Machine Learning 441 Assignment 1

David Nicolay (26296918) 6 August 2025

Task 1: Analytics Base Table

1. 581012
2. 61
3. Continuous features

Table 1: Data Quality Report for Continuous Features

Feature	**Count**	**% Miss.**	**Card.**	**Min.**	**1^{st} Qrt.**	**Mean**	**Median**	**3^{rd} Qrt.**	**Max.**	**Std. Dev.**
0	581012	0.00	1978	2054845.65	3104928.15	3271134.43	3311628.60	3496222.05	4264440.30	309481.13
1	581012	0.51	361	0.00	58.00	155.66	127.00	260.00	360.00	111.91
2	581012	0.00	576099	0.00	145.49	389.92	318.12	652.53	903.48	280.34
3	581012	0.05	67	0.00	9.00	14.10	13.00	18.00	66.00	7.49
4	581012	0.51	569	-691.00	108.00	269.42	218.00	384.00	1397.00	212.56
5	581012	0.00	581012	-173.07	6.99	46.42	29.91	68.97	600.95	58.30
6	581012	0.51	577988	-1.00	-0.50	-0.00	-0.00	0.50	1.00	0.58
7	581012	0.00	5811	0.00	1106.00	8158.11	1997.00	3328.00	510165098.00	1185156.02
8	581012	0.51	207	0.00	198.00	212.14	218.00	231.00	254.00	26.77
9	581012	1.50	185	0.00	213.00	223.32	226.00	237.00	254.00	19.77
10	581012	0.00	255	0.00	119.00	142.53	143.00	168.00	254.00	38.27
11	581012	0.51	5826	0.00	1024.00	1980.43	1710.00	2550.00	7173.00	1324.25
60	581012	0.00	581012	1.00	145253.75	290506.50	290506.50	435759.25	581012.00	167723.86

4. Target feature data quality report:

Table 3: Data Quality Report for target feature

Feature	Count	Miss.	Card.	Min.	1st Qrt.	Mean	Median	3rd Qrt.	Max.	Std. Dev.
T	580952	60	7	1.00	1.00	2.05	2.00	2.00	7.00	1.40

Task 2: Data Quality Issues

Feature	**Data Quality Issue**	**Justification**
A2	Missing Values	Many records have nulls, impacting model accuracy.
A7	Outlier	The maximum value of this feature is much higher than the mean and the 3rd quartile, indicating that there may be an error.
A15	Cardinality of 1	This categorical variable is redundant and can be removed since all the observations have the same value
A16	Cardinality of 1	This categorical variable is redundant since all the observations have the same value.
A61	Cardinality of 1	This categorical variable is redundant since all the observations have the same value.

Task 3: Addressing The Data Quality Issues

References

Table 2: Categorical Data Quality Report

Feature	**Count**	**%miss**	**Card.**	**Mode**	**Mode Freq.**	**Mode %**	**2^{nd} Mode**	**2^{nd} Mode Freq**	**2^{nd} Mode %**
12	581012	0.00	2	0.00	318579	54.83	1.00	259490	44.66
13	581012	0.00	2	0.00	551128	94.86	1	29884	5.14
14	581012	0.00	2	0.00	327648	56.39	1	253364	43.61
15	581012	0.00	1	0.00	578069	99.49		0	0.00
16	581012	0.00	1	0.00	581012	100.00		0	0.00
17	581012	0.00	3	0.00	538608	92.70	1.00	36589	6.30
18	581012	0.00	2	0.00	291278	50.13	1	289734	49.87
19	581012	0.00	2	0.00	574553	98.89	1.00	3013	0.52
20	581012	0.00	2	0.00	172455	29.68	1.00	900	0.15
21	581012	0.00	2	0.00	573487	98.70	1	7525	1.30
22	581012	0.00	2	0.00	576189	99.17	1	4823	0.83
23	581012	0.00	2	0.00	568616	97.87	1	12396	2.13
24	581012	0.00	2	0.00	579415	99.73	1	1597	0.27
25	581012	0.00	2	0.00	574437	98.87	1	6575	1.13
26	581012	0.00	2	0.00	580907	99.98	1	105	0.02
27	581012	0.00	2	0.00	580833	99.97	1	179	0.03
28	581012	0.00	2	0.00	579865	99.80	1	1147	0.20
29	581012	0.00	2	0.00	548378	94.38	1	32634	5.62
30	581012	0.00	2	0.00	568602	97.86	1	12410	2.14
31	581012	0.00	2	0.00	551041	94.84	1	29971	5.16
32	581012	0.00	2	0.00	563581	97.00	1	17431	3.00
33	581012	0.00	2	0.00	580413	99.90	1	599	0.10
34	581012	0.00	2	0.00	581009	100.00	1	3	0.00
35	581012	0.00	2	0.00	578167	99.51	1	2845	0.49
36	581012	0.00	2	0.00	577590	99.41	1	3422	0.59
37	581012	0.00	2	0.00	579113	99.67	1	1899	0.33
38	581012	0.00	2	0.00	576991	99.31	1	4021	0.69
39	581012	0.00	2	0.00	571753	98.41	1	9259	1.59
40	581012	0.00	2	0.00	580174	99.86	1	838	0.14
41	581012	0.00	2	0.00	547639	94.26	1	33373	5.74
42	581012	0.00	2	0.00	523260	90.06	1	57752	9.94
43	581012	0.00	2	0.00	559734	96.34	1	21278	3.66
44	581012	0.00	2	0.00	580538	99.92	1	474	0.08
45	581012	0.00	2	0.00	578423	99.55	1	2589	0.45
46	581012	0.00	2	0.00	579926	99.81	1	1086	0.19
47	581012	0.00	2	0.00	580066	99.84	1	946	0.16
48	581012	0.00	2	0.00	465765	80.16	1	115247	19.84
49	581012	0.00	2	0.00	550842	94.81	1	30170	5.19
50	581012	0.00	2	0.00	555346	95.58	1	25666	4.42
51	581012	0.00	2	0.00	528493	90.96	1	52519	9.04
52	581012	0.00	2	0.00	535858	92.23	1	45154	7.77
53	581012	0.00	2	0.00	579401	99.72	1	1611	0.28
54	581012	0.00	2	0.00	579121	99.67	1	1891	0.33
55	581012	0.00	2	0.00	580893	99.98	1	119	0.02
56	581012	0.00	2	0.00	580714	99.95	1	298	0.05
57	581012	0.00	2	0.00	565439	97.32	1	15573	2.68
58	581012	0.00	2	0.00	567206	97.62	1	13806	2.38
59	581012	0.00	2	0.00	572262	98.49	1	8750	1.51
61	581012	0.00	1	1.00	581012	100.00		0	0.00