# Modeling House Price Proposal

Corbin Christiansen AND Michael Bjornberg AND George Westover

## 1 Group Members

The members of our group are Corbin Christiansen, Michael Bjornberg, George Westover

## 2 Data Set

This data set is a compilation of houses, their various features(house size, floors, bathroom count, ect…), and the details around which it was sold(sale date, price, etc…). This data set contains categorical and numerical data, which could prove a challenge for data analysis. This data set can be found online at the CMU S&DS data repository site.

```r
ames_housing <- read_csv("ames-housing.csv")
```

```
Rows: 2930 Columns: 82
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (43): MS.Zoning, Street, Alley, Lot.Shape, Land.Contour, Utilities, Lot....
dbl (39): Order, PID, MS.SubClass, Lot.Frontage, Lot.Area, Overall.Qual, Ove...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
head(ames_housing)
```

```
# A tibble: 6 x 82
  Order       PID MS.SubClass MS.Zoning Lot.Frontage Lot.Area Street Alley
  <dbl>     <dbl>       <dbl> <chr>            <dbl>    <dbl> <chr>  <chr>
1     1 526301100          20 RL                 141    31770 Pave   <NA>
```

```
2     2 526350040          20 RH                  80     11622 Pave   <NA>
3     3 526351010          20 RL                  81     14267 Pave   <NA>
4     4 526353030          20 RL                  93     11160 Pave   <NA>
5     5 527105010          60 RL                  74     13830 Pave   <NA>
6     6 527105030          60 RL                  78      9978 Pave   <NA>
# i 74 more variables: Lot.Shape <chr>, Land.Contour <chr>, Utilities <chr>,
#   Lot.Config <chr>, Land.Slope <chr>, Neighborhood <chr>, Condition.1 <chr>,
#   Condition.2 <chr>, Bldg.Type <chr>, House.Style <chr>, Overall.Qual <dbl>,
#   Overall.Cond <dbl>, Year.Built <dbl>, Year.Remod.Add <dbl>,
#   Roof.Style <chr>, Roof.Matl <chr>, Exterior.1st <chr>, Exterior.2nd <chr>,
#   Mas.Vnr.Type <chr>, Mas.Vnr.Area <dbl>, Exter.Qual <chr>, Exter.Cond <chr>,
#   Foundation <chr>, Bsmt.Qual <chr>, Bsmt.Cond <chr>, ...
```

# 3 Potential Questions of Interest

1. **How does the age of a house effect the price at which it was sold?**

This relationship is important to know for families because knowing the timing of when you should sell your house for maximum value can help families get bigger homes when they need it to accommodate bigger family sizes.

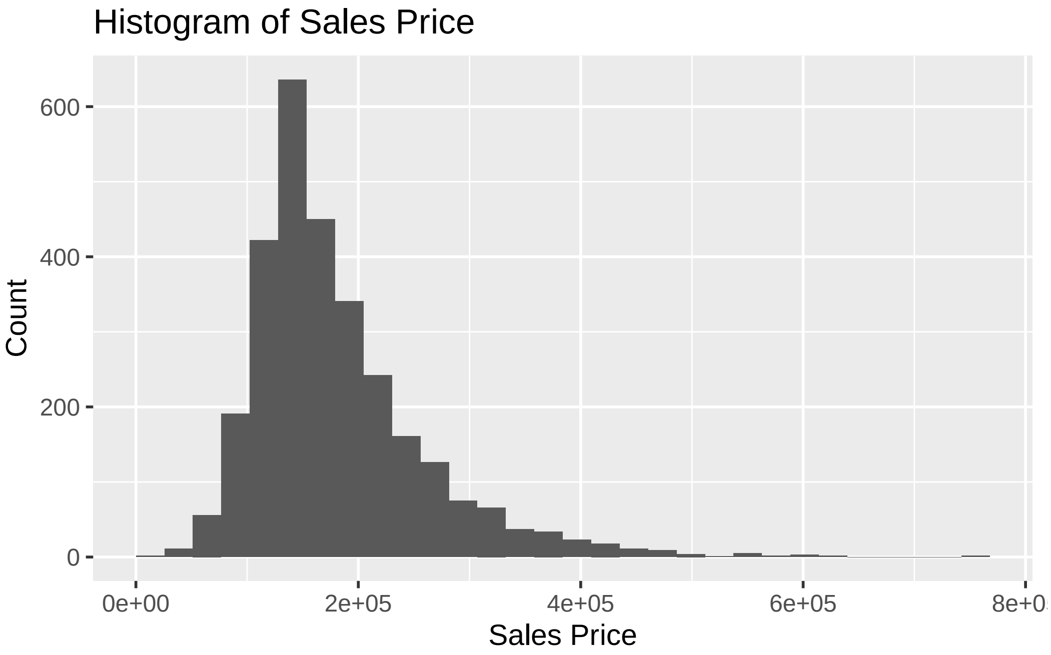
2. **How do other features of the house(ie. lot size, number of floors, number of bedrooms, etc.) effect the price at which it was sold?**

This relationship could be important for house builders to estimate the profit they could make on potential houses, based on the features they plan on adding.

# 4 Exploratory Data Analysis

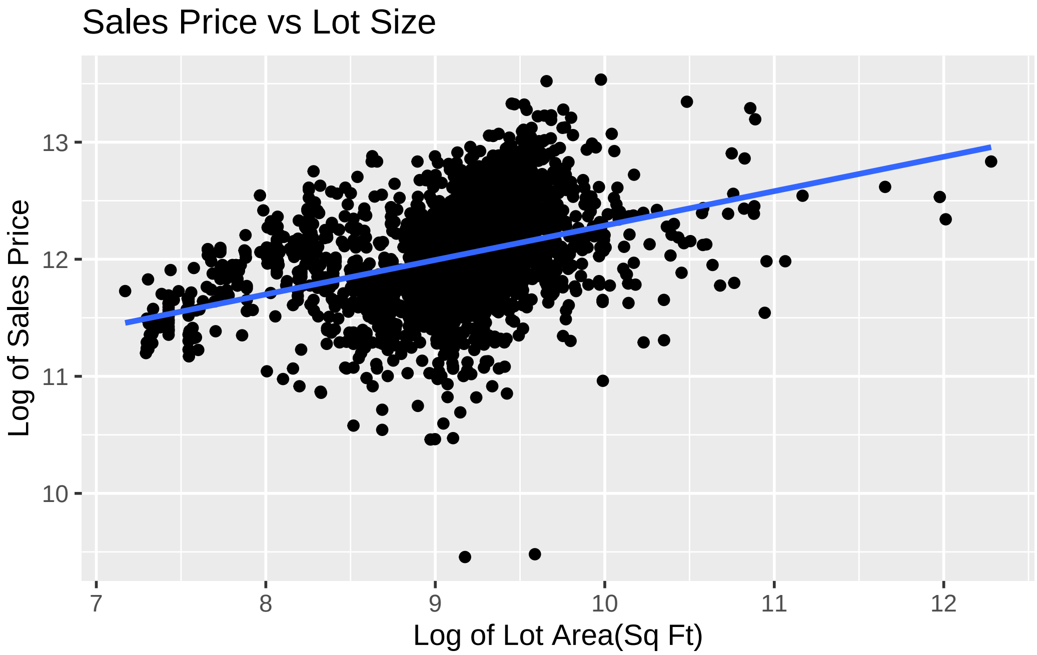
```
ames_housing %>%
  ggplot(aes(x = SalePrice)) +
  geom_histogram() +
  labs(title = "Histogram of Sales Price",
       x = "Sales Price",
       y = "Count")
```

```
`stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```



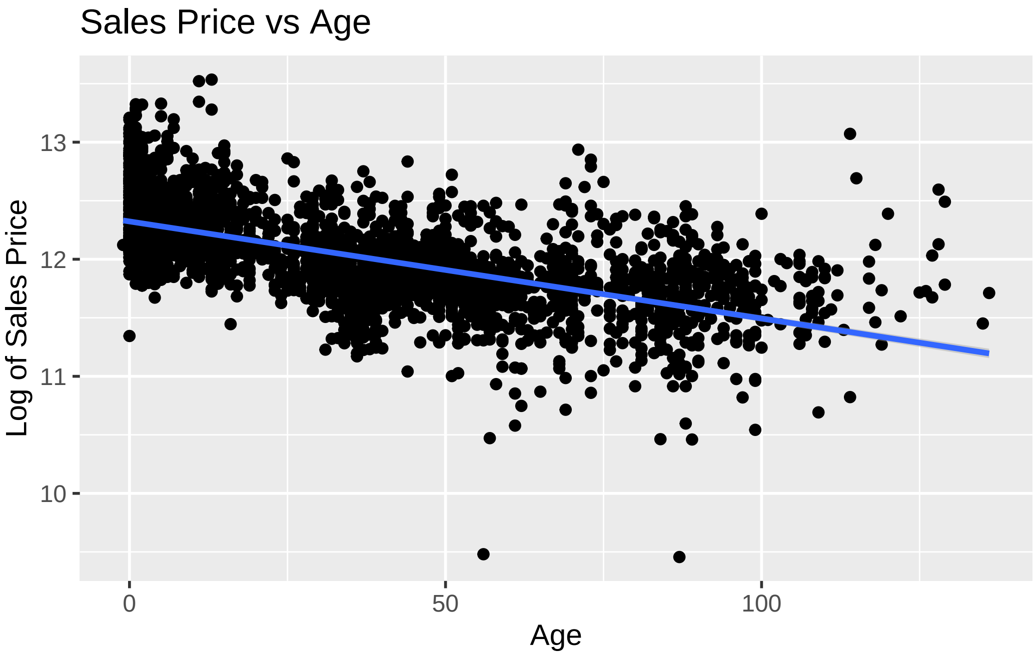
```
ames_housing %>%
  ggplot(aes(x = log(Lot.Area), y = log(SalePrice))) +
  geom_point() +
  geom_smooth(method = 'lm', se = F) +
  labs(title = "Sales Price vs Lot Size",
       x = "Log of Lot Area(Sq Ft)",
       y = "Log of Sales Price")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



```
ames_housing %>%
  ggplot(aes(x = (Yr.Sold - Year.Built), y = log(SalePrice))) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(title = "Sales Price vs Age",
       x = "Age",
       y = "Log of Sales Price")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



1st Graph. We found that the distribution of the sales price is right-skewed by plotting a histogram of the sales price. This is important because it means we will probably have to do some transformations to get linear relationships.

2nd Graph. With some log transformations on both the response and explanatory variables, we were able to find a strong positive linear relationship with sales price and lot size. This is noteworthy because this means the first assumption of linear regression is met

3rd Graph: This shows the relationship between the age of the house and the price at which it was sold. In order to find a linear relationship we needed to take the log of the sales price. We found a week, negitive relationship between the age and the price, meaning that as the house gets older it will sell for less money.

## 5 Preliminary Model Fit

```
house.lm = lm(I(log(SalePrice))~I(Yr.Sold - Year.Built), data = ames_housing)
summary(house.lm)
```

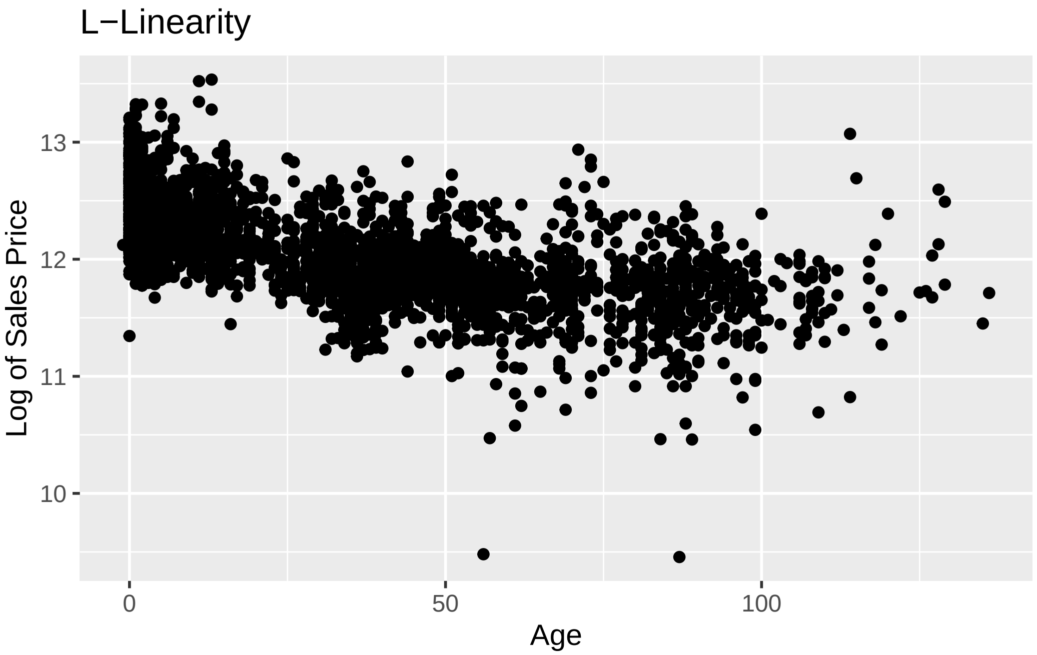
```
Call:
lm(formula = I(log(SalePrice)) ~ I(Yr.Sold - Year.Built), data = ames_housing)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-2.37843 -0.18949 -0.02904  0.17106  1.69300 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             12.3229524  0.0092808 1327.79   <2e-16 ***
I(Yr.Sold - Year.Built) -0.0082885  0.0001959  -42.31   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3211 on 2928 degrees of freedom
Multiple R-squared:  0.3794,    Adjusted R-squared:  0.3792 
F-statistic:  1790 on 1 and 2928 DF,  p-value: < 2.2e-16
```

$$\log(\widehat{\text{Sale Price}}_i) = 12.3229524 - 0.0082885\widehat{\text{Age}}_i$$

This relationship is important to know for families because knowing the timing of when you should sell your house for maximum value can help families get bigger homes when they need it to accommodate bigger family sizes.

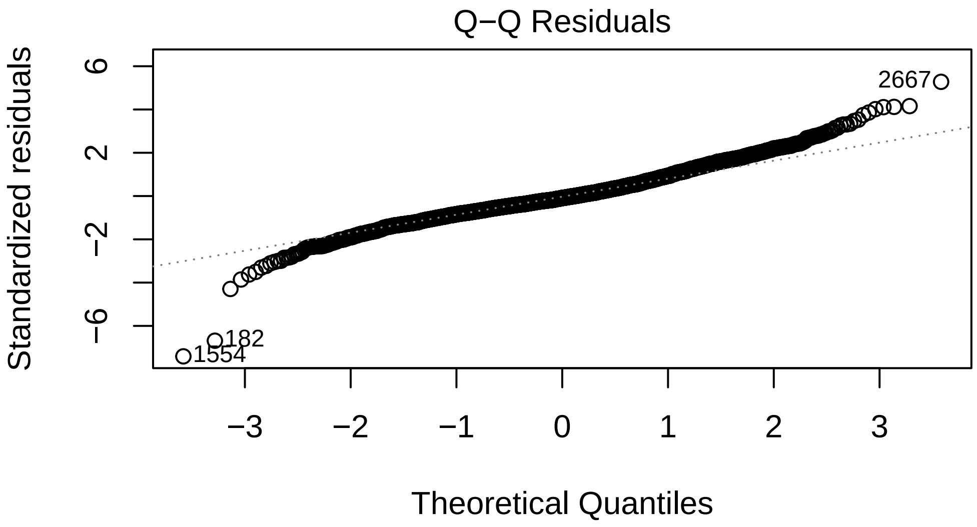
```r
#L- linearity
ames_housing %>%
  ggplot(aes(x = (Yr.Sold - Year.Built), y = log(SalePrice))) +
  geom_point() +
  labs(title = "L-Linearity",
       x = "Age",
       y = "Log of Sales Price")
```



Linearity - Linearity is met because the scatter plot of the age of the house against the log of the Sale Price looks linear. The residuals vs fitted values plot also shows a linear trend.

Independence - Is not met because the houses sampled are from the same area in Ames, Iowa.

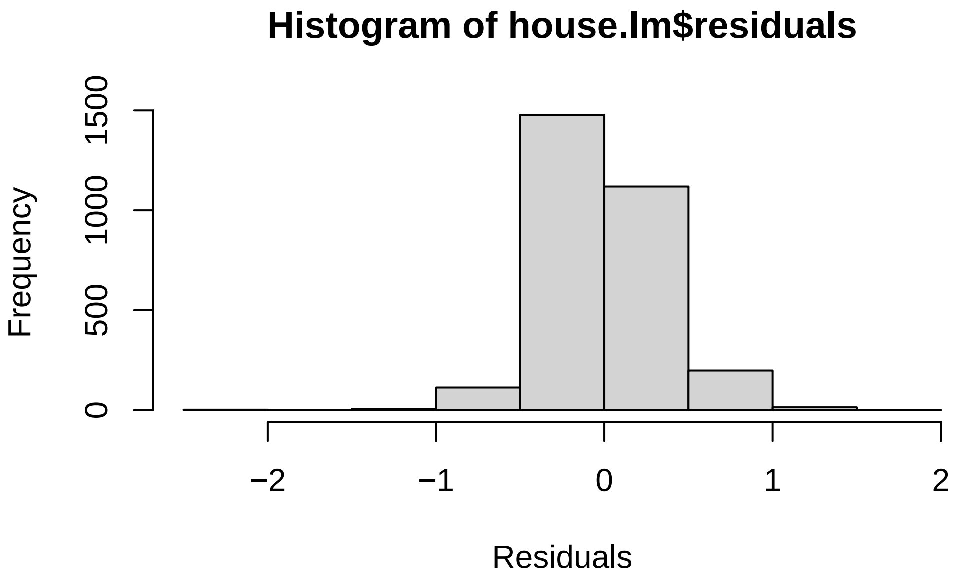
```r
#N- Normality
plot(house.lm, which = 2, sub.caption = "")
```



```r
shapiro.test(house.lm$residuals)
```

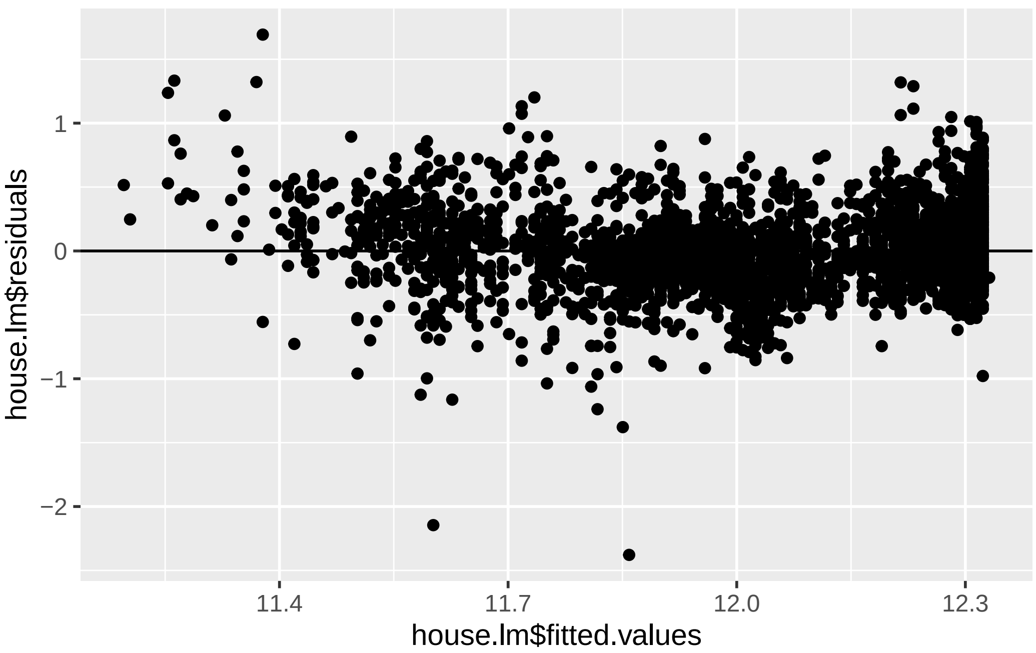
```
	Shapiro-Wilk normality test

data:  house.lm$residuals
W = 0.97212, p-value < 2.2e-16
```

```r
hist(house.lm$residuals, xlab = "Residuals")
```



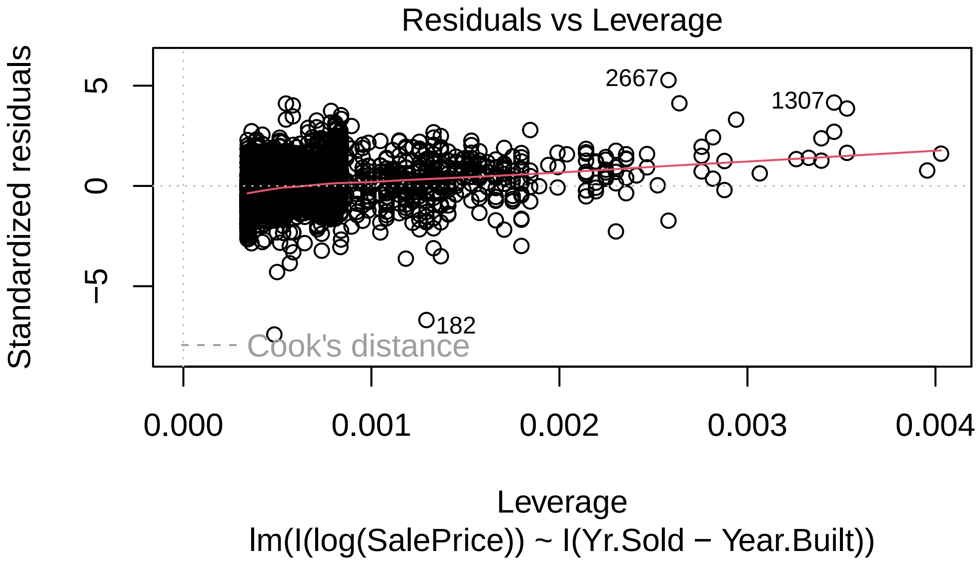
Normality - Through the Shapiro-Wilk test and the Q-Q plot, and a histogram of the residuals, it seems as if the data is short-tailed and not normal. Although this assumption seems violated, because of how large our data set is, 2930 data points, this violation is not a big problem.

```
#E-Equal Variance
ggplot(mapping = aes(y = house.lm$residuals, x = house.lm$fitted.values)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Equal Variance - The residuals vs fitted values plot seems to show homoskedasticity except for a slight curve and truncation at the end, and a few outlines here and there. All in all, the assumption of Equal Variance seems to be met

```
#Outliers
plot(house.lm, which = 5)
```



Outlines - Based on the Cook's distance, we have no undue influential points or outlines

Our estimates of $\beta_0$ and $\beta_1$ tell us that for every year increase in the age of the house we estimate that the average log price of the house will increase by 0.0082885.

We also estimate that a brand new house will have an average log sell price of 12.3229524.

Based on the summary we received from r about our lm model, our p-value for our $\beta_1$s is nearly 0, meaning we reject the null hypothesis that there is no relationship.

Our $R^2$ statistics is 0.3793 which means that 37.93% of the variance we see in the log of Sales Price can be explained by the age of the house.