data_mining1

2022 年 3 月 18 日

1 引入必要的库

[1]:
```
import numpy as np
import pandas as pd
from pandas import DataFrame
from typing import *
from enum import Enum
import matplotlib.pyplot as plt
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import SimpleImputer, IterativeImputer, KNNImputer
```

2 属性类型，对应无关属性、标称属性、数值属性

[2]:
```
class AttributeType(Enum):
    Nonsense = 0
    Nominal = 1
    Numeric = 2
```

3 缺失值处理方法

[3]:
```
class MissingProcessing(object):
    # 将缺失部分剔除
    @staticmethod
    def eliminate(data_array: np.ndarray) -> list:
        return [data for data in data_array if not np.isnan(data)]
```

```
    # 用最高频率值来填补缺失值
    @staticmethod
    def frequencyFill(data_array: np.ndarray) -> np.ndarray:
        if len(data_array.shape) == 1:
            data_array = data_array.reshape(-1, 1)
        return SimpleImputer(strategy='most_frequent').
↪fit_transform(data_array).reshape(-1)

    # 通过属性的相关关系来填补缺失值，这里使用贝叶斯回归算法
    @staticmethod
    def relevanceFill(data_array: np.ndarray) -> np.ndarray:
        if len(data_array.shape) == 1:
            data_array = data_array.reshape(-1, 1)
        return IterativeImputer().fit_transform(data_array).reshape(-1)

    # 通过数据对象之间的相似性来填补缺失值，这里使用 knn 算法
    @staticmethod
    def similarityFill(data_array: np.ndarray) -> np.ndarray:
        if len(data_array.shape) == 1:
            data_array = data_array.astype(float).reshape(-1, 1)
        return KNNImputer().fit_transform(data_array).reshape(-1)

    def __init__(self,
                 method: Callable[[np.ndarray], Union[list, np.ndarray]]):
        self.method = method

    def __call__(self, data_array: np.ndarray) -> Union[list, np.ndarray]:
        return self.method(data_array)
```

4 数据摘要

[4]:
```
def dataSummary(data_frame: DataFrame,
                attribute_types: List[AttributeType],
                missing_method: Callable[[np.ndarray], Union[list, np.
↪ndarray]]) -> List:
```

```
    # 统计数据取值的频数
    def getFrequency(data_array: np.ndarray) -> dict:
        frequency_dict = {}
        for data in data_array:
            try:
                frequency_dict[data] += 1
            except KeyError:
                frequency_dict[data] = 1
        return frequency_dict

    # 获得数据的 5 数概括、nan 值个数以及处理 nan 值后的数据
    def statistics(data_array: np.ndarray) -> Tuple[float, float, float, float,␣
↪float, int, list]:
        nan_sum = sum(1 for data in data_array if np.isnan(data))
        if nan_sum > 0:
            data_array = missing_method(data_array)
        describe = pd.Series(list(data_array)).describe()
        return describe['min'], describe['25%'], describe['50%'],␣
↪describe['75%'], describe['max'],  nan_sum, data_array

    summary_results = []
    values = data_frame.values

    for i in range(len(data_frame.columns)):
        # 根据数据属性类别获得不同的摘要结果
        if attribute_types[i] == AttributeType.Nominal:
            summary_results.append(getFrequency(values[:, i]))
        elif attribute_types[i] == AttributeType.Numeric:
            summary_results.append(statistics(values[:, i]))
        else:
            summary_results.append(None)

    return summary_results
```

在数据摘要的输出结果中，**标称属性**以 **词典 (dict)** 的形式输出，**数值属性**以 **元组 (tuple)** 的形式输出。其中字典中以标称属性 **所有可能取值**为 **键 (key)** ，以 **频数**为 **值 (value)** ；元组中 **前 5 个**

数字为数值属性的 **5 数概括，第 6 个数字为 缺失值的个数。** 举例：> country { 'US' : 62397, 'Italy' : 23478, 'France' : 21098, 'Spain' : 8268, 'Chile' : 5816, 'Argentina' : 5631, 'Portugal' : 5322, 'Australia' : 4957, 'New Zealand' : 3320, 'Austria' : 3057}

表示标称属性 **country** 的可能取值 **'US'** 的频数为 **62397** ， **'Italy'** 的频数为 **23478** .....以此类推。为了便于查看，只显示前 **max_display_sum(processing** 函数的参数) 条统计结果。

price (4.0, 16.0, 24.0, 40.0, 2300.0, 13695)

表示数值属性 **price** 的 5 数概括为 **(4.0, 16.0, 24.0, 40.0, 2300.0)** ， 缺失值的个数为 **13695** 。

5 数据可视化

[5]:
```
def visualize(summary_results: List[Union[dict, Tuple, None]],
              attribute_types: List[AttributeType],
              attribute_names: List[str]) -> None:
    visualization_count = attribute_types.count(AttributeType.Numeric) * 3
    fig_x_sum = 3
    fig_y_sum = (visualization_count + fig_x_sum - 1) // fig_x_sum
    fig_index = 1
    plt.figure(figsize=(fig_x_sum * 3.5, fig_y_sum * 2))
    for summary_result, attribute_type, attribute_name in zip(summary_results,␣
 ↪attribute_types, attribute_names):
        # 直方图、盒图只对数值属性的数据有效
        if attribute_type == AttributeType.Numeric:
            plt.subplot(fig_y_sum, fig_x_sum, fig_index)
            plt.title(f'Hist of {attribute_name}')
            plt.hist(summary_result[-1])
            fig_index += 1
            plt.subplot(fig_y_sum, fig_x_sum, fig_index)
            plt.title(f'Boxplot of {attribute_name}\n(with outliers)')
            plt.boxplot(summary_result[-1], showfliers=True)
            fig_index += 1
            plt.subplot(fig_y_sum, fig_x_sum, fig_index)
            plt.title(f'Boxplot of {attribute_name}\n(without outliers)')
            plt.boxplot(summary_result[-1], showfliers=False)
            fig_index += 1
        else:
```

```
            continue
    plt.tight_layout()
    plt.show()
```

由于部分偏差过大的异常值的存在，显示盒图时盒体可能过小，因此将显示包含异常值 (**with outliers**) 和不包含异常值 (**without outliers**) 的两种盒图。

6 数据处理流程

[6]:
```
def processing(csv_file_path: str,
               attribute_types: List[AttributeType],
               missing_method: Callable[[np.ndarray], Union[list, np.ndarray]]␣
↪= MissingProcessing.eliminate,
               max_display_sum: int = 10) -> None:
    data_frame = pd.read_csv(csv_file_path)
    summary_results = dataSummary(data_frame, attribute_types, missing_method)
    for i, summary_result in enumerate(summary_results):
        if summary_result is not None:
            if isinstance(summary_result, dict):
                # 为了便于查看，只显示前 max_display_sum 条统计结果
                print(data_frame.columns[i],
                      dict(sorted(summary_result.items(), key=lambda x: x[1],␣
↪reverse=True)[:max_display_sum]))
            else:
                print(data_frame.columns[i], summary_result[:6])
    visualize(summary_results, attribute_types, data_frame.columns)
```

---

7 数据分析流程

7.1 1. 分析 Wine Reviews 数据集的 winemag-data_first150k.csv’ 文件

7.1.1 csv 文件路径

[7]:
```
csv_file_path = 'D:/Data/data_mining/1/Wine Reviews/winemag-data_first150k.csv'
```

7.1.2 标注出无关属性、数值属性的列序号，其余的为标称属性

[8]:
```
nonsense_columns = [0, 2]
Numeric_columns = [4, 5]
attributeType = [AttributeType.Nonsense if i in nonsense_columns
                 else AttributeType.Numeric if i in Numeric_columns
                 else AttributeType.Nominal
                 for i in range(11)]
```

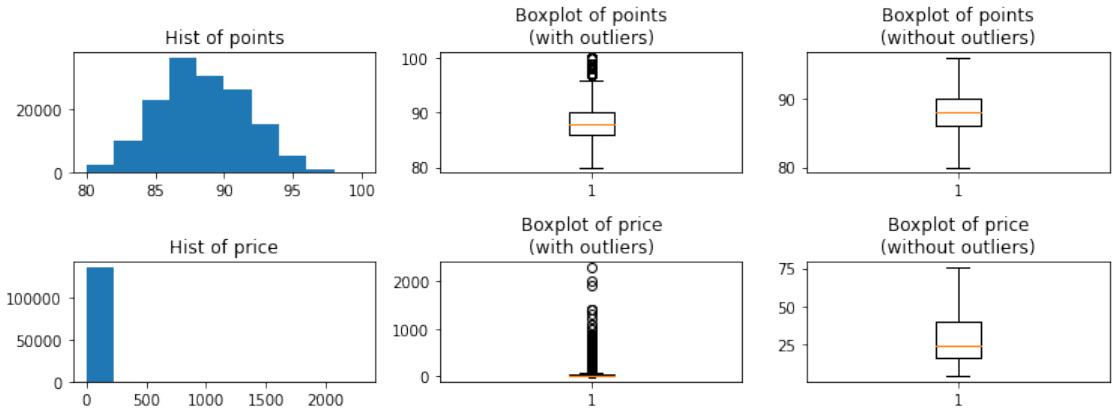
7.1.3 依次采用不同的缺失数据处理方法进行处理

- 将缺失部分剔除

[9]:
```
processing(csv_file_path, attributeType, MissingProcessing.eliminate)
```

```
country {'US': 62397, 'Italy': 23478, 'France': 21098, 'Spain': 8268, 'Chile':
5816, 'Argentina': 5631, 'Portugal': 5322, 'Australia': 4957, 'New Zealand':
3320, 'Austria': 3057}
designation {nan: 45735, 'Reserve': 2752, 'Reserva': 1810, 'Estate': 1571,
'Barrel sample': 1326, 'Riserva': 754, 'Barrel Sample': 639, 'Brut': 624,
'Crianza': 503, 'Estate Grown': 449}
points (80.0, 86.0, 88.0, 90.0, 100.0, 0)
price (4.0, 16.0, 24.0, 40.0, 2300.0, 13695)
province {'California': 44508, 'Washington': 9750, 'Tuscany': 7281, 'Bordeaux':
6111, 'Northern Spain': 4892, 'Mendoza Province': 4742, 'Oregon': 4589,
'Burgundy': 4308, 'Piedmont': 4093, 'Veneto': 3962}
region_1 {nan: 25060, 'Napa Valley': 6209, 'Columbia Valley (WA)': 4975,
'Mendoza': 3586, 'Russian River Valley': 3571, 'California': 3462, 'Paso
Robles': 3053, 'Willamette Valley': 2096, 'Rioja': 1893, 'Toscana': 1885}
region_2 {nan: 89977, 'Central Coast': 13057, 'Sonoma': 11258, 'Columbia
Valley': 9157, 'Napa': 8801, 'California Other': 3516, 'Willamette Valley':
```

3181, 'Mendocino/Lake Counties': 2389, 'Sierra Foothills': 1660, 'Napa-Sonoma': 1645}
variety {'Chardonnay': 14482, 'Pinot Noir': 14291, 'Cabernet Sauvignon': 12800, 'Red Blend': 10062, 'Bordeaux-style Red Blend': 7347, 'Sauvignon Blanc': 6320, 'Syrah': 5825, 'Riesling': 5524, 'Merlot': 5070, 'Zinfandel': 3799}
winery {'Williams Selyem': 374, 'Testarossa': 274, 'DFJ Vinhos': 258, 'Chateau Ste. Michelle': 225, 'Columbia Crest': 217, 'Concha y Toro': 216, 'Kendall-Jackson': 216, 'Trapiche': 205, 'Bouchard Père & Fils': 203, 'Kenwood': 191}

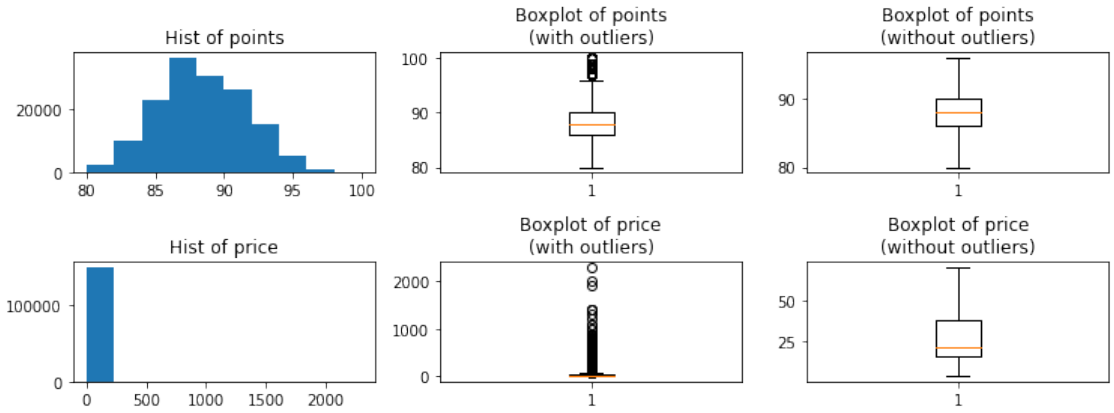


- 用最高频率值来填补缺失值

[10]:
```
processing(csv_file_path, attributeType, MissingProcessing.frequencyFill)
```

country {'US': 62397, 'Italy': 23478, 'France': 21098, 'Spain': 8268, 'Chile': 5816, 'Argentina': 5631, 'Portugal': 5322, 'Australia': 4957, 'New Zealand': 3320, 'Austria': 3057}
designation {nan: 45735, 'Reserve': 2752, 'Reserva': 1810, 'Estate': 1571, 'Barrel sample': 1326, 'Riserva': 754, 'Barrel Sample': 639, 'Brut': 624, 'Crianza': 503, 'Estate Grown': 449}
points (80.0, 86.0, 88.0, 90.0, 100.0, 0)
price (4.0, 16.0, 22.0, 38.0, 2300.0, 13695)
province {'California': 44508, 'Washington': 9750, 'Tuscany': 7281, 'Bordeaux': 6111, 'Northern Spain': 4892, 'Mendoza Province': 4742, 'Oregon': 4589, 'Burgundy': 4308, 'Piedmont': 4093, 'Veneto': 3962}
region_1 {nan: 25060, 'Napa Valley': 6209, 'Columbia Valley (WA)': 4975,

```
'Mendoza': 3586, 'Russian River Valley': 3571, 'California': 3462, 'Paso
Robles': 3053, 'Willamette Valley': 2096, 'Rioja': 1893, 'Toscana': 1885}
region_2 {nan: 89977, 'Central Coast': 13057, 'Sonoma': 11258, 'Columbia
Valley': 9157, 'Napa': 8801, 'California Other': 3516, 'Willamette Valley':
3181, 'Mendocino/Lake Counties': 2389, 'Sierra Foothills': 1660, 'Napa-Sonoma':
1645}
variety {'Chardonnay': 14482, 'Pinot Noir': 14291, 'Cabernet Sauvignon': 12800,
'Red Blend': 10062, 'Bordeaux-style Red Blend': 7347, 'Sauvignon Blanc': 6320,
'Syrah': 5825, 'Riesling': 5524, 'Merlot': 5070, 'Zinfandel': 3799}
winery {'Williams Selyem': 374, 'Testarossa': 274, 'DFJ Vinhos': 258, 'Chateau
Ste. Michelle': 225, 'Columbia Crest': 217, 'Concha y Toro': 216, 'Kendall-
Jackson': 216, 'Trapiche': 205, 'Bouchard Père & Fils': 203, 'Kenwood': 191}
```

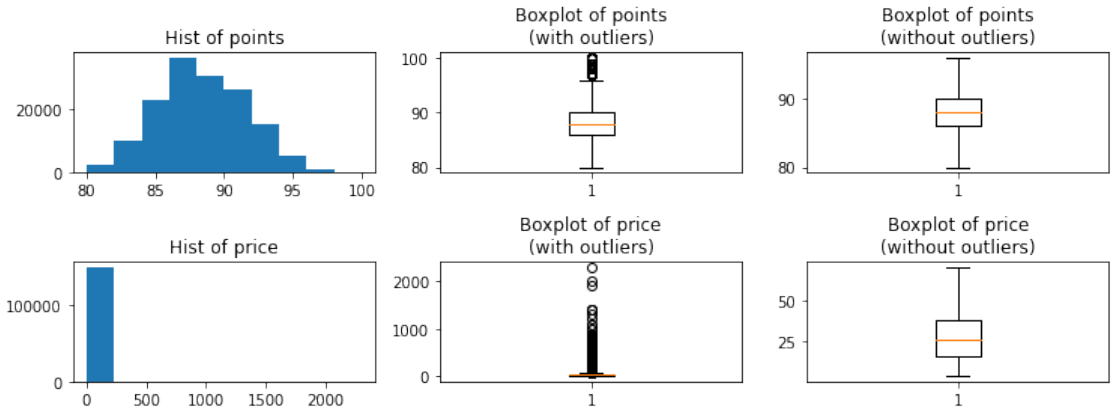


- 通过属性的相关关系来填补缺失值

[11]:
```
processing(csv_file_path, attributeType, MissingProcessing.relevanceFill)
```

```
country {'US': 62397, 'Italy': 23478, 'France': 21098, 'Spain': 8268, 'Chile':
5816, 'Argentina': 5631, 'Portugal': 5322, 'Australia': 4957, 'New Zealand':
3320, 'Austria': 3057}
designation {nan: 45735, 'Reserve': 2752, 'Reserva': 1810, 'Estate': 1571,
'Barrel sample': 1326, 'Riserva': 754, 'Barrel Sample': 639, 'Brut': 624,
'Crianza': 503, 'Estate Grown': 449}
points (80.0, 86.0, 88.0, 90.0, 100.0, 0)
price (4.0, 16.0, 26.0, 38.0, 2300.0, 13695)
```

```
province {'California': 44508, 'Washington': 9750, 'Tuscany': 7281, 'Bordeaux':
6111, 'Northern Spain': 4892, 'Mendoza Province': 4742, 'Oregon': 4589,
'Burgundy': 4308, 'Piedmont': 4093, 'Veneto': 3962}
region_1 {nan: 25060, 'Napa Valley': 6209, 'Columbia Valley (WA)': 4975,
'Mendoza': 3586, 'Russian River Valley': 3571, 'California': 3462, 'Paso
Robles': 3053, 'Willamette Valley': 2096, 'Rioja': 1893, 'Toscana': 1885}
region_2 {nan: 89977, 'Central Coast': 13057, 'Sonoma': 11258, 'Columbia
Valley': 9157, 'Napa': 8801, 'California Other': 3516, 'Willamette Valley':
3181, 'Mendocino/Lake Counties': 2389, 'Sierra Foothills': 1660, 'Napa-Sonoma':
1645}
variety {'Chardonnay': 14482, 'Pinot Noir': 14291, 'Cabernet Sauvignon': 12800,
'Red Blend': 10062, 'Bordeaux-style Red Blend': 7347, 'Sauvignon Blanc': 6320,
'Syrah': 5825, 'Riesling': 5524, 'Merlot': 5070, 'Zinfandel': 3799}
winery {'Williams Selyem': 374, 'Testarossa': 274, 'DFJ Vinhos': 258, 'Chateau
Ste. Michelle': 225, 'Columbia Crest': 217, 'Concha y Toro': 216, 'Kendall-
Jackson': 216, 'Trapiche': 205, 'Bouchard Père & Fils': 203, 'Kenwood': 191}
```



- 通过数据对象之间的相似性来填补缺失值

[12]:
```
processing(csv_file_path, attributeType, MissingProcessing.similarityFill)
```

```
country {'US': 62397, 'Italy': 23478, 'France': 21098, 'Spain': 8268, 'Chile':
5816, 'Argentina': 5631, 'Portugal': 5322, 'Australia': 4957, 'New Zealand':
3320, 'Austria': 3057}
designation {nan: 45735, 'Reserve': 2752, 'Reserva': 1810, 'Estate': 1571,
```

'Barrel sample': 1326, 'Riserva': 754, 'Barrel Sample': 639, 'Brut': 624, 'Crianza': 503, 'Estate Grown': 449}

points (80.0, 86.0, 88.0, 90.0, 100.0, 0)

price (4.0, 16.0, 26.0, 38.0, 2300.0, 13695)

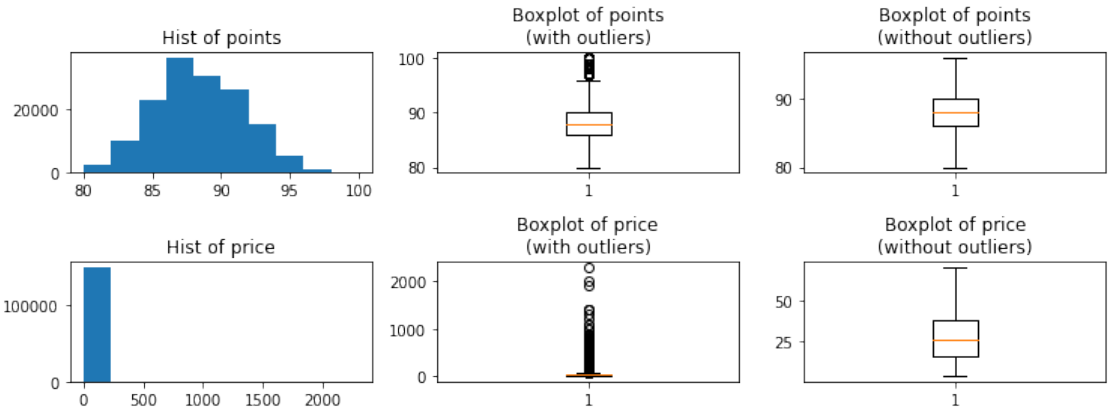
province {'California': 44508, 'Washington': 9750, 'Tuscany': 7281, 'Bordeaux': 6111, 'Northern Spain': 4892, 'Mendoza Province': 4742, 'Oregon': 4589, 'Burgundy': 4308, 'Piedmont': 4093, 'Veneto': 3962}

region_1 {nan: 25060, 'Napa Valley': 6209, 'Columbia Valley (WA)': 4975, 'Mendoza': 3586, 'Russian River Valley': 3571, 'California': 3462, 'Paso Robles': 3053, 'Willamette Valley': 2096, 'Rioja': 1893, 'Toscana': 1885}

region_2 {nan: 89977, 'Central Coast': 13057, 'Sonoma': 11258, 'Columbia Valley': 9157, 'Napa': 8801, 'California Other': 3516, 'Willamette Valley': 3181, 'Mendocino/Lake Counties': 2389, 'Sierra Foothills': 1660, 'Napa-Sonoma': 1645}

variety {'Chardonnay': 14482, 'Pinot Noir': 14291, 'Cabernet Sauvignon': 12800, 'Red Blend': 10062, 'Bordeaux-style Red Blend': 7347, 'Sauvignon Blanc': 6320, 'Syrah': 5825, 'Riesling': 5524, 'Merlot': 5070, 'Zinfandel': 3799}

winery {'Williams Selyem': 374, 'Testarossa': 274, 'DFJ Vinhos': 258, 'Chateau Ste. Michelle': 225, 'Columbia Crest': 217, 'Concha y Toro': 216, 'Kendall-Jackson': 216, 'Trapiche': 205, 'Bouchard Père & Fils': 203, 'Kenwood': 191}



7.2 2. 分析 Wine Reviews 数据集的 winemag-data-130k-v2.csv’ 文件

7.2.1 csv 文件路径

[13]:
```
csv_file_path = 'D:/Data/data_mining/1/Wine Reviews/winemag-data-130k-v2.csv'
```

7.2.2 标注出无关属性、数值属性的列序号，其余的为标称属性

[14]:
```
nonsense_columns = [0, 2, 11]
Numeric_columns = [4, 5]
attributeType = [AttributeType.Nonsense if i in nonsense_columns
                 else AttributeType.Numeric if i in Numeric_columns
                 else AttributeType.Nominal
                 for i in range(14)]
```

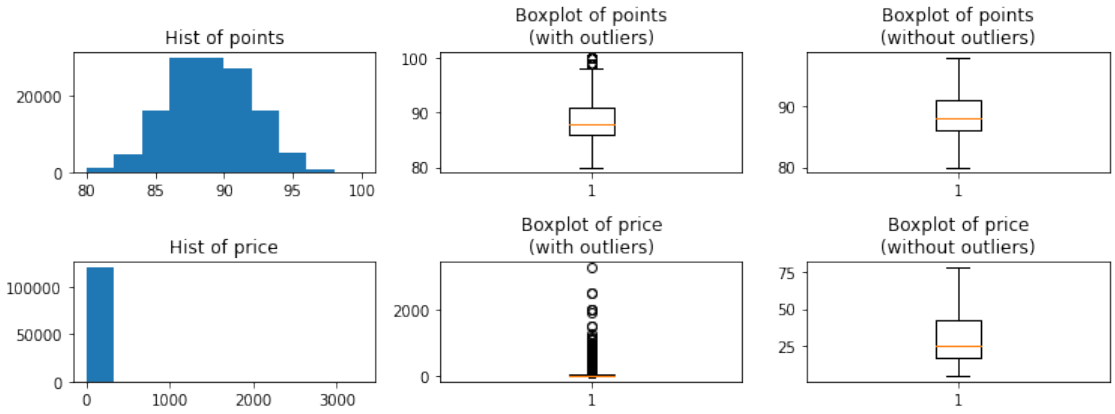
7.2.3 依次采用不同的缺失数据处理方法进行处理

- 将缺失部分剔除

[15]:
```
processing(csv_file_path, attributeType, MissingProcessing.eliminate)
```

```
country {'US': 54504, 'France': 22093, 'Italy': 19540, 'Spain': 6645,
'Portugal': 5691, 'Chile': 4472, 'Argentina': 3800, 'Austria': 3345,
'Australia': 2329, 'Germany': 2165}
designation {nan: 37465, 'Reserve': 2009, 'Estate': 1322, 'Reserva': 1259,
'Riserva': 698, 'Estate Grown': 621, 'Brut': 513, 'Dry': 413, 'Barrel sample':
375, 'Crianza': 343}
points (80.0, 86.0, 88.0, 91.0, 100.0, 0)
price (4.0, 17.0, 25.0, 42.0, 3300.0, 8996)
province {'California': 36247, 'Washington': 8639, 'Bordeaux': 5941, 'Tuscany':
5897, 'Oregon': 5373, 'Burgundy': 3980, 'Northern Spain': 3851, 'Piedmont':
3729, 'Mendoza Province': 3264, 'Veneto': 2716}
region_1 {nan: 21247, 'Napa Valley': 4480, 'Columbia Valley (WA)': 4124,
'Russian River Valley': 3091, 'California': 2629, 'Paso Robles': 2350,
'Willamette Valley': 2301, 'Mendoza': 2301, 'Alsace': 2163, 'Champagne': 1613}
region_2 {nan: 79460, 'Central Coast': 11065, 'Sonoma': 9028, 'Columbia Valley':
8103, 'Napa': 6814, 'Willamette Valley': 3423, 'California Other': 2663, 'Finger
Lakes': 1777, 'Sierra Foothills': 1462, 'Napa-Sonoma': 1169}
taster_name {nan: 26244, 'Roger Voss': 25514, 'Michael Schachner': 15134, 'Kerin
```

```
O’Keefe': 10776, 'Virginie Boone': 9537, 'Paul Gregutt': 9532, 'Matt Kettmann':
6332, 'Joe Czerwinski': 5147, 'Sean P. Sullivan': 4966, 'Anna Lee C. Iijima':
4415}
taster_twitter_handle {nan: 31213, '@vossroger': 25514, '@wineschach': 15134,
'@kerinokeefe': 10776, '@vboone': 9537, '@paulgwine\xa0': 9532, '@mattkettmann':
6332, '@JoeCz': 5147, '@wawinereport': 4966, '@gordone_cellars': 4177}
variety {'Pinot Noir': 13272, 'Chardonnay': 11753, 'Cabernet Sauvignon': 9472,
'Red Blend': 8946, 'Bordeaux-style Red Blend': 6915, 'Riesling': 5189,
'Sauvignon Blanc': 4967, 'Syrah': 4142, 'Rosé': 3564, 'Merlot': 3102}
winery {'Wines & Winemakers': 222, 'Testarossa': 218, 'DFJ Vinhos': 215,
'Williams Selyem': 211, 'Louis Latour': 199, 'Georges Duboeuf': 196, 'Chateau
Ste. Michelle': 194, 'Concha y Toro': 164, 'Columbia Crest': 159, 'Kendall-
Jackson': 130}
```

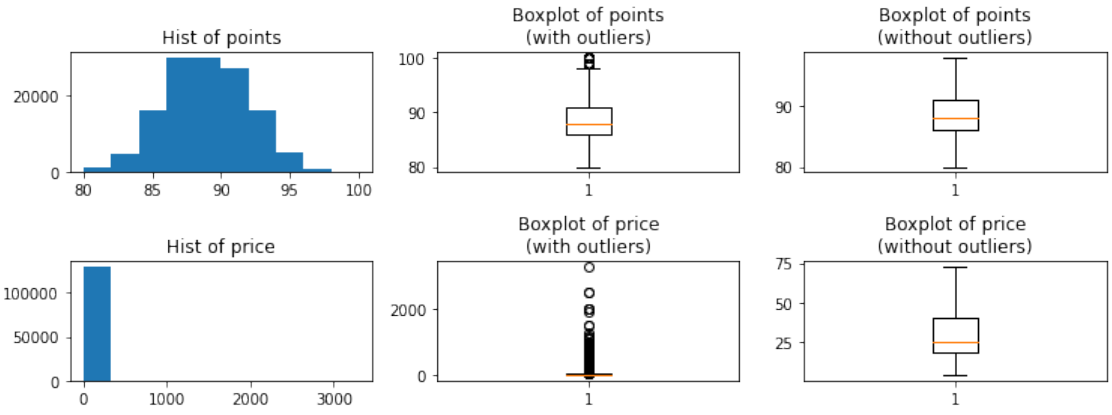


- 用最高频率值来填补缺失值

[16]:
```
processing(csv_file_path, attributeType, MissingProcessing.frequencyFill)
```

```
country {'US': 54504, 'France': 22093, 'Italy': 19540, 'Spain': 6645,
'Portugal': 5691, 'Chile': 4472, 'Argentina': 3800, 'Austria': 3345,
'Australia': 2329, 'Germany': 2165}
designation {nan: 37465, 'Reserve': 2009, 'Estate': 1322, 'Reserva': 1259,
'Riserva': 698, 'Estate Grown': 621, 'Brut': 513, 'Dry': 413, 'Barrel sample':
375, 'Crianza': 343}
points (80.0, 86.0, 88.0, 91.0, 100.0, 0)
```

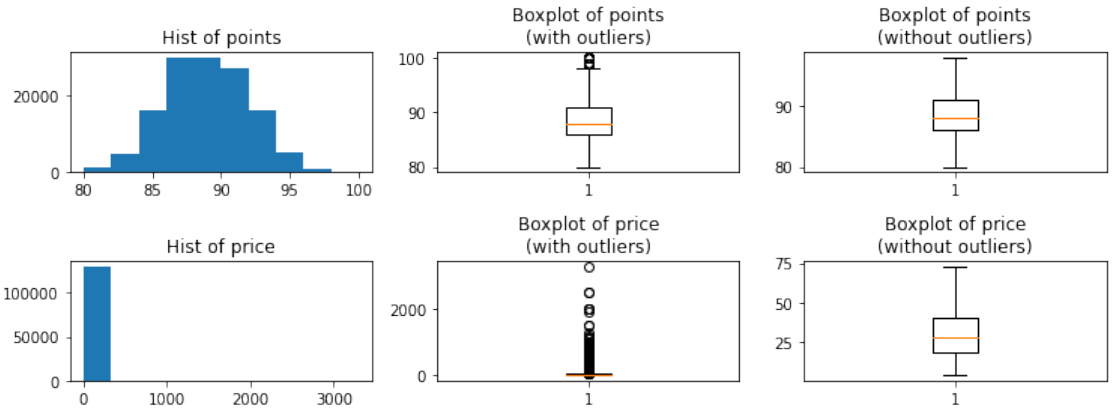
price (4.0, 18.0, 25.0, 40.0, 3300.0, 8996)
province {'California': 36247, 'Washington': 8639, 'Bordeaux': 5941, 'Tuscany': 5897, 'Oregon': 5373, 'Burgundy': 3980, 'Northern Spain': 3851, 'Piedmont': 3729, 'Mendoza Province': 3264, 'Veneto': 2716}
region_1 {nan: 21247, 'Napa Valley': 4480, 'Columbia Valley (WA)': 4124, 'Russian River Valley': 3091, 'California': 2629, 'Paso Robles': 2350, 'Willamette Valley': 2301, 'Mendoza': 2301, 'Alsace': 2163, 'Champagne': 1613}
region_2 {nan: 79460, 'Central Coast': 11065, 'Sonoma': 9028, 'Columbia Valley': 8103, 'Napa': 6814, 'Willamette Valley': 3423, 'California Other': 2663, 'Finger Lakes': 1777, 'Sierra Foothills': 1462, 'Napa-Sonoma': 1169}
taster_name {nan: 26244, 'Roger Voss': 25514, 'Michael Schachner': 15134, 'Kerin O’ Keefe': 10776, 'Virginie Boone': 9537, 'Paul Gregutt': 9532, 'Matt Kettmann': 6332, 'Joe Czerwinski': 5147, 'Sean P. Sullivan': 4966, 'Anna Lee C. Iijima': 4415}
taster_twitter_handle {nan: 31213, '@vossroger': 25514, '@wineschach': 15134, '@kerinokeefe': 10776, '@vboone': 9537, '@paulgwine\xa0': 9532, '@mattkettmann': 6332, '@JoeCz': 5147, '@wawinereport': 4966, '@gordone_cellars': 4177}
variety {'Pinot Noir': 13272, 'Chardonnay': 11753, 'Cabernet Sauvignon': 9472, 'Red Blend': 8946, 'Bordeaux-style Red Blend': 6915, 'Riesling': 5189, 'Sauvignon Blanc': 4967, 'Syrah': 4142, 'Rosé': 3564, 'Merlot': 3102}
winery {'Wines & Winemakers': 222, 'Testarossa': 218, 'DFJ Vinhos': 215, 'Williams Selyem': 211, 'Louis Latour': 199, 'Georges Duboeuf': 196, 'Chateau Ste. Michelle': 194, 'Concha y Toro': 164, 'Columbia Crest': 159, 'Kendall-Jackson': 130}



- 通过属性的相关关系来填补缺失值

[17]:
```
processing(csv_file_path, attributeType, MissingProcessing.relevanceFill)
```

```
country {'US': 54504, 'France': 22093, 'Italy': 19540, 'Spain': 6645,
'Portugal': 5691, 'Chile': 4472, 'Argentina': 3800, 'Austria': 3345,
'Australia': 2329, 'Germany': 2165}
designation {nan: 37465, 'Reserve': 2009, 'Estate': 1322, 'Reserva': 1259,
'Riserva': 698, 'Estate Grown': 621, 'Brut': 513, 'Dry': 413, 'Barrel sample':
375, 'Crianza': 343}
points (80.0, 86.0, 88.0, 91.0, 100.0, 0)
price (4.0, 18.0, 28.0, 40.0, 3300.0, 8996)
province {'California': 36247, 'Washington': 8639, 'Bordeaux': 5941, 'Tuscany':
5897, 'Oregon': 5373, 'Burgundy': 3980, 'Northern Spain': 3851, 'Piedmont':
3729, 'Mendoza Province': 3264, 'Veneto': 2716}
region_1 {nan: 21247, 'Napa Valley': 4480, 'Columbia Valley (WA)': 4124,
'Russian River Valley': 3091, 'California': 2629, 'Paso Robles': 2350,
'Willamette Valley': 2301, 'Mendoza': 2301, 'Alsace': 2163, 'Champagne': 1613}
region_2 {nan: 79460, 'Central Coast': 11065, 'Sonoma': 9028, 'Columbia Valley':
8103, 'Napa': 6814, 'Willamette Valley': 3423, 'California Other': 2663, 'Finger
Lakes': 1777, 'Sierra Foothills': 1462, 'Napa-Sonoma': 1169}
taster_name {nan: 26244, 'Roger Voss': 25514, 'Michael Schachner': 15134, 'Kerin
O'Keefe': 10776, 'Virginie Boone': 9537, 'Paul Gregutt': 9532, 'Matt Kettmann':
6332, 'Joe Czerwinski': 5147, 'Sean P. Sullivan': 4966, 'Anna Lee C. Iijima':
4415}
taster_twitter_handle {nan: 31213, '@vossroger': 25514, '@wineschach': 15134,
'@kerinokeefe': 10776, '@vboone': 9537, '@paulgwine\xa0': 9532, '@mattkettmann':
6332, '@JoeCz': 5147, '@wawinereport': 4966, '@gordone_cellars': 4177}
variety {'Pinot Noir': 13272, 'Chardonnay': 11753, 'Cabernet Sauvignon': 9472,
'Red Blend': 8946, 'Bordeaux-style Red Blend': 6915, 'Riesling': 5189,
'Sauvignon Blanc': 4967, 'Syrah': 4142, 'Rosé': 3564, 'Merlot': 3102}
winery {'Wines & Winemakers': 222, 'Testarossa': 218, 'DFJ Vinhos': 215,
'Williams Selyem': 211, 'Louis Latour': 199, 'Georges Duboeuf': 196, 'Chateau
Ste. Michelle': 194, 'Concha y Toro': 164, 'Columbia Crest': 159, 'Kendall-
Jackson': 130}
```

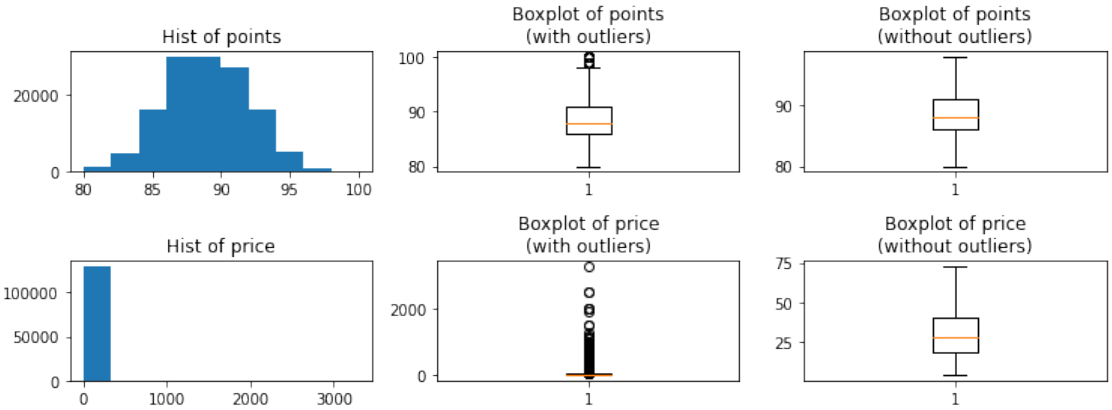


- 通过数据对象之间的相似性来填补缺失值

[18]:
```
processing(csv_file_path, attributeType, MissingProcessing.similarityFill)
```

```
country {'US': 54504, 'France': 22093, 'Italy': 19540, 'Spain': 6645,
'Portugal': 5691, 'Chile': 4472, 'Argentina': 3800, 'Austria': 3345,
'Australia': 2329, 'Germany': 2165}
designation {nan: 37465, 'Reserve': 2009, 'Estate': 1322, 'Reserva': 1259,
'Riserva': 698, 'Estate Grown': 621, 'Brut': 513, 'Dry': 413, 'Barrel sample':
375, 'Crianza': 343}
points (80.0, 86.0, 88.0, 91.0, 100.0, 0)
price (4.0, 18.0, 28.0, 40.0, 3300.0, 8996)
province {'California': 36247, 'Washington': 8639, 'Bordeaux': 5941, 'Tuscany':
5897, 'Oregon': 5373, 'Burgundy': 3980, 'Northern Spain': 3851, 'Piedmont':
3729, 'Mendoza Province': 3264, 'Veneto': 2716}
region_1 {nan: 21247, 'Napa Valley': 4480, 'Columbia Valley (WA)': 4124,
'Russian River Valley': 3091, 'California': 2629, 'Paso Robles': 2350,
'Willamette Valley': 2301, 'Mendoza': 2301, 'Alsace': 2163, 'Champagne': 1613}
region_2 {nan: 79460, 'Central Coast': 11065, 'Sonoma': 9028, 'Columbia Valley':
8103, 'Napa': 6814, 'Willamette Valley': 3423, 'California Other': 2663, 'Finger
Lakes': 1777, 'Sierra Foothills': 1462, 'Napa-Sonoma': 1169}
taster_name {nan: 26244, 'Roger Voss': 25514, 'Michael Schachner': 15134, 'Kerin
O' Keefe': 10776, 'Virginie Boone': 9537, 'Paul Gregutt': 9532, 'Matt Kettmann':
6332, 'Joe Czerwinski': 5147, 'Sean P. Sullivan': 4966, 'Anna Lee C. Iijima':
4415}
```

```
taster_twitter_handle {nan: 31213, '@vossroger': 25514, '@wineschach': 15134,
'@kerinokeefe': 10776, '@vboone': 9537, '@paulgwine\xa0': 9532, '@mattkettmann':
6332, '@JoeCz': 5147, '@wawinereport': 4966, '@gordone_cellars': 4177}
variety {'Pinot Noir': 13272, 'Chardonnay': 11753, 'Cabernet Sauvignon': 9472,
'Red Blend': 8946, 'Bordeaux-style Red Blend': 6915, 'Riesling': 5189,
'Sauvignon Blanc': 4967, 'Syrah': 4142, 'Rosé': 3564, 'Merlot': 3102}
winery {'Wines & Winemakers': 222, 'Testarossa': 218, 'DFJ Vinhos': 215,
'Williams Selyem': 211, 'Louis Latour': 199, 'Georges Duboeuf': 196, 'Chateau
Ste. Michelle': 194, 'Concha y Toro': 164, 'Columbia Crest': 159, 'Kendall-
Jackson': 130}
```



7.3 3. 分析 Chicago Building Violations 数据集的 building-violations.csv' 文件

7.3.1 csv 文件路径

[19]:
```
csv_file_path = 'D:/Data/data_mining/1/Chicago Building Violations/
↪building-violations.csv'
```

7.3.2 标注出无关属性、数值属性的列序号，其余的为标称属性

[20]:
```
nonsense_columns = [0, 1, 2, 5, 8, 25]
Numeric_columns = [22, 23, 24, 26, 29, 30, 31]
attributeType = [AttributeType.Nonsense if i in nonsense_columns
                 else AttributeType.Numeric if i in Numeric_columns
                 else AttributeType.Nominal
```

```
                for i in range(32)]
```

7.3.3 依次采用不同的缺失数据处理方法进行处理

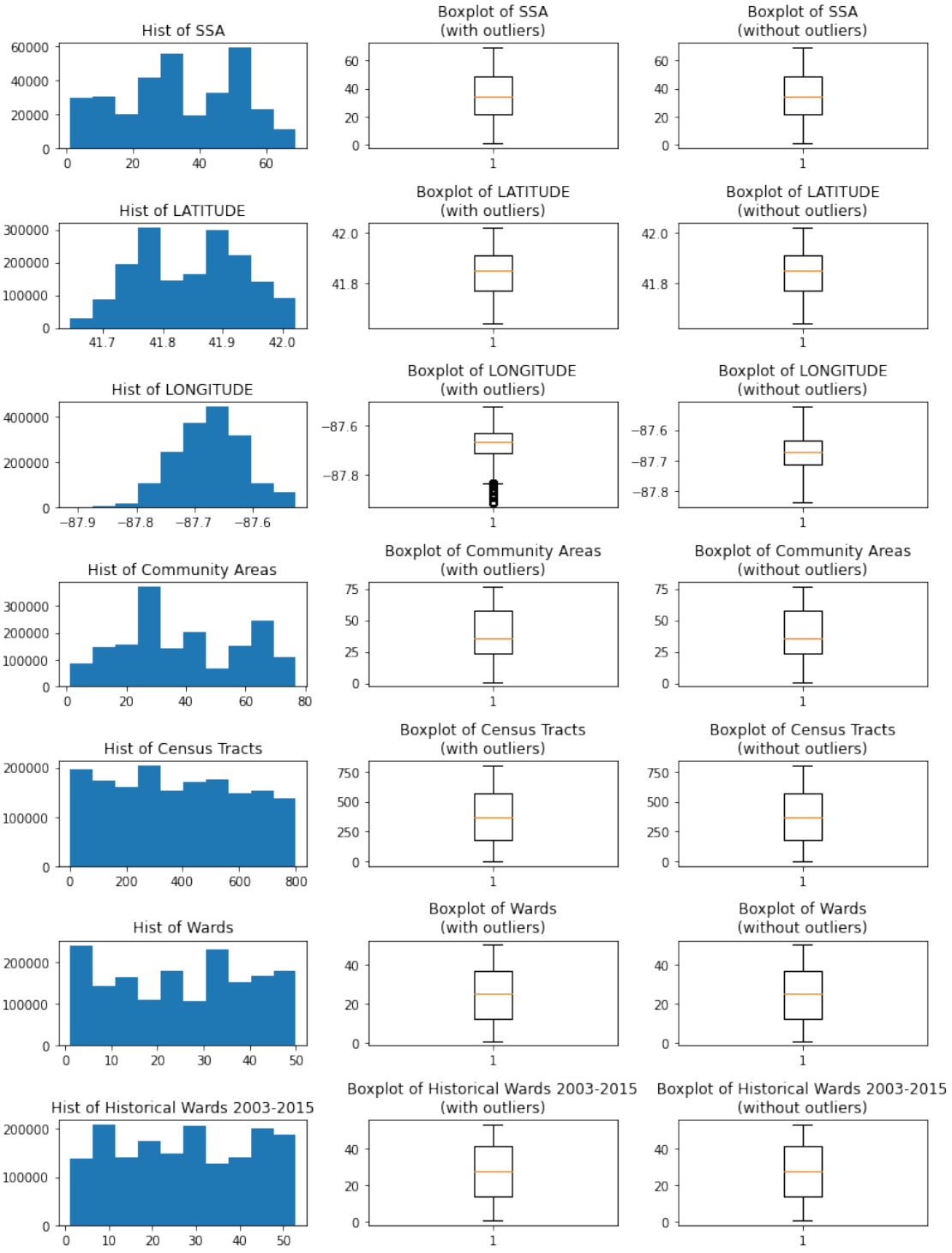
- 将缺失部分剔除

[21]:
```
processing(csv_file_path, attributeType, MissingProcessing.eliminate)
```

```
VIOLATION CODE {'CN190019': 89995, 'CN196029': 58136, 'CN061014': 51946,
'EV1110': 43700, 'CN070024': 43673, 'CN193110': 37127, 'CN104015': 34641,
'CN070014': 32093, 'CN197019': 30793, 'NC2011': 28750}
VIOLATION STATUS {'OPEN': 1030958, 'COMPLIED': 641247, 'NO ENTRY': 5583}
VIOLATION DESCRIPTION {'ARRANGE PREMISE INSPECTION': 90004, 'POST OWNER/MANAGERS
NAME/#': 58136, 'REPAIR EXTERIOR WALL': 51946, 'MAINTAIN OR REPAIR ELECT ELEVA':
43700, 'REPAIR PORCH SYSTEM': 43673, 'VACANT BUILDING - REGISTER': 37127,
'REPLCE WINDOW PANES, PLEXGLAS': 34641, 'REPAIR EXTERIOR STAIR': 32093, 'INSTALL
SMOKE DETECTORS': 30846, 'PLANS & PERMITS REQ - CONTRCTR': 28750}
VIOLATION LOCATION {nan: 897282, 'OTHER   :    :': 284277, 'OTHER   :
:OTHER': 35182, 'OTHER   :    :BUILDING': 21934, 'EXTERIOR:E   :': 21522,
'EXTERIOR:W   :': 19460, 'EXTERIOR:S   :': 17053, 'OTHER   :    :BL00006':
16951, 'EXTERIOR:N   :': 16004, 'INTERIOR:001 :': 13016}
VIOLATION ORDINANCE {'Arrange for inspection of premises. (13-12-100)': 89995,
"Post name, address, and telephone of owner, owner's agent for managing,
controlling or collecting rents, and any other person managing or controlling
building conspicuously where accessible or visible to public way. (13-12-030)":
58136, 'Failed to maintain the exterior walls of a building or structure free
from holes, breaks, loose or rotting boards or timbers and any other conditions
which might admit rain or dampness to the walls.  (13-196-530(b), 13-196-641)':
51946, nan: 47581, 'Failed to maintain electric elevator equipment provided at
premises in safe and sound working condition. (13-196-590, 13-196-630(b),
18-30-001)': 43700, 'Failed to repair or replace defective or missing members of
porch system.  (13-196-570, 13-196-641)': 43673, 'Register vacant building
within 30 days of it becoming vacant, or within 30 days after assuming ownership
of an existing vacant building. (13-12-125(a)). Building must be kept in
compliance with all vacant building requirements pursuant to 13-12-135. See
Vacant Building Ordinance and registration form at
https://ipiweb.cityofchicago.org/VBR': 37127, 'Replace broken, missing or
```

defective window panes. (13-196-550 A)': 34641, 'Failed to maintain exterior stairways in safe condition and in sound repair. (13-196-570, 13-196-641)': 32093, 'Install and maintain approved smoke detectors. (13-196-100 thru 13-196-160) Install a smoke detector in every dwelling unit. Install one on any living level with a habitable room or unenclosed heating plant, on the uppermost ceiling of enclosed porch stairwell, and within 15 feet of every sleeping room. Be sure the detector is at least 4 inches from the wall, 4 to 12 inches from the ceiling, and not above door or window.': 30846}
INSPECTOR ID {'BL00444': 79336, 'BL01000': 58497, 'BL00831': 52450, 'BL00746': 48797, 'BL00941': 48717, 'BL00812': 44556, 'BL00722': 43139, 'BL01037': 39108, 'BL01041': 38331, 'BL01039': 34976}
INSPECTION NUMBER {11247474: 535, 11587123: 209, 12964992: 135, 10669347: 123, 2323972: 123, 12955264: 120, 11356407: 106, 12253306: 104, 11995426: 92, 11041943: 82}
INSPECTION STATUS {'FAILED': 1159758, 'PASSED': 293076, 'CLOSED': 224784, 'HOLD': 154, nan: 16}
INSPECTION WAIVED {'N': 1677788}
INSPECTION CATEGORY {'COMPLAINT': 1186426, 'PERIODIC': 415176, 'PERMIT': 73600, 'REGISTRATION': 2586}
DEPARTMENT BUREAU {'CONSERVATION': 1110911, 'DEMOLITION': 125464, 'SPECIAL TASK FORCE': 115885, 'ELEVATOR': 85805, 'ELECTRICAL': 37243, 'VENTILATION': 32108, 'BOILER': 31235, 'NEW CONSTRUCTION': 29938, 'REFRIGERATION': 29681, 'PLUMBING': 28199}
ADDRESS {'1900 N AUSTIN AVE': 639, '11601 W TOUHY AVE': 605, '2545 W FITCH AVE': 316, '7200 S COLES AVE': 294, '727 E 60TH ST': 274, '1124 W WILSON AVE': 258, '4408 S DREXEL BLVD': 256, '8228 S SOUTH SHORE DR': 253, '55 E WASHINGTON ST': 253, '5501 W WASHINGTON BLVD': 250}
STREET NUMBER {200: 1704, 400: 1669, 500: 1542, 1000: 1501, 300: 1483, 600: 1462, 1900: 1442, 901: 1431, 1: 1419, 100: 1389}
STREET DIRECTION {'S': 683917, 'W': 500418, 'N': 395246, 'E': 98207}
STREET NAME {'MICHIGAN': 17980, 'ASHLAND': 16992, 'WESTERN': 13327, 'KEDZIE': 13034, 'HALSTED': 11268, 'WABASH': 10941, 'CALIFORNIA': 10747, 'CLARK': 10671, 'DR MARTIN L KING JR': 10614, 'INDIANA': 10612}
STREET TYPE {'AVE': 940725, 'ST': 523743, 'BLVD': 59536, 'PL': 57665, 'RD': 41100, 'DR': 27145, nan: 13541, 'PKWY': 6605, 'CT': 3287, 'TER': 2222}
PROPERTY GROUP {6124: 639, 21995: 605, 654907: 359, 13851: 320, 20890: 316, 25993: 313, 20386: 294, 1655: 275, 19382: 274, 2991: 269}

SSA (1.0, 22.0, 34.0, 49.0, 69.0, 1356267)
LATITUDE (41.644670132, 41.770896504250004, 41.854002336, 41.913504192, 42.02268599, 1510)
LONGITUDE (-87.914435848, -87.713917698, -87.6698535045, -87.632882744, -87.524679151, 1510)
Community Areas (1.0, 24.0, 36.0, 58.0, 77.0, 2279)
Zip Codes {21569.0: 79463, 21861.0: 69403, 21546.0: 67402, 21202.0: 63458, 22257.0: 61235, 21554.0: 60879, 21559.0: 60507, 14924.0: 59678, 4299.0: 58145, 21867.0: 57676}
Boundaries - ZIP Codes {57.0: 80508, 19.0: 68579, 61.0: 67501, 25.0: 63330, 59.0: 62103, 11.0: 61950, 58.0: 60268, 5.0: 59985, 32.0: 58309, 37.0: 57971}
Census Tracts (1.0, 179.0, 374.0, 572.0, 801.0, 1545)
Wards (1.0, 12.0, 25.0, 37.0, 50.0, 2279)
Historical Wards 2003-2015 (1.0, 14.0, 28.0, 41.0, 53.0, 2279)



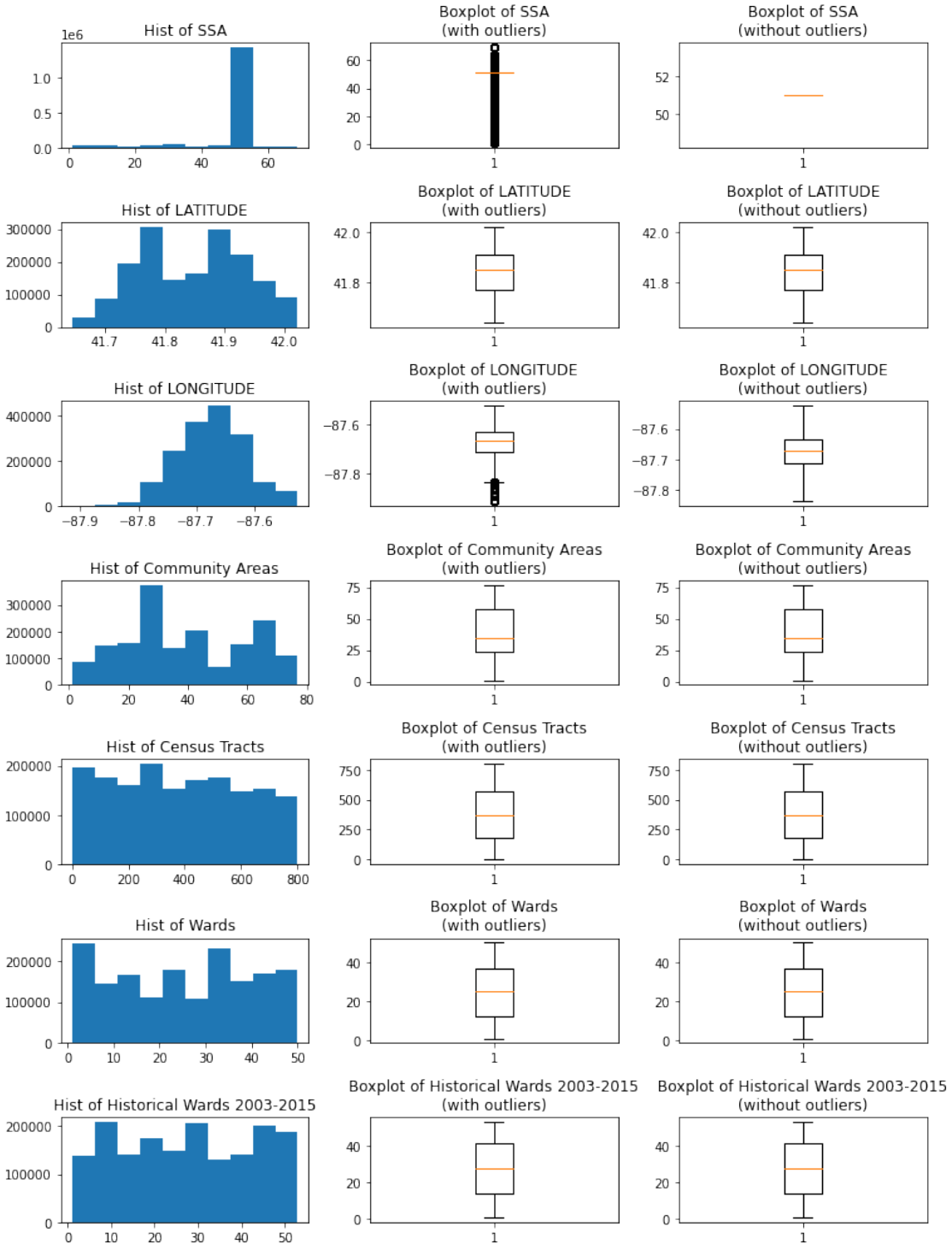
- 用最高频率值来填补缺失值

[22]:
```
processing(csv_file_path, attributeType, MissingProcessing.frequencyFill)
```

```
VIOLATION CODE {'CN190019': 89995, 'CN196029': 58136, 'CN061014': 51946,
'EV1110': 43700, 'CN070024': 43673, 'CN193110': 37127, 'CN104015': 34641,
'CN070014': 32093, 'CN197019': 30793, 'NC2011': 28750}
VIOLATION STATUS {'OPEN': 1030958, 'COMPLIED': 641247, 'NO ENTRY': 5583}
VIOLATION DESCRIPTION {'ARRANGE PREMISE INSPECTION': 90004, 'POST OWNER/MANAGERS
NAME/#': 58136, 'REPAIR EXTERIOR WALL': 51946, 'MAINTAIN OR REPAIR ELECT ELEVA':
43700, 'REPAIR PORCH SYSTEM': 43673, 'VACANT BUILDING - REGISTER': 37127,
'REPLCE WINDOW PANES, PLEXGLAS': 34641, 'REPAIR EXTERIOR STAIR': 32093, 'INSTALL
SMOKE DETECTORS': 30846, 'PLANS & PERMITS REQ - CONTRCTR': 28750}
VIOLATION LOCATION {nan: 897282, 'OTHER   :    :': 284277, 'OTHER   :
:OTHER': 35182, 'OTHER   :    :BUILDING': 21934, 'EXTERIOR:E   :': 21522,
'EXTERIOR:W   :': 19460, 'EXTERIOR:S   :': 17053, 'OTHER   :    :BL00006':
16951, 'EXTERIOR:N   :': 16004, 'INTERIOR:001 :': 13016}
VIOLATION ORDINANCE {'Arrange for inspection of premises. (13-12-100)': 89995,
"Post name, address, and telephone of owner, owner's agent for managing,
controlling or collecting rents, and any other person managing or controlling
building conspicuously where accessible or visible to public way. (13-12-030)":
58136, 'Failed to maintain the exterior walls of a building or structure free
from holes, breaks, loose or rotting boards or timbers and any other conditions
which might admit rain or dampness to the walls.  (13-196-530(b), 13-196-641)':
51946, nan: 47581, 'Failed to maintain electric elevator equipment provided at
premises in safe and sound working condition. (13-196-590, 13-196-630(b),
18-30-001)': 43700, 'Failed to repair or replace defective or missing members of
porch system.  (13-196-570, 13-196-641)': 43673, 'Register vacant building
within 30 days of it becoming vacant, or within 30 days after assuming ownership
of an existing vacant building. (13-12-125(a)). Building must be kept in
compliance with all vacant building requirements pursuant to 13-12-135. See
Vacant Building Ordinance and registration form at
https://ipiweb.cityofchicago.org/VBR': 37127, 'Replace broken, missing or
defective window panes. (13-196-550 A)': 34641, 'Failed to maintain exterior
stairways in safe condition and in sound repair.  (13-196-570, 13-196-641)':
32093, 'Install and maintain approved smoke detectors. (13-196-100 thru
13-196-160) Install a smoke detector in every dwelling unit. Install one on any
living level with a habitable room or unenclosed heating plant, on the uppermost
ceiling of enclosed porch stairwell, and within 15 feet of every sleeping room.
```

Be sure the detector is at least 4 inches from the wall, 4 to 12 inches from the ceiling, and not above door or window.': 30846}
INSPECTOR ID {'BL00444': 79336, 'BL01000': 58497, 'BL00831': 52450, 'BL00746': 48797, 'BL00941': 48717, 'BL00812': 44556, 'BL00722': 43139, 'BL01037': 39108, 'BL01041': 38331, 'BL01039': 34976}
INSPECTION NUMBER {11247474: 535, 11587123: 209, 12964992: 135, 10669347: 123, 2323972: 123, 12955264: 120, 11356407: 106, 12253306: 104, 11995426: 92, 11041943: 82}
INSPECTION STATUS {'FAILED': 1159758, 'PASSED': 293076, 'CLOSED': 224784, 'HOLD': 154, nan: 16}
INSPECTION WAIVED {'N': 1677788}
INSPECTION CATEGORY {'COMPLAINT': 1186426, 'PERIODIC': 415176, 'PERMIT': 73600, 'REGISTRATION': 2586}
DEPARTMENT BUREAU {'CONSERVATION': 1110911, 'DEMOLITION': 125464, 'SPECIAL TASK FORCE': 115885, 'ELEVATOR': 85805, 'ELECTRICAL': 37243, 'VENTILATION': 32108, 'BOILER': 31235, 'NEW CONSTRUCTION': 29938, 'REFRIGERATION': 29681, 'PLUMBING': 28199}
ADDRESS {'1900 N AUSTIN AVE': 639, '11601 W TOUHY AVE': 605, '2545 W FITCH AVE': 316, '7200 S COLES AVE': 294, '727 E 60TH ST': 274, '1124 W WILSON AVE': 258, '4408 S DREXEL BLVD': 256, '8228 S SOUTH SHORE DR': 253, '55 E WASHINGTON ST': 253, '5501 W WASHINGTON BLVD': 250}
STREET NUMBER {200: 1704, 400: 1669, 500: 1542, 1000: 1501, 300: 1483, 600: 1462, 1900: 1442, 901: 1431, 1: 1419, 100: 1389}
STREET DIRECTION {'S': 683917, 'W': 500418, 'N': 395246, 'E': 98207}
STREET NAME {'MICHIGAN': 17980, 'ASHLAND': 16992, 'WESTERN': 13327, 'KEDZIE': 13034, 'HALSTED': 11268, 'WABASH': 10941, 'CALIFORNIA': 10747, 'CLARK': 10671, 'DR MARTIN L KING JR': 10614, 'INDIANA': 10612}
STREET TYPE {'AVE': 940725, 'ST': 523743, 'BLVD': 59536, 'PL': 57665, 'RD': 41100, 'DR': 27145, nan: 13541, 'PKWY': 6605, 'CT': 3287, 'TER': 2222}
PROPERTY GROUP {6124: 639, 21995: 605, 654907: 359, 13851: 320, 20890: 316, 25993: 313, 20386: 294, 1655: 275, 19382: 274, 2991: 269}
SSA (1.0, 51.0, 51.0, 51.0, 69.0, 1356267)
LATITUDE (41.644670132, 41.770945238, 41.854079166, 41.913845946, 42.02268599, 1510)
LONGITUDE (-87.914435848, -87.714004538, -87.669900427, -87.632909287, -87.524679151, 1510)
Community Areas (1.0, 24.0, 35.0, 58.0, 77.0, 2279)

Zip Codes {21569.0: 79463, 21861.0: 69403, 21546.0: 67402, 21202.0: 63458, 22257.0: 61235, 21554.0: 60879, 21559.0: 60507, 14924.0: 59678, 4299.0: 58145, 21867.0: 57676}
Boundaries - ZIP Codes {57.0: 80508, 19.0: 68579, 61.0: 67501, 25.0: 63330, 59.0: 62103, 11.0: 61950, 58.0: 60268, 5.0: 59985, 32.0: 58309, 37.0: 57971}
Census Tracts (1.0, 178.0, 374.0, 572.0, 801.0, 1545)
Wards (1.0, 12.0, 25.0, 37.0, 50.0, 2279)
Historical Wards 2003-2015 (1.0, 14.0, 28.0, 41.0, 53.0, 2279)



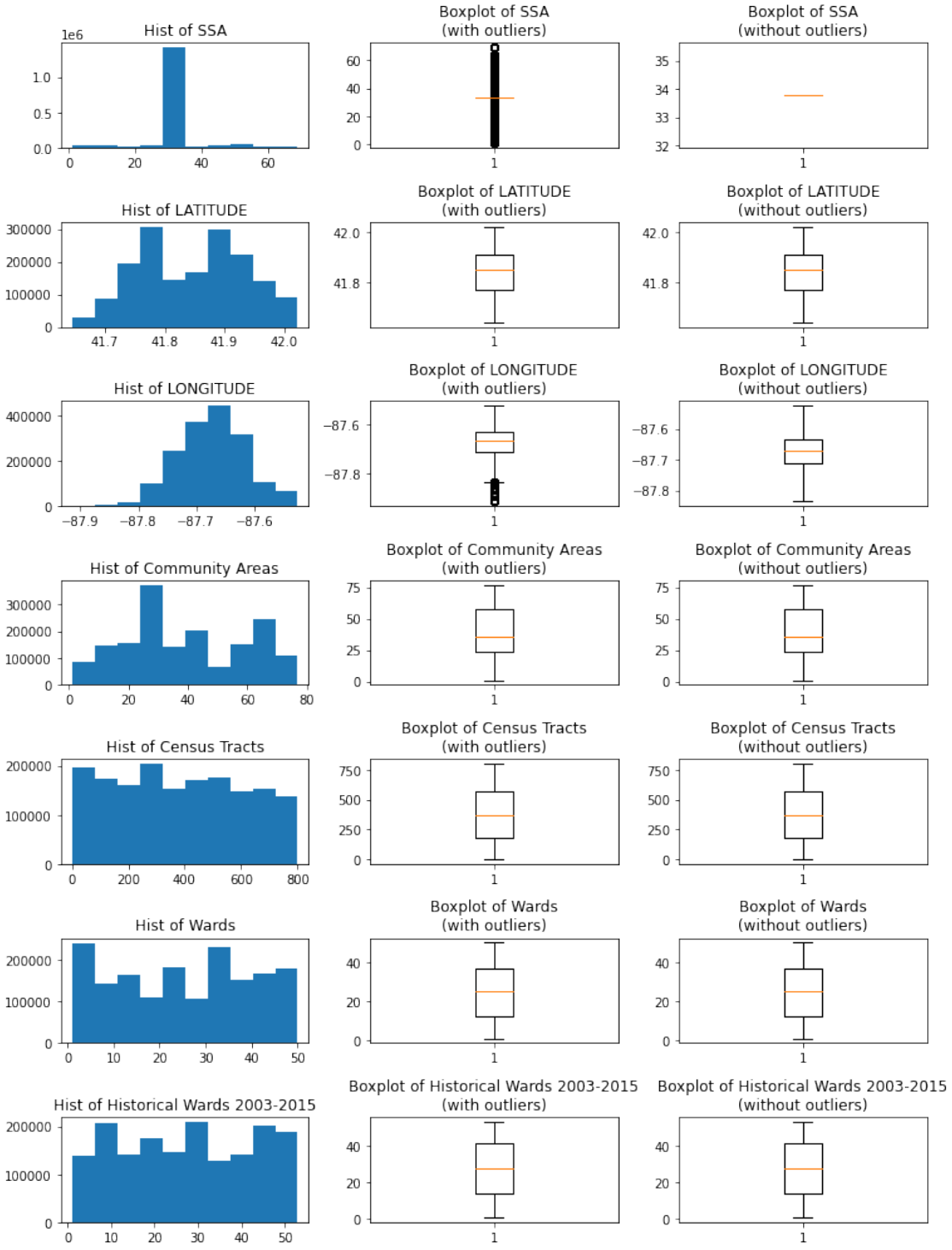
- 通过属性的相关关系来填补缺失值

[23]:
```
processing(csv_file_path, attributeType, MissingProcessing.relevanceFill)
```

```
VIOLATION CODE {'CN190019': 89995, 'CN196029': 58136, 'CN061014': 51946,
'EV1110': 43700, 'CN070024': 43673, 'CN193110': 37127, 'CN104015': 34641,
'CN070014': 32093, 'CN197019': 30793, 'NC2011': 28750}
VIOLATION STATUS {'OPEN': 1030958, 'COMPLIED': 641247, 'NO ENTRY': 5583}
VIOLATION DESCRIPTION {'ARRANGE PREMISE INSPECTION': 90004, 'POST OWNER/MANAGERS
NAME/#': 58136, 'REPAIR EXTERIOR WALL': 51946, 'MAINTAIN OR REPAIR ELECT ELEVA':
43700, 'REPAIR PORCH SYSTEM': 43673, 'VACANT BUILDING - REGISTER': 37127,
'REPLCE WINDOW PANES, PLEXGLAS': 34641, 'REPAIR EXTERIOR STAIR': 32093, 'INSTALL
SMOKE DETECTORS': 30846, 'PLANS & PERMITS REQ - CONTRCTR': 28750}
VIOLATION LOCATION {nan: 897282, 'OTHER   :    :': 284277, 'OTHER   :
:OTHER': 35182, 'OTHER   :    :BUILDING': 21934, 'EXTERIOR:E   :': 21522,
'EXTERIOR:W   :': 19460, 'EXTERIOR:S   :': 17053, 'OTHER   :    :BL00006':
16951, 'EXTERIOR:N   :': 16004, 'INTERIOR:001 :': 13016}
VIOLATION ORDINANCE {'Arrange for inspection of premises. (13-12-100)': 89995,
"Post name, address, and telephone of owner, owner's agent for managing,
controlling or collecting rents, and any other person managing or controlling
building conspicuously where accessible or visible to public way. (13-12-030)":
58136, 'Failed to maintain the exterior walls of a building or structure free
from holes, breaks, loose or rotting boards or timbers and any other conditions
which might admit rain or dampness to the walls.  (13-196-530(b), 13-196-641)':
51946, nan: 47581, 'Failed to maintain electric elevator equipment provided at
premises in safe and sound working condition. (13-196-590, 13-196-630(b),
18-30-001)': 43700, 'Failed to repair or replace defective or missing members of
porch system.  (13-196-570, 13-196-641)': 43673, 'Register vacant building
within 30 days of it becoming vacant, or within 30 days after assuming ownership
of an existing vacant building. (13-12-125(a)). Building must be kept in
compliance with all vacant building requirements pursuant to 13-12-135. See
Vacant Building Ordinance and registration form at
https://ipiweb.cityofchicago.org/VBR': 37127, 'Replace broken, missing or
defective window panes. (13-196-550 A)': 34641, 'Failed to maintain exterior
stairways in safe condition and in sound repair.  (13-196-570, 13-196-641)':
32093, 'Install and maintain approved smoke detectors. (13-196-100 thru
13-196-160) Install a smoke detector in every dwelling unit. Install one on any
living level with a habitable room or unenclosed heating plant, on the uppermost
ceiling of enclosed porch stairwell, and within 15 feet of every sleeping room.
```

Be sure the detector is at least 4 inches from the wall, 4 to 12 inches from the ceiling, and not above door or window.': 30846}
INSPECTOR ID {'BL00444': 79336, 'BL01000': 58497, 'BL00831': 52450, 'BL00746': 48797, 'BL00941': 48717, 'BL00812': 44556, 'BL00722': 43139, 'BL01037': 39108, 'BL01041': 38331, 'BL01039': 34976}
INSPECTION NUMBER {11247474: 535, 11587123: 209, 12964992: 135, 10669347: 123, 2323972: 123, 12955264: 120, 11356407: 106, 12253306: 104, 11995426: 92, 11041943: 82}
INSPECTION STATUS {'FAILED': 1159758, 'PASSED': 293076, 'CLOSED': 224784, 'HOLD': 154, nan: 16}
INSPECTION WAIVED {'N': 1677788}
INSPECTION CATEGORY {'COMPLAINT': 1186426, 'PERIODIC': 415176, 'PERMIT': 73600, 'REGISTRATION': 2586}
DEPARTMENT BUREAU {'CONSERVATION': 1110911, 'DEMOLITION': 125464, 'SPECIAL TASK FORCE': 115885, 'ELEVATOR': 85805, 'ELECTRICAL': 37243, 'VENTILATION': 32108, 'BOILER': 31235, 'NEW CONSTRUCTION': 29938, 'REFRIGERATION': 29681, 'PLUMBING': 28199}
ADDRESS {'1900 N AUSTIN AVE': 639, '11601 W TOUHY AVE': 605, '2545 W FITCH AVE': 316, '7200 S COLES AVE': 294, '727 E 60TH ST': 274, '1124 W WILSON AVE': 258, '4408 S DREXEL BLVD': 256, '8228 S SOUTH SHORE DR': 253, '55 E WASHINGTON ST': 253, '5501 W WASHINGTON BLVD': 250}
STREET NUMBER {200: 1704, 400: 1669, 500: 1542, 1000: 1501, 300: 1483, 600: 1462, 1900: 1442, 901: 1431, 1: 1419, 100: 1389}
STREET DIRECTION {'S': 683917, 'W': 500418, 'N': 395246, 'E': 98207}
STREET NAME {'MICHIGAN': 17980, 'ASHLAND': 16992, 'WESTERN': 13327, 'KEDZIE': 13034, 'HALSTED': 11268, 'WABASH': 10941, 'CALIFORNIA': 10747, 'CLARK': 10671, 'DR MARTIN L KING JR': 10614, 'INDIANA': 10612}
STREET TYPE {'AVE': 940725, 'ST': 523743, 'BLVD': 59536, 'PL': 57665, 'RD': 41100, 'DR': 27145, nan: 13541, 'PKWY': 6605, 'CT': 3287, 'TER': 2222}
PROPERTY GROUP {6124: 639, 21995: 605, 654907: 359, 13851: 320, 20890: 316, 25993: 313, 20386: 294, 1655: 275, 19382: 274, 2991: 269}
SSA (1.0, 33.76919703534139, 33.76919703534139, 33.76919703534139, 69.0, 1356267)
LATITUDE (41.644670132, 41.770945238, 41.853903366, 41.913405744, 42.02268599, 1510)
LONGITUDE (-87.914435848, -87.713879745, -87.669900427, -87.632909287, -87.524679151, 1510)

Community Areas (1.0, 24.0, 36.0, 58.0, 77.0, 2279)
Zip Codes {21569.0: 79463, 21861.0: 69403, 21546.0: 67402, 21202.0: 63458, 22257.0: 61235, 21554.0: 60879, 21559.0: 60507, 14924.0: 59678, 4299.0: 58145, 21867.0: 57676}
Boundaries - ZIP Codes {57.0: 80508, 19.0: 68579, 61.0: 67501, 25.0: 63330, 59.0: 62103, 11.0: 61950, 58.0: 60268, 5.0: 59985, 32.0: 58309, 37.0: 57971}
Census Tracts (1.0, 179.0, 374.0, 572.0, 801.0, 1545)
Wards (1.0, 12.0, 25.0, 37.0, 50.0, 2279)
Historical Wards 2003-2015 (1.0, 14.0, 27.508575602995865, 41.0, 53.0, 2279)



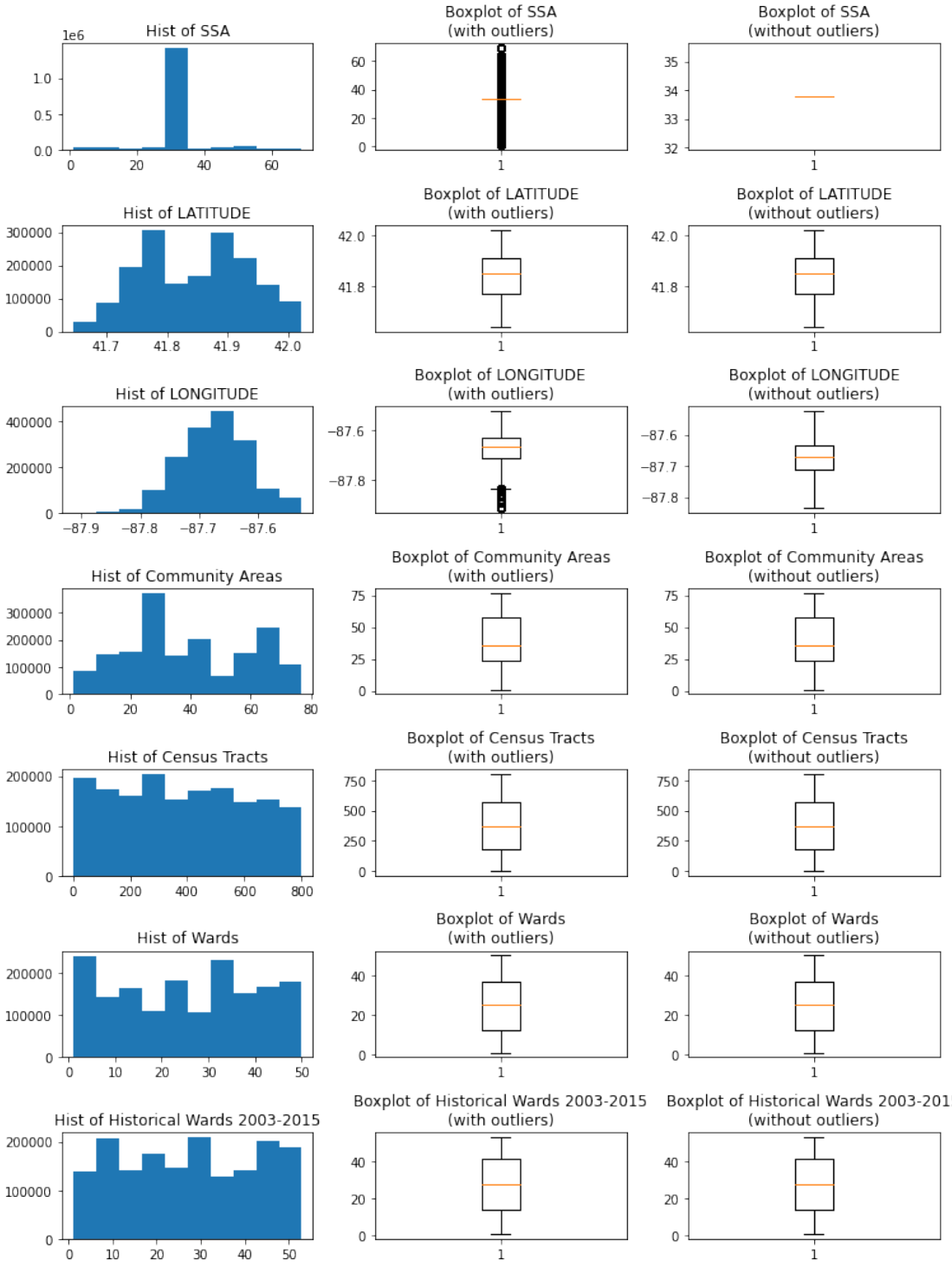
- 通过数据对象之间的相似性来填补缺失值

[23]:
```
processing(csv_file_path, attributeType, MissingProcessing.similarityFill)
```

```
VIOLATION CODE {'CN190019': 89995, 'CN196029': 58136, 'CN061014': 51946,
'EV1110': 43700, 'CN070024': 43673, 'CN193110': 37127, 'CN104015': 34641,
'CN070014': 32093, 'CN197019': 30793, 'NC2011': 28750}
VIOLATION STATUS {'OPEN': 1030958, 'COMPLIED': 641247, 'NO ENTRY': 5583}
VIOLATION DESCRIPTION {'ARRANGE PREMISE INSPECTION': 90004, 'POST OWNER/MANAGERS
NAME/#': 58136, 'REPAIR EXTERIOR WALL': 51946, 'MAINTAIN OR REPAIR ELECT ELEVA':
43700, 'REPAIR PORCH SYSTEM': 43673, 'VACANT BUILDING - REGISTER': 37127,
'REPLCE WINDOW PANES, PLEXGLAS': 34641, 'REPAIR EXTERIOR STAIR': 32093, 'INSTALL
SMOKE DETECTORS': 30846, 'PLANS & PERMITS REQ - CONTRCTR': 28750}
VIOLATION LOCATION {nan: 897282, 'OTHER   :    :': 284277, 'OTHER   :
:OTHER': 35182, 'OTHER   :    :BUILDING': 21934, 'EXTERIOR:E   :': 21522,
'EXTERIOR:W   :': 19460, 'EXTERIOR:S   :': 17053, 'OTHER   :    :BL00006':
16951, 'EXTERIOR:N   :': 16004, 'INTERIOR:001 :': 13016}
VIOLATION ORDINANCE {'Arrange for inspection of premises. (13-12-100)': 89995,
"Post name, address, and telephone of owner, owner's agent for managing,
controlling or collecting rents, and any other person managing or controlling
building conspicuously where accessible or visible to public way. (13-12-030)":
58136, 'Failed to maintain the exterior walls of a building or structure free
from holes, breaks, loose or rotting boards or timbers and any other conditions
which might admit rain or dampness to the walls.  (13-196-530(b), 13-196-641)':
51946, nan: 47581, 'Failed to maintain electric elevator equipment provided at
premises in safe and sound working condition. (13-196-590, 13-196-630(b),
18-30-001)': 43700, 'Failed to repair or replace defective or missing members of
porch system.  (13-196-570, 13-196-641)': 43673, 'Register vacant building
within 30 days of it becoming vacant, or within 30 days after assuming ownership
of an existing vacant building. (13-12-125(a)). Building must be kept in
compliance with all vacant building requirements pursuant to 13-12-135. See
Vacant Building Ordinance and registration form at
https://ipiweb.cityofchicago.org/VBR': 37127, 'Replace broken, missing or
defective window panes. (13-196-550 A)': 34641, 'Failed to maintain exterior
stairways in safe condition and in sound repair.  (13-196-570, 13-196-641)':
32093, 'Install and maintain approved smoke detectors. (13-196-100 thru
13-196-160) Install a smoke detector in every dwelling unit. Install one on any
living level with a habitable room or unenclosed heating plant, on the uppermost
ceiling of enclosed porch stairwell, and within 15 feet of every sleeping room.
```

Be sure the detector is at least 4 inches from the wall, 4 to 12 inches from the ceiling, and not above door or window.': 30846}
INSPECTOR ID {'BL00444': 79336, 'BL01000': 58497, 'BL00831': 52450, 'BL00746': 48797, 'BL00941': 48717, 'BL00812': 44556, 'BL00722': 43139, 'BL01037': 39108, 'BL01041': 38331, 'BL01039': 34976}
INSPECTION NUMBER {11247474: 535, 11587123: 209, 12964992: 135, 10669347: 123, 2323972: 123, 12955264: 120, 11356407: 106, 12253306: 104, 11995426: 92, 11041943: 82}
INSPECTION STATUS {'FAILED': 1159758, 'PASSED': 293076, 'CLOSED': 224784, 'HOLD': 154, nan: 16}
INSPECTION WAIVED {'N': 1677788}
INSPECTION CATEGORY {'COMPLAINT': 1186426, 'PERIODIC': 415176, 'PERMIT': 73600, 'REGISTRATION': 2586}
DEPARTMENT BUREAU {'CONSERVATION': 1110911, 'DEMOLITION': 125464, 'SPECIAL TASK FORCE': 115885, 'ELEVATOR': 85805, 'ELECTRICAL': 37243, 'VENTILATION': 32108, 'BOILER': 31235, 'NEW CONSTRUCTION': 29938, 'REFRIGERATION': 29681, 'PLUMBING': 28199}
ADDRESS {'1900 N AUSTIN AVE': 639, '11601 W TOUHY AVE': 605, '2545 W FITCH AVE': 316, '7200 S COLES AVE': 294, '727 E 60TH ST': 274, '1124 W WILSON AVE': 258, '4408 S DREXEL BLVD': 256, '8228 S SOUTH SHORE DR': 253, '55 E WASHINGTON ST': 253, '5501 W WASHINGTON BLVD': 250}
STREET NUMBER {200: 1704, 400: 1669, 500: 1542, 1000: 1501, 300: 1483, 600: 1462, 1900: 1442, 901: 1431, 1: 1419, 100: 1389}
STREET DIRECTION {'S': 683917, 'W': 500418, 'N': 395246, 'E': 98207}
STREET NAME {'MICHIGAN': 17980, 'ASHLAND': 16992, 'WESTERN': 13327, 'KEDZIE': 13034, 'HALSTED': 11268, 'WABASH': 10941, 'CALIFORNIA': 10747, 'CLARK': 10671, 'DR MARTIN L KING JR': 10614, 'INDIANA': 10612}
STREET TYPE {'AVE': 940725, 'ST': 523743, 'BLVD': 59536, 'PL': 57665, 'RD': 41100, 'DR': 27145, nan: 13541, 'PKWY': 6605, 'CT': 3287, 'TER': 2222}
PROPERTY GROUP {6124: 639, 21995: 605, 654907: 359, 13851: 320, 20890: 316, 25993: 313, 20386: 294, 1655: 275, 19382: 274, 2991: 269}
SSA (1.0, 33.76919703534139, 33.76919703534139, 33.76919703534139, 69.0, 1356267)
LATITUDE (41.644670132, 41.770945238, 41.853903366, 41.913405744, 42.02268599, 1510)
LONGITUDE (-87.914435848, -87.713879745, -87.669900427, -87.632909287, -87.524679151, 1510)

Community Areas (1.0, 24.0, 36.0, 58.0, 77.0, 2279)
Zip Codes {21569.0: 79463, 21861.0: 69403, 21546.0: 67402, 21202.0: 63458, 22257.0: 61235, 21554.0: 60879, 21559.0: 60507, 14924.0: 59678, 4299.0: 58145, 21867.0: 57676}
Boundaries - ZIP Codes {57.0: 80508, 19.0: 68579, 61.0: 67501, 25.0: 63330, 59.0: 62103, 11.0: 61950, 58.0: 60268, 5.0: 59985, 32.0: 58309, 37.0: 57971}
Census Tracts (1.0, 179.0, 374.0, 572.0, 801.0, 1545)
Wards (1.0, 12.0, 25.0, 37.0, 50.0, 2279)
Historical Wards 2003-2015 (1.0, 14.0, 27.508575602995865, 41.0, 53.0, 2279)


Hist of SSA
1e6
Boxplot of SSA
(with outliers)
Boxplot of SSA
(without outliers)
Hist of LATITUDE
Boxplot of LATITUDE
(with outliers)
Boxplot of LATITUDE
(without outliers)
Hist of LONGITUDE
Boxplot of LONGITUDE
(with outliers)
Boxplot of LONGITUDE
(without outliers)
Hist of Community Areas
Boxplot of Community Areas
(with outliers)
Boxplot of Community Areas
(without outliers)
Hist of Census Tracts
Boxplot of Census Tracts
(with outliers)
Boxplot of Census Tracts
(without outliers)
Hist of Wards
Boxplot of Wards
(with outliers)
Boxplot of Wards
(without outliers)
Hist of Historical Wards 2003-2015
Boxplot of Historical Wards 2003-2015
(with outliers)
Boxplot of Historical Wards 2003-2015
(without outliers)