House Price Prediction

Machine Learning for Advanced Regression

Shunling Guo, PhD
Last updated Oct 10, 2019

Overview:

- **Problem to be solved:**

  Financial media focus most of their attention on stocks and bonds, but the world's biggest asset class is actually residential property. With an estimated value of about $200trn, homes are collectively worth about three times as much as all publicly traded shares (cited from '*The Economics*'). According to Zillow and Redfin data, house prices almost doubled globally during the past five years in the USA. Houses could not only mean your dream home, but also could be valuable investments in this fast growing economic era.

  What variables/parameters are most important in determining house price is very hard to predict manually due to too many features should be taken into consideration. For example, intuitively, we would assume house size, how old is the house, neighborhood, house style, quality, etc, they are all valuable attributes, and the house price would be an overall reflection of all those values. Correctly predict/evaluate the house price would be necessary to make wise investment for investors, or to generate maximum revenue for real estate developers.

  In this project, we compared polynomial regression, random forest regression and xgboost model, and determined xgboost is best in terms of cross-validation performance and testing accuracy. Random forest get very close performance, and using principal components analysis to reduce feature dimension didn't show model improvement. We therefore used the best model for feature analysis and case study to draw business insight.

- **Business insight:**

As investigated in this project (case study), bedroom number showed association with house prices, with an inner connection of house age, living area and house quality, etc, and is not a causal attribute to drag down house prices (as shown a negative correlation in the raw data), and didn't contribute much to increase house value.

Based on the data, we would suggest for real estate developers, there is no need to design houses with more bedrooms, fewer rooms are easier to design and could cut down budget. But we need to keep another possible reason in mind, whether the bedroom number correlate with house quality and age is a selected results under market economy regulation, or is just simply because there is not enough choice in the market that people have no selection opportunity to choose newer and better houses with more bedrooms (for example new houses), and to test this hypothesis, we could design an A/B test or at least do some marketing investigation to see whether increasing bedroom number in new houses would generate more revenue or not, and to see whether increasing newer and better houses in the market with more bedrooms would affect the data and results in a different model. The model would be adjusted accordingly to market changes. The model could also predict marketing value of the new houses, to calculate revenue generation in advance.

For marketing purposes, we surprisingly see 'wood deck square feet' and 'open porch square feet' are of high importance among the features, above 'overall quality' feature, which we already showed is one of the driving forces to affect house prices. Therefore, we would suggest to highlight these features in advertisements to draw the attention of buyers.

For investors, we could use the model to predict the house price to see whether the house price is over or under market value. Also since in this particular data set, the house prices didn't change much

along time, we would not suggest this is a good region for investing in real estate due to lack of appreciation space. But it also indicate this region is quite safe to invest if you just want to avoid investment risk because the house price is quite stable and unlikely to drop much, even after the financial crisis in 2008. It depends on your purpose for investment whether is to pursue profit or to avoid risk.

Statistical Inference:

- **Hypothesis Testing:**
  In this section, I first observed whether there are some differences between different groups, then form hypotheses to see whether the difference is high enough to give any statistical significance.

  ❖ Null hypothesis 1: one story or multi-story houses have the same price per square feet of living area.

  ➢ Data:
  story1_mean,story1_sd,story2_mean,story2_sd
  (132.7, 31.3, 108.6, 26.5)

  ➢ t-test:

```
ttest_ind(story1.PPLSF,story2.PPLSF,equal_var = True)
```

```
Ttest_indResult(statistic=15.9259878990733, pvalue=9.015848818873706e-53)
```

  ➢ Bootstrapping testing:

  - Bootstrapping the mean of the two groups and center it at 0 to leave only raviance distribution.
  - Get distribution sample of the difference between the two groups.
  - Calculate observed difference.
  - Calculate the probability (p-value) for seeing the observed difference or bigger under the null hypothesis (the hypothesized difference is 0)

```
# the probability to see this observation
p = np.sum(np.abs(dif) > obs_dif) / len(dif)
p
```

```
0.0
```

➢ Conclusion:
Either from t-testing or from bootstrapping testing, the p value under the null hypothesis is two small (type-1 error level alpha < 0.001), therefore, we should reject the null hypothesis, and the alternative hypothesis is True: 1 story house and multi-story houses have different price per square feet.

❖ Null hypothesis 2: house price (in terms of price per square feet of living area) from different years is the same.

➢ Data:

	year	count	mean_price	sd
0	2006	314	119.85	29.27
1	2007	329	121.61	30.64
2	2008	304	121.35	30.91
3	2009	338	119.47	33.77
4	2010	175	120.69	32.76

➢ It looks like the difference is very small, we only choose one pair to test for example: Year 2007 and Year 2009

```
ttest_ind(Y2007,Y2009,equal_var = True)
```

```
Ttest_indResult(statistic=0.8548212202635521, pvalue=0.39295798040706165)
```

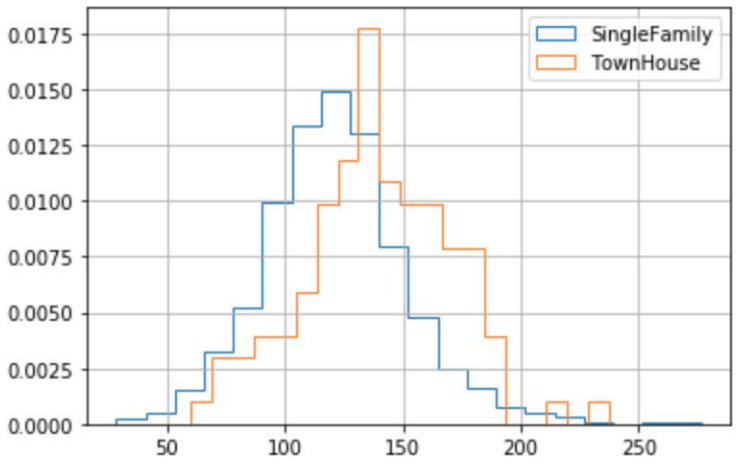
```
ttest_ind(Y2007,Y2009,equal_var = False)
```

```
Ttest_indResult(statistic=0.8559427222757352, pvalue=0.3923392786528672)
```

➢ Conclusion: the p value is big enough (>0.05), and we could not reject the null hypothesis.

❖ Null hypothesis 3: Townhouses have the same price per square feet of living area with single family houses.

➢ Data:histogram distribution of house price per square feet in living area (x-axis) and frequency (y-axis)



➢ T-test:

```
ttest_ind(Single,TownH,equal_var = True)
```

```
Ttest_indResult(statistic=-5.95334927414826, pvalue=3.355238005363966e-09)
```

```
ttest_ind(Single,TownH,equal_var = False)
```

```
Ttest_indResult(statistic=-5.801430756457911, pvalue=4.531556117720209e-08)
```

➢ Conclusion: p value is small enough (<0.001) to reject the null hypothesis. Difference is significant for the price per square feet in living area between single family houses and townhouses.

Machine Learning:

In this section, we compared three models (polynomial, random forest, xgboost) and choose Xgboost as the best model to train data, and fine tuned parameters to reach maximum testing accuracy power.

- **Comparison of different regression models:**

regression	Best Cross validation r2 score	Best Testing score
Polynomial	0.84	0.76
Random forest	0.87	0.83
Xgboost	>0.99	0.87

- **Random grid search to get a reasonable starting point:**

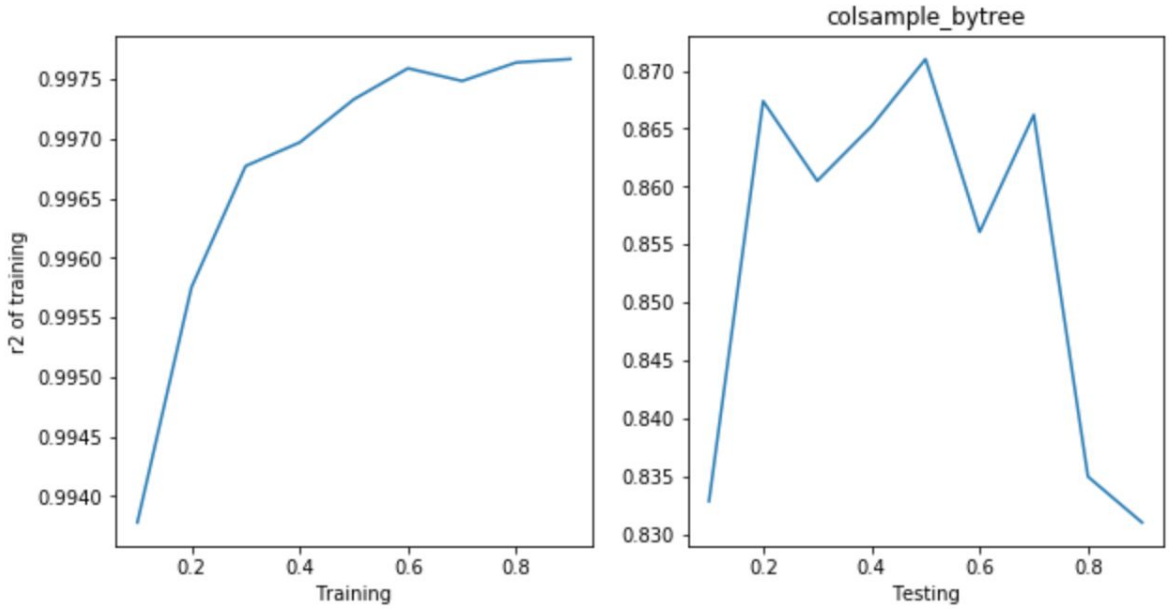
```
model = xgb.XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
          colsample_bynode=1, colsample_bytree=1, gamma=0,
          importance_type='gain', learning_rate=0.15000000000000002,
          max_delta_step=0, max_depth=4, min_child_weight=1, missing=None,
          n_estimators=50*n, n_jobs=1, nthread=None,
          objective='reg:squarederror', random_state=0, reg_alpha=0,
          reg_lambda=1, scale_pos_weight=1, seed=None, silent=None,
          subsample=0.8500000000000001, verbosity=1)
```

- **Fine tune parameters:**
  - ❖ I plotted accuracy curve to find the best value for the following parameters:
    - ➢ N_estimator
    - ➢ Learning_rate
    - ➢ Maximum tree depth
    - ➢ Min tree nodes
    - ➢ subsample/colsample
    - ➢ Base score
  - ❖ Experiment outcome was documented into a dataframe, and saved as '.csv' file.

	parameter_name	best_parameter(train,test)	best_score(train,test)
0	n_estimators	(400, 350)	(0.999807244639819, 0.8621939187139758)
1	Learning Rate	(0.16, 0.06)	(0.9997778081287881, 0.8630069628715896)
2	max_depth	(7, 2)	(0.999982526524428, 0.8579016329547222)
3	subsample	(0.64, 0.56)	(0.9979672879949777, 0.8571673758577336)
4	colsample_bytree	(0.8, 0.4)	(0.997669999620488, 0.8710150491582573)
5	min_child_weight	(0.0, 0.2)	(0.9970036828809768, 0.8691081680526537)

❖ Plotting accuracy performance example:

Fig3. y axis is accuracy score r-square, x axis is colsample_bytree value. Left panel is training process, right panel is testing process.



- **Conclusion: The best model found is:**

```
# Best model for testing accuracy:
model = xgb.XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=0.4, gamma=0,
             importance_type='gain', learning_rate=0.2,
             max_delta_step=0, max_depth=3, min_child_weight=1, missing=None,
             n_estimators= 350, n_jobs=1, nthread=None,
             objective='reg:squarederror', random_state=0, reg_alpha=0,
             reg_lambda=1, scale_pos_weight=1, seed=None, silent=None,
             subsample=0.65, verbosity=1)
```

Saved model to ‘best_xgb_model.dat’ for future prediction.

Case Study Story:

In this section, I chose one feature from each interesting feature keywords (Quality, age, size) to investigate causal relationships: House overall quality, age, living area. I also investigated a more ambiguous feature, the bedroom number, since intuitively we would think it is important.

- ### Time Series View:

  Fig4. Age of house and Sale Price correlation trends over time.
  Y axis is average house sale price (red, SalePrice) or average home age (black, YearBuilt) of sold houses within a month. X axis is Sold Time with precision of a month.
  We see a negative correlation along time between the two.

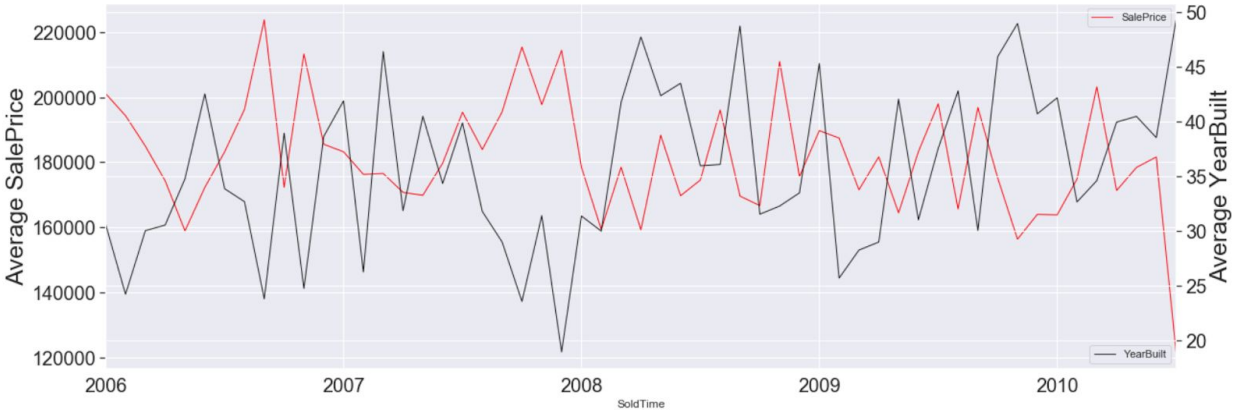


Fig5. House quality and Sale Price correlation trends over time.
Y axis is average house sale price (red, SalePrice) or average home quality (black, OverallQual) of sold houses within a month. X axis is Sold Time with precision of a month.
We see a positive correlation along time between the two.

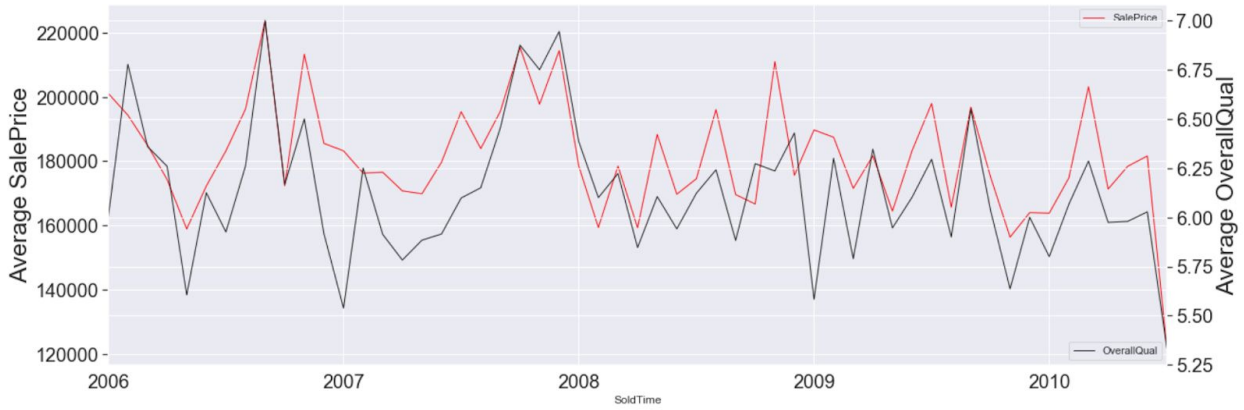
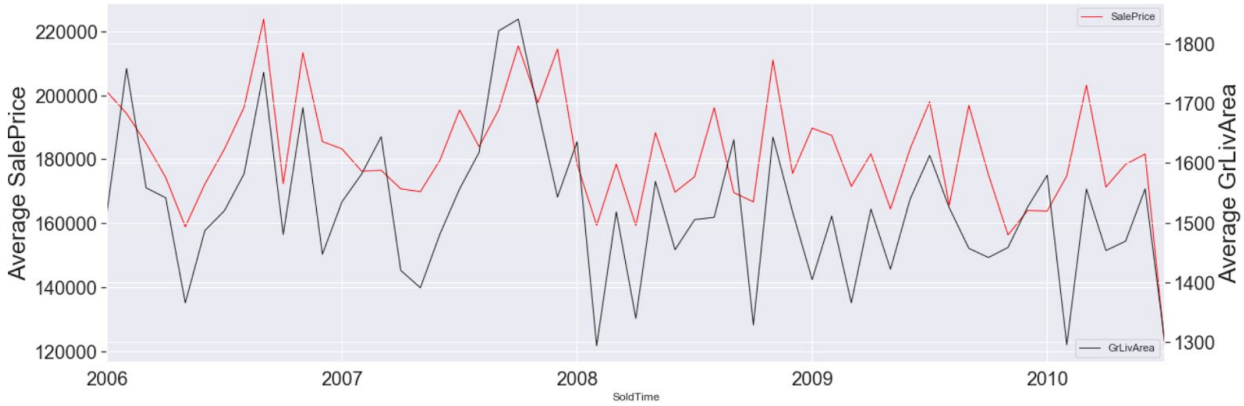
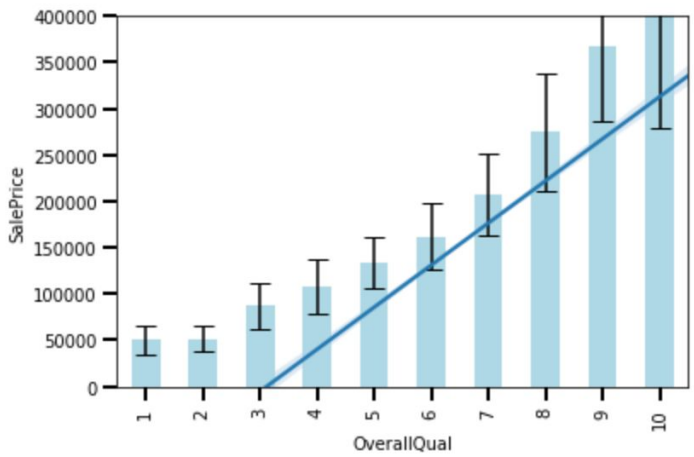


Fig5. Living area and Sale Price correlation trends over time.
Y axis is average house sale price (red, SalePrice) or average living area (black, GrLivArea) of sold houses within a month. X axis is Sold Time with precision of a month.
The correlation is harder to predict, most times are positively correlated, but sometimes see negative effect.



- **Statistical inference:**
  Time series are most suitable for us to draw a trend insight, which unfortunately, from this dataset, we didn't see a trend pattern, therefore, more suitable way to visualize correlation is to use mean and standard deviation plot, and fit data using regression. I also used ANOVA test difference and confidence interval between groups.

Fig6. plot the mean and standard deviation of house prices grouped by overall quality of house. Fit with linear regression.



OLS Regression Results

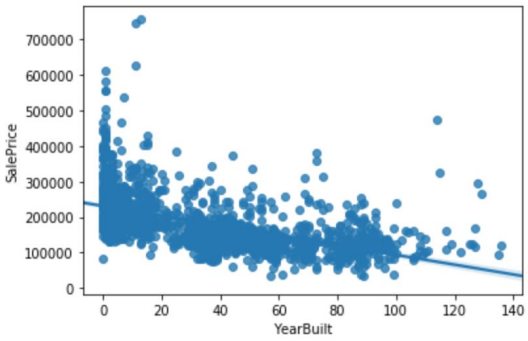
Dep. Variable:	OverallQual	R-squared (uncentered):	0.928
Model:	OLS	Adj. R-squared (uncentered):	0.928
Method:	Least Squares	F-statistic:	1.879e+04
Date:	Wed, 04 Sep 2019	Prob (F-statistic):	0.00
Time:	10:30:10	Log-Likelihood:	-2828.1
No. Observations:	1460	AIC:	5658.
Df Residuals:	1459	BIC:	5663.
Df Model:	1		
Covariance Type:	nonrobust		

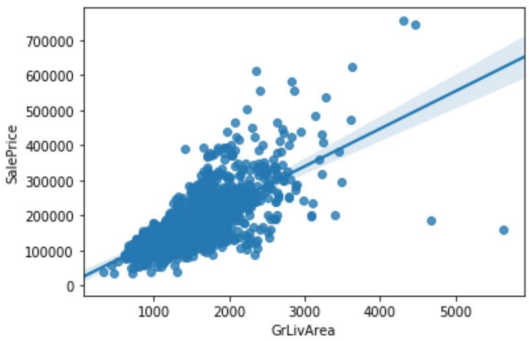
	coef	std err	t	P>|t|	[0.025	0.975]
SalePrice	3.049e-05	2.22e-07	137.068	0.000	3.01e-05	3.09e-05

Omnibus:	812.455	Durbin-Watson:	1.794
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10480.098
Skew:	-2.319	Prob(JB):	0.00
Kurtosis:	15.278	Cond. No.	1.00

The above statistics showed strong correlation of x, y in data, r-squared = 0.928, and the probability to get this statistics given the null hypothesis that x,y is not related is 0.00 (Prob of F-statistic). The coefficient (line slope) within 95% confidence interval is [3.01e-5, 3.09e-5].

Fig7. Correlation for house age and house size with sale price:





We now can have a good sense of the positive and negative correlation of features and sale prices.

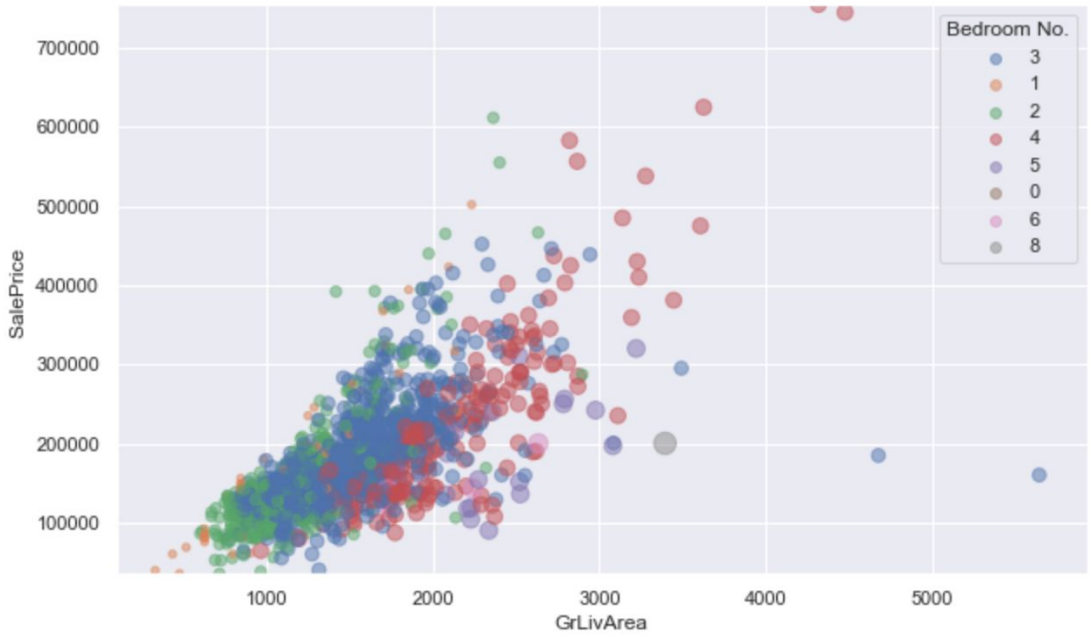
- **Investigate the relationship between bedroom number and sale price.**

  Intuitively, we would think bedroom numbers could affect house prices a lot, due to the market impression that renting a 3 bedroom house is much more expensive than renting a 2 bedroom or 1 bedroom house. Is that true in metadata?

  EDA:

  We used scatter plot to show this information with bedroom numbers in different colors:

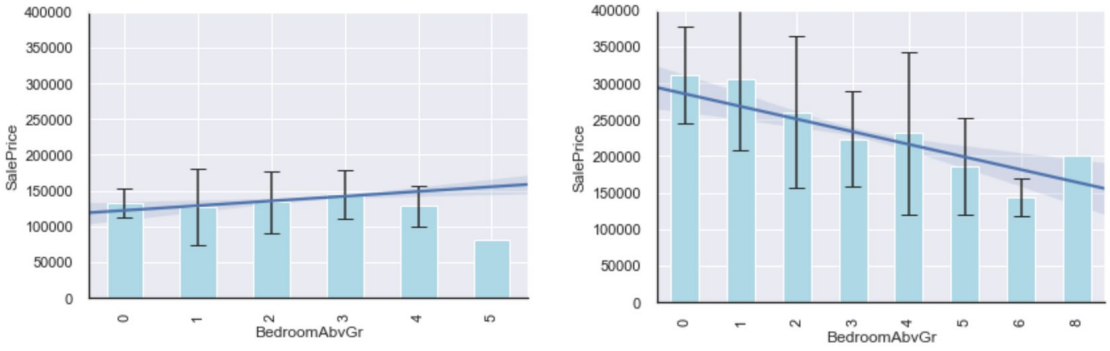
  Fig8.



Surprisingly, we would see that generally speaking, bedroom numbers negatively correlated with SalePrice, especially after living area bigger than 1500 square feet.

We separate the data for two parts, df1500U contain house data with living area more than 1500 square feet, and df1500L contain house data with living area less than 1500 square feet.

Fig9. left: df1500L, right: df1500U.



We can see, for smaller houses, bedroom number slightly positively correlated with house price, but the effect is very subtle, but for bigger rooms, bedroom number strongly negatively correlated with house

price. ANONA analysis with post Tukey Hsd analysis showed significant group difference, with p value smaller than 0.002.

Multiple Comparison of Means - Tukey HSD, FWER=0.03

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-5463.3333	0.9	-183353.6785	172427.0118	False
0	2	-50298.0145	0.9	-214094.9474	113498.9185	False
0	3	-86825.1878	0.6258	-247820.3807	74170.0052	False
0	4	-79340.5218	0.7199	-240943.0847	82262.041	False
0	5	-124523.3333	0.2594	-296477.6105	47430.9438	False
0	6	-166554.3333	0.0873	-358207.0777	25098.4111	False
0	8	-110333.3333	0.9	-431029.7138	210363.0471	False
1	2	-44834.6812	0.6362	-128806.8547	39137.4924	False
1	3	-81361.8544	0.0167	-159728.5747	-2995.1342	True
1	4	-73877.1885	0.0511	-153484.2171	5729.8401	False
1	5	-119060.0	0.0023	-218005.2582	-20114.7418	True
1	6	-161091.0	0.0015	-291293.4006	-30888.5994	True
1	8	-104870.0	0.9	-393085.2999	183345.2999	False
2	3	-36527.1733	0.0242	-72938.7536	-115.593	True
2	4	-29042.5073	0.227	-68051.985	9966.9703	False
2	5	-74225.3188	0.0143	-144756.3365	-3694.3012	True
2	6	-116256.3188	0.0138	-226424.9417	-6087.696	True

Question: Correlation or Causation?

Is this a random association or actually driven by other factors or the bedroom number is a driven force to drag down house price for bigger rooms?

To answer this question, we would take advantage of machine learning models, to do a Monte Carlo simulation, which is to use random simulation to randomly assign bedroom numbers to houses, and then predict the house price to see whether altering bedroom number alone would drive the house price to change, or say whether more rooms really can drag the house price down.

We need a clean background to show the difference, otherwise since house price is a multi-factor driven outcome, we don't want other features to bother the outcome and bury the signal in background.

Keep in mind, the negative correlation only happens in bigger houses

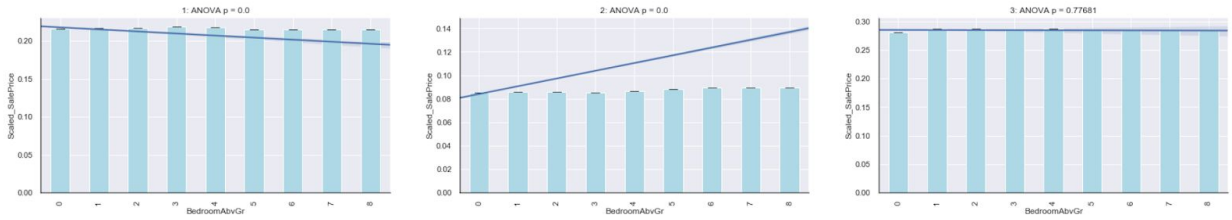
Simulation Algorithm:

- ❖ Choose a random house with living area bigger than 1500 as a simulation base input
- ❖ Keep all other features the same with the chosen house, assign bedroom numbers randomly to house data
- ❖ Predict house price after processing data accordlingly to protocol
- ❖ Draw mean and standard deviation and plot linear regression to visualize the data

Results:

- ❖ Results showed random trend. We don't see a strong correlation between bedroom number with house price.

Fig10. Simulation results of 3 independent simulation experiments.



Question:

Then what caused this observation that bedroom number is negatively correlated with house price, which is very counter-intuitive.

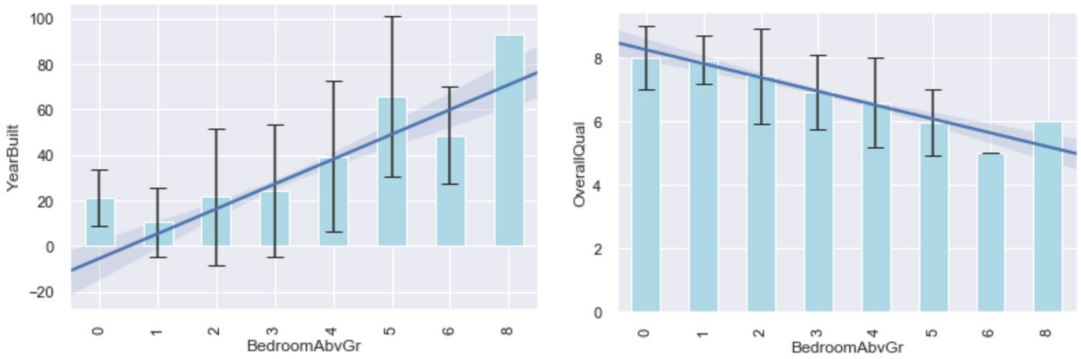
Hypothesis:

Bedroom number is actually correlated with other more important features, for example: house age. It could be possible that older houses tend to have more bedrooms, because family size is generally bigger in older times, and nowadays, family sizes get

smaller because people tend to have fewer kids in the more modern era (this should be supported by other data), therefore, newer houses tend to have fewer bedrooms.
Another more important feature is: house quality. Bedroom number might be negatively correlated with house quality, because it's easier to maintain or fix for fewer bedroom houses, house quality could also correlate with house age because older rooms tend to have worse condition than newer rooms.

Testing:
First of all, we need to see whether bedroom numbers are correlated with house age or overall quality.

Fig11. Plot of mean and standard deviation with linear regression of bedroom number and Yearbuilt or OverallQual.



	coef	std err	t	P>|t|	[0.025	0.975]
YearBuilt	0.0545	0.002	26.296	0.000	0.050	0.059

Omnibus:	89.762	Durbin-Watson:	1.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	123.564
Skew:	-1.023	Prob(JB):	1.47e-27
Kurtosis:	3.458	Cond. No.	1.00

	coef	std err	t	P>|t|	[0.025	0.975]
OverallQual	0.4488	0.006	69.923	0.000	0.436	0.461

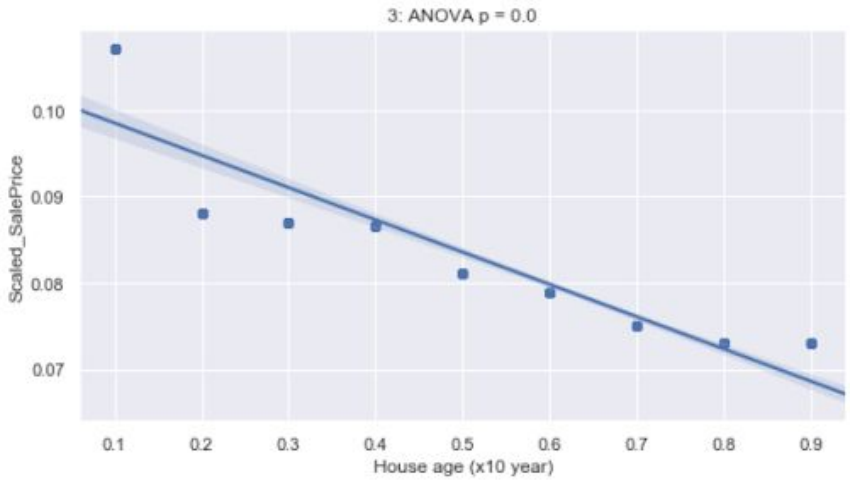
Omnibus:	26.071	Durbin-Watson:	1.921
Prob(Omnibus):	0.000	Jarque-Bera (JB):	61.411
Skew:	0.139	Prob(JB):	4.62e-14
Kurtosis:	4.451	Cond. No.	1.00

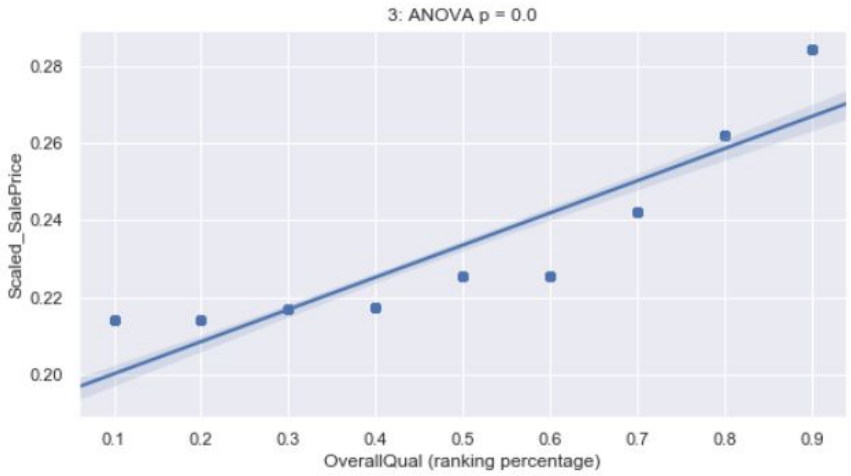
The answer is yes! Bedroom number is positively correlated with house age, which is negatively correlated with house price. Bedroom number is also negatively correlated with house quality, which is positively correlated with house price.

Question:
Are house age and quality driven forces to affect house age?

Testing:
We used the same simulation algorithm and change bedroom number to ‘Yearbuilt’ or ‘OverallQual’.
The Answer is Yes!

Fig12. Simulation results examples.

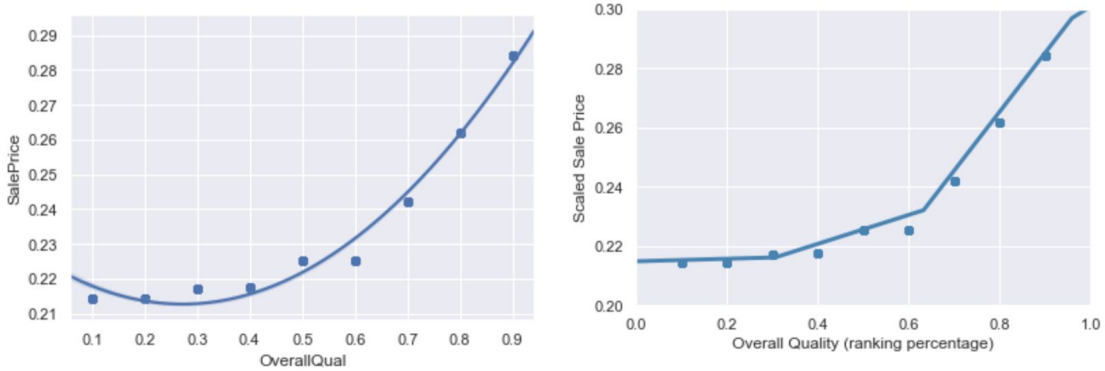




House price is strongly correlated with house age and Overall quality, although the correlation might not be linear, it showed the house prices would be capped to some point. It also gave us insight that house age would only affect house prices within a certain range, and overall quality below a certain point would also not contribute too much to house price difference.

We could use other functions to better fit with the simulation data. For example, higher order of polynomial regression (Fig13 left) or sigmoid curve (Fig13 right).

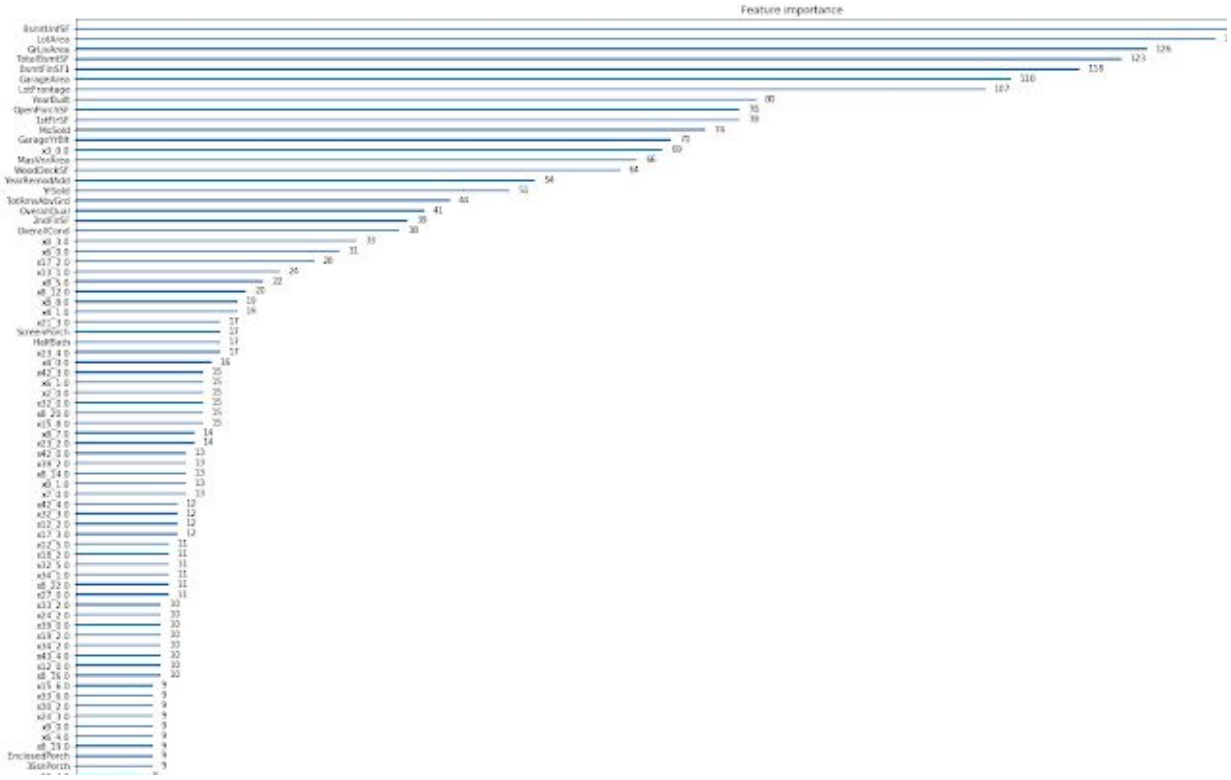
Fig13. Other regression to fit simulation results for better fitting. Left: order of 2 polynomial regression. Right: sigmoid curve regression.



Conclusion:

House ages and House quality are two driving forces to determine house prices, and bedroom number is not, it is just correlated with the important features therefore correlated with the house price, but is not a driving force or causal feature. Actually, with machine learning, we could extract the importance hierarchy precisely.

Fig14. Importance score of each feature: (partial)

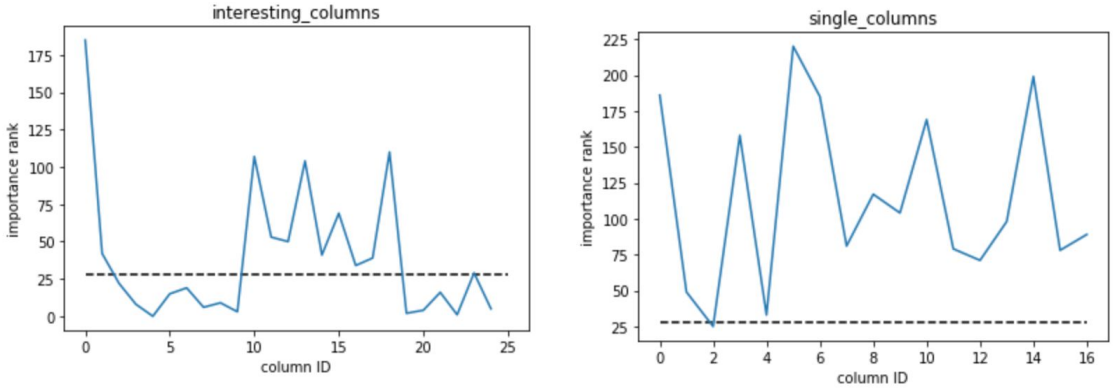


According to the importance score, bedroom number is one of the least considered features during the modeling (score =13, ranked 28), while living area, house quality and age are among the top 20 important features (score > 30). This is a very good explanation of what is causally correlated with outcome, and what correlation is not a cause at all due to much less important to affect the outcome.

Let’s also verify whether our initial hypothesis is right:

Initially we thought the features in ‘single_column’ tend to be not important because not enough variance could be introduced, and in ‘interesting_columns’ tend to be important, because they showed correlation with sale price. We plotted the importance score (the lower the more important), and set 28 as a threshold, which is equal to the importance of bedroom number.
The results supported our hypothesis. Although in ‘interesting columns’, some features pop out to be not important(features with rank above threshold), but indeed many of them are important, while in ‘single columns’, almost all of them are not important. Machine learning helped us to make unbiased conclusions.

Fig15. Plot importance of each column in single_columns(right) or interesting_columns(left).



Discussion:

Classification of problem:

House price regression represents one type of problem that share the following essences:

- ❖ Only one outcome feature, could be either categorical (discrete) or numeric (continuous).
- ❖ The parameters/features/attributes/variables which could affect the outcome are too many to fit manually predictable model or function with manageable estimators (10 is already a lot if manually).

- ❖ The feature number to affect the outcome should be less than the training sample number.
- ❖ There are a mix of categorical and numeric features.

This project is an example of attribution analysis to extract importance of attributes, and understand the causal or correlation relationship between attribute and outcome.

This project also showed the power of machine learning in extracting important attributes among hundreds of attributes. Without machine learning algorithm, we could easily mistakenly thought some features are important due to seeing a strong correlation of the feature and the outcome.

Similar problems could be applied to (not limited to):

- ❖ Clinical data with patient background (genetic and environmental) as attributes/features, and patient's disease or drug response as outcome. Note: If features numbers is bigger than sample size, then it's another kind of problem.
- ❖ Multi-channel marketing with multiple marketing/advertising strategies as attributes and CTR (Click-through rate) or sales as outcome.
- ❖ Fraud detection.

Successfully solving a problem and provide the correct data-driven solution includes at least three parts of work:

- ❖ Deep understanding of problems, perform careful investigation to gather enough information to make sure the driving forces/features are all included in the dataset for machine learning to mine them out. These require people to form all possible hypotheses for data curation.
- ❖ Deep understanding of machine learning algorithms and choose the best model.

- ❖ Careful validation and interpretation of data (data bias, data balance, etc), distinguish the difference of causal reason and correlation.

Further resources to strengthen model:

This data set contains 79 attributes that potentially could affect house prices, but are there any other attributes that could contribute too? We would simply thought the crime information and school district information would also be valuable attributes, although this information might be reflected in neighborhood feature, but how closely they are correlated would not be clear until we see the data. Walking score would also worthy considerate, and even flood probability. Other data curation work needs to be done to further enrich our understanding of this problem and improve model accuracy.

Appendix:

- **Data:**
  This project would use 'Ames Housing Dataset' from kaggle competition as training and testing data. Data from here.
  Explanation:
  The dataset contains 79 explanatory variables, with 44 categorical variables and 35 numerical variables.
  Detailed explanation of each variable and its categorical groups is uploaded into github (info from here).

- **Data Cleaning:**
  First, load data to pandas dataframe. Fill missing values (NA) with appropriate value according to explanatory of variables, for example, NA could be filled either with a 0 or a string indicate for a category.

Deal with abnormal data if any.

- **Data Wrangling:**
  - ❖ Change year to age
  - ❖ Convert selling year and month data to datetime object, and set ‘datetime’ to index for ‘timeseries’ analysis
  - ❖ Re-organize data with categorical data grouped together and numeric data grouped together
  - ❖ Preprocess categorical data using LabelEncoder() and OneHotEncoder(), save mapping dictionaries for standardized data processing protocol
  - ❖ Use MinMaxScaler to normalize data
  - ❖ Split data to training and testing set for standardized data input for different machine learning algorithms
  - ❖ Save training and testing input into csv files for future load
  - ❖ Standardize data processing of new data for future prediction

- **Exploratory Data Analysis - Distribution plot:**

  In this section, I first explored the data distribution to get a general idea of data complexity. Track features which doesn’t have enough complexity, they could be potentially not important due to not enough diversity or variance generated to contribute to difference. Store these columns in ‘single_columns’ list.

  Generally speaking, the features that are in ‘single_columns’ are prone to be miscellaneous, key words can be summarized as:
  - ❖ Porch
  - ❖ Fence
  - ❖ Pool
  - ❖ Utility

- ❖ Outside condition

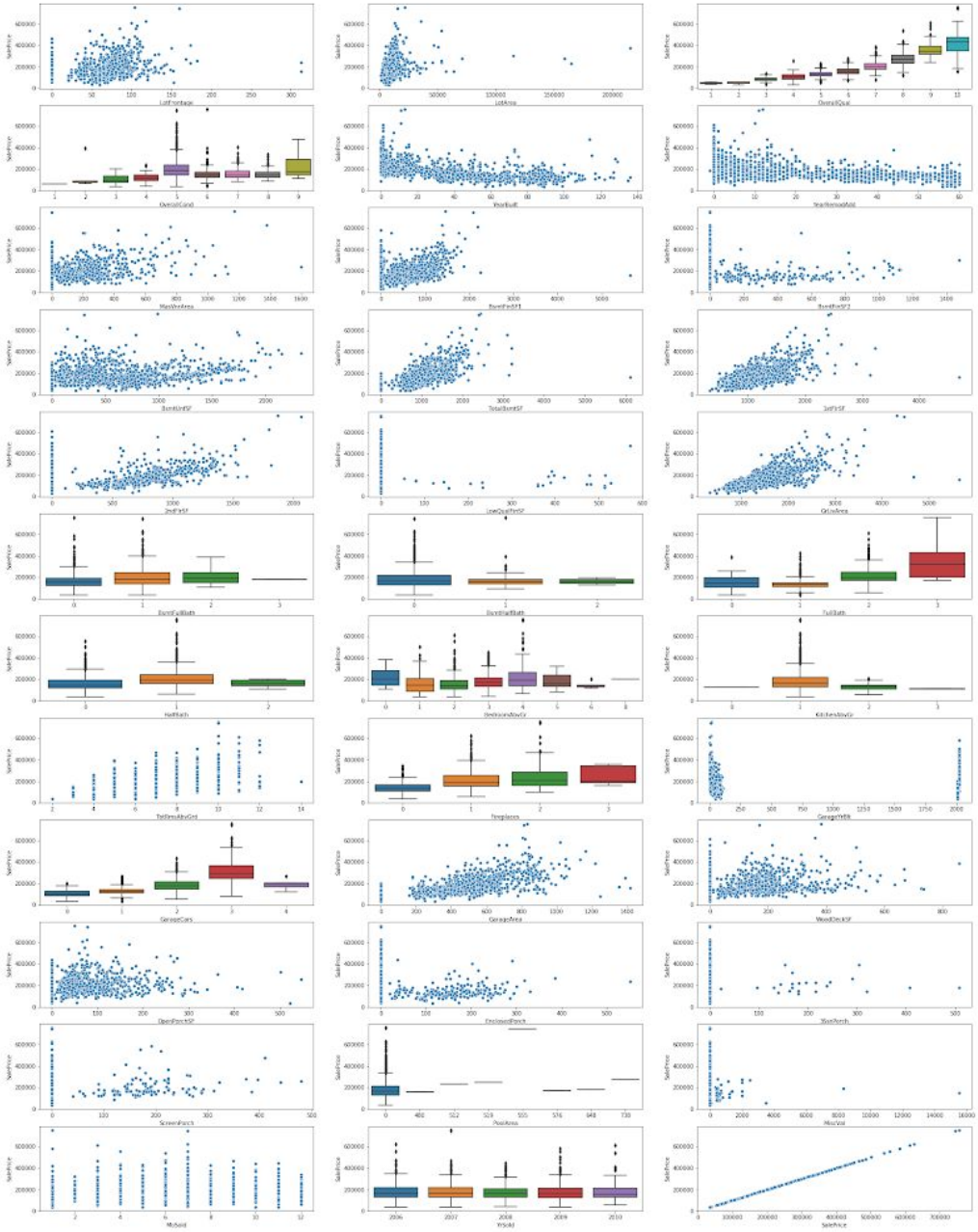
- **EDA - Correlation plot:**
  In this section, I plotted the relationship between ‘SalePrice’ and each feature, track the interesting columns that have noticeable correlation with ‘SalePrice’, those are potential important features to be analyzed. Store the information in ‘interesting_columns’ list.

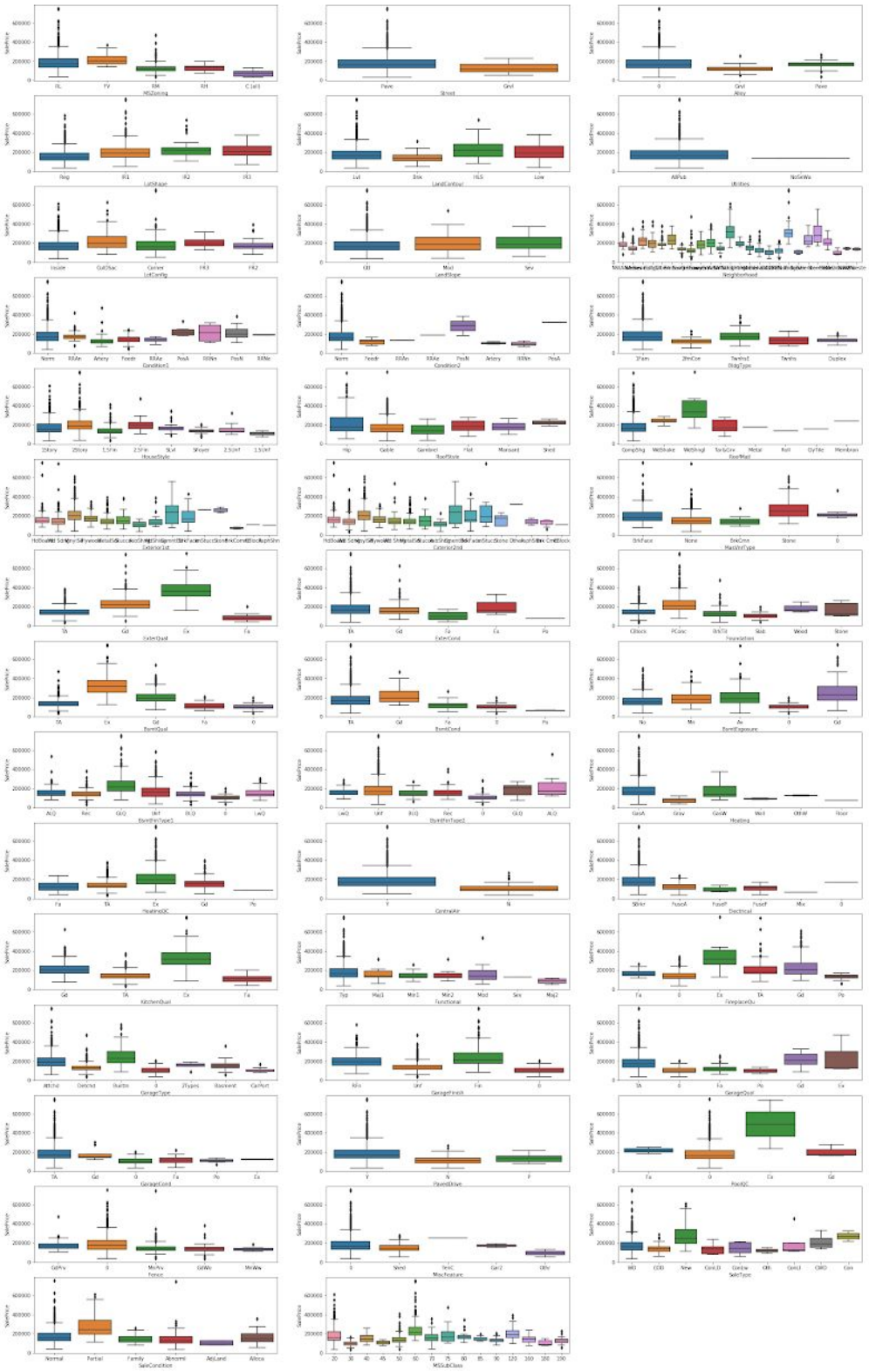
  There are four key words can be summarized to describe the common attributes of features in ‘interesting_columns’:
  - ❖ Quality
  - ❖ Age
  - ❖ Size
  - ❖ Neighborhood

Fig s1: Distribution of variables. For categorical data, mean and standard deviation were drawn for each category, and for continuous data, histogram was used.

Fig s1

Fig s2: Correlation plot of variables. For more discrete data (unique values <= 10), box plots of median and percentiles were used, and for more continuous data, scatter plots were used.





- **Prediction of new data:**
  - ❖ Preprocess new data accordingly as training data
  - ❖ Load best model and predict
  - ❖ Scale final sale price back to real number as original input
  - ❖ Save results in dataframe and '.csv' file

- **Raw codes in Github repository:**

  Jupyter notebooks: Shirley_HousePriceRegression_{*}.ipynb
  - ❖ Data cleaning, wrangling, processing protocol
    - ➢ DataCleaning_Wrangling
    - ➢ DataPreprocessing
  - ❖ Machine learning with parameter fine tuning
    - ➢ ML_Polynomial
    - ➢ ML_RandomForest
    - ➢ ML_Xgboost
    - ➢ ML_Xgboost_2
    - ➢ ML_Xgboost_Fine_tuning
  - ❖ Statistical analysis and storytelling
    - ➢ Statistical_Analysis
    - ➢ EDA_StoryTelling
    - ➢ more_EDA
    - ➢ Data_insight_price_dif
  - ❖ Prediction of new data
    - ➢ Xgboost_Predict_Preprocess_Test_Data
    - ➢ Xgboost_Prediction