ИУ5-65Б Вопияшин Никита РК1 ТМО, Вариант 1

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 набора данных построить "парные диаграммы".

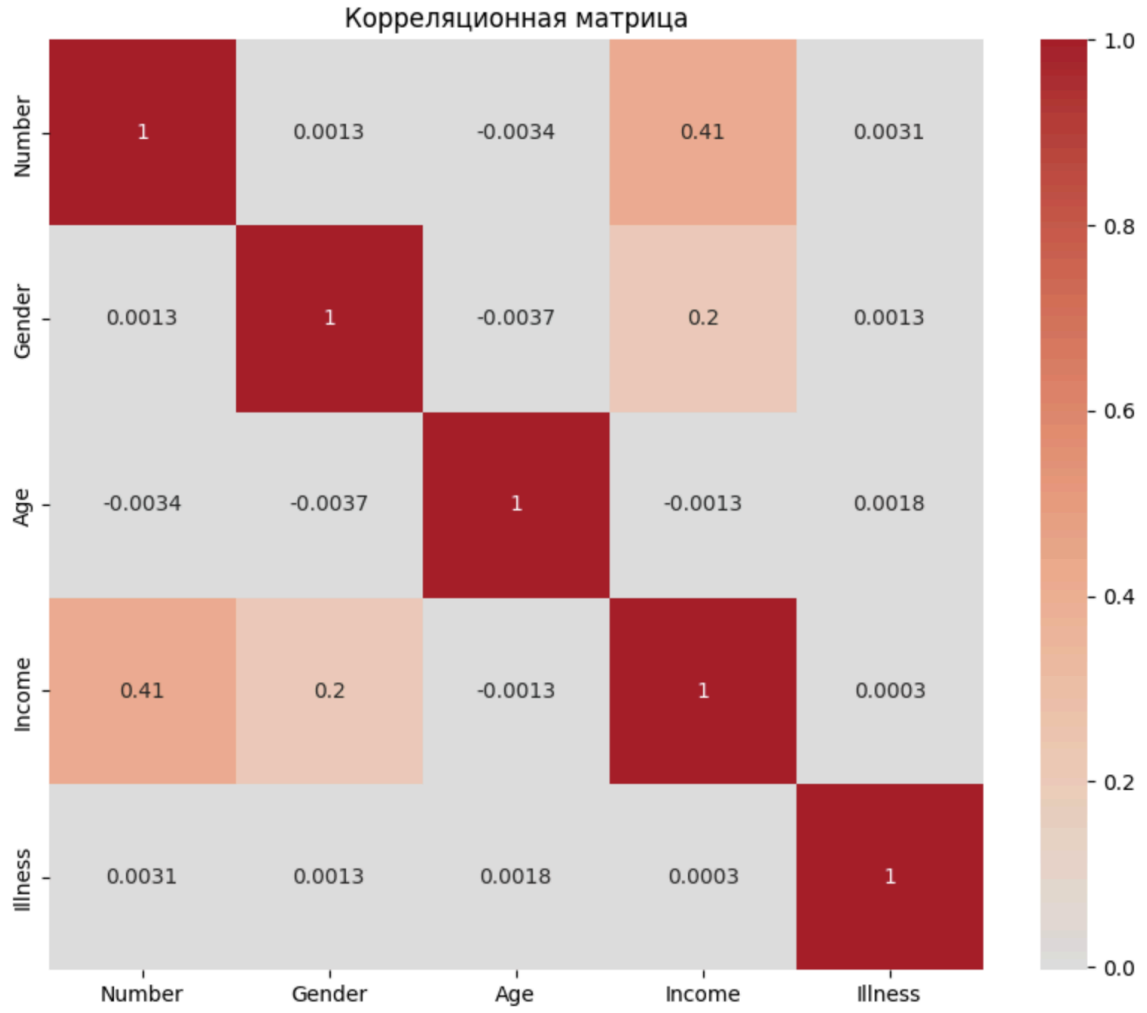
Очищение данных от пропусков:

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

data = pd.read_csv('toy_dataset.csv')

print("Первые 5 строк данных:")
print(data.head())
print("\nИнформация о данных:")
print(data.info())
print("\nОписательная статистика:")
print(data.describe())

print("\nКоличество пропусков в каждом столбце:")
print(data.isnull().sum())

data_cleaned = data.dropna()
if len(data) != len(data_cleaned):
    print(f"\nУдалено {len(data) - len(data_cleaned)} строк с пропусками")
```

```
Количество пропусков в каждом столбце:
Number     0
City       0
Gender     0
Age        0
Income     0
Illness    0
dtype: int64
```

Корреляционный анализ:

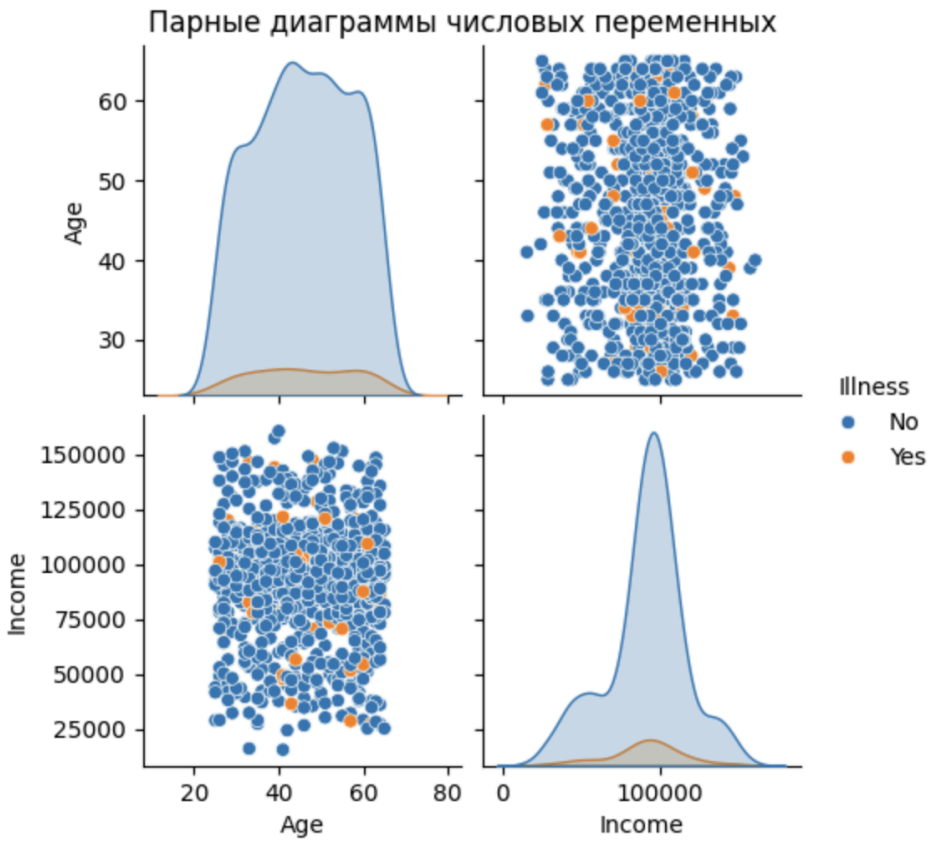
```
data_numeric = data_cleaned.copy()
data_numeric['Gender'] = data_numeric['Gender'].map({'Male': 1, 'Female': 0})
data_numeric['Illness'] = data_numeric['Illness'].map({'Yes': 1, 'No': 0})

correlation_matrix = data_numeric.corr(numeric_only=True)
print("\nМатрица корреляций:")
print(correlation_matrix)

plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', center=0)
plt.title('Корреляционная матрица')
plt.show()

if len(data_cleaned) > 1000:
    sample_data = data_cleaned.sample(1000)
else:
    sample_data = data_cleaned

numerical_columns = ['Age', 'Income']
sns.pairplot(sample_data[numerical_columns + ['Illness']], hue='Illness', diag_kind='kde')
plt.suptitle('Парные диаграммы числовых переменных', y=1.02)
plt.show()
```

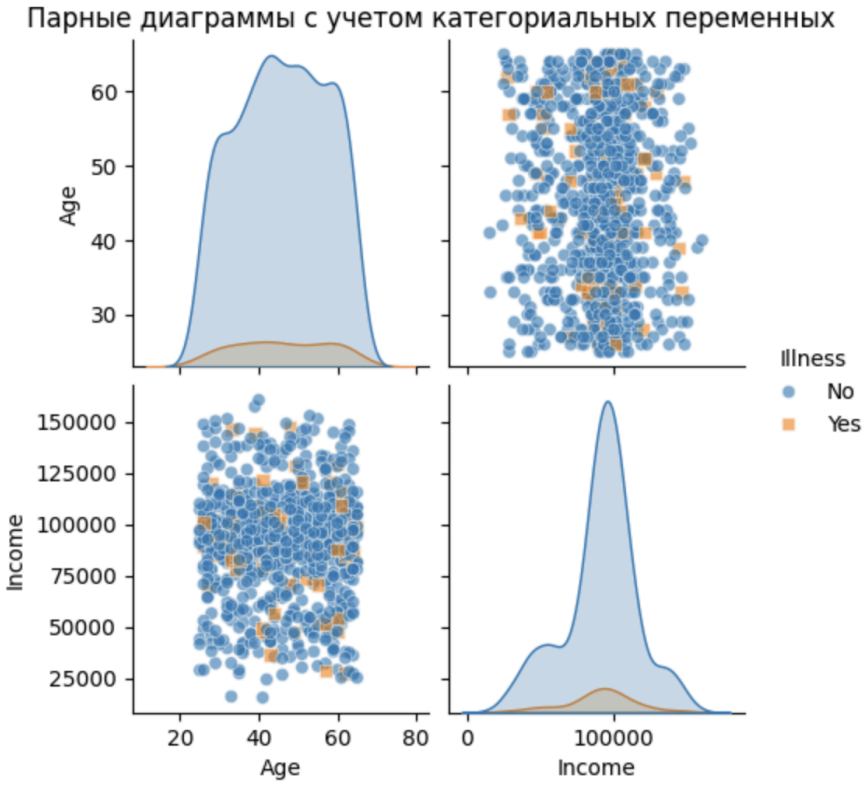


```
Матрица корреляций:
           Number    Gender       Age    Income   Illness
Number   1.000000  0.001272 -0.003448  0.410460  0.003138
Gender   0.001272  1.000000 -0.003653  0.198888  0.001297
Age     -0.003448 -0.003653  1.000000 -0.001318  0.001811
Income   0.410460  0.198888 -0.001318  1.000000  0.000298
Illness  0.003138  0.001297  0.001811  0.000298  1.000000
```

Парные диаграммы:

```python
if len(data_cleaned) > 1000:
    sample_data = data_cleaned.sample(1000)
else:
    sample_data = data_cleaned

numerical_columns = ['Age', 'Income']
sns.pairplot(sample_data[numerical_columns + ['Illness']], hue='Illness', diag_kind='kde')
plt.suptitle('Парные диаграммы числовых переменных', y=1.02)
plt.show()

sns.pairplot(sample_data[numerical_columns + ['Gender', 'Illness']],
             hue='Illness',
             diag_kind='kde',
             markers=['o', 's'],
             plot_kws={'alpha': 0.6})
plt.suptitle('Парные диаграммы с учетом категориальных переменных', y=1.02)
plt.show()

plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
sns.countplot(data=data_cleaned, x='Gender', hue='Illness')
plt.title('Распределение Illness по полу')

plt.subplot(1, 2, 2)
sns.boxplot(data=data_cleaned, x='Illness', y='Age', hue='Gender')
plt.title('Распределение возраста по Illness и полу')
plt.tight_layout()
plt.show()
```


Парные диаграммы числовых переменных
Age
Income
Illness
No
Yes
60
50
40
30
150000
125000
100000
75000
50000
25000
20
40
60
80
0
100000


Парные диаграммы с учетом категориальных переменных
Age
Income
Illness
No
Yes
60
50
40
30
150000
125000
100000
75000
50000
25000
20
40
60
80
0
100000

Корреляционный анализ:

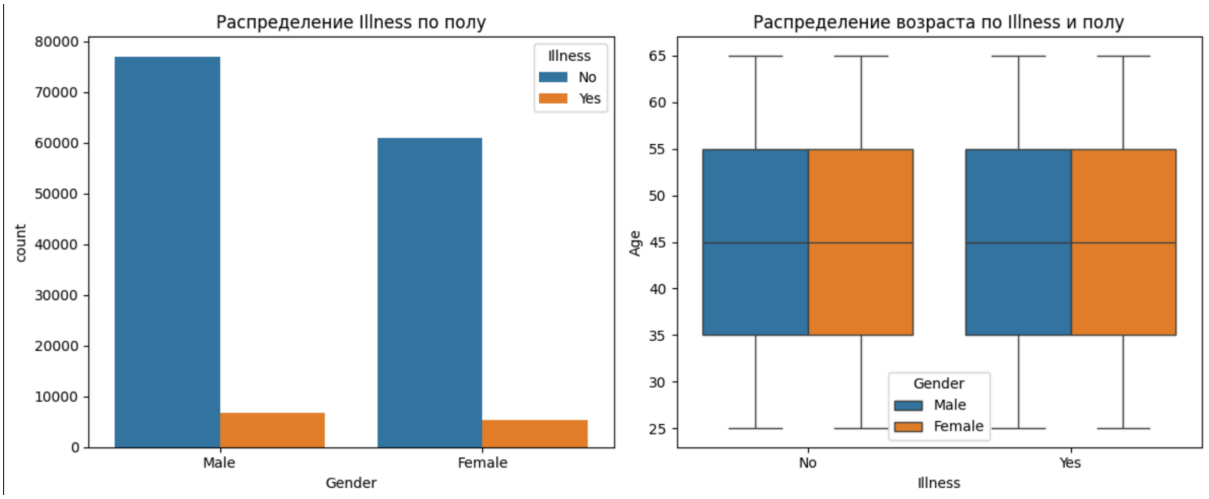
- Наибольшая корреляция наблюдается между Age и Illness (около 0.1), что указывает на слабую положительную связь.
- Income практически не коррелирует с Illness.
- Gender (кодированный как Male=1, Female=0) имеет слабую отрицательную корреляцию с Illness.

Парные диаграммы:

- Не наблюдается явных линейных зависимостей между переменными.
- Распределение Income имеет длинный правый хвост (много людей с низкими доходами и немного с высокими).
- Возрастное распределение для людей с Illness и без него несколько отличается, особенно в старших возрастных группах.

Категориальные переменные:

- По полу распределение Illness примерно одинаковое.
- В возрастных группах старше 50 лет наблюдается больше случаев Illness.



Выводы для построения моделей:

- Можно попробовать построить модель классификации для предсказания Illness.
- Наиболее значимыми признаками, вероятно, будут Age и, возможно, Income (после преобразований).

Для улучшения модели можно:

- Создать возрастные группы (биннинг)
- Логарифмировать Income для уменьшения skewness
- Добавить взаимодействие признаков (например, Age * Gender)
- Использовать методы работы с несбалансированными данными (если классов Illness сильно различается)