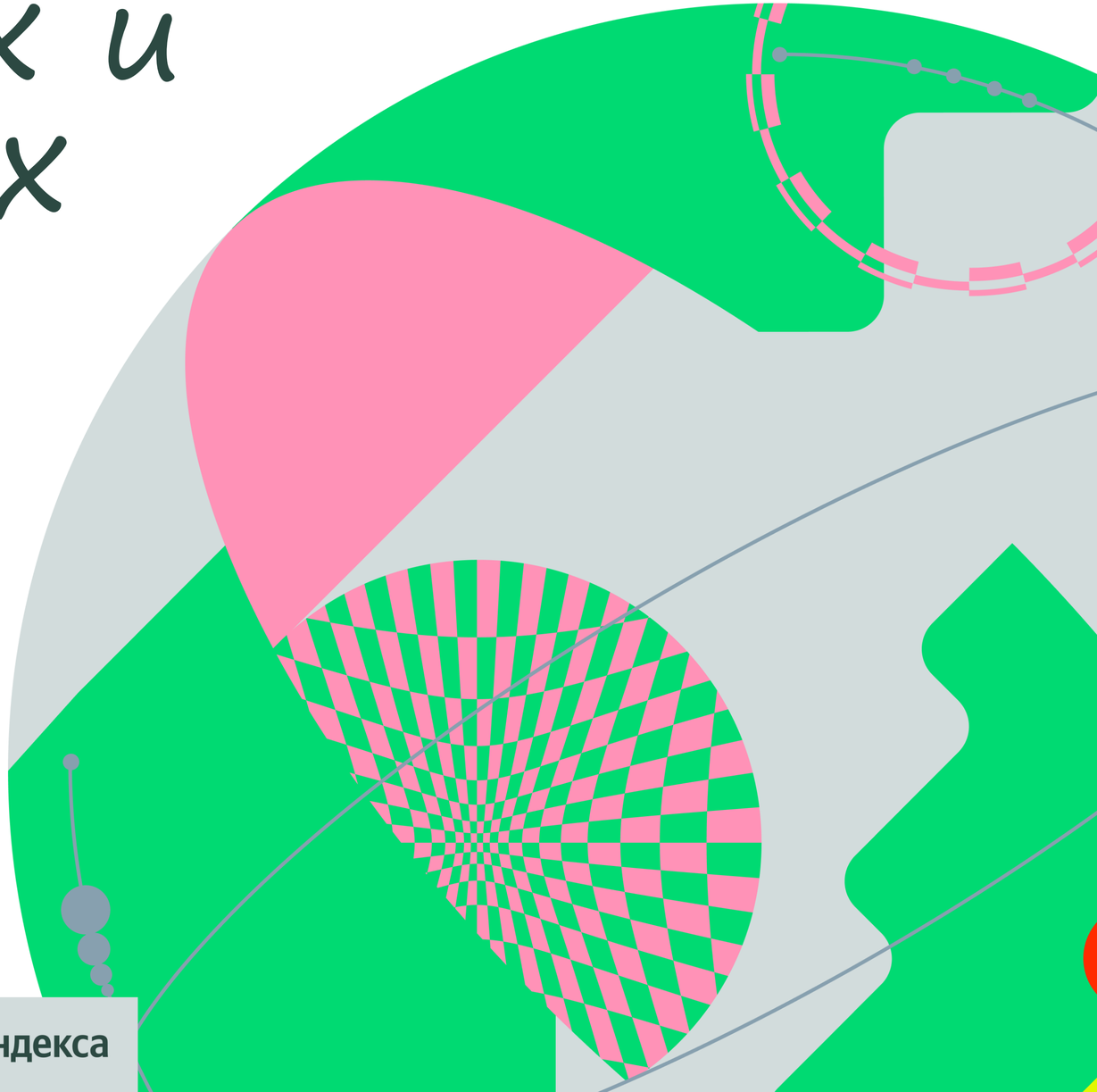
Описательный анализ количественных и категориальных данных



Академия Яндекса

Начало аналитической работы в питоне

**Считайте, что поддается подсчету, измеряйте,
что поддается измерениям,
а не измеряемое делайте измеряемым.**

Галилео Галилей

Обычный план аналитического проекта

1. Проведение обзора данных (EDA)

Первичное исследование данных:
- Импорт необходимых библиотек;
- Чтение файлов и сохранение полученных данных в переменные;
- Получение общей информации о таблицах
- Графическое представление данных из таблиц
- Выводы

Предобработка данных:
.....

Предыстория

Вышло 3 части «Стражи галактики». В каждой – команда героев выполняет разные миссии. Всем они известны. Но мало кто знает, что есть секретная миссия «Стражей галактики». Они и сами про неё не очень любят рассказывать. Но в летописях всё записано.

Полёт был к спутнику планеты Ваканда. (Ваканда. Именно здесь был найден металл вибраниум из которого изготовлены: щит Капитана Америки, когти Черной Пантеры, костюм Железного Человека, рука Баки). А название спутника Сирис.

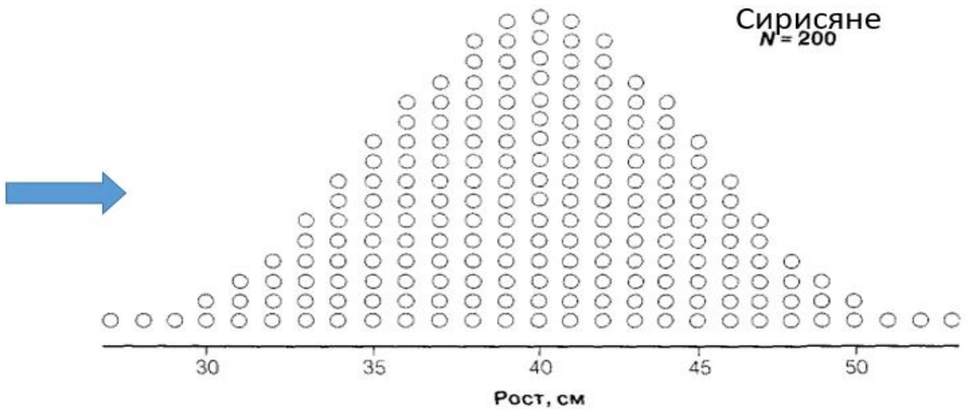
Было известно, что там живут уникальные существа. Похожие на людей. Но невероятно развитые по уровню интеллекта. Задача была как можно больше узнать о них.

Что происходило на планете

И первое, что было сделано – измерен рост существ. Их всего было пару сотен. Результаты показаны ниже. Сначала занесли их в табличный вид. Глядя на столбец данных стало понятно, что многого просто смотря на 200 чисел увидеть не удастся. И тогда изобразили тот же самый столбец, но в виде графика.

Если долго мучить данные
Они признаются (в чём угодно)
……

Рост Сирисянин
47
38
40
36
41
44
35
35
43
41
43
39
34
41
50
31
43
48
34
47
38
39
48
43
45
41
44
36
36
38
41
35
39

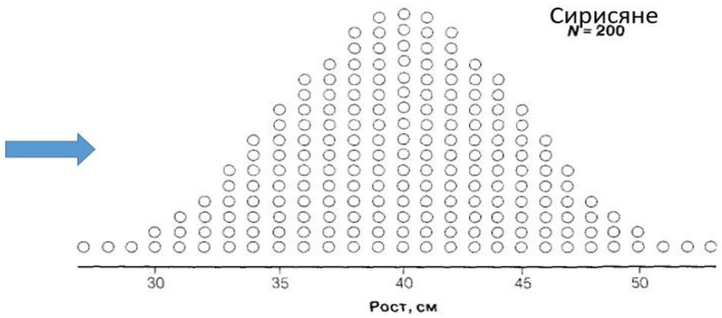


На картинке, каждому сирисянину соответствует кружочек (ячейка в столбце данных). Например, два кружочка над числом 30 означают, что имеется два сирисянина ростом 30 см. А десять кружочков над цифрой 35 означает, что имеется 10 представителей Сириса ростом 35 см.

В таком виде исходных данных (помните про цепочку «данные – информация») можно будет сделать некоторые выводы. Видно, что рост большинства сирисянов - от 35 до 45. Коротышек (ниже 30 см) совсем немного – всего трое. И столько же великанов (выше 50 см)

Что происходило на планете

Рост Сирисянин
47
38
40
36
41
44
35
35
43
41
43
39
34
41
50
31
43
48
34
47
38
39
48
43
45
41
44
36
36
38
41
35
39



Если долго мучить данные
Они признаются (в чём угодно)
……

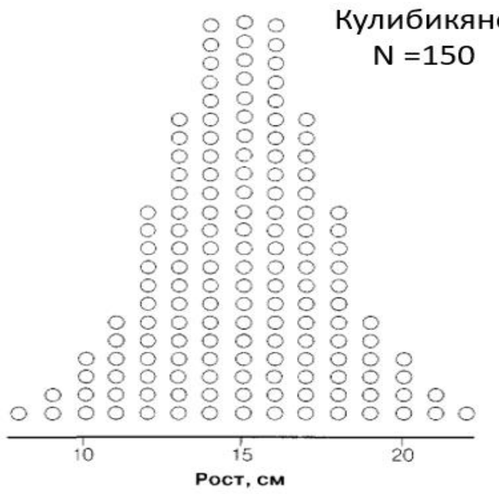
Такой (колокол с кружочками) приято называть – распределением. Пик колокола соответствует наибольшему количеству сирисян с данным ростом.

Такой пик называют – средним значением. И среднее значение роста жителей сириса равняется – 40 см. Также, из графика распределения можно увидеть, что оно симметрично относительно пика (т.е. среднего). Т.е. Левая часть равна правой относительно 40 см.

Что было дальше

Миссия по измерению жителей Сириса прошла успешно. Никто не пострадал. И было решено слетать на второй спутник планеты Ваканда – Кулибика. И измерить рост жителей этой планеты, взяв всё необходимое – сделали это. Результат оформили аналогично:

Рост Кулибни
14
18
21
12
17
19
14
13
18
14
10
15
13
14
13
14
17
17
13
17
17
18
12
14
16
18
17
15
16
14
18
15
17

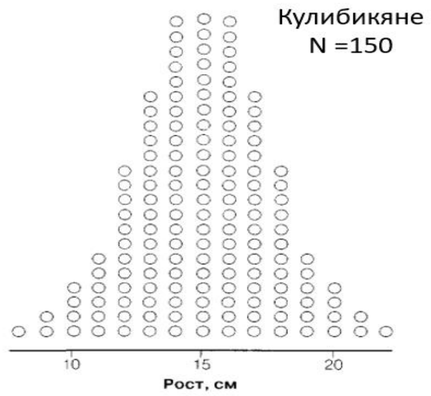


Что было дальше

А тут видно, что рост большинства кулибикян, примерно, - от 13 до 17. Коротышек (ниже 10 см) трое. Ну и местных великанов тоже трое (выше 20 см). В целом, распределение кулибикян схоже по форме с распределением сирисян – колокол и там и там.

Про вид распределения роста – увидели сразу, прямо на корабле. А когда прилетели на базу, то приступили к более детальному изучению данных. Из графика видно, что в среднем сирисяне выше кулибикян. И что интервал, в который умещается рост всех сирисян, шире, чем соответствующий интервал кулибикян. Ширина интервала, в который попадают почти все жители сириса (194 из 200 – это можно посчитать по кружочкам) – 20 см. Рост большинства кулибикян (144 из 150 – тоже можно посчитать по кружочкам) умещается в интервал от 10 до 20 см. То есть имеет ширину всего 10 см.

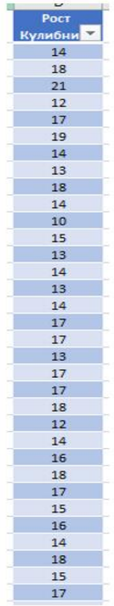
Рост Кулибяни
14
18
21
12
17
19
14
13
18
14
10
15
13
14
13
14
17
17
13
17
17
18
12
14
16
18
17
15
16
14
18
15
17



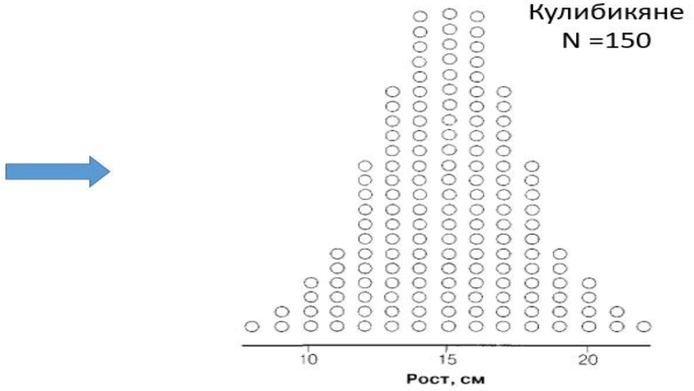
Академия Яндекса

Что было дальше

А тут видно, что рост большинства кулибикян, примерно, - от 13 до 17. Коротышек (ниже 10 см) трое. Ну и местных великанов тоже трое (выше 20 см). В целом, распределение кулибикян схоже по форме с распределением сирисян – колокол и там и там.



Рост Кулибики
14
18
21
12
17
19
14
13
18
14
10
15
13
14
13
14
17
17
13
17
17
18
12
14
16
18
17
15
16
14
18
15
17



Про вид распределения роста – увидели сразу, прямо на корабле. А когда прилетели на базу, то приступили к более детальному изучению данных. Из графика видно, что в среднем сирисяне выше кулибикян. И что интервал, в который умещается рост всех сирисян, шире, чем соответствующий интервал кулибикян. Ширина интервала, в который попадают почти все жители сириса (194 из 200 – это можно посчитать по кружочкам) – 20 см. Рост большинства кулибикян (144 из 150 – тоже можно посчитать по кружочкам) умещается в интервал от 10 до 20 см. То есть имеет ширину всего 10 см.

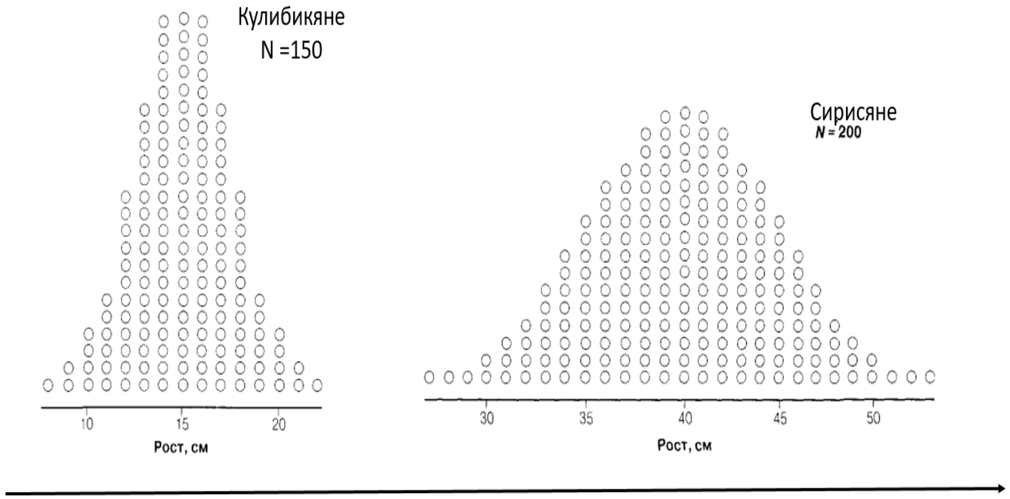
Что там с жителями планет

Несмотря на эти различия, между жителями 2-х планет есть и существенное сходство. В обоих рост любого жителя скорее близок к середине распределения, нежели удалён от неё. При этом форма распределения примерно одинаковая по форме. Т.е. можно предположить, что они (формы распределения) определяются одной и той же формулой).

И это интересное наблюдение и предположение. Получается, что жителей двух планет мы сравниваем друг с другом, опираясь не на все данные из таблицы. А на всего несколько значений (параметров) из распределения. Это среднее и ширина интервала в котором находятся большинство значений.(по другому ещё называют ширину интервала – разброс данных).

Посмотрим детальнее на эти понятия.

Информация



Видим, что центр распределения (пик графика) кулибякян расположено в начале числовой оси, а распределение сирисян – дальше по оси. Центр распределения принято называть – среднее.

Формулу среднего запишем так:

$$\text{Среднее по всем данным} = \frac{\text{Сумма всех значений фактора}}{\text{Число значений фактора}}$$

А в математике принято писать так:

$$\mu = \frac{\sum X}{N}$$

, где X – значение фактора (в нашем случае это рост каждого жителя планеты),
N – число значений фактора (в нашем случае - количество жителей на планете),
мю – обозначение среднего (греческая буква мю)

Теперь можем численно посчитать средние рост жителя каждой планеты и их разницу. Сирисяне в среднем 40 см. Кулибикян – 15 см. И сирисяне выше кулибикян в среднем на 25 см. (40 минус 15)

Академия Яндекса

Всегда ли справедливо среднее

Команда «Стражей галактики» состоит из шести (назовём человек:) Пришло время обедать. А Квиллу дедушка прислал с Земли 12 пирожков (да 12, а что?)).

И вот такую ситуацию рассмотрим, что Дракс взял всё и съел. С кем не бывает, если пирожки вкусные, а огромное тело просит вкусняшек)

Если задаться вопрос: «сколько в среднем съел пирожков каждый?».

Имя	Съел пирожков	
Квилл	0	1
Гамора	0	2
Ракета	0	3
Небула	0	4
"я есть"Брут"	0	5
Дракс	12	6
	0+0+0+0+0+12/ 6'	
Среднее =	2	

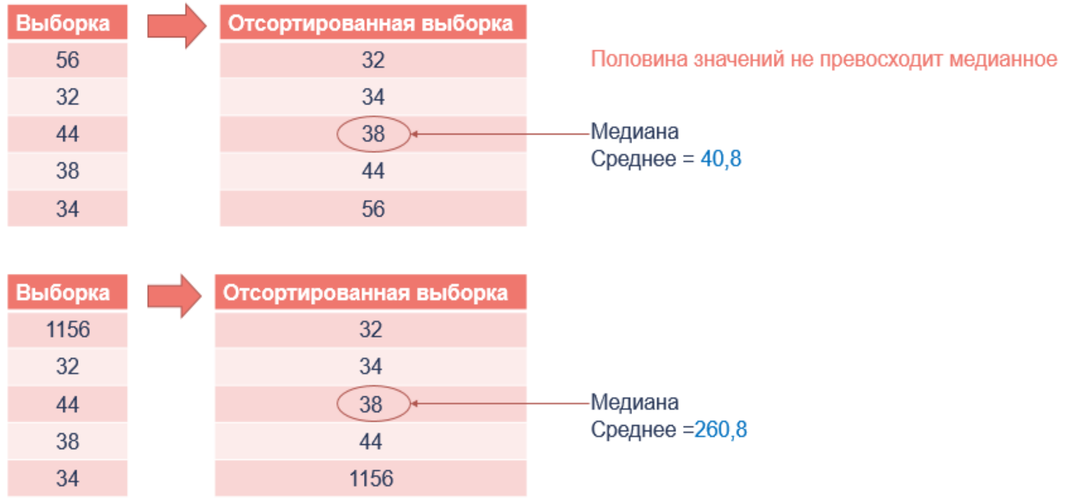
Применим формулу среднего, и получим, что в среднем каждый съел по два пирожка.

Драксу хорошо. А остальные не понимают, почему им так голодно, ведь все съели по 2 пирожка? :))

Всегда ли справедливо среднее

Чтобы избегать таких, прямо скажем, обманов, разработали аналог среднего, но более честный - медиана.

Давайте вместе посмотрим, что такое. По определению, медиана – это число, которое делит отсортированную выборку на две равные части.

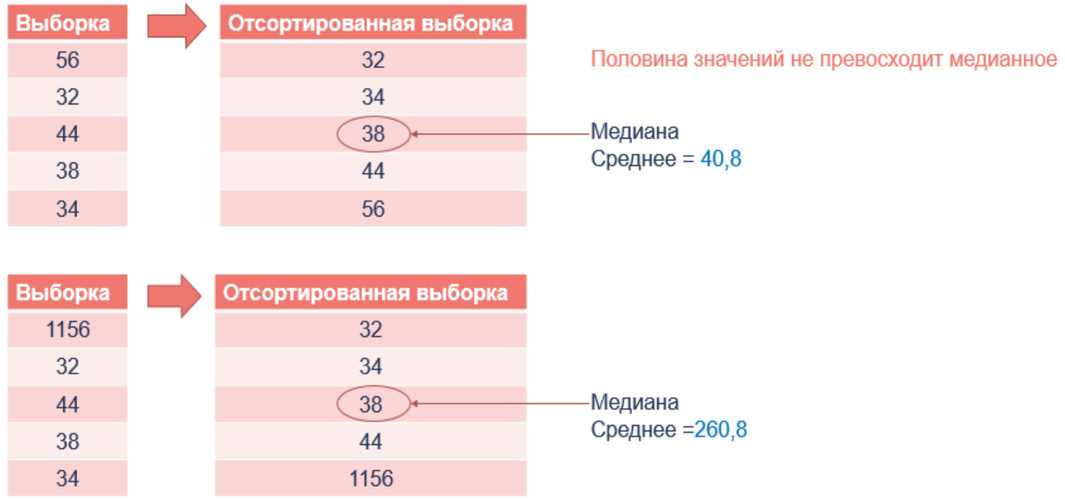


КАКОВА МЕДИАНА В СЛУЧАЕ С ПИРОЖКАМИ?

Всегда ли справедливо среднее

Чтобы избегать таких, прямо скажем, обманов, разработали аналог среднего, но более честный - медиана.

Давайте вместе посмотрим, что такое. По определению, медиана – это число, которое делит отсортированную выборку на две равные части.



Имя	Съел пирожков
Квилл	0
Гамора	0
Ракета	0
Небула	0
"я есть"Брут"	0
Дракс	12

середина

Медиана равна = 0.
И это более честный ответ ведь.
Потому что пирожками подкрепился только Дракс.

Всегда ли справедливо среднее

Заодно, рассмотрим такой вот график, который применяется тогда, когда речь заходит о медиане. Его называют «Ящик с усами» (боксплот). Позже посмотрим, как его построить. А сейчас просто рассмотрим его суть.

Разброс данных

Вернёмся к жителям планет Сирис и Кулибик. Было видно, что графики распределения отличаются - один уже другого. Мы даже это обозначали. Как этот факт принято обозначать в статистике (в описательном анализе).

Во-первых

можно посмотреть разницу мин/макс по каждому распределению.

1) Сирис: 53 – 27 = 25
2) Кулибик: 22 – 8 = 14

Во-вторых

математики придумали более формулу.

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}.$$

Обозначается «сигма в квадрате».
И называется – дисперсия. Это тоже мера разброс данных.

Но как видим, не просто мин/макс, а то, как высоко/низко каждое измерение удалено от среднего.
Т.е. на сколько суммарно рост всех жителей планеты отклоняется от среднего роста.

- Сирис: 25
- Кулибик: 6,25

В-третьих

Квадрат тяжело понять и анализировать.
Поэтому решили взять корень квадратный и получили понятие «стандартное отклонение»

$$\text{Стандартное отклонение} = \sqrt{\frac{(\text{i-е значение} - \text{среднее значение})^2}{\text{кол-во значений}}}$$

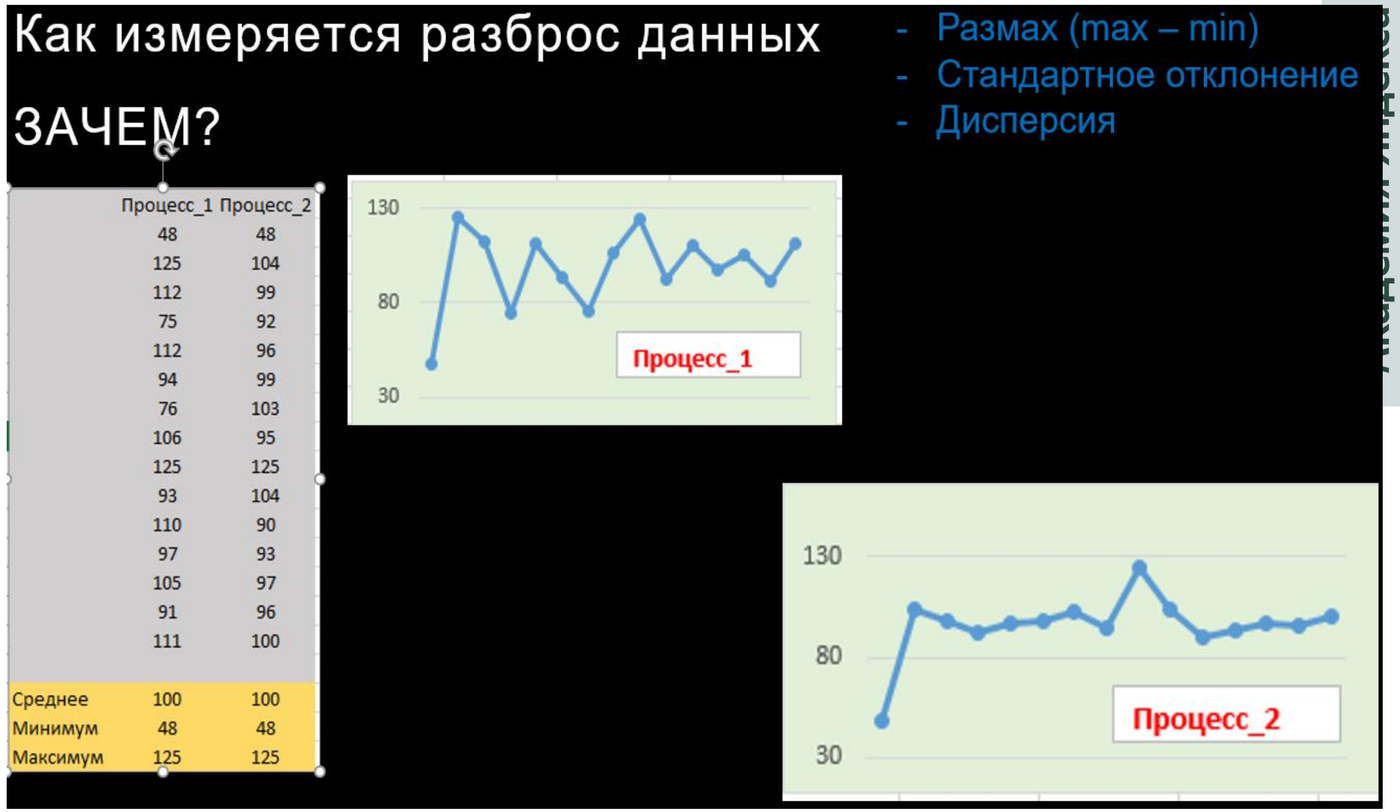
- Сирис: 5
- Кулибик: 2,5

Разброс данных

Вот мы и получили три меры, которые помогут нам понять как сильно изменяются данные. (помните, зачем нам всё это? Изначально ведь мы измерили всех жителей одной планеты и другой планеты. Получили длинный столбец с данными один и другой. Как из данных легко и не принуждённо «выудить информацию? Нужно научиться описывать их несколькими параметрами. Которые можно сравнивать друг с другом и тем самым получать знания о процессе.)

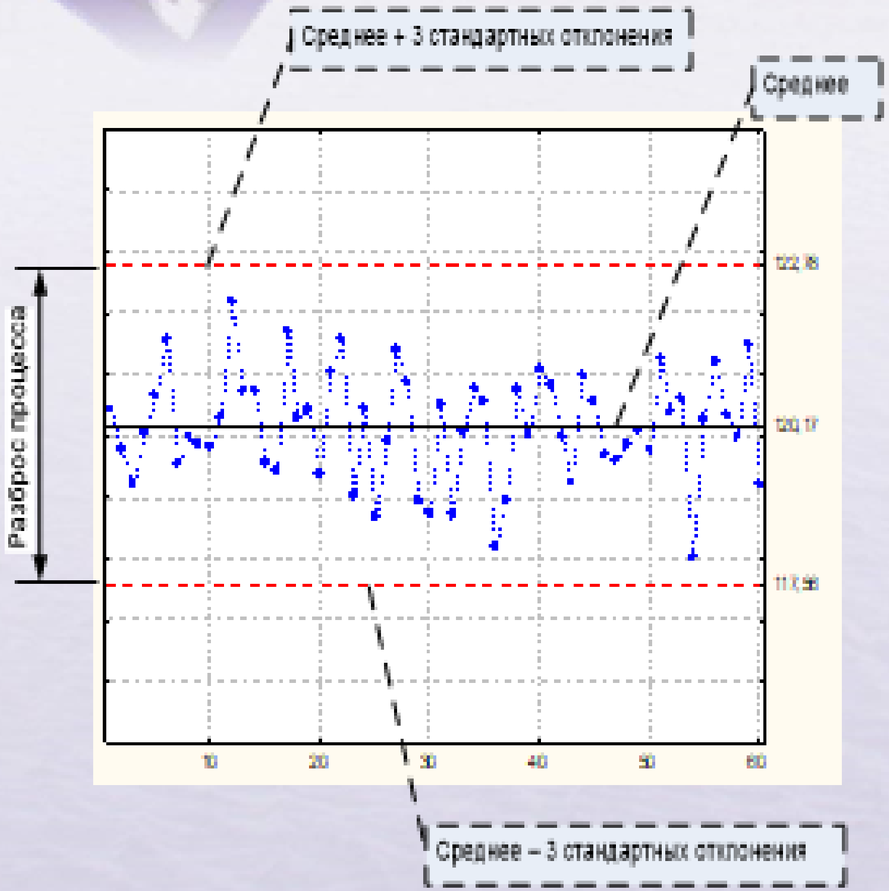
И так, имеем меры разброса данных:
- размах
- дисперсия
- стандартное отклонение

На практике для одних задач используется одно, для других другое.
Давайте посмотрим в чём практическая разница этих трёх мер разброса:
(далее рассуждаем о разнице)



Разброс данных

1.Золотое правило 3-х сигм



$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

Правило «трех сигм»:

Среднее ± σ - 68,3%
Среднее ± 2σ - 95,4%
Среднее ± 3σ - 99,7%

С помощью стандартного отклонения можно измерить реальный разброс процесса, а не только фактический разброс выборки

Академия Яндекса

Разброс данных

1. Складывание дисперсий(стандартные отклонения не складываем). Пример из M&M’s, Orbit, Драже

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}$$

Дисперсии можно складывать и вычитать

Дисперсия – это стандартное отклонение в квадрате.

Чем больше дисперсия, тем больше разброс данных вокруг среднего.

Основная формула для статистического изучения как существующих процессов, так и проектирование новых:

$$\sigma^2_{процесса} = \sigma^2_{оборуд.} + \sigma^2_{люди} + \sigma^2_{изм.сист.} + \sigma^2_{материал} + \sigma^2_{прочее}$$

Разброс данных (задачки)

Задача 1. Оценить реальный разброс процесса

Средний вес одной шоколадки = 120 г
σ = 2,0 г
Среднее ± 3σ = 120 ± 6 = [114…126] – реальный разброс процесса

Задача 2. Сравнить процесс до и после замены механизма

Средний вес одной шоколадки = 120 г
$\sigma_{до}$ = 2,0 г
$\sigma_{после}$ = 2,4 г
После замены механизма процесс стал хуже, так как реальный разброс процесса увеличился с **[114…126]** до **[112,8…127,2]**

Разброс данных (задачки)

Средний вес одной шоколадки = 120 г

$\sigma_1 = 2$ г

Вопрос. *Сколько будут весить две шоколадки и какие спецификации по весу двух шоколадок необходимо установить?*

Средний вес двух шоколадок = 240 г

$$\sigma_2 \neq \sigma_1 + \sigma_1$$

$$\sigma_2^2 = \sigma_1^2 + \sigma_1^2 = 4 + 4 = 8$$

$$\sigma_2 = \sqrt{8} = 2,83$$

Среднее ± 3σ = 240 ± 8,49 – реальный разброс процесса

Ответ. *Рекомендуемые спецификации не должны быть меньше, чем полученный реальный разброс процесса.*

Распределение данных

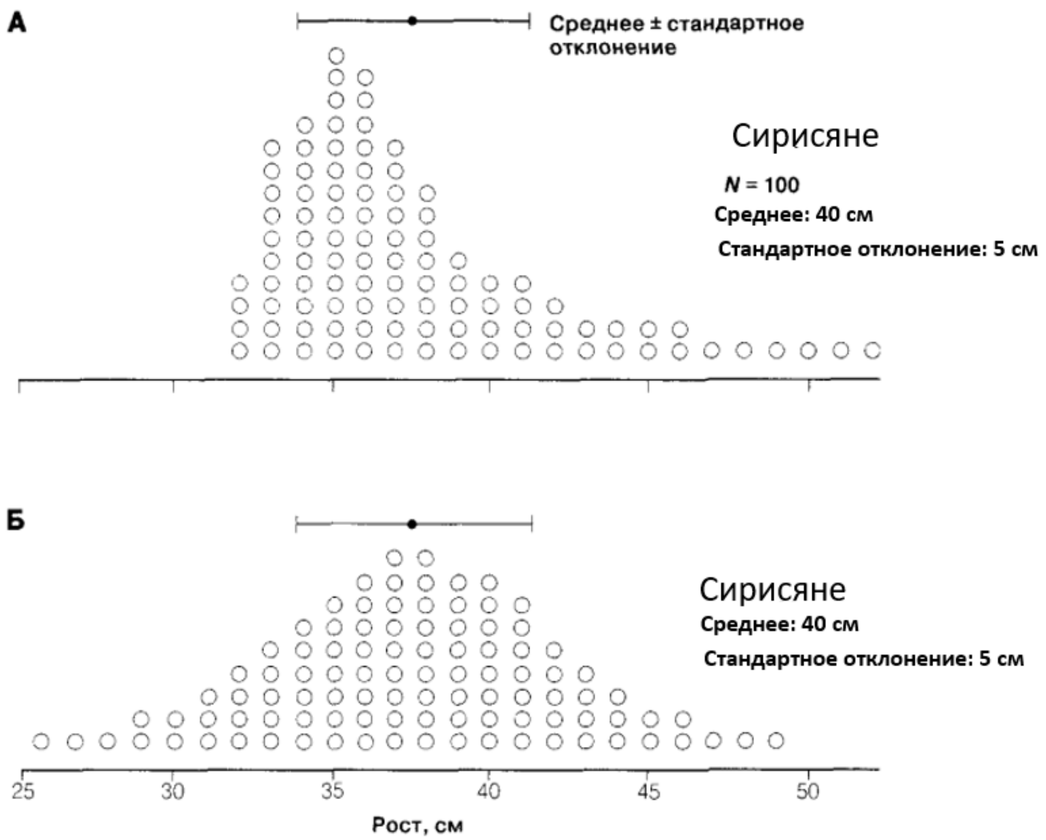
Всё это работает в промышленной аналитике на УРА. Но в том случае, если данные имеют так называемое нормальное распределение данных. Вы его уже видели – это колокол Сириса и Кулибика выше. Такое распределение носит название – нормальное или «распределение Гаусса».
Есть даже математическая формула. Посмотрим на неё (но практиковать конечно же не будем)

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

Смотрите, там участвуют параметр среднего и стандартного отклонения.

Как не покажется странным, распределения данных играют в нашей практической жизни бОльшую роль, чем нам кажется. Давайте ещё посмотрим, что это такое:

Распределение данных



Если распределение асимметрично (т.е. не нормальное, не Гауссово), то на среднее и стандартное отклонение полагаться нельзя.

Вот ниже рисунки А и Б:
Рисунок Б: Реальное распределение роста сирисян, которое мы уже смотрели. У него среднее и стандартное отклонение посчитано.

Рисунок А: мы создали новое распределение с теми же параметрами среднего и стандартного отклонения. Но оно ни чего не имеет общего с реальным распределением сирисян.

Как с пирожками и видели, в случае не симметричного распределения для аналитических выводов правильнее пользоваться медианой и перцентилями.

Распределение данных вокруг нас

Сотрудник	Выполнение плана за квартал
1	91%
2	114%
3	88%
4	104%
5	96%
6	106%
7	100%
8	116%
9	96%
10	110%
11	91%
12	103%
13	92%
14	88%
15	106%
16	80%
17	98%
18	103%
19	114%
20	103%
...	...

Номер интервала	Нижняя граница	Верхняя граница	Частота
1	80%	85%	3
2	85%	91%	7
3	91%	96%	9
4	96%	101%	13
5	101%	106%	13
6	106%	112%	10
7	112%	117%	5

$n = 60$ чел.

Число интервалов определяется автоматически
(есть много формул)

Гистограмму НЕ СТРОЯТ, если

- данных «мало» (менее 30 значений)
- **данные имеют тренд и / или сезонность**

Академия Яндекса

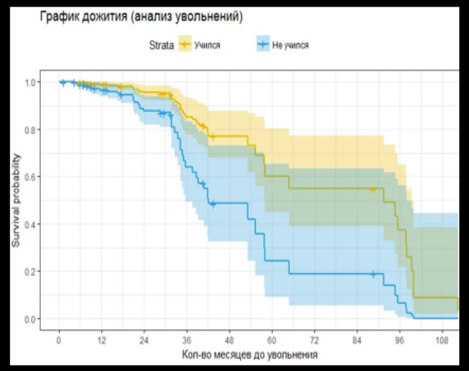
Распределение данных вокруг нас

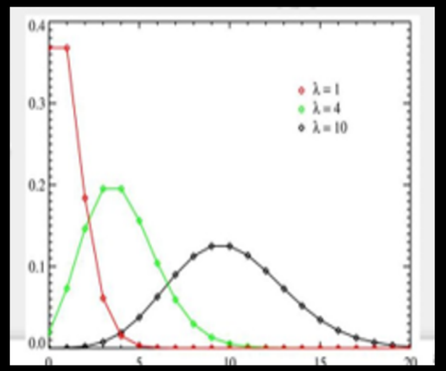
Этому вопросу посвящали много времени разные учёные мужи. Сначала наблюдали жизнь, а потом открывали закон распределения в математической формуле. (Пуассон, Вейбул, Бернули….)

Что такое распределение

Любое явление нашей жизни подчиняется тому или иному закону распределения.

- Приход на работу в офис (Гаусса, нормальное)
- Уровень регистрируемого брака на заводе
- Время обслуживания клиента
- Количество человек в очереди в единицу времени (Пуассона, теория массового обслуживания
- Время дожития (в медицине)
- Надёжность оборудования (Вейбула, производство изделий)





Академия Яндекса

Описание данных в качественной (категориальной) шкале.

Предположим космопорт Асгарда (окунёмся опять в мир комиксов Марвела) ведёт учёт прибывших к ним из всех возможных 9 миров. И табличка имеет вид:

но	Откуда	Кто
1	Асгард	…
2	Мидгард	..
3	Ванахейм	..
4	Йотунхейм	
5	Асгард	
6	Асгард	
7	Ванахейм	
8	Мидгард	
….	…	..
100	Альфахейм	…

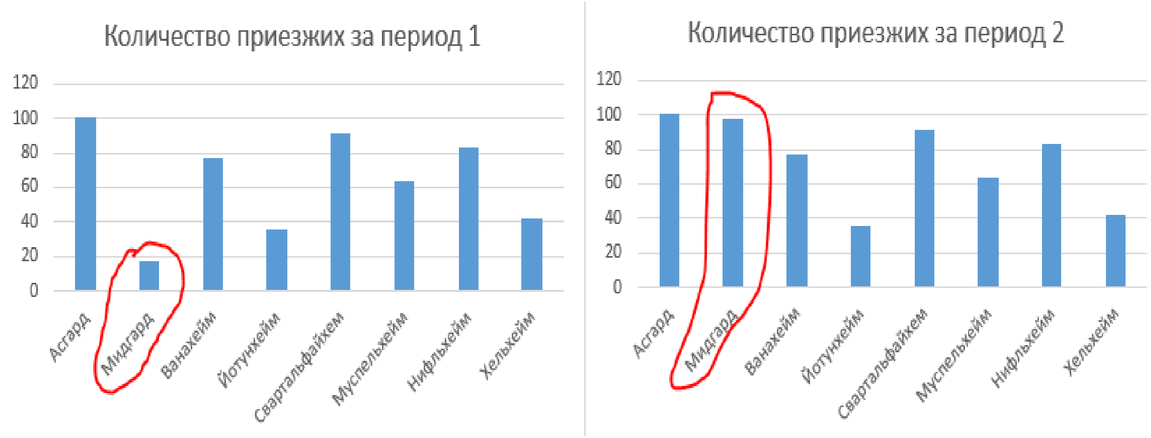
Строк много, глазами не объять всё сразу. Что можно с ними сделать, как свернуть, чтобы быстро зрительно оценить?

Тут только один метод – подсчёт. И оценить его легко графически через столбчатые диаграммы и круговые диаграммы. В нашем случае построим столбчатые.

Академия Яндекса

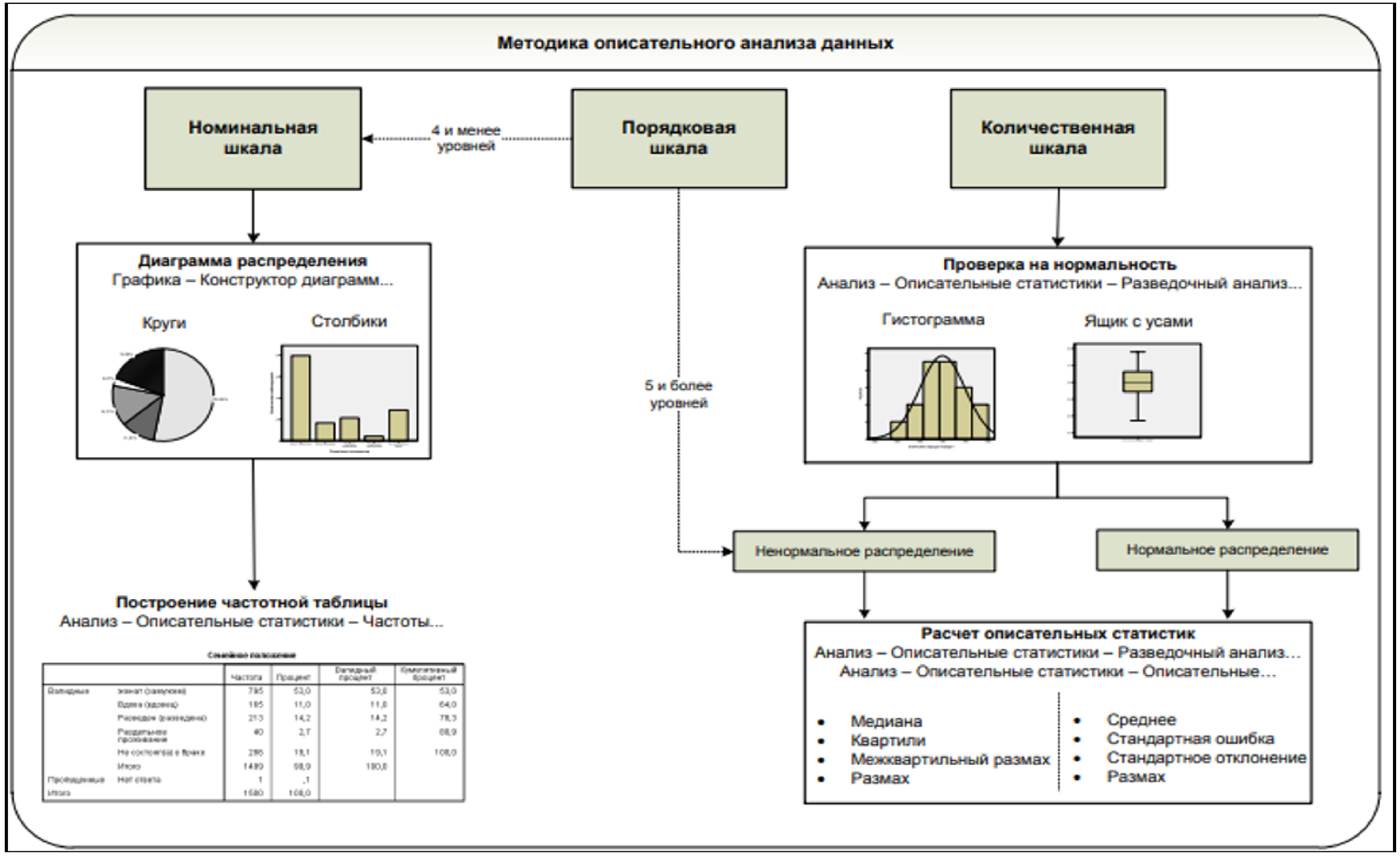
Описание данных в качественной (категориальной) шкале.

Можно выводить в процентах, в штуках. Но идея одна – подсчёт. И сравнение подсчёта. Например, можно сопоставить два периода и посмотреть, есть ли прирост приезжающих?



Видно, что с Мидгарда прибыло необычно много! Ситуационный отдел получил команду разобраться почему/отчего/зачем?

Итоги по описательному анализу



Считайте, что поддается подсчету, измеряйте, что поддается измерениям, а не измеряемое делайте измеряемым.

Галилео Галилей

Академия Яндекса

Начало аналитической работы в питоне

**Считайте, что поддается подсчету, измеряйте,
что поддается измерениям,
а не измеряемое делайте измеряемым.**

Галилео Галилей

Обычный план аналитического проекта

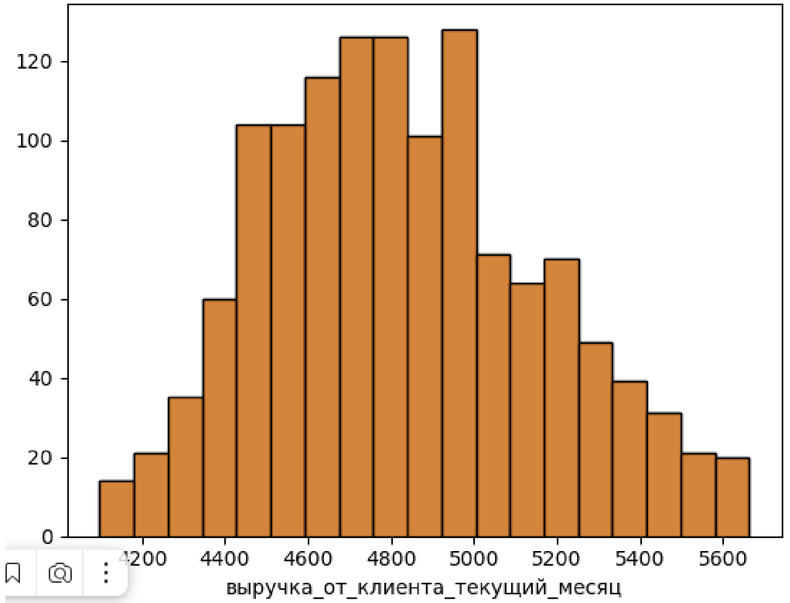
1. Проведение обзора данных (EDA)

Первичное исследование данных:
- Импорт необходимых библиотек;
- Чтение файлов и сохранение полученных данных в переменные;
- Получение общей информации о таблицах (head, info, **describe**);
- Графическое представление данных из таблиц
- Выводы

Предобработка данных:
.....

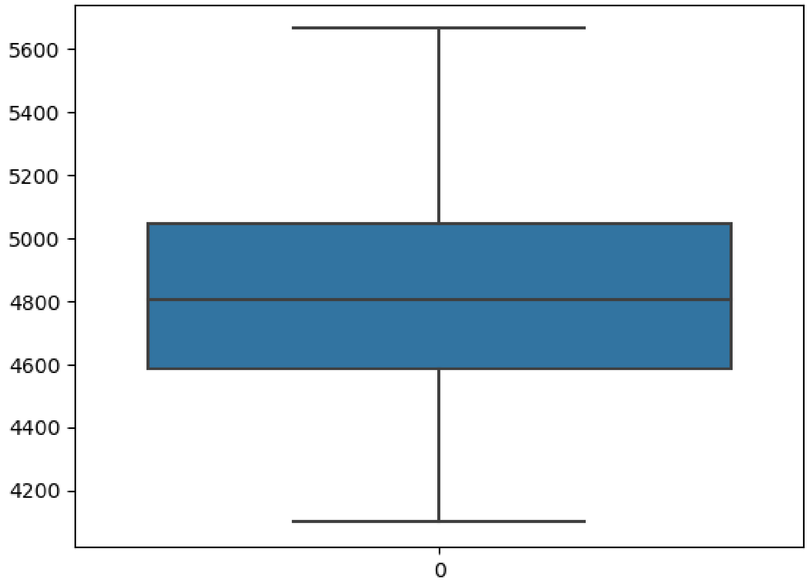
Что понадобится для выполнения задач

```
sns.histplot(df['выручка_от_клиента_текущий_месяц'])
plt.show()
```

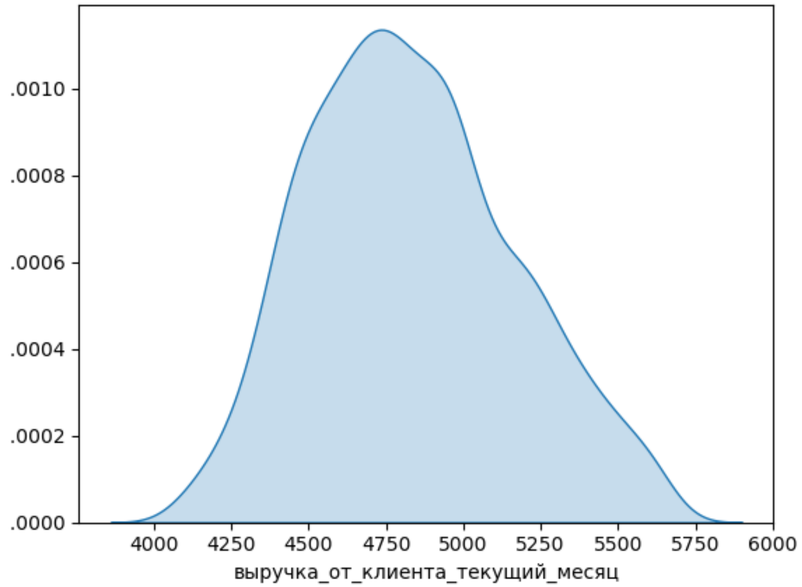


seaborn

```
1 sns.boxplot(df['выручка_от_клиента_текущий_месяц'])
2 plt.show()
```



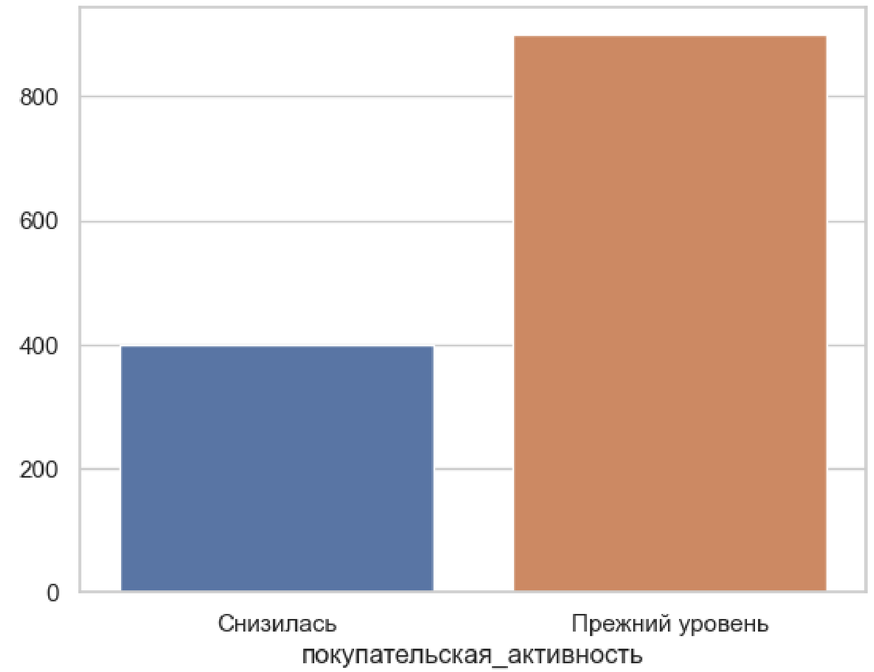

```
sns.kdeplot(df, x="выручка_от_клиента_текущий_месяц", fill = True)
plt.ylabel('')
plt.show()
```



Что понадобится для выполнения задач

```
y = df['тип_сервиса'].value_counts()
plt.pie(y)
plt.show()
```

```
sns.countplot(data=df, x="покупательская_активность")
plt.show()
```



Интересные статьи. Важно при ответе на вопросы

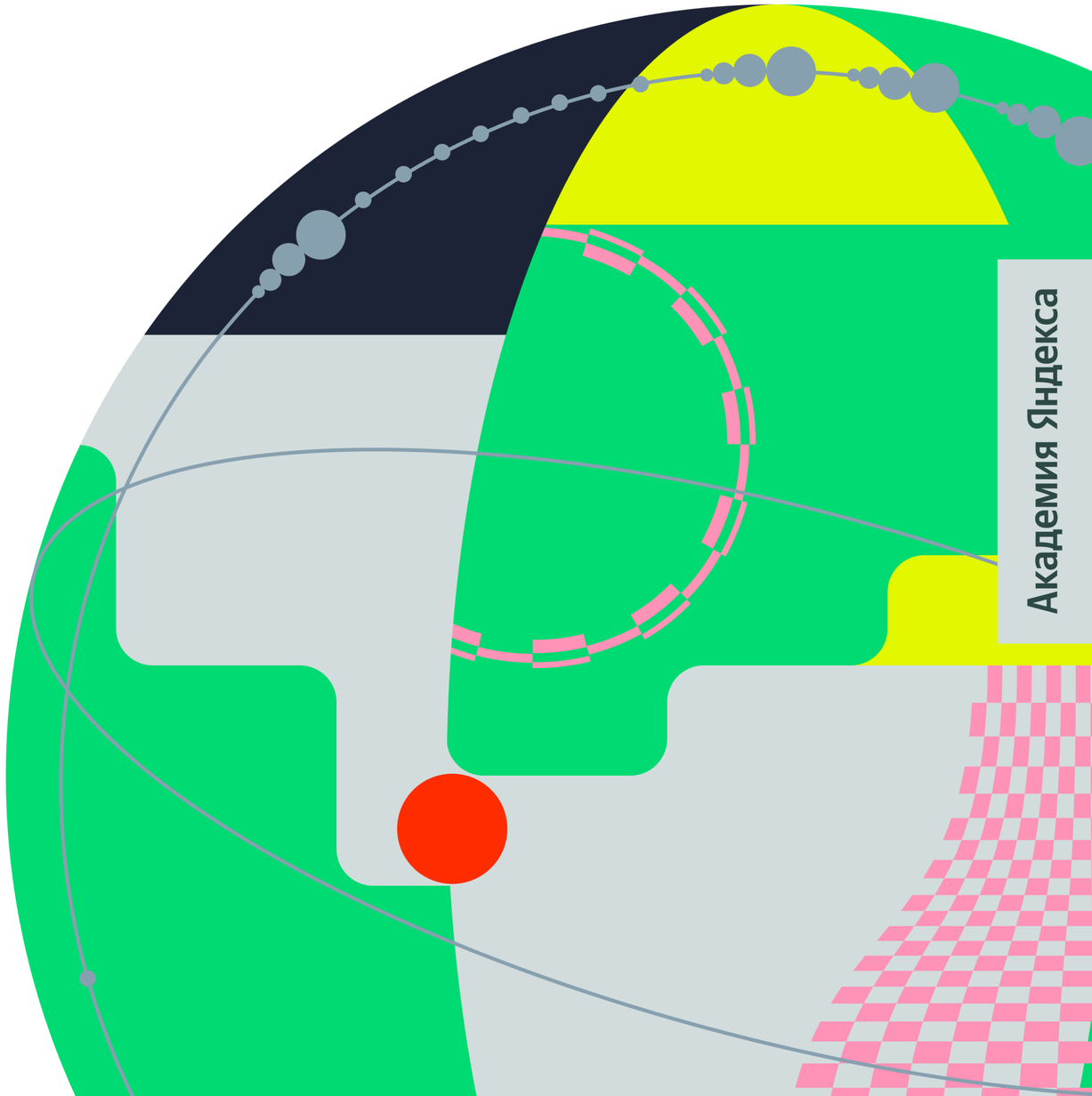
О распределениях: https://hr-portal.ru/statistica/gl3/gl3.php

Просто для эрудиции

Сборник готовых задач на различные виды распределений дискретной случайной величины

https://mathprofi.net/files/zadachi_dsv.pdf

Спасибо за внимание



Академия Яндекса позволяет школьникам и студентам освоить востребованные IT-профессии по программам, разработанным экспертами компании