Oriol Gallego Vazquez: 229408
Cinta Arnau Arasa: 231629

upf. Universitat Pompeu Fabra Barcelona

# Report final project Visual Analytics

**Introduction:**
For the final project of visual analytics we decided to analyze the house pricing in Seattle given a dataset we found online. The goal will be given the characteristics of the dataset, we will determine whether a house is more or less expensive than the median. Then we will use **Shap values** to understand the weight of each variable in the predictions, **tableau** to visualize the dataset and behaviors and **geopandas** to visualize the data in a map and help visualize where the houses are located in real life and the correlation between location and price.

**Data cleaning:**
This dataset has 5529 entries and none of them are null. We selected some columns to drop because they had confusing names and some because they were redundant. Then we saved a copy of the dataset before removing the spatial information to work with predictions, this copy will be used in the geolocation section.

**Dataset Exploratory Analisis:**
Most of the variables of our dataset have self explanatory names (price, sqft_living, bedrooms) the only strange thing from the dataset is that some variables that should be integer are floats (bathrooms, floors) we assumed that the decimals represent not completed forms of that variables (a bathroom without a skin, a balcony which is almost a floor). To understand each variable we used describe() which gives us the quartiles, min, max and standard deviation.
Next we plotted the distribution of the 3 most important variables: 'floors', 'sqft_living' and 'price'. **[figures 1 to 3].** The insight gained is that all distributions are similar, floors is the least similar distribution because it has only 6 possible values, but the rest follow the normal distribution.
Next we split the dataset into two, those with grade>7 and those with grade<=7 and compared the same distributions for each splitted dataset. **[figures 4 to 6]** We got no surprises from these histograms, for houses with low grade it is more common to have cheaper and smaller living rooms than houses with high grade.
For the next section we are analyzing the boxplots of the same 3 variables, and we can see how most of them look different, in the case of boxplots for floor we see that high grade has no outliers while all the others have some. Most notably sqft_living with high grades has the most outliers, with values at 10000 when the median is a bit higher than 2000, and quartiles at 3000 and 2000.
The last exploratory analysis we did was the correlation matrix **[figure 7],** the output that the matrix presents us fits into what we expected, we see how sqft_living and sqft_above are highly correlated and the number of floors with the condition are not correlated almost (which also tells us that big houses don't really imply worse condition because they take longer to clean).

**Modeling:**
For modeling our dataset, as we explained in the introduction, we are going to create a target based on the house price. To do this we replace the price column for a new price column which will be our target and will be 0 or 1 depending if the price of the house is greater or less than the median. Now that we have our target and data frame ready, we create our x and y test/train. and fit it into an XGBClassifier. We obtained some results, and to analyze them we built a confusion matrix **[figure 8]** . We have great precision as most of the predictions are true positives and negatives and almost no false negatives. But we have a 10% of the predicted values that are false positives, meaning that we predict that some houses should be more expensive than the median, and they are not in reality, this is good to find good prices for houses. On the other hand having almost no false negatives means that there are almost no overpriced houses which means safety when buying a house, as it is less likely that you are paying more than what it should cost.
To visualize better this we created the density chart **[figure 9]**. The output density char is very standart as most of the density is accumulated in the respective class, class 0 has more density at 0 and vice versa. If you wanted to make a cutoff for the model, it would depend on whether you are the buyer or the seller, if you want to sell the house, a smaller cutoff would make it more probable that the model classifies your house as more expensive than the median. If you wanted to buy the house, a higher cutoff would make it more probable that the model classifies the house you want to buy as less expensive than the median. After this explanation, we would like to remember

that we created this target by choosing a variable and splitting it into above the median and below, which makes the intersections of densities be near 0.5 and it is the case for our density chart.

**Explainability and interpretability using SHAP:**

For this section we used the Shap library that was introduced to us in lab 6. We analyzed 2 predictions of houses with index 0 and 4, and saw how each feature affects our model's prediction. Then we created the summary plot which gives us insight about which Shap values are more common for each variable, and how does the original value affect its Shap value **[figure 10]**. We can see how high grades imply high Shap values, and surprisingly, newly built houses imply low Shap values. Next we can see in **[figure 11]** the bar plot which tells us the most important variables to our model, in this case we have that the top 3 most important are grade, year built and square footage of the living room. We expected this output more or less, as in real life these are the things you look for in a house. Next we plotted some scatter plots to compare the shap value and the original value of years built and square foot living **[figures 12 and 13]**. Let's start analyzing the years built scatter plot, we see how its shap value decreases as the year built increases which makes sense with the summary plot, as we saw that they are negatively correlated. The opposite happens with square foot living, which is more akin to the human mind because a bigger house means a more expensive house.

**Spatial visualization using GeoPandas**

For the spatial visualization part of our project we decided to use GeoPandas and find a dataset that contains latitude and longitude variables, and the one we selected does have them. First we searched where the houses were located from the coordinates, and discovered that our dataset has information about house pricing in Seattle, Washington. The next thing we did was load the buildings at that location with osmnx and filter the houses that are contained in the geodataframe provided by osmnx, this turned out to be a deception to us because there were only 30 matches in the beginning, but we decided to scale up the buildings by 1.8 to make it more likely to have more matches, but we only increased the number of matches to 56 which is not significant enough, to show this we show the map figure with the labels that we filtered. Those labels are the following: Big recent house (higher than median sqft_living and newer than 1990), Small recent house, Big old house,Small old house, Building footprints(the buildings that didn't match with our houses). In the first figure we see how the map is overwhelmed by Building footprints which we are not interested in, in the second map we filtered out the Building footprints but the result shows almost no houses. Finally we decided to use scatter mapbox to scatter the houses from our original data without using osmnx and set the color and size to the price value. We see a much better output with lots of points representing houses, and those points go from green(cheap) to red(expensive).

**Streamlit:**

There is an explanatory video in the submission and insights section in the streamlit app.

**TABLEAU**

**DASHBOARD 1: Price per Location and Year based on Condition**

We believe it is important to check the condition (and given grade) of the house when considering its price, as, for example, no one would want to buy a really expensive house if its condition is really bad. Furthermore, we also want to take into account the year the house was built, as it is 'common knowledge' that older houses tend to have worse conditions and, therefore, their prices are lower.

If we take a look at *Figure 14*, we can see how the condition of a house and the grade given to it are positively correlated, meaning that the worse the condition of the house, the lower the grade it has been given. Also, older houses have worse conditions and, hence, lower grades than newer houses.

Furthermore, if we take a look at both *Figure 15* and *Figure 16*, it can be seen that all the newer houses, which have better average grades and conditions, are on the right side of Seattle, Washington. On the other hand, all the houses with worse conditions and, hence, lower grades, are all on the left side of Seattle.

Finally, we would like to make a few comments about the *Average Price per Year* graph (*Figure 17)*. As we had previously mentioned, *it is 'common knowledge' that older houses tend to have worse conditions and, therefore,*

*their prices are lower*. However, in this graph we can see that this does not exactly apply. It only applies for those houses that are in the interval of Year Built [1941, 2015], as the houses that were built between 1941 and 1981 are all under the average price, meaning that they have a low price, while the newer houses - built between 1982 and 2015 - mostly have a high price which is over the average. Nevertheless, we have an exception with those houses that were built between 1900 and 1940, as they are very old but have a high price which is over the average. We have thought about the possible reason for this and we have come up with the conclusion that older houses tend to be 'worn' and deteriorated, which means that they have bad conditions and are in need of being 'renovated'.

*"Older homes are made of older materials, so it follows that the aging construction in these homes would come with a need for frequent maintenance. From faulty plumbing to sloping floors, there's no shortage of projects to do in existing homes – and these projects don't come cheap."*

Therefore, we have guessed that those older houses that were built in the interval [1900, 1940] have most probably been 'renovated', which implies an increment in its price, hence its high price over the average.

**DASHBOARD 2: Bedrooms, Bathrooms & Floors Based on Square foot Living**

We believe it is important to know how many bedrooms, bathrooms and floors a house has based on its square foot living. The reason for this is because maybe there is a family with lots of children that cannot afford to buy a big house (large amount of square foot living) but needs a house with many bedrooms, while on the other hand there might be a couple that wants a big house but with just the needed bedrooms (i.e., only 2).

As a curiosity, one might think that the bigger the house, the more bedrooms it has, but if we check the graph *Number of Bedrooms depending on Square foot Living* (*Figure 17*), it can be seen that the houses with higher average square foot (adding up the living, basement and above square foot) - 7.948 square foot - have in general a total of 7 bedrooms, while those that have 9 bedrooms have instead a total average of 7.550 square foot, which is not much of a difference however. On the other hand, it is quite true that the smaller the house, the less bedrooms it has, as, for example, the houses with lower average square foot (1.769) have only 1 bedroom.

Similarly, as it could have also been initially guessed, it can be seen in *Figure 19* and *20* how, the larger the house, the more floors and bathrooms it has, and vice versa.

Nonetheless, not always having a large house implies having lots of bedrooms, bathrooms and floors. For example, in *Figure 21* we can see an example of a large house with 7 bedrooms that has only 1 bathroom and 1 floor.

On the other hand, we have in *Figure 22* a perfect example of a small house with lots of bedrooms, which could be perfect for the *"family with lots of children that cannot afford to buy a big house (large amount of square foot living) but needs a house with many bedrooms"*, while in *Figure 23* we have a perfect example of the opposite situation *"a couple that wants a big house but with just the needed bedrooms (i.e., only 2)"*.

Oriol Gallego Vazquez: 229408
Cinta Arnau Arasa: 231629

Universitat Pompeu Fabra Barcelona

# ANNEXES

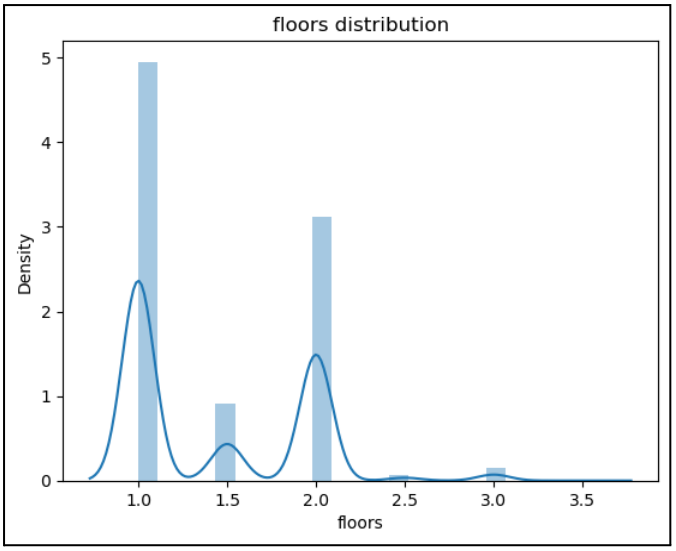


**Figure 1.** Floors distribution (House Price).

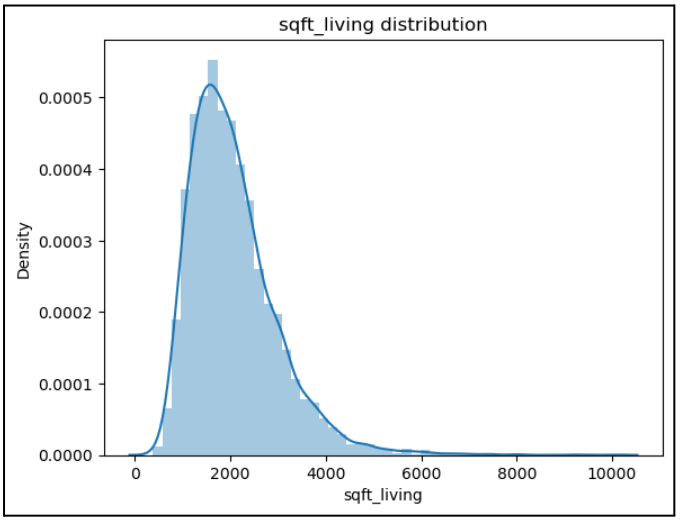


**Figure 2.** Square foot Living distribution (House Price).

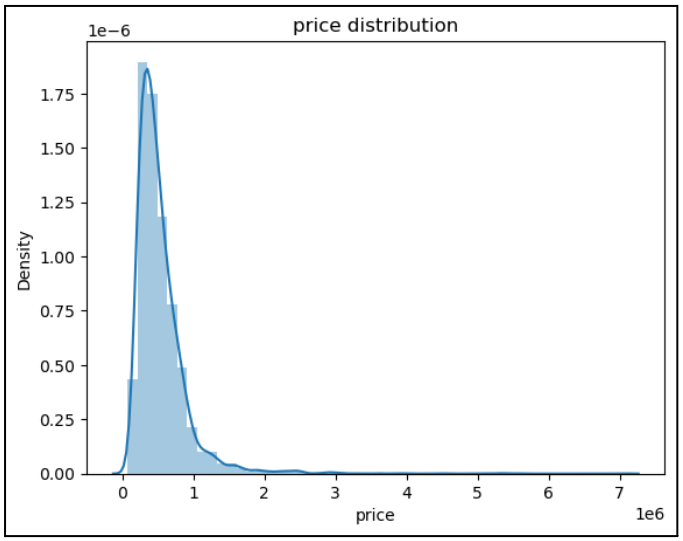


**Figure 3.** Price distribution (House Price).

Oriol Gallego Vazquez: 229408
Cinta Arnau Arasa: 231629

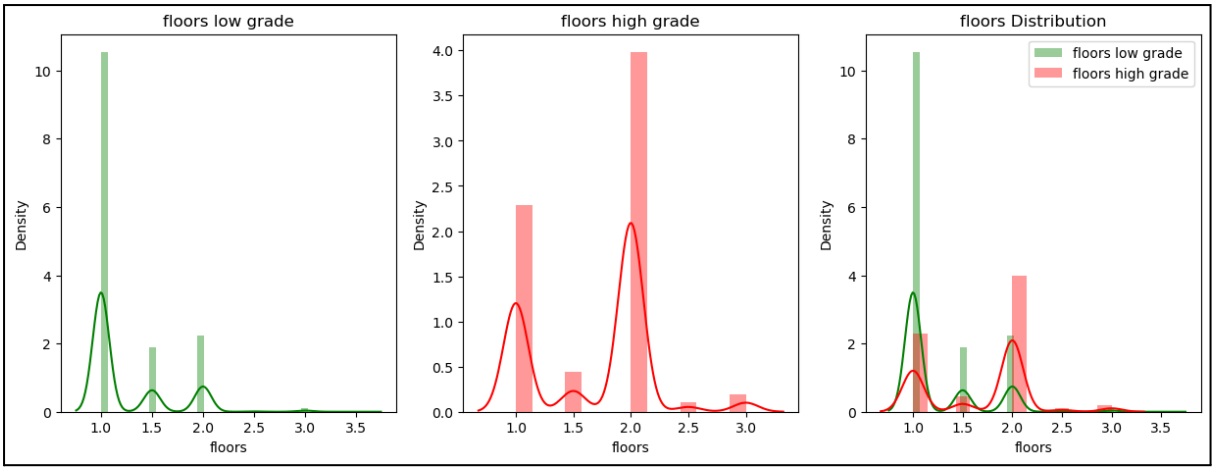


**Figure 4.** Comparison of number of floors distribution between high and low graded houses.

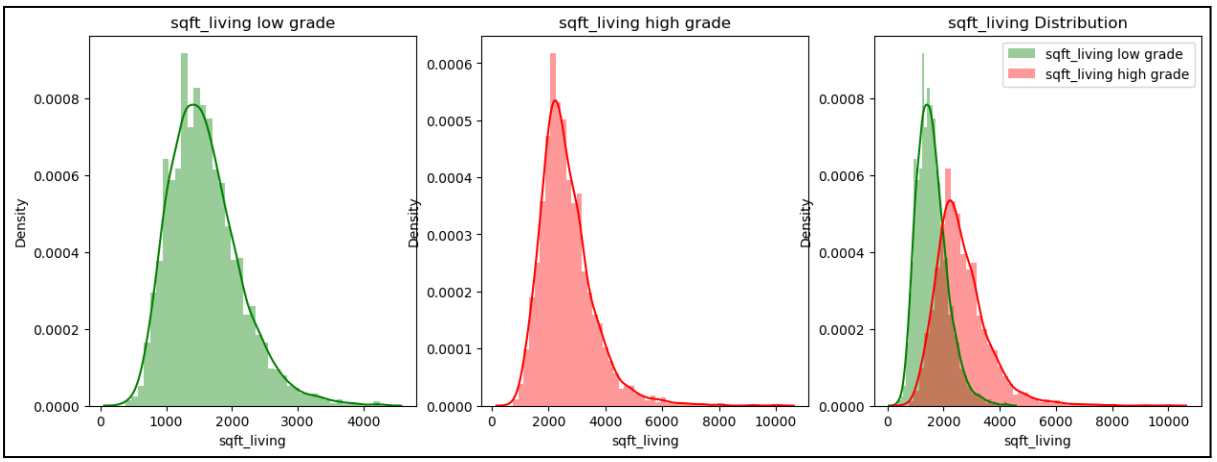


**Figure 5.** Comparison of square footage distribution between high and low graded houses.

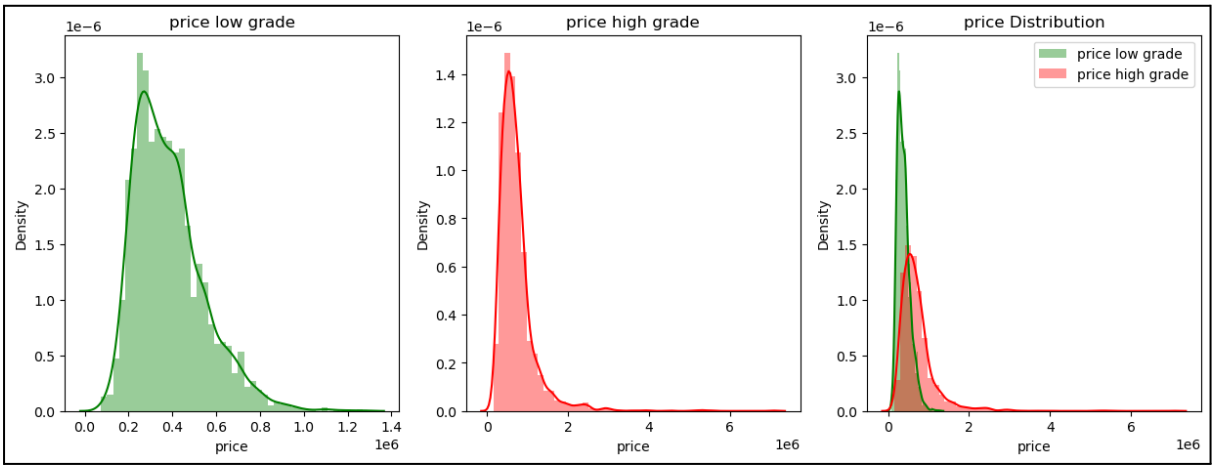


**Figure 6.** Comparison of price distribution between high and low graded houses.

Oriol Gallego Vazquez: 229408
Cinta Arnau Arasa: 231629



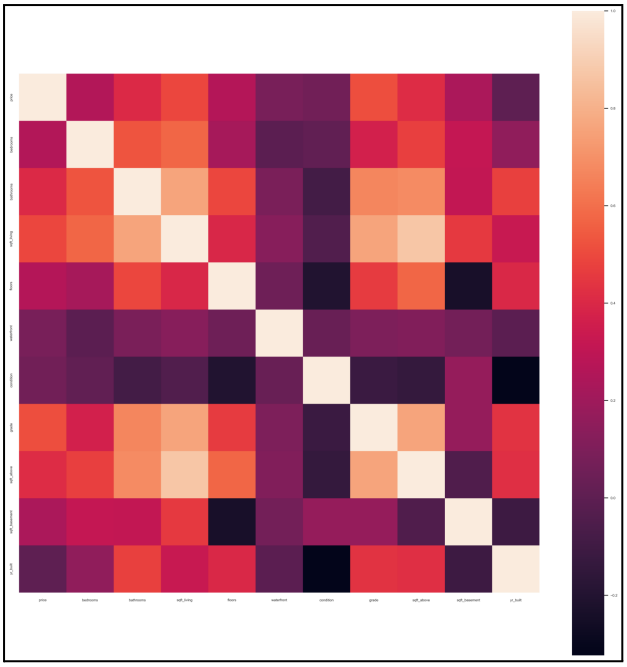

**Figure 7.** Correlation matrix.

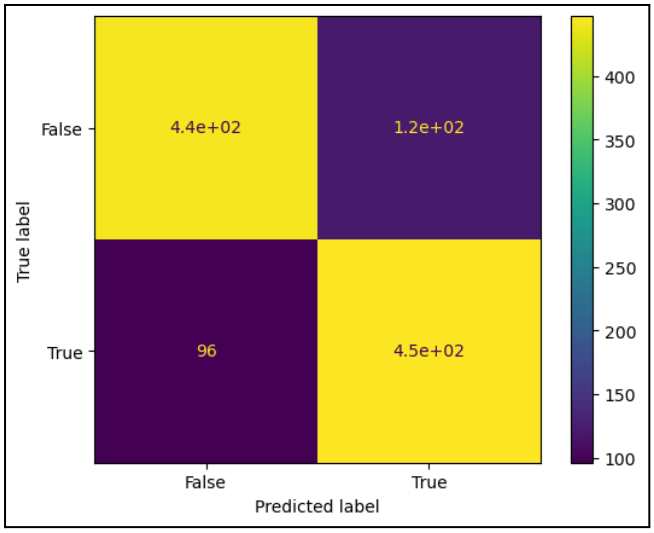


**Figure 8.** Confusion matrix.

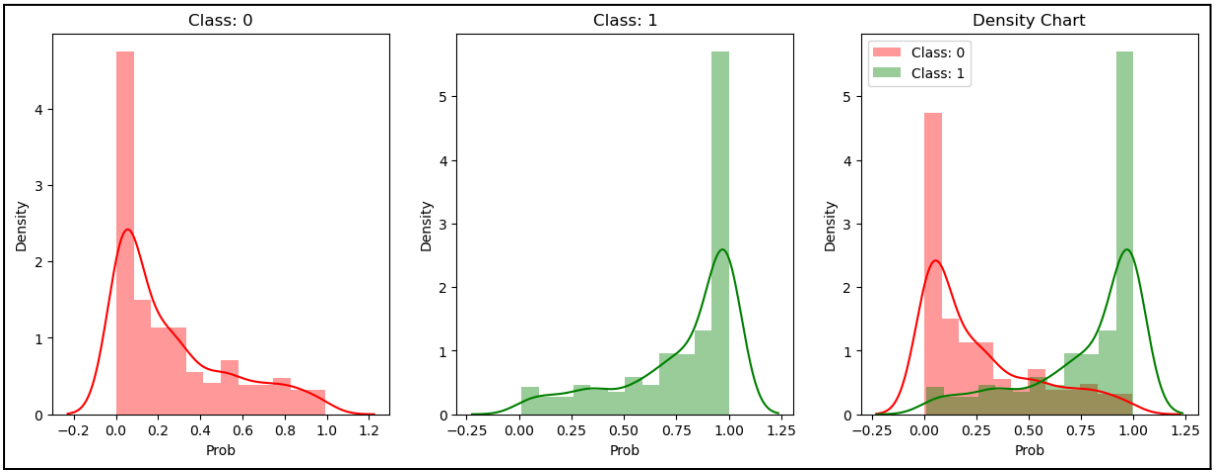


**Figure 9.** Density chart.

Oriol Gallego Vazquez: 229408
Cinta Arnau Arasa: 231629

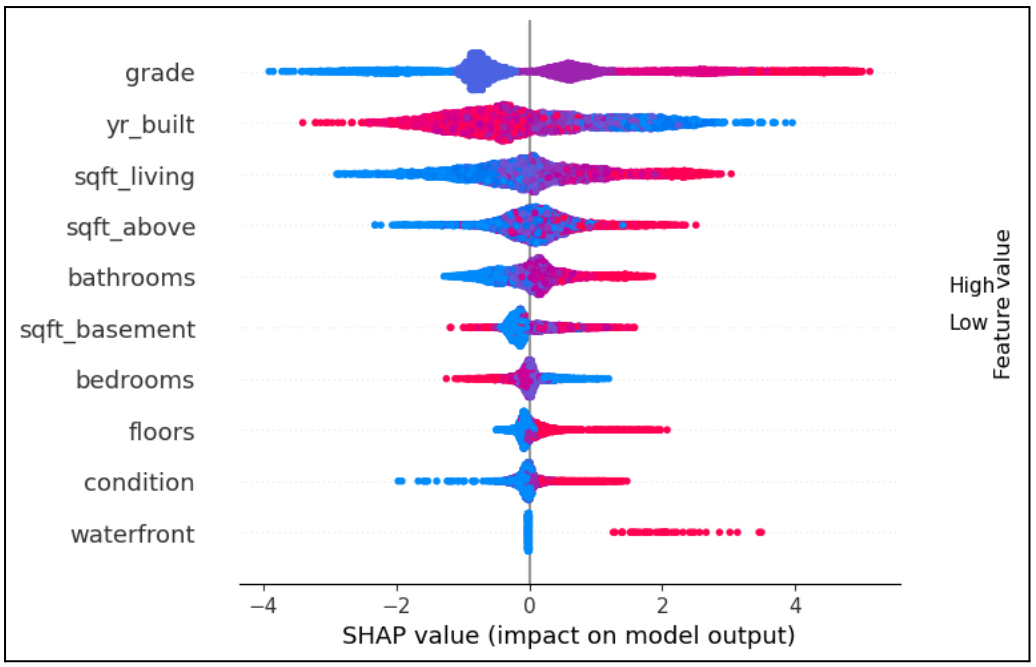


**Figure 10.** Summary plot.

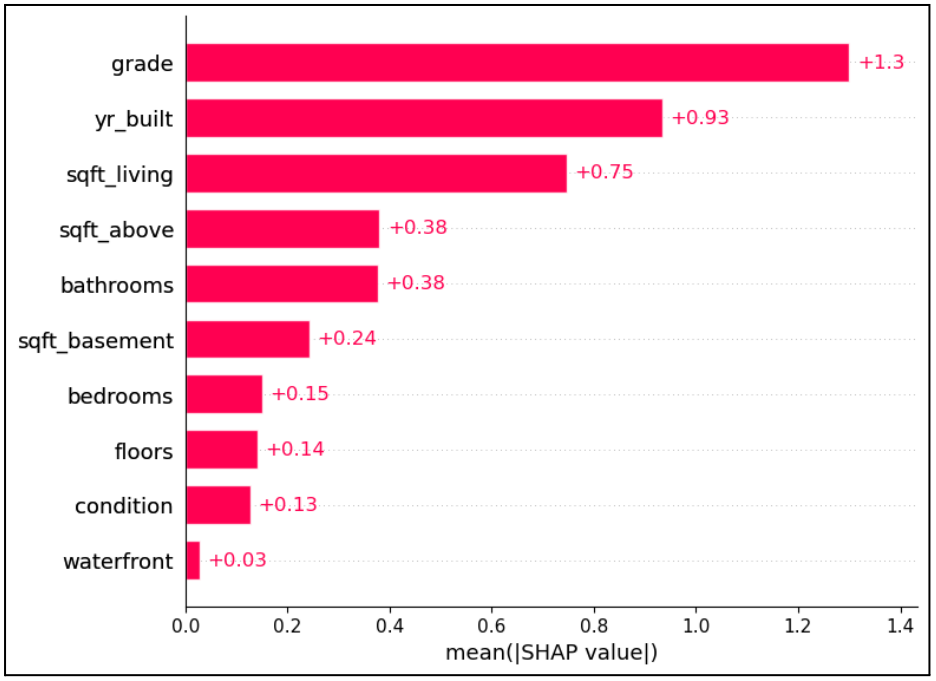


**Figure 11.** Bar plot.

Oriol Gallego Vazquez: 229408
Cinta Arnau Arasa: 231629



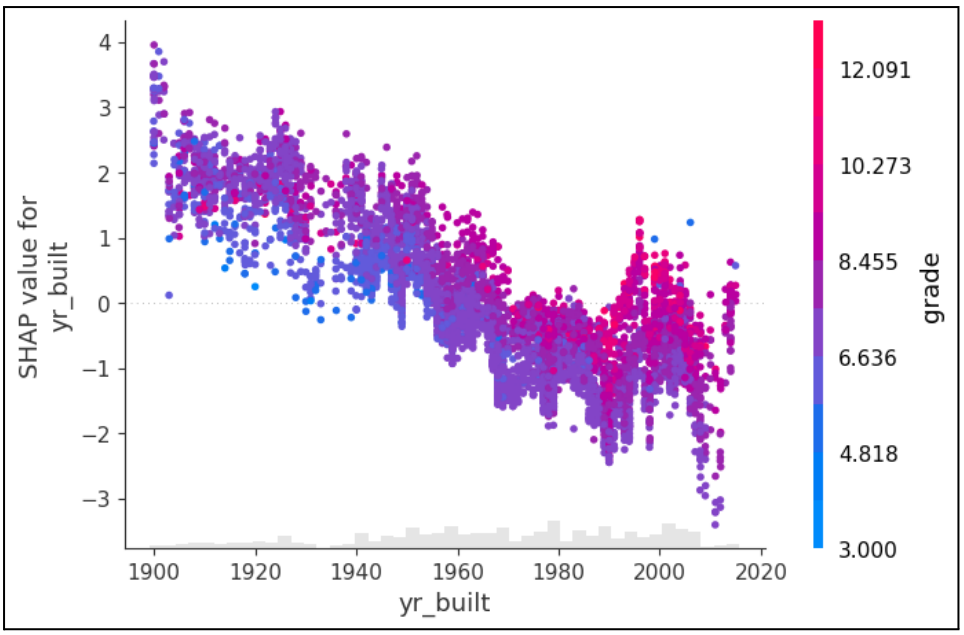


**Figure 12.** Scatter plot between year built and its shap value.

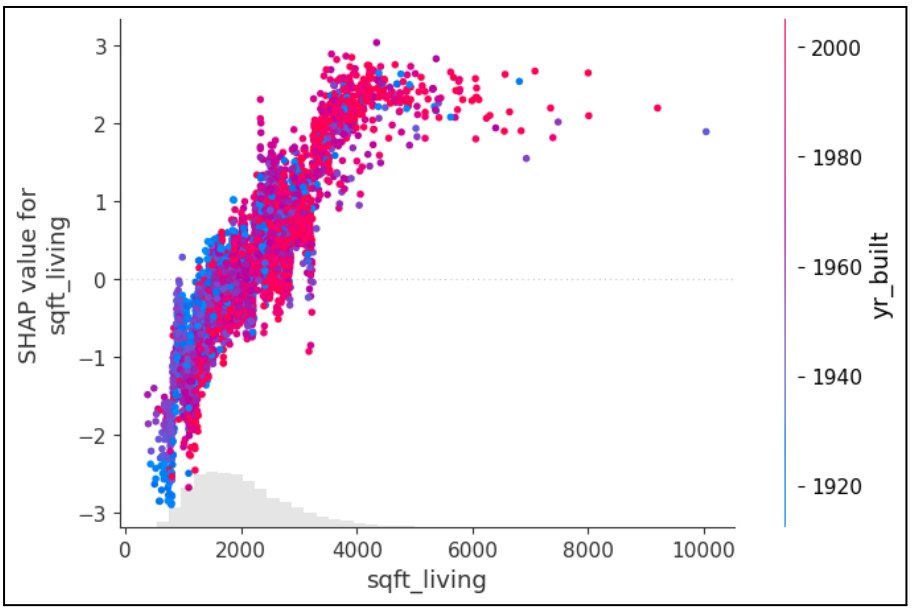


**Figure 13.** Scatter plot between square footage and its shap value.

Oriol Gallego Vazquez: 229408
Cinta Arnau Arasa: 231629



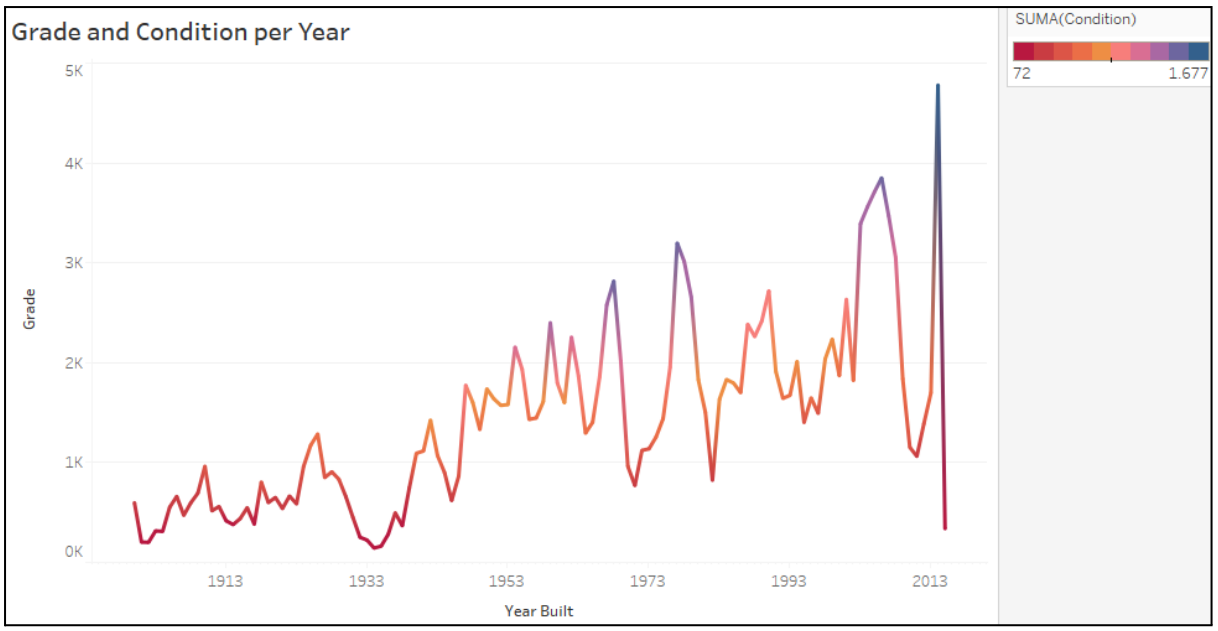


**Figure 14.** Grade and Condition of a house based on the Year Built.

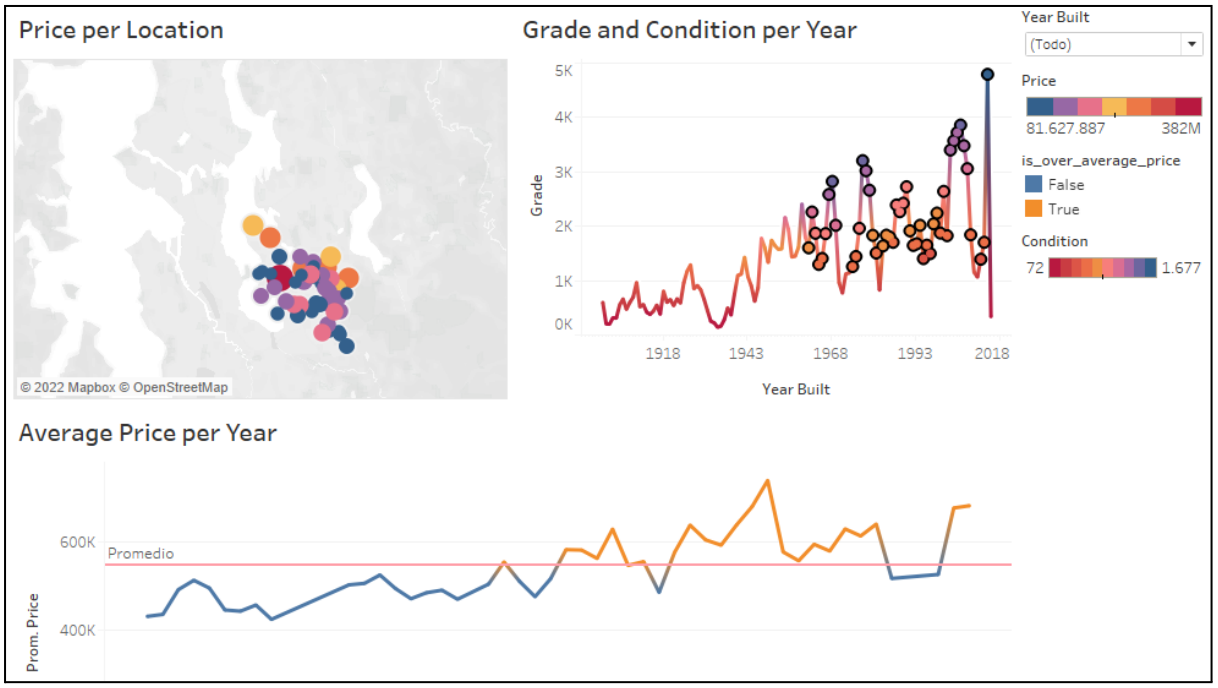


**Figure 15.** Dashboard 1. Location of houses with better grade and condition.

Oriol Gallego Vazquez: 229408
Cinta Arnau Arasa: 231629

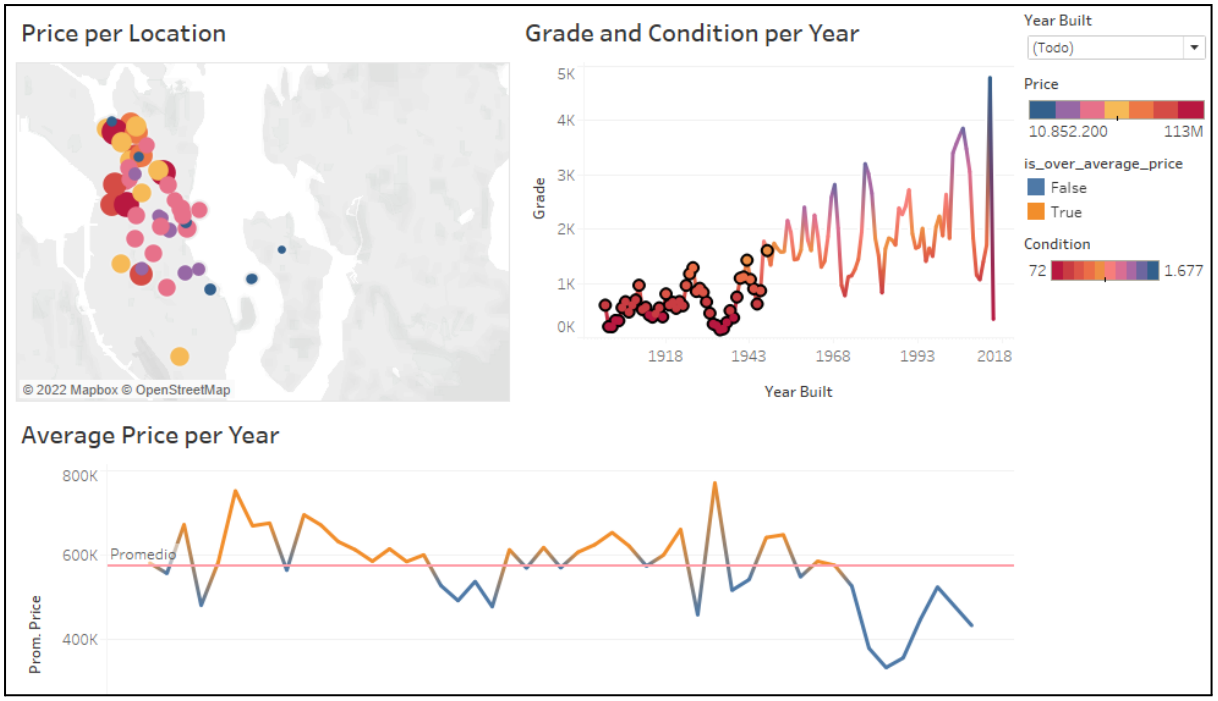


**Figure 16.** Dashboard 1. Location of houses with worse grades and conditions.

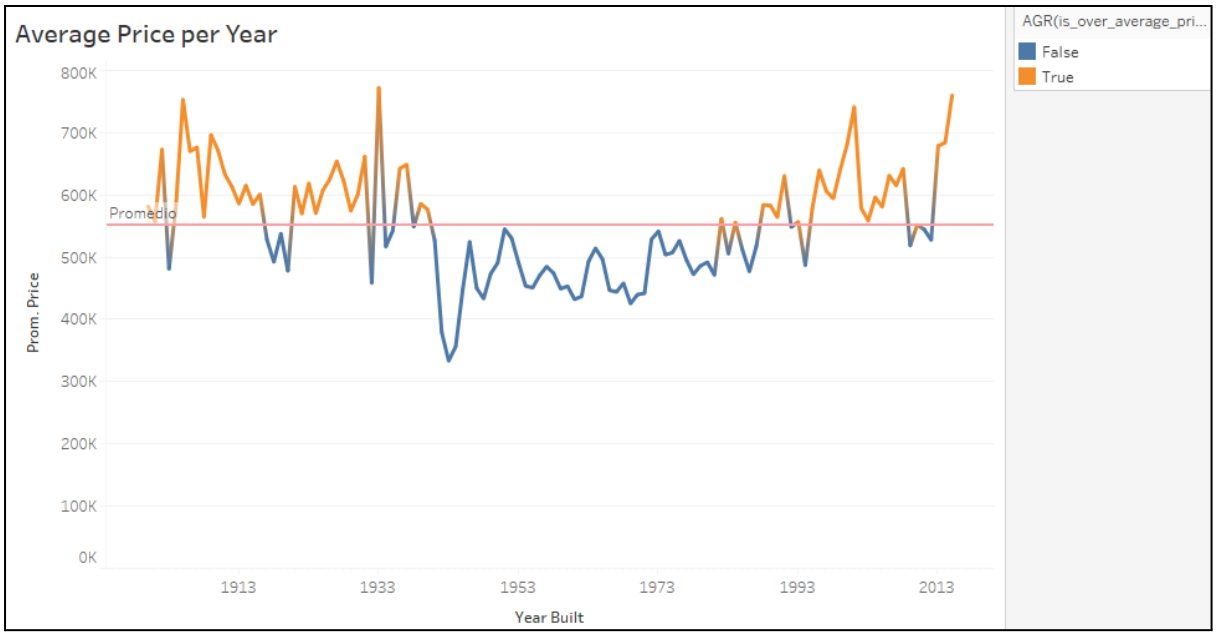


**Figure 17.** Average Price per Year.

Oriol Gallego Vazquez: 229408
Cinta Arnau Arasa: 231629

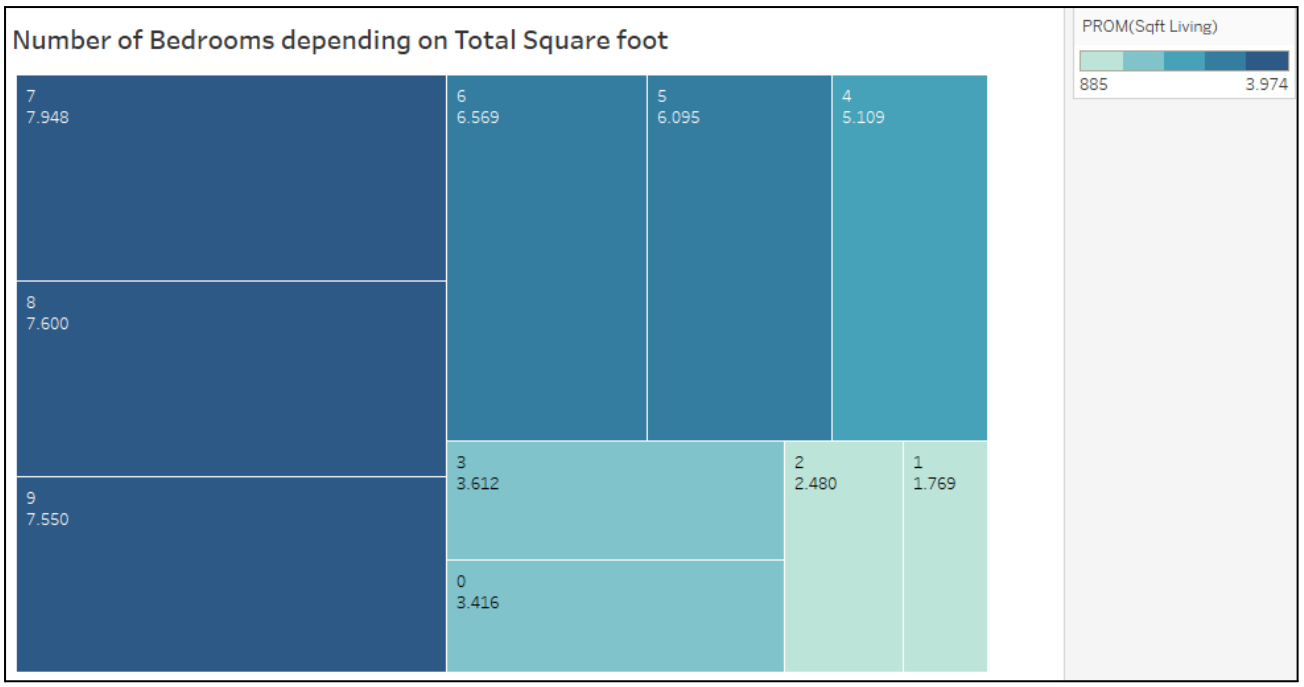


**Figure 18.** Number of Bedrooms depending on Total Square foot.

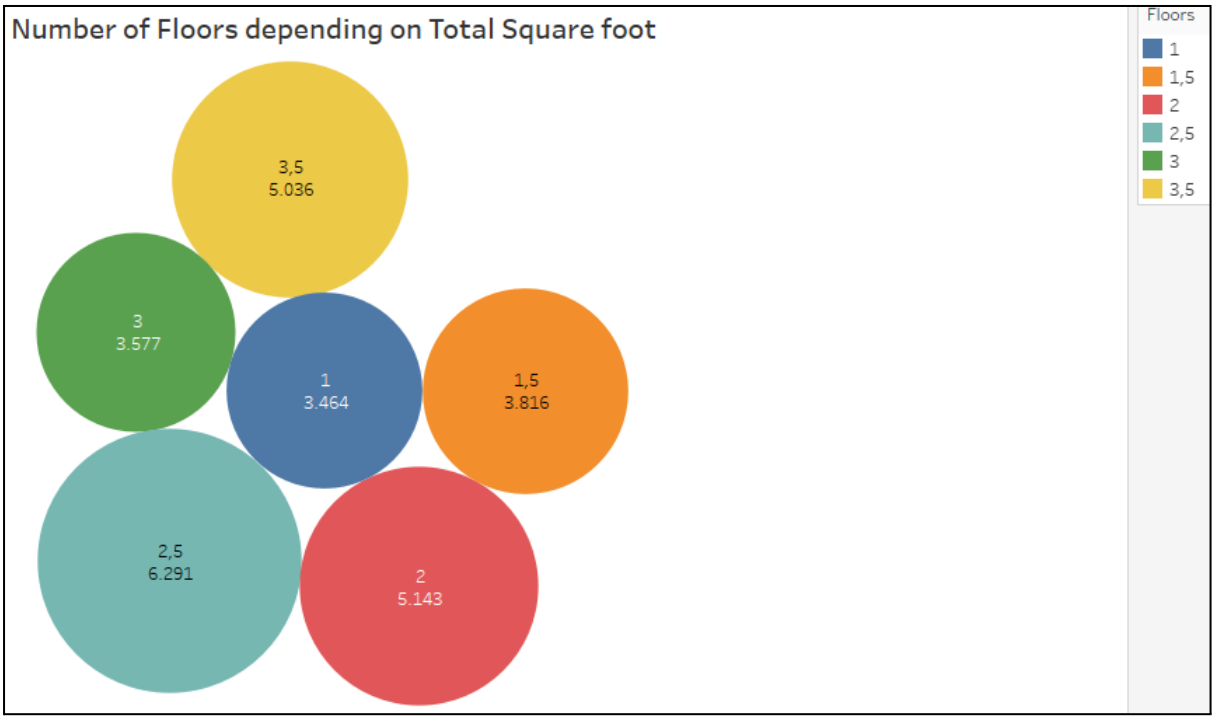


**Figure 19.** Number of Floors depending on Total Square foot.

Oriol Gallego Vazquez: 229408
Cinta Arnau Arasa: 231629



| Bathrooms | total_average_sqft |
|---|---|
| 0 | 3.274 |
| 0,5 | 2.470 |
| 0,75 | 1.793 |
| 1 | 2.362 |
| 1,25 | 3.051 |
| 1,5 | 3.089 |
| 1,75 | 3.574 |
| 2 | 3.593 |
| 2,25 | 4.195 |
| 2,5 | 4.756 |
| 2,75 | 5.340 |
| 3 | 5.472 |
| 3,25 | 6.448 |
| 3,5 | 6.823 |
| 3,75 | 7.559 |
| 4 | 8.144 |
| 4,25 | 9.219 |
| 4,5 | 8.812 |
| 4,75 | 10.551 |
| 5 | 9.704 |
| 5,25 | 10.032 |
| 5,5 | 13.012 |
| 5,75 | 13.925 |
| 6 | 12.887 |
| 6,25 | 16.690 |
| 6,5 | 13.530 |
| 6,75 | 17.120 |

1.793 — 17.120

**Figure 20.** Number of Bathrooms depending on Total Square foot.

Oriol Gallego Vazquez: 229408
Cinta Arnau Arasa: 231629

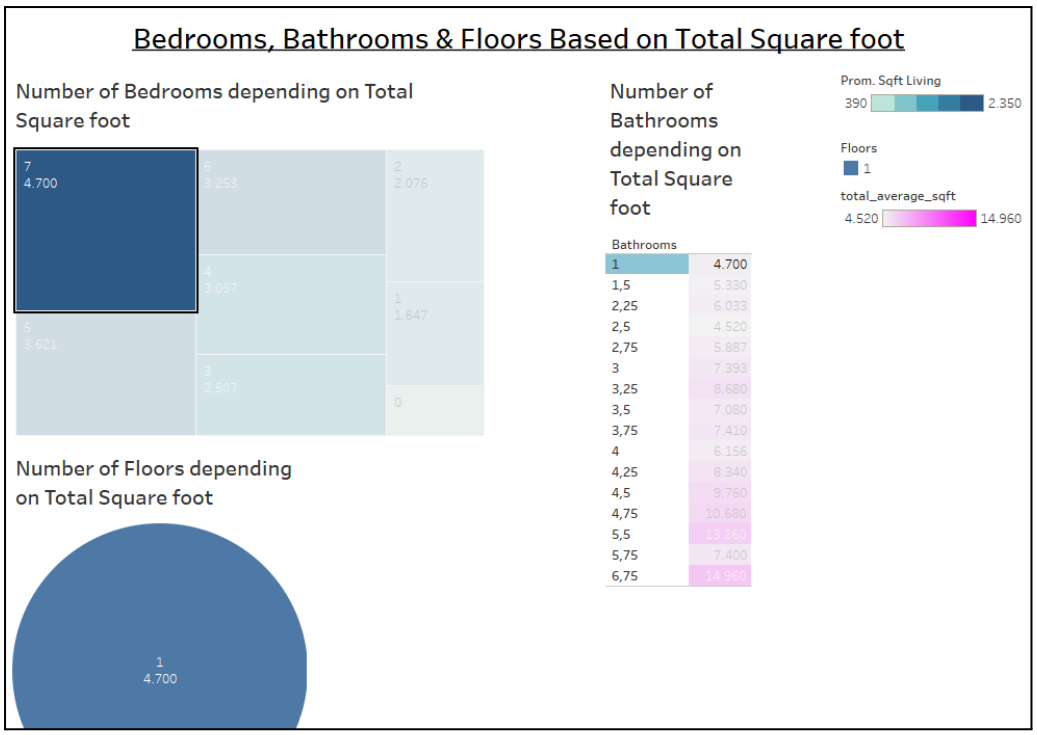


**Figure 21.** Example 1. Bedrooms, Bathrooms & Floors Based on Total Square foot.

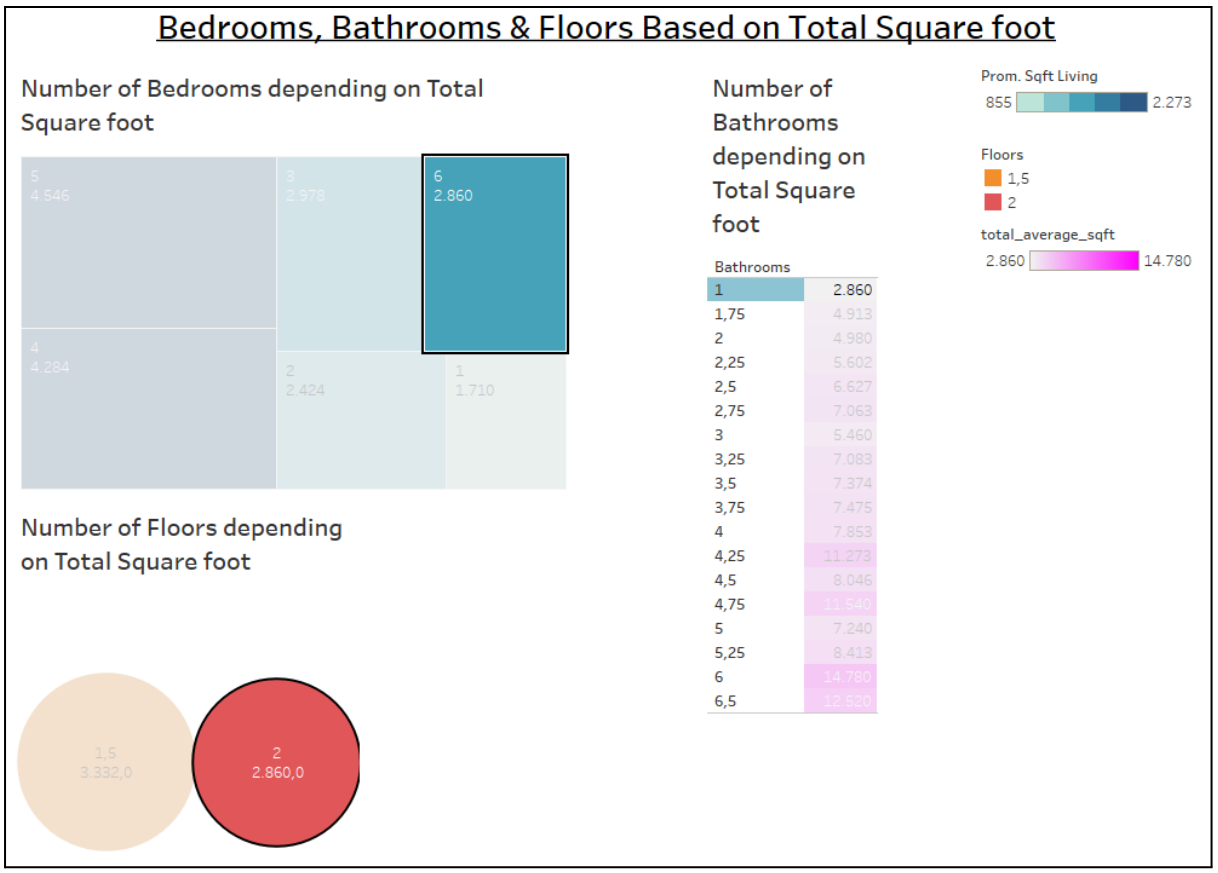


**Figure 22.** Example 2. Bedrooms, Bathrooms & Floors Based on Total Square foot.

Oriol Gallego Vazquez: 229408
Cinta Arnau Arasa: 231629

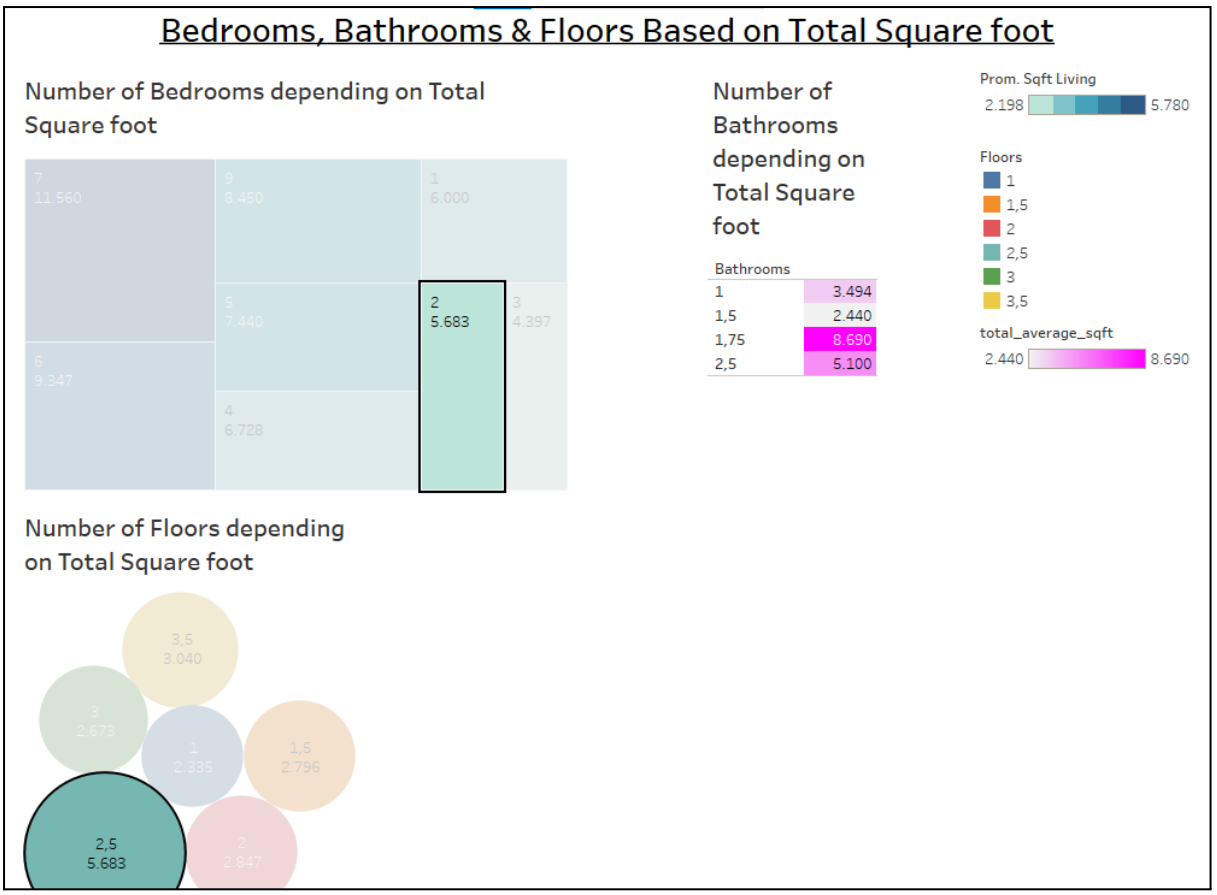


**Figure 23.** Example 3. Bedrooms, Bathrooms & Floors Based on Total Square foot.