|  |  |
| --- | --- |
|  | **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   
Папин А.В.

подпись, дата

Проверил:

к.т.н., доц., Ю.Е. Гапанюк

подпись, дата

2024 г.

**СОДЕРЖАНИЕ ОТЧЕТА**

[1. Цель лабораторной работы: 3](#_Toc5)

[2. Описание задание 3](#_Toc6)

[3. Основные характеристики датасета 3](#_Toc7)

[4. Изучение данных 5](#_Toc8)

[5. Описательная статистика 5](#_Toc9)

[6. Предобработка данных 6](#_Toc10)

[6.1. Пропущенные значения 6](#_Toc11)

[6.2. Дубликаты 7](#_Toc12)

[6.3. Выбросы - Ящик с усами 8](#_Toc13)

[6.4. Создание новых признаков - высота и ширина 8](#_Toc14)

[6.5. Преобразование категорий в числа 9](#_Toc15)

[7. Итог 9](#_Toc16)

[7.1. Предобработка данных 9](#_Toc17)

# **Цель лабораторной работы:**

Изучение способов предварительной обработки данных для дальнейшего формирования моделей.

# **Описание 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: Hotels Dataset (Датасет отелей)

Ссылка: https://www.kaggle.com/datasets/raj713335/tbo-hotels-dataset/data

**О датасетах**

Этот датасет содержит информацию о 1 000 000+ отелях из различных стран и регионов, таких как их тарифы, отзывы, удобства, местоположение и звездный рейтинг. Данные были собраны из различных источников, таких как веб-сайты отелей, онлайн-агентства по бронированию и платформы отзывов. Датасет может использоваться для различных целей, таких как:

* Исследовательский анализ данных для понимания характеристик и распределения отелей по разным рынкам и сегментам.
* Анализ настроений для извлечения идей из отзывов и рейтингов гостей отеля и выявленидаций для предоставления персонализированных предложений по бронированию отелей на основе предпочтений и поведения пользователя.
* Прогнозирование цен для оценки оптимальных тарифов для отелей на основе спроса, сезонности и конкуренции.
* Классификация для определения типа и категории отелей на основе их характеристик и атрибутов.

**Структура данных**

Датасет состоит из 16 столбцов и 1 000 000+ строк, где каждая строка представляет собой отель. Столбцы включают в себя:

countyCode: Код страны, к которой принадлежит отель.

countyName: Название страны, к которой принадлежит отель.

cityCode: Код города, где расположен отель.

cityName: Город, где расположен отель.

HotelCode: Уникальный идентификатор каждого отеля.

hotel\_name: Название отеля.

HotelRating: Звездный рейтинг отеля от 1 до 5.

Address: Адрес отеля.

Attractions: Достопримечательности рядом с отелем.

Description: Подробное описание отеля.

FaxNumber: Номер факса отеля.

HotelFacilities: Доступные в отеле удобства.

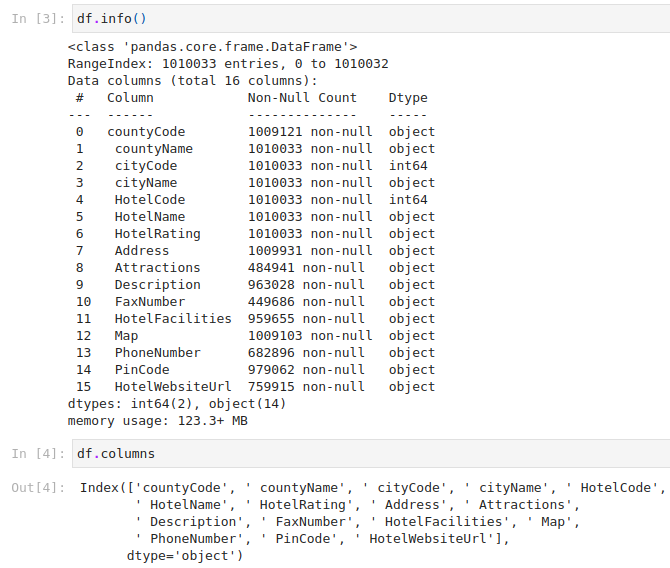
Map: Местоположение отеля в формате GPS (широта и долгота).

PhoneNumber: Телефонный номер отеля.

PinCode: Почтовый индекс адреса отеля.

HotelWebsiteUrl: Веб-ссылка для бронирования отеля.

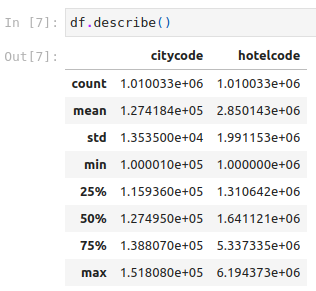
# **Изучение данных**



Как и видим, что присуствует лишний пробел перед названием, устраним и приводим их к нижнему регистру



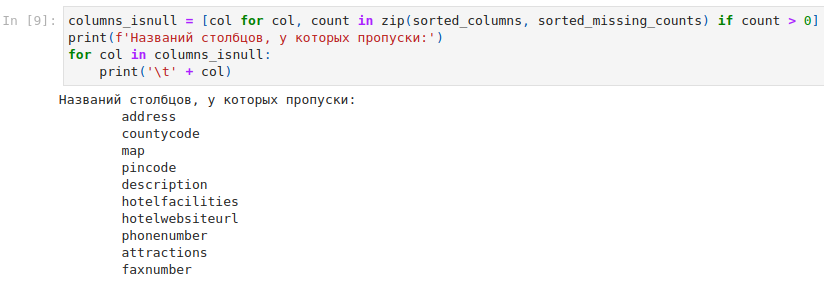
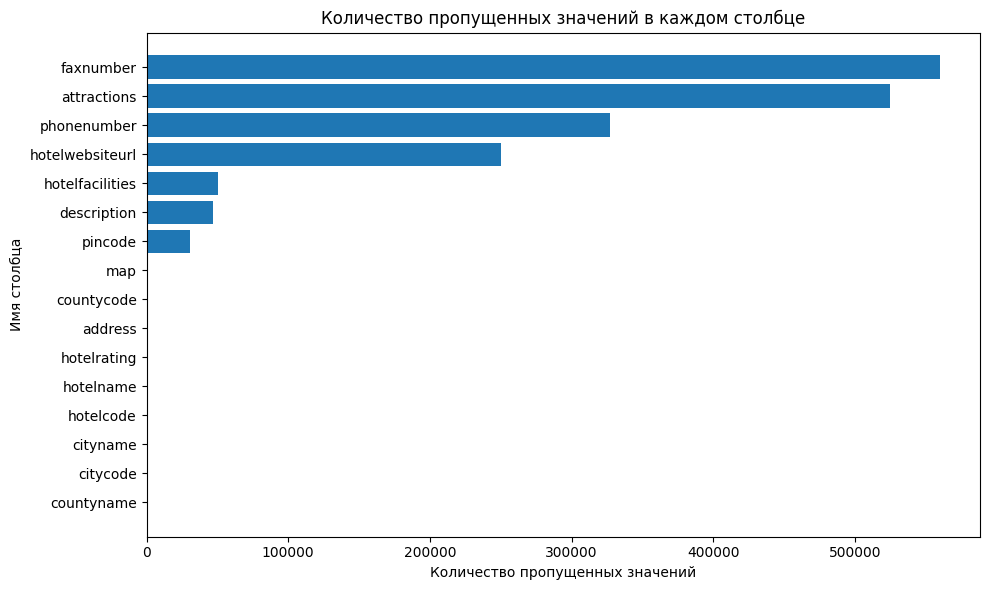
# **Описательная статистика**



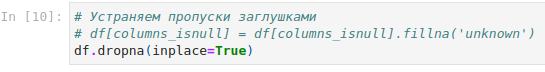
В датасете слишком много категориальных признаков, поэтому в описательной статистике мало дает информацию. Можно сделать кодирование признаков, т.е. OHE или OH, или попроще - get\_dummies()

# **Предобработка данных**

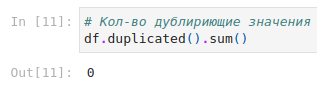
# **Пропущенные значения**



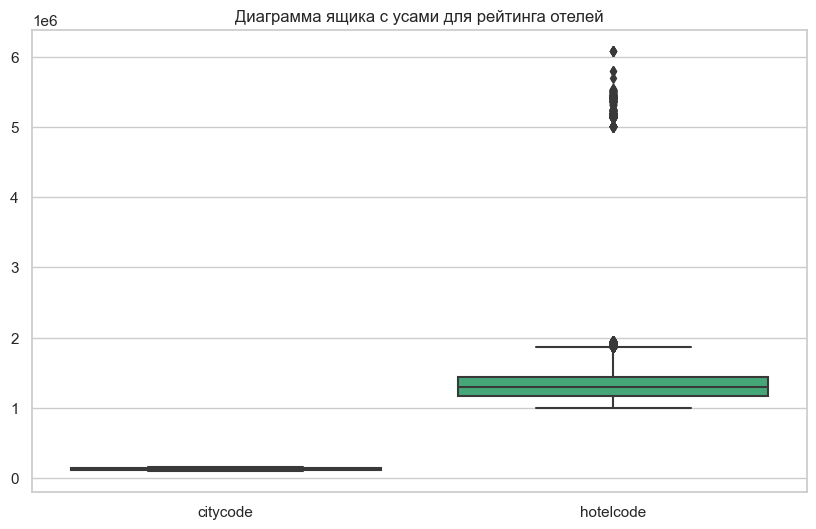
Пропущенные значения колонки в основном категориальные, поэтому не получится заполнить их медианой. Можно их удалять, но лучше всего заполнить заглушку как - unknown. Если устранить их, то более 50% данных мы потеряем, а это не очень хорошо



# **Дубликаты**

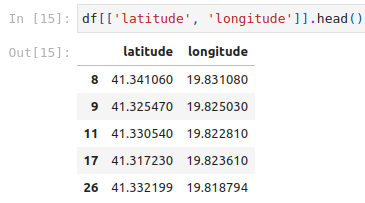
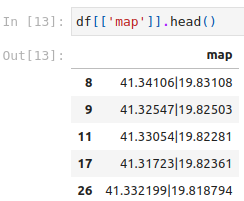


# **Выбросы - Ящик с усами**

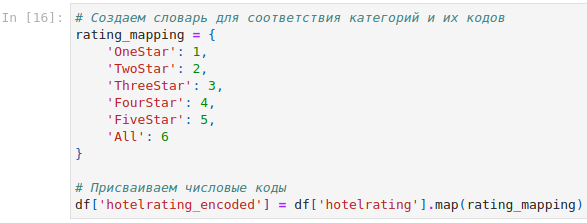


# **Создание новых признаков - высота и ширина**

Как и видим, что можно разделить на высоту и ширины с колонки map



# **Преобразование категорий в числа**



# **Итог**

# **Предобработка данных**

* Присутствовали огромные пропуски, было решено устранять их. Была мысль заполнять их заглушками, но это можно реализовать, если потребуется
* Дубликатов нет
* Присуствует значительное кол-во выбросов - код отеля
* Создали новые признаки - долгота и ширина, разбив с колонки map
* В ходе предобработки данных было вявлено, что есть колонка, которая дает информацию ближайщих достопримчательности отелей, но однако информация носит HTML формата, что на парсинг уходит много времени и необходимо сопровождать кода. Для тщательного исследования будет полезно
* Также много категориальных признаков, можно преобразовать их в численным через OHE или OH