BUSINESS ANALYTICS --PROJECT I HOUSING DEMAND ANALYSIS--SAMPLE ANSWER

Instructor：Zoe

TA：Merlin & 包子

REQUIREMENTS 要求

- Demonstrate your thought process of translating a business problem to a statistical problem
  展示将商业问题转化为统计问题的思维过程

- Create data visualization
  制作数据可视化

- Apply technical methods (e.g., Regression) 应用技术方法（例：回归）
  - No limitation on software used 不限定软件使用
  - No limitation on methodologies used 不限定方法

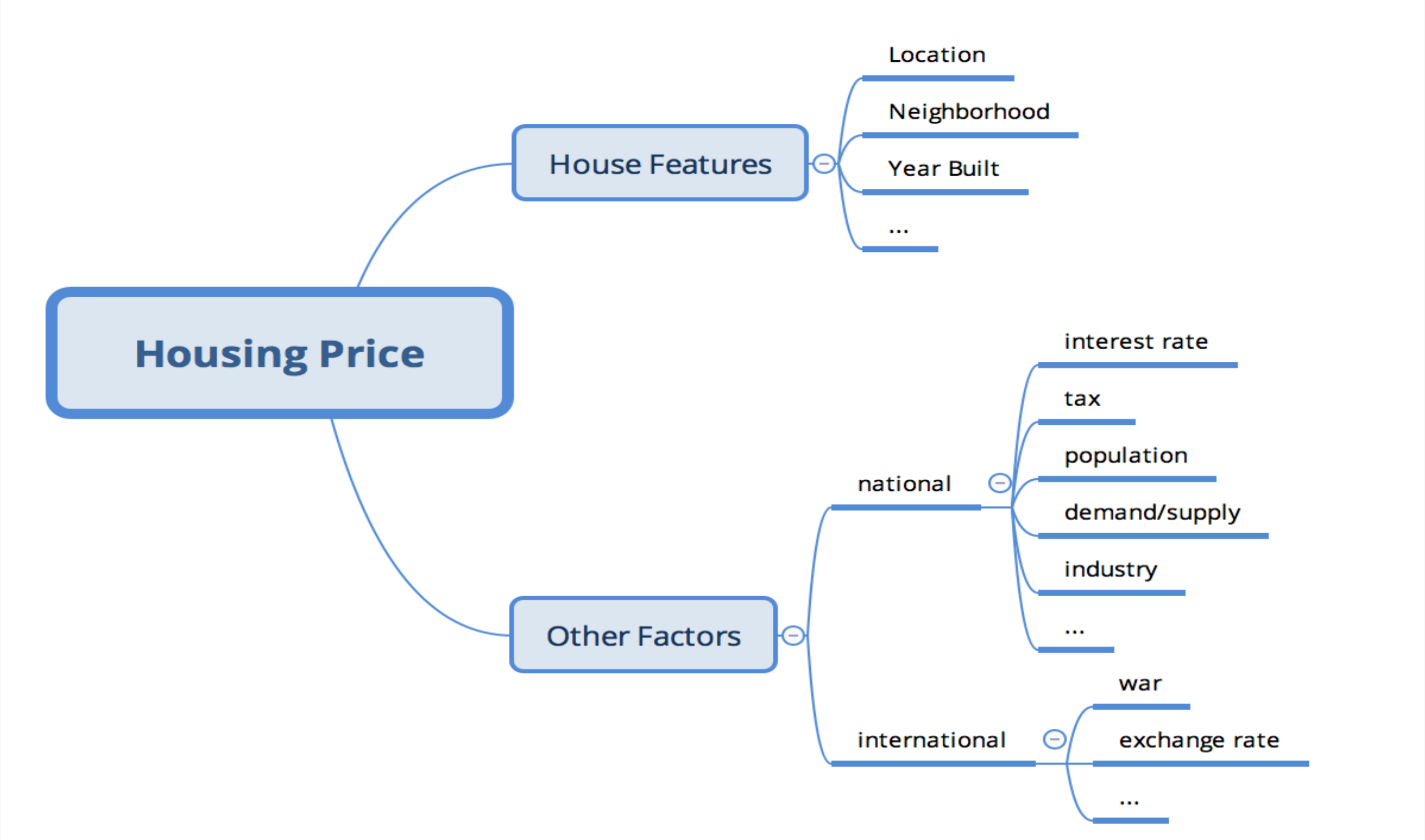
- Create a presentation illustrating a complete business story 课堂展示并讲述商业故事

PROJECT1 课题1：

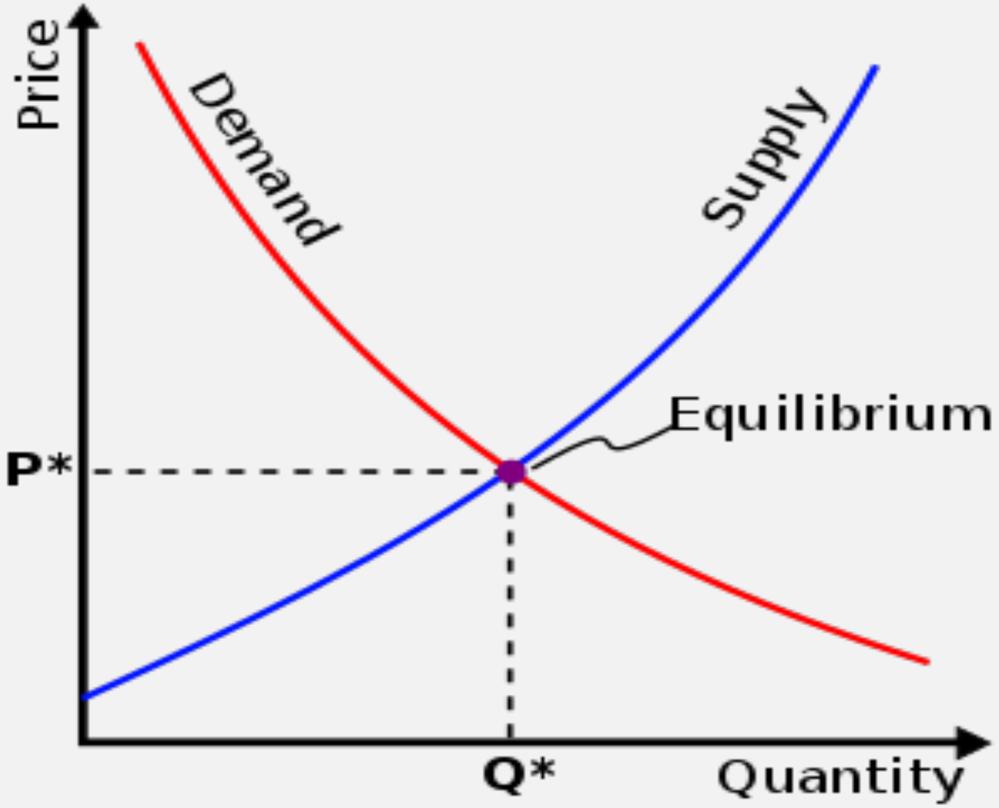
- Suppose you are a business analyst in a real estate agency. You are assigned to predict the next season housing price with your 4 colleges using statistical methods. You need to submit your results to the stakeholders within one week.
- Data: Historical housing features information is given. You need to use but not limited the given "housing.csv" data to to predict the future housing prices in the "predict.csv" package.
- 假设你是某房地产公司的商业分析师，现在公司任命你为房价预测组的负责人之一，你将会和你的四个组员一起预测将来的房价走势。你有一个星期来完成这项任务，提交研究报告，3个优秀组将获得推荐并进行成果展示。
- 数据：你的数据来源要包括但不限于给定的公司内部房屋特征和测试信息。

THINKING PROCESS 思维过程

- First, we ask ourselves: What determine the house price ?



HOUSING PRICE IS DETERMINED BY DEMAND AND SUPPLY
经济学：房价由市场供需关系决定



DATA COLLECTION
数据 I

- Previous Customer／Client information
- 考虑到宏观数据的收集比较困难，并且房地产公司会有大量的历史客户信息，我们将数据锁定在微观房产特征上，共有79个类别

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN

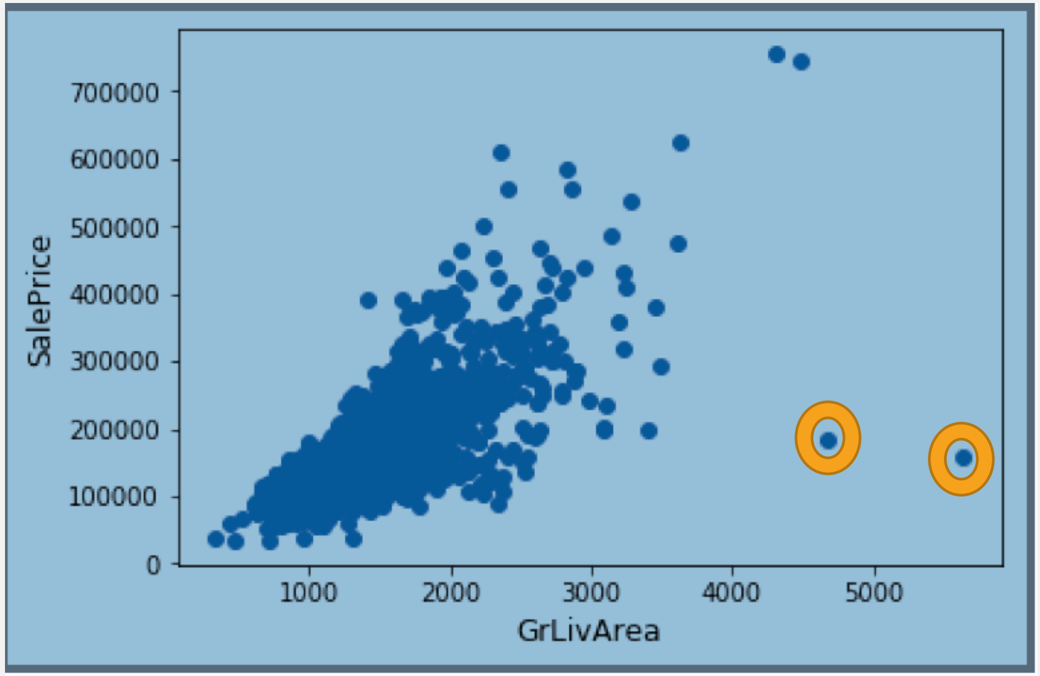
-- 79 explanatory variables : （Approx.all） Features of Houses in Ames, Iowa
-- Objective:predict the sale prices by the features given
-- Software used: Python -- Jupiter notebook

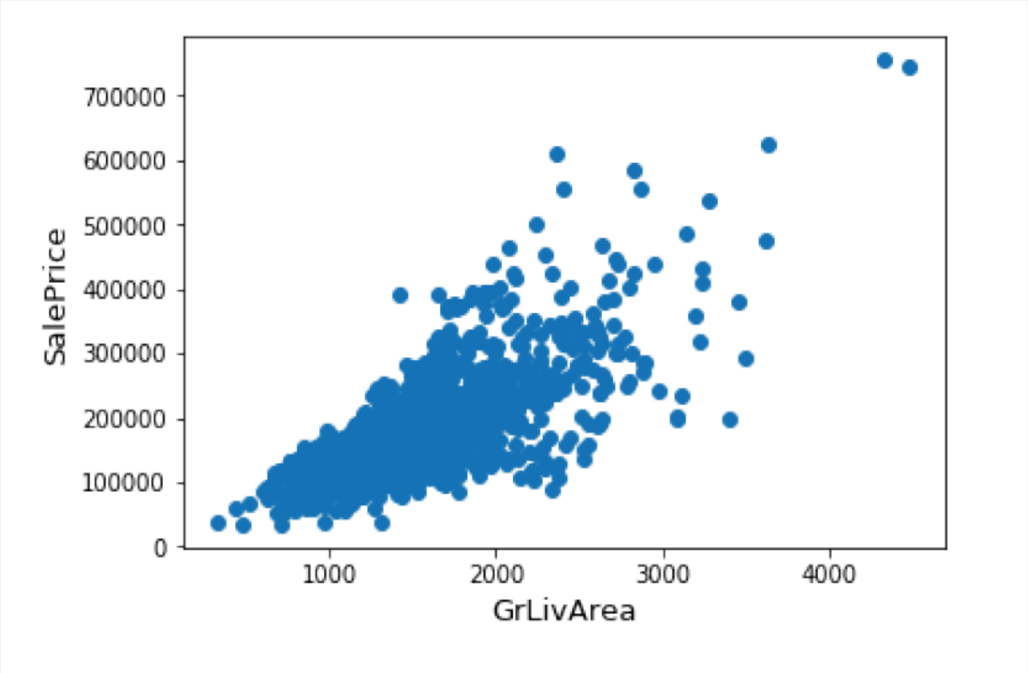
DATA INTUITION
数据假设

- 居住面积越大房价越贵 -- 可能引发你去画一个x轴是面积，y轴是房屋销售价格的散点图 see the next page: GrLivArea
- 房子越老旧越便宜 -- ? Build year
- 房子越…-- ?
- …

DATA CLEANING--OUTLIERS

散点图 -- 数据探索可视化的一种
我们从散点图上能看出一些异常值 我们将异常值限定在常规范围内并进行标记





DATA CLEANING

现实生活中收集的数据并不能直接拿来使用，现在需要进行数据清洗（Data Cleaning）：

常用的清洗流程有哪些呢？

1. 丢失值（missing value）

如果某一列内的信息丢失超过一定百分比，可以选择弃掉整列的数据；如果数据丢失率较低，可以取这列的平均值（mean）/众数（mode）来补上。

2. 异常值（outlier）

画出图像的散点图，可以选择肉眼识别过于不合理的数据，或者用标准的统计学方法；在选出outlier以后，删除此信息所在的案例。

3. 相关性（correlation）

确定特征（feature）之间是否有正或者反的线性关系，如果有，去除掉相对不重要的特征。比如预测房价的信息有房屋建成年份和车库建成年份，这两个之间如果有线性关系，可以去除掉车库建成年份这个相对较窄的信息。

4. 调整为正态分布（normal distribution）

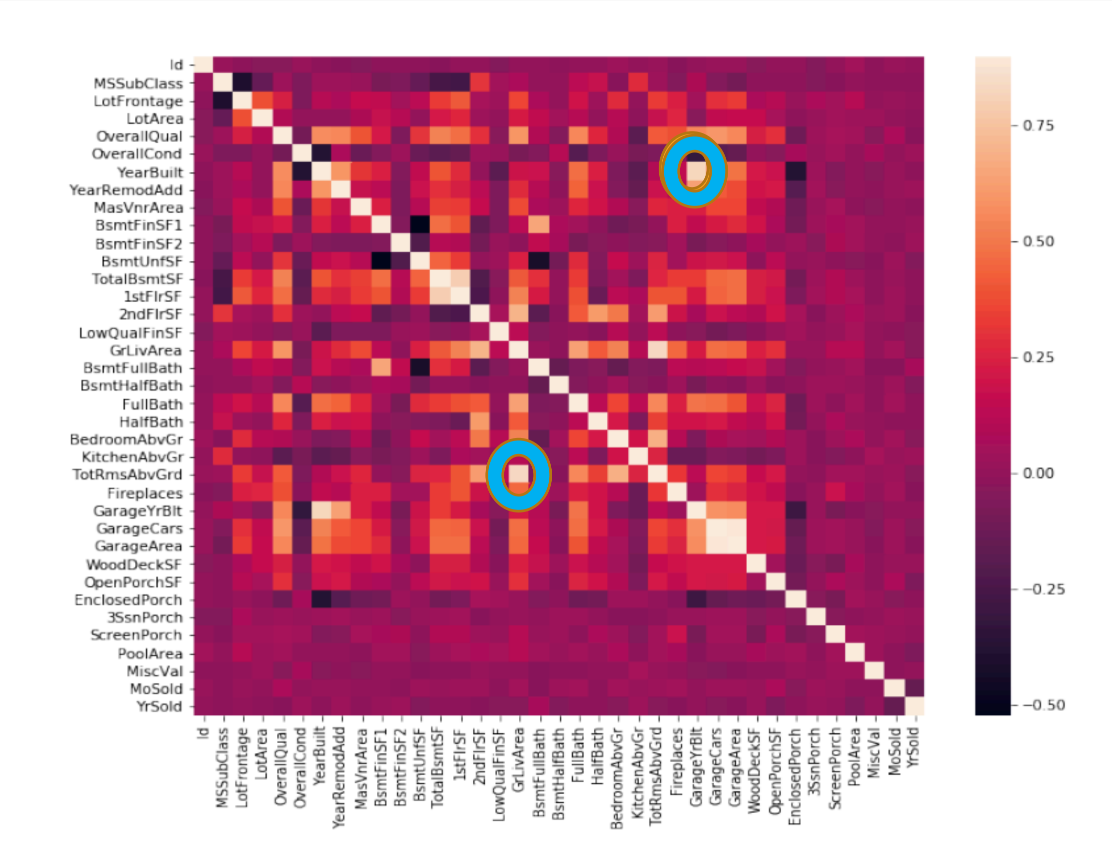
很多统计学的性质都是建立在正态分布上的，在对数据进行变换和修改时，如果数据是正态分布或是修正的正态分布，能够极大的减小误差。（进阶方法）

DATA CLEANING--MISSING VALUES

- NA substitution methods 遗失值替换
- \- Replace by Mean, mode 用中位数或平均值进行替换
- \- Replace by 0 用0替换
- \- Converting categorical data to dummies 用最多的种类构成替换

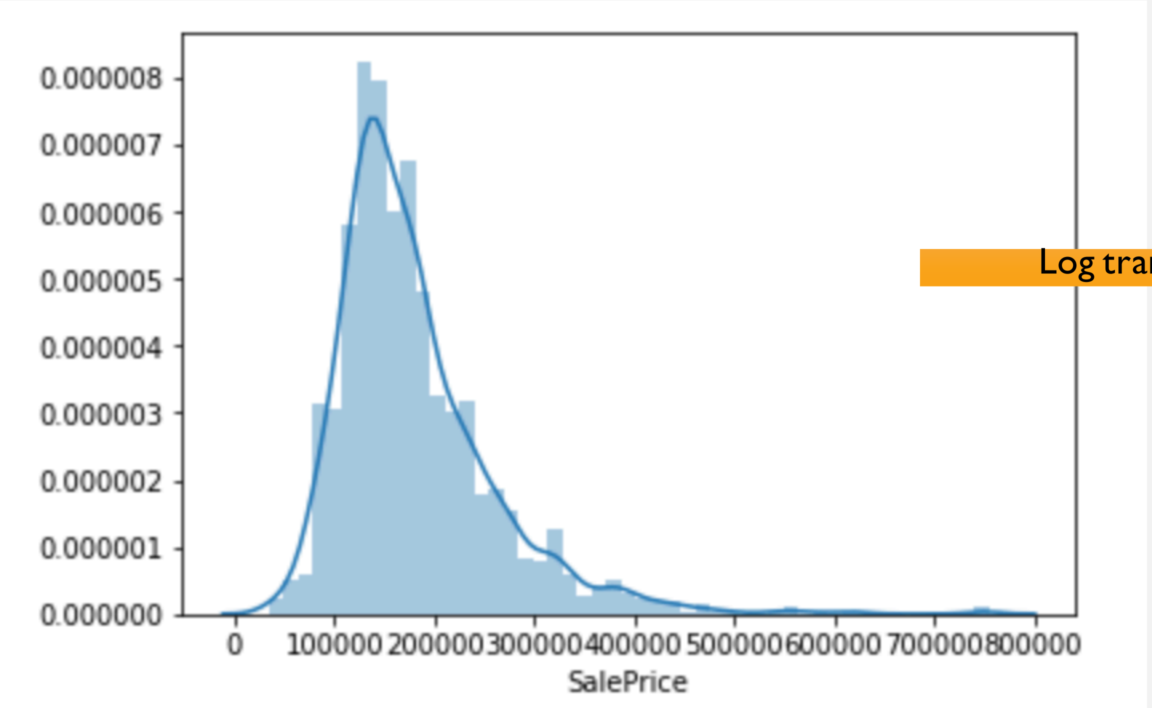
```
# KitchenAbvGr to categorical
features['KitchenAbvGr'] = features['KitchenAbvGr'].astype(str)
# Electrical NA in pred. filling with most popular values
features['Electrical'] = features['Electrical'].fillna(features['Electrical'].mode()[0])
# TotalBsmtSF  NA in pred. I suppose NA means 0
features['TotalBsmtSF'] = features['TotalBsmtSF'].fillna(0)
```

CORRELATIONS—AVOIDING COLLINEARITY

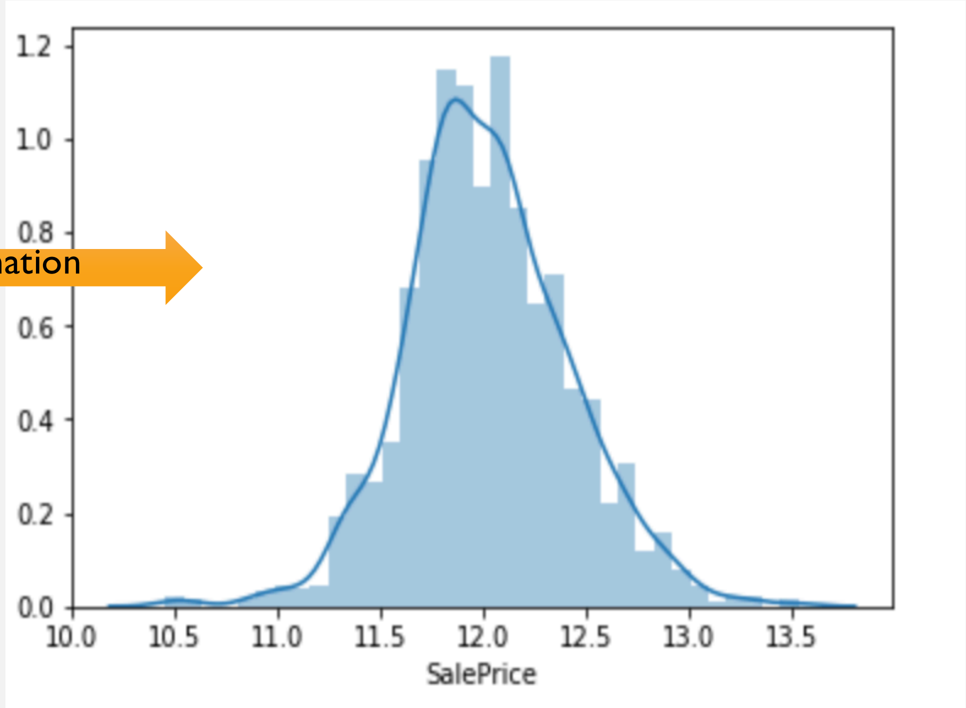


找到相关性强的变量，
看彼此之间可否取代，
缩小training set

DATA CLEANING DISTRIBUTION TRANSFORMATION



Log transformation



MODEL TRAINING

- Model： Linear regression model 线性回归（多变量）
- Algorithm: Gradient Descent or Random Forest
- Advanced Algorithm: Gradient Boost， Elastic Net , Random Forest…
- Try different learning rates and do more iterations

MODEL TEST – REGRESSION

- In regression model, the most commonly known evaluation metrics include:
- **R-squared** (R2), which is the proportion of variation in the outcome that is explained by the predictor variables. In multiple regression models, R2 corresponds to the squared correlation between the observed outcome values and the predicted values by the model. The Higher the R-squared, the better the model.
- **Adjusted R-squared**, which adjusts the R2 for having too many variables in the model. Not sensitive to sample size.
- **Root Mean Squared Error** (RMSE), which measures the average error performed by the model in predicting the outcome for an observation. MSE = mean((observeds - predicteds)^2) and RMSE = sqrt(MSE). The lower the RMSE, the better the model.
- **Residual Standard Error** (RSE), also known as the *model sigma*, is a variant of the RMSE adjusted for the number of predictors in the model. The lower the RSE, the better the model.
- **Mean Absolute Error** (MAE), like the RMSE, the MAE measures the prediction error. Mathematically, it is the average absolute difference between observed and predicted outcomes, MAE = mean(abs(observeds - predicteds)). MAE is less sensitive to outliers compared to RMSE.

MODEL TEST – SELECTION

- 在多个模型中对比选择
- Additionally, there are four other important metrics - **AIC**, **AICc**, **BIC** and **Mallows Cp** - that are commonly used for model evaluation and selection. These are an unbiased estimate of the model prediction error MSE. The lower these metrics, he better the model.
- **AIC** stands for (*Akaike's Information Criteria*), a metric developped by the Japanese Statistician, Hirotugu Akaike, 1970. The basic idea of AIC is to penalize the inclusion of additional variables to a model. It adds a penalty that increases the error when including additional terms. The lower the AIC, the better the model.
- **AICc** is a version of AIC corrected for small sample sizes.
- **BIC** (or *Bayesian information criteria*) is a variant of AIC with a stronger penalty for including additional variables to the model.
- **Mallows Cp**: A variant of AIC developed by Colin Mallows.

RETRAINING MODELS

```
: # Retraining models
  GB_model = GBest.fit(train_features, train_labels)
  ENST_model = ENSTest.fit(train_features_st, train_labels)
```

由于本题涉及变量较多，没有一个“标准模型”，
所以我们采用advanced supervise learning technique

选用模型：Gradient Boost(python)
GBDT概述
GBDT（Gradient Boosting Decison Tree）中的树都是回归树，GBDT用来做回归预测，调整后也可以用于分类（设定阈值，大于阈值为正例，反之为负例），可以发现多种有区分性的特征以及特征组合。GBDT是把所有树的结论累加起来做最终结论的。

“SALE PRICE” ESTIMATIONS

```
## Getting our SalePrice estimation
Final_labels = (np.exp(GB_model.predict(test_features)) + np.exp(ENST_model.predict(test_feature
s_st))) / 2
```

Average the result calculated by 2 models

```
print (Final_labels)
```

```
[ 118947.59269167  154399.35950148  177924.15871109 ...,  154936.82476979
  121416.88349977  218399.08316018]
```

FUTURE IMPROVEMENTS/RECOMMENDATION

- 1. 加入除分析历史数据以外的外界shocks： interest rate，供需关系，income level
- 2. 分析公司的市场份额，分析竞争性定价
- 3. … (你可以根据自己的模型提出可能存在的问题，风险和未来改进意见）

THOUGHTS/SHARING 课程感想

- 如果你有什么学习此课程的感想或意见，欢迎与我们分享。
- 你觉得你在课程中学到了什么?
- *如*：在此次课程中我学到了如何将商业问题转化为统计问题，并用数学方法解决它们
- 未来有什么其他希望学习的?