РК 1

По дисциплине «Технология машинного обучения»

Вариант 6

Выполнил:

ФИО __Павлов Сергей Алексеевич_

Рецензент:

ФИО _Гапанюк Юрий Евгеньевич_

2024

Технологии разведочного анализа и обработки данных.

Описание данных. Toy dataset представляет собой вымышленные данные для исследовательского анализа данных (EDA) и тестирования простых моделей прогнозирования.
Источник: https://www.kaggle.com/datasets/carlolepelaars/toy-dataset/data
Набор данных включает в себя 150 000 строк и 6 столбцов.

Столбцы:

- Number: Простой индексный номер для каждой строки.
- City: Местоположение человека (Даллас, Нью-Йорк, Лос-Анджелес, Маунтин-Вью, Бостон, Вашингтон, Сан-Диего и Остин).
- Gender: Пол человека (Мужской или Женский).
- Age: Возраст человека (в диапазоне от 25 до 65 лет).
- Income: Годовой доход человека (в диапазоне от -674 до 177 175).
- Illness: Болен ли человек? (Да или Нет).

Ход работы:

Проверим данные на наличие пропусков и дубликатов:

```
missing_values = df.isnull().sum()
print("Пропущенные значения в каждом столбце:")
print(missing_values)

duplicate_rows = df.duplicated().sum()
print("\nКоличество дубликатов строк:", duplicate_rows)
```

```
Пропущенные значения в каждом столбце:
Number     0
City       0
Gender     0
Age        0
Income     0
Illness    0
dtype: int64

Количество дубликатов строк: 0
```

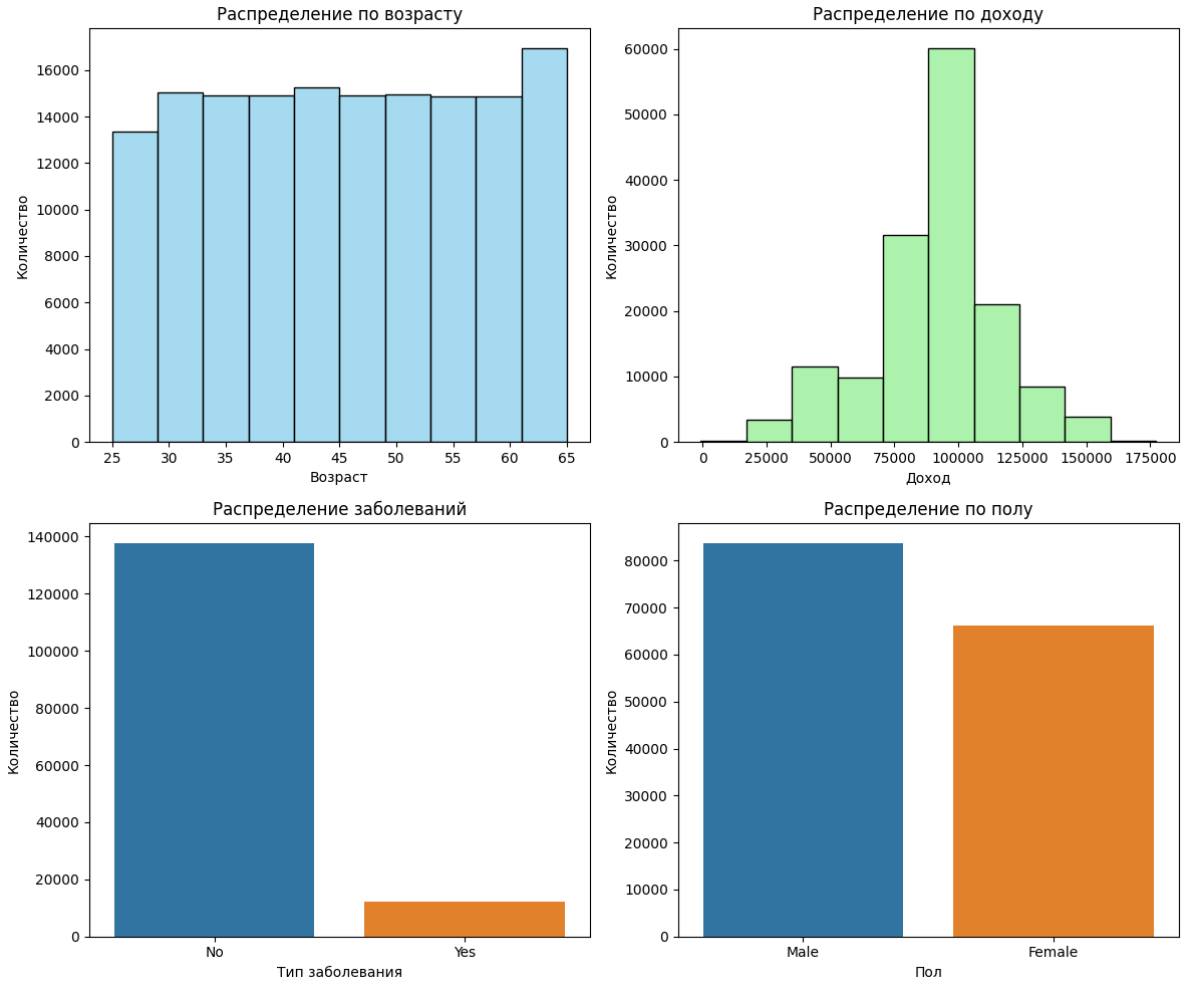
Все чисто, переходим к следующему этапу. Рассчитаем статистические характеристики переменных:

```
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

Полученные характеристики могут быть полезны для быстрого обзора основных характеристик данных до более глубокого анализа.

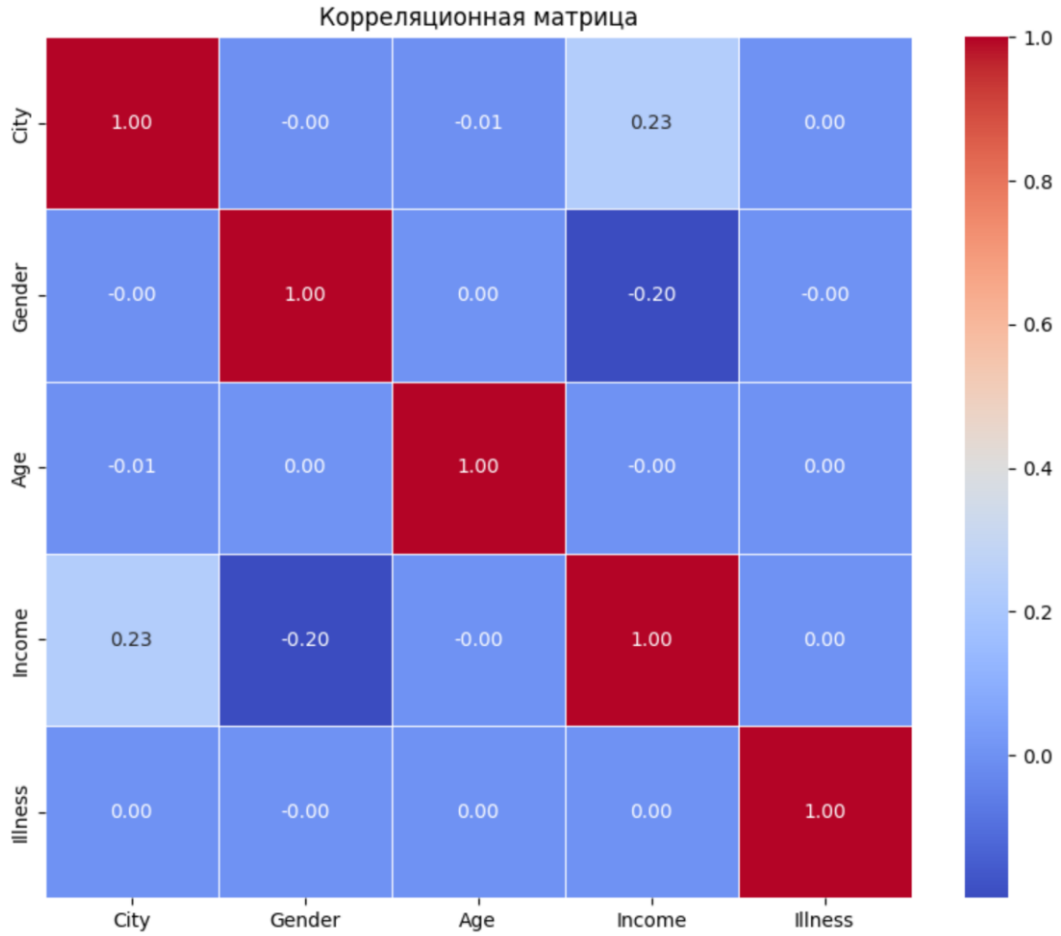
Далее приступим непосредственно к самому EDA. Посмотрим на распределение переменных. Для этого на одном полотне, визуализируем гистограммы распределений:



Видно, что основную часть в данных занимают Нью Йорк и Лос Анджелес. В плане сравнения мужчин и женщин, данные более менее

сбалансированы: 55.9 и 44.1 процентов соответственно. В разрезе заболеваний же данные не сбалансированы. Количество больных во много раз меньше здоровых. По возрасту данные распределены равномерно. Годовой доход подчиняется нормальному закону.

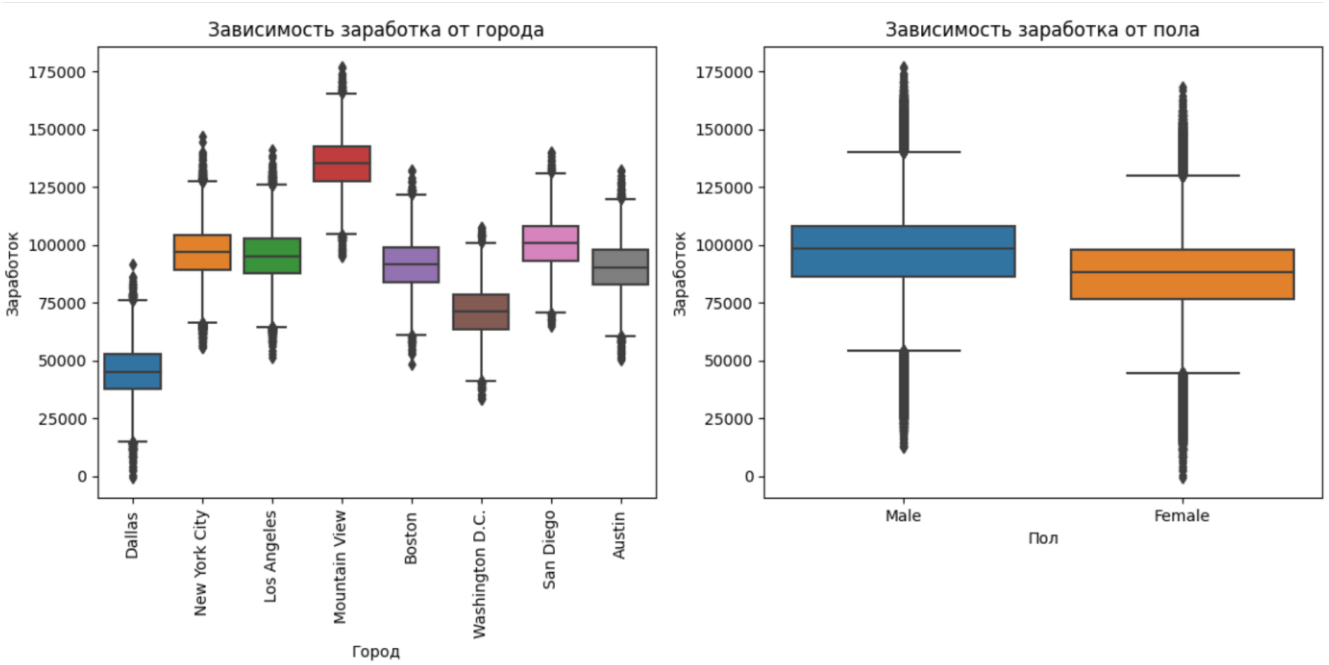
Далее для численных признаков построим корреляционную матрицу в виде тепловой карты:



По рассчитанным значениям, мы можем оценить наличие линейной взаимосвязи между переменными. Из нашей матрицы следует следующие выводы:

- Между городом и заработком существует малая линейная зависимость (коэффициент корреляции 0.23).
- Между полом и доходом также наблюдается малая корреляция (0.2).
- Между остальными парами переменных коэффициент корреляции равен нулю - линейной взаимосвязи нет.

Так как была найдена хоть и незначительная взаимосвязь между вышеприведенными признаками. Рассмотрим эти зависимости более подробно:



Тут можно сказать, например, что заработок в Mountain View больше чем в Dallas, а в остальных городах приблизительно одинаковый.

Также, из второго графика видно, что мужчины хоть и совсем незначительно, но в среднем зарабатывают больше женщин.

Вывод: таким образом, был проведен исследовательский анализ данных синтетического набора данных. Были исследованы распределения каждой переменной, оценены взаимосвязи между ними, каждая значимая взаимосвязь проанализирована в своем разрезе.