# SQL questions - regression

(Use sub queries or views wherever necessary)

1. Create a database called `house\_price\_regression`. NO

2. Create a table `house\_price\_data` with the same columns as given in the csv file. Please make sure you use the correct data types for the columns. You can find the names of the headers for the table in the `regression\_data.xls` file. Use the same column names as the names in the excel file. Please make sure you use the correct data types for each of the columns.

3. Import the data from the csv file into the table. Before you import the data into the empty table, make sure that you have deleted the headers from the csv file. (in this case we have already deleted the header names from the csv files). To not modify the original data, if you want you can create a copy of the csv file as well. Note you might have to use the following queries to give permission to SQL to import data from csv files in bulk:

```sql

SHOW VARIABLES LIKE 'local\_infile'; -- This query would show you the status of the variable ‘local\_infile’. If it is off, use the next command, otherwise you should be good to go

SET GLOBAL local\_infile = 1;NO

```

4. Select all the data from table `house\_price\_data` to check if the data was imported correctly

5. Use the alter table command to drop the column `date` from the database, as we would not use it in the analysis with SQL. Select all the data from the table to verify if the command worked. Limit your returned results to 10.

6. Use sql query to find how many rows of data you have.PYTHON

7. Now we will try to find the unique values in some of the categorical columns:

- What are the unique values in the column `bedrooms`?

- What are the unique values in the column `bathrooms`?

- What are the unique values in the column `floors`?

- What are the unique values in the column `condition`?

- What are the unique values in the column `grade`?

8. Arrange the data in a decreasing order by the price of the house. Return only the IDs of the top 10 most expensive houses in your data.

9. What is the average price of all the properties in your data?

10. In this exercise we will use simple group by to check the properties of some of the categorical variables in our data

- What is the average price of the houses grouped by bedrooms? The returned result should have only two columns, bedrooms and Average of the prices. Use an alias to change the name of the second column.

- What is the average `sqft\_living` of the houses grouped by bedrooms? The returned result should have only two columns, bedrooms and Average of the `sqft\_living`. Use an alias to change the name of the second column.

- What is the average price of the houses with a waterfront and without a waterfront? The returned result should have only two columns, waterfront and `Average` of the prices. Use an alias to change the name of the second column.

- Is there any correlation between the columns `condition` and `grade`? You can analyse this by grouping the data by one of the variables and then aggregating the results of the other column. Visually check if there is a positive correlation or negative correlation or no correlation between the variables.

You might also have to check the number of houses in each category (ie number of houses for a given `condition`) to assess if that category is well represented in the dataset to include it in your analysis. For eg. If the category is under-represented as compared to other categories, ignore that category in this analysis

11. One of the customers is only interested in the following houses:

- Number of bedrooms either 3 or 4

- Bathrooms more than 3

- One Floor

- No waterfront

- Condition should be 3 at least

- Grade should be 5 at least

- Price less than 300000

For the rest of the things, they are not too concerned. Write a simple query to find what are the options available for them?

12. Your manager wants to find out the list of properties whose prices are twice more than the average of all the properties in the database. Write a query to show them the list of such properties. You might need to use a sub query for this problem.

13. Since this is something that the senior management is regularly interested in, create a view called `Houses\_with\_higher\_than\_double\_average\_price` of the same query.NO

14. Most customers are interested in properties with three or four bedrooms. What is the difference in average prices of the properties with three and four bedrooms? In this case you can simply use a group by to check the prices for those particular houses

15. What are the different locations where properties are available in your database? (distinct zip codes)

16. Show the list of all the properties that were renovated.

17. Provide the details of the property that is the 11th most expensive property in your database.