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xls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

[12]:
```
df=pd.read_excel('D:\Pokemon.xls')
df
```

[12]:

	#	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	Ivysaur	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...
805	721	Volcanion	Fire	Water	600	80	110	120	130	90	70	6	1
806	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
807	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
808	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
809	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

810 rows × 13 columns

去除最后的无意义行和空行

[13]:
```
df=df[:-4]
df
```

[13]:

	#	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	Ivysaur	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

type2 存在异常的数值取值，可清空

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

[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	Ivysaur	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
```

[15]:

	#	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	Ivysaur	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

801 rows × 13 columns

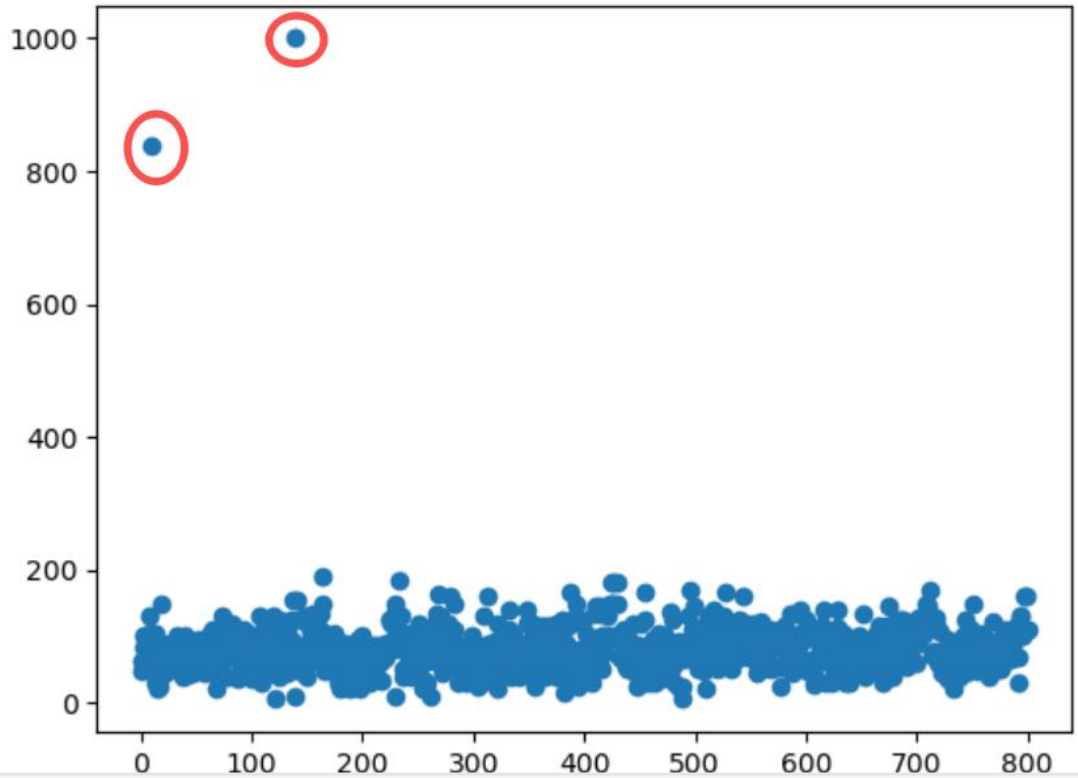
[16]:
```
has_dup=df_clean.duplicated().any()
if has_dup:
    print("仍有重复值")
else:
    print("无重复值")
```

无重复值

Attack 属性存在过高的异常值

[17]:
```
import matplotlib.pyplot as plt
plt.scatter(range(0,df_clean.shape[0]),df_clean.iloc[:,6])
```

[17]:
```
<matplotlib.collections.PathCollection at 0x1f9236d8350>
```



[21]:
```
df_clean['Attack'] = pd.to_numeric(df_clean['Attack'], errors='coerce')
df_clean_2 = df_clean[df_clean['Attack'] <= 400].dropna(subset=['Attack'])
df_clean_2
```

[21]:

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	318	45	49.0	49	65	65	45	1	False
1	2	Ivysaur	Grass	Poison	405	60	62.0	63	80	80	60	1	False
2	3	Venusaur	Grass	Poison	525	80	82.0	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100.0	123	122	120	80	1	False
4	4	Charmander	Fire	NaN	309	39	52.0	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	Dark	680	80	160.0	60	170	130	80	6	1
805	721	Volcanion	Fire	Water	600	80	110.0	120	130	90	70	6	1

798 rows × 13 columns

有数据的 generation 与 Legendary 属性被置换

[28]:
```
condition=df_clean_2['Generation'].apply(lambda x: isinstance(x,(int)))
row_n=df_clean_2.index[condition==False].tolist()
row_n
```

[28]: [771]

[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	1	Bulbasaur	Grass	Poison	318	45	49.0	49	65	65	45	1	False
1	2	Ivysaur	Grass	Poison	405	60	62.0	63	80	80	60	1	False
2	3	Venusaur	Grass	Poison	525	80	82.0	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100.0	123	122	120	80	1	False
4	4	Charmander	Fire	NaN	309	39	52.0	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	Dark	680	80	160.0	60	170	130	80	6	1
805	721	Volcanion	Fire	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。