Ex3

April 15, 2020

1 Exercise 3

1.1 Draw histogram for all of the numerical variables

[1]:

```
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sb
from math import sqrt
import numpy as np
from os import system
```

[2]:

```
df = pd.read_csv('/home/hakim/Documents/semester 8/DM/HW_2/adult.csv')
```

1.1.1 divide numerical values

First we divide numerical values and save them in num_colls.

[3]:

```
colls = df.columns

num_colls = []
for c in colls:
    try:
        float(df[c][0])
        print(c, ' can be histogram')
        num_colls.append(str(c))
    except:
        pass
```

```
age  can be histogram
demogweight  can be histogram
education-num  can be histogram
capital-gain  can be histogram
capital-loss  can be histogram
hours-per-week  can be histogram
```

Calculating the plot matrix size for drawing numerical histogram plot.

[4]:

```
coll_count = int(len(num_colls))
row_number = sqrt(coll_count)

if row_number % 1 > 0: row_number = int(row_number) + 1

print('numerical collumns count: ', coll_count)
print('hist row numbers: ', row_number)
print(f'hist matrix {row_number} x {row_number}')
```

```
numerical collumns count:  6
hist row numbers:  3
hist matrix 3 x 3
```

1.1.2 Histogram of all numerical values

below here we draw a histogram of all numerical values of the dataset and see the variety of each of them.

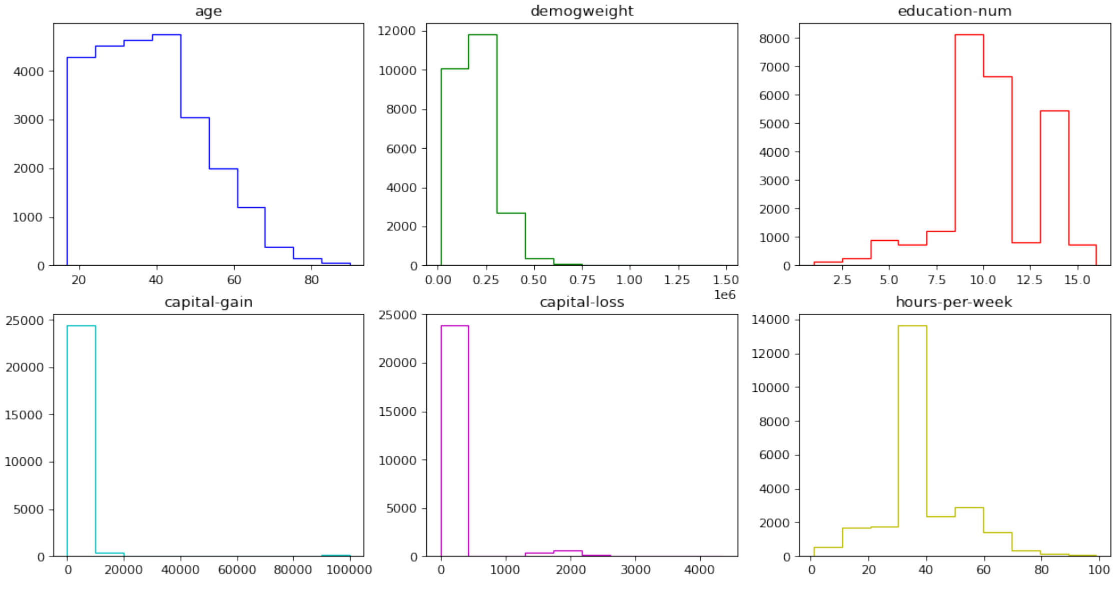
Age As shown in the plot, younger persons have a higher chance of being hired because of the large number of people who are hired in jobs.

education-num most of the people in this dataset have an education number between 8 to 15. so that means that higher education numbers can give the person more chance to be hired.

[8]:

```
plt.figure(figsize=(15,12), dpi=80)
colores = ['b','g', 'r', 'c', 'm', 'y', 'k', 'w']
for i in range(len(num_colls)):
    plt.subplot(row_number,row_number,i+1)
    plt.hist(df[num_colls[i]], histtype='step', color=colores[i])
    plt.title(str(num_colls[i]))

plt.show()
```



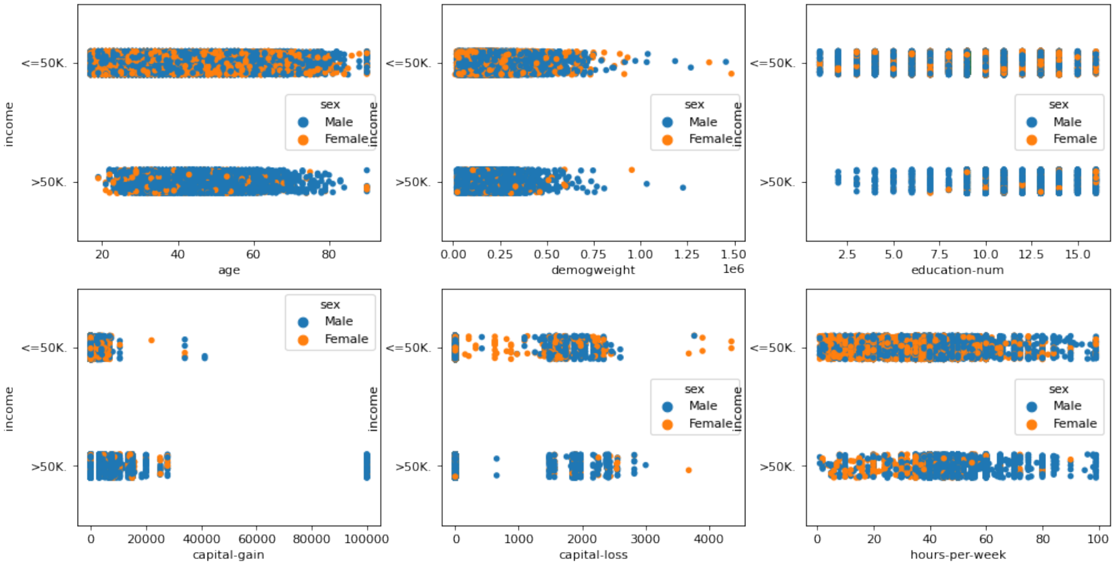
1.1.3 Strip Plot

For a better understanding of income values in this dataset we draw a strip plot for each numerical values with income in the Y position. also for better understanding the relationship between gender and income in each plot.

[6]:

```
target_col = 'income'
plt.figure(figsize=(15,12), dpi=80)
for i in range(len(num_colls)):
    plt.subplot(row_number,row_number,i+1)
    sb.stripplot(x= df[num_colls[i]], y= df[target_col], hue=df['sex'])

plt.show()
```



1.2 conclusion

As shown above, in the strip plot, most of the low income is for men and for higher-income men and women have an equal variety