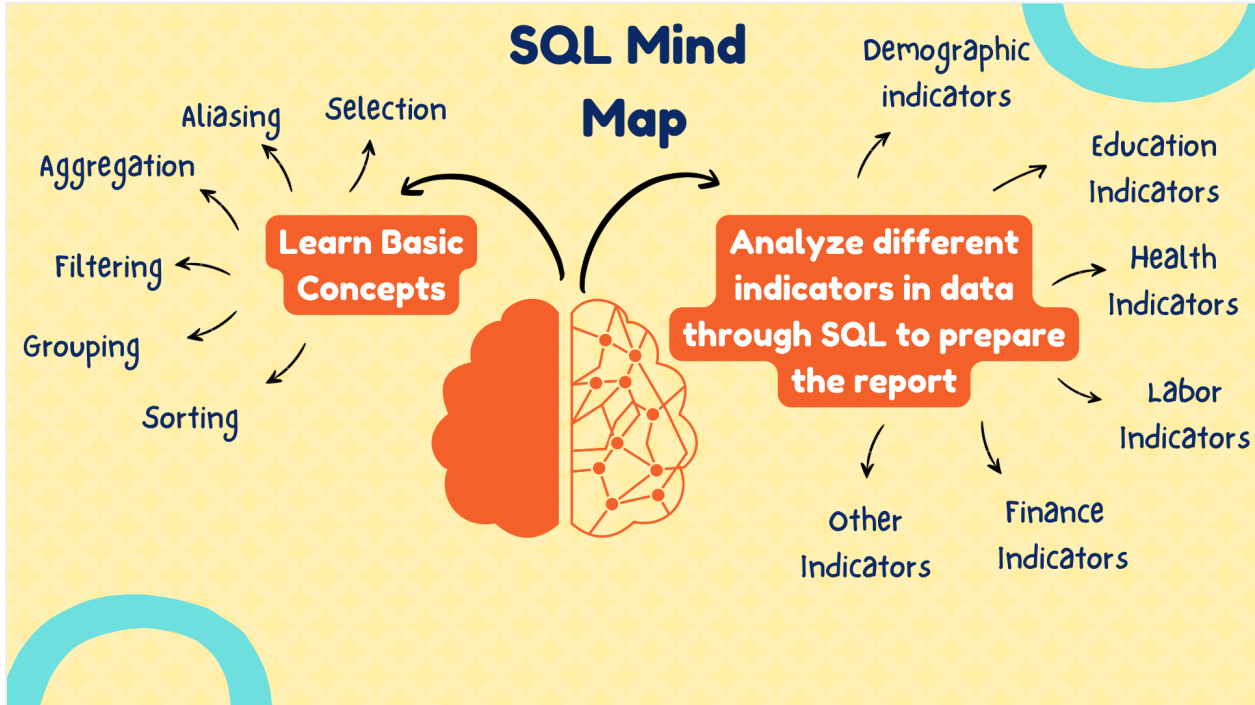# Gender Case Study through SQL

We will first go through basic concepts in SQL and then move to [solve report questions](#).

Basic SQL Concepts

Let's first learn some basic concepts for extracting and analyzing data in SQL

Question 1: **Level: Easy**

Get an overview of the dataset and output 100 rows from the dataset

Method:

- The "**SELECT * "** function Selects all rows & columns from the table
  - If specific columns are required in the output they should be specified after the **SELECT** statement
- The **"FROM"** keyword defines the table from which we want to query the data from
- The **"LIMIT"** keyword defines how many Rows of data we want to display from the results

Answer 1:

```
SELECT
  *
FROM
  career_nub.gender_data
LIMIT 100;
```

Output 1:

We extracted 100 rows of data with all 5 columns in the dataset.

Question 2: `Level: Easy`

Get specific columns from the data.
Get the Country, Year, Indicator, and Value from the dataset *(only 100 rows)*

Method:

- The "**SELECT**" function chooses the specified columns from the table
  - Since we need specific columns, we can specify the name of each column after the **SELECT** statement
- The **"FROM"** keyword defines the table from which we want to query the data from
- The **"LIMIT"** keyword defines how many Rows of data we want to display from the results

Answer 2:

```
SELECT
  country,
  year,
  indicator,
  value
FROM
  career_nub.gender_data
LIMIT 100;
```

Output 2:

We extracted 100 rows of data with columns - country, year, indicator, and value - from the dataset.

Question 3: <mark>Level: Easy</mark>

Get specific columns from the data.

Get the Country, Year, Indicator, and Value from the dataset *(only 100 rows)*, and assign different column names to the Indicator and Value columns.

Method:

- **ALIASING**
  - "**AS**" function: We can define and use aliases for table names and column names using the "AS" function

Answer 3:

```
SELECT
  country,
  year,
  indicator AS metric,
  value AS number
FROM
  career_nub.gender_data AS data
LIMIT 100;
```

Output 3:

We extracted 100 rows of data with columns - country, year, indicator, and value - from the dataset while renaming some of the columns.

Question 4: <mark>Level: Easy</mark>

How many total rows are present in the dataset?

Method:

- The "**COUNT**" function is used to count the number of rows returned in a SELECT statement
  - **COUNT(*)** in bracket counts all rows from a specified table
- Use the **AS** function to alias this column as total_rows

Answer 4:

```
SELECT
      COUNT(*) AS total_rows
```

FROM career_nub.gender_data;

We see that there are 64,534 rows in the dataset.

Question 5: <mark>Level: Easy</mark>

How many countries does our dataset contain?

Method:

- Since each country will have multiple rows of data (for different indicatorts and years), we need to calculate the distinct number of countries from the dataset
- The **COUNT()** function with the **DISTINCT** clause eliminates the repetitive appearance of the same data
  - Use **COUNT(DISTINCT column_name)** to calculate the number of distinct values of a column
- Use the **AS** function to alias this column as total_countries

Answer 5:

```
SELECT
  COUNT(DISTINCT country) AS total_countries
FROM
  career_nub.gender_data;
```

Output 5:

We see that there are 193 unique countries in the dataset.

Question 6: <mark>Level: Easy</mark>

List down all the countries present in our dataset.

Method:

- Since each country will have multiple rows of data, we need to list the distinct countries from the dataset
- **DISTINCT** is used with the **SELECT** statement to eliminate all duplicate records and fetch only unique records

○ Use **DISTINCT column_name** to list the distinct values of a column
- Use the **AS** function to alias this column as total_countries

Answer 6:

SELECT
  DISTINCT country
FROM
  career_nub.gender_data;

Output 6:

We see that there are 193 unique countries in the dataset and can see the names of each of these countries.

Question 7: <mark>Level: Easy</mark>

How many indicators does our dataset contain?

Method:

- Since each indicator will have multiple rows of data, we need to calculate the distinct number of indicators from the dataset
- **COUNT()** function with the **DISTINCT** clause eliminates the repetitive appearance of the same data
  ○ Use **COUNT(DISTINCT column_name)** to calculate the number of distinct values of a column

Answer 7:

SELECT
  COUNT(DISTINCT indicator) AS total_indicators
FROM
  career_nub.gender_data;

Output 7:

We see that there are 74 unique indicators in the dataset.

Question 8: Level: Easy

List down all the indicators present in our dataset.

Method:

- Since each indicator will have multiple rows of data, we need to list the distinct indicators from the dataset
- **DISTINCT** is used in with the **SELECT** statement to eliminate all duplicate records and fetch only unique records
  - Use **DISTINCT column_name** to list the distinct values of a column

Answer 8:

SELECT
  DISTINCT indicator
FROM
  career_nub.gender_data;

Output 8:

We see that there are 74 unique indicators in the dataset and can see the names of each of these indicators.

CONDITIONAL FILTERING: Filter data based on a condition

Question 9: Level: Easy

List all indicators within the category of Education

Method:

- Since we only want indicators from a certain category, we will have to filter the dataset using this condition
  - The "**WHERE**" function is used to filter the data based on conditions
  - Use **WHERE column_name = condition value** to filter the dataset
- Also, since each indicator will have multiple rows of data, we need to list the distinct indicators from the dataset
- **DISTINCT** is used in with the **SELECT** statement to eliminate all duplicate records and fetch only unique records
  - Use **DISTINCT column_name** to list the distinct values of a column

Answer 9:

```
SELECT
  DISTINCT indicator
FROM
        career_nub.gender_data
WHERE
        category = 'Education';
```

Output 9:

We see that there are 12 indicators in the Education category. We get the list of all indicators within the category Education using this code.

Question 10: **Level: Easy**

Get all data post-2019

Method:

- "**SELECT * "** Selects all rows & columns from the the table
- Since we only want indicators from a certain category, we will have to filter the dataset using this condition
    - The "**WHERE**" function is used to filter the data based on conditions
    - Use **WHERE column_name > condition value** to filter the dataset

Answer 10:

```
SELECT
 *
FROM
career_nub.gender_data
WHERE
 year > 2019;
```

Output 10:

We filter out the data for all years post-2019 and we see there are 3,845 rows of data.

Question 11: **Level: Easy**

Get all data from these five Nordic countries:

- Denmark, Finland, Iceland, Norway, and Sweden

Method:

- "**SELECT * **" Selects all rows & columns from the the table
- Since we only want all data for only certain countries, we will have to filter the dataset using this condition
  - The "**WHERE**" function is used to filter the data based on conditions
  - Use **WHERE column_name IN ('condition value1', 'condition value2' …)** to filter the dataset
  - **WHERE column_name IN ('condition value1', 'condition value2' …)** is used when there are multiple values to be filtered within the same column

Answer 11:

SELECT
 *
FROM
career_nub.gender_data
WHERE
  country IN ('Denmark', 'Finland', 'Iceland', 'Norway', 'Sweden');

Output 11:

We filter out the data for these Nordic countries and we see there are 1,824 rows of data.

MULTIPLE CONDITIONAL FILTERING: Filter data based on multiple conditions

Question 12: Level: Medium

Get all data from 2015 onwards for India

Method:

- Since we only want data from a given country for a certain time period, we will have to filter the dataset using these conditions
  - The "**WHERE**" function is used to filter the data based on conditions
  - The "**AND**" Operator is used to stack multiple conditions together in the **WHERE** clause which will output a result only if all conditions are met
  - Use the below function to filter the dataset
    WHERE

column_name1 = condition value
AND column_name2 >= condition value

Answer 12:

SELECT
  *
FROM
career_nub.gender_data
WHERE
  country = ('India')
  AND year >= 2015;

Output 12:

We filter out the data for all years post-2015 in India and we see there are 238 rows of data.

Question 13: Level: Medium

Get all rows of data which are either in Australia or from the Health category

Method:

- Since we only want data from either a given country or a particular category, we will have to filter the dataset using these conditions
    - The "**WHERE**" function is used to filter the data based on conditions
    - The "**OR**" Operator is used to stack multiple conditions together in the **WHERE** clause which will output a result **if either of the conditions is met**
    - Use the below function to filter the dataset
      WHERE
          column_name1 = condition value
          OR column_name2 = condition value

Answer 13:

SELECT
  *
FROM
career_nub.gender_data
WHERE
  country = ('Australia')
  OR category = 'Health';

Output 13:

We filter out the data for rows in either Australia or in the Health category and we see there are 13,858 rows of data.

Question 14: **Level: Easy**

What is the range of time period in the dataset?

Method:

- The **MIN** & **MAX** are functions used to determine the maximum and the minimum value of a particular column
    - These functions can be applied to integer datatype columns
- Use the below function to get the maximum and minimum values
  MIN(column_name) AS min_value,
  MAX(column_name) AS max_value

Answer 14:

SELECT
  MIN(year) as start_year,
  MAX(year) as end_year
FROM
  career_nub.gender_data;

Output 14:

We see that the data is from 2012 to 2020.

Report Questions

Let's now move to some of the advanced topics in SQL and solve questions to prepare our report.

Report Question 1: Understanding demographic indicators

Report Question 1.1: <mark>Level: Easy</mark>

Find out the Total female population per year
  ● Indicator Names: Population, female

Method:

- ● Since we only want data for a particular indicator we will filter the dataset using
     WHERE column_name = condition value
- ● The "**GROUP BY**" function is used with the SELECT statement to arrange identical data into groups
    - ○ Since we want the sum of values of a column by each year we will use the GROUP BY function
- ● The **SUM** function is used to calculate the total value of a particular column

- ○ This function can be applied to integer datatype columns
- ○ Use the below function to get the sum of value
  SUM(column_name) AS total_value

```
SELECT
  year,
  SUM(value) AS total_female_population
FROM
career_nub.gender_data
WHERE
  indicator = 'Population, female'
GROUP BY 1
ORDER BY 1
```

Report Output 1.1:

The total female population per year increased from 3.5Bn in 2012 to 3.8Bn in 2020

Report Question 1.2: <mark>Level: Medium</mark>

Find the yearly total and % of Female population across all countries
- ● Indicator Names:
  - ○ Population, female
  - ○ Population, male
  - ○ *Total Population = (Population, female) + (Population, male)*

Method:
- ● Since we only want data for particular indicators we will filter the dataset using
  WHERE column_name IN ('value1', 'value2')
- ● To calculate the sum of a column for particular values we can use this function
  SUM(CASE WHEN column_name = condition THEN value END)
- ● We can then divide the female population by the total population to get the % of female population
- ● Since we want the sum of the column with a condition by each year we will use the "**GROUP BY**" function
- ● Use the "**ORDER BY**" function to sort the dataset by the year

Report Answer 1.2:

```
SELECT
  year,
  SUM(CASE WHEN indicator = 'Population, female' THEN value END) AS female_population,
  SUM(value) AS total_population,
  100.0 * SUM(CASE WHEN indicator = 'Population, female' THEN value END) / SUM(value)
AS perc_female_population
FROM
  career_nub.gender_data
WHERE
  indicator IN ('Population, female', 'Population, male')
GROUP BY 1
ORDER BY 1;
```

Report Output 1.2:

We see that the global female population has increased from 3.5Bn or 49.57% of the total population in 2012 to 3.8 Bn or 49.58% of the overall population in 2020

Report Question 1.3: Level: Hard

Find the top 5 countries with the highest percentage of the female population in the year 2020
- Indicator Names:
    - Population, female
    - Population, male

Method:
- Since we only want data for particular indicators for a year we will filter the dataset using
  WHERE column_name1 IN ('value1', 'value2')
  AND column_name2 = value
- To calculate the sum of a column for particular values we can use this function
  SUM(CASE WHEN column_name = condition THEN value END)
- We can then divide the female population by the total population to get the % of female population
- Since we want the sum of the column with a condition by each country we will use the "**GROUP BY**" function
- Use the "**ORDER BY DESC**" function to sort the countries by the percentage of female population in descending order
- Use the "**LIMIT**" function to restrict it to 5 rows to get the top 5 countries

Report Answer 1.3:

```
SELECT
  country,
  SUM(CASE WHEN indicator = 'Population, female' THEN value END) AS female_population,
  SUM(value) AS total_population,
  100.0 * SUM(CASE WHEN indicator = 'Population, female' THEN value END) / SUM(value)
AS perc_female_population
FROM
  career_nub.gender_data
WHERE
  year = 2020
  AND indicator IN ('Population, female', 'Population, male')
GROUP BY 1
ORDER BY 4 DESC
LIMIT 5;
```

Report Output 1.3:

Countries with the highest percentage of female population in 2020 are:
Nepal, Hong Kong, Latvia, Lithuania, and Ukraine

Report Question 1.4: <mark>Level: Medium</mark>

Which country had the highest sex ratio at birth in 2020?
  ● Indicator Name: Sex ratio at birth (male births per female births)

Method:
  ● Since we want data for a particular indicator and a year we will filter the dataset using
    WHERE column_name1 = condition value
          AND column_name2 = condition value
  ● Use the "**ORDER BY DESC**" function to sort the countries by sex ratio at birth in
    descending order
  ● Use the "**LIMIT**" function to restrict to the top row

Report Answer 1.4:

```
SELECT
  *
FROM
  career_nub.gender_data
WHERE
```

```
    indicator = 'Sex ratio at birth (male births per female births)'
    AND year = '2020'
ORDER BY value DESC
LIMIT 1
```

Report Output 1.4:

The sex ratio at birth (male births per female births) was highest for Azerbaijan in 2020

Report Question 2: Understanding Education indicators

Report Question 2.1: **Level: Medium**

Find the bottom 2 countries with the lowest adult female literacy rate in 2019
- Indicator Names:
  - Literacy rate, adult female

Method:
- Since we want data for a particular indicator and a year we will filter the dataset using
  WHERE column_name1 = condition value
        AND column_name2 = condition value
- Use the "**ORDER BY**" function to sort the countries by adult female literacy rate in ascending order
  - *Default for order by function is ascending*
- Use the "**LIMIT**" function to restrict to the top 3 rows

Report Answer 2.1:

```
SELECT
 *
FROM
  career_nub.gender_data
WHERE
  indicator in ('Literacy rate, adult female')
  AND year = 2019
ORDER BY value
LIMIT 2;
```

Report Output 2.1:

We see that the countries with the lowest adult female literacy rate in the year 2019 are Pakistan and Togo.

Report Question 2.2:

Find the change in adult female literacy rate between 2012 and 2020 for Bangladesh
- Indicator Name: Literacy rate, adult female

Method:
- To filter the dataset for an indicator for a particular country and a range of years we can use the "BETWEEN" condition
  WHERE column_name1 BETWEEN value1 AND value2
  AND column_name2 = 'value'
  AND column_name3 = 'value'
- Select all columns using **SELECT ***
- Order the final dataset by the year column using
  ORDER BY column_name

Report Answer 2.2:

```
SELECT
 *
FROM career_nub.gender_data
WHERE
  year BETWEEN 2012 AND 2019
  AND indicator='Literacy rate, adult female'
  AND country = 'Bangladesh'
```

ORDER BY year;

Between 2012 and 2019, the adult female literacy rate increased from 54% to 72% in Bangladesh

Report Question 2.3: Level: Hard

Find the number of Female Children out of primary school as a percentage of overall children out of primary school in Cameroon by each year
- Indicator Names:
  - Children out of school, primary, female
  - Children out of school, primary, male

Method:
- Since we only want data for particular indicators we will filter the dataset using **WHERE column_name1 IN ('value1', 'value2')**
- To calculate the sum of a column for particular values we can use this function **SUM(CASE WHEN column_name = condition THEN value END)**
- We can then divide the female indicator by total to get the % of female out of primary school
- Since we want the sum of the column with a condition by each country we will use the "**GROUP BY**" function
- Use "**ORDER BY**" function to sort countries by the % of female out of primary school

Report Answer 2.3:

```
SELECT
 year,
 SUM(CASE WHEN indicator = 'Children out of school, primary, female' THEN value END) AS female_OutOfPrimarySchool_population,
 SUM(value) AS total_children_OutOfPrimarySchool_population,
 100.0 * SUM(CASE WHEN indicator = 'Children out of school, primary, female' THEN value END) / SUM(value) AS perc_female_OutOfPrimarySchool_population
FROM
 career_nub.gender_data
WHERE
 indicator IN ('Children out of school, primary, female', 'Children out of school, primary, male')
 AND country = 'Cameroon'
GROUP BY 1;
```

The percentage of female children out of school *(of the total children out of primary school)* in Cameroon has improved from 92% in 2016 to 76% in 2019.

Report Question 3: Understanding Health indicators

Report Question 3.1: Level: Hard

Which are the best/worst countries based on the female & male Life expectancy among SAARC countries in 2020
Indicator Names:
- Life expectancy at birth, female
- Life expectancy at birth, male

SAARC countries: Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka

Method:
- Use **WHERE column_name IN ('condition value1','condition value2'...)** function to filter the dataset for the countries and for a particular year
- To calculate the average of a column for particular values we can use this function **AVG(CASE WHEN column_name = condition THEN value END)**
- Since we want the average of the column with a condition by each country we will use the "**GROUP BY**" function
- Use the "**ORDER BY**" function to sort the dataset by the highest value

Report Answer 3.1:

```
SELECT
  country,
  AVG(CASE WHEN indicator = 'Life expectancy at birth, female' THEN value END) AS "Life
expectancy at birth for females",
  AVG(CASE WHEN indicator = 'Life expectancy at birth, male' THEN value END) AS "Life
expectancy at birth for males"
FROM
  career_nub.gender_data
WHERE
  country IN ('Afghanistan', 'Bangladesh', 'Bhutan', 'India', 'Maldives', 'Nepal', 'Pakistan', 'Sri
Lanka')
  AND year = 2020
GROUP BY 1
ORDER BY 2 DESC;
```

Report Output 3.1:

We see that among SAARC countries, Maldives has the highest life expectancy for both females at 81 years and males at 77 yrs, with the lowest being in Afghanistan for both females (67) and males (64).

Report Question 3.2: Level: Hard

Compare the female and male Under-5 mortality rates in the year 2020 among SAARC countries
Indicator Names:
- Mortality rate, under-5, female
- Mortality rate, under-5, male

SAARC countries: Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka

Method:

- Use **WHERE column_name IN ('condition value1','condition value2'...)** function to filter the dataset for the countries
- To calculate the average of a column for particular values we can use this function **AVG(CASE WHEN column_name = condition THEN value END)**
- Since we want the average of the column with a condition by each country we will use the "**GROUP BY**" function
- Use the "**ORDER BY DESC**" function to sort the dataset by the highest value

Report Answer 3.2:

SELECT
  country,
  AVG(CASE WHEN indicator = 'Mortality rate, under-5, female' THEN value END) AS "Average under-5 mortality rate for females ",
  AVG(CASE WHEN indicator = 'Mortality rate, under-5, male' THEN value END) AS "Average under-5 mortality rate for males "
FROM
  career_nub.gender_data
WHERE
  country IN ('Afghanistan', 'Bangladesh', 'Bhutan', 'India', 'Maldives', 'Nepal', 'Pakistan', 'Sri Lanka')
  AND year = 2020
GROUP BY 1
ORDER BY 2 DESC;

Report Output 3.2:

Pakistan has the worst under-5 mortality rate for both females and males with Maldives being the best for both genders.

Report Question 3.3: `Level: Medium`

Find the countries that had the highest Prevalence of HIV among females in the age group 15-24) in the year 2020:
- Indicator Name: Prevalence of HIV, female (% ages 15-24)

Method:
- Since we only want data from a given year for an indicator, we will filter the dataset using **WHERE**
  > **column_name1 = condition value**
  > **AND column_name2 >= condition value**
- The "**ORDER BY**" function is used to sort the result set in ascending or descending order based on the mentioned column
  - Since we want the worst countries that means a high % of HIV prevalence, we will use the DESC clause which orders the dataset in a descending order
  - ORDER BY column_name DESC
- Let's use the "**LIMIT**" function to restrict it to 5 rows to get the bottom 5 countries

Report Answer 3.3:

```
SELECT
  country,
  indicator,
  value
FROM
  career_nub.gender_data
WHERE
  indicator = 'Prevalence of HIV, female (% ages 15-24)'
  AND year = 2020
ORDER BY value DESC
LIMIT 5;
```

Report Output 3.3:

The bottom 5 countries in terms of the prevalence of HIV in females are South Africa, Lesotho, Botswana, Mozambique, and Zambia. We can share this analysis with the health vertical to focus on these countries and take steps to reduce HIV prevalence.

Report Question 4: Understanding Labor indicators

Report Question 4.1: Level: Hard

How has the percentage of female employment in Agriculture changed from 2012 to 2019
- Indicator Names:
    - Employment in agriculture, female
    - Population employed, female

Method:

- Since we only want data for a range of years, we can filter the dataset using **WHERE column_name1 BETWEEN value1 AND value2**
- To calculate the sum of a column for particular values we can use this function **SUM(CASE WHEN column_name = condition THEN value END)**
- We can then divide the female employed in agriculture by total females employed to get the % of female employed in agriculture
- Since we want the sum of the column with a condition by each year we will use the "**GROUP BY**" function
- Use "**ORDER BY**" function to sort the dataset by year

Report Answer 4.1:

```
SELECT
  year,
  SUM(CASE WHEN indicator = 'Employment in agriculture, female' THEN value END) AS
"Females Employed in Agriculture",
  SUM(CASE WHEN indicator = 'Population employed, female' THEN value END) AS "Total
females employed",
  100*SUM(CASE WHEN indicator = 'Employment in agriculture, female' THEN value END) /
SUM(CASE WHEN indicator = 'Population employed, female' THEN value END) AS "% Female
Employment in Agriculture"
FROM
  career_nub.gender_data
WHERE
  year BETWEEN 2012 AND 2019
GROUP BY 1
ORDER BY 1;
```

Report Output 4.1:

Employment in agriculture for females has reduced from 32% to 27% between 2012 and 2019

Report Question 4.2: <mark>Level: Hard</mark>

How has the percentage of female employment in the Industry sector changed from 2012 to 2019

- Indicator Names:
    - Employment in industry, female
    - Population employed, female

Method:

- Since we only want data for a range of years, we can filter the dataset using
- WHERE column_name1 BETWEEN value1 AND value2
- To calculate the sum of a column for particular values we can use this function
- SUM(CASE WHEN column_name = condition THEN value END)
- We can then divide the female employed in industry by total females employed to get the % of female employed in industry
- Since we want the sum of the column with a condition by each year we will use the "GROUP BY" function
- Use "ORDER BY" function to sort the dataset by year

Report Answer 4.2:

```
SELECT
  year,
  SUM(CASE WHEN indicator = 'Employment in industry, female' THEN value END) AS
"Females Employed in Industry",
  SUM(CASE WHEN indicator = 'Population employed, female' THEN value END) AS "Total
females employed",
  100*SUM(CASE WHEN indicator = 'Employment in industry, female' THEN value END) /
SUM(CASE WHEN indicator = 'Population employed, female' THEN value END) AS "% Female
Employment in Industry"
FROM
  career_nub.gender_data
WHERE
  year BETWEEN 2012 AND 2019
GROUP BY 1
ORDER BY 1;
```

Report Output 4.2:

Employment in the industry sector for females has also dropped a bit from 16.3% to 15% between 2012 and 2019

Report Question 4.3: <mark>Level: Hard</mark>

How has the percentage of female employment in the Service sector changed from 2012 to 2019

- Indicator Names:
    - Employment in services, female
    - Population employed, female

Method:

- Since we only want data for a range of years, we can filter the dataset using
- WHERE column_name1 BETWEEN value1 AND value2
- To calculate the sum of a column for particular values we can use this function
- SUM(CASE WHEN column_name = condition THEN value END)
- We can then divide the female employed in services by total females employed to get the % of female employed in services
- Since we want the sum of the column with a condition by each year we will use the "GROUP BY" function
- Use "ORDER BY" function to sort the dataset by year

Report Answer 4.3:

```
SELECT
  year,
  SUM(CASE WHEN indicator = 'Employment in services, female' THEN value END) AS
"Females Employed in Services",
  SUM(CASE WHEN indicator = 'Population employed, female' THEN value END) AS "Total
females employed",
  100*SUM(CASE WHEN indicator = 'Employment in services, female' THEN value END) /
SUM(CASE WHEN indicator = 'Population employed, female' THEN value END) AS "% Female
Employment in Services"
FROM
  career_nub.gender_data
WHERE
  year BETWEEN 2012 AND 2019
GROUP BY 1
ORDER BY 1;
```

Report Output 4.3:

Employment in the service sector for females has increased from 51% to 57% between 2012 and 2019

Overall we see a shift in female employment from Agriculture to the Services sector which now has a 57% share of female employment.

Report Question 5: Understanding Financial Indicators

Report Question 5.1:

How has female debit card ownership changed in 2017 compared to 2014?
- Indicator Names:
  - Debit card ownership, female, 15+
  - Population, female, 15+

Method:
- To filter the dataset for a particular indicator
  WHERE column_name1 = 'value'
- Since we want the Average value of a column by each year we will use the "**GROUP BY**" function
- To calculate the average of a column we can use this function
  AVG(column_name)
- The "**ORDER BY**" function is used to sort the result set in ascending or descending order based on the mentioned column

Report Answer 5.1:

SELECT
 year,
 SUM(CASE WHEN indicator = 'Debit card ownership, female, 15+' THEN value END) AS "Females who own a debit card",

```
  SUM(CASE WHEN indicator = 'Population, female, 15+' THEN value END) AS "Adult female
population",
  100*SUM(CASE WHEN indicator = 'Debit card ownership, female, 15+' THEN value END) /
SUM(CASE WHEN indicator = 'Population, female, 15+' THEN value END) AS "% Females who
own a debit card"
FROM
  career_nub.gender_data
WHERE
  year IN (2014,2017)
GROUP BY 1
ORDER BY 1;
```

Report Output 5.1:

There is an increase in the percentage of females, owning a debit card from 36% in 2014 to
43% in 2017 indicating an increase in financial inclusion.

Report Question 5.2: Level: Medium

Find the five countries with the lowest female debit card ownership in 2017
  ● Indicator: Debit card ownership, female (% age 15+)

Method:

  ● To filter the dataset for a given year for a certain indicator, we will use
    WHERE column_name1 = condition value
        AND column_name2 = condition value
  ● The "**ORDER BY**" function is used to sort the result set in ascending since we want the
    worst countries which means a low female debit card ownership %
  ● Let's use the "**LIMIT**" function to restrict it to 5 rows to get the worst 5 countries

Report Answer 5.2:

```
SELECT
  country,
  indicator,
  value
FROM
  career_nub.gender_data
WHERE
  indicator = 'Debit card ownership, female (% age 15+)'
  AND year = 2017
ORDER BY value
```

LIMIT 5;

Report Output 5.2:

The countries with the lowest percentage of females with debit card ownership are:
South Sudan, Afghanistan, Chad, Sierra Leone, and Liberia

Report Question 6: Understanding Other Indicators

Report Question 6.1: Level: Easy

How has the Human Capital Index for females in India changed from 2017 to 2020
- Indicator Names: Human Capital Index (HCI), Female

Method:

- To filter the dataset for an indicator in a country across multiple years, we will use
  WHERE column_name1 = 'value'
      AND column_name2 = 'value'
      AND column_name3 IN (value1, value2)
- Select all columns by using **"SELECT *"**

Report Answer 6.1:

```
SELECT
 *
FROM
 career_nub.gender_data
WHERE indicator = 'Human Capital Index (HCI), Female'
 AND country = 'India'
 AND year IN (2017, 2020);
```

Report Output 6.1:

The Human Capital Index for females in India has improved from 0.45 in 2017 to 0.5 in 2020

Report Question 6.2:

Which are the best 5 countries based on the Human Capital Index (HCI) for females in 2020
  ● Indicator Name: Human Capital Index (HCI), Female

Method:

  ● To filter the dataset for a given year for a certain indicator, we will use
    WHERE column_name1 = condition value
        AND column_name2 = condition value
  ● The "**ORDER BY DESC**" function is used to sort the result set in descending order since
    we want the best countries which means a high HCI score
    ORDER BY column_name DESC
  ● Let's use the "**LIMIT**" function to restrict it to 5 rows to get the best 5 countries

Report Answer 6.2:

```
SELECT
 country,
 indicator,
 value
FROM
 career_nub.gender_data
WHERE
 indicator = 'Human Capital Index (HCI), Female'
 AND year = 2020
ORDER BY value DESC
LIMIT 5;
```

Report Output 6.2:

We see that the best 5 countries in terms of the Human Capital Index for females are Singapore, Hong Kong, Finland, Korea, and Estonia.

Report Question 6.3: Level: Medium

Which are the worst 5 countries based on the Human Capital Index (HCI) for females in 2020
- Indicator Name: Human Capital Index (HCI), Female

Method:
- To filter the dataset for a given year for a certain indicator, we will use
  WHERE column_name1 = condition value
          AND column_name2 = condition value
- The "**ORDER BY**" function is used to sort the result set in ascending since we want the worst countries which means a low HCI score
- Let's use the "**LIMIT**" function to restrict it to 5 rows to get the worst 5 countries

Report Answer 6.3:

```
SELECT
  country,
  indicator,
  value
FROM
  career_nub.gender_data
WHERE
  indicator = 'Human Capital Index (HCI), Female'
  AND year = 2020
ORDER BY value
LIMIT 5;
```

Report Output 6.3:

We see that the worst 5 countries in terms of the Human Capital Index for females are: Chad, Niger, Mali, Liberia, and Angola.

Report Question 6.4: Level: Medium

Find the number of countries each year where a woman cannot get a job the same way a man can

- Indicator Name: A woman can get a job in the same way as a man (1=yes; 0=no)

Method:

- Use **WHERE column_name = condition value** function to filter the dataset for the indicator
- To calculate the count of a column for particular values we can use this function **COUNT(CASE WHEN column_name = condition THEN value END)**
- Since we want the count of countries with a condition by each year we will use the "**GROUP BY**" function
- Use the "**ORDER BY**" function to sort the dataset by year

Report Answer 6.4:

SELECT
  year,
  COUNT(CASE WHEN value = 0 THEN country END) AS "Countries where women cannot get job same way as a man",
  COUNT(country) AS "Total Countries"
FROM
  career_nub.gender_data
WHERE
  indicator = 'A woman can get a job in the same way as a man (1=yes; 0=no)'
GROUP BY 1
ORDER BY 1;

Report Output 6.4:

We can see that the number of countries where a woman cannot get the job in the same way as man has dropped from 23 in 2012 to 17 in 2021

Report Question 6.5: Level: Easy

List all countries where this still happens

- Indicator Name: A woman can open a bank account in the same way as a man (1=yes; 0=no)

Method:

- Use this function to filter the dataset for the indicator
  WHERE
        column_name1 = condition value
        AND column_name2 = condition value

Report Answer 6.5:

SELECT
  year,
  country,
  indicator,
  value
FROM
  career_nub.gender_data
WHERE
  indicator = 'A woman can open a bank account in the same way as a man (1=yes; 0=no)'
  AND year = 2020
  AND value = 0;

Report Output 6.5:

We see that there are still 5 countries where a woman cannot open a bank account in the same way as a man. These are Bhutan, Equatorial Guinea, Guinea-Bissau, Kenya, and Suriname.

→ *Congratulations on completing the SQL tutorial! You can now move to the Python tutorial*