DataSet features

The Dataset: Albuquerque House Pricing

- **PRICE** (Target Variable) : House Selling Price in dollars
- **SquareFeet :** Square feet of living space
- **AgeYear :** The age of the house (years)
- **NumberFeatures :** Sort the of 11 features (dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access)
- **Northeast:** If the building is located in the northeast sector of city (Yes or No)
- **CustomBuild :** Custom built (Yes or No)
- **CornerLot :** If the building is located in a Corner location (Yes or No)

There are 117 rows and 7 variables in this data set.

Please note the 49 null-values in the AgeYear column

Objectives and Issues

I wanna try to highlight an eventual relation between the target Price and the features. I will look for linear regression models to fit the data.

Missing Data: I will fill the missing data with a simple linear interpolation

Categorical Variables: I will perform one-hot encoding for the categorical variables (Northeast, CustomBuild, CornerLot)

Summary Statistics

Of the 7 variables, Northeast, CustomBuild, CornerLot are categorical variables, Price, SquareFeet and NumberFeatures are Int, AgeYear is a float.

```
RangeIndex: 117 entries, 0 to 116
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Price           117 non-null    int64
 1   SquareFeet      117 non-null    int64
 2   AgeYear         68 non-null     float64
 3   NumberFeatures  117 non-null    int64
 4   Northeast       117 non-null    object
 5   CustomBuild     117 non-null    object
 6   CornerLot       117 non-null    object
dtypes: float64(1), int64(3), object(3)
```

After the linear interpolation to fill the missing values and a change in the dytpe of AgeYear:

```
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Price           117 non-null    int64
 1   SquareFeet      117 non-null    int64
 2   AgeYear         117 non-null    int64
 3   NumberFeatures  117 non-null    int64
 4   Northeast       117 non-null    object
 5   CustomBuild     117 non-null    object
 6   CornerLot       117 non-null    object
dtypes: int64(4), object (3)
memory usage: 6.5+ KB
```

After the one-hot encoding:

	0	1	2	3	4
Northeast	Yes	Yes	Yes	Yes	Yes
CustomBuild	Yes	Yes	Yes	Yes	Yes
CornerLot	No	No	No	No	No

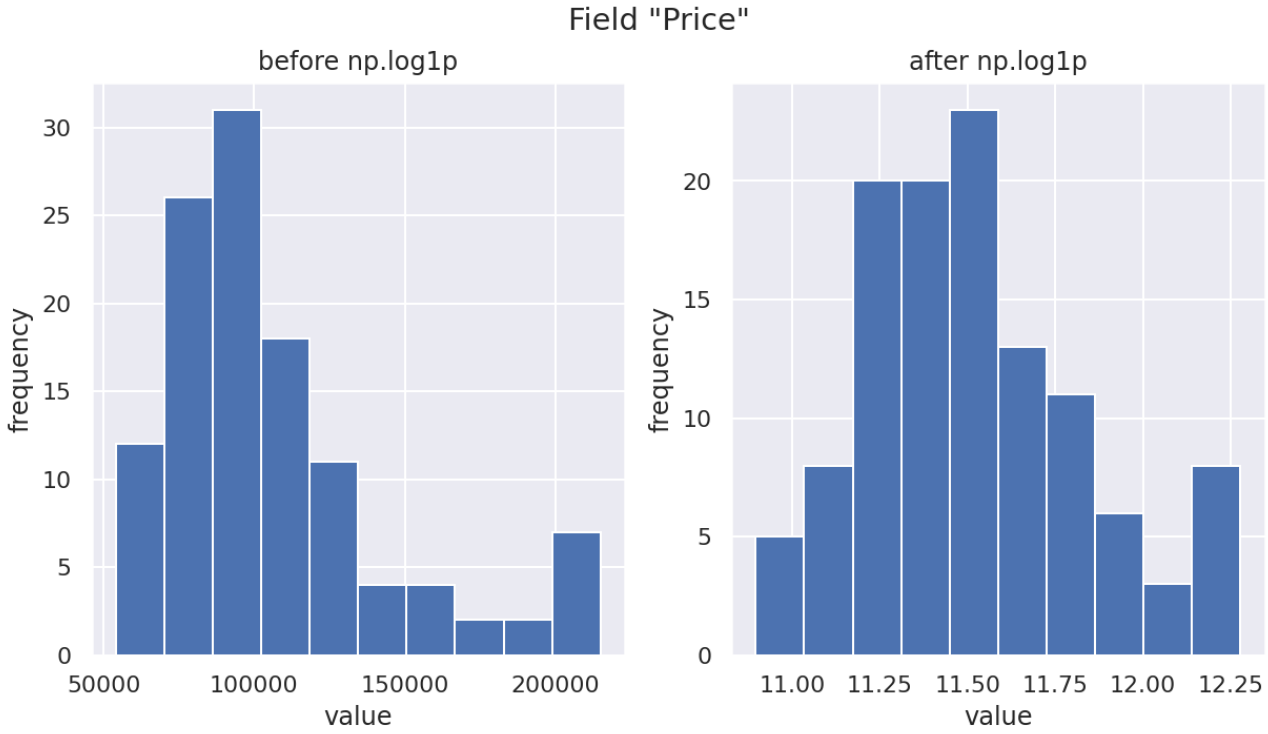
Let's see the common statistics values:

	count	mean	std	min	25%	50%	75%	max
Price	117.0	106273.504274	38043.698543	54000.0	78000.0	96000.0	120000.0	215000.0
SquareFeet	117.0	1653.854701	523.722802	837.0	1280.0	1549.0	1894.0	3750.0
AgeYear	117.0	18.068376	13.370533	1.0	6.0	15.0	27.0	53.0
NumberFeatures	117.0	3.529915	1.405486	0.0	3.0	4.0	4.0	8.0
Northeast_Yes	117.0	0.666667	0.473432	0.0	0.0	1.0	1.0	1.0
CustomBuild_Yes	117.0	0.230769	0.423137	0.0	0.0	0.0	0.0	1.0
CornerLot_Yes	117.0	0.188034	0.392420	0.0	0.0	0.0	0.0	1.0

Regarding skewed values (with a skew limit `0.75)`:

	Skew
Price	1.375404
SquareFeet	1.187560
AgeYear	0.765807

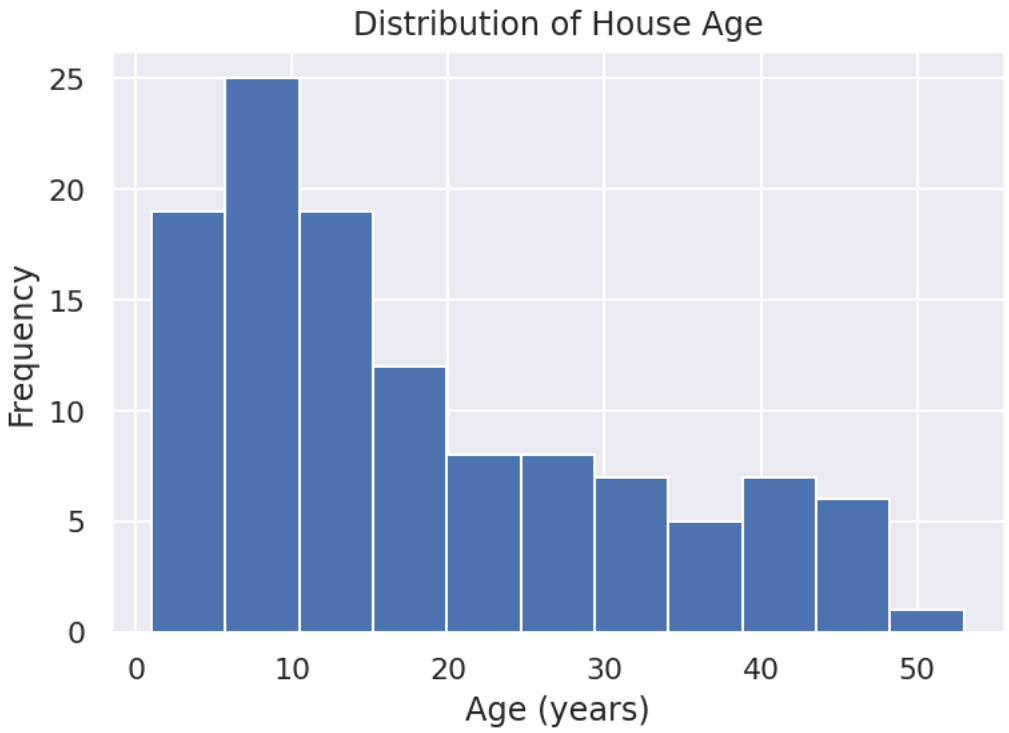
Let's perform a log transformation for the field Price.

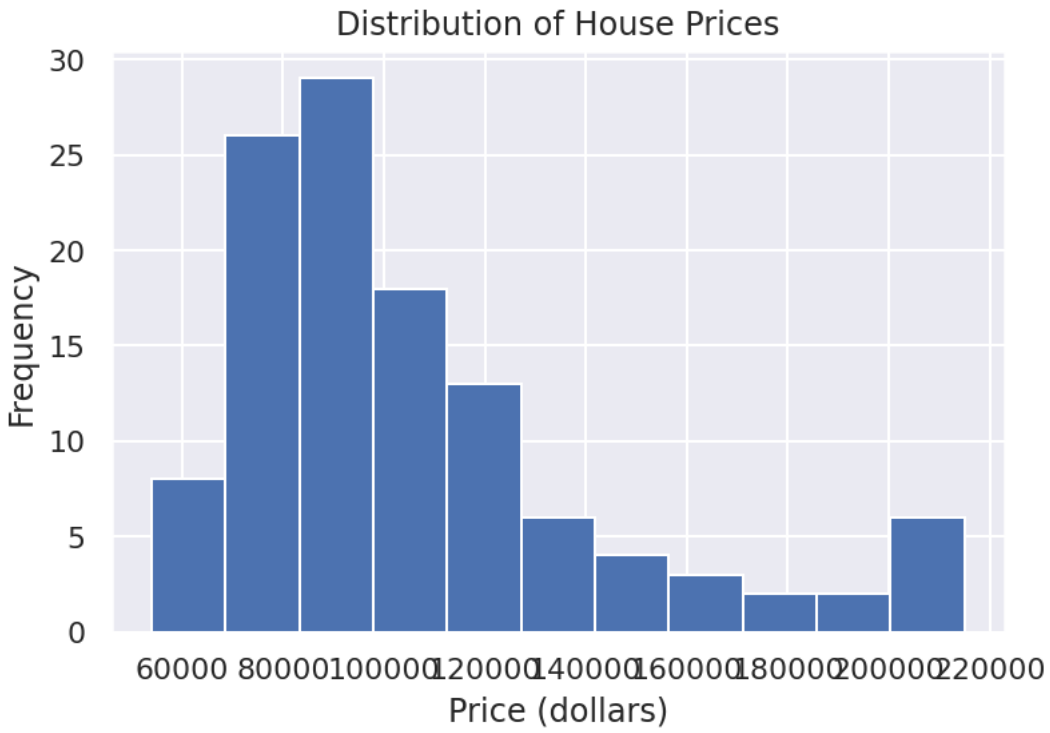


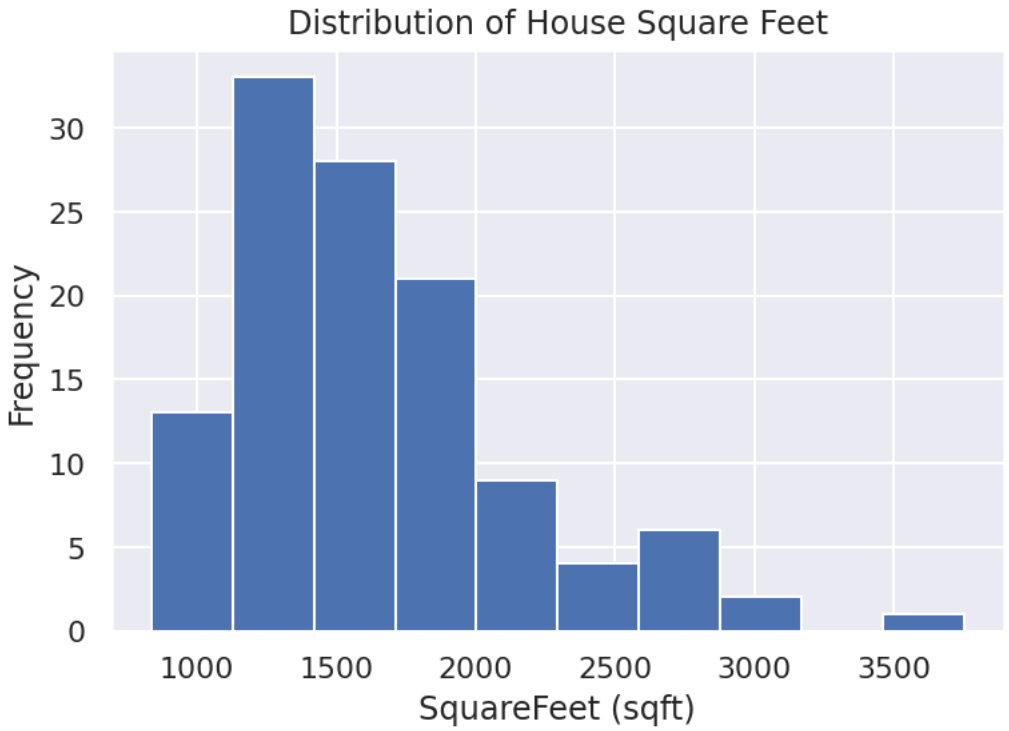
And after the other 2 log transformations, let's see the statistics for the cleaned and transformed dataset

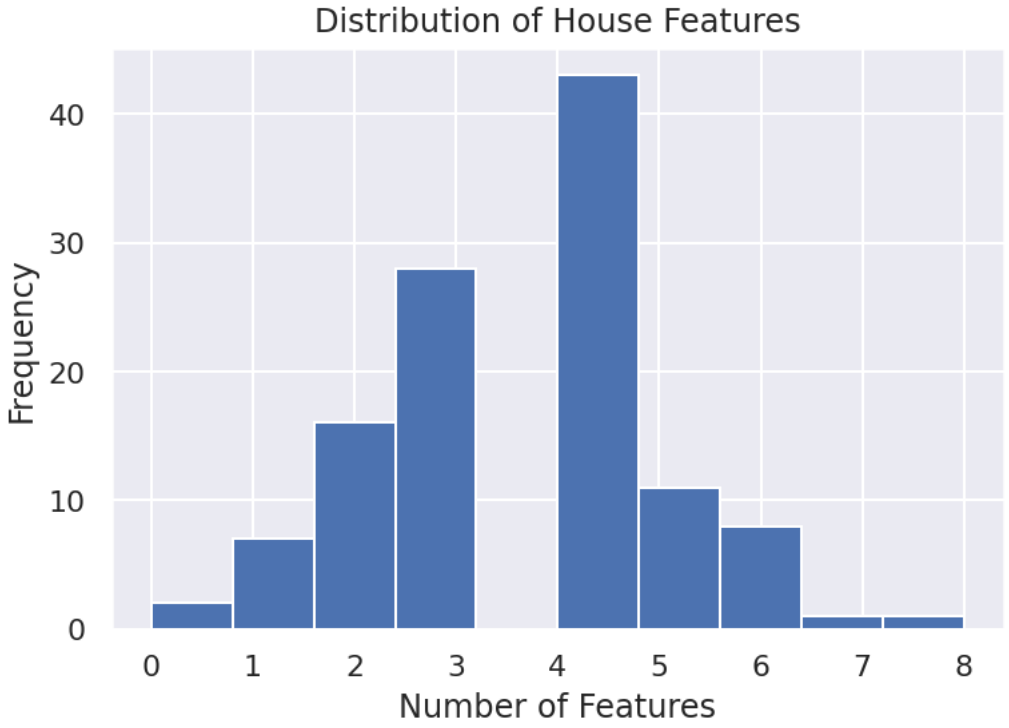
	count	mean	std	min	25%	50%	75%	max
Price	117.0	11.519675	0.319825	10.896758	11.264477	11.472114	11.695255	12.278398
SquareFeet	117.0	7.366834	0.294580	6.731018	7.155396	7.346010	7.546974	8.229778
AgeYear	117.0	2.670314	0.793985	0.693147	1.945910	2.772589	3.332205	3.988984
NumberFeatures	117.0	3.529915	1.405486	0.000000	3.000000	4.000000	4.000000	8.000000
Northeast_Yes	117.0	0.666667	0.473432	0.000000	0.000000	1.000000	1.000000	1.000000
CustomBuild_Yes	117.0	0.230769	0.423137	0.000000	0.000000	0.000000	0.000000	1.000000
CornerLot_Yes	117.0	0.188034	0.392420	0.000000	0.000000	0.000000	0.000000	1.000000

Let’s see the Histograms for each feature:

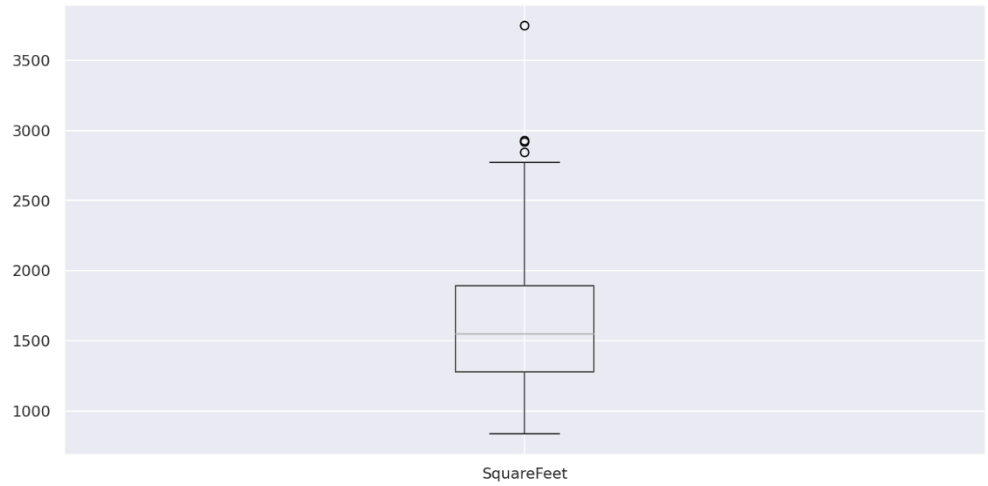





Distribution of House Square Feet
Frequency
0
5
10
15
20
25
30
1000
1500
2000
2500
3000
3500
SquareFeet (sqft)

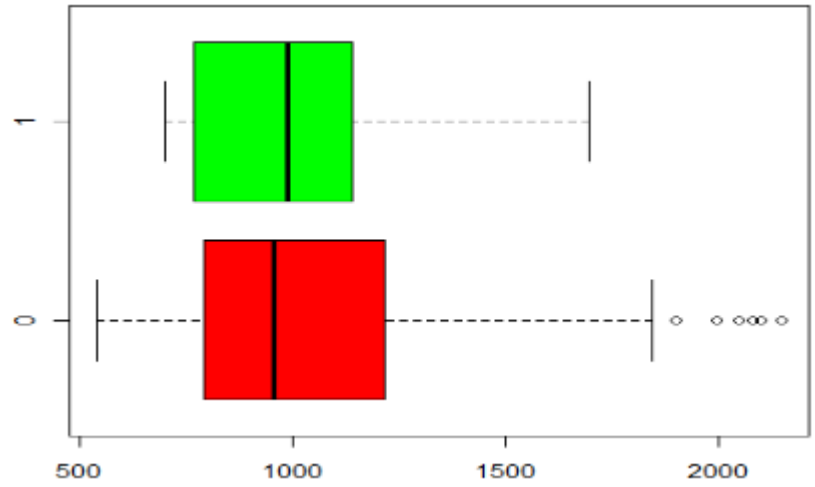

Distribution of House Features
Frequency
0
10
20
30
40
0
1
2
3
4
5
6
7
8
Number of Features

And now the BoxPlot

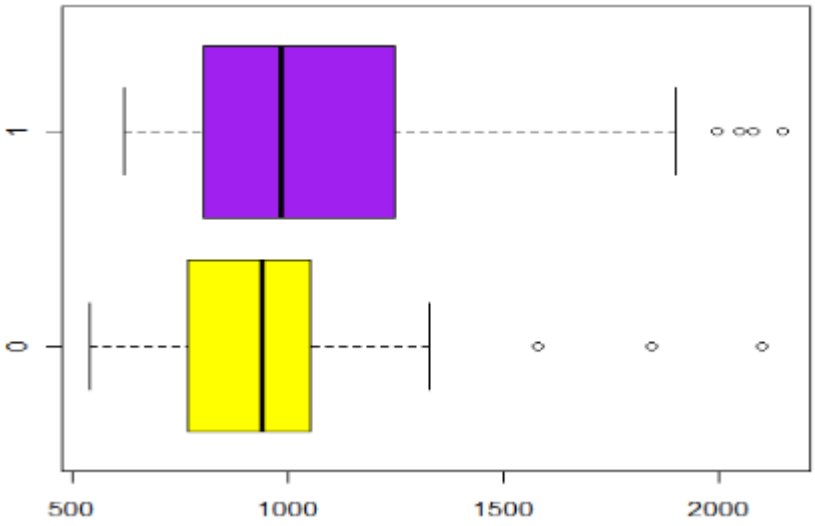


We can see Outliers in Price and SquareFeet Variables.

Here we can see the **Price Vs NorthEast Esposition BoxPlot**



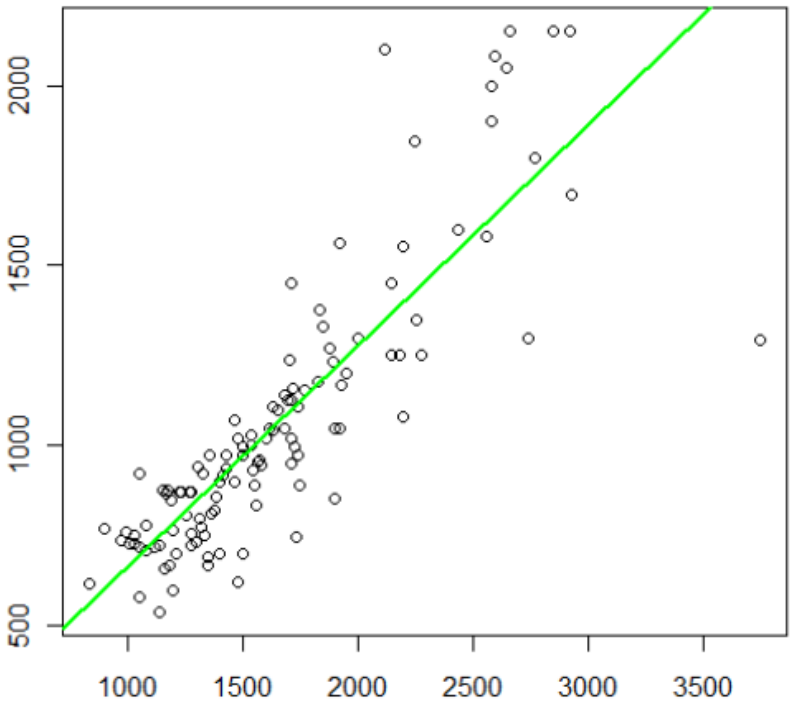
And here **Price Vs Corner Position**



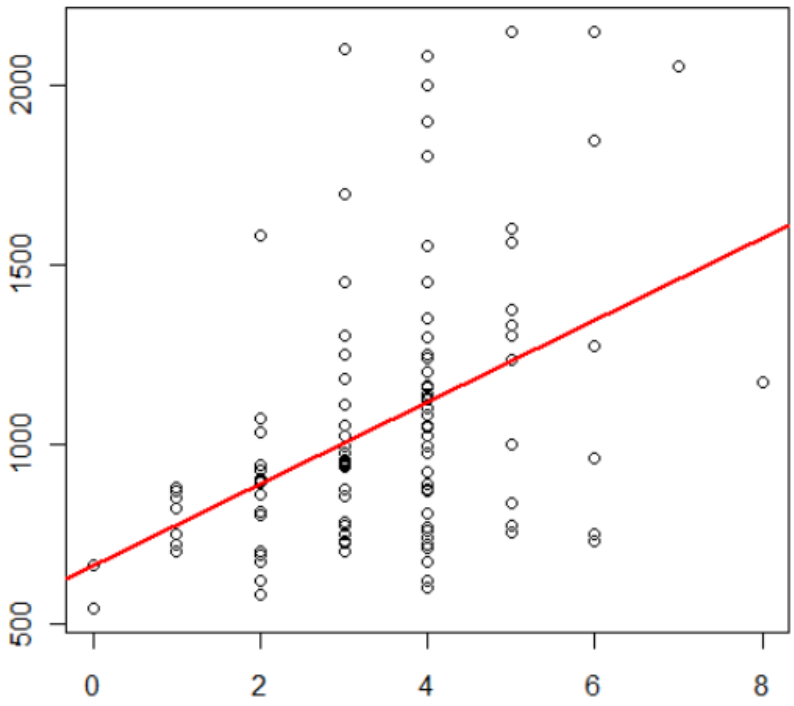
So far we've seen from the histograms none variables have a recognizable distribution (i.e gaussian)

From the boxplots we can se a relation between Price vs NorthEast position and Price vs Corner location.

I've tried to fit a linear regression between **Price and SquareFeet**



And a linear regression between **Price and NumberFeatures**



In conclusion we can't say the linear model is a proper model to describe the relations between our dataset variables. Maybe this due to the fact that the outliers are not normally distributes.

Hypothesis:

(H_0) for 3 linear regressions:

Linear Regressions with H_0 :

- Price Vs SquareFeet
- Price Vs NumberFeatures
- Price Vs Age

H_0 = the data can be modeled by setting all our Betas to zero.

In a linear regression usually Betas are the coefficients for each one of our features. We will reject the null Hypo if the p-value is small enough.

We will use F-Statistic to test the Hypos.

I wil try to test null H for Price Vs Surface

```
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.714
Model:                            OLS   Adj. R-squared:                  0.711
Method:                 Least Squares   F-statistic:                     286.6
Date:                Sun, 22 Nov 2020   Prob (F-statistic):           5.15e-33
Time:                        15:10:31   Log-Likelihood:                -787.49
No. Observations:                 117   AIC:                             1579.
Df Residuals:                     115   BIC:                             1584.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     47.8193     62.855      0.761      0.448     -76.684     172.323
SquareFeet     0.6137      0.036     16.931      0.000       0.542       0.685
==============================================================================
Omnibus:                       30.950   Durbin-Watson:                   1.536
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              216.286
Skew:                          -0.540   Prob(JB):                     1.08e-47
Kurtosis:                       9.573   Cond. No.                     5.77e+03
==============================================================================
```

P-value very very low**, so we reject the Null Hypothesis**.

Conclusion

Overall's data set is, in my opinion, poor. We should have more data.

The linear model doesn't fit very well our data, this is due to the non-normal distribution of the data and outliers, even with a log transformation. We can't find a linear regression just in the case of relation between Price and Number of Features. We should try with a multiple regression model.