Hackathon : Data Science & Analytics

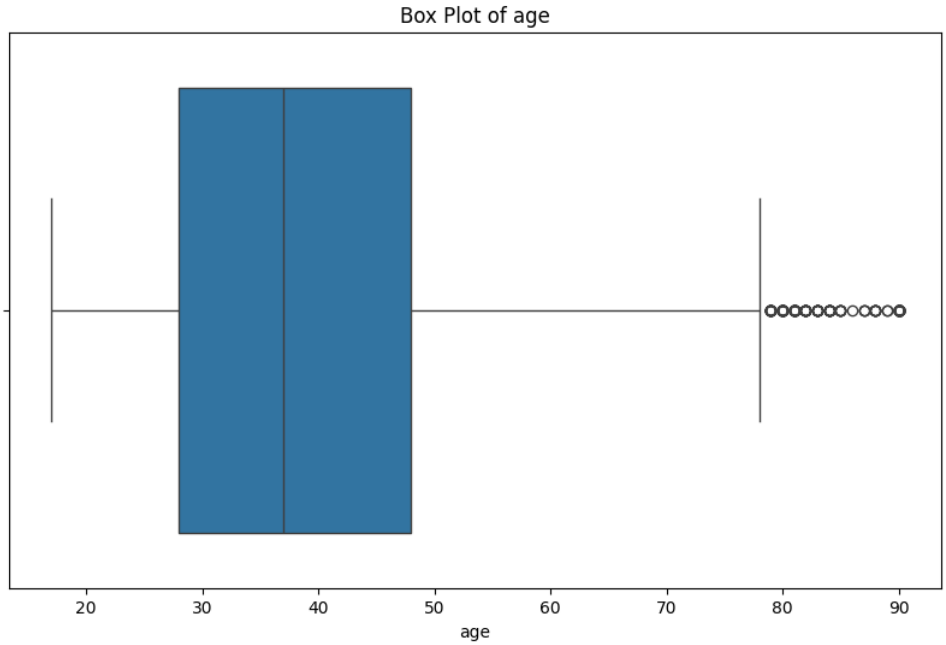
Challenge: From Data to Insights

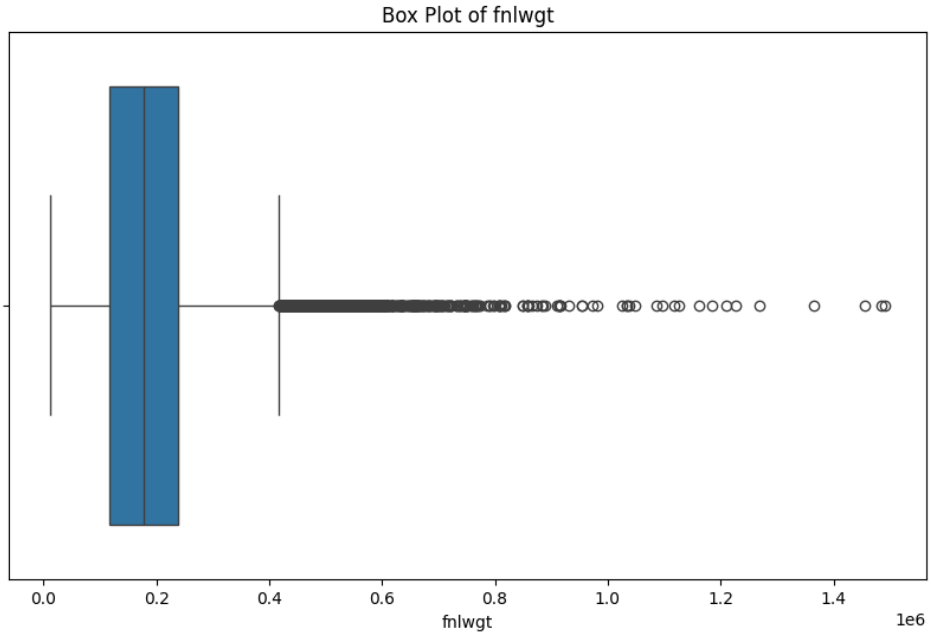
1. Automated Data Cleaning and Preprocessing Challenge

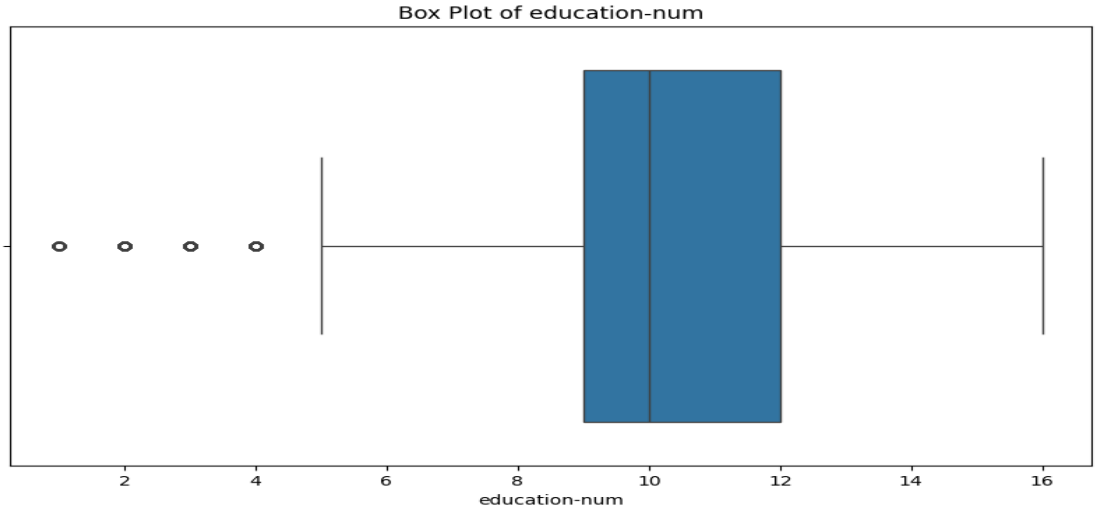
Dataset: https://archive.ics.uci.edu/dataset/2/adult

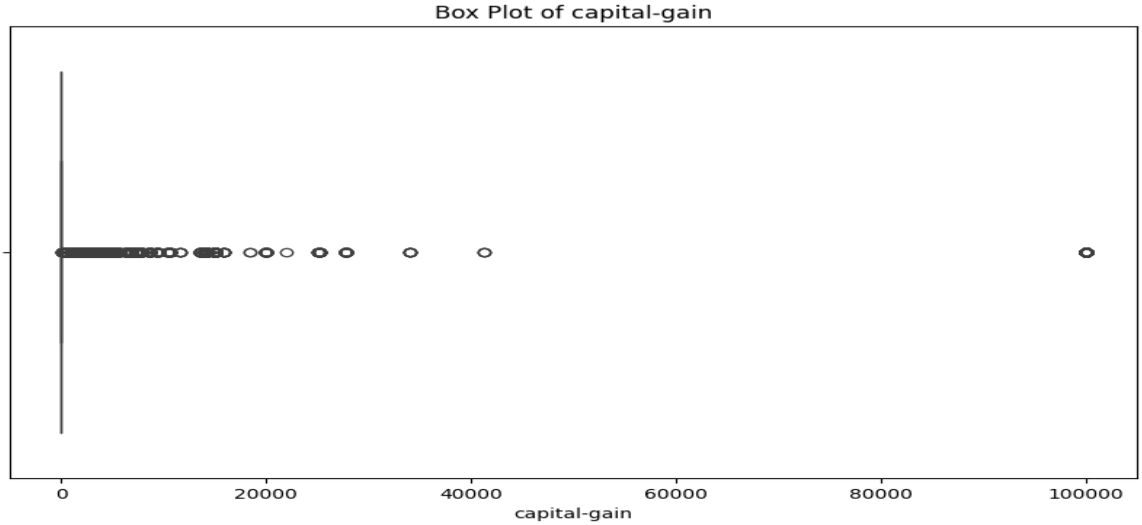
name	role	type	demographic	description	units	missing_values
age	Feature	Integer	Age	N/A		no
workclass	Feature	Categorical	Income	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked		yes
fnlwgt	Feature	Integer		None		no
education	Feature	Categorical	Education Level	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool		no
education-num	Feature	Integer	Education Level	None		no
marital-status	Feature	Categorical	Other	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse		no
occupation	Feature	Categorical	Other	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces		yes
relationship	Feature	Categorical	Other	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried		no
race	Feature	Categorical	Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black		no
sex	Feature	Binary	Sex	Female, Male.		no
capital-gain	Feature	Integer		None		no
capital-loss	Feature	Integer		None		no
hours-per-week	Feature	Integer		None		no
native-country	Feature	Categorical	Other	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands		yes
income	Target	Binary	Income	>50K, <=50K.		no

BOX PLOT OF DATA

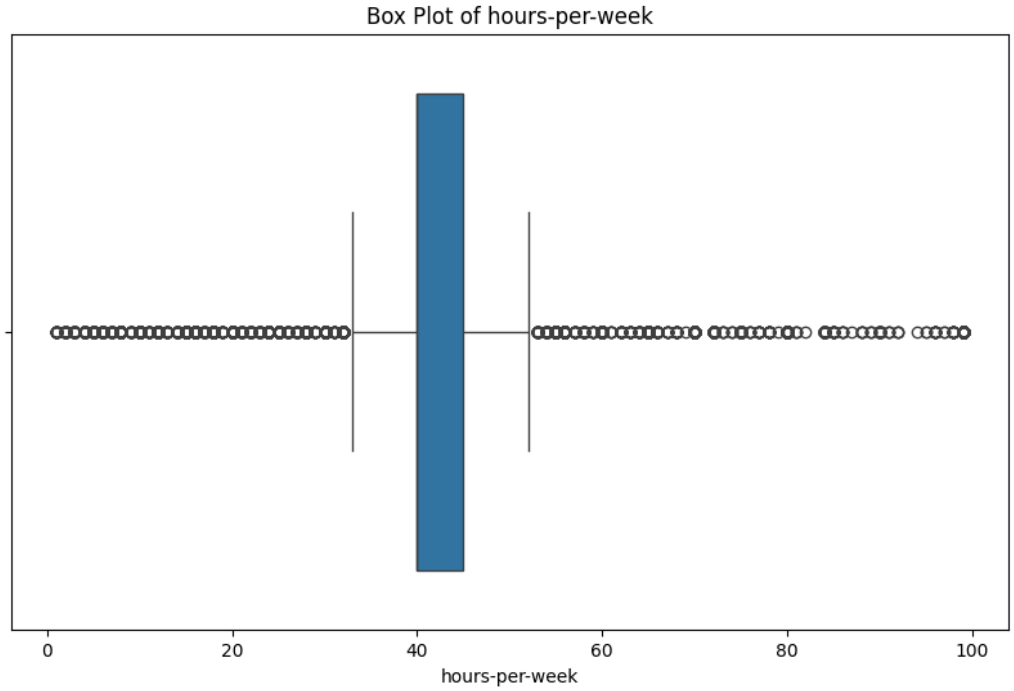




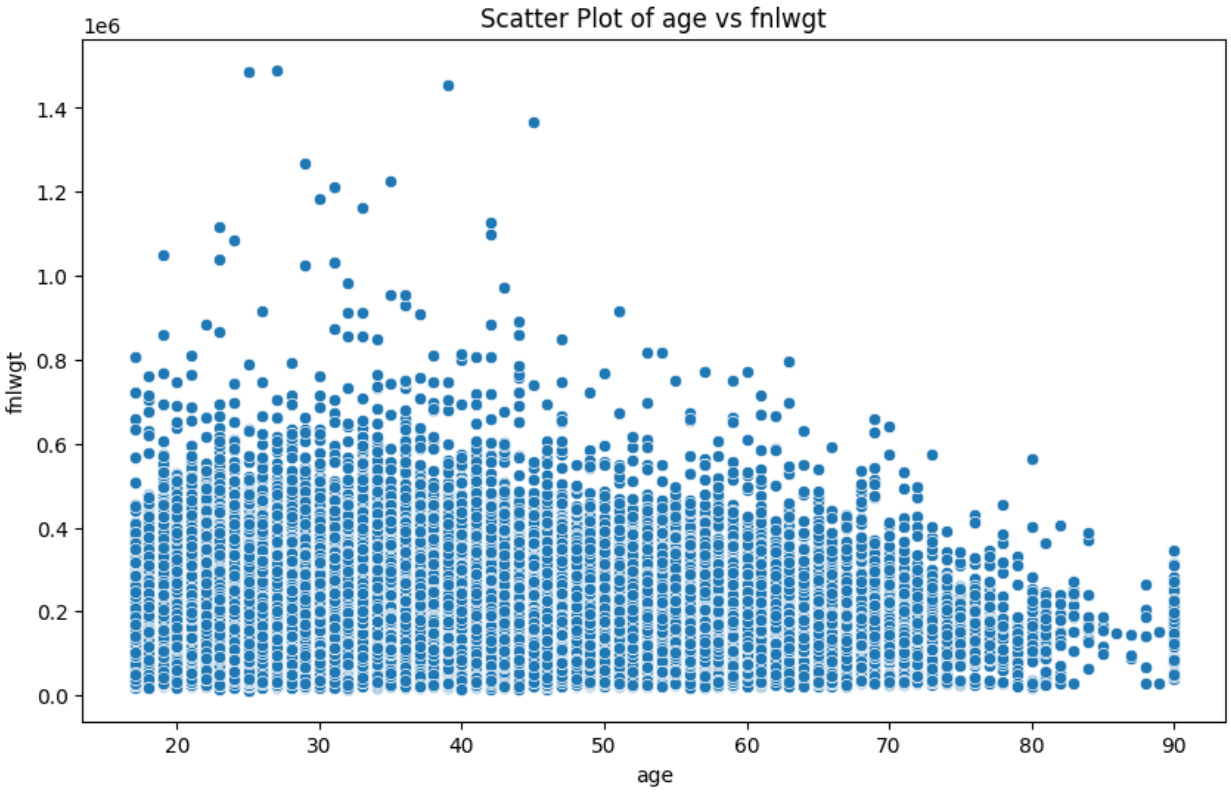

Box Plot of education-num
2
4
6
8
10
12
14
16
education-num

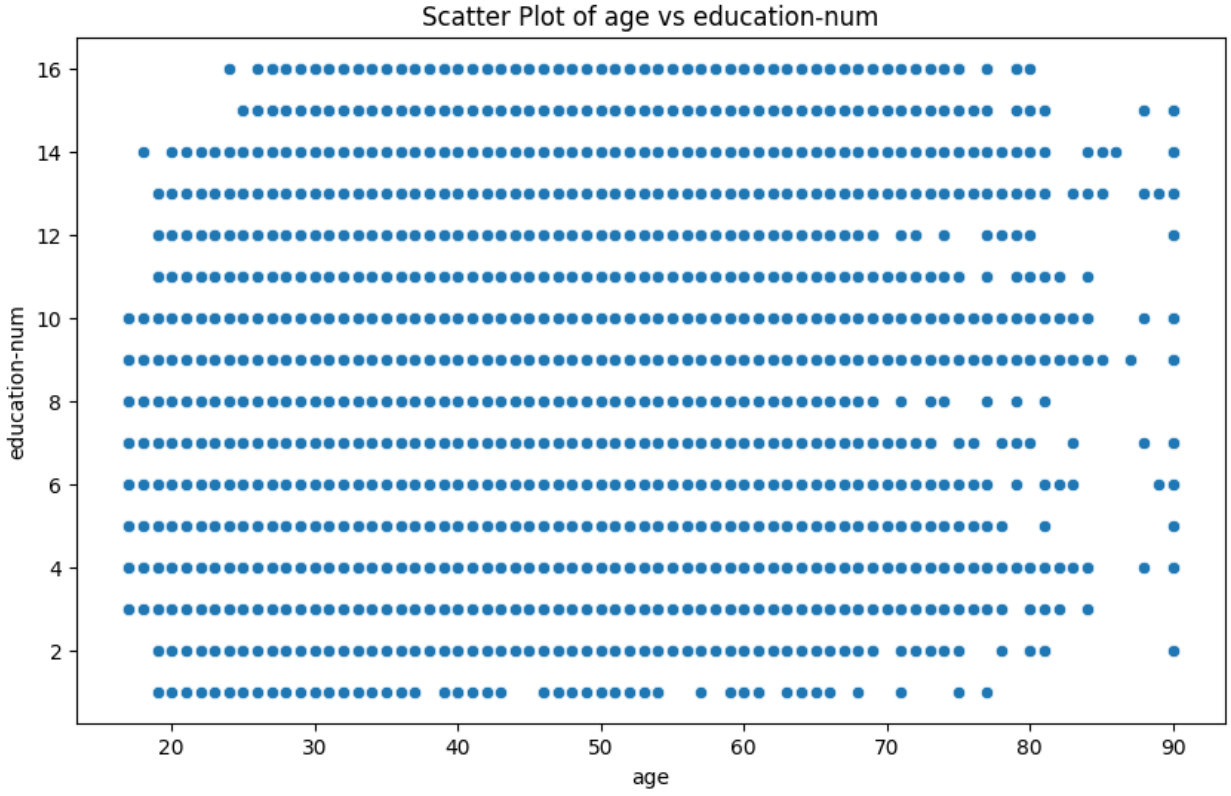

Box Plot of capital-gain
0
20000
40000
60000
80000
100000
capital-gain

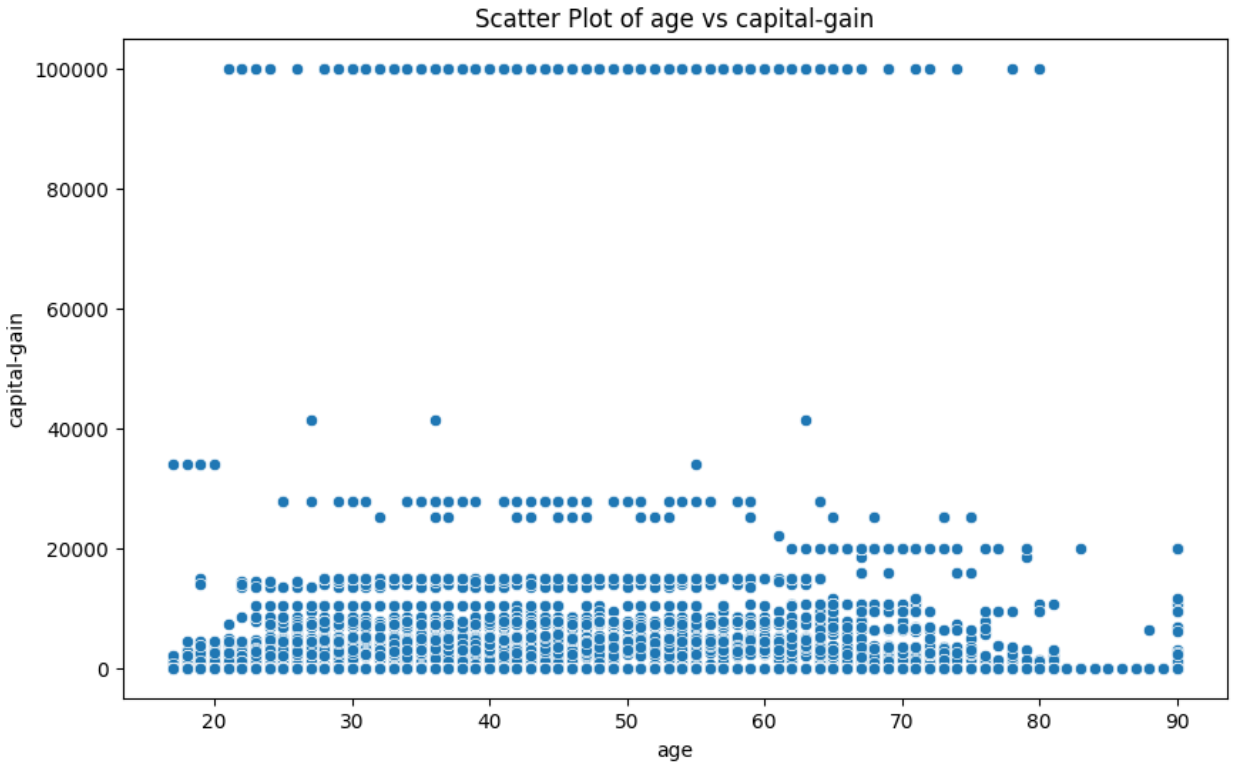

Box Plot of capital-loss
0
1000
2000
3000
4000
capital-loss

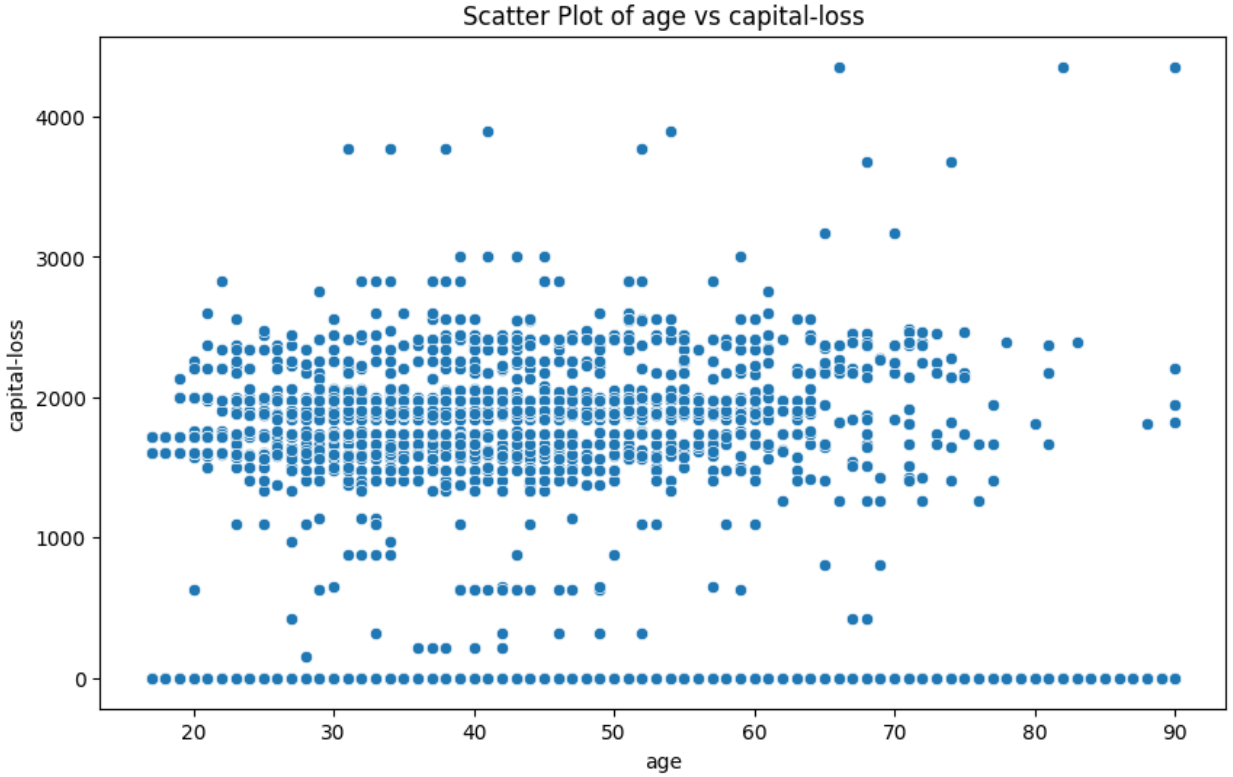

Box Plot of hours-per-week
0
20
40
60
80
100
hours-per-week

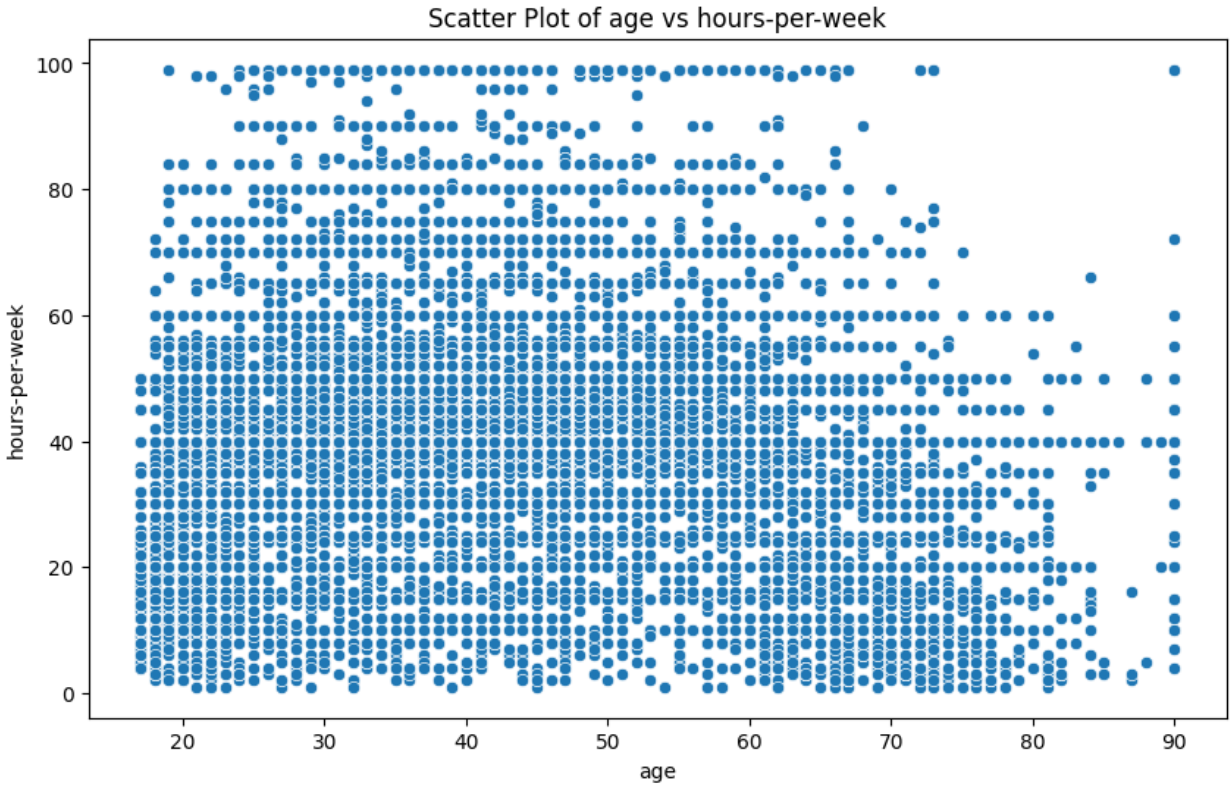
SCATTER PLOTS FOR PAIRWISE COMPARISON

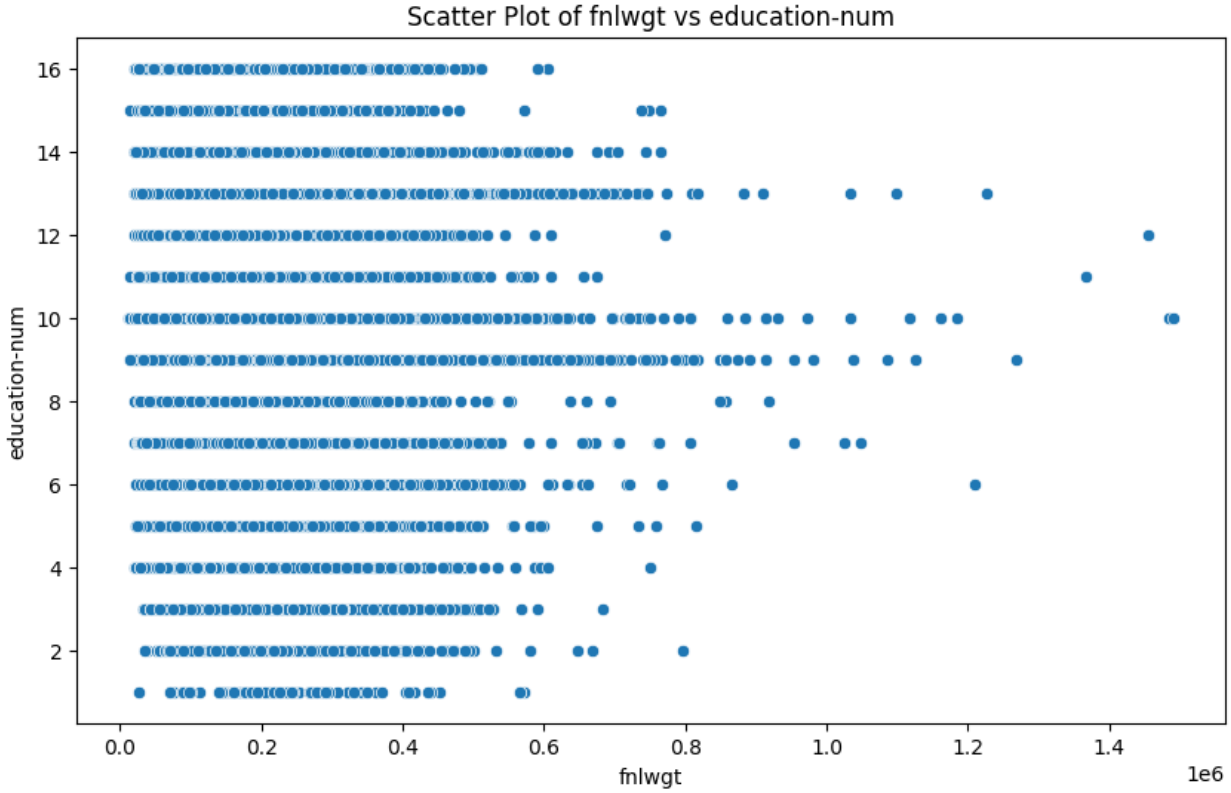


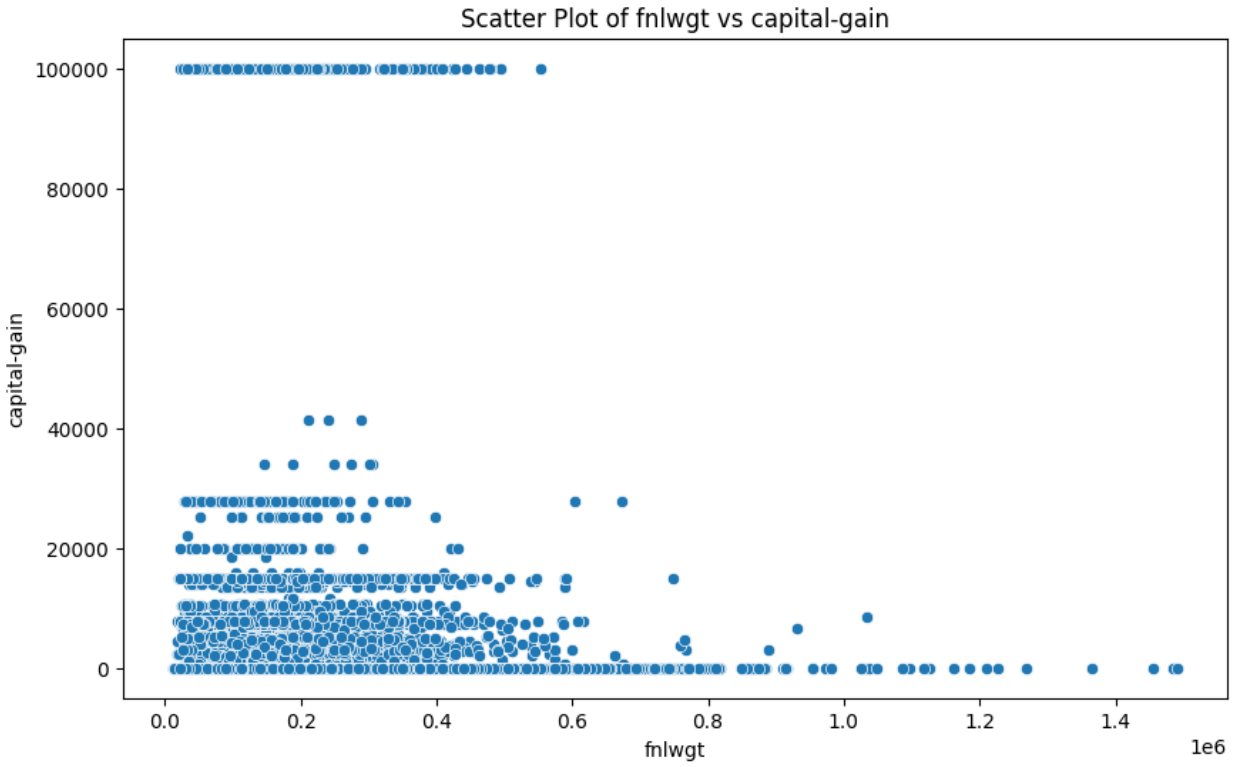


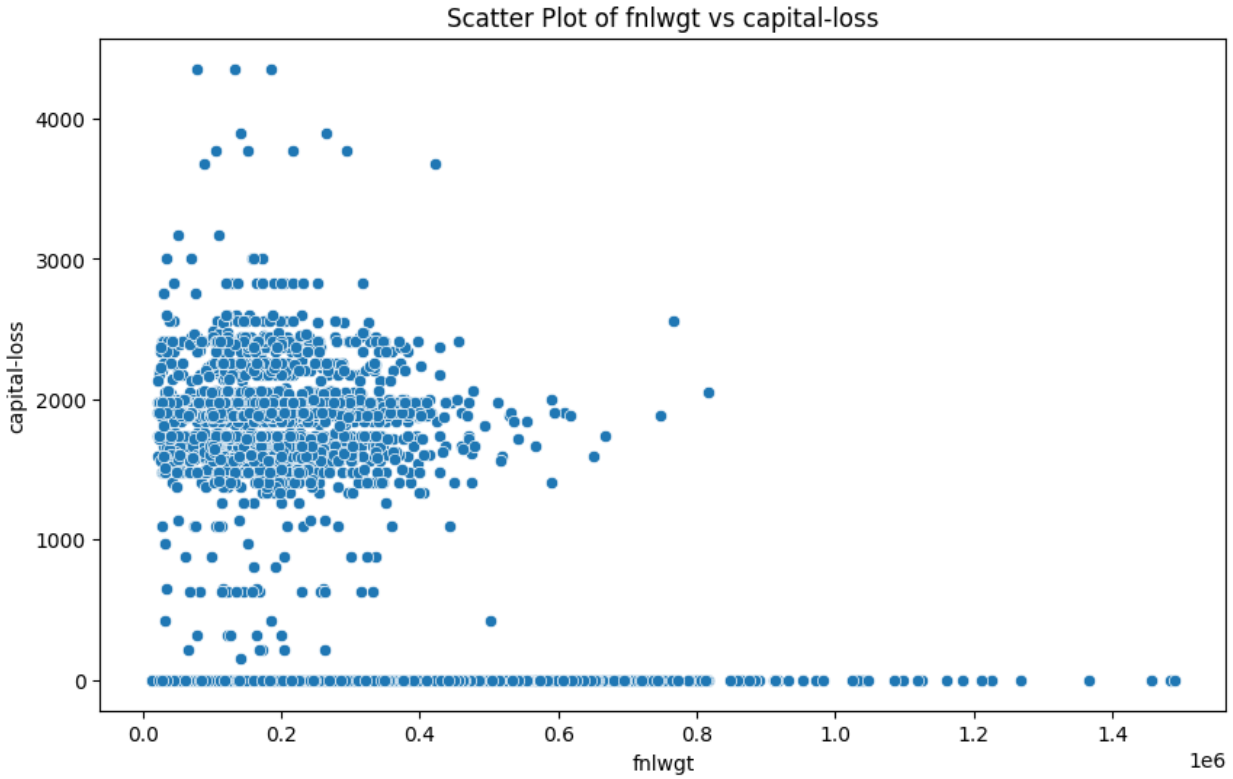

Scatter Plot of age vs capital-gain
capital-gain
age
100000
80000
60000
40000
20000
0
20
30
40
50
60
70
80
90

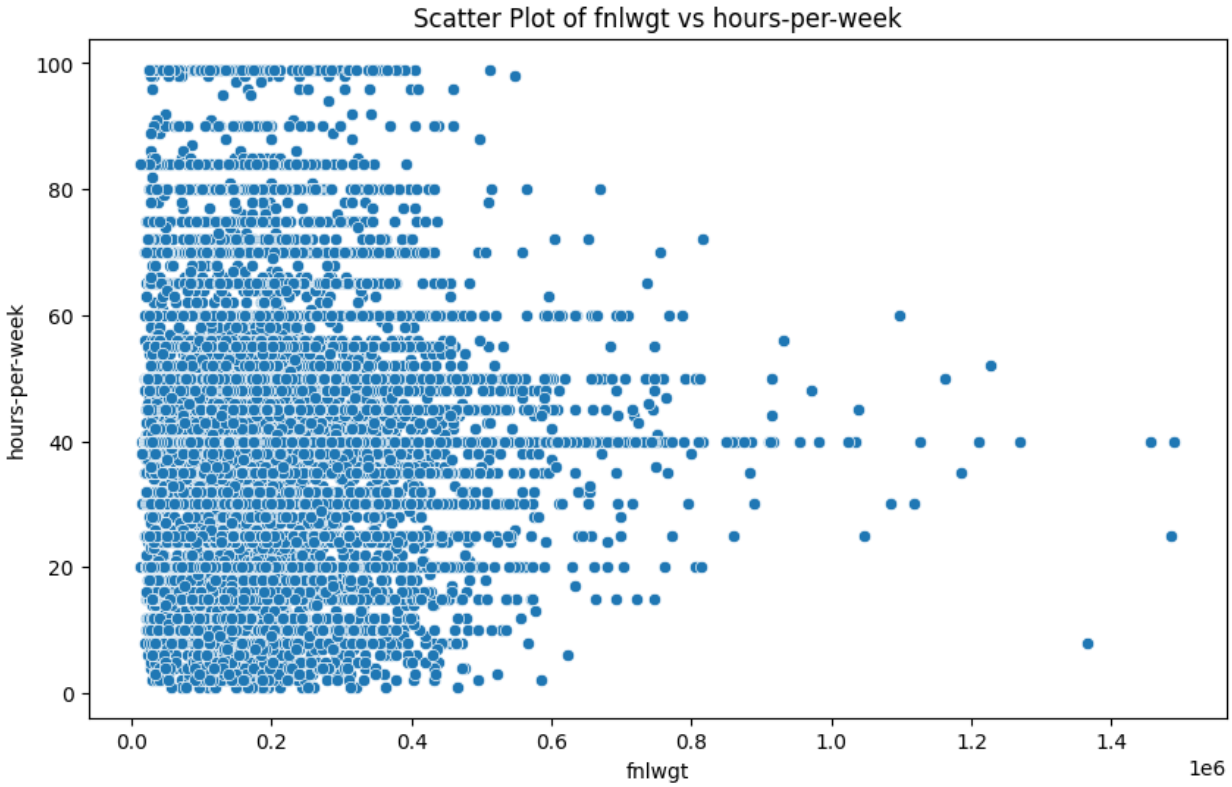

Scatter Plot of age vs capital-loss
capital-loss
age
4000
3000
2000
1000
0
20
30
40
50
60
70
80
90

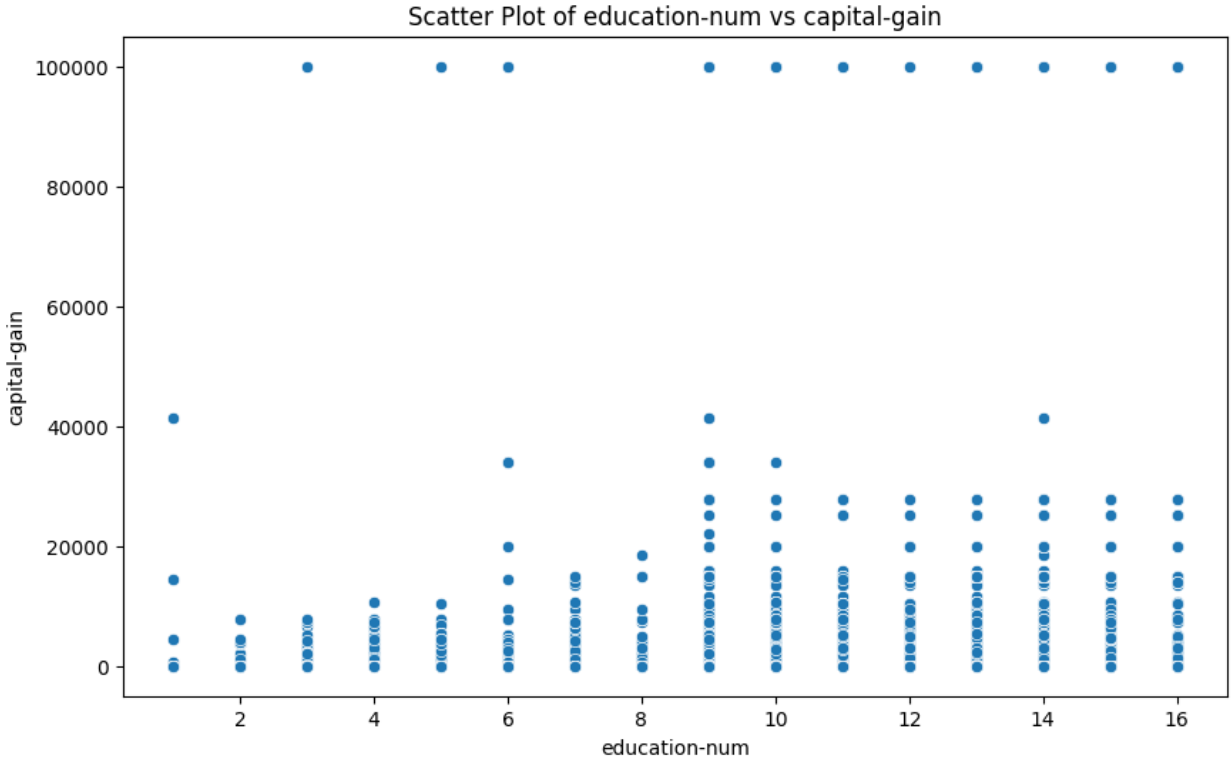

Scatter Plot of age vs hours-per-week
hours-per-week
age
0
20
40
60
80
100
20
30
40
50
60
70
80
90

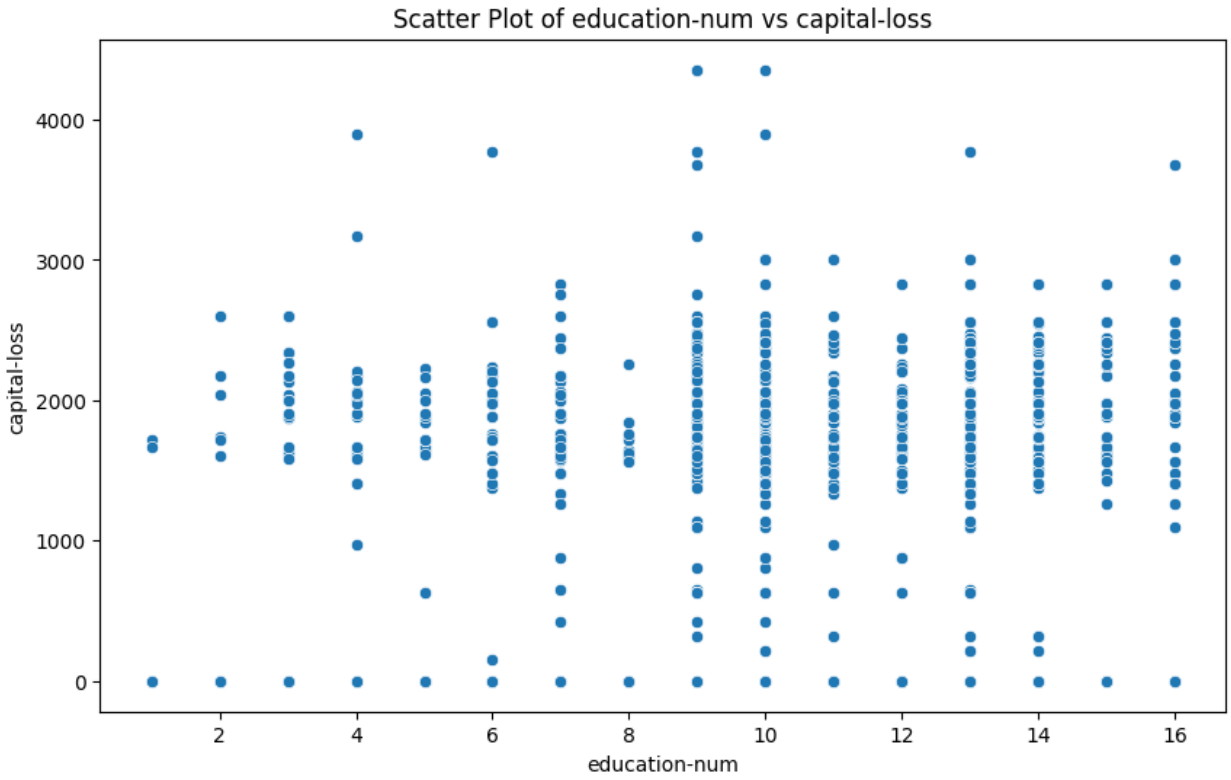

Scatter Plot of fnlwgt vs education-num
education-num
fnlwgt
1e6
2
4
6
8
10
12
14
16
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

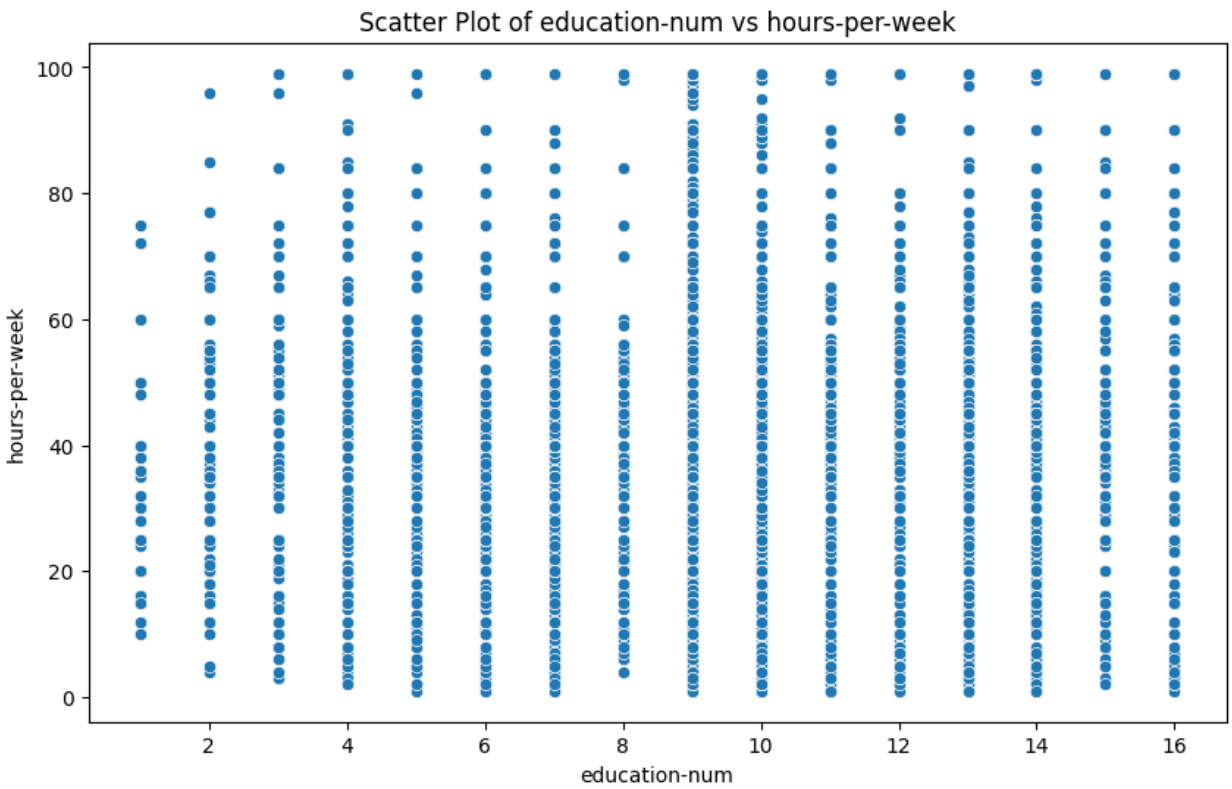

Scatter Plot of fnlwgt vs capital-gain
capital-gain
fnlwgt
1e6
0
20000
40000
60000
80000
100000
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

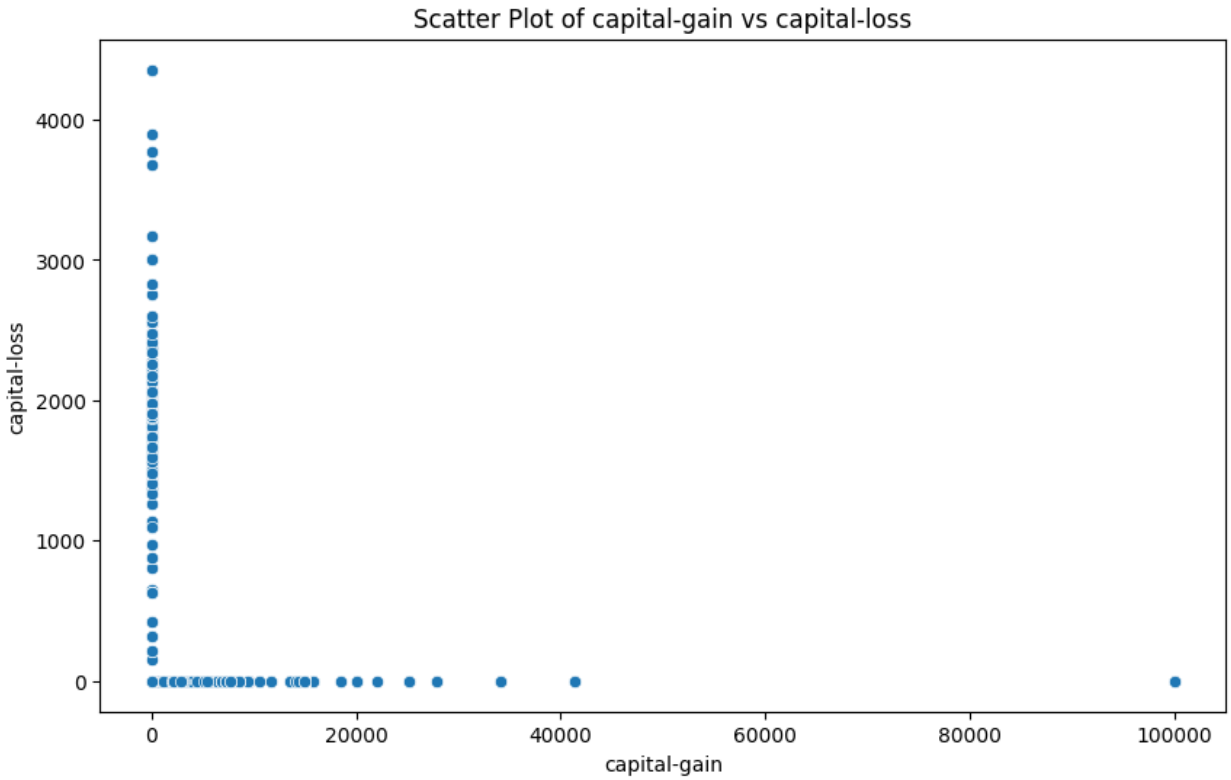

Scatter Plot of fnlwgt vs capital-loss
capital-loss
fnlwgt
1e6
0
1000
2000
3000
4000
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

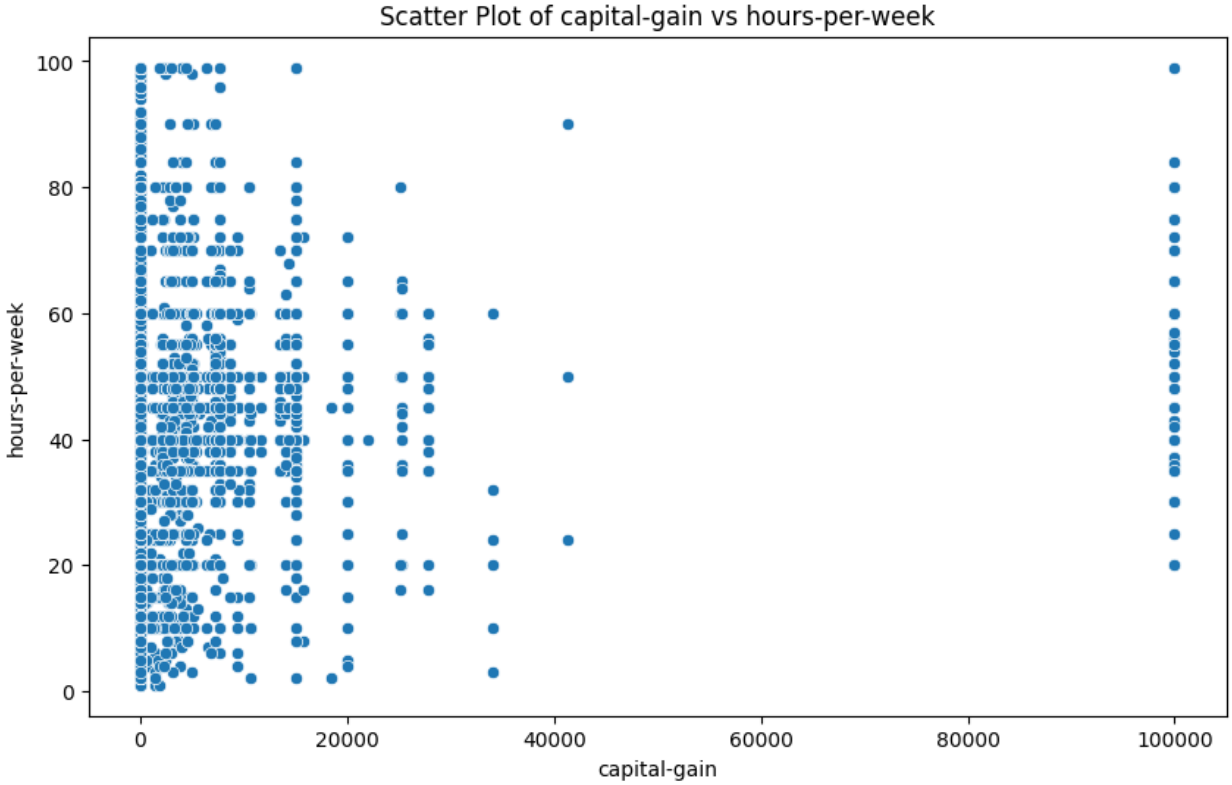

Scatter Plot of fnlwgt vs hours-per-week
hours-per-week
fnlwgt
1e6
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

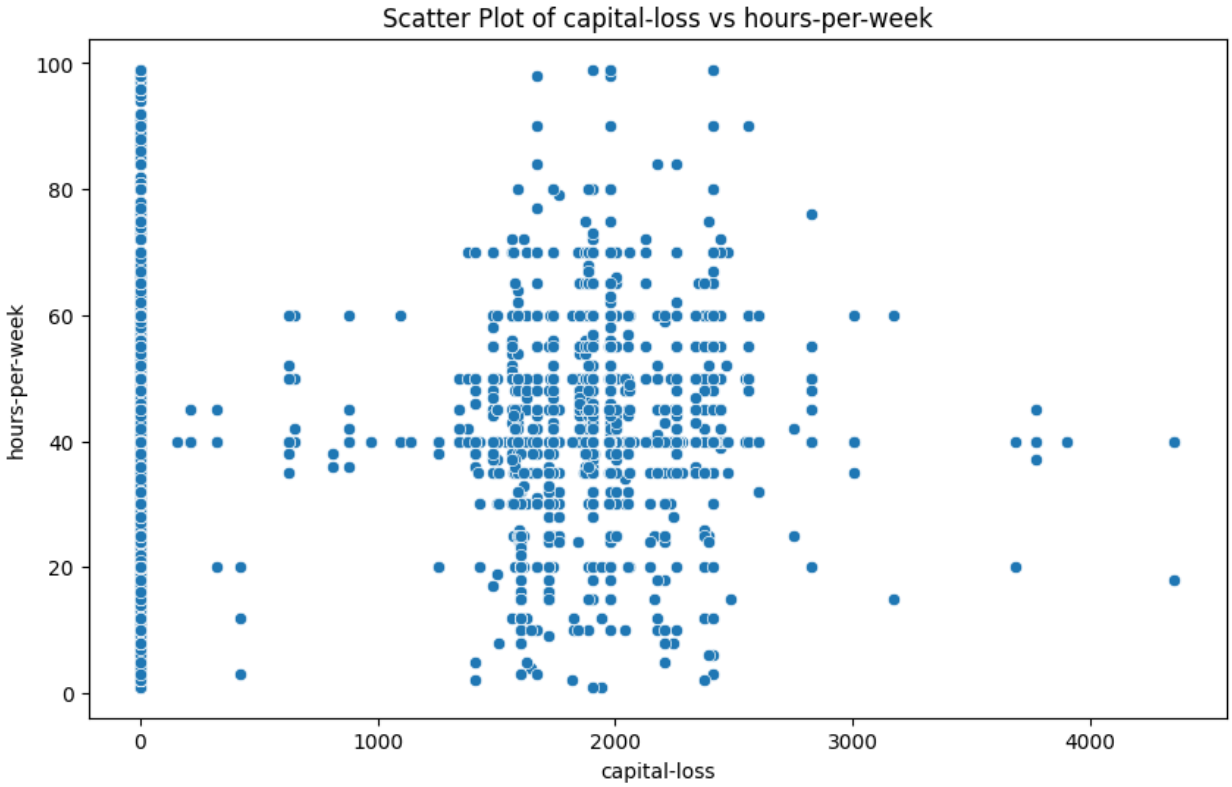

Scatter Plot of education-num vs capital-gain
capital-gain
education-num
0
20000
40000
60000
80000
100000
2
4
6
8
10
12
14
16


Scatter Plot of education-num vs capital-loss
capital-loss
education-num
0
1000
2000
3000
4000
2
4
6
8
10
12
14
16

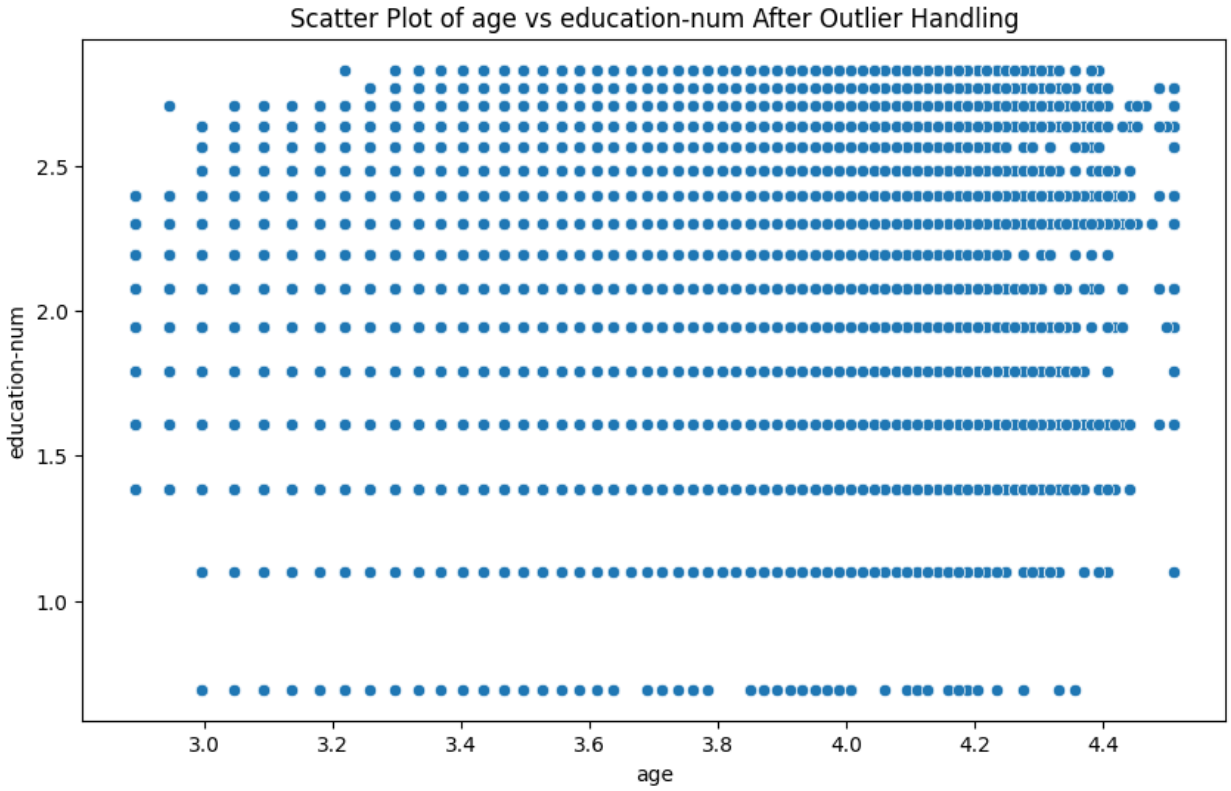

Scatter Plot of education-num vs hours-per-week
hours-per-week
education-num
0
20
40
60
80
100
2
4
6
8
10
12
14
16

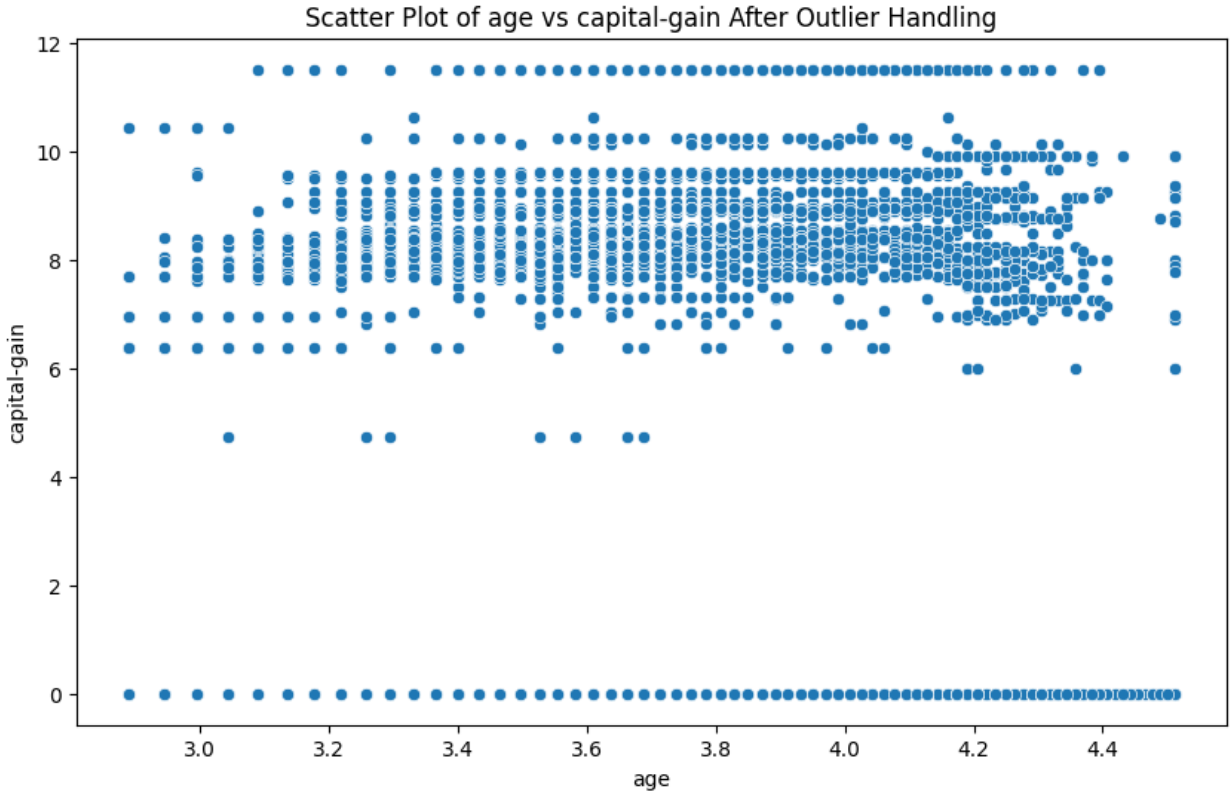

Scatter Plot of capital-gain vs capital-loss
capital-loss
capital-gain
0
1000
2000
3000
4000
0
20000
40000
60000
80000
100000

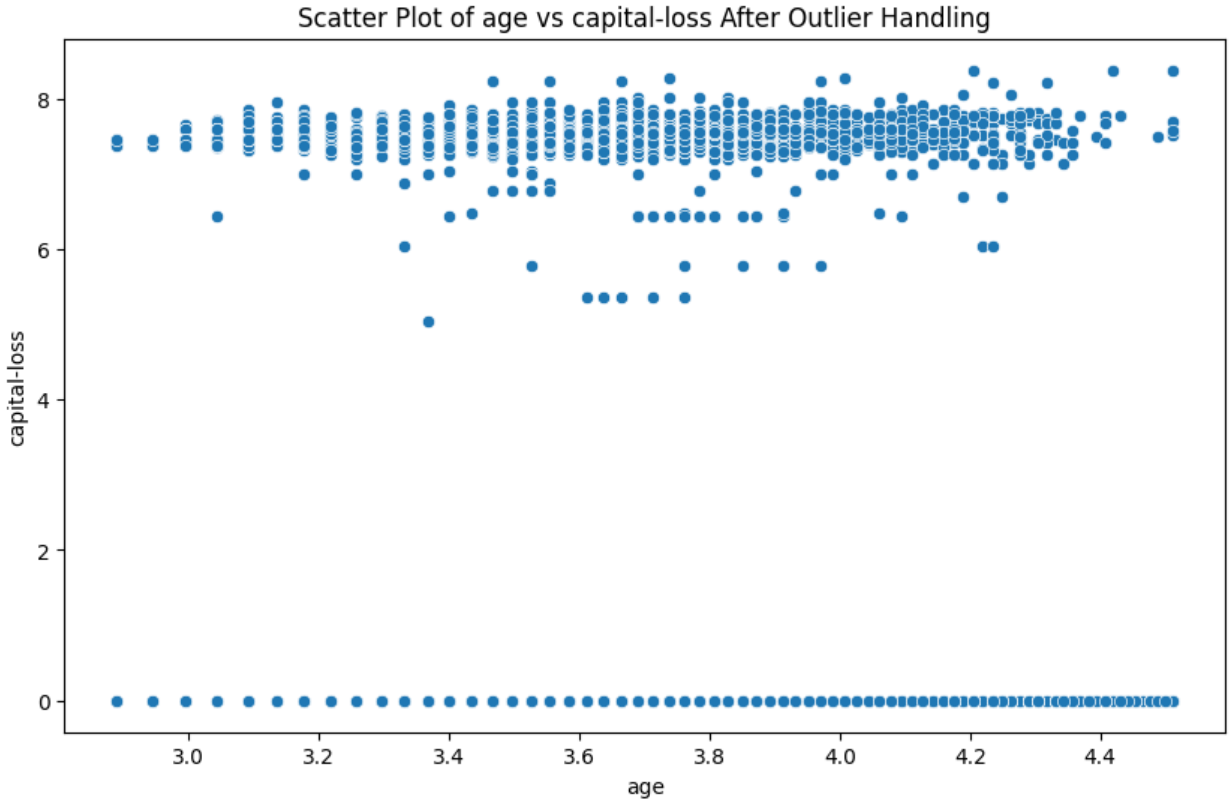

Scatter Plot of capital-gain vs hours-per-week
hours-per-week
capital-gain
0
20
40
60
80
100
0
20000
40000
60000
80000
100000

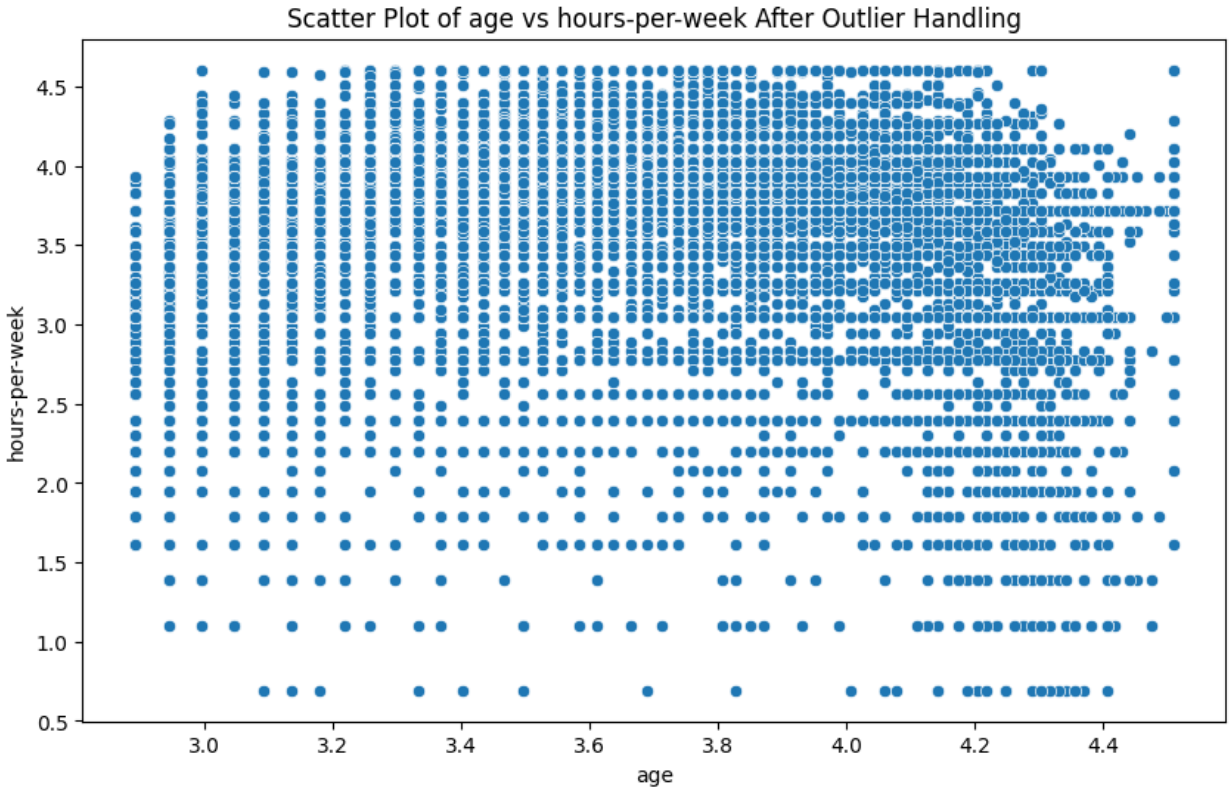

Scatter Plot of capital-loss vs hours-per-week
hours-per-week
capital-loss
0
20
40
60
80
100
0
1000
2000
3000
4000

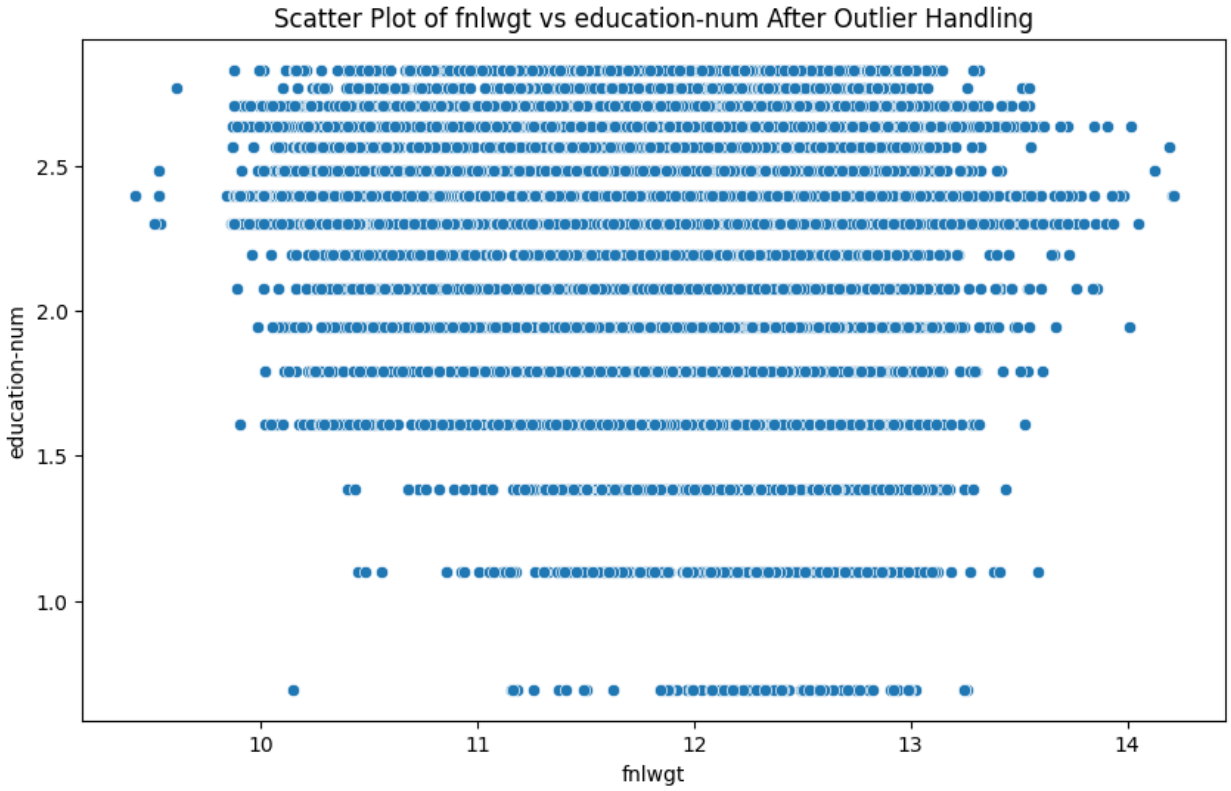

Scatter Plot of age vs fnlwgt After Outlier Handling
fnlwgt
age
10
11
12
13
14
3.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4

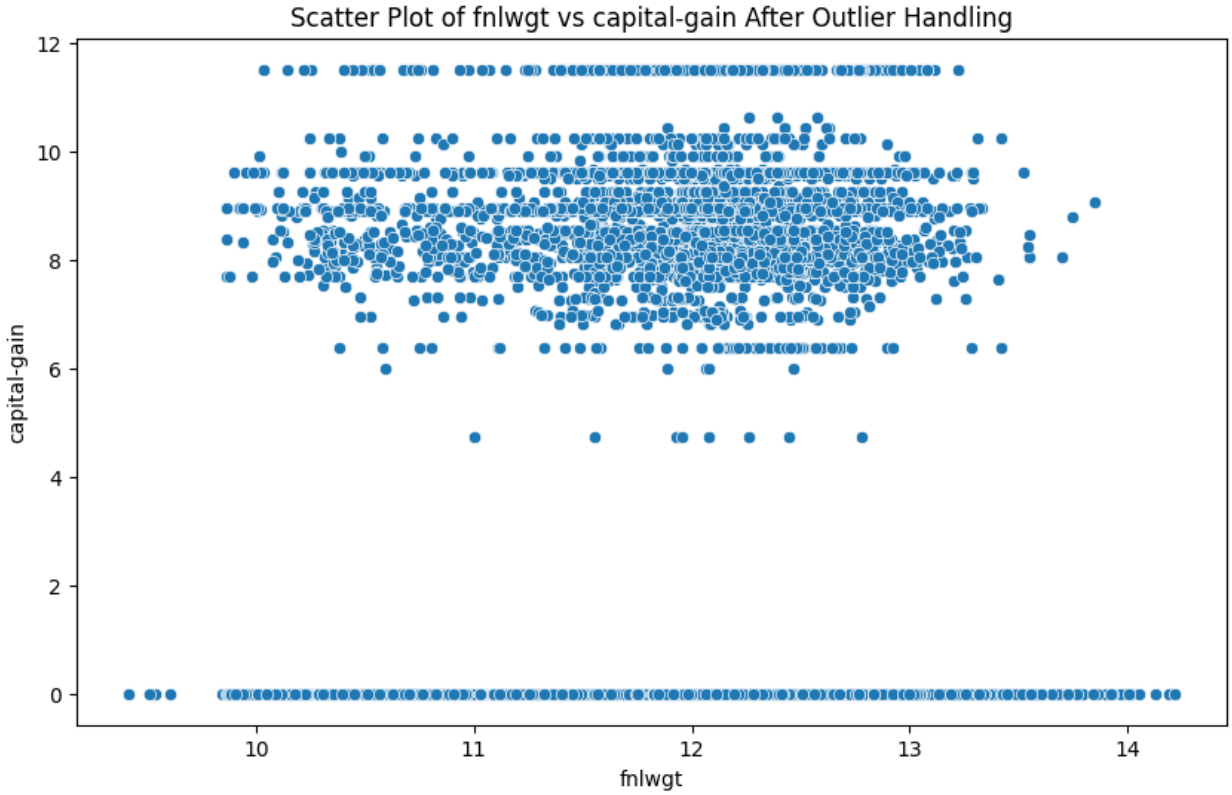

Scatter Plot of age vs education-num After Outlier Handling
education-num
2.5
2.0
1.5
1.0
3.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4
age

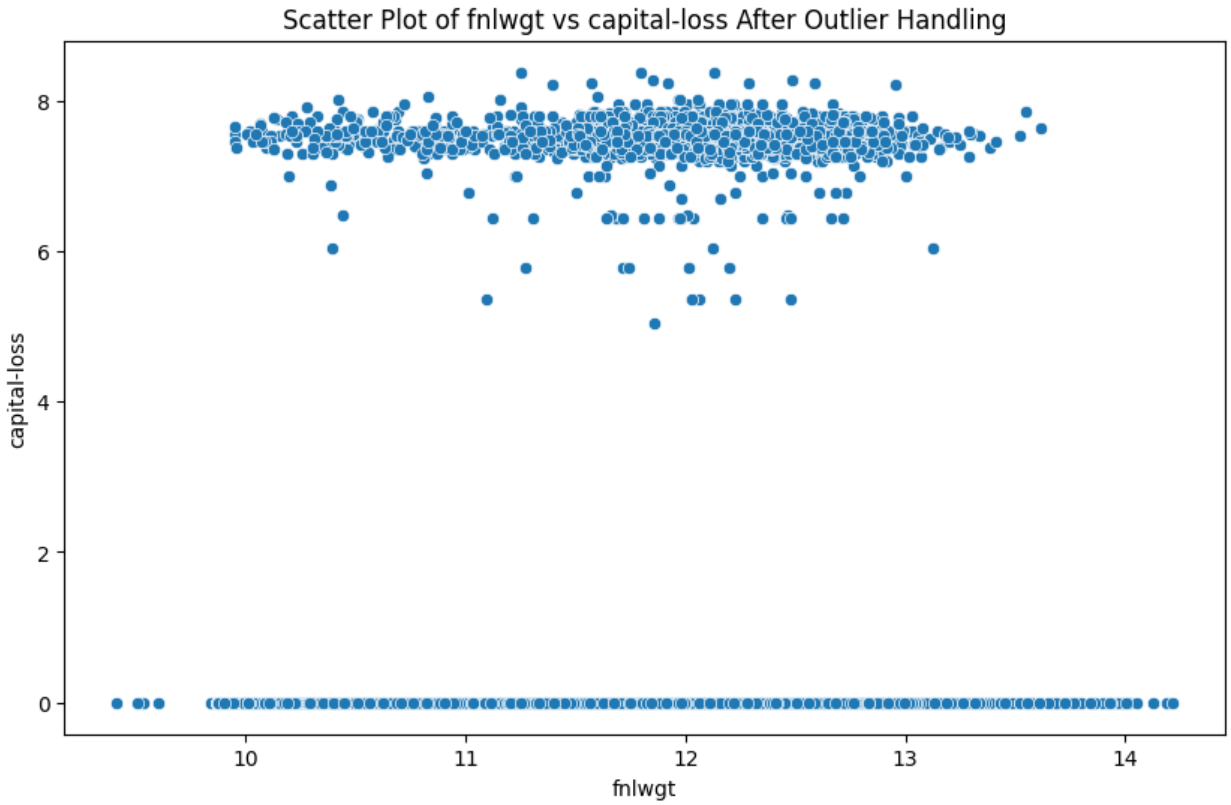

Scatter Plot of age vs capital-gain After Outlier Handling
capital-gain
12
10
8
6
4
2
0
3.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4
age

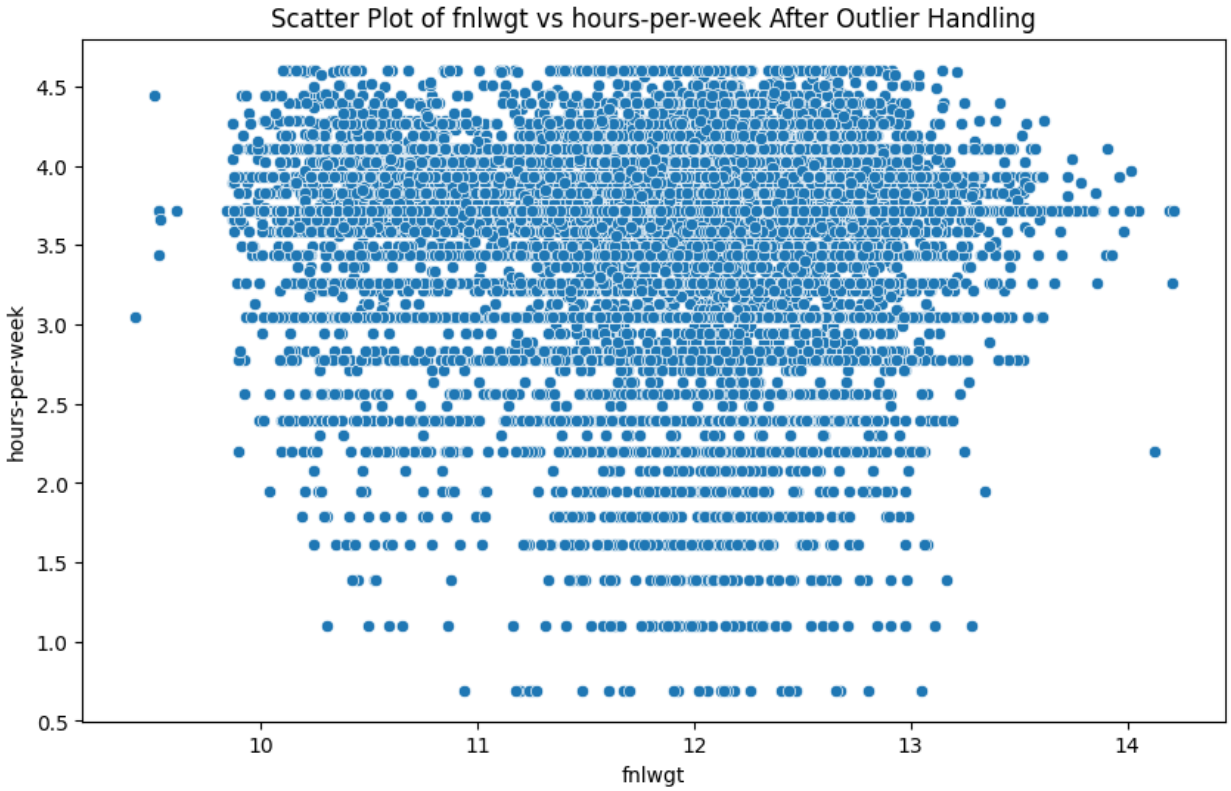

Scatter Plot of age vs capital-loss After Outlier Handling
capital-loss
age
0
2
4
6
8
3.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4

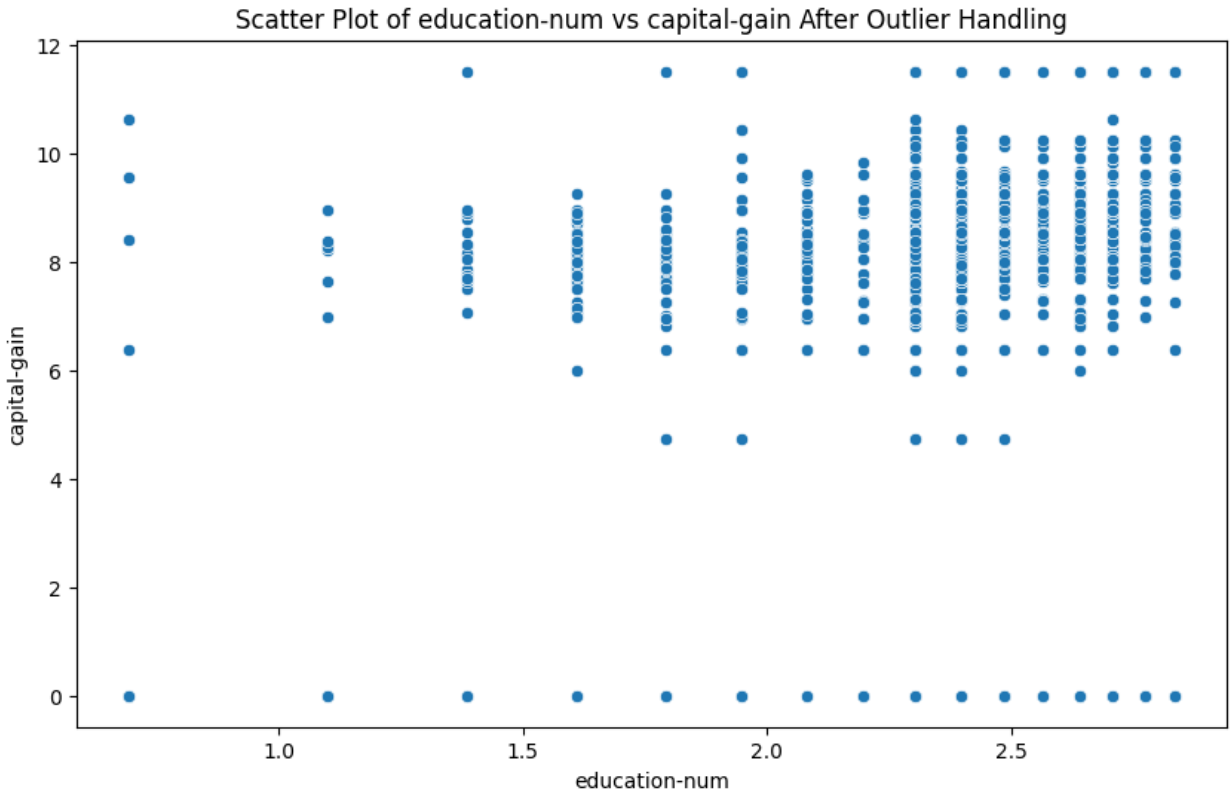

Scatter Plot of age vs hours-per-week After Outlier Handling
hours-per-week
age
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
3.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4

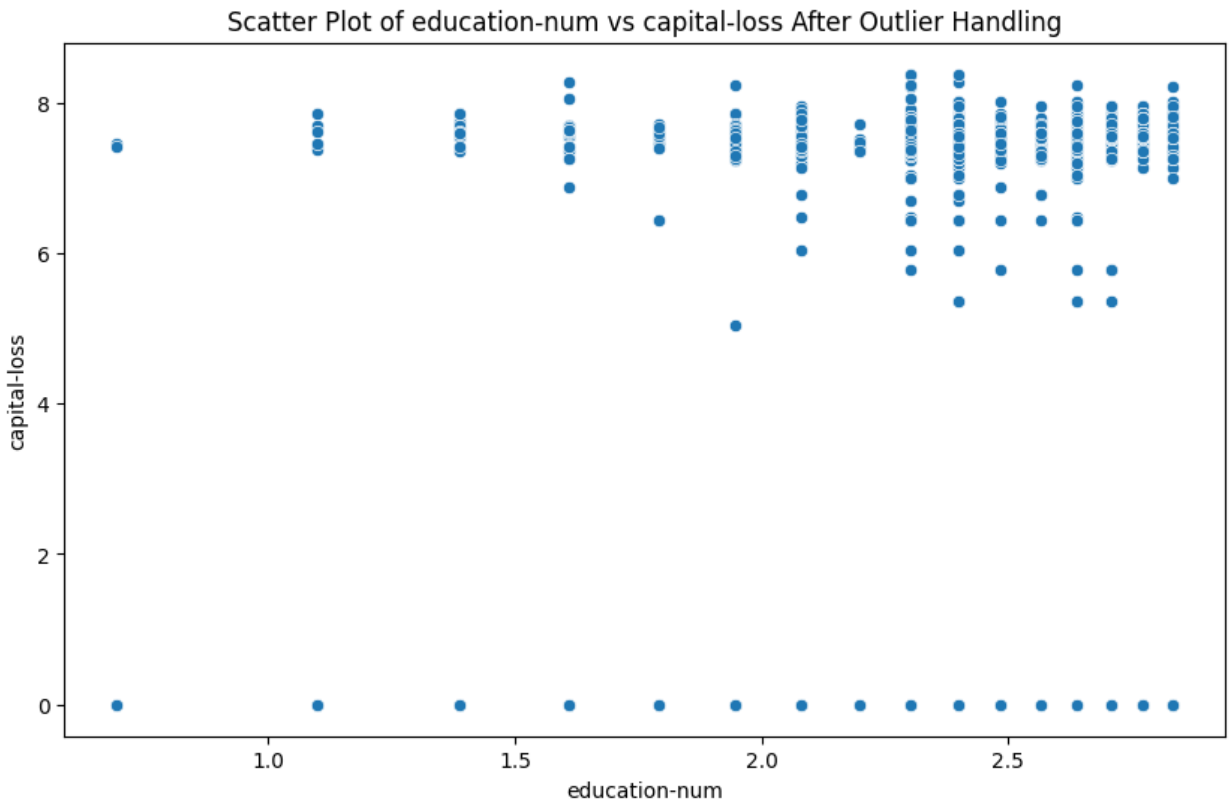

Scatter Plot of fnlwgt vs education-num After Outlier Handling
education-num
fnlwgt
2.5
2.0
1.5
1.0
10
11
12
13
14

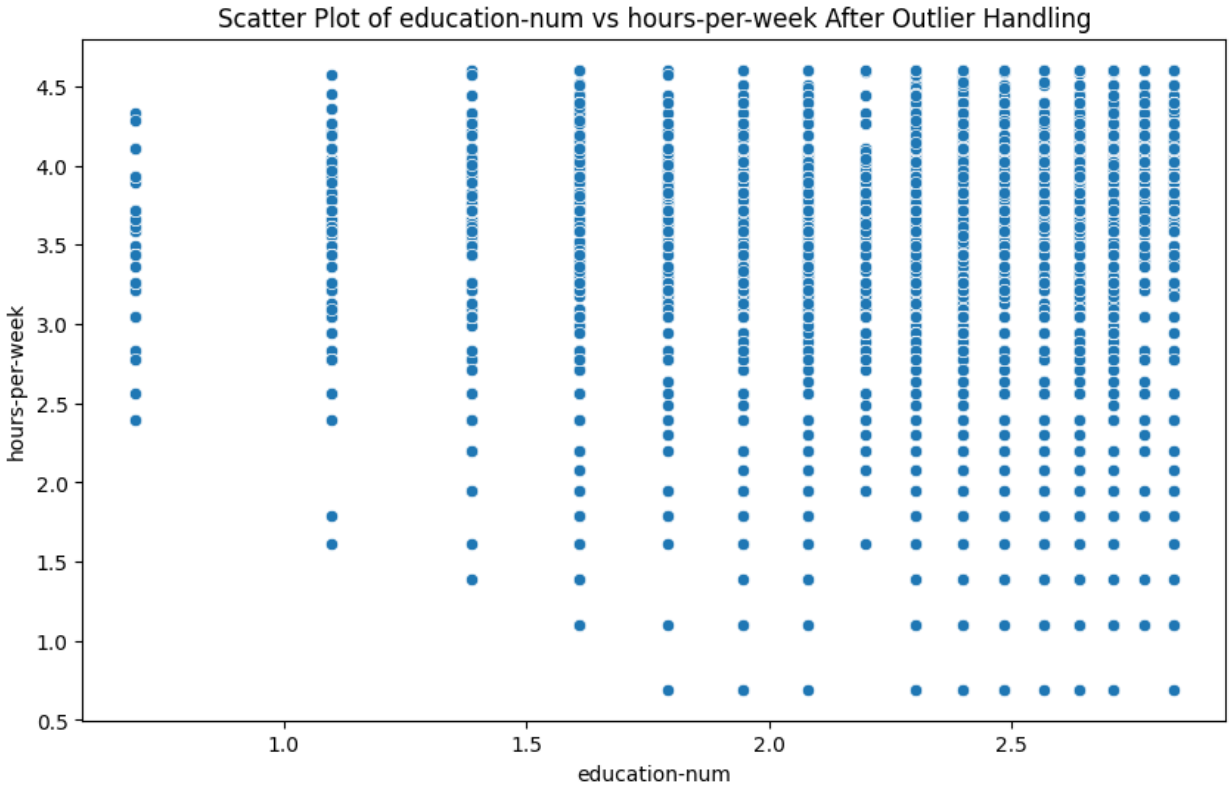

Scatter Plot of fnlwgt vs capital-gain After Outlier Handling
capital-gain
fnlwgt
12
10
8
6
4
2
0
10
11
12
13
14

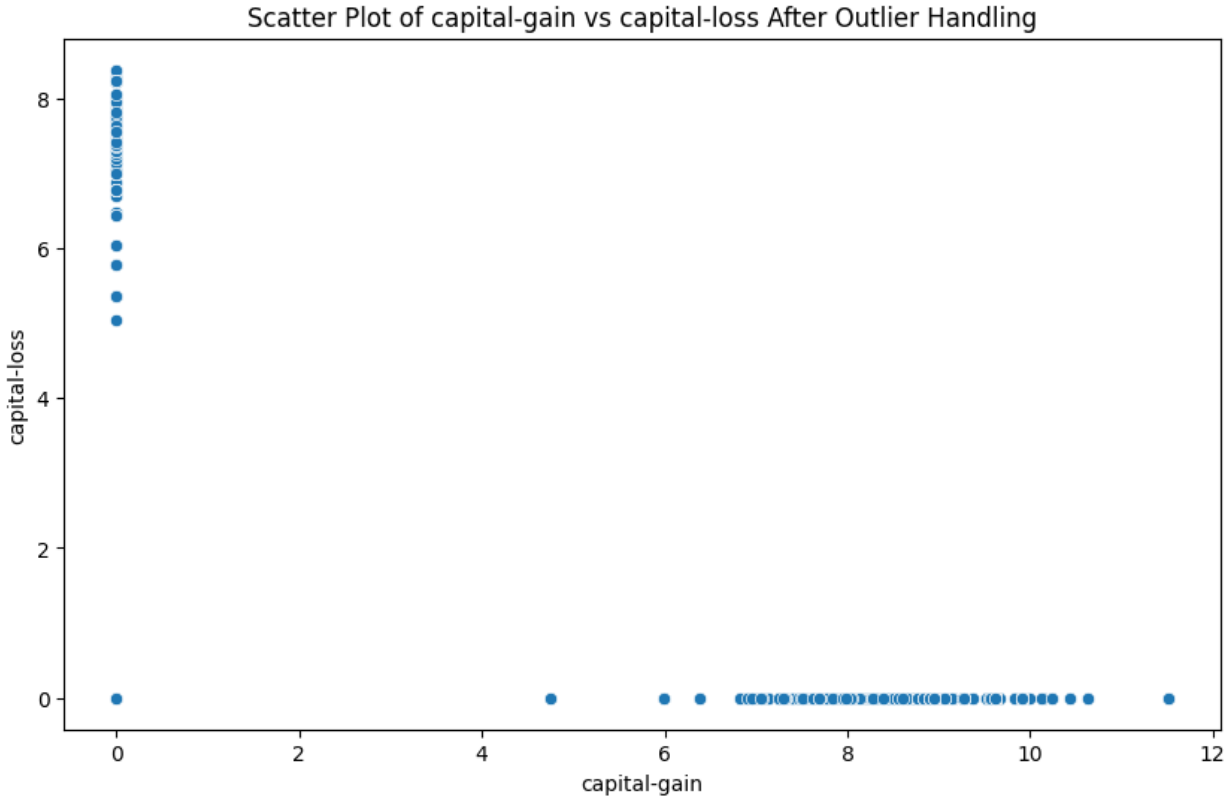

Scatter Plot of fnlwgt vs capital-loss After Outlier Handling
capital-loss
fnlwgt
0
2
4
6
8
10
11
12
13
14

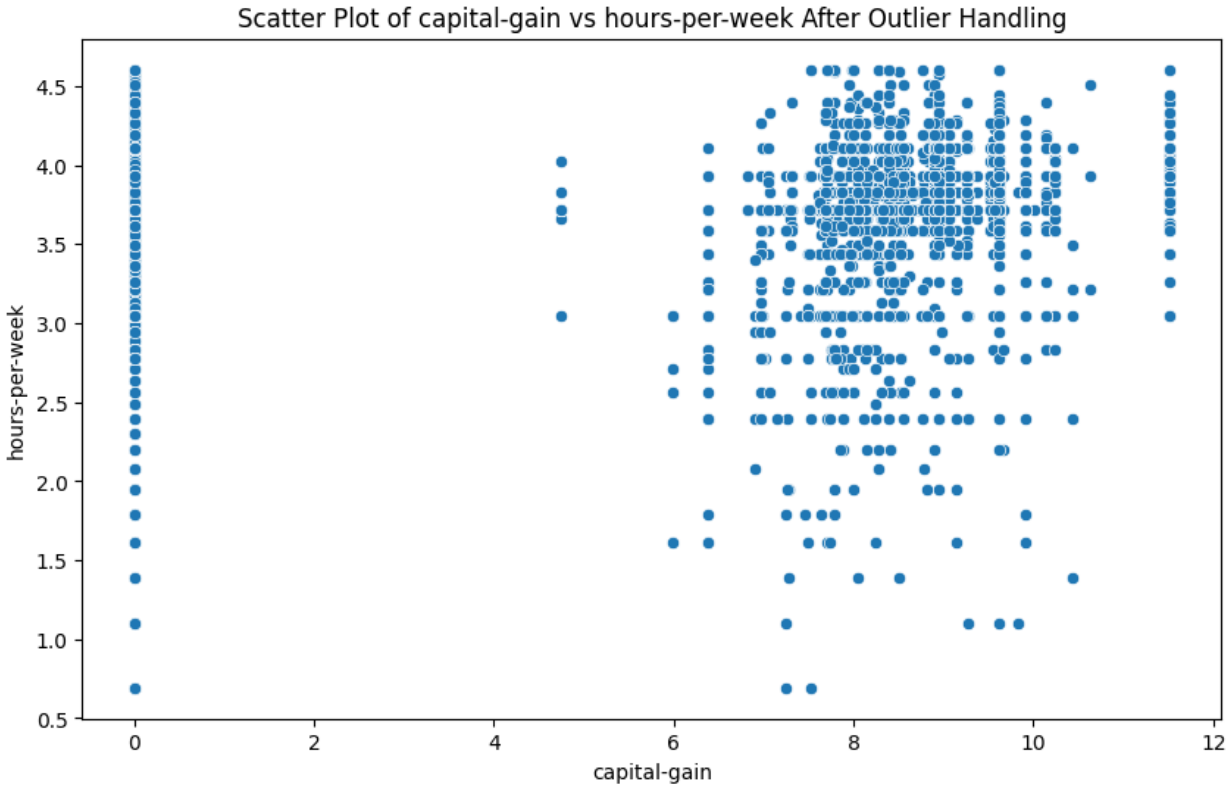

Scatter Plot of fnlwgt vs hours-per-week After Outlier Handling
hours-per-week
fnlwgt
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
10
11
12
13
14

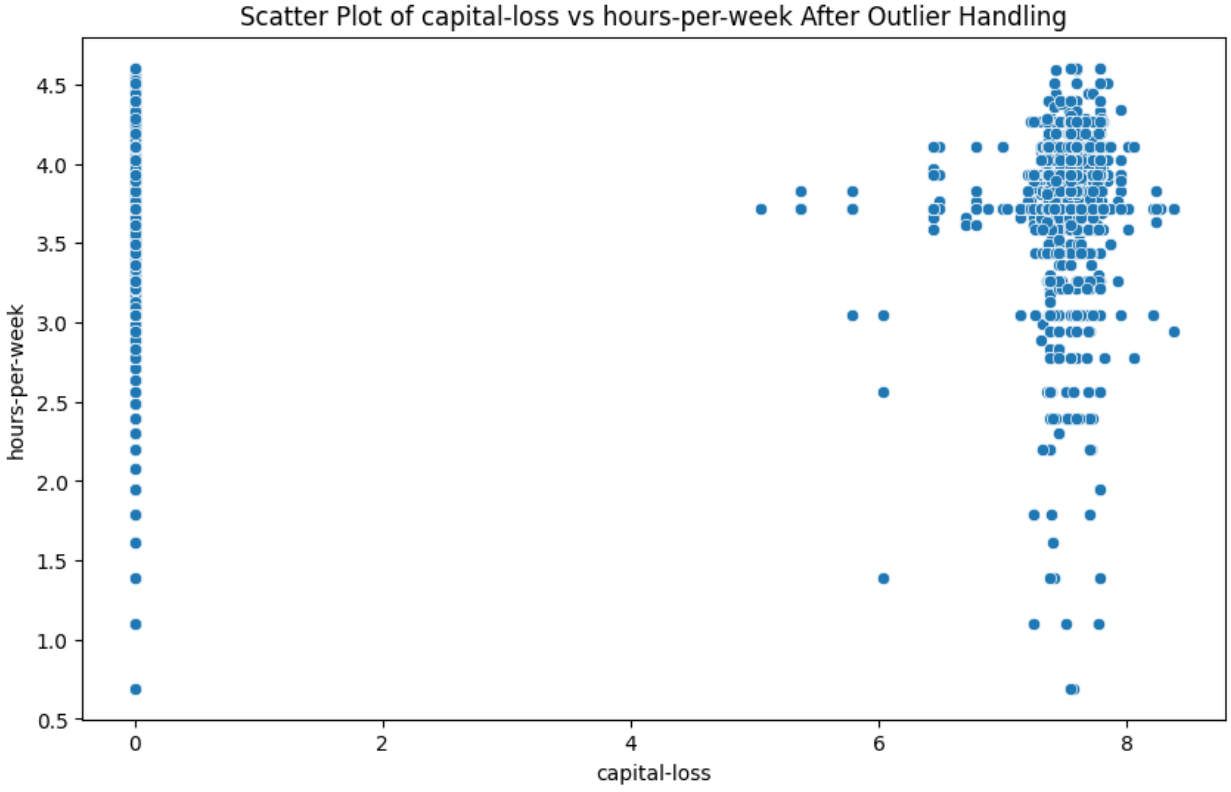

Scatter Plot of education-num vs capital-gain After Outlier Handling
capital-gain
education-num
0
2
4
6
8
10
12
1.0
1.5
2.0
2.5


Scatter Plot of education-num vs capital-loss After Outlier Handling
capital-loss
education-num
0
2
4
6
8
1.0
1.5
2.0
2.5


Scatter Plot of education-num vs hours-per-week After Outlier Handling
hours-per-week
education-num
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
1.0
1.5
2.0
2.5


Scatter Plot of capital-gain vs capital-loss After Outlier Handling
capital-loss
capital-gain
8
6
4
2
0
0
2
4
6
8
10
12


Scatter Plot of capital-gain vs hours-per-week After Outlier Handling
hours-per-week
capital-gain
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
0
2
4
6
8
10
12


Scatter Plot of capital-loss vs hours-per-week After Outlier Handling
hours-per-week
capital-loss
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
0
2
4
6
8

Accuracy with all features: 0.4385300440167878

Classification Report with all features:

	precision	recall	f1-score	support
<=50K	0.63	0.52	0.57	4936
<=50K.	0.34	0.31	0.33	2478
>50K	0.35	0.42	0.38	1562
>50K.	0.18	0.35	0.24	793
accuracy			0.44	9769
macro avg	0.38	0.40	0.38	9769
weighted avg	0.48	0.44	0.45	9769