University of Ottawa
School of Electrical Engineering and Computer Science
CSI4142 Fundamentals of Data Science – Winter 2023
Project Phase 2: Conceptual Design – Dimensional Model

Instructor : Yazan Otoum, Ph.D.
TA: Paritosh Singh
Due Date: March 24 th, 2023 11:59 PM

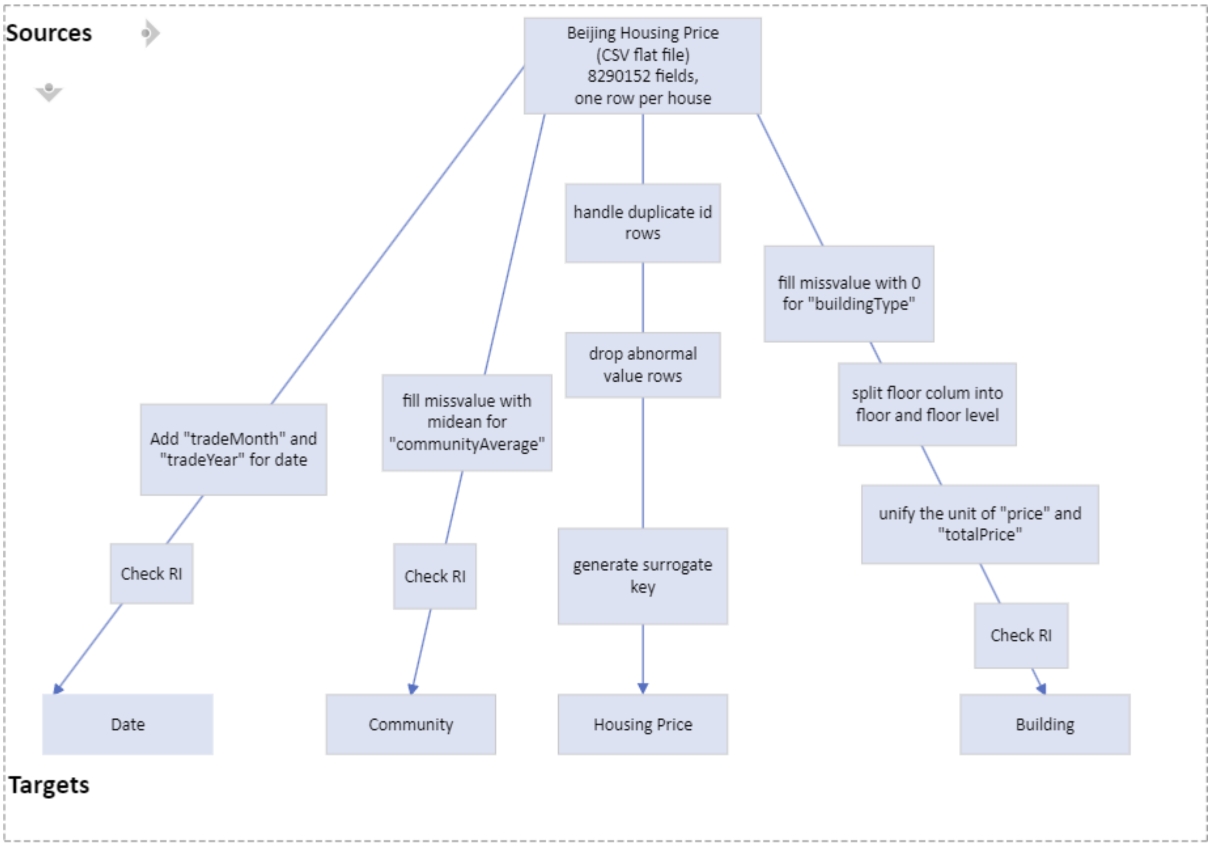
Group 1
Members:

Gary Gao	300124236
Yingqi Feng	300077437
Binxuan Wu	300142301

Source code of ETL process
https://colab.research.google.com/drive/1YxEvK4Qd4yHWbZUbHNjsd2ssXt5QfJ5C?usp=sharing

GitHub Repository:
https://github.com/BinxuanWu/CSI4142-Project

A. A one-page schematic with your high-level data staging plan.



Sources
Beijing Housing Price (CSV flat file) 8290152 fields, one row per house
handle duplicate id rows
drop abnormal value rows
generate surrogate key
fill missvalue with 0 for "buildingType"
split floor colum into floor and floor level
unify the unit of "price" and "totalPrice"
Check RI
fill missvalue with midean for "communityAverage"
Check RI
Add "tradeMonth" and "tradeYear" for date
Check RI
Date
Community
Housing Price
Building
Targets

B. A list of data quality issues you encountered and how you handled them (i.e., how did you detect and handle missing or noisy data (if any). How did you integrate the data from different sources etc.?

quality issues you encountered and how you handled them

1. loading the data into our DBMS.
   solution: after trying with phpmyadmin, we chose to use MySQLWorkbench, it has more visual and more efficient, also, faster to load.
2. dealing with messy data in the table.
   solution: we removed the rows that were completely messed up, corrected values that were in a different language, changed data types and split one column to multiple columns.

3. dealing with uppercase letters in the table
   solution: We replaced all uppercase letters into lowercase letter and added under slash between words in order for the database to function.
4. Exporting .sql file
   solution: since our database's name is capital letter, it causes trouble to export, so we migrated data into another database and successfully exported .sql file.

How did you integrate the data from different sources

Our data comes from one source so there is no trouble integrating the data from different sources.

C. Fill out the attached excel sheet (Team Planning) and include it in the PDF.
Appendix - Phase 2-Team Planning_W23

Deliverable checklist	Responsible	Expected completion date	Actual completion date	Estimated	Actual
	team member(s)			time (hours) to complete	time (hours) to complete
high level plan	Gary Gao, Yingqi Feng, Binxuan Wu	March 10 th	March 10 th	3 hrs	4.5 hrs
Create database instance	Gary Gao	March 14 th	March 14 th	2 hrs	3 hrs
Create <building> dimension	Yingqi Feng	March 17 th	March 17 th	0.5 hrs	1 hrs
Create <date> dimension	Yingqi Feng	March 17 th	March 17 th	0.5 hrs	1 hrs
Create <house_price> dimension	Binxuan Wu	March 17 th	March 18 th	0.5 hrs	0.5 hrs
Create <community> dimension	Gary Gao	March 17 th	March 18 th	0.5 hrs	0.5 hr
Staging of fact table	Binxuan Wu	March 18 th	March 18 th	0.5 hrs	0.5 hrs
Staging of dimension <commnunity>	Gary Gao	March 18 th	March 18 th	0.5 hrs	1 hr
Staging of dimension <building>	Yingqi Feng	March 18 th	March 18 th	0.5 hrs	1 hr

Staging of dimension <date>	Yingqi Feng	March 18 th	March 18 th	0.5 hrs	0.5 hrs
Surrogate key pipeline	Binxuan Wu	March 18 th	March 21 st	1 hr	1 hr
Staging of fact table – including FKs and measures	Gary Gao, Yingqi Feng, Binxuan Wu	March 21 st	March 21 st	2 hrs	2 hrs
Data quality handling and reporting	Gary Gao, Yingqi Feng, Binxuan Wu	March 21 st	March 24 th	1 hr	2 hr
Others – if any					