ЛАБОРАТОРНАЯ РАБОТА 7. РАЗРАБОТКА ЕДИНОГО ШАБЛОНА ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ ДАННЫХ

**Исследование набора данных**

**Цели и задачи**

Цель лабораторной работы: изучение теоретических принципов и инструментальных редств для построения пайплайна для предварительной обработки данных.

Основные задачи:

* предварительная обработка данных;
* изучение библиотек для предварительной обработки данных;
* масштабирование признаков;
* представление категориальных данных;
* построение пайплайна для предварительной обработки данных.

**Индивидуальное задание**

**Описание набора данных**

Данный набор данных является коллекцией пользовательских данных из социальных сетей. Он содержит информацию о отдельных пользователях, их паттернах использования и характеристиках.

Цель набора данных:

Целью этого набора данных является анализ поведения пользователей в социальных сетях, понимание их поведения, использования и выявление трендов. Этот набор данных может быть использован для построения моделей, предсказывающих вовлеченность пользователей, идентификации пользователей в сети и оптимизации маркетинга в социальных сетях.

Описание каждого признака и его тип:

* UserId: Уникальный идентификатор для каждого пользователя в наборе данных (целое число)
* UsageDuration: Общее время, проведенное пользователем в социальных сетях, в часах (число)
* Age: Возраст пользователя в годах (число)
* Country: Страна проживания пользователя (категориальная переменная)
* TotalLikes: Общее количество лайков, поставленных пользователем за день (число)

**Индивидуальное задание**

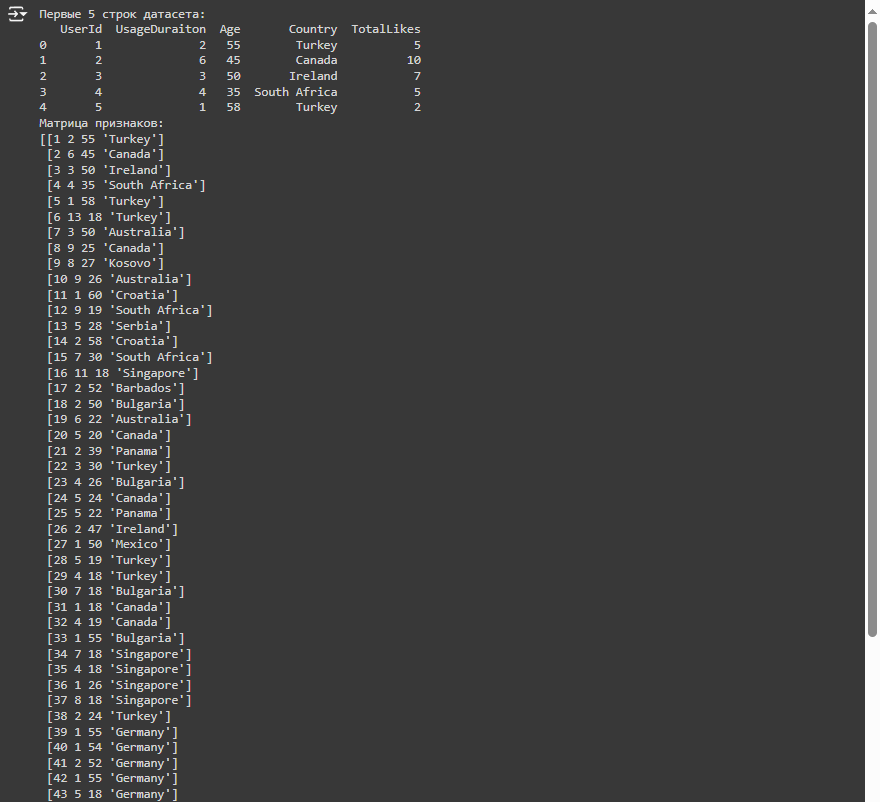
Реализуйте первичную обработку данных загруженного набора. Выполните полный спектр операций для загруженного набора данны: загрузка, визуализация, обработка пропущенных значений, обработка категориальных данных и разделение выборки на тестовую и тренировочную.

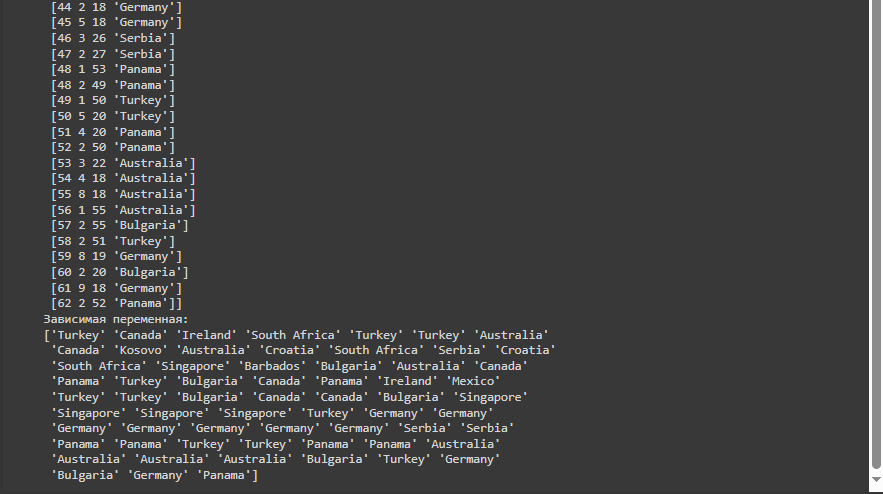
1. Импорт необходимых библиотек.



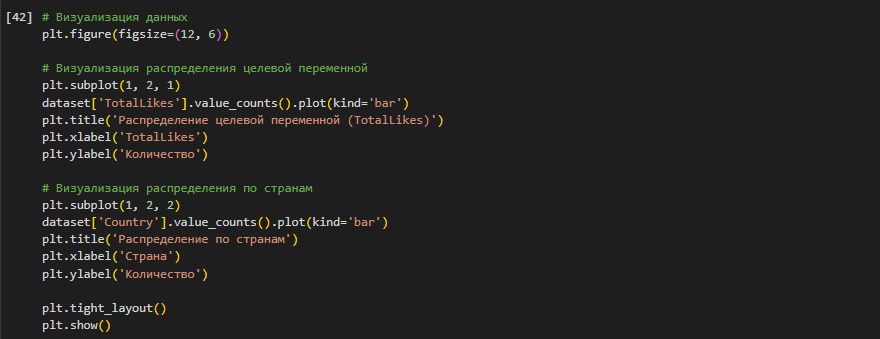
1. Загрузка датасета. Создание матрицы признаков и зависимой переменной.



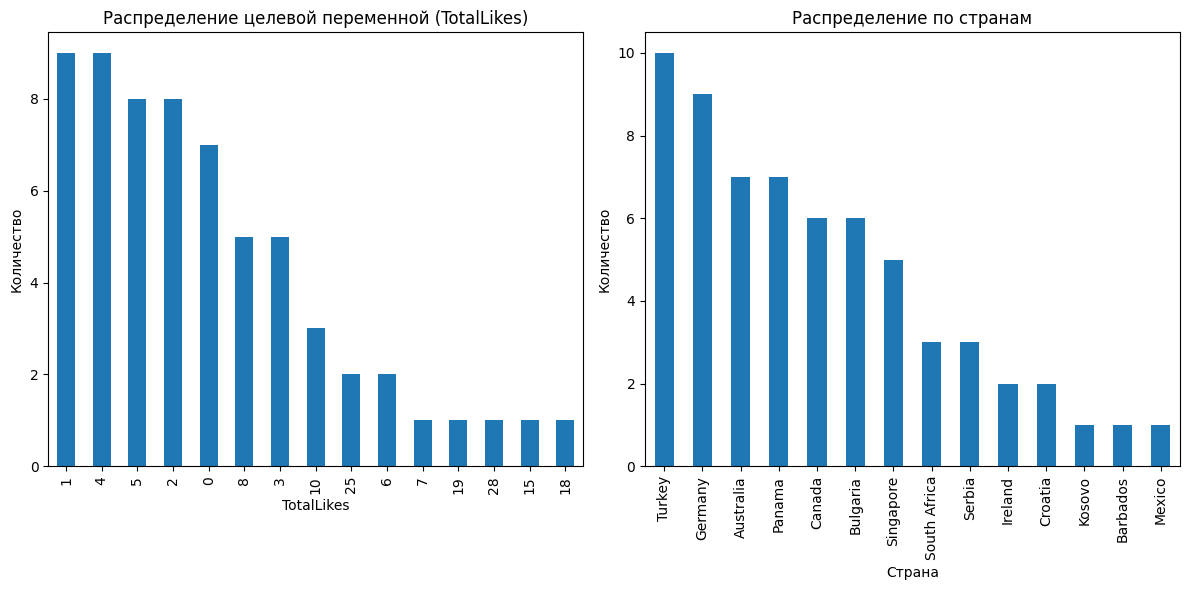




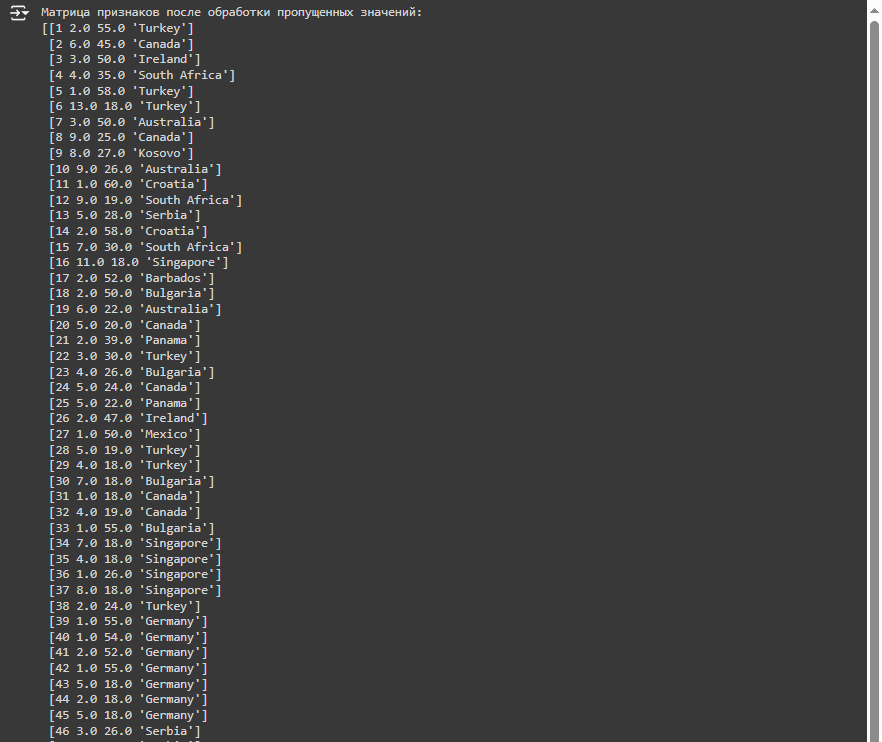
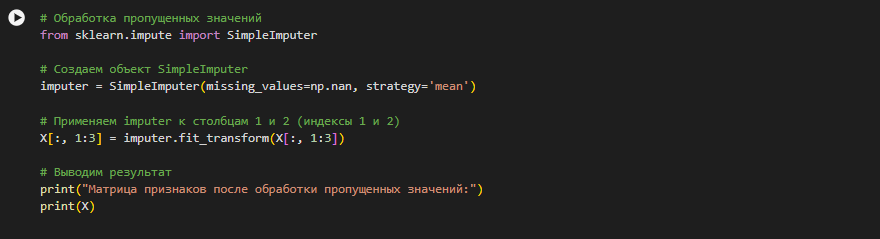
1. Визуализация данных. Распределение целевой переменной. Распределение по странам



Графики:

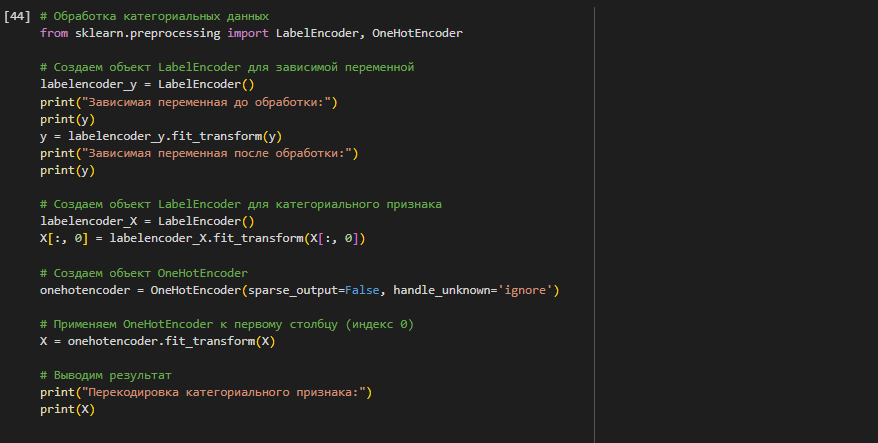


1. Обработка пропущенных значений.



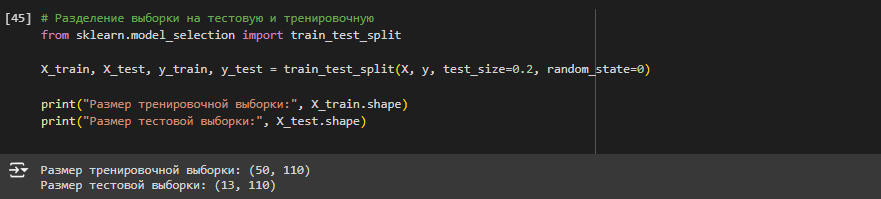


1. Обработка категориальных значений





1. Разделение выборки на тестовую и тренировочную



**Контрольные вопросы**

1. Какая библиотека python предназначена для управления наборами данных: numpy, pandas, sklearn, opencv, matplotlib?

Ответ:

Для управления наборами данных используется библиотека pandas. Она предоставляет мощные инструменты для обработки и анализа данных, такие как DataFrame и Series.

1. Какая стратегия является нежелательной при обработке пропусков в данных?

а) замена пропущенных значений в столбце медианным значением по данному столбцу;

б) удаление строк, содержащих пропуски в данных;

в) замена пропущенных значений в столбце средним арифметическим значением по данному столбцу;

г) замена пропущенных значений в столбце наиболее часто встречающимся значением по данному столбцу;

Удаление строк, содержащих пропуски в данных, является нежелательной стратегией, так как это может привести к потере значительного количества информации, особенно если пропуски встречаются часто.

1. Обоснуйте ответ на следующую проблему предварительной обработки данных: имеется независимая категориальная переменная y, которая представляет собой категориальный признак, опеределнный на домене {C#, Java, Python, R}. Нужно ли применять к данному целевому признаку OneHotEncoder?

Ответ:

К категориальной переменной, такой как язык программирования (C#, Java, Python, R), следует применять OneHotEncoder. Это необходимо, чтобы преобразовать категориальные значения в числовые, которые можно использовать в модели, и избежать создания искусственной иерархии между категориями.

1. Поясните принцип разбиения набора данных на обучающую и тестовую выборку. Какое соотношение «тестовая:обучающая» наиболее оптимально: 20:80, 50:50, 25:75, 5:95, 40:30?

Ответ:

Принцип разбиения набора данных на обучающую и тестовую выборку заключается в том, чтобы модель обучалась на одной части данных и проверялась на другой, чтобы оценить ее обобщающую способность. Наиболее оптимальное соотношение "тестовая:обучающая" обычно составляет 20:80, так как это балансирует между количеством данных для обучения и количеством данных для проверки.

1. Какой код лучше использовать при загрузке данных из csv-файла?

а) dataset = read\_csv(“data.csv”)

б) dataset = import(“data.csv”)

в) dataset = read.csv(“data.csv”)

г) dataset = import.csv(“data.csv”) д) dataset = read\_xls(“data.csv”)

Этот метод является стандартным и наиболее распространенным способом загрузки данных в pandas.