Predicting Property’s Total Value

Group 3
Jaeyeon Won
Coskun Erden
Wanying Fu

Dataset

- **Dataset**: Home Value Data
- **Sample**: 100 random residential properties in a specific county in a specific state in the U.S. in March, 2008
- **Population**: All residential properties in a specific county in a specific state in the U.S. in March, 2008

Variables

- **Dependent Variable**:
  - **Total**: the total assessed value of the property (in U.S dollars)
- **Independent Variable**:
  - **Age***: the age of the structure (in years)
  - **Sq.Ft**: the area of the floor plan (in square feet)
  - **Acres**: how many acres is included in the plot (in acres)
  - **StoryNew**: how many stories the structure has (in stories)

    categories: 1 Story, 2 Stories
  - **BathsNew**: the number of bathrooms at the residence (in bathrooms)

    categories: 1 Bathroom, 2 Bathrooms, 3 Bathrooms

*Age was calculated by (2008 - YearBuilt)

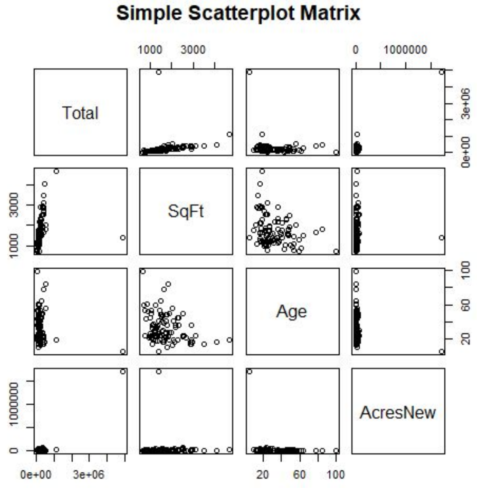
Research Questions

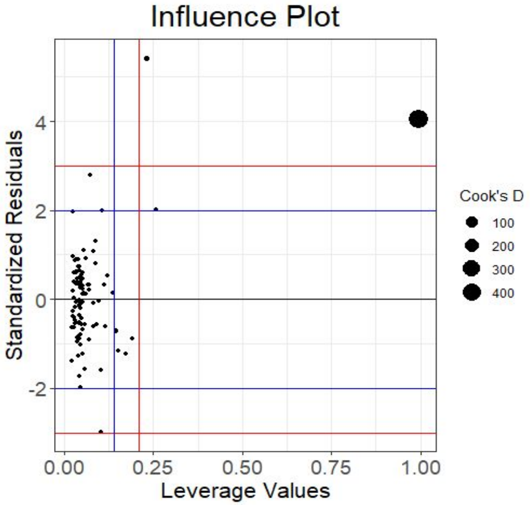
1. Are Age, Sq.Ft, StoryNew, Acres, and BathsNew predictors of Total? If so, is the relationship statistically reliable?

2. How does the relationship between SqFt and Total change depending on StoryNew?

Dataset

- Have a highly influential outlier
- Remove the outlier (YearBuilt = 2003)
- Run the model with sample size of 99





Descriptive Statistics (Quantitative Variables)

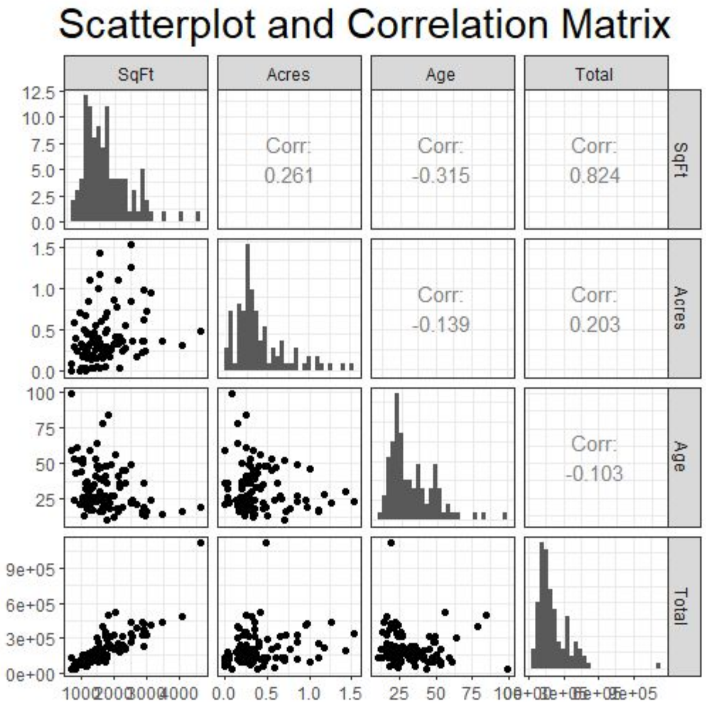
Type of Variable	Name of Variable	Center	Variability
Dependent Variable	Total	Median: $163,800	IQR: $141,758
Independent Variable	Age	Median: 27 years	IQR: 21 years
	SqFt	Median: 1589 sqft	IQR: 870 sqft
	Acres	Median: 0.3 acres	IQR: 0.295 acres

Descriptive Statistics (Categorical Variables)

Name of Variable	Categories	Counts	Proportions
StoryNew	1 Story	74	0.7475
	2 Stories	25	0.2525
BathsNew	1 Bathroom	25	0.2525
	2 Bathrooms	64	0.6465
	3 Bathrooms	10	0.1010

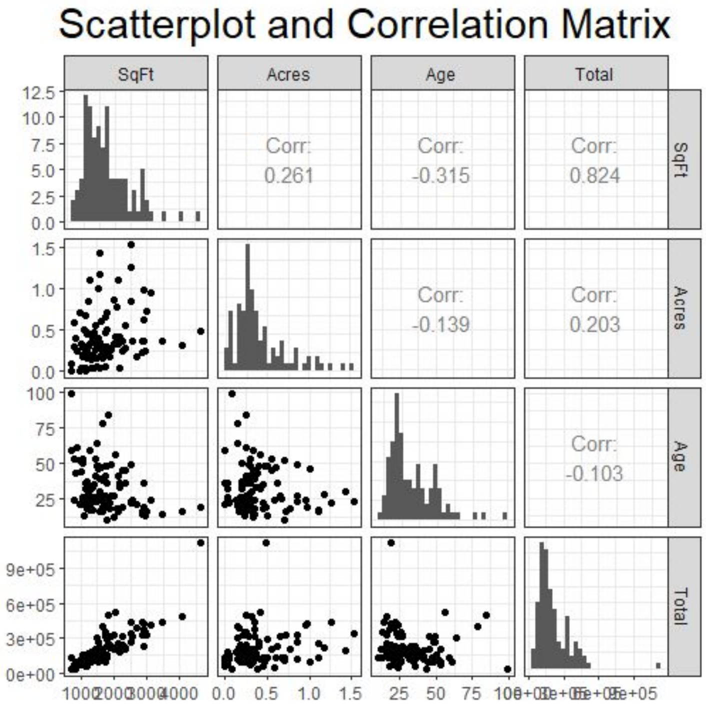
Scatterplot and Correlation Matrix

- **Total** and **SqFt**
  - a strong, positive, linear relationship
- **Total** and **Acres**
  - a weak, positive, linear relationship
- **Total** and **Age**
  - concave up relationship
- **Independent variables**
  - not highly correlated with each other



Scatterplot and Correlation Matrix

- Correlation between **Total** and **SqFt**
  - r = 0.824
  - Strong and positive
- Correlation between **Total** and **Acres**
  - r = 0.203
  - Weak and positive
- Correlation between **Total** and **Age**
  - r = -0.103
  - Weak and negative
- Correlations between **Independent variables**
  - Weak



Full Model

$$\begin{aligned}\widehat{Total} = -105617.72 - 1345.11(Age) + 31.62(Age^2) + 8063.32(Acres) + 181.74(SqFt)\\ + 10461.54(BathsNew2) + 19933.34(BathsNew3) + 47499.56(StoryNew2)\\ - 36.94(SqFt * StoryNew2)\end{aligned}$$

- Adjusted R-squared = 0.6984
- F-statistic = 29.37 on 8 and 90 DF, p-value = 2.2e-16

Full Model

- **Model Summary**

	Intercept	Age	Age^2	Acres	SqFt	BathsNew2	BathsNew3	StoryNew2	SqFt*StoryNew2
Estimate	-105617.72	-1345.11	31.62	8063.32	181.74	10461.54	19933.34	47499.56	-36.94
***p*-value**	0.0484	0.4783	0.1161	0.7801	< 2e-16	0.6484	0.6153	0.4369	0.1985
Confidence Interval	[-210469.0, -766.42]	[-5098.47, 2408.25]	[-7.97, 71.21]	[-49143.73, 65270.38]	[147.40, 216.08]	[-34965.98, 55889,06]	[-5859.67, 98463.36]	[-733345.22, 168344.33]	[-93.59, 19.71]

- **Significance of Categorical Variables (Type II Test)**

	F - statistic	p-value
BathsNew	0.1549	0.8567
StoryNew	1.5204	0.2208
SqFt*StoryNew	1.6780	0.1985

- Only variable SqFt is significant
- None of the categorical variables is significant
- Interaction term between SqFt and StoryNew is not significant

Reduced Model

- Both Forward Selection and Backward Elimination methods give the model

$$\widehat{Total} =- 116875.71 + 18.142(Age^2) + 175.840(SqFt)$$

- Adding the first order term of Age gives the model

$$\widehat{Total} =- 100370.47 - 776.91(Age) + 26.21(Age^2) + 174.71(SqFt)$$

with F-statistic = 77.84 on 3 and 95 DF, p-value = 2.2 e-16 and Adjusted R-squared=0.7017

Reduced Model

- **Partial F-Test**
  - Compare Full model and Reduced model using Partial F-test
  - F-test: 0.7922
  - P-value: 0.5581
  - Conclusion: We do not have sufficient evidence to conclude that the model with Acres, StoryNew, BathsNew and StoryNew*SqFt, Age, Age^2 and SqFt (Full model), is significantly better than the model with Age, Age^2 and SqFt (Reduced model) in predicting Total.

Reduced Model

- Assumptions for Reduced Model
    - **Violation of Constant Variance**
    - Levene’s Test: F-statistic = 22.235, p-value = 8.081 e-06 <0.0001
    - **Violation of Normality**
    - Shapiro-Wilk Test: W = 0.90114, p-value = 1.773e-06 < 0.0001

- Applied Log transformation to obtain Final Model

Final Model

$$\ln\left(\widehat{Total}\right) = 10.70 + 0.005943(Age) - 0.0000268(Age^2) + 0.000706(SqFt)$$

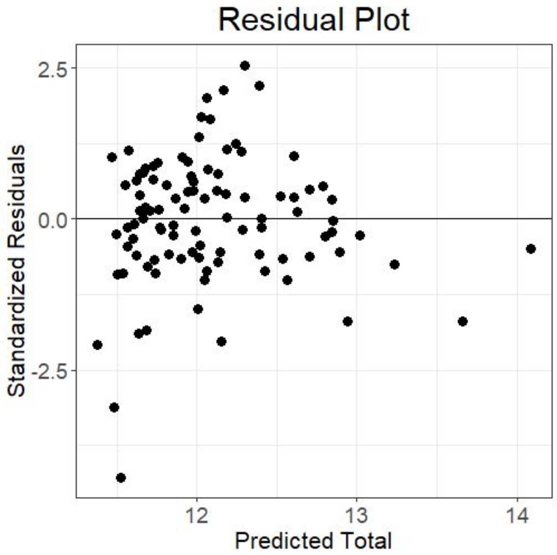
- **Description :**
  - F-statistic: 61.68 on 3 and 95 DF, p-value: < 2.2e-16
  - Adjusted R-squared: 0.6501

Final Model

	Intercept	Age	Age^2	SqFt
Estimate	10.70	0.0059	-0.0000268	0.000706
***p*-value**	< 2e-16	0.460	0.60573	< 2e-16
Confidence Interval	[10.2973, 11.1017]	[-0.0099, 0.2183]	[-0.0002, 0.00015]	[0.0006, 0.0008]

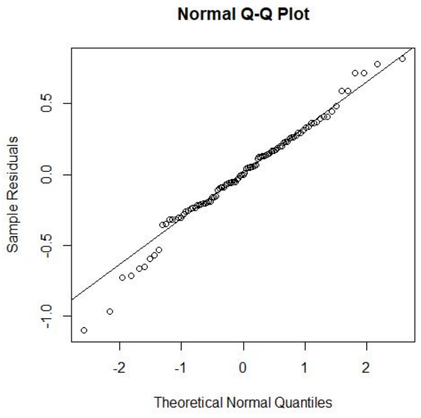
Assumptions for Final Model

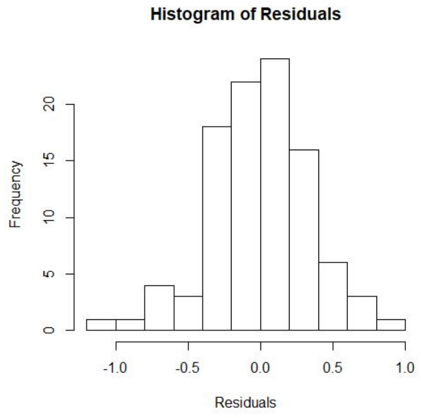
- **Independence** ✓
  - Randomly selected residential properties
- **Linearity** ✓
  - No curved pattern in residual plot
  - Several outliers at the bottom
- **Constant Variance** ✓
  - No distinct fan shape in residual plot
  - Levene’s Test: F statistic = 0.7128,   p-value= 0.4006



Assumptions for Final Model

- **Normality** ✓
  - Follows the straight line in QQ plot (but points outside the line)
  - Fair bell shape in histogram
  - Shapiro-Wilk Test: W=0.98096, p-value = 0.1625





Multicollinearity

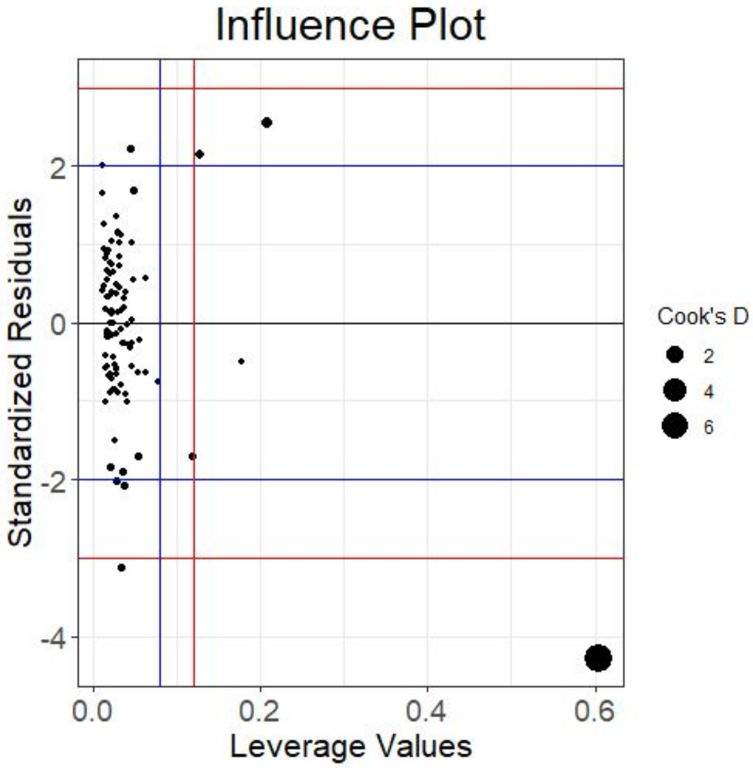
- **VIF**

	Age	Age^2	SqFt
VIF	13.1094	12.7115	1.1298

- Multicollinearity between Age and Age^2 is inevitable
- No issue with multicollinearity between other variables

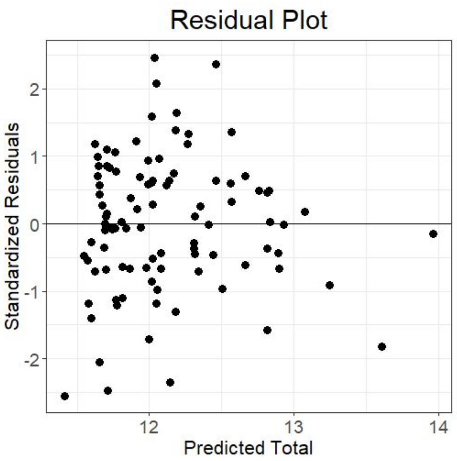
Influential Points

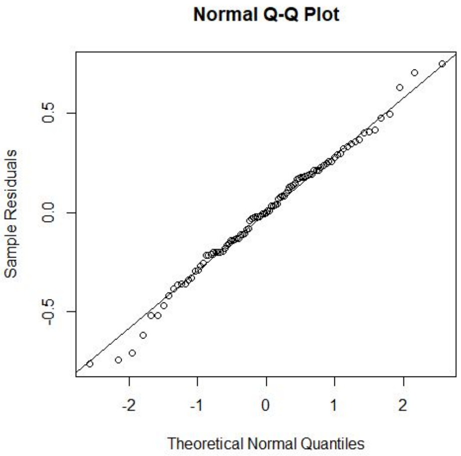
- **Standardized Residuals**
  - 6 somewhat unusual points (>2 or <-2)
  - 2 very unusual points (<-3)
- **Leverage Points**
  - 1 somewhat high leverage point (>0.081)
  - 4 very high leverage points (>0.121)
- **Cook's D**
  - Several somewhat and very influential points (>0.5 or >1)

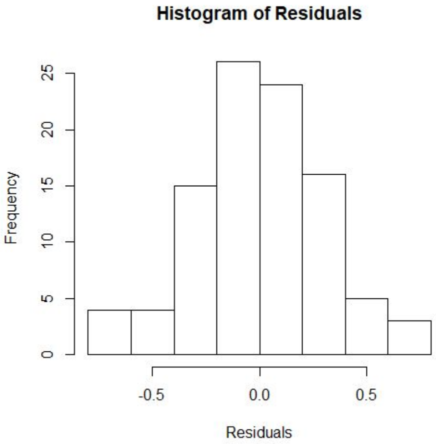


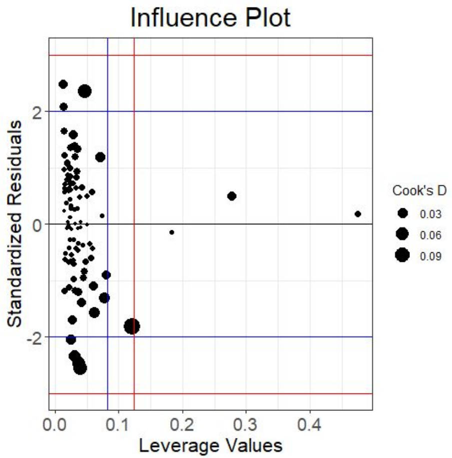
Removing Two Very Unusual Points

- **Adjusted R-squared**: 0.7081
- **F-statistic**: 78.62 on 3 and 93 DF, **p-value** = 2.2e-16< 0.0001
- **Levene's Test** (Constant Variance): p=0.5081
- **Shapiro-Wilk Test** (Normality): p=0.7478









Summary

- **Research Question 1**: With all variables, full model is significant. However, not all variables are good predictors of Total (only SqFt is significant).
- **Research Question 2**: There is no significant interaction between SqFt and StoryNew (the relationship between SqFt and Total does not depend on StoryNew).
- **Final model** including Age, Age^2 and SqFt is statistically significant in predicting log of Total.
- **Improvement:**
  - Datasets with more observations can contribute to development of a new model with higher Adjusted R-Squared.
  - Although it does not contribute to Adjusted R-squared, Acres can be added to the model to explain the relationship between the plot and the Total.