数据挖掘exp01

实验一：Python预处理函数

实验目的

实验目的： 1、搭建好Windows版本的Python开发平台，并在该平台上安装一系列第三方的扩展库； 2、通过检验数据集的数据质量、绘制图表、计算某些特征等手段，对样本数据集的结构和规律进行分析，学会使用第三方拓展库对数据进行探索； 3、学会对原始数据中的缺失值、异常值进行处理； 2、针对原始数据集中的脏数据，能够通过Python及其拓展库对数据进行集成、规约、转换等操作	

实验内容

实验内容：

1、 从Python官网https://www.python.org/ 下载3.6版本的Python；

2、 下载Python环境管理软件Anaconda并熟悉它的使用；

3、 对abalone.data数据文件进行预处理，使得其转化为包含表头的csv文件（命名为abalone.csv），读取并打印前10行数据；

4、 查看是否有属性存在空值；

5、 绘制每一个数值型属性的直方图，查看其分布，输出符合正态分布的属性及其直方图，从中挑一个属性，利用3σ原则分析异常值；

6、 针对带壳肉高度，画出其箱型图，并运用箱型图分析法进行异常检测，将异常值输出；

7、 判断外壳长度与高度、整只鲍鱼重量是否相关，若相关，输出相关系数的值，并判断哪一个相关性更大；

8、 针对外壳长度和整只鲍鱼重量两个属性建立关系，使用拉格朗日插值方法针对外壳长度为0.137、0.172的重量进行插补，将插补结果进行输出，并画出程序流程图；

9、 使用Z-score对肉的重量属性进行规范化处理；

10、 附加题：尝试使用PCA对原始数据样本进行降维，α设置为0.6，并将降维结果进行输出；

11、 将上述实验内容的核心代码及实验结果截图放到“实验过程及分析”中。

代码实现

```
data_file_path = "abalone/abalone.data"
csv_file_path = "abalone/abalone.csv"

import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import skew


# 3、对abalone.data数据文件进行预处理，使得其转化为包含表头的csv文件（命名为
abalone.csv），读取并打印前10行数据；
columns = ["Sex", "Length", "Diameter", "Height", "Whole weight",
           "Shucked weight", "Viscera weight", "Shell weight", "Rings"]
data = pd.read_csv(data_file_path, header=None, names=columns)
```

```
data.to_csv(csv_file_path, index=False)
print(data.head(10))

# 4、查看是否有属性存在空值；
data = pd.read_csv(csv_file_path)
null_counts = data.isnull().sum()
print(null_counts)

# 5、绘制每一个数值型属性的直方图，查看其分布，输出符合正态分布的属性及其直方图，从中挑一个属性，利用3σ原则分析异常值；
data = pd.read_csv(csv_file_path)
numeric_columns = ["Length", "Diameter", "Height", "Whole weight",
                   "Shucked weight", "Viscera weight", "Shell weight",
"Rings"]
plt.style.use('seaborn-v0_8-darkgrid')
normal_attributes = []
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False  # 用于正常显示负号

# 绘制所有数值型属性的直方图
plt.figure(figsize=(16, 12))
for i, col in enumerate(numeric_columns):
    plt.subplot(3, 3, i+1)
    plt.hist(data[col], bins=30, color='skyblue', edgecolor='black')
    plt.title(col)
    # 计算偏度值，通常偏度绝对值较小（例如 < 0.5）可以认为近似正态分布
    skewness = skew(data[col])
    plt.xlabel(f'Skewness={skewness:.2f}')
    # 如果偏度绝对值小于0.5，则认为近似正态
    if abs(skewness) < 0.5:
        normal_attributes.append(col)

plt.tight_layout()
plt.show()

print("近似符合正态分布的属性：", normal_attributes)

attr = "Length"
attr_data = data[attr]

mean_val = np.mean(attr_data)
std_val = np.std(attr_data)

lower_bound = mean_val - 3 * std_val
upper_bound = mean_val + 3 * std_val

outliers = data[(attr_data < lower_bound) | (attr_data > upper_bound)]
print(f"属性 '{attr}' 的均值：{mean_val:.2f}，标准差：{std_val:.2f}")
print(f"根据3σ原则，阈值为 [{lower_bound:.2f}, {upper_bound:.2f}]")
print(f"'{attr}' 属性中异常值的数量：{len(outliers)}")
```

```
print("异常值数据预览：")
print(outliers[[attr]].head())

# 绘制选择属性的直方图，并标出3σ范围
plt.figure(figsize=(8, 5))
plt.hist(attr_data, bins=30, color='lightgreen', edgecolor='black')
plt.axvline(lower_bound, color='red', linestyle='dashed', linewidth=2,
label='下界')
plt.axvline(upper_bound, color='red', linestyle='dashed', linewidth=2,
label='上界')
plt.title(f"{attr} 的直方图及3σ异常值界限")
plt.xlabel(attr)
plt.ylabel("频数")
plt.legend()
plt.show()

# 6、针对带壳肉高度，画出其箱型图，并运用箱型图分析法进行异常检测，将异常值输出；
data = pd.read_csv(csv_file_path)
height_data = data["Height"]
# 计算第一四分位数、第三四分位数及四分位距
Q1 = height_data.quantile(0.25)
Q3 = height_data.quantile(0.75)
IQR = Q3 - Q1
# 根据箱型图异常值判断标准：小于 Q1-1.5*IQR 或大于 Q3+1.5*IQR 的数据认为是异常值
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
# 筛选出异常值
outliers = data[(height_data < lower_bound) | (height_data > upper_bound)]
print("带壳肉高度异常值：")
print(outliers["Height"])
# 绘制箱型图
plt.figure(figsize=(8, 6))
plt.boxplot(height_data, vert=False, patch_artist=True,
            boxprops=dict(facecolor="lightblue", color="blue"),
            medianprops=dict(color="red"))
plt.title("带壳肉高度箱型图")
plt.xlabel("Height")
plt.show()

# 7、判断外壳长度与高度、整只鲍鱼重量是否相关，若相关，输出相关系数的值，并判断哪一个相
关性更大；
data = pd.read_csv(csv_file_path)
# 计算外壳长度（Length）与高度（Height）的皮尔逊相关系数
corr_length_height = data['Length'].corr(data['Height'])
# 计算外壳长度（Length）与整只鲍鱼重量（Whole weight）的皮尔逊相关系数
corr_length_whole = data['Length'].corr(data['Whole weight'])
# 输出相关系数
print("外壳长度与高度的相关系数：", corr_length_height)
print("外壳长度与整只鲍鱼重量的相关系数：", corr_length_whole)
# 判断哪一个相关性更大（比较相关系数绝对值）
```

```
if abs(corr_length_height) > abs(corr_length_whole):
    print("外壳长度与高度的相关性更强。")
elif abs(corr_length_height) < abs(corr_length_whole):
    print("外壳长度与整只鲍鱼重量的相关性更强。")
else:
    print("外壳长度与高度和整只鲍鱼重量的相关性相近。")

# 针对外壳长度和整只鲍鱼重量两个属性建立关系，使用拉格朗日插值方法针对外壳长度为0.137、
0.172的重量进行插补，将插补结果进行输出；
from scipy.interpolate import lagrange

data = pd.read_csv(csv_file_path)
# 按外壳长度排序，并选取较小外壳长度的4个数据点用于局部插值
sorted_data = data.sort_values("Length")
subset = sorted_data.head(4)
# 提取外壳长度和整只鲍鱼重量的值
x = subset["Length"].values
y = subset["Whole weight"].values
# 显示选取的数据点
print("选取用于插值的数据点：")
for xi, yi in zip(x, y):
    print(f"Length = {xi:.3f}, Whole weight = {yi:.3f}")
# 利用拉格朗日插值方法构建插值多项式
poly = lagrange(x, y)
# 对指定的外壳长度进行插值，并匹配最近的数据点
lengths_to_interp = [0.137, 0.172]
interp_results = {}
for length_val in lengths_to_interp:
    # 查找与目标外壳长度最接近的实际数据点
    closest_idx = np.argmin(np.abs(x - length_val))
    closest_length = x[closest_idx]
    closest_weight = y[closest_idx]
    # 如果插值点在数据范围外，使用最接近的点的值
    if length_val < min(x) or length_val > max(x):
        interp_results[length_val] = closest_weight
    else:
        interp_weight = poly(length_val)
        interp_results[length_val] = interp_weight
    print(f"外壳长度为 {length_val} 时，插补的整只鲍鱼重量为：
{interp_results[length_val]:.3f}")

# 9、使用Z-score对肉的重量属性进行规范化处理；
data = pd.read_csv(csv_file_path)
shucked = data["Shucked weight"]
# 计算均值和标准差
mean_val = shucked.mean()
std_val = shucked.std()
# 使用Z-score公式进行规范化处理
data["Shucked weight_zscore"] = (shucked - mean_val) / std_val
```

```
# 输出前几行查看规范化后的结果
print(data[["Shucked weight", "Shucked weight_zscore"]].head())
```

运行结果

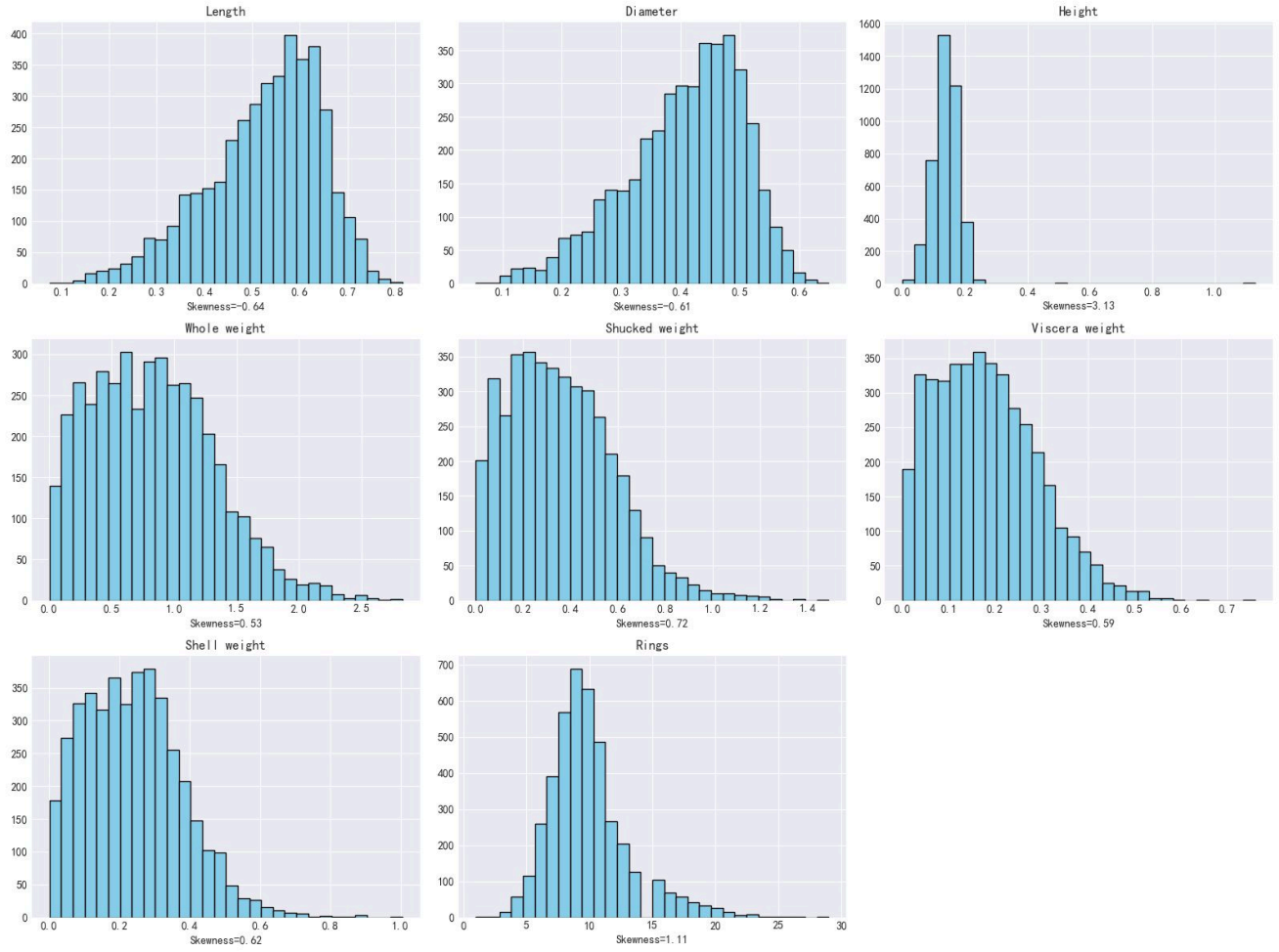
终端输出

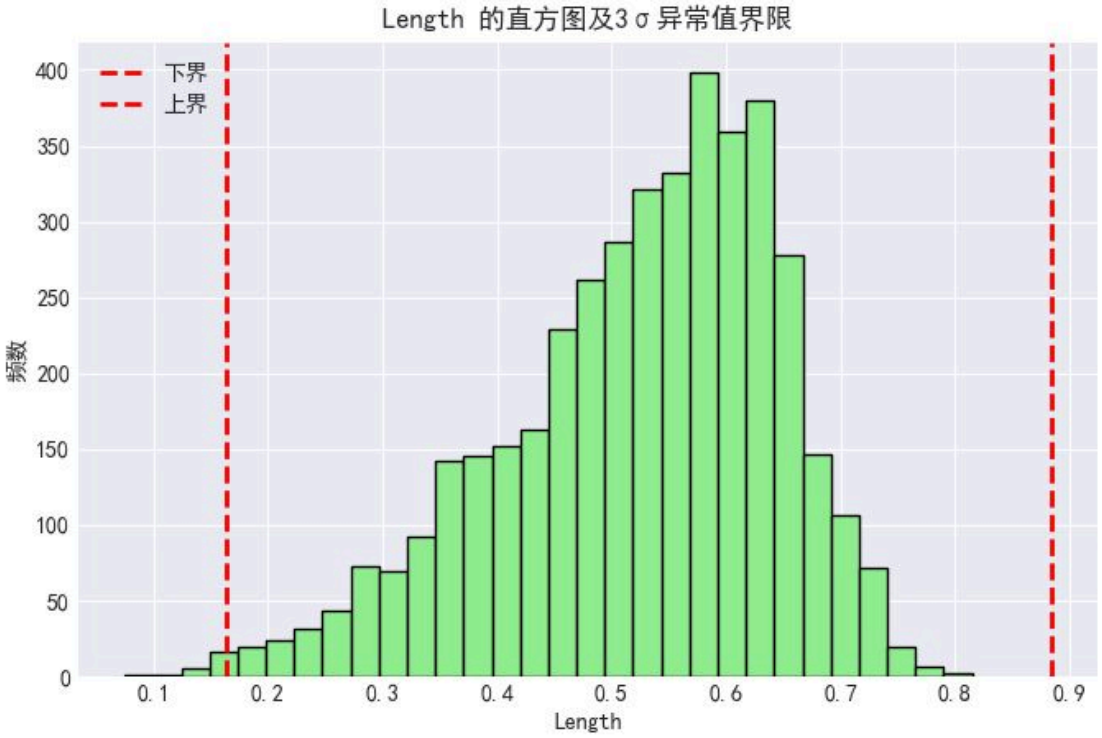
```
  Sex  Length  Diameter  ...  Viscera weight  Shell weight  Rings
0   M   0.455     0.365  ...          0.1010         0.150     15
1   M   0.350     0.265  ...          0.0485         0.070      7
2   F   0.530     0.420  ...          0.1415         0.210      9
3   M   0.440     0.365  ...          0.1140         0.155     10
4   I   0.330     0.255  ...          0.0395         0.055      7
5   I   0.425     0.300  ...          0.0775         0.120      8
6   F   0.530     0.415  ...          0.1415         0.330     20
7   F   0.545     0.425  ...          0.1495         0.260     16
8   M   0.475     0.370  ...          0.1125         0.165      9
9   F   0.550     0.440  ...          0.1510         0.320     19

[10 rows x 9 columns]
Sex               0
Length            0
Diameter          0
Height            0
Whole weight      0
Shucked weight    0
Viscera weight    0
Shell weight      0
Rings             0
dtype: int64
近似符合正态分布的属性： []
属性 'Length' 的均值：0.52，标准差：0.12
根据3σ原则，阈值为 [0.16, 0.88]
'Length' 属性中异常值的数量：15
异常值数据预览：
     Length
236   0.075
237   0.130
238   0.110
239   0.160
526   0.155
带壳肉高度异常值：
236     0.010
```

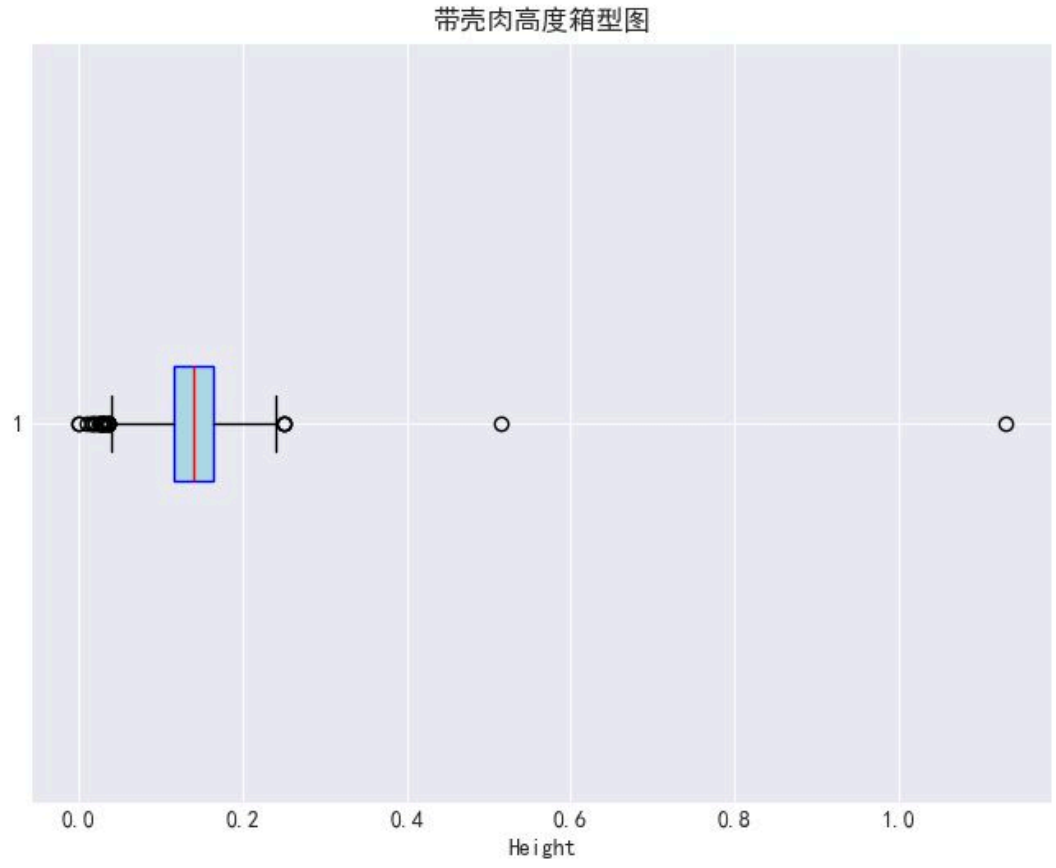
```
237     0.030
238     0.030
239     0.035
306     0.030
694     0.020
718     0.035
719     0.025
720     0.025
1174    0.015
1257    0.000
1417    0.515
1428    0.250
1429    0.035
1763    0.250
1987    0.025
2051    1.130
2114    0.035
2169    0.015
2171    0.030
2172    0.030
2179    0.250
2381    0.025
2711    0.030
3190    0.025
3837    0.035
3899    0.035
3902    0.020
3996    0.000
Name: Height, dtype: float64
外壳长度与高度的相关系数： 0.8275536093192105
外壳长度与整只鲍鱼重量的相关系数： 0.9252611721489453
外壳长度与整只鲍鱼重量的相关性更强。
选取用于插值的数据点：
Length = 0.075, Whole weight = 0.002
Length = 0.110, Whole weight = 0.008
Length = 0.130, Whole weight = 0.013
Length = 0.130, Whole weight = 0.011
外壳长度为 0.137 时，插补的整只鲍鱼重量为：0.013
外壳长度为 0.172 时，插补的整只鲍鱼重量为：0.013

  return poly1d(self.coeffs/other)
   Shucked weight  Shucked weight_zscore
0          0.2245              -0.607613
```

```
1           0.0995          -1.170770
2           0.2565          -0.463444
3           0.2155          -0.648160
4           0.0895          -1.215822
```

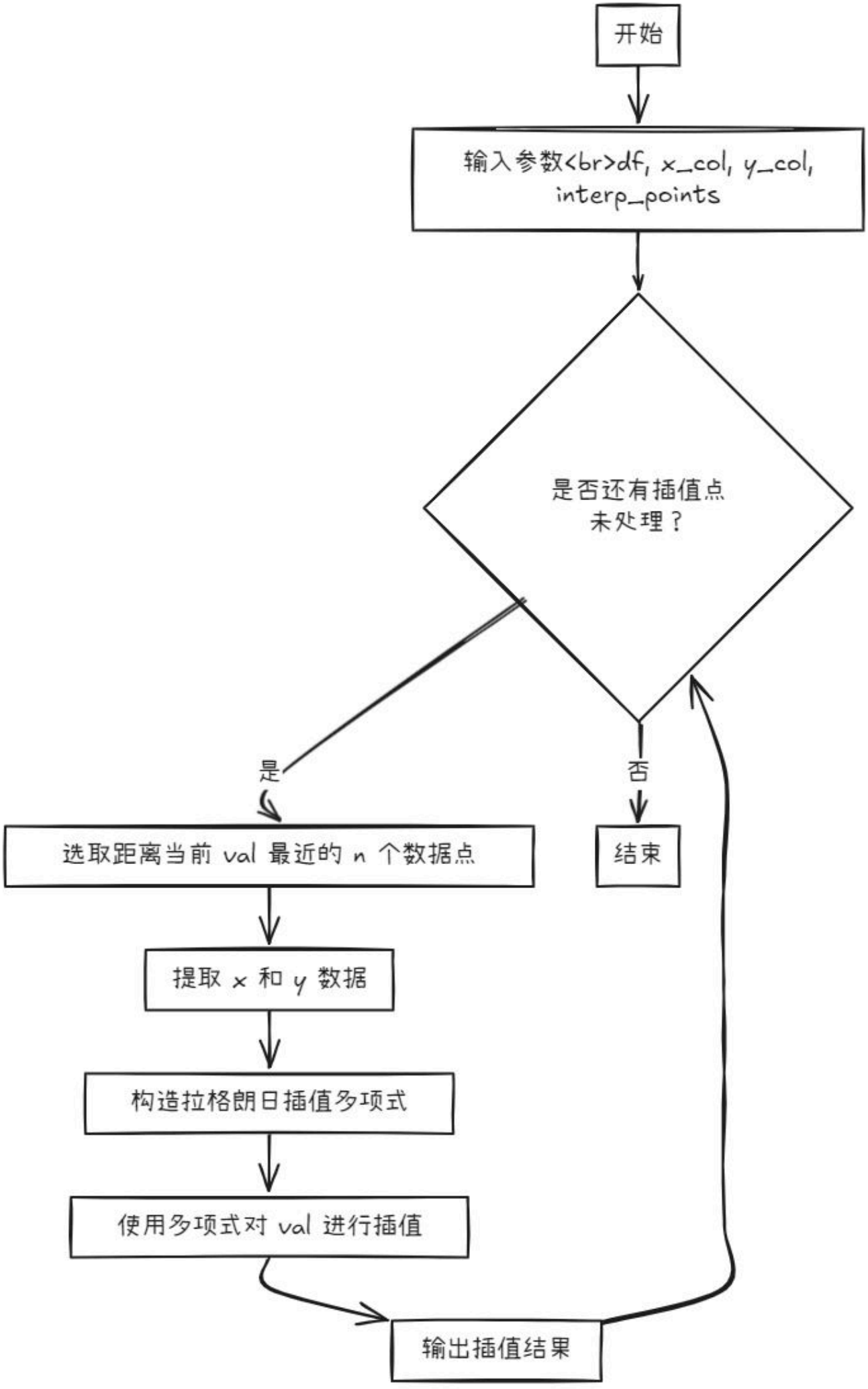
图表输出




Length 的直方图及3σ异常值界限
下界
上界
频数
Length
0
50
100
150
200
250
300
350
400
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9


带壳肉高度箱型图
1
0.0
0.2
0.4
0.6
0.8
1.0
Height

流程图绘制


开始
输入参数
df, x_col, y_col, interp_points
是否还有插值点未处理？
是
否
选取距离当前 val 最近的 n 个数据点
结束
提取 x 和 y 数据
构造拉格朗日插值多项式
使用多项式对 val 进行插值
输出插值结果