



HR Analytics

DSC 424 – Advanced Data Analysis
(Winter 2020)



By – Andy Huang, Arun Gopal, Shweta Gujrathi, Sahana Nataraja



Agenda

- ☐ Predict factors influencing employee attrition at a company
- ☐ Predict monthly income of the employee

Dataset Introduction

Dataset name : HR Analytics

Dependent variables:

- Attrition – Binary variable
- Monthly income – Numeric variable

Number of independent variables:

- Numeric Variables – 9
- Categorical Variables – 7
- Ordinal Variables - 8

Number of Observations: 4410

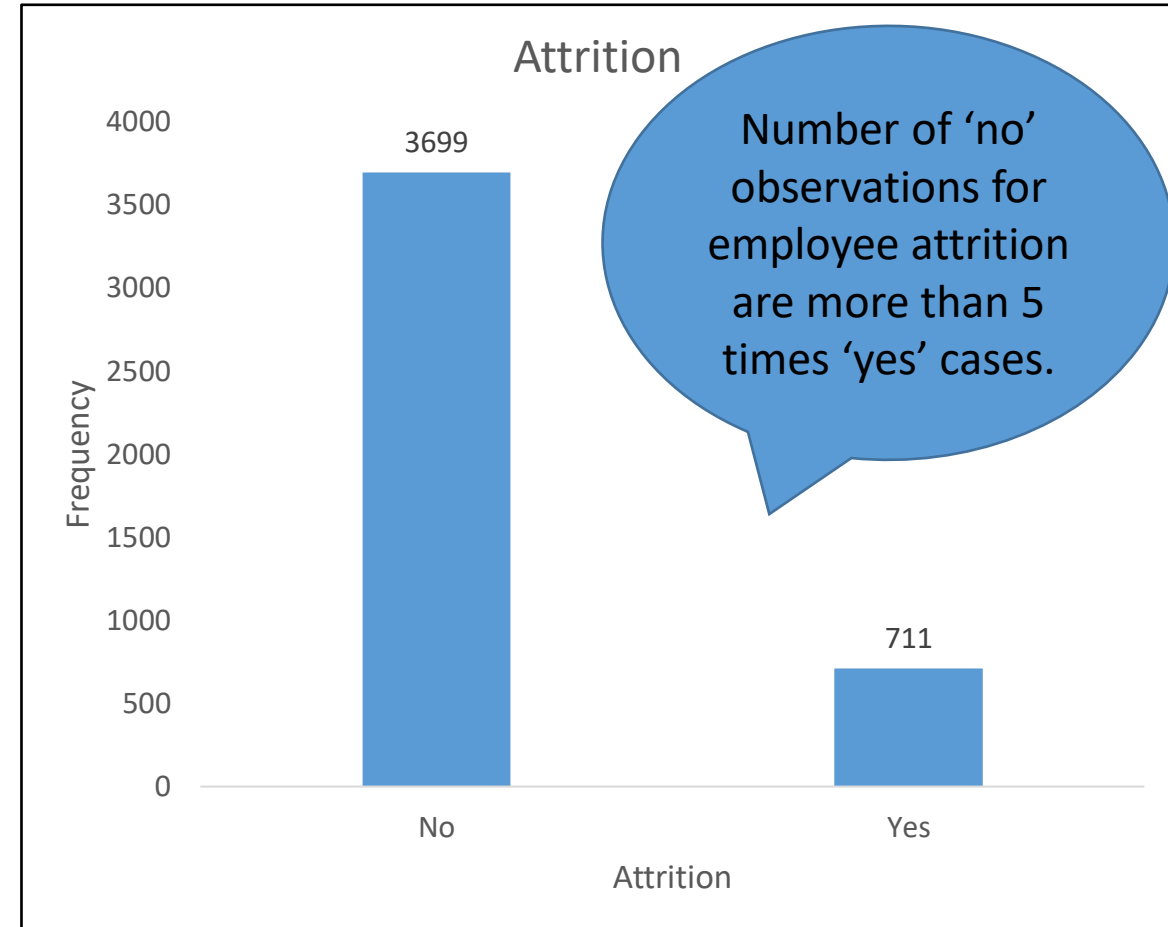
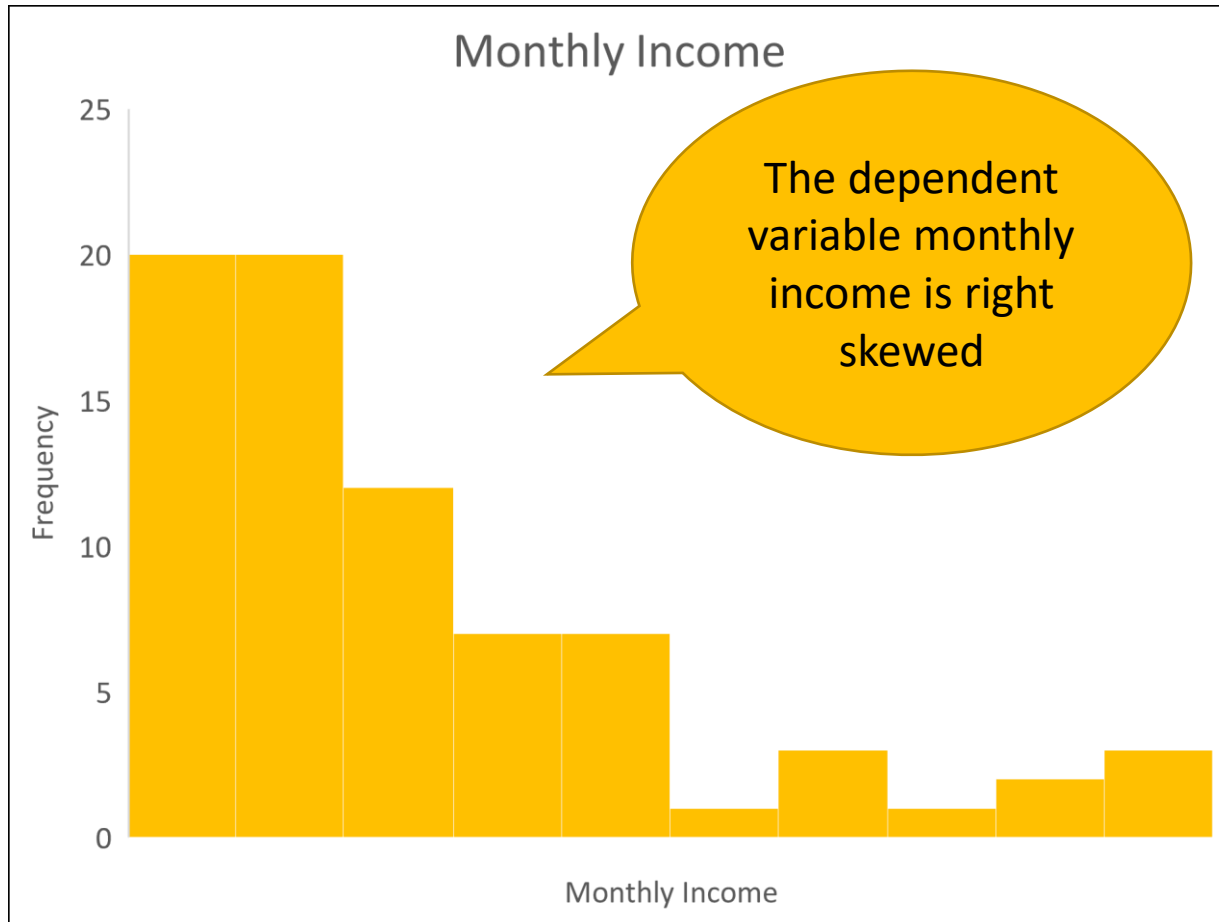
Independent Variables

Numeric	Categorical	Ordinal
Age	Attrition	Education level
Distance from home	Business Travel	Job Involvement
% salary hike	Department	Performance Rating
Training Times Last Year	Education Field	Job Level
Years at Company	Gender	Stock Option Level
Years since last promotion	Job Role	Environmental Satisfaction
Years with current manager	Marital Status	Job Satisfaction
Total working hours		Work-Life Balance
Number of companies worked		

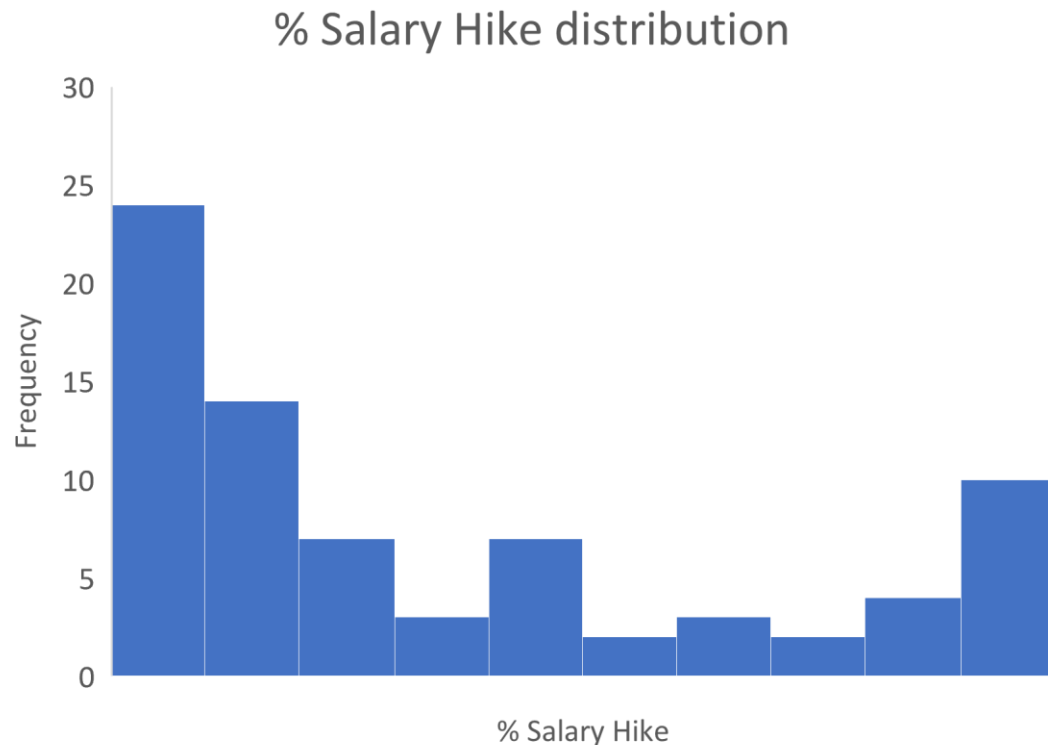
Exploratory Analysis



Distribution of dependent variables – Monthly Income and Employee Attrition



% Salary Hike

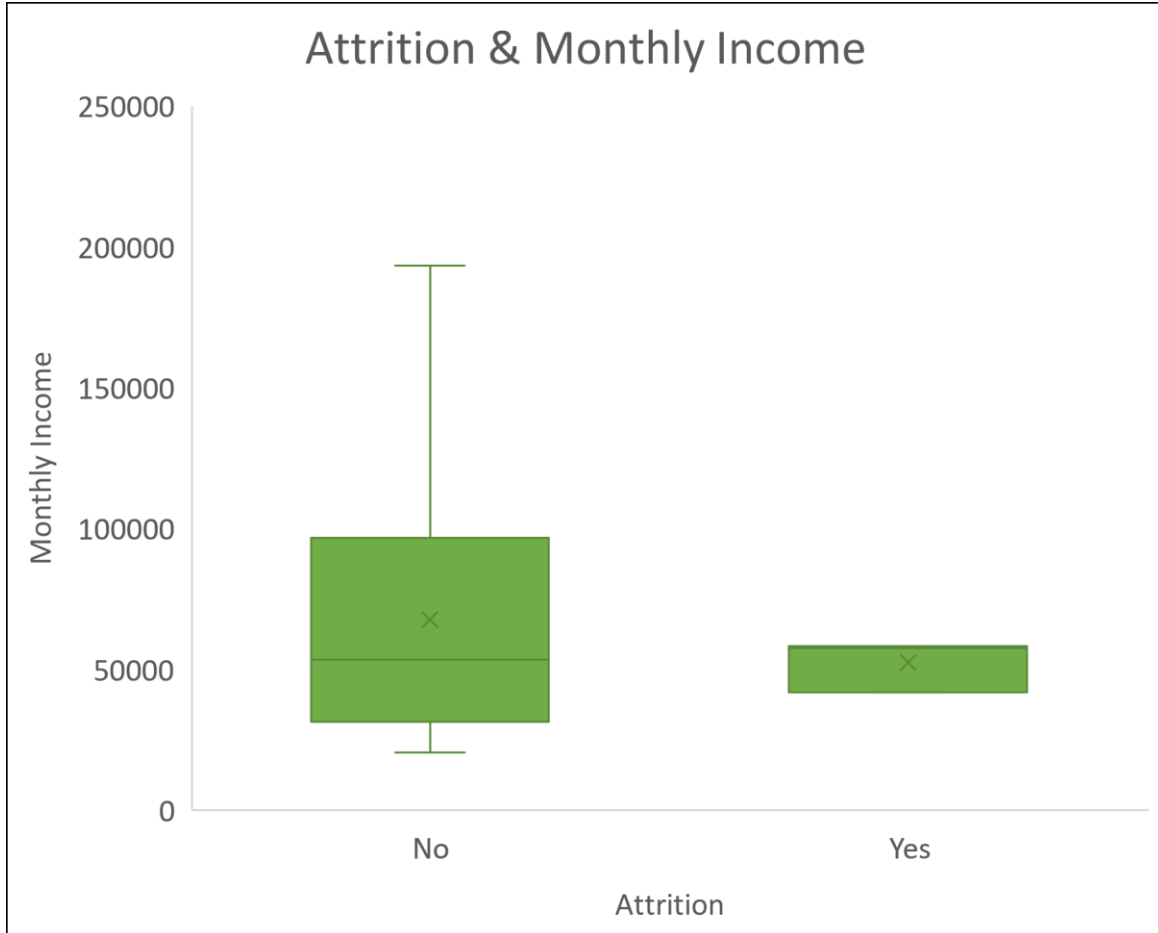


- The data for % Salary Hike is right skewed, thus needing log transformation in linear and logistic regression to normalize the data

Attrition - Monthly Income & % Salary Hike

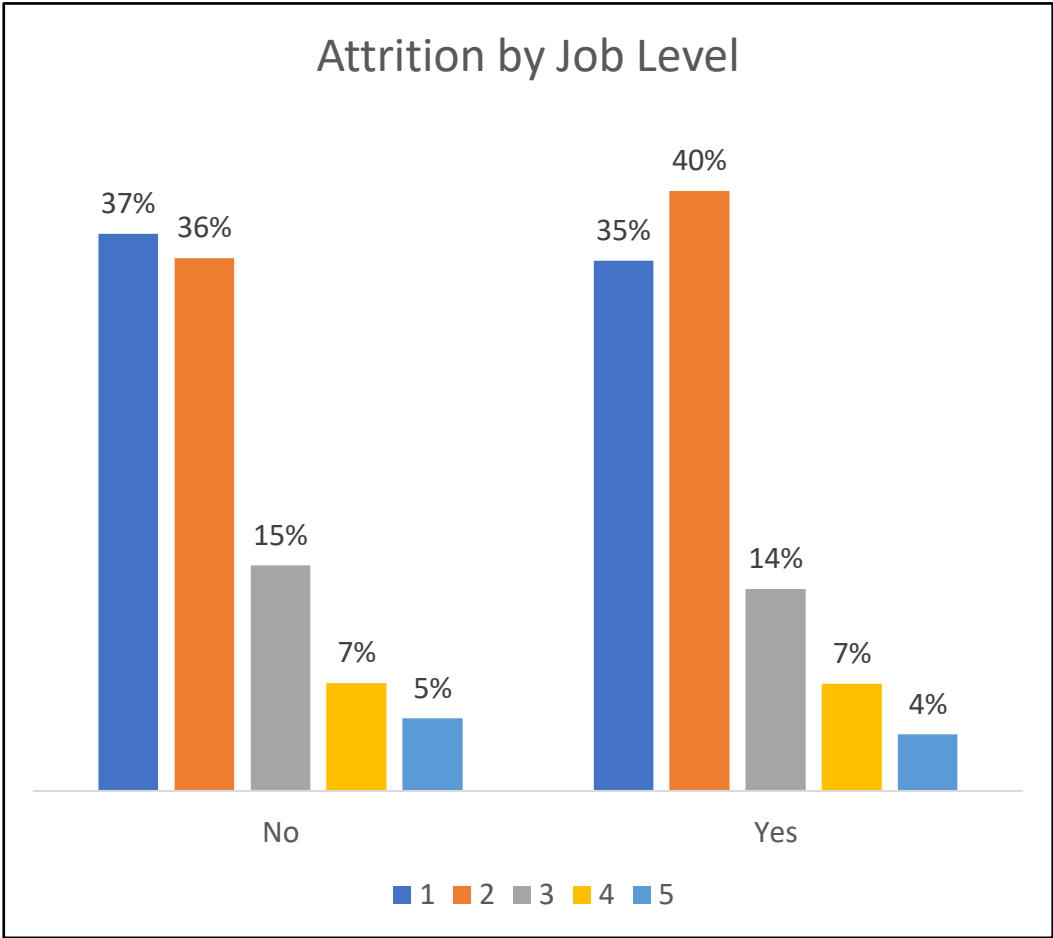
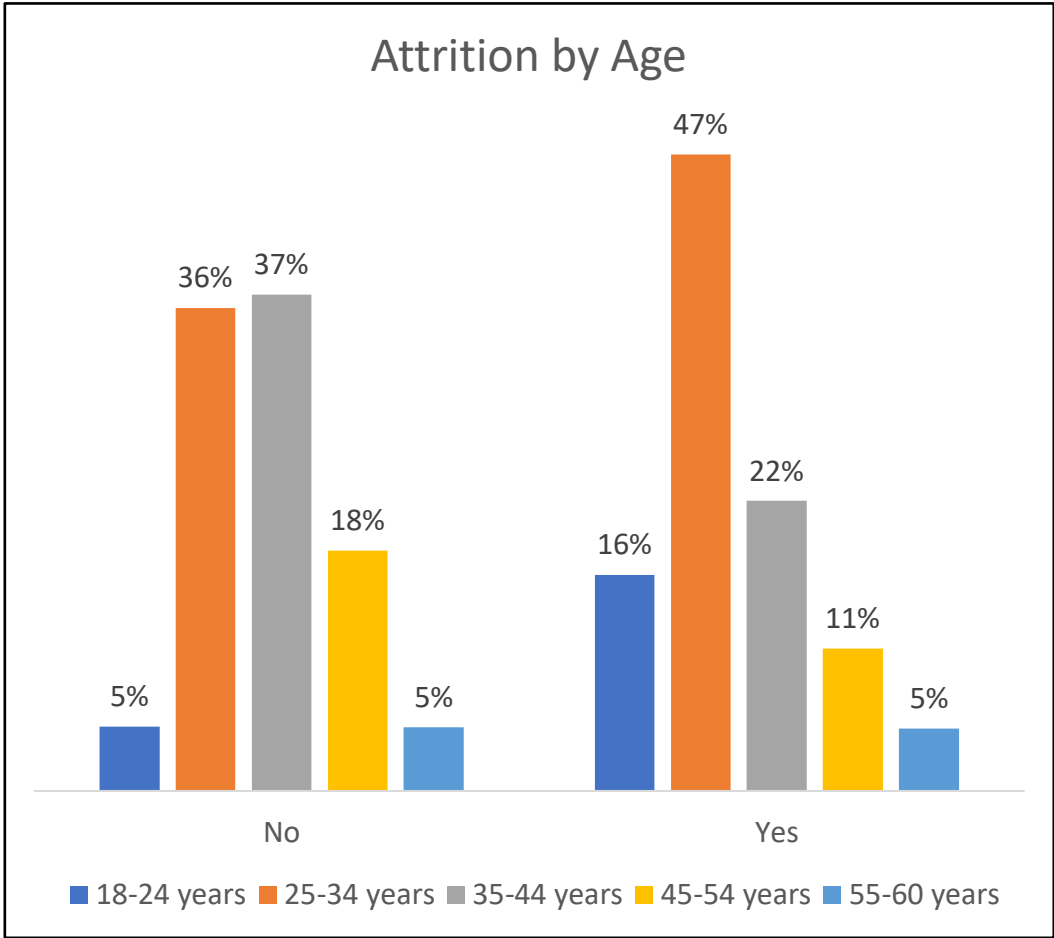


Employees with lower salaries tend to leave the company even if they are given much higher % salary hikes than the employees who are not leaving the company



Attrition - Age & Job Level

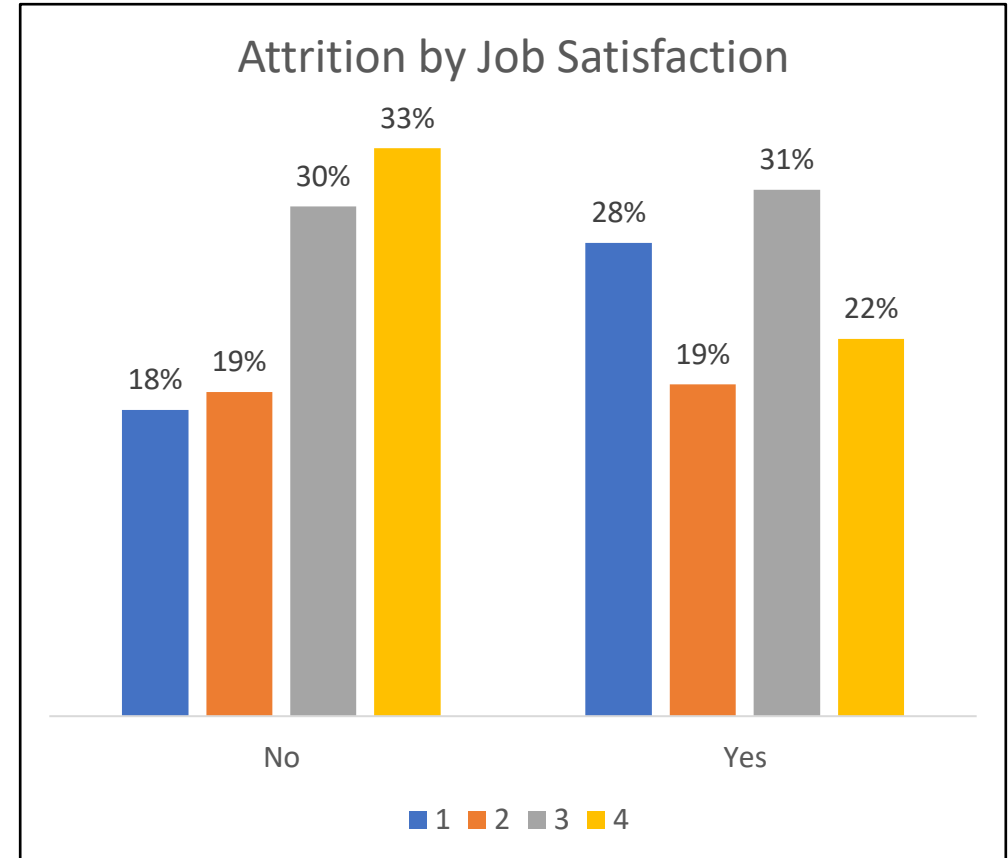
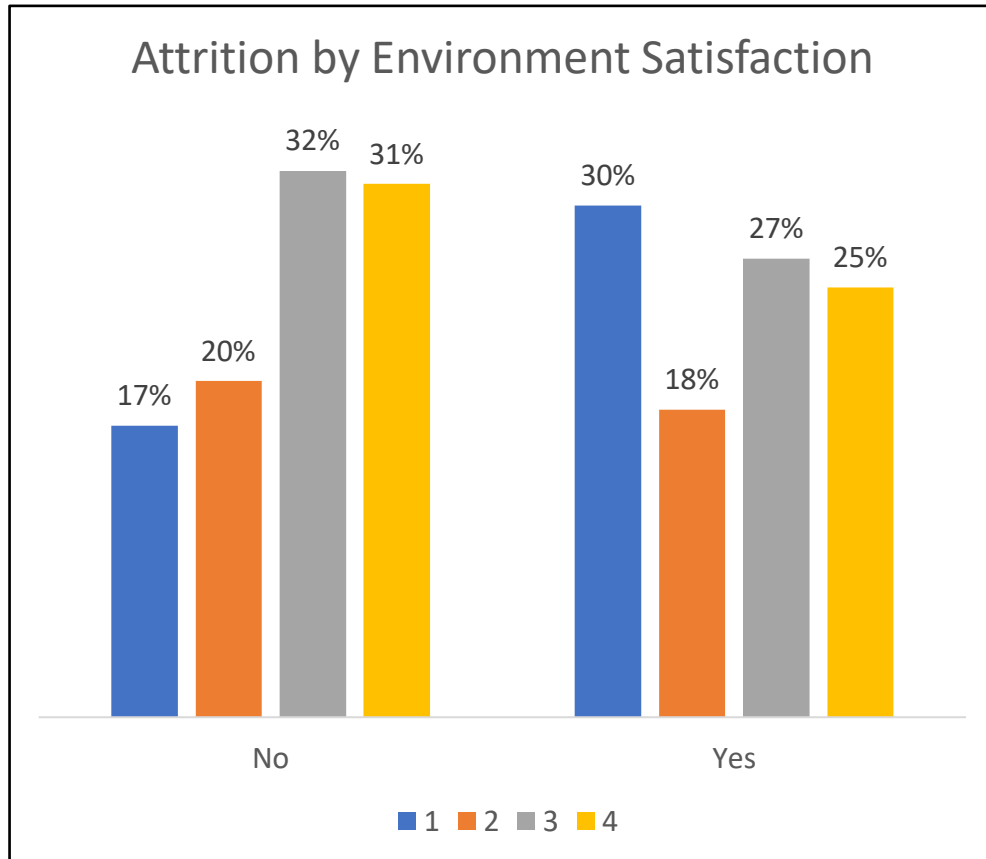
Out of the people who left the company, 47% were from the 25-34 age group. Additionally, attrition was high in employees in job levels 1 & 2



Attrition - Environment & Job Satisfaction

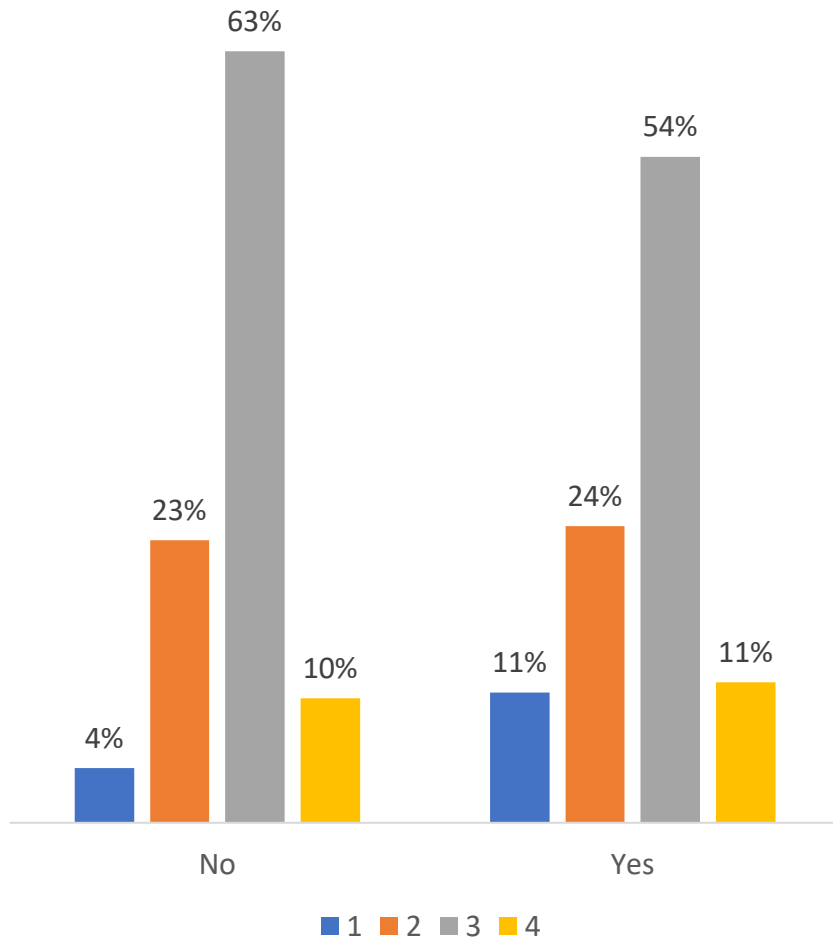


Environment and Job Satisfaction doesn't differ significantly between employees who left the company and the ones who stayed.



Attrition and Work-Life Balance

Attrition by Work-Life Balance

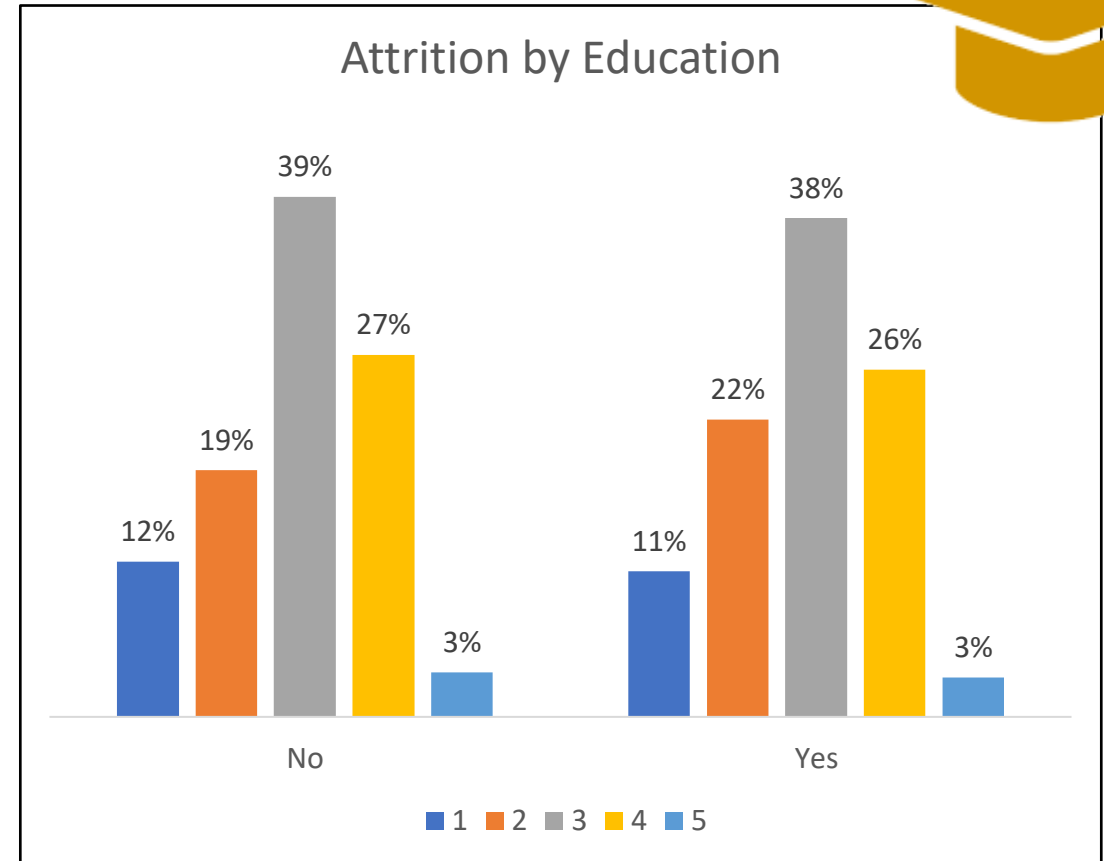
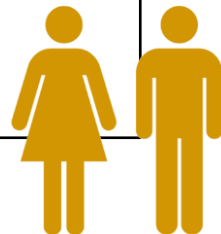
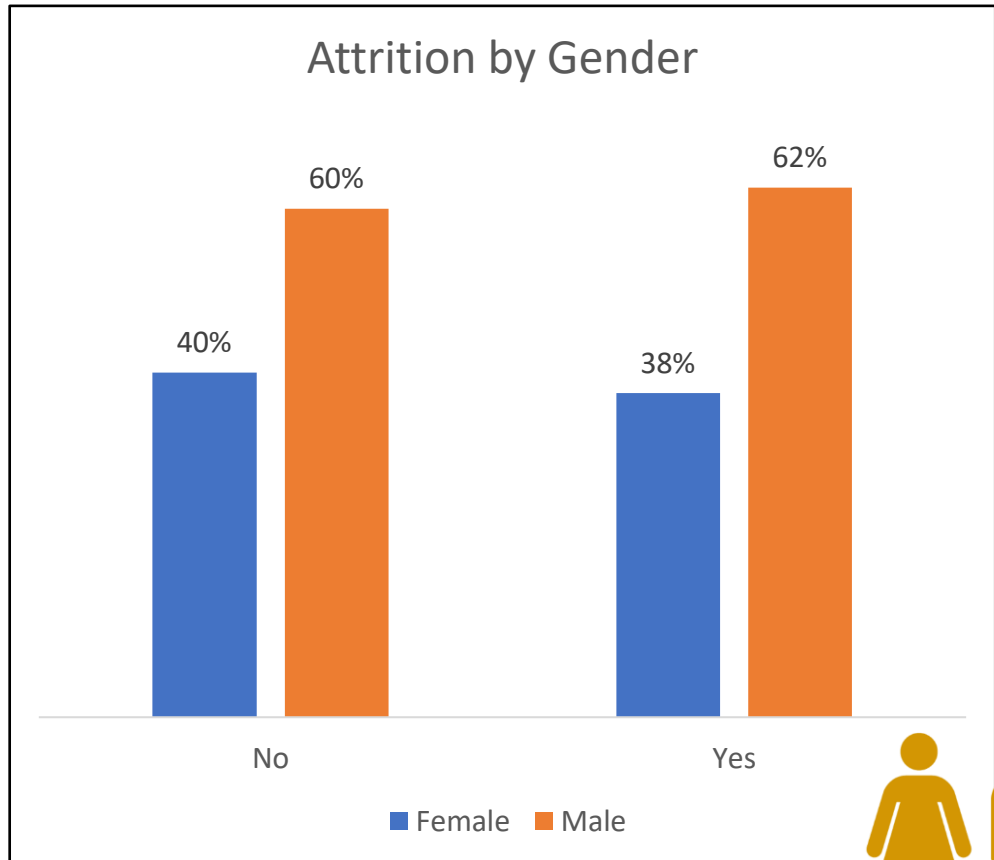


- Amongst the employees who left the company, almost 65% reported high work-life balance in the survey.



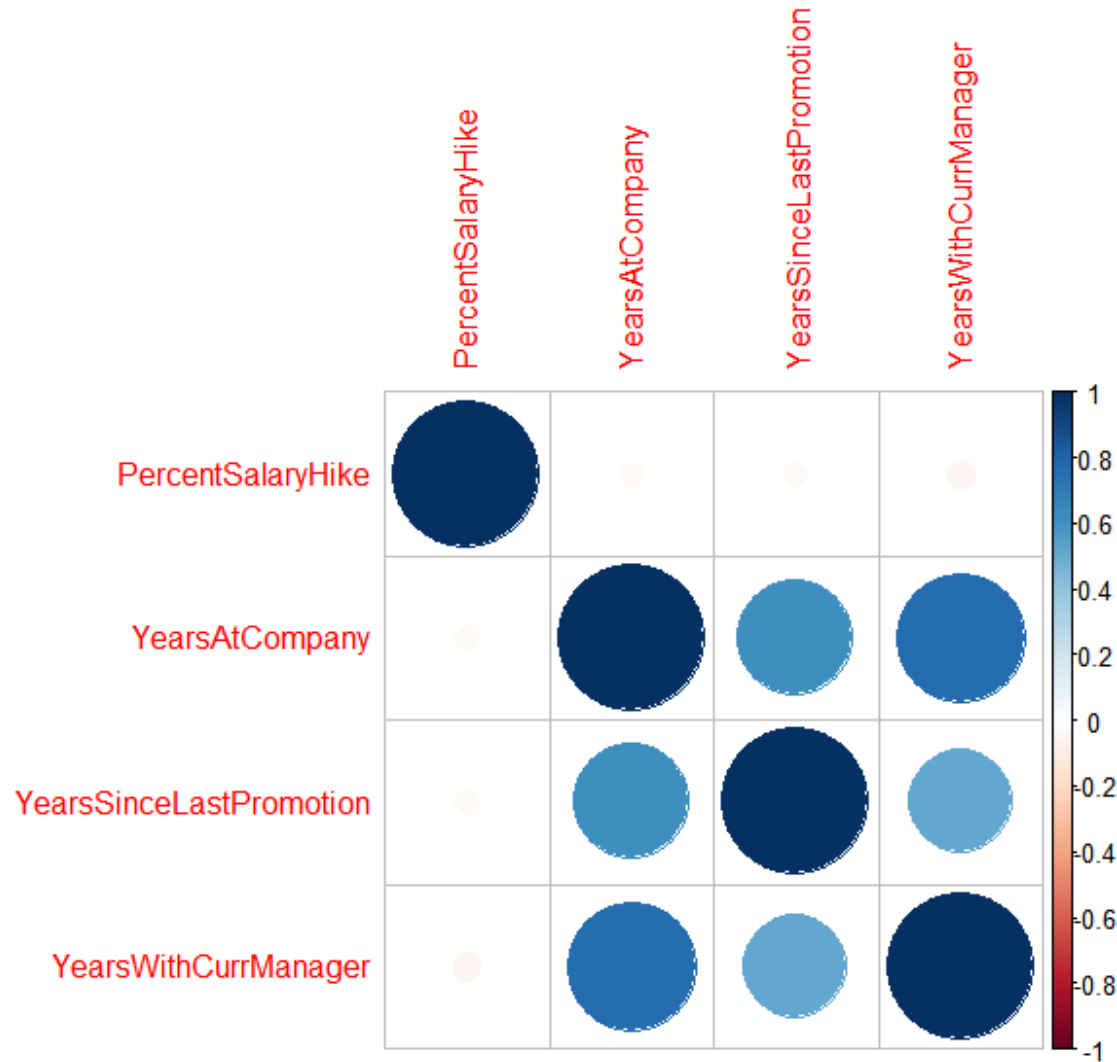
Attrition by Gender and Education

Gender and Education Level is not playing a significant role in attrition



Correlation

- Out of 15 numeric variables, only 4 variables have high correlation between them.



Data Analysis Techniques

- Linear and Logistic Regression
- Principal Component Analysis
- Factor Analysis
- Ordinal Factor Analysis
- Correspondence Analysis
- Partial Least Square Regression

Ordinary Least Square Regression

```
Call:
lm(formula = loglp(monthlyIncome) ~ ., data = hr_num_ord)
```

Residuals:

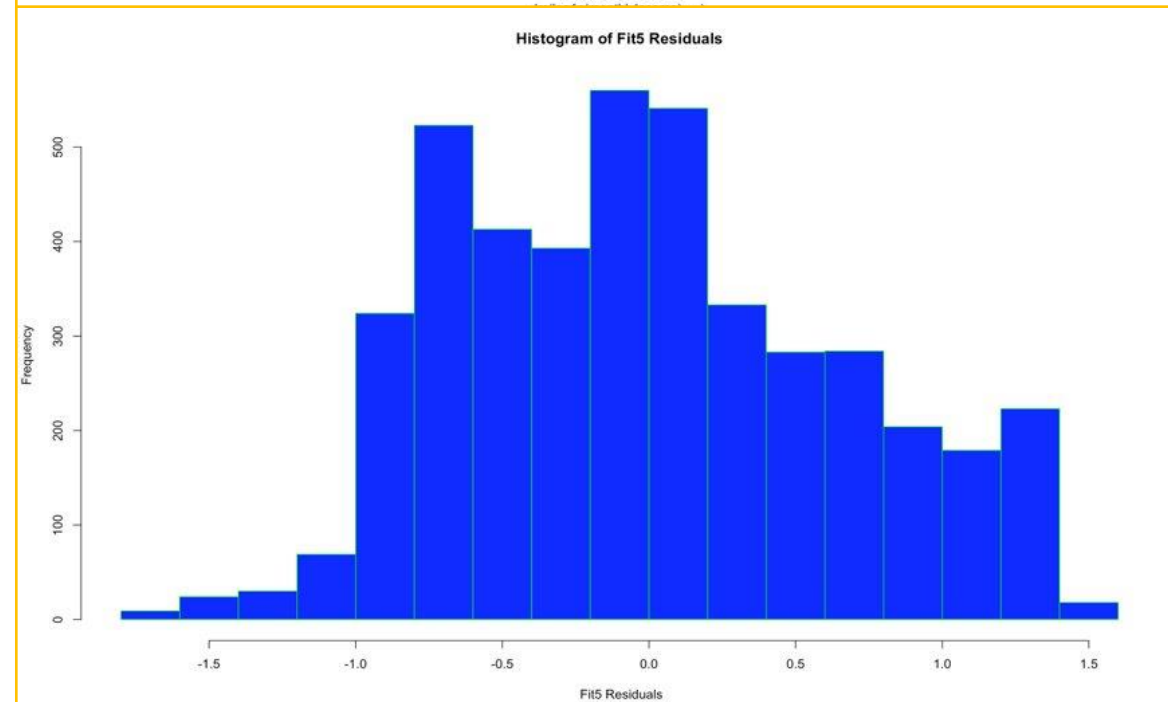
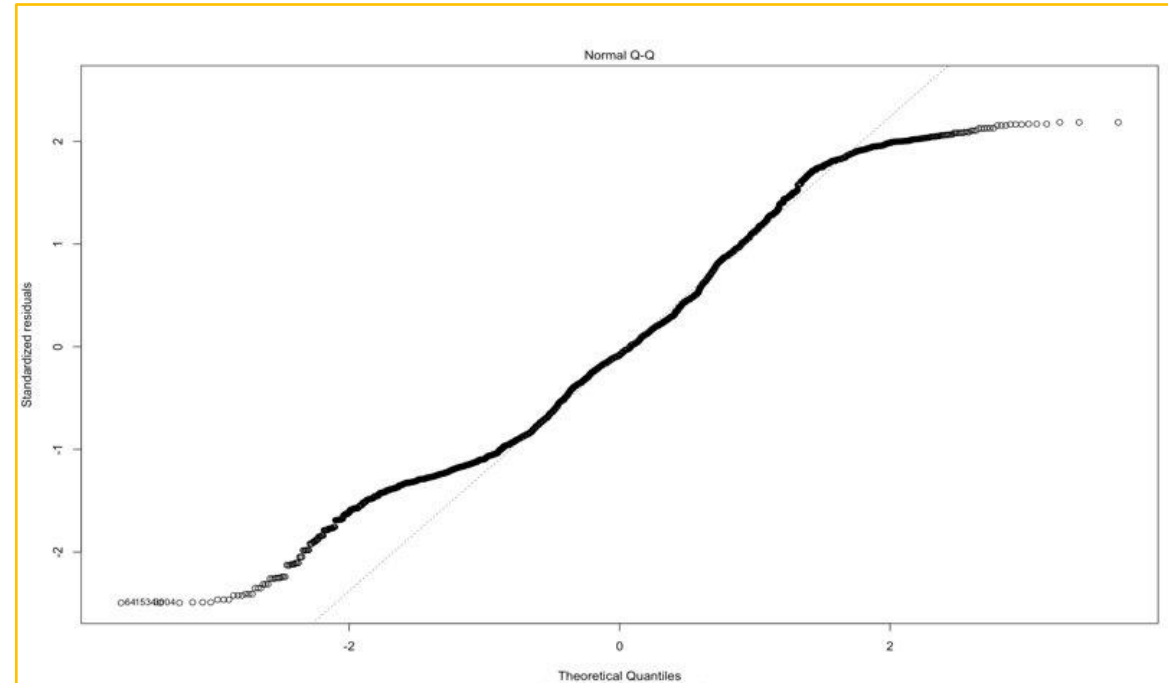
Min	1Q	Median	3Q	Max
-1.64337	-0.55968	-0.05147	0.46757	1.44029

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	10.5744741	0.1346718	78.520	< 2e-16	***
Age	-0.0014122	0.0015332	-0.921	0.35707	
DistanceFromHome	-0.0014231	0.0012306	-1.156	0.24756	
PercentSalaryHike	-0.0009915	0.0043149	-0.230	0.81826	
TrainingTimesLastYear	0.0196607	0.0077636	2.532	0.01136	*
YearsAtCompany	-0.0049731	0.0032141	-1.547	0.12186	
YearsSinceLastPromotion	0.0194486	0.0039521	4.921	8.92e-07	***
YearsWithCurrManager	0.0065376	0.0043924	1.488	0.13672	
TotalWorkingYears	-0.0035097	0.0022229	-1.579	0.11443	
NumCompaniesWorked	-0.0039066	0.0044304	-0.882	0.37795	
Education	-0.0019660	0.0097562	-0.202	0.84031	
JobInvolvement	0.0191982	0.0140074	1.371	0.17058	
PerformanceRating	0.0442362	0.0437304	1.012	0.31180	
JobLevel	0.0291194	0.0090443	3.220	0.00129	**
StockOptionLevel	0.0273503	0.0117717	2.323	0.02020	*
EnvSat	-0.0021381	0.0091501	-0.234	0.81525	
JobSat	-0.0011768	0.0090726	-0.130	0.89680	
WrkLifBal	0.0225428	0.0141773	1.590	0.11189	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6604 on 4392 degrees of freedom
Multiple R-squared: 0.01547, Adjusted R-squared: 0.01166
F-statistic: 4.06 on 17 and 4392 DF, p-value: 3.75e-08



```
Call:
lm(formula = log1p(monthlyIncome) ~ YearsSinceLastPromotion +
    TotalWorkingYears + JobLevel + TrainingTimesLastYear + StockOptionLevel +
    WrkLifBal, data = hr_num_ord)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.62038	-0.55254	-0.04896	0.46368	1.41248

Coefficients:

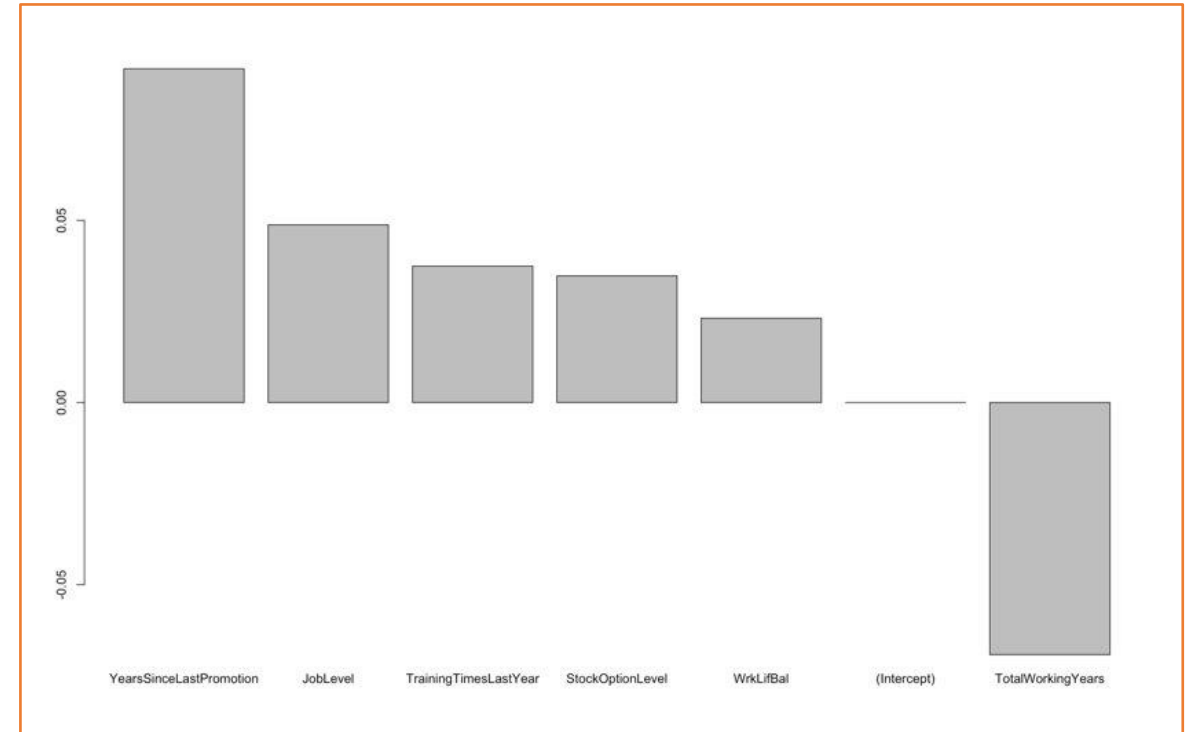
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.683837	0.054118	197.418	< 2e-16 ***
YearsSinceLastPromotion	0.018890	0.003381	5.588	2.44e-08 ***
TotalWorkingYears	-0.005900	0.001397	-4.223	2.46e-05 ***
JobLevel	0.029284	0.009011	3.250	0.00116 **
TrainingTimesLastYear	0.019298	0.007751	2.490	0.01282 *
StockOptionLevel	0.027119	0.011708	2.316	0.02059 *
WrkLifBal	0.021852	0.014146	1.545	0.12249

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6603 on 4403 degrees of freedom

Multiple R-squared: 0.01334, Adjusted R-squared: 0.012

F-statistic: 9.922 on 6 and 4403 DF, p-value: 6.668e-11



OLS – Feature Selection

- Forward, Backward, and Stepwise all have chosen the same variables
- Variable Chosen: Years Since Last Promotion, Total Working Years, Job Level, Training Times Last Year, Stock Option Level, Work Life Balance
- Total Working Years – the only predictor with negative sign

Logistic Regression model with numeric variables

- Logistic Regression Model with just the numeric predictors has the accuracy of 83% on the test set.
- The model has a chi-square of 227 with 8 df and p-value < 0.05, thus proving it is a better fit than an empty model.
- The significant predictors as per chi-square in the order of importance are –
 - Age
 - Distance from home
 - Training Times since last year
 - Years since last promotion
 - Years with current manager
 - Total Working Years
 - Number of companies worked
 - Log (Monthly Income)

```
Call:
glm(formula = Attrition ~ Age + DistanceFromHome + TrainingTimesLastYear +
    YearsSinceLastPromotion + YearsWithCurrManager + TotalWorkingYears +
    NumCompaniesWorked + LogMI, family = "binomial", data = train)

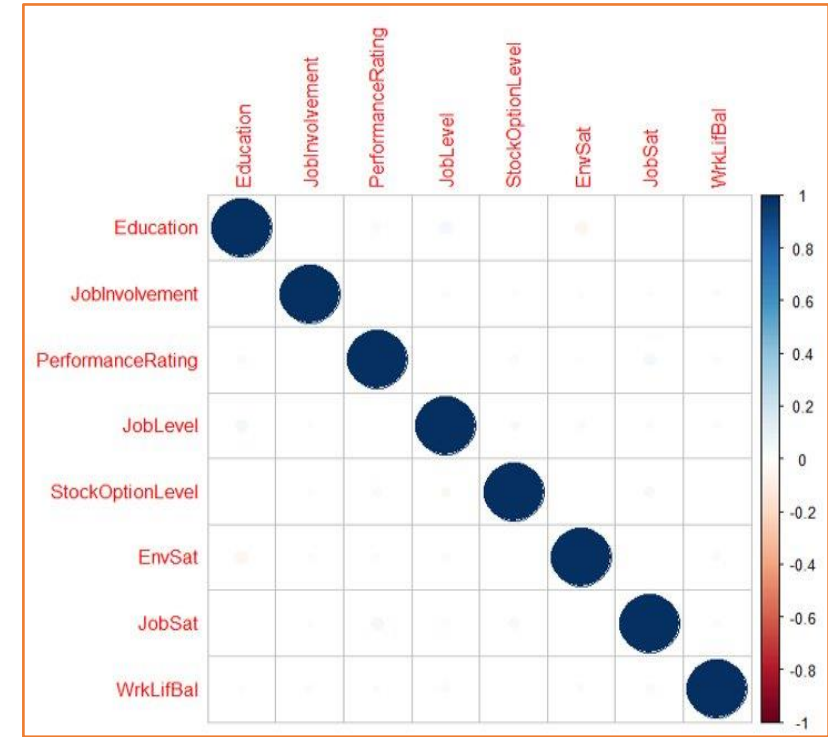
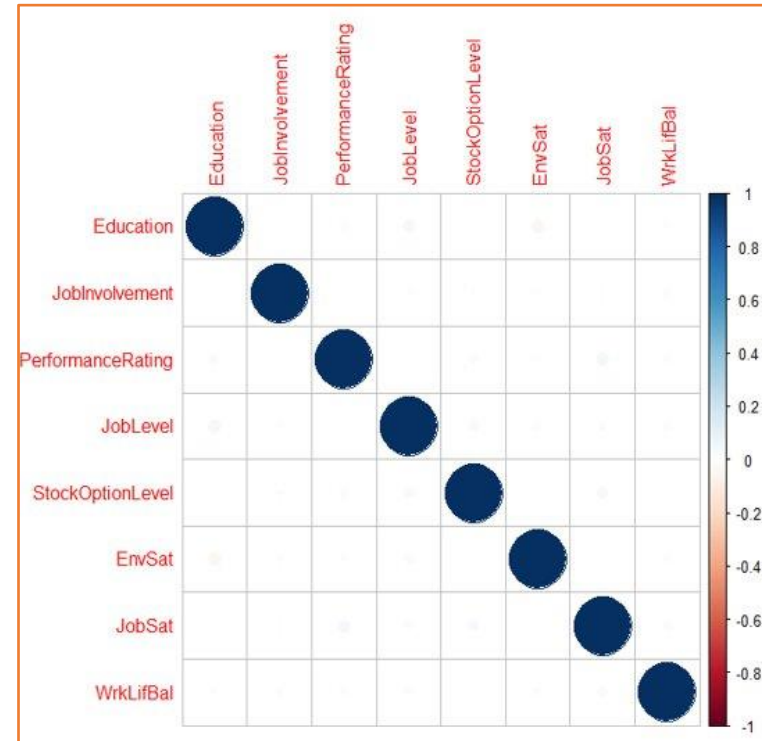
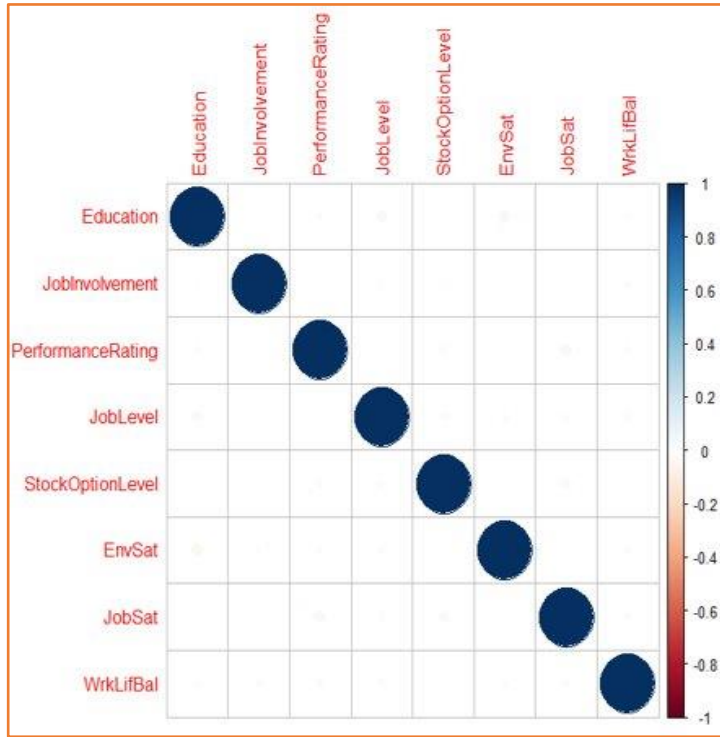
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2088  -0.6467  -0.4781  -0.3037   3.0474

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.449523   0.884455   2.770 0.005614 **
Age          -0.040847   0.008088  -5.050 4.41e-07 ***
DistanceFromHome -0.009341  0.006538  -1.429 0.153068
TrainingTimesLastYear -0.140496  0.041242  -3.407 0.000658 ***
YearsSinceLastPromotion 0.116342  0.021641   5.376 7.62e-08 ***
YearsWithCurrManager -0.119393  0.021635  -5.519 3.42e-08 ***
TotalWorkingYears -0.051877  0.012497  -4.151 3.31e-05 ***
NumCompaniesWorked  0.128307  0.020594   6.230 4.66e-10 ***
LogMI         -0.171202  0.077456  -2.210 0.027084 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Principal Component Analysis

Loadings:				
		RC1	RC2	RC3
Age			0.80	
DistanceFromHome				0.87
PercentSalaryHike				0.52
TrainingTimesLastYear				
YearsAtCompany		0.93		
YearsSinceLastPromotion		0.76		
YearsWithCurrManager		0.86		
TotalWorkingYears		0.61	0.66	
NumCompaniesWorked			0.77	
		RC1	RC2	RC3
SS loadings	2.712	1.682	1.037	1.030
Proportion Var	0.301	0.187	0.115	0.114
Cumulative Var	0.301	0.488	0.603	0.718

- The initial principal factor analysis with only numeric variables resulted in the following factors.
- Overall 71% variance could be captured .
- The factor loadings gave us a clear idea about the similarity of variables.



Ordinal Factor Analysis

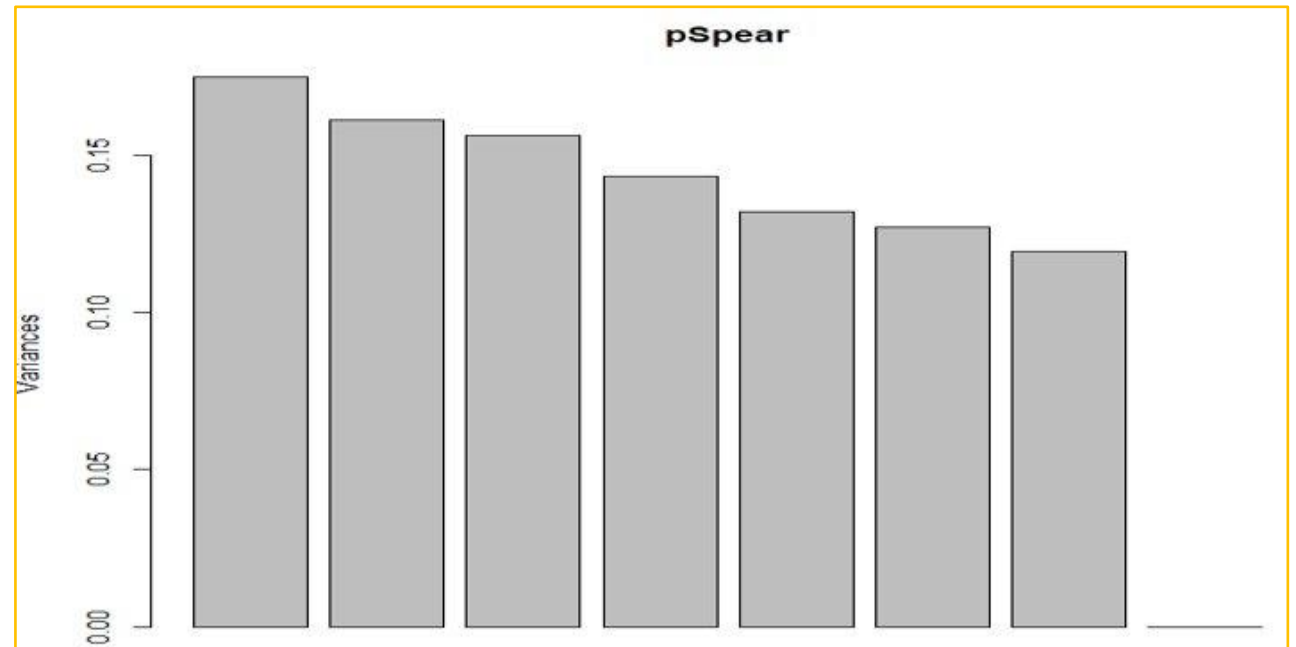
- We used Pearson, Spearman and Kendall methods to find correlations between Ordinal features and perform Factor analysis
- There are no correlation between Ordinal features

Ordinal Factor Analysis

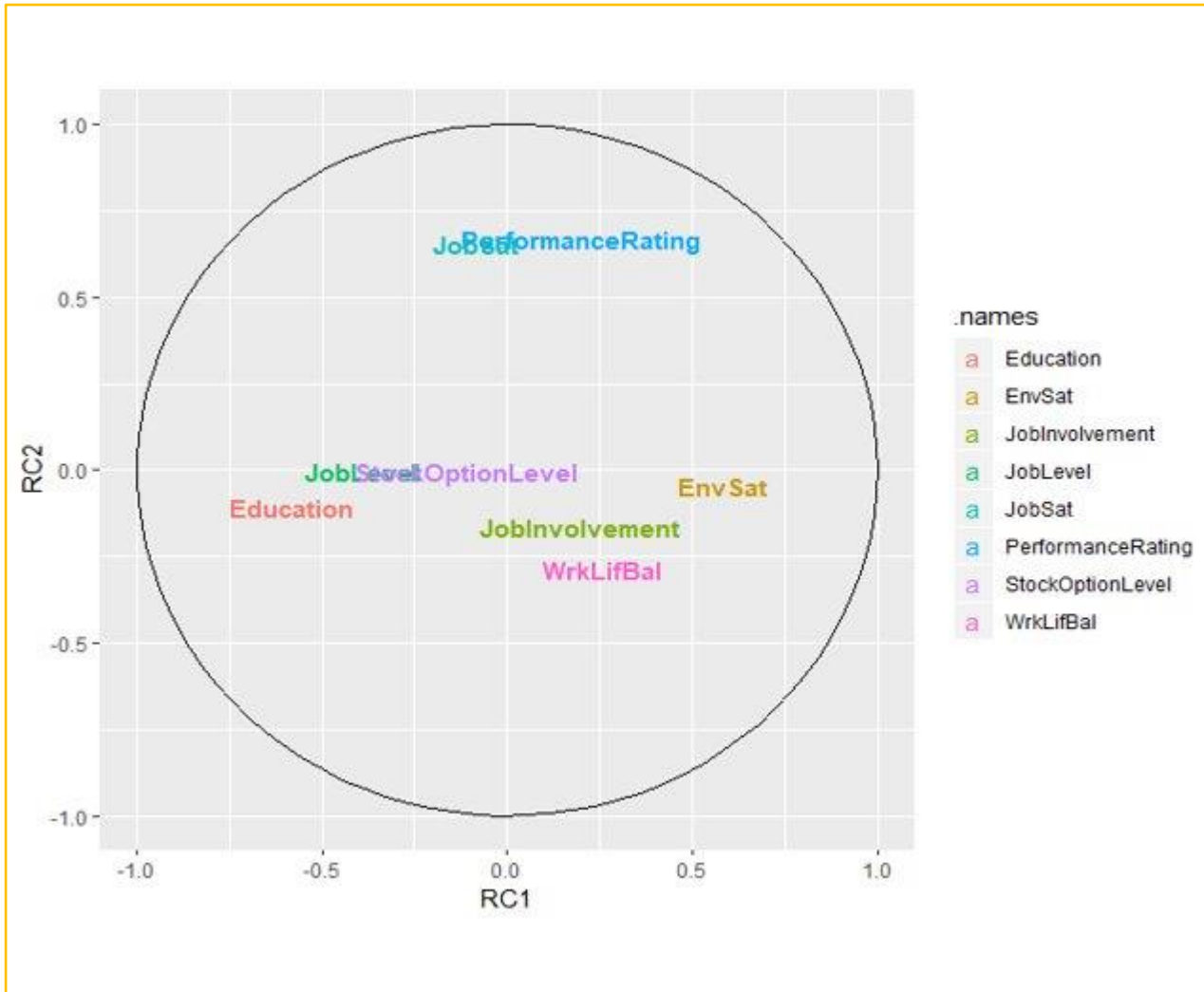
Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.1277	1.1044	1.0797	1.0117	0.8509	0.71445	0.69849	0.3596
Proportion of Variance	0.1947	0.1867	0.1785	0.1567	0.1108	0.07814	0.07469	0.0198
Cumulative Proportion	0.1947	0.3814	0.5599	0.7166	0.8274	0.90552	0.98020	1.0000

- PRComp summary indicate 6 components are required to account of 90% variance in data
- Scree plot does not show a clear knee, indicating more features are needed to account to 90% variance



Ordinal Features – Psych plot




```
> print(pspear2$loadings, cutoff=.4)
```

Loadings:

	RC1	RC3	RC2	RC4
Education	-0.607			
JobInvolvement				0.826
PerformanceRating			0.689	
JobLevel	-0.407	-0.497		
StockOptionLevel		0.734		
EnvSat	0.614			
JobSat			0.671	
WrkLifBal				-0.566


	RC1	RC3	RC2	RC4
SS loadings	1.094	1.057	1.055	1.022
Proportion var	0.137	0.132	0.132	0.128
Cumulative var	0.137	0.269	0.401	0.529

- Using 4 components with rotation using “Spearman” we get some clear groupings



Ordinal Features Vs. Dependent Variables

- Attrition
 - Job Satisfaction
 - Environment Satisfaction
 - Work Life Balance
 - Monthly Income
 - Job Level
-



PCA with Numeric and Ordinal Features

- PCA by including both numeric and ordinal variables resulted in the following factors which could capture 50% of variance in the data.
 - The factor loadings were close to what we saw using numeric variables only. Here the factors have additional contributions from the ordinal variables.
 - The interpretation of the variable contribution to the factors were very clear.
-

PCA with Numeric and Ordinal Features

- Factor loadings
 - Factor 1 – Experience With Same Company
 - Factor 2 – Overall. Experience
 - Factor 3 – Job. Satisfaction
 - Factor 4 – Environmental. Satisfaction

Loadings:

	RC1	RC2	RC3	RC4
Age		0.79		
DistanceFromHome				0.57
PercentSalaryHike			0.68	
TrainingTimesLastYear			-0.44	
YearsAtCompany	0.93			
YearsSinceLastPromotion	0.76			
YearsWithCurrManager	0.86			
TotalWorkingYears	0.63	0.65		
NumCompaniesWorked		0.78		
JobLevel				-0.59
EnvSat				0.51
JobSat			0.43	
WrkLifBal				

	RC1	RC2	RC3	RC4
SS loadings	2.751	1.688	1.077	1.076
Proportion Var	0.212	0.130	0.083	0.083
Cumulative Var	0.212	0.341	0.424	0.507

PFA with numeric and ordinal features – Including monthly income

Loadings:				
	RC1	RC2	RC3	RC4
Age		0.78		
DistanceFromHome			0.49	
PercentSalaryHike				-0.59
TrainingTimesLastYear				0.52
YearsAtCompany	0.93			
YearsSinceLastPromotion	0.76			
YearsWithCurrManager	0.86			
TotalWorkingYears	0.63	0.64		
NumCompaniesWorked		0.77		
JobLevel			-0.60	
EnvSat				
JobSat				-0.43
WrkLifBal				
monthlyIncome			-0.48	
	RC1	RC2	RC3	RC4
SS loadings	2.752	1.667	1.083	1.085
Proportion Var	0.197	0.119	0.077	0.078
Cumulative Var	0.197	0.316	0.393	0.471

- Monthly income was included in the analysis of factors and this did not change much.
- Env. Satisfaction was replaced by monthly income in the factor which we named env. Satisfaction.
- Now looking at these variables, the factor actually seems to be explaining the job level of the employee.

```
Call:
lm(formula = log1p(monthlyIncome) ~ Overall_Exp + Exp_with_Company +
    Env_Sat, data = ols_data)

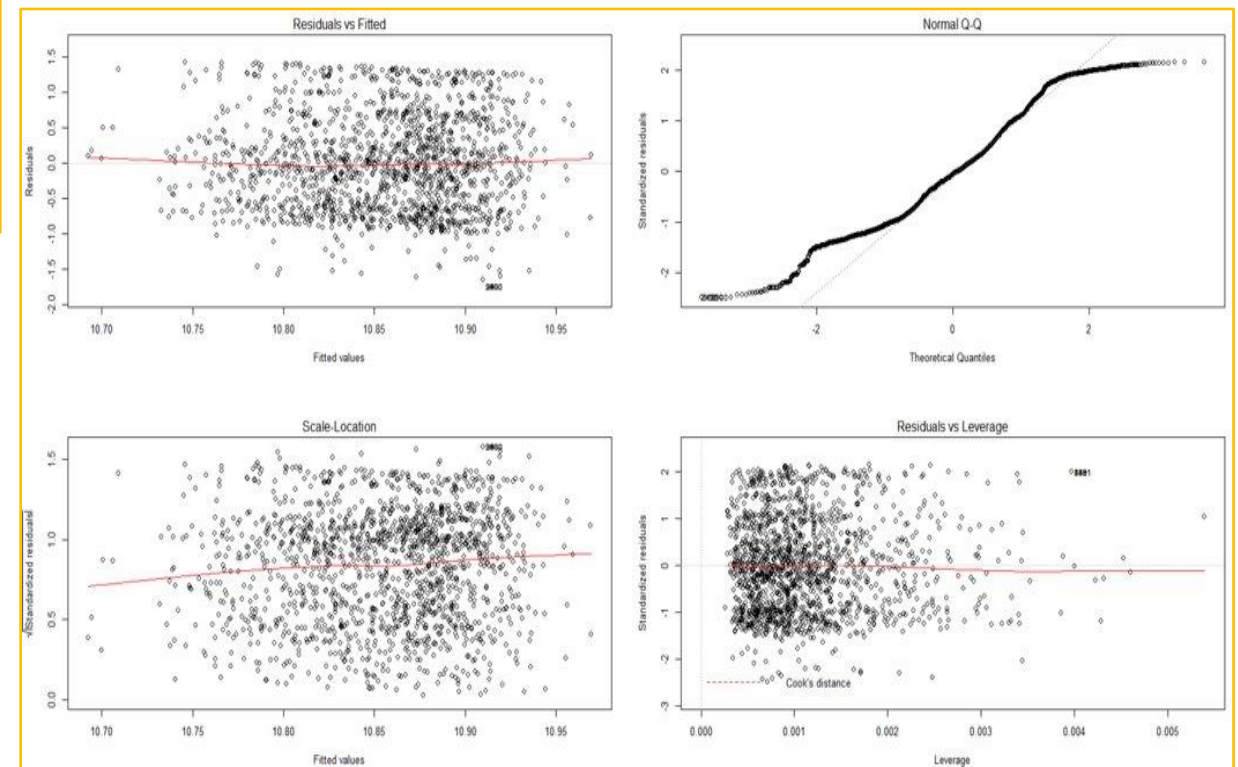
Residuals:
    Min       1Q   Median       3Q      Max
-1.64104 -0.57169 -0.04738  0.45972  1.42385

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.855126   0.009984 1087.253 < 2e-16 ***
Overall_Exp   -0.037966   0.009985  -3.802 0.000145 ***
Exp_with_Company 0.017798   0.009985   1.782 0.074747 .
Env_Sat      -0.015081   0.009987  -1.510 0.131078
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

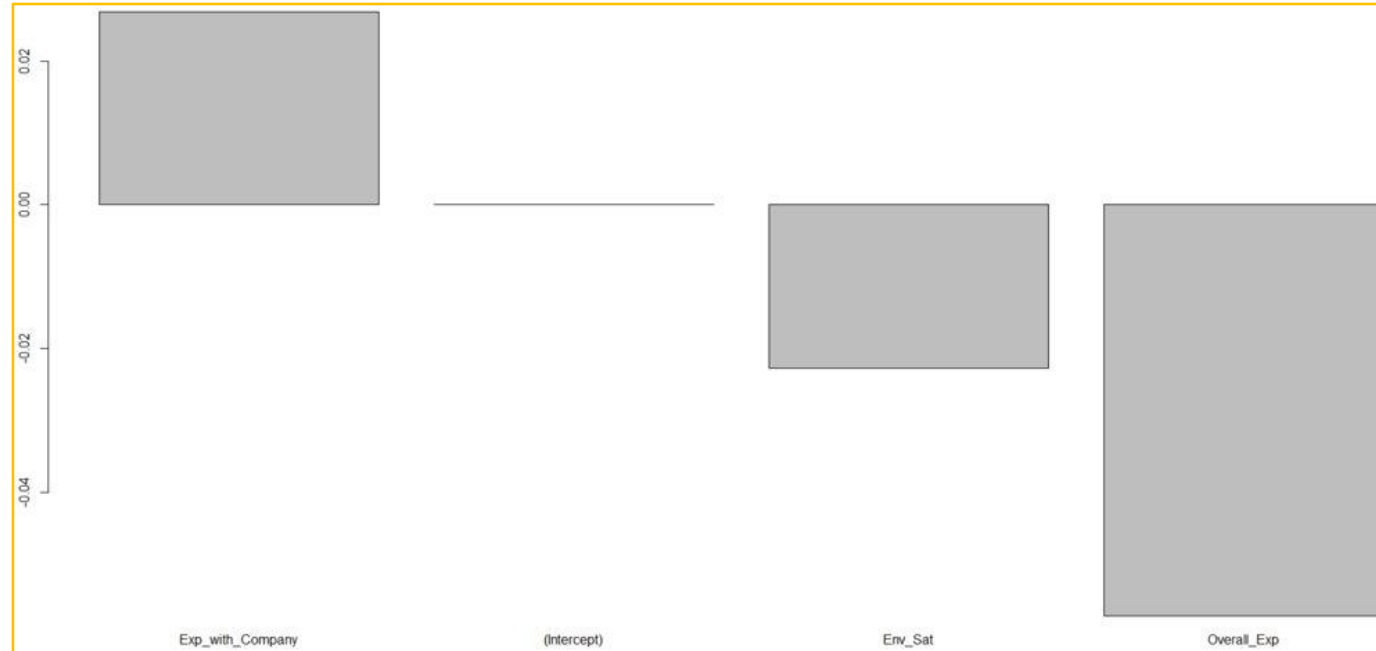
Residual standard error: 0.663 on 4406 degrees of freedom
Multiple R-squared:  0.0045,    Adjusted R-squared:  0.003822
F-statistic: 6.639 on 3 and 4406 DF,  p-value: 0.0001803
```

OLS with Factor Data (Numeric and Ordinal)

- The residuals looks almost normal and does not show any pattern or violate assumptions.



OLS with Factor Data – Important Features



- Experience with company positively influences Monthly Income
- Environment Satisfaction and Overall Experience negatively influences Monthly Income
- Overall Experience is the strongest influencer

Logistic Regression with Factor Analysis

```
Call:
glm(formula = Attrition ~ Experience_with_Company + Overall_Experience +
    Job_Satisfaction, family = "binomial", data = train1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9036  -0.6626  -0.5452  -0.3663   2.9617

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.76112    0.05458  -32.268  < 2e-16 ***
Experience_with_Company -0.60322    0.06649   -9.072  < 2e-16 ***
Overall_Experience  -0.17539    0.05086   -3.449  0.000564 ***
Job_Satisfaction   -0.07255    0.04988   -1.455  0.145748
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

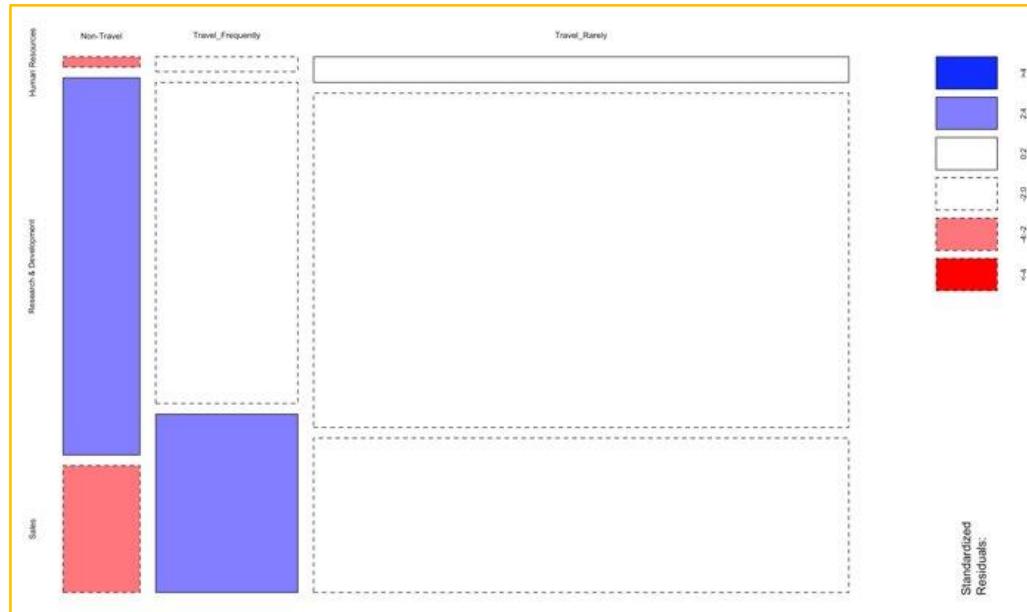
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2728.0  on 3086  degrees of freedom
Residual deviance: 2615.5  on 3083  degrees of freedom
AIC: 2623.5
```

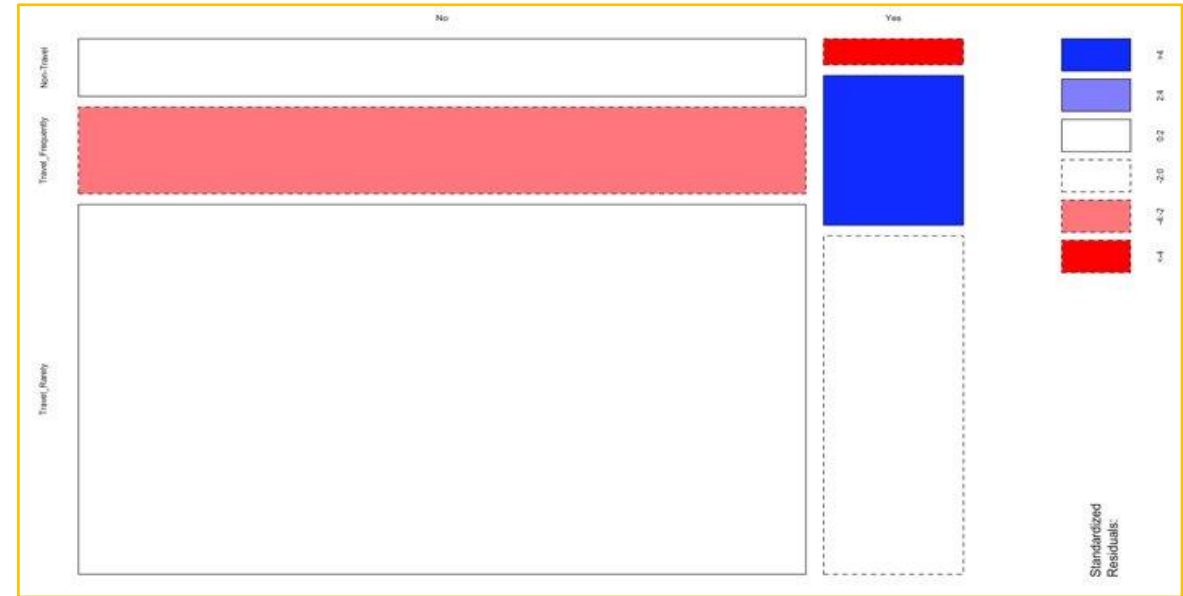
- Logistic Regression Model with just the factors from PCA also has the accuracy of 83% on the test set.
- The model has a chi-square of 113 with 3 df and p-value < 0.05, thus proving it is a better fit than an empty model.
- The significant predictors as per chi-square in the order of importance are –
 - Experience with Company (includes years at company, years since last promotion, years with current manager)
 - Overall Experience (includes Age, total working years, number of companies worked)
 - Job Satisfaction (includes distance from home, Job Level, monthly income)

Correspondence Analysis

Business Travel versus Department



Attrition versus Business Travel

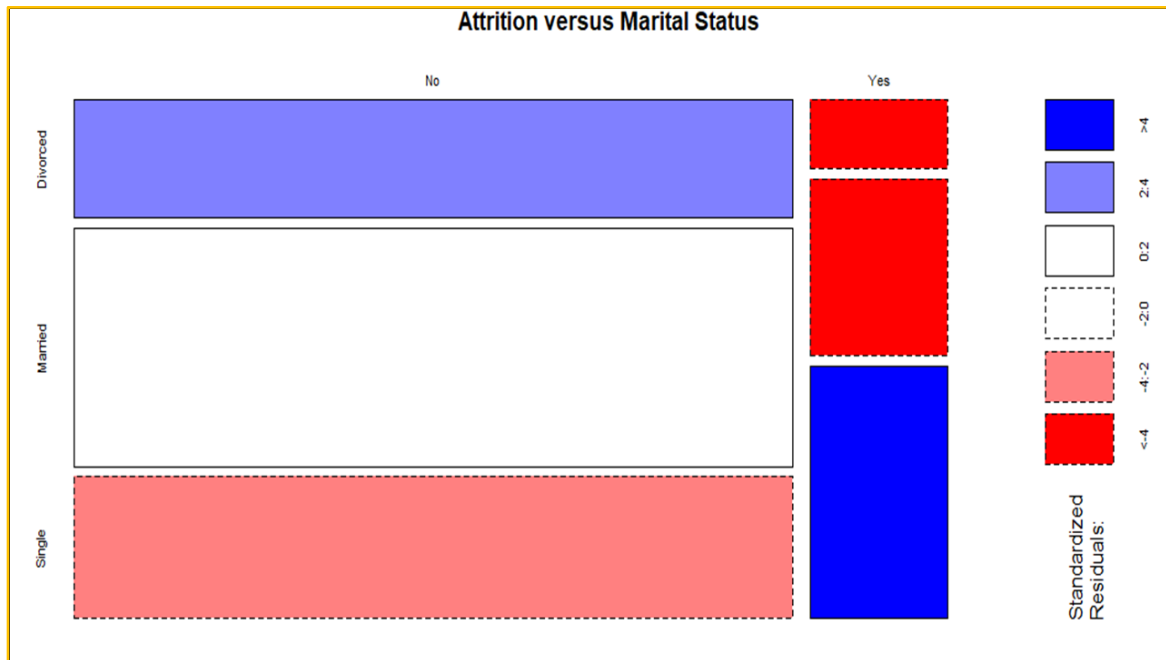


Attrition versus Department



Correspondence Analysis

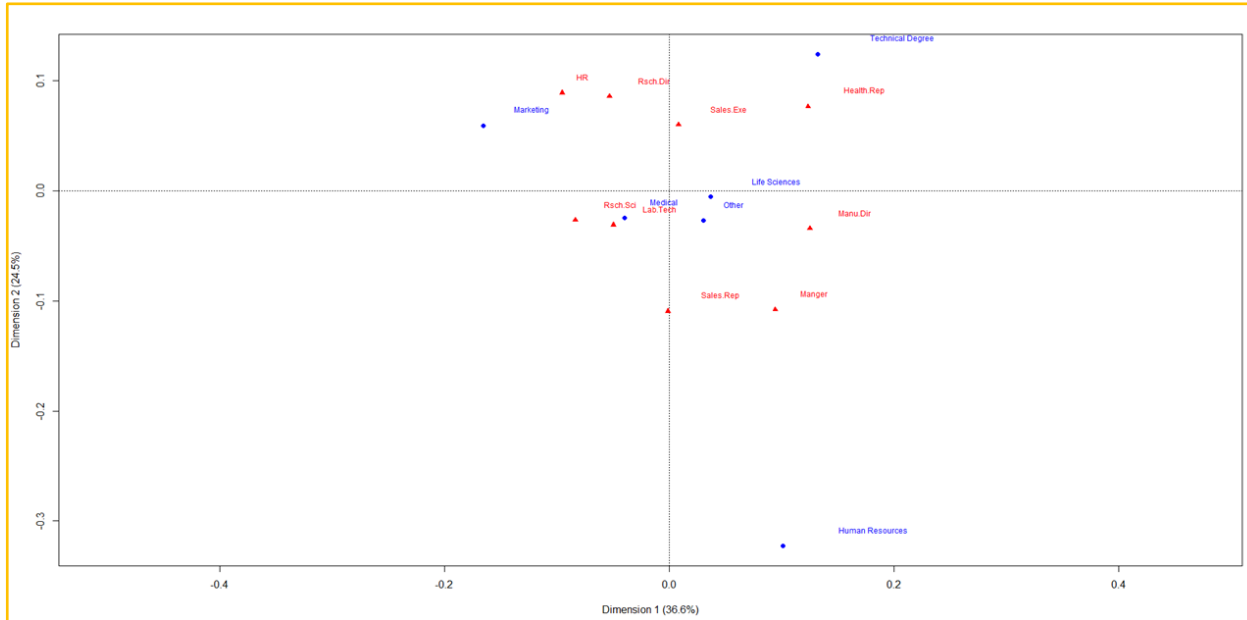
- Single population had a high correspondence with the attrition rate of the company .
- Divorced population staying in the company seems to be highly likely to happen



- High number of employees with a technical degree are working as a Sales Executive
- Employees with technical degree are highly unlikely to work as a Research Scientist

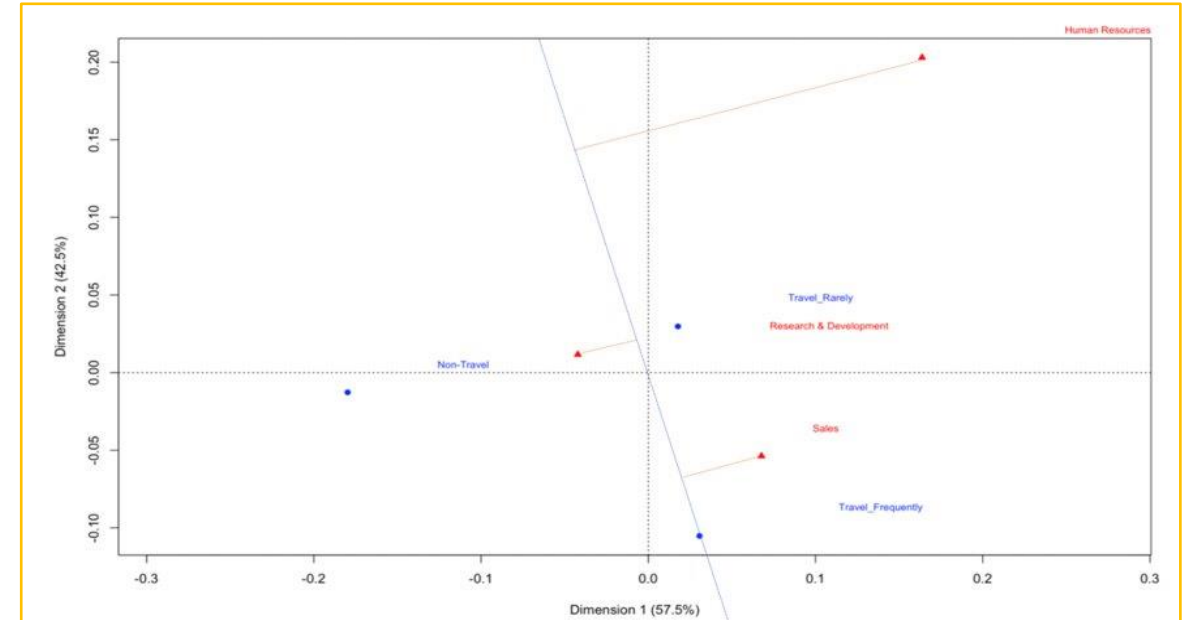


Correspondence Analysis



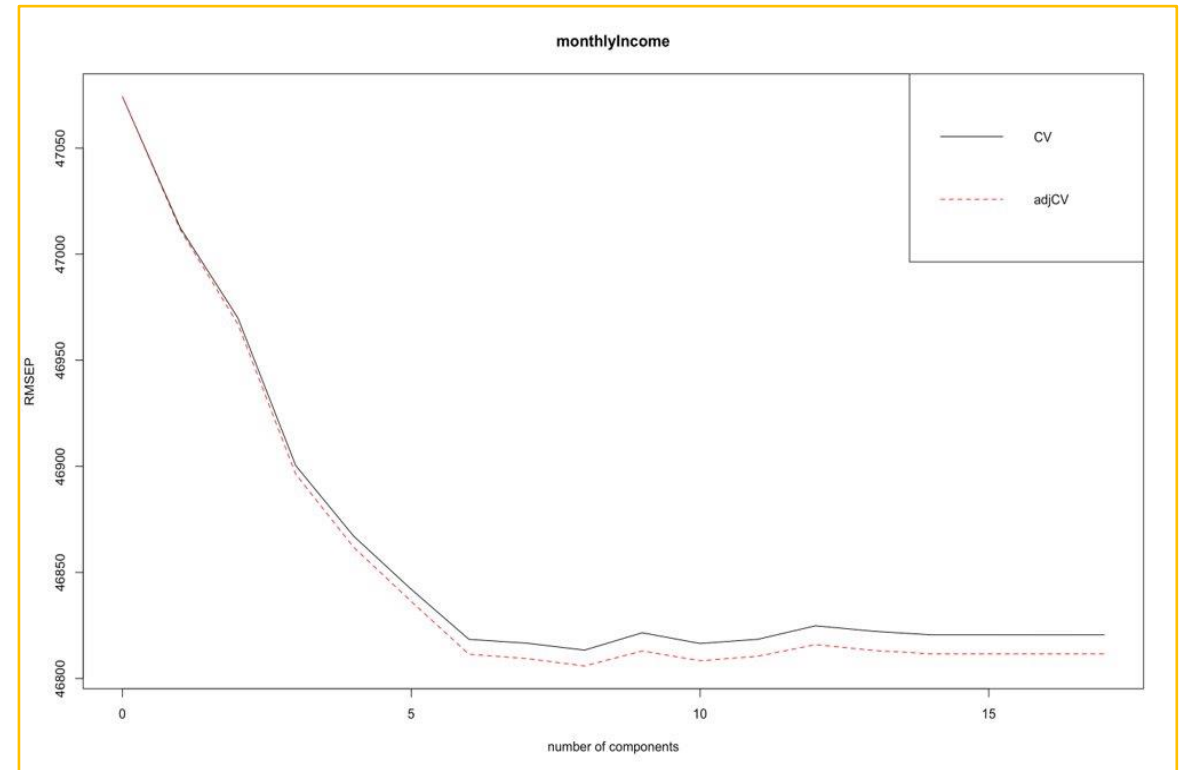
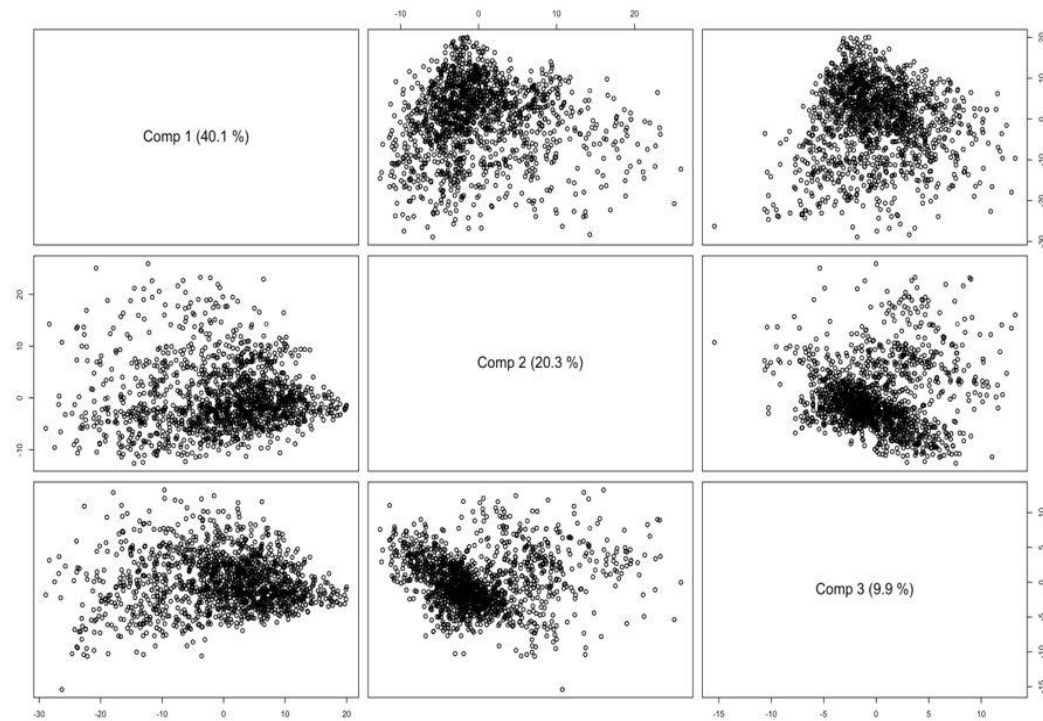
Job Role versus Education Field

- Medical Background - research scientists & lab technicians
- Technical degree - health representative & manufacturing director
- Marketing background - HR, research scientist, and research director.



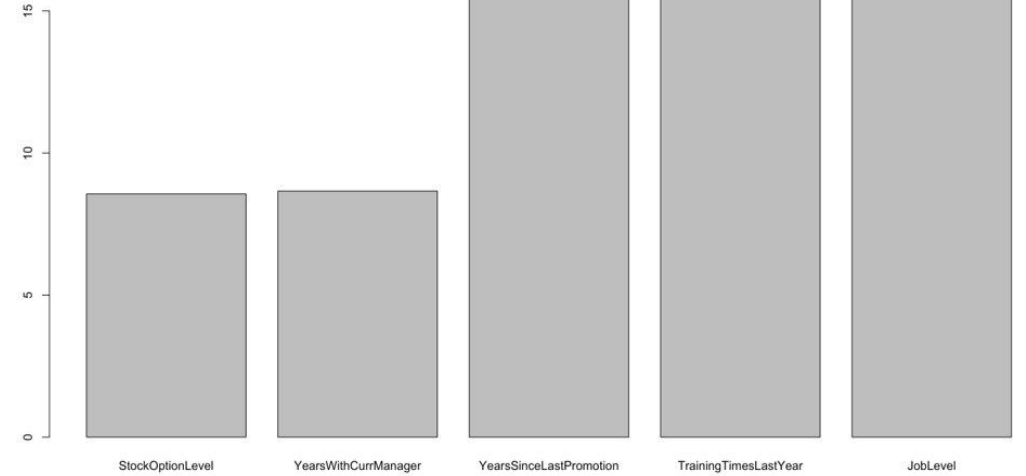
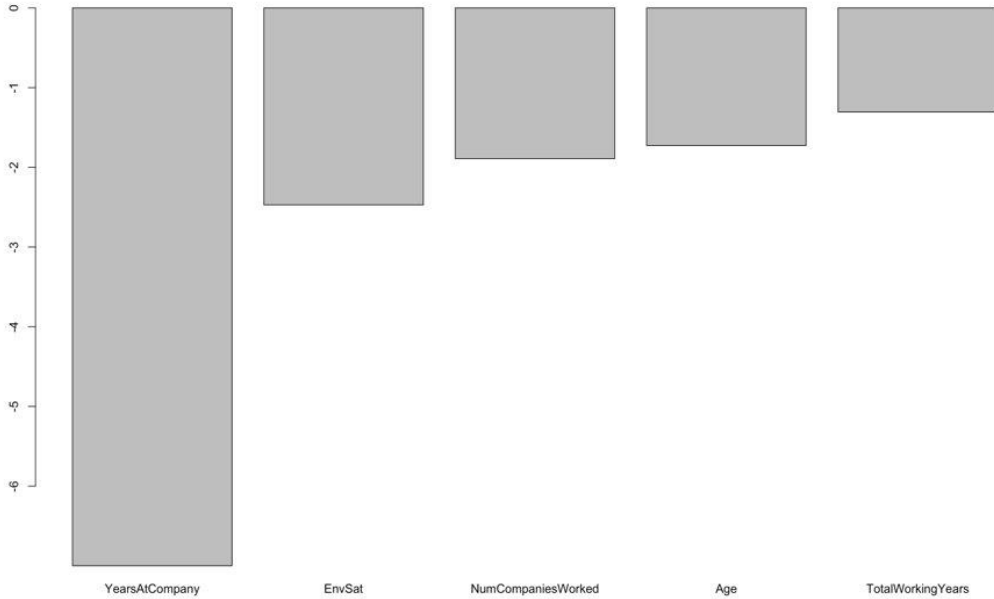
Business Travel versus Department

- Sales department corresponds most to travel frequently
- HR department corresponds the least to travel frequently



Partial Least Square Regression

- Cross validation – find the optimal number of retained dimensions
- Root Mean Square Error of Prediction – seems 6 components



Partial Least Square Regression

- Extract the useful information
- Job Level, Training Time Last Year, and Years Since Last Promotion are positive predictors of Monthly Income
- Years at Company, Environmental Satisfaction are negative predictors of Monthly Income

Linear Regression Feature Selection vs Partial Least Square

Technique	Automate Model Selection	Partial Least Square
Variables Selected	<i>Training Time Last Year</i> <i>Years Since Last Promotion</i> <i>Job Level</i> <i>Stock Option Level</i> <i>Work Life Balance</i> <i>Total Working Years</i>	<i>Job Level</i> <i>Training Times Last Year</i> <i>Year Since Last Promotion</i> <i>Years At Company</i> <i>Environmental Satisfaction</i>

Summary



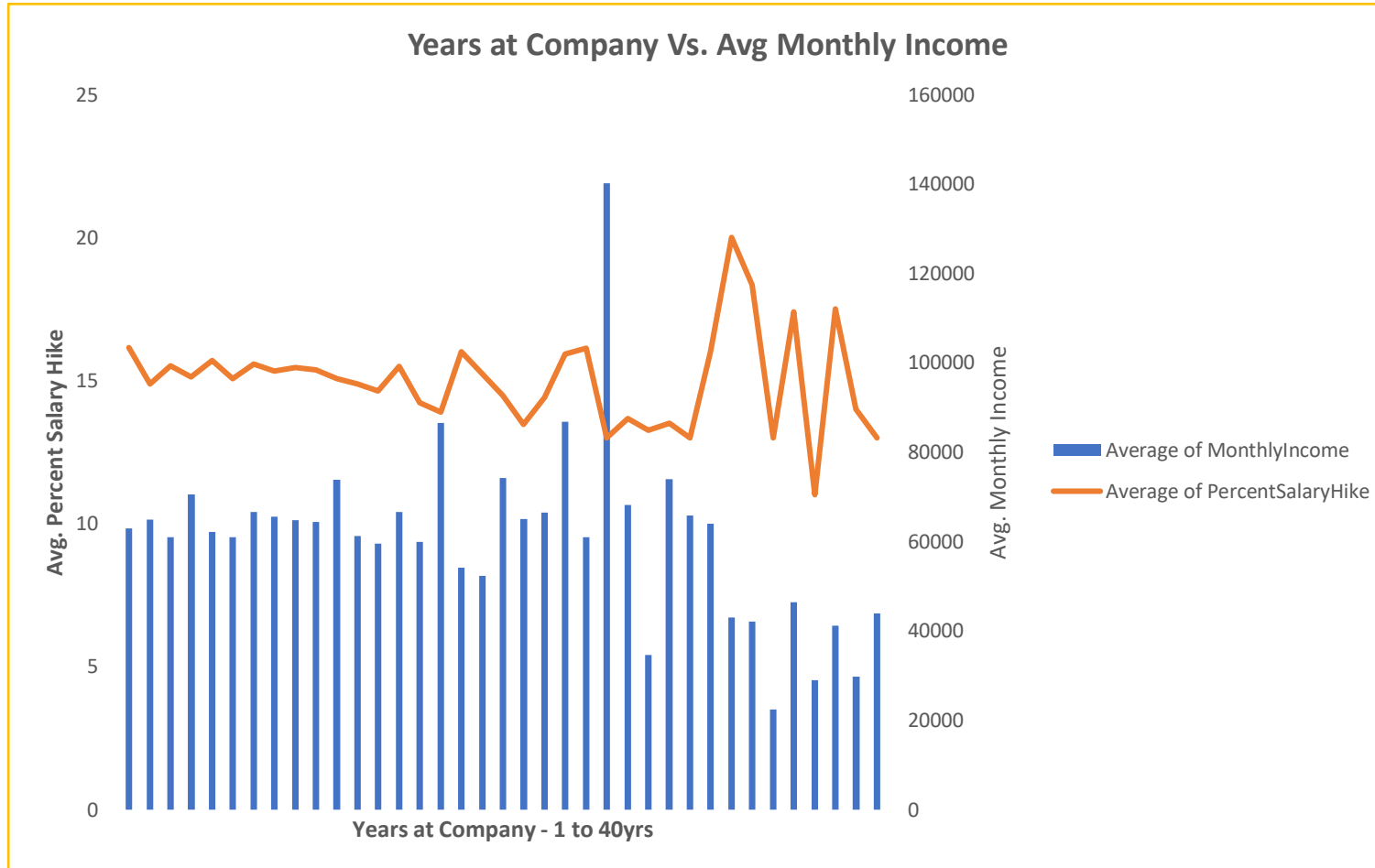
- The HR dataset contained ordinal and categorical variables in addition to numeric variables therefore ordinary least square regression and logistic regression were not suitable.
- No correlation found between ordinal variables therefore were treated as numeric variables.
- Numeric and ordinal variables were reduced using factor analysis to get predictors for linear and logistic regression.
- These factors were used for linear and logistic regression models.

Conclusion



- Factors which had an impact on monthly income – Overall Experience, Experience within Company and Environment satisfaction.
- Factors which had an impact on attrition – Overall Experience, Experience within Company and Job Satisfaction

Key Take Aways



- Years at company Vs. Avg Monthly Income
- Average Salary hike over the years in the company
- Experience within company plays an important role in whether an employee will leave the company or not. The company can use this insight to decrease attrition rate by monitoring promotions, employee relationship with their managers.

Next Steps



- Combine factors from correspondence analysis in with the numeric and ordinal factors.



**THANK
YOU**

