House Sales in King County

Kevin Huang, Joyce Hung, Afshan Ijaz, Hamza Kiani

Purpose of Analysis

Real Estate Price Prediction in King County:

- To predict the sale price of houses in King County with high accuracy

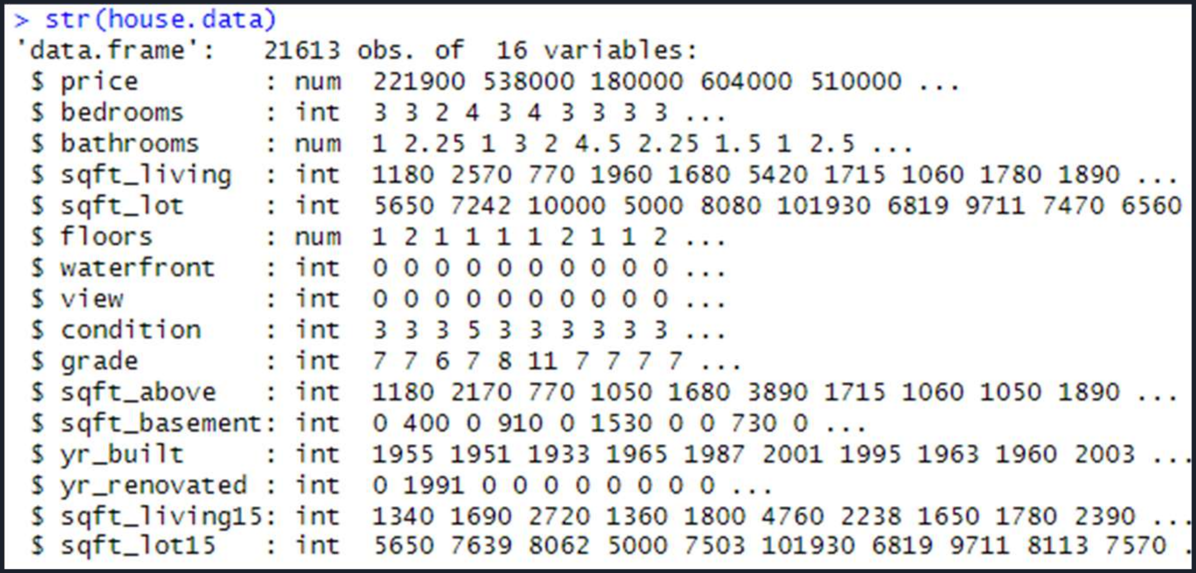
Why this Dataset?

Dataset contains house sale prices in King County for the year 2014-2015

- A big collection of variables for a house price prediction
- All variables are numeric which is convenient for linear regression
- At least two categorical variables

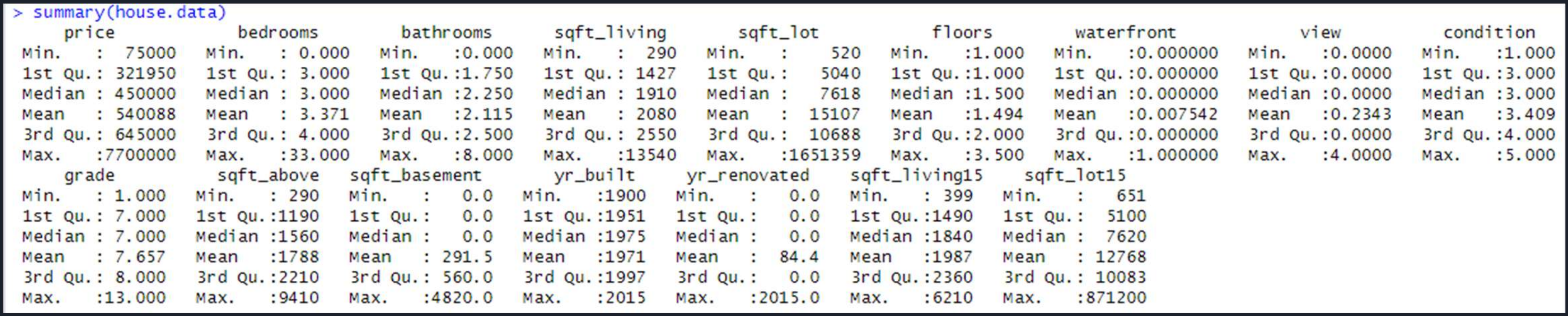
Some attributes like Data, Zip-code, Latitude, and Longitude were removed to make the dataset more manageable

Exploratory Data Analysis: Structure



```
> str(house.data)
'data.frame':   21613 obs. of  16 variables:
 $ price        : num  221900 538000 180000 604000 510000 ...
 $ bedrooms     : int  3 3 2 4 3 4 3 3 3 3 ...
 $ bathrooms    : num  1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
 $ sqft_living  : int  1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
 $ sqft_lot     : int  5650 7242 10000 5000 8080 101930 6819 9711 7470 6560
 $ floors       : num  1 2 1 1 1 1 2 1 1 2 ...
 $ waterfront   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ view         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ condition    : int  3 3 3 5 3 3 3 3 3 3 ...
 $ grade        : int  7 7 6 7 8 11 7 7 7 7 ...
 $ sqft_above   : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
 $ sqft_basement: int  0 400 0 910 0 1530 0 0 730 0 ...
 $ yr_built     : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ..
 $ yr_renovated : int  0 1991 0 0 0 0 0 0 0 0 ...
 $ sqft_living15: int  1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ..
 $ sqft_lot15   : int  5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 .
```

Exploratory Data Analysis: Summary

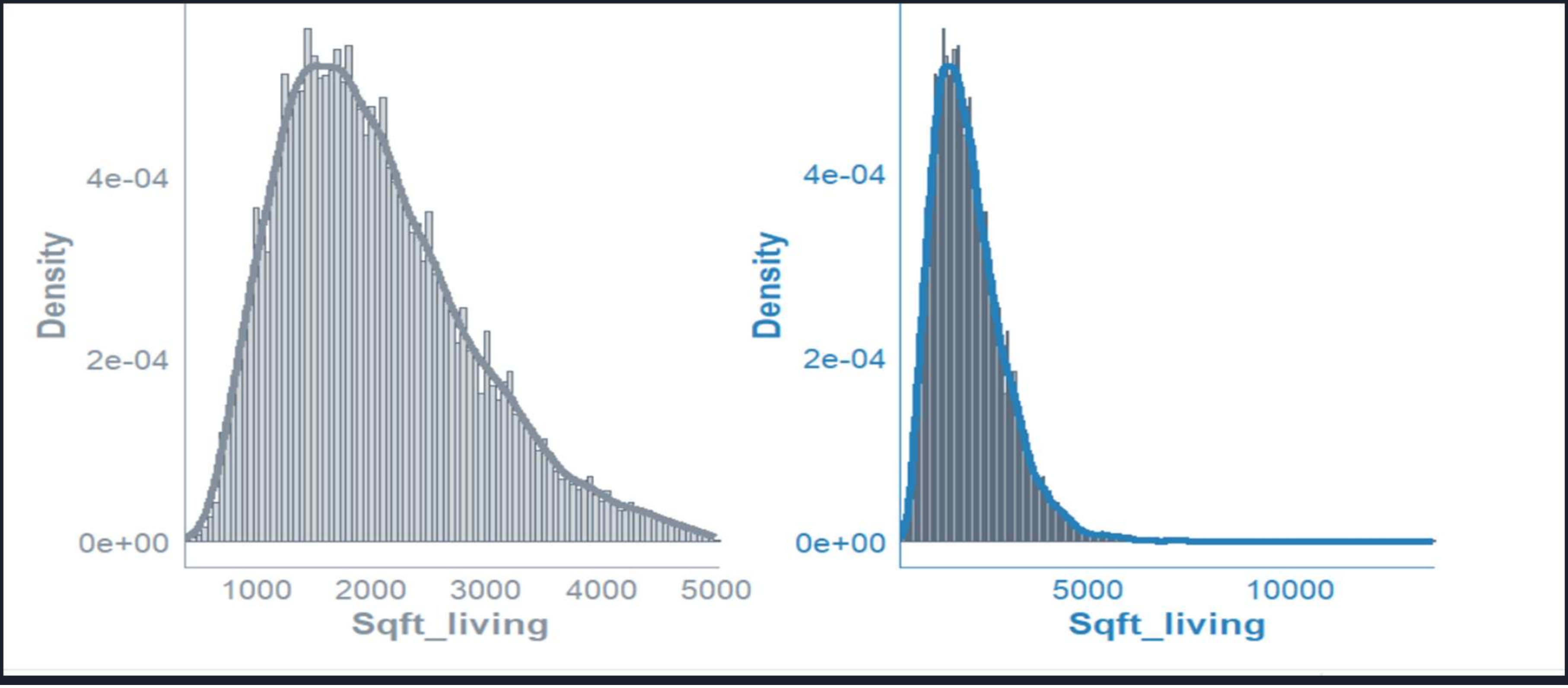


```
> summary(house.data)
     price            bedrooms        bathrooms      sqft_living       sqft_lot           floors        waterfront            view          condition    
 Min.   :  75000   Min.   : 0.000   Min.   :0.000   Min.   :  290   Min.   :    520   Min.   :1.000   Min.   :0.000000   Min.   :0.0000   Min.   :1.000  
 1st Qu.: 321950   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1427   1st Qu.:   5040   1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000  
 Median : 450000   Median : 3.000   Median :2.250   Median : 1910   Median :   7618   Median :1.500   Median :0.000000   Median :0.0000   Median :3.000  
 Mean   : 540088   Mean   : 3.371   Mean   :2.115   Mean   : 2080   Mean   :  15107   Mean   :1.494   Mean   :0.007542   Mean   :0.2343   Mean   :3.409  
 3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10688   3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000  
 Max.   :7700000   Max.   :33.000   Max.   :8.000   Max.   :13540   Max.   :1651359   Max.   :3.500   Max.   :1.000000   Max.   :4.0000   Max.   :5.000  
     grade          sqft_above   sqft_basement       yr_built     yr_renovated     sqft_living15    sqft_lot15    
 Min.   : 1.000   Min.   : 290   Min.   :   0.0   Min.   :1900   Min.   :   0.0   Min.   : 399   Min.   :   651  
 1st Qu.: 7.000   1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.0   1st Qu.:1490   1st Qu.:  5100  
 Median : 7.000   Median :1560   Median :   0.0   Median :1975   Median :   0.0   Median :1840   Median :  7620  
 Mean   : 7.657   Mean   :1788   Mean   : 291.5   Mean   :1971   Mean   :  84.4   Mean   :1987   Mean   : 12768  
 3rd Qu.: 8.000   3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.0   3rd Qu.:2360   3rd Qu.: 10083  
 Max.   :13.000   Max.   :9410   Max.   :4820.0   Max.   :2015   Max.   :2015.0   Max.   :6210   Max.   :871200  
```

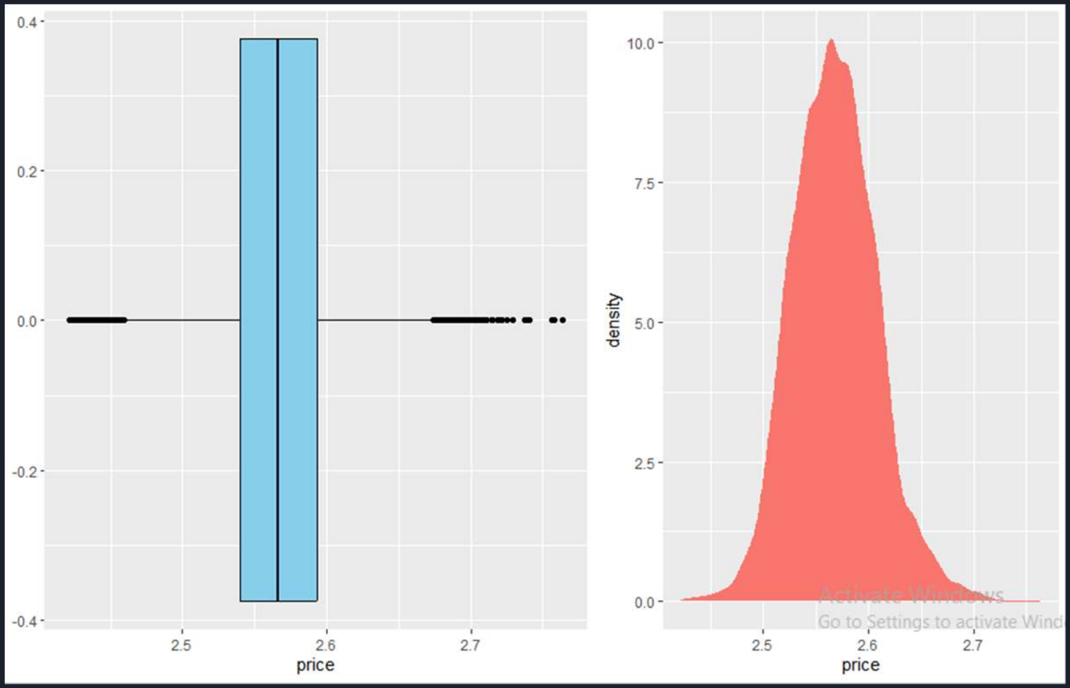
Exploratory Data Analysis: Histograms/Density Plots

Distribution of Sqft_living (without outliers)

Distribution of Sqft_living (with outliers)



Exploratory Data Analysis:
Response Variable - Outliers and Normality

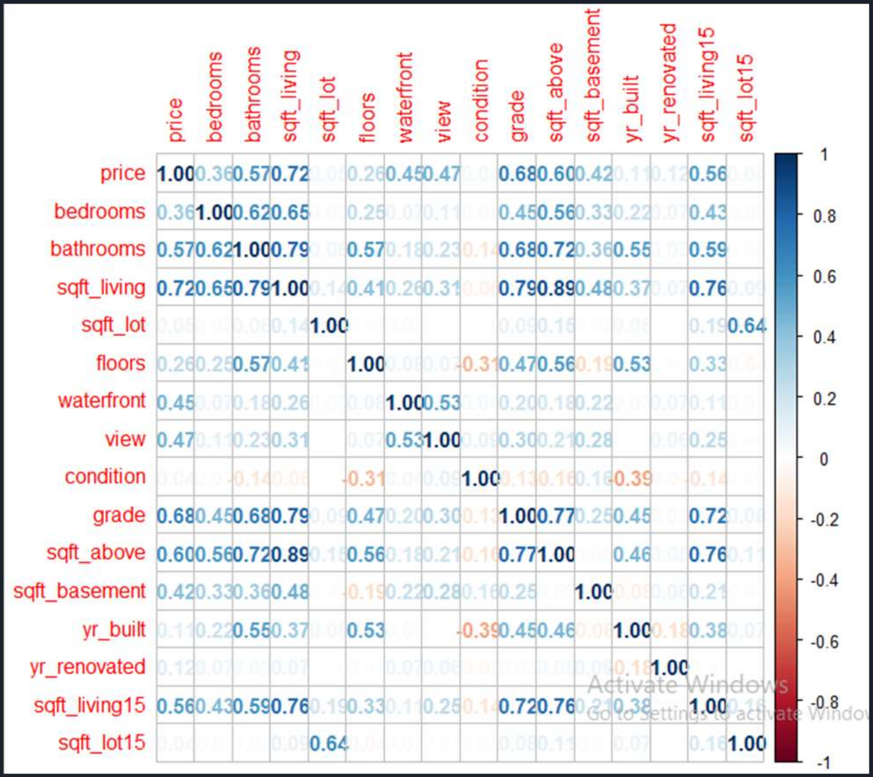


Multiple Linear Regression Assumptions

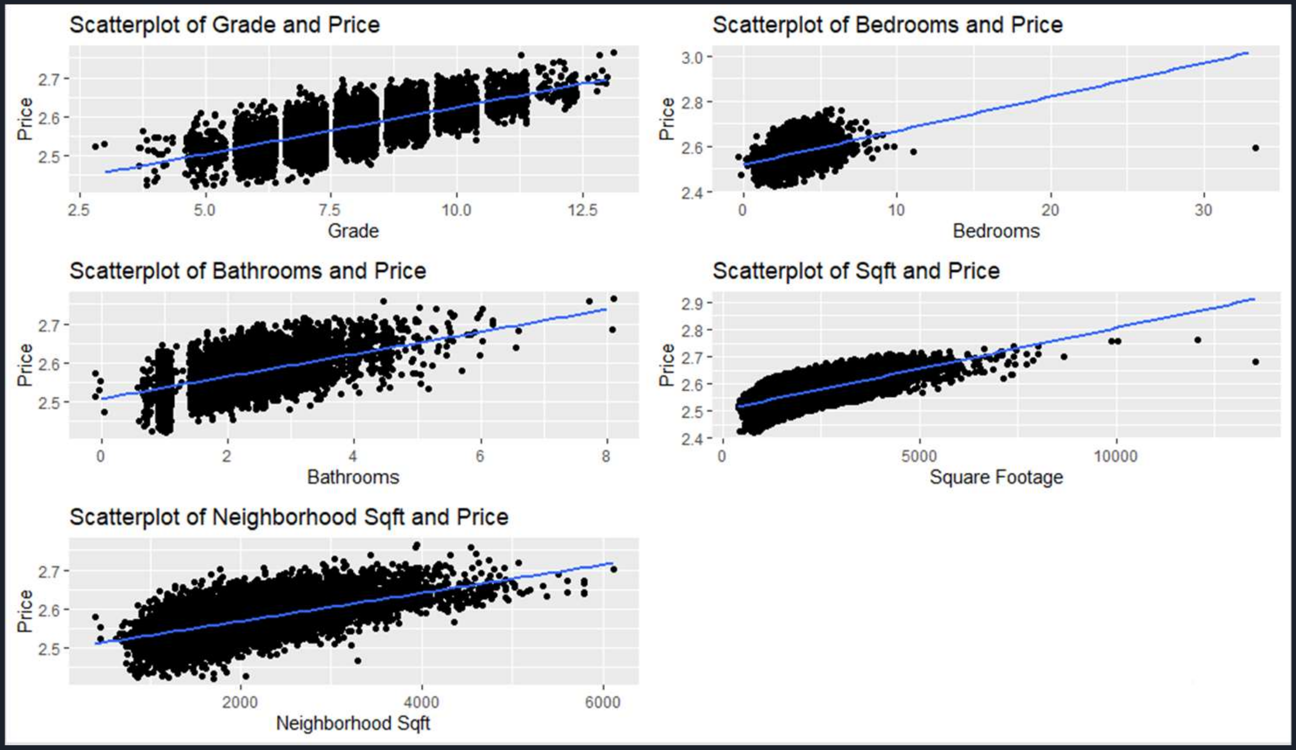
- Independence of data
- Linear relationship between explanatory and response variables
- Homoscedasticity (Variation of observations (residual SE) around regression line is constant)
- Normal distribution of model residuals for a given value of x

Multicollinearity: Correlation Plot

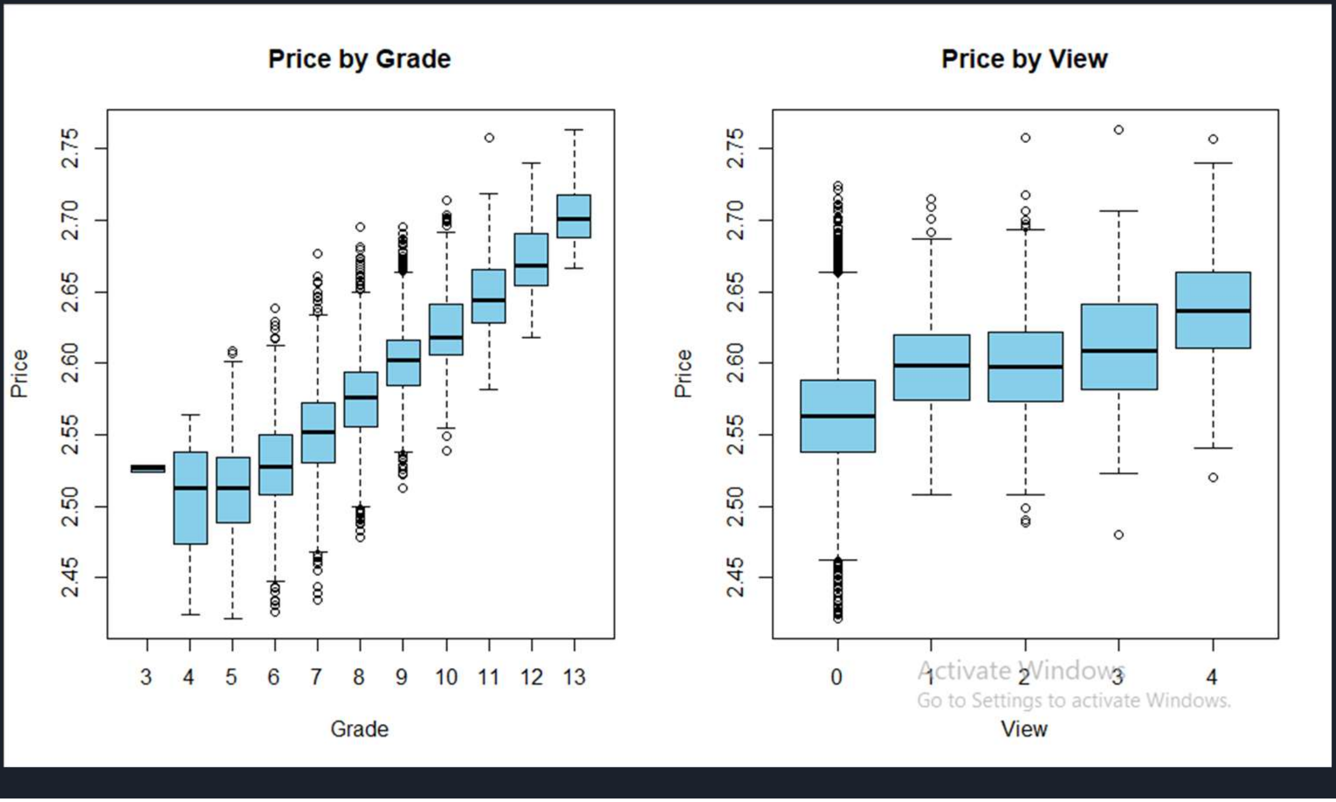
- Some independent variables highly correlated with each other (unsuitable for multiple regression)
- Based on this plot, independent variables with mutual correlation less than 0.8 and high correlation with the response variable (price) were chosen for multiple regression model



Linearity: Scatterplots

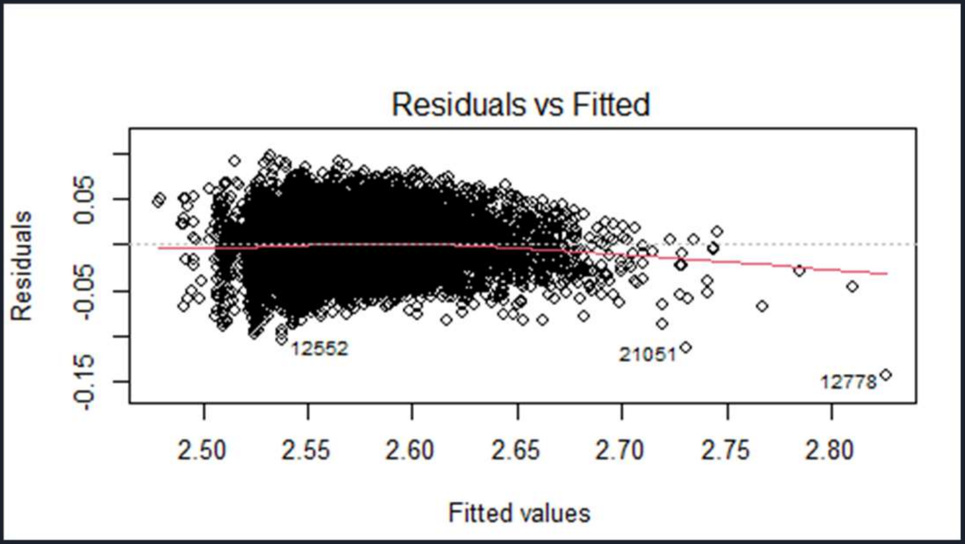


Relationship with Categorical Variables: Box Plots



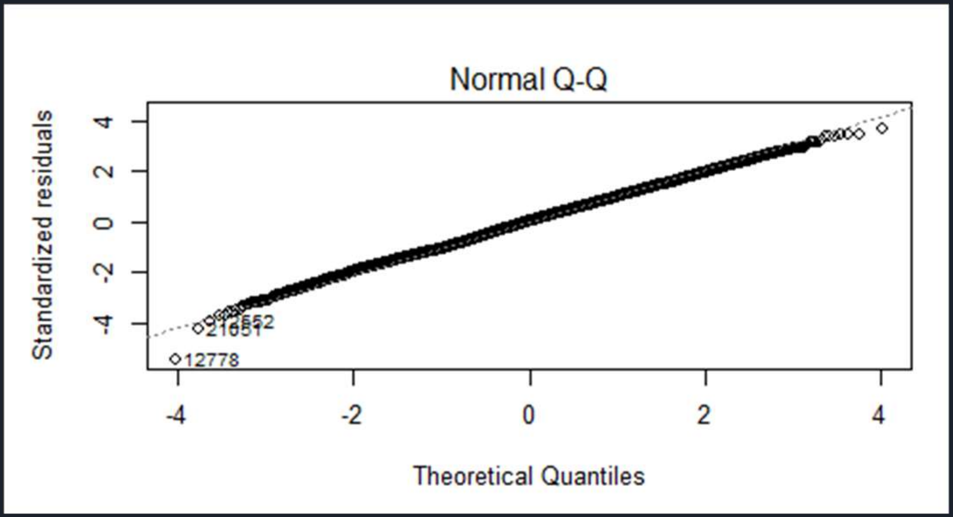
Linearity and Independence: Residuals vs Fitted Plot

- The scatterplots show a positive linear relationship between the independent variables and the response variable
- The scatterplots, together with the residuals vs. fitted plot show that the linearity assumption is fulfilled
- No pattern to the measurements so independence assumption is also fulfilled



Normality:
Q-Q Plot

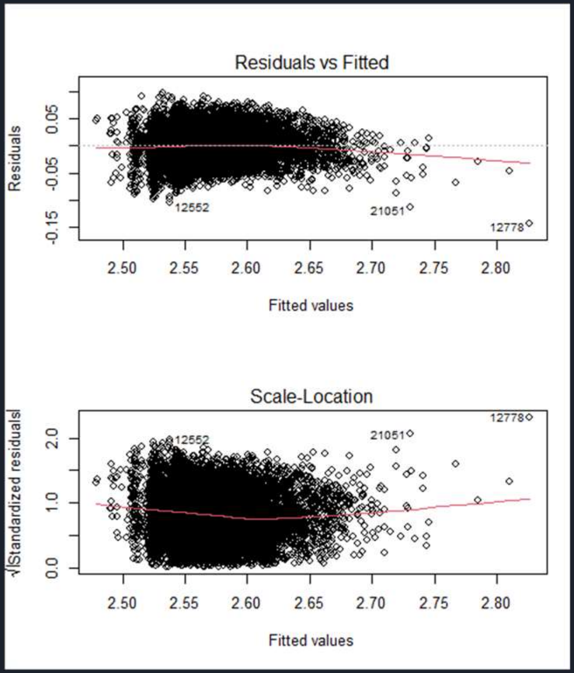
- Y axis is the ordered, observed, and standardized residuals
- X axis is the ordered, theoretical residuals
- The residuals fall along with Q-Q line, so the normality assumption is satisfied



Homoscedasticity:
Residual vs. Fitted and Scale-Location Plot

Both plots fall along a roughly horizontal line, indicating the variance of the residuals is the the same

Homoscedasticity assumption is satisfied



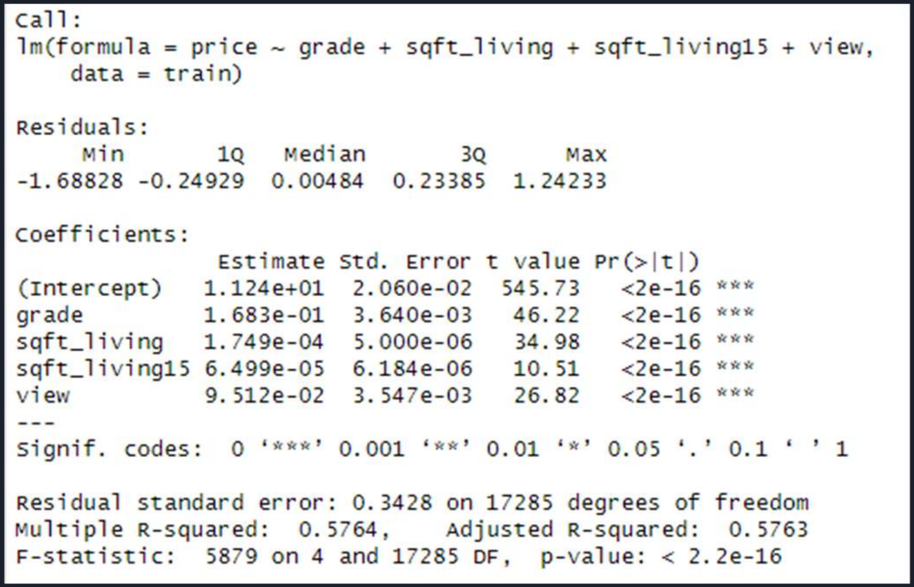
Steps for Multiple Linear Regression

1. Identify the independent and response variables:
    a. Y = House Price
    b. X = Square footage of the house, number of bedrooms, number of bathrooms, grade assigned to the house by King County (categorical variable), view, square footage of 15 nearest neighborhood houses
2. Test different combinations
3. Select the best model out of the tested combinations

Multiple Linear Regression

Combination with the smallest RMSE (0.34):

- Price ~ grade, bedrooms, bathrooms, sqft_living, view, sqft_living15
- All p-values are less than 0.05 indicating all parameters are significant
- R-squared shows moderate correlation (58%)



```
Call:
lm(formula = price ~ grade + sqft_living + sqft_living15 + view,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.68828 -0.24929  0.00484  0.23385  1.24233

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.124e+01  2.060e-02  545.73   <2e-16 ***
grade         1.683e-01  3.640e-03   46.22   <2e-16 ***
sqft_living   1.749e-04  5.000e-06   34.98   <2e-16 ***
sqft_living15 6.499e-05  6.184e-06   10.51   <2e-16 ***
view          9.512e-02  3.547e-03   26.82   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3428 on 17285 degrees of freedom
Multiple R-squared:  0.5764,    Adjusted R-squared:  0.5763
F-statistic:  5879 on 4 and 17285 DF,  p-value: < 2.2e-16
```

Interpretation of the Multiple Linear Regression Model:

- **MLR Equation:**

Price = 11.24 + 1.683e-01 * grade + 1.748e-04 * sq. ft. living + 9.512e-02 * view + 6.499e-05 * sq. ft. of 15 nearest neighborhood houses

For each increase of one unit in the parameter, the model estimates the increase in price by the parameter’s corresponding coefficient. We can see that grade carries the most influence on Price amongst these parameters

Given these parameters, the price of a house in King County can be predicted

Note: The price estimate calculated from this equation will give a log-price value. Taking an inverse log of this value will provide us with the actual sale price.

Multiple Linear Regression: Actual vs. Predicted

The actual values are close to the predicted price but there are some values with high differences

Actual Price ($)	Predicted Price ($)
180,000	151,027
291,850	287,616
468,000	302,926
252,700	287,687
535,000	384,979
322,500	453,819

Limitations and Possible Improvements

Limitations

- Given that the response variable needed to be log-transformed to satisfy regression assumptions, any new data in the model will also need to undergo similar processing
- Given regional differences in house prices, model's applicability to data from other counties might be limited

Improvements

- Adding additional variables (house location, neighborhood comp, household income)
- Removing extreme outliers
- Try different train/test ratios
- Incorporating the zip-code in the combination to account for price variation by location