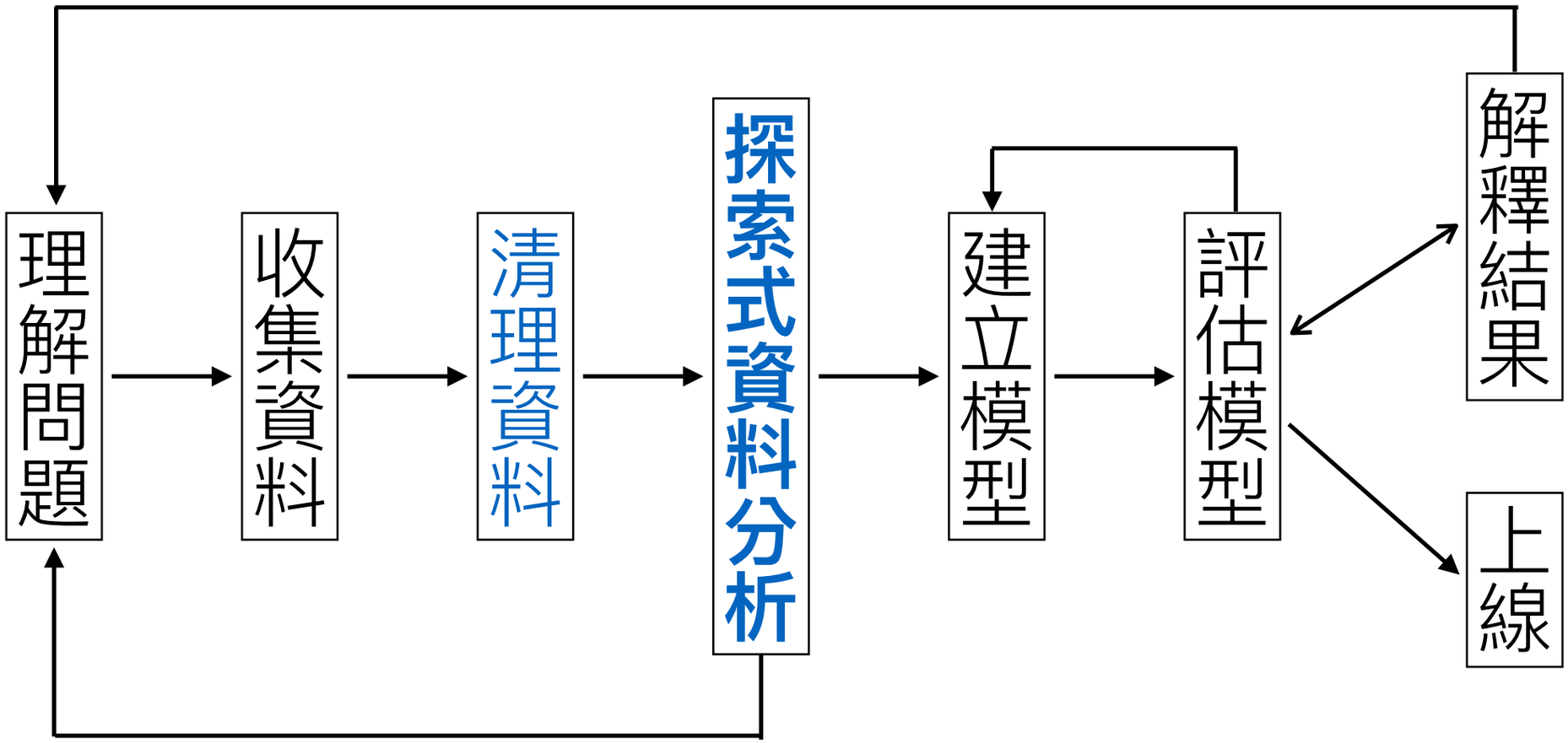
Explorative Data Analysis

探索式資料分析

Data science pipeline



目的

利用敘述統計量及資料視覺化:

- 用眼睛初步分析資料
- 建立研究假說
- 檢查資料異常（e.g. 離群值）
- 檢驗統計模型假設

Example: 台中空氣品質是否改善

- Mean (or median) with standard deviation (or IQR)
- kurtosis and skew
- Box plots of daily means (or 3rd quantile, etc.)
- Histograms (or density plots) of daily means
- Line plots at different times

Example: 空氣品質 vs 風力發電量

- Correlation coefficients
- Scatter plot
- Heat map

Example: 影像辨識

- ~~Pie chart~~
- Bar chart
- Histogram (of gradients, etc.)

Agenda

- 利用Pandas計算敘述統計量
- Data visualization in Python
  - Box plot
  - Histogram and density plot
  - Pie and bar charts
  - Scatter plot
  - Heat map

Descriptive statistics

```
import pandas as pd
df = pd.read_csv('train.csv')
df.info()
```

House Price data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 81 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Id              1460 non-null   int64
 1   MSSubClass      1460 non-null   int64
 2   MSZoning        1460 non-null   object
 3   LotFrontage     1201 non-null   float64
 4   LotArea         1460 non-null   int64
```

Descriptive statistics

```
df['SalePrice'].describe()
```

```
count      1460.000000
mean     180921.195890
std       79442.502883
min       34900.000000
25%      129975.000000
50%      163000.000000
75%      214000.000000
max      755000.000000
Name: SalePrice, dtype: float64
```

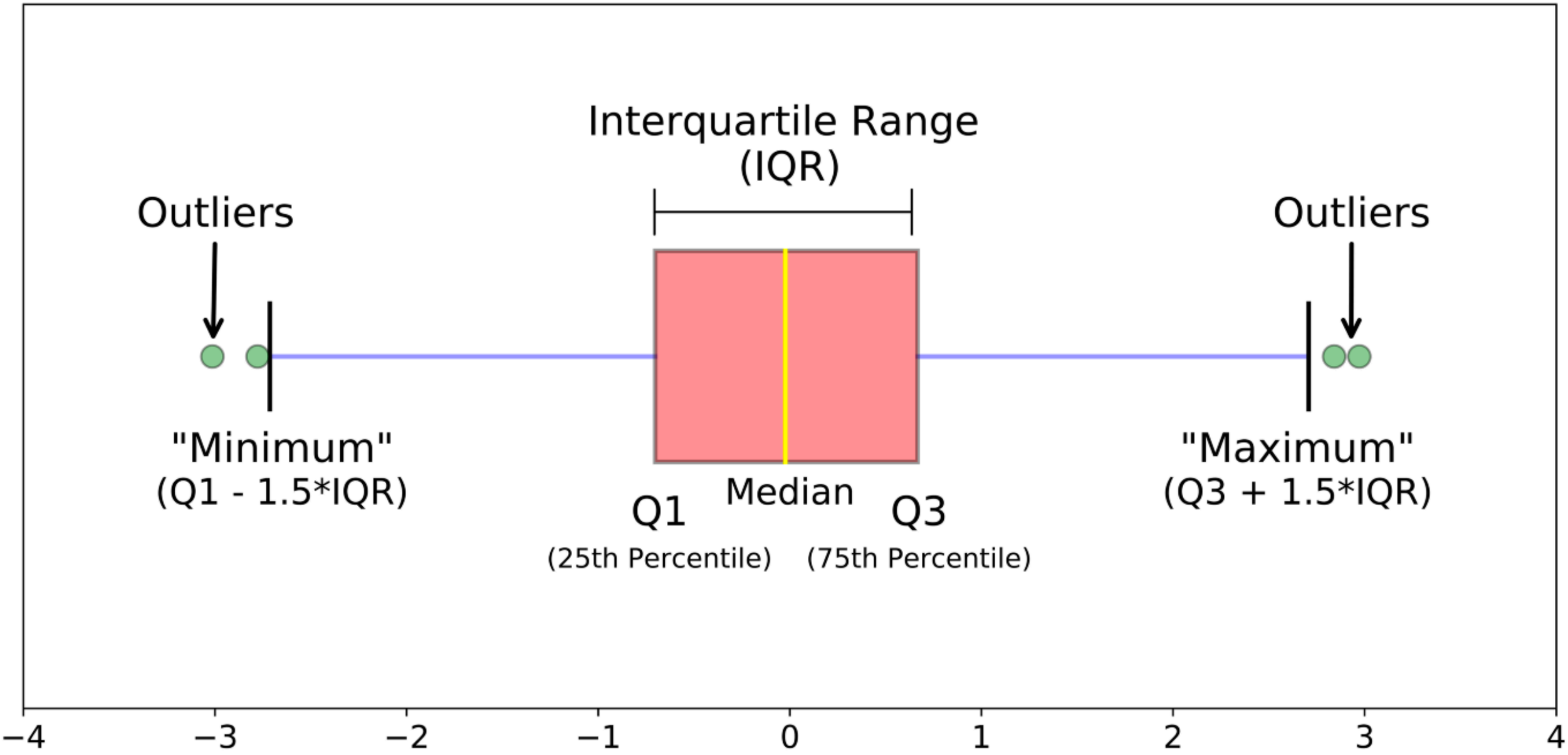
Descriptive statistics

- mean, median, and mode
- min and max
- var and std
- quantile
- apply custom functions

Recap: modules

- **pandas for data manipulation**
- ibis and dask for database and big data
- **matplotlib and seaborn for visualization**
- scikit-learn and statsmodels for machine learning and statistical models
- An IDE such as jupyterlab
- Tensorflow 2 for deep learning

Box plot



https://medium.com/@agarwal.vishal819/outlier-detection-with-boxplots-1b6757fafa21

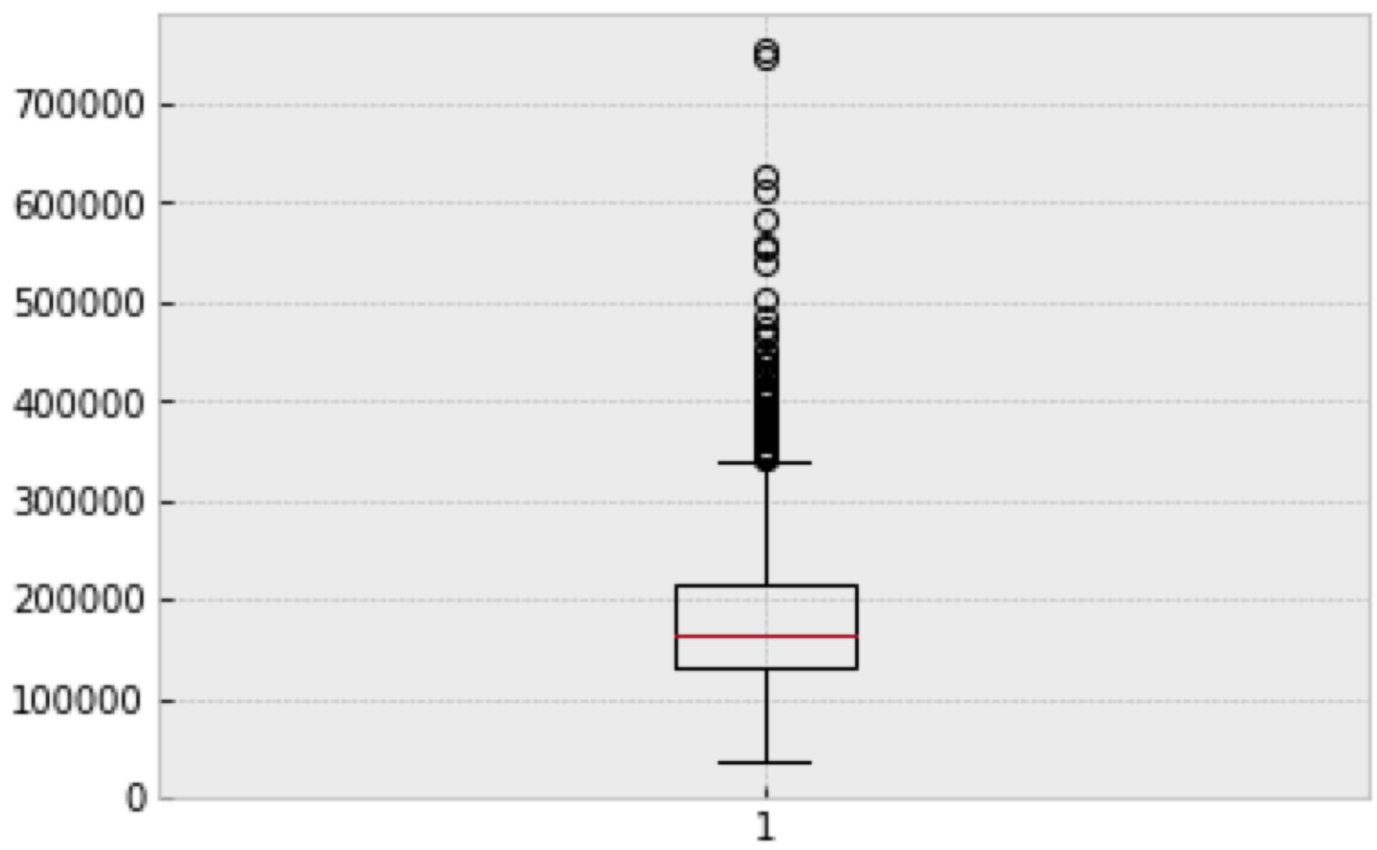
Box plot

```
import matplotlib.pyplot as plt
plt.style.use('bmh')
%matplotlib inline
```

設定matplotlib繪圖風格(style)

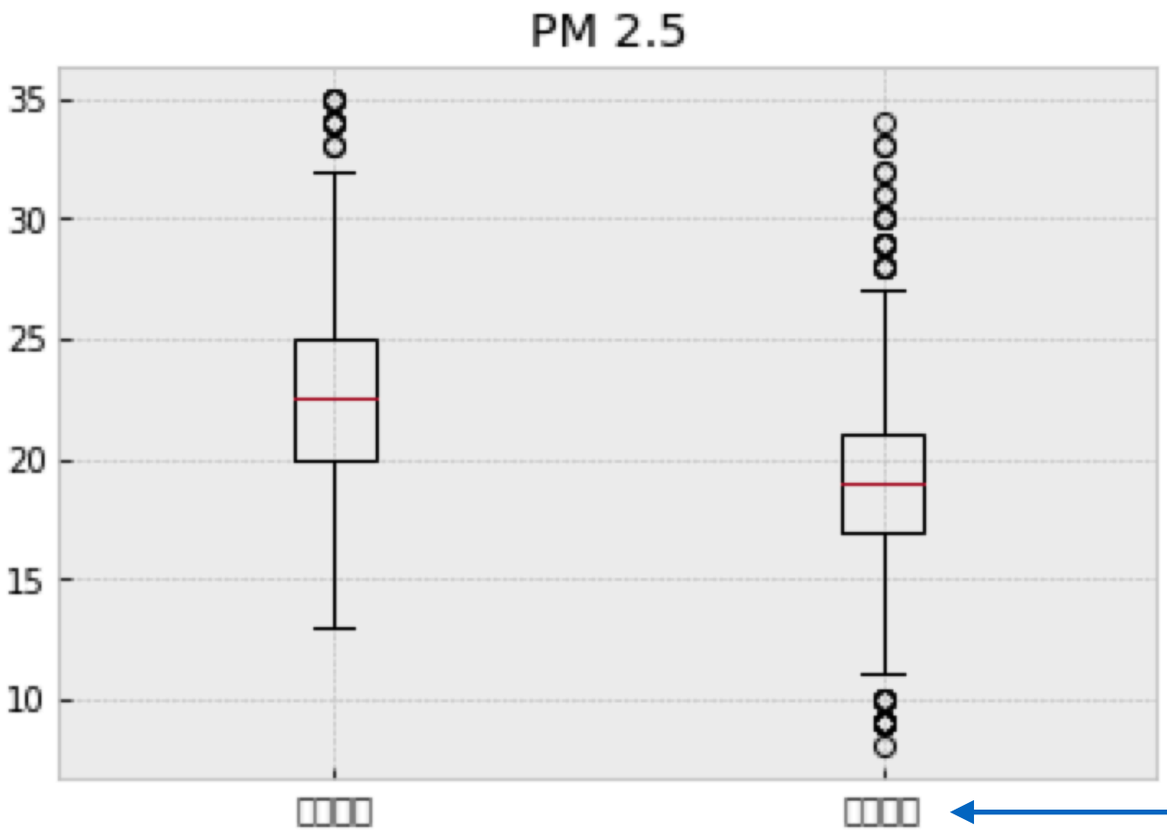
在jupyter notebook中直接顯示圖片

```
plt.boxplot(df['SalePrice'])
plt.show()
```



Box plot

```
df2 = pd.read_csv('iis_airbox_20200912.csv')
x = df2[df2['SiteName'].str.contains('臺中市立國光國小')]['PM25']
y = df2[df2['SiteName'].str.contains('臺中市立省三國小')]['PM25']
fig, ax = plt.subplots()
ax.set_title('PM 2.5')
ax.boxplot([x,y], labels=['國光國小', '省三國小'])
plt.show()
```



中文字形

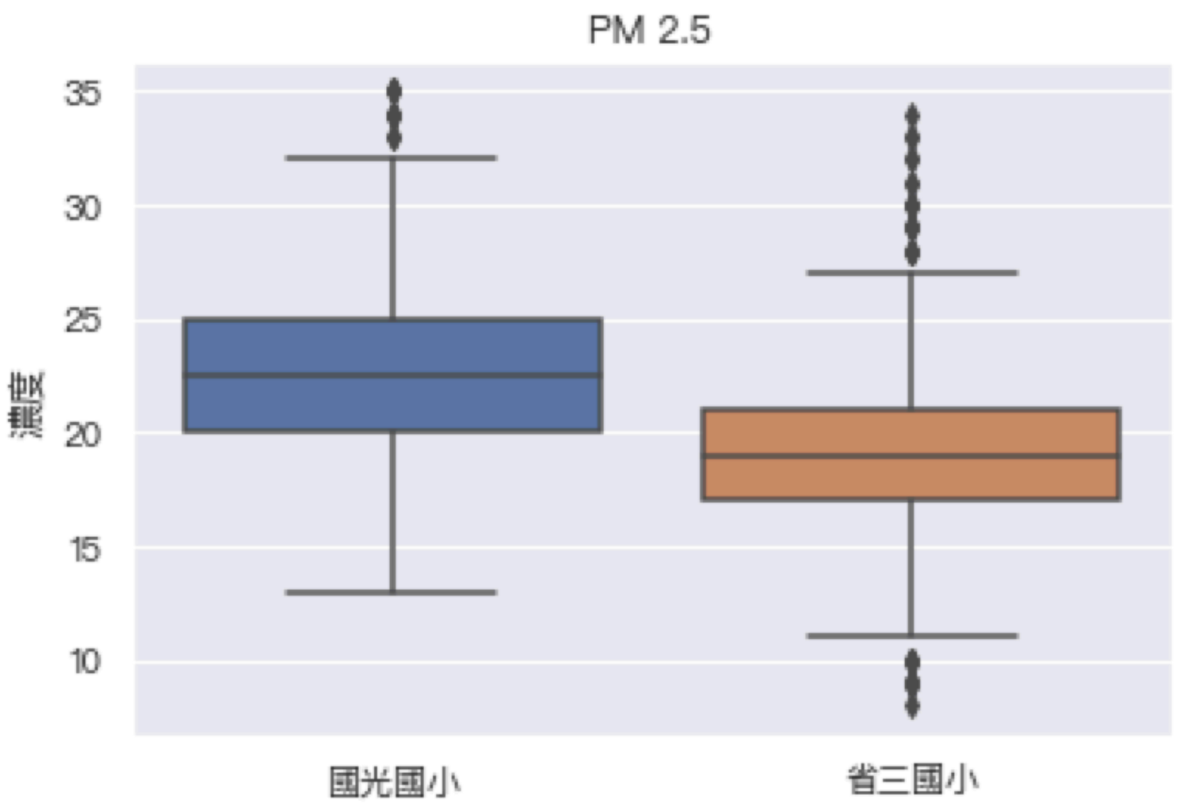
Box plot

```
import seaborn as sns
from matplotlib.font_manager import FontProperties
myfont=FontProperties(fname='/System/Library/Fonts/PingFang.ttc',size=14)
sns.set(font=myfont.get_name())
data = pd.DataFrame({'國光國小':x, '省三國小':y})
sns.boxplot(data=data)
plt.ylabel('濃度')
plt.title('PM 2.5')
plt.show()
```

設定中文字形

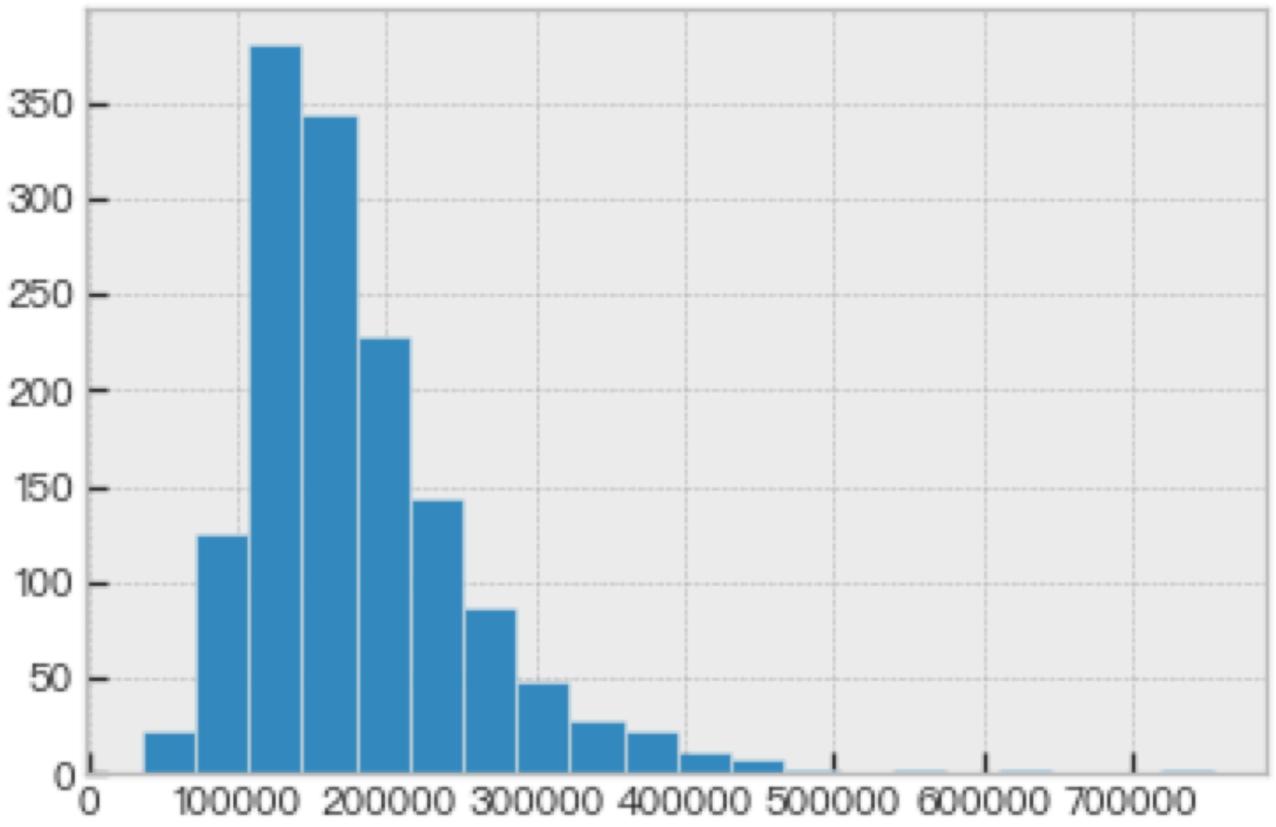
seaborn資料需包成DataFrame

設定title與y軸名稱



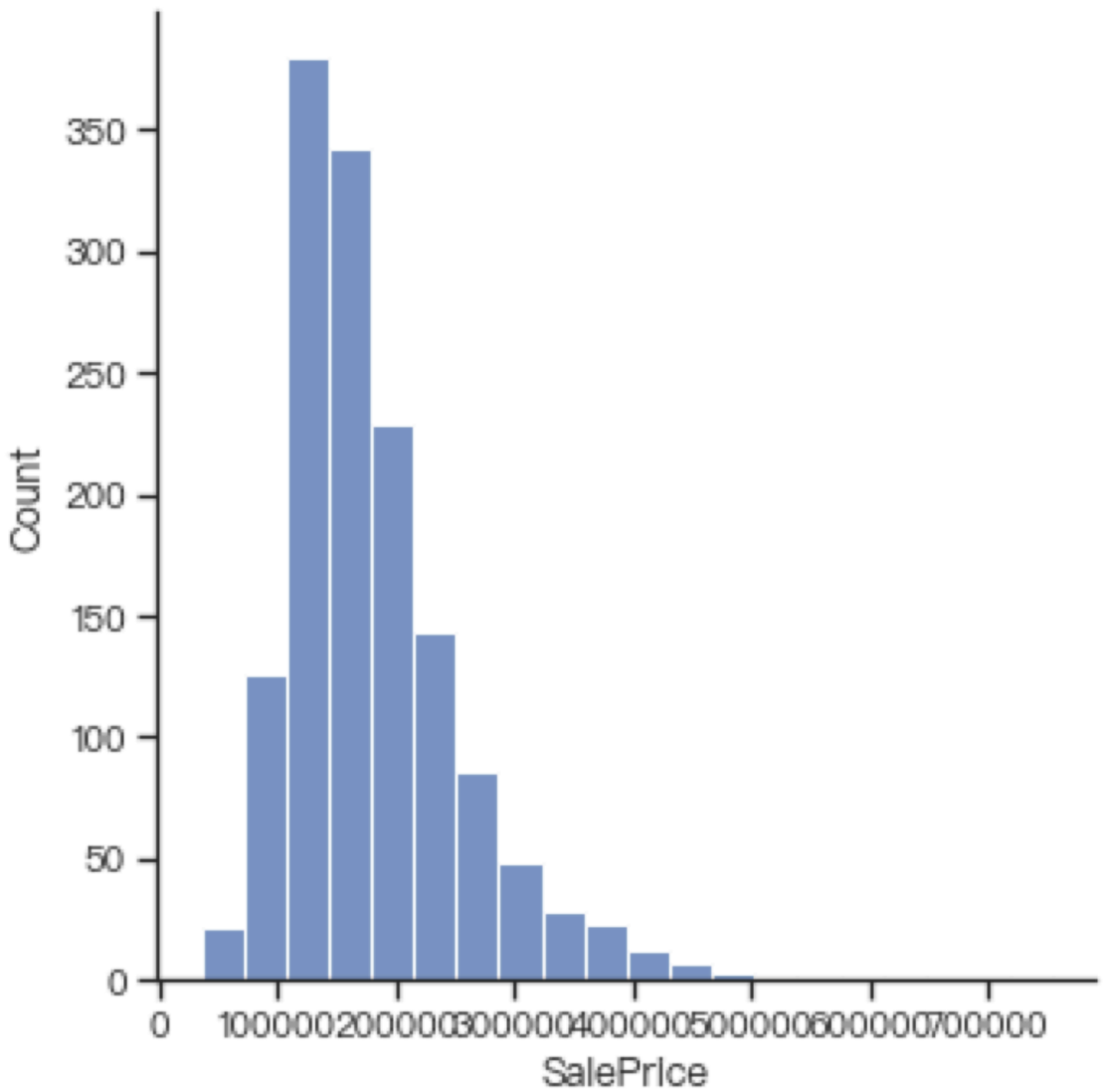
Histogram

```
plt.hist(df['SalePrice'], bins=20)
plt.show()
```



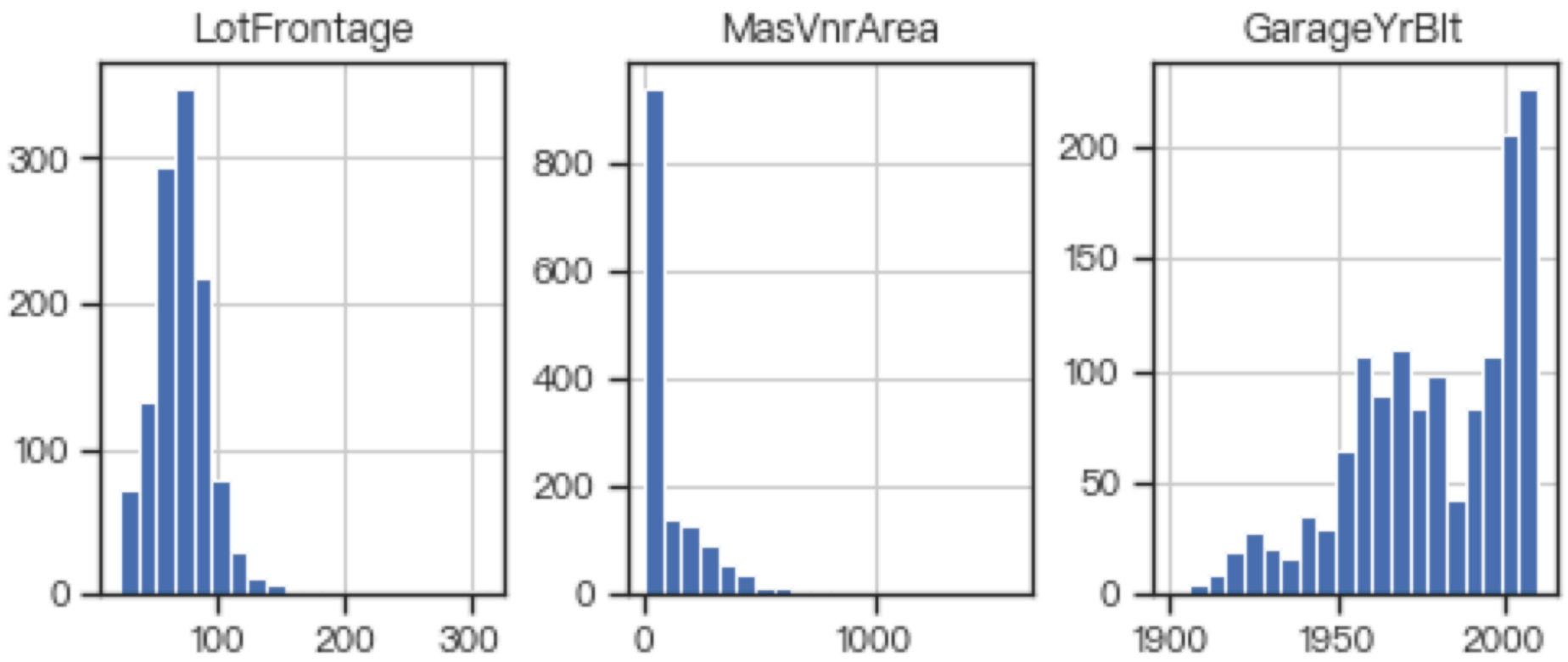
Histogram

```
sns.set_theme(font=myfont.get_name(), style='ticks')
sns.displot(df['SalePrice'], bins=20, kind='hist')
plt.show()
```



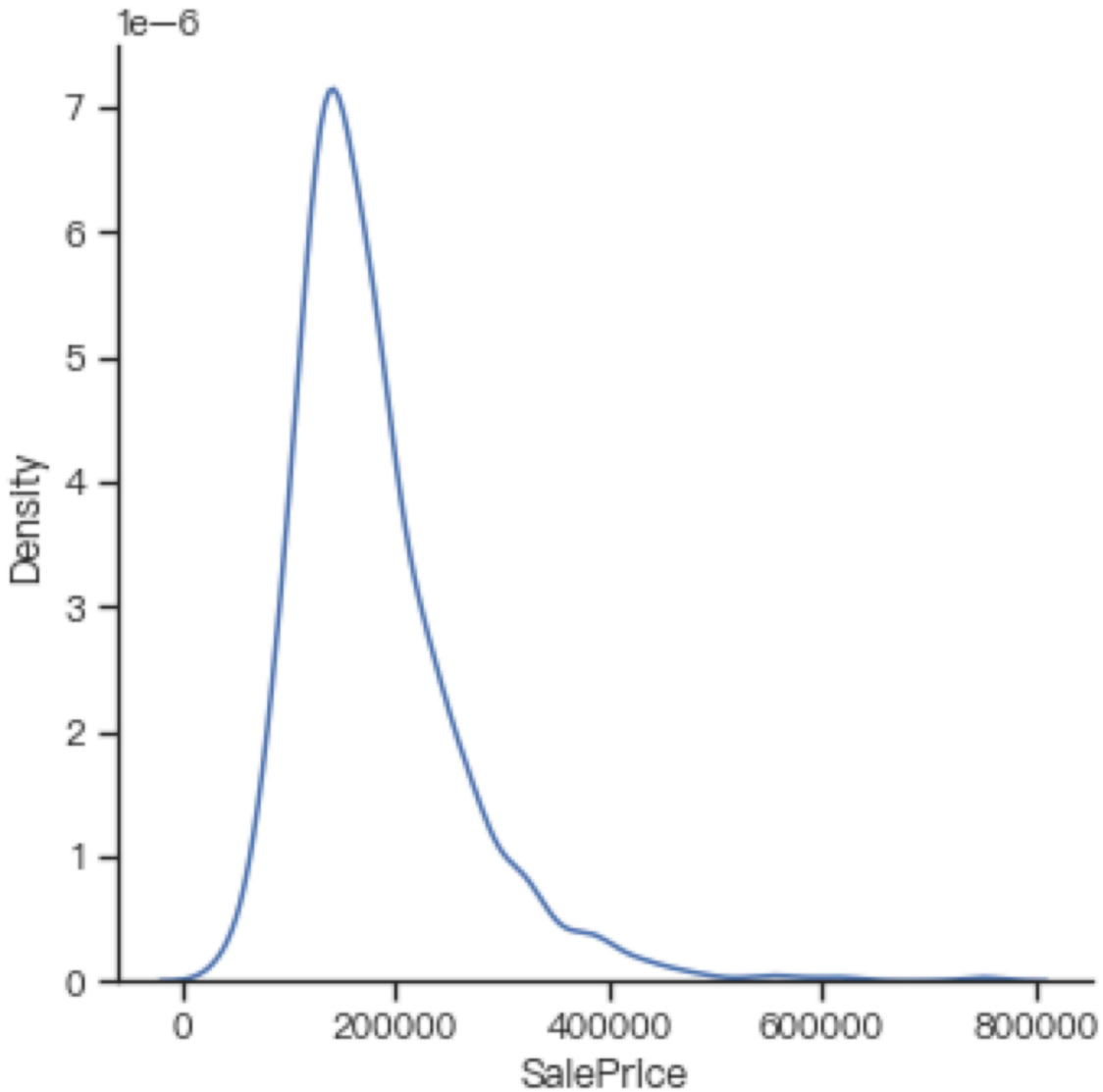
Histogram

```
df_num = df.select_dtypes(include = ['float64'])
df_num.hist(figsize=(8, 3), bins=20, layout=(1,3))
plt.show()
```



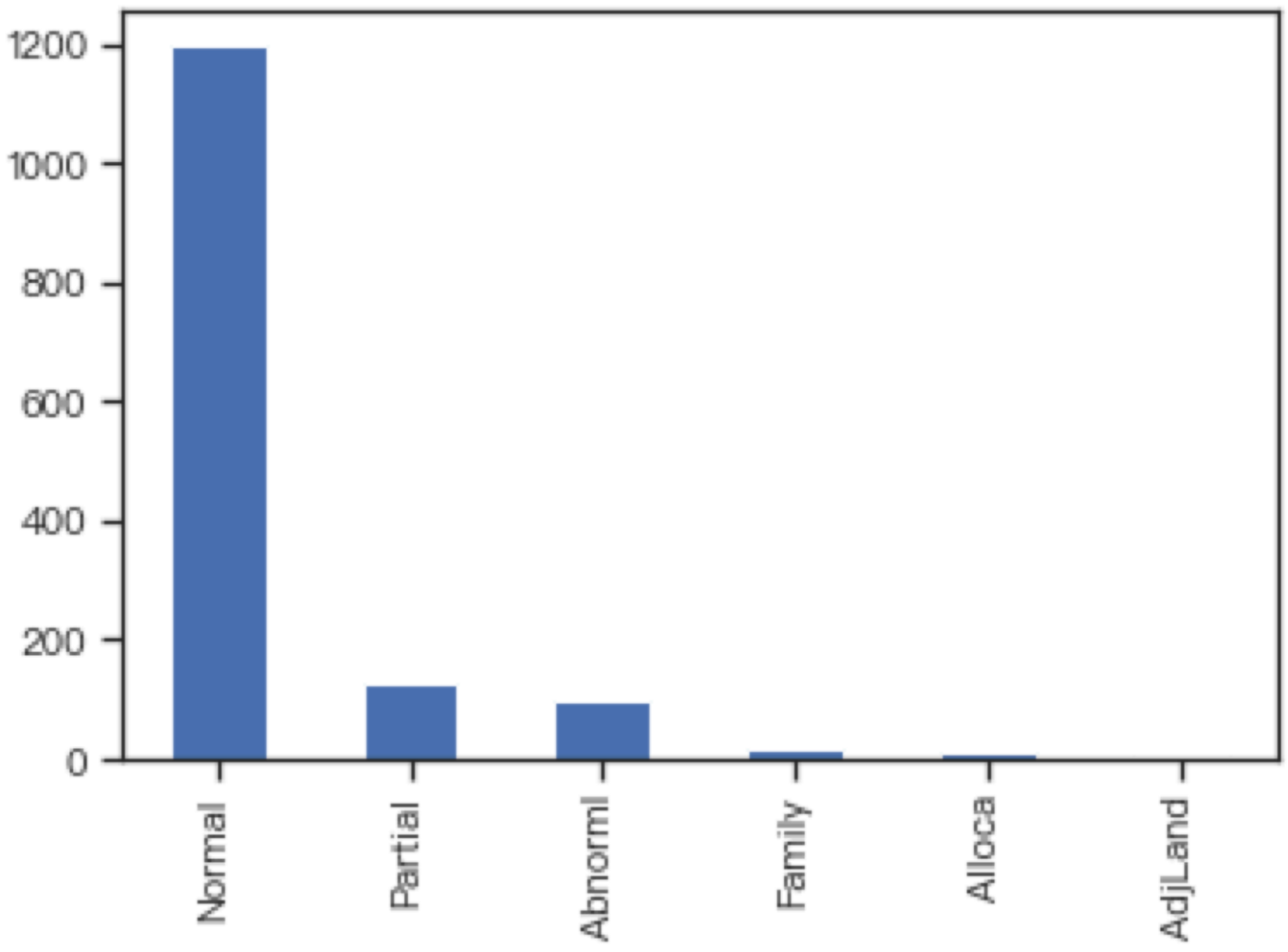
Density plot

```
sns.set_theme(font=myfont.get_name(), style='ticks')
sns.displot(df['SalePrice'], kind='kde')
plt.show()
```



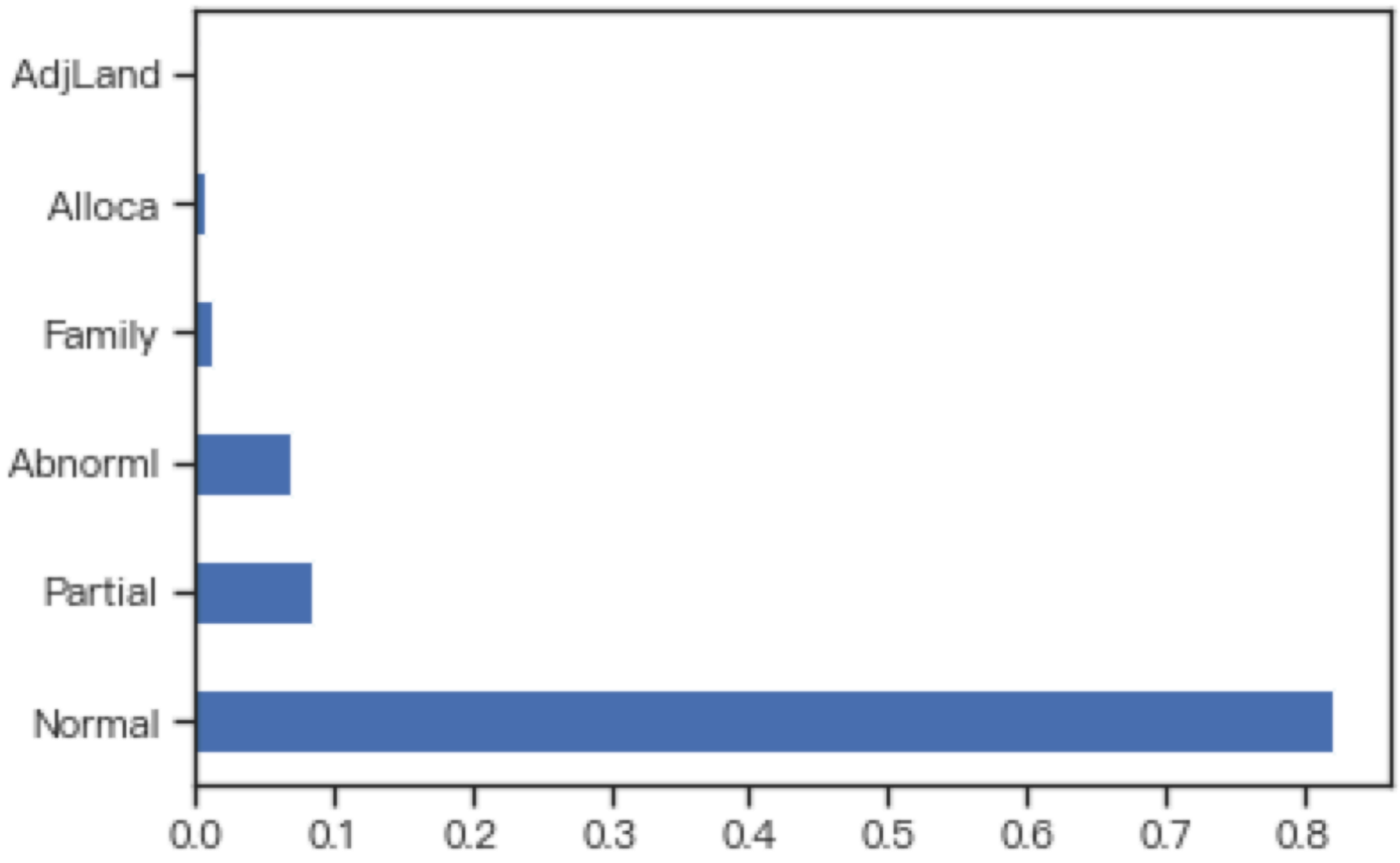
Bar chart

```
df['SaleCondition'].value_counts().plot.bar()
plt.show()
```



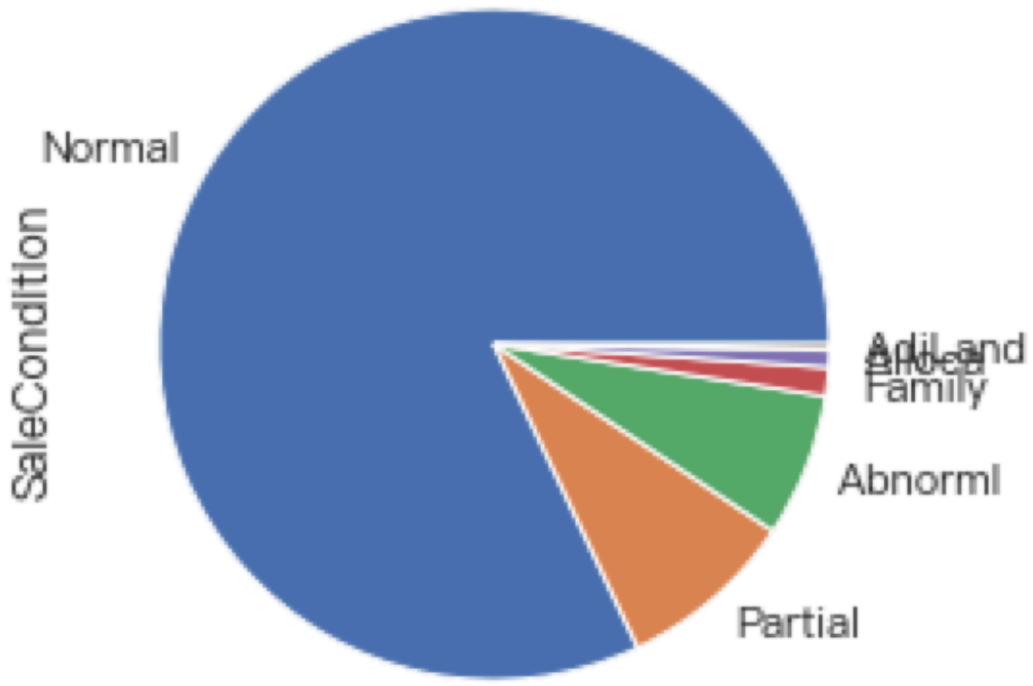
Bar chart

```
df['SaleCondition'].value_counts(normalize=True).plot.barh()
plt.show()
```



Pie chart

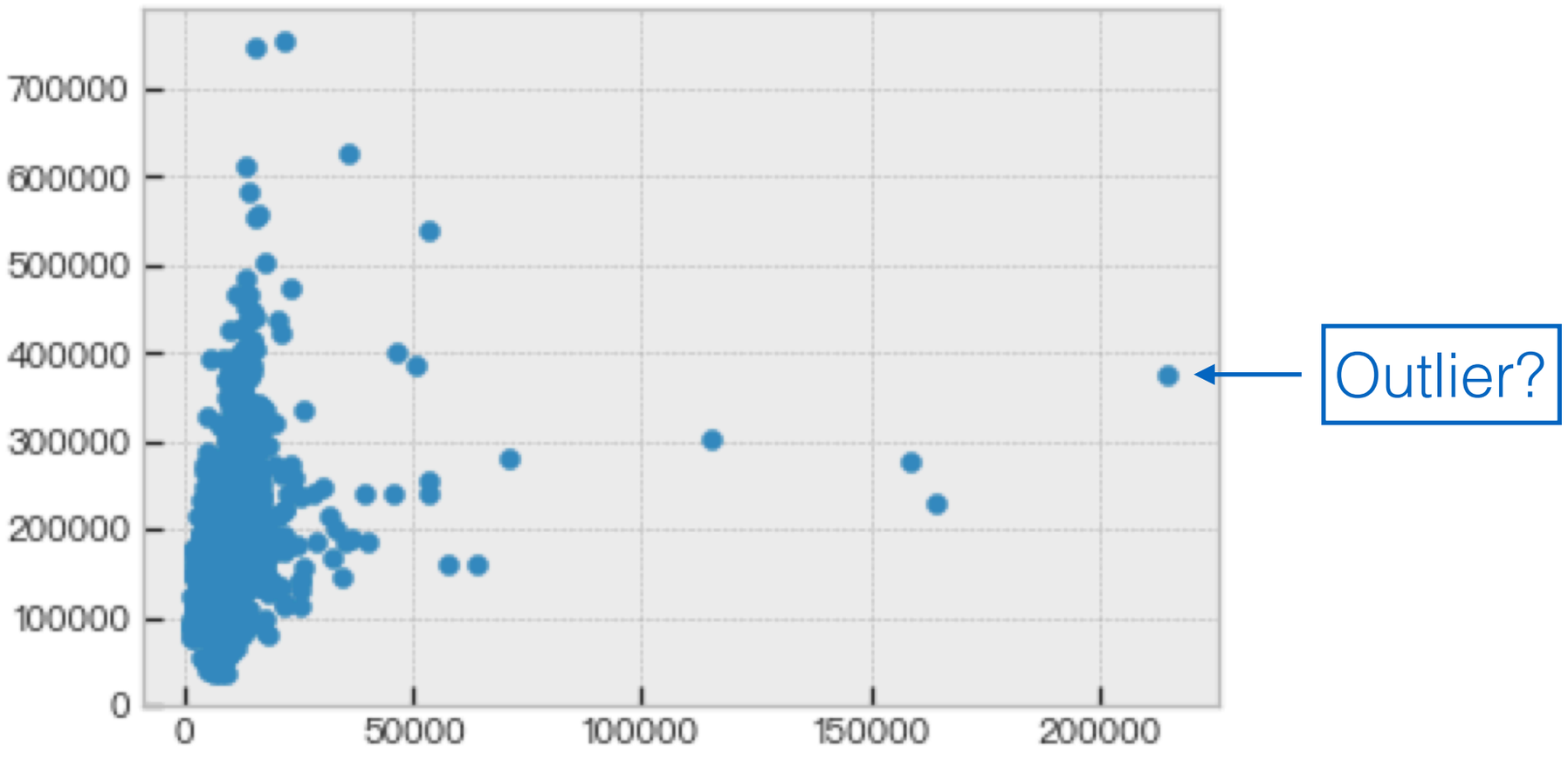
```
df['SaleCondition'].value_counts().plot.pie()
plt.show()
```



盡量少用！人眼對面積大小不敏感！

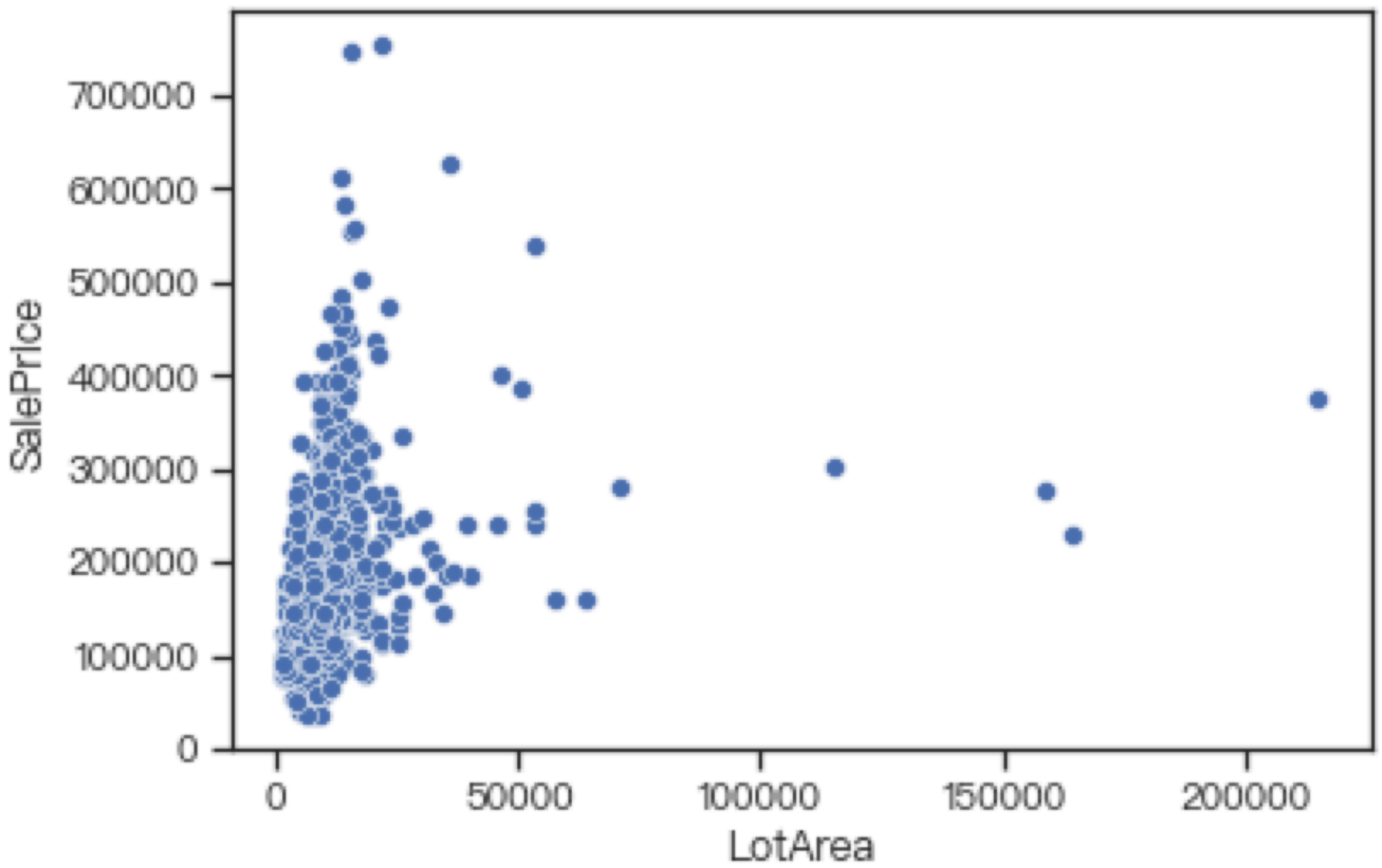
Scatter plot

```
plt.style.use('bmh')
plt.scatter(df['LotArea'],df['SalePrice'])
plt.show()
```



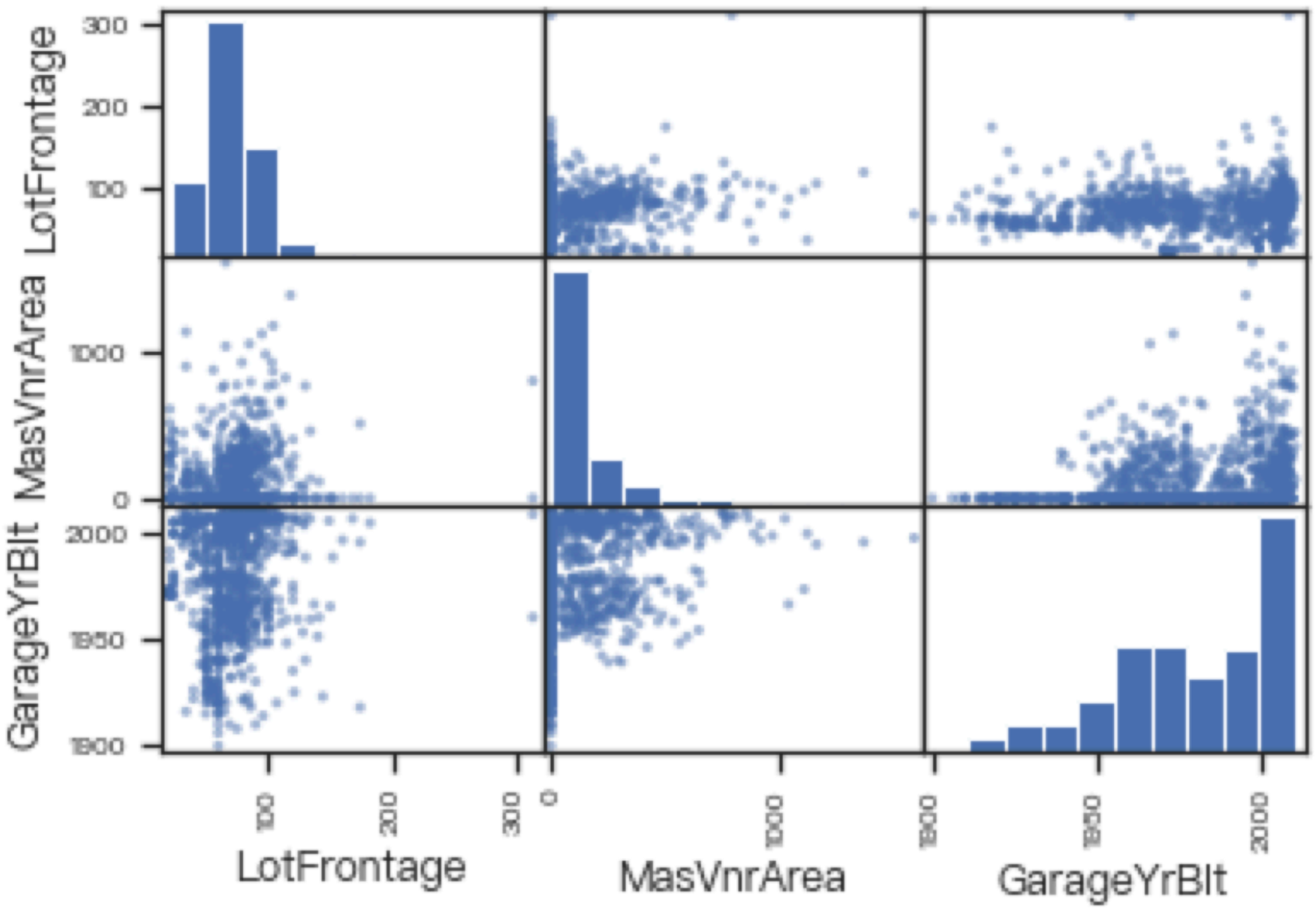
Scatter plot

```
sns.set_theme(font=myfont.get_name(), style='ticks')
sns.scatterplot(x='LotArea', y='SalePrice', data=df)
plt.show()
```



Scatter plot

```
pd.plotting.scatter_matrix(df_num)
plt.show()
```



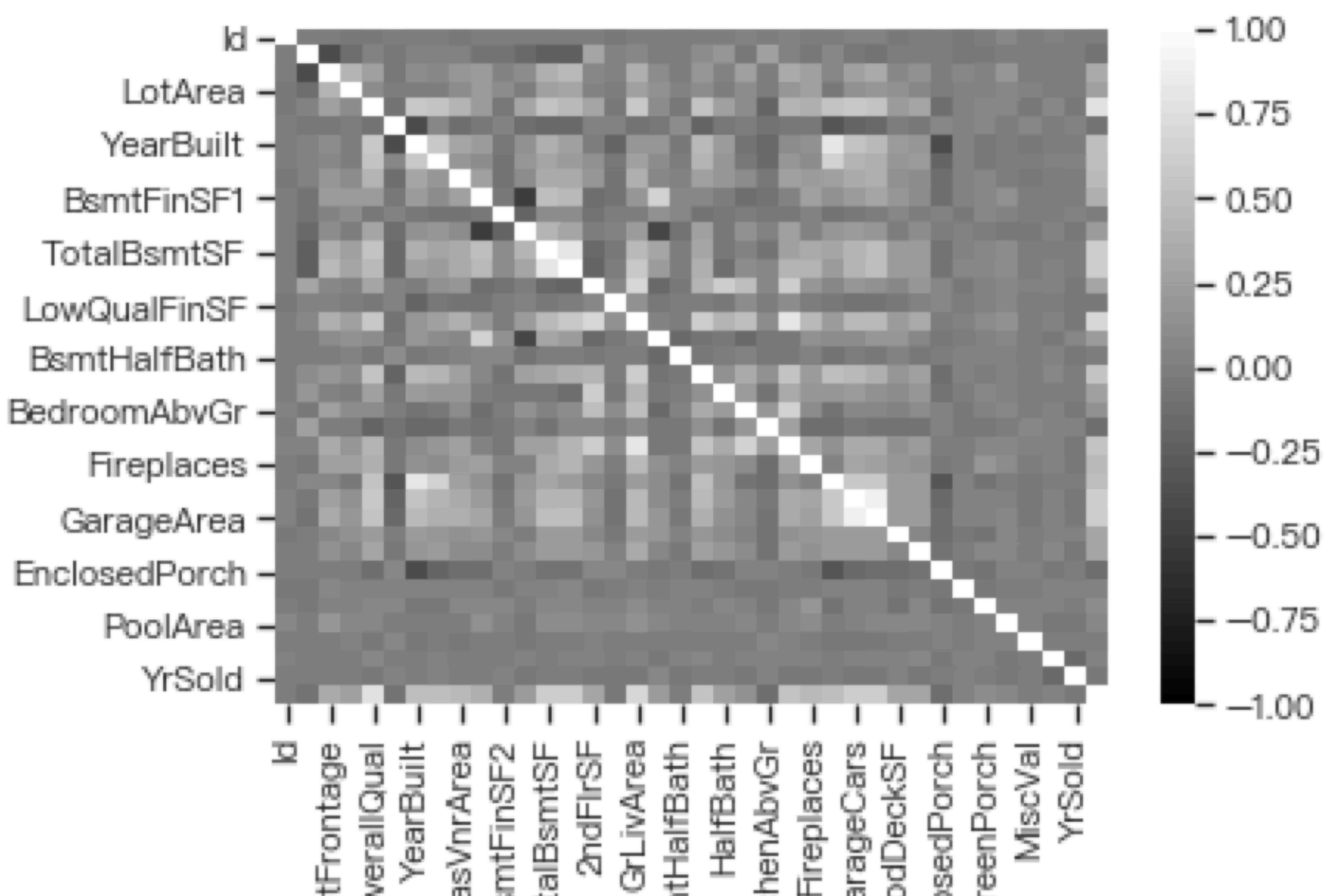
Correlations

```
corr = df_num.corr()
corr
```

	LotFrontage	MasVnrArea	GarageYrBlt
LotFrontage	1.000000	0.193458	0.070250
MasVnrArea	0.193458	1.000000	0.252691
GarageYrBlt	0.070250	0.252691	1.000000

Heat map

```
df_num = df.select_dtypes(include = ['float64', 'int64'])
corr = df_num.corr()
sns.heatmap(corr, cmap='gray', vmax=1.0, vmin=-1.0)
plt.show()
```



Readings

- Chapter 6 of Principles and Techniques of Data Science

- Chapter 9 of Python for Data Analysis, 2nd Edition

Y763

Homework

1. 利用2020/10/1–12的airbox資料計算台中市與台北市的逐時平均PM 2.5濃度(hint: 善用os.listdir)

2. 繪製下列圖表以比較台中市與台北市空氣污染程度：

   - scatter plot （台中市 v.s. 台北市）
   - box plot
   - 台中市與台北式的逐時折線圖（畫在同一張圖內）