Dựa vào dữ liệu ‘data_hw08.csv’ thực hiện các bước EDA như sau

Question 1: Liệt kê các feature nào là feature categorical? Feature nào là feature numerical

Kết quả mong đợi: categorical features là những features sau

workclass	education	marital_status	occupation	relationship	race	sex	native_country	income_bracket

Numerical features là những features sau

age	functional_weight	education_num	capital_gain	capital_loss	hours_per_week

Question 2: Đối với mỗi feature cần cho biết pandas_dtype, python_type, số record bị trống và tỷ lệ phần trăm các dòng trống.

Kết quả mong đợi như hình sau

	Pandas_Dtype	python_type	Missing_Value	% Missing_Values
age	int64	int	0	0
workclass	object	str	0	0
functional_weight	int64	int	0	0
education	object	str	0	0
education_num	int64	int	0	0
marital_status	object	str	0	0
occupation	object	str	0	0
relationship	object	str	0	0
race	object	str	0	0
sex	object	str	0	0
capital_gain	int64	int	0	0
capital_loss	int64	int	0	0
hours_per_week	int64	int	0	0
native_country	object	str	0	0
income_bracket	object	str	0	0

Question 3: đối với features dạng numerical, tính toán các đặc trưng thống kê như hình (lưu ý nếu chưa biết các định nghĩa này cần tìm hiểu thêm trên internet), cụ thể cần tính các thông số sau cho mỗi feature dạng numerical: max, range, IQR, mode, mad, kurtosis, skewness, mean, std, min, 25%, 50%, 75%.

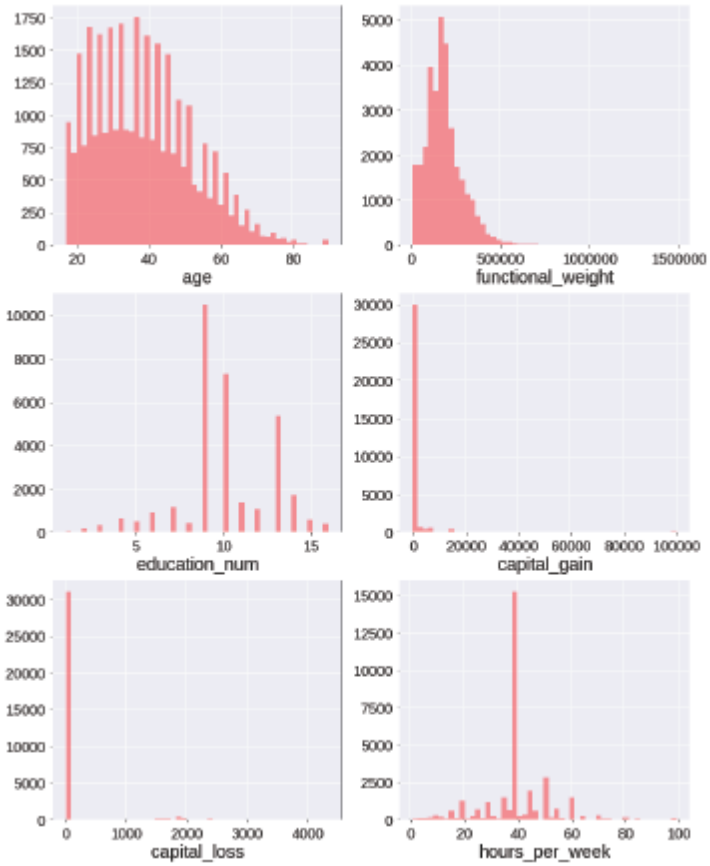
Kết quả mong đợi

	max	range	IQR	mode	mad	kurtosis	skewness
age	90	73	20	36	11.189182	-0.166127	0.558743
functional_weight	1484705	1472420	119224	123011	77608.21854	6.218811	1.44698
education_num	16	15	3	9	1.903048	0.623444	-0.311676
capital_gain	99999	99999	0	0	1977.373437	154.799438	11.953848
capital_loss	4356	4356	0	0	166.462055	20.376802	4.594629
hours_per_week	99	98	5	40	7.583228	2.916687	0.227643

	mean	std	min	25%	50%	75%
age	38.581647	13.640433	17	28	37	48
functional_weight	189778.3665	105549.9777	12285	117827	178356	237051
education_num	10.080679	2.57272	1	9	10	12
capital_gain	1077.648844	7385.292085	0	0	0	0
capital_loss	87.30383	402.960219	0	0	0	0
hours_per_week	40.437456	12.347429	1	40	40	45

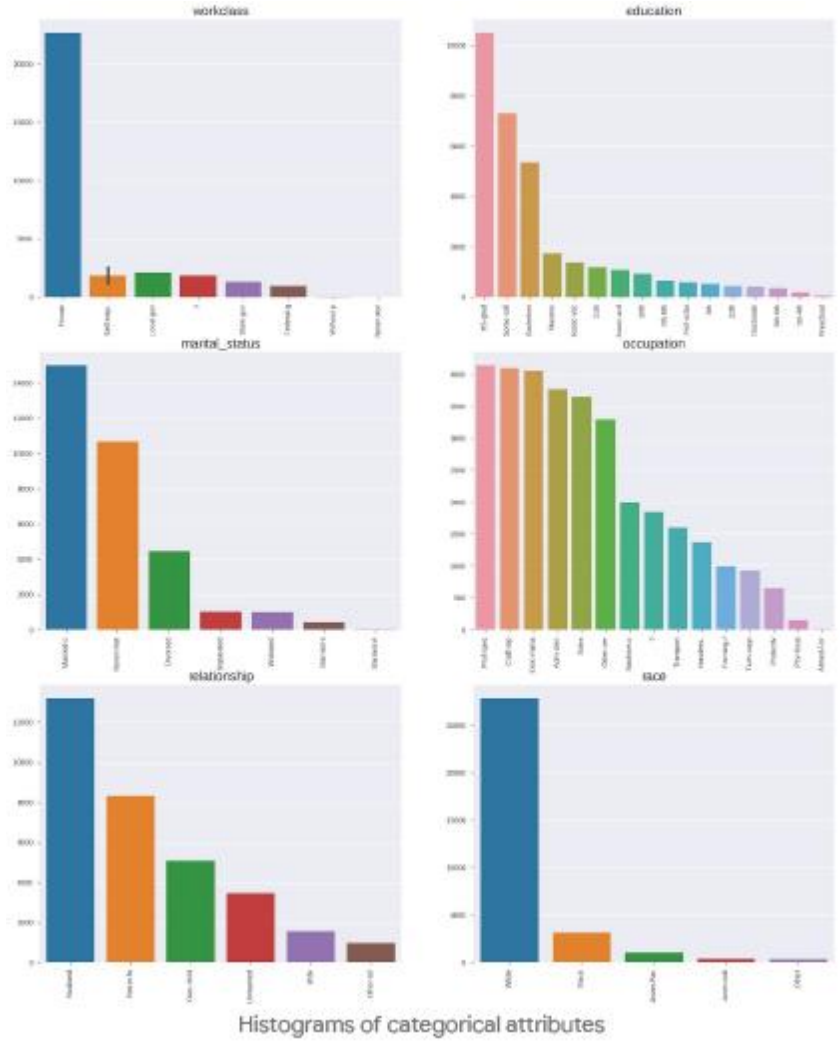
Question 4: Khảo sát histogram của các feature numerical

Kết quả mong đợi

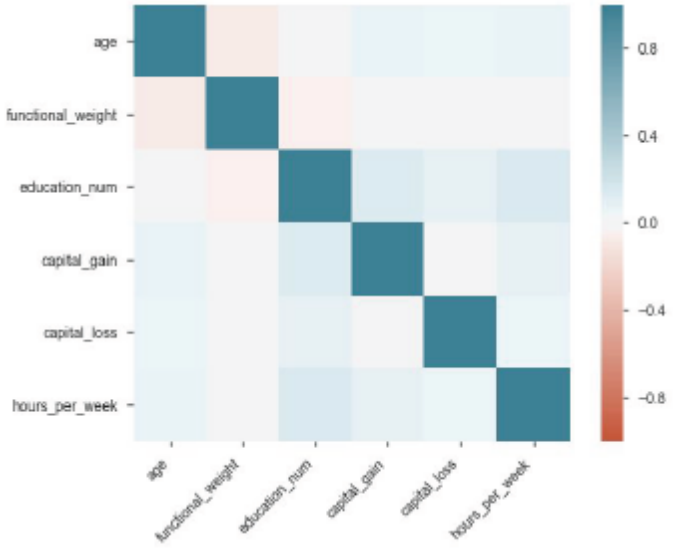


Question 5: Vẽ histogram (countplot) cho feature dạng categorical.

Kết quả mong đợi



Question 6: vẽ heatmap so sánh mối tương quan giữa các numerical feature. Lưu ý heatmap chỉ thực hiện trên các numerical features.



	age	functional_ weight	education_ num	capital_ gain	capital_ loss	hours_ per_week
age	1	-0.076646	0.036527	0.077674	0.057775	0.068756
functional_weight	-0.076646	1	-0.043195	0.000432	-0.010252	-0.018768
education_num	0.036527	-0.043195	1	0.12263	0.079923	0.148123
capital_gain	0.077674	0.000432	0.12263	1	-0.031615	0.078409
capital_loss	0.057775	-0.010252	0.079923	-0.031615	1	0.054256
hours_per_week	0.068756	-0.018768	0.148123	0.078409	0.054256	1