Московский Государственный технический университет имени Н. Э. Баумана

Рубежный контроль № 1 по курсу: «Технология машинного обучения»

Работу выполнил студент группы ИУ5-63

Федорова Антонина_____

Работу проверил:

Гапанюк Ю.Е._____

Москва 2019

Задание:

Для заданного набора данных постройте основные графики, входящие в этап разведочного анализа данных. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Какие графики Вы построили и почему? Какие выводы о наборе данных Вы можете сделать на основании построенных графиков?

Текст программы с примерами выполнения программы:

```
#Импорт библиотек
import numpy as np
import pandas as pd
import seaborn as sns
from google.colab import drive
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

```
#Монтирую гугл диск, чтобы взять оттуда датасет
drive.mount("/content/gdrive", force_remount=True)
```

```
Mounted at /content/gdrive
```

```
#Загружаю данные с гугл диска
data = pd.read_csv('/content/gdrive/My Drive/toy_dataset.csv', sep=",")
```

```
data.head(10)
```

	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
5	6	Dallas	Female	36	50786.0	No
6	7	Dallas	Female	32	33155.0	No
7	8	Dallas	Male	39	30914.0	No
8	9	Dallas	Male	51	68667.0	No
9	10	Dallas	Female	30	50082.0	No

```
data.shape
```

```
(150000, 6)
```

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

```
Всего строк: 150000
Number - 0
City - 0
Gender - 0
Age - 0
Income - 0
Illness - 0
```

```
[ ] #Различные метрики по моим данным
    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

```
[ ] #Типы данных значений датасета
    data.dtypes
```

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

```
[ ] print(data['Illness'].unique().size)
    data['Illness'].unique()
```

```
2
array(['No', 'Yes'], dtype=object)
```

```
[ ] from sklearn.preprocessing import  LabelEncoder
```

```
    le = LabelEncoder()
    data['Illness'] = le.fit_transform(data[['Illness']])
```

```
[ ] data.head(10)
```

	Number	City	Gender	Age	Income	Illness
0	1	Dallas	Male	41	40367.0	0
1	2	Dallas	Male	54	45084.0	0
2	3	Dallas	Male	42	52483.0	0
3	4	Dallas	Male	40	40941.0	0
4	5	Dallas	Male	46	50289.0	0
5	6	Dallas	Female	36	50786.0	0
6	7	Dallas	Female	32	33155.0	0
7	8	Dallas	Male	39	30914.0	0
8	9	Dallas	Male	51	68667.0	0
9	10	Dallas	Female	30	50082.0	0

Add code

```
[ ] data.corr()['Illness'].abs().sort_values(ascending=False)
```

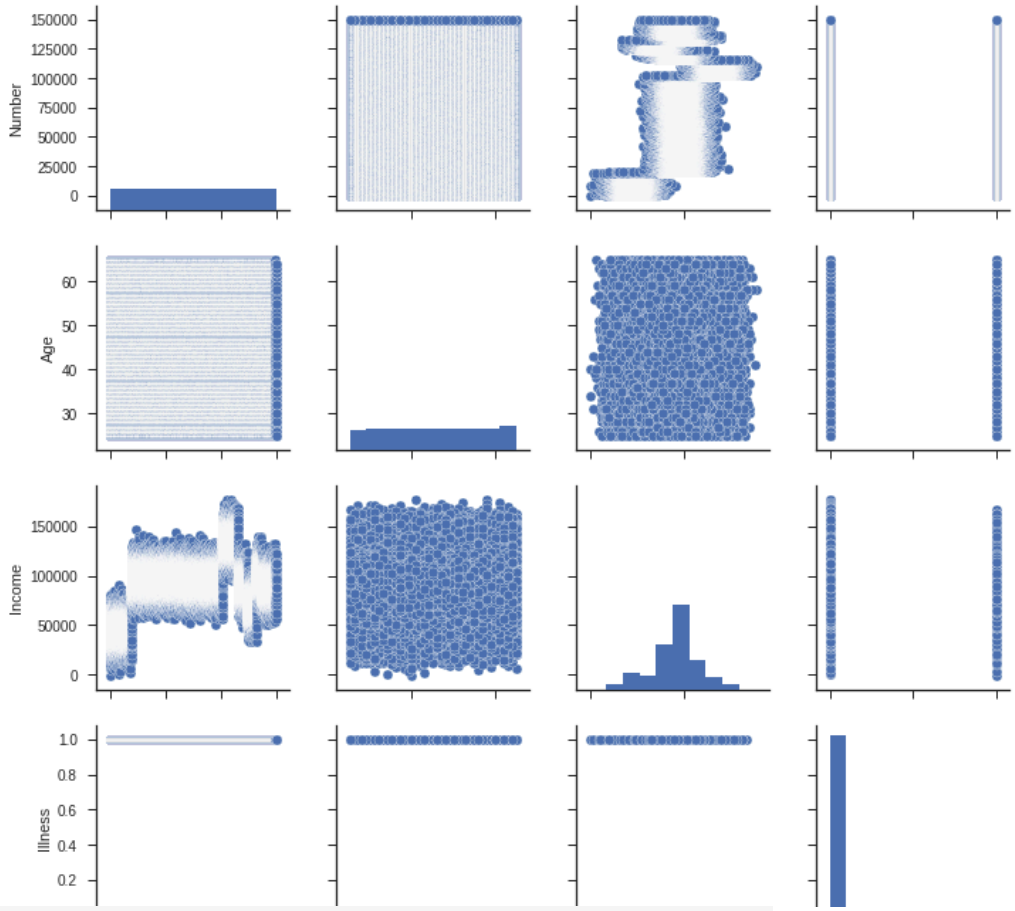
```
Illness    1.000000
Number     0.003138
Age        0.001811
Income     0.000298
Name: Illness, dtype: float64
```

```
[ ]  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.pairplot(data)
```

```
<seaborn.axisgrid.PairGrid at 0x7fd2ca6c8be0>
```




```
sns.heatmap(data.corr(method='pearson'), annot=True, fmt='.2f')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd2c3c94ac8>
```