DATA EXPLORATION AND ANALYSIS

CASE STUDY

AN INVESTIGATION OF FACTORS AFFECTING

HOUSE PRICES IN SARATOGA, NY

Contents

LIST OF FIGURES

LIST OF TABLES

1.0 INTRODUCTION

1.1 Background

Population: This report examines house price data for a random sample of 1734 homes from Saratoga County, New York. The dataset is sourced from Dasl with 16 features relating to house sales.

Response Variable: Sale price in dollars

Explanatory Variables:

Table 1. List of Possible Explanatory Variables

Variable	Measurement
Lot Size	Acres
Waterfront Location	Y/N
Age	Years
Land Value	Dollars
New Construction	Y/N
Central Air	Y/N
Fuel Type	Gas, Electric etc.
Heat Type	Hot Air, Electric etc.
Sewer Type	Public, private
Living Area	Square Feet
Pct. College Educated	%
Bedrooms, bathrooms, rooms	Number
Fireplaces	Number

1.2 Determining the Research Question

It is expected that the house price is related to one or more of the other variables listed in Table 1 and that these relationships will be:

- linear or non-linear,
- positive or negative,
- different strengths.

It is anticipated that land value and living area will have a relatively strong linear relationship with price. There may be negative linear some collinear relationships, since several of the explanatory variables are related.

Research Question:

Is there a linear relationship between any of the explanatory variables and house prices?

Hypothesis:

There is a linear relationship between one or more of the response variables and house prices.

Code, analysis and detailed commentary are in the Colab Notebook (Appendix 1).

2.0 PROCEDURE AND STATISTICAL METHODS

Descriptive statistics and plots were used to identify patterns, trends and relationships and Pearson's Correlation Coefficient calculated to quantify the direction and strength of relationships.

Binary and multi-level categorical features were converted into dummy variables and one variable dropped from each category to avoid the 'dummy trap'. A base multiple linear regression model was constructed and Statsmodels OLS library used to run the regression and produce summary results. The base model was a linear function with unknown parameters B_0, B_1 etc, the response variable Y_i and predictor variables x_1, x_2 etc of the form:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$$

The coefficient of determination (r-squared) and adjusted r-squared, were examined to determine the proportion of the variance in the response that could be predicted from the independent variables. The F-statistic and associated p-value at 5% significance identified whether all variables included had improved the model, against a null hypothesis that all the regression coefficients were equal to zero (an intercept only model).

Model coefficients were inspected including the value, sign and effect, i.e. how much the response is expected to increase when that variable increases by one unit, holding all the other independent variables constant. The significance of each coefficient was tested against the null hypothesis that the true coefficient(β) was zero by reference to the associated t-test and p-values at the 5% level of significance.

Multicollinearity was assessed through Variance Inflation Factors and the effect of dropping one of the collinear variables was investigated. Summary, individual residual plots and statistical tests were used to determine whether the linear regression assumptions of linearity, independence, normality and homoscedasticity had been met. Cook's Distance plots and z-scores identified outliers, which were then reviewed.

Non-significant variables were dropped from the model using a back wise step selection process based upon the Akaike Information Criterion (AIC). This method balances goodness of fit with minimum complexity, helping to reduce the likelihood of over or under-fitting.

Model improvement included examining interaction terms and one extra term was added to the preferred model. A further term was tested to address an apparent non-linear pattern in the residual plot for one predictor variable.

Heteroscedasticity identified from the Breusch-pagan test suggested the need for transformation. A log transformation on the two 'best' models allowed the respective effect on performance to be compared. Taking the natural log of the response variable a log-linear model were produced of the form:

$$\log(y_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + e_i,$$

where B1 is the expected change in log Y from a one unit increase in X1, assuming all other variables remain constant. The effect on the response was calculated from the exponentiated regression coefficients as a reverse log:

$$\exp(\beta)$$

Comparing the log-linear models using summary statistics, the preferred model was identified and fitted. The log-linear and preferred linear model were then compared. As direct comparison cannot be made between a linear and log model due to different scales, the predicted values for the response of the log-linear model were back-transformed and the r-squared and AIC recalculated.

A simple predictor tool was used to determine the confidence intervals for the mean response and prediction intervals for new observations. The results of the multiple linear regression with respect to the initial research question were discussed and conclusions made.

3.0 DESCRIPTIVE STATISTICS

3.1 Summary Statistics

Table 2. summarises basic statistics for numerical variables.

Table 2. Saratoga Study – Basic Descriptive Statistics - Numerical Variables

	Price	Lot Size	Age	Land Value	Living Area	% College	Beds	Baths	Fires	Rooms
Mean	211,545	0.5	28	34,536	1,753	56	3	2	0.6	7
Median	189,700	0.4	19	25,000	1,632	57	3	2	1	7
Mode	n/a	0.5	19	27,000	1,480	64	3	2.5	1	7
Min	5,000	*0.0	0	200	616	20	1	0	0	2
Max	775,000	12.2	225	412,600	5,228	82	7	4.5	4	12
Std	98,554	0.7	30	34,981	620	10	1	0.7	0.6	2.3

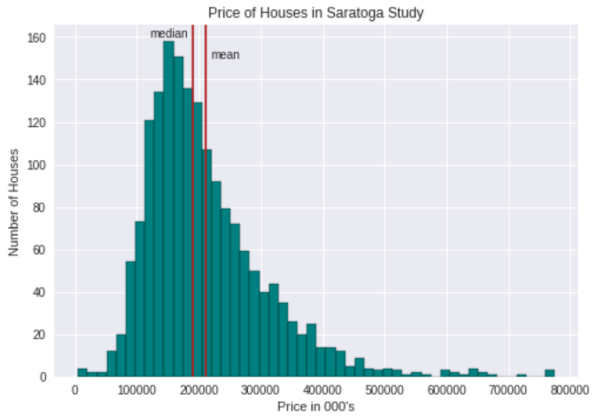
Note: Half bathrooms are toilet and pedestal
**Assumed to be flats*

- The mean house price of $211,500 is above the median, indicating positive skew.
- Most properties are under 20 years old, with the oldest being 225 years old and others under a year, suggesting a wide variety in age of homes.
- Most properties have 3 bedrooms and 2.5 bathrooms and total rooms range from 7 to 12.

3.2 Exploratory and Comparison Plots for All Variables

3.2.1 House prices

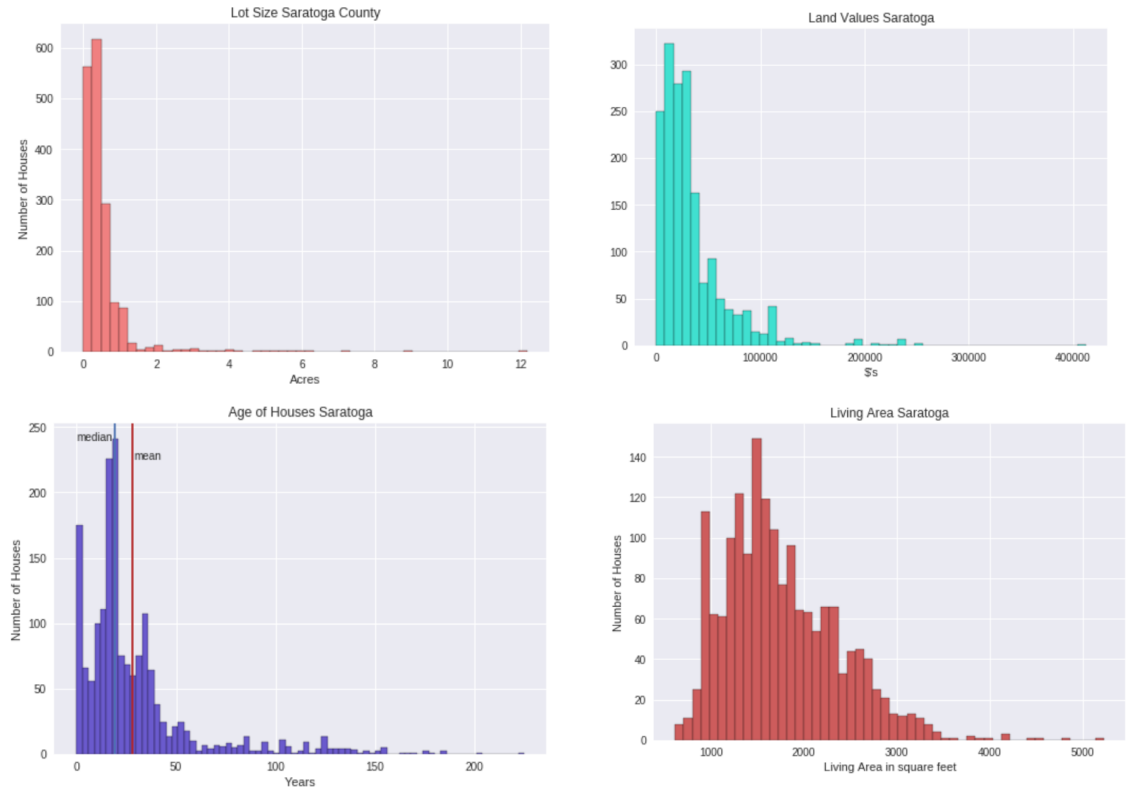
Figure 3.1 House Prices Saratoga Study



Prices are positively skewed.

3.2.2 Lot Size, Land Values, House Age and Living Area

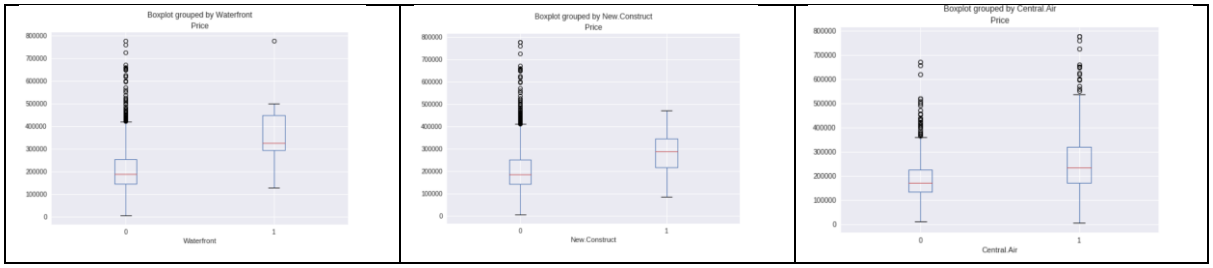
Fig. 3.2 Distribution Plots for Lot Size, Land Values, Age of Houses and Living Area



All plots are positively skewed.

3.3.3 New Constructions, Waterfront Location and Central Air Conditioning

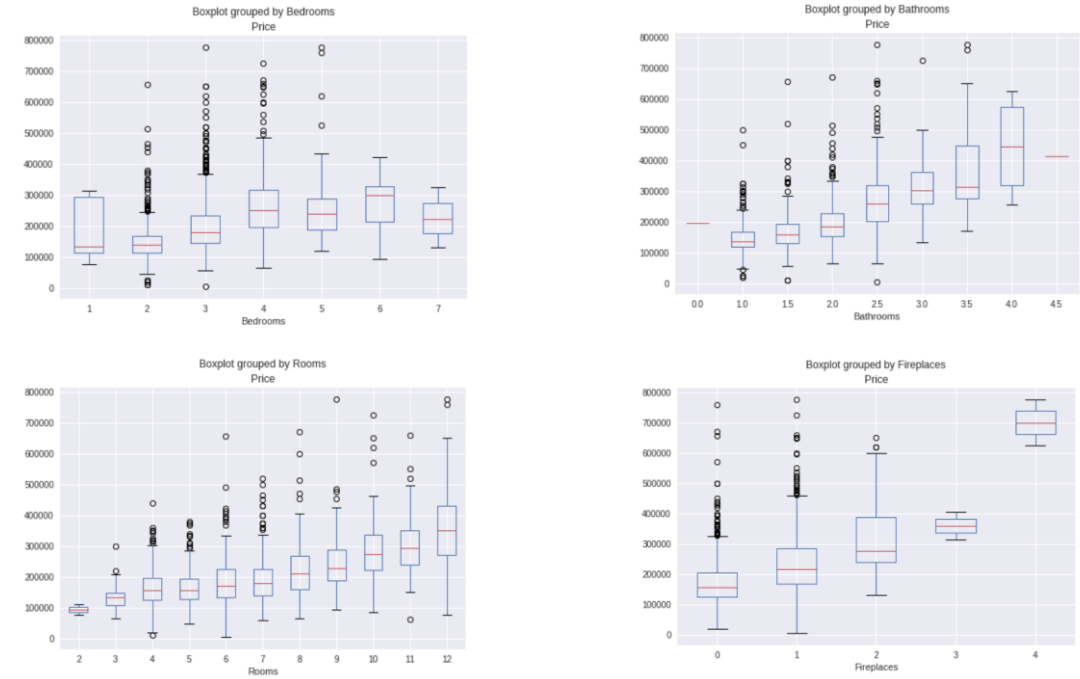
Fig. 3.3 Construction Type and Location



The median price is higher where each binary variable is present.

3.3.4 Bedrooms, Bathrooms, Rooms and Fireplaces

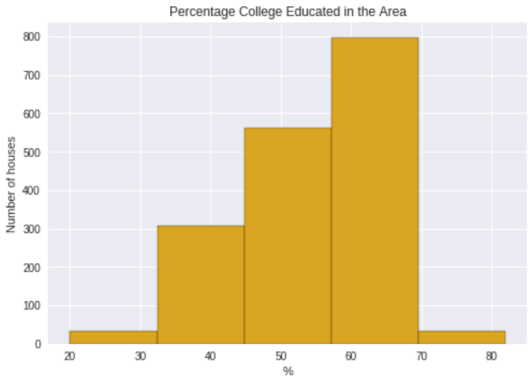
Fig 3.4 Bedrooms, Bathrooms, Total Rooms and Fireplaces



Median price generally increases with number of rooms, except beyond 4 bedrooms. There is increasing variance in the bathrooms plot and a jump up in price above 4 fireplaces.

3.3.5 Percent College Educated

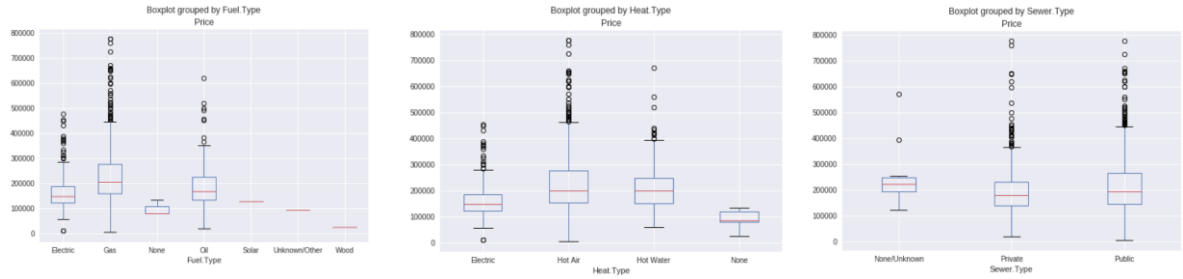
Fig. 3.5 College Educated



The histogram is negatively skewed and few areas below 30% or above 70%, suggesting it might be a middle-class area.

3.3.6 Utilities

Fig. 3.6 Fuel, Heat and Sewer Type



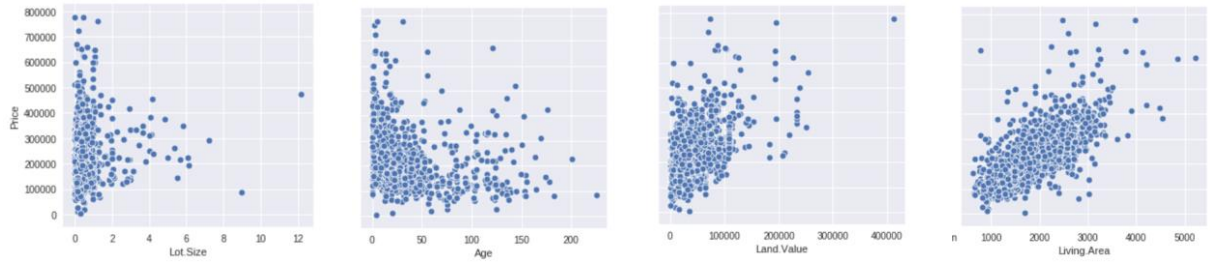
There are no significant observations for these categories.

4.0 CORRELATION

4.1 Scatter Plots

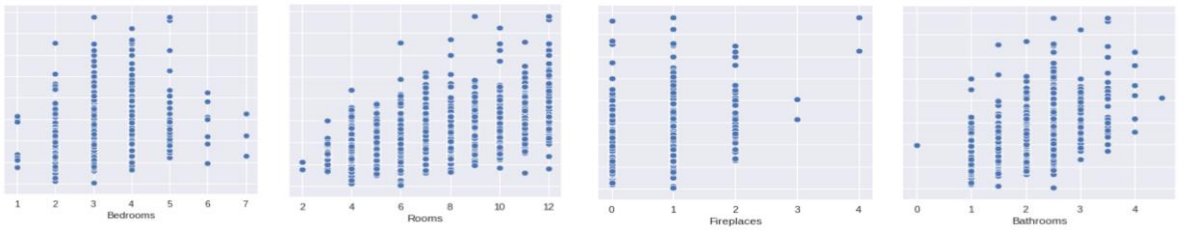
The strength and presence of relationships between house price and explanatory variables were examined.

Fig. 3.7 Plots of house prices and continuous variables



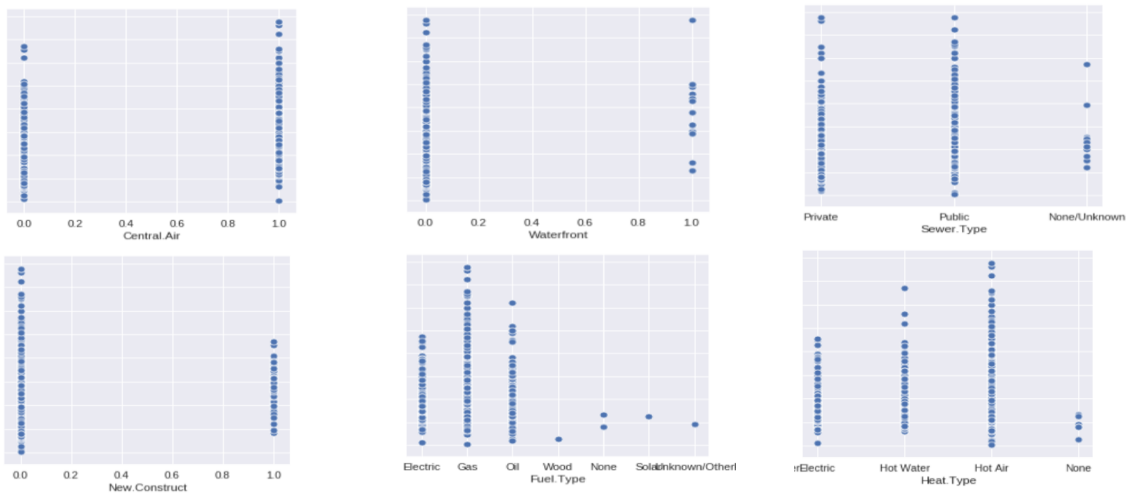
Land value and living area appear quite strongly positively linearly related to price. Age is negatively related and there may be some curvature indicating a non-linear relationship.

Fig 3.8 Plots of house prices and discrete numerical variables



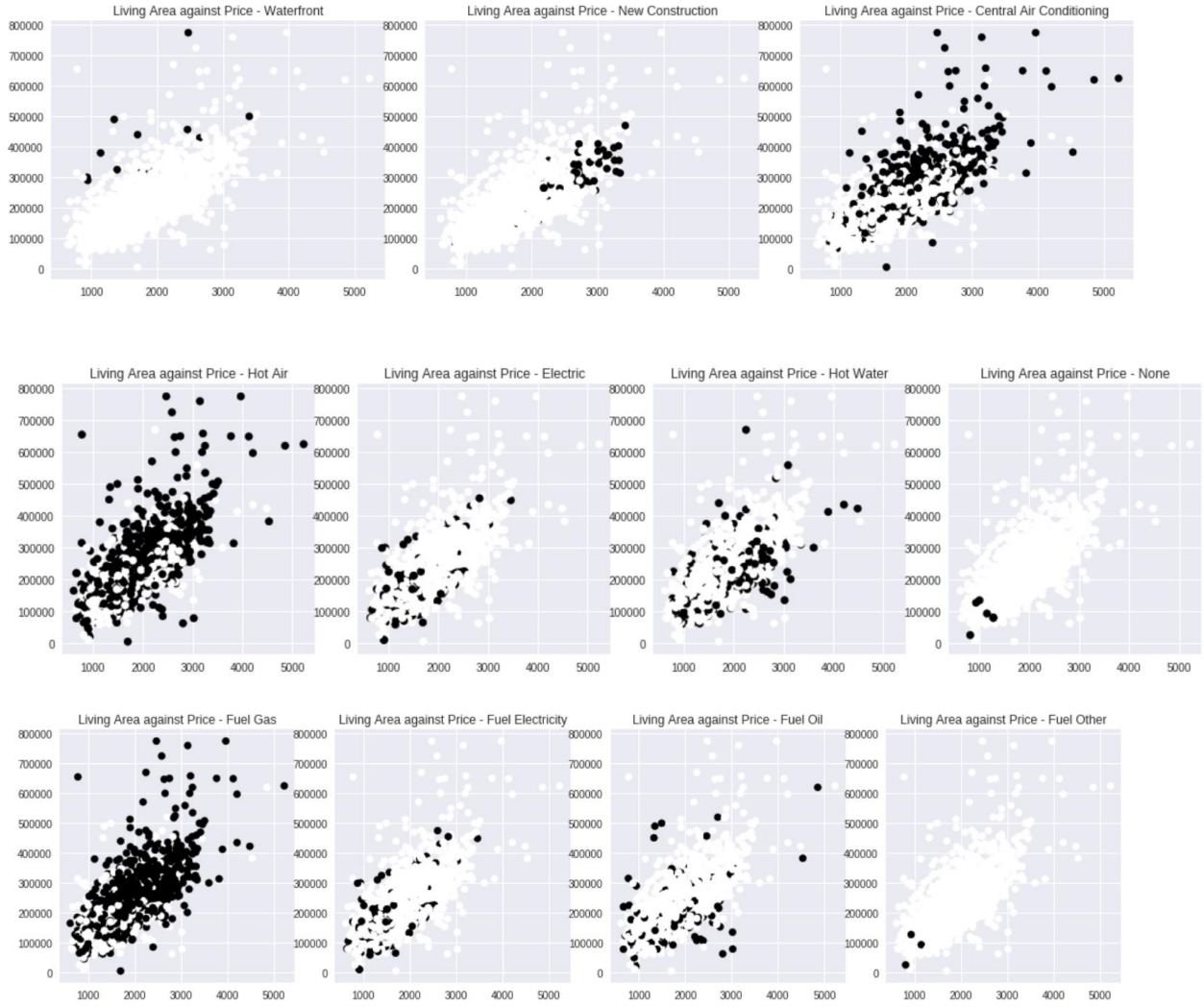
Relationships between price, rooms and bathrooms appear positive with some curvature in the bedrooms plot suggesting differing relationships with price.

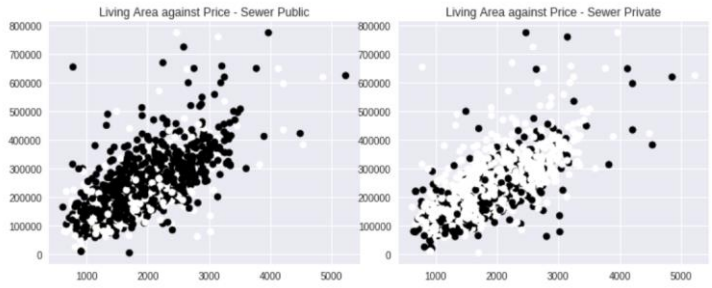
Fig. 3.9 Plots of house prices and categorical variables



Plotting price against a continuous variable such as living area and then using colour for the categorical variable highlights some interesting trends.

Fig. 3.10 House Prices and Living Area with Categorical Variables Overlaid





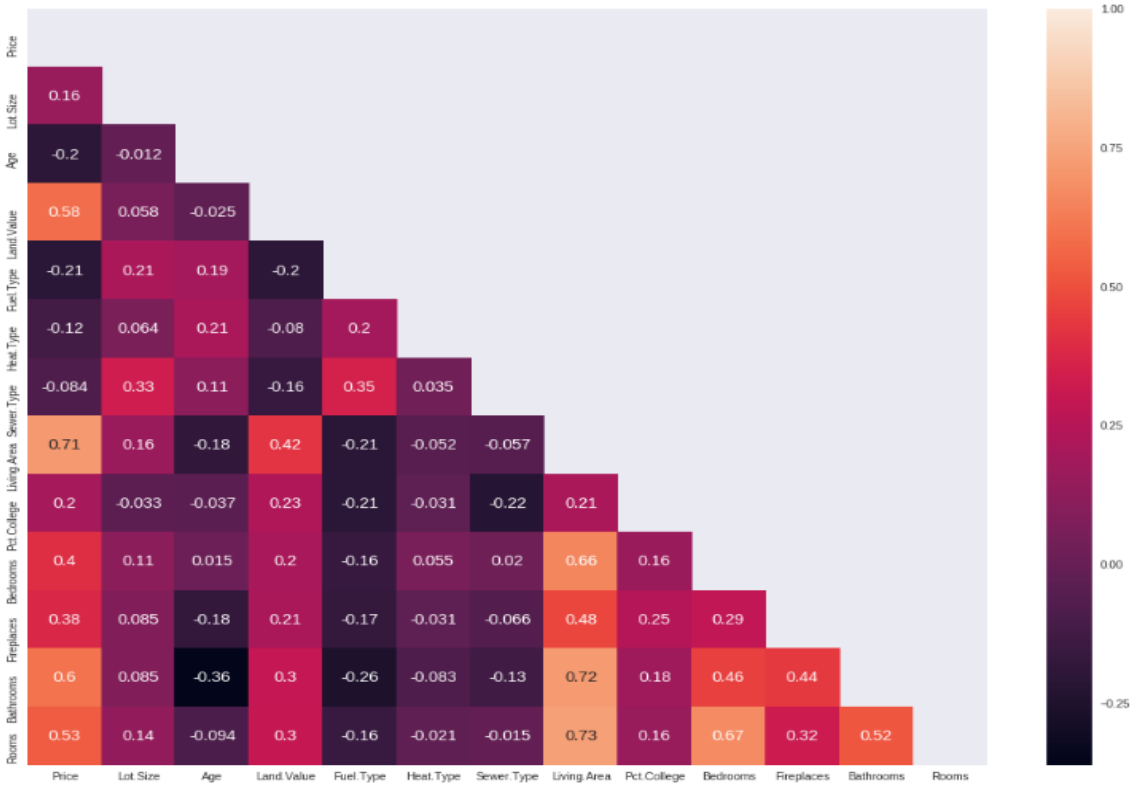
Observations:

- Waterfront properties are generally priced higher.
- New construction houses are usually larger than 2000 square feet
- Larger, expensive homes generally have central air conditioning.
- Electric heating is mostly in lower priced smaller homes
- Higher priced larger homes having mainly hot air heating
- Gas is the most used fuel type at all price levels and electricity and oil found in smaller lower price homes

This indicates that some categorical variables are likely to have an influence on the price of the property.

4.2 Correlation Coefficients

Fig 3.11 Pearson's Correlation Coefficients Heatmap



- There are strong positive relationships between price and land value, living area, bathrooms and rooms and negative relationships with age, fuel and heat type.
- There are correlations amongst some of the explanatory variables suggesting multicollinearity.

5.0 BUILDING A LINEAR MODEL

5.1 Ordinary Least Squares Regression Base Model

Fig. 4.1 OLS Base Model Results

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.655
Model:	OLS	Adj. R-squared:	0.651
Method:	Least Squares	F-statistic:	162.8
Date:	Thu, 09 Jan 2020	Prob (F-statistic):	0.00
Time:	12:16:48	Log-Likelihood:	-21475.
No. Observations:	1734	AIC:	4.299e+04
Df Residuals:	1713	BIC:	4.311e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.396e+04	1.03e+04	2.317	0.021	3680.876	4.42e+04
Lot.Size	7585.0362	2238.119	3.389	0.001	3195.302	1.2e+04
Age	-132.5030	58.055	-2.282	0.023	-246.369	-18.637
Land.Value	0.9231	0.047	19.439	0.000	0.830	1.016
Living.Area	69.9994	4.608	15.191	0.000	60.962	79.037
Pct.College	-111.9048	151.242	-0.740	0.459	-408.542	184.733
Bedrooms	-7868.6859	2563.082	-3.070	0.002	-1.29e+04	-2841.585
Fireplaces	1052.8200	2982.800	0.353	0.724	-4797.494	6903.133
Bathrooms	2.301e+04	3360.707	6.846	0.000	1.64e+04	2.96e+04
Rooms	3024.7569	960.974	3.148	0.002	1139.951	4909.562
Waterfront	1.201e+05	1.55e+04	7.735	0.000	8.96e+04	1.51e+05
New.Construct	-4.551e+04	7299.618	-6.235	0.000	-5.98e+04	-3.12e+04
Central.Air	9924.6806	3474.222	2.857	0.004	3110.516	1.67e+04
Heat_electric	-107.6183	1.23e+04	-0.009	0.993	-2.42e+04	2.4e+04
Heat_hotwater	-1.042e+04	4185.085	-2.491	0.013	-1.86e+04	-2214.549
Heat_none	-2.972e+04	1.65e+04	-1.806	0.071	-6.2e+04	2547.836
Fuel_elec	-1.093e+04	1.21e+04	-0.901	0.368	-3.47e+04	1.29e+04
Fuel_oil	-4248.3564	5006.189	-0.849	0.396	-1.41e+04	5570.531
Fuel_none	-1.298e+04	2.53e+04	-0.513	0.608	-6.26e+04	3.67e+04
Fuel_other	-1.673e+04	2.51e+04	-0.667	0.505	-6.6e+04	3.25e+04
Sewer_priv	1323.0310	3656.863	0.362	0.718	-5849.357	8495.419
Sewer_none	-3407.8757	1.71e+04	-0.200	0.842	-3.69e+04	3e+04

Omnibus:	602.197	Durbin-Watson:	1.659
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4883.746
Skew:	1.398	Prob(JB):	0.00
Kurtosis:	10.731	Cond. No.	3.25e+21

Predictor variables explain 66% of the variation in house prices. Equation for the model:

y = 23,960 + 7,585(Lot Size) – 132.5(Age) + 0.92(Land Value) + 70(Living Area) -112(Pct.College) – 7,867(Bedrooms) + 1,053(Fireplaces) + 23,010(Bathrooms) +3,025(Rooms) + 120,100(Watefront) – 45,510(New Construct) + 9,925(Central Air) – 108(Heat_electric) – 10,420(Heat_hotwater) – 29,720(Heat_none) – 10,930(Fuel_elec) – 4,248(Fuel_oil) – 12,980(Fuel_none) – 16,730(Fuel_other) + 1,323(Sewer_private) - 3,408(Sewer_none)

For each square foot of living area, price increases by $70 on average - ceteris paribus.

With a large F-statistic and associated small p-value the alternative hypothesis is accepted, the model with all the independent variables fits the data better than one based solely on the intercept.

Table 3. Base Model – Coefficients Sorted by Significance (5%)

	unit	coef	p-values
New.Construct	Y/N	-45513.040043	0.000
Waterfront	Y/N	120099.910773	0.000
Land.Value	$	0.923139	0.000
Living.Area	sq ft	69.999377	0.000
Bathrooms	No.	23007.766521	0.000
Lot.Size	Acres	7585.036217	0.001
Rooms	No.	3024.756920	0.002
Bedrooms	No.	-7868.685865	0.002
Central.Air	Y/N	9924.680632	0.004
Heat_hotwater	Y/N	-10422.964814	0.013
Age	yrs	-132.502973	0.023
Heat_none	Y/N	-29717.748562	0.071
Fuel_elec	Y/N	-10926.252047	0.368
Fuel_oil	Y/N	-4248.356357	0.396
Pct.College	%	-111.904758	0.459
Fuel_other	Y/N	-16733.069900	0.505
Fuel_none	Y/N	-12984.678663	0.608
Sewer_priv	Y/N	1323.031005	0.718
Fireplaces	No.	1052.819961	0.724
Sewer_none	Y/N	-3407.875683	0.842
Heat_electric	Y/N	-107.618301	0.993

Significant coefficients < 0.05

5.2 Multicollinearity

Multicollinearity measured by the Variance Inflation Factor (VIF) is shown in Table 4. Values of 1 are not correlated, 1 to 5 moderately and above 5, highly correlated. Collinearity amongst dummy variables collinearity is ignored, (Statisticalhorizons.com, 2019).

Table 4. Variance Inflation Factors

	VIF Factor	features
18	inf	Fuel_none
15	inf	Heat_none
19	inf	Fuel_other
13	11.239626	Heat_electric
16	11.183578	Fuel_elec
4	4.178358	Living.Area
9	2.538467	Rooms
8	2.504243	Bathrooms
6	2.247742	Bedrooms
2	1.537379	Age
12	1.433950	Central.Air
20	1.416177	Sewer_priv
3	1.411678	Land.Value
7	1.408329	Fireplaces
17	1.398903	Fuel_oil
14	1.289455	Heat_hotwater
1	1.249173	Lot.Size
5	1.246514	Pct.College
11	1.214523	New.Construct
10	1.058124	Waterfront
21	1.022738	Sewer_none

6.0 MODEL EVALUATION

6.1 Residual Analysis

Fig. 5.1 Residuals Against Predicted Values Plots

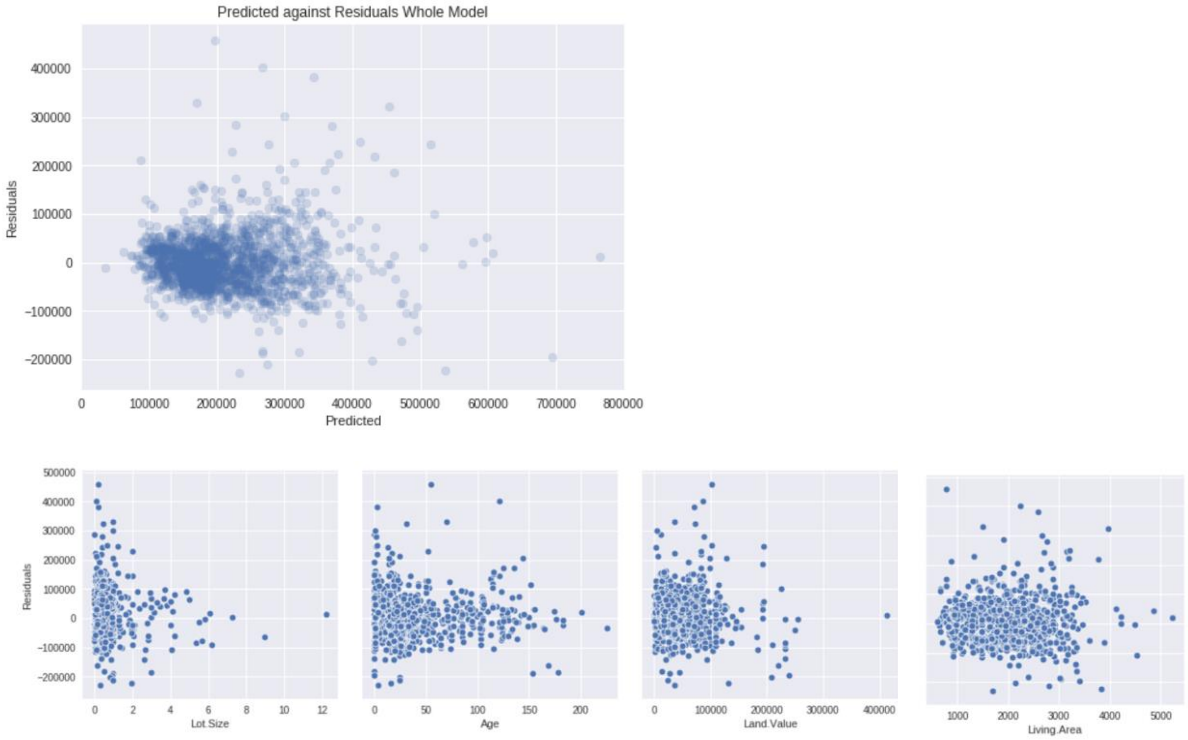


Fig. 5.2 Residual Plots for Normality

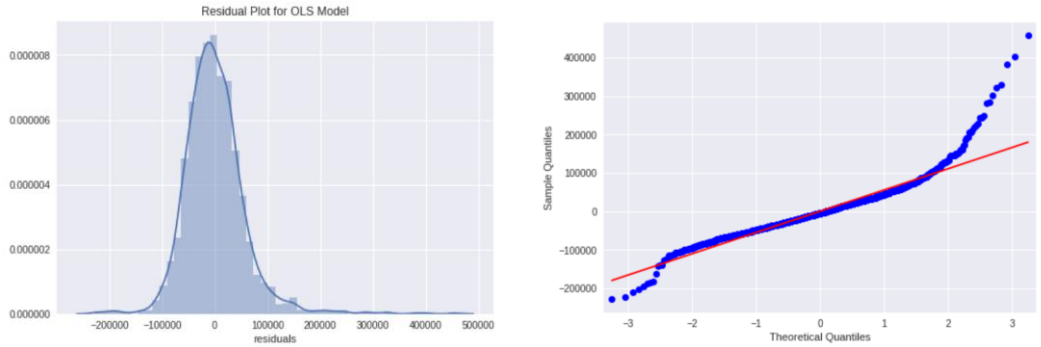


Table 5. Statistical Tests for Residual Normality

Statistical Tests	Test-stat	p-value	Comment
Skew	1.4	na	Positive skew
Shapiro Wilks	0.912	0.0	Reject Null Hyp. Residuals not normally distributed
Jarque-Bera	4884	0.0	Reject Null Hyp. Residuals not normally distributed
Anderson Darling	20.8	0.0	Reject Null Hyp. Residuals not normally distributed

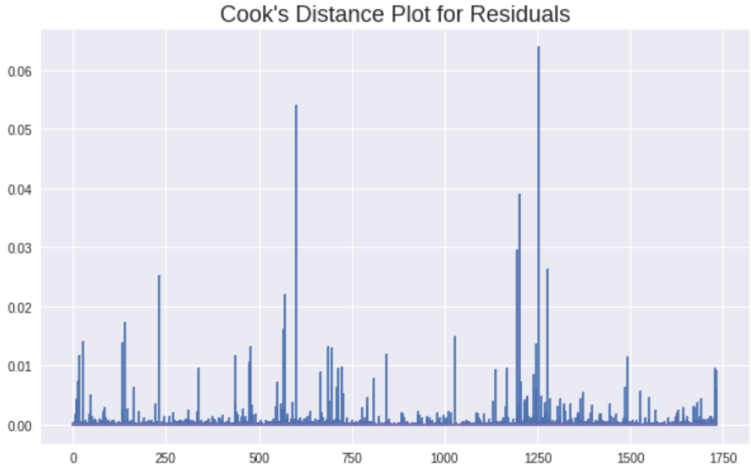
Table 6. Statistical Tests for Residual Homeoscasdicity

Statistical Tests	Test-stat	p-value	Comment
Breusch pagan	121	0.0	Reject Null Hyp. Non-constant variance.

There is no trend or curvature except maybe for 'Age'. The histogram has positive skew and the qq plot tails are drifting. Statistical tests reject the null hypothesis of a normal distribution. Non-normality can affect significance of the coefficients and the confidence intervals but with large sample sizes, the linear model can be robust due to the central limit theorem. However, the residuals do not exhibit homoscedasticity, which can affect the reliability of the coefficients and accuracy of confidence intervals.

6.2 Outliers

Fig. 5.3 Cook's Distance Plot



Cook's distance plot shows several influential outliers and 26 values were above 3 standard deviations from the mean (99.7% of values should lie within). Investigation of datapoints showed these were large properties with living area well above the mean. Lack of locational information in the data, does not allow further analysis and without domain knowledge, so these points were not removed.

7.0 IMPROVING THE BASE MODEL

7.1 Back-wise Selection

Removing one variable at a time and rerunning the function to obtain the result with the lowest AIC, 8 of the non-significant variables were dropped. The new model was 'fitted' and R_2 increased slightly indicating that removing the variables did not reduce the performance of the model.

Table 7. Model 2 Comparison with Base

	Description	R2	Adj. R2	Non-sig var	F-Stat(p-value)
Base	Base	0.655	0.651	9	0.00
Model 2	Improved	0.655	0.652	1	0.00

Table 8. Model 2 predictors

	Adj R-sqr	AIC
Fuel_other	0.651925	42980.2
Heat_hotwater	0.651037	42984.6
Fuel_elec	0.650904	42985.3
Age	0.650704	42986.3
Central.Air	0.650584	42986.8
Bedrooms	0.650393	42987.8
Rooms	0.650225	42988.6
Lot.Size	0.649609	42991.7
New.Construct	0.64425	43018
Bathrooms	0.642479	43026.6
Waterfront	0.639749	43039.8
Living.Area	0.602986	43208.3
Land.Value	0.571795	43339.4

The equation for Model 2:

House Price = 18,080 + 7,397(Lot Size) - 155(Age) + 0.92(Land Value) + 70(Living Area) - 7,744(Bedrooms) + 23,010(Bathrooms) + 3,044(Rooms) + 120,500(Waterfront) - 44,910(New Construction) + 9,740(Central Air) - 10,210(Hot Water) - 10,280(Fuel Electric) - 43,920(Fuel Other)

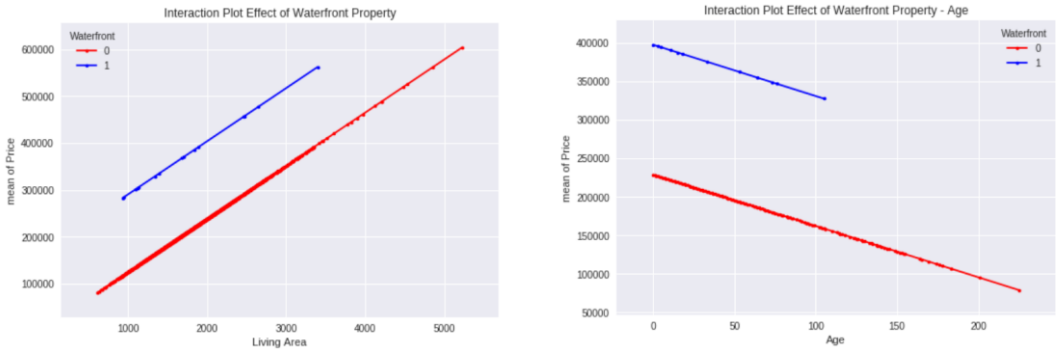
- Extra bathroom - $23,010 increase in price.
- Waterfront location - $120,500 more than a non-waterfront location.
- Electric Fuel - $10,280 lower than a home with gas ceteris paribus.

7.2 Interaction Terms

A range of interaction terms were added to the model. Waterfront homes were investigated:

- Waterfront homes are more expensive but do smaller homes in waterfront locations still command higher prices?
- Older properties are valued lower but do age and waterfront location interact so older waterfront homes are still expensive?

Fig. 6.1 Waterfront properties against living area and age of property



$$y = 113(LivingArea) + 165,600(Waterfront)$$
$$y = 119(LivingArea)$$

An interaction term 'waterfront x living area' was significant but 'waterfront x age' was not. Other interactions were investigated and summarised in Appendix 2.

Table 9. Model 3 Comparison

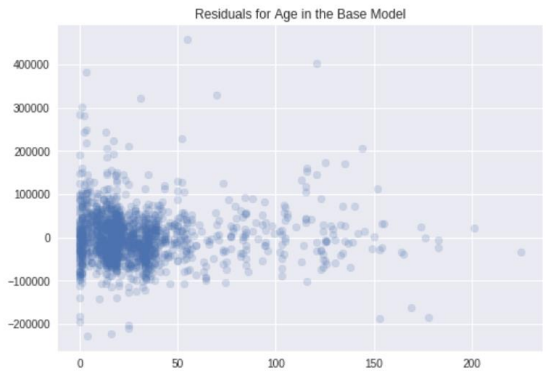
	Description	R2	Adj. R2	Non-sig. Coeff.	F-Stat(p-value)
Model 1	Base	0.655	0.651	9	0.00
Model 2	Improved	0.655	0.652	1	0.00
Model 3	Interaction	0.656	0.653	1	0.00

Adding an extra term only improved the model slightly this should be balanced against increasing complexity and the risk of overfitting.

7.3 Transformations

7.3.1 Independent Variable - Age

Fig. 6.2 Residuals



The residuals plot for Age appeared non-linear. Various transformations (zero values preclude using reciprocals and logs) were tried with a root transformation increasing R^2 slightly.

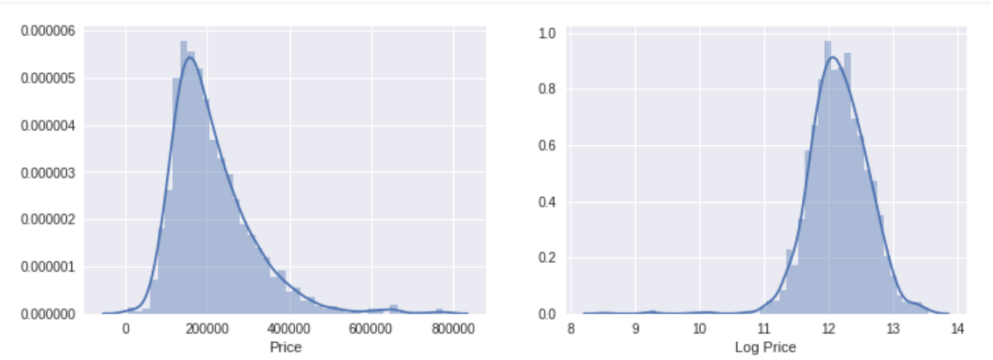
Table 10. Model 4 Comparison

	Description	R^2	Adj. R^2	Non-sig. Coeff.	F-Stat(p-value)
Model 1	Base	0.655	0.651	9	0.00
Model 2	Improved	0.655	0.652	1	0.00
Model 3	Interaction	0.656	0.653	1	0.00
Model 4	Transform Age	0.658	0.655	2	0.00

7.3.2 Response Variable - Price

To stabilise the variance and reduce skew, the response variable was log transformed.

Fig. 6.3 Log Transformation of the Response Distribution Plots



Regressions were re-run for the two 'best' models.

Table 11. Comparison of Log-linear(semi-log) models

	R^2	Adj. R^2	Non-sig. Coeff.	F-Stat(p-value)	AIC
Model 3 Transform	0.589	0.585	3	0.00	686.2
Model 4 Transform	0.589	0.586	*5	0.00	686.6

**The root Age and Age variables in Model 4 were non-significant. Log Model 3 is the preferred model.*

7.3.3 The Log-Linear Model

Following back wise selection, one further variable was dropped, and the final model fitted.

Table 12. Log Price Regression Coefficients

	unit	coef	p-values	% chnge
Lot.Size	Acres	0.036949	0.000	3.76
Age	yrs	-0.001590	0.000	-0.16
Land.Value	$	0.000004	0.000	0.00
Living.Area	sq ft	0.000281	0.000	0.03
Bathrooms	No.	0.104337	0.000	11.00
Waterfront	Y/N	1.139563	0.000	212.54
New.Construct	Y/N	-0.177109	0.000	-16.23
WaterLiving	Y/N	-0.000393	0.000	-0.04
Fuel_elec	Y/N	-0.067789	0.001	-6.55
Fuel_other	Y/N	-0.594193	0.001	-44.80
Central.Air	Y/N	0.050202	0.002	5.15
Rooms	No.	0.007931	0.102	0.80
Bedrooms	No.	0.016019	0.211	1.61

Equation:

Log Price = 0.036949(Lot Size) – 0.0016(Age) + 0.000004(Land Value) + 0.000281(Living Area) + 0.104377(Bathrooms) + 1.139563(Waterfront) – 0.17719(New Construction) – 0.00393(WaterLiving) - 0.067789(Fuel_elec) -0.594193(Fuel_other) + 0.050202(Central Air) + 0.007931(Rooms) + 0.016019(Bedrooms)

Coefficients are interpreted differently from the linear model. The exponent of the coefficient is used to calculate the percentage change for one unit increase in house prices. E.g. each acre increase of lot size increases the untransformed house price by a multiple of e0.0376 = 1.0376 or a 3.7% increase on average. Note: under back transformation the mean is the geometric mean rather than the arithmetic. Of interest is that bedrooms now have a positive coefficient compared to the linear model.

Table 13. Comparison of Log linear Models

	R_2	Adj. R_2	Non-sig. Coeff.	F-Stat(p-value)	AIC
Model 3 Transform	0.589	0.585	3	0.00	686.2
Model 4 Transform	0.589	0.586	5	0.00	686.6
Model 5	0.589	0.585	2	0.00	685.2

Fig. 6.4 Residuals Plot for Model 5

Table 14. Statistical Tests for Model 5

Statistical Test	Comment
Shapiro-Wilks Normality	Reject Null Hyp. Residuals not normally distributed
Breusch-Pagan	Fail to reject null hypothesis. Constant variance

Log transformation removed variance in the residuals.

8.0 LINEAR VERSUS LOG MODELS

The log-linear model cannot be compared to a non-log model by comparing R_2 values or AIC since the scales are different. The predicted Y values were back-transformed and the metrics recalculated.

Table 15. Model Comparison – Model 5 and Model 3

	Best Log-Linear Model	Best Linear Model
Adjusted R2	0.57	0.65
Adjusted AIC	51,349	42,977

Although the linear model appears better, Model 5 removed the variance in the residuals. With non-constant variance coefficients are unbiased but errors may be unreliable resulting in smaller p-values, so some variables will appear incorrectly significant. Confidence intervals may also be too affected. Therefore, the log transformed model is preferred:

9.0 HOUSE PRICE PREDICTION TOOL

A simple prediction tool was built for values for explanatory variables within the value range covered by the data. This returns the 95% back-transformed confidence and prediction intervals for the mean response. The exponential of the log data is used to produce estimates on the original scale. Table 17 uses the first datapoint in the dataset.

*Table 16 Back transformed confidence and Prediction Intervals Using Prediction Tool And Sample Datapoint**

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	74649.572742	1.186436	53383.690689	104386.913655	38335.296643	145363.651738

The actual price of $132,500 is well above the model mean response and outside the confidence interval. Confidence intervals are for the geometric mean so are narrower and unsymmetrical. Comparing this to the best linear model:

Table 17 Confidence and Prediction Intervals Using Adapted Tool and the Linear Model 3

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	101210.960878	33817.723225	34882.739533	167539.182224	-30585.317062	233007.238819

The untransformed model appears to perform better but further analysis is required with test and training data.

10.0 MODEL EVALUATION

House price models are hedonic, relying on internal and external factors. This model explains only 59% of the variation in price, with 41% explainable by other variables. Location is missing and excluding important variables can result in model misspecification and significant omitted variables bias. Inclusion of geo-location or zip-codes would likely improve performance.

The log transformed model was chosen over the linear model partly due to failure of the Breusch-pagan test. However, this test can be unreliable where data is non-normal and where there is misspecification of the model. The linear model may be therefore be preferable for estimating, even with the effect of wider confidence intervals. The use of test and training data sets would help to refine this position.

11.0 CONCLUSIONS

The final multiple regression model accounts for 59% of the variation in house price and thirteen predictors. Coefficients of the model express the relationship between the explanatory variable and the response variable assuming all other variables remain constant, but this does not imply a causal relationship. As the response was log transformed, the relationships are expressed in terms of percentage increases rather than absolute values.

- Choosing a waterfront home over a non-waterfront home will result in a 213% increase in house price on average*, assuming all other variables remain constant.
- Extra bathrooms are also important resulting in an 11% increase.
- New constructions however will realise 16% less than old
- An extra bedroom represents a 1.6% increase
- Other fuel types return a price 45% less than the base – gas

The model is far from ideal, with 41% of the variation in house price explainable by other factors. The inclusion of zip codes or geo-locational data would likely improve performance. The use of test and training data would also help to assess performance.

**the mean relates to the geometric and not arithmetic mean due to the logarithmic transformation.*

APPENDIX 1

All code, analysis and detailed comment found at this Colab Link:

APPENDIX 2

Comparison of Various Interaction Terms added to Model 2

	R	Adj R	AIC	p-value 1	p-value 2	p-value Int
Model 2	0.655	0.652	42990	n/a	n/a	n/a
Model With Waterfront and Living Area	0.656	0.653	42980	0.000	0.000	0.027
Rooms and Bathrooms	0.658	0.654	42970	0.310	0.515	0.004
Rooms and Bedrooms	0.656	0.653	42980	0.019	0.720	0.244
Rooms and Living Area	0.657	0.654	42970	0.608	0.000	0.016
Bedrooms and Living Area	0.656	0.653	42980	0.188	0.000	0.718
Bedrooms and Bathrooms	0.656	0.653	42980	0.183	0.005	0.821
Bathroom and Living Area	0.659	0.655	42970	0.472	0.000	0.000
Adding Age^2 Interaction Variable	0.660	0.657	42960	n/a	n/a	0.013

Note: collinearity does not imply interaction, but these variables were investigated due to interesting effects such as the negative bedrooms coefficient in Model 2. As the number of bedrooms increases we would expect the price to increase and living area to increase but for smaller homes, more bedrooms would imply smaller bedrooms, so would have a negative effect on price. This indicates a different relationship with price depending on the number of bedrooms.

REFERENCES

Bhalla, D. (2020). *Predicting Transformed Dependent Variable*. [online] ListenData. Available at: https://www.listendata.com/2015/09/predicting-transformed-dependent.html

Bland, J. and Altman, D. (2020). *Statistics notes: Transformations, means, and confidence intervals*.

Cazaar.com. (2019). *Interpret Regression Coefficient Estimates - {level-level, log-level, level-log & log-log regression} - Curtis Kephart*. [online] Available at: http://www.cazaar.com/ta/econ113/interpreting-beta

coefficients, B., Comtois, D. and Monica, g. (2019). *Back-transformation of regression coefficients*. [online] Cross Validated. Available at: https://stats.stackexchange.com/questions/27067/back-transformation-of-regression-coefficients

Community.amstat.org. (2019). *ASA Community*. [online] *"How Much is a Fireplace Worth?".* Available at: ttps://community.amstat.org/stats101/resources/viewdocument?DocumentKey=e4f8d3f1-41a3-4f01-9f8b-f8fbe1562c15&tab=librarydocuments&CommunityKey=5ad27b39-58d0-49e9-9f6f-0c39c82a0401

Data.library.virginia.edu. (2019). *Is R-squared Useless?* | *University of Virginia Library Research Data Services + Sciences*. [online] Available at: https://data.library.virginia.edu/is-r-squared-useless

Frost, J. (2019). *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions - Statistics By Jim*. [online] Statistics By Jim. Available at: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

GitHub. (2019). *issues with Saratoga · Issue #26 · ProjectMOSAIC/mosaicData*. [online] Available at: https://github.com/ProjectMOSAIC/mosaicData/issues/26

Investopedia. (2019). *What Is a Zero-Lot-Line House?* [online] Available at: https://www.investopedia.com/terms/z/zero-lot-line-house.asp

Is it acceptable to log transform an independent variable even if the residuals of the non-transformed model are normal? - Statalist. [online] Statalist.org. Available at: https://www.statalist.org/forums/forum/general-stata-discussion/general/1413967-is-it-acceptable-to-log-transform-an-independent-variable-even-if-the-residuals-of-the-non-transformed-model-are-normal

Medium. (2020). *How do you check the quality of your regression model in Python?*. [online] Available at: https://towardsdatascience.com/how-do-you-check-the-quality-of-your-regression-model-in-python-fa61759ff685

Medium. (2020). *Tests for Heteroskedasticity in Python*. [online] Available at: https://medium.com/@remycanario17/tests-for-heteroskedasticity-in-python-208a0fdb04ab

Medium. (2020). *Ways to Detect and Remove the Outliers*. [online] Available at: https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba

model?, H. and Rajan, V. (2020). *How to decide which interaction terms to include in a multiple regression model?*. [online] Cross Validated. Available at: https://stats.stackexchange.com/questions/171151/how-to-decide-which-interaction-terms-to-include-in-a-multiple-regression-model

Oecd.org. (2020). *Handbook on Residential Property Price Indices - en - OECD*. [online] Available at: https://www.oecd.org/publications/handbook-on-residential-property-price-indices-9789264197183-en.htm.

Online.stat.psu.edu. (2020). *6.4 - Tests for Constant Error Variance* | *STAT 462*. [online] Available at: https://online.stat.psu.edu/stat462/node/148/

Online.stat.psu.edu. (2020). *6.5 - Confidence Interval for the Mean Response | STAT 462*. [online] Available at: https://online.stat.psu.edu/stat462/node/150

package, F., docs, R. and browser, R. (2019). *SaratogaHouses: Houses in Saratoga County (2006) in mosaicData: Project MOSAIC Data Sets*. [online] Rdrr.io. Available at: https://rdrr.io/cran/mosaicData/man/SaratogaHouses.html

Packages, O., Power, S., Output, A., Examples, D., Questions, F., Examples, T., Test?, W., Loan, B., Policies, S., Consulting, W., Consulting, E., Service, F., Updating, S., Hire, C., Centers, O., Center, D., Clinic, D. and US, A. (2019). *FAQ How do I interpret a regression model when some variables are log transformed?*. [online] Stats.idre.ucla.edu. Available at: https://stats.idre.ucla.edu/other/mult-kg/faq/general/faqhow-do-i-interpret-a-regression-model-when-some-variables-are-log-transformed/

regression, C. and Pape, T. (2019). *Confidence interval of a log-linear regression*. [online] Cross Validated. Available at: https://stats.stackexchange.com/questions/388706/confidence-interval-of-a-log-linear-regression

Researchgate, (2020). [online] Available at: https://www.researchgate.net/post/When_I_use_AIC_akaike_information_criterion_to_find_the_model_of_the_best_fit_do_I_need_to_consider_p-values

Socratic.org. (2019). *Is age continuous or discrete data? + Example*. [online] Available at: https://socratic.org/questions/is-age-continuous-or-discrete-data.

Statisticalhorizons.com. (2019). *When Can You Safely Ignore Multicollinearity? | Statistical Horizons*. [online] Available at: https://statisticalhorizons.com/multicollinearity

Statistics How To. (2020). *Adjusted R2 / Adjusted R-Squared: What is it used for? - Statistics How To*. [online] Available at: https://www.statisticshowto.datasciencecentral.com/adjusted-r2/ Statistics Solutions. (2019). *Normality - Statistics Solutions*. [online] Available at: https://www.statisticssolutions.com/normality/ Statmath.wu.ac.at. (2020). [online] Available at: http://statmath.wu.ac.at/~fruehwirth/Oekonometrie_I/Folien_Econometrics_I_teil6.pdf

Stattrek.com. (2019). *Regression Slope Test*. [online] Available at: https://stattrek.com/regression/slope-test.aspx.

Taylor & Francis. (2020). *Modelling Home Prices Using Realtor Data*. [online] Available at: https://tandfonline.com/doi/full/10.1080/10691898.2008.11889569.

The Analysis Factor. (2019). *When to leave insignificant effects in a model - The Analysis Factor*. [online] Available at: https://www.theanalysisfactor.com/insignificant-effects-in-model/.

Toptal Finance Blog. (2019). *Real Estate Valuation Using Regression Analysis – A Tutorial*. [online] Available at: https://www.toptal.com/finance/real-estate/real-estate-valuation.

Vincentarelbundock.github.io. (2019). *R: Houses in Saratoga County (2006)*. [online] Available at: https://vincentarelbundock.github.io/Rdatasets/doc/mosaicData/SaratogaHouses.html.