EDA 탐색 가이드라인

1. 간단한 탐색

```
train_data.head()
```

	Unnamed: 0	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race
0	0	40	Private	168538	HS-grad	9	Married-civ-spouse	Sales	Husband	Whit
1	1	17	Private	101626	9th	5	Never-married	Machine-op-inspct	Own-child	Whit
2	2	18	Private	353358	Some-college	10	Never-married	Other-service	Own-child	Whit
3	3	21	Private	151158	Some-college	10	Never-married	Prof-specialty	Own-child	Whit
4	4	24	Private	122234	Some-college	10	Never-married	Adm-clerical	Not-in-family	Blac

2. 기본적인 데이터 정보 확인

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26049 entries, 0 to 26048
Data columns (total 16 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Unnamed: 0      26049 non-null  int64
 1   age             26049 non-null  int64
 2   workclass       26049 non-null  object
 3   fnlwgt          26049 non-null  int64
 4   education       26049 non-null  object
 5   education_num   26049 non-null  int64
 6   marital_status  26049 non-null  object
 7   occupation      26049 non-null  object
 8   relationship    26049 non-null  object
 9   race            26049 non-null  object
 10  sex             26049 non-null  object
 11  capital_gain    26049 non-null  int64
 12  capital_loss    26049 non-null  int64
 13  hours_per_week  26049 non-null  int64
 14  native_country  26049 non-null  object
 15  income          26049 non-null  object
dtypes: int64(7), object(9)
memory usage: 3.2+ MB
```

]:
```
train_data.describe()
```

]:

	Unnamed: 0	age	fnlwgt	education_num	capital_gain	capital_loss	hours_per_week
count	26049.000000	26049.000000	2.604900e+04	26049.000000	26049.00000	26049.000000	26049.000000
mean	13024.000000	38.569235	1.903045e+05	10.088372	1087.68970	87.732734	40.443126
std	7519.842917	13.671489	1.059663e+05	2.567610	7388.85469	403.230205	12.361850
min	0.000000	17.000000	1.376900e+04	1.000000	0.00000	0.000000	1.000000
25%	6512.000000	28.000000	1.181080e+05	9.000000	0.00000	0.000000	40.000000
50%	13024.000000	37.000000	1.788660e+05	10.000000	0.00000	0.000000	40.000000
75%	19536.000000	48.000000	2.377350e+05	12.000000	0.00000	0.000000	45.000000
max	26048.000000	90.000000	1.484705e+06	16.000000	99999.00000	4356.000000	99.000000

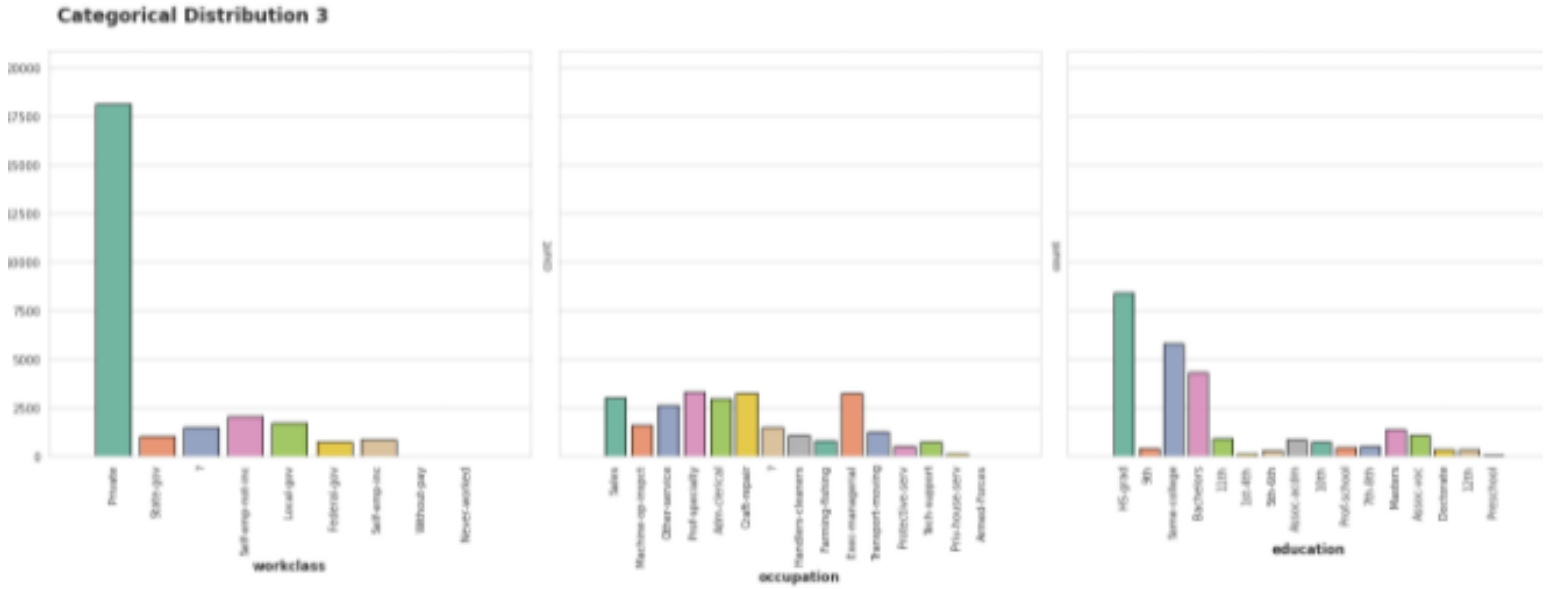
3. 각 feature 들의 의미 이해

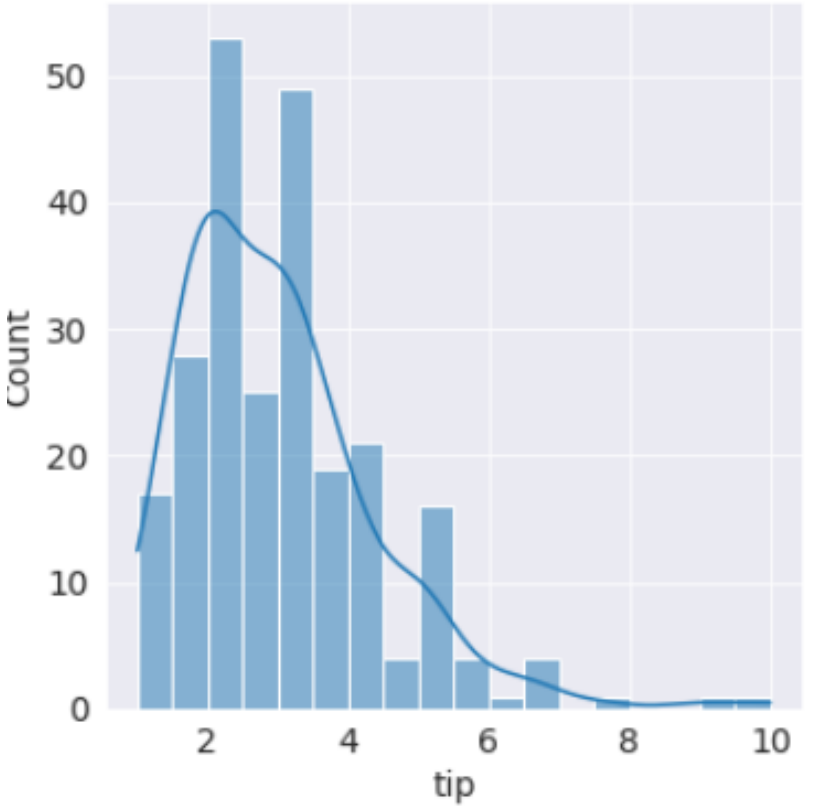
- `age` : 나이
- `workclass` : 고용 형태
- `fnlwgt` : 사람 대표성을 나타내는 가중치 (final weight의 약자)
- `education` : 교육 수준
- `education_num` : 교육 수준 수치
- `marital_status` : 결혼 상태
- `occupation` : 업종
- `relationship` : 가족 관계
- `race` : 인종
- `sex` : 성별
- `capital_gain` : 양도 소득
- `capital_loss` : 양도 손실
- `hours_per_week` : 주당 근무 시간
- `native_country` : 국적
- `income` : **수익 (예측해야 하는 값)**

범주형 자료	수치형 자료
race, education …	Income, age, …

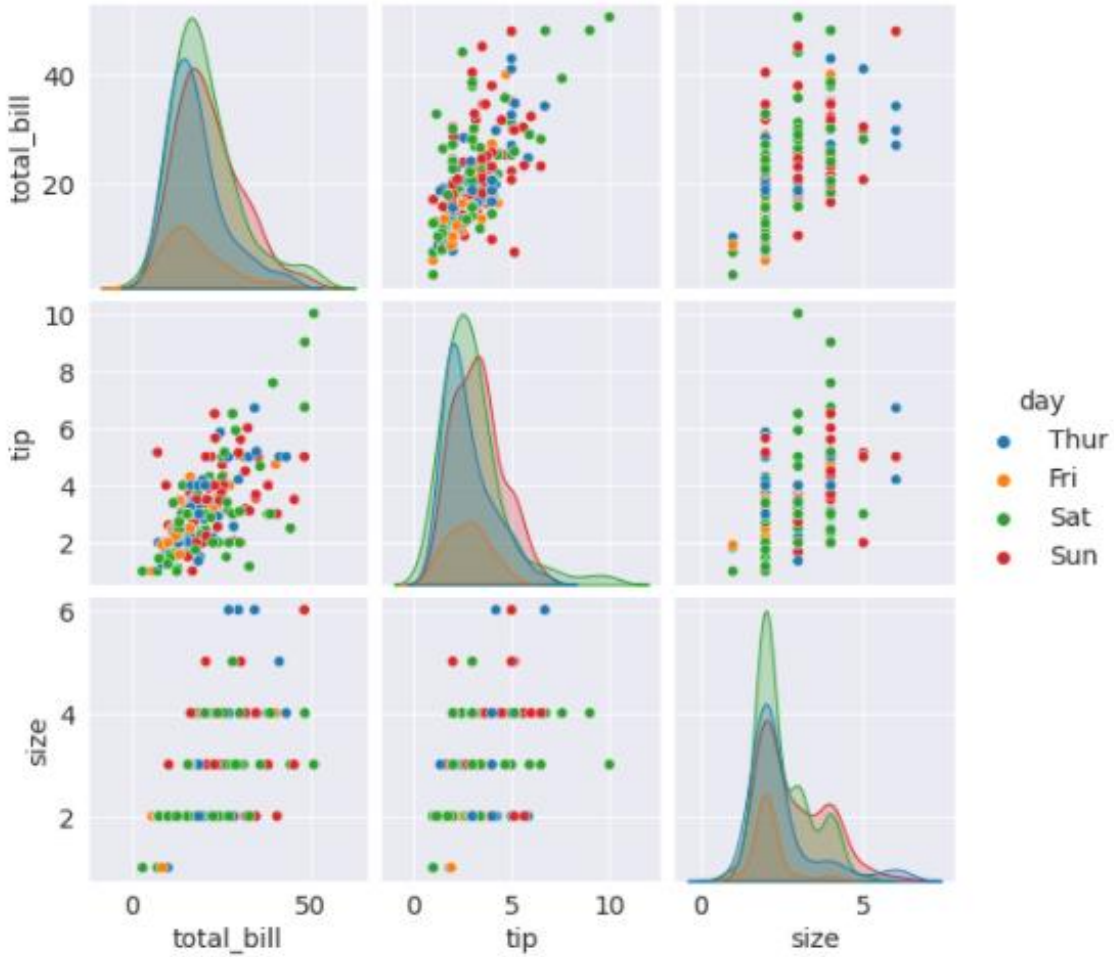
Y(예측값)	X
income	race, education, age,

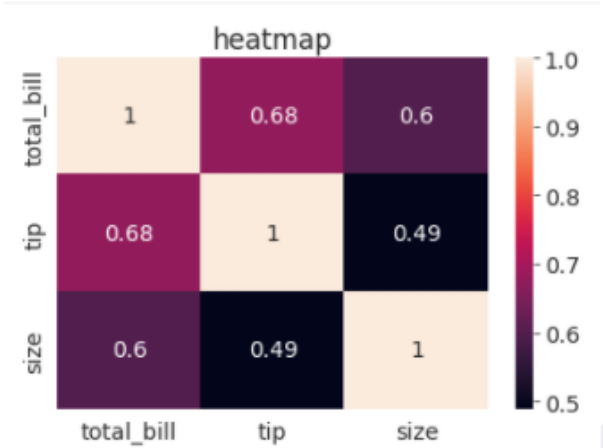
4. 각 속성별 분석 및 시각화

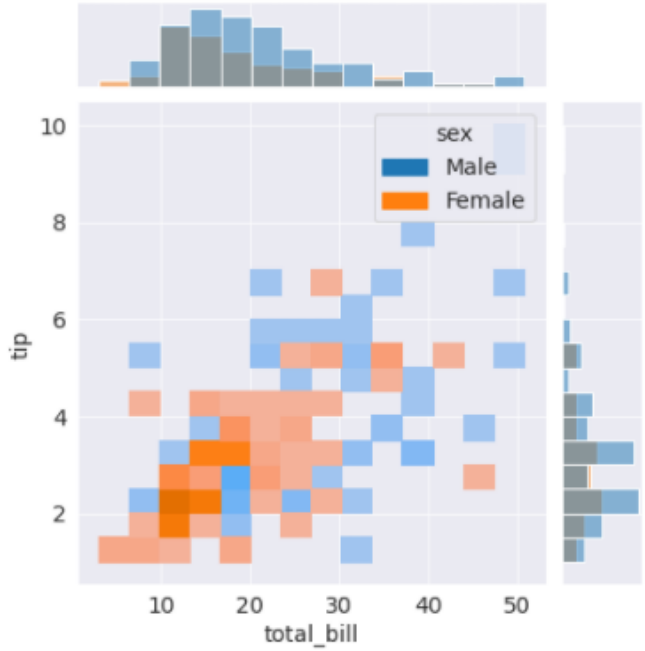




5. 속성간 분석 및 시각화


total_bill
tip
size
day
Thur
Fri
Sat
Sun


heatmap
total_bill
tip
size
1
0.68
0.6
0.68
1
0.49
0.6
0.49
1


sex
Male
Female
tip
total_bill

6. Summary 및 데이터 설명

- 여러 독립변수 중 종속변수와 상관관계가 큰 것은?
- 문제가 있는 속성은?
- 결측치, 이상치 등이 있다면 처리할 방법은?
- 전반적인 데이터의 경향은?
- 각 속성별 두드러진 특징 혹은 속성간 유의미한 의미가 있는지?
- …