Analysis of the Ames Housing Dataset

Palash Jain

11/30/2018

Executive Summary

The purpose of this analysis report was to analyse the Ames Housing dataset and use it to fit a linear regression model which predicts Sale prices of houses based on the features of the house. After thorough exploratory analysis, 6 out of 8 available features were chosen for fitting a predicitve linear regression model. From multiple models, the optimal model was chosen based on ANOVA and the simplicity of the model. Using linear regression, I was able to explain approximately 80% of the variation shown by the outcome variable using a simple and interpretable model using just 4 predictor variables. The key insights obtained from this analysis are as follows:

- LotArea & SqFt both drive the sale prices for the houses up.
- A positive correlation between YearBuilt and SalePrices suggests that newly built houses are sold for more value than older ones.
- The style of a house has a significant impact on SalePrice.
- The number of rooms/bedrooms/bathrooms do not have a significant impact on the house sale price given that everything else is fixed.

Exploratory Data Analysis

```
# Loading the dataset
data<-read.csv('ames_housing_data_PROJECT2018.csv')
dim(data)
```

```
## [1] 1598    9
```

```
str(data)
```

```
## 'data.frame':    1598 obs. of  9 variables:
##  $ LotArea  : int  31770 11622 14267 11160 13830 9978 7500 10000 7980 8402 ...
##  $ Style    : Factor w/ 2 levels "1Story","2Story": 1 1 1 1 2 2 2 2 1 2 ...
##  $ YearBuilt: int  1960 1961 1958 1968 1997 1998 1999 1993 1992 1998 ...
##  $ SqFt     : int  1656 896 1329 2110 1629 1604 1804 1655 1187 1465 ...
##  $ FullBath : int  1 1 1 2 2 2 2 2 2 2 ...
##  $ HalfBath : int  0 0 1 1 1 1 1 1 0 1 ...
##  $ Bedrooms : int  3 2 3 3 3 3 3 3 3 3 ...
##  $ Rooms    : int  7 5 6 8 6 7 7 7 6 7 ...
##  $ SalePrice: num  215000 105000 172000 244000 189900 ...
```

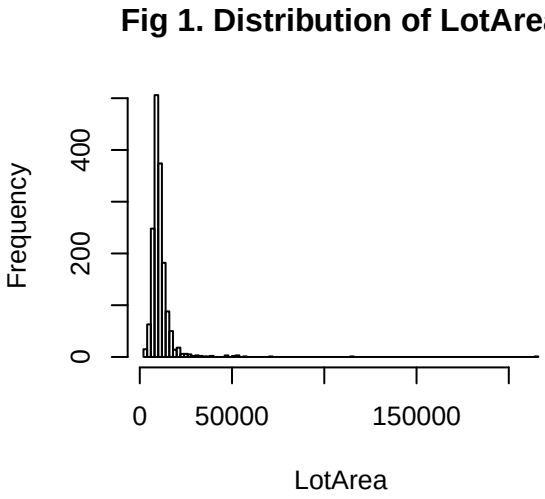
As seen above, the dataset contains information about 1598 houses across 9 features. Out of the 9 features, 8 are ordinal while 1 is categorical. These features describe the basic characteristics of houses that anyone looking to purchase a house will be interested in such as the lot area, square footage, number of rooms/bathrooms/bedrooms, the year the house was built, and the sale price of the house. The dataset has no missing values in any of the columns and thus we can move to further exploratory analysis.

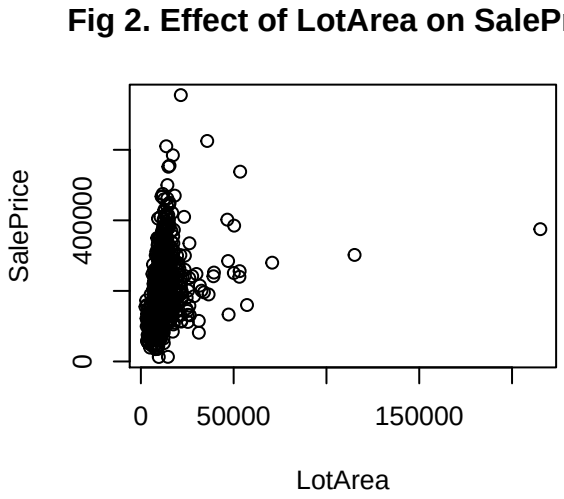
Lot Area

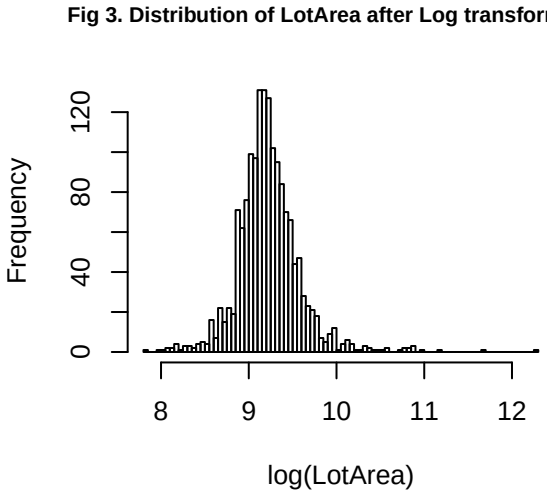
```
par(mfrow=c(2,2))
summary(data$LotArea)
```

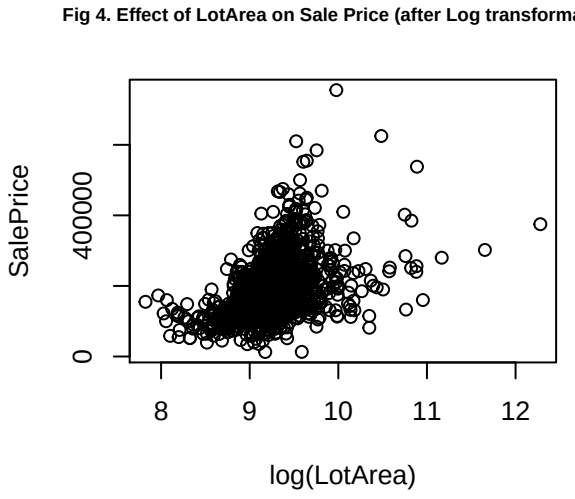
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2500    8372    9876   11047   11956  215245
```

```
hist(data$LotArea,breaks = 100, main = 'Fig 1. Distribution of LotArea',xlab = 'LotArea')
plot(SalePrice~LotArea,data = data,main = 'Fig 2. Effect of LotArea on SalePrice')
hist(log(data$LotArea),breaks = 100,
     main = 'Fig 3. Distribution of LotArea after Log transformation',
     xlab = 'log(LotArea)',cex.main = 0.85)
plot(SalePrice~log(LotArea),data = data,
     main = 'Fig 4. Effect of LotArea on Sale Price (after Log transformation)',
     cex.main = 0.75)
```









LotArea is a numeric variable that ranges between 2,500 and 215,245 units. As seen above, it has a distribution that is skewed to the left (Fig 1). This is because the majority of the houses are sold for low to moderate houses and there are only a few houses which are sold for extraordinary amounts of money. To deal with this

skewness we can apply a logarithmic transformation on the variable.

The logarithmic transformation eliminates the skewness of the distribution and hence this variable is now ready to be used in an ordinary least squares regression model (Fig 3). I also see a better positive correlation with SalePrice (Fig 2 & 4).

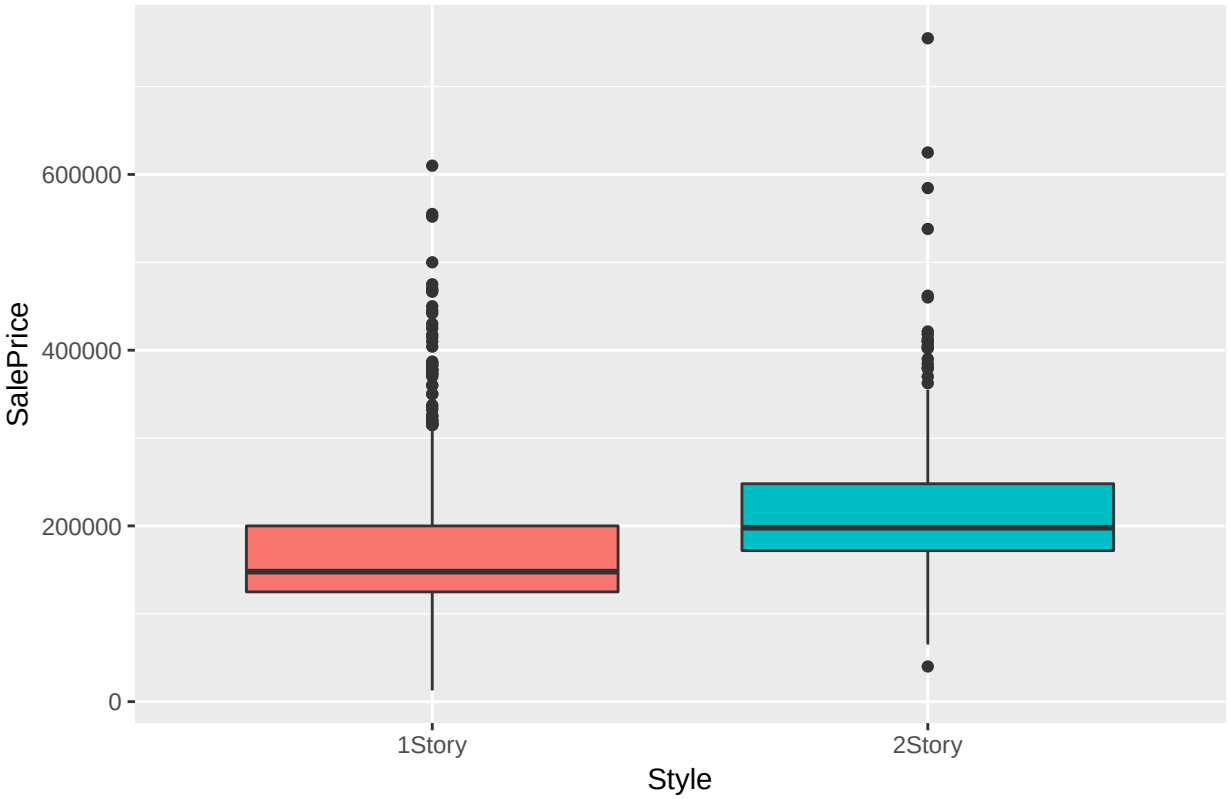
Style

```
table(data$Style)
```

```
##
## 1Story 2Story
##   1006    592
```

```
ggplot(data,aes(Style,SalePrice,fill = Style)) + geom_boxplot() +
  ggtitle('Fig 5. Distribution of SalePrice across different Styles of houses') +
  theme(legend.position = 'none',plot.title = element_text(size = 12))
```

Fig 5. Distribution of SalePrice across different Styles of houses



Style is a categorical variable with two levels describing the style of the house. As shown above there are 1006 one-story houses and 592 two-story houses. The median price for two-story houses is higher than that of one-story houses (Fig 5).

Year Built

```
yearly<-group_by(data,YearBuilt)%>%
  summarise(avg_sale_price = mean(SalePrice),count=n())
```

```
par(mfrow = c(2,1))
barplot(height = yearly$avg_sale_price,names.arg = yearly$YearBuilt,
        ylab = 'Average SalePrice',space = 0.3,
        main = 'Fig 6. Average sale prices aggregated according to YearBuilt')
barplot(height = yearly$count,names.arg = yearly$YearBuilt,
        ylab = 'Number of houses sold',space = 0.3,
        main = 'Fig 7. Number of houses sold summed for YearBuilt')
```

Fig 6. Average sale prices aggregated according to YearBuilt

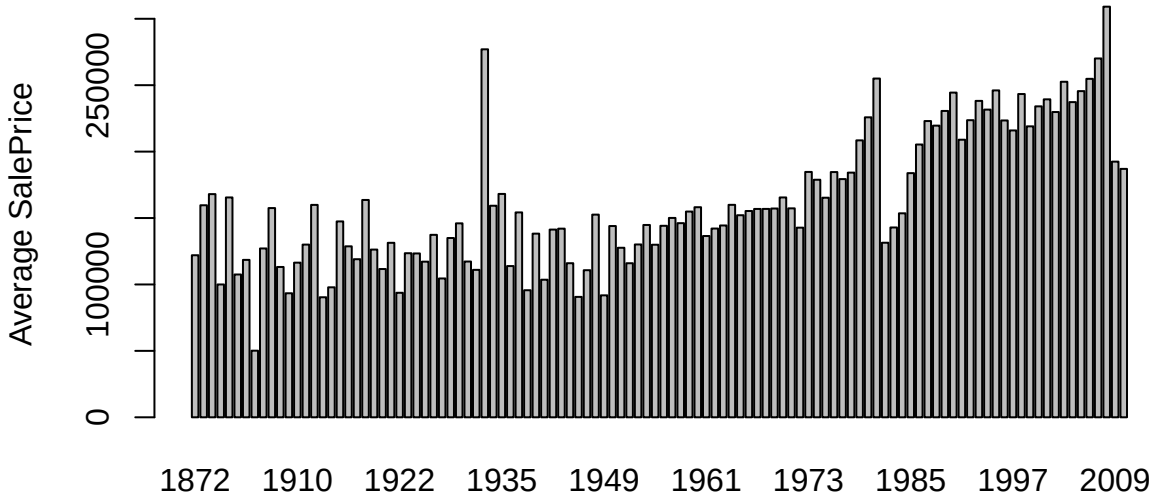
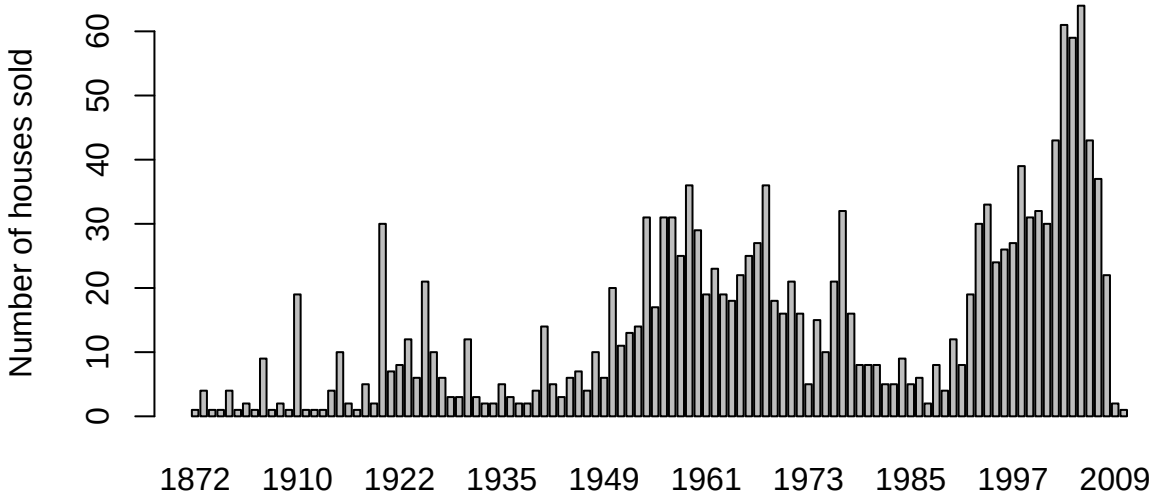


Fig 7. Number of houses sold summed for YearBuilt

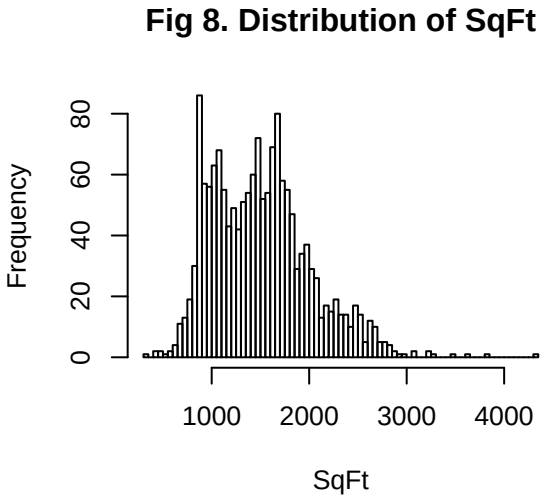


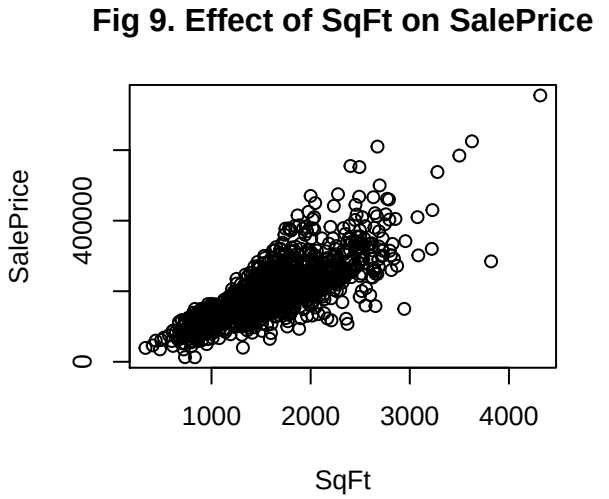
This numerical variable shows the year in which the houses were built. We can aggregate the data to show average sale price of a house for individual years. As seen above, the average house prices seem to increase

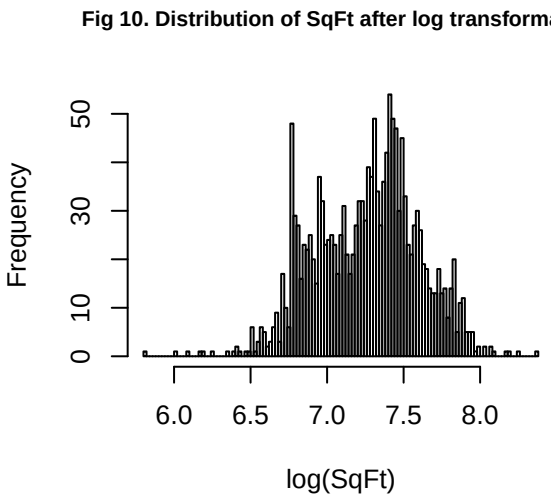
over the years until they hit a peak and there is a big dip in the prices. This dip is again followed by gradual increase until the next peak (Fig 6 & 7).

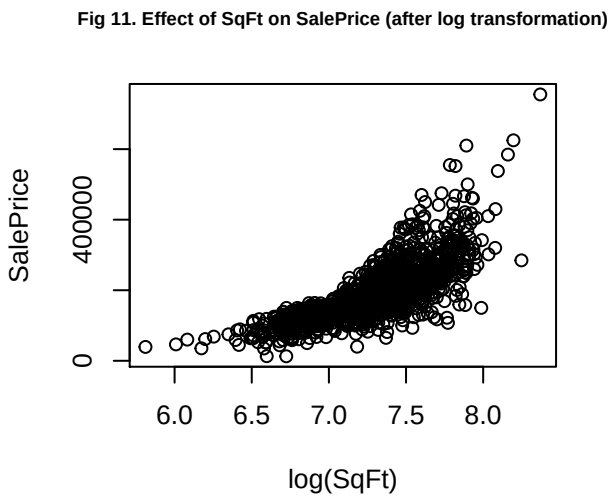
SqFt

```
par(mfrow=c(2,2))
hist(data$SqFt,breaks=100, main = 'Fig 8. Distribution of SqFt',xlab = 'SqFt')
plot(SalePrice~SqFt,data = data,main = 'Fig 9. Effect of SqFt on SalePrice')
hist(log(data$SqFt),breaks=100,xlab = 'log(SqFt)',cex.main=0.85,
     main = 'Fig 10. Distribution of SqFt after log transformation')
plot(SalePrice~log(SqFt),data = data,cex.main = 0.75,
     main = 'Fig 11. Effect of SqFt on SalePrice (after log transformation)')
```









This numerical variable shows the area of the property in square feet. In theory, the more the area the higher the sale price should be. As seen above, while the distribution is slightly skewed to the left (Fig 8 & 10), the skewness is not that pronounced and hence a log transformation is not needed. We also see that there seems to be a positive correlation between SqFt and SalePrice variables (Fig 9 & 11) .

FullBath

This variable represents the number of full bathrooms within the property. There seems to be a clear correlation between sale prices and number of full bathrooms for a house if the house has at least one full bathroom. However, the median sale price houses with zero full bathrooms is almost equal to that for houses with three bathrooms (Fig 12).

```
p1<-ggplot(data,aes(factor(FullBath),SalePrice,fill = factor(FullBath))) + geom_boxplot()+
  ggtitle('Fig 12. Effect of number of Full Bathrooms in a house on SalePrice') +
  theme(legend.position = 'none',plot.title = element_text(size = 12)) + xlab('FullBath')
p2<-ggplot(data,aes(factor(HalfBath),SalePrice,fill = factor(HalfBath))) + geom_boxplot()+
  ggtitle('Fig 13. Effect of number of half Bathrooms in a house on SalePrice')+
  theme(legend.position = 'none',plot.title = element_text(size = 12)) + xlab('HalfBath')
grid.arrange(p1,p2)
```

Fig 12. Effect of number of Full Bathrooms in a house on SalePrice

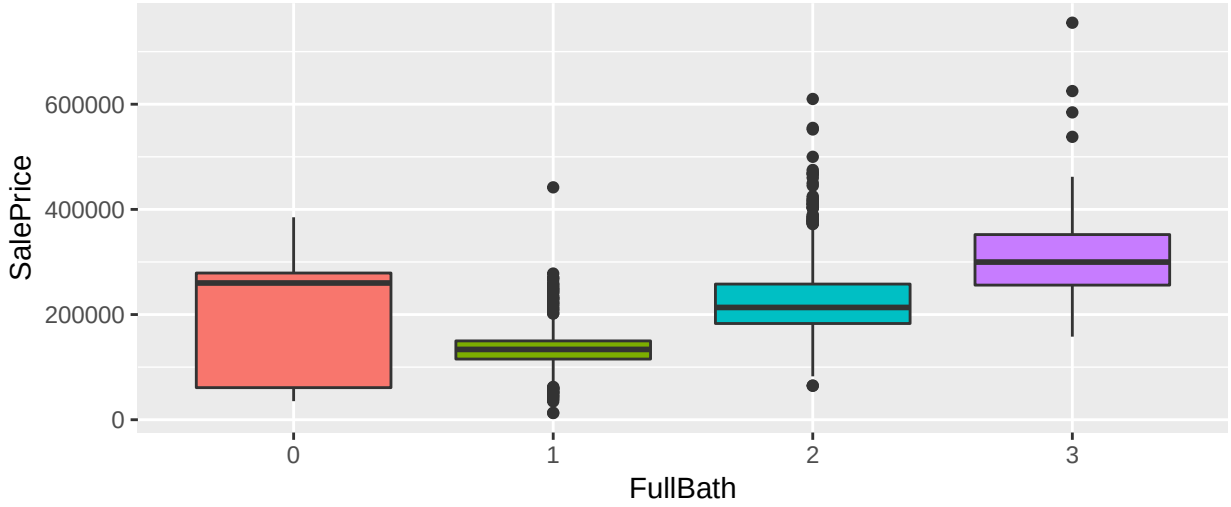
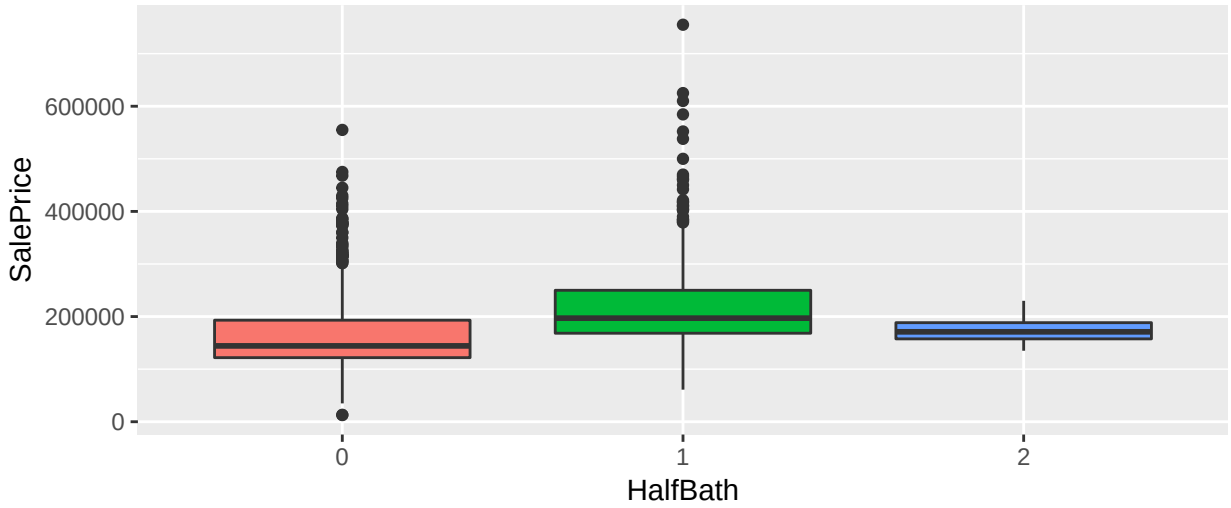


Fig 13. Effect of number of half Bathrooms in a house on SalePrice



HalfBath

HalfBath describes the number of half bathrooms within the property. There doesn't seem to be significant variation of Sale price as the number of half bathrooms within the property vary (Fig 13).

Bedrooms

This is a variable showing the number of Bedrooms within the property. There is not a lot of variation in the median sale prices as the number of bedrooms change. However, the median sale price for houses with zero bedrooms again is high compared to the rest (Fig 14).

```
p3<-ggplot(data,aes(factor(Bedrooms),SalePrice,fill = factor(Bedrooms))) + geom_boxplot() +
  ggtitle('Fig 14. Effect of number of Bedrooms in a house to SalePrice') +
  theme(legend.position  = 'none',plot.title = element_text(size = 12)) + xlab('Number of Bedrooms')
p4<-ggplot(data,aes(factor(Rooms),SalePrice,fill = factor(Rooms))) + geom_boxplot() +
  ggtitle('Fig. 15 Effect of number of Rooms in a house to SalePrice') +
  theme(legend.position = 'none',plot.title = element_text(size = 12)) + xlab('Number of Rooms')
grid.arrange(p3,p4)
```

Fig 14. Effect of number of Bedrooms in a house to SalePrice

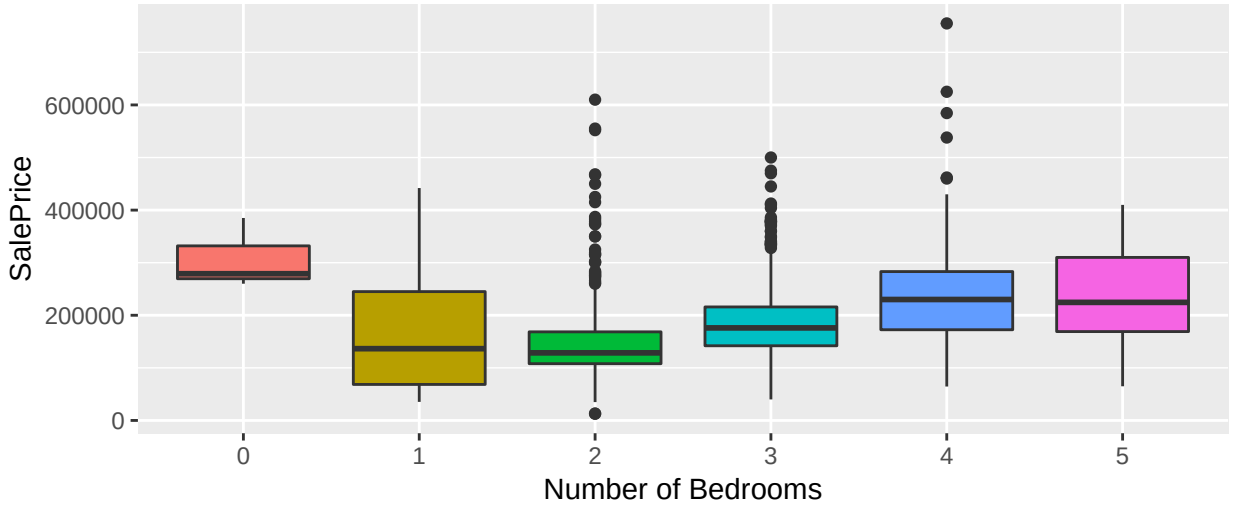
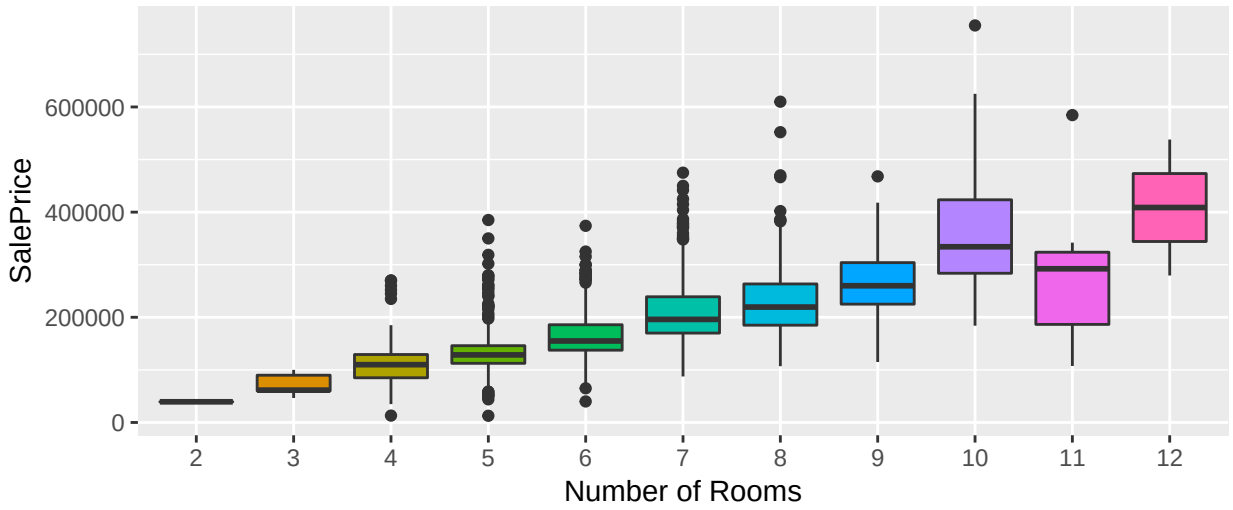


Fig. 15 Effect of number of Rooms in a house to SalePrice



Rooms

This variable shows the total number of rooms in a house. There seems to be a clear positive correlation between Sale Price and the total number of rooms in a property (Fig 14).

Exploratory Data Analysis Summary

From the analysis above, the following variables seem predicitve of SalePrice :

- LotArea (Log-transformed)
- Style
- YearBuilt
- SqFt
- FullBath
- Rooms

In the next section we will build simple ordinary least squares linear regression models to predict the SalePrice of a house based on these features.

Model fitting

```
fit <- lm(SalePrice~LotArea,data = data)
fit1 <- lm(SalePrice~LotArea+Style,data=data)
fit2 <- lm(SalePrice ~ LotArea + Style + YearBuilt, data = data)
fit3 <- lm(SalePrice ~ LotArea + Style + YearBuilt + SqFt, data = data)
fit4 <- lm(SalePrice ~ LotArea + Style + YearBuilt + SqFt + FullBath, data = data)
fit5 <- lm(SalePrice ~ LotArea + Style + YearBuilt + SqFt + FullBath + Rooms, data = data)
fit6 <- lm(SalePrice ~ LotArea + Style + YearBuilt + SqFt + Rooms, data = data)
calls <- data.frame(call=as.character(c(fit$call,fit1$call,fit2$call,fit3$call,fit4$call,fit5$call,fit6$
                    r_squared = round(c(summary(fit)$r.squared,summary(fit1)$r.squared,summary(fit2)$r.s
                                        summary(fit3)$r.squared,summary(fit4)$r.squared,summary(fit5)$r.squared
                                        summary(fit6)$r.squared) ,3))
calls$call<-gsub('lm\\(formula \\=','',calls$call)
calls$call<-gsub(', data = data)','',calls$call)
kable(calls)
```

call	r_squared
SalePrice ~ LotArea	0.085
SalePrice ~ LotArea + Style	0.167
SalePrice ~ LotArea + Style + YearBuilt	0.452
SalePrice ~ LotArea + Style + YearBuilt + SqFt	0.781
SalePrice ~ LotArea + Style + YearBuilt + SqFt + FullBath	0.781
SalePrice ~ LotArea + Style + YearBuilt + SqFt + FullBath + Rooms	0.782
SalePrice ~ LotArea + Style + YearBuilt + SqFt + Rooms	0.782

As seen above we are able to explain almost 80% of the variation in Sale Price using just 4 variables (fit3). Furthermore, addition of any new variables to that model doesn't significantly improve our r-squared value. In the next section we will evaluate which model is optimal using ANOVA.

Model selection using ANOVA

```
anova(fit,fit1,fit2,fit3,fit4,fit5)
```

```
## Analysis of Variance Table
##
## Model 1: SalePrice ~ LotArea
## Model 2: SalePrice ~ LotArea + Style
## Model 3: SalePrice ~ LotArea + Style + YearBuilt
## Model 4: SalePrice ~ LotArea + Style + YearBuilt + SqFt
## Model 5: SalePrice ~ LotArea + Style + YearBuilt + SqFt + FullBath
## Model 6: SalePrice ~ LotArea + Style + YearBuilt + SqFt + FullBath + Rooms
##   Res.Df           RSS Df     Sum of Sq         F               Pr(>F)
## 1   1596 8751462715991
## 2   1595 7960209774581  1  791252941410  603.6032 < 0.0000000000000002 ***
## 3   1594 5240844475136  1 2719365299445 2074.4539 < 0.0000000000000002 ***
```

```
## 4   1593 2095700089183  1 3145144385954 2399.2572 < 0.0000000000000002 ***
## 5   1592 2093714404744  1    1985684439    1.5148              0.21859
## 6   1591 2085614086725  1    8100318019    6.1793              0.01303 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit,fit1,fit2,fit3,fit6)
```

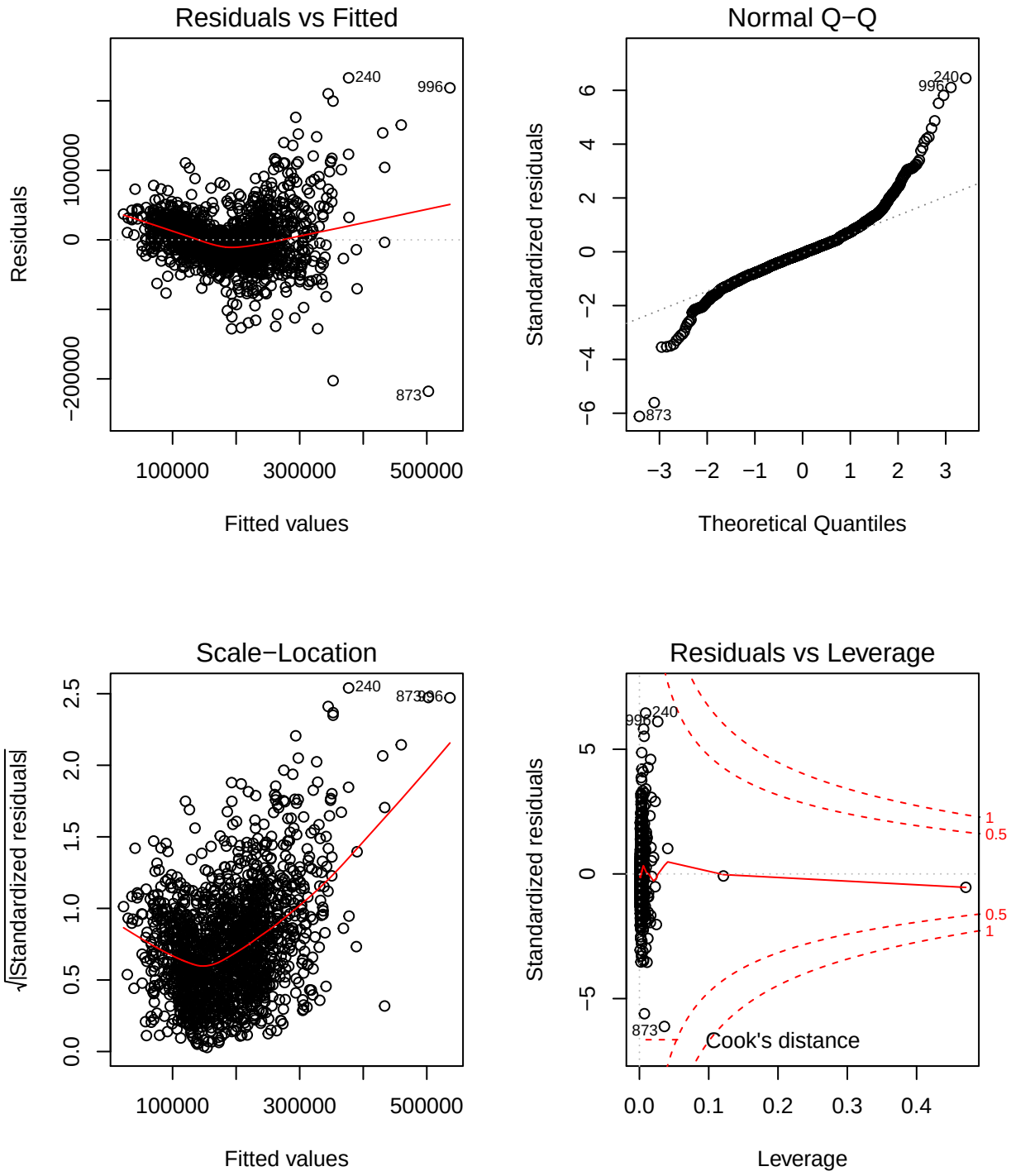
```
## Analysis of Variance Table
##
## Model 1: SalePrice ~ LotArea
## Model 2: SalePrice ~ LotArea + Style
## Model 3: SalePrice ~ LotArea + Style + YearBuilt
## Model 4: SalePrice ~ LotArea + Style + YearBuilt + SqFt
## Model 5: SalePrice ~ LotArea + Style + YearBuilt + SqFt + Rooms
##   Res.Df           RSS Df     Sum of Sq         F               Pr(>F)
## 1   1596 8751462715991
## 2   1595 7960209774581  1  791252941410  603.4764 < 0.0000000000000002 ***
## 3   1594 5240844475136  1 2719365299445 2074.0181 < 0.0000000000000002 ***
## 4   1593 2095700089183  1 3145144385954 2398.7532 < 0.0000000000000002 ***
## 5   1592 2087363490159  1    8336599023    6.3582              0.01178 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As seen above, the addition of the variable 'FullBath' to the model doesn't signigficantly improve the model. On the other hand, the addition of the variable 'Rooms' does seem to improve the model in a statistically significant manner. However, As seen from the table above, adding 'Rooms' to the model doesn't improve the amount of variation explained for the SalePrice. Since we want to understand the most variation in our outcome variable using the simplest model possible, the model 'fit3' is our best choice.

```
summary(fit3)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + Style + YearBuilt + SqFt,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -217823  -19317   -2233   15077  232983
##
## Coefficients:
##                  Estimate    Std. Error t value             Pr(>|t|)
## (Intercept) -1595644.1365    67169.4024 -23.756 < 0.0000000000000002 ***
## LotArea            0.6261        0.1245   5.028          0.000000551 ***
## Style2Story   -40718.4401     2434.0009 -16.729 < 0.0000000000000002 ***
## YearBuilt        811.6619       34.6652  23.414 < 0.0000000000000002 ***
## SqFt             125.3113        2.5629  48.895 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36270 on 1593 degrees of freedom
## Multiple R-squared:  0.7808, Adjusted R-squared:  0.7802
## F-statistic:  1418 on 4 and 1593 DF,  p-value: < 0.00000000000000022
```

```
par(mfrow=c(2,2))
plot(fit3)
```



After running diagnostics on the linear regression model fit3, I recognized the following :

- Residuals v/s fitted - The residuals are equally spread around a horizontal line. This means that there is no non-linear relationship between the predictor and outcome variables.
- Normal Q-Q - The residuals follow the dashed straight line for the most part and only deviate from it at extremely high or low values. This means that our residuals can be approximated to have a normal

distribution.

- Scale Location - At lower values the residuals are spread evenly along the predictor ranges but at high values we see that the spread is higher towards the x-axis.
- Residuals v/s Leverage - We see that there is only one outlier (extreme right of the plot) that influences our model heavily as it lies far beyond the Cook's distance lines. This case should revisited and evaluated.

Model Fitting Summary

The linear regression model 'fit3' gives the perfect balance between predictive power and simplicity while also performing well on diagnostics. Thus this should be our model of choice.