

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 2 Final Project

Authors: Elena Kazakova

Cohort: DS02222021

Instructor: James Irving



Summary

This Project analyses the data on the properties sold

- In King County, WA
- From May 2014 till May 2015

The resulting model provides insight into

- what controllable property features increase its' value
- what external variables affect the price of a house.



Outline of the presentation

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



Business Problem

Inference Analysis of King County, WA house prices

This study aims to build a model(s) of a house sale price versus features of the property

This information can be helpful to

- house owners
- house buyers
- and real estate agents in the county

Project scope is limited, and further in-depth analysis would be beneficial



Data

- King County of Washington State
- Between **May 2014** and **May 2015**
- **21597** records
- **19** house features:

date
bedrooms
bathrooms
floors
sqft_living
sqft_above
sqft_basement

sqft_lot
sqft_lot15
sqft_living15
yr_built
yr_renovated

view
waterfront
condition
grade
lat
long
zipcode



Modifications to the Data

Added features:

- **month**
- **Distance**
- **basement_exists**
- **renovation_done**

Limit values in data

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
between 100 and 40000

Square footage of an average neighborhood lot
between 100 and 40000

Distance < 30 miles

Removed features:

- **zipcode**
- **distance**
- **sqft_basement**
- **lat**
- **long**
- **date**



Final Model

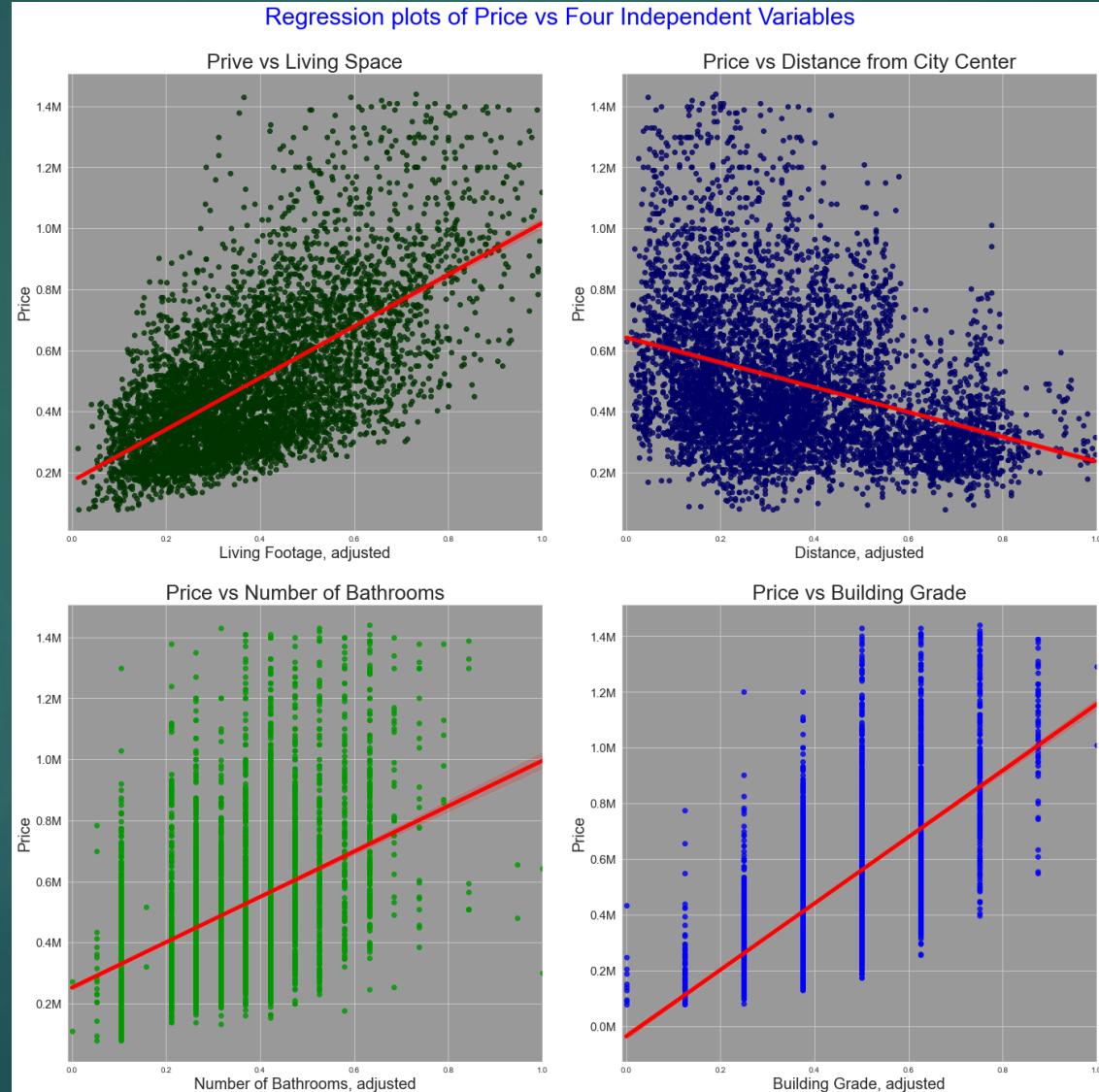
“It's better to solve the right problem approximately than to solve the wrong problem exactly”
Quote by John Tukey

- Several models were tested
- Final model has the best statistical
- The model is a Multiple Linear Regression model
The predictive variables have a linear correlation with the price

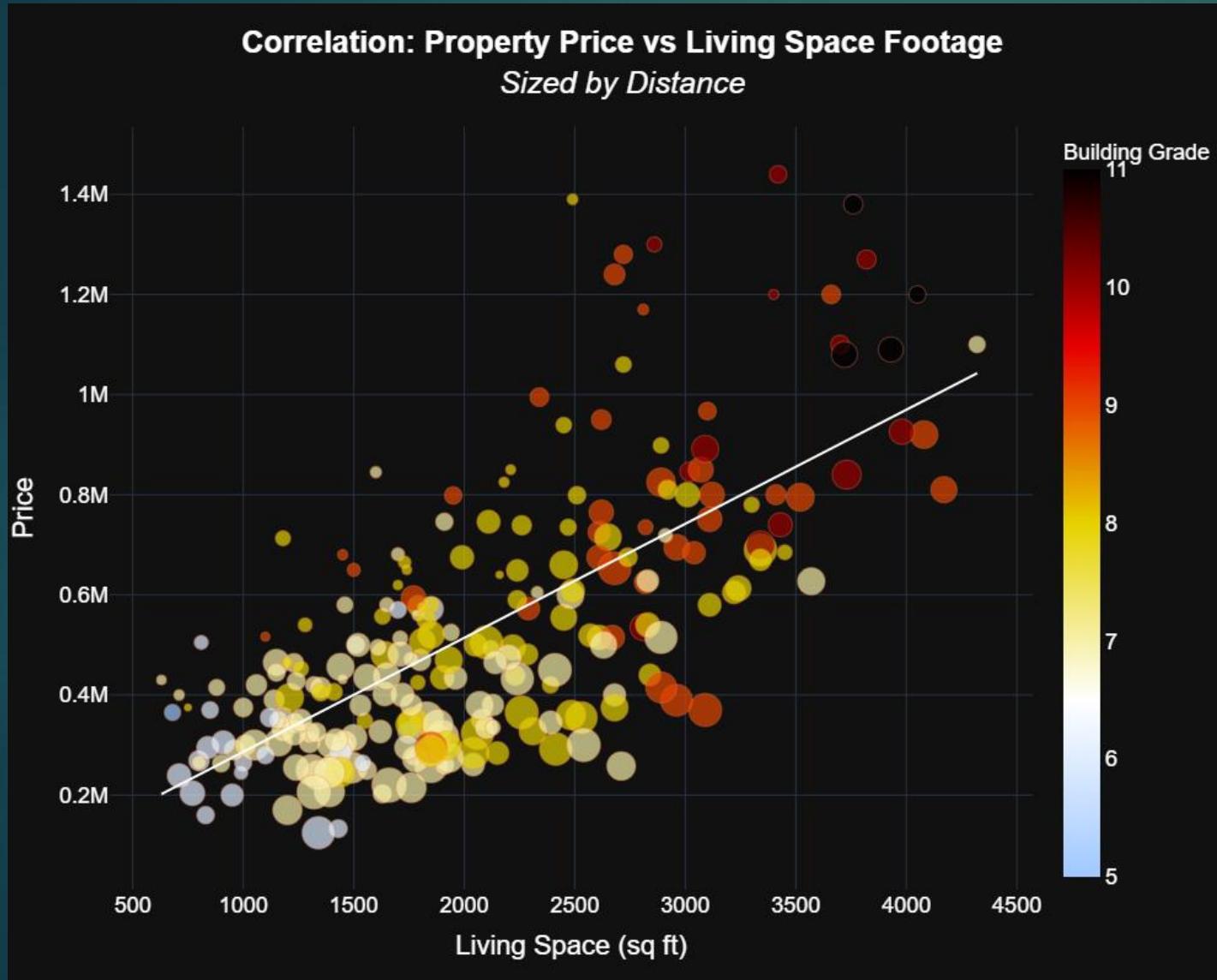
*Price=grade+sqft_living-distance+bathrooms
-recent_renovation_new*



Visualizing linear Regression Model



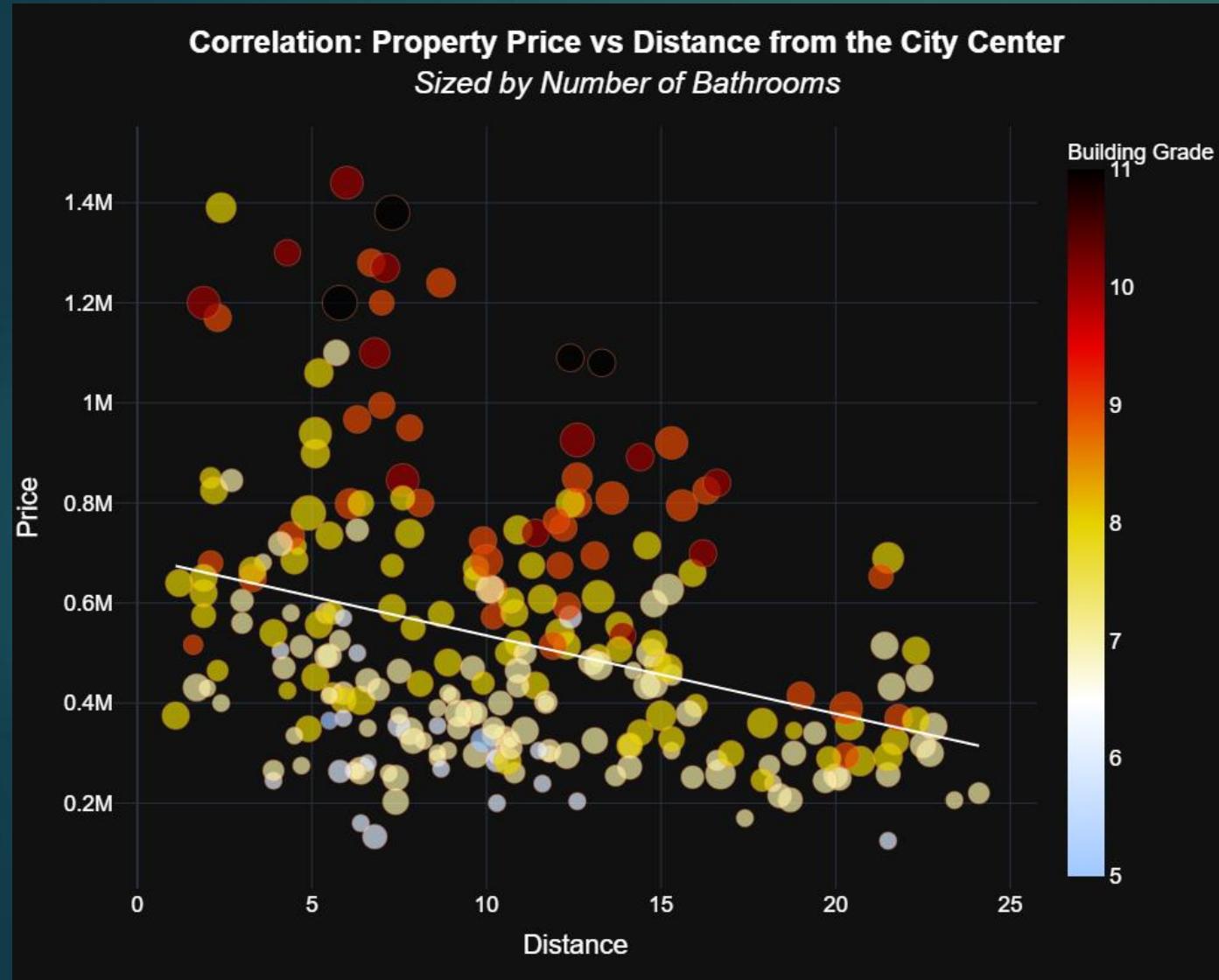
Visualizing the Correlations



Marker color represents **the building grade**

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

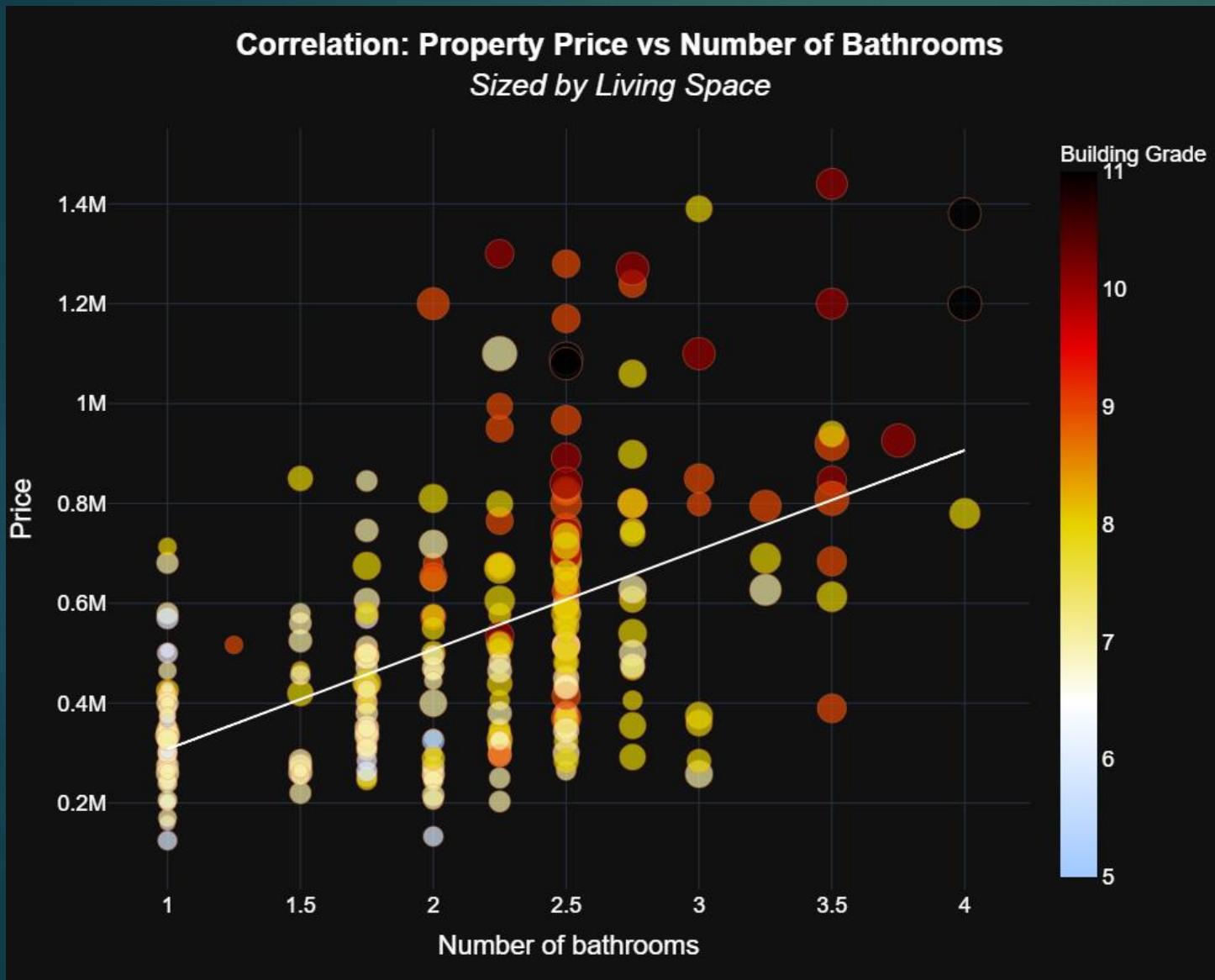
Visualizing the Correlations



Marker color represents **the building grade**

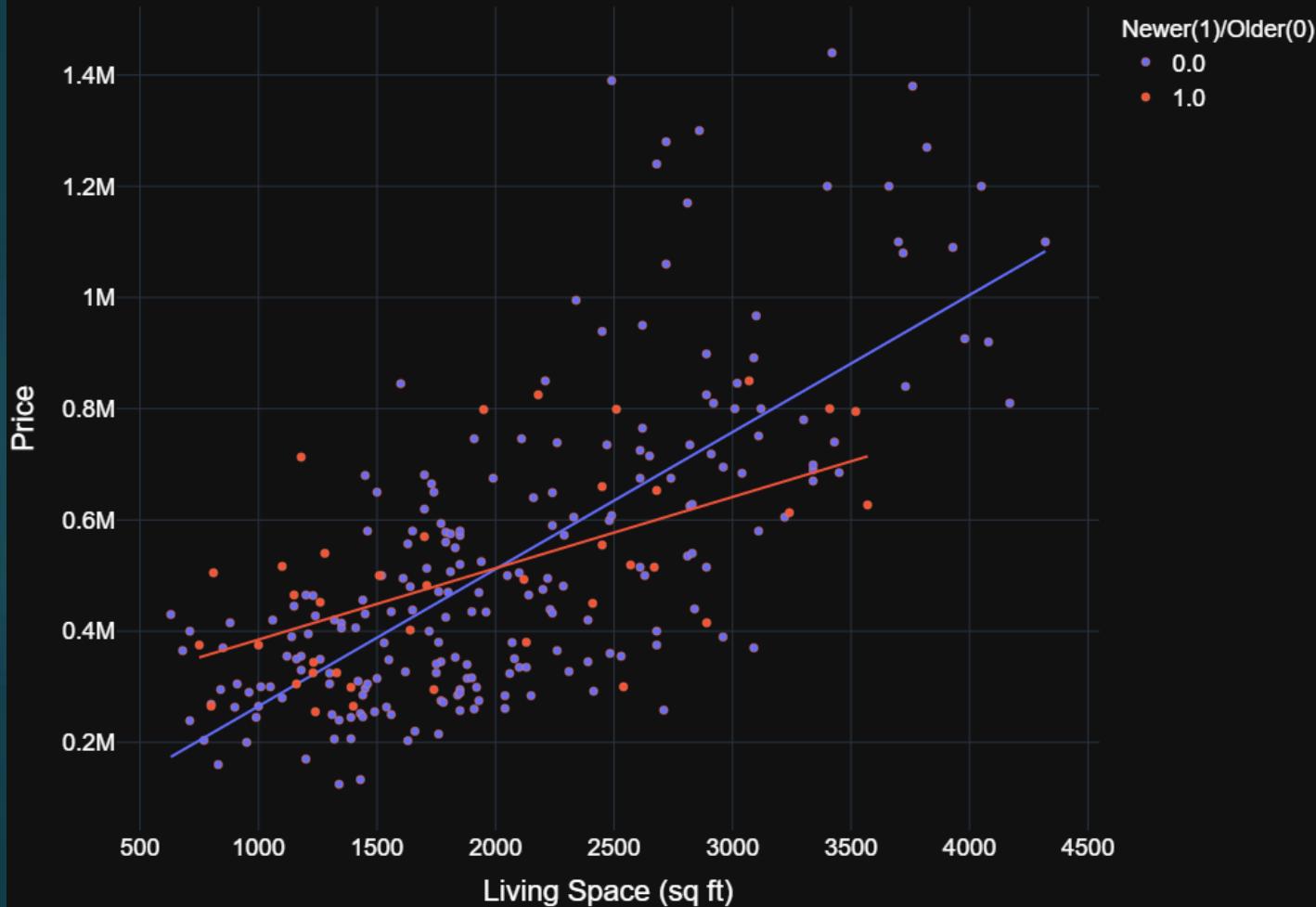
Marker size represent **the number of bathrooms**

Visualizing the Correlations



Visualizing the Correlations

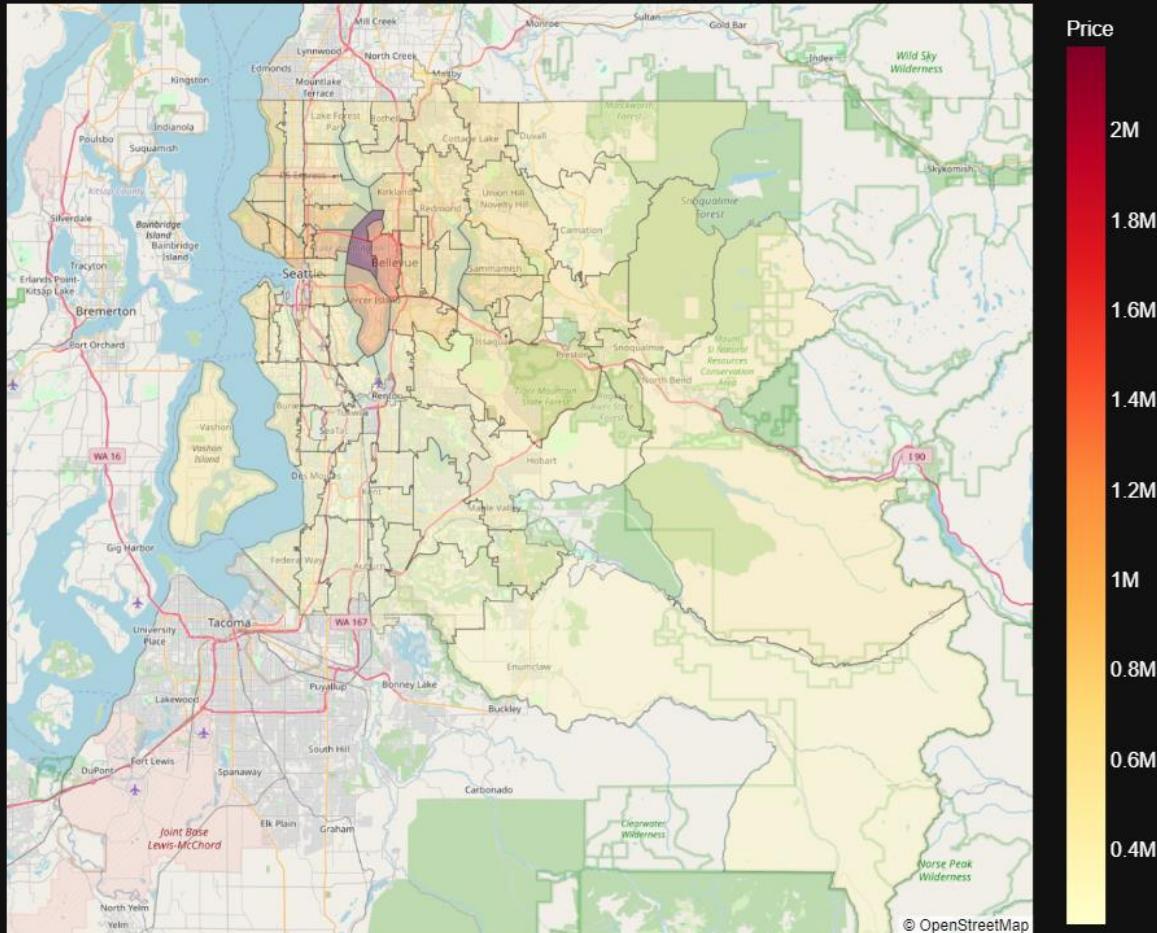
Correlation: Property Price vs Living Space Footage of Newer vs Older Properties



Marker color represents if a property has been
renovated or built between 2006 and 2015

Visualizing the Differences between Zipcodes

Average Prices of Sold Properties per Zipcode (King County, 2014-2015)



Several zipcodes around Bellevue area are the most expensive ones

Bellevue zipcodes

Zipcodes 98039

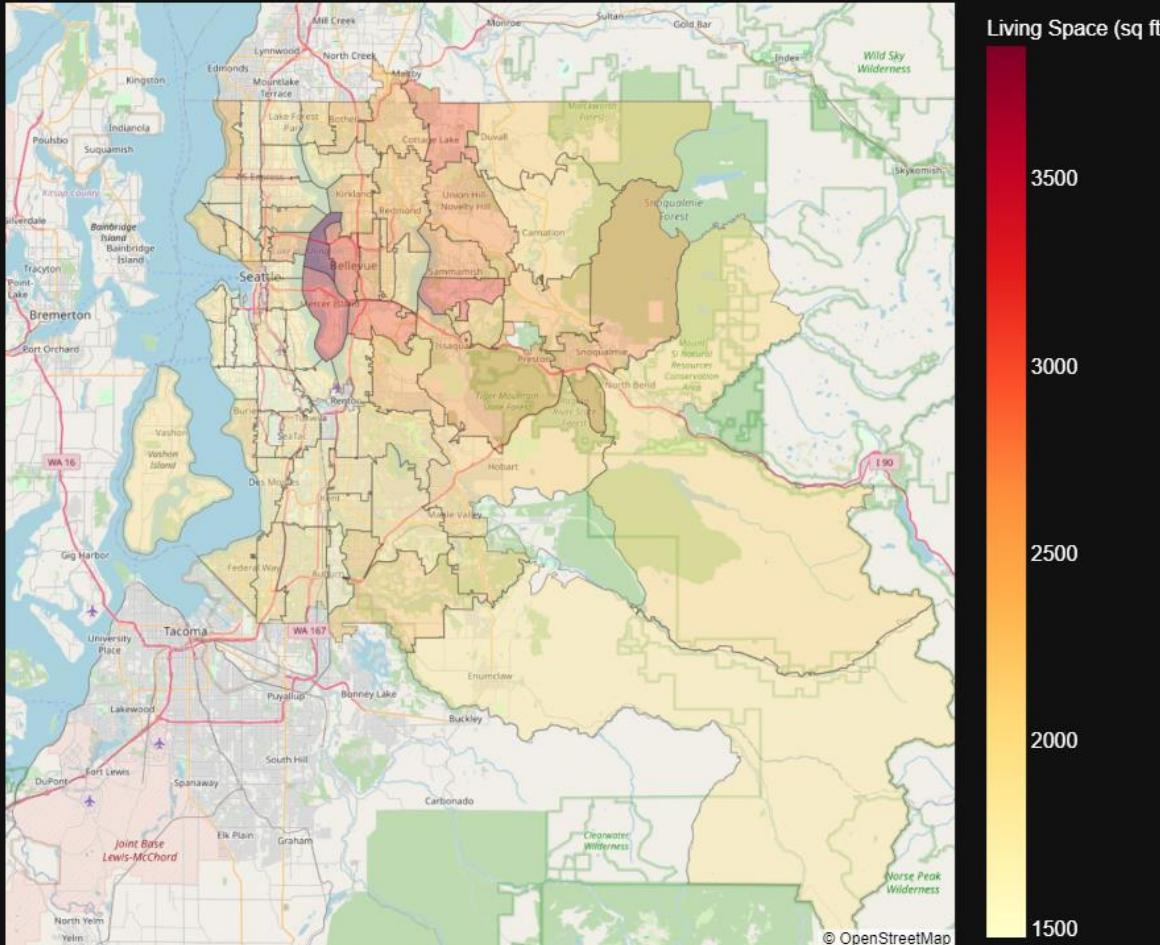
Zipcode 98004

Zip[ode] 98040



Visualizing the Differences between Zipcodes

Average Living Space of Sold Properties per Zipcode (King County, 2014-2015)



Zipcodes around Bellevue area are the most expensive ones

Zipcodes with similar living space footage but more affordable properties

Zipcodes 98077

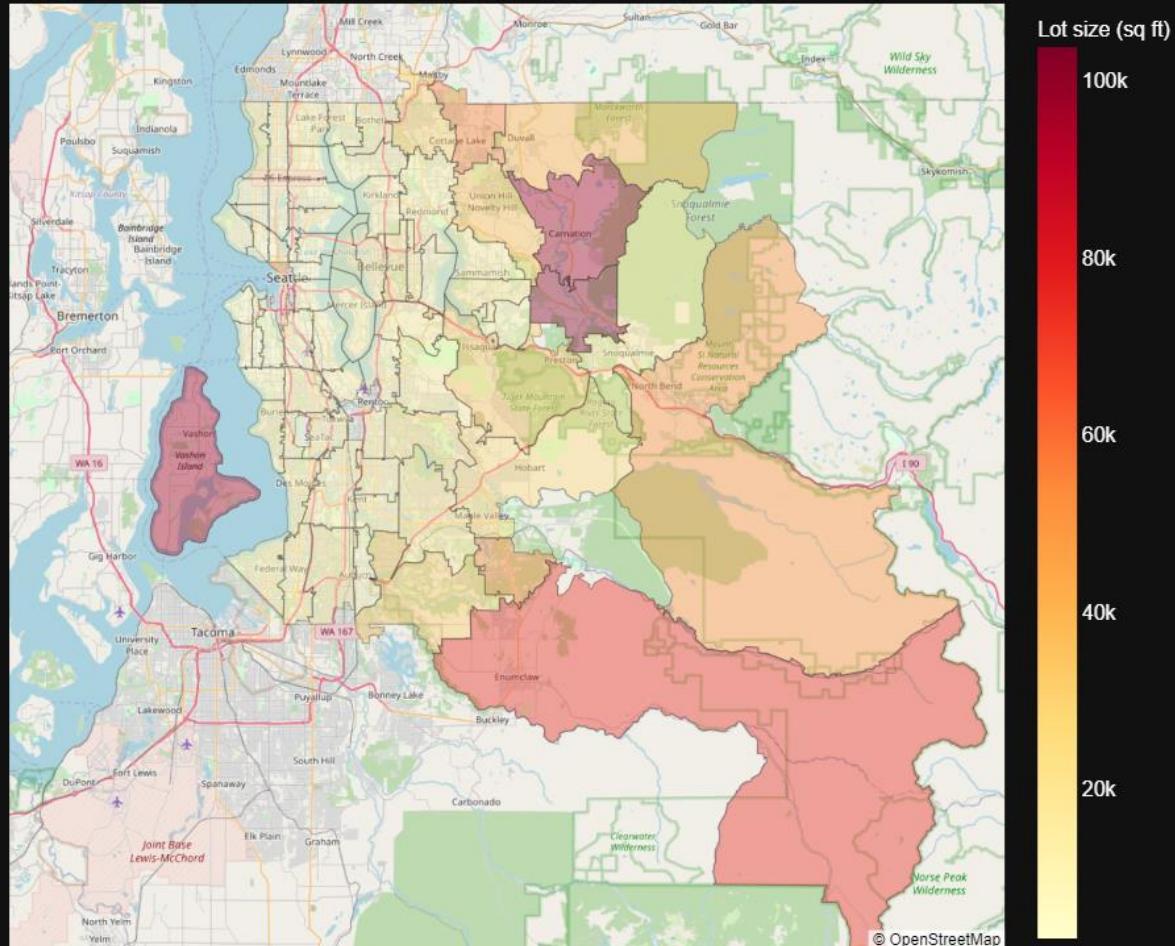
Zipcode 98075

Zip[ode 98006



Visualizing the Differences between Zipcodes

Average Lot Size of Sold Properties per Zipcode (King County, 2014-2015)



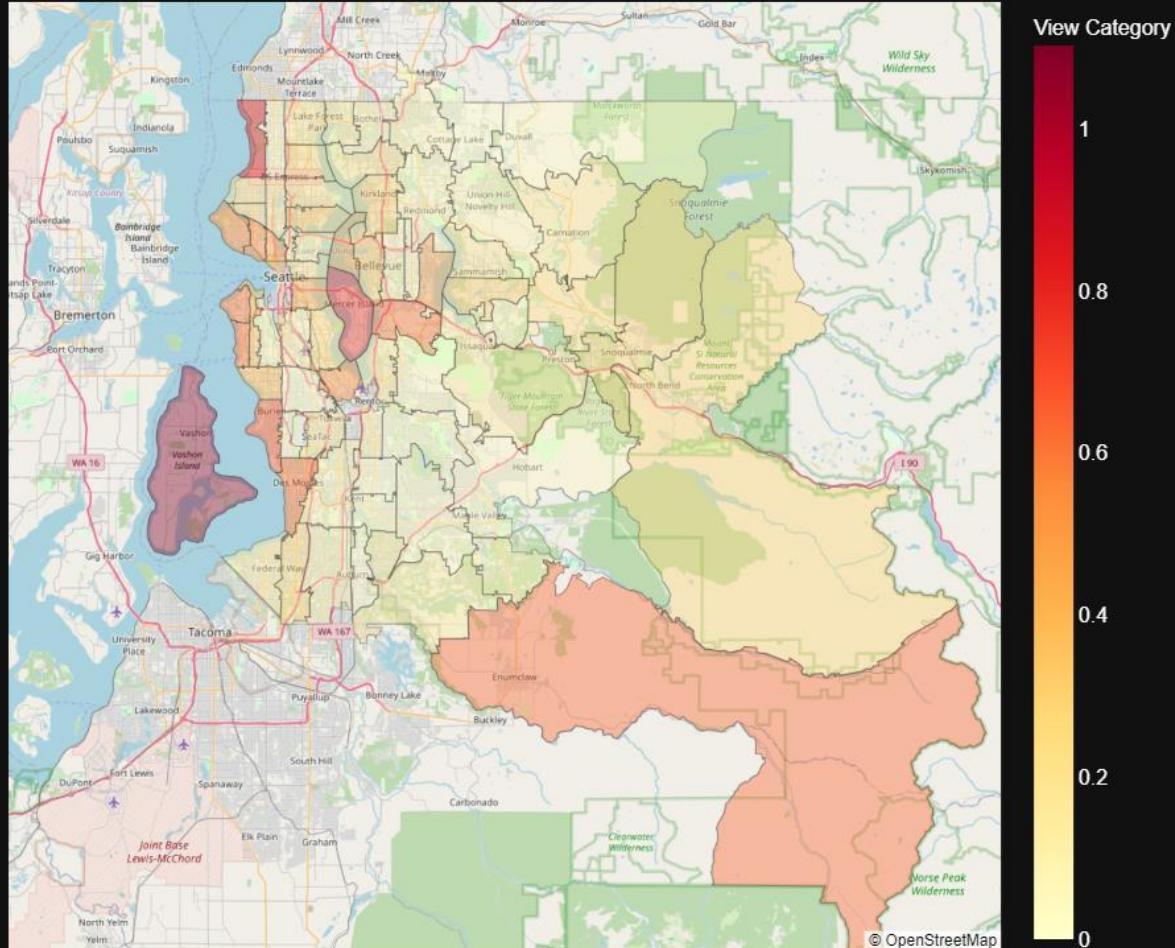
Zipcodes with bigger lots

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



Visualizing the Differences between Zipcodes

Average View Category of Sold Properties per Zipcode (King County, 2014-2015)



Zipcodes with a better view

Zipcodes 98198
Zipcode 98166
Zipcode 98022
Zipcode 98070
Zipcode 98177



Recommendations

Recommendations to property owners:

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

Recommendations to potential home buyers:

- Properties further away from the city with the same features more affordable
- Properties in some zipcodes of the city are more affordable than others, with
 - better view
 - more considerable property lots
 - better quality construction and character
 - same distance



Conclusions:

Limitations of the model:

- Other important factors are not included in dataset
- Multiple linear regression models are limited in their predictive ability
- Strong unavoidable correlation of property features with each other

Additional analysis suggested:

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



Thank you!

Email: e.v.kazakova@gmail.com

GitHub: [@sealaurel](https://github.com/sealaurel)

LinkedIn: <https://www.linkedin.com/in/elenavkazakova/>



Appendix: data description

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



Appendix: data description (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



Appendix: 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)



Appendix: 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