Analysis of Sold Properties Data (King County, WA (May 2014-May 2015)

INCREASE YOUR PROPERTY VALUE AND STRETCH YOUR HOUSE BUYING BUDGET FURTHER

FLATIRON SCHOOL

Phase 1 Final Project

Authors: Elena Kazakova

Cohort: DS02222021

Instructor: James Irving

Summary

This Project analyses the data on the properties sold in King County, WA (not including the Seattle inner-city areas) over one year, from May 2014 till May 2015. The resulting model provides insight into what features of a property increase its' value and what variables outside of property owners' control affect the price of a house.

The Project is built using OSEMN workflow methodology for Inferential Data Analysis (estimating parameters of a numeric model)

Outline of the presentation

- Business Problem
- Data
- Final Model
- Results
- Conclusions

Business Problem

This project is the Inference Analysis project of King County, WA house prices, and various factors that might affect the sales price. This study aims to build a model(s) of house sale prices depending on the features of the property in the dataset provided. This information can be helpful for house owners, house buyers, and real estate agents in the county. This project's scope is limited, and further in-depth analysis would be beneficial; however, the analysis results reveal several valuable recommendations to house owners and house buyers.

Data

The dataset used in this project has been downloaded from KAGGLE site. The dataset includes the information about properties sold in King County of Washington State between **May 2014** and **May 2015**. The area consists of Seattle city area but does not include the inner city. The dataset contains **21597** records and has **21** dependent and independent variables. The description of the data is as follows:

id - Unique ID for each home sold

date - Date of the sale

Between: May 2014 and May 2015

price - Price of each home sold

Minimum price: 78000
Maximum price: 7700000

bedrooms - Number of bedrooms

Values between 1 and 33

bathrooms - Number of bathrooms, where .5 accounts for a room with a toilet but no shower

Values between 0.5 and 8.0

sqft_living - Square footage of the house interior living space

Minimum value: 370
Maximum value: 13540

Data (continued)

view - A categorical variable describing how good the view of the property was

Values: 1.0, 2.0, 3.0, 4.0

condition - A categorical variable describing the condition of the house

Values: 1, 2, 3, 4, 5

grade - A categorical variable describing the quality of construction, from 1 to 13; 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.

Values between 3 and 13

sqft_above - The square footage of the interior housing space that is above ground level

Minimum value: 370
Maximum value: 9410

sqft_basement - The square footage of the interior housing space that is below ground level

Minimum value: 370
Maximum value: 9410

yr_built - The year the house was initially built

Minimum value: 1900
Maximum value: 2015

Modifications to the Data

- Added a new variable **'month'** – a month of a year a property was sold
- Added a new variable **'distance'** – a distance at which a property is located from the center of Seattle
- Added a new variable **'basement_exists'**, values **1 and 0** – an indicator if a property has a basement
- Added a new variable **'renovation_done'** with values **[0,1,2,3,4]**
  - **0** representing renovation never done on houses more than 9 years old (built between 2015 and 2006)
  - **1** representing renovation done more than or equal 50 years ago
  - **2** representing renovation done between 30 and 49 years ago
  - **3** representing renovation done between 29 and 10 years ago
  - **4** representing renovation done between 9 and 1 year ago OR houses built less or equal 9 years ago (built between 2015 and 2006)

Modifications to the Data (continued)

- Removed **'zipcode'** variable because there are better indicators of a location that can be used
  - Zipcodes boundaries are usually drawn out of convenience for postal services or other more formal reasons than geographic location
- Removed **'id'** field, it is not useful
- Removed **extreme values** from **'sqft_lot', 'sqft_lot15', 'bedrooms', and 'price'**
  - Price < 1.5 M
  - Number of bedrooms < 9
  - Number of bathrooms < 5.5
  - Number of floors < 3.5
  - Square footage of a property lot between 100 and 40000
  - Square footage of an average neighborhood lot between 100 and 40000
  - Distance < 30 miles
- Removed the original **'sqft_basement'** variable
- Removed the original **'lat' and 'long'** variables
- Removed the original **'date'** field

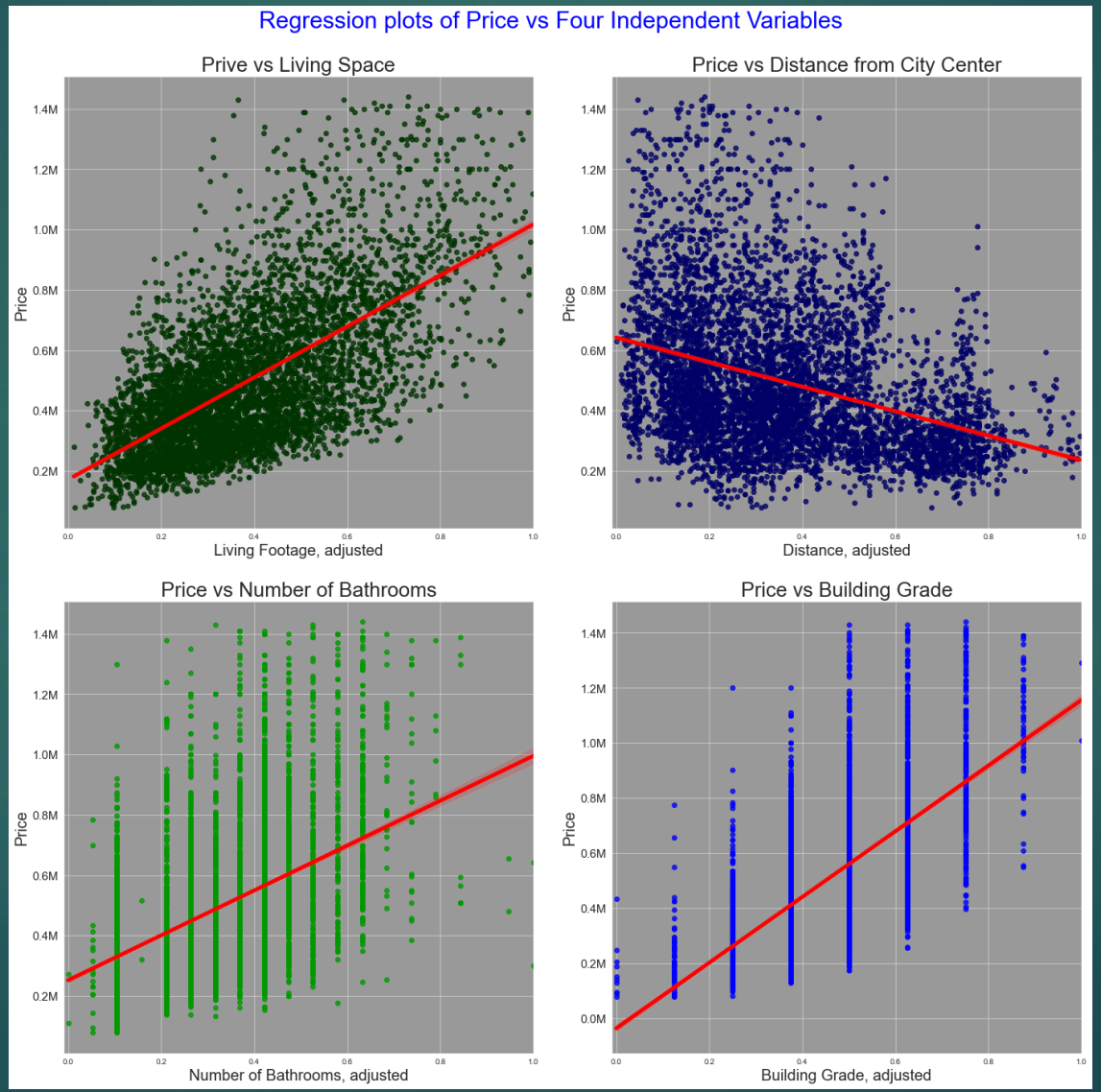
Final Model

"It's better to solve the right problem approximately than to solve the wrong problem exactly"
Quote by John Tukey

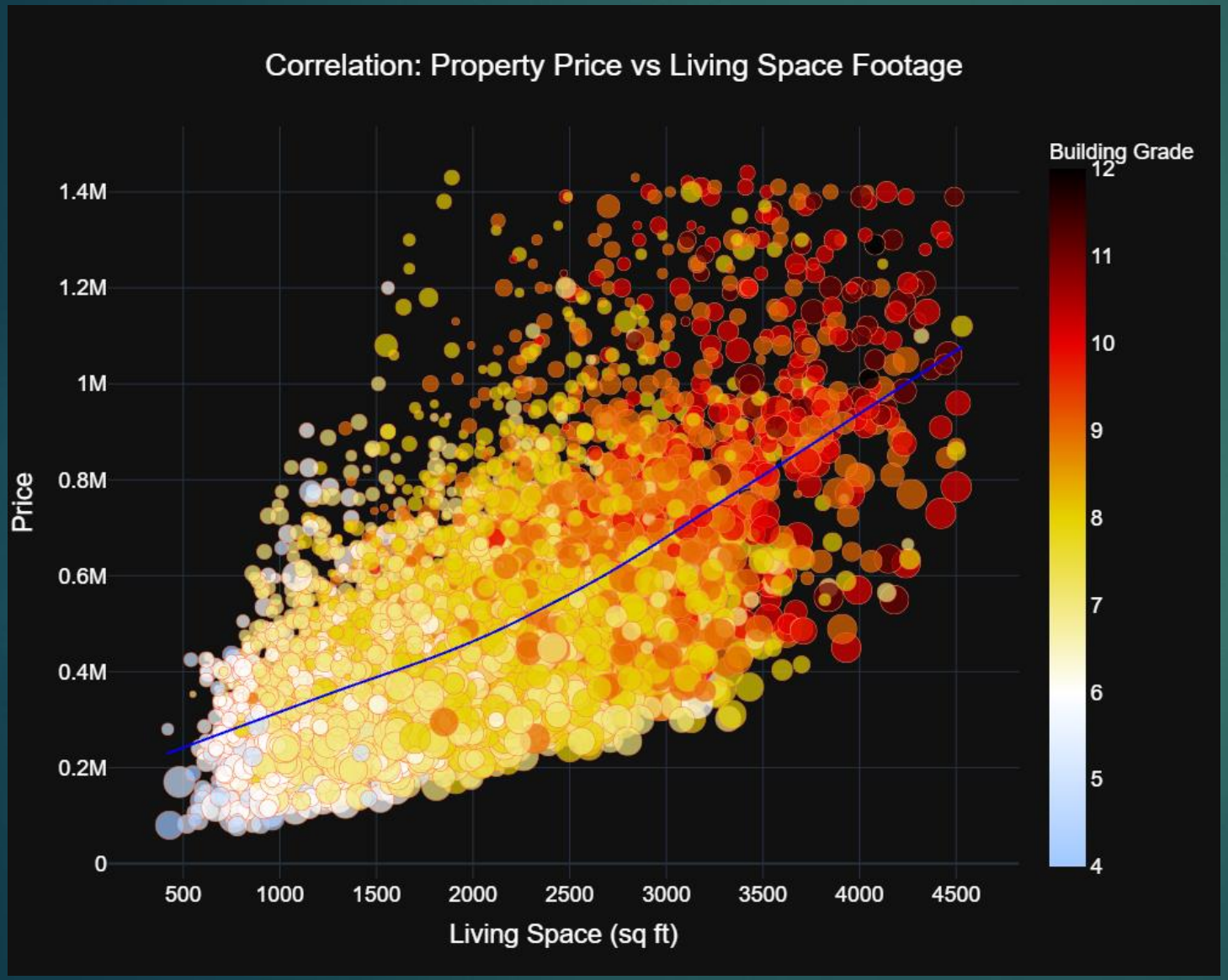
In the Project several models were tested, and a model with the best statistical characteristics was chosen as the final one. The model is a Multiple Linear Regression model. It means that the predictive variables have a linear correlation with the target – **a price** of a sold property

$$\ln(Price)=12.366+1.265\cdot(grade)+1.080\cdot(sqft\_living)-0.989\cdot(distance)+0.060\cdot(bathrooms)$$
$$-0.035\cdot(recent\_renovation\_new)$$

Visualizing linear Regression Model



Visualizing the Correlations



Marker color represents **the building grade**

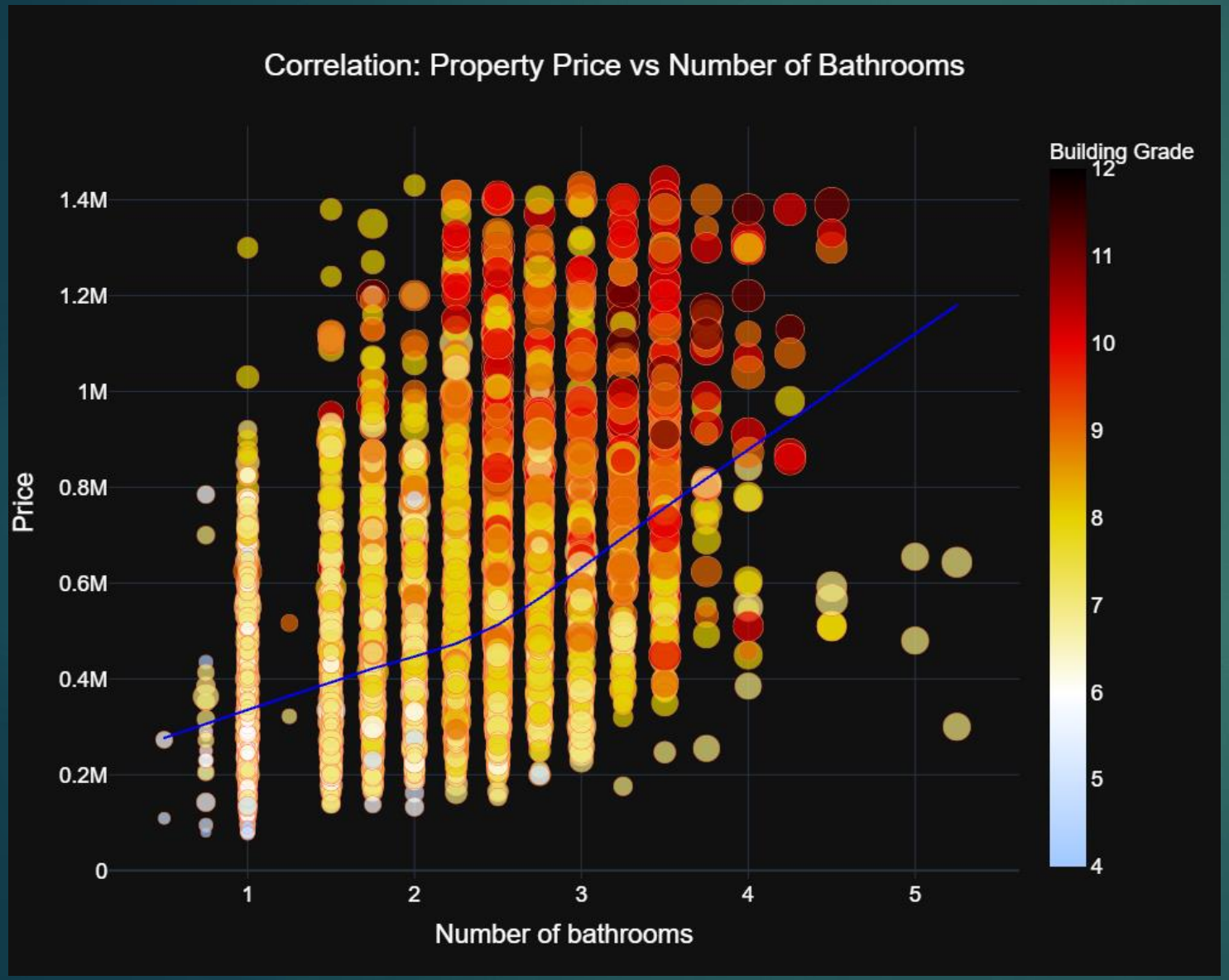
Marker size represents **the distance from the center of the city**

Visualizing the Correlations

Marker color represents **the building grade**

Marker size represent **the number of bathrooms**

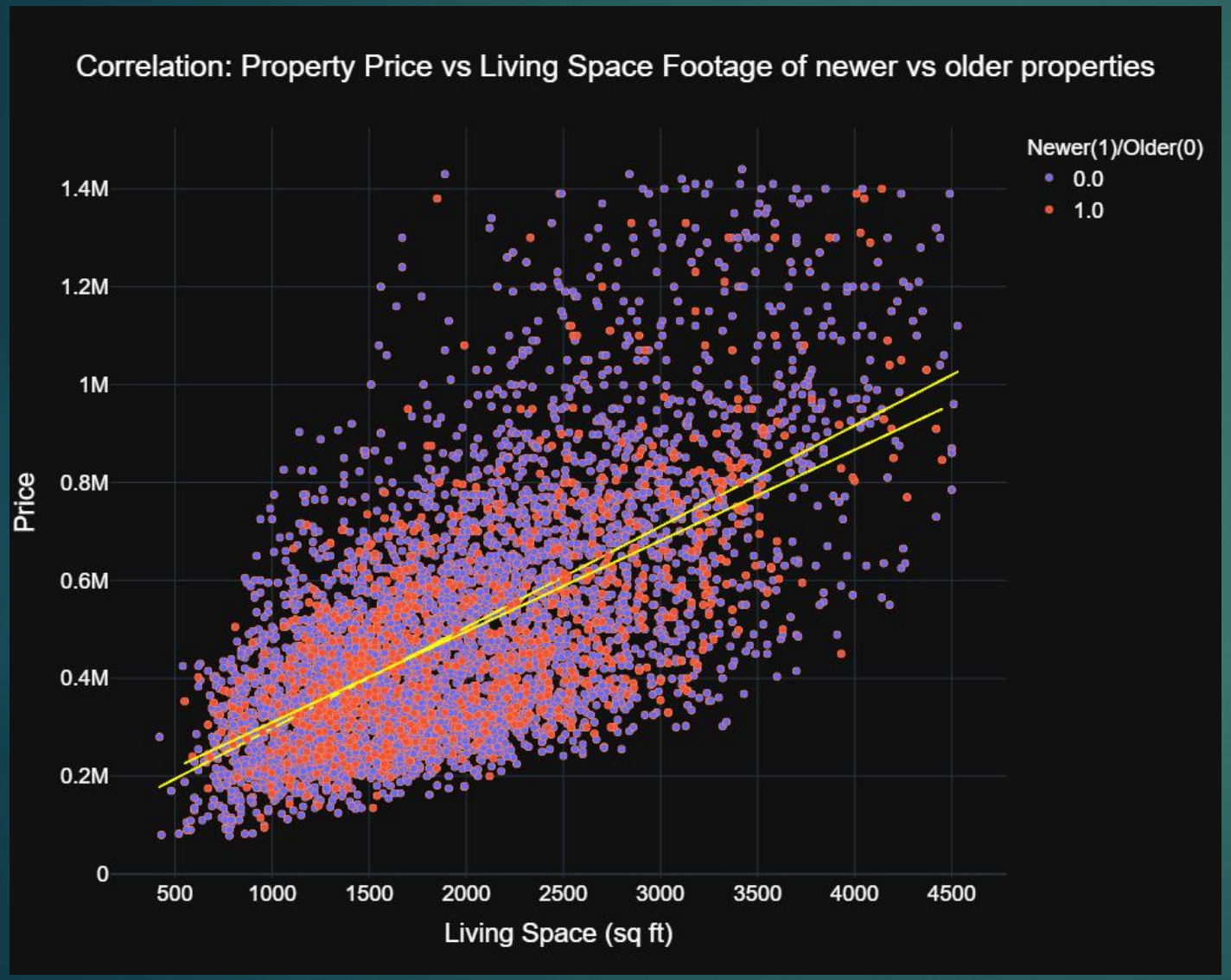
Visualizing the Correlations



Marker color represents **the building grade**

Marker size represents **the living space footage**

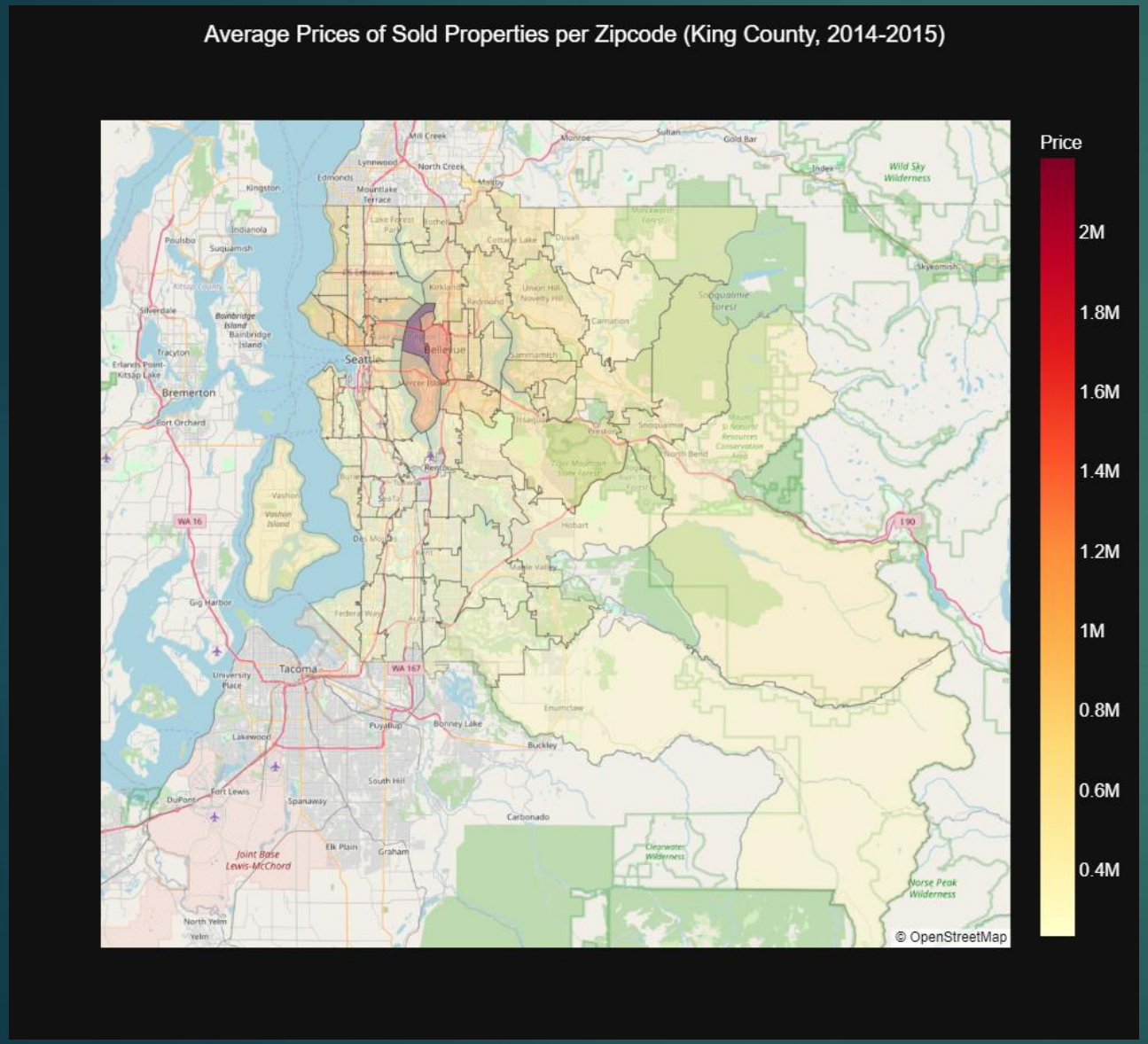
Visualizing the Correlations



Marker color represents if a property has been

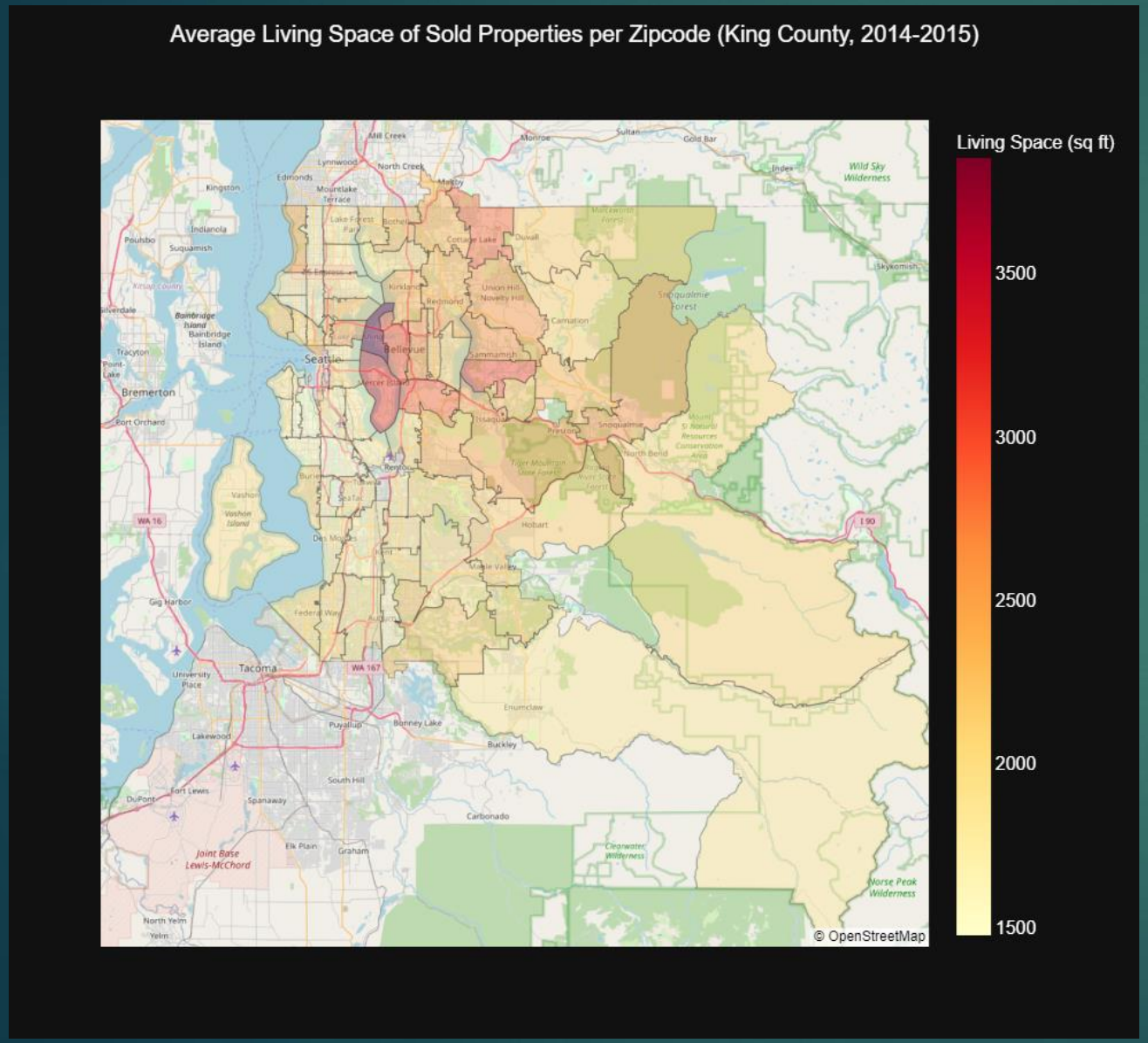
renovated or built between 2006 and 2015

Visualizing the Differences between Zipcodes



- Several zipcodes around Bellevue area are the most expensive ones
- There are other zipcodes that are as close to the center of the city but have better pricing and comparable characteristics of properties

Visualizing the Differences between Zipcodes



- Given that zipcodes around Bellevue area are the most expensive ones

If an amount of living space is essential to you, it is better to look for properties in the zipcodes that have **comparable living space footage** but **lower prices**

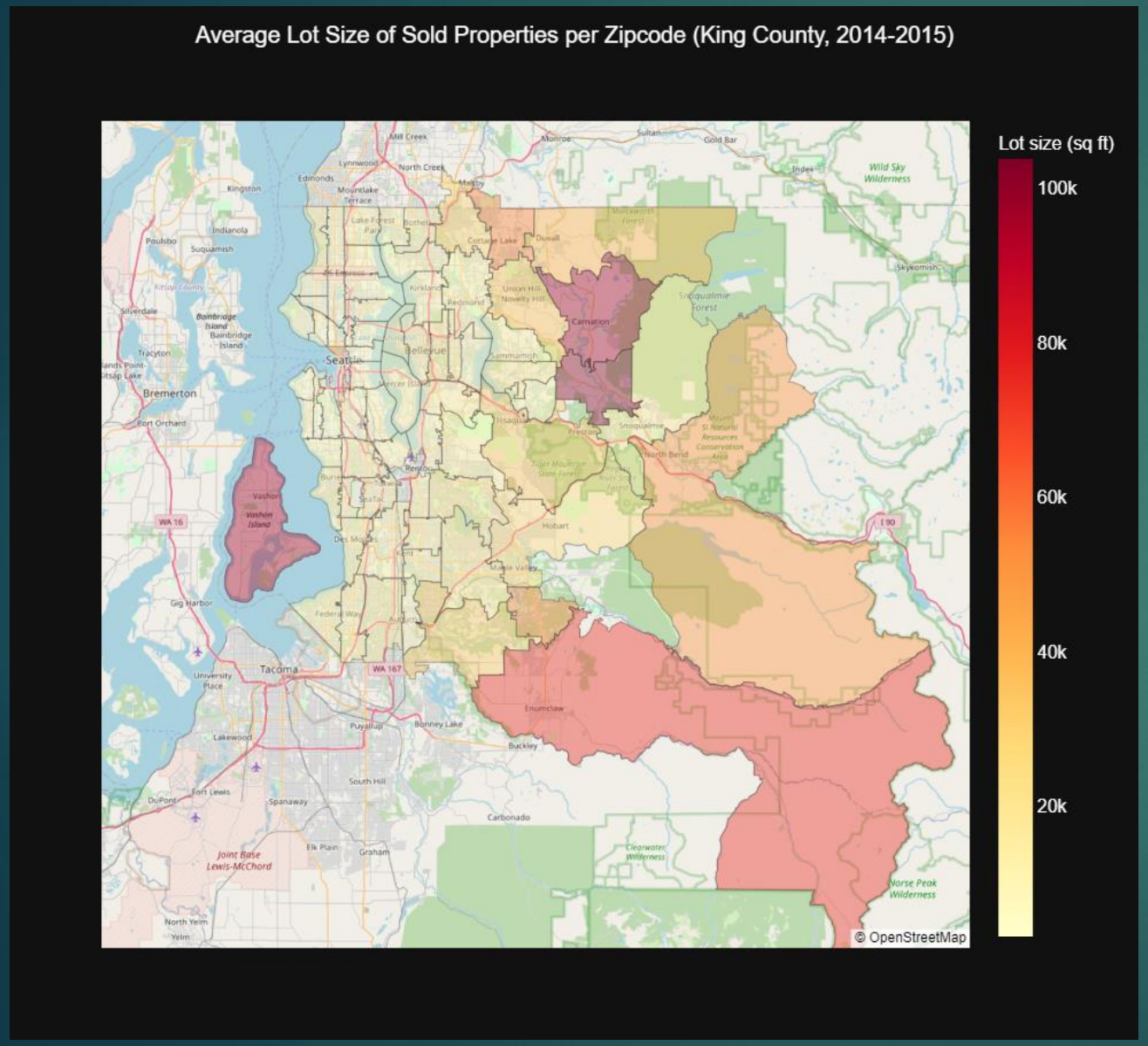
Zipcodes with similar living space footage as Bellevue area properties

Zipcodes 98077

Zipcode 98075

Zip[code 98006

Visualizing the Differences between Zipcodes

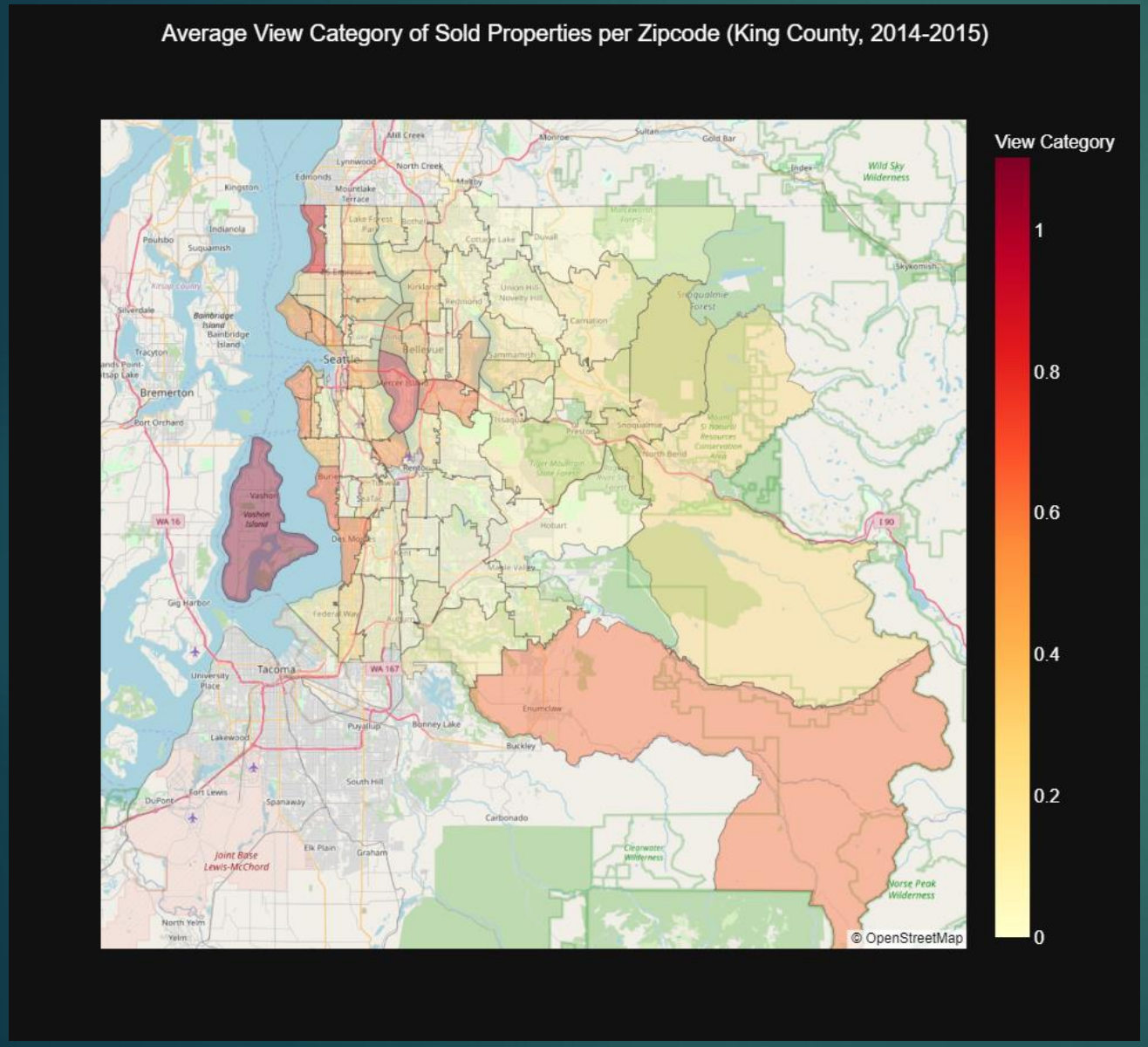


If a size of a property lot is essential to you, it is better to look for properties in the zipcodes that have bigger lots but **comparable prices** with other zipcodes

Zipcodes with bigger lots

- Zipcodes 98014
- Zipcode 98024
- Zipcode 98022
- Zip[code 98070

Visualizing the Differences between Zipcodes



If a view is essential to you, it is better to look for properties in the zipcodes either closer to water or in the mountain areas. Properties in these zipcodes have a better view but **comparable prices** with other zipcodes

Zipcodes with a better view

- Zipcodes 98198
- Zipcode 98166
- Zipcode 98022
- Zipcode 98070
- Zipcode 98177

Recommendations

Recommendations to property owners planning a renovation to their properties:

- Increase the living space of your property
- Do the renovation with higher building quality
- Consider adding a bathroom

Recommendations to potential home buyers:

- Look for properties further away from the city center to make the best out of your property buying budget
- Properties in some zipcodes of the city are more affordable than others at the same distance from the city center
- Properties in some zipcodes of the city are more affordable than others, with a better view, with more considerable property lots, and with older houses of better quality construction and character if these factors are essential to you

Conclusions:

Limitations of the model:

- The original dataset does not include other important factors, and therefore the model is biased
- Multiple linear regression models, while easily interpretable, are limited in their predictive ability
- Some variables in the dataset are strongly correlated with each other, and that affect the predictive power of the model

Additional analysis suggested:

- Add variables to the original dataset like kitchen renovation, average commute time, crime index, average nearby public school quality, etc.
- Update the dataset with more current data

Thank you!

Email: e.v.kazakova@gmail.com

GitHub: @sealaurel

LinkedIn: https://www.linkedin.com/in/elena-v-kazakova/