**HR Analytics – Statistical Learning**

**Problem Statement**

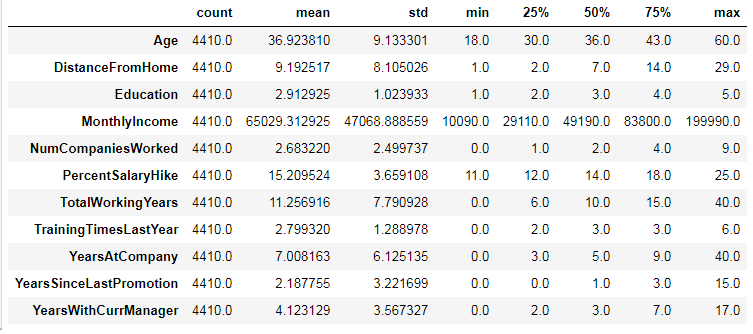
A large company named XYZ, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company, because of the following reasons -

The former employees’ projects get delayed, which makes it difficult to meet timelines, resulting in a reputation loss among consumers and partners A sizeable department has to be maintained, for the purposes of recruiting new talent More often than not, the new employees have to be trained for the job and/or given time to acclimatise themselves to the company Hence, the management has contracted an HR analytics firm to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay. Also, they want to know which of these variables is most important and needs to be addressed right away.

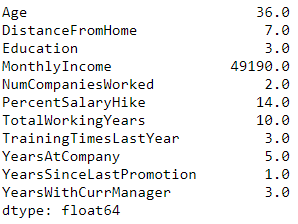
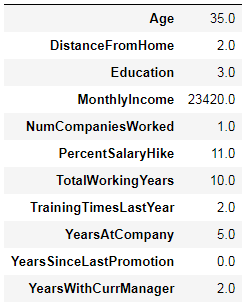
Since you are one of the star analysts at the firm, this project has been given to you.

Goal of the case study is You are required to model the probability of attrition. The results thus obtained will be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay.

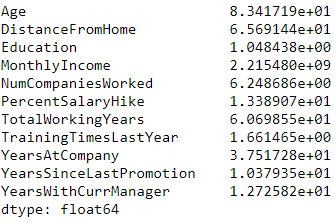
### **Univariate Analysis:**



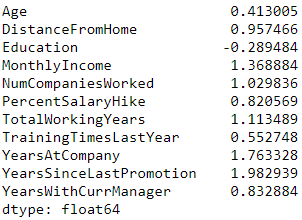
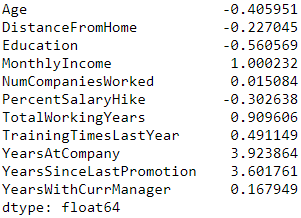
**Median: Mode:**

**Variance:**



**Skewness: Kurtosis:**

### **From above Description we have to see the continuous features that can be causes to Attrition** i.e 'Age','DistanceFromHome','MonthlyIncome', 'PercentSalaryHike', 'TotalWorkingYears', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager'

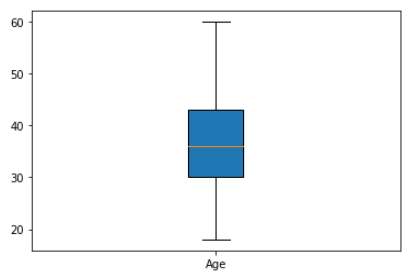
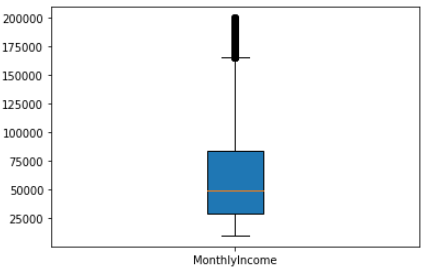
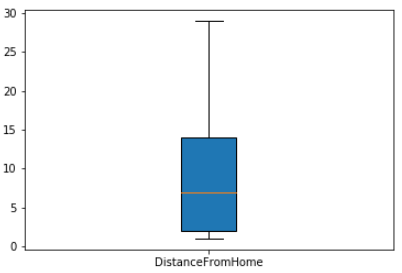
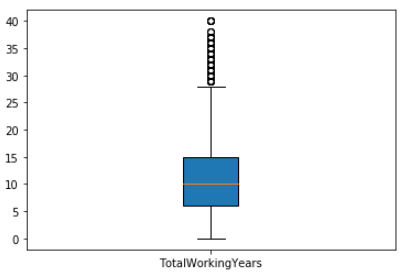
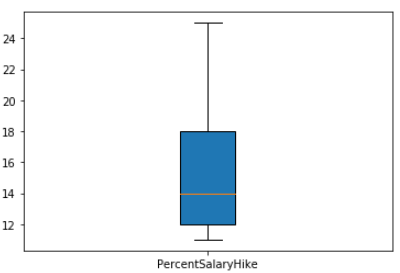
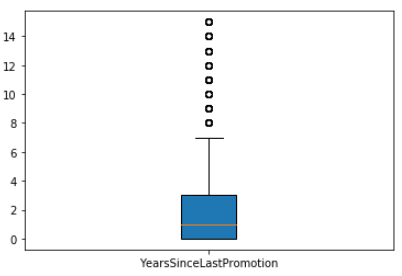
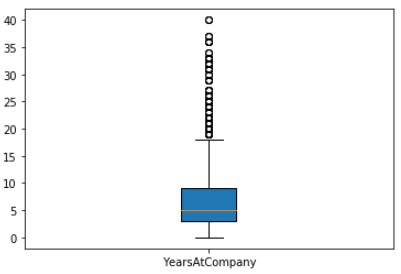
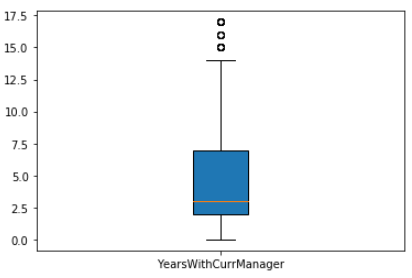
### **From these features all of them showing positive Skewness**

### **Age, DistanceFromHome, PercentSalaryHike are leptokurtic and all other are platykurtic.**

## Attrition rate %:

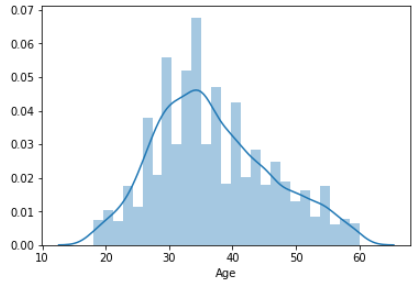
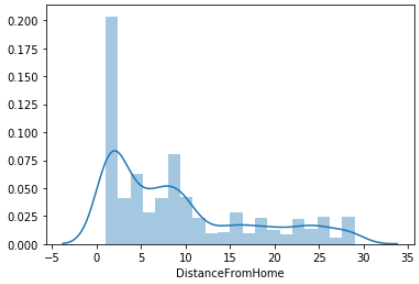
## 

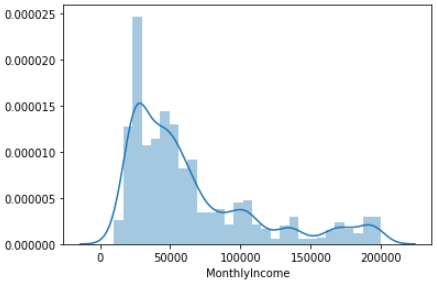
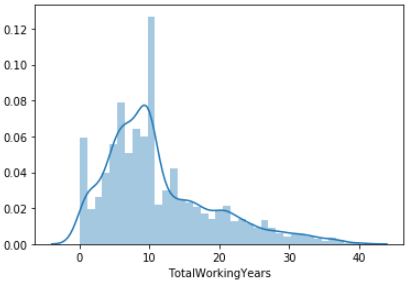
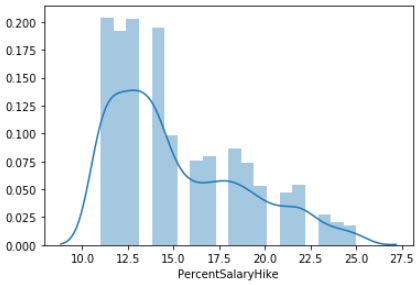
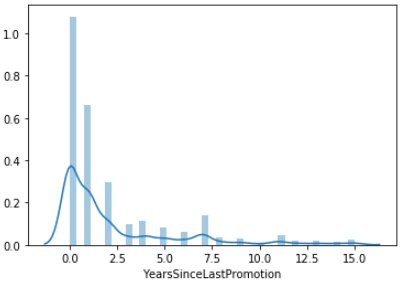
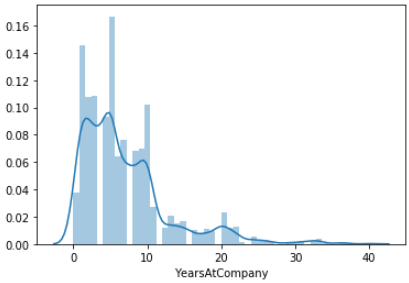
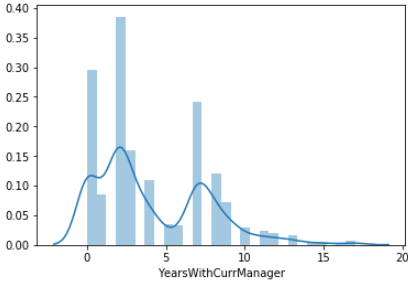
#### **Let's check outliers through box plot:**

    **From above boxplots we can see that Age, DistanceFromHome, PercentSalaryHike don't have Outliers others have outliers.**

**IQR of Age is 30 to 43 years (13 years)**

**Checking Distribution of Data:**

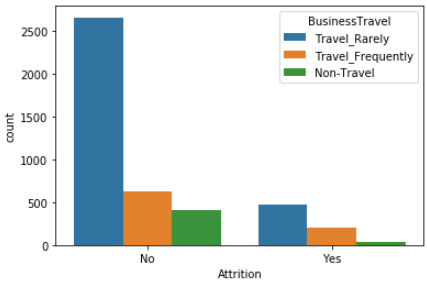
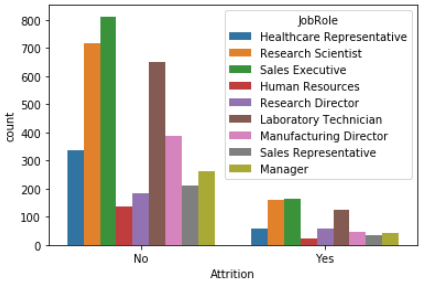
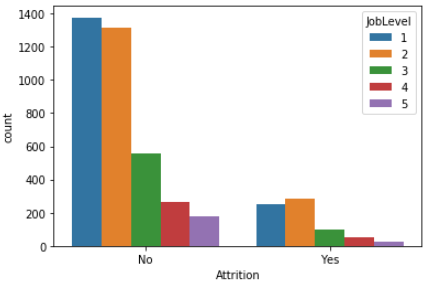
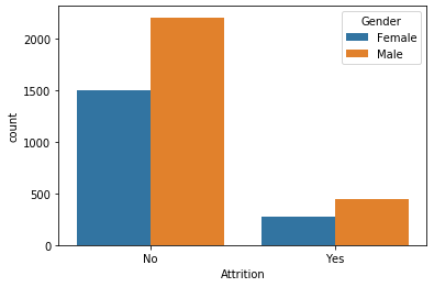
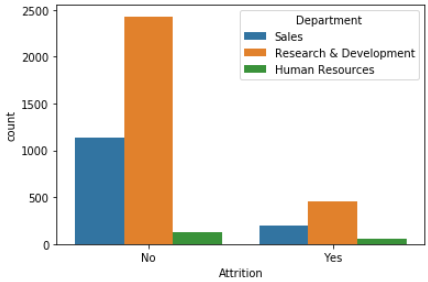
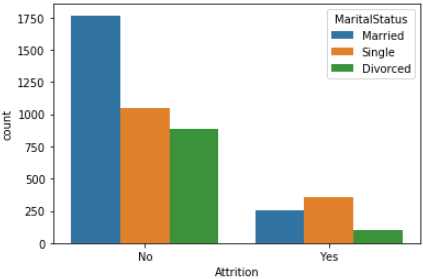
 

### **From above histograms we can see Distribution & Skewness of individual variables**

## Distribution with Categorical Features:

#### **Categorical features which can have dependency on Attrition are**: BusinessTravel, Department, Gender, JobLevel, JobRole, MaritalStatus

**From above plots we can make some conclusions:**

#### 1. The employee who Rarely travels have high attrition rate

#### 2. Research & Development Department has high Attrition rate

#### 3. Male Employees have high Attrition rate as compared to Female Employees

#### 4. Job level 1 & 2 has high Attrition rate

#### 5. Research & Development, Sales Executives & Laboratory Technician has high Attrition rate

#### 6. Single employees has high Attrition rate than married and Divorced Employees

## Correlations of all independent features with dependent feature Attrition:

## 

**Correlation & P – Value Check:**

Null Hypothesis: There is no Significant Correlation between Attrition and Age

Alternate Hypothesis: There is Significant Correlation between Attrition and Age

Correlation: -0.15920500686577965 P Value: 1.996801615886744e-26

P-Value < 0.05 hence Null Hypothesis is rejected, Accepting Alternate Hypothesis

There is Negative Correlation between Attrition and Age

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and BusinessTravel

Alternate Hypothesis: There is Significant Correlation between Attrition and BusinessTravel

Correlation: 7.377694602225034e-05 P Value: 0.9960919945437704

P-Value >= 0.05 hence Null hypothesis is Accepted

There is no Significant Correlation between Attrition and BusinessTravel

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and DistanceFromHome

Alternate Hypothesis: There is Significant Correlation between Attrition and DistanceFromHome

Correlation: -0.009730141010179674 P Value: 0.5182860428050771

P-Value >= 0.05 hence Null hypothesis is Accepted

There is no Significant Correlation between Attrition and DistanceFromHome

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and Education

Alternate Hypothesis: There is Significant Correlation between Attrition and Education

Correlation: -0.015111167710968713 P Value: 0.3157293177118575

P-Value >= 0.05 hence Null hypothesis is Accepted

There is no Significant Correlation between Attrition and Education

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and EducationField

Alternate Hypothesis: There is Significant Correlation between Attrition and EducationField

Correlation: -0.05794031241568039 P Value: 0.00011819790920714883

P-Value < 0.05 hence Null Hypothesis is rejected, Accepting Alternate Hypothesis

There is Negative Correlation between Attrition and EducationField

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and Gender

Alternate Hypothesis: There is Significant Correlation between Attrition and Gender

Correlation: 0.018125078877010238 P Value: 0.22881970951795952

P-Value >= 0.05 hence Null hypothesis is Accepted

There is no Significant Correlation between Attrition and Gender

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and JobLevel

Alternate Hypothesis: There is Significant Correlation between Attrition and JobLevel

Correlation: -0.010289713287495042 P Value: 0.49451717271828405

P-Value >= 0.05 hence Null hypothesis is Accepted

There is no Significant Correlation between Attrition and JobLevel

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and JobRole

Alternate Hypothesis: There is Significant Correlation between Attrition and JobRole

Correlation: 0.025808853490975857 P Value: 0.08658208267560671

P-Value >= 0.05 hence Null hypothesis is Accepted

There is no Significant Correlation between Attrition and JobRole

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and MaritalStatus

Alternate Hypothesis: There is Significant Correlation between Attrition and MaritalStatus

Correlation: 0.162070234657015 P Value: 2.4449148399198393e-27

P-Value < 0.05 hence Null Hypothesis is rejected, Accepting Alternate Hypothesis

There is Positive Correlation between Attrition and MaritalStatus

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and MonthlyIncome

Alternate Hypothesis: There is Significant Correlation between Attrition and MonthlyIncome

Correlation: -0.031176281698115007 P Value: 0.03842748490600132

P-Value < 0.05 hence Null Hypothesis is rejected, Accepting Alternate Hypothesis

There is Negative Correlation between Attrition and MonthlyIncome

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and NumCompaniesWorked

Alternate Hypothesis: There is Significant Correlation between Attrition and NumCompaniesWorked

Correlation: 0.04150332971190174 P Value: 0.00584142424757884

P-Value < 0.05 hence Null Hypothesis is rejected, Accepting Alternate Hypothesis

There is Positive Correlation between Attrition and NumCompaniesWorked

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and PercentSalaryHike

Alternate Hypothesis: There is Significant Correlation between Attrition and PercentSalaryHike

Correlation: 0.03253259489105349 P Value: 0.030743386433355353

P-Value < 0.05 hence Null Hypothesis is rejected, Accepting Alternate Hypothesis

There is Positive Correlation between Attrition and PercentSalaryHike

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and StockOptionLevel

Alternate Hypothesis: There is Significant Correlation between Attrition and StockOptionLevel

Correlation: -0.006838852403261513 P Value: 0.6498072937475723

P-Value >= 0.05 hence Null hypothesis is Accepted

There is no Significant Correlation between Attrition and StockOptionLevel

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and TotalWorkingYears

Alternate Hypothesis: There is Significant Correlation between Attrition and TotalWorkingYears

Correlation: -0.17023794049182428 P Value: 4.959687383334802e-30

P-Value < 0.05 hence Null Hypothesis is rejected, Accepting Alternate Hypothesis

There is Negative Correlation between Attrition and TotalWorkingYears

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and TrainingTimesLastYear

Alternate Hypothesis: There is Significant Correlation between Attrition and TrainingTimesLastYear

Correlation: -0.049430576244255 P Value: 0.0010247061915365072

P-Value < 0.05 hence Null Hypothesis is rejected, Accepting Alternate Hypothesis

There is Negative Correlation between Attrition and TrainingTimesLastYear

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and YearsAtCompany

Alternate Hypothesis: There is Significant Correlation between Attrition and YearsAtCompany

Correlation: -0.1343922139899772 P Value: 3.1638831224877484e-19

P-Value < 0.05 hence Null Hypothesis is rejected, Accepting Alternate Hypothesis

There is Negative Correlation between Attrition and YearsAtCompany

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and YearsSinceLastPromotion

Alternate Hypothesis: There is Significant Correlation between Attrition and YearsSinceLastPromotion

Correlation: -0.03301877514258434 P Value: 0.028330336189396753

P-Value < 0.05 hence Null Hypothesis is rejected, Accepting Alternate Hypothesis

There is Negative Correlation between Attrition and YearsSinceLastPromotion

-----------------------------------------------------------------------------------------------------------------

Null Hypothesis: There is no Significant Correlation between Attrition and YearsWithCurrManager

Alternate Hypothesis: There is Significant Correlation between Attrition and YearsWithCurrManager

Correlation: -0.15619931590162847 P Value: 1.7339322652896276e-25

P-Value < 0.05 hence Null Hypothesis is rejected, Accepting Alternate Hypothesis

There is Negative Correlation between Attrition and YearsWithCurrManager

-----------------------------------------------------------------------------------------------------------------

### **From above calculations we can have some conclusion based on correlation & P value:**

#### **Features don't have relationship with Attrition** : BusinessTravel, DistanceFromHome, Education, Gender, JobLevel, JobRole, StockOptionLevel

#### **Features have relationship with Attrition and have Positive Correlation:** MaritalStatus, NumCompaniesWorked, PercentSalaryHike

#### **Features have relationship with Attrition and have Negative Correlation:** Age, EducationField, MonthlyIncome, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsSinceLastPromotion, YearsWithCurrManager

# Statistical tests:

## As we seen in distplots none of features Normally Distributed so we can perform only Non-Parametric tests.

## Dependent variable is Attrition and that is Categorical so in Non-Parametric we can perform below Tests:

### **1. Mann-Whitney Test -** 1 Dependent categorical variable and other continuous varibles

### **2. CHI Square test -** Only for Categorical Variables

## 1. Mann-Whitney Test:

#### For Mann-Whitney Test we need to separate data as Attrition Yes & Attrition No

##### **Continuous Variables we should check with Attrition as per above correlation results :** Age, MonthlyIncome, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsSinceLastPromotion, YearsWithCurrManager, NumCompaniesWorked, PercentSalaryHike

H0 = There is no significant difference between Attrition\_yes with Age and Attrition\_No with Age

H1 = There is significant difference between Attrition\_yes with Age and Attrition\_No with Age

961731.0 P Value: 2.9951588479067175e-30

P-Value < 0.05 hence H0 rejected, Accepting H1 Hypothesis

-----------------------------------------------------------------------------------------------------------------

H0 = There is no significant difference between Attrition\_yes with MonthlyIncome and Attrition\_No with MonthlyIncome

H1 = There is significant difference between Attrition\_yes with MonthlyIncome and Attrition\_No with MonthlyIncome

1264900.5 P Value: 0.053577283839938566

P-Value >= 0.05 hence H0 Accepted

-----------------------------------------------------------------------------------------------------------------

H0 = There is no significant difference between Attrition\_yes with NumCompaniesWorked and Attrition\_No with NumCompaniesWorked

H1 = There is significant difference between Attrition\_yes with NumCompaniesWorked and Attrition\_No with NumCompaniesWorked

1259144.0 P Value: 0.03266173775282211

P-Value < 0.05 hence H0 rejected, Accepting H1 Hypothesis

-----------------------------------------------------------------------------------------------------------------

H0 = There is no significant difference between Attrition\_yes with PercentSalaryHike and Attrition\_No with PercentSalaryHike

H1 = There is significant difference between Attrition\_yes with PercentSalaryHike and Attrition\_No with PercentSalaryHike

1250640.0 P Value: 0.018660129917539733

P-Value < 0.05 hence H0 rejected, Accepting H1 Hypothesis

-----------------------------------------------------------------------------------------------------------------

H0 = There is no significant difference between Attrition\_yes with TotalWorkingYears and Attrition\_No with TotalWorkingYears

H1 = There is significant difference between Attrition\_yes with TotalWorkingYears and Attrition\_No with TotalWorkingYears

907502.5 P Value: 1.0203529765342384e-39

P-Value < 0.05 hence H0 rejected, Accepting H1 Hypothesis

-----------------------------------------------------------------------------------------------------------------

H0 = There is no significant difference between Attrition\_yes with TrainingTimesLastYear and Attrition\_No with TrainingTimesLastYear

H1 = There is significant difference between Attrition\_yes with TrainingTimesLastYear and Attrition\_No with TrainingTimesLastYear

1238940.0 P Value: 0.005167954938699059

P-Value < 0.05 hence H0 rejected, Accepting H1 Hypothesis

-----------------------------------------------------------------------------------------------------------------

H0 = There is no significant difference between Attrition\_yes with YearsAtCompany and Attrition\_No with YearsAtCompany

H1 = There is significant difference between Attrition\_yes with YearsAtCompany and Attrition\_No with YearsAtCompany

923238.0 P Value: 6.047598261692858e-37

P-Value < 0.05 hence H0 rejected, Accepting H1 Hypothesis

-----------------------------------------------------------------------------------------------------------------

H0 = There is no significant difference between Attrition\_yes with YearsSinceLastPromotion and Attrition\_No with YearsSinceLastPromotion

H1 = There is significant difference between Attrition\_yes with YearsSinceLastPromotion and Attrition\_No with YearsSinceLastPromotion

1209366.0 P Value: 0.0002021180346719736

P-Value < 0.05 hence H0 rejected, Accepting H1 Hypothesis

-----------------------------------------------------------------------------------------------------------------

H0 = There is no significant difference between Attrition\_yes with YearsWithCurrManager and Attrition\_No with YearsWithCurrManager

H1 = There is significant difference between Attrition\_yes with YearsWithCurrManager and Attrition\_No with YearsWithCurrManager

957253.5 P Value: 1.2365483142169853e-31

P-Value < 0.05 hence H0 rejected, Accepting H1 Hypothesis

-----------------------------------------------------------------------------------------------------------------

### **From above Mann-Whitney Tests we can have conclusion:**

#### There is no significant difference between Attrition Yes Monthly Income & Attrition No Monthly Income

#### **There is significant difference between Attrition Yes & Attrition No with following Variables:** Age, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsSinceLastPromotion, YearsWithCurrManager, NumCompaniesWorked, PercentSalaryHike

## 2. CHI Square Test:

#### **For CHI Square test we are checking dependency of Categorical Variables with Attrition**

#### **Categorical Variables we should check based on correlation results**: BusinessTravel , EducationField , Gender , Department , JobRole , MaritalStatus , JobLevel , StockOptionLevel

Ho= There is no dependency betweem Attrition and BusinessTravel

H1= There is dependency betweem Attrition and BusinessTravel

BusinessTravel 0 1 2

Attrition

0 414 624 2661

1 36 207 468

72.54724105696552 P Value: 1.764276972983189e-16

P-Value < 0.05 hence H0 rejected, Accepting H1 Hypothesis

-----------------------------------------------------------------------------------------------------------------

Ho= There is no dependency betweem Attrition and EducationField

H1= There is dependency betweem Attrition and EducationField

EducationField 0 1 2 3 4 5

Attrition

0 48 1515 402 1167 216 351

1 33 303 75 225 30 45

46.194921001730584 P Value: 8.288917469574179e-09

P-Value < 0.05 hence H0 rejected, Accepting H1 Hypothesis

-----------------------------------------------------------------------------------------------------------------

Ho= There is no dependency betweem Attrition and Gender

H1= There is dependency betweem Attrition and Gender

Gender 0 1

Attrition

0 1494 2205

1 270 441

1.349904410246582 P Value: 0.24529482862926827

P-Value >= 0.05 hence H0 Accepted

-----------------------------------------------------------------------------------------------------------------

Ho= There is no dependency betweem Attrition and Department

H1= There is dependency betweem Attrition and Department

Department 0 1 2

Attrition

0 132 2430 1137

1 57 453 201

29.090274924488266 P Value: 4.820888218170406e-07

P-Value < 0.05 hence H0 rejected, Accepting H1 Hypothesis

-----------------------------------------------------------------------------------------------------------------

Ho= There is no dependency betweem Attrition and JobRole

H1= There is dependency betweem Attrition and JobRole

JobRole 0 1 2 3 4 5 6 7 8

Attrition

0 336 135 651 264 387 183 717 813 213

1 57 21 126 42 48 57 159 165 36

25.116313674604072 P Value: 0.001485544744815264

P-Value < 0.05 hence H0 rejected, Accepting H1 Hypothesis

-----------------------------------------------------------------------------------------------------------------

Ho= There is no dependency betweem Attrition and MaritalStatus

H1= There is dependency betweem Attrition and MaritalStatus

MaritalStatus 0 1 2

Attrition

0 882 1767 1050

1 99 252 360

138.49102962254608 P Value: 8.45385940605786e-31

P-Value < 0.05 hence H0 rejected, Accepting H1 Hypothesis

-----------------------------------------------------------------------------------------------------------------

Ho= There is no dependency betweem Attrition and JobLevel

H1= There is dependency betweem Attrition and JobLevel

JobLevel 1 2 3 4 5

Attrition

0 1377 1317 558 267 180

1 252 285 96 51 27

6.2691759264759925 P Value: 0.1799276801337184

P-Value >= 0.05 hence H0 Accepted

-----------------------------------------------------------------------------------------------------------------

Ho= There is no dependency betweem Attrition and StockOptionLevel

H1= There is dependency betweem Attrition and StockOptionLevel

StockOptionLevel 0 1 2 3

Attrition

0 1575 1518 390 216

1 318 270 84 39

3.046265305068262 P Value: 0.38454683657380506

P-Value >= 0.05 hence H0 Accepted

-----------------------------------------------------------------------------------------------------------------

### **From above performed CHI Square tests we can have following conclusions:**

#### **Variables have dependency with Attrition are:** BusinessTravel , EducationField, Department , JobRole , MaritalStatus

#### **Variables don't have dependency with Attrition are:** Gender, JobLevel , StockOptionLevel