Predicting King County, WA Home Sales Prices

Dennis Murray

November 7, 2016

#### Introduction

Home prices have become one of the most examined sets of data in the American culture. As homeowners, people are innately interested in the potential market value of their home. For real estate agents, the pricing of the home is a key to whether the property sells in a reasonable amount of time, or if the homeowner loses some of the potential value in their home by underpricing. Conventionally, homeowners and agents have used a rough system of "Comparable" sales of houses in the same neighborhood with similar feature sets. Websites like Zillow have been developed solely for the purpose of showing their own estimates of an individual home based on a black-box model.

library(ggplot2)  
library(GGally)  
library(pls)

##   
## Attaching package: 'pls'

## The following object is masked from 'package:stats':  
##   
## loadings

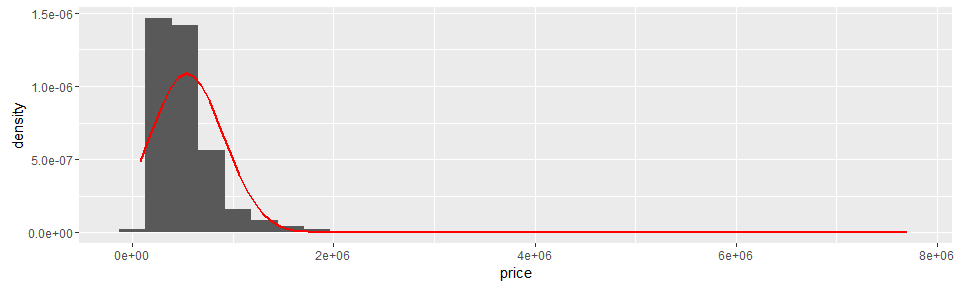
library(ggplot2)  
HomeSales<-read.csv("DATA/kc\_house\_data.csv")  
HomeSales$LogLotSqFt<-log(HomeSales$sqft\_lot)  
HomeSales$LogSalePrice<-log(HomeSales$price)  
HomeSales$ScaledLogPrice<-scale(HomeSales$LogSalePrice, scale=TRUE, center = TRUE)  
HomeSales$LogSqFtLiving<-log(HomeSales$sqft\_living)  
HomeSales$LogSqFtAbove<-log(HomeSales$sqft\_above)  
HomeSales$LogSqFtBasement<-log(HomeSales$sqft\_basement+1)  
HomeSales$SqRtSalesPrice<-sqrt(HomeSales$price)

#### Examining the Data

The King County home sales data set contains 21,613 observations of sales of homes in King County, Washington over a 12 month time from 2014 to 2015. In addition to price, the dataset includes many potential explanatory variables that may help to project the price of an individual property. For purposes of this evaluation, we will only include variables that can be evaluated on a continuous scale, and not any purely categorical variables.

Price: The dollar value paid by the purchaser. With a mean of $540,088 and a standard deviation of $367,127 there is a clear right skew of the data with an extremely long tail to the data. The maximum value of the price is $7.7MM, more than 200 standard deviations above the mean home sale value. This data will be transformed using a log transformation to reduce the skewness of the data, and return it closer to a normal distribution. Examination of a histogram after the transformation shows a clear improvement in the distribution.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

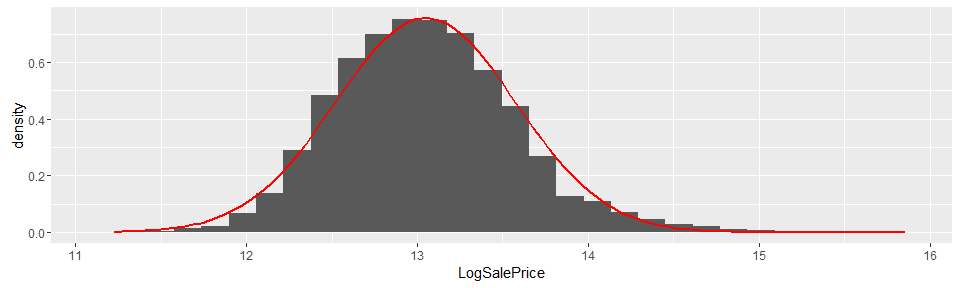


## [1] 540088.1

## [1] 367127.2

## [1] 450000

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



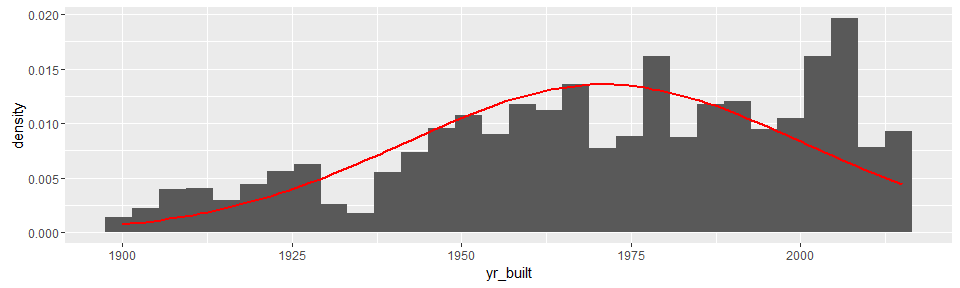
## [1] 13.04782

## [1] 0.5266845

## [1] 13.017

Year Built: The year that the house was constructed. Houses in the data set range from 1900 to 2015, with an increasing frequency from 1975 to 2005 and a clear deficit of houses built in the Depression and World War II era. As an explanatory variable, this may prove to be with lesser value - as the value of an older home could also be related to a desirable neighborhood, design or condition of the house.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## [1] 1971.005

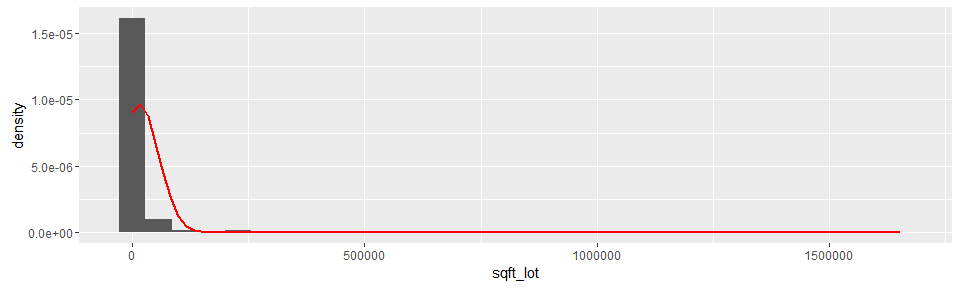
## [1] 29.37341

## [1] 1975

## [1] 1900 2015

Lot Size: In square feet, the size of the property that the home sits upon. Similar to price, there is an extremely large range of values for lot size - from just over 500 square feet to more than 1.6MM square feet (more than 39 standard deviations above the mean of 15,106). The mean value is heavily affected versus the median by these outliers, with the median at just more than half of the mean at 7,618 square feet. For these reasons, the data used in the model will be log transformed to return some degree of normality.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



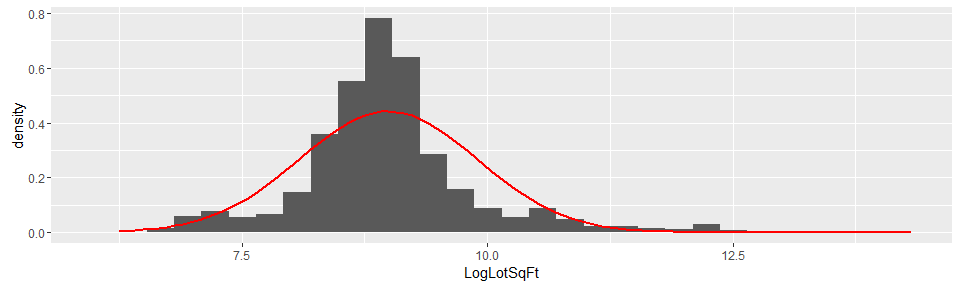
## [1] 15106.97

## [1] 41420.51

## [1] 7618

## [1] 520 1651359

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## [1] 8.989956

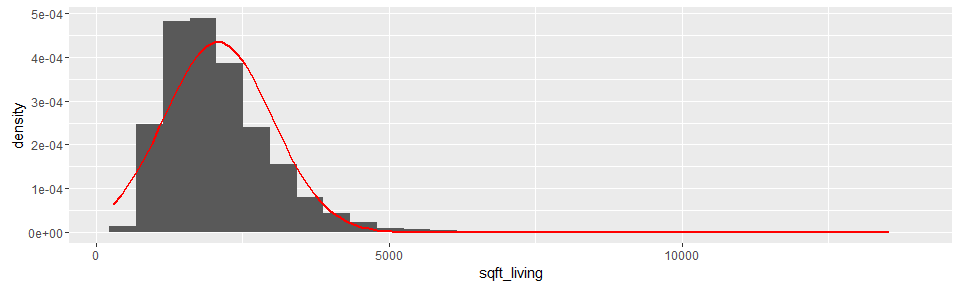
## [1] 0.9024251

## [1] 8.938269

## [1] 6.253829 14.317109

Square Footage: The data set includes several measures of interior square footage. A total square footage of living space is included, with secondary measures that separate square footage above grade from basement square footage. All three of these measures exhibit a similar skew to the lot size, with a large variance within the data and all three will be log transformed for the modeling.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



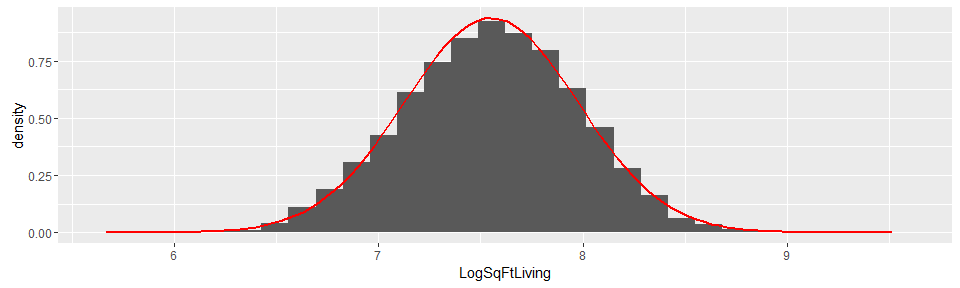
## [1] 2079.9

## [1] 918.4409

## [1] 1910

## [1] 290 13540

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## [1] 7.550335

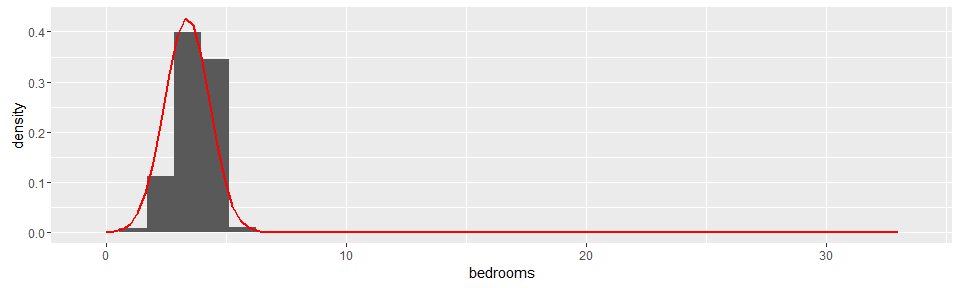
## [1] 0.4248069

## [1] 7.554859

## [1] 5.669881 9.513404

Bedrooms and Bathrooms: The data reports both the number of bedrooms as well as the number of bathrooms. We would expect both of these to increase the sale price of the home in a positive correlation. The number of bedrooms has a mean value of 3.4 with a standard deviation of <1 bedroom. Bathrooms average 2.1, with a standard deviation 0.77 and a median of 2.25.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



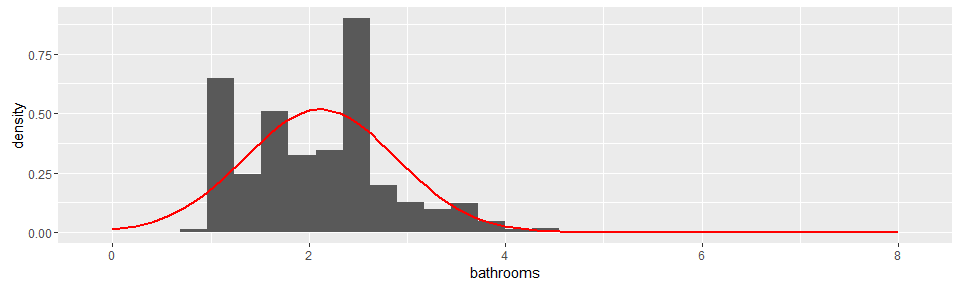
## [1] 3.370842

## [1] 0.9300618

## [1] 3

## [1] 0 33

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## [1] 2.114757

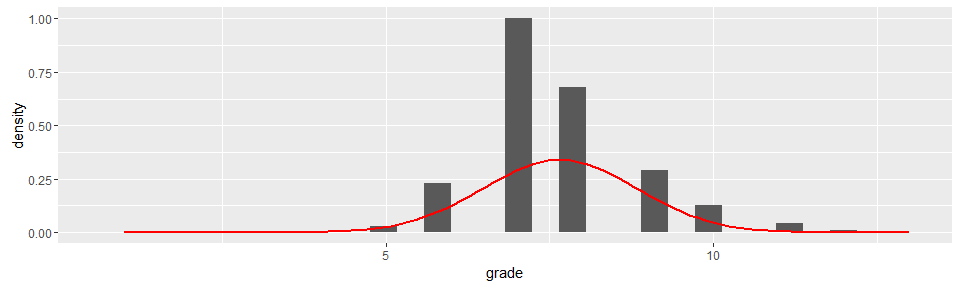
## [1] 0.7701632

## [1] 2.25

## [1] 0 8

Condition and Grade: These two measures are coded via inspector decision based on condition of the house at time of survey as well as the grade of materials used. Grade is measured on a scale of one to thirteen. Condition is measured on a scale of 1:5 with most homes surveyed in the range of three to five.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



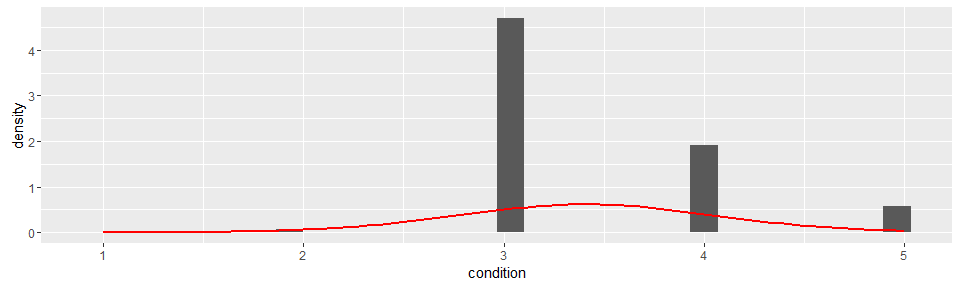
## [1] 7.656873

## [1] 1.175459

## [1] 7

## [1] 1 13

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## [1] 3.40943

## [1] 0.650743

## [1] 3

## [1] 1 5

Latitude and Longitude: An examination of a map of Seattle with the pricing superimposed gives the clear impression that there are some strong geographic components to pricing. While a categorical variable measure like neighborhood, or a distance from the city center might better represent the differences in pricing, use of latitude and longitude might suffice to help refine the model's projections.

round(cor(HomeSales[c(3:8,11:16)]),3)

## price bedrooms bathrooms sqft\_living sqft\_lot floors  
## price 1.000 0.308 0.525 0.702 0.090 0.257  
## bedrooms 0.308 1.000 0.516 0.577 0.032 0.175  
## bathrooms 0.525 0.516 1.000 0.755 0.088 0.501  
## sqft\_living 0.702 0.577 0.755 1.000 0.173 0.354  
## sqft\_lot 0.090 0.032 0.088 0.173 1.000 -0.005  
## floors 0.257 0.175 0.501 0.354 -0.005 1.000  
## condition 0.036 0.028 -0.125 -0.059 -0.009 -0.264  
## grade 0.667 0.357 0.665 0.763 0.114 0.458  
## sqft\_above 0.606 0.478 0.685 0.877 0.184 0.524  
## sqft\_basement 0.324 0.303 0.284 0.435 0.015 -0.246  
## yr\_built 0.054 0.154 0.506 0.318 0.053 0.489  
## yr\_renovated 0.126 0.019 0.051 0.055 0.008 0.006  
## condition grade sqft\_above sqft\_basement yr\_built  
## price 0.036 0.667 0.606 0.324 0.054  
## bedrooms 0.028 0.357 0.478 0.303 0.154  
## bathrooms -0.125 0.665 0.685 0.284 0.506  
## sqft\_living -0.059 0.763 0.877 0.435 0.318  
## sqft\_lot -0.009 0.114 0.184 0.015 0.053  
## floors -0.264 0.458 0.524 -0.246 0.489  
## condition 1.000 -0.145 -0.158 0.174 -0.361  
## grade -0.145 1.000 0.756 0.168 0.447  
## sqft\_above -0.158 0.756 1.000 -0.052 0.424  
## sqft\_basement 0.174 0.168 -0.052 1.000 -0.133  
## yr\_built -0.361 0.447 0.424 -0.133 1.000  
## yr\_renovated -0.061 0.014 0.023 0.071 -0.225  
## yr\_renovated  
## price 0.126  
## bedrooms 0.019  
## bathrooms 0.051  
## sqft\_living 0.055  
## sqft\_lot 0.008  
## floors 0.006  
## condition -0.061  
## grade 0.014  
## sqft\_above 0.023  
## sqft\_basement 0.071  
## yr\_built -0.225  
## yr\_renovated 1.000

The strongest correlations for price are Grade, Square Footage above ground, Square Feet of Living Space and Bathrooms. Surprisingly, lot size offers nearly no correlation to price - although this could also be reflective of confounding variables like location. Based on the correlations, I would expect the four variables mentioned will have the largest influence on the final model.

#### Splitting the Observations

As the data set contains more than 20,000 observations of price and the explanatory variables, there is an adequate supply of observations to use part of the data to create the model and part of the data to test the model. Random sampling of about one-third of the observations will be used to create the model.

#set.seed(1202104298)  
SampleCount<-round(length(HomeSales$id)/3.5)  
TrainDataSample<-sample(nrow(HomeSales), SampleCount, replace=FALSE)  
TrainDataSample2<-HomeSales[TrainDataSample,]  
TestData<-HomeSales[-TrainDataSample,]

#### Modeling With Principal Component Analysis

Our objective is to find the explanatory variables that explain the variation in price best, and then model the price equation. We will use principal component analysis and regression to complete this objective. There is likely several colinear variables when you consider the integrated development of a house. Larger houses are likely to have more bedrooms and bathrooms, and in return are likely to be on a larger lot. In a purely regression solution to this question, we could have severe variance inflation. Principal component analysis should help resolve these complex colinearity problems.

In this case, the principal component analysis will center and scale all variables based on their mean and standard deviation. In the results, it's important to consider that our input values for the variables are expressed in terms of one standard deviation of the value.

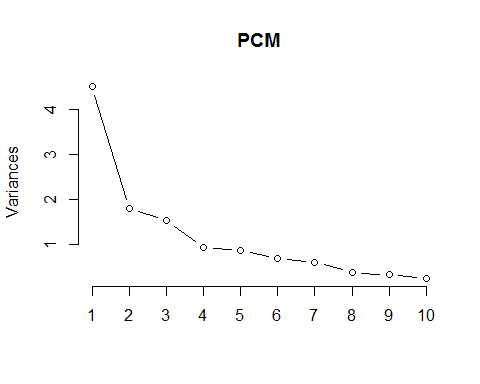
PCM<-prcomp(TrainDataSample2[c(4:5,8,11:12,15,18:19,22,25,26,27)], center=TRUE, scale=TRUE)  
PCM

## Standard deviations:  
## [1] 2.1233008 1.3364926 1.2337194 0.9620570 0.9275070 0.8227721 0.7700605  
## [8] 0.6076817 0.5590223 0.4853765 0.4387724 0.1328646  
##   
## Rotation:  
## PC1 PC2 PC3 PC4  
## bedrooms 0.284807804 0.31298514 0.06390571 -0.178292084  
## bathrooms 0.401204297 0.12096569 -0.13948881 -0.150164781  
## floors 0.284596202 -0.31037463 -0.35727117 0.023140684  
## condition -0.127990469 0.36711929 0.19173542 0.080184433  
## grade 0.389808716 0.07516150 -0.10755974 0.138767533  
## yr\_built 0.296189557 -0.35944578 -0.04601065 -0.219618259  
## lat -0.008658918 0.20376116 -0.38905219 0.808614395  
## long 0.202854990 -0.24663928 0.44670852 0.197405489  
## LogLotSqFt 0.130430282 0.14560406 0.63992159 0.227919500  
## LogSqFtLiving 0.419356505 0.27022416 0.02126877 -0.004936355  
## LogSqFtAbove 0.432157037 -0.01673771 0.07784841 0.152308417  
## LogSqFtBasement 0.002686115 0.56973394 -0.17732796 -0.321268021  
## PC5 PC6 PC7 PC8  
## bedrooms 0.159163246 -0.216311738 -0.74418615 -0.22660792  
## bathrooms 0.001607650 0.153128248 0.02711012 0.00721070  
## floors 0.317897855 0.003909662 0.02353411 0.41375666  
## condition 0.699219546 0.514446594 0.15445692 -0.14338320  
## grade -0.021098984 -0.022213131 0.44229177 -0.12571457  
## yr\_built -0.195701162 0.409365982 0.02728788 -0.64434524  
## lat -0.200757103 0.139729729 -0.20253505 -0.17873292  
## long -0.189589711 0.530534666 -0.31833502 0.42956516  
## LogLotSqFt -0.199751512 -0.248255975 0.24191719 -0.11829266  
## LogSqFtLiving -0.005793282 -0.070362478 0.08443274 0.14962066  
## LogSqFtAbove 0.198100776 -0.215169667 0.06761208 0.06638418  
## LogSqFtBasement -0.436494588 0.288348982 0.10996699 0.26191709  
## PC9 PC10 PC11 PC12  
## bedrooms -0.02794334 0.219655648 0.24148837 -0.0204659322  
## bathrooms 0.27415559 -0.782512765 0.26583219 -0.0019390775  
## floors 0.50070637 0.386811445 0.14201325 0.0342625851  
## condition 0.05797963 0.059943485 -0.02384582 -0.0121414358  
## grade -0.49442418 0.227441848 0.55019779 0.0175202082  
## yr\_built 0.14902837 0.198127975 -0.22214192 -0.0102597513  
## lat 0.12887956 -0.015538051 -0.05946624 -0.0024979004  
## long -0.21424525 0.004157218 0.10151882 -0.0015537190  
## LogLotSqFt 0.53723750 0.127735847 0.14611965 -0.0003922231  
## LogSqFtLiving -0.10982215 0.034536107 -0.49977198 0.6731396176  
## LogSqFtAbove -0.15802917 -0.072220661 -0.45817714 -0.6701241866  
## LogSqFtBasement 0.12310741 0.268776186 -0.05765410 -0.3092760071

summary(PCM)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 2.1233 1.3365 1.2337 0.96206 0.92751 0.82277  
## Proportion of Variance 0.3757 0.1489 0.1268 0.07713 0.07169 0.05641  
## Cumulative Proportion 0.3757 0.5245 0.6514 0.72852 0.80021 0.85662  
## PC7 PC8 PC9 PC10 PC11 PC12  
## Standard deviation 0.77006 0.60768 0.55902 0.48538 0.43877 0.13286  
## Proportion of Variance 0.04942 0.03077 0.02604 0.01963 0.01604 0.00147  
## Cumulative Proportion 0.90604 0.93681 0.96285 0.98249 0.99853 1.00000

plot(PCM, type="l")



Based on the summary of principal component analysis, we see the first component comprising 0.37 of cumulative variance, the second with 0.15, the third with 0.13, the fourth with 0.08, with a total of six principal components required to reach 0.80 of the overall variance and seven principal components to reach on 0.90 of the overall variance.

We will map eight of the principal components in the linear regression model.

traindata<-as.data.frame(predict(PCM,TrainDataSample2[c(4:5,8,11:12,15,18:19,22,25,26,27)]))  
model<-lm(TrainDataSample2$LogSalePrice ~ traindata$PC1 + traindata$PC2 + traindata$PC3 + traindata$PC4 + traindata$PC5 + traindata$PC6 + traindata$PC7+traindata$PC8)  
summary(model)

##   
## Call:  
## lm(formula = TrainDataSample2$LogSalePrice ~ traindata$PC1 +   
## traindata$PC2 + traindata$PC3 + traindata$PC4 + traindata$PC5 +   
## traindata$PC6 + traindata$PC7 + traindata$PC8)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.02055 -0.17405 0.00285 0.15963 1.14574   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.053479 0.003450 3783.631 <2e-16 \*\*\*  
## traindata$PC1 0.150321 0.001625 92.508 <2e-16 \*\*\*  
## traindata$PC2 0.156258 0.002582 60.528 <2e-16 \*\*\*  
## traindata$PC3 -0.098787 0.002797 -35.324 <2e-16 \*\*\*  
## traindata$PC4 0.201893 0.003586 56.295 <2e-16 \*\*\*  
## traindata$PC5 0.004823 0.003720 1.297 0.1948   
## traindata$PC6 -0.008830 0.004193 -2.106 0.0353 \*   
## traindata$PC7 0.119945 0.004481 26.770 <2e-16 \*\*\*  
## traindata$PC8 0.062209 0.005678 10.957 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2711 on 6166 degrees of freedom  
## Multiple R-squared: 0.7392, Adjusted R-squared: 0.7389   
## F-statistic: 2185 on 8 and 6166 DF, p-value: < 2.2e-16

Mapping eight of the principal components into a linear model, we have an adjusted R-Squared value of 0.74 showing that 74% of the variability of the Log Sale price is explained by the eight principal components in the model. However, with the p-value for principal component 5 equal to 0.886, we would fail to reject the null hypothesis that the coefficient for principal component five and above are all zero. We will model again, this time with only four principal components.

traindata2<-as.data.frame(predict(PCM,TestData[c(4:5,8,11:12,15,18:19,22,25,26,27)]))  
  
results<-predict(model,traindata, interval='predict')  
results<-cbind(results, TrainDataSample2$LogSalePrice)  
results<-as.data.frame(results)  
  
results$Price<-(TrainDataSample2$price)  
results$Predicted<-exp(results$fit)  
results$difference<-(results$Price-results$Predicted)  
results$percentage<-abs(results$difference/results$Price)  
MAPE1<-(100/length(results$Price))\*sum(results$percentage)  
MAPE1

## [1] 21.29856

model2<-lm(TrainDataSample2$LogSalePrice ~ traindata$PC1 + traindata$PC2 + traindata$PC3 + traindata$PC4)  
summary(model2)

##   
## Call:  
## lm(formula = TrainDataSample2$LogSalePrice ~ traindata$PC1 +   
## traindata$PC2 + traindata$PC3 + traindata$PC4)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.1226 -0.1820 -0.0049 0.1651 1.2770   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.053479 0.003677 3550.01 <2e-16 \*\*\*  
## traindata$PC1 0.150321 0.001732 86.80 <2e-16 \*\*\*  
## traindata$PC2 0.156258 0.002751 56.79 <2e-16 \*\*\*  
## traindata$PC3 -0.098787 0.002981 -33.14 <2e-16 \*\*\*  
## traindata$PC4 0.201893 0.003822 52.82 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2889 on 6170 degrees of freedom  
## Multiple R-squared: 0.7036, Adjusted R-squared: 0.7034   
## F-statistic: 3662 on 4 and 6170 DF, p-value: < 2.2e-16

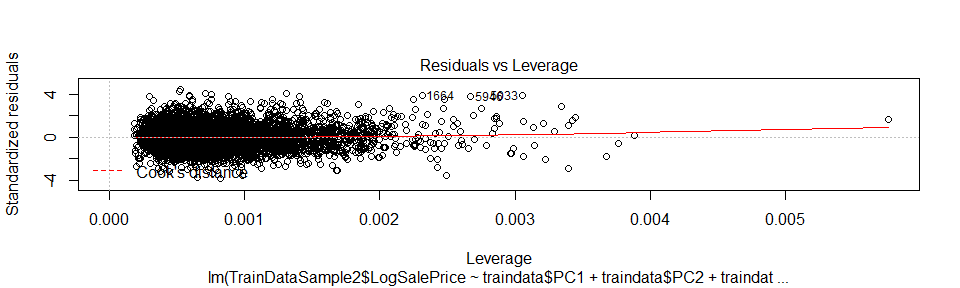
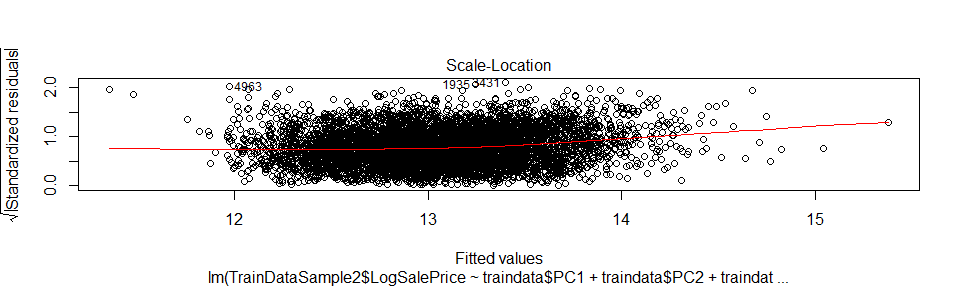
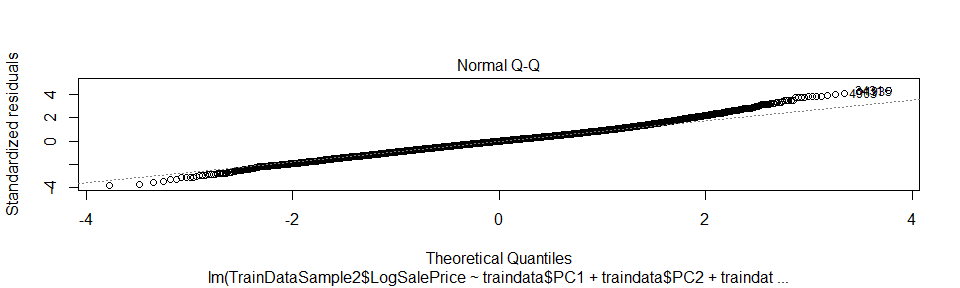
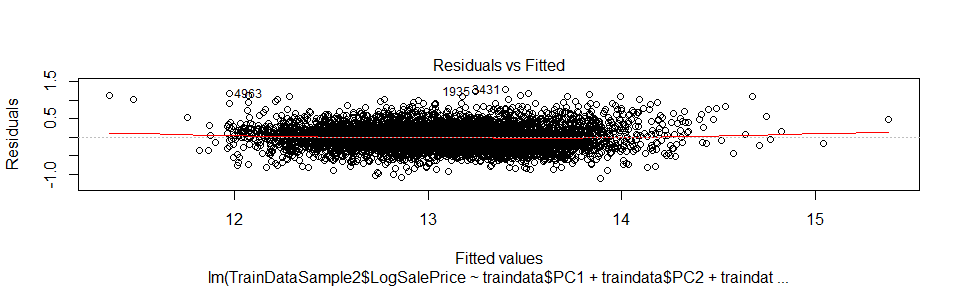
With only 4 principal components, our p-value for each of the coefficients is now <0.001, as well as for our intercept. The adjusted R-squared value still registers a good 0.71, meaning our model represented 71% of variance in the price of homes from the 4 principal components and the explanatory variables and data underlying them.

traindata<-as.data.frame(predict(PCM,TestData[c(4:5,8,11:12,15,18:19,22,25,26,27)]))  
  
results2<-predict(model2,TestData, interval='predict')  
results2<-as.data.frame(results2)  
  
results2$Price<-(TestData$price)  
results2$Predicted<-exp(results2$fit)  
results2$difference<-(results2$Price-results2$Predicted)  
results2$percentage<-abs(results2$difference/results2$Price)  
MAPE2<-(100/length(results2$Price))\*sum(results2$percentage)  
MAPE2

## [1] 22.67032

#sqrt(mean((results2$difference)^2))

within our 15,438 observations used to test the model, the model had increasing residuals for the range above $1MM of predicted value. Compared against the actual sale price, the MAPE is 21.5%.



The model shows mostly a good distribution of residuals, but an increasing variance from normality exhibited at the limits of the expected value. Another model using the square root of price as the response variable was also used, but did not offer any advantage in terms of residual fit versus the log transformation.

#extract a final equation  
Final<-PCM$rotation[,1:5]\*(model2$coefficients[2:5])  
rowSums(Final)

## bedrooms bathrooms floors condition   
## 0.096591993 0.036583751 0.004150154 0.244344311   
## grade yr\_built lat long   
## 0.071414542 -0.081970796 -0.040888741 0.082925406   
## LogLotSqFt LogSqFtLiving LogSqFtAbove LogSqFtBasement   
## 0.141922027 0.109399307 -0.083344466 -0.073220527

colMeans(HomeSales[c(4:8,11:15,18:19, 22,25:28)], na.rm=TRUE)

## bedrooms bathrooms sqft\_living sqft\_lot   
## 3.370842 2.114757 2079.899736 15106.967566   
## floors condition grade sqft\_above   
## 1.494309 3.409430 7.656873 1788.390691   
## sqft\_basement yr\_built lat long   
## 291.509045 1971.005136 47.560053 -122.213896   
## LogLotSqFt LogSqFtLiving LogSqFtAbove LogSqFtBasement   
## 8.989956 7.550335 7.394877 2.529293   
## SqRtSalesPrice   
## 706.295186

In explaining the final model, it's important to consider the process of principal component analysis as applied in this paper is using centered and scaled data and the coefficients in the model are expressing the price from that assumption.

Intercept: The intercept is 13.05, that when reverse log transforms equates to a home price of $466,308 if we assume that all other variable take the value of 0 (giving the average house in Seattle based on the criteria selected into the model.)

Bedrooms: For each standard deviation (0.65) above the mean number of bedrooms (3.4), the price increases by 9.2%.

Bathrooms: For each standard deviation (0.77) above the mean number of bathrooms (2.1), the price increases by 12.2%.

Floors: For each standard deviation (0.54) above the mean number of bathrooms (1.5), the price decreases by 0.3%.

Lot Square Footage: For each standard deviation (0.90) above the mean log lot square footage (9.0), the price decreases by 5.5%.

Living Square Footage: For each standard deviation (0.42) above the mean log living square footage (7.5), the price increases by 10.2%.

Above Grade Square Footage: For each standard deviation (0.43) above the mean log above grade living space (7.4), the price decreases by 0.1%.

Basement Square Footage: For each standard deviation (3.1) above the mean log basement square footage (2.5), the price decreases 32%.

Latitude: For each standard deviation (0.1) of latitude going North from the mean (47.56 degrees North), price decreases 0.1%.

Longitude: For each standard deviation (0.1) of longitude going West from the mean (122.21 Degrees West), price decreases 9.5%.

Year Built: For each standard deviation (29.4 years) of year built above the mean (1971), price increases by 5%.

Condition: For each standard deviation (0.65) of score above the mean (3.4), price increases by 13%.

Grade: For each standard deviation (1.2) of score above the mean (7.7), price increases by 6.5%.

#### Conclusions

Home prices in King County, WA follow a basic pattern, with a mean price of $466,308 and increases coming from Square Footage, number of bathrooms and number of bedrooms primarily.