HR ANALYTICS

USING SQL

ABHINAV SREENIVAS

INTRODUCTION

This HR analysis offers a deep dive into the key elements that shape workforce dynamics within the organization, leveraging a range of attributes stored in the *project_hr* database. By examining employee demographics, training and development data, performance metrics, and recruitment channel effectiveness, the analysis uncovers meaningful insights into patterns that influence employee engagement, retention, and productivity.

Through carefully crafted SQL queries, this analysis empowers HR professionals with data-driven insights that inform strategic decision-making. Insights gained can help guide talent management strategies, optimize training investments, and enhance performance across departments and regions. This approach enables a holistic understanding of the workforce, allowing HR to make more targeted and impactful interventions that support both employee development and organizational success.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an approach in data science used to analyze datasets to summarize their main characteristics, often with the help of visualizations and descriptive statistics. The primary goal of EDA is to uncover patterns, detect anomalies, test hypotheses, and check assumptions, providing a foundational understanding of the data before proceeding with more complex modeling or data processing tasks.

EDA involves several key steps:

1. **Data Collection And Understanding**

First, we gather data from relevant sources and examine its structure, types, and dimensions. Understanding the context and variables within the data is crucial, as it helps shape the direction of analysis and decision-making.

2. **Data Cleaning**

This step includes handling missing values, correcting data types, dealing with outliers, and ensuring data consistency.

Data cleaning is essential to enhance the quality and reliability of insights derived from the data.

3. Feature Engineering

It involves transforming raw data into meaningful features that improve the performance of machine learning models and enhances the dataset by creating, modifying, or selecting variables that reveal patterns and improve the interpretability of data.

4. Univariate Analysis

This involves examining each variable individually, often through visualizations like histograms, box plots, and bar charts, to understand its distribution and key characteristics.

5. Bivariate Analysis

It is a statistical method used to explore the relationship between two variables. It aims to understand how the two variables interact, influence each other, or show correlations.

6. Multivariate Analysis

It is a statistical approach used to examine more than two variables simultaneously to understand complex relationships within a dataset. Unlike univariate or bivariate analysis, which focuses on one or two variables, respectively, multivariate analysis considers multiple variables to gain deeper insights, detect patterns, and capture interactions that may be missed when examining variables individually.

DATA SET DESCRIPTION

Data set source : Kaggle
Rows :17420
Columns :13
The data set contains the following feature :
1. *employee_id*: Unique identifier for each employee in the organization. This column is often used for counting employees or grouping them in different queries.

2. *gender*: Indicates the gender of the employee, typically coded as 'm' for male and 'f' for female. Used in queries to understand gender distribution and analyze gender-based performance or award trends.

3. *age*: Represents the age of the employee. This column is useful for analyzing age distribution, grouping employees into age ranges, and understanding the relationship between age and training or performance metrics.

4. *department*: Denotes the department in which the employee works, such as HR, IT, or Sales. This column helps to identify trends, such as average age, training scores, and award counts within specific departments.

5. *region*: Geographic region where the employee is based. This field is used to understand the distribution of employees by location and to analyze training effectiveness and KPI achievement by region.

6. *no_of_trainings*: Indicates the number of training sessions an employee has attended. This column helps analyze training participation and its correlation with KPI achievement, awards, and length of service.

7. *avg_training_score*: The average score an employee has achieved across all attended training programs. It is used to evaluate training effectiveness by department, education level, and recruitment channel.

8. *KPIs_met_more_than_80*: A binary indicator where 1 signifies that an employee has met more than 80% of their key performance indicators (KPIs), and 0 indicates otherwise. This column is crucial in analyzing high-performing employees and examining how training impacts KPI achievement.

9. *length_of_service*: The number of years the employee has been with the organization. This column is used to analyze retention, correlate with training and KPI achievement, and assess average service length by department and recruitment channel.

10. *previous_year_rating*: Performance rating from the previous year, typically on a scale from 1 to 5, with 5 being the highest. This metric is used to analyze trends in awards and training effectiveness, as well as performance trends across different demographics.

11. *awards_won*: Indicates the number of awards an employee has won. Used in analyses to understand award distribution by department, gender, and other demographics.

12. *education*: Represents the educational background of the employee, such as Bachelor's, Master's, or Doctorate. This column is analyzed to understand how education affects training scores, performance ratings, and award trends.

13. *recruitment_channel*: Specifies the recruitment source through which the employee joined the organization, such as internal referral or third-party agency. This column is valuable for analyzing the effectiveness of recruitment channels in terms of retention, training scores, and performance.

DATA CLEANING AND PREPROCESSING

1. Identify and Remove Duplicates

- **Counting Total Rows:** Counts the total number of records to assess the dataset size.
- **Identifying Duplicate Records:** Finds the number of duplicate `employee_id` entries.
- **Removing Duplicates:** Uses row numbers and a temporary table to remove duplicate records, retaining only the first instance.

```
SELECT COUNT(employee_id) FROM project_hr;

SELECT (COUNT(employee_id) - COUNT(DISTINCT employee_id)) AS no_of_duplicates FROM project_hr;

SELECT employee_id, COUNT(employee_id)  FROM project_hr  GROUP BY employee_id

HAVING COUNT(employee_id) > 1;

DELETE FROM project_hr

WHERE employee_id IN (SELECT employee_id FROM (SELECT employee_id, ROW_NUMBER() OVER (PARTITION BY employee_id ORDER BY employee_id) AS row_num FROM project_hr) AS duplicate_rowsWHERE row_num > 1);
```

2. Handling Missing Values

- **Filling Missing Values in Education Field:** Updates rows where `education` is blank, setting them to "Not Specified."
- **Handling Missing Values for `previous_year_rating`:**
  - **Data Imputation Strategy:** Calculates mean, mode, and median for `previous_year_rating` (excluding blanks), which are all close to 3, making 3 a reasonable choice for imputation.
  - **Imputation:** Replaces missing `previous_year_rating` values with 3 to maintain consistency.
  - **Type Consistency:** Ensures that `previous_year_rating` is correctly set as an integer.

```
UPDATE project_hr  SET education = 'Not Specified'  WHERE education = '';

SELECT AVG(previous_year_rating) AS MEAN  FROM project_hr  WHERE
previous_year_rating != 'None';

SELECT previous_year_rating, COUNT(*) AS frequency FROM project_hr GROUP BY
previous_year_rating  ORDER BY frequency DESC  LIMIT 1;
```

Impute Missing Values:

```
UPDATE project_hr  SET previous_year_rating = 3 WHERE previous_year_rating = ' ';

SELECT COUNT(previous_year_rating)  FROM project_hr  WHERE previous_year_rating
= '';
```

3. Data Consistency Checks

- **Checking Column Data Types:** Confirms the data type of each column in the `project_hr` table for consistency and compatibility.
- **Null Values Check by Column:** Counts missing values for each critical column to ensure data completeness.

```
ALTER TABLE project_hr MODIFY COLUMN previous_year_rating INT;
```

4. Final Null Check and Data Validation

- **Re-validate Nulls for `previous_year_rating`:** Verifies that all previously null values in `previous_year_rating` have been addressed.

```
SELECT COUNT(*) FROM project_hr WHERE employee_id IS NULL;
SELECT COUNT(*) FROM project_hr WHERE department IS NULL;
SELECT COUNT(*) FROM project_hr WHERE region IS NULL;
SELECT COUNT(*) FROM project_hr WHERE education IS NULL;
SELECT COUNT(*) FROM project_hr WHERE gender IS NULL;
SELECT COUNT(*) FROM project_hr WHERE recruitment_channel IS NULL;
SELECT COUNT(*) FROM project_hr WHERE no_of_trainings IS NULL;
SELECT COUNT(*) FROM project_hr WHERE age IS NULL;
SELECT COUNT(*) FROM project_hr WHERE previous_year_rating IS NULL;
SELECT COUNT(*) FROM project_hr WHERE length_of_service IS NULL;
SELECT COUNT(*) FROM project_hr WHERE KPIs_met_more_than_80 IS NULL;
SELECT COUNT(*) FROM project_hr WHERE awards_won IS NULL;
SELECT COUNT(*) FROM project_hr WHERE avg_training_score IS NULL;
```

UNIVARIATE ANALYSIS

Univariate analysis in HR focuses on examining individual variables to understand the fundamental characteristics of the workforce. By analyzing variables such as age distribution, gender ratio, department size, and average training scores, HR can identify key patterns in employee demographics, training participation, and performance levels. This approach provides a foundational understanding that helps guide more targeted HR interventions, from optimizing recruitment strategies to enhancing employee engagement and retention efforts.

```
SELECT COUNT(*) AS total_employees FROM project_hr;

SELECT MAX(previous_year_rating) AS max_previous_rating FROM project_hr;

SELECT AVG(age) AS average_age FROM project_hr;

SELECT gender, COUNT(*) AS gender_count FROM project_hr GROUP BY gender;

SELECT department, AVG(age) AS average_age FROM project_hr GROUP BY department;

SELECT age, COUNT(*) AS age_count FROM project_hr GROUP BY age ORDER BY age;

SELECT AVG(avg_training_score) AS average_training_score FROM project_hr;

SELECT AVG(length_of_service) AS average_length_of_service FROM project_hr;

SELECT AVG(no_of_trainings) AS average_trainings FROM project_hr;

SELECT AVG(previous_year_rating) AS average_previous_rating FROM project_hr;

SELECT AVG(KPIs_met_more_than_80) * 100 AS kpi_achievement_rate FROM project_hr;

SELECT MIN(avg_training_score) AS min_training_score FROM project_hr;

SELECT COUNT(*) AS no_awards_count FROM project_hr WHERE awards_won = 0;

SELECT AVG(KPIs_met_more_than_80) * 100 AS kpi_achievement_new_hires
```

```
FROM project_hr WHERE length_of_service < 1;


SELECT AVG(age) AS avg_age_top_rated FROM project_hr WHERE previous_year_rating = 5;

SELECT ROUND(COUNT(CASE WHEN avg_training_score < (SELECT AVG(avg_training_score) FROM project_hr) THEN 1 END) * 100.0 / COUNT(*), 2) AS pct_below_avg_score FROM project_hr;

SELECT SUM(no_of_trainings) AS total_trainings FROM project_hr;

SELECT COUNT(DISTINCT recruitment_channel) AS unique_recruitment_channels FROM project_hr;

SELECT AVG(KPIs_met_more_than_80) * 100 AS kpi_achievement_new_hires FROM project_hr WHERE length_of_service < 1;

SELECT education, COUNT(*) AS count FROM project_hr GROUP BY education ORDER BY count DESC LIMIT 3;

SELECT ROUND(COUNT(CASE WHEN avg_training_score > 80 THEN 1 END) * 100.0 / COUNT(*), 2) AS pct_above_80  FROM project_hr;

SELECT AVG(KPIs_met_more_than_80) * 100 AS avg_kpi_award_winners FROM project_hr WHERE awards_won > 0;

SELECT region, COUNT(*) AS region_count FROM project_hr GROUP BY region ORDER BY region_count DESC;
```

Result :

- The total number of employees is 17,412.
- The average age of employees is 34.81 years.
- The majority of employees are male (70.94%), with females accounting for 29.06%.
- The average training score is 63.18 out of 100.
- The average length of service for employees is 5.8 years.
- The average number of trainings attended per employee is 1.25.
- The average previous year rating is 3.32 out of 5.
- The KPI achievement rate is 35.88%.
- The minimum training score is 39.
- A significant number of employees (98.20%) have not won any awards.

- The majority of employees hold a Bachelor's degree (66.10%), followed by Masters or higher degrees (27.77%). A small percentage have not specified their education level (4.43%).
- Only 16.69% of employees have training scores above 80.
- The average age of top-rated employees (previous year rating of 5) is 35.63 years.
- The total number of training sessions attended by all employees is 21,778.

BIVARIATE ANALYSIS

In human resource analysis Bivariate analysis helps examine the relationship between two variables, such as employee performance and tenure or job satisfaction. and type of department By analyzing these variable pairs HR professionals can discover trends and relationships that may affect employee performance, turnover, or engagement. This analysis provides valuable insights for making data-driven decisions in talent management and organizational strategy. In human resource analysis Bivariate analysis helps examine the relationship between two variables, such as employee performance and tenure or job satisfaction. and type of department By analyzing these variable pairs HR professionals can discover trends and relationships that may affect employee performance, turnover, or engagement. This analysis provides valuable insights for making data-driven decisions in talent management and organizational strategy

```
SELECT department, gender, COUNT(employee_id) AS employee_count  FROM project_hr     GROUP BY department, gender;

SELECT department, AVG(age) AS avg_age  FROM project_hr  GROUP BY department;

SELECT education, AVG(KPIs_met_more_than_80) * 100 AS kpi_achievement_rate

FROM project_hr   GROUP BY education;

SELECT recruitment_channel, AVG(length_of_service) AS avg_service_years

FROM project_hr GROUP BY recruitment_channel;

SELECT recruitment_channel, AVG(previous_year_rating) AS avg_rating

FROM project_hr GROUP BY recruitment_channel;

SELECT education, AVG(avg_training_score) AS avg_training_score

FROM project_hr  GROUP BY education;

SELECT education,

        COUNT(CASE WHEN awards_won > 0 THEN 1 END) * 100.0 / COUNT(*) AS award_winning_rate FROM project_hr  GROUP BY education;
```

```
SELECT
        CASE
        WHEN age < 30 THEN 'Under 30'
        WHEN age BETWEEN 30 AND 40 THEN '30-40'
        WHEN age BETWEEN 41 AND 50 THEN '41-50'
        ELSE 'Over 50'
        END AS age_group,
    AVG(no_of_trainings) AS avg_trainings,
        AVG(awards_won) * 100 AS award_winning_rate
FROM project_hr  GROUP BY age_group;
 SELECT recruitment_channel, AVG(length_of_service) AS avg_length_of_service
FROM project_hr  GROUP BY recruitment_channel;
SELECT region, AVG(KPIs_met_more_than_80) * 100 AS kpi_achievement_rate
FROM project_hr  GROUP BY region;
SELECT recruitment_channel, AVG(avg_training_score) AS avg_training_score
FROM project_hr GROUP BY recruitment_channel;
SELECT no_of_trainings, AVG(avg_training_score) AS avg_training_score
FROM project_hr GROUP BY no_of_trainings;
```

Result :

- The **Technology Department** has a total of **1,348 male staff** members.
- The **average age** of employees in the Technology Department is **35.0364 years**.
- Among employees with a **Bachelor's degree**, the **average KPI achievement rate** is **35.7912%**.
- The **recruitment channel** named **Sourcing** has an **average service duration** of **5.7812 years**.
- For employees with a **Bachelor's degree**, the **average training score** is **63.0334**.
- The **award-winning rate** for employees with a **Bachelor's degree** is **2.29286**.
- For the **other recruitment channel**, the **average length of service** is **5.8415 years**.
- For the **Referred recruitment channel**, the **average training score** is **64.4032**.

MULTIVARIATE ANALYSIS

Multivariate analysis in HR provides a comprehensive view of factors that impact employee performance, retention, and engagement. By examining multiple variables simultaneously, such as demographics (department, region, education), recruitment (channel, number of trainings), and performance metrics (ratings, KPIs met, awards, training scores), we can uncover complex patterns and interactions. This approach enables HR teams to make data-driven decisions, optimize talent management strategies, and improve organizational outcomes through a deeper understanding of what drives employee success and retention.

```
SELECT region, gender, ROUND(AVG(age), 2) AS avg_age FROM project_hr
GROUP BY region, gender;
SELECT department, region, COUNT(employee_id) AS employee_count
FROM project_hr  GROUP BY department, region  ORDER BY region DESC;
SELECT department, COUNT(employee_id) AS num_under_30, region
FROM project_hr  WHERE age < 30  GROUP BY department, region ORDER BY
num_under_30 DESC;
SELECT education, gender, ROUND(AVG(length_of_service), 2) AS
avg_len_of_service
FROM project_hr  WHERE no_of_trainings > 2 AND avg_training_score > 75
GROUP BY education, gender;
SELECT department, region, COUNT(employee_id) AS no_of_employees
FROM project_hr  WHERE KPIs_met_more_than_80 > 0 AND awards_won > 0
GROUP BY department, region  ORDER BY no_of_employees DESC  LIMIT 3;

SELECT
   CASE
      WHEN age < 30 THEN 'Under 30'
      WHEN age BETWEEN 30 AND 40 THEN '30-40'
      WHEN age BETWEEN 41 AND 50 THEN '41-50'
      ELSE 'Over 50'
   END AS age_group,
   AVG(avg_training_score) AS avg_training_score,
   AVG(KPIs_met_more_than_80) AS kpi_achievement_rate
```

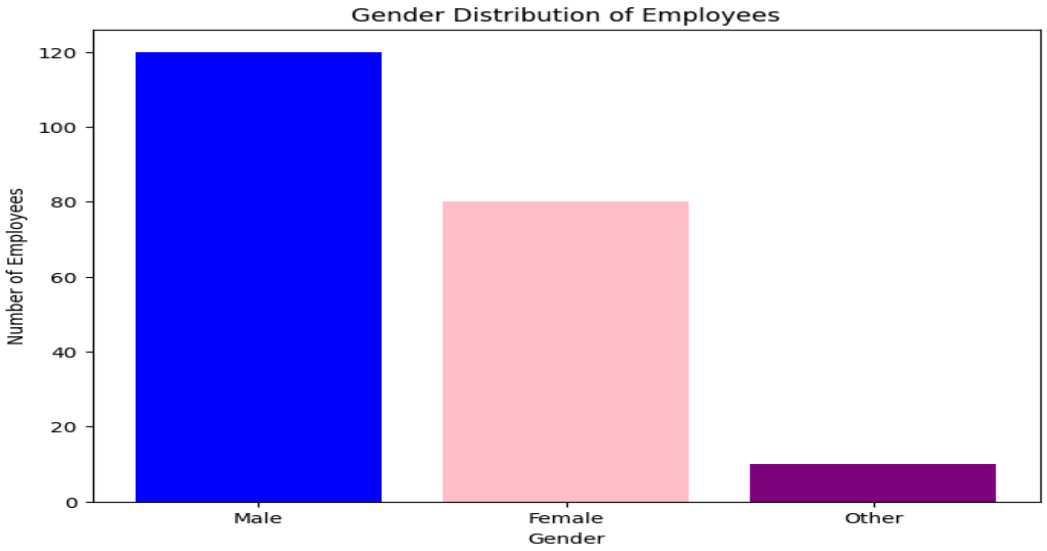
```
FROM project_hr
GROUP BY age_group;
SELECT no_of_trainings,   AVG(length_of_service) AS avg_tenure,
AVG(previous_year_rating) AS avg_performance FROM project_hr  GROUP BY
no_of_trainings  ORDER BY no_of_trainings DESC;



SELECT department,
     ROUND((COUNT(CASE WHEN gender = 'f' AND awards_won > 0 THEN 1
END))/(COUNT(CASE WHEN gender = 'f' THEN 1 END)) * 100, 2) AS
total_F_awards_percent,
     COUNT(CASE WHEN gender = 'f' AND awards_won > 0 THEN 1 END) AS
total_F_awards,
     COUNT(CASE WHEN gender = 'f' THEN 1 END) AS total_F_employees
FROM project_hr GROUP BY department;
SELECT recruitment_channel, education, COUNT(employee_id) AS
no_of_employees_having_KPIs_80plus
FROM project_hr  WHERE KPIs_met_more_than_80 > 0  GROUP BY
recruitment_channel, education;
```

Result :

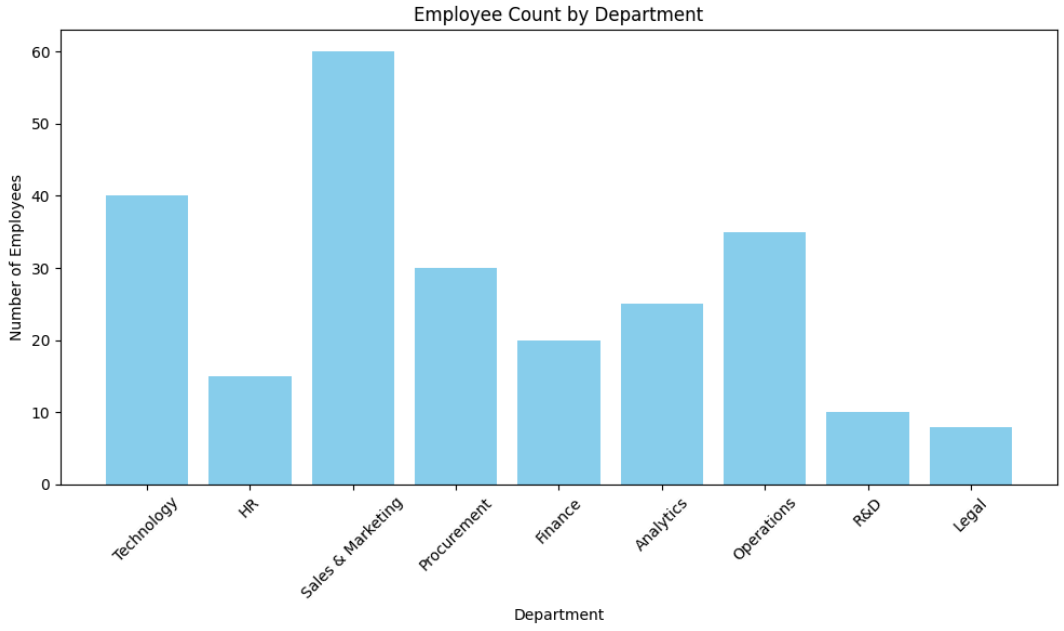
- The 'Sales & Marketing' department has the highest number of employees under 30 years old, with 231 employees in 'region_22'.
- Masters-level male employees have the highest average length of service (6.7 years).
- Bachelors-level female employees have a higher average length of service (4.81 years) compared to Bachelors-level male employees (3.97 years).
- The 'Sales & Marketing' department in 'region_2' has the highest number of high-performing employees (KPIs met > 80% and awards won).
- The 'Under 30' age group has the highest KPI achievement rate (35.95%).
- The '30-40' age group has the highest average training score (63.27)
- The 'Operations' department has the highest percentage of female employees who have won awards (2.94%).
- The 'other' recruitment channel has the highest number of employees with KPIs met more than 80% across all education levels.

VISUALISATION

BAR PLOT :

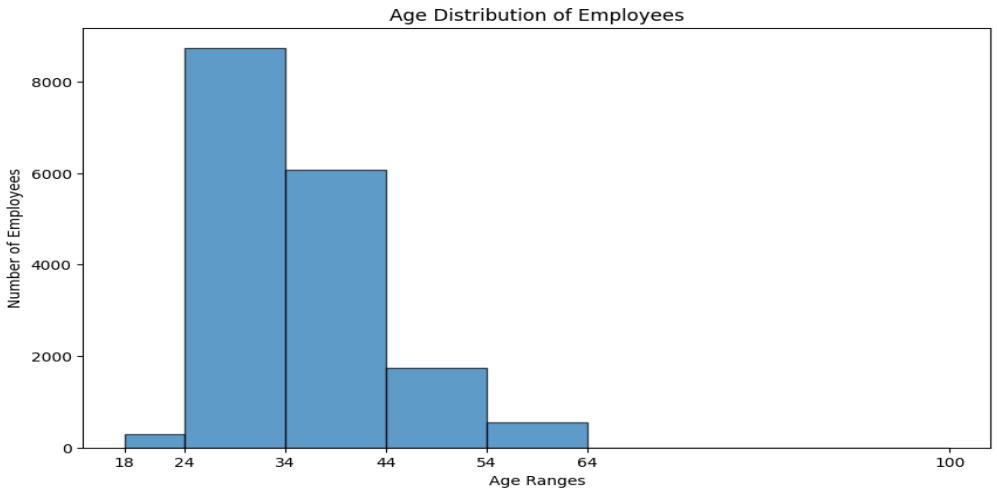


The bar chart presents a gender distribution among employees. The majority of employees are male, followed by females. The number of employees identifying as 'Other' is significantly lower compared to the other two categories.



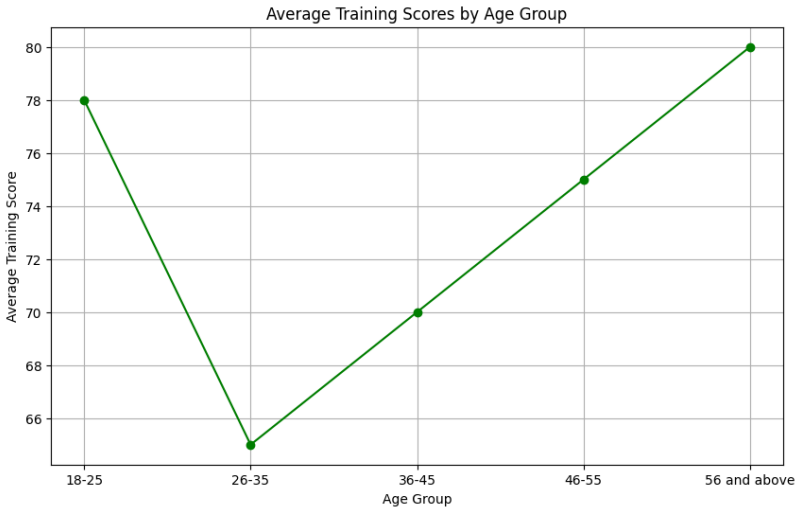
The bar chart presents the employee count across different departments. The Technology department has the highest number of employees, followed by Sales & Marketing. The Legal and R&D departments have the lowest number of employees.

Histogram



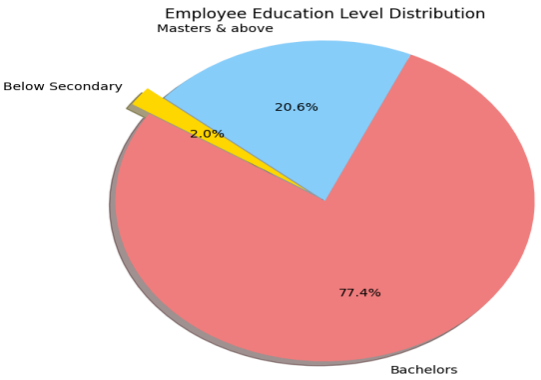
The histogram displays the age distribution of employees. The majority of employees are in the 25-34 age group. The number of employees decreases as the age range increases, with the fewest employees in the oldest age group (64 and above).

Line Chart



The line chart illustrates the average training scores across different age groups. The average training score increases with age, with the highest score observed in the 56 and above age group and the lowest score in the 26-35 age group.

PIE CHART



The pie chart illustrates the distribution of employee education levels within an organization. The majority of employees (77.4%) hold a Bachelor's degree, followed by a smaller proportion with Masters or higher degrees (20.6%). A very small percentage (2.0%) have education below secondary level.

CONCLUSION

- Gender imbalance, with a much higher count of male employees.

- To promote inclusivity, prioritizing female recruitment, especially for leadership roles, could improve diversity. This approach will help cultivate varied perspectives and enhance overall organizational performance..
- The diverse age range across departments highlights an opportunity for leveraging cross-generational knowledge-sharing and mentorship programs.
- Differences in age profiles across departments may reflect unique hiring practices, which could inform tailored retention and career development strategies for each department.
- Training programs seem to be effective, as evidenced by the positive correlation between the number of trainings attended and performance metrics.
- The 'Referred' recruitment channel appears to yield employees with higher average training scores.
- Implement targeted training programs to enhance performance in specific age groups and departments.
- KPI achievement and awards are unevenly distributed across regions, with some areas demonstrating stronger performance and recognition cultures. Sharing best practices from high-performing regions could improve overall company culture and results.
- Longer-tenured employees generally show better KPI achievement, and those with higher education levels also tend to perform well. These insights reinforce the value of retaining experienced employees and fostering continuous education.