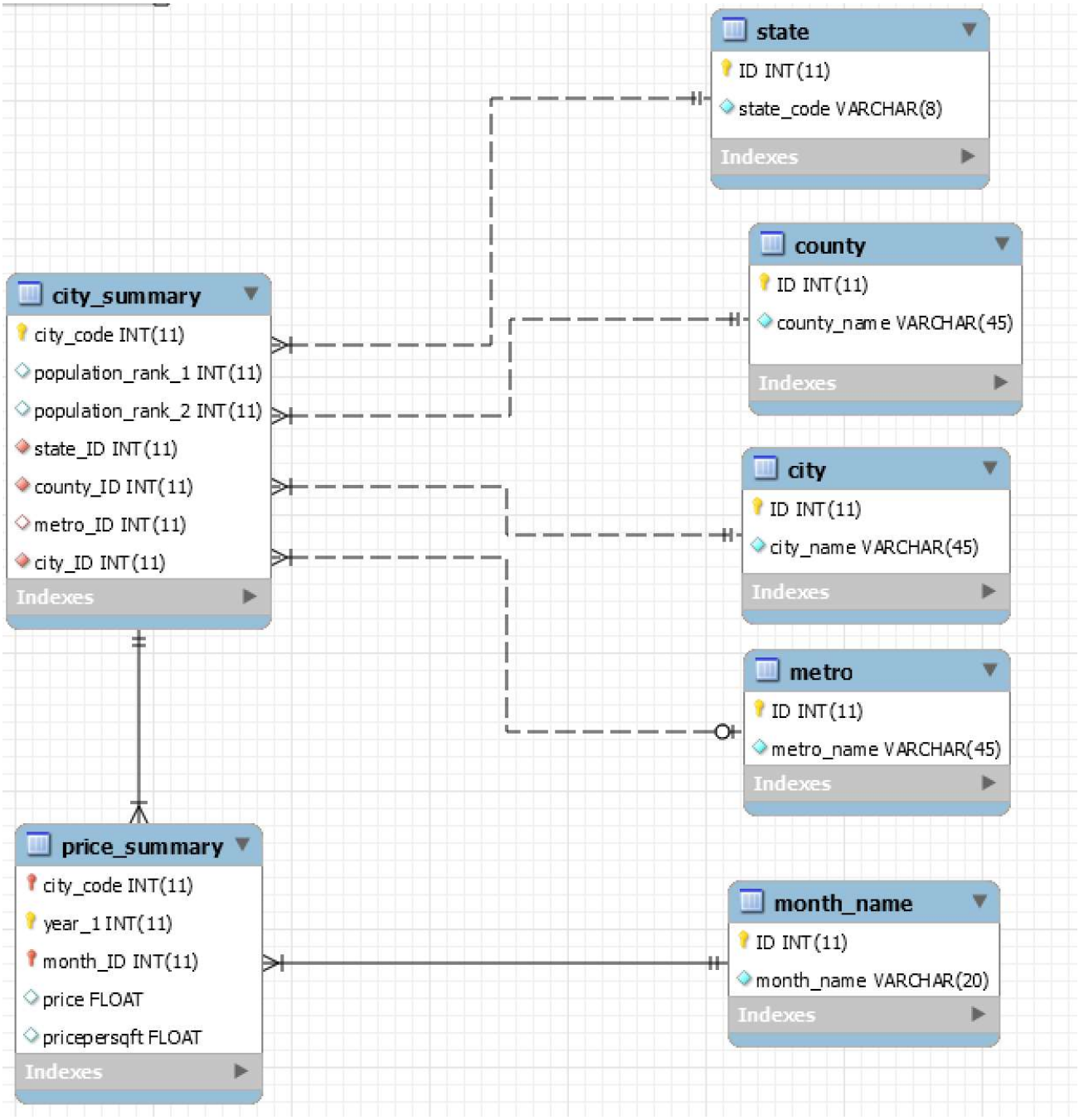
BUAN 6320.003
Final project
Group 35

Question 1: Create a data model to fit the data provided in the two text files provided by Zillow. Put screenshots of the diagram along with notes on your assumptions and a paragraph describing the normalization level of the model.



Notes:

1. The states, counties, cities and metros tables.

The states, counties, cities and metros will have their own tables to remove redundancy - This will eliminate the names being repeated continuously for the corresponding city codes.
In the event of a state/county/city/metro has a name change or any new field needs to be added, only the respective parent table column needs to change.
The use of “int” data types as identifiers also helps in better memory utilization.

The relationship to the “city_summary” table is 1:m for state, county and city tables:

- 1 state will have many city codes; 1 city code will be only in 1 state.
- 1 county will have many city codes; 1 city code will be only in 1 county
- 1 city will have many city codes; 1 city code will be only in 1 city

The relationship between the metro table and city summary is slightly different as 1 metro can have many city codes whilst a city code may or may not have a metro.

2. The city summary table

This table consists of all the city codes. There are 13131 codes in “price.csv” and 11919 city codes in “pricepersqft.csv”. Both tables combined, there are 13196 unique city codes.
Both the population rank as per the "price.csv" file (included in population_rank_1 column in table) and population rank as per the "pricepersqft.csv" file (included in population_rank_2 column in table) are included because the corresponding values of the 2 tables are different in majority of the city codes.

This table also connects all the locations by consisting the respective foreign keys from state, county, metro and city.

3. Price summary table

Has a compound primary key consisting of the city code, year and month. The “price” and “pricepersqft” for each city code, for each year and for each month will be added to the data table as entries (rows).

Since RDBMS allows vertical scalability, it is a better method to store these seasonally updating, increasing data.
Hence the schema of the database will not require frequent change.
This will help ensure integrity.

4. Month table

We have included this as having an “int” data type as primary key is more efficient than characters.
Int requires less memory than varchar.

Normalization level of the model

The model is normalized up to the 3NF level.

1NF level in the model:
The tables for state, count, city, metro and month consists of a unique identifier and the respective name.
This is to ensure there is no duplication of data with in any rows of a column.
It ensures that in case any changes should happen to a name, or any additional data regarding a state/county/city/metro/month needs to be added, the change should be done only once with in that respective parent table.

2NF level in the model:
The table "price_summary" consists of a composite primary key. The non-key columns "price" and "pricepersqft" depend on the entire composite primary key for unique identification.

3NF level in the model:
All the non-key columns in every table depends on the respective primary keys only.