King County House Price Prediction

Data Analysis and Multiple Regression in Python

Presenter: Bonnie Ma

Agenda

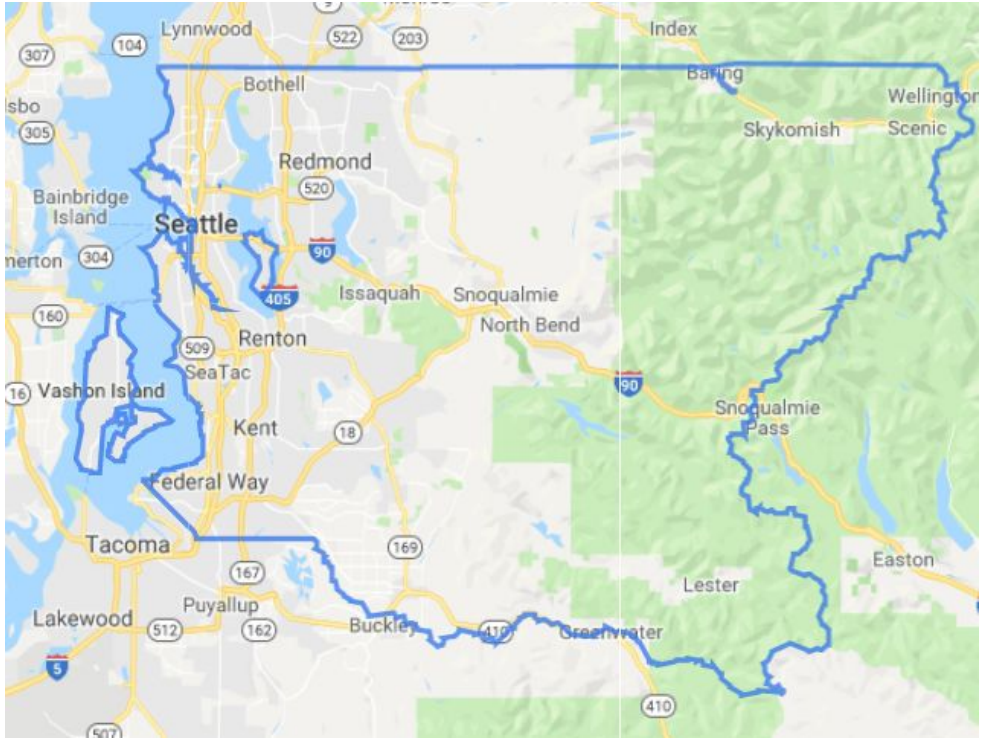
- Business Background
- Analytics Approach
- Data Exploration
- Model Overview
- Model Results
- Insights

Business Background

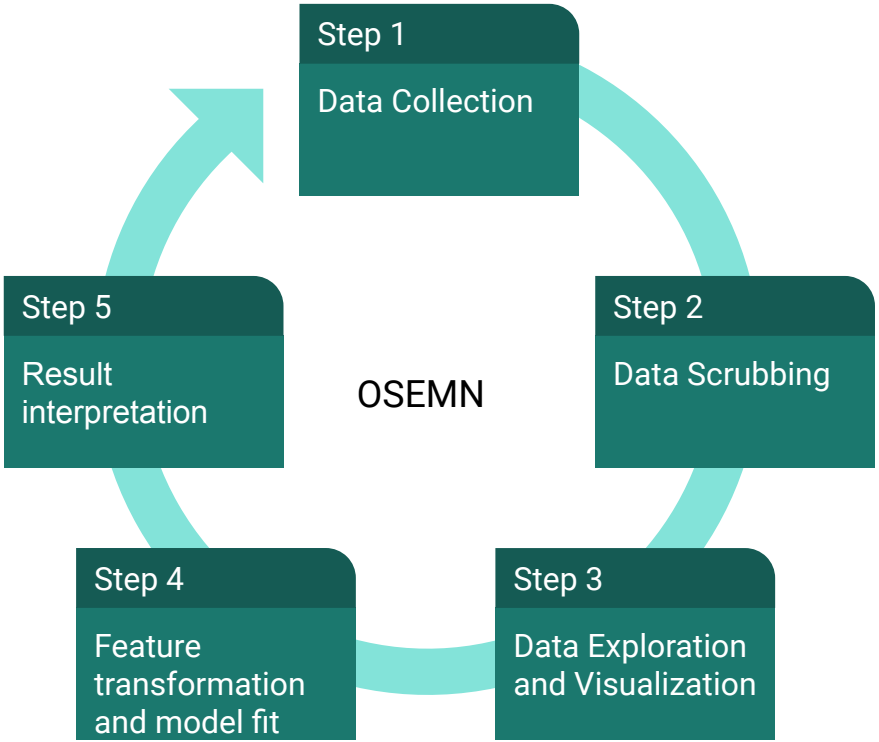
King County is one of the most beautiful and populous area in the United States. It is located in the state of Washington and has the population of 2.3 million as of 2018.

The growth of companies like Amazon and Microsoft has lifted the house price significantly over the past few years.

The goal of this project is to understand the drivers of King County house price in order to predict house price in the future and identify investment opportunities.



Our Approach - OSEMN

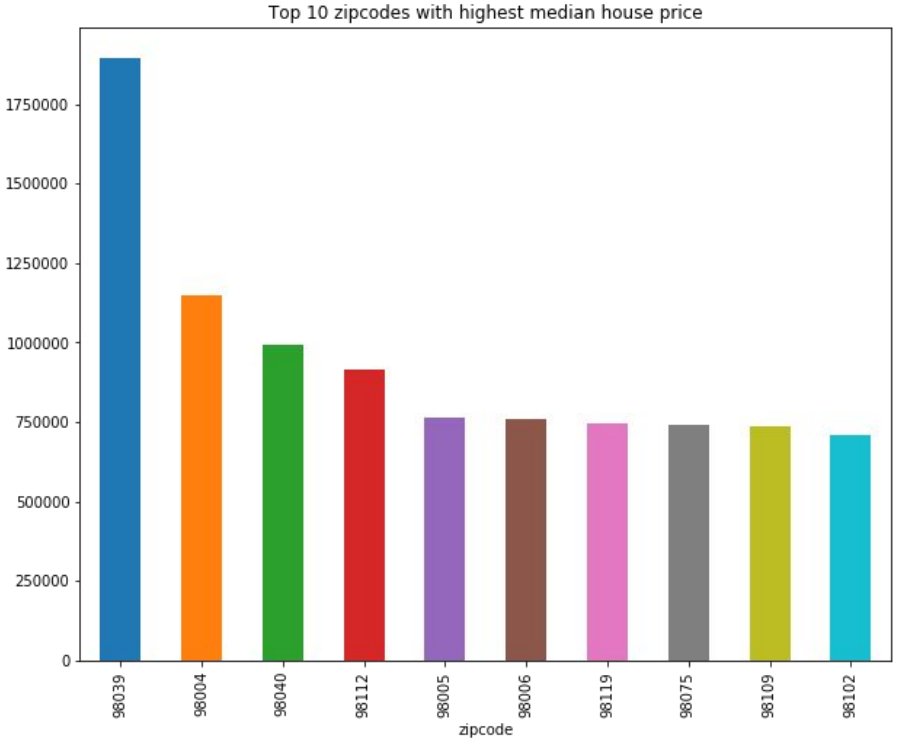


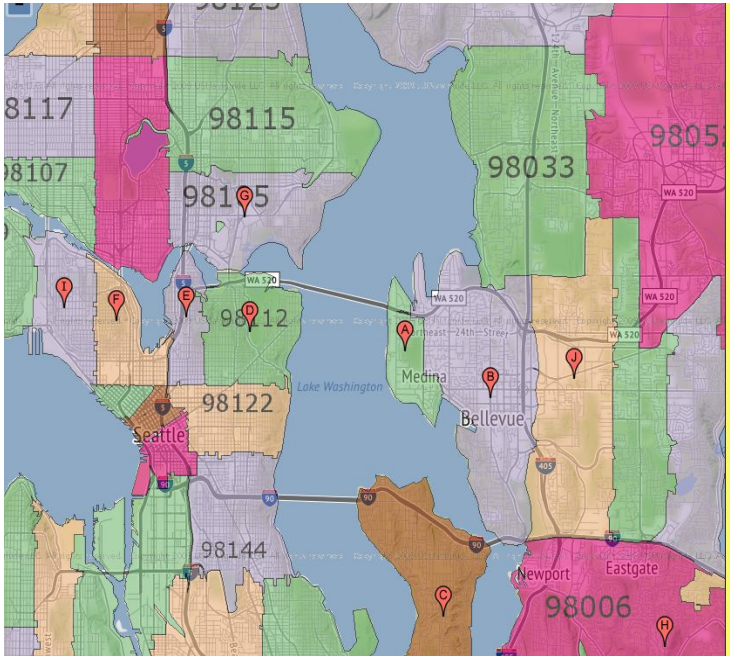
Goal: build a robust regression model to predict King County house price as accurate as possible

KPIs: Adjusted R-square, Mean Squared Error (MSE)

The data set has 21,597 records and 18 variables.

Top 10 Areas with Highest Median House Price





J 98005 Bellevue, WA
County: King
Area Code: 425
2000 Households: 7786
2008 Households: 8403

I 98119 Seattle, WA
County: King
Area Code: 206
2000 Households: 10040
2008 Households: 10844

H 98006 Bellevue, WA
County: King
Area Code: 425
2000 Households: 13039
2008 Households: 14071

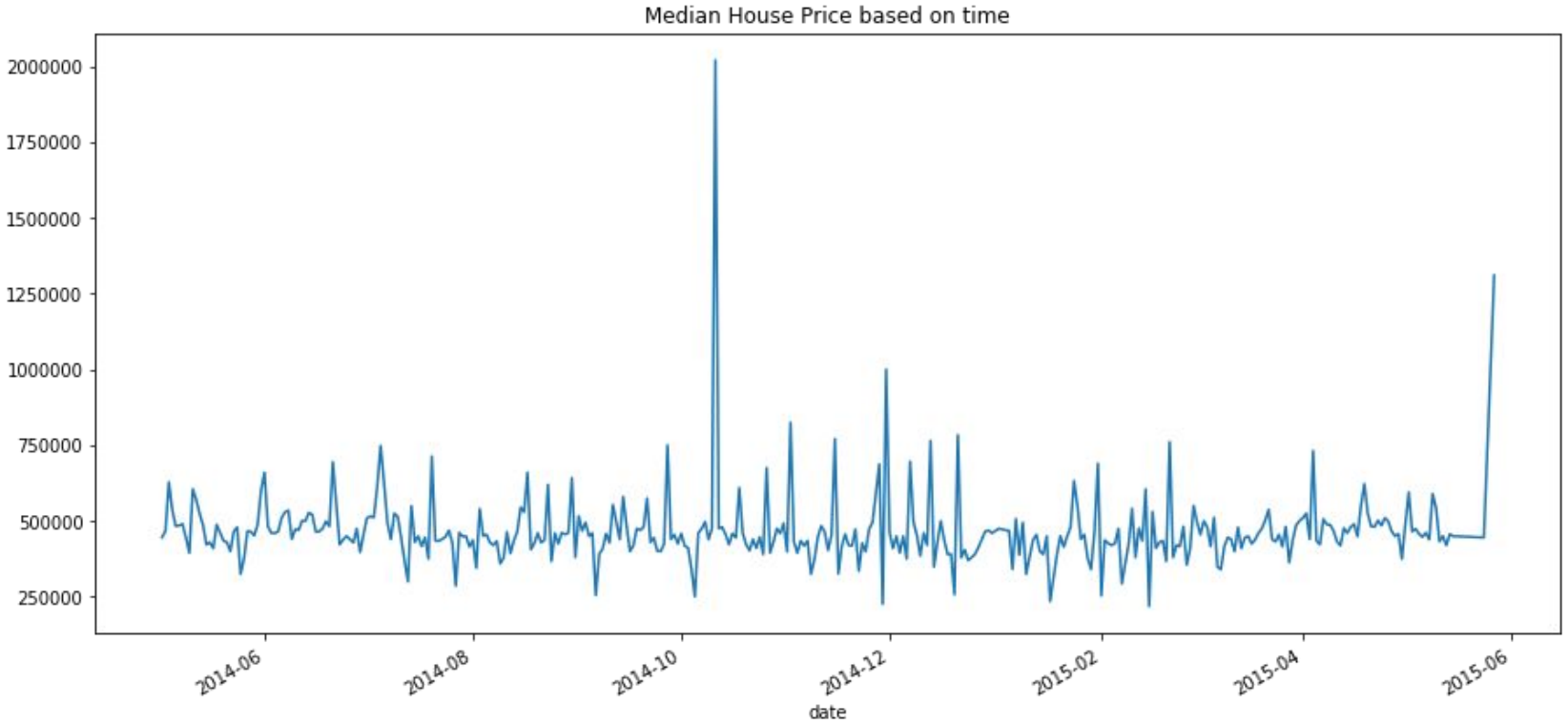
G 98105 Seattle, WA
County: King
Area Code: 206
2000 Households: 14965
2008 Households: 16155

F 98109 Seattle, WA
County: King
Area Code: 206
2000 Households: 9472
2008 Households: 10225

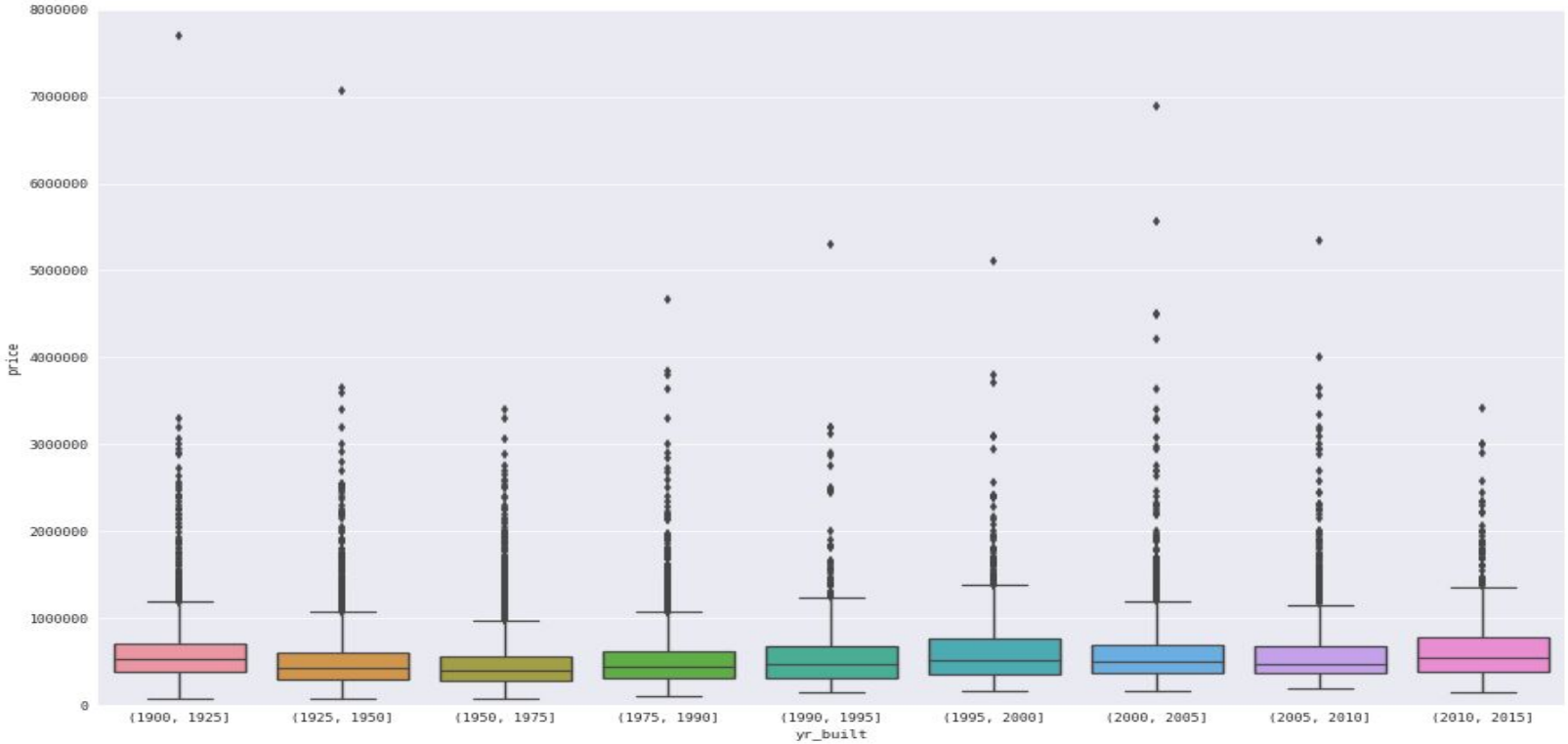
E 98102 Seattle, WA
County: King
Area Code: 206
2000 Households: 13126
2008 Households: 14166

D 98112 Seattle, WA
County: King
Area Code: 206
2000 Households: 9374

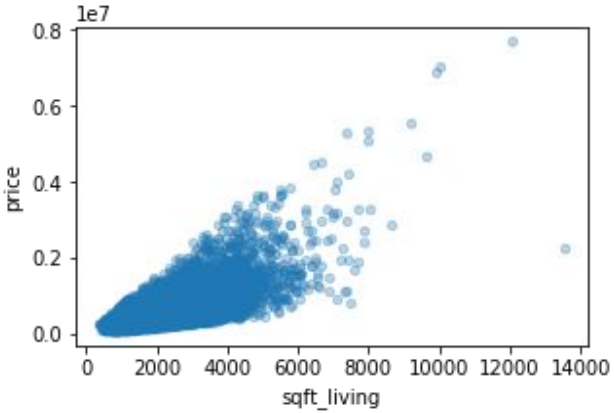
House Price Change Over Time

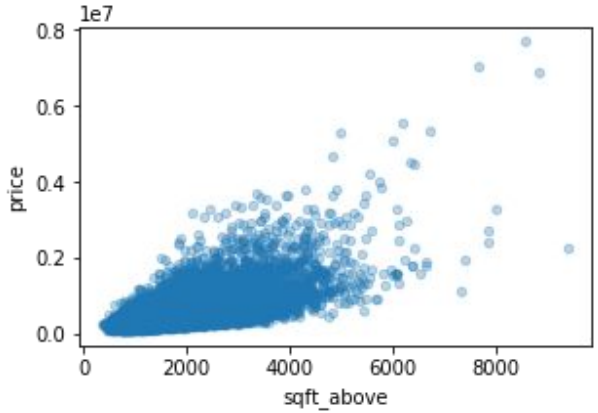


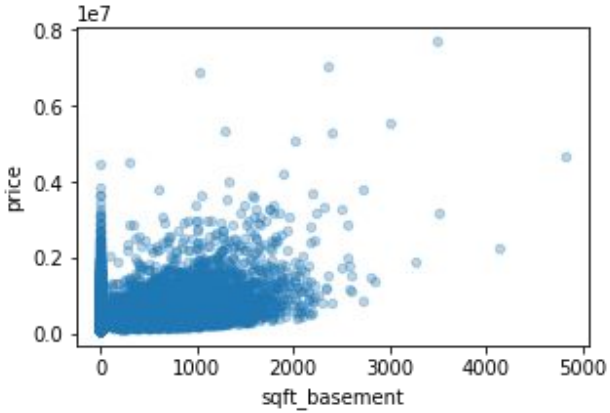
House Price Change Over Time



House price vs Footage

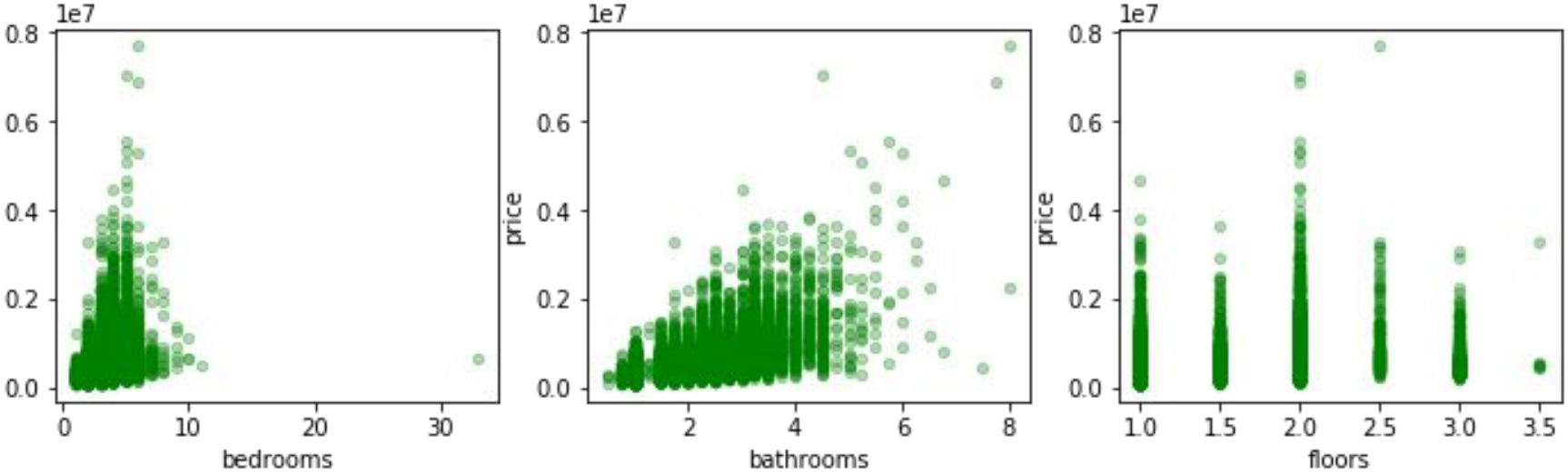






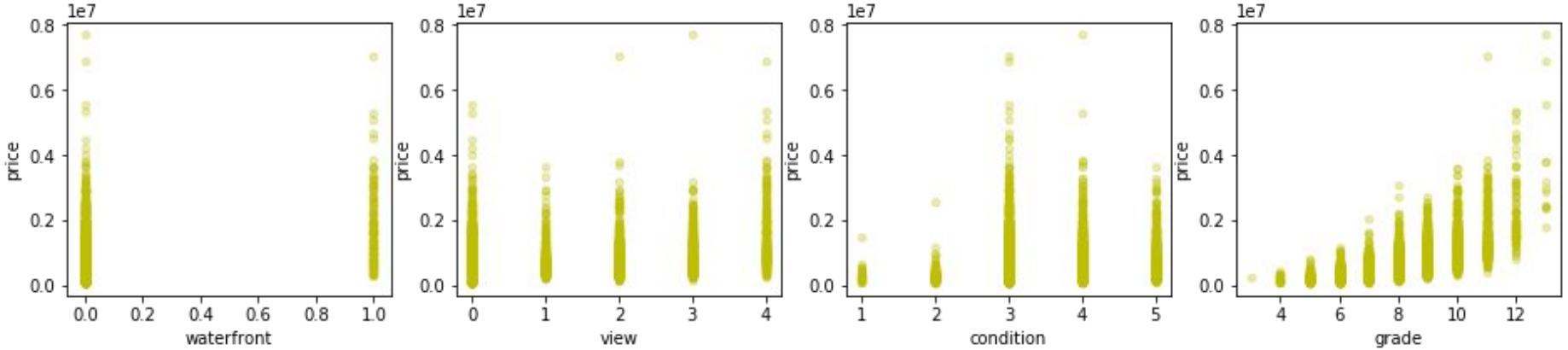
Sqft_living: Footage of the home
Sqft_above: footage of house apart from basement
Sqft_basement: footage of the basement

House price vs Bed/Bathrooms/Floors



No obvious correlation observed between variable and target.

House price vs other Categorical Variables



Grade has a positive correlation with price.

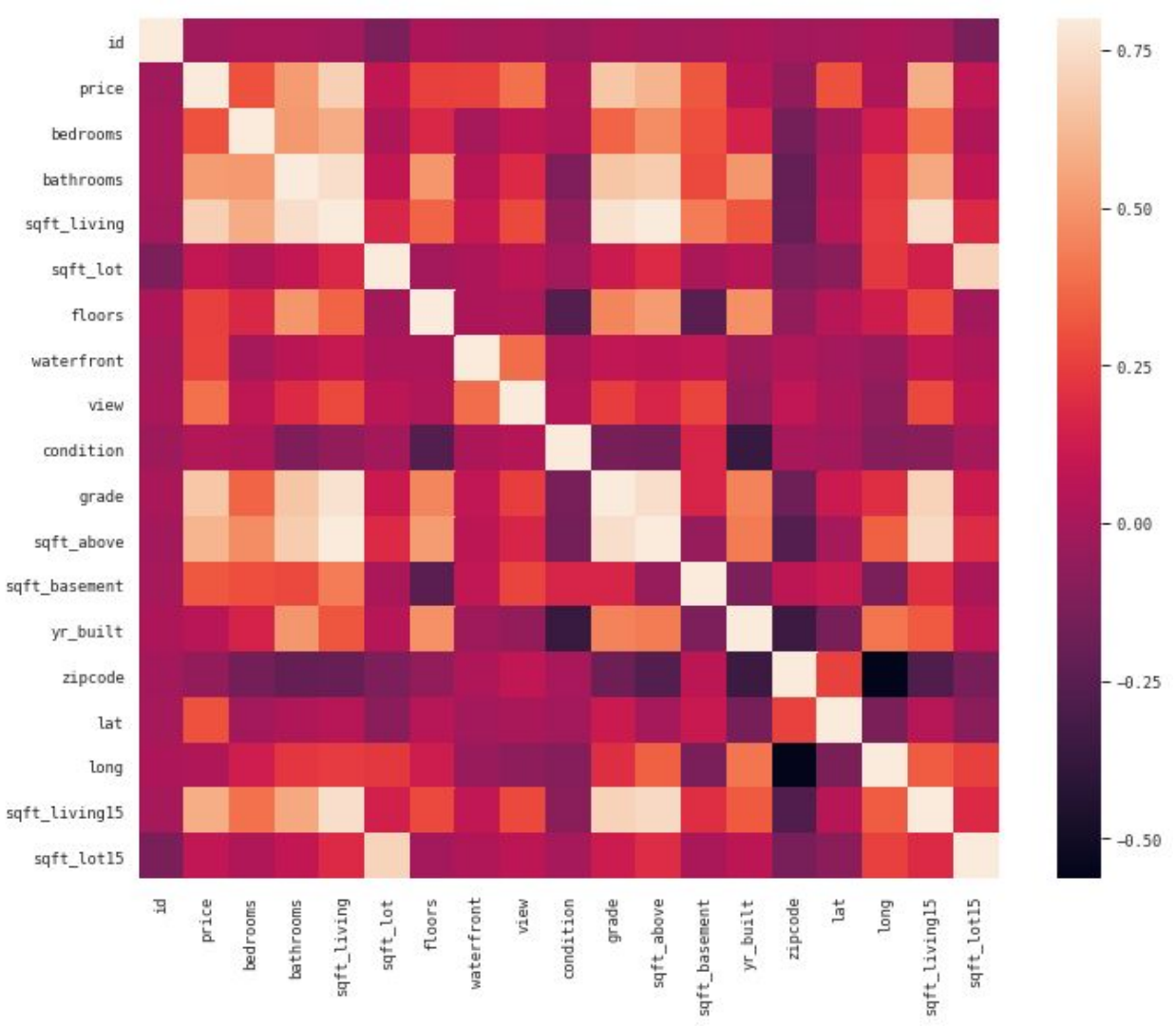
Correlation between variables

Variables that are highly correlated to price are:

- sqft_living
- sqft_above
- grade

Independent variables which are highly correlated with each other:

- sqft_living with sqft_living15, bathroom, grade, sqft_above
- grade with bathroom, sqft_above, sqft_living15, sqft_living



Model Set up

- Multiple Linear Regression on Price
- Data Transformation:
  - Drop variables with high collinearity or poor quality (sqft_living15, sqft_lot15, yr_renovated)
  - Turn binary and ordinal variables into categorical values (bedrooms, bathrooms, waterfront, view, condition, grade, yr_built, zip code)
  - Log transformation and standardization on continuous variables
- Package used for model: statsmodels, sklearn
- Model selection method: Forward selection and Stepwise selection
- Model validation: cross-validation

Model Result - 83 variables

Variables that significant impact to house price:

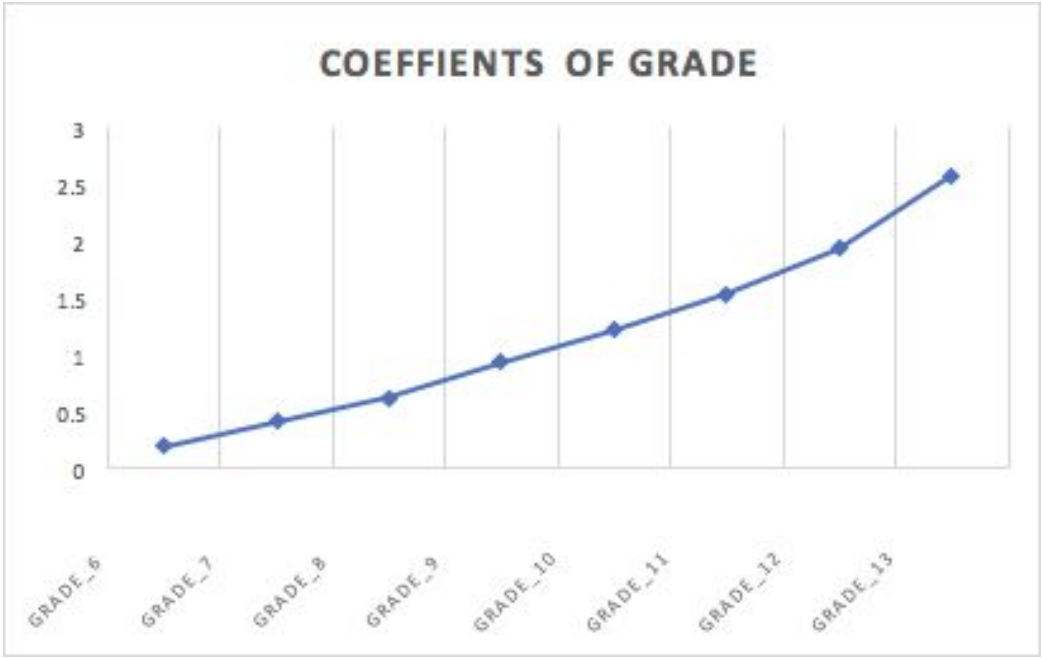
- Footage of area above basement
- Bathroom: 1.5-2, 2-2.5, 5-8
- Bedroom: 1-2, 3-4, 4-5, 5-6
- Grade: 6, 7, 8, 9, 10, 11, 12, 13
- Zipcode: some of zipcodes
- Year: 1925-1950, 1950-1975, 1975-1990, 1990-1995, 1995-2000

Model KPIs:

- Adjusted R-sqaure: 0.839
- P values: all p values is less than 0.05
- Train MSE: 0.1615
- Test MSE: 0.1607

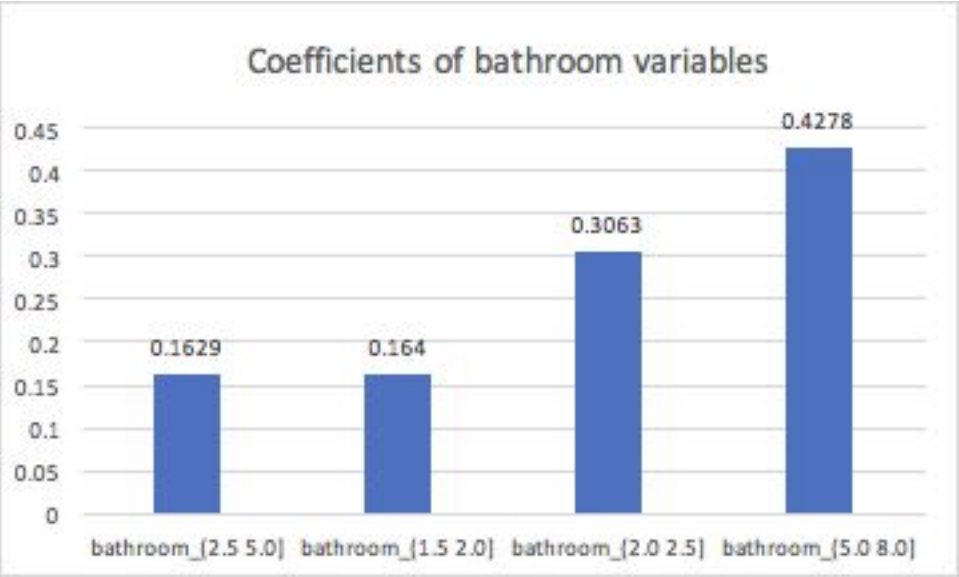
Insights - grade

Grade begins to have significant impact on the house price till the level reaches 6 and the coefficient accelerates as the grade increases. It is important to pay attention to house grades when making investment decisions.



Insights - bathrooms

Number of bathrooms per bedroom has different impacts on house price in different bins. When it reaches 5-8 bathrooms per bedroom, the coefficient increased the most which may indicate a different kind of building.



Insights - zip code

Coefficients of zip code tell which area is more referable by the market. For long term investment, it is better to choose zip code with a positive coefficient.

Fastest growing areas	Fastest declining areas
98039	98022
98004	98023
98112	98001
98109	98003
98119	98092