

Описательные статистики в контексте EDA. Корреляция и корреляционный анализ

Урок 2

Мы рассмотрим различные методы анализа данных, такие как:

Описательная статистика

Корреляция и корреляционные методы



Булгакова Татьяна

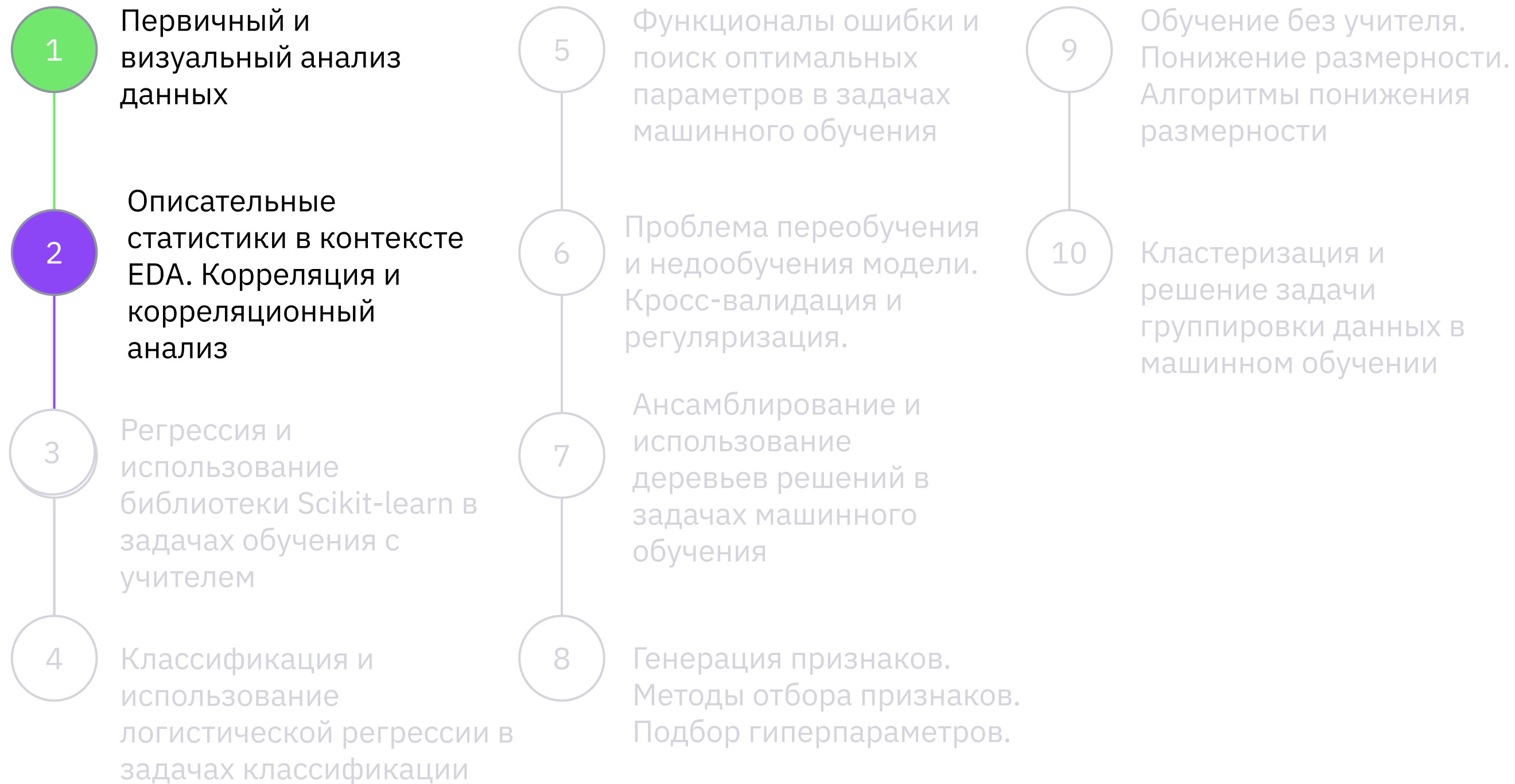
Преподаватель в GeekBrains, Нетология, Skillfactory

С 2010 года занимаюсь DataScience и NN. Фрилансер

- Участвовала в разработке программы по настройке оборудования для исследования пространственного слуха китообразных НИИ ИПЭЭ РАН
- Участвую в разработке рекомендательных систем по настройке нейростимуляторов для медицинских центров
- Работаю над курсом по нейронным сетям



План курса





Что будет на уроке сегодня



Описательная статистика



Корреляция и корреляционный анализ



Описательная статистика

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





Ключевые идеи:

Статистика

наука о данных

Данные

набор наблюдений за
интересующей нас
генеральной совокупностью

Статистика

предоставляет конкретный способ
сравнения генеральных совокупностей с
помощью чисел, а не неоднозначных
описаний.



Описательная статистика



Меры центральной тенденции эти числа описывают, где расположен центр набора данных. Примеры включают среднее и медиана .



Меры дисперсии

эти числа описывают, насколько разбросаны значения в наборе данных. Примеры включают размах , межквартильный размах , стандартное отклонение и дисперсию .



Мера центральной тенденции

Среднее арифметическое значение - сумма значений признака, деленная на общее количество объектов, называется средним значением.

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

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




Мера центральной тенденции

Медиана - центральное значение атрибута известно как медиана. Чтобы вычислить медианное значение, сначала отсортируйте данные столбца в порядке возрастания или убывания.

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



Мера центральной тенденции

Мода - это то значение, которое чаще всего встречается (самое модное значение).

```
1 # mode
2 u = [2, 3, 2, 8, 12, 6, 4, 2, 8]
3 mode = max((u.count(item), item) for item in set(u))[1]
4 mode
5
```



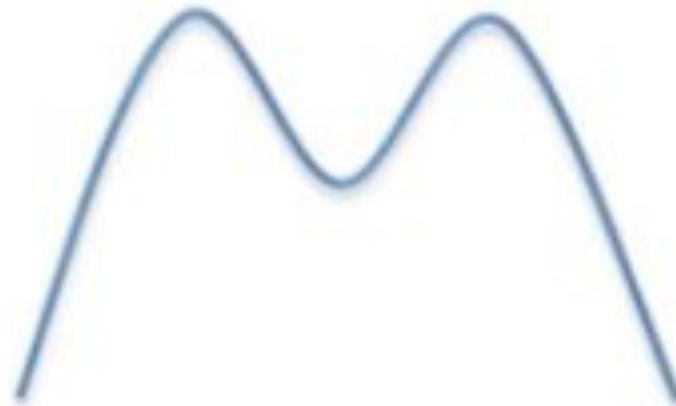
Мера центральной тенденции

Мода - это то значение, которое чаще всего встречается (самое модное значение).

Unimodal



Bimodal



Multimodal





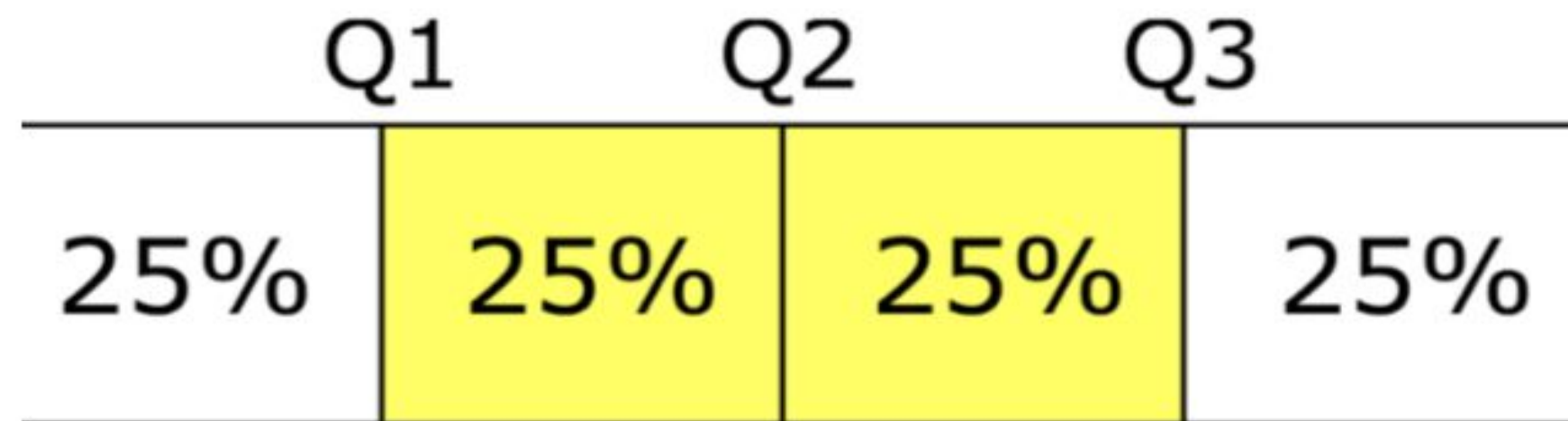
Метрики вариативности данных

- ✦ **Межквартильный диапазон (IQR)** является мерой статистического разброса между верхним (75-м) и нижним (25-м) квартилями.
- ✦ **Дисперсия** - среднеквадратичное отклонение значений от среднего арифметического, показывающее разброс данных относительно него.
- ✦ **Стандартное отклонение** представляет собой квадратный корень из дисперсии



Метрики вариативности данных

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



Interquartile Range
 $= Q3 - Q1$

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




Метрики вариативности данных

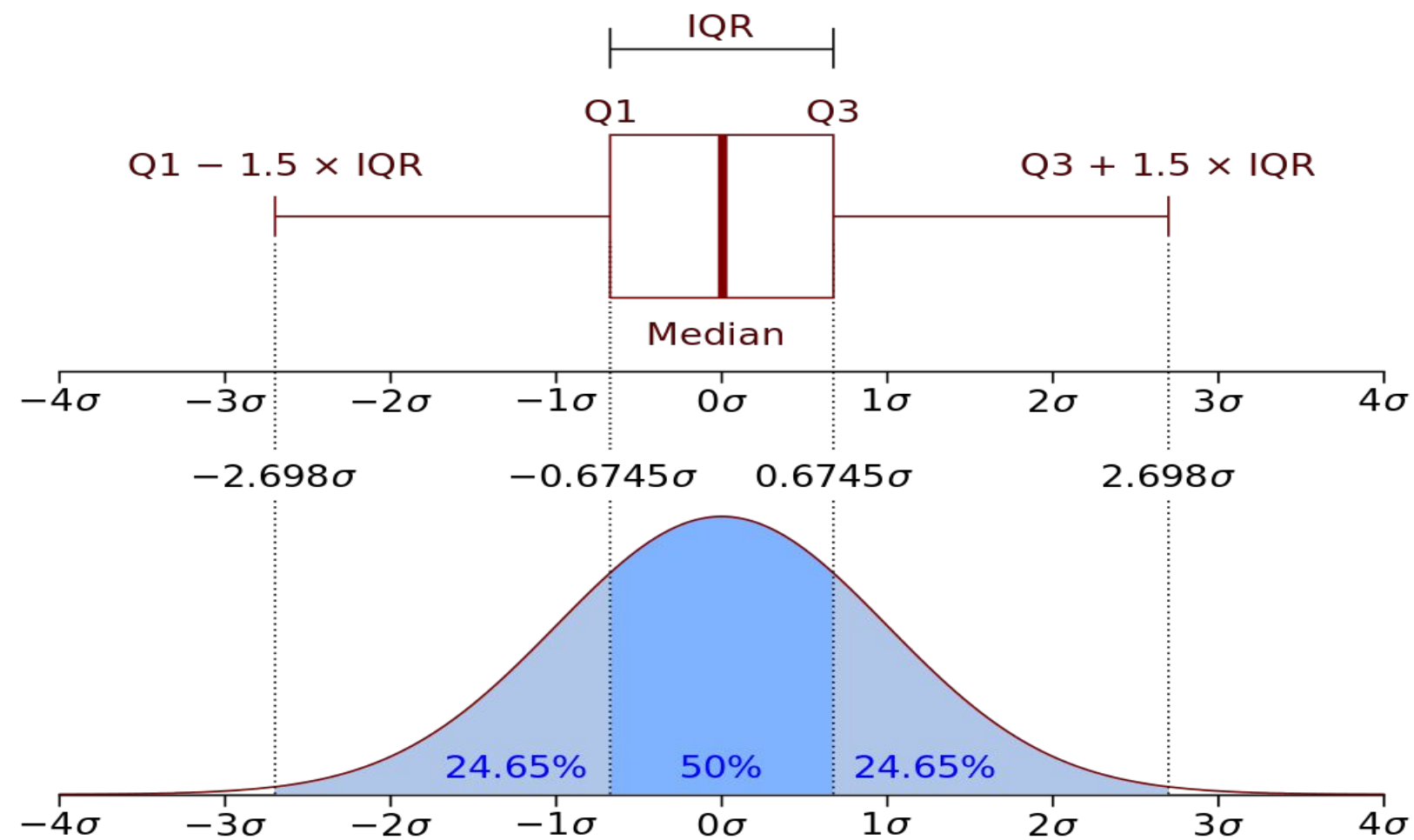
Дисперсия - среднеквадратичное отклонение значений от среднего арифметического, показывающее разброс данных относительно него.

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

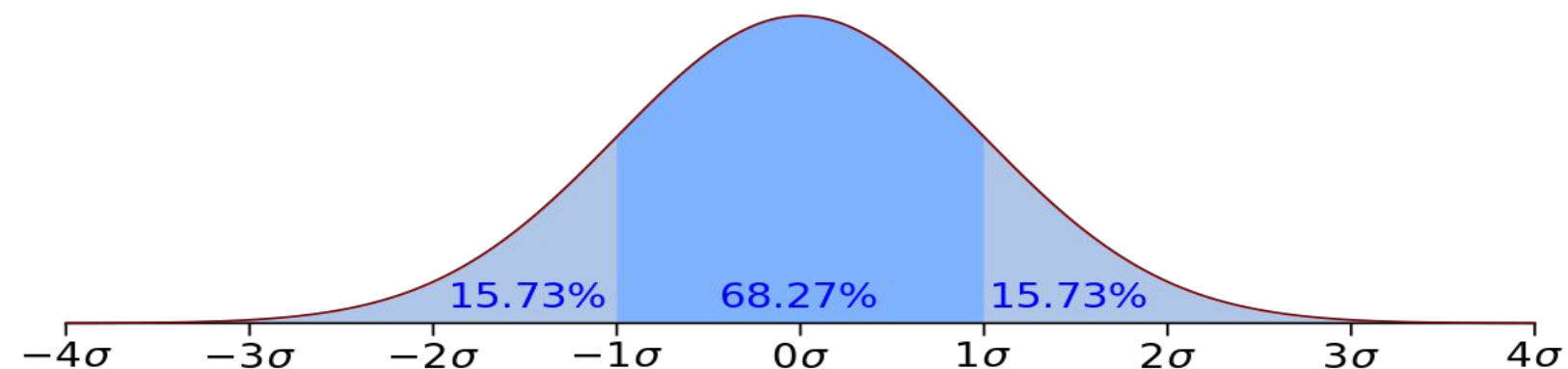
```
1 n = len(x)
2 mean = sum(x) / n
3 var = sum((x - mean)**2 for x in x) / (n - 1)
4 var
```



Описательная статистика. Визуализация



Так как мы знаем, что 99,7 процентов наблюдений лежат в пределах трех СКО от среднего, то можем предположить, что выбросами будут оставшиеся 0,3 процента.





Описательная статистика. Визуализация

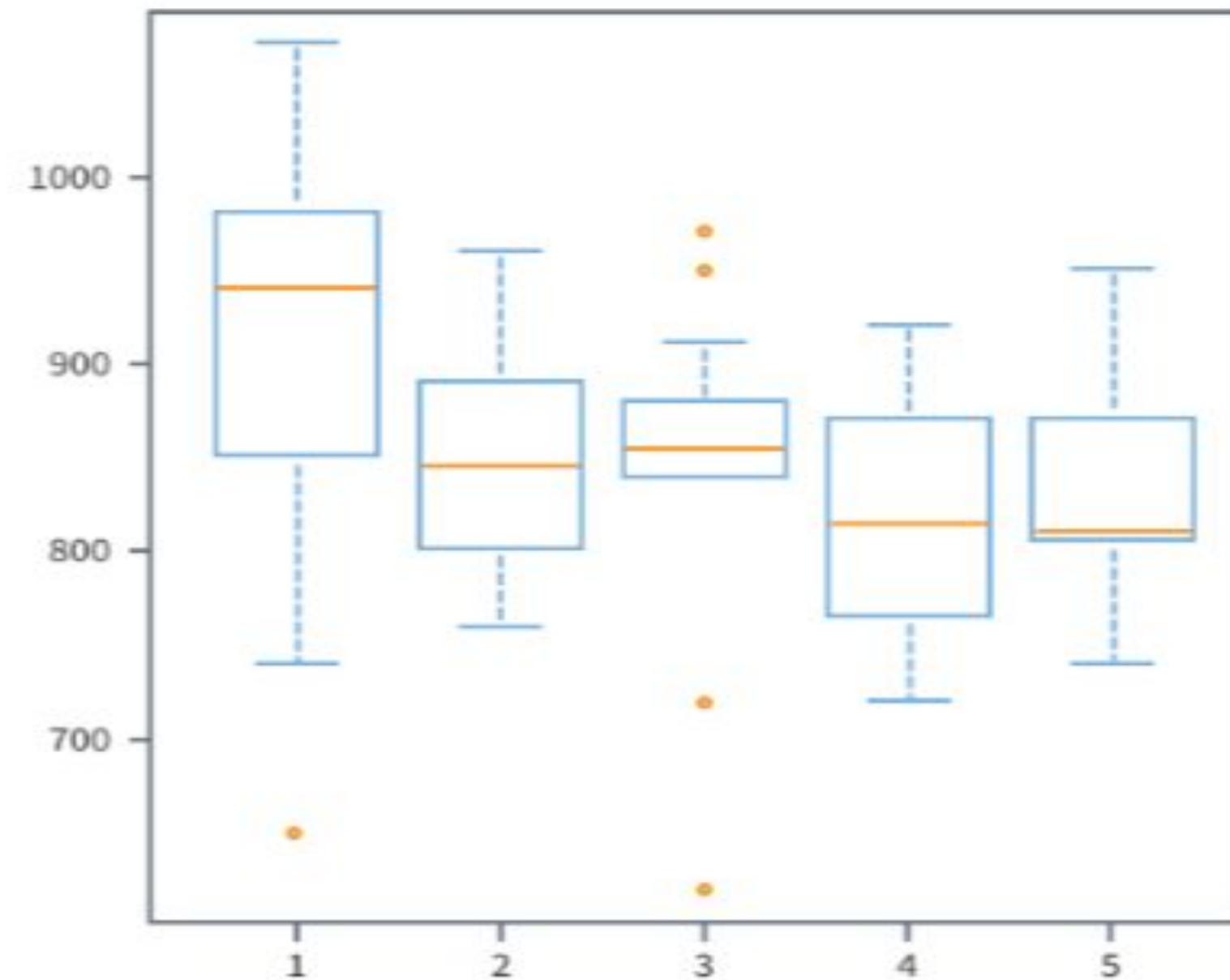




Диаграмма с переменной шириной ящика

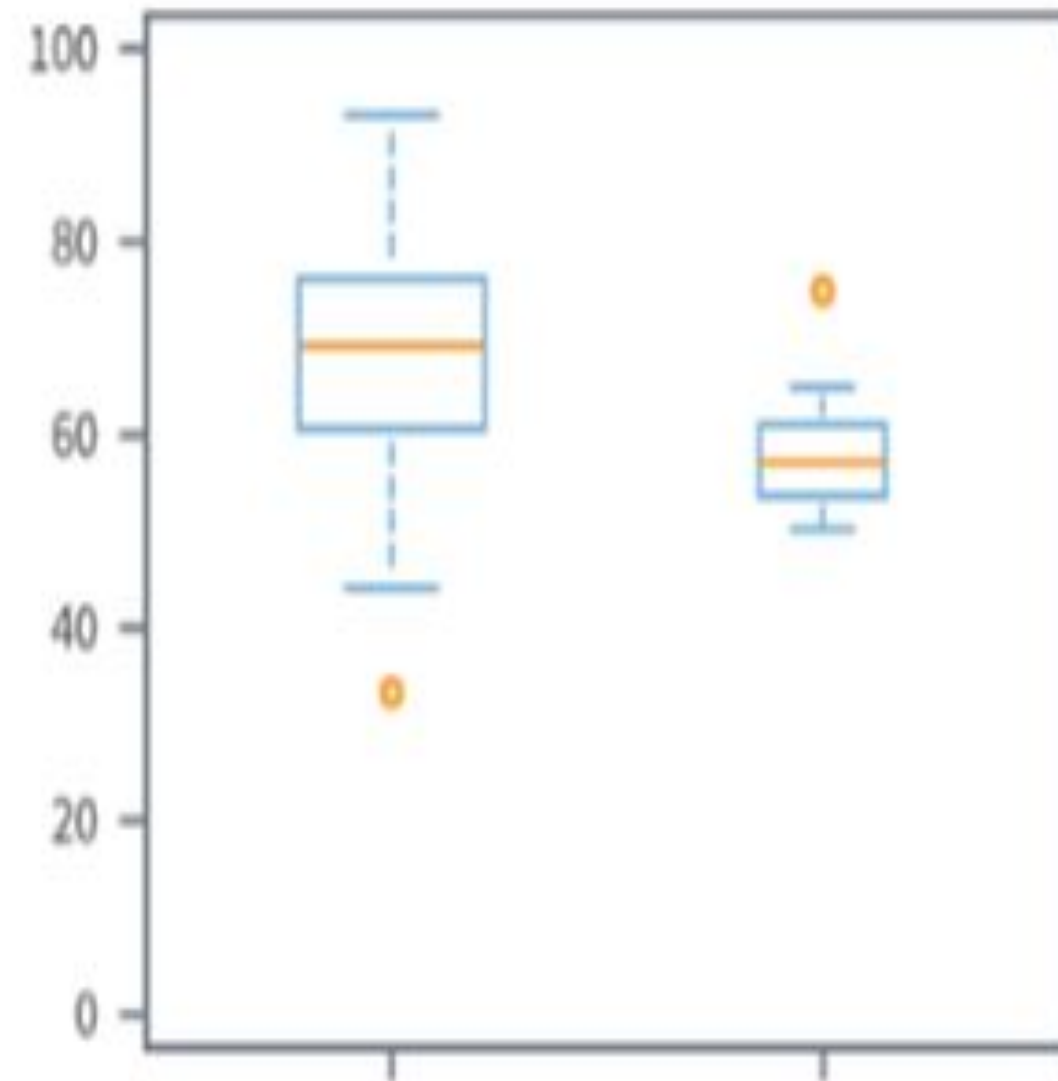
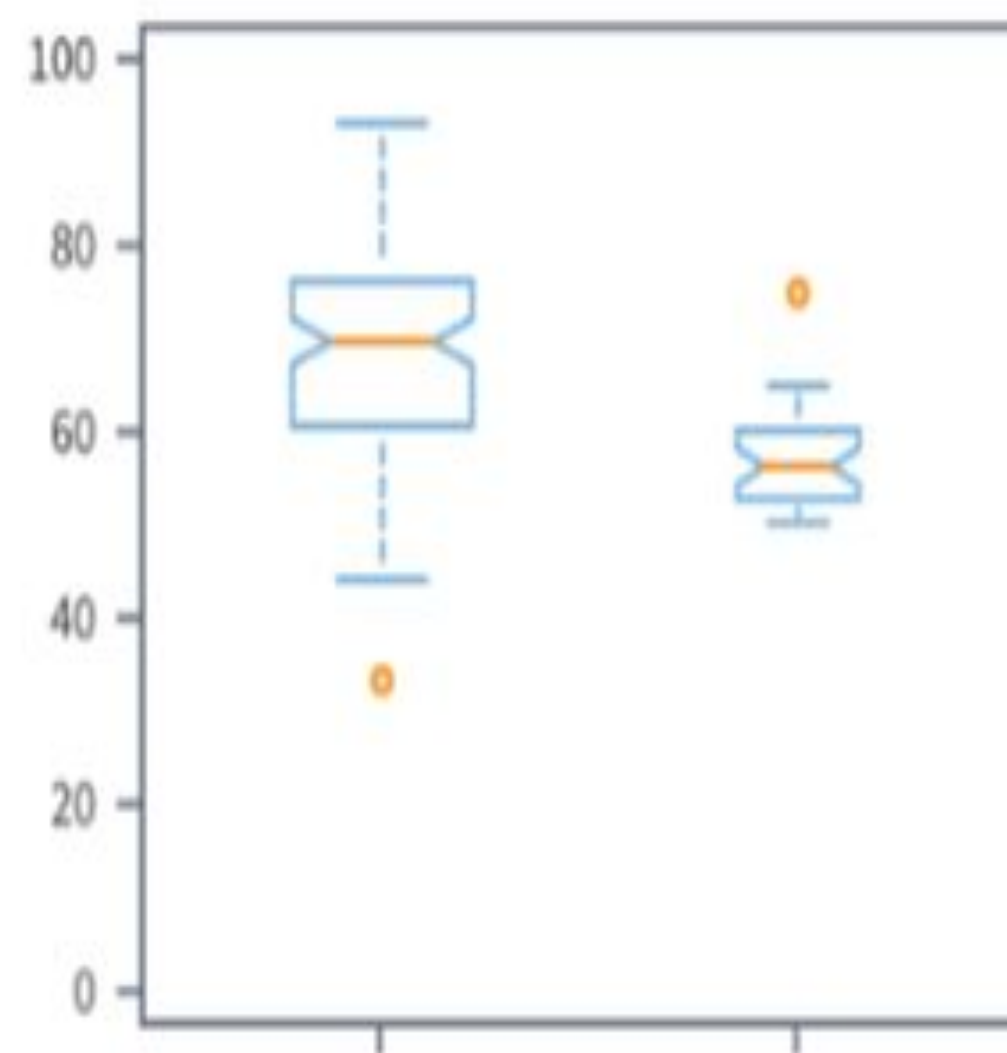


Диаграмма с выемками





Корреляция и корреляционный анализ

Корреляция - это простая взаимосвязь между двумя переменными в контексте, при которой одна переменная влияет на другую, ковариация или ассоциация между двумя или более переменными.

Меры корреляции



Коэффициент корреляции Пирсона



Коэффициент корреляции Спирмена



Коэффициент корреляции Пирсона

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.

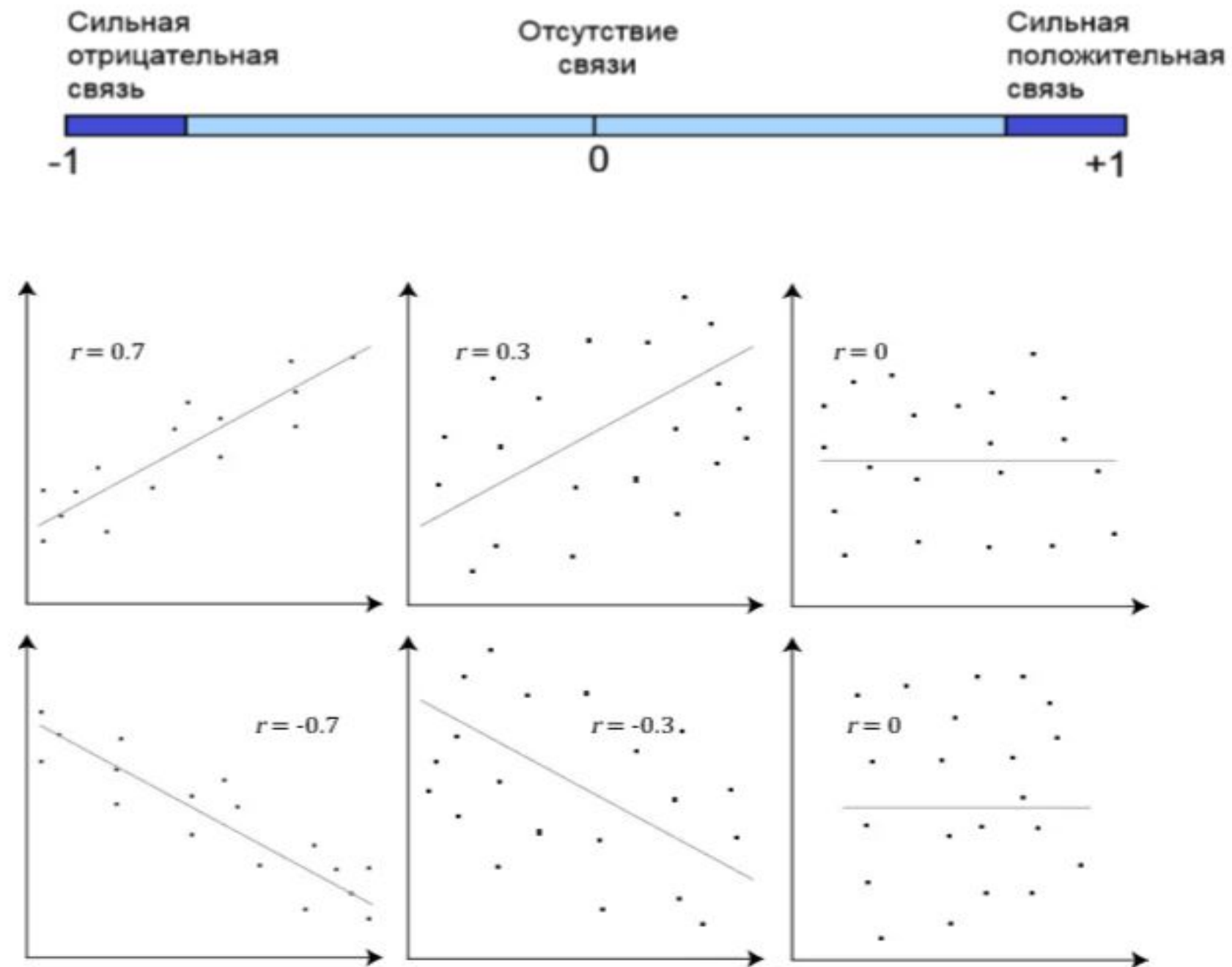
$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Свойства:

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



Коэффициент корреляции Пирсона





Коэффициент корреляции Спирмена

Коэффициент корреляции Спирмена - это непараметрический показатель силы и направления связи, которая существует между двумя категориальными переменными

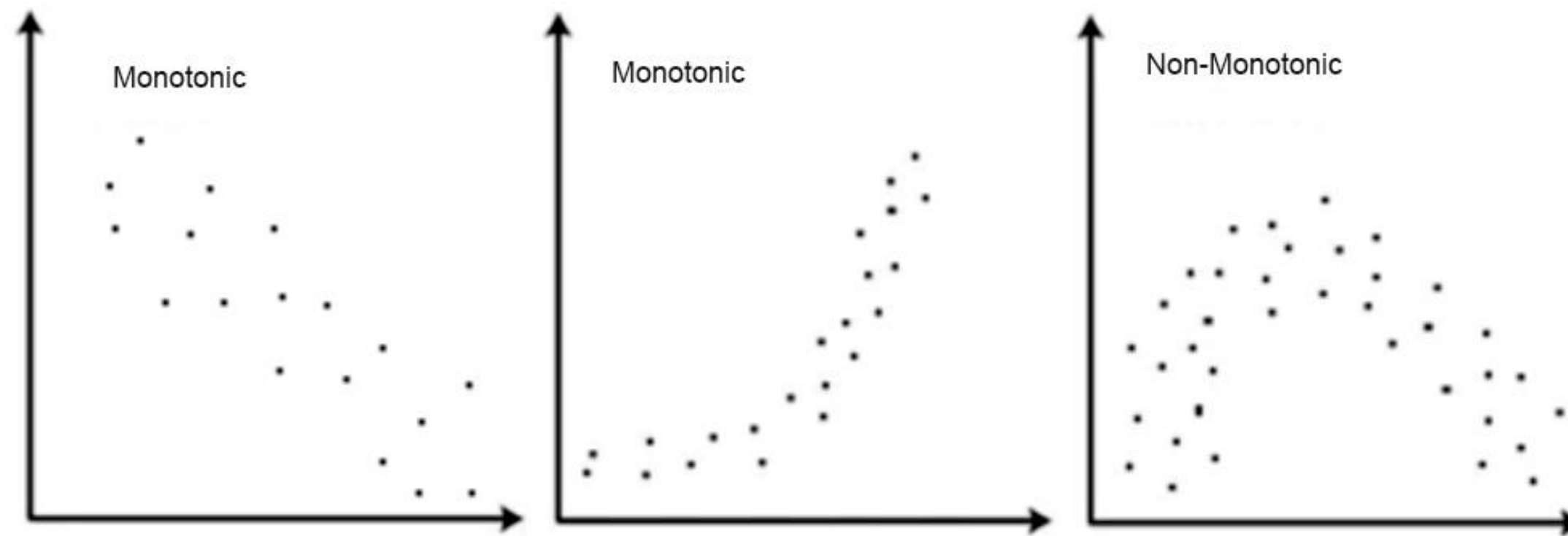
$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Свойства:

- Диапазон r находится в пределах $[-1, 1]$.
- Сохраняет все свойства r .
- Поскольку коэффициент основан на порядковых данных, он не зависит от какого-либо конкретного распределения (поэтому называется непараметрической мерой)



Коэффициент корреляции Спирмена





Корреляция и корреляционный анализ

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






Решение:

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

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








Описательная статистика позволяет:

-  Определить основные характеристики данных, такие как центральную тенденцию и разброс значений.
-  Изучить распределение данных и определить, являются ли они нормальными, скошенными или имеют аномальные значения.
-  Определить наличие выбросов (значений, сильно отклоняющихся от остальных данных) и разработать стратегию их обработки.
-  Определить корреляцию между переменными и изучить зависимость между ними.
-  Сравнить характеристики у разных групп данных и выявить различия.



Корреляционный анализ можно использовать для:

-  Определения сильных и слабых связей между переменными
-  Отбора наиболее значимых переменных для модели
-  Исключения мультиколлинеарности между переменными в модели
-  Обнаружения выбросов и ошибок в данных
-  Построения графиков и визуализации зависимостей между переменными.



Спасибо за внимание

