HR DATA – Employee Distribution Report (SQL + Power BI)

Data Analysed by: Gokulakrishnan K

Project Overview:

Objective:

In today's data-driven world, organizations rely on data analysis to gain valuable insights into various aspects of their operations. One crucial area is human resources (HR), where analyzing employee data can provide valuable information about the workforce. In this project, we will explore a comprehensive analysis of HR data, including the gender breakdown, race/ethnicity distribution, age demographics, location distribution, employment length, gender distribution across departments and job titles, job title distribution, turnover rate by department, employee distribution across locations, and changes in employee count over time. By employing SQL for data cleaning and analysis and Power BI for visualization, the project aims to deliver a holistic view of employee distribution trends within the organization.

Executive Summary:

The project undertakes a detailed exploration of HR data spanning from the year 2000 to 2020, encompassing over 22,000 rows. The analysis is structured around answering specific questions related to employee demographics, organizational structure, and historical trends. MySQL Workbench is employed for data cleaning and analysis, ensuring data integrity and relevance. Power BI is utilized for its robust visualization capabilities, enabling stakeholders to gain actionable insights from the analyzed HR data.

Data Used:

Data - HR Data with over 22000 rows from the year 2000 to 2020.

Data Cleaning & Analysis - MySQL Workbench

Data Visualization – Power BI

Problem Statement:

Our Client, a leading multi-national corporation has been experiencing some puzzling trends in their employee turnover rates they suspect that there may be hidden factors causing valuable employees to leave, which aim to address the following questions:

1. What is the gender breakdown of employees in the company?
2. What is the race/ethnicity 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 state?
10. How has the company's employee count changed over time based on hire and termination dates?
11. What is the tenure distribution for each department?

Data Cleaning - MySQL Workbench:

The provided MySQL code performs various data cleaning and transformation steps on the 'hr' table in the 'hr_db' database. Here's a breakdown of each step:

Creating a Database:

```
CREATE DATABASE hr_db;
```

This query creates a new database named 'hr_db'.

Selecting the Database:

```
USE hr_db;
```

This query selects the 'hr_db' database for further operations.

Importing Raw Data:

- Import the csv file to the database using the Table Data Import Wizard option and name the table as 'hr'.

Viewing Table Structure:

```
DESCRIBE hr;
```

This query displays the structure of the 'hr' table, including the column names and data types.

Cleaning Id Column:

- Id column name needs correction so we are going to provide new column name to the id Column,

```
ALTER TABLE hr
RENAME COLUMN ï»¿id to emp_id;
```

- The query below changes the data type of emp_id column to varchar data type,

```
ALTER TABLE hr
MODIFY COLUMN emp_id VARCHAR(20);
```

✓ Before updating the records or fields in the table we need to temporarily disable safe update mode for our current session,

```
SET sql_safe_updates = 0;
```

Cleaning Birthdate Column:

```
UPDATE hr
SET birthdate =
CASE
WHEN birthdate LIKE '%/%' THEN date_format(str_to_date(birthdate,'%m/%d/%Y'),'%Y-%m-%d')
WHEN birthdate LIKE '%-%' THEN date_format(str_to_date(birthdate,'%m-%d-%Y'),'%Y-%m-%d')
ELSE NULL
END;
```

This query converts the 'birthdate' column values to the 'YYYY-MM-DD' date format. It handles different date formats and sets any invalid dates as NULL.

Modifying Birthdate Column Data Type:

```
ALTER TABLE hr
MODIFY COLUMN birthdate DATE;
```

This query modifies the data type of the 'birthdate' column to DATE.

Cleaning Hiredate Column:

```
UPDATE hr
SET hire_date =
CASE
WHEN hire_date LIKE '%/%' THEN date_format(str_to_date(hire_date,'%m/%d/%Y'),'%Y-%m-%d')
WHEN hire_date LIKE '%-%' THEN date_format(str_to_date(hire_date,'%m-%d-%Y'),'%Y-%m-%d')
ELSE NULL
END;
```

This query converts the 'hire_date' column values to the 'YYYY-MM-DD' date format. It handles different date formats and sets any invalid dates as NULL.

Modifying Hiredate Column Data Type:

```
ALTER TABLE hr
MODIFY COLUMN hire_date DATE;
```

This query modifies the data type of the 'hire_date' column to DATE.

Cleaning Termdate Column:

```
UPDATE hr
SET termdate = date(str_to_date(termdate,'%Y-%m-%d %H:%i:%s UTC'))
WHERE termdate IS NOT NULL AND termdate != '';
```

This query converts the 'termdate' column values to the 'YYYY-MM-DD' date format. It handles a specific date format containing UTC information and excludes empty or spaces-only values.

- The below query updates NULL values in the empty spaces present in the termdate column,

```
UPDATE hr
SET termdate = NULL
WHERE termdate = '';
```

Modifying Termdate Column Data Type:

```
ALTER TABLE hr
MODIFY COLUMN termdate DATE;
```

This query modifies the data type of the 'termdate' column to DATE.

Adding Age Column:

```
ALTER TABLE hr
ADD COLUMN age int;
```

This query adds a new column named 'age' with the INT data type to the 'hr' table.

Calculating Age:

```
UPDATE hr
SET age = timestampdiff(YEAR,birthdate,curdate());
```

This query calculates the age of each employee based on their birthdate and the current date.

- ✓ After updating the records or fields in the table we can enable the safe update mode for our current session,

```
SET sql_safe_updates = 1;
```

Limitations in the data:

- Some termination dates were far into the future and were not included in the analysis. The only termination dates used were those less than or equal to the current date.

Data Analysis - MySQL Workbench:

The provided MySQL code answers various questions about the employee data in the 'hr' table. Here's the corresponding query for each requirement:

Total Employees working:

```
SELECT COUNT(*) AS "Total Employees"  FROM hr
WHERE termdate IS NULL;
```

- This query counts the total number of employees in the 'hr' table with employees not having termination date.

Minimum and Maximum Age of Employees:

```
SELECT MIN(age) AS "Minimum Age", MAX(age) AS "Maximum Age" FROM hr;
```

- This query retrieves the minimum and maximum age values from the 'age' column, providing the youngest and oldest employee ages.

Gender Breakdown:

```
SELECT gender, COUNT(*) AS "Employee Count" FROM hr
WHERE termdate IS NULL
GROUP BY gender;
```

- This query calculates the count of employees in the 'hr' table based on their gender.

Race/Ethnicity Breakdown:

```
SELECT race, COUNT(*) AS "Employee Count" FROM hr
WHERE termdate IS NULL
GROUP BY race;
```

- This query calculates the count of employees in the 'hr' table based on their race/ethnicity.

Age Group Distribution:

```
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 "Employee Count" FROM hr
WHERE termdate IS NULL
GROUP BY Age_Group
ORDER BY Age_Group;
```

- This query calculates the count of employees in different age groups (18–24, 25–34, 35–44, 45–54, 55–64, 65+) and presents the results in ascending order of age groups.

Gender Distribution by Age Group:

```
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, gender, COUNT(*) AS "Employee Count" FROM hr
WHERE termdate IS NULL
GROUP BY Age_Group, gender
ORDER BY Age_Group;
```

- This query calculates the count of employees in the 'hr' table based on their gender and age groups. It provides the gender distribution across age groups.

Headquarters vs. Remote Locations:

```
SELECT location, COUNT(*) AS "Employee Count" FROM hr
WHERE termdate IS NULL
GROUP BY location;
```

- This query calculates the count of employees in the 'hr' table based on their location (headquarters or remote).

Average Length of Employment for Terminated Employees:

```
SELECT ROUND(AVG(YEAR(termdate)-YEAR(hire_date)),0) AS "Average Length of Employment"
FROM hr
WHERE termdate IS NOT NULL AND termdate <= curdate();
```

- This query calculates the average length of employment for employees who have been terminated. The result is rounded to the nearest whole number.

Gender Distribution across Departments:

```
SELECT department, gender, COUNT(*) AS "Employee Count" FROM hr
WHERE termdate IS NULL
GROUP BY department, gender
ORDER BY department;
```

- This query calculates the count of employees in the 'hr' table based on their gender and department. It provides the gender distribution across departments.

Distribution of Employees by Job Title:

```
SELECT jobtitle, COUNT(*) AS "Employee Count" FROM hr
WHERE termdate IS NULL
GROUP BY jobtitle;
```

- This query calculates the count of employees in the 'hr' table based on their job titles.

Department with the Highest Turnover / Termination Rate:

```
SELECT department, COUNT(*) AS Total_Count,
COUNT(CASE
        WHEN termdate IS NOT NULL AND termdate <= curdate() THEN 1
        END) AS Termination_Count,
ROUND((COUNT(CASE
                WHEN termdate IS NOT NULL AND termdate <= curdate() THEN 1
                END) / COUNT(*))*100,2) AS Termination_rate FROM hr
GROUP BY department
ORDER BY Termination_rate DESC;
```

- This query calculates the termination rate (as a percentage) for each department in the 'hr' table. It sorts the departments based on the termination rate in descending order.

Distribution of Employees across Locations (State):

```
SELECT location_state, COUNT(*) AS "Employee Count" FROM hr
WHERE termdate IS NULL
GROUP BY location_state;
```

- This query calculates the count of employees in the 'hr' table based on their location (state).

Employee Count 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 net_change_percentage
FROM (SELECT YEAR(hire_date) AS year, COUNT(*) AS hires,
COUNT(CASE
        WHEN termdate IS NOT NULL AND termdate <= curdate() THEN 1
        END) AS terminations FROM hr
        GROUP BY YEAR(hire_date)) AS sub_query
ORDER BY year;
```

- This query calculates the employee count change over time based on the hire and termination dates. It provides the number of hires, terminations, net change (hires — terminations), and net change percentage (net change divided by hires, multiplied by 100) for each year. The results are ordered in ascending order of the year.

Tenure Distribution of Terminated Employees for Each Department:

```
SELECT department, ROUND(AVG(YEAR(termdate)-YEAR(hire_date)),0) AS "Average Tenure"
FROM hr
WHERE termdate IS NOT NULL AND termdate <= curdate()
GROUP BY department;
```

- This query calculates the average tenure (in years) for each department in the 'hr' table. The results provide the average tenure rounded to the nearest whole number for each department.

Key Findings:

The following are some of the key findings from the data analysis:

- ✓ **Total Employees working:**
  - The Total Employees working in the Organization were **18285**.
- ✓ **Finding Minimum and Maximum Age of Employees:**
  - The youngest employee is 21 years old and the oldest is 58 years old.
- ✓ **Gender Breakdown:**
  - There are **more male employees**, outnumbering females by almost a thousand, while non-conforming individuals comprise around five hundred.
- ✓ **Race/Ethnicity Breakdown:**
  - **White race** is the most dominant with over five thousand employees falling into this category while **Native Hawaiian and American Indian** are the least dominant.
- ✓ **Age Group Distribution:**
  - 5 age groups were created (18-24, 25-34, 35-44, 45-54, 55-64).
  - A large number of employees were between **35-44** followed by **25-34** while the smallest group was **55-64**.
- ✓ **Gender Distribution by Age Group:**
  - Notably, the **35–44 age groups** tilts towards **more male and fewer female** employees compared to the **25–34 age groups**.
- ✓ **Headquarters vs. Remote Locations:**
  - A large number of employees work at the **Headquarters (74.98%)** comparing employees work at **Remote (25.02%).**
- ✓ **Average Length of Employment for Terminated Employees:**
  - The average length of employment for terminated employees is **8 years**, providing a glimpse into the longevity of past staff members.
- ✓ **Gender Distribution across Departments:**
  - The gender distribution across departments is fairly balanced but there are generally more male than female employees.
- ✓ **Distribution of Employees by Job Title:**
  - Research Assistant II job title has the **highest** employee count followed by Business Analyst and Human Resources Analyst II job titles.
- ✓ **Department with the Highest Turnover / Termination Rate:**
  - The **Auditing department** has the highest turnover rate followed by Legal department.
  - The least turnover rate was in the Support, Marketing and Business Development departments.
- ✓ **Distribution of Employees across Locations (State):**
  - **Ohio** emerges as the nucleus, hosting 14788 employees.
  - Pennsylvania takes second place with 930 employees, followed by Illinois with 730.
  - Michigan and Indiana boast around 570 employees each, while Wisconsin and Kentucky have fewer than 400 employees each.

- ✓ **Employee Count Change over Time (Based on Hire and Term Dates):**
  - Tracking the journey through hire and termination dates, the net change in employees has **increased over the years**.
- ✓ **Tenure Distribution of Terminated Employees for Each Department:**
  - The average tenure for each department is about 8 years with **Sales and Accounting** having the highest and Production Management, Support and Training having the lowest.

Recommendations:

Retention Strategies:

- Develop targeted retention strategies, especially focusing on departments with high turnover rates like Auditing and Legal.
- Analyze the factors contributing to termination in these departments and implement measures to address them.

Diversity and Inclusion:

- Promote diversity and inclusion initiatives, given the noticeable gender imbalances across departments.
- Consider implementing programs or policies to enhance gender equality, especially in departments where the gender distribution is skewed.

Training and Development:

- Identify departments with lower tenure and invest in training and development programs to enhance employee satisfaction and loyalty.
- Tailor programs based on the specific needs and characteristics of each department.

Remote Work Policies:

- Evaluate the success of remote work arrangements and consider expanding or optimizing remote work policies, given the significant percentage of employees working at headquarters.
- Assess the impact of remote work on employee satisfaction and productivity.

Succession Planning:

- Given the aging workforce, implement succession planning initiatives to ensure a smooth transition of knowledge and skills as employees retire or leave the organization.
- Identify key roles and individuals critical to the organization's success for targeted succession planning efforts.

Employee Engagement Surveys:

- Conduct regular employee engagement surveys to understand the sentiments and concerns of the workforce.

- Use survey results to address specific issues and enhance overall employee satisfaction.

Data Quality Improvement:

- Address data quality issues related to termination dates to ensure accurate and comprehensive analysis.
- Regularly update and validate termination dates to provide a more accurate representation of turnover trends.

Departmental Analysis:

- Conduct a detailed analysis of departments with low turnover rates (e.g., Support, Marketing, Business Development) to identify best practices that can be implemented in other departments.
- Share success stories and strategies across departments to foster a positive work environment.

Conclusion:

The comprehensive analysis of HR data has unveiled valuable insights into the organization's workforce, providing a foundation for strategic decision-making. The findings shed light on demographic patterns, departmental dynamics, and historical trends, enabling the organization to address critical issues related to turnover, diversity, and employee satisfaction.

While the organization demonstrates strengths in certain areas, such as employee distribution and tenure, there are clear opportunities for improvement in addressing turnover rates, gender imbalances, and remote work policies. The project's use of SQL for data analysis and Power BI for visualization has proven effective in presenting actionable insights to stakeholders.

Moving forward, the organization is well-positioned to implement targeted strategies to enhance employee retention, foster diversity and inclusion, and optimize workforce management practices. The limitations identified, particularly related to termination dates, highlight the importance of ongoing data quality improvement initiatives.

In conclusion, the HR data analysis project equips the organization with the knowledge needed to navigate the evolving landscape of its workforce, fostering a workplace environment that promotes employee satisfaction, engagement, and long-term success.

Power BI Report: