2018

House Prices: Predictive Analytics

STATISTICAL AND PROBABILISTIC APPROACH TO HOUSE PREDICTION WITH REGRESSION MODELING

RAFAL DECOWSKI

CUNY | Computational Mathematics

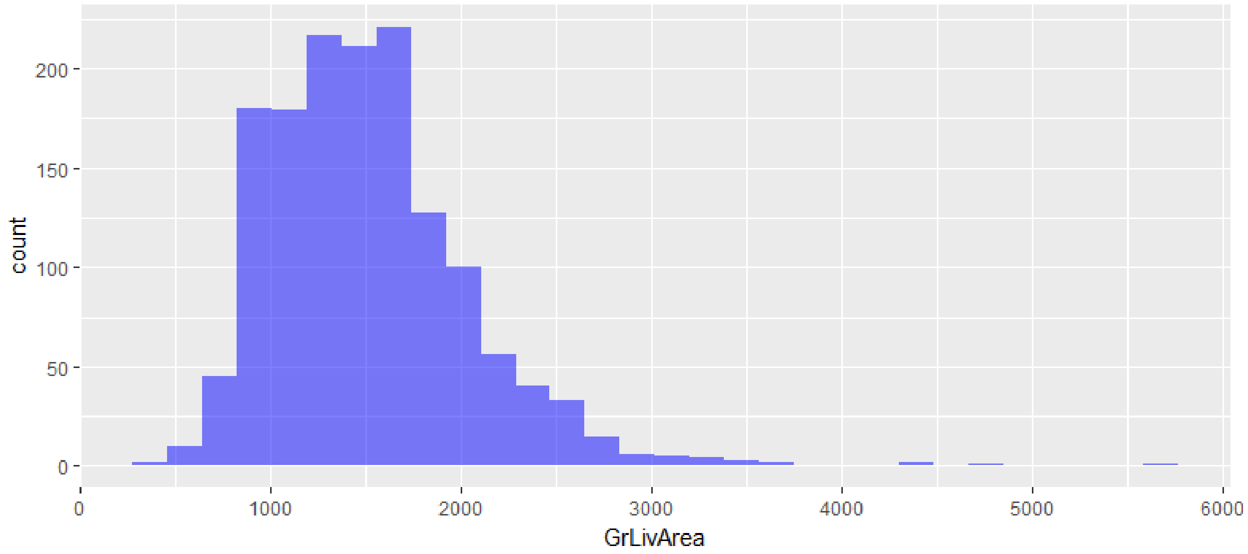
Contents

Conditional Probability

Chosen Variable

GrLivArea - Above grade (ground) living area square feet

Skewness - Right



P(X>x | Y>y)

What is the probability of X being greater than 1129.5 (x) given that Y is greater than 129975 (y)

P(A | B) = P(A & B) / P(B)

P(X>x | Y>y) = 0.65 / 0.75 = 0.8667

P(X>x, Y>y)

P(X>x, Y>y) = P(X>x) * P(Y>y)

What is the probability of X being greater than 1129.5 (x) and that Y is greater than 129975 (y)

P(X<x | Y>y)

What is the probability of X being less than 1129.5 (x) given that Y is greater than 129975 (y)

Confusion Matrix

Confusion matrix for values either smaller or equal first quartile value OR larger than.

x/y	<= 1Q	> 1Q	Total
<= 1Q	224	141	*365*
> 1Q	141	954	*1095*
Total	*365*	*1095*	*1460*

Variable dependence

Does splitting the training data in this fashion make them independent?

When knowledge of one event does not change the probability of another event happening, the two events are called statistically independent. In this case, we still can comfortably assume that the bigger the property is the more expensive it may be which makes the variables dependent.

A be the new variable counting those observations above the 1st quartile for X,
B be the new variable counting those observations above the 1st quartile for Y.

Does P(AB) = P(A)P(B)?

Calculated with Q3, 75% percentile as well as counts of values and the confusion matrix.
P(A)P(B) = 0.75 * 0.75 = 0.5625
P(AB) = 954/1460 = 0.6534247

P(AB) != P(A)P(B)

Chi Square test for association

Check mathematically, and then evaluate by running a Chi Square test for association.

X-squared = 340.75
df = 1
p-value < 2.2e-16

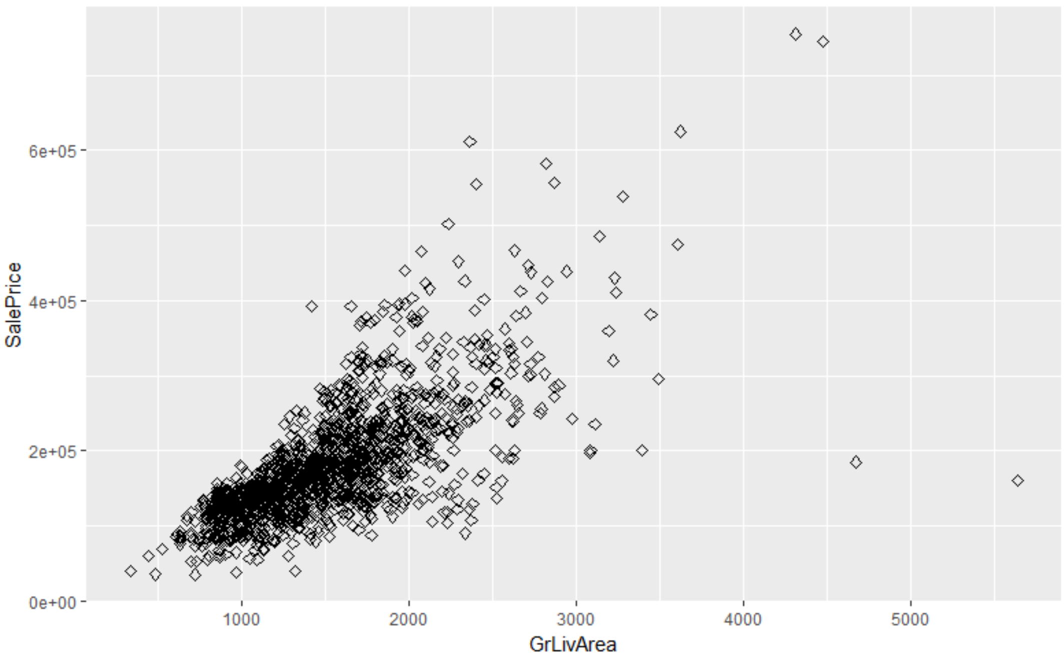
Since the p-value is less than 0.05, we have a statistically significant evidence to reject the null hypothesis and thus there is strong evidence to suggest an association between above grade (ground) living area square feet and the selling price

Descriptive and Inferential Statistics

Descriptive Statistics & Scatter Plot

Provide <u>univariate descriptive statistics</u> and appropriate <u>plots</u> for the training data set.

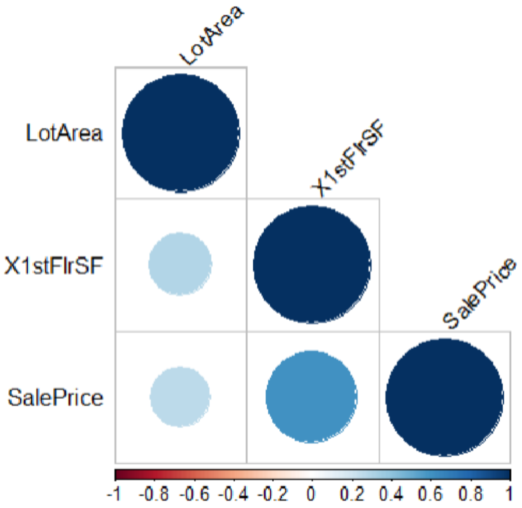
Min.	334
1st Q	1130
Median	1464
Mean	1515
3rd Q	1777
Max.	5642



Correlation

Derive a correlation matrix for any THREE quantitative variables in the dataset.

	X1stFlrSF	LotArea	Sale-Price
X1stFlrSF	1.00	0.30	0.61
LotArea	0.30	1.00	0.26
Sale-Price	0.61	0.26	1.00



Hypothesis Test

Test the hypotheses that the correlations between each pairwise set of variables is 0 and provide a 92% confidence interval.

	X1stFlrSF	LotArea	SalePrice
X1stFlrSF	0.00E+00	1.23E-31	5.39E-147
LotArea	1.23E-31	0.00E+00	1.12E-24
SalePrice	5.39E-147	1.12E-24	0.00E+00

OR

	X1stFlrSF	LotArea	SalePrice
X1stFlrSF	0	0	0
LotArea	0	0	0
SalePrice	0	0	0

Analysis Summary

Discuss the meaning of your analysis. Would you be worried about familywise error? Why or why not?

Correlation is a statistical formula that measures the strength between variables and relationships. We used it to determine the association between.

- First Flow Square Footage
- Lot Area
- Sale Price

In statistics, family-wise error rate (FWER) is the probability of making one or more false discoveries, or type I errors, among all the hypotheses when performing multiple hypotheses tests. In this case I would not worry about type-1 error because the only consequence would be adding a feature to predict the sale price that does not have a significant relationship.

The test confirms that all 3 variables have small-to-moderate, positive correlation between them. To prove that it is a statistically significant relationship we look at the **p-value**. P-value helps us with testing hypotheses. In this case we can pose two:

Hypothesis Null	**H0**: There **is no** correlation between the variables in question.
Alternative	**HA**: There **is** correlations between the variables in question

The smaller the p-value is the more confidently we can **reject H0**. We tested the values to be not only less than 0.05, which is the most popular threshold, but so close to 0 that we can comfortably round to 0. **Conclusion**: there is a statistically significant evidence that there is a relationship between the variables.

Linear Algebra and Correlation

Precision Matrix

Invert your 3 x 3 correlation matrix from above. This is known as the precision matrix and contains variance inflation factors on the diagonal.

	X1stFlrSF	LotArea	SalePrice
X1stFlrSF	1.63	-0.25	-0.93
LotArea	-0.25	1.11	-0.14
SalePrice	-0.93	-0.14	1.60

Mul - Correlation by Precision & Precision by Correlation

Multiply the correlation matrix by the precision matrix.
Multiply the precision matrix by the correlation matrix.

Both are identity matrices (diagonal 1’s)

	X1stFlrSF	LotArea	SalePrice
X1stFlrSF	1	0	0
LotArea	0	1	0
SalePrice	0	0	1

LU Matrix Decomposition

Conduct LU decomposition on the matrix.

L	X1stFlrSF	LotArea	SalePrice
X1stFlrSF	1.00	0.00	0.00
LotArea	0.30	1.00	0.00
SalePrice	0.61	0.09	1.00

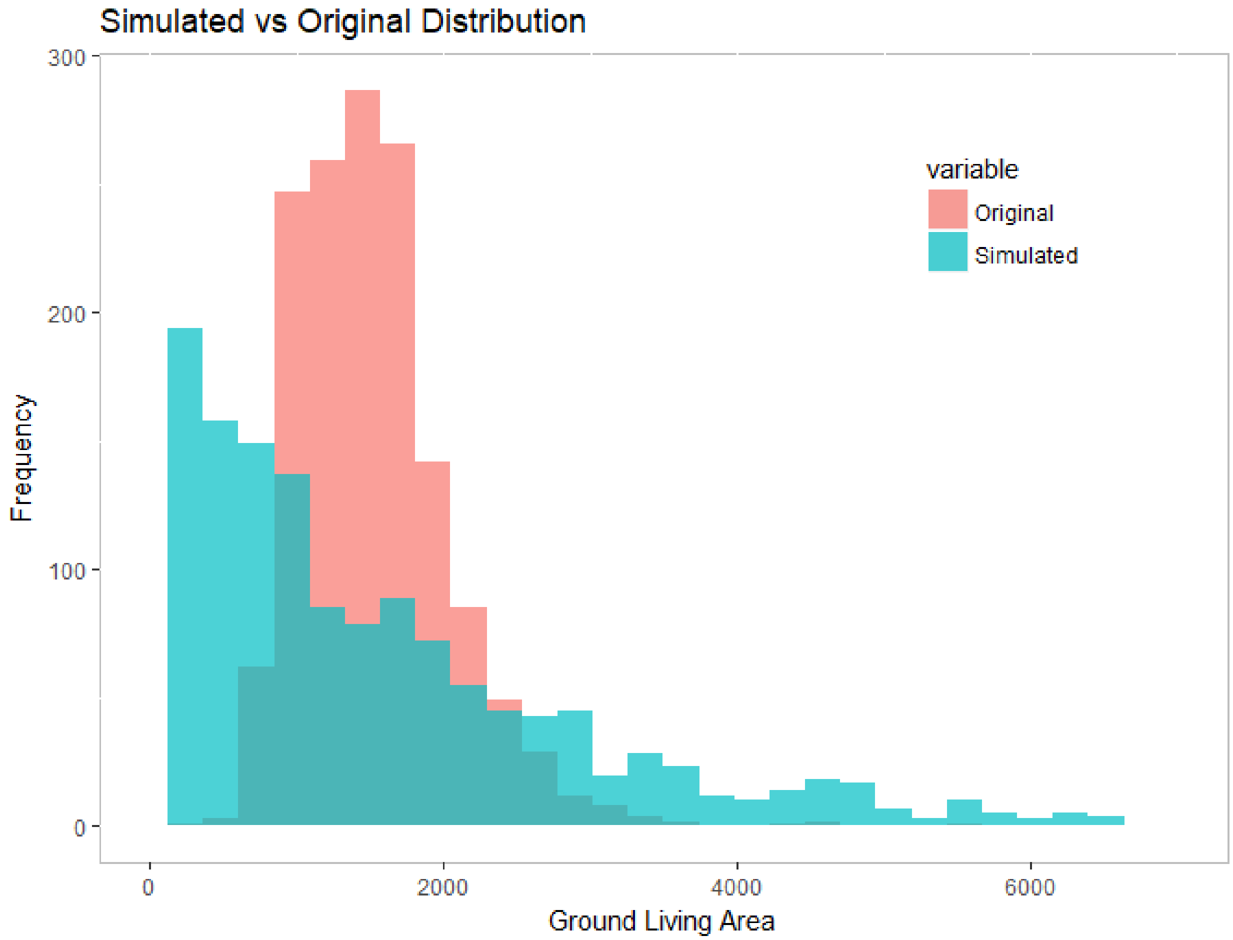
U	X1stFlrSF	LotArea	SalePrice
X1stFlrSF	1.00	0.30	0.61
LotArea	0.00	0.91	0.08
SalePrice	0.00	0.00	0.63

Calculus-Based Probability & Statistics

To make it an exponential distribution I used fitdistr() function in R with the ‘exponential’ flag. This provided the optimal **λ** value for this distribution: 0.000659864

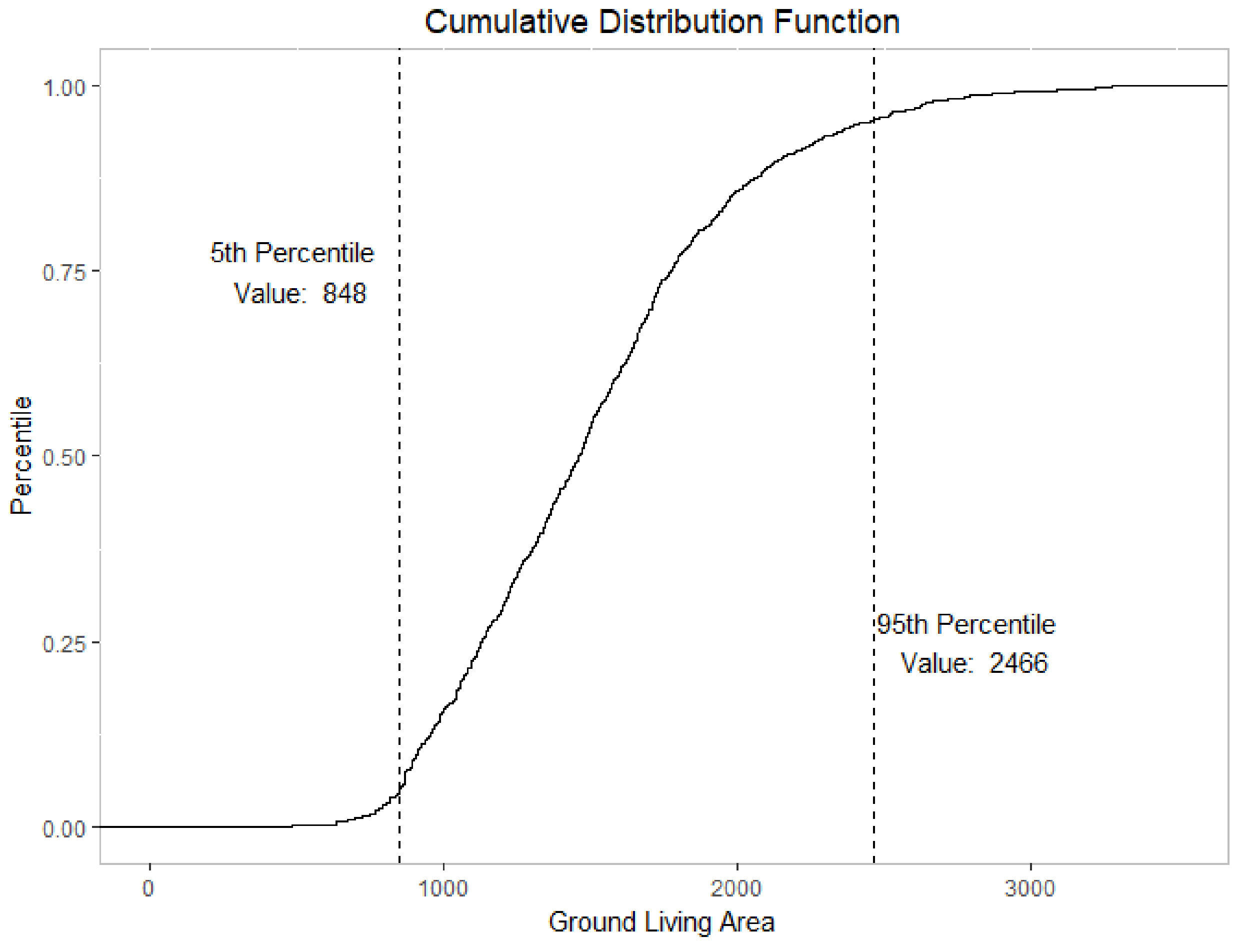
Histograms

Plot a histogram and compare it with a histogram of your original variable.



Percentile values - Observation

Using the exponential pdf, find the 5th and 95th percentiles using the cumulative distribution function (CDF).



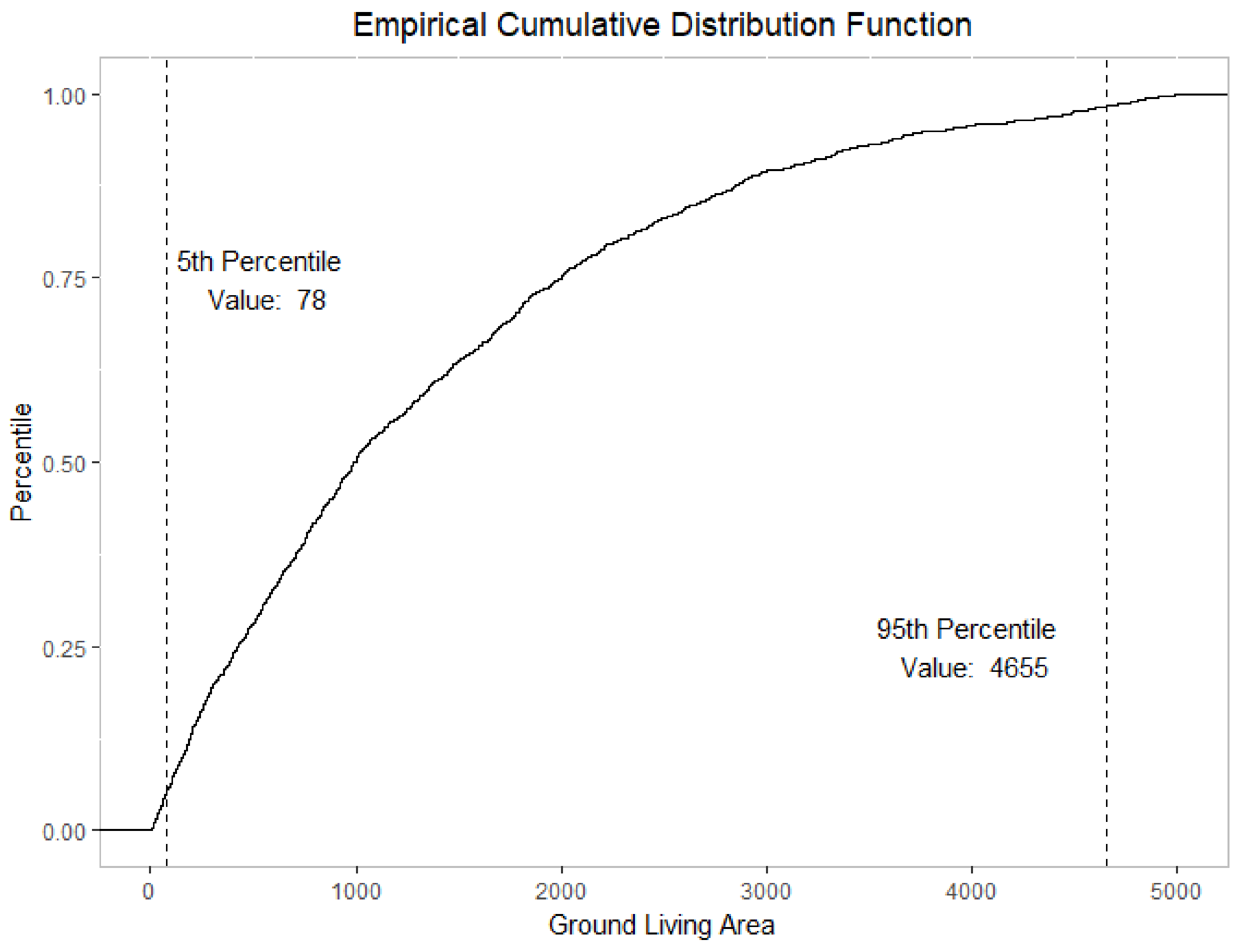
Confidence interval

Generate a 95% confidence interval from the empirical data, assuming normality.

Lower Confidence Interval: 1465
Upper Confidence Interval: 1625

Percentile values - Simulation

Provide the empirical 5th percentile and 95th percentile of the data. Discuss.



The simulated exponential distribution with the optimal lambda gradient shows a less steep slope and a wider 5th and 95th percentile range.

Modeling

Variable Grouping

There are 80 unique variables in the dataset. The approach for multiple regression modeling was to select the features into more meaningful categories.

Model Feature categories:

- Size
- Age/Time
- Quality and Condition
- House Features
- Rooms
- House Layout/Architecture

All models were fit using the Generalized Linear Models function – glm() with their respective categorized variables.

Performance

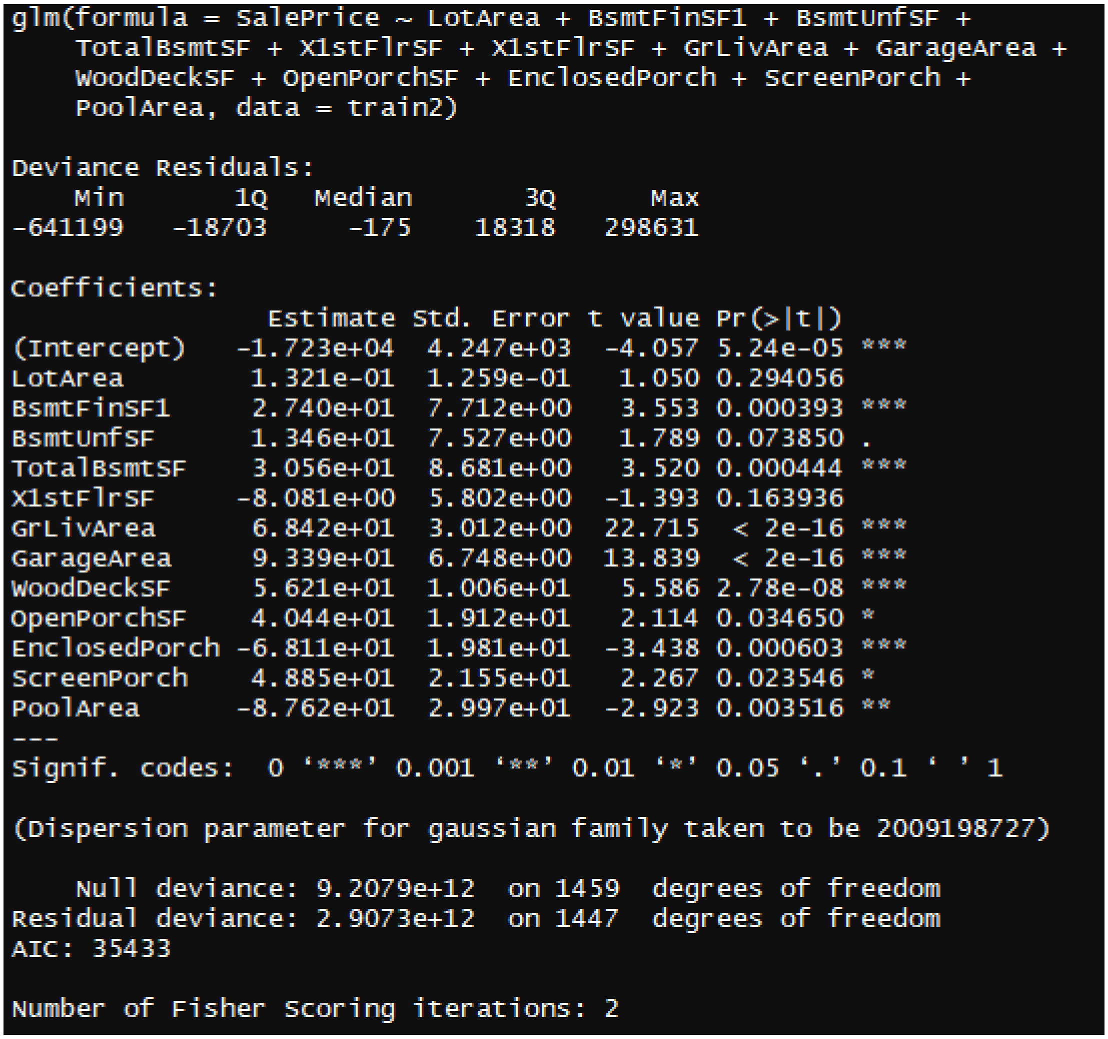
Since there are 6 very different models, as a performance comparative metric I used Min-Max Accuracy as well as Mean Absolute Percent Error (MAPE). The two values were collected for each model with the following results:

	MinMax	**MAPE**
Size	87%	17%
Rooms	83%	22%
Age	80%	26%
House Layout	70%	46%
Quality and Condition	70%	47%
House Features	69%	47%

Size related features provide the most accurate results with the smallest error. This is the model used for the final prediction and for the competition submission.

Further investigation of the model with the summary() function shows that there are some features that may not be statistically significant. This opens an opportunity to further tune and optimize the model.

Furthermore, the other models must have contained important variables as well. A combination of the most statistically significant features would be the optimal construction of a model for the highest accuracy.



```
glm(formula = SalePrice ~ LotArea + BsmtFinSF1 + BsmtUnfSF +
    TotalBsmtSF + X1stFlrSF + X1stFlrSF + GrLivArea + GarageArea +
    WoodDeckSF + OpenPorchSF + EnclosedPorch + ScreenPorch +
    PoolArea, data = train2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-641199   -18703     -175    18318   298631

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.723e+04  4.247e+03  -4.057 5.24e-05 ***
LotArea        1.321e-01  1.259e-01   1.050 0.294056
BsmtFinSF1     2.740e+01  7.712e+00   3.553 0.000393 ***
BsmtUnfSF      1.346e+01  7.527e+00   1.789 0.073850 .
TotalBsmtSF    3.056e+01  8.681e+00   3.520 0.000444 ***
X1stFlrSF     -8.081e+00  5.802e+00  -1.393 0.163936
GrLivArea      6.842e+01  3.012e+00  22.715  < 2e-16 ***
GarageArea     9.339e+01  6.748e+00  13.839  < 2e-16 ***
WoodDeckSF     5.621e+01  1.006e+01   5.586 2.78e-08 ***
OpenPorchSF    4.044e+01  1.912e+01   2.114 0.034650 *
EnclosedPorch -6.811e+01  1.981e+01  -3.438 0.000603 ***
ScreenPorch    4.885e+01  2.155e+01   2.267 0.023546 *
PoolArea      -8.762e+01  2.997e+01  -2.923 0.003516 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 2009198727)

    Null deviance: 9.2079e+12  on 1459  degrees of freedom
Residual deviance: 2.9073e+12  on 1447  degrees of freedom
AIC: 35433

Number of Fisher Scoring iterations: 2
```