МИНОБРНАУКИ РОССИИ

Федеральное государственное автономное образовательное учреждение высшего образования «Южный федеральный университет»

Институт математики, механики и компьютерных наук им. И. И. Воровича

Кафедра информатики и вычислительного эксперимента

Иванов Александр

Андреева Екатерина

Пустовойт Максим

Лавронов Арсений

Научно-исследовательская работа

получение первичных навыков научно-исследовательской работы

«Системы компьютерной верстки и визуализации данных»

на примере датасета “Титаник”

по направлению подготовки 01.03.02 – Прикладная математика и информатика

Научный руководитель – Чикина Любовь Григорьевна

Ростов-на-Дону – 2023

# Оглавление

[Оглавление 2](#_Toc140567767)

[1. Введение 4](#_Toc140567768)

[Цели: 4](#_Toc140567769)

[Задачи: 4](#_Toc140567770)

[2. Предварительная обработка данных 5](#_Toc140567771)

[2.1. Первичная визуализация данных 5](#_Toc140567772)

[2.1.1 Анализ числовых показателей 9](#_Toc140567773)

[2.1.2 Анализ категориальных показателей 12](#_Toc140567774)

[2.2. Очистка данных 14](#_Toc140567775)

[2.2.1 Метод межквартильного размаха 14](#_Toc140567776)

[2.2.2. Замена выбросов на концы интервала 16](#_Toc140567777)

[2.2.3 Метод доверительных интервалов 16](#_Toc140567778)

[2.2.4. Ориентируясь на здравый смысл 18](#_Toc140567779)

[2.3. Заполнение пропущенных данных 19](#_Toc140567780)

[2.3.1. Заполнение столбца “Возраст” средними 19](#_Toc140567781)

[2.3.2. Z-стандартизация данных 20](#_Toc140567782)

[2.3.3. Min-Max нормализация данных данных 22](#_Toc140567783)

[2.4. Корреляционно-регрессионный анализ 23](#_Toc140567784)

[2.4.1. Регуляризация линейной регрессии 23](#_Toc140567785)

[2.4.2. Сравнительный анализ кластеризации методом К-средних,среднего сдвига, иерархической. 25](#_Toc140567786)

[3. Анализ данных 30](#_Toc140567787)

[3.1. Составление выборки для анализа 30](#_Toc140567788)

[3.2. Анализ данных при помощи тестов 32](#_Toc140567789)

[3.2.1 Тест Андерсона - Дарлинга 32](#_Toc140567790)

[3.2.2 Тест хи-квадрат 37](#_Toc140567791)

[Заключение 40](#_Toc140567792)

[Список литературы 40](#_Toc140567793)

# 1. Введение

## Цели:

* Получение данных
* Очистка данных
* Составление выборки для анализа
* Обработка и анализ выборки

## Задачи:

* Чтение данных из csv файла
* Предобработка полученных данных
* Создание описательной статистики
* Визуализация данных
* Проведение тестов для анализа данных

**Используемые библиотеки:**

pandas - добыча информации, обработка и анализ

matplotlib, seaborn - построение геометрической интерпретации

numpy, scipy - выполнение математических вычислений

**Используемый язык программирования:** python 3.9.13

**Среда разработки:** Jupyter Notebook

# 2. Предварительная обработка данных

## 2.1. Первичная визуализация данных

Датасет ”Титаник” состоит из 11 столбцов, не считая порядкового номера. Все они названы на английском языке, поэтому для удобства восприятия переведём их и сократим названия, для удобства использования.

Рис. 1 обзор датасета

При помощи встроенных в библиотеку pandas инструментов проведём общий анализ всех столбцов датасета.  
Рис. 2 общая статистика датасета по столбцам

Из описательно статистики видно, что в столбце “Возраст”, в отличии от остальных, среднее близко к медиане, что может говорить о, например, симметричном или даже нормальном распределении, требуется дальнейшая проверка.

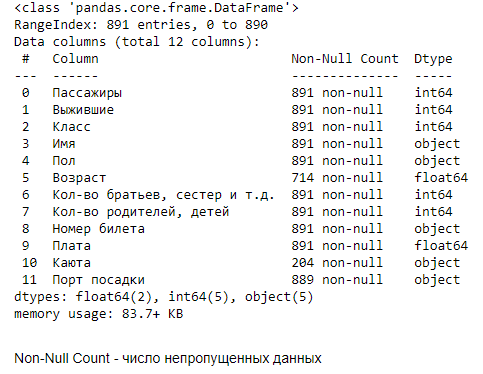
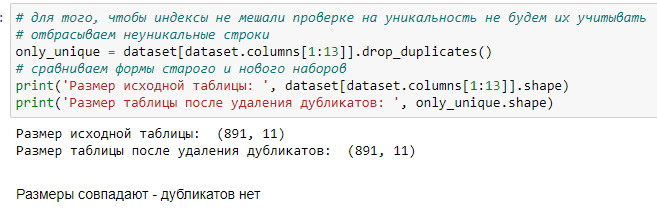


Рис. 3 подробная информация о каждом столбце

Проверим таблицу на наличие повторяющихся записей

Рис. 4

Проверим пропущенные или некорректные данные  


Рис. 5

Красный означает пропущенные записи, зелёный - непропущенные.

Видно, что в столбцах “Возраст”, “Порт посадки” и особенно “Каюта” пропущены данные.

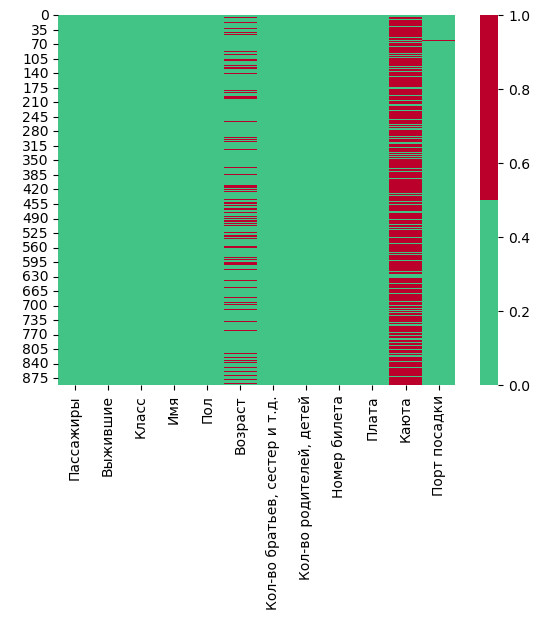


Рис. 6 карта пропущенных данных

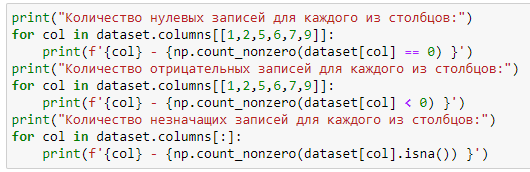
Проверка на наличие некорректных данных  


Рис. 7

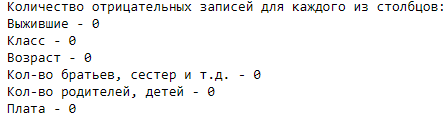


Рис. 8

Проверка показало, что некорректные, а в случае числовых показателей - отрицательные, данные отсутствуют.

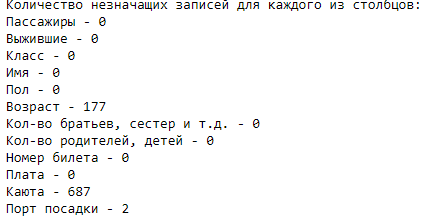
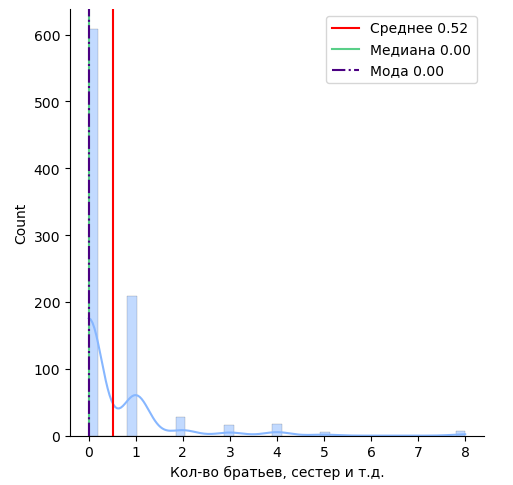
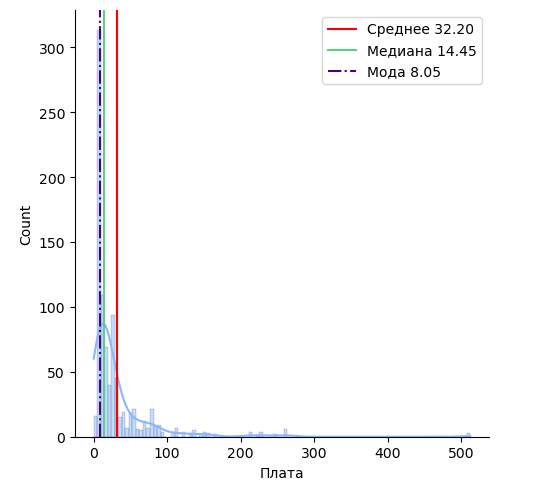
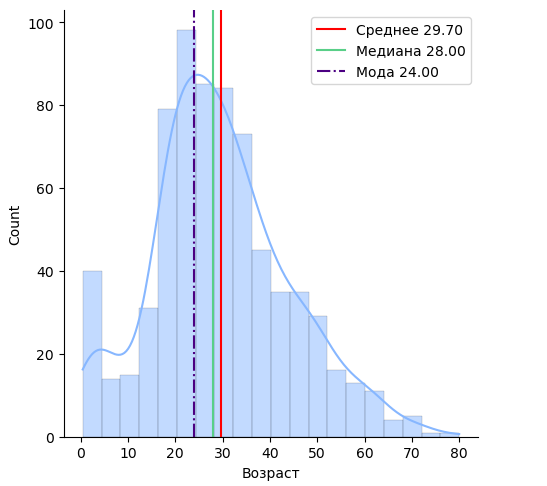
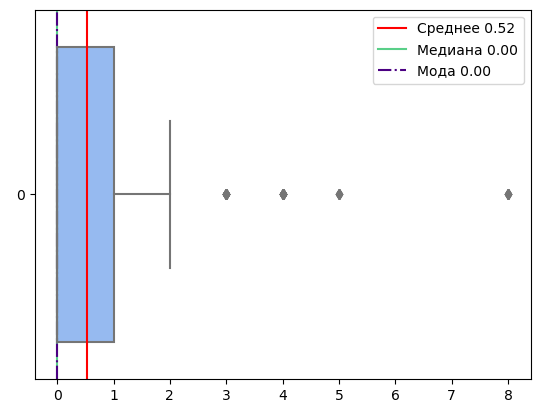
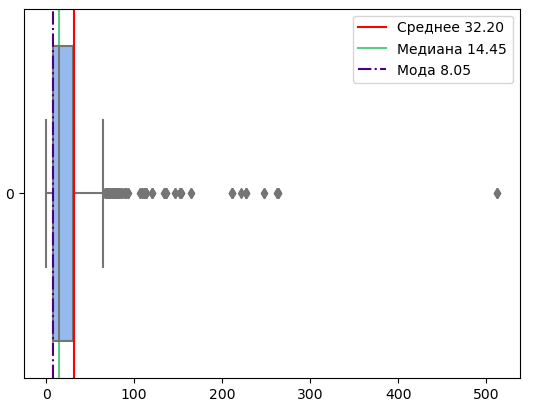
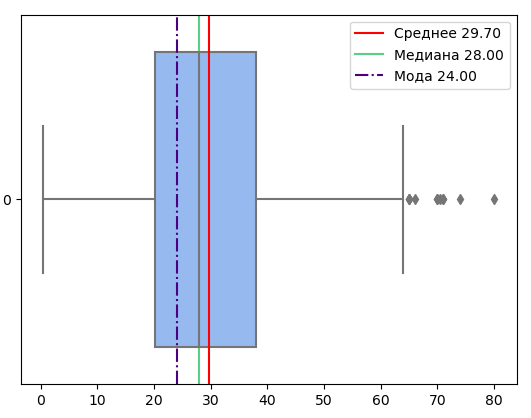


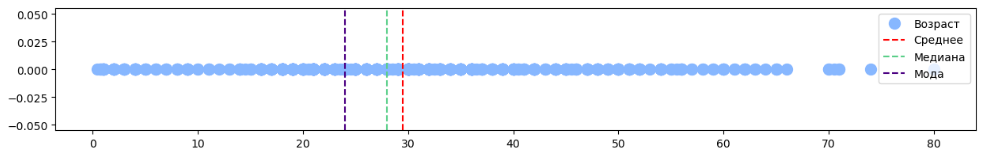
Рис. 9

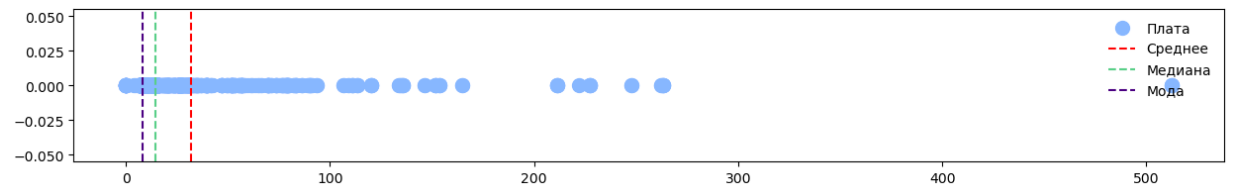
В случае с незначащими данными (NaN) - численные показатели соответствуют графическим.

### 2.1.1 Анализ числовых показателей









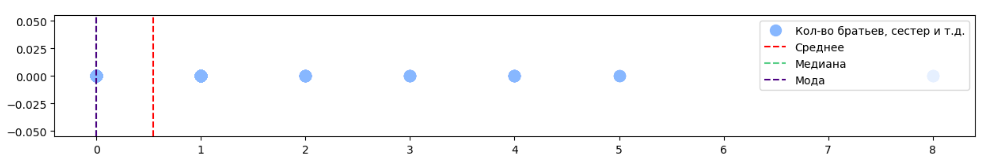


Рис. 10-18 гистограммы, коробки с усами и графики дисперсии самых информативных столбцов



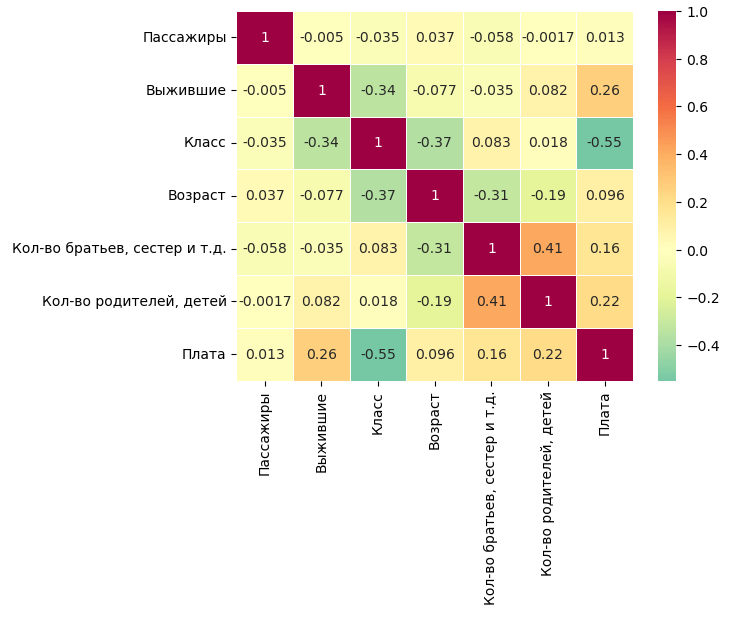
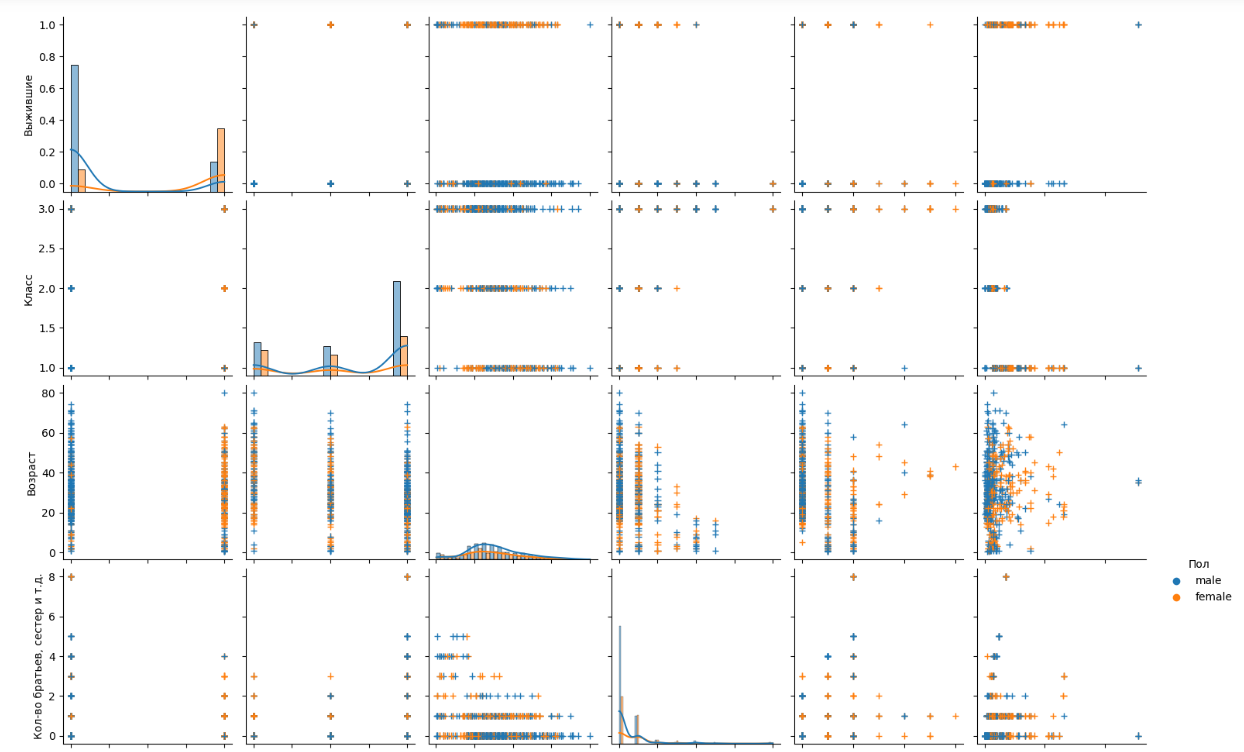


Рис. 19 тепловая карта корреляции

Тепловая карта корреляции показывает, что большинство параметров не связаны линейной зависимостью.



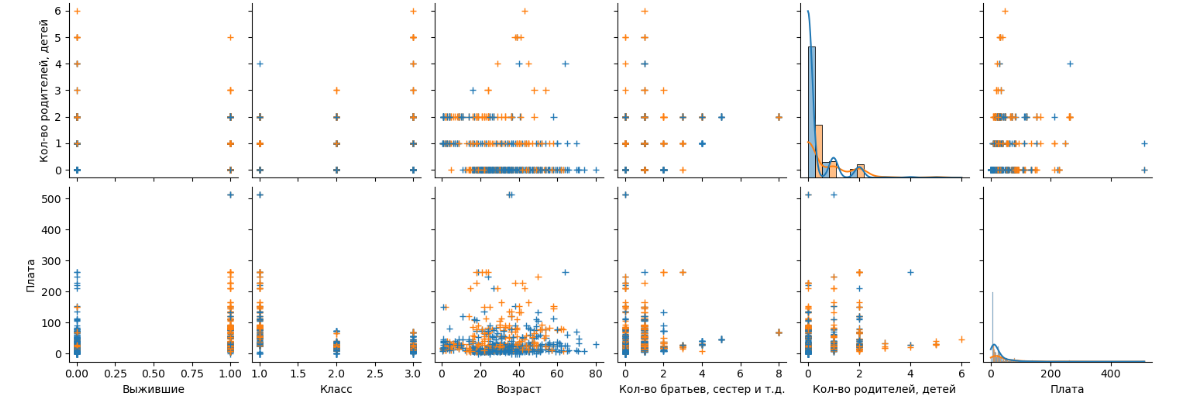


Рис. 20-21 таблица зависимости параметров, где на диагонали количественная гистограмма с графиком плотности.

За счёт большого количества столбцов сложно анализировать таблицу.

### 2.1.2 Анализ категориальных показателей

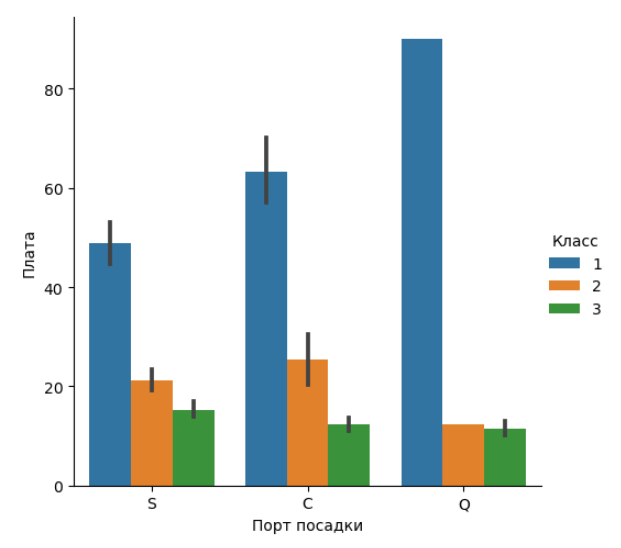
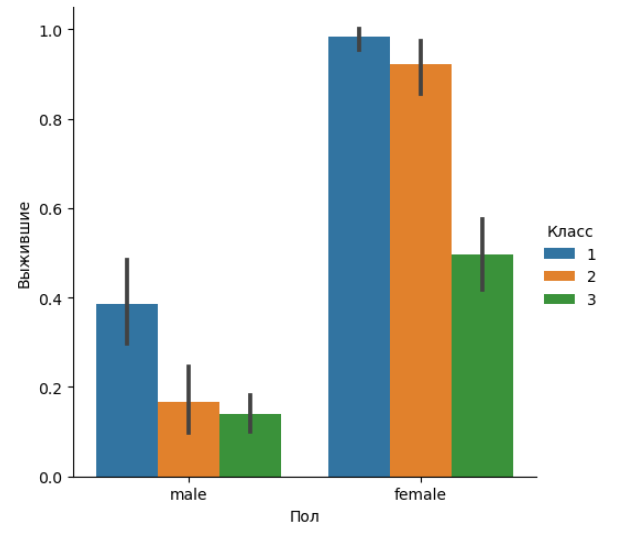


Рис. 22-23 соотношение процента выживших к полу, относительно класса и соотношение платы к порту посадки, относительно класса

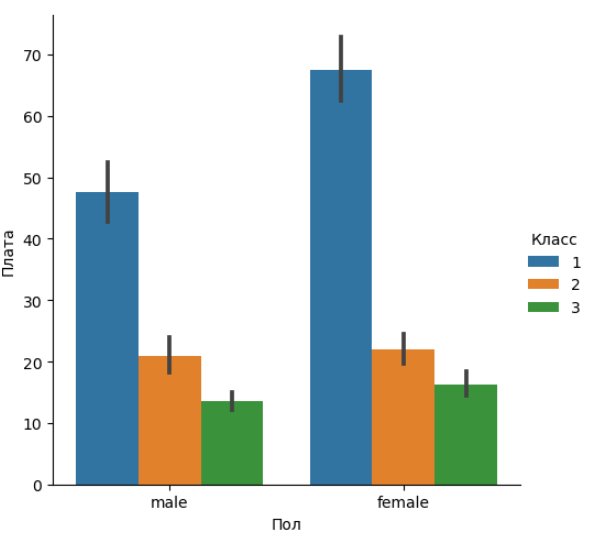
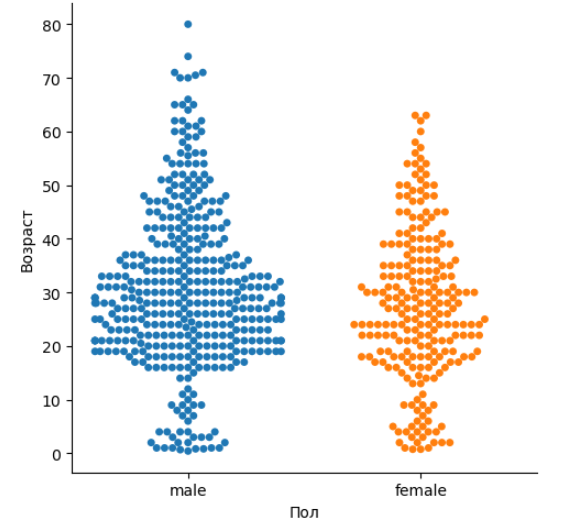


Рис. 24-25 Распределение по возрасту, относительно пола и средняя плата относительно пола по классу.

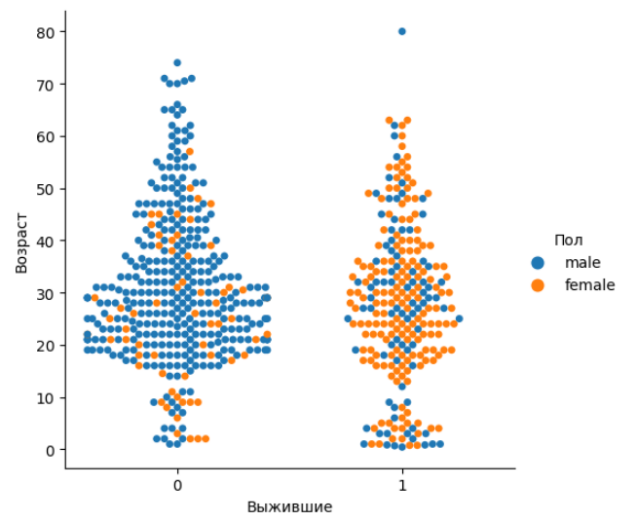
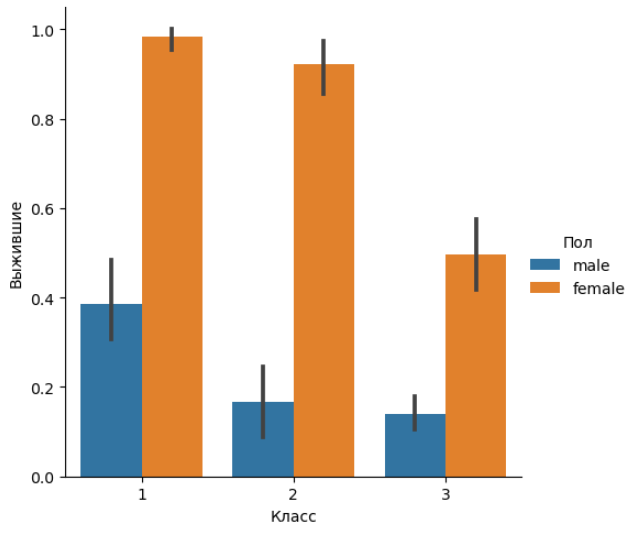


Рис. 26-27 Соотношение выживших по классу, относительно пола и соотношение выживших по возрасту, относительно пола

Из графиков 22, 26 видно, что женщин спаслось больше, чем мужчин.

Из графика 27 видно, что выжившие были в основном в возрасте от 20 до 40.

Из графика 23, 25 видно, что стоимость билета не зависела от пола, и зависела от класса и порта посадки.

График 24 показывает, что распределение по возрастам приняло классический вид “ёлочки” для обоих полов.

Всё вышеперечисленное косвенно говорит о “нормальности” данных.

## 2.2. Очистка данных

### 2.2.1 Метод межквартильного размаха

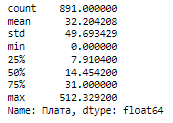
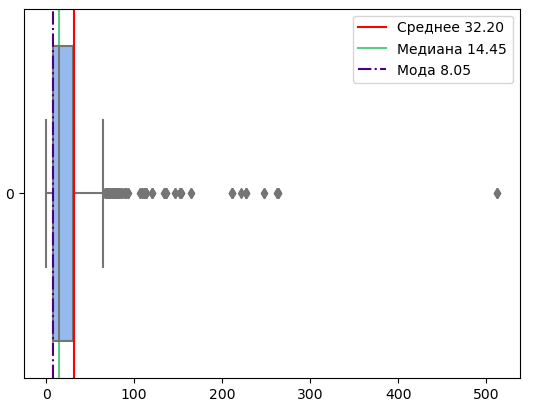
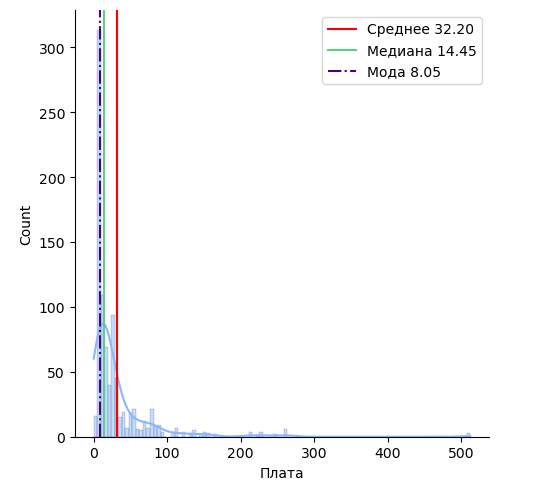
**Выброс** — это наблюдение, которое лежит аномально далеко от других значений в наборе данных. Выбросы могут быть проблематичными, поскольку они могут повлиять на результаты анализа.

Один из распространенных способов найти выбросы в наборе данных — использовать межквартильный диапазон.

Межквартильный диапазон, часто сокращенно IQR, представляет собой разницу между 25-м процентилем (Q1) и 75-м процентилем (Q3) в наборе данных. Он измеряет разброс средних 50% значений.

Один из популярных методов состоит в том, чтобы объявить наблюдение выбросом, если его значение в 1,5 раза больше, чем IQR, или в 1,5 раза меньше, чем IQR.

Так как только столбец “Плата” обладает выбросами, работать будем с ним.



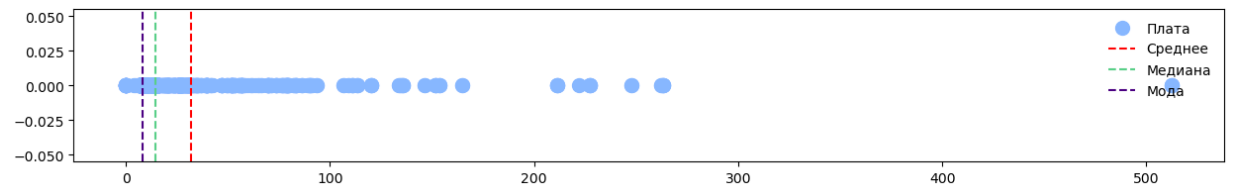
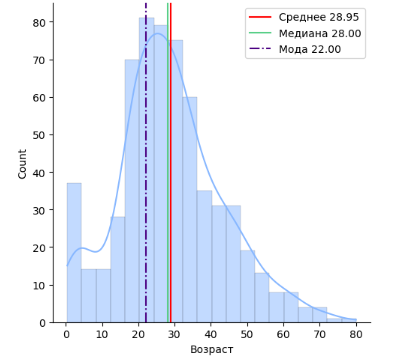
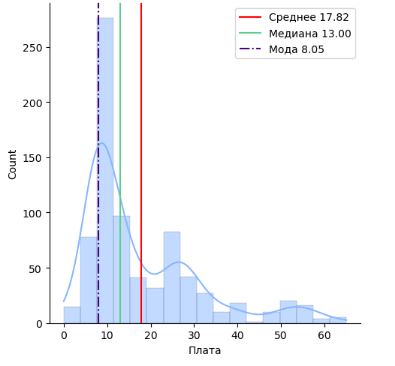


Рис. 28-31

Видно, что Q1 = 7,9, Q3 = 31, тогда IQR = 31–7,9 = 23,1  
Нижний предел = Q1 – 1,5\*IQR = 7.9–1,5\*23.1 = -26,75  
Верхний предел = Q3 + 1,5\*IQR = 31 + 1,5\*23.1 = 65,65





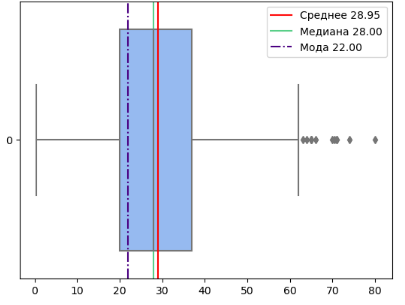
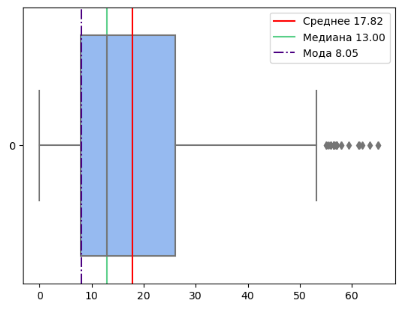
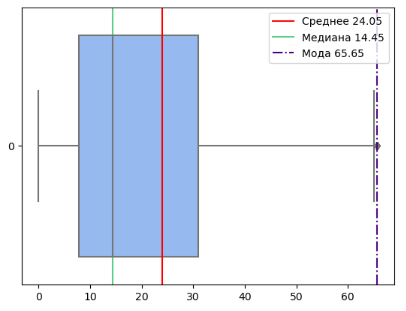
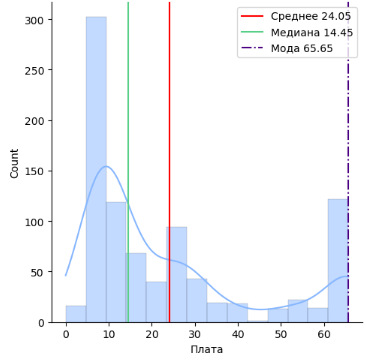


Рис. 32 – 35 изменённые данные

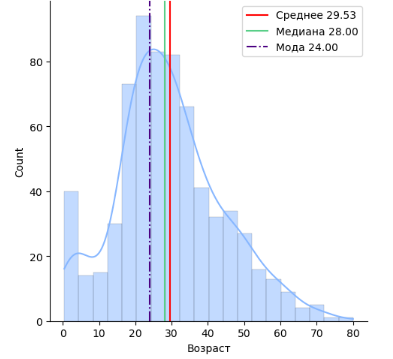
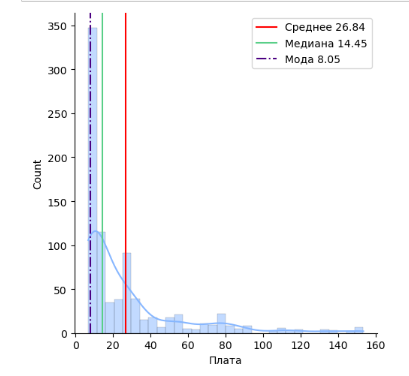
Видно, что количество выбросов в столбце “Плата” сильно сократилось и распределение в других столбцах не сильно изменилось, но отсечение половины датасета может повлечь какие-нибудь проблемы в будущем, поэтому лучше применять такой метод как один из вариантов для проверки данных, не изменяя при этом исходный датасет.

### 2.2.2. Замена выбросов на концы интервала

Возьмём интервал из предыдущего примера - [-26,75; 65,65]  
  
 Рис. 36 - 37  
Легко заметить искажение действительности в данных, поэтому данный способ нам не подходит.

### 2.2.3 Метод доверительных интервалов

Для границ доверительно интервала возьмём 2 и 3.



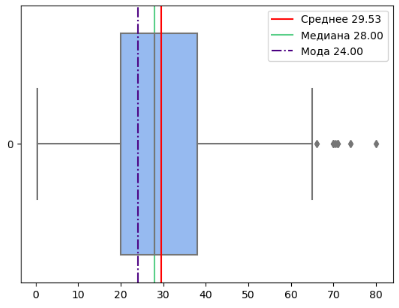
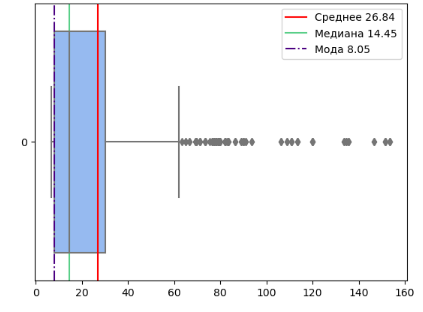


Рис. 38-41 результат отсечения для

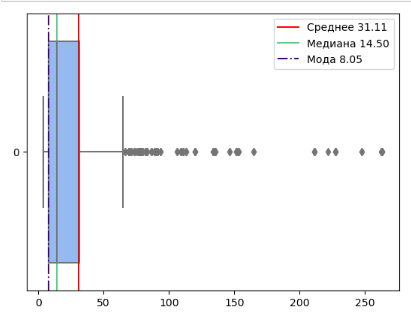
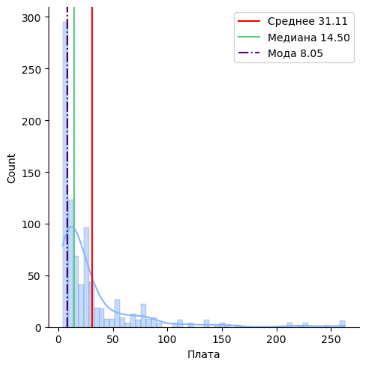
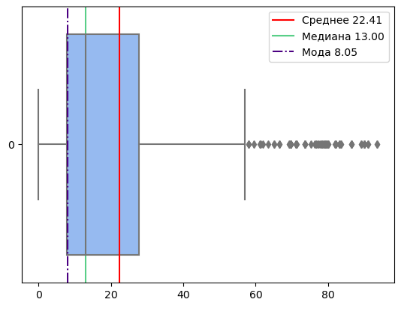
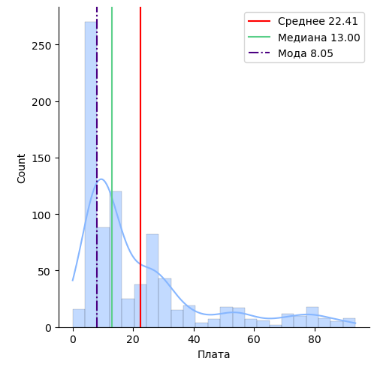
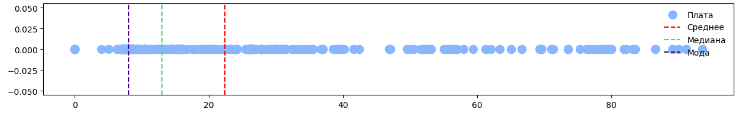


Рис. 42-43 результат отсечения для

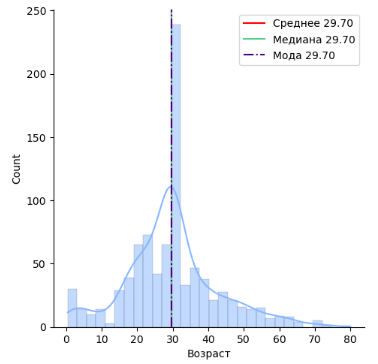
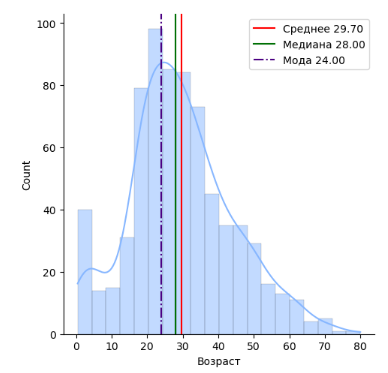
Оба варианта имеют свои достоинства и недостатки, распределение для более непрерывное, а для сохраняется больше информации об исходных данных.

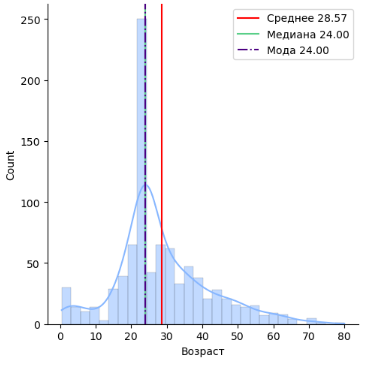
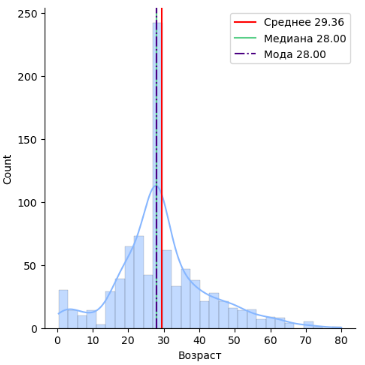
### 2.2.4. Ориентируясь на здравый смысл

Заметим, что из всех столбцов "Плата" имеет очень большой хвост и некоторое количество выбросов, которые могут являться ошибкой ввода. Также можно заметить, что до значения 100 имеет большую плотность и ведёт себя одинаково, но затем данные становятся реже относительно друг друга и их поведение становится непредсказуемым.  
  
  
 Рис. 44-46 данные столбца "Плата" после изменения  
  
  
Данная выборка может быть полезна при анализе распределения записей о плате, так как распределение похоже на распределение Гамбела.

## 2.3. Заполнение пропущенных данных

### 2.3.1. Заполнение столбца “Возраст” средними





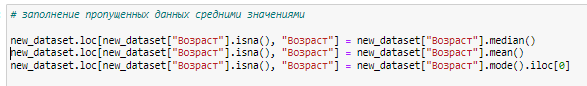
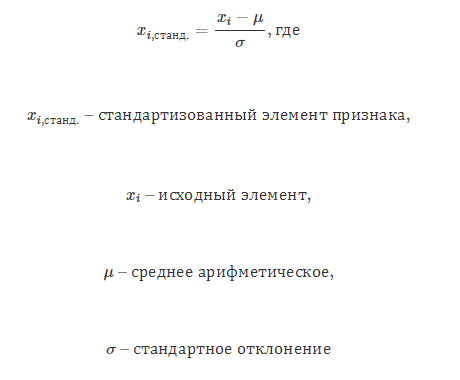


Рис. 47 – 51 исходные данные, а также с заполненными средним, медианой и модой пропусками

Видно, что заполнение пропущенных данных каким-либо одним значением приводит к искажению действительности.

### 2.3.2. Z-стандартизация данных

**Z-стандартизация** (Z-score Normalization) – техника преобразования значений признака, адаптирующая признаки с разными диапазонами значений к Моделям (Model) Машинного обучения (ML), использующих дистанцию для прогнозирования. Эта разновидность нормализации с использованием Стандартизированной оценки (Z-Score) преобразует значения таким образом, что из каждого Наблюдения каждого Признака вычитается Среднее значение и результат делится на Стандартное отклонение этого признака. Как правило, каждый стандартизованный элемент признака вычисляется следующим образом:



Такое преобразование необходимо, поскольку признаки датасета могут иметь большие различия между своими диапазонами, и для моделей Машинного обучения, основанных на вычислении дистанции между точками на графике как основу прогнозирования. **Z-оценка** – один из самых популярных методов стандартизации, однако не единственный. После стандартизации все столбцы будут иметь среднее значение, равное нулю, стандартное отклонение, равное единице, и, следовательно, одинаковый масштаб влияния на модель.

Теперь проведем z-стандартизацию нашего датасета. Для этого мы добавим одну библиотеку:



Далее для понимания насколько важна z-стандартизация мы сделаем два графика:

До стандартизации и После стандартизации.



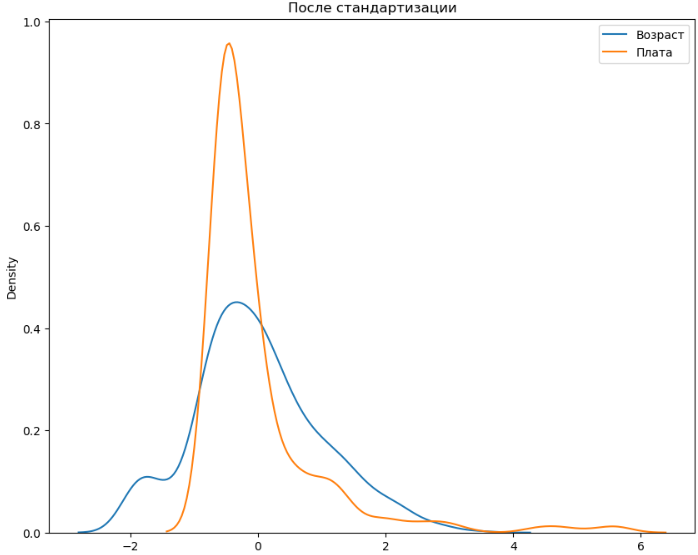
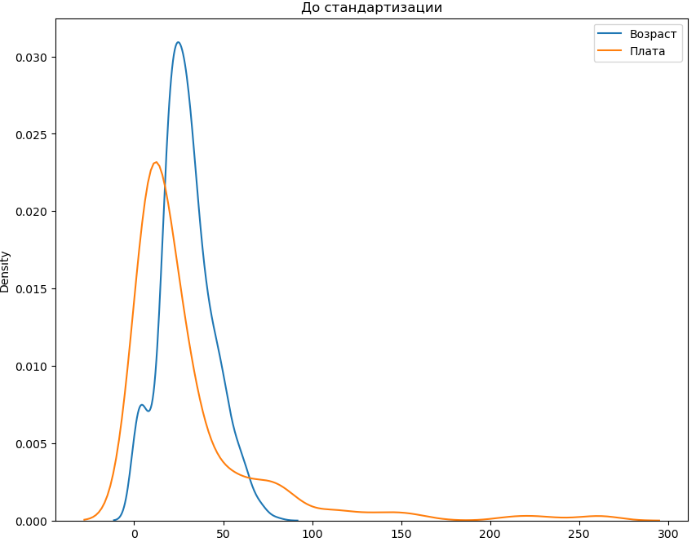
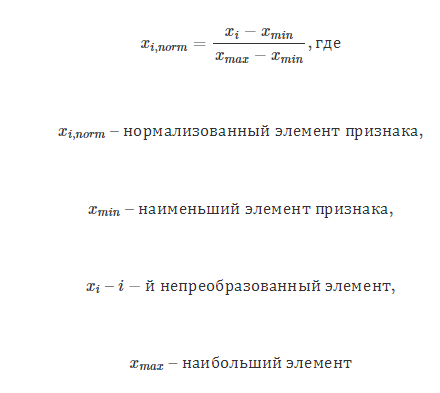


Рис 52-53 график кривых до стандартизации и после.

На рис. 53 После стандратизации мы получили преобразованные кривые, и теперь их масштабы влияния на выбранную модель будут соответствовать друг другу.

### 2.3.3. Min-Max нормализация данных данных

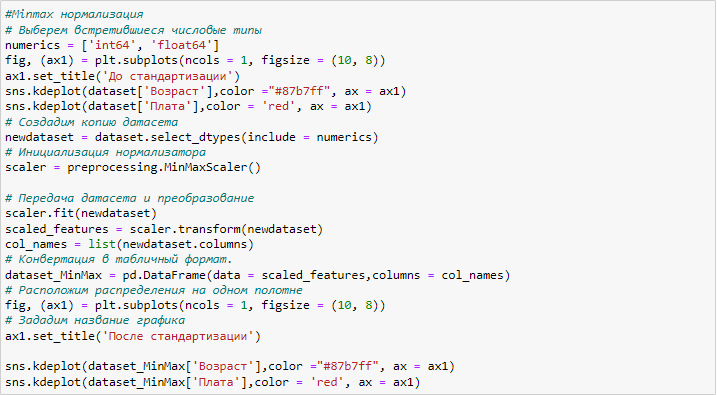
Нормализация (Max-Min Normalization, Min-Max Scaling) – техника преобразования значений признака, масштабирующая значения таким образом, что они располагаются в диапазоне от 0 до 1. Вычисляется каждый нормализованный элемент признака с помощью формулы:



Цель такого преобразования – изменить значения числовых столбцов в наборе данных так, чтобы сохранить различия их диапазонов. В Машинном обучении датасет требует нормализации, когда признаки имеют разные диапазоны и тем самым способствуют искажению восприятия взаимоотношений между Переменными-предикторами и Целевой переменной.

Теперь проведем нормализация нашего датасета.

Воспользуемся той же самой библиотекой, что мы пользовались при стандартизации и также построим 2 графика для того, чтобы заметить разницу.



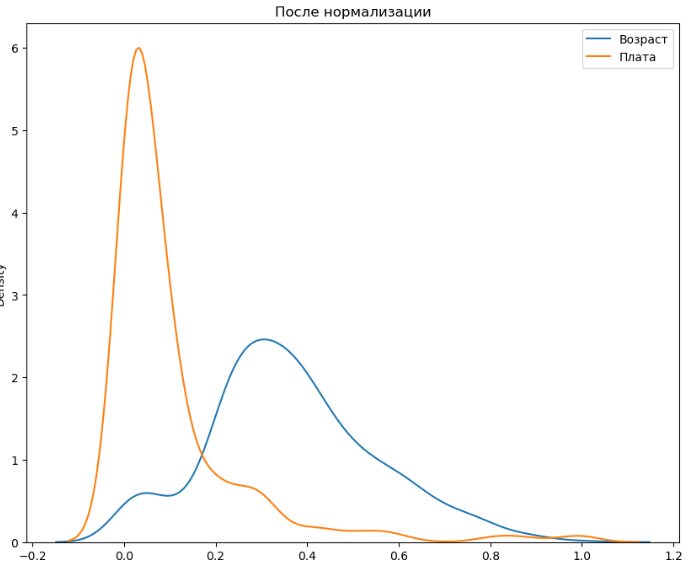
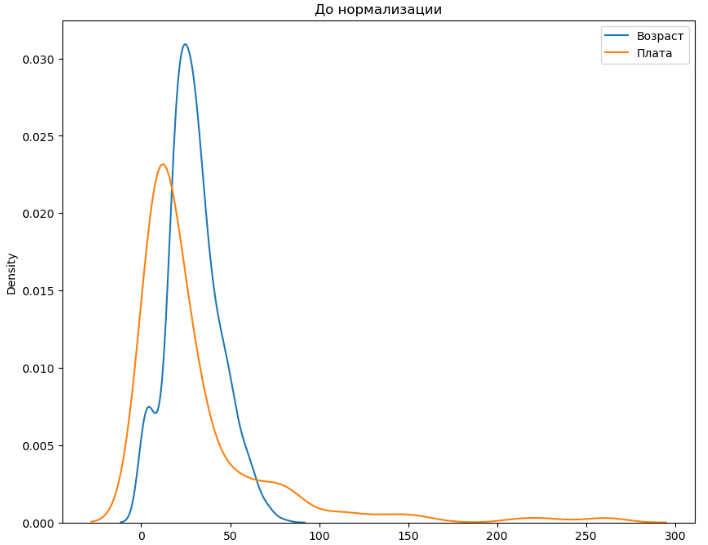


Рис 54-55 график кривых до нормализации и после.

Можно заметить как сильно повлияла нормализация на отношение между признаками «Возраст» и «Плата» на рис 54-55.

## 2.4. Корреляционно-регрессионный анализ

### 2.4.1. Регуляризация линейной регрессии

**Регуляризация** (англ. *regularization*) в статистике, машинном обучении, теории обратных задач — метод добавления некоторых дополнительных ограничений к условию с целью решить некорректно поставленную задачу или предотвратить переобучение. Чаще всего эта информация имеет вид штрафа за сложность модели.

L1 регуляризация также известна как Lasso (Least Absolute Shrinkage and Selection Operator) регуляризация. Она основана на добавлении штрафа, равного абсолютному значению коэффициентов модели.

Формально, L1 регуляризация добавляет в функцию потерь дополнительное слагаемое налагающее штраф за сложность модели, то есть высокие веса:

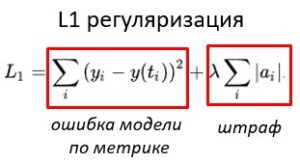


Рис 56-57 – формула L1 регуляризации и ее использование в коде при обучении линейной регрессии

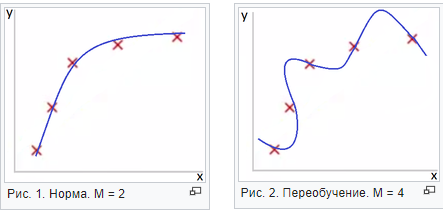
L1 регуляризация склонна к отбору признаков, так как она может уменьшить веса признаков до нуля. Это позволяет убрать неинформативные признаки из модели, что может уменьшить сложность модели и улучшить ее обобщающую способность.

L1 регуляризация является эффективным методом борьбы с переобучением модели в машинном обучении. Однако, при использовании метода градиентного спуска, который является одним из самых популярных алгоритмов оптимизации модели, L1 регуляризация может привести к некоторым проблемам.

В частности, L1 регуляризация имеет несколько «острых» углов (разрывов) в окрестности нуля, где производная не определена. Это усложняет вычисление градиента функции потерь, когда используется L1 регуляризация. Метод градиентного спуска требует, чтобы градиент был гладким и непрерывным, чтобы правильно работать, и поэтому L1 регуляризация может быть менее эффективна при использовании градиентного спуска.

Вместо L1 регуляризации в методе градиентного спуска часто используется L2 регуляризация, так как она имеет более гладкую производную и может лучше работать с градиентным спуском. Однако, в некоторых случаях L1 регуляризация может все же использоваться в методе градиентного спуска с использованием. различных техник оптимизации, таких как координатный спуск или L-BFGS, которые могут лучше обрабатывать разрывы в функции потерь.

В качестве наглядного примера рассмотрим линейные регрессионные модели. Восстановить зависимость для нескольких точек можно пытаться полиномами разной степени M.



На Рис. 1 представлена зависимость, которая хорошо подходит для описания данных, а на Рис. 2 — модель, слишком сильно заточенная под обучающую выборку.

Однин из способов бороться с негативным эффектом излишнего подстраивания под данные — использование регуляризации, т. е. добавление некоторого штрафа за большие значения коэффициентов у линейной модели. Тем самым запрещаются слишком "резкие" изгибы, и предотвращается переобучение.

**Выводы:**

* если регуляризация слишком сильная, то решением задачи минимизации логистической функции потерь может оказаться то, когда многие веса занулились или стали слишком малыми. Еще говорят, что модель недостаточно "штрафуется" за ошибки. В таком случае модель окажется недообученной.
* наоборот, если регуляризация слишком слабая,то решением задачи оптимизации может стать вектор с большими по модулю компонентами. В таком случае больший вклад в оптимизируемый функционал имеет и, вольно выражаясь, модель слишком "боится" ошибиться на объектах обучающей выборки, поэтому окажется переобученной.

### 2.4.2. Сравнительный анализ кластеризации методом К-средних,среднего сдвига, иерархической.

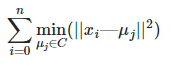
Для начала опишем эти три метода кластеризации.

**Методом К-средних**

Данные алгоритмы кластеров пытаются отдельными образцы в п групп одинаковой дисперсии, сводя к минимуму критерия , известный как *инерция* или внутри-кластера сумм квадратов (см ниже). Этот алгоритм требует указания количества кластеров. Он хорошо масштабируется для большого количества образцов и используется в широком диапазоне областей применения во многих различных областях.

Алгоритм k-средних делит набор N образцы X в K непересекающиеся кластеры C, каждый из которых описывается средним  образцов в кластере. Средние значения обычно называют «центроидами» кластера; обратите внимание, что это, как правило, не баллы из X, хотя они живут в одном пространстве.

Алгоритм K-средних нацелен на выбор центроидов, которые минимизируют **инерцию** или **критерий суммы квадратов внутри кластера** :



Инерцию можно определить как меру того, насколько кластеры внутренне связаны. Он страдает различными недостатками:

* Инерция предполагает, что кластеры выпуклые и изотропные, что не всегда так. Он плохо реагирует на удлиненные кластеры или коллекторы неправильной формы.
* Инерция — это не нормализованная метрика: мы просто знаем, что более низкие значения лучше, а ноль — оптимально. Но в очень многомерных пространствах евклидовы расстояния имеют тенденцию становиться раздутыми (это пример так называемого «проклятия размерности»). Выполнение алгоритма уменьшения размерности, такого как анализ главных компонентов (PCA) перед кластеризацией k-средних, может облегчить эту проблему и ускорить вычисления.

В общих чертах алгоритм состоит из трех шагов. На первом этапе выбираются начальные центроиды, а самый простой метод — выбрать k образцы из набора данных X. После инициализации K-средних состоит из цикла между двумя другими шагами. Первый шаг присваивает каждой выборке ближайший центроид. На втором этапе создаются новые центроиды, взяв среднее значение всех выборок, назначенных каждому предыдущему центроиду. Вычисляется разница между старым и новым центроидами, и алгоритм повторяет эти последние два шага, пока это значение не станет меньше порогового значения. Другими словами, это повторяется до тех пор, пока центроиды не переместятся значительно.

K-средних эквивалентно алгоритму максимизации ожидания с маленькой, все равной диагональной ковариационной матрицей.

Важно также учесть:

* При работе с алгоритмами кластеризации, включая K-Means, рекомендуется стандартизировать данные, поскольку такие алгоритмы используют измерения на основе расстояний для определения сходства между точками данных.
* Из-за итеративной природы K-средних и случайной инициализации центроидов K-средние могут придерживаться локального оптимума и могут не сходиться к глобальному оптимуму. Вот почему рекомендуется использовать разные инициализации центроидов.

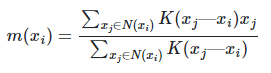
**Средний сдвиг**

Данная кластеризация направлена ​​на обнаружение капель в образцах с плавной плотностью. Это алгоритм на основе центроидов, который работает, обновляя кандидатов в центроиды, чтобы они были средними точками в данном регионе. Затем эти кандидаты фильтруются на этапе постобработки, чтобы исключить почти дубликаты и сформировать окончательный набор центроидов.

Учитывая центроид кандидата  для итерации t, кандидат обновляется в соответствии со следующим уравнением:



Где  это соседство образцов на заданном расстоянии вокруг  а также m — вектор среднего сдвига, который вычисляется для каждого центроида, который указывает на область максимального увеличения плотности точек. Это вычисляется с использованием следующего уравнения, эффективно обновляющего центроид до среднего значения выборок в его окрестности:

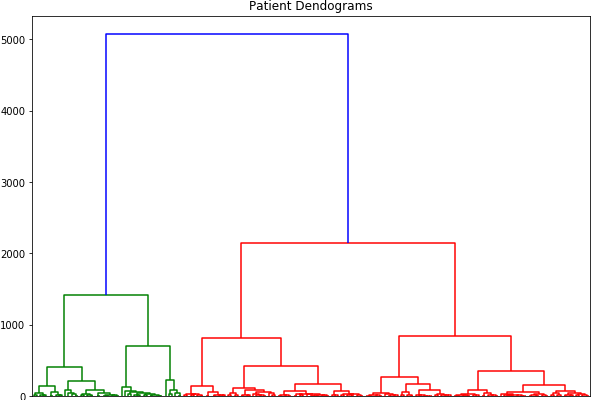


**Иерархическая кластеризация**

Иерархическая кластеризация — это общее семейство алгоритмов кластеризации, которые создают вложенные кластеры путем их последовательного слияния или разделения. Эта иерархия кластеров представлена ​​в виде дерева (или дендрограммы). Корень дерева — это уникальный кластер, который собирает все образцы, а листья — это кластеры только с одним образцом.

В Иерархической кластеризации объекте выполняет иерархическую кластеризацию с использованием подхода снизу вверх: каждый начинает наблюдения в своем собственном кластере, и кластеры последовательно объединены вместе. Критерии связывания определяют метрику, используемую для стратегии слияния:

* **Ward** минимизирует сумму квадратов разностей во всех кластерах. Это подход с минимизацией дисперсии, и в этом смысле он аналогичен целевой функции k-средних, но решается с помощью агломеративного иерархического подхода.
* **Максимальное** или **полное связывание** сводит к минимуму максимальное расстояние между наблюдениями пар кластеров.
* **Среднее связывание** минимизирует среднее расстояние между всеми наблюдениями пар кластеров.
* **Одиночная связь** минимизирует расстояние между ближайшими наблюдениями пар кластеров.

 Иерархическая кластеризация может также масштабироваться до большого количества выборок, когда он используется вместе с матрицей связности, но требует больших вычислительных затрат, когда между выборками не добавляются ограничения связности: он рассматривает на каждом шаге все возможные слияния.

Теперь зная 3 алгоритма кластеризации, построим таблицу, которая покажет нам, где выигрывают эти алгоритмы.

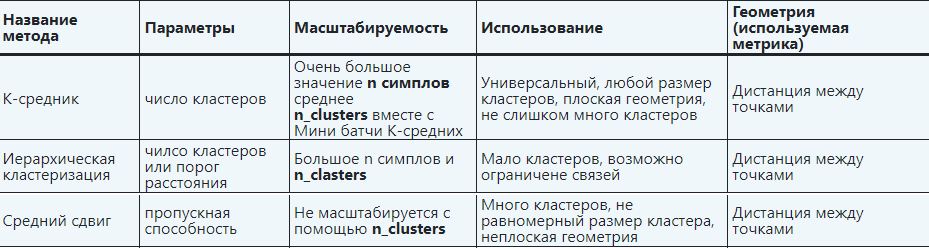


Рис 58-таблица методов кластеризации, их масштабируемости и их использования.

# 3. Анализ данных

# 3.1. Составление выборки для анализа

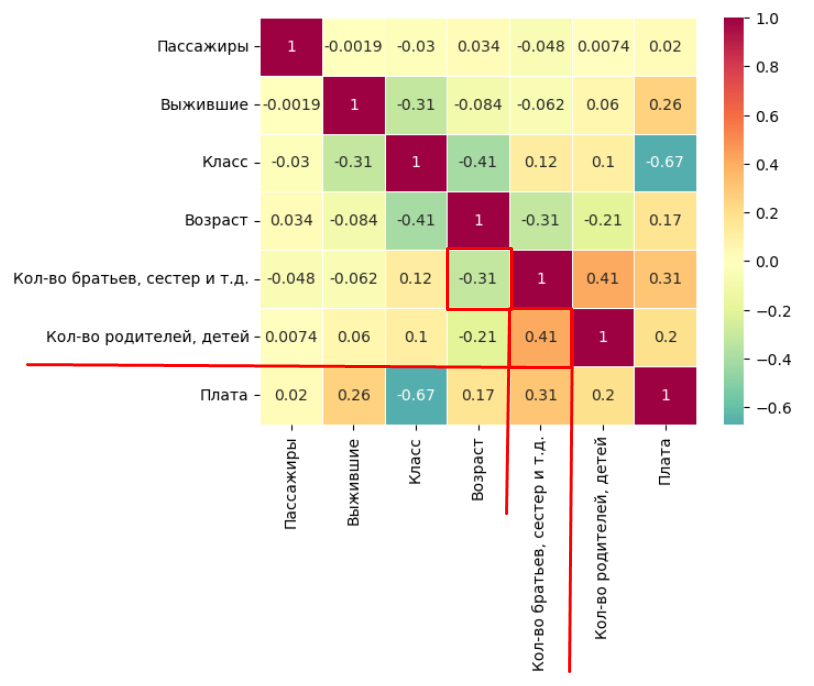


Рис.

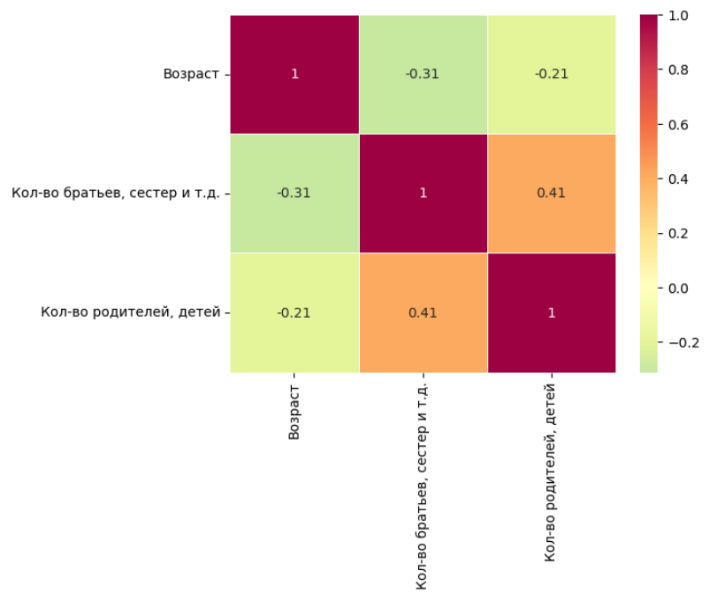
Так как набор из параметров “Возраст”, “Кол-во братьев, сестер и т. д.”, “Кол-во родителей, детей” состоит из пар, имеющих достаточную положительную и отрицательную корреляцию и с её отсутствием, то набора хватит для последующего анализа и проведения тестов. Столбец “Класс” не берётся, так как он состоит из нечисловых данных и анализу не подлежит.  


Рис.

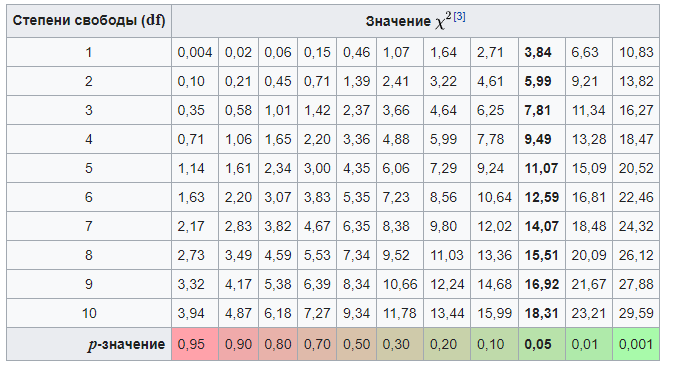
## 3.2. Анализ данных при помощи тестов

### 3.2.1 Тест хи-квадрат

Процедура Критерий хи-квадрат табулирует переменную по категориям и рассчитывает статистику хи-квадрат. Данный критерий согласия сравнивает наблюденные и ожидаемые частоты в каждой категории, чтобы проверить, что либо все категории содержат одинаковые доли значений, либо каждая категория содержит заданную пользователем долю значений.

**Распределение �2 (хи-квадрат) с степенями свободы** — распределение суммы квадратов � независимых стандартных нормальных случайных величин.



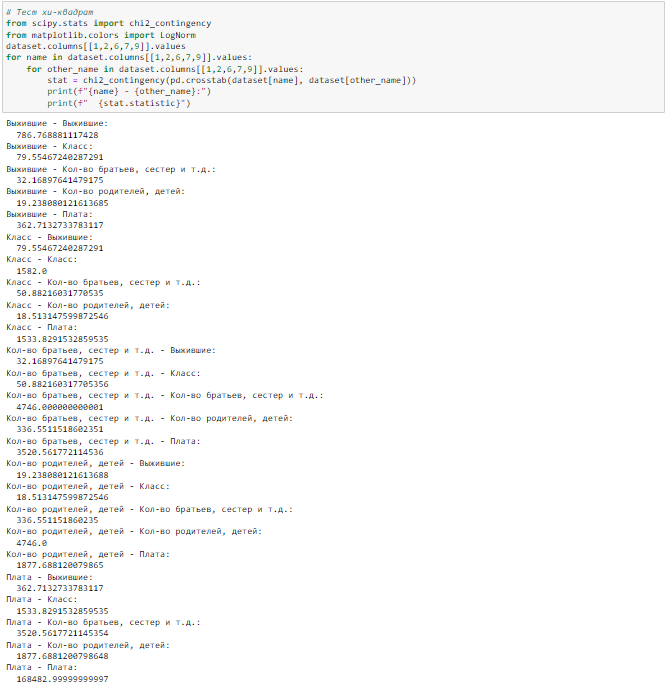


**Критерий хи-квадрат** — любая статистическая проверка гипотезы, в которой выборочное распределение критерия имеет распределение хи-квадрат при условии верности нулевой гипотезы. Считается, что критерий хи-квадрат — это критерий, который *асимптотически* верен, то есть, выборочное распределение можно сделать, как угодно, близким к распределению хи-квадрат путём увеличения размера выборки

Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описаниеПлотность вероятностиИзображение выглядит как линия, текст, График, диаграмма

Автоматически созданное описаниеФункция распредедения



**Ссылка на гит репозиторий:** [**https://github.com/PustovoitMaxim/Titanic**](https://github.com/PustovoitMaxim/Titanic)

# Список литературы

1.[Регуляризация линейной регрессии](https://scikit-learn.ru/clustering/#overview-of-clustering-methods)

2. [Кластеризация датафреймов](https://scikit-learn.ru/clustering/#overview-of-clustering-methods)

3.[L1 регуляризация датафреймов](https://python-school.ru/blog/regularization-l1-l2/)