1. Рубежный контроль №1

Наинг Ко Ко Линн, группа ИУ5-21М.Вариант №3, набор данных №2.

1.1. Задание

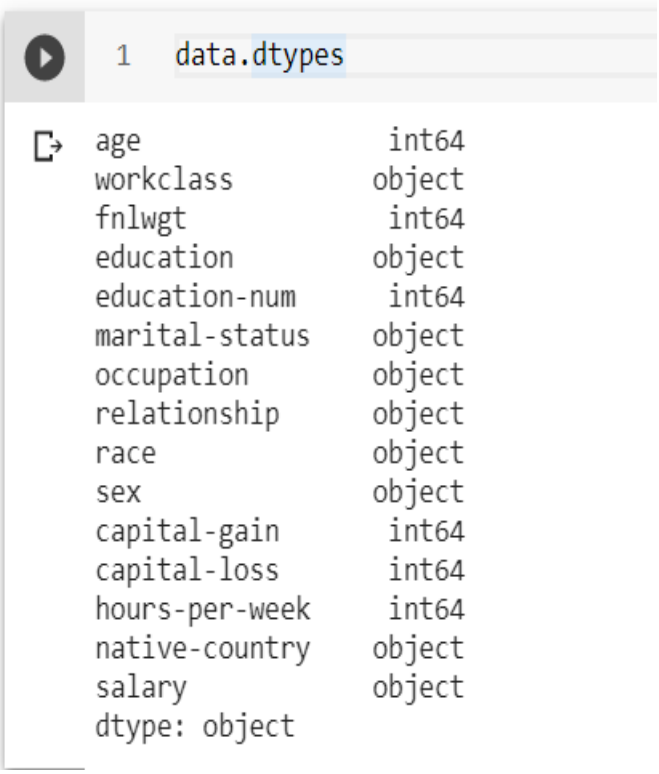
In [1]:
```
!pip install pandasql
from pandasql import sqldf
import pandas as pd
```

In [2]:
```
import numpy as np

import pandas as pd
import seaborn as sns
%matplotlib inline
```

In [3]:
```
data = pd.read_csv("adult.data.csv")
```

In [4]:
```
data.dtypes
```



```
In [5]: data.head()
```

```
1 data.head()
```

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

Для заданного набора данных произведите масштабирование данных (для одного признака) и преобразование категориальных признаков в количественные двумя способами (label encoding, one hot encoding) для одного признака. Какие методы Вы использовали для решения задачи и почему?

1.2. Решение

1.2.1. Загрузка и предобработка данных

```
In [5]: data.shape
```

```
1 data.shape

(32561, 15)
```

```
In [6]: data.isnull().sum()
```

```
1 data.isnull().sum()

age               0
workclass         0
fnlwgt            0
education         0
education-num     0
marital-status    0
occupation        0
relationship      0
race              0
sex               0
capital-gain      0
capital-loss      0
hours-per-week    0
native-country    0
salary            0
dtype: int64
```

```
In [7]: d = data[["age", "sex", "fnlwgt"]]

d = d.dropna(axis=0, how="any")
In [8]: d.head()
```

```
[ ]  1  d.head()
```

	age	sex	fnlwgt
0	39	Male	77516
1	50	Male	83311
2	38	Male	215646
3	53	Male	234721
4	28	Female	338409

```
In [9]: d.shape
```

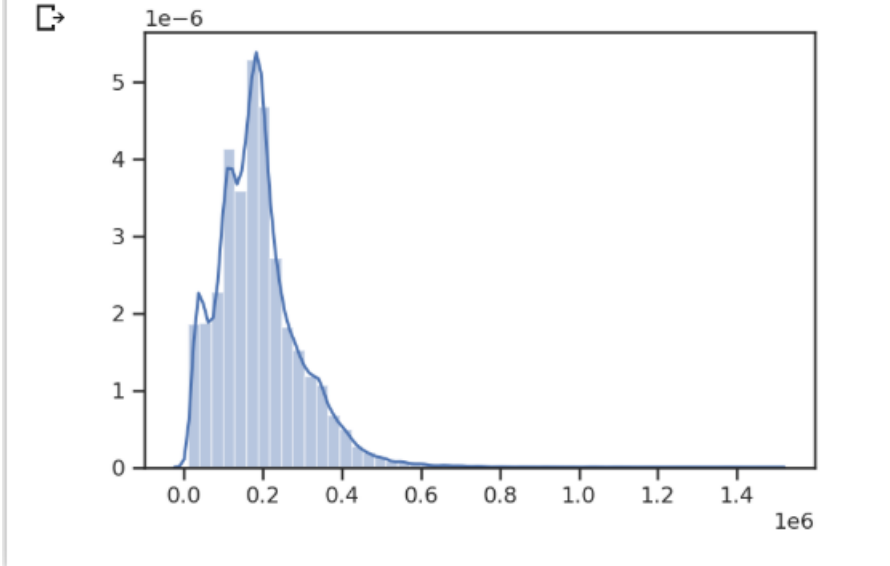
```
[ ]  1  d.shape
```

```
(32561, 3)
```

1.2.2. Масштабирование данных

```
In [10]: sns.distplot(d[["fnlwgt"]]);
```

```
1  sns.distplot(d[["fnlwgt"]]);
```

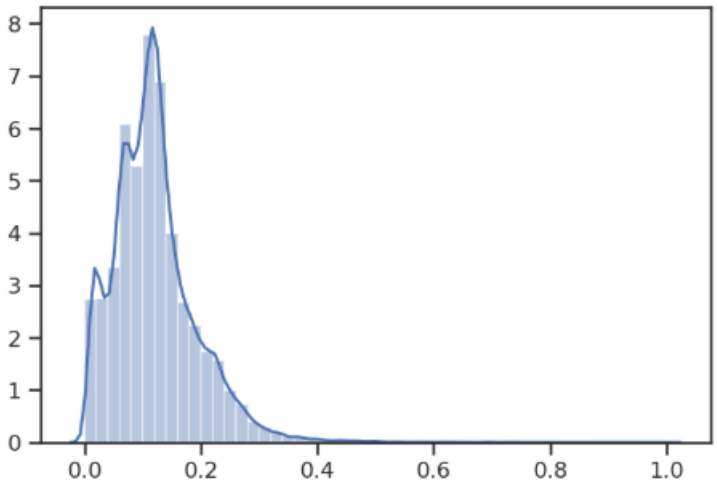



```
In [11]:  from sklearn.preprocessing import MinMaxScaler

          sc = MinMaxScaler()
          sc_data = sc.fit_transform(d[["fnlwgt"]])
          sns.distplot(sc_data)
```

```
from sklearn.preprocessing import MinMaxScaler
sc = MinMaxScaler()
sc_data = sc.fit_transform(d[["fnlwgt"]])
sns.distplot(sc_data)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe6cb386668>
```




```
In [12]: d["APPEARANCES_SCALED"] = sc_data
```

1.2.3. Преобразование категориальных признаков

```
In [13]: from sklearn.preprocessing import LabelEncoder, OneHotEncoder
```

Label encoding

```
In [14]: le = LabelEncoder()

sex_le = le.fit_transform(d["sex"])
In [15]: np.unique(sex_le)
```

```
np.unique(sex_le)
```

```
array([0, 1])
```

```
In [16]: le.inverse_transform(np.unique(sex_le))
```

```
le.inverse_transform(np.unique(sex_le))
```

```
array(['Female', 'Male'], dtype=object)
```

```
In [17]: d["sex_INDEX"] = sex_le
```

One Hot encoding

```
In [18]: ohe = OneHotEncoder()
sex_ohe = ohe.fit_transform(d[["sex"]])
```

In [19]: sex_ohe.todense()[0:10]

```
1  sex_ohe.todense()[0:10]
```

```
matrix([[0., 1.],
        [0., 1.],
        [0., 1.],
        [0., 1.],
        [1., 0.],
        [1., 0.],
        [1., 0.],
        [0., 1.],
        [1., 0.],
        [0., 1.]])
```

In [20]: d["sex"].head(10)

```
1  d["sex"].head(10)
```

```
0      Male
1      Male
2      Male
3      Male
4    Female
5    Female
6    Female
7      Male
8    Female
9      Male
Name: sex, dtype: object
```

In [21]: ohe_names = ohe.get_feature_names()

ohe_names

```
1  ohe_names = ohe.get_feature_names()
2  ohe_names
```

```
array(['x0_Female', 'x0_Male'], dtype=object)
```

In [22]: for idx, name in enumerate(ohe_names):

  d[name] = sex_ohe[:, idx].todense()

1.2.4. Получившийся набор данных

```
In [23]: d.head(10)
```

```
1  d.head(10)
```

	age	sex	fnlwgt	fnlwgt_SCALED	sex_INDEX	x0_Female	x0_Male
0	39	Male	77516	0.044302	1	0.0	1.0
1	50	Male	83311	0.048238	1	0.0	1.0
2	38	Male	215646	0.138113	1	0.0	1.0
3	53	Male	234721	0.151068	1	0.0	1.0
4	28	Female	338409	0.221488	0	1.0	0.0
5	37	Female	284582	0.184932	0	1.0	0.0
6	49	Female	160187	0.100448	0	1.0	0.0
7	52	Male	209642	0.134036	1	0.0	1.0
8	31	Female	45781	0.022749	0	1.0	0.0
9	42	Male	159449	0.099947	1	0.0	1.0