|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

**Факультет «Информатика и системы управления»**

**Кафедра «Системы обработки информации и управления»**

Отчет по лабораторной работе №2

**«Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных»**

по дисциплине «Технологии машинного обучения»

Выполнил:

студент группы ИУ5Ц-84Б

Перевощиков Н.Д.

подпись, дата

22.02.2024

Проверил:

Гапанюк Ю.Е.

подпись, дата

2024 г.

**СОДЕРЖАНИЕ ОТЧЕТА**

[1. Цель лабораторной работы 3](#_Toc160112766)

[2. Задание 3](#_Toc160112767)

[3. Основные характеристики датасета 3](#_Toc160112768)

[4. Изучение данных 4](#_Toc160112770)

[5. Описательная статистика 5](#_Toc160112771)

[6. Предобработка данных 5](#_Toc160112772)

[7. Итог 8](#_Toc160112773)

[8. Вывод 8](#_Toc160112774)

## Цель лабораторной работы

Изучение способов предварительной обработки данных для дальнейшего формирования моделей.

## Задание

1. Выбрать набор данных (датасет), содержащий категориальные признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.)
2. Для выбранного датасета (датасетов) на основе материалов лекции решить следующие задачи:
   * + обработку пропусков в данных;
     + кодирование категориальных признаков;
     + масштабирование данных.

## Основные характеристики датасета

## Название датасета: Netflix Userbase Dataset

Ссылка: <https://www.kaggle.com/datasets/arnavsmayan/netflix-userbase-dataset>

**О датасетах**

Набор данных предоставляет снимок примерной пользовательской базы Netflix, демонстрирующий различные аспекты пользовательских подписок, доходов, сведений об учетной записи и активности. Каждая строка представляет уникального пользователя, идентифицируемого по его идентификатору пользователя. Набор данных включает такую информацию, как тип подписки пользователя (базовая, Стандартная или Премиум), ежемесячный доход, получаемый от подписки, дата, когда он присоединился к Netflix (Join Date), дата его последнего платежа (Last Payment Date) и страна, в которой он находится.

Добавлены дополнительные столбцы, позволяющие получить представление о поведении и предпочтениях пользователя. В эти столбцы входят тип устройства (например, Smart TV, мобильный, настольный компьютер, планшет) и состояние учетной записи (активна учетная запись или нет). Набор данных служит синтетическим представлением и не отражает фактические данные о пользователях Netflix. Его можно использовать для анализа и моделирования, чтобы понять тенденции пользователей, предпочтения и получение дохода в рамках гипотетической базы пользователей Netflix.

**Структура данных**

User ID: Идентификатор пользователя

Subscription Type: Тип подписки

Monthly Revenue: Ежемесячный доход

Join Date: Дата присоединения

Last Payment Date: Дата последнего платежа

Country: Страна

Age: Возраст

Gender: Пол

Device: Устройство

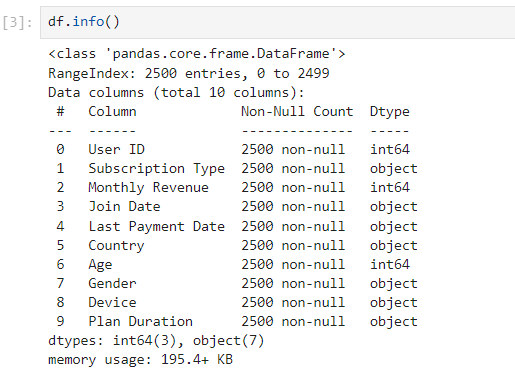
Plan Duration: Срок действия тарифного плана

## Изучение данных

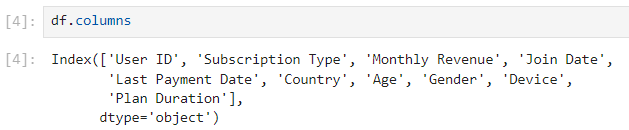
Подключаем необходимые библиотеки.



Выводим информацию.



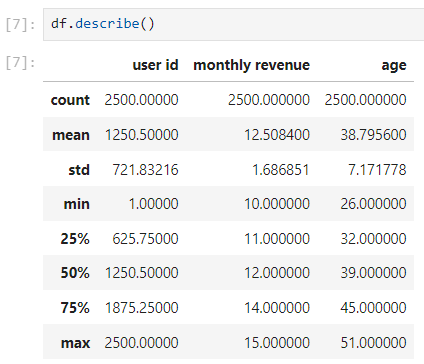
Выводим названия столбцов датасета.



Устраним и приводим их к нижнему регистру.

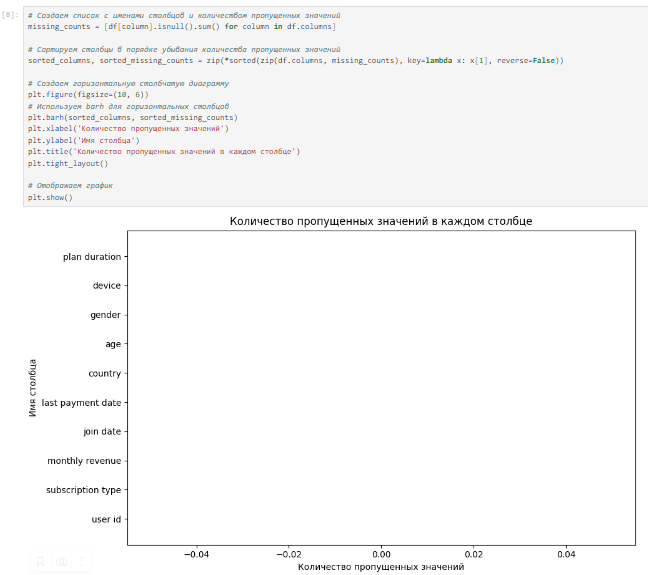


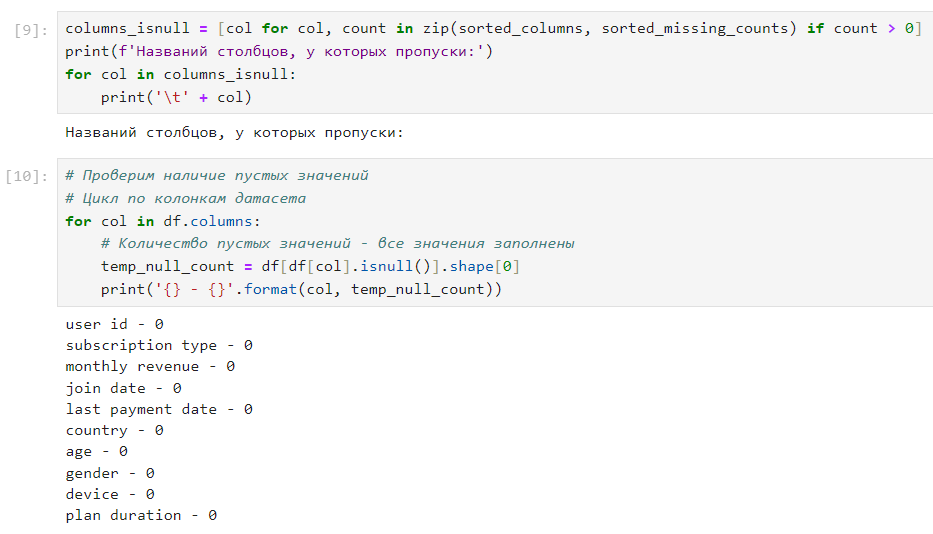
## Описательная статистика



## Предобработка данных

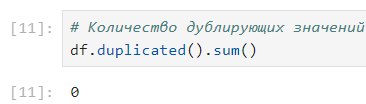
* 1. **Пропуск значения**

****

****

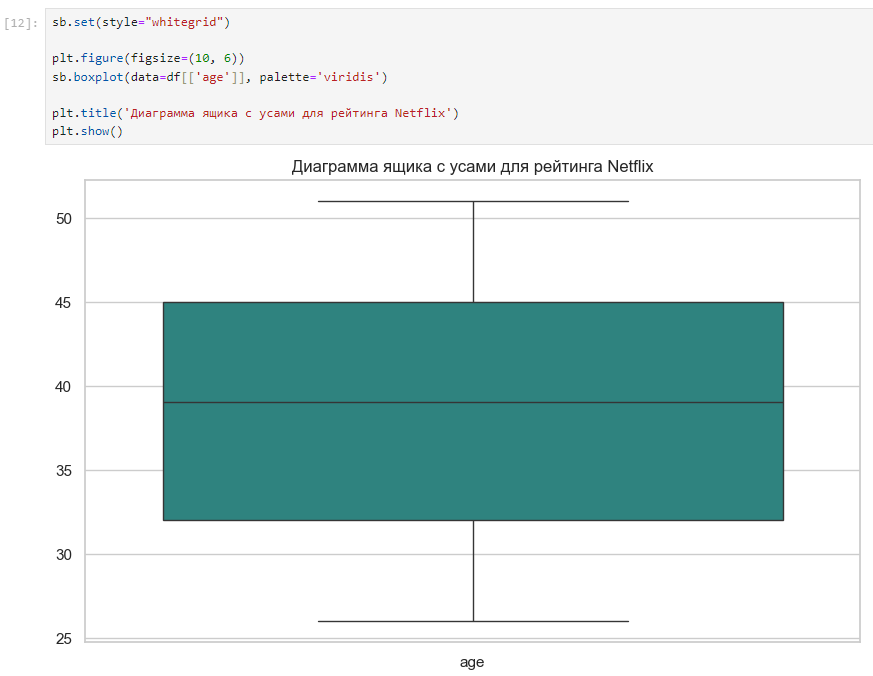
В нашем датасете нет пропусков значений, это говорится о том, что датасет сделали идеальным образом.

* 1. **Дубликаты**

****

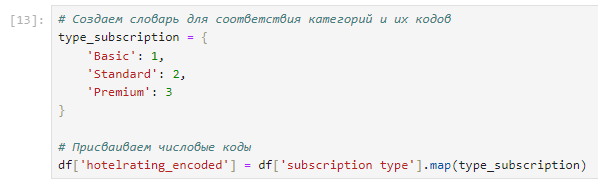
Дубликатов тоже нет.

* 1. **Выбросы (Ящик с усами)**

****

Часто просматриваемые это возраст от 32 до 45 лет

* 1. **Преобразование категорий в числа**

****

## Итог

* 1. **Предобработка данных**
* Пропусков нет
* Дубликатов нет
* Количество выбросов отсутствует
* В ходе предобработки данных было вявлено, что есть колонка, которая дает информацию типов подписок, но однако информация носит HTML формата, что на парсинг уходит много времени и необходимо сопровождать кода. Для тщательного исследования будет полезно.

## Вывод

В ходе выполнения лабораторной работы изучили способы предварительной обработки данных для дальнейшего формирования моделей.