Greater Seattle Property Sale Price

Final Project MSDS 430
Alisher Siddikov

(1) An introduction to your project explaining why the problem you chose is interesting or relevant to you. You should explain the problem in context and detail. In particular, you should explain any variables or formulas used in the problem.

- I am using King County (Greater Seattle) property sale price data and obtained from
- https://www.kaggle.com/harlfoxem/housesalesprediction. It has about 20 property features/details for 21613 residential property transactions between May 2014 and May 2015. I'm interested in knowing if the house sale price correlates to additional amenities such as the number of bedrooms, sq. ft. of living space, and other features.
- Here is the variable explanations:
    - id - Unique ID for each home sold
    - date - Date of the home sale
    - price - Price of each home sold
    - bedrooms - Number of bedrooms
    - bathrooms - Number of bathrooms, where .5 accounts for a room with a toilet but no shower
    - sqft_living - Square footage of the apartments interior living space
    - sqft_lot - Square footage of the land space
    - floors - Number of floors
    - waterfront - A dummy variable for whether the apartment was overlooking the waterfront or not
    - view - An index from 0 to 4 of how good the view of the property was
    - condition - An index from 1 to 5 on the condition of the apartment,
    - grade - An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.
    - sqft_above - The square footage of the interior housing space that is above ground level
    - sqft_basement - The square footage of the interior housing space that is below ground level
    - yr_built - The year the house was initially built
    - yr_renovated - The year of the house's last renovation
    - zipcode - What zipcode area the house is in
    - lat - Lattitude
    - long - Longitude
    - sqft_living15 - The square footage of interior housing living space for the nearest 15 neighbors
    - sqft_lot15 - The square footage of the land lots of the nearest 15 neighbors
- Since there are a lot of features, I was able to down select property features from 21 to 9 (reduced columns by 57%) and able to remove outliers (reduced rows by 5%). I have the following variables:

- date
- price
- bedrooms
- bathrooms
- sqft_living
- waterfront
- view
- geo indicators: latitude and longitude

- Here is the conclusion of my analysis:

  According to the linear regression model, the property sale price is changed by additional amenities such as the number of bedrooms, bathrooms, living space, waterfront, and views:
  Formula/Charts: Linear Regression Model
  - Each one-unit increase in bedrooms causes a decrease of $25,490 in the price.
  - Each one-unit increase in bathrooms causes an increase of $12,946 in the price.
  - Each waterfront property causes an increase of $52,312 in its price.
  - Each one-unit increase in the view of a property causes an increase of $43,819 in its price.
  - Each square-footage increase in the area causes an increase of $174 in the price.

  Price:
  Formula/Charts: pairplot, displot, histogram, mean, median, skewness, kurtosis
  - Most expensive property was sold at a price of $7.7M million
  - Around 2014 and 2015 at King County (greater Seattle Area), the most frequently sold properties were at price range of $300k and $400k.
  - After $400k price, the property availability decreases for each $100k price increase.

  Waterfront:
  Formula/Charts: stripplot, violinplot, boxplot
  - Waterfront properties are more expensive. However, fewer waterfront properties are sold.
  - Most common non-waterfront property price is between $250k and $450k.

  View:
  Formula/Charts: stripplot, violinplot, boxplot
  - The highest view score which is 4 had waterfront properties.
  - However, view 4 is more expensive than waterfront properties.

  Bedrooms:
  Formula/Charts: plot hist, plot bar, violin plot
  - As we can see that 3-bedroom houses were most commonly sold followed by 4 bedrooms.
  - Builders can make a new building with more 3 and 4 bedroom’s to attract more buyers.
  - So now we know that 3 and 4 bedrooms were the highest selling.
  - Most of the bedrooms were sold between the price range of $200k and $600k.

Bathrooms:
Formula/Charts: plot hist, plot bar, violin plot

- As we can see that 2.5-bathroom houses are most commonly sold followed by 1 bathroom.
- Builders can make a new building with more 2.5 and 1 bathrooms to attract more buyers.
- So now we know that 2.5 and 1 bathrooms are the highest selling.

Living Area:
Formula/Charts: jointplot

- Around 2014 and 2015 at King County (greater Seattle Area), the most frequently sold properties were at living space range of 1500 sqft and 2000 sqft.
- After 2000 sqft, the property availability decreases for each additional 500 sqft increase.
- It makes sense that people would pay for the more living area.
- Most common living area is between 1000 and 2500 sq. ft.
- Most common price range is between $250k and $500k.
- As we can see that 2.5-bathroom houses are most commonly sold followed by 1 bathroom.

Time:
Formula/Charts: datetime, pivot_table, plot bar

- The slowest months are during fall and winter seasons. April, June, and July months are the busiest times.

Location:
Formula/Charts: jointplot, folium.Map, lmplot

- We can find out ideal locations and here are the denser locations where most of properties are sold.
- There are many properties sold at these latitudes: around 47.7 and between 47.6 and 47.5.
- Also, there are many properties sold at these longitudes: between -122.4 to -122.2
- Most of the properties were sold in West Seattle, North Seattle, and close to the waterfront.
- The most expensive neighborhoods were North Seattle, University of Washington, Mercer Island, Medina, Bellevue, Redmond, and properties along the waterfront area.
- Least expensive neighborhoods were South Seattle, further North Seattle, and the properties away from water.

(2) A very brief description of your algorithmic process (your approach) for solving the problem. Your code, of course, should implement this algorithm (thought process).

- I was able to use feature extraction and percentile to extract important features of the dataset and now I have the following features date, price of sold property, number of bedrooms and

bathrooms, square footage of living area, waterfront properties, views, and geographic indicators (latitude and longitude).

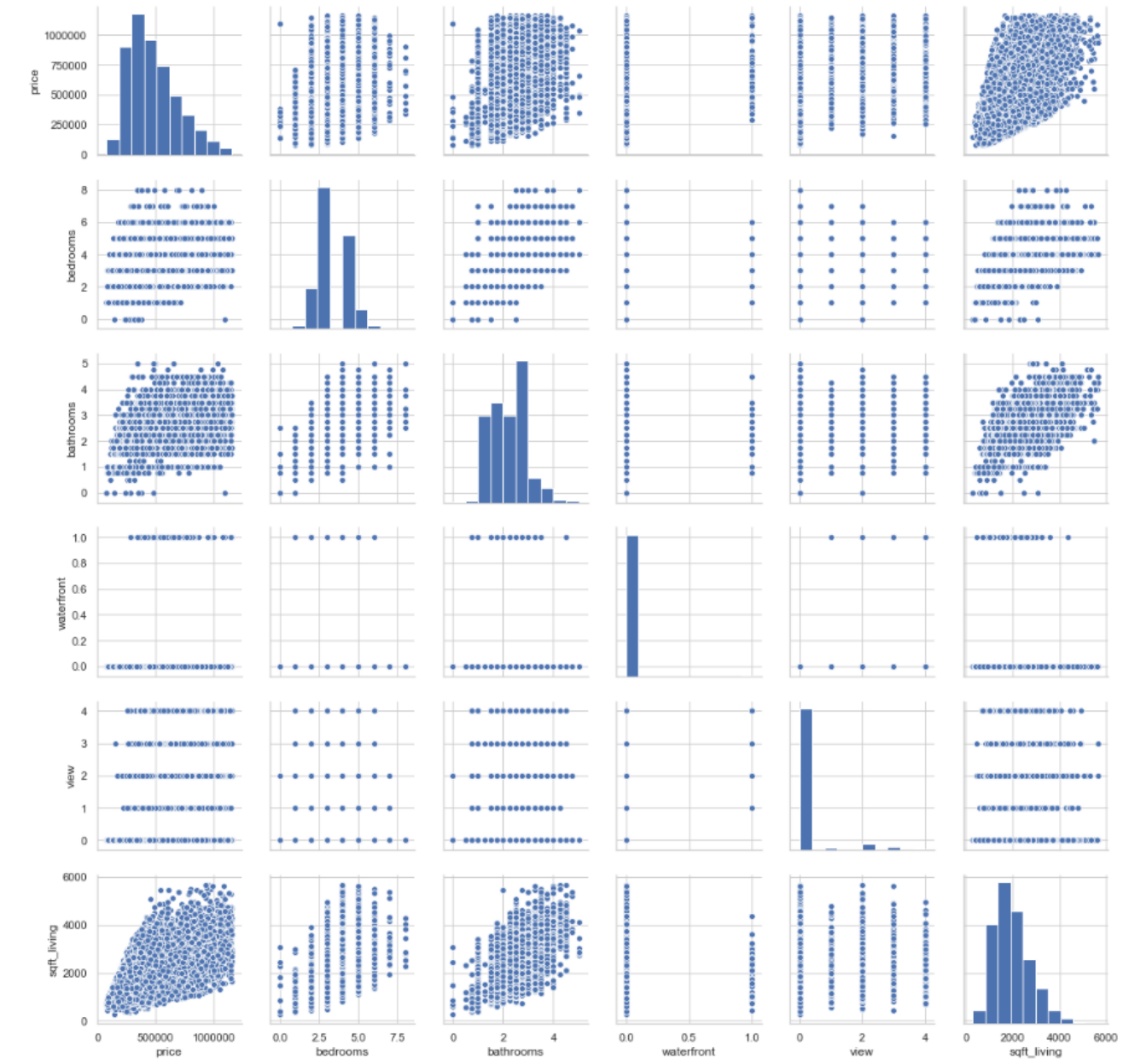
- I was also able to remove extreme outlier values
- After I prepared the data for analysis, I started with correlation analysis and histograms to understand the dataset better. I used the following packages:
    - Seaborn
        - sns.pairplot(house_df4_analysis, kind="scatter")
        - sns.distplot(house_df['price'])
        - sns.violinplot(x='view', y='price', hue = 'waterfront', data = house_df4, scale = 'count', split = **True**)
        - sns.jointplot(x=house_df4['sqft_living'], y=house_df4['price'], kind='reg')
        - sns.lmplot(x='long', y='lat', hue='price_range', data=house_df4, fit_reg=**False**, height=10)
    - Pandas
        - pd.read_csv('kc_house_data.csv')
        - pd.qcut(house_df4.price, 5, labels=["min", "25%", "50%","75%", "max"])
        - pd.to_datetime(house_df4.date, format='%Y%m**%d**', errors='ignore')
        - pd.DataFrame(reg.coef_, X.columns, columns=['Coefficient'])
    - NumPy
        - np.round(house_df_analysis.describe())
        - np.percentile(house_df.price, 95)]
        - np.mean(house_df.price)
        - np.median(house_df.price)
    - Matplotlib
        - plt.subplots(figsize=(12, 8))
        - plt.xlabel('price without outliers', fontsize = 15)
        - plt.show()
- I also run linear regression model to understand price and property feature correlation relationship
    - reg = LinearRegression()
    - reg.fit(X, y)
    - pd.DataFrame(reg.coef_, X.columns, columns=['Coefficient'])
- I used bar plot, hist plot, joint plots, boxplot, strip plot, violin plot, and lm plot for further analysis.

(3) Any relevant results such as tables, graphs, etc. along with their purpose and description.

- Statistical description of data (outliers has not yet removed at this stage).

	price	bedrooms	bathrooms	waterfront	view	sqft_living
count	21613.0	21613.0	21613.00	21613.0	21613.0	21613.0
mean	540182.0	3.0	2.00	0.0	0.0	2080.0
std	367362.0	1.0	1.00	0.0	1.0	918.0
min	75000.0	0.0	0.00	0.0	0.0	290.0
25%	321950.0	3.0	2.00	0.0	0.0	1427.0
50%	450000.0	3.0	2.00	0.0	0.0	1910.0
75%	645000.0	4.0	2.00	0.0	0.0	2550.0
max	7700000.0	33.0	8.00	1.0	4.0	13540.0
median	450000.0	3.0	2.25	0.0	0.0	1910.0

- Correlation and histogram analysis



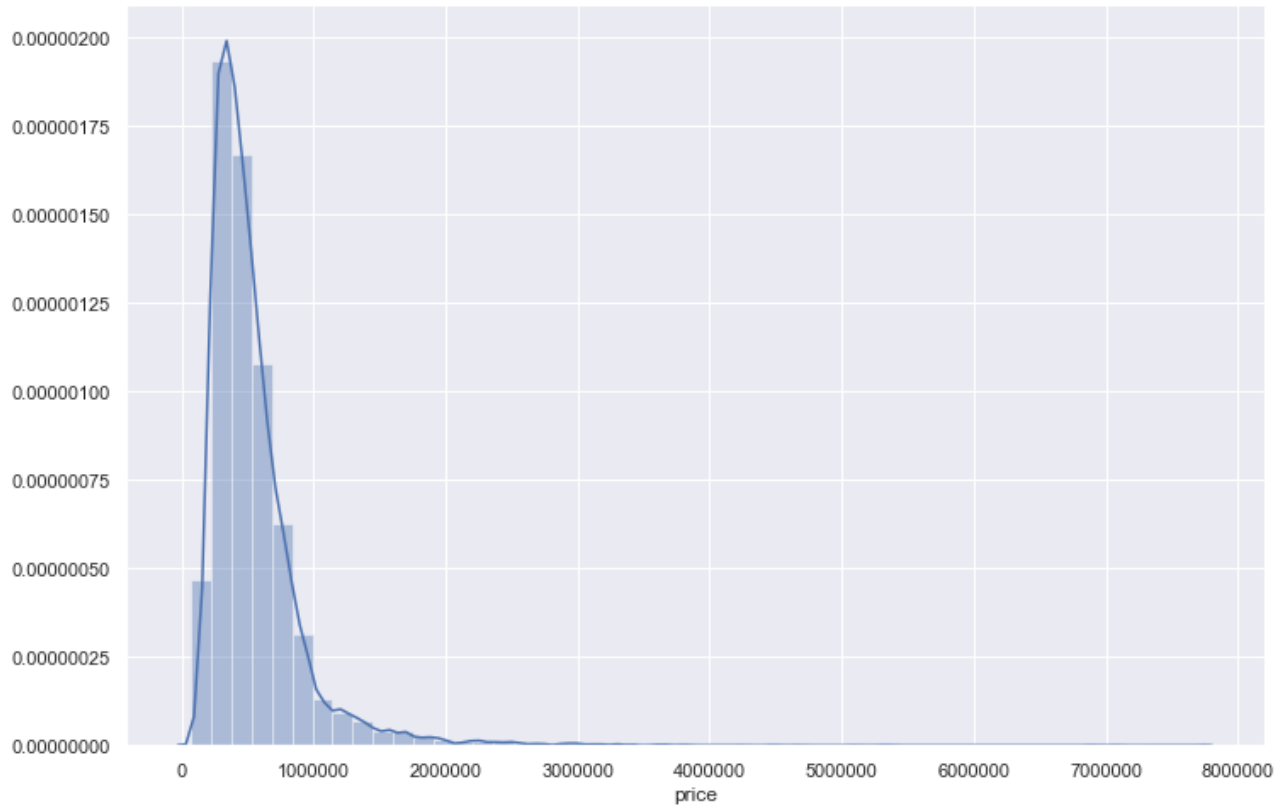
- Linear Regression Analysis

	Coefficient
bedrooms	-25490.210923
bathrooms	12945.978668
waterfront	52312.229508
view	43819.312150
sqft_living	174.350976

- Each one-unit increase in bedrooms causes a decrease of $25,490 in the price.
- Each one-unit increase in bathrooms causes an increase of $12,946 in the price.
- Each waterfront property causes an increase of $52,312 in its price.
- Each one-unit increase in the view of a property causes an increase of $43,819 in its price.
- Each square-footage increase in area causes an increase of $174 in the price.

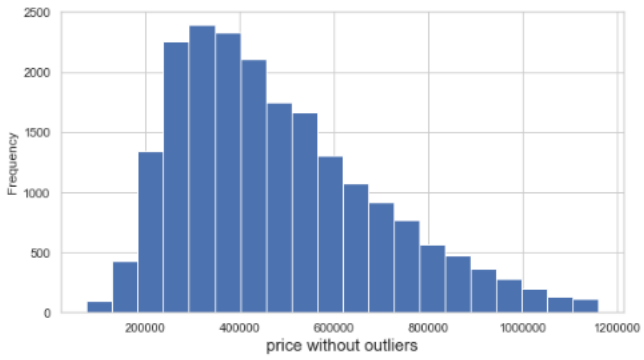
- Price chart with outliers

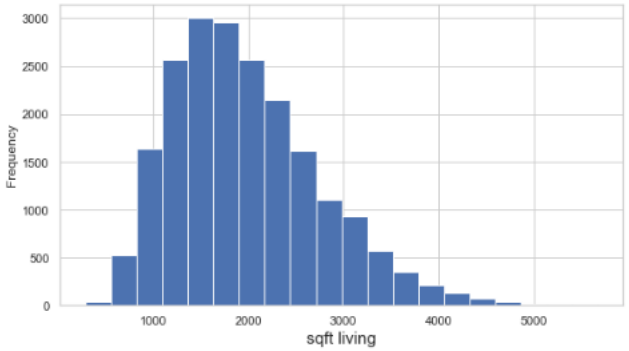
```
Skewness of price: 4.021716
Kurtosis of price: 34.522444
Average price paid = $ 540182
Median price paid  = $ 450000
```



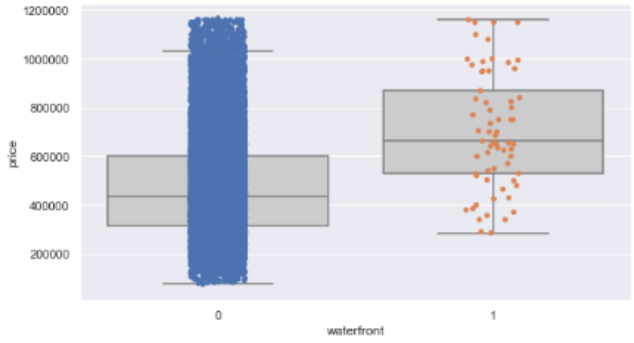
- Price chart with outliers

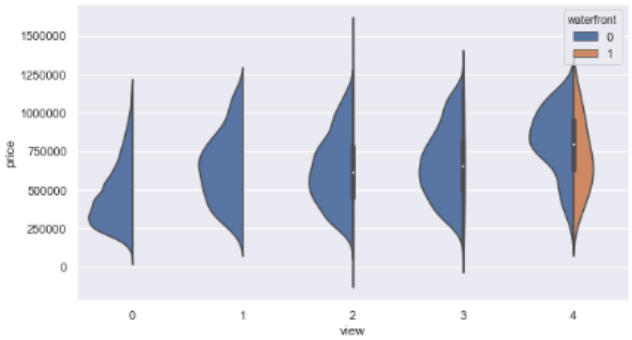
```
Skewness of price: 0.793138
Kurtosis of price: 0.129848
Average price paid = $ 479105
Median price paid  = $ 438400
```



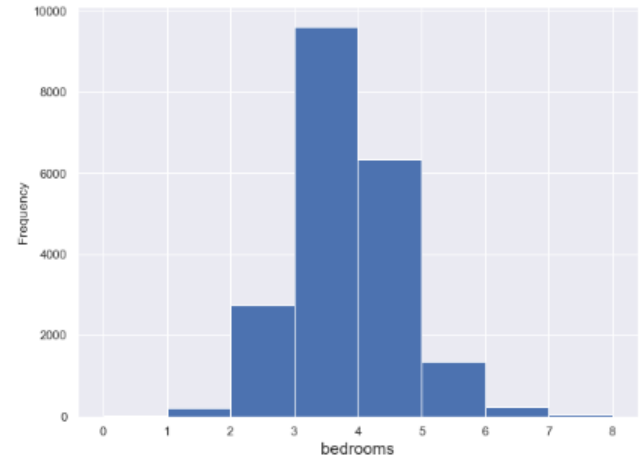


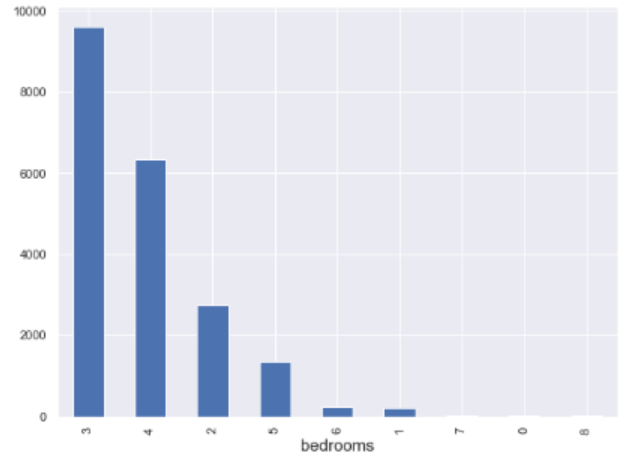
- Price vs. Waterfront & View

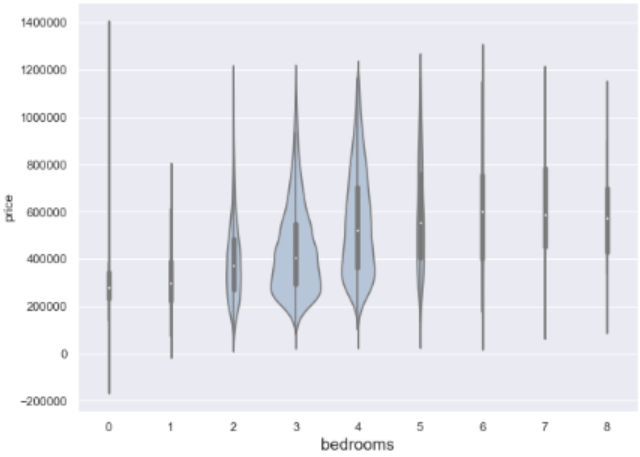


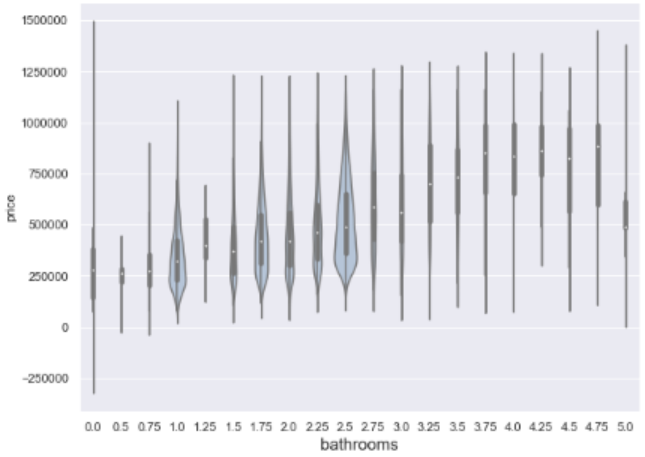


- Price vs. Bedrooms & Bathrooms

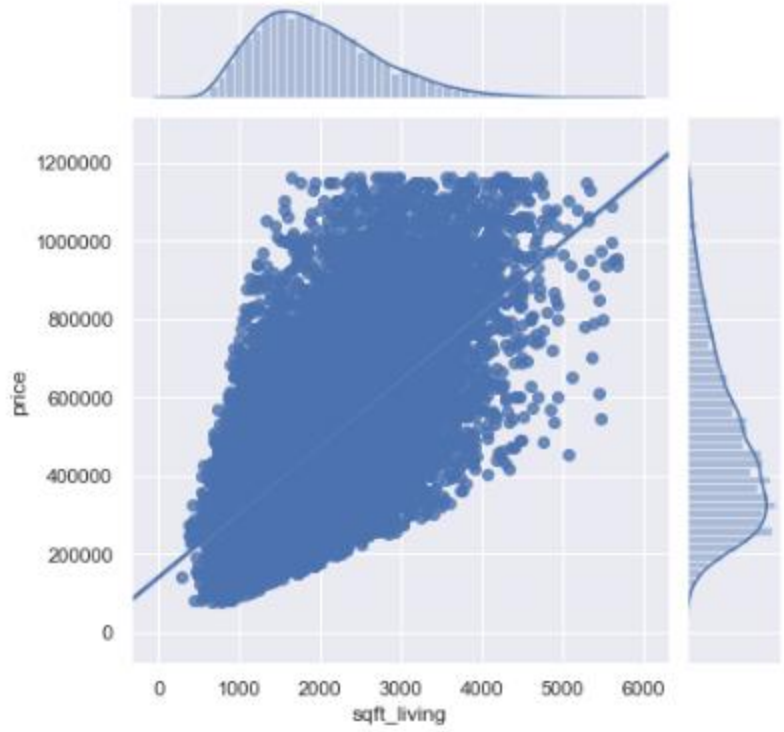






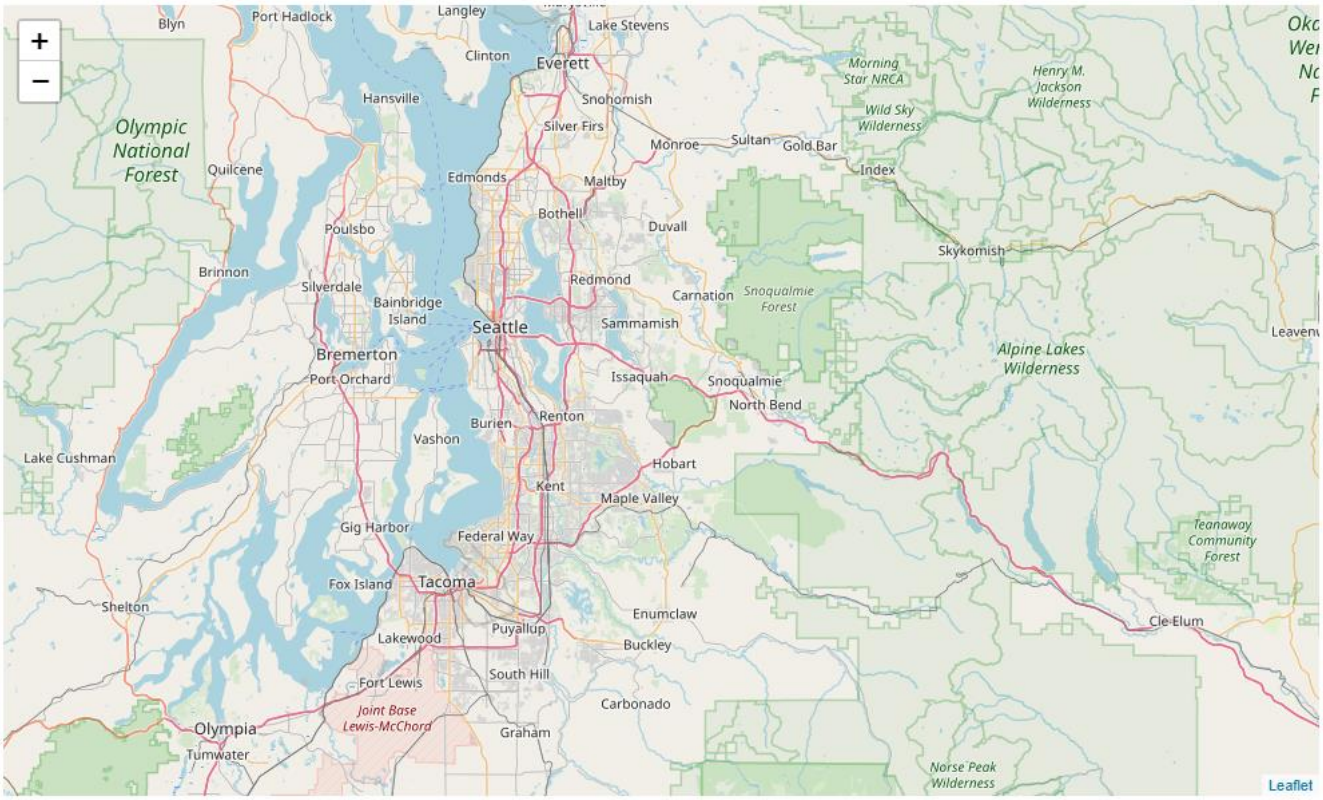


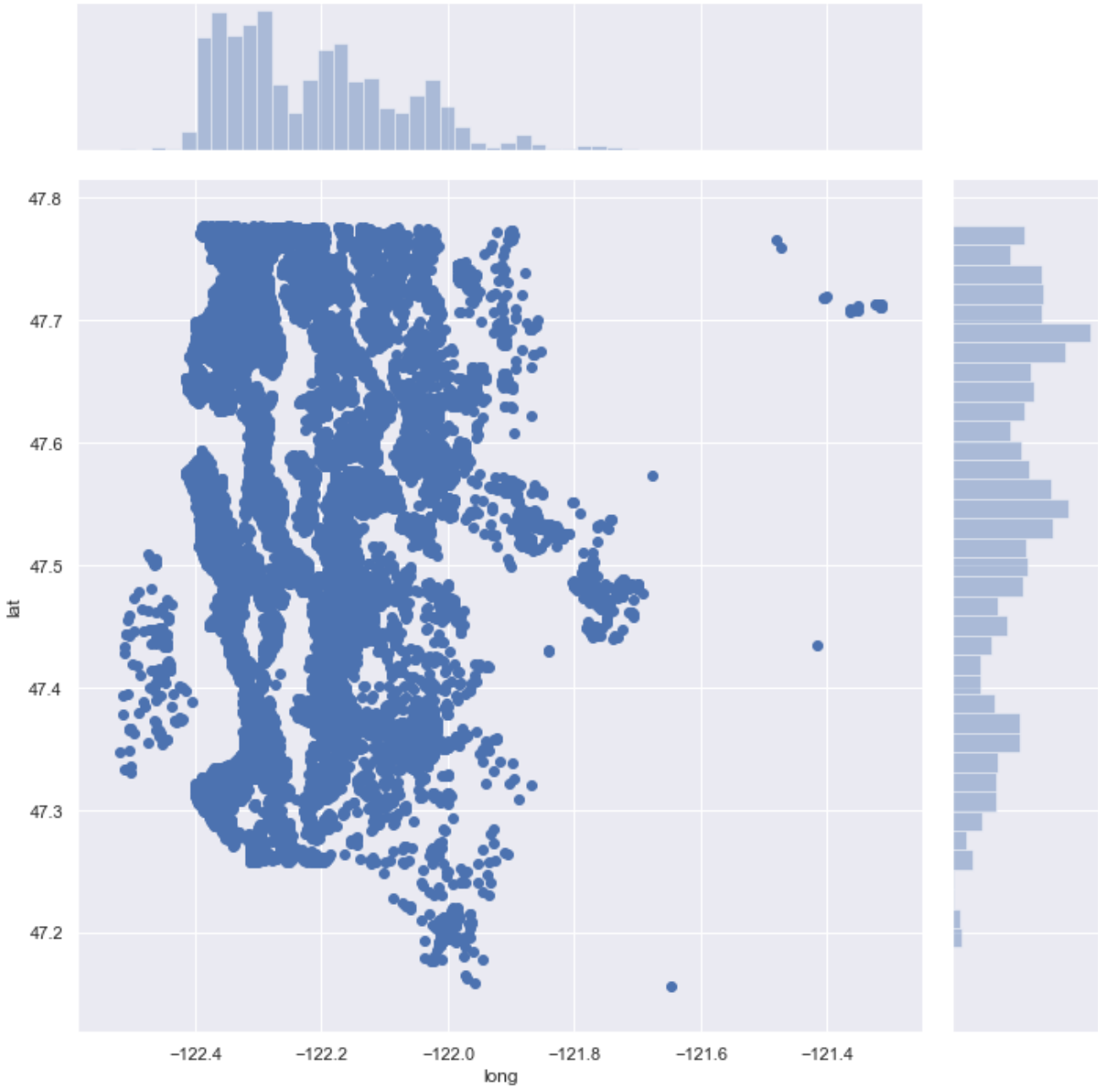
- Price vs. Living Room

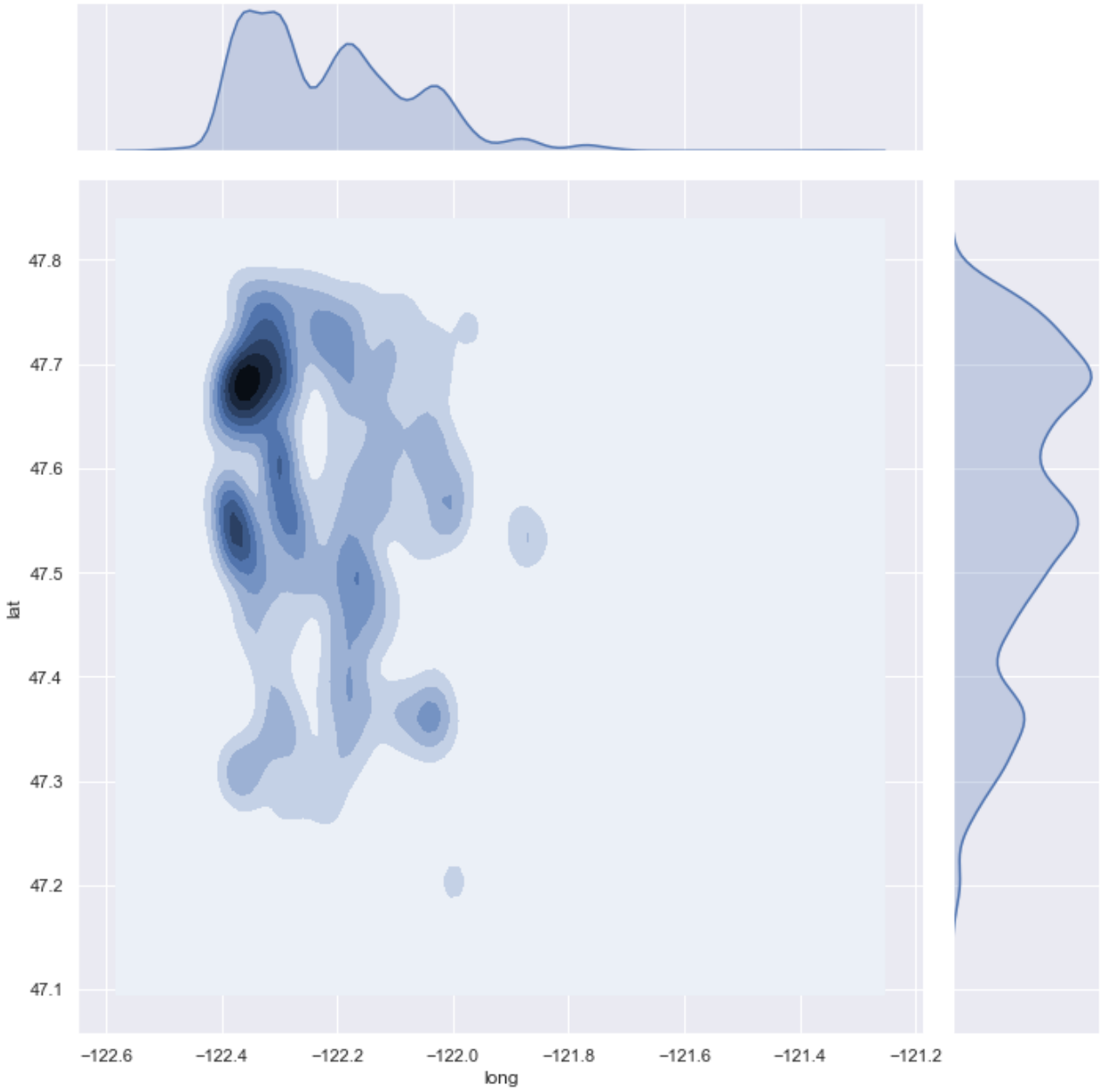


- Price vs. Date

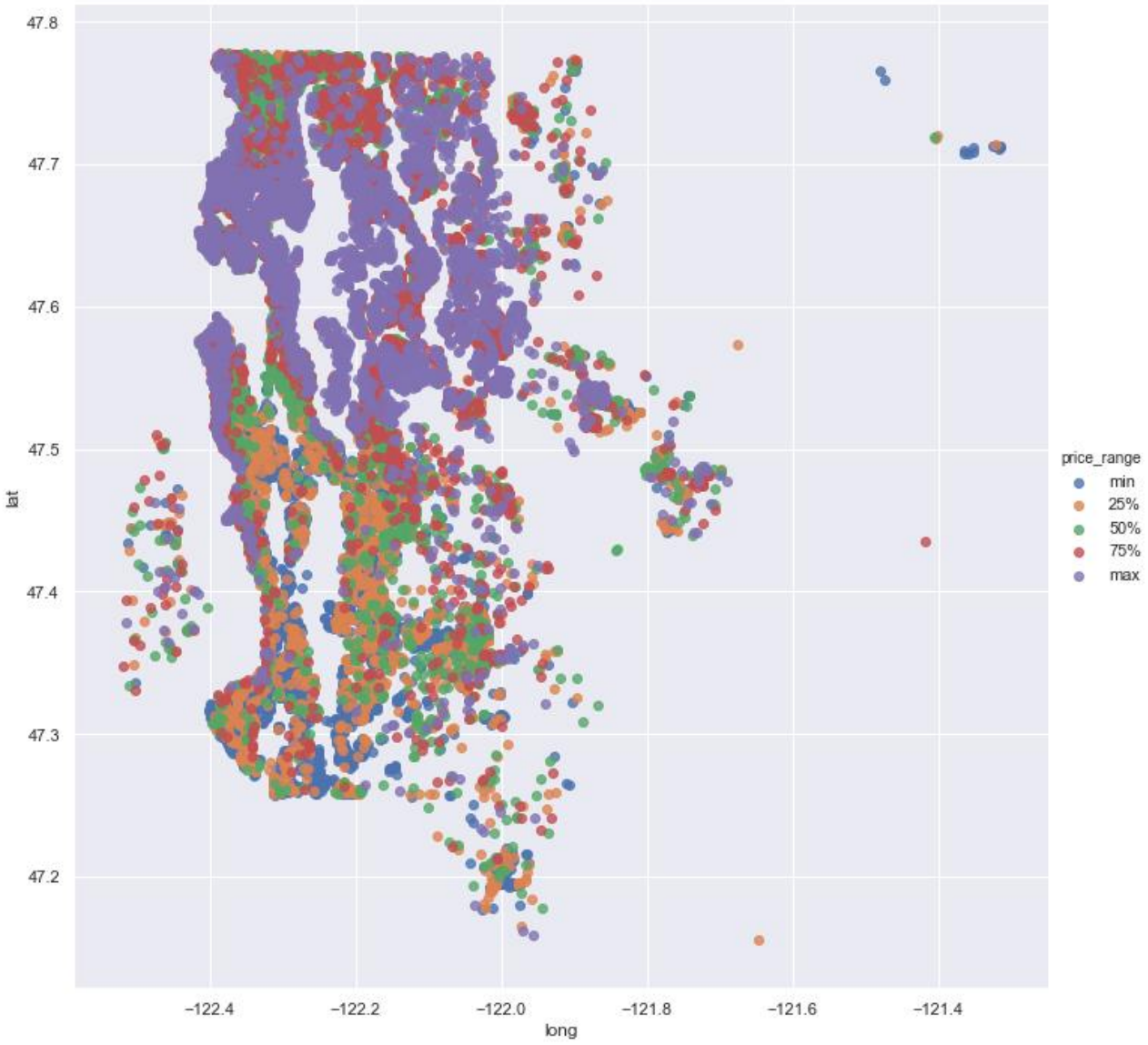
- **Price vs. Location (density)**


Blyn
Port Hadlock
Langley
Lake Stevens
Clinton
Everett
Hansville
Snohomish
Silver Firs
Monroe
Sultan
Gold Bar
Index
Morning Star NRCA
Wild Sky Wilderness
Henry M. Jackson Wilderness
Olympic National Forest
Quilcene
Edmonds
Maltby
Bothell
Duvall
Poulsbo
Skykomish
Brinnon
Silverdale
Redmond
Carnation
Snoqualmie Forest
Bainbridge Island
Seattle
Sammamish
Bremerton
Port Orchard
Issaquah
Snoqualmie
North Bend
Alpine Lakes Wilderness
Renton
Burien
Vashon
Lake Cushman
Hobart
Kent
Maple Valley
Gig Harbor
Federal Way
Teanaway Community Forest
Fox Island
Tacoma
Shelton
Enumclaw
Cle Elum
Puyallup
Lakewood
Buckley
South Hill
Fort Lewis
Carbonado
Joint Base Lewis-McChord
Olympia
Graham
Tumwater
Norse Peak Wilderness
Leaflet


47.8
47.7
47.6
47.5
47.4
47.3
47.2
lat
−122.4
−122.2
−122.0
−121.8
−121.6
−121.4
long



- **Price vs. Location (Heatmap)**


47.8
47.7
47.6
47.5
47.4
47.3
47.2
lat
−122.4
−122.2
−122.0
−121.8
−121.6
−121.4
long
price_range
min
25%
50%
75%
max