Tableau analysis report: key findings

Summary 1
Introduction 3
Distribution plot 3
a) price vs. number of bedrooms 3
b) price vs. number of bathrooms 4
c) price vs. condition 5
d) price vs. floors 7
e) price vs. grade 8
f) price vs. view 9
g) price vs. waterfront 10
Relationship between variables 11
Linear regression 12
a) price vs. sqft_above 12
b) Price vs. sqft_basement 13
c) Price vs. living15 14
d) Price vs. sqft_lot15 14
Zip codes 15
Additional plotting 16
Time series analysis 17
Filtered data visualization, calculated fields, and bucketing 17
Further considerations, data source, methodology and limitations 22

Summary

Here we provide a brief of the findings that we gathered from our analysis:

- Distribution plots

- We found that, on average, **8-bedroom housing unit** has the highest price among the rest, which exceeds $1,1 million. This plot resembles a ‘left-skewed’ distribution.

- The average **housing unit price** tends to increase with the number of bathrooms; housing units with 7-8 bathrooms are worth between $5 and $7 million.
- In general, the better the **condition,** the better the price, we can see from the below graph that the average price per condition (1-5) moves in a range between 350K and 650K.
- Price varies across the **number of floors**, the highest average housing unit price is concentrated in housing units with 2,5 floors, between $1 and $1,1 million.
- The better the **grading,** the higher the average price, which ranges from $260K for grading 3 housing units to $3,7 million for housing units with the highest grading (13).
- Price also improves the better the **view**, ranging from $500K the housing unit with the poorest view to $1,5 million for the housing unit with the best view (scale from 0 to 4).
- Housing units with **waterfront views** present a significantly higher average level price to housing units with no waterfront view, the average difference in price exceeds $1 million. This is one of the categorical variables that we made the regression with in our machine learning model, to determine whether waterfront view had a significant impact on the model in terms of R2.

- Linear regression

- Even though the fit is not perfect the R2 approaches 0.6, which is a good fit in the univariate linear regression between **price and sqft above**.
- There is an R2 slightly below 0,5 (0,45769) between **price and square feet of the basement** and we observe some outliers when we move to a square feet surface between 2.400 and 3.400 sqft, where the average housing unit price moves between $7 and $7.7 million.
- There is a good adjustment (R2 = 0.62) between **sqft living 15 and average price**, even though this R2 indicator is slightly higher when plotting the **price vs. sqft living** (that is, without taking into account the works made in 2015), this indicator yields an R2 of 0.67.
- consumers mostly buy **3 and 4-bedroom housing units**, a total of 16.000 housing units, which accounts for 77% of the total number of housing units that we have in our dataset.

- Filtered data visualization, calculated fields, and bucketing

- Nearly **79% of the total number of housing units** were built **between 1900 and 2000,** significantly above the other two categories, with an important decrease after the year 2010

- From the category analysis of the three groups of housing unit, we find in general that, as we move across categories, the average price is higher, so does the average number of bedrooms, condition, grade, sqft above, and sqft living, however, we find that the **average square feet of the basement decreases**, this could indicate a trend of fewer houses with a basement being built

Introduction

1. Convert the necessary measures to dimensions (the variables that are categorical in nature)

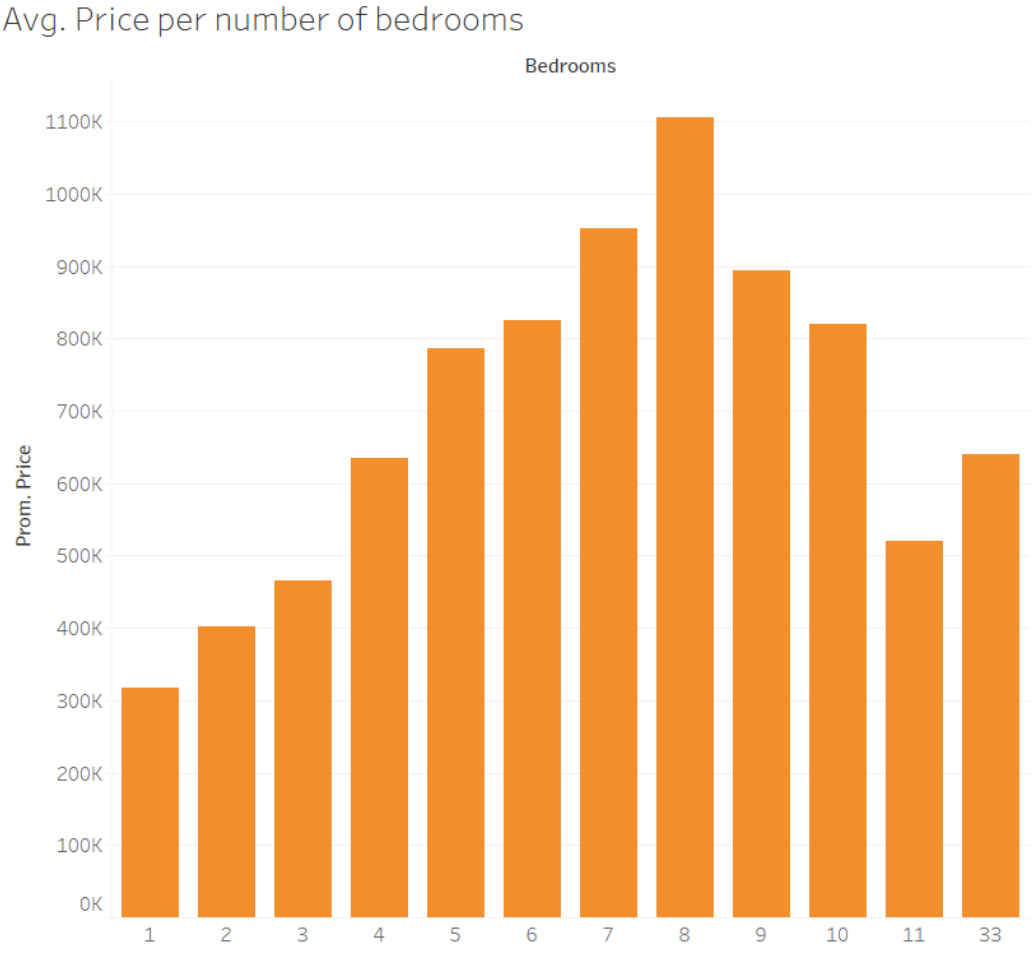
This has been done for variables like number of bedrooms, conditions, waterfront view and zip code. We found that variables such as 'number of bathrooms' contained 'non-discrete' values.

Distribution plot

2. Plot the distribution of price vs. number of bedrooms, price vs. number of bathrooms, price vs. condition, price vs. floors, price vs. grade, price vs. view, and price vs. waterfront. Done

a) price vs. number of bedrooms

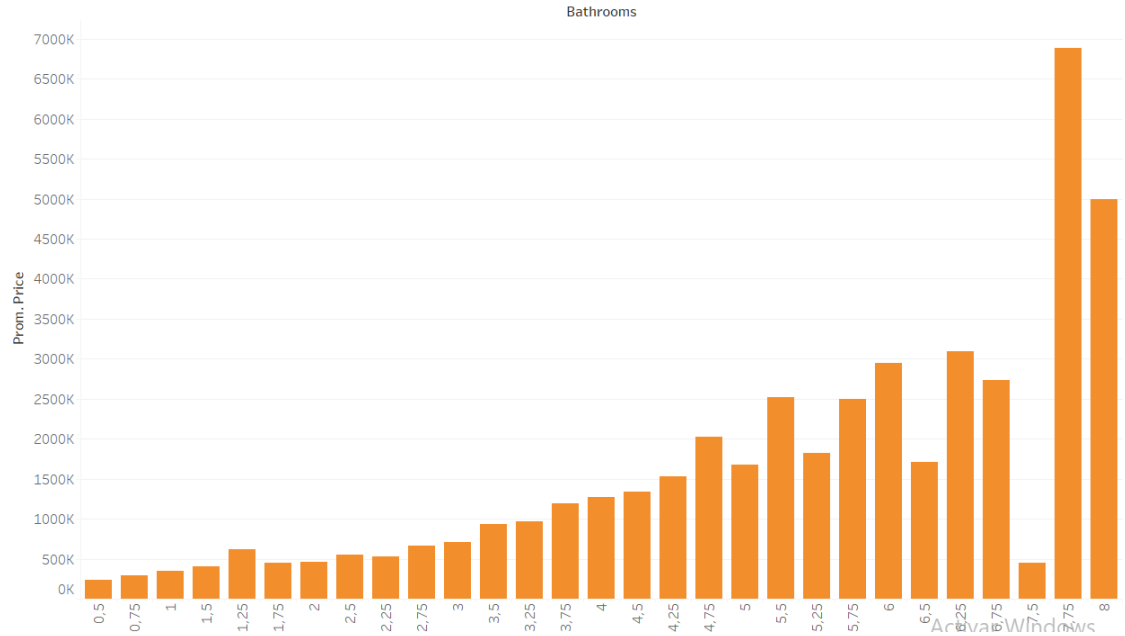
We found that, on average, 8-bedroom housing unit has the highest price among the rest, which exceeds $1,1 million. This plot resembles a 'left-skewed' distribution.



b) price vs. number of bathrooms

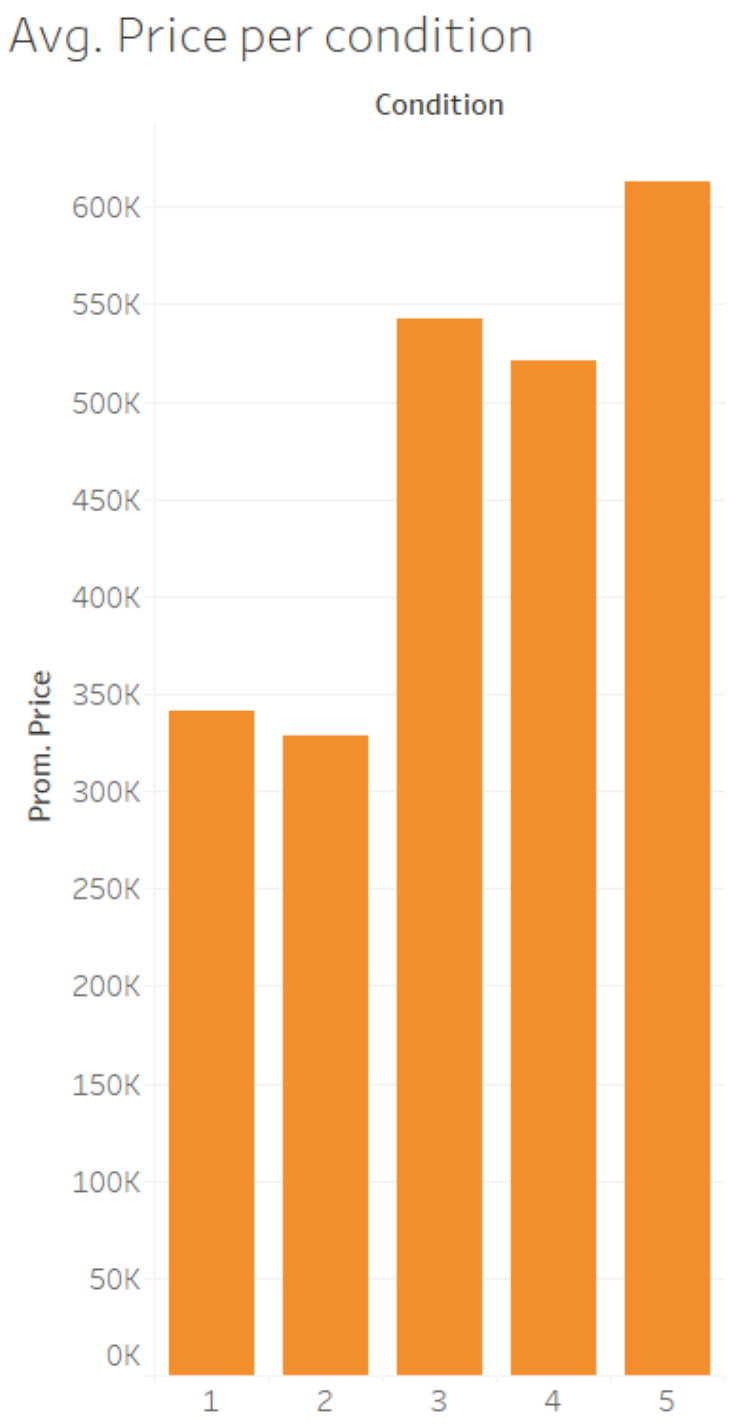
The average housing unit price tends to increase with the number of bathrooms; housing unit with 7-8 bathrooms are worth between $5 and $7 million.

Avg. price per number of bathrooms



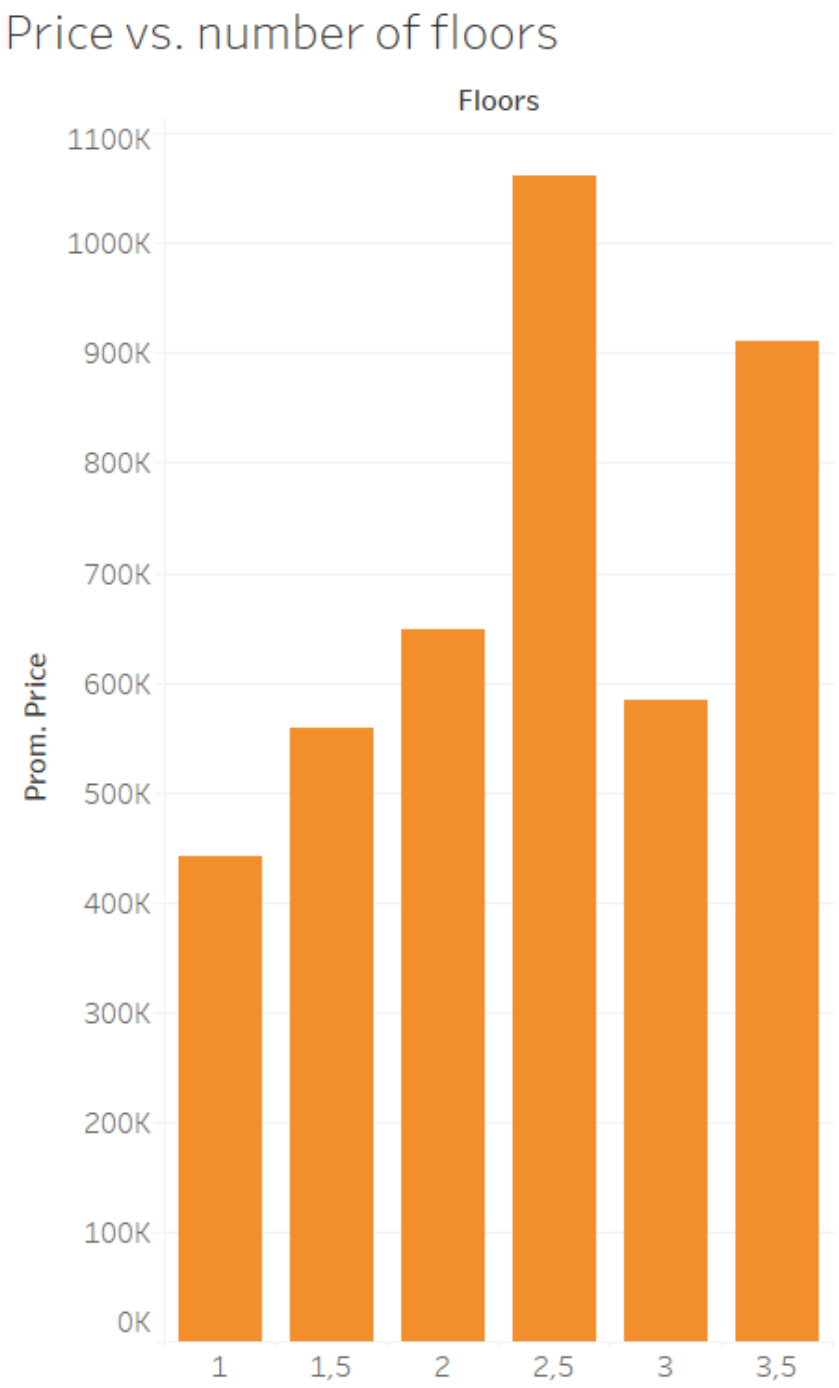
c) price vs. condition

In general, the better the condition, the better the price, we can see from the below graph that the average price per condition (1-5) moves in a range between 350K and 650K.


Avg. Price per condition
Condition
Prom. Price
600K
550K
500K
450K
400K
350K
300K
250K
200K
150K
100K
50K
0K
1
2
3
4
5

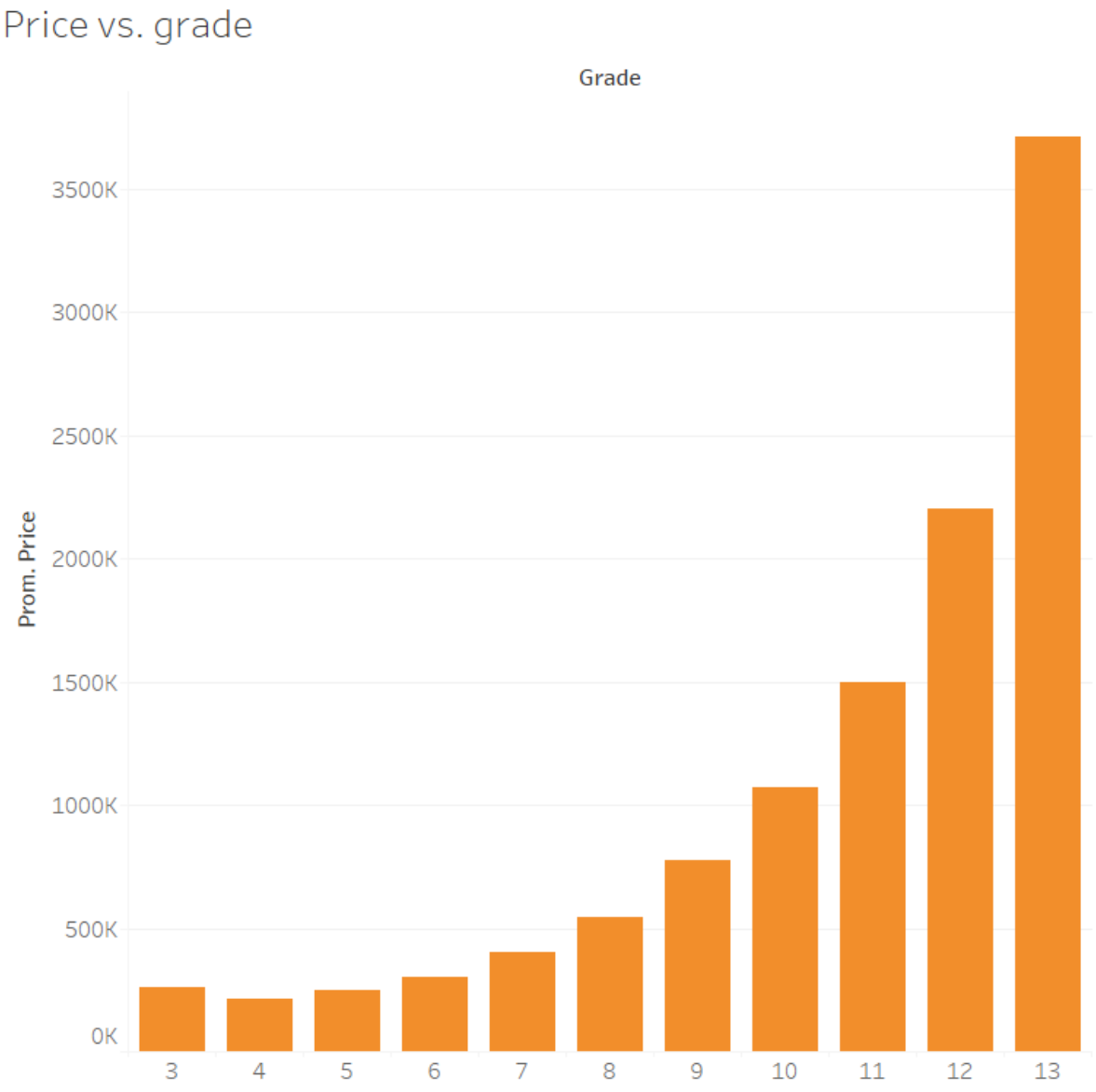
d) price vs. floors

Price varies across the number of floors, the highest average housing unit price is concentrated in housing unit with 2,5 floors, between $1 and $1,1 million.



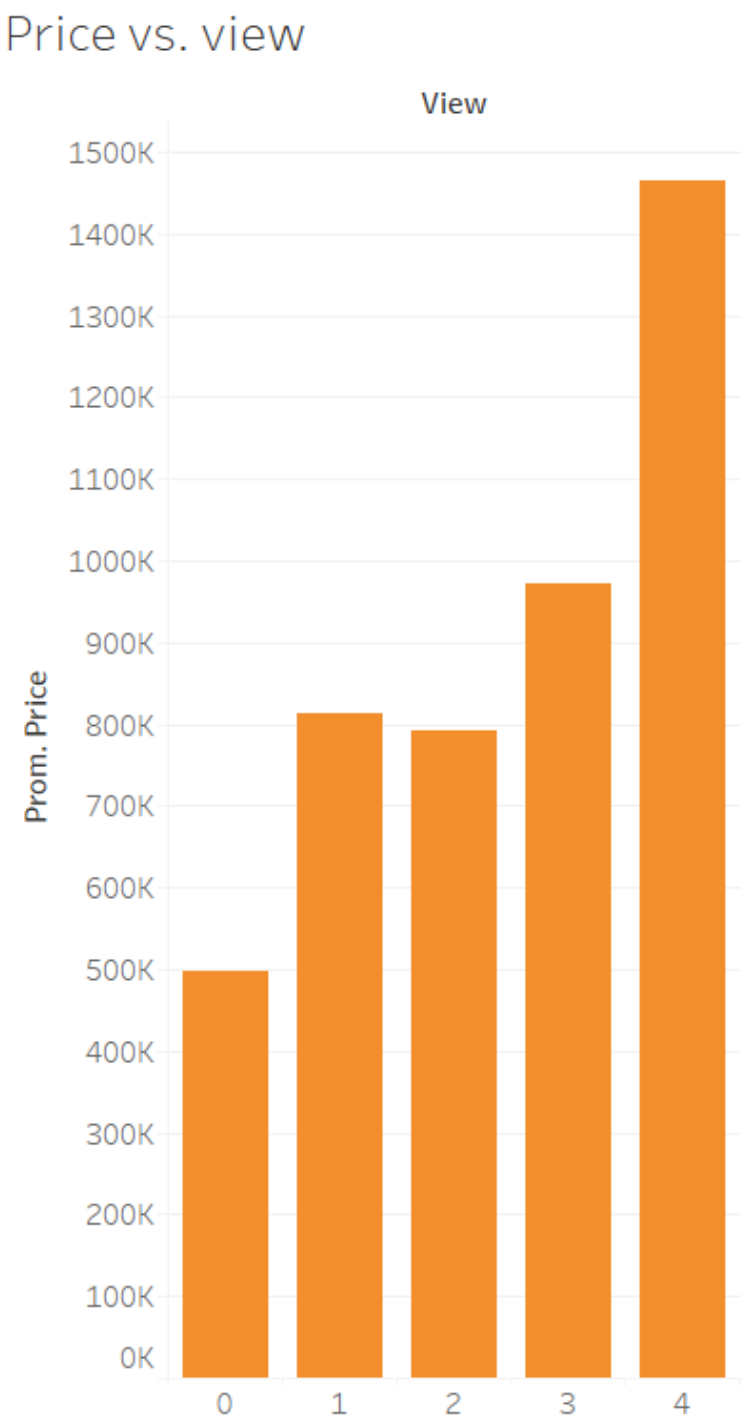
e) price vs. grade

The better the grading, the higher the average price, which ranges from $260K for grading 3 housing unit to $3,7 million+ for housing unit with the highest grading (13).



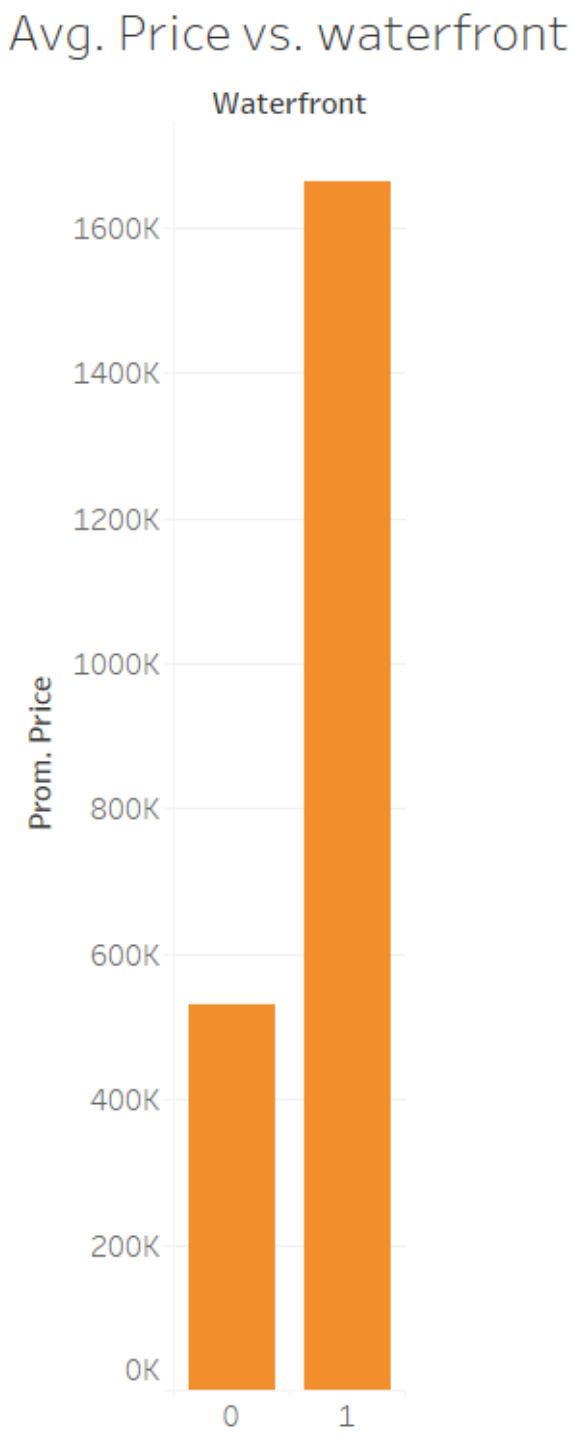
f) price vs. view

Price also improves the better the view, ranging from $500K the housing unit with the poorest view to $1,5 million for the housing unit with the best view (scale from 0 to 4).



g) price vs. waterfront

housing unit with waterfront view present a significantly higher average level price to housing unit with no waterfront view, the average difference in price exceeds $1 million.



Relationship between variables

3. State your observation for each one of those graphs. Do you see any trends in prices vs the rest of those variables individually? This can also

be used for EDA to identify some data cleaning operations that you might need to perform further.

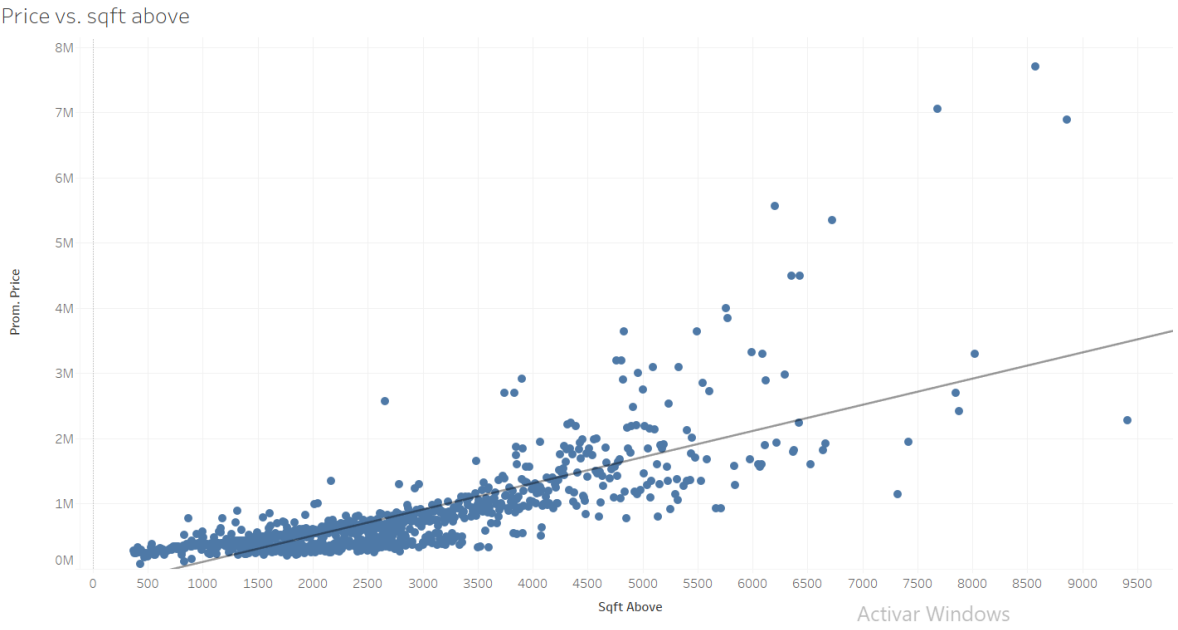
In general, from the graphs that we have plotted from data visualization with Tableau, we observe a positive relationship between prices and the previously stated variables (price vs. number of bedrooms, price vs. number of bathrooms, price vs. condition, price vs. floors, price vs. grade, price vs. view, and price vs. waterfront).

The above is true, even though we observe some decreasing returns: as for the average number of bedrooms, the highest average value is concentrated among 8-bedroom housing unit, out of 33, which is the maximum number of possible bedrooms in our dataset, which comes from a single outlier. This also happens with number of floors, price is highest for housing unit between 2 and 3 floors (2,5).

Linear regression

4. Draw scatter plots for price vs. sqft_above, price vs. sqft_basement, price vs. living15, price vs. sqft_lot15.

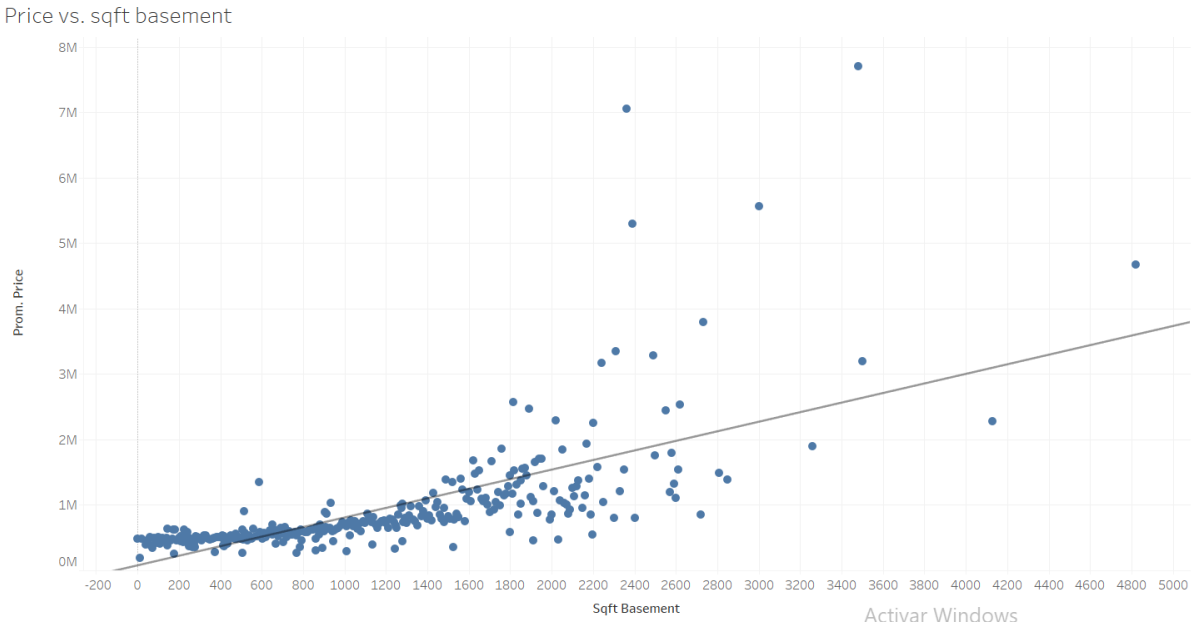
a) price vs. sqft_above



R2 = 0.5991

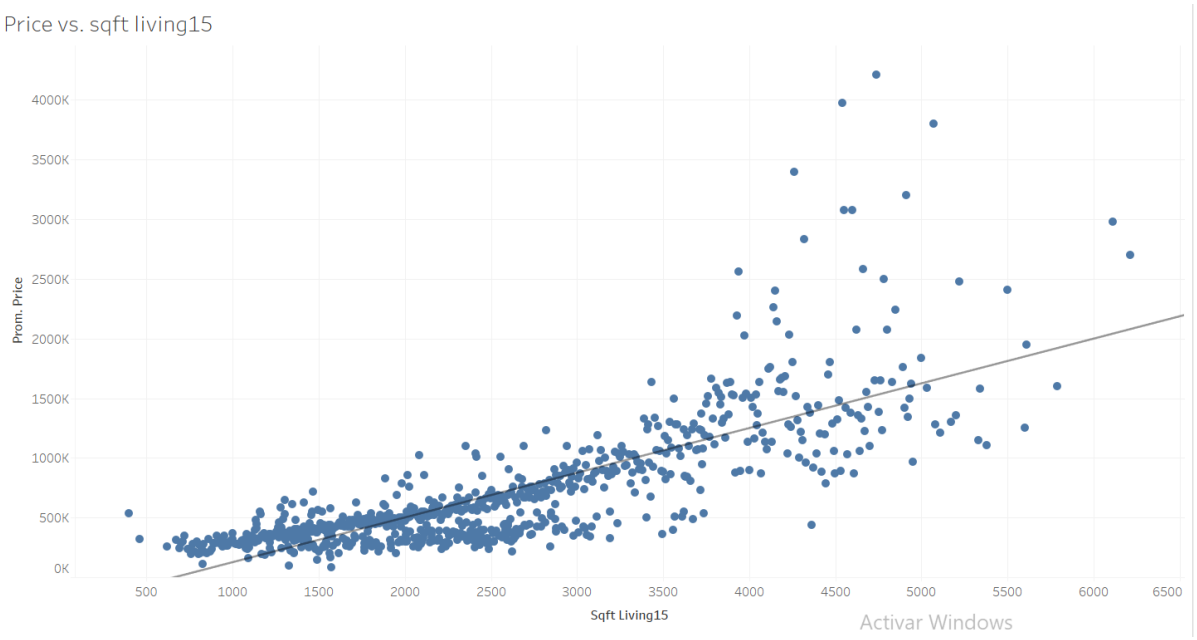
This R2 approaches 0.6, which is a good fit in the univariate linear regression between price and sqft above

b) Price vs. sqft_basement



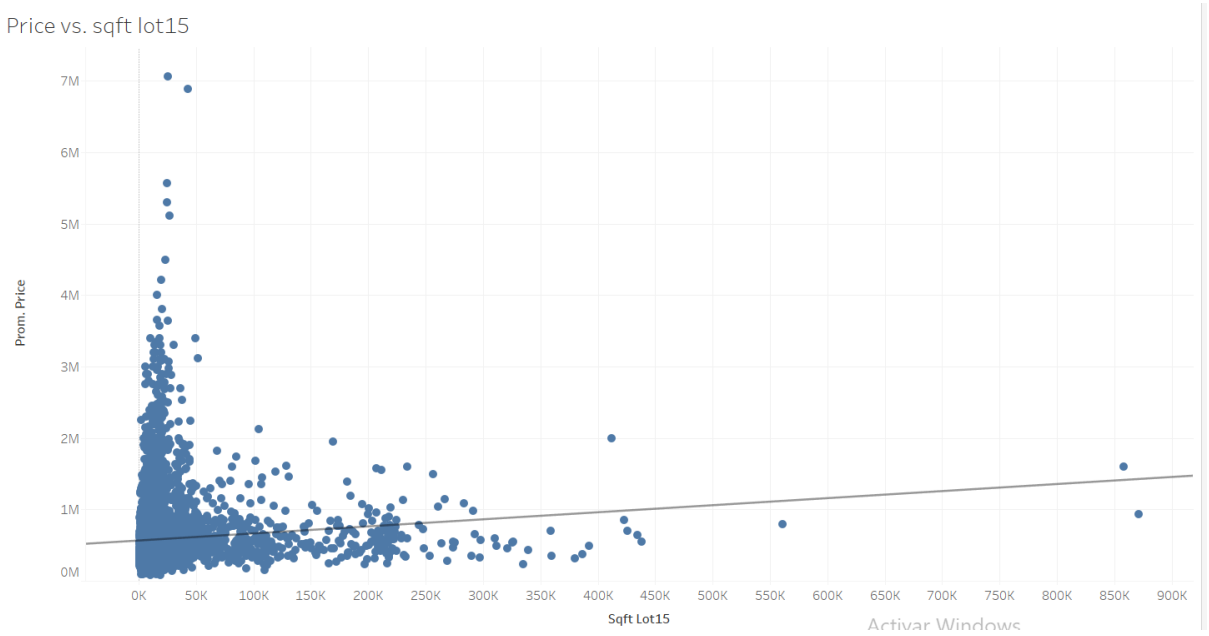
There is an R2 slightly below 0,5 (0,45769) between price and square feet of the basement and we observe some outliers when we move to a square feet surface between 2.400 and 3.400 sqft, where the average housing unit price moves between $7 and $7.7 million.

c) Price vs. living15



There is a good adjustment (R2 = 0.62) between sqft living 15 and average price, even though this R2 indicator is slightly higher when plotting the price vs. sqft living (that is, without taking into account the works made in 2015), this indicator yields an R2 of 0.67.

d) Price vs. sqft_lot15



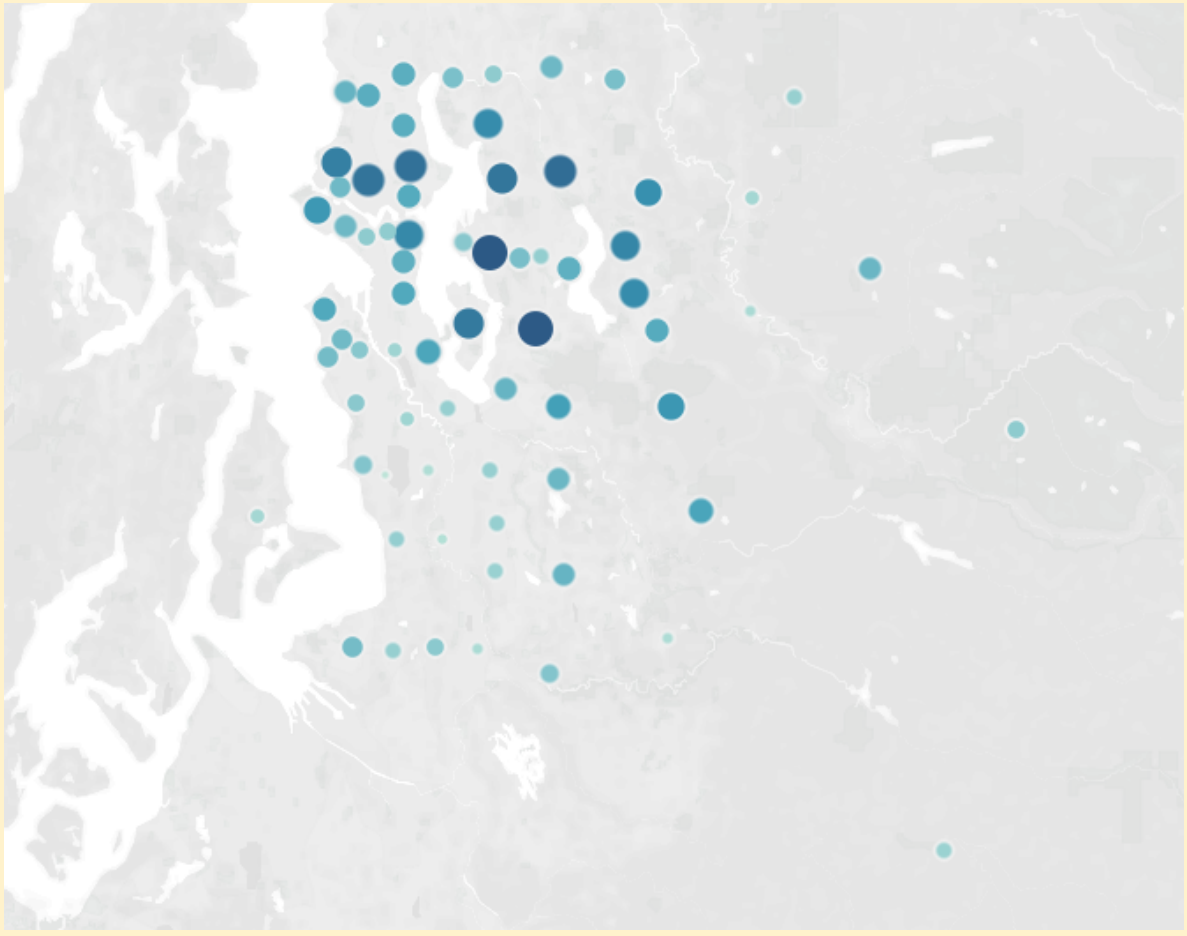
There is a really weak (R2 = 0.008) between price and square feet lot 2015.

Zip codes

5. Identify using tableau which state data is presented to you. Use latitude (generated), longitude (generated), and zip code for this.

We are presented with zipcodes from 98001 to 98199, which belong to the State of Washington, US, across different cities.

By crossing the lat, long and zipcode values we generate the corresponding map:



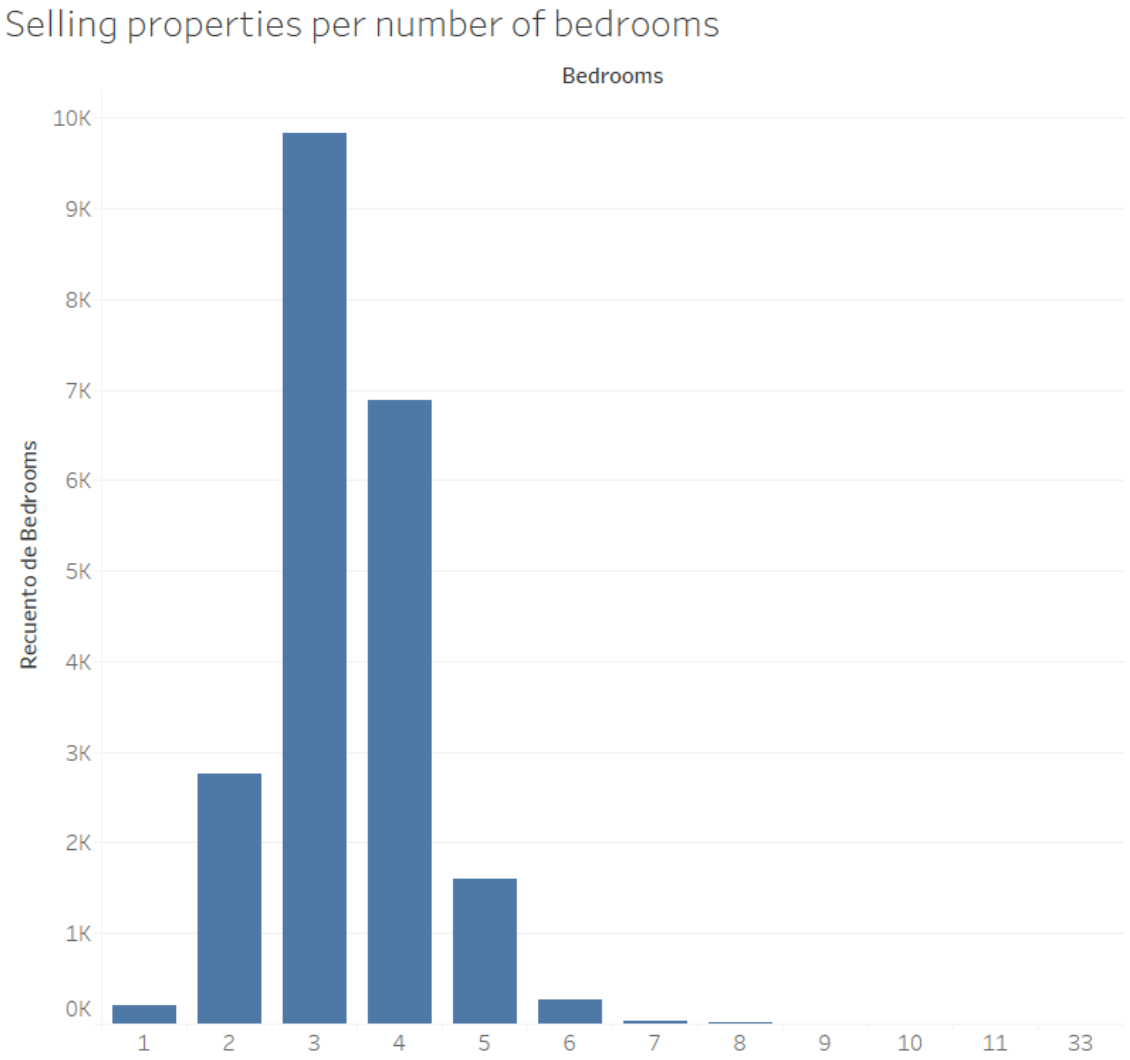
From the zipcode map, we can interpret that average housing unit is smaller as we go further to the outer locations in the map; on the other hand, housing unit with the highest price are concentrated next to one another.

6. Color code the zip codes based on the prices to see which areas are more expensive than the others.

Done above, highlighted the areas with highest total value of the housing unit.

Additional plotting

7. Create a plot to check which are the more selling properties based on the number of bedrooms in the house. Create a plot of bedrooms vs. count of data points.

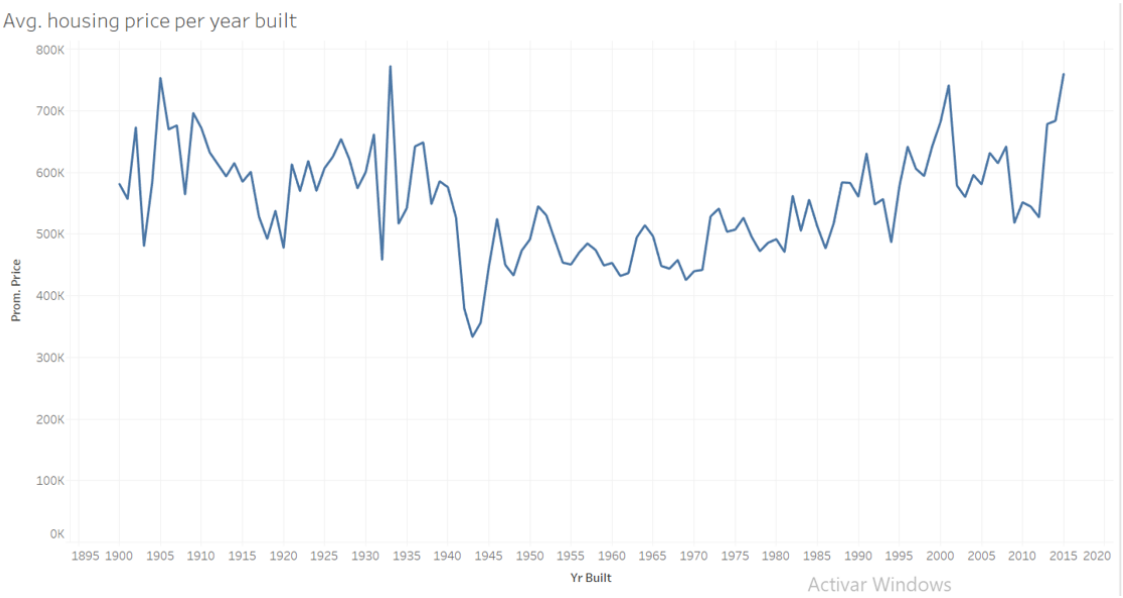


We see from the chart that consumers mostly buy 3 and 4-bedroom housing unit, total of 16.000 housing unit, which accounts for 77% of the total number of housing unit that we have in our dataset.

Time series analysis

8. We want to see the trend in price of houses based on the year built.

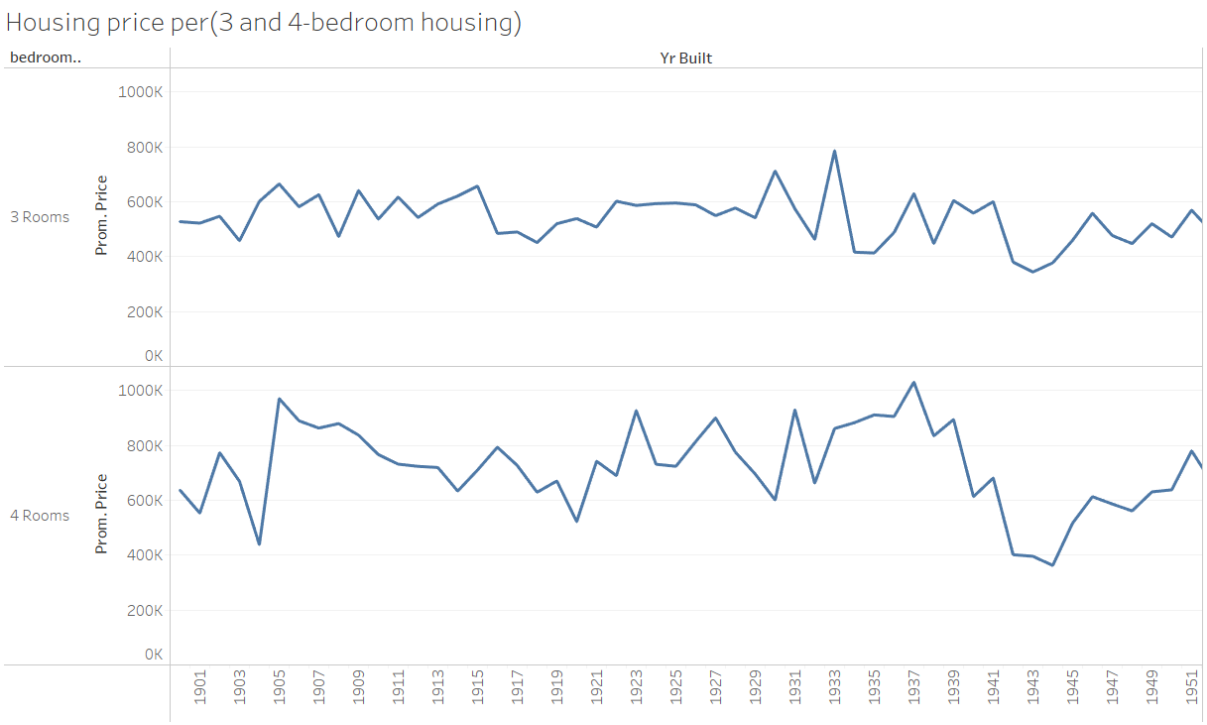
Housing price from 1900 to 2010 follows a distribution with an average of 540296.57usd and a standard deviation of 367368.14usd, a high variation for the time periods, identifying some periods of high volatility (early 1940s and 2010).



Filtered data visualization, calculated fields, and bucketing

9. From our previous plot, we know that most of our customers are interested in three and four bedroom houses. Create a filter on bedroom features to select those properties and compare the trends in prices using line charts.

We have created a chart to see the average housing unit price per year only for 3 and 4-bedroom housing unit



In general we see the same trend in housing unit prices for 3 and 4-bedroom housing unit. A slight observation allows us to interpret that the average price ranges from 400K to 800K in 3-bedroom housing units, whereas this range is wider in 4-bedroom housing, from an average above 400K to 1 million, which means this dataset has a higher variation among the values.

10. Create calculated field year_built_bins for the column year_built by creating buckets as follows, for:
    - houses built between 1900 and 2000 - category A,
    - for houses built between 2000 and 2010 - category B,
    - and for houses built after 2010 - category C.

    Use IF-ELSE statement to create the bins/buckets. Compare the prices of houses for the three categories.

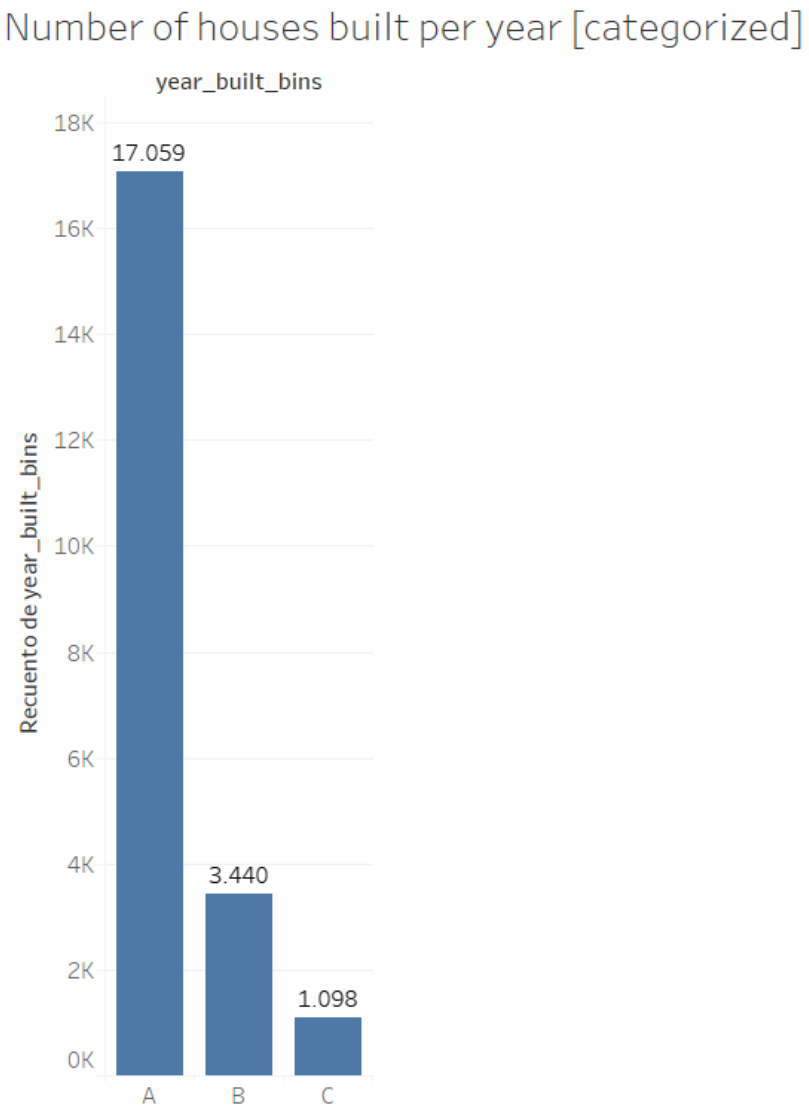
    We have used the following conditional to create the bin/bucket for this visualisation, where we separate the housing unit by category depending on the construction year:

```
IF [Yr Built] >= 1900 AND [Yr Built] <= 2000 THEN "A"
ELSEIF [Yr Built] > 2000 AND [Yr Built] <= 2010 THEN "B"
```

```
ELSEIF [Yr Built] > 2010 THEN "C"
END
```

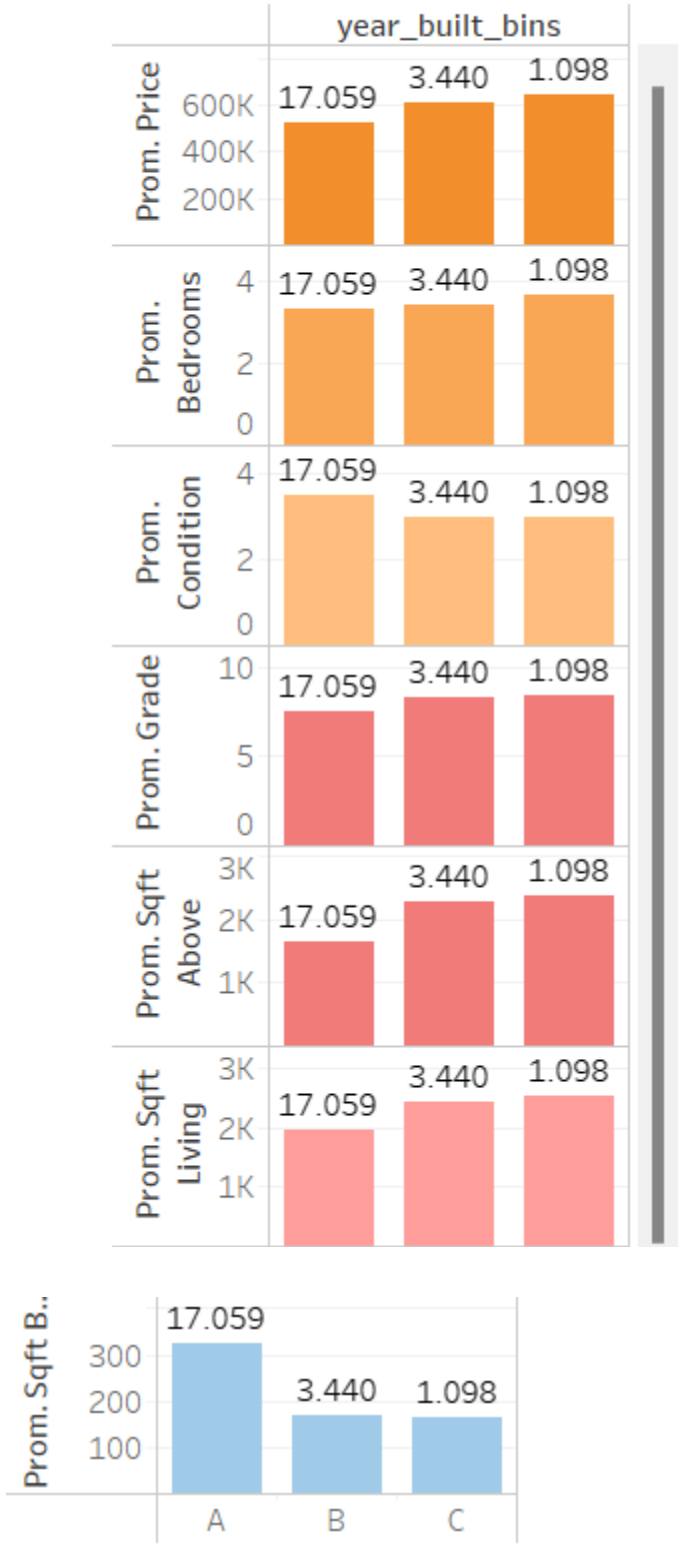
11. Now we want to deep dive into the categories we created in the last question. Let's see how many properties are in each of the categories. Indicate the numbers as labels on each of the three categories

From the chart that we have plotted with the 3 categories, we see that near 79% of the total number of housing unit were built between 1900 and 2000, significantly above the other two categories, with an important decrease after the year 2010, which is explained by the housing unit bubble arising in 2008.



12. Deep dive in category A, category B and category C using filters. Identify different characteristics/trends for each of the three categories.

From the category analysis of the three groups of housing unit, we find in general that, as we move across categories, we find that the average price is higher, so does the average number of bedrooms, condition, grade, sqft above, and sqft living, however, we find that the average square feet of the basement decreases, this could indicate a trend of fewer houses with a basement being built as we move forward.


Variable analysis per housing category
year_built_bins
Prom. Price
600K
400K
200K
17.059
3.440
1.098
Prom. Bedrooms
4
2
0
Prom. Condition
Prom. Grade
10
5
Prom. Sqft Above
3K
2K
1K
Prom. Sqft Living
Prom. Sqft B..
300
200
100
A
B
C

Further considerations, data source, methodology and limitations

- Further considerations

This report is a summary of a deeper data visualization performed with Tableau Public, meaning the information generated is available to the general public, as we are working with the free version of the tool.

On the other hand, this piece of work is part of an assignment which consists of an extensive data analysis with Pandas / Jupyter notebook, where data transformation is performed when necessary in order to improve the model accuracy, and where there is machine learning (ML) involved. To keep this report simple, we have skipped this part, as well as the sql queries that we have implemented in order to display certain information as requested by our stakeholders.

- Data source:

This is the original link to the requirements of the data visualization project: https://github.com/ironhack-edu/data_mid_bootcamp_project_regression/blob/master/tableau_regression.md

This is the link to the public tableau data visualization with all the sheets and dashboards
https://public.tableau.com/views/MID_BOOTCAMP_03-05/Hoja26?:language=es-ES&publish=yes&:display_count=n&:origin=viz_share_link

- Limitations

Due to some time constraints, we could not look into some aspects resulting from the data visualization, including the following:

- As for the number of housing built per category (A, B and C, depending on the year of construction, between 1900 - 2000, 2001 - 2010 and 2010 onwards) we miss some information on the reason why the number of housing units sold decreases as we move forward, we can assume that there was a housing bubble between those periods which generated a credit crisis (like the one we had in 2009-2010) but there might well be other factors explaining this trend, that we have missed and we have limited information to explain this
- Looking into the reason why housing price varies among the different postcodes, it would be more complete if we had information on demographic variables such GDP per capital, density of population or average daily income among other

variables and cross them in order to gain more valuable insights into our analysis

- Methdology

For this report we were provided with a dataset that contained 21597 observation of housing units sold for the period between 01/06/2014 and 12/05/2015.

Link to the original csv is the following:

https://github.com/ironhack-edu/data_mid_bootcamp_project_regression/blob/c2c811ad2a101cdd245bc04d0e71450c7940ac5f/regression_data.csv

Here is a set of the possible explanatory variables, being housing price, in usd, the target variable:

- bedrooms
- bathrooms
- sqft_living
- sqft_lot
- floors
- waterfront
- view
- condition
- grade sqft_above
- sqft_basement
- yr_built
- yr_renovated
- zipcode
- lat
- long
- sqft_living15
- sqft_lot15