EN 605.42762 Data Visualization Project #3

Sophie Sackstein

Overview

Dean De Cock created the Ames Housing dataset for use in data science teaching. It's a fantastic option for data scientists seeking for a more updated and enlarged version of the Boston and California Housing dataset. The study's data includes property prices in the Greater Bristol region from 2001 to 2013. To make the models appropriate for national use, secondary data was gathered from the Land Registry, the Population Census, and Neighborhood Statistics.

The authors cited a variety of reasons for choosing the Greater Bristol area, including its unique urban and rural mix, as well as its various housing kinds. Each entry in the collection provides information about a home in the region, including the location, unit postcode, property type, tenure (freehold or leasehold), selling price, sale date, and whether the house was freshly constructed at the time of sale. The dataset has about 65,000 items in total. I chose 10 variables for testing, based on variable importance and least null values for this project. The data set includes the average lot size and total residence square footage seen on most frequent house listings. Room measures in the basement, main living area, and even porches are classified according to quality and kind. These 20 continuous variables for each observation correspond to different area dimensions.

The 14 discrete variables quantify the number of items attached/included in the house. The data set includes the number of kitchens, bedrooms, and bathrooms (full and half) located in the basement and above grade (ground) living areas of the home. The garage capacity and construction/remodeling dates are also recorded. There are a large number of categorical variables (23 nominal, 23 ordinal) associated with this data set.

The nominal variables are used to identify different types of homes, garages, materials, and environmental conditions, while the ordinal variables are used to assess different aspects of the property. The original data used an eight-character name that was important to the categorization, although several of the original class levels were difficult to comprehend. Many class levels have been recoded into more useable forms for convenience of usage (see the documentation file https://www.openml.org/d/42165).

The Variables

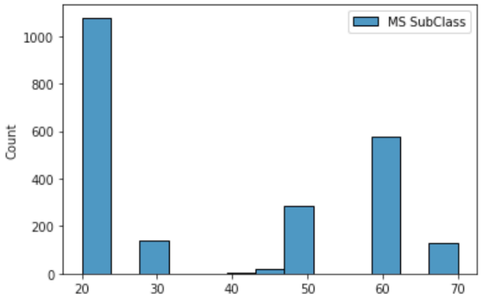
REFERENCE: https://www.openml.org/d/42165

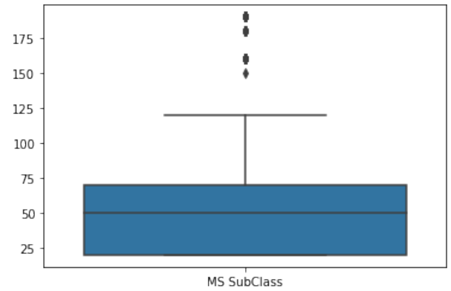
```
Data columns (total 11 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   MS SubClass    2930 non-null   int64
 1   MS Zoning      2930 non-null   object
 2   Year Built    2930 non-null   int64
 3   Lot Area      2930 non-null   int64
 4   Neighborhood  2930 non-null   object
 5   House Style   2930 non-null   object
 6   Overall Cond  2930 non-null   int64
 7   Full Bath     2930 non-null   int64
 8   Gr Liv Area   2930 non-null   int64
 9   Yr Sold       2930 non-null   int64
 10  SalePrice     2930 non-null   int64
```

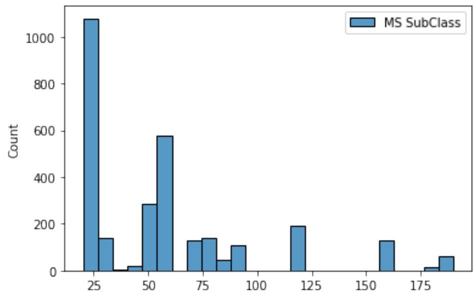
1. MSSubClass: Identifies the type of dwelling involved in the sale. (Nominal, Categorical)

20	1-STORY 1946 & NEWER ALL STYLES
30	1-STORY 1945 & OLDER
40	1-STORY W/FINISHED ATTIC ALL AGES
45	1-1/2 STORY - UNFINISHED ALL AGES
50	1-1/2 STORY FINISHED ALL AGES
60	2-STORY 1946 & NEWER
70	2-STORY 1945 & OLDER
75	2-1/2 STORY ALL AGES
80	SPLIT OR MULTI-LEVEL
85	SPLIT FOYER
90	DUPLEX - ALL STYLES AND AGES
120	1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150	1-1/2 STORY PUD - ALL AGES
160	2-STORY PUD - 1946 & NEWER
180	PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190	2 FAMILY CONVERSION - ALL STYLES AND AGES

Most houses are 1-story, and also many are newer (<1945), so the “distribution” is skewed towards that category. This is categorical variable is also split further into style and age type, which are arguably also categorical. One interesting note here is that while there’s a big difference between the ages types (pre- and post-1945) in 1-story styles, there’s less of a difference for 1.5 story, and some don’t have pre-1945, like split foyer and PUD, 2-family conversion and duplexes. Perhaps those styles didn’t exist in Ames, Iowa during that time? According to the Atomic Heritage Foundation (https://www.atomicheritage.org/location/ames-ia) the Ames project employed thousands of new workers and new architects for Iowa campus lab and uranium storage facility. This may have seen in an in-flux of 1950’s style pre-fab houses (https://behost.lib.iastate.edu/DR/Dikis_NA730.I8-D569i.pdf) that were mostly single story ramblers. I show the zoomed in histogram on the left, for clear pre-1945 to post-1945 differences.







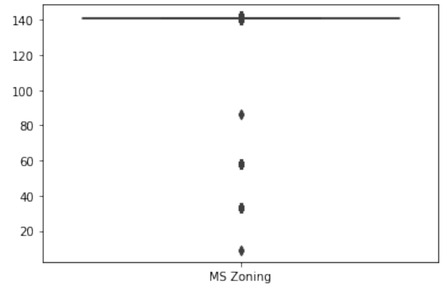
2. MSZoning: Identifies the general zoning classification of the sale (Nominal, Categorical Non- Numeric)

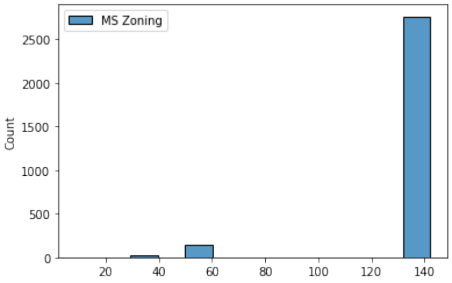
(ONLY including R and FV in later analysis)*

A Agriculture
C Commercial
FV Floating Village Residential
I Industrial
RH Residential High Density
RL Residential Low Density
RP Residential Low Density Park
140- RM Residential Medium Density

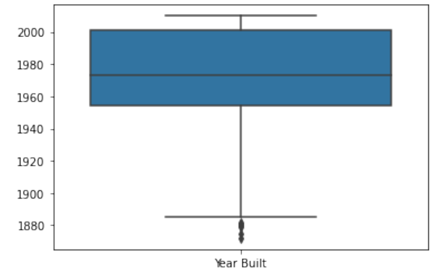
```
{9.0: 'RL',
 33.0: 'RH',
 58.0: 'FV',
 86.0: 'RM',
 140.0: 'C (all)',
 141.0: 'I (all)',
 142.0: 'A (agr)'}
```

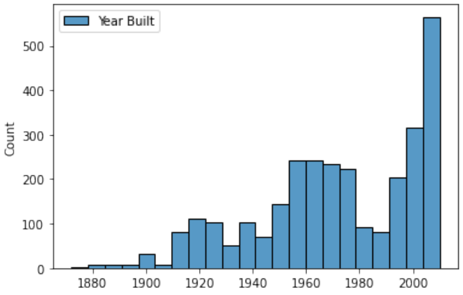
Most are skewed the right, being 140- RM, residential medium density, which makes sense given the above explanation of the 50’s era suburban growth in Iowa.





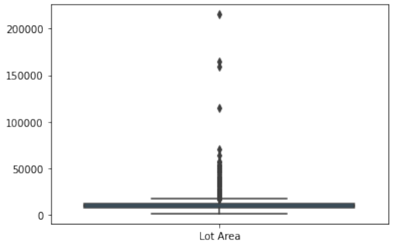
3. YearBuilt: Original construction date (Quantitative , Interval)

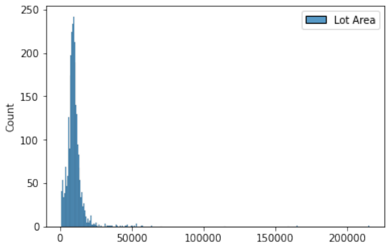




4. LotArea: Lot size in square feet (Quantitative, Ratio)

This is a quantitative variable of type ratio. In theory, the value of a home could be zero (mobile or floating home), but it’s unlikely to ever happen and its evident in the data that few cases exists. The distribution of the histogram appears unimodal with a skew towards the left.

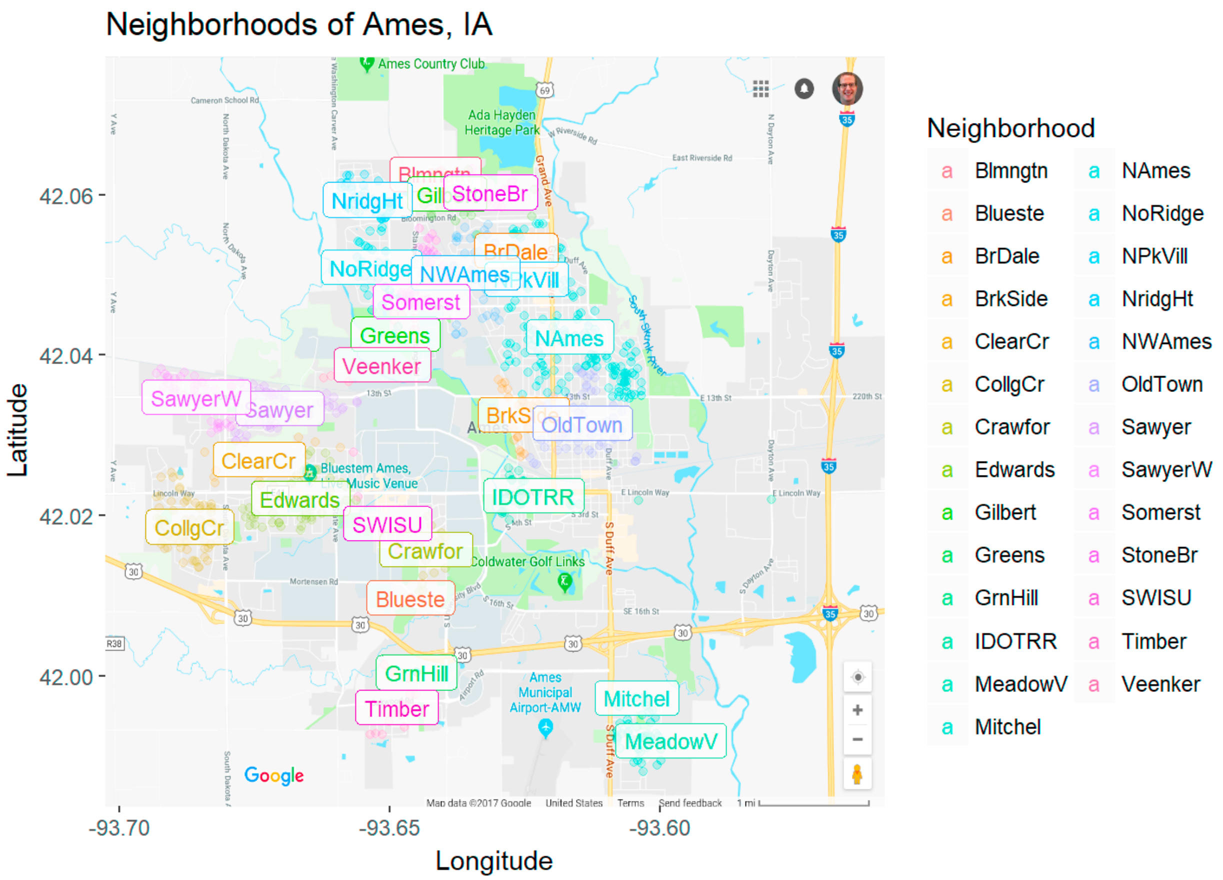


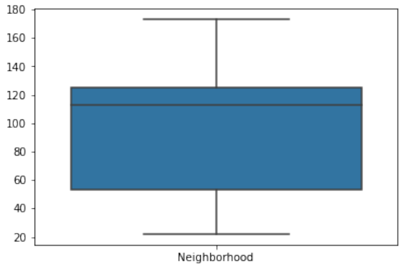


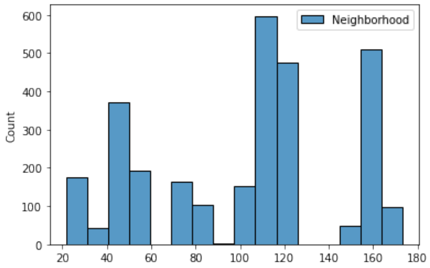
5. Neighborhood: Physical locations within Ames city limits (Nominal, Categorical Non-Numeric)

```
{22.0: 'NAmes',23.0: 'Gilbert',25.0: 'StoneBr',30.0: 'NWAmes',39.0: ‘Somerst',
42.0: 'BrDale', 49.0: ‘NPkVill', 53.0: 'NridgHt', 77.0: 'Blmngtn', 79.0: 'NoRidge',
 80.0: 'SawyerW',87.0: 'Sawyer',93.0: 'Greens',100.0: 'BrkSide',
 106.0: 'OldTown',113.0: 'IDOTRR',114.0: 'ClearCr',
 115.0: 'SWISU',118.0: 'Edwards',124.0: ‘CollgCr', 125.0: 'Crawfor',
 153.0: 'Blueste',155.0: 'Mitchel',156.0: 'Timber',160.0: 'MeadowV',
 162.0: 'Veenker',167.0: 'GrnHill',173.0: 'Landmrk'}
```

Most are located near city center with small clusters around the airport. It can also consider arbitrary instead of categorical as the number of towns increases.
image source:https://rstudio-pubs-static.s3.amazonaws.com/337439_24918eaefe724411be93e41ede48b256.html



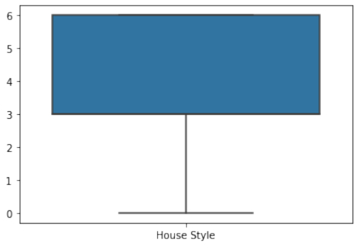


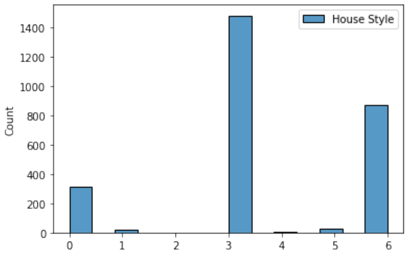


6. HouseStyle: Style of dwelling (similar to the first, Nominal, Categorical)

```
{0.0: ‘1 Story',
 1.0: ‘2 Story',
 3.0: ‘1.5 Finished',
 4.0: ‘1.5 Unfinished',
 5.0: ‘2.5 Finished',
 6.0: ‘2.5 Unfinished'}
```

Most are one story or split foyer, and some have an unfinished second level. Cleary Ames is skewed towards 1-story and 1.5-story houses, with more 1.5 unfinished and 2.5 unfinished houses then finished ones.

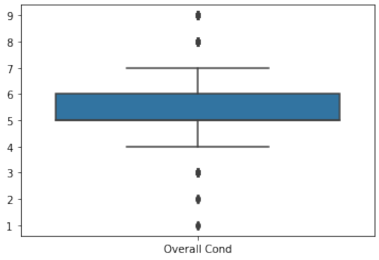


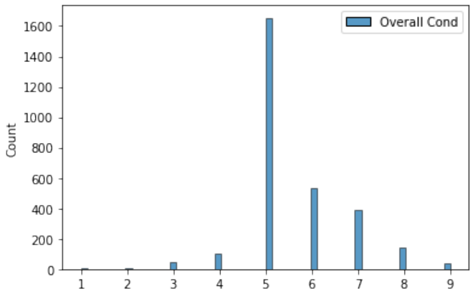


7. OverallCond: Rates the overall condition of the house (Ordinal, Numeric)

10 Very Excellent
9 Excellent
8 Very Good
7 Good
6 Above Average
5 Average
4 Below Average
3 Fair
2 Poor
1 Very Poor

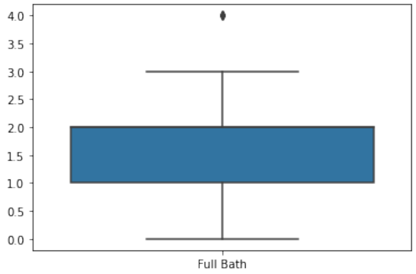
Most are in average, tending towards excellent.

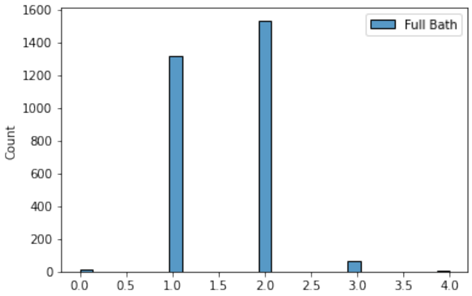




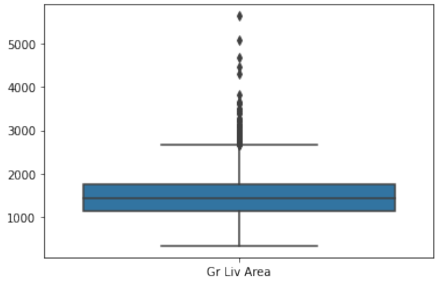
8. FullBath: Full bathrooms above grade (quantitative, ratio)

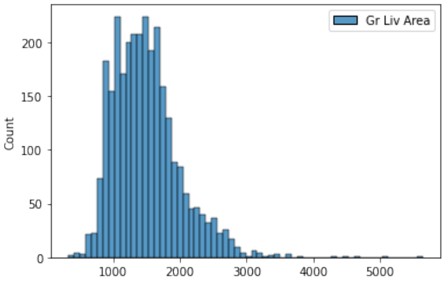
This variable seems to indicate that each record exists within a defined address and each address has a discrete number of full bathrooms. It is a quantitative variable also of type ratio.



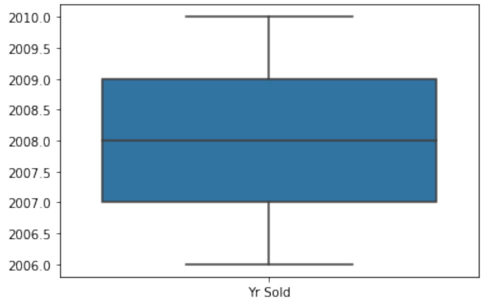


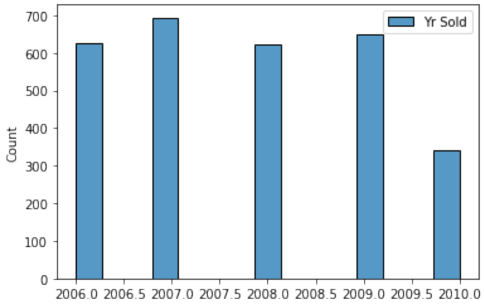
9. GrLivArea: Above grade (ground) living area square feet (qualitative, ratio)
This is a quantitative variable that is a ratio. This variable is a ratio since the value zero is meaningful in that it would indicate that the house has no above grade area.





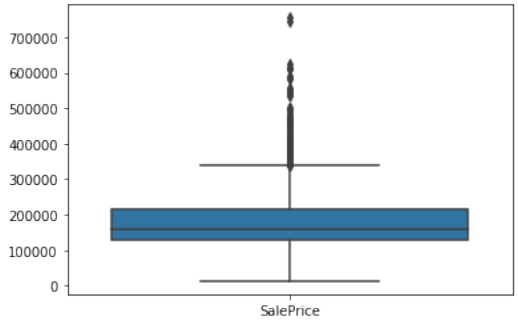
10. YrSold: Year Sold (YYYY) (Quantitative, Interval)
This dataset only covers 2006-2013, and it seems like its pretty uniform, besides the dip in 2010.

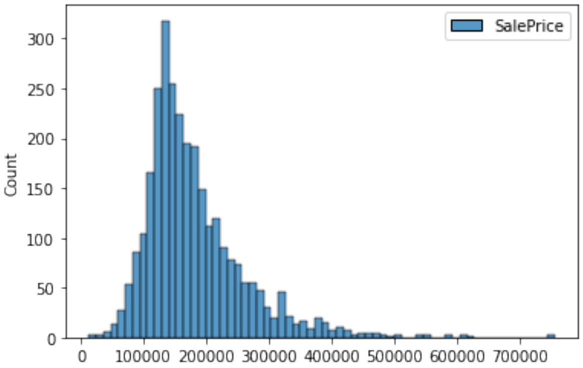




11. SalePrice (TARGET for Study): final price of each homeData: SalePrice (Quantitative, Ratio)
This is a quantitative variable of type ratio. In theory, the value of a home could be zero, but it's unlikely to ever happen and it's evident in the data that no such case exists. The distribution of the histogram appears unimodal with a skew towards the right. However, at the right end of the

distribution is a small spike showing that there is a density of extreme values towards the higher end. The boxplot shows a similar picture with values centered in an area, and some outliers towards the end. The median of this data is 180,796 which represents the main peak in the distribution of the histogram. The outliers in the boxplot seem to be around 300,000. There's possibly some cap on this value, as the max is also 755,000, identical to the maximum value. Then any areas with median values above that would all be lumped into this value.





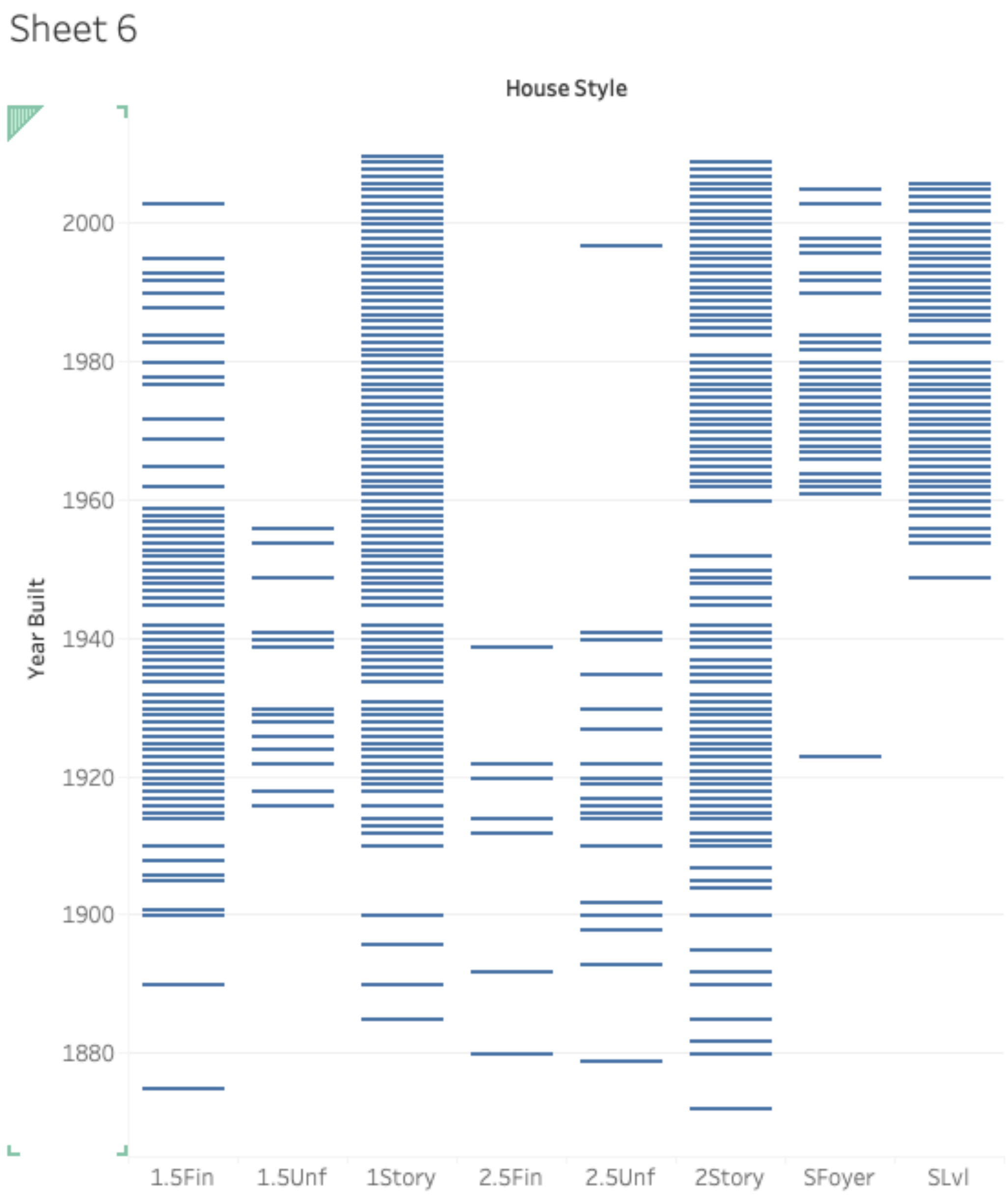
	mean	std	min	50%	max
MS SubClass	57.387372	42.638025	20.0	50.0	190.0
MS Zoning	136.161775	20.425412	9.0	141.0	142.0
Year Built	1971.356314	30.245361	1872.0	1973.0	2010.0
Lot Area	10147.921843	7880.017759	1300.0	9436.5	215245.0
Neighborhood	102.713311	43.090990	22.0	113.0	173.0
House Style	14.280205	38.305705	0.0	3.0	152.0
Overall Cond	5.563140	1.111537	1.0	5.0	9.0
Full Bath	1.566553	0.552941	0.0	2.0	4.0
Gr Liv Area	1499.690444	505.508887	334.0	1442.0	5642.0
Yr Sold	2007.790444	1.316613	2006.0	2008.0	2010.0
SalePrice	180796.060068	79886.692357	12789.0	160000.0	755000.0

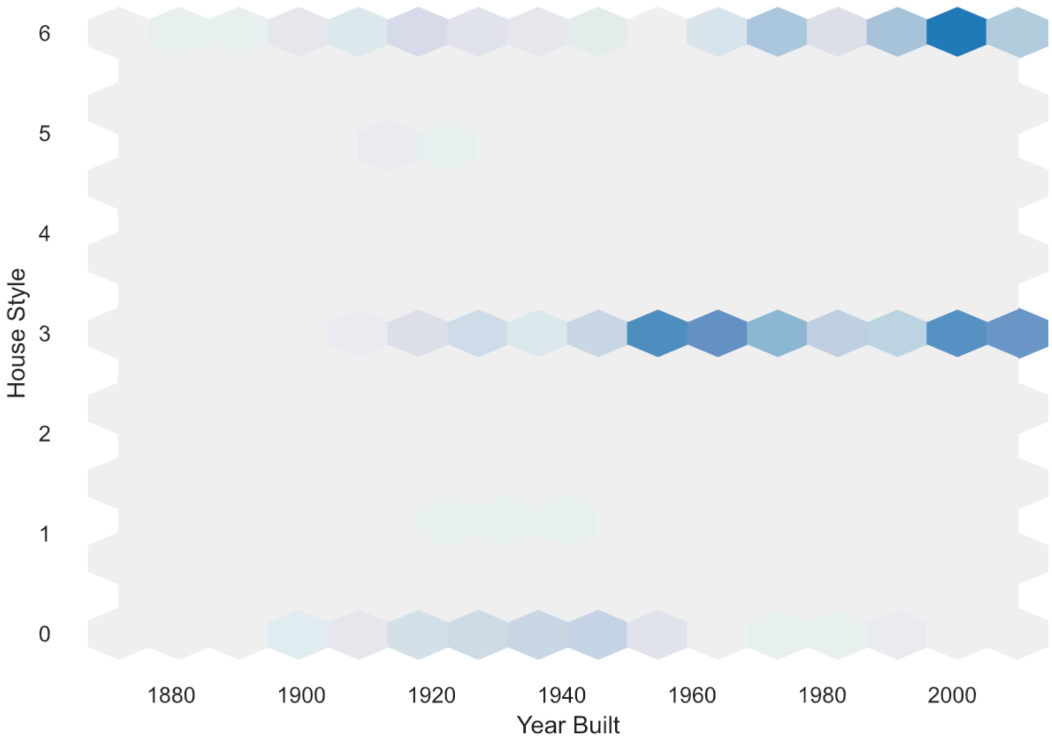
Questions:

1. How did house style change over the years?

```
{0.0: ‘1 Story',
 1.0: ‘2 Story',
 3.0: ‘1.5 Finished',
 4.0: ‘1.5 Unfinished',
 5.0: ‘2.5 Finished',
 6.0: ‘2.5 Unfinished'}
```

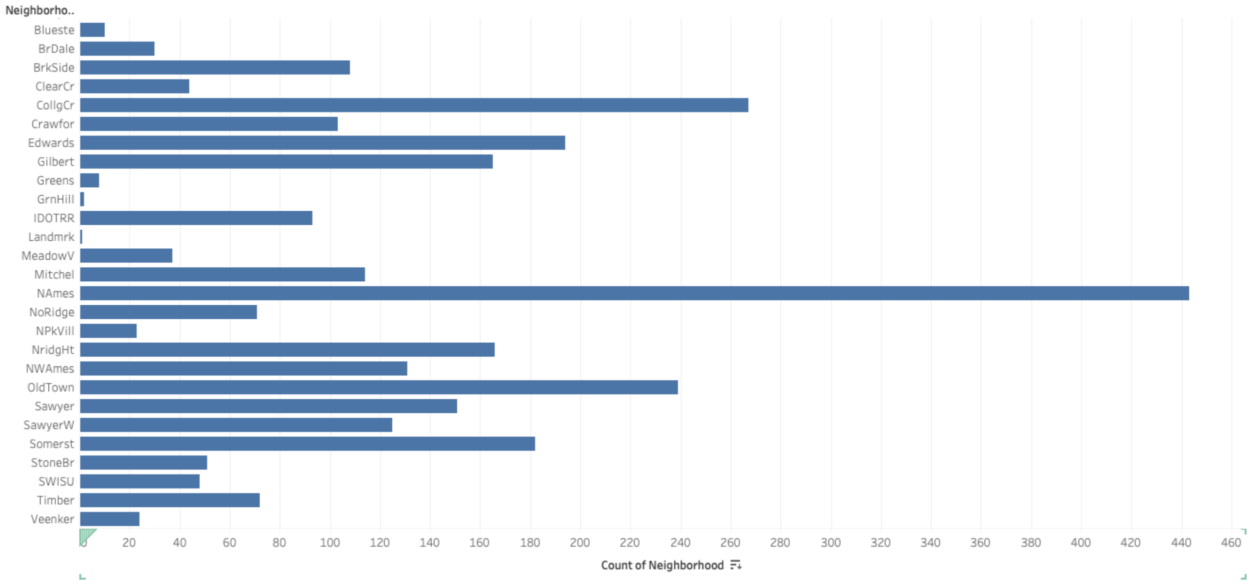
As expected, there was a boom in 1-1.5 story houses around 1950, I used color hex-bins to show this since it’s a bit easier to read than a scatter pairplot, and it shows trends (like increase in 2.5 story unfinished houses the 2000s) well.





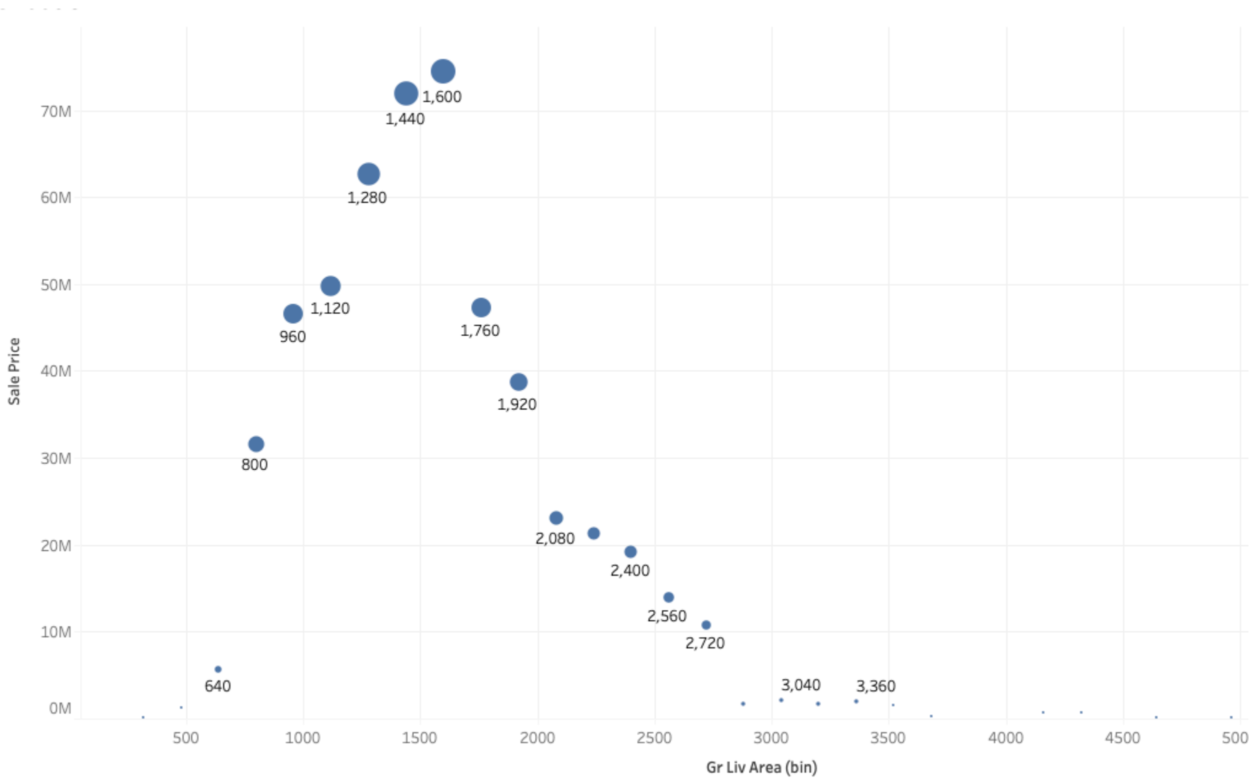
2. What area are most records concentrated in?

The first question to be analyzed with visualizations is, “What area are most records concentrated in?” The most straightforward method was to plot per neighborhood counts for each.

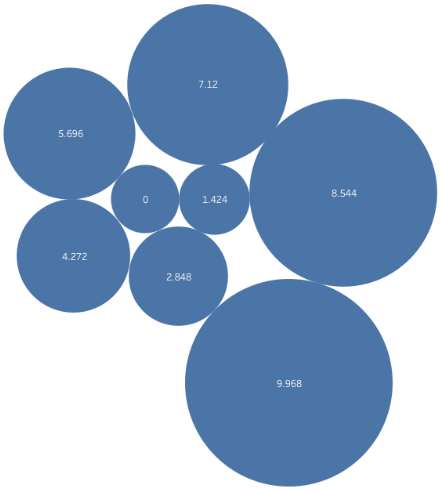


3. Whats the relationship between sales price and living area?

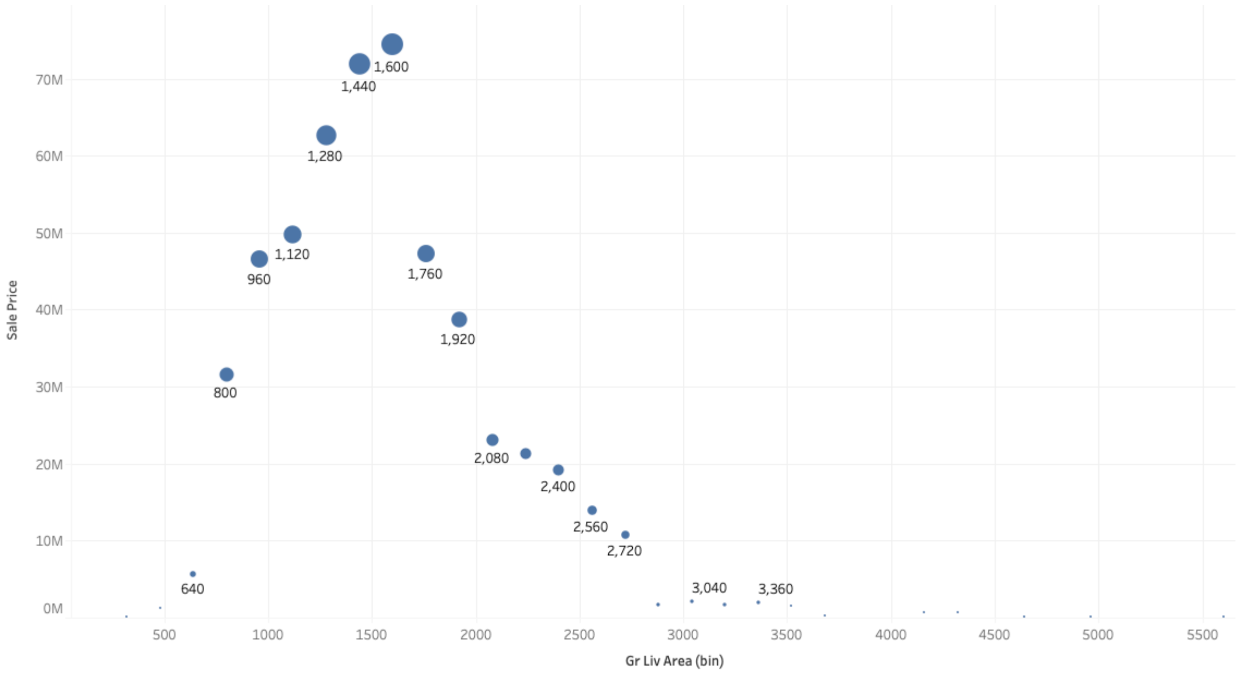
We can see that the above-ground living area falls approximately between 800 and 1800 ft2. Now, let us see the relationship between Gr Liv Area and the target variable:
The scatter plot above shows clearly the strong positive correlation between Gr Liv Area and SalePrice verifying what the trend I hypothesized.

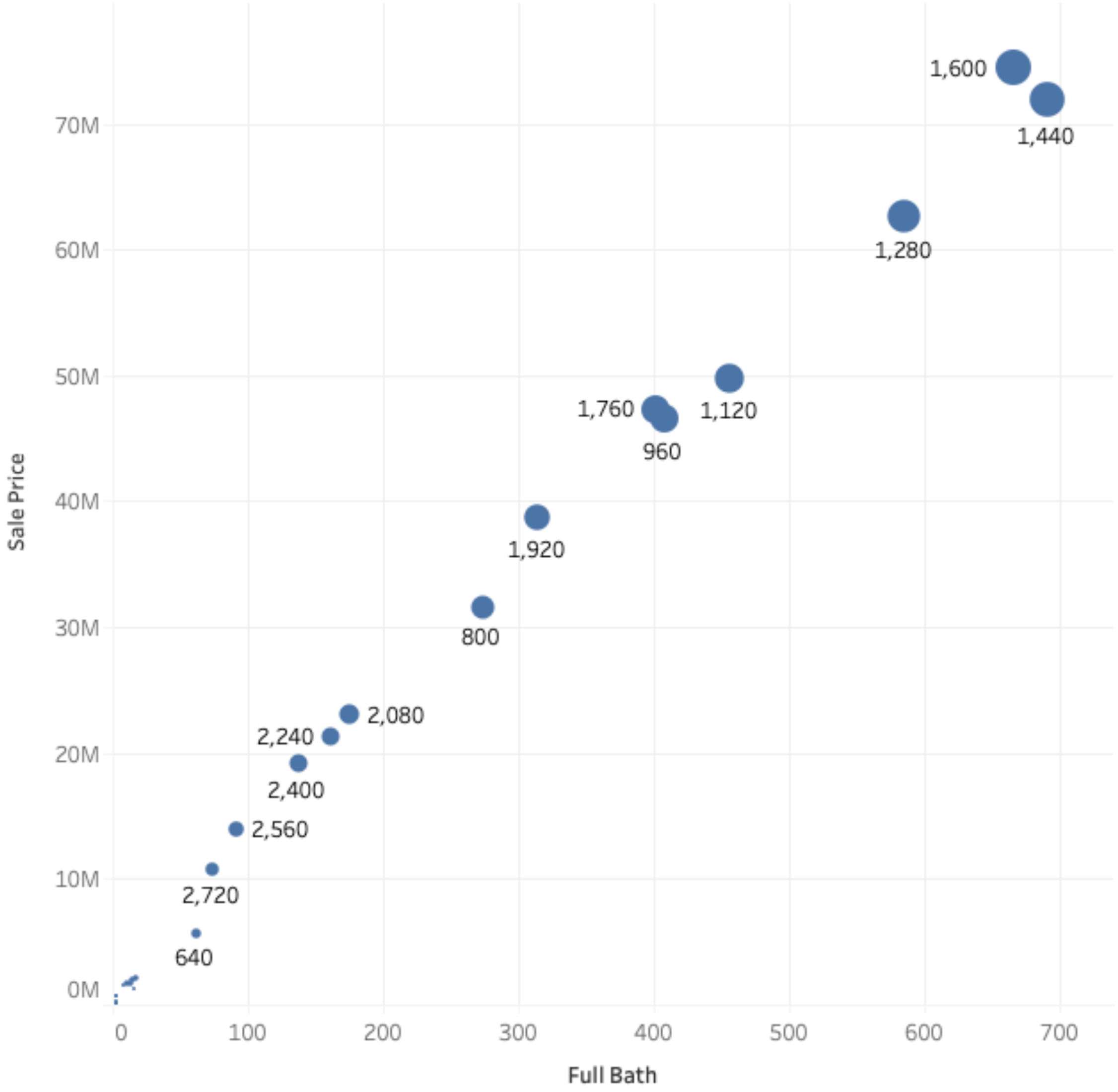


4. **How does overall quality influence price?** Overall Qual is an integer number between 1 and 10, and the majority of houses have an overall quality of 5 to 7. To illustrate the link between SalePrice and Overall Qual, we plot the circle plot:



5. **Are larger homes in better condition? Is the condition also effected by the number of bathrooms?** Yes! Condition is effected by the number of full bathrooms, with higher concentrations of 2 bathroom houses correlated with above average to excellent condition. This is also the case with the above grade living area, but it tapers off at around 2500 sq ft, maybe it’s too much to maintain for a home-owner, or there could be a lack of concentrated data around that number.





1,600
1,440
1,280
1,760
1,120
960
1,920
800
2,080
2,240
2,400
2,560
2,720
640
70M
60M
50M
40M
30M
20M
10M
0M
Sale Price
0
100
200
300
400
500
600
700
Full Bath