Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика, искусственный интеллект и системы управления»

Кафедра «Системы обработки информации и управления»



**Лабораторная работа № 2**

**по дисциплине «Методы машинного обучения»**

Обработка признаков (часть 1)

ИСПОЛНИТЕЛЬ:

студентка ИУ5-23М

Морозевич М.А.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

\_\_\_ "\_\_\_\_\_\_\_\_" 2024 г.

Москва, 2024

# **Задание**

Для наборов данных необходимо решить задачи ниже:

* устранение пропусков в данных;
* кодирование категориальных признаков;
* нормализация числовых признаков.

# **Описание данных**

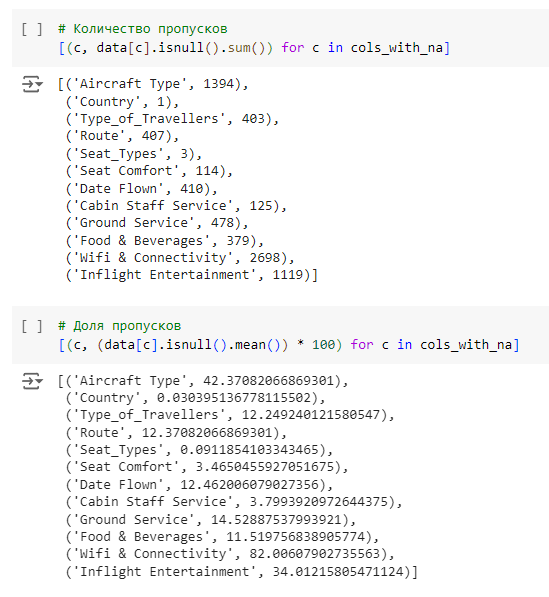
Набор данных представляет собой данные с отзывами и рейтингами авиакомпаний пассажирами о различных аспектах их впечатлений от полетов в различных авиакомпаниях.

Названия и описания колонок набора данных:

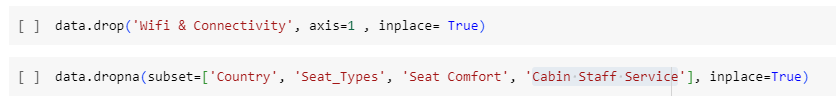
* Aircraft Type: Тип воздушного судна, используемого для полета
* Users Reviews: Тексты отзывов, предоставленных пользователями
* Country: Страна авиакомпании или отправления/пункта назначения рейса
* Type of Travellers: Классификация путешественников (например, одиноких, семейных, деловых...)
* Route: Выбранный маршрут полета
* Seat Types: Класс места (Эконом, Бизнес, Первый класс...)
* Seat Comfort: Оценка комфортности места
* Date Flown: Дата полета
* Cabin Staff Service: Оценка сервиса, предоставляемого персоналом салона
* Ground Service/Floor: Оценка наземного обслуживания, включая регистрацию на рейс и посадку
* Food & Beverages: Оценка качества питания и напитков
* Wifi & Connectivity: Рейтинг доступных возможностей Wi-Fi и подключений
* Inflight Entertainment: Рейтинг вариантов развлечений в полете
* Value For Money: Общий рейтинг соотношения цены и качества
* Recommended: Итоговая рекомендация авиакомпании пользователем

# **Устранение пропусков в данных**

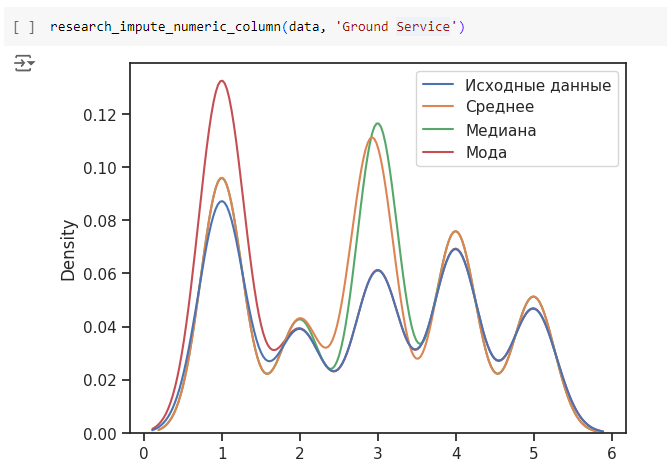
Определим колонки с пропусками и их долю.



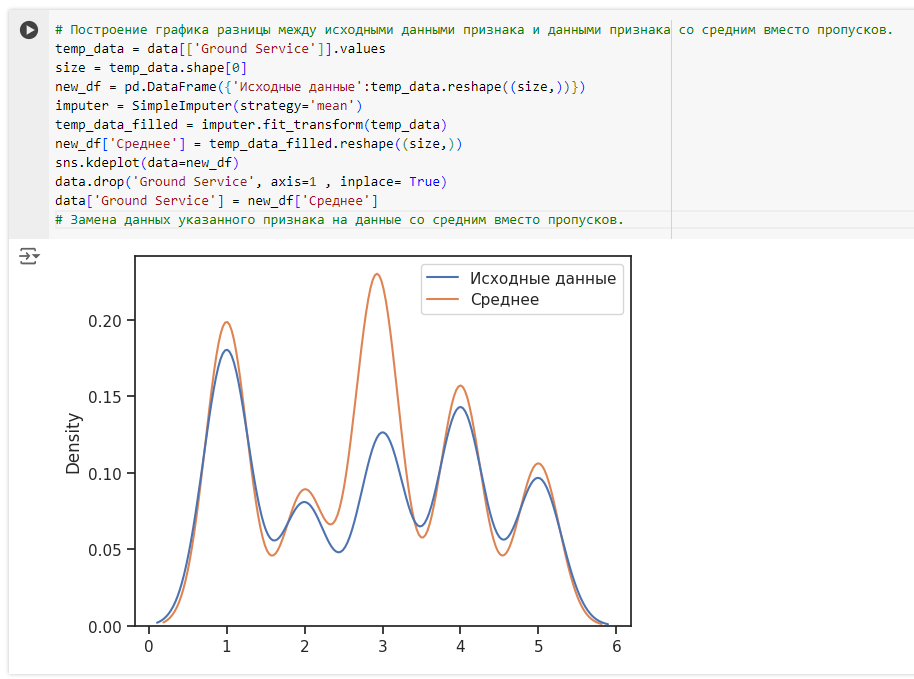
В столбце «Wifi & Connectivity» доля пропусков превышает 80%, поэтому ее было решено удалить из датасета. С другой стороны, в столбцах «Country», «Seat\_Types», «Seat Comfort» и «Cabin Staff Service» доля пропусков меньше 5% и для них было решено пропуски удалить.



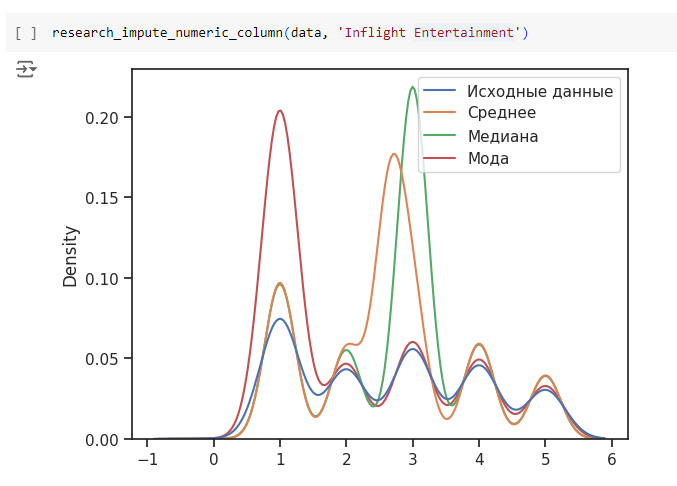
Для столбца «Ground Service» построим графики со сравнением замены пропусками разными способами.

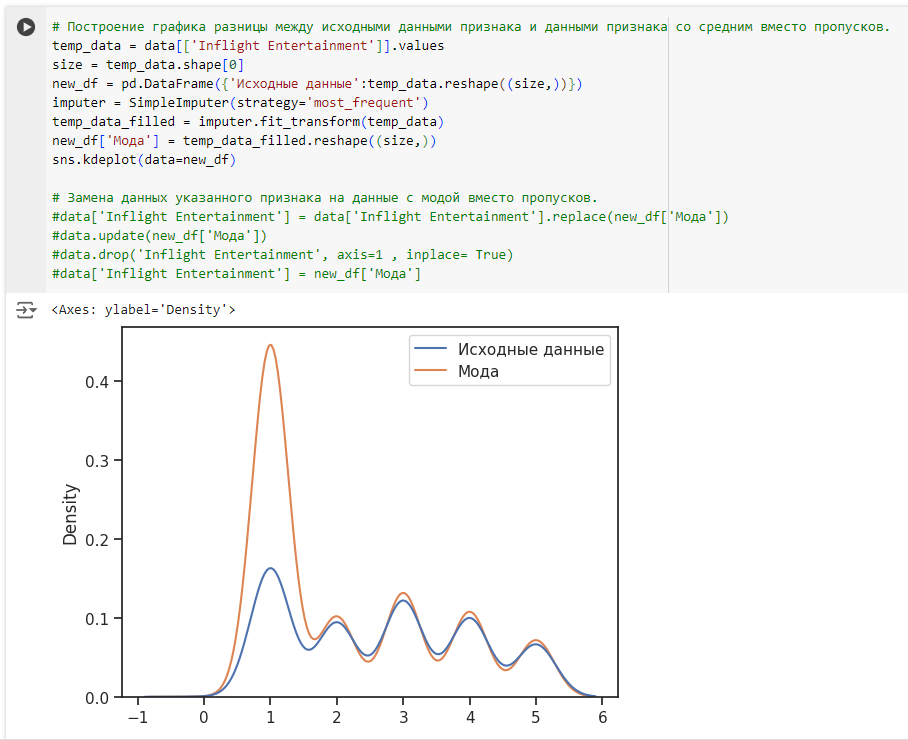


Из графиков можно отметить, что ближе всего распределение с заменой на основе среднего, поэтому применим его.

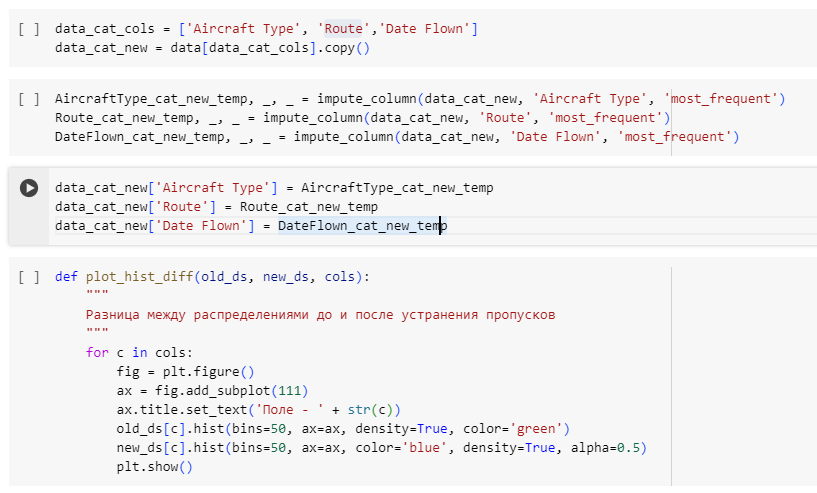


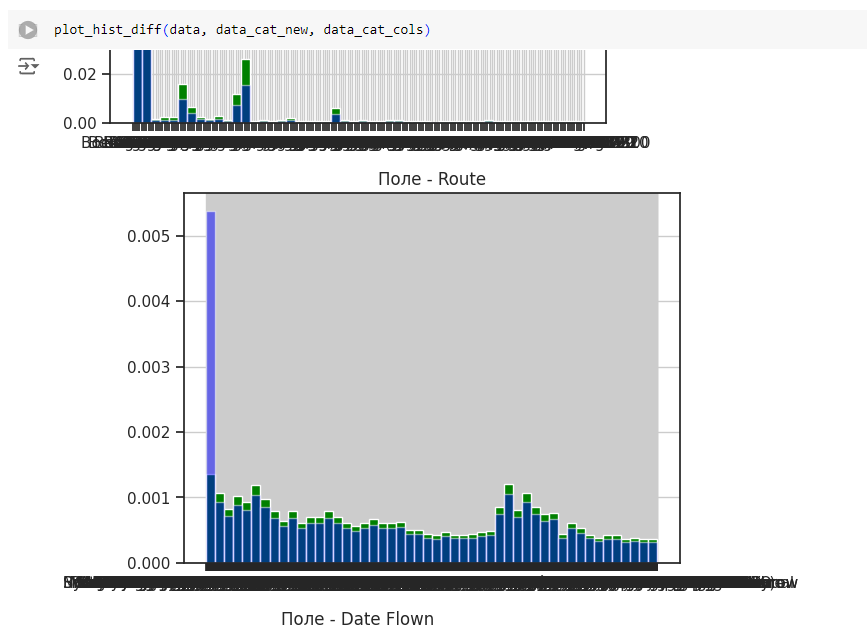
Аналогично исследуем столбец «Inflight Entertainment», для него распределение, которое ближе всего к оригинальному, на основе моды, соответственно, применим его.





Для категориального столбца «Route» используем заполнение наиболее распространенным значением категории.



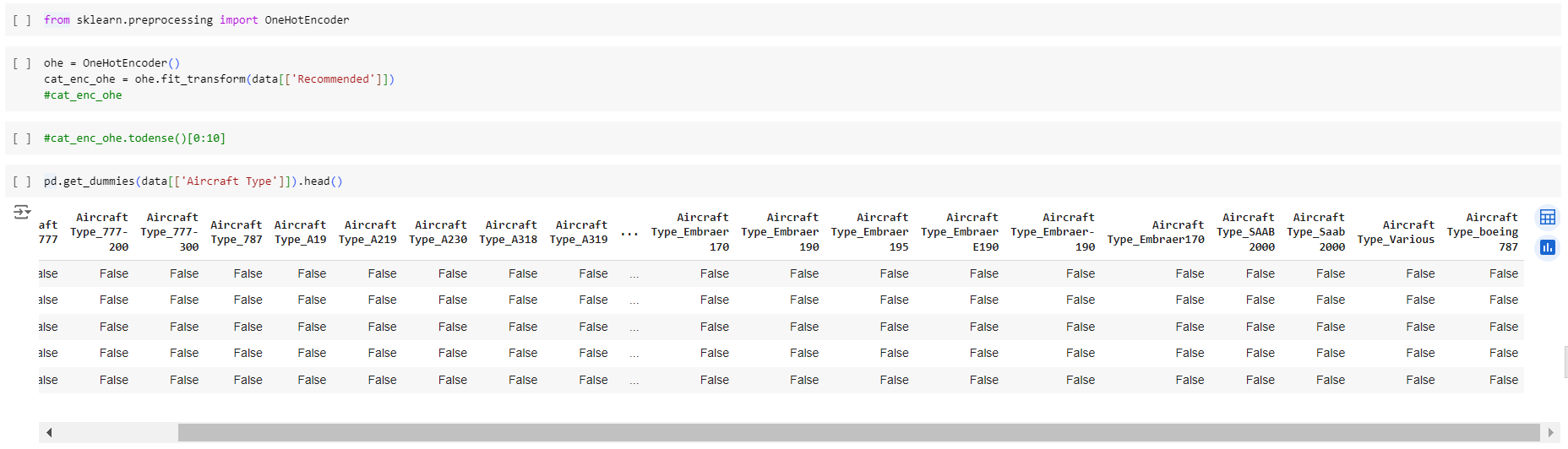


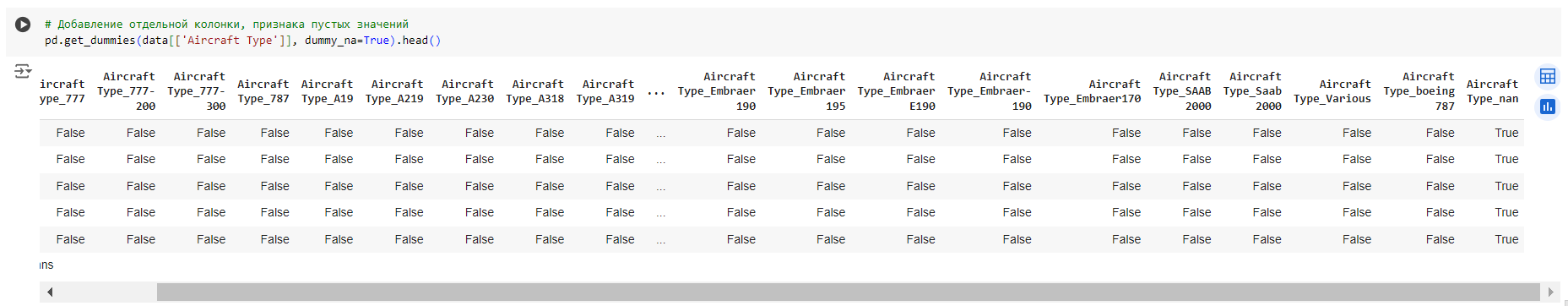
Таким образом, были заменены пропуски в числовых данных на основе среднего и моды, а для категориальных данных – на основе наиболее распространенного значения категории.

1. **Кодирование категориальных признаков**

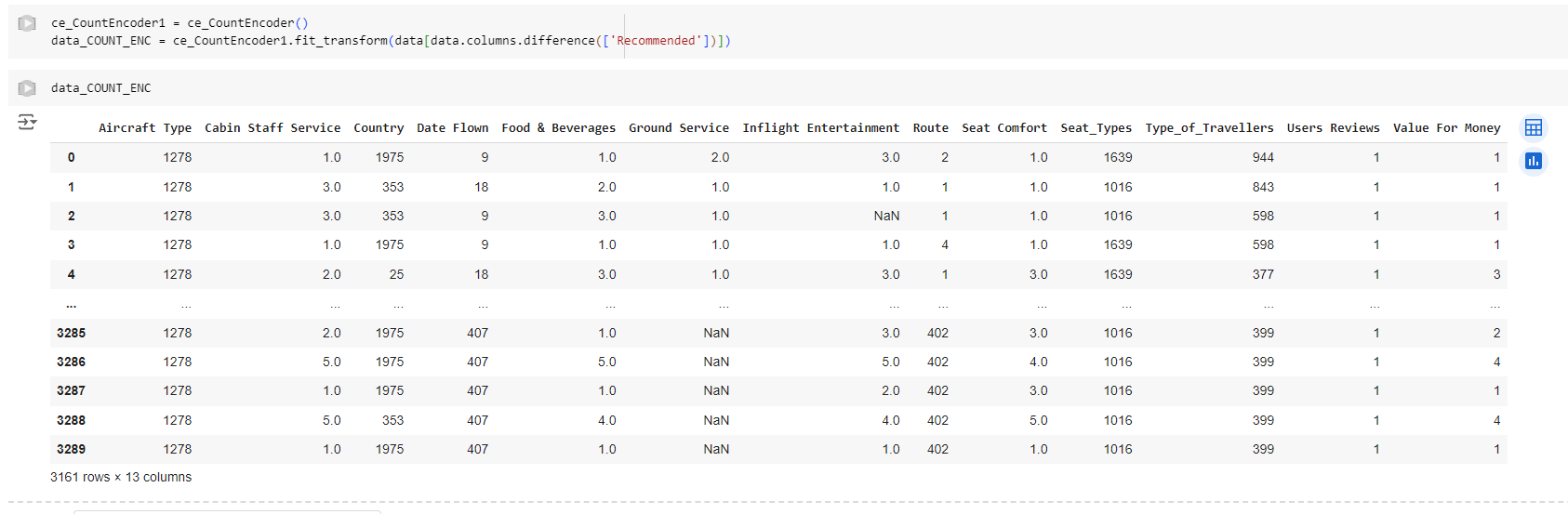
Для кодирования категориальных признаков будут использованы методы one-hot encoding, count encoding и frequency encoding.

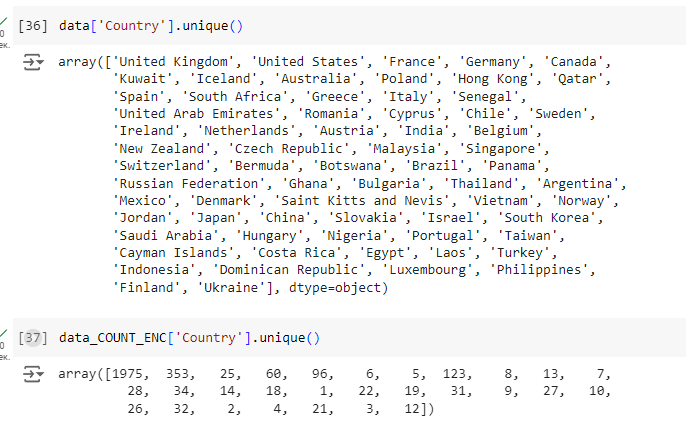
Для столбца «Aircraft Type» применим кодирование категорий наборами бинарных значений – one-hot encoding. Получим два варианта кодирования без и с отдельным столбцом для пустых значений, выведем примеры значений.



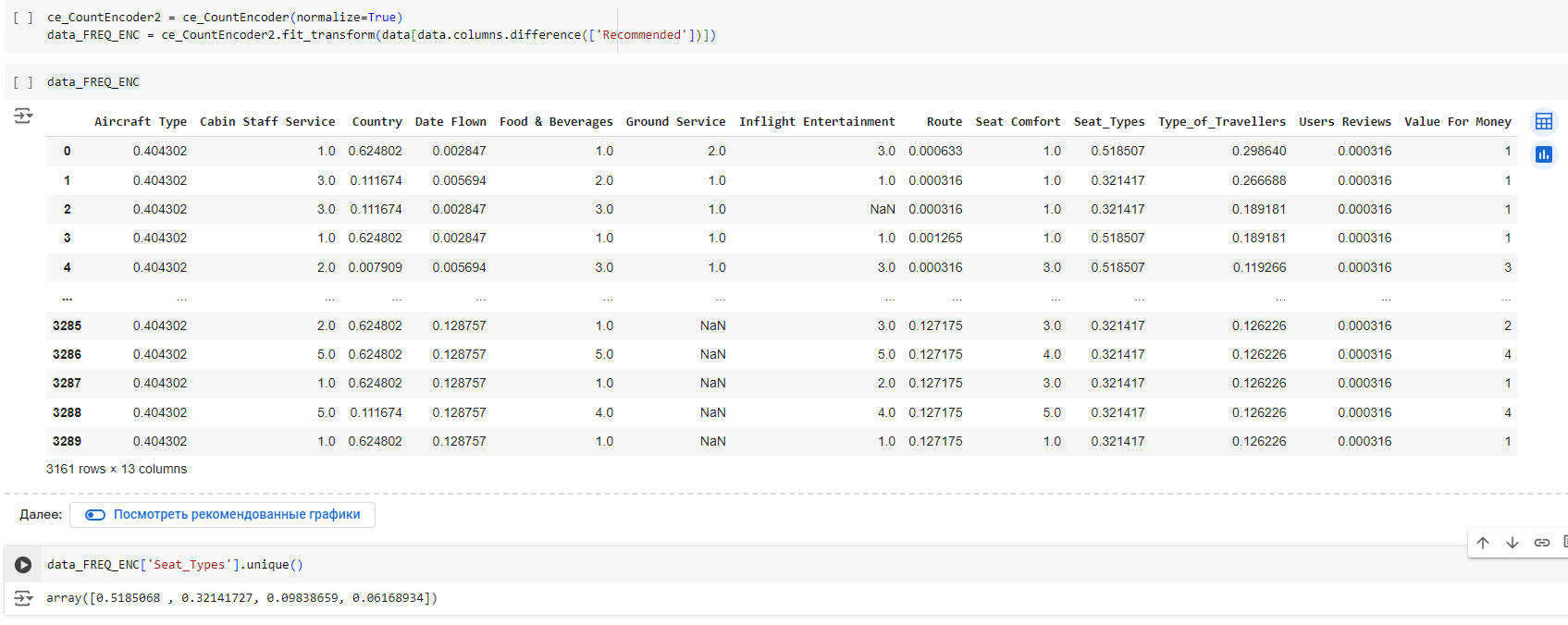


Для столбца «Country» применим кодирование размером категорий – count encoding.





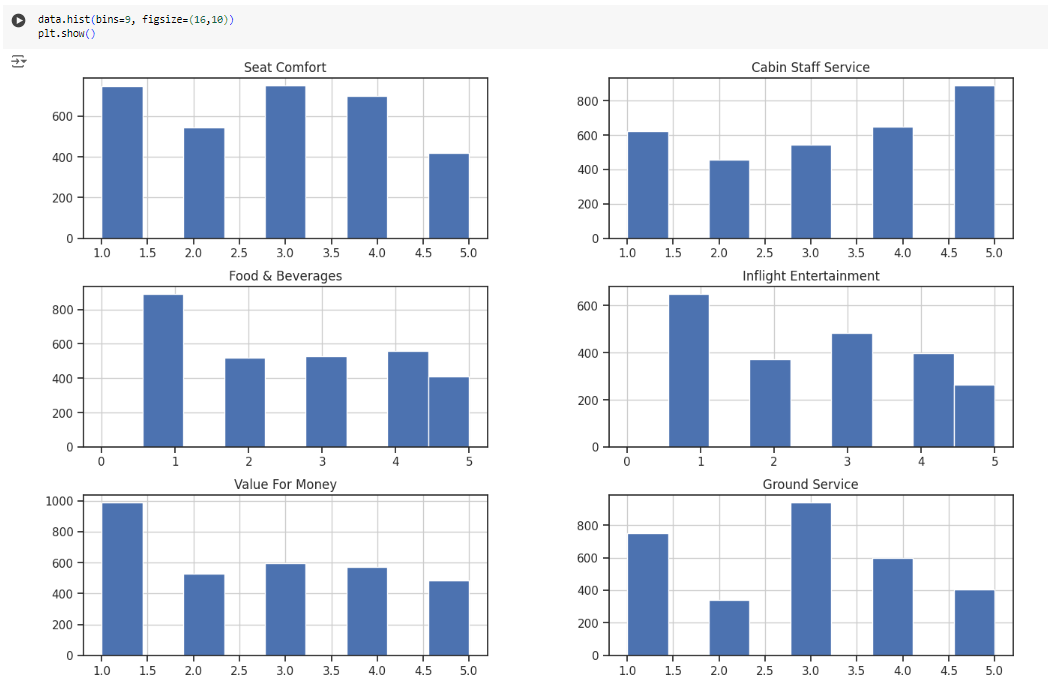
Для столбца «Seat\_Types» применим кодирование частотой категорий – frequency encoding.



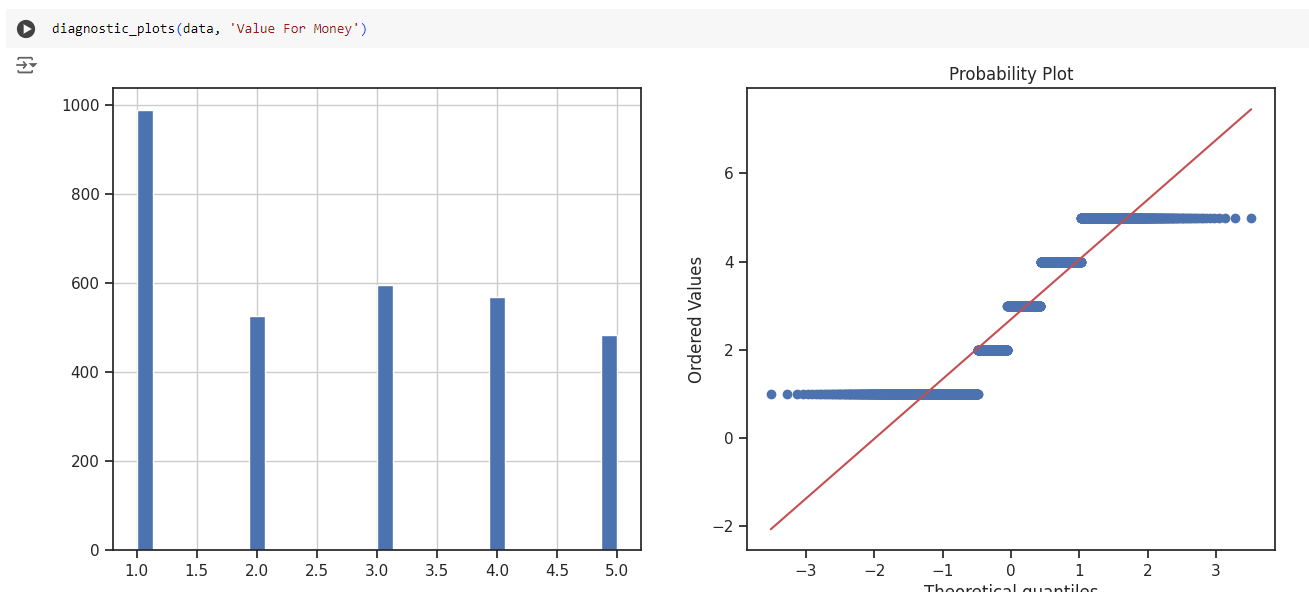
Таким образом, были использовано три метода для кодирования категориальных данных.

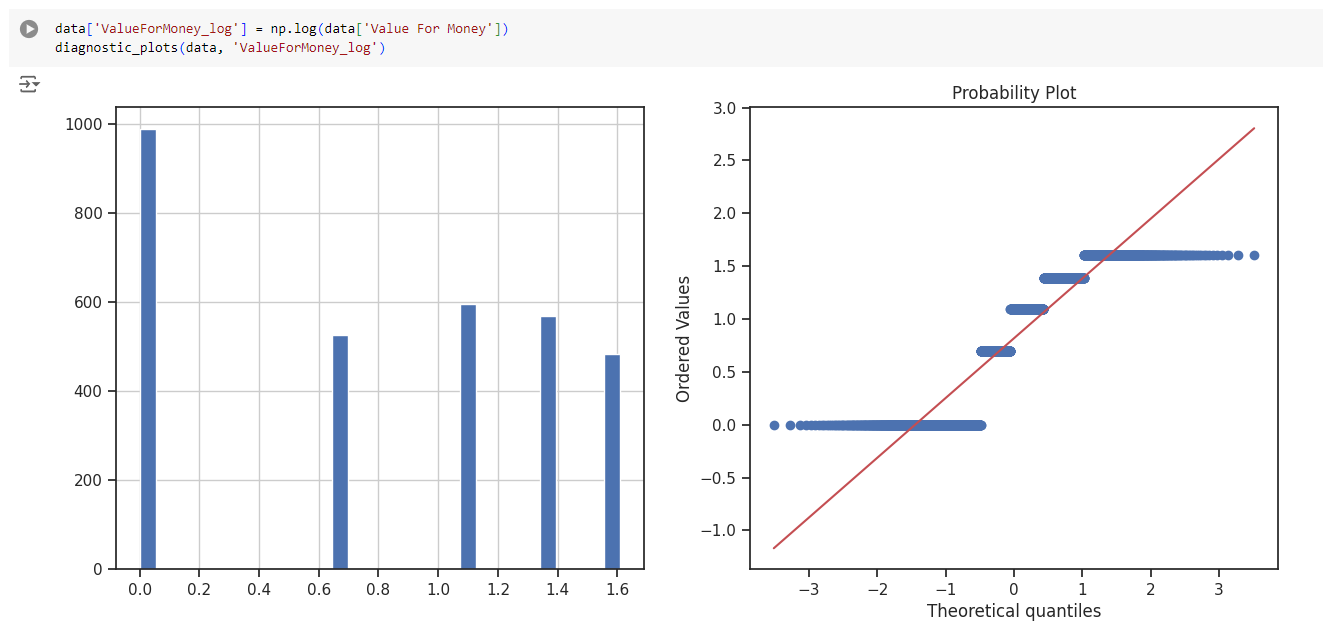
# **Нормализация числовых признаков**

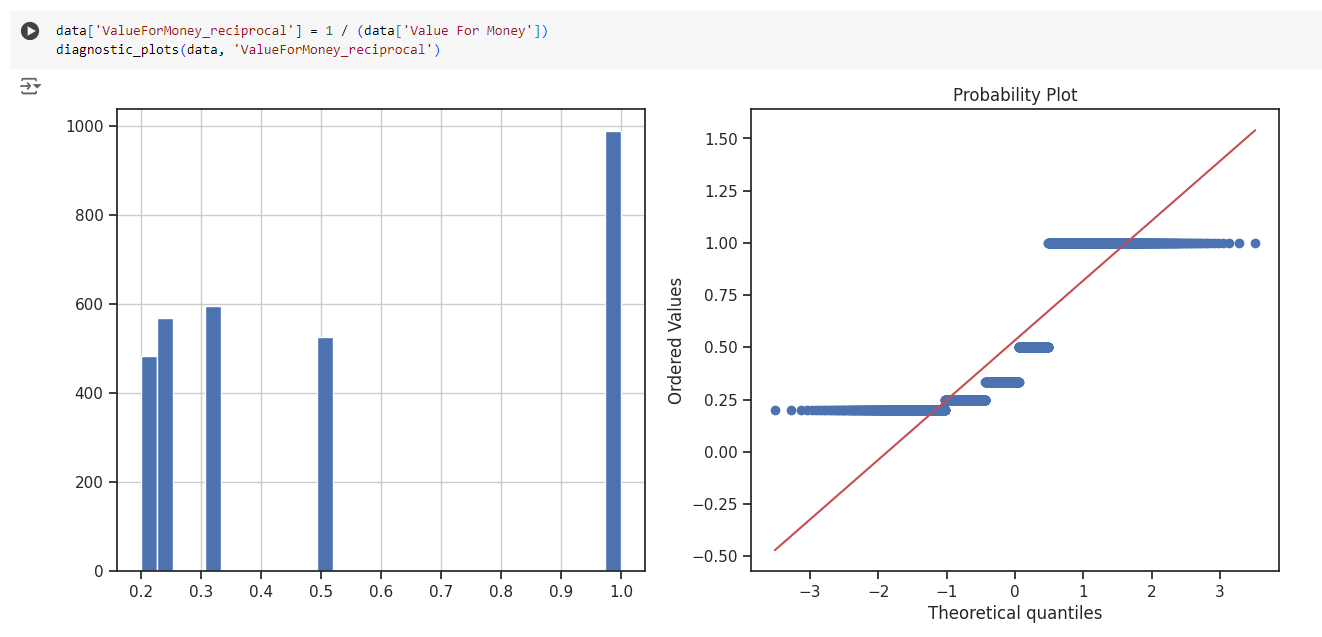
Для числовых признаков были построены гистограммы для обзора их распределений.

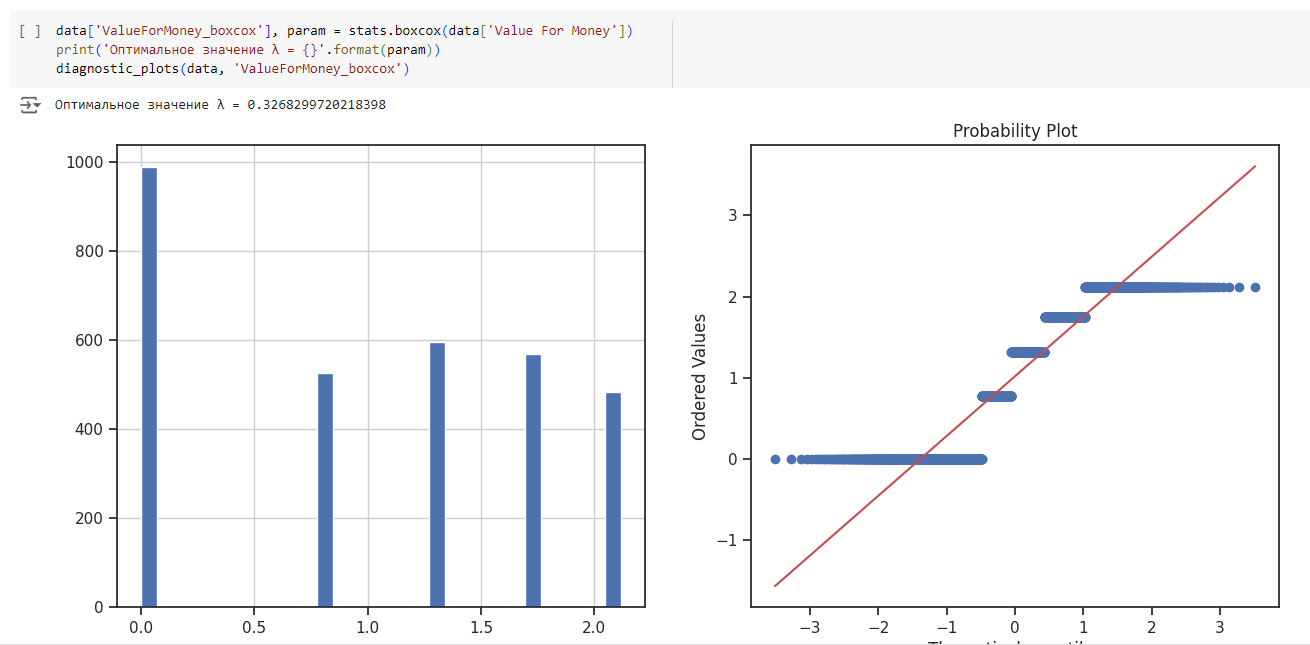


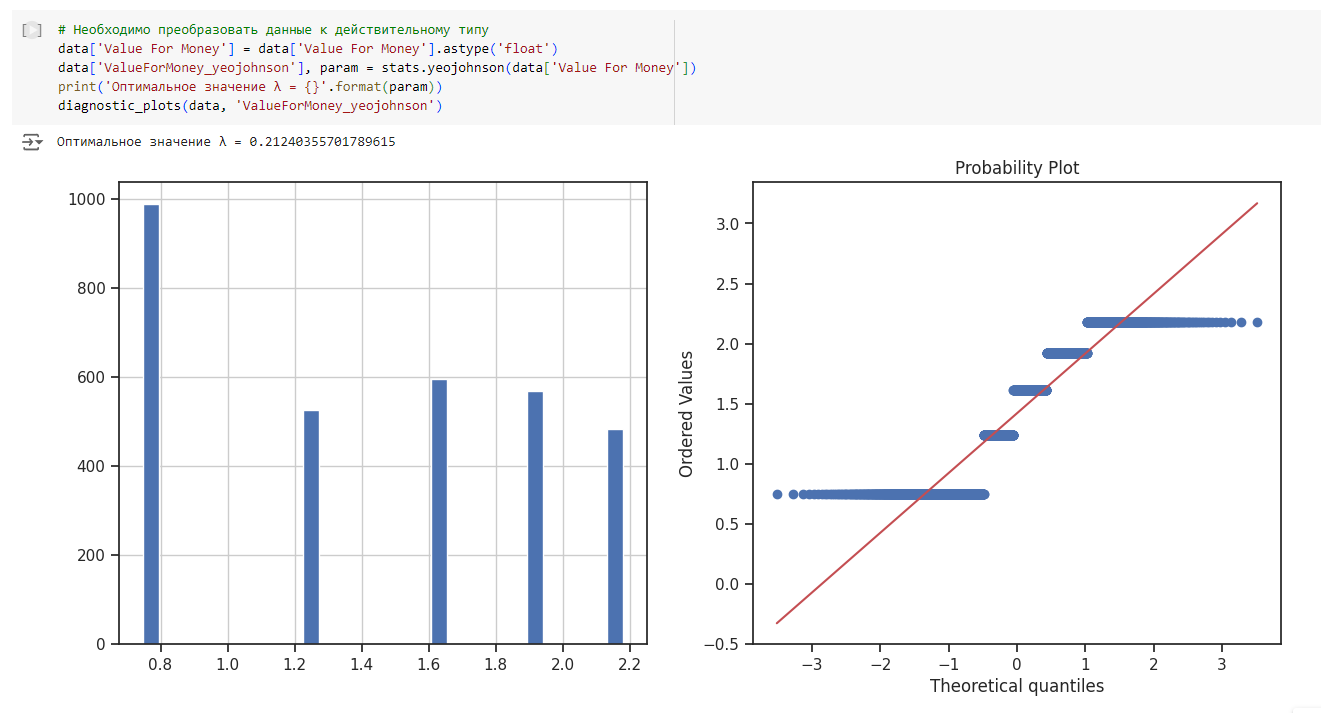
Из графиков можно отметить, что все столбцы имеют распределения неблизкие к нормальному. Следовательно, признаки необходимо нормализовать. Для столбца «Value For Money» применим логарифмическое преобразование, обратное преобразование, преобразование Бокса-Кокса и преобразование Йео-Джонсона.





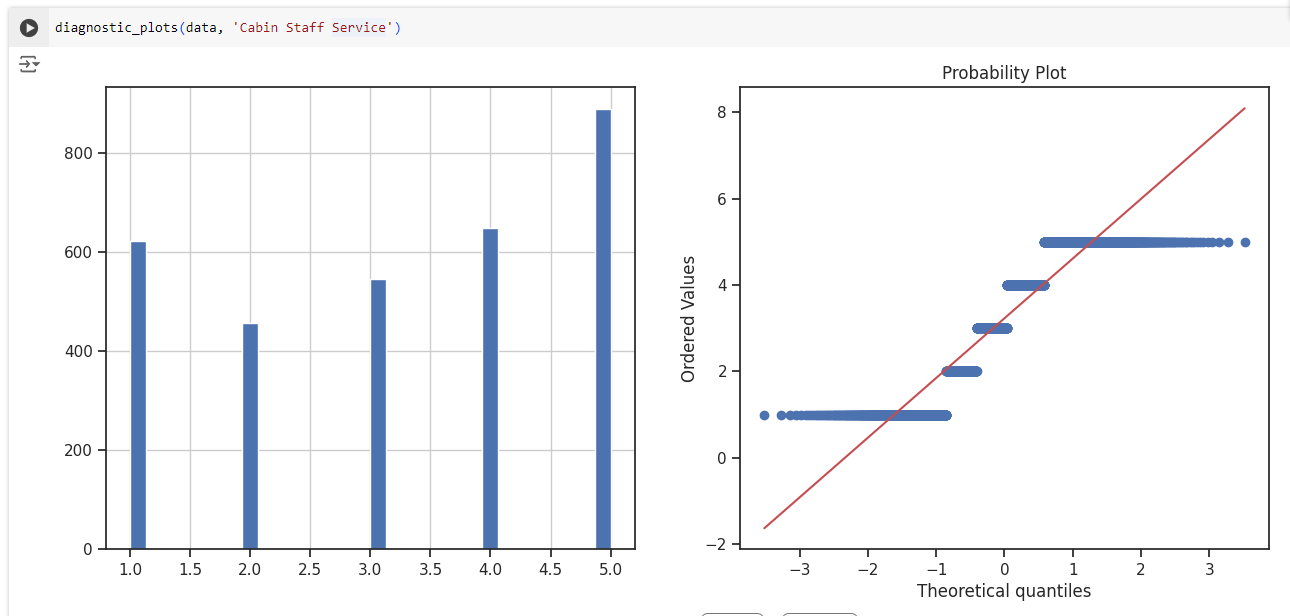


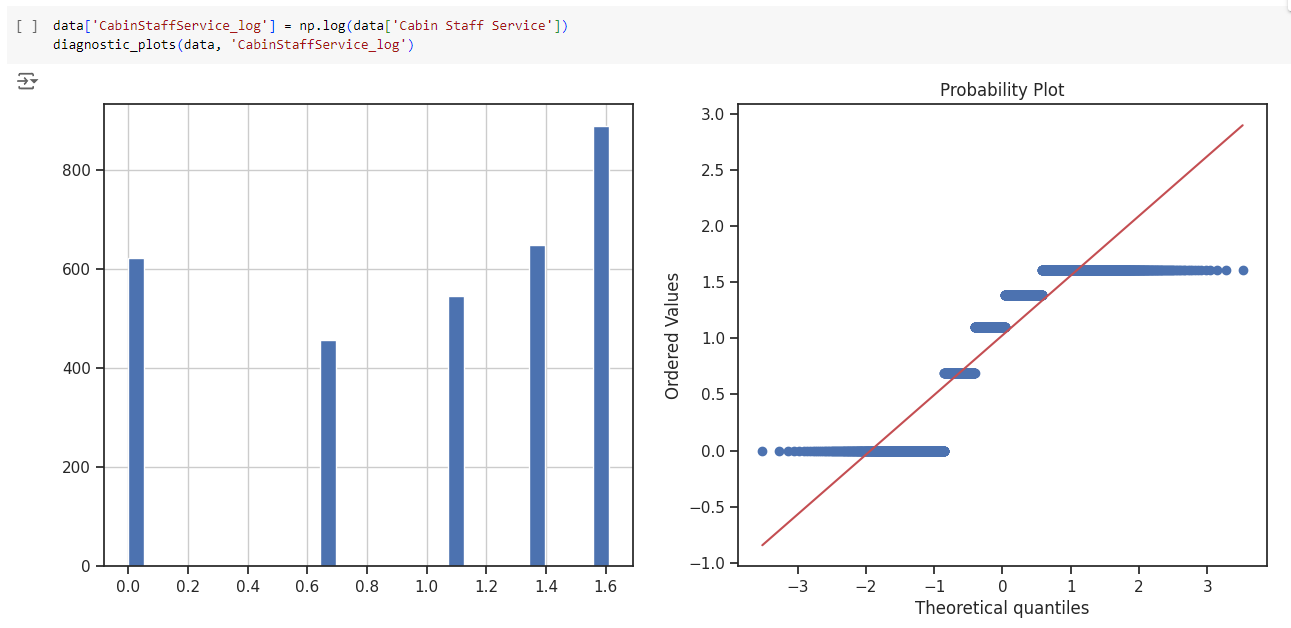


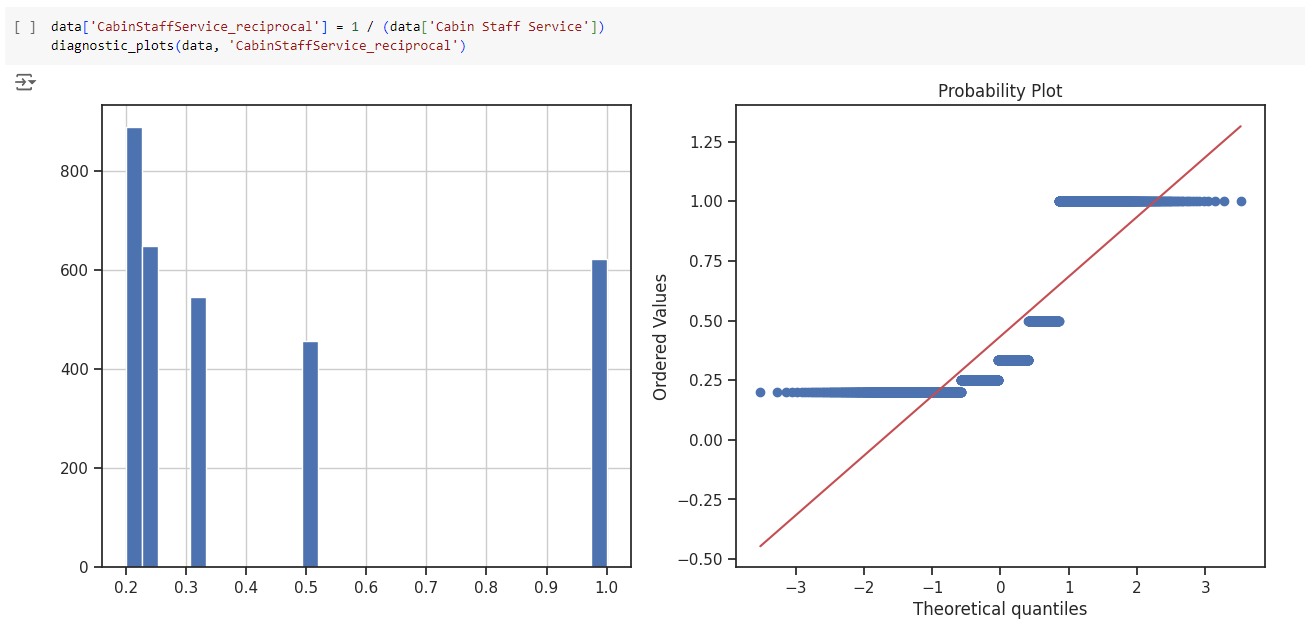


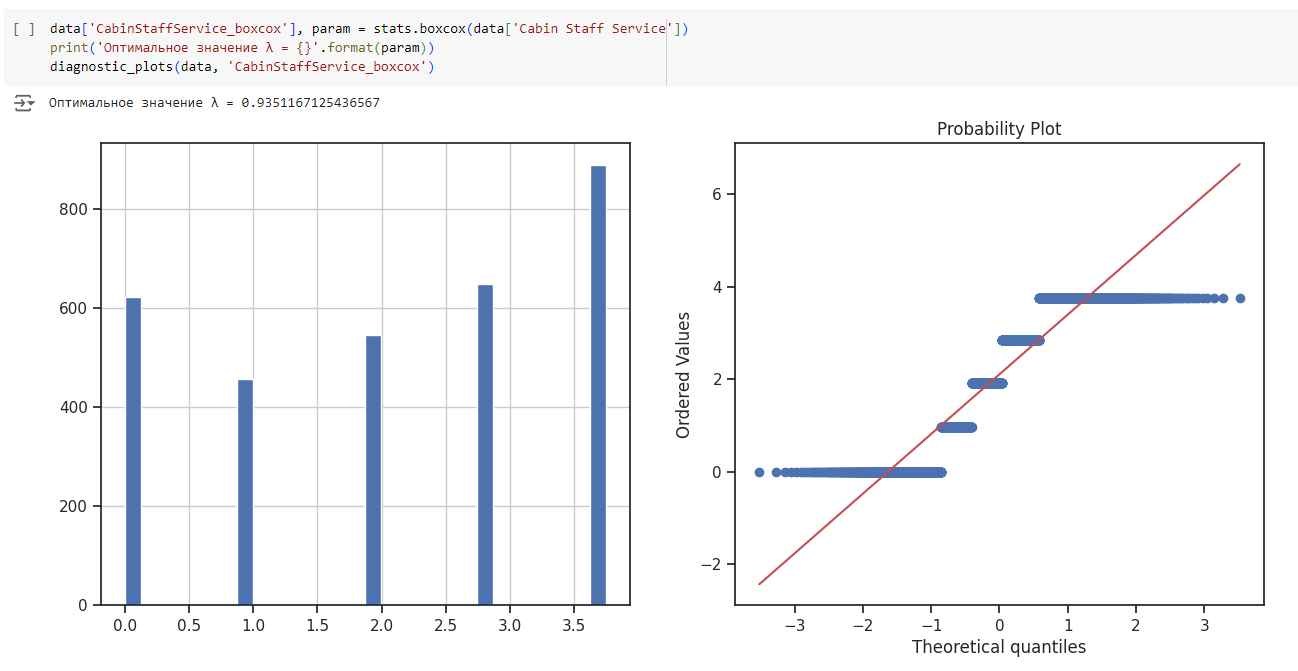
Из графиков можно отметить, что существенно приблизить распределение признака к нормальному не удалось.

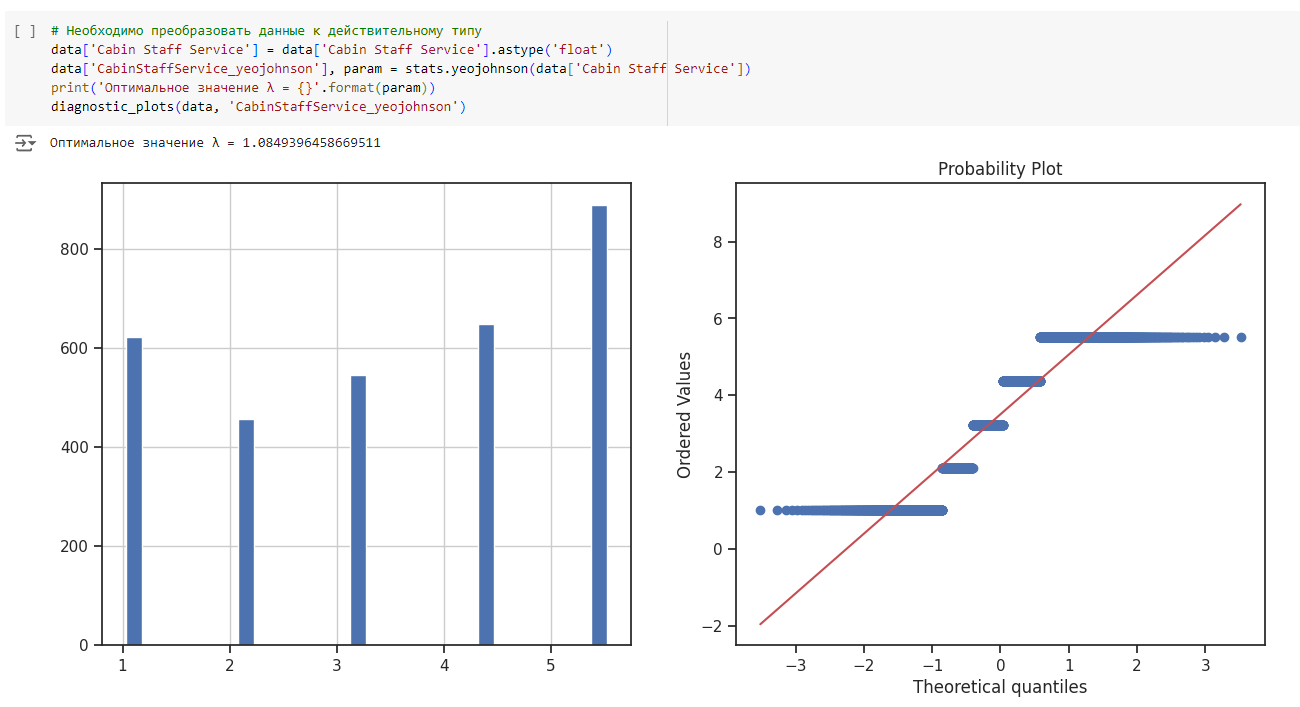
Те же преобразование применим к столбцу «Cabin Staff Service».











Таким образом, были использованы несколько преобразований для нормализации числовых данных, однако были получены не очень хорошие результаты.