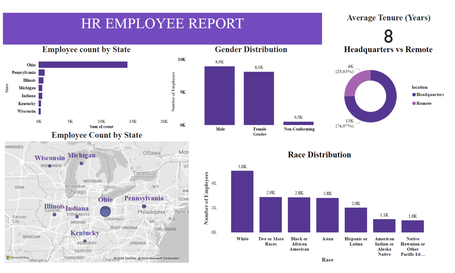
HR Employee Data Analysis using MySQL-PowerBI



This project provides an analysis of SQL queries used to examine employee data from the HR database. The aim is to uncover insights into various aspects of the workforce, such as demographic distributions, location-based employee statistics, turnover rates, and more. By conducting Exploratory Data Analysis (EDA) and calculating relevant metrics, the project assists in informing decisions related to human resource management, workforce planning, and diversity initiatives.

Objectives:

1. Understand demographic distributions (e.g., gender, race, age).
2. Analyze employee turnover rates and trends over time.
3. Investigate the distribution of employees across different locations and departments.
4. Identify workforce trends like employee growth, tenure etc.

Data Source

- **Data** - HR Data with over 22000 rows from the year 2000 to 2020.
- **Data Cleaning & Analysis** - MySQL Workbench
- **Data Visualization** - PowerBI

The dataset used in this analysis is the `HR` table, which contains employee information such as:

- `emp_id`: Unique identifier for each employee.
- `first_name`, `last_name`: Employee's personal information.
- `birthdate`: Date of birth of the employee.
- `gender`, `race`: Demographic information.
- `hire_date`: The date the employee was hired.
- `termdate`: The termination date, if applicable.
- `age`: The employee's age.
- `department`: The department where the employee works.
- `jobtitle`: Employee's job title.
- `location`: The employee's work location (e.g., headquarters or remote).
- `location_city`, `location_state`: City and state information for each employee's work location.

Data Cleaning

Before conducting the analysis, several data-cleaning steps were performed:

1. **Handling Null Values**: Columns such as `termdate` that had null values (for active employees) were either left blank or converted to `'0000-00-00'` for consistency.
2. **Date type Formatting**: The `birthdate`, `termdate` and `hire_date` columns were cleaned and formatted to ensure proper date values were stored.
3. **Age Calculation**: The employee's age was pre-calculated based on the birthdate, and discrepancies in age values were corrected to ensure accuracy for queries involving age distribution.

Key Cleaning Operations:

- **Replacing invalid date formats** in the `birthdate`, `termdate` and `hire_date` columns.
- **Ensuring age values are accurate** and reflect real-world employee ages.

Exploratory Data Analysis (EDA)

The EDA was performed using SQL queries to uncover trends and insights into employee demographics, turnover rates, and other HR metrics. Below are the types of EDA conducted:

1. Gender Breakdown of Employees

This query shows how the employees are distributed by gender, focusing only on employees currently working at the company.

```
SELECT gender, COUNT(*) AS count
FROM HR
WHERE age >= 18 AND termdate = '0000-00-00'
GROUP BY gender;
```

2. Race Breakdown of Employees

This query calculates the racial distribution of employees.

```
SELECT race, COUNT(*) AS count
FROM hr
WHERE age >= 18 AND termdate = '0000-00-00'
GROUP BY race
ORDER BY COUNT(*) DESC;
```

3. Age Distribution of Employees

To understand the age demographics, we calculate the youngest and oldest employees, as well as group employees into age ranges.

```
SELECT MIN(age) AS youngest, MAX(age) AS oldest
FROM HR
WHERE age >= 18 AND termdate = '0000-00-00';
```

Grouping employees into age brackets:

```
SELECT
  CASE
    WHEN age >= 18 AND age <= 24 THEN '18-24'
    WHEN age >= 25 AND age <= 34 THEN '25-34'
    WHEN age >= 35 AND age <= 44 THEN '35-44'
    WHEN age >= 45 AND age <= 54 THEN '45-54'
    WHEN age >= 55 AND age <= 64 THEN '55-64'
    ELSE '65+'
  END AS age_group,
  COUNT(*) AS count
FROM hr
WHERE age >= 18 AND termdate = '0000-00-00'
GROUP BY age_group
ORDER BY age_group;
```

4. Employee Distribution by Location

To understand where employees are located, this query shows how many employees are located in various locations by state.

```
SELECT location, count(*) AS count
FROM hr
WHERE age >= 18 AND termdate = '0000-00-00'
GROUP BY location;
```

5. The average length of employment for employees who have been terminated

```
SELECT
  round(avg(datediff(termdate, hire_date))/365, 0) AS avg_length_of_employment
FROM HR
WHERE termdate <= curdate() AND termdate != '0000-00-00' AND age >= 18;
```

6. The gender distribution across departments

```
SELECT department,gender, count(*) AS count
FROM hr
WHERE age >= 18 AND termdate = '0000-00-00'
GROUP BY department, gender
ORDER BY department, gender;
```

7. The distribution of job titles across the company

```
SELECT jobtitle, COUNT(*) AS count
FROM HR
WHERE age >= 18 AND termdate = '0000-00-00'
GROUP BY jobtitle
ORDER BY jobtitle DESC;
```

8. Employee Turnover Rate by Department

This query helps determine which department has the highest turnover rate by comparing the number of terminated employees to the total count.

```
SELECT department,
 total_count,
    terminated_count,
    terminated_count/total_count AS termination_rate
FROM (
  SELECT department,
   count(*) AS total_count,
      SUM(CASE WHEN termdate != '0000-00-00' AND termdate <= curdate()THEN 1 ELSE 0 END) AS terminated_count
  FROM hr
    WHERE age >= 18
    GROUP BY department
    ) AS dept_stats
ORDER BY termination_rate DESC;
```

9. The distribution of employees across locations by city and state

```
SELECT location_state, count(*) AS count
FROM hr
WHERE age >= 18 AND termdate = '0000-00-00'
GROUP BY location_state
ORDER BY count DESC;
```

10. The company's employee change over time based on hire and term dates

```
SELECT
  YEAR,
  hires,
    terminations,
    hires - terminations AS net_change,
    round((hires - terminations)/hires*100, 2) AS percent_net_change
FROM (
  SELECT YEAR(hire_date) AS YEAR,
   count(*) AS hires,
      SUM(CASE WHEN termdate <> '0000-00-00' AND termdate <= curdate() THEN 1 ELSE 0 END) AS terminations
  FROM HR
    WHERE age >= 18
    GROUP BY year(hire_date)
    ) AS hire_metrics
ORDER BY YEAR ASC;
```

11. The average tenure distribution for each department

```
SELECT department, round(avg(datediff(termdate, hire_date)/365), 0) AS avg_tenure
FROM hr
WHERE age >= 18 AND termdate <= curdate() AND termdate <> '0000-00-00'
GROUP BY department
ORDER BY avg_tenure DESC;
```

12. Department with the highest number of employees

```
SELECT department,count(*) AS employee_count
FROM hr
WHERE age >= 18 AND termdate = '0000-00-00'
GROUP BY department
ORDER BY employee_count DESC
LIMIT 1;
```

Questions

1. What is the gender breakdown of employees in the company?
2. What is the race breakdown of employees in the company?
3. What is the age distribution of employees in the company?
4. How many employees work at headquarters versus remote locations?
5. What is the average length of employment for employees who have been terminated?
6. How does the gender distribution vary across departments?
7. What is the distribution of job titles across the company?
8. Which department has the highest turnover rate?
9. What is the distribution of employees across locations by city and state?
10. How has the company's employee count changed over time based on hire and term dates?
11. What is the tenure distribution for each department?
12. Which department has the highest number of employees based on hire and term dates?
13. What is the tenure distribution for each department?
14. Which department has the highest number of employees?

Key Insights

1. **Gender Breakdown**: The gender distribution is relatively balanced across the company, with males higher in number than females.
2. **Race Breakdown**: Certain races may be underrepresented, with the White race being the most dominant.
3. **Age Distribution**: The majority of employees fall within the 25-34,35-44 and 45-54 age ranges, with a minimum age of 22 and a maximum age of 58.
4. **Headquarters vs. Remote**: A clear distinction exists between employees working at headquarters and those in remote locations. A large number of employees work at the headquarters versus remotely.
5. **Average Length of Employment**: On average, terminated employees stayed with the company for about 8 years.
6. **Gender Across Departments**: Some departments have notable gender imbalances, which may require attention to promote diversity.
7. **Job Title Distribution**: Job titles are distributed widely, with some job titles having significantly higher counts than others.
8. **Department with Highest Turnover**: Certain departments experience higher turnover rates, which may signal dissatisfaction or other issues. Auditing has the highest termination rate, followed by Legal and Marketing with the lowest.
9. **Location Distribution**: Most employees come from the state of Ohio.
10. **Employee Growth**: Over the years, the company has experienced fluctuating hiring and termination trends, with some years showing a net positive change or an increase in workforce size.
11. **Tenure by Department**:The average tenure across most departments is around 8 years, with Sales having the highest at 9 years, while Support, Training, Product Management, and Legal have slightly lower averages at 7 years.
12. **Department with Most Employees**: The Engineering department stands out as having the highest number of employees.

Limitations

Here are three limitations of the project:

1. **Data Quality**: The accuracy of the insights depends on the quality of the underlying data. Some records contained negative ages, leading to the exclusion of 967 entries from the analysis, while only ages of 18 years and above were considered.
2. **Scope of Analysis**: The project focuses on specific metrics, such as demographics, job titles, and tenure, which may not encompass all relevant factors affecting workforce dynamics. Additionally, 1,341 records with termination dates set far into the future were excluded, limiting the analysis to only those term dates on or before the current date.
3. **Interpretation of Results**: insights derived from the data can be subjective and open to various interpretations. Without adequate context or an understanding of the organizational culture, there is a risk of misapplying findings, potentially leading to ineffective decision-making.

Conclusion

In conclusion, this project has provided valuable insights into employee data, highlighting key aspects such as demographics, location, job titles, and tenure. By employing SQL queries and visualizing the results with Power BI, the analysis has enhanced the understanding of workforce dynamics and their implications. Despite some limitations, such as data quality issues and the exclusion of certain records, the findings can inform future human resource policies, diversity initiatives, and workforce planning strategies. Ultimately, these insights can help organizations optimize their workforce for improved efficiency and inclusivity.