GeekBrains

# Описательные статистики в контексте EDA. Корреляция и корреляционный анализ

## Библиотеки Python для Data Science

# Оглавление



# Введение

EDA – это явление анализа данных, используемое для лучшего понимания таких аспектов данных, как:
– основные характеристики данных
– переменные и связи, которые существуют между ними
– определение того, какие переменные важны для нашей проблемы.

Основная цель исследовательского анализа данных — добиться уверенности в своих данных до такой степени, чтобы вы были готовы задействовать алгоритм машинного обучения. Этот шаг очень важен, особенно когда мы приступаем к моделированию данных с целью применения машинного обучения. Исследовательский анализ данных — это не что иное, как полная картина данных

**Преимущества:**

Он используется для понимания и обобщения содержимого набора данных, чтобы гарантировать, что функции, которые мы передаем нашим алгоритмам машинного обучения, уточнены, и мы получаем достоверные, правильно интерпретированные результаты.

Дает некоторое представление о наборе данных и некоторое понимание базовой структуры.

Помогает нам извлекать важные параметры и взаимосвязи, которые сохраняются между ними.
Помогает нам проверять лежащие в основе предположения.

Мы рассмотрим различные методы анализа данных, такие как:

- Описательная статистика, которая является способом дать краткий обзор набора данных, с которым мы имеем дело, включая некоторые показатели и особенности выборки
- Группировка данных [Базовая группировка с помощью group by] ANOVA, дисперсионный анализ, который представляет собой вычислительный метод для разделения вариаций в наборе наблюдений на разные компоненты.
- Корреляция и корреляционные методы

# Термины, используемые в лекции

**Среднее арифметическое значение** — сумма значений признака, деленная на общее количество объектов.
**Медиана** — центральное значение атрибута.
**Мода** — это то значение, которое чаще всего встречается (самое модное значение).
**Разброс** описывает разницу между максимальной и минимальной точками в ваших данных.
**Межквартильный размах** (IQR) является мерой статистического разброса между верхним (75-м) и нижним (25-м) квартилями.
**Квартили** — это значения, которые делят отсортированную выборку на четыре примерно равные части.
**Дисперсия** — среднеквадратичное отклонение значений от среднего арифметического, показывающее разброс данных относительно него.
**Корреляция** – это статистическая зависимость между случайными величинами, при которой изменение одной из случайных величин приводит к изменению математического ожидания другой.
**Корреляция Спирмена** — это непараметрическая версия коэффициента корреляции Пирсона, которая измеряет степень связи между двумя переменными на основе их рангов.
**Коэффициент корреляции Пирсона** — это мера силы линейной связи между двумя переменными

**Мультиколлинеарность** — явление, при котором наблюдается сильная корреляция между признаками.

# Описательная статистика

**Описательная статистика**

Описательная статистика — это стандартная процедура анализа данных. Исследовательский анализ данных (EDA) невозможен без описательной статистики.

Описательная статистика — это описание и интегральные параметры наборов данных. Если говорить о метриках, то в этой части изучаются центральные метрики (которые говорят нам о центрах концентрации данных, таких как среднее, медиана и мода) и метрики вариативности данных (которые говорят о разбросе значений, таких как дисперсия и стандартное отклонение).

Начнем с мер центральной тенденции. Мера центральной тенденции - это способ описать центральную позицию частотного распределения из заданных наборов данных:

Среднее арифметическое значение — сумма значений признака, деленная на общее количество объектов, называется средним значением. Он также известен как средние значение.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i\,.$$

Для расчета среднего можно использовать базовые методы python, а также функцию mean() библиотеки Numpy

```
1 x = [8.0, 1, 2.5, 4, 28.0]
2 sum(x) / len(x)
3 np.mean(x)
```

Среднее арифметическое очень чувствительно к выбросам (неадекватным значениям данных). Среднее значение хороший способ описания, но у него есть важный недостаток: что произойдет, если в наборе данных будет ошибка со значением, сильно отличающимся от остальных? Например, если учесть часы, отработанные в неделю, они обычно бывают в диапазоне от 20 до 80; но что

случилось бы, если бы по ошибке было значение 1000? Элемент данных, который значительно отличается от остальных данных, называется выбросом. В этом случае среднее значение будет резко изменено в сторону выброса. Одним из решений этого недостатка является статистическая медиана, которая дает середину выборки. В этом случае все значения располагаются по величине и медианой является значение в середине этого списка. Следовательно, это значение является гораздо более устойчивым перед выбросами.

Медиана — центральное значение атрибута известно как медиана. Чтобы вычислить медианное значение, сначала отсортируйте данные столбца в порядке возрастания или убывания. Затем найдите общее количество строк и разделите его на 2.

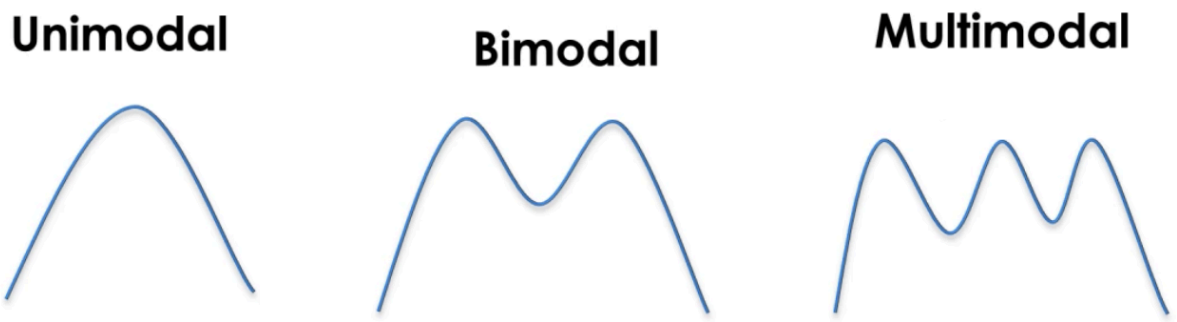
```
n = len(x)
if n % 2: # нечетное
    median = sorted(x)[round(.5*(n-1))]
else:
    x_ord, index = sorted(x), round(.5*n)
    median = .5 * (x_ord[index-1] + x_ord[index])
median
```

Также можно использовать функцию median() библиотеки Numpy

```
np.median(x)
```

Мода — это то значение, которое чаще всего встречается (самое модное значение). Существует только одно среднее и одна медиана для каждого признака. Но признаки могут иметь более одного значения mode. Но может быть набор данных, в котором нет моды вообще, поскольку все значения встречаются одинаковое количество раз.

Если два значения появились одновременно и больше, чем остальные значения, то набор данных бимодальный, если три значения появились одновременно и больше, чем остальные значения, тогда набор данных тримодальный и для n режимов этот набор данных мультимодальные. В этом случае мода считается как среднее от всех значений моды.




Штатными методами можно найти моду:

```
u = [2, 3, 2, 8, 12, 6, 4, 2, 8]
mode = max((u.count(item), item) for item in set(u))[1]
mode
```

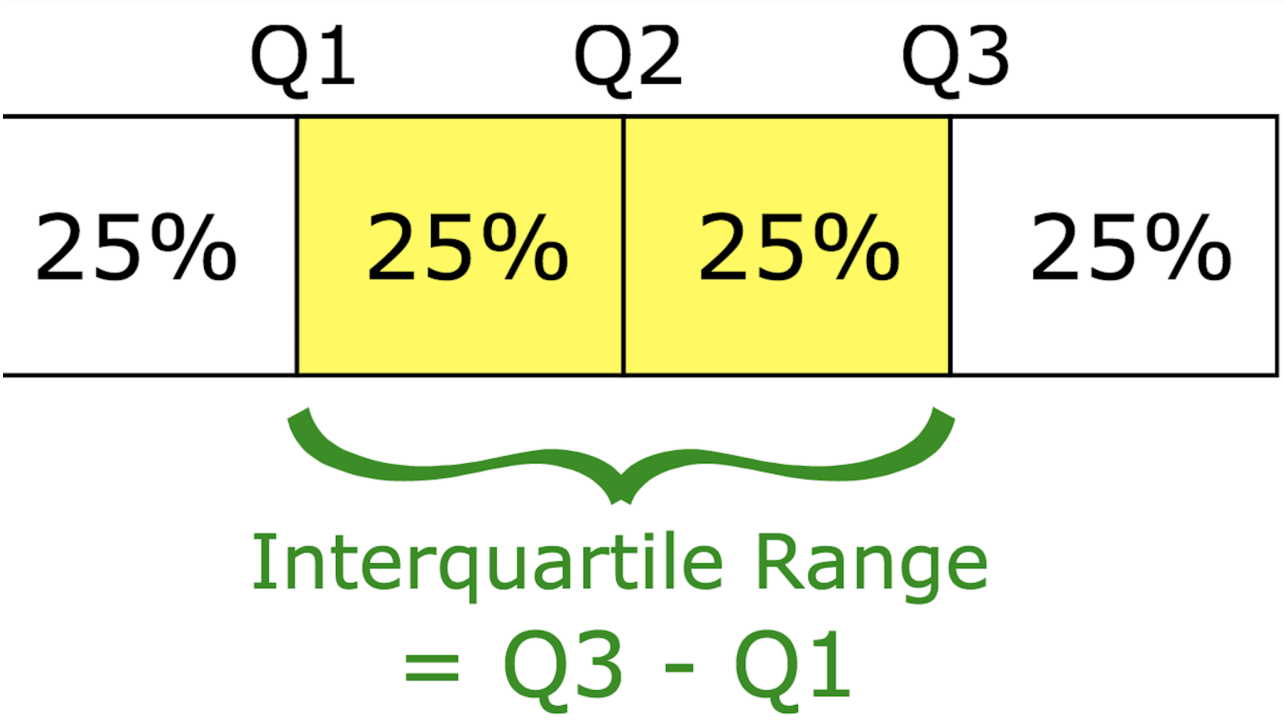
В библиотеке Pandas можно использовать:

```
df0.mode()
```

Метрики вариативности данных используются для измерения разброса или изменчивости в данных.

Разброс описывает разницу между максимальной и минимальной точками в ваших данных. Помимо характеристики границ разброса признаков, этот показатель может быть использован для обнаружения ошибок. Если в данных есть слишком большое (или слишком маленькое) значение, изменчивость быстро и немедленно увеличится (или уменьшится), что потребует изучения и корректировки исходных данных. Недостатком этого показателя является то, что он оценивает только пределы изменения характеристики и не отражает изменчивость в этих пределах.

Межквартильный диапазон (IQR) является мерой статистического разброса между верхним (75-м) и нижним (25-м) квартилями.




Квартили — это значения, которые делят отсортированную выборку на четыре примерно равные части. Первая часть содержит первые 25% данных, вторая часть — следующие 25% данных и так далее.

Второй квартиль отделяет 25% значений в вариационном ряду, третий квартиль — первые 50% значений в вариационном ряду, и, наконец, четвертый квартиль отделяет 100% значений, т.е. все наблюдения в выборке.

Легко видеть, что медиана — это значение второго квартиля. То есть это, то значение, которое отделяет первую половину (0-50%) значений от второй половины в отсортированной выборке (50-100%).

```
1 np.percentile(y,25)
2 np.percentile(y,75)
```

Размах измеряет, где начинаются и заканчиваются значения ваших данных, а межквартильный размах измеряет, где находится большинство значений.

Дисперсия и стандартное отклонение

Стандартное отклонение и дисперсия также измеряют, как размах и IQR, насколько разбросаны наши данные.

Дисперсия – среднеквадратичное отклонение значений от среднего арифметического, показывающее разброс данных относительно него. В случае экстремальных значений величина дисперсии больше и более заметна.

Следовательно, на нее также влияют выбросы.

$$\bar{\sigma}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$$

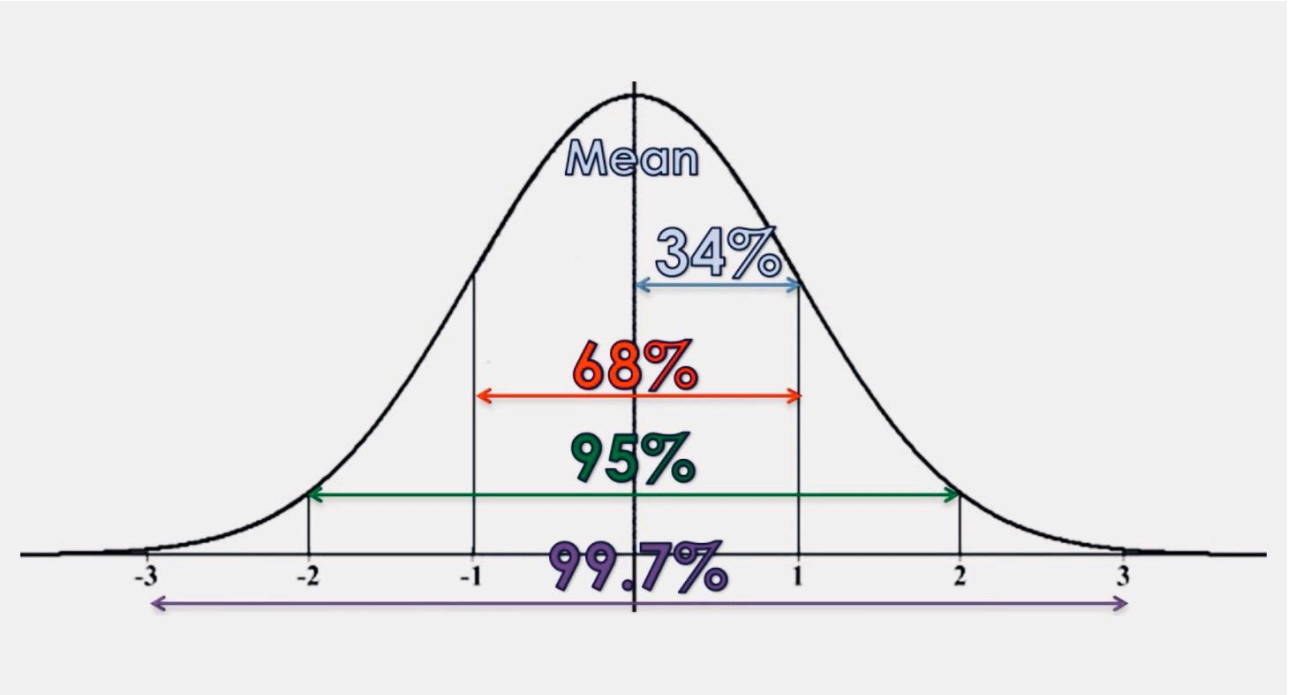
```
n = len(x)
mean = sum(x) / n
var = sum((x - mean)**2 for x in x) / (n - 1)
var
```

Интерпретировать дисперсию достаточно сложно, так как мы получаем квадрат значения. Предположим, что вы имеете дело с набором данных, содержащим значения в сантиметрах. Разница составляет сантиметры в квадрате, что не является оптимальным измерением.

Именно поэтому чаще используется стандартное отклонение, поскольку оно в исходной единице измерения. Стандартное отклонение представляет собой квадратный корень из дисперсии и поэтому возвращается к исходной единице измерения.

Если стандартное отклонение низкое, то точки данных обычно близки к среднему значению. Если стандартное отклонение высокое, это означает, что точки данных распределены в широком диапазоне.

Стандартное отклонение лучше всего использовать, когда данные унимодальны. В случае нормального распределения около 34% точек данных лежат между средним значением и стандартным отклонением выше и ниже него. Поскольку нормальное распределение симметрично, 68% точек данных лежат между стандартным отклонением выше и стандартным отклонением ниже среднего. Около 95% лежат между двумя стандартными отклонениями ниже и двумя стандартными отклонениями выше среднего. И около 99,7% лежат между тремя стандартными отклонениями выше и тремя стандартными отклонениями ниже среднего.




**Задание:** Что такое нормальное или гауссовское распределение?


**Ответ:** Нормальное распределение – это совокупность объектов, в которой крайние значения некоторого признака – наименьшее и наибольшее – появляются редко; чем ближе значение признака к математическому ожиданию, тем чаще оно встречается. Имеет форму колокола.

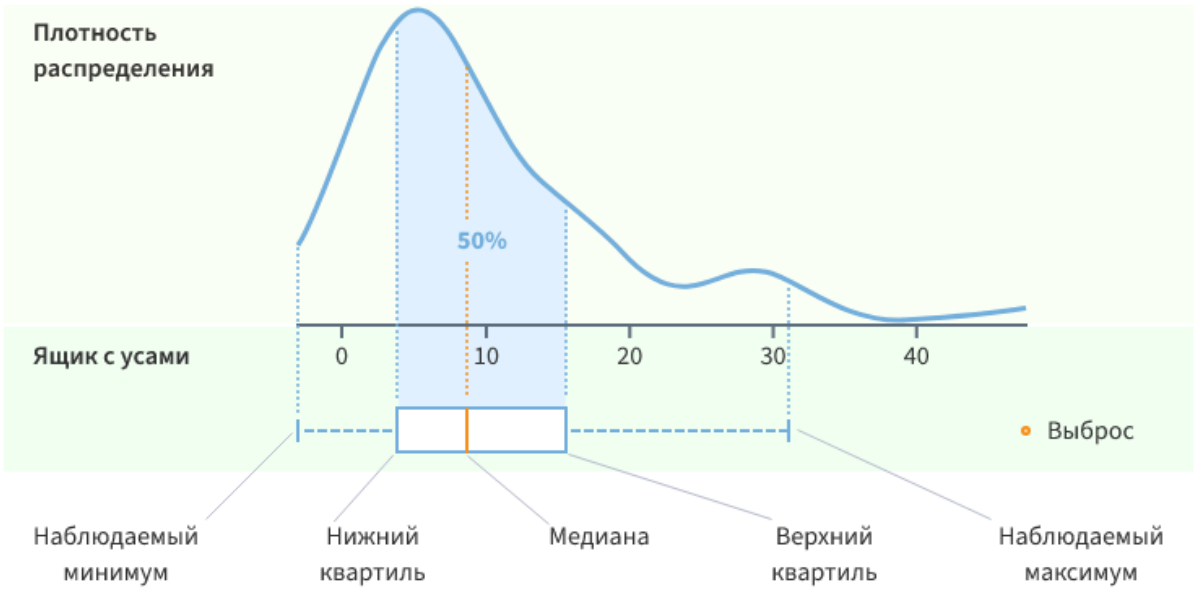
Описательная статистика — это полезный способ понять и проанализировать характеристики ваших данных. Python Pandas предоставляет интересный метод describe(). Функция describe() применяет к набору данных основные статистические расчеты, такие как выбросы, количество точек данных и стандартное отклонение.

Пропущенные значения и значения NaN автоматически опускаются; функция describe() дает хорошее представление о распределении данных.

Еще один полезный метод value_counts(), который позволяет получить количество каждой категории в наборе категориальных значений атрибутов.

Еще одним полезным инструментом является boxplot, который доступен через модуль matplotlib. Boxplot – это графическое представление распределения данных, показывающее выбросы, медианы и квартили. Графики помогают найти выбросы.

Графики Boxplot обеспечивают очень компактное и четкое представление упорядоченной статистики закона распределения (показывая квартили, медиану, минимум, максимум и выбросы, наблюдаемые в выборке). Boxplot можно рассматривать как инструмент непараметрической статистики, поскольку они не делает никаких предположений относительно закона распределения для выборки.




Ящики с усами изначально задумывались как способ представления сводки 5 чисел, то есть набора описательных статистик, описывающих распределение исследуемой выборки. Сводка, представленная в ящике с усами, содержит следующие элементы:

- ➔ минимальное наблюдаемое значение (0-й квартиль или 0-й процентиль)

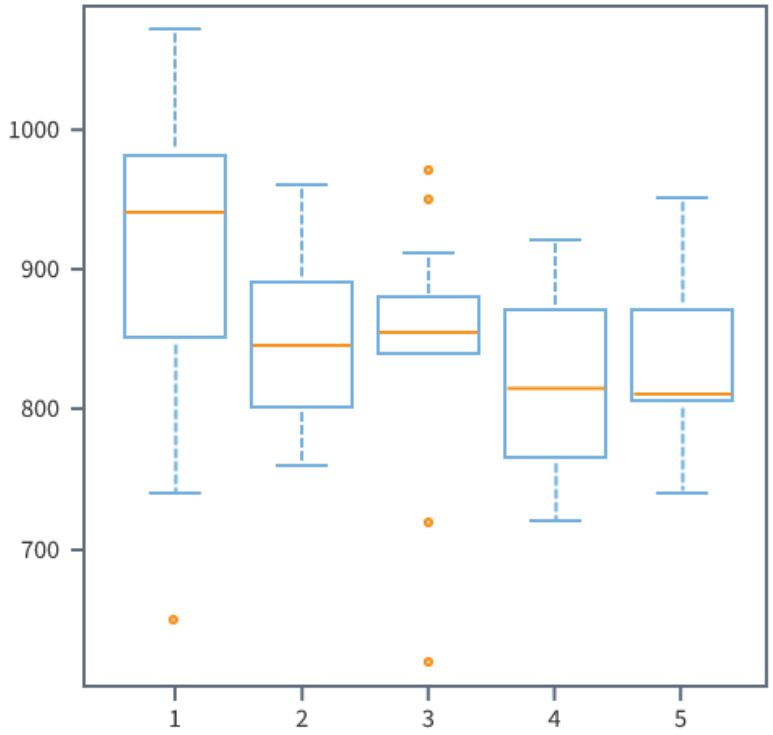
- ➔ максимальное наблюдаемое значение выборки (4-й квартиль или 100-й процентиль)
- ➔ медиана (2-й квартиль или 50-процентиль) — отображается чертой, разделяющей «ящик» на две части
- ➔ 1-й квартиль или 25-й процентиль — представляется левой стороной «ящика»
- ➔ 3-й квартиль или 75-й процентиль — представляется правой стороной «ящика»

Таким образом, длина "ящика" представляет собой межквартильный размах.

Длина "усов" на диаграмме представляет разброс (изменчивость) выборочных значений. Таким образом, расстояние между двумя концами "усов" представляет собой вариабельность. Если длины "усов" одинаковы, это указывает на симметричность распределения выборки. Если длины "усов" неравны, это указывает на то, что распределение асимметрично.

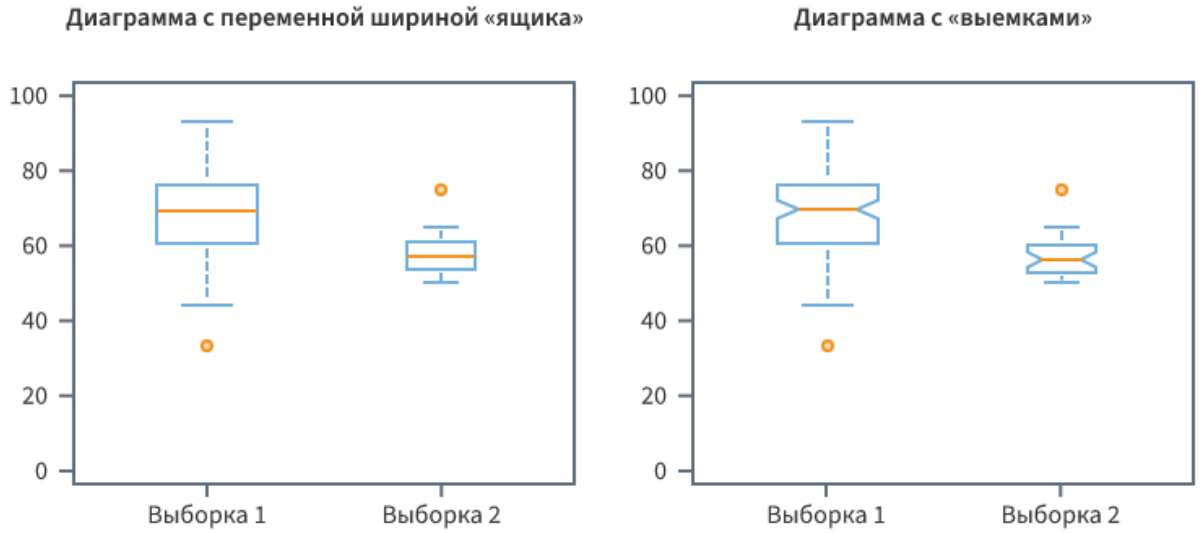
Круги или точки со звездочками на диаграмме указывают на выбросы в данных.

Обычно диаграмма "ящик с усами" строится вертикально, при этом несколько "ящиков" для разных образцов отображаются рядом. Это очень полезно для сравнения статистических свойств различных выборок.




Наиболее распространенными модификациями классической диаграммы являются:

- Диаграмма с переменной шириной «ящика» – используются, когда несколько "ящиков с усами" отображаются вертикально и сравниваются распределения нескольких выборок. В обычных графиках длина "ящика" представляет собой четверть интервала каждой выборки и имеет одинаковую ширину; в графиках переменной шириной "ящика" меняется в зависимости от размера выборки и обычно определяется как квадратный корень из числа элементов;
- Графики с "выемками" - "насечка" вокруг медианы, или узость ширины ящика, используется для получения приблизительного представления о значимости разницы между медианами; если две прямоугольные насечки не пересекаются, это указывает на статистически значимую разницу между медианами двух выборок. Ширина выемки пропорциональна межквартильному размаху выборки и обратно пропорциональна квадратному корню из ее размера.




💡 Описательная статистика — это процесс по умолчанию при анализе данных. Исследовательский анализ данных (EDA) не является полным без анализа описательной статистики.

# Корреляция и корреляционный анализ

**Корреляция и вычисления корреляции**

Корреляция — это простая взаимосвязь между двумя переменными в контексте, при которой одна переменная влияет на другую, ковариация или ассоциация между двумя или более переменными. Оно касается не изменений в x или y по отдельности, а измерения одновременных изменений в обеих переменных.

Рост родителей и детей коррелирует, но это нельзя объяснить тем, что рост ребенка определяется исключительно генетическими факторами. Существует еще несколько факторов, включая экологические и генетические.

Причинно-следственная связь является функциональной и, естественно, отражает корреляцию, но она не выходит за рамки исследования ковариации.

Различают парную, частную и множественную корреляцию.

Парная корреляция – это связь между двумя признаками (результативным и факторным или между двумя факторными).

Частная корреляция – это связь между двумя признаками (результативным и факторным или между двумя факторными) при фиксированном значении других факторных признаков.

Множественная корреляция – это связь между результативным и двумя или более факторными признаками, включенными в исследование.

**Меры корреляции**

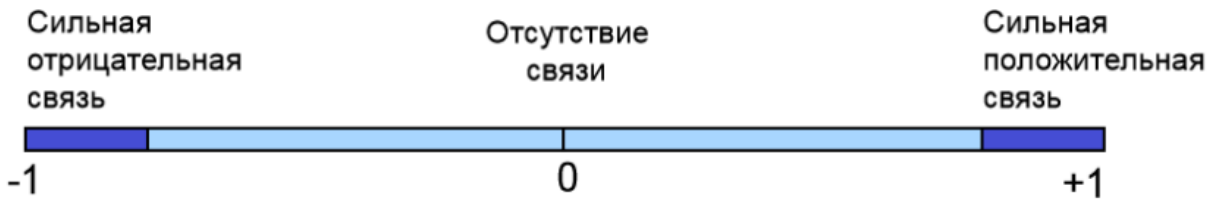
1. Коэффициент корреляции Пирсона — это мера силы линейной связи между двумя переменными, выраженная в виде r. По сути, корреляция Пирсона пытается провести линию наилучшего соответствия через данные двух переменных. Коэффициент корреляции Пирсона r показывает, насколько далеко все точки данных находятся от этой линии наилучшего соответствия.

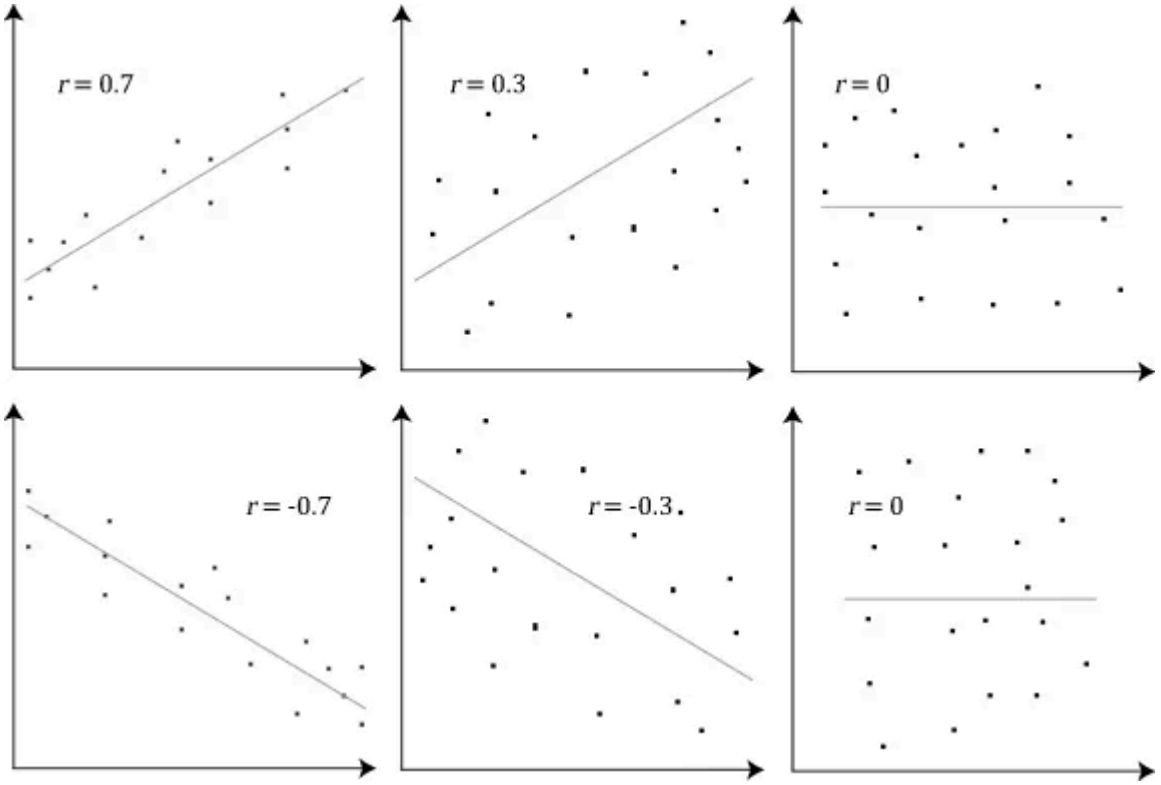
- В коэффициенте корреляции Пирсона переменные могут измеряться в совершенно разных единицах. Например, мы можем соотнести рост человека с его весом. Он разработан таким образом, что единица измерения не может повлиять на исследование ковариации.
- Коэффициент корреляции Пирсона (r) является безразмерной мерой корреляции и не изменяется при измерении происхождения или сдвига шкалы.
- При этом не учитывается, была ли переменная классифицирована как зависимая или независимая переменная. Оно одинаково обрабатывает все переменные. Возможно, мы захотим выяснить, коррелируют ли результаты в баскетболе с ростом человека. Но если мы определим, определялся ли рост человека его баскетбольными показателями (что не имеет смысла), результат будет тем же самым.

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Свойства:

1. Диапазон r находится в пределах [-1,1].
2. Вычисление r не зависит от изменения источника и масштаба измерения.
3. r = 1 (абсолютно положительная корреляция), r = -1 (абсолютно отрицательная корреляция), r = 0 (корреляции нет)



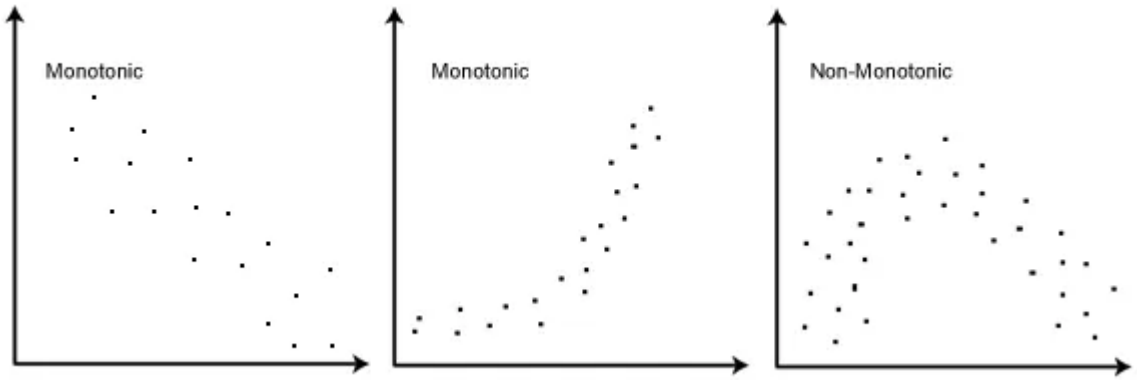

2. Коэффициент корреляции Спирмена

Коэффициент корреляции Спирмена — это непараметрический показатель силы и направления связи, которая существует между двумя категориальными переменными. Обозначается символом rs или ρ. Например:

Возможно, нам захочется выяснить корреляцию между рангами, присвоенными двум кандидатам на собеседовании, оценками, полученными группой студентов по пяти предметам, и т.д.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Корреляция Спирмена определяет силу и направление монотонной взаимосвязи между двумя переменными, а не силу и направление линейной зависимости между двумя переменными, как корреляция Пирсона.




Свойства:

1. Диапазон r находится в пределах [-1,1].
2. Сохраняет все свойства r.
3. Поскольку это основано на порядковых данных, это не зависит от какого-либо конкретного распределения (поэтому называется непараметрической мерой)

💡 Примечание: Корреляция Спирмена может использоваться, когда допущения корреляции Пирсона заметно нарушаются.

Корреляция — это широко применяемый метод машинного обучения при анализе данных и их интеллектуальном анализе. Он может извлекать ключевые проблемы из заданного набора функций, которые впоследствии могут нанести значительный ущерб во время подгонки модели.

Данные, имеющие некоррелированные характеристики, имеют много преимуществ. Например:

1. Изучение алгоритма будет быстрее
2. Интерпретируемость будет высокой
3. Смещение будет меньше

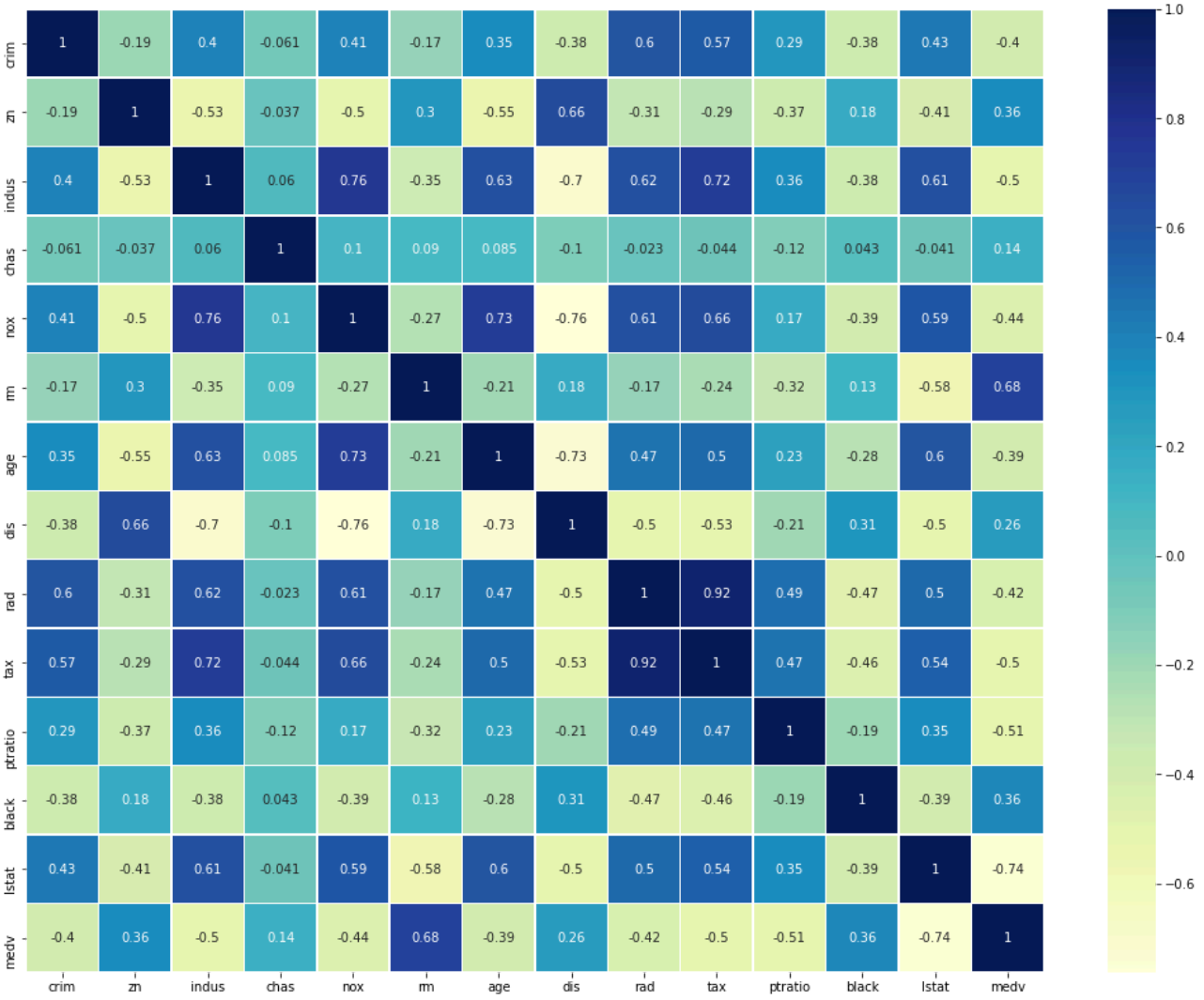
Посмотрим датасет о ценах на жилье в Бостоне:

```
df = pd.read_csv(r"\boston-dataset\boston_data.csv")
df.head()
```

| | crim | zn | indus | chas | nox | rm | age | dis | rad | tax | ptratio | black | lstat | medv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.15876 | 0.0 | 10.81 | 0.0 | 0.413 | 5.961 | 17.5 | 5.2873 | 4.0 | 305.0 | 19.2 | 376.94 | 9.88 | 21.7 |
| **1** | 0.10328 | 25.0 | 5.13 | 0.0 | 0.453 | 5.927 | 47.2 | 6.9320 | 8.0 | 284.0 | 19.7 | 396.90 | 9.22 | 19.6 |
| **2** | 0.34940 | 0.0 | 9.90 | 0.0 | 0.544 | 5.972 | 76.7 | 3.1025 | 4.0 | 304.0 | 18.4 | 396.24 | 9.97 | 20.3 |
| **3** | 2.73397 | 0.0 | 19.58 | 0.0 | 0.871 | 5.597 | 94.9 | 1.5257 | 5.0 | 403.0 | 14.7 | 351.85 | 21.45 | 15.4 |
| **4** | 0.04337 | 21.0 | 5.64 | 0.0 | 0.439 | 6.115 | 63.0 | 6.8147 | 4.0 | 243.0 | 16.8 | 393.97 | 9.43 | 20.5 |

Посмотрим на тепловую карту матрицы корреляции

```
df_corr = df.corr(method='pearson', min_periods=1)
pt.figure(figsize=(18,14))
sns.heatmap(df_corr, xticklabels = df_corr.columns, yticklabels =
df_corr.columns, annot=True,
            linewidths=0.5, cmap = "YlGnBu")
```




Наблюдения

Признаки “tax” и “rad” имеют высокую корреляцию со значением 0,92 (положительная корреляция).
Некоторые функции имеют отрицательную корреляцию, и значение их корреляции отрицательно велико. Например, “lstat” против “medv”, “dis” против “indus”, “dis” против “age”.

Из приведенных выше наблюдений мы можем заключить, что существует скрытая ковариация между налоговой ставкой и индексом доступности радиальных магистралей. Мы можем вывести зависимость, согласно которой, если доступность дома к автомагистралям высока, то ставка налога на недвижимость в полном объеме также будет выше.

Это логично, потому что дома с высокой ценой, как правило, ближе к рынку, хорошие удобства, автомагистрали и т.д.

Посмотрим на коэффициент корреляции Спирмена (Корреляция Спирмена — это непараметрическая версия коэффициента корреляции Пирсона, которая измеряет степень связи между двумя переменными на основе их рангов.)

```
df_corr = df.corr(method='spearman', min_periods=1)
pt.figure(figsize=(18,14))
sns.heatmap(df_corr, xticklabels = df_corr.columns, yticklabels =
df_corr.columns, annot=True,
            linewidths=0.5)
```


Наблюдения

“chas” - это категориальный признак, и поскольку мы принимаем во внимание коэффициент корреляции Спирмена, он также был включен в корреляцию.

🔥 В случае числовых признаков всегда используйте коэффициент корреляции Пирсона.

**Эффект мультиколлинеарности** (Мультиколлинеарность — явление, при котором наблюдается сильная корреляция между признаками).

Ключевой целью регрессионного анализа в машинном обучении является выделение взаимосвязи каждой независимой переменной и зависимой переменной. Таким образом, изменение одной независимой переменной не должно влиять на какие-либо другие переменные в данных. Однако, когда независимые переменные коррелируют, это указывает на то, что изменения одной переменной связаны со сдвигами в другой переменной. По мере увеличения серьезности мультиколлинеарности увеличивается влияние этой проблемы.

Таким образом, во время подгонки модели небольшое изменение в одной переменной может привести к значительному изменению выходных данных модели. Однако эти проблемы затрагивают только те независимые переменные, которые коррелированы.

**Решение:**

Серьезность проблем возрастает со степенью мультиколлинеарности. Следовательно, если у вас только умеренная мультиколлинеарность ее не понадобиться устранять.

Мультиколлинеарность влияет только на конкретные независимые переменные, которые коррелированы. Следовательно, если мультиколлинеарность отсутствует для независимых переменных, которые вас особенно интересуют, возможно, вам не нужно ее разрешать. Предположим, что ваша модель содержит интересующие экспериментальные переменные и некоторые контрольные переменные.

Если для контрольных переменных существует высокая мультиколлинеарность, но не для экспериментальных переменных, вы можете интерпретировать экспериментальные переменные без проблем.

Если одна из коллинеарных функций не имеет большого вклада в прогнозирование или классификацию для алгоритмов, основанных на расстоянии, мы можем отказаться от функции анализа.

# Что можно почитать еще?

1. Exploratory Data Analysis(EDA) In Python
2. First Step in EDA: Descriptive Statistics Analysis
3. Машинное обучение — 1. Корреляция и регрессия.

# Используемая литература

1. https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

2. https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php
3. https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/
4. https://www.kaggle.com/code/shoose80/russian-notes-pythondataanalysis-5-2/edit