Analyse of Adult dataset.nbconvert

January 19, 2020

1 Analyse dataset

This notebook main goal is to : - Understand what is in the data : plot variables one by one, missing values, etc. - See which data are correlated

1.1 Load packages

[1]:
```
import pandas as pd
import numpy as np

from IPython.display import display, Markdown

# transparentai package : https://github.com/Nathanlauga/transparentai
import transparentai.explore as explore
from transparentai.utils import remove_var_with_one_value
```

1.2 Load dataset

[2]:
```
dataset = pd.read_csv('../data/adult.csv', sep=',')
```

[3]:
```
target = 'income'
target = None if target not in dataset.columns else target
```

1.3 Quick overview

[4]:
```
display(Markdown(f'#### {dataset.shape}'))
display(dataset.head())
```

(48842, 15)

```
   age  workclass  fnlwgt     education  educational-num      marital-status  \
0   25    Private  226802          11th                7       Never-married
1   38    Private   89814       HS-grad                9  Married-civ-spouse
2   28  Local-gov  336951    Assoc-acdm               12  Married-civ-spouse
3   44    Private  160323  Some-college               10  Married-civ-spouse
4   18          ?  103497  Some-college               10       Never-married
```

```
         occupation relationship   race  gender  capital-gain  capital-loss  \
0  Machine-op-inspct    Own-child  Black    Male             0             0
1    Farming-fishing      Husband  White    Male             0             0
2    Protective-serv      Husband  White    Male             0             0
3  Machine-op-inspct      Husband  Black    Male          7688             0
4                  ?    Own-child  White  Female             0             0

   hours-per-week native-country income
0              40  United-States  <=50K
1              50  United-States  <=50K
2              40  United-States   >50K
3              40  United-States   >50K
4              30  United-States  <=50K
```

1.4 Analyse : missing values

[5]:
```python
display(Markdown('#### Missing values for adult dataset'))
explore.show_missing_values(dataset)
```

Missing values for adult dataset No missing value.

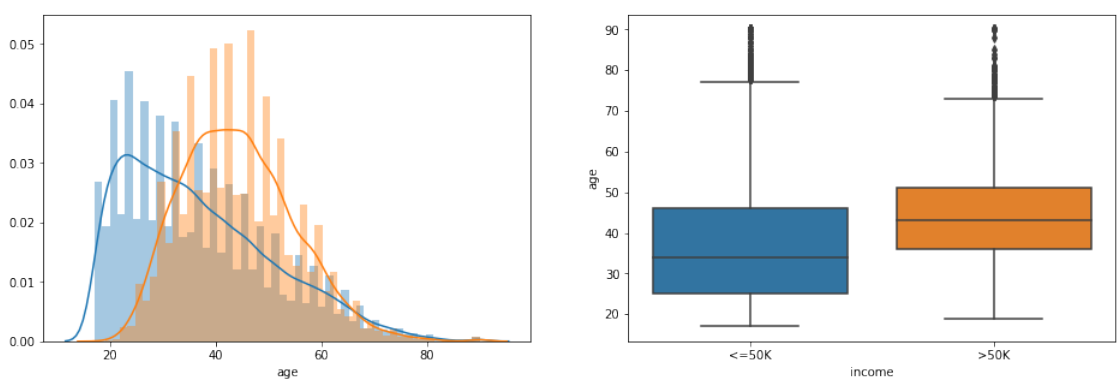
1.5 Analyse : each variable

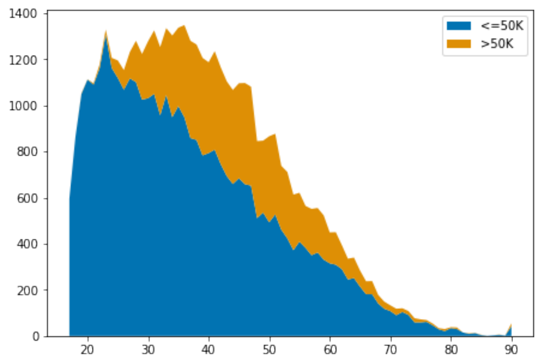
[6]:
```python
dataset = remove_var_with_one_value(dataset)
```

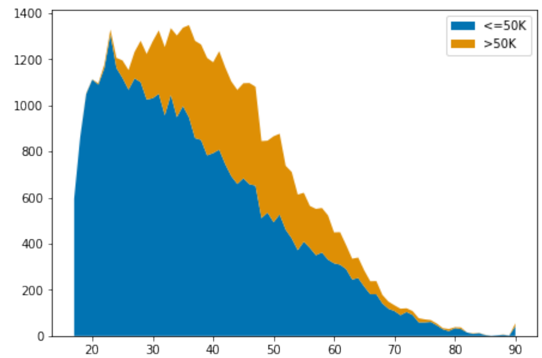
[7]:
```python
explore.show_df_vars(df=dataset, target=target)
```

1.5.1 Numerical variables

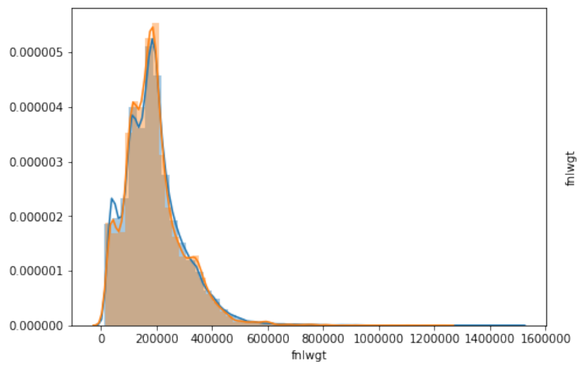
age : 0 nulls, 74 unique vals, most common: {36: 1348, 35: 1337}

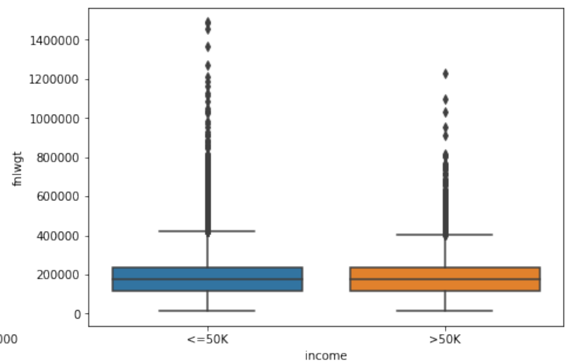


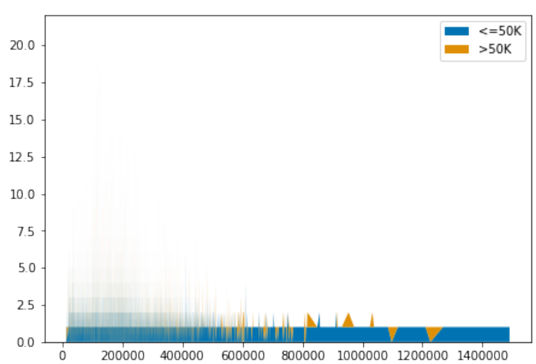


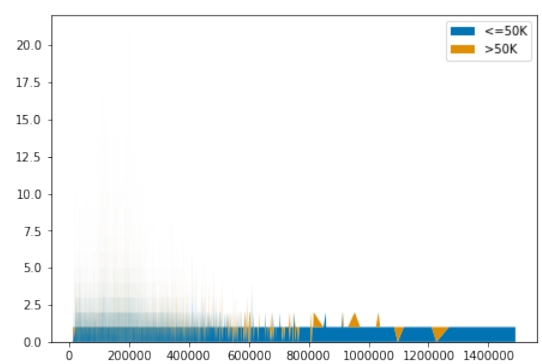


fnlwgt : 0 nulls, 28523 unique vals, most common: {203488: 21, 190290: 19}

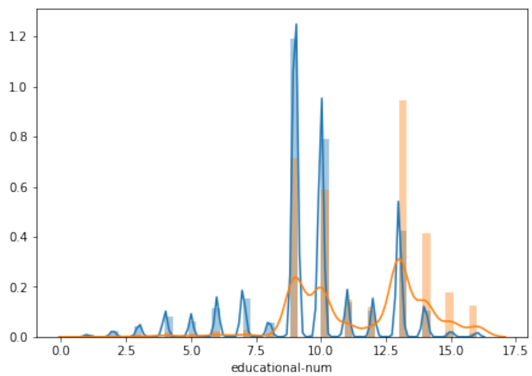


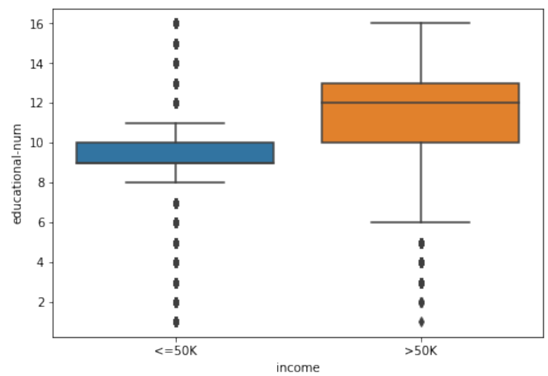


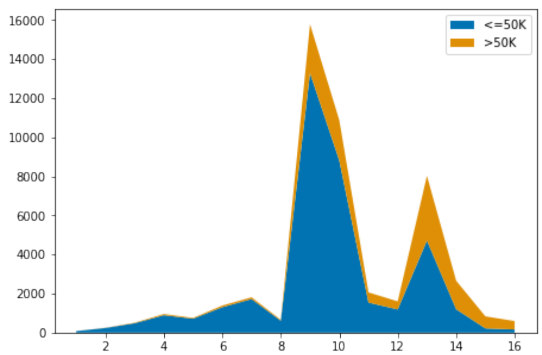


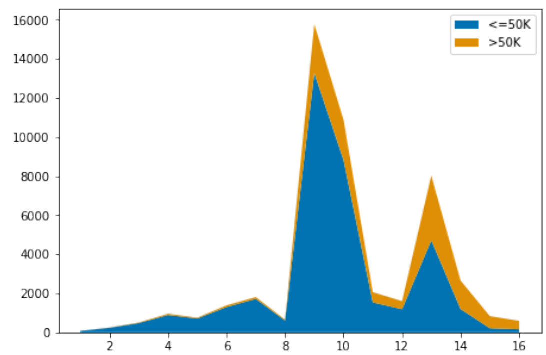


educational-num : 0 nulls, 16 unique vals, most common: {9: 15784, 10: 10878}

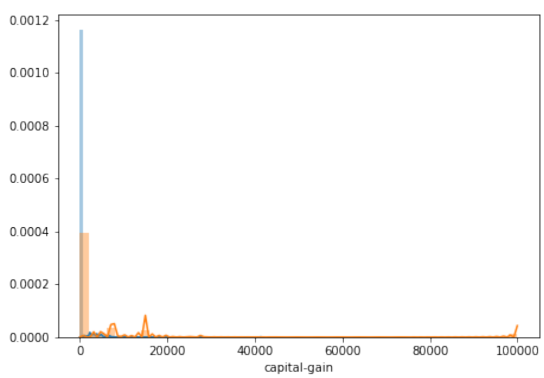


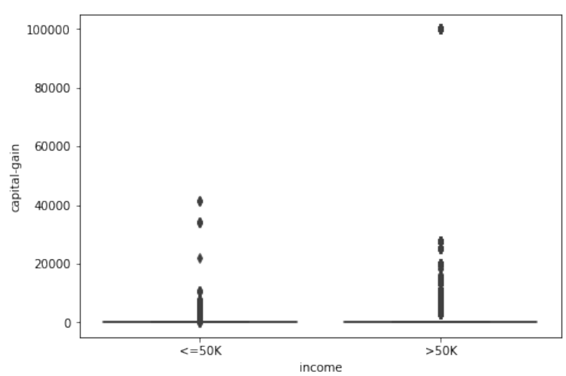


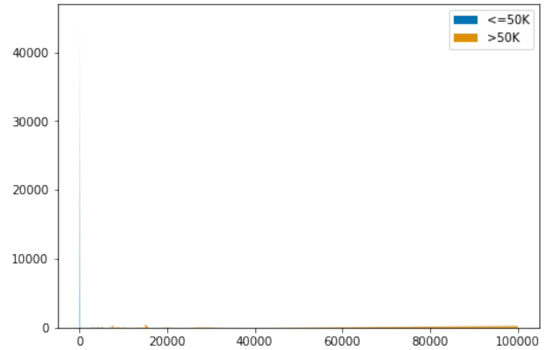


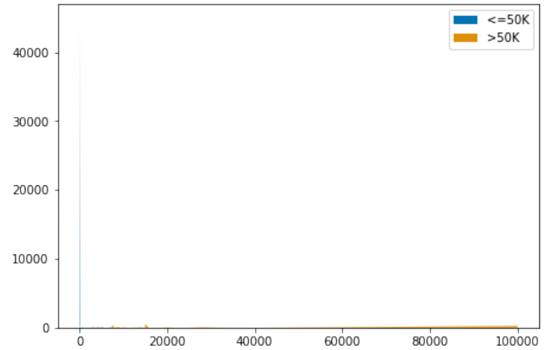


capital-gain : 0 nulls, 123 unique vals, most common: {0: 44807, 15024: 513}

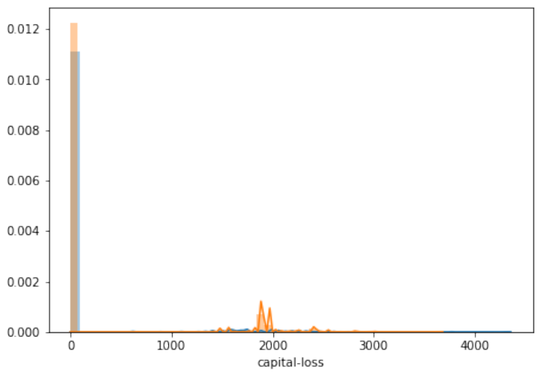


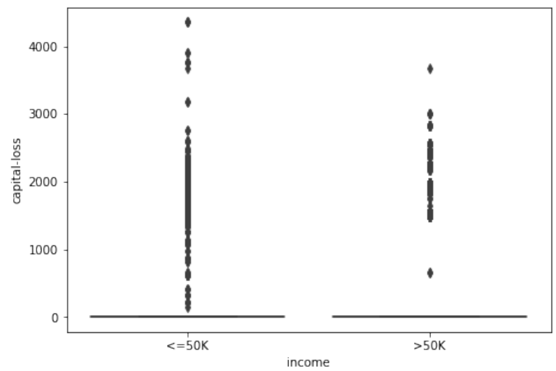


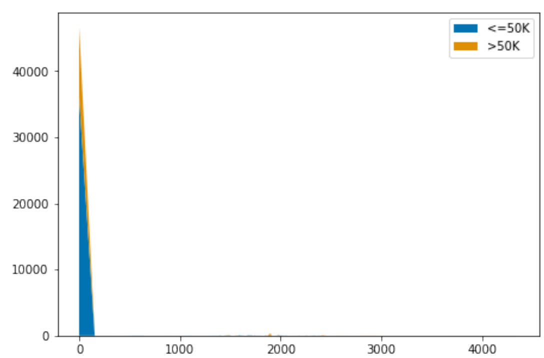


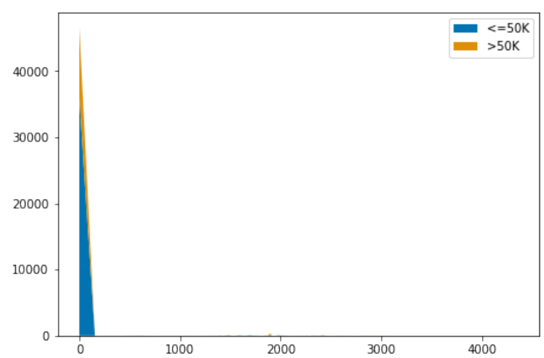


capital-loss : 0 nulls, 99 unique vals, most common: {0: 46560, 1902: 304}

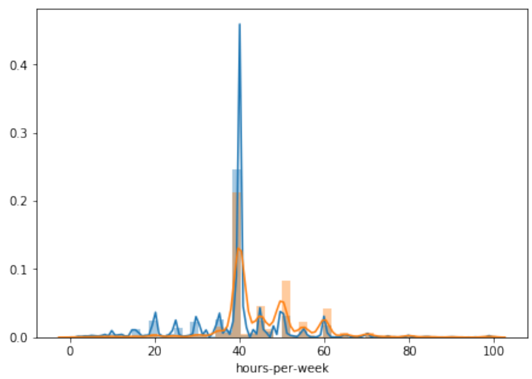


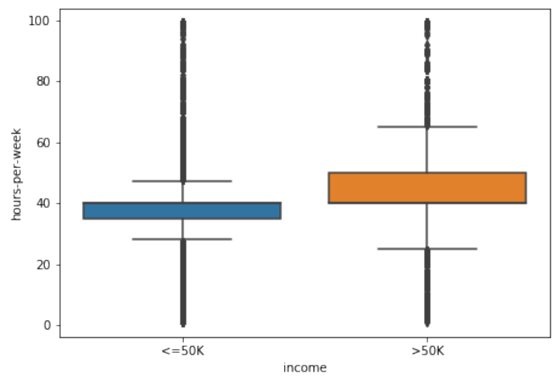


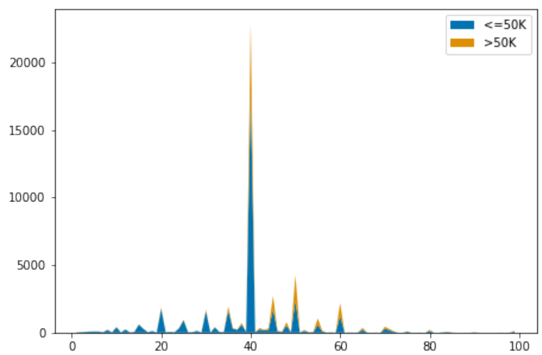


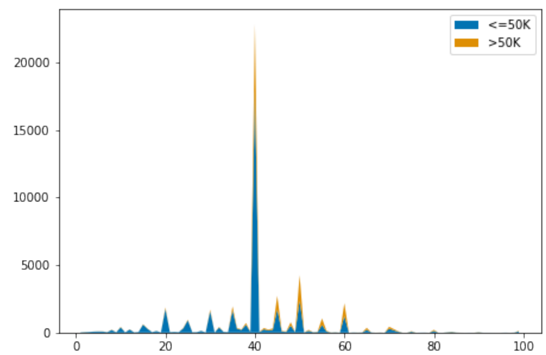


hours-per-week : 0 nulls, 96 unique vals, most common: {40: 22803, 50: 4246}



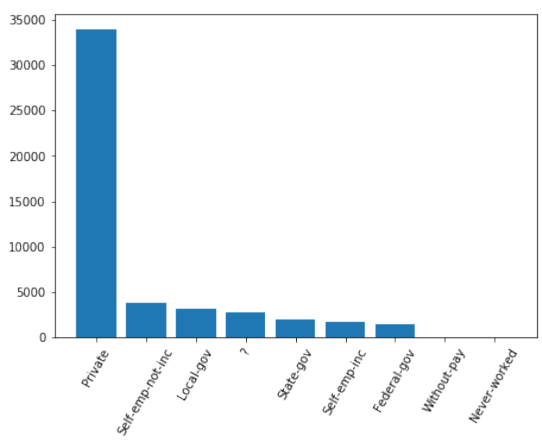


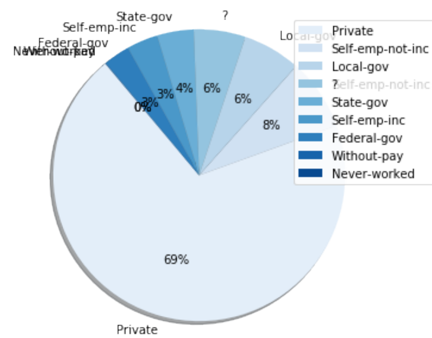


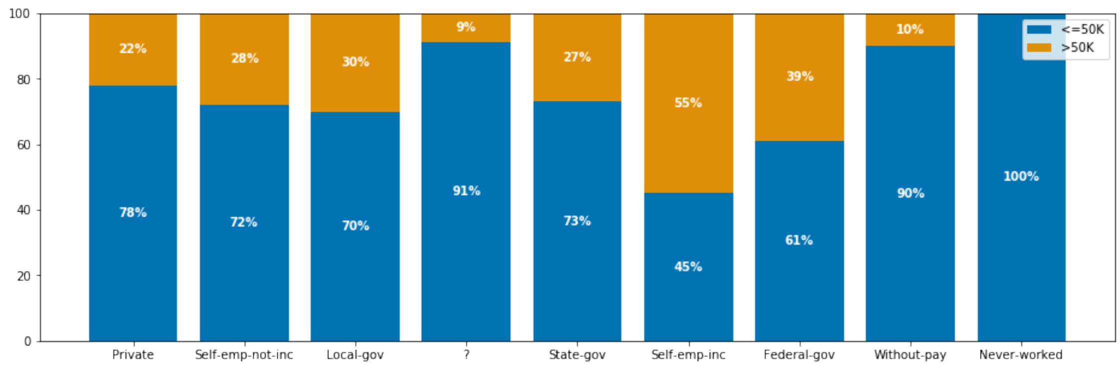


1.5.2 Categorical variables

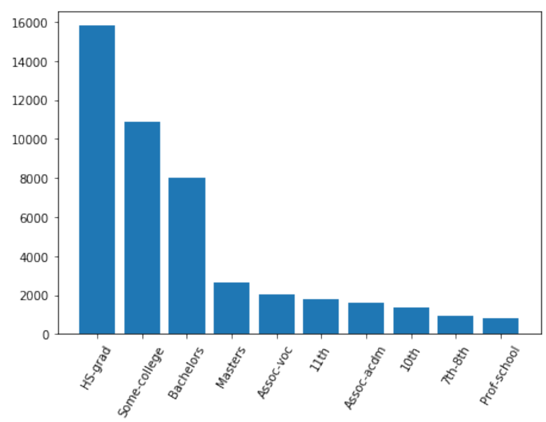
workclass : 0 nulls, 9 unique vals, most common: {'Private': 33906, 'Self-emp-not-inc': 3862}

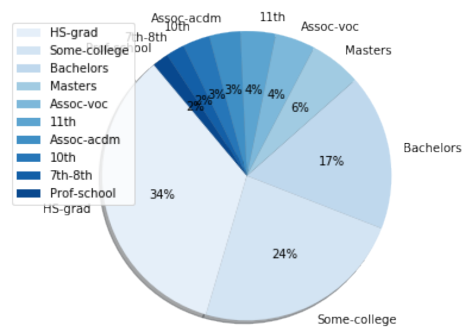


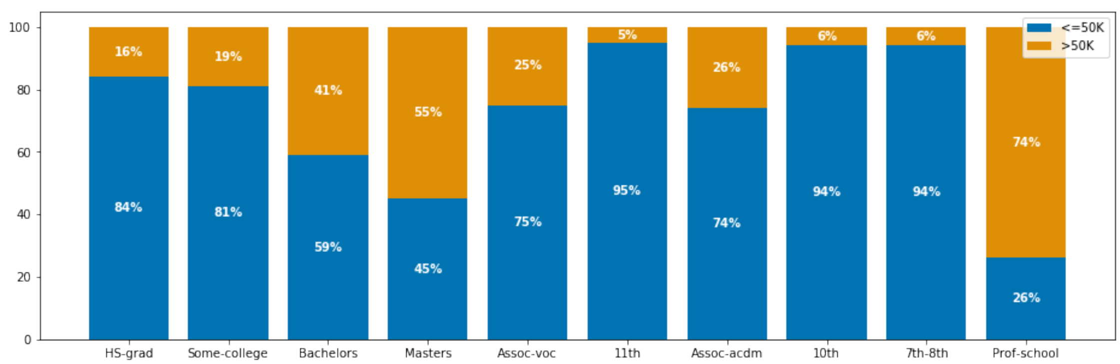




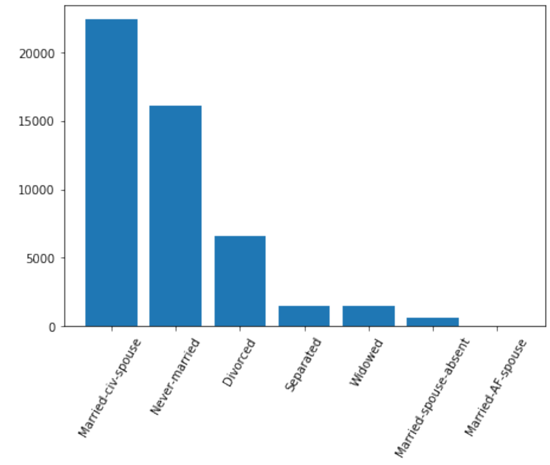
education : 0 nulls, 16 unique vals, most common: {'HS-grad': 15784, 'Some-college': 10878}

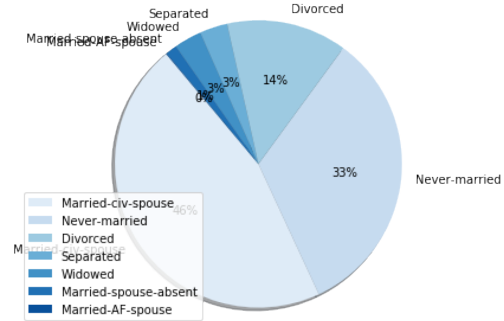


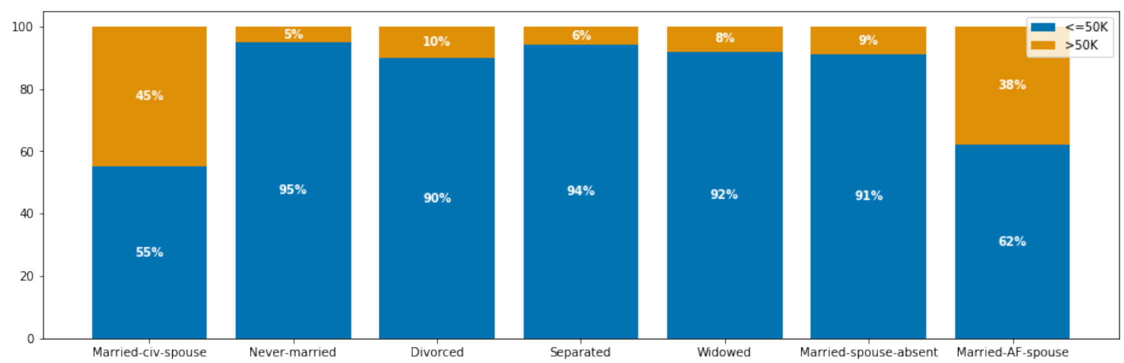




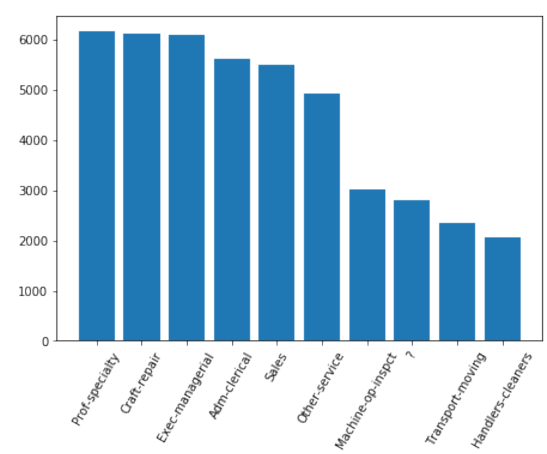
marital-status : 0 nulls, 7 unique vals, most common: {'Married-civ-spouse': 22379, 'Never-married': 16117}

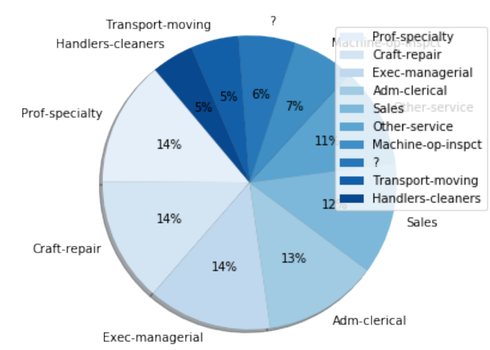


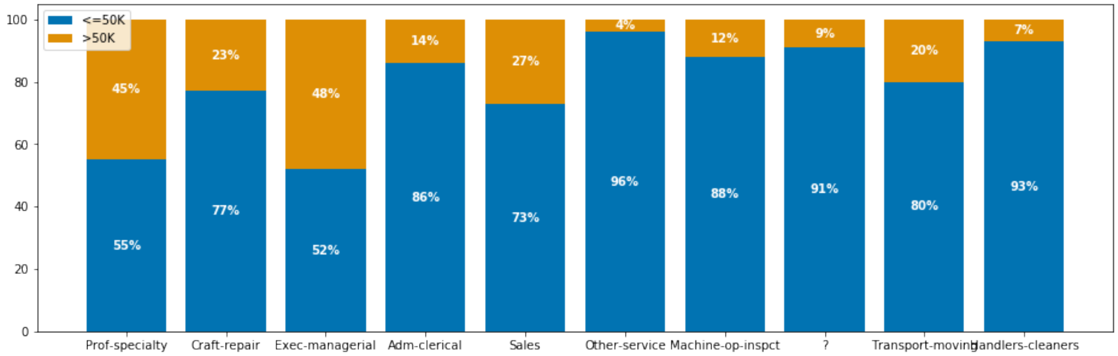




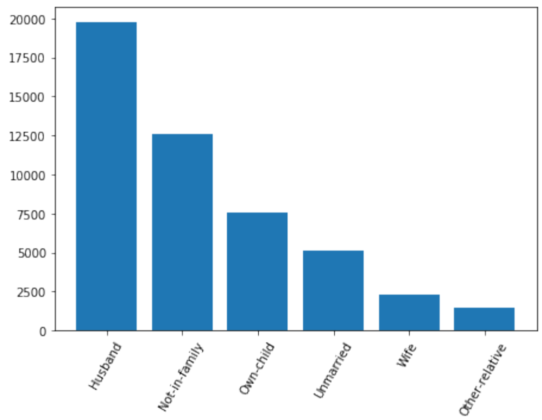
occupation : 0 nulls, 15 unique vals, most common: {'Prof-specialty': 6172, 'Craft-repair': 6112}

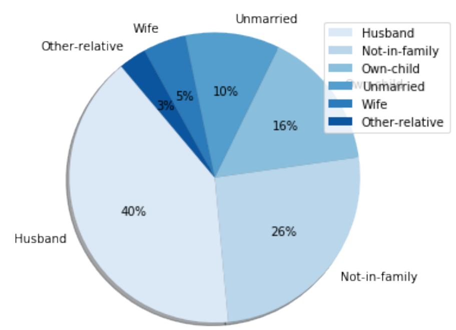


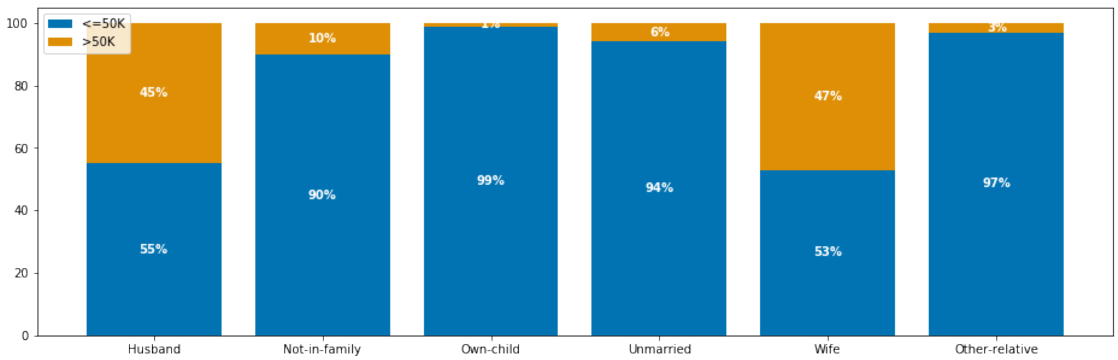




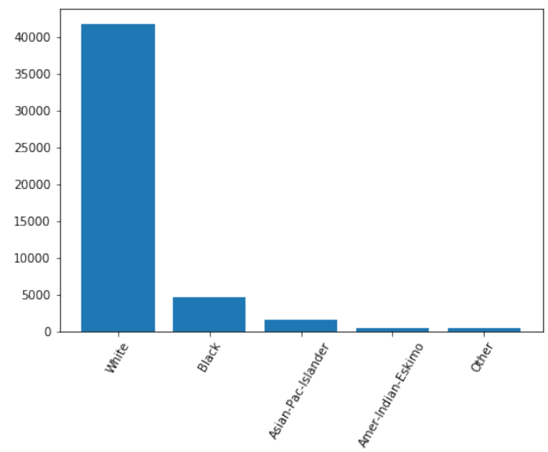
relationship : 0 nulls, 6 unique vals, most common: {'Husband': 19716, 'Not-in-family': 12583}

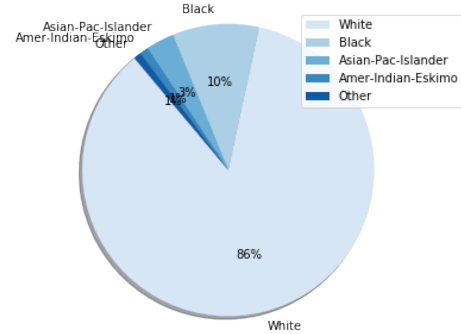


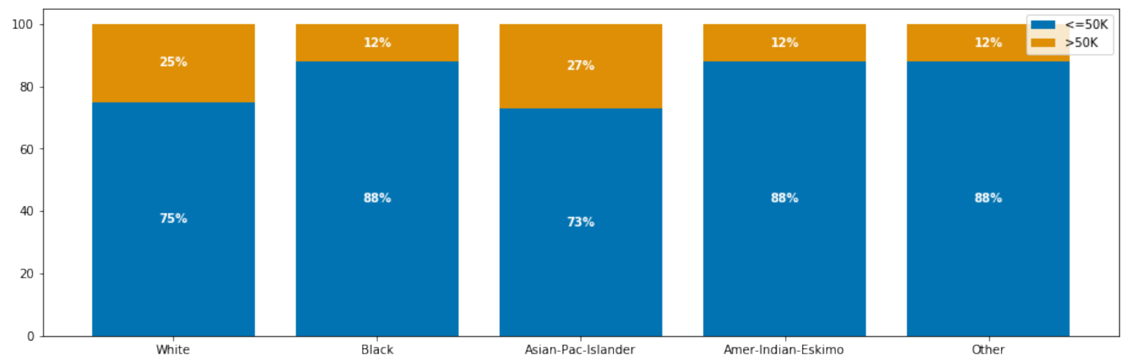




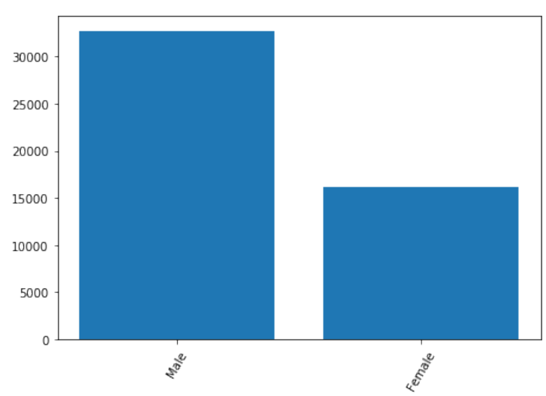
race : 0 nulls, 5 unique vals, most common: {'White': 41762, 'Black': 4685}

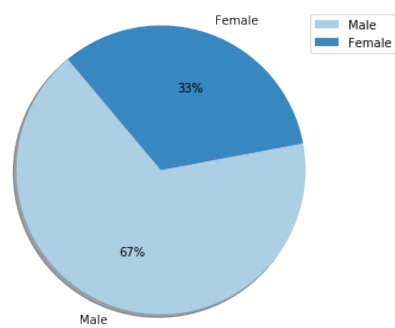


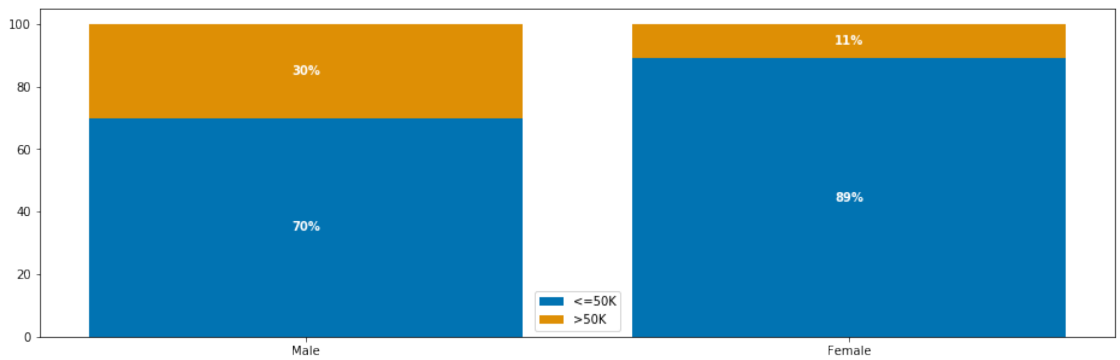




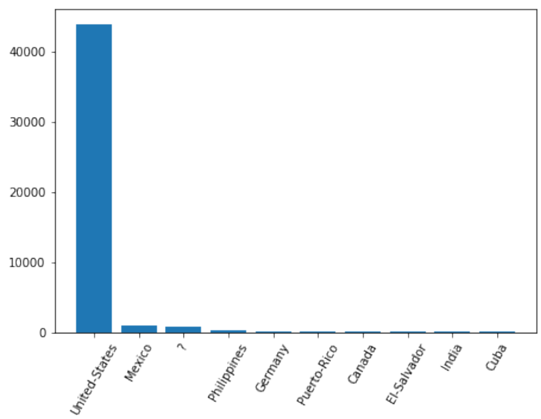
gender : 0 nulls, 2 unique vals, most common: {'Male': 32650, 'Female': 16192}

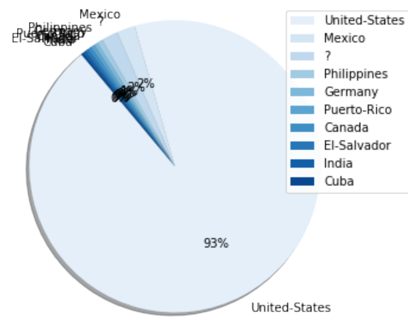


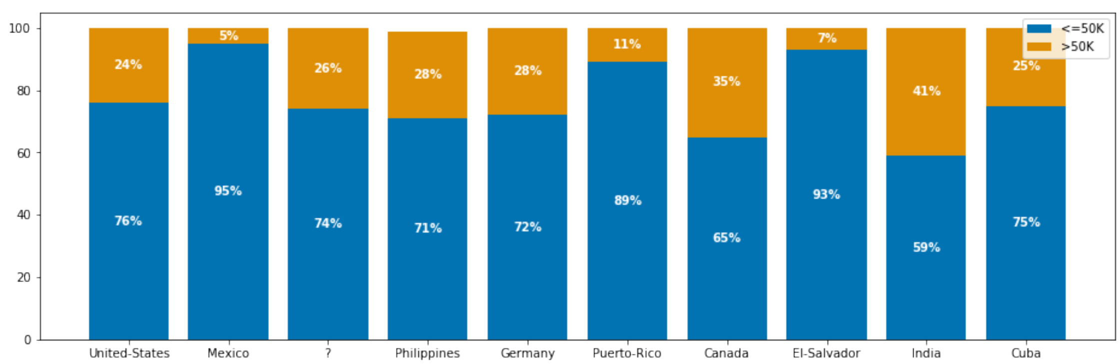




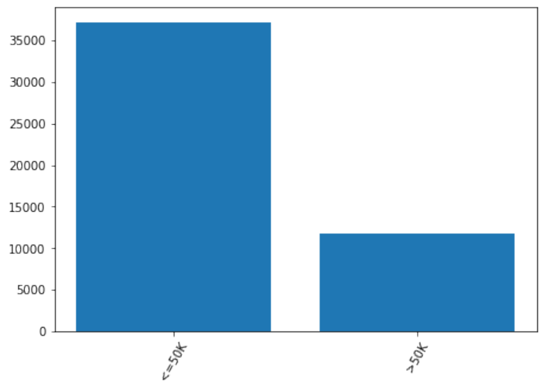
native-country : 0 nulls, 42 unique vals, most common: {'United-States': 43832, 'Mexico': 951}

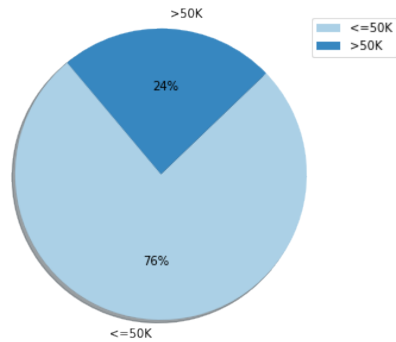


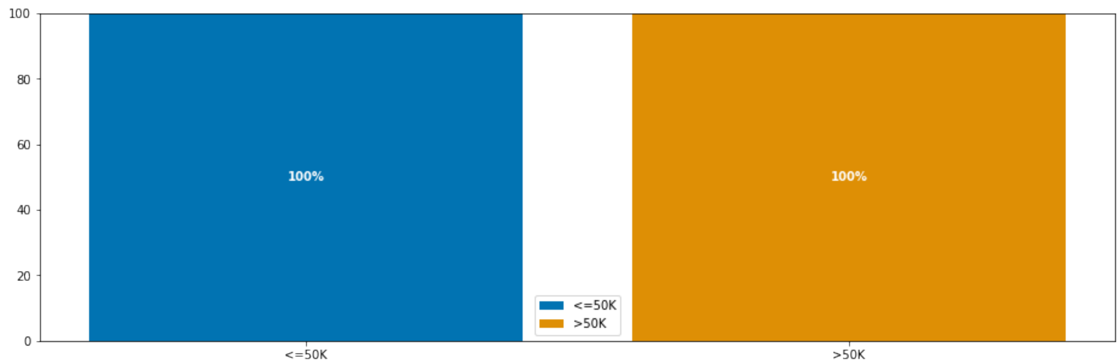




income : 0 nulls, 2 unique vals, most common: {'<=50K': 37155, '>50K': 11687}





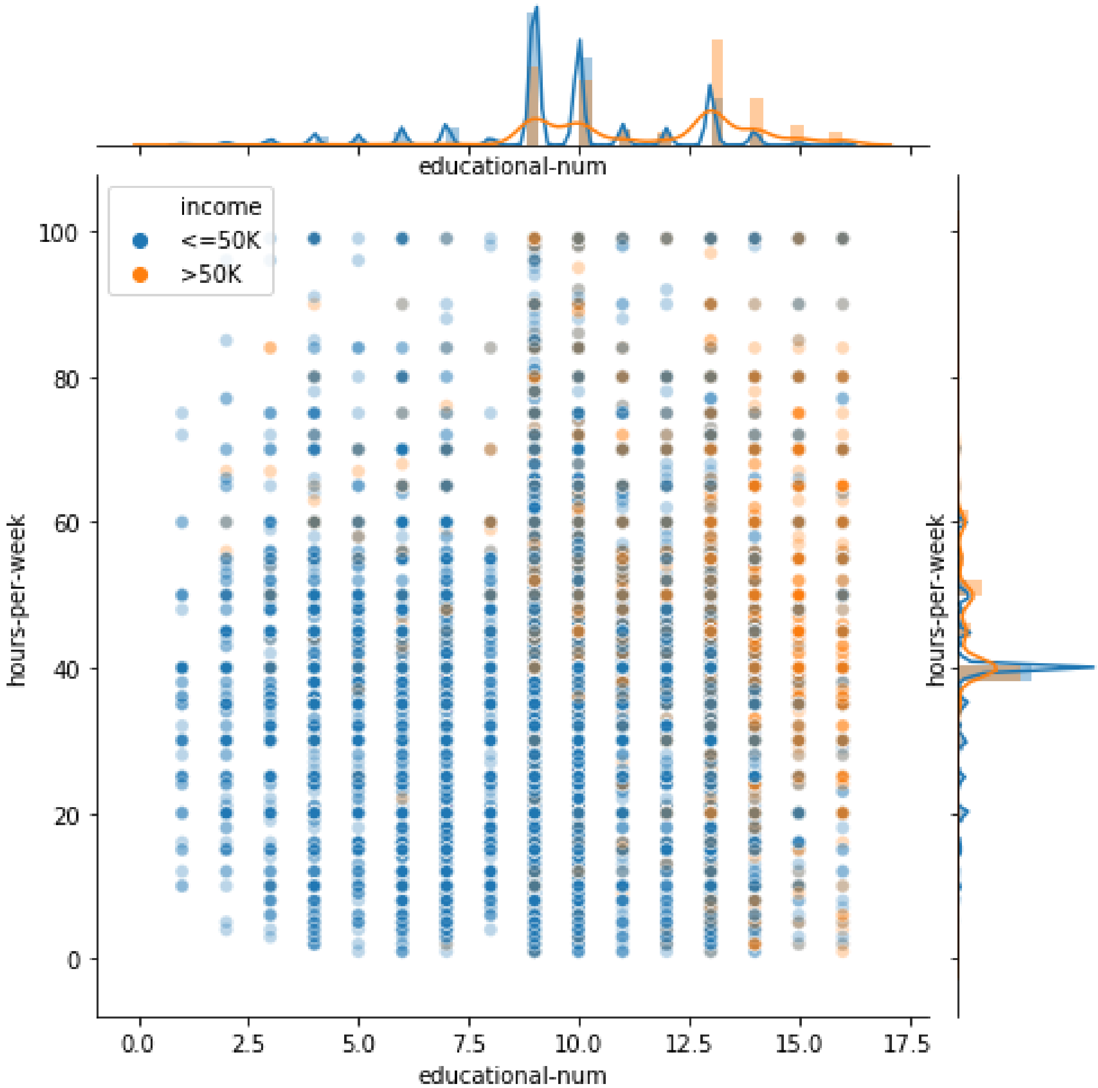


1.5.3 Datetime variables

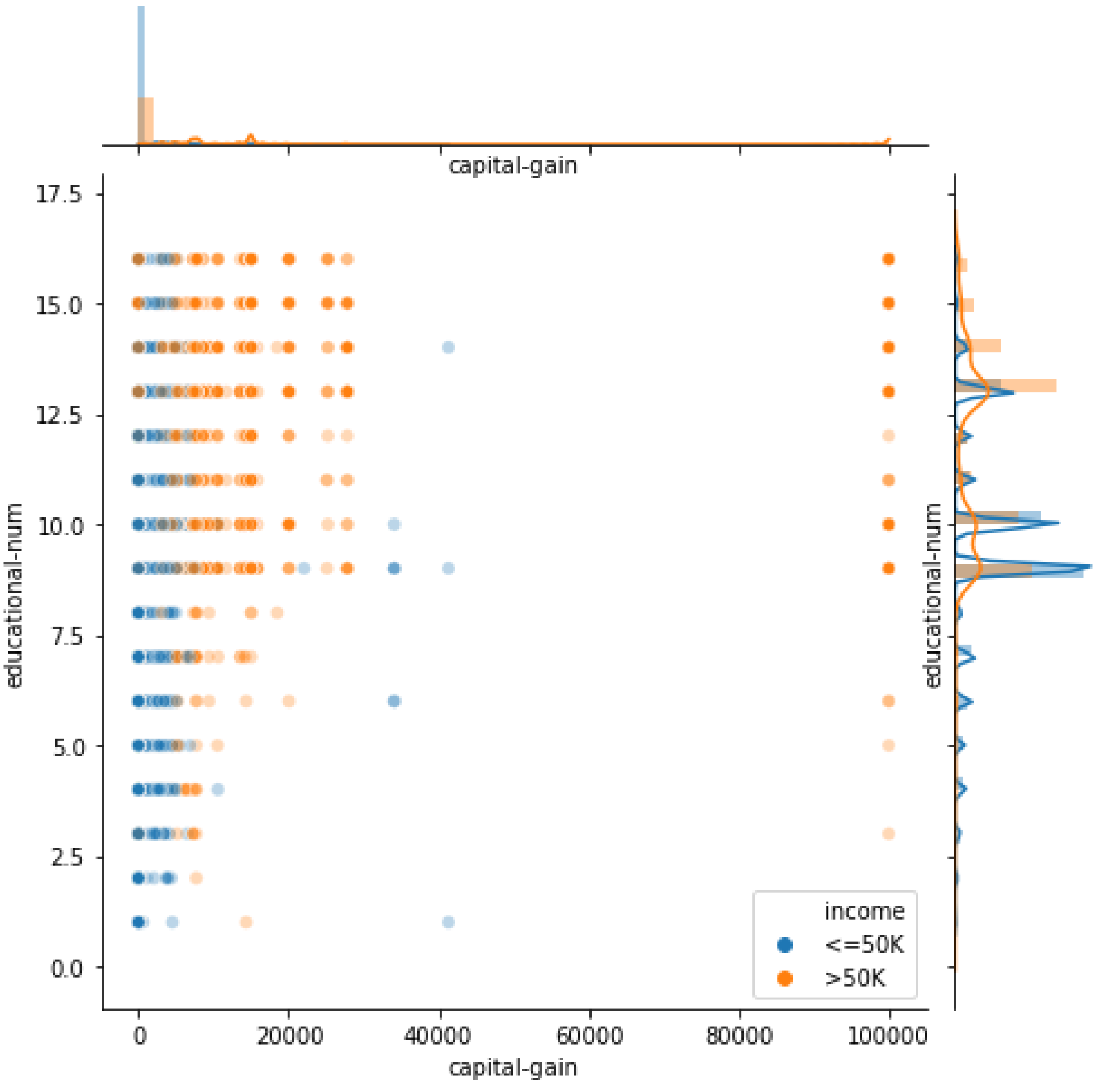
1.6 Analyse : numericals variables relations

[8]:
```
explore.show_df_numerical_relations(df=dataset, target=target)
```

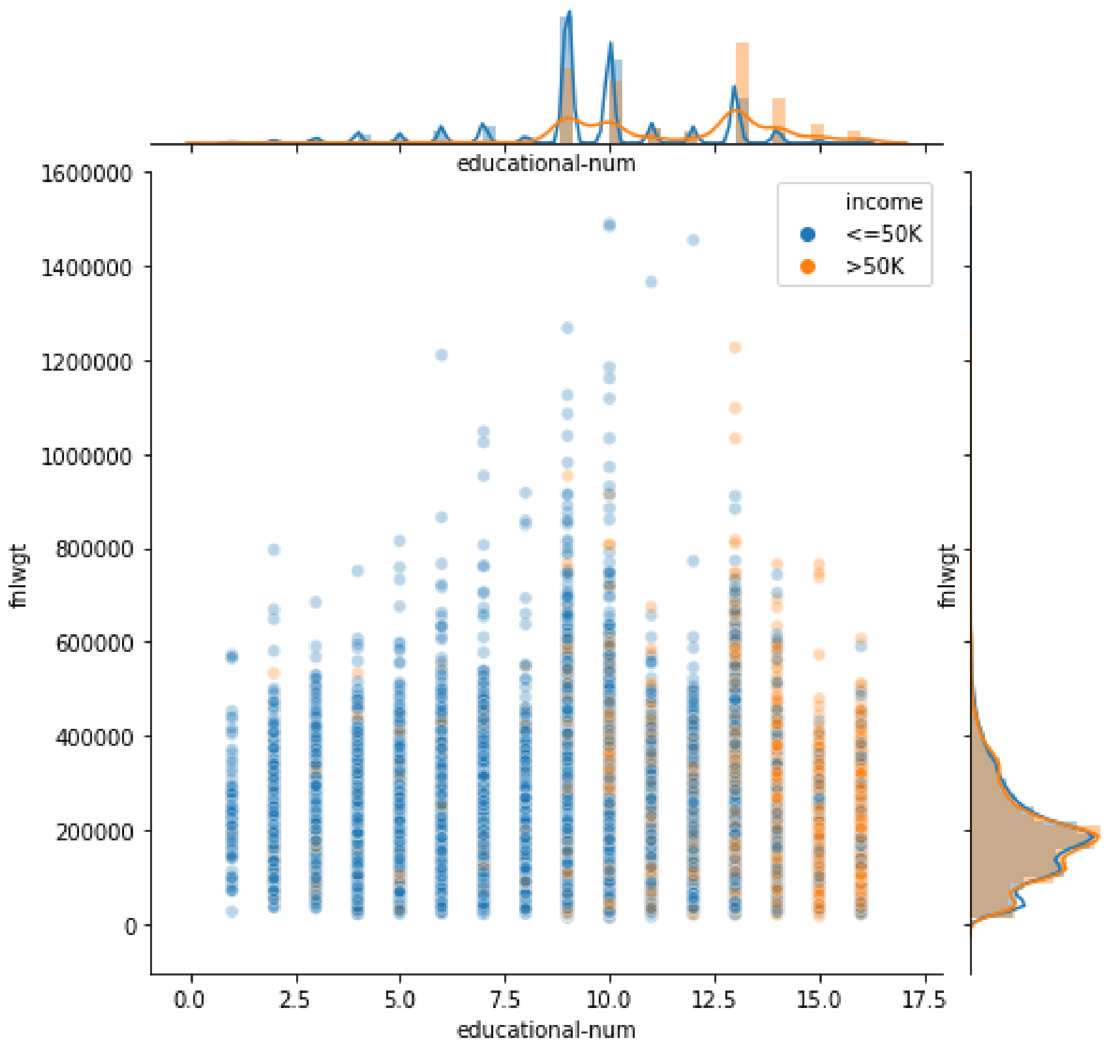
Joint plot for **educational-num** & **hours-per-week**



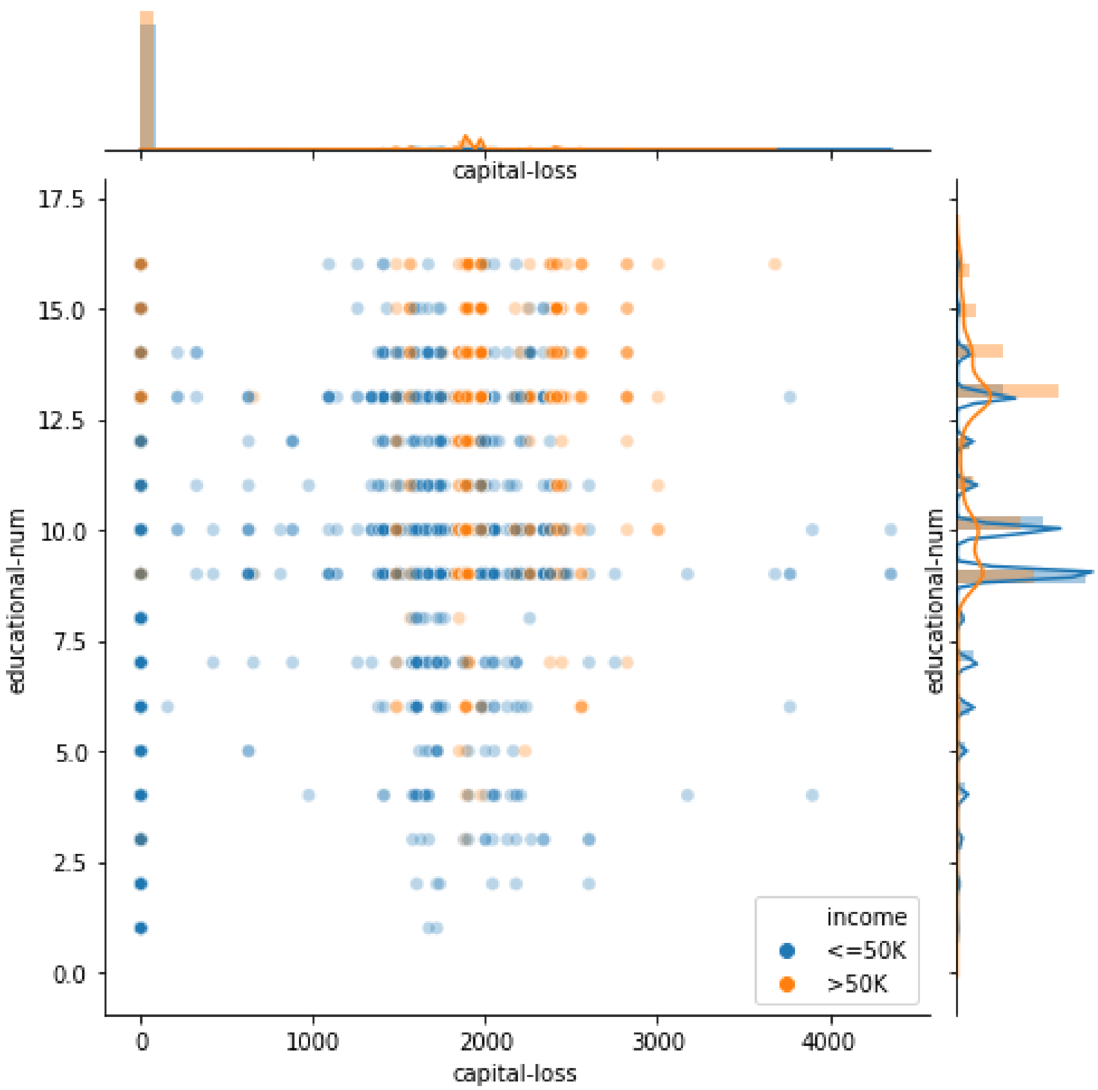
Joint plot for **capital-gain** & **educational-num**



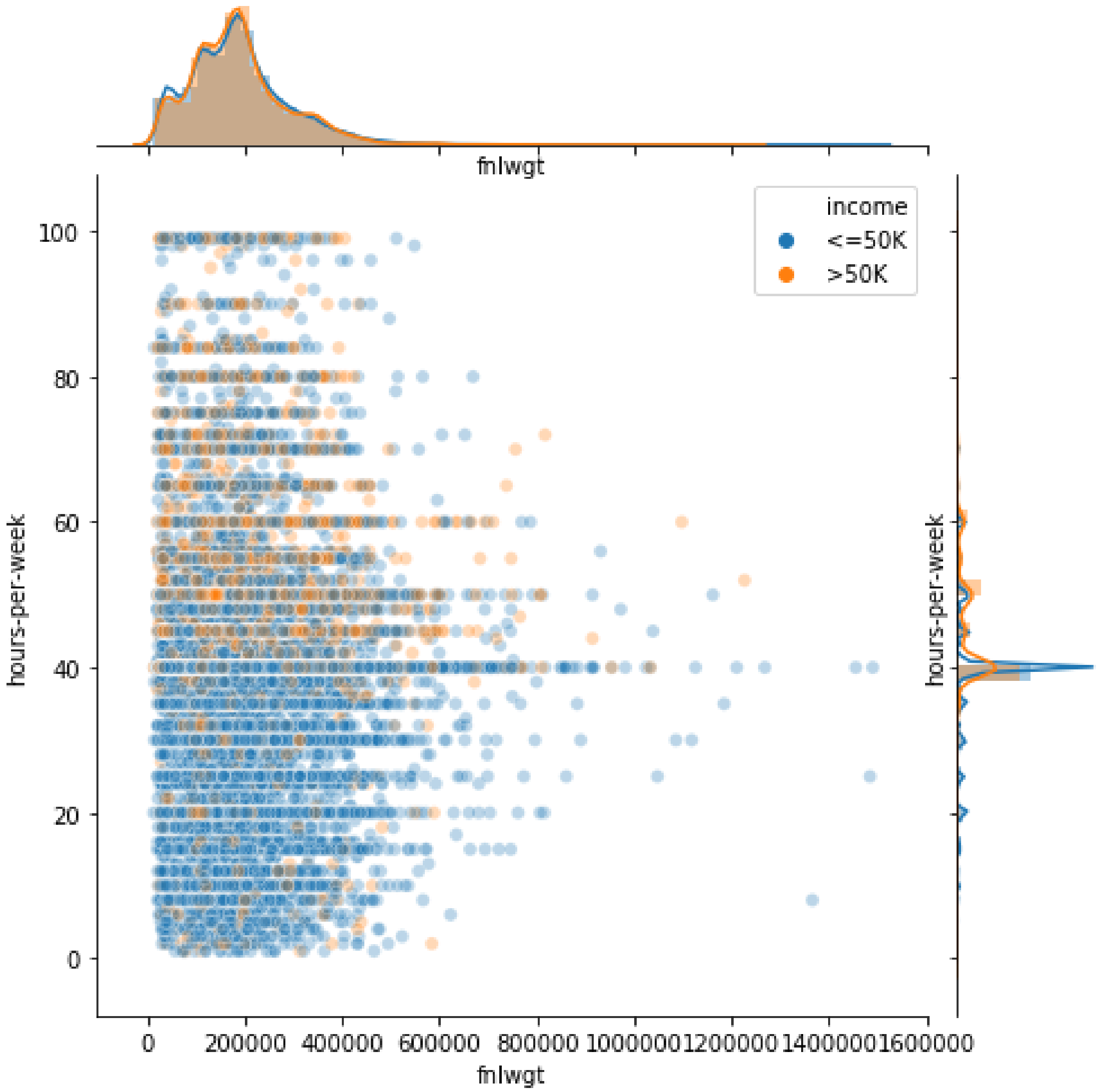
Joint plot for **educational-num** & **fnlwgt**



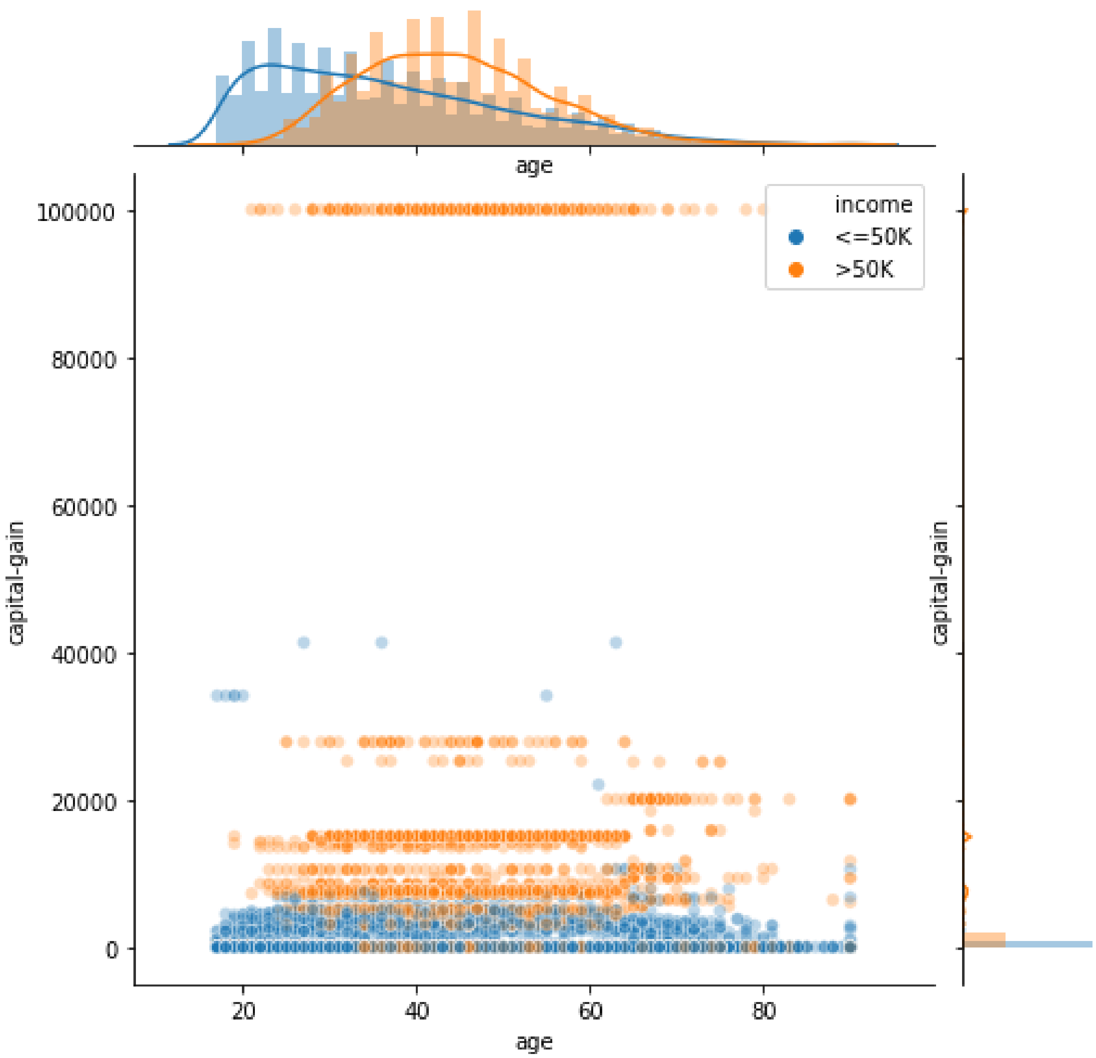
Joint plot for **capital-loss** & **educational-num**



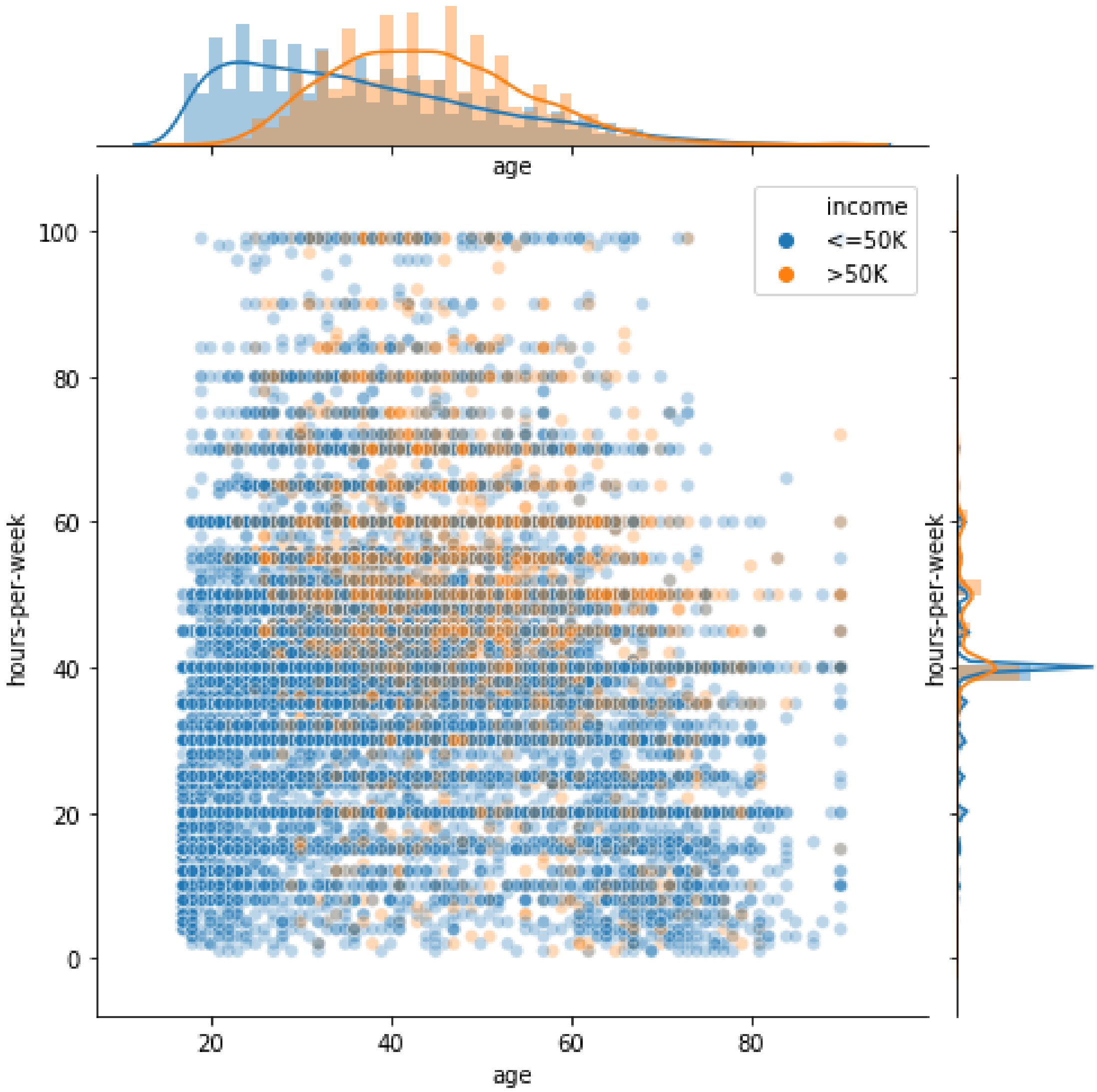
Joint plot for **fnlwgt** & **hours-per-week**



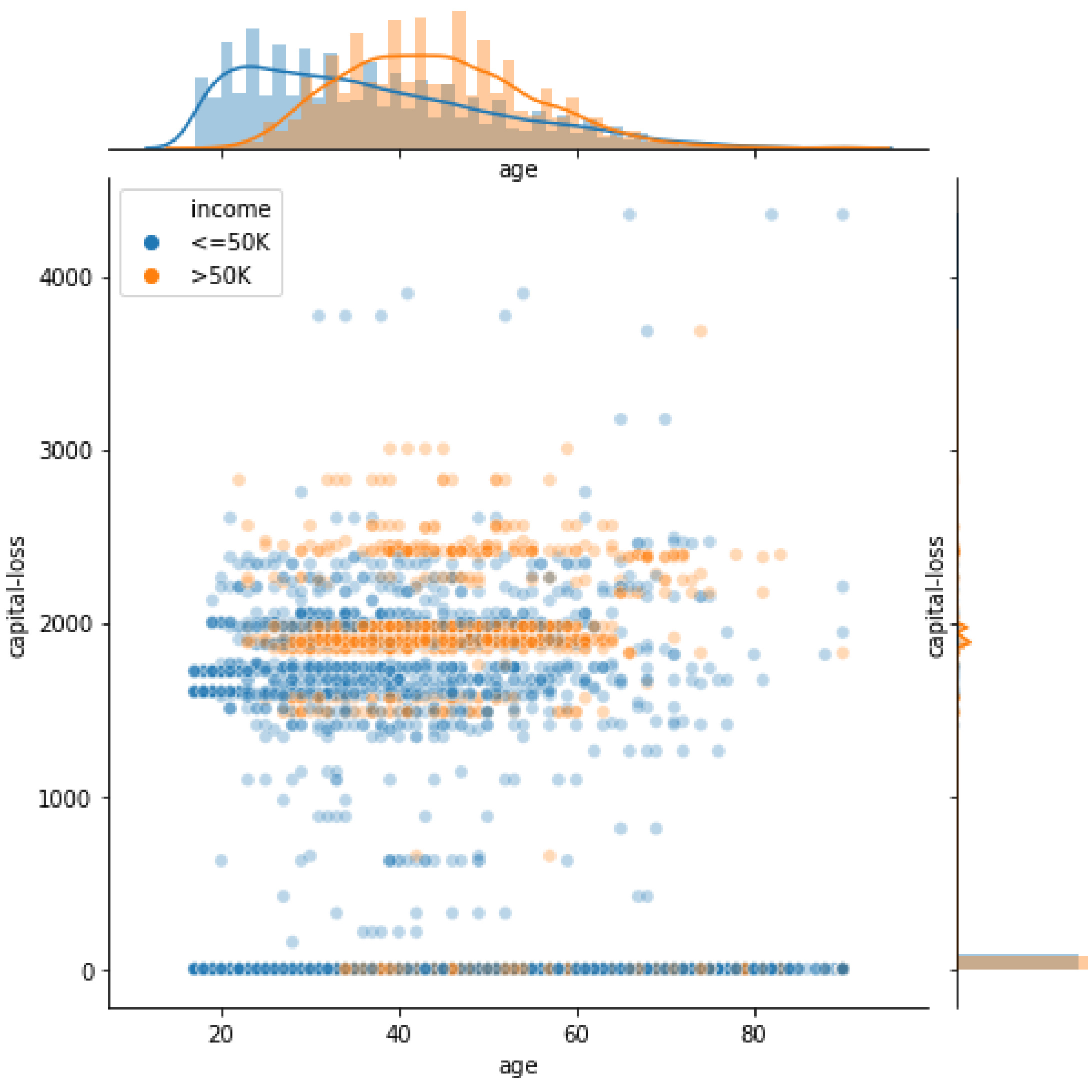
Joint plot for **age** & **capital-gain**



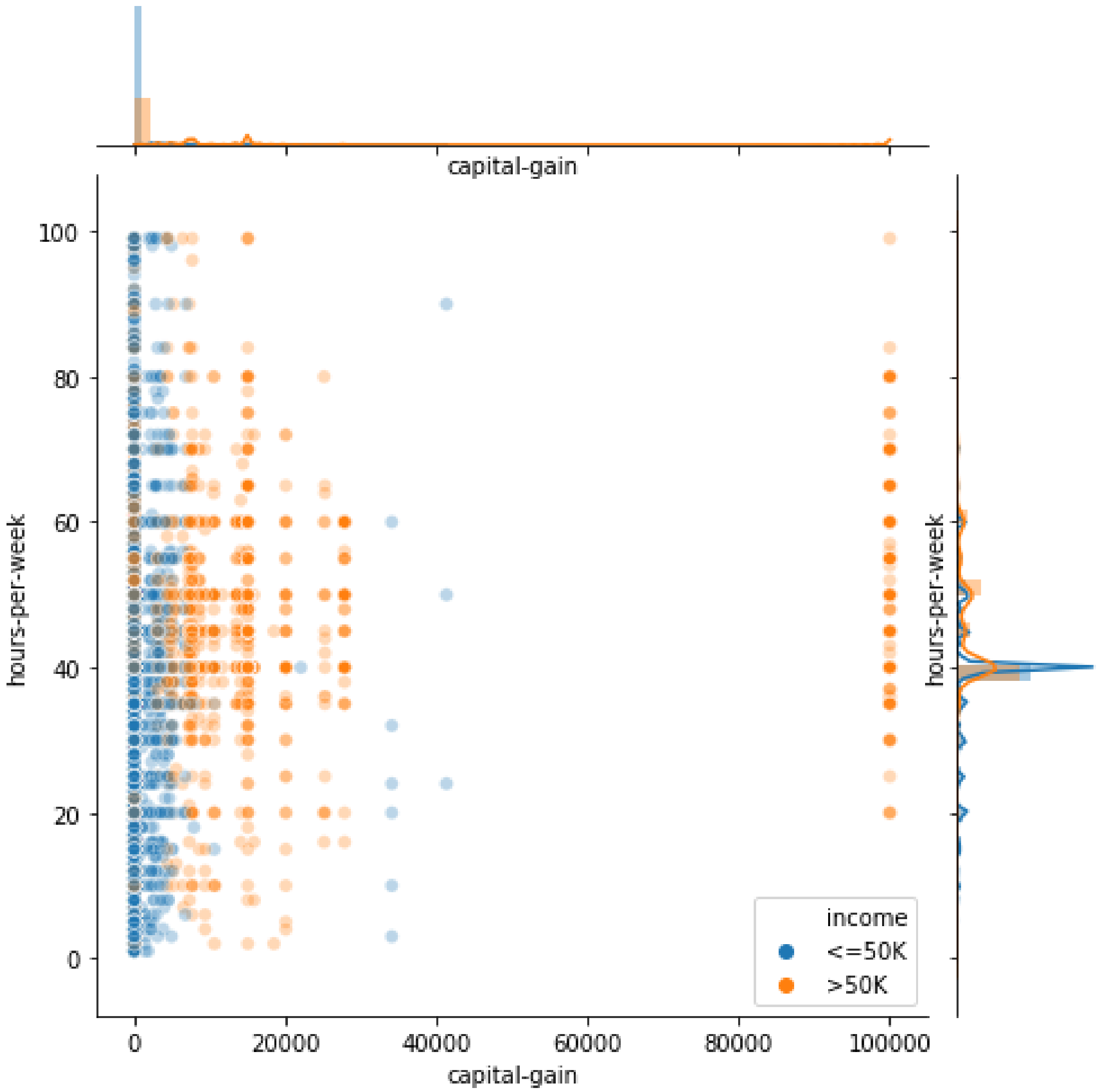
Joint plot for **age** & **hours-per-week**



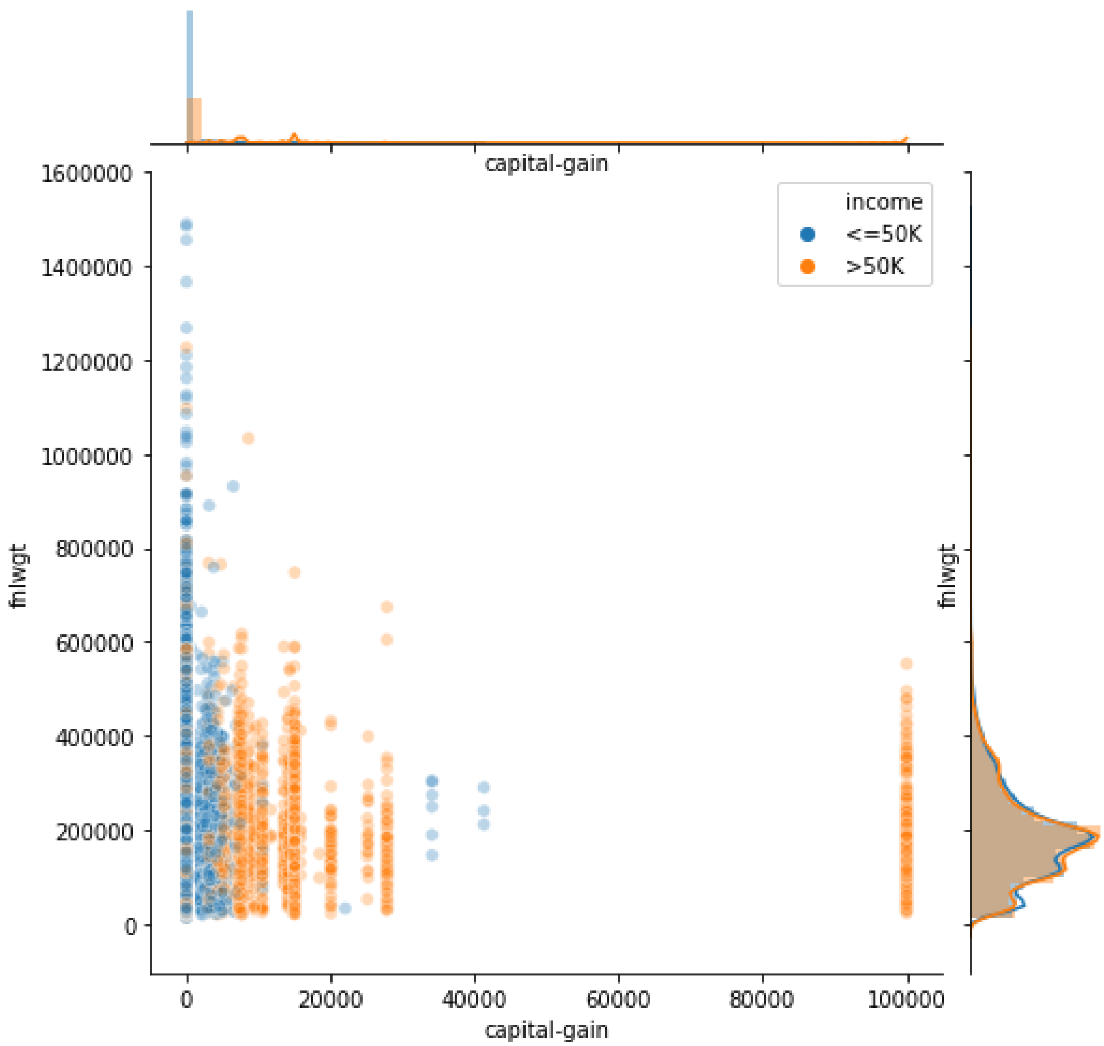
Joint plot for **age** & **capital-loss**



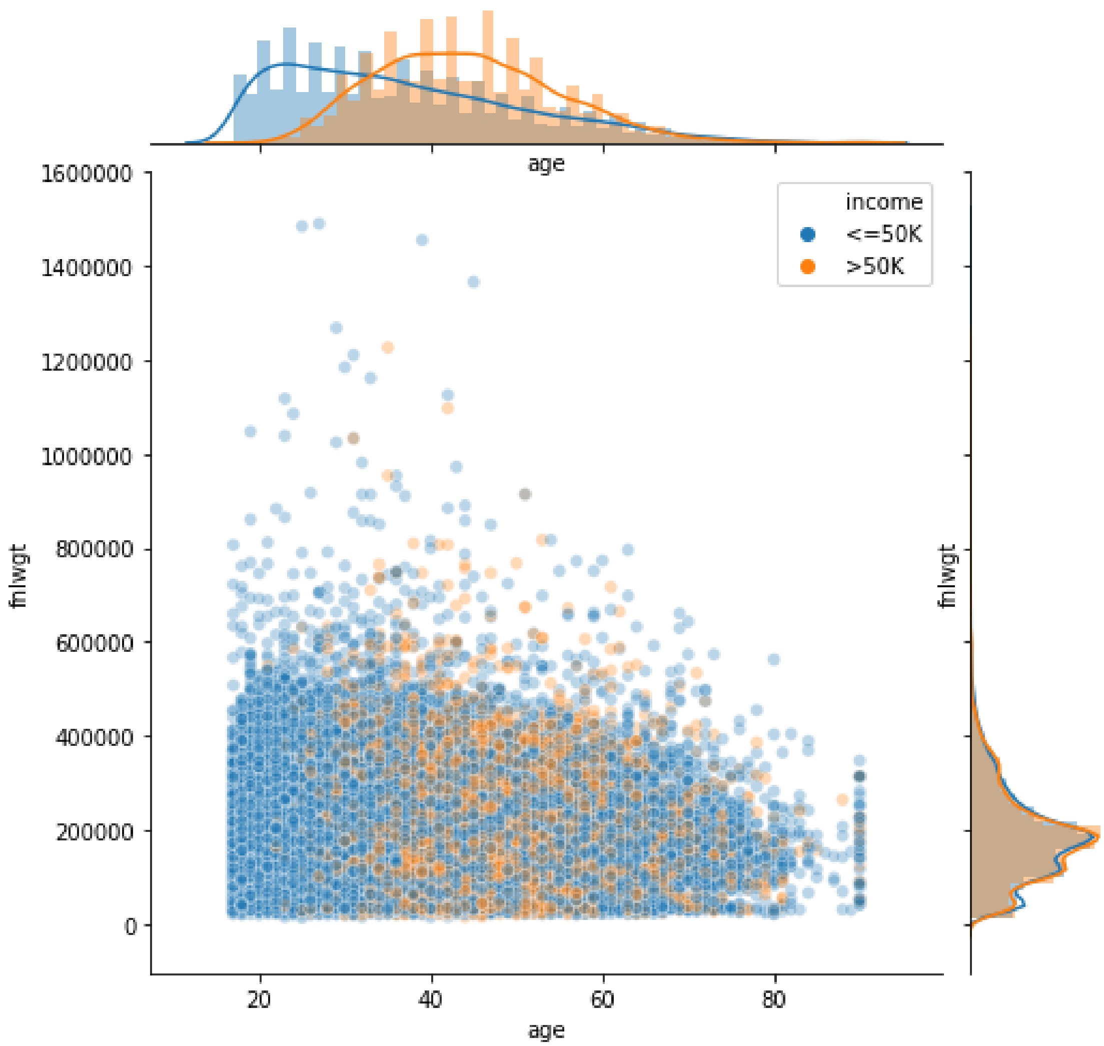
Joint plot for **capital-gain** & **hours-per-week**



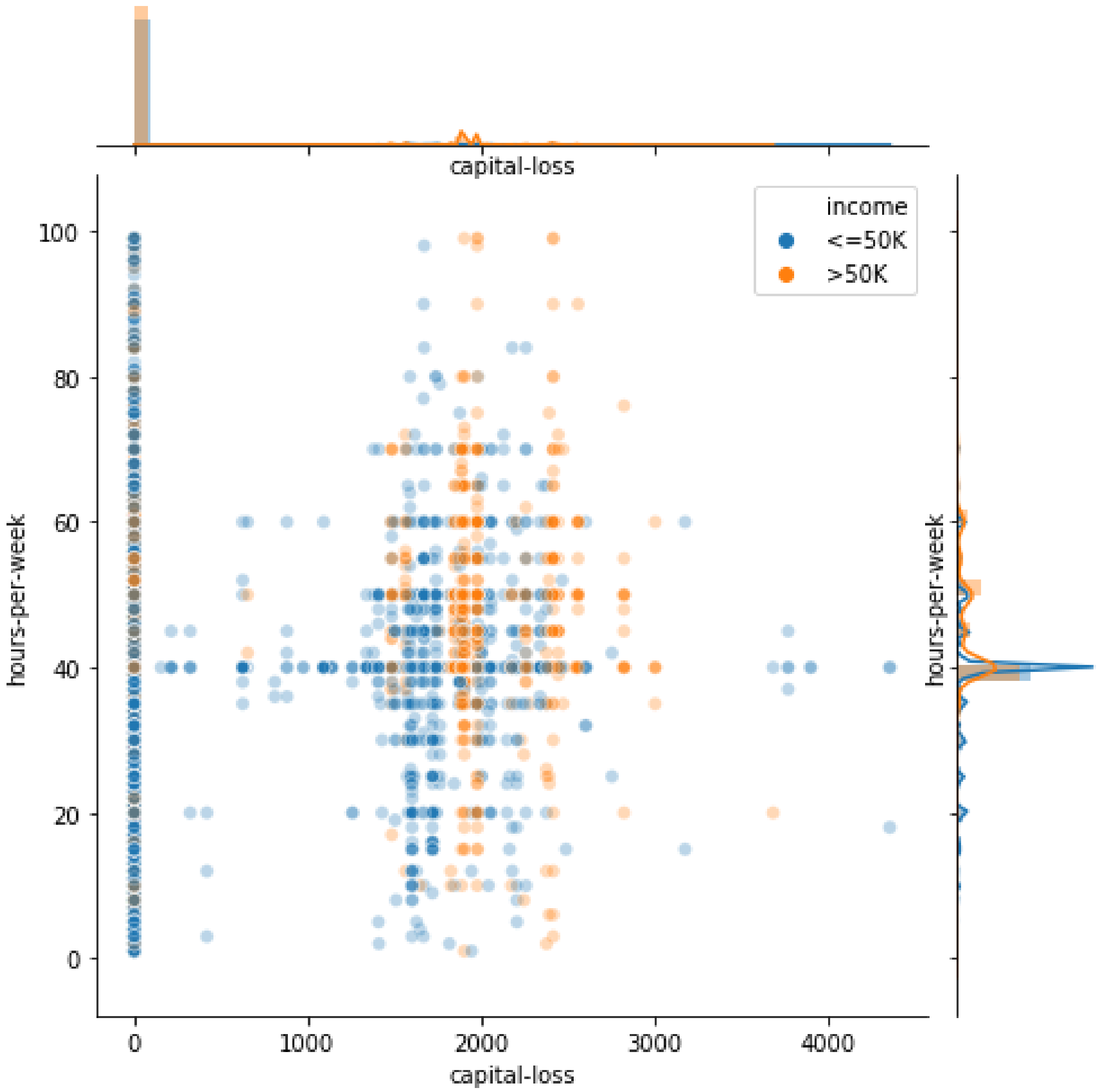
Joint plot for **capital-gain** & **fnlwgt**



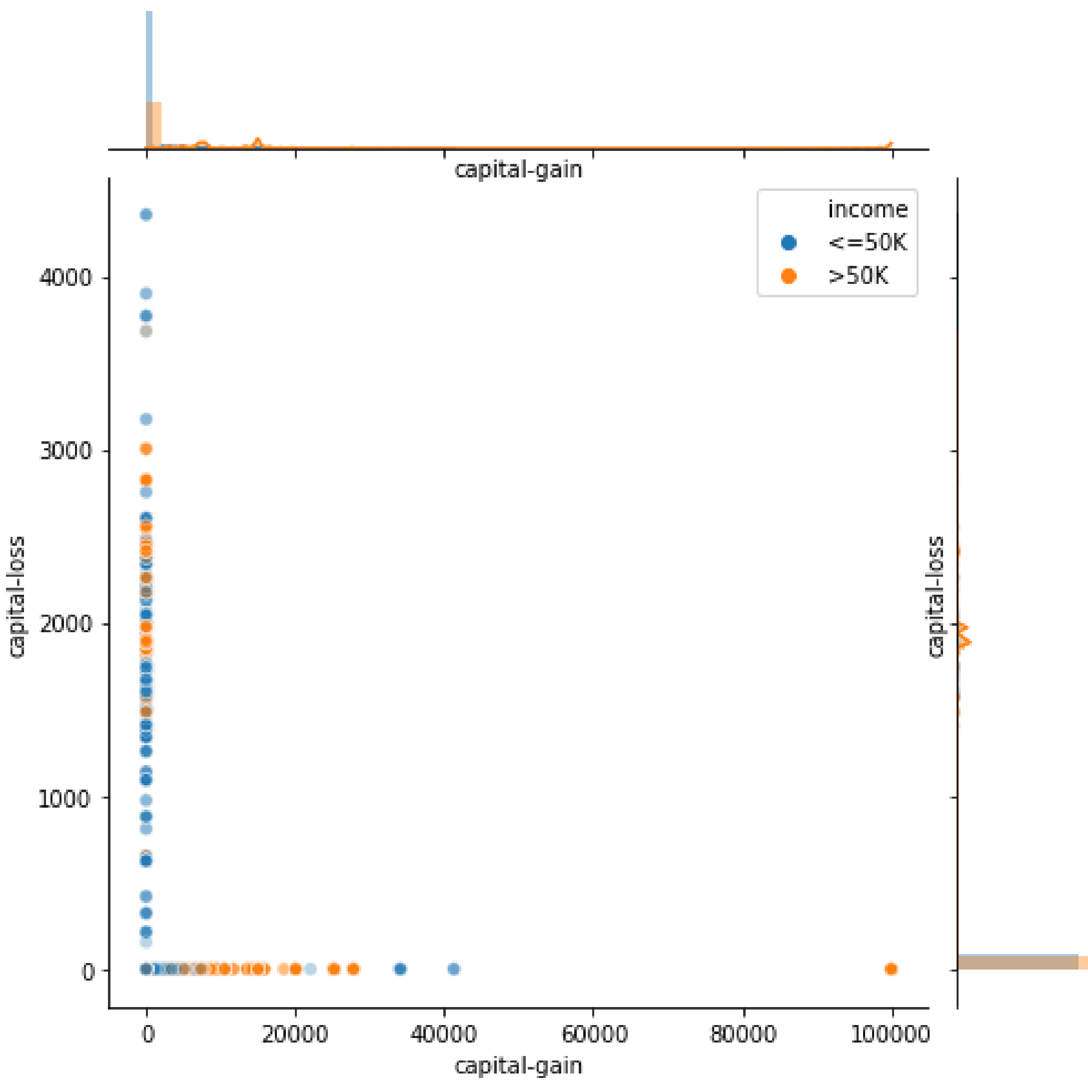
Joint plot for **age** & **fnlwgt**



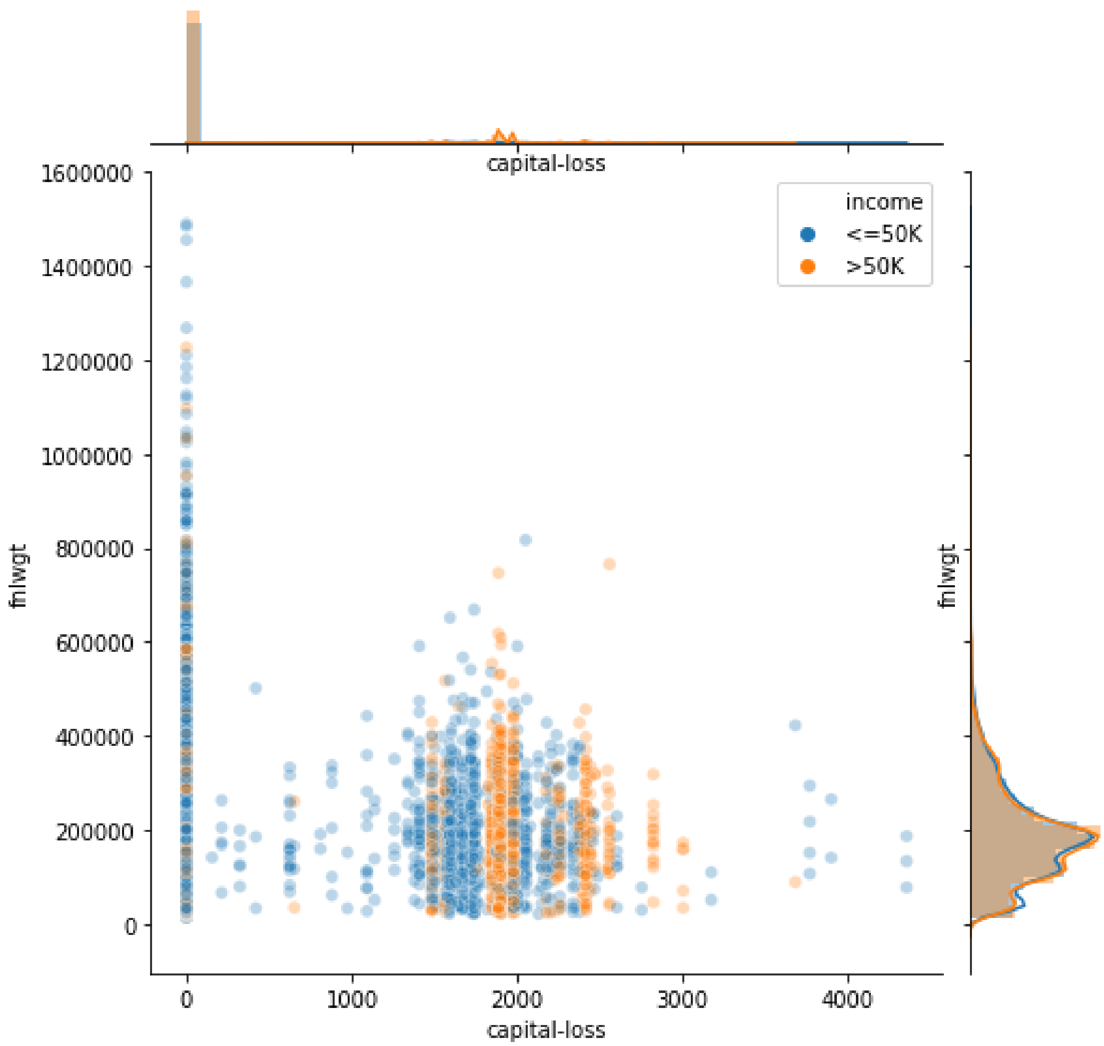
Joint plot for **capital-loss** & **hours-per-week**



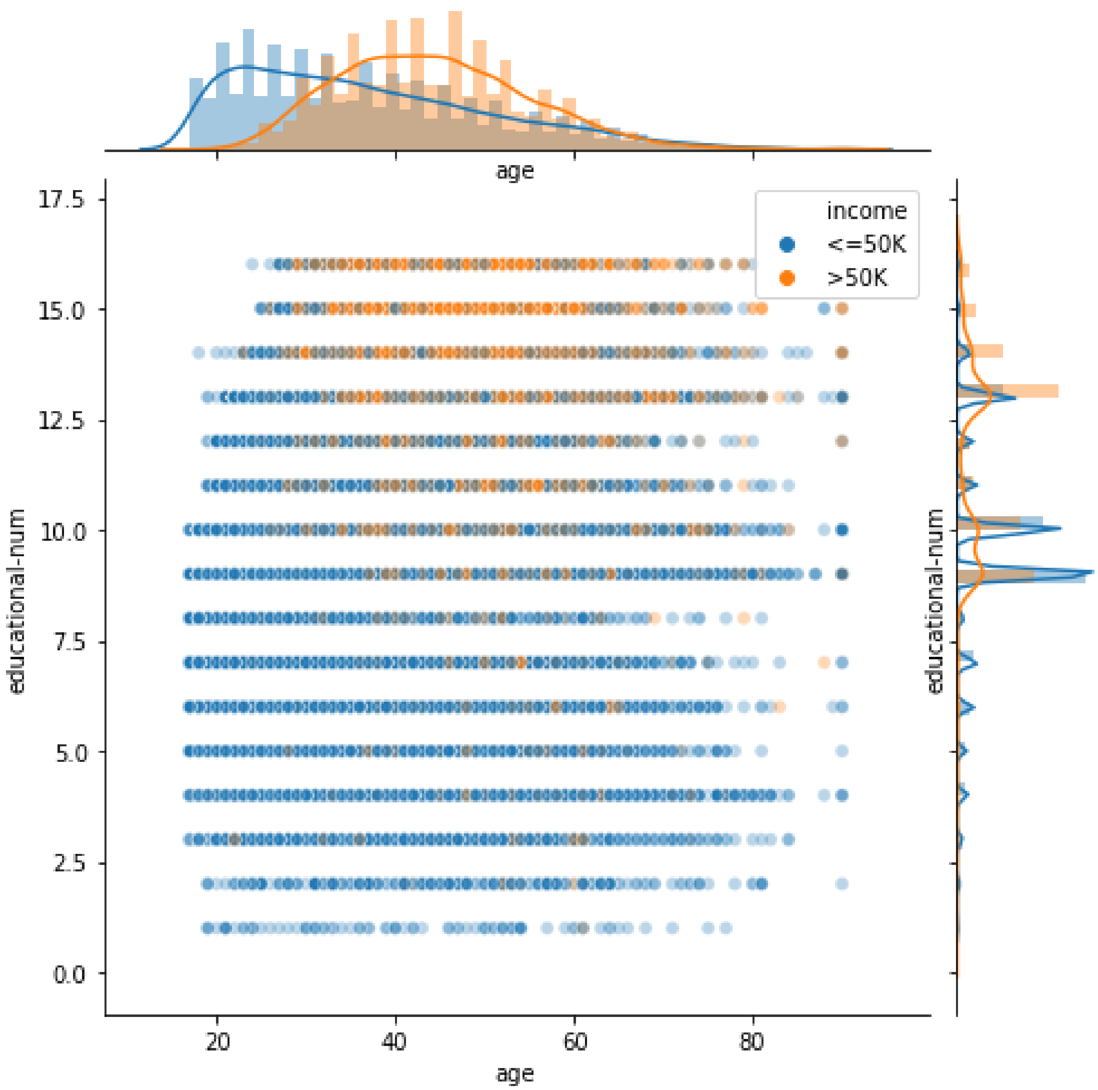
Joint plot for **capital-gain** & **capital-loss**



Joint plot for **capital-loss** & **fnlwgt**



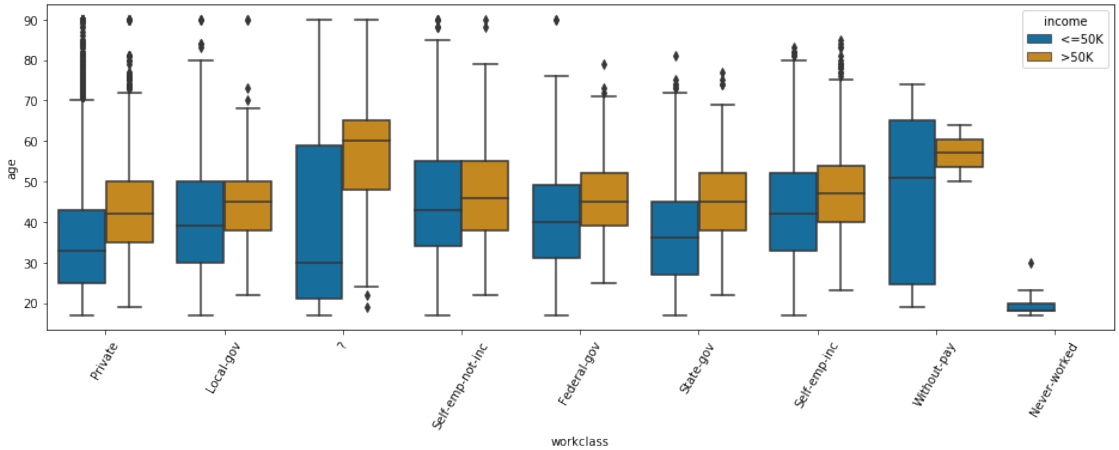
Joint plot for **age** & **educational-num**



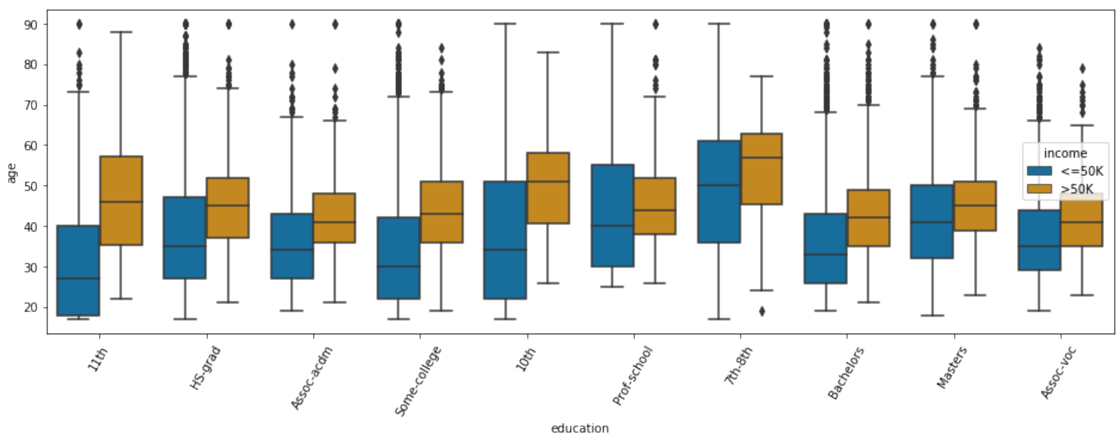
1.7 Analyse : numericals & categoricals variables relations

[9]:
```
explore.show_df_num_cat_relations(df=dataset, target=target)
```

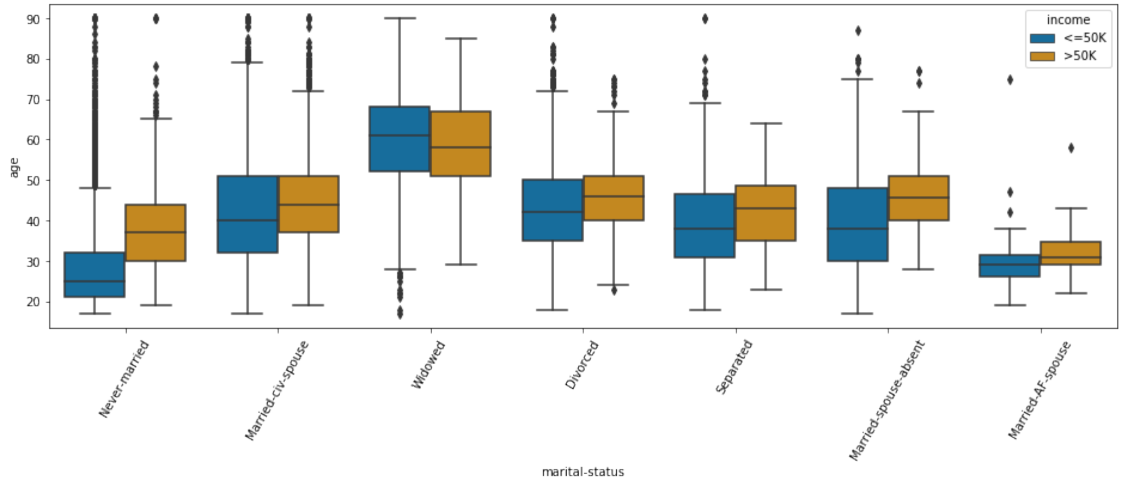
Box plot for **workclass** & **age**



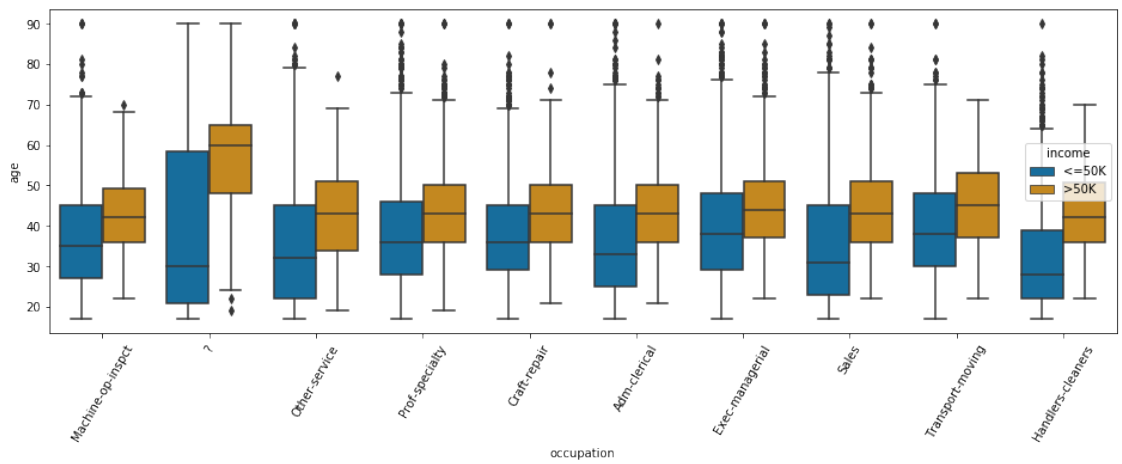
Box plot for **education** & **age**



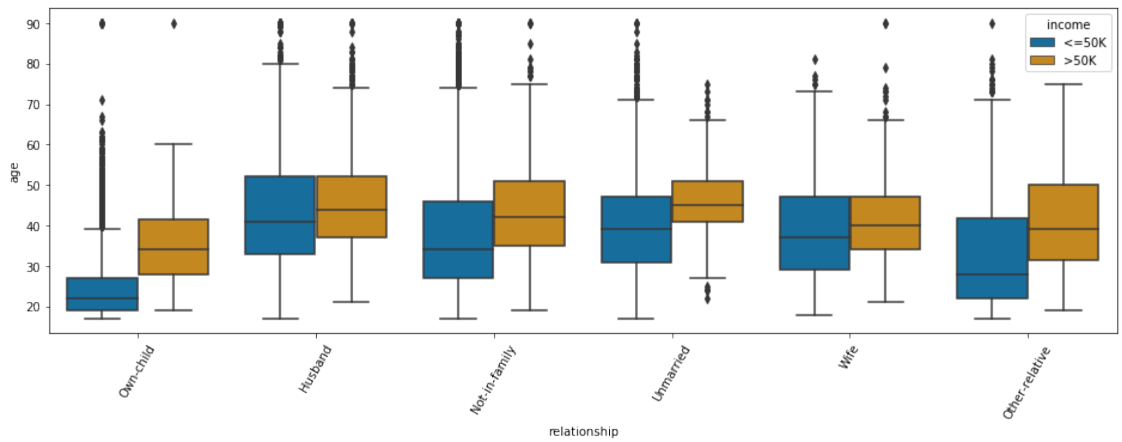
Box plot for **marital-status** & **age**



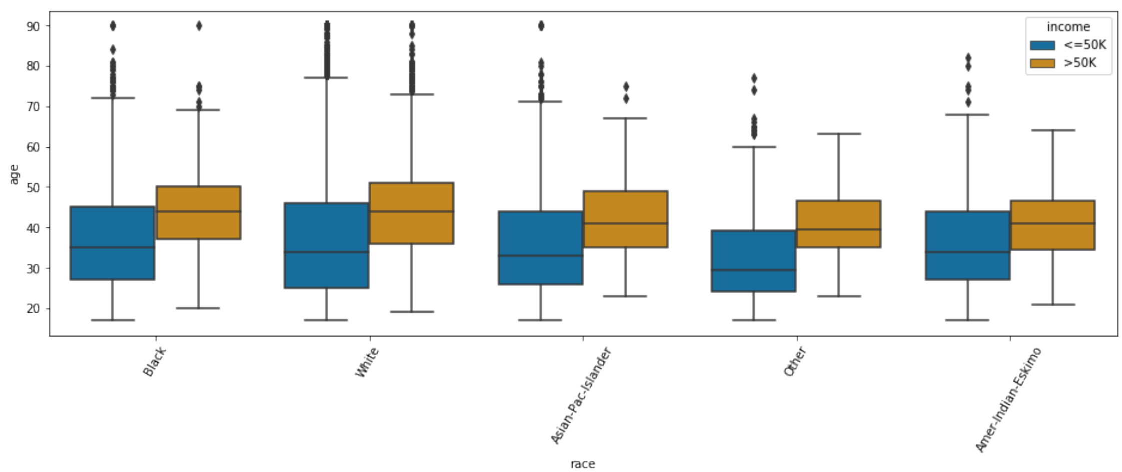
Box plot for **occupation** & **age**



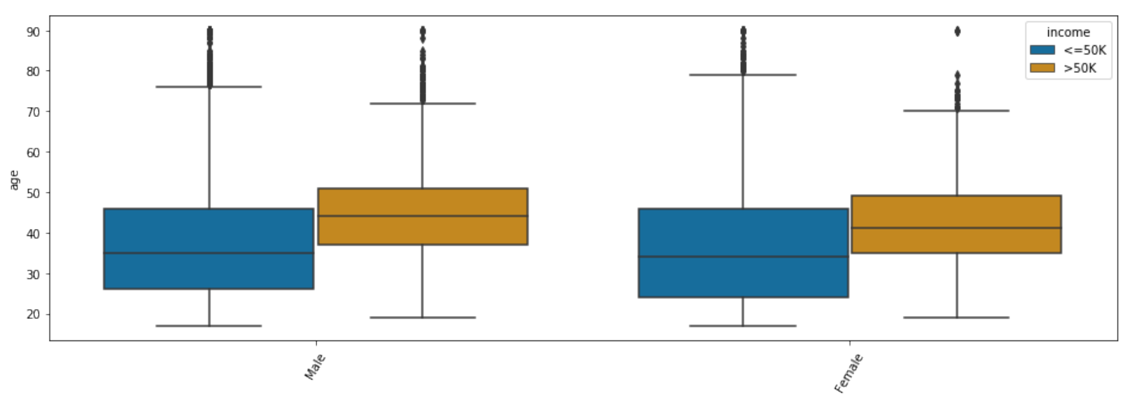
Box plot for **relationship** & **age**



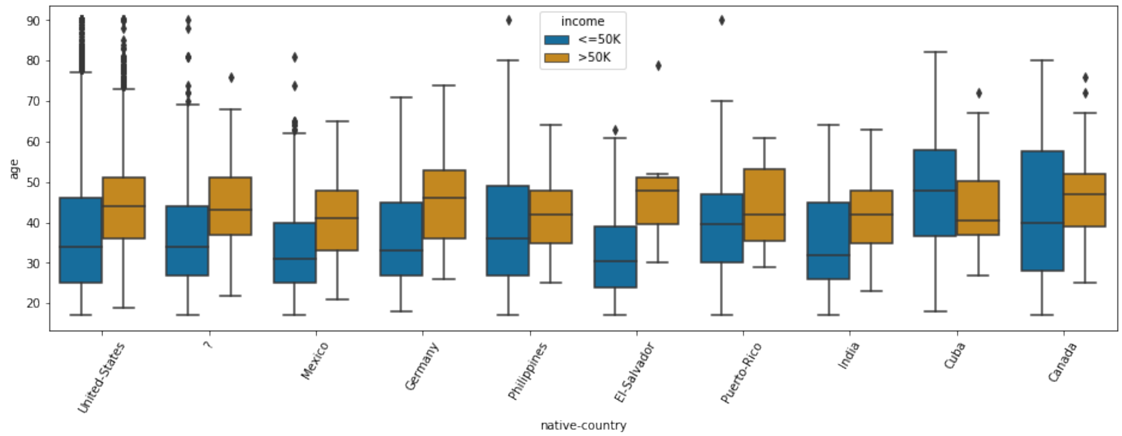
Box plot for **race** & **age**



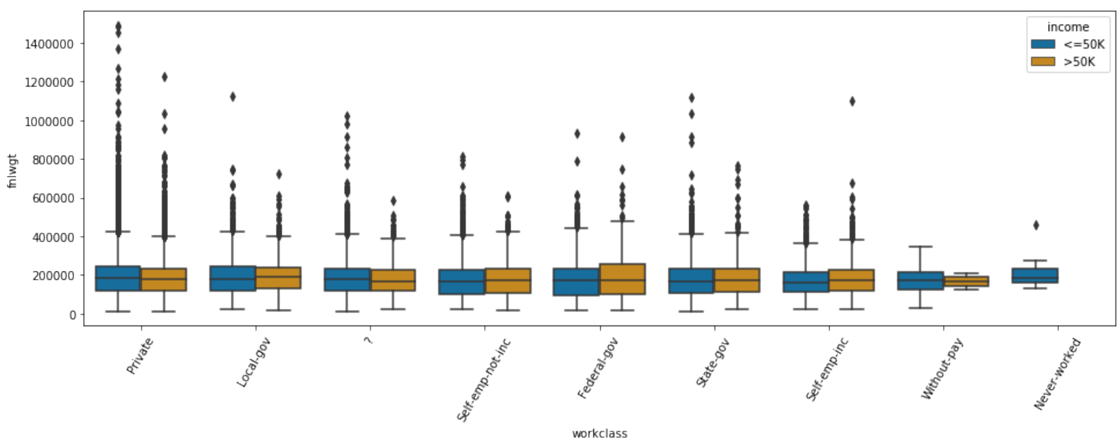
Box plot for **gender** & **age**



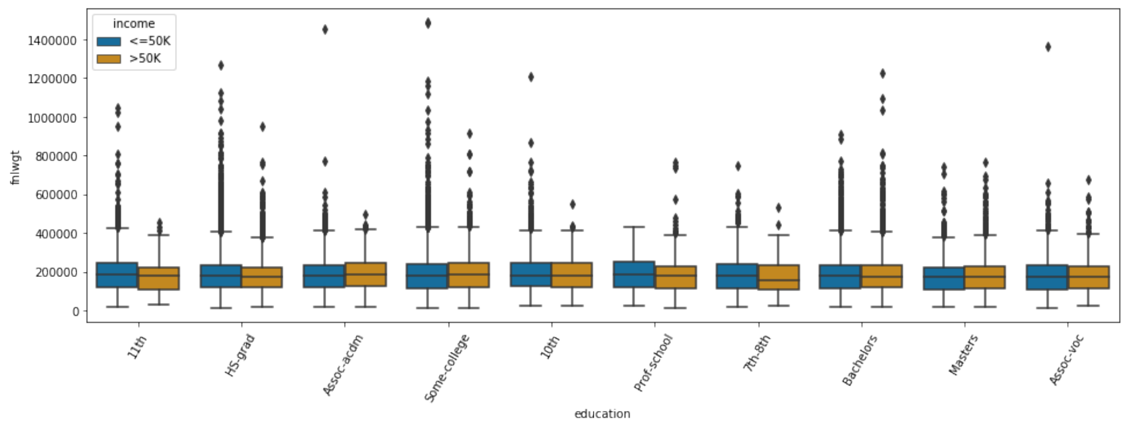
Box plot for **native-country** & **age**



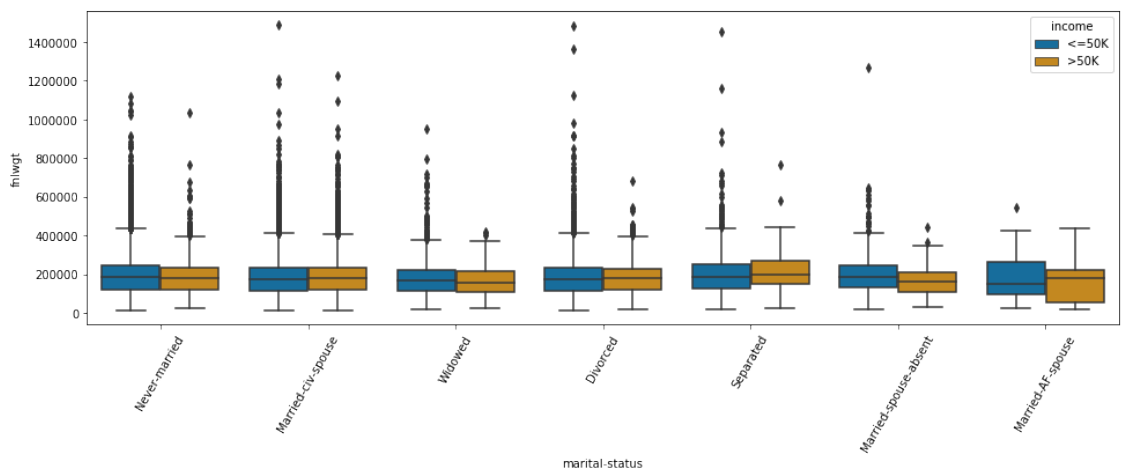
Box plot for **workclass** & **fnlwgt**



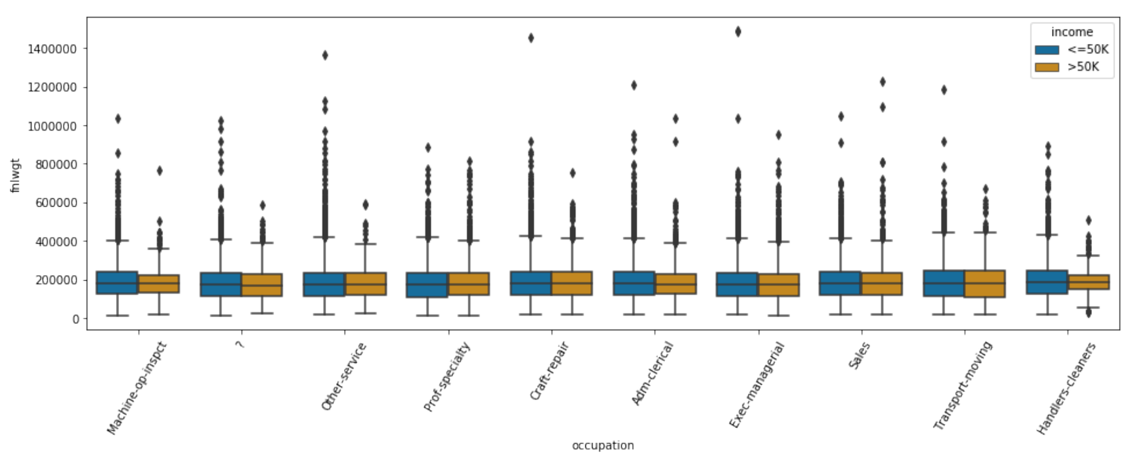
Box plot for **education** & **fnlwgt**



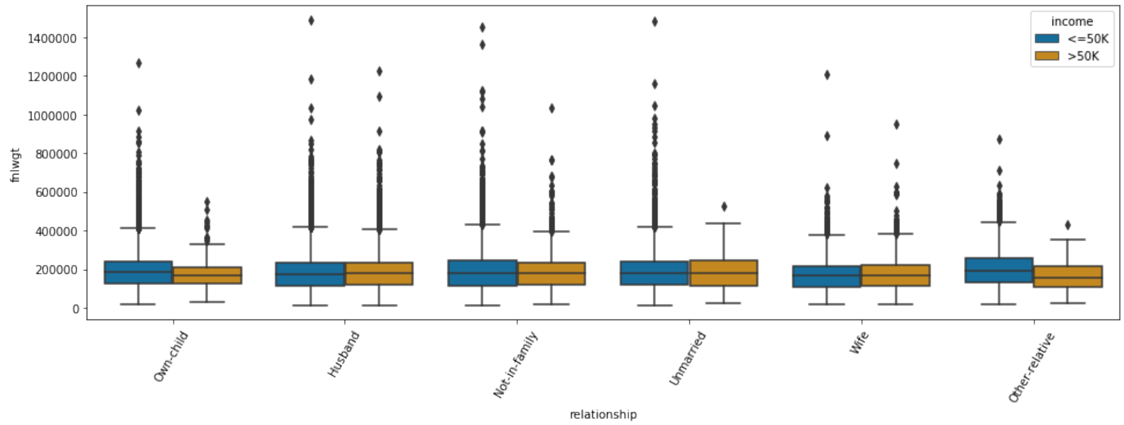
Box plot for **marital-status** & **fnlwgt**



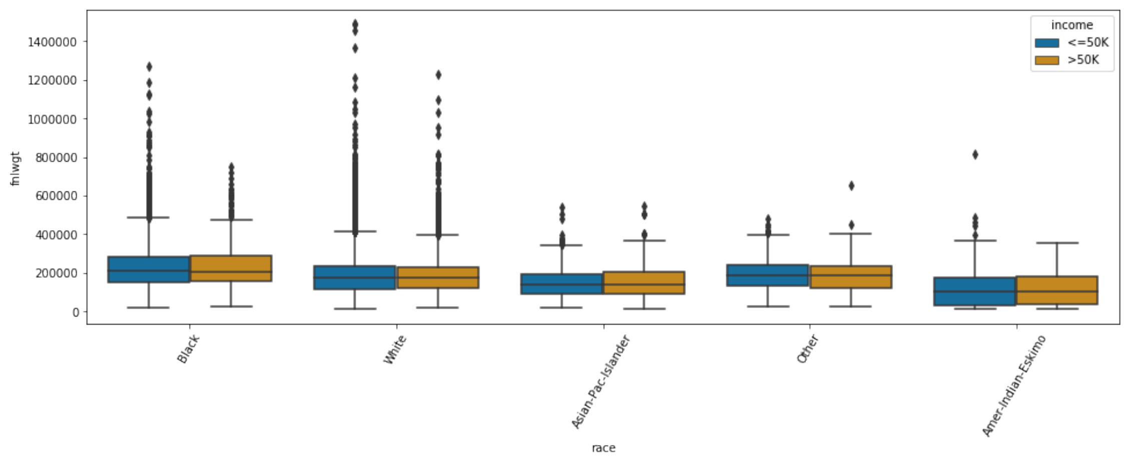
Box plot for **occupation** & **fnlwgt**



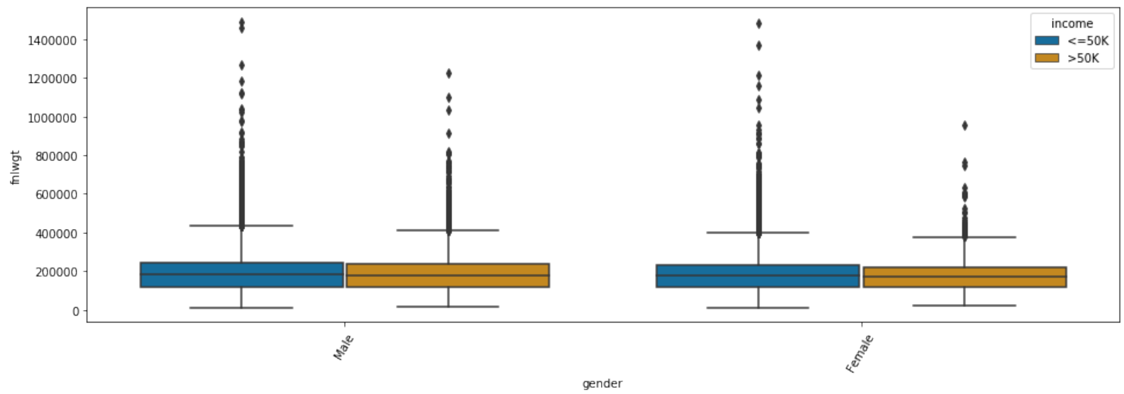
Box plot for **relationship** & **fnlwgt**



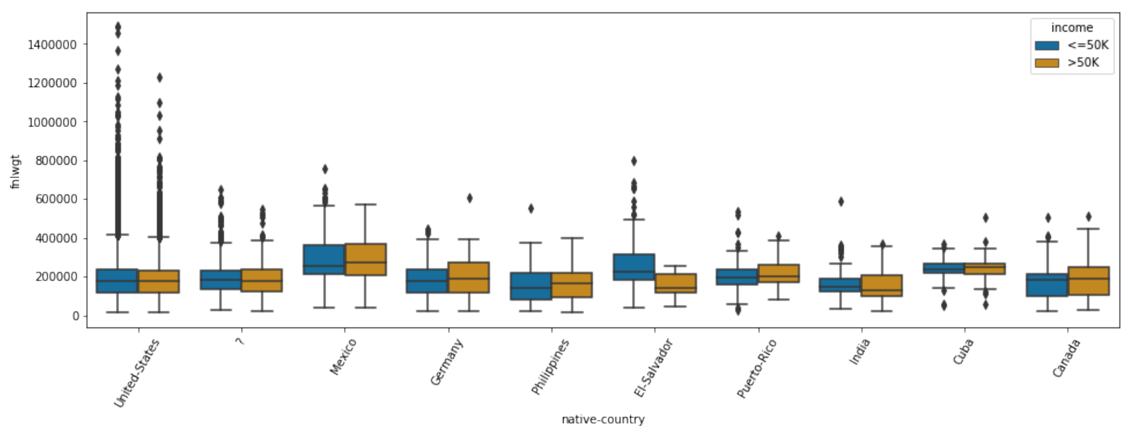
Box plot for **race** & **fnlwgt**



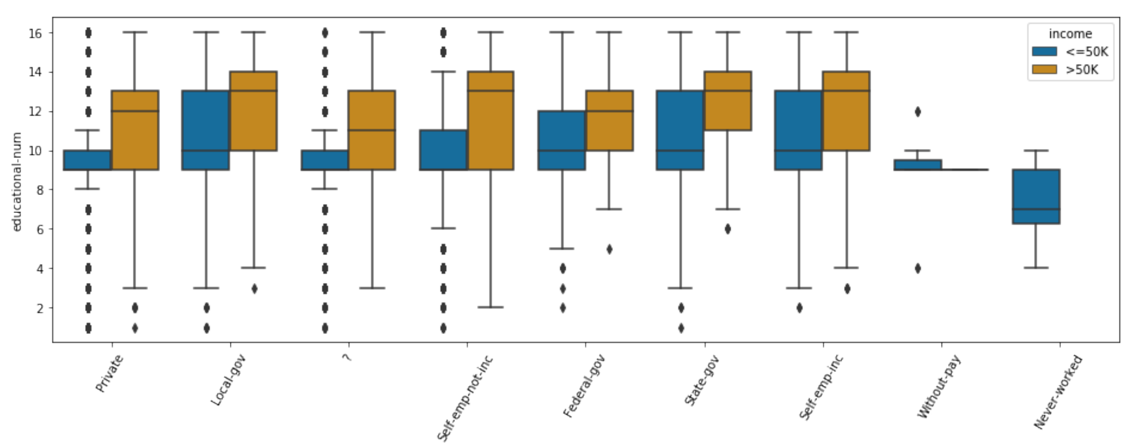
Box plot for **gender** & **fnlwgt**



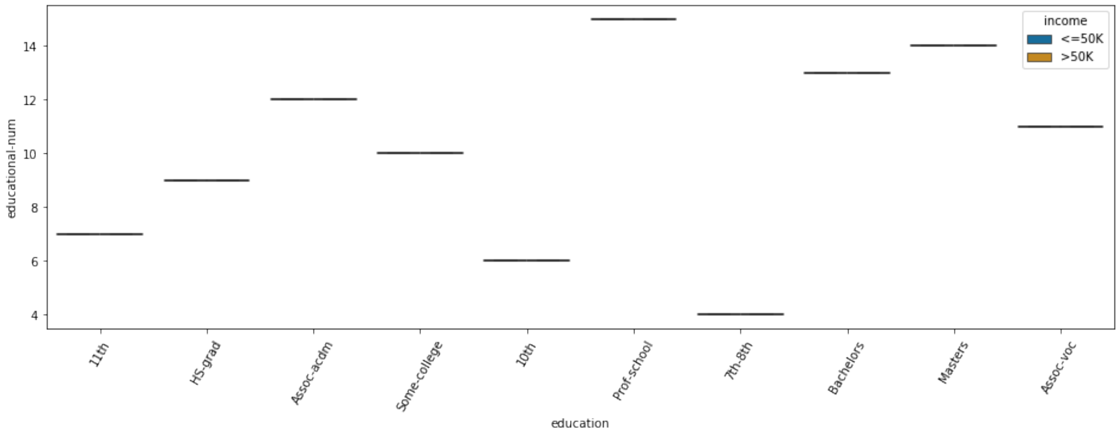
Box plot for **native-country** & **fnlwgt**



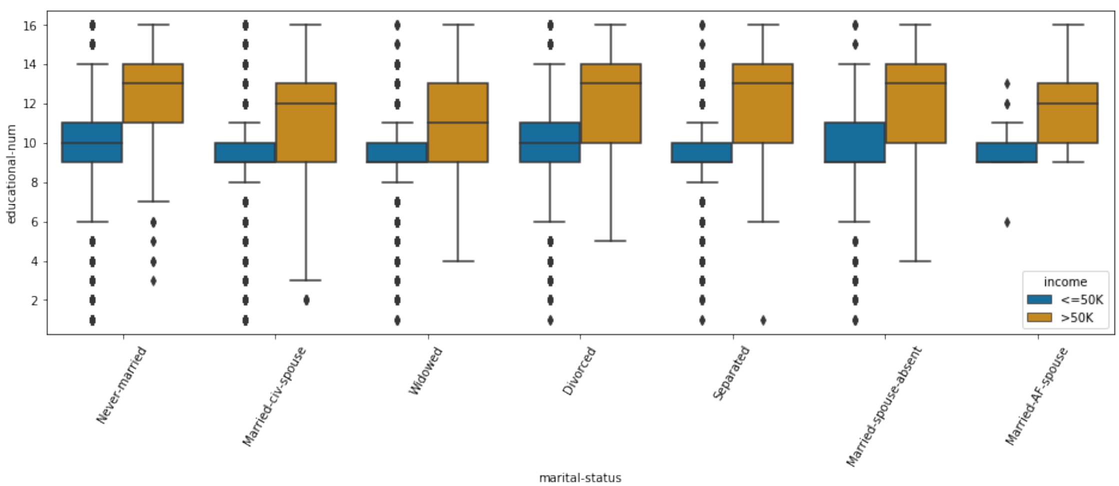
Box plot for **workclass** & **educational-num**



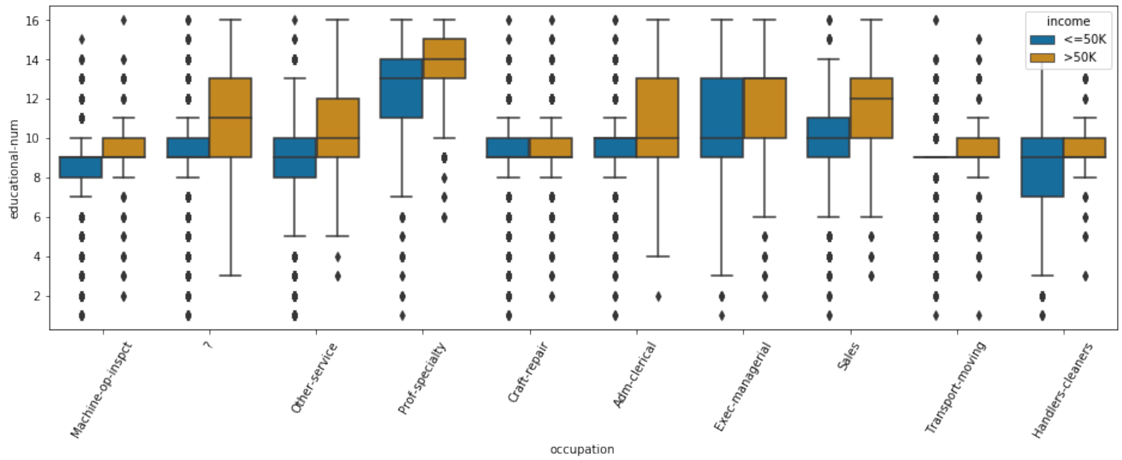
Box plot for **education** & **educational-num**



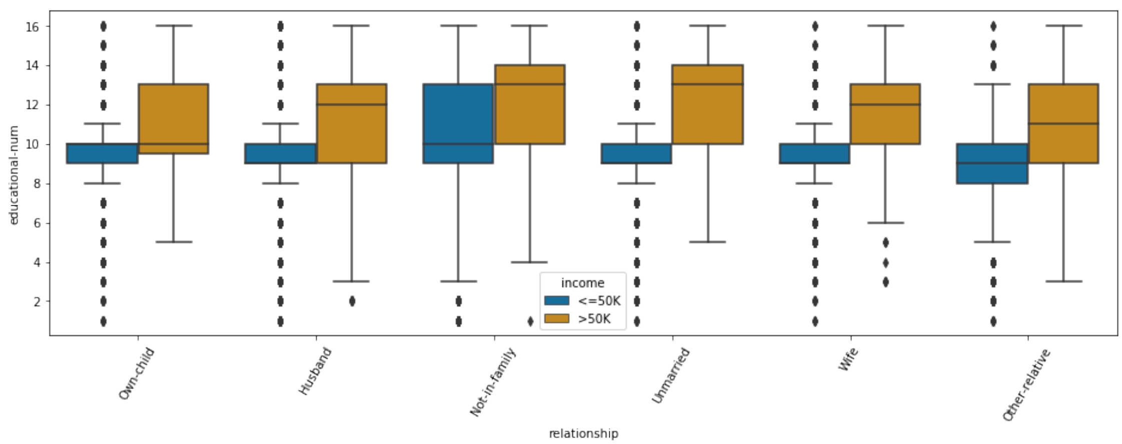
Box plot for **marital-status** & **educational-num**



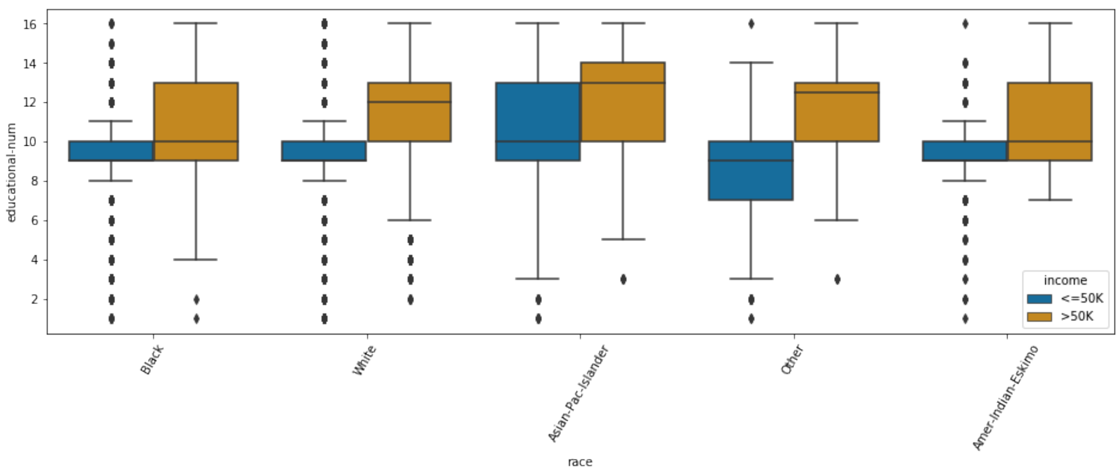
Box plot for **occupation** & **educational-num**



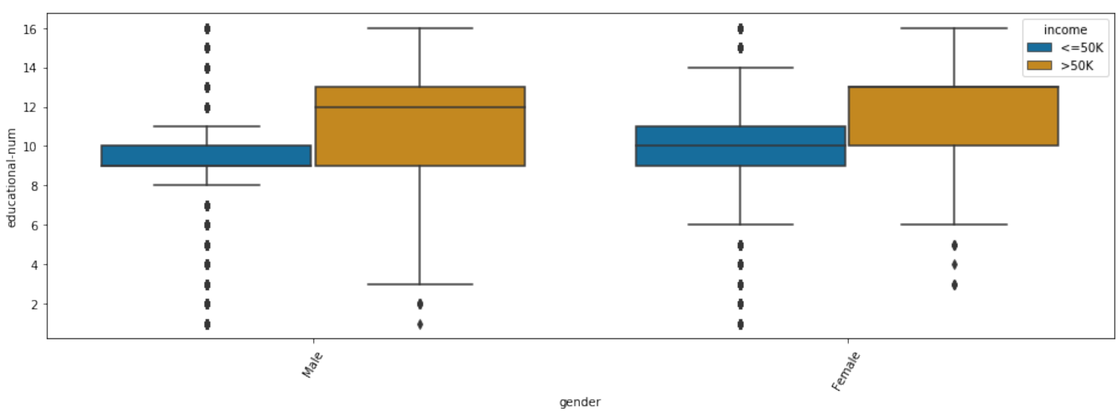
Box plot for **relationship** & **educational-num**



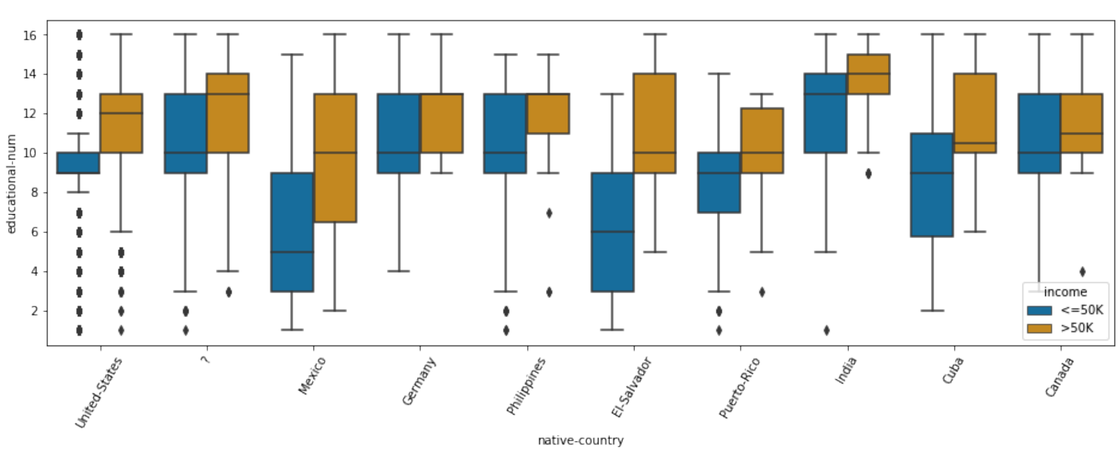
Box plot for **race** & **educational-num**



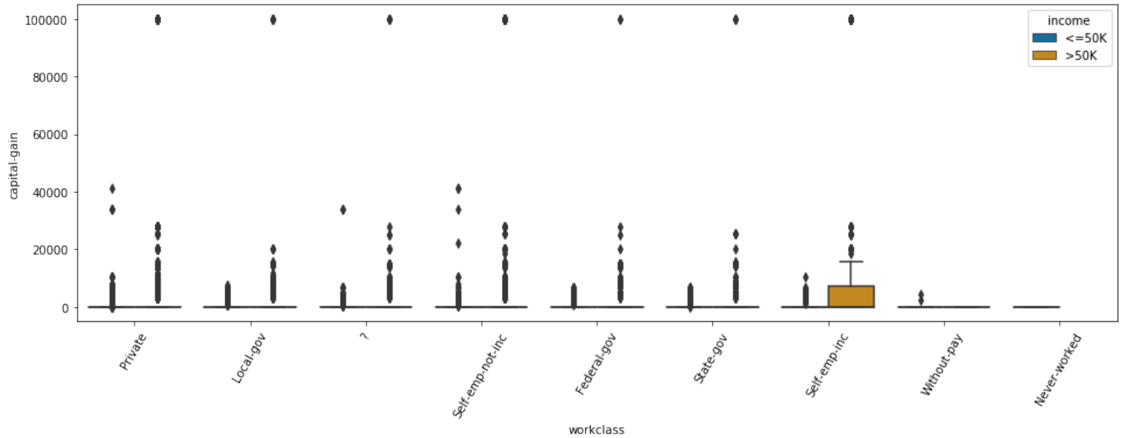
Box plot for **gender** & **educational-num**



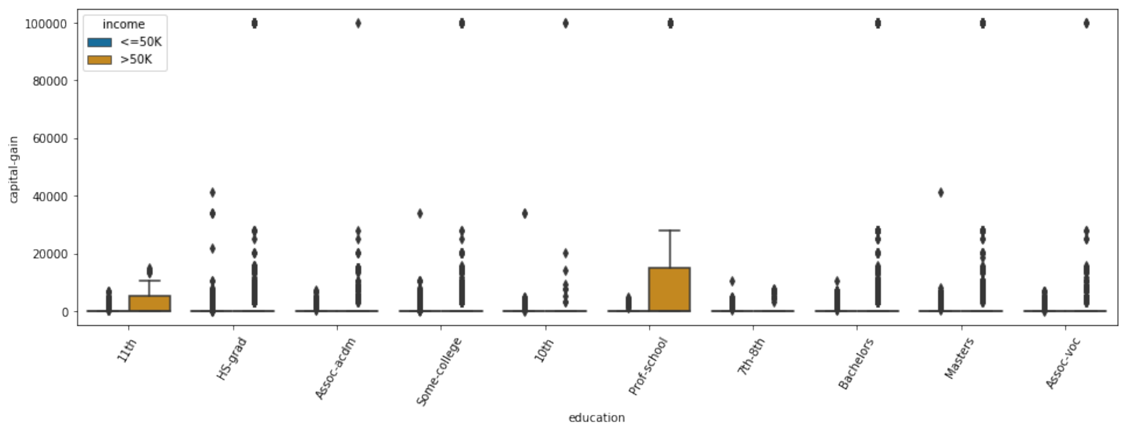
Box plot for **native-country** & **educational-num**



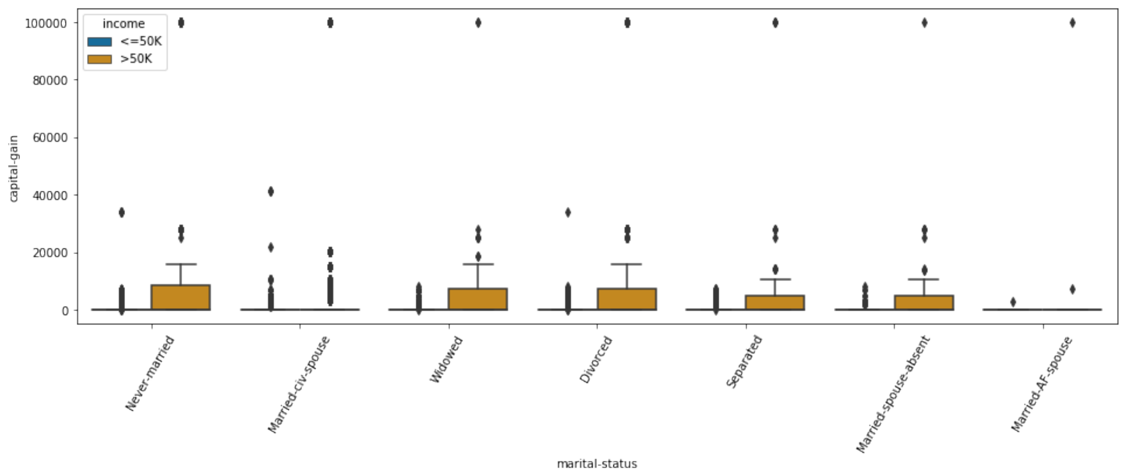
Box plot for **workclass** & **capital-gain**



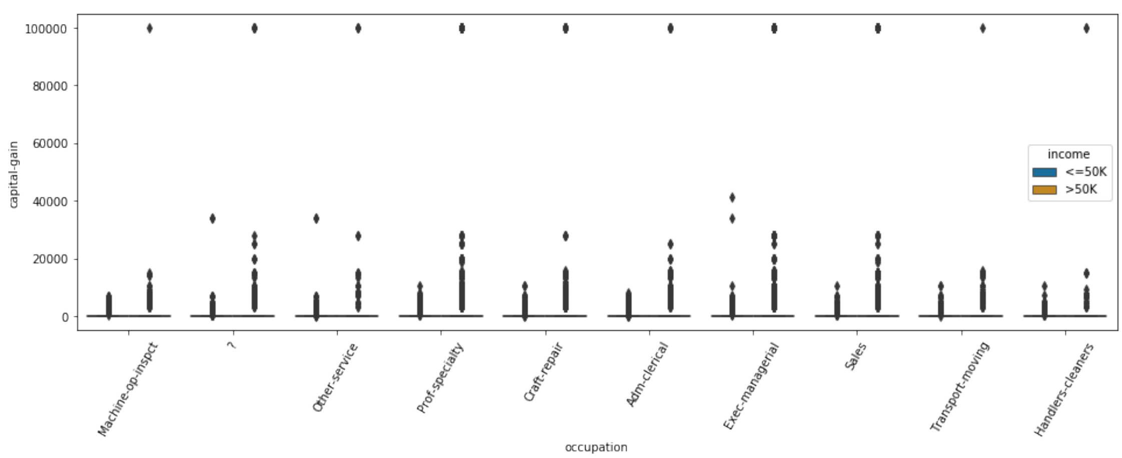
Box plot for **education** & **capital-gain**



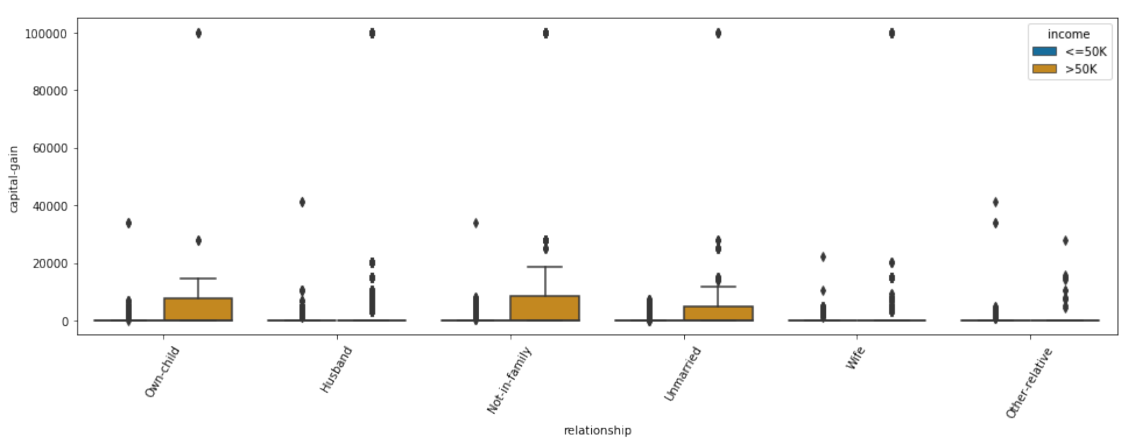
Box plot for **marital-status** & **capital-gain**



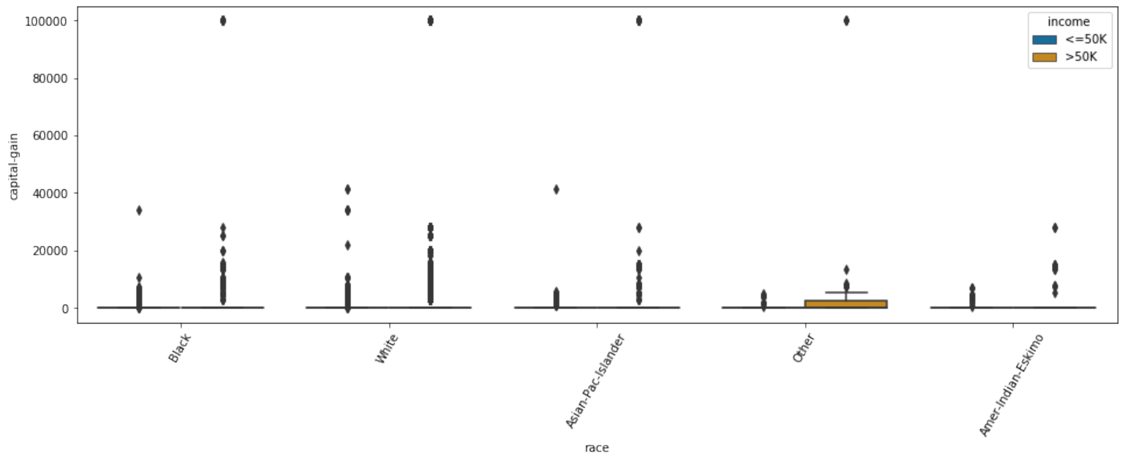
Box plot for **occupation** & **capital-gain**



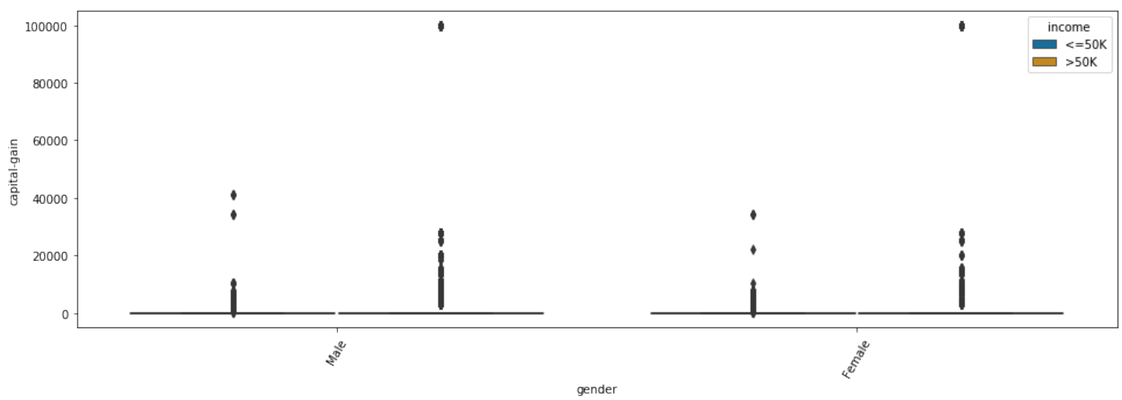
Box plot for **relationship** & **capital-gain**



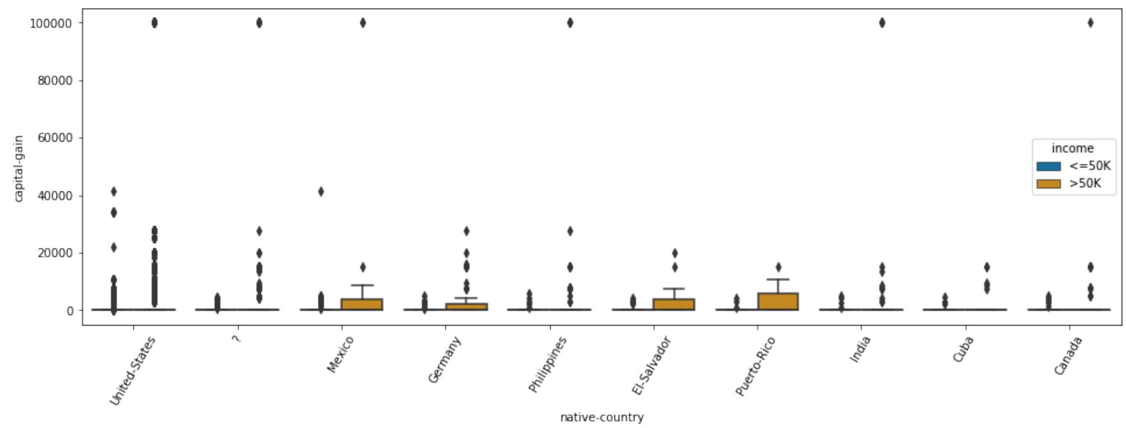
Box plot for **race** & **capital-gain**



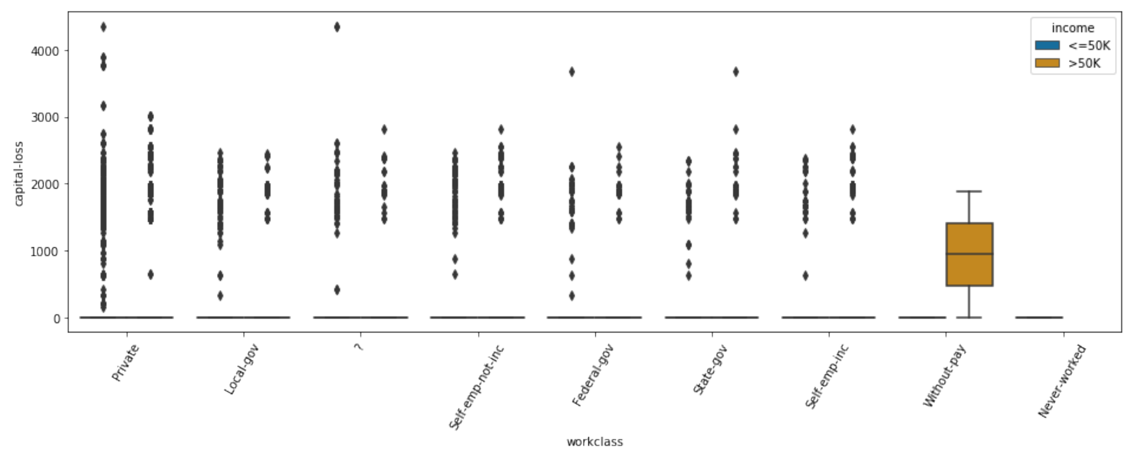
Box plot for **gender** & **capital-gain**



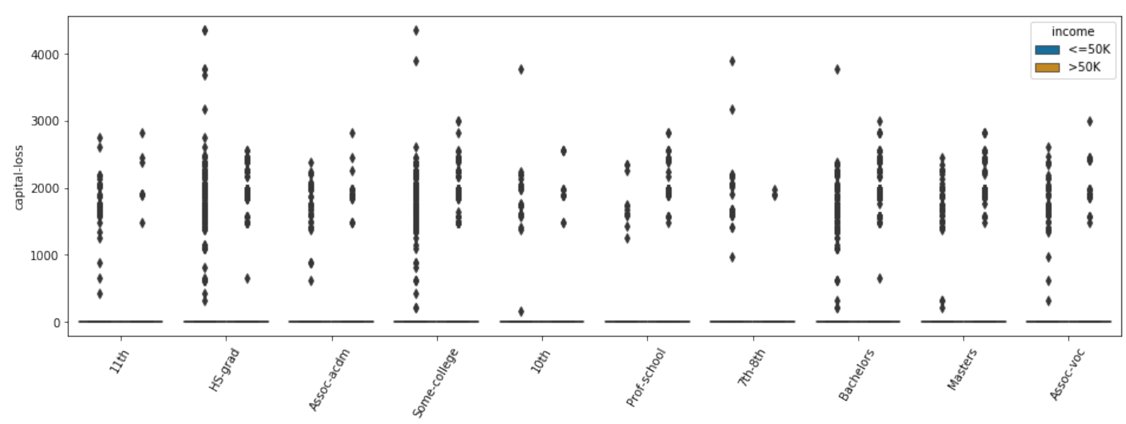
Box plot for **native-country** & **capital-gain**



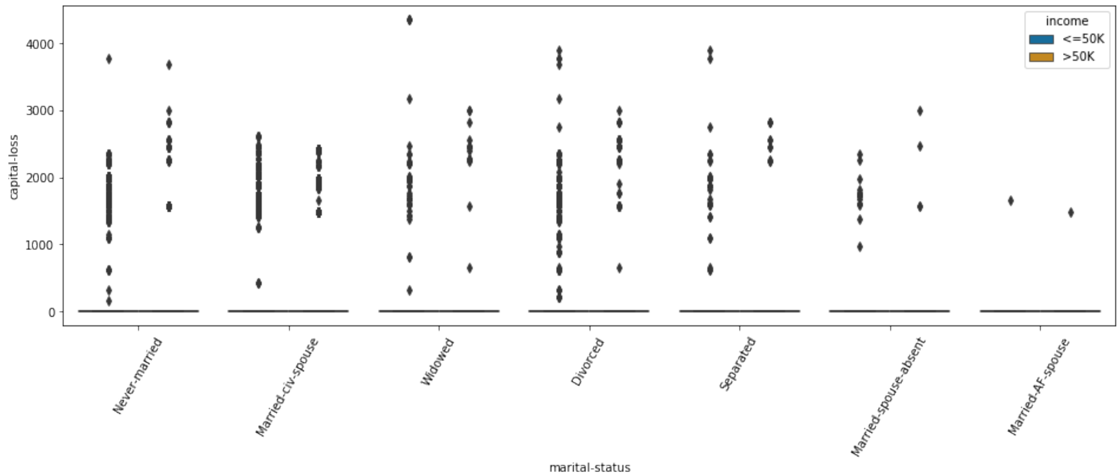
Box plot for **workclass** & **capital-loss**



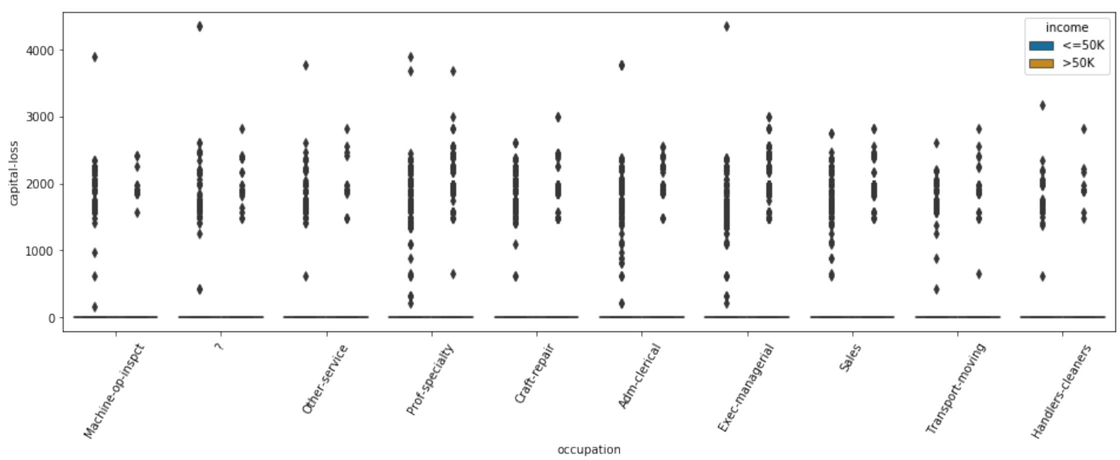
Box plot for **education** & **capital-loss**



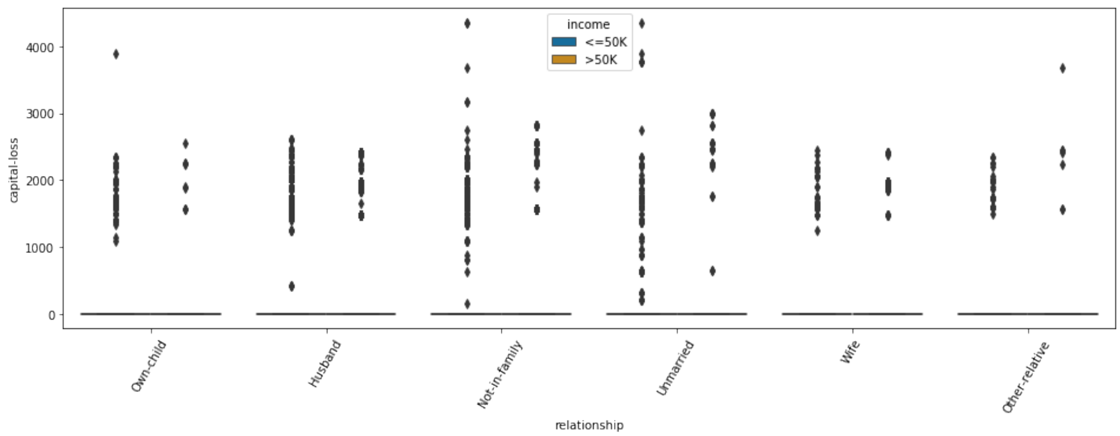
Box plot for **marital-status** & **capital-loss**



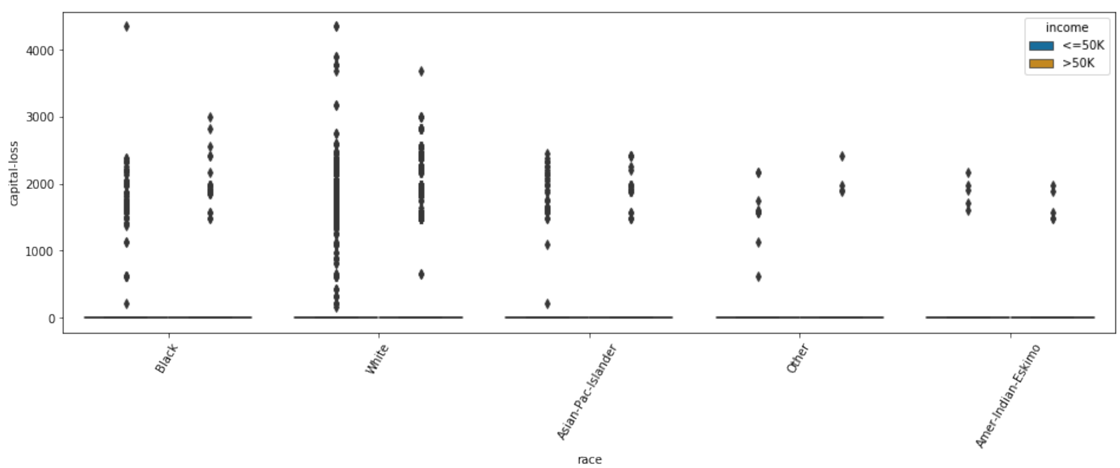
Box plot for **occupation** & **capital-loss**



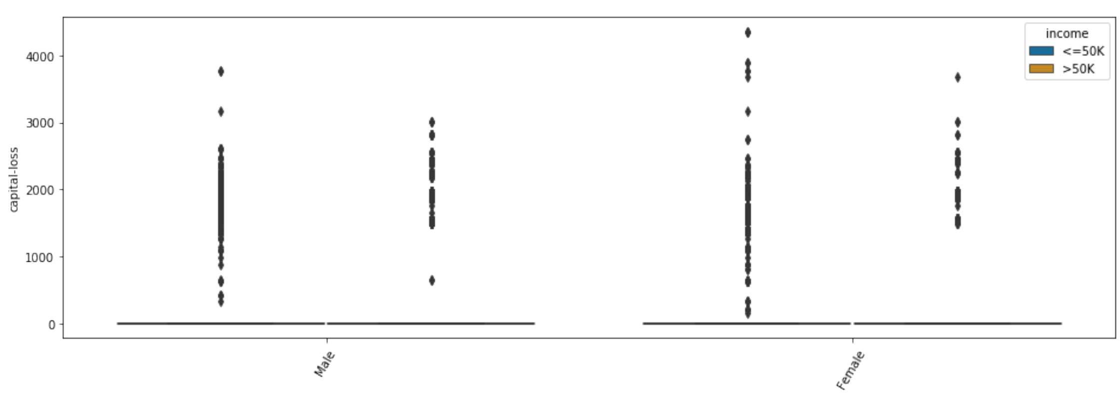
Box plot for **relationship** & **capital-loss**



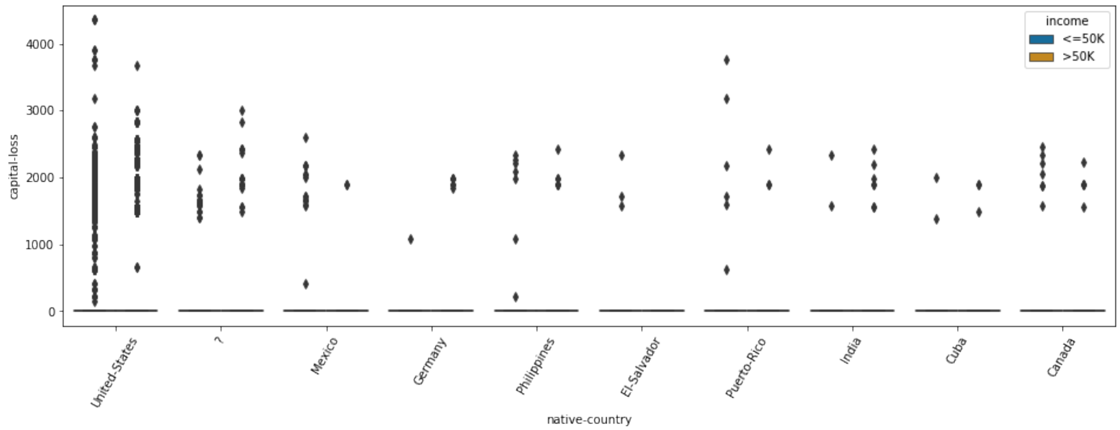
Box plot for **race** & **capital-loss**



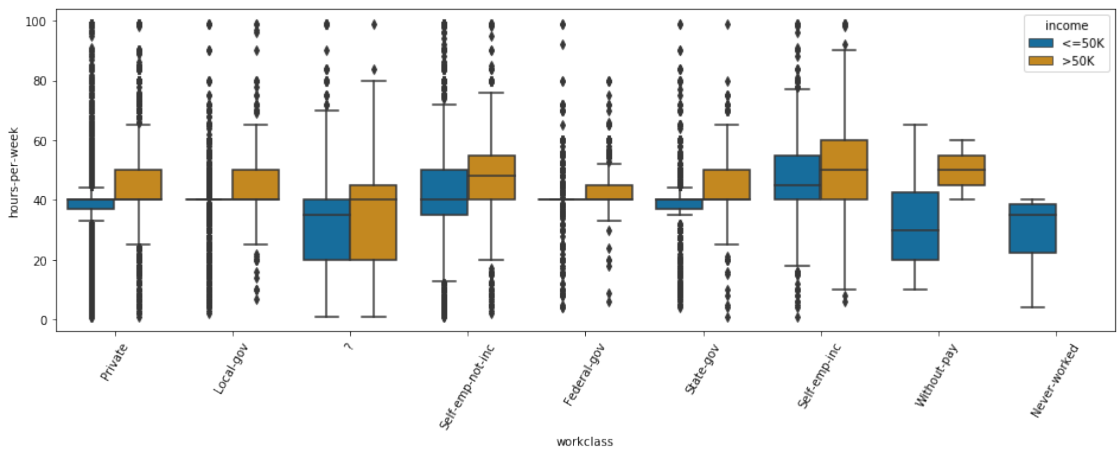
Box plot for **gender** & **capital-loss**



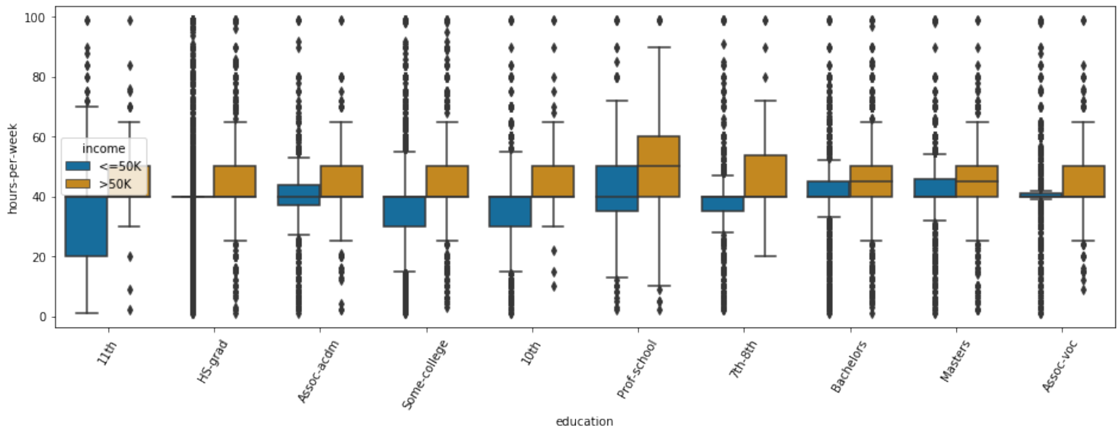
Box plot for **native-country** & **capital-loss**



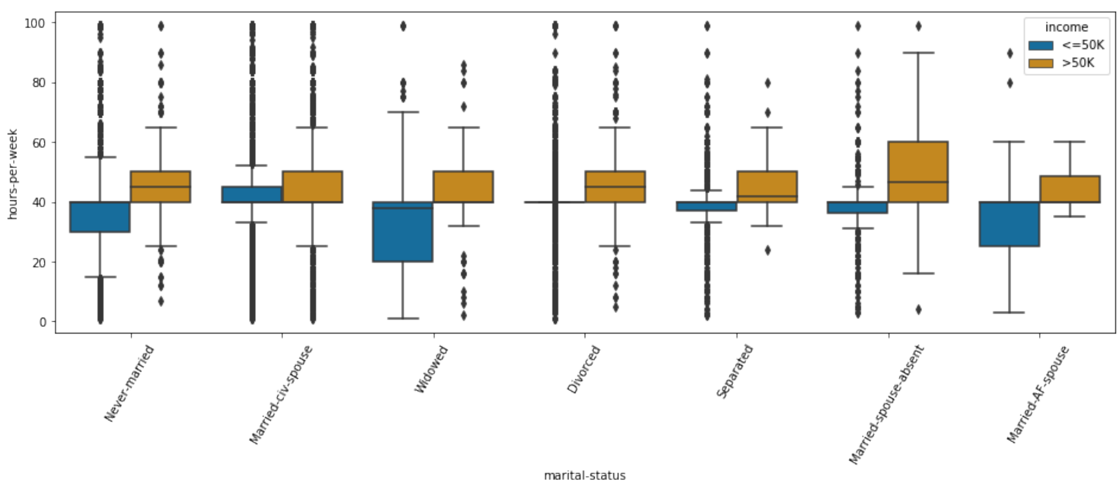
Box plot for **workclass** & **hours-per-week**



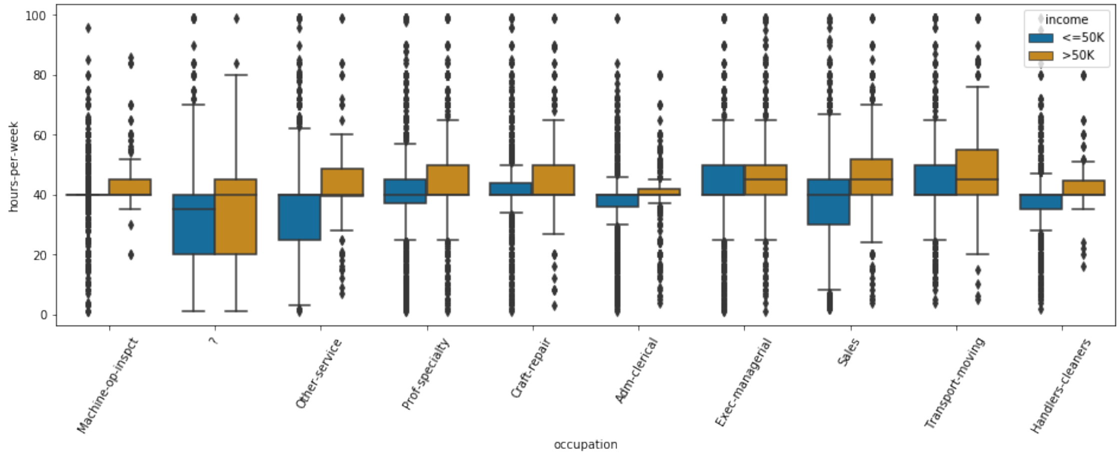
Box plot for **education** & **hours-per-week**



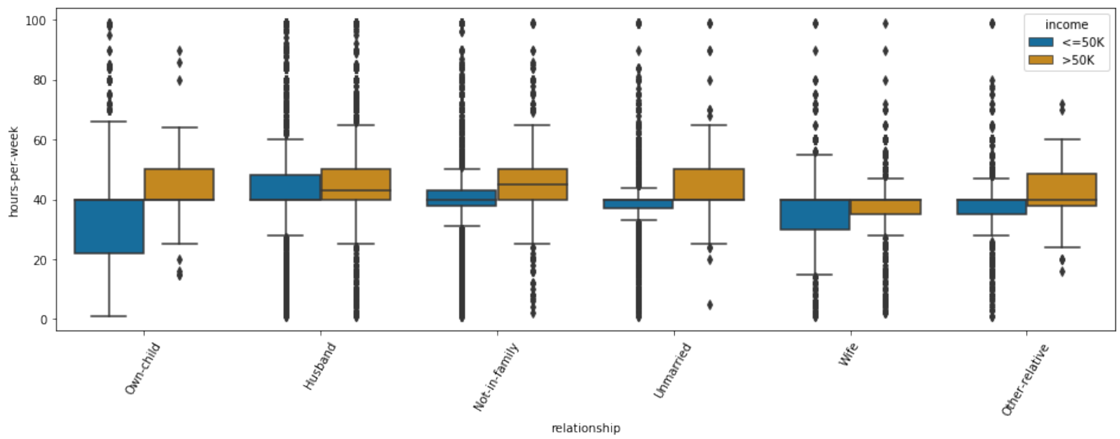
Box plot for **marital-status** & **hours-per-week**



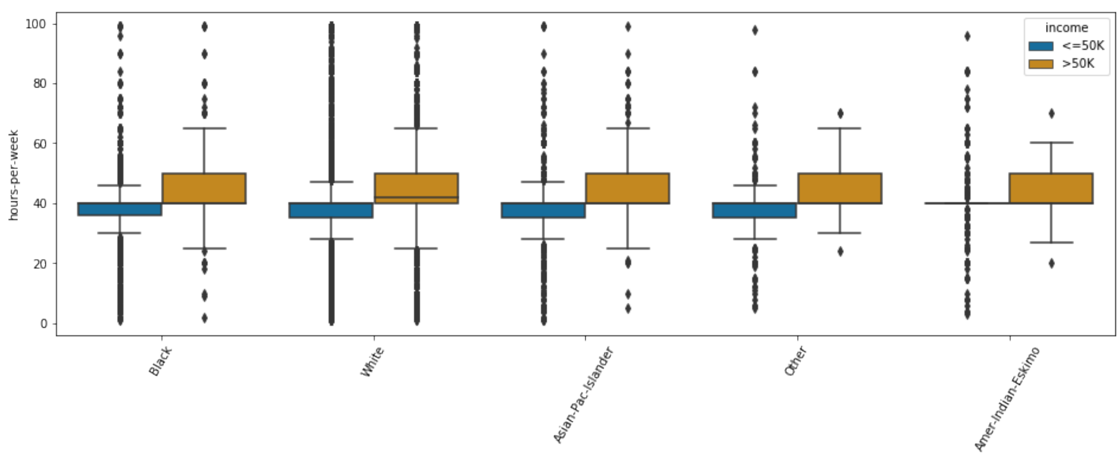
Box plot for **occupation** & **hours-per-week**



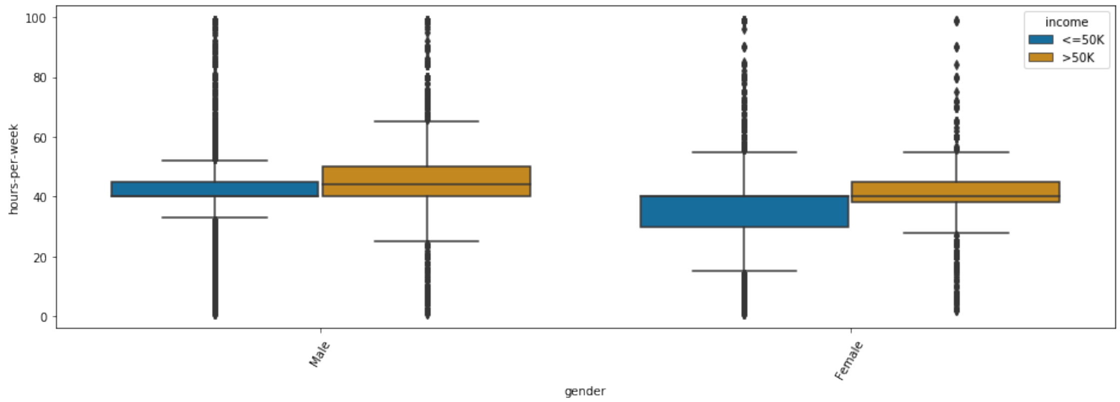
Box plot for **relationship** & **hours-per-week**



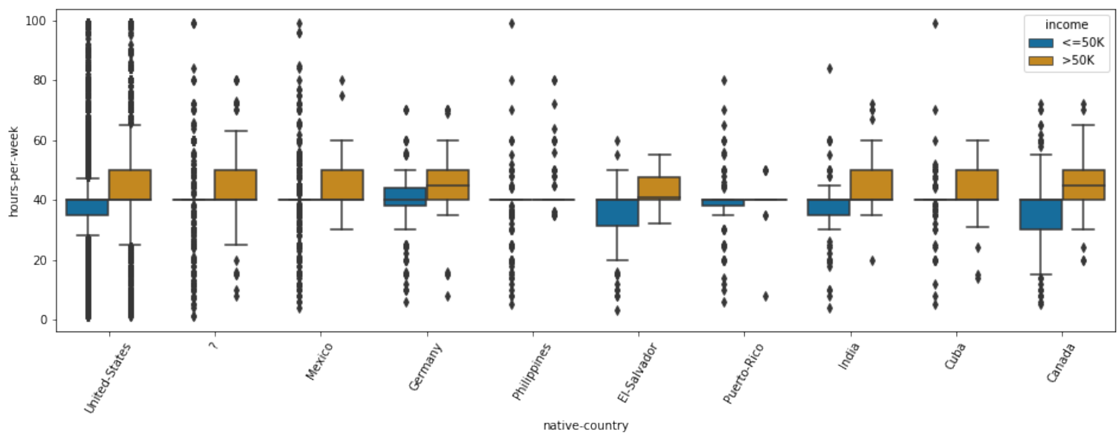
Box plot for **race** & **hours-per-week**



Box plot for **gender** & **hours-per-week**



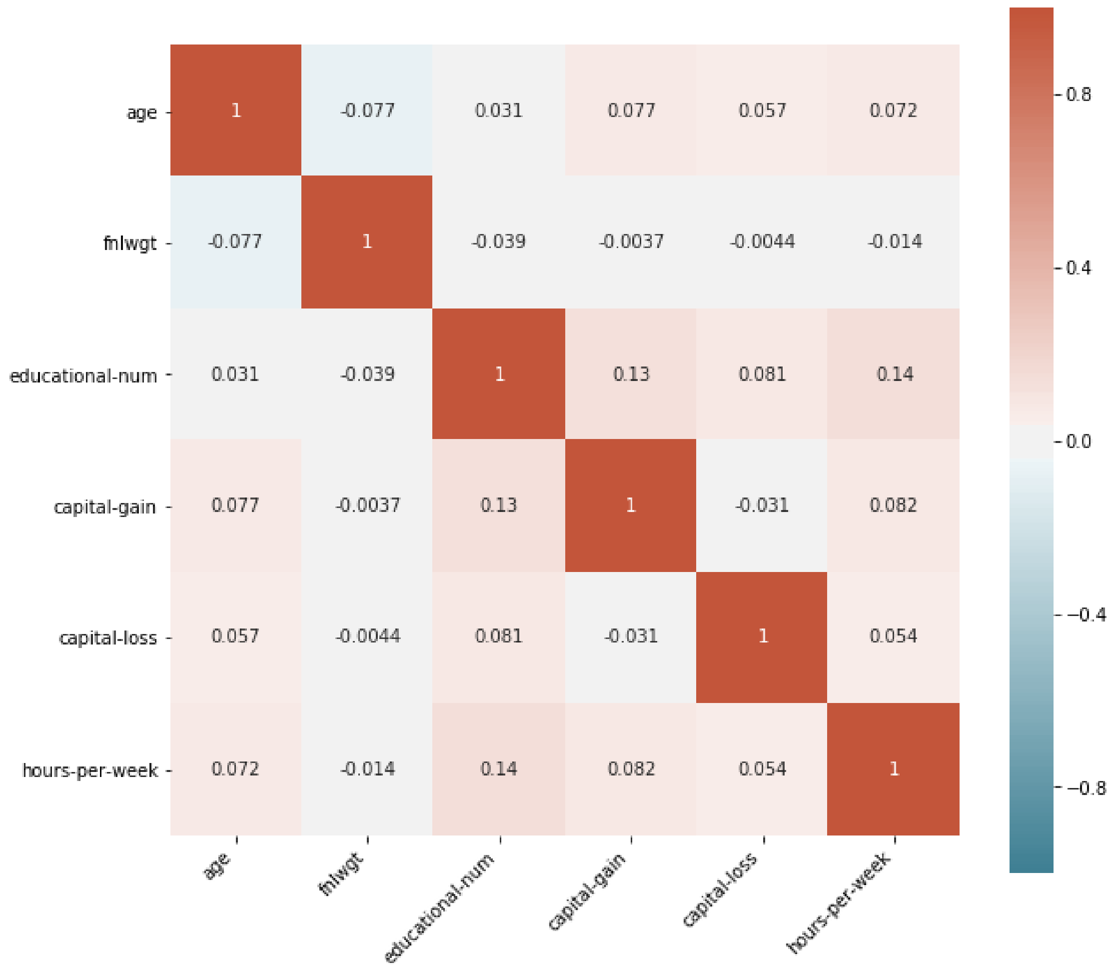
Box plot for **native-country** & **hours-per-week**



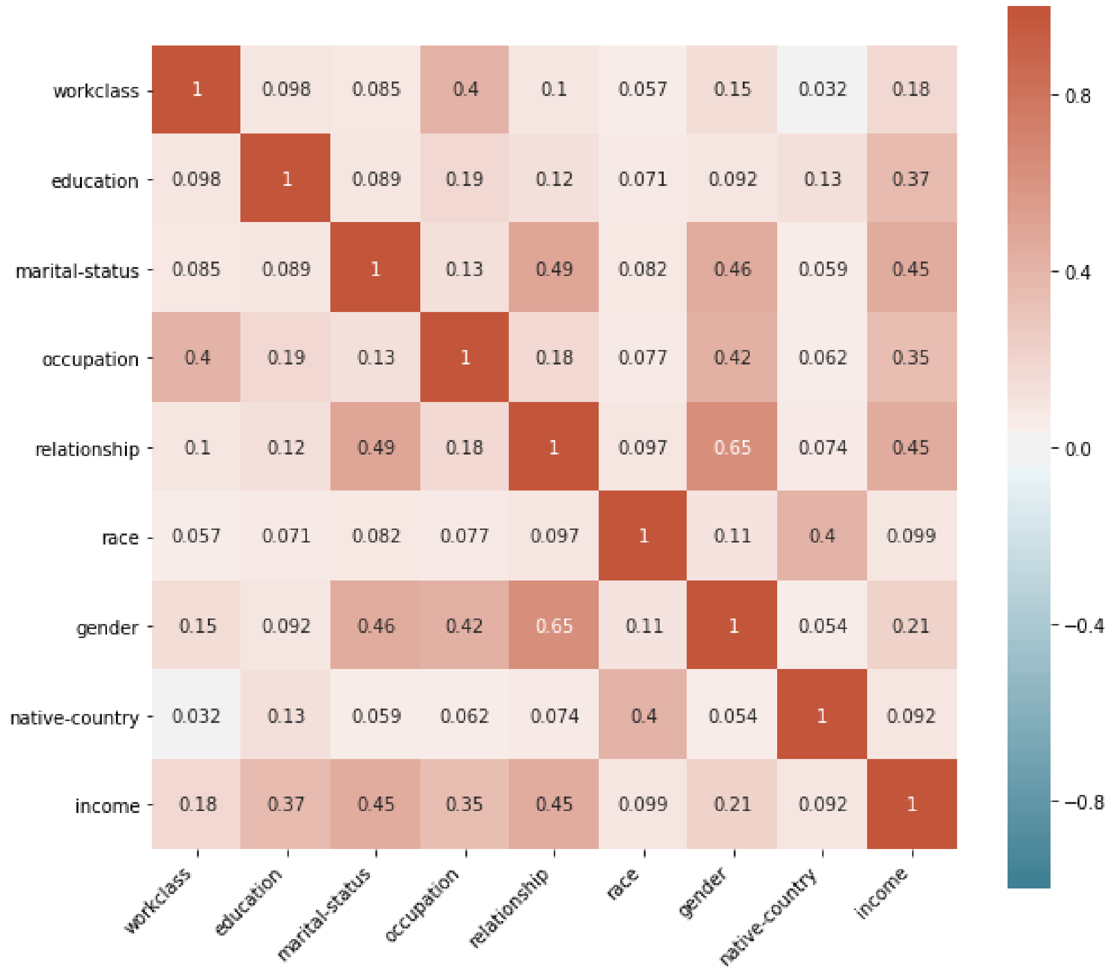
1.8 Analyse variables correlations

[10]:
```
explore.show_df_correlations(df=dataset)
```

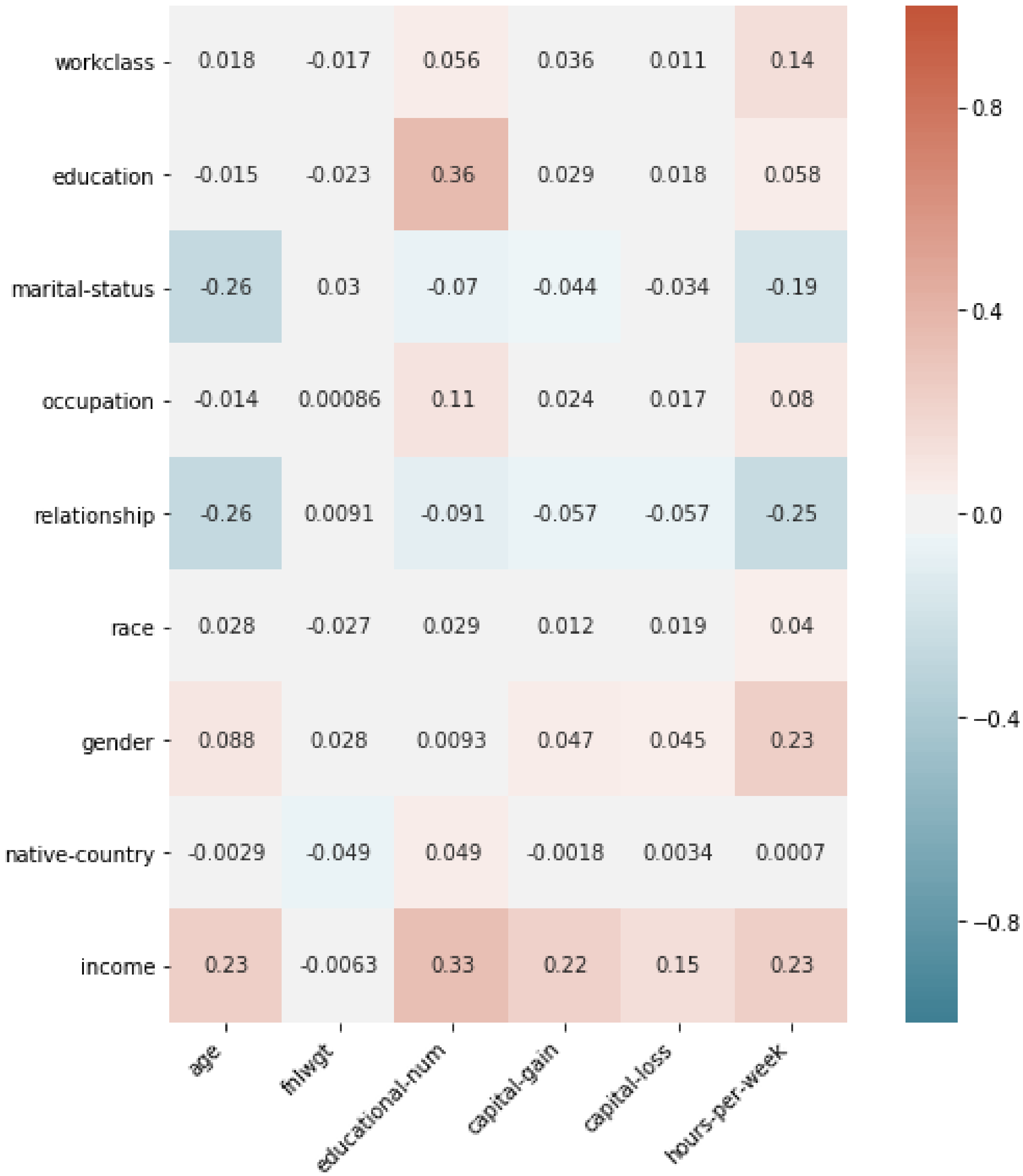
Pearson correlation matrix for numerical variables



Cramers V correlation matrix for categorical variables



Point Biserial correlation matrix for numerical & categorical variables


workclass
education
marital-status
occupation
relationship
race
gender
native-country
income
age
fnlwgt
educational-num
capital-gain
capital-loss
hours-per-week
0.018
-0.017
0.056
0.036
0.011
0.14
-0.015
-0.023
0.36
0.029
0.018
0.058
-0.26
0.03
-0.07
-0.044
-0.034
-0.19
-0.014
0.00086
0.11
0.024
0.017
0.08
-0.26
0.0091
-0.091
-0.057
-0.057
-0.25
0.028
-0.027
0.029
0.012
0.019
0.04
0.088
0.028
0.0093
0.047
0.045
0.23
-0.0029
-0.049
0.049
-0.0018
0.0034
0.0007
0.23
-0.0063
0.33
0.22
0.15
0.23
0.8
0.4
0.0
−0.4
−0.8

The end.