Data analysis on Human Resource using MYSQL AND POWER BI

A report and note made by Tenzin Delek

Data Used

Data - HR Data with over 5000 rows from the year 2000 to 2020.

Data Cleaning & Analysis - MySQL Workbench

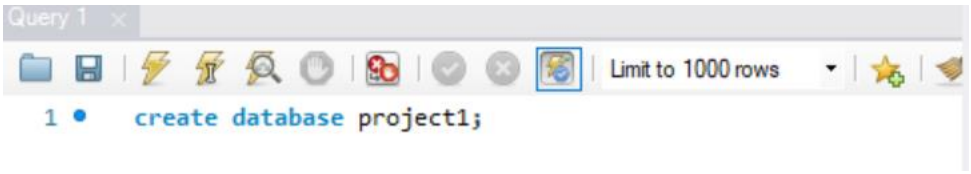
Data Visualization – PowerBI

Questions for Analysing

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. How does the gender distribution vary across departments and job titles?
6. What is the distribution of job titles across the company?
7. What is the distribution of employees across locations by state?

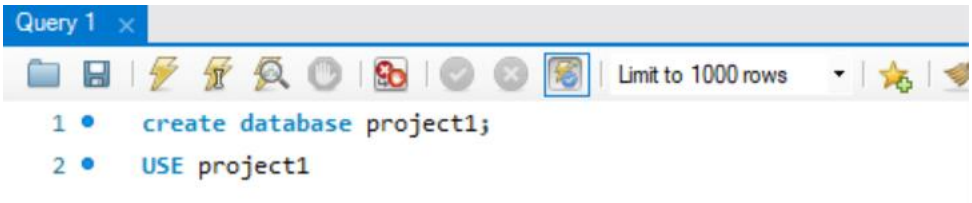
Steps-

Created database (project1) in workbench



Data imported from HR datasets excel to mysql

Select the database



A simple query run to get all the data

```
create database project1;
USE project1;
select * from hr;
```

ï»¿id	first_name	last_name	birthdate	gender	race	department	jobtitle	location	hire_date	termdate	location_
00-0037846	Kimmy	Walczynski	06-04-91	Male	Hispanic or Latino	Engineering	Programmer Analyst I	Headquarters	1/20/2002		Cleveland
00-0041533	Ignatius	Springett	6/29/1984	Male	White	Business Development	Business Analyst	Headquarters	04-08-19		Cleveland
00-0045747	Corbie	Bittlestone	7/29/1989	Male	Black or African American	Sales	Solutions Engineer Manager	Headquarters	10-12-10		Cleveland
00-0055274	Baxy	Matton	9/14/1982	Female	White	Services	Service Tech	Headquarters	04-10-05		Cleveland
00-0076100	Terrell	Suff	04-11-94	Female	Two or More Races	Product Management	Business Analyst	Remote	9/29/2010	2029-10-29 06:09:38 UTC	Flint
00-0116166	Kacie	Offiler	1/18/1971	Male	Asian	Engineering	Developer III	Headquarters	09-01-18		Cleveland
00-0363185	Sandro	Admans	11/19/1979	Male	Two or More Races	Product Management	Quality Engineer	Headquarters	11-08-12		Cleveland
00-0380704	Eugene	Lehrahan	10/14/1988	Female	Black or African American	Engineering	Developer I	Headquarters	6/27/2007		Cleveland
00-0381660	Wainwright	Corfield	12/13/1996	Male	Asian	Engineering	Business Systems Development Analyst	Headquarters	2/20/2001	2008-12-05 01:21:48 UTC	Cleveland
00-0419202	Dyann	Isoldi	3/27/1980	Male	Two or More Races	Engineering	Web Developer I	Headquarters	1/27/2005		Cleveland
00-0472287	Grantley	Oret	09-06-75	Male	Two or More Races	Services	Service Tech II	Headquarters	11-01-04		Cleveland
00-0472832	Elmore	Worner	01-07-66	Female	White	Engineering	Business Systems Development Analyst	Headquarters	12-05-00		Cleveland
00-0566380	Dud	Brain	3/17/1984	Male	Two or More Races	Business Development	Business Analyst	Headquarters	9/17/2008		Cleveland
00-0571075	Aguie	Conford	11-02-71	Male	White	Business Development	Research Assistant II	Headquarters	11/25/2015		Cleveland
00-0624189	Katerina	Rosborough	8/20/1967	Male	Hispanic or Latino	Engineering	Analyst Programmer	Headquarters	5/17/2019		Cleveland
00-0715212	Alida	Longley	1/28/1973	Female	American Indian or Alask...	Accounting	Staff Accountant III	Headquarters	02-04-02		Cleveland

Data cleaning-the id column name has been wrongly typed

```
create database project1;
USE project1;
select * from hr;
alter table hr
change column ï»¿id emp_id varchar(20) null;
```

emp_id	first_name	last_name	birthdate	gender	race	department	jobtitle	location	hire_date	termdate	location_
00-0037846	Kimmy	Walczynski	06-04-91	Male	Hispanic or Latino	Engineering	Programmer Analyst I	Headquarters	1/20/2002		Clevelanc
00-0041533	Ignatius	Springett	6/29/1984	Male	White	Business Development	Business Analyst	Headquarters	04-08-19		Clevelanc
00-0045747	Corbie	Bittlestone	7/29/1989	Male	Black or African American	Sales	Solutions Engineer Manager	Headquarters	10-12-10		Clevelanc
00-0055274	Baxy	Matton	9/14/1982	Female	White	Services	Service Tech	Headquarters	04-10-05		Clevelanc
00-0076100	Terrell	Suff	04-11-94	Female	Two or More Races	Product Management	Business Analyst	Remote	9/29/2010	2029-10-29 06:09:38 UTC	Flint
00-0116166	Kacie	Offiler	1/18/1971	Male	Asian	Engineering	Developer III	Headquarters	09-01-18		Clevelanc
00-0363185	Sandro	Admans	11/19/1979	Male	Two or More Races	Product Management	Quality Engineer	Headquarters	11-08-12		Clevelanc
00-0380704	Eugene	Lehrahan	10/14/1988	Female	Black or African American	Engineering	Developer I	Headquarters	6/27/2007		Clevelanc
00-0381660	Wainwright	Corfield	12/13/1996	Male	Asian	Engineering	Business Systems Development Analyst	Headquarters	2/20/2001	2008-12-05 01:21:48 UTC	Clevelanc
00-0419202	Dyann	Isoldi	3/27/1980	Male	Two or More Races	Engineering	Web Developer I	Headquarters	1/27/2005		Clevelanc
00-0472287	Grantley	Oret	09-06-75	Male	Two or More Races	Services	Service Tech II	Headquarters	11-01-04		Clevelanc
00-0472832	Elmore	Worner	01-07-66	Female	White	Engineering	Business Systems Development Analyst	Headquarters	12-05-00		Clevelanc
00-0566380	Dud	Brain	3/17/1984	Male	Two or More Races	Business Development	Business Analyst	Headquarters	9/17/2008		Clevelanc
00-0571075	Aguie	Conford	11-02-71	Male	White	Business Development	Research Assistant II	Headquarters	11/25/2015		Clevelanc
00-0624189	Katerina	Rosborough	8/20/1967	Male	Hispanic or Latino	Engineering	Analyst Programmer	Headquarters	5/17/2019		Clevelanc
00-0715212	Alida	Longley	1/28/1973	Female	American Indian or Alask...	Accounting	Staff Accountant III	Headquarters	02-04-02		Clevelanc

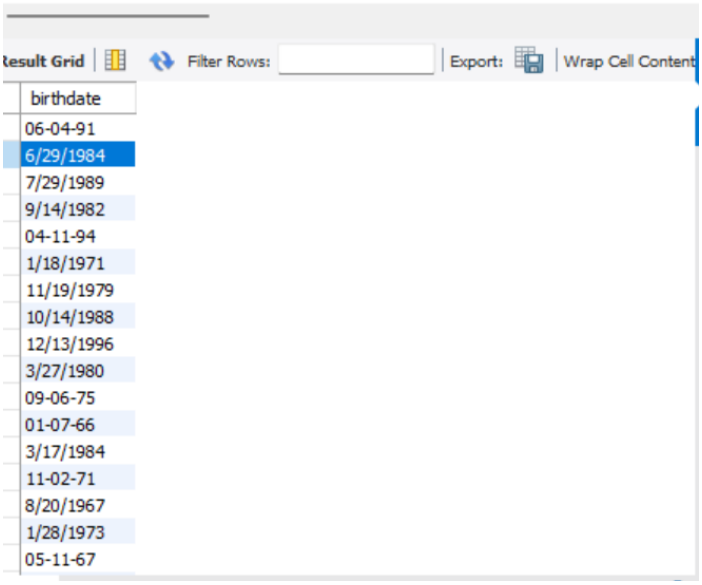
Describing each column and its type

```
describe hr;
```

Field	Type	Null	Key	Default	Extra
emp_id	varchar(20)	YES		NULL	
first_name	text	YES		NULL	
last_name	text	YES		NULL	
birthdate	text	YES		NULL	
gender	text	YES		NULL	
race	text	YES		NULL	
department	text	YES		NULL	
jobtitle	text	YES		NULL	
location	text	YES		NULL	
hire_date	text	YES		NULL	
termdate	text	YES		NULL	
location_city	text	YES		NULL	
location_st...	text	YES		NULL	

All the birthdates are in different format

```
9 •   select birthdate from hr;
```



birthdate
06-04-91
6/29/1984
7/29/1989
9/14/1982
04-11-94
1/18/1971
11/19/1979
10/14/1988
12/13/1996
3/27/1980
09-06-75
01-07-66
3/17/1984
11-02-71
8/20/1967
1/28/1973
05-11-67

So we change it to one

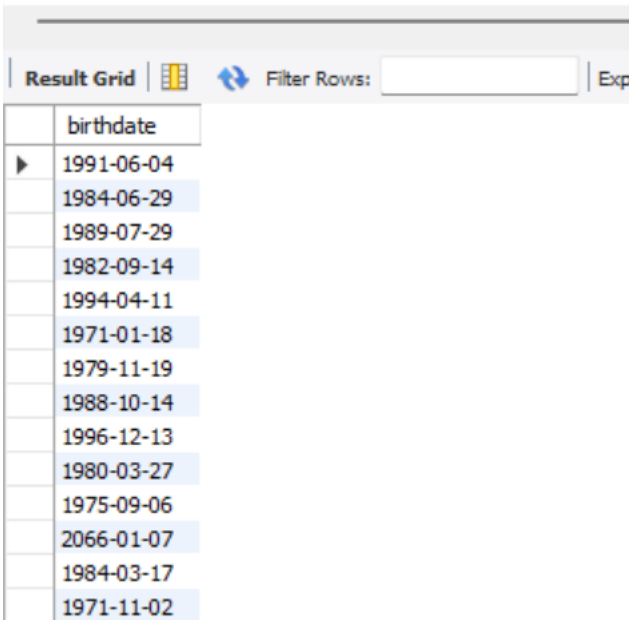
```
•   set sql_safe_updates=0;
•   update hr set birthdate=case
    when birthdate like '%/%'then date_format(str_to_date(birthdate,'%m/%d/%Y'),'%Y-%m-%d')
    when birthdate like '%-%'then date_format(str_to_date(birthdate,'%m-%d-%Y'),'%Y-%m-%d')
    else null
    end;

•   select birthdate from hr;
```

It is important to note that when we try to change something from the datebase we need to set the safe update to 0 to makes changes in the datasets as in default the datasets are being protected

Now the birthdate are all in same formats-

```
18 •   select birthdate from hr;
```



birthdate
1991-06-04
1984-06-29
1989-07-29
1982-09-14
1994-04-11
1971-01-18
1979-11-19
1988-10-14
1996-12-13
1980-03-27
1975-09-06
2066-01-07
1984-03-17
1971-11-02

Now the current birthdate datatype is set to text when imported so we need to alter it to date

```
18 •    alter table hr
19      modify column birthdate date;
```

Similar to that we did the same for the hire_date

```
21 • update hr set hire_date=case
22   when hire_date like '%/%'then date_format(str_to_date(hire_date,'%m/%d/%Y'),'%Y-%m-%d')
23   when hire_date like '%-%'then date_format(str_to_date(hire_date,'%m-%d-%Y'),'%Y-%m-%d')
24   else null
25   end;
26 • select hire_date from hr;
```

Now in termdate we don't want the timestamp to be shown from the date

```
29 •    select termdate from hr;
```

Result Grid | Filter Rows:

termdate
2029-10-29 06:09:38 UTC
2008-12-05 01:21:48 UTC

```
30 •    update hr
31      set termdate=date(str_to_date(termdate,'%Y-%m-%d %H:%i:%s UTC'))
32      WHERE termdate is not null and termdate !='';
33
34 •    select termdate from hr;
```

So by using above query the time is being remove and it is turn into a proper date format

Result Grid | Filter Rows: | Export:

termdate
2029-10-29
2008-12-05

Now after that we want to add a new column call age where the initial value will be null so we use

A timestampdiff to calculate the age of each person in the date

```
29 •  alter table hr
30    add column age int;
31
32 •  update hr
33    set age=timestampdiff(YEAR,birthdate,CURDATE());
34 •  SELECT age from hr;
```

Result Grid | Filter Rows: | Export:

age
32
39
34
41
29
52
44
35
27
43

But when we closely see the values we get to know that some age are coming in negative values.

```
36 •  select min(age) as youngest,
37    max(age) as oldest from hr;
```

Result Grid | Filter Rows:

youngest	oldest
-45	58

We can count the values that are below the age of 18

```
39 •  select count(*) from hr where age <18;
```

Result Grid | Filter Rows: | Export: | Wrap

count(*)
268

NOW WE COMES TO THE DATA ANAYLSIS PART

1. WHAT IS THE GENDER BREAKDOWN OF EMPLOYEE IN THE COMPANY

```
42 •  select gender, count(*) as count
43    from hr where age >=18 group by gender;
```

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

gender	count
Male	3015
Female	2677
Non-Conforming	154

2. What is the race/ethnicity breakdown of employee in the company

```
45 •  select race,count(*) as count from hr
46    where age>=18 group by race order by count(*) desc;
```

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

race	count
White	1644
Two or More Races	965
Asian	949
Black or African American	921
Hispanic or Latino	725
American Indian or Alaska Native	348
Native Hawaiian or Other Pacific Islander	294

3. What is the age distribution of employee in the company?

```
48 •  select case
49    when age>=18 and age<=24 then '18-24'
50    when age>=25 and age<=34 then '25-34'
51    when age>=35 and age<=44 then '35-44'
52    when age>=45 and age<=54 then '45-54'
53    when age>=55 and age<=64 then '55-64'
54    else '65+'
55    end as age_group,
56    count(*) as count from hr where age>=18 group by
57    age_group order by age_group;
58
```

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

age_group	count
18-24	640
25-34	1638
35-44	1669
45-54	1588
55-64	311

4. How many employees work at headquarter versus remote location?

```
71 •  select location,count(*) as count
72    from hr where age>=18 group by location;
73
```

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

location	count
Headquarters	4388
Remote	1458

5. How does the gender distribution vary across the departments and job title?

```
74 •  select department,gender,count(*) as count
75    from hr where age>=18 group by department,gender
76    order by department ;
77
```

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

department	gender	count
Accounting	Female	415
Accounting	Male	438
Accounting	Non-Conforming	16
Auditing	Female	7
Auditing	Male	11
Business Development	Female	196
Business Development	Male	239

6. What is the distribution of job title across the company?

```
78 •  select jobtitle,count(*) as count from hr
79    where age >=18 group by jobtitle order by
80    jobtitle desc;
81
```

Result Grid | Filter Rows: | Export:

jobtitle	count
Web Developer IV	17
Web Developer III	27
Web Developer II	19
Web Developer I	28
Web Designer IV	3
Web Designer III	2
Web Designer I	11

7. What is the distribution of employee across location by state?

```
82 •  select location_state,count(*) as count from hr
83    where age>=18 group by location_state order by
84    count desc;
```

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

location_state	count
Ohio	4742
Pennsylvania	302
Illinois	225
Michigan	196
Indiana	159
Kentucky	113
Wisconsin	109

Now after this we will visualize our query in power bi

Conclusion

- There are more male employees
- White race is the most dominant while Native Hawaiian and American Indian are the least dominant.
- The youngest employee is 20 years old and the oldest is 57 years old
- 5 age groups were created (18-24, 25-34, 35-44, 45-54, 55-64). A large number of employees were between 25-34 followed by 35-44 while the smallest group was 55-64.
- A large number of employees work at the headquarters versus remotely.
- The average length of employment for terminated employees is around 7 years.
- The gender distribution across departments is fairly balanced but there are generally more male than female employees.
- A large number of employees come from the state of Ohio.

Limitations

- Some records had negative ages and these were excluded during querying (967 records). Ages used were 18 years and above.