FINAL PROJECT

*A brief recap of my data, goals, and tasks

A dataset that I'm interested in is the house data which derived from data in the cities of the state of Washington USA from May1' to July 9', 2014. It includes information such as: date, price, the number of bedrooms, the number of bathrooms, square feet living, square feet lot, the number of floors, square feet above, square feet basement, year built, year renovated, street, city, state zip. There are 4600 homes in the data set.
This data was downloaded from the link: https://www.kaggle.com/shree1992/housedata

I think to decide for buying, selling or holding a home, we should understand the trend of the house price in the real estate market cycle and figure out factors that we need to take into consideration. Since if the asking price is too high, the likelihood of the home being sold greatly decreases. On the other hand, if a home we are looking to sell is priced too low, we'll miss out on potential profits. Therefore, my goal is to find out the fair value of the property, know the trend of the house price in the period and how those attributes such as: the number of bedrooms, the number of bathrooms, or square feet living, year built, year renovated affect to the house price.

For existing visualizations, there are some vertical bar charts of each attribute like: date, price, bedrooms, bathrooms, square feet living, square feet lot, floors, waterfront, view and condition. There are not any graphs display the correlations between price and these attributes as well as the trend of the house price. I think I'll plot line chart because it can help to look clearly the trend of the house price and the scatterplot will support seeing the correlations in the cities. Moreover, I would like to visualize different charts for different combinations of dimensions, I think it will help to show which dimension appear to correlate most to the price.

Some tasks I would wish to complete with my data:

1. I analyze to know the trend of the house price by date.
2. Explore data to differentiate the house's price at different cities.
3. Find out the correlation between house's price and other attributes.
4. Search for properties that are comparable to the one we want or have.
5. Determine a price range.

Identifying:

There are several reasons to do this type of study, the main one is to have information that allows us to buy or sell house safely. Other reasons also include:
+ To know if in the future they will have a development that will increase the value of the properties.
+ To know if we should invest in one city instead of another.
+ By knowing what kind of house that we wish such as how many bedrooms, bathrooms, which city it is, etc.., based on the visualizations, with comparing similar properties on the market, we will be able to estimate the housing market prices. So, we can negotiate a lower price, avoid overpaying the market value of a property and vice versa.

I will build these visualizations by python, use the Altair library to create and share some interactive visualizations:

1. To explore the trend of the house price by date, I'll use mark line function to create a line graph with x = "date" and y = "price".
2. I think a scatterplot will help to visualize the house price were over the cities by adding in some color to differentiate different cities of the state of Washington USA.
3. To see the correlation between house's price and other attributes, I'll use the repeat function to look at the house price (mapped to color) across bedrooms, bathrooms, square feet living, year built and city.

+ To search for the comparable properties, I'll find the house that should have the same number of bedrooms and bathrooms, square feet, square lot, located in the same city, built year and features.

+ I'll determine the price range by setting the ceiling value and floor price. First, I'll choose one property from the comparable properties I found that is worth more than mine, maybe it's on a better street or newer. This number will be my ceiling value. On the other hand, I'll pick a house that located on a busy street or has fewer features, and this number will be my floor price.

These tasks seek to learn the trend of the housing price over time and estimate the reasonable price of a property. Target data is a list of 4600 homes in the cities of the state of Washington USA and the current listings of comparable homes. The task is performed following these steps:

+ First, collecting data is one of the most important steps in conducting house market research, as it allows, we to gather information that will help we determine if the project is viable, and our investment is secure.

+ After that, analyzing and visualizing data will help we explore the trend of property's price and find the correlations between house price and other attributes.

+ Next, looking for the current listings of comparable homes.

+ Finally, examining the price range

And as a data scientist, I am executing this task.

* Screenshots of and a link to your visualization implementation, a summary of the key elements of my design and accompanying justification

Because I see that there are some missing values with price equal 0, so I remove those rows from the data. And my target data points are the houses have price equal or below $10,000,000, I removed 2 rows that have price were over 10,000,000.

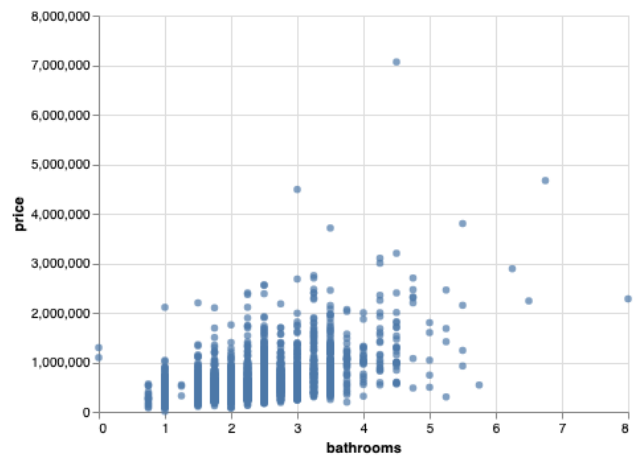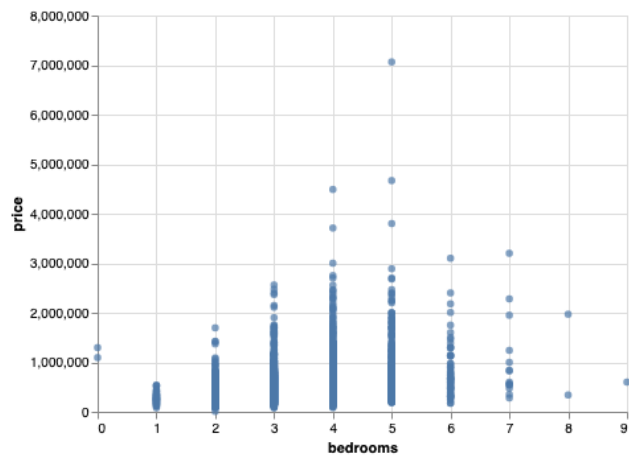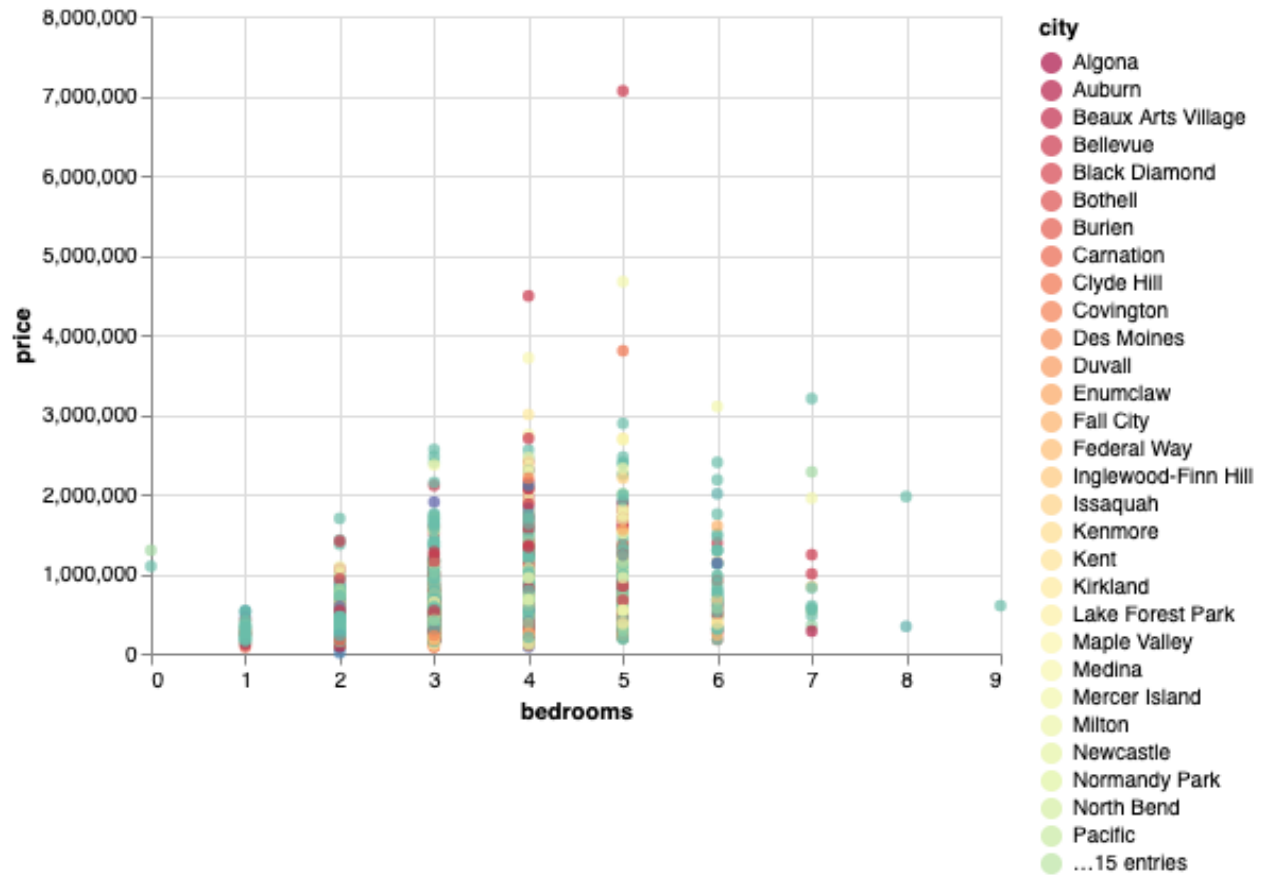This is the link to my visualization implementation:

https://colab.research.google.com/github/thuylinh225/House_price_visualizations/blob/main/House_Price_Visualizations.ipynb
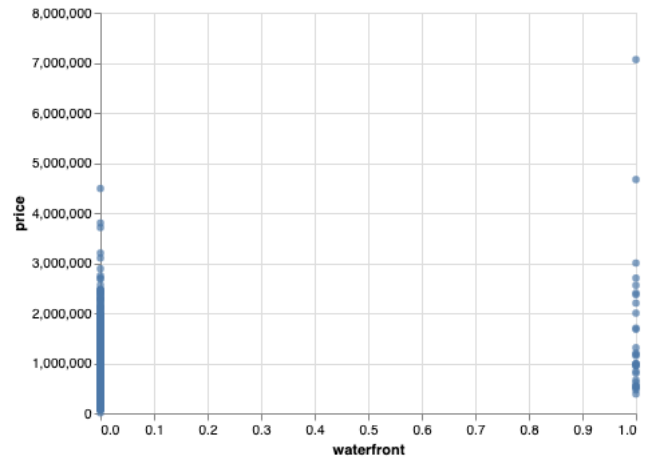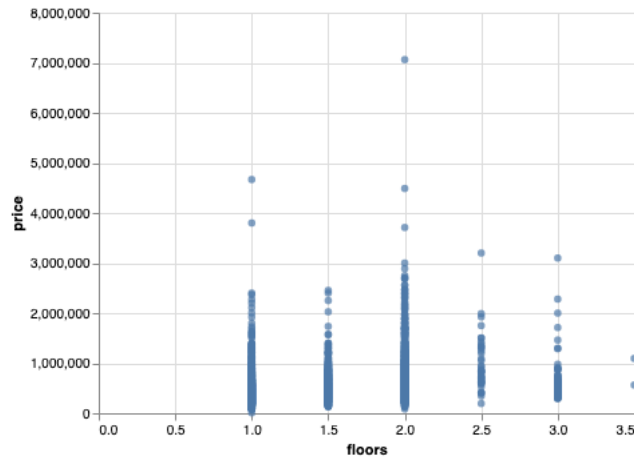
Here are screenshots of my visualizations:
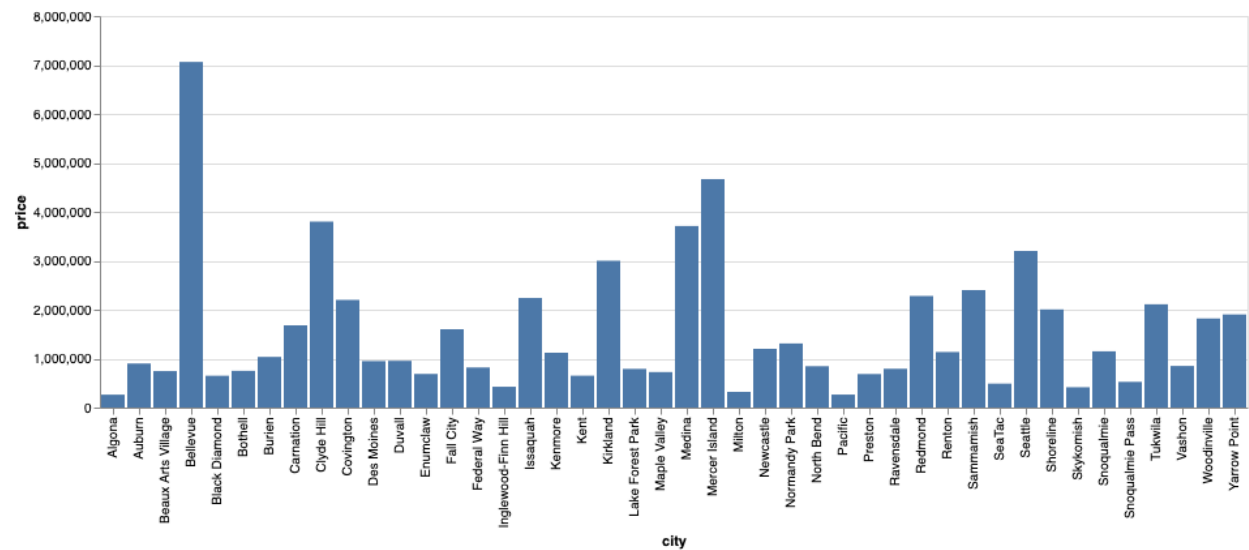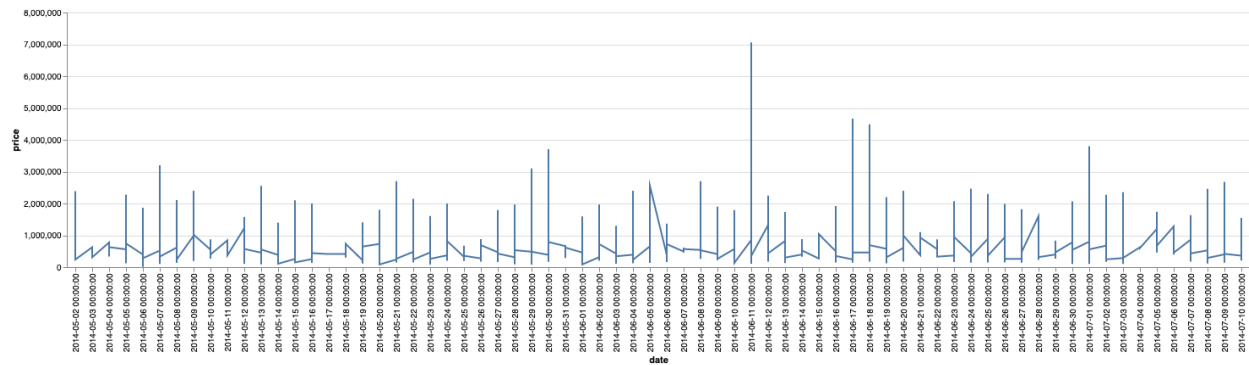
From two plots of house price by date and by city above, we can see that:

+ The house price didn't fluctuate much in the period time we chose. The maximum house price we see is over $7,000,000 on Jun 11' 2014.

+ Bellevue is the city where the highest house price was in that period time.

+ Mercer Island, Medina, Clyde Hill were the cities have the higher house price.

+ Algona is the city where the lowest house price was in that period time.

+ Pacific and Milton were the cities have the lower house price.

city
- Algona
- Auburn
- Beaux Arts Village
- Bellevue
- Black Diamond
- Bothell
- Burien
- Carnation
- Clyde Hill
- Covington
- Des Moines
- Duvall
- Enumclaw
- Fall City
- Federal Way
- Inglewood-Finn Hill
- Issaquah
- Kenmore
- Kent
- Kirkland
- Lake Forest Park
- Maple Valley
- Medina
- Mercer Island
- Milton
- Newcastle
- Normandy Park
- North Bend
- Pacific
- …15 entries

First plot is the visualization of the house price by bedrooms at each city with interactive points showing information of the city and the price. Next, they are 4 visualizations of the house price by bedrooms, bathrooms, floors and waterfront. These plots show that:

+ In general, the houses that have from 1 to 5 bedrooms, the more bedrooms they had the more price they got. However, there are some houses had more than 5 bedrooms, but they have the lower prices because of their locations.

+ In general, the more bathrooms the houses had the more price they got.
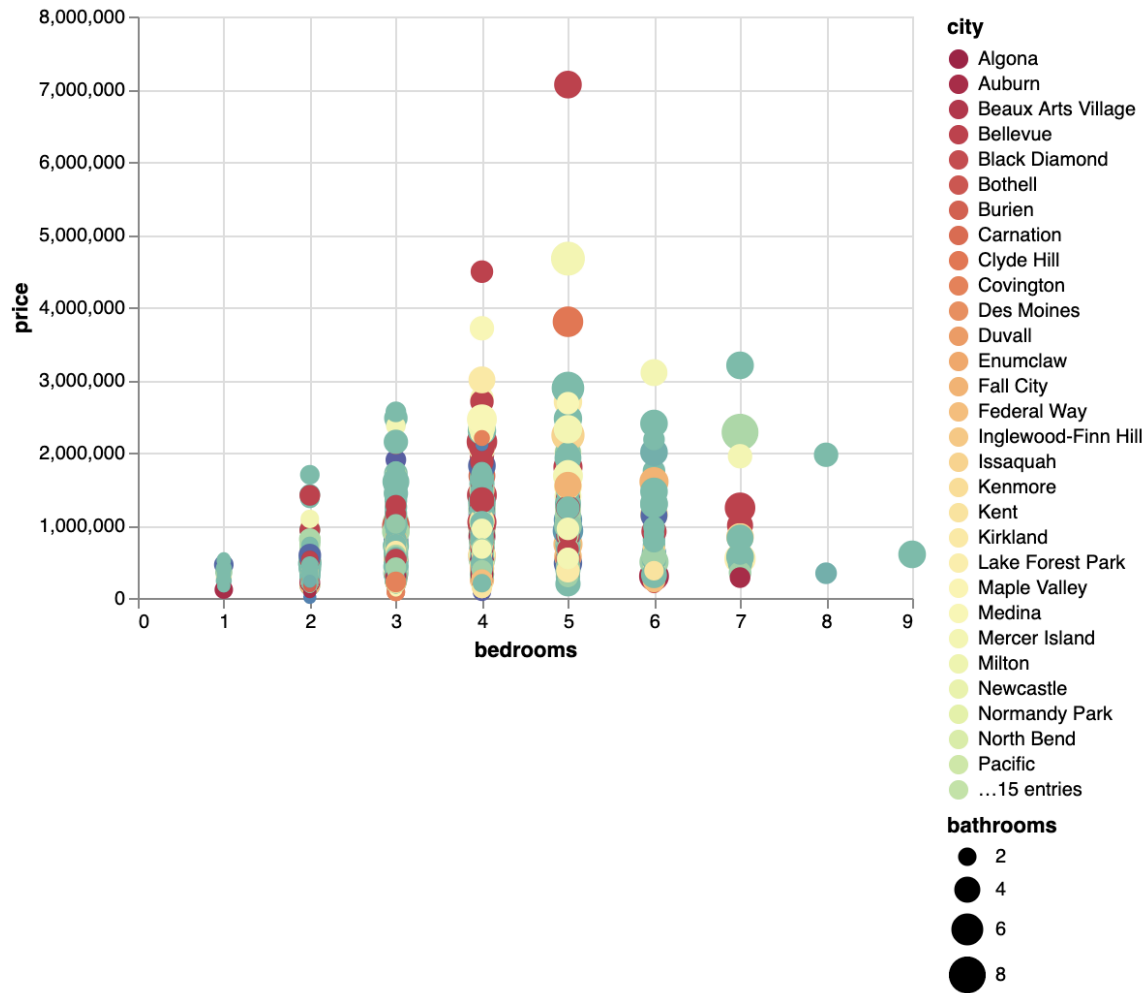
+ The most houses had 2 floors and no waterfront.
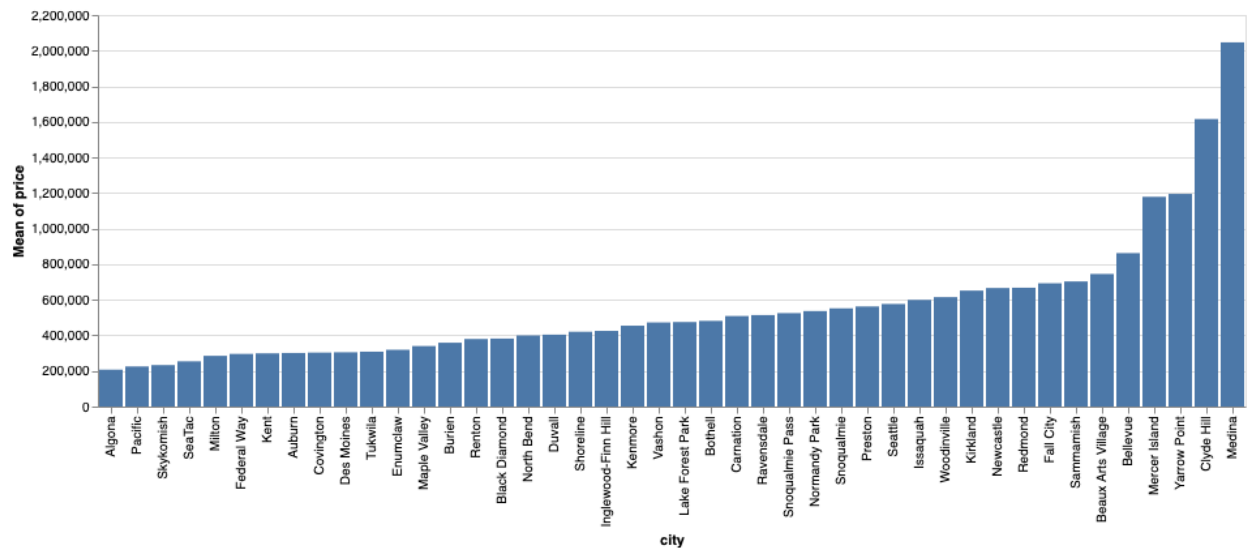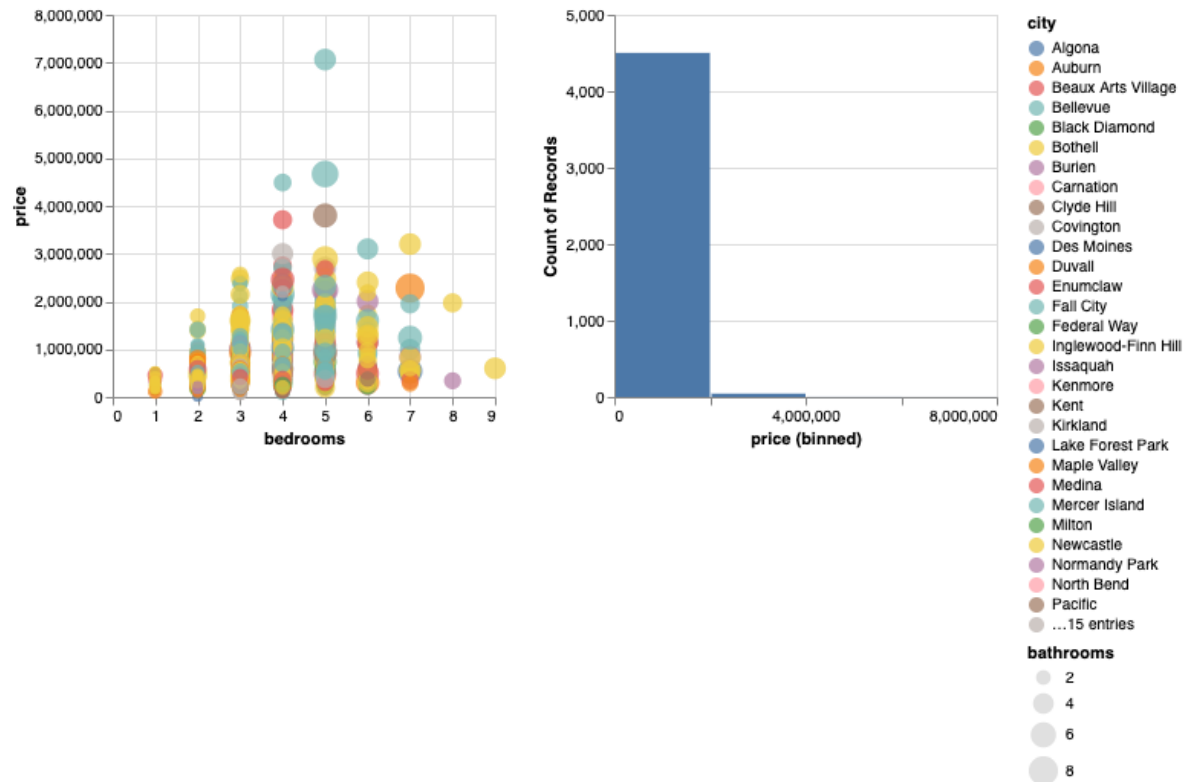
From above visualizations, I can conclude that:

    + Medina is the city has the highest mean of house price and Medina is the city has the lowest mean of house price.

    + The most houses were at price equal or below $2,000,000.

In this plot, I implemented filtering city using dynamic queries. If you select a city, the plot will display the house price points at that city with information of bedrooms, bathrooms and price.
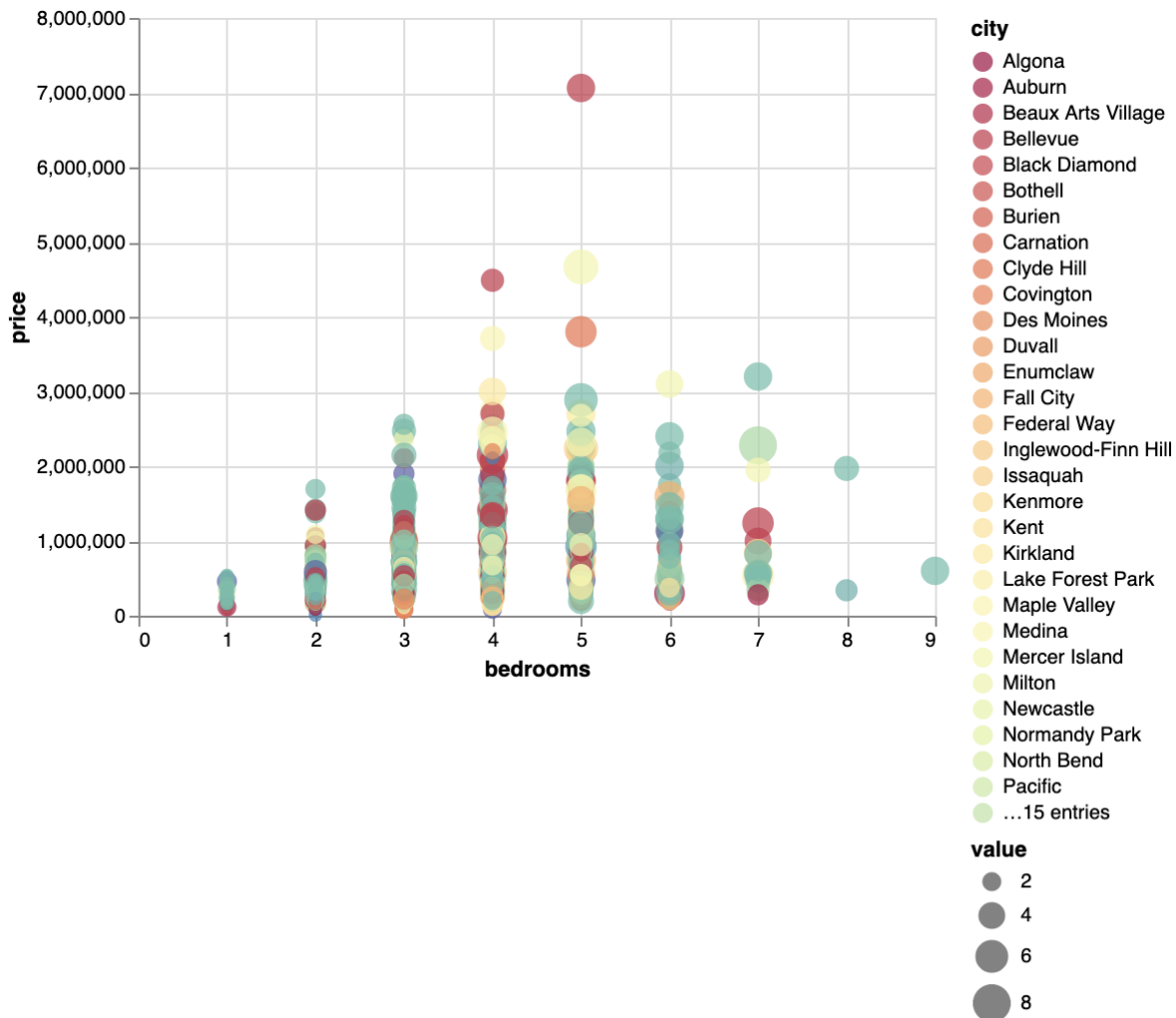
Select a size variable: [bathrooms ˅]

In this plot, I implemented filtering a selection of variables includes bathrooms, floors and waterfront. If you select one of the selections, the plot will display house price points at each city with information of bedrooms and chosen variable.

*A discussion of my final evaluation approach

The target question I want to answer is I should buy, sell or hold a house at a time and what the fair price should be. To answer that question, I will use the journaling study by inviting two friends who are real estate agents to help me evaluate my design. They are professionals and have many experiences can help me successfully navigate the real estate market. Moreover, I also receive the support from family trying to use this design. The measures I would use to answer my data are the listing of house prices were accumulated over time. By my visualizations and my participants knowledge, we'll review the results that we got from that and revisualize for more accuracy. The criteria I would use to indicate my visualization was successful or not if the users can decide buy, sell or hold their houses and they can estimate the price of their properties based on my visualizations.

*A synthesis of my findings

In general, after discussing with them, I received some suggestions about adding more attributes into my design such as: built year, built renovated because they are some important factors that have big influence on the house prices. They like my interactive visualizations because of easy and practical using. I think my approach worked well with the past data. However, in future iterations, I would update data, get more data because the house prices changed too much recently.

This study can help us learn the trend of housing prices in the market and create visualizations to predict upcoming situations based on market trends. Using the results from this report, we can make decisions when to buy, sell and hold properties. This study also designs and implements ways to minimize real estate investment risks, in order to guarantee the stability of the user's overall investment portfolio.