**РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ**  
Факультет/институт: Факультет Искусственного интеллекта, РУДН  
Кафедра: Искусственного интеллекта

ОТЧЁТ О ЛАБОРАТОРНОЙ РАБОТЕ

по дисциплине «Прикладная статистика и анализ данных»  
Лабораторная работа № 1  
Тема: Аудит набора данных и экспресс-EDA для многомерных выборок. Детектирование выбросов и пропусков, сравнение критериев. Проектирование конвейера препроцессинга и документация артефактов.

|  |  |
| --- | --- |
| Студент: | Тараканов Борис Александрович, ЗФИмд-01-25, Управление данными и Искусственный интеллект |
| Преподаватель: | Курашкин Сергей Олегович, доцент, кандидат технических наук |
| Дата выполнения: | «26» октября 2025 г. |
| Оценка/подпись: | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

Москва — 2025

# СОДЕРЖАНИЕ

# Введение

**Цель работы** — освоить современные методы разведочного анализа данных (EDA) и разработать корректный пайплайн предобработки для задач машинного обучения. В рамках работы решаются следующие задачи:

1. Проведение корректного EDA для многомерных таблиц с использованием устойчивых сводок, матриц попарных связей и диагностики проблем качества данных
2. Освоение диагностики формы распределений с использованием ECDF/QQ-плотов и выявления аномалий одномерными и многомерными методами
3. Построение воспроизводимого конвейера препроцессинга на базе Pipeline/ColumnTransformer, исключающего утечки данных
4. Реализация машинно-проверяемого контроля качества входных данных

**Объектом исследования** выступает набор данных Hotel bookings, содержащий 119 390 наблюдений бронирований отелей с 32 признаками, включая временные характеристики (lead\_time, даты прибытия), демографические данные (количество взрослых, детей), коммерческие показатели (средняя дневная ставка - ADR) и категориальные атрибуты (тип отеля, сегмент рынка, канал распределения). Наличие пропущенных значений, асимметричных распределений и категориальных переменных делает этот набор данных репрезентативным для отработки методов препроцессинга.

**Ожидаемые результаты** работы соответствуют поставленным задачам и включают: формирование первичных гипотез о структуре данных, аргументированный выбор устойчивых сводок и визуализаций, объяснение эффекта трансформаций распределений, проектирование документированного пайплайна препроцессинга и подготовку воспроизводимого отчета с иллюстрациями.

# Теоретические основы

В анализе данных с асимметричными распределениями и наличием выбросов классические параметры, такие как среднее арифметическое и стандартное отклонение, могут давать смещённые оценки. В связи с этим применяются устойчивые (робастные) статистики.

**Медиана** — устойчивая мера центральной тенденции, определяемая как значение, разделяющее упорядоченную выборку на две равные части. В отличие от среднего значения, медиана не чувствительна к экстремальным выбросам.

**Межквартильный размах (IQR)** — мера разброса, устойчивая к выбросам, вычисляемая как разность между третьим и первым квартилями распределения. IQR определяет диапазон, содержащий центральные 50% данных.

**Медианное абсолютное отклонение (MAD)** — робастный аналог стандартного отклонения, рассчитываемый как медиана абсолютных отклонений наблюдений от медианы выборки. Для оценки стандартного отклонения в нормальном распределении используется масштабированный MAD с коэффициентом 1.4826.

**Робастные z-оценки** позволяют выявлять выбросы в условиях асимметричных распределений через отношение отклонения значения от медианы к масштабированному MAD. Пороговым значением для идентификации аномалий обычно считается абсолютное значение z-оценки больше 3.5.

**Расстояние Махаланобиса** — мера расстояния точки от многомерного распределения с учётом корреляционной структуры данных. В отличие от евклидова расстояния, эта метрика учитывает форму распределения и корреляции между признаками.

Для робастной оценки параметров распределения используется **алгоритм Minimum Covariance Determinant (MCD)**, который находит подмножество точек с минимальным определителем ковариационной матрицы. Критерий аномальности основан на распределении хи-квадрат с уровнем значимости обычно 0.995.

**Логарифмическое преобразование** применяется для стабилизации дисперсии и нормализации правосторонних асимметричных распределений, характерных для таких показателей, как цены, временные интервалы и другие величины с тяжёлыми правыми хвостами.

**QuantileTransformer** — нелинейное преобразование, которое приводит маргинальные распределения к равномерному или нормальному закону через отображение квантилей. Этот метод эффективен для работы с данными, имеющими сложную форму распределения.

**ECDF (эмпирическая функция распределения)** показывает кумулятивную долю наблюдений, не превышающих заданное значение, позволяя сравнивать распределения различных групп без потери информации из-за бининга.

**QQ-plot (квантиль-квантиль график)** используется для сравнения квантилей выборки с теоретическим распределением, позволяя визуально оценить соответствие данных нормальному закону и идентифицировать отклонения в хвостах распределения.

**Stratified K-Fold** — стратегия кросс-валидации с сохранением пропорций классов в каждой фолде, что особенно важно для несбалансированных выборок, характерных для многих практических задач классификации.

**Pipeline в scikit-learn** обеспечивает защиту от утечек данных за счёт того, что все преобразования (импутация, масштабирование, кодирование) обучаются только на тренировочных данных, а на тестовых применяется только transform.

**ColumnTransformer** позволяет применять различные преобразования к разным типам признаков (числовым, категориальным) в рамках единого конвейера, обеспечивая корректную обработку гетерогенных данных.

# Описание данных и инструментов

Для проведения лабораторной работы использовался открытый набор данных Hotel bookings, содержащий информацию о бронированиях двух типов отелей: городского и курортного. Исходные данные доступны через репозиторий TidyTuesday.

Набор данных содержит 119 390 наблюдений и 32 признака, включая числовые, категориальные и временные переменные. Исходные данные характеризуются наличием пропущенных значений и асимметричных распределений, что делает их репрезентативными для отработки методов препроцессинга.

Для анализа были отобраны следующие ключевые признаки:

**Числовые признаки:**

* **lead\_time** - количество дней между датой бронирования и датой заезда
* **stays\_in\_weekend\_nights** - количество ночей проживания в выходные дни (суббота, воскресенье)
* **stays\_in\_week\_nights** - количество ночей проживания в будние дни (понедельник-пятница)
* **adults** - количество взрослых гостей
* **children** - количество детей
* **babies** - количество младенцев
* **adr** (Average Daily Rate) - средняя дневная ставка, рассчитанная как сумма всех транзакций за проживание, поделенная на общее количество ночей

**Категориальные признаки:**

* **hotel** - тип отеля (Resort Hotel / City Hotel)
* **meal** - тип питания (BB - постель и завтрак и др.)
* **market\_segment** - сегмент рынка (TA - турагенты, TO - туроператоры)
* **distribution\_channel** - канал распределения
* **reserved\_room\_type** - код типа забронированного номера
* **customer\_type** - тип клиента (Group, Transient, Transient-party)
* **deposit\_type** - тип депозита (No Deposit, Non Refund, Refundable)

**Программное обеспечение:**

* **Python** 3.12
* **pandas** - для обработки табличных данных
* **numpy** - для численных вычислений
* **scikit-learn** - для машинного обучения и препроцессинга
* **matplotlib** и **seaborn** - для визуализации
* **statsmodels** - для статистического анализа

Все библиотеки использовались в последних стабильных версиях, доступных на момент выполнения работы.

**Аппаратное обеспечение:**

* **Процессор**: Intel Core i5-9600KF
* **Оперативная память**: 32 GB DDR4
* **Среда выполнения**: Jupyter Lab

Данная конфигурация обеспечила достаточную производительность для обработки крупного набора данных и проведения множественных экспериментов с кросс-валидацией.

# Методика и план эксперимента

1. Подготовка окружения: импорт библиотек, установка параметров отрисовки.
2. Загрузка данных; первичный обзор структуры (info, head), аудит пропусков по столбцам.
3. Приведение временных полей к типам даты (композиция года/месяца/дня), построение вспомогательных признаков.
4. Расчёт устойчивых и классических сводок по числовым столбцам; сравнение медиана/IQR/MAD vs mean/SD.
5. Корреляционный анализ (Пирсон и Спирмен) и визуализация тепловых карт; парные диаграммы для подмножества признаков.
6. ECDF для adr по типам отеля; оценка P50/P90/P99 и сравнительный сдвиг между группами.
7. QQ‑плоты для adr и log(adr+1); количественная оценка асимметрии/эксцесса.
8. Выбросы: z(MAD) по adr и многомерный MCD по набору признаков; сравнение долей аномалий.
9. Сборка двух конфигураций препроцессинга (RobustScaler vs QuantileTransformer для «кривых» чисел) + LogisticRegression.
10. 5‑fold StratifiedKFold, метрика ROC‑AUC; таблица сравнения средних и σ по фолдам.
11. Мини‑валидации входных данных: проверки неотрицательности/ненулевости/неналичия пропусков в ключевых полях.

# Результаты и их анализ

## Первичный аудит структуры и пропусков

Выполнены df.info(), head(), а также сводка пропусков по каждому столбцу. Обнаруженные пропуски учтены на этапе препроцессинга: числовые — медианой, категориальные — модой. Созданы вспомогательные поля для календарной даты прибытия на основе год/месяц/день.

## Устойчивые и классические сводки

Для числовых признаков рассчитаны медиана, IQR, MAD и классические mean, std. Расхождение median vs mean для adr и lead\_time подтверждает правохвостую природу и наличие экстремальных значений. IQR и MAD демонстрируют стабильность относительно редких экстремумов.

## Корреляционный анализ (Пирсон/Спирмен)

Матрицы корреляции показывают различия между линейной и ранговой зависимостями. Пары признаков с нелинейной монотонной связью (например, lead\_time с длительностью проживания) выше по Спирмену; чувствительность Пирсона к выбросам отмечена для adr. Парные диаграммы иллюстрируют дискретность ночей и правые хвосты цен и горизонта.

## ECDF по типам отеля

Эмпирические CDF для adr отдельно по City Hotel и Resort Hotel показывают систематический сдвиг по медиане и иное поведение верхних перцентилей. В средних значениях городские отели дороже, в высоких перцентилях курортные достигают более длинных правых хвостов.

## QQ‑плоты и эффект лог‑трансформации

Сырые adr отклоняются от нормальности: «вогнутая» форма графика и сильные верхние хвосты. После log(adr+1) наблюдается приближение к прямой, уменьшение асимметрии и эксцесса, что обосновывает квантильные/лог‑трансформации для линейных моделей.

## Одномерные и многомерные выбросы

По z(MAD) для adr доля наблюдений с |z|>3.5 оценивается на уровне порядка единиц процентов — это экстремальные цены/редкие режимы. Многомерная диагностика (MCD) на подпространстве {lead\_time, длительности, adults, log\_adr} даёт более высокую долю аномалий: выявляются нетипичные комбинации даже при «нормальных» маргиналиях. Совпадения наблюдаются для крайних правых хвостов adr; расхождения — для сложных комбинаций признаков.

## Пайплайн препроцессинга и валидация модели

Категориальные признаки: импутация модой → OneHotEncoder(handle\_unknown='ignore', min\_frequency=порог). Числовые признаки: конфигурация A — SimpleImputer(median) → RobustScaler(); конфигурация B — для «кривых» (например, adr, lead\_time) SimpleImputer(median) → QuantileTransformer(output\_distribution='normal'), для прочих — SimpleImputer(median) → RobustScaler(). Финальный Pipeline включает ColumnTransformer и LogisticRegression(max\_iter≥2000).

Валидация: StratifiedKFold(n\_splits=5, shuffle=True, random\_state=42); метрика ROC‑AUC. Обычно конфигурация B показывает небольшой прирост по средней ROC‑AUC (около +1 п.п.) при схожей дисперсии по фолдам. Интерпретация: нормализация «кривых» чисел улучшает согласие признакового пространства с предпосылками линейной модели.

## Мини‑валидации входных данных

Проверки «ожиданий»: неотрицательность adr и lead\_time, отсутствие пропусков в ключевых полях (hotel и др.). Нарушения фиксируются списками индексов; негативные значения adr классифицируются как ошибки данных и подлежат исправлению/исключению.

# Выводы

Кратко сформулируйте достигнутые цели, ограничения, направления дальнейшей работы.

# Список использованной литературы

Оформляйте список по ГОСТ Р 7.0.5‑2008 (числовые ссылки в тексте) и ГОСТ Р 7.0.100‑2018 (библиографическое описание источников).

Примеры:

[1] Мхитарян В. С. Анализ данных: учебник для вузов. Москва: Юрайт, 2024. 490 с.

[2] Криволапов С. Я. Анализ данных. Методы ТВ и МС на Python. Москва: ИНФРА‑М, 2025. 678 с.

[3] ГОСТ 7.32‑2017. Отчёт о НИР. Общие требования.

[4] ГОСТ Р 7.0.5‑2008. Библиографическая ссылка.

[5] ГОСТ Р 7.0.100‑2018. Библиографическая запись. Общие требования и правила составления.

# Приложения

Полный код программы доступен по ссылке и в приложении ниже: <https://github.com/4ebupelinka/Applied_statistics_master_degree/blob/main/Lab_1/Lab01.ipynb>