Project Exam 2

BUAN 6320.002 - DBM 25

Weiyang Sun, Danqin Shang, Lyuqihui Shi

Question 1 - Normalization

The first step for this project required us to create a model for the given dataset. After looking through the data provided in the two CSV files (Price.csv & Pricepersqft.csv), we created 5 main tables and 5 intermediate tables to build the model. The primary guiding factors are based on obtaining the 3rd Normal Form (3NF), while also, maintaining the relationships between each field/column. For example, since city code can uniquely determine the location information, including city, metro, county and state, we will use it as the primary key for the "City" table. For other tables, refer to Figure 1 below, which shows the full Entity-Relations (ER) diagram.

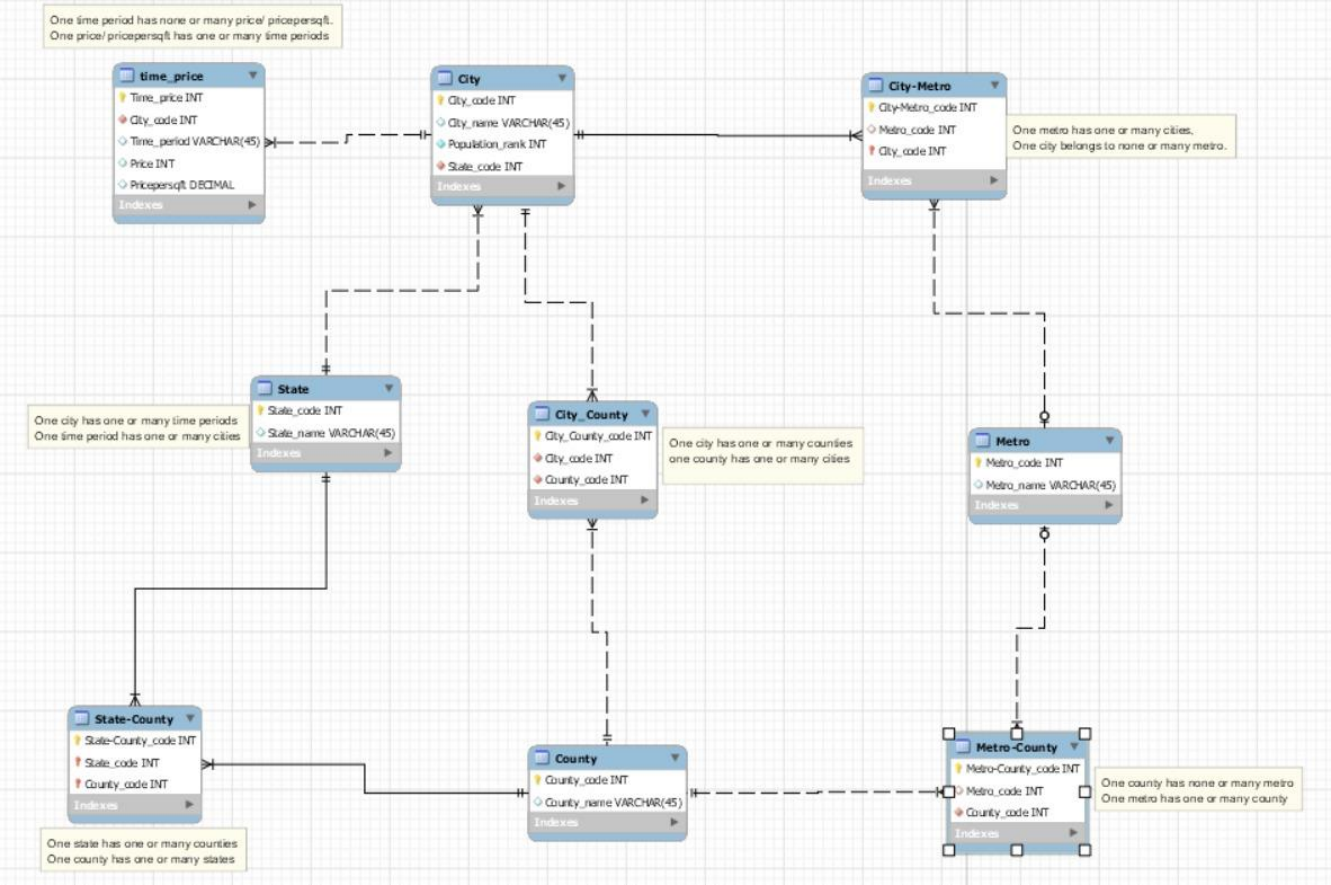

Figure 1

Since in normalization, it is necessary to first pass through the first normal form and second normal form, to get to the third normal form, we will begin by explaining using 1NF first.

To achieve the first normal form (1NF), we need to have all data in one table and each entity should only have a single value. We created "Time period" column to replace all the time-reference columns, and another two columns, price and pricepersqft in a new table. This was done through R stacking function. In all, we would expect 989700 rows (because some cities would have price but not price per sqft and vice-versa).

```
transpose <- t(interested_data)

transpose <- data.frame(transpose = unlist(transpose))

str(transpose)

total <- 20
# create progress bar
pb <- txtProgressBar(min = 0, max = total, style = 3)
for(i in 1:total){
  Sys.sleep(0.1)
  # update progress bar
  setTxtProgressBar(pb, i)

  trying <- data.frame(context2[1], stack(context2[2:ncol(context2)]))

}
close(pb)
```

Figure 2

A sample of the dataset which has been stacked, can be seen below (Figure 3):-

	A	B	C	D	E	F	G	H	I	J
1	City Code	City	Metro	County	State	Population Rank	Time Period	Price	Pricepersqft	
2	6181	New York	New York	Queens	NY	1	10-Nov	N/A	N/A	
3	6181	New York	New York	Queens	NY	1	10-Dec	N/A	N/A	
4	6181	New York	New York	Queens	NY	1	11-Jan	N/A	N/A	
5	6181	New York	New York	Queens	NY	1	11-Feb	N/A	N/A	
6	6181	New York	New York	Queens	NY	1	11-Mar	N/A	N/A	
7	6181	New York	New York	Queens	NY	1	11-Apr	N/A	N/A	
8	6181	New York	New York	Queens	NY	1	11-May	N/A	N/A	
9	6181	New York	New York	Queens	NY	1	11-Jun	N/A	N/A	
10	6181	New York	New York	Queens	NY	1	11-Jul	N/A	N/A	
11	6181	New York	New York	Queens	NY	1	11-Aug	N/A	N/A	
12	6181	New York	New York	Queens	NY	1	11-Sep	N/A	N/A	
13	6181	New York	New York	Queens	NY	1	11-Oct	N/A	N/A	
14	6181	New York	New York	Queens	NY	1	11-Nov	N/A	N/A	
15	6181	New York	New York	Queens	NY	1	11-Dec	1746	1.39	
16	6181	New York	New York	Queens	NY	1	12-Jan	1752	1.388	
17	6181	New York	New York	Queens	NY	1	12-Feb	1764	1.392	
18	6181	New York	New York	Queens	NY	1	12-Mar	1778	1.396	
19	6181	New York	New York	Queens	NY	1	12-Apr	1792	1.402	
20	6181	New York	New York	Queens	NY	1	12-May	1804	1.412	
21	6181	New York	New York	Queens	NY	1	12-Jun	1813	1.42	
22	6181	New York	New York	Queens	NY	1	12-Jul	1814	1.424	
23	6181	New York	New York	Queens	NY	1	12-Aug	1810	1.422	
24	6181	New York	New York	Queens	NY	1	12-Sep	1805	1.42	
25	6181	New York	New York	Queens	NY	1	12-Oct	1806	1.426	
26	6181	New York	New York	Queens	NY	1	12-Nov	1817	1.442	
27	6181	New York	New York	Queens	NY	1	12-Dec	1831	1.458	
28	6181	New York	New York	Queens	NY	1	13-Jan	1851	1.468	
29	6181	New York	New York	Queens	NY	1	13-Feb	1870	1.478	

Figure 3

Then we worked on the second normal form (2NF). We split the 1NF table into 2 parts, one includes location information and the other one includes City Code, time and price information. Since city, metro, county and state, as well as population rank are uniquely determined by city code, these columns are all depend on city code, no matter one-to-one or one-to-many relationship. So, city code is the primary key for this sub table, no

functional dependency. As for price information, each price is determined by the combination of city and time, so we created Time_Price_code as primary key of this table. The sample refers to Figure 4.

	A	B	C	D	E	F	G	H	I	J	K	L
1	City Code	City	Metro	County	State	Population Rank		Time_price_code	Time_Period	City_City_	Price	Pricepersqft
2	6181	New York	New York	Queens	NY	1		1	10-Nov	6181	0	0
3	12447	Los Angeles	Los Angeles	Los Angeles	CA	2		2	10-Dec	6181	0	0
4	17426	Chicago	Chicago	Cook	IL	3		3	11-Jan	6181	0	0
5	39051	Houston	Houston	Harris	TX	4		4	11-Feb	6181	0	0
6	13271	Philadelphia	Philadelphia	Philadelphia	PA	5		5	11-Mar	6181	0	0
7	40326	Phoenix	Phoenix	Maricopa	AZ	6		6	11-Apr	6181	0	0
8	18959	Las Vegas	Las Vegas	Clark	NV	7		7	11-May	6181	0	0
9	6915	San Antonio	San Antonio	Bexar	TX	8		8	11-Jun	6181	0	0
10	54296	San Diego	San Diego	San Diego	CA	9		9	11-Jul	6181	0	0
11	38128	Dallas	Dallas-Fort Worth	Dallas	TX	10		10	11-Aug	6181	0	0
12	33839	San Jose	San Jose	Santa Clara	CA	11		11	11-Sep	6181	0	0
13	25290	Jacksonville	Jacksonville	Duval	FL	12		12	11-Oct	6181	0	0
14	20330	San Francisco	San Francisco	San Francisco	CA	13		13	11-Nov	6181	0	0
15	32149	Indianapolis	Indianapolis	Marion	IN	14		14	11-Dec	6181	1746	1.39
16	10221	Austin	Austin	Travis	TX	15		15	12-Jan	6181	1752	1.388
17	18172	Fort Worth	Dallas-Fort Worth	Tarrant	TX	16		16	12-Feb	6181	1764	1.392
18	17762	Detroit	Detroit	Wayne	MI	17		17	12-Mar	6181	1778	1.396
19	10920	Columbus	Columbus	Franklin	OH	18		18	12-Apr	6181	1792	1.402
20	32811	Memphis	Memphis	Shelby	TN	19		19	12-May	6181	1804	1.412
21	24043	Charlotte	Charlotte	Mecklenburg	NC	20		20	12-Jun	6181	1813	1.42
22	17933	El Paso	El Paso	El Paso	TX	21		21	12-Jul	6181	1814	1.424
23	44269	Boston	Boston	Suffolk	MA	22		22	12-Aug	6181	1810	1.422

Figure 4

The last normal form is third normal form (3NF). We split the 2NF table into 10 tables. The short snippet below in Figure 5 shows the many-to-many relationship between city and metro, which has been broken down through an intermediate table resulting in two 1-to-many relationships.

City Code	City name	Population Rank
3300	Aberdeen	7597
3301	Aberdeen	3325
3304	Absecon	4594
3305	Accokeek	5596
3310	Ada	1912
3312	Adairsville	3410
3315	Adams	12264
3319	Addison	1225
3320	Addison	11967
3322	Addy	12767
3333	Ahwahnee	9330
3334	Aiken	671
3340	Alamogordo	1232
3341	Albany	11247
3342	Albany	11016
3344	Albion	3562
3352	Alexandria	9718
3379	Alta	12600
3403	Andalusia	12713
3404	Anderson	407
3406	Andersonville	9967
3412	Angola	2992
3418	Anthem	1287
3419	Anthony	9650
3425	Apple Valley	905
3427	Aransas Pass	4519
3436	Ardmore	3395
3438	Arena	12959

City_Metro_code	Metro_code	City_code
1	M1	6181
2	M2	12447
3	M3	17426
4	M4	39051
5	M5	13271
6	M6	40326
7	M7	18959
8	M8	6915
9	M9	54296
10	M10	38128
11	M11	33839
12	M12	25290
13	M13	20330
14	M14	32149
15	M15	10221
16	M10	18172
17	M16	17762
18	M17	10920
19	M18	32811
20	M19	24043
21	M20	17933
22	M21	44269
23	M22	16037
24	M23	3523
25	M24	11093
26	M25	41568
27	M26	6118
28	M27	5976

Metro_code	Metro_name
M1	New York
M2	Los Angeles
M3	Chicago
M4	Houston
M5	Philadelphia
M6	Phoenix
M7	Las Vegas
M8	San Antonio
M9	San Diego
M10	Dallas-Fort Worth
M11	San Jose
M12	Jacksonville
M13	San Francisco
M14	Indianapolis
M15	Austin
M16	Detroit
M17	Columbus
M18	Memphis
M19	Charlotte
M20	El Paso
M21	Boston
M22	Seattle
M23	Baltimore
M24	Denver
M25	Washington
M26	Nashville
M27	Milwaukee
M28	Tucson

Figure 5