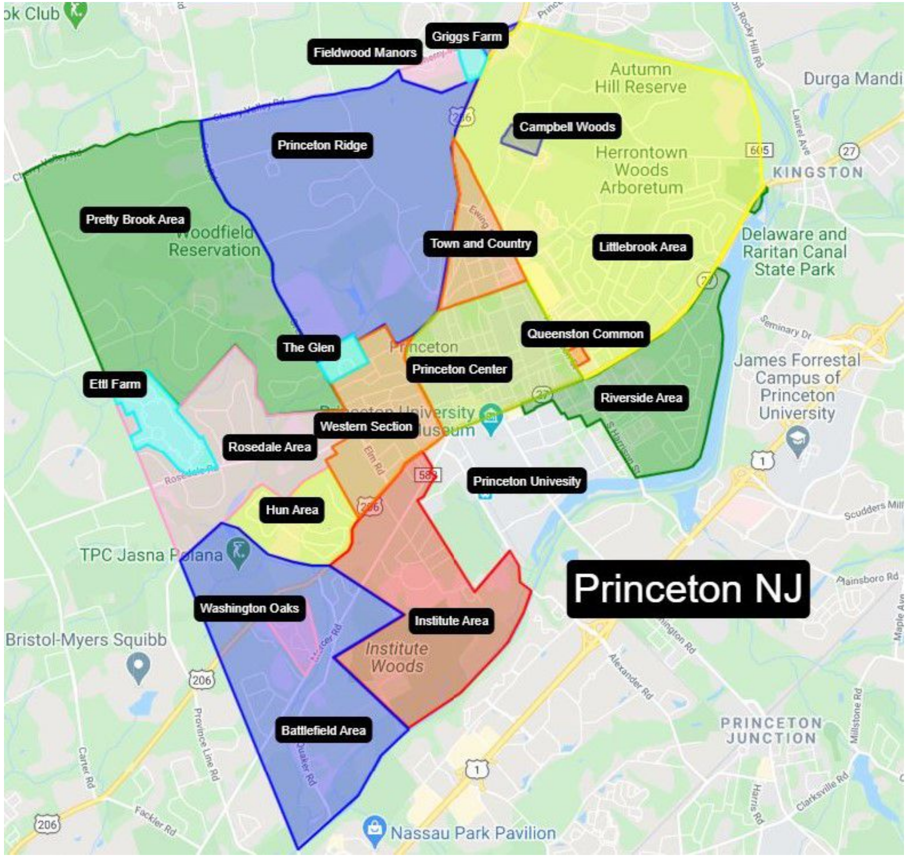
Princeton House Price Predictors

Brandon Feder & Atticus Wang

Our Motivating Question

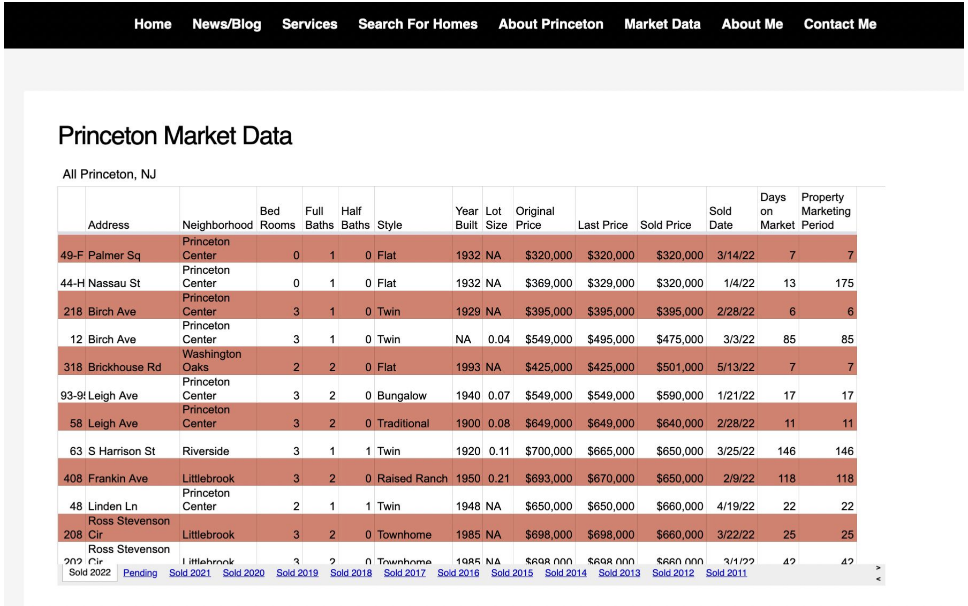
The goal of our research project was to answer the question “What characteristics of a house in Princeton most greatly influences it’s price?”

These “characteristics” include the style of the house, the neighborhood it is in, the number and type of bathrooms the house, and the house’s’ age.



Data

Beatrice Bloom, a Princeton Residential Specialist, provides many great resources about the Princeton housing market including a table of houses sold in Princeton since 2011.



Home News/Blog Services Search For Homes About Princeton Market Data About Me Contact Me

Princeton Market Data

All Princeton, NJ

	Address	Neighborhood	Bed Rooms	Full Baths	Half Baths	Style	Year Built	Lot Size	Original Price	Last Price	Sold Price	Sold Date	Days on Market	Property Marketing Period
49-F	Palmer Sq	Princeton Center	0	1	0	Flat	1932	NA	$320,000	$320,000	$320,000	3/14/22	7	7
44-H	Nassau St	Princeton Center	0	1	0	Flat	1932	NA	$369,000	$329,000	$320,000	1/4/22	13	175
218	Birch Ave	Princeton Center	3	1	0	Twin	1929	NA	$395,000	$395,000	$395,000	2/28/22	6	6
12	Birch Ave	Princeton Center	3	1	0	Twin	NA	0.04	$549,000	$495,000	$475,000	3/3/22	85	85
318	Brickhouse Rd	Washington Oaks	2	2	0	Flat	1993	NA	$425,000	$425,000	$501,000	5/13/22	7	7
93-9!	Leigh Ave	Princeton Center	3	2	0	Bungalow	1940	0.07	$549,000	$549,000	$590,000	1/21/22	17	17
58	Leigh Ave	Princeton Center	3	2	0	Traditional	1900	0.08	$649,000	$649,000	$640,000	2/28/22	11	11
63	S Harrison St	Riverside	3	1	1	Twin	1920	0.11	$700,000	$665,000	$650,000	3/25/22	146	146
408	Frankin Ave	Littlebrook	3	2	0	Raised Ranch	1950	0.21	$693,000	$670,000	$650,000	2/9/22	118	118
48	Linden Ln	Princeton Center	2	1	1	Twin	1948	NA	$650,000	$650,000	$660,000	4/19/22	22	22
208	Ross Stevenson Cir	Littlebrook	3	2	0	Townhome	1985	NA	$698,000	$698,000	$660,000	3/22/22	25	25
202	Ross Stevenson Cir	Littlebrook	3	2	0	Townhome	1985	NA	$698,000	$698,000	$660,000	3/1/22	42	42

Sold 2022 Pending Sold 2021 Sold 2020 Sold 2019 Sold 2018 Sold 2017 Sold 2016 Sold 2015 Sold 2014 Sold 2013 Sold 2012 Sold 2011

Regression Analysis

We tried many different models including a full linear model, forward and backward stepwise regressions, a ridge regression, LASSO regression, elastic net regression, and finally, a PCR regression.

However, the LASSO regression seemed most promising because it has the lowest RMSE and good R^2.

Table 2: Performance of all models we tried

	R2	RMSE
full lm	0.5448	454579.1
forward	0.5449	454880.4
backward	0.5449	454880.4
ridge	0.5170	455475.4
lasso	0.5227	452814.7
elastic net	0.5192	454296.0
pcr	0.5996	507588.0

LASSO Regression

In a Least Absolute Shrinkage and Selection Operator (LASSO) model, predictions are enhanced by performing both variable selection (selecting which predictors to use) AND shrinkage (shrinking coefficients of less important predictors).

More formally, a LASSO regression attempts to minimize

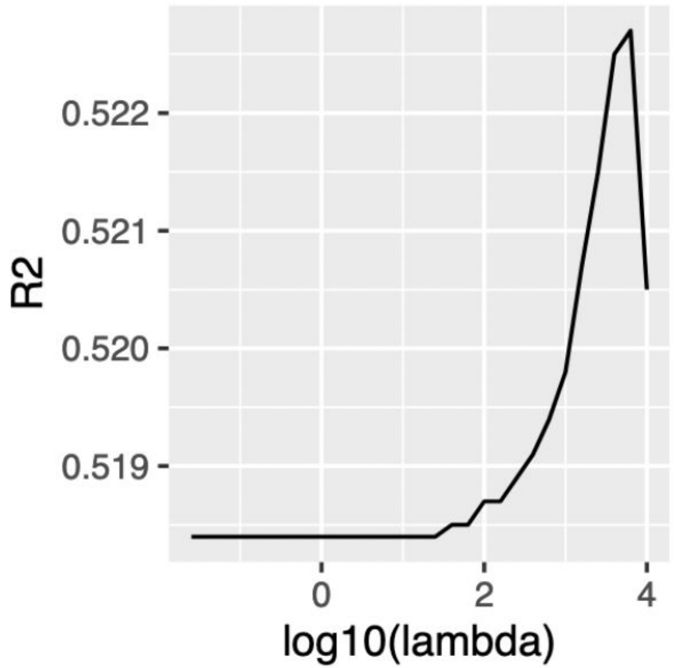
$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

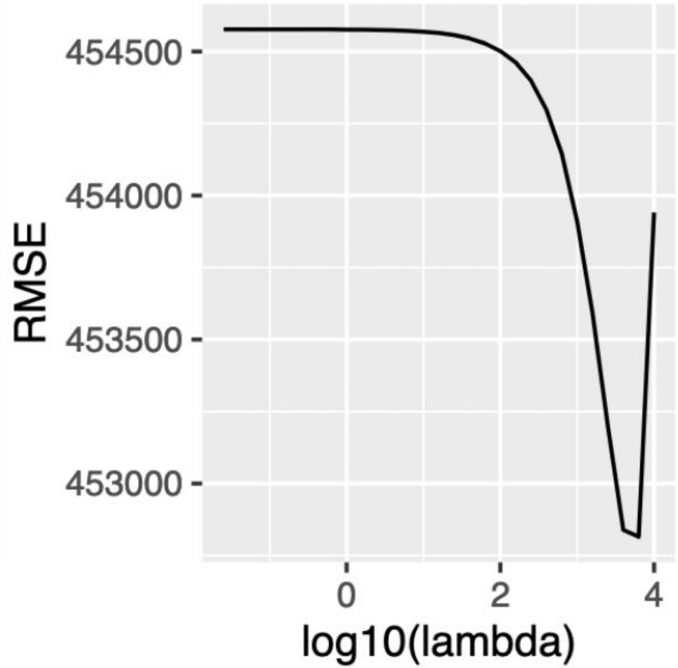
for all non-negative real λs.

Refer to our report for explanation of other models.

LASSO Model Output

Here are the R^2 and RMSE values vs λ. The best choice of λ occurred around 6309.





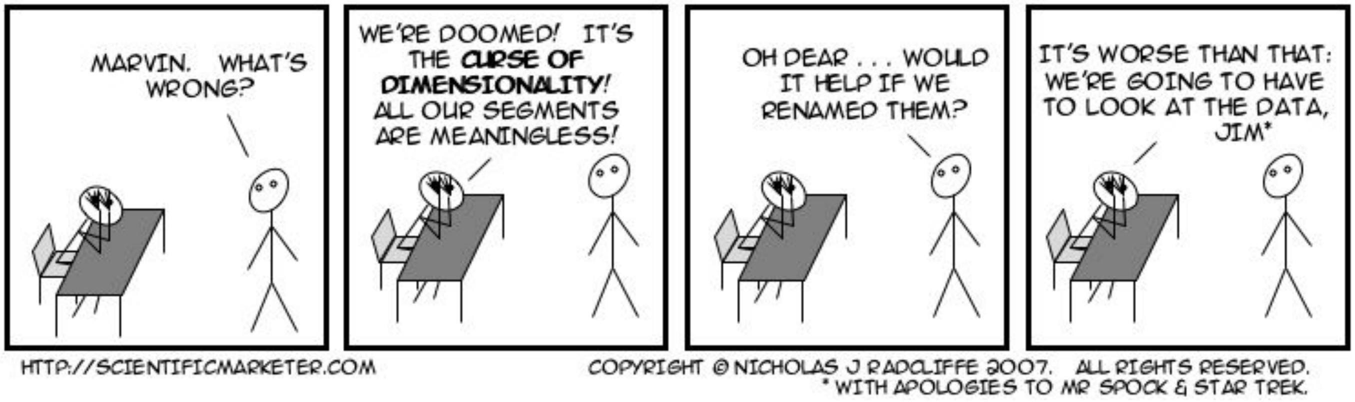
Interpreting Model Coefficients

Of the 61 variables considered in the regression (many of which were dummies), the variables with the largest coefficients were year the hose was sold (412690), the number of full bathrooms (232332), and the number of half bathrooms (57121).

This means, in general, the price of a house increases by $400K dollars each year, each additional full bathroom increases the price by $230K, and finally, each half bathroom increases the price of the house by $60K dollars.

Discussion and Limitations

Though our model fits the data well, we were restricted by the curse of dimensionality: a required sample size will grow exponentially with the number of dimensions of the data. That is, we may not have had sufficient data to guarantee without a reasonable doubt that our model did not “detect” a coincidence in the data that would not be prevalent with more data.



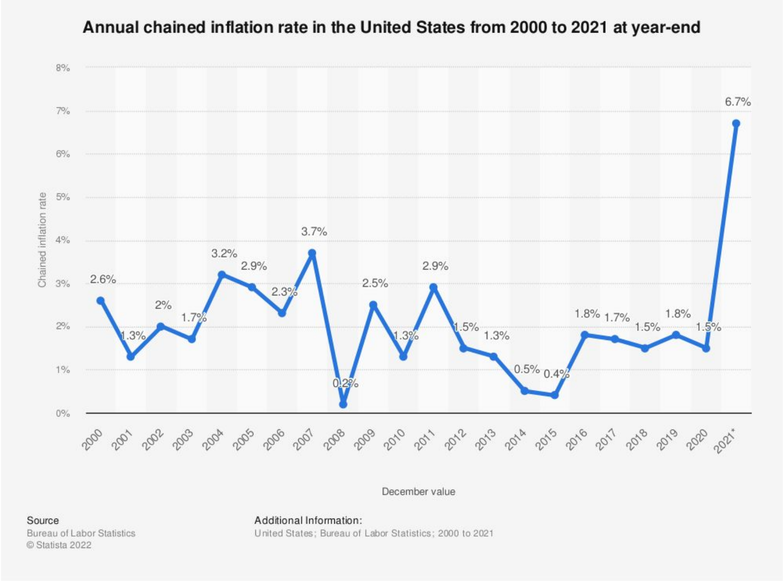
Discussion and Limitations

In addition, we did not have time to incorporate US census data as we discussed in our proposal. From personal experience, we believe that this data would not particularly insightful because “like” individuals tend to congregate in neighborhoods which were already analyzed in the data.

That being said, incorporating this data is something that would be useful for a more in-depth analysis.

Discussion and Limitations

Finally, we should have adjusted house prices for inflation. The LASSO model shows that the year the house sold greatly affects the houses price. However, we are not sure if houses actually “age like wine,” or if they simply get more expensive inversely with the dollar’s relative worth.



Conclusion

Overall, our analysis was successful in that it answered the question: “What characteristics of a house in Princeton most greatly influences it’s price?” However, with more time we would have considered more characteristics of the house, attempted to use other models (maybe a neural network), and performed actual predictions based off our data.

THANKS!