Importing Libraries

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Adult Salary DataSet -- Taken from UCI's Machine Learning Repository

Importing Dataset

In [2]:

```
df = pd.read_csv("adult_salary.csv")
df.head()
```

Out[2]:

	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
0	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
1	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
2	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
3	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
4	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K

Exploring Dataset ----

In [3]:

```
df.columns
```

Out[3]:

```
Index(['39', ' State-gov', ' 77516', ' Bachelors', ' 13', ' Never-married',
       ' Adm-clerical', ' Not-in-family', ' White', ' Male', ' 2174', ' 0',
       ' 40', ' United-States', ' <=50K'],
      dtype='object')
```

AS we cannot identify columns name properly --- so Redefining Column Name

In [4]:

```
col = ['age','workclass','fnlwgt','education','education-num','marital-
status','occupation','relationship','race','sex',
       'capital-gain','capital-loss','hours-per-week','native-country','Salary']
```

In [5]:

```
df = pd.read_csv("adult_salary.csv",names = col,na_values = ' ?')
```

In [6]:

```
df.head()
```

Out[6]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba

In [7]:

```
df.shape
```

Out[7]:

```
(32561, 15)
```

In [8]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   age             32561 non-null  int64 
 1   workclass       30725 non-null  object
 2   fnlwgt          32561 non-null  int64 
 3   education       32561 non-null  object
 4   education-num   32561 non-null  int64 
 5   marital-status  32561 non-null  object
 6   occupation      30718 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   sex             32561 non-null  object
 10  capital-gain    32561 non-null  int64 
 11  capital-loss    32561 non-null  int64 
 12  hours-per-week  32561 non-null  int64 
 13  native-country  31978 non-null  object
 14  Salary          32561 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

From the "df.info" we got to know that we have 9 String/object columns (not int or float)

Stastical Summary Of DataSet

In [9]:

```
df.describe(include='all')
```

Out[9]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain
count	32561.000000	30725	3.256100e+04	32561	32561.000000	32561	30718	32561	32561	32561	32561.00000
unique	NaN	8	NaN	16	NaN	7	14	6	5	2	Na
top	NaN	Private	NaN	HS-grad	NaN	Married-civ-spouse	Prof-specialty	Husband	White	Male	Na
freq	NaN	22696	NaN	10501	NaN	14976	4140	13193	27816	21790	Na
mean	38.581647	NaN	1.897784e+05	NaN	10.080679	NaN	NaN	NaN	NaN	NaN	1077.64884
std	13.640433	NaN	1.055500e+05	NaN	2.572720	NaN	NaN	NaN	NaN	NaN	7385.29208
min	17.000000	NaN	1.228500e+04	NaN	1.000000	NaN	NaN	NaN	NaN	NaN	0.00000
25%	28.000000	NaN	1.178270e+05	NaN	9.000000	NaN	NaN	NaN	NaN	NaN	0.00000
50%	37.000000	NaN	1.783560e+05	NaN	10.000000	NaN	NaN	NaN	NaN	NaN	0.00000
75%	48.000000	NaN	2.370510e+05	NaN	12.000000	NaN	NaN	NaN	NaN	NaN	0.00000
max	90.000000	NaN	1.484705e+06	NaN	16.000000	NaN	NaN	NaN	NaN	NaN	99999.00000

From the descrption we came to know--

- Maximum people have salary less than 50k
- Maximum people are from United states
- Most of the data involved Male
- Most of them are White (doesn't Influence Salry)

In [10]:

```
df.sort_values(by=['age','education','occupation'],ascending=[False,False,False]).head(10)
```

Out[10]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week
19212	90	Private	139660	Some-college	10	Divorced	Sales	Unmarried	Black	Female	0	0	37
2303	90	Private	52386	Some-college	10	Never-married	Other-service	Not-in-family	Asian-Pac-Islander	Male	0	0	35
5104	90	Private	52386	Some-college	10	Never-married	Other-service	Not-in-family	Asian-Pac-Islander	Male	0	0	35
10210	90	Self-emp-not-inc	282095	Some-college	10	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	40
2891	90	Private	171956	Some-college	10	Separated	Adm-clerical	Own-child	White	Female	0	0	40
12451	90	NaN	225063	Some-college	10	Never-married	NaN	Own-child	Asian-Pac-Islander	Male	0	0	10
8806	90	Private	87372	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	20051	0	72
20610	90	Private	206667	Masters	14	Married-civ-spouse	Prof-specialty	Wife	White	Female	0	0	40
5370	90	Local-gov	227796	Masters	14	Married-civ-spouse	Exec-managerial	Husband	White	Male	20051	0	60
5406	90	Private	51744	Masters	14	Never-married	Exec-managerial	Not-in-family	Black	Male	0	0	50

DROPING THE IRRELEVANT COLUMNS THAT DOEST SHOW ANY EFFECT ON SALARY

In [11]:

```
df.drop(['fnlwgt','marital-status','race','relationship'],axis=1,inplace=True)
```

In [12]:

```
df.head()
```

Out[12]:

	age	workclass	education	education-num	occupation	sex	capital-gain	capital-loss	hours-per-week	native-country	Salary
0	39	State-gov	Bachelors	13	Adm-clerical	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	Bachelors	13	Exec-managerial	Male	0	0	13	United-States	<=50K
2	38	Private	HS-grad	9	Handlers-cleaners	Male	0	0	40	United-States	<=50K
3	53	Private	11th	7	Handlers-cleaners	Male	0	0	40	United-States	<=50K
4	28	Private	Bachelors	13	Prof-specialty	Female	0	0	40	Cuba	<=50K

In [13]:

```
df.mean()
```

Out[13]:

```
age                38.581647
education-num      10.080679
capital-gain     1077.648844
capital-loss       87.303830
hours-per-week     40.437456
dtype: float64
```

In [14]:

```
df.groupby(['sex','Salary']).mean()
```

Out[14]:

		age	education-num	capital-gain	capital-loss	hours-per-week
sex	**Salary**					
Female	**<=50K**	36.210801	9.820475	121.986134	47.364470	35.916701
	>50K	42.125530	11.787108	4200.389313	173.648855	40.426633
Male	**<=50K**	37.147012	9.452142	165.723823	56.806782	40.693879
	>50K	44.625788	11.580606	3971.765836	198.780396	46.366106

Checking Missing Values ---

In [15]:

```
df.isnull().sum()
```

Out[15]:

```
age                0
workclass       1836
education          0
education-num      0
occupation      1843
sex                0
capital-gain       0
capital-loss       0
hours-per-week     0
```

```
native-country     583
Salary               0
dtype: int64
```

This dataset has missing values

In [16]:

```
df.dropna(inplace=True)
```

In [17]:

```
df.isnull().sum()
```

Out[17]:

```
age               0
workclass         0
education         0
education-num     0
occupation        0
sex               0
capital-gain      0
capital-loss      0
hours-per-week    0
native-country    0
Salary            0
dtype: int64
```

All the missing values are removed --- NOW

DataVizualization

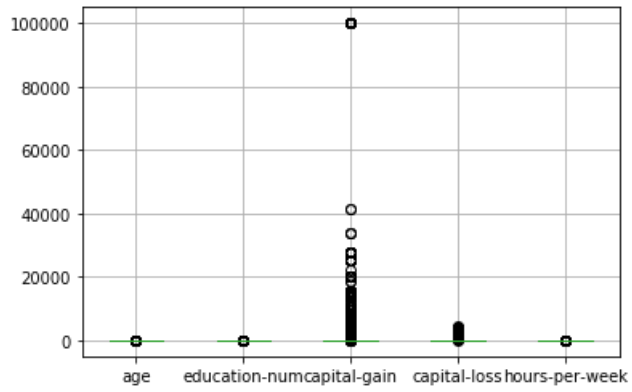
1. BOXPLOT

In [18]:

```
df.boxplot()
```

Out[18]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1edf8c141c8>
```



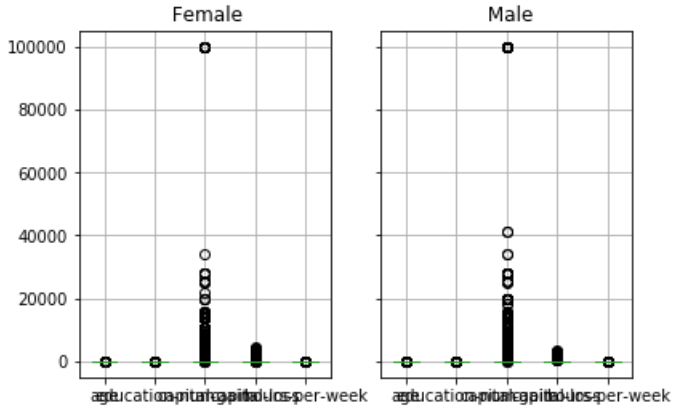
In [19]:

```
df.groupby('sex').boxplot()
```

Out[19]:

```
Female         AxesSubplot(0.1,0.15;0.363636x0.75)
Male      AxesSubplot(0.536364,0.15;0.363636x0.75)
```

```
Male      AxesSubplot(0.536364,0.15;0.363636x0.75)
dtype: object
```



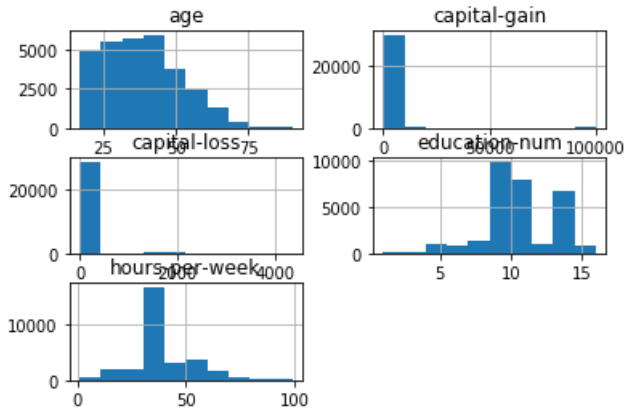
2. HISTOGRAM

In [20]:

```
df.hist(grid='off')
```

Out[20]:

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFA786EC8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFA81CE88>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFA856608>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFA890048>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFA8C5A08>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFA8FEA48>]],
      dtype=object)
```

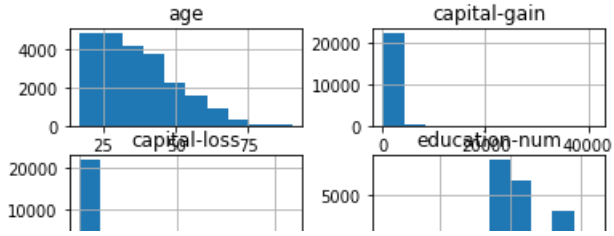


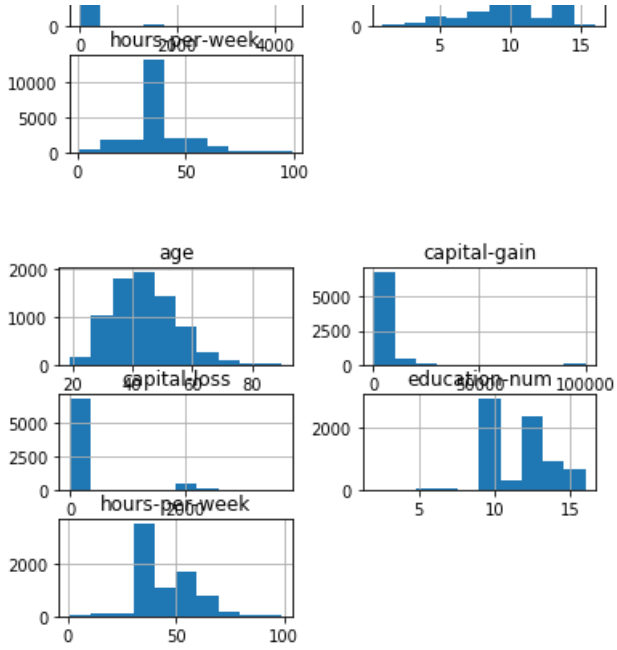
In [21]:

```
df.groupby('Salary').hist(grid='off')
```

Out[21]:

```
Salary
 <=50K    [[AxesSubplot(0.125,0.670278;0.336957x0.209722...
 >50K     [[AxesSubplot(0.125,0.670278;0.336957x0.209722...
dtype: object
```





In [ ]:

```
```

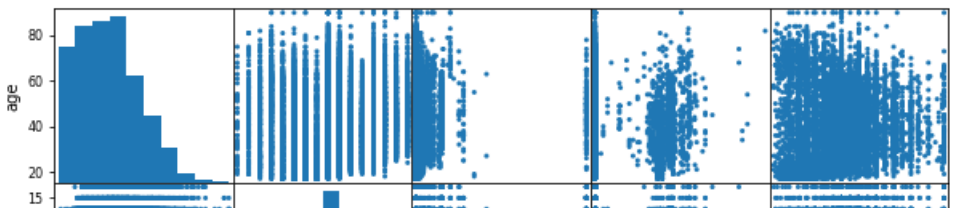
3.SCATTERPLOT

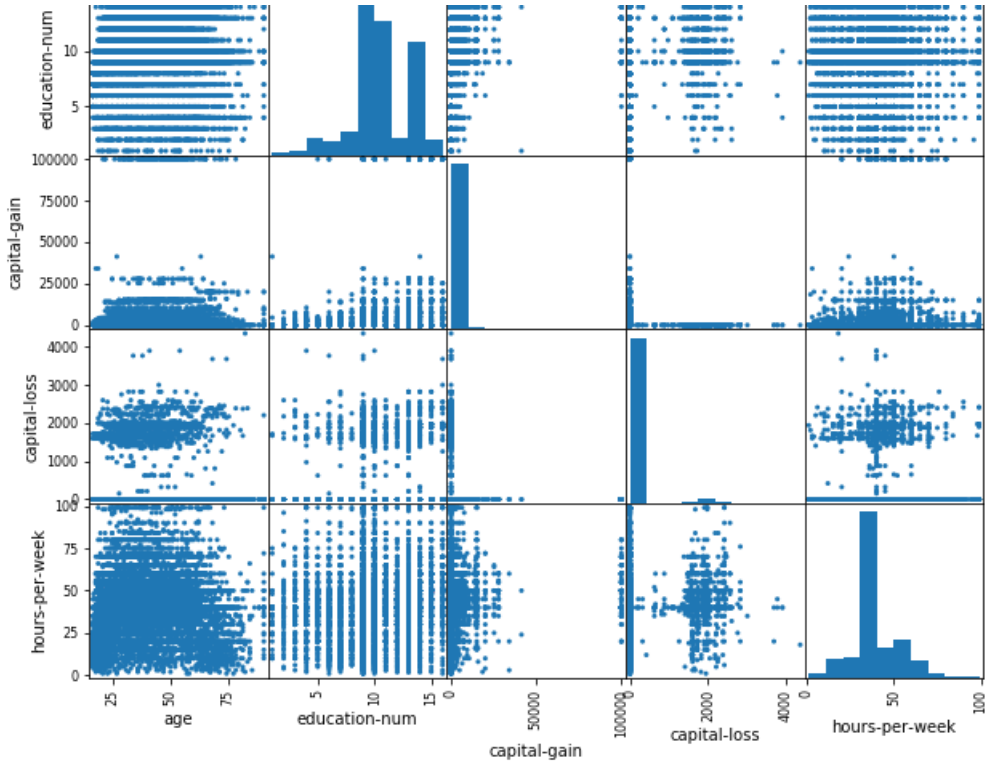
In [22]:

```
from pandas.plotting import scatter_matrix
scatter_matrix(df,alpha=1,figsize=(10,10),diagonal='hist')
```

Out[22]:

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFA6CA088>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBC090C8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBC37888>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBC6F2C8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBCA7C88>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBCE5E88>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBD1EFC8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBD516C8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBD5D2C8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBD97488>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBDFCA08>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBE35A88>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBE6EB88>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBEA8CC8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBEE2DC8>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBF1AEC8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBF51F88>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBF8C0C8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBFC61C8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFBFFF308>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFC036408>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFC06F4C8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFC0A8048>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFC0DF148>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001EDFC118248>]],
      dtype=object)
```




education-num
capital-gain
capital-loss
hours-per-week
age
education-num
capital-gain
capital-loss
hours-per-week

In [ ]: