

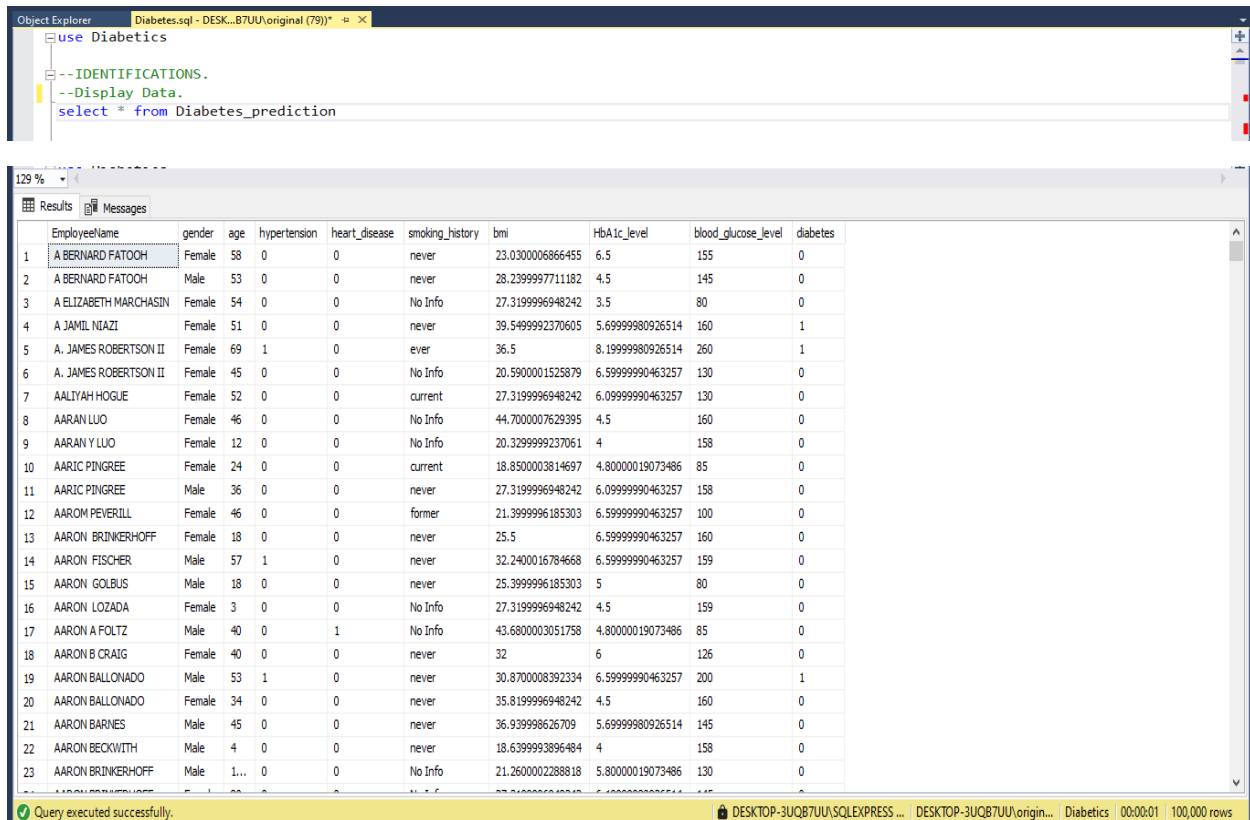
# Patient Data Overview

## DTA SUMMARY

I am working with a dataset that include information on patients with diabetes. Thie dataset contain 100000 rows and include the following columns:

- **Employee Name**: The name of the employee.
- **Gender**: The gender of the patient.
- **Age**: The age of the patient
- **Hypertension**: Weather the patient has hypertension(0: does not have hypertension, 1: has hypertension)
- **Heart Disease**: Weather the patient has heart disease(0: does not have heart disease, 1: has heart disease)
- **Smoking History**: The patient's smoking history.
- **BMI**: The Body Mass Index.
- **HbA1c Level**: The HbA1c level in the blood.
- **Blood Glucose Level**: The blood Glucose Level.
- **Diabetes**: Weather the patient has diabetes(0: does not have diabetes, 1: has diabetes)

## DISPLAY THE DATA



The screenshot displays a SQL Server environment. The top pane shows a query window with the following SQL code:

```
use Diabetics
--IDENTIFICATIONS.
--Display Data.
select * from Diabetes_prediction
```

The bottom pane shows the results of the query, displaying 23 rows of patient data. The columns are: EmployeeName, gender, age, hypertension, heart\_disease, smoking\_history, bmi, HbA1c\_level, blood\_glucose\_level, and diabetes. The status bar at the bottom indicates the query was executed successfully and returned 100,000 rows.

|    | EmployeeName          | gender | age  | hypertension | heart_disease | smoking_history | bmi              | HbA1c_level      | blood_glucose_level | diabetes |
|----|-----------------------|--------|------|--------------|---------------|-----------------|------------------|------------------|---------------------|----------|
| 1  | A BERNARD FATOOH      | Female | 58   | 0            | 0             | never           | 23.0300006866455 | 6.5              | 155                 | 0        |
| 2  | A BERNARD FATOOH      | Male   | 53   | 0            | 0             | never           | 28.2399997711182 | 4.5              | 145                 | 0        |
| 3  | A ELIZABETH MARCHASIN | Female | 54   | 0            | 0             | No Info         | 27.3199996948242 | 3.5              | 80                  | 0        |
| 4  | A JAMIL NIAZI         | Female | 51   | 0            | 0             | never           | 39.5499992370605 | 5.69999980926514 | 160                 | 1        |
| 5  | A. JAMES ROBERTSON II | Female | 69   | 1            | 0             | ever            | 36.5             | 8.19999980926514 | 260                 | 1        |
| 6  | A. JAMES ROBERTSON II | Female | 45   | 0            | 0             | No Info         | 20.5900001525879 | 6.59999990463257 | 130                 | 0        |
| 7  | AALIYAH HOGUE         | Female | 52   | 0            | 0             | current         | 27.3199996948242 | 6.09999990463257 | 130                 | 0        |
| 8  | AARAN LUO             | Female | 46   | 0            | 0             | No Info         | 44.7000007629395 | 4.5              | 160                 | 0        |
| 9  | AARAN Y LUO           | Female | 12   | 0            | 0             | No Info         | 20.3299999237061 | 4                | 158                 | 0        |
| 10 | AARIC PINGREE         | Female | 24   | 0            | 0             | current         | 18.8500003814697 | 4.80000019073486 | 85                  | 0        |
| 11 | AARIC PINGREE         | Male   | 36   | 0            | 0             | never           | 27.3199996948242 | 6.09999990463257 | 158                 | 0        |
| 12 | AAROM PEVERILL        | Female | 46   | 0            | 0             | former          | 21.3999996185303 | 6.59999990463257 | 100                 | 0        |
| 13 | AARON BRINKERHOFF     | Female | 18   | 0            | 0             | never           | 25.5             | 6.59999990463257 | 160                 | 0        |
| 14 | AARON FISCHER         | Male   | 57   | 1            | 0             | never           | 32.2400016784668 | 6.59999990463257 | 159                 | 0        |
| 15 | AARON GOLBUS          | Male   | 18   | 0            | 0             | never           | 25.3999996185303 | 5                | 80                  | 0        |
| 16 | AARON LOZADA          | Female | 3    | 0            | 0             | No Info         | 27.3199996948242 | 4.5              | 159                 | 0        |
| 17 | AARON A FOLTZ         | Male   | 40   | 0            | 1             | No Info         | 43.6800003051758 | 4.80000019073486 | 85                  | 0        |
| 18 | AARON B CRAIG         | Female | 40   | 0            | 0             | never           | 32               | 6                | 126                 | 0        |
| 19 | AARON BALLONADO       | Male   | 53   | 1            | 0             | never           | 30.8700008392334 | 6.59999990463257 | 200                 | 1        |
| 20 | AARON BALLONADO       | Female | 34   | 0            | 0             | never           | 35.8199996948242 | 4.5              | 160                 | 0        |
| 21 | AARON BARNES          | Male   | 45   | 0            | 0             | never           | 36.939998626709  | 5.69999980926514 | 145                 | 0        |
| 22 | AARON BECKWITH        | Male   | 4    | 0            | 0             | never           | 18.6399993896484 | 4                | 158                 | 0        |
| 23 | AARON BRINKERHOFF     | Male   | 1... | 0            | 0             | No Info         | 21.2600002288818 | 5.80000019073486 | 130                 | 0        |

## Identifying Data Type in the Table:

```
--IDENTIFYING DATA TYPE
select column_name,
       data_type
from INFORMATION_SCHEMA.columns
where TABLE_NAME='Diabetes_Prediction'
```

Results

|    | column_name         | data_type |
|----|---------------------|-----------|
| 1  | EmployeeName        | nvarchar  |
| 2  | gender              | nvarchar  |
| 3  | age                 | float     |
| 4  | hypertension        | bit       |
| 5  | heart_disease       | bit       |
| 6  | smoking_history     | nvarchar  |
| 7  | bmi                 | float     |
| 8  | HbA1c_level         | float     |
| 9  | blood_glucose_level | smallint  |
| 10 | diabetes            | bit       |

## Identifying Constraints.

There are no constraints (P.K....)

```
--IDENTIFYING CONSTRAINTS.
SELECT TABLE_NAME,
       CONSTRAINT_NAME,
       CONSTRAINT_TYPE
FROM INFORMATION_SCHEMA.TABLE_CONSTRAINTS
WHERE TABLE_SCHEMA='DBO'
```

Results

| TABLE_NAME | CONSTRAINT_NAME | CONSTRAINT_TYPE |
|------------|-----------------|-----------------|
|------------|-----------------|-----------------|

## Data Cleaning.

### NULL VALUES

```
SELECT * FROM Diabetes_prediction
WHERE EmployeeName =NULL
```

Results

| EmployeeName | gender | age | hypertension | heart_disease | smoking_history | bmi | HbA1c_level | blood_glucose_level | diabetes |
|--------------|--------|-----|--------------|---------------|-----------------|-----|-------------|---------------------|----------|
|--------------|--------|-----|--------------|---------------|-----------------|-----|-------------|---------------------|----------|

There are Null Values.

## Identifying Duplicates:

When analyzing the dataset, it was discovered that there are approximately 35000 duplicate names.

```
select
  distinct (employeeename)
from
  Diabetes_prediction
ORDER BY
  EmployeeName ASC
```

| Results      |                       | Messages |
|--------------|-----------------------|----------|
| employeeName |                       |          |
| 1            | A BERNARD FATOOH      |          |
| 2            | A ELIZABETH MARCHASIN |          |
| 3            | A JAMIL NIAZI         |          |
| 4            | A. JAMES ROBERTSON II |          |
| 5            | AALIYAH HOGUE         |          |
| 6            | AARAN LUO             |          |
| 7            | AARAN Y LUO           |          |
| 8            | AARIC PINGREE         |          |
| 9            | AAROM PEVERILL        |          |
| 10           | AARON BRINKERHOFF     |          |
| 11           | AARON FISCHER         |          |
| 12           | AARON GOLBUS          |          |
| 13           | AARON LOZADA          |          |
| 14           | AARON A FOLTZ         |          |
| 15           | AARON R CRAIG         |          |

Query executed successfully. DESKTOP-3UQB7UU\SQLEXPRESS ... DESKTOP-3UQB7UU\origin... Diabetics 00:00:02 65,468 rows

## Names and Numbr of Duplicate Names:

I identified how many times each name s repeated in the dataset. The analysis reveals the frequency of each duplicate name.

```
SELECT
    EmployeeName,
    COUNT(*) AS Total
FROM
    Diabetes_prediction
GROUP BY
    (EmployeeName)
HAVING COUNT(*)>0
ORDER BY EmployeeName ASC
```

| Results      |                       | Messages |
|--------------|-----------------------|----------|
| EmployeeName | Total                 |          |
| 1            | A BERNARD FATOOH      | 2        |
| 2            | A ELIZABETH MARCHASIN | 1        |
| 3            | A JAMIL NIAZI         | 1        |
| 4            | A. JAMES ROBERTSON II | 2        |
| 5            | AALIYAH HOGUE         | 1        |
| 6            | AARAN LUO             | 1        |
| 7            | AARAN Y LUO           | 1        |
| 8            | AARIC PINGREE         | 2        |
| 9            | AAROM PEVERILL        | 1        |
| 10           | AARON BRINKERHOFF     | 1        |
| 11           | AARON FISCHER         | 1        |
| 12           | AARON GOLBUS          | 1        |
| 13           | AARON LOZADA          | 1        |
| 14           | AARON A FOLTZ         | 1        |

Query executed successfully. DESKTOP-3UQB7UU\SQLEXPRESS ... DESKTOP-3UQB7UU\origin... Diabetics 00:00:00 65,468 rows

## Extract All Data Related to Employees Who Have Duplicate:

```
select * from Diabetes_prediction
where EmployeeName IN(
    select
        DISTINCT(EmployeeName)
    from Diabetes_prediction
    group by EmployeeName
    having count (*)>1)
ORDER BY EmployeeName ASC
```

|    | EmployeeName          | gender | age              | hypertension | heart_disease | smoking_history | bmi              | HbA1c_level      | blood_glucose_level | diabetes |
|----|-----------------------|--------|------------------|--------------|---------------|-----------------|------------------|------------------|---------------------|----------|
| 1  | A BERNARD FATOOH      | Female | 58               | 0            | 0             | never           | 23.0300006866455 | 6.5              | 155                 | 0        |
| 2  | A BERNARD FATOOH      | Male   | 53               | 0            | 0             | never           | 28.2399997711182 | 4.5              | 145                 | 0        |
| 3  | A. JAMES ROBERTSON II | Female | 69               | 1            | 0             | ever            | 36.5             | 8.19999980926514 | 260                 | 1        |
| 4  | A. JAMES ROBERTSON II | Female | 45               | 0            | 0             | No Info         | 20.5900001525879 | 6.59999990463257 | 130                 | 0        |
| 5  | AARIC PINGREE         | Female | 24               | 0            | 0             | current         | 18.8500003814697 | 4.80000019073486 | 85                  | 0        |
| 6  | AARIC PINGREE         | Male   | 36               | 0            | 0             | never           | 27.3199996948242 | 6.09999990463257 | 158                 | 0        |
| 7  | AARON BALLONADO       | Male   | 53               | 1            | 0             | never           | 30.8700008392334 | 6.59999990463257 | 200                 | 1        |
| 8  | AARON BALLONADO       | Female | 34               | 0            | 0             | never           | 35.8199996948242 | 4.5              | 160                 | 0        |
| 9  | AARON BRINKERHOFF     | Male   | 1.08000004291534 | 0            | 0             | No Info         | 21.2600002288818 | 5.80000019073486 | 130                 | 0        |
| 10 | AARON BRINKERHOFF     | Female | 80               | 0            | 0             | No Info         | 27.3199996948242 | 6.19999980926514 | 145                 | 0        |
| 11 | AARON CHEN            | Male   | 20               | 0            | 0             | No Info         | 36.6100006103516 | 5.69999980926514 | 145                 | 0        |
| 12 | AARON CHEN            | Male   | 66               | 0            | 0             | current         | 32.7200012207031 | 3.5              | 160                 | 0        |
| 13 | AARON COWHIG          | Female | 19               | 0            | 0             | No Info         | 22.0799999237061 | 5.69999980926514 | 126                 | 0        |
| 14 | AARON COWHIG          | Female | 19               | 0            | 0             | never           | 33.8899993896484 | 6.19999980926514 | 159                 | 0        |
| 15 | AARON CRAIG           | Female | 46               | 0            | 0             | never           | 38.6100006103516 | 3.5              | 159                 | 0        |

## Analyzing Duplicate Names with Different Data.

Performed additional analysis to gather more information about the duplicate names. However, it was found that while the names appeared to be duplicates, the associated data was different. This indicates that the apparent duplicates are not actual duplicates but rather different records with similar names.

```

select employeeName,
       gender,
       age,
       smoking_history,
       hypertension,
       diabetes,
       bmi,
       heart_disease,
       HbA1c_level, |
       count(*) as Count_Name
from Diabetes_prediction
group by EmployeeName,
       gender,
       age,
       smoking_history,
       hypertension,
       diabetes,
       bmi,
       heart_disease,
       HbA1c_level
having count(*)>1

```

| employeeName | gender | age | smoking_history | hypertension | diabetes | bmi | heart_disease | HbA1c_level | Count_Name |
|--------------|--------|-----|-----------------|--------------|----------|-----|---------------|-------------|------------|
|--------------|--------|-----|-----------------|--------------|----------|-----|---------------|-------------|------------|

## Data Validating:

Ensuring Age is Within a Valid Range

```

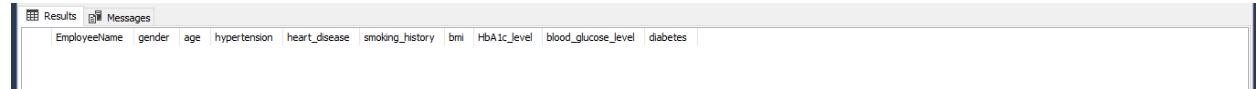
SELECT *
FROM Diabetes_prediction
WHERE age<0 OR age>100

```

| EmployeeName | gender | age | hypertension | heart_disease | smoking_history | bmi | HbA1c_level | blood_glucose_level | diabetes |
|--------------|--------|-----|--------------|---------------|-----------------|-----|-------------|---------------------|----------|
|--------------|--------|-----|--------------|---------------|-----------------|-----|-------------|---------------------|----------|

## Ensuring Binary Values (0 OR 1) Are Correct

```
--ENSURING BINARY VALUES (0 OR 1) ARE CORRECT
SELECT * FROM Diabetes_prediction
WHERE diabetes NOT IN (0,1) OR
hypertension NOT IN (0,1) OR
heart_disease NOT IN (0,1)
```



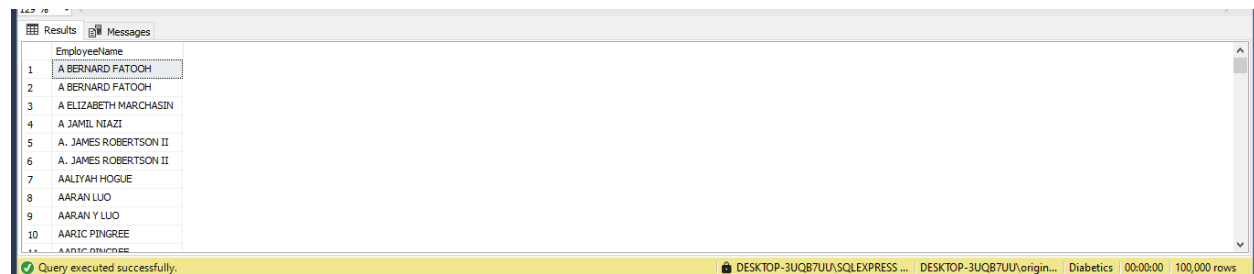
## Improve the performance

### Converting All Names to Uppercase for Easier Search:

To simplify the search and comparison of names, all names in the dataset were converted to uppercase. This standardization helps to ensure consistency and make search more straightforward.

```
UPDATE Diabetes_prediction
SET EmployeeName=UPPER(EmployeeName);

SELECT EmployeeName
FROM Diabetes_prediction
```

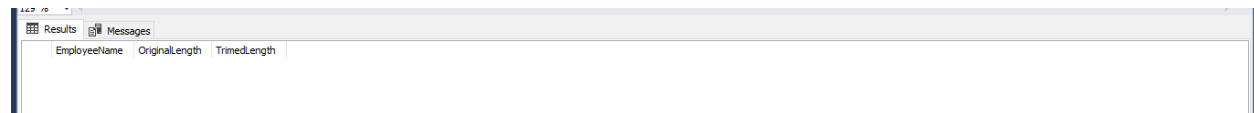


This transformation makes it easier to perform case-insensitive search and comparison across the dataset.

### Removing Extra Spaces in Names:

To clean up the names and ensure there are no leading, trailing, or multiple spaces between words, I used TRIM to remove extra spaces. This helps in standardizing the names for accurate data analysis and searching.

```
SELECT
EmployeeName,
LEN(EmployeeName) AS OriginalLength,
LEN(TRIM(RTRIM(EmployeeName))) AS TrimedLength
FROM Diabetes_prediction
WHERE
LEN(EmployeeName)<> LEN(LTRIM(RTRIM(EmployeeName)))
```



## CREATING INDEXES:

I created a clustered index to facilitate faster search operations and improve overall query performance. A clustered index orders the rows in the table, which can significantly speed up data retrieval.

```
--CREATING INDEXES
CREATE CLUSTERED INDEX idx_EmployeeName
ON Diabetes_Prediction(EmployeeName)
```

## CREATING STORED PROCEDURE:

```
CREATE PROC
GetEmployeesByAge
@MinAge FLOAT,
@MaxAge FLOAT
AS
BEGIN
SELECT
    EmployeeName,
    age,
    diabetes, heart_disease
FROM Diabetes_prediction
WHERE age BETWEEN @MinAge AND @MaxAge;
END;
```

```
EXEC GetEmployeesByAge
@MinAge=0, @MaxAge=35
```

|    | EmployeeName      | age                | diabetes | heart_disease |
|----|-------------------|--------------------|----------|---------------|
| 1  | AARAN Y LUO       | 12                 | 0        | 0             |
| 2  | AARIC PINGREE     | 24                 | 0        | 0             |
| 3  | AARON BRINKERHOFF | 18                 | 0        | 0             |
| 4  | AARON GOLBUS      | 18                 | 0        | 0             |
| 5  | AARON LOZADA      | 3                  | 0        | 0             |
| 6  | AARON BALLONADO   | 34                 | 0        | 0             |
| 7  | AARON BECKWITH    | 4                  | 0        | 0             |
| 8  | AARON BRINKERHOFF | 1.08000004291534   | 0        | 0             |
| 9  | AARON C BALLONADO | 27                 | 0        | 0             |
| 10 | AARON C STEVENSON | 0.2399999994635... | 0        | 0             |
| 11 | AARON CHAPMAN     | 19                 | 0        | 0             |
| 12 | AARON CHEN        | 20                 | 0        | 0             |
| 13 | AARON COWHIG      | 19                 | 0        | 0             |
| 14 | AARON COWHIG      | 19                 | 0        | 0             |
| 15 | AARON CRAMER      | 31                 | 0        | 0             |
| 16 | AARON D LOUKONEN  | 19                 | 0        | 0             |
| 17 | AARON D STARR     | 14                 | 0        | 0             |
| 18 | AARON DEL TREDICI | 3                  | 0        | 0             |
| 19 | AARON DURAN       | 6                  | 0        | 0             |
| 20 | AARON FLORES      | 4                  | 0        | 0             |

Query executed successfully. DESKTOP-3UQB7UU\SQLEXPRESS ... | DESKTOP-3UQB7UU\origin... | Diabetics | 00:00:00 | 39,988 rows

## Exploratory Data Analysis (EDA)

### 1. Count of Rows

To begin our exploratory data analysis, we performed a count of the total number of rows in the Diabetes\_prediction table. This step is crucial to understand the dataset's size and ensure that we are working with a substantial amount of data. The total count of rows in the table is **100,000**.

### 2. Age Statistics

We analyzed the Age column to understand the distribution of ages in the dataset. The key statistics obtained were:

- **Minimum Age:** The youngest individual in the dataset is **7 days**.
- **Maximum Age:** The oldest individual in the dataset is **80 years old**.
- **Average Age:** The average age of individuals in the dataset is **41.8 years**.

These statistics help in understanding the age range and the central tendency of the age distribution among the individuals.

### 3. BMI (Body Mass Index) Statistics

Next, we examined the BMI column to gain insights into the body mass index of individuals. The statistics include:

- **Minimum BMI:** The lowest BMI recorded is **10.01**.
- **Maximum BMI:** The highest BMI recorded is **95.6**.
- **Average BMI:** The average BMI of individuals is **27.3**.

These metrics are important for assessing the general health status and potential obesity issues within the population.

### 4. HbA1c Level Statistics

We analyzed the HbA1c\_level column, which is a crucial indicator for diabetes management. The key statistics are:

- **Minimum HbA1c Level:** The lowest HbA1c level recorded is **3.5**.
- **Maximum HbA1c Level:** The highest HbA1c level recorded is **9**.
- **Average HbA1c Level:** The average HbA1c level is **5.5**.

Understanding the distribution of HbA1c levels helps in assessing the diabetes control among the individuals.

### 5. Blood Glucose Level Statistics

Finally, we analyzed the Blood\_glucose\_level column to understand the blood sugar levels in the dataset. The statistics include:

- **Minimum Blood Glucose Level:** The lowest blood glucose level recorded is **80**.
- **Maximum Blood Glucose Level:** The highest blood glucose level recorded is **300**.
- **Average Blood Glucose Level:** The average blood glucose level is **130**.

These statistics are crucial for identifying individuals with potential blood sugar management issues and for overall diabetes care.

```
SELECT
    COUNT(*) AS TotalRows,
    MIN(age) AS MinAge,
    MAX(age) AS MaxAge,
    AVG(age) AS AvgAge,
    MIN(bmi) AS MinBmi,
    MAX(bmi) AS MaxBmi,
    AVG(bmi) AS AvgBmi,
    MIN(HbA1c_level) AS MinHbA1cLevel,
    MAX(HbA1c_level) AS MaxHbA1cLevel,
    AVG(HbA1c_level) AS AvgHbA1cLevel,
    MIN(blood_glucose_level) AS MinBloodGlucoseLevel,
    MAX(blood_glucose_level) AS MaxBloodGlucoseLevel,
    AVG(blood_glucose_level) AS AvgBloodGlucoseLevel
FROM Diabetes_prediction
```

|   | TotalRows | MinAge              | MaxAge | AvgAge           | MinBmi           | MaxBmi           | AvgBmi           | MinHbA1cLevel | MaxHbA1cLevel | AvgHbA1cLevel    | MinBloodGlucoseLevel | MaxBloodGlucoseLevel | AvgBloodGlucoseLevel |
|---|-----------|---------------------|--------|------------------|------------------|------------------|------------------|---------------|---------------|------------------|----------------------|----------------------|----------------------|
| 1 | 100000    | 0.07999999982118607 | 80     | 41.8858559999676 | 10.0100002288818 | 95.6900024414063 | 27.3207670177078 | 3.5           | 9             | 5.52750698394775 | 80                   | 300                  | 138                  |

## Count and Percentage of Gender

Next, we examined the Gender column to understand the distribution of gender in the dataset. This analysis helps us to identify any gender imbalance that might exist. The key findings were:

- **Count of Males:** There are **41430** males in the dataset.
- **Count of Females:** There are **58552** females in the dataset.
- **Percentage of Males:** Males make up **41%** of the dataset.
- **Percentage of Females:** Females make up **58%** of the dataset.

This gender distribution analysis is important for understanding the demographic composition of our dataset, which can influence the interpretation of subsequent analyses.

```
SELECT gender ,
    COUNT(*) AS TotalCount,
    COUNT(*)*100/SUM(COUNT(*)) OVER () AS GenderPercentage
FROM Diabetes_prediction
GROUP BY gender
```

|   | gender | TotalCount | GenderPercentage |
|---|--------|------------|------------------|
| 1 | Male   | 41430      | 41               |
| 2 | Female | 58552      | 58               |
| 3 | Other  | 18         | 0                |

## Count and Percentage of Diabetes

In this step, we analyzed the Diabetes column to understand the distribution of individuals with and without diabetes in the dataset. This analysis helps us to identify the prevalence of diabetes within our population. The key findings were:

- **Count of Individuals with Diabetes:** There are **8500** individuals diagnosed with diabetes in the dataset.



- **Count of Individuals without Diabetes:** There are **91500** individuals without diabetes in the dataset.
- **Percentage of Individuals with Diabetes:** Individuals with diabetes make up **8%** of the dataset.
- **Percentage of Individuals without Diabetes:** Individuals without diabetes make up **91%** of the dataset.

This analysis is crucial for understanding the proportion of individuals affected by diabetes, which can influence the focus of subsequent analyses and the development of targeted interventions.

```
SELECT
    diabetes,
    COUNT(*) AS TotalCount,
    COUNT(*)*100/SUM(COUNT(*)) OVER () AS DiabetesPercentage
FROM Diabetes_prediction
GROUP BY diabetes
```

|   | diabetes | TotalCount | DiabetesPercentage |
|---|----------|------------|--------------------|
| 1 | 0        | 91500      | 91                 |
| 2 | 1        | 8500       | 8                  |

### Count and Percentage of Hypertension

In this step, we analyzed the Hypertension column to understand the distribution of individuals with and without hypertension in the dataset. This analysis helps us to identify the prevalence of hypertension within our population. The key findings were:

- **Count of Individuals with Hypertension:** There are **7485** individuals diagnosed with hypertension in the dataset.
- **Count of Individuals without Hypertension:** There are **92515** individuals without hypertension in the dataset.
- **Percentage of Individuals with Hypertension:** Individuals with hypertension make up **7%** of the dataset.
- **Percentage of Individuals without Hypertension:** Individuals without hypertension make up **92%** of the dataset.

This analysis is important for understanding the proportion of individuals affected by hypertension, which can inform healthcare strategies and resource allocation for managing hypertension within the population.

```
SELECT
    hypertension,
    COUNT(*) AS TotalCount,
    COUNT(*)*100/SUM(COUNT(*)) OVER () AS HypertensionPercentage
FROM Diabetes_prediction
GROUP BY hypertension
```

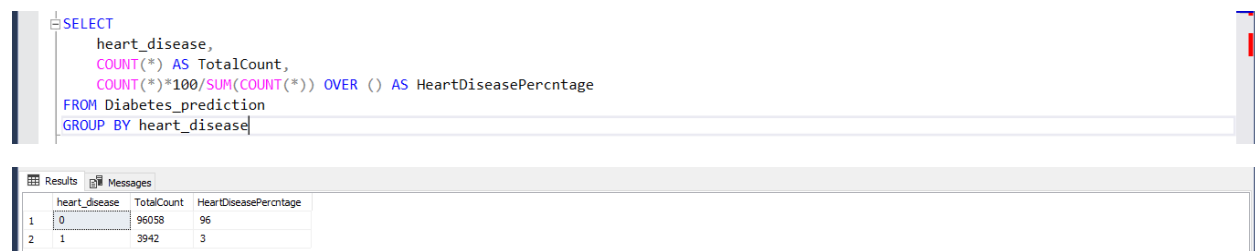
|   | hypertension | TotalCount | HypertensionPercentage |
|---|--------------|------------|------------------------|
| 1 | 0            | 92515      | 92                     |
| 2 | 1            | 7485       | 7                      |

## Count and Percentage of Heart Disease

In this step, we examined the heart disease column to determine the distribution of individuals with and without heart disease in the dataset. This analysis provides insight into the prevalence of heart disease among the population. The key findings were:

- **Count of Individuals with Heart Disease:** There are **3942** individuals diagnosed with heart disease in the dataset.
- **Count of Individuals without Heart Disease:** There are 96058 individuals without heart disease in the dataset.
- **Percentage of Individuals with Heart Disease:** Individuals with heart disease make up **3%** of the dataset.
- **Percentage of Individuals without Heart Disease:** Individuals without heart disease make up **96%** of the dataset.

This information is crucial for identifying the burden of heart disease within the population and can help in planning preventive measures and treatment strategies for those at risk.



```
SELECT
    heart_disease,
    COUNT(*) AS TotalCount,
    COUNT(*)*100/SUM(COUNT(*)) OVER () AS HeartDiseasePercentage
FROM Diabetes_prediction
GROUP BY heart_disease
```

|   | heart_disease | TotalCount | HeartDiseasePercentage |
|---|---------------|------------|------------------------|
| 1 | 0             | 96058      | 96                     |
| 2 | 1             | 3942       | 3                      |

Data Distribution and Outliers.

## Distribution of BMI

In this step, we analyzed the Bmi (Body Mass Index) column to understand the distribution of BMI values in the dataset. We ran a query to count the frequency of each unique BMI value present in the dataset. The results of this query provide insights into how BMI values are distributed among the individuals.

Here's a breakdown of what the results tell us:

- **Frequency Distribution:** The query returns each unique BMI value along with the count of individuals having that specific BMI. This allows us to see how many people fall into each BMI category.
- **Identification of Common BMI Values:** By looking at the frequency of BMI values, we can identify the most common BMI values in the dataset. This helps in understanding the general health status of the population.
- **Health Risk Assessment:** The distribution of BMI values can be used to assess the risk of health conditions. For example, a higher frequency of high BMI values may indicate a higher prevalence of overweight and obesity, which are risk factors for diseases like diabetes, hypertension, and heart disease.

Understanding the distribution of BMI values is essential for public health analysis and for designing interventions aimed at improving the health and nutritional status of the population.



The screenshot shows a SQL query editor with the following query:

```
SELECT
    bmi,
    COUNT(*) AS Frequency
FROM Diabetes_prediction
GROUP BY bmi
```

Below the query editor, the 'Results' tab displays a table with two columns: 'bmi' and 'Frequency'. The table contains 16 rows of data. The status bar at the bottom indicates 'Query executed successfully.' and '4,247 rows'.

|    | bmi              | Frequency |
|----|------------------|-----------|
| 1  | 16.7000007629395 | 50        |
| 2  | 41.439998626709  | 1         |
| 3  | 41.6199989318848 | 6         |
| 4  | 52.1100006103516 | 3         |
| 5  | 49.3400001525879 | 1         |
| 6  | 24.1399993896484 | 65        |
| 7  | 15.2299995422363 | 15        |
| 8  | 40.75            | 16        |
| 9  | 15.039999961853  | 14        |
| 10 | 35.0400009155273 | 13        |
| 11 | 18.0300006866455 | 17        |
| 12 | 30.9899997711182 | 34        |
| 13 | 19.8999996185303 | 26        |
| 14 | 16.8600006103516 | 11        |
| 15 | 46.6800003051758 | 1         |
| 16 | 32.2700004577637 | 20        |

## CORRELATION.

### Correlation Between Hypertension and Diabetes

#### Correlation Coefficient: 0.19

- **Interpretation:** The correlation coefficient of 0.19 indicates a weak positive relationship between Hypertension and Diabetes. This suggests that there is a slight tendency for individuals with hypertension to also have diabetes, but the relationship is not strong.
- **Implications:**
  - **Weak Relationship:** Since the correlation is relatively low, it implies that while there is some association between hypertension and diabetes, it's not a strong or consistent one. This means that having hypertension does not necessarily mean an individual will have diabetes, and vice versa.
  - **Further Investigation:** Given the weak correlation, additional factors or variables might be influencing the relationship between hypertension and diabetes. Further investigation could explore other contributing factors or interactions that might better explain the connection between these conditions.
  - **Health Strategies:** While the weak correlation indicates that hypertension and diabetes are not strongly linked in this dataset, it's still important for health interventions to consider both conditions, especially in populations where other risk factors may be present.

**In summary,** the correlation of 0.19 reflects a weak association between hypertension and diabetes, suggesting that the two conditions are somewhat related but not strongly so.

```

WITH CTE AS (
    SELECT
        AVG(CAST(hypertension AS FLOAT)) AS AvgHypertension,
        AVG(CAST(diabetes AS FLOAT)) AS AvgDiabetes,
        STDEV(CAST(hypertension AS FLOAT)) AS StdevHypertension,
        STDEV(CAST(diabetes AS FLOAT)) AS StdevDiabetes
    FROM Diabetes_prediction
),
Covariance AS(
    SELECT
        SUM((hypertension-AvgHypertension)*(diabetes- AvgDiabetes))/
        (COUNT(*)-1) AS Covariance
    FROM Diabetes_Prediction,
    CTE
)
SELECT Covariance/(StdevHypertension*StdevDiabetes) AS CorrelationHypertensionDiabetes
FROM Covariance,CTE;

```

| Results                         | Messages          |
|---------------------------------|-------------------|
| CorrelationHypertensionDiabetes |                   |
| 1                               | 0.197823246407987 |

## Correlation Between Heart Disease and Diabetes

### Correlation Coefficient: 0.17

- **Interpretation:** The correlation coefficient of 0.17 signifies a very weak positive relationship between heart disease and Diabetes. This indicates a slight tendency for individuals with heart disease to also have diabetes, but the association is minimal.
- **Implications:**
  - **Very Weak Relationship:** The low correlation coefficient suggests that heart disease and diabetes have only a minor relationship with each other in the dataset. The presence of one condition does not strongly predict the presence of the other.
  - **Further Analysis:** The weak correlation calls for a deeper investigation to understand if other factors are influencing the link between heart disease and diabetes. It might be helpful to examine additional variables or consider interaction effects that could provide more insights.
  - **Clinical Considerations:** Although the correlation is weak, individuals with either condition should still be monitored for the other as part of a comprehensive health assessment, especially in the context of other risk factors.

In summary, the correlation of **0.17** reflects a very weak association between heart disease and diabetes, suggesting that the two conditions are weakly related.

```

WITH CTE AS (
    SELECT
        AVG(CAST(heart_disease AS FLOAT)) AS AvgHeartDisease,
        AVG(CAST(diabetes AS FLOAT)) AS AvgDiabetes,
        STDEV(CAST(heart_disease AS FLOAT)) AS StdevHeartDisease,
        STDEV(CAST(diabetes AS FLOAT)) AS StdevDiabetes
    FROM Diabetes_prediction
),
Covariance AS (
    SELECT
        SUM((heart_disease-AvgHeartDisease)*(diabetes- AvgDiabetes))/
        (COUNT(*)-1) AS Covariance
    FROM Diabetes_Prediction,
    CTE
)
SELECT Covariance/((StdevDiabetes*StdevHeartDisease) AS CorrelationHeartDiseaseDiabetes
FROM Covariance,CTE;

```

| Results                         | Messages |
|---------------------------------|----------|
| CorrelationHeartDiseaseDiabetes |          |
| 1 0.171726849549217             |          |

## Total Employees and Disease Percentage

### 1. Total Employees:

- The total number of employees in the dataset is **100,000**. This represents the complete count of individuals included in the analysis.

### 2. Diabetes Percentage:

- The percentage of employees diagnosed with diabetes is **8%**. This means that **8%** of the employees have been diagnosed with diabetes.

### 3. Hypertension Percentage:

- The percentage of employees with hypertension is **7%**. This indicates that **7%** of the employees suffer from hypertension.

### 4. Heart Disease Percentage:

- The percentage of employees with heart disease is **3%**. This reflects that **3%** of the employees have been diagnosed with heart disease.

## Summary:

- This analysis provides insights into the prevalence of diabetes, hypertension, and heart disease among the employees. Understanding these percentages helps in assessing the overall health status of the workforce and can guide the development of health and wellness programs tailored to these conditions.

```

SELECT
    COUNT(*) AS TotalEmployees,
    SUM(CASE
        WHEN diabetes=1 THEN 1
        ELSE 0
    END)*100/COUNT(*) AS DiabetesPercentage,
    SUM(CASE
        WHEN hypertension=1 THEN 1
        ELSE 0
    END)*100/COUNT(*) AS HypertensionPercentage,
    SUM(CASE
        WHEN heart_disease=1 THEN 1
        ELSE 0
    END)*100/COUNT(*) AS HeratDiseasPercentage
FROM Diabetes_prediction

```

|   | TotalEmployees | DiabetesPercentage | HypertensionPercentage | HeartDiseasePercentage |
|---|----------------|--------------------|------------------------|------------------------|
| 1 | 100000         | 8                  | 7                      | 3                      |

## Total Number of Diabetes by Gender

### 1. Male:

- **No Diabetes: 37,391** males do not have diabetes.
- **Diabetes: 4,039** males have been diagnosed with diabetes.

### 2. Female:

- **No Diabetes: 54,091** females do not have diabetes.
- **Diabetes: 4,461** females have been diagnosed with diabetes.

### 3. Other:

- **No Diabetes: 18** individuals categorized as "Other" do not have diabetes.
- **Diabetes:** No individuals in this category have been diagnosed with diabetes.

## Summary:

- This analysis reveals the distribution of diabetes among different genders in the dataset. It shows that while the number of females with diabetes is slightly higher than males, the majority of individuals across all gender categories do not have diabetes. This information is useful for understanding the prevalence of diabetes within different gender groups and can inform targeted health interventions.

```
WITH CTE AS (
    SELECT
        gender,
        diabetes
    FROM
        Diabetes_prediction
)
SELECT gender,
    ISNULL([0], 0) AS NoDiabetes,
    ISNULL([1], 0) AS Diabetes
FROM
    CTE
PIVOT (
    COUNT(diabetes) FOR diabetes IN ([0], [1])
) AS pt;
```

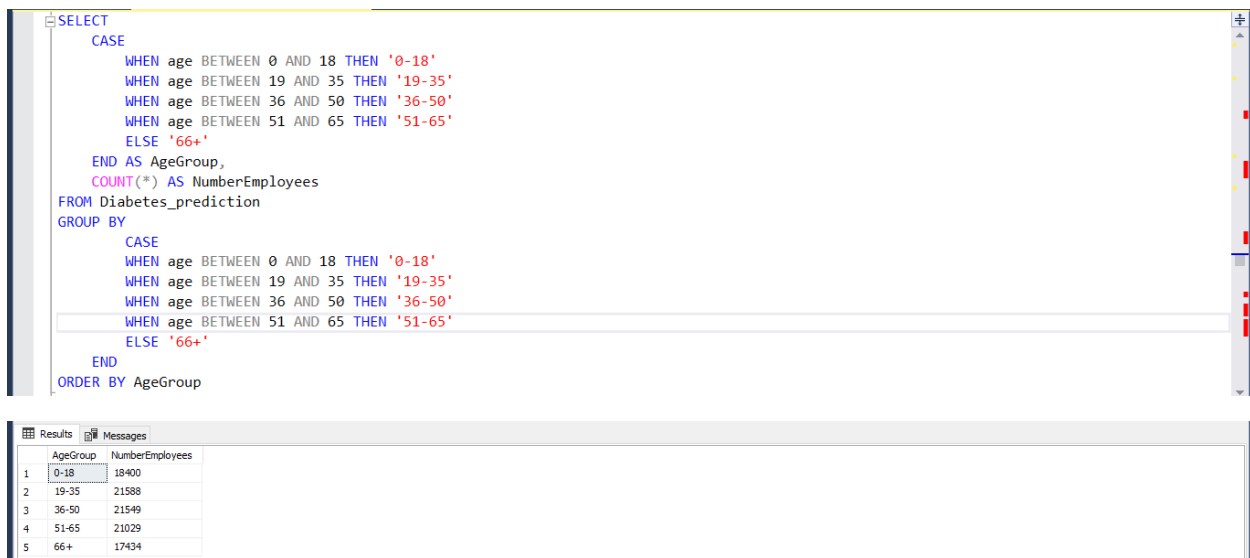
|   | gender | NoDiabetes | Diabetes |
|---|--------|------------|----------|
| 1 | Male   | 37391      | 4039     |
| 2 | Female | 54091      | 4461     |
| 3 | Other  | 18         | 0        |

## Age Group Distribution

1. **Age Group 0-18:**
  - **Number of Employees: 18,400** individuals are within the age range of 0 to 18.
2. **Age Group 19-35:**
  - **Number of Employees: 21,588** individuals are within the age range of 19 to 35.
3. **Age Group 36-50:**
  - **Number of Employees: 21,549** individuals are within the age range of 36 to 50.
4. **Age Group 51-65:**
  - **Number of Employees: 21,029** individuals are within the age range of 51 to 65.
5. **Age Group 66+:**
  - **Number of Employees: 17,434** individuals are aged 66 or older.

#### Summary:

- The distribution of employees across different age groups shows a relatively balanced spread, with the highest number of employees falling into the age groups 19-35 and 36-50. The number of employees decreases slightly in the age groups 51-65 and 66+, which could reflect the typical retirement age and the natural reduction in workforce participation among older age groups. This information can be useful for understanding the age demographics within the dataset and for planning targeted health initiatives or policies.



```
SELECT
CASE
  WHEN age BETWEEN 0 AND 18 THEN '0-18'
  WHEN age BETWEEN 19 AND 35 THEN '19-35'
  WHEN age BETWEEN 36 AND 50 THEN '36-50'
  WHEN age BETWEEN 51 AND 65 THEN '51-65'
  ELSE '66+'
END AS AgeGroup,
COUNT(*) AS NumberEmployees
FROM Diabetes_prediction
GROUP BY
CASE
  WHEN age BETWEEN 0 AND 18 THEN '0-18'
  WHEN age BETWEEN 19 AND 35 THEN '19-35'
  WHEN age BETWEEN 36 AND 50 THEN '36-50'
  WHEN age BETWEEN 51 AND 65 THEN '51-65'
  ELSE '66+'
END
ORDER BY AgeGroup
```

|   | AgeGroup | NumberEmployees |
|---|----------|-----------------|
| 1 | 0-18     | 18400           |
| 2 | 19-35    | 21588           |
| 3 | 36-50    | 21549           |
| 4 | 51-65    | 21029           |
| 5 | 66+      | 17434           |

#### Categorical Data Analysis.

##### Distribution by Gender

1. **Male:**

- **Total Diabetes Cases: 4,039** males have been diagnosed with diabetes.
- **Total Heart Disease Cases: 2,380** males have been diagnosed with heart disease.
- **Total Hypertension Cases: 3,288** males have been diagnosed with hypertension.

## 2. Female:

- **Total Diabetes Cases: 4,461** females have been diagnosed with diabetes.
- **Total Heart Disease Cases: 1,562** females have been diagnosed with heart disease.
- **Total Hypertension Cases: 4,197** females have been diagnosed with hypertension.

## 3. Other:

- **Total Diabetes Cases:** No cases of diabetes reported.
- **Total Heart Disease Cases:** No cases of heart disease reported.
- **Total Hypertension Cases:** No cases of hypertension reported.

## Summary:

- **Diabetes:** The number of diabetes cases is slightly higher among females compared to males, with no cases reported for individuals classified as "Other".
- **Heart Disease:** Males have a higher number of heart disease cases compared to females, with no cases reported for the "Other" category.
- **Hypertension:** Females show a higher prevalence of hypertension compared to males, with no cases reported for the "Other" category.

This distribution helps in understanding the prevalence of various health conditions among different genders and can be useful for targeted health interventions and policy planning.

```
SELECT
  gender,
  SUM(CASE
    WHEN diabetes=1 THEN 1
    ELSE 0
  END) AS TotalDiabetes,
  SUM(CASE
    WHEN heart_disease=1 THEN 1
    ELSE 0
  END) AS TotalHeartDisease,
  SUM(CASE
    WHEN hypertension =1 THEN 1
    ELSE 0
  END) AS TotalHypertension
FROM Diabetes_prediction
GROUP BY gender;
```

|   | gender | TotalDiabetes | TotalHeartDisease | TotalHypertension |
|---|--------|---------------|-------------------|-------------------|
| 1 | Male   | 4039          | 2380              | 3288              |
| 2 | Female | 4461          | 1562              | 4197              |
| 3 | Other  | 0             | 0                 | 0                 |



## Disease Percentage by Age Group

### 1. Age Group 0-18:

- **Diabetes Percentage: 0%** of individuals in this age group have been diagnosed with diabetes.
- **Hypertension Percentage: 0%** of individuals in this age group have been diagnosed with hypertension.
- **Heart Disease Percentage: 0%** of individuals in this age group have been diagnosed with heart disease.
- **Total Employees: 18,400** individuals are in this age group.

### 2. Age Group 19-35:

- **Diabetes Percentage: 1%** of individuals in this age group have been diagnosed with diabetes.
- **Hypertension Percentage: 1%** of individuals in this age group have been diagnosed with hypertension.
- **Heart Disease Percentage: 0%** of individuals in this age group have been diagnosed with heart disease.
- **Total Employees: 21,588** individuals are in this age group.

### 3. Age Group 36-50:

- **Diabetes Percentage: 6%** of individuals in this age group have been diagnosed with diabetes.
- **Hypertension Percentage: 6%** of individuals in this age group have been diagnosed with hypertension.
- **Heart Disease Percentage: 1%** of individuals in this age group have been diagnosed with heart disease.
- **Total Employees: 21,549** individuals are in this age group.

### 4. Age Group 51-65:

- **Diabetes Percentage: 14%** of individuals in this age group have been diagnosed with diabetes.
- **Hypertension Percentage: 12%** of individuals in this age group have been diagnosed with hypertension.
- **Heart Disease Percentage: 5%** of individuals in this age group have been diagnosed with heart disease.

- **Total Employees: 21,029** individuals are in this age group.

## 5. Age Group 66+:

- **Diabetes Percentage: 20%** of individuals in this age group have been diagnosed with diabetes.
- **Hypertension Percentage: 18%** of individuals in this age group have been diagnosed with hypertension.
- **Heart Disease Percentage: 13%** of individuals in this age group have been diagnosed with heart disease.
- **Total Employees: 17,434** individuals are in this age group.

## Summary:

- **Diabetes:** The prevalence of diabetes increases with age, with the highest percentage observed in the 66+ age group.
- **Hypertension:** Hypertension also shows a higher prevalence in older age groups, peaking in the 66+ age group.
- **Heart Disease:** Heart disease percentage rises with age, with the highest incidence in the 66+ age group.

This analysis helps in understanding how different diseases are distributed across various age groups and can be useful for healthcare planning and resource allocation.

```
SELECT
CASE
  WHEN age BETWEEN 0 AND 18 THEN '0-18'
  WHEN age BETWEEN 19 AND 35 THEN '19-35'
  WHEN age BETWEEN 36 AND 50 THEN '36-50'
  WHEN age BETWEEN 51 AND 65 THEN '51-65'
  ELSE '66+'
END AS AgeGroup,
SUM(CASE
  WHEN diabetes=1 THEN 1
  ELSE 0
END)*100/COUNT(*) AS DiabetesPercentage,
SUM(CASE
  WHEN hypertension=1 THEN 1
  ELSE 0
END)*100/COUNT(*) AS HypertensionPercentage,
SUM(CASE
  WHEN heart_disease=1 THEN 1
  ELSE 0
END)*100/COUNT(*) AS HeartDiseasesPercentage,
COUNT(*) AS TotalEmployees
FROM Diabetes_prediction
GROUP BY
CASE
  WHEN age BETWEEN 0 AND 18 THEN '0-18'
  WHEN age BETWEEN 19 AND 35 THEN '19-35'
  WHEN age BETWEEN 36 AND 50 THEN '36-50'
  WHEN age BETWEEN 51 AND 65 THEN '51-65'
  ELSE '66+'
END
ORDER BY AgeGroup
```

|   | AgeGroup | DiabetesPercentage | HypertensionPercentage | HeartDiseasePercentage | TotalEmployees |
|---|----------|--------------------|------------------------|------------------------|----------------|
| 1 | 0-18     | 0                  | 0                      | 0                      | 18400          |
| 2 | 19-35    | 1                  | 1                      | 0                      | 21588          |
| 3 | 36-50    | 6                  | 6                      | 1                      | 21549          |
| 4 | 51-65    | 14                 | 12                     | 5                      | 21029          |
| 5 | 66+      | 20                 | 18                     | 13                     | 17434          |

Filtering and Analyzing Based on Conditions.

### Percentage of Diabetes and Hypertension

The query calculates the percentage of individuals who have both diabetes and hypertension among the total number of individuals in the dataset.

- **Calculation:** The query counts the number of records where both conditions (Diabetes = 1 and Hypertension = 1) are true. It then calculates the percentage of these individuals relative to the total number of records in the dataset.
- **Result:** The percentage of individuals who have both diabetes and hypertension is **2%**.

### Explanation:

- The result of **2%** indicates that among all individuals in the dataset, **2%** are diagnosed with both diabetes and hypertension. This percentage helps understand the overlap between these two health conditions within the dataset.

```
SELECT
SUM(
CASE
WHEN Diabetes=1 AND Hypertension =1 THEN 1
else 0
END) *100 /COUNT(*) AS DiabetesAndHypertensionPercentage
FROM Diabetes_prediction
```

| DiabetesAndHypertensionPercentage |
|-----------------------------------|
| 2                                 |

### Diabetes and Hypertension as PIVOT Table

The provided SQL query generates a pivot table to analyze the distribution of individuals based on their diabetes and hypertension status, segmented by gender and age group. Here's a detailed explanation:

### Breakdown:

### 1. Query Overview:

- **CTE Definition:** The Common Table Expression (CTE) categorizes each individual into one of four conditions (Both, None, Diabetes Only, hypertension Only) based on their diabetes and hypertension status. It also groups individuals into age categories.
- **PIVOT Operation:** The pivot table aggregates the counts of individuals falling into each condition category for different gender and age groups.

### 2. Categories:

- **Conditions:**
  - Both: Individuals with both diabetes and hypertension.
  - None: Individuals with neither condition.
  - Diabetes Only: Individuals with diabetes but without hypertension.
  - hypertension Only: Individuals with hypertension but without diabetes.
- **Age Groups:**
  - 0-18
  - 19-35
  - 36-55
  - 56-65
  - 66+

### 3. Table Structure:

- **Rows:** Each row represents a combination of gender and age group.
- **Columns:** Each column shows the count of individuals in one of the four condition categories (Both, None, Diabetes Only, hypertension Only).

### Explanation of Results:

- **For Females:**
  - **0-18:** There are no females with both conditions; however, there are **9083** with only hypertension and a small number with diabetes only.
  - **19-35:** Shows moderate counts for each category with the highest count for individuals with only hypertension.
  - **36-55:** A larger number of individuals with both conditions and a significant number with diabetes only.
  - **56-65:** A notable count of individuals with both conditions and a considerable number with diabetes only.

- **66+:** Higher counts across all categories, especially for those with both conditions.
- **For Males:**
  - **0-18:** Similar to females, with no individuals having both conditions but a significant number with only hypertension.
  - **19-35:** A lower count of individuals with both conditions compared to females, with more having hypertension only.
  - **36-55:** Higher counts in all categories, especially for diabetes only.
  - **56-65:** Significant number of individuals with diabetes only and those with both conditions.
  - **66+:** A relatively high number in all categories, particularly for individuals with both conditions.
- **For Others:**
  - **0-18:** Very few records, all with only hypertension.
  - **19-35 and 36-55:** Minimal counts across all conditions.

### Summary:

This pivot table provides a clear view of how diabetes and hypertension are distributed among different genders and age groups. It helps in understanding the prevalence of these conditions and can aid in targeted health interventions based on demographic factors.

```
WITH CTE AS (
    SELECT gender, diabetes,
           CASE
               WHEN diabetes=1 AND hypertension=1 THEN 'Both'
               WHEN diabetes=0 AND hypertension=0 THEN 'None'
               WHEN diabetes=1 AND hypertension=0 THEN 'Diabetes_Only'
               WHEN diabetes=0 AND hypertension=1 THEN 'hypertension_Only'
           END AS Condition,
           CASE
               WHEN age BETWEEN 0 AND 18 THEN '0-18'
               WHEN age BETWEEN 19 AND 35 THEN '19-35'
               WHEN age BETWEEN 36 AND 55 THEN '36-55'
               WHEN age BETWEEN 56 AND 65 THEN '56-65'
               ELSE '66+'
           END AS AgeGroup
    FROM Diabetes_prediction
)
SELECT
    gender,
    AgeGroup,
    ISNULL([Both], 0) AS Both,
    ISNULL([None], 0) AS None,
    ISNULL([Diabetes_Only], 0) AS OnlyDiabetes,
    ISNULL([hypertension_Only], 0) AS OnlyHypertension
FROM
    CTE
PIVOT (
    COUNT(diabetes) FOR Condition IN ([Both], [None], [Diabetes_Only], [hypertension_Only])
) AS pt
```

|    | gender | AgeGroup | Both | None  | OnlyDiabetes | OnlyHypertension |
|----|--------|----------|------|-------|--------------|------------------|
| 1  | Female | 0-18     | 0    | 9083  | 52           | 5                |
| 2  | Female | 19-35    | 26   | 13665 | 194          | 117              |
| 3  | Female | 36-55    | 231  | 15526 | 939          | 861              |
| 4  | Female | 56-65    | 305  | 5843  | 815          | 714              |
| 5  | Female | 66+      | 556  | 6895  | 1343         | 1382             |
| 6  | Male   | 0-18     | 0    | 9210  | 40           | 5                |
| 7  | Male   | 19-35    | 18   | 7283  | 149          | 130              |
| 8  | Male   | 36-55    | 249  | 9849  | 829          | 763              |
| 9  | Male   | 56-65    | 265  | 3956  | 824          | 602              |
| 10 | Male   | 66+      | 438  | 4775  | 1227         | 818              |
| 11 | Other  | 0-18     | 0    | 5     | 0            | 0                |
| 12 | Other  | 19-35    | 0    | 6     | 0            | 0                |
| 13 | Other  | 36-55    | 0    | 7     | 0            | 0                |

## Smoking History vs. Disease Prevalence

The SQL query analyzes the relationship between smoking history and the prevalence of diabetes and heart disease among individuals in the dataset. Here's a detailed explanation of the results:

### Explanation of Results:

#### 1. Overview:

- The query groups individuals by their smoking history and calculates the percentage of individuals with diabetes and heart disease for each group.

#### 2. Categories:

- **Smoking History:**
  - current: Individuals who are currently smoking.
  - not current: Individuals who were previously smoking but are not currently.
  - former: Individuals who have smoked in the past but are not current smokers.
  - ever: Individuals who have ever smoked.
  - No Info: Individuals for whom smoking history information is not available.
  - never: Individuals who have never smoked.

#### 3. Disease Prevalence Percentages:

- **Current Smokers:**
  - **Diabetes: 10%**
  - **Heart Disease: 4%**
  - Current smokers have a moderate percentage of diabetes and a lower percentage of heart disease compared to some other groups.
- **Not Current Smokers:**
  - **Diabetes: 10%**
  - **Heart Disease: 4%**

- Individuals in this category show similar percentages to current smokers, indicating similar prevalence rates of diabetes and heart disease.
- **Former Smokers:**
  - **Diabetes: 17%**
  - **Heart Disease: 9%**
  - Former smokers have higher percentages of both diabetes and heart disease compared to current and not current smokers, suggesting a potential long-term impact of smoking history on disease prevalence.
- **Ever Smokers:**
  - **Diabetes: 11%**
  - **Heart Disease: 7%**
  - This category includes all individuals who have smoked at any point in their life, and the disease prevalence percentages are relatively high, reflecting the cumulative effect of smoking on health.
- **No Info:**
  - **Diabetes: 4%**
  - **Heart Disease: 2%**
  - Individuals with no available information on smoking history have the lowest disease prevalence percentages, which might reflect either a lack of data or a lower disease prevalence in this group.
- **Never Smokers:**
  - **Diabetes: 9%**
  - **Heart Disease: 3%**
  - Never smokers show a relatively low percentage of both diseases compared to former smokers and ever smokers, which aligns with the expectation that smoking history may contribute to higher disease rates.

### **Summary:**

The analysis highlights how smoking history correlates with the prevalence of diabetes and heart disease. Former smokers tend to have higher disease percentages, indicating that the health effects of smoking may persist even after quitting. In contrast, current and not current smokers show similar disease rates, while individuals with no smoking history generally have lower disease prevalence.

```

SELECT
    smoking_history,
    SUM(CASE
        WHEN Diabetes=1 THEN 1
        ELSE 0
    END)*100 / COUNT(*) AS DiabetesPercentage,
    SUM(CASE
        WHEN heart_disease =1 THEN 1
        ELSE 0
    END)*100 /COUNT(*) AS HeartDiseasePercentage
FROM Diabetes_prediction
GROUP BY smoking_history

```

| smoking_history | DiabetesPercentage | HeartDiseasePercentage |
|-----------------|--------------------|------------------------|
| current         | 10                 | 4                      |
| not current     | 10                 | 4                      |
| former          | 17                 | 9                      |
| ever            | 11                 | 7                      |
| No Info         | 4                  | 2                      |
| never           | 9                  | 3                      |

## Diabetes & Heart Disease by Gender

The SQL query examines the relationship between diabetes and heart disease across different genders. Here's a breakdown of the results:

### Explanation of Results:

#### 1. Overview:

- The query categorizes individuals based on their diabetes and heart disease status and then pivots this data to show the counts for each category by gender.

#### 2. Categories:

- **Both:** Individuals with both diabetes and heart disease.
- **None:** Individuals with neither diabetes nor heart disease.
- **Diabetes Only:** Individuals with diabetes but no heart disease.
- **Heart\_Disease\_Only:** Individuals with heart disease but no diabetes.

#### 3. Results by Gender:

- **Female:**
  - **Both: 526** individuals have both diabetes and heart disease.
  - **None: 53,055** individuals have neither diabetes nor heart disease.
  - **Diabetes Only: 3,935** individuals have diabetes but no heart disease.



- **Heart\_Disease\_Only: 1,036** individuals have heart disease but no diabetes.
- **Summary:** The largest group is females with neither condition. The number of females with both conditions is relatively low compared to those with only one condition or none.
- **Male:**
  - **Both: 741** individuals have both diabetes and heart disease.
  - **None: 35,752** individuals have neither diabetes nor heart disease.
  - **Diabetes\_Only: 3,298** individuals have diabetes but no heart disease.
  - **Heart\_Disease\_Only: 1,639** individuals have heart disease but no diabetes.
  - **Summary:** The number of males with both conditions is slightly higher compared to females. Like females, most males fall into the category of having neither condition.
- **Other:**
  - **Both: 0** individuals have both conditions.
  - **None: 18** individuals have neither condition.
  - **Diabetes\_Only: 0** individuals have diabetes only.
  - **Heart\_Disease\_Only: 0** individuals have heart disease only.
  - **Summary:** The "Other" gender category has very few individuals, and none of them have either condition.

### **Summary:**

The data reveals that a significant number of individuals across genders do not have either diabetes or heart disease. Among those who do have one of the conditions, females and males show similar patterns, with more males having both conditions compared to females. The "Other" gender category has minimal representation and no cases of diabetes or heart disease, indicating it may not have sufficient data for meaningful analysis.

```

WITH CTE AS (
    SELECT
        gender, diabetes,
        CASE
            WHEN diabetes=1 AND heart_disease=1 THEN 'Both'
            WHEN diabetes=0 AND heart_disease=0 THEN 'None'
            WHEN diabetes=1 AND heart_disease=0 THEN 'Diabetes_Only'
            WHEN diabetes=0 AND heart_disease=1 THEN 'Heart_Disease_Only'
        END AS Condition
    FROM
        Diabetes_prediction
)
SELECT
    gender,
    ISNULL([Both], 0) AS Both,
    ISNULL([None], 0) AS None,
    ISNULL([Diabetes_Only], 0) AS OnlyDiabetes,
    ISNULL([Heart_Disease_Only], 0) AS Only_Heart_Disease
FROM
    CTE
PIVOT (
    COUNT(diabetes) FOR Condition IN ([Both], [None], [Diabetes_Only], [Heart_Disease_Only])
) AS pt
ORDER BY gender;

```

|   | gender | Both | None  | OnlyDiabetes | Only_Heart_Disease |
|---|--------|------|-------|--------------|--------------------|
| 1 | Female | 526  | 53055 | 3935         | 1036               |
| 2 | Male   | 741  | 35752 | 3298         | 1639               |
| 3 | Other  | 0    | 18    | 0            | 0                  |

## Proportion Test Results

The SQL query calculates the proportion of individuals with and without diabetes across different genders. Here's an explanation of the results:

### Explanation of Results:

#### 1. Overview:

- The query calculates the count and proportion of individuals with (diabetes = 1) and without diabetes (diabetes = 0) for each gender.

#### 2. Results by Gender:

- **Female:**
  - **Diabetes (1):**
    - **Count: 4,461** individuals.
    - **Proportion: 4.46%** of the total population.
  - **No Diabetes (0):**
    - **Count: 54,091** individuals.
    - **Proportion: 54.09%** of the total population.
- **Male:**
  - **Diabetes (1):**
    - **Count: 4,039** individuals.
    - **Proportion: 4.04%** of the total population.

- **No Diabetes (0):**
  - **Count: 37,391** individuals.
  - **Proportion: 37.39%** of the total population.
- **Other:**
  - **Diabetes (1):**
    - **Count: 0** individuals.
    - **Proportion: 0.02%** of the total population.
  - **No Diabetes (0):**
    - **Count: 18** individuals.
    - **Proportion: 0.02%** of the total population.

### Summary:

- **Female:** A higher proportion of females do not have diabetes compared to those who have it.
- **Male:** A similar pattern is observed where a higher proportion of males do not have diabetes compared to those who do.
- **Other:** The "Other" gender category has minimal representation, and the data shows no individuals with diabetes in this category.

Overall, the proportions reflect the distribution of diabetes across different genders in the dataset, with both females and males having a higher proportion of non-diabetic individuals compared to those with diabetes. The "Other" category has very few individuals, leading to very small proportions.

```
SELECT
  gender,
  diabetes,
  COUNT(*) AS TotalCount,
  (COUNT(*)*1.0/
   (SELECT COUNT(*)
    FROM Diabetes_prediction)) AS Proportion
FROM Diabetes_prediction
GROUP BY
  gender, diabetes
order by gender
```

|   | gender | diabetes | TotalCount | Proportion     |
|---|--------|----------|------------|----------------|
| 1 | Female | 1        | 4461       | 0.044610000000 |
| 2 | Female | 0        | 54091      | 0.540910000000 |
| 3 | Male   | 0        | 37391      | 0.373910000000 |
| 4 | Male   | 1        | 4039       | 0.040390000000 |
| 5 | Other  | 0        | 18         | 0.000180000000 |

- **Recommendation:** The 51-65 age group has the highest percentages of diabetes and hypertension. Implementing health screenings and educational programs in this demographic could be beneficial.
- **Recommendation:** Develop and promote health programs that are tailored to different genders. the data shows that females have a higher percentage of diabetes compared to males. Gender-specific campaigns may improve outreach and effectiveness.