基于房价预测模型的data exploration and cleaning 部分

1.Understand the problem

2.Hypothesis generation

3.Get data

4.Data exploration and data cleaning process from kaggle

# Variable relationship—seaborn

## Correlation matrix(heatmap)

-using heatmaps to explore correlationships among variables

\*\* sometimes find multicollinearity among them, used for feature selection

## ‘saleprice ’ correlation matrix (zoomed (放大版)heatmap style)

## //只选择与target variable相关系数大的

## Scatter plot between the most correlated variables

# Missing data-pd.isnull

首先，将有missing value的feature list出来，并从大到小排序

然后分析，delete what：

Delete principle：delete variable or delete observation?

-  when more than 15% of the data is missing, we should delete the corresponding variable and pretend it never existed.

- not so much related with target variables(for last analysis—correlation coefficient)

-if just one obserbvation value missing, just delete the observation, not the variable.

最后，pd.drop；check一下

**Analysis** of variance (**ANOVA**) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. **ANOVA** checks the impact of one or more f  factors by comparing the means of different samples.

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

Example:

It can be used to determine whether there are any statistically ﻿ significant﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿ differences between average salary among different companies for the job software engineering.﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿

determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

# Out liars

## Univariate analysis—'saleprice’

Standardize data， convert values to have mean of 0 and a standard deviation of 1.

Print(ourter range (low/high) of the distribution)

## Univariate analysis

A set of Scatter plots

最后：delete points

# Focus on target variables

## Normality--sns

* **Histogram** - Kurtosis and skewness.峰度和偏度
* **Normal probability plot** - Data distribution should closely follow the diagonal that represents the normal distribution.

---if not normal, data transformation to solve the problem

for positive skewness, use log transformation

--values have zero cannot using log transformation

To solve: create binary variable, for zero, ignore, for non-zero ,log transform

# Homoscedasticity

--scatter plot

Norm之后，基本都会满足

# dummy variables

*convert categorical variable into dummy*

5.feature engineering

6.model training and evaluation

相关代码

<https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python>

利用xgboost,lasso regression and neutral network 三种算法

<https://www.hackerearth.com/fr/practice/machine-learning/machine-learning-projects/python-project/tutorial/>s