Elixir Project Report

Project Description:

This project mainly provides details of all the Diseases, Drugs, Hospitals and the Diagnosis details of the patient. It includes data gathering from various data sources, which includes data repositories from various sites, web scraping using BeautifulSoup and twitter scraping utilizing Twitter API. The gathered data is preprocessed according to the database structure and schema. A detailed audit of the data sources is made using Pandas Profiling. Preprocessed the database using python scripts and normalization is done for all the tables.

GitHub: https://github.com/nikhil-enni/Project_Elixir.git

Audit:

Please find the visualized audit report of the data sources gathered in the python notebook

1. **Diagnosis Data :**
    a. Relevancy: *The data should meet the requirements for the intended use.*

    b. Comments: *Data created by the Elixir team to manage the functioning of Hospitals in a professional and optimal manner.*

    c. Completeness: *The data should not have missing values or miss data records.*

    d. Comments:

Bed Grade	*113*
City_Code_Patient	*4532*

    Above are empty value counts for bed grade and city_code _patient.
    city_code_patient: Filling with "null" value
    Bed Grade: Filling with mean value of bed grade

    e. Consistency: *The data should have the data format as expected and can be cross reference-able with the same results.*

    f. Comments:
    Age and Stay have interval values
    Age - Mean of the interval taken
    Stay - Mean of the interval taken

2. **Hospitals**
    a. Relevancy: *The data should meet the requirements for the intended use.*

    b. Comments: *This dataset is provided by the Homeland Infrastructure Foundation-Level Data (HIFLD) without a license and for Public Use.*

    c. Completeness: *The data should not have missing values or miss data records.*

    d. Comments:
    TTL_STAFF: Have constant value -999.
    Diagnosis - *We have dropped the column as there is no information in TTL_STAFF*

    WEBSITE: 377 ARE NOT AVAILABLE.
    Diagnosis: *Mode Imputation will lead to wrong websites to hospitals. So filled null values with empty string.*

e. Consistency: *The data should have the data format as expected and can be cross reference-able with the same results.*

f. Comments: *Data format is as needed to directly join with diagnosis data. No preprocessing is done.*

3. **Drugs:**

a. Relevancy: *The data should meet the requirements for the intended use.*

b. Comments: *This dataset is created from web scraping of various health websites*

c. Completeness: *The data should not have missing values or miss data records.*

d. Comments: *There are no missing values in diseases-drugs data.*

e. Consistency: *The data should have the data format as expected and can be cross reference-able with the same results.*

f. Comments: *Data format is as needed to directly join with diagnosis data.No preprocessing is done.*

Elixir Twitter Bot:

Description:
This twitter bot which is a python script, scrapes the data utilizing Twitter API. The bot essentially extracts the Tweets related to Diseases, Drugs and Hospitals, this also stores the user information like user name, user handle, twitter joining date, followers count etc. associated with the Tweets that were extracted. Additional feature of this bot is to store the tags used by the users to Tweet in Twitter. One important functionality is to add a sentiment score to all the tweets, by this it would be possible to know the user's opinion on the medicines and hospitals. Extraction is done on the basis of the details stored in the Diseases, Drugs and Hospitals table. All the extracted details are committed in the Database used using a python script. Below are the details that are being scrapped from Twitter.

Below are the data fields that were scrapped using Twitter API
Tweet Details:

1. Tweet ID
2. Twitter Handle
3. Tweet
4. Tweet creation date
5. Retweet count
6. Likes count

User Details:

1. User ID
2. User Name
3. User Handle
4. User Profile Picture link
5. Followers Count

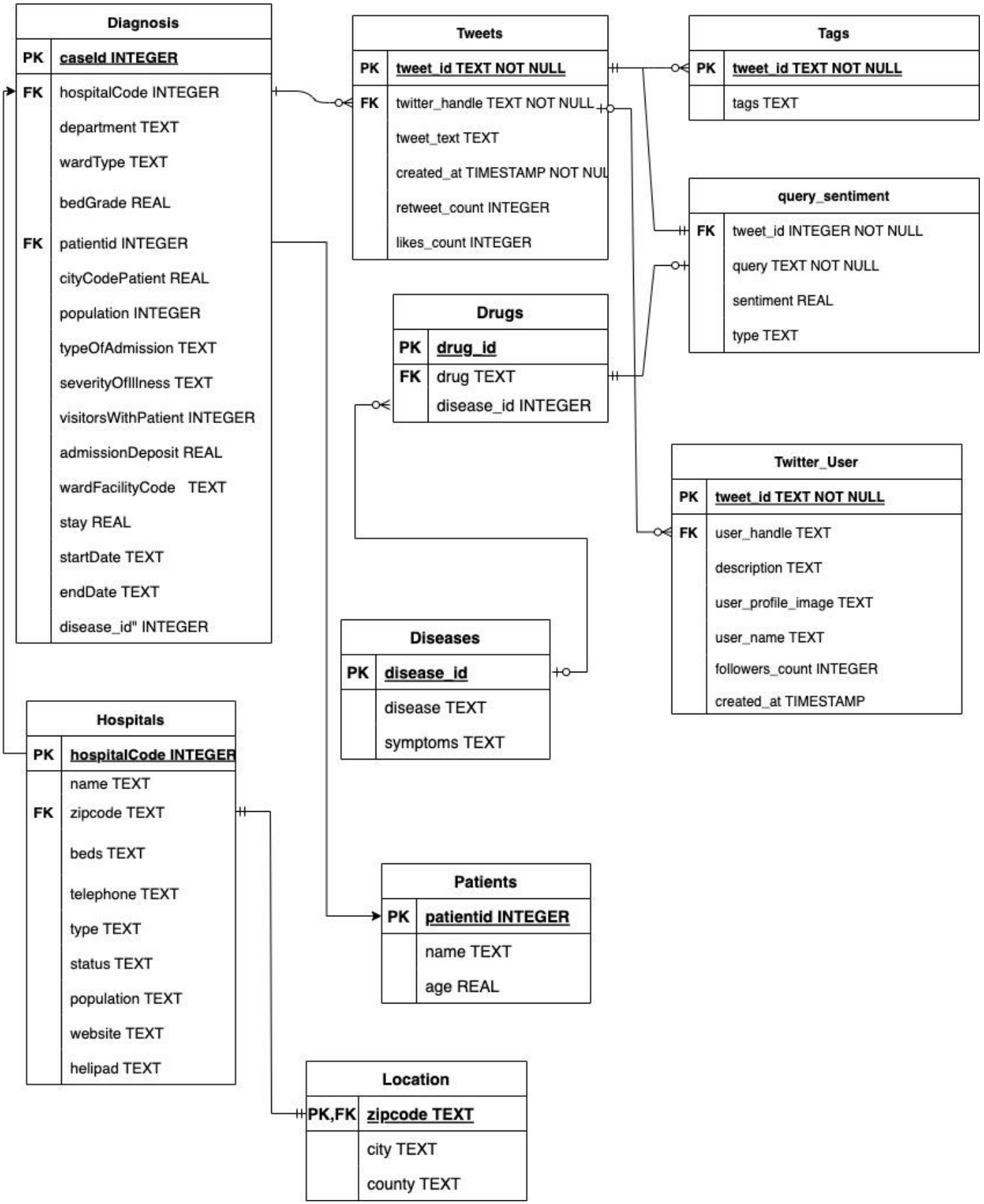
Tag Details:

1. Tag Name

Sentiment score:
Sentiment score of all the tweets are calculated

1. score>0 implies positive opinion
2. score<0 implies negative opinion
3. score=0 implies neutral opinion

ER Diagram:



Steps followed:

1. Created problem statements by considering a real world problem, that is hospital management and drug distribution which involves Hospitals, Drugs, Diseases, Patients, Diagnosis and public opinion from twitter.
2. Scrapped the twitter data using Twitter API and collected data like tweets, usernames, user handles, twitter tags, likes and followers count.

3. Gathered data repositories from various medical data sources to use this while scraping and also to construct the database.
4. And also gathered some data fields using Beautiful Soup
5. A detailed audit is done utilizing Pandas Profiling to get a clear understanding of the data and to clean the database.
6. Preprocessed the data frames to fit into the desired database structure and schema.
    a. Removed all the empty values and replaced with nulls
    b. Changed the intervals into mean values, to make it convenient for the constructed queries
    c. Removed all the unwanted data fields which don't add value to the project in the interest of storage.
7. Connection to a database is made and all the data is loaded to the SQL tables.
8. Normalized the end database
    a. Maintained unique values and atomicity. Every table has a primary key.
    b. No partial dependency of the data fields
    c. No transitive dependency of the data fields
9. Provided result set for all the use cases using SQL queries.

Code:

Please refer to the below link for the well commented python scripts,

Scripts for:

1. Twitter Bot
2. Audit report
3. Preprocessing of the data

GitHub: https://github.com/nikhil-enni/Project_Elixir.git

SQL:

Create and Insert Statements:

Create Statements:

```
CREATE TABLE "Diagnosis" (
        "caseId"        INTEGER,
        "hospitalCode"INTEGER,
        "department"   TEXT,
        "wardType"     TEXT,
        "wardFacilityCode"     TEXT,
        "bedGrade"     REAL,
        "patientid"       INTEGER,
        "cityCodePatient"       REAL,
        "typeOfAdmission"      TEXT,
        "severityOfIllness"      TEXT,
        "visitorsWithPatient"   INTEGER,
        "admissionDeposit"    REAL,
        "stay"   REAL,
        "startDate"      TEXT,
        "endDate"       TEXT,
        "disease_id"    INTEGER
);

CREATE TABLE "Diseases" (
        "disease"        TEXT,
```

```
        "symptoms"    TEXT,
        "disease_id"    INTEGER
);

CREATE TABLE "Drugs" (
        "drug"  TEXT,
        "drug_id"       INTEGER,
        "disease_id"    INTEGER
);

CREATE TABLE "Hospitals" (
        "hospitalCode"INTEGER,
        "name"TEXT,
        "city"    TEXT,
        "zipcode"       TEXT,
        "beds"  TEXT,
        "telephone"     TEXT,
        "type"  TEXT,
        "status"        TEXT,
        "population"    INTEGER,
        "county"        TEXT,
        "website"       TEXT,
        "helipad"       TEXT
);

CREATE TABLE "Patients" (
        "patientid"     INTEGER,
        "name"TEXT,
        "age"   REAL
);

CREATE TABLE "Tags" (
        "tweet_id"      TEXT,
        "tags"  TEXT
);

CREATE TABLE "Tweets" (
        "tweet_id"      TEXT,
        "twitter_handle"        TEXT,
        "tweet_text"    TEXT,
        "created_at"    TIMESTAMP,
        "retweet_count"         INTEGER,
        "likes_count"   INTEGER
);

CREATE TABLE "Twitter_User" (
        "user_id"       TEXT,
        "user_handle" TEXT,
        "user_name"   TEXT,
        "user_profile_image"  TEXT,
        "description"   TEXT,
        "followers_count"       INTEGER,
        "created_at"    TIMESTAMP
);
```

```
CREATE TABLE "hospitals_sentiment_by_disease" (
	"Tweet Text"	TEXT,
	"sentiment"	REAL,
	"hospital"	TEXT,
	"disease"	TEXT
);

CREATE TABLE "medicines_sentiment_by_disease" (
	"Tweet Text"	TEXT,
	"sentiment"	REAL,
	"medicine"	TEXT,
	"disease"	TEXT
);

CREATE TABLE "query_sentiment" (
	"tweet_id"	TEXT,
	"query"TEXT,
	"sentiment"	REAL,
	"type"	TEXT
);
```

Insert Statements:

Diagnosis table:

```
INSERT INTO "main"."Diagnosis"
("caseId","hospitalCode","department","wardType","wardFacilityCode","bedGrade","patientid","cityCodePatient
","typeOfAdmission","severityOfIllness","visitorsWithPatient","admissionDeposit","stay","startDate","endDate","
disease_id") VALUES
("1","8","radiotherapy","R","F","2.0","31397","7","Emergency","Extreme","2","4911","5","2022-03-05","2022-03-0
8","31");
```

Diseases table:

```
INSERT INTO "main"."Diseases"("disease","symptoms","disease_id") VALUES ('vulvodynia','pelvic pain,sharp
abdominal pain,lower abdominal pain','201');
```

Drugs table:

```
INSERT INTO "main"."Drugs"("drug","drug_id","disease_id") VALUES ("clonazepam","145","134");
```

Hospitals table:

```
INSERT INTO
"main"."Hospitals"("hospitalCode","name","city","zipcode","beds","telephone","type","status","population","coun
ty","website","helipad") VALUES (“0”,”east jefferson general hospital”,”metairie”,”70006”,”420.0”,”(504)
454-4000”,”GENERAL ACUTE CARE”,”OPEN”,”420”,”JEFFERSON”,”http://www.ejgh.org”,”Y”);
```

Patients table:

```
INSERT INTO "main"."Patients"("patientid","name","age") VALUES ("1","Jack","42");
```

Tags table:

INSERT INTO "main"."Tags"("tweet_id","tags") VALUES ("1597781304250028032","weightloss");

Tweets table:

INSERT INTO "main"."Tweets"("tweet_id","twitter_handle","tweet_text","created_at","retweet_count","likes_count") VALUES ("1596686265692590080","bettylo52207153","@Stickit2Stage4 Thank you so much Susan. I will let her know and its good to hear it's not unusual. I never experi… https://t.co/WJRFWOv4SG","2022-11-27 02:04:05+00:00","0","1");

Twitter_User table:

INSERT INTO "main"."Twitter_User"("user_id","user_handle","user_name","user_profile_image","description","followers_count","created_at") VALUES ("1359714993357393923","bettylo52207153","bettylou,http://abs.twimg.com/sticky/default_profile_images/default_profile_normal.png","Ocean State, D/x TNBC breast cancer", "BRCA1+
Kind always!"
,"56","2021-02-11 04:05:03+00:00");

Hospitals_sentiment_by_disease table:

INSERT INTO "main"."hospitals_sentiment_by_disease"("Tweet Text","sentiment","hospital","disease") VALUES ("Please reduce the price of perjeta as perjeta is life saving drug. Now Zydus and Intas in race of launching the par… https://t.co/SKTiQZ9Cey","0.0","Massachusetts General Hospital","Alzheimers disease");

Medicines_sentiment_by_disease table:

INSERT INTO "main"."medicines_sentiment_by_disease"("Tweet Text","sentiment","medicine","disease") VALUES ("Please reduce the price of perjeta as perjeta is life saving drug. Now Zydus and Intas in race of launching the par… https://t.co/SKTiQZ9Cey","0.0","Perjeta","Cancer");

Query_sentiment table:

INSERT INTO "main"."query_sentiment"("tweet_id","query","sentiment","type") VALUES ("1596686265692590080","Perjeta","0.266666666666667","medicine");

Use Cases:

Below are the SQL queries for the real world problems/questions

1. Use Case: View the most popular drug and its associated disease based on the twitter users's opinion (utilizing sentimental analysis)
   Actor: User
   Step:
   Actor action: User executes the below query against the database
   System Responses: Drug Name and its associated diseases.

   SELECT Drugs.DiseaseId, DiseaseName, DrugName from Drugs join Diseases on Diseases.DiseaseId= Drugs.DiseaseId where DrugName in (select DISTINCT query from

query_sentiment where sentiment = (SELECT max(sentiment) from query_sentiment where type='medicine'));

2. Use Case: View the best hospital for given disease
   Actor: User
   Step:
   Actor Action: Users tweets about the best hospital
   System Responses: Best hospital for given disease is displayed
   Alternate Path: No Hospital is displayed.

   SELECT max(avg_score), query FROM (SELECT avg(sentiment) AS avg_score, query FROM query_sentiment WHERE type='hospital' GROUP BY query HAVING query like '%Cancer');

```
SELECT max(avg_score), query FROM (SELECT avg(sentiment) AS avg_score, query FROM query_sentiment WHERE type='hospital' GROUP BY query HAVING query like '%Cancer');
```

	max(avg_score)	query
1	0.5	Cape Cod Hospital AND Cancer

3. Use Case:View the city that needs more health care attention
   Actor: User
   Step:
   Actor Action: Users tweets about the cases
   System Responses: Best hospital for given disease is displayed

   SELECT max(cases_count) as cases_count, cityCodePatient
   from (SELECT count(caseId) as cases_count, cityCodePatient from Diagnosis group by cityCodePatient);

```
SELECT max(cases_count) as cases_count, cityCodePatient
from (SELECT count(caseId) as cases_count, cityCodePatient from Diagnosis group by cityCodePatient);
```

	cases_count	cityCodePatient
1	124011	8.0

4. Use Case: Patients who referred to multiple hospitals for a given city
   Actor: User
   Step:
   Actor Action: Users tweets about the cases
   System Responses: Best hospital for given disease is displayed

   SELECT count(DISTINCT Diagnosis.hospitalCode) as hospital_count, Diagnosis.patientid, Patients.name, Hospitals.city FROM Diagnosis JOIN Patients on Patients.patientid=Diagnosis.patientid JOIN Hospitals on Hospitals.hospitalCode=Diagnosis.hospitalCode
   GROUP by Diagnosis.patientid HAVING hospital_count>3 AND Hospitals.city='tillamook';

```
SELECT count(DISTINCT Diagnosis.hospitalCode) as hospital_count, Diagnosis.patientid, Patients.name, Hospitals.city FROM Diagnosis JOIN Patients on Patients.patientid=Diagnosis.patientid
JOIN Hospitals on Hospitals.hospitalCode=Diagnosis.hospitalCode
GROUP by Diagnosis.patientid HAVING hospital_count>3 AND Hospitals.city='tillamook';
```

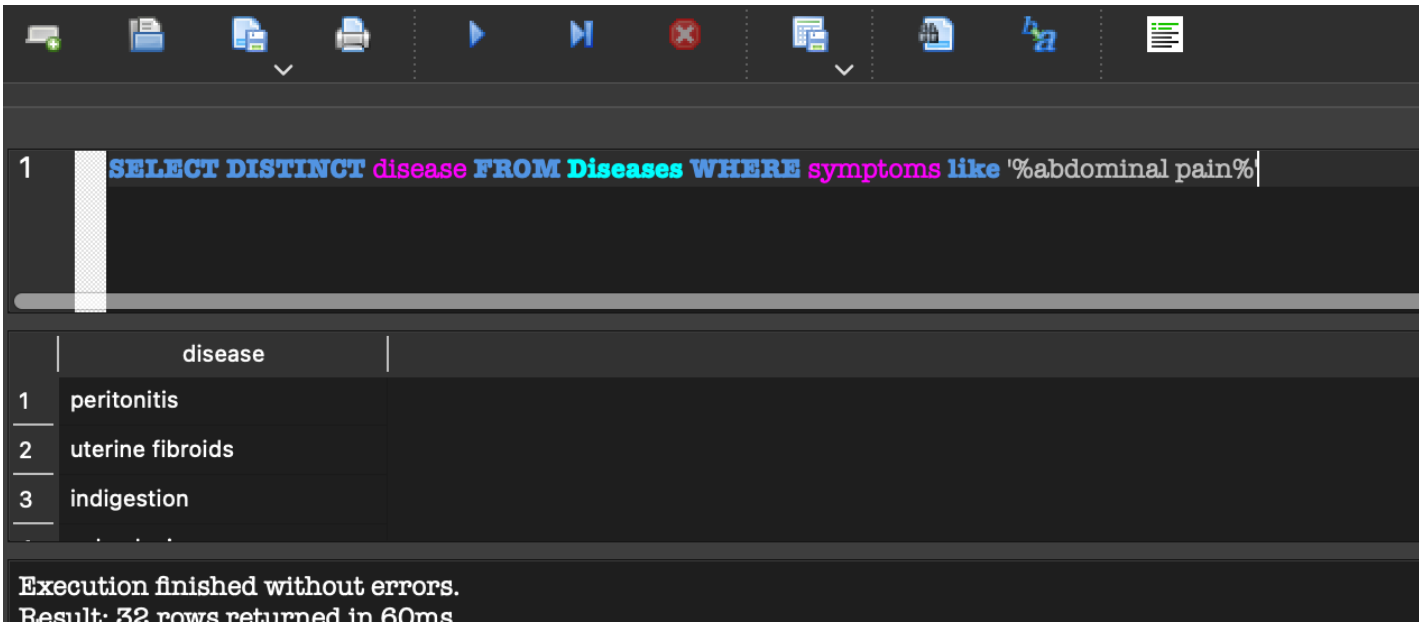
	hospital_count	patientid	name	city
1	8	33	johnny thompson	tillamook
2	4	43	nathan hughes	tillamook
3	5	92	sarah singh	tillamook

Execution finished without errors.
Result: 1173 rows returned in 829ms

5. Use Case: Prediction of disease based on a particular symptom

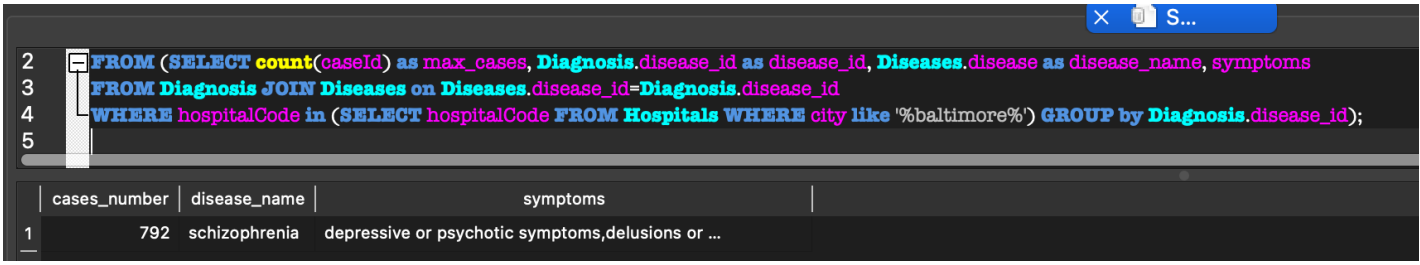
Action: Disease
Step:
Actor Action: predict disease based on symptoms
System Responses:Disease is displayed along with the symptoms

SELECT DISTINCT disease FROM Diseases WHERE symptoms like '%abdominal pain%';



6. Use Case:Most prevalent disease in the city and symptoms
   Action: Disease
   Step:
   Actor Action:
   System Responses:Disease is displayed along with the symptoms

   SELECT max(max_cases) as cases_number, disease_name, symptoms
   FROM (SELECT count(caseId) as max_cases, Diagnosis.disease_id as disease_id, Diseases.disease as disease_name, symptoms
   FROM Diagnosis JOIN Diseases on Diseases.disease_id=Diagnosis.disease_id
   WHERE hospitalCode in (SELECT hospitalCode FROM Hospitals WHERE city like '%baltimore%')
   GROUP by Diagnosis.disease_id);



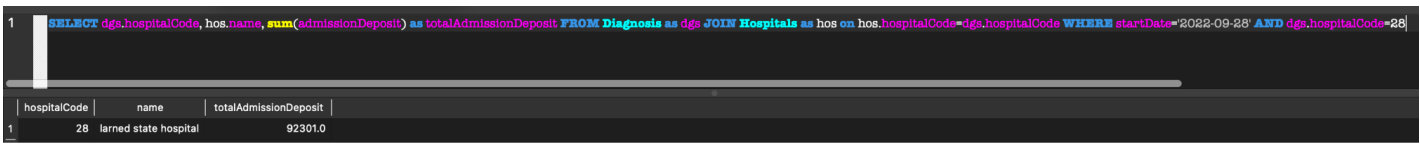
7. Use Case:Identify the list of all the patients less than a particular age affected by a particular disease in extreme severity and urgent admission.
   Action: Disease
   Step:
   Actor Action:
   System Responses:Disease is displayed along with the symptoms

   SELECT DISTINCT Patients.patientid, Patients.name FROM Diagnosis as dgs JOIN Diseases as dis on dis.disease_id=dgs.disease_id

JOIN Patients on Patients.patientid=dgs.patientid WHERE dis.disease like '%bipolar disorder%' AND severityOfIllness='Extreme' AND typeOfAdmission='Urgent' AND age<30;

8. Use Case:Total revenue generated by a particular hospital with admission deposits on a particular date.
   Actor: User
   Actor Action: Total revenue generated by a particular hospital
   System Responses: To display revenue

   SELECT dgs.hospitalCode, hos.name, sum(admissionDeposit) as totalAdmissionDeposit FROM Diagnosis as dgs JOIN Hospitals as hos on hos.hospitalCode=dgs.hospitalCode WHERE startDate='2022-09-28' AND dgs.hospitalCode=28;

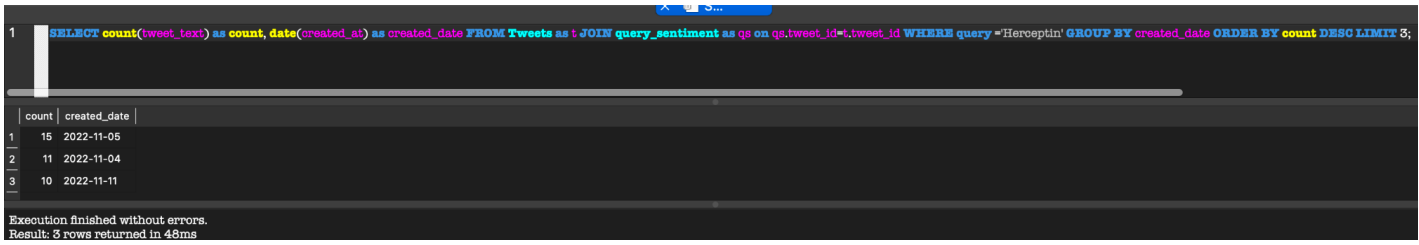


9. Use Case:List of patients whose length of stay is maximum in a given hospital along with age and name of the patient.
   Actor: User
   Actor Action: Total revenue generated by a particular hospital
   System Responses: To display revenue

   SELECT DISTINCT disease, Patients.name, age FROM Diagnosis JOIN Diseases on Diseases.disease_id=Diagnosis.disease_id
   JOIN Patients on Patients.patientid=Diagnosis.patientid WHERE stay=120 AND hospitalCode=5;

10. Use Case: Top 3 dates when the users got the attention of a particular drug/hospital
    Actor: User
    Step:
    Actor Action: Top 3 dates
    System Responses: To display top 3 dates when users got attention of a particular drug.

    SELECT count(tweet_text) as count, date(created_at) as created_date FROM Tweets as t JOIN query_sentiment as qs on qs.tweet_id=t.tweet_id WHERE query ='Herceptin' GROUP BY created_date ORDER BY count DESC LIMIT 3;



11. Use Case:List of all the users who used a particular hashtag
    Actor: User
    Step:
    Actor Action: Used particular tags
    System Responses: To display the users who used a particular hashtag.

    select user_name from Twitter_User where user_handle in (select twitter_handle from Tags join Tweets on Tags.tweet_id=Tweets.tweet_id where tags is not null and tags='EAPCI');

```
select user_name from Twitter_User where user_handle in (select twitter_handle from Tags join Tweets on Tags.tweet_id=Tweets.tweet_id where tags is not null and tags='EAPCI');
```

	user_name
1	Marta Kaluzna-Oleksy

Relation Algebra:
π user_name
 σ user_handle = "π twitter_handle
 σ NOT (tags = NULL)AND tags = "EAPCI" (tags ⋈ tags . tweet_id = tweets . tweet_id tweets)"
twitter_user

12. Use Case:Which disease requires attention
    Step:
    System Responses: To display the disease

    SELECT DiseaseName from Diseases where DiseaseId in (SELECT DiseaseId from Drugs where DrugName in (select query from query_sentiment where sentiment=(select min(sentiment) from query_sentiment where type='medicine')));

13. Use Case:What tags are being promoted by a particular user
    Actor: User
    Step:
    Actor Action: To promote tags
    System Responses: To display the tags that are been promoted

    select tags from Tags as tg join Tweets as t on tg.tweet_id=t.tweet_id join Twitter_User as tu on tu. user_handle=t.twitter_handle where user_name='Dr. Thomas Ichim';

```
select tags from Tags as tg join Tweets as t on tg.tweet_id=t.tweet_id join Twitter_User as tu on tu. user_handle=t.twitter_handle where user_name='Dr. Thomas Ichim';
```

	tags
1	exosomes,herceptin

Relation Algebra:
π tags
 σ user_name = "Dr. Thomas Ichim"
  (ρ tg tags ⋈ tg . tweet_id = t . tweet_id
   ρ t tweets ⋈ tu . user_handle = t . twitter_handle
    ρ tu twitter_user)

14. Use Case:Retrieve all the tweets that Doctors have posted about a particular medicine/drug to get accurate information about a drug.
    Actor: Doctors
    Step:
    Actor Action: To post about a particular drug
    System Responses: To show the accurate information about the drug

    SELECT tweet_text, query, user_name from Twitter_User as tu join Tweets as t on tu.user_handle=t.twitter_handle join query_sentiment as qs on qs. tweet_id=t.tweet_id where user_name like 'Dr%' and type='medicine' and query='Albuterol';

```
SELECT tweet_text, query, user_name from Twitter_User as tu join Tweets as t on tu.user_handle=t.twitter_handle join query_sentiment as qs on qs. tweet_id=t.tweet_id where user_name like 'Dr%' and type='medicine' and query
```

	tweet_text	query	user_name
1	@BCBSMAservice out here making Albuterol the next ...	Albuterol	Dr. Nicola Chamberlain
2	ProAir, Ventolin, Proventil (albuterol) is a Short Acting ...	Albuterol	Drugs and Conditions
3	@Ideas4Russillo @afivey86 @barstoolsports Albuterol	Albuterol	Dr Dot Em

Execution finished without errors.
Result: 13 rows returned in 58ms

Relation Algebra:

π tweet_text, query, user_name
 σ user_name LIKE "Dr%" AND type = "medicine" AND query = "Albuterol"
  (ρ tu twitter_user ⋈ tu . user_handle = t . twitter_handle
   ρ t tweets ⋈ qs . tweet_id = t . tweet_id
    ρ qs query_sentiment)

15. Use Case:Identify all the negative reviews by the users about a hospital and it's treatment for a disease to improve the performance of the hospital
    Action: User
    Step:
    Actor Action: To post their views about a hospital
    System Responses: Shows the hospitals which have negative reviews and corresponding diseases

SELECT DISTINCT tweet_text, query FROM Tweets AS t JOIN query_sentiment AS qs ON t.tweet_id=qs.tweet_id WHERE type='hospital' AND query like 'Baystate Medical Center%Cancer' AND sentiment<0;

```
SELECT DISTINCT tweet_text, query FROM Tweets AS t JOIN query_sentiment AS qs ON t.tweet_id=qs.tweet_id WHERE type='hospital' AND query like 'Baystate Medical Center%Cancer' AND sentiment
```

	tweet_text	query
1	RT @masslivenews: The Baystate Medical Center ...	Baystate Medical Center AND Cancer

Relation Algebra:

δ
 π tweet_text, query
  σ type = "hospital" AND query LIKE "Baystate Medical Center%Cancer" AND sentiment < 0
   (ρ t tweets ⋈ t . tweet_id = qs . tweet_id
    ρ qs query_sentiment)

16. Use Case: Which user posted this tweet.
    Action: User
    Step:
    Actor Action: To post
    System Responses: Best hospital for given disease is displayed

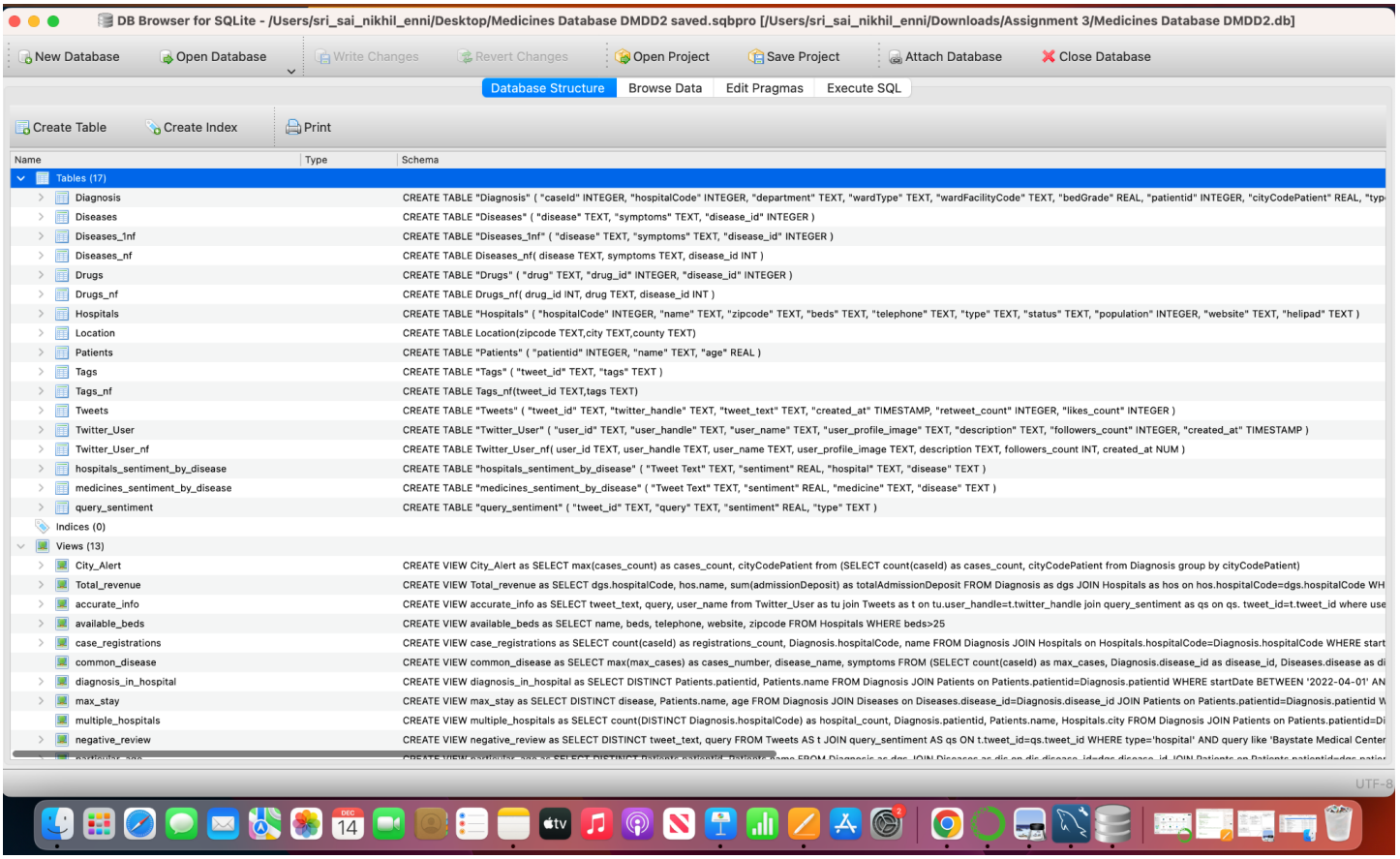
SELECT DISTINCT tu.user_name FROM Tweets as t
JOIN Twitter_User as tu on tu.user_handle=t.twitter_handle
where tweet_text='Please reduce the price of perjeta as perjeta is life saving drug. Now Zydus and Intas in race of launching the par… https://t.co/SKTiQZ9Cey';

Relation Algebra:

δ
 π tu . user_name

σ tweet_text = "Please reduce the price of perjeta as perjeta is life saving drug. Now Zydus and Intas in race of launching the par...https://t.co/SKTiQZ9Cey"
(ρ t tweets ⋈ tu . user_handle = t . twitter_handle
ρ tu twitter_user)

Database snippents:



Diseases_nf, Drugs_nf, Tags_nf, Twitter_User_nf tables depict that the tables are modified to bring the database to normal forms. All the databases satisfy 3NF.

Normalization of the tables:

1 NF:

Diseases:

1. Removed duplicate values, reduced the record count from 3439 to 202.
2. Having atomicity
3. No repetitive columns
4. Primary key: disease_id
5. Query used:
   CREATE TABLE Diseases_nf as SELECT DISTINCT disease, symptoms, disease_id FROM Diseases;

```
SELECT disease, symptoms, disease_id, count(*) FROM Diseases_nf GROUP by disease, symptoms, disease_id HAVING count(*)>1;
SELECT DISTINCT disease FROM Diseases;
```

```
Execution finished without errors.
Result: 0 rows returned in 16ms
At line 72:
SELECT disease, symptoms, disease_id, count(*) FROM Diseases_nf GROUP by disease, symptoms, disease_id HAVING count(*)>1;
```

6. Removed comma separated values for symptoms

```
SELECT * FROM Diseases_1nf
```

	disease	symptoms	disease_id
1	panic disorder	anxiety and nervousness	134
2	panic disorder	depression	134
3	panic disorder	shortness of breath	134
4	panic disorder	anxiety and nervousness	134
5	panic disorder	depression	134
6	panic disorder	shortness of breath	134
7	panic disorder	anxiety and nervousness	134
8	panic disorder	depression	134
9	panic disorder	shortness of breath	134
10	panic disorder	anxiety and nervousness	134
11	panic disorder	depression	134

```
WITH Diseases_1nf(disease, disease_id, symptoms_nf, symptoms) AS
(SELECT
    disease,
    disease_id,
    LEFT(symptoms, CHARINDEX(',', symptoms + ',') - 1),
    STUFF(symptoms, 1, CHARINDEX(',', symptoms + ','), '')
  FROM Diseases
  UNION all

  SELECT
    disease,
    disease_id,
    LEFT(symptoms, CHARINDEX(',', symptoms + ',') - 1),
    STUFF(symptoms, 1, CHARINDEX(',', symptoms + ','), '')
  FROM tmp
  WHERE
    symptoms > ''
)
```

Drugs:

1. Removed duplicate values, reduced the record count from 3439 to 1141.
2. Having atomicity
3. No repetitive columns
4. Primary key: drug_id
5. Query used:
   CREATE TABLE Drugs_nf as SELECT DISTINCT drug_id, drug, disease_id FROM Drugs;

```
SELECT count(*) FROM Drugs;
SELECT drug_id, drug, disease_id, count(*) FROM Drugs_nf GROUP by drug_id, drug, disease_id HAVING count(*)>1;
```

```
Execution finished without errors.
Result: 0 rows returned in 6ms
At line 56:
SELECT drug_id, drug, disease_id, count(*) FROM Drugs_nf GROUP by drug_id, drug, disease_id HAVING count(*)>1;
```

Patients:

1. No duplicate values
2. Having atomicity
3. No repetitive columns
4. Primary Key: patientid

```
SELECT patientid, name, age, count(*) FROM Patients GROUP by patientid, name, age HAVING count(*)>1;
```

```
Result: 0 rows returned in 209ms
At line 1:
SELECT patientid, name, age, count(*) FROM Patients GROUP by patientid, name, age HAVING count(*)>1;
```

Hospitals:

1. No duplicate values
2. Having atomicity
3. No repetitive columns
4. Primary key: hospital_id

```
SELECT hospitalCode, name, city, zipcode, beds, telephone, type, status, population, county, website, helipad, count(*) FROM Hospitals GROUP by
hospitalCode, name, city, zipcode, beds, telephone, type, status, population, county, website, helipad HAVING count(*)>1;
```

```
Execution finished without errors.
Result: 0 rows returned in 13ms
At line 1:
SELECT hospitalCode, name, city, zipcode, beds, telephone, type, status, population, county, website, helipad, count(*) FROM Hospitals GROUP by
hospitalCode, name, city, zipcode, beds, telephone, type, status, population, county, website, helipad HAVING count(*)>1;
```

Diagnosis:

1. No duplicate records
2. Having atomicity
3. No repetitive columns
4. Primary key: user_id

```
SELECT caseId, hospitalCode,department ,wardType,wardFacilityCode,bedGrade,patientid,cityCodePatient,typeOfAdmission,severityOfIllness,visitorsWithPatient,admissionDeposit,stay,startDate,endDate,disease_id,count(*) FROM Diag
```

```
Execution finished without errors.
Result: 0 rows returned in 1324ms
At line 1:
SELECT caseId, hospitalCode,department ,wardType,wardFacilityCode,bedGrade,patientid,cityCodePatient,typeOfAdmission,severityOfIllness,visitorsWithPatient,admissionDeposit,stay,startDate,endDate,disease_id,count(*) FROM Diagnosis
GROUP by caseId, hospitalCode,department ,wardType,wardFacilityCode,bedGrade,patientid,cityCodePatient,typeOfAdmission,severityOfIllness,visitorsWithPatient,admissionDeposit,stay,startDate,endDate,disease_id HAVING count(*)>1;
```

Twitter_User:

1. Removed duplicate values, reduced the record count from 4653 to 3997.
2. Having atomicity
3. No repetitive columns
4. Query used:
   CREATE TABLE Twitter_User_nf as SELECT DISTINCT user_id, user_handle, user_name, user_profile_image, description, followers_count, created_at FROM Twitter_User;

```
86
87  SELECT user_id, user_handle, user_name, user_profile_image, description, followers_count, created_at, count(*) FROM Twitter_User_nf
88  GROUP by user_id, user_handle, user_name, user_profile_image, description, followers_count, created_at HAVING count(*)>1;
89  SELECT * FROM Twitter_User;
```

```
Result: 0 rows returned in 47ms
At line 87:
SELECT user_id, user_handle, user_name, user_profile_image, description, followers_count, created_at, count(*) FROM Twitter_User_nf GROUP by user_id, user_handle, user_name, user_profile_image, description, followers_count, created_at HAVING count(*)>1;
```

Tweets:

1. No duplicate records
2. Having atomicity
3. No repetitive columns
4. Primary Key: tweet_id

Tags:

1. Removed duplicate values and reduced records from 4653 to 4608
   Query used:
   CREATE TABLE Tags_nf as SELECT DISTINCT tweet_id, tags FROM Tags;

2. Having atomicity
3. No repetitive columns
4. Primary Key: tag_id

query_sentiment:

5. No duplicate records
6. Having atomicity
7. No repetitive columns
8. Primary Key: tweet_id

```
1  SELECT tweet_id, query, sentiment,type, count(*) FROM query_sentiment GROUP by tweet_id, query, sentiment,type HAVING count(*)>1;
```

```
Execution finished without errors.
Result: 0 rows returned in 39ms
At line 1:
SELECT tweet_id, query, sentiment,type, count(*) FROM query_sentiment GROUP by tweet_id, query, sentiment,type HAVING count(*)>1;
```

2NF:

Diseases:

1. Satisfied 1NF
2. No calculated data

3. No partial dependency

Drugs:
1. Satisfied 1NF
2. No partial dependency
3. No calculated data

	disease	symptoms	disease_id
1	panic disorder	anxiety and nervousness,depression,shortness of breath	134
2	turner syndrome	groin mass,leg pain,hip pain	189
3	atrophic vaginitis	vaginal itching,vaginal dryness,painful urination	17
4	glaucoma	diminished vision,pain in eye,symptoms of eye	68
5	transient ischemic attack	loss of sensation,dizziness,headache	185
6	diabetic retinopathy	diminished vision,spots or clouds in vision,pain in eye	49
7	fibromyalgia	back pain,ache all over,neck pain	64

Patients:
1. Satisfied 1NF
2. No partial dependency
3. No calculated data

	patientid	name	age
1	31397	lydia blackwell	55.5
2	63418	ronald johnson	75.5
3	8088	michelle gonzales	35.5
4	28843	miguel carter	45.5

Hospitals:
1. Satisfied 1NF
2. Hospitals table has **city** and **county** columns, which are partially dependent on **zipcode**. New table **Location** is created to accommodate **city** and **county** along with zipcode

   Queries used:
   CREATE TABLE Location as SELECT zipcode, city, county FROM Hospitals;

   ALTER TABLE Hospitals DROP COLUMN city;
   ALTER TABLE Hospitals DROP COLUMN county;
3. No partial dependency
4. No calculated data

	hospitalCode	name	zipcode	beds	telephone	type	status	population	website	helipad
1	0	saint elizabeths hospital	20032	292.0	(202) 562-4000	PSYCHIATRIC	OPEN	292	http://dmh.dc.gov/page/saint-elizabeth...	NOT AVAILABLE
2	1	saint thomas river park hospital	37110	125.0	(931) 815-4101	GENERAL ACUTE CARE	OPEN	125	http://www.sthealth.com/locations/saint...	Y
3	2	vibra hospital of richmond llc	23230	60.0	(804) 678-7000	LONG TERM CARE	OPEN	60	www.vibrahealthcare.com	NOT AVAILABLE
4	3	pelican rehabilitation hospital, llc	70131	-999.0	(504) 378-5060	REHABILITATION	CLOSED	-999	NOT AVAILABLE	NOT AVAILABLE
5	4	sparks regional medical center	72901	476.0	(479) 441-4000	GENERAL ACUTE CARE	OPEN	476	http://www.sparks.org	Y
6	5	shannon west texas memorial hospital	76902	295.0	(325) 653-6741	GENERAL ACUTE CARE	OPEN	295	http://www.shannonhealth.com/	Y
7	6	scotland memorial hospital and edwin morgan center	28352	104.0	(910) 291-7000	GENERAL ACUTE CARE	OPEN	104	http://www.scotlandhealth.org/...	Y
8	7	mount washington pediatric hospital	21209	61.0	(410) 578-5050	GENERAL ACUTE CARE	OPEN	61	http://www.mwph.org	NOT AVAILABLE
9	8	weatherford rehabilitation hospital llc	76086	26.0	(214) 472-4101	REHABILITATION	OPEN	26	http://www.weatherfordrehab.com/	N
10	9	broaddus hospital	26416	72.0	(304) 457-1760	CRITICAL ACCESS	OPEN	72	http://www.davishealthsystem.org/	Y
11	10	geisinger healthsouth rehabilitation hospital	17822	42.0	(570) 271-6110	REHABILITATION	OPEN	42	http://www.geisingerhealthsouth.com/	N
12	11	highlands regional rehabilitation hospital	79936	41.0	(915) 298-7222	REHABILITATION	OPEN	41	http://www.highlandsrehab.com	NOT AVAILABLE

Locations:

1. Satisfied 1NF
2. No partial dependency
3. No calculated data

	zipcode	city	county
1	20032	washington	DISTRICT OF COLUMBIA
2	37110	mc minnville	WARREN
3	23230	richmond	HENRICO
4	70131	new orleans	ORLEANS
5	72901	fort smith	SEBASTIAN
6	76902	san angelo	TOM GREEN

Diagnosis:

1. Satisfied 1NF
2. No partial dependency
3. No calculated data

	caseId	hospitalCode	department	wardType	wardFacilityCode	bedGrade	patientid	cityCodePatient	typeOfAdmission	severityOfIllness	visitorsWithPatient	admissionDeposit	stay	startDate	endDate	disease_id
1	1	8	radiotherapy	R	F	2.0	31397	7.0	Emergency	Extreme	2	4911.0	5.0	2022-03-05	2022-03-08	31
2	2	2	radiotherapy	S	F	2.0	31397	7.0	Trauma	Extreme	2	5954.0	45.5	2022-04-20	2022-04-21	43
3	3	10	anesthesia	S	E	2.0	31397	7.0	Trauma	Extreme	2	4745.0	35.5	2019-05-15	2019-05-16	129
4	4	26	radiotherapy	R	D	2.0	31397	7.0	Trauma	Extreme	2	7272.0	45.5	2021-05-28	2021-06-03	112

Tags:

1. Satisfied 1NF
2. No calculated data
3. No partial dependency

Tweets:

1. Satisfied 1NF
2. No calculated data
3. No partial dependency

Tweets_User_nf:

1. Satisfied 1NF
2. No calculated data
3. No partial dependency

	user_id	user_handle	user_name	user_profile_image	description	followers_count	created_at
1	2835840133	vimal_madani	vimal madani	http://abs.twimg.com/sticky/...		0	2014-09-30 09:47:39+00:00
2	4330690359	maperdoo	Cmon Warnock	http://pbs.twimg.com/...	Resistor. Bereaved ...	1165	2015-11-30 16:42:48+00:00
3	1426846147922989056	MikeBromley15	Mike 🪑	http://pbs.twimg.com/...	Here for the challeng...	99	2021-08-15 10:00:19+00:00
4	1359714993357393923	bettylo52207153	bettylou	http://abs.twimg.com/sticky/...	Ocean State, D/x TNB...	53	2021-02-11 04:05:03+00:00
5	19163612	jamesstout	james stout	http://pbs.twimg.com/...	hack investigative ...	6972	2009-01-19 00:11:54+00:00
6	15211869	jamie_love	James Love	http://pbs.twimg.com/...	Director, Knowledge ...	10720	2008-06-23 20:59:59+00:00
7	1889492676	1stOncology	1stOncology	http://pbs.twimg.com/...	Follow top-line ...	2202	2013-09-21 09:10:17+00:00

3NF:

Diseases:

1. Satisfied 2NF
2. No transitive dependencies

Drugs:

1. Satisfied 2NF

2.No transitive dependencies

Patients:

1.Satisfied 2NF
2.No transitive dependencies

Hospitals:

1.Satisfied 2NF
2.No transitive dependencies

Diagnosis:

1.Satisfied 2NF
2.No transitive dependencies

Tags:

1.Satisfied 2NF
2.No transitive dependencies

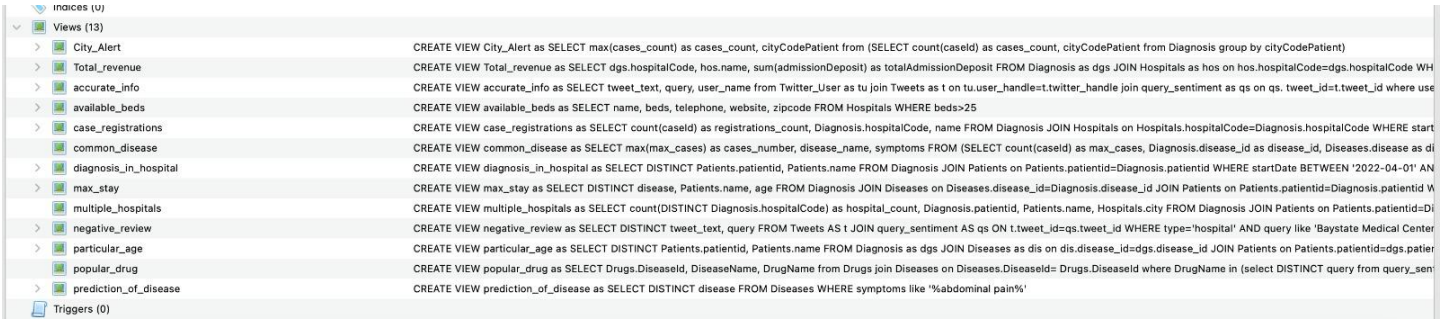
Tweets:

1.Satisfied 2NF
2.No transitive dependencies

Tweets_User:

1.Satisfied 2NF
2.No transitive dependencies

Snippet of Views and Tables:



Indices (0)	
Views (13)	
City_Alert	CREATE VIEW City_Alert as SELECT max(cases_count) as cases_count, cityCodePatient from (SELECT count(caseId) as cases_count, cityCodePatient from Diagnosis group by cityCodePatient)
Total_revenue	CREATE VIEW Total_revenue as SELECT dgs.hospitalCode, hos.name, sum(admissionDeposit) as totalAdmissionDeposit FROM Diagnosis as dgs JOIN Hospitals as hos on hos.hospitalCode=dgs.hospitalCode WH
accurate_info	CREATE VIEW accurate_info as SELECT tweet_text, query, user_name from Twitter_User as tu join Tweets as t on tu.user_handle=t.twitter_handle join query_sentiment as qs on qs. tweet_id=t.tweet_id where use
available_beds	CREATE VIEW available_beds as SELECT name, beds, telephone, website, zipcode FROM Hospitals WHERE beds>25
case_registrations	CREATE VIEW case_registrations as SELECT count(caseId) as registrations_count, Diagnosis.hospitalCode, name FROM Diagnosis JOIN Hospitals on Hospitals.hospitalCode=Diagnosis.hospitalCode WHERE start
common_disease	CREATE VIEW common_disease as SELECT max(max_cases) as cases_number, disease_name, symptoms FROM (SELECT count(caseId) as max_cases, Diagnosis.disease_id as disease_id, Diseases.disease as di
diagnosis_in_hospital	CREATE VIEW diagnosis_in_hospital as SELECT DISTINCT Patients.patientid, Patients.name FROM Diagnosis JOIN Patients on Patients.patientid=Diagnosis.patientid WHERE startDate BETWEEN '2022-04-01' AN
max_stay	CREATE VIEW max_stay as SELECT DISTINCT disease, Patients.name, age FROM Diagnosis JOIN Diseases on Diseases.disease_id=Diagnosis.disease_id JOIN Patients on Patients.patientid=Diagnosis.patientid W
multiple_hospitals	CREATE VIEW multiple_hospitals as SELECT count(DISTINCT Diagnosis.hospitalCode) as hospital_count, Diagnosis.patientid, Patients.name, Hospitals.city FROM Diagnosis JOIN Patients on Patients.patientid=Di
negative_review	CREATE VIEW negative_review as SELECT DISTINCT tweet_text, query FROM Tweets AS t JOIN query_sentiment AS qs ON t.tweet_id=qs.tweet_id WHERE type='hospital' AND query like 'Baystate Medical Center
particular_age	CREATE VIEW particular_age as SELECT DISTINCT Patients.patientid, Patients.name FROM Diagnosis as dgs JOIN Diseases as dis on dis.disease_id=dgs.disease_id JOIN Patients on Patients.patientid=dgs.patier
popular_drug	CREATE VIEW popular_drug as SELECT Drugs.DiseaseId, DiseaseName, DrugName from Drugs join Diseases on Diseases.DiseaseId= Drugs.DiseaseId where DrugName in (select DISTINCT query from query_sen
prediction_of_disease	CREATE VIEW prediction_of_disease as SELECT DISTINCT disease FROM Diseases WHERE symptoms like '%abdominal pain%'
Triggers (0)	

Views:

Table: City_Alert

	cases_count	cityCodePatient
1	124011	8.0

Table: Total_revenue

	hospitalCode	name	totalAdmissionDeposit
1	28	larned ...	92301.0

Table: available_beds

	name	beds	telephone	website	zipcode
1	saint elizabeths hospital	292.0	(202) 562-4000	http://dmh.dc.gov/page/saint-elizabeths-hospital	20032
2	vibra hospital of richmond llc	60.0	(804) 678-7000	www.vibrahealthcare.com	23230
3	sparks regional medical center	476.0	(479) 441-4000	http://www.sparks.org	72901
4	shannon west texas memorial hospital	295.0	(325) 653-6741	http://www.shannonhealth.com/	76902

Table: negative_review

	tweet_text	query
1	RT ...	Baystat...
2	The ...	Baystat...

Table: diagnosis_in_hospital

	patientid	name
1	115513	ryan ...
2	74865	joseph ...
3	75068	gerald ...
4	88532	jeff ...
5	17161	phillip ...
6	128549	don ...
7	23912	cory ...

Table: accurate_info

	tweet_text	query	user_name
1	@BCBSMAservice out here making Albuterol the next ...	Albuterol	Dr. Nicola Chamberlain
2	ProAir, Ventolin, Proventil (albuterol) is a Short Acting ...	Albuterol	Drugs and Conditions
3	@Ideas4Russillo @afivey86 @barstoolsports Albuterol	Albuterol	Dr Dot Em
4	@Ideas4Russillo @afivey86 @barstoolsports Albuterol	Albuterol	Dr Dot Em
5	@barstoolsports Did he slide him some albuterol? 🤔🤔	Albuterol	Dr Dot Em
6	@barstoolsports Did he slide him some albuterol? 🤔🤔	Albuterol	Dr Dot Em

Table: case_registrations

	registrations_count	hospitalCode	name
1	110	1	saint ...
2	88	2	vibra ...
3	150	3	pelican...
4	23	4	sparks ...
5	101	5	shanno...

Tables:

Table: Diagnosis

	caseId	hospitalCode	department	wardType	wardFacilityCode	bedGrade	patientid	cityCodePatient	typeOfAdmission	severityOfIllness	visitorsWithPatient	admissionDeposit	stay	startDate	endDate	disease_id
1	1	8	radiotherapy	R	F	2.0	31397	7.0	Emergency	Extreme	2	4911.0	5.0	2022-03-05	2022-03-08	31
2	2	2	radiotherapy	S	F	2.0	31397	7.0	Trauma	Extreme	2	5954.0	45.5	2022-04-20	2022-04-21	43
3	3	10	anesthesia	S	E	2.0	31397	7.0	Trauma	Extreme	2	4745.0	35.5	2019-05-15	2019-05-16	129
4	4	26	radiotherapy	R	D	2.0	31397	7.0	Trauma	Extreme	2	7272.0	45.5	2021-05-28	2021-06-03	112
5	5	26	radiotherapy	S	D	2.0	31397	7.0	Trauma	Extreme	2	5558.0	45.5	2019-10-05	2019-10-10	199
6	6	23	anesthesia	S	F	2.0	31397	7.0	Trauma	Extreme	2	4449.0	15.5	2021-10-01	2021-10-04	190

Table: Diseases

	disease	symptoms	disease_id
1	panic disorder	anxiety and nervousness,depression,shortness of breath	134
2	panic disorder	anxiety and nervousness,depression,shortness of breath	134
3	panic disorder	anxiety and nervousness,depression,shortness of breath	134
4	panic disorder	anxiety and nervousness,depression,shortness of breath	134
5	panic disorder	anxiety and nervousness,depression,shortness of breath	134

	drug	drug_id	disease_id
1	clonazepam	145	134
2	alprazolam	18	134
3	lorazepam	349	134

```
select * from Hospitals;
```

hospitalCode	name	city	zipcode	beds	telephone	type	status	population	county	website	helipad
0	saint elizabeths hospital	washington	20032	292.0	(202) 562-4000	PSYCHIATRIC	OPEN	292	DISTRICT OF COLUMBIA	http://dmh.dc.gov/page/saint-elizabeths-hospital	NOT AVAILABLE
1	saint thomas river park hospital	mc minnville	37110	125.0	(931) 815-4101	GENERAL ACUTE CARE	OPEN	125	WARREN	http://www.sthealth.com/locations/saint-thomas-river-par...	Y
2	vibra hospital of richmond llc	richmond	23230	60.0	(804) 678-7000	LONG TERM CARE	OPEN	60	HENRICO	www.vibrahealthcare.com	NOT AVAILABLE

```
1  select * from Patients;
```

	patientid	name	age
1	31397	lydia blackwell	55.5
2	63418	ronald johnson	75.5
3	8088	michelle gonzales	35.5

```
1  select * from Tweets;
```

	tweet_id	twitter_handle	tweet_text	created_at	retweet_count	likes_count
1	1590701484282241024	vimal_madani	Please reduce the price of perjeta as perjeta is life saving ...	2022-11-10 13:42:42+00:00	0	0
2	1589430238651764736	maperdoo	@WilsonMaryAnne1 @ThanksCancer I got congestive hear...	2022-11-07 01:31:13+00:00	0	0
3	1588501059747840001	MikeBromley15	@Proxima_Project @JCautec It's bs. We pay drug ...	2022-11-04 11:59:00+00:00	0	5

```
1  select * from Twitter_User;
```

	user_id	user_handle	user_name	user_profile_image	description	followers_count	created_at
1	2835840133	vimal_madani	vimal madani	http://abs.twimg.com/sticky/default_profile_images/...		0	2014-09-30 09:47:39+00:0
2	4330690359	maperdoo	Cmon Warnock	http://pbs.twimg.com/profile_images/...	Resistor. Bereaved Mom. Reluctant Pharmacist....	1165	2015-11-30 16:42:48+00:0
3	1426846147922989056	MikeBromley15	Mike 🥷	http://pbs.twimg.com/profile_images/...	Here for the challenge. Business owner, dad, BJJ Brown ...	99	2021-08-15 10:00:19+00:0

```
1  select * from query_sentiment;
```

	tweet_id	query	sentiment	type
1	1590701484282241024	Perjeta	0.0	medicine
2	1589430238651764736	Perjeta	-0.316666666666667	medicine
3	1588501059747840001	Perjeta	-0.125	medicine

Group Members:

Team 44- Elixir

NUID	Name	email
002725894	Sri Sai Nikhil Enni	enni.s@northeastern.edu
002725890	Vaishnavi Yadamreddy	yadamreddy@northeastern.edu

References:

Hospitals List:
https://health.usnews.com/best-hospitals/area/ma
https://www.kaggle.com/datasets/carlosaguayo/usa-hospitals

Medicine_prescription_records:
https://www.kaggle.com/datasets/manncodes/drug-prescription-to-disease-dataset

Hospitals:
https://www.kaggle.com/datasets/carlosaguayo/usa-hospitals

Diseases associated with drug:
https://www.kaggle.com/datasets/manncodes/drug-prescription-to-disease-dataset?resource=download

Patients disease (diagnosis):
https://datahack.analyticsvidhya.com/contest/janatahack-healthcare-analytics-ii/

Diagnosis:
https://datahack.analyticsvidhya.com/contest/janatahack-healthcare-analytics-ii/#ProblemStatement