DSCI 4700/5260 ITDS CAPSTONE
BUSINESS PROCESS ANALYSIS PROJECT REPORT TEMPLATE

IBM HR EMPLOYEE ATTRITION: HOW IBM HUMAN RESOURCES CAN PREDICT AND CONTROL KEY EMPLOYEE ATTRITION RATE

Submitted to:

Dr. Valerie Bell
ITDS Department/UNT

Prepared by:

Sonakshi Sharma, Brooke Woods, Ugochi Madumere
ITDS Department, College of Business/UNT

Date Submitted
02-08-2020

EXECUTIVE SUMMARY

Recruiting, hiring, onboarding, and training new employees costs IBM billions each year due to employee attrition. They incur losses involving the time, money, and efforts for training in their hiring processes. They also suffer productivity/profit losses when there is constant shed in the workforce, especially top talents and they are very difficult and expensive to replace. [5] Generally, businesses are better off when they can retain good employees and the organizational experience they have. The purpose of this project is to understand what factors affect attrition rates in the employee hiring process and provide valuable insight to IBM's managers in which they can use to better their hiring processes.

We analyzed the "IBM HR Analytics Employee Attrition & Performance"; a fictional dataset from IBM data scientists downloaded from the data source Kaggle. [4] It contained variables that we further analyzed such as "age", "sex", "marital status", "department", "education field", "job level", "job role", "#companiesworked", "%salaryhike", "stockoption", and "performance rating" that we have understood to be determinants of Employee attrition. Our goal is not only reducing the employee attrition rates but providing solutions to management that improve the hiring process. [4]

"Employee Benefit News (EBN) reports that it cost employers 33% of a worker's annual salary to hire a replacement if that worker leaves. In dollar figures, the replacement cost is $15,000 per person for an employee earning a median salary of $45,000 a year, according to the Work Institute's 2017 Retention Report." [3] During 2019, IBM had 352,600 employees. Based on the analysis about the "IBM HR Analytic Employee Attrition & Performance", We discovered some valuable insights, most of the employees who left the company have worked for 0-5 years with relatively low monthly income. [4] Although some employees seem satisfied with the work environment, colleagues, and culture, they still choose to leave. Another interesting finding about the company's employees is that age and work experience are highly correlated with income in the company. This partly explains

why employees leave. Most people who left have worked for a short period of time, which relates to lower salary.

The focus of our is aimed at helping IBM human resources make better hiring decisions and find resignation trend/numbers to understand the latent factors which contribute to employee attrition in the organization. We have used a logistic regression model to predict which variables affect employee attrition rates. We have used the DMAIC problem solving method to define, measure, analyze, improve, and control IBMs hiring process. Our study provides insight to reduce financial losses due to employee attrition as we assist with proactiveness in employee welfare planning.

1.0 Improvement Opportunity: Define Phase

1.1 Problem Statement/Discussion of Process being Examined

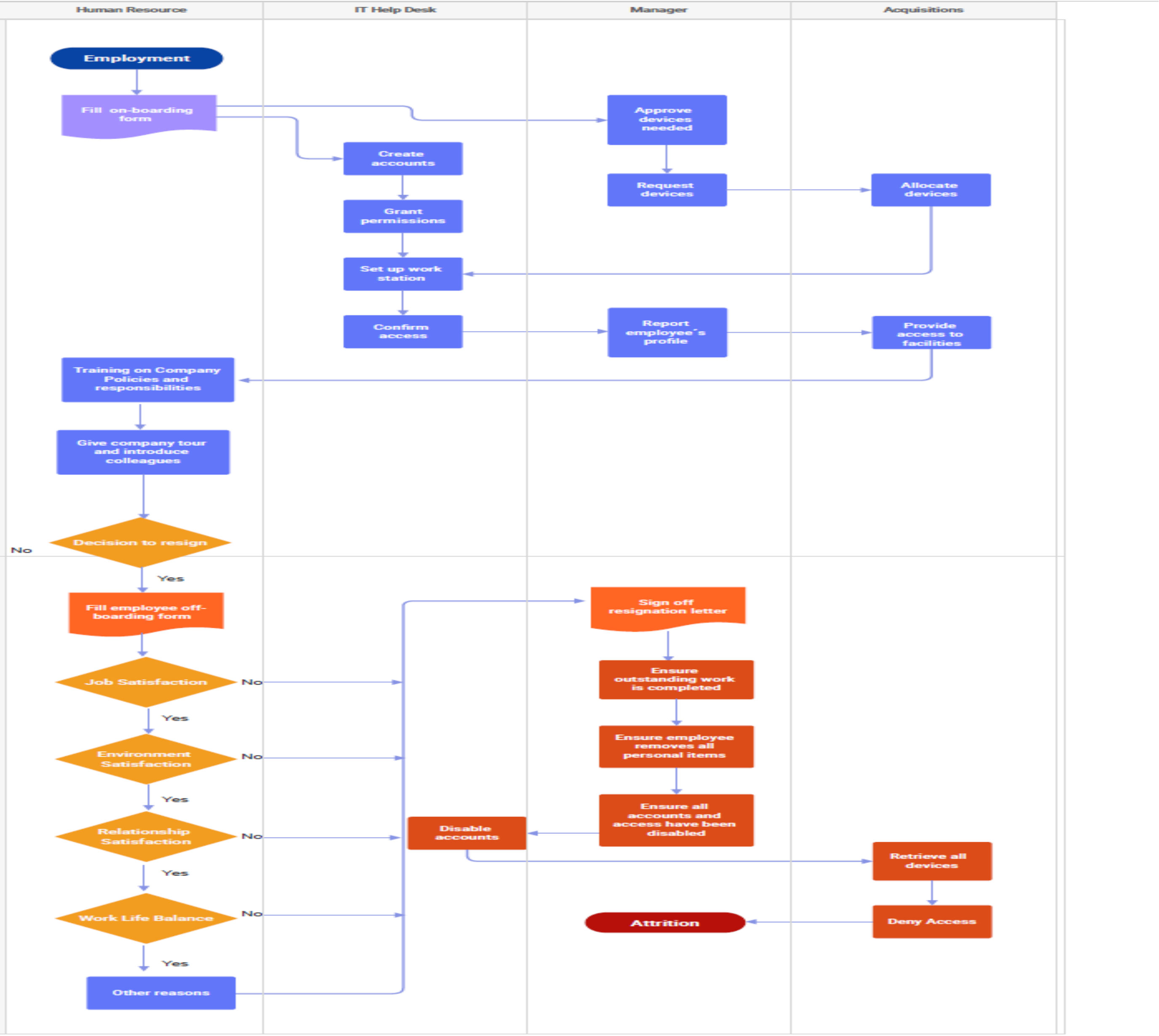
Our dataset contained qualitative variables such as gender, marital status, department, and job role and quantitative variables such as the number of companies they have worked, the number of years they have worked, etc. [4] However, the problem is that only some of these input variables explain why employees are leaving, contribute to future attrition, and can provide adjusting suggestions where necessary. Below are the details of our problem-solving approach and Exploratory phase of our analysis.

1.2 Problem Solving Approach/ Process

For this project, we looked specifically at the onboarding vs the recruiting process map for employees of IBS. We ran binary logistic regression using SPPS and Python. From there, we used metrics such as R squared, goodness of fit, and odds ratio to determine which variables contribute the most to attrition, and therefore the variables which influence employee behavior. We then suggested these latent factors IBM managers, provided insight to Mangers on the controllable factors that should be reconsidered, and directed them to better control their recruiting and onboarding processes.

With high attrition for IBS, the inputs and resources used in the recruiting and onboarding processes are of no use and can become a loss after an employee leaves the company. We will provide a solution using various analytical techniques.

First, we looked at the univariate box plots to determine if there were any outliers present. Then we find the bivariate correlation regression which evaluates the degree of the relationship between two quantitative variables using model comparison making sure that we did not have bias from correlation. Our process flow chart is shown below.



1.3 Identification of Key Metrics

Key Measures used

- R squared

Indicates variance in the DV explained by the independent variables.

- Goodness of Fit

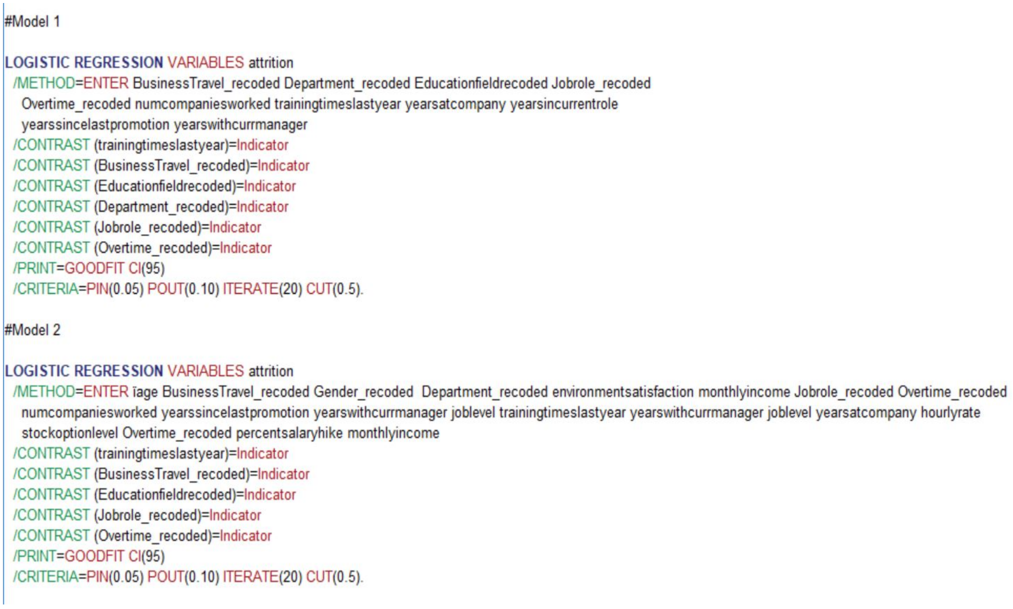
To test robustness of the model

- Odds Ratio

1.4 Project Scope

- Performing stepwise Logistic Regression in SPSS and Python
- Find trends by modeling relationships between Attrition and variables in Bar charts
- Providing summary statistics about employees
- Understanding how the company's employees perceive the working environment
- Investigating possible factors that affect attrition
- Using statistical analysis tools like SPSS, and Python
- Developing a model that can detect future attrition based on HR employee survey data collection
- Providing insight to Managers
- Constrained only by factors that can be managed and controlled.

Input Variables Code:



```
#Model 1

LOGISTIC REGRESSION VARIABLES attrition
  /METHOD=ENTER BusinessTravel_recoded Department_recoded Educationfieldrecoded Jobrole_recoded
    Overtime_recoded numcompaniesworked trainingtimeslastyear yearsatcompany yearsincurrentrole
    yearssincelastpromotion yearswithcurrmanager
  /CONTRAST (trainingtimeslastyear)=Indicator
  /CONTRAST (BusinessTravel_recoded)=Indicator
  /CONTRAST (Educationfieldrecoded)=Indicator
  /CONTRAST (Department_recoded)=Indicator
  /CONTRAST (Jobrole_recoded)=Indicator
  /CONTRAST (Overtime_recoded)=Indicator
  /PRINT=GOODFIT CI(95)
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).

#Model 2

LOGISTIC REGRESSION VARIABLES attrition
  /METHOD=ENTER ïage BusinessTravel_recoded Gender_recoded  Department_recoded environmentsatisfaction monthlyincome Jobrole_recoded Overtime_recoded
    numcompaniesworked yearssincelastpromotion yearswithcurrmanager joblevel trainingtimeslastyear yearswithcurrmanager joblevel yearsatcompany hourlyrate
    stockoptionlevel Overtime_recoded percentsalaryhike monthlyincome
  /CONTRAST (trainingtimeslastyear)=Indicator
  /CONTRAST (BusinessTravel_recoded)=Indicator
  /CONTRAST (Educationfieldrecoded)=Indicator
  /CONTRAST (Jobrole_recoded)=Indicator
  /CONTRAST (Overtime_recoded)=Indicator
  /PRINT=GOODFIT CI(95)
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

2.0 CURRENT STATE OF THE PROCESS: MEASURE PHASE

2.1 Current Performance Level

We used a dataset called “IBM HR Analytics Employee Attrition & Performance”; a fictional dataset from IBM data scientists downloaded from the data source Kaggle. This dataset is composed of 1479 variables and 35 columns and is available for download at https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset/data#.[4] It took about two days to find this dataset after a rigorous search. On average, it took about three days to explore and test the data. This data had been preprocessed, so there were no significant outliers or missing data found when we attempted to clean our dataset.

2.2 Descriptive Statistics Summary Table

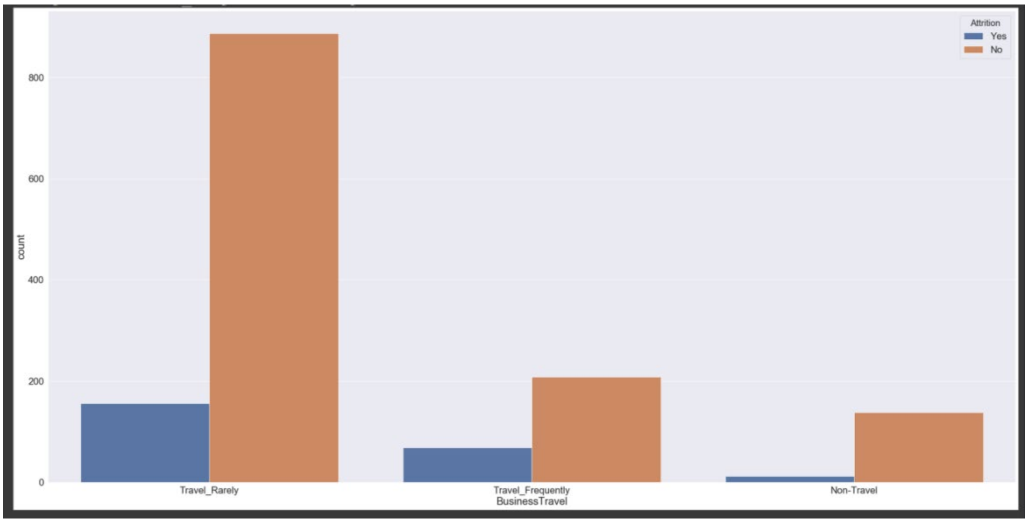
The descriptive statistics are summarized below

	N	Minimum	Maximum	Mean	Std. Deviation
DailyRate	1470	102	1499	802.49	403.509
DistanceFromHome	1470	1	29	9.19	8.107
Education	1470	1	5	2.91	1.024
EmployeeCount	1470	1	1	1.00	.000
EmployeeNumber	1470	1	2068	1024.87	602.024

EnvironmentSatisfaction	1470	1	4	2.72	1.093
HourlyRate	1470	30	100	65.89	20.329
JobInvolvement	1470	1	4	2.73	.712
elationshipSatisfaction	1470	1	4	2.71	1.081
StandardHours	1470	80	80	80.00	.000
StockOptionLevel	1470	0	3	.79	.852
TotalWorkingYears	1470	0	40	11.28	7.781
TrainingTimesLastYear	1470	0	6	2.80	1.289
WorkLifeBalance	1470	1	4	2.76	.706
YearsAtCompany	1470	0	40	7.01	6.127
YearsInCurrentRole	1470	0	18	4.23	3.623
YearsSinceLastPromotion	1470	0	15	2.19	3.222
YearsWithCurrManager	1470	0	17	4.12	3.568
Valid N (listwise)	1470				

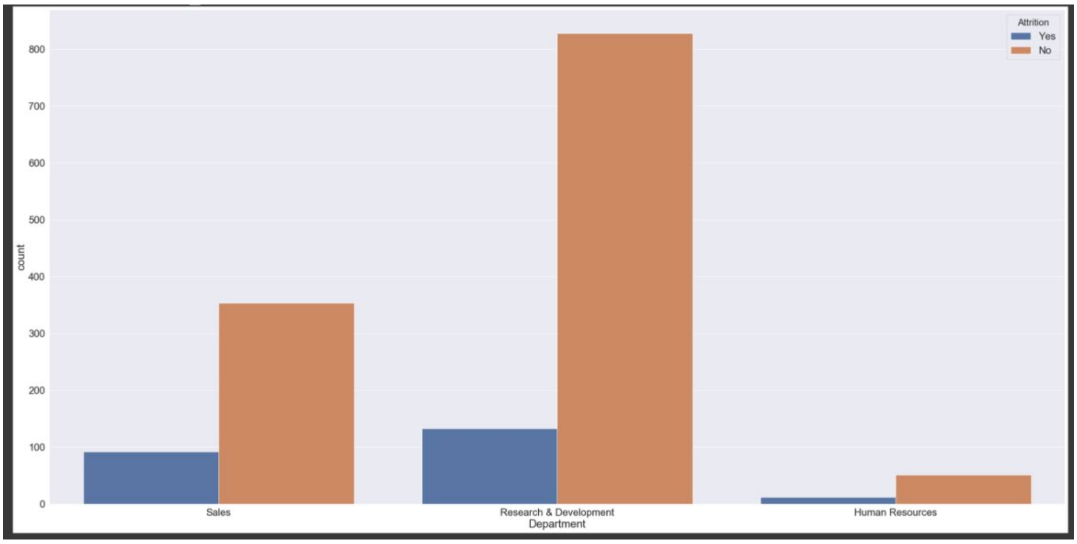
- Distribution/Patterns of Key Y Outputs

BUSINESS TRAVEL



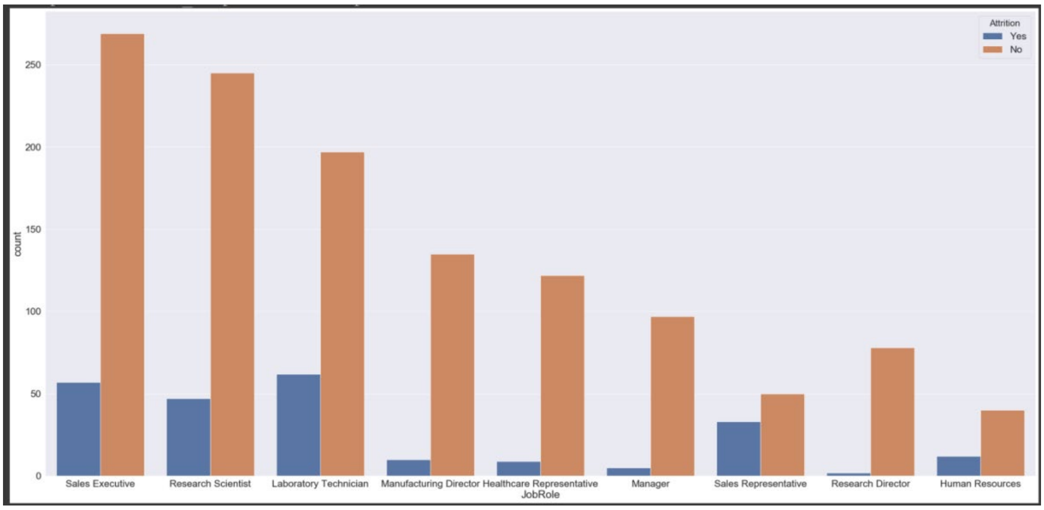
When employees travel rarely, they have a significantly higher attrition than when compared to when they travel frequently or do not travel at all.

DEPARTMENT



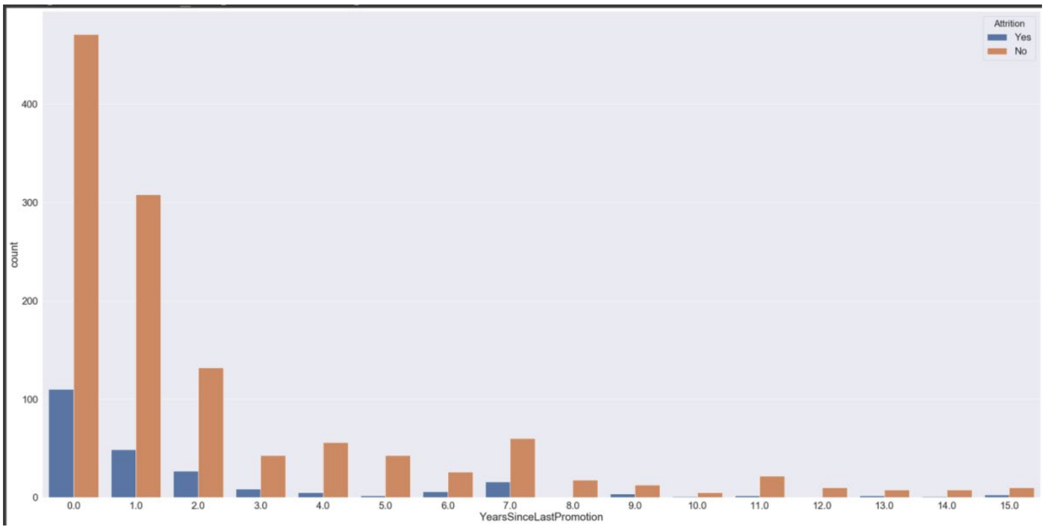
The python bar chart shows that department influences attrition rates. There issignificantly much higherattrition in Research and Development departments than HR.

JOB ROLE



This bar chart shows that Lab technicans, Sales executives, Sales representatives, and Research scientiest have much higher attrition than other Job roles.

YEARS SINCE LAST PROMOTION



We can see that the years there is higher attrition from yrs 0-2 for YearsSinceLastPromotion.

2.3 Identification of Target Performance Levels or Project Goals

Desired performance level: Minimum attrition rate amongst key employees.

In order to find the variables which impact IBMS attrition rate, we analyzed our findings using the Metrics for Evaluation in SPSS and Python. Before performing these tests, the data was tested to make sure no assumptions were violated. All violated assumptions were corrected before running logistic regression models.

Assumption testing:

Normality assumption:

Normality of the residuals means that the residuals follow a normal distribution. The assumption of normality was tested using Q-Q plots and histograms. When plotting Q-Q plot of observed versus expected values, the points should be symmetrically distributed around a diagonal line. When plotting the histogram of residuals, the shape should be a shape of a normal distribution. Moreover, normality assumption was tested using Kolmogorov-Smirnov and Shapiro-Wilk test for all three models. If the test was significant at 5%, there was no evidence of violation of the normality assumption.

Histograms:

The graphs were displaying a combination of different shapes to check for violations of the normality assumption. Some histograms were both left skewed and some were right skewed. There were also some bell-shaped formations in different variable histograms. Since statistical tests would reflect a better investigation of the normality assumption. These histograms are shown in Appendix C

Statistical Test:

We see that the normality tests are statistically significant in the statistical test, which further reinstates violation of normality assumption.

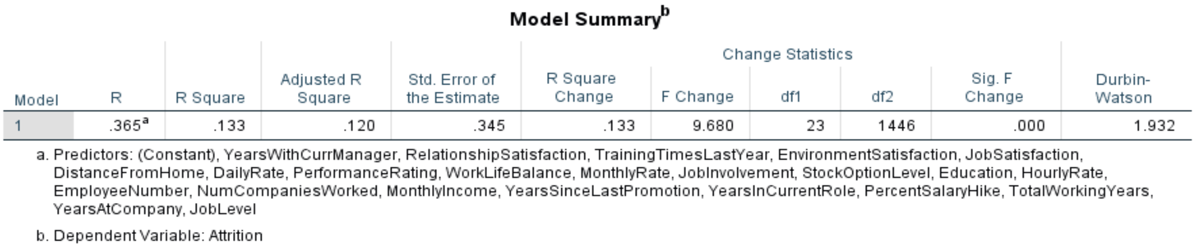
The tests showed that the data was non-normal and so transformation was necessary and shown in appendix D. Therefore, we had to make transformed variables like natural log. We created new variables which were logged versions of the original variables. Before we transform the data, we need to calculate the ratio between of mean to standard deviation ratio. If the ratio is less than 4 then the transformation will be impactful. After transforming the variables, the normality tests were performed again. The normality tests showed significance again, which again showed violation of normality assumption.

Independence of Residuals:

Independence of the residuals means that the residuals are not autocorrelated. The assumption of independence was tested using scatter plots shown in Appendix E, plotting residuals versus case number. The points should be symmetrically distributed around a horizontal line. The Durbin-Watson statistic was also used to test for independence of the residuals. If the Durbin-Watson statistic is close to 2, there was no autocorrelation.

- Values from 0 to less than 2 indicated positive autocorrelation and values from 2 to 4 indicated negative autocorrelation.
- Multicollinearity was also detected with the help of tolerance and its reciprocal, called variance inflation factor (VIF).
- If the value of tolerance was less than 0.2 or 0.1 and, simultaneously, the value of VIF 10 and above, then the multicollinearity was problematic.

One of the violations occur when we have time relevant or longitudinal data. The best way to analyze such data is General Estimation Equation (GEE). The DW statistic falls within the range of 1.5 and 2.5 which showed that the independence of residuals assumption was met.



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.365[a]	.133	.120	.345	.133	9.680	23	1446	.000	1.932

a. Predictors: (Constant), YearsWithCurrManager, RelationshipSatisfaction, TrainingTimesLastYear, EnvironmentSatisfaction, JobSatisfaction, DistanceFromHome, DailyRate, PerformanceRating, WorkLifeBalance, MonthlyRate, JobInvolvement, StockOptionLevel, Education, HourlyRate, EmployeeNumber, NumCompaniesWorked, MonthlyIncome, YearsSinceLastPromotion, YearsInCurrentRole, PercentSalaryHike, TotalWorkingYears, YearsAtCompany, JobLevel

b. Dependent Variable: Attrition

Multicollinearity assumption

Collinearity exists if there is an approximate linear relationship (i.e., shared variance) among some of IVs in the data.

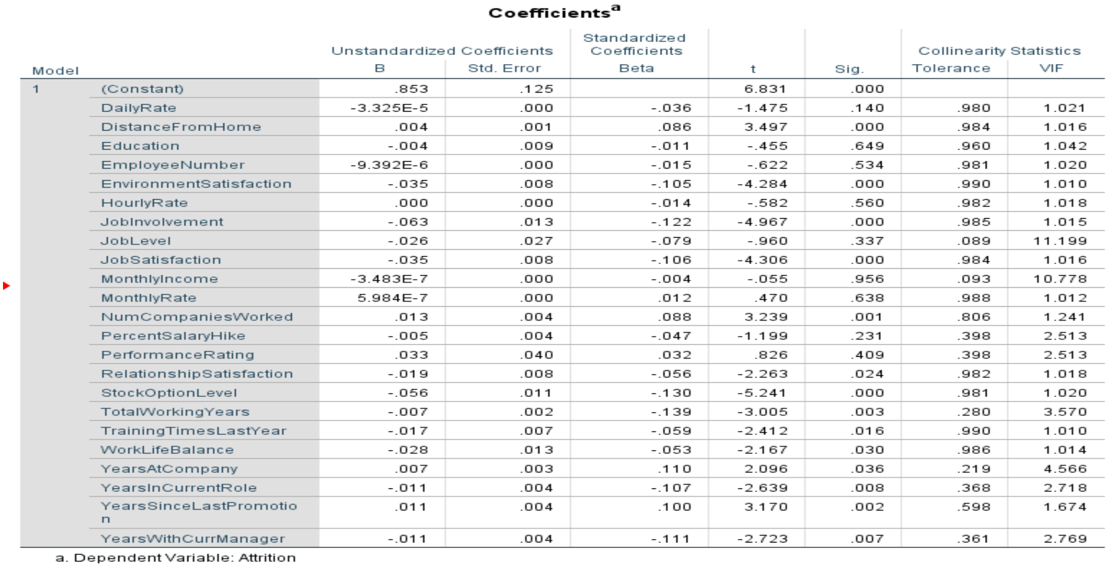
Problems with collinearity: Collinearities inflate the variances of the regression coefficients. This could have the following consequences: Parameter estimates that fluctuate dramatically with negligible changes in the sample; parameter estimates with signs that are "wrong" in terms of theoretical considerations; theoretically "important" variables with insignificant coefficients; the inability to determine the relative importance of collinear variables.

It produces large SE of Bs which results in coefficients being non-significant; it produces bizarre β estimates (e.g., wrong direction); Removal or addition of one IV results in enormous change to the models.

The indicators of multicollinearity are Tolerance ≤ .1 (or .2) [VIF ≥ 10 (or 5)].; even if tolerance ≥ .2 (or VIF ≤ 5), multicollinearity could be problematic when: There is a bivariate correlation of .7 or more between two IVs; the bivariate correlation between two IVs are greater than either of IV's correlation with DV.

SPSS Logistic Regression Output model 1:

The results of VIF lack multicollinearity. Tolerance looked better in most of the variables.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	.853	.125		6.831	.000		
	DailyRate	-3.325E-5	.000	-.036	-1.475	.140	.980	1.021
	DistanceFromHome	.004	.001	.086	3.497	.000	.984	1.016
	Education	-.004	.009	-.011	-.455	.649	.960	1.042
	EmployeeNumber	-9.392E-6	.000	-.015	-.622	.534	.981	1.020
	EnvironmentSatisfaction	-.035	.008	-.105	-4.284	.000	.990	1.010
	HourlyRate	.000	.000	-.014	-.582	.560	.982	1.018
	JobInvolvement	-.063	.013	-.122	-4.967	.000	.985	1.015
	JobLevel	-.026	.027	-.079	-.960	.337	.089	11.199
	JobSatisfaction	-.035	.008	-.106	-4.306	.000	.984	1.016
	MonthlyIncome	-3.483E-7	.000	-.004	-.055	.956	.093	10.778
	MonthlyRate	5.984E-7	.000	.012	.470	.638	.988	1.012
	NumCompaniesWorked	.013	.004	.088	3.239	.001	.806	1.241
	PercentSalaryHike	-.005	.004	-.047	-1.199	.231	.398	2.513
	PerformanceRating	.033	.040	.032	.826	.409	.398	2.513
	RelationshipSatisfaction	-.019	.008	-.056	-2.263	.024	.982	1.018
	StockOptionLevel	-.056	.011	-.130	-5.241	.000	.981	1.020
	TotalWorkingYears	-.007	.002	-.139	-3.005	.003	.280	3.570
	TrainingTimesLastYear	-.017	.007	-.059	-2.412	.016	.990	1.010
	WorkLifeBalance	-.028	.013	-.053	-2.167	.030	.986	1.014
	YearsAtCompany	.007	.003	.110	2.096	.036	.219	4.566
	YearsInCurrentRole	-.011	.004	-.107	-2.639	.008	.368	2.718
	YearsSinceLastPromotion	.011	.004	.100	3.170	.002	.598	1.674
	YearsWithCurrManager	-.011	.004	-.111	-2.723	.007	.361	2.769

a. Dependent Variable: Attrition

2.4 Identification of Key Variables

We then checked for influential, outliers and leverage points.[4]

Normalized score:

There were some outliers as everything was not between -3 and 3 shown in the first graph of Appendix E. Based on the standardized scores we could see that all the scores for the continuous variables were within the range of -3 to 3 and so could not be considered as outliers.

Leverage points:

Observations that are substantially different from the remaining observations (shown in the third graph of Appendix E) on one or more independent variables such that they may "lever" the relationship in their direction.

Leverage distance:

(2*P)/sample size. the value we get.

Leverage – H hat/leverage – 2p/N (P = number of predictors+1), use 3p/N if you have larger sample size

p = 31, N= 1470. Leverage = 0.211.

Interpretation:

We found that Points 1341-1470 were all leverage points.

Influence Points:

These are the observations that have a disproportionate effect on the regression results.

Cook's Distance: 1.0 or 4/(n-k-1)

where n=sample size and k= number of predictors

4/(1470-30-1) = 0.027

Influential points existed between 1079-1479

3.0 ANALYSIS AND FINDINGS: THE ANALYZE PHASE

In order to find the variables which, influence attrition for IBM, we must use a logistic regression model. We referred to (Anderson, C. L., and Agarwal, R. 2010. "Practicing Safe Computing: A Multimethod Empirical Examination of Home Computer User Security Behavioral Intentions," MIS Quarterly (34:3), pp. 613-643.)

[4] for getting an idea of how analysis is performed. The target for our regression contains a binary variable named attrition. This variable indicated whether an employee chooses to leave the company (1, "yes") or not leaves the company (0,"no"). The categorical variables were first coded and then the variables are entered, shown in Appendix F. The different independent variables (both categorical as well as continuous) which could have a potential impact or precisely impact the odds of attrition for an employee are Business travel, Education, education field, environmental satisfaction, monthly rate, monthly income etc.

Model 1

Variables entered

```
LOGISTIC REGRESSION VARIABLES attrition
  /METHOD=ENTER
BusinessTravel_recodedDepartment_recodedEducationfieldrecodedJobrole_recoded
Overtime_recodednumcompaniesworkedtrainingtimeslastyearyearsatcompanyyearsincurrentrole
yearssincelastpromotionyearswithcurrmanager
  /CONTRAST (trainingtimeslastyear)=Indicator
  /CONTRAST (BusinessTravel_recoded)=Indicator
  /CONTRAST (Educationfieldrecoded)=Indicator
  /CONTRAST (Department_recoded)=Indicator
  /CONTRAST (Jobrole_recoded)=Indicator
  /CONTRAST (Overtime_recoded)=Indicator
  /PRINT=GOODFIT CI(95)
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

Output of **Model 1**

The Cox & Snell R square tests were used as well as Nagelkerke R square tests in the model summary table and indicated that 16.8% to 28.6% variance in the DV is explained by the independent variables.

Model Summary

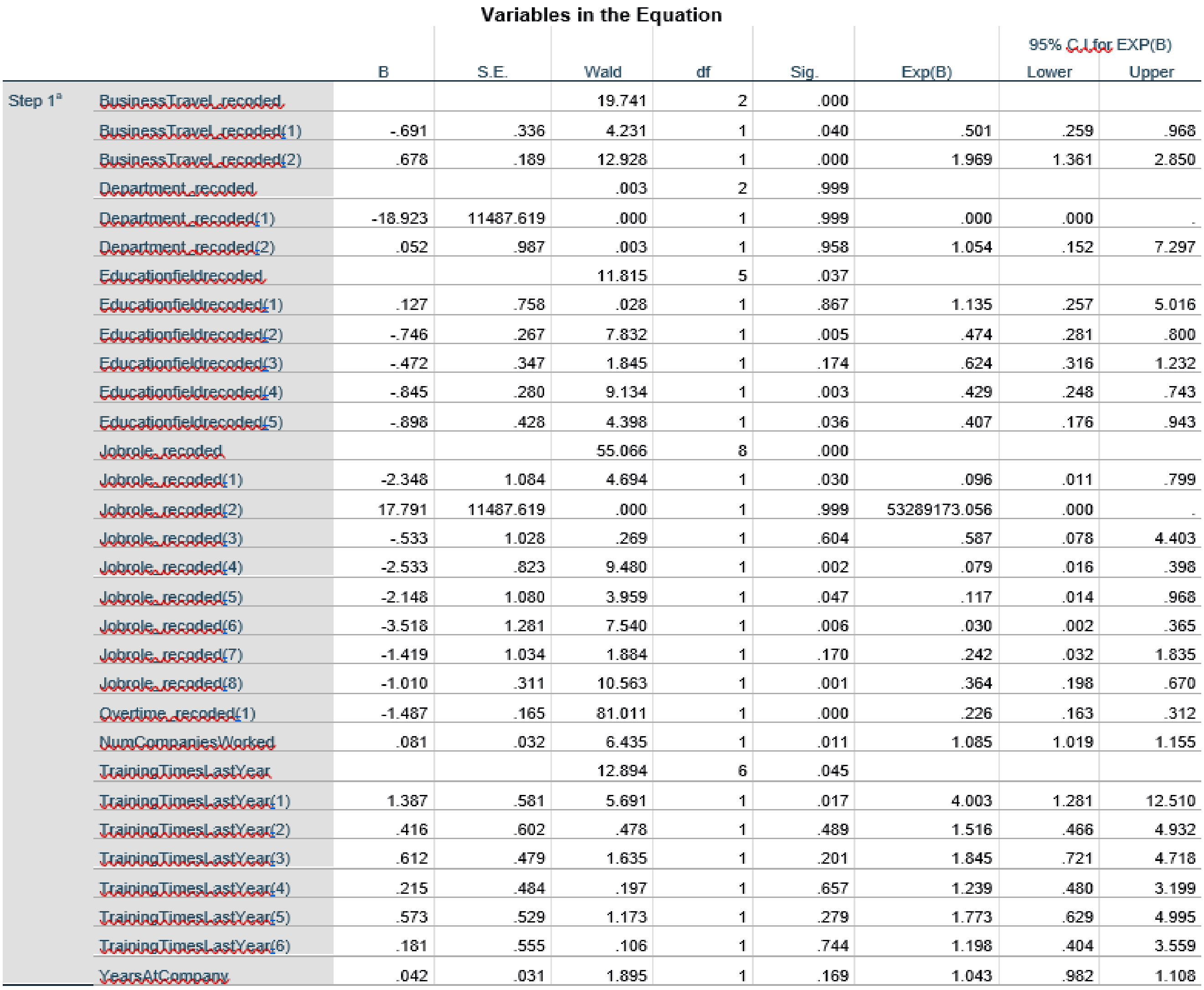
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1028.242[a]	.168	.286

a. Estimation terminated at iteration number20 because maximum iterations has been reached. Final solution cannot be found.

Hosmer and Lemeshow Test

Step	Chi-square	Df	Sig.
1	5.693	8	.682

The Hosmer and Lemeshow test is non- significant which indicates a good fit of the model.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	BusinessTravel_recoded			19.741	2	.000			
	BusinessTravel_recoded(1)	-.691	.336	4.231	1	.040	.501	.259	.968
	BusinessTravel_recoded(2)	.678	.189	12.928	1	.000	1.969	1.361	2.850
	Department_recoded			.003	2	.999			
	Department_recoded(1)	-18.923	11487.619	.000	1	.999	.000	.000	.
	Department_recoded(2)	.052	.987	.003	1	.958	1.054	.152	7.297
	Educationfieldrecoded			11.815	5	.037			
	Educationfieldrecoded(1)	.127	.758	.028	1	.867	1.135	.257	5.016
	Educationfieldrecoded(2)	-.746	.267	7.832	1	.005	.474	.281	.800
	Educationfieldrecoded(3)	-.472	.347	1.845	1	.174	.624	.316	1.232
	Educationfieldrecoded(4)	-.845	.280	9.134	1	.003	.429	.248	.743
	Educationfieldrecoded(5)	-.898	.428	4.398	1	.036	.407	.176	.943
	Jobrole_recoded			55.066	8	.000			
	Jobrole_recoded(1)	-2.348	1.084	4.694	1	.030	.096	.011	.799
	Jobrole_recoded(2)	17.791	11487.619	.000	1	.999	53289173.056	.000	.
	Jobrole_recoded(3)	-.533	1.028	.269	1	.604	.587	.078	4.403
	Jobrole_recoded(4)	-2.533	.823	9.480	1	.002	.079	.016	.398
	Jobrole_recoded(5)	-2.148	1.080	3.959	1	.047	.117	.014	.968
	Jobrole_recoded(6)	-3.518	1.281	7.540	1	.006	.030	.002	.365
	Jobrole_recoded(7)	-1.419	1.034	1.884	1	.170	.242	.032	1.835
	Jobrole_recoded(8)	-1.010	.311	10.563	1	.001	.364	.198	.670
	Overtime_recoded(1)	-1.487	.165	81.011	1	.000	.226	.163	.312
	NumCompaniesWorked	.081	.032	6.435	1	.011	1.085	1.019	1.155
	TrainingTimesLastYear			12.894	6	.045			
	TrainingTimesLastYear(1)	1.387	.581	5.691	1	.017	4.003	1.281	12.510
	TrainingTimesLastYear(2)	.416	.602	.478	1	.489	1.516	.466	4.932
	TrainingTimesLastYear(3)	.612	.479	1.635	1	.201	1.845	.721	4.718
	TrainingTimesLastYear(4)	.215	.484	.197	1	.657	1.239	.480	3.199
	TrainingTimesLastYear(5)	.573	.529	1.173	1	.279	1.773	.629	4.995
	TrainingTimesLastYear(6)	.181	.555	.106	1	.744	1.198	.404	3.559
	YearsAtCompany	.042	.031	1.895	1	.169	1.043	.982	1.108

Interpreting the odds ratio for Model 1:

- The employees who travel frequently have lower odds of attrition than employees who do not travel at all. The odds of attrition for such employees are 0.50 times that of employees who do not travel.
- The employees who travel rarely have higher odds of attrition than employees who do not travel at all. The odds of attrition for such employees are 1.969 times that of employees who do not travel.
- The employees who have human resources, manufacturing director, research director, research scientist and sales representative as job roles have lower odds (.096, .079, .117, .030 and .364 respectively) of attrition than those employees who have a job role as a healthcare representative.

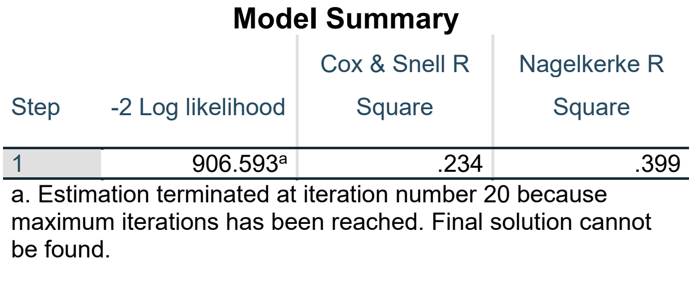
- The employees who have marketing, other and Technical Degree education fields have lower odds (.474, .429 and .407 respectively) of attrition than those employees with human resources education field.
- A unit increase in the number of companies an employee works with increases the odds of attrition of an employee 1.085 times.
- The employees who work overtime have lower odds (.226) of attrition than those employees who do not work overtime.
- The employees who have trained once in last year have higher odds (4.003) of attrition than those employees who not trained at all last year.
- A unit increase in the number of years an employee spent in a current role lower the odds (.861) of attrition of an employee 1.085 times.
- A unit increase in the number of years an employee spends after last promotion increases the odds of attrition of an employee 1.149 times.
- A unit increase in the number of years an employee spends with the current manager lowers the odds of attrition of an employee .890 times.

MODEL 2

Variables entered

```
LOGISTIC REGRESSION VARIABLES attrition
  /METHOD=ENTER
ïageBusinessTravel_recodedGender_recodedDepartment_recodedenvironmentsatisfactionmonthl
yincomeJobrole_recodedOvertime_recoded
numcompaniesworkedyearssincelastpromotionyearswithcurrmanagerjobleveltrainingtimeslastyea
ryearswithcurrmanagerjoblevelyearsatcompanyhourlyrate
stockoptionlevelOvertime_recodedpercentsalaryhikemonthlyincome
  /CONTRAST (trainingtimeslastyear)=Indicator
  /CONTRAST (BusinessTravel_recoded)=Indicator
  /CONTRAST (Educationfieldrecoded)=Indicator
  /CONTRAST (Jobrole_recoded)=Indicator
  /CONTRAST (Overtime_recoded)=Indicator
  /PRINT=GOODFIT CI(95)
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

Model 2 used only the variables that we found to influence Attrition. The variables were entered shown in The Cox & Snell R square tests as well as Nagelkerke R square tests in the model summary table indicate the 23.4% to 39.9% variance in the DV is explained by the independent variables.



Model Summary

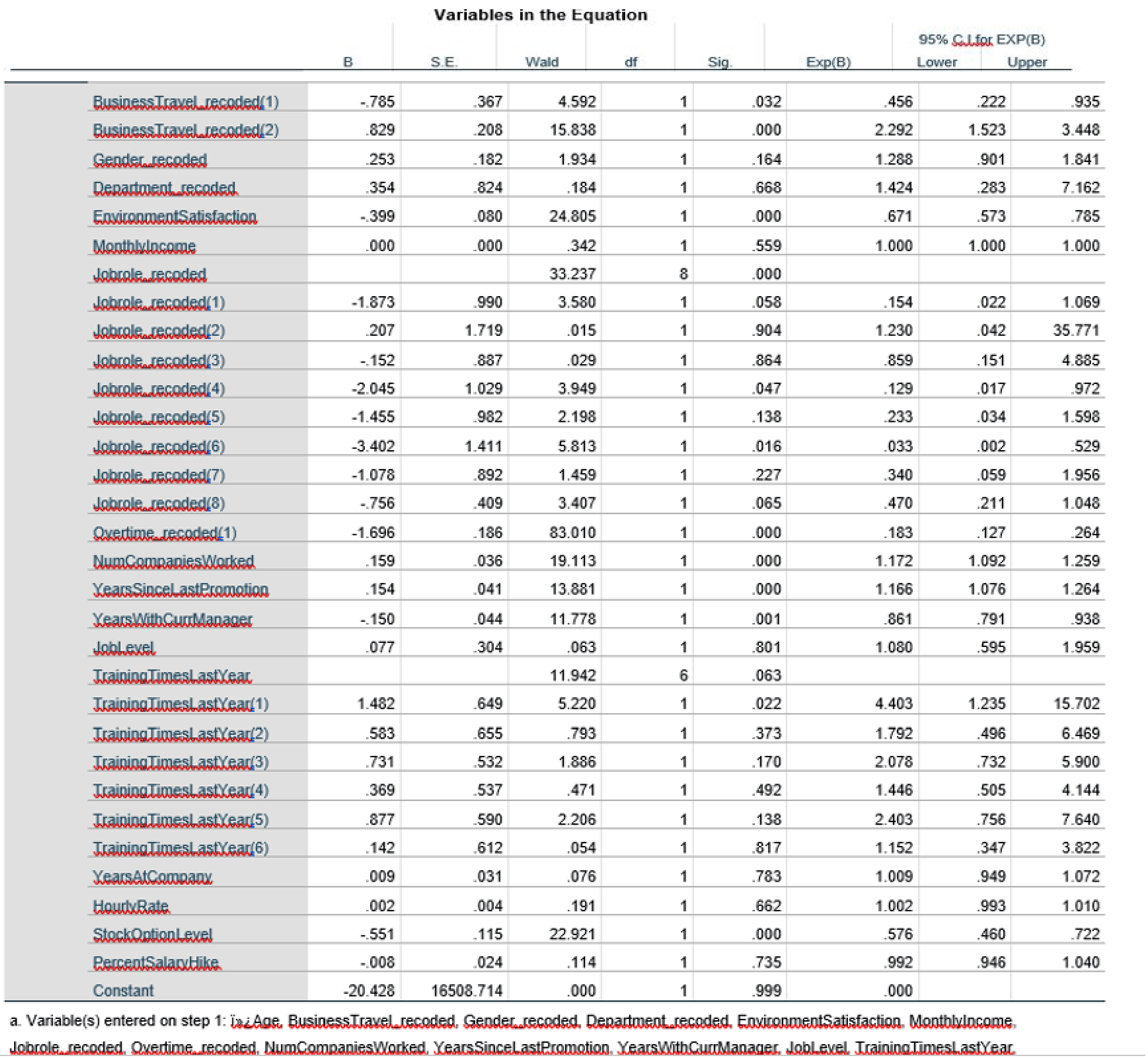
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	906.593[a]	.234	.399

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

The Hosmer and Lemeshow test is non- significant which indicates a good fit of the model.

Hosmer and Lemeshow Test

Step	Chi-square	Df	Sig.
1	14.711	8	.065



Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
BusinessTravel_recoded(1)	-.785	.367	4.592	1	.032	.456	.222	.935
BusinessTravel_recoded(2)	.829	.208	15.838	1	.000	2.292	1.523	3.448
Gender_recoded	.253	.182	1.934	1	.164	1.288	.901	1.841
Department_recoded	.354	.824	.184	1	.668	1.424	.283	7.162
EnvironmentSatisfaction	-.399	.080	24.805	1	.000	.671	.573	.785
MonthlyIncome	.000	.000	.342	1	.559	1.000	1.000	1.000
Jobrole_recoded			33.237	8	.000			
Jobrole_recoded(1)	-1.873	.990	3.580	1	.058	.154	.022	1.069
Jobrole_recoded(2)	.207	1.719	.015	1	.904	1.230	.042	35.771
Jobrole_recoded(3)	-.152	.887	.029	1	.864	.859	.151	4.885
Jobrole_recoded(4)	-2.045	1.029	3.949	1	.047	.129	.017	.972
Jobrole_recoded(5)	-1.455	.982	2.198	1	.138	.233	.034	1.598
Jobrole_recoded(6)	-3.402	1.411	5.813	1	.016	.033	.002	.529
Jobrole_recoded(7)	-1.078	.892	1.459	1	.227	.340	.059	1.956
Jobrole_recoded(8)	-.756	.409	3.407	1	.065	.470	.211	1.048
Overtime_recoded(1)	-1.696	.186	83.010	1	.000	.183	.127	.264
NumCompaniesWorked	.159	.036	19.113	1	.000	1.172	1.092	1.259
YearsSinceLastPromotion	.154	.041	13.881	1	.000	1.166	1.076	1.264
YearsWithCurrManager	-.150	.044	11.778	1	.001	.861	.791	.938
JobLevel	.077	.304	.063	1	.801	1.080	.595	1.959
TrainingTimesLastYear			11.942	6	.063			
TrainingTimesLastYear(1)	1.482	.649	5.220	1	.022	4.403	1.235	15.702
TrainingTimesLastYear(2)	.583	.655	.793	1	.373	1.792	.496	6.469
TrainingTimesLastYear(3)	.731	.532	1.886	1	.170	2.078	.732	5.900
TrainingTimesLastYear(4)	.369	.537	.471	1	.492	1.446	.505	4.144
TrainingTimesLastYear(5)	.877	.590	2.206	1	.138	2.403	.756	7.640
TrainingTimesLastYear(6)	.142	.612	.054	1	.817	1.152	.347	3.822
YearsAtCompany	.009	.031	.076	1	.783	1.009	.949	1.072
HourlyRate	.002	.004	.191	1	.662	1.002	.993	1.010
StockOptionLevel	-.551	.115	22.921	1	.000	.576	.460	.722
PercentSalaryHike	-.008	.024	.114	1	.735	.992	.946	1.040
Constant	-20.428	16508.714	.000	1	.999	.000		

a. Variable(s) entered on step 1: ï»¿Age, BusinessTravel_recoded, Gender_recoded, Department_recoded, EnvironmentSatisfaction, MonthlyIncome, Jobrole_recoded, Overtime_recoded, NumCompaniesWorked, YearsSinceLastPromotion, YearsWithCurrManager, JobLevel, TrainingTimesLastYear,

Interpreting the odds ratio for Model 2

- The employees who travel frequently have lower odds of attrition than employees who do not travel at all. The odds of attrition for such employees are 0.456 times that of employees who do not travel.
- The employees who travel rarely have higher odds of attrition than employees who do not travel at all. The odds of attrition for such employees are 2.292 times that of employees who do not travel.
- A unit increase in environment satisfaction of an employee lowers the odds of attrition of an employee .671 times.
- The employees who have a manufacturing director, scientist and sales representative as job roles have lower odds (.129, .033 and .470 respectively) of attrition than those employees who have a job role as a healthcare representative.

- The employees who work overtime have lower odds (.183) of attrition than those employees that do not work overtime.
- A unit increase in the number of companies an employee works increases the odds of attrition of an employee 1.172 times.
- A unit increase in the number of years an employee spends after last promotion increases the odds of attrition of an employee 1.166 times.
- A unit increase in the number of years an employee spends with the current manager lowers the odds of attrition of an employee .861 times.
- The employees who have trained once in last year have higher odds (4.403) of attrition than those employees who not trained at all last year.
- A unit increase in the stock option level of an employee lowers the odds of attrition of an employee .576 times.

BUSINESS SOLUTION

Based on the results shown in the statistical analysis, we propose the following solutions to the HR unit of IBM.

We recommend that IBM changes its business process so that their controllable variables can be managed to prevent their negative impact on employee attrition. The employees should be given better opportunities to travel. We suggest that employees are aware of these travel opportunities in the hiring process.

We also recommend that Managers work to create a comfortable work environment. For example, commodities such as vending machines, break rooms, and other luxuries could be added to improve the current work environment. During the hiring process, the new employee should be aware of these luxuries to see that IBM provides a comfortable work environment for its employees.

The working condition of Employees under the following roles: Manufacturing director, sales executive and research scientist should be investigated and improved upon to lower the chances of attrition and enhance satisfaction.

Offering more overtime to employees and new hires is another factor which could lower the chances of attrition. This could be considered in the hiring process since offering overtime to employees could reduce the number of employees needed.

Interestingly employees who have been exposed to different company experiences have higher chances of attrition. A good step would be to strike a balance between employees having high experience in different companies and employees who spend great number of years working with one manager.

Another important factor to consider for reducing the attrition rate of employees is frequent promotion. Increases training and stock option level are also very important factors for lowering the chances of attrition for employees.

4.0 RECOMMENDATIONS: THE IMPROVE PHASE

The following variables play an important role in determining the attrition of an employee

- Business travel
- Environment Satisfaction
- Overtime
- Job Satisfaction
- Job role
- Stock option level
- Promotion
- Number of companies worked

The HR department should focus on these variables and redesign their business process to accommodate measures that will curb the rate of attrition

Anonymous questionnaires should be deployed frequently to employees in order to understand the root cause attrition in the company.

Data collected via questionnaires ought to be analyzed from time to time to discover new findings that will be of importance to HRs decision regarding attrition.

The analytics team can take this a step further by predicting the likelihood and rate of attrition based on historical data collected. This will be useful for planning amongst the HR team.

Alternative Solutions Considered

Recommended Solution

We recommend training for IBM managers adjust the setting of key process input variables to control the input variables which they know will increase their employee attrition. The employee hiring process will need to be redesigned so that variables are only considered in which we deemed not to affect employee attrition.

Historically, improvement recommendations fall into a few general categories.

5.0 MONITORING AND CONTROL: THE CONTROL PHASE

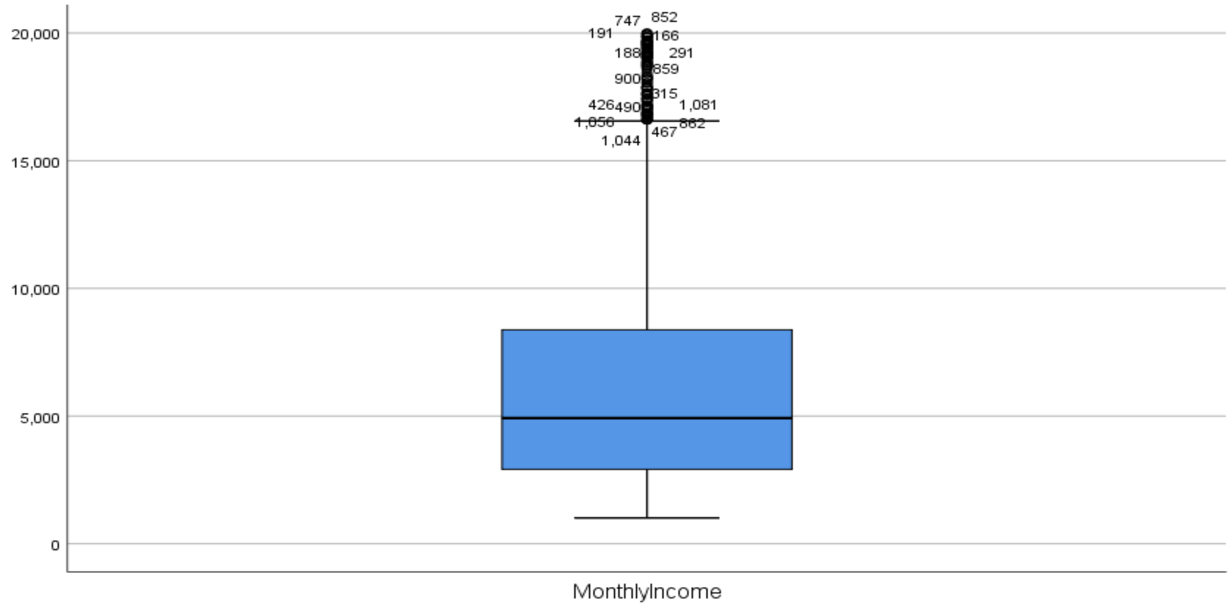
- Training of the HR team will be required in view of the findings from this report.
- The HR quality control procedures will have to be reviewed to accommodate new measures in respect to the result of this analysis.
- The process owner will be responsible for the review and execution of these measures.
- Control charts should be redesigned and made visible to every member of the HR team to serve as a reminder and monitor performance. [2]
- Weekly review meetings should be conducted to check progress based on feedback from employees.
- The improved process and steps listed above should be standardized.
- Quality Control Procedures must be documented in line with the review. HR Manager and Process Owner should append their signature after each review and date captured. [2]
- External auditors can be deployed to check the activities of the company in order to erase any form of bias.
- The HR process must be reviewed to accommodate new measures put in place to prevent attrition of key employees.

6.0 SUMMARY/CONCLUSION

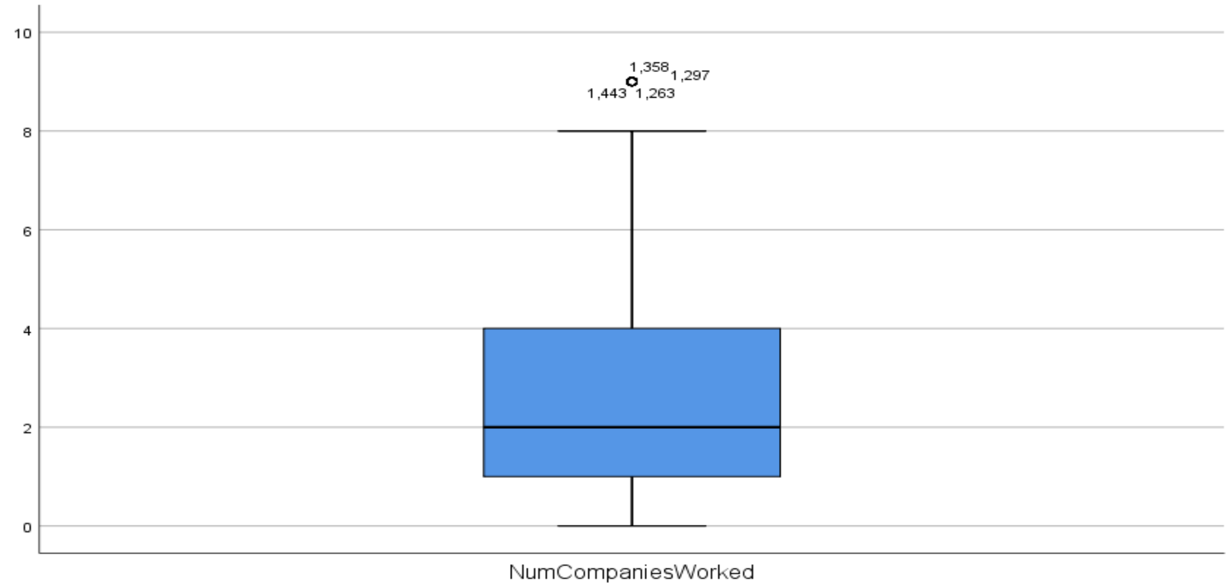
Employees are the backbone of any organization and the performance of an organization is dependent on retaining quality employees. Attrition is a problem that affects all businesses irrespective of their processes as it is has an impact on productivity, profit and time. Our focus on this report was to help the IBM human resources find resignation trend/numbers to understand the factors leading to employee attrition in the organization and provide solutions to management that will control attrition. We looked specifically at the on boarding vs the recruiting process map for employees of IBS and ran binary logistic regression using SPPS and Python after which we used metrics such as R squared, goodness of fit, and odds ratio to determine which variables contribute the most to attrition, and influence employee behavior. The following contributing factors were discovered after running our analysis, Business travel, Environment Satisfaction, Overtime, Job Satisfaction, job role, Stock option level, Promotion and Number of companies worked. We then suggested these latent factors to IBM managers, provided solution and recommendation to Managers on the controllable factors that should be reconsidered, and directed them to better control their recruiting and on boarding processes. The analysis can be manipulated from time to time to factor in new data from employees and discover more insights.

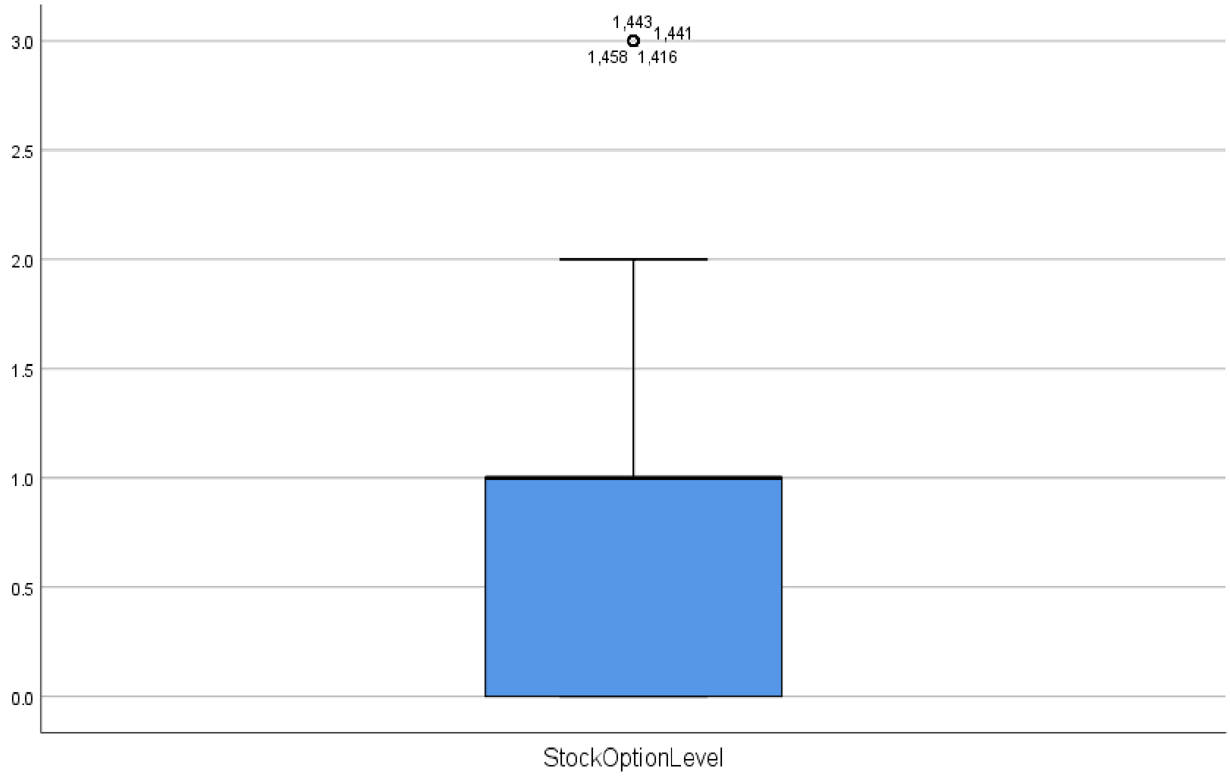
APPENDICES

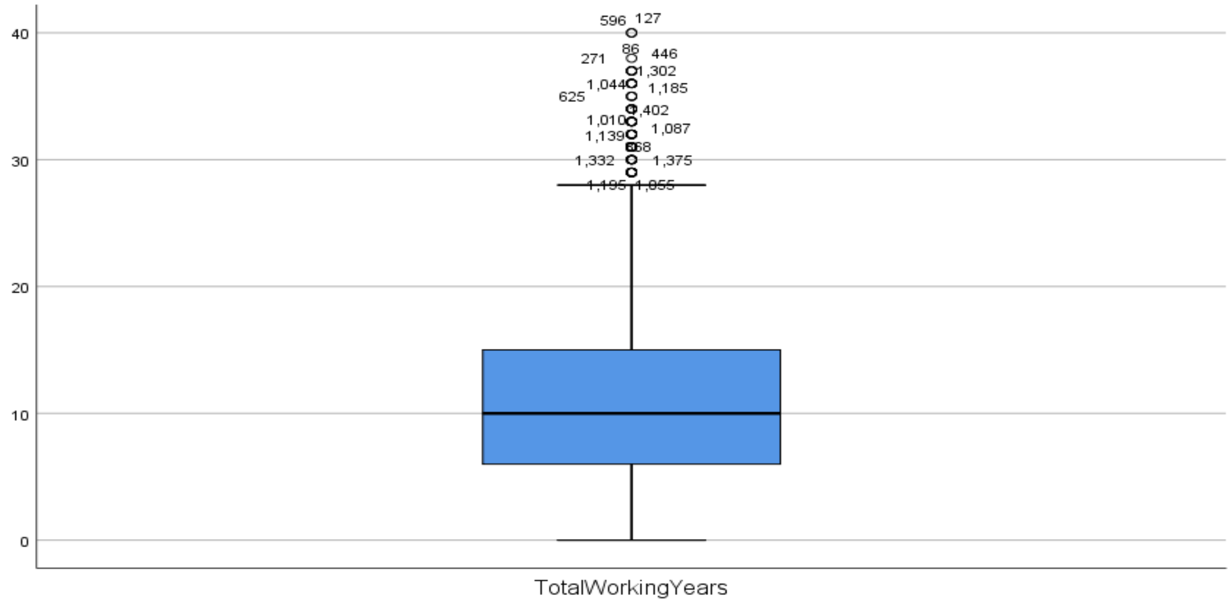
APPENDIX A

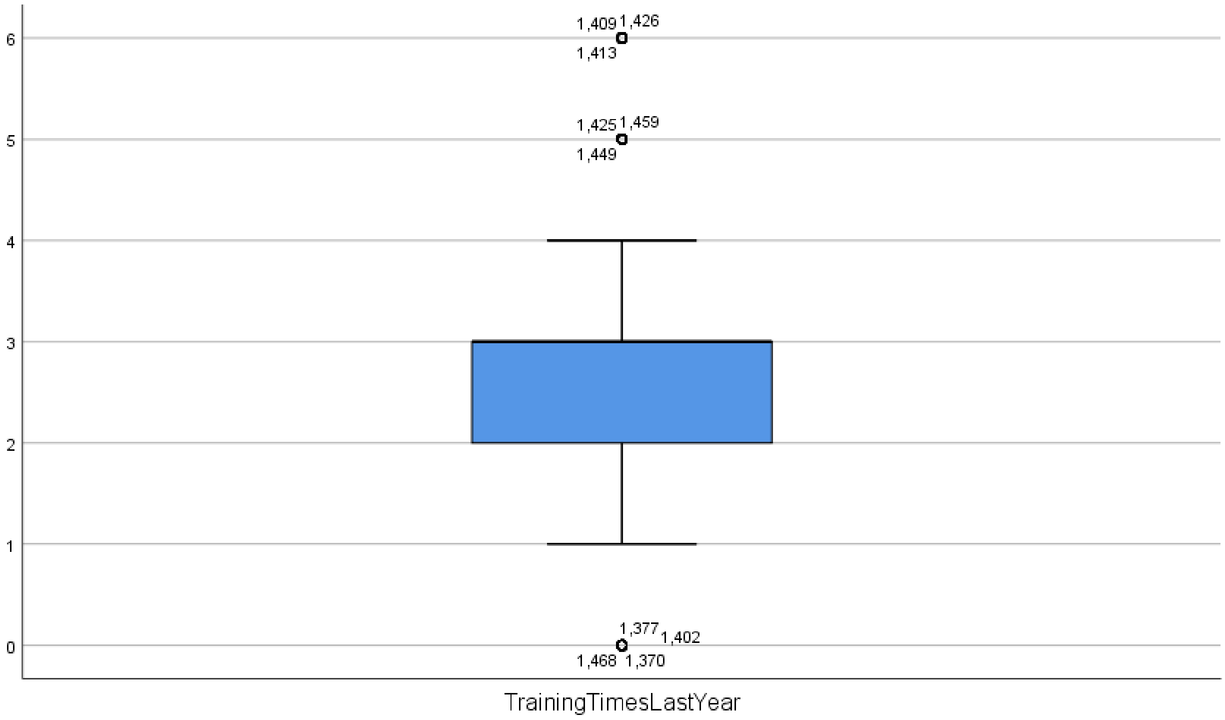


There were outliers according to the univariate analysis since the scores were beyond upper and lower quadrantile, this is shown in the box plots in the rest of appendix B.

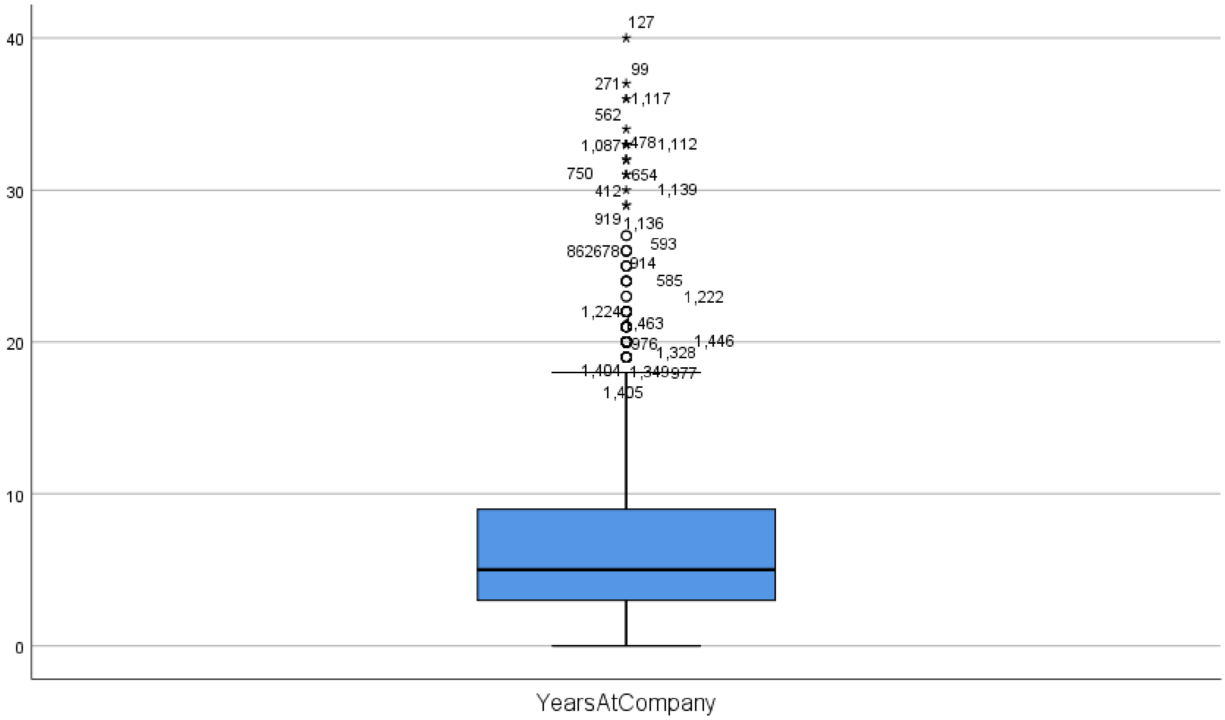



1,443
1,441
1,458
1,416
3.0
2.5
2.0
1.5
1.0
0.5
0.0
StockOptionLevel

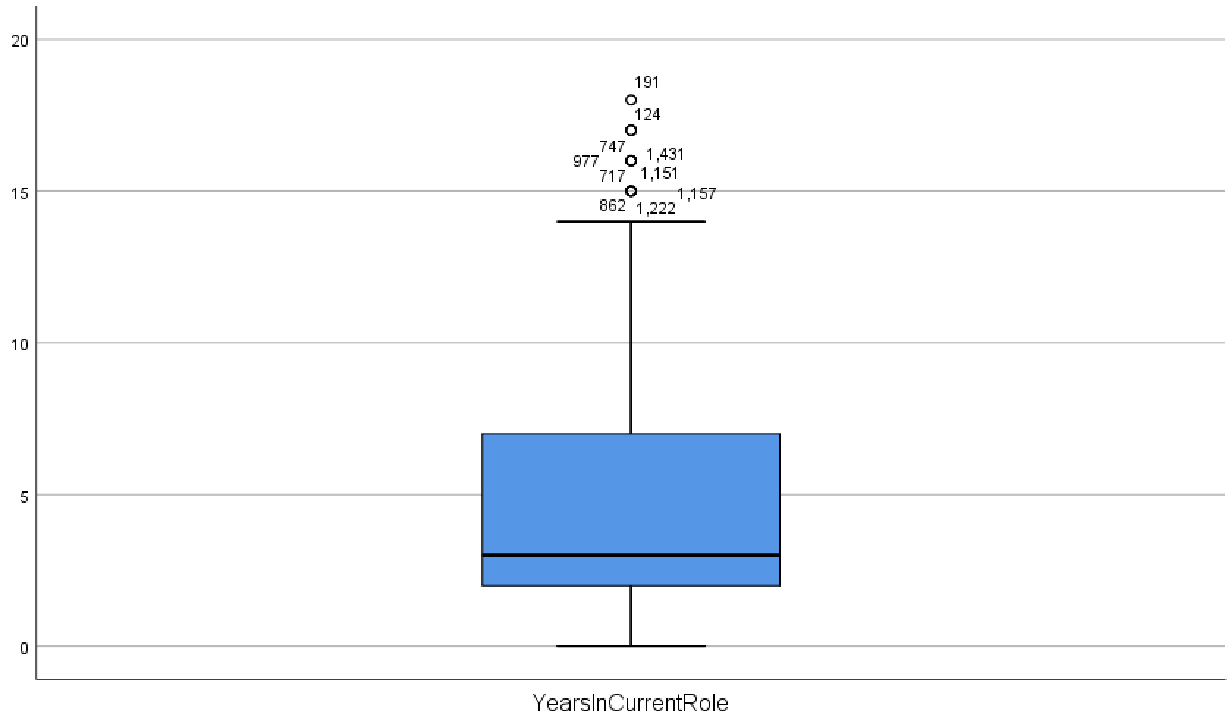

596
127
86
271
446
1,302
1,044
1,185
625
1,402
1,010
1,087
1,139
868
1,332
1,375
1,195
1,055
40
30
20
10
0
TotalWorkingYears

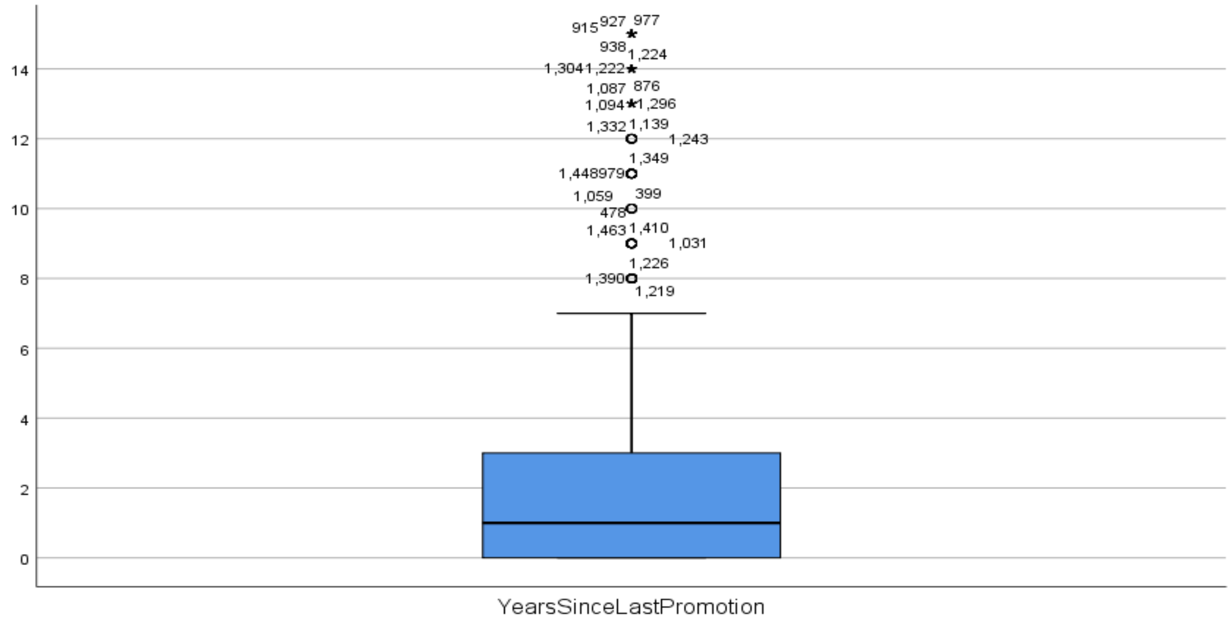


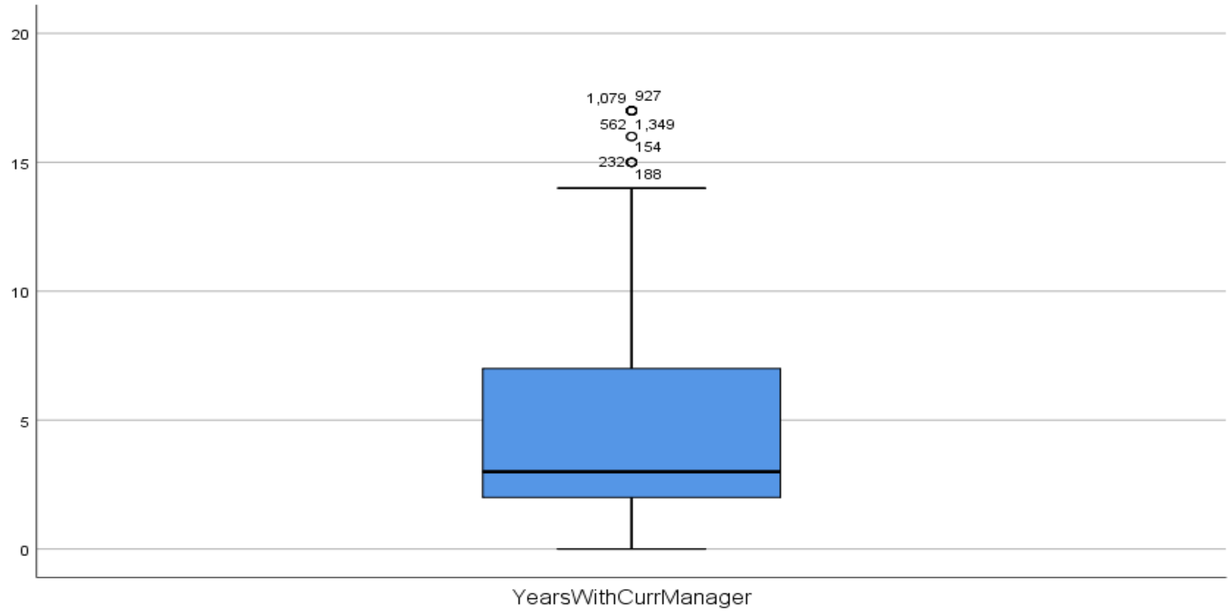
6
5
4
3
2
1
0
1,409
1,426
1,413
1,425
1,459
1,449
1,377
1,402
1,468
1,370
TrainingTimesLastYear



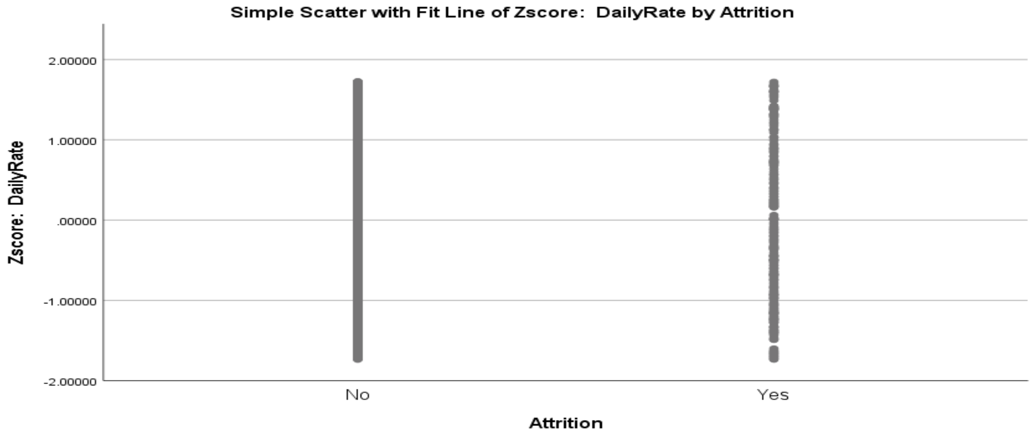
40
30
20
10
0
127
99
271
1,117
562
1,087
478
1,112
750
654
412
1,139
919
1,136
593
862
678
914
585
1,222
1,224
1,463
976
1,328
1,446
1,404
1,349
977
1,405
YearsAtCompany


20
15
10
5
0
191
124
747
977
1,431
717
1,151
1,157
862
1,222
YearsInCurrentRole

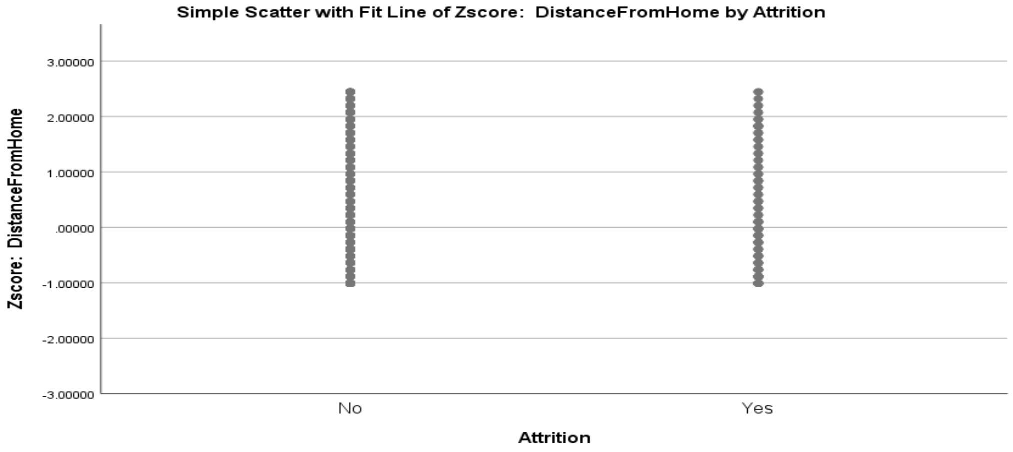

14
12
10
8
6
4
2
0
915
927
977
938
1,224
1,3041,222
1,087
876
1,094
1,296
1,332
1,139
1,243
1,349
1,448979
1,059
399
478
1,463
1,410
1,031
1,226
1,390
1,219
YearsSinceLastPromotion



Appendix B

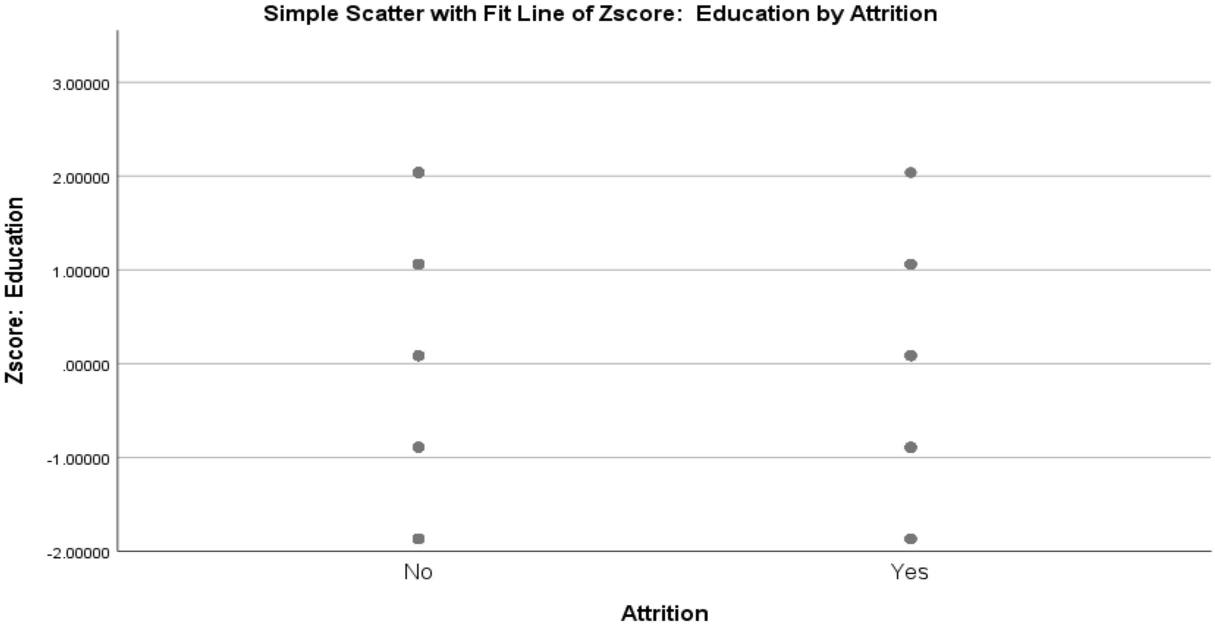


#Attrition, #DailyRate
There are no outliers detected since every point has a z score falling between -3 and 3 shown in the graph above.



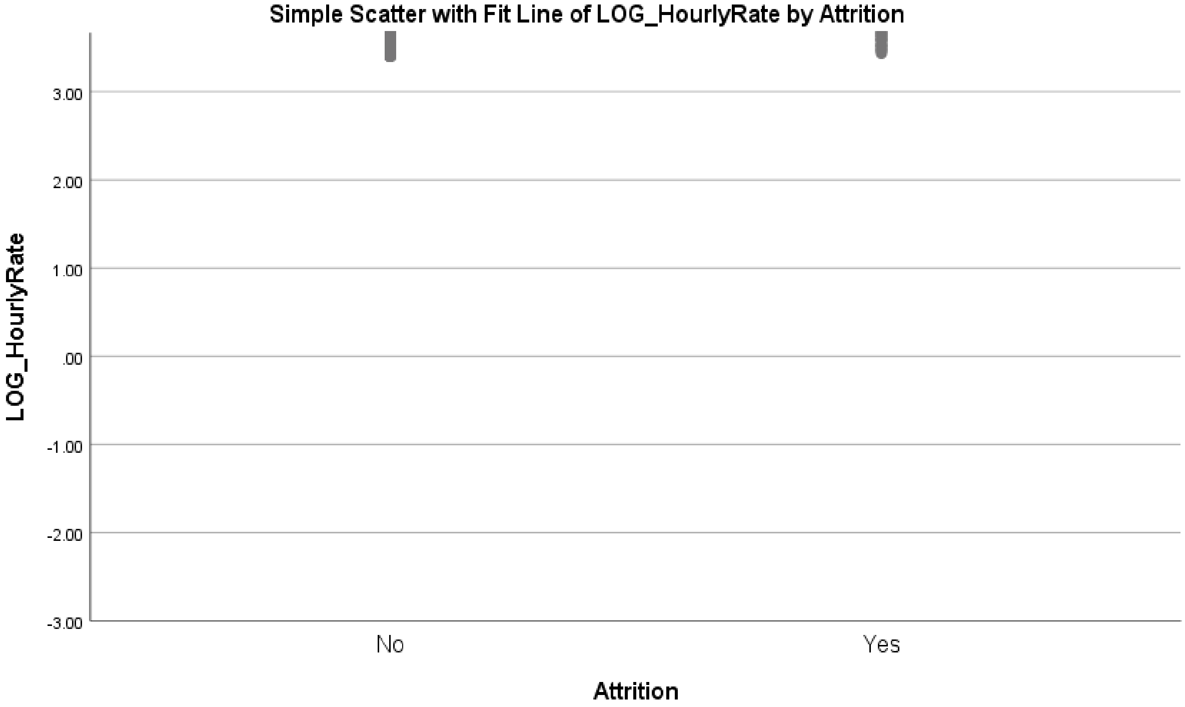
#Attrition,#DistanceFromHome

There were no outliers because every z score fell between -3 and 3.

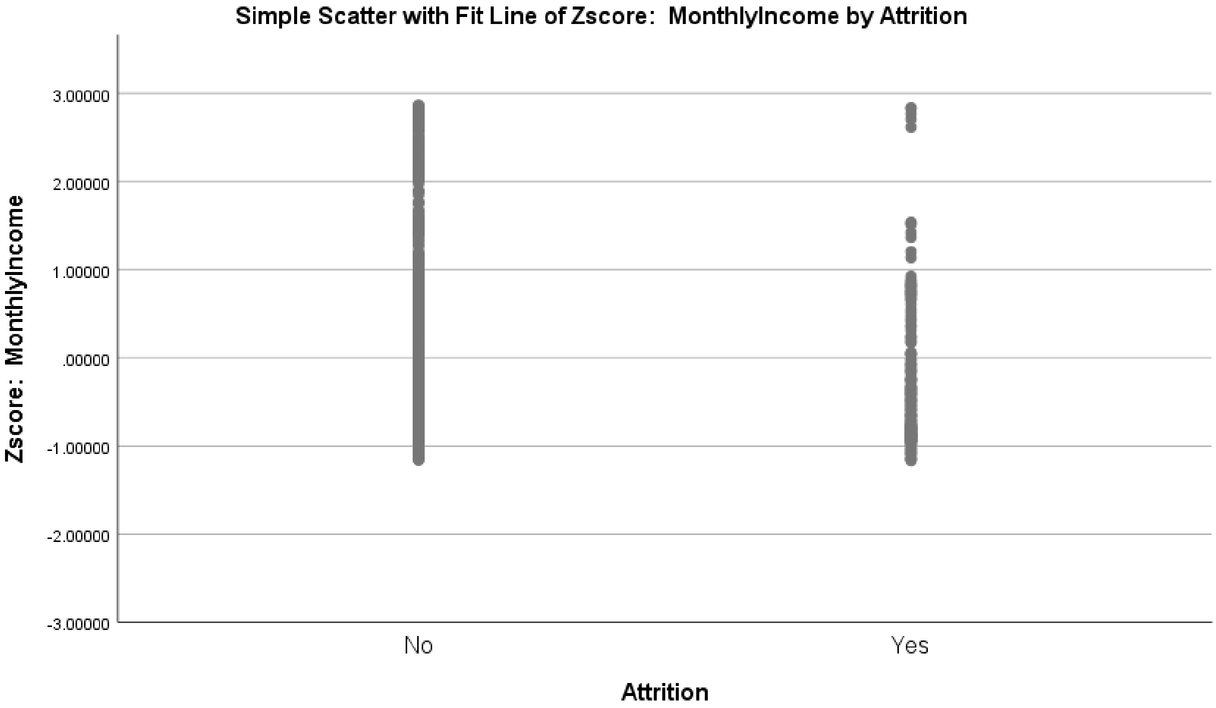


# Attrition,#education

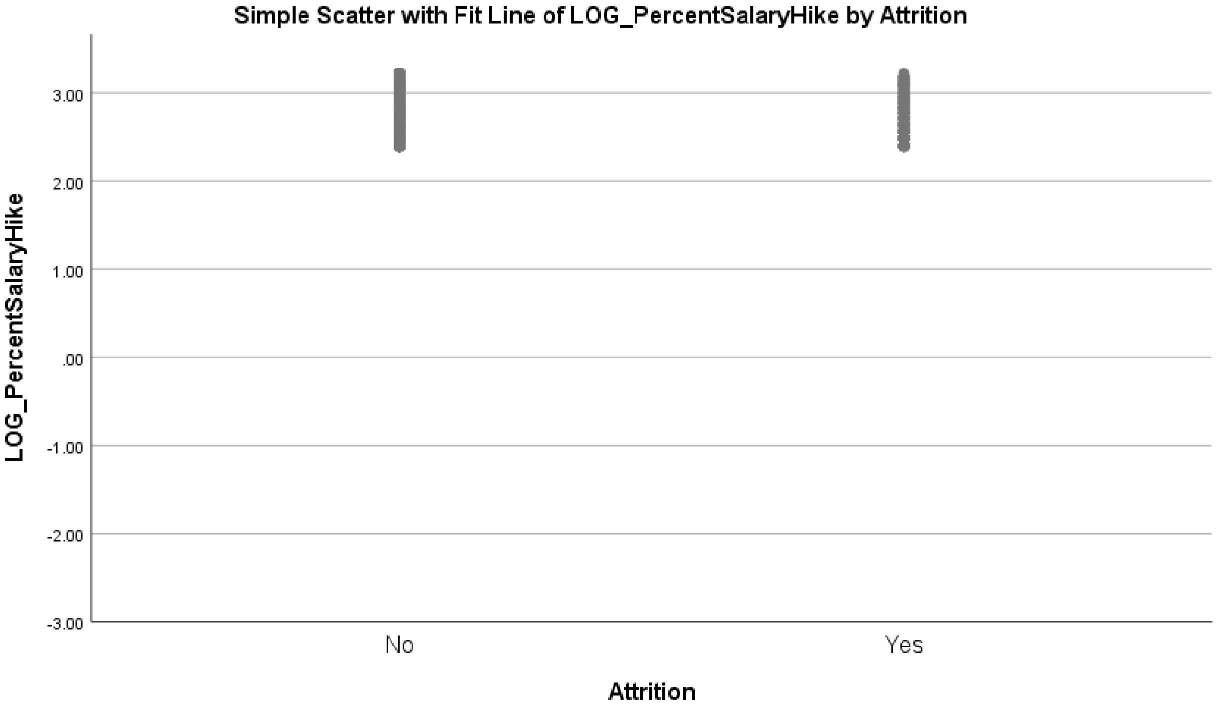
# Result: No outlier because every point is between -3 and 3.



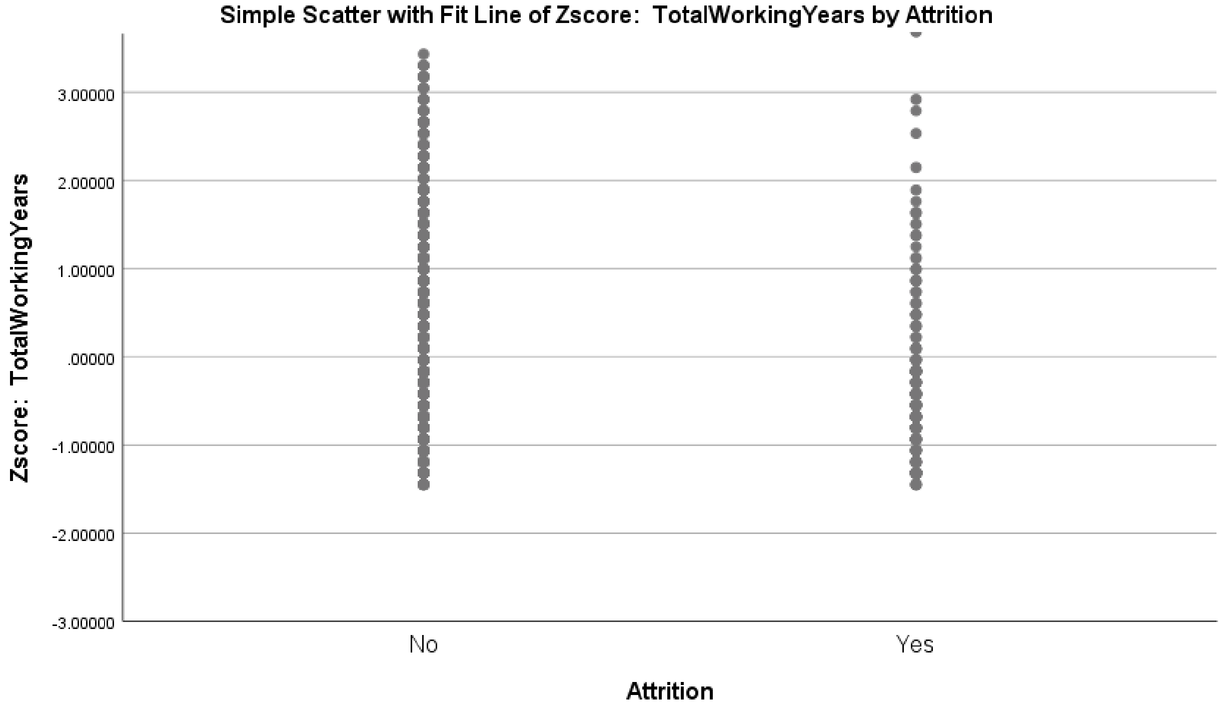
#Attrition,#Hourlyrate



#Attrition,#MonthlyIncome

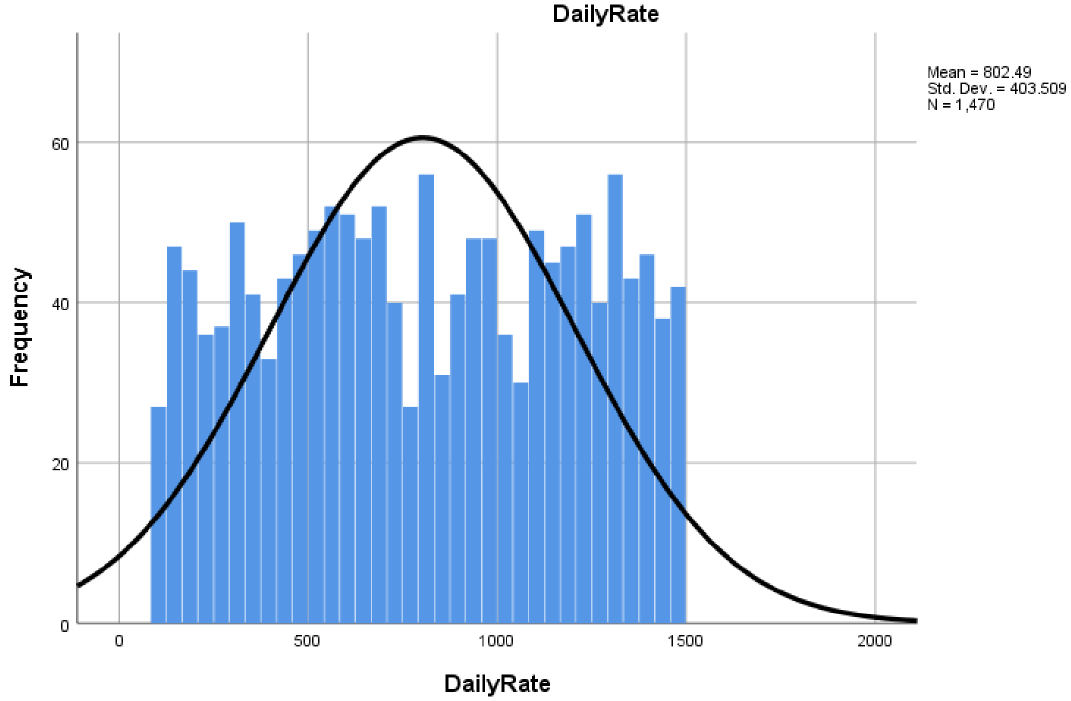


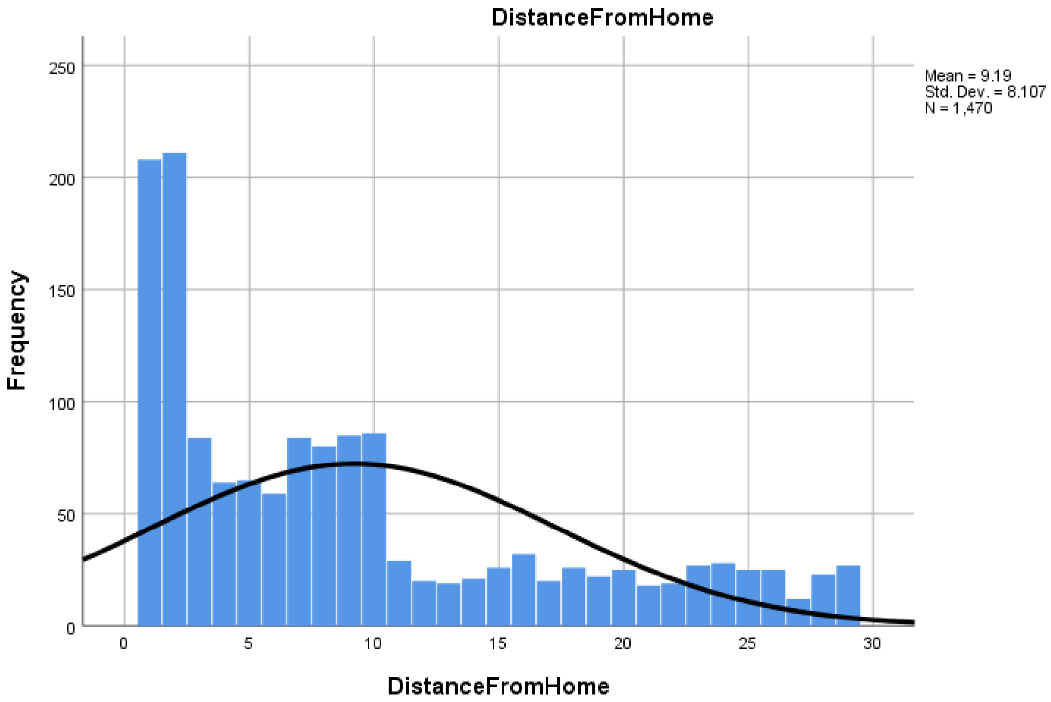
#Attriton, #PercentSalaryHike

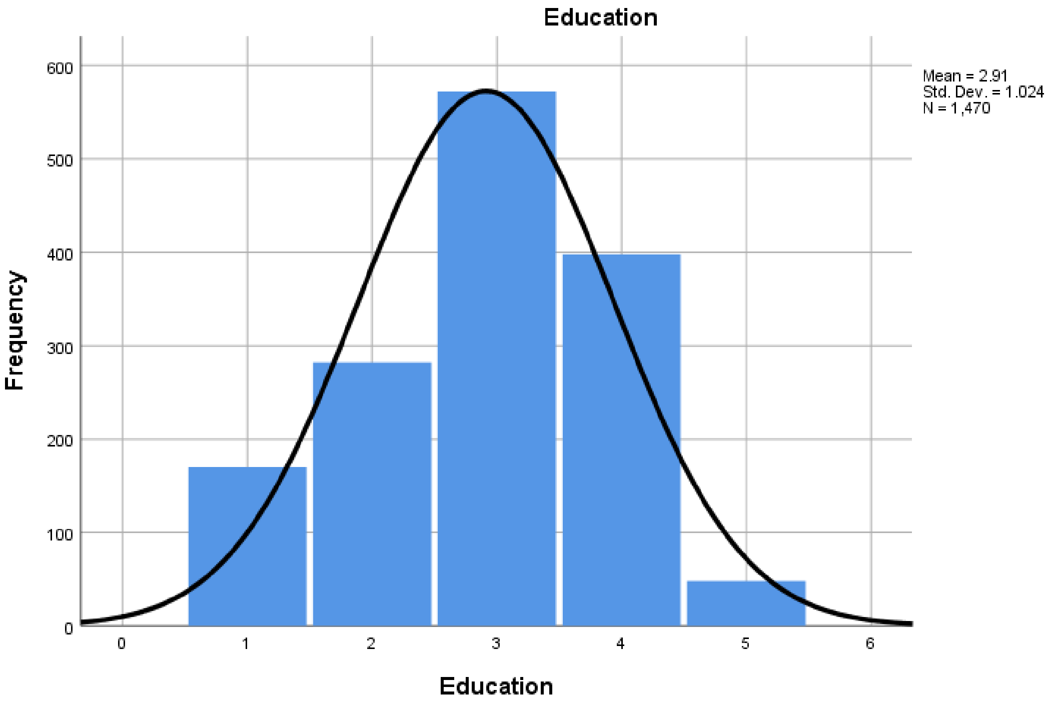


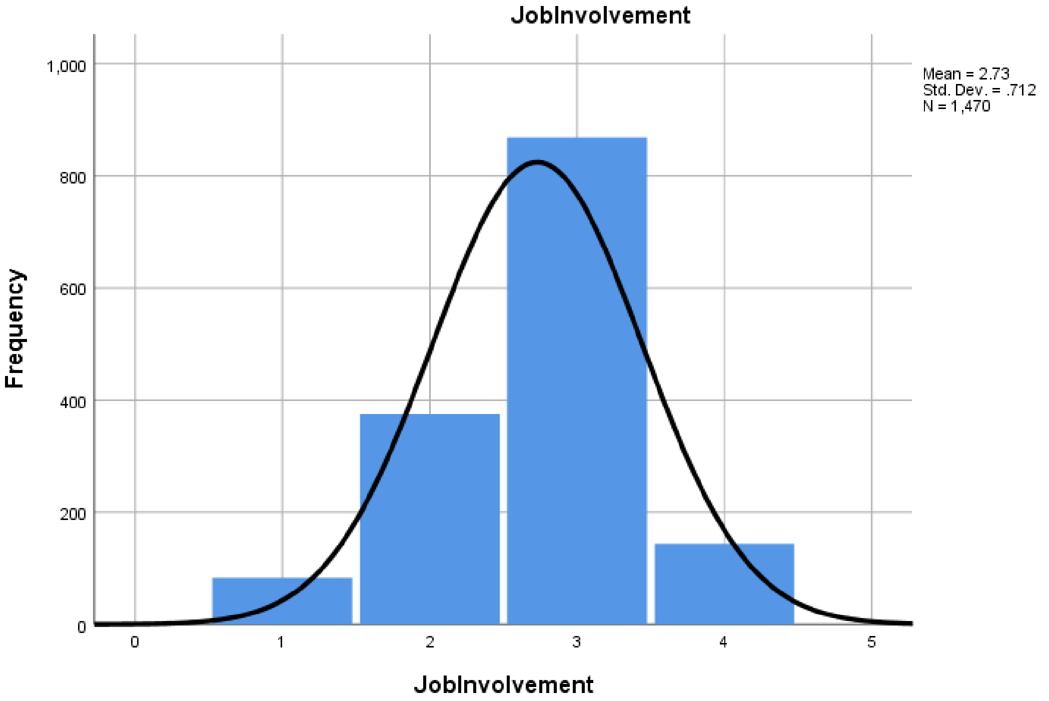
#Attrition, #TotalWorking Years

Appendix C




DistanceFromHome
Mean = 9.19
Std. Dev. = 8.107
N = 1,470
Frequency
0
50
100
150
200
250
0
5
10
15
20
25
30
DistanceFromHome


Education
Frequency
0
100
200
300
400
500
600
0
1
2
3
4
5
6
Education
Mean = 2.91
Std. Dev. = 1.024
N = 1,470


JobInvolvement
Frequency
0
200
400
600
800
1,000
0
1
2
3
4
5
JobInvolvement
Mean = 2.73
Std. Dev. = .712
N = 1,470

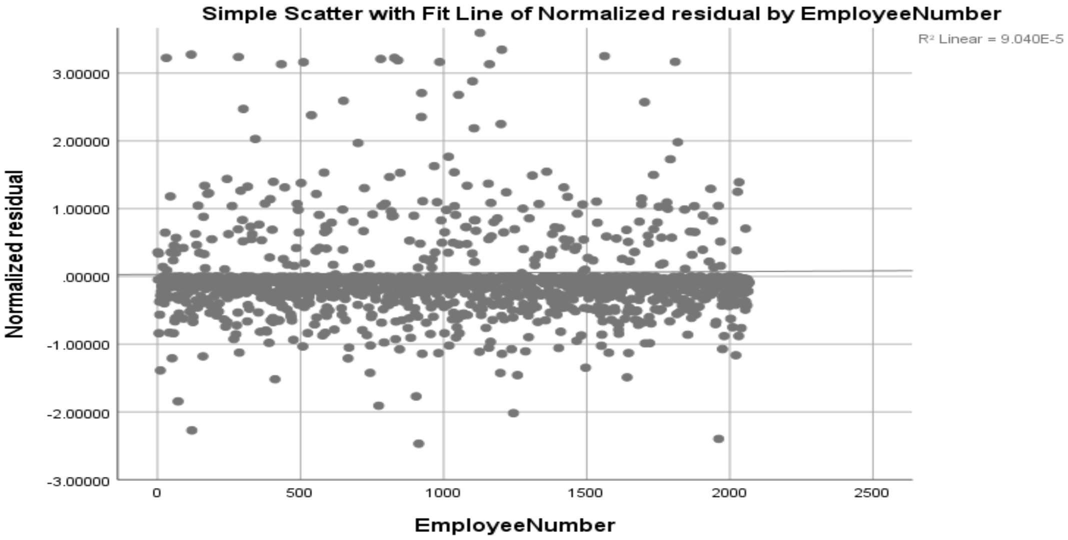
Appendix D

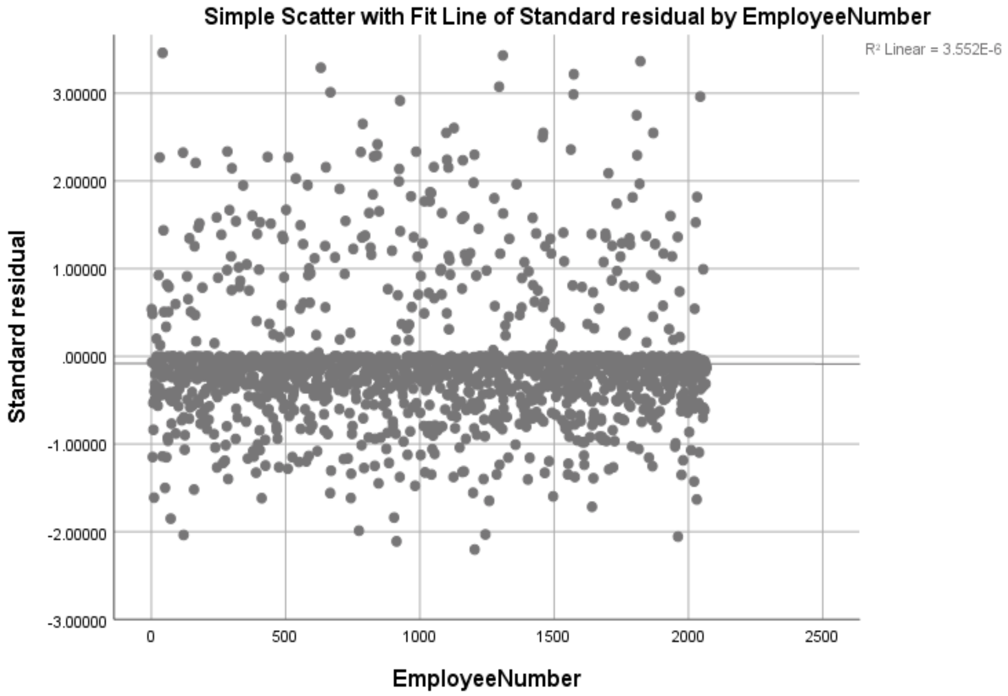
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
LOG_DailyRate	.113	1426	.000	.904	1426	.000
LOG_DistanceFromHome	.112	1426	.000	.934	1426	.000
LOG_EmployeeNumber	.157	1426	.000	.825	1426	.000
LOG_HourlyRate	.095	1426	.000	.943	1426	.000
LOG_MonthlyIncome	.063	1426	.000	.967	1426	.000
LOG_monthlyrate	.110	1426	.000	.917	1426	.000
LOG_yearsatcompany	.114	1426	.000	.958	1426	.000
LOG_PercentSalaryHike	.155	1426	.000	.929	1426	.000
LOG_StandardHours	.	1426	.	.	1426	.
LOG_totalworkingyears	.112	1426	.000	.934	1426	.000

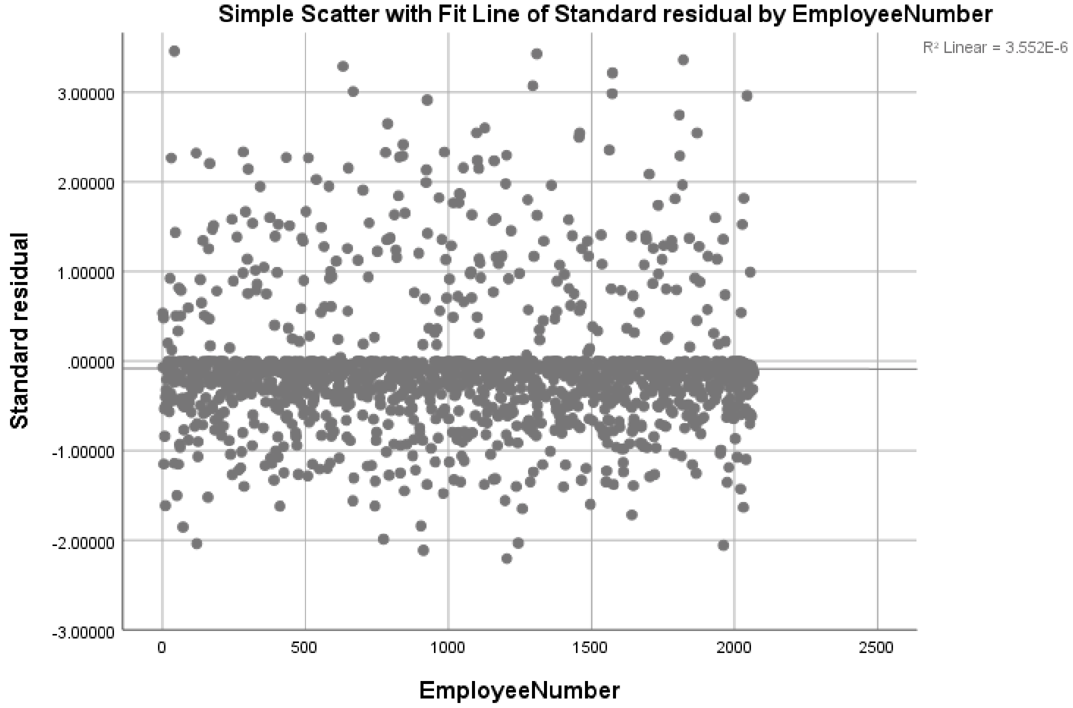
a. Lilliefors Significance Correction

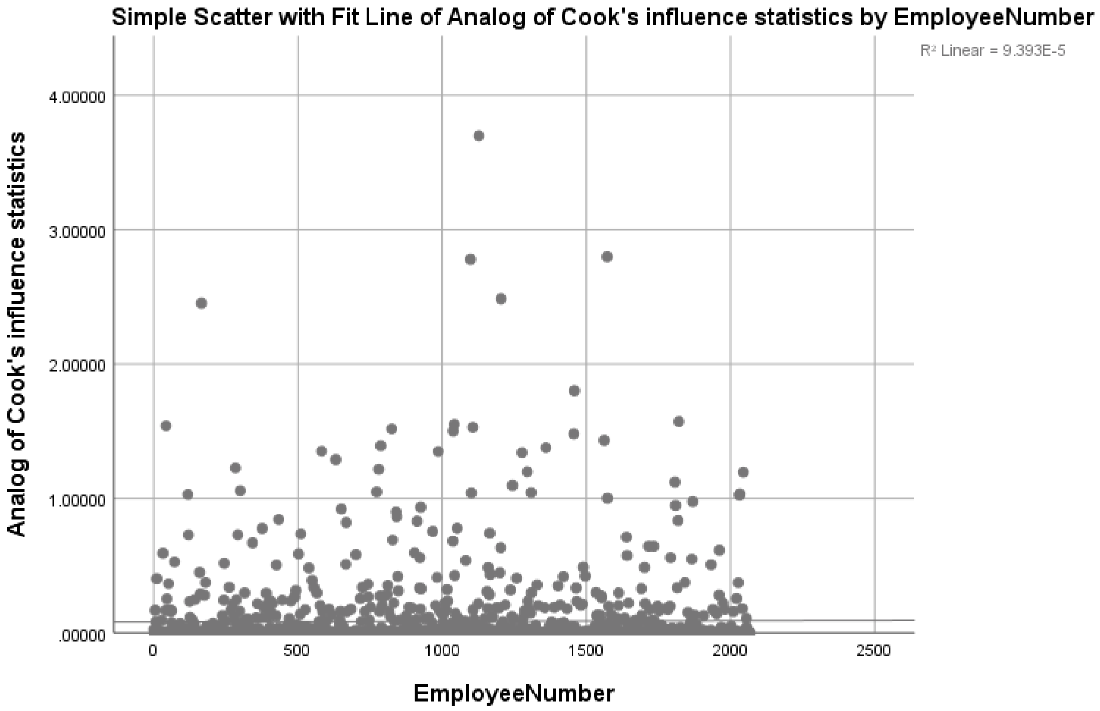
Appendix E





There were some outliers as none of the observation had a studentized residual that were some observation were larger than 3 (in absolute value).





Coding

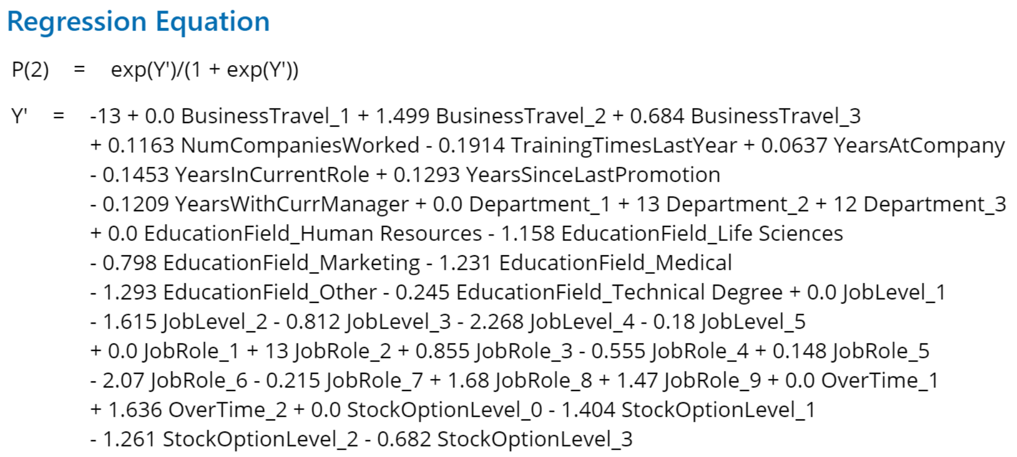
Appendix F

Coding scheme of categorical variables:

```
RECODE attrition ('No'='0') ('Yes'='1').

EXECUTE.


RECODE businesstravel ('Non-Travel'=1) ('Travel_Frequently'=2) ('Travel_Rarely'=3) INTO

BusinessTravel_recoded.

EXECUTE.


RECODE educationfield ('Human Resources'=1) ('Life Sciences'=2) ('Marketing'=3) ('Medical'=4)

   ('Other'=5) ('Technical Degree'=6) INTO Educationfieldrecoded.

EXECUTE.


RECODE department ('Human Resources '=1) ('Research & Development '=2) ('Sales '=3) INTO

Department_recoded.
```

```
EXECUTE.


RECODE gender ('Female'=1) ('Male'=2) INTO Gender_recoded.

EXECUTE.


RECODE jobrole ('Healthcare Representative'=1) ('Human Resources '=2) ('Laboratory Technician '=3)

  ('Manager '=4) ('Manufacturing Director '=5) ('Research Director '=6) ('Research Scientist '=7)

  ('Sales Executive '=8) ('Sales Representative '=9) INTO Jobrole_recoded.

EXECUTE.


RECODE overtime ('Yes'=1) ('No'=0) INTO Overtime_recoded.

EXECUTE.
```

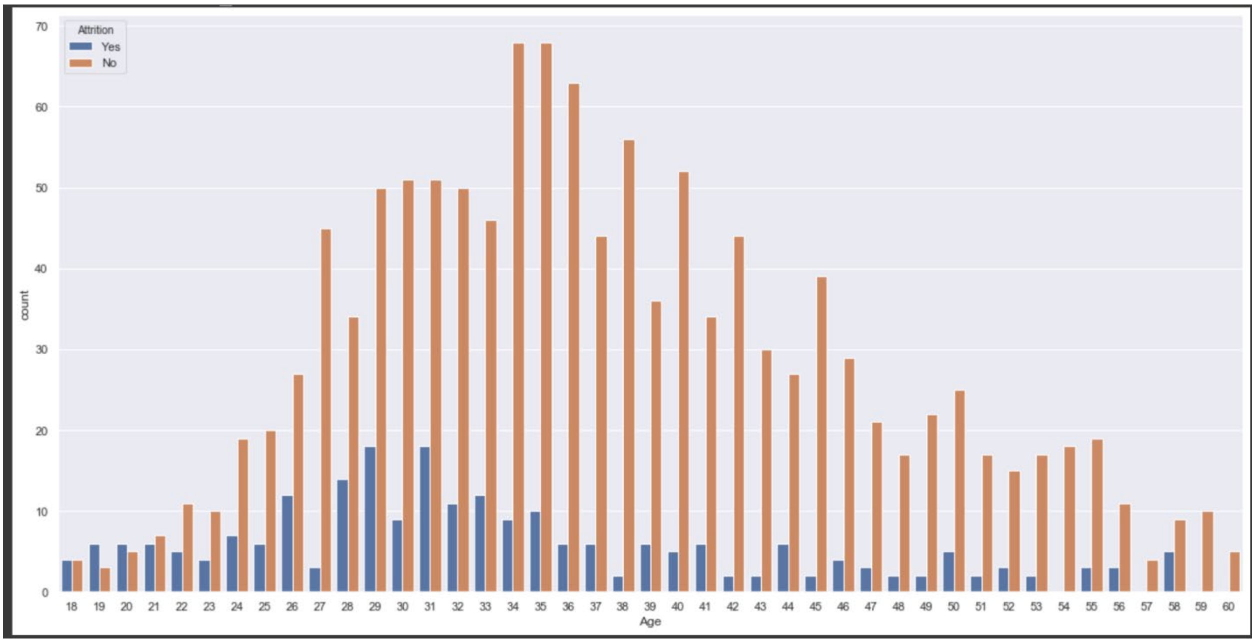
Appendix G



Regression Equation

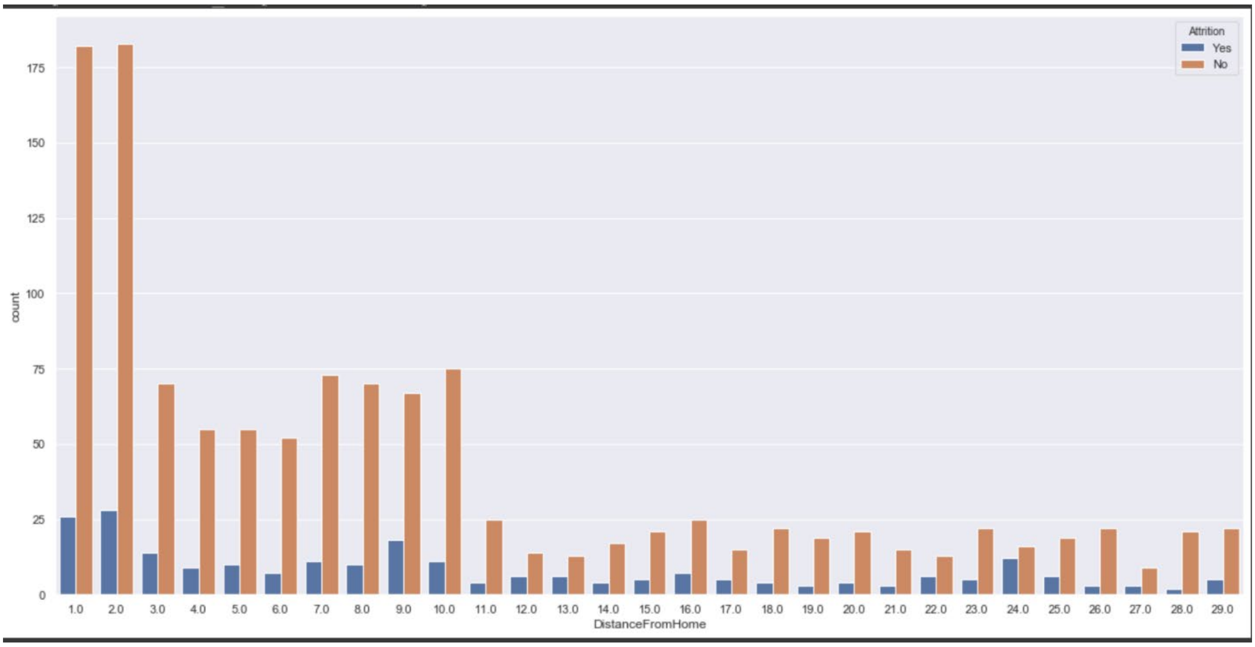
```
P(2)   =   exp(Y')/(1 + exp(Y'))

Y'   =   -13 + 0.0 BusinessTravel_1 + 1.499 BusinessTravel_2 + 0.684 BusinessTravel_3
         + 0.1163 NumCompaniesWorked - 0.1914 TrainingTimesLastYear + 0.0637 YearsAtCompany
         - 0.1453 YearsInCurrentRole + 0.1293 YearsSinceLastPromotion
         - 0.1209 YearsWithCurrManager + 0.0 Department_1 + 13 Department_2 + 12 Department_3
         + 0.0 EducationField_Human Resources - 1.158 EducationField_Life Sciences
         - 0.798 EducationField_Marketing - 1.231 EducationField_Medical
         - 1.293 EducationField_Other - 0.245 EducationField_Technical Degree + 0.0 JobLevel_1
         - 1.615 JobLevel_2 - 0.812 JobLevel_3 - 2.268 JobLevel_4 - 0.18 JobLevel_5
         + 0.0 JobRole_1 + 13 JobRole_2 + 0.855 JobRole_3 - 0.555 JobRole_4 + 0.148 JobRole_5
         - 2.07 JobRole_6 - 0.215 JobRole_7 + 1.68 JobRole_8 + 1.47 JobRole_9 + 0.0 OverTime_1
         + 1.636 OverTime_2 + 0.0 StockOptionLevel_0 - 1.404 StockOptionLevel_1
         - 1.261 StockOptionLevel_2 - 0.682 StockOptionLevel_3
```

Appendix I



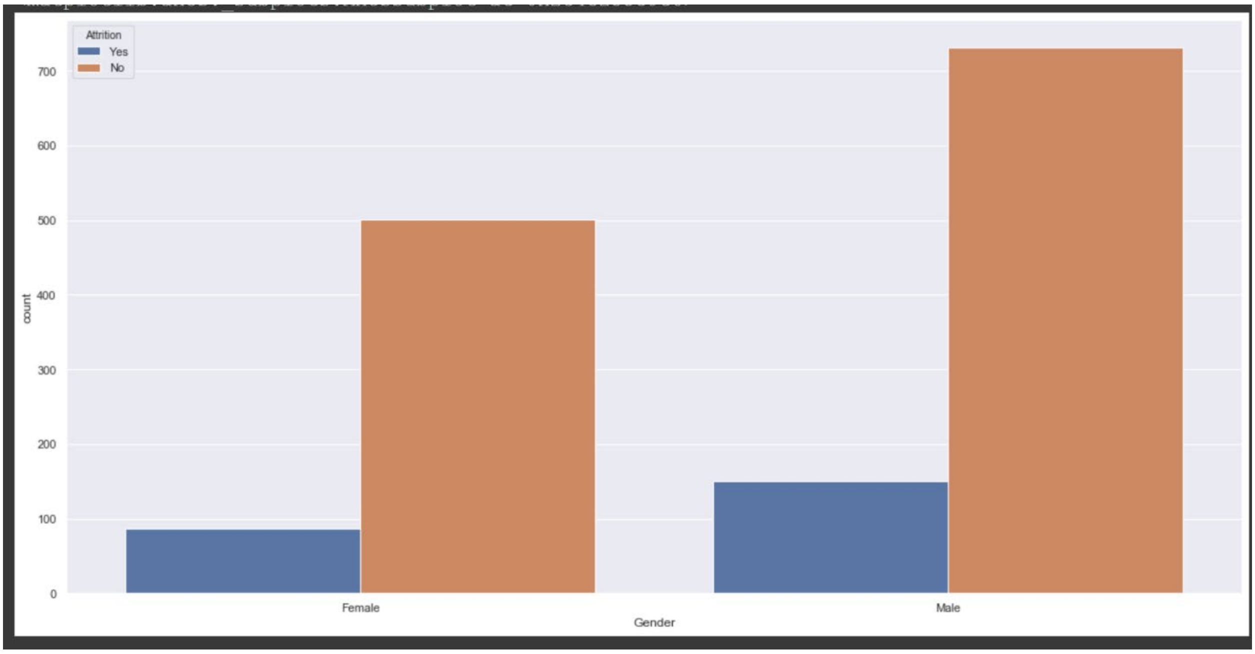
We can see a few upward trend lines with age and attriton. Attrition rates seem to increase for IBM from ages 25-26, 28-29 and between 30-31

DISTANCE FROM HOME



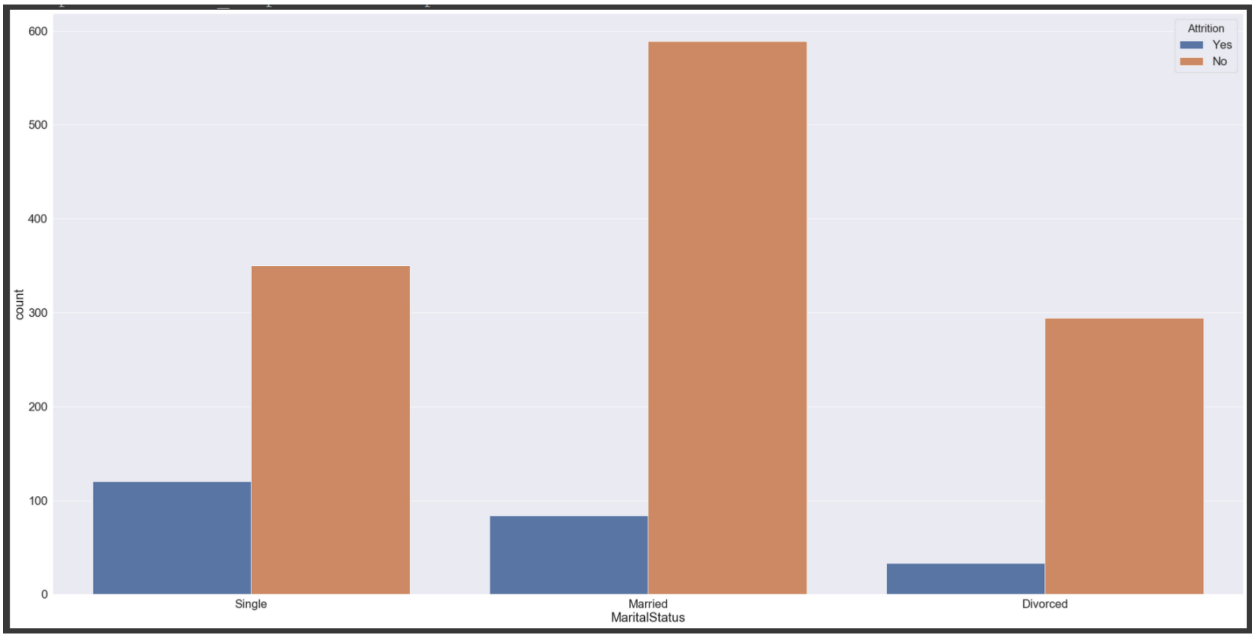
Distance from home seems to have no affect on attrition for IBM.

SEX



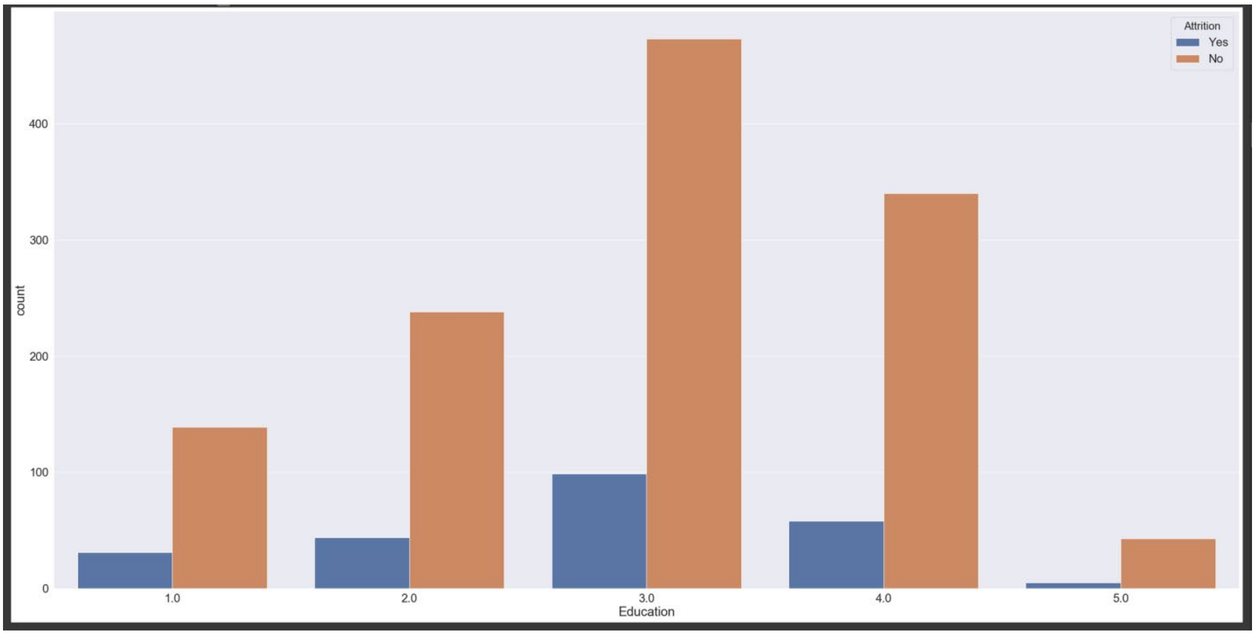
Gender is shown to affect attrition rates for IBM. Male employees for IBM tend to leave at higher rates than females.

MARITAL STATUS



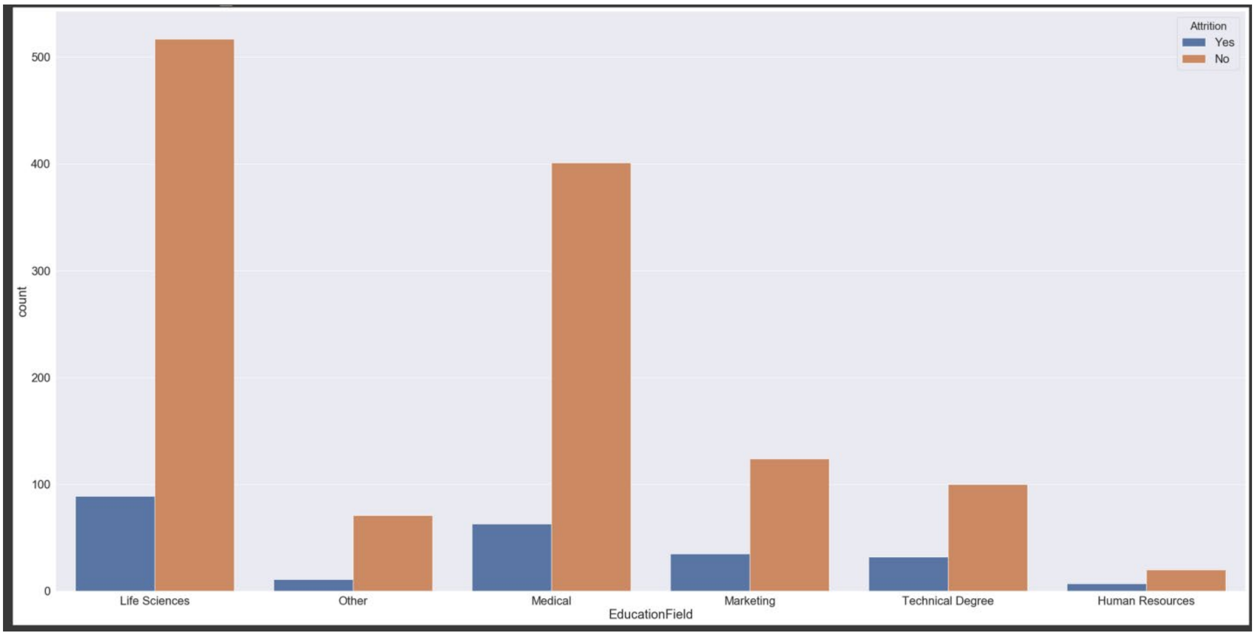
Here we can see that marital status influences attrition for IBM. Single employees tend to have higher attrition rates compared to married employees.

EDUCATION



We can see that education level influences IBM employee attrition rates. People that hold a bachelors degree show the highest attrition rates.

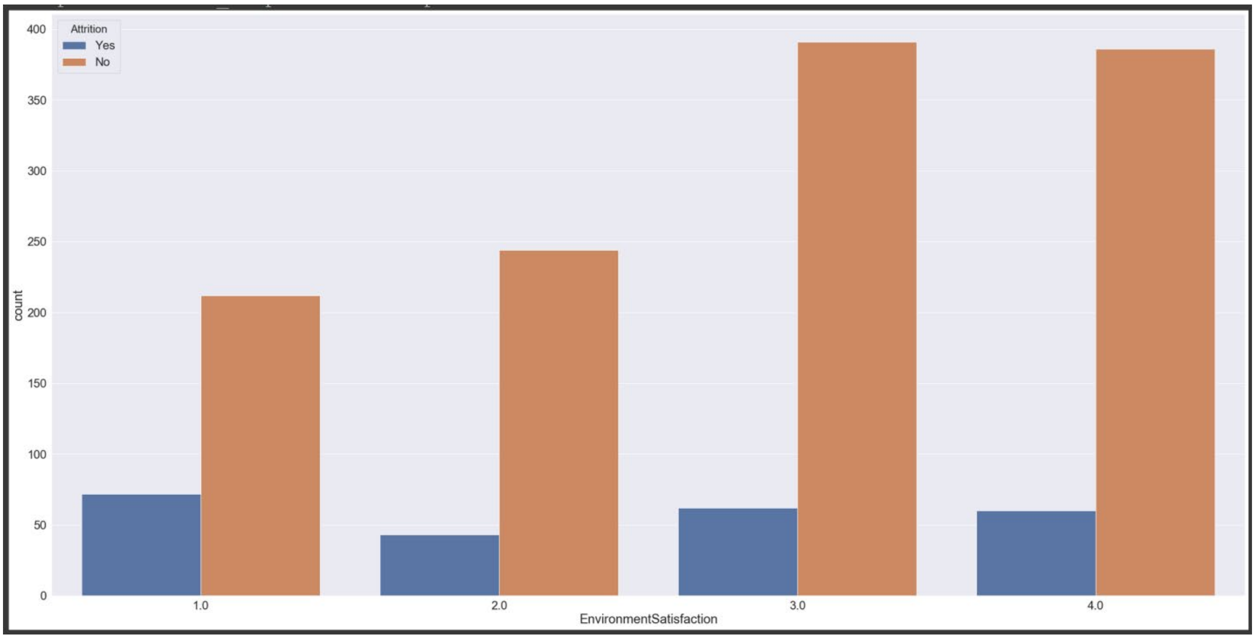
EDUCATION FIELD



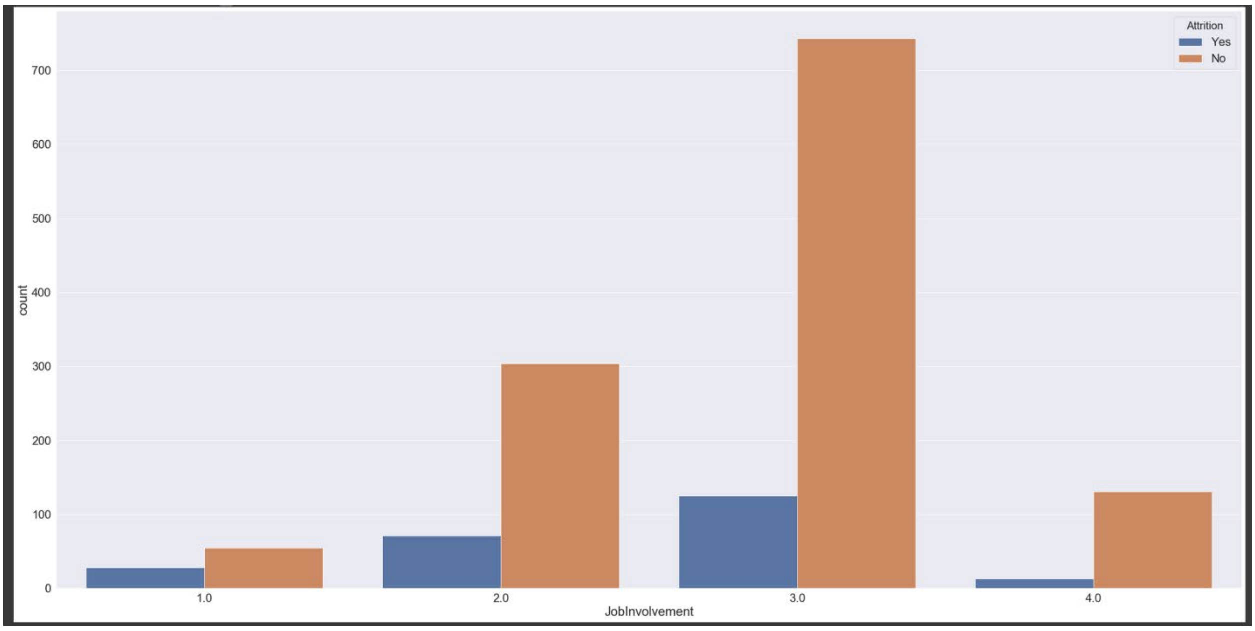
The education field affects IBM employee attrition rates. Employees that hold degree in life sciences and Medical related background show significant attrition rate

ENVIRONMENT SATISFACTION

No effect on Attrition (low to very high)



JOB INVOLVEMENT

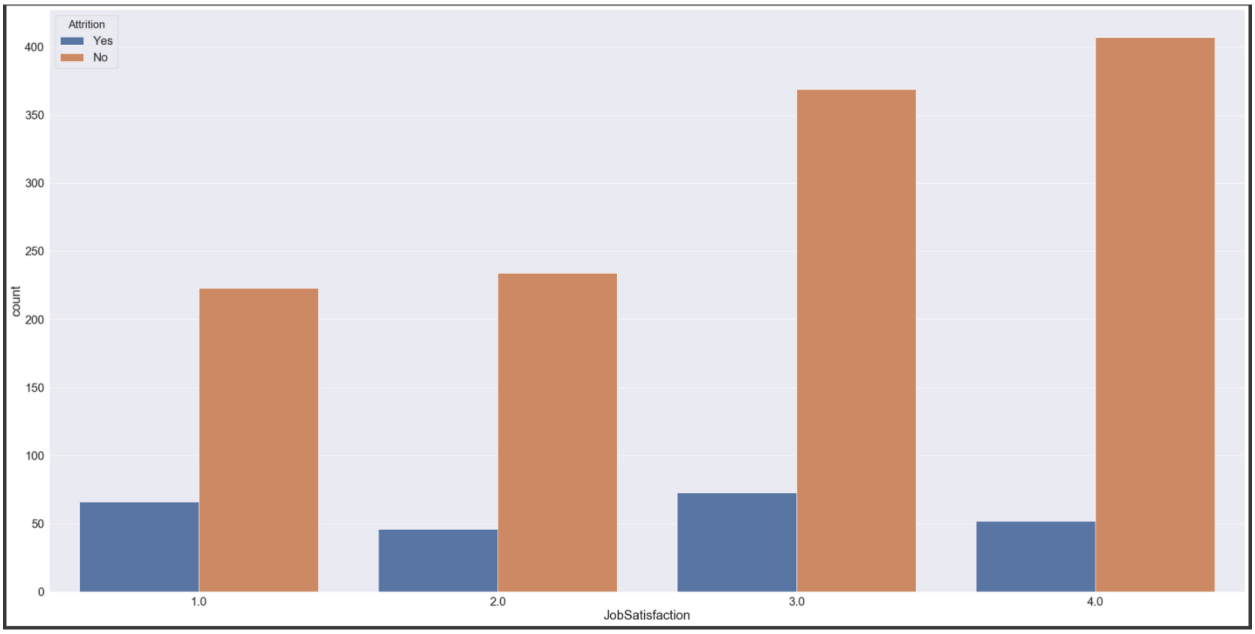


From the bar chart, we can see that JOB INVOLVEMENT has little to not effect on attriton.

JOB ROLE

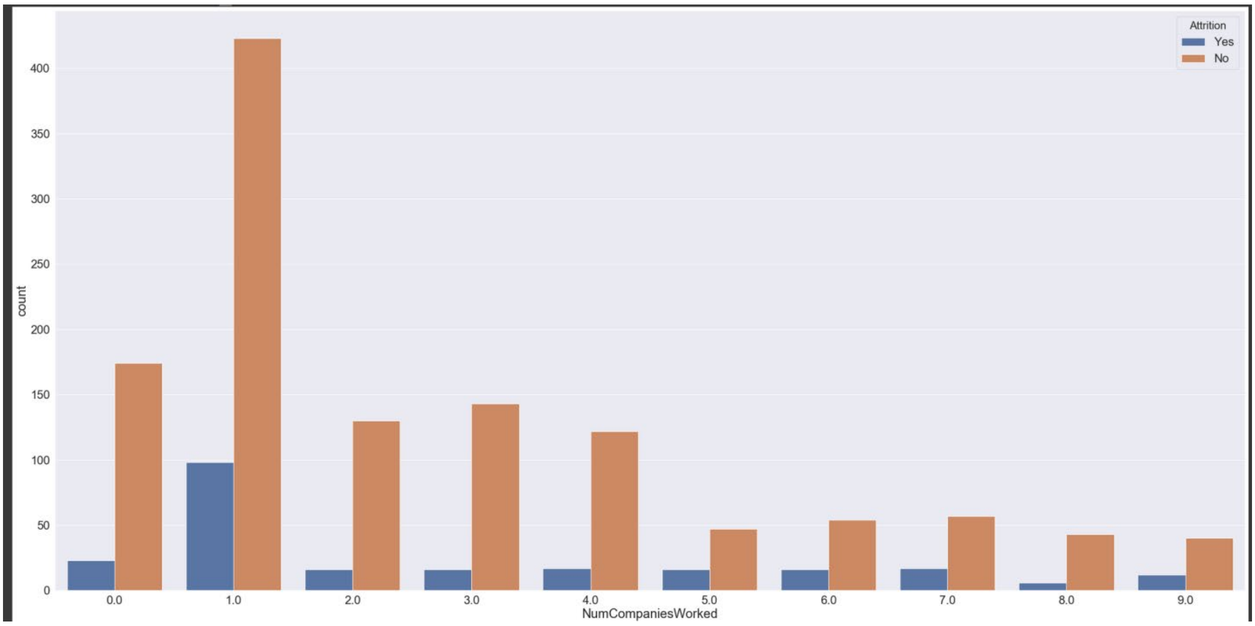
JOB STATISFACTION

No Impact



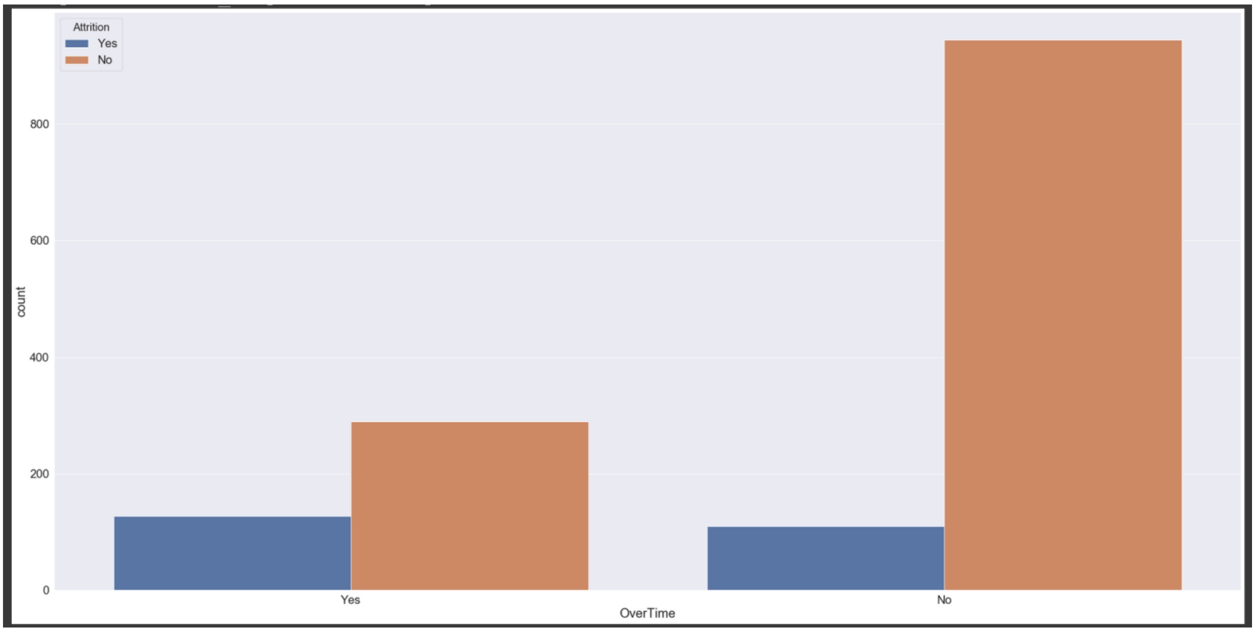
NUMBER OF COMPANIES WORKED

First timers leave faster than old time employees



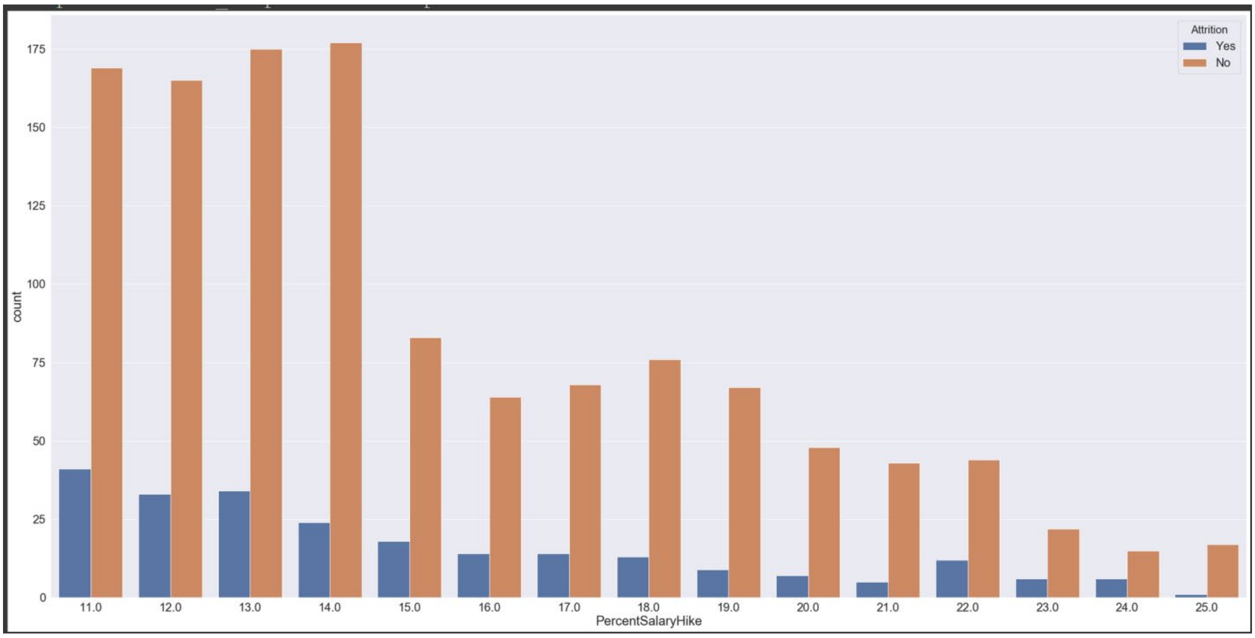
OVERTIME

Employees with no overtime payment leave more



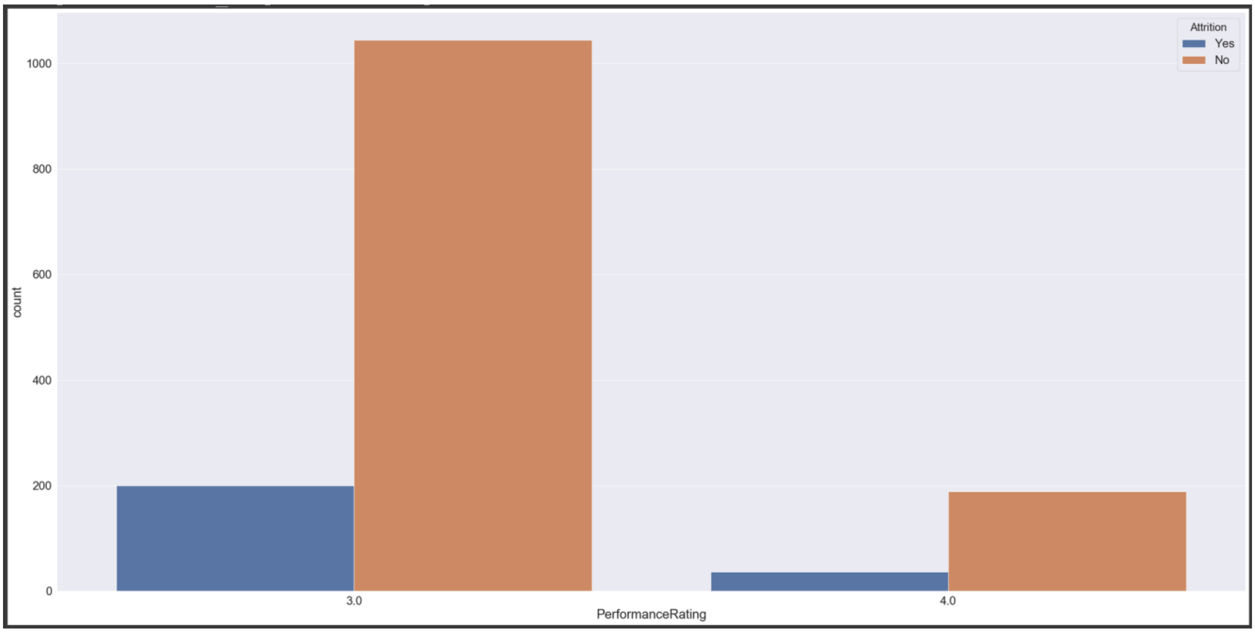
PERCENT SALARY HIKE

The lower the percent salary hike, the higher the attrition rate



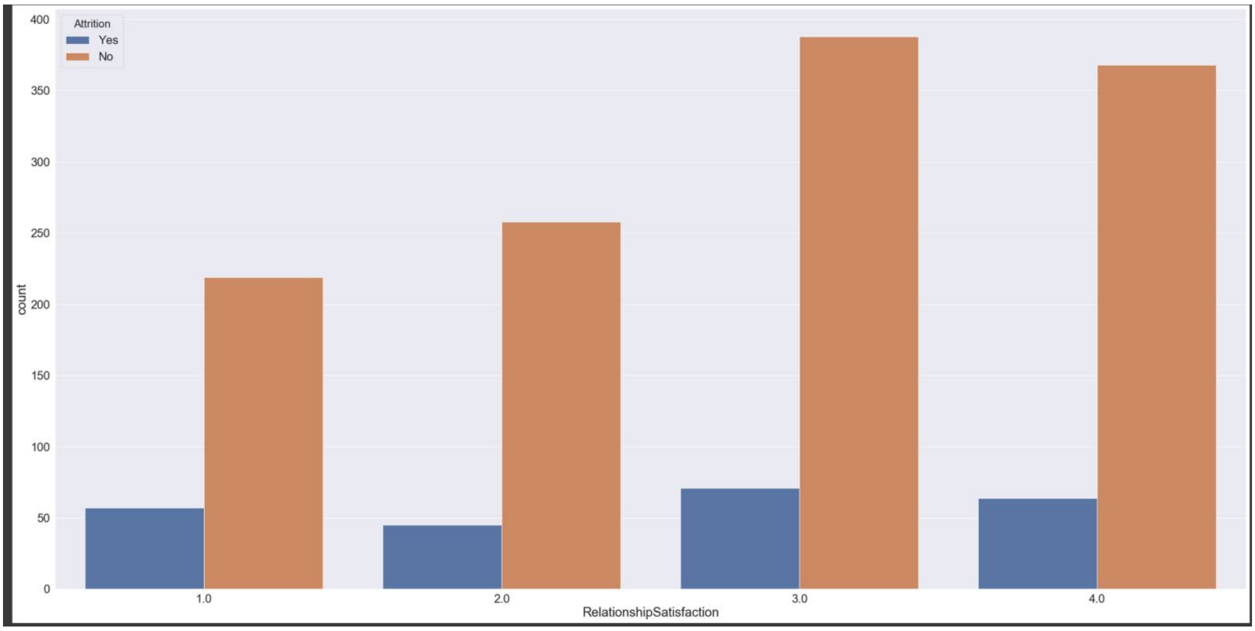
PERFORMANCE RATING

Staff with an average performance rating tend to leave more



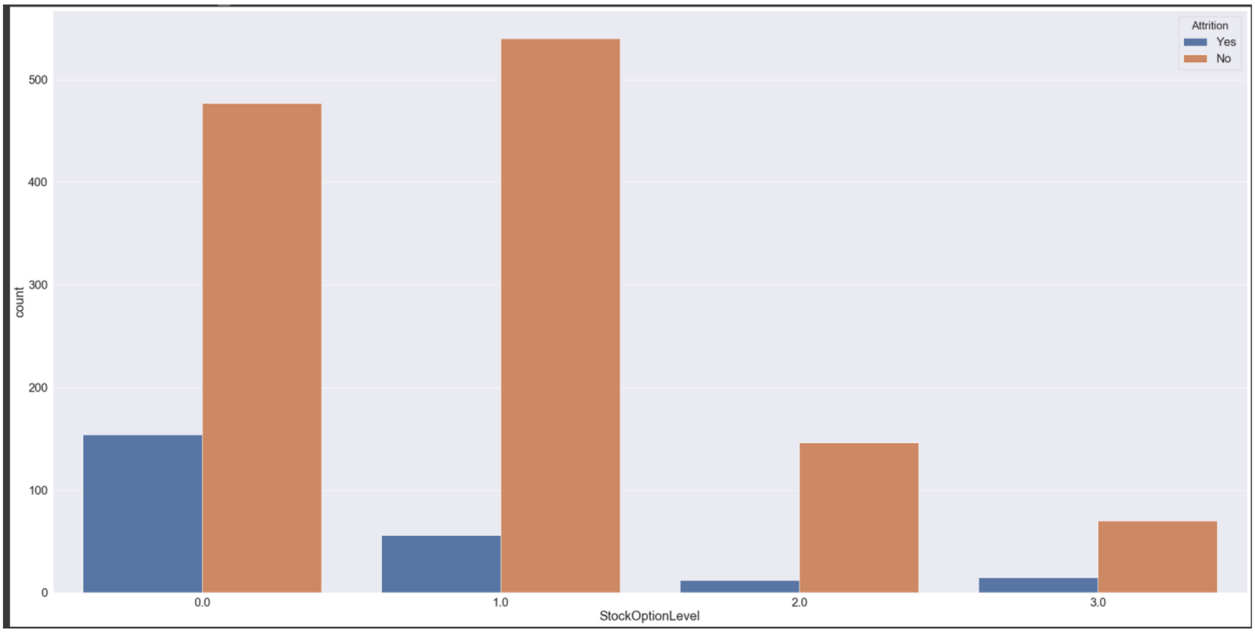
RELATIONSHIP SATISFACTION

No Impact (low to high)



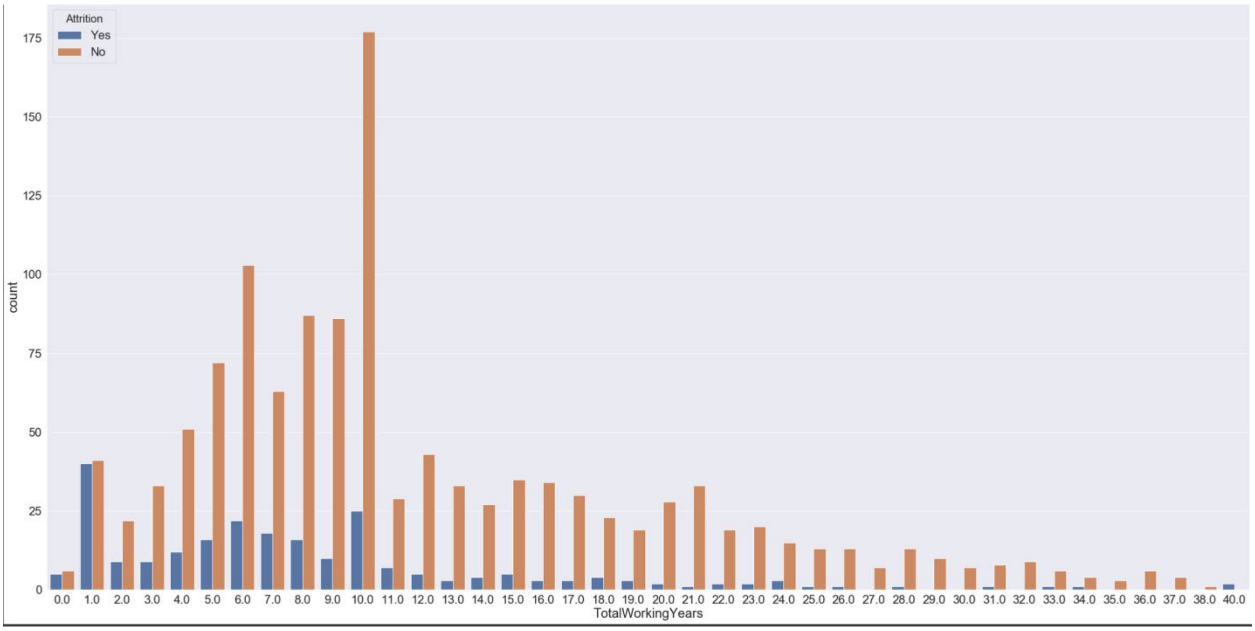
STOCK OPTION LEVEL

Employees with little or no stakes in the company have higher attrition rate



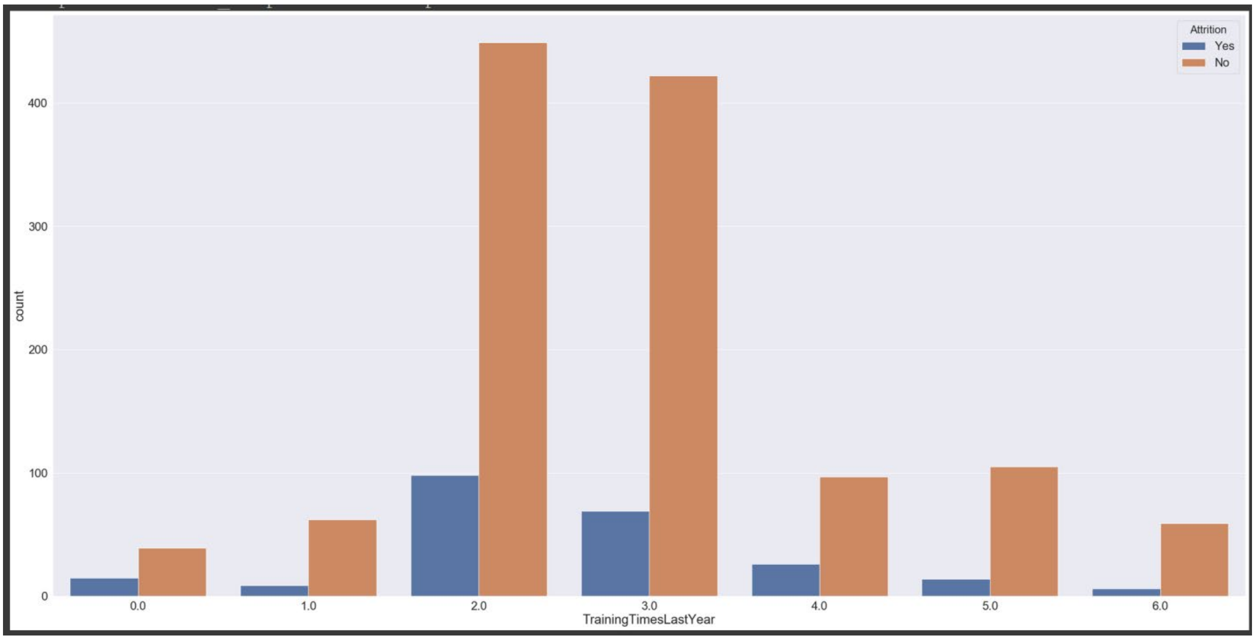
TOTAL WORKING YEARS

Employees that worked for 10years of their life tend to leave more



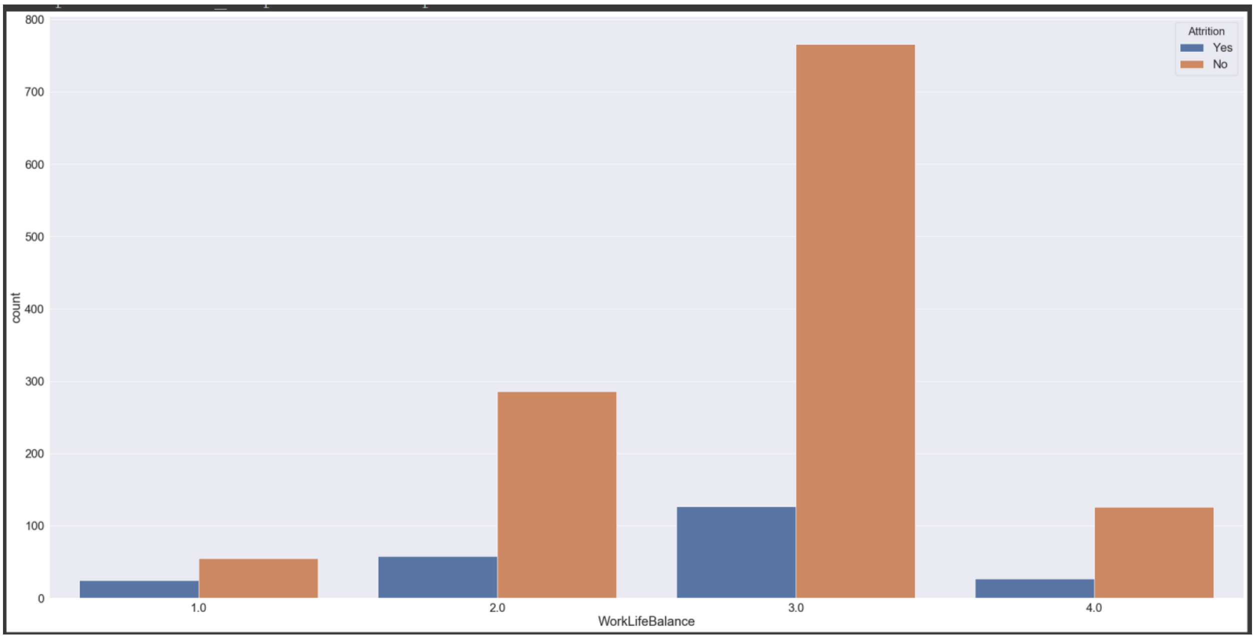
TRAINING TIMES SINCE LAST YEAR

Employees that have had 2/3 times training leave compared to those with more training



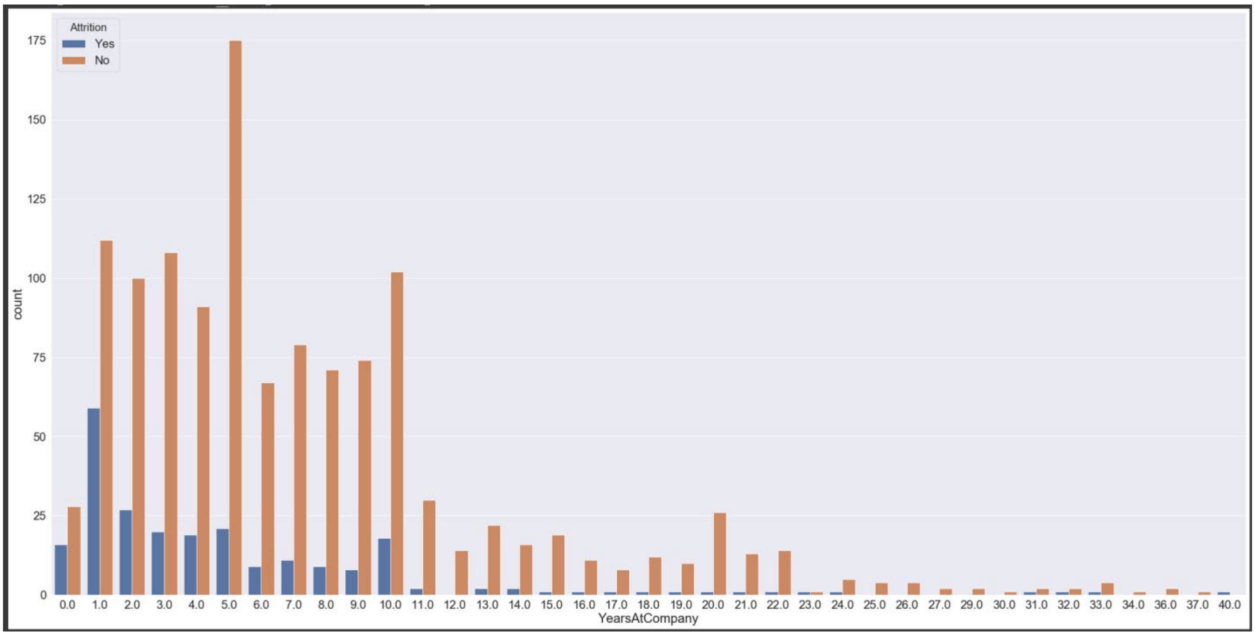
WORK LIFE BALANCE

No Impact



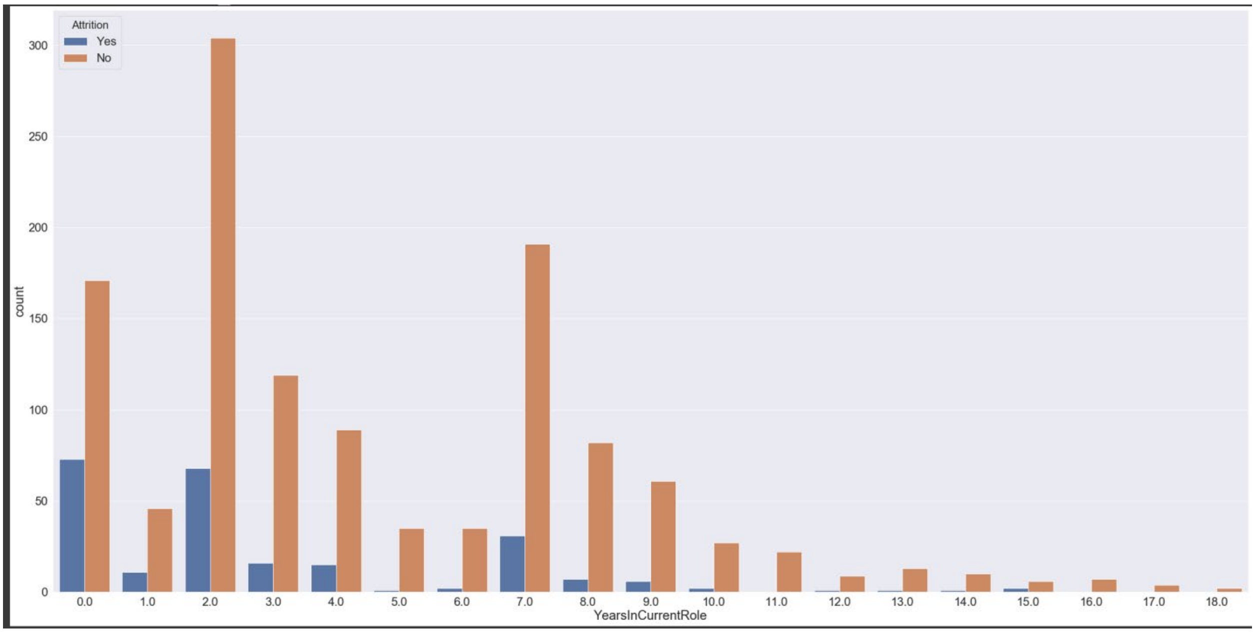
YEARS AT COMPANY

Employees that have spent 5 years in IBM have the highest attrition rate



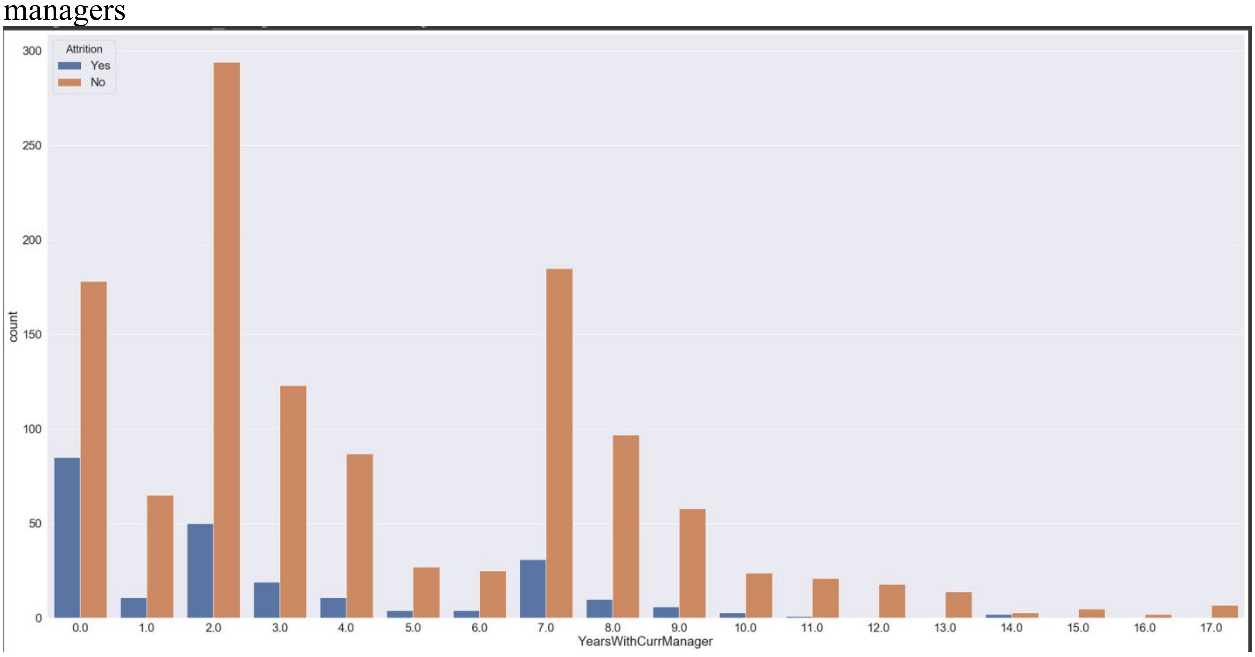
YEARS IN CURRENT ROLE

Employees with 2yrs in their current role leave more



YEARS WITH CURRENT MANAGER

Managers have an impact on Employee attrition. Most employees leave after 2yrs with their managers



Works Cited

Anderson, C. L., and Agarwal, R. 2010. “Practicing Safe Computing: A Multimethod Empirical Examination of Home Computer User Security Behavioral Intentions,” *MIS Quarterly* (34:3), pp. 613-643. .[1]

Automation, SAGE. *The Essential Guide to Six Sigma DMAIC: Phase 5 (of 5) - Control*, www.sageautomation.com/blog/the-essential-guide-to-six-sigma-dmaic-phase-5-of-5-control. [2]

Bolden-Barrett, Valerie. “Study: Turnover Costs Employers $15,000 per Worker.” *HR Dive*, 11 Aug. 2017, www.hrdive.com/news/study-turnover-costs-employers-15000-per-worker/449142/. [3]

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate data analysis* (Vol. 5, No. 3, pp. 207-219). Upper Saddle River, NJ: Prentice hall.

“Lack of Career Development Drives Employee Attrition.” *Smarter With Gartner*,www.gartner.com/smarterwithgartner/lack-of-career-development-drives-employee-attrition/.

Pavansubhash. “IBM HR Analytics Employee Attrition & Performance.” *Kaggle*, 31 Mar. 2017, www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset.[4]

Rohan’s Four - Rohan Jain, Ali Shahid. *IBM HR Analytics Employee Attrition & Performance*,inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/Januar y2018FBL/IBM_Attrition_VSS.html[4]

Terpening, William. *Business Process Design, Management, and Improvement*. Hercher, 2017, *vitalsource*, bookshelf.vitalsource.com/#/.[5]

Zaal, T.M.E, and Steve Newton. *Integrated Design and Engineering: as a Business Improvement Process*. Maj Engineering Publishing, 2014.[6]

Statistical Analysis Plan for Firm 3

<version 1>
<Sonkashi Sharma, Brooke Woods, UgochiMadumere>
<July 20, 2020>

Contents

Introduction

"Global Talent Monitor's report on workforce activity in 2018 shows that the lack of future career development remains a key driver of employee attrition — cited by 40% of departing employees as a dissatisfying factor in their job. At the same time, 28% of employees are actively seeking a job and 42% are passively open to new opportunities (Gartner 2018)". "Employee attrition costs large organizations millions of dollars each year and the loss of a particularly conscientious employee can be debilitating, not just to culture and morale, but to employee productivity(Gartner 2018)".This study intends to understand the factors which contribute to employee attrition in an organization. The study would contribute towards providing an insight for organizations which could eventually reduce their financial losses caused due to employee attrition.
IBM, also known as Business Process Manager on cloud, is known for producing and selling computer hardware, middleware and software, and providing hosts and consulting services to businesses. It is also well known for its research organization and its subscription services that provide clients a full lifecycle business process management (BPM) environment which includes the development, test, and production with tooling and run times for the process design, process execution, process monitoring,, and optimization. IBS offers the visualization and management of business processes, with low startup costs and a high return on investment for businesses. [1]
For this project, we will look specifically at the onboarding vs the recruiting process to map out the process for the employees of IBS. We will look at the univariate and multivariate analytics of employees and the industryingeneraltofindvariableswhichcontributetohighemployeeattritionrates. We will then use software like SAS Enterprise Miner, SAS Studio and SPSS to run a binary logistic regression on our model which will target the variables contributing to high attrition rates and therefore, the variables which influenceemployeebehavior.Wewillcomparetheactivitiesoftheemploymentprocessusingthemetrics of design, management, andimprovement.
Predictive analytics can be to better help organizations understand and design interventions that will be most effective in reducing unwanted attrition. Financial considerations aside, businesses are better off when they can retain good employees and the organizational knowledge they possess. Our findings will be oriented to helping organizations understand what is most important to their employees, with the goal of making improvements to increase employee engagement and productivity and reduce unwanted attrition. The comparison of engagement survey data to termination data can reveal areas of the employee experience in need of improvement. Ultimately, employee attrition can design an employee retention model that will work even if attrition is not expected to be a big issue soon. The model will still help the organization determine if it is experiencing the right kind of attrition. [4]

Data Source

For this project, we will analyze a dataset called "IMB HR Analytics Employee Attrition & Performance" which is a fictional dataset from IBM data scientists downloaded from the data source Kaggle. This dataset is composed of 1479 variables and 35 columns, containing variables such as age, travel frequency, daily rate, distance from home, education, employee count, department, etc. It is available for download at https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset/data#. [4]

Exit surveys are another potential data source that can provide richer information. Comparing responses on exit surveys to employees' engagement survey responses can reveal how the employees' perceptions changed over time. Correlating exit and engagement survey data can yield additional capability to predict attrition risk.

Analysis Objectives

- *Understandingthefactorswhichwillinfluencetheemployerandleadthemtoemployeeattrition*

- *Inthecomparison,lookingintoparameterslikemodelfit,Rsquaredcomparison, multicollinearity (VIF, tolerance), and P valuessignificance.*

Analysis Description

-*Logistic regression* must be used since the target is a binary variable. Logistic regression is used to predict the odds of being a case based on the values on the independent variables (predictor variables). Theoddsaredefinedastheprobabilitythataparticularoutcomeisacasedividedbytheprobabilitythatit is a noncase.

So, in our project we test robustness using the relative insensitivity of the statistical test to violations of the underlying internal assumptions:

- *Independence of the residuals* means that the residuals are not autocorrelated. The assumptionofindependencecanbetestedusingscatterplots:plotofresidualsversuscase number. The points should be symmetrically distributed around a horizontal line. The Durbin-Watson statistic can also be used to test for independence of the residuals. If the Durbin-Watson statistic is close to 2, there is no autocorrelation. Violation of this assumption can create bias in the significance tests and confidenceintervals. [6]

Remedy: One of the examples of violation occurs when we have time relevant or longitudinal data. The best way to analyze such data is General Estimation Equation (GEE).

- *Normality of the residuals* means that the residuals follow a normal distribution. Assumes 99% of values will fall within 3 standard deviations of the mean. The assumption of normality can be tested using plots: Q-Q plots; histograms. When plotting Q-Q plot of observed versus expected values, the points should be symmetrically distributed around a diagonal line with z values greater to or less than 3 being considered outliers. When plotting the histogram of residuals, the shape should be of a normal distribution. Observations will be arranged in increasing magnitude and plotted against normal expected distribution. If the test is non-statistically significant at 5%, there is no evidence of violation of the normality assumption. Violation of these assumptions can lead to bias in the estimation of coefficients and standard errors especially when the sample size is small. Nonnormality does not lead to severe problems in the interpretation whenthesamplesizeislarge.Oursamplesizeisconsideredlargesinceitis>30. [6]

Remedy: Data transformation, and particularly the Box-Cox power transformation, is one of the remedial actions that may help to make data normal.

A) *Collinearity of residuals* exists if there is an approximate linear relationship (i.e., shared variance) among some of IVs in thedata.

Problem with collinearity: Collinearities inflate the variances of the regression coefficients. This could have the following consequences: Parameter estimates that fluctuate dramatically with negligible changes in the sample; parameter estimates with signs that are "wrong" in terms of theoretical considerations; theoretically "important" variables with insignificant coefficients; the inability to determine the relative importance of collinear variables. [6]

What can multicollinearity do? It produces large SE of Bs which results in coefficients being non-significant; it produces bizarre β estimates (e.g., wrong direction); Removal or addition of one IV results in enormous change to the models.

When are multicollinearity problems certain? Tolerance ≤ .1 (or .2). [VIF ≥ 10 (or 5)]; even if tolerance ≥ .2 (or VIF ≤ 5), multicollinearity could be problematic when: There is a bivariate correlation of .7 or more between two IVs; the bivariate correlation between two IVs are greater than either of IV's correlation withDV. [6]

Remedies for multicollinearity: Omit the redundant IV [or aggregate the two similar ones] (Hair et al. 2010); Increase the sample size (Hair et al. 2010); Transform the raw-data X to create a new, orthogonal matrix (Mason and Perreault 1991); Mean center or Scale center raw X.

Statistical methodology/procedures

Beforetheimplementationofanyalgorithm,wemustsplitthedataintotrainingset(70%),validationset (20%) and testing set (10%) using SAS EM's partitionnode.

In the logit model the log odds of the outcome is modeled as a linear combination of predictor variables. We will name the node SAS EM node Logit. With Logistic regression, the training and predicting speed is relatively veryfast.

Using Multivariate analytics to detect the outliers

Some outliers could be detected from the Mahala Nobis distance which detects outliers of any kind. The Mahala Nobis distance should be divided by degrees of freedom (1 subtracted from the variables which would be entered as IV's).

For multivariate outliers, mahala Nobis distance is evaluated on a chi squared (x2) statistics with degrees of freedom=# of variables in the analysis. Testing the significance of the model can be done using chi square significant values with p values <.05 accepted and considered significant.[6]

All hypothesis testing will be done at 5% (95% confidence interval) with a 2-sided test with a significance level of <.05 unless otherwise specified for categorical parameters.

P values are rounded using 3 decimal places and values <.001 will be reported as <.001 Normality

assumption can be tested using Kolmogorov-Smirnov and Shapiro-Wilk. [6]

Handling of missing data/outliers

An initial cleaning of this dataset will be needed before an analysis can be performed. The dataset will be examined for missing data patterns. We will estimate missing values using a logistic approach and later use them during the main analysis. We will find univariate (extreme values of one variable) and multivariate (unusual combinations of 2 or more z scores) outliers. Outliers can be detected by visually inspecting the data with frequency distribution/histograms of the residuals and with standardization. Standardization is applied to transform the data to z scores. Z scores are tested using the normality assumption, and outliers are detected by values with Z scores falling greater than 3 or less than -3.

For outlier detection we are going to perform the following analysis:

Graph Examination & Outlier detections:
Outlier detection:

1) ***Apply a Standardized score to identify outliers for the continuousvariables.***

Applying a standardized score to identify outliers transforms the skewness of our model for the continuous variables falling within the range of -3 to 3. Such variables would not be considered asoutliers in this business process. We would also look at the Box plot for these variables also to check foroutliers. [6]

2) ***Apply box-whisker plot to identify the outliers for the continuousvariables***

We would analyze the values shown in the dataset that approach the threshold -3 and 3. Since the dataset is so small, it is possible that the values above and below 2 could be considered as outliers. A graphical method can be used to examine outliers in the boxplot. The lines extend to the outliers in the box-whiskers plot. [6]

3) ***Apply scatter plot matrix to identify the outliers for the continuousvariables.***

We will choose individual bivariate scatterplots to spot the outliers and choose the Z scores instead of using original scores. This could give a clearer picture of the observations. [6]

OUR FIRM'S ETHICALCODE

We pledge in writing to abide by the American Statistical Association's (ASA) and INFORMS' Codes of Ethics. Our adherence to these Codes signifies voluntary assumption of self-discipline. As the professional associations for our firm in the United States, the ASA and INFORMS requires adherence to their Codes of Ethics as a condition of membership. The standards of conduct set forth in these Codes provide basic principles in the ethical practice of data analysis consulting. The purpose of these Codes is to help us maintain our professionalism and adhere to high ethical standards in the conduct of providing services to clients and in our dealings withourcolleaguesandthepublic.Ourindividualjudgmentrequiresweapplytheseprinciples.Weareliableto disciplinaryactionundertheASA'sandINFORMS'RulesofProcedureforEnforcementofthisCodeifourconductis foundbytheASA'sorINFORMS'respectiveEthicsCommitteestobeinviolationoftheirrespectiveCodesortobring discredittotheprofessionortoASAandINFORMS.

Our Commitment to Our Clients

1) Wewillserveourclientswithintegrity,competence,independence,objectivity,andprofessionalism.
2) Wewillmutuallyestablishwithourclientsrealisticexpectationsofthebenefitsandresultsofourservices.
3) Wewillonlyacceptassignmentsforwhichwepossesstherequisiteexperienceandcompetencetoperformandwillonly assignstafforengagecolleagueswiththeknowledgeandexpertiseneededtoserveourclientseffectively.
4) Beforeacceptinganyengagement,wewillensurethatwehaveworkedwithourclientstoestablishamutualunderstanding oftheobjectives,scope,workplan,andfeearrangements.
5) Wewilltreatappropriatelyallconfidentialclientinformationthatisnotpublicknowledge,takereasonablestepsto preventitfromaccessbyunauthorizedpeople,andwillnottakeadvantageofproprietaryorprivileged information, either for use by ourselves, the client's firm, or another client, without the client's permission.
6) Wewillavoidconflictsofinterestortheappearanceofsuchandwillimmediatelydisclosetotheclientcircumstancesor intereststhatwebelievemayinfluencemyjudgmentorobjectivity.
7) Wewilloffertowithdrawfromaconsultingassignmentwhenwebelievemyobjectivityorintegritymaybeimpaired.
8) Wewillrefrainfrominvitinganemployeeofanactiveorinactiveclienttoconsideralternativeemploymentwithout prior discussion with theclient.

Our Commitment to Fiscal Integrity

9) Wewillagreeinadvancewithaclientonthebasisforfeesandexpensesandwillchargefeesthatarereasonableand commensuratewiththeservicesdeliveredandtheresponsibilityaccepted.
10) Wewillnotacceptcommissions,remuneration,orotherbenefitsfromathirdpartyinconnectionwiththe recommendationstoaclientwithoutthatclient'spriorknowledgeandconsent,andwilldiscloseinadvanceanyfinancial interestsingoodsorservicesthatformpartofsuchrecommendations.

Our Commitment to the Public and the Profession

11) Ifwithinthescopeofmyengagement,wewillreporttoappropriateauthoritieswithinorexternaltotheclientorganization anyoccurrencesofmalfeasance,dangerousbehavior,orillegalactivities.
12) Wewillrespecttherightsofconsultingcolleaguesandconsultingfirmsandwillnotusetheirproprietaryinformationor methodologies withoutpermission.
13) Wewillrepresenttheprofessionwithintegrityandprofessionalisminmyrelationswithourclients,colleagues,andthe generalpublic.
14) Wewillnotadvertiseourservicesinadeceptivemannernormisrepresentordenigrateindividualconsultingpractitioners, consultingfirms,ortheconsultingprofession.
15) IfweperceiveaviolationoftheCode,wewillreportittotheAPAandINFORMSandwillpromoteadherencetotheCode byothermemberconsultantsworkingonourbehalf.

SonakshiSharma

BrookeWoods

UgochiMadumere

Works Cited

*"*Gartner Survey Shows 29 Percent of Employees Witnessed At Least One Compliance Violation In The Last Two Years*." Gartner,* ww.gartner.com/en/newsroom/press-releases/2018-08-02-gartner-survey-shows-29-percent-of-employees-witnessed-at-least-one-compliance-violation-in-the-last-two-years.[1]

"Here's How IBM Predicts 95% of Its Turnover Using Data." *LinkedIn Talent Blog*, business.linkedin.com/talent-solutions/blog/artificial-intelligence/2019/IBM-predicts-95-\ percent-of-turnover-using-AI-and-data.[2]

"Lack of Career Development Drives Employee Attrition." *Smarter With Gartner*,www.gartner.com/smarterwithgartner/lack-of-career-development-drives-employee-attrition/. [3]

Pavansubhash. "IBM HR Analytics Employee Attrition & Performance." *Kaggle*, 31 Mar. 2017, www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset.[4]

Rohan's Four - Rohan Jain, Ali Shahid. *IBM HR Analytics Employee Attrition &Performance*,inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2018FBL/IBM_Attrition_VSS.html[5]

Zaal, T.M.E, and Steve Newton. *Integrated Design and Engineering: as a Business Improvement Process*. Maj Engineering Publishing, 2014.[6]