РК1 Бахрамов Никита Андреевич ИУ5-63Б

Задание 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.

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

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

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

```
print('Количество пропущенных значений')
data.isnull().sum()
```

```
Количество пропущенных значений

Number     0
City       0
Gender     0
```

```
Age        0
Income     0
Illness    0
dtype: int64
```

обнаружены пропуски данных

```
data.fillna("No", inplace=True)
print('Количество пропущенных значений')
data.isnull().sum()
```

```
Количество пропущенных значений

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

Корреляционный анализ

```
data['Illness'].replace('No', 0,inplace=True)
data['Illness'].replace('Yes', 1, inplace=True)
data['Illness'].astype(int)
```

```
0         0
1         0
2         0
3         0
4         0
         ..
149995    0
149996    0
149997    0
149998    0
149999    0
Name: Illness, Length: 150000, dtype: int64
```

```
data.corr()
```

```
           Number       Age    Income   Illness
Number   1.000000 -0.003448  0.410460  0.003138
Age     -0.003448  1.000000 -0.001318  0.001811
Income   0.410460 -0.001318  1.000000  0.000298
Illness  0.003138  0.001811  0.000298  1.000000
```

```
sns.heatmap(data.corr(), annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4751164d10>
```

Дополнительное задание

для произвольной колонки данных построить график "Ящик с усами (boxplot)".

```
sns.boxplot(x=data['Age'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4750646950>
```