РК №1 ИУ5-22М Демьянчук Г.В. Вариант 4 - Задача 1 - Набор данных 4

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

```
/usr/local/lib/python3.6/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning:
  import pandas.util.testing as tm
```

```
data = pd.read_csv('/content/toy_dataset.csv', sep=",")
```

New Section

```
data.head()
```

	Number	City	Gender	Age	Income	Illness
0	1	Dallas	Male	41	40367.0	No
1	2	Dallas	Male	54	45084.0	No
2	3	Dallas	Male	42	52483.0	No
3	4	Dallas	Male	40	40941.0	No
4	5	Dallas	Male	46	50289.0	No

```
data.shape
```

```
(150000, 6)
```

```
total_count = data.shape[0]
print('Всего строк: {}'.format(total_count))
```

```
Всего строк: 150000
```

```
data.dtypes
```

```
Number       int64
City        object
Gender      object
Age          int64
Income     float64
Illness     object
dtype: object
```

```
# Проверим наличие пустых значений
# Цикл по колонкам датасета
for col in data.columns:
    # Количество пустых значений - все значения заполнены
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

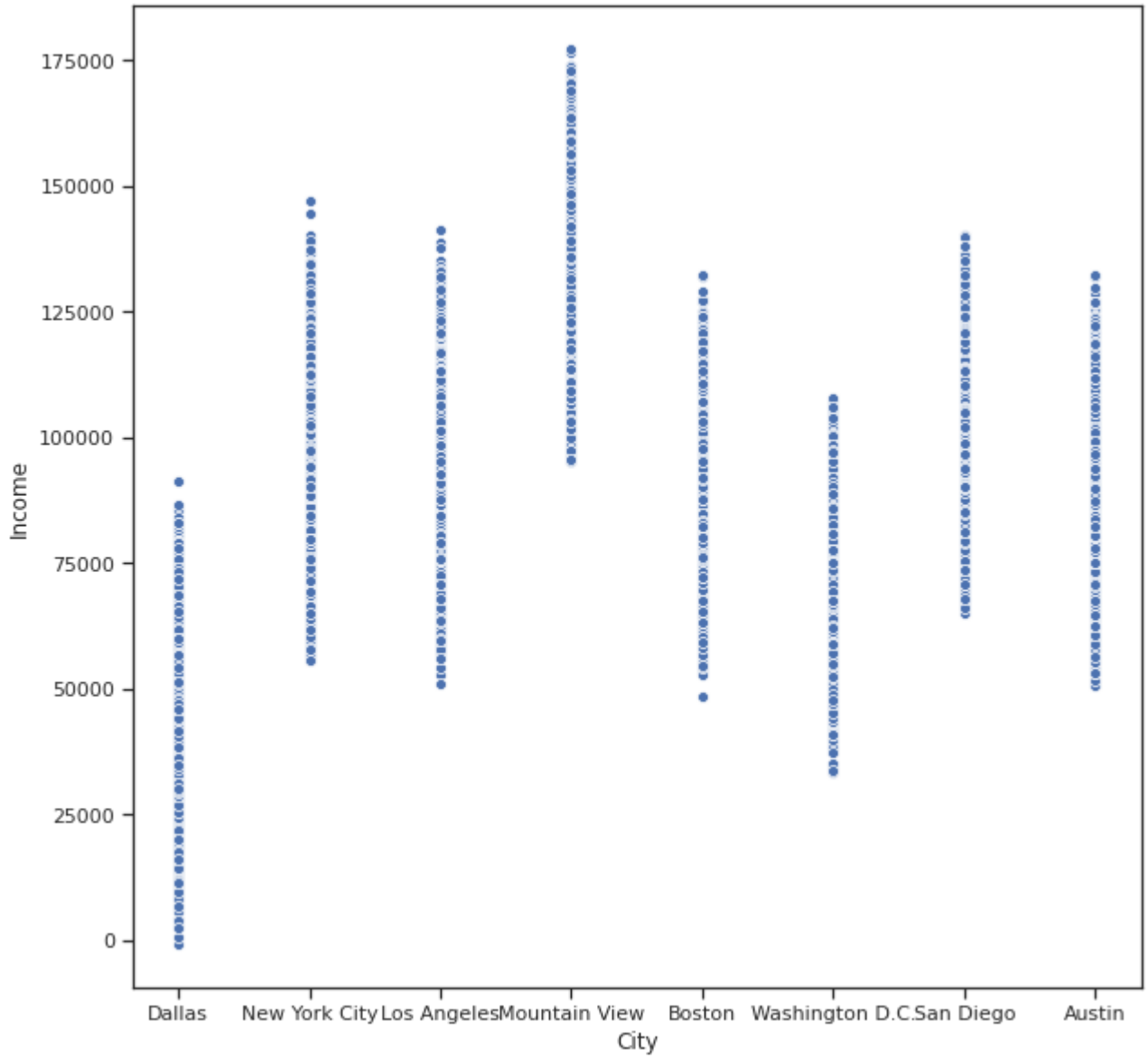
```
Number - 0
City - 0
Gender - 0
Age - 0
Income - 0
Illness - 0
```

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

```
Number     0
City       0
Gender     0
Age        0
Income     0
Illness    0
dtype: int64
```

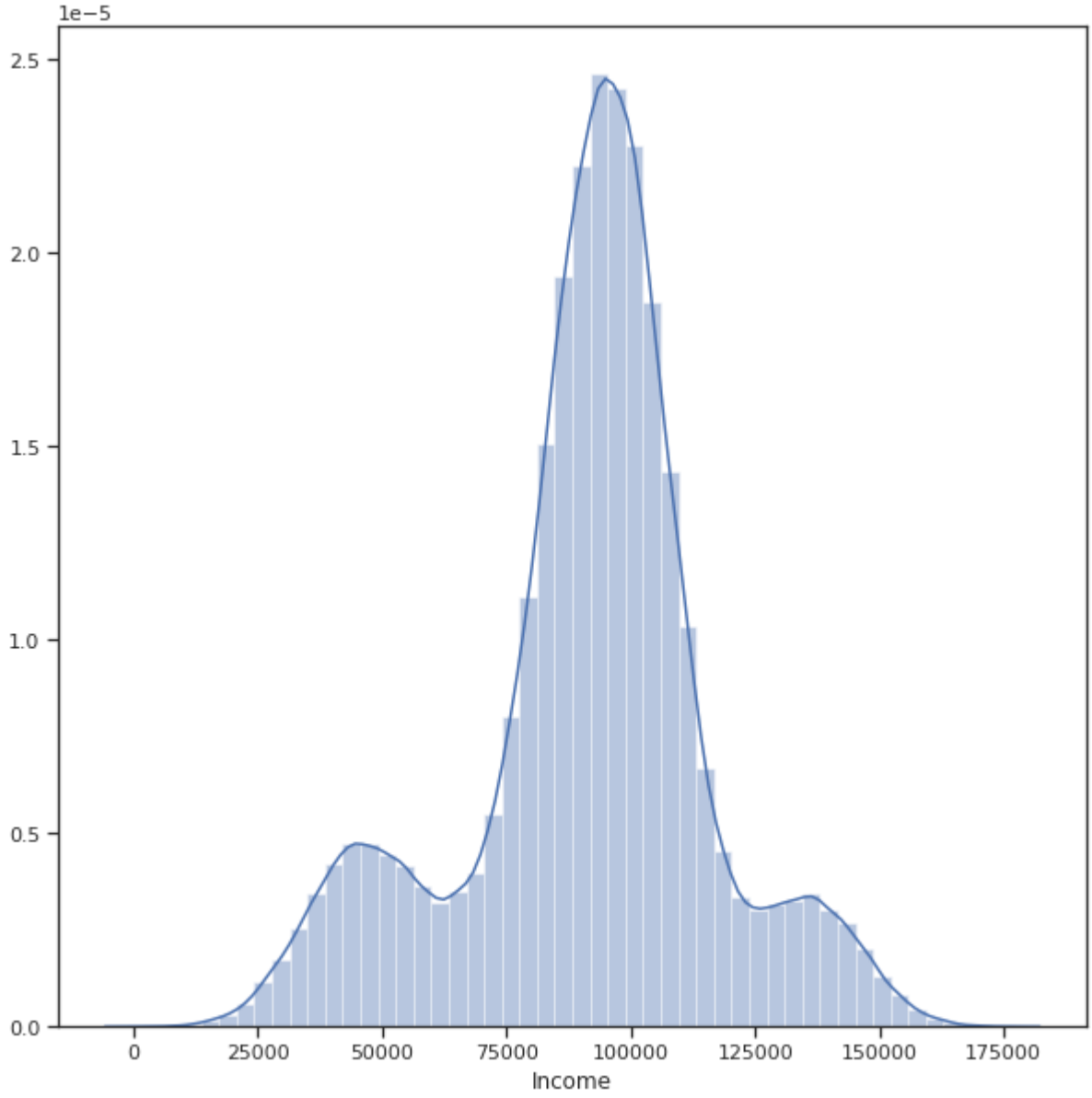
```
#Диаграмма рассеяния для City и Income
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='City', y='Income', data=data)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7838332e48>
```



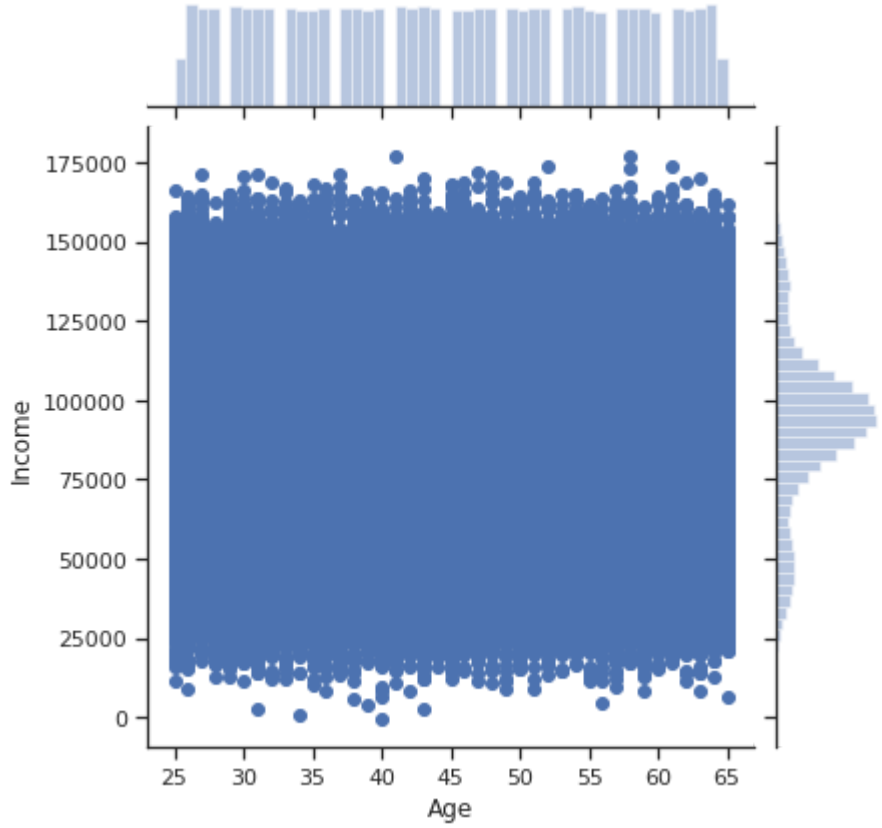

```
#Гистограмма
fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(data['Income'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7835a8ccf8>
```




```
#Комбинация гистограмм и диаграмм рассеивания
sns.jointplot(x='Age', y='Income', data=data)
```

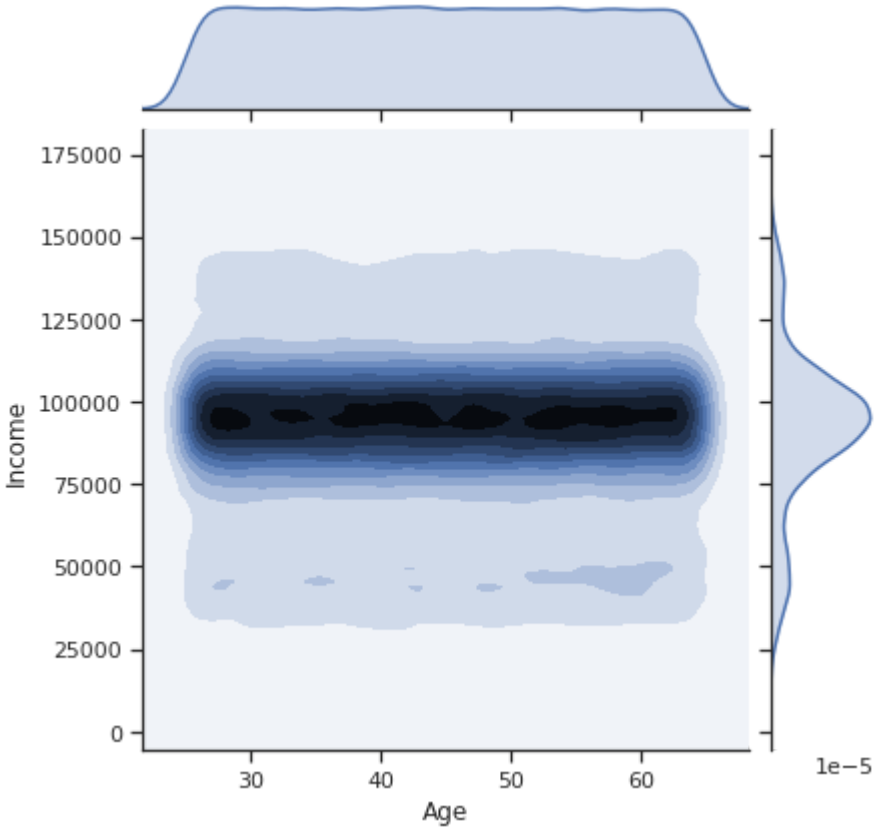
```
<seaborn.axisgrid.JointGrid at 0x7f7835544278>
```




```
sns.jointplot(x='Age', y='Income', data=data, kind="hex")
```

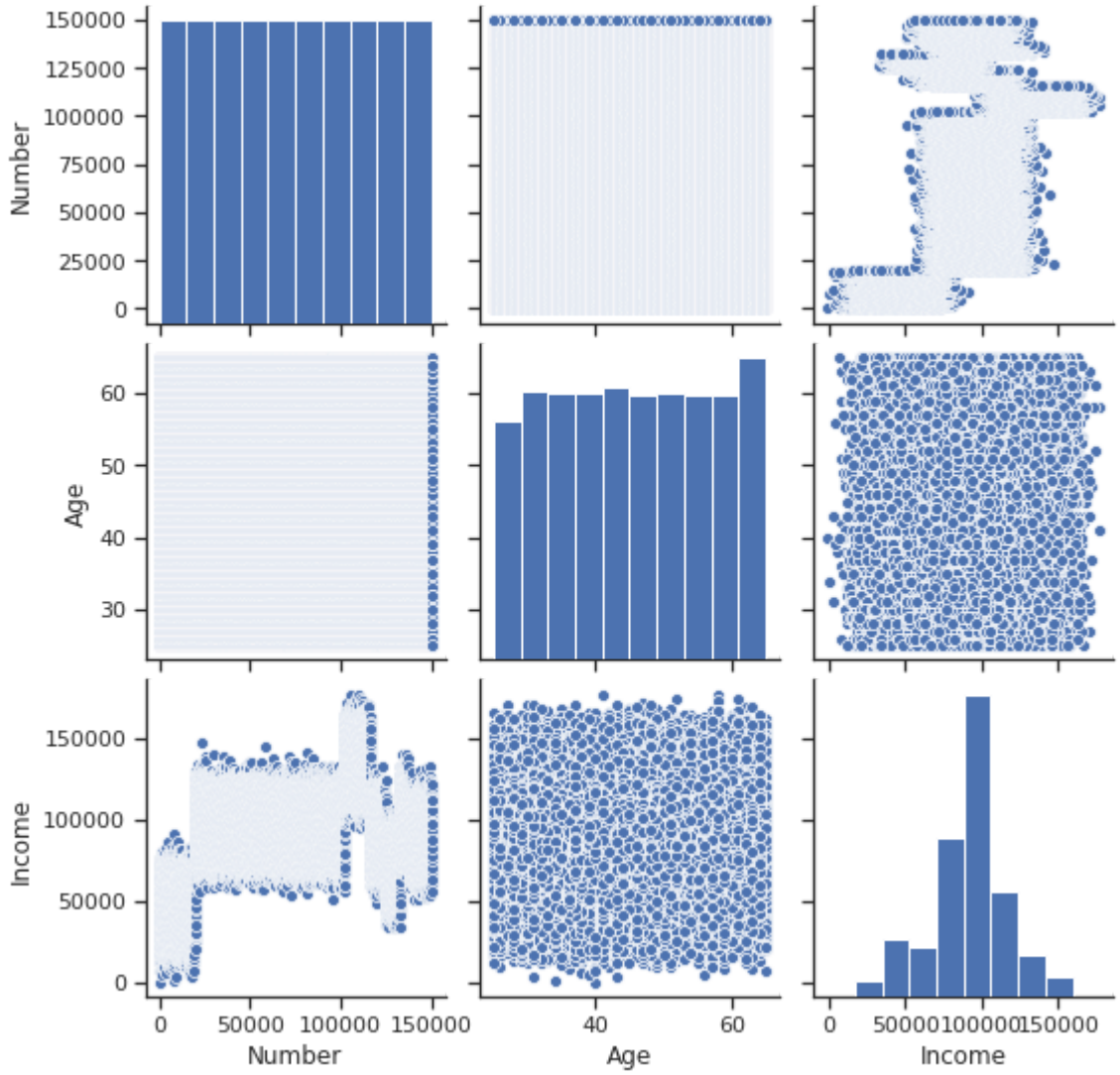
```
sns.jointplot(x='Age', y='Income', data=data, kind="kde")
```

```
<seaborn.axisgrid.JointGrid at 0x7f78351510b8>
```



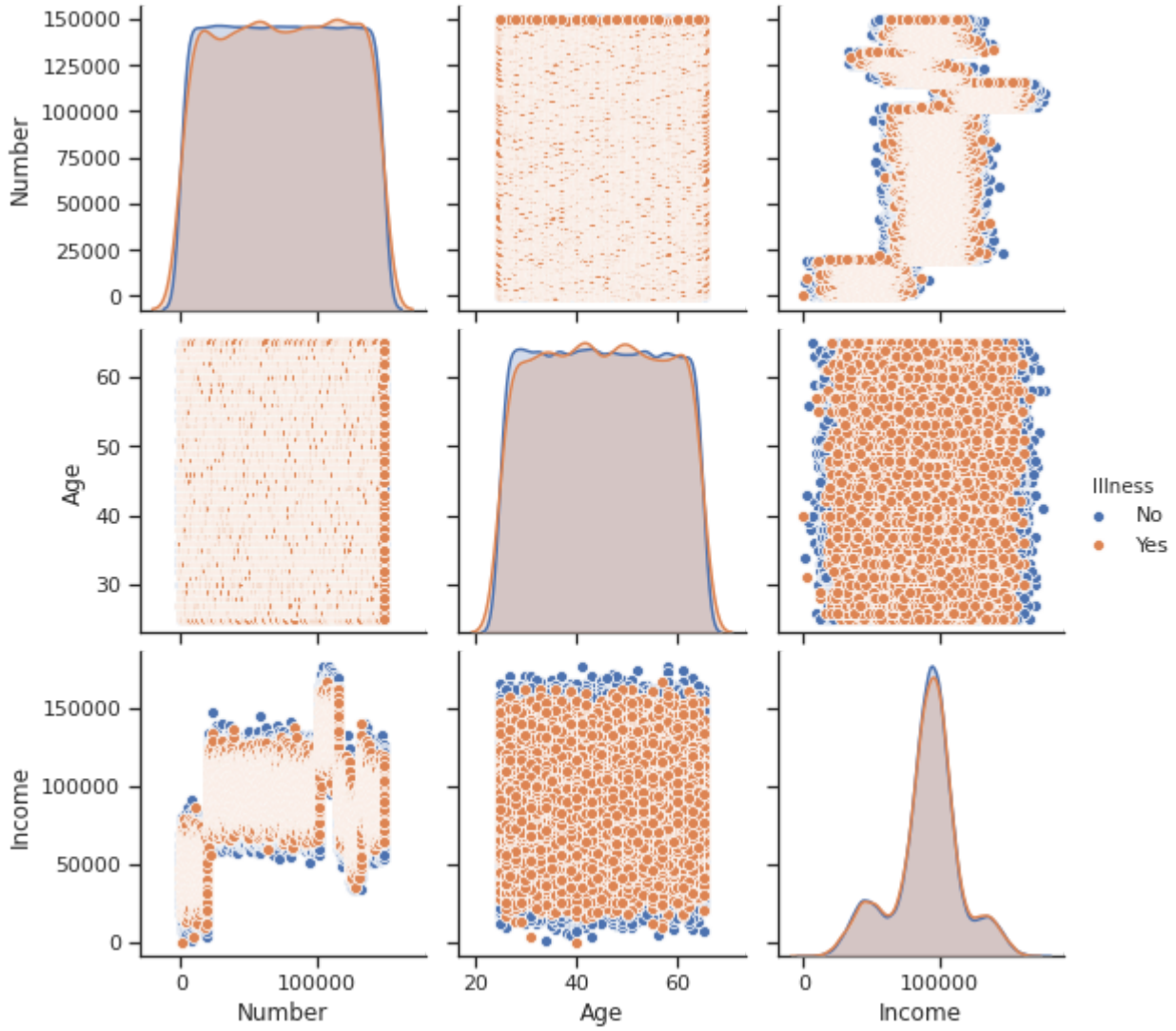

```
#Парные диаграммы - комбинация гистограмм и диаграмм рассеивания для всего набора данных
sns.pairplot(data)
```

```
<seaborn.axisgrid.PairGrid at 0x7f78351515f8>
```



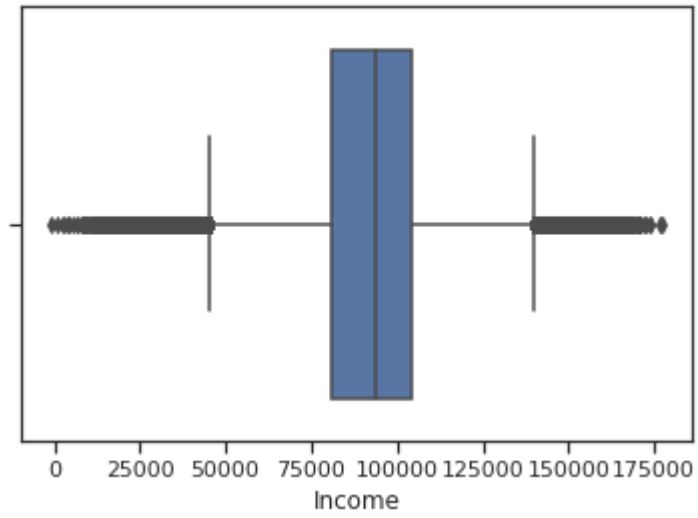

```
#С помощью параметра "hue" возможна группировка по значениям какого-либо признака
sns.pairplot(data, hue="Illness")
```

```
<seaborn.axisgrid.PairGrid at 0x7f78331d2ba8>
```



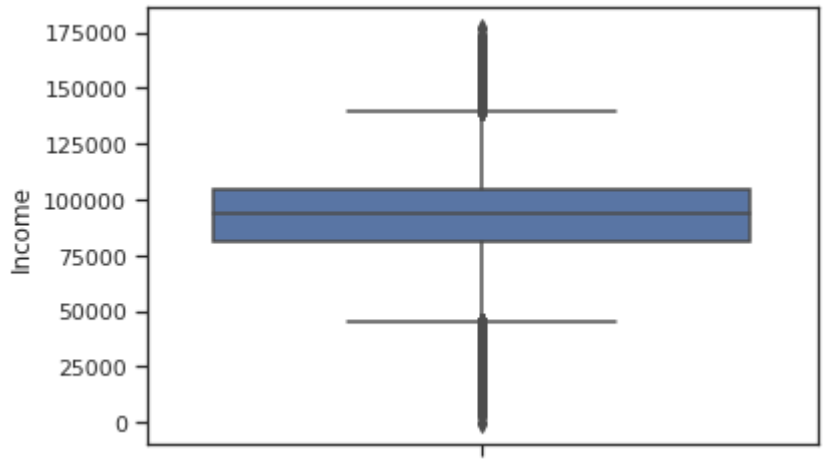

```
#Ящик с усами
sns.boxplot(x=data['Income'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7832f7f8d0>
```



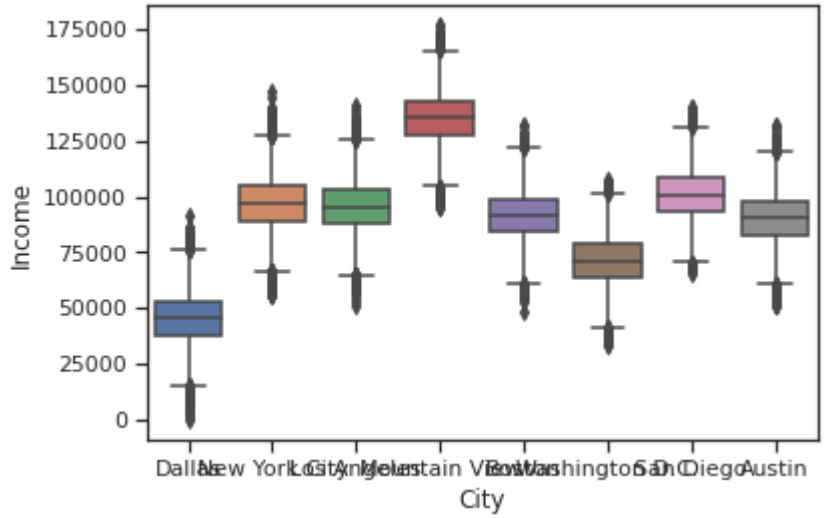
```
#Ящик с усами
sns.boxplot(y=data['Income'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7832db1da0>
```



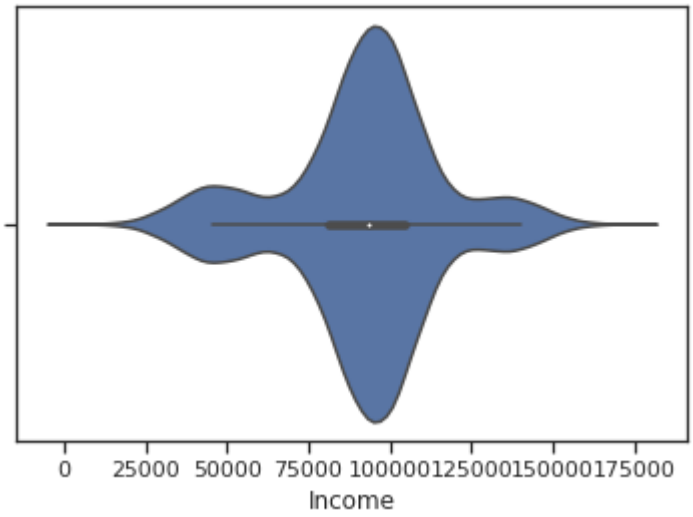

```
# Распределение параметра Income сгруппированные по City.
sns.boxplot(x='City', y='Income', data=data)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7832d1b7f0>
```



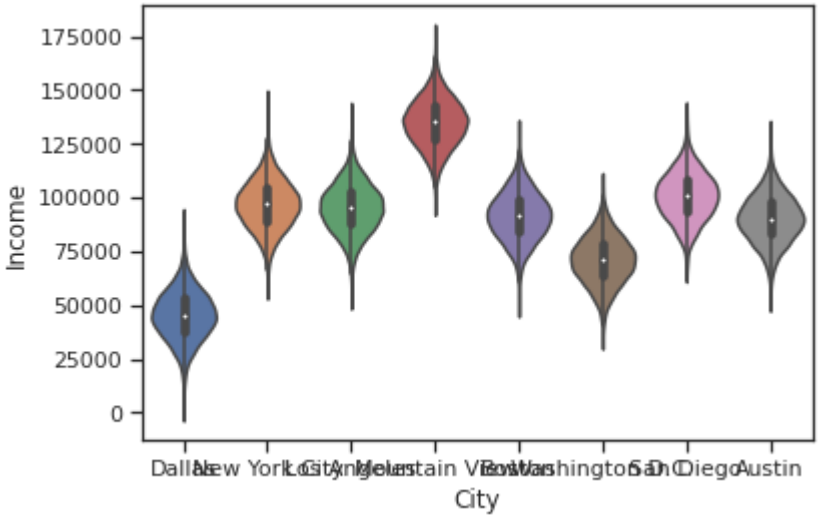

```
#Violin plot
sns.violinplot(x=data['Income'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7832d136a0>
```




```
# Распределение параметра Income сгруппированные по City.
sns.violinplot(x='City', y='Income', data=data)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7832bfc6a0>
```

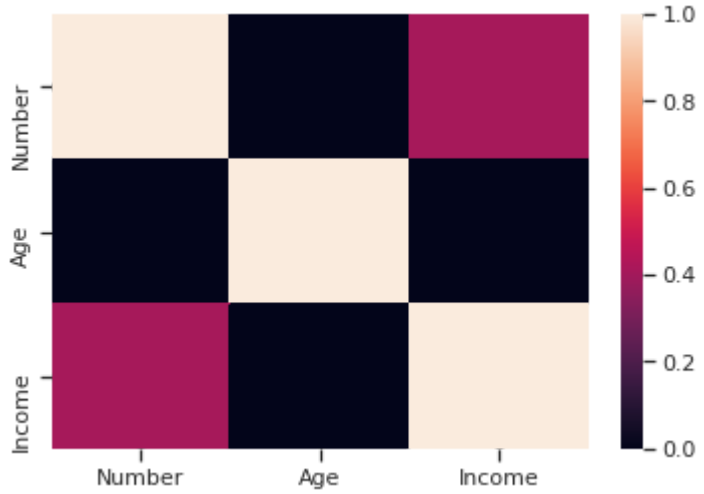



```
#Корреляция
data.corr()
```

	Number	Age	Income
Number	1.000000	-0.003448	0.410460
Age	-0.003448	1.000000	-0.001318
Income	0.410460	-0.001318	1.000000

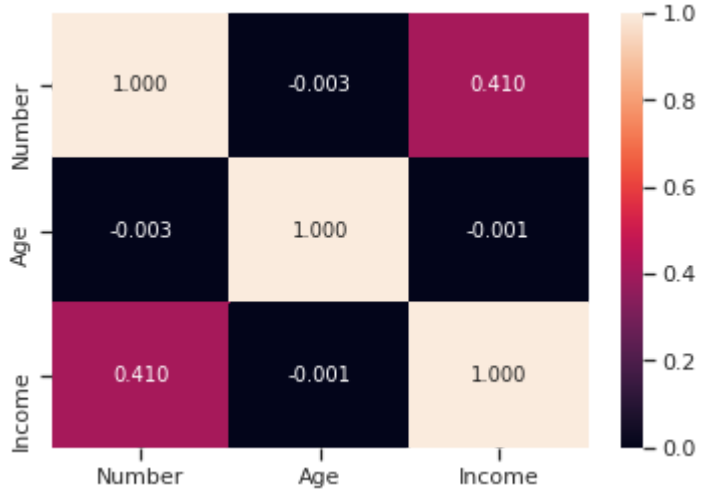
```
sns.heatmap(data.corr())
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7832b36e10>
```




```
# Вывод значений в ячейках
sns.heatmap(data.corr(), annot=True, fmt='.3f')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f78317c7240>
```



Я построил: диаграмму рассеяния - позволяет обнаружить наличие зависимости, гистограмму вероятности распределения данных, jointplot - комбинация гистограмм и диаграмм рассеивания гистограмм и диаграмм рассеивания для всего набора данных, ящик с усами - отображает одн

Выводы по диаграммам: наибольший Income - в городе Mountain View, наименьший - Dallas. На 100000.

Income отчасти коррелирует с Number, слабо коррелирует с Age.