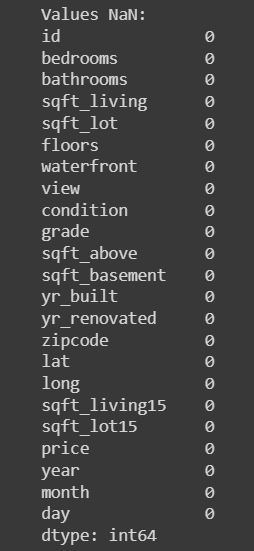
**Housing Regression - Mini Project**

**EDA & ML Models**

**Data Cleaning**

On our first approach to the data que saw that we don't need so much data cleaning. We did the usual approach to the data, with the info, columns, categorical and numerical values to have an idea on how the data is stored and managed.

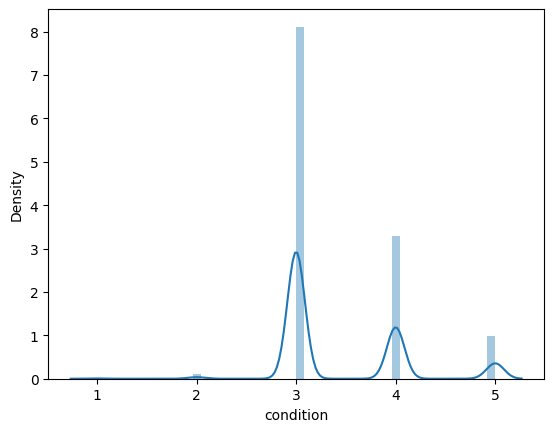
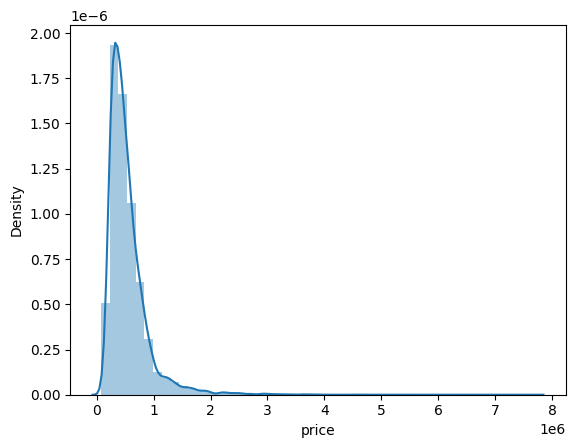


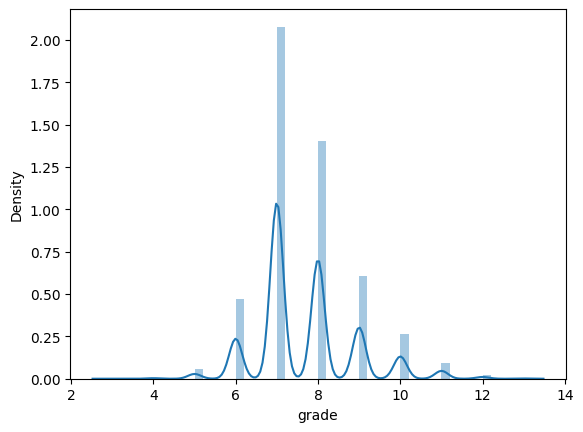
The data is well organised, and we notice that we don't have nulls and duplicates.

So, we proceed to prepare all for the EDA with some plots.

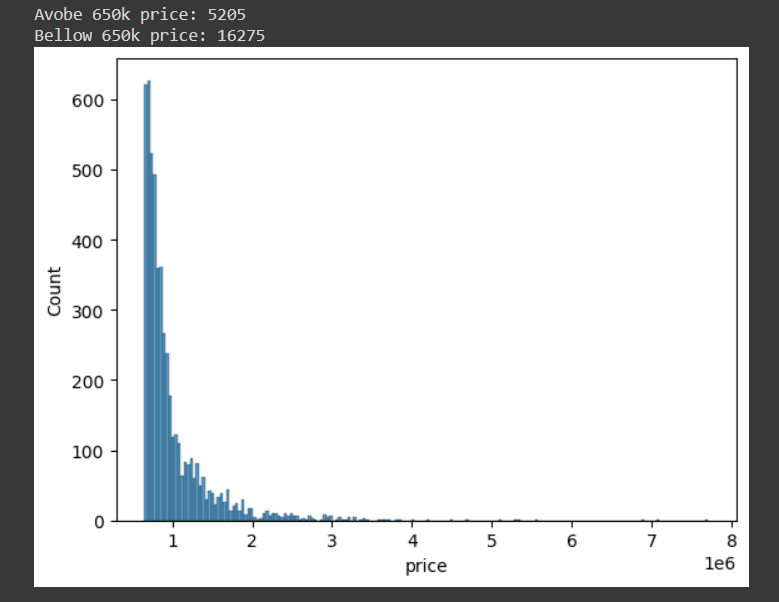
**Plots**

Some of the early plots we did like this gave us some insights like the density related with the price, condition and grade. Values we considered important for their relation with the price.

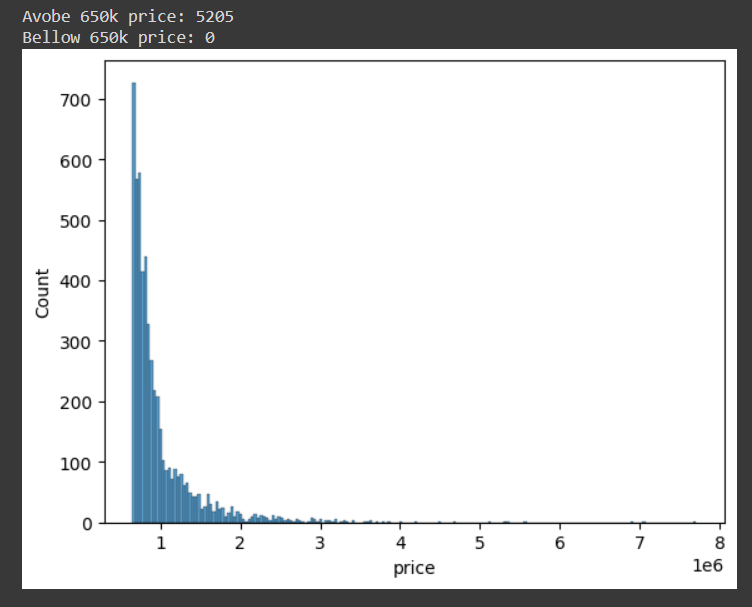




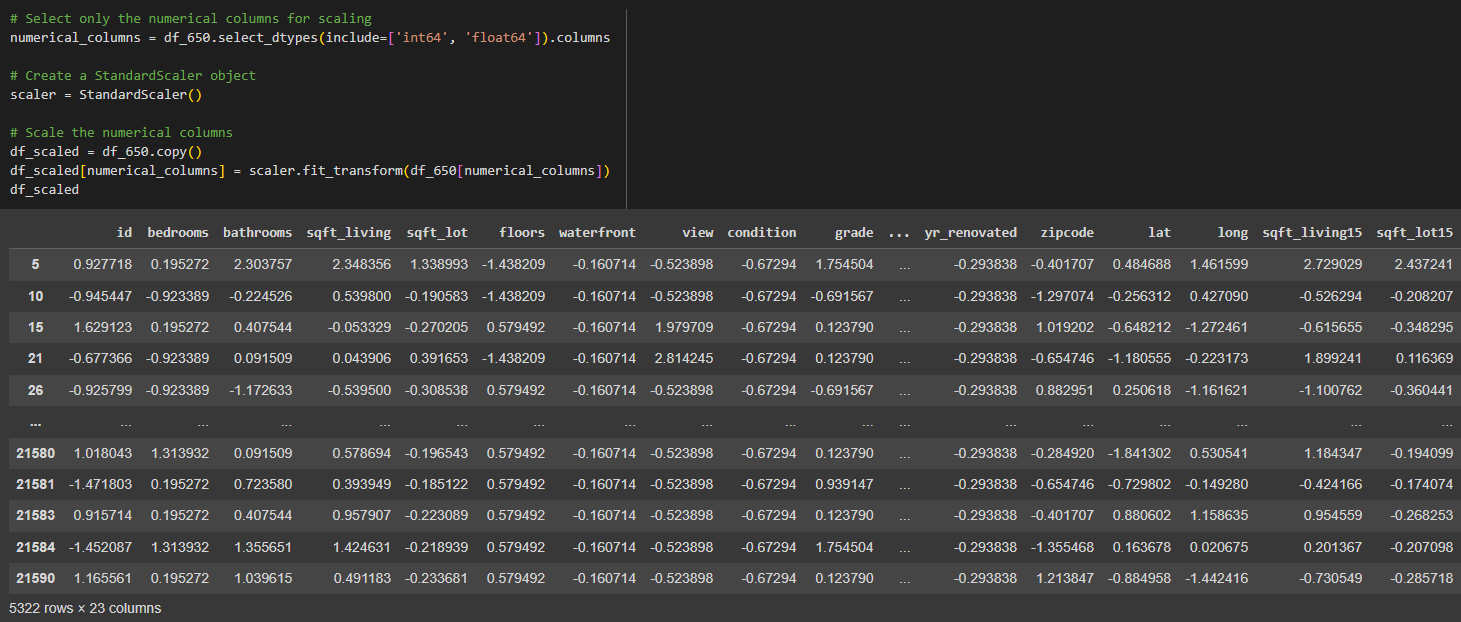
One of our principal concerns was to determine the amount of houses with the price above 650k. We created a code to determine how many were above and below that price and show it with a plot:

****

After that we determine that we could work with only the amount above 650k for our training model as our client requested:

****

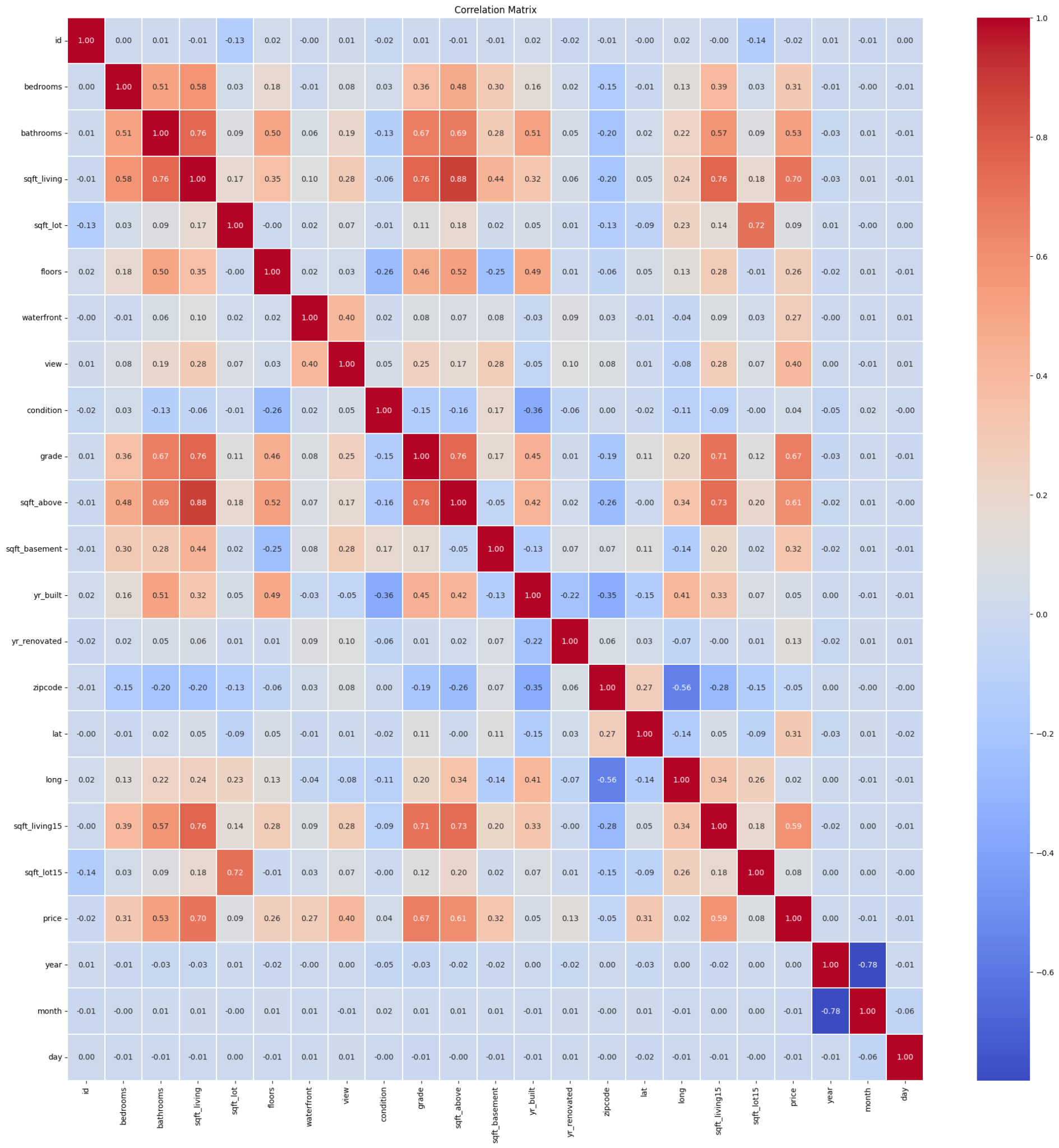
During our exploration of the data, we proceed to scale the categorical and numerical columns:



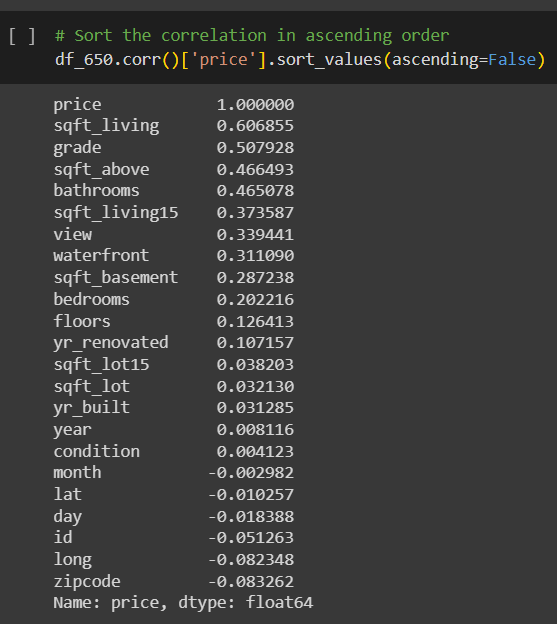
**Machine Learning**

As usual we started creating the heat map for the clean data. Followed by our first linear regression and benchmarking the data before the training.

We compared the price value with several other values to have an idea on how important it was for our client and determine some insights of the housing problems we could have.

****

We sorted also the correlation between the values too:

****

We build a machine learning model to predict the selling prices of houses based on a variety of features.

We extracted the features that most influenced the prediction of house prices, based on the same features with our target variable (price). We created our linear regression model to have a good understanding of it. The purpose of this machine learning model is to predict the price of houses by evaluating the metrics between Mean Squared Error (MSE) and R-squared (R²):

Where the MeanSquaredError(MSE): Indicates the average squared difference between predicted and actual values. Our model's MSE is $61,254,344,139.52, reflecting the accuracy of predictions. And the R-squared (R²): A measure of how well the model explains the variance in the target variable. Our model has an R² of 0.53, indicating moderate predictive power.

A number with black text

Description automatically generated with medium confidence

Then we highlight the features that clearly contributes to the house prices of 650k and above.

Key features.

* **Waterfront:** Strong positive influence with a coefficient of $855,195.76.
* **Condition:** Positive impact with a coefficient of $80,527.38.
* **Bedrooms:** Positive impact with a coefficient of $4,475.77.
* **Grade:** Significant positive influence with a coefficient of $138,214.37.
* **Sqft\_Above:** Positive impact with a coefficient of $79.35.
* **Sqft\_Living15:** Positive impact with a coefficient of $77.30.
* **Sqft\_Lot15:** Negative influence with a coefficient of -$0.57

A screenshot of a computer

Description automatically generated

Scatter Plot for our Evaluation model.

A graph with a line and a dotted line

Description automatically generated

* Each point on the plot represents an individual house property.
* The diagonal red line represents a perfect prediction scenario where actual price and predicted prices are equal.
* Points above the red line indicate overestimation, while points below the line indicate a massive error(underestimation).
* Wider dispersion from the line indicates the areas where the model struggles to predict accurately.

Predicted Prices for Testing Set:

[ $91,241.98 $244,942.26 $348,952.01 ..$372,512.91

$421,628.05 $358,472.98]

**SQL Queries**

On our project we started to create the queries necessary for our client.

During the resolution of the queries, we created tables, learnt how to produce and explore the data and show some insights for the project.

CREATE TABLE house\_price\_data (

`id` bigint DEFAULT NULL,

`date` text,

`bedrooms` int DEFAULT NULL,

`bathrooms` text,

`sqft\_living` int DEFAULT NULL,

`sqft\_lot` int DEFAULT NULL,

`floors` int DEFAULT NULL,

`waterfront` int DEFAULT NULL,

`view` int DEFAULT NULL,

`condition` int DEFAULT NULL,

`grade` int DEFAULT NULL,

`sqft\_above` int DEFAULT NULL,

`sqft\_basement` int DEFAULT NULL,

`yr\_built` int DEFAULT NULL,

`yr\_renovated` int DEFAULT NULL,

`zipcode` int DEFAULT NULL,

`lat` text,

`long` text,

`sqft\_living15` int DEFAULT NULL,

`sqft\_lot15` int DEFAULT NULL,

`price` int DEFAULT NULL

) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4\_0900\_ai\_ci.

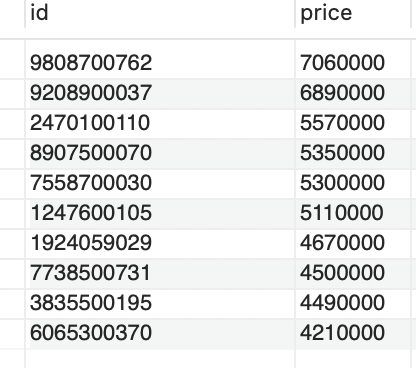
With this query we found detailed data of the 10 most expensive property in the database

***SELECT id, price.***

***FROM house\_price\_data***

***ORDER BY price DESC***

***LIMIT 10;***

******

We calculated the average price of all the properties in the database.

***SELECT AVG(price) AS average\_price***

***FROM house\_price\_data;***

***A screenshot of a phone

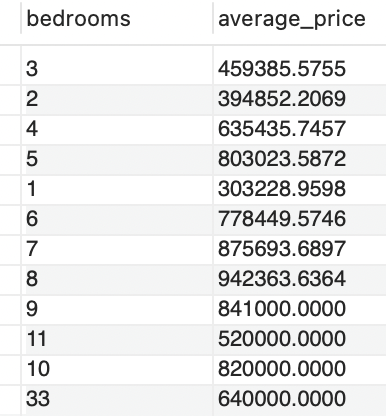
Description automatically generated***

We also calculated what is the average price of the houses grouped by bedrooms?

***SELECT bedrooms, AVG(price) AS average\_price***

***FROM house\_price\_data***

***GROUP BY bedrooms;***

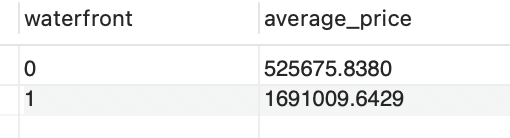
******

What is the average price of the houses with a waterfront and without a waterfront?

***SELECT waterfront, AVG (price) AS average price***

***FROM house\_price\_data***

***GROUP BY waterfront;***



**Tableau**

