Семенова А.А. ИУ5Ц-84Б
25 + 3 = 28 вариант РК-1
Номер задачи – 4, номер набора данных – 4.

Для студентов группы ИУ5-64Б, ИУ5Ц-84Б - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".

Задача №4.
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?
Используемый набор данных: Toy Dataset | Kaggle

In [1]:
```
#Подключаем Dataset
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib
import matplotlib_inline
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
from IPython.display import set_matplotlib_formats
matplotlib_inline.backend_inline.set_matplotlib_formats("retina")
```

In [2]:
```
#Размер набора данных
data = pd.read_csv('toy_dataset.csv', sep=",")
```

In [3]:
```
data.shape
```

Out[3]:
```
(150000, 6)
```

In [4]:
```
#Типы колонок
data.dtypes
```

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

In [5]:
```
#Проверяем, есть ли пропущенные значения
data.isnull().sum()
#Первые 5 строк датасета
```

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

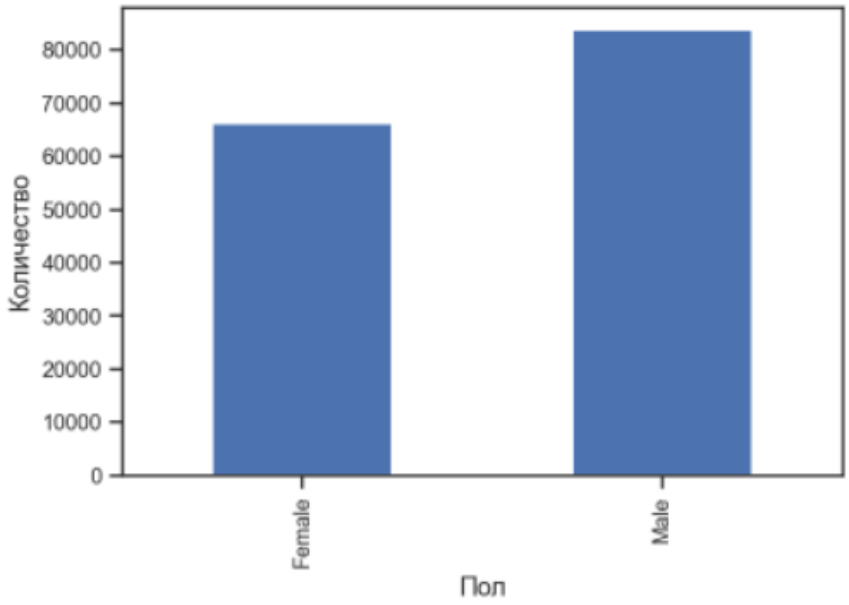
In [6]:
```
#Зададим ширину текста, чтобы он влезал на А4
data.head()
```

Out[6]:

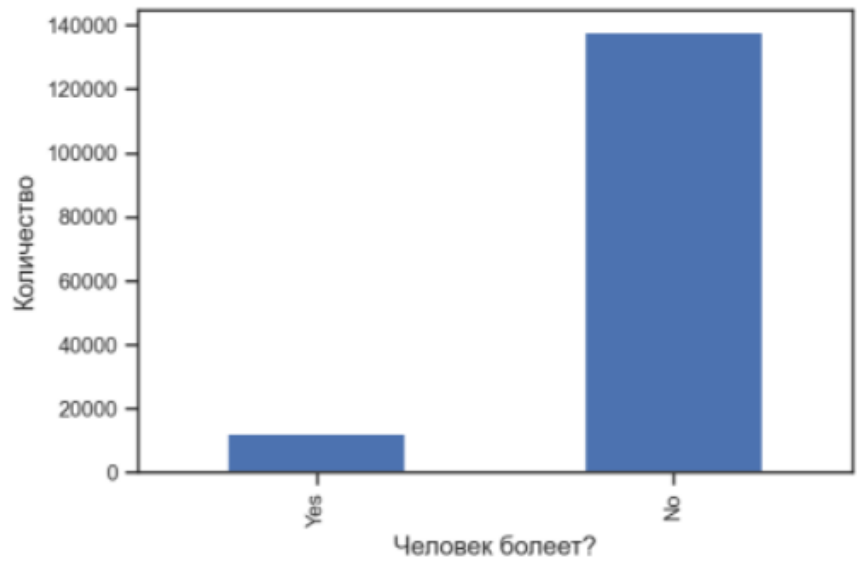
	Number	City	Gender	Age	Income	Illness
0	1	Dallas	Male	41	40367.0	No
1	2	Dallas	Male	54	45084.0	No
2	3	Dallas	Male	42	52483.0	No
3	4	Dallas	Male	40	40941.0	No
4	5	Dallas	Male	46	50289.0	No

In [7]:
```
pd.set_option("display.width", 70)
```

In [8]:
```
#Визуальное исследование датасета
#Оценим наиболее распространённый пол
count_full = data.groupby("Gender")["Gender"].count().sort_values()
count_full.plot(x="Пол", y="Количество", kind="bar", fontsize=10)
plt.xlabel("Пол")
plt.ylabel("Количество")
plt.show()
#Видно, что количество женьшин больше количества мужчин
```

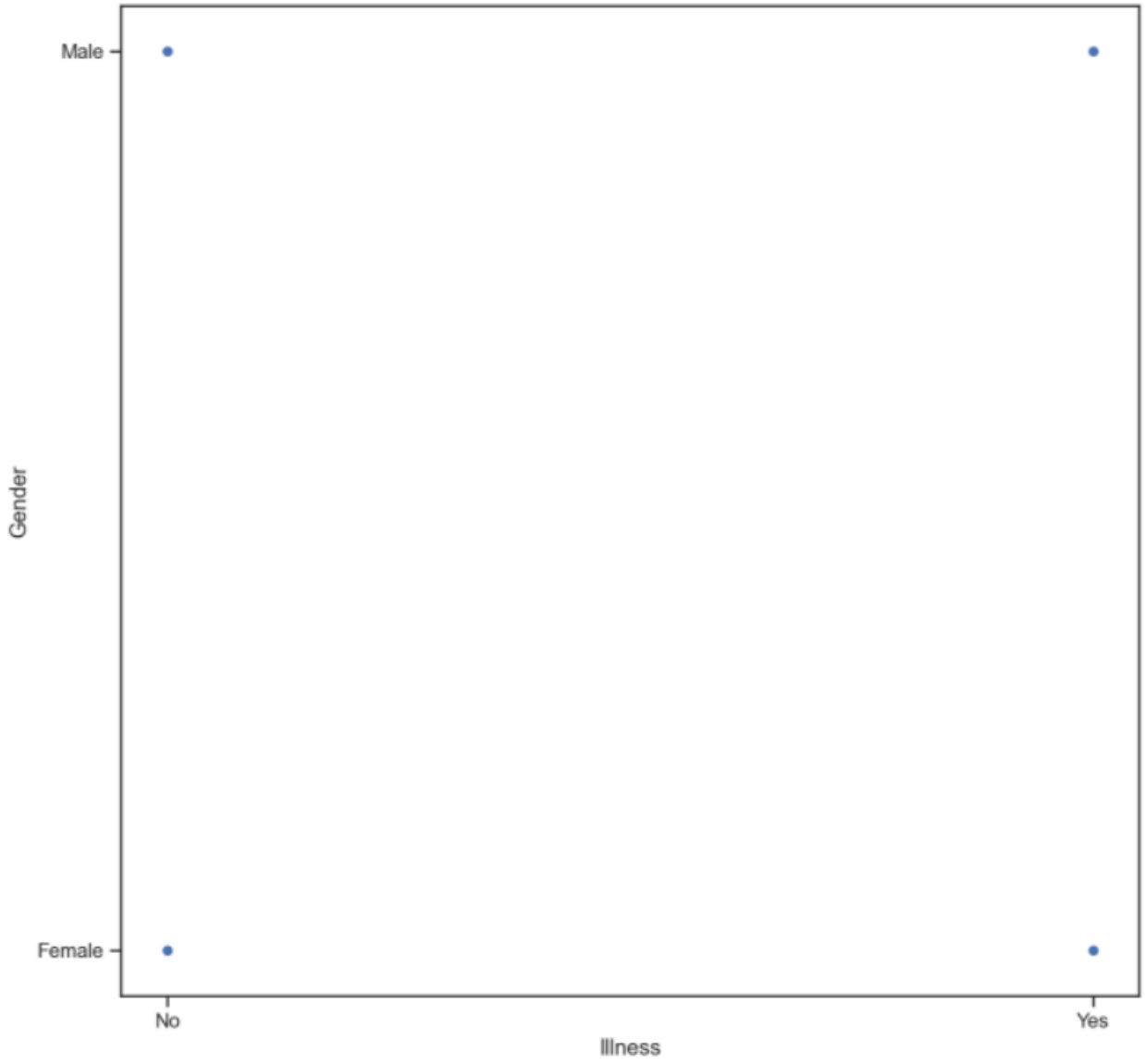


In [9]:
```
#Оценим соотношение здоровых и больных
count_full = data.groupby("Illness")["Illness"].count().sort_values()
count_full.plot(x="Человек болен?", y="Количество", kind="bar", fontsize=10)
plt.xlabel("Человек болеет?")
plt.ylabel("Количество")
plt.show()
#Видно, что из всей выборки больных меньше 20000 человек
```



In [10]:
```
#Диаграммы рассеяния
#Диаграмма рассеяния, показывающая зависимость пола от наличия заболевания
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='Illness', y='Gender', data=data)
#Из данной диаграммы (на ней всего 4 точки) следует, что есть в любом поле как здоровые, так и больные люди
```

Out[10]:
```
<AxesSubplot:xlabel='Illness', ylabel='Gender'>
```



In [11]:
```
#Скрипичная диаграмма по столбцу "age"
fig, ax = plt.subplots(figsize=(10,10))
sns.violinplot(data['Age'])
```

Out[11]: <AxesSubplot:xlabel='Age'>