# House Sales in King County USA Regression

Done by

1- Dr.Saleh AL-Frhan

2- Mohammad Alzabyedi

3- Saif Sultan

# OutLines

- Introduction
- Overview of the Study
- Data
- Data Cleaning
- Visualizing The Data
- Data Preparing
- Models and Tests

# Introduction

King County is a country located in the U.S. state of Washington. The population was 2,149,970 in a 2016 census estimate. King is the most populous county in Washington, and the 13th-most populous in the United States. The county seat is Seattle, which is the state's largest city. King County is one of three Washington counties that are included in the Seattle-Tacoma-Bellevue metropolitan statistical area. About two-thirds of King County's population lives in the city's suburbs. As of 2011, King County was the 86th highest-income county in the United States. This paper addresses the factors concerning the "house sale prices" in King County sold between May 2014 and May 2015.

# Overview of the Study

Our field study concerns house prices in King County, USA. The county comprises houses with varied features. The features include bedrooms/house, bathrooms/bedroom, area of the house and lot, presence of a waterfront, views, condition of the house, grade assigned by the county, built year, renovated year and the location of the house. We empirically study how the various factors influence the house prices. Our regression analysis revealed the best fit model to predict the price of the house. We found that the houses with no waterfront and fewer bedrooms were the cheapest and the houses which comprised a waterfront had more views than the ones which didn't.

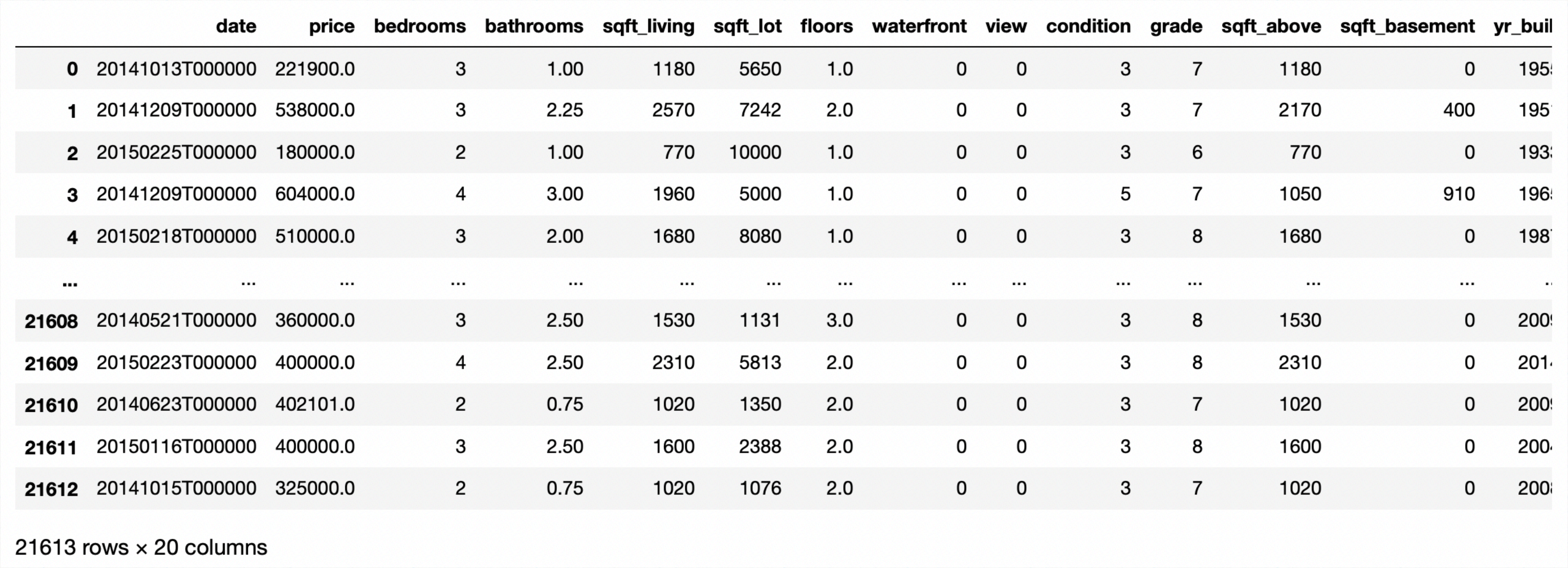
# Data

For this study, we collected data from the website named “Kaggle”-(https://www.kaggle.com/harlfoxem/housesalesprediction). Kaggle is a platform for predictive modelling and analytics competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users.

# Data

- id :a notation for a house
- date: Date house was sold
- price: Price is prediction target
- bedrooms: Number of Bedrooms/House
- bathrooms: Number of bathrooms/bedrooms
- sqft_living: square footage of the home
- sqft_lot: square footage of the lot
- floors :Total floors (levels) in house
- waterfront :House which has a view to a waterfront
- view: Has been viewed
- condition :How good the condition is Overall
- grade: overall grade given to the housing unit, based on King County grading system
- sqft_above :square footage of house apart from basement
- sqft_basement: square footage of the basement
- yr_built :Built Year
- yr_renovated :Year when house was renovated
- zipcode:zip code
- lat: Latitude coordinate
- long: Longitude coordinate
- sqft_living15 :Living room area in 2015(implies-- some renovations) This might or might not have affected the lot size area
- sqft_lot15 :lotSize area in 2015(implies-- some renovations)ased on King County grading system
- sqft_above :square footage of house apart from basement
- sqft_basement: square footage of the basement
- yr_built :Built Year
- sqft_above :square footage of house apart from basement

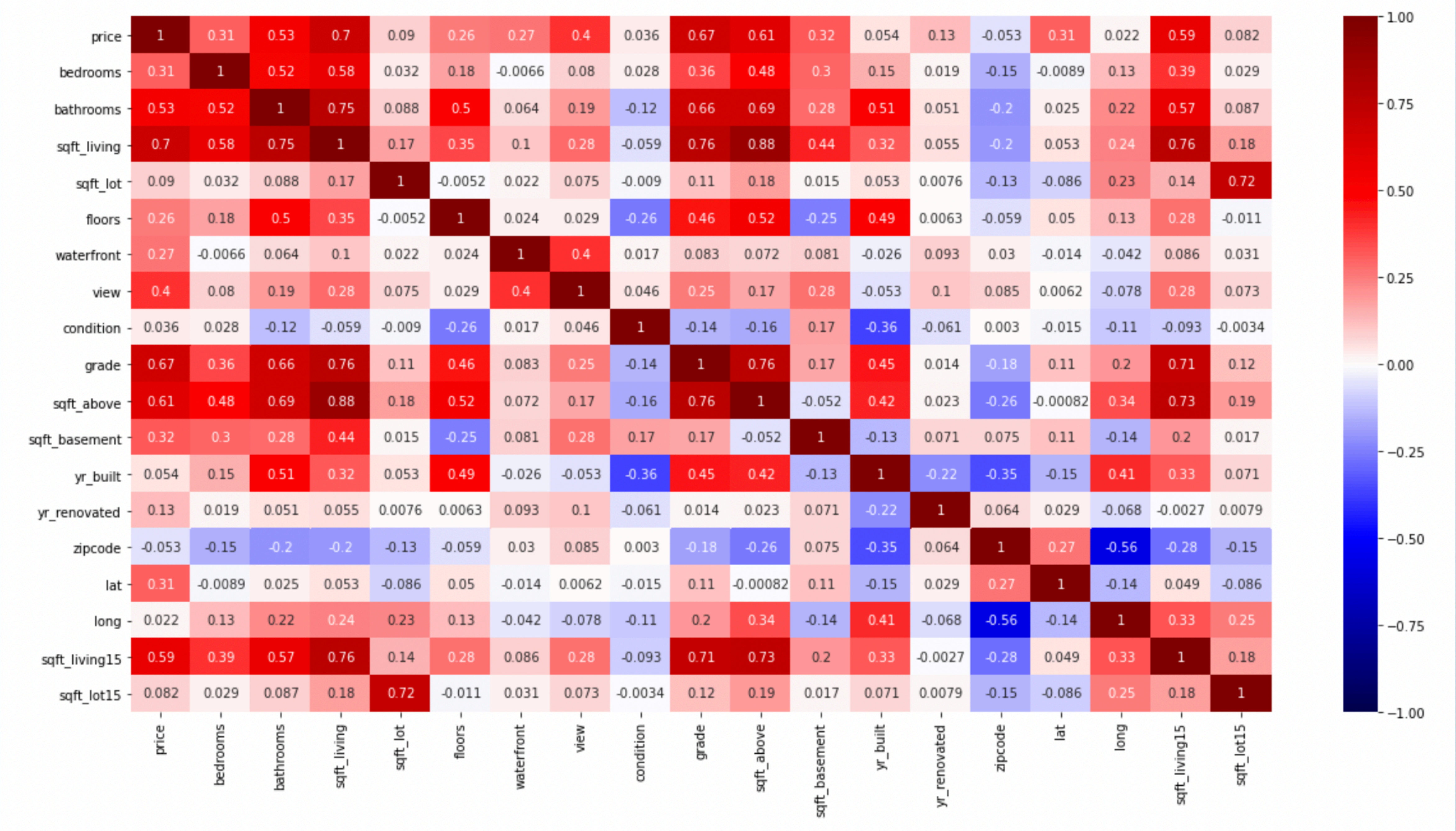
# Data Cleaning



| | date | price | bedrooms | bathrooms | sqft_living | sqft_lot | floors | waterfront | view | condition | grade | sqft_above | sqft_basement | yr_buil |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 20141013T000000 | 221900.0 | 3 | 1.00 | 1180 | 5650 | 1.0 | 0 | 0 | 3 | 7 | 1180 | 0 | 195 |
| **1** | 20141209T000000 | 538000.0 | 3 | 2.25 | 2570 | 7242 | 2.0 | 0 | 0 | 3 | 7 | 2170 | 400 | 195 |
| **2** | 20150225T000000 | 180000.0 | 2 | 1.00 | 770 | 10000 | 1.0 | 0 | 0 | 3 | 6 | 770 | 0 | 193 |
| **3** | 20141209T000000 | 604000.0 | 4 | 3.00 | 1960 | 5000 | 1.0 | 0 | 0 | 5 | 7 | 1050 | 910 | 196 |
| **4** | 20150218T000000 | 510000.0 | 3 | 2.00 | 1680 | 8080 | 1.0 | 0 | 0 | 3 | 8 | 1680 | 0 | 198 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | .. |
| **21608** | 20140521T000000 | 360000.0 | 3 | 2.50 | 1530 | 1131 | 3.0 | 0 | 0 | 3 | 8 | 1530 | 0 | 200 |
| **21609** | 20150223T000000 | 400000.0 | 4 | 2.50 | 2310 | 5813 | 2.0 | 0 | 0 | 3 | 8 | 2310 | 0 | 201 |
| **21610** | 20140623T000000 | 402101.0 | 2 | 0.75 | 1020 | 1350 | 2.0 | 0 | 0 | 3 | 7 | 1020 | 0 | 200 |
| **21611** | 20150116T000000 | 400000.0 | 3 | 2.50 | 1600 | 2388 | 2.0 | 0 | 0 | 3 | 8 | 1600 | 0 | 200 |
| **21612** | 20141015T000000 | 325000.0 | 2 | 0.75 | 1020 | 1076 | 2.0 | 0 | 0 | 3 | 7 | 1020 | 0 | 200 |

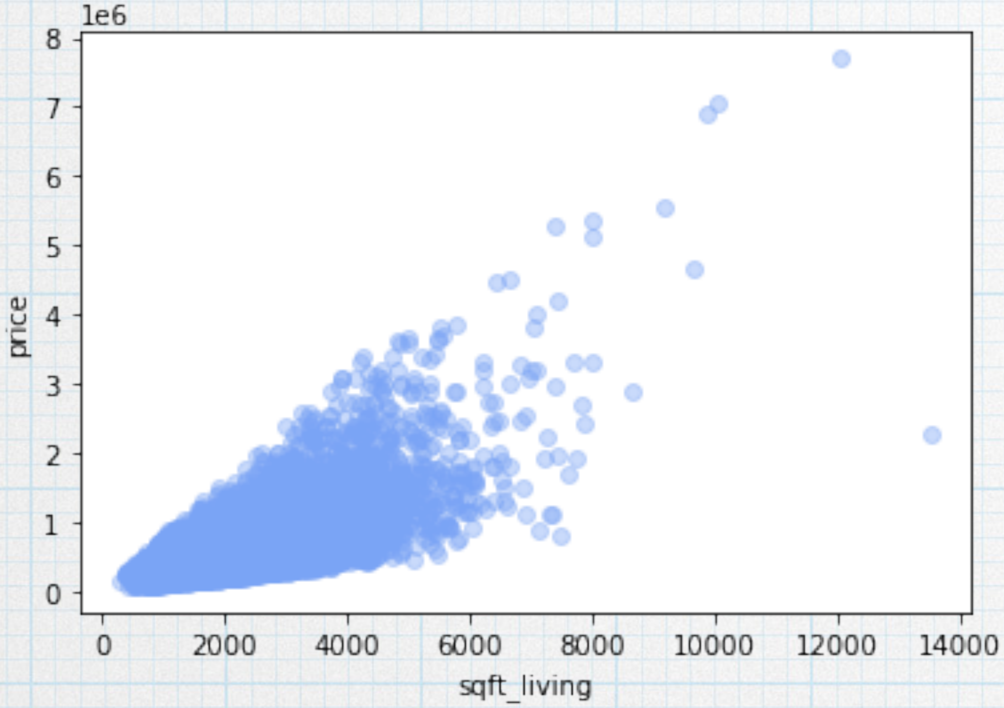
21613 rows × 20 columns

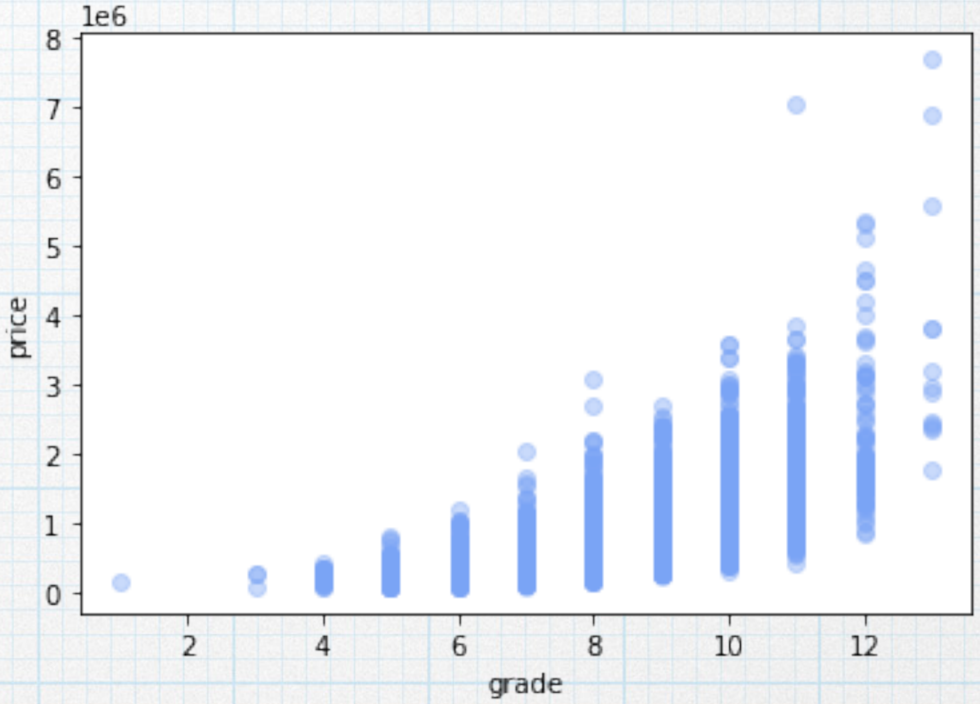
# Visualizing The Data

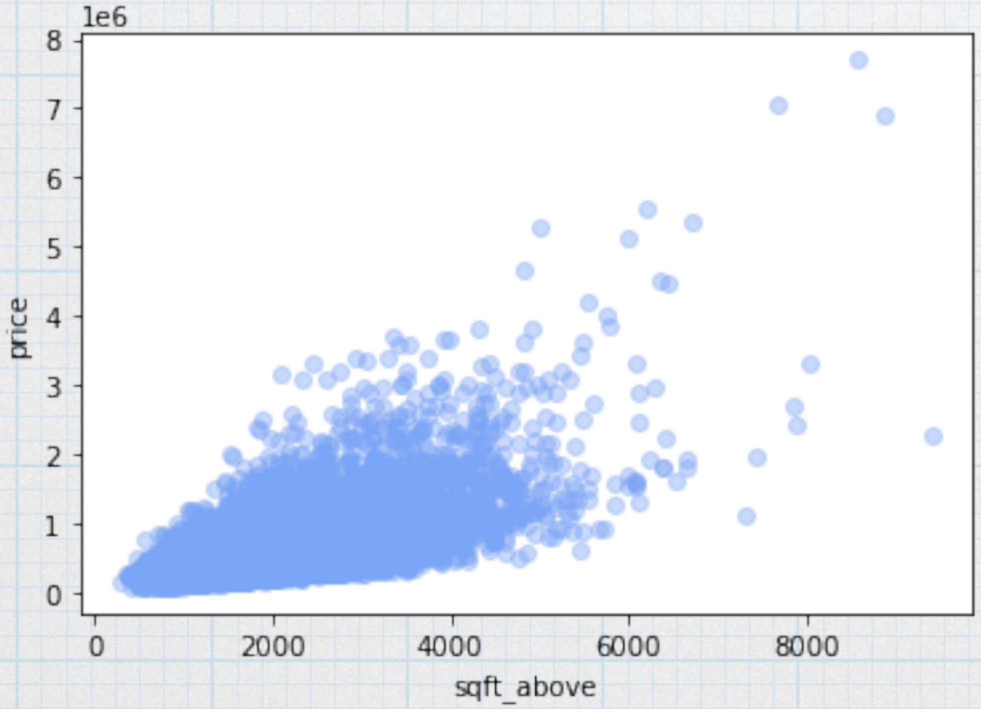


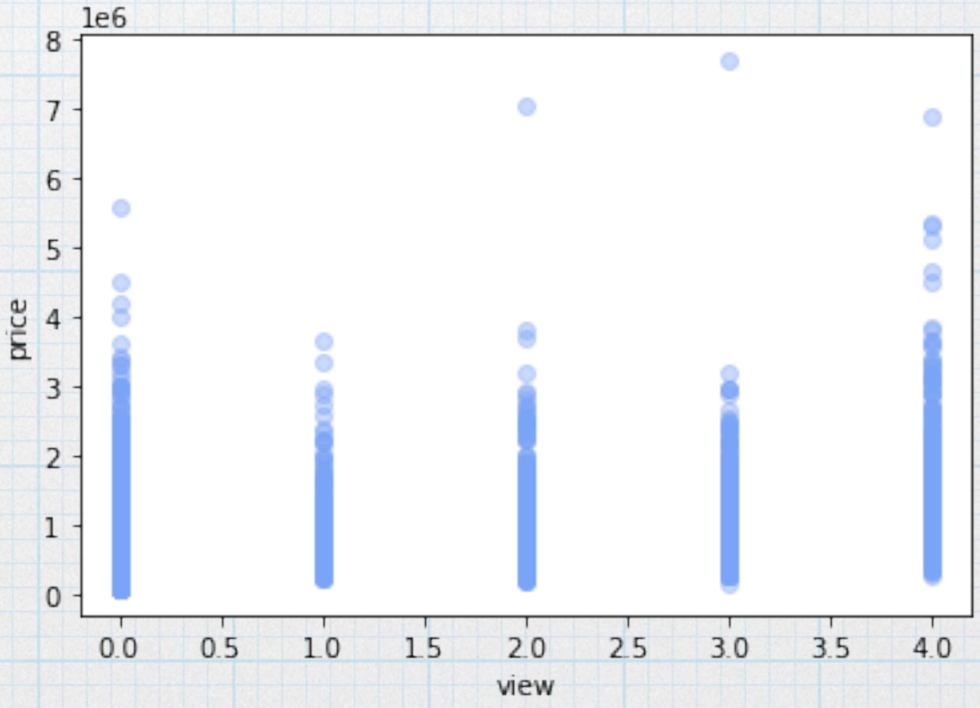
| | price | bedrooms | bathrooms | sqft_living | sqft_lot | floors | waterfront | view | condition | grade | sqft_above | sqft_basement | yr_built | yr_renovated | zipcode | lat | long | sqft_living15 | sqft_lot15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| price | 1 | 0.31 | 0.53 | 0.7 | 0.09 | 0.26 | 0.27 | 0.4 | 0.036 | 0.67 | 0.61 | 0.32 | 0.054 | 0.13 | -0.053 | 0.31 | 0.022 | 0.59 | 0.082 |
| bedrooms | 0.31 | 1 | 0.52 | 0.58 | 0.032 | 0.18 | -0.0066 | 0.08 | 0.028 | 0.36 | 0.48 | 0.3 | 0.15 | 0.019 | -0.15 | -0.0089 | 0.13 | 0.39 | 0.029 |
| bathrooms | 0.53 | 0.52 | 1 | 0.75 | 0.088 | 0.5 | 0.064 | 0.19 | -0.12 | 0.66 | 0.69 | 0.28 | 0.51 | 0.051 | -0.2 | 0.025 | 0.22 | 0.57 | 0.087 |
| sqft_living | 0.7 | 0.58 | 0.75 | 1 | 0.17 | 0.35 | 0.1 | 0.28 | -0.059 | 0.76 | 0.88 | 0.44 | 0.32 | 0.055 | -0.2 | 0.053 | 0.24 | 0.76 | 0.18 |
| sqft_lot | 0.09 | 0.032 | 0.088 | 0.17 | 1 | -0.0052 | 0.022 | 0.075 | -0.009 | 0.11 | 0.18 | 0.015 | 0.053 | 0.0076 | -0.13 | -0.086 | 0.23 | 0.14 | 0.72 |
| floors | 0.26 | 0.18 | 0.5 | 0.35 | -0.0052 | 1 | 0.024 | 0.029 | -0.26 | 0.46 | 0.52 | -0.25 | 0.49 | 0.0063 | -0.059 | 0.05 | 0.13 | 0.28 | -0.011 |
| waterfront | 0.27 | -0.0066 | 0.064 | 0.1 | 0.022 | 0.024 | 1 | 0.4 | 0.017 | 0.083 | 0.072 | 0.081 | -0.026 | 0.093 | 0.03 | -0.014 | -0.042 | 0.086 | 0.031 |
| view | 0.4 | 0.08 | 0.19 | 0.28 | 0.075 | 0.029 | 0.4 | 1 | 0.046 | 0.25 | 0.17 | 0.28 | -0.053 | 0.1 | 0.085 | 0.0062 | -0.078 | 0.28 | 0.073 |
| condition | 0.036 | 0.028 | -0.12 | -0.059 | -0.009 | -0.26 | 0.017 | 0.046 | 1 | -0.14 | -0.16 | 0.17 | -0.36 | -0.061 | 0.003 | -0.015 | -0.11 | -0.093 | -0.0034 |
| grade | 0.67 | 0.36 | 0.66 | 0.76 | 0.11 | 0.46 | 0.083 | 0.25 | -0.14 | 1 | 0.76 | 0.17 | 0.45 | 0.014 | -0.18 | 0.11 | 0.2 | 0.71 | 0.12 |
| sqft_above | 0.61 | 0.48 | 0.69 | 0.88 | 0.18 | 0.52 | 0.072 | 0.17 | -0.16 | 0.76 | 1 | -0.052 | 0.42 | 0.023 | -0.26 | -0.00082 | 0.34 | 0.73 | 0.19 |
| sqft_basement | 0.32 | 0.3 | 0.28 | 0.44 | 0.015 | -0.25 | 0.081 | 0.28 | 0.17 | 0.17 | -0.052 | 1 | -0.13 | 0.071 | 0.075 | 0.11 | -0.14 | 0.2 | 0.017 |
| yr_built | 0.054 | 0.15 | 0.51 | 0.32 | 0.053 | 0.49 | -0.026 | -0.053 | -0.36 | 0.45 | 0.42 | -0.13 | 1 | -0.22 | -0.35 | -0.15 | 0.41 | 0.33 | 0.071 |
| yr_renovated | 0.13 | 0.019 | 0.051 | 0.055 | 0.0076 | 0.0063 | 0.093 | 0.1 | -0.061 | 0.014 | 0.023 | 0.071 | -0.22 | 1 | 0.064 | 0.029 | -0.068 | -0.0027 | 0.0079 |
| zipcode | -0.053 | -0.15 | -0.2 | -0.2 | -0.13 | -0.059 | 0.03 | 0.085 | 0.003 | -0.18 | -0.26 | 0.075 | -0.35 | 0.064 | 1 | 0.27 | -0.56 | -0.28 | -0.15 |
| lat | 0.31 | -0.0089 | 0.025 | 0.053 | -0.086 | 0.05 | -0.014 | 0.0062 | -0.015 | 0.11 | -0.00082 | 0.11 | -0.15 | 0.029 | 0.27 | 1 | -0.14 | 0.049 | -0.086 |
| long | 0.022 | 0.13 | 0.22 | 0.24 | 0.23 | 0.13 | -0.042 | -0.078 | -0.11 | 0.2 | 0.34 | -0.14 | 0.41 | -0.068 | -0.56 | -0.14 | 1 | 0.33 | 0.25 |
| sqft_living15 | 0.59 | 0.39 | 0.57 | 0.76 | 0.14 | 0.28 | 0.086 | 0.28 | -0.093 | 0.71 | 0.73 | 0.2 | 0.33 | -0.0027 | -0.28 | 0.049 | 0.33 | 1 | 0.18 |
| sqft_lot15 | 0.082 | 0.029 | 0.087 | 0.18 | 0.72 | -0.011 | 0.031 | 0.073 | -0.0034 | 0.12 | 0.19 | 0.017 | 0.071 | 0.0079 | -0.15 | -0.086 | 0.25 | 0.18 | 1 |

# Correlation between Features



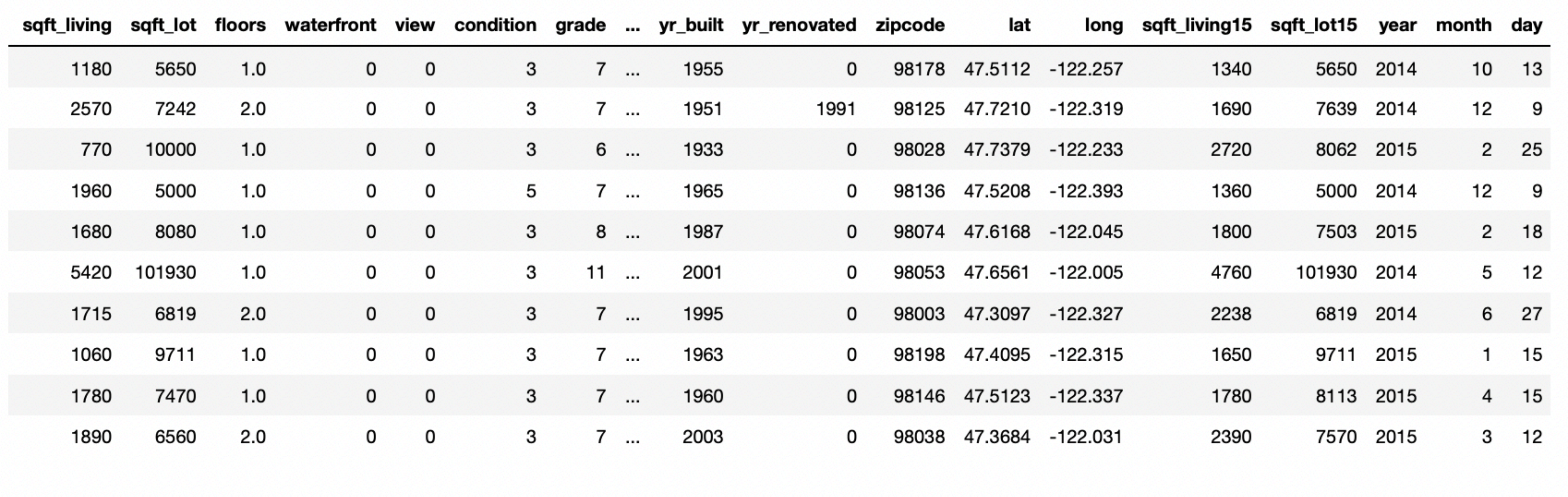






# Data Preparing

The date column is not numerical so we will split it into three columns which are year, month, day. Then we dropped the date column



| sqft_living | sqft_lot | floors | waterfront | view | condition | grade | ... | yr_built | yr_renovated | zipcode | lat | long | sqft_living15 | sqft_lot15 | year | month | day |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1180 | 5650 | 1.0 | 0 | 0 | 3 | 7 | ... | 1955 | 0 | 98178 | 47.5112 | -122.257 | 1340 | 5650 | 2014 | 10 | 13 |
| 2570 | 7242 | 2.0 | 0 | 0 | 3 | 7 | ... | 1951 | 1991 | 98125 | 47.7210 | -122.319 | 1690 | 7639 | 2014 | 12 | 9 |
| 770 | 10000 | 1.0 | 0 | 0 | 3 | 6 | ... | 1933 | 0 | 98028 | 47.7379 | -122.233 | 2720 | 8062 | 2015 | 2 | 25 |
| 1960 | 5000 | 1.0 | 0 | 0 | 5 | 7 | ... | 1965 | 0 | 98136 | 47.5208 | -122.393 | 1360 | 5000 | 2014 | 12 | 9 |
| 1680 | 8080 | 1.0 | 0 | 0 | 3 | 8 | ... | 1987 | 0 | 98074 | 47.6168 | -122.045 | 1800 | 7503 | 2015 | 2 | 18 |
| 5420 | 101930 | 1.0 | 0 | 0 | 3 | 11 | ... | 2001 | 0 | 98053 | 47.6561 | -122.005 | 4760 | 101930 | 2014 | 5 | 12 |
| 1715 | 6819 | 2.0 | 0 | 0 | 3 | 7 | ... | 1995 | 0 | 98003 | 47.3097 | -122.327 | 2238 | 6819 | 2014 | 6 | 27 |
| 1060 | 9711 | 1.0 | 0 | 0 | 3 | 7 | ... | 1963 | 0 | 98198 | 47.4095 | -122.315 | 1650 | 9711 | 2015 | 1 | 15 |
| 1780 | 7470 | 1.0 | 0 | 0 | 3 | 7 | ... | 1960 | 0 | 98146 | 47.5123 | -122.337 | 1780 | 8113 | 2015 | 4 | 15 |
| 1890 | 6560 | 2.0 | 0 | 0 | 3 | 7 | ... | 2003 | 0 | 98038 | 47.3684 | -122.031 | 2390 | 7570 | 2015 | 3 | 12 |

# Models and Tests

## 1- Linear Regression

RMSE : 395.28421

MAE : 330.71470

R^2 Train : 0.568

R^2 Test : 0.558

## 2- Linear Regression

RMSE : 395.28421

MAE : 330.71470

R^2 Train : 0.718

R^2 Test : 0.702

## Polynomial Regression

RMSE : 368.5382

MAE : 305.3620

R^2 Train : 0.805

R^2 Test : 0.775

# Models and Tests

## Random Forest

RMSE : 323.7482

MAE : 255.6092

R^2 Train : 0.982

R^2 Test : 0.866

## Decision Tree

RMSE : 379.26861

MAE : 301.9369

R^2 Train : 0.999

R^2 Test : 0.748

# Thank you