**DSCI D532 : HOMEWORK 1 (**[**schekka@iu.edu**](mailto:schekka@iu.edu)**)**

CREATE DATABASE adt\_hw1;

-- Creating the hospitals table by following all the constraints properly

CREATE TABLE IF NOT EXISTS hospitals (

hospital\_id INT AUTO\_INCREMENT PRIMARY KEY,

hospital\_name VARCHAR(255) NOT NULL,

state CHAR(2) NOT NULL,

city VARCHAR(100) NOT NULL,

doctor VARCHAR(255) NOT NULL

);

-- Creating the doctors table by following all the constraints properly

CREATE TABLE IF NOT EXISTS doctors (

doctor\_id INT AUTO\_INCREMENT PRIMARY KEY,

name VARCHAR(255) NOT NULL,

gender CHAR(7) CHECK (gender IN ('Male', 'Female', 'O')),

insurance CHAR(3) CHECK (insurance IN ('Yes', 'No')),

new\_patients CHAR(3) CHECK (new\_patients IN ('Yes', 'No')),

speciality\_one VARCHAR(100),

speciality\_two VARCHAR(100),

speciality\_three VARCHAR(100),

license CHAR(3) CHECK (license IN ('MFT', 'PhD', 'MD')),

phone CHAR(10)

);

INSERT INTO

doctors(name, gender, insurance, new\_patients, speciality\_one,

speciality\_two, speciality\_three, license, phone)

VALUES

('Flora Martinez', 'Female', 'Yes', 'Yes', 'Diabetes', 'Cholesterol', 'immunology', 'MD', '8495776489'),

('Andy James', 'Male', 'Yes', 'No', 'Hypertension', 'Diabetes', 'PTSD', 'PhD', '2340894766'),

('Hannah Myers', 'Female', 'No', 'Yes', 'Diabetes', 'Hypertension', 'immunology', 'MD', '9907846574'),

('Jane Huang', 'Female', 'Yes', 'Yes', 'Dermatology', 'Hypertension', 'immunology', 'MD', '4507856797'),

('April Adams', 'Female', 'No', 'Yes', 'OCD', 'Hypertension', 'PTSD', 'MFT', '4507856797'),

('Jon Schaffer', 'Male', 'Yes', 'No', 'BPD', 'immunology', 'Dermatology', 'PhD', '9907846574'),

('Shauna West','Female', 'Yes', 'Yes', 'ADHD', 'immunology', 'OCD', 'MD', '8495776480'),

('Juan Angelo', 'Male', 'No', 'Yes', 'Diabetes', 'immunology', 'Dermatology', 'MD', '4507856797'),

('Christie Yang', 'Female', 'Yes', 'Yes', 'Autism', 'ADHD', 'OCD', 'PhD', '4507856796'),

('Annika Neusler', 'Female', 'Yes', 'No', 'Addiction', 'Dermatology', 'PTSD', 'MFT', '9907846575'),

('Simone Anderson', 'Female', 'No', 'No', 'Hypertension', 'Dermatology', 'PTSD', 'MD', '8304498765'),

('Ted Nyguen', 'Male', 'Yes', 'Yes', 'ADHD', 'Hypertension', 'Allergy', 'PhD', '4301239990'),

('Valentino Rossi', 'Male', 'Yes', 'Yes', 'Autism', 'Hypertension', 'Dermatology', 'MD', '8304498765'),

('Jessica Armer', 'Female', 'No', 'Yes', 'PTSD', 'immunology', 'Dermatology', 'MD', '3330456612'),

('Sid Michaels', 'Female', 'Yes', 'Yes', 'OCD', 'Allergy', 'Hypertension', 'MFT', '4301239997'),

('Yen Waters', 'Male', 'Yes', 'Yes', 'Hypertension', 'Dermatology', 'ADHD', 'PhD', '4507856796'),

('Ru Izaelia', 'Female', 'No', 'Yes', 'immunology', 'BPD', 'Allergy', 'MD', '4301239990'),

('Vishal Rao', 'Male', 'Yes', 'Yes', 'Dermatology', 'Diabetes', 'Hypertension', 'MD', '7305557894'),

('Lana John', 'Female', 'Yes', 'Yes', 'Hypertension', 'Allergy', 'OCD', 'MFT', '7305557894'),

('Izzie Geralt', 'Female', 'Yes', 'Yes', 'Dermatology', 'Addiction', 'Hypertension', 'MD', '4301239990');

INSERT INTO

hospitals(hospital\_name, state, city, doctor)

VALUES

('Van Holsen Community Hospital', 'CA', 'San Francisco', 'Flora Martinez'),

('Clear Water Services', 'CA', 'San Diego', 'Andy James'),

('Imagery Health', 'CA', 'Sacramento', 'Hannah Myers'),

('Blue Cross Clinic', 'CA', 'Los Angeles', 'Jane Huang'),

('Blue Cross Clinic', 'CA', 'Los Angeles', 'April Adams'),

('Imagery Health', 'CA', 'Sacramento', 'Jon Schaffer'),

('Van Holsen Community Hospital', 'CA', 'Long Beach', 'Shauna West'),

('Blue Cross Clinic', 'CA', 'Santa Barbara', 'Juan Angelo'),

('Blue Cross Clinic', 'CA', 'San Francisco', 'Christie Yang'),

('Imagery Health', 'CA', 'Auburn', 'Annika Neusler'),

('Holistic Health Services', 'CA', 'Santa Barbara', 'Simone Anderson'),

('Open Clinic', 'CA', 'San Jose', 'Ted Nyguen'),

('Holistic Health Services', 'CA', 'Santa Barbara', 'Valentino Rossi'),

('Clark Jamison Hospitals', 'CA', 'Fresno', 'Jessica Armer'),

('Open Clinic', 'CA', 'Oakland', 'Sid Michaels'),

('Blue Cross Clinic', 'CA', 'San Francisco', 'Yen Waters'),

('Open Clinic', 'CA', 'San Jose', 'Ru Izaelia'),

('Clear Minds Community', 'CA', 'Sacramento', 'Vishal Rao'),

('Clear Minds Community', 'CA', 'Sacramento', 'Lana John'),

('Open Clinic', 'CA', 'San Jose', 'Izzie Geralt');

SELECT \* FROM doctors;

SELECT \* FROM hospitals;

-- Alter the hospitals table to add the doctor\_id column

ALTER TABLE hospitals

ADD COLUMN doctor\_id INT NOT NULL;

UPDATE hospitals

INNER JOIN doctors

ON hospitals.doctor = doctors.name

SET hospitals.doctor\_id = doctors.doctor\_id;

CREATE TABLE IF NOT EXISTS specialties(

specialties\_key serial PRIMARY KEY,

speciality\_one VARCHAR(100),

speciality\_two VARCHAR(100),

speciality\_three VARCHAR(100),

doctor\_id INTEGER,

CONSTRAINT fk\_doctor

FOREIGN KEY(doctor\_id)

REFERENCES doctors(doctor\_id));

INSERT INTO specialties(doctor\_id, speciality\_one, speciality\_two, speciality\_three)

SELECT doctor\_id, speciality\_one, speciality\_two, speciality\_three

FROM doctors;

ALTER TABLE doctors

DROP COLUMN speciality\_one,

DROP COLUMN speciality\_two,

DROP COLUMN speciality\_three;

-- Exploration

-- 1. How many doctors are currently accepting new patients?

SELECT COUNT(\*) AS doctors\_accepting\_new\_patients

FROM doctors

WHERE new\_patients = 'Yes';

-- 2. What is the distribution of doctors across different cities in California?

SELECT city, COUNT(doctor\_id) AS number\_of\_doctors

FROM hospitals

GROUP BY city;

-- 3. How many male and female doctors have each type of license?

SELECT

license,

SUM(CASE WHEN gender = 'Male' THEN 1 ELSE 0 END) AS male\_doctors,

SUM(CASE WHEN gender = 'Female' THEN 1 ELSE 0 END) AS female\_doctors,

SUM(CASE WHEN gender = 'O' THEN 1 ELSE 0 END) AS other\_gender\_doctors

FROM doctors

GROUP BY license;

-- 4. How many doctors have a license in 'MD' and are treating 'Diabetes'?

SELECT COUNT(\*) AS md\_diabetes\_doctors

FROM doctors

WHERE license = 'MD' AND doctor\_id IN (

SELECT doctor\_id

FROM specialties

WHERE speciality\_one = 'Diabetes' OR speciality\_two = 'Diabetes' OR speciality\_three = 'Diabetes'

);

-- 5. What is the average number of doctors per hospital in the database?

SELECT AVG(doctor\_count) as average\_doctors\_per\_hospital

FROM (

SELECT hospital\_id, COUNT(doctor\_id) as doctor\_count

FROM hospitals

GROUP BY hospital\_id

) AS doctor\_counts;

-- Visualization

-- Visualization 1: Distribution of Doctors by Specialty

SELECT speciality\_one, COUNT(\*) AS number\_of\_doctors

FROM specialties

GROUP BY speciality\_one

ORDER BY number\_of\_doctors DESC;

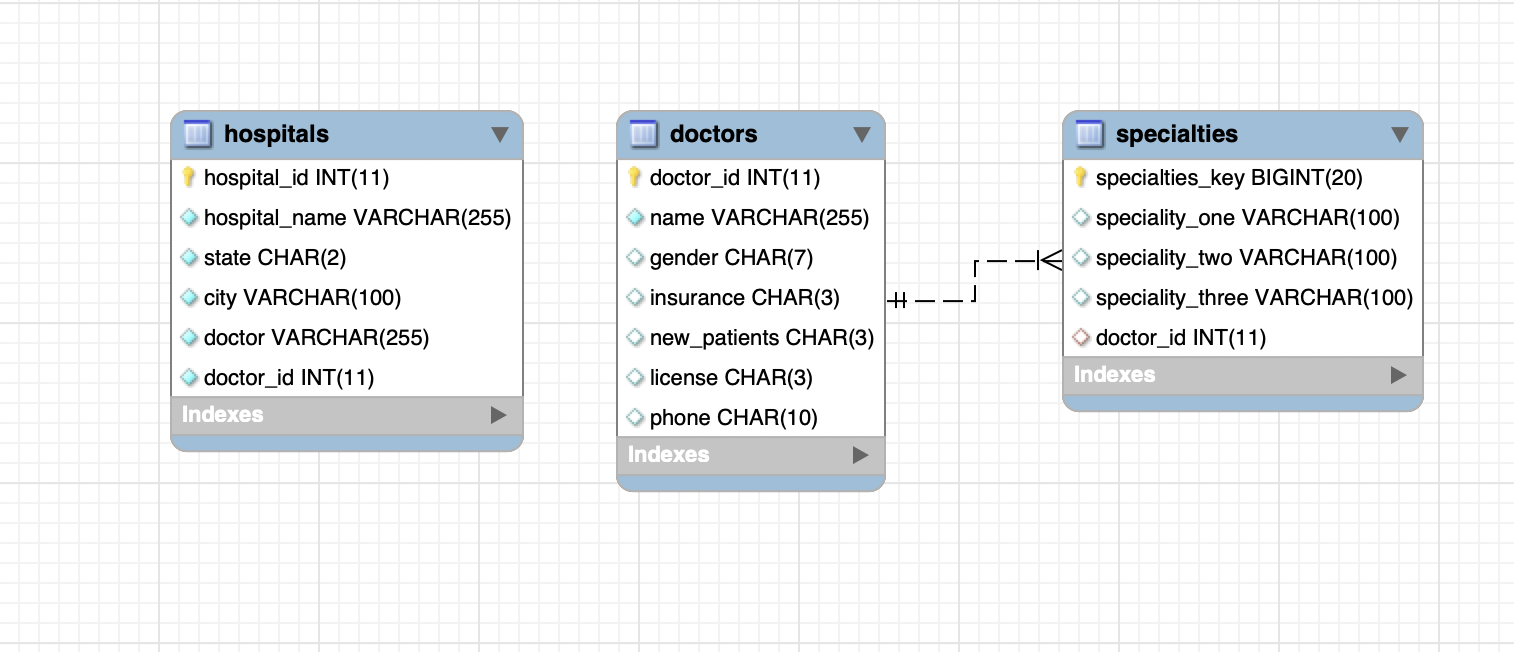
-- Visualization 2: Acceptance of New Patients by Gender

SELECT gender, new\_patients, COUNT(\*) AS number\_of\_doctors

FROM doctors

GROUP BY gender, new\_patients;

**Schema of the normalized database(EER Diagram):**

****

**Python notebook: Exploration**

**config = {**

**'user': 'root',**

**'password': 'root',**

**'host': '127.0.0.1',**

**'port': 8889,**

**'database': 'adt\_hw1',**

**'raise\_on\_warnings': True**

**}**

**pip install mysql-connector-python**

**# create a MySQL server connection object**

**import mysql.connector**

**mydb = mysql.connector.connect(\*\*config)**

**# cursor object**

**my\_cursor = mydb.cursor(dictionary=True)**

**import mysql.connector**

**import pandas as pd**

**# Function to run a single query and return the results as a pandas DataFrame**

**def run\_query(query):**

**my\_cursor.execute(query)**

**rows = my\_cursor.fetchall()**

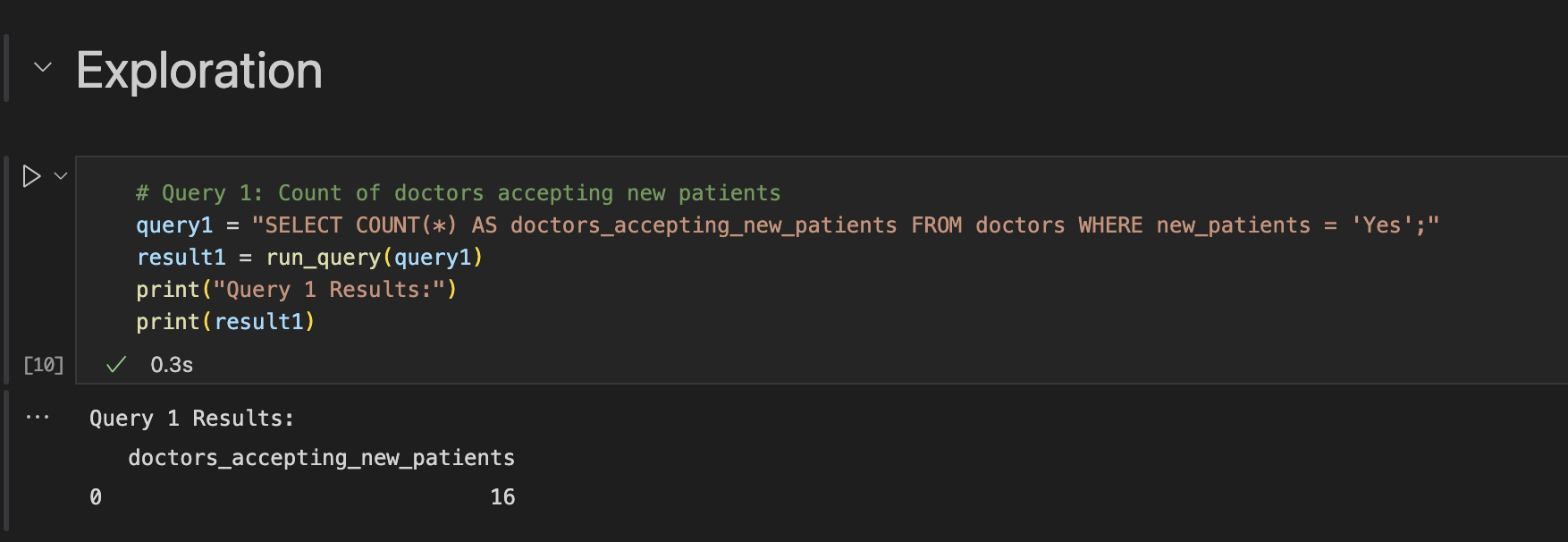
**return pd.DataFrame(rows)**

**Query 1:**

**SELECT COUNT(\*) AS doctors\_accepting\_new\_patients**

**FROM doctors**

**WHERE new\_patients = 'Yes';**

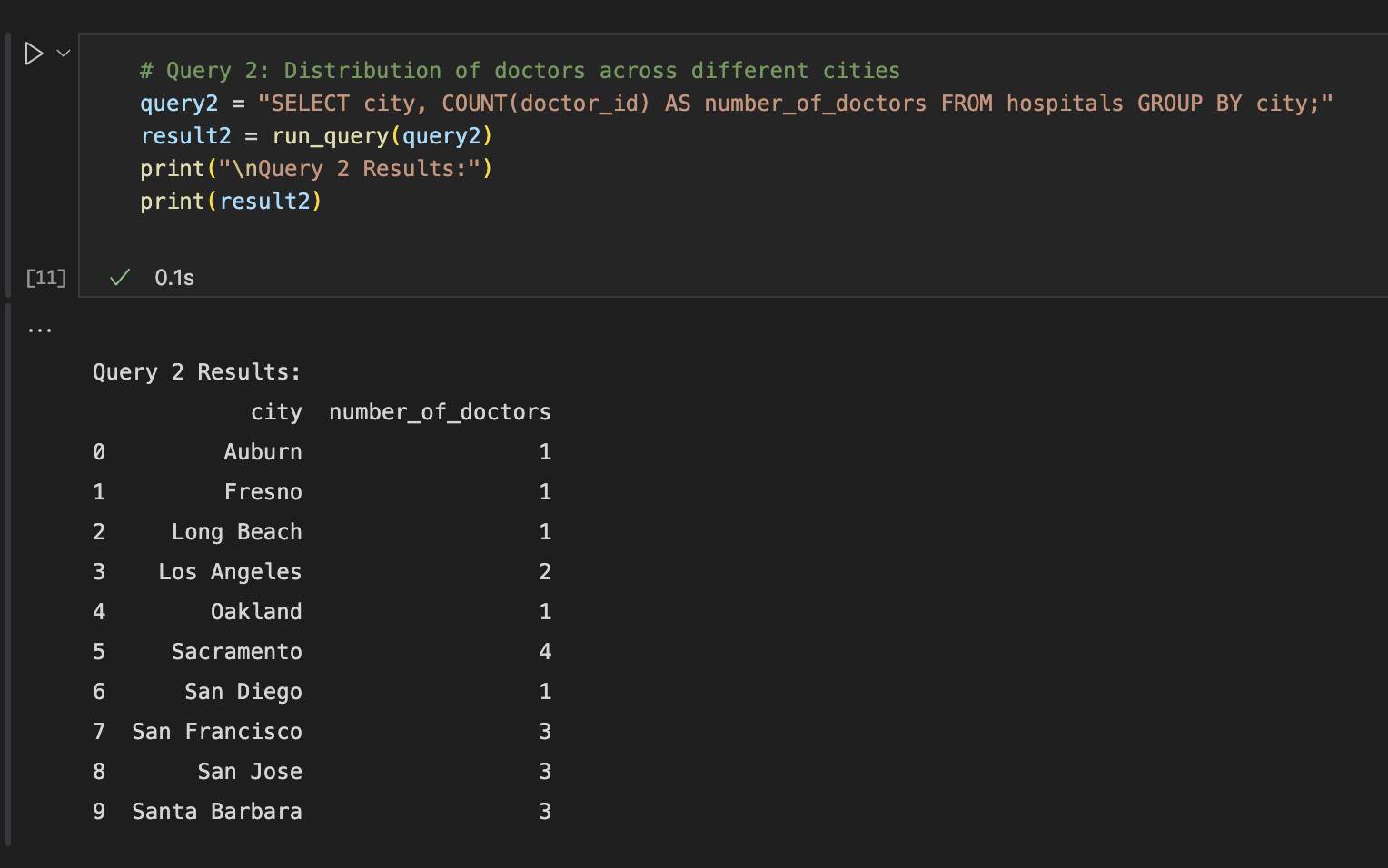
****

**Query 2**

**SELECT city, COUNT(doctor\_id) AS number\_of\_doctors**

**FROM hospitals**

**GROUP BY city;**

****

**Query 3:**

**SELECT**

**license,**

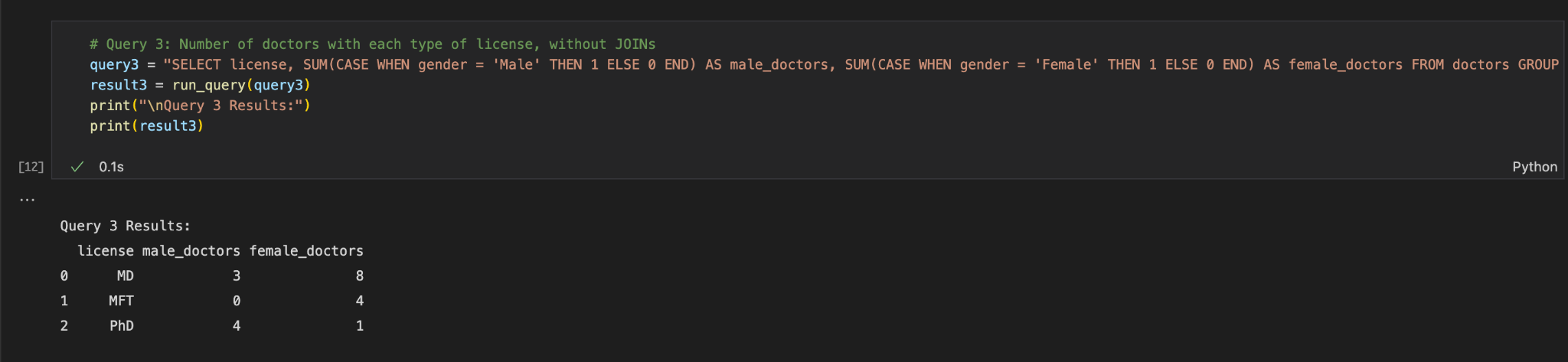
**SUM(CASE WHEN gender = 'Male' THEN 1 ELSE 0 END) AS male\_doctors,**

**SUM(CASE WHEN gender = 'Female' THEN 1 ELSE 0 END) AS female\_doctors,**

**SUM(CASE WHEN gender = 'O' THEN 1 ELSE 0 END) AS other\_gender\_doctors**

**FROM doctors**

**GROUP BY license;**

****

**Query 4 :**

**SELECT COUNT(\*) AS md\_diabetes\_doctors**

**FROM doctors**

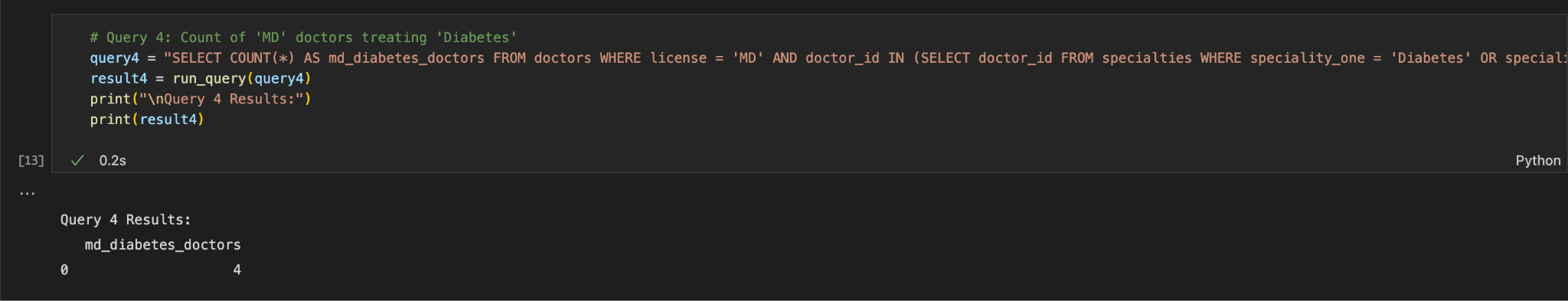
**WHERE license = 'MD' AND doctor\_id IN (**

**SELECT doctor\_id**

**FROM specialties**

**WHERE speciality\_one = 'Diabetes' OR speciality\_two = 'Diabetes' OR speciality\_three = 'Diabetes'**

**);**

****

**Query 5**

**SELECT AVG(doctor\_count) as average\_doctors\_per\_hospital**

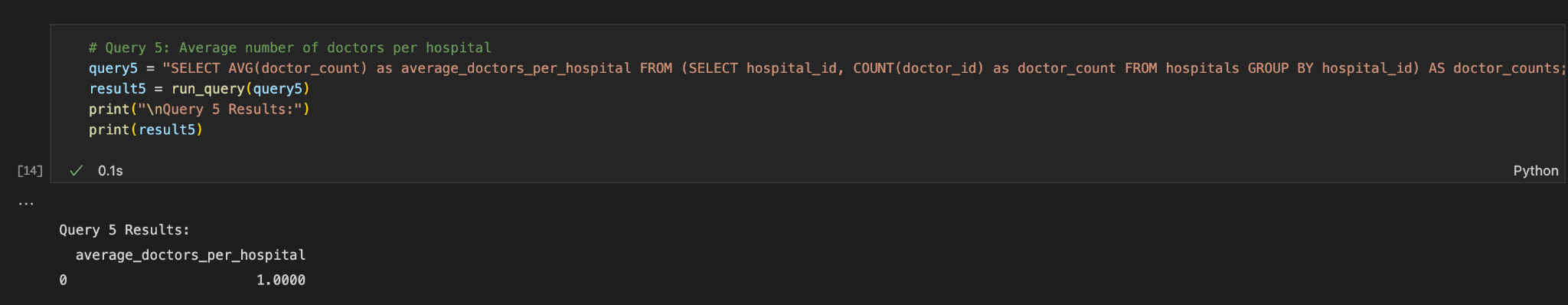
**FROM (**

**SELECT hospital\_id, COUNT(doctor\_id) as doctor\_count**

**FROM hospitals**

**GROUP BY hospital\_id**

**) AS doctor\_counts;**

****

**Visualization 1 : Distribution of Doctors by Specialty**

**# Executing the query to get data**

**query\_specialty\_distribution = """**

**SELECT speciality\_one, COUNT(\*) AS number\_of\_doctors**

**FROM specialties**

**GROUP BY speciality\_one**

**ORDER BY number\_of\_doctors DESC;**

**"""**

**specialty\_distribution = run\_query(query\_specialty\_distribution)**

**# Visualization with matplotlib**

**import matplotlib.pyplot as plt**

**plt.figure(figsize=(12, 8))**

**plt.barh(specialty\_distribution['speciality\_one'], specialty\_distribution['number\_of\_doctors'])**

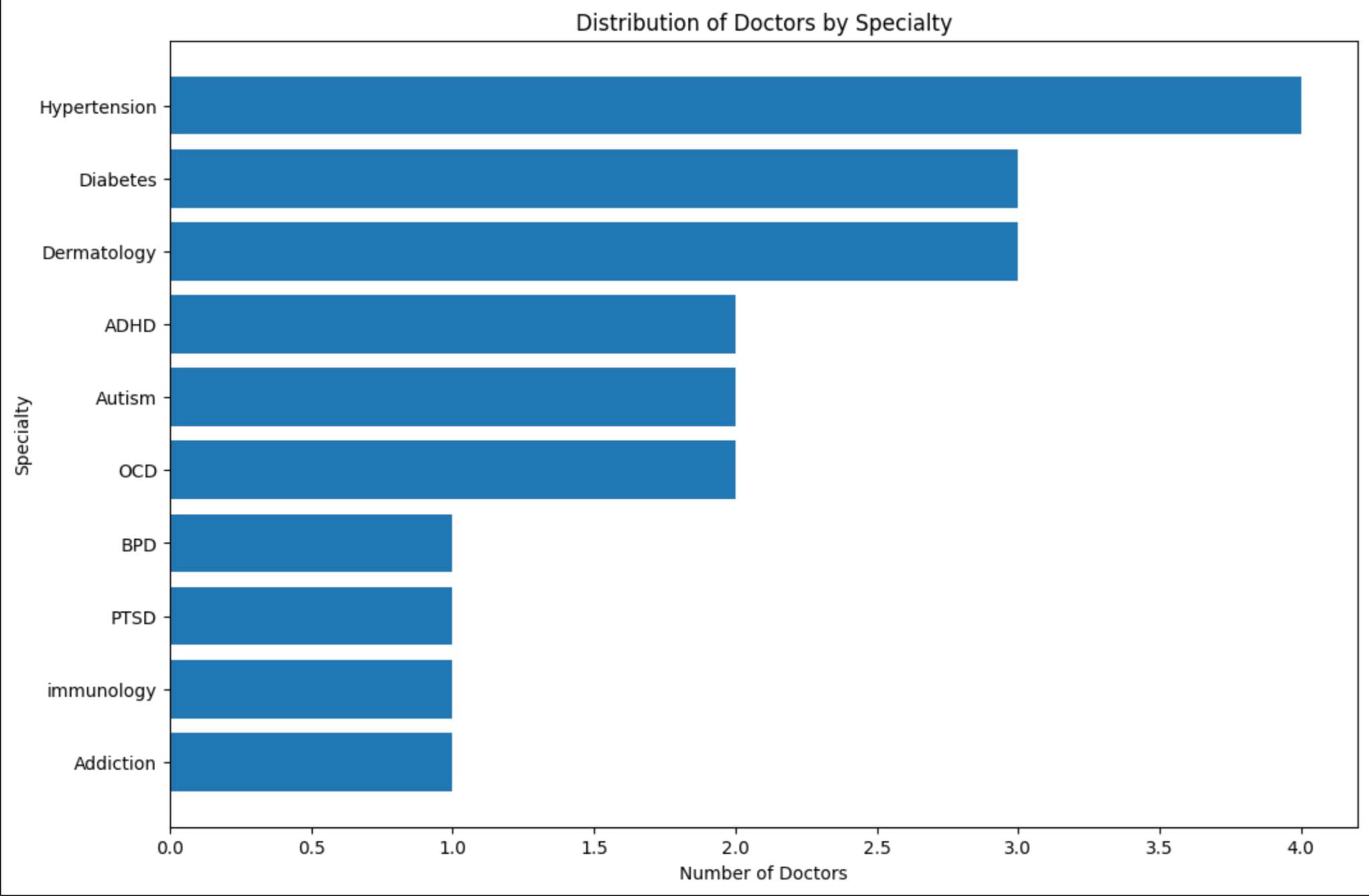
**plt.xlabel('Number of Doctors')**

**plt.ylabel('Specialty')**

**plt.title('Distribution of Doctors by Specialty')**

**plt.gca().invert\_yaxis() # To display the highest count at the top**

**plt.show()**

****

**Visualization 2 : Acceptance of New Patients by Gender**

**# Executing the query to get data**

**query\_new\_patients\_by\_gender = """**

**SELECT gender, new\_patients, COUNT(\*) AS number\_of\_doctors**

**FROM doctors**

**GROUP BY gender, new\_patients;**

**"""**

**new\_patients\_by\_gender = run\_query(query\_new\_patients\_by\_gender)**

**# Visualization with seaborn**

**import seaborn as sns**

**plt.figure(figsize=(10, 6))**

**sns.barplot(x='gender', y='number\_of\_doctors', hue='new\_patients', data=new\_patients\_by\_gender)**

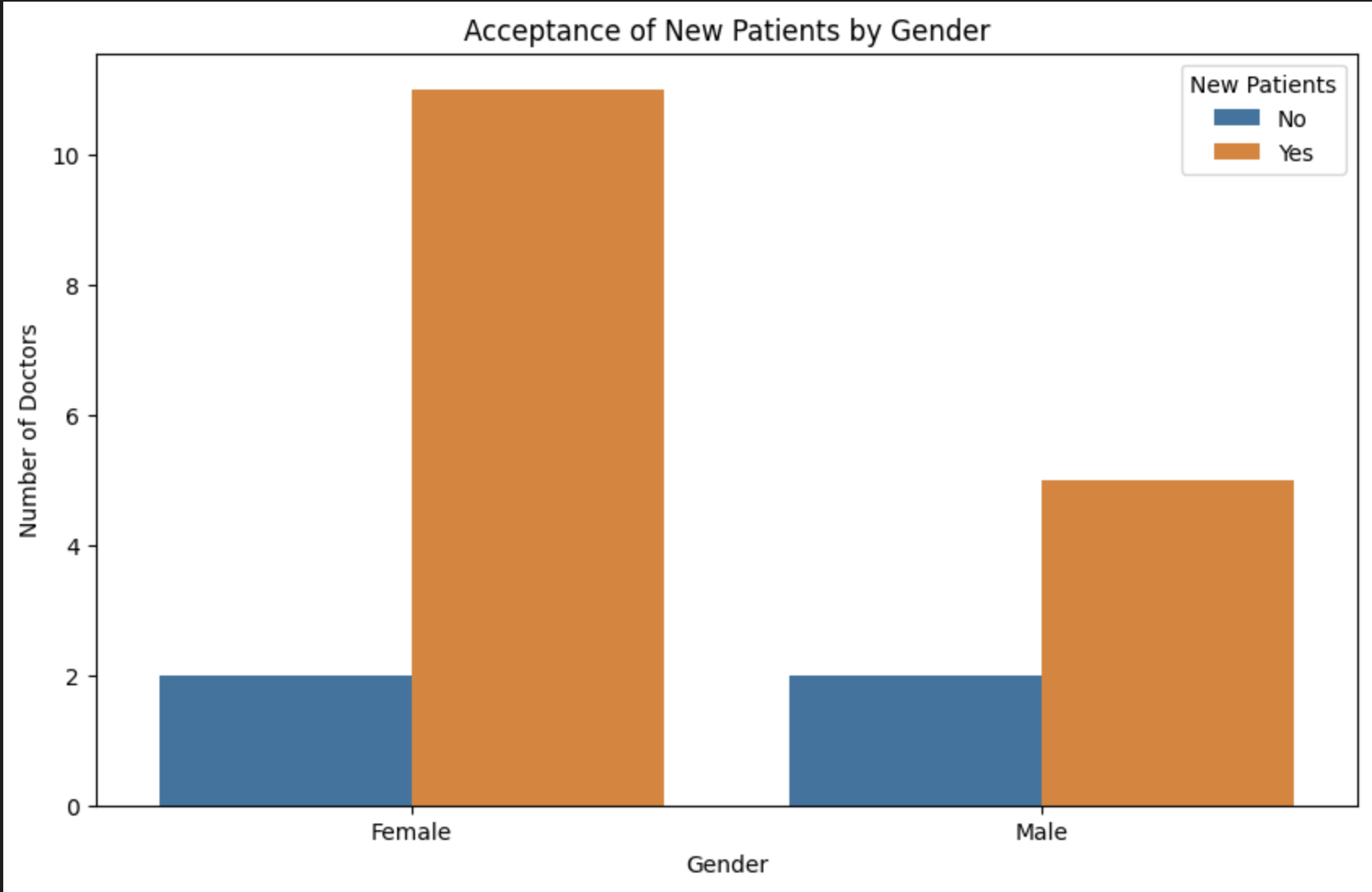
**plt.xlabel('Gender')**

**plt.ylabel('Number of Doctors')**

**plt.title('Acceptance of New Patients by Gender')**

**plt.legend(title='New Patients')**

**plt.show()**

****