Home Sales Prices in King County, Washington

Ethan Ancell, Christine Huffman, Samuel Rands

December 19, 2019

1 Introduction

Our group selected a data set describing several features of homes sold in King County, Washington. Most of these measures were quantitative, such as the price at which the house sold (*price*), grade of the land on which the house was built (*grade*), number of bedrooms (*bedrooms*), number of bathrooms (*bathrooms*), the year the home was built (*yr_built*), square footage of the living space (*sqft_living*), and square footage of the lot (*sqft_lot*). The remaining variable included in the dataset, waterfront (*waterfront*), is qualitative and reports if the property is located on a waterfront. With *price* as the response variable with the other variables as predictors, we were able to build a prediction model with 21,000 observation points.

2 Data Summary

From the scatterplots shown in Figures 1 and 2, *price* increased at varying degrees as each predictor variable increased. This suggested there could be a linear relationship between the predictor variables and the response, meaning a linear modeling approach is appropriate. We began by exploring an ordinary least squares modeling approach.

Figure 1: Scatterplot of Price vs. Bedrooms, Bathrooms and Living Square Footage

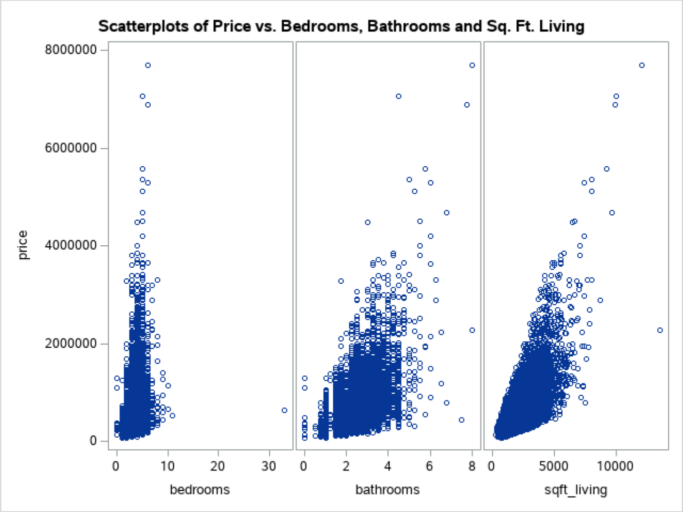
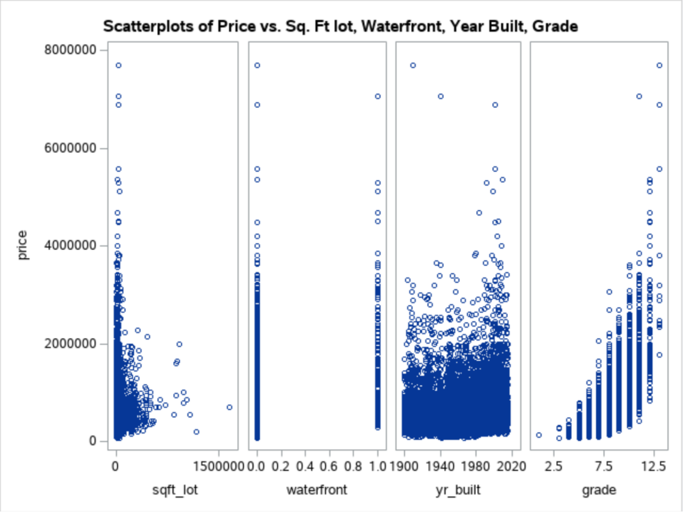


Figure 2: Scatterplot of Price vs. Lot Square Footage, Waterfront, Year Built and Grade



3 Preliminary Ordinary Least Squares Model

We regressed predictors on *price* to produce the following model:

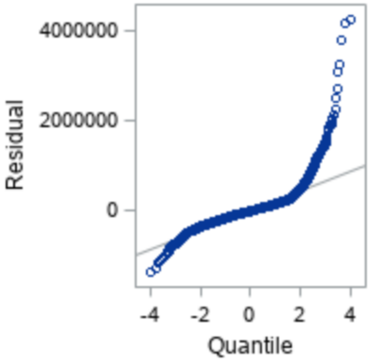
$$\hat{Y} = 6840365 - 41956(\text{bedrooms}) + 55747(\text{bathrooms}) + 176.621(\text{sqft\_living}) - 0.251(\text{sqft\_lot}) \\ +722210(\text{waterfront}) - 3879.494(\text{yr\_built}) + 130701(\text{grade})$$

4 Residual Assumptions

Ordinary least squares minimizes the sum square error, however the approach is only valid when certain assumptions regarding residuals are met. These assumptions are that the residuals are normally distributed, have constant variance and are independent.

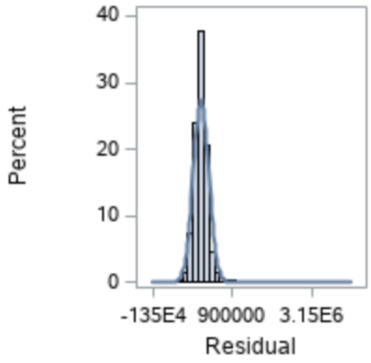
We assessed whether the residuals were normally distributed by examining residual graphs (the QQ plot and residual histogram). The QQ plot plotted the residuals from smallest to greatest against their quantile. The residual histogram plotted residual values on the horizontal axis and percentage of residuals on the vertical axis.

Figure 3: QQ plot for OLS pre-transformation



From the QQ plot in Figure 3, our data followed the normal distribution until quantile 2, then residuals surged far outside normal quantiles. From this, we suspected that our data is right-skewed.

Figure 4: Residual Histogram for OLS pre-transformation

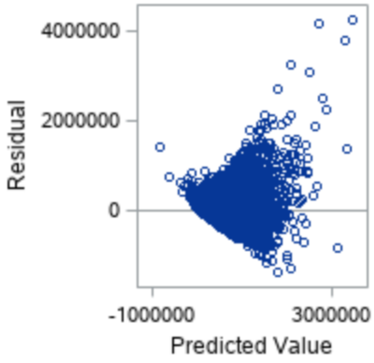


The residual histogram in Figure 4 confirmed our suspicion that the ordinary least squares model is right-skewed, as much of the residuals fall before the mean of the normal distribution.

From the QQ plot and residual histogram, there was reasonable evidence that this model violated the assumption that residuals be normally distributed.

We used the residual plot and the Brown-Forsythe test of constant variance to test whether our residuals had constant variance. The residual plot plotted residuals against the model's predicted value.

Figure 5: Residual Plot for OLS pre-transformation

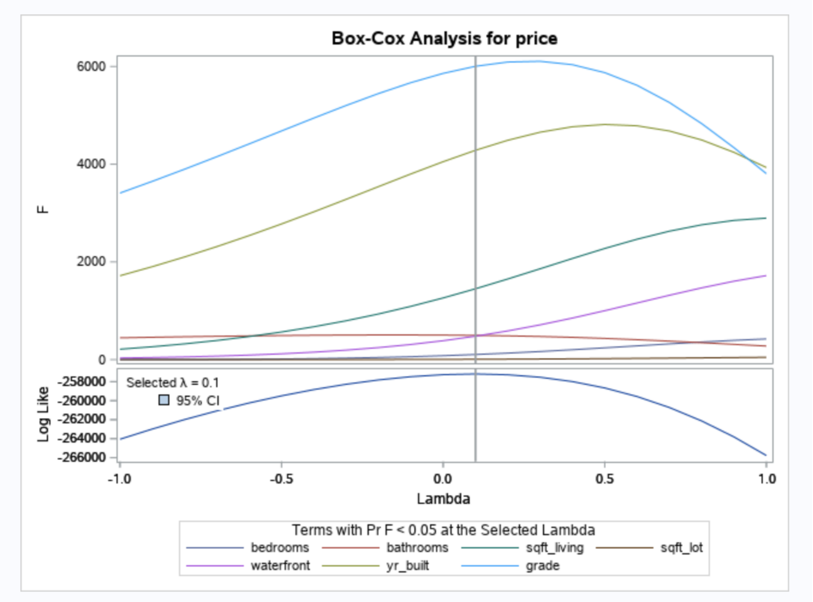


From Figure 5, this model's residuals are shaped like a "megaphone", meaning that points with lower predicted values had lower variance than points with higher predicted values. Thus, the residuals are heteroskedastic.

We assumed that the residuals were independent, since there didn't appear to be any pattern in the residuals.

To fix the non-normality distribution and non-constant variance of residuals, we transformed *price*. We used the Box-Cox approach to choose a transformation for *price*. The Box-Cox analysis shown in Figure 6 suggested using $\lambda = 0$, or a log transformation of *price*.

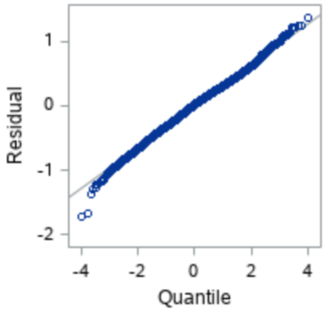
Figure 6: Box-Cox graph to decide transformation



We used the QQ plot and residual histogram to test whether the transformed model had normally distributed residuals.

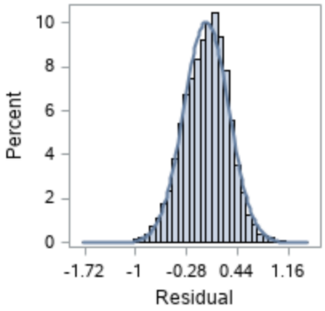
The post-transformation QQ plot in Figure 7 showed data that follow the normal distribution closely.

Figure 7: QQ plot after log-transformation



The post-transformation residual histogram in Figure 8 confirmed that the transformed model's residuals closely followed the normal distribution.

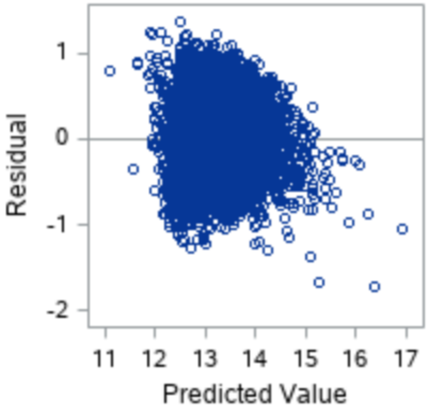
Figure 8: Residual Histogram after log-transformation



From these two plots, there was sufficient evidence that the transformed model did not violate the assumption that residuals be normally distributed.

We tested whether the transformed model's residuals were heteroskedastic using the residual plot and Brown-Forsythe test of variance.

Figure 9: Residual plot after log-transformation



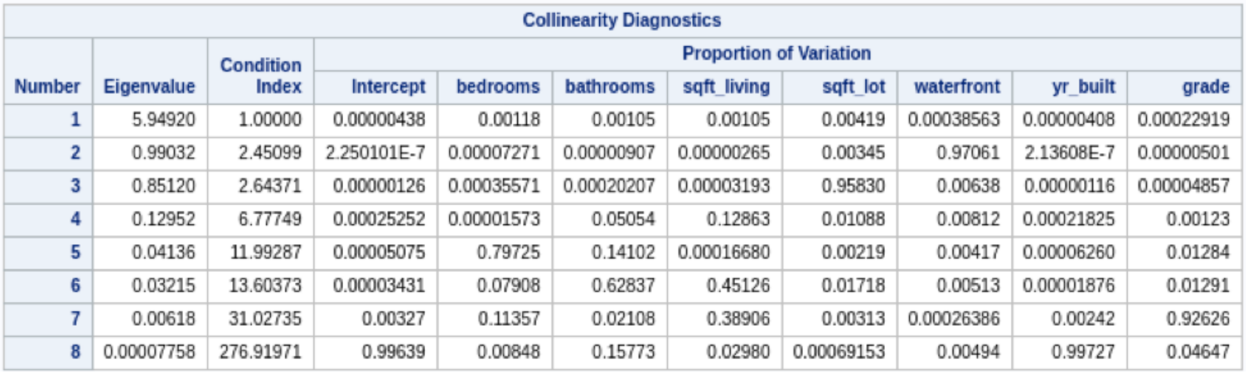
The residual plot showed a nebulous cloud of points, however it did seem that variance of higher predicted values were lower than the variance of lower predicted values.

The Brown-Forsythe test of constant variance also showed non-constant variance. The p-value for the Brown-Forsythe test of constant variance is 0.0005, meaning that there is a minute chance of observing the residuals we had if they had constant variance. In this case, the best option for dealing with residual non-constant variance is Weighted Least Squares. However, as this course dealt with linear models, we used the transformed Ordinary Least Squares model.

5 Multicollinearity, Influential Points and Outliers

After checking residual assumptions, we examined whether our predictors were collinear and whether our data contained outliers or influential points. We used the condition index test to find any predictors in our model that were collinear. Whenever there is a condition index higher than ten and two variables that explain more than 50% of the variance, the test suggests that those two variables are collinear. As seen in Figure 10, we do not have multicollinearity.

Figure 10: Collinearity diagnostics for log-transformed model



Collinearity Diagnostics

			Proportion of Variation							
Number	Eigenvalue	Condition Index	Intercept	bedrooms	bathrooms	sqft_living	sqft_lot	waterfront	yr_built	grade
1	5.94920	1.00000	0.00000438	0.00118	0.00105	0.00105	0.00419	0.00038563	0.00000408	0.00022919
2	0.99032	2.45099	2.250101E-7	0.00007271	0.00000907	0.00000265	0.00345	0.97061	2.13608E-7	0.00000501
3	0.85120	2.64371	0.00000126	0.00035571	0.00020207	0.00003193	0.95830	0.00638	0.00000116	0.00004857
4	0.12952	6.77749	0.00025252	0.00001573	0.05054	0.12863	0.01088	0.00812	0.00021825	0.00123
5	0.04136	11.99287	0.00005075	0.79725	0.14102	0.00016680	0.00219	0.00417	0.00006260	0.01284
6	0.03215	13.60373	0.00003431	0.07908	0.62837	0.45126	0.01718	0.00513	0.00001876	0.01291
7	0.00618	31.02735	0.00327	0.11357	0.02108	0.38906	0.00313	0.00026386	0.00242	0.92626
8	0.00007758	276.91971	0.99639	0.00848	0.15773	0.02980	0.00069153	0.00494	0.99727	0.04647

We tested for influential points using the Cook's D diagnostic. The graph obtained is shown in Figure 11.

Figure 11: Cook's D for log-transformed model

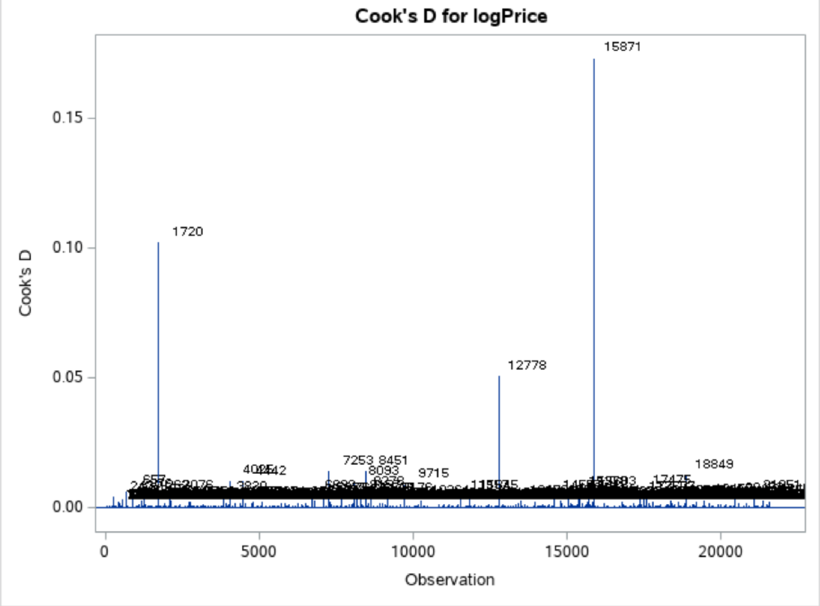


Figure 11 showed that observations 1720, 12778, and 15871 were influential points. To address these points, we examined each of these houses in Google Earth using the latitudinal and longitudinal information contained in the data set.

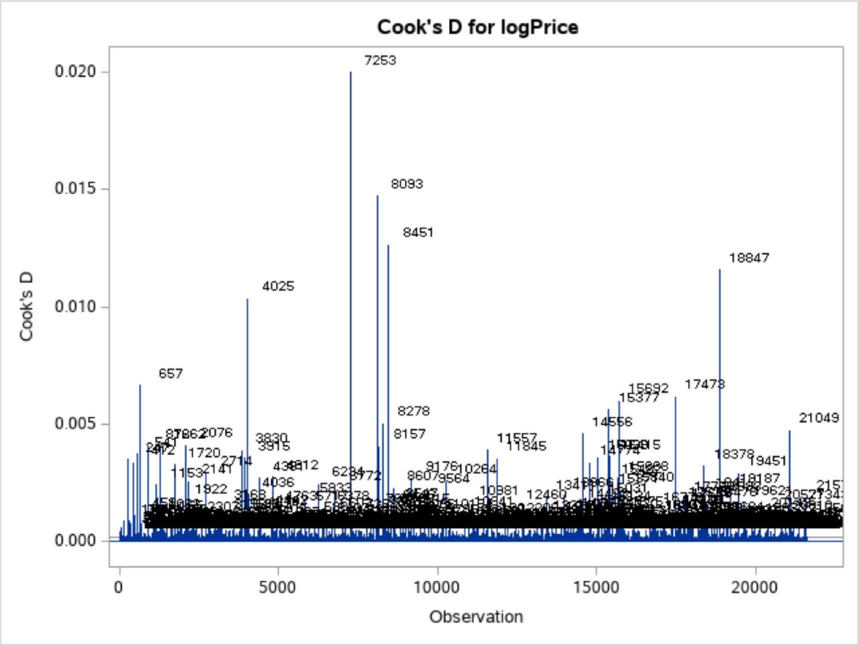
In observation 1720, we saw an abnormal lot square footage of around 1.65 million square feet. In Google Earth, we found that this observation was a farm. Since this point was only abnormal in lot square footage, we used a log-transformation on *sqft_lot*.

Observation 12778 was above average in nearly all predictor variables. As this observation was far above average in many different predictors, a transformation was unlikely to reduce its influence. However, since this house was atypical of a possible King County House, we removed it from our model.

Observation 15871 had typical numbers for all predictors except bedrooms, where it was recorded as having 33 bedrooms. Using Google Earth, we determined that the house was in a suburban neighborhood and likely did not have 33 bedrooms. As this observation is probably an outlier, we decided to remove it.

After removing observations 12778 and 15871 and transforming *sqft_lot*, we produced the following Cook's D graph shown in Figure 12.

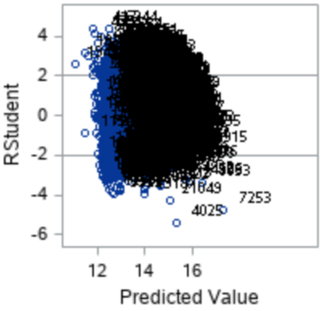
Figure 12: Cook's D for log-transformed model



We were satisfied with these influential point diagnostics. The transformation and removal (due to different populations) addressed the problem of influential points in the model.

We tested for the existence of outliers using the RStudent plot. The transformed model's highest RStudent values, shown in Figure 13, were approximately 4.5.

Figure 13: RStudent Plot for log-transformed model



We tested if these high RStudent values were outliers using the normal distribution. We used the normal distribution over the t-distribution since the normal distribution follows the t-distribution for large values of n. We used a threshold value of 0.01 to test for outliers, and because of multiple hypothesis testing, we used the Bonferroni adjustment to get a threshold of $\frac{0.01}{2 \cdot n} = \frac{0.01}{2 \cdot 21611} = 0.00000023146$. The critical value in the normal distribution for this threshold was 5.0411. Since none of the RStudent values exceeded this critical value, we determined that our data set didn't contain outliers.

6 Potential Interaction Terms

The two potential interactions we chose to look at were between *waterfront* and *logSqftLot* and between *year built* and *sqftliving*. Square footage probably affected sale price more for waterfront properties than non-waterfront properties. We also examined whether the living room's square footage had a higher increase on the sale price for older homes.

We used the t-test to decide whether the interaction terms ought to remain in the model. The p-value for the t-test concerning *yrBuiltSqftLiving* was less than 0.0001 and the p-value for the t-test concerning *waterfrontlogSqftLot* was 0.0024. Since the tests for *yrBuiltSqftLiving* and *waterfrontlogSqftLot* were both statistically significant, there is ample evidence to keep these two interaction terms in our model.

7 Interpretation of Final Ordinary Least Squares Model

Our final model equation is:

$$\begin{aligned}\ln \hat{Y} = 21.17433 - 0.03139(\text{bedrooms}) + 0.08624(\text{bathrooms}) + 0.00088244(\text{sqft\_living}) - 0.05240(\ln(\text{sqft\_lot}))\\ +1.43285(\text{waterfront}) - 0.0050(\text{yr\_built}) + 0.22916(\text{grade}) - 0.00000034(\text{year\_built} \cdot \text{sqft\_living})\\ -0.09477(\text{waterfront} \cdot \ln(\text{sqft\_lot}))\end{aligned}$$

These terms were all selected because of their impact on the slope. All variables included in our final model had a highly significant p-value. In fact, the only variable with a p-value greater than 0.0001 is the interaction term $waterfront \cdot ln(sqft\_lot)$, which had a value of 0.0024. This p-value is still well below the $\alpha = 0.05$ significance cutoff level and was thus included in the model.

Our final model's R^2 was 0.6428. This means our model explains 0.6428 variation in the response variable.

8 Alternative Model

As identified in the residual assumptions section, we had non-constant variance in the residuals. We solved this problem using weighted least squares regression.

The weighted model in equation form was:

$$\begin{aligned}\ln \hat{Y} = 21.8286 - 0.0314(\text{bedrooms}) + 0.08397(\text{bathrooms}) + 0.00061231(\text{sqft\_living}) - 0.06597(\text{sqft\_lot})\\ +1.29660(\text{waterfront}) - 0.00531(\text{yr\_built}) + 0.22947(\text{grade}) - 0.0000002003(\text{year\_built} \cdot \text{sqft\_living})\\ -0.08011(\text{waterfront} \cdot \ln(\text{sqft\_lot}))\end{aligned}$$

We cross-validated the predictions to test if the weighted model performed better than the transformed ordinary least squares model. A training set was created using 14000 observations and a test set was created from the other 7000 (or so) observations. We observed an MSPR of 150,255,146,757 for the ordinary least squares regression, and an MSPR of 127,867,708,592 for the weighted least squares regression.

These numbers were large because of the log-transformation on the *price* variable, so an outlier in the positive direction artificially inflated the MSPR. Because of this, we found the MSPR on the log-scale as well. We observed an MSPR of 0.096667 for the ordinary least squares regression and an MSPR of 0.096583 for the weighted least squares regression.

Nonetheless, the weighted model outperformed the ordinary least squares model on the test data. Furthermore, the ordinary least squares model didn't even meet residual assumptions. For these reasons, we opted to use the weighted least squares model.

9 Future Research Directions and Conclusion

The price of a home is location dependent. Two houses with the same X-profile might sell for different prices if one is located in a richer city like Medina, and the other is located in a poorer city like Renton. A powerful step forward in research would be to look at spatial statistic techniques to utilize the longitude and latitude points of the data set for predictions. Given the amount of data points, a heat map of high selling homes could be used to adjust the predicted sale price by a given home's proximity to other expensive houses.

Given the final model obtained from both ordinary least squares and weighted least squares, we can accurately predict the sale price of a home in King County.

We originally expected that all quantitative aspects of the house would have positive coefficients. It surprised us to see that bedrooms and square footage of lot had negative coefficients.

One would assume that the more bedrooms a house has, the more it will sell for, on average, because more rooms means you have a bigger house. However, it had a negative coefficient which means that, on average, per unit increase in number of bedrooms, the house would sell for less. Most likely, this is because houses with less bedrooms are bought and sold by single people who have more money. Those who have families will probably look for homes that are more family friendly and not in as expensive areas.

We would also assume that a house with more lot square footage would sell for more, because bigger lots cost more if you hold all else constant. However, the information that is contained in lot square footage is probably contained within location data. Houses that have a bigger lot size are typically not near the city and are in the more rural parts of King County, and will sell for less due to not being right downtown.

The other variable that we put in just to see what happened, but ended up having a coefficient that was large in magnitude was the grade of the land. It surprised me to see that the angle that your land is on is positively associated with having a more expensive home. This is one of those situations where grade doesn't cause the increase in price, but is associated with some other factors like location on a mountainside, where expensive homes happen to be.

This model is useful to someone selling their home in King County and desiring an estimate of their sale price, assuming that the home was sold in the years 2014 to 2015. Spatial data was not taken into consideration, so this home would be better used as a ballpark estimate of how much the home costs.

10 Appendix

Source for the data: https://www.kaggle.com/harlfoxem/housesalesprediction

10.1 Data Import

```
proc import datafile='/folders/myfolders/EPG194/stat5100project/kc_house_data.csv' replace
        out = house
        dbms = CSV
        ;
run;
```

10.2 Data Summary

```
proc sgscatter data=house;
    compare y=price
            x=(bedrooms bathrooms sqft_living);
    title1 'Scatterplots of Price vs. Bedrooms, Bathrooms and Sq. Ft. Living';
run;

proc sgscatter data=house;
    compare y=price
            x=(sqft_lot waterfront yr_built grade);
title1 'Scatterplots of Price vs. Sq. Ft lot, Waterfront, Year Built, Grade';
run;
```

10.3 Preliminary Ordinary Least Squares

```
proc reg data=house plots(maxpoints=22000);
        model price = bedrooms bathrooms sqft_living sqft_lot waterfront yr_built grade;
run;
```

10.4 Residual Assumptions

```
/* Box-cox transformation. */
proc transreg data=house;
        model boxcox(price / lambda=-1 to 1 by 0.1) = identity(bedrooms bathrooms
        sqft_living sqft_lot waterfront yr_built grade);
        title1 'Box-cox transformation of predicting price';
run;

/* Fixed model violations with log transform */
data house; set house;
        logPrice = log(price);
run;
proc reg data=house plots(maxpoints=220000 label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
        model logPrice = bedrooms bathrooms sqft_living logSqftLot waterfront yr_built grade;
        output out = house_out r=resid p=pred;
run;

/** Brown-Forsythe and Correlation Test of Normality (shortcut) **/
filename macrourl url "http://www.stat.usu.edu/jrstevens/stat5100/resid_num_diag.sas";
```

```
    \%include macrourl;

/* Numeric diagnostics to support probably doing WLS */
\%resid_num_diag(dataset=house_out, datavar=resid,
   label='residual', predvar=pred, predlabel='predicted');
```

10.5 Multicollinearity, Influential Points, and Outliers

```
/* Multicollinearity diagnostics */
proc reg data=house;
        model logPrice = bedrooms bathrooms sqft_living
        sqft_lot waterfront yr_built grade / collin;
run;

/* Influential points */
proc reg data=house plots(maxpoints=220000 label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
        model logPrice = bedrooms bathrooms sqft_living
        sqft_lot waterfront yr_built grade;
run;

/* Remedial measures of influential points */
data house; set house;
        logSqftLot = log(sqft_lot);
        order = _n_;
run;

/* Refit model */
proc reg data=house plots(maxpoints=220000 label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
        where order NE 15871 and order NE 12778;
        model logPrice = bedrooms bathrooms sqft_living
        logSqftLot waterfront yr_built grade;
run;
```

10.6 Potential Interaction Terms

```
/* Look at interaction terms */
data house; set house;
        waterfrontLogSqftLot = logSqftLot * waterfront;
        yrBuiltSqftLiving = yr_built * sqft_living;
run;

/* Test for significance */
proc reg data=house;
        where order NE 15871 and order NE 12778;
        model logPrice = bedrooms bathrooms sqft_living logSqftLot waterfront
                    yr_built grade
                    yrBuiltSqftLiving waterfrontLogSqftLot;
run;
```

10.7 Alternative Model

```
/* Remove the influential observations. */
data house;
        set house;
        if id NE 2402100895 and id NE 5315100874;
run;

data house;
        set house;
        rand = ranuni(12);
run;
proc sort data=house;
        by rand;
run;

/* Transformations */
data house;
        set house;
        order=_n_;
        logPrice = log(price);
        logSqftLot = log(sqft_lot);
        waterfrontLogSqftLot = logSqftLot * waterfront;
        yrBuiltSqftLiving = yr_built * sqft_living;
run;

data train;
        set house;
        where order < 14000;
run;

data test;
        set house;
        where order GE 14000;
run;

data train;
        set train;
        train_logPrice = logPrice;
run;

data combine;
        set train test;
run;

/* Try OLS */
proc reg data=combine;
        model train_logPrice = bedrooms bathrooms sqft_living
        logSqftLot waterfront yr_built grade yrBuiltSqftLiving waterfrontLogSqftLot;
        output out = house_out r=resid p=pred;
run;
```

```
/* Try WLS */
data house_out2; set house_out;
        abs_resid = abs(resid);
run;
proc reg data=house_out2 noprint;
        model abs_resid = bedrooms bathrooms sqft_living
        logSqftLot waterfront yr_built grade yrBuiltSqftLiving waterfrontLogSqftLot;
        output out=house_out3 p=estSD;
run;
data house_out3; set house_out3;
        useWeight = 1/estSD**2;
run;
proc reg data=house_out3;
        model logPrice = bedrooms bathrooms sqft_living
        logSqftLot waterfront yr_built grade yrBuiltSqftLiving waterfrontLogSqftLot;
        weight useWeight;
        output out=house_out4 r=resid p=pred2;
run;

/* Compare the two */
/* OLS */
data compareOLS;
        set house_out;
        where train_logPrice=.;
        serr = (exp(pred) - price)**2;
        l_serr = (pred - logPrice)**2;
run;
proc means data=compareOLS;
        var serr;
run;

/* WLS */
data compareWLS;
        set house_out4;
        where train_logPrice=.;
        serr2 = (exp(pred2) - price)**2;
        l_serr2 = (prep2 - logPrice)**2;
run;
proc means data=compareWLS;
        var serr2;
run;

/* Now on log scale */
proc means data=compareOLS;
        var l_serr;
        title1 'MSPR for OLS on log-scale';
run;
proc means data=compareWLS;
        var l_serr2;
        title1 'MSPR for WLS on log-scale';
run;
```