Московский государственный технический университет имени Н. Э. Баумана

Факультет «Информатика и системы управления»

Кафедра ИУ5

Отчёт по

Рубежному контролю № 1

Выполнил:

Елисеев В. Б

Группа РТ5-61Б

Москва

2021

Рубежный контроль №1

Вариант 4. Задание 1.
Задача №1.

Для заданного набора данных проведите корреляционный анализ. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

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

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

# размер датасета
data.shape

(150000, 6)

# Number: A simple index number for each row
# City: The location of a person (Dallas, New York City, Los Angeles,
Mountain View, Boston, Washington D.C., San Diego and Austin)
# Gender: Gender of a person (Male or Female)
# Age: The age of a person (Ranging from 25 to 65 years)
# Income: Annual income of a person (Ranging from -674 to 177175)
# Illness: Is the person Ill? (Yes or No)
data.dtypes

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

# проверка на пробелы
data.isnull().sum()

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

```
# посмотрим первые значения
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

```
# стат. хар-ки датасета
data.describe()
```

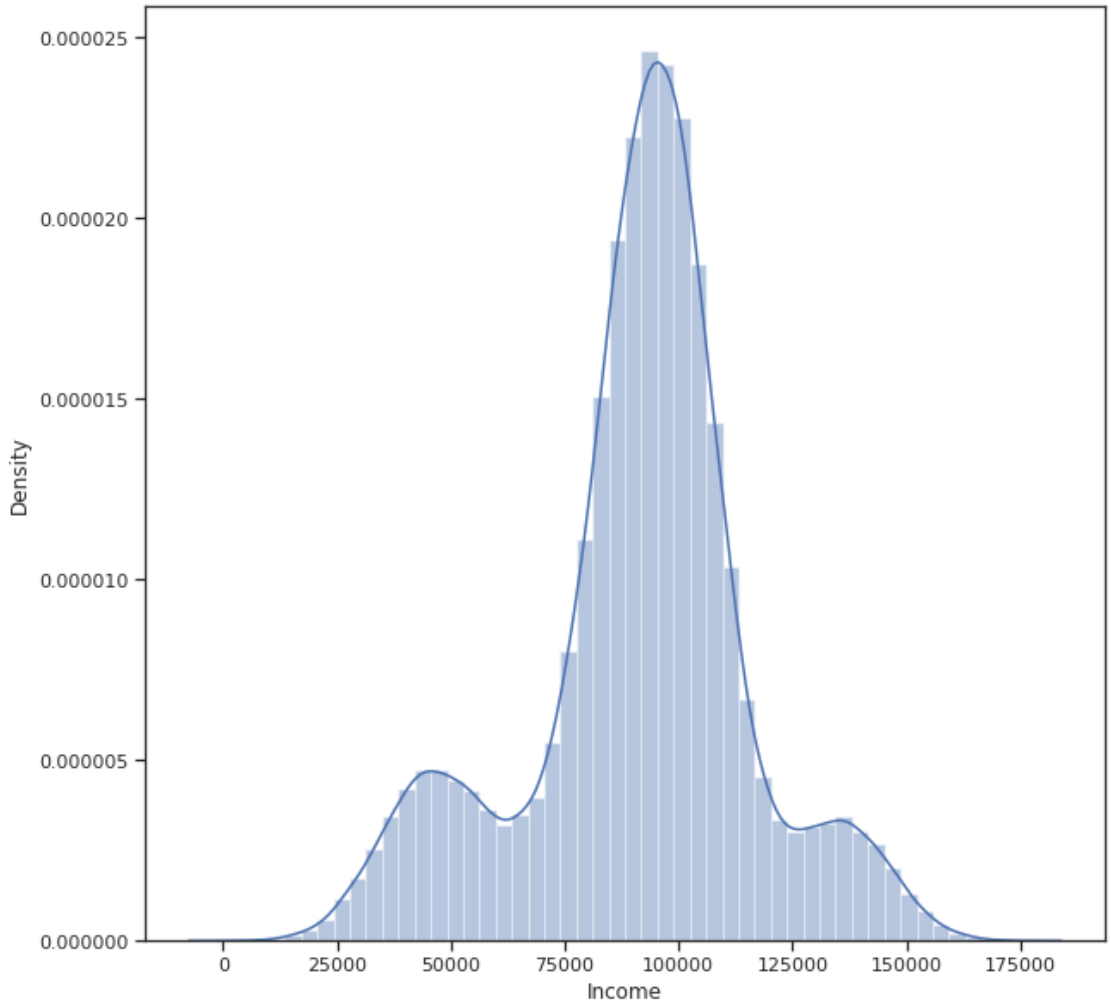
	Number	Age	Income
count	150000.000000	150000.000000	150000.000000
mean	75000.500000	44.950200	91252.798273
std	43301.414527	11.572486	24989.500948
min	1.000000	25.000000	-654.000000
25%	37500.750000	35.000000	80867.750000
50%	75000.500000	45.000000	93655.000000
75%	112500.250000	55.000000	104519.000000
max	150000.000000	65.000000	177157.000000

```
# гисотграмма

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

```
/srv/conda/envs/notebook/lib/python3.7/site-
packages/seaborn/distributions.py:2557: FutureWarning: `distplot` is a
deprecated function and will be removed in a future version. Please adapt
your code to use either `displot` (a figure-level function with similar
flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb0b63fc450>
```

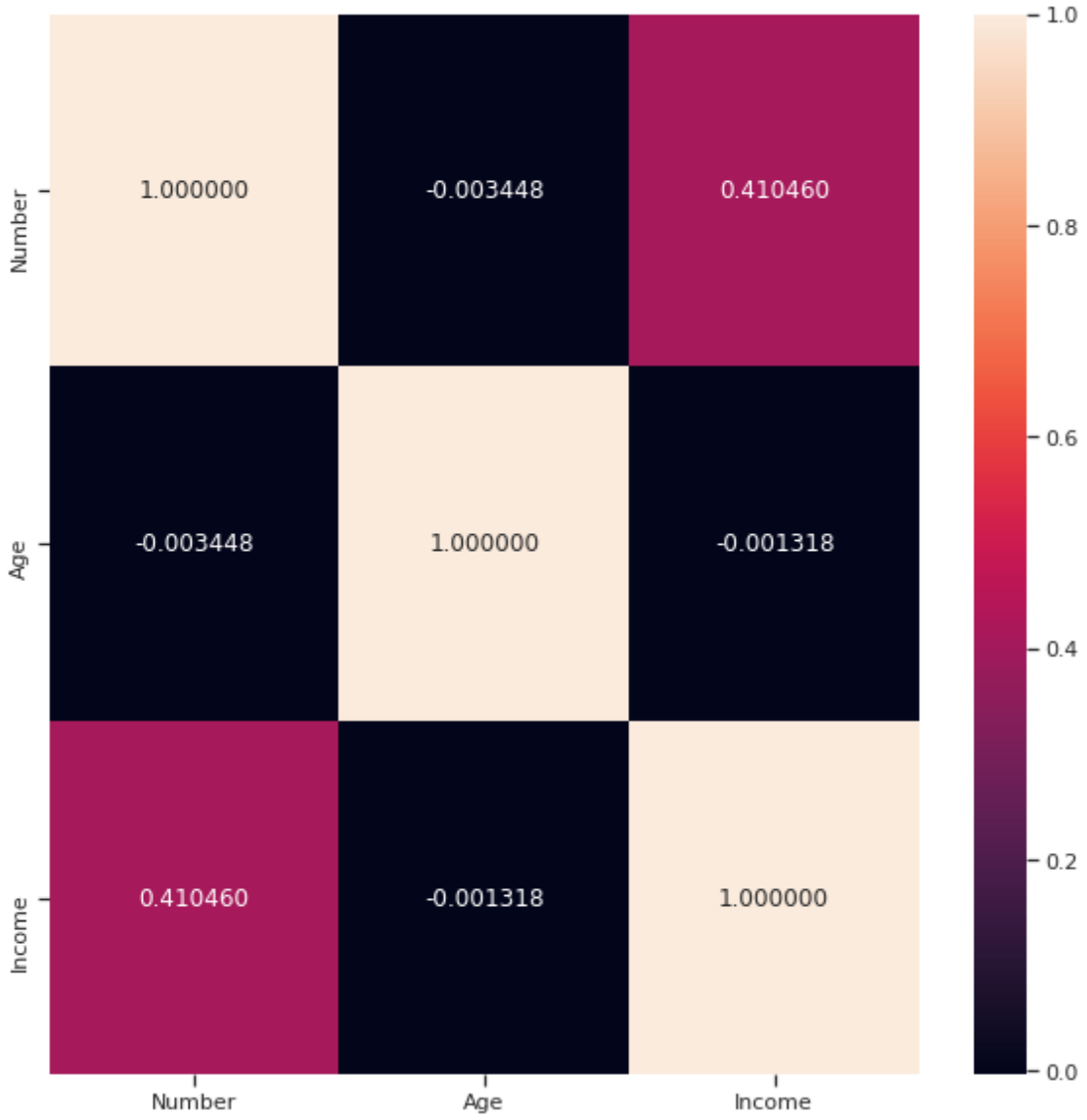



```
# корреляция
data.corr()
```

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

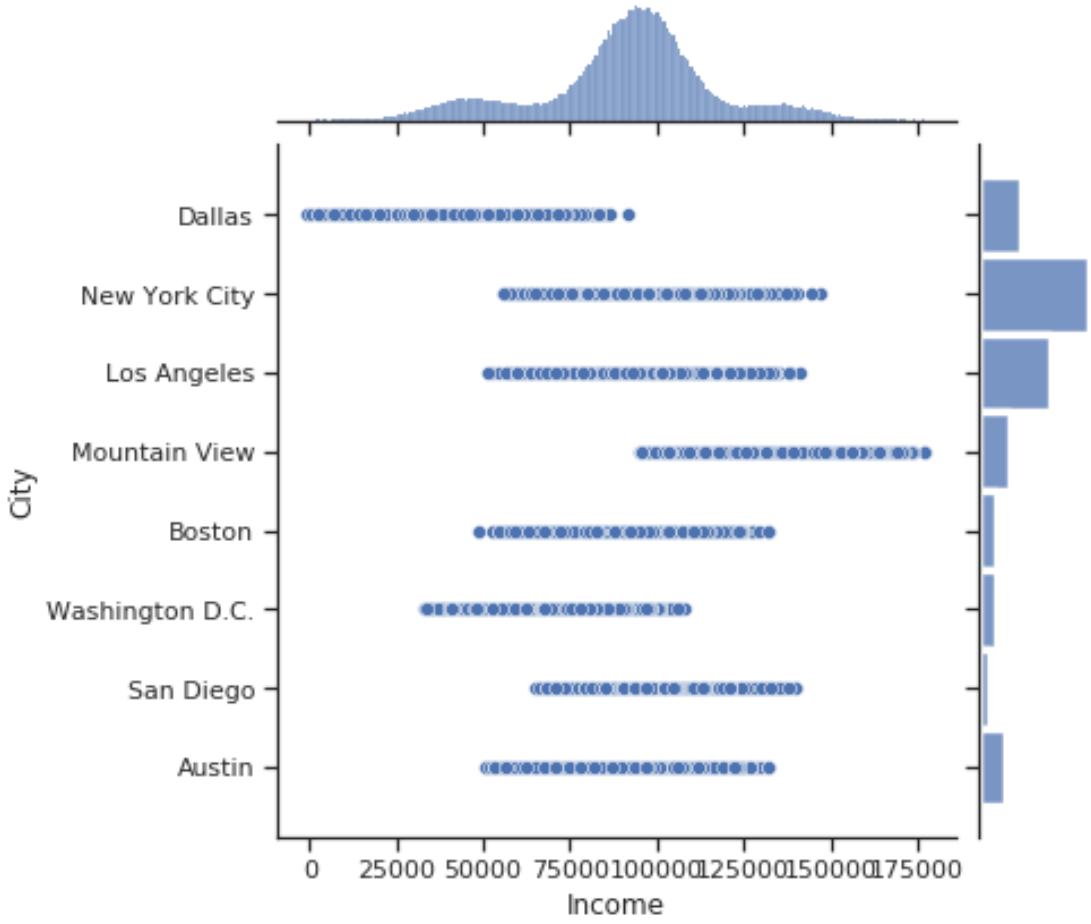
```
fig, ax = plt.subplots(figsize=(10,10))
sns.heatmap(data.corr(), annot=True, fmt='.6f')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb0ae23ddd0>
```




```
# Jointplot для Income, City
sns.jointplot(data=data, x='Income', y='City')
```

```
<seaborn.axisgrid.JointGrid at 0x7fb0ada17590>
```


Dallas
New York City
Los Angeles
Mountain View
Boston
Washington D.C.
San Diego
Austin
City
0
25000
50000
75000
100000
125000
150000
175000
Income