

گزارش فاز اول پروژه مبانی داده کاوی

مبینا صولتی ۱۲۶۲۰۶۸

گروه ۴ دیتاست : adult

عنوان پروژه:

تحلیل داده‌های جمعیت‌شناسی از دیتاست Adult جهت آماده‌سازی برای مدل‌سازی داده‌کاوی

هدف پروژه:

هدف این فاز، آماده‌سازی داده‌های خام برای تحلیل‌های پیشرفته‌تر از طریق مراحل پیش‌پردازش است. این شامل بررسی کیفیت داده، پاک‌سازی داده‌های گمشده، پرت، نویزی، قالب‌بندی اشتباہ، و آماده‌سازی برای تحلیل‌های بعدی است.

مشخصات کلی داده‌ها

• تعداد رکوردها: ۵۰۰۶۴

• تعداد ستون‌ها: ۱۵ ویژگی

• نوع داده‌ها: ترکیبی از عددی و متغیر

• نمونه ویژگی‌ها:

(باید عددی باشد، ولی به صورت شیء ذخیره شده) ○

workclass, education, marital-status: ○

capital-gain, capital-loss: ○

income: ○

بخش اول : بررسی اولیه داده ها

اطلاعات ساختاری:

با استفاده از `adults.describe()` و `adults.info()` ساختار کلی داده ها بررسی شد.
ستون های عددی و متنی به تفکیک تحلیل شده اند.

بررسی مقادیر گمشده:

هیچ مقدار NaN استاندارد وجود ندارد اما مقادیر مشکوک مثل ? در برخی ستون ها شناسایی شدند.

بررسی مقادیر تکراری:

تعدادی ردیف تکراری (duplicated) وجود دارد که مشخص شده و می توان آنها را حذف کرد.

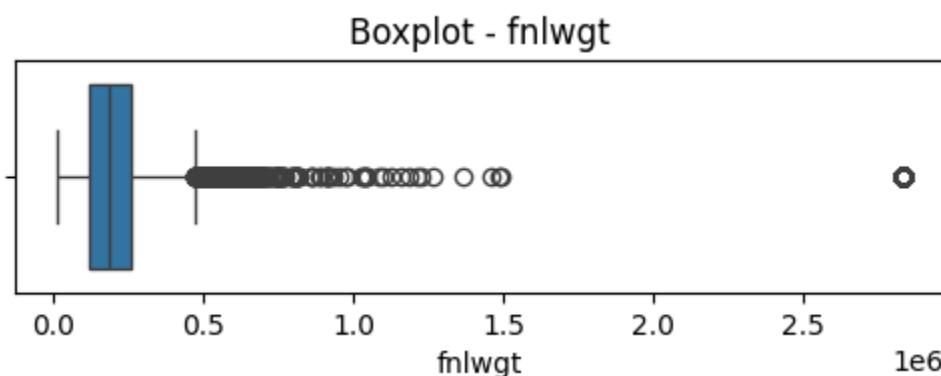
تحلیل داده های پرت و نویزی

مقادیر پرت:

- با استفاده از نمودارهای Boxplot برای هر ستون عددی، مقادیر پرت بررسی شده اند.
- متغیرهایی مثل capital-loss و capital-gain مقادیر پرت قابل توجهی هستند.

داده های نویزی:

- مقادیر دارای نویز متنی مانند فاصله اضافی، کاراکترهای خاص، و کلمات مشابه بررسی شده اند.
- از کتابخانه RapidFuzz جهت تشخیص شباهت های واژگانی استفاده شده.

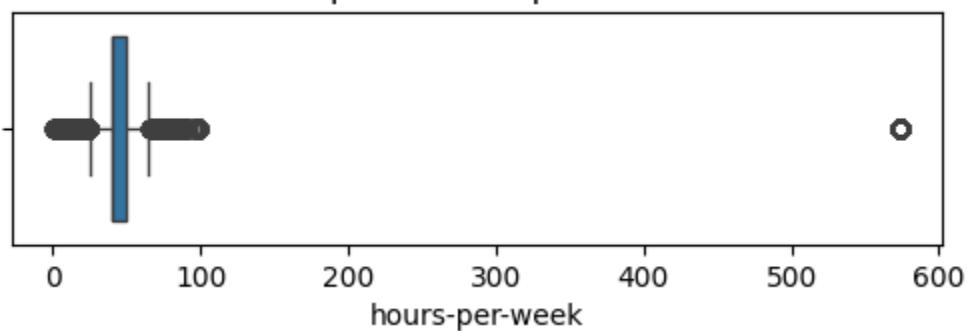


متغیری که نشان دهنده وزن نمونه ای هر فرد توی آمارگیری سرشماری هست (یعنی چقدر "Final Weight" یا نماینده جمعیت بزرگ ترہ)

Boxplot - capital-loss



Boxplot - hours-per-week

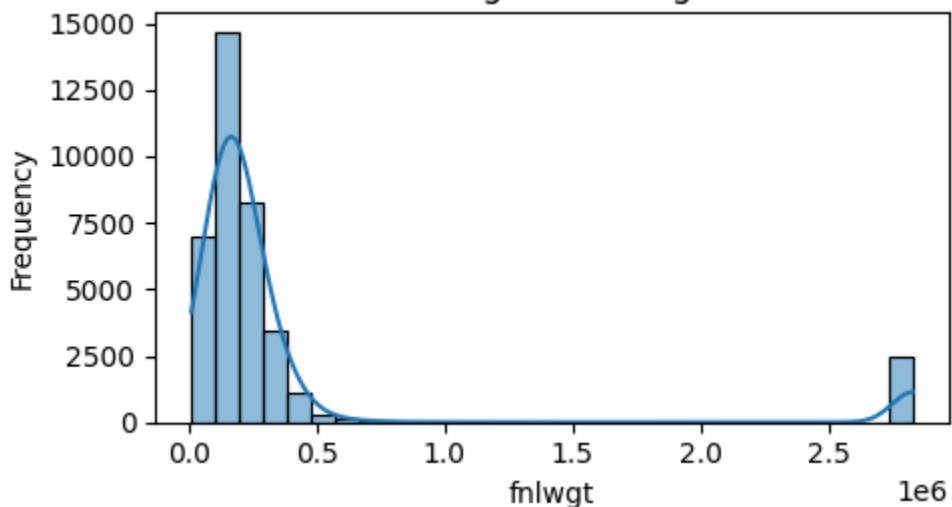


کیفیت داده ها:

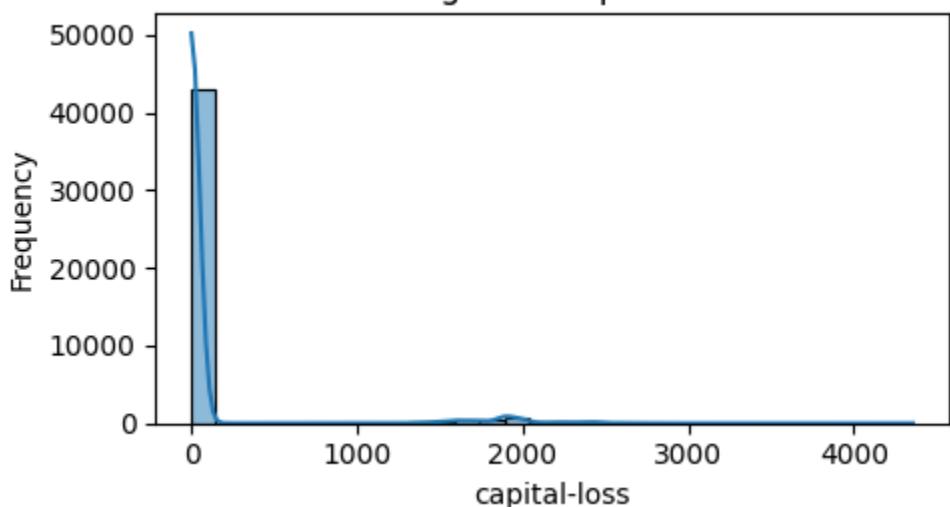
	feature_name	record_count	null_values	accuracy	completeness	validity	consistency	currentness
0	age	50064	2494	100	95.018376	NaN	100	None
1	workclass	50064	11910	100	76.210451	NaN	100	None
2	fnlwgt	50064	12681	100	74.670422	NaN	100	None
3	education	50064	11936	100	76.158517	71.310722	100	None
4	education-num	50064	11596	100	76.837648	NaN	100	None
5	marital-status	50064	11959	100	76.112576	NaN	100	None
6	occupation	50064	11967	100	76.096596	NaN	100	None
7	relationship	50064	11945	100	76.140540	NaN	100	None
8	race	50064	11962	100	76.106584	NaN	100	None
9	sex	50064	11894	100	76.242410	71.360658	100	None
10	capital-gain	50064	1	100	99.998003	NaN	100	None
11	capital-loss	50064	4782	100	90.448226	NaN	100	None
12	hours-per-week	50064	12664	100	74.704378	NaN	100	None
13	native-country	50064	11930	100	76.170502	NaN	100	None
14	income	50064	1	100	99.998003	66.662672	100	None

نمایش هیستوگرام برای ستونهای عددی:

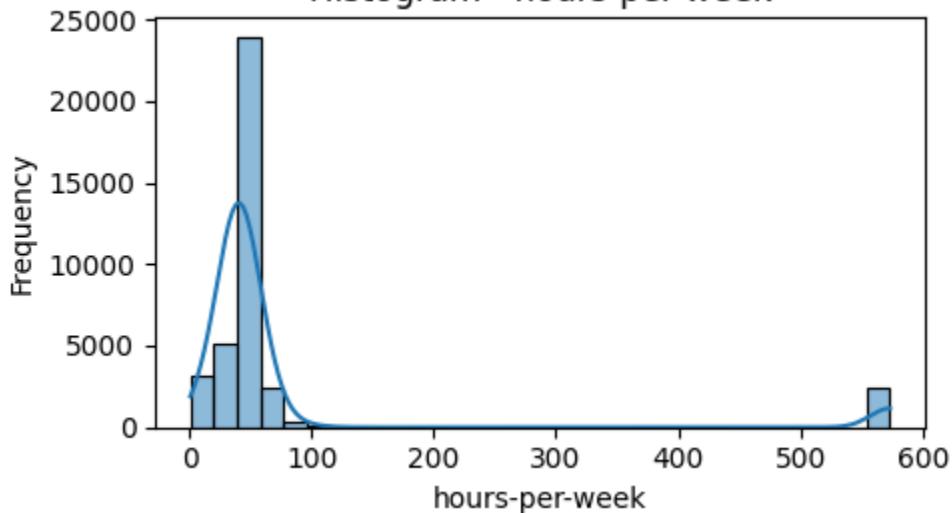
Histogram - fnlwgt



Histogram - capital-loss

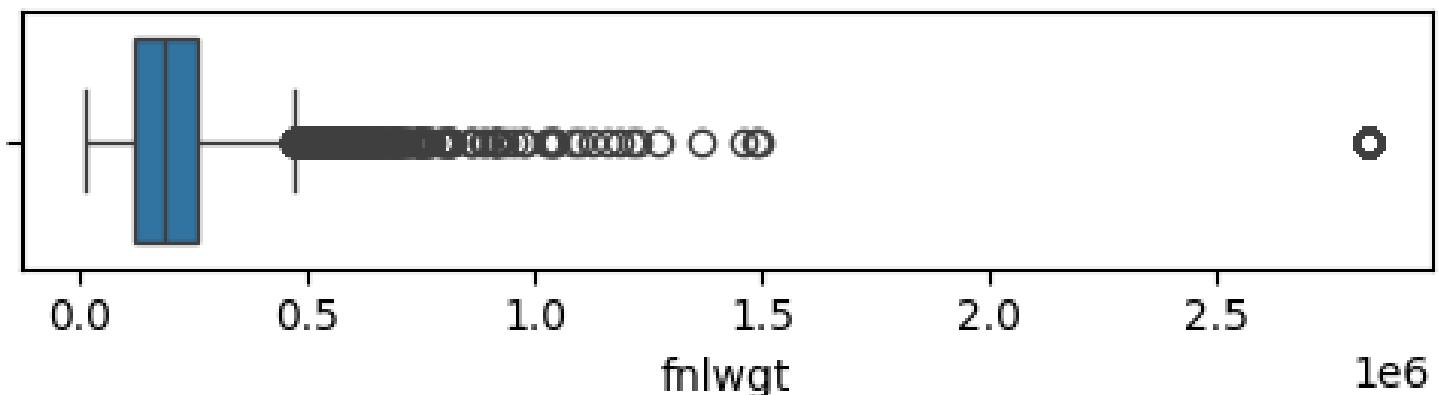


Histogram - hours-per-week

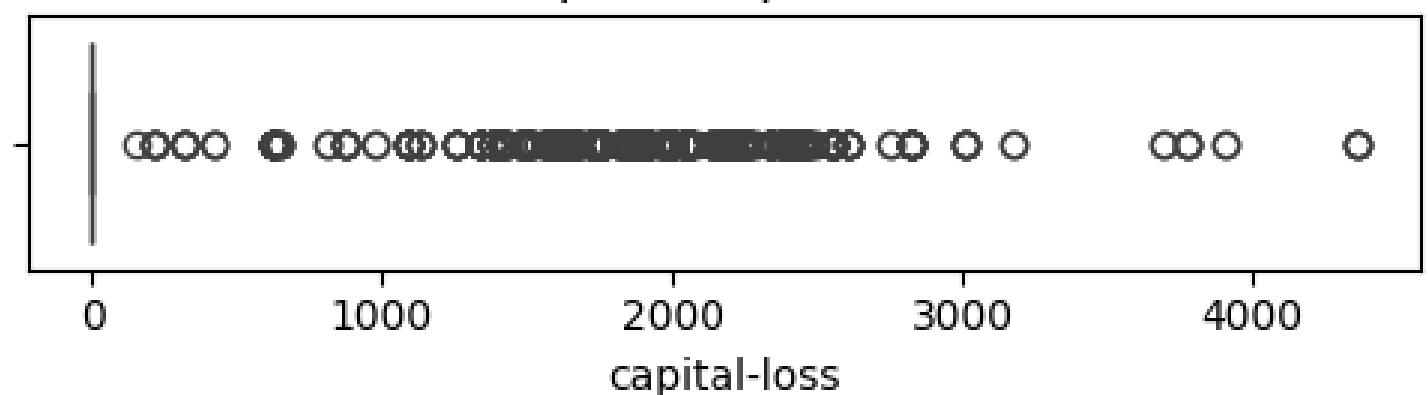


باکس پلات برای هر ویژگی عددی:

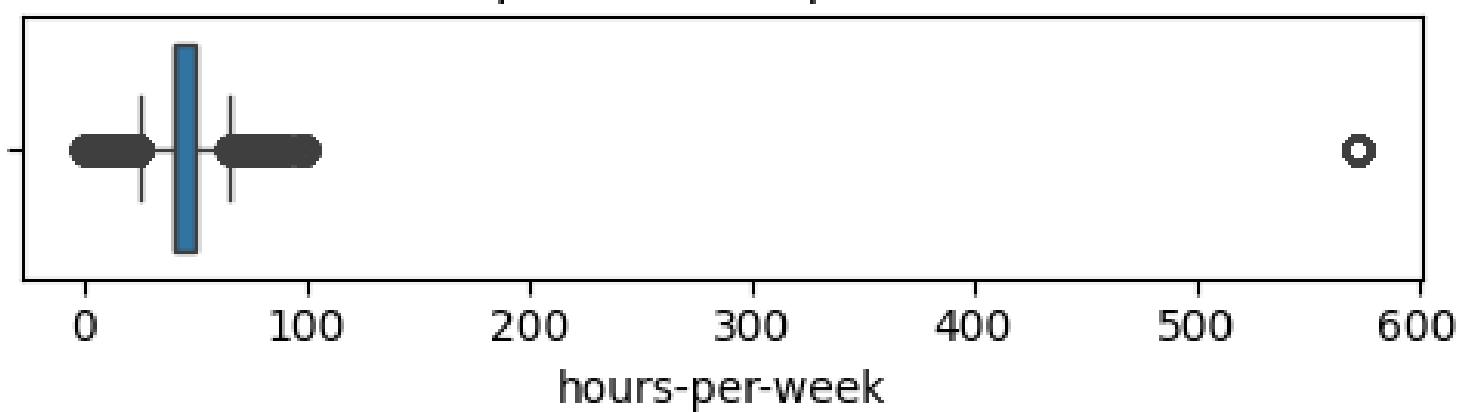
Boxplot - fnlwgt



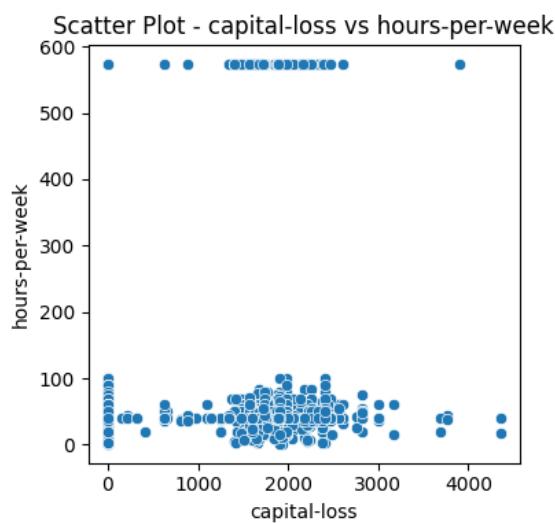
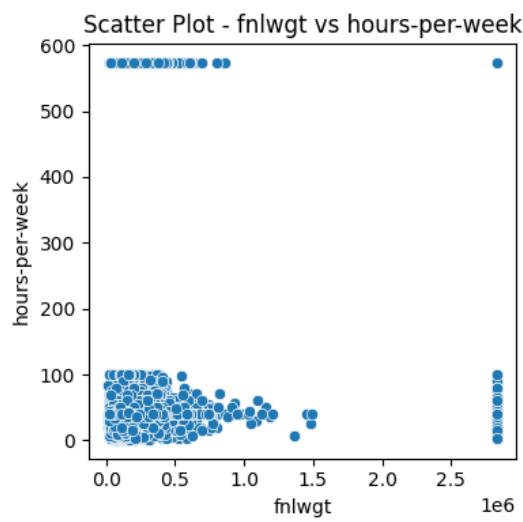
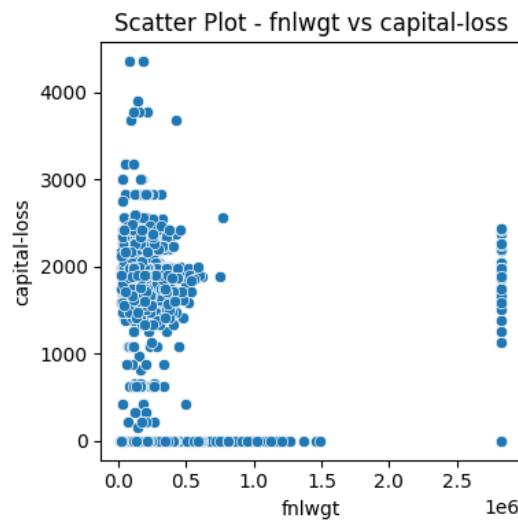
Boxplot - capital-loss



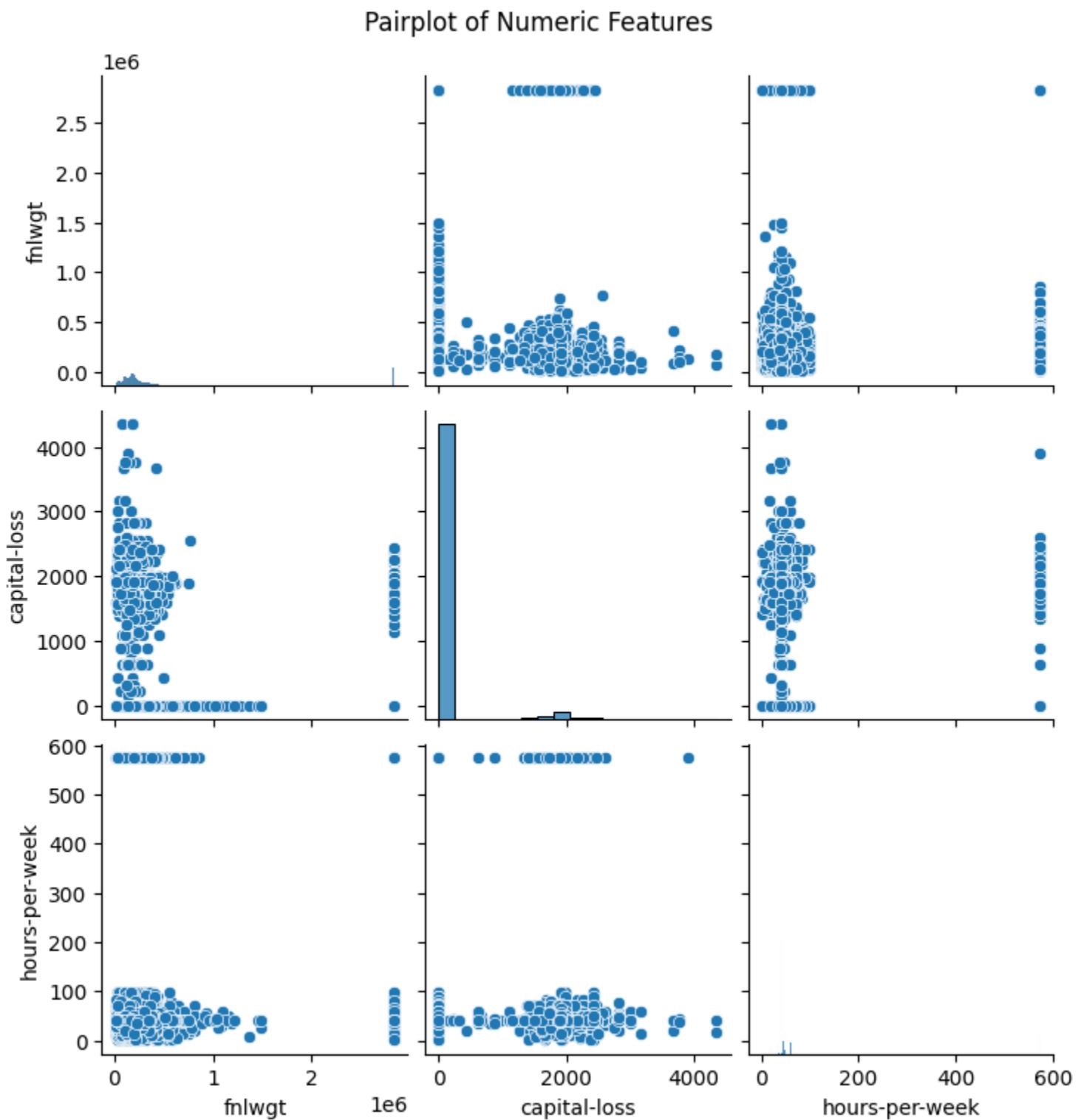
Boxplot - hours-per-week



بین هر دو ستون عددی :Scatter Plot



: Pairplot of Numeric Features



بخش دوم : پیش پردازش داده ها

مقادیر گمشده قبل از جایگزینی:

```
age           2494
workclass    14005
fnlwgt       12681
education    11936
education-num 11596
marital-status 11959
occupation   14085
relationship  11945
race          11962
sex           11894
capital-gain 1
capital-loss  4782
hours-per-week 12664
native-country 12569
income         1
dtype: int64
```

مقادیر گمشده بعد از جایگزینی:

```
age           0
workclass    0
fnlwgt       0
education    0
education-num 0
marital-status 0
occupation   0
relationship  0
race          0
sex           0
capital-gain 0
capital-loss  0
hours-per-week 0
native-country 0
income         0
dtype: int64
```

نرمالسازی داده ها:

مشخصه که نرمال سازی به درستی انجام شده چون:

- تمامی مقادیر عددی بین ۰ و ۱ قرار دارند (محدوده نرمال شده).
- هیچ مقداری خارج از بازه استاندارد دیده نمی شود.
- همه ویژگی هایی که باید نرمال شوند (age, fnlwgt, education-num, ...) به درستی نرمال شده اند.

نمونه ای از داده های نرمال شده:

	age	fnlwgt	education-num	capital-gain	capital-loss	\
0	0.000000	0.071689	0.533333	0.0	0.0	
1	0.574713	0.078463	0.533333	0.0	0.0	
2	0.287356	1.000000	0.800000	0.0	0.0	
3	0.528736	0.051995	0.266667	0.0	0.0	
4	0.000000	0.069557	0.533333	0.0	0.0	

	hours-per-week
0	0.397959
1	0.397959
2	0.397959
3	0.397959
4	0.397959

حذف داده‌های «شبه‌تکراری» (اختیاری):

این مورد پیشرفت‌تره و بسته به پروژه ممکنه نیاز باشه. مثلا:

- ردیف‌های که مقدار ویژگی‌های مهمشون (education) مشابه هستن.
- یا با درصد شباهت بالا در رشته‌ها.

ردیف تکراری حذف شدند پس از اصلاح مقادیر مشابه 2						
age	workclass	fnlwgt	education	education-num		\
0	154.0	Private	77516.00	Bachelors	13.0	
1	154.0	Self-emppnot-inc	83311.00	Bachelors	13.0	
2	3.0	Private	215646.00	HS-grad	9.0	
3	53.0	Private	234721.00	11th	9.0	
4	28.0	Private	2829764.77	Bachelors	13.0	
	marital-status	occupation	relationship	race	sex	\
0	Married-civtspouse	Adm-clerical	Not-iv-family	White	Male	
1	Married-civtspouse	Exic-managerial	Husband	White	Male	
2	Divorced	Handlers-cleaners	Not-iv-family	White	Male	
3	Married-civtspouse	Handlers-cleaners	Husband	Black	Male	
4	Married-civtspouse		198	Wife	White	Male
	capital-gain	capital-loss	hours-per-week	native-country	income	
0	2174.0	0.0	40.0	270	<=50K	
1	0.0	0.0	13.0	United-States	<=50K	
2	0.0	0.0	40.0	United-States	<=50K	
3	0.0	0.0	40.0	United-States	<=50K	
4	0.0	0.0	40.0	United-States	<=50K	

مقایسه توزیع داده ها:

هدف:

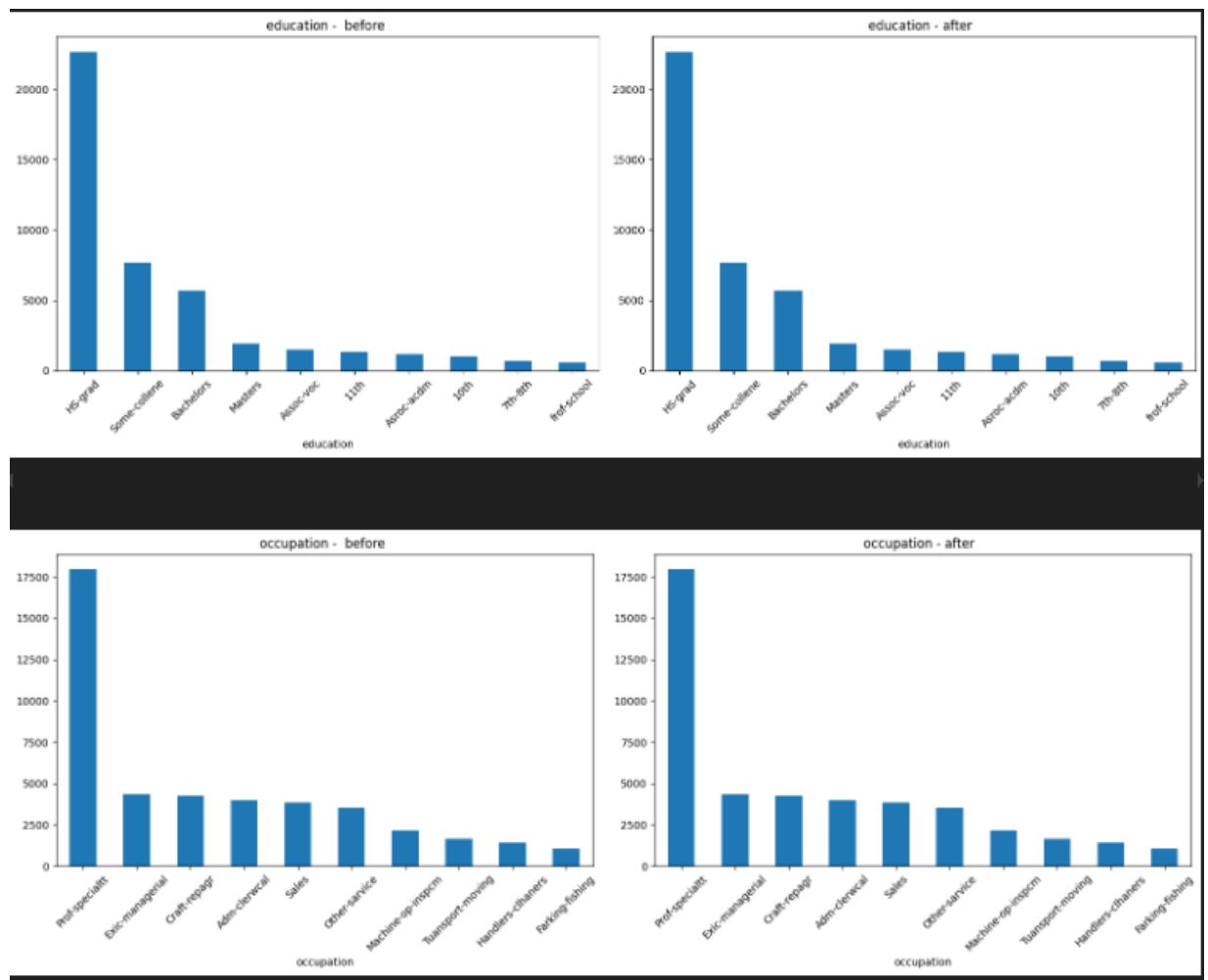
بررسی کنیم که آیا توزیع دو ویژگی (occupation و education) مثلاً چقدر با هم تفاوت دارند؟ اگر توزیع شون خیلی متفاوته یا خیلی شبیه، با KL Divergence اینو می فهمیم. مقدار KL Divergence اگر به . نزدیک باشه یعنی توزیع ها خیلی شبیه‌اند.

هر چی مقدار بزرگ‌تر باشه → تفاوت توزیع ها بیشتره.

KL Divergence بین 'education' و 'occupation': 9.3133

مقدار KL Divergence بین ویژگی‌های occupation و education برابر 9.3133 شده که عدد نسبتاً بزرگیه — این نشون می‌ده که توزیع این دو ویژگی تفاوت قابل توجهی با هم دارن (و شباهت کمی بینشونه).

نمایش داده های پاکسازی شده:



تحلیل نمودار مقایسه‌ای ستون: **education**:

نقاط قوت نمودار:

فرمت ساید-بای-ساید (before/after): خیلی خوبه برای مقایسه دیداری سریع.

محورها و برچسب‌ها واضح‌تر.

مقدارهای پرتکرار مثل HS-grad و Some-college در هر دو نمودار دیده می‌شون و ثبات دارند.

نکات قابل توجه:

اشتباهات اصلاح نشده هنوز باقی‌اند!

مثلاً **BacheLors** و **Bachelors** در سمت چپ و راست یکی شدن، اما هنوز مواردی مثل:

Some-college به جای Some-college

Assoc-voc و Assoc-acdm که ممکنه اشکال تایپی داشته باشن.

که احتمالاً باید Preschool یا Pre-school باشد.

KL Divergence بالای ۹/۳۱ که قبلاً محاسبه شد، نشون می‌ده توزیع خیلی تغییر کرده؛ اما نمودار نشون می‌ده که تغییرات بیشتر در ریزدسته‌ها بوده نه در گروه‌های پرتکرار.

پیشنهاد برای بهبود تحلیل:

برای اینکه دقیق‌تر تحلیل کنیم:

ستون‌های اصلاح نشده یا مشکوک به تایپی رو جدا چاپ کنیم.

یا پیش و پس از اصلاح مقدارهایی که خیلی کم تکرار شدن رو جدا نشون بدیم، چون اون‌ها بیشتر منشأ هستن.