Document Based Database

Part 1: MongoDB CRUD

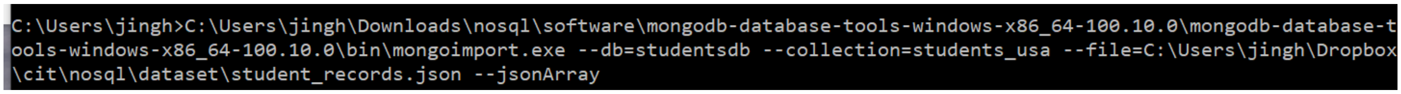
Download the bank.json file from Canvas. Complete the following tasks and screenshot each command and its output.

1. Create a new database “banksdb” and a new collection “clients”. Manually insert all documents in the json file into this collection. List out all collections in your database and documents to confirm that you have all documents in the correct collection. Then insert the following client to your collection:

   ```
   {
           "isActive": true,
           "balance": 2042.37,
           "age": 38,
           "eyeColor": "green",
           "name": "Margie Ayala",
           "gender": "female",
           "company": "VOIPA",
           "email": "margieayala@voipa.com",
           "phone": "+1 (941) 569-2231",
           "address": "111 Schroeders Avenue, Suitland, Louisiana, 7042"
    }
   ```
2. Explain why the profile of Margie has been inserted twice? How do we fix this problem? What happens if we try to reinsert the same profile of Margie again after you fixed the problem?
3. Update the final balance after applying one year’s 1% interest rate to the account of all active young female clients whose age below 30. [balance = principal * (1 + rate)]
4. All inactive middle aged male clients share some common hobbies “reading” and “video games” and they all love “Chinese”, “Korean”, and “French” food, keep this extra information in their profiles.
5. Some of inactive male clients who have green eyes have lost their passion on Korean food, update their profiles to reflect this change.

Part 2: MongoDB Queries and Aggregation (json file)

Download the student_records.json file from Canvas. Use mongoimport command to import the data into the students_usa collection of the studentsdb database with the following syntax:



```
C:\Users\jingh>C:\Users\jingh\Downloads\nosql\software\mongodb-database-tools-windows-x86_64-100.10.0\mongodb-database-tools-windows-x86_64-100.10.0\bin\mongoimport.exe --db=studentsdb --collection=students_usa --file=C:\Users\jingh\Dropbox\cit\nosql\dataset\student_records.json --jsonArray
```

Complete the following tasks and screenshot each command and its output.

Region	City

Northeastern	Boston, New York
Westsouth	San Antonio, New Orleans
Southwestern	Austin, San Diego, Los Angeles, Palm Springs, Lahaina, Sedona, San Francisco, Nashville, Las Vegas, Honolulu
Midwestern	St. Louis, Branson, Chicago
Southernmost	Miami Beach
Pacific Northwest	Portland, Seattle, Washington
Southeastern	Charleston, Savannah
Northernmost	Saint Augustine
Southern	Orlando

City	Class ID	Student ID
Austin	2	0, 6, 60
	13	25, 67
	23	38, 191, 4, 8
Boston	0	57, 127, 43, 126, 144
	16	117, 160, 53, 132
	25	70, 122, 1, 187, 73, 103
San Diego	5	2, 142
	10	120
	21	101
	30	59
St. Louis	16	3, 91
	30	5
San Antonio	6	7, 163, 78
	10	196, 27
	28	13, 116, 195
New Orleans	11	9, 112, 10, 93, 66, 37
	25	16, 121, 140, 133, 69
	3	75, 173, 147, 20
Portland	18	11, 179, 33,
	22	45, 46
	30	123
Charleston	22	12, 124, 21
	11	85
	30	115
	8	130, 153
Los Angeles	16	14, 146, 169, 72
	29	156, 183, 90, 152
Chicago	13	15, 50

	20	81, 18, 47, 109
	1	118, 188, 157
Palm Springs	27	17, 199, 44
	7	26, 82, 186, 83, 65
	17	49, 189
New York City	10	19, 125
	25	151
	0	34, 181, 114
	30	36
Miami Beach	12	22, 23, 137, 54
	25	167, 29
	30	98
	4	52, 192, 143
Saint Augustine	3	24, 131, 150, 174
	17	84, 148, 62
	20	56, 92
Honolulu	5	28, 63, 154, 39
	20	128, 89
	18	129
San Francisco	16	64, 145
	6	176, 134
Seattle	24	30, 185, 102, 155, 94, 164, 197
	1	178, 87, 48, 96
	17	86, 58
Savannah	8	31, 158
	17	135
	25	111, 71
Las Vegas	19	32, 166
	25	194
	30	193
	5	107, 40
Orlando	11	76, 80, 119, 74, 35, 104
	21	42, 161, 139
	7	172, 198, 68, 61, 180
Nashville	16	41, 175, 171, 141, 97
	22	182, 177
	4	170, 88
Lahaina	29	51, 77
	12	190
	2	168, 162, 55
Sedona	22	79, 136
	12	100, 149
Branson	15	95

	4	99, 159, 113, 105
Washington	3	106
	16	184, 165
	22	108, 138, 110

1. List out all cities located in the southwest region of United States.
2. Find all adolescents aged under 20, lived in the Southern area of the United States, and have only one hobby. Evaluate the performance of this query, then add an index to improve the performance of this query.
3. How many female students passed all three assessment components in each region? The pass grade is 40 percent.
4. How many students have passed all three assessment components in each class of each region?
5. What is the most popular hobby, and which class do students have it?

Part 3 MongoDB Queries and Aggregation (CSV file)

This dataset has been carefully curated to support research in stroke risk prediction, helping develop models that estimate:

1. Whether a person is at risk of a stroke (Binary Classification).
2. The percentage likelihood of stroke occurrence (Regression Analysis).

It is designed for machine learning and deep learning applications in medical AI and predictive healthcare. The dataset is balanced, ensuring that 50% of the records belong to individuals at risk and 50% belong to those not at risk.

The dataset was constructed based on medical literature, expert consultations, and statistical modeling. The feature distributions and relationships were inspired by real-world clinical observations, ensuring medical validity.

The dataset structure is based on established risk factors documented in leading medical textbooks, research papers, and guidelines from health organizations. Key references include:

- **American Stroke Association (ASA):** Guidelines on stroke risk factors and early warning symptoms.
- **Mayo Clinic & Cleveland Clinic**: Medical literature on cardiovascular diseases and stroke risk factors.
- "**Harrison's Principles of Internal Medicine**" (20th Edition): Provides in-depth insights into stroke etiology and risk factors.
- "**Stroke Prevention, Treatment, and Rehabilitation**" (2021, Oxford University Press): A comprehensive guide on stroke mechanisms and preventive strategies.
- "**The Stroke Book" (Cambridge Medicine, 2nd Edition)**: Clinical insights into the symptoms and early predictors of stroke.

- World Health Organization (WHO) Reports on Stroke Risk and Prevention.

Each record represents an individual's medical condition, symptoms, and risk assessment. The dataset includes the following features:

1 **Symptoms (Primary Predictors)**
The presence of these symptoms significantly influences stroke risk. These features are binary (1 = symptom present, 0 = absent).

- Chest Pain
- Shortness of Breath
- Irregular Heartbeat
- Fatigue & Weakness
- Dizziness
- Swelling (Edema)
- Pain in Neck/Jaw/Shoulder/Back
- Excessive Sweating
- Persistent Cough
- Nausea/Vomiting
- High Blood Pressure
- Chest Discomfort (Activity)
- Cold Hands/Feet
- Snoring/Sleep Apnea
- Anxiety/Feeling of Doom

2 **Target Variables (Predicted Outcomes)**

- At Risk (Binary) → 1 if the person is at risk of stroke, 0 otherwise.
- Stroke Risk (%) → The estimated probability of stroke occurrence, ranging from 0 to 100.

3 Demographic Feature

- Age → A key risk factor, as stroke prevalence increases with age.

⚡Why This Dataset is Accurate and Useful?

1. Balanced Data Distribution:

- 50% of the data represents individuals at risk of stroke.
- 50% represents those who are not at risk.

- Ensures no model bias towards a specific class.
- Medically Inspired Feature Engineering:

1. Features are derived from real-world stroke risk factors, validated through medical literature.

- Age is incorporated as a major determinant of risk.
- Symptom severity is considered through a weighted scoring approach.
- Diverse Risk Factors Considered:

1. Cardiovascular symptoms like chest pain, irregular heartbeat, high blood pressure.

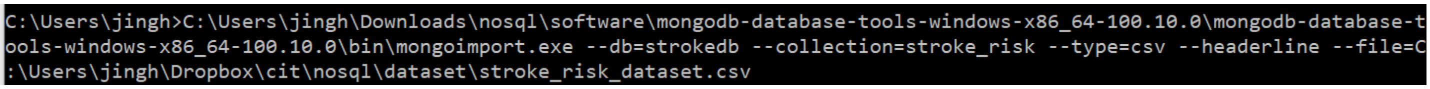
- Neurological symptoms such as dizziness, fatigue, and anxiety.
- Sleep-related issues like snoring and sleep apnea, which are linked to increased stroke risk.

Column Name	**Description**
Chest Pain	Binary (0/1): Indicates whether the individual experiences chest pain, a common symptom of cardiovascular conditions.
Shortness of Breath	Binary (0/1): Represents whether the person has difficulty breathing, which may indicate heart or lung problems.
Irregular Heartbeat	Binary (0/1): Shows if the person has an irregular heartbeat, a potential stroke risk factor.
Fatigue & Weakness	Binary (0/1): Indicates persistent fatigue and muscle weakness, common signs of cardiovascular issues.
Dizziness	Binary (0/1): Reports whether the individual frequently experiences dizziness, which may be linked to poor circulation.

Column Name	Description
Swelling (Edema)	Binary (0/1): Indicates swelling in extremities due to fluid retention, a potential cardiovascular issue.
Pain in Neck/Jaw/Shoulder/Back	Binary (0/1): Describes pain in these areas, which can be a warning sign of stroke or heart attack.
Excessive Sweating	Binary (0/1): Shows whether the individual experiences unusual sweating, which may indicate cardiovascular distress.
Persistent Cough	Binary (0/1): Indicates chronic coughing, which can be associated with heart failure.
Nausea/Vomiting	Binary (0/1): Reports frequent nausea or vomiting, which may be linked to cardiovascular events.
High Blood Pressure	Binary (0/1): Represents whether the person has high blood pressure, a major risk factor for stroke.
Chest Discomfort (Activity)	Binary (0/1): Shows if the individual experiences chest discomfort during physical activity.
Cold Hands/Feet	Binary (0/1): Indicates whether the person often has cold extremities, a possible sign of circulation problems.

Column Name	Description
Snoring/Sleep Apnea	Binary (0/1): Reports whether the individual has sleep apnea, which can increase stroke risk.
Anxiety/Feeling of Doom	Binary (0/1): Captures whether the person experiences frequent anxiety or a sense of impending doom, which can be related to cardiovascular distress.
Stroke Risk (%)	Continuous (0-100%): The estimated percentage risk of having a stroke, based on symptom severity and medical indicators.
At Risk (Binary)	Binary (0/1): Indicates whether the person is classified as at risk of stroke (1) or not (0).
Age	Integer: The age of the individual, an essential factor in assessing stroke risk.

Download the stroke_risk_dataset file from Canvas. Use mongoimport command to import the data into the stroke_risk collection of the strokedb database with the following syntax:



```
C:\Users\jingh>C:\Users\jingh\Downloads\nosql\software\mongodb-database-tools-windows-x86_64-100.10.0\mongodb-database-t
ools-windows-x86_64-100.10.0\bin\mongoimport.exe --db=strokedb --collection=stroke_risk --type=csv --headerline --file=C
:\Users\jingh\Dropbox\cit\nosql\dataset\stroke_risk_dataset.csv
```

Complete the following tasks and screenshot each command and its output.

1. There are three groups of patients - patients who have cardiovascular symptom, patients who have neurological symptoms, patients who have sleep-related issues. Create a view for each group of patients. The main cardiovascular symptoms are chest pain, irregular heartbeat, swelling, excessive sweating, nausea/vomiting, and high-blood pressure. The main neurological symptoms are fatigue and weakness, dizziness, persistent cough, anxiety, pain in various parts of the body.

2. Dr. James Smith is a cardiologist, he wants to have an account to directly access the information of patients who suffer from cardiovascular symptoms. He wants to know how many young patients are under 30 years old, who have suffered from chest pain, irregular heartbeats, high blood pressure, and have over 40% chance of getting a stroke.
3. Switch back to the normal user account, how many pensioners aged between 60 and 70 do not have a risk to get a stroke even though they constantly feel dizzy, fatigue, and aches everywhere.
4. Stay as a normal user account, how many patients aged between 40 and 50 have over 45% chance of not getting a stroke when they snore heavily.
5. Stay as a normal user account, find the chances of getting a stroke and the age of a patient when s/he participates some activities without any anxieties, s/he constant feels discomfort in his/her chest, hands/feet are cold, and sometimes short of breath.