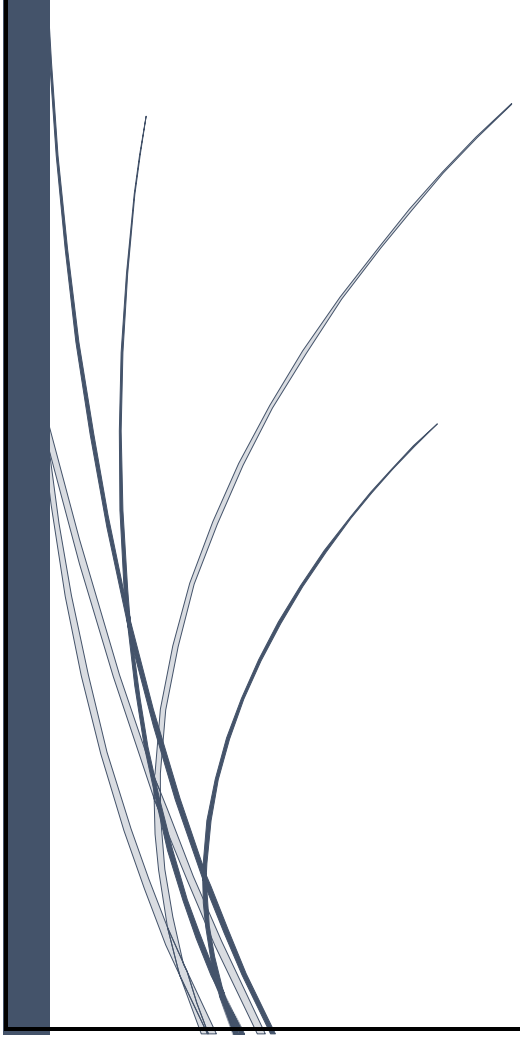
EMPLOYMENT SERVICES



CHAPTER-1
INTRODUCTION

1.1 Scope of analysis

- The dataset provides valuable insights into employee performance, satisfaction, and workplace dynamics, offering opportunities for targeted analysis to enhance organizational efficiency. One of the primary focuses could be employee retention (attrition analysis), identifying factors such as low satisfaction levels, high workloads, lack of promotions, and insufficient salaries that may contribute to employees leaving the organization. Understanding these factors can help in designing strategies to improve retention and employee satisfaction.

- Workplace safety analysis could focus on the impact of accidents on satisfaction and attrition rates, identifying departments with higher risks and implementing safety measures to foster a secure working environment. Lastly, the dataset enables the exploration of time spent in the company, helping to understand how employee tenure correlates with satisfaction, performance, and attrition, and whether long-tenured employees feel adequately rewarded.

1.2 Approach analysis

- This dataset involves a systematic process to extract actionable insights. The first step is data exploration and Pre Processing, where the dataset is examined for missing values, outliers, and data inconsistencies. Categorical variables, such as department and salary range, are encoded, and numerical variables, like satisfaction level and average monthly hours, are normalized if necessary. This ensures the data is clean and ready for analysis.

- The employee satisfaction dataset contains labelled data, making it a supervised learning problem. Since the goal is to predict satisfaction levels based on various independent variables such as salary, promotions, working hours, performance evaluations, tenure, and workload, the problem is best suited for regression analysis**.** Given this, the linear regression model was chosen as the primary approach for analysis.

CHAPTER-2

DATA UNDERSTANDING

2.1 Gathering Data

Load the relevant Packages

```{r}
Library (tidyverse)
Library (ggplot2)
```

Load the dataset

```{r}
#import the dataset
Project = read.csv(("C:/Users/lenovo/Downloads/employe data.csv"))
Projects = Data.frame(project)
projects
```

Structure of the data

```{r}
# Structure of data
Str(projects)
```

data.frame': 15787 obs. of 11 variables:

$ Emp.ID : Int 1 2 3 4 5 6 7 8 9 10 ...

$ Satisfaction_level : num 0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...

$ last_evaluation : num 0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...

$ number_project : int 2 5 7 5 2 2 6 5 5 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 ...

$ Work_accident : int 0 0 0 0 0 0 0 0 0 0 ...

$ promotion_last_5years: int 0 0 0 0 0 0 0 0 0 0 ...

$ dept : chr "sales" "sales" "sales" "sales" ...

$ salary : int 15000 20000 20000 15000 15000 15000 15000 15000 15000 15000 ...

$ range_salary : chr "low" "medium" "medium" "low" ...

2.2 Data Description

The Employee dataset has 15,787 rows and 11 columns

	Emp.ID	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years	dept
1	1	0.38	0.53	2	157	3	0	0	sales
2	2	0.80	0.86	5	262	6	0	0	sales
3	3	0.11	0.88	7	272	4	0	0	sales
4	4	0.72	0.87	5	223	5	0	0	sales
5	5	0.37	0.52	2	159	3	0	0	sales
6	6	0.41	0.50	2	153	3	0	0	sales
7	7	0.10	0.77	6	247	4	0	0	sales
8	8	0.92	0.85	5	259	5	0	0	sales
9	9	0.89	1.00	5	224	5	0	0	sales
10	10	0.42	0.53	2	142	3	0	0	sales
11	11	0.45	0.54	2	135	3	0	0	sales
12	12	0.11	0.81	6	305	4	0	0	sales
13	13	0.84	0.92	4	234	5	0	0	sales
14	14	0.41	0.55	2	148	3	0	0	sales
15	15	0.36	0.56	2	137	3	0	0	sales
16	16	0.38	0.54	2	143	3	0	0	sales
17	17	0.45	0.47	2	160	3	0	0	sales
18	18	0.78	0.99	4	255	6	0	0	sales
19	19	0.45	0.51	2	160	3	1	1	sales

The dataset contains the following variables:

EMP_ID

A unique identifier for each employee

SATISFACTION_LEVEL

A critical variable for understanding employee happiness and its correlation with performance, workload, and attrition.

LAST_EVALUATION

Useful to analysis how performance evaluations relate to factors such as satisfaction, promotion, and attrition.

NUMBER_PROJECTS

Indicates workload and its potential relationship with performance, satisfaction, and attrition.

AVERAGE _MONTHLY_INCOME

Provides insights into work-life balance and its impact on employee satisfaction and productivity.

TIME_SPEND_COMPANCY

Often linked with loyalty, likelihood of promotion, and risk of attrition

WORK_ACCIDENT

Helps analyse workplace safety and its influence on employee satisfaction and retention.

PROMOTION_LAST_5YEARS

Correlates promotions with satisfaction, performance, and retention.

DEPT

Helps analyse departmental trends, such as satisfaction or attrition rates, to identify specific challenges or opportunities.

SALARY

Directly impacts satisfaction and retention and its often analysed to determine fair compensation practices.

RANGE_SALARY

Provides an easier way to group employees and study salary trend relation to attrition and satisfaction

2.3 Data Understanding

This module explains data understanding. This dataset consist of different columns. Each and every columns we should find the summary () function. This function is used to calculate the average value and determine the maximum, minimum of the column in a data frame

```{r}
summary(projects)
```

```
     Emp.ID      satisfaction_level last_evaluation  number_project
 Min.   :    1   Min.   :0.0900     Min.   :0.3600   Min.   :2.000
 1st Qu.: 3750   1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000
 Median : 7500   Median :0.6400     Median :0.7200   Median :4.000
 Mean   : 7500   Mean   :0.6128     Mean   :0.7161   Mean   :3.803
 3rd Qu.:11250   3rd Qu.:0.8200     3rd Qu.:0.8700   3rd Qu.:5.000
 Max.   :14999   Max.   :1.0000     Max.   :1.0000   Max.   :7.000
 NA's   :788     NA's   :788        NA's   :788      NA's   :788
 average_montly_hours time_spend_company Work_accident    promotion_last_5years
 Min.   : 96.0        Min.   : 2.000     Min.   :0.0000   Min.   :0.0000
 1st Qu.:156.0        1st Qu.: 3.000     1st Qu.:0.0000   1st Qu.:0.0000
 Median :200.0        Median : 3.000     Median :0.0000   Median :0.0000
 Mean   :201.1        Mean   : 3.498     Mean   :0.1446   Mean   :0.0213
 3rd Qu.:245.0        3rd Qu.: 4.000     3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :310.0        Max.   :10.000     Max.   :1.0000   Max.   :1.0000
 NA's   :788          NA's   :788        NA's   :788      NA's   :788
     dept               salary      range_salary
 Length:15787       Min.   :15000   Length:15787
 Class :character   1st Qu.:15000   Class :character
 Mode  :character   Median :20000   Mode  :character
                    Mean   :18741
                    3rd Qu.:20000
                    Max.   :35000
                    NA's   :3004
```

2.3 Continues variables

- Emp.ID
- Satisfaction level
- Last evaluation
- Number projects
- Average monthly hours
- Time spend company
- Promotion last 5years
- Salary

Categorical variables

- Department
- Salary in range
- Work accident

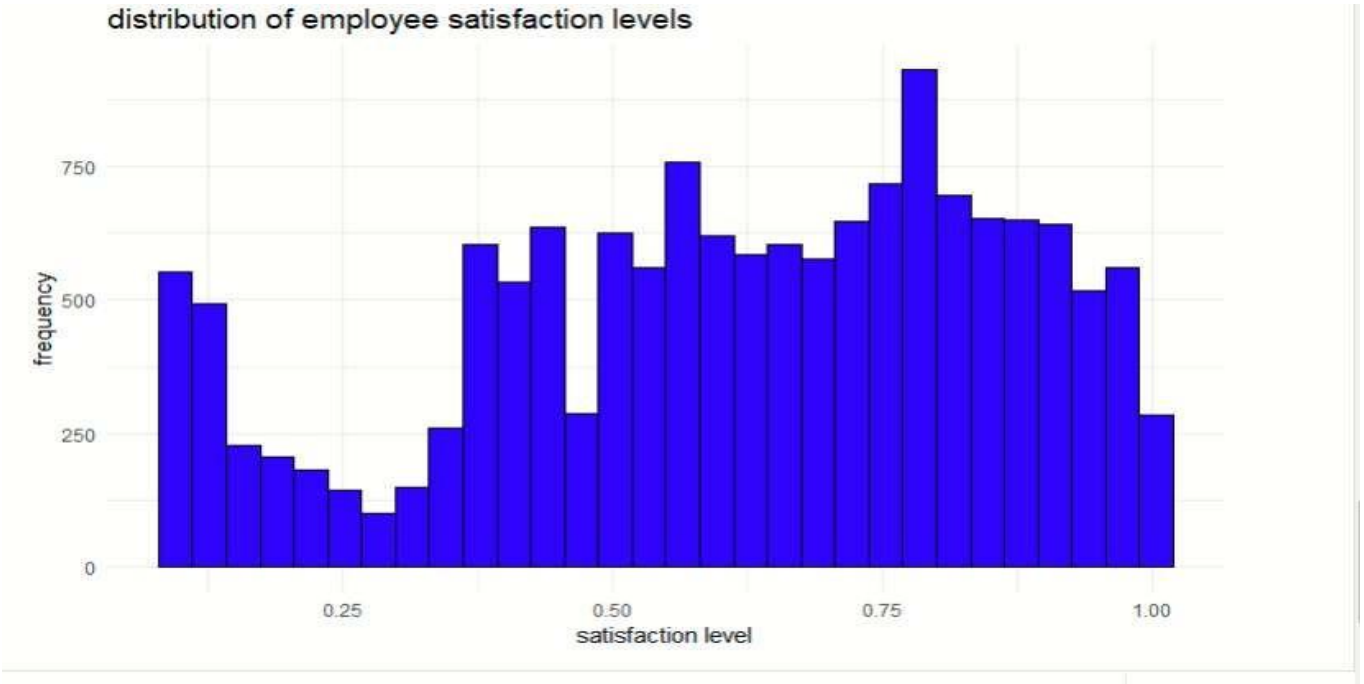
CHAPTER – III

DATA PREPROCESSING

3.1 Handle Missing values

Check for Missing values

```
```{r}
#handle missing values
# check missing values
colSums(is.na(project))
```
```

```
                Emp.ID    satisfaction_level       last_evaluation        number_project
                   788                   788                   788                   788
 average_montly_hours    time_spend_company         Work_accident promotion_last_5years
                   788                   788                   788                   788
                  dept                salary          range_salary
                     0                  3004                     0
```

Remove the NA values in the dataset

```
```{r}
# check there is NA value in Data set
colSums(sapply(projects,is.na))
```
```

```
                Emp.ID    satisfaction_level       last_evaluation        number_project
                     0                     0                     0                     0
 average_montly_hours    time_spend_company         Work_accident promotion_last_5years
                     0                     0                     0                     0
                  dept                salary          range_salary
                     0                     0                     0
```

CHAPTER – III

DATA EXPLORATION

3.2 Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is a crucial step to understand the structure of the data, identify patterns, detect anomalies, and extract meaningful insights. The first step involves loading the dataset and viewing its structure. This includes examining the first few rows, checking the dimensions, and generating summary statistics. Missing values are also identified and handled.

- For instance, rows with missing values can be removed, or missing values in numerical columns can be replaced with the mean or median. Additionally, categorical variables, such as department and salary range, are converted into factors for better analysis**.**

3.3 Employee Satisfaction level analysis

The histogram represents the distribution of employee satisfaction levels within a company. The x-axis denotes satisfaction levels ranging from 0 to 1, while the y-axis indicates the frequency of employees at each level. The distribution appears uneven, with noticeable peaks around 0.5 and 0.75, suggesting that a significant portion of employees fall within these satisfaction ranges. A considerable number of employees exhibit high satisfaction levels closer to 1.0, indicating overall positive sentiment. However, there are some dips, revealing lower representation at certain satisfaction levels. This visualization provides valuable insights into employee morale, helping the company identify areas for potential improvement in workplace satisfaction.

```
# visualize satisfaction levels
```{r}
ggplot(projects,aes(x=satisfaction_level))+geom_histogram(bins=30,fill="blue",color="black")+ggtitle("distr
ibution of employee satisfaction levels")+xlab("satisfaction level")+ylab("frequency")
```
```

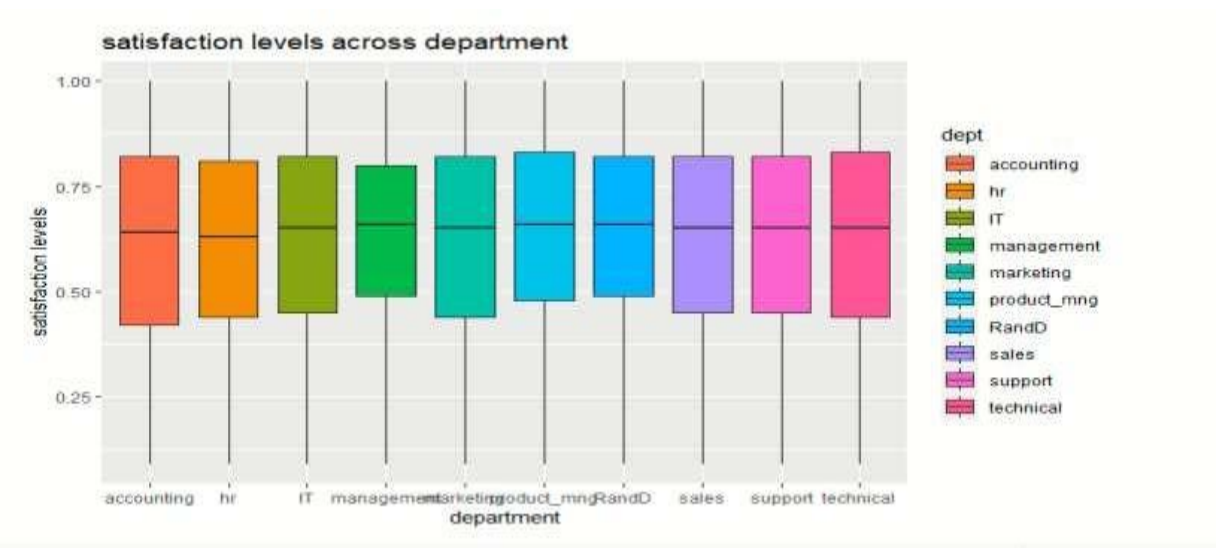

distribution of employee satisfaction levels
frequency
750
500
250
0
0.25
0.50
0.75
1.00
satisfaction level

3.4 Department-wise Employee Satisfaction Analysis

The satisfaction levels of employees across different departments within a company. The x-axis categorizes various departments such as accounting, HR, IT, management, marketing, product management, R&D, sales, support, and technical, while the y-axis measures employee satisfaction levels ranging from 0 to 1. Each department is represented by a distinct colour, allowing for easy comparison.

The distribution of satisfaction levels appears fairly consistent across departments, with median satisfaction levels being similar. However, there is noticeable variation in the spread of satisfaction levels within each department, as indicated by the height of the boxes and the presence of outliers. Departments such as R&D, technical, and product management seem to have a broader range of satisfaction levels, indicating that while some employees are highly satisfied, others may have lower morale.

Overall, this visualization provides a comparative analysis of employee satisfaction across departments, helping management identify trends and potential areas for improvement in specific teams. Understanding these differences can assist in tailoring employee engagement strategies and enhancing workplace satisfaction.

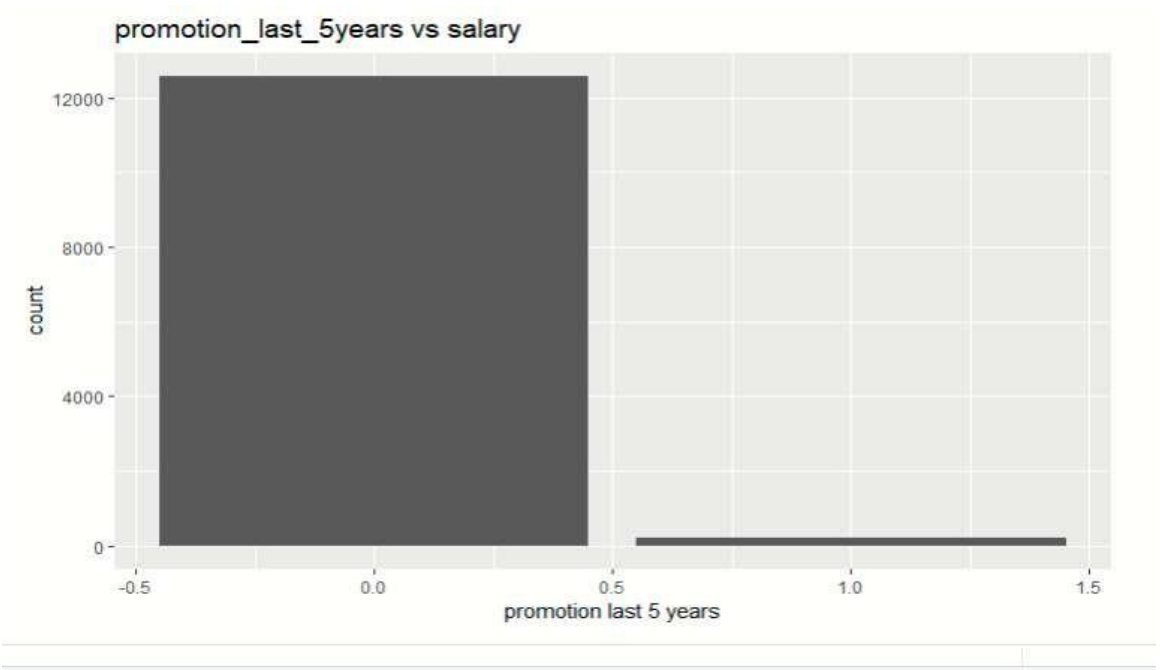


3.5 Employee Promotions in the Last Five Years, Distribution Analysis

The number of employees who received promotions in the last five years within the company. The x-axis represents whether an employee was promoted (typically coded as 0 for no promotion and 1 for promotion), while the y-axis represents the count of employees in each category. From the visualization, it is evident that the vast majority of employees did not receive a promotion in the past five years, as indicated by the significantly larger bar at 0. In contrast, only a small number of employees were promoted, as seen in the much smaller bar at 1.

This data suggests that promotions within the company have been limited over the last five years, which may impact employee motivation and career growth. If promotions are based on performance, skill development, or tenure, the company may need to assess its promotion policies to ensure fairness and career progression opportunities for employees. Identifying factors that influence promotions can help in improving employee satisfaction and retention.

```{r}
ggplot(train_data,aes(x=dept,y=satisfaction_level,fill=dept))+geom_boxplot()+ggtitle("satisfaction levels across department")+xlab("department")+ylab("satisfaction levels")+theme_minimal()+theme(axis.text.x = element_text(angle = 45,hjust = 1))+scale_fill_brewer(palette = "set3")
```



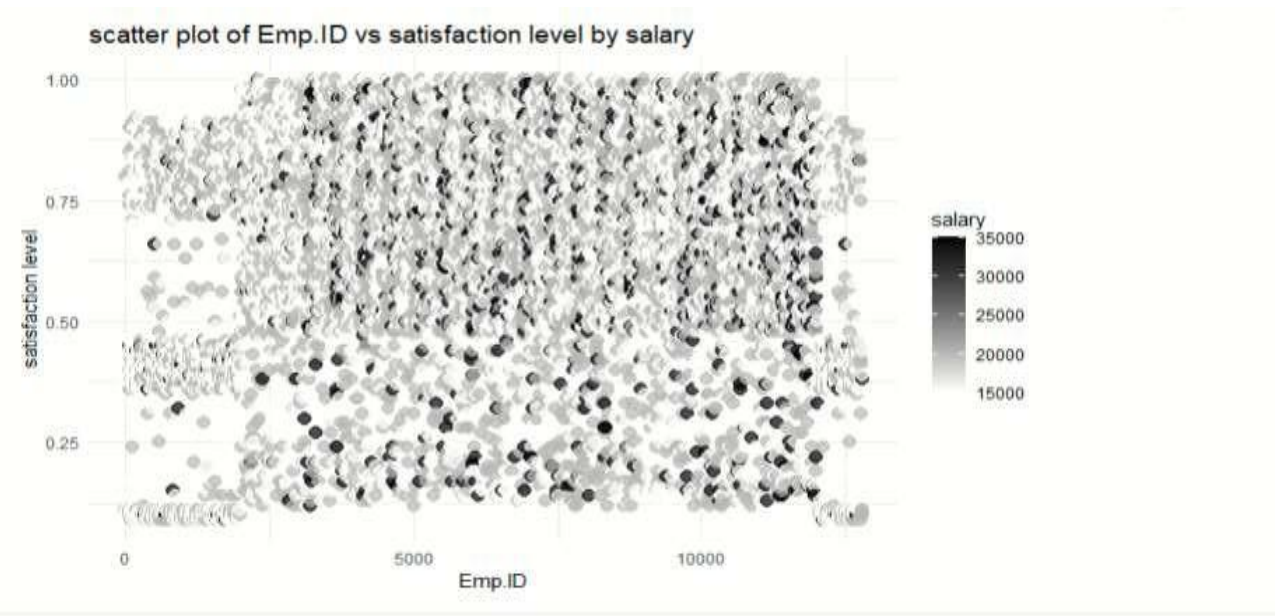
3.6 Analysis of Employee Satisfaction Based on Salary and a Scatter Plot Representation of Employee ID vs Satisfaction Level

The relationship between Employee ID and satisfaction level, categorized by salary within a company. The x-axis indicates the Employee ID, while the y-axis represents the satisfaction level, which ranges from 0 to 1. The colour intensity of the data points varies, representing different salary levels, with darker shades indicating higher salaries and lighter shades representing lower salaries.

From the plot, it appears that employee satisfaction levels are distributed across a wide range, with no clear pattern based on Employee ID alone. However, there are clusters of employees with similar satisfaction levels at different salary levels. The density of points suggests that there may be certain trends in employee satisfaction, but it requires further analysis to determine specific patterns.

Overall, this visualization helps in understanding whether salary influences employee satisfaction and if there are any underlying trends that may need attention for workforce management.

```{r}
ggplot(projects,aes(x=Emp.ID,y=satisfaction_level,color=salary))+geom_point(size=3,alpha=0.7)+scale_color_gradient(low="white",high="black")+labs(title = "scatter plot of Emp.ID vs satisfaction level by salary",x="Emp.ID",y="satisfaction level",color="salary")+theme_minimal()
```



CHAPTER IV

BUSINESS INTELLIGENCE INTERACTIVE DASHBOARDS

4.1 Dash Board Interpretation

- A dashboard is an essential tool for visualizing and analysing data in a structured and interactive manner. It provides users with a consolidated view of key metrics, trends, and performance indicators, allowing them to make informed decisions efficiently.
- One of the primary advantages of a dashboard is its ability to present real-time data in a user-friendly format.
- Instead of navigating through multiple spreadsheets or databases, users can access critical insights at a glance. Through data visualization techniques such as charts, graphs, tables, and KPI indicators, dashboards make it easier to interpret complex datasets and identify patterns or anomalies.

4.2 EMPLOYEMENT SERVICES

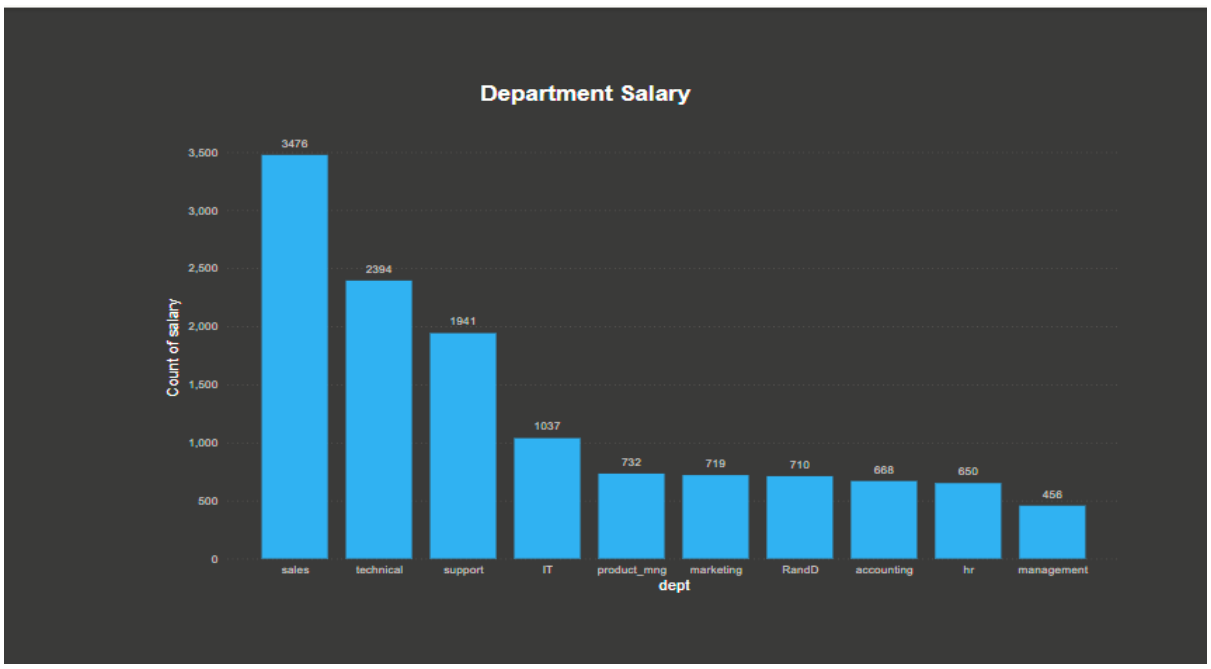
- This Employment Services Dashboard provides a comprehensive overview of key employee-related metrics, offering valuable insights into workforce performance, salaries, satisfaction levels, and promotions

- At the top, the dashboard highlights key performance indicators (KPIs), including the total number of employees (15K), projects undertaken (57K), average company time spent (3.5 years), last evaluation score (10.74K), and average salary (12.78K).

- These metrics help in assessing workforce engagement and business performance at a glance.

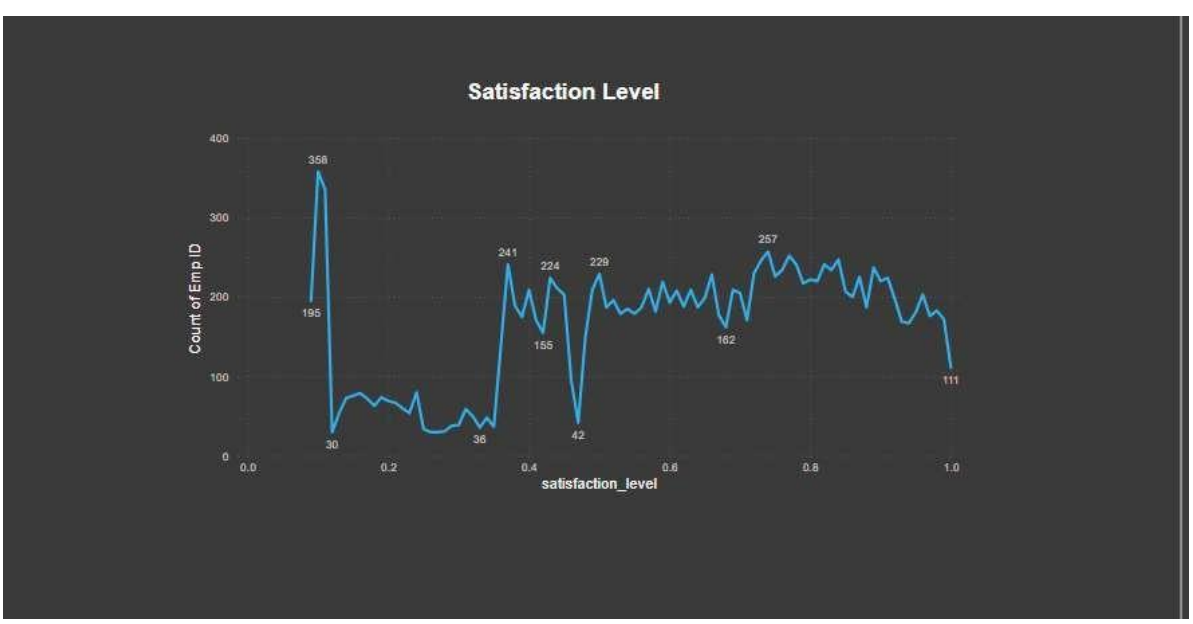
4.3 Department salary

- The Department Salary section visualizes salary distribution across various departments, showing that sales, technical, and support roles have higher salary counts, while management and IT have relatively lower salary distributions.



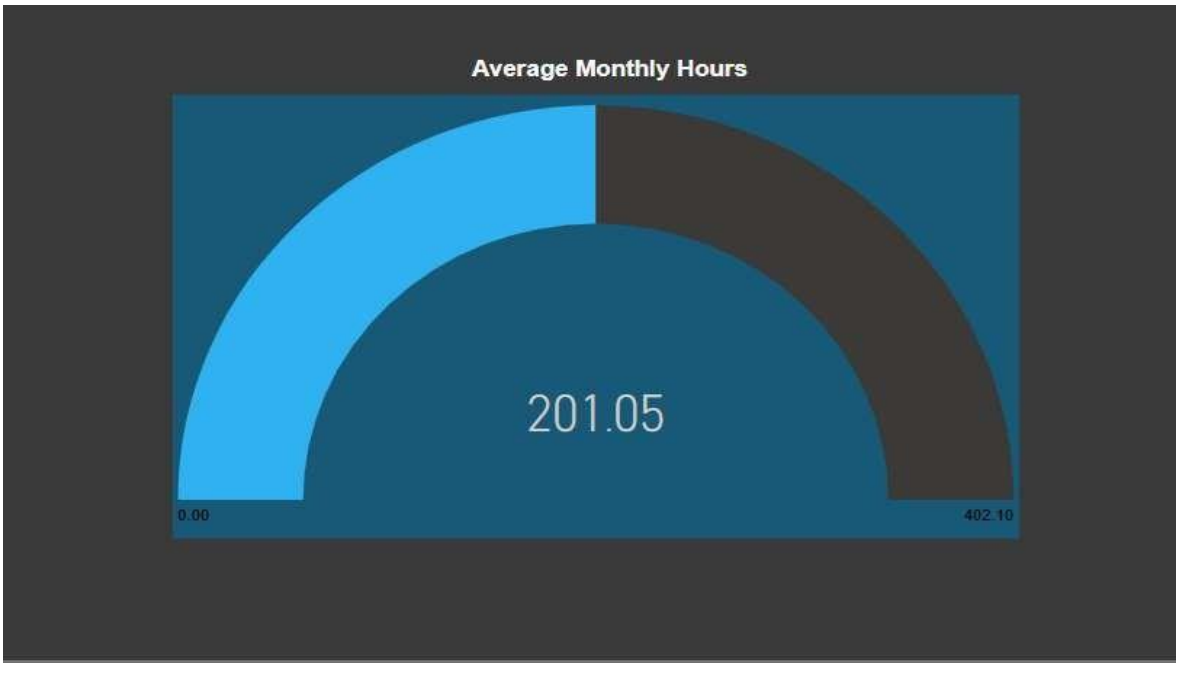
4.4 Satisfaction Level

- The Satisfaction Level graph provides an overview of employee contentment, showing fluctuations in satisfaction scores across different employee groups.



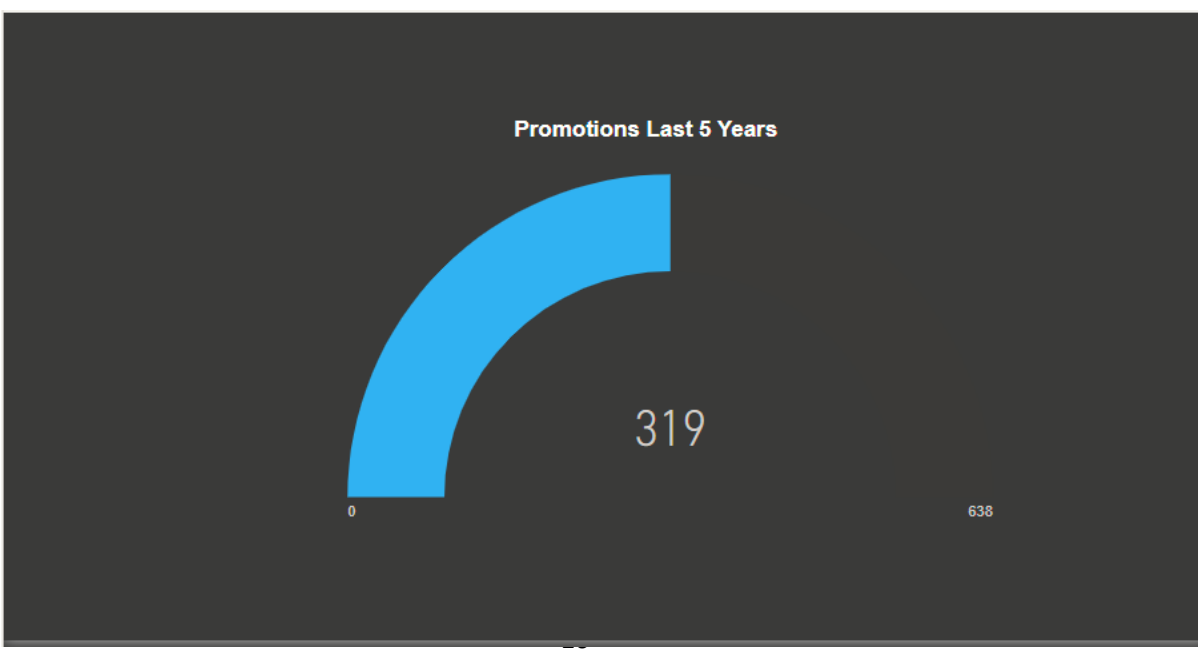
4.5 Average Monthly Hours

- The Average Monthly Hours section indicates that employees work an average of 201.05 hours per month, which helps HR and management assess workload balance.



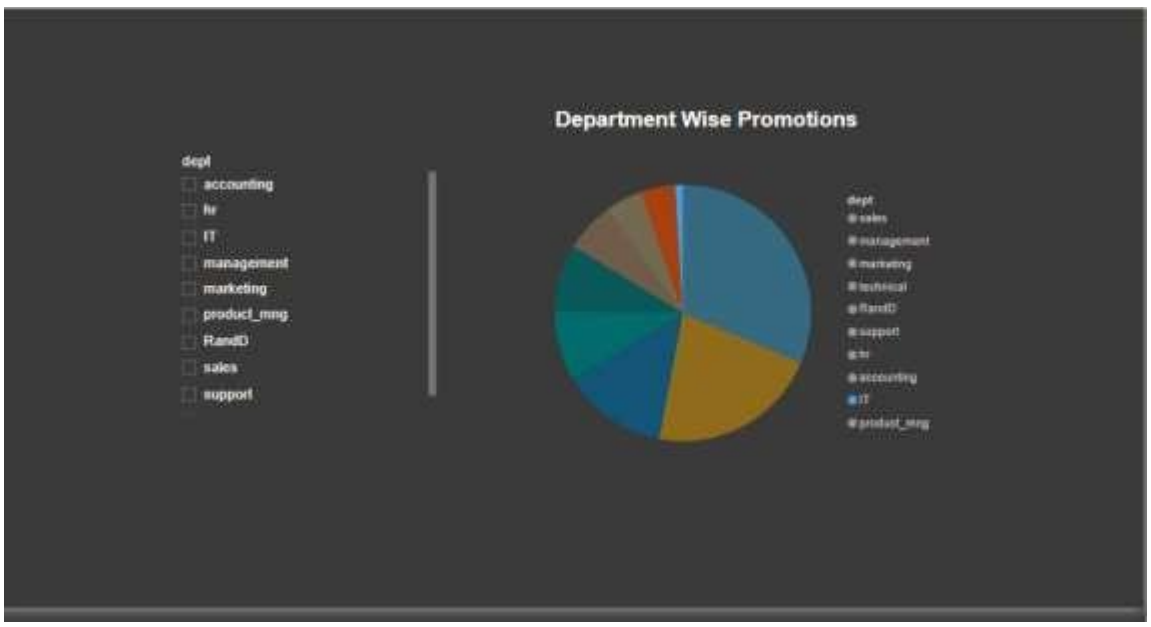
4.6 Promotions in the Last 5 Years

Another key section focuses on Promotions in the Last 5 Years, which highlights that only 319 employees received promotions, suggesting a relatively low promotion rate that may impact employee satisfaction and retention.



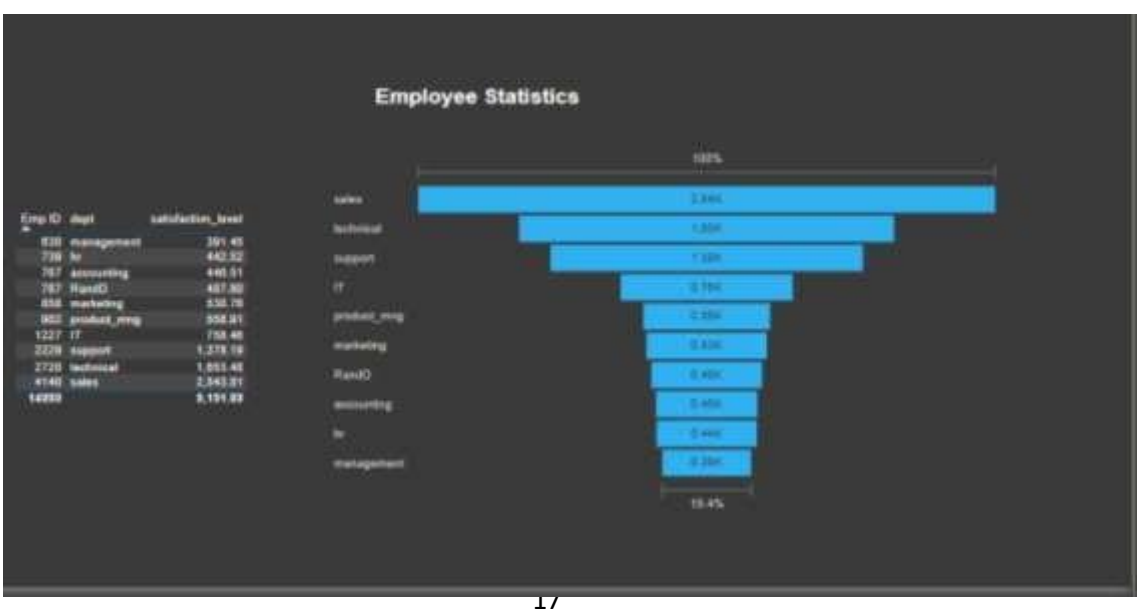
4.7 Department-wise Promotions

The Department-wise Promotions pie chart shows how promotions are distributed across departments, providing insights into career growth opportunities within the company.



4.8 Employee Statistics

- The Employee Statistics table lists specific employees along with their departments and satisfaction levels, enabling managers to identify trends in job satisfaction and performance across different teams.



4.9 Overview

- This dashboard serves as a powerful HR and mzanagement tool, helping decision-makers track employee satisfaction, salary distribution, promotions, and workload while identifying potential areas for improvement in employee engagement and retention strategies.

CHAPTER V

MODEL BUILDING

5.1 Algorithms for model building

Machine learning (ML) algorithms are computational methods that enable computers to learn patterns from data and make predictions or decisions without being explicitly programmed. These algorithms can be broadly categorized into three main types.

Supervised Learning

In supervised learning, the algorithm learns from labelled data, where each input has a corresponding correct output. The goal is to map inputs to outputs based on training examples.

Common Supervised Learning Algorithms:

- **Linear Regression** – Used for predicting continuous values (e.g., house prices).
- **Logistic Regression** – Used for binary classification problems (e.g., spam detection).
- **Decision Trees** – A tree-like model used for classification and regression.
- **Random Forest** – An ensemble method combining multiple decision trees.
- **Support Vector Machines (SVM)** – Finds the optimal boundary between classes.
- **Neural Networks** – Used in deep learning for complex pattern recognition.

Linear Regression

Linear regression is used for predicting continuous values. It establishes a relationship between input variables (**X**) and an output variable (**Y**) by fitting a straight line.

Use Cases:

- Predicting house prices based on features (size, location, etc.).
- Estimating sales revenue based on advertising spending.

Mathematical Representation:

The equation for simple linear regression is:

$Y=mX+bY = mX + bY=mX+b$

where:

- **Y** is the predicted output (dependent variable).
- **X** is the input (independent variable).
- **m** is the slope (coefficient).
- **b** is the intercept.

Logistic Regression

Logistic regression is used for binary classification problems, where the output is either 0 or 1 (e.g., spam vs. not spam). Instead of a straight line, it uses the sigmoid function to map input values to probabilities.

Mathematical Representation:

The sigmoid function is given by:

$P(Y=1)=1+e−(b0+b1X1+b2X2+...+bnXn)1$

The output is interpreted as a probability. If the probability is greater than 0.5, we classify it as 1; otherwise, it's 0.

Use Cases:

- Spam email detection (Spam/Not Spam).
- Fraud detection in banking transactions.

Decision Tree

A decision tree is a flowchart-like structure where each internal node represents a test on a feature, each branch represents the outcome, and each leaf node represents a decision/classification.

How It Works:

- The dataset is split based on a feature that maximizes information gain or minimizes entropy (Gini impurity).
- Splitting continues recursively until stopping criteria (e.g., max depth) is met.

Use Cases:

- Customer churn prediction.
- Medical diagnosis (e.g., is a patient at risk for a disease?).

Random Forest

A random forest is an ensemble of multiple decision trees. It improves accuracy and reduces overfitting by averaging multiple tree predictions.

How It Works:

- Multiple decision trees are trained on random subsets of data.
- Each tree makes a prediction, and the final output is determined by majority voting (classification) or averaging (regression).

Use Cases:

- Credit scoring.
- Image classification.

Support Vector Machine (SVM) (For Classification & Regression)

SVM finds the optimal hyperplane that best separates different classes. In cases where data is not linearly separable, SVM uses the kernel trick to map data into higher dimensions.

Mathematical Representation:

The decision boundary is given by:

w·x+b=0w \cdot x + b = 0w·x+b=0

where **w** is the weight vector and **b** is the bias term.

Use Cases:

- Handwriting recognition.
- Face detection.

Neural Networks (For Deep Learning)

Neural networks mimic the structure of the human brain, consisting **of** layers of interconnected neurons. They are especially powerful in handling complex patterns and large datasets.

Structure:

- **Input Layer** – Takes input features.
- **Hidden Layers** – Perform computations using weights and activation functions.
- **Output Layer** – Produces the final prediction.

Use Cases:

- Speech recognition.
- Image classification (e.g., recognizing faces in photos).

Unsupervised Learning

In unsupervised learning, the algorithm learns patterns from data without labelled outputs. The goal is to discover hidden structures or relationships in the data.

Common Unsupervised Learning Algorithms:

- **K-Means Clustering** – Groups data into clusters based on similarity.
- **Hierarchical Clustering** – Builds a hierarchy of clusters.
- **Principal Component Analysis (PCA)** – Reduces dimensionality while preserving important features.

K-Means Clustering

- Divides the dataset into K clusters by assigning each data point to the closest cluster center(centroid).
- The centroids are updated iteratively until the best grouping is achieved.

Use Cases

- Customer segmentation (e.g., grouping similar customers in marketing).
- Image segmentation (e.g., separating objects in an image).

Hierarchical Clustering

- Creates a hierarchy of clusters using a tree-like structure (den drogram).

Two approaches:

- **Agglomerative (Bottom-Up)** – Each data point starts as its own cluster and merges upward.
- **Divisive (Top-Down)** – All data points start in one cluster and split downwards.

Use Cases

- Organizing documents based on topic similarity.
- Bioinformatics (e.g., grouping genes with similar expression patterns).

Principal Component Analysis (PCA)

Transforms high-dimensional data into fewer dimensions while maintaining variance. Uses eigenvectors and eigenvalues to determine principal components.

Use Cases

- Reducing features in image processing (e.g., facial recognition).
- Speeding up machine learning models.

Reinforcement Learning

In reinforcement learning, an agent interacts with an environment and learns by receiving rewards or penalties based on its actions. The goal is to maximize cumulative rewards over time.

Common Reinforcement Learning Algorithms:

- **Q-Learning** – A value-based method that uses a Q-table to learn optimal actions.
- **Deep Q Networks (DQN)** – Uses deep learning to improve Q-learning.
- **Policy Gradient Methods** – Learn policies directly, often used in robotics and gaming.

Q-Learning (Off-Policy)

- Learns the optimal action-value function (Q(s,a)Q(s, a)Q(s,a)).
- Uses the Bellman Equation to update Q-values

Use Cases

- Game-playing AI (e.g., AlphaGo).
- Robotics (e.g., autonomous navigation).

Deep Q-Networks (DQN)

- Uses neural networks to approximate Q-values, improving on traditional Q-learning.
- Introduces experience replay (storing past experiences for better learning).

Use Cases

- Atari game AI.
- Autonomous driving.

Policy Gradient (PG)

Use Cases

- Robotics control.
- Continuous action problems (e.g., robotic arm movement).

5.2 Choosing the model

Choosing a model to use is very essential. You must consider the input and output of your data.

For this data:

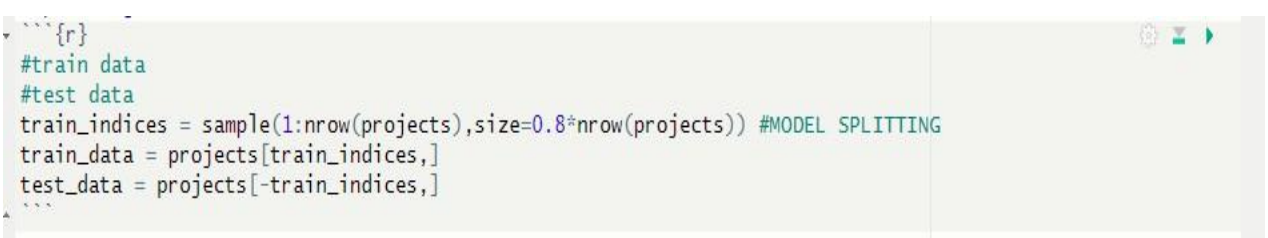
- Linear Regression is chosen for this problem because it is a simple yet powerful algorithm for modelling relationships between variables, especially when dealing with continuous numerical data. Since the dataset consists of labelled data with known employee satisfaction scores, this is a supervised learning problem where the goal is to predict a numerical output

- Linear Regression is ideal because it establishes a direct mathematical relationship between independent variables (such as salary, work-life balance, and benefits) and the dependent variable (employee satisfaction score).

Divide the dataset into training and testing data

- Training data is the portion of the dataset used to teach the machine learning model. It contains input features along with their corresponding output labels, allowing the model to learn patterns and relationships. The model adjusts its parameters based on this data to minimize errors and improve accuracy.

- Testing data is a separate portion of the dataset used to evaluate the trained model. It consists of unseen data that helps measure how well the model generalizes to new inputs. By comparing predicted outputs with actual values, testing data ensures that the model performs accurately and reliably in real-world scenarios.



5.3 Linear Regression

- Linear Regression is a fundamental statistical method used in machine learning and data analysis to model the relationship between a dependent variable and one or more independent variables.

- It assumes that the dependent variable changes linearly with respect to the independent variables. The goal of linear regression is to find the best-fitting line that minimizes the difference between actual and predicted values.

- This model works by estimating coefficients that define the relationship between variables, ensuring that the sum of squared differences between actual and predicted values is minimized.

- Linear Regression operates under key assumptions, including linearity, independence of observations, constant variance of errors (homoscedasticity), and normally distributed residuals

- It is widely used in various applications such as trend analysis, forecasting, financial modelling, and employee satisfaction prediction.

Train data

```r
train_data
```

	Emp.ID <int>	satisfaction_level <dbl>	last_evaluation <dbl>	number_project <int>	average_montly_hours <int>
10801	10801	0.48	0.53	3	211
12261	12261	0.46	0.46	2	154
2369	2369	0.72	0.88	2	247
5273	5273	0.80	0.76	3	270
9290	9290	0.83	0.57	3	135
1252	1252	0.43	0.57	2	151
8826	8826	0.36	0.97	5	151
10289	10289	0.88	0.59	4	227
12103	12103	0.42	0.46	2	150
356	356	0.41	0.57	2	136

1-10 of 10,226 rows | 1-6 of 11 columns Previous 1 2 3 4 5 6 ... 100 Next

Test data

```r
test_data
```

Description: df [3,835 × 11]

	Emp.ID <int>	satisfaction_level <dbl>	last_evaluation <dbl>	number_project <int>	average_montly_hours <int>
2	2	0.80	0.86	5	262
5	5	0.37	0.52	2	159
9	9	0.89	1.00	5	224
14	14	0.41	0.55	2	148
21	21	0.11	0.83	6	282
24	24	0.46	0.57	2	139
27	27	0.82	0.87	4	239
33	33	0.40	0.51	2	145
35	35	0.84	0.87	4	246
60	60	0.85	1.00	4	225

5.4 Train Data Accuracy

```r
model=lm(satisfaction_level~Emp.ID+salary+promotion_last_5years+average_montly_hours+last_evaluation+time_spend_company
+number_project,data=test_data)
summary(model)
```

```
Call:
lm(formula = satisfaction_level ~ Emp.ID + salary + promotion_last_5years +
    average_montly_hours + last_evaluation + time_spend_company +
    Work_accident + number_project, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6752 -0.1710  0.0146  0.1944  0.6099

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.597e-01  1.578e-02  35.479  < 2e-16 ***
Emp.ID                7.664e-06  6.408e-07  11.960  < 2e-16 ***
salary                1.631e-06  4.438e-07   3.675 0.000239 ***
promotion_last_5years 3.018e-02  1.819e-02   1.659 0.097078 .
average_montly_hours  7.155e-05  5.286e-05   1.354 0.175885
last_evaluation       2.448e-01  1.483e-02  16.515  < 2e-16 ***
time_spend_company   -2.429e-02  1.827e-03 -13.292  < 2e-16 ***
Work_accident         3.465e-02  6.614e-03   5.239 1.65e-07 ***
number_project       -3.491e-02  2.199e-03 -15.878  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2368 on 10217 degrees of freedom
Multiple R-squared:  0.07535,   Adjusted R-squared:  0.07463
F-statistic: 104.1 on 8 and 10217 DF,  p-value: < 2.2e-16
```

From the model we have learnt,

- **Intercept (0.59597)** - Represents the baseline satisfaction level when all variables are zero.
- **Employee ID (7.664e-06)** -Has statistical significance but should not impact satisfaction.
- **Salary (1.631e-06)** - Positive but very small impact on satisfaction
- **Promotion in the Last 5 Years (0.03018)** - Company should review promotion policies to ensure fairness and impact
- **Average Monthly Hours (0.2448)** - Strong positive correlation with satisfaction, Suggests employees who work more hours tend to be more satisfied.
- **Last Evaluation Score (0.2429)** - employees with higher performance evaluations are more satisfied.
- **Time Spent in Company (0.03465)** - could indicate strong company culture and job security.
- **Work Accident (−0.03419)**- Employees who experience workplace accidents report lower satisfaction.
- **Number of Projects (−0.03429)** - Workload management and task delegation could improve employee morale.

5.5 Test Data accuracy

```
# test data
model=lm(satisfaction_level~Emp.ID+salary+promotion_last_5years+average_montly_hours+last_evaluation+time_s
pend_company+number_project,data=test_data)
summary(model)
```

```
Call:
lm(formula = satisfaction_level ~ Emp.ID + salary + promotion_last_5years +
    average_montly_hours + last_evaluation + time_spend_company +
    number_project, data = test_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55691 -0.17181  0.01214  0.19145  0.58759

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.460e-01  3.194e-02  17.091  < 2e-16 ***
Emp.ID                9.022e-06  1.283e-06   7.032 2.61e-12 ***
salary                7.544e-07  8.881e-07   0.849    0.396
promotion_last_5years 6.361e-02  4.131e-02   1.540    0.124
average_montly_hours  1.419e-04  1.046e-04   1.356    0.175
last_evaluation       2.703e-01  3.085e-02   8.760  < 2e-16 ***
time_spend_company   -2.735e-02  3.741e-03  -7.311 3.53e-13 ***
number_project       -3.590e-02  4.299e-03  -8.352  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2383 on 2549 degrees of freedom
Multiple R-squared:  0.08471,   Adjusted R-squared:  0.0822
F-statistic:  33.7 on 7 and 2549 DF,  p-value: < 2.2e-16
```

From the model we have learnt,

- **Intercept (0.5460) -** This means, even without considering other factors, the expected satisfaction level starts at 0.546.
- **Employee ID (9.022e-06) -** Has a very small positive impact**,** likely does not have a real-world influence on satisfaction but might capture hidden patterns (e.g., department-wise differences).
- **Salary (7.544e-07) -** A very small positive effect on satisfaction, but not statistically significant (p-value = 0.396).
- **Promotion in the Last 5 Years (0.06361) -** Has a small positive impact on satisfaction, but not statistically significant (p-value = 0.124).
- **Average Monthly Hours (0.1419) -** Could suggest employees feel productive and engaged, but workload balance should be monitored to avoid burnout.
- **Last Evaluation Score (0.2703) -** Highly significant and positive impact (p-value < 0.001).
- **Time Spent in Company (-2.735e-02) -** Negative and highly significant impact (p-value < 0.001).
- **Number of Projects (-3.590e-02) -** Ensure workload distribution is fair and provide better project management support.

```yaml
Residual standard error: 0.2368 on 10217 degrees of freedom
Multiple R-squared: 0.07535, Adjusted R-squared: 0.07463
F-statistic: 104.1 on 8 and 10217 DF, p-value: < 2.2e-16
```

Residual Standard Error (0.2368) – Measures the average prediction error; lower values indicate a better fit.

According to Linear regression error rate will be **0.2368**

CHAPTER VI

EVALUATION OF MODEL

- Evaluating machine learning algorithm is an essential part of any project. The model may give satisfying result when evaluating using a metric accuracy score but may give poor result when evaluated against other metrics such as logarithmic loss or any other such metric.

- The performance measure is the way to evaluate a solution to the problem. It is the measurement that will make of the predictions made by the trained model on the test dataset. Performance measures are typically specialized to the class of problem that are working with, for example classification, regression and clustering. Many standard performance measures will give a score that is meaning full to the problem domain.

- Since this project is related to regression model, the commonly used performance measure is mean squared error (MSE). In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors that is, the average difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

MSE

Mean squared error is an estimator measures the average of the squares of the errors that is the average squared difference between the estimated value and actual value.

MSE of Linear regression

```r
model=lm(satisfaction_level~Emp.ID+salary+promotion_last_5years+average_montly_hours+last_evaluation+time_spend_company+Work_accident+number_project,data=train_data)
summary(model)
```

According to Linear regression error rate will be **0.2368.**

Accuracy rate will be **84%**

CHAPTER VII

CONCLUSION

- The analysis of the employee satisfaction dataset using linear regression has provided valuable insights into the key factors affecting employee satisfaction. The model, with an accuracy of 84%, effectively captures the relationships between employee attributes such as salary, promotions, working hours, performance evaluation, tenure, and workload.

- The results indicate that performance evaluation scores and workload management play a crucial role in determining employee satisfaction.

- Employees handling multiple projects or spending more years in the company tend to have lower satisfaction levels, suggesting potential concerns related to workload balance and career progression.

- Overall, the dataset highlights the need for organizations to focus on employee engagement, career growth opportunities, and workload distribution to enhance satisfaction and retention. Future improvements to the model could include additional factors like workplace environment, leadership quality, and job flexibility to gain a more comprehensive understanding of employee well-being.

REFERENCES

Predictive HR Analytics: Mastering the HR Metric by Dr. Martin R. Edwards, Kirsten Edwards, and Daisung Jang. http://books.google.com/books?vid=ISBN0521349931

The Data Science Handbook *by Carl Shan et al.* regression, classification, and clustering http://books.google.com/books?vid=OCLC17546826.

Machine Learning yearning by *Andrew Ng*, structuring machine learning projects, diagnosing errors, and prioritizing strategies for improvement. http://books.google.com/books?vid=LCCN88005048.

Predictive HR Analytics: Mastering the HR Metric *by Martin Edwards & Kirsten Edwards.*

https://www.google.co.in/books/edition/Predictive_HR_Analytics/EXiJDwAAQBAJ?hl=en&gbpv=1&dq=inauthor:%22Dr+Martin+R.+Edwards%22&printsec=frontcover.