山东大学计算机科学与技术学院

可视化技术课程实验报告

学号: 202300130220 姓名: 刘傲宇 班级: 数据科学与大数据技术班

实验题目:数据质量实践

实验学时: 2 实验日期: 2025/9/26

实验目标:

本次实验主要围绕宝可梦数据集进行分析,考察在拿到数据后如何对现有的数据进行预处理清洗操作,建立起对于脏数据、缺失数据等异常情况的一套完整流程的认识。

实验步骤与内容:

csv 格式文件不好处理, 转换为 xIs 格式并下载相关库

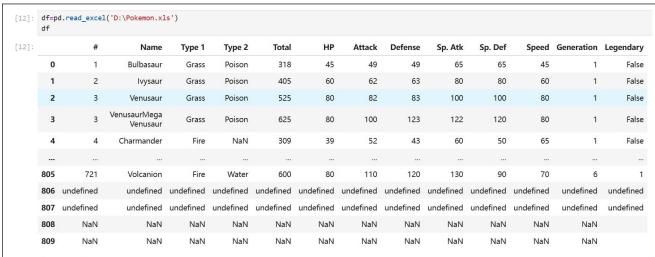
```
[10]: !pip install xlrd --upgrade
    Collecting xlrd
     Downloading xlrd-2.0.2-py2.py3-none-any.whl.metadata (3.5 kB)
    Downloading xlrd-2.0.2-py2.py3-none-any.whl (96 kB)
      ----- 0.0/96.6 kB ? eta -:--:-
      ---- 10.2/96.6 kB ? eta -:--:--
      --- ----- 10.2/96.6 kB ? eta -:--:--
         ------ 10.2/96.6 kB ? eta -:--:--
         ------ 10.2/96.6 kB ? eta -:--:--
          ----- 10.2/96.6 kB ? eta -:--:--
         ------ 10.2/96.6 kB ? eta -:--:--
         ----- 10.2/96.6 kB ? eta -:--:--
         ------ 10.2/96.6 kB ? eta -:--:--
         ------ 10.2/96.6 kB ? eta -:--:--
           ----- 10.2/96.6 kB ? eta -:--:--
           ----- 10.2/96.6 kB ? eta -:--:--
            ----- 10.2/96.6 kB ? eta -:--:--
             ----- 10.2/96.6 kB ? eta -:--:--
             ----- 10.2/96.6 kB ? eta -:--:--
           ----- 10.2/96.6 kB ? eta -:--:--
           ----- 10.2/96.6 kB ? eta -:--:--
           ----- 10.2/96.6 kB ? eta -:--:--
             ----- 10.2/96.6 kB ? eta -:--:--
            ----- 71.7/96.6 kB 71.4 kB/s eta 0:00:01
                   ----- ---- 71.7/96.6 kB 71.4 kB/s eta 0:00:01
      ----- 71.7/96.6 kB 71.4 kB/s eta 0:00:01
```

Installing collected packages: xlrd Successfully installed xlrd-2.0.2

导入库

[11]: import pandas as pd from pandas import DataFrame import numpy as np

读取文件



810 rows × 13 columns

去除最后的无意义行和空行

df=d·	f[:-4]											
	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendar
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	Fals
1	2	lvysaur	Grass	Poison	405	60	62	63	80	80	60	1	Fals
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	Fals
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	Fals
4	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	Fals

801	719	Diancie	Rock	Fairy	600	50	100	150	100	150	50	6	
802	719	DiancieMega Diancie	Rock	Fairy	700	50	160	110	160	110	110	6	
803	720	HoopaHoopa Confined	Psychic	Ghost	600	80	110	60	150	130	70	6	
804	720	HoopaHoopa Unbound	Psychic	Dark	680	80	160	60	170	130	80	6	

Fire Water 600 80

806 rows × 13 columns

805 721

type2 存在异常的数值取值,可清空

Volcanion

```
[14]: condition=df['Type 2'].apply(lambda x: isinstance(x, (int, float)))
    df.loc[condition,'Type 2']=np.nan
    df
```

110

120

130

90

70

14]:		#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
	0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False
	1	2	lvysaur	Grass	Poison	405	60	62	63	80	80	60	1	False
	2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False
	3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
	4	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False
			(•••					
	801	719	Diancie	Rock	Fairy	600	50	100	150	100	150	50	6	1
	802	719	DiancieMega Diancie	Rock	Fairy	700	50	160	110	160	110	110	6	1
	803	720	HoopaHoopa Confined	Psychic	Ghost	600	80	110	60	150	130	70	6	1
	804	720	HoopaHoopa Unbound	Psychic	Dark	680	80	160	60	170	130	80	6	1
	805	721	Volcanion	Fire	Water	600	80	110	120	130	90	70	6	1

806 rows × 13 columns

数据集中存在重复值

[15]: df_clean=df.drop_duplicates(subset=['Name'])
df_clean

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False
1	2	lvysaur	Grass	Poison	405	60	62	63	80	80	60	1	False
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
4	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False
801	719	Diancie	Rock	Fairy	600	50	100	150	100	150	50	6	1
802	719	DiancieMega Diancie	Rock	Fairy	700	50	160	110	160	110	110	6	1
803	720	HoopaHoopa Confined	Psychic	Ghost	600	80	110	60	150	130	70	6	1
804	720	HoopaHoopa Unbound	Psychic	Dark	680	80	160	60	170	130	80	6	1
805	721	Volcanion	Fire	Water	600	80	110	120	130	90	70	6	

801 rows × 13 columns

[16]: has_dup=df_clean.duplicated().any()
 if has_dup:

print("仍有重复值")

else:

print("无重复值")

无重复值

Attack 属性存在过高的异常值

```
import matplotlib.pyplot as plt
  [17]:
          plt.scatter(range(0,df_clean.shape[0]),df_clean.iloc[:,6])
          <matplotlib.collections.PathCollection at 0x1f9236d8350>
           1000
            800
            600
            400
            200
               0
                     0
                              100
                                       200
                                                 300
                                                           400
                                                                    500
                                                                              600
                                                                                        700
                                                                                                 800
         df_clean['Attack'] = pd.to_numeric(df_clean['Attack'], errors='coerce')
         df_clean_2 = df_clean[df_clean['Attack'] <= 400].dropna(subset=['Attack'])</pre>
         df_clean_2
                                  Type 1 Type 2 Total HP Attack Defense Sp. Atk Sp. Def Speed Generation Legendary
                            Name
        0
                         Bulbasaur
                                                 318
                                                     45
                                                           49.0
                                                                    49
                                                                           65
                                                                                  65
                                                                                         45
                                                                                                           False
                                    Grass
       1
            2
                                                 405
                                                     60
                                                           62.0
                                                                    63
                                                                           80
                                                                                  80
                                                                                         60
                                                                                                           False
                           lvysaur
                                   Grass
                                         Poison
            3
       2
                          Venusaur
                                   Grass
                                         Poison
                                                 525
                                                     80
                                                           82.0
                                                                    83
                                                                          100
                                                                                 100
                                                                                         80
                                                                                                    1
                                                                                                           False
       3
                                                          100.0
                                                                   123
                                                                          122
                                                                                 120
                                                                                         80
            3 VenusaurMega Venusaur
                                         Poison
                                                 625
                                                     80
                                                                                                           False
                                   Grass
                                                                    43
                                                                           60
                                                                                  50
                                                                                                    1
        4
            4
                       Charmander
                                                     39
                                                           52.0
                                                                                         65
                                                                                                           False
                                     Fire
                                           NaN
                                                 309
                                                          100.0
                                                                   150
                                                                          100
                                                                                 150
                                                                                         50
                                                                                                    6
     801 719
                                                 600
                                                     50
                                                                                                              1
                           Diancie
                                    Rock
                                           Fairy
     802 719
                                                                   110
                                                                                        110
                                                                                                    6
                 DiancieMega Diancie
                                           Fairy
                                                 700
                                                     50
                                                          160.0
                                                                          160
                                                                                 110
                                                                                                              1
                                                                    60
                                                                                         70
                                                                                                    6
     803 720
               HoopaHoopa Confined
                                  Psychic
                                                 600
                                                     80
                                                          110.0
                                                                          150
                                                                                 130
                                                                                                              1
                                          Ghost
                                                                    60
                                                                          170
                                                                                                    6
         720
               HoopaHoopa Unbound
                                                          160.0
                                                                                 130
                                                                                         80
                                                                                                              1
     804
                                  Psychic
                                           Dark
                                                 680
                                                     80
     805 721
                         Volcanion
                                                                   120
                                                                          130
                                                                                  90
                                                                                         70
                                                                                                    6
                                          Water
                                                 600
     798 rows × 13 columns
有数据的 generation 与 Legendary 属性被置换
```

```
condition=df_clean_2['Generation'].apply(lambda x: isinstance(x,(int)))
 [28]:
           row_n=df_clean_2.index[condition==False].tolist()
           row_n
          [771]
[28]:
[29]: def swap_values(df,row1,col1,col2):
         df.loc[row1,col1],df.loc[row1,col2]=df.loc[row1,col2],df.loc[row1,col1]
     swap_values(df_clean_2,771,'Generation','Legendary')
     df_clean_2
[29]:
                             Name Type 1 Type 2 Total HP Attack Defense Sp. Atk Sp. Def Speed Generation Legendary
       0
                           Bulbasaur
                                                                                65
                                                                                              45
                                     Grass Poison
                                                    318
                                                         45
                                                               49.0
                                                                                        65
                                                                                                                  False
       1
            2
                                                    405
                                                         60
                                                               62.0
                                                                        63
                                                                                80
                                                                                        80
                                                                                               60
                                                                                                                  False
                             lvysaur
                                     Grass Poison
       2
            3
                           Venusaur
                                     Grass Poison
                                                    525
                                                         80
                                                               82.0
                                                                        83
                                                                               100
                                                                                       100
                                                                                               80
                                                                                                          1
                                                                                                                  False
                                                                                               80
            3 VenusaurMega Venusaur
                                     Grass Poison
                                                    625
                                                         80
                                                              100.0
                                                                       123
                                                                               122
                                                                                       120
                                                                                                                  False
       4
            4
                         Charmander
                                       Fire
                                             NaN
                                                    309
                                                         39
                                                               520
                                                                        43
                                                                                60
                                                                                        50
                                                                                               65
                                                                                                          1
                                                                                                                  False
       •••
     801 719
                            Diancie
                                      Rock
                                             Fairy
                                                    600
                                                         50
                                                              100.0
                                                                       150
                                                                               100
                                                                                       150
                                                                                              50
                                                                                                          6
                                                                                                                     1
     802 719
                  DiancieMega Diancie
                                      Rock
                                             Fairy
                                                    700
                                                         50
                                                              160.0
                                                                       110
                                                                               160
                                                                                       110
                                                                                              110
                                                                                                          6
                                                                                                                     1
     803 720
                HoopaHoopa Confined Psychic
                                            Ghost
                                                    600
                                                         80
                                                              110.0
                                                                        60
                                                                               150
                                                                                       130
                                                                                              70
                                                                                                          6
                                                                                                                     1
     804 720
                HoopaHoopa Unbound Psychic
                                                              160.0
                                                                        60
                                                                               170
                                                                                              80
                                                                                                          6
                                                                                                                     1
                                             Dark
                                                    680
                                                         80
                                                                                       130
     805 721
                          Volcanion
                                            Water
                                                    600
                                                         80
                                                              110.0
                                                                       120
                                                                               130
                                                                                        90
                                                                                              70
                                                                                                          6
                                                                                                                     1
     798 rows × 13 columns
```

结论分析与体会:

熟悉了分析数据质量的方法,掌握了多种数据清洗的方法。