Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Информатика и системы управления

КАФЕДРА Системы обработки информации и управления

Отчёт по рубежному контролю №1

По дисциплине:
«Технологии машинного обучения»

Выполнила:

Студентка группы ИУ5 _____ **Быкова Д.И.**

(Подпись, дата) (Фамилия И.О.)

Проверил:

_____ **Гапанюк Ю. Е.**

(Подпись, дата) (Фамилия И.О.)

Москва, 2021

Задание

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.

Для студентов групп ИУ5-61Б, ИУ5Ц-81Б - для пары произвольных колонок данных построить график "Диаграмма рассеяния".

Набор данных:

https://www.kaggle.com/carlolepelaars/toy-dataset

РК ИУ5-61Б

Импорт библиотек

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from pandas.plotting import scatter_matrix
import warnings
warnings.filterwarnings('ignore')
sns.set(style="ticks")
%matplotlib inline
```

In [2]:
```
data = pd.read_csv('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.isnull().sum()
# проверим есть ли пропущенные значения
```

Out[5]:
```
Number     0
City       0
Gender     0
Age        0
Income     0
Illness    0
dtype: int64
```

In [6]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150000 entries, 0 to 149999
Data columns (total 6 columns):
 #   Column   Non-Null Count   Dtype
---  ------   --------------   -----
 0   Number   150000 non-null  int64
 1   City     150000 non-null  object
 2   Gender   150000 non-null  object
 3   Age      150000 non-null  int64
 4   Income   150000 non-null  float64
 5   Illness  150000 non-null  object
dtypes: float64(1), int64(2), object(3)
memory usage: 6.9+ MB
```

Корреляционный анализ

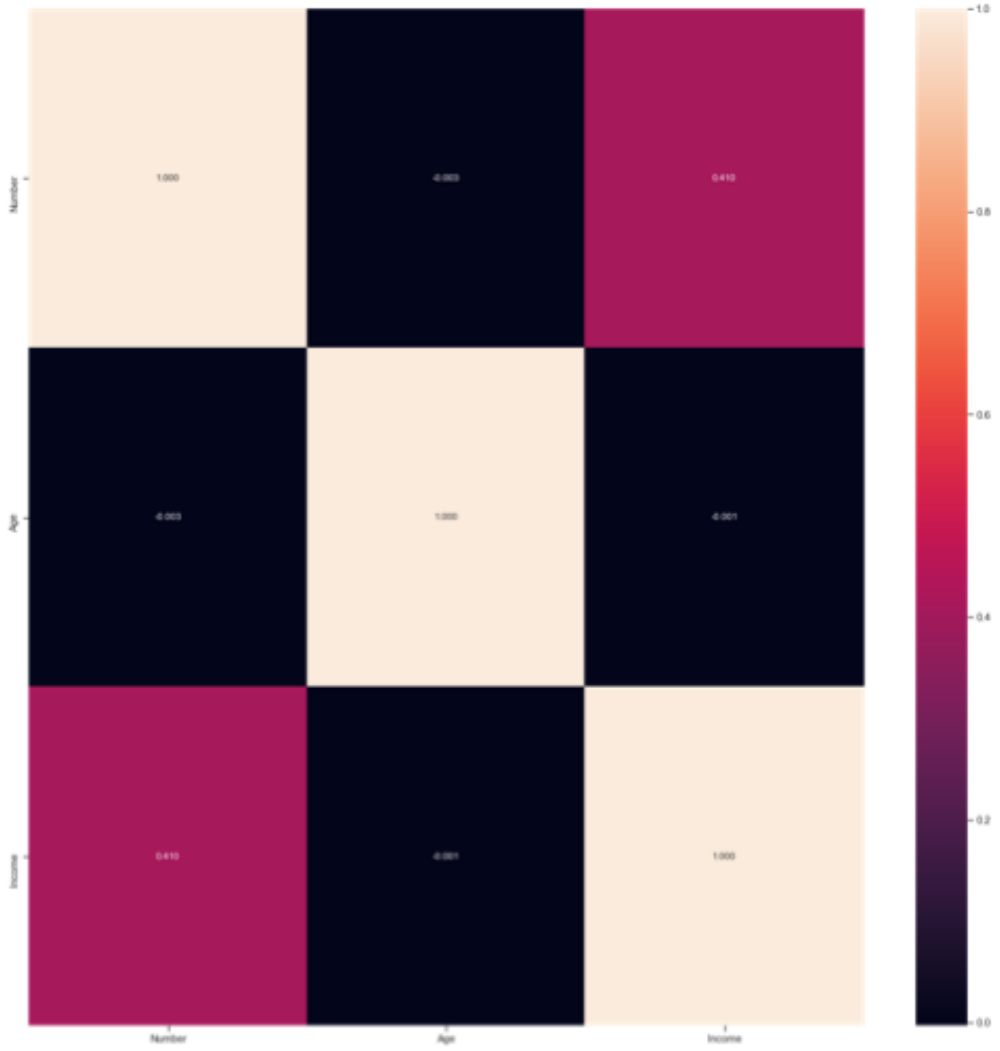
In [7]:
```
corr_matrix = data.corr()
```

In [8]:
```
corr_matrix['Income']
```

Out[8]:
```
Number    0.410460
Age      -0.001318
Income    1.000000
Name: Income, dtype: float64
```

In [9]:
```
plt.figure(figsize=(20,20))
sns.heatmap(corr_matrix, annot=True, fmt='.3f')
```

Out[9]:
```
<AxesSubplot:>
```

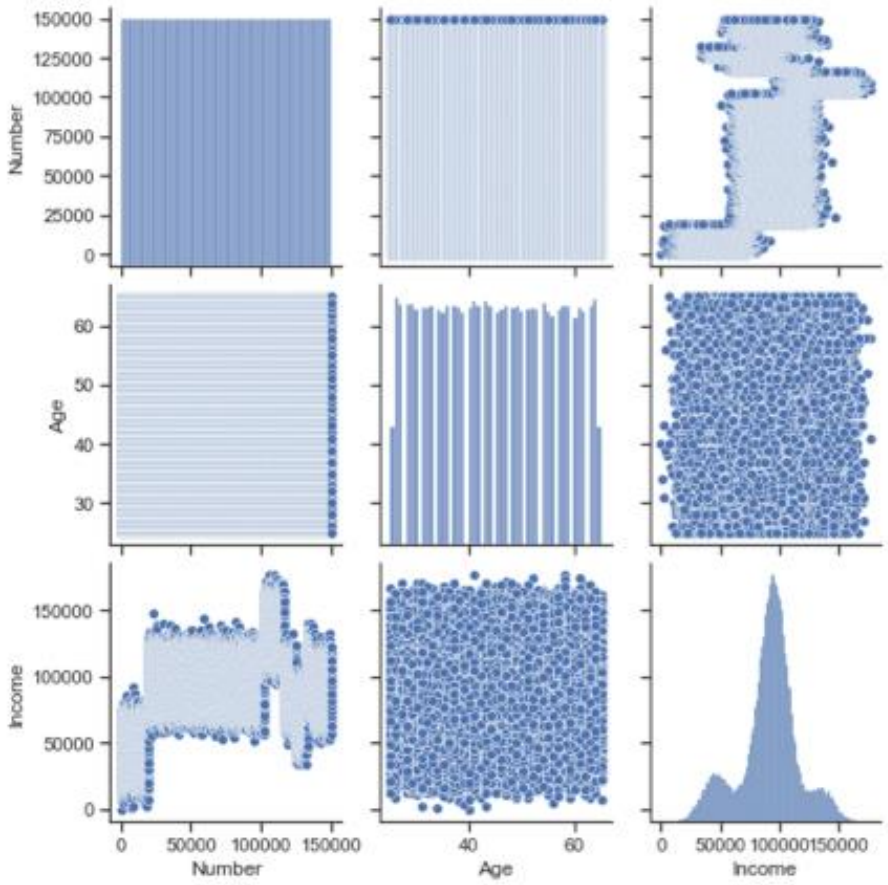


In [10]:

```
# Диаграммы рассеяние для всех признаков
plt.figure(figsize=(12,6))
sns.pairplot(data)
```

Out[10]: <seaborn.axisgrid.PairGrid at 0x7f985b7e0df0>

<Figure size 864x432 with 0 Axes>



In [11]:
```
# Увеличенные диаграммы рассеяния
sns.jointplot(x = "Age", y = "Income", kind="scatter", data = data)
```

Out[11]: <seaborn.axisgrid.JointGrid at 0x7f985bdc8f40>