Human Resource Dataset

Abstract. In this document, the human resource dataset is explored by applying various statistical techniques. For this, extensive exploratory data analysis is performed to analyze and understand the hidden patterns behind the key attributes and observations. Based on some important stats, data is passed through the parametric and non-parametric tests based on the normality check of the concerned variables. A set of machine learning models is applied to predict the satisfaction level of the employee using linear regression and also to predict whether the employee will leave the company or not using logistic regression. After reading this document, we will have a better grasp of the necessity for statistical tools in Human Resource Management, as well as an overview of many statistical techniques that are appropriate for specific HR functions.

Keywords: Shapiro-Wilk, Anderson-darling, Hypothesis Test, Parametric, Non-Parametric Chi-squared, Man-Whitney U Test, Wilcoxon Test, ANOVA, Spearman Correlation, Linear Regression, and Logistic Regression.

Table of Contents

1 Introduction

HR stands for Human Resource. Human resource is the department in any well-established organization that particularly handles employee-related operations and entertains them with their matters among an organization. Human resource management now become quantitative in its decision making which eventually enable businesses to make the data-driven decision using useful insights and stats. Furthermore, human resource management is now based on facts and figures with embryonic statistical techniques in the field of analytics.[3]

In the context of this article, data from Kaggle is extracted which gives information of employee key attributes in an organization. Using these attributes we shall explore some effective applications of statistical approaches, both basic and advanced, in human resource management for an anonymous organization, as well as illuminate some of the most widely used statistical techniques. **Table 1** shows an overview of the attributes and their types used in this study.

This document aims to apply the key statistical concepts in the analysis of data of different types and evaluate solutions for statistical problems. The key objective of this report includes:

- Implementation of Parametric or Non-parametric test based on the normality check.
- Predict the satisfaction level of the employee based on key attributes.
- Predict employees will be left the company based on time_spend_company, satisfaction_level, number_project, last_evaluation, promotion_last_5years, and salary.

Several tests and algorithms are performed to meet the above-mentioned objectives. These tests include Shapiro-Wilk, Anderson-darling, Hypothesis test, T-test, ANOVA, Chi-squared, Spearman Correlation while for prediction linear and logistic regression are used.

2 Exploratory Data Analysis

Exploratory data analysis is an essential analysis before proceeding to any algorithm or test. In exploratory data analysis, we analyze data to find hidden patterns and trends that exist in the dataset. Furthermore, data cleaning is a subprocess of EDA in which null values are dealt with appropriately and converted data into standard data types to perform further tests.

In this article, a broad exploratory data analysis is performed step by step to grasp the concealed development among the dataset. Initially, data is analyzed in terms of dimensionality, variable names, and their data types. The selected dataset consists of **14999** observations and **10** unique variables. Among these variables, 8 are numerical and 2 are categorical. Collectively all of the analysis can be summarized using a *glimpse* that is a part of a *dplyr* package in R. **Fig 1.**

```
## Rows: 14,999
## Columns: 10
## $ satisfaction_level    <dbl> 0.38, 0.80, 0.11, 0.72, 0.37, 0.41, 0.10, 0.92, ~
## $ last_evaluation       <dbl> 0.53, 0.86, 0.88, 0.87, 0.52, 0.50, 0.77, 0.85, ~
## $ number_project        <int> 2, 5, 7, 5, 2, 2, 6, 5, 5, 2, 2, 6, 4, 2, 2, 2, ~
## $ average_montly_hours  <int> 157, 262, 272, 223, 159, 153, 247, 259, 224, 142~
## $ time_spend_company    <int> 3, 6, 4, 5, 3, 3, 4, 5, 5, 3, 3, 4, 5, 3, 3, 3, ~
## $ Work_accident         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ left                  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ promotion_last_5years <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Department            <fct> sales, sales, sales, sales, sales, sales, sales,~
## $ salary                <fct> low, medium, medium, low, low, low, low, low, lo~
```

Fig 1: Summary of the Selected Dataset

Before proceeding further, we need to check the null values in the dataset and tackle them appropriately in such a way that our data doesn't become biased or skewed. The selected dataset doesn't contain any null values so we will consider the same dataset for processing. **Fig 2.**

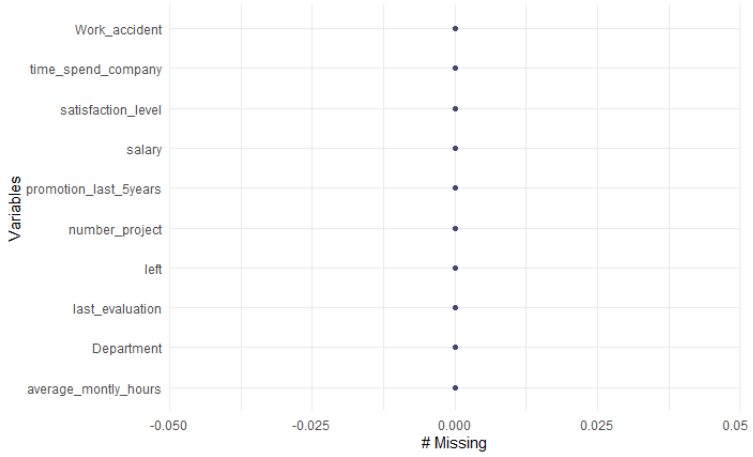

Fig 2: Checking Null values in a dataset

In this dataset, there are 2 categorical variables named department and salary. To visualize different categories and the percentage of each category in a department variable, a pie chart is selected. Pie chart for such scenarios gives a better and clear understanding. **Fig 3.**

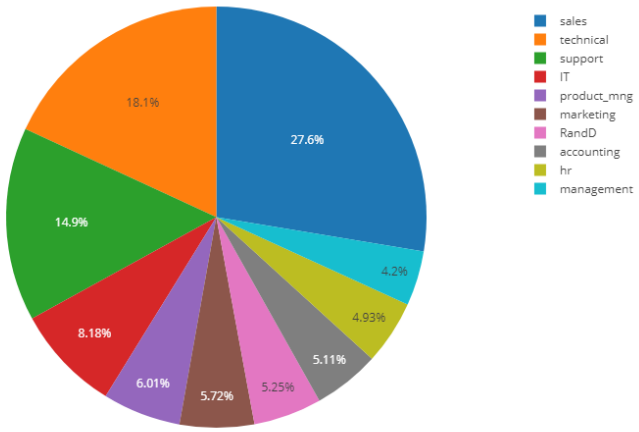

Fig 3: Various departments and their percentage in a company

This pie chart depicts that there are in total 10 departments in a company and sales department has the most observations followed by technical and support department while management department is the least common in the dataset which results in a conclusion that sales, technical and support department is the largest department among all and have a significant impact in the company performance.

In EDA, salary analysis across each department is made to analyze, employees of which department are taking the highest, lowest and medium salary. **Fig 4.**

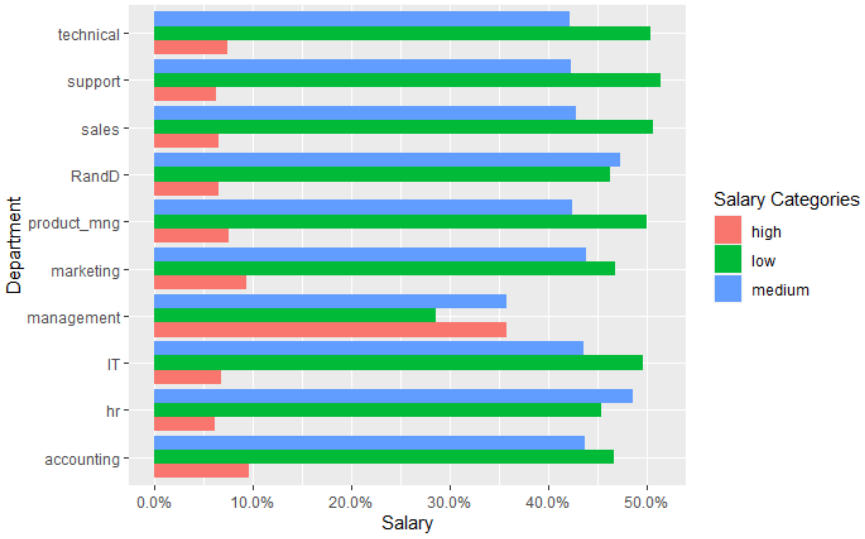

Fig 4: Salaries of the employee in each department

In the above bar graph, employees with medium and low salaries are most common among the company. On the other hand, employees in the management department have the highest salaries as compared to the employees in other departments.

In this dataset, the satisfaction level is one of the key attributes among all the variables. Keeping in view its importance, various comparisons are made to explore the behaviour of employees of different departments. **Fig 5.**

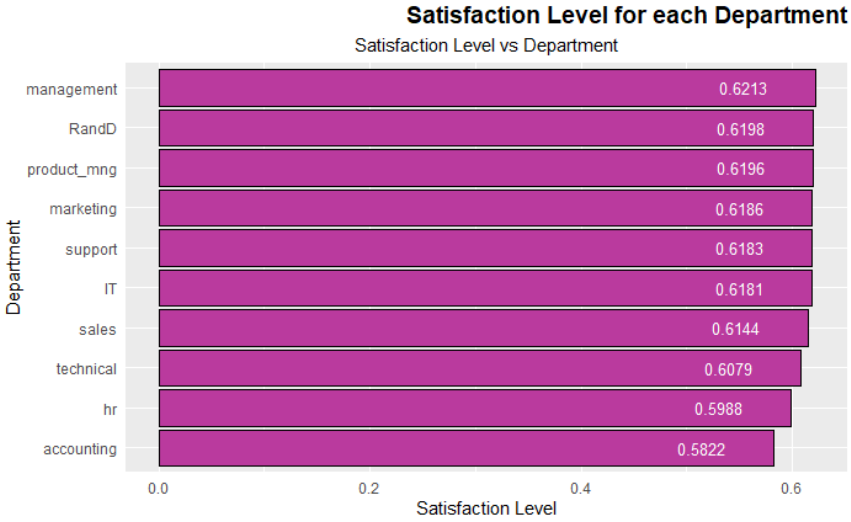

Fig 5: Satisfaction Level among different departments

The above bar graph gives us an insight that almost every department on average has equal satisfaction with few exceptions. Accounting and HR departments have the least satisfaction level amongst all.

Multiple unique key features describe the employee in each department. To find a relationship between the features, a correlation plot is required to get appropriate visuals and understanding of the features. **Fig 6.**

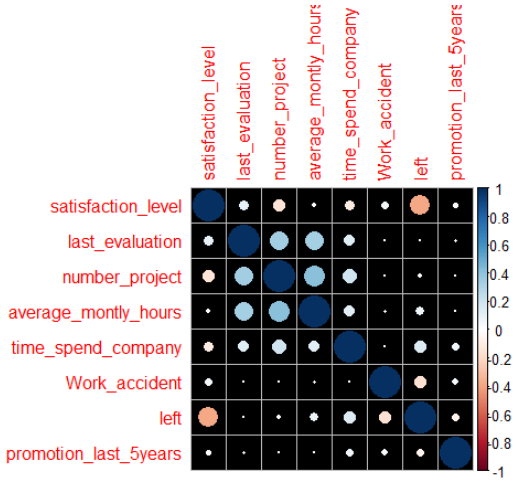

Fig 6: Correlation between the features of the dataset

The correlation plot shows that satisfaction_level has a significant impact on employee attrition. Moreover, some features are greatly influenced by others which include several cases which include number_project and last_evaluation, average_monthly_hours and last_evaluation, and average_montly_hours and number_project.

3 Statistical Techniques

Every data analysis approach relies on the information available about the subject to be effective. Data obtained using one or more of the standard data gathering methods or questionnaires is referred to as information. However, understanding various data analytics methodologies is required to convert the massive amount of data into useable information. There are two types of data collected: qualitative and quantitative. Qualitative data methodologies, despite their limits, provide for useful analysis. To get insight into the business, quantitative tools are frequently applied. Exploratory research can be applied to almost any qualitative or quantitative study. Graphical representations are commonly employed in exploratory inquiry. Confirmatory analysis, on the other hand, is based on more stringent statistical techniques.

In statistical concepts, there are some categories of data and tests. In data categorie,s there are six types of data present in any dataset. Following are the categories of those datasets:

- One Sample
- Two Dependent Samples
- Two Independent Samples
- More than two independent samples
- Correlation between two variables
- More than two dependent samples

Just like categories of data, there are two types of tests in statistical techniques through which respective data goes through. These types of tests are known as parametric tests and non-parametric tests. These two tests are further divided into six tests. To decide which type of test has to be implemented we perform a normality test to check the normality of the data. If data is normal or normally distributed, tests from the parametric category will be applied according to the given sample. On the contrary, if data isn't normally distributed, tests from the non-parametric category will be applied keeping in view the given sample. In the following sections, we will discuss statistical techniques on the qualitative and quantitative data.

3.1. Two Independent Samples

There are different categories of samples in our dataset. One of them is known as Two Independent Samples, in which two independent samples are taken from the dataset and undergo a statistical test. In this article, two independent variables are taken, named as, satisfaction_level and left for the respective statistical test. First, we perform a normality check to analyze whether the data is normally distributed or not. For this, two hypotheses are made Null Hypothesis and Alternative Hypothesis.

Null Hypothesis: Data is normally distributed.
Alternative Hypothesis: Data isn't normally distributed.

Histogram and Q-Q plot are drawn for graphical visualization of data distribution while Shapiro-Wilk and Anderson-Darling statistical tests are used for statistically analyzing the normality within the data. **Fig 7.**

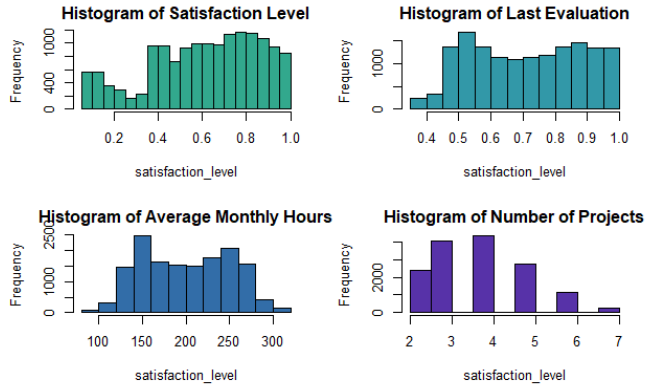

Fig 7: Distribution of Variables using Histogram

Fig 7 shows the distribution graph of multiple variables. For the independent sample test, the variable of concern is satisfaction_level, which is not normally distributed instead of this it's right-skewed.

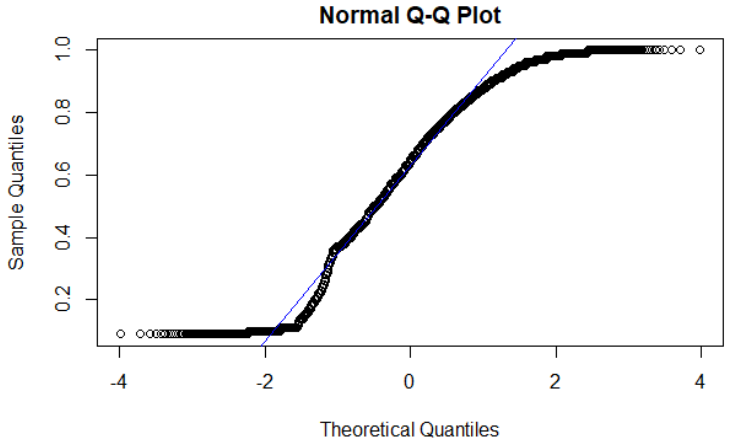

Fig 8: QQpplot for satisfaction_level

Shapiro-Wilk test has a limitation that it can only be implemented on 5000 observations. Keeping in view the deficiency an alternate test Anderson-Darling is applied to check the normality statistically. Shapiro-Wilk test is also applied by dividing the data into 3 chunks of 5000 observations to get the required results.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0900  0.4400  0.6400  0.6128  0.8200  1.0000

        Anderson-Darling normality test

data:  hr$satisfaction_level
A = 168.37, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  hr$satisfaction_level[0:5000]
W = 0.94543, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  hr$satisfaction_level[5001:10000]
W = 0.9563, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  hr$satisfaction_level[10001:14999]
W = 0.95056, p-value < 2.2e-16
```

Fig 8: Normality Tests

Fig 8 depicts the results of the Anderson-Darling and Shapiro-Wilk tests respectively. In both the tests, p-values in less than the significance level of 0.05 which shows that data isn't normally distributed.

From the above plots and tests, one can conclude that data for satisfaction level is right-skewed, which means that distribution is not normally distributed. Moreover, the p-value is much smaller than the threshold value i.e. 0.05. So we can reject the null hypothesis and go with the alternative hypothesis. The alternative hypothesis says that data is not normally distributed. Hence, we will now implement a non_parametric test i.e. Mann-Whitney U-test instead of T-test (Parametric Test).

3.1.1. Mann-Whitney U-Test

This test is a non-parametric test used for two independent samples and an alternative of a t-test. In this test, selected dimensions are satisfaction_level as a metric variable and left as a nominal variable.
Null Hypothesis: Left does not influence satisfaction_level
Alternative Hypothesis: Left influences satisfaction_level

```
        Wilcoxon rank sum test with continuity correction

data:  hr$satisfaction_level by hr$left
W = 30522915, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Fig 9: Mann-Whitney U-test

A Mann-Whitney U-test showed that left has significant influence (W = 30522915, p-value = 2.2e-16) on satisfaction level.

3.2. Chi-Squared Test

Qualitative data is predominantly categorical which means they are all about characters, names or categories that are not amenable to analysis. As a result, for such data, normal summary statistic measures are ruled out. Frequencies can only be analyzed with categorical data. A cross-tabulation of the frequencies is an excellent method to begin the inquiry.[2]

In this test we consider two variables, department and promotion_last_5years to perform the required test and find a relationship among the selected variables. Before performing the test, null and alternative hypotheses are made in which either of them can be a conclusion.

Null Hypothesis: No relation between department and promotion_last_5years
Alternative Hypothesis: There is a relationship between department and promotion_last_5years

A cross-tabulation displaying the frequencies of the departments and promotion_last_5years, whether they are promoted in the last 5 years or not. **Fig 10.**

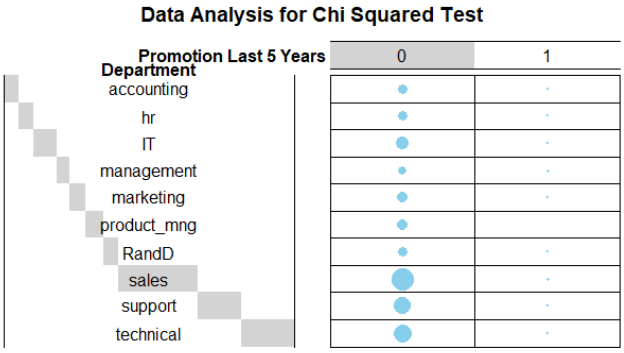

Fig 10: Correlation plot among departments and promotion_last_5years

If we want to look at the data descriptively, we will look at the frequencies for all conceivable combinations of the categories of the two variables in a cross-table. **Fig 11 a and b.**

	0	1
accounting	753	14
hr	724	15
IT	1224	3
management	561	69
marketing	815	43
product_mng	902	0
RandD	760	27
sales	4040	100
support	2209	20
technical	2692	28

	0	1
accounting	751	16
hr	723	16
IT	1201	26
management	617	13
marketing	840	18
product_mng	883	19
RandD	770	17
sales	4052	88
support	2182	47
technical	2662	58

Fig 11

a: Observed Frequencies *b: Expected Frequencies*

A Chi-square test of hypothesis would be used to evaluate the probability of any association between employees in a certain department and advancement. **Fig 12.**

```
        Pearson's Chi-squared test

data:  chisq_table
X-squared = 350.91, df = 9, p-value < 2.2e-16
```

Fig 12: Chi-Squared Test, the p-value is less than 0.05. Alternative Hypothesis Accepted.

An alternative hypothesis is accepted as the p-value is approximately approaching zero which depicts that there is no relationship among the selected variables.

3.3. More than Two Independent Samples (Kruskal-Wallis Test)

This is another type of data sample that is considered for finding some useful results. Using this data we can implement ANOVA from the parametric test and Kruskal-Wallis test from the non-parametric test. As the data in our dataset is not normally distributed as concluded by the preliminary tests in exploratory data analysis. Keeping in view the distribution of data, the Kruskal-Wallis test is implemented. For this test, Department and satisfaction_level are the selected dimensions.

Null Hypothesis: No significant differences between the satisfaction level of employees among various department
Alternative Hypothesis: There's a significant difference between the satisfaction_level of employees among various department

To implement the required test, data is prepared in such a way that it can be easily used in a particular test. A graphical representation of the graph is shown in **Fig 13 and Fig 14.**

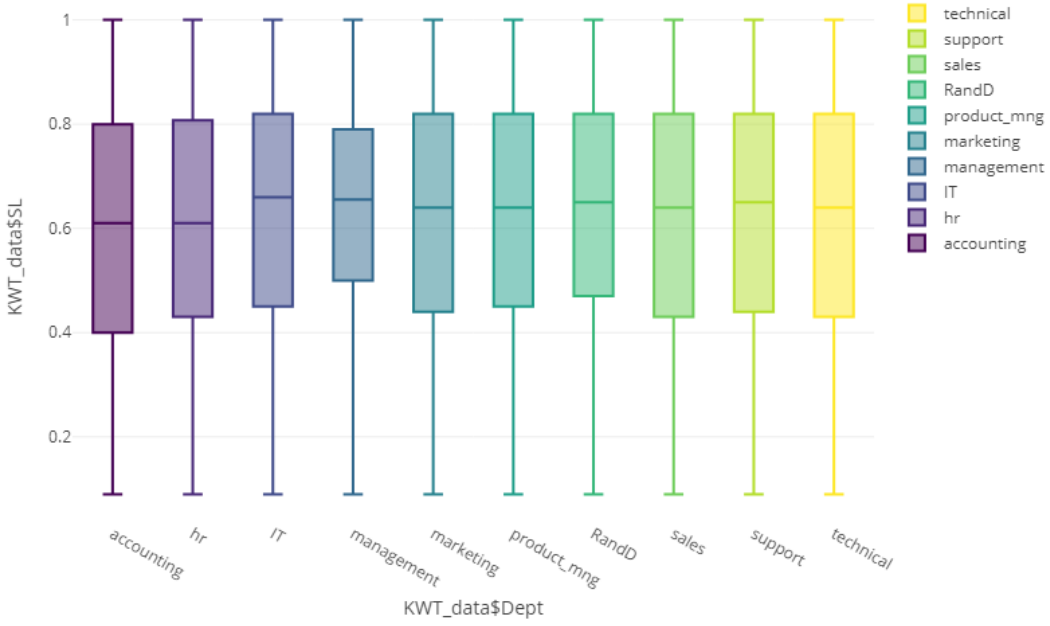

Fig 13: Boxplot visualization for departments and their satisfaction_level

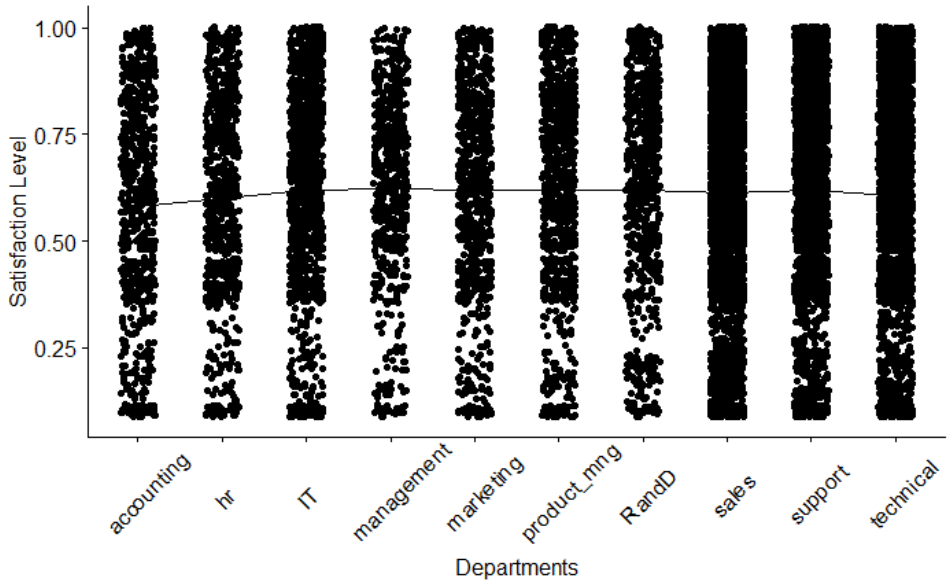

Fig 14: A line is drawn to see the difference of satisfaction_level between departments

A Kruskal-Wallis test is as follows in which we are computing if there is any significant difference between the satisfaction_level of employees in different departments. **Fig 15** shows a Kruskal-Wallis test using the R package.

```
        Kruskal-Wallis rank sum test

data:  KWT_data$SL by KWT_data$Dept
Kruskal-Wallis chi-squared = 18.296, df = 9, p-value = 0.03189
```

Fig 15: Kruskal-Wallis Test

The result depicts that the p-value is less than the significance level, so we can conclude that there are significant differences in satisfaction level among employees of different departments but as we can see that there isn't a big difference within calculated p-value and significance level. Keeping in view the current scenario, we can say that there is a small difference in satisfaction level among employees of different departments but not a massive one.

3.4. Correlation (Spearman Correlation)

Correlation is one of the statistical techniques which is used to see if two continuous variables are connected. This is an important measurement in the development of prediction algorithms. The correlation coefficient test indicates how closely two continuous variables are associated with each other. The correlation coefficient ranges from -1 to +1. The direction of the link between the two variables is indicated by the coefficient sign. When the value of one variable increases, the value of the other increases as well. If the sign is negative, an increase in one variable will cause a decrease in the other. As the coefficient value approaches 1, stronger will be the relation between the variables.

In the context of this document, for correlation, two variables are taken under consideration which includes "time_spend_company", "work_accident" and "number_project". Furthermore, we have to find whether to implement the non-parametric or parametric test. For this, the normality of considered variables will be checked. **Fig 16, 17 a and b.**

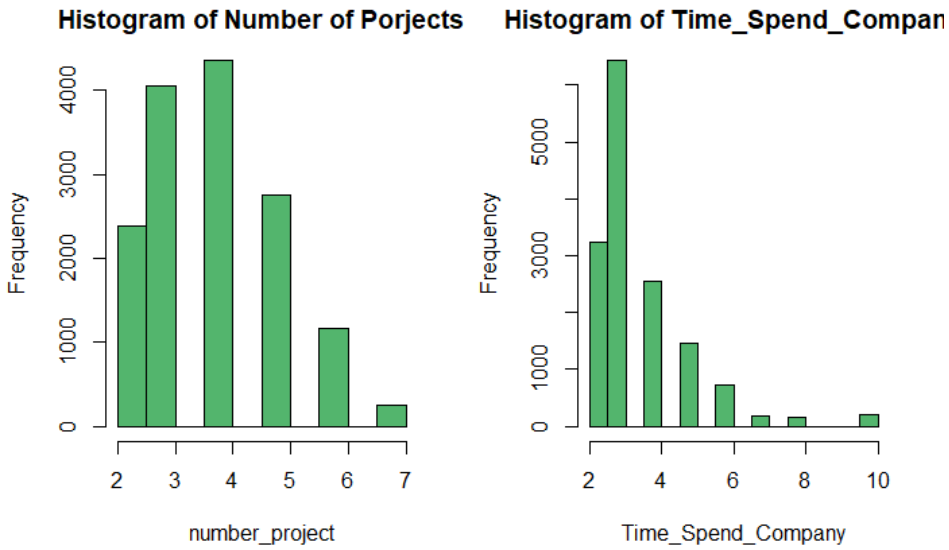

Fig 16: Data distribution of Number of Projects and Time Spend in Company

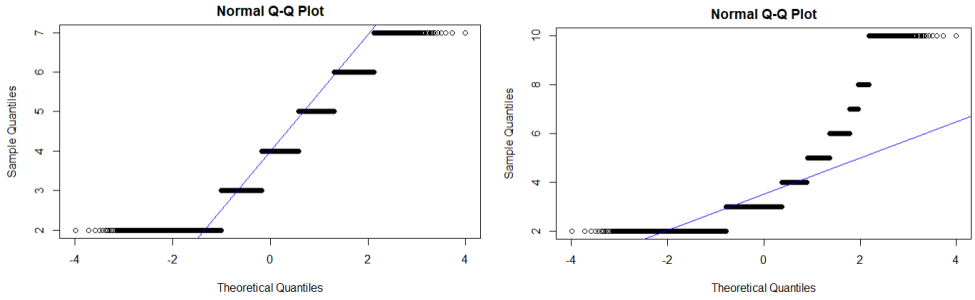

Fig 17 QQ Plots

a: number_project *b: time_spend_company*

The above graphs depict that data is not normally distributed it's left-skewed. For more clarification, we visualize the QQ plot and perform a normality test as well. This concludes that a non-parametric test will be implemented for the correlation analysis.

To visualize the data points for the correlation a scatter plot is drawn to have an understanding of the data separated. **Fig 18.**

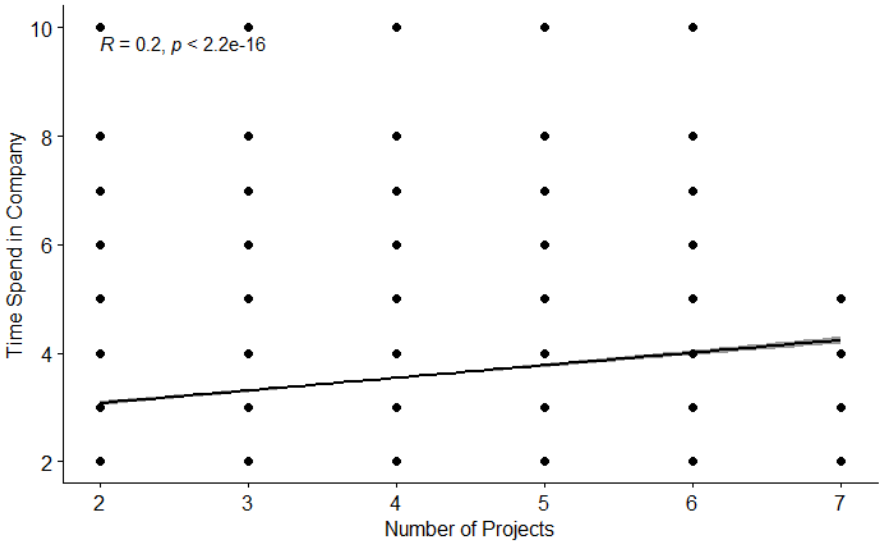

Fig 18: Scatter Plot for correlation variables

The correlation test is implemented as follows using the R package. **Fig 19.**

```
        Spearman's rank correlation rho

data:  hr$number_project and hr$time_spend_company
S = 4.2068e+11, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.251971
```

Fig 19: Spearman Correlation

In the result above, considerable values are spearman value *S* is 4.2068*10^11, the *p-value* is almost equal to zero and Spearman coefficient *rho* is 0.251971 which shows that there is a positive relationship but not as much strong among the variables.

3.5. Linear Regression

Regression models are frequently used in predictive analysis. The purpose of regression is to develop a mathematical equation that depicts the inter-relationship between the variables. Linear regression is widely used in the data science sector because it performs well when the target variable is quantitative or continuous, and it may take both qualitative and quantitative predictors to impact the linear model's conclusion. The main goal is to choose a line that fits the data the best. The best fit line is the one with the smallest overall prediction error (across all data points). The gap between the point and the regression line is called error.

In the linear equation, each input value or column is given a scaling factor, also known as a coefficient and indicated by Beta (B). Another coefficient, known as the intercept or bias coefficient, is added to give the line an extra degree of freedom.[1] Whereas "E" epsilon describes the random component of the linear relationship among X and Y. The comprehensive linear model can be denoted as shown below;

$$Y = a + \beta * X + \epsilon \qquad (1)$$

Where "a" is an intercept and "β" is the slope, under a collective name of regression coefficients, ϵ is known as the error term and "X" is the predictor variable. Collectively by computing this equation, they describe Y that is known as the target variable.

The Linear Regression technique is useful to find the relationship among the predictive and target variables. It also summarizes a whole model with the mathematical equation which takes in the predictive variable as an input and in response predict the target value. It is useful when dealing with continuous variables. Keeping in view its effectiveness on the particular set of data, it is used in this project for the sole purpose to predict the satisfaction level of the employee based on key predictive variables in the dataset.

A selected dataset of Human Resources is divided in a ratio of ¾. 75% of data is assigned for training while 25% is kept for testing the model. Afterwards, the Linear Regression model is trained on the training dataset using the key predictive variables which include number_project, salary, last_evaluation, time_spend_company, and Work_accident. After the training, useful parameters are observed using the summary of the linear model. A detailed summary is shown in **Fig 20.** The Linear model is also plotted against key parameters. **Fig 21.**

```
Call:
lm(formula = satisfaction_level ~ number_project + time_spend_company +
    salary + Work_accident + last_evaluation, data = HR_training)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63834 -0.18607  0.01932  0.19455  0.60152

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.644361   0.013383  48.149  < 2e-16 ***
number_project     -0.037271   0.001989 -18.741  < 2e-16 ***
time_spend_company -0.014319   0.001594  -8.985  < 2e-16 ***
salarylow          -0.039653   0.008516  -4.656 3.26e-06 ***
salarymedium       -0.017722   0.008597  -2.061   0.0393 *
Work_accident       0.043528   0.006414   6.786 1.21e-11 ***
last_evaluation     0.252203   0.014177  17.790  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2401 on 11242 degrees of freedom
Multiple R-squared:  0.05756,   Adjusted R-squared:  0.05706
F-statistic: 114.4 on 6 and 11242 DF,  p-value: < 2.2e-16
```

Fig 20: Linear Model Summary

The Output of the linear model gives some useful insights about the important factors like F-statistic, p-value Adjusted R-Squared and R-squared. The *R-squared value* of the linear model is **0.0575** while the *Adjusted R-Squared* is **0.0570**, which is above 0.05 and tells about the variance in the dependent variable which is demonstrated by the predictors. This value increases with an increase in the predictive variables but in this scenario, the calculated values are the best value representing the dependent variable.

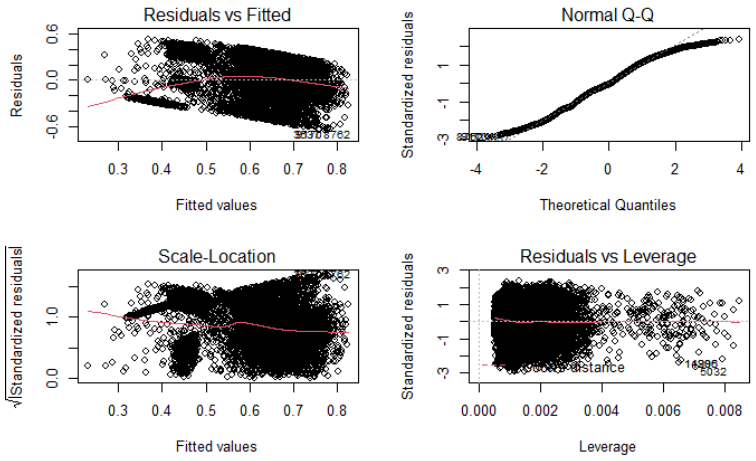

Fig 21: Linear Model Plot

The trained linear model is now undergoing the testing phase. In the testing phase, 25% of test data is used to predict the target variable i.e. satisfaction_level. The regression line on the training and testing dataset fits as shown in **Fig 22.**

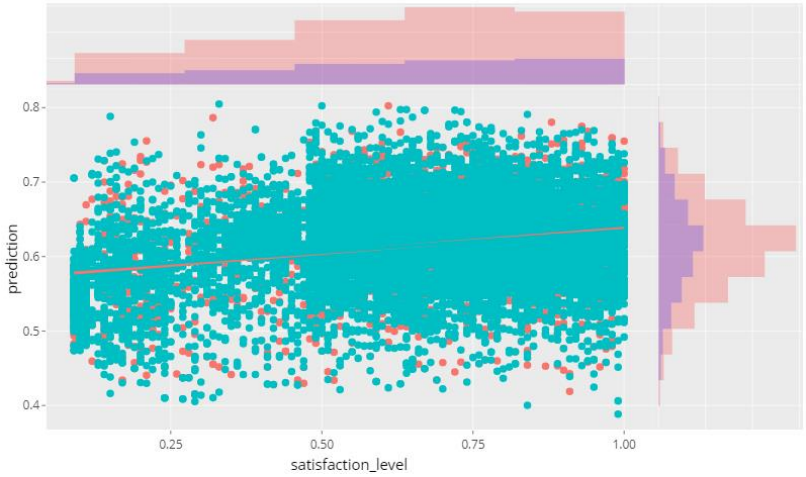

Fig 22: Linear Model

To check the difference in the training and testing, and accuracy test is performed to check the accuracy of the predicted model. An accuracy result is calculated using MAE. Calculated MAE indicates that the average absolute difference between the observed values and the predicted values is approximately 0.2 as shown in **Fig 23.**

```
[1] 0.205784
```

Fig 23: Mean Absolute Error

3.6. Logistic Regression

As a statistical approach, machine learning has incorporated logistic regression. It's a sophisticated statistical strategy for modelling a binomial outcome using one or more explanatory variables. It estimates probabilities using a logistic function, which is the cumulative logistic distribution, to quantify the connection between the categorical dependent variable and one or more independent variables.

In the context of the model implementation in this document, we consider "left" as a target variable and all others as predictive variables except department and work accident as they are not much related to the left variable as seen in the correlational plot. Moreover, the salary variable is converted to a dummy variable as salary_0, salary_1 and salary_2 corresponding to low, medium and high respectively. After that, a dataset is divided into two halves with a ratio of 8:2. 80% of data is used for training the model while 20% is for testing the model. **Fig 24** shows the training of the logistic model on the training dataset.

	Estimate	Std. Error	t value	Pr(>|t|)
satisfaction_level	-0.6587603939	1.446215e-02	-45.550665	0.000000e+00
last_evaluation	0.0925642591	2.266551e-02	4.083927	4.456870e-05
number_project	-0.0353490138	3.296096e-03	-10.724511	1.032118e-26
average_montly_hours	0.0006283654	7.903627e-05	7.950343	2.026222e-15
time_spend_company	0.0359188298	2.450352e-03	14.658643	3.098454e-48
promotion_last_5years	-0.1262921114	2.462190e-02	-5.129259	2.954252e-07
salary_0	0.5150207116	2.092876e-02	24.608277	1.722938e-130
salary_1	0.4284465899	2.122426e-02	20.186647	3.852667e-89
salary_2	0.2979233185	2.388676e-02	12.472320	1.754324e-35

Fig 24: Logistic Regression Model Training

After training the model, probabilities and prediction is made which result as shown in the figures below.

5	10	24	25	39	46
0.4563813	0.4136868	0.3891538	0.4369158	0.6488003	0.1821503

Fig 25: Probabilities

For prediction, a threshold of 0.5 is set. The values above the threshold are grouped in categories of an employee who doesn't leave the company while below the threshold are the ones who left the company.

5	10	24	25	39	46
0	0	0	0	1	0

Fig 26: Predictions

Finally, the last step is to get the accuracy of the model. The accuracy of our model is about 77%. **Fig 27.**

```
[1] 0.7732578
```

Fig 27: Accuracy of the model is 77%

4 Statistical Techniques

1. Brownlee Jason: Machine Learning Mastery, https://machinelearningmastery.com/linear-regression-for-machine-learning/, [Last accessed 2022/01/10]
2. Pooja Sengupta: Application of Statistics in Human Resource Management, https://www.researchgate.net/publication/322819702_Application_of_Statistics_in_Human_Resource_Management, [Last Accessed 2022/01/15]
3. Scott Mondore, Douthitt and Marisa Carson: Maximizing the impact and effectiveness of HR Analytics to Drive Business Outcomes, Strategic Management Decisions, http://datascienceassn.org/sites/default/files/Maximizing%20the%20Impact%20and%20%20Effectiveness%20of%20HR%20Analytics%20%20to%C2%A0Drive%20Business%20Outcomes.pdf, [Last Accessed 2022/01/16]