

Data Cleaning and Preprocessing

Data Analytics Internship

Anisha More

Dataset Overview

Initial State

Raw dataset with inconsistencies

- Missing values present
- Duplicate records found
- Formatting issues

Missing Value Handling

Identified null values

Used mean/median imputation

Removed rows with excessive missing data



Duplicate Removal



Detected duplicate records



Removed duplicates



Ensured data uniqueness

DATABASE						
	TabLs	頁数	PAVS	古指	DAR	背効
1	241	1	300	125	3	000
2	0A1K	6	350	225	0	409
3	DA5300	8	850	440	0	443
6	200,400	5	970	435	0	140
4	571.00	40	6,510	148	30	800

6	SA06_	35	7450	110	70	430
61	S44,000	90	6100	105	80	140
45	Z7100	16	4410	132	10	445
25	Z0720	18	1000	02	10	150
25	Z0,50	18	1865	413	30	440
12	27,00	45	1200	7,333	0	120
2	1falno0	15	1800	3,674	9	220
20	D0poo	1A	12X7	1,40	5	420
90	EA068	46	1700	1233	3	320
20	Fk3t80	3/4	100T	128	20	340
20	DA60	S12	1000	847	4	440

	NAA5HCO	琵琶	EA0C	琵琶	TBR	琵琶
1	6E0.5	16	3700	535	5	1300
3	2.1661	65	2408	317	0	400
7	23500	45	2300	239	6	300
5	2/200	14	3700	25	0	430
51	78300	37	2,30	5	0	021
00	47300	32	3100	2810	9	401
37	75903	35	2500	-56	6	330
29	25896	47	2100	5	0	455
78	27393	33	2810	17	1	300
28	17040	34	2201	21	6	300
166	111002	26	2900	842	1	300
1821	0,300	22	7510	143	8	365
120	DA300C2	22	3760		3	200

262	25,00	32	8300	2,20	3	474
7	69400	474	14310	145	20	435
310	X804	9/3	3000	3,735	75	420

TABLSE					
RAC	DSANGL	10ANO	FS	440KT	S4906
200008	100000	6,56	20	30	00
290004	80,000	2,56	50	20	40
232008	3AL010	1,55	30	40	24
26008	39,000	1,53	60	20	28
100093	894042	1,06	30	20	15

242008	64000	1,00	80	350	26
244006	38000	1,33	80	21	30
180004	22000	3,20	20	40	20
119056	2,805	1,50	230	20	245
35008	0,53	5,05	60	30	515
201096	5,59	2,55	48	93	423
230198	8,60	10398	20	90	85
350196	1843	3,65	40	20	10
280096	8,85	24006	20	24	44
1,5098	048	3,00	60	34	4
340,08	028	1,3,5	45	45	2

	FBALS	TIBD	E	EdonET	
13PDS	S0ER	30	44	3L	35
10056	3139	2,98	21	10	60
56004	5355	2,88	30	27	19
29058	4457	72	41	40	19
22056	3458	- 90	40	40	150
150908	4,555	1,60	16	25	890
90033	5785	4,02	44	40	140
90094	5745	2,00	170	240	160
150004	4285	3,50	38	40	262
10308	5464	4,60	41	40	190
210705	12,3007	2,47	50	42	681
86096	144,006	2,00	28	227	600
24776	14,006	3,00	20	40	150

104708	14,1422	2,92	30	40	3
166706	140062	7,53	44	25	3
1,6038	15,90	4,09	80	48	20

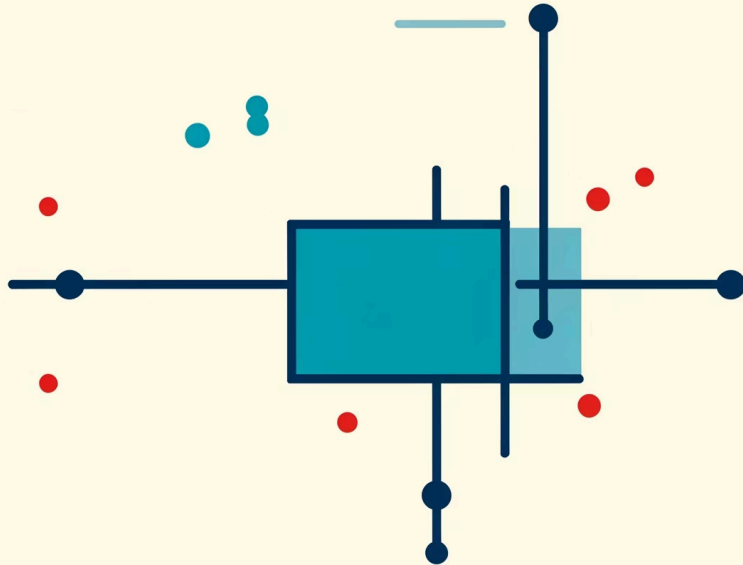
Standardization

Standardized date formats

Unified categorical values

Corrected inconsistent entries

Outlier Detection



01

Used boxplots for detection

02

Identified extreme values

03

Treated outliers appropriately

Final Outcome



Clean and structured
dataset



Improved data quality



Ready for analysis

 Speak: