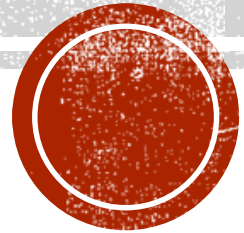
YOGESH'S DATA ANALYSIS SQL PROJECT

DATA CLEANING ,TRANSFORM AND ANALYSIS



A Case Study on company employees

Duration: 5-10minutes

PROBLEM STATEMENT / PROJECT SCOPE

- Our client wants the analysis of their employees in the company
- **Questions:**

1. What is the gender breakdown in the company ?
2. What is the race breakdown of employees in the company ?
3. What is the age distribution of employees in company ?
4. How many employees work at headqauters vs remote ?
5. What is the average length of employment for employees who have been terminated ?
6. What is the gender distribution vary across department ?
7. What is the gender distribution vary across job title ?
8. What is the distribution of employees across locations by city?
9. What is the distribution of employees across locations by state?

SOLUTION APPROACH

- There is one table provided which has 22214 rows and 14 columns in it .
- MYSQL was the tool used for Analysis and cleaning the data
- The data was imported, analysed and transformed as per necessity
- Birthdate ,hire_date ,termdate was formatted to right date format for the further analysis
- Adding a age column to further do our analysis
- KEY SKILLS
- Importing data using command line
- Data analysis
- Data cleaning
- Using queries
- Exporting data for visualization
- Report

IMPORTING DATA

- As MYSQL takes some time to import large amount of data .We have used command line to import the data using local_infile command in command prompt.

```
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> show tables
    -> ;
ERROR 1046 (3D000): No database selected
mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| housing            |
| how                |
| information_schema |
| luna               |
| mysql              |
| performance_schema |
| sakila             |
| sys                |
| world              |
+--------------------+
9 rows in set (0.00 sec)

mysql> use how;
Database changed
mysql> show tables;
+---------------+
| Tables_in_how |
+---------------+
| hrr           |
+---------------+
1 row in set (0.00 sec)
```

```
mysql> LOAD DATA LOCAL INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/hire.csv'
    -> INTO TABLE hrr
    -> FIELDS TERMINATED BY ','
    -> ENCLOSED BY '"'
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 ROWS;
Query OK, 22214 rows affected, 26143 warnings (1.40 sec)
Records: 22214  Deleted: 0  Skipped: 0  Warnings: 26143

mysql>
```

DATA TRANSFORMATION/CLEANING

```
--  Formating dates to correct format
```

```
UPDATE hrr
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;
```

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

```
UPDATE hrr
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;
```

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

```
UPDATE hrr
SET termdate = date(str_to_date(termdate, '%Y-%m-%d %H:%i:%s UTC'))
WHERE termdate is NOT null and termdate != '';
```

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

```
-- cleaning and fixing header of column 1

ALTER TABLE hrr
CHANGE COLUMN ï»¿id emp_id VARCHAR(20) NULL;

set SQL_SAFE_UPDATES = 0;
```

ADDING A AGE COLUMN FOR FURTHER ANALYSIS

- Adding age column and calculating the age
- Using timestampdiff function in MYSQL

```
-- Adding an age column

ALTER TABLE hrr ADD COLUMN age INT;
select * from hrr;

UPDATE hrr
SET age = timestampdiff(YEAR,birthdate,CURDATE());

select
 MIN(age) as youngest,
 max(age) as oldest
from hrr;
```

DATA TRANSFORMATION

emp_id	first_name	last_name	birthdate	gender	race	department	jobtitle	location	hire_date	termdate	location_city	location_state	age
00-0037846	Kimmy	Walczynski	1991-06-04	Male	Hispanic or Latino	Engineering	Programmer Analyst I	Headquarters	2002-01-20	0000-00-00	Cleveland	Ohio	32
00-0041533	Ignatius	Springett	1984-06-29	Male	White	Business Development	Business Analyst	Headquarters	2019-04-08	0000-00-00	Cleveland	Ohio	39
00-0045747	Corbie	Bittlestone	1989-07-29	Male	Black or African American	Sales	Solutions Engineer Manager	Headquarters	2010-10-12	0000-00-00	Cleveland	Ohio	34
00-0055274	Baxy	Matton	1982-09-14	Female	White	Services	Service Tech	Headquarters	2005-04-10	0000-00-00	Cleveland	Ohio	41
00-0076100	Terrell	Suff	1994-04-11	Female	Two or More Races	Product Management	Business Analyst	Remote	2010-09-29	2029-10-29	Flint	Michigan	29
00-0116166	Kacie	Offiler	1971-01-18	Male	Asian	Engineering	Developer III	Headquarters	2018-09-01	0000-00-00	Cleveland	Ohio	52
00-0363185	Sandro	Admans	1979-11-19	Male	Two or More Races	Product Management	Quality Engineer	Headquarters	2012-11-08	0000-00-00	Cleveland	Ohio	44
00-0380704	Eugene	Lehrahan	1988-10-14	Female	Black or African American	Engineering	Developer I	Headquarters	2007-06-27	0000-00-00	Cleveland	Ohio	35
00-0381660	Wainwright	Corfield	1996-12-13	Male	Asian	Engineering	Business Systems Development Analyst	Headquarters	2001-02-20	2008-12-05	Cleveland	Ohio	27
00-0419202	Dyann	Isoldi	1980-03-27	Male	Two or More Races	Engineering	Web Developer I	Headquarters	2005-01-27	0000-00-00	Cleveland	Ohio	43
00-0472287	Grantley	Oret	1975-09-06	Male	Two or More Races	Services	Service Tech II	Headquarters	2004-11-01	0000-00-00	Cleveland	Ohio	48
00-0472832	Elmore	Worner	1966-01-07	Female	White	Engineering	Business Systems Development Analyst	Headquarters	2000-12-05	0000-00-00	Cleveland	Ohio	57
00-0566380	Dud	Brain	1984-03-17	Male	Two or More Races	Business Development	Business Analyst	Headquarters	2008-09-17	0000-00-00	Cleveland	Ohio	39
00-0571075	Aguie	Conford	1971-11-02	Male	White	Business Development	Research Assistant II	Headquarters	2015-11-25	0000-00-00	Cleveland	Ohio	52
00-0624189	Katerina	Rosborough	1967-08-20	Male	Hispanic or Latino	Engineering	Analyst Programmer	Headquarters	2019-05-17	0000-00-00	Cleveland	Ohio	56
00-0715212	Alida	Longley	1973-01-28	Female	American Indian or Alask...	Accounting	Staff Accountant III	Headquarters	2002-02-04	0000-00-00	Cleveland	Ohio	50
00-0755645	Laraine	Petre	1967-05-11	Male	White	Engineering	Software Engineer I	Headquarters	2010-09-30	0000-00-00	Cleveland	Ohio	56
00-0778934	Gareth	MacCook	1987-02-21	Female	Black or African American	Legal	Senior Attorney	Remote	2010-02-18	0000-00-00	Flint	Michigan	36
00-0794247	Scottie	Chestney	1972-02-23	Female	Black or African American	Engineering	Data Visualization Specialist	Headquarters	2002-07-03	0000-00-00	Cleveland	Ohio	51
00-0948136	Christoph...	Boseley	1983-05-23	Female	Two or More Races	Marketing	Senior Editor	Headquarters	2007-09-06	0000-00-00	Cleveland	Ohio	40
00-0971612	Arleyne	Froome	1999-08-01	Male	Two or More Races	Engineering	Software Consultant	Headquarters	2015-04-09	0000-00-00	Cleveland	Ohio	24
00-1051096	Todd	Cashen	1999-06-19	Female	White	Accounting	Financial Analyst	Remote	2014-06-20	0000-00-00	Pittsburgh	Pennsylvania	24
00-1052230	Elmo	McNee	1988-12-16	Female	Black or African American	Services	Service Manager	Headquarters	2002-07-18	2006-05-22	Cleveland	Ohio	34
00-1100714	Regen	Nafzger	1990-04-06	Male	White	Human Resources	Senior Recruiter	Headquarters	2016-07-23	2022-07-12	Cleveland	Ohio	33
00-1147503	Penelope	Wenman	1994-12-19	Male	White	Business Development	Research Assistant II	Headquarters	2019-09-10	0000-00-00	Cleveland	Ohio	28
00-1189819	Rudolf	Reichardt	1979-06-05	Male	Black or African American	Business Development	Business Analyst	Remote	2007-05-07	2022-02-25	Philadelphia	Pennsylvania	44
00-1222963	Tobiah	Fruchon	1991-10-17	Female	Asian	Services	Service Tech II	Headquarters	2011-02-13	0000-00-00	Cleveland	Ohio	32
00-1268049	Fay	Monnelly	1966-07-09	Male	Native Hawaiian or Othe...	Engineering	Software Engineer I	Headquarters	2010-02-24	2030-03-21	Cleveland	Ohio	57
00-1277358	Lola	Burrells	1973-10-02	Male	Black or African American	Human Resources	Senior Recruiter	Remote	2008-03-08	2024-10-09	Pittsburgh	Pennsylvania	50
00-1284831	Hamel	Edgeler	1973-12-22	Male	Two or More Races	Human Resources	HR Manager	Headquarters	2001-08-08	2007-10-31	Cleveland	Ohio	49

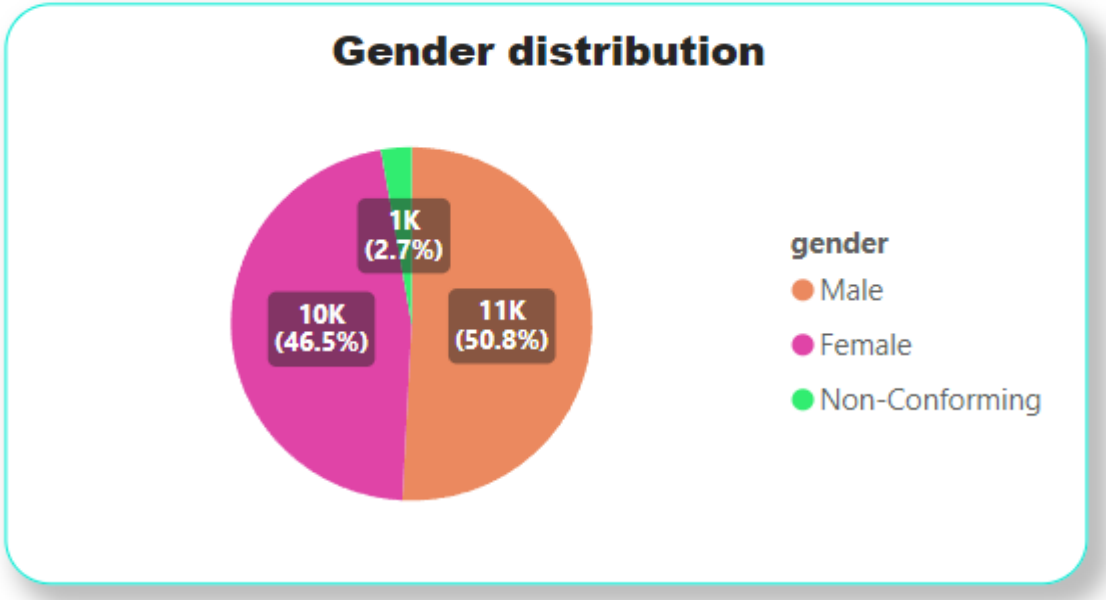
SOLUTION APPROACH

- What is the gender breakdown in company ?

```
3     -- What is the gender breakdown in company ?
4
5 •   SELECT DISTINCT gender, count(*) as count from hrr
6     group by gender;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

gender	count
Male	11288
Female	10321
Non-Conforming	605



- From the insight and Pie chart we can say that Male employee is more than Female and Non-conforming.

SOLUTION APPROACH

- What is the average length of employment for employees who have been terminated ?

```
27    -- What is the average length of employment for employees who have been terminatd ?
28 •  SELECT hire_date, termdate from hrr;
29 •  SELECT round(avg(datediff(termdate, hire_date))/365,0) as average_length
30    from hrr;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

average_length
10

Average length of employement

10

- The average length of employment for the employees in the company is 10 years .

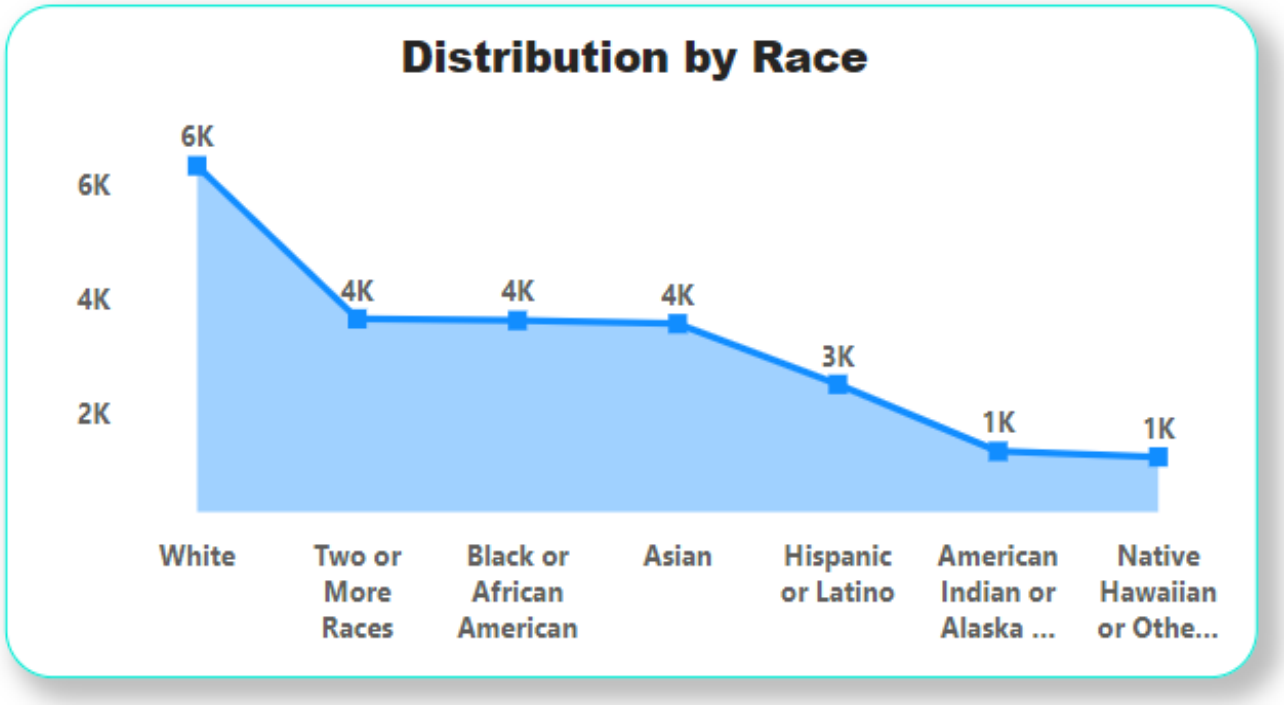
SOLUTION APPROACH

- What is the race breakdown of employees in the company?

```
7     -- What is the race breakdown of employees in the company?
8 •   SELECT race,count(race) from hrr
9     GROUP BY race;
10
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

race	count(race)
Hispanic or Latino	2501
White	6328
Black or African American	3619
Two or More Races	3648
Asian	3562
American Indian or Alaska Native	1327
Native Hawaiian or Other Pacific Islander	1229



- White peoples are the in the majority in the company .
- Native Hawaiian or others are the in the minority in the company .

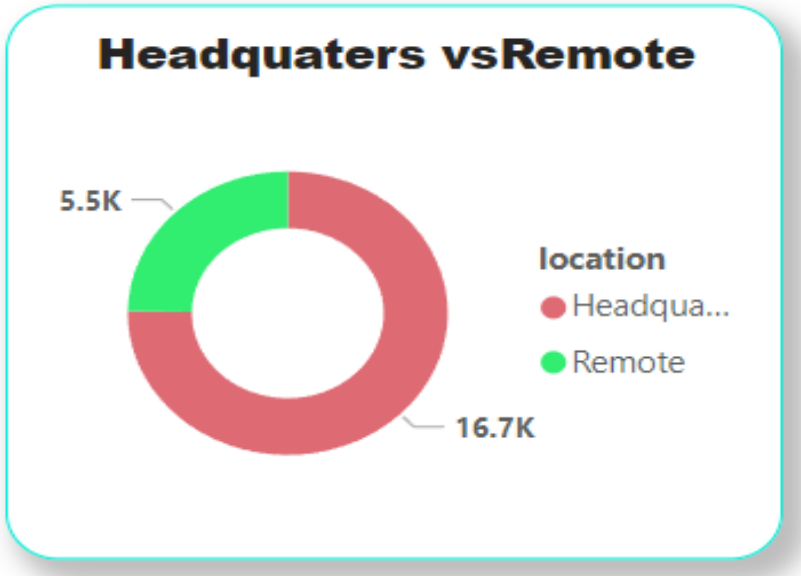
SOLUTION APPROACH

- How many employees work at headquaters vs remote ?

```
22
23      -- How many employees work at headquaters vs remote ?
24 •    select location, count(emp_id) as count from hrr
25      group by location;
26
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

location	count
Headquarters	16715
Remote	5499



- Majority of peoples work from headquarters .

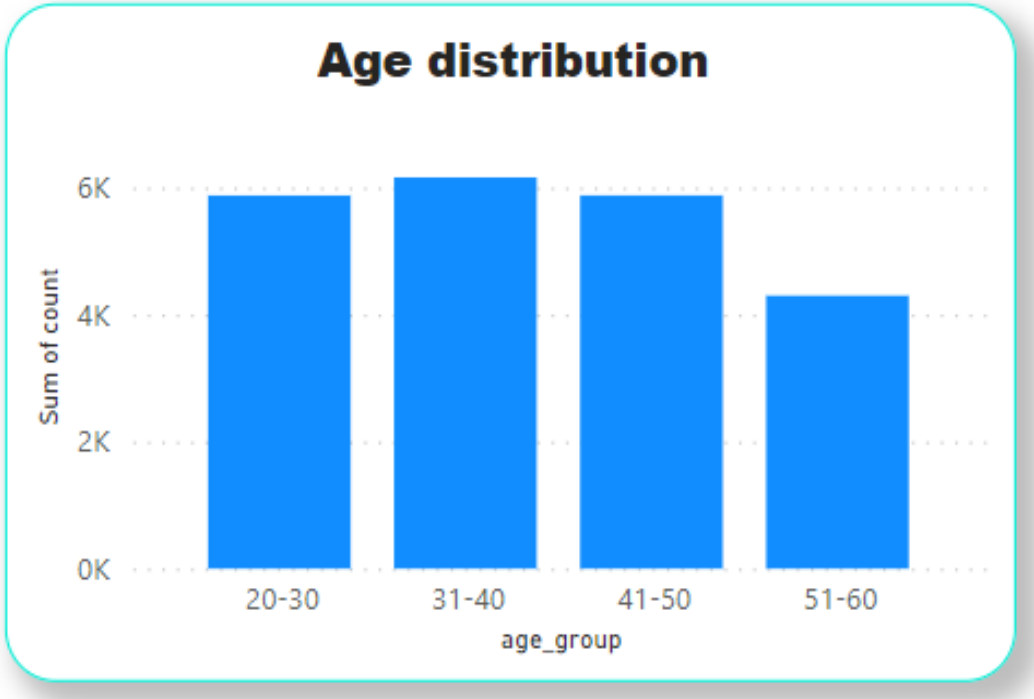
SOLUTION APPROACH

- What is the age distribution of employees in company ?

```
10
11      -- Whatr is the age distribution of employees in company ?
12      SELECT CASE
13          when age >= 20 and age <= 30 then '20-30'
14          when age >= 31 and age <= 40 then '31-40'
15          when age >= 41 and age <= 50 then '41-50'
16          when age >= 51 and age <= 60 then '51-60'
17          Else '61+'
18          END as age_group,
19          count(*) as count
20          from hrr GROUP BY age_group;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

age_group	count
31-40	6160
41-50	5878
20-30	5877
51-60	4299



- Company consists mainly of employees from age group 31-40

- While 51-60 age group employees are the least in the company

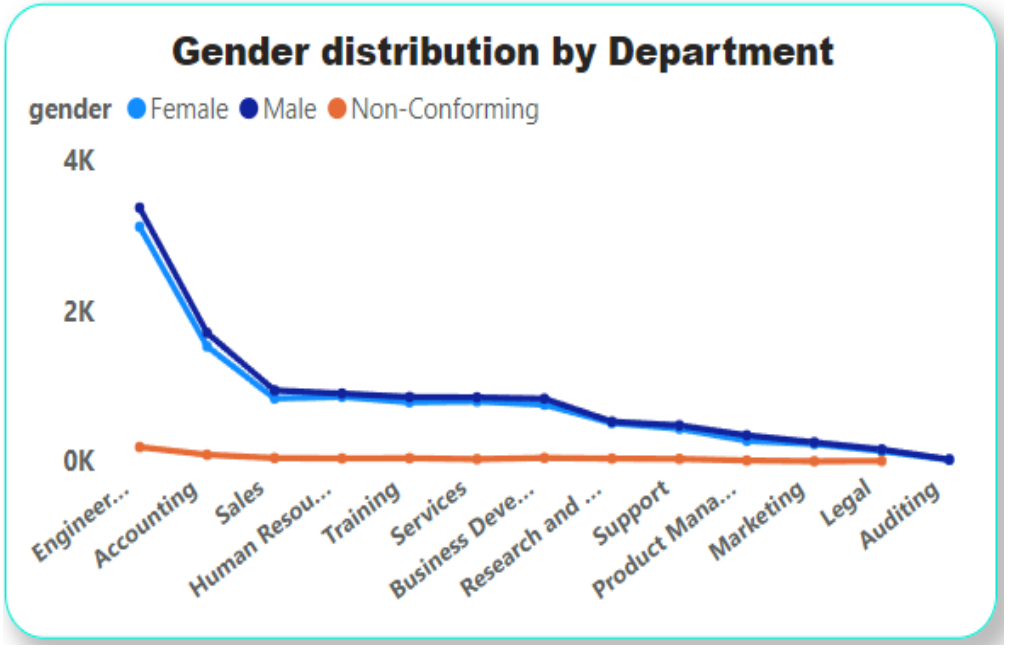
SOLUTION APPROACH

- What is the gender distribution vary across department ?

```
33
34 •  select department , gender, count(*) from hrr
35    group by gender, department
36    order by department;
37
38
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

department	gender	count(*)
Accounting	Male	1711
Accounting	Female	1531
Accounting	Non-Conforming	91
Auditing	Male	28
Auditing	Female	24
Business Development	Male	836
Business Development	Non-Conforming	49
Business Development	Female	757
Engineering	Non-Conforming	193
Engineering	Female	3120
Engineering	Male	3373
Human Resources	Male	904
Human Resources	Non-Conforming	42
Human Resources	Female	861
Legal	Female	140
Legal	Male	162



- Female and Male are more in each departments compare to Non –Conforming genders.

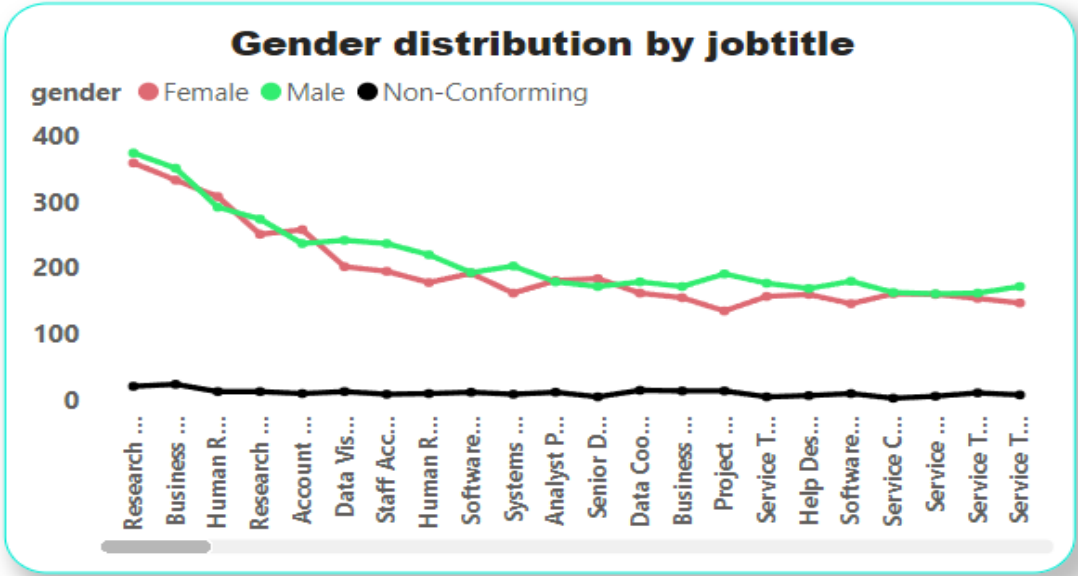
SOLUTION APPROACH

- What is the gender distribution vary across jobtitle ?

```
37
38 •  select jobtitle, gender, count(*) from hrr
39    group by gender, jobtitle
40    order by jobtitle;
41
42
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

jobtitle	gender	count(*)
Account Coordinator	Male	2
Account Executive	Female	258
Account Executive	Male	237
Account Executive	Non-Conforming	10
Account Manager	Non-Conforming	5
Account Manager	Female	102
Account Manager	Male	107
Accountant I	Female	30
Accountant I	Non-Conforming	3
Accountant I	Male	46
Accountant II	Female	48
Accountant II	Male	39
Accountant II	Non-Conforming	2
Accountant III	Non-Conforming	3
Accountant III	Male	38
Accountant III	Female	45



- Female and Male are more in each job titles compare to Non –Conforming genders.

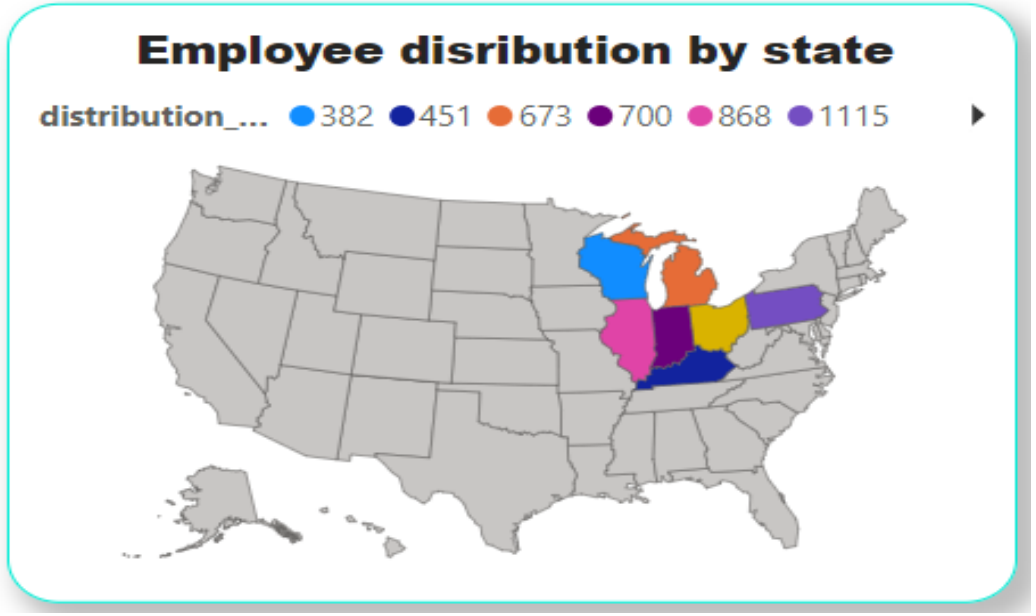
SOLUTION APPROACH

- What is the distribution of employees across locations by State ?

```
42    -- What is the distribution of employees across locations by city and state?
43 •  select location_state,count(*) as distribution_rate
44    from hrr
45    group by  location_state;
46
47
48
49
50
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

location_state	distribution_rate
Ohio	18025
Michigan	673
Pennsylvania	1115
Wisconsin	382
Illinois	868
Indiana	700
Kentucky	451



- Ohio states contains the most employees.

- **Note : Currently I don't have a Microsoft business account for power BI so maps won't display correctly, apologies for the maps.**

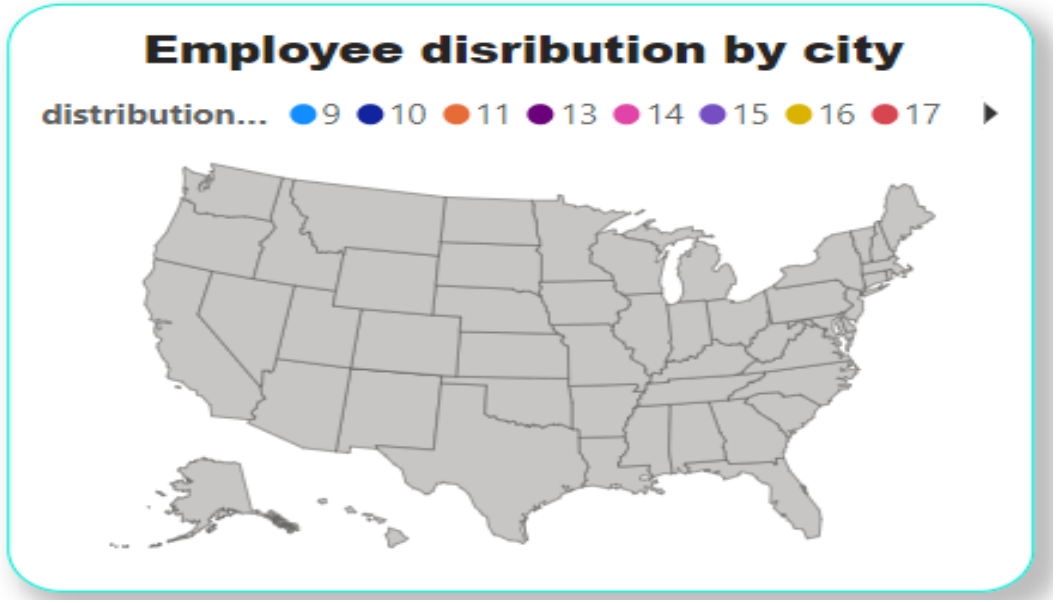
SOLUTION APPROACH

- What is the distribution of employees across locations by city ?

```
47 •  select location_city,count(*) as distribution_rate
48    from hrr
49    group by  location_city;
50
51
52
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

location_city	distribution_rate
Milwaukee	168
Harrisburg	96
Springfield	168
Dayton	202
Aurora	22
Fort Wayne	143
Lexington	199
Louisville	228
Mansfield	22
Cincinnati	285



- Milwaukee city contains the most employees.

- **Note : Currently I don't have a Microsoft business account for power BI so maps won't display correctly, apologies for the maps.**

CONCLUSION

- A Report was built for Stakeholders depicting their various questions.
- The report was cleaned , transformed and then analyzed to get the relevant answers .
- Each Fields were calculated and visual representation were made for better understanding of the analysis.
- This report can be used for both high-level and in-depth analysis of Employees.
- This report is just a pseudo/sample report and does not contain real information of any company out in the market.

THANK YOU!