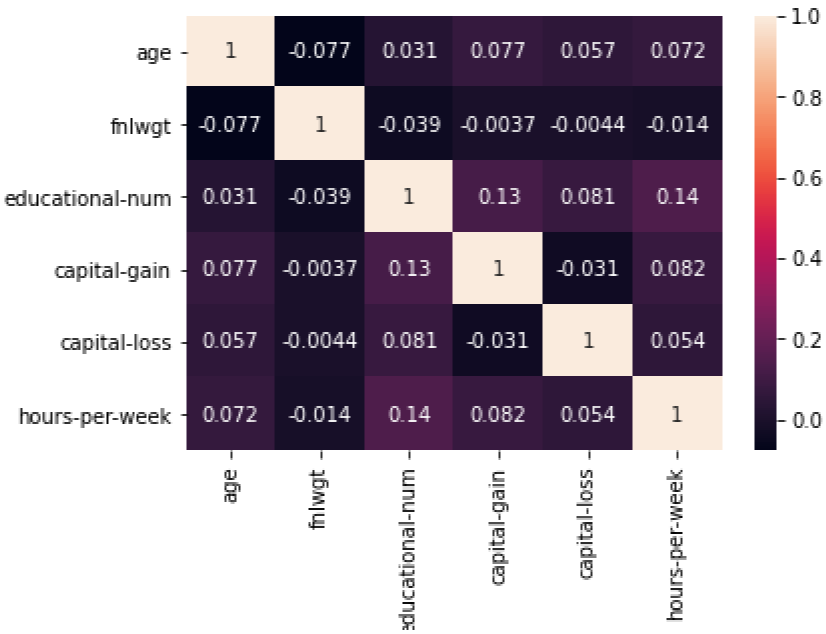
Krittamate Cherdpongtagit 6310545884

1. Use the Adult data set, where the target variable is income, and the goal is to classify income based on the other variables.
Answer the following questions.
a. Which variables are categorical, and which are continuous?

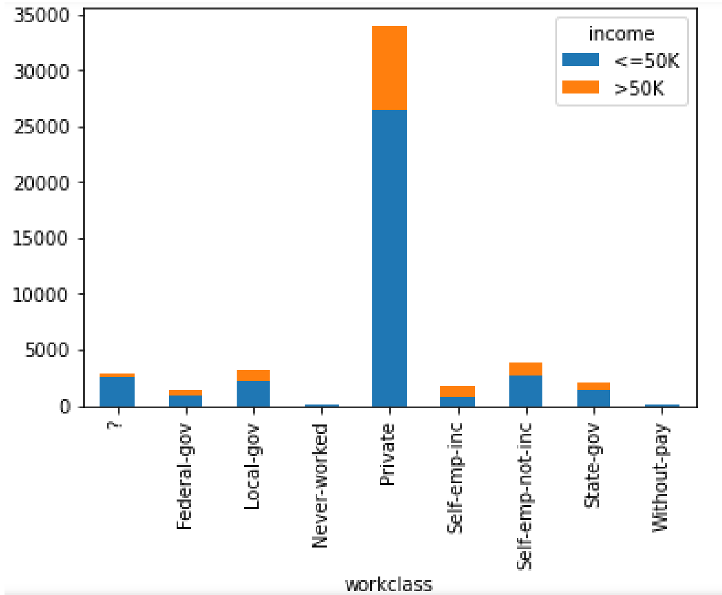
Categorical : 'workclass', 'education', 'marital-status', 'occupation', 'relationship', 'race', 'gender', 'native-country', 'income'
Continuous : 'age', 'fnlwgt', 'educational-num', 'capital-gain', 'capital-loss', 'hours-per-week'

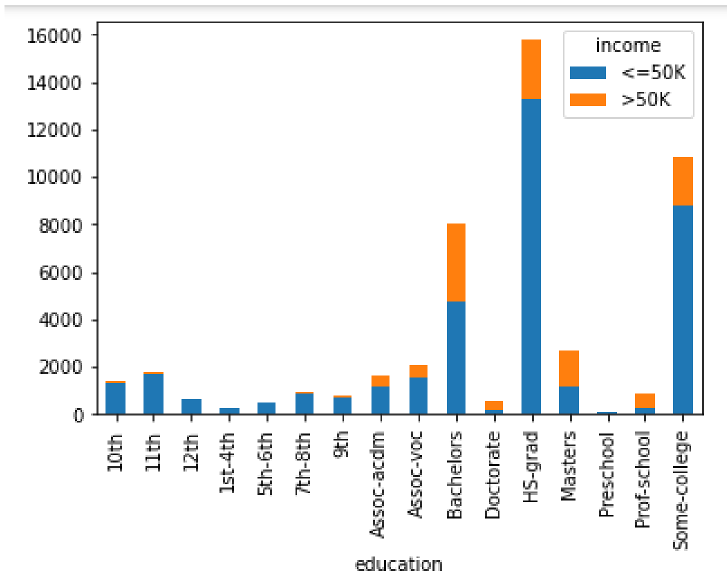
b. Investigate whether we have any correlated variables.

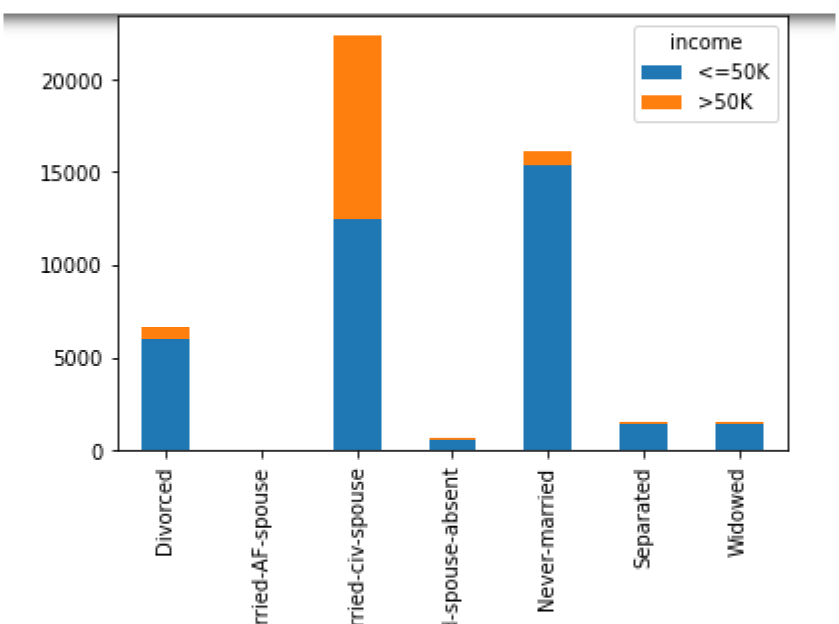
We have correlated variables between age,fnlwgt, education-num, capital gain, capital-loss, hours-per-week



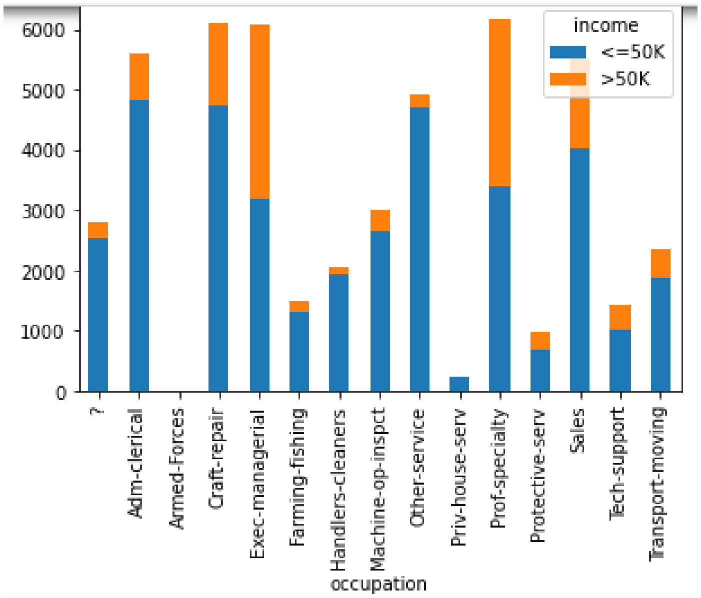
c. For each of the categorical variables, construct a bar chart of the variable, with an overlay of the target variable. Normalize if necessary.

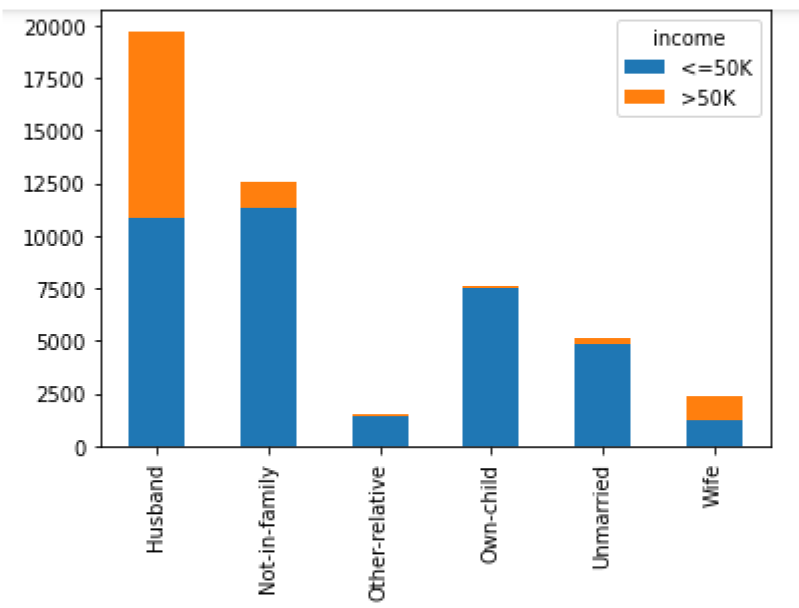


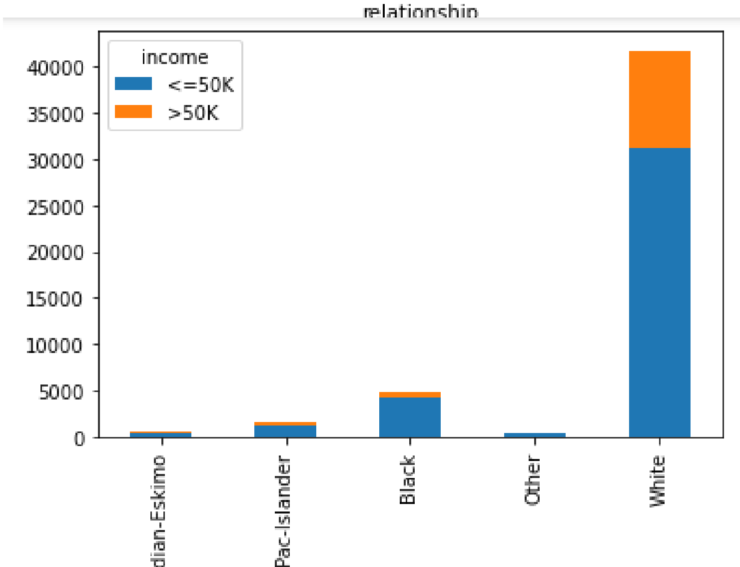




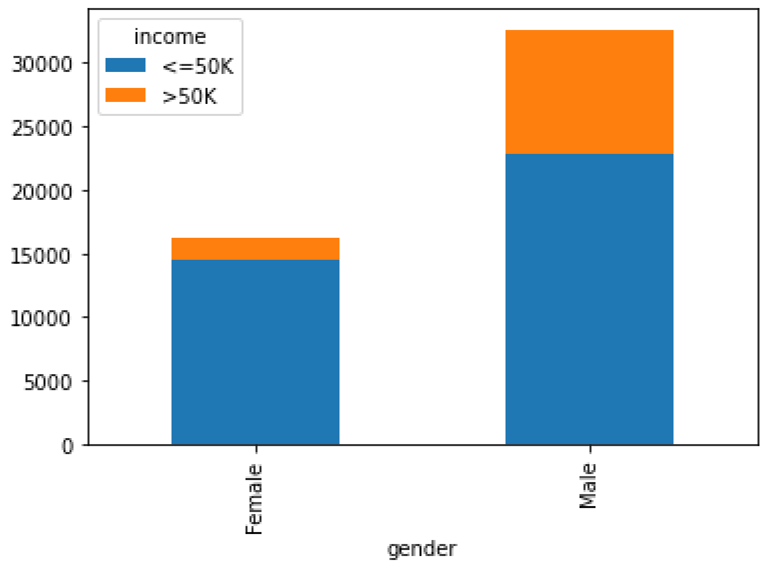
Krittamate Cherdpongtagit 6310545884

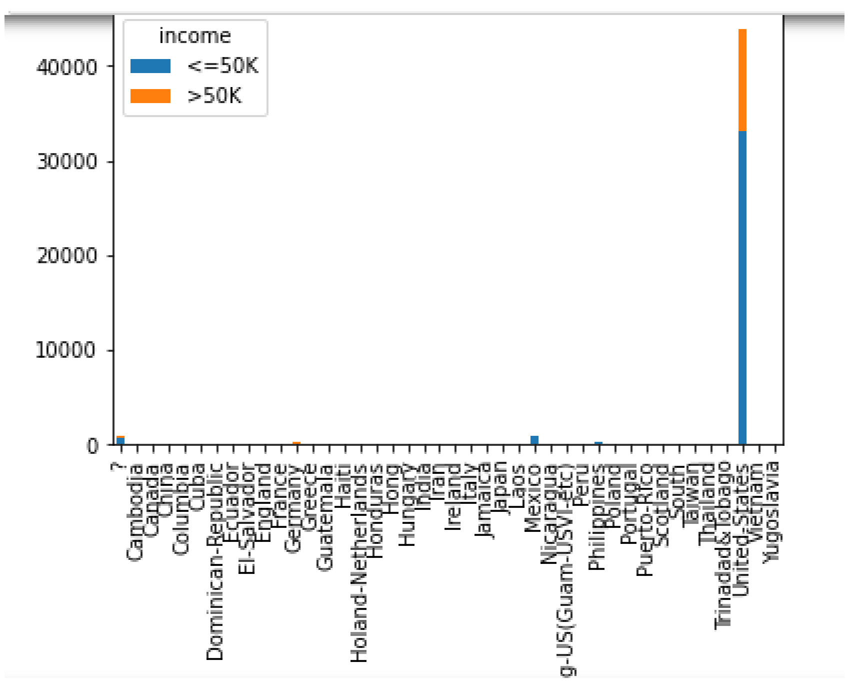






Krittamate Cherdpongtagit 6310545884





Krittamate Cherdpongtagit 6310545884

i. Discuss the relationship, if any, each of these variables has with the target Variables.

-in work class, private has the most frequency of both > and <= 50k

-in education bachelors has the most frequency in > 50k and HS-grad has the most in <= 50K

-in marital-status married af spouse has the most in > 50k and never married has the most in <=50k

-in occupation has average in all categories

-in relationship, husband has the most in >50k and almost the same as not in family in <=50k

-in race, white has the most in both > and <= 50k

-in gender, male has more frequency than female in both > and <= 50k

ii. Which variables would you expect to make a significant appearance in any classification model we work with?

workclass, education, marital status, race, gender

d. Report on whether anomalous fields exist in this dataset, and what we should do about it.

Replace missing value

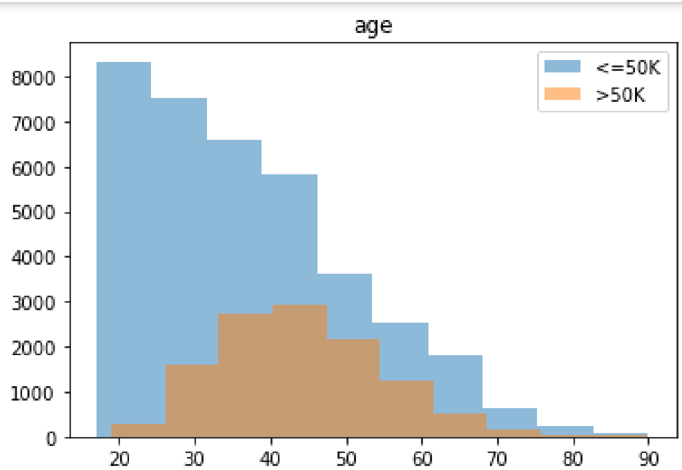
In [37]:
```
df.replace("?", float("NaN"), inplace=True)

print("Number of missing values:")
print(df.isnull().sum())
```

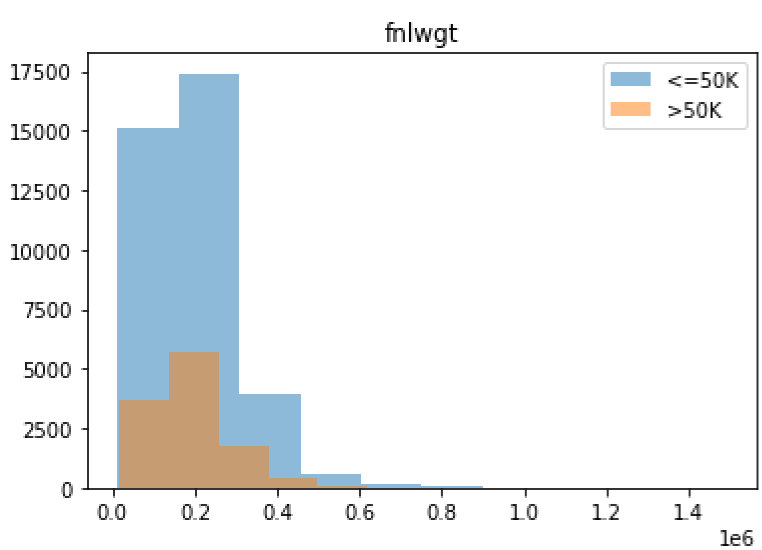
```
Number of missing values:
age                  0
workclass         2799
fnlwgt               0
education            0
educational-num      0
marital-status       0
occupation        2809
relationship         0
race                 0
gender               0
capital-gain         0
capital-loss         0
hours-per-week       0
native-country     857
income               0
dtype: int64
```

e. Report the mean, median, minimum, maximum, and standard deviation for each
of the numerical variables.

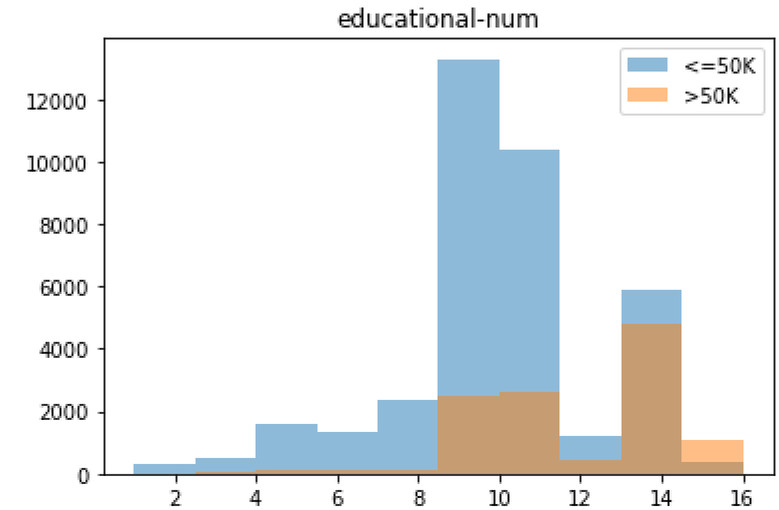
	age	fnlwgt	educational-num	capital-gain	capital-loss	hours-per-week
count	48842.000000	4.884200e+04	48842.000000	48842.000000	48842.000000	48842.000000
mean	38.643585	1.896641e+05	10.078089	1079.067626	87.502314	40.422382
std	13.710510	1.056040e+05	2.570973	7452.019058	403.004552	12.391444
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.175505e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.781445e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.376420e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.490400e+06	16.000000	99999.000000	4356.000000	99.000000

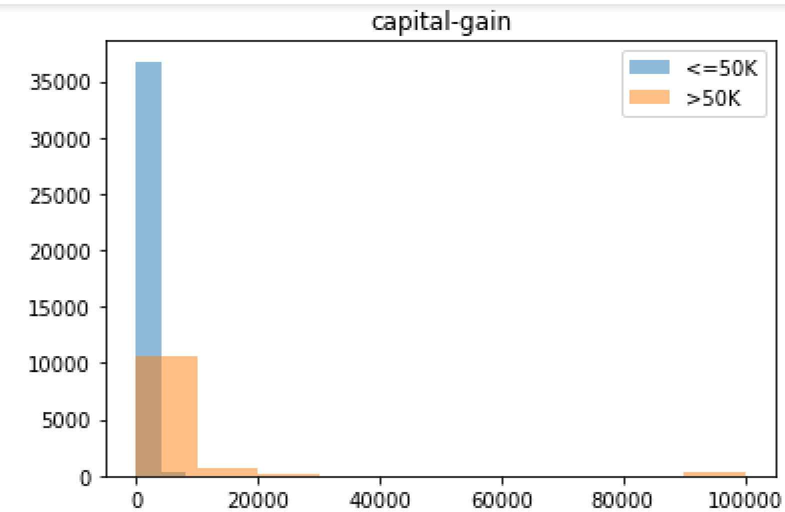
f. Construct a histogram of each numerical variables, with an overlay of the target variable income. Normalize if necessary.

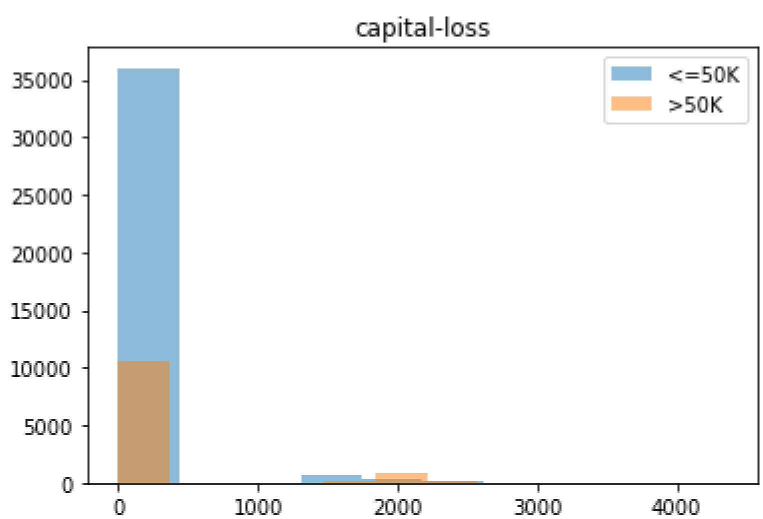


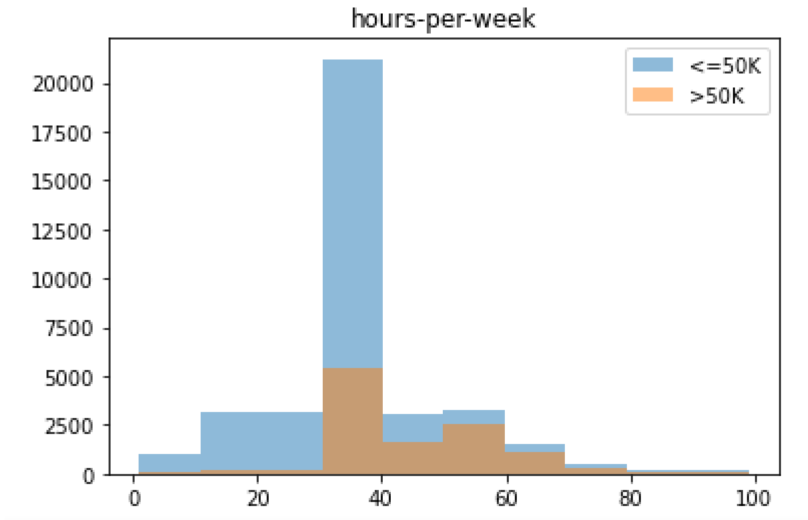


Krittamate Cherdpongtagit 6310545884









i. Discuss the relationship, if any, each of these variables has with the target Variables.

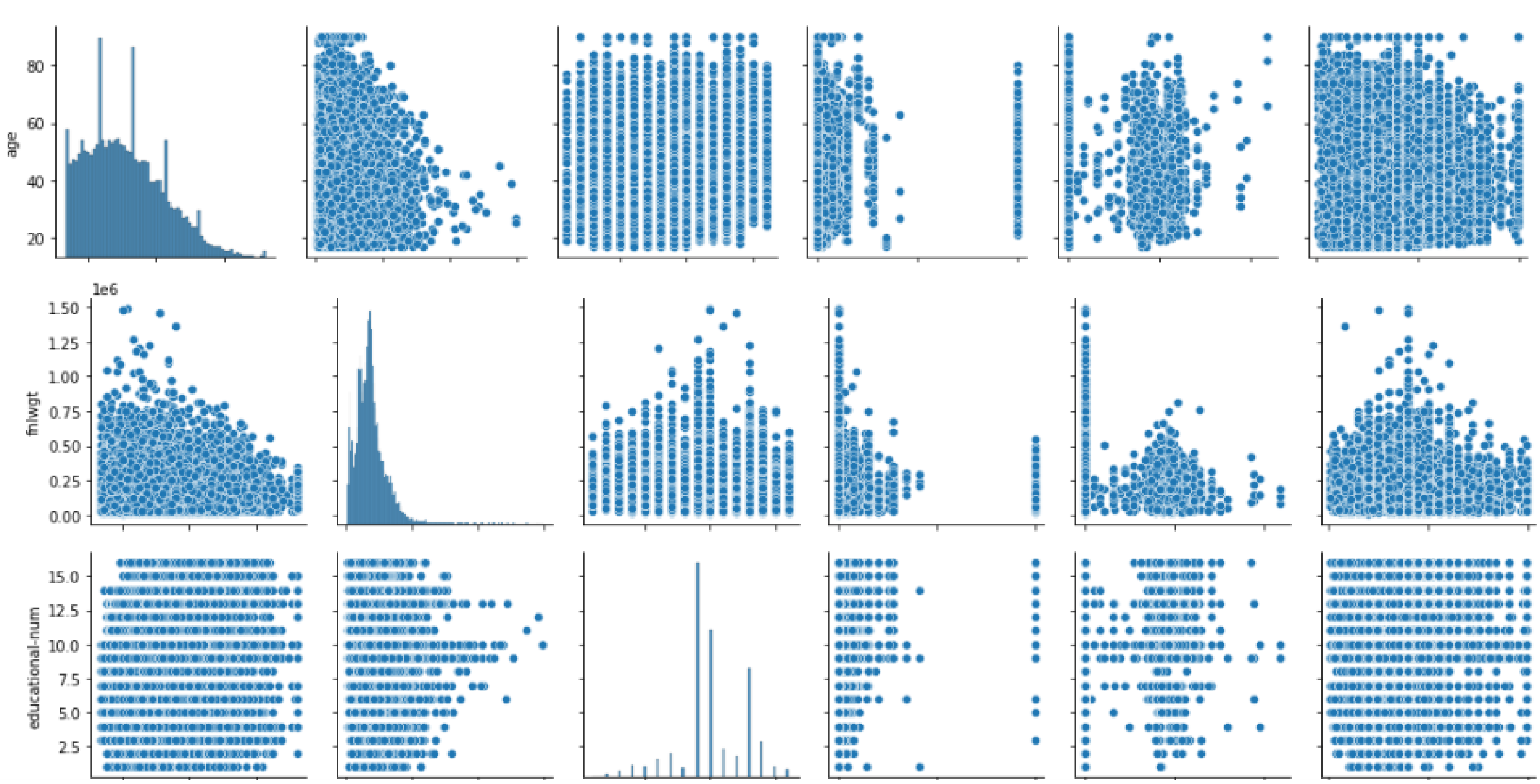
- **the higher the age is, the lower the income is. > 50k has the most frequency at 45 years old**
- **The fnlwgt has the peak at 0.2 and become lower as it increases**
- **the histogram of education-num has bell shaped. The education num 14 has the most frequency of >50k**
- **Hours-per-week has bell shaped that peak at 40**

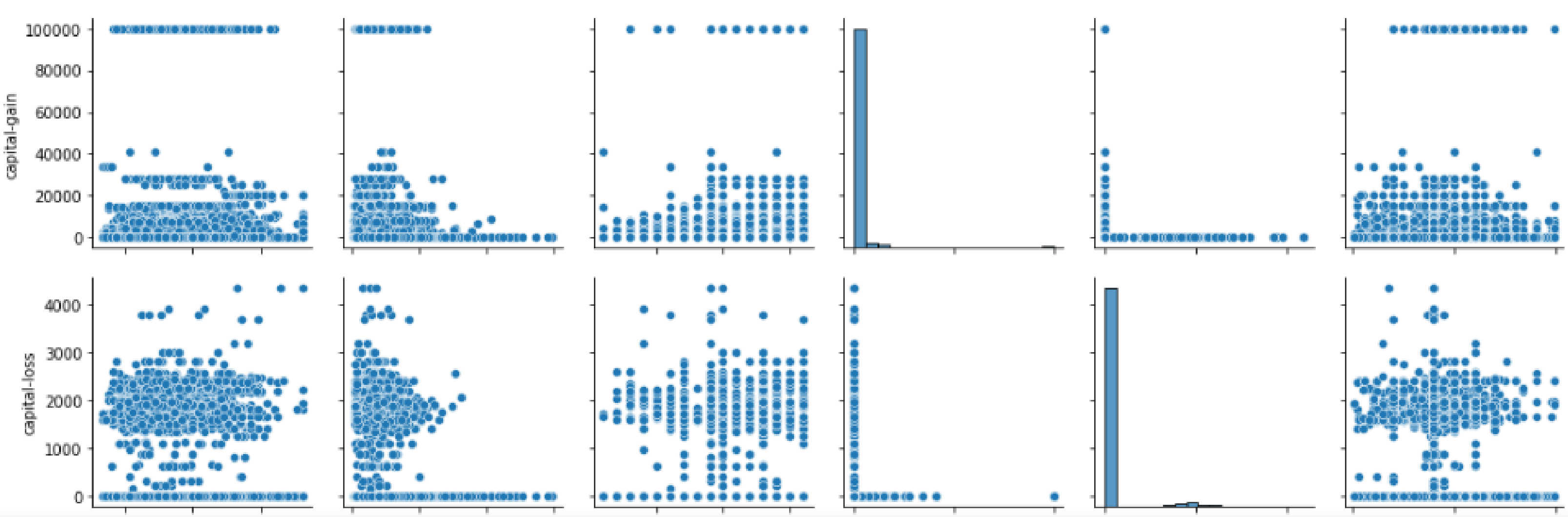
ii. Which variables would you expect to make a significant appearance in any classification model we work with?

Age and educational-num

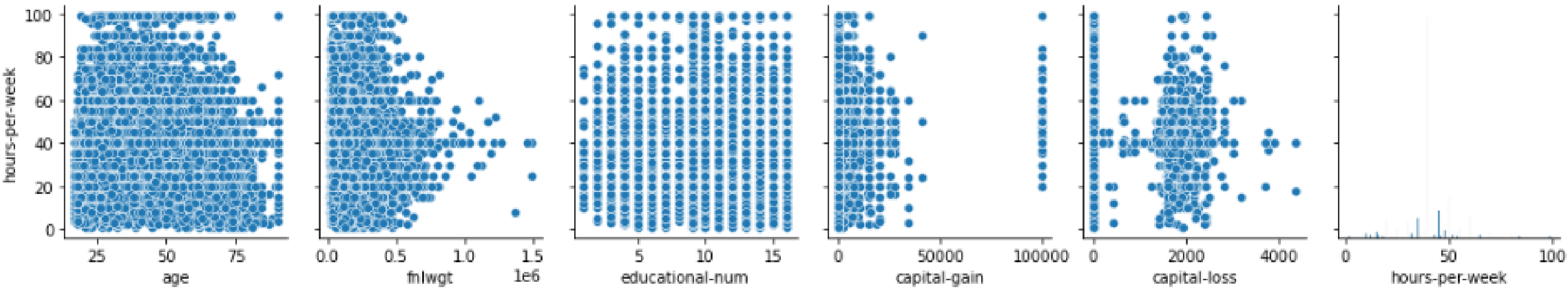
g. For each pair of numerical variables, construct a scatter plot of the variables. Discuss your salient results.

Krittamate Cherdpongtagit 6310545884





Krittamate Cherdpongtagit 6310545884



-between Age and fnlwgt has negative correlation, between age and the rest has positive

-between fnlwgt and all numerical variables has negative correlation

-between educational-num and fnlwgt has negative correlation, between educational-num and the rest has positive correlation

-between capital-gain and fnlwgt and capital-loss has negative correlation and the rest has positive

-between capital-loss and fnlwgt and capital-gain has negative correlation and the rest has positive

-between hours-per-week and fnlwgt has negative correlation and the rest has positive

Krittamate Cherdpongtagit 6310545884

2. Explore your own dataset

	Title	Genre	Director	Year	Runtime	Rating	Votes
0	Guardians of the Galaxy	Action	James Gunn	2014	121	8.1	757074
1	Prometheus	Adventure	Ridley Scott	2012	124	7.0	485820
2	Split	Horror	M. Night Shyamalan	2016	117	7.3	157606
3	Sing	Animation	Christophe Lourdelet	2016	108	7.2	60545
4	Suicide Squad	Action	David Ayer	2016	123	6.2	393727

a. Explain meaning of each attribute?

Title: title of movies
Genre : genre of movies
Director: director of movies
Year: year of release of movies
Runtime : runtime of movies
Rating: score of movies from 0-10
Votes: number of votes of movies

b. Indicate the target variable

Rating of movies

c. Explore your dataset

-Dataset.head()

```
## print the top5 records
dataset.head()
```

	Title	Genre	Director	Year	Runtime	Rating	Votes
0	Guardians of the Galaxy	Action	James Gunn	2014	121	8.1	757074
1	Prometheus	Adventure	Ridley Scott	2012	124	7.0	485820
2	Split	Horror	M. Night Shyamalan	2016	117	7.3	157606
3	Sing	Animation	Christophe Lourdelet	2016	108	7.2	60545
4	Suicide Squad	Action	David Ayer	2016	123	6.2	393727

Krittamate Cherdpongtagit 6310545884

-Don't have missing value

Missing Values

```
In [128]: ## Here we will check the percentage of nan values present in each feature
          ## 1 -step make the list of features which has missing values
          features_with_na=[features for features in dataset.columns if dataset[features].isnull().sum()>1]
          ## 2- step print the feature name and the percentage of missing values

          for feature in features_with_na:
              print(feature, np.round(dataset[feature].isnull().mean(), 4),  ' % missing values')
```

From the above dataset some of the features like Id is not required

-Numerical Variable

Numerical Variables ¶

```
In [130]: # list of numerical variables
          numerical_features = [feature for feature in dataset.columns if dataset[feature].dtypes != 'O']

          print('Number of numerical variables: ', len(numerical_features))

          # visualise the numerical variables
          dataset[numerical_features].head()

          Number of numerical variables:  4
```

Out[130]:

	Year	Runtime	Rating	Votes
0	2014	121	8.1	757074
1	2012	124	7.0	485820
2	2016	117	7.3	157606
3	2016	108	7.2	60545
4	2016	123	6.2	393727

-Target Variable

Target Variables

```
In [131]: # list of variables that contain year information
          year_feature = [feature for feature in numerical_features if 'Year' in feature]

          year_feature
Out[131]: ['Year']
```

```
In [132]: # let's explore the content of these year variables
          for feature in year_feature:
              print(feature, dataset[feature].unique())

          Year [2014 2012 2016 2015 2007 2011 2008 2006 2009 2010 2013]
```

```
In [133]: ## Lets analyze the Temporal Datetime Variables
          ## We will check whether there is a relation between year and rating

          dataset.groupby('Year')['Rating'].median().plot()
          plt.xlabel('Year')
          plt.title("Rating vs Year")
Out[133]: Text(0.5, 1.0, 'Rating vs Year')
```

Krittamate Cherdpongtagit 6310545884

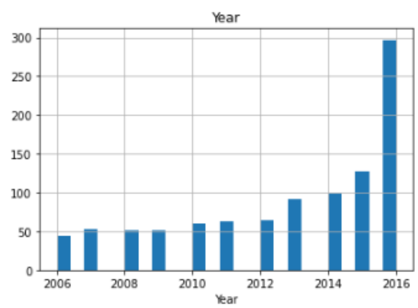
-Continuous variable

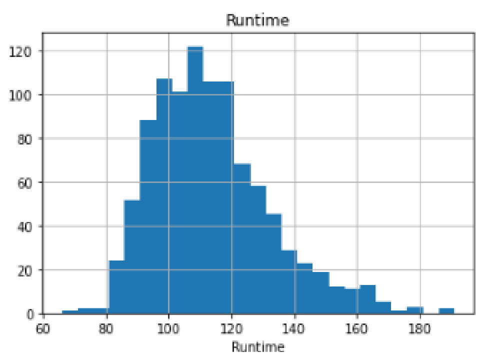
Continuous Variable

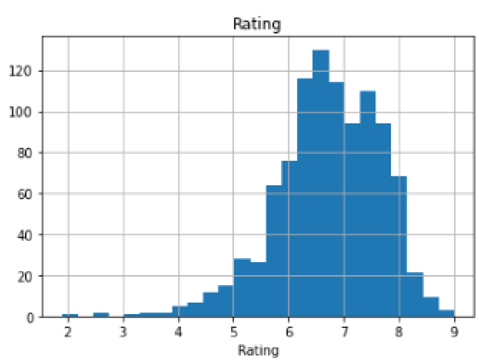
In [134]:
```
continuous_feature=[feature for feature in numerical_features]
print("Continuous feature Count {}".format(len(continuous_feature)))
```

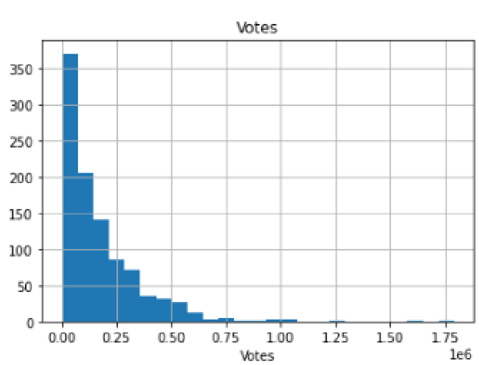
```
Continuous feature Count 4
```

In [135]:
```
## Lets analyse the continuous values by creating histograms to understand the distribution

for feature in continuous_feature:
    data=dataset.copy()
    data[feature].hist(bins=25)
    plt.xlabel(feature)
    plt.title(feature)
    plt.show()
```









Krittamate Cherdpongtagit 6310545884

-Categorical variable

Categorical Variables

In [136]:
```
categorical_features=[feature for feature in dataset.columns if data[feature].dtypes=='O']
categorical_features
```

Out[136]: ['Title', 'Genre', 'Director']

In [137]:
```
dataset[categorical_features].head()
```

Out[137]:

	Title	Genre	Director
0	Guardians of the Galaxy	Action	James Gunn
1	Prometheus	Adventure	Ridley Scott
2	Split	Horror	M. Night Shyamalan
3	Sing	Animation	Christophe Lourdelet
4	Suicide Squad	Action	David Ayer

In [138]:
```
for feature in categorical_features:
    print('The feature is {} and number of categories are {}'.format(feature,len(dataset[feature].unique())))
```

```
The feature is Title and number of categories are 999
The feature is Genre and number of categories are 13
The feature is Director and number of categories are 644
```

- Find out the relationship between categorical variable and dependent feature Rating

Krittamate Cherdpongtagit 6310545884

Find out the relationship between categorical variable and dependent feature Rating

In [139]:
```
for feature in categorical_features:
    data=dataset.copy()
    data.groupby(feature)['Rating'].median().plot.bar()
    plt.xlabel(feature)
    plt.title(feature)
    plt.show()
```