

山东大学计算机科学与技术学院

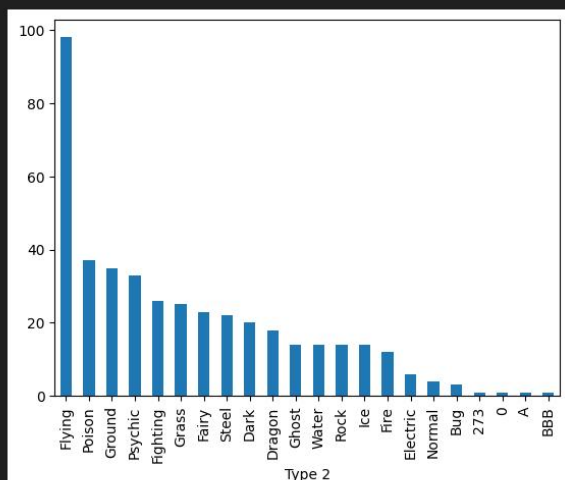
大数据分析实践 课程实验报告

学号：202300130086	姓名：张珈恺	班级：23 数据																																																																																																																																																																							
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。																																																																																																																																																																									
实验环境： Python3.9, Jupyter notebook																																																																																																																																																																									
实验步骤与内容： 1、导入数据集																																																																																																																																																																									
<div><pre>import numpy as np import pandas as pd import matplotlib data = pd.read_csv("C:\\SDU\\cloudCache\\张珈恺\\大数据分析实践\\Pokemon.csv", encoding='Windows-1252') data</pre><div>✓ 62s</div><table><thead><tr><th>#</th><th>Name</th><th>Type 1</th><th>Type 2</th><th>Total</th><th>HP</th><th>Attack</th><th>Defense</th><th>Sp. Atk</th><th>Sp. Def</th><th>Speed</th><th>Generation</th><th>Legendary</th></tr></thead><tbody><tr><td>0</td><td>1</td><td>Bulbasaur</td><td>Grass</td><td>Poison</td><td>318</td><td>45</td><td>49</td><td>49</td><td>65</td><td>65</td><td>45</td><td>1</td><td>FALSE</td></tr><tr><td>1</td><td>2</td><td>Ivysaur</td><td>Grass</td><td>Poison</td><td>405</td><td>60</td><td>62</td><td>63</td><td>80</td><td>80</td><td>60</td><td>1</td><td>FALSE</td></tr><tr><td>2</td><td>3</td><td>Venusaur</td><td>Grass</td><td>Poison</td><td>525</td><td>80</td><td>82</td><td>83</td><td>100</td><td>100</td><td>80</td><td>1</td><td>FALSE</td></tr><tr><td>3</td><td>3</td><td>VenusaurMega Venusaur</td><td>Grass</td><td>Poison</td><td>625</td><td>80</td><td>100</td><td>123</td><td>122</td><td>120</td><td>80</td><td>1</td><td>FALSE</td></tr><tr><td>4</td><td>4</td><td>Charmander</td><td>Fire</td><td>NaN</td><td>309</td><td>39</td><td>52</td><td>43</td><td>60</td><td>50</td><td>65</td><td>1</td><td>FALSE</td></tr><tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr><tr><td>805</td><td>721</td><td>Volcanion</td><td>Fire</td><td>Water</td><td>600</td><td>80</td><td>110</td><td>120</td><td>130</td><td>90</td><td>70</td><td>6</td><td>TRUE</td></tr><tr><td>806</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td></tr><tr><td>807</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td><td>undefined</td></tr><tr><td>808</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr><tr><td>809</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr></tbody></table><div>810 rows x 13 columns</div></div>			#	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	TRUE	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	NaN	809	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
#	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	TRUE																																																																																																																																																												
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	NaN																																																																																																																																																												
809	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN																																																																																																																																																												
2、删除无意义数据 最后四行数据无意义，直接删去。																																																																																																																																																																									
<div><pre>data.drop([806, 807, 808, 809], axis=0, inplace=True) data</pre><div>✓ 0.0s</div><table><thead><tr><th>#</th><th>Name</th><th>Type 1</th><th>Type 2</th><th>Total</th><th>HP</th><th>Attack</th><th>Defense</th><th>Sp. Atk</th><th>Sp. Def</th><th>Speed</th><th>Generation</th><th>Legendary</th></tr></thead><tbody><tr><td>0</td><td>1</td><td>Bulbasaur</td><td>Grass</td><td>Poison</td><td>318</td><td>45</td><td>49</td><td>49</td><td>65</td><td>65</td><td>45</td><td>1</td><td>FALSE</td></tr><tr><td>1</td><td>2</td><td>Ivysaur</td><td>Grass</td><td>Poison</td><td>405</td><td>60</td><td>62</td><td>63</td><td>80</td><td>80</td><td>60</td><td>1</td><td>FALSE</td></tr><tr><td>2</td><td>3</td><td>Venusaur</td><td>Grass</td><td>Poison</td><td>525</td><td>80</td><td>82</td><td>83</td><td>100</td><td>100</td><td>80</td><td>1</td><td>FALSE</td></tr><tr><td>3</td><td>3</td><td>VenusaurMega Venusaur</td><td>Grass</td><td>Poison</td><td>625</td><td>80</td><td>100</td><td>123</td><td>122</td><td>120</td><td>80</td><td>1</td><td>FALSE</td></tr><tr><td>4</td><td>4</td><td>Charmander</td><td>Fire</td><td>NaN</td><td>309</td><td>39</td><td>52</td><td>43</td><td>60</td><td>50</td><td>65</td><td>1</td><td>FALSE</td></tr><tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr><tr><td>801</td><td>719</td><td>Diancie</td><td>Rock</td><td>Fairy</td><td>600</td><td>50</td><td>100</td><td>150</td><td>100</td><td>150</td><td>50</td><td>6</td><td>TRUE</td></tr><tr><td>802</td><td>719</td><td>DiancieMega Diancie</td><td>Rock</td><td>Fairy</td><td>700</td><td>50</td><td>160</td><td>110</td><td>160</td><td>110</td><td>110</td><td>6</td><td>TRUE</td></tr><tr><td>803</td><td>720</td><td>HoopaHoopa Confined</td><td>Psychic</td><td>Ghost</td><td>600</td><td>80</td><td>110</td><td>60</td><td>150</td><td>130</td><td>70</td><td>6</td><td>TRUE</td></tr><tr><td>804</td><td>720</td><td>HoopaHoopa Unbound</td><td>Psychic</td><td>Dark</td><td>680</td><td>80</td><td>160</td><td>60</td><td>170</td><td>130</td><td>80</td><td>6</td><td>TRUE</td></tr><tr><td>805</td><td>721</td><td>Volcanion</td><td>Fire</td><td>Water</td><td>600</td><td>80</td><td>110</td><td>120</td><td>130</td><td>90</td><td>70</td><td>6</td><td>TRUE</td></tr></tbody></table><div>806 rows x 13 columns</div></div>			#	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	TRUE	802	719	DiancieMega Diancie	Rock	Fairy	700	50	160	110	160	110	110	6	TRUE	803	720	HoopaHoopa Confined	Psychic	Ghost	600	80	110	60	150	130	70	6	TRUE	804	720	HoopaHoopa Unbound	Psychic	Dark	680	80	160	60	170	130	80	6	TRUE	805	721	Volcanion	Fire	Water	600	80	110	120	130	90	70	6	TRUE
#	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	TRUE																																																																																																																																																												
802	719	DiancieMega Diancie	Rock	Fairy	700	50	160	110	160	110	110	6	TRUE																																																																																																																																																												
803	720	HoopaHoopa Confined	Psychic	Ghost	600	80	110	60	150	130	70	6	TRUE																																																																																																																																																												
804	720	HoopaHoopa Unbound	Psychic	Dark	680	80	160	60	170	130	80	6	TRUE																																																																																																																																																												
805	721	Volcanion	Fire	Water	600	80	110	120	130	90	70	6	TRUE																																																																																																																																																												
3、删除存在异常值的数据 对 Type2 列的取值频次进行可视化，发现存在少数无意义异常值，删除。																																																																																																																																																																									

```
type2 = data['Type 2'].value_counts(dropna=True)
type2.plot(kind='bar')
```

✓ 0.4s

<Axes: xlabel='Type 2'>



```
data.drop(data[(data['Type 2']=='273') | (data['Type 2']=='0') | (data['Type 2']=='A') | (data['Type 2']=='BBB')].index, inplace=True, axis=0)
data
```

✓ 0.0s

#	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	TRUE
802	719	DiancieMega Diancie	Rock Fairy	700	50	160	110	160	110	110	6	TRUE
803	720	HoopaHoopa Confined	Psychic Ghost	600	80	110	60	150	130	70	6	TRUE
804	720	HoopaHoopa Unbound	Psychic Dark	680	80	160	60	170	130	80	6	TRUE
805	721	Volcanion	Fire Water	600	80	110	120	130	90	70	6	TRUE

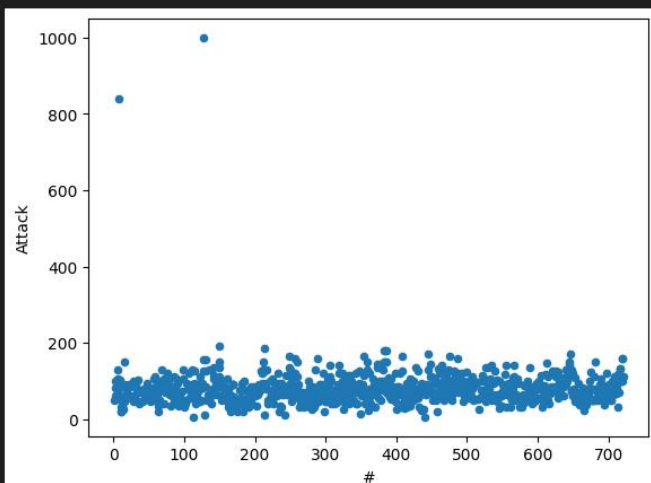
802 rows x 13 columns

对 Attack 列的值进行可视化，发现存在少数无意义异常值，删除

```
data['#'] = pd.to_numeric(data['#'], errors='coerce')
data['Attack'] = pd.to_numeric(data['Attack'], errors='coerce')
data[['#', 'Attack']].plot(kind='scatter', x='#', y='Attack')
```

✓ 0.0s

<Axes: xlabel='#', ylabel='Attack'>



```
data.drop(data[data['Attack']>800].index, axis=0, inplace=True)
data
```

✓ 0.0s

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1.0	Bulbasaur	Grass	Poison	318	45	49.0	49	65	65	45	1	FALSE
1	2.0	Ivysaur	Grass	Poison	405	60	62.0	63	80	80	60	1	FALSE
2	3.0	Venusaur	Grass	Poison	525	80	82.0	83	100	100	80	1	FALSE
3	3.0	VenusaurMega Venusaur	Grass	Poison	625	80	100.0	123	122	120	80	1	FALSE
4	4.0	Charmander	Fire	NaN	309	39	52.0	43	60	50	65	1	FALSE
...
801	719.0	Diancie	Rock	Fairy	600	50	100.0	150	100	150	50	6	TRUE
802	719.0	DiancieMega Diancie	Rock	Fairy	700	50	160.0	110	160	110	110	6	TRUE
803	720.0	HoopaHoopa Confined	Psychic	Ghost	600	80	110.0	60	150	130	70	6	TRUE
804	720.0	HoopaHoopa Unbound	Psychic	Dark	680	80	160.0	60	170	130	80	6	TRUE
805	721.0	Volcanion	Fire	Water	600	80	110.0	120	130	90	70	6	TRUE

800 rows × 13 columns

4、删除重复值

```
data.drop_duplicates(inplace=True)
data
```

✓ 0.0s

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1.0	Bulbasaur	Grass	Poison	318	45	49.0	49	65	65	45	1	FALSE
1	2.0	Ivysaur	Grass	Poison	405	60	62.0	63	80	80	60	1	FALSE
2	3.0	Venusaur	Grass	Poison	525	80	82.0	83	100	100	80	1	FALSE
3	3.0	VenusaurMega Venusaur	Grass	Poison	625	80	100.0	123	122	120	80	1	FALSE
4	4.0	Charmander	Fire	NaN	309	39	52.0	43	60	50	65	1	FALSE
...
801	719.0	Diancie	Rock	Fairy	600	50	100.0	150	100	150	50	6	TRUE
802	719.0	DiancieMega Diancie	Rock	Fairy	700	50	160.0	110	160	110	110	6	TRUE
803	720.0	HoopaHoopa Confined	Psychic	Ghost	600	80	110.0	60	150	130	70	6	TRUE
804	720.0	HoopaHoopa Unbound	Psychic	Dark	680	80	160.0	60	170	130	80	6	TRUE
805	721.0	Volcanion	Fire	Water	600	80	110.0	120	130	90	70	6	TRUE

795 rows × 13 columns

5、有两条数据的 generation 与 Legendary 属性被置换

```
data[~(data['Generation'].isin(['1', '2', '3', '4', '5', '6']))]
```

✓ 0.0s

	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
11	9.0	Blastoise	Water	NaN	530	79	83.0	100	85	105	78	FALSE	1
32	25.0	Pikachu	Electric	NaN	320	35	55.0	40	50	50	90	FALSE	0
408	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
771	695.0	Heliolisk	Electric	Normal	481	62	55.0	52	109	94	109	undefined	FALSE

结论与体会：

本次实验围绕宝可梦数据集完成数据质量优化，通过删除无意义行、异常值、重复值，修正属性置换问题，构建了完整数据清洗流程。我掌握了用 **Pandas** 处理脏数据的方法，也认识到可视化能高效定位异常。数据质量是分析基础，规范预处理步骤才能保障后续结果可靠。