ML final report

Group 18
B06502167 盧志賢、B06901080 楊欣睿、B06901105 劉佳婷

I. Introduction & Motivation

我們對這個題目最有興趣，想要在比起作業有更多特徵的資料中，實作一些 Regression Model，並了解其運作原理．

II. Data Preprocessing

i. 資料理解

可將 79 個 feature 大致分為以下五類：

1.)建物面積：1stFlrSF(一樓面積)、GrLivArea(一樓以上總生活空間)、TotalBsmtSF(地下室總面積)等．

2.)裝置數量：FullBath、Kitchen、Bedroom 數量．

3.)量表型類別特徵：ExterCond(外部狀態)、ExterQual(外觀材質)、GarageQual(車庫品質)等，這類型特徵的值為 Excellent/Good/Average/Fair/Poor 等具有順序特性．

4.)一般類別特徵：Neighborhood(所在區域)、HouseStyle(房屋型式)等．

5.)日期相關特徵：YrSold(售出年度)、Mosold(售出月份)、GarageYrBlt(車庫建造年度等)．

ii. 數值資料分析

1.) 大概分佈情形

Tools:

- select_dtypes(['int64','float64'])
- describe()

	count	mean	std	...	50%	75%	max
Id	2919.0	1460.000000	842.787043	...	1460.0	2189.5	2919.0
MSSubClass	2919.0	57.137718	42.517628	...	50.0	70.0	190.0
LotFrontage	2433.0	69.305795	23.344905	...	68.0	80.0	313.0
LotArea	2919.0	10168.114080	7886.996359	...	9453.0	11570.0	215245.0
OverallQual	2919.0	6.089072	1.409947	...	6.0	7.0	10.0
OverallCond	2919.0	5.564577	1.113131	...	5.0	6.0	9.0
YearBuilt	2919.0	1971.312778	30.291442	...	1973.0	2001.0	2010.0
YearRemodAdd	2919.0	1984.264474	20.894344	...	1993.0	2004.0	2010.0
MasVnrArea	2896.0	102.201312	179.334253	...	0.0	164.0	1600.0
BsmtFinSF1	2918.0	441.423235	455.610826	...	368.5	733.0	5644.0
BsmtFinSF2	2918.0	49.582248	169.205611	...	0.0	0.0	1526.0
BsmtUnfSF	2918.0	560.772104	439.543659	...	467.0	805.5	2336.0
TotalBsmtSF	2918.0	1051.777587	440.766258	...	989.5	1302.0	6110.0
1stFlrSF	2919.0	1159.581706	392.362079	...	1082.0	1387.5	5095.0
2ndFlrSF	2919.0	336.483727	428.701456	...	0.0	704.0	2065.0
LowQualFinSF	2919.0	4.694416	46.396825	...	0.0	0.0	1064.0
GrLivArea	2919.0	1500.759849	506.051045	...	1444.0	1743.5	5642.0
BsmtFullBath	2917.0	0.429894	0.524736	...	0.0	1.0	3.0
BsmtHalfBath	2917.0	0.061364	0.245687	...	0.0	0.0	2.0
FullBath	2919.0	1.568003	0.552969	...	2.0	2.0	4.0
HalfBath	2919.0	0.380267	0.502872	...	0.0	1.0	2.0
BedroomAbvGr	2919.0	2.860226	0.822693	...	3.0	3.0	8.0
KitchenAbvGr	2919.0	1.044536	0.214462	...	1.0	1.0	3.0
TotRmsAbvGrd	2919.0	6.451524	1.569379	...	6.0	7.0	15.0
Fireplaces	2919.0	0.597122	0.646129	...	1.0	1.0	4.0
GarageYrBlt	2760.0	1978.113406	25.574285	...	1979.0	2002.0	2207.0
GarageCars	2918.0	1.766621	0.761624	...	2.0	2.0	5.0
GarageArea	2918.0	472.874572	215.394815	...	480.0	576.0	1488.0
WoodDeckSF	2919.0	93.709832	126.526589	...	0.0	168.0	1424.0
OpenPorchSF	2919.0	47.486811	67.575493	...	26.0	70.0	742.0
EnclosedPorch	2919.0	23.098321	64.244246	...	0.0	0.0	1012.0
3SsnPorch	2919.0	2.602261	25.188169	...	0.0	0.0	508.0
ScreenPorch	2919.0	16.062350	56.184365	...	0.0	0.0	576.0
PoolArea	2919.0	2.251799	35.663946	...	0.0	0.0	800.0
MiscVal	2919.0	50.825968	567.402211	...	0.0	0.0	17000.0
MoSold	2919.0	6.213087	2.714762	...	6.0	8.0	12.0
YrSold	2919.0	2007.792737	1.314964	...	2008.0	2009.0	2010.0

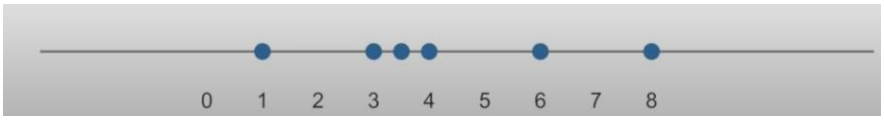
2.) 詳細分佈情形

Tools: sns.distplot()

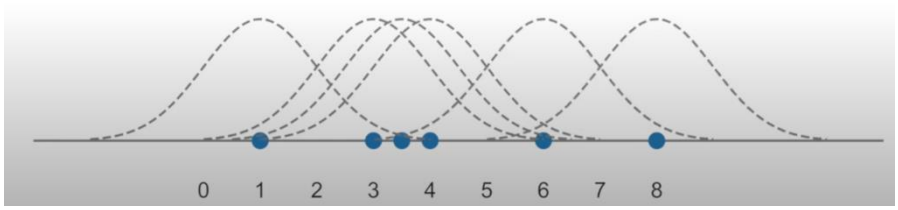
- Matplotlib 的 hist() + sns.kdeplot()

<Note>: Kernel Density Estimation

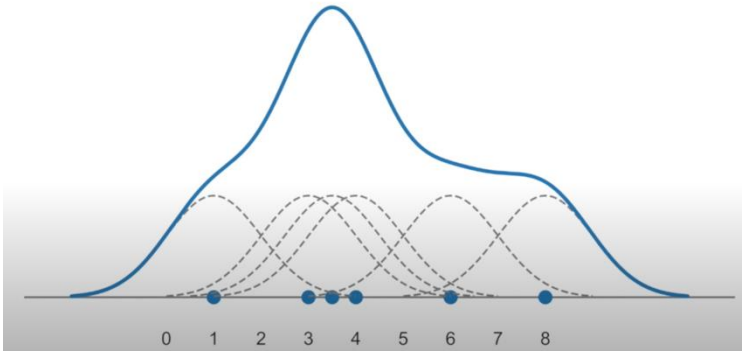
- Estimate probability density function
- Non-parametric: not assuming underlying distribution
- Intuitive interpretation



(Assume there are six data points)

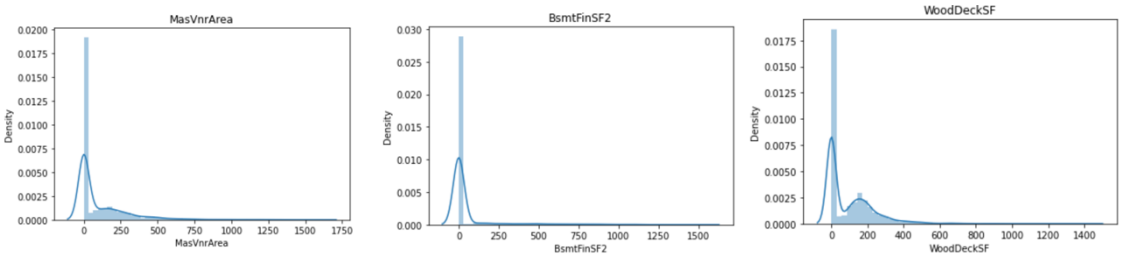


（以 6 個資料點為中心加 kernel，Seaborn 的 default 是 Gaussian，也可以加其他 kernel，像是 triangular kernel 或是 cosine kernel，資料點越多，kernel 種類的選取就愈不重要。）



（將所有 kernel 疊加就可以得到類似機率分佈的函數。）

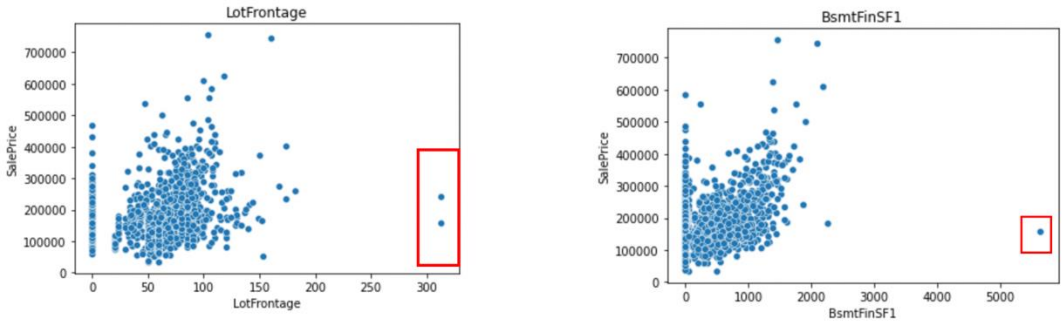
✧ **極端值**



✧ 有大量值為 0，轉成 categorical data（有/無）

✧ 極端值處理

- 極端值對於小資料集（training+testing 不到 3000）的模型效果影響極大
- 畫出各數值特徵與房價的二維散佈分佈圖



(sns.scatterplot)

- 設 threshold 移除極端值

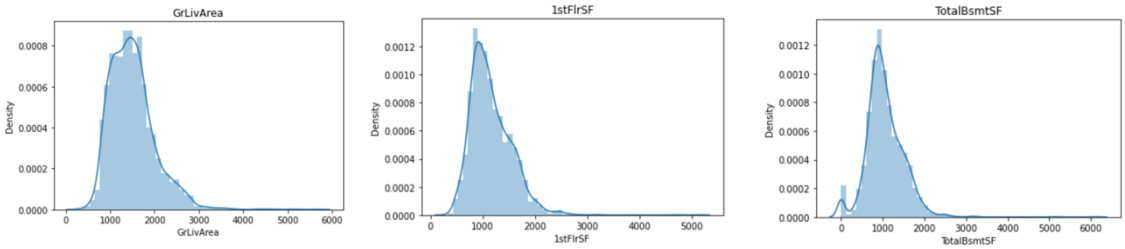
✧ Correlation heatmap：觀察數值特徵和房價的相關係數

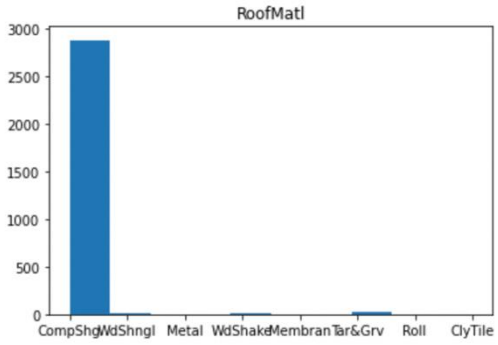
Tools:

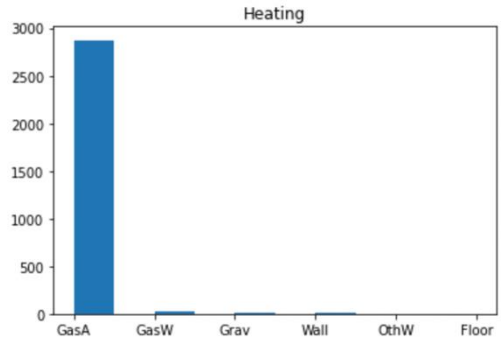
- corr()
- sns.heatmap()

最相關

OverallQual	0.790982
GrLivArea	0.708624
GarageCars	0.640409
GarageArea	0.623431
TotalBsmtSF	0.613581
1stFlrSF	0.605852
FullBath	0.560664
TotRmsAbvGrd	0.533723
YearBuilt	0.522897
YearRemodAdd	0.507101







iii. 類別資料分析

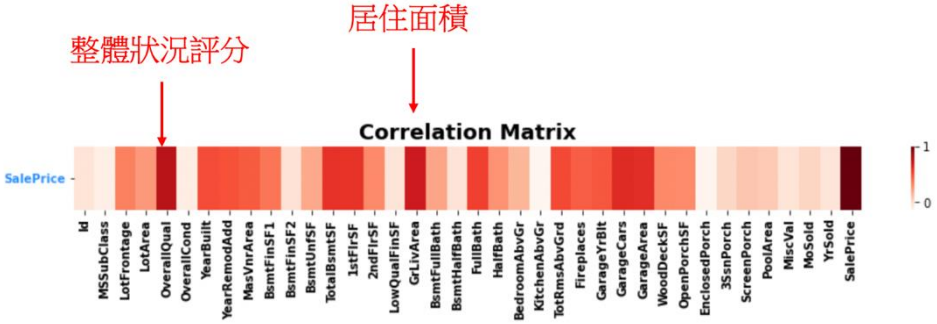
Tools:

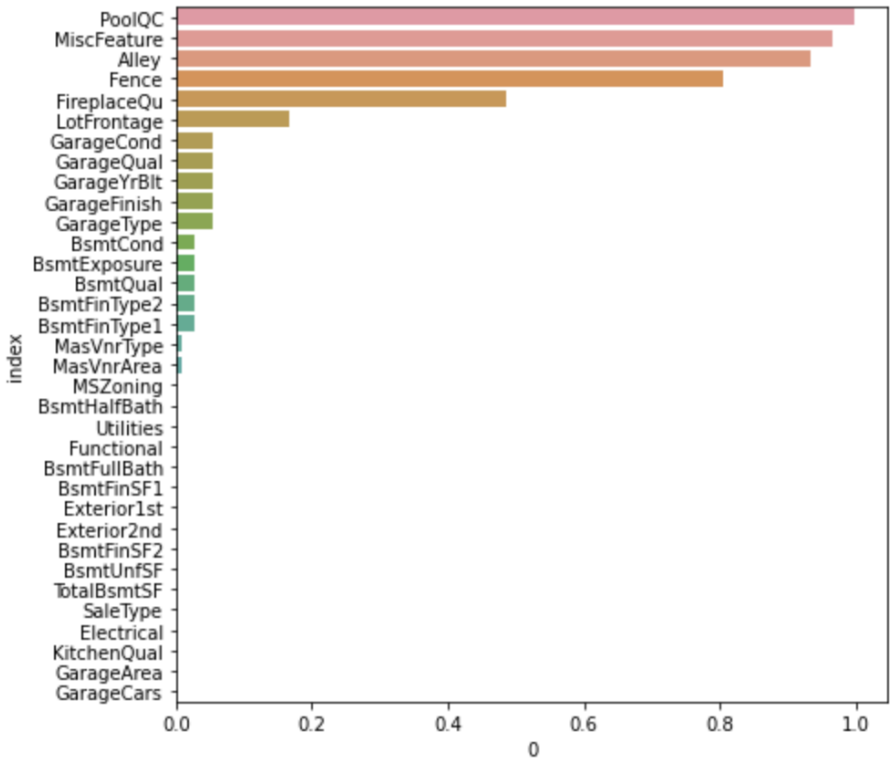
- select_dtypes([‘object’])
- hist()

✧ 過度集中單一欄位：考慮排除

✧ 遺漏值處理

- 計算個別特徵的遺漏率，排序後繪圖，找出遺漏率特別大的特徵．
- Tools: isnull().mean().sort_values(), fillna()





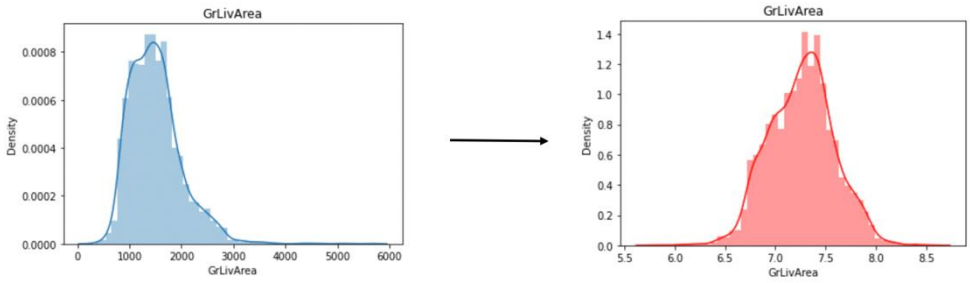
- 分三種情況處理
  1. 遺漏率 > 80 % → 移除
  2. 數值特徵補 0，類別特徵補 None
  3. 車庫建造年(GarageYrBlt)補中位數

III. Feature Engineering

i. 對數轉換

面積相關的特徵，大多為左偏分佈，故全部對數化處理，使其趨近常態分佈．

Tools: np.log1p()



ii. 類別特徵轉順序特徵

```
ExterQual: Evaluates the quality of the material on the exterior

   5
   4  Ex       Excellent
   3  Gd       Good
   2  TA       Average/Typical
   1  Fa       Fair
   0  Po       Poor                    None:0
```

```
BsmtExposure: Refers to walkout or garden level walls

   4  Gd       Good Exposure
   3  Av       Average Exposure (split levels or foyers typically score average or above)
   2
   1  Mn       Mimimum Exposure
   0  No       No Exposure
      NA       No Basement
```

```
BsmtFinType1: Rating of basement finished area

   6
   5  GLQ      Good Living Quarters
   4  ALQ      Average Living Quarters
   3  BLQ      Below Average Living Quarters
   2  Rec      Average Rec Room
   1  LwQ      Low Quality
   0  Unf      Unfinshed
      NA       No Basement
```

iii. 年份特徵計算

1.)銷售時屋齡：Yrsold – YearBuilt

2.)多久前整修：Yrsold –YearRemodAdd

3.)幾年前建造車庫：Yrsold - GarageYrBlt

iv. 數值特徵轉類別特徵

Tools: astype(str)

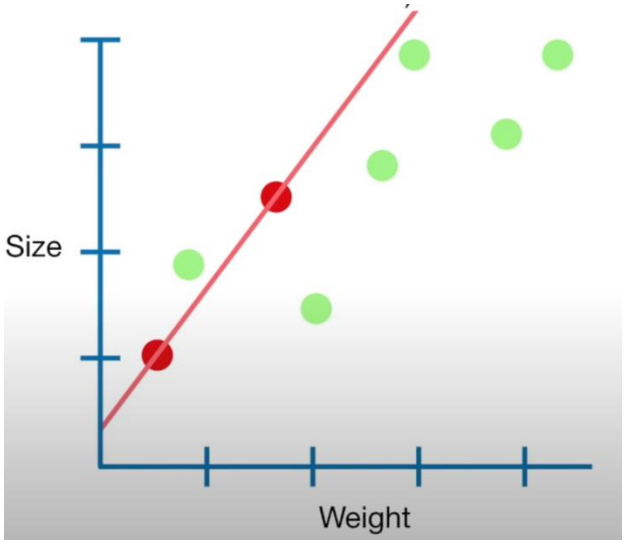
```
MSSubClass: Identifies the type of dwelling involved in the sale.

        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
```

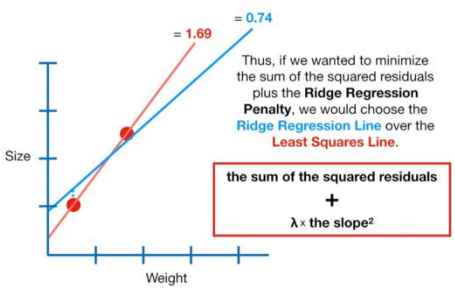
IV. Model Description

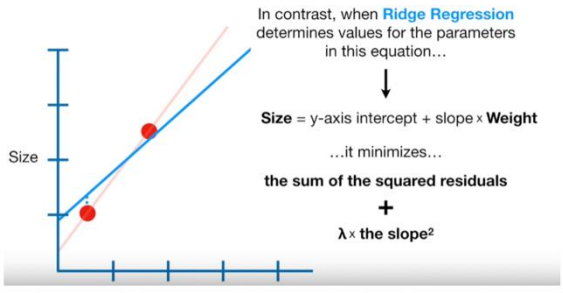
Ridge Regression

- Tool: from sklearn.linear_model import RidgeCV
- Implementation details:
  5-fold cross validation
- alpha=[0.0001,0.001,0.01,0.1,1,10,100]
  (maximizes the generalization score on different subsets of the data)
- Result
  Training RMSE: 0.09574
  Validation RMSE: 0.11697
  Testing RMSE: 0.13155
- Model 原理及特性
  ✧ 能有效避免 overfit



（紅點代表 training data，綠點代表 testing data，顯然 overfit．）





（加入 penalty，讓 regression model 不只 minimize squared residual．）

DNN

以下是我們用 Keras 搭建出來的模型:

```
model = Sequential()

model.add(Dense(input_dim = x_train.shape[1], units = 256, activation = 'relu'))
model.add(Dense(units = 256, activation = 'relu'))
model.add(Dense(units =128, activation = 'relu'))
model.add(Dense(units = 1))
model.compile(loss = 'mse', optimizer = 'adam', metrics = ['mse'])
model.fit(x_train, y_train, batch_size = 200, epochs = 200,validation_data = (x_test,y_test))
```

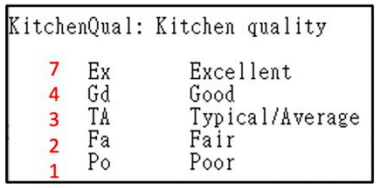
得到的結果為:

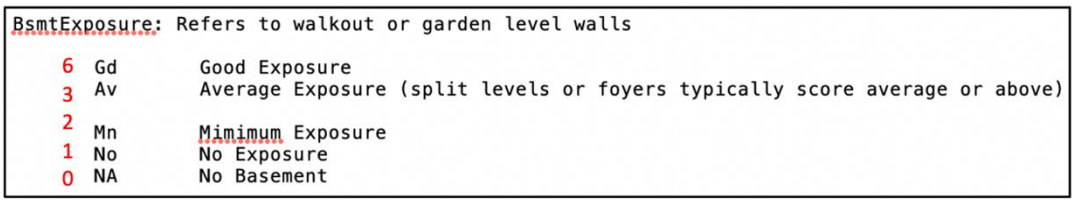
Training RMSE:0.19235

Validation RMSE:0.19313

Testing RMSE:0.19876

Experiment & Discussion





我們發現有三個特徵(車庫、廚房、地下室)，調高最佳品質的分數，可以提昇 performance，如下所示．

我們推測會有如此現象的原因：

廚房的部分，推測因為美國人比起外食較常在家開火，比較重視廚房的品質．至於地下室，美國人較不喜歡打擾到鄰居，因此會把熱水器、洗衣機等噪音較大的物品放在地下室，此外美國人也有裝修地下室的習慣，除了可以當成堆雜物的儲藏室外，也有改造成酒窖、書房等功能，甚至也有人改造成娛樂間、家庭影院．

至於只調高最佳品質的分數才能得到更好的預測結果，可能是因為品質最高的設施通常是更高等級的房子，跟其他房型相比會高出不少房價，是以這樣的調整能讓模型進步．

而在兩種不同模型(RidgeCV 和 DNN)的比較中，在訓練資料數較少，只有 1460 筆的情況下，Ridge Regression 的會較深度網路更容易達到好的結果。

V. Conclusion

在這次的 project 中，可以發現雖然極端值可能存在於現實，但去除之後能夠讓模型更加進步。而如果能在資料處理的部分，先預想哪些的特徵能夠對結果造成較大的影響，進而拉大其數值或差距，可以有更準確的結果。在訓練資料數約在 1000 多筆的情況下，比起使用 DNN 模型，套用 Regression 套件可能可以更輕易地得到更好的模型。

至於是否會在購屋時參考這樣的模型，如果是在人生地不熟的地區購買，會讓這樣的數值模型納入更多考慮。然而在較為熟悉的地區購屋時，可能會考慮一些較無法用數值表示的特徵，例如考量該區的物價、天災發生頻率、治安和環境整潔等問題，或是針對購屋者的狀況納入考量，例如與該購屋者的工作地點的交通方便與否，和購屋者的親友距離等。

VI. Reference

i. Data Preprocessing & Feature Engineering:
https://medium.com/@permoonzz/kaggle-house-prices-advanced-regression-techniques-python-ensemble-learning%E5%AF%A6%E5%81%9A-99f757f4d326

ii. KDE:
https://www.youtube.com/watch?v=DCgPRaIDYXA&t=128s

iii. Ridge Regression:
https://www.youtube.com/watch?v=Q81RR3yKn30&t=387s

iv. Keras:
https://www.youtube.com/watch?v=L8unuZNpWw8