## Brief description of the data set and a summary of its attributes

Given Ames Housing data on 2925 properties sold during 2006~2010, with 82 feature measurements covering the following:

1) construction history (e.g. year built, year remodeled, lot area);
2) infrastructure (e.g. #total rooms, garage area, utility, pool);
3) overall assessment (e.g. overall quality, overall condition, neighborhood, house style);
4) sale information (e.g. sale type, sale condition, year/month sold, sale price).

## Initial plan for data exploration

1) Identify key business metrics (i.e. outcome variables): Sale Price, #Houses Sold
2) Check data quality: missing values, outliers, statistics summary
3) Explore trends in preliminary plots: scatter plots, histograms, pair plots
4) Focus on high-level attributes (i.e. feature variables): Overall Quality, Overall Condition, Lot Area, Total Rooms, Year Built, Year Remodeled, etc.
5) Conduct data cleaning & feature engineering on focused fields
6) Reevaluate correlation/causation relationship between features and outcome variables and draw preliminary hypotheses/insights
7) Run significance test to further confirm/reject hypotheses
8) Conclude actionable business insights and/or lay out additional data/experiments for follow-up

## Actions taken for data cleaning and feature engineering

1) Impute missing values:
   - Garage Cars set to 0 if missing, due to small volume
   - Garage Type set to "Unknown" if missing, due to relatively large volume and potential indicative power

2) Remove outliers:
   - Gr Liv Area > 4000 excluded, per data set author's recommendation
   - Lot Area > 20000 excluded, to be able to capture the vast majority of the sample while differentiating the impact of Lot Area on Sale Price

3) Combine least-represented values to "Other" category:
   - House Style with counts < 100 grouped into "Other"

4) Log transform skew variables for symmetrization:

– Mas Vnr Area was positively skewed and np.log1p was applied to convert it to an approximately Normally distributed shape

5) Conducted one–hot encoding on categorical variables

– Neighborhood, House Style, Garage Type, Sale Type, Sale Condition converted to binary dummies

6) Polynomial features

– Overall Qual ^2, Garage Area ^2 created due to the curved positive correlation observed in their pair plots against Sale Price

7) Feature Interactions

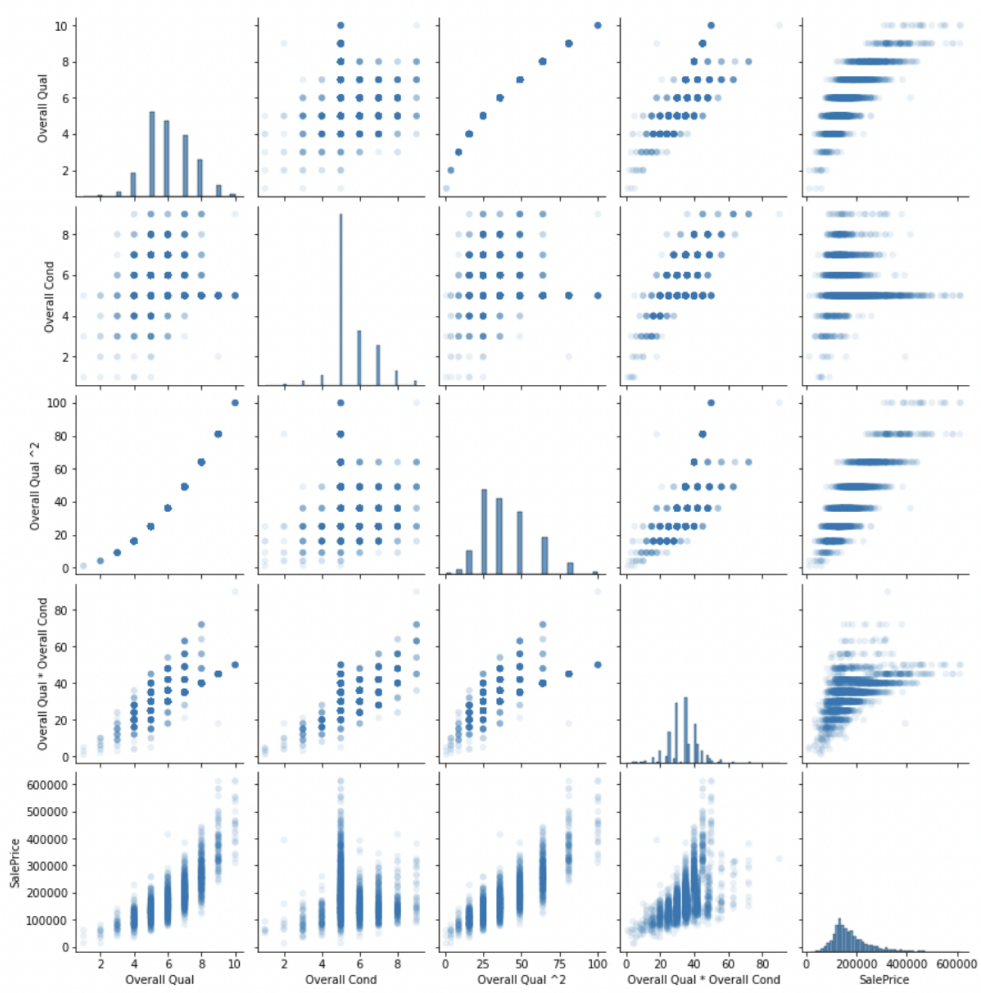
– Lot Area / TotRms AbvGrd created to capture average lot area per room, which could be a measure for house spaciousness that consumers value and could affect Sale Price

– Overall Qual * Overall Cond created to capture overall assessment of property quality & condition in one measure, where multiplication is applied due to the potentially complementary interaction between the two features

**Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner**

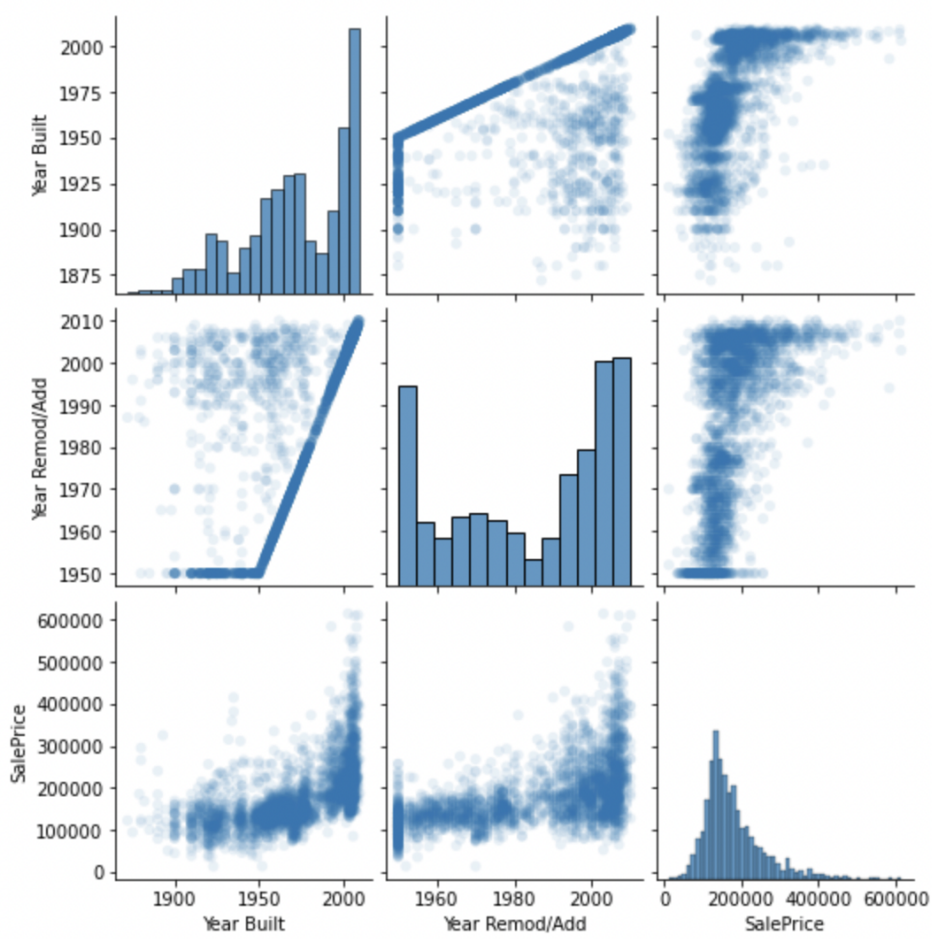
1) Overall Qual is identified to be positively correlated with Sale Price, in that better quality may have materially raised the lower bound of the sale price along with a wider wiggle room for the final price settlement. On the other hand, Overall Cond plays a limited part in raising the Sale Price, in that once the condition has passed a minimum threshold (>=5) the price can settle within a relatively wide range and beyond this point better condition does not seem to sway the price range further.

Overall Qual * Overall Cond may serve as a reasonable aggregated metric for assessing the property as a whole, and its positive correlation with Sale Price can be more consistently observed.

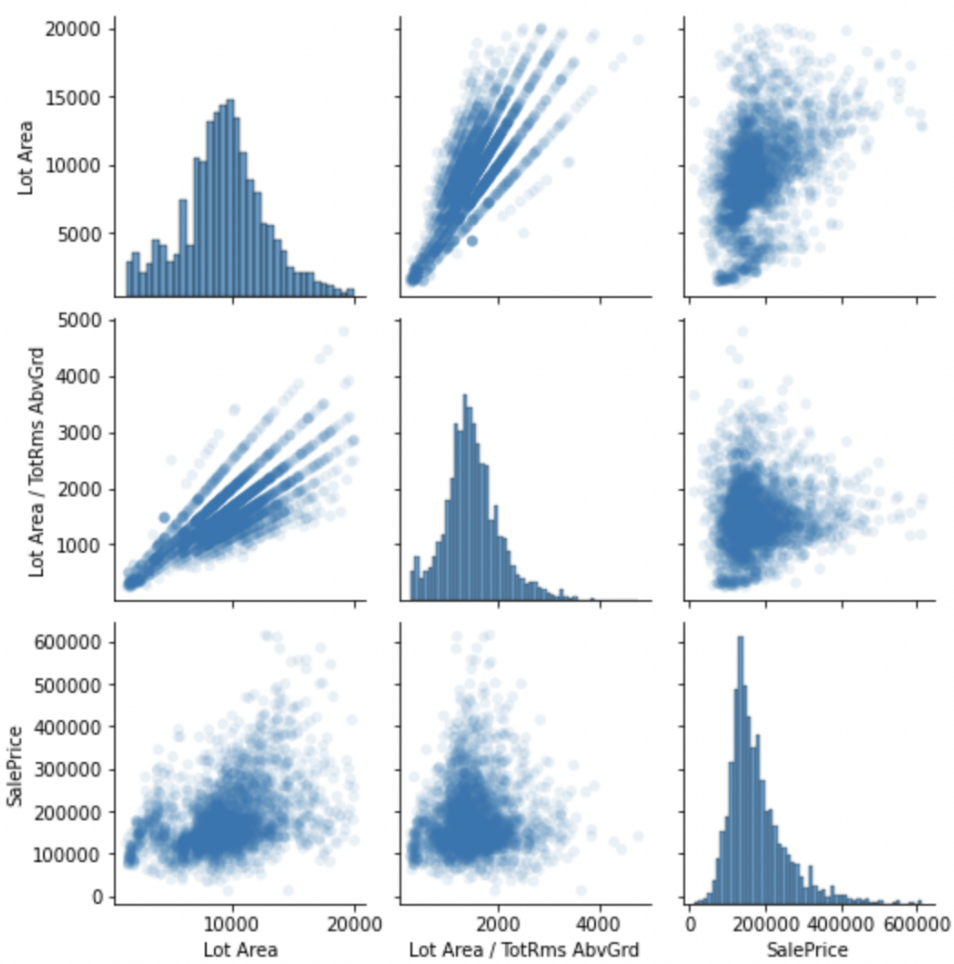



2) There's a clear separation in Sale Price between properties built before 1974 and after, where older models seems to have been settled at a similar price range over the years, while the recent models are being purchased at a higher and wider price range.

One additional note is that while a more recent Year Built may be related to a higher price bottom line, remodeling seems to provide a longer lasting power for price negotiation in favor of Ames.




3) Compared with the above features, Sale Price is not as sensitive to Lot Area or Lot Area/TotRms AbvGrd. The Sale Price seems to be bounded at the low end when the property does not come with sufficient lot area (<5000) or unit room area (<500). However, once passed that threshold, spaciousness does not translate directly to a crucial selling point, and the Sale Price seem to be driven by other features rather than the area/rooms alone.




While the above preliminary insights are still subject to further significance tests, it is recommended that Ames prioritize properties built/remodeled post–1974 with overall condition >=5 and lot area > 5000 in terms of property supplies and allocation of sales talents. Overall quality may be the crucial selling point to advertise on in order to attract consumers.

On the other hand, properties with insufficient lot area per room (<500) may not be worth spending too much resources or time on negotiation for better price. However, a finer consumer segmentation may be applied to offer promotion on the low value–generating properties and attract more traffic to Ames in general.

### Formulating at least 3 hypothesis about this data

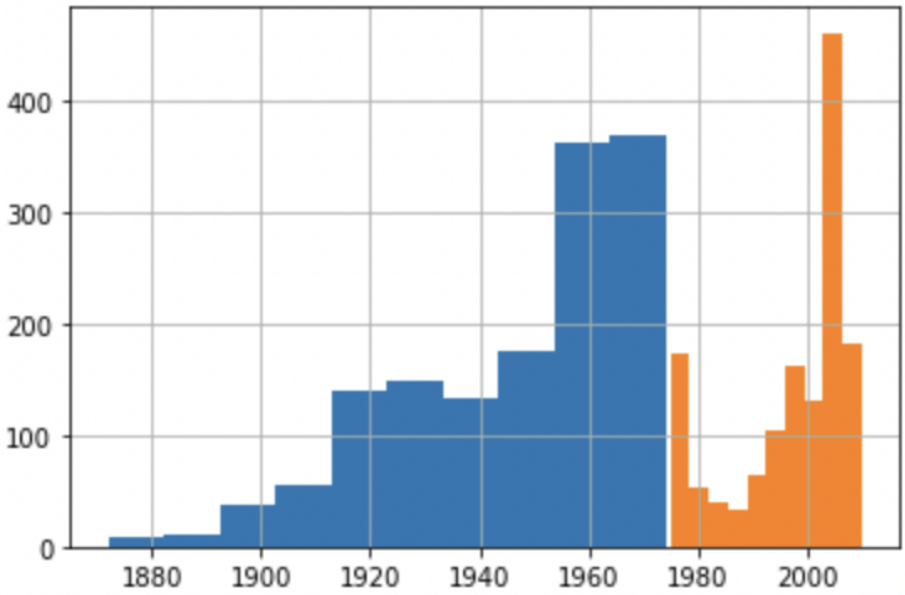
Hypothesis 1: Properties with Overall Qual * Overall Cond > 30 are sold at a better price.
Hypothesis 2: Properties with Year Built > 1974 are sold at a better price.
Hypothesis 3: Properties with Lot Area < 5000 are sold at a worse price.

**Conducting a formal significance test for one of the hypotheses and discuss the results**

Hypothesis 2: Properties with Year Built > 1974 are sold at a better price.

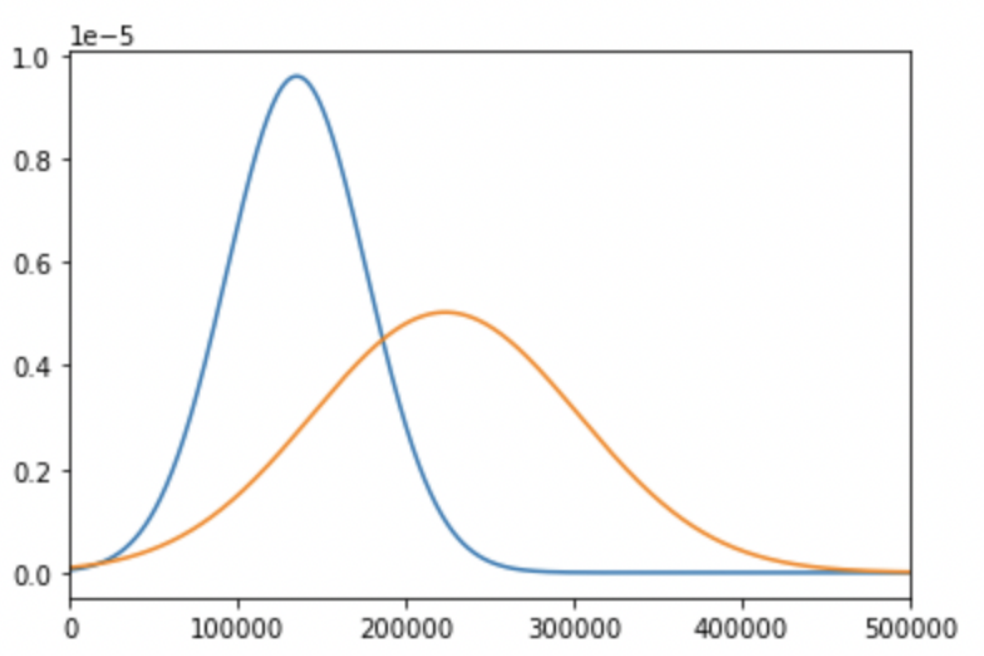
Null Hypothesis: Properties with Year Built > 1974 are not sold at a different price.
Alternative Hypothesis: Properties with Year Built > 1974 are sold at a different (better) price.

Control Group: 1436 properties built in 1974 or prior
Test Group: 1403 properties built after 1974




The Sale Price for the two groups can be fitted with the below two Normal distributions, where the blue curve represents the Control Group (with mean 135,160 and standard deviation 41,584) , and the orange curve presents the Test Group (with mean 223,597 and standard deviation 79,287).

Since the mean of the Test Group is more than 2 standard deviations away from the mean of the Control Group, the Null Hypothesis can be rejected at 95% confidence level. That is, properties built post–1974 are being sold at different (better) price that is statistically significant.




### Suggestions for next steps in analyzing this data

While the above test shows a significant relationship between Year Built and Sale Price, it is unknown whether there's indeed a causal effect (as opposed to pure correlation) or there could be a third confounding variable affecting them both. In particular, it's worth further analyzing to see if it is construction recency that consumers value, or is there a major change in property design in recent years that has caused consumers to be willing to pay at a higher price.

A retrospective analysis combined with focus group research may prove useful to gather insights on what the factors consumers regard as more important and could have resulted in the wide price range or scattered price points we see.

### A paragraph that summarizes the quality of this data set and a request for additional data if needed

This data set contains a sufficiently large sample and a wide variety of feature measurements of descent data quality for analytical purposes. One major issue with the data quality may be its lack of timeliness, as it stores data on properties sold more than 10 years ago and may not reflect a trend that's still applicable to the housing industry today.

For that matter, additional data on properties sold between 2010 and now would definitely prove useful. And any information on property purchaser should also help with consumer segmentation and the analyses / business recommendations could be more tailored and accurate.