1. Diseases
    - using “disease_symptom” dataset https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset?select=dataset.csv
    - Unique diseases and their descriptions, rating of severity, and a maximum of 4 precautions
    - “Disease” determines every other feature and no other FD
        - 1 key satisfies both 3nf and bcnf
    - 41 entries
    - Has one to many relationships with Symptoms
    - Diseases(name:VARCHAR(50) [PK], description:VARCHAR(400), severity:INT, precaution_1:VARCHAR(50), precaution_2:VARCHAR(50), precaution_3:VARCHAR(50), precaution_4:VARCHAR(50))
    - Assumptions and Explanation of Modeling Choices:
        - Given that we are making a project centered around disease identification and diagnosis, we would naturally need to store information about our diseases.
        - This is an entity because each patient we diagnose will have symptoms and those symptoms will be mapped to a disease. We cannot store disease as an attribute of symptoms because many diseases have overlapping symptoms, hence a many to one relationship. Each disease has their own unique attributes, so it needs a separate table with each disease as a primary key to be identified. We also want to associate precautions and severity with the disease rather than the symptoms as that provides a clear answer to the user. Otherwise there could be multiple precautions possible for a symptom and they could also conflict. As explained before cardinality is many to one. Many symptoms map to one disease. This is because that's how it works in the real world. Having many diseases map onto a single symptom will not allow us to diagnose the user and thus the advice would be meaningless.
        - There is also a many to at most 1 relation between drug table and diseases table.
        - This is because the dataset we use lists many drugs that can help alleviate a particular disease. And when querying for the user we would like to match a specific diagnosis with all potential drugs which means that a disease can have more than 1 drug that cures it. Also it is potentially true that a drug(s) won't work on any of the diseases hence the 0 both ways in the relation.
        - While it is true that a single drug can cure multiple diseases it is our design choice to map a set of drugs to at most a single disease so that it is easier to display drug information to the user after we have finished the diagnosis phase.
2. Symptoms

- using “symptom_checker” dataset https://www.kaggle.com/datasets/rabisingh/symptom-checker?select=Testing.csv
- data entries of patients’ symptoms and the disease they were diagnosed with
- patient_id is the only key and no other FD
    - 1 key and not other FD satisfies both 3nf and bcnf
    - patient_id is added only to uniquely identify each tuple, otherwise, since there is no FD, there will be no way to identify them
- 4961 entries
- disease referencing Diseases.name followed by 132 boolean (TINYINT(1)) symptoms
- Symptoms(patient_id:INT [PK], disease:VARCHAR(50) [FK to Diseases.name], itching:TINYINT(1), skin_rash:TINYINT(1), nodal_skin_eruptions:TINYINT(1), continuous_sneezing:TINYINT(1), shivering:TINYINT(1), chills:TINYINT(1), joint_pain:TINYINT(1), stomach_pain:TINYINT(1), acidity:TINYINT(1), ulcers_on_tongue:TINYINT(1), muscle_wasting:TINYINT(1), vomiting:TINYINT(1), burning_micturition:TINYINT(1), spotting_ urination:TINYINT(1), fatigue:TINYINT(1), weight_gain:TINYINT(1), anxiety:TINYINT(1), cold_hands_and_feets:TINYINT(1), mood_swings:TINYINT(1), weight_loss:TINYINT(1), restlessness:TINYINT(1), lethargy:TINYINT(1), patches_in_throat:TINYINT(1), irregular_sugar_level:TINYINT(1), cough:TINYINT(1), high_fever:TINYINT(1), sunken_eyes:TINYINT(1), breathlessness:TINYINT(1), sweating:TINYINT(1), dehydration:TINYINT(1), indigestion:TINYINT(1), headache:TINYINT(1), yellowish_skin:TINYINT(1), dark_urine:TINYINT(1), nausea:TINYINT(1), loss_of_appetite:TINYINT(1), pain_behind_the_eyes:TINYINT(1), back_pain:TINYINT(1), constipation:TINYINT(1), abdominal_pain:TINYINT(1), diarrhoea:TINYINT(1), mild_fever:TINYINT(1), yellow_urine:TINYINT(1), yellowing_of_eyes:TINYINT(1), acute_liver_failure:TINYINT(1), fluid_overload:TINYINT(1), swelling_of_stomach:TINYINT(1), swelled_lymph_nodes:TINYINT(1), malaise:TINYINT(1), blurred_and_distorted_vision:TINYINT(1), phlegm:TINYINT(1), throat_irritation:TINYINT(1), redness_of_eyes:TINYINT(1), sinus_pressure:TINYINT(1), runny_nose:TINYINT(1), congestion:TINYINT(1), chest_pain:TINYINT(1), weakness_in_limbs:TINYINT(1), fast_heart_rate:TINYINT(1), pain_during_bowel_movements:TINYINT(1), pain_in_anal_region:TINYINT(1), bloody_stool:TINYINT(1), irritation_in_anus:TINYINT(1), neck_pain:TINYINT(1),

dizziness:TINYINT(1), cramps:TINYINT(1), bruising:TINYINT(1), obesity:TINYINT(1), swollen_legs:TINYINT(1), swollen_blood_vessels:TINYINT(1), puffy_face_and_eyes:TINYINT(1), enlarged_thyroid:TINYINT(1), brittle_nails:TINYINT(1), swollen_extremeties:TINYINT(1), excessive_hunger:TINYINT(1), extra_marital_contacts:TINYINT(1), drying_and_tingling_lips:TINYINT(1), slurred_speech:TINYINT(1), knee_pain:TINYINT(1), hip_joint_pain:TINYINT(1), muscle_weakness:TINYINT(1), stiff_neck:TINYINT(1), swelling_joints:TINYINT(1), movement_stiffness:TINYINT(1), spinning_movements:TINYINT(1), loss_of_balance:TINYINT(1), unsteadiness:TINYINT(1), weakness_of_one_body_side:TINYINT(1), loss_of_smell:TINYINT(1), bladder_discomfort:TINYINT(1), foul_smell_of urine:TINYINT(1), continuous_feel_of_urine:TINYINT(1), passage_of_gases:TINYINT(1), internal_itching:TINYINT(1), toxic_look_(typhos):TINYINT(1), depression:TINYINT(1), irritability:TINYINT(1), muscle_pain:TINYINT(1), altered_sensorium:TINYINT(1), red_spots_over_body:TINYINT(1), belly_pain:TINYINT(1), abnormal_menstruation:TINYINT(1), dischromic _patches:TINYINT(1), watering_from_eyes:TINYINT(1), increased_appetite:TINYINT(1), polyuria:TINYINT(1), family_history:TINYINT(1), mucoid_sputum:TINYINT(1), rusty_sputum:TINYINT(1), lack_of_concentration:TINYINT(1), visual_disturbances:TINYINT(1), receiving_blood_transfusion:TINYINT(1), receiving_unsterile_injections:TINYINT(1), coma:TINYINT(1), stomach_bleeding:TINYINT(1), distention_of_abdomen:TINYINT(1), history_of_alcohol_consumption:TINYINT(1), fluid_overload:TINYINT(1), blood_in_sputum:TINYINT(1), prominent_veins_on_calf:TINYINT(1), palpitations:TINYINT(1), painful_walking:TINYINT(1), pus_filled_pimples:TINYINT(1), blackheads:TINYINT(1), scurring:TINYINT(1), skin_peeling:TINYINT(1), silver_like_dusting:TINYINT(1), small_dents_in_nails:TINYINT(1), inflammatory_nails:TINYINT(1), blister:TINYINT(1), ed_sore_around_nose:TINYINT(1), yellow_crust_ooze:TINYINT(1))

- Assumptions and Explanation of Modeling Choices:
    - We need to be able to store the symptoms profile of a disease in order to diagnose. Hence the justification of having some sort of symptom information/ component.
    - The user will enter their symptoms where it will be mapped to an approximate set of symptoms, given by the symptom ID.

- We cannot store symptom ID with diseases as that would lead to varying numbers of extra columns in diseases and this would be inefficient and lead to redundancies.
- We have also added a symptomID to uniquely identify the set of symptoms. This will be used later down the line when we are querying using SELECT UNIQUE.
- The cardinality of the relationship is many to one as explained in the disease relation section: We have multiple symptoms that help us accurately diagnose a single cause of these symptoms, so the disease.

3. Drugs
    - 3959 entries
    - Drug names and Diseases: https://www.kaggle.com/datasets/jithinanievarghese/drugs-related-to-common-treatments
        - the “side effects” section is pure shit
    - Side effects https://www.kaggle.com/datasets/shudhanshusingh/250k-medicines-usage-side-effects-and-substitutes
        - maximum of 3 side effects
        - need some string parsing to map the side effect from this dataset to the one above
    - Contains information about commercial drugs, including the medical condition it’s used for, side effects, pregnancy rating (see description in the link), interacts with alcohol, and rating (used for sorting), and the class of the drug
    - has many-to-1 relationship with Diseases
    - “drug” determines every other columns, and no other FD
        - each drugs will have a disease it’s associated to
            - eventhough in reality one drug can be used for multiple diseases, this dataset is limited to 1 disease per drugs
        - each drugs will have their own combination of side effect, might not be unique, and this dataset limits to at most 3 side effects per drugs
        - each drup will have their own average user rating, pregnancy category (how safe it is to use during pregnancy), and alcohol interaction boolean value (whether the drug interacts with alcohol or not)
        - 1 key, satisfies both 3nf and bcnf
    - Drugs(name:VARCHAR(100) [PK], disease:VARCHAR(50) [FK to Diseases.name], side_effect1:VARCHAR(50), side_effect2:VARCHAR(50), side_effect3:VARCHAR(50), rating: FLOAT, pregnancy_category:VARCHAR(1), alcohol:TINYINT(1))
    - has many-to-many recursive relationship with other drugs in the form of substitue drugs
        - relationship portrayed using a separate table
        - Drug_Relations(drug1:VARCHAR(100) [FK to Drugs.name], drug2:VARCHAR(100) [FK to Drugs.name])

- Assumptions and Explanation of Modeling Choices:
    - Part of our project is to not only IDENTIFY the disease plaguing our user, but also what medicines they can use to CURE themselves. Hence we have a Entity set for all drugs we could find in our dataset.
    - We have not combined the drug information as an attribute for diseases as like the symptoms example it would lead to redundancies and multiple uneven columns in the disease table as many drugs map onto a single disease in our model.
    - We have explained the relation between drugs and disease in the disease section.
    - There is a self relation in the drugs table. This is because many drugs have substitute drugs and drugs with generic names. We want to capture this information to provide alternatives. Having to store this information separately would lead to redundancy in that we would have all columns in the substitution list linked to the Drug table anyway, so we can just capture this as a relationship set.

4. User_Queries
    - user input
    - query_id is the key and no other functional dependencies
        - 1 key satisfies both 3nf and bcnf
        - This is because the lists of symptoms, location, and date alone have no relation to 1 another (it will be too complicated to determine which symptom will result in another symptom), and the query_id is just an identifier for each tuple containing those attributes
    - User_Queries(query_id:INT [PK], itching:TINYINT(1), skin_rash:TINYINT(1), nodal_skin_eruptions:TINYINT(1), continuous_sneezing:TINYINT(1), shivering:TINYINT(1), chills:TINYINT(1), joint_pain:TINYINT(1), stomach_pain:TINYINT(1), acidity:TINYINT(1), ulcers_on_tongue:TINYINT(1), muscle_wasting:TINYINT(1), vomiting:TINYINT(1), burning_micturition:TINYINT(1), spotting_ urination:TINYINT(1), fatigue:TINYINT(1), weight_gain:TINYINT(1), anxiety:TINYINT(1), cold_hands_and_feets:TINYINT(1), mood_swings:TINYINT(1), weight_loss:TINYINT(1), restlessness:TINYINT(1), lethargy:TINYINT(1), patches_in_throat:TINYINT(1), irregular_sugar_level:TINYINT(1), cough:TINYINT(1), high_fever:TINYINT(1), sunken_eyes:TINYINT(1), breathlessness:TINYINT(1), sweating:TINYINT(1), dehydration:TINYINT(1), indigestion:TINYINT(1), headache:TINYINT(1), yellowish_skin:TINYINT(1), dark_urine:TINYINT(1), nausea:TINYINT(1), loss_of_appetite:TINYINT(1), pain_behind_the_eyes:TINYINT(1), back_pain:TINYINT(1), constipation:TINYINT(1), abdominal_pain:TINYINT(1), diarrhoea:TINYINT(1),

mild_fever:TINYINT(1), yellow_urine:TINYINT(1),
yellowing_of_eyes:TINYINT(1), acute_liver_failure:TINYINT(1),
fluid_overload:TINYINT(1), swelling_of_stomach:TINYINT(1),
swelled_lymph_nodes:TINYINT(1), malaise:TINYINT(1),
blurred_and_distorted_vision:TINYINT(1), phlegm:TINYINT(1),
throat_irritation:TINYINT(1), redness_of_eyes:TINYINT(1),
sinus_pressure:TINYINT(1), runny_nose:TINYINT(1),
congestion:TINYINT(1), chest_pain:TINYINT(1),
weakness_in_limbs:TINYINT(1), fast_heart_rate:TINYINT(1),
pain_during_bowel_movements:TINYINT(1),
pain_in_anal_region:TINYINT(1), bloody_stool:TINYINT(1),
irritation_in_anus:TINYINT(1), neck_pain:TINYINT(1),
dizziness:TINYINT(1), cramps:TINYINT(1), bruising:TINYINT(1),
obesity:TINYINT(1), swollen_legs:TINYINT(1),
swollen_blood_vessels:TINYINT(1), puffy_face_and_eyes:TINYINT(1),
enlarged_thyroid:TINYINT(1), brittle_nails:TINYINT(1),
swollen_extremeties:TINYINT(1), excessive_hunger:TINYINT(1),
extra_marital_contacts:TINYINT(1), drying_and_tingling_lips:TINYINT(1),
slurred_speech:TINYINT(1), knee_pain:TINYINT(1),
hip_joint_pain:TINYINT(1), muscle_weakness:TINYINT(1),
stiff_neck:TINYINT(1), swelling_joints:TINYINT(1),
movement_stiffness:TINYINT(1), spinning_movements:TINYINT(1),
loss_of_balance:TINYINT(1), unsteadiness:TINYINT(1),
weakness_of_one_body_side:TINYINT(1), loss_of_smell:TINYINT(1),
bladder_discomfort:TINYINT(1), foul_smell_of urine:TINYINT(1),
continuous_feel_of_urine:TINYINT(1), passage_of_gases:TINYINT(1),
internal_itching:TINYINT(1), toxic_look_(typhos):TINYINT(1),
depression:TINYINT(1), irritability:TINYINT(1), muscle_pain:TINYINT(1),
altered_sensorium:TINYINT(1), red_spots_over_body:TINYINT(1),
belly_pain:TINYINT(1), abnormal_menstruation:TINYINT(1), dischromic
_patches:TINYINT(1), watering_from_eyes:TINYINT(1),
increased_appetite:TINYINT(1), polyuria:TINYINT(1),
family_history:TINYINT(1), mucoid_sputum:TINYINT(1),
rusty_sputum:TINYINT(1), lack_of_concentration:TINYINT(1),
visual_disturbances:TINYINT(1), receiving_blood_transfusion:TINYINT(1),
receiving_unsterile_injections:TINYINT(1), coma:TINYINT(1),
stomach_bleeding:TINYINT(1), distention_of_abdomen:TINYINT(1),
history_of_alcohol_consumption:TINYINT(1), fluid_overload:TINYINT(1),
blood_in_sputum:TINYINT(1), prominent_veins_on_calf:TINYINT(1),
palpitations:TINYINT(1), painful_walking:TINYINT(1),

pus_filled_pimples:TINYINT(1), blackheads:TINYINT(1), scurring:TINYINT(1), skin_peeling:TINYINT(1), silver_like_dusting:TINYINT(1), small_dents_in_nails:TINYINT(1), inflammatory_nails:TINYINT(1), blister:TINYINT(1), ed_sore_around_nose:TINYINT(1), yellow_crust_ooze:TINYINT(1), location: VARCHAR(200), date:DATE)

- Assumptions and Explanation of Modeling Choices:
    - We would like to do some metadata analysis. To do this we would need to store information on the users that use our product. Since user information is a separate object than the ones we have discussed such as diseases and drugs, it gets its own table. It also does not relate to our other entity sets as it is derived from user inputs, not the datasets we have prepared. Hence it is ever changing and cannot be stored within a static table like the other ones.
    - We can use the anonymous information of the query given by query id to track the outbreak and prevalence of certain diseases in certain areas over a given time frame. This is another feature we wish to implement in our product. Hence the particular attribute selection. We have also tried to keep user data anonymous so that the user doesn't need to input personal details.

5. Drug_Reviews
    - https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018
    - 214800 entries
    - review_id is the only key and no other FD
        - drug does not always determine the condition it's used for, since people could be using the wrong drug for the wrong condition.
        - drug and condition alone cannot determine the review since there are other factor such as the allergy and state of mind of the user that will affect the review they give
        - 1 key and no other FD satisfies both 3nf and bcnf
    - note that "condition" is not the same thing as "disease", since something like "birth control" counts as a condition.
    - the "useful_count" is the number of users who found the review useful
    - Drug_Reviews(review_id:INT [PK], drug:VARCHAR(100) [FK to Drugs.name], condition:VARCHAR(100), review:VARCHAR(255), rating:INT, date:DATE, useful_count:INT)
    - Assumptions and Explanation of Modeling Choices:
        - We want users to know the experience of other consumers with certain drugs. Hence we have a dataset of reviews patients have given for certain drugs and their experience with it. This is a feature we want to put along with our output for which drugs the user should buy for a certain disease.
        - There is a many to one relationship between the reviews and drugs table as a single drug can have multiple views but it is unlikely for a single review to mention multiple classes of drugs.

- Because of the many to one relationship we cannot store the drug review as an attribute of drugs. This is because if we did, there would be multiple columns or a list of reviews in the drugs table and this would violate the 3NF/ BCNF.