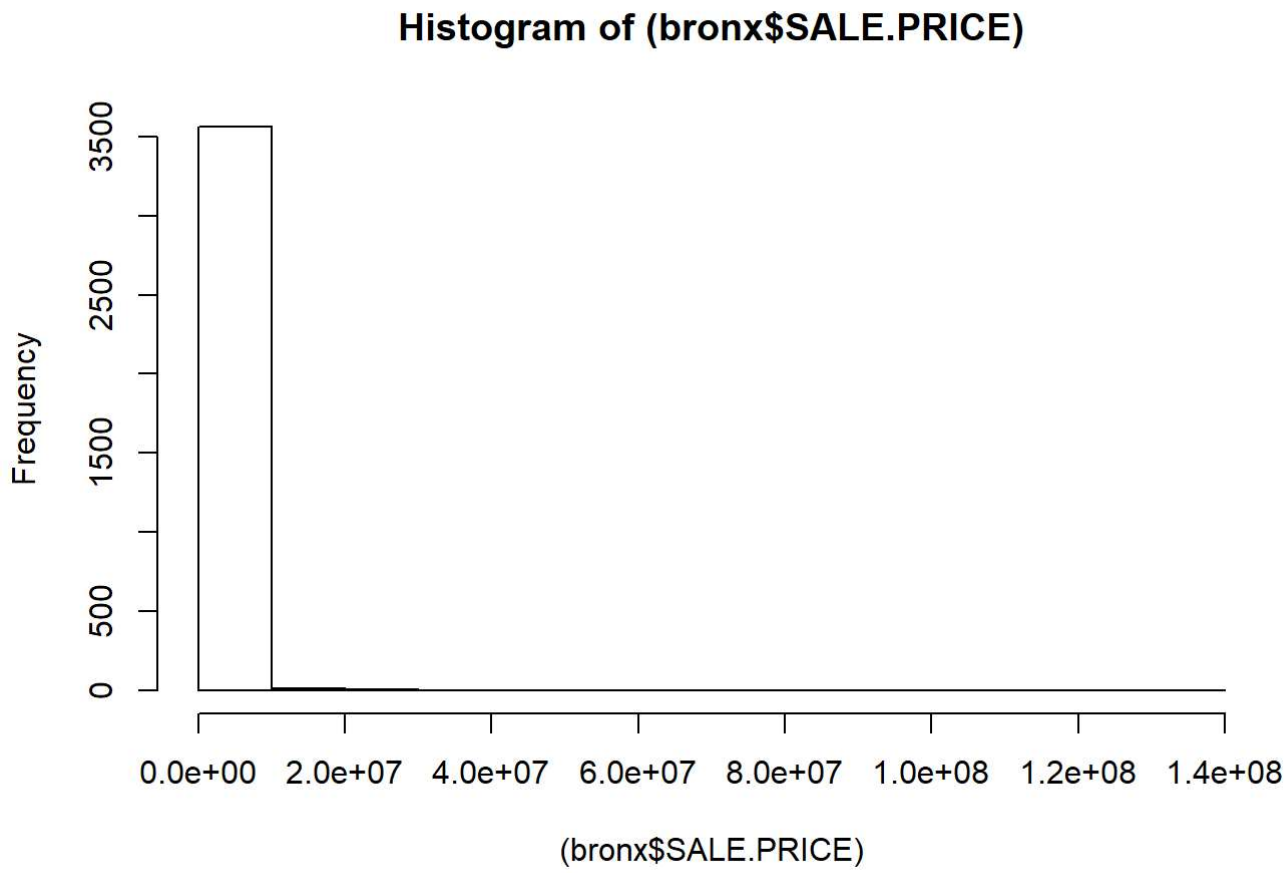
```
library(openxlsx)
bronx <- read.xlsx("rollingsales_bronx.xlsx")
```

1.

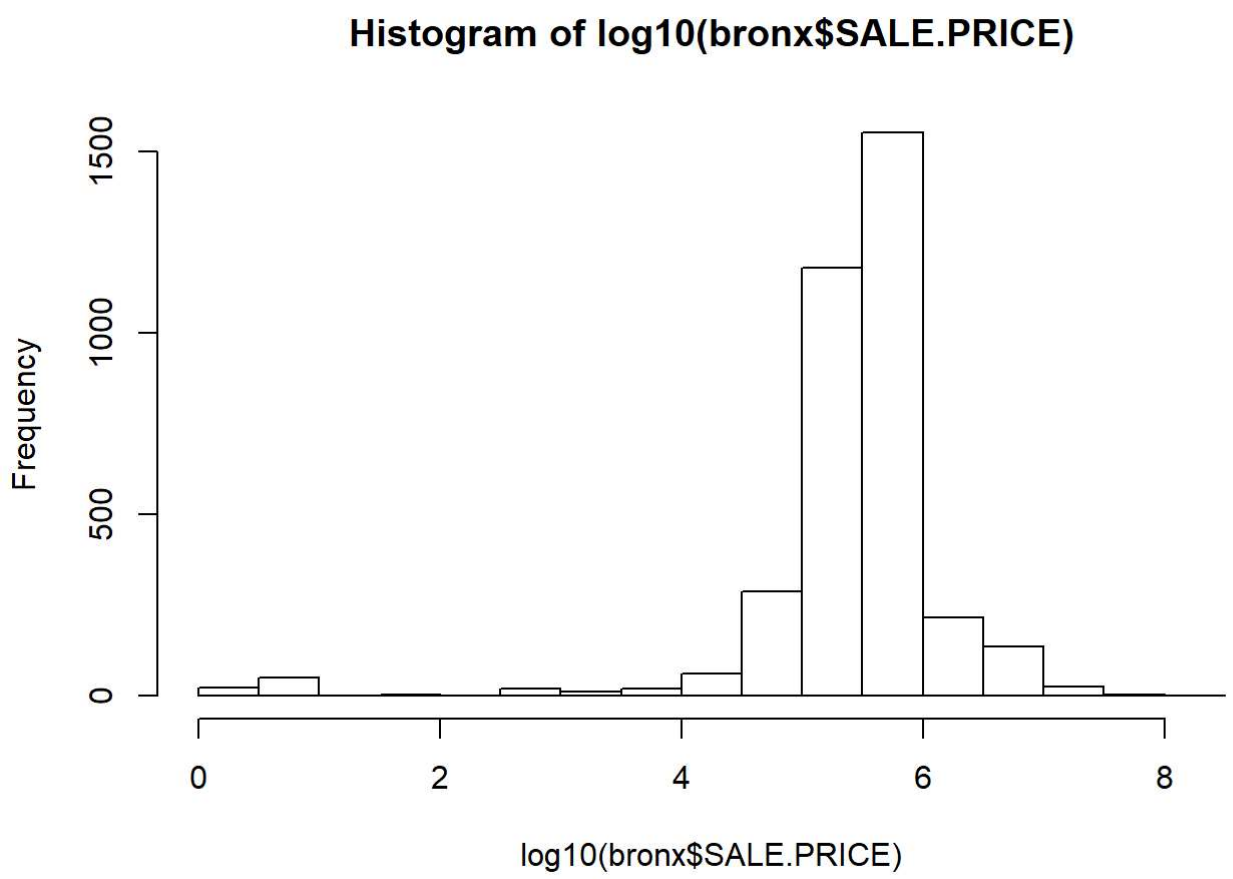
a. I would try to predict the sale price. This could be done based on factors like year built and square footage. I plan to model them with a linear regression.

Descriptive Analysis: Many of the the data points had 0 as a sale price, so I filtered them out. They probably represent missing data.

```
bronx <- bronx[bronx$SALE.PRICE > 0, ]
hist((bronx$SALE.PRICE))
```

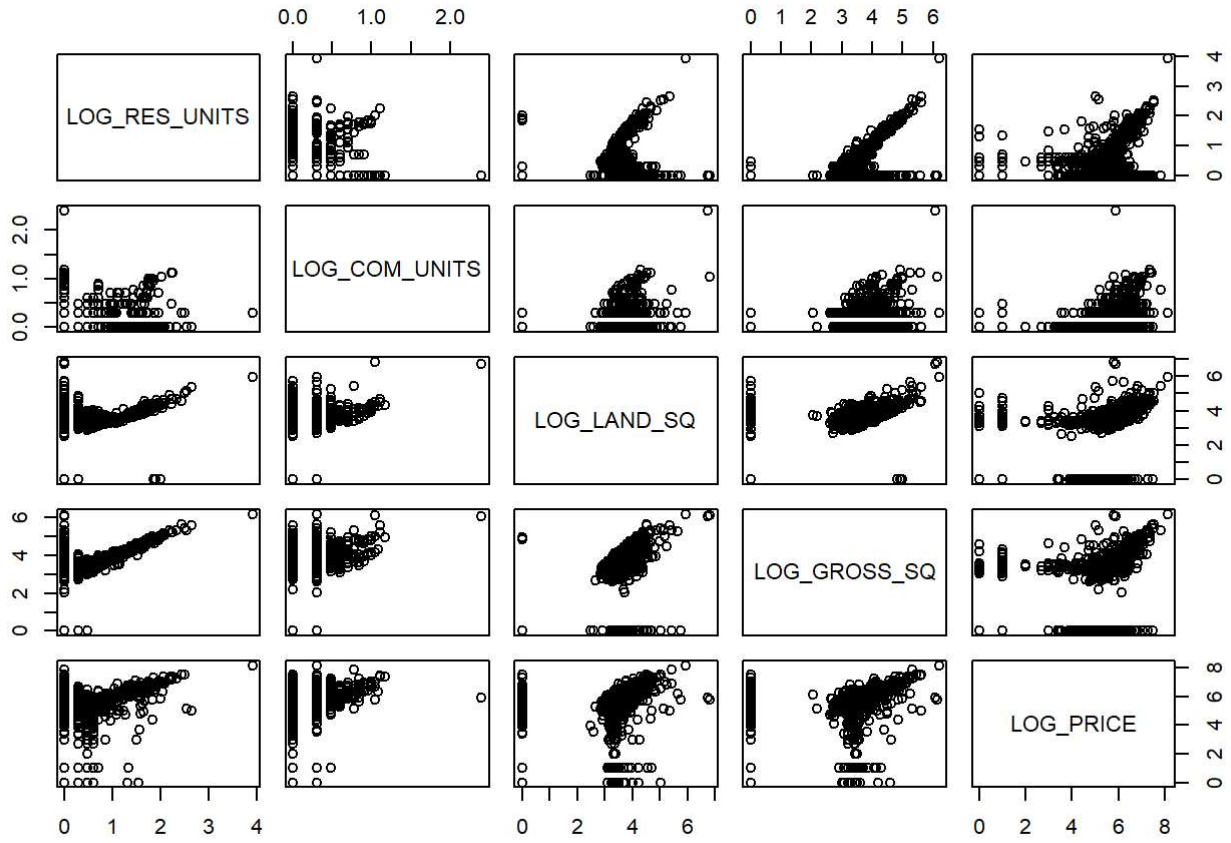



```
hist(log10(bronx$SALE.PRICE))
```



Many of the variables become more normal distributions after taking a logarithm.

```
bronx$LOG_RES_UNITS <- log10(bronx$RESIDENTIAL.UNITS + 1)
bronx$LOG_COM_UNITS <- log10(bronx$COMMERCIAL.UNITS + 1)
bronx$LOG_LAND_SQ <- log10(bronx$LAND.SQUARE.FEET + 1)
bronx$LOG_GROSS_SQ <- log10(bronx$GROSS.SQUARE.FEET + 1)
bronx$LOG_PRICE <- log10(bronx$SALE.PRICE)
vars = c("LOG_RES_UNITS", "LOG_COM_UNITS", "LOG_LAND_SQ", "LOG_GROSS_SQ", "LOG_PRICE")
pairs(bronx[,vars])
```



The pairs plot seems to indicate linear relationships between the log-re-expressed variables. Therefore, multivariate regression might be an appropriate choice,

b. We will be using linear regression to predict housing prices.

```
model <- lm(LOG_PRICE~., data = bronx[, vars])
summary(model)
```

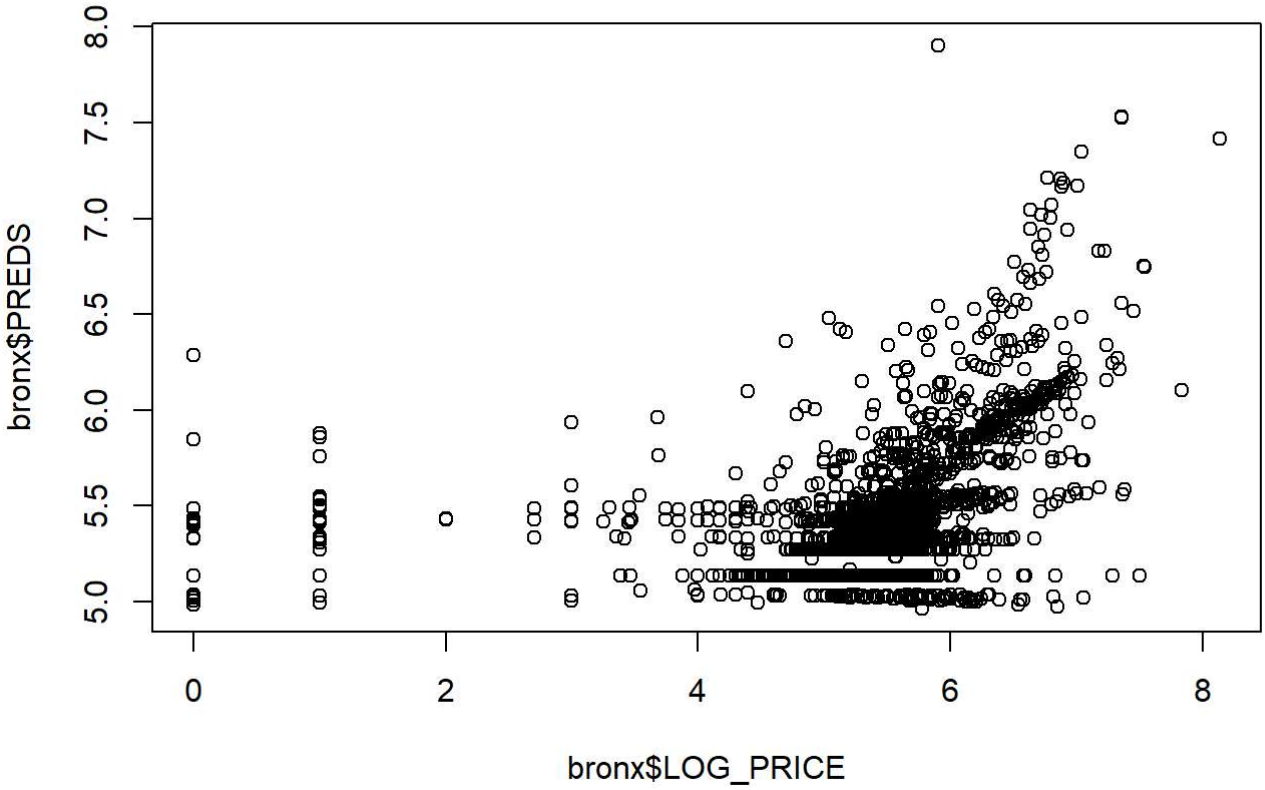
```
##
## Call:
## lm(formula = LOG_PRICE ~ ., data = bronx[, vars])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2871 -0.0738  0.1454  0.2885  2.3670
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.13471    0.02519 203.855   <2e-16 ***
## LOG_RES_UNITS  0.46218    0.04541  10.178   <2e-16 ***
## LOG_COM_UNITS  1.10781    0.09595  11.545   <2e-16 ***
## LOG_LAND_SQ   -0.03011    0.01988  -1.515   0.1299
## LOG_GROSS_SQ   0.05053    0.02224   2.272   0.0231 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8136 on 3584 degrees of freedom
## Multiple R-squared:  0.1154, Adjusted R-squared:  0.1144
## F-statistic: 116.8 on 4 and 3584 DF,  p-value: < 2.2e-16
```

The R^2 of 0.1154 does not indicate a very tight fit. However, the p-value generated from the F-statistic indicates that the linear relationship is indeed significant even if it does not explain most of the variation. For cleanining, I had to remove all entries with a sale price of zero. This was because, the number of zeroes was significant and did not fit the overall lognormal distribution. I believe the zeroes represent missing data.

2.

a.

```
bronx$PREDS <- predict(model, bronx[,vars])
plot(bronx$LOG_PRICE, bronx$PREDS)
```



We can see that our model works generally pretty well. However, it overestimates sales with low prices pretty significantly. This indicates that the data has a nonlinearity as the price drops.

b.

```
summary(model)
```

```
##
## Call:
## lm(formula = LOG_PRICE ~ ., data = bronx[, vars])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2871 -0.0738  0.1454  0.2885  2.3670
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.13471    0.02519 203.855   <2e-16 ***
## LOG_RES_UNITS  0.46218    0.04541  10.178   <2e-16 ***
## LOG_COM_UNITS  1.10781    0.09595  11.545   <2e-16 ***
## LOG_LAND_SQ   -0.03011    0.01988  -1.515   0.1299
## LOG_GROSS_SQ   0.05053    0.02224   2.272   0.0231 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8136 on 3584 degrees of freedom
## Multiple R-squared:  0.1154, Adjusted R-squared:  0.1144
## F-statistic: 116.8 on 4 and 3584 DF,  p-value: < 2.2e-16
```

We performed an F-test to test the significances of the linear model. The p-value of 2.2e-16 indicates a very high level of significance, This means that the total linear combination of predictors found by our model does have a positive relationship with the actual sale price. The low correlation indicates that although we are sure a relationship exists, it might not be that strong.

c. I am concerned about omitting the zeroes from our data. Our filtered data set might not be representative of the original. In general, there is something nonlinear occuring when housing prices drop that our model is not able to predict correctly. Understanding that nonlinearity warrants additional study.