Final: Getting Real with Real Estate

Section 4 Team 2 - Patrick Barranger, Mark Paluta, Subha Paravastu

Our goal for this analysis was to identify trends in a publicly available Iowa housing data set and verify the trends in two independent data sets from California and North Carolina. Our analysis uncovered some of the houses recorded in each of the three datasets might have errors. These data points were identified and removed since we could not validate the source. We identified and confirmed correlation between the final sale price of a house and its: square footage, number of bathrooms, and year built. Interestingly, we were unable to detect strong correlations in bedrooms or lot area.

Introduction

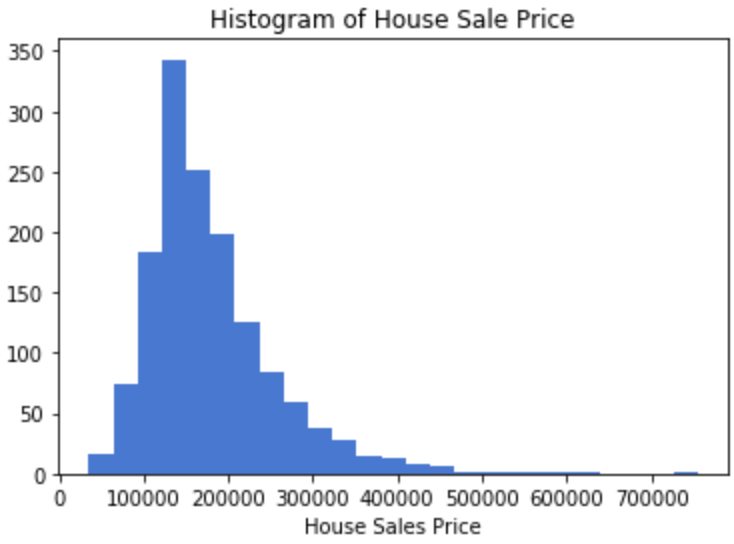
We started by identifying trends in our Iowa housing dataset. Our data comes from the Ames Housing Dataset built by Dean De Cock and hosted on Kaggle.com. While the dataset contained 79 variables we focused on the handful that overlapped with at least one of the variables in our other two datasets (CA and NC). With each data set we start with a sanity-check and EDA before moving onto identifying or confirming trends.

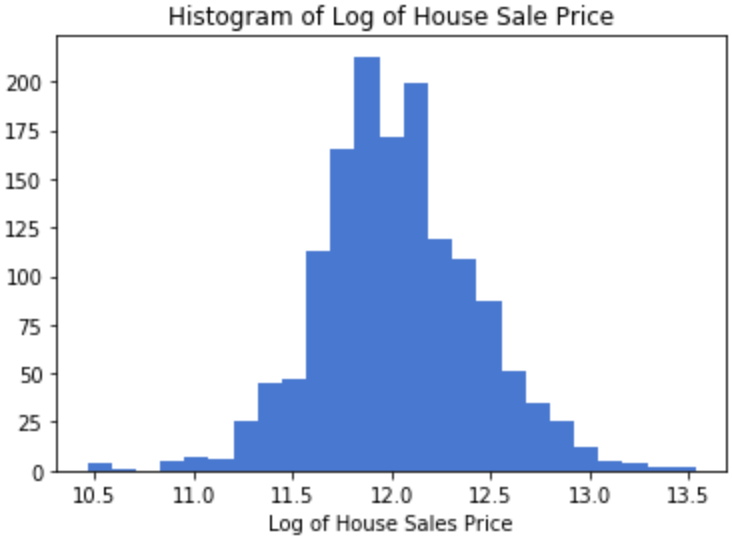
Iowa Dataset

We investigated the bed, bathroom, square footage, lot area, and year built data points in the Iowa dataset and how they related to the final sale price. We started with a sanity-check of the data and more clearly defining our outcome variable as Log Sale Price. We then close the section by deep-diving into each of the above variables. For each, we perform and EDA, handle outliers and missing values, perform feature engineering, and conclude with identifying any relations with log sale price.

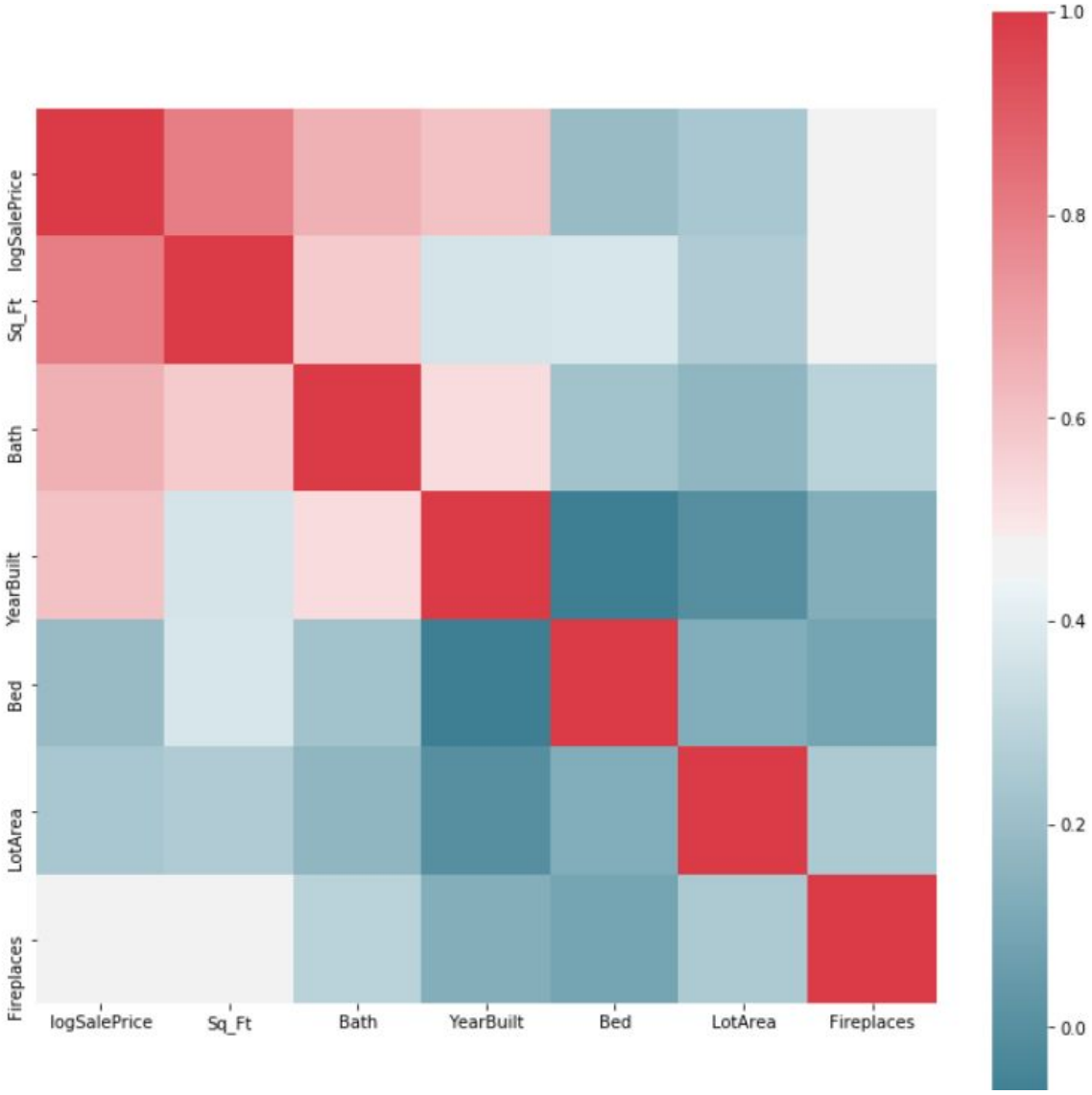
EDA

We began our analysis by looking at the head of the data as well as a summary of the key data points. Reviewing our data, we noted that our sale prices had a heavy right skew. To aid in visualization and future modeling we applied a log transform to result in a more normal distribution. For the rest of the analysis we measured correlation relative to Log Sales Price.


Histogram of House Sale Price
350
300
250
200
150
100
50
0
0
100000
200000
300000
400000
500000
600000
700000
House Sales Price


Histogram of Log of House Sale Price
200
175
150
125
100
75
50
25
0
10.5
11.0
11.5
12.0
12.5
13.0
13.5
Log of House Sales Price

We also visualized the correlation among our variables to get some early insights.



Year Built

The hypothesis is ‘Year Built’ is a key variable in determining the Sale Price of a house.

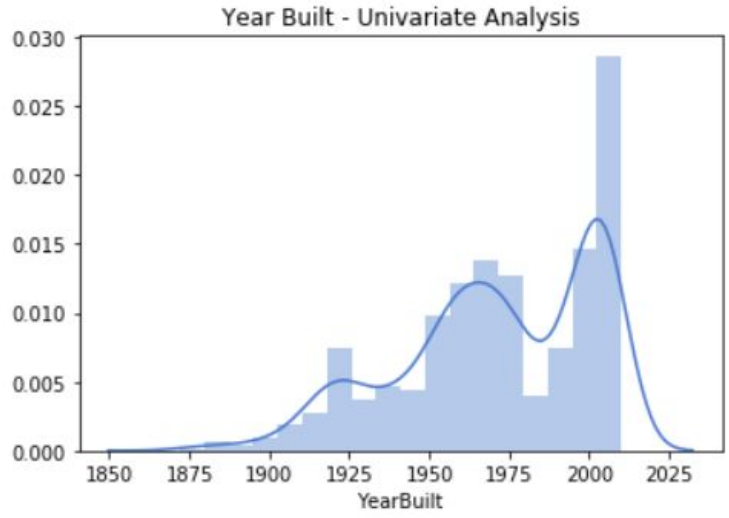
Data Exploration and Missing Values

There is no missing data for ‘Year Built’. The oldest house that was sold was built in 1872 and the newest one was built in 2010.

count	1,460.00
mean	1,971.27
std	30.20
min	1,872.00
25%	1,954.00
50%	1,973.00
75%	2,000.00
max	2,010.00

Univariate Analysis

About 26.5% of the houses were built in the 2000s, 72.5% of the houses were built in 1900s and about 1% of the houses were built in 1800s

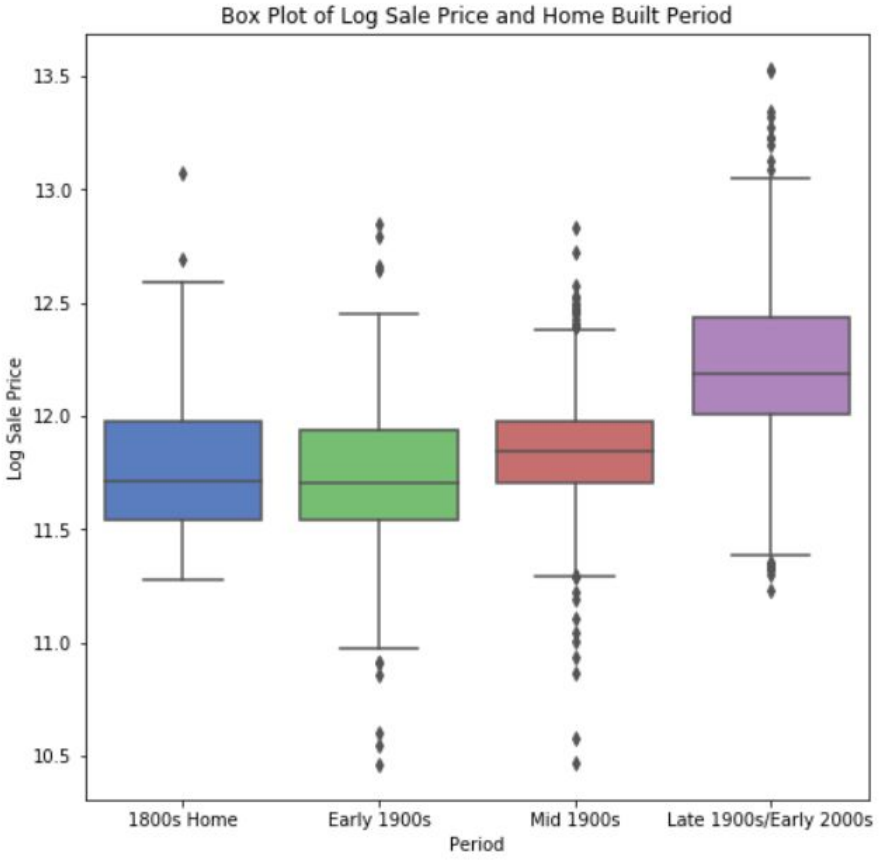


Feature Engineering & Relationship

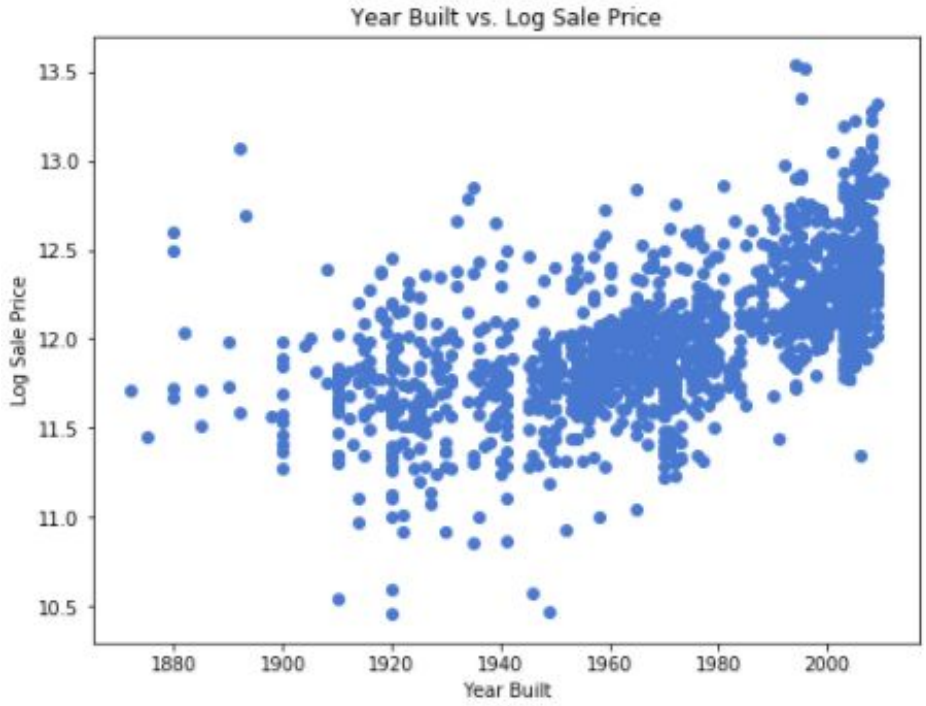
Year Built has been converted from a continuous variable into 4 Categories of periods:

- Built in the 1800s
- Built in the early 1900s (1900 - 1940)
- Mid 1900s (1940 - 1970)
- Late 1900s & Early 2000s (1970 - 2010)

The categories are plotted against Log of Sales Price. As expected, the homes built in the period 1970s to early 2000s sell at a higher price than the rest. There is variation in Sale Price shown in the chart below for homes built in different periods. Also shown below is the variation in median price of homes from 1900s to present.



Year Built was plotted against the Log of Sale Price. There is a medium correlation between Year Built and Sale Price.



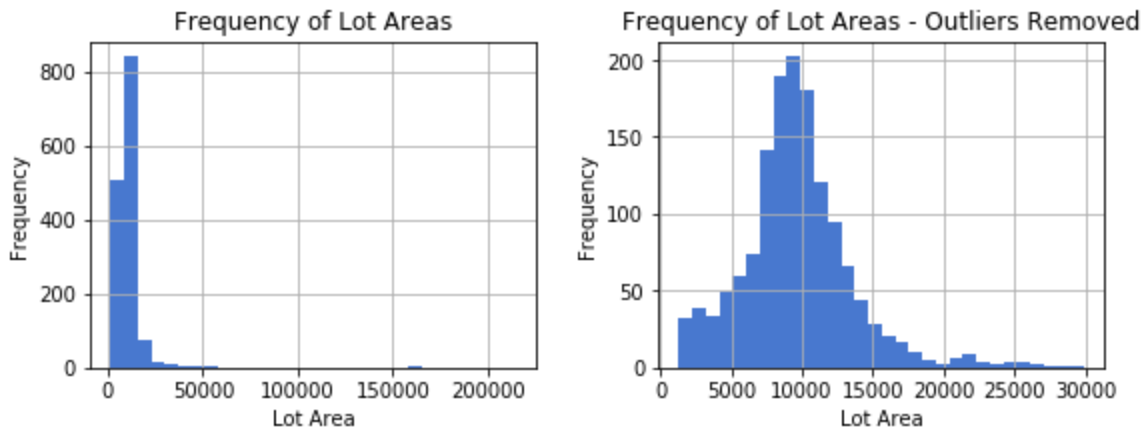
From the two charts above, it appears that there is not much difference in Sale Price between the newer homes and the homes built 15-20 years back. However, there is some variation when it comes to homes that are over 50 - 100 years old.

Lot Area

Lot area is the size of property that the home occupies. We would expect a positive correlation, but not necessarily a strong correlation as it doesn't indicate livable area and price per square foot of property can vary significantly geographically.

Data Exploration & Missing Values

A quick evaluation of the data shows a right skewed distribution with no missing data or mixed variable types. By removing outliers greater than two standard deviations above the mean, we can get a more roughly normal distribution of lot sizes.

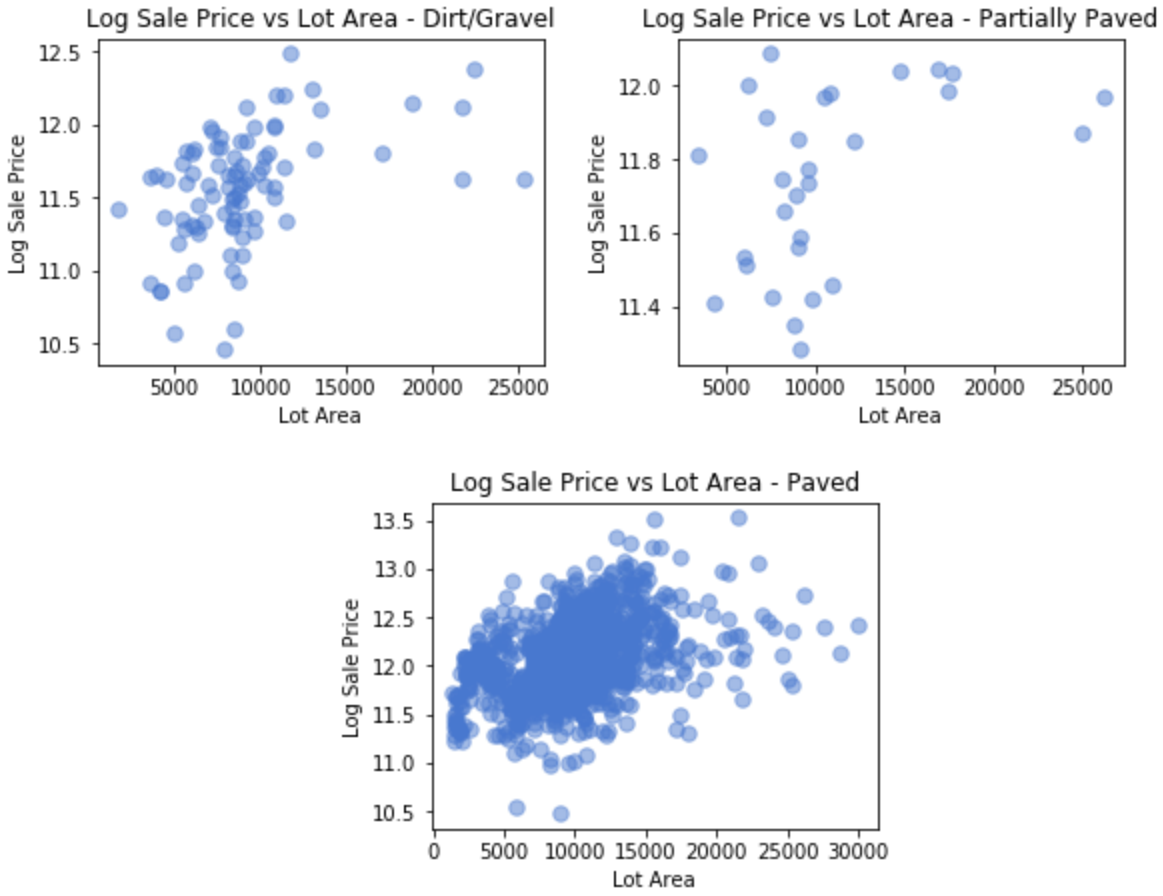


Feature Engineering and Relationships

A scatter plot of Log(Price) versus Lot Area shows a positive but fairly poor correlation (r = .26).

This plot illustrates that for a given Lot Area, there can still be a wide band of prices for the home. Thinking deeper about why this could be, we generated a hypothesis that in urban communities, price may be more strongly impacted by Lot Area as land is more scarce. In rural communities the opposite impact may be true. Thus, this combined plot may be muddying a deeper trend which is worth investigating.

The data doesn't split up rural versus urban explicitly, but we attempted to use whether or not the driveway is paved as a proxy. Dissecting by paved, partially paved, and unpaved yields the following scatterplots.



The associated correlation values are .39 for Dirt/Gravel, .55 for Partially Paved, and .26 for Paved. This improved a few correlations to medium strength, but the plots still indicate a fairly weak correlation and wide variation. Furthermore, the trend somewhat contradicts the hypothesis because it is the unpaved driveways (theoretically more rural geographies) that exhibit the stronger relationship. Due to the relatively small sample size for those data and the likely imperfect relationship between driveway paving and ruralness of the community, this deep-dive will not be investigated further in the later datasets.

Living Area

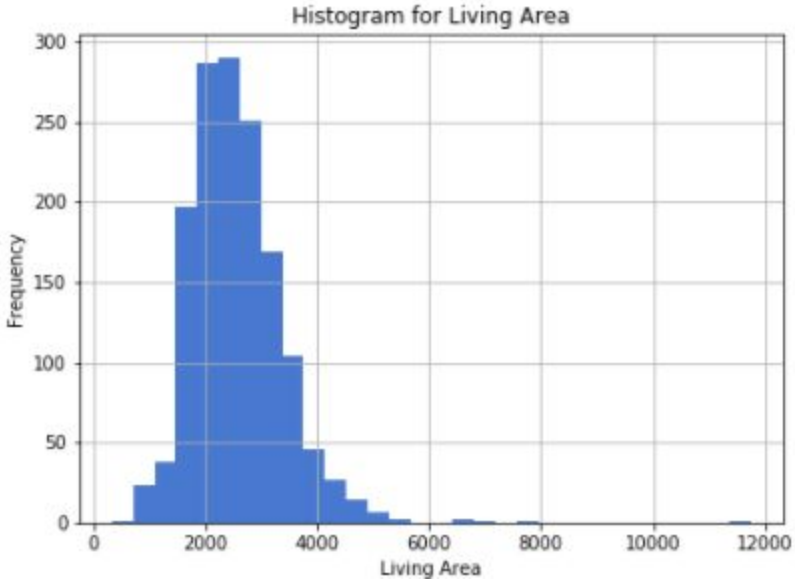
Data Exploration & Missing Values

The Iowa dataset does not have a single variable that corresponds to Living Area. There are two separate variables, one 'GrLivArea' and another 'TotalBsmtSF'. There are some missing

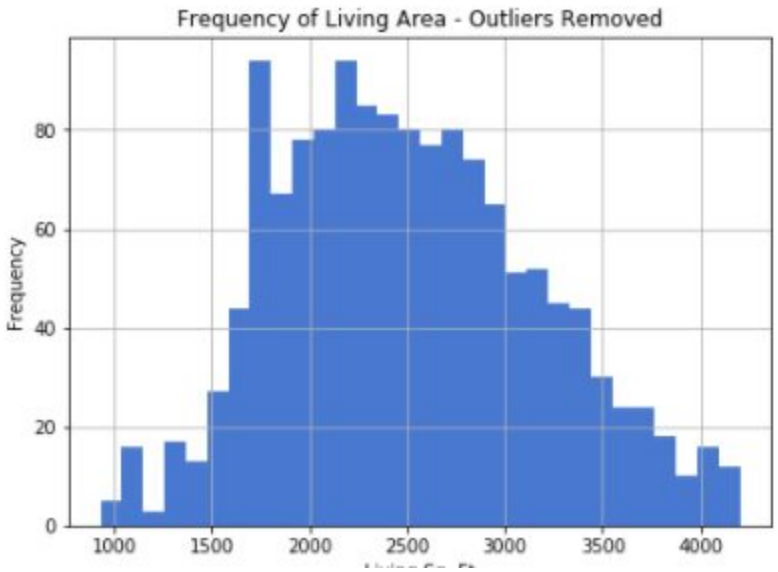
values in these fields. However, when summed together, no missing value was identified. It appears that some of the records had either ‘GrLivArea’ or ‘TotalBsmtSF’, not both.

count	1,460.00	count	1,460.00	count	1,460.00
mean	1,515.46	mean	1,057.43	mean	2,572.89
std	525.48	std	438.71	std	823.60
min	334.00	min	0.00	min	334.00
25%	1,129.50	25%	795.75	25%	2,014.00
50%	1,464.00	50%	991.50	50%	2,479.00
75%	1,776.75	75%	1,298.25	75%	3,008.50
max	5,642.00	max	6,110.00	max	11,752.00

There were some outliers found in the Living Area revealed in the histogram below. The Living area outside 926 - 4219 sq. ft were omitted for further plotting and analysis.



Here is the histogram after removing outliers.



Feature Engineering and Relationships

High correlation (0.77) between Living Area and Sale Price is shown in the Iowa DS.

Bed/Bath

Bed/bath metrics are key data points when marketing a house. Through exploring our data set we discovered 4 houses that had 0 bedrooms, most likely in error. Attempts to engineer

a bed bath feature, similar to how houses are colloquially talked about, failed to show stronger correlation than was observed between bathrooms and log sale price. Interestingly, bedrooms were not as useful in predicting sales price as bathrooms or our engineered features.

Data Exploration & Missing Values

In exploring the bedroom dataset, four houses were identified with 0 bedrooms. These houses had all other fields logically filled out in terms of sale value, square footage, bathrooms and other key features. If we had access to where this data originated we would have ideally gone back and confirm if the bedroom count was recorded correctly. Since we don't have this access, we decided to remove these 4 data points from the data set.

Looking at bed/bath listings for the houses we observed that while most of the houses fell into standard configurations there were some outliers (8/2 and 3/6 among others). Our hypothesis was that by combining bath and bed information we would be able to identify unique trends missed by bed or bath alone.

BedroomAbvGr bath	1	2	3	4	5	6	8
1.0	8	108	105	5	2	0	0
1.5	2	23	86	17	0	0	0
2.0	22	131	233	56	9	3	1
2.5	12	26	189	61	3	1	0
3.0	5	57	98	21	2	2	0
3.5	1	12	87	41	3	0	0
4.0	0	1	3	8	1	0	0
4.5	0	0	2	3	1	1	0
5.0	0	0	0	1	0	0	0
6.0	0	0	1	0	0	0	0

Crosstab table of Bedrooms vs Bathrooms

Feature Engineering

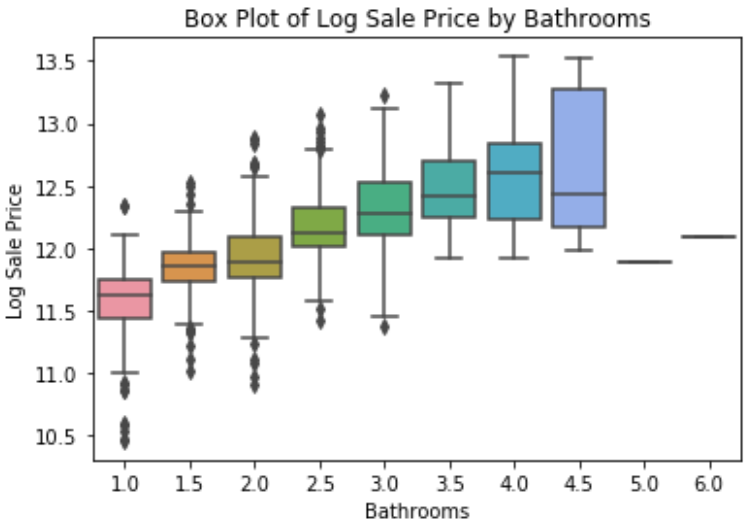
3 new features were calculated to aid in analysis:

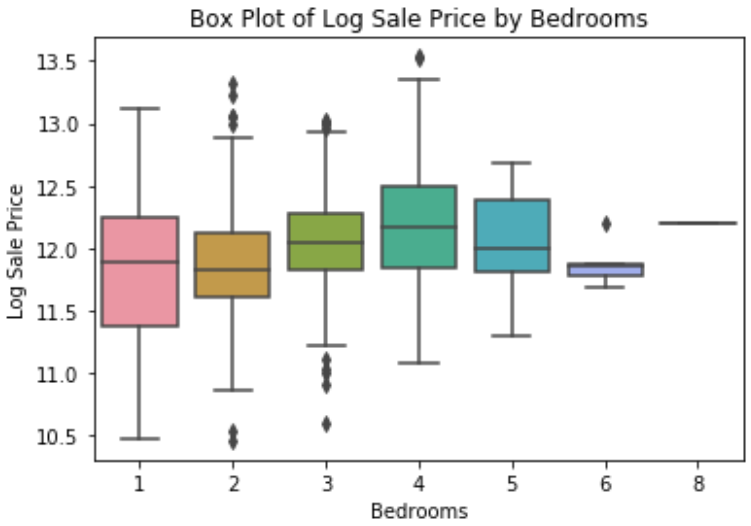
- Total Bath
    - There are several columns in the Iowa data set that serve as binary variables. We combined these into a single discrete column of bathroom count in 0.5 increments.
- Bed/Bath Tuple
    - Houses are typically talked about and listed in terms of their Bed/Bath numbers. To emulate this colloquial use we created a (bed, bath) tuple
- Bed/Bath Sum

- Tuples have no intrinsic order and we hypothesized that we could capture any information gain by summing beds and baths for a house.

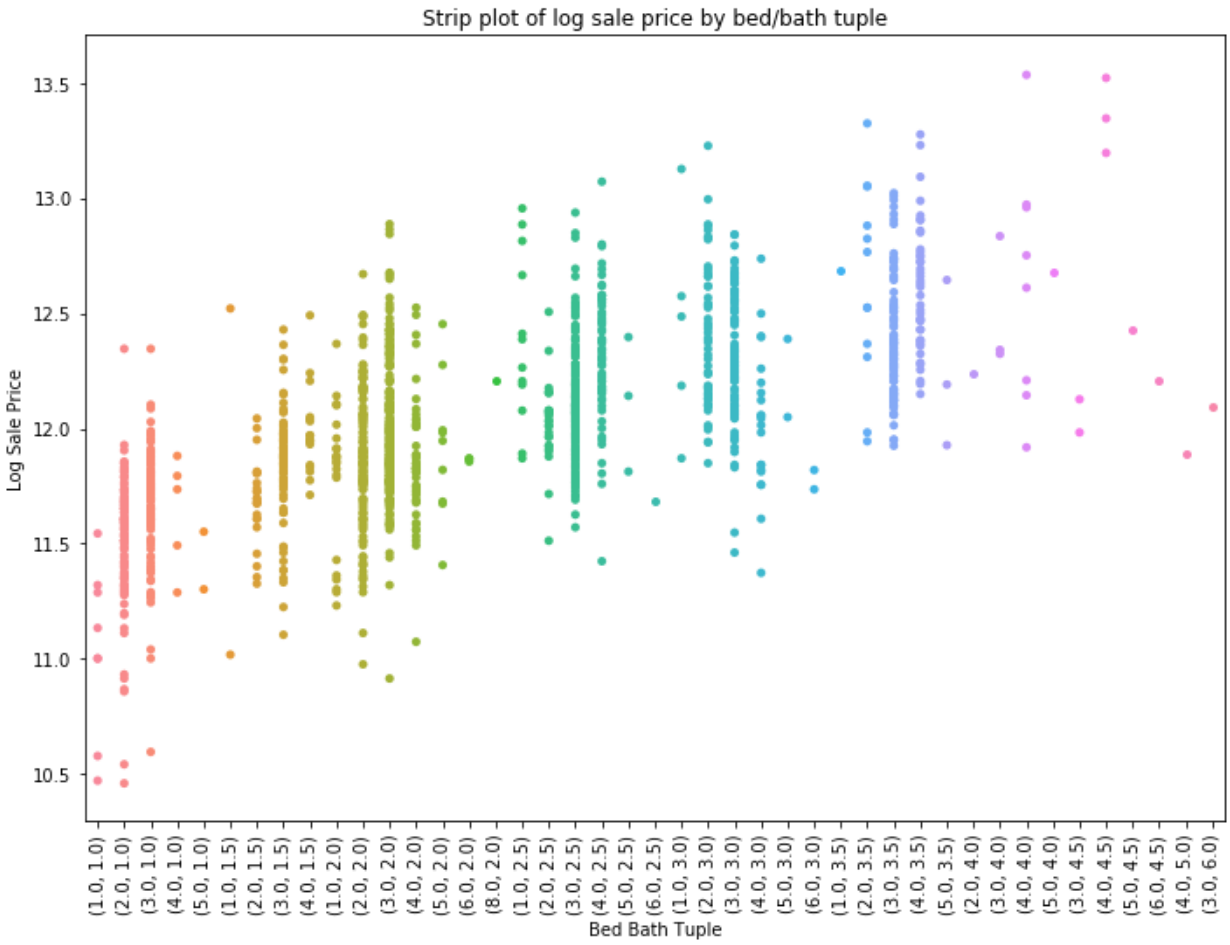
Relationships

Beds and baths were first evaluated independently. A strong correlation is observed between log sale price and bathrooms. The relationship between bedrooms and log sale price is not as apparent or strong.

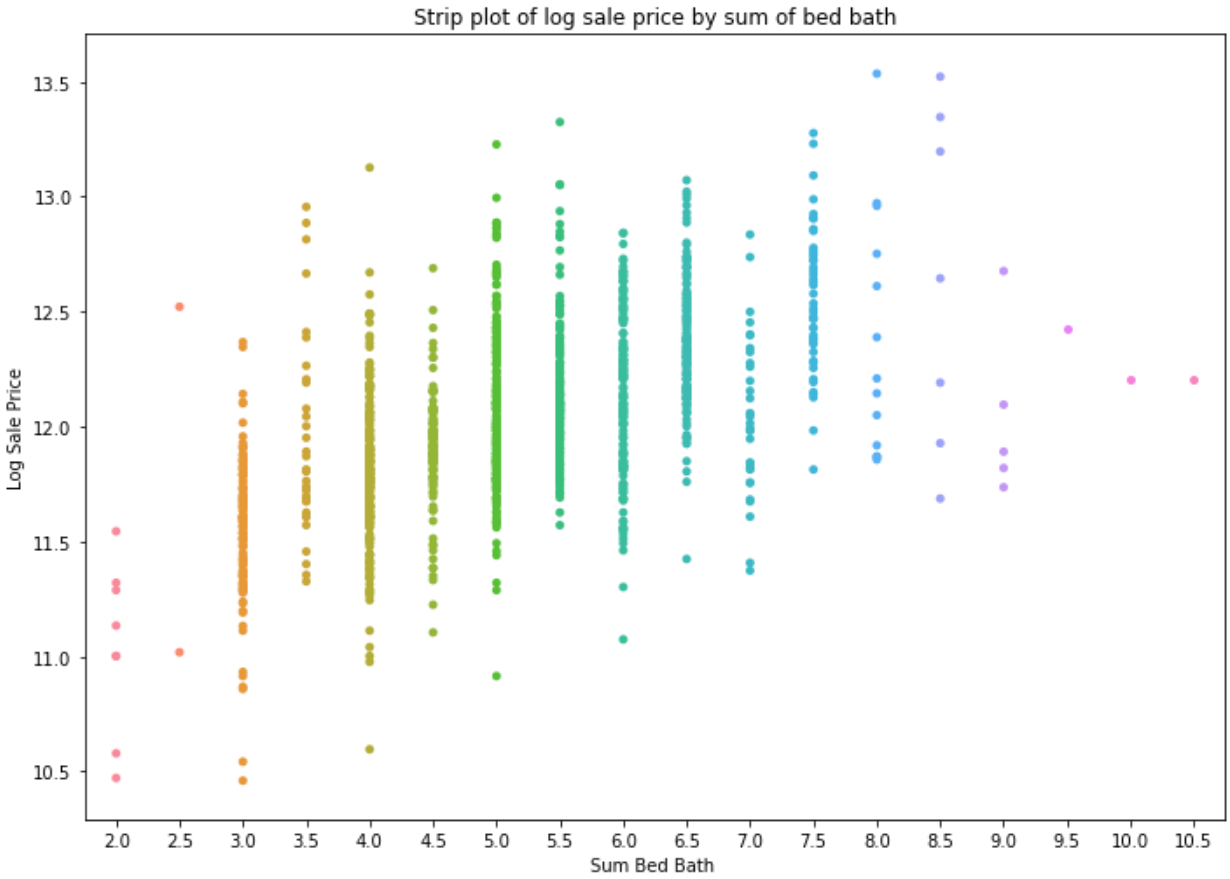




Our hypothesis was that combining bed and baths into a tuple we would be able to capture the way houses are typically listed (ex. 3 bed / 2 bath). Given the trend in bathrooms, we expected that an increase of 1 bedroom would increase the sale price. The below strip plot shows that an increase in bedroom, holding bathroom constant, does not always lead to an increase in price.



Given the unordered nature of tuples, we next tested the idea of summing bedrooms and bathrooms. This metric would be more easily leveraged in a future model and ideally capture the bed/bath affect better than either independently. The strip plot below shows a strong correlation with an dip at 7. Quantified, the correlation between summed bed baths and log sale price is 0.56.



Ultimately both attempts to engineer a bed/bath feature resulted in lower correlation than observed between bathrooms and log sale price of 0.67. This indicates that looking at the relationship between bathrooms and sales price may be more valuable than including number of bedrooms in a model.

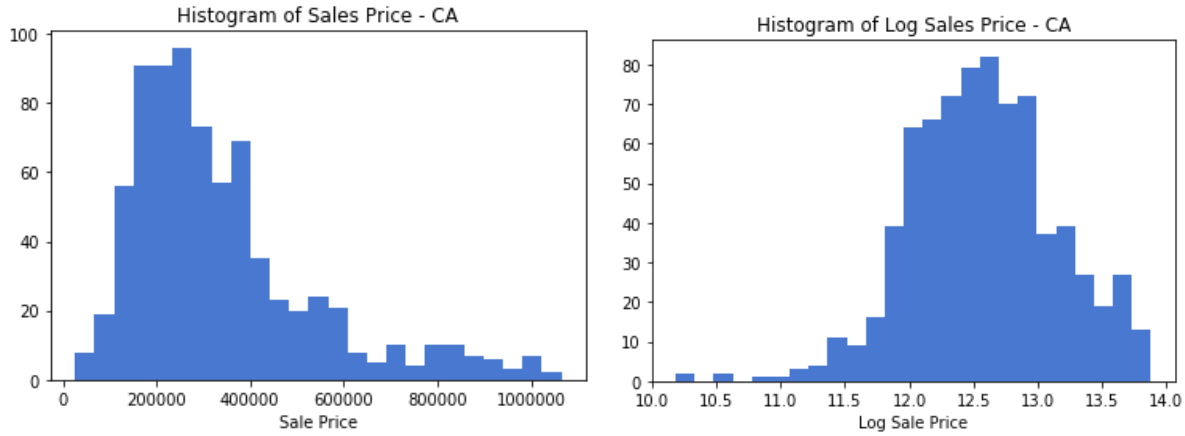
Confirming the Trends

To validate our findings from the Iowa data set we decided to compare our insights to two independent data sources, one from California and one from North Carolina.

Data Exploration

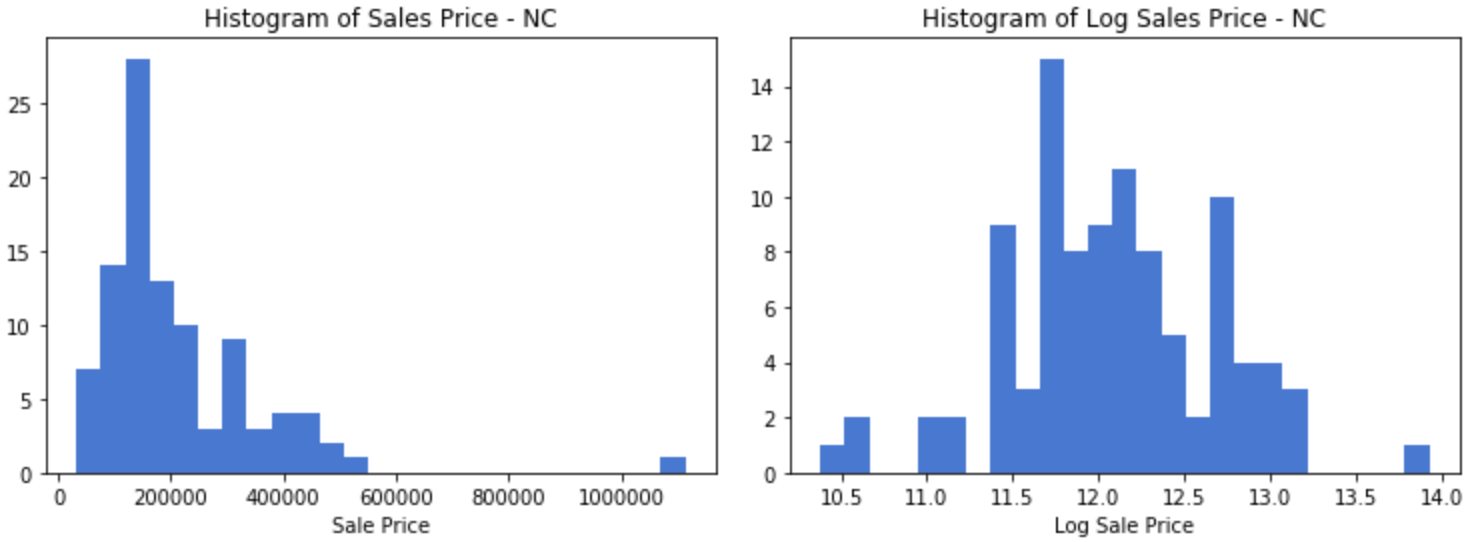
Before comparing our insights, we performed an EDA on both data sets.

California



The house sale prices in the CA dataset had a right skew. Applying a log transform to the data helped to remove the skew.

North Carolina



The house sale prices in the NC dataset had a right skew. Applying a log transform to the data helped to remove the skew.

Missing Values and Outliers

California

Via the EDA it was discovered that there were two houses with 0 bedrooms. Ideally we would be able to go back to the original data source and confirm this was not recorded in error. Because we can't do that currently, these two houses were removed from the data set. Additionally, there were a few houses with extremely high sale prices. Our Iowa data set had none comparable and these CA prices were over two standard deviations from the mean. These houses were also removed for the combined analysis.

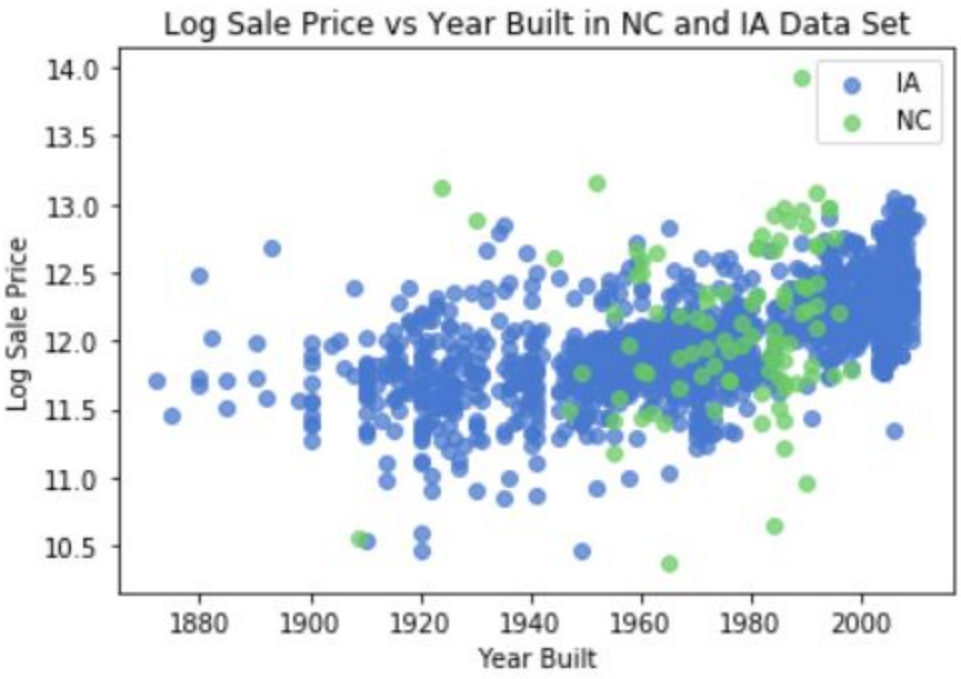
North Carolina

Via the EDA it was discovered that there were three houses with no recorded lot size. Ideally we would be able to go back to the original data source and confirm this was not recorded in error. Because we can't do that currently, these three houses were removed from the data set. Additionally, there was one house with an extremely high sales price. Our Iowa dataset had none comparable and the house was over two standard deviations from the mean. The house was removed from the combined analysis.

Do the Trends Persist?

Year Built

Only two datasets (IA and NC) out of the three have Year Built. Plotting these two together it appears that trend does not persist. The Iowa Dataset is showing a medium correlation (0.59) with Log of Sale Price. However, NC dataset shows very little correlation with Log of Sale Price (0.17)



Lot Area

Both the Iowa and North Carolina datasets have Lot Area. Plotting the two together yields the following.

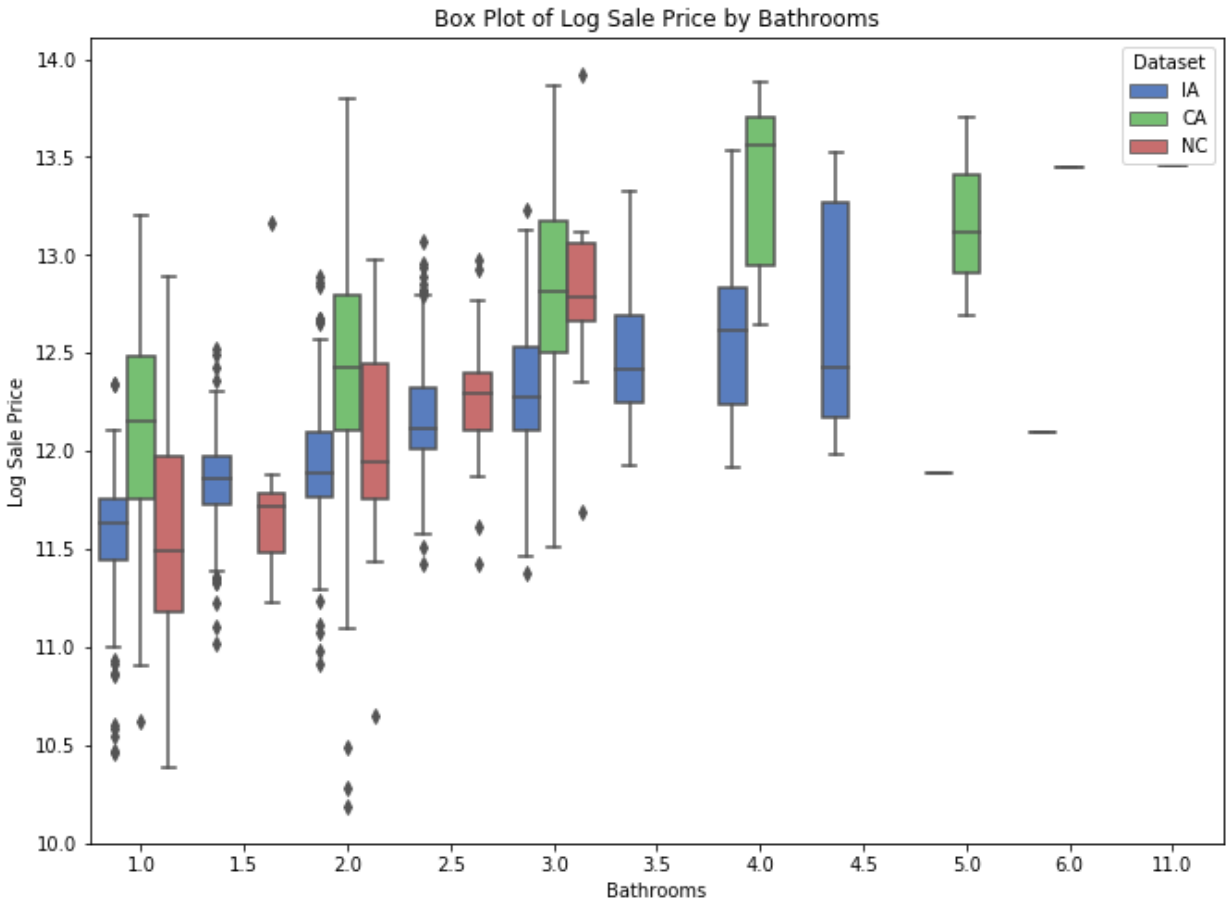
This confirms the (low) trend seen in the Iowa dataset. The data appears to be more of a cloud than a firm trend, and the datasets occupy roughly the same bounds in space. North Carolina is slightly less correlated at r = .20.

Living Area

The high correlation seen in the IA dataset persists in the other two datasets for Living Area. The NC and CA datasets have a similar trend showing high correlation between Living Area and Log Sale Price. The correlation is strongest in the NC dataset (0.81) followed closely by IA (0.80). CA dataset showed the least correlation of the three (0.63).

Bathrooms

Replotting our Iowa data we see that the observed correlation persists in both the California and North Carolina data sets. Additional things to note, house prices are higher in California and the California data sets does not have any half bathrooms.



Conclusion

This analysis showed that house price can be correlated to several of the explored factors, namely square footage, bathroom count, and year built. However, the correlations to bedrooms and lot area were much weaker. We also determined that these trends persisted across multiple geographies, with the exception of year built which had a large range in correlation. Although the correlations persisted, it is also worth noting that the different geographic regions had their own peculiarities, and scatter plots illustrated that the slopes of the

relationships differed. Similar methods could be used on the unstudied variables in each dataset to determine what other factors influence home price.