Datathon FY'25

Team 3

CONVERTING THE CLEAN DATA CSV FILE TO A SQLITE DATABASE

Python Imports

In [1]:
```
# sqlite library
import sqlite3 as sql
import pandas as pd

import warnings
warnings.filterwarnings('ignore')
```

Conversion of CSV to SQLite

The conversion of the dataset within CSV to an SQLite Database involves creating a pandas Dataframe; creating a new SQLite database, then loading the Dataframe into this new database as a new table.

We have determined that the dataset as it is, is in 3NF and all column parameters are independent of each other, with only the Diabetes column being the primary reference. We will create a column of primary keys as an index for the `diabetes_patient_data` table. We will also create 2 columns to bin or group age into age groups (`0-9` , `10-19` , etc), and BMI into BMI levels (`Underweight` , `Normal` , `Overweight` , `Obese`).

In [2]:
```
# Importing Dataset
clean_data_import = pd.read_csv('../data/cleaned_data/diabetes_data_cleansed.csv')

#Creating & connecting to the sqlite db
connection = sql.connect("data.db")
```

```
#Load CSV data to SQLite
clean_data_import.to_sql('diabetes_patient_data', connection, if_exists='replace')

cursor = connection.cursor()
query = "SELECT * FROM diabetes_patient_data LIMIT 5"
cursor.execute(query)
result = cursor.fetchall()
for item in result:
    print(item)
```

```
(0, 'male', 59.0, 0.0, 0.73, None, 17.1, 156.5, 0.0, 'Occasionally', 'Lightly Active', 5.4, 'Moderate', 0.0, 'none', 1.0)
(1, 'female', 31.0, None, 0.4649999999999999, None, 22.5, 137.8, None, 'Occasionally', 'Lightly Active', 7.6, 'Low', 0.0, 'ligh
t', 1.0)
(2, 'male', 48.1, 0.0, 0.2099999999999999, 'vegetarian', 24.8, 91.1, 1.0, 'Never', 'Sedentary', 1.0999999999999996, 'Moderate',
0.0, None, 1.0)
(3, 'female', 67.0, 0.0, 0.27, 'mediterranean', 21.2, 135.83, 0.0, 'Moderate', 'Moderately Active', 6.8, 'Moderate', 1.0, None,
1.0)
(4, None, 24.0, 0.0, 0.31, 'raw food', 30.7, 210.5, 1.0, 'Never', 'Sedentary', 10.1, 'Extreme', 0.0, 'moderate', 1.0)
```

Updating data types

SQLite DB data types are limited to the following storage classes: `NULL`, `INTEGER`, `REAL`, `TEXT`, `BLOB`. Data conversion of data types to be used in Power BI will be done using DAX.

Adding Columns

Importing the data into SQLite has added an `index` column, so we can focus on adding the columns for binning.

Adding binned BMI & Age columns

In [3]:
```
# Add a new column 'bmi_bins' to the table
#  Update the new column 'bmi_bins' with bmi ranges
update_query_bmi = """
    ALTER TABLE diabetes_patient_data ADD COLUMN bmi_bins TEXT;
    UPDATE diabetes_patient_data
    SET bmi_bins = CASE
        WHEN bmi < 18.5 THEN 'Underweight'
        WHEN bmi >= 18.5 AND bmi < 25 THEN 'Normal'
```

```
        WHEN bmi >= 25 AND bmi < 30 THEN 'Overweight'
        WHEN bmi >= 30 THEN 'Obese'
        ELSE 'Unknown'
    END;
    SELECT * FROM diabetes_patient_data LIMIT 5;
"""

cursor.executescript(update_query_bmi)
result = cursor.fetchall()
for item in result:
    print(item)
```

In [4]:
```
# Add a new column 'age_bins' to the table
#  Update the new column 'age_bins' with categorized age ranges
update_query_age = """
    ALTER TABLE diabetes_patient_data ADD COLUMN age_bins TEXT;
    UPDATE diabetes_patient_data
    SET age_bins = CASE
        WHEN age <= 10 THEN '0-10'
        WHEN age > 10 AND age <= 20 THEN '11-20'
        WHEN age > 20 AND age <= 30 THEN '21-30'
        WHEN age > 30 AND age <= 40 THEN '31-40'
        WHEN age > 40 AND age <= 50 THEN '41-50'
        WHEN age > 50 AND age <= 60 THEN '51-60'
        WHEN age > 60 AND age <= 70 THEN '61-70'
        WHEN age > 70 AND age <= 80 THEN '71-80'
        WHEN age > 80 AND age <= 90 THEN '81-90'
        WHEN age > 90 AND age <= 100 THEN '91-100'
        ELSE 'Unknown'
    END;
    SELECT * FROM diabetes_patient_data LIMIT 5;
"""

cursor.executescript(update_query_age)
result = cursor.fetchall()
for item in result:
    print(item)
```

In [5]:
```
select_query = "SELECT * FROM diabetes_patient_data WHERE diabetes IS 1 LIMIT 5"
cursor.execute(select_query)
```

```
result = cursor.fetchall()
for item in result:
    print(item)
```

```
(0, 'male', 59.0, 0.0, 0.73, None, 17.1, 156.5, 0.0, 'Occasionally', 'Lightly Active', 5.4, 'Moderate', 0.0, 'none', 1.0, 'Unde
rweight', '51-60')
(1, 'female', 31.0, None, 0.4649999999999999, None, 22.5, 137.8, None, 'Occasionally', 'Lightly Active', 7.6, 'Low', 0.0, 'ligh
t', 1.0, 'Normal', '31-40')
(2, 'male', 48.1, 0.0, 0.2099999999999999, 'vegetarian', 24.8, 91.1, 1.0, 'Never', 'Sedentary', 1.0999999999999996, 'Moderate',
0.0, None, 1.0, 'Normal', '41-50')
(3, 'female', 67.0, 0.0, 0.27, 'mediterranean', 21.2, 135.83, 0.0, 'Moderate', 'Moderately Active', 6.8, 'Moderate', 1.0, None,
1.0, 'Normal', '61-70')
(4, None, 24.0, 0.0, 0.31, 'raw food', 30.7, 210.5, 1.0, 'Never', 'Sedentary', 10.1, 'Extreme', 0.0, 'moderate', 1.0, 'Obese',
'21-30')
```

Creating relevant views

In [6]:
```
diabetes_view_query = """
    CREATE VIEW diabetes_prevalence_view AS
    SELECT diabetes, COUNT(*)
    FROM diabetes_patient_data WHERE diabetes IN (0,1) GROUP BY diabetes;
"""
cursor.executescript(diabetes_view_query)
result = cursor.fetchall()
for item in result:
    print(item)
```

In [7]:
```
comorbidity_view_query = """
    CREATE VIEW comorbidity_view AS
    SELECT diabetes, hypertension, age, gender, diet_type
    FROM diabetes_patient_data WHERE diabetes = 1;
"""
cursor.executescript(comorbidity_view_query)
result = cursor.fetchall()
for item in result:
    print(item)
```

In [8]:
```
social_view_query = """
    CREATE VIEW social_media_usage_view AS
    SELECT social_media_usage, diabetes, age, gender
```

```
    FROM diabetes_patient_data WHERE diabetes = 1;
"""
cursor.executescript(social_view_query)
result = cursor.fetchall()
for item in result:
    print(item)
```

In [9]:
```
family_history_view_query = """
    CREATE VIEW family_history_view AS
    SELECT diabetes, age, gender, family_diabetes_history, pregnancies
    FROM diabetes_patient_data WHERE diabetes = 1;
"""
cursor.executescript(family_history_view_query)
result = cursor.fetchall()
for item in result:
    print(item)
```

In [10]:
```
lifestyle_view_query = """
    CREATE VIEW lifestyle_view AS
    SELECT diabetes, physical_activity_level, stress_level, alcohol_consumption, diet_type, age, gender
    FROM diabetes_patient_data WHERE diabetes = 1;
"""
cursor.executescript(lifestyle_view_query)
result = cursor.fetchall()
for item in result:
    print(item)
```

In [11]:
```
bmi_view_query = """
    CREATE VIEW bmi_view AS
    SELECT diabetes, bmi_bins, weight
    FROM diabetes_patient_data WHERE diabetes = 1;
"""
cursor.executescript(bmi_view_query)
result = cursor.fetchall()
for item in result:
    print(item)
```

In [12]:
```
demographic_query = """
    CREATE VIEW pop_demographic_view AS
```

```
    SELECT diabetes, bmi_bins, weight
    FROM diabetes_patient_data WHERE diabetes = 1;
"""
cursor.executescript(demographic_query)
result = cursor.fetchall()
for item in result:
    print(item)
```

In [13]:
```
diabetes_select_query = """
    SELECT *
    FROM diabetes_patient_data LIMIT 5;
"""
cursor.execute(diabetes_select_query)
result = cursor.fetchall()
for item in result:
    print(item)
```

```
(0, 'male', 59.0, 0.0, 0.73, None, 17.1, 156.5, 0.0, 'Occasionally', 'Lightly Active', 5.4, 'Moderate', 0.0, 'none', 1.0, 'Underweight', '51-60')
(1, 'female', 31.0, None, 0.4649999999999999, None, 22.5, 137.8, None, 'Occasionally', 'Lightly Active', 7.6, 'Low', 0.0, 'light', 1.0, 'Normal', '31-40')
(2, 'male', 48.1, 0.0, 0.2099999999999999, 'vegetarian', 24.8, 91.1, 1.0, 'Never', 'Sedentary', 1.0999999999999996, 'Moderate', 0.0, None, 1.0, 'Normal', '41-50')
(3, 'female', 67.0, 0.0, 0.27, 'mediterranean', 21.2, 135.83, 0.0, 'Moderate', 'Moderately Active', 6.8, 'Moderate', 1.0, None, 1.0, 'Normal', '61-70')
(4, None, 24.0, 0.0, 0.31, 'raw food', 30.7, 210.5, 1.0, 'Never', 'Sedentary', 10.1, 'Extreme', 0.0, 'moderate', 1.0, 'Obese', '21-30')
```

In [14]:
```
comorbidity_select_query = "SELECT * FROM comorbidity_view LIMIT 5;"
cursor.execute(comorbidity_select_query)
result = cursor.fetchall()
for item in result:
    print(item)
```

```
(1.0, 0.0, 59.0, 'male', None)
(1.0, None, 31.0, 'female', None)
(1.0, 0.0, 48.1, 'male', 'vegetarian')
(1.0, 0.0, 67.0, 'female', 'mediterranean')
(1.0, 0.0, 24.0, None, 'raw food')
```

In [15]:
```
#Close connection
connection.close()
```