Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Рубежный контроль №1
по дисциплине
«Методы машинного обучения»
на тему
«Технологии разведочного анализа и обработки данных.»

Вариант 4

Выполнил:
студент группы ИУ5-61Б
Белоусов Е. А.

Москва — 2020 г.

```
In [1]: import numpy as np
        import pandas as pd
        from matplotlib import pyplot as plt
        import seaborn as sns
        %matplotlib inline

In [2]: data = pd.read_csv('../data/toy_dataset.csv')

In [3]: data.head()

Out[3]:   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

In [4]: data.dtypes

Out[4]: Number       int64
        City        object
        Gender      object
        Age          int64
        Income     float64
        Illness     object
        dtype: object

In [5]: data.shape

Out[5]: (150000, 6)

In [6]: # Проверим наличие пустых значений
        for col in data.columns:
            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


In [7]: data.corr()

Out[7]:           Number       Age    Income
        Number  1.000000 -0.003448  0.410460
        Age    -0.003448  1.000000 -0.001318
        Income  0.410460 -0.001318  1.000000

In [8]: sns.heatmap(data.corr(), annot=True, fmt='.3f')
```

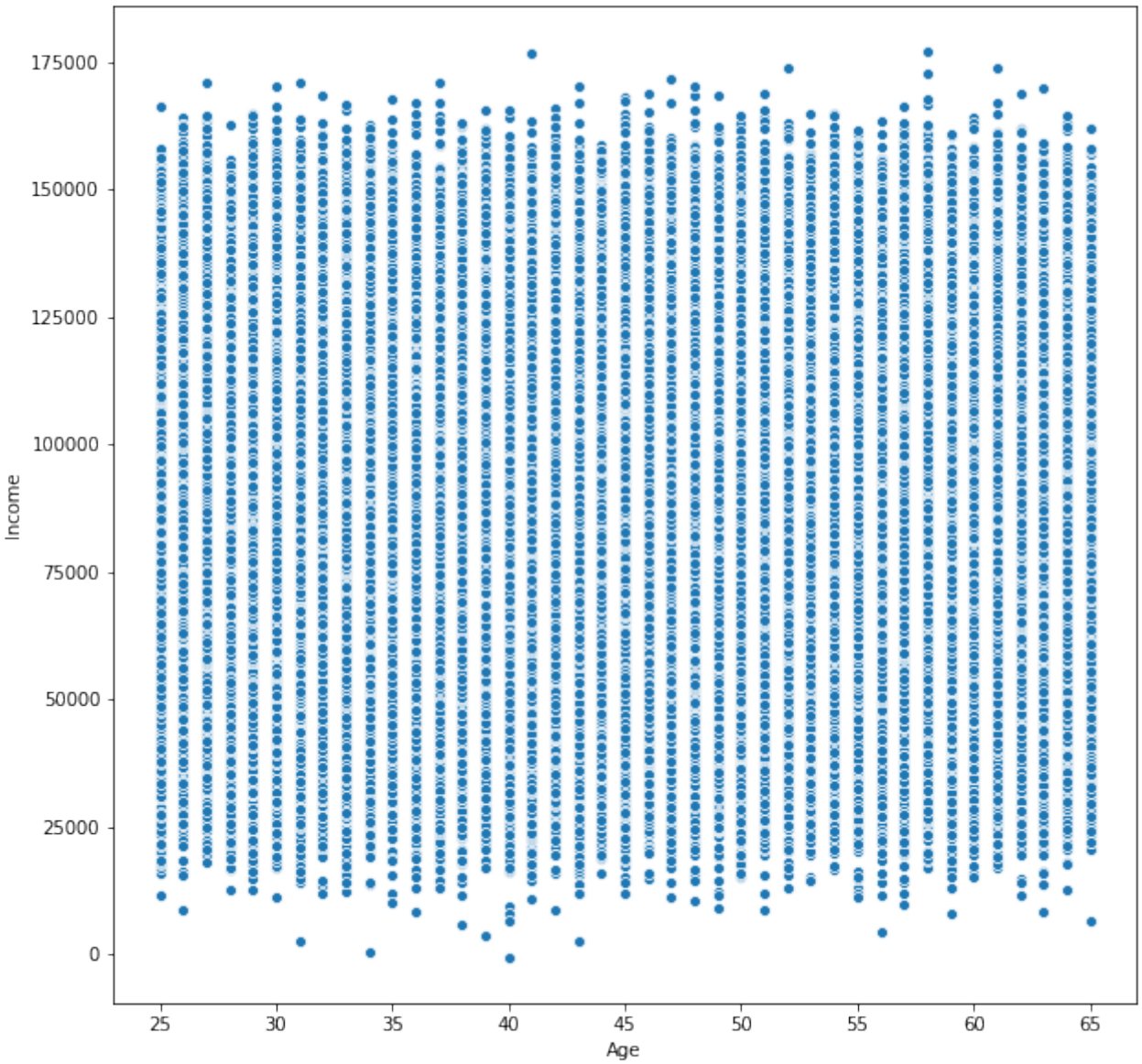
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x7f870d5f62b0>

0.1. Вывод по корреляционному анализу

Исходя из кореляционного анализа нельзя сделать вывод о зависимости дохода от возраста или номера в таблице, соответственно, эти данные ненужно использовать для построения линейных моделей. К остальным атрибутам, имеющим тип object, нельзя применить корреляционный анализ.

In [9]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='Age', y='Income', data=data)
```

Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x7f870aee20b8>



как и ожидалось, корреляции между данными не наблюдается