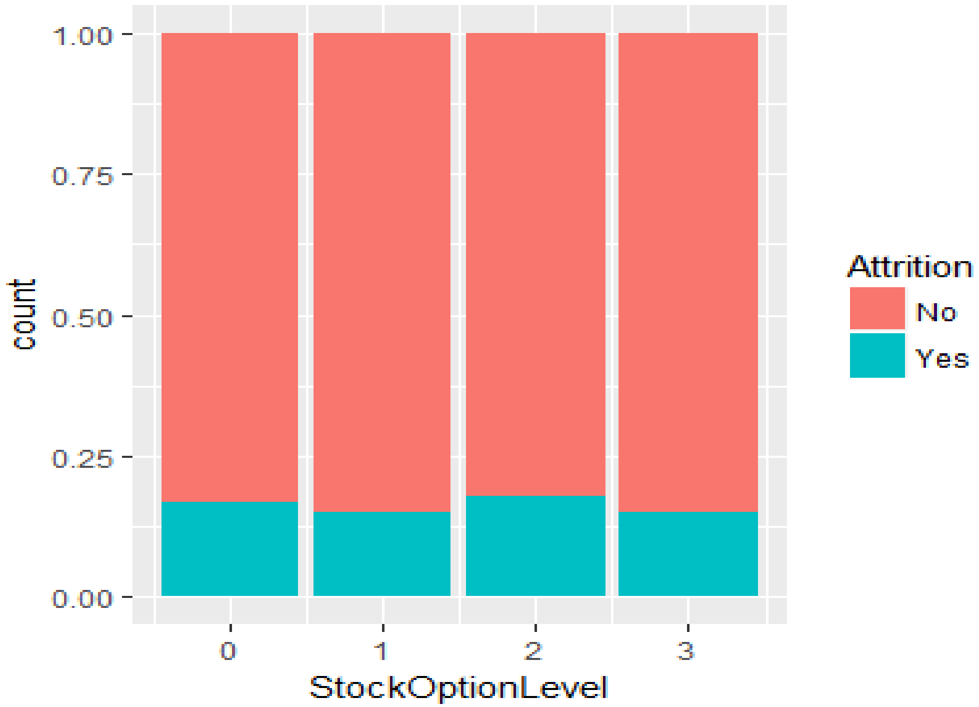


- ✓ Employees who are single have high attrition rate but it is correlated
  with lower age.
  Employees with less years of experience and over 40 have considerable high attrition rate

Univariate/Bivariate Analysis of Data



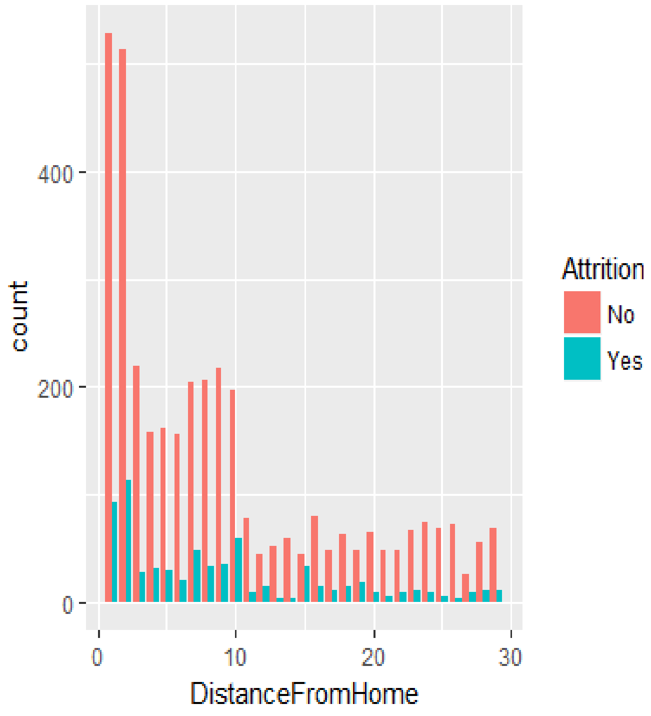


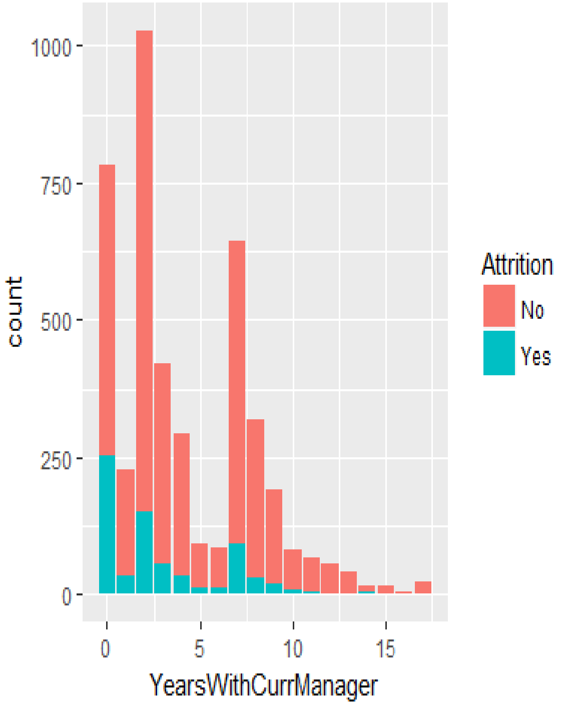


✓ Training & Leave count also has no positive impact

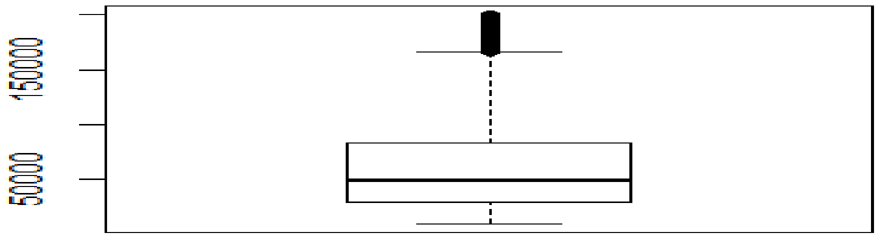
✓ No apparent effect of stock option on attrition

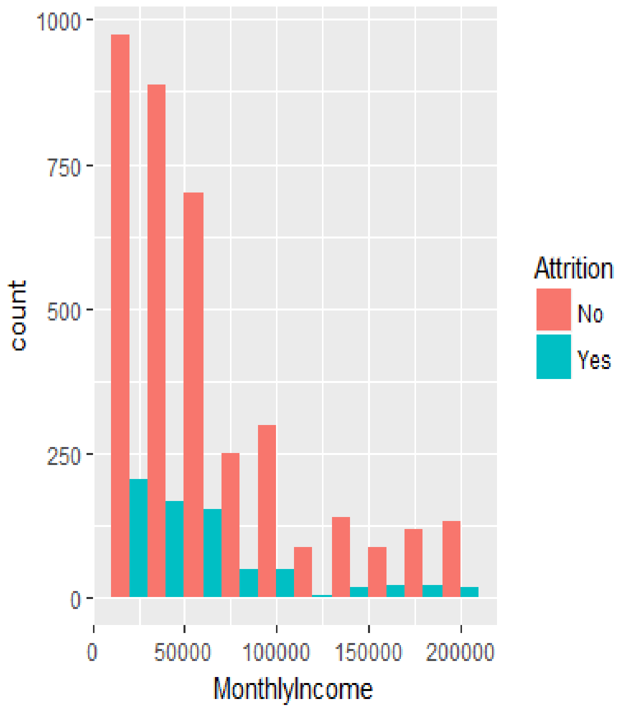
Continuous Variable Analysis





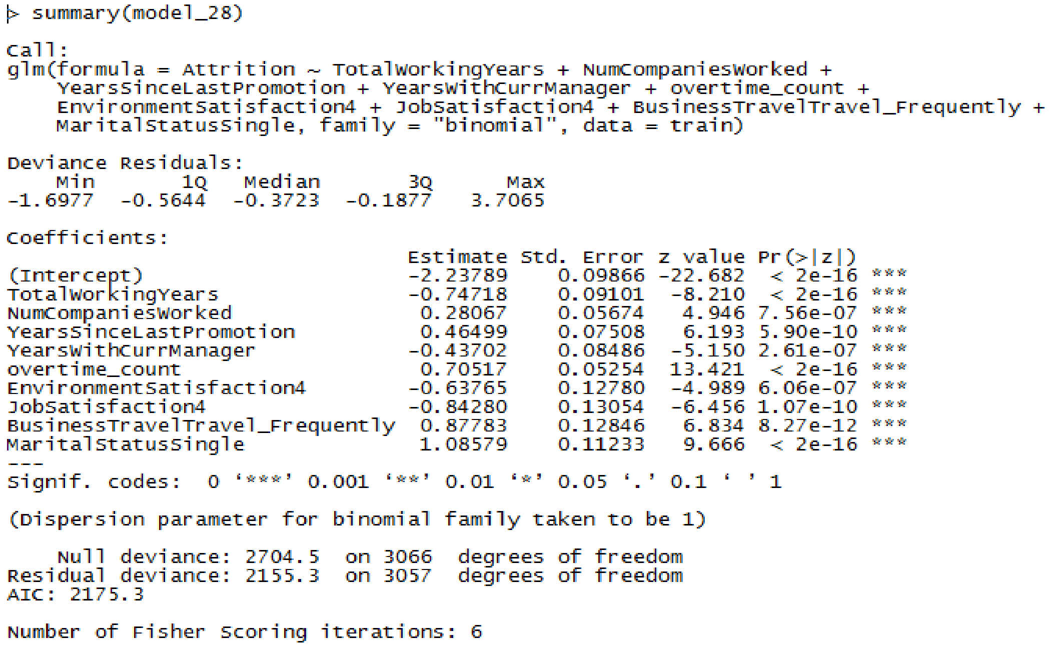
- ✓ Distance from home is not a factor affecting attrition
- ✓ attrition rate for employees working less than assigned working hour is quite small than those who are working overtime, though such overtime working employees are less.



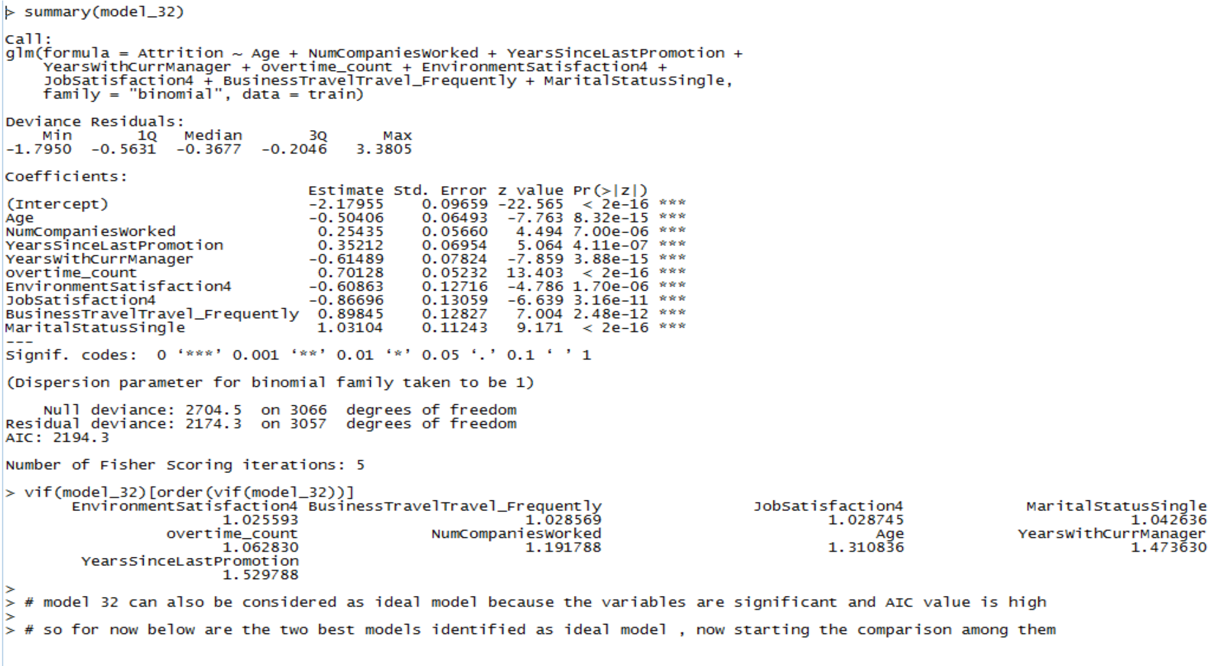


- ✓ Monthly income also not a major influencing factor for attrition

Model Building



```
> summary(model_28)

Call:
glm(formula = Attrition ~ TotalWorkingYears + NumCompaniesWorked +
    YearsSinceLastPromotion + YearsWithCurrManager + overtime_count +
    EnvironmentSatisfaction4 + JobSatisfaction4 + BusinessTravelTravel_Frequently +
    MaritalStatusSingle, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6977  -0.5644  -0.3723  -0.1877   3.7065

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -2.23789    0.09866 -22.682  < 2e-16 ***
TotalWorkingYears               -0.74718    0.09101  -8.210  < 2e-16 ***
NumCompaniesWorked               0.28067    0.05674   4.946 7.56e-07 ***
YearsSinceLastPromotion          0.46499    0.07508   6.193 5.90e-10 ***
YearsWithCurrManager            -0.43702    0.08486  -5.150 2.61e-07 ***
overtime_count                   0.70517    0.05254  13.421  < 2e-16 ***
EnvironmentSatisfaction4        -0.63765    0.12780  -4.989 6.06e-07 ***
JobSatisfaction4                -0.84280    0.13054  -6.456 1.07e-10 ***
BusinessTravelTravel_Frequently  0.87783    0.12846   6.834 8.27e-12 ***
MaritalStatusSingle              1.08579    0.11233   9.666  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2704.5  on 3066  degrees of freedom
Residual deviance: 2155.3  on 3057  degrees of freedom
AIC: 2175.3

Number of Fisher Scoring iterations: 6
```

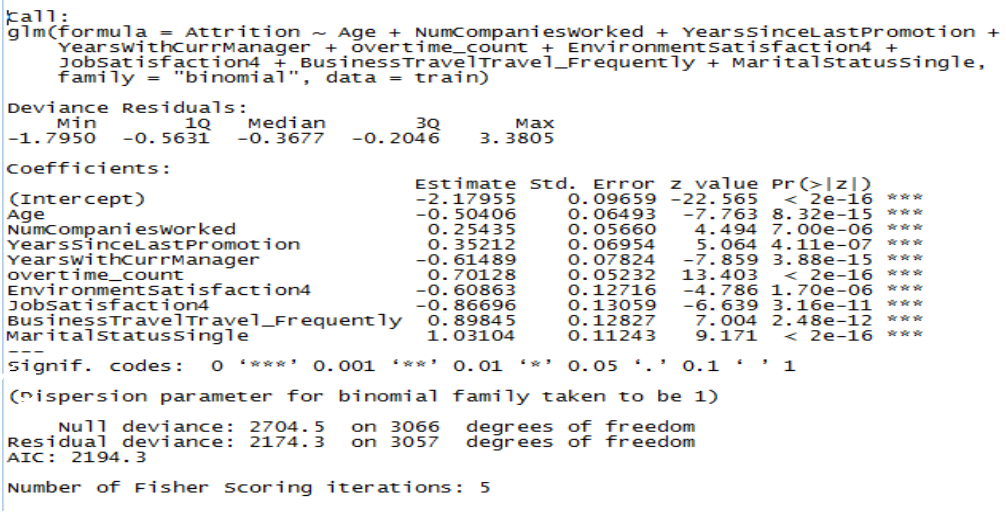


```
> summary(model_32)

Call:
glm(formula = Attrition ~ Age + NumCompaniesWorked + YearsSinceLastPromotion +
    YearsWithCurrManager + overtime_count + EnvironmentSatisfaction4 +
    JobSatisfaction4 + BusinessTravelTravel_Frequently + MaritalStatusSingle,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7950  -0.5631  -0.3677  -0.2046   3.3805

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -2.17955    0.09659 -22.565  < 2e-16 ***
Age                             -0.50406    0.06493  -7.763 8.32e-15 ***
NumCompaniesWorked               0.25435    0.05660   4.494 7.00e-06 ***
YearsSinceLastPromotion          0.35212    0.06954   5.064 4.11e-07 ***
YearsWithCurrManager            -0.61489    0.07824  -7.859 3.88e-15 ***
overtime_count                   0.70128    0.05232  13.403  < 2e-16 ***
EnvironmentSatisfaction4        -0.60863    0.12716  -4.786 1.70e-06 ***
JobSatisfaction4                -0.86696    0.13059  -6.639 3.16e-11 ***
BusinessTravelTravel_Frequently  0.89845    0.12827   7.004 2.48e-12 ***
MaritalStatusSingle              1.03104    0.11243   9.171  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2704.5  on 3066  degrees of freedom
Residual deviance: 2174.3  on 3057  degrees of freedom
AIC: 2194.3

Number of Fisher Scoring iterations: 5

> vif(model_32)[order(vif(model_32))]
       EnvironmentSatisfaction4 BusinessTravelTravel_Frequently                JobSatisfaction4             MaritalStatusSingle
                       1.025593                        1.028569                        1.028745                        1.042636
                 overtime_count              NumCompaniesWorked                             Age            YearsWithCurrManager
                       1.062830                        1.191788                        1.310836                        1.473630
        YearsSinceLastPromotion
                       1.529788
>
> # model 32 can also be considered as ideal model because the variables are significant and AIC value is high
>
> # so for now below are the two best models identified as ideal model , now starting the comparison among them
```

✓Based on VIF and p values model 28 is the most optimal model for consideration.

✓ Drilling down the analysis and adding and removing few variables and checking their effect on analysis model 32 is another optimal model for consideration.

Final Model Solution



```
Call:
glm(formula = Attrition ~ Age + NumCompaniesWorked + YearsSinceLastPromotion + 
    YearsWithCurrManager + overtime_count + EnvironmentSatisfaction4 + 
    JobSatisfaction4 + BusinessTravelTravel_Frequently + MaritalStatusSingle, 
    family = "binomial", data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7950  -0.5631  -0.3677  -0.2046   3.3805  

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                     -2.17955    0.09659 -22.565  < 2e-16 ***
Age                             -0.50406    0.06493  -7.763 8.32e-15 ***
NumCompaniesWorked               0.25435    0.05660   4.494 7.00e-06 ***
YearsSinceLastPromotion          0.35212    0.06954   5.064 4.11e-07 ***
YearsWithCurrManager            -0.61489    0.07824  -7.859 3.88e-15 ***
overtime_count                   0.70128    0.05232  13.403  < 2e-16 ***
EnvironmentSatisfaction4        -0.60863    0.12716  -4.786 1.70e-06 ***
JobSatisfaction4                -0.86696    0.13059  -6.639 3.16e-11 ***
BusinessTravelTravel_Frequently  0.89845    0.12827   7.004 2.48e-12 ***
MaritalStatusSingle              1.03104    0.11243   9.171  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2704.5  on 3066  degrees of freedom
Residual deviance: 2174.3  on 3057  degrees of freedom
AIC: 2194.3

Number of Fisher Scoring iterations: 5
```

Final Observations

✓model_32" CONTAINS THE VARIABLES "Age + NumCompaniesWorked + YearsSinceLastPromotion + YearsWithCurrManager + overtime_count + EnvironmentSatisfaction4 + JobSatisfaction4 + BusinessTravelTravel_Frequently + MaritalStatusSingle“

✓BIVARIATE ANALYSIS ALSO JUSTIFIES THE PRESENCE OF VARIABLE LIKE Age, MaritalStatusSingle, BusinessTravelTravel_Frequently etc.

Performance Measurement

✓Sensitivity for model_28 is 69.34% and for model_32 is 71.2%
✓Specificity for model_28 is 69.45% and for model_32 is 71.60%
✓Accuracy for model_28 is 69.40% and for model_32 is 71.56%
✓K-statistics for model_28 is 41.7%(3rd decile) and for model_32 is 44.5%(3rd decile)

Based on above statistics we can clearly see that sensitivity,specificity,Accuracy and k statistics it is clear that model_32 is the best model for our prediction case as it will correctly and more accurately predict each case of employee attrition.

Major Factors Influencing Attrition and recommendation

✓Employees who are <30 are more likely to switch jobs as per the trend . Company should focus on such employees and likely to concentrate on the more skilled ones and should provide them rewards and incentives to retain them.

✓The employees with higher job satisfaction will likely to stay in the job and continue. Company should have a job rotation policy

✓The trend shows that those who are travelling more frequently are likely to quit more as they may not be happy travelling often and company should focus on those employee and try to manage there work life balance.

✓The trend shows that those who are single are more likely to leave as they can easily relocate to other places if they have better offers in hand as they don't have to think about their family and children's school and other stuff

✓The more the duration of the employee's tenure with the current manager the less likely he/she will leave the job.

✓Company should focus on those employees who were not promoted for a long period of time. Company should look into their past performance appraisals and if they are good then they should have a discussion with their respective supervisors to understand the specific reasons of delay in promotion.

✓Those employees who are doing overtime and are spending more than 8 hours in office more often they are more likely to quit. It can be due to work pressure and improper work life balance. Management should focus on them and should ensure that the skill matrix and talent pool is competent so that the work load is balanced

✓For employees doing overtime there should be attractive incentive proposal.