Отчет по результатам выполнения заданий практикума

Студент: Жидкова Ю.М. 316

9 декабря 2024

Содержание

1 Описание первого датасета: Данные о поездках в такси

1.1 Общие сведения

Датасет содержит информацию о 9 998 поездках на такси в определенном регионе. Каждая строка представляет собой отдельную поездку и включает в себя различные характеристики, такие как дата и время начала и окончания поездки, количество пассажиров, расстояние поездки, координаты начальной и конечной точек, тип оплаты, стоимость поездки и другие финансовые показатели.

1.2 Структура датасета

Датасет состоит из 18 столбцов:

1. **Unnamed: 0**: Индекс строки (не имеет практического значения).
2. **VendorID**: Идентификатор поставщика услуг такси.
3. **tpep_pickup_datetime**: Дата и время начала поездки.
4. **tpep_dropoff_datetime**: Дата и время окончания поездки.
5. **passenger_count**: Количество пассажиров в поездке.
6. **trip_distance**: Расстояние поездки в милях.
7. **pickup_longitude**: Долгота начальной точки поездки.
8. **pickup_latitude**: Широта начальной точки поездки.
9. **RateCodeID**: Код тарифа.
10. **dropoff_longitude**: Долгота конечной точки поездки.
11. **dropoff_latitude**: Широта конечной точки поездки.
12. **payment_type**: Тип оплаты (наличные, кредитная карта и т.д.).
13. **fare_amount**: Стоимость поездки.
14. **extra**: Дополнительные сборы (например, за ночное время).
15. **mta_tax**: Налог MTA.
16. **tip_amount**: Сумма чаевых.
17. **tolls_amount**: Сумма платы за проезд по платным дорогам.
18. **improvement_surcharge**: Сбор на улучшение.
19. **total_amount**: Общая сумма оплаты за поездку.

1.3 Статистические характеристики

- **Количество строк**: 9 998
- **Среднее значение passenger_count**: 1.67
- **Среднее значение trip_distance**: 2.82 мили
- **Среднее значение fare_amount**: 11.97 долларов
- **Среднее значение total_amount**: 14.90 долларов
- **Минимальное значение fare_amount**: 0.00 долларов
- **Максимальное значение fare_amount**: 129.50 долларов
- **Минимальное значение trip_distance**: 0.00 мили
- **Максимальное значение trip_distance**: 38.50 мили

1.4 Выводы

Датасет предоставляет подробную информацию о поездках на такси, включая географические координаты, финансовые показатели и характеристики поездки. Данные могут быть использованы для анализа паттернов поездок, определения наиболее популярных маршрутов, оценки финансовой эффективности и других исследований в области такси.

2 Анализ второго датасета: Данные о здоровье студентов

2.1 Общие сведения

Датасет содержит информацию о 1000 студентах, включая их идентификатор, возраст, гендер, показатели сердечного ритма, артериального давления, уровень стресса (биосенсорный и самоотчет), часы занятий и работы над проектами, физическую активность, качество сна, настроение и уровень риска для здоровья.

2.2 Структура датасета

Датасет состоит из 14 столбцов:

1. **Student_ID**: Уникальный идентификатор студента.
2. **Age**: Возраст студента.
3. **Gender**: Пол студента (M - мужской, F - женский).
4. **Heart_Rate**: Сердечный ритм студента (ударов в минуту).
5. **Blood_Pressure_Systolic**: Систолическое артериальное давление (мм рт. ст.).
6. **Blood_Pressure_Diastolic**: Диастолическое артериальное давление (мм рт. ст.).
7. **Stress_Level_Biosensor**: Уровень стресса, измеренный биосенсором.
8. **Stress_Level_Self_Report**: Уровень стресса, сообщенный студентом.
9. **Physical_Activity**: Уровень физической активности (Low, Moderate, High).
10. **Sleep_Quality**: Качество сна (Poor, Moderate, Good).
11. **Mood**: Настроение (Happy, Neutral, Stressed).
12. **Study_Hours**: Количество часов, потраченных на учебу.
13. **Project_Hours**: Количество часов, потраченных на работу над проектами.
14. **Health_Risk_Level**: Уровень риска для здоровья (Low, Moderate, High).

2.3 Статистические характеристики

- **Количество строк**: 1000
- **Средний возраст**: 20.96 лет
- **Средний уровень стресса (биосенсор)**: 5.48
- **Средний уровень стресса (самоотчет)**: 5.36
- **Среднее количество часов на учебу**: 30.23 часа
- **Среднее количество часов на проекты**: 14.89 часа

2.4 Максимумы и минимумы

- **Минимальный возраст**: 18 лет
- **Максимальный возраст**: 24 лет
- **Минимальный уровень стресса (биосенсор)**: 1.01
- **Максимальный уровень стресса (биосенсор)**: 9.99
- **Минимальный уровень стресса (самоотчет)**: 1.00
- **Максимальный уровень стресса (самоотчет)**: 9.96
- **Минимальное количество часов на учебу**: 5 часов
- **Максимальное количество часов на учебу**: 60 часов
- **Минимальное количество часов на проекты**: 0 часов
- **Максимальное количество часов на проекты**: 32.72 часа

2.5 Выводы

Датасет предоставляет подробную информацию о студентах, включая их физиологические показатели, уровень стресса, физическую активность, качество сна, настроение и уровень риска для здоровья. Данные могут быть использованы для анализа влияния стресса и других факторов на здоровье студентов, определения факторов риска, оценки эффективности методов снижения стресса и других исследований в области психологии и медицины.

3 Аппроксимация распределений данных с помощью ядерных оценок

Kernel Density Estimation (KDE) — это непараметрический метод оценки плотности вероятности, который позволяет строить гладкую аппроксимацию распределения данных.

3.1 Формула

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

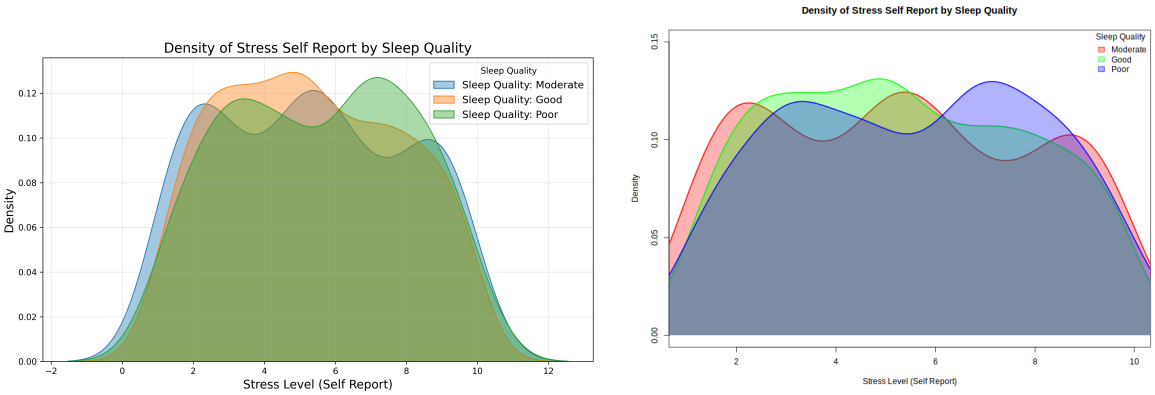
где $x_1, \ldots, x_n$ — данные, h — ширина окна, $K(u)$ — функция ядра (например, гауссовское ядро).

3.2 Основные характеристики

- **Ядро $K(u)$:** определяет форму сглаживания (например, гауссовское ядро $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$).
- **Ширина окна h:** управляет степенью сглаживания (маленькое h выявляет детали, большое h сглаживает).

3.3 Преимущества и недостатки

- **Преимущества:** не требует предположений о распределении, гибкость, удобство визуализации.
- **Недостатки:** выбор h влияет на результаты, высокая вычислительная сложность.

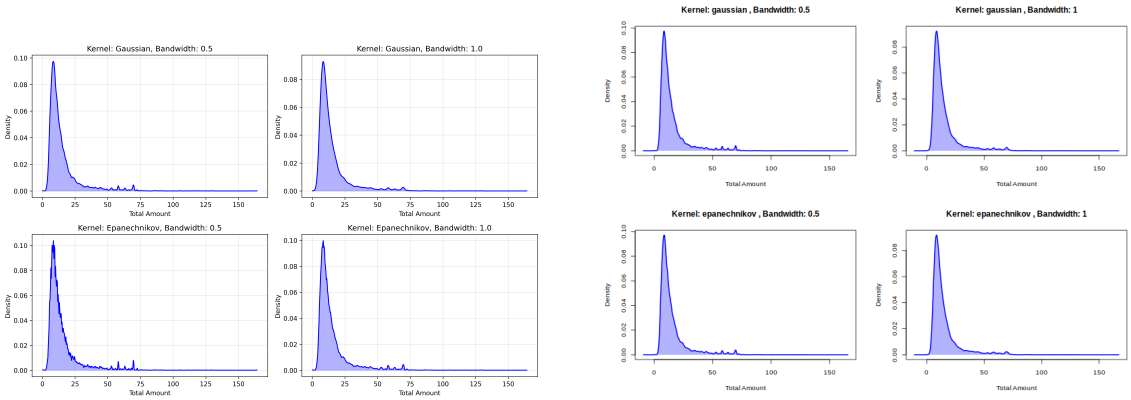


(a) график на Python (b) график на R

Рис. 1: KDE Уровня стресса в зависимости от качества сна

На графике изображены плотности распределения уровня стресса в зависимости от качества сна. Они построены с нормальным ядром и автоматически подобранной шириной окна. Мы можем увидеть по графику, как влияет качество сна на уровень стресса - для плохого качества сна распределение имеет больший пик в области значений 6-10, а для среднего и хорошего - в области 2-6.

На R и Python ядерные оценки получились одинаковые, но в графике Python не обрезана область значений по тому, какие на самом деле значений принимает переменная, а оценщик аппроксимирует так, что плотность ненулевая даже вне границ.

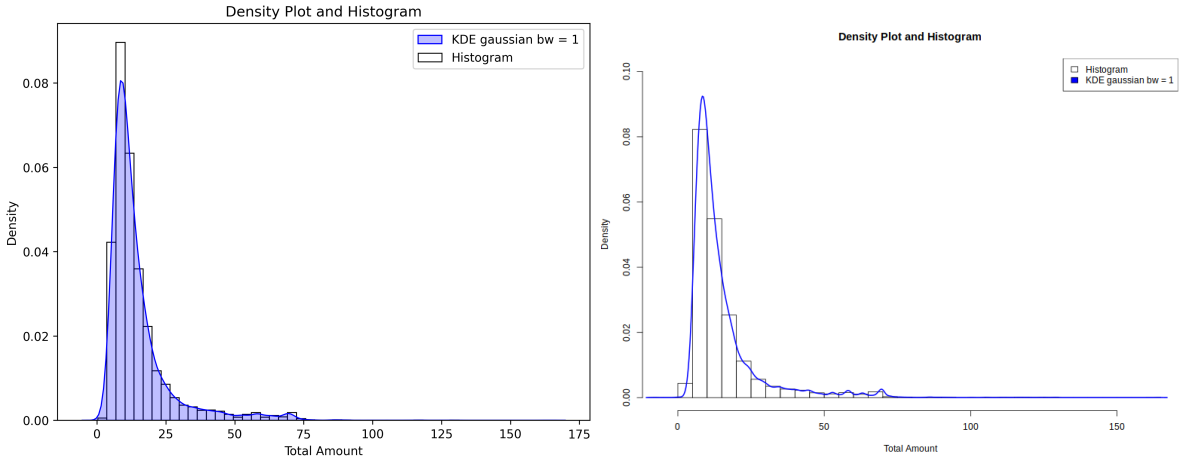


(a) график на Python (b) график на R

Рис. 2: KDE с различными ядрами и шириной окна

На этом графике мы пробуем аппроксимировать плотноть распределения переменной total_amount - стоимость поездки. Пробуем два ядра - нормальное ядро и ядро Епанечникова. При одинаковой ширине можем увидеть, что аппроксимация с гауссовским ядром более гладкая, это связано с свойствами самого ядра. Так же мы можем увидеть, что на сглаживание и чувствительность влияет ширина окна. Величина распределена так, что у нее один пик, одна мода, поэтому целом все графики не очень отличаются.

На R и на Python аппроксимации распределений получились одинаковые.



(a) график на Python (b) график на R

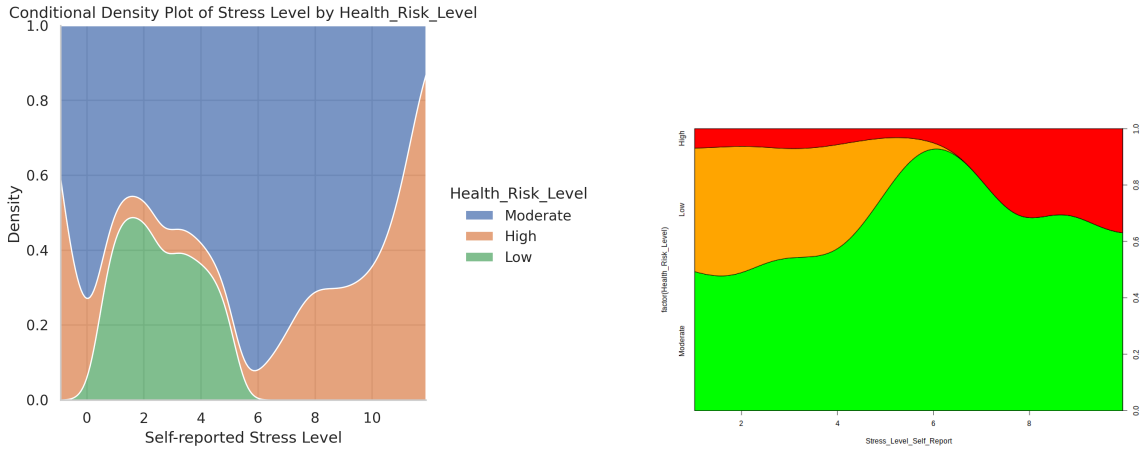
Рис. 3: KDE суммы оплаты поездки

На графике изображены гистограмма и ядерная оценка плотности (с гауссовским ядром, шириной окна = 1) распределения total_amount.

4 Анализ графиков

4.1 Conditional Density Plot

Описание графика: Условный график плотности показывает распределение уровней стресса (по оси X) в зависимости от категорий “Уровня риска для здоровья” (`Health_Risk_Level`).



(a) график на Python (b) график на R

Рис. 4: CD plot

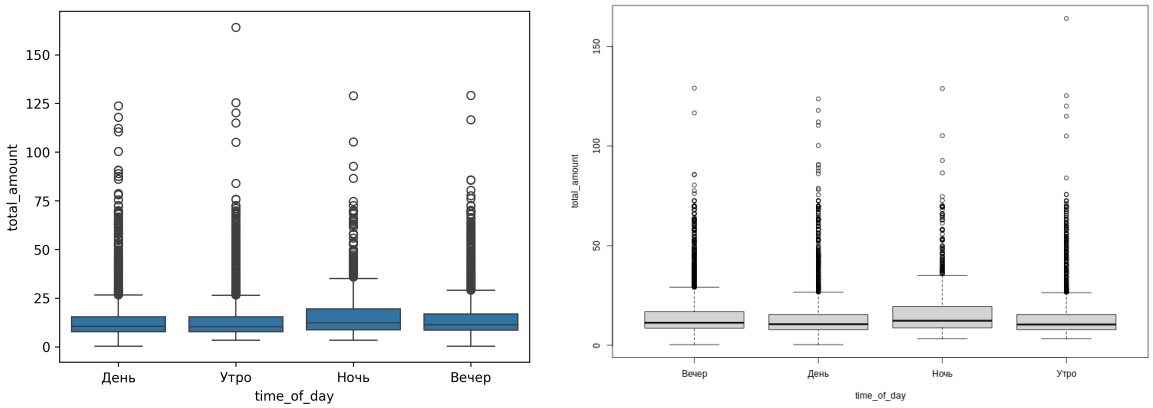
Выводы:

- Люди с **низким уровнем риска для здоровья** в основном сообщают о низких уровнях стресса (около 0–3).
- У людей с **умеренным риском** наблюдается почти равномерное распределение стресса, но пик заметен ближе к высоким значениям (около 9–10).
- У тех, кто относится к категории **высокого риска**, распределение стресса сильно сдвинуто в сторону высоких значений (6–10), что может свидетельствовать о том, что высокий уровень стресса коррелирует с ухудшением здоровья.

На R и на Python графики выглядят по разному, но к обоим графикам подходит написанная выше интерпретация.

4.2 Box Plot

Описание графика: Диаграммы размаха показывают распределение значений переменной “`total_amount`” (общее количество) в зависимости от времени суток (`time_of_day`), представленное категориями: День, Утро, Ночь, Вечер.



(a) график на Python

(b) график на R

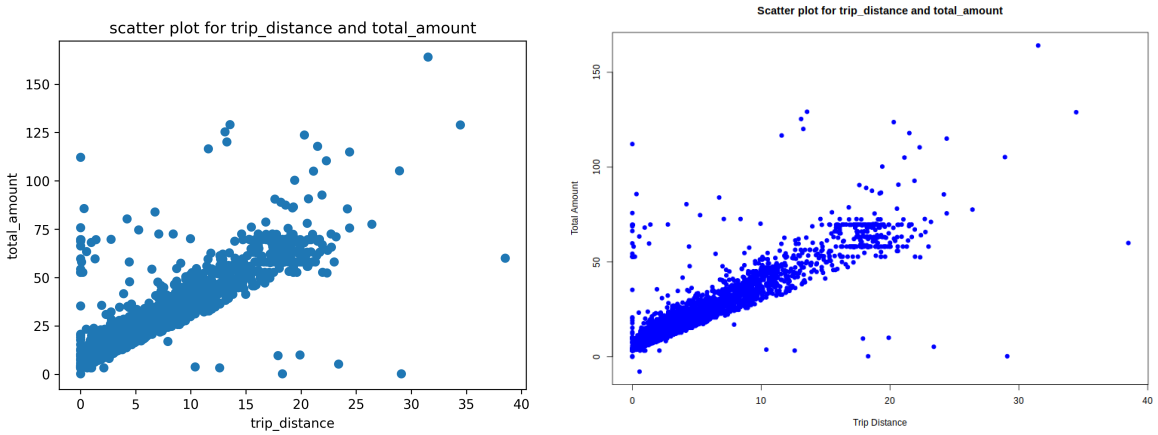
Рис. 5: CD plot

Выводы:

- Вне зависимости от времени суток медианное значение “`total_amount`” (величина в центре коробки) остаётся примерно одинаковым.
- Наибольшее количество выбросов (значения далеко за пределами усов) наблюдается в категориях “Ночь” и “Утро”. Это может говорить о более значительных колебаниях в сумме в эти периоды.
- Визуально разброс данных (размер межквартильного размаха) также примерно одинаков для всех категорий времени суток.
- В ночное время стоимость в целом выше, но ненамного.

4.3 Scatter Plot

Описание графика: Диаграмма рассеяния использует точки для представления значений двух разных числовых переменных. Положение каждой точки на горизонтальной и вертикальной оси указывает значения для отдельной точки данных. Диаграммы рассеяния используются для наблюдения за взаимосвязями между переменными. Диаграммы рассеяния показывает взаимосвязь значений переменных “`total_amount`” и `trip_distance`, - стоимость и длина поездки.



(a) график на Python

(b) график на R

Рис. 6: Scatter Plot

Выводы:

- Из графика видно, что переменные линейно зависят - это вплоне логично, так как сумма поездки, очевидно, будет зависеть от длины пути.

5 Критерии выявления выбросов

5.1 Критерий Граббса

Описание: Критерий Граббса используется для обнаружения одиночных выбросов в выборке с нормальным распределением. Он основан на расчете статистики, которая сравнивает абсолютное отклонение каждого значения от среднего с стандартным отклонением. Если эта статистика превышает определенный порог, наблюдение считается выбросом.

Как работает:

1. Рассчитывается среднее и стандартное отклонение выборки.
2. Для каждого значения вычисляется статистика Граббса:

$$G = \frac{|x_i - \bar{x}|}{s}$$

3. Сравнивается статистика с критическим значением для выбранного уровня значимости. Если статистика превышает критическое значение, наблюдение является выбросом.

Применимость:

- Применим только для нормально распределенных данных.
- Подходит для обнаружения одиночных выбросов.

5.2 Q-тест Диксона

Описание: Q-тест Диксона используется для выявления одиночных выбросов в малых выборках (обычно до 30 наблюдений). Он сравнивает разницу между соседними значениями с диапазоном всех данных.

Как работает:

1. Данные сортируются по возрастанию.
2. Для каждой пары соседних значений вычисляется значение Q:

$$Q = \frac{x_{i+1} - x_i}{x_n - x_1}$$

3. Вычисляется критическое значение Q для выбранного уровня значимости. Если вычисленное Q больше критического, то наблюдение считается выбросом.

Применимость:

- Работает только для малых выборок (около 30 наблюдений).
- Применим для поиска одиночных выбросов, а не для множества выбросов.

6 Проверка выброса

Рассмотрим данные о количестве вызовов такси в день в течение месяца. В данных 31 строка. Минимальное значение в выборке составляет 93. Проверим, является ли это значение выбросом с использованием статистических критериев.

6.1 Гистограмма данных

На рисунке ниже представлена гистограмма распределения данных:

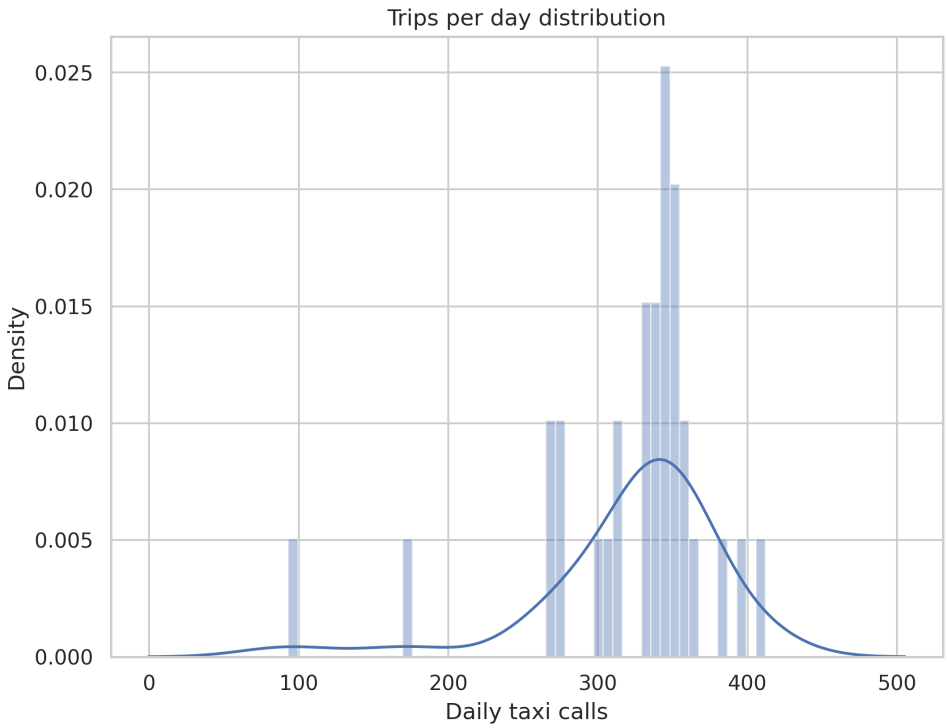


Рис. 7: Гистограмма распределения количества вызовов такси

6.2 Проверка на нормальность распределения

Сначала проверим данные на нормальность с помощью теста Шапиро-Уилка, так как мы будем применять к данным критерий Граббса, для которого требуется нормальность. Результаты теста:

- Статистика: $W = 0.9376$
- p-значение: $p = 0.0866$

Поскольку $p > 0.05$, гипотеза о нормальности данных не отвергается. Можно применять критерии на основе нормального распределения.

6.3 Проверка выброса тестом Диксона

Тест Диксона используется для проверки выброса в крайних значениях выборки. Результаты теста:

- $Q = 0.59727$
- p-значение: $p < 2.2 \times 10^{-16}$
- Альтернативная гипотеза: минимальное значение 93 является выбросом.

6.4 Проверка выброса критерием Граббса

Критерий Граббса применяется для выявления одиночных выбросов в выборке. Результаты теста:

- $G = 3.71271$
- $U = 0.52521$
- p-значение: $p = 0.0002815$
- Альтернативная гипотеза: минимальное значение 93 является выбросом.

6.5 Бокс-плот

На рисунке ниже представлен бокс-плот, который наглядно показывает распределение данных и наличие выбросов.

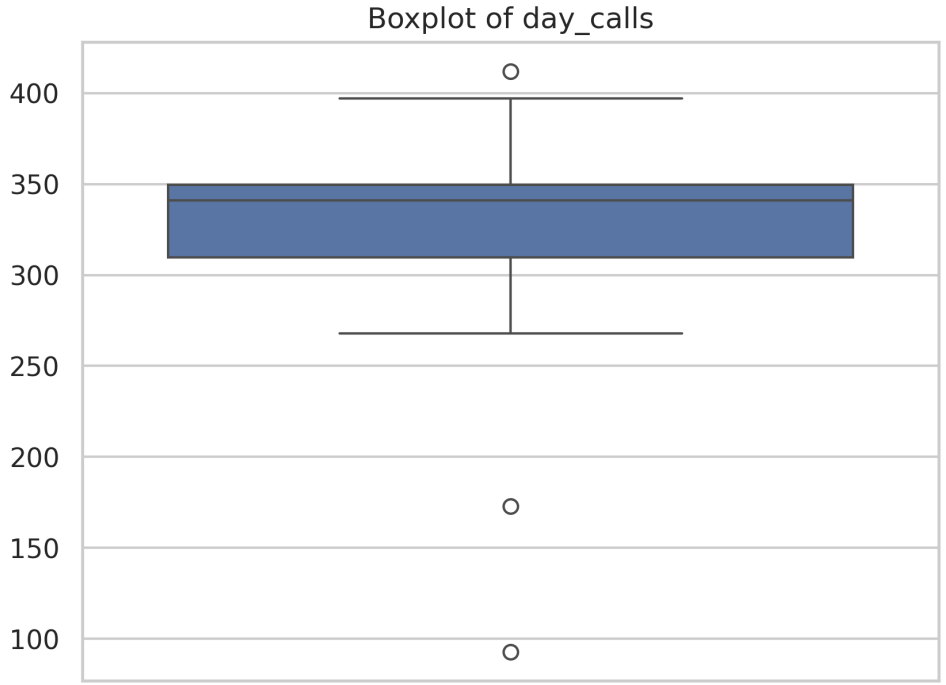


Рис. 8: Бокс-плот распределения количества вызовов такси

6.6 Выводы

На Python и R тесты дали одинаковый ответ. На питоне расчет критериев был реализованы вручную. На основе тестов Диксона и Граббса можно сделать вывод, что минимальное значение 93 является выбросом. Это подтверждается низкими p-значениями в обоих тестах, а также визуальным анализом на бокс-плоте.

7 Заполнение пропусков

Заполнение пропусков в данных — важный этап предобработки, который помогает сохранить полноту информации. Среди методов заполнения можно выделить использование моды, медианы и среднего значения. Мода подходит, когда необходимо заполнить пропуски наиболее частыми значениями, медиана — для данных с выбросами, так как она менее чувствительна к аномальным значениям, а среднее значение часто используется в случае нормального распределения данных. Каждый метод имеет свои особенности, и выбор зависит от характера данных.

7.1 Заполнение пропусков модой

Рассмотрим заполнение пропусков модой для переменной passenger_count, которая является категориальной и принимает значения от 1 до 6. (Пропуски созданы вручную, 100) Мы заменим все пропущенные значения на моду, то есть на наиболее часто встречающееся значение в данных. После анализа мы увидим, что значение 1 встречается гораздо чаще остальных, что делает его ярко выраженной модой. В данном случае использование моды для заполнения пропусков оправдано, так как это значение значительно преобладает, и его выбор позволяет сохранить статистическую целостность данных. На графике мы можем увидеть гистограммы данных до внесения пропусков и после заполнения.

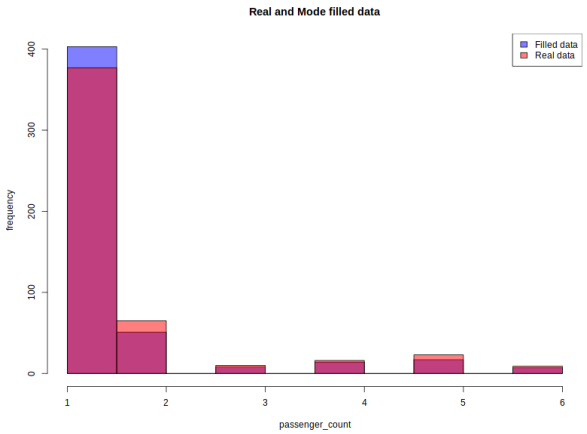
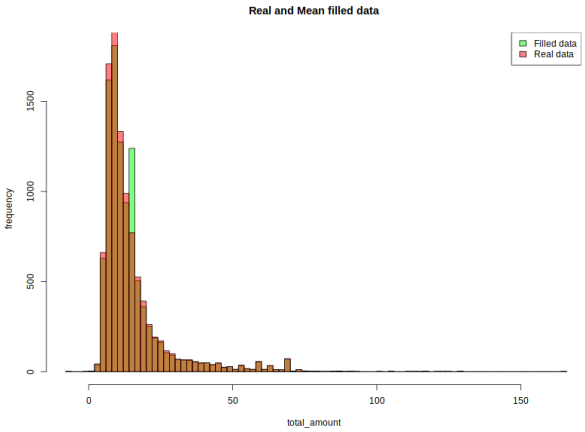


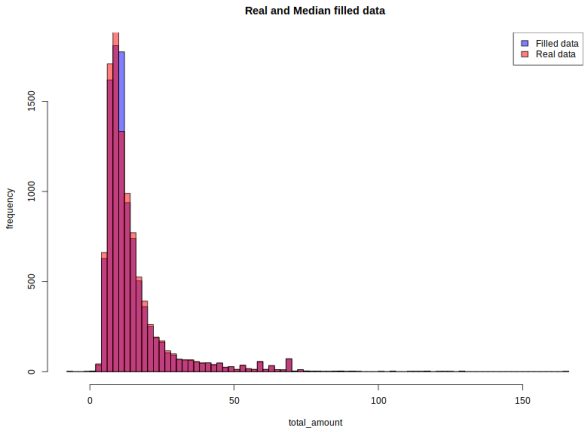
Рис. 9: Заполнение пропусков модой

7.2 Заполнение медианным значением. Заполнение средним значением

Теперь заполним пропуски в переменной total_amount (стоимость поездки) с помощью среднего значения и медианы. Медиана оказалась немного лучше, потому что среднее значение оказалось чувствительным к выбросам и могло искажать данные. Медиана, в свою очередь, не так подвержена влиянию аномальных значений, что сделало её более устойчивым выбором для заполнения пропусков в данном случае. На графике мы можем увидеть гистограммы данных до внесения пропусков и после заполнения.



(a) средним значением



(b) медианой

Рис. 10: Заполнение пропусков

8 Проверка на нормальность

8.1 1. Графики эмпирических функций распределения (ЭФР)

Эмпирическая функция распределения (ЭФР) — это кумулятивная частота наблюдений в выборке. Она строится путем упорядочивания данных и нанесения на график кумулятивной доли наблюдений, которые меньше или равны каждому значению. График ЭФР позволяет визуально оценить, насколько данные соответствуют теоретической функции распределения, например, нормальному распределению.

8.2 2. Графики квантилей (Q-Q plot)

График квантилей (Q-Q plot) — это графический метод, который сравнивает квантили эмпирического распределения с квантилями теоретического распределения. Для проверки нормальности данных на графике квантилей по оси X откладываются квантили нормального распределения, а по оси Y — квантили эмпирического распределения. Если данные распределены нормально, точки на графике должны лежать близко к прямой линии.

8.3 3. Метод огибающих (Envelope Method)

Метод огибающих — это графический метод, который сравнивает ЭФР с теоретической функцией распределения (ФР) нормального распределения. Для оценки близости ЭФР к ФР строятся огибающие, которые представляют собой доверительные интервалы вокруг ФР. Если ЭФР находится внутри этих огибающих, то данные считаются соответствующими нормальному распределению с заданным уровнем значимости.

8.4 4. Стандартные процедуры проверки гипотез о нормальности

8.4.1 Критерий Колмогорова-Смирнова

Критерий Колмогорова-Смирнова — это непараметрический критерий, который сравнивает эмпирическую функцию распределения с теоретической функцией распределения. Критерий основан на максимальном расстоянии между ЭФР и ФР. Если это расстояние мало, то данные могут считаться распределенными согласно теоретическому распределению.

8.4.2 Критерий Шапиро-Уилка

Критерий Шапиро-Уилка — это параметрический критерий, который проверяет нормальность распределения данных. Он основан на сравнении выборочных коэффициентов асимметрии и эксцесса с их теоретическими значениями для нормального распределения. Критерий Шапиро-Уилка особенно эффективен для небольших выборок.

8.4.3 Критерий Андерсона-Дарлинга

Критерий Андерсона-Дарлинга — это модификация критерия Колмогорова-Смирнова, которая более чувствительна к отклонениям от нормальности в хвостах распределения. Он часто используется в качестве альтернативы критерию Колмогорова-Смирнова, особенно когда важно проверить нормальность в хвостах распределения.

8.4.4 Критерий Крамера фон Мизеса

Критерий Крамера фон Мизеса — это непараметрический критерий, который сравнивает эмпирическую функцию распределения с теоретической функцией распределения. Он основан на квадрате разности между ЭФР и ФР, интегрированной по всему диапазону данных. Критерий Крамера фон Мизеса менее чувствителен к отклонениям в хвостах распределения по сравнению с критерием Колмогорова-Смирнова.

8.4.5 Критерий Колмогорова-Смирнова в модификации Лиллиефорса

Критерий Колмогорова-Смирнова в модификации Лиллиефорса — это модификация критерия Колмогорова-Смирнова, которая используется для проверки нормальности данных с учетом их стандартизации. Он особенно полезен для проверки нормальности данных, которые уже были стандартизированы.

8.4.6 Критерий Шапиро-Франсия

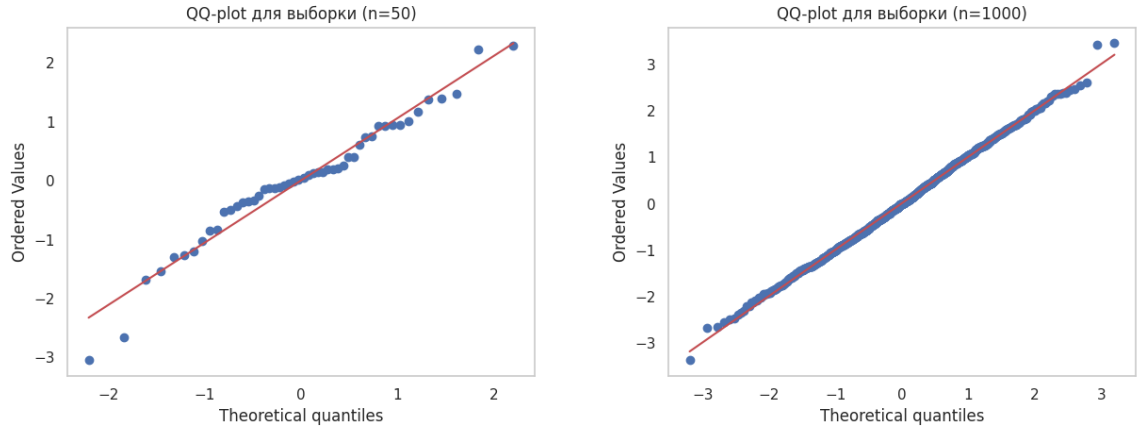
Критерий Шапиро-Франсия — это модификация критерия Шапиро-Уилка, которая используется для проверки нормальности данных в случае, когда выборка содержит повторяющиеся значения. Он более устойчив к наличию повторяющихся значений в данных по сравнению с критерием Шапиро-Уилка.

8.5 Заключение

Выбор метода проверки нормальности зависит от конкретных задач и особенностей данных. Графические методы, такие как графики ЭФР, Q-Q plot и метод огибающих, предоставляют наглядную информацию о соответствии данных нормальному распределению. Стандартные процедуры проверки гипотез, такие как критерии Колмогорова-Смирнова, Шапиро-Уилка, Андерсона-Дарлинга, Крамера фон Мизеса, Колмогорова-Смирнова в модификации Лиллиефорса и Шапиро-Франсия, предоставляют более строгие статистические выводы о нормальности распределения данных.

8.6 Применение к сгенерированным данным

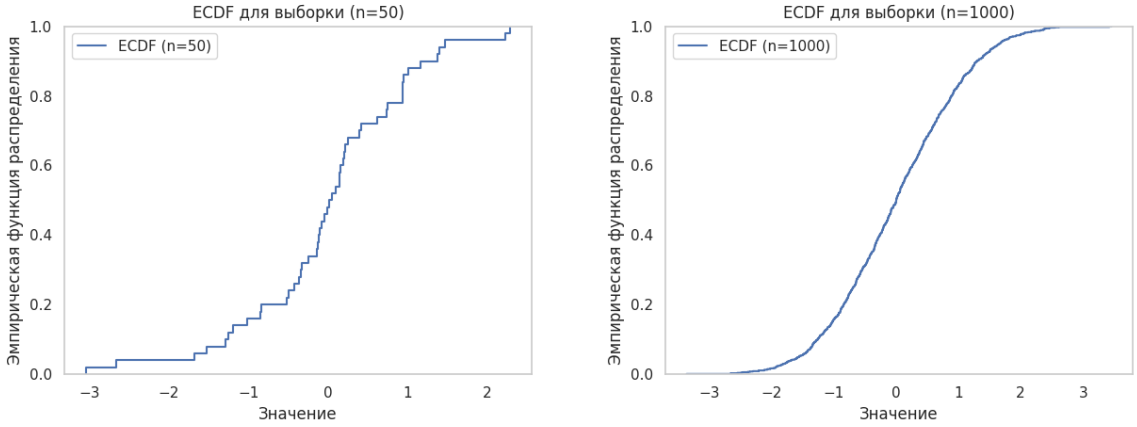
Применим наши методы к сгенерированным из стандартного нормального распределения данным. Маленькая выборка - 50 значений, умеренная - 1000.



(a) n=50

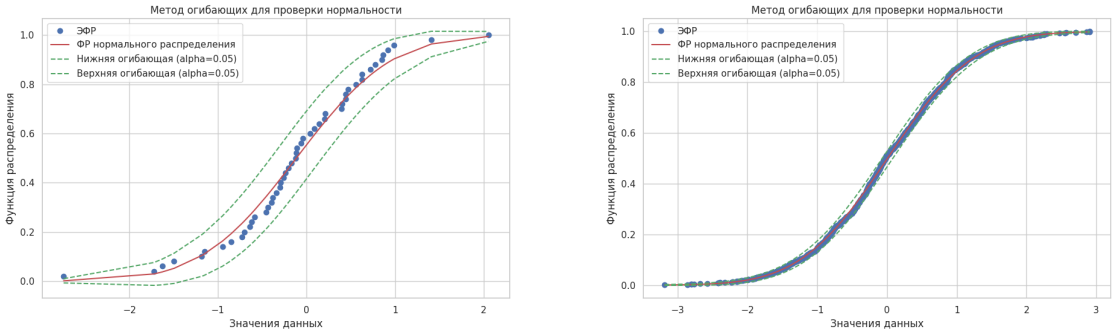
(b) n=1000

Рис. 11: QQ-plot



(a) n=50 (b) n=1000

Рис. 12: Эмпирическая функция распределения



(a) n=50 (b) n=1000

Рис. 13: Метод огибающих

Таблица 1: Результаты тестов на нормальность

Критерий	**Статистика**	**p-value**
Маленькая выборка (n=50)		
Критерий Колмогорова-Смирнова	0.08239	0.85877
Критерий Шапиро-Уилка	0.98396	0.72607
Критерий Андерсона-Дарлинга	0.29908	-
Критерий Крамера фон Мизеса	0.2616	0.1742
Критерий Лиллиефорса	0.07996	0.58085
Критерий Шапиро-Франция	-	0.7985
Средняя выборка (n=1000)		
Критерий Колмогорова-Смирнова	0.02004	0.80889
Критерий Шапиро-Уилка	0.99768	0.17164
Критерий Андерсона-Дарлинга	0.50713	-
Критерий Крамера фон Мизеса	0.1582	0.3656
Критерий Лиллиефорса	0.01993	0.51275
Критерий Шапиро-Франция	-	0.2145

Вывод:
По результатам всех критериев можем принять гипотезу, что наши данные из нормального распределения. Попробуем объяснить, почему разные критерии выдали различный уровень значимости.

Объяснение различий в p-value

- **Различия в чувствительности:**
  - **Критерий Шапиро-Уилка** и **Критерий Шапиро-Франция** могут давать более высокие p-value для небольших выборок, так как они более чувствительны к отклонениям в центральной части распределения.
  - **Критерий Андерсона-Дарлинга** и **Критерий Лиллиефорса** могут давать более низкие p-value, если данные имеют отклонения в хвостах распределения.
- **Размер выборки:**
  - Для маленькой выборки (n=50) **Критерий Шапиро-Уилка** и **Критерий Шапиро-Франция** могут быть более надежными, что отражается в более высоких p-value.
  - Для средней выборки (n=1000) **Критерий Колмогорова-Смирнова** и **Критерий Крамера фон Мизеса** могут быть более надежными, что отражается в более высоких p-value.

8.7 Применение к данным

Проверка на нормальность данных о количестве вызовов такси в день
У нас есть данные о количестве вызовов такси в день, размер выборки 29 строк, проверим данные на нормальность.

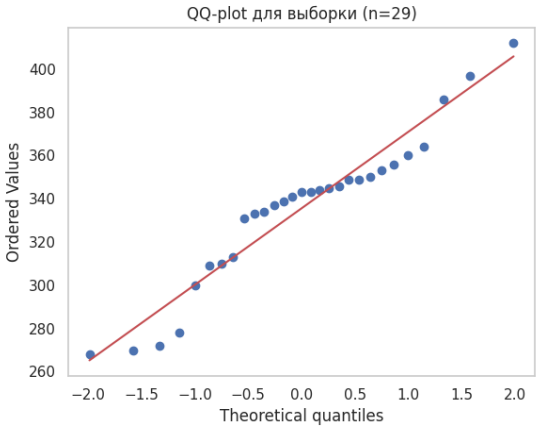


Рис. 14: QQ-plot

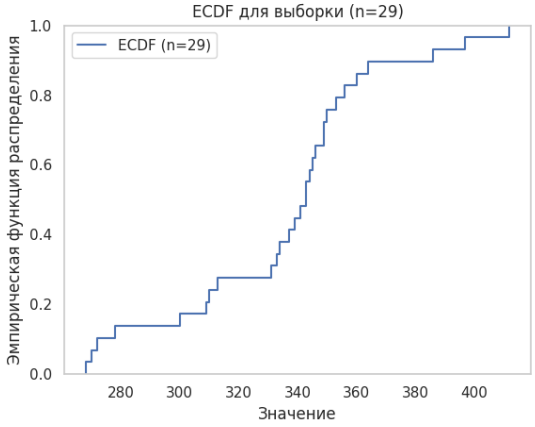


Рис. 15: Эмпирическая функция распределения

Выводы:
Критерий Колмогорова-Смирнова, Критерий Шапиро-Уилка и Критерий Шапиро-Франция не отвергают гипотезу о нормальности распределения данных, так как их p-value больше 0.05.

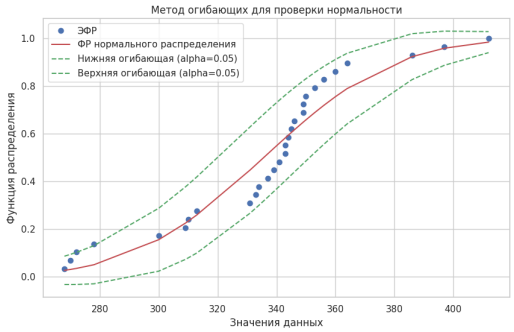


Рис. 16: Метод огибающих

Критерий Андерсона-Дарлинга также указывает на нормальность распределения, так как значение статистики мало.

Критерий Крамера фон Мизеса и Критерий Лиллиефорса отвергают гипотезу о нормальности распределения данных, так как их p-value меньше 0.05.

Таким образом, результаты тестов на нормальность показывают, что данные по количеству вызовов такси в день не совсем соответствуют нормальному распределению, особенно с учетом результатов критериев Крамера фон Мизеса и Лиллиефорса. Однако, учитывая, что некоторые критерии не отвергают гипотезу о нормальности, можно предположить, что данные близки к нормальному распределению, но имеют некоторые отклонения. По графикам можно сделать такой же вывод.

Таблица 2: Результаты тестов на нормальность для маленькой выборки (n = 29)

Критерий	**Статистика**	**p-value**
Критерий Колмогорова-Смирнова	0.17137	0.32406
Критерий Шапиро-Уилка	0.93757	0.08661
Критерий Андерсона-Дарлинга	0.85025	-
Критерий Крамера фон Мизеса	9.6667	0.0000
Критерий Лиллиефорса	0.17228	0.02782
Критерий Шапиро-Франция	-	0.0843

Проверка на нормальность данных о стоимости поездки такси

Проверим на выборке размера 1000, в которой содержится информаиця о стоимости поездки в такси, является ли она нормальной.

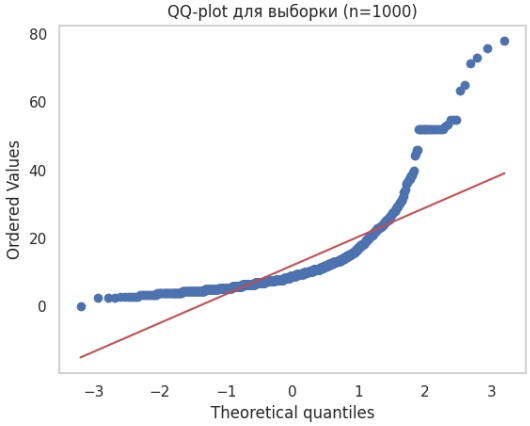


Рис. 17: QQ-plot

Выводы:

По всем результатам очевидно, что выборка не может быть из нормального распределения.

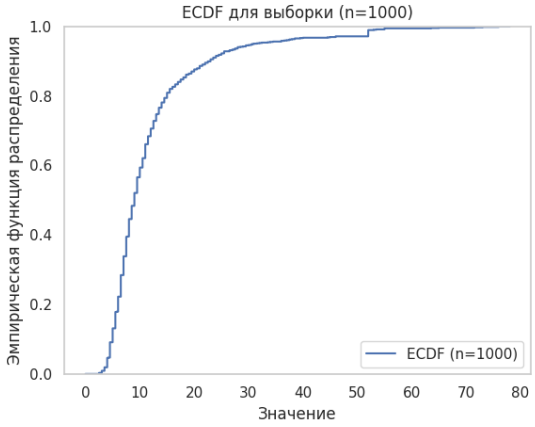


Рис. 18: Эмпирическая функция распределения

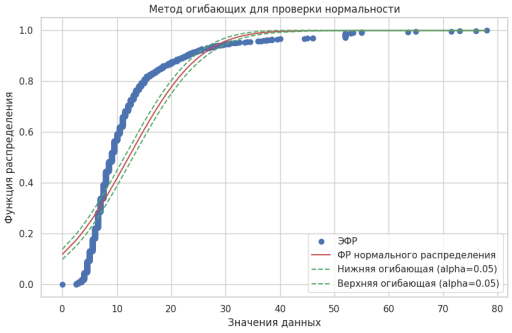


Рис. 19: Метод огибающих

Таблица 3: Результаты тестов на нормальность для средней выборки (n = 1000)

Критерий	**Статистика**	**p-value**
Критерий Колмогорова-Смирнова	0.21424	9.18986e-41
Критерий Шапиро-Уилка	0.67128	3.84628e-40
Критерий Андерсона-Дарлинга	94.09886	-
Критерий Крамера фон Мизеса	332.5241	0.0000
Критерий Лиллиефорса	0.21425	0.0010
Критерий Шапиро-Франция	-	0.0000

9 Критерии проверки гипотез

Критерий Стьюдента

Критерий Стьюдента используется для проверки гипотез о равенстве средних значений двух выборок. Он предполагает, что данные распределены нормально, а дисперсии выборок равны (в случае независимых выборок). Критерий может быть двусторонним и односторонним. Для малых выборок (обычно $n < 30$) особенно важно выполнение предположения о нормальности распределения.
Тип: Не является ранговым.
Требования к данным: Нормальное распределение, равенство дисперсий для независимых выборок.

Критерий Уилкоксона-Манна-Уитни

Критерий Уилкоксона-Манна-Уитни — это непараметрический тест для проверки гипотезы о равенстве двух медиан. Он используется, когда данные не удовлетворяют предположениям нормальности. Сравниваются ранги элементов двух выборок, а не сами значения.
Тип: Ранговый.
Требования к данным: Отсутствие сильных выбросов, данные могут быть не нормальными.

Критерий Фишера

Критерий Фишера используется для проверки гипотезы об однородности дисперсий двух выборок. Он основан на отношении дисперсий и требует нормального распределения данных.
Тип: Не является ранговым.
Требования к данным: Нормальное распределение.

Критерий Левене

Критерий Левене применяется для проверки равенства дисперсий нескольких выборок. Он менее чувствителен к отклонениям от нормальности по сравнению с критерием Фишера.
Тип: Не является ранговым.
Требования к данным: Допускается отклонение от нормальности.

Критерий Бартлетта

Критерий Бартлетта используется для проверки гипотезы об однородности дисперсий нескольких выборок. Чувствителен к отклонению от нормального распределения, поэтому может быть ненадежным на данных с сильной асимметрией.
Тип: Не является ранговым.
Требования к данным: Нормальное распределение.

Критерий Флигнера-Килина

Критерий Флигнера-Килина — это непараметрический тест, который проверяет равенство дисперсий нескольких выборок. Он основывается на рангах данных.
Тип: Ранговый.
Требования к данным: Отсутствие сильных выбросов, данные могут быть не нормальными.

Ранговые критерии

Ранговые критерии — это методы статистической проверки гипотез, которые используют порядковые ранги вместо исходных числовых значений. Такие критерии применяются, когда данные не соответствуют предположениям параметрических тестов (например, нормальности). Ранговые методы устойчивы к выбросам и подходят для анализа данных, измеренных на порядковой шкале.

10 Проверка гипотез на данных о количестве вызовов такси в день

Рассматриваем данные о количестве вызовов такси в день, всего 31 строка. Данные разделены на две группы: количество вызовов такси в день до 15 числа месяца и после 15 числа месяца.

Проверка на нормальность

Результаты теста Колмогорова-Смирнова:
Статистика = 0.2324, p-значение = 0.0590

Тест Колмогорова-Смирнова показал p-значение 0.059, что больше уровня значимости $\alpha = 0.05$. Это говорит о том, что мы не отвергаем гипотезу о нормальности распределения данных. Однако, p-значение близко к пороговому значению, что может свидетельствовать о легком отклонении от нормальности.

10.1 Критерий Стьюдента

Двусторонний тест: (средние значения равны)
t-статистика = 0.49, p-значение = 0.6262

Односторонний тест: (среднее значение 1-й группы $\leq$ среднего значения другой)
t-статистика = 0.49, p-значение = 0.6869

Выводы для различных доверительных уровней:

- **Доверительный уровень: 0.9**
  Не отвергаем нулевую гипотезу (для обоих тестов).
- **Доверительный уровень: 0.96**
  Не отвергаем нулевую гипотезу (для обоих тестов).
- **Доверительный уровень: 0.99**
  Не отвергаем нулевую гипотезу (для обоих тестов).

Двусторонний тест и **односторонний тест** не отвергают нулевую гипотезу при всех доверительных уровнях. Это означает, что статистически значимой разницы между средними числами вызовов такси до 15 числа и после 15 числа месяца нет.

10.2 Критерий Уилкоксона-Манна-Уитни

Результаты:
Статистика = 93.00, p-значение = 0.2948

Выводы для различных доверительных уровней:

- **Доверительный уровень: 0.9**
  Не отвергаем нулевую гипотезу.
- **Доверительный уровень: 0.96**
  Не отвергаем нулевую гипотезу.
- **Доверительный уровень: 0.99**
  Не отвергаем нулевую гипотезу.

Этот критерий также не выявил статистически значимой разницы между группами при всех доверительных уровнях. Поскольку критерий является ранговым, он менее чувствителен к отклонениям от нормальности, но его результаты согласуются с тестом Стьюдента. Это говорит о том, что различия между группами действительно минимальны.

10.3 Проверка равенства дисперсий

Критерий Левена:
Статистика = 1.01, p-значение = 0.3231
Критерий Бартлетта:
Статистика = 8.15, p-значение = 0.0043
Критерий Флигнера-Килина:
Статистика = 0.02, p-значение = 0.8792

Выводы для различных доверительных уровней:

- **Доверительный уровень: 0.9**
  - Критерий Левена: не отвергаем нулевую гипотезу.
  - Критерий Бартлетта: отвергаем нулевую гипотезу.
  - Критерий Флигнера-Килина: не отвергаем нулевую гипотезу.
- **Доверительный уровень: 0.96**

- Критерий Левена: не отвергаем нулевую гипотезу.
- Критерий Бартлетта: отвергаем нулевую гипотезу.
- Критерий Флигнера-Килина: не отвергаем нулевую гипотезу.

- **Доверительный уровень: 0.99**
  - Критерий Левена: не отвергаем нулевую гипотезу.
  - Критерий Бартлетта: отвергаем нулевую гипотезу.
  - Критерий Флигнера-Килина: не отвергаем нулевую гипотезу.

Анализ результатов

- **Критерий Левена:** не отвергает нулевую гипотезу, что означает, что дисперсии групп можно считать равными. Это подтверждает предположение, необходимое для использования критерия Стьюдента.
- **Критерий Бартлетта:** отвергает нулевую гипотезу (p-значение 0.0043), указывая на возможное различие дисперсий. Однако он чувствителен к отклонениям от нормального распределения, поэтому его результат может быть связан с легкой ненормальностью данных.
- **Критерий Флигнера-Килина:** показывает, что различия в дисперсиях статистически незначимы. Это подтверждает выводы критерия Левена и указывает на устойчивость результатов.

Вывод Проведенные тесты не выявили статистически значимой разницы между группами или дисперсиями.

10.4 Мощность критерия Стьюдента и расчет объема выборки

10.4.1 Мощность критерия Стьюдента

Рассчитанная мощность критерия Стьюдента для текущей выборки составила 0.08. Это означает, что вероятность правильно отвергнуть нулевую гипотезу при наличии реального различия между средними значениями крайне низка. Такая мощность недостаточна для уверенных статистических выводов.

Причины низкой мощности:

- Небольшой объем выборки, что снижает вероятность обнаружения реального различия.
- Возможность того, что реальное различие между группами минимально.

10.4.2 Расчет необходимого объема выборки

Для достижения стандартной статистической мощности 0.8 (80%), необходимый объем выборки составляет $n = 509$ наблюдений. Увеличение объема выборки до 509 позволит существенно повысить вероятность обнаружения реального различия между средними значениями, если оно существует.
Выводы:
Текущая выборка недостаточна для получения статистически обоснованных выводов из теста Стьюдента, текущие результаты теста Стьюдента следует интерпретировать с осторожностью, так как существует высокая вероятность ошибки второго рода (пропустить существующее различие), но мы сделали несколько тестов, которые показали такой же результат, поэтому можно сделать вывод, что и при такой мощности этому результату можно доверять.

11 Исследование корреляционных взаимосвязей

11.1 Коэффициент корреляции Пирсона

Описание:

- Коэффициент корреляции Пирсона измеряет **линейную** взаимосвязь между двумя переменными.

- Принимает значения от -1 (идеальная отрицательная линейная связь) до 1 (идеальная положительная линейная связь). Значение 0 указывает на отсутствие линейной зависимости.

Требования к данным:

- Переменные должны быть количественными (интервальные или шкалы отношений).
- Данные должны быть нормально распределены (желательно, но не строго обязательно для больших выборок). .

11.2 Коэффициент корреляции Спирмена

Описание:

- Коэффициент корреляции Спирмена измеряет **монотонную** связь между переменными, основанную на ранговой зависимости.
- Принимает значения от -1 (идеальная отрицательная монотонная связь) до 1 (идеальная положительная монотонная связь). Значение 0 указывает на отсутствие монотонной зависимости.
- Устойчив к выбросам и нелинейным зависимостям.

Требования к данным:

- Переменные могут быть количественными или порядковыми.
- Распределение данных может быть любым (нет требований к нормальности).
- Подходит для нелинейных монотонных зависимостей.

11.3 Коэффициент корреляции Кендалла (тау)

Описание:

- Коэффициент Кендалла измеряет силу **монотонной** связи между переменными, основываясь на сравнении пар значений.
- Принимает значения от -1 до 1. Чем ближе значение к 1, тем сильнее положительная монотонная связь; чем ближе к -1, тем сильнее отрицательная монотонная связь.
- Менее чувствителен к выбросам по сравнению с коэффициентом Спирмена.

Требования к данным:

- Переменные могут быть количественными или порядковыми.
- Данные должны быть монотонно зависимыми (не обязательно линейными).
- Устойчив к выбросам, но требует минимального объема данных.

12 Коэффициенты корреляции между переменными

Используем коэффициенты корреляции для исследования взаимосвязей в сгенерированных данных.

- X, Z - независимые, из нормального распределения,
- Y линейно зависит от X (по формуле $Y = 2X + \text{шум}$),
- W зависит от X экспоненциально (по формуле $W = \exp(X) + \text{шум}$).

1. Коэффициенты корреляции между X и Y:

$$\text{Пирсон: } r = 0.97, \quad \text{Спирмен: } \rho = 0.97, \quad \text{Кендалл: } \tau = 0.84$$

Так как Y линейно зависит от X, это приводит к очень сильной положительной линейной связи между переменными X и Y, что подтверждается высокими значениями всех коэффициентов корреляции (особенно по Пирсону и Спирмену).

2. Коэффициенты корреляции между X и Z:

$$\text{Пирсон: } r = -0.13, \quad \text{Спирмен: } \rho = -0.13, \quad \text{Кендалл: } \tau = -0.09$$

Z генерируется независимо от X, и корреляции между ними должны быть близки к нулю. Как мы и видим, значения всех коэффициентов близки к нулю, что подтверждает отсутствие значимой линейной и монотонной зависимости.

3. Коэффициенты корреляции между X и W:

$$\text{Пирсон: } r = 0.80, \quad \text{Спирмен: } \rho = 0.96, \quad \text{Кендалл: } \tau = 0.84$$

W зависит от X экспоненциально, что означает, что зависимость между этими переменными не является линейной. Однако, так как экспоненциальная зависимость монотонная, методы Спирмена и Кендалла показывают сильную монотонную связь (высокие значения корреляции), в то время как Пирсон (метод для линейной зависимости) имеет несколько более низкое значение (0.80), так как экспоненциальная зависимость не является строго линейной.

4. Коэффициенты корреляции между Y и Z:

$$\text{Пирсон: } r = -0.13, \quad \text{Спирмен: } \rho = -0.14, \quad \text{Кендалл: } \tau = -0.10$$

Поскольку Y зависит от X, а Z генерируется независимо, корреляция между Y и Z должна быть очень слабой или отсутствовать. Эти значения подтверждают это: все коэффициенты близки к нулю, что свидетельствует об отсутствии значимой зависимости.

5. Коэффициенты корреляции между Y и W:

$$\text{Пирсон: } r = 0.76, \quad \text{Спирмен: } \rho = 0.92, \quad \text{Кендалл: } \tau = 0.76$$

Поскольку Y и W оба зависят от X, но Y — линейно, а W — экспоненциально, мы видим умеренную корреляцию между этими переменными. Спирмен и Кендалл дают более высокие значения, так как зависимость монотонная, но не линейная. Пирсон, как линейный метод, оценивает зависимость ниже (0.76), что подтверждает, что зависимость между Y и W не является строго линейной.

6. Коэффициенты корреляции между Z и W:

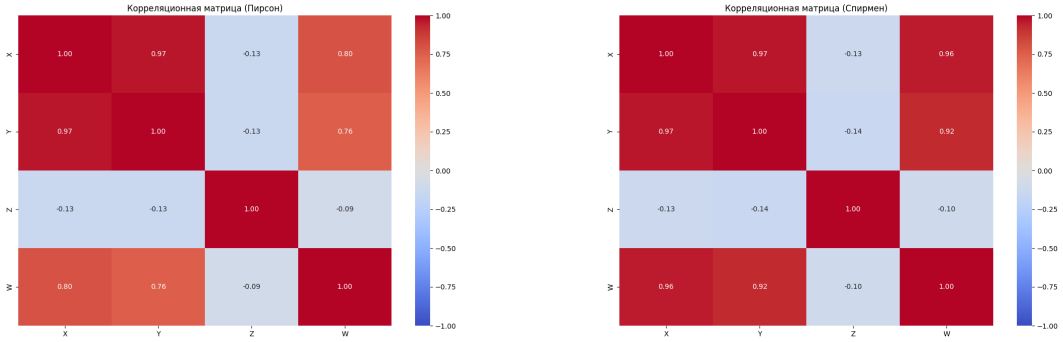
$$\text{Пирсон: } r = -0.09, \quad \text{Спирмен: } \rho = -0.10, \quad \text{Кендалл: } \tau = -0.07$$

Поскольку Z и W независимы, мы ожидаем, что корреляция между ними будет близка к нулю. Эти результаты подтверждают, что зависимости между Z и W нет.

Выводы: - Линейная зависимость между X и Y (очень высокие коэффициенты корреляции). - Монотонная зависимость между X и W, которая не является линейной, объясняет высокие значения для методов Спирмена и Кендалла, но относительно более низкое значение для Пирсона. - Отсутствие зависимости между независимыми переменными, такими как X и Z, или Y и Z, подтверждается низкими коэффициентами корреляции.

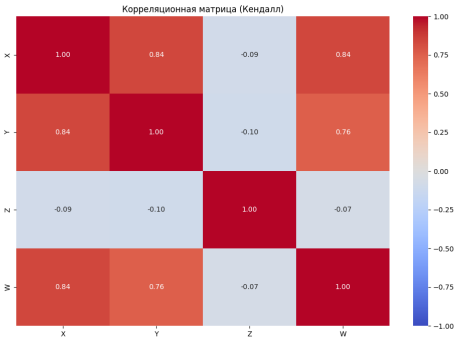
13 Исследование корреляций на данных о такси

В этом разделе рассмотрен анализ таблицы корреляций Кендалла для набора данных, связанного с поездками на такси. Корреляционная матрица Кендалла позволяет выявить монотонные зависимости между переменными и оценить степень их взаимосвязи.



(a) Пирсон

(b) Спирмен



(c) Кендалл

Рис. 20: Корреляционные матрицы для сгенерированных данных

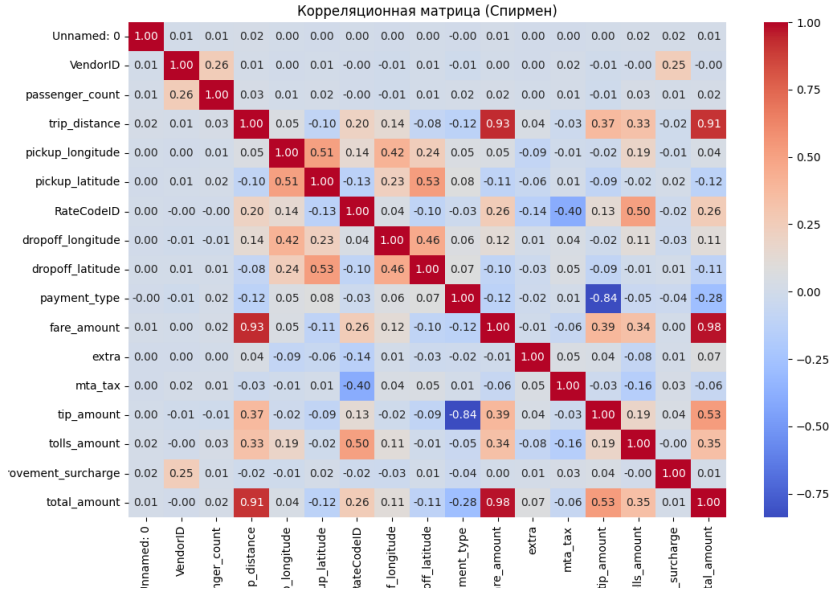


Рис. 21: Корреляционная матрица (Спирмен)

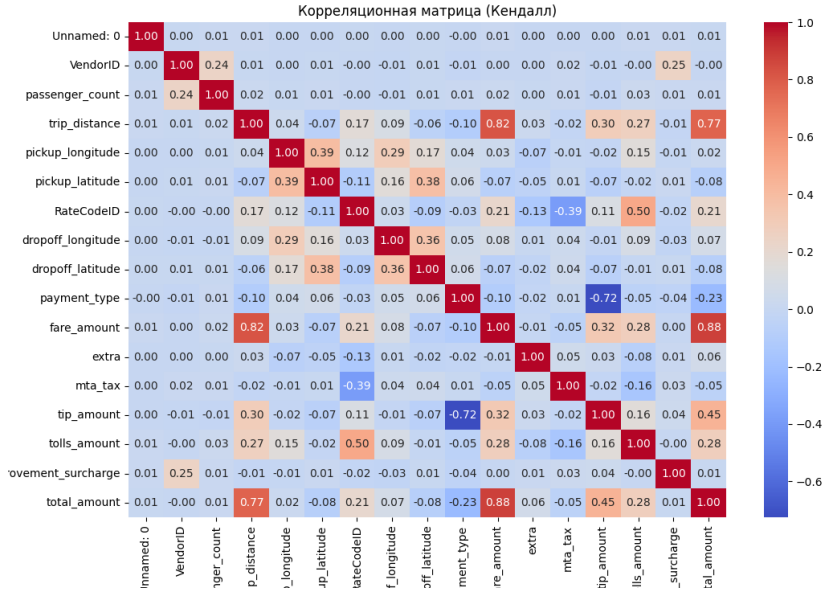


Рис. 22: Корреляционная матрица (Кендалл)

13.1 Переменные с сильной корреляцией

- **fare_amount** и **total_amount** (корреляция 0.88): Эти переменные имеют сильную положительную монотонную зависимость, что логично, так как `total_amount` вряд ли может быть значительно выше, чем `fare_amount`, если не учитывать дополнительные сборы, такие как чаевые, налоги и другие.

- **trip_distance** и **fare_amount** (корреляция 0.82): Очевидно, что длина поездки будет прямо пропорциональна стоимости поездки. Чем больше расстояние, тем выше стоимость.

- **tip_amount** и **total_amount** (корреляция 0.45): Чаевые имеют умеренную положительную зависимость с общей суммой, так как больше чаевых обычно означает более высокая общая сумма.

- **RateCodeID** и **tolls_amount** (корреляция 0.50): Это может отражать то, что для разных категорий поездок (RateCodeID) могут быть разные сборы за проезд по платным дорогам (tolls).

13.2 Переменные с низкой или отсутствующей корреляцией

- **VendorID** и **fare_amount** (корреляция 0.00): Здесь нет значимой корреляции, что подтверждает независимость этих переменных.
- **mta_tax** и **fare_amount** (корреляция -0.05): Минимальная отрицательная корреляция, что также говорит об отсутствии сильной зависимости.
- **pickup_latitude** и **dropoff_latitude** (корреляция 0.38): Эти переменные имеют небольшую положительную корреляцию, что указывает на то, что в некоторых случаях местоположение посадки и высадки могут быть близкими, но это не является жестким правилом.

13.3 Некоторые интересные зависимости

- **tip_amount** и **payment_type** (корреляция -0.72): Отрицательная корреляция, так как при оплате наличными чаевые не указываютя - то есть считаются ранвыми 0.
- **pickup_longitude** и **pickup_latitude** (корреляция 0.39): Логично ожидать, что координаты посадки будут в некотором роде коррелировать, поскольку они отражают географическое местоположение.

13.4 Предсказания по данным корреляциям

- На основе высокой корреляции между переменными **trip_distance** и **fare_amount** можно сделать вывод, что стоимость поездки будет пропорциональна пройденному расстоянию.
- Также сильная корреляция между **fare_amount** и **total_amount** подтверждает, что общая сумма всегда будет включать стоимость поездки как основной компонент.

13.5 Низкая корреляция между не связанными переменными

- **fare_amount** и **mta_tax** (-0.05), **fare_amount** и **payment_type** (-0.10) показывают низкие корреляции, что указывает на то, что эти переменные вряд ли влияют друг на друга напрямую.

Вывод: Корреляционная матрица Кендалла позволяет выявить как сильные, так и слабые взаимосвязи между переменными, а также понять, какие переменные могут быть полезными для построения модели, например, для прогнозирования стоимости поездки или общей суммы.

14 Методы статистических тестов для категориальных данных

14.1 Критерий хи-квадрат χ^2

Описание: Критерий хи-квадрат используется для проверки гипотезы о том, что наблюдаемые данные соответствуют некоторому теоретическому распределению. Применим его в качестве теста на независимость, проверим, независимы ли две категориальные переменные.
Требования к данным:

- Данные должны быть категориальными (например, в виде таблицы сопряженности).
- Каждая ячейка таблицы должна содержать количество наблюдений, а не доли или проценты.
- Ожидаемое количество наблюдений в каждой ячейке должно быть не менее 5.

14.2 Точный тест Фишера

Описание: Точный тест Фишера — это статистический метод для проверки гипотезы о независимости между двумя категориальными переменными, когда размер выборки маленький.
Требования к данным:

- Данные должны быть категориальными.
- Подходит для небольших таблиц (обычно таблица 2×2).
- Предполагается, что данные независимы.

14.3 Тест МакНемара

Описание: Тест МакНемара используется для анализа парных данных, когда необходимо проверить, изменилось ли распределение признака до и после воздействия на одну и ту же группу.
Требования к данным:

- Данные должны быть парными (например, измерения до и после воздействия на одних и тех же объектах).
- Переменные должны быть категориальными (обычно бинарными).
- Необходима таблица 2×2 с изменениями до и после.

14.4 Тест Кохрана-Мантеля-Хензеля

Описание: Тест Кохрана-Мантеля-Хензеля используется для проверки взаимосвязи между переменными, когда есть несколько групп и нужно учитывать один или несколько дополнительных факторов (например, ковариат). Это обобщение теста хи-квадрат для стратифицированных данных.
Требования к данным:

- Данные должны быть категориальными.
- Необходимо, чтобы данные были разделены на несколько групп или стратифицированы по какому-либо признаку.
- Требуются стратифицированные таблицы.

Когда использовать:

- Когда данные имеют несколько групп и нужно проверить зависимость между переменными, контролируя возможные дополнительные факторы.
- Когда необходимо контролировать влияние одного или нескольких факторов на зависимость между двумя переменными.

15 Анализ результатов статистических тестов

Хи-квадрат тест 1 Исследуем связь между категориальными переменными *Sleep_ Quality* (Качество сна) и *Health_ Risk_ Level* (Уровень риска для здоровья).

- **Результаты:** $\chi^2 = 50.13$, p-value $= 0.0000$, $dof = 4$.
- **Вывод:** Связь между переменными значима, так как p-value < 0.05. Это означает, что существует статистически значимая связь между качеством сна студентов и их уровнем риска для здоровья.

Хи-квадрат тест 2 Исследуем связь между категориальными переменными *Gender* (Пол) и *Health_ Risk_ Level* (Уровень риска для здоровья).

- **Результаты:** $\chi^2 = 0.80$, p-value $= 0.6719$, $dof = 2$.

- **Вывод:** Нет значимой связи между переменными, так как p-value > 0.05. Это означает, что пол студентов не оказывает статистически значимого влияния на их уровень риска для здоровья.

Точный тест Фишера 1 Исследуем связь между категориальными переменными *Mood* (Настроение) и *Gender* (Пол).

- **Результаты:** p-value $= 0.2497$.
- **Вывод:** Нет значимой связи между *Mood* (Настроение) и *Gender* (Пол), так как p-value > 0.05. Это говорит о том, что настроение студентов не связано с их полом.

Точный тест Фишера 2 (для маленькой выборки) Исследуем связь между категориальными переменными *Mood* (Настроение) и *Gender* (Пол) в маленькой выборке (n = 100)

- **Результаты:** p-value $= 0.6686$.
- **Вывод:** Нет значимой связи между переменными в выборке с малым числом данных, так как p-value > 0.05. Это подтверждает отсутствие связи между настроением и полом в данной группе студентов.

Точный тест Фишера 3 (для маленькой выборки) Исследуем связь между категориальными переменными *Sleep_ Quality* (Качество сна) и *Health_ Risk_ Level* (Уровень риска для здоровья) в маленькой выборке (n = 100)

- **Результаты:** p-value $= 0.0300$.
- **Вывод:** Связь между переменными значима, так как p-value < 0.05. Это означает, что в данной выборке существует статистически значимая зависимость между качеством сна и уровнем риска для здоровья студентов.

Тест МакНемара 1 Исследуем связь между признаками *Stress_ Level_ Biosensor* (Уровень стресса с использованием биосенсора - числовой) и *HRL_ Bin* (Уровень риска для здоровья — бинарный)

- **Результаты:** p-value $= 0.0000$.
- **Вывод:** Признаки зависимы. p-value $= 0.0000$ указывает на высокую вероятность того, что стресс, измеренный с помощью биосенсора, связан с уровнем риска для здоровья студентов.

Тест МакНемара 2 Признаки *Study_ Hours* (Часы учебы) и *Project_ Hours* (Часы работы над проектом) - числовые. Рассмотрим связь между ними.

- **Результаты:** p-value $= 0.3094$.
- **Вывод:** Признаки независимы, так как p-value > 0.05. Это означает, что количество часов, которые студенты тратят на учебу, не связано с количеством времени, которое они посвящают работе над проектами.

Тест Кохрана-Мантеля-Хензеля Исследуем связь между признаками *SLB_ Bin* (Стрессовый уровень в бинарной форме) и *HRL_ Bin* (Уровень риска для здоровья — бинарный), с учетом переменной *Mood* (Настроение).

- **Результаты:** p-value $= 0.0000$.
- **Вывод:** Существует значимая связь между *SLB_ Bin* (Стрессовый уровень в бинарной форме) и *HRL_ Bin* (Уровень риска для здоровья — бинарный), с учетом переменной *Mood* (Настроение). Это говорит о том, что настроение оказывает влияние на зависимость между уровнем стресса и уровнем риска для здоровья студентов.

16 Методы проверки мультиколлинеарности

Мультиколлинеарность возникает, когда в данных существует высокая корреляция между независимыми переменными, что может искажать результаты регрессионного анализа. Для её обнаружения применяются следующие методы:

16.1 Корреляционная матрица

Описание: Корреляционная матрица показывает парные корреляции между всеми независимыми переменными в наборе данных. Если коэффициенты корреляции между переменными приближаются к ± 1, это может свидетельствовать о наличии мультиколлинеарности.

Ограничения:

- Корреляционная матрица выявляет только парную корреляцию и не учитывает взаимосвязи между несколькими переменными одновременно.
- Может недооценивать мультиколлинеарность, вызванную взаимодействием нескольких переменных.

16.2 Фактор инфляции дисперсии (VIF — Variance Inflation Factor)

Описание: VIF измеряет, насколько дисперсия коэффициента регрессии переменной увеличивается из-за мультиколлинеарности. Если VIF высок, это свидетельствует о сильной зависимости между переменной и другими предикторами.

Формула:

$$VIF_i = \frac{1}{1 - R_i^2},$$

где R_i^2 — коэффициент детерминации при регрессии i-й переменной на все остальные независимые переменные.

Преимущества:

- Учитывает множественную корреляцию между переменными, а не только парные зависимости.
- Позволяет количественно оценить влияние мультиколлинеарности на каждый предиктор.

Интерпретация VIF:

- $VIF = 1$: отсутствие корреляции между переменной и другими.
- $1 < VIF \leq 5$: допустимая корреляция, мультиколлинеарность незначительна.
- $VIF > 5$: сильная мультиколлинеарность, которая может исказить результаты регрессии.
- $VIF > 10$: критический уровень мультиколлинеарности, требуется пересмотр модели.

Ограничения:

- VIF нельзя использовать для категориальных переменных, если они не представлены в виде фиктивных переменных.
- Требует, чтобы данные подходили для линейной регрессии.

17 Корреляционная матрица и Анализ VIF (Фактора инфляции дисперсии)

В таблице представлены значения VIF для каждой переменной, которые используются для оценки степени мультиколлинеарности. Рассмотрим интерпретацию и выводы на основе полученных значений.

17.1 Общие наблюдения

- **VIF $<$ 5**: Незначительная мультиколлинеарность.
  - Пример: `passenger_count` (VIF = 2.85), `Unnamed: 0` (VIF = 4.03).
- **5 VIF 10**: Умеренная мультиколлинеарность.
  - Пример: `VendorID` (VIF = 11.96), `trip_distance` (VIF = 12.02).

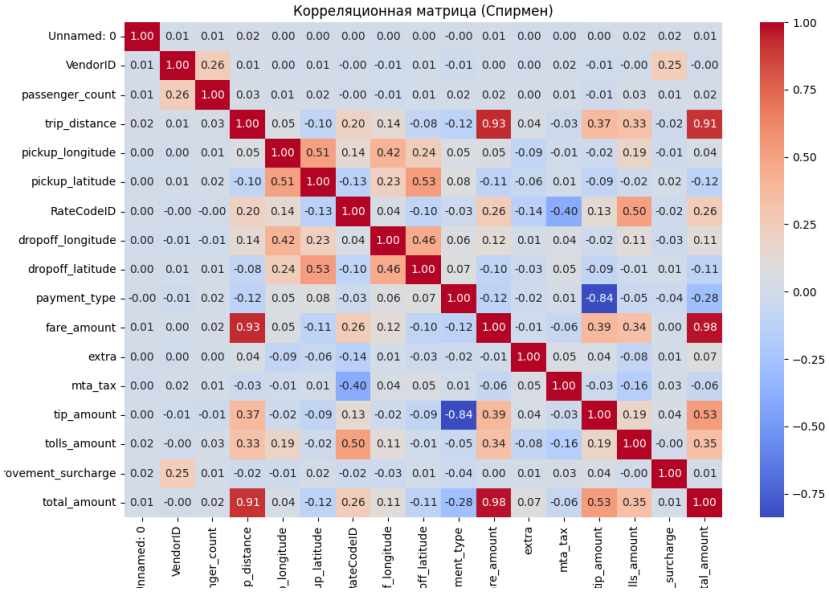


Рис. 23: Корреляционная матрица (Спирмен)

Переменная (feature)	**VIF**
Unnamed: 0	4.03
VendorID	11.96
passenger_count	2.85
trip_distance	12.02
pickup_longitude	2,121,007.00
pickup_latitude	2,120,637.00
RateCodeID	21.18
dropoff_longitude	1,409,546.00
dropoff_latitude	1,409,753.00
payment_type	14.79
fare_amount	158,670.00
extra	145.97
mta_tax	400.83
tip_amount	4,732.71
tolls_amount	1,128.68
improvement_surcharge	17.38
total_amount	243,727.20

Таблица 4: Значения фактора инфляции дисперсии (VIF) для переменных

- **VIF $>$ 10**: Сильная мультиколлинеарность, требует внимания.
  - Пример: большинство переменных, такие как `pickup_longitude` (VIF = 2,121,007.00), `total_amount` (VIF = 243,727.20).

17.2 Ключевые переменные с высоким VIF

- `pickup_longitude` **и** `pickup_latitude` **(VIF $>$ 2,120,000)**: Очень высокая мультиколлинеарность, скорее всего, из-за сильной взаимосвязи между координатами (например, географической близости).
- `dropoff_longitude` **и** `dropoff_latitude` **(VIF $>$ 1,409,000)**: Аналогичная ситуация с координатами места высадки.

- `fare_amount` и `total_amount` **(VIF $>$ 158,670 и 243,727 соответственно)**: Эти переменные, вероятно, включают общие компоненты (например, `fare_amount` входит в `total_amount`), что вызывает высокую мультиколлинеарность.
- `mta_tax`, `extra`, `tip_amount`, `tolls_amount`: Значения VIF выше 100 для этих переменных свидетельствуют о мультиколлинеарности, вызванной их прямыми или косвенными взаимосвязями с общей суммой (`total_amount`).

17.3 Вывод

Мультиколлинеарность в данных выражена очень сильно, особенно в переменных, связанных с географическими координатами и оплатой.

18 Дисперсионный анализ (ANOVA)

Дисперсионный анализ (ANOVA, *Analysis of Variance*) используется для определения наличия статистически значимых различий между средними значениями групп. Существуют два основных типа ANOVA: однофакторный и многофакторный.

18.1 Однофакторный ANOVA

Описание: Однофакторный ANOVA исследует влияние одной независимой переменной (фактора) на одну зависимую переменную. Используется, когда есть несколько групп, и необходимо проверить, различаются ли их средние значения.
Требования к данным:

- **Независимость данных:** Наблюдения в группах должны быть независимыми.
- **Нормальность распределения:** Зависимая переменная должна быть нормально распределена в каждой группе.
- **Гомогенность дисперсий:** Дисперсии зависимой переменной в группах должны быть примерно равными (проверяется тестом Левена или Бартлетта).

Методика:

1. Вычисляется общее среднее значение по всем данным.
2. Рассчитываются:
   - **Межгрупповая дисперсия:** отражает вариацию между средними групп.
   - **Внутригрупповая дисперсия:** отражает вариацию внутри групп.
3. Сравнивается отношение межгрупповой дисперсии к внутригрупповой (F-статистика).
4. Проверяется гипотеза:
   - **Нулевая гипотеза (H_0):** Средние значения групп равны.
   - **Альтернативная гипотеза (H_a):** Средние значения хотя бы одной из групп отличаются.

18.2 Многофакторный ANOVA

Описание: Многофакторный ANOVA используется для анализа влияния нескольких независимых переменных (факторов) на одну зависимую переменную, а также для изучения взаимодействия между факторами.
Требования к данным:

- Независимость данных.
- Нормальность распределения зависимой переменной.

- Гомогенность дисперсий.

Методика:

1. Аналогично однофакторному ANOVA, рассчитываются меж- и внутригрупповые дисперсии, но отдельно для каждого фактора.
2. Учитывается взаимодействие между факторами (*interaction effect*).
3. Проверяются гипотезы:
   - H_0 **для каждого фактора:** У него нет влияния на зависимую переменную.
   - H_0 **для взаимодействия:** Нет взаимодействия между факторами.

19 Применение дисперсионного анализа

Рассмотрим влияние различных факторов на переменную - `Heart_Rate` (Уровень сердечного ритма)

19.1 Проверка условий для применения ANOVA

- **Тест Колмогорова-Смирнова**
  Значение $p = 0.9122$ указывает на то, что распределение `Heart_Rate` не отклоняется от нормального. Это подтверждает выполнение условия нормальности.
- **Тест Левена**
  Значение $p = 0.9326$ указывает, что дисперсии в группах равны. Это соответствует предположению гомогенности дисперсий, необходимому для ANOVA.

19.2 Однофакторный ANOVA

Однофакторный ANOVA для `Health_Risk_Level`

- **Результаты:**
  - F-статистика: 0.67
  - p-value: 0.5130

  Поскольку $p > 0.05$, мы не можем отвергнуть нулевую гипотезу. Это означает, что нет статистически значимых различий в среднем значении `Heart_Rate` между группами, определёнными уровнем риска здоровья (`Health_Risk_Level`).
- **Вывод:**
  Уровень риска здоровья (`Health_Risk_Level`) не оказывает значимого влияния на средний уровень сердечного ритма (`Heart_Rate`).

19.3 Многофакторный ANOVA

Многофакторный ANOVA для `Mood` и `Sleep_Quality`

- **Результаты:**
  - Влияние фактора `Mood`:
    - $F = 0.187$, $p = 0.829$
    - Нет статистически значимого влияния настроения (`Mood`) на уровень сердечного ритма.
  - Влияние фактора `Sleep_Quality`:
    - $F = 1.374$, $p = 0.254$
    - Качество сна (`Sleep_Quality`) также не оказывает значимого влияния на уровень сердечного ритма.

  - Остаточная дисперсия (*Residual*) составляет основную часть общей вариации, что указывает на сильное влияние других факторов, не учтённых в модели.

- **Вывод:**
  Ни настроение (`Mood`), ни качество сна (`Sleep_Quality`) не оказывают статистически значимого влияния на уровень сердечного ритма (`Heart_Rate`).

20 Результаты моделей и их интерпретация

20.1 Результаты моделей

Обучили три модели на данных о такси, предсказывали стоимость поездки, и получили следующие метрики качества:

- Линейная регрессия:
  - $R^2 = 0.8251$
  - MSE = 16.7651
- Случайный лес:
  - $R^2 = 0.8534$
  - MSE = 14.0527
- Полиномиальная регрессия:
  - $R^2 = 0.8767$
  - MSE = 11.8167

Реализации на R и на Python дали разные результаты: На R:

- Линейная регрессия:
  - $R^2 = 0.8727$
  - MSE = 19.1073
- Полиномиальная регрессия:
  - $R^2 = 0.9015$
  - MSE = 169539.3441 - очень сильно отличается результат, так как функция реализована вручную.

20.2 Обсуждение качества моделей

Несмотря на то, что модели показывают достаточно высокое качество ($R^2 > 0.8$), результаты не идеальны. Возможные причины и пути улучшения:

- **Неучтённые факторы:** Модели не учитывают временные характеристики (например, час дня или день недели), погодные условия, пробки и другие важные параметры, которые могут существенно влиять на стоимость поездки.
- **Нелинейные зависимости:** Линейная и полиномиальная регрессии могут не полностью захватывать сложные зависимости между признаками, например, географическими координатами.
- **Мультиколлинеарность:** Признаки, такие как долгота и широта, могут быть сильно коррелированы, что усложняет обучение моделей.

20.3 Вывод

Модели показывают хорошие результаты, но есть потенциал для улучшения. К сожалению, удалением мультиколлинеарных признаков улучшить модель не удалось, возможно требуются наоборот добавить больше признаков, чтобы учитывать сложные зависимости.

21 Вывод

В ходе выполнения работы были проанализированы два датасета: данные о поездках на такси и данные о здоровье студентов. На них были применены различные статистические методы. Были исследованы распределения данных с помощью ядерных оценок плотности, проведены статистические тесты на выявление выбросов и проверка на нормальность распределения. Также были изучены методы заполнения пропусков в данных и проверки гипотез, включая критерии Стьюдента, Уилкоксона-Манна-Уитни, Фишера, Левене, Бартлетта и Флигнера-Килина.

В ходе анализа корреляционных взаимосвязей были выявлены сильные и слабые зависимости между переменными, а также проведена проверка на мультиколлинеарность с использованием фактора инфляции дисперсии (VIF). Был проведен дисперсионный анализ (ANOVA) для изучения влияния различных факторов на уровень сердечного ритма студентов.

На основе данных о такси были обучены три модели: линейная регрессия, случайный лес и полиномиальная регрессия, для предсказания стоимости поездки. Результаты показали, что модели имеют достаточно высокое качество, но есть потенциал для улучшения, особенно за счет учета неучтенных факторов и устранения мультиколлинеарности.

В целом, работа продемонстрировала важность тщательного анализа данных и выбора подходящих статистических методов для получения корректных и значимых результатов.