Modeling the Impact of Workplace Factors on Employee Performance

STAT 31631 – Statistical Modelling

Department of Statistics & Computer Science

University of Kelaniya

Academic Year - 2023/2024

By

Group 03

- PS/2021/236 - T.N.N.Hendawitharana
- PS/2021/145 - P.G.N.I.Ananda
- PS/2021/056 - D.C.A.S.Madhusari
- PS/2021/227 - L.G.R.Hasaranga
- PS/2021/002 - I.D.S.K.Jayarathna
- PS/2021/095 - K.N.Chandraraja
- PS/2021/057 - W.T.N.Jayasekara
- PS/2021/244 - S.M.D.T.Bandara
- PS/2021/058 - P.D.N.Thathsarani
- PS/2021/158 - P.K.R.I. Buddhima
- PS/2021/144 - H.M.C.P.Bandara

Abstract

In today’s dynamic business landscape, organizations face persistent challenges in optimizing employee performance and job satisfaction, both of which are critical to productivity and long-term success. Despite the availability of extensive employee data, actionable insights into the drivers of performance remain elusive for many companies. This study addresses this gap by employing advanced statistical modeling and machine learning techniques in RStudio to systematically analyze the relationship between workplace factors—such as job satisfaction, compensation, workload, and leadership quality—and employee performance, using real-world HR datasets.

The research identifies and quantifies key demographic and workplace variables influencing performance, develops predictive models for forecasting employee outcomes, and categorizes employees based on absenteeism and productivity patterns. Notably, the study introduces a novel, regression-based framework that models complex interactions among 22 variables, with department-specific analyses revealing nuanced differences across organizational functions.

Key findings demonstrate that data-driven approaches can significantly enhance HR decision-making, reduce operational costs, and foster employee engagement by enabling targeted, evidence-based interventions. The study’s adaptable methodology offers broad applicability across industries and organizational sizes, providing a scalable solution for workforce optimization. These insights empower organizations to move beyond intuition, leveraging robust analytics to drive sustainable improvements in employee performance and organizational effectiveness.

Introduction

- **Background of the study**

In today's competitive business environment, organizations are increasingly recognizing that their workforce is their most valuable asset. Employee performance and job satisfaction directly impact productivity, innovation, and overall business success. However, many companies struggle with high turnover rates, inconsistent performance, and employee dissatisfaction—issues that lead to significant financial and operational costs.

Recent advances in **people analytics** and **data science** have made it possible to analyze workplace dynamics systematically. By leveraging **statistical modeling and machine learning in RStudio**, businesses can now uncover hidden patterns in employee behavior, predict critical outcomes like attrition and performance, and make data-driven decisions to optimize their workforce strategies.

This study focuses on **modeling the relationship between workplace factors (e.g., job satisfaction, compensation, workload, leadership quality) and employee performance** using real-world HR datasets. The analysis will be conducted in **RStudio**, a powerful open-source tool for statistical computing, ensuring reproducibility and scalability.

- **Problem Statement**

Despite collecting vast amounts of employee data, many organizations lack actionable insights into how workplace factors influence individual performance. Decisions regarding promotions, training and employee engagement are often made without fully understanding the underlying patterns in employee behavior. This project aims to address this gap by modeling the relationship between workplace factors and employee performance to help organizations make data-informed improvements in workforce management and to guide actionable HR strategies.

I. Primary Research Objectives

To identify key workplace and demographic factors that influence employee performance and productivity

II. Secondary Research Objectives

1. To quantify the strength and direction of relationships between these factors and performance outcomes using statistical techniques.

2. To develop predictive models (statistical or machine learning) that forecast employee performance based on measurable variables.

3. To analyze how job roles, work conditions, and personal characteristics contribute to variations in absenteeism and performance.

4. To categorize employees into distinct groups based on their patterns of absenteeism and levels of job performance.

5. To propose actionable and data-driven interventions aimed at reducing absenteeism and enhancing employee performance.

- **Significance of the Study**

Novelty : This study presents a novel, regression-based framework that examines 22 interconnected variables across demographic, professional, and workplace domains to predict employee performance. This dataset has previously been analyzed primarily with a focus on employee attrition,aiming to identify factors associated with employees leaving the organization.By shifting the focus from attrition to performance ,this study offers new insights into how various employee attributes relate to their performance outcome.

Methodology

Data Collection & Preparation :

The dataset utilized in this study consists of Employee records, each containing information on a range of workplace factors and Employee performance ratings. The data was obtained from https://www.kaggle.com/datasets/ziya07/employee-attrition-prediction-dataset . Each observation represents a unique employee.

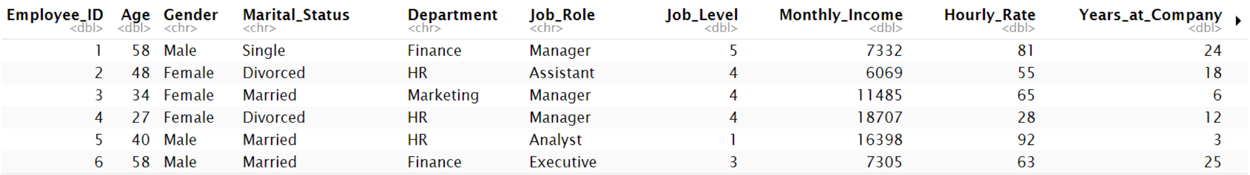
- ❖ This dataset included variables as follows,

1. Employee_ID: Unique identifier for each employee.
2. Age: Age of the employee.
3. Gender: Gender of the employee.
4. Marital_Status: Marital status of the employee (Single, Married, Divorced).
5. Department: Department the employee works in (e.g., HR, IT, Sales, Marketing).
6. Job_Role: Specific role within the department (e.g., Manager, Analyst).
7. Job_Level: Level in the organizational hierarchy.
8. Monthly_Income: Monthly salary of the employee.
9. Hourly_Rate: Rate per hour for hourly employees.
10. Years_at_Company: Number of years the employee has been with the company.
11. Years_in_Current_Role: Number of years the employee has been in their current role.
12. Years_Since_Last_Promotion: Time since the employee's last promotion.
13. Work_Life_Balance: Rating of work-life balance.
14. Job_Satisfaction: Rating of job satisfaction (1-5 scale).
15. Performance_Rating: Performance rating (1-5 scale).
16. Training_Hours_Last_Year: Number of training hours completed in the past year.
17. Overtime: Whether the employee works overtime (Yes/No).
18. Project_Count: Number of projects managed by the employee.
19. Average_Hours_Worked_Per_Week: Average working hours per week.
20. Absenteeism: Number of days the employee was absent in the past year.
21. Work_Environment_Satisfaction: Rating of work environment satisfaction.
22. Relationship_with_Manager: Rating of the relationship with the manager.
23. Job_Involvement: Rating of job involvement.
24. Distance_From_Home: Distance from home to the workplace (in kilometers).
25. Number_of_Companies_Worked: Total number of companies the employee has worked for.

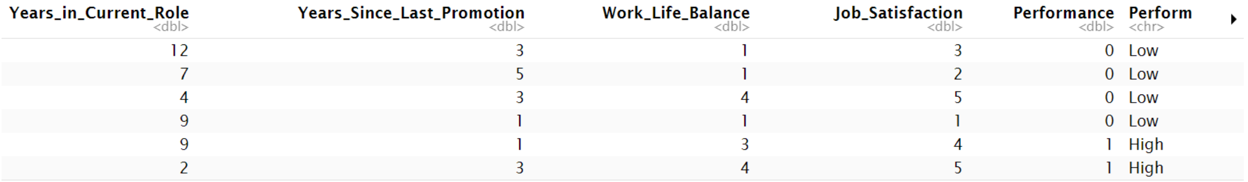
26. Attrition: The target column (Yes/No) indicating whether the employee left the company.

After obtaining the dataset, the research topic was selected by considering the nature of the variables available in the dataset. So selected topic is **Modeling the Impact of Workplace Factors on Employee Performance**. Accordingly, selected **performance** as response variable.

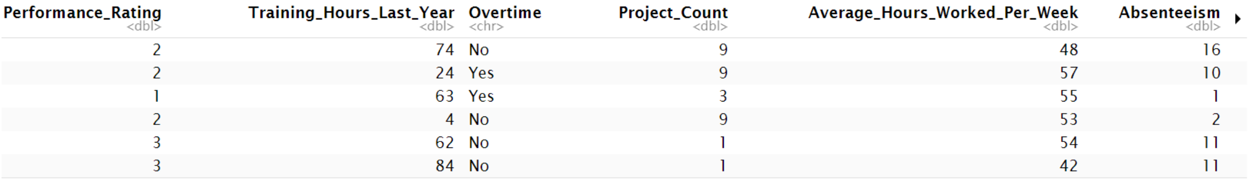
The performance variable in the dataset was originally recorded as ratings from 1 to 4. To simplify the analysis rating were selected such that values 1 and 2 were classified as “low” performance , and values 3 and 4 were classified as “high” performance. As a result, the response variable became binary. Accordingly , we selected a logistic regression model to evaluate the relationship between the Employee performance and various workplace factors.



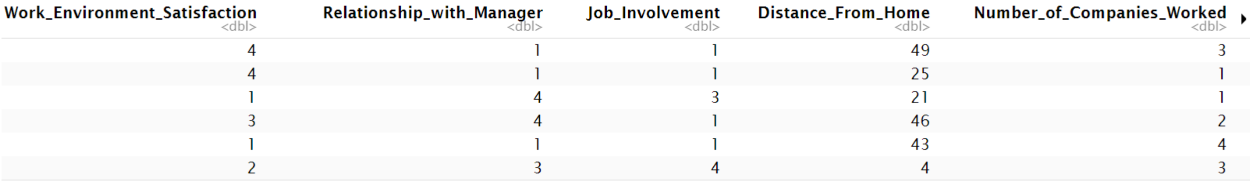
Employee_ID	Age	Gender	Marital_Status	Department	Job_Role	Job_Level	Monthly_Income	Hourly_Rate	Years_at_Company
<dbl>	<dbl>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	58	Male	Single	Finance	Manager	5	7332	81	24
2	48	Female	Divorced	HR	Assistant	4	6069	55	18
3	34	Female	Married	Marketing	Manager	4	11485	65	6
4	27	Female	Divorced	HR	Manager	4	18707	28	12
5	40	Male	Married	HR	Analyst	1	16398	92	3
6	58	Male	Married	Finance	Executive	3	7305	63	25



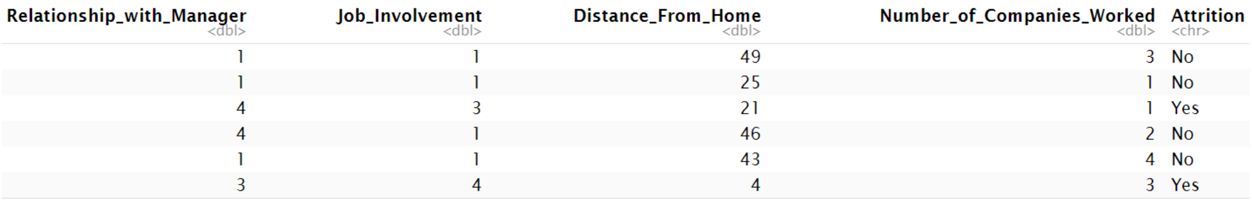
Years_in_Current_Role	Years_Since_Last_Promotion	Work_Life_Balance	Job_Satisfaction	Performance	Perform
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
12	3	1	3	0	Low
7	5	1	2	0	Low
4	3	4	5	0	Low
9	1	1	1	0	Low
9	1	3	4	1	High
2	3	4	5	1	High



Performance_Rating	Training_Hours_Last_Year	Overtime	Project_Count	Average_Hours_Worked_Per_Week	Absenteeism
<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>
2	74	No	9	48	16
2	24	Yes	9	57	10
1	63	Yes	3	55	1
2	4	No	9	53	2
3	62	No	1	54	11
3	84	No	1	42	11



Work_Environment_Satisfaction	Relationship_with_Manager	Job_Involvement	Distance_From_Home	Number_of_Companies_Worked
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
4	1	1	49	3
4	1	1	25	1
1	4	3	21	1
3	4	1	46	2
1	1	1	43	4
2	3	4	4	3



Relationship_with_Manager	Job_Involvement	Distance_From_Home	Number_of_Companies_Worked	Attrition
<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	1	49	3	No
1	1	25	1	No
4	3	21	1	Yes
4	1	46	2	No
1	1	43	4	No
3	4	4	3	Yes

Preprocessing

Prior to model development, the data underwent preprocessing procedure to ensure completeness, accuracy and suitability for regression analysis :

- Checked for missing values.
    - According to the R output this dataset has not any missing values.
- Categorical variables were recoded into factor levels suitable for analysis in R.
- Outliers were detected using boxplots.
- Skewed continuous variables were transformed were necessary to approximate normality.

Model Fitting

- The logistic regression model was fitted using R-studio. The model estimates the log odds of an Employee achieving a high performance rating as a function of various workplace factors. The general form of the logistic regression model is :

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{r-1} x_{r-1} + \beta_{r+1} x_{r+1} + \cdots + \beta_p x_p$$

Where;

π = Probability of High Performance

X_1, X_2 ,............, X_p = Predictor variables

$B_{0,}$ β_1, β_2, ..., β_p = Regression coefficients

- The coefficients were estimated using the R-stuido.
- Before fitting the model, the dataset was split into two datasets as “train” and “test” By 0.8:0.2 split ratio.
- The model was fitted using the test dataset with all predictors an then multicollinearity was checked among predictor variables.
- The model parameters were interpreted in terms of odds ratios.
- Used stepwise subset selection for select best model.
- Fitted last model with significance predictors.

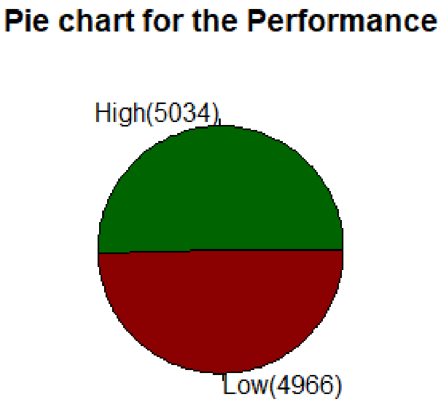
Model Evaluation

- The Chi-square goodness of fit test and the likelihood ratio test were performed to assess the model’s adequacy.
- A confusion matrix was created to evaluate how well the model performed on the test dataset.
- Model accuracy was calculated using confusion matrix.

.

Descriptive Analysis

- Pie chart for performance variable(response variable)



This pie chart illustrates the distribution of the performance variable within dataset.5034 number of people classified as having High Performance while 4966 no of people classified as having Low Performance. This indicates the High Performance is more prevalent in the population than Low Performance

- Used summary() function to find descriptive statistics such as Minimum, Maximum, Mean, Median, Quartiles to summarize both numerical and categorical variables.
  - **Job Role**

• Top Roles: Analyst (2572), Assistant (2538), Executive (2476), Manager (2414)

• Interpretation: Job roles are well-distributed, with no overwhelming concentration in a single role.

- **Job Level**

• Min: 1

• 1st Quartile (Q1): 2

• Median (Q2): 3

• Mean: 2.991

• 3rd Quartile (Q3): 4

• Max: 5

• Interpretation: Most employees fall between Job Levels 2 to 4, suggesting mid-level Position dominate.

- **Monthly Income**

• Min: 3000

• 1st Quartile (Q1): 7182

• Median (Q2): 11402

• Mean: 11437

• 3rd Quartile (Q3): 15680

• Max: 19999

• Interpretation: Monthly income ranges widely, from 3000 to nearly 20000, with most employees earning between 7182 and 15680.

- **Hourly Rate**

• Min: 15

• 1st Quartile (Q1): 36

• Median (Q2): 57

• Mean: 57.03

• 3rd Quartile (Q3): 78

• Max: 99

• Interpretation: Hourly rates very significantly, but the median and mean are closely aligned at 57, showing a balanced wage structure.

- **Years at Company**

• Min: 1

• 1st Quartile (Q1): 8

• Median (Q2): 15

• Mean: 14.94

• 3rd Quartile (Q3): 22

• Max: 29

• Interpretation: Many employees have long tenures, with a median of 15 years, indicating work forces stability and loyalty.

➢ Years in Current Role

• Min: 1

• 1st Quartile (Q1): 4

• Median (Q2): 7

• Mean: 7.45

• 3rd Quartile (Q3): 11

• Max: 14

• Interpretation: Employees typically stay in the same role for a long period, with most spending between 4 to 11 years.

➢ Years Since Last Promotion

• Min: 0

• 1st Quartile (Q1): 2

• Median (Q2): 4

• Mean: 4.47

• 3rd Quartile (Q3): 3

• Max: 64

• Interpretation: Promotion are relatively spaced out, with many employees promoted 2 to 7 years ago.

➢ Work-life Balance

• Scale : 1 to 4

• Mean: 2.25

• Interpretation: The average work-life balance is moderate, with most employees rating it between 2 and 3.

➢ Job Satisfaction

- Min: 1
- 1st Quartile (Q1): 2
- Median (Q2): 3
- Mean: 3.038
- 3rd Quartile (Q3): 4
- Max: 5
- Interpretation: Job satisfaction is generally high, with the majority rating 3 or 4 on a 5-point scale.

➢ Performance

- High : 5034
- Low: 4966
- Interpretation: The dataset has a nearly equal distribution between high and low performers.

➢ Training Hours Last Year

- Min: 0
- 1st Quartile (Q1): 25
- Median (Q2): 49
- Mean: 49.59
- 3rd Quartile (Q3): 75
- Max: 99
- Interpretation: Training hours are well-distributed, with most employees receiving between 25 and 75 hours of training annually.

➢ Overtime

- Yes: 4897
- No: 5103
- Interpretation: Overtime is evenly distributed among employees, indicating moderate work demands.

➢ Project Count

- Min: 1
- 1st Quartile (Q1): 3
- Median (Q2): 5
- Mean: 4.984
- 3rd Quartile (Q3): 7
- Max: 9
- Interpretation: Most employees handle between 3 to 7 projects, showing a reasonable workload.

➢ Average Hours Worked Per Week

- Min: 30
- 1st quartile (Q1): 37
- Median (Q2): 45
- Mean: 44.47
- 3rd Quartile (Q3): 52
- Max: 59
- Interpretation: The average workweek is around 44-45 hours, consistent with full-time work expectations.

➢ Absenteeism

- Min: 0
- 1st Quartile (Q1): 4
- Median (Q2): 9
- Mean: 9.41
- 3rd Quartile (Q3): 14

- Max: 19
- Interpretation: Most employees miss 4 to 14 days per year, indicating moderate levels of absenteeism.

➢ Distance form Home

- Min: 1
- 1st Quartile (Q1): 13
- Median (Q2): 25
- Mean: 25.27
- 3rd Quartile (Q3): 37
- Max: 49
- Interpretation: Commute distances vary, but most employees live within 13 to 37 units from their workplace.

➢ Job Involvement

- Scale: 1-4
- Mean: 2.505
- Interpretation: Employee involvement in their jobs is generally moderate.

➢ Relationship with Manager

- Scale: 1-4
- Mean: 2.491
- Interpretation: The relationship with managers is also average, with most employees rating it between 2 and 3.

➢ Work Environment Satisfaction

- Scale: 1-4
- Mean: 2.493
- Interpretation: Work environment satisfaction tends to be average, with many employees giving it mid-range ratings.

- **Number of Companies Worked**

• Min: 1

• 1st Quartile (Q1): 2

• Median (Q2): 2

• Mean: 2.517

• 3rd Quartile (Q3): 4

• Max: 5

• Interpretation: Most employees have experience with 2 to 4 companies,showing a moderately diverse work history

- **Attrition**

• Yes: 1997

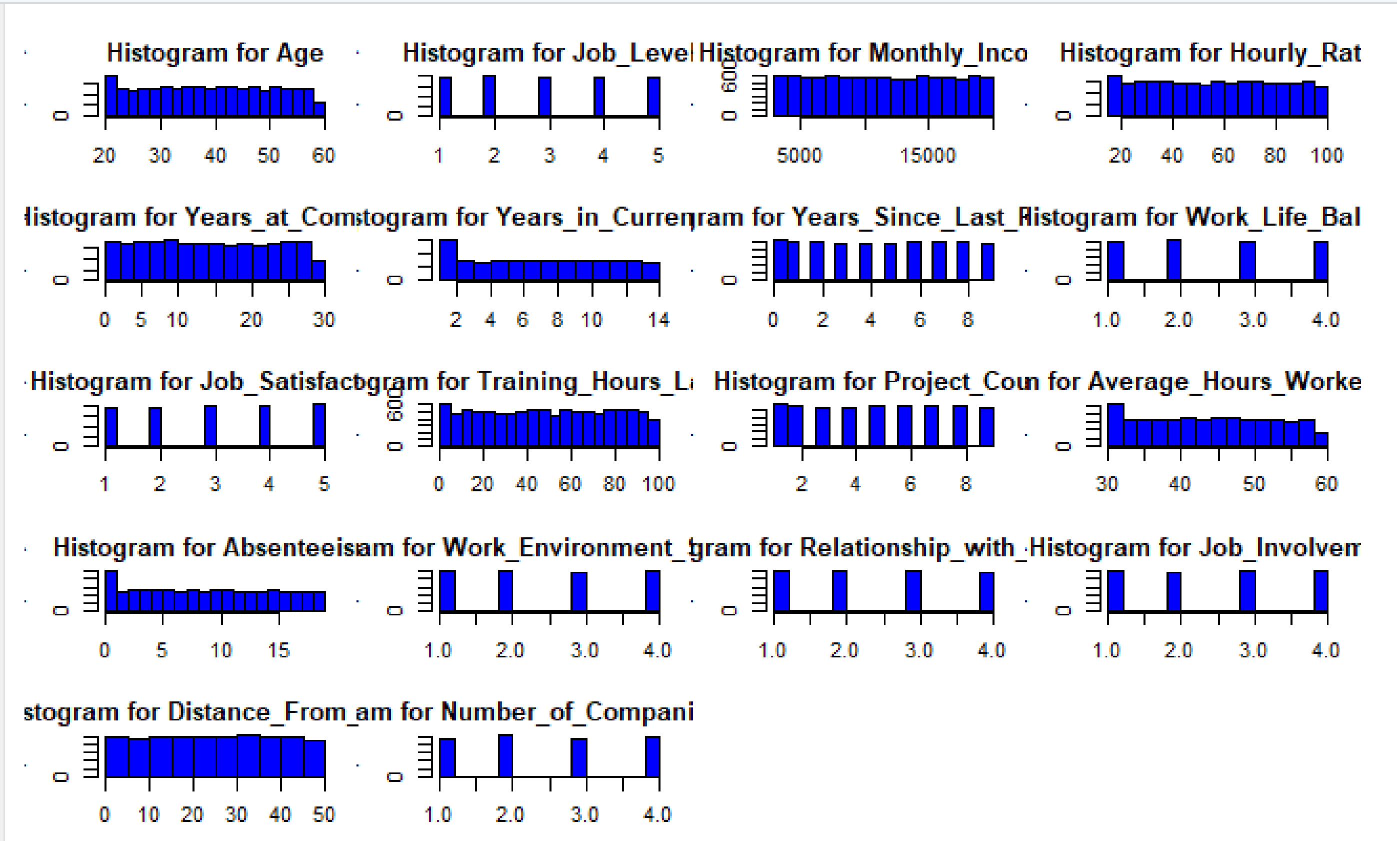
• No: 8003

• Interpretation: About 20% of the employees have left the organization, indicating a relatively stable workforce.

- Mean and Standard deviation of numerical variables.

Variable	Mean	Sd
Age	39.5618	11.454986
Job level	2.9908	1.410643
Monthly Income	11436.7167	4926.528302
Hourly Rate	57.0323	24.703261
Years at Company	14.9362	8.431657
Years in Current Role	7.4513	4.042903
Years since Last Promotion	4.4719	2.891617
Work Life Balance	2.5024	1.112348
Job Satisfaction	3.0380	1.414764
Training Hours Last Year	49.5889	28.801393

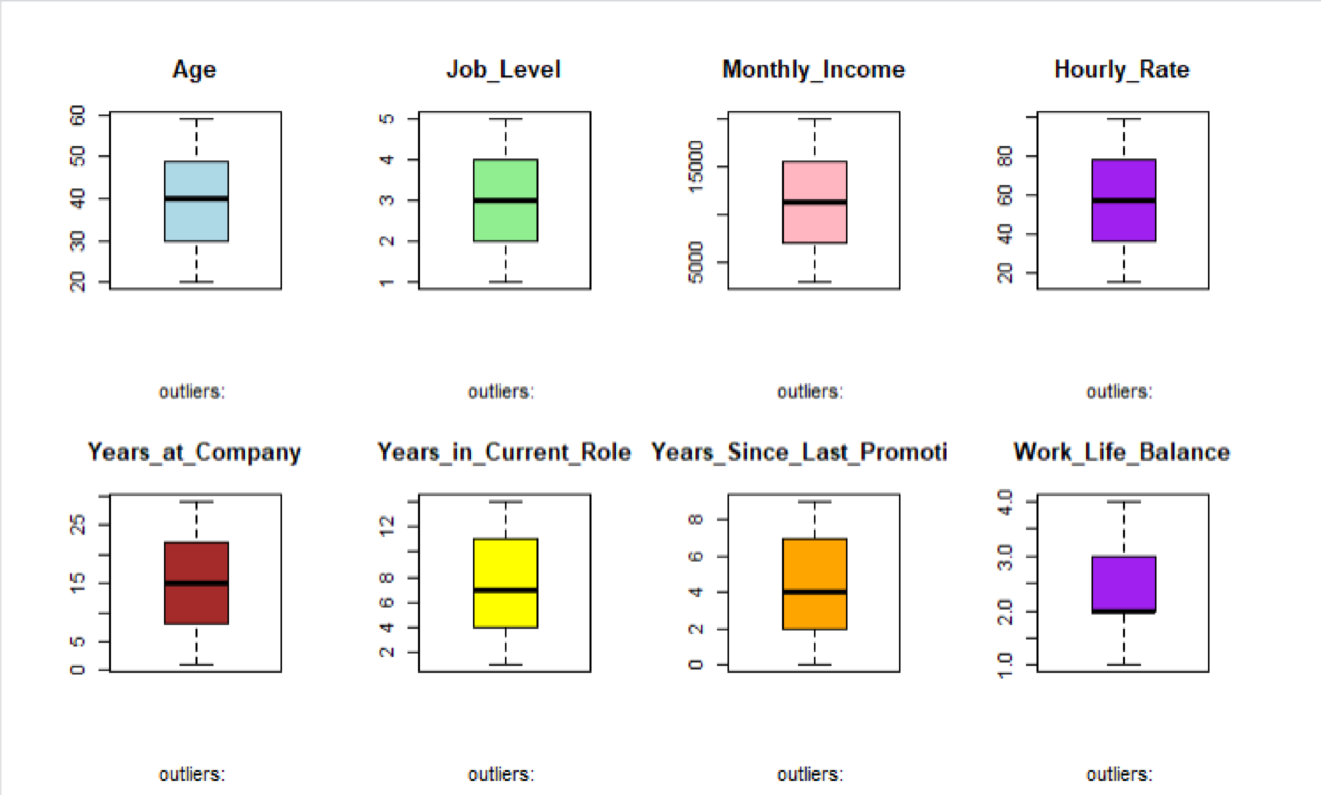
Histograms were used to visualize numerical variables.

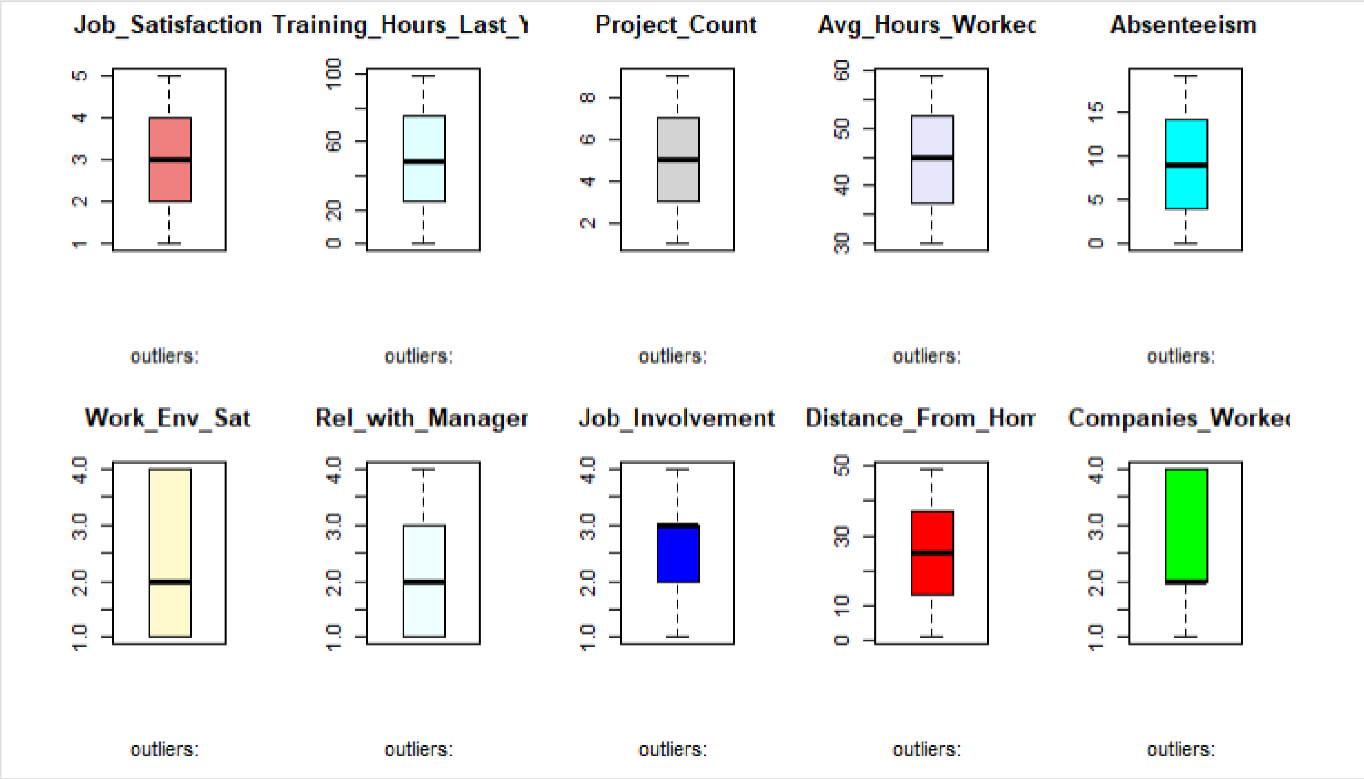


Age,monthly_income,Hourly_Rate,Training_Hours_Last_Year,Absenteesim,Distance_from_Home variables have relatively balanced distributions,suggesting no significant skewness and outliers.

Histrogram Job_level ,Work_life_balance, job_satisfaction_, Work_environment_satisfaction, Relationship with manager,Years sice last promotion,job_involvement and no of company variables show distinct bar-like patterns corresponding to their rating or category levels.

- **Boxplot** were used to identify outliers.





Accordingly to the boxplot there are no any outliers.

- Calculate skewness for variables

variable	skewness
Years_Since_Last_Promotion	0.015217678
Job_Satisfaction	-0.034540544
Relationship_with Manager	0.014011848
Work_Life_Balance	0.015038630
Training_Hours_Last_Year	-0.005605580
Job_Involvement	-0.011169225
Job_Level	0.007873296
Project_Count	-0.008448485
Distance_From_Home	-0.012448076
Monthly_Income	0.003241685
Average_Hours_Worked_Per_week	-0.003639286
Number_of_Companies_Worked	-0.005456899
Hourly_Rate	-0.001873735
Absenteeism	0.030083060
Years_in_Current_Role	0.001810751
Work_Environment_Satisfaction	0.013832390

The result indicated that all variables had skewness values close to zero,suggesting that their distributions are approximately symmetric.This implies that the data for these variables are evenly distributed around their mean values,with no significant long tails on either side.

As a result,no transformation or corrective measures were necessary to address skewness prior to model fitting.

Results and discussion

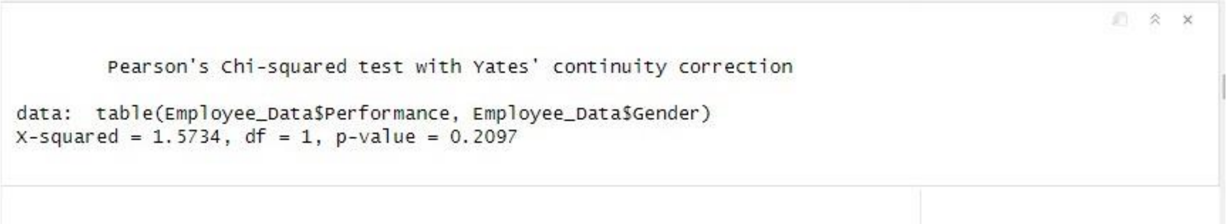
The aim is to model the Impact of Workplace Factors on Employee Performance. The data set has 10000 rows. And according to the preprocessing part, founded that there is no any null values, missing values, outliers, and also there is no any skewness. The data types of the data set was checked and then converted the char data types into factor data type. Then perform column was removed as the Y variable and the ID column. Also created the histograms for the numeric columns. In this way , checked whether the data set was suitable for fitting a model.

- **Bivariate analysis (Chi-square tests)**

  Null Hypothesis (H0): There is no association between the **Perform** variable and the predictor variables. (The variables are independent)

  Alternative Hypothesis (H1): There is an association between the **Perform** variable and the predictor variables. (The variables are not independent)

  Gender variable

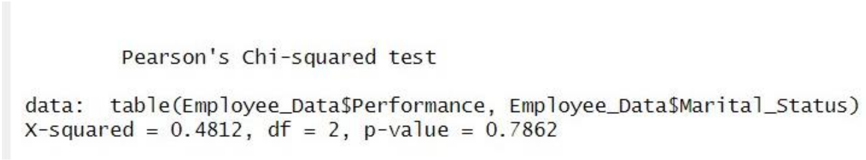


```
        Pearson's Chi-squared test with Yates' continuity correction

data:  table(Employee_Data$Performance, Employee_Data$Gender)
X-squared = 1.5734, df = 1, p-value = 0.2097
```

- P- value > 0.05. Therefore do not reject null hypothesis. Can conclude that there is no association between the Perform variable and the Gender variable.

  Marital_Status variable

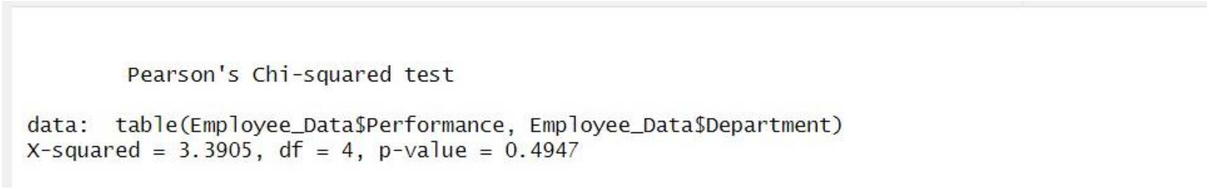


```
        Pearson's Chi-squared test

data:  table(Employee_Data$Performance, Employee_Data$Marital_Status)
X-squared = 0.4812, df = 2, p-value = 0.7862
```

- P-value > 0.05.Therefore do not reject the null hypothesis. Can conclude that there is no association between the perform variable and the Marital_Status variable.

Department variable

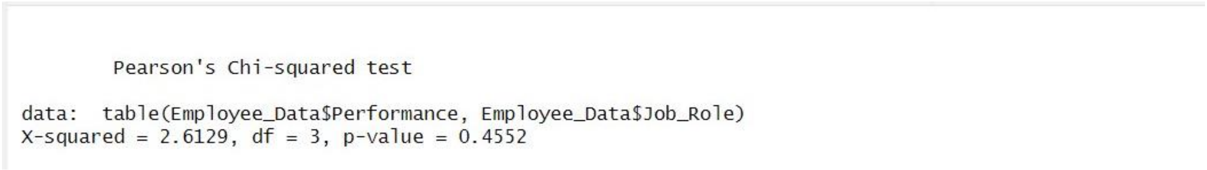


```
        Pearson's Chi-squared test

data:  table(Employee_Data$Performance, Employee_Data$Department)
X-squared = 3.3905, df = 4, p-value = 0.4947
```

- P-value > 0.05.Therefore do not reject the null hypothesis. Can conclude that there is no association between the perform variable and the Department variable.

<u>Job Role variable</u>

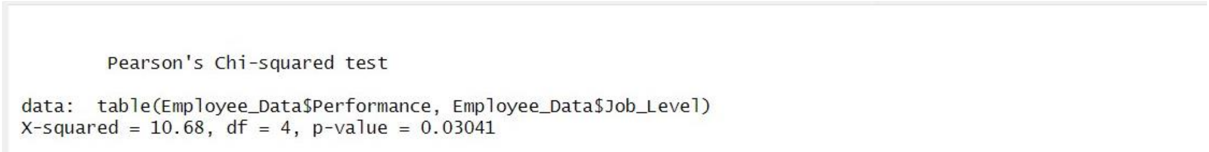


```
        Pearson's Chi-squared test

data:  table(Employee_Data$Performance, Employee_Data$Job_Role)
X-squared = 2.6129, df = 3, p-value = 0.4552
```

- P-value > 0.05.Therefore do not reject the null hypothesis. Can conclude that there is no association between the perform variable and the Job_Role variable.

<u>Job levelvariable</u>

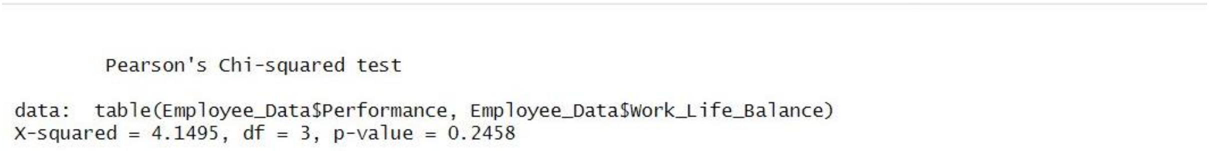


```
        Pearson's Chi-squared test

data:  table(Employee_Data$Performance, Employee_Data$Job_Level)
X-squared = 10.68, df = 4, p-value = 0.03041
```

- P-value <0.05.Therefore reject the null hypothesis. Can conclude that there is association between the perform variable and the Job_Level variable.

Work_life balance variable

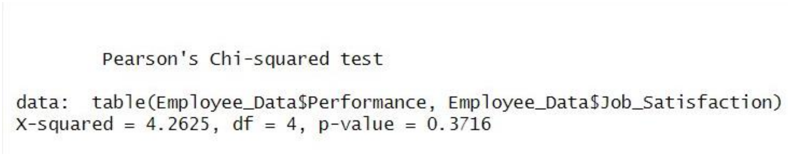


```
        Pearson's Chi-squared test

data:  table(Employee_Data$Performance, Employee_Data$Work_Life_Balance)
X-squared = 4.1495, df = 3, p-value = 0.2458
```

- P-value > 0.05.Therefore do not reject the null hypothesis. Can conclude that there is no association between the perform variable and the work_Life_Balance variable.

Job_satisfaction variable

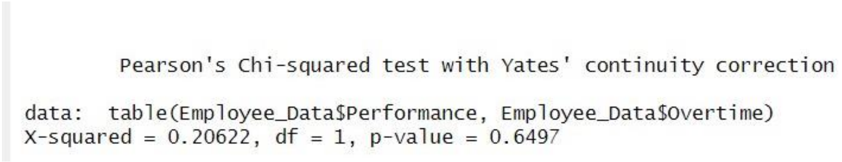


```
        Pearson's Chi-squared test

data:  table(Employee_Data$Performance, Employee_Data$Job_Satisfaction)
X-squared = 4.2625, df = 4, p-value = 0.3716
```

- P-value > 0.05.Therefore do not reject the null hypothesis. Can conclude that there is no association between the perform variable and the work Job_Satisfaction variable.

Overtime variable

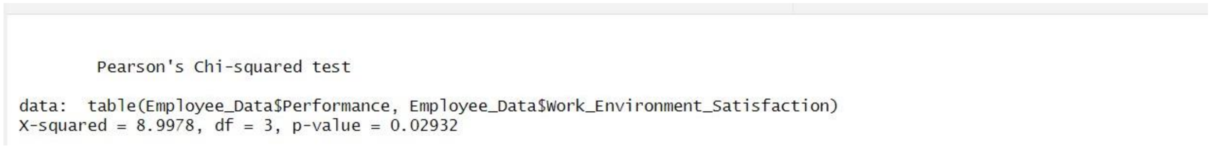


```
        Pearson's Chi-squared test with Yates' continuity correction

data:  table(Employee_Data$Performance, Employee_Data$Overtime)
X-squared = 0.20622, df = 1, p-value = 0.6497
```

- P-value > 0.05 .Therefore do not reject the null hypothesis. Can conclude that there is no association between the perform variable and the overtime variable.

Envioronment_satisfaction variable

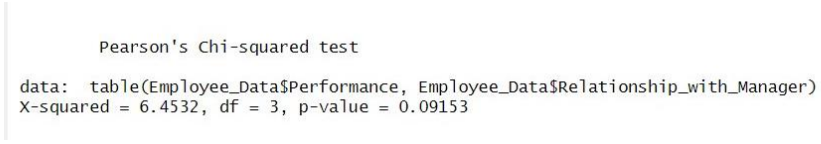


```
        Pearson's Chi-squared test

data:  table(Employee_Data$Performance, Employee_Data$Work_Environment_Satisfaction)
X-squared = 8.9978, df = 3, p-value = 0.02932
```

- P-value < 0.05 .Therefore reject the null hypothesis. Can conclude that there is association between the perform variable and the Work_Environment_Satisfaction variable.

Relationship with manager variable

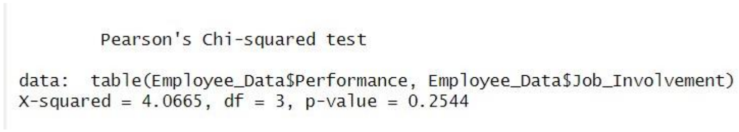


```
        Pearson's Chi-squared test

data:  table(Employee_Data$Performance, Employee_Data$Relationship_with_Manager)
X-squared = 6.4532, df = 3, p-value = 0.09153
```

- P-value > 0.05 .Therefore do not reject the null hypothesis. Can conclude that there is no association between the perform variable and the Relationship_with_Manager variable.

Job_involvement Variable

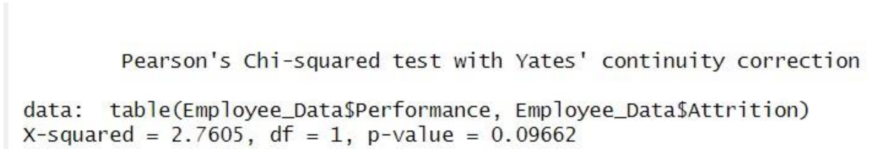


```
        Pearson's Chi-squared test

data:  table(Employee_Data$Performance, Employee_Data$Job_Involvement)
X-squared = 4.0665, df = 3, p-value = 0.2544
```

- P-value > 0.05 .Therefore do not reject the null hypothesis. Can conclude that there is no association between the perform variable and the Job_Involvement variable.

Attrition variable

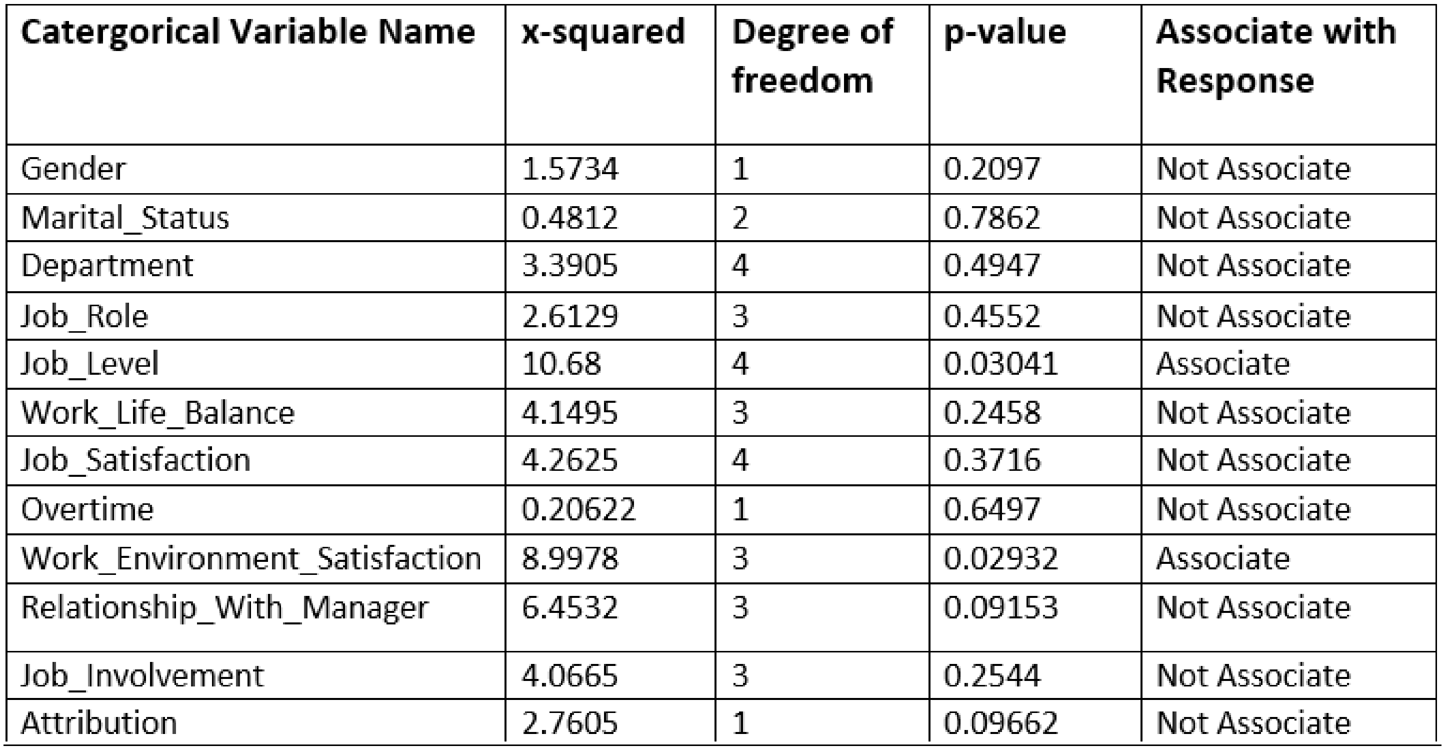


```
        Pearson's Chi-squared test with Yates' continuity correction

data:  table(Employee_Data$Performance, Employee_Data$Attrition)
X-squared = 2.7605, df = 1, p-value = 0.09662
```

- P-value > 0.05 .Therefore do not reject the null hypothesis. Can conclude that there is no association between the perform variable and the Attrition variable.

Summary of the Chi-Square test



Catergorical Variable Name	x-squared	Degree of freedom	p-value	Associate with Response
Gender	1.5734	1	0.2097	Not Associate
Marital_Status	0.4812	2	0.7862	Not Associate
Department	3.3905	4	0.4947	Not Associate
Job_Role	2.6129	3	0.4552	Not Associate
Job_Level	10.68	4	0.03041	Associate
Work_Life_Balance	4.1495	3	0.2458	Not Associate
Job_Satisfaction	4.2625	4	0.3716	Not Associate
Overtime	0.20622	1	0.6497	Not Associate
Work_Environment_Satisfaction	8.9978	3	0.02932	Associate
Relationship_With_Manager	6.4532	3	0.09153	Not Associate
Job_Involvement	4.0665	3	0.2544	Not Associate
Attribution	2.7605	1	0.09662	Not Associate

Univariable logistic regression model fit for each predictor(summary)

Variable Name	**Estimate**	**Standard Error**	**P value**	**AIC**
Age	-0.002295	0.001949	0.239	11093
Job_Level	-0.01655	0.01585	0.296	11093
Monthly_Income	-2.179e-06	4.541e-06	0.631	11094
Hourly_Rate	-0.0004842	0.0009062	0.593	11094
Years_at_Company	0.002696	0.002649	0.309	11093
Years_in_current_Role	-0.003523	0.005518	0.523	11094
Average_hours_woked_per_week	-0.001839	0.002603	0.480	11094
Absenteeism	-0.0002164	0.0038944	0.956	11094
Work_Environment_satisfaction	-0.04245	0.01993	0.0331	11090
Training_Hours_Last_Year	0.0016329	0.0007768	0.0356	11090
Project_count	-0.006528	0.008660	0.451	11094
Years_since_Last_promotion	0.001959	0.007749	0.800	11094
Work_life_Balance	-0.009102	0.020088	0.650	11094
Job_Satisfaction	0.008812	0.15805	0.577	11094
Relationship_with_manager	0.03238	0.20000	0.105	11092
Job_involvment	0.04324	0.2007	0.312	11090
Distance_from_Home	-0.001748	0.001576	0.267	11093
Number_of_companies_worked	-0.007261	0.020014	0.717	11094

The tables shows the summary of univariable logistic regression models for each predictor variables.

It seems p values of Work_Environment_satisfaction, Training_Hours_Last_Year are lower than 0.05 and it indicate that only these predictors are significant at their univariable regression models.

Splitting the data set

- The data set divided into two parts using 0.8 as the split ratio.
- Got 8000 rows as train data set. Therefore got 2000 rows as test data set.

Fitting binary logistic regression using "glm" function for all predictor variables

- The logistic regression model fitted to predict Performance using the all predictor variable(Excluded removed variables) in the dataset.
- The model summary indicates that Work Environment Satisfaction, Job Involvement, and Training Hours Last Year were statistically significant at the 5% significance level(p_value<0.05).
- Its Null deviance: 11090 on 7999 degrees of freedom
- Residual deviance: 11056 on 7969 degrees of freedom
- AIC: 11118

Multicollinearity among predictors

	GVIF
Age	1.004510
Gender	1.003941
Marital_Status	1.005901
Department	1.011895
Job_Role	1.014737
Job_Level	1.003363
Monthly_Income	1.004296
Hourly_Rate	1.002746
Years_at_Company	1.002920
Years_in_Current_Role	1.003898
Years_Since_Last_Promotion	1.002473
Work_Life_Balance	1.004363
Job_Satisfaction	1.003548
Training_Hours_Last_Year	1.005468
Overtime	1.002513
Project_Count	1.004298
Average_Hours_Worked_Per_Week	1.003357
Absenteeism	1.002704
Work_Environment_Satisfaction	1.004086
Relationship_with_Manager	1.005147
Job_Involvement	1.002396
Distance_From_Home	1.004135
Number_of_Companies_Worked	1.003903
Attrition	1.003702

All vif values are lower than 2.Thats mean there are no any linear relationship between predictor variables.So this model has not multicollinearity issue.

Variable selection using Stepwise method.

Using stepwise variable selection method, got 6 variables such as,

	Estimated coefficients
Training_Hours_Last_Year	-0.0016465
Work_Environment_Satisfaction	0.0417642
Relationship_with_Manager	-0.0327912
Job_Involvement	-0.0422470
AttritionYes	-0.0957194

but there is only 3 significant variables which are Training_Hours_Last_Year , Work_Environment_Satisfaction, Job_Involvement

Null deviance: 11090 on 7999 degrees of freedom
Residual deviance: 11071 on 7994 degrees of freedom
AIC: 11083

- **Fit the model with significance variables**

This is the final model with the significance variables from stepwise selection method.

	Estimated coefficients
Training_Hours_Last_Year	-0.0016385
Work_Environment_Satisfaction	0.0424220
Job_Involvement	-0.0423240

Null deviance: 11090 on 7999 degrees of freedom
Residual deviance: 11077 on 7996 degrees of freedom
AIC: 11085

Here the Residual deviance < Null deviance. The difference between these two is greater than the previous one. Therefore can say this model fitted good.

Summary of all fitted Models.

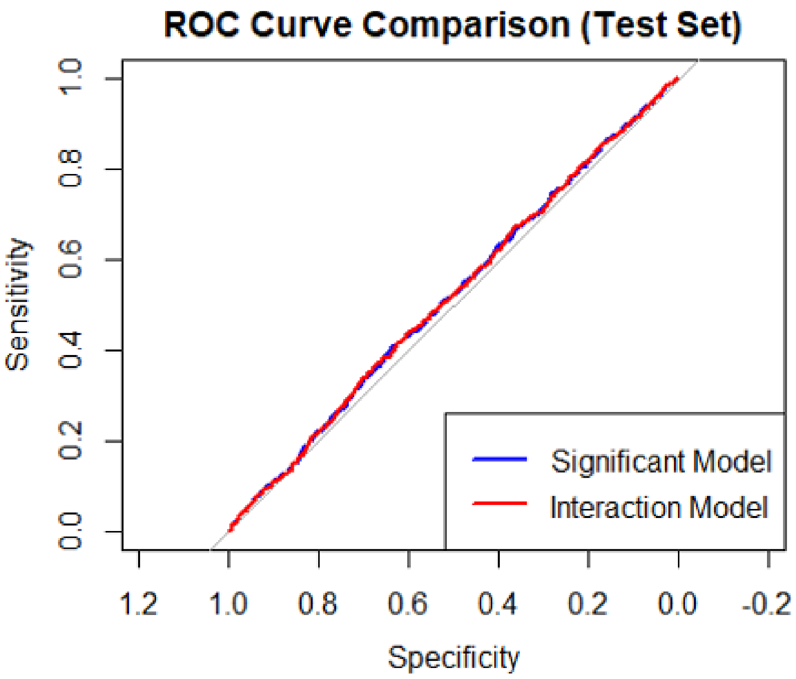
Model	Predictors	Significant Predators	Null deviance	Residuals deviance	AIC
Full model	All predictors in the data set	Training_Hours_Last_Year Work_Environment_Satisfaction Job_Involvement	11090	11056	11118
Step_model	Training_Hours_Last_Year Work_Environment_Satisfaction Relationship_with_Manager Job_Involvement Attrition	Training_Hours_Last_Year Work_Environment_Satisfaction Job_Involvement	11090	11071	11083
Significant model	Training_Hours_Last_Year Work_Environment_Satisfaction Job_Involvement	Training_Hours_Last_Year Work_Environment_Satisfaction Job_Involvement	11090	11070	11085

In step_model the AIC values is greater than the significant of the stepwise selection method but difference is 2.So it can say both model are equaly well supported. But Attrition and Relationship with Manager variables are not significant.So significant model was choosed as final model.

Final fitted model:

Performance=0.0788169 -0.0016385*(Training_Hours_Last_Year)+ 0.0424220*(Work_Environment_Satisfaction) -0.0423240*(Job_Involvement)

- **AUC (Area Under the Receiver Operating Characteristic (ROC) Curve)**



This graph strongly suggests that both logistic regression models are not effectively discriminating between the positive and negative classes in the dataset

Area under the curve of the step model = 0.5177

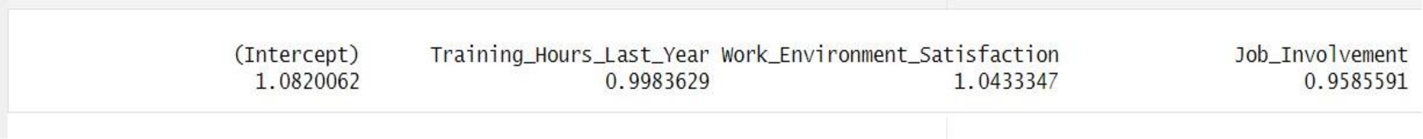
Area under the curve of the interaction model = 0.518

Therefore, adding interaction terms did not significantly improve the predictive power of the model, as both models have nearly identical AUC values. So, we used the model without interactions as the optimal model.

- **Check multicollinearity using VIF values**

All the VIF values < 5. Therefore can conclude that there is no multicollinearity. Therefore there is no correlation among predictors.

- **Parameter interpretation in terms of odds ratios**



```
      (Intercept) Training_Hours_Last_Year Work_Environment_Satisfaction       Job_Involvement
        1.0820062                0.9983629                     1.0433347             0.9585591
```

If odds ratio > 1 → increases chance of Y = 1

If odds ratio < 1 → decreases chance of Y = 1

- Odds ratio of intercept >1 ; Therefore intercept increase chance of **perform** variable = 1.
- Odds ratio of Training_Hours_Last_Year < 1; Therefore Training_Hours_Last_Year decrease chance of **perform** variable = 1.
- Odds ratio of Work_Environment_Satisfaction >1 ; Therefore Work_Environment_Satisfaction increase chance of **perform** variable = 1.
- Odds of Job_Involvement < 1 ; Therefore Job_Involvementdecrease chance of **perform** = 1.

- **Goodness of fit test**

Goodness of fit using the Hosmer-Lemeshow test to confirm that good fit of the final model to the data.

Null hypothesis (H0): The model fits the data well.

Alternative hypothesis (Ha): The model does not fit the data well.

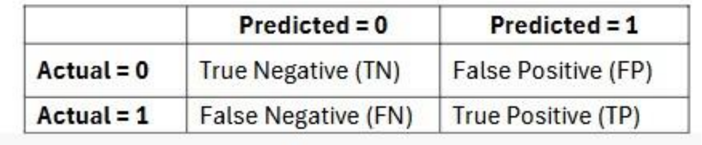
X-squared = 3.9359, df = 8, p-value = 0.8629

p-value > 0.05, At 5% significance level there is no evidence to reject the null hypothesis. In conclusion the model fit the data well.

▯

Model Evaluation

A confusion matrix



	Predicted = 0	Predicted = 1
Actual = 0	True Negative (TN)	False Positive (FP)
Actual = 1	False Negative (FN)	True Positive (TP)

	0	1
High	516	484
Low	487	513

Confusion matrix show the details in the table above.

According to the output confusion matrix ;

There are 516 values are actually 0 in the real data set & the predicted values are also 0 called True Negative.

There are 484 values are actually 0 in the real data set & the predicted values are 1 called False Positive There are 487 values are actually 1 in the real data set & the predicted values are 0 called False Negative.

There are 513 values are actually 1 in the real data set & the predicted values are also 1 called True Positive.

Accuracy

- There is 0.5145 accuracy in the fitted model. 51.45% accuracy is there.
- Accuracy is the proportion of **correct predictions** (both true positives and true negatives) out of all predictions.
- A 51.45% accuracy means the model correctly predicted performance about **half the time**.
- **51.45% is considered quite low.**

▯

▪ Relationship Between Results and Research Objectives

- **Primary Objective**

The logistic regression analysis identified **Work Environment Satisfaction**, **Job Involvement**, and **Training Hours Last Year** as statistically significant predictors of employee performance. Among these, Work Environment Satisfaction had a **positive** association with performance, while Training Hours and Job Involvement showed **negative** associations.

This supports the primary objective partially, while most variables showed no strong effect, a few **workplace-related factors** did emerge as statistically significant, suggesting that aspects of job satisfaction and training **do influence** performance outcomes.

Secondary objectives

The relationships were quantified successfully using appropriate statistical methods.

While models were developed and statistically valid, they had **limited predictive accuracy**, suggesting the need for better features or more complex modeling techniques.

The objective, to analyze how job roles, work conditions, and personal characteristics contribute to variations in absenteeism and performance, was not strongly supported by the results. These factors did not contribute meaningfully to variation in performance in the dataset.

Though limited in predictive accuracy, your findings support targeted interventions to enhance workplace conditions and employee experience.

▯

Conclusion

The final logistic regression model identified training hours last year, work environment satisfaction, and job involvement as significant predictors of employee performance. However, the model's predictive power was weak (auc = 0.52, accuracy = 51.45%), indicating these factors alone are not strong predictors. Removing insignificant variables improved model interpretability but did not enhance predictive accuracy. For better results, future analyses should include additional relevant predictors and consider advanced or ordinal modeling techniques

References

Bender, r., & grouven, u. (1998). Using binary logistic regression models for ordinal data with non proportional odds.

Journal of clinical epidemiology, 51(10), 809-816. Li, l., & lin, d. (2006). Ordinal regression analysis using generalized estimating equations.

Biometrics, 62(3), 688-695. Include key references used in the report, such as the kaggle dataset and important research papers

▯

Individual Contribution

Task Completed	PS/2021/236	PS/2021/145	PS/2021/056	PS/2021/227	PS/2021/002	PS/2021/095	PS/2021/057	PS/2021/244	PS/2021/058	PS/2021/158	PS/2021/144
find the dataset	🟦	🟦	🟦					🟦			
Creating Activity 1	🟦	🟦									
Creating Activity 2	🟦	🟦						🟦		🟦	
Creating Activity 3	🟦	🟦	🟦					🟦			
Methodology	🟦	🟦	🟦					🟦			
Descriptive Analysis	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Write an Interpretations			🟦				🟦	🟦			
Model Evaluation	🟦	🟦	🟦					🟦			
Fitting Binary Logistic regression		🟦						🟦			
Discussion & Results	🟦	🟦	🟦					🟦			
Creating Presentation Slides	🟦	🟦			🟦						
Final Report Creating	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦