

---

# Занятие № 5

## Проблема качества данных



---

# Содержание

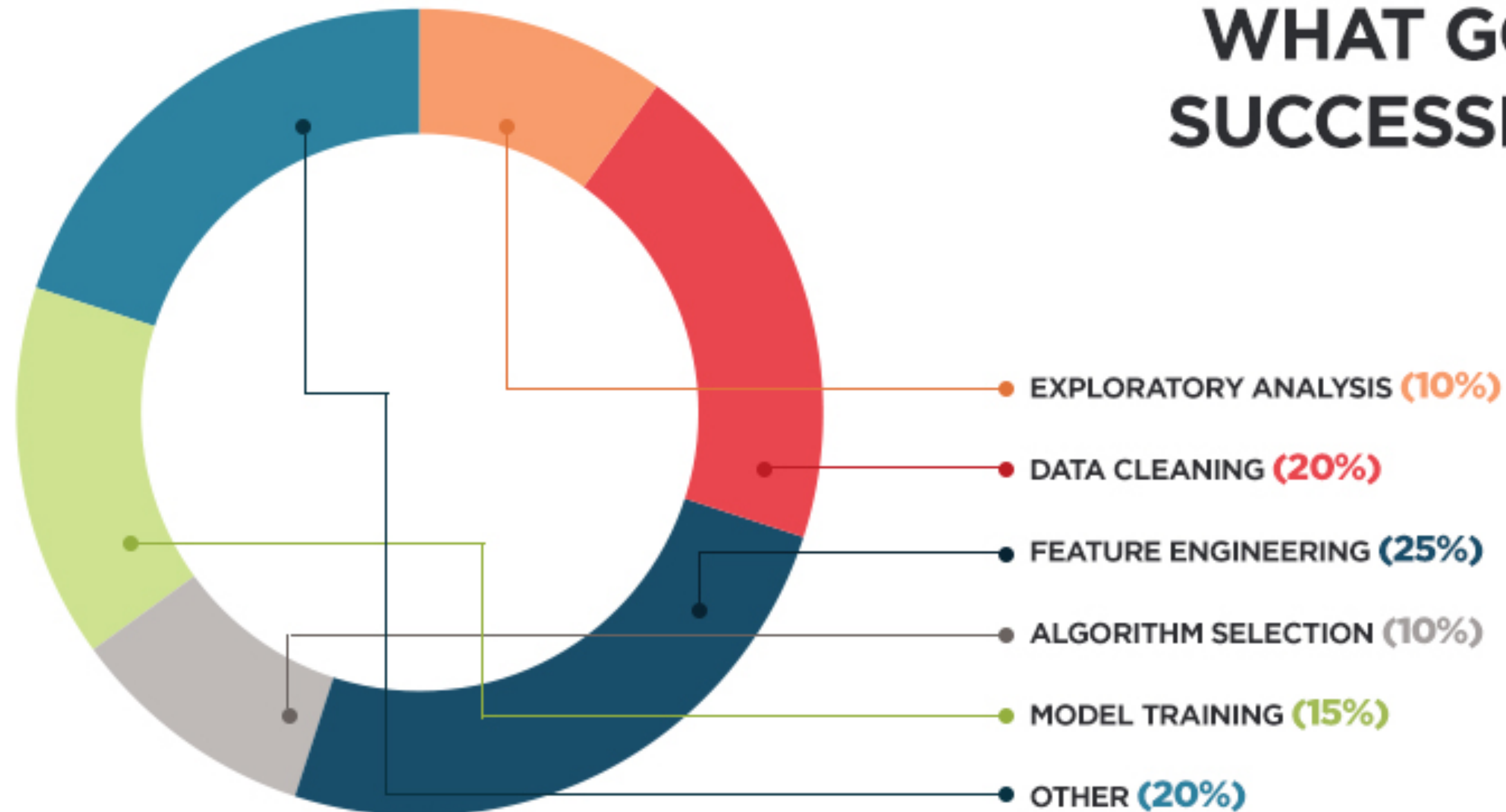
---

- 1 Описание основных проблем с данными
- 2 Определение валидности и правильности данных
- 3 Работа с пропущенными значениями
- 4 Обработка категориальных переменных
- 5 Практика.



# Введение

## WHAT GOES INTO A SUCCESSFUL MODEL

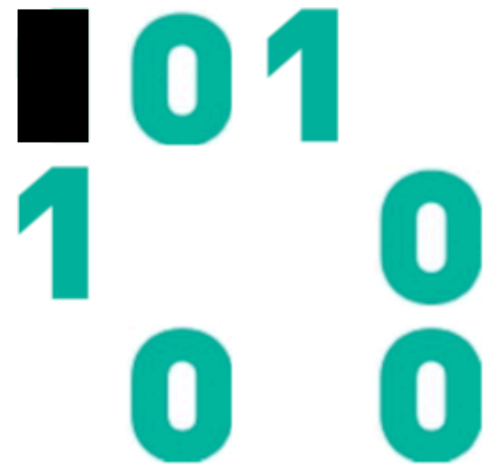


# Создание новых параметров

1. Метод мозгового штурма или проверка признаков;
2. Решение, какие признаки создавать;
3. Создание признаков;
4. Проверка, какие признаки работают с вашей моделью;
5. Улучшение признаков, если требуется;
6. Возврат к методу мозгового штурма/создание других признаков, пока работа не будет завершена.



# Данные и сопутствующие проблемы



A 3x3 matrix of data points. The top-left cell is a solid black square. The other cells contain binary digits: top row is 0, 1, 0; middle row is 1, 0, 0; bottom row is 0, 0, 0.

Недостаточное  
количество  
данных



A 3x4 matrix of data points. All cells contain the binary sequence 1010, arranged in three rows.

Нерепрезентативные  
данные



A 3x4 matrix of data points. All cells contain the binary sequence 1010, arranged in three rows.

Данные  
плохого качества  
аномалии,  
выбросы,  
нулевые значения



A 14x14 matrix of data points. All cells contain the binary sequence 10101010101010, arranged in 14 rows.

Большая  
размерность  
пространства  
данных



# Обработка нулевых значений

**Не стоит делать в большинстве случаев:**

- Удалять столбец содержащий нулевое значение (потеря информации)
- Удалять строки, в которых атрибут равен нулевому значению (потеря информации)



# Обработка нулевых значений



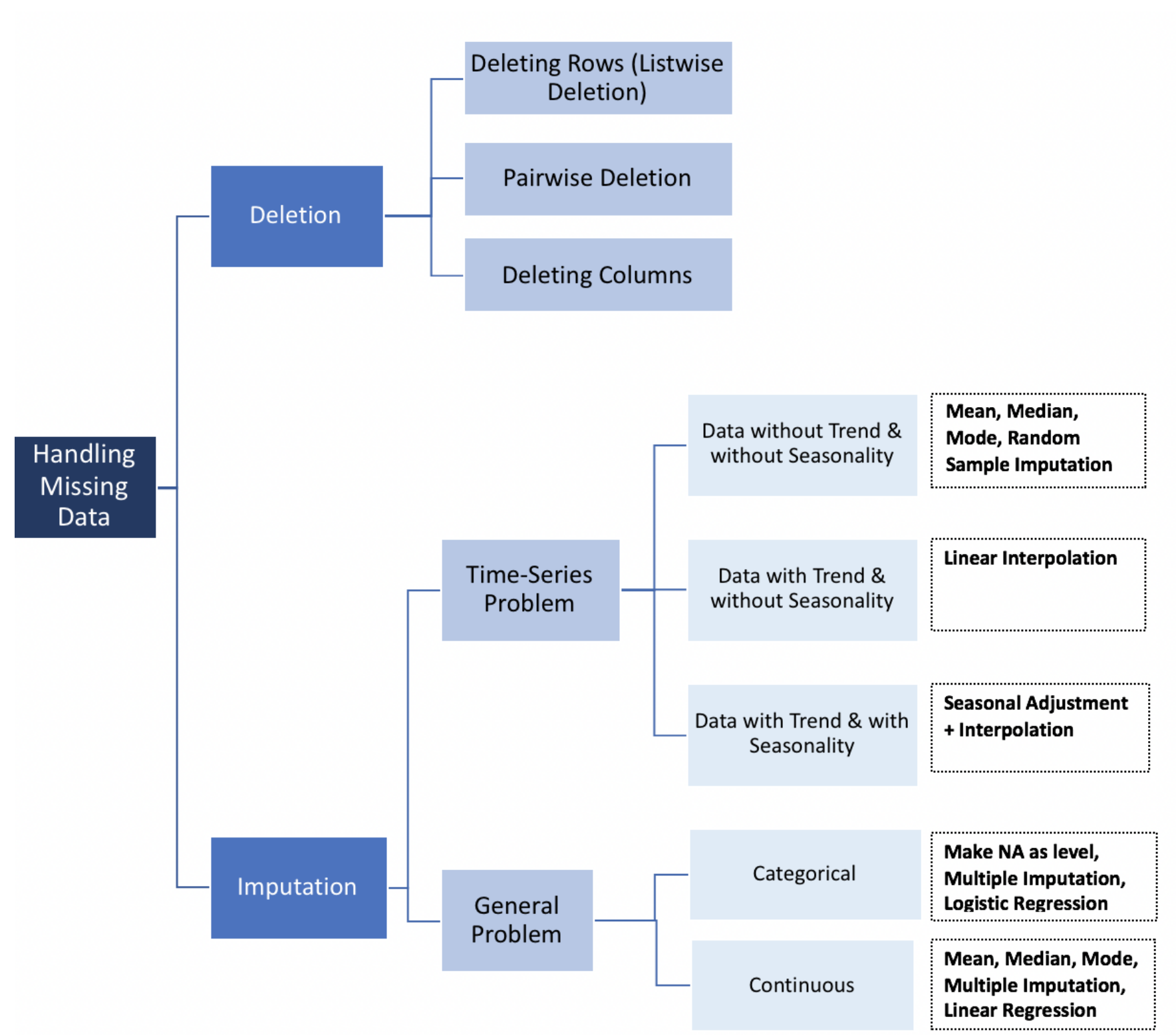
## Что же делать?

- Заменять на среднее значение, медиану, моду
- Indicator Method - замена пропущенных значений нулями и создание новой переменной индикатора (где она принимает значение 1 при наличие пропуска и 0 в остальных случаях)
- Повторить результат последнего наблюдения (среднее между
- Восстановление пропусков на основе моделей





# Обработка нулевых значений





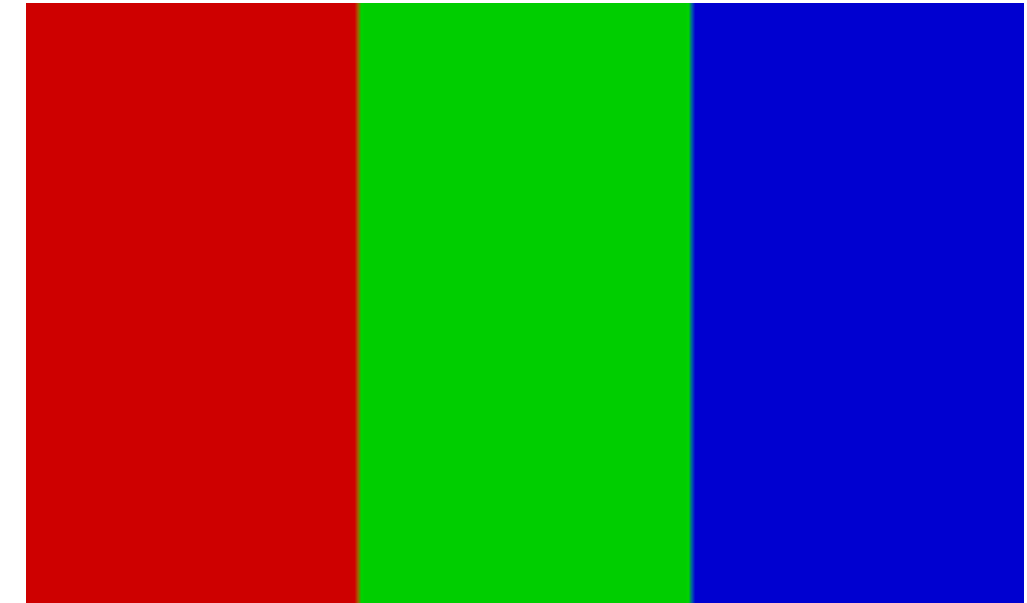
# Первичный анализ данных. Типы признаков

0 1 2 3 4  
5 6 7 8 9

**Количественные**



**Бинарные**



**Категориальные**

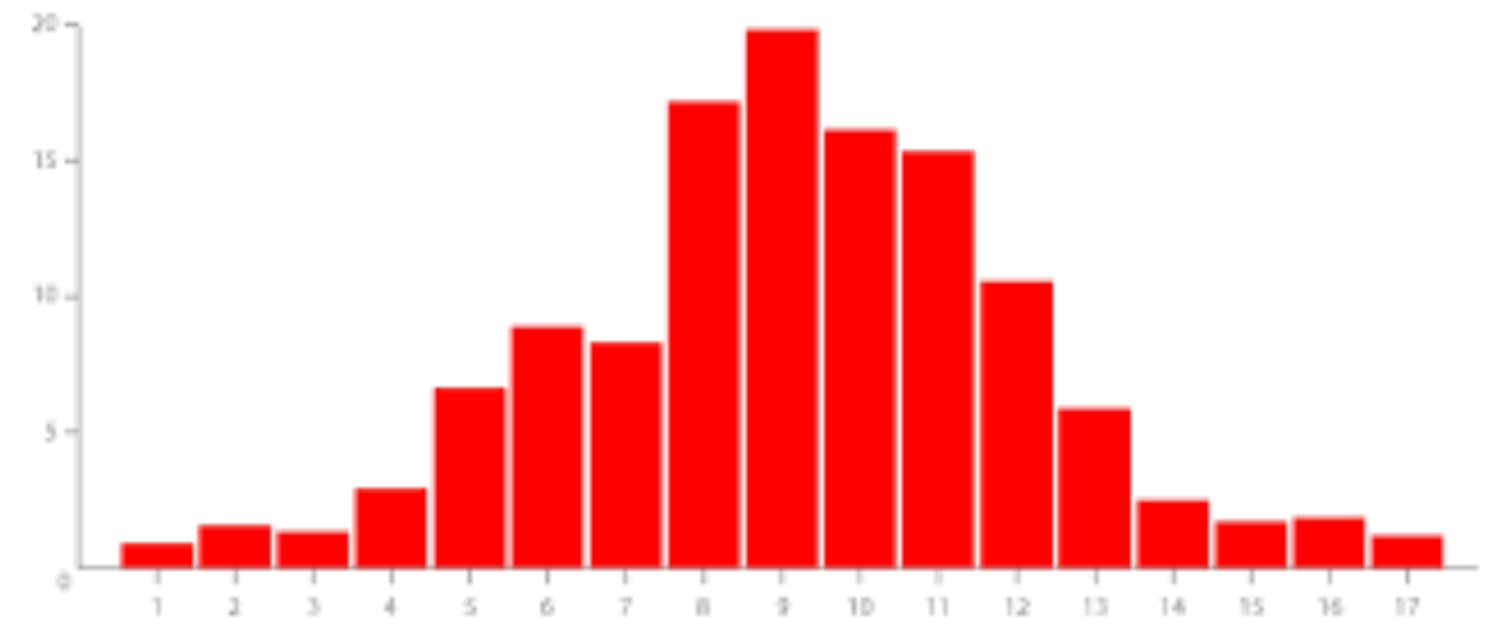
- Номинальные
- Порядковые



# Первичный анализ данных. Визуальный анализ данных

**EDA** - это критически важный процесс первоначального исследования данных с помощью сводной статистики и визуализаций с 4 основными целями:

- Выявить паттерны/зависимости
- Заметить аномалии
- Сформировать гипотезы
- Проверить первичные предположения



# Первичный анализ данных.

## Визуальный анализ данных

1. Как собираются данные?
2. Сколько и каких переменных?
3. Что обозначает каждая переменная, какие единицы измерения и как она собирается?
4. Есть ли пропущенные значения и как они появились?
5. Есть ли аномалии в распределениях?
6. Есть ли корреляции и другие зависимости?



# ПРАКТИКА



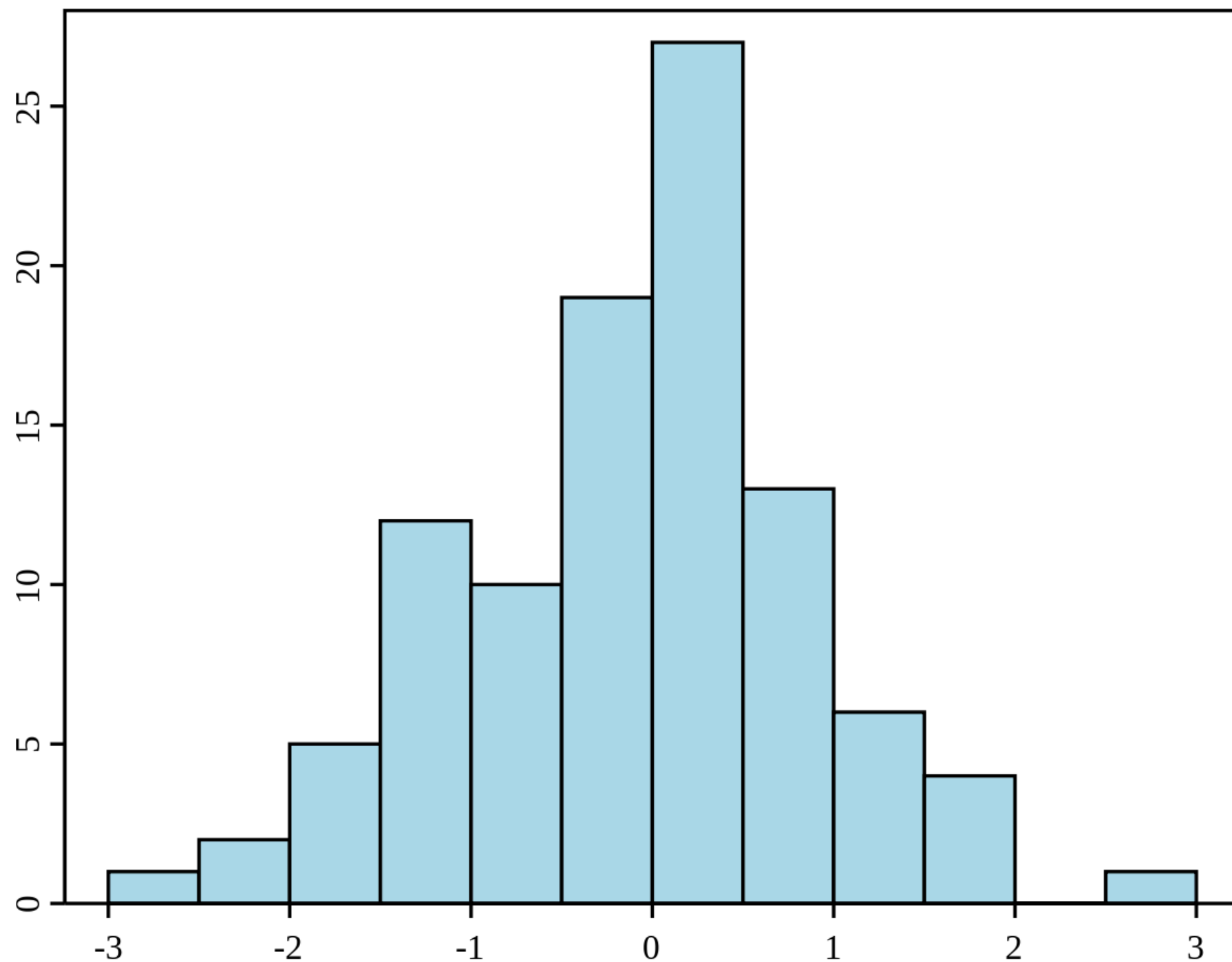
# Обработка категориальных переменных. One-hot/ Label encoding

Company Name	Categorical value	Price
VW	1	20.000
Acura	2	10.011
Honda	3	50.000
Honda	3	10.000

VW	Acura	Honda	Price
1	0	0	20.000
0	1	0	10.011
0	0	1	50.000
0	0	1	10.000



# Обработка категориальных переменных. Bins to number



# Обработка категориальных переменных.

## Bins to number

User_ID	Product_ID	Gender	Age	C
1000001	P00069042	F	0-17	
1000001	P00248942	F	0-17	
1000001	P00087842	F	0-17	
1000001	P00085442	F	0-17	
1000002	P00285442	M	55+	
1000003	P00193542	M	26-35	
1000004	P00184942	M	46-50	
1000004	P00346142	M	46-50	
1000004	P0097242	M	46-50	
1000005	P00274942	M	26-35	
1000005	P00251242	M	26-35	

Средняя или медиана



Верхняя или нижняя граница



User_ID	Product_ID	Gender	Age	New_Age	Occ
1000001	P00069042	F	0-17	14	
1000001	P00248942	F	0-17	14	
1000001	P00087842	F	0-17	14	
1000001	P00085442	F	0-17	14	
1000002	P00285442	M	55+	60	
1000003	P00193542	M	26-35	30	
1000004	P00184942	M	46-50	47	
1000004	P00346142	M	46-50	47	
1000004	P0097242	M	46-50	47	
1000005	P00274942	M	26-35	30	
1000005	P00251242	M	26-35	30	

User_ID	Product_ID	Gender	Age	Lower_Age	Upper_Age
1000001	P00069042	F	0-17	0	17
1000001	P00248942	F	0-17	0	17
1000001	P00087842	F	0-17	0	17
1000001	P00085442	F	0-17	0	17
1000002	P00285442	M	55+	55	80
1000003	P00193542	M	26-35	26	35
1000004	P00184942	M	46-50	46	50
1000004	P00346142	M	46-50	46	50
1000004	P0097242	M	46-50	46	50
1000005	P00274942	M	26-35	26	35
1000005	P00251242	M	26-35	26	35





# Обработка категориальных переменных.

## По бизнес-логике

По бизнес логике

Zip Code	District
110044	South Delhi
110048	South Delhi
110049	South Delhi
110006	North Delhi
110007	North Delhi
110058	West Delhi
110059	West Delhi
110063	West Delhi
110064	West Delhi

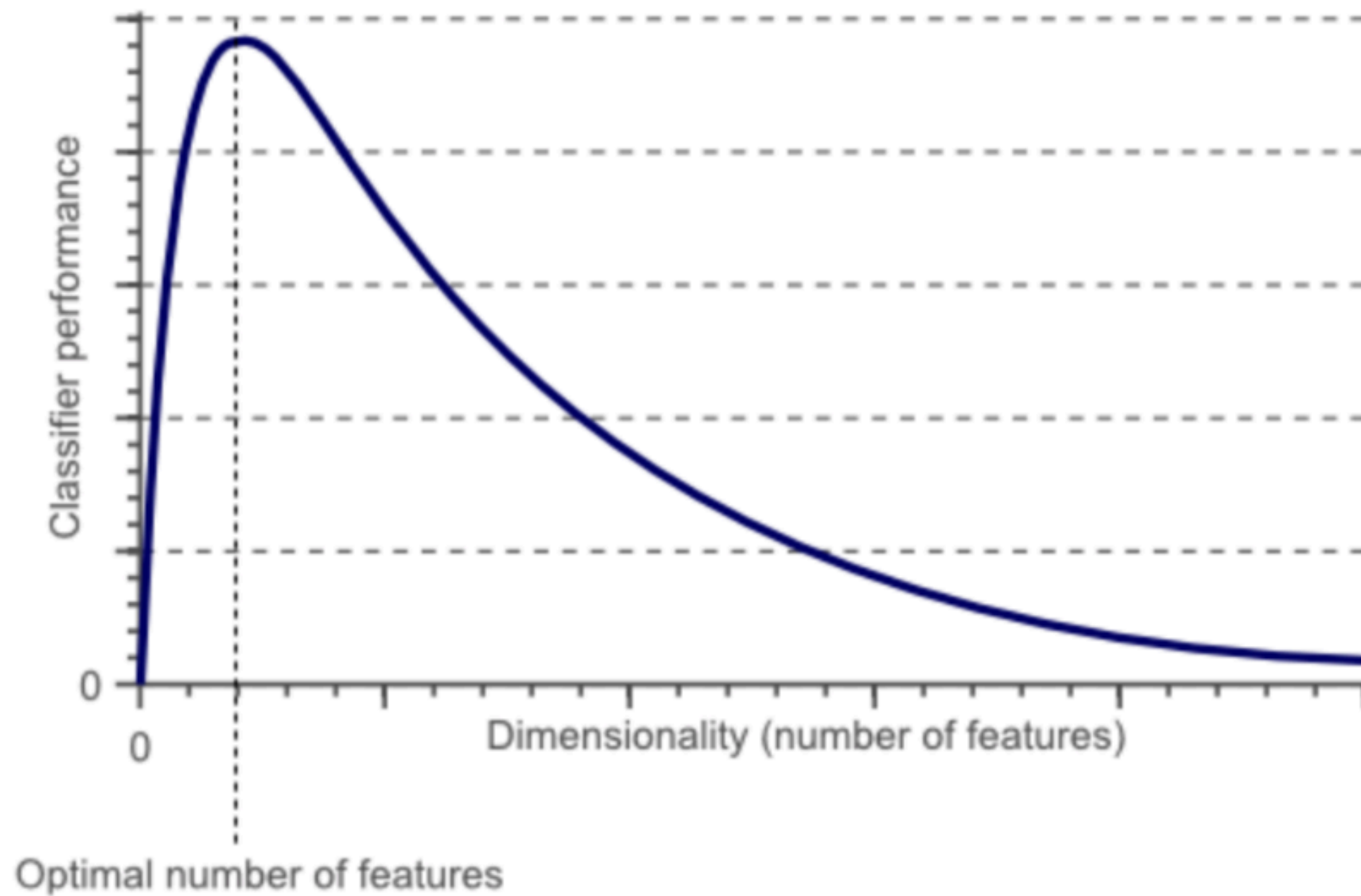
По доле таргетинга

Based on Response Rate

Levels	Response_Rate	New_Level
HA014	98%	1
HA001	97%	1
HA003	93%	1
HA009	81%	2
HA015	75%	3
HA010	73%	3
HA006	66%	4
HA017	60%	4
HA007	49%	5
HA004	36%	6
HA005	31%	6
HA012	28%	7



# Проклятие размерности



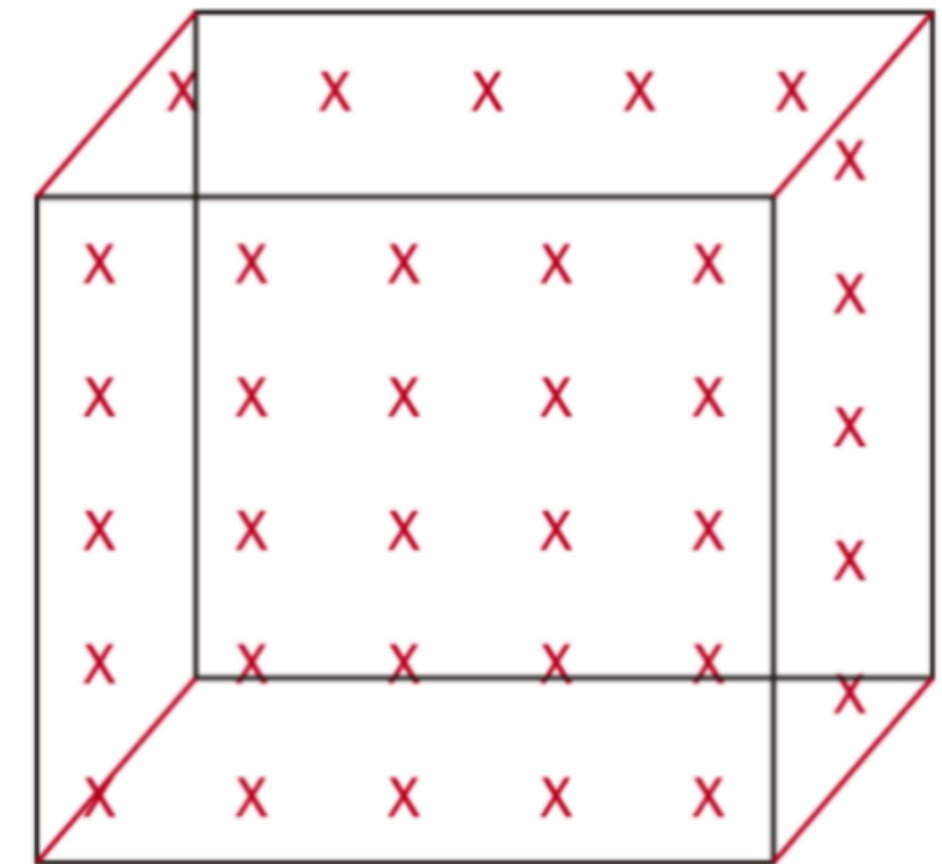
# Проклятие размерности



Одно измерение - 5 точек



Два измерения - 25 точек



Три измерения - 125 точек



---

# Спасибо за внимание!

---

**Сапрыкин Артур**  
Data Scientist



[fb.com/asaprykin92](https://fb.com/asaprykin92)



[asaprykin92@gmail.com](mailto:asaprykin92@gmail.com)

