Kumar - SDS 291 Homework 3

HOMEWORK 3: CONDITIONS, CONFOUNDING, AND TRANSFORMATIONS

```
library(Stat2Data)
library(performance)
library(ggplot2)
library(equatiomatic)
library(gridExtra)
```

Load libraries

EXERCISE 1

you're concerned that their data do not meet the necessary conditions for inference with a linear regression model

In order to fit any linear model, all models have to pass the LINE test:

- Linearity test –> plot fitted values for each observation against the residuals of each observation there should be no relationship between the two variables.
- **Independence test** –> regarding the data collection, the points should not be connected – this is violated when the same unit is measured over time.
- Normality test –> a histogram of the residuals should be normal.
- Equal Variance test –> the standard deviation of the errors should be the same regardless of if the model is predicting large or small values

The way data was collected for this model violates the Independence Test.

EXERCISE 2

The first thing this analyst wants to investigate is the hypothesis that a home's sale price is linearly related to its square footage.

```
data("GrinnellHouses")
head(GrinnellHouses)
```

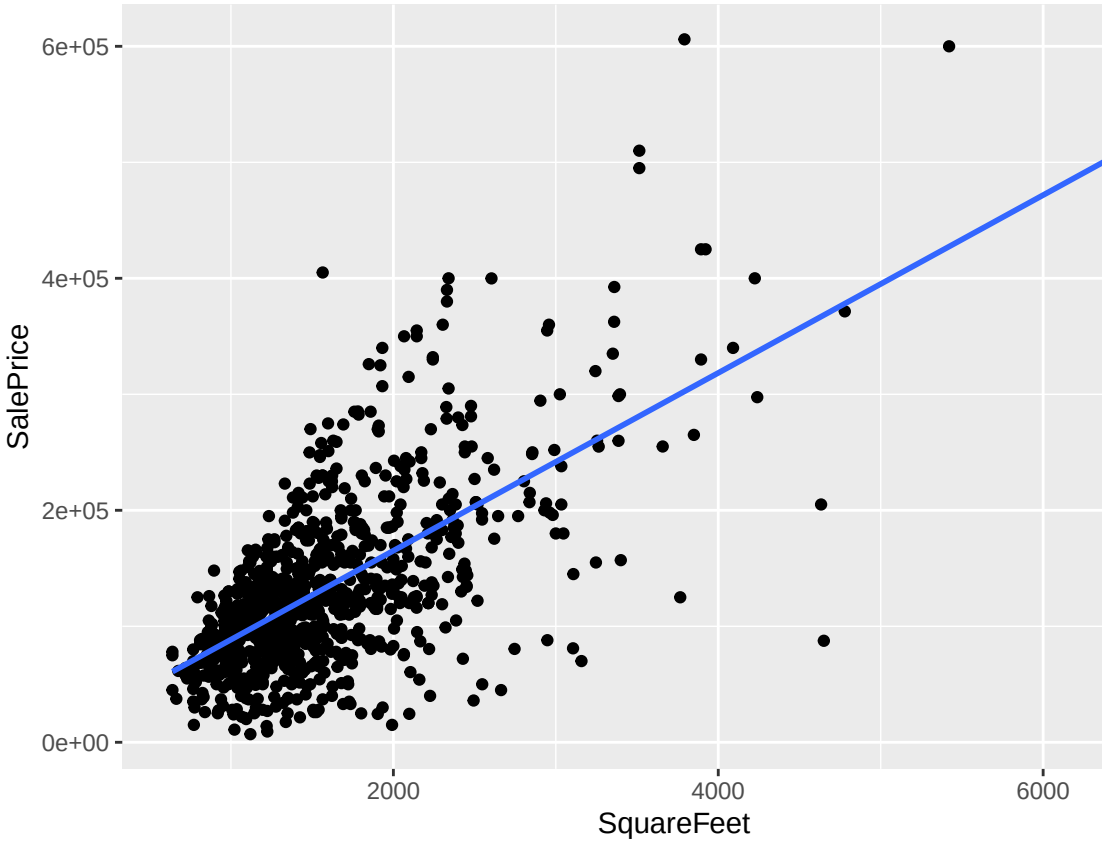
Load data

```
##    Date                 Address Bedrooms Baths SquareFeet   LotSize YearBuilt
## 1 16695    1510 First Ave #112         2     1       1120        NA      1993
## 2 16880         1020 Center St         3     1       1224 0.1721763      1900
## 3 16875      918 Chatterton St         4     1       1540        NA      1970
## 4 16833 1023 & 1025 Spring St.         3     1       1154        NA      1900
## 5 16667            503 2nd Ave         3     1       1277 0.2066116      1900
## 6 16583           9090 Clay St         3     1       1079 0.1993572      1900
##   YearSold MonthSold DaySold CostPerSqFt OrigPrice ListPrice SalePrice SPLPPct
## 1     2005         9      16        6.25     17000     10500      7000   66.67
## 2     2006         3      20       22.06     35000     35000     27000   77.14
## 3     2006         3      15       18.18     54000     47000     28000   59.57
## 4     2006         2       1       26.00     65000     49000     30000   61.22
## 5     2005         8      19       24.08     35000     35000     30750   87.86
## 6     2005         5      27       38.92     45900     45900     42000   91.50
```

EXERCISE 2A:

Fit a model that is suitable for the analyst's needs. Then, summarize this model's goodness of fit using the R^2 statistic, and interpret its value for the analyst.

```
ggplot(data = GrinnellHouses, mapping = aes(y = SalePrice,
                                            x = SquareFeet)) +
  geom_point() +
  geom_smooth(method = lm, formula = y ~ x, se = F)
```



Visualize data + model

```
price_footage_model = lm(SalePrice ~ SquareFeet, data = GrinnellHouses)
summary(price_footage_model)$coefficients
```

Generate model

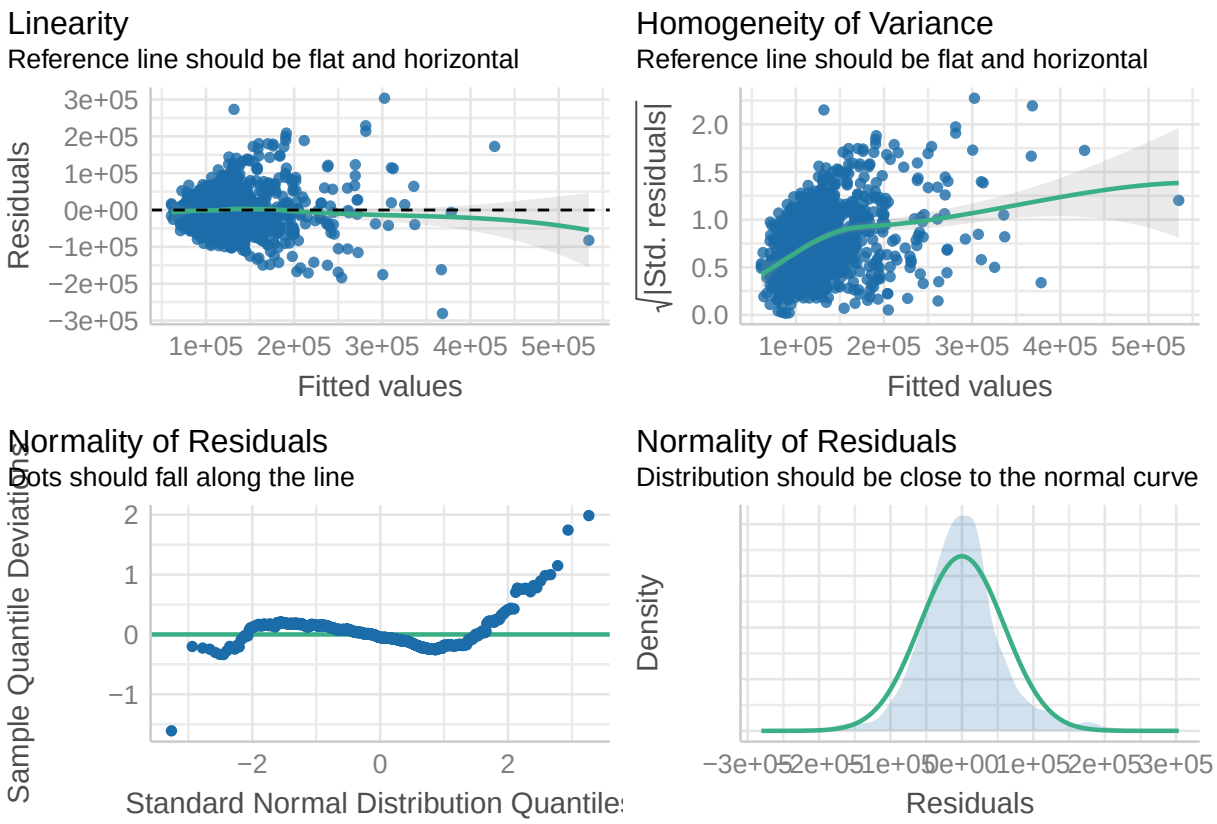
```
##                Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) 11836.23402 5116.455090  2.313366  2.092488e-02
## SquareFeet     76.65055    2.986012 25.669875 9.996874e-110
```

Interpreting model This model's R^2 value is 0.4203, meaning 42.03% of the variance in the data can be explained by the model. Because the model explains so little of the variance (low R^2), this model (and thus square footage) do not explain Sale Price very well.

EXERCISE 2B:

Investigate whether the assumptions that would support using a t-test on the slope coefficient of this model are reasonable to make in this case, and advise the analyst on whether they should consider the t-tests they see in the regression table as reliable.

```
check_model(price_footage_model, check = c("linearity", "homogeneity",
                                           "qq", "normality"))
```



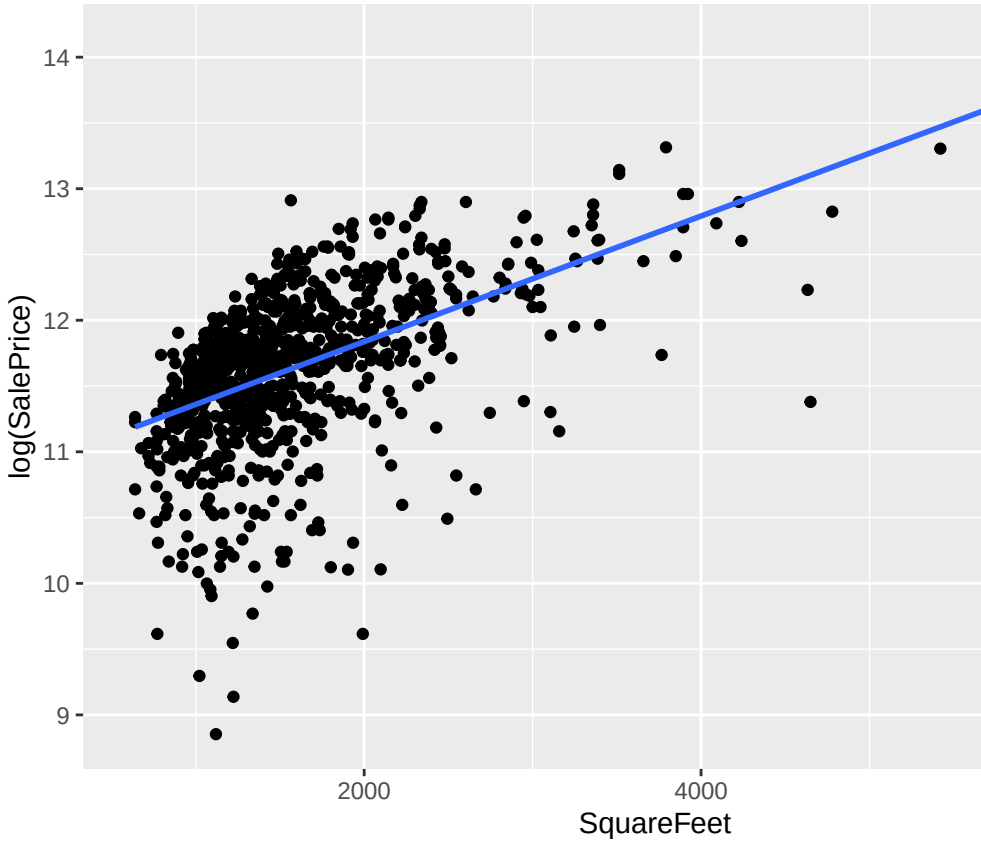
Assess fit

From these checks, we can see there are only mild violations of the assumptions necessary to make t-tests. When plotted against the fitted values, the residuals have a mostly flat+horizontal spread. While the homogeneity of variance is not entirely constant, the other tests are largely unviolated. The normality of residuals is almost exact, with exceptions at the high end of the bottom left plot. Thus, we can trust the t-tests and view them as reliable.

EXERCISE 3

How does a logarithmic transformation of the outcome variable affect your conclusions about whether the model's assumptions are reasonably met?

```
ggplot(data = GrinnellHouses, mapping = aes(x = SquareFeet, y = log(SalePrice))) +
  geom_point() +
  geom_smooth(method = lm, formula = y ~ x, se = F)
```



Visualize the transformed model

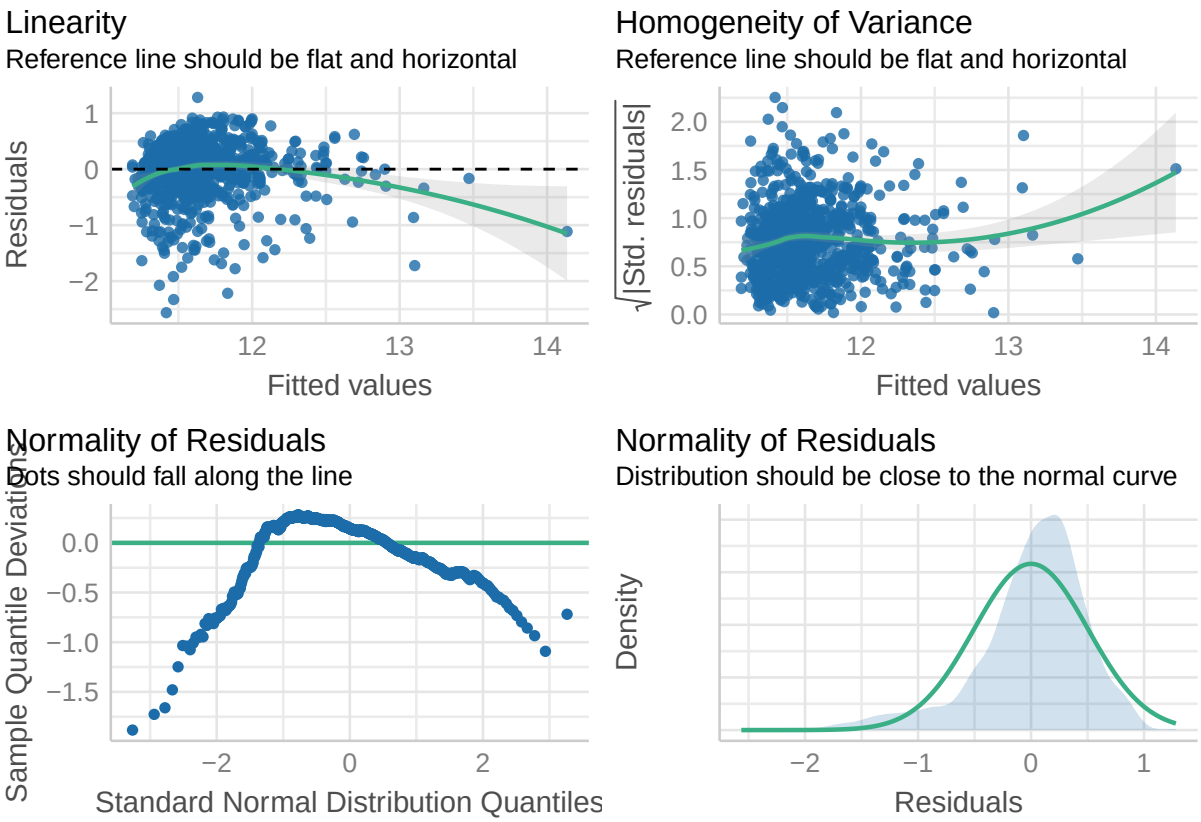
```
price_logft_model = lm(log(SalePrice) ~ SquareFeet, data = GrinnellHouses)
summary(price_logft_model)
```

Generate model

```
##
## Call:
## lm(formula = log(SalePrice) ~ SquareFeet, data = GrinnellHouses)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.56394 -0.21188  0.07232  0.31144  1.28167
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.088e+01  4.378e-02  248.60   <2e-16 ***
## SquareFeet  4.772e-04  2.555e-05   18.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5053 on 909 degrees of freedom
##   (18 observations deleted due to missingness)
## Multiple R-squared:  0.2774, Adjusted R-squared:  0.2766
```

```
## F-statistic: 348.9 on 1 and 909 DF,  p-value: < 2.2e-16
```

```
check_model(price_logft_model, check = c("linearity", "homogeneity",
                                         "qq", "normality"))
```



Assess fit

Upon completing a log transformation of the outcome variable (SalePrice), the assumptions necessary to run t-tests are strongly violated. While the homogeneity of variance has improved, the linearity has gotten slightly worse, and the residuals are nowhere near normal. Thus, we cannot trust the t-tests generated by this model.

Interpret what the slope coefficient of this model tells you about the relationship between home price (measured in dollars) and size (measured in square feet).

The slope coefficient of this model is 4.772e-04, which implies that for each increase in house size by one square foot, the price of the house increases by $0.0004772.

EXERCISE 4

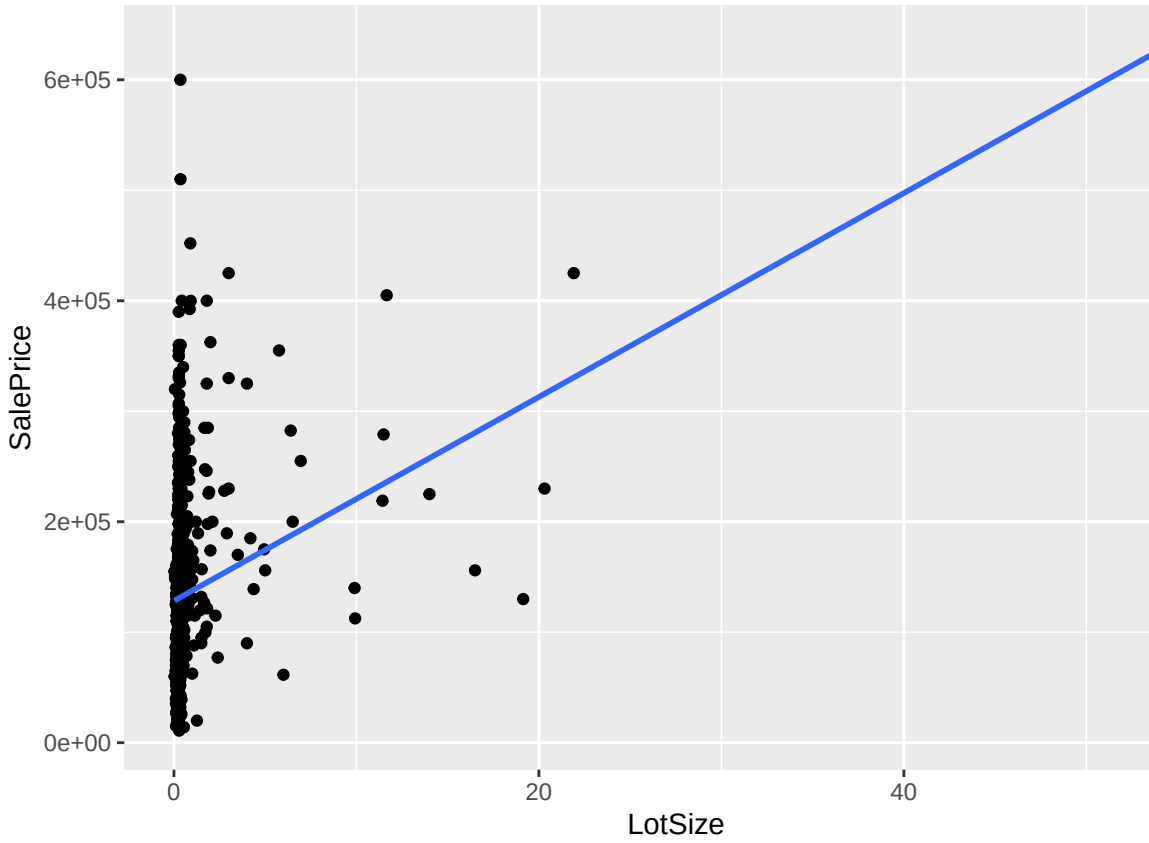
Fit a model that accounts for confounding variables: the size of the lot and the year the home was built

I will begin this exercise by visualizing the relationships of these possibly confounding variables with our outcome data.

Visualize data

```
#### NO CURVE, LOG IS BEST

ggplot(data = GrinnellHouses, mapping = aes(x = LotSize,
                                            y = SalePrice)) +
  geom_point() +
  geom_smooth(method = lm, formula = y ~ x, se = F)
```

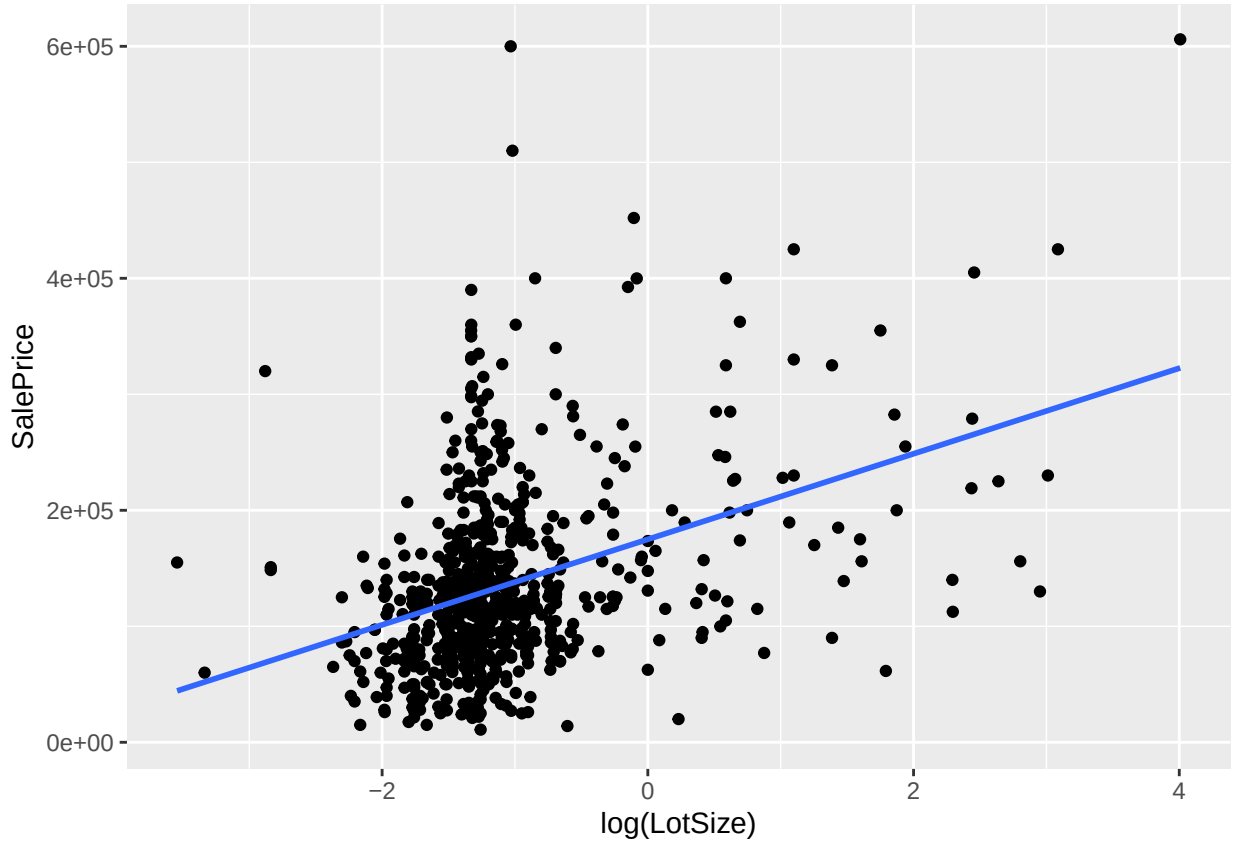


Size of lot and Price

This model would be easier to interpret and view if we logged the explanatory variable:

```
size_price_vis = ggplot(data = GrinnellHouses, mapping = aes(x = log(LotSize),
                                                             y = SalePrice)) +
  geom_point() +
  geom_smooth(method = lm, formula = y ~ x, se = F)

size_price_vis
```



We also know this logged model fits the data better than the untransformed data:

```
lotsize_model = lm(SalePrice ~ LotSize, data = GrinnellHouses)
summary(lotsize_model)$r.squared
```

```
## [1] 0.1005096
```

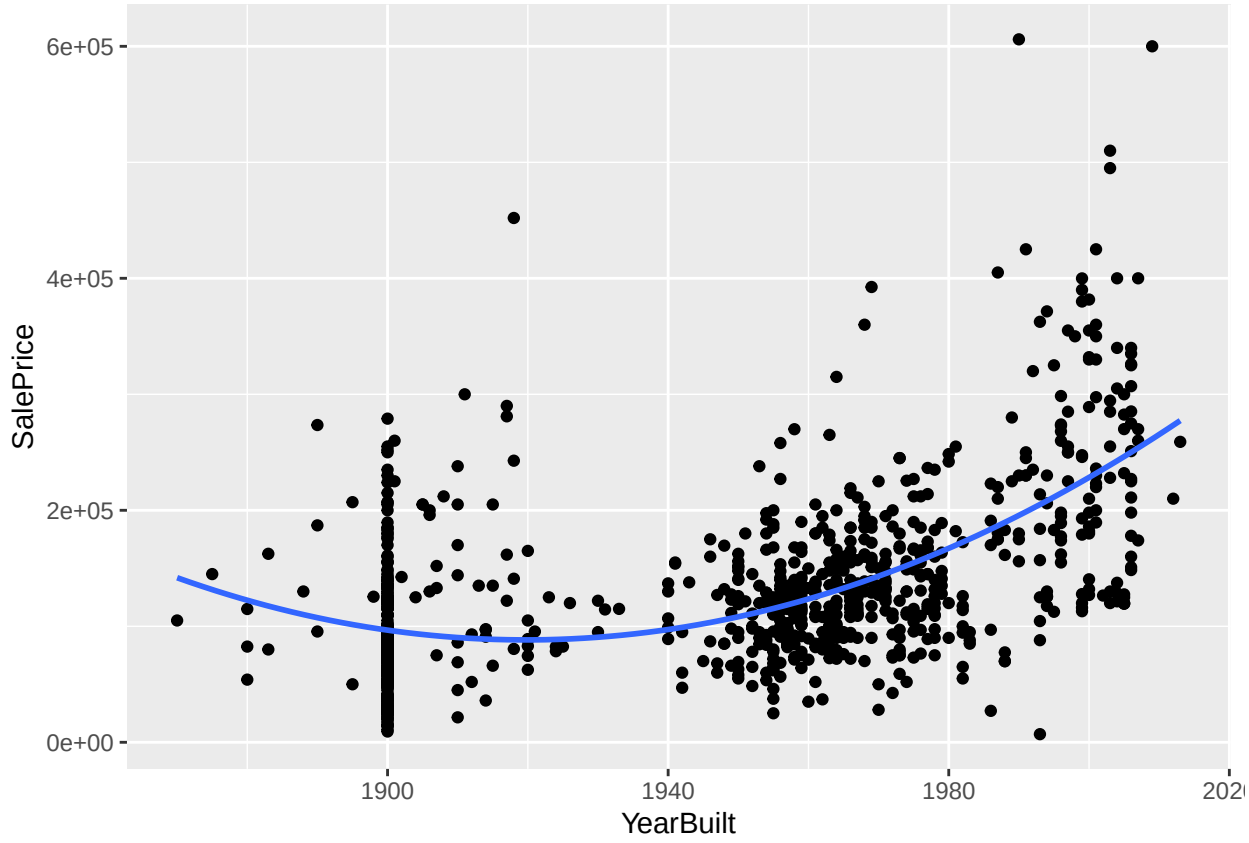
```
log_model = lm(SalePrice ~ log(LotSize), data = GrinnellHouses)
summary(log_model)$r.squared
```

```
## [1] 0.146651
```

Year home was built and Price This data has a clear curve, so we generate a quadratic model

```
year_price_vis = ggplot(data = GrinnellHouses, mapping = aes(x = YearBuilt,
                                                             y = SalePrice)) +
  geom_point() +
  geom_smooth(method = lm, formula = y ~ x + I(x^2), se = F)

year_price_vis
```



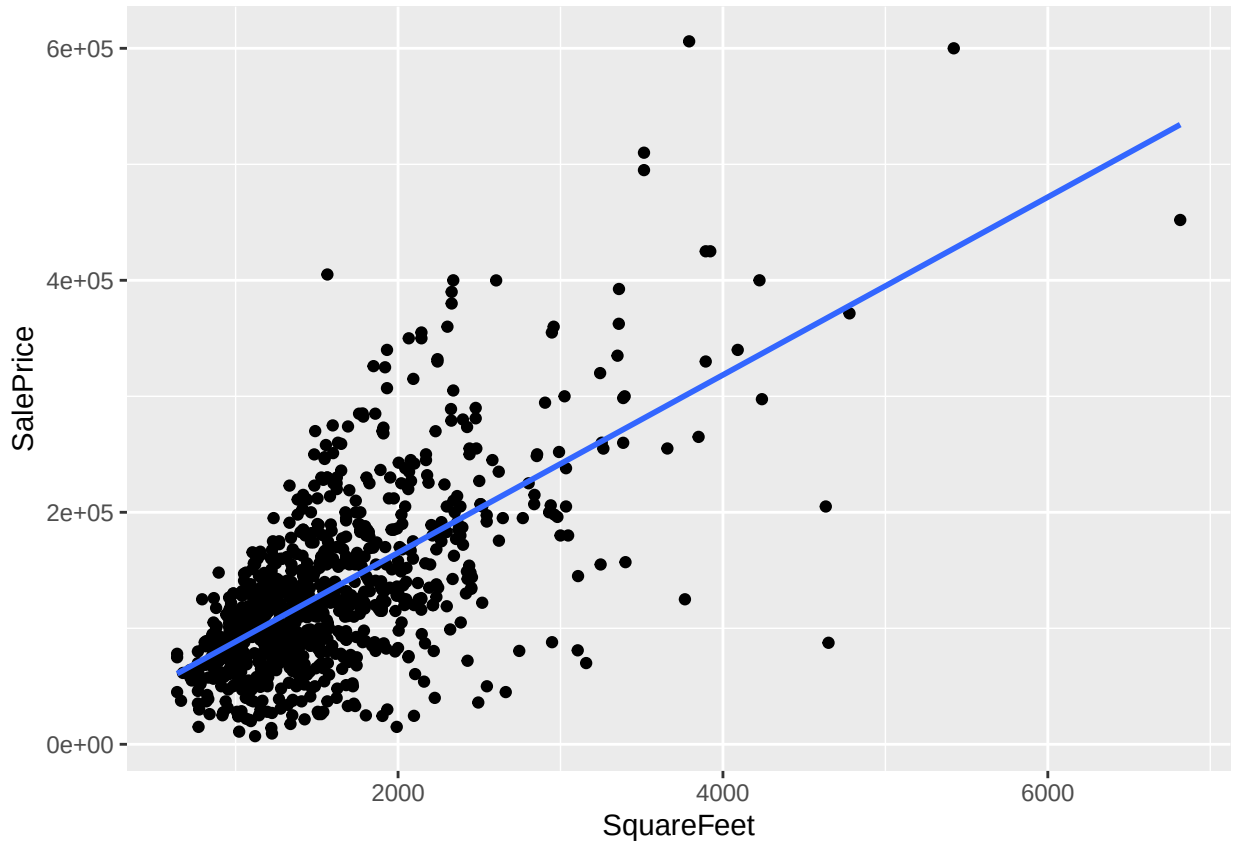

```
quad_model = lm(SalePrice ~ YearBuilt + I(YearBuilt^2), data = GrinnellHouses)
summary(quad_model)$r.squared
```

```
## [1] 0.3380766
```

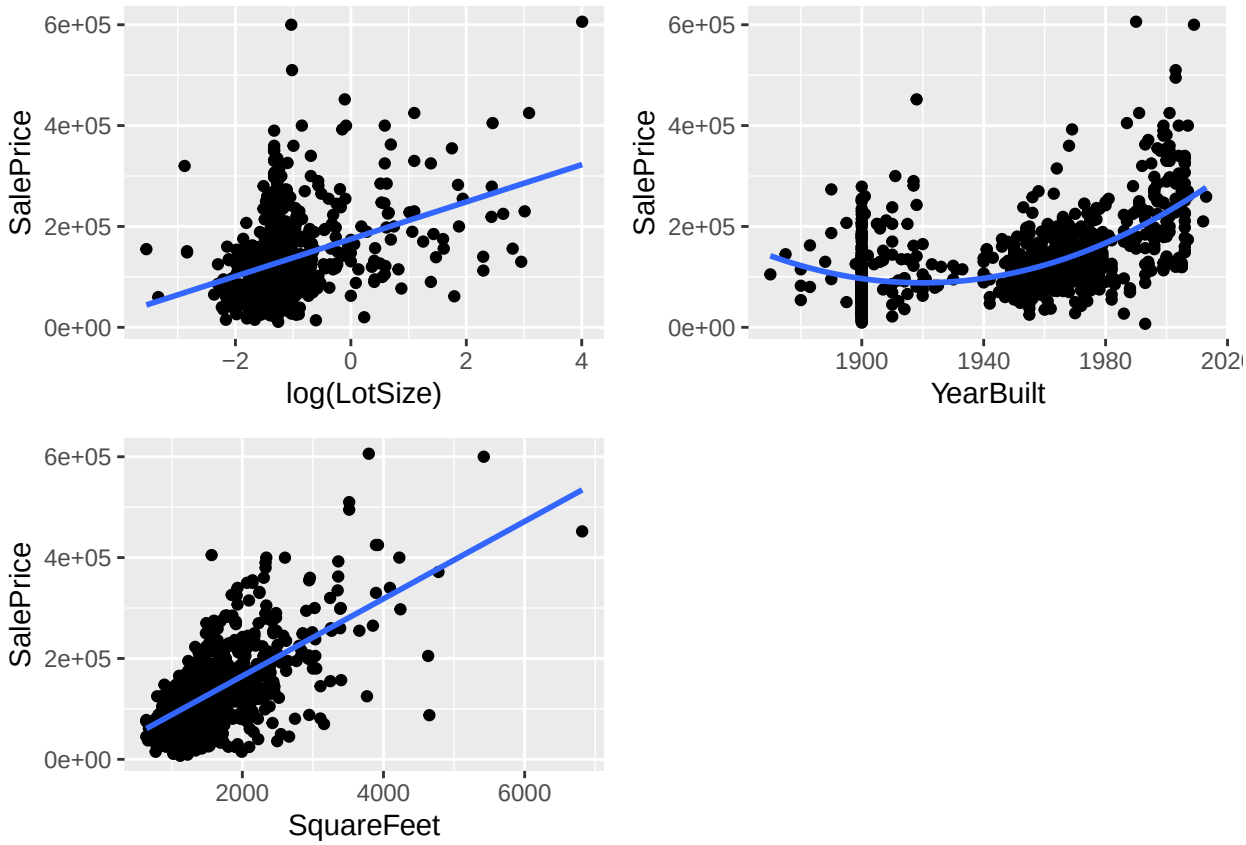
Square footage and Price These two have a linear relationship which we explored earlier, best left untransformed

```
footage_price_vis = ggplot(data = GrinnellHouses, mapping = aes(x = SquareFeet,
                                                                y = SalePrice)) +
  geom_point() +
  geom_smooth(method = lm, formula = y ~ x, se = F)

footage_price_vis
```




```
grid.arrange(size_price_vis, year_price_vis, footage_price_vis, nrow = 2)
```



Visualize

Generate new model If we combine these three explanatory variables, we can generate a possible new model

```
compound_model = lm(SalePrice ~ SquareFeet + log(LotSize) + YearBuilt + I(YearBuilt^2), data = Grinnell

summary(compound_model)$adj.r.squared
```

```
## [1] 0.690093
```

However, we should check to ensure that, when these three variables are combined, our transformations are still necessary:

Without log This version of the model has a slightly higher adjusted R^2 value

```
compound_model = lm(SalePrice ~ SquareFeet + LotSize + YearBuilt + I(YearBuilt^2), data = GrinnellHouse

summary(compound_model)$coefficients
```

```
##                    Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept)     2.799623e+07 6.009690e+06  4.658514  3.787526e-06
## SquareFeet      6.816152e+01 2.571621e+00 26.505270 7.581415e-109
## LotSize         4.721145e+03 6.010836e+02  7.854389  1.444195e-14
## YearBuilt      -2.980583e+04 6.185075e+03 -4.818992  1.757134e-06
## I(YearBuilt^2)  7.926759e+00 1.591235e+00  4.981513  7.889630e-07
```

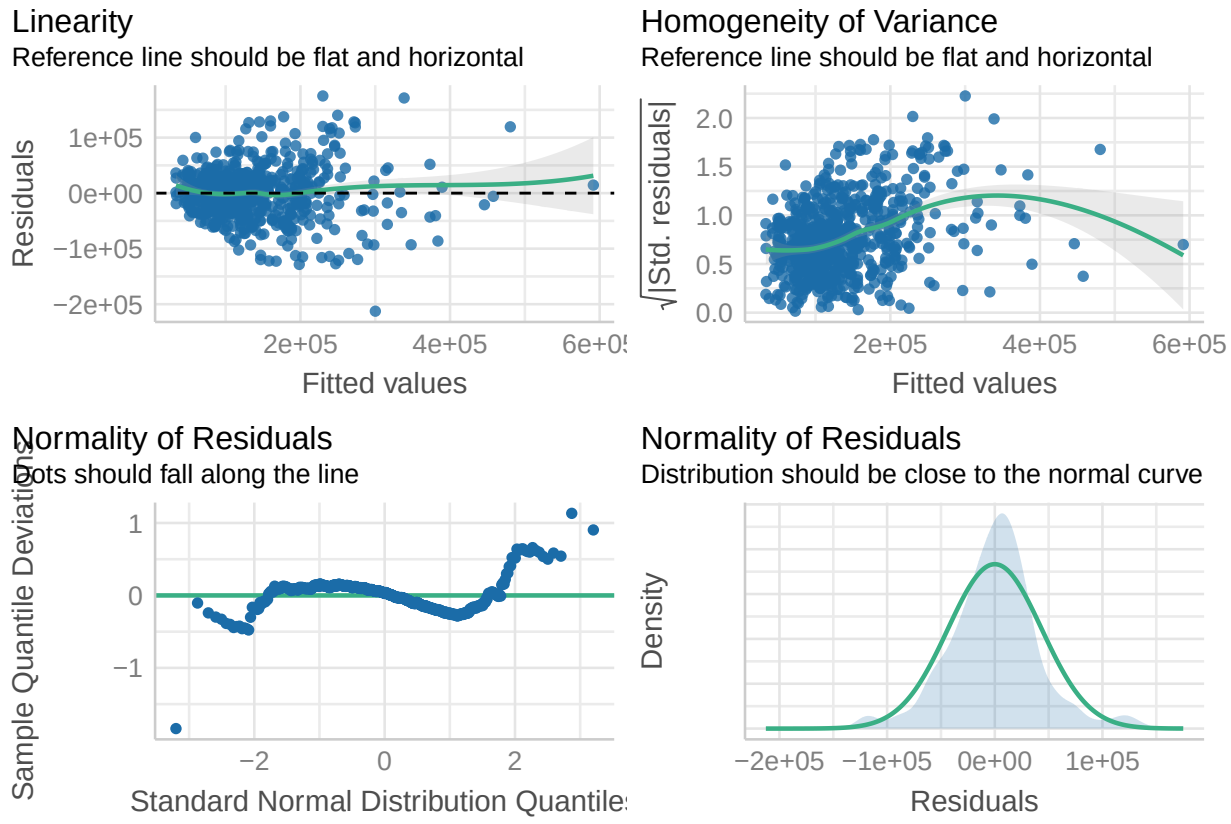
We can view this model:

```
extract_eq(compound_model, use_coefs = T)
```

$$\widehat{\text{SalePrice}} = 27996226.26 + 68.16(\text{SquareFeet}) + 4721.14(\text{LotSize}) - 29805.83(\text{YearBuilt}) + 7.93(\text{YearBuilt\^{}2}) \quad (1)$$

Based on the following, we can see this model doesn't too badly violate our assumptions. Thus, we can continue to interpret our coefficients.

```
check_model(compound_model, check = c("linearity", "homogeneity",
                                      "qq", "normality"))
```



Interpret coefficients

- `(Intercept)` = 2.799623e+07 is SalePrice when SquareFeet, LotSize, and YearBuilt are all 0 (not very realistic).
- `SquareFeet` = 68.16152 represents the average increase in SalePrice for each increase by 1 of SquareFeet, when LotSize and YearBuilt are both 0 or don't change.
- `LotSize` = 4721.14 is the average change in SalePrice for each increase by 1 of LotSize when SquareFeet and YearBuilt are both 0 or don't change.
- `YearBuilt` = −29805.83 is the average change in SalePrice for each increase by 1 of YearBuilt, when SquareFeet and LotSize are both 0 or don't change.

- `I(YearBuilt^2)` = 7.93 is the average change in SalePrice for each increase in $YearBuilt^2$ by 1, given the SquareFeet and LotSize are both 0 or don't change.

side note, not sure if this model is correct because we haven't learned about interpreting a complete second order model (not that this is one because it doesn't have a $\beta_1(X_1 * X_2)$ term) even though we learned about them in the textbook. i feel sort of confident in my interpretations, although this might have gotten too sophisticated.