BOUCHER Valentin ZANKOWITCH Alexis

Continuous features

	Count	Miss %	Card	Min	1st Qrt	Mean	Median	3rd Qrt	Max	Std
age	30940	0	72	17	28	38.5607627666	37	48	90	13.6394030688
fnlwgt	30940	0	20880	12285	117849	189786.401422107	178384	237318	1484705	105406.394386105
education-num	30940	0	16	1	9	10.0812540401	10	12	16	2.569966837
capital-gain	30940	0	119	0	0	1081.8129928895	0	0	99999	7443.7730412872
capital-loss	30940	0	91	0	0	86.5699741435	0	0	4356	401.7060231858
hours-per-week	30940	0	93	1	40	40.4089204913	40	45	99	12.336944975

Categorical features

	Count	Miss %	Card	Mode	Mode freq	Mode %	2nd Mode	2nd Mode Freq	2nd Mode %
workclass	30940	5.6	9	Private	21576	69.7	Self-emp-not-inc	2406	7.8
education	30940	0	16	HS-grad	9976	32.2	Some-college	6938	22.4
marital-status	30940	0	7	Married-civ-spouse	14201	45.9	Never-married	10167	32.9
occupation	30940	5.6	15	Prof-specialty	3932	12.7	Craft-repair	3887	12.6
relationship	30940	0	6	Husband	12496	40.4	Not-in-family	7904	25.5
race	30940	0	5	White	26442	85.5	Black	2965	9.6
sex	30940	0	2	Male	20705	66.9	Female	10235	33.1
native-country	30940	1.8	42	United-States	27719	89.6	Mexico	607	2

Missing values :

There's no need to worry about that because the rate of missing value is never above 60%

Outliers :

- Capital gain : it seems a little bit weird to only have 9 we should check that
- Hours per week : it also seems weird to have 9

Cardinality

- We have kept the missing values into it, but since we also have the missing values on our figures, it is not an issue.

Figure issues

- Capital loss & Capital Gain there is a lot of 0, maybe we should convert it into a categorical features

Features	Data quality issue	Potential Handling Strategies
Capital gain	Outliers (high)	
Hours per week	Outliers (high)	