**Лекция №3 Разведочный анализ или EDA   
(exploratory data analysis)**

**Цель лекции:**

* познакомиться с основными понятиями выборки и генеральной совокупности
* научиться рассчитывать и интерпретировать параметры описательной статистики
* ознакомиться с принципами построения графиков
* научиться читать графики, извлекать полезную информацию из них

**Материал прошлого урока:**

Прошлое занятие было посвящено дискретным распределениям: биномиальному и распределению Пуассона. Закончили мы изучение этих распределений расчетами параметров описательной статистики. Сегодня изучим общие формулы для этих величин и будем их использовать далее, работая с нормальным распределением, которое является, наверное, самым важным распределением статистики. Но о нем речь пойдет на следующем занятии.

**План урока:**

1. Понятия
2. Вероятность VS Статистика
3. Математическое ожидание
4. Дисперсия и среднее квадратичное отклонение
5. Смещенное и несмещенное среднее квадратичное отклонение в Python
6. Медиана и мода
7. Квартили, перцентили
8. Размах
9. Графический анализ
10. Правила визуализации

На прошлом уроке, при изучении дискретных распределений, мы уже затронули описательную статистику. Описательная статистика помогает нам узнать, как можно больше о случайной величине (далее СВ), которую мы изучаем.

С помощью математического ожидания мы можем узнать, вокруг какого значения группируется большая доля значений СВ, с помощью дисперсии оценить рассеянность данных вокруг среднего арифметического (математического ожидания).

Мы познакомились со «специфическими» формулами для среднего арифметического и для дисперсии дискретных СВ. Сегодня посмотрим их в общем виде в рамках изучения темы разведочного анализа.

Разведочный анализ, основоположником которого является американский математик Джон Тьюки, является важной неотъемлемой частью всего статистического анализа. Это может быть достаточно большой пласт работы, которым не стоит пренебрегать. Потому что, в ходе разведочного анализа, вы можете обнаружить выбросы, которые, как мы помним, очень влияют на среднее арифметическое. Они также влияют на дисперсию. Мы можем обнаружить и другие интересные моменты, а также предварительно оценить результат, который мы хотим подтвердить тестированием гипотез, разведочный анализ поможет выбрать правильные методы статистического анализа.

Если не учесть все эти нюансы, то выводы, полученные при статистическом анализе, будут очень далеки от реального положения вещей. Более того, иногда разведочного анализа бывает достаточно, чтобы сделать выводы и не прибегать к каким-то дальнейшим вычислениям.

Разведочный анализ или EDA от английского названия exploratory data analysis включает в себя описательную статистику и графический анализ. Начнем с первой его части, с описательной статистики, но для начала введем некоторые понятия

**Понятия**

**Генеральная совокупность - это множество, которое содержит данные обо всех объектах, соответствующих определенным характеристикам.**

Например, мы хотим измерить рост всех людей России. Если мы сможем измерить абсолютно каждого, то мы получим данные обо всей генеральной совокупности. Но в реальной жизни обычно у нас нет доступа ко всей генеральной совокупности, поскольку, во-первых, измерить абсолютно всех людей России займет очень много времени, во-вторых, это дорого, потому что нужно нанимать людей и платить, а, в - третьих, есть такие СВ, при измерении которых, нельзя вернуть объект обратно в генеральную совокупность. Например, краш-тест автомобиля. Мы не можем перебить всю партию и отправить её на продажу. Та же история и с испытаниями на износ деталей.

Когда мы измеряем всю генеральную совокупность, то мы проводим сплошной контроль. Но как вы видите, в реальной жизни это совсем невыгодно. Поэтому практически всегда мы прибегаем к выборочному контролю, т.е. берем репрезентативную (обычно «максимально» случайную) выборку.

**Выборка - это случайным образом выбранная часть генеральной совокупности.**

По выборке мы можем получить точечные оценки истинных значений параметров генеральной совокупности. Т.е. чтобы узнать, например, среднее генеральной совокупности, мы должны измерить все ее объекты. Но т.к. обычно доступа у нас нет ко всем объектам по выше перечисленным причинам, то мы берем репрезентативную выборку необходимого объема и измеряем среднее по ней. Это и будет точечная оценка истинного математического ожидания.

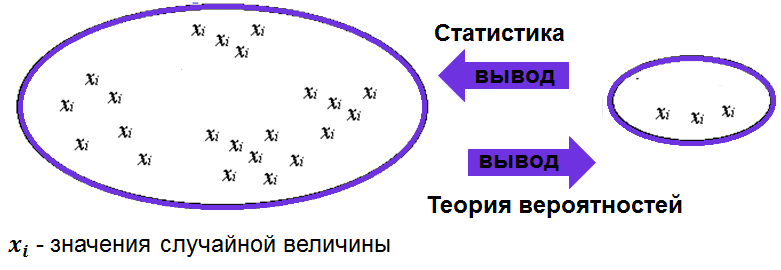
В зависимости от ситуации данные могут быть выборкой, а могут быть и генеральной совокупностью. Например, если мы хотим измерить рост всех людей на Земле и взяли людей из города Х, то люди, живущие в городе Х, будут являться выборкой. Но, если мы хотим оценить рост людей, живущих в этом городе Х, и взяли случайных 100 человек, то теперь все люди города Х будут генеральной совокупностью.

**Вероятность VS Статистика**

Чем же отличается теория вероятностей от статистики?

В теории вероятностей нам уже известна математическая модель. И мы ее используем для того, чтобы предсказать, какой результат мы увидим. Например, вероятность, что выпадет орел составляет ½, мы используем это значение, чтобы предсказать исход броска.

А в статистике мы не знаем поведения генеральной совокупности, которую мы изучаем. И набрав статистические данные, мы ищем закономерности, подбирая математическую модель, и затем эту модель экстраполируем на всю генеральную совокупность. Т.е. с помощью подобранной модели мы будем описывать поведение всей популяции.



**Математическое ожидание**

Одной из самых важных характеристик описательной статистики является математическое ожидание, которое находится, как среднее арифметическое всех элементов. Иногда его еще называют «среднее». В реальной жизни мы обычно измеряем этот параметр по выборке.

Недостатком среднего арифметического, напомню, является то, что оно очень чувствительно к выбросам, особенно при небольших объемах выборки.

Пример:

Заработные платы:

Получили среднее арифметическое

А мы помним, что первоочередная задача математического ожидания - это описать, а вокруг какого значения группируется большая доля значений СВ. Но в данной ситуации математическое ожидание не справилось с этой задачей, потому что мы в маленькой выборке имеем выброс (257). Получается, что 80% значений выборки (4 из 5) лежат почти в 2 раза ниже среднего арифметического.

В несимметричных распределениях можно пользоваться другой мерой центральной тенденции – медианой. О ней поговорим чуть позже.

**Дисперсия и среднее квадратичное отклонение**

Знакомясь с дискретными распределениями, мы уже давали определение дисперсии.

**Дисперсия характеризует степень рассеянности значений случайной величины относительно ее математического ожидания.**

Обозначается D(X) или

Общая формула для нахождения дисперсии:

Эта формула для дисперсии генеральной совокупности, но опять же, в реальной жизни доступа ко всей генеральной совокупности у нас нет, поэтому и этот параметр мы оцениваем обычно по репрезентативной выборке. Для выборки дисперсию обычно обозначают , но иногда в книгах можно встретить и для дисперсии по выборке. Понять, о какой именно дисперсии идет речь можно по контексту.

И работая с выборками, мы сталкиваемся с такими понятиями, как смещенная и несмещенная дисперсия.

Если объем выборки менее 100, то обязательно используют формулу несмещенной дисперсии.

Сравним с формулой смещенной дисперсии

, n – объем выборки

Проще, просто всегда использовать формулу несмещенной дисперсии.

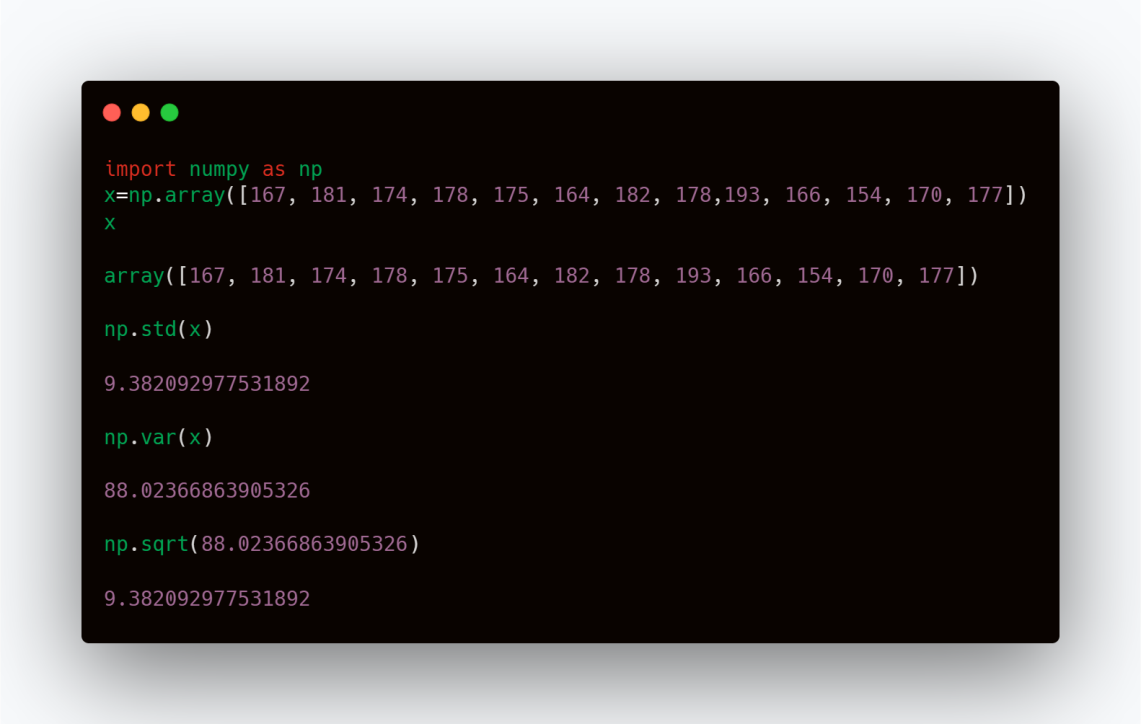
Дисперсия – квадратичная величина, поэтому пользоваться ей не очень удобно. В статистике используют квадратный корень из дисперсии, стандартное отклонение. Его еще называют среднее квадратичное отклонение (𝞼 –для генеральной совокупности, S – для выборки).

**Среднее квадратичное отклонение (далее СКО) показывает, насколько далеко наблюдения могут быть "разбросаны" относительно среднего значения.**

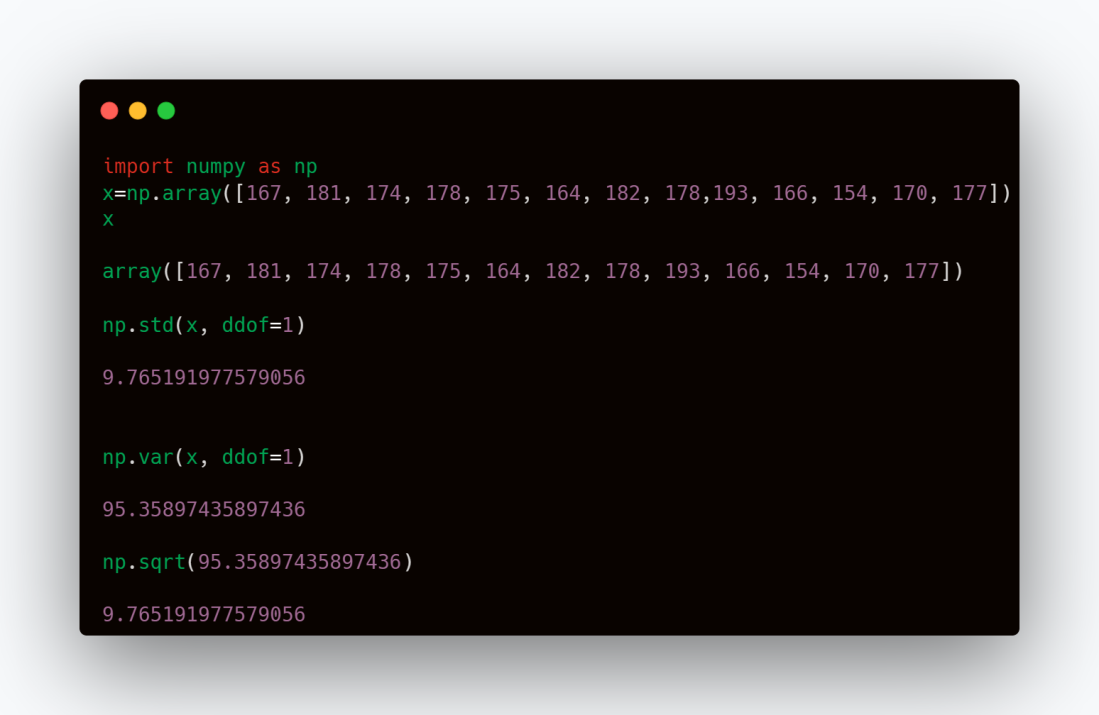
**Смещенное и несмещенное среднее квадратичное отклонение в Python**

Для вычисления СКО используется функция std(). std расшифровывается как standard deviation, что в переводе означает стандартное отклонение. Функция var() используется для расчета дисперсии (на английском дисперсия – variance, отсюда и название функции). По умолчанию Python рассчитывает смещенные параметры. Чтобы рассчитать несмещенные значения, в функции нужно прописать параметр ddof=1 (степени свободы).

Посчитаем смещенное стандартное отклонение и дисперсию для выборки Х



А теперь посчитаем несмещенное стандартное отклонение и дисперсию для выборки Х

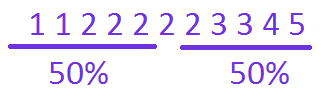


**Медиана и мода**

Не всегда мы работаем с симметричными распределениями. Поэтому в качестве меры центральной тенденции иногда используют медиану. Она не чувствительна к выбросам и показывает, ниже какого значения лежат 50% наблюдений, т.е.

**медиана – это значение, которое делит выборку на две равные части так, что значения, которые меньше медианы, составляют 50% выборки**.

Чтобы посчитать медиану нужно сначала разместить значения по возрастанию. Медиана немного по-разному рассчитывается для четных и нечетных размеров выборки.

Возьмем нечетную выборку объемом 11. Сначала разместим значения по возрастанию. Ниже 6-го элемента (число 2) лежит 50% наблюдений , т.е. 5 значений и выше тоже 50% наблюдений, 5 значений.

Чтобы найти этот 6-й элемент мы воспользуемся формулой где n –объем выборки, а N – элемент, соответствующий медиане.

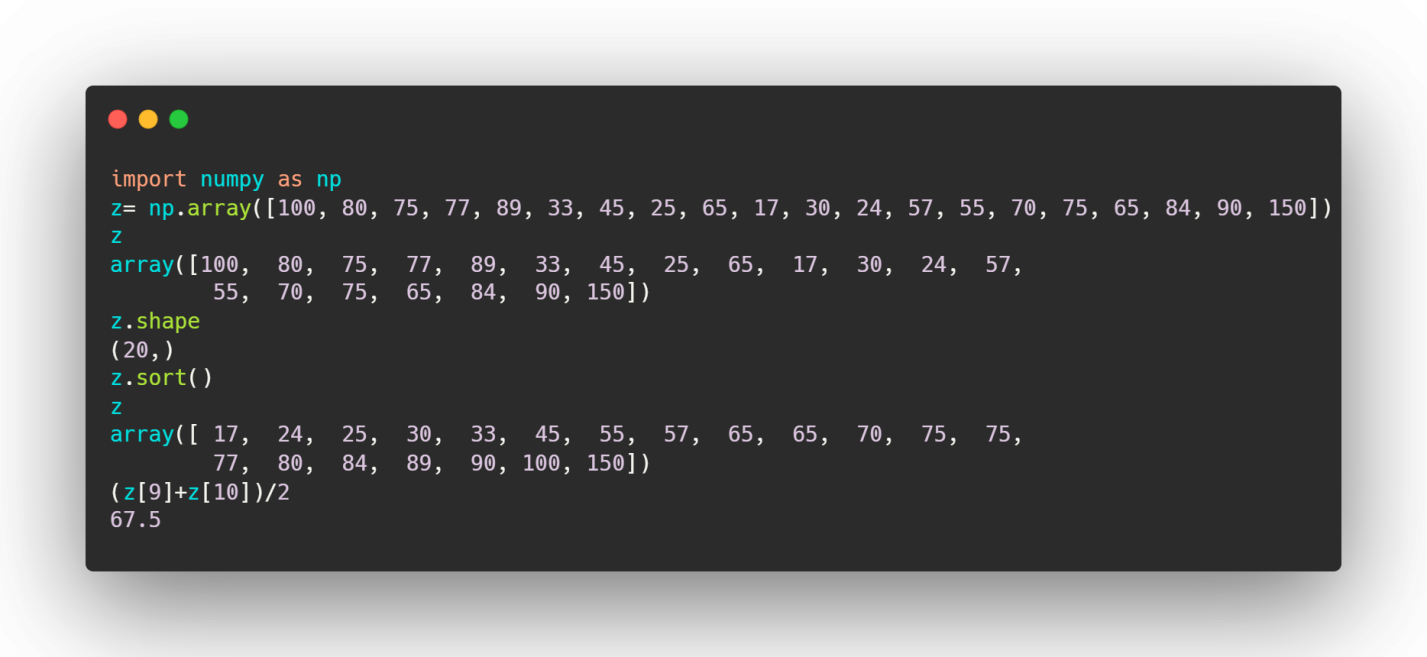
(11+1)/2=6 - номер элемента, который является медианой в выборке нечетного размера.

Для четной выборки формула:

n – четный объем выборки. Медиана для ряда 1,3,7,8 равна 5



**Рассчитаем медиану в Python**



Еще одной мерой центральной тенденции является мода. Используется для категориальных данных. Например, голосование за кандидатов Иванова, Петрова, Сидорова. Кто наберет большее число голосов, тот и будет модой.

**Квартили, перцентили**

Еще важные параметры, которые нам очень помогают в разведочном анализе, - это перцентили.

Если речь идет о каждом 25-ом перцентиле, мы называем их квартилями, а о каждом 10-ом, то децентили

Первый квартиль показывает значение, ниже которого лежат 25% наблюдений.

Второй квартиль - это синоним медианы.

Третий квартиль показывает значение, ниже которого лежат 75% наблюдений.

Т.е. сначала вам нужно отсортировать значения по возрастанию, а потом вычислить, ниже какого значения лежат 25% наблюдения или 50%, или 75%.

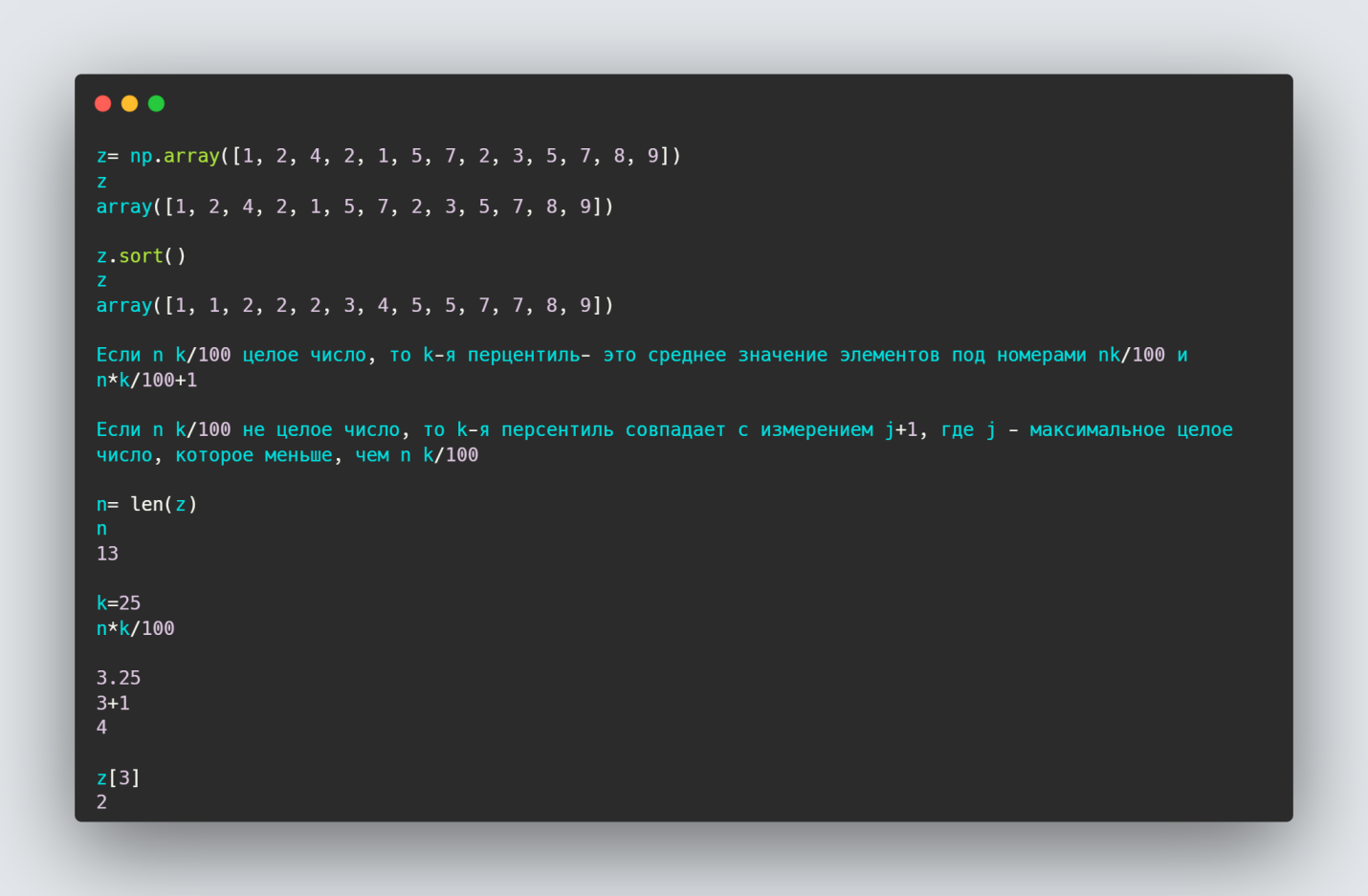
Интерквартильное расстояние (еще его называют межквартильное ) - отрезок, равный разности 3-го и 1-го квартиля.

Есть формулы, которые позволяют рассчитать нам любой перцентиль.

Если n\*k/100 – целое число, тогда k-я перцентиль – это среднее значение элементов под номерами n\*k/100 и (n\*k /100)+1. Здесь n – объем выборки, k – перцентиль.

Если n\*k/100 – нецелое число, то k-я перцентиль совпадает с измерением j+1, где j – максимальное целое число, которое меньше, чем n\*k/100. Здесь n – объем выборки, k – перцентиль.

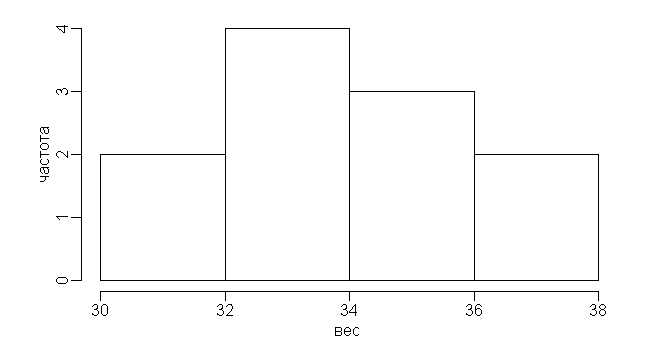
Посчитаем 1 квартиль (25%) в Python.



Поскольку нумерация в Python с нуля, мы берем не 4-й, а 3-й индекс.

**Размах**

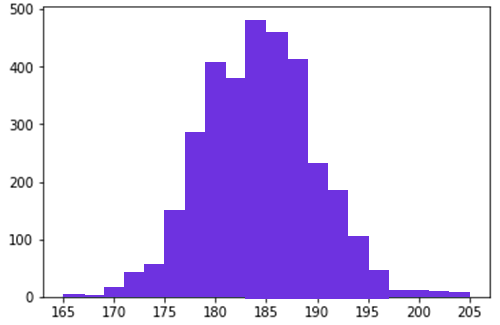
**Размах** - очень чувствительный к выбросам параметр, т.к. это разность между максимальным и минимальным значением. Например, используется для построения гистограмм



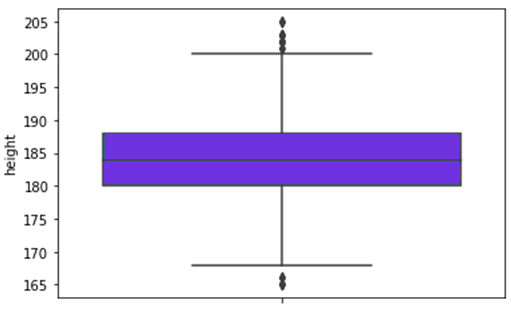
**Графический анализ**

Важной частью разведочного анализа (EDA) является графический анализ. С помощью графика мы можем взглянуть на данные целиком. График особенно ценен, когда он позволяет нам увидеть того, чего мы вообще не ожидали увидеть. Гистограмма и боксплот (еще его называют ящик с усами) являются самыми часто используемыми графиками, потому что они просты в построении и интерпретации результата и к тому же очень информативны

Гистограмма:



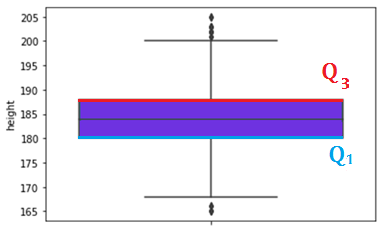
Боксплот



Давайте более подробно остановимся на ящиках с усами. Они бывают как с выбросами, так и без. Края усов в ящике без выбросов просто совпадают с максимальным и минимальным значением. Для ящика с выбросами края усов считаются по следующей формуле:

= Q1 - 1.5\*(Q3-Q1) ;

= Q3+ 1.5\*(Q3-Q1)

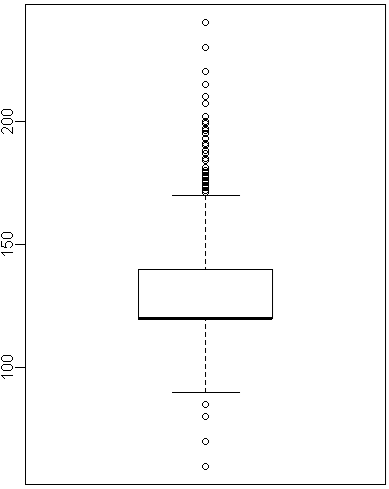


Q1 – это 25-й перцентиль

Q3 – это 75-й перцентиль

Линия посередине соответствует медиане. Т.е. на нашем графике ниже, приблизительно, 182 см лежит 25% значений, а приблизительно ниже 188 лежит 75% значений. Межквартильное расстояние 188-182 =6 см Т.е. 50% значений попадают в диапазон между 182 и 188 см. Линия между Q1 и Q3 – это линия, соответствующая медиане. Так мы можем приблизительно оценить по графику.

Давайте попробуем интерпретировать информацию с этого боксплота.



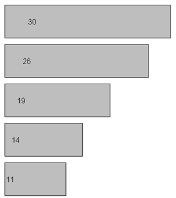
Здесь медиана совпадает с 1-м квартилем, т.е. ниже, приблизительно, 120 – ти лежат 50 % наблюдений и видим, что 1 квартиль соответствует тем же 120, т.е. ниже 120 лежат 25 % наблюдений. Что получается? Что между 120, которое соответствуют 1-му квартилю и 120, которое соответствуют медиане, лежит 25 % и они все будет равны 120. Также видим большое количество выбросов - точки, расположенные за краями усов. Распределение ассиметрично, медиана не проходит посередине ящика.

**Правила визуализации**

При визуализации данных следует соблюдать некоторые правила, которые сделают ваш график «читабильным». Визуализируя данные, помните, что график должен сочетать в себе максимальную простоту и информативность.

Вот правила, которые помогут вам добиться этого:

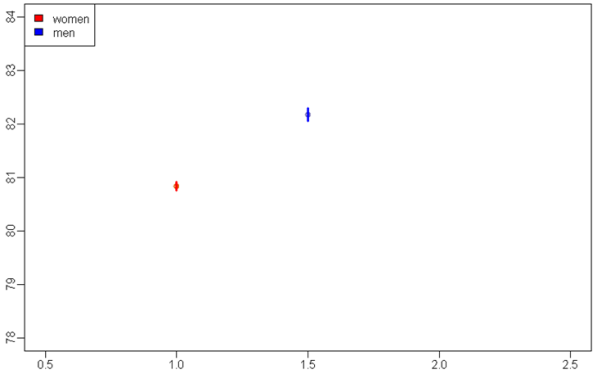
1. Располагать значения в определенном порядке



1. Избегать круговых диаграмм

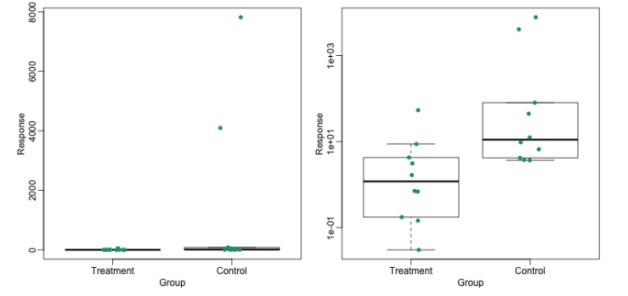
Нам проще сравнивать линейные величины, чем площади или объемы.

1. Не использовать псевдотрехмерную графику (Не надо строить объемные барплоты)
2. Стараться максимально просто изображать данные
3. Использовать одинаковые единицы измерения
4. Не оставлять много знаков после запятой
5. Добавлять легенду на график



1. При необходимости прибегать к масштабированию данных (например, применять log)

Когда полезно масштабирование данных? Взглянем на данные. На левом графике из-за 2 выбросов мы не видим, как изменяются (как рассеянны) остальные данные. Применив логарифмическое масштабирование данных, получаем боксплоты на правом графике, с которыми уже можно работать.



На следующем уроке мы будем знакомиться с непрерывной случайной величиной и с самым главным распределением статистики, с распределением Гаусса.