Case Study Business Report

Tim Lee, Taylor Pellerin, Jake Toffler, Ian Smeenk, Alex Howard

10/4/2017

Contents

Introduction: **1**

Task 1A: Explanatory Modeling **1**

Task 1B: Prospective Customer Report - Morty's House **8**

Task 2: Predictive Modeling **12**

Introduction:

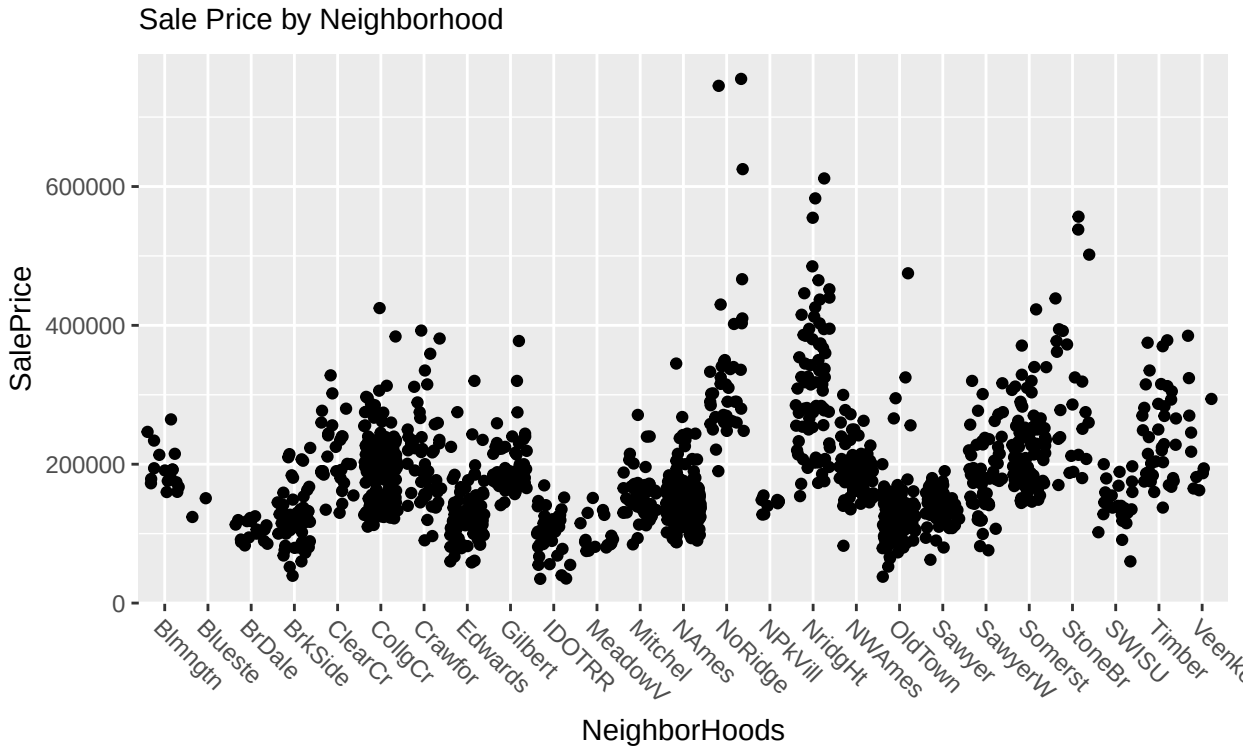
We have conducted in-depth regression analysis on a set of data from 1400+ house sales in Ames, Iowa with 80+ associated features. Our report into this data focuses on 3 key sections:

- Part 1A: Dominant Features related to different housing prices
- Part 1B: Top Recommendations for Morty, a new home seller
- Part 2B: Construction of a robust predictive model to be used going forward

Task 1A: Explanatory Modeling

Macro Data Summary:

Field	Value
Min Year	2006
Max Year	2010
Min Sale Price	$34,900
Median Sale Price	$163,000
Max Sale Price	$625,000
Number of Neighborhoods	25
Number of Rows (houses)	1460
Number of Rows (houses), excluding outliers	1349



Dataset Considerations:

- **Simplicity**: While it is enlightening to know exactly how big a pool is in square feet, the scarcity of pools makes this distinguishing detail unnecessary. A number of variables were simplified due to scarcity, or reduce the complexity of the problem.
- **Subjective to numeric**: how does one rate the condition of a basement? What is the definition of good? A large number of the features in the data covering "conditions" of walls, kitchens and exteriors were recorded in phrases such as "excellent" to "poor". Many of these variables were codified with integer scores to simplify the model.
- **Duplicate information**: Being overly detailed, there were many fields that recorded the same information, such as number of bedrooms upstairs, number of bedrooms downstairs, and then total bedrooms. There were similar fields recording square footage that were functionally added for sales sake, but these introduce collinearity problems for statistical analysis. Some of these fields were examined and dropped completely from the analysis.
- **Missing or Null Values**: with manual data collection or with the house not having all the features that in the database structure, empty fields are inevitable. These issues were either translated into something like "no garage", or were given the correct value based on the description in the data dictionary.
- **Outliers**: every generation has its Lebron James, and this housing dataset is no different. It contained a few massive outliers that were identified using DFFITs methods and were dealt with accordingly in the respective explanatory and predictive sections.
- **Normality Considerations**: for the method ordinary least squares regression to apply, a few statistical assumptions must be met. Y must be normally distributed, and the resulting residual (Y_predicted - Y_true) should be normal as well. This test was performed using Kolmogorov_Smirnov. It turns out that sale price distribution is skewed left (figure left). After running a Box-Cox analysis, it was determined to use the log of price as the baseline (right)

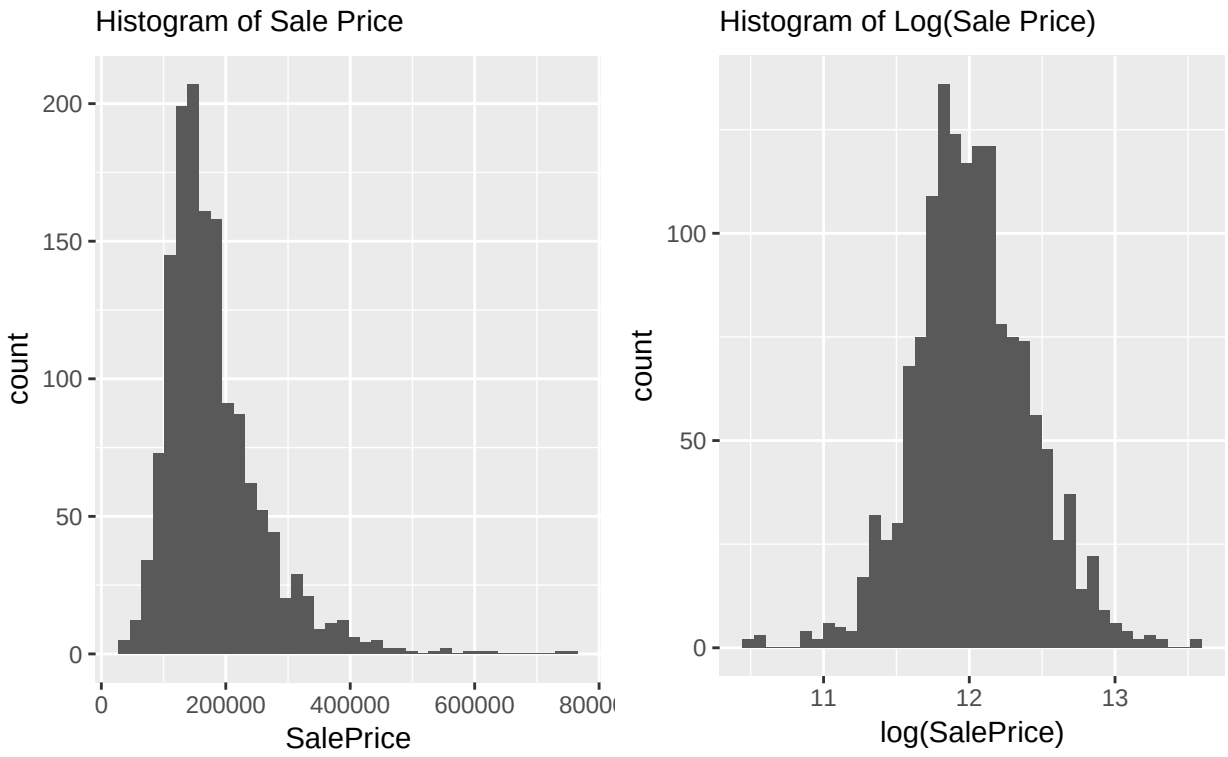


Figure: SalePrice .vs Log(SalePrice) comparison of counts

Fitting a Linear Model: Features Analysis

The previously described dataset was analyzed use R Studio, and a linear model (no regularization) was fit. The following is a summary of the optimal model.

Linear Model Summary: |Metric | Score| |–|—| |Rˆ2 |0.962| |MSE (log sale) | 0.0052| |MSE(saleprice) | 237,269,622| |RMSE(saleprice) | 15,403|

Important Features

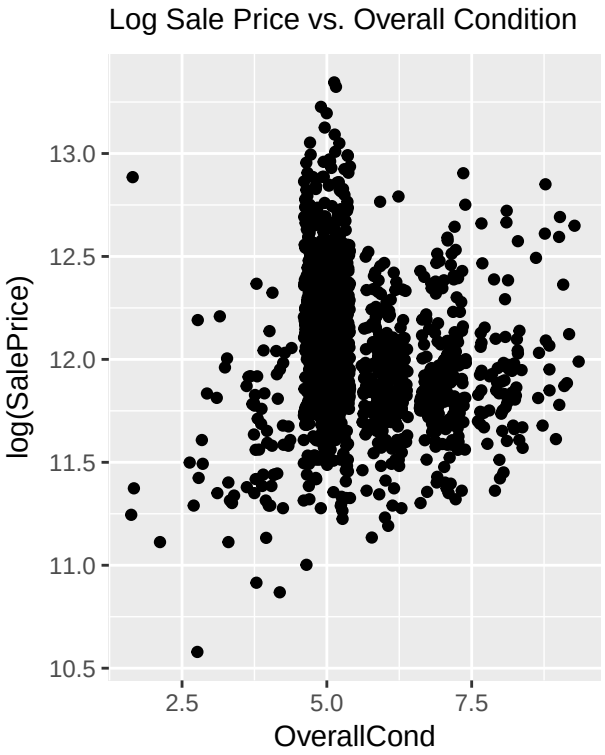
After data transformation and cleaning, a Ordinary Least Squares (OLS) model was fit onto the data. A statistical analysis of the fields show the following fields as the top significant features:

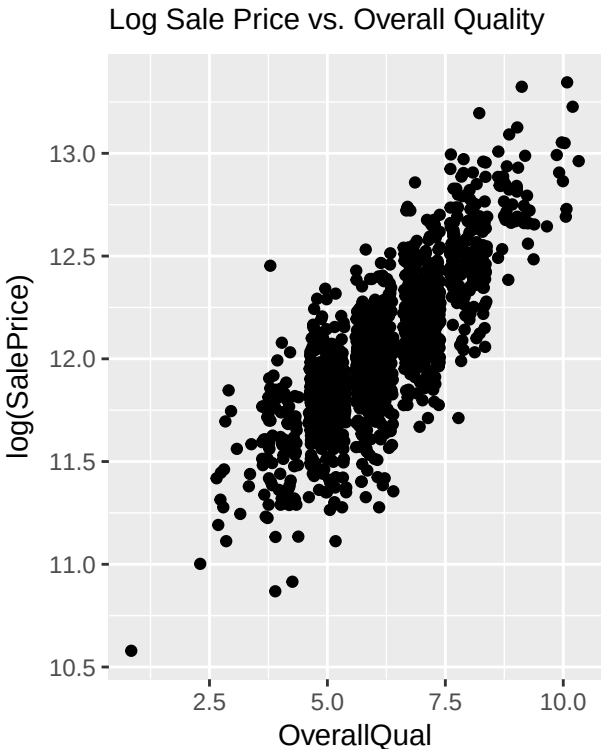
Sig. Rank	Field	Description
1	OverallCond	Rates the overall condition of the house
2	OverallQual	Rates the overall material and finish of the house
3	LotArea	Lot size in square feet
4	GrLivArea	Above grade (ground) living area square feet
5	Functional	Home functionality (Assume typical unless deductions are warranted)
6	YearBuilt	Original construction date
7	Condition1Norm	Condition1: Proximity to various conditions, specifically if near a major railway. Norm = Normal
8	BsmtExposure	BsmtExposure: Refers to walkout or garden level walls
9	MSZoningRL	Residential Low Density Zoning
10	BsmtFullBath	How many Full baths in the basement
11	MSZoningFV	Floating Village Residential
12	MSZoningRM	Residential Medium Density Zoning
13	MSZoningRH	Residential High Density Zong
14	Condition1PosN	Near positive feature such as parks
15	GarageCars	How many cars fit in the garage
16	MasVnrTypeNone	Masonry veneer type
17	Fireplaces	How many fireplaces the house has
18	BsmtFinSF1	Type 1 finished square feet
19	BsmtQual	BsmtQual: Evaluates the height of the basement

For full details of parameter estimates, see the extended report.

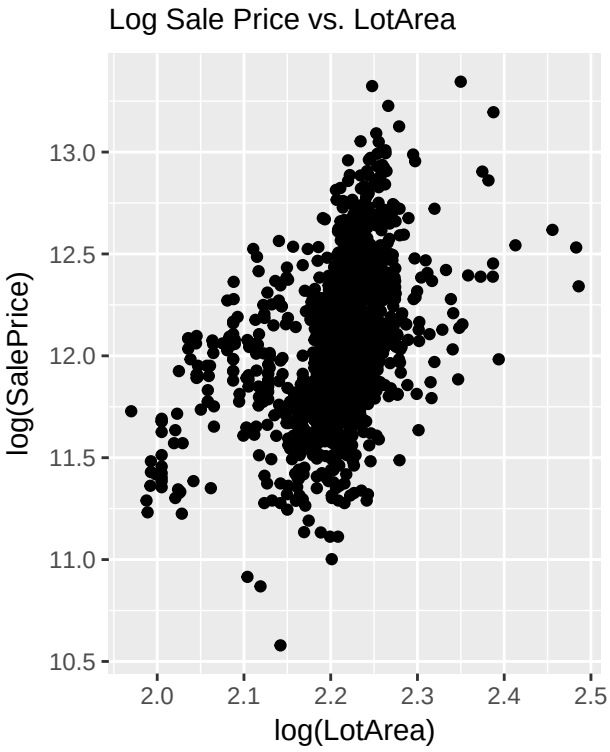
Trends between Sale price and Important Features:

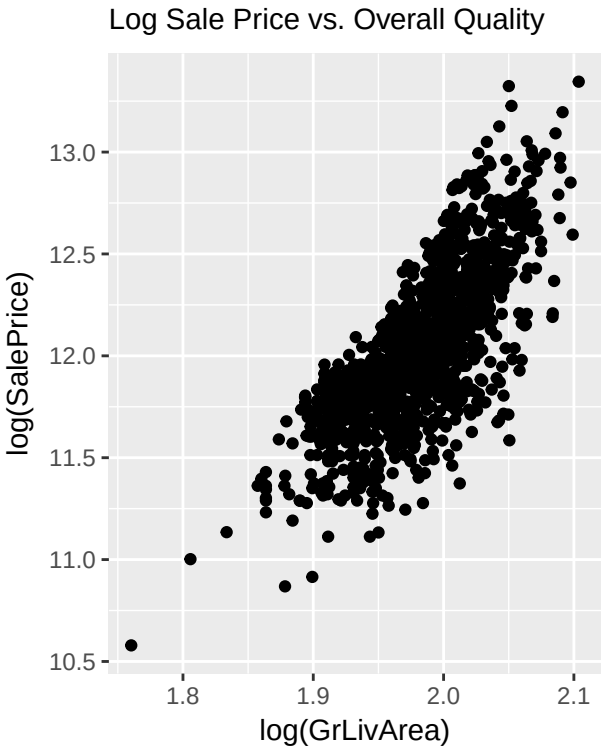
Select features from the above list of significant features were plotted versus sale price below:





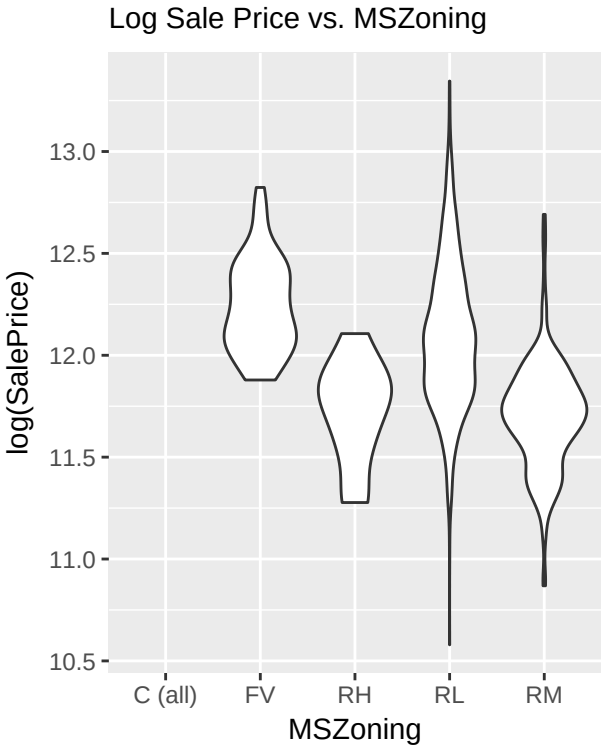
The overall condition has a loose linear correlation with Sale price, with a large number of houses showing up in the "normal" condition. It is not clear whether this is due to default choices, or all houses are generally rated normally, and only exceptional achieve higher ratings. The overall quality, on the other had shows a strong linear relationship with the log Home Price.

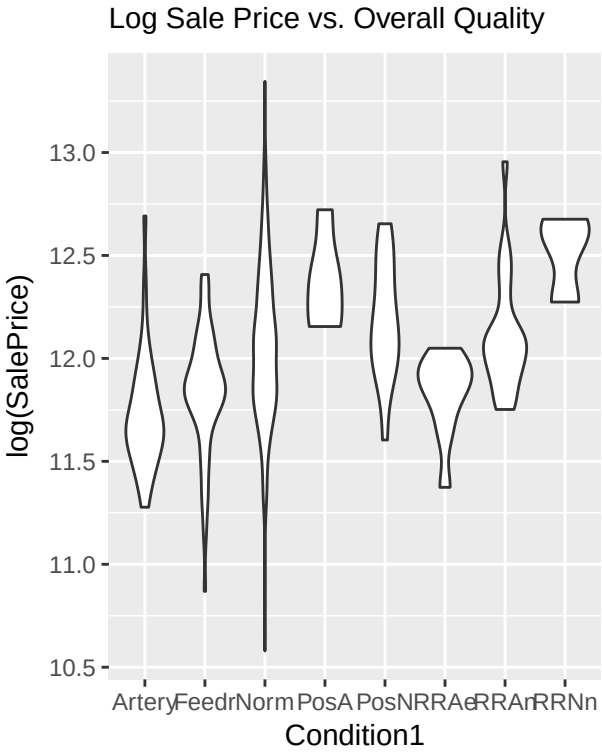




Comparing both the lot area and the general living area. Positive correlations can be seen in the scatter plots.

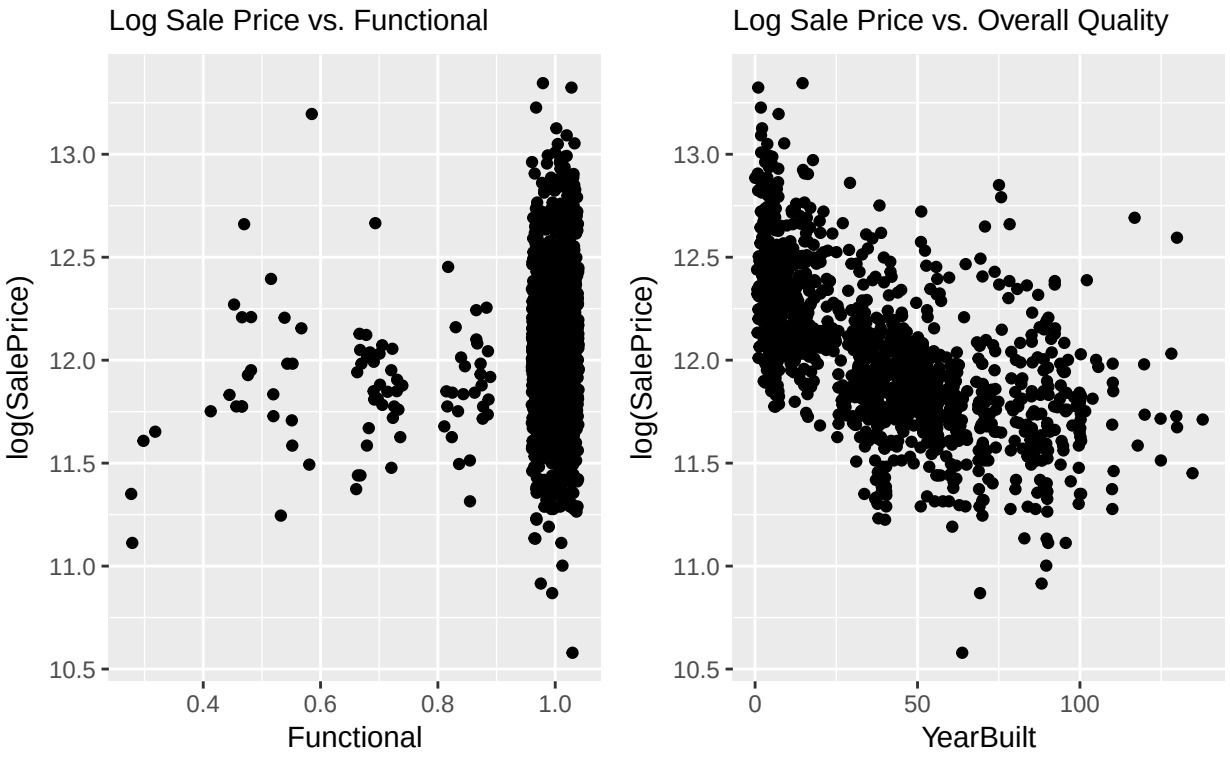
The lot area is a much weaker relationship to sale price. This is most likely due to lots not being directly proportional to house sizes. A small house can have no backyard or can be on a farm. The below plot shows the interactions between lot area, and living area. The size of the dots are the sale prices of the houses.





Two of the categorical features, MS Zoning and Condition1 (position of house relative to other local landmarks are shown below). The fatter parts of the violin plots show that in zoning, the floating village has the higher home prices, and low residential coming in 2nd. It should also be shown that it has the higher range of

prices, which makes sense considering the house to land ratio of a mansion and a farm are very similar; few buildings, lots of area. For the condition feature, normal contains most of the data, but being close to an intersection "artery" is associated with decreased prices, and being PosA - near a park or beltway showed a high increase in price. And interestingly being close to North-South railway was very high priced compared to being near East West railroad, or directly adjacent to the North-South railroad. Perhaps this is due to being near commuter traffic, but not right next to the train.



Looking at the functional field, this can be interpreted almost like deductions. 1 means there are few deductions, and 0 means severe deductions. A rough approximation is 100% functional. The way to interpret this plot is as the house loses functionality, the sale price is on average much lower. Finally, looking at YearBuilt, this field was translated into age by subtracting the report date of 2010. The number on the x-axis represents how old the house is. And as intuition would lend, the younger the house, the more expensive the house. We clearly see the inverse relationship to log SalePrice.

The above plots are a comparison of y_predicted from our OLS model (x-axix) and y_true (y-axis). The left plot units is the log version and the right plot units is the regular SalePrice. We can qualitatively observe a very tight relationship on the predicted vs. the actual sale price.

Task 1B: Prospective Customer Report - Morty's House

Our client, Morty, received an estimate from another firm stating that he could likely sell his house for $143,000. Ever the skeptics, we looked to make a rough guess at his sale price via our own explanatory model, and give him a few tips as to how he could increase the interpolated value of his home.

Our model predicts an expected house sale price of $ 143000. However, using standard regression Confidence Intervals, we are able to report to Morty that the maximum price we can reasonably expect his house to sell for (at the 95% level) is **$ 176966**. Any value more extreme than this would be considered to be a statistically significant outlier.

Features we recommend upgrading:

From our analysis, we identified 3 key areas that would most benefit from Morty's attention:
1. **Fireplaces** - the addition of an electrical fireplace is one of the most significant factors in raising a house price.
2. **Kitchen Quality** - Morty's kitchen quality is currently below market average. This, again, is one of the most significant factors in determining a house price and a refurbishment could generate significant profits.
3. **Garage Finish** - Morty's garage finish is also currently below market average. Similarly, we identified this as an easy feature to change that may add great value to the property.

Upgrading the kitchen and garage to the market average and adding a single fireplace adjusts our predictions as follows:
- Expected house sale price climbs from **$ 143000** to **$ 148406**
- The maximum value for which we could reasonably expect to sell Morty's house climbs from **$ 176966** to **$ 183714**.

Given that we're most concerned by the significance of our coefficients, let's rank by the SL and exclude any not significant at 95% level:

In the below table, we summarise the key statistics regarding these 3 features in our model:

field	beta	SL	Morty	Mean	exp_beta
Fireplaces	0.0189290	0.0000323	0.00	0.6189770	1.019109
KitchenQual	0.0107696	0.0001037	5.00	6.0392884	1.010828
GarageFinish	0.0281241	0.0193024	0.33	0.5783173	1.028523

Note that the final column in this table, the exponent of the coefficient, represents the approximate factor that value would increase if we were to increment the corresponding feature value by 1 unit.

Some other considerations, which are mostly out of Morty's control include:

1. Pick up his house and move it somewhere else, as neighborhood is quite significant
2. Forge the documentation to change the date that the garage was constructed
3. Buy some property off of a neighbor to fix the irregular shape of his lot

Final Notes on Morty House:

As noted before, we perfectly predicted the sell price of Morty's house using an explanatory model, which is a symptom of overfitting. This model was too flexible and as such perfectly tuned in on the data provided, which will in turn result in large variability in the predicted price when we feed the model an observation that it has not seen before. Generally, a model that is over specified for explanation will perform terribly on new data. We deal with this by building a more appropriately flexible, and as a result, more robust model which can better handle new input. This will be covered further in the following section, Task 2.

Cosine Similarity

Once all of the explanatory feature framework was setup, all of the houses in the dataset were converted into vectors. Categorical features were turned into dummy binary variables, so the entire vector was integer. With this vector in hand, housing similarity was calculated to find similar houses to morty. The 6 closest houses are listed below.

Wow! It turns out there's an exact copy of Morty's House in the dataset! Id#6.

Show Similar Houses - Cosine Similarity

	1	2	3	4	5	6
Id	6	445	1085	1111	1323	140
MSSubClass	1.5LVL_FIN	2LVL>1946	2LVL>1946	2LVL>1946	2LVL>1946	2LVL>1946
MSZoning	RL	RL	RL	RL	RL	RL
LotArea	9.555064	9.076923	9.475163	8.987322	9.228868	9.643875
Street	Pave	Pave	Pave	Pave	Pave	Pave
LotShape	IR1	Reg	IR2	Reg	IR1	IR1
LandContour	Lvl	Lvl	Lvl	Lvl	Lvl	Lvl
LotConfig	Inside	Inside	Corner	Inside	Inside	Inside
LandSlope	Gtl	Gtl	Gtl	Gtl	Gtl	Gtl
Neighborhood	Mitchel	CollgCr	Gilbert	Gilbert	NoRidge	CollgCr
Condition1	Norm	Norm	Norm	Norm	Norm	Norm
Condition2	Norm	Norm	Norm	Norm	Norm	Norm
OverallQual	5	7	6	6	7	6
OverallCond	5	5	5	5	5	5
YearBuilt	17	16	15	15	18	13
YearRemodAdd	15	15	14	14	18	13
RoofStyle	Gable	Gable	Gable	Gable	Gable	Gable
RoofMatl	250	250	250	250	250	250
MasVnrType	None	None	None	None	None	None
MasVnrArea	0	0	0	0	0	0
ExterQual	5	7	5	5	7	5
ExterCond	5	7	5	5	5	5
Foundation	Wood	PConc	PConc	PConc	PConc	PConc
BsmtQual	7	7	7	7	7	7
BsmtCond	5	5	5	5	5	5
BsmtExposure	1	1	1	1	1	1
BsmtFinType1	10	10	7	10	10	10
BsmtFinSF1	6.597146	6.466145	6.385194	5.393628	6.514713	6.711740
BsmtFinType2	0	0	0	0	0	0
BsmtFinSF2	0	0	0	0	0	0
BsmtUnfSF	4.174387	5.613128	4.605170	6.318968	4.343805	4.682131
Heating	GasA	GasA	GasA	GasA	GasA	GasA
HeatingQC	9	9	7	7	9	9
CentralAir	Y	Y	Y	Y	Y	Y
Electrical	SBrkr	SBrkr	SBrkr	SBrkr	SBrkr	SBrkr
X1stFlrSF	6.680855	6.839476	6.539586	6.651572	6.967909	6.834109
X2ndFlrSF	6.340359	6.883463	6.694562	6.786717	6.760415	6.729824
LowQualFinSF	0	0	0	0	0	0
GrLivArea	7.217443	7.554335	7.312553	7.413970	7.562162	7.475906
BsmtFullBath	1	1	0	1	1	1
BsmtHalfBath	0	0	0	0	0	0
FullBath	1	2	2	2	2	2
HalfBath	1	1	1	1	1	1
BedroomAbvGr	1	4	3	3	3	3
KitchenAbvGr	1	1	1	1	1	1
KitchenQual	5	7	5	5	7	7
TotRmsAbvGrd	5	8	6	8	8	7
Functional	1	1	1	1	1	1
Fireplaces	0	1	1	1	1	0
GarageType	Attchd	Attchd	Attchd	Attchd	Attchd	Attchd
GarageYrBlt	17	16	15	15	18	13
GarageFinish	0.33	0.33	1.00	1.00	0.66	0.66
GarageCars	2	2	2	2	2	2
GarageArea	6.175867	6.202536	6.016157	6.068426	6.336826	6.154858
GarageQual	5	5	5	5	5	5

	1	2	3	4	5	6
PavedDrive	Y	Y	Y	Y	Y	Y
WoodDeckSF	3.713572	4.976734	5.755742	5.416100	5.484797	5.624018
OpenPorchSF	3.433987	4.897840	3.806662	4.442651	3.688879	4.605170
EnclosedPorch	0	0	0	0	0	0
X3SsnPorch	5.771441	0.000000	0.000000	0.000000	0.000000	0.000000
ScreenPorch	0	0	0	0	0	0
PoolArea	0	0	0	0	0	0
MiscVal	6.552508	0.000000	0.000000	0.000000	0.000000	0.000000
MoSold	10	7	7	6	6	8
YrSold	1	2	4	2	0	1
SaleType	WD	WD	WD	WD	WD	WD
SalePrice	143000	210000	187500	188000	190000	231500
OLS_price	143000.0	211193.5	174310.5	184109.1	242765.4	207599.8
proxy_score	1.0000000	0.9989289	0.9989237	0.9989091	0.9989052	0.9988966

Task 2: Predictive Modeling

Here, we employed a handful of modeling techniques, iteratively testing out how well they performed as measured by the mean squared prediction error on a set of hold out data. Parameters for the models considered were generated via OLS, Ridge, LASSO, and Elastic Net algorithms. Besides sharing the same model type as explanatory modeling, extrapolation was markedly different in the following ways:

Normality Conditions: When optimizing our model for explanation, we used a number of functions to clean and alter the data in order to reduce bias and meet several assumptions when performing regression. When optimizing our model for prediction, however, we relaxed these assumptions to focus our concerns on prediction performance. Since we were only focused on the closeness of our predictions, our main criterion for selecting our best regression model was minimizing MSPE.

Variable Selection: To do this, we took raw data and tested out different subsets of our earlier data cleaning functions. Some of the steps taken in the variable selection for interpolation could not be omitted, such as cleaning up any null values, imputing values where necessary, and changing qualitative variables to numeric variables. For the other cleaning procedures, such as translating the year a house was built to age or removing collinear variables, we exhausted all combinations to create "partially clean" datasets.

Parameter Selection: With the data cleaned, we then ran several iterations of OLS, Ridge, LASSO, and Elastic Net on these variables to find the parameter estimates whose model had the lowest MSPE. After much consideration, we finally settled on a model with parameter estimates generated by L1 norm penalized regression (Lasso). Despite the fact that the table below has Lasso listed as having the highest MSPE, when averaged out, Lasso actually had the lowest mean MSPE. It also proved to be useful for decisive variable selection. The model in all its glory is detailed below.

Ridge Modeling

The ridge regularized regression was used as a possible model. Due to the high number of features and often redundant data, there could be a high level of collinearity between different fields. As mentioned previously, total area in square feet is redundant if all the parts are found in the data. For example, the MSSubClass is really a mashup of the year, and the number of levels of the house. The ridge should automatically minimize the impact of these redundant features on the predictive model.

Lasso Modeling

Lasso modeling was also applied against the dataset. This was ideal because we suspect that over 80+ features, there are probably a large number of features that do not have an affect on sales price. Things such as paved roads, or electrical breakers will most likely have little effect. Instead of plotting each of these fields against the main response variable SalePrice, we will apply a Lasso Regularized model which will drop some of these unnecessary variables.

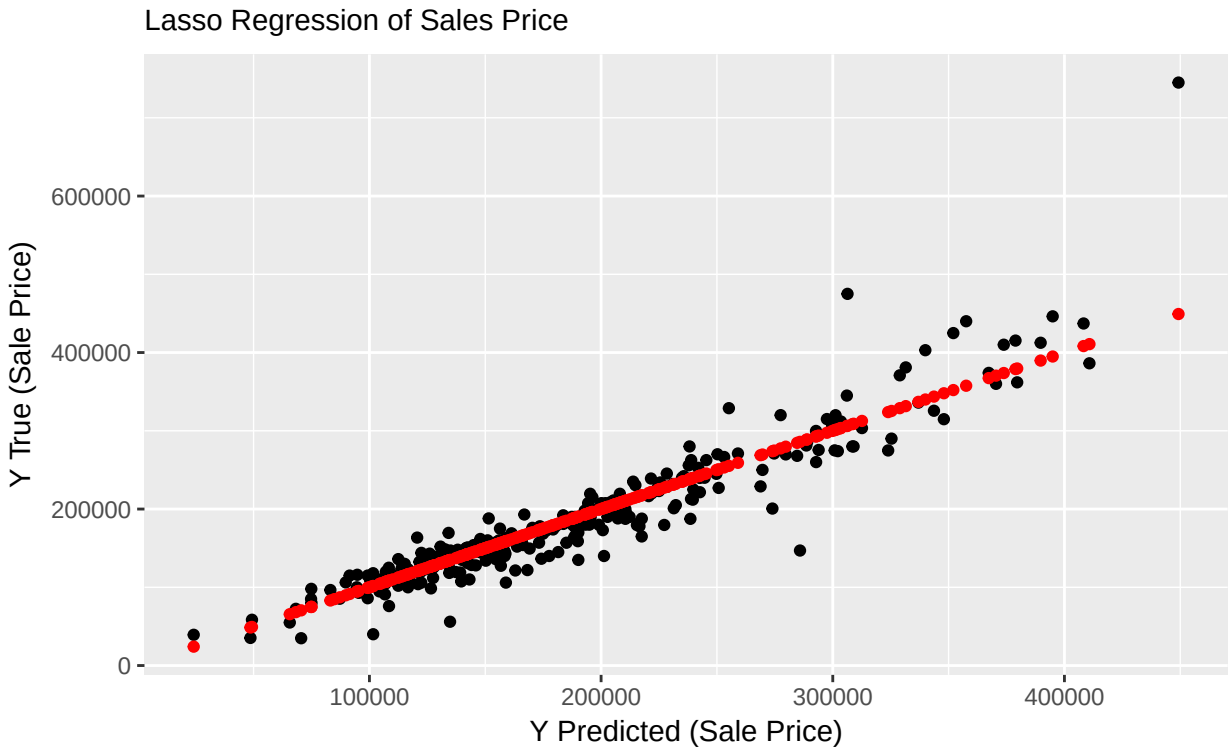
ElasticNet Modeling

ElasticNet was also used, which has both regularization terms from both Ridge and Lasso, with an alpha blend factor. The model that was tried here was with alpha = 0.5

OLS Modeling informed by lasso model

OLS was re-run again, only pulling features that were selected after running the Lasso model. This was used as a baseline score to compare the other models against.

model	MSPEscores	r.squared	best.lambda	RMSPE
Ridge	1163258208	0.8715369	27054.7059	34106.57
Lasso	947954186	0.8776744	639.7057	30788.86
Elastic.net	958787802	0.8787657	1165.7519	30964.30
OLS	12943797335	0.8815906	0.0000	113770.81



Appendix: Optimal Lasso Beta Values

	.
(Intercept)	-600290.7636491
NeighborhoodStoneBr	47546.9486765
NeighborhoodNoRidge	45779.6111614
NeighborhoodNridgHt	38212.1084479
LotShapeIR3	-21485.3073009
NeighborhoodCrawfor	20254.0858641
Exterior2ndImStucc	19487.1112858
BsmtExposureGd	18993.3766632
KitchenQualTA	-17375.2193842
KitchenQualGd	-17187.4198699
Exterior1stBrkFace	15701.6486998
SaleTypeNew	15262.0204982
BsmtQualGd	-15117.6529234
MSSubClass2-STORY PUD - 1946 & NEWER	-13136.4967108
NeighborhoodSomerst	13098.7013498
KitchenQualFa	-12153.7986527
Exterior1stImStucc	-12059.7856599
BldgTypeDuplex	-11957.7778370
OverallQual	11927.8826802
NeighborhoodVeenker	11855.4212477
BsmtQualTA	-11366.4169936
BldgTypeTwnhs	-11192.5274359
MSSubClass1-STORY PUD (Planned Unit Development) - 1946 & NEWER	-10682.3402985
GarageCars	10371.6810543
Exterior2ndStucco	-9845.1280125
MSSubClass2 FAMILY CONVERSION - ALL STYLES AND AGES	-9720.8194963
LotConfigCulDSac	8629.5756805
BsmtExposureNone	-8580.8213685
NeighborhoodEdwards	-8503.6788059
LotConfigFR3	-8121.0734210
Condition1Norm	8058.3231051
Condition1RRAe	-7634.4441414
BldgTypeTwnhsE	-7271.3180334
HouseStyle1Story	7208.4578325
LotShapeIR2	7206.1339055
BsmtQualFa	-6768.7201393
SaleTypeCon	6439.2655846
Exterior2ndStone	-6418.2711167
GarageYrBlt1910s	-6114.3461017
BsmtFullBath	5458.0109132
OverallCond	4991.1468029
FullBath	4882.2988619
BsmtFinType1GLQ	4850.4019974
NeighborhoodMitchel	-4640.4008961
GarageTypeBasment	-4579.0023548
Fireplaces	4486.4610517
BsmtExposureNo	-4453.8171613
BsmtFinType1Unf	-4403.5465758
MSSubClassDUPLEX - ALL STYLES AND AGES	-4210.4166754
NeighborhoodOldTown	-4108.8019758

	.
BsmtFinType1None	-3768.1253603
Exterior1stCemntBd	3747.3140442
ExterQualTA	-3722.4575117
NeighborhoodSWISU	-3696.9892680
NeighborhoodBrkSide	3639.9367766
GarageTypeNone	3192.3666590
FoundationSlab	-3030.2383726
HalfBath	2987.2925021
LotConfigFR2	-2819.7683656
GarageTypeCarPort	-2629.7683697
MSZoningRM	-2566.8397425
Exterior2ndWd Shng	-2560.3506571
Condition1PosN	-2491.6998485
NeighborhoodNPkVill	2154.5828191
GarageYrBlt1930s	1919.9793449
GarageFinishRFn	-1854.7698970
Condition1Feedr	-1841.8005922
HeatingQCGd	-1736.6116220
HeatingQCTA	-1636.8344329
SaleConditionFamily	-1618.2394974
NeighborhoodNAmes	-1569.4744598
FoundationPConc	1528.7068681
Exterior2ndVinylSd	1298.4717934
TotRmsAbvGrd	1284.6394750
BsmtFinType1LwQ	-1280.5045482
SaleConditionNormal	1213.1950440
Exterior1stHdBoard	-1157.1840456
GarageYrBlt1970s	-1088.1777938
RoofStyleHip	1056.7033349
NeighborhoodNWAmes	-872.9369427
GarageTypeBuiltIn	806.4840364
BsmtFinType2None	-760.2472000
SaleTypeCWD	637.1008566
MSZoningRL	547.1498084
BedroomAbvGr	-520.1901375
NeighborhoodIDOTRR	-489.0813775
GarageYrBlt1960s	-464.8747145
RoofStyleGable	-455.7720508
MasVnrTypeNone	355.4749340
YearBuilt	216.9891812
Condition1RRAn	204.8301857
BsmtQualNone	-200.4517811
MasVnrTypeBrkFace	-163.5089968
MoSold	-125.8062820
YearRemodAdd	77.7062775
GarageYrBltNAs	65.2256599
GrLivArea	42.2969084
GarageFinishUnf	-26.1244250
WoodDeckSF	17.2188849
MasVnrArea	13.1429611
TotalBsmtSF	4.7769917
BsmtFinSF1	3.5359027

	.
LotArea	0.3089301