Database creation, table population, and business questions

1. Converting the dataset into tables

The dataset was taken from the kaggle: Zillow prize (Zestimate) which has ~1 Million rows and ~50 columns. This is a single dataset. Below are the approaches I followed to convert the dataset into tables

- Identified Core Tables: There were 6 major categories in which I can divide the whole dataset into. They were property, building, pool, yard, region, price
- Identified Lookup Tables: For some attributes in these core tables, lookup tables acted as catalogs, providing detailed descriptions. These are referred to as quality tables (Building quality, Heating System Quality, Pool Quality, Air Conditioning Quality). For example, the Building Quality table served as a lookup table for the Property table
- Identified Attributes and Allocated Them to Core Tables: Each table was carefully examined, and its data elements were mapped to the appropriate attributes. For instance, the Property table included columns such as latitude, longitude, lotsizesquarefeet, and yearbuilt
- Primary Key in Core Tables: The parcelid was designated as the primary key for all core tables
- Established Relationships: All core tables are related to each other via primary key - parcelid. Lookup tables have their own primary key depending on the attribute they are describing
- Data Cleanup and Sampling: The data was cleaned by removing irrelevant columns, such as neighborhood_id and construction_type_id, which did not align with any core table. Given the dataset's size of 1 million rows, I sampled 10% of the data (100,000 rows) to manage processing efficiently. I used a simple random sampling technique
- Structured the Data into Tables: With the tables, attributes, and relationships clearly defined, the data was organized into structured database tables, ensuring normalization

2. Challenges faced during importing of data

Firstly I have created a Database - Zillow. The tables were stored in my local folder in CSV format. I used the Table Data Import Wizard in MySQL Workbench to load all the tables into the Zillow database. Given the amount of rows, it took around 2 mins for each table to import. Some of the challenges faced while importing:

- Inflated values for latitude and longitude: In the Property table, each value of latitude and longitude were populated with e+06 factor multiplied. Corrected them to original values.
- Defining data types: Some of the data types were incorrectly specified for the following columns. I have changed the datatype.
    - Year Built: Changed from double to int
    - Assessment Year - Changed from double to int

- Handling null values: In some of the catalog tables, the description was not given if a particular type of amenity is not present. I have converted all those Null values to below values:
    - Pool Quality Table - Changed null value to ‘no pool’
    - Heating System Quality Table - Changed null value to ‘no heating system’
    - Air Conditioning Quality Table - Changed null value to ‘no air conditioning’

3. Data dictionary

The database - Zillow has 10 tables and 36 unique columns in total. Each core table contains 100000 rows. Below table provides data dictionary for all 10 tables:

Column Name	Data Type	Key Type	Description	Min Value	Max Value
Property Table - property_table					
parcelid	int	PK	Unique identifier for each property	10711770	162960800
latitude	double	-	Geographic latitude coordinate	33.3396	34.81965
longitude	double	-	Geographic longitude coordinate	-119.4739	-117.572
lotsizesquarefeet	double	-	Total lot size in square feet	283	6971010
yearbuilt	int	-	Year the property was constructed	1824	2016
roomcnt	double	-	Total number of rooms	0	16
unitcnt	double	-	Number of units in the property	1	4
calculatedfinishedsquarefeet	double	-	Total finished living area in square feet	1	35360
buildingqualitytypeid	int	FK	Foreign key linking to Building Quality	1	12
propertylandusetypeid	int	-	Identifier for property land use type	31	275
Building Table - building_table					
parcelid	int	PK/FK	Property identifier	10711770	162960800
bathroomcnt	double	-	Number of bathrooms	1	16
bedroomcnt	double	-	Number of bedrooms	0	18
fullbathcnt	double	-	Number of full bathrooms	1	16
calculatedbathnbr	double	-	Calculated total number of bathrooms	1	16
threequarterbathnbr	double	-	Number of 3/4 bathrooms	0	5
fireplacecnt	double	-	Number of fireplaces	0	6
garagecarcnt	double	-	Garage capacity in number of cars	0	19
garagetotalsqft	double	-	Total garage area in square feet	0	5998
airconditioningtypeid	int	FK	Foreign key linking to Air Conditioning	1	13
heatingorsystemtypeid	int	FK	Foreign key linking to Heating System	1	20
decktypeid	int	-	Identifier for deck type	0	66

Building Quality Table - building_quality_table					
buildingqualitytypeid	int	PK	Unique identifier for building quality	1	12
quality_description	text	-	Description of building quality level	N/A	N/A
Air Conditioning Quality Table - air_conditioning_quality_table					
airconditioningtypeid	int	PK	Unique identifier for AC type	1	13
ac_type	text	-	Description of air conditioning system	N/A	N/A
Heating System Quality Table - heating_system_quality_table					
heatingsystemtypeid	int	PK	Unique identifier for heating system	1	25
hs_type	text	-	Description of heating system type	N/A	N/A
Pool Table - pool_table					
parcelid	int	PK/FK	Property identifier	10711770	162960800
poolcnt	double	-	Number of pools	0	1
poolsizesum	double	-	Total pool size in square feet	0	2576
pooltypeid	int	FK	Foreign key linking to Pool Quality	0	10
Pool Quality Table - pool_quality_table					
pooltypeid	int	PK	Unique identifier for pool type	0	10
pool_type_description	text	-	Description of pool type	N/A	N/A
Region Table - region_table					
parcelid	int	PK/FK	Property identifier	10711770	162960800
regionidcity	double	-	City region identifier	3491	396556
regionidcounty	double	-	County region identifier	1286	3101
regionidzip	double	-	ZIP code region identifier	95982	399675
Price Table - price_table					
parcelid	int	PK/FK	Property identifier	10711770	162960800
taxvaluedollarcnt	double	-	Total property value in dollars	7310	56M
structuretaxvaluedollarcnt	double	-	Cost of construction in dollars	0	17M
landtaxvaluedollarcnt	double	-	Value of the land in dollars	59.5	49M
taxamount	double	-	Total property tax amount	20.24	651677
assessmentyear	int	-	Year of tax assessment	2015	2016
Yard Table - yard_table					
parcelid	int	PK/FK	Property identifier	10711770	162960800
yardbuildingsqft17	double	-	Size of yard building type 17 in sqft	0	3080
yardbuildingsqft26	double	-	Size of yard building type 26 in sqft	0	1144

4. Business Questions

1. Give counts of following for the properties for each region, to understand which regions prefer pools and fireplaces
    a. Pool Count
    b. Fireplace Count
2. List Average Property Value for each city, to assess the living expenses
3. List Average Property Size for each region, what is the trend looking like?
4. Find Average price for different property types and find premium property types
5. Average Property Tax Rate by County
6. What is the average price per square foot for properties by year built (for properties built after 2000)
7. List most popular Heating System Type and Air condition quality type
8. What are the distribution statistics of bedroom and bathroom counts in the building table (Min, Q1, Q2, Q3, Max)
9. Find number of properties and their average price in the top 5% of their city by price
10. Calculate year-over-year price changes for any two cities in california. Consider years from 2000 to 2015
11. What is the percentage share of properties in each property type for the entire dataset, and how does it rank among other types?