

# 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						
	TABLES	行数	PAWS	否倍	DAR	肯均
1	241	1	900	120	8	000
2	0A1K	6	250	225	0	409
3	DA5300	8	850	440	0	413
5	200400	5	970	435	0	140
4	57100	40	6,510	149	30	800

6	SAQB	36	T480	110	70	430
81	S444000	90	6100	105	80	140
45	Z7100	16	4410	132	10	445
25	Z0720	18	1000	02	10	150
25	Z0750	18	1865	413	30	440
12	27.00	45	1200	1,833	0	120
2	1f1000	15	1800	3,64	8	420
20	D4000	1A	19X7	1,40	5	420
90	14059	45	1900	1233	3	320
20	FKEBO	3/4	100T	128	20	880
20	DA60	SI2	1000	847	4	440

	NAA6BCD	凭据	EAIC	民需	TBR	肯怒
1	6605	18	5700	535	5	1300
3	2.1661	65	2408	317	0	400
7	Z8800	45	2300	239	6	800
5	Z/200	14	8700	25	0	430
51	78800	37	2.30	日	0	021
00	4/300	32	8100	280	9	401
37	75608	35	2500	-56	6	300
29	25896	47	2300	5	0	155
78	27.338	33	280	17	1	300
28	1/2040	34	220I	21	6	300
166	111000	26	2900	642	1	300
1321	0..300	22	7510	148	8	365
120	DA30002	22	3760		3	200
262	26,00	32	8300	2,20	3	474
7	G360	474	1420	145	20	435
30	R8004	9.8	3000	3,795	75	420

TABLES					
RAC	D5ANGL	10AND	F5	A40KT	S4906
200008	100000	6,56	20	30	00
290008	80.000	2,56	50	20	40
232008	3A.010	1,55	30	40	24
26009	39,000	1,56	60	20	28
100098	89042	1,36	30	20	15

242008	64000	1,00	80	450	26
244006	38000	1,93	80	21	80
130004	22000	3,20	20	40	20
190066	2,805	1,50	230	20	245
35008	0,55	5,55	60	30	515
201096	5,59	2,55	40	93	423
290198	8,60	10368	20	90	85
360096	1043	3,65	40	20	10
2800%	8,85	24,016	20	20	44
1,50%	048	0,00	60	34	4
34008	028	1,35	48	46	2

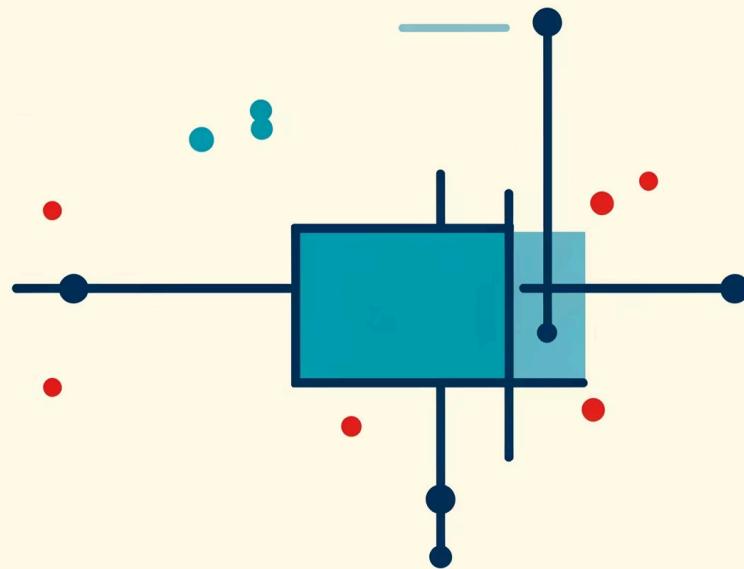
# Standardization

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: