

Biomedical Data Analysis / Анализ данных в биологии и медицине

Lecturer: Katerina Oleynikova

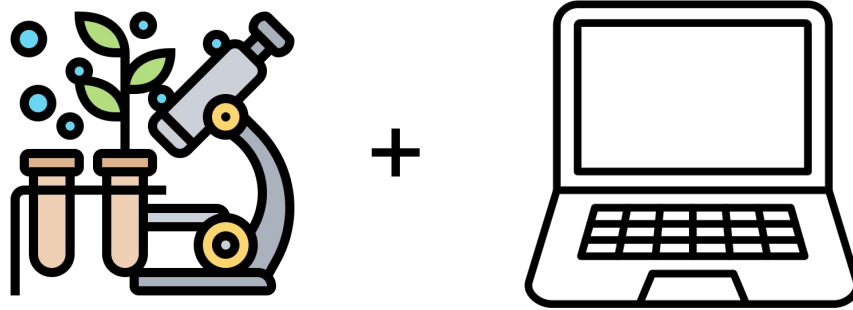
October, 2023

Rus. version



Для чего необходима **биоинформатика**?

На сегодня существует **большой объем данных в области молекулярной биологии**



Знание биологии и умение решать биологические задачи компьютерными методами

Exploratory Data Analysis (EDA) / Разведочный анализ данных

- Что на входе (что у нас имеется): какие-то данные в текстовом, табличном виде, содержащую совершенно любую информацию, например, статистика по пациентам, уровни экспрессии генов, и т.д.
- Примеры данных из различных областей (в том числе, с разбором того, как можно работать над ними) можно найти на kaggle:
<https://www.kaggle.com>
- Что хотим получить на выходе (какой результат мы ожидаем): понять структуру данных, выявить аномалию в них (выбросы / outliers), преобразовать данные в наиболее информативный вид - преимущественно методами визуализации.

Примеры данных (взяты с kaggle)

- [Home](#)
- [Competitions](#)
- [Datasets](#)
- [Models](#)
- [Code](#)
- [Discussions](#)
- [Learn](#)
- [More](#)
- [Your Work](#)
- [View Active Events](#)

VIKAS UKANI · UPDATED 3 YEARS AGO

Parkinson's Disease Data Set

To Detecting Parkinson's Disease – Python Machine Learning Project

88
New Notebook
Download (16 kB)

Data Card

Code (25)

Discussion (0)

About Dataset

Parkinson's Data Set

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds to one of 195 voice recordings from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to the "status" column which is set to 0 for healthy and 1 for PD.

The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient, the name of the patient is identified in the first column. For further information or to pass on comments, please contact Max Little (little '@' robots.ox.ac.uk).

Further details are contained in the following reference -- if you use this dataset, please cite:

[Download \(195 MB\)](#)

DARSHAN DESHPANDE · UPDATED 4 YEARS AGO

COVID-19 Detection X-Ray Dataset

X-rays of COVID-19, Bacterial/Viral Pneumonia Patients and Normal People

42
New Notebook
Download (3 MB)

Data Card

Code (8)

Discussion (1)

About Dataset

Content

Following the new Coronavirus(COVID-19) outbreak, data scientists are looking for patterns and methods to speed up testing and diagnostics. This dataset consists of X-ray(PA-CXR) images of COVID-19,bacterial and viral pneumonia patients and Normal people. This dataset is uploaded with hopes that some pattern can be detected which could help Medical Professionals.

Content

The dataset has three directories- TrainData,ValData and NonAugmentedTrain.

[Download \(195 MB\)](#)

HIRTEHRTER · UPDATED A YEAR AGO

HIV / AIDS

HIV = AIDS? IDK I am not a medical guy lol (For reasons : This is a joke)

1
New Notebook
Download (3 MB)

Data Card

Code (1)

Discussion (0)

Usability

Attribution 4.0 International (CC BY-SA 4.0)


Expected update frequency

Weekly

Tags

Health Conditions
Coronavirus

Также на kaggle есть **ноутбуки** самих пользователей сайта - примеры работ, как можно производить анализ на сетах данных.

About Dataset	Usability 
<p>HIV (human immunodeficiency virus) is a virus that attacks the body's immune system. If HIV is not treated, it can lead to AIDS (acquired immunodeficiency syndrome). There is currently no effective cure. Once people get HIV, they have it for life. But with proper medical care, HIV can be controlled.</p> <p>Symptoms: Influenza-like illness; Fatigue...</p> <p>Treatments: Management of HIV/AIDS</p> <p>Type of infectious agent: Virus</p> <p>AIDS (acquired immune deficiency syndrome) is the name used to describe a number of potentially life-threatening infections and illnesses that happen when your immune system has been severely damaged by the HIV virus. While AIDS cannot be transmitted from 1 person to another, the HIV virus can.</p>	<p>8.24</p> <p>Licence</p> <p>CC0: Public Domain</p>
	<p>Expected update frequency</p> <p>Never</p> <p>Tags</p> <div> Tabular Beginner Intermediate Advanced </div>



MRIDUL GUPTA · UPDATED 3 YEARS AGO



27

New Notebook



Download (82 kB)



Dota 2 - all hero data - 7.28b

Hero stats, roles, att. for Dota 2: 7.27d - 7.28b



Data Card

Code (4)

Discussion (1)

About Dataset

Context

I couldn't find any data related to hero attributes, roles, etc. on Kaggle. Hence, I decided to write a simple web scrapper and put in here.

Content

It has 3 files.

1. dota_heroes.json - This file has data of all heroes ranging from hp to legs.
2. hero_category.json - This file has info on classes of heroes - Strength, Int, Agi. You know what I'm talking about.
3. hero_roles.json - This file contains what roles the hero fits in - Carry, Nuker, Supports, etc.

Usability ⓘ

9.12

License

CC0: Public Domain

Expected update frequency

Monthly

Tags

Games

Video Games

Что из себя представляет **notebook**?

- **Notebook** - простыми словами - *это то окошко, где мы пишем наш код* (в зависимости от того языка программирования, который мы выбрали (например, **python**, **R**, **SQL**, etc.)
- **Jupyter Notebook** (также известен как IPython Notebook) для работы на языке **python***



*Внутри jupyter notebook можно писать не только на python, но и, например, на R. Для того, чтобы это сделать, предварительно качаются специальные библиотеки для поддержки R языка внутри нашего jupyter ноутбука.

Contents

- 1 Intro
- 2 Load step tracker data
- 3 Analyze and visualize the raw data
- 4 Analyze and visualize transforms of the data
- 5 **Infer Steps**
- 6 Simulate the "real-time" system

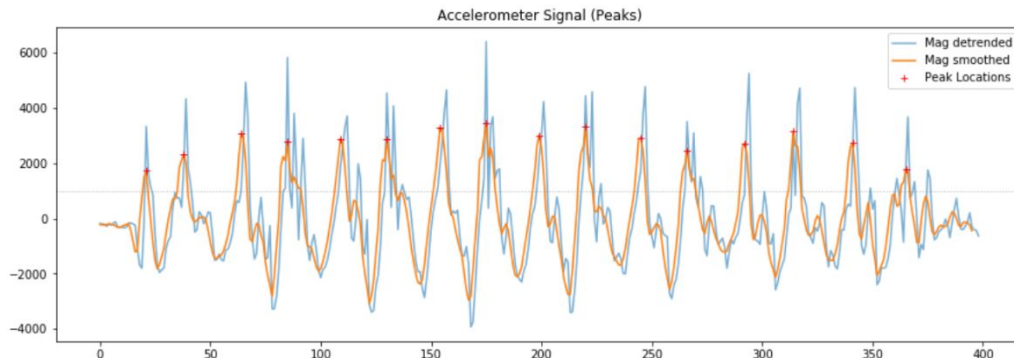
```
peak_indices, peak_properties = sp.signal.find_peaks(mag_filtered, height=min_peak_height, distance=min_distance_between_peaks)
print("We've detected:", len(peak_indices), "peaks")
print("Average peak:", np.average(mag_filtered[peak_indices]), "SD=", np.std(mag_filtered[peak_indices]))
print("Min peak:", np.min(mag_filtered[peak_indices]))
print("Max peak:", np.max(mag_filtered[peak_indices]))
```

```
# Plots the peaks
axes.axhline(y=min_peak_height, linewidth=1, linestyle=":", alpha=0.6, color='gray')
axes.plot(peak_indices, mag_filtered[peak_indices], 'y+', color="red", label="Peak Locations")

# set the title and show the legend
axes.set_title("Accelerometer Signal (Peaks)")
axes.legend()
```

```
Min num samples between peaks: 14.50909090909091
We've detected: 16 peaks
Average peak: 2774.901806021425 SD= 484.5612644848081
Min peak: 1722.5138963612142
Max peak: 3448.133316689275
```

Out[109]: <matplotlib.legend.Legend at 0x1a16bb594a8>



Google Colab



<https://colab.research.google.com>

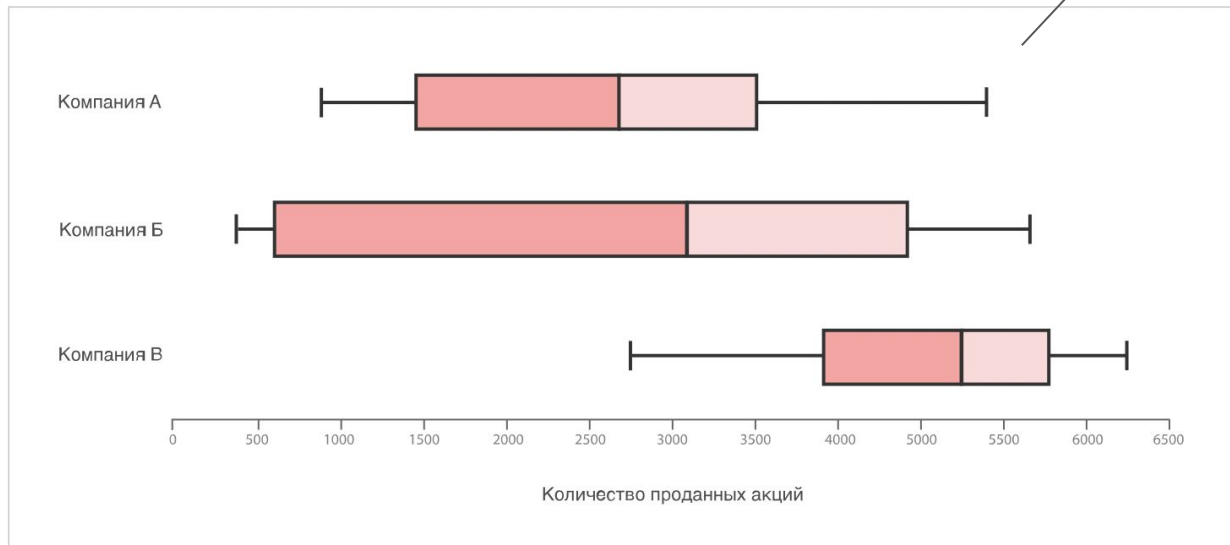
- + не нужно устанавливать никакие пакеты и сам python на свой компьютер
- + идеален для анализа небольших сетов данных или для пробы в первые разы анализа
- + можно подкачивать свои данные с локального компьютера
- не очень хорош для анализа больших данных - может просто отрубиться соединение
- для анализа данных по работе или там, где требуется конфиденциальность, и т.д., лучше все же писать код не здесь, а локально либо через рабочий сервер

Exploratory Data Analysis (EDA)

I. Инструменты для визуализации:

1. Box plot / Ящик с усами / Диаграмма размаха

подходит для визуального представления как одного, так и нескольких одномерных наборов данных



Box plot

- представление данных через **квартили**

Квартіль (от фр. quartile, «четверть»)

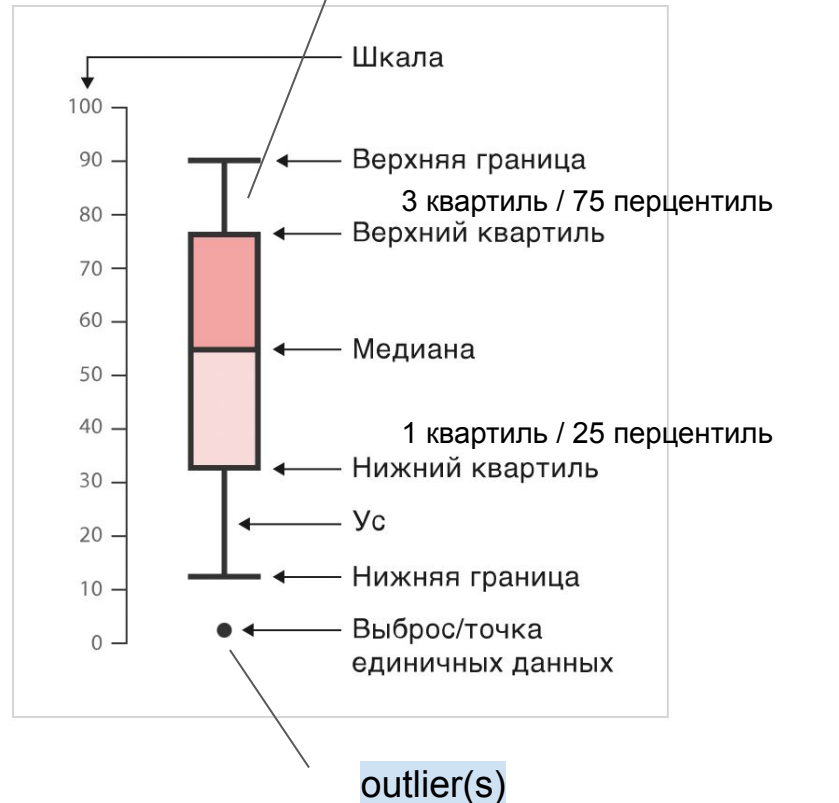
Квартили — числовые значения некого признака, которые делят упорядоченную по возрастанию совокупность данных на четыре равных части.

Квартили делят совокупность на четыре части, поэтому квартилей бывает три варианта: первый (нижний), второй(средний), третий (верхний).

Второй квартиль это и есть **медиана**.

Медиана - это срединное значение набора чисел, то есть число, которое находится в середине этого набора, если его упорядочить по возрастанию.

Структура



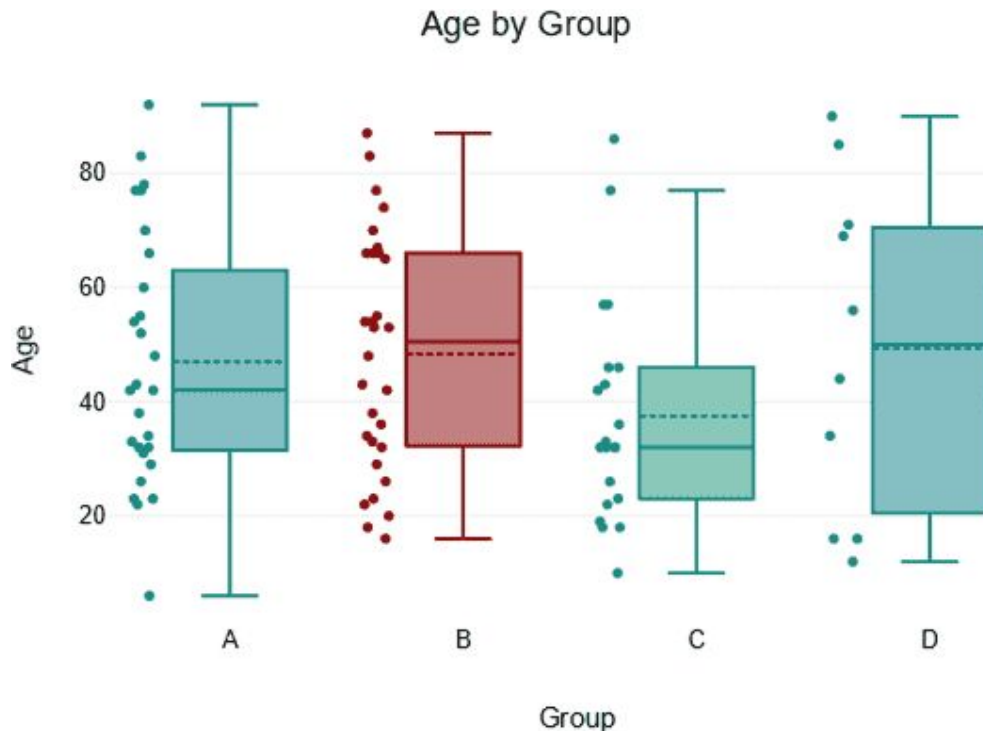
Box plot

Основной + такого графика - его компактность

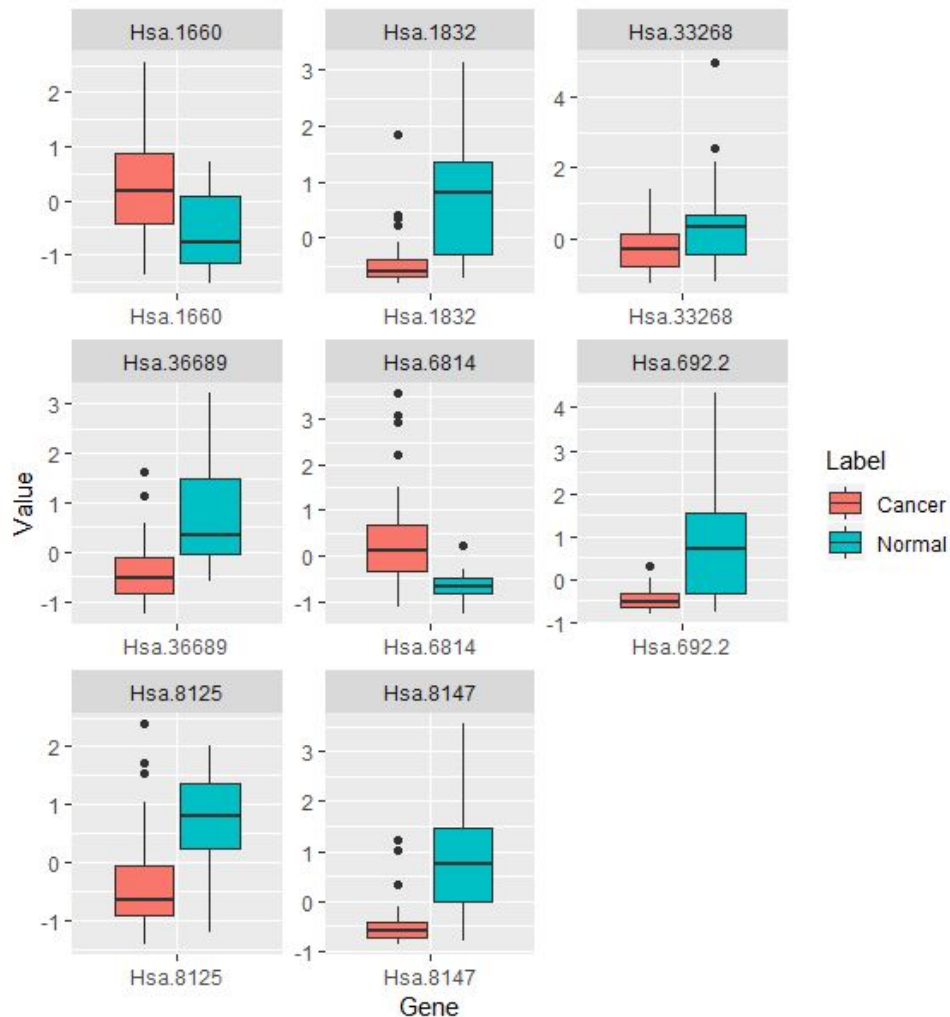
Виды наблюдений, которые можно сделать на основе ящика с усами:

- Каковы ключевые значения, например: средний показатель, медиана etc.
- Есть ли выбросы и каковы их значения.
- Симметричны ли данные.
- Насколько плотно сгруппированы данные.
- Смещены ли данные и, если да, то в каком направлении.

Прямая линия внутри ящика - медиана, пунктирной показывают среднее значение набора данных



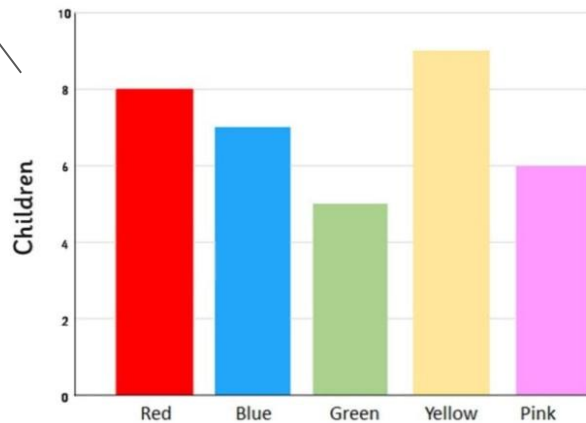
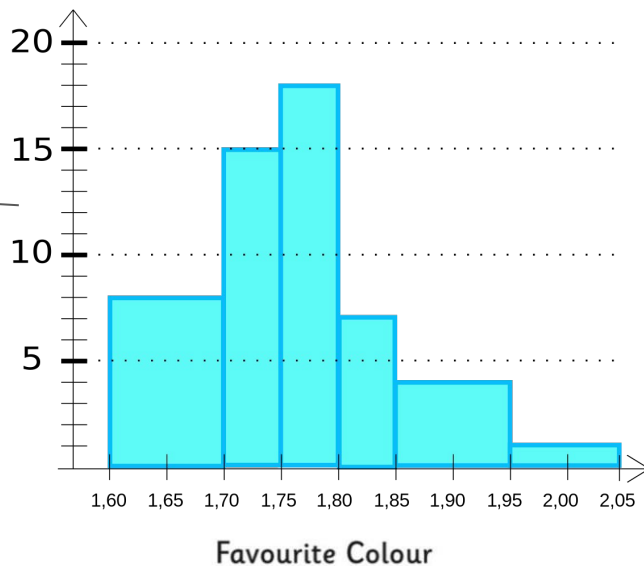
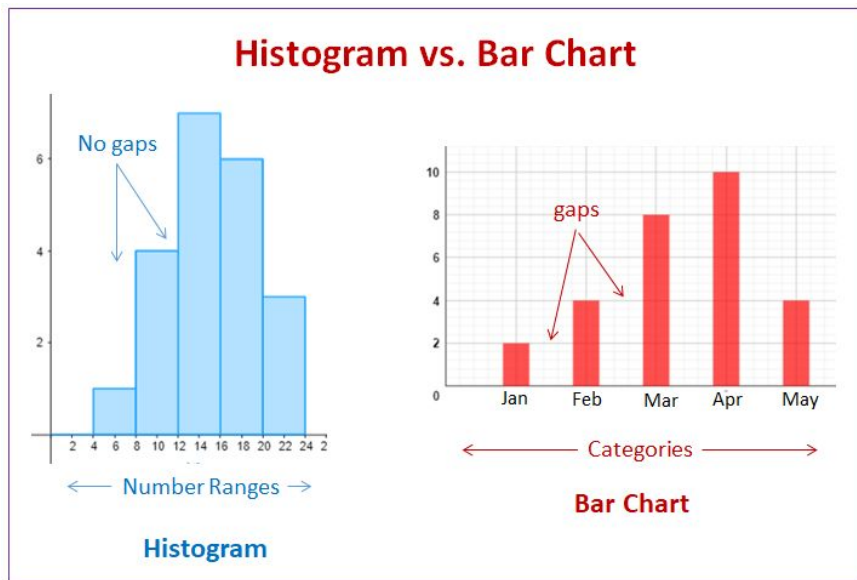
Boxplots of differential expression level between normal and cancer samples on eight genes



- еще способ представления одномерных данных в графическом виде

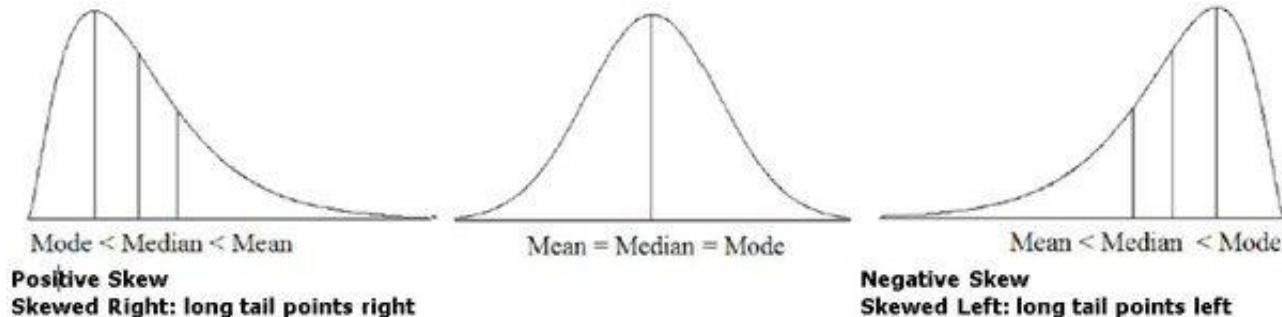
2. Histogram / Гистограмма

3. Bar chart / Столбчатая диаграмма



Гистограмма

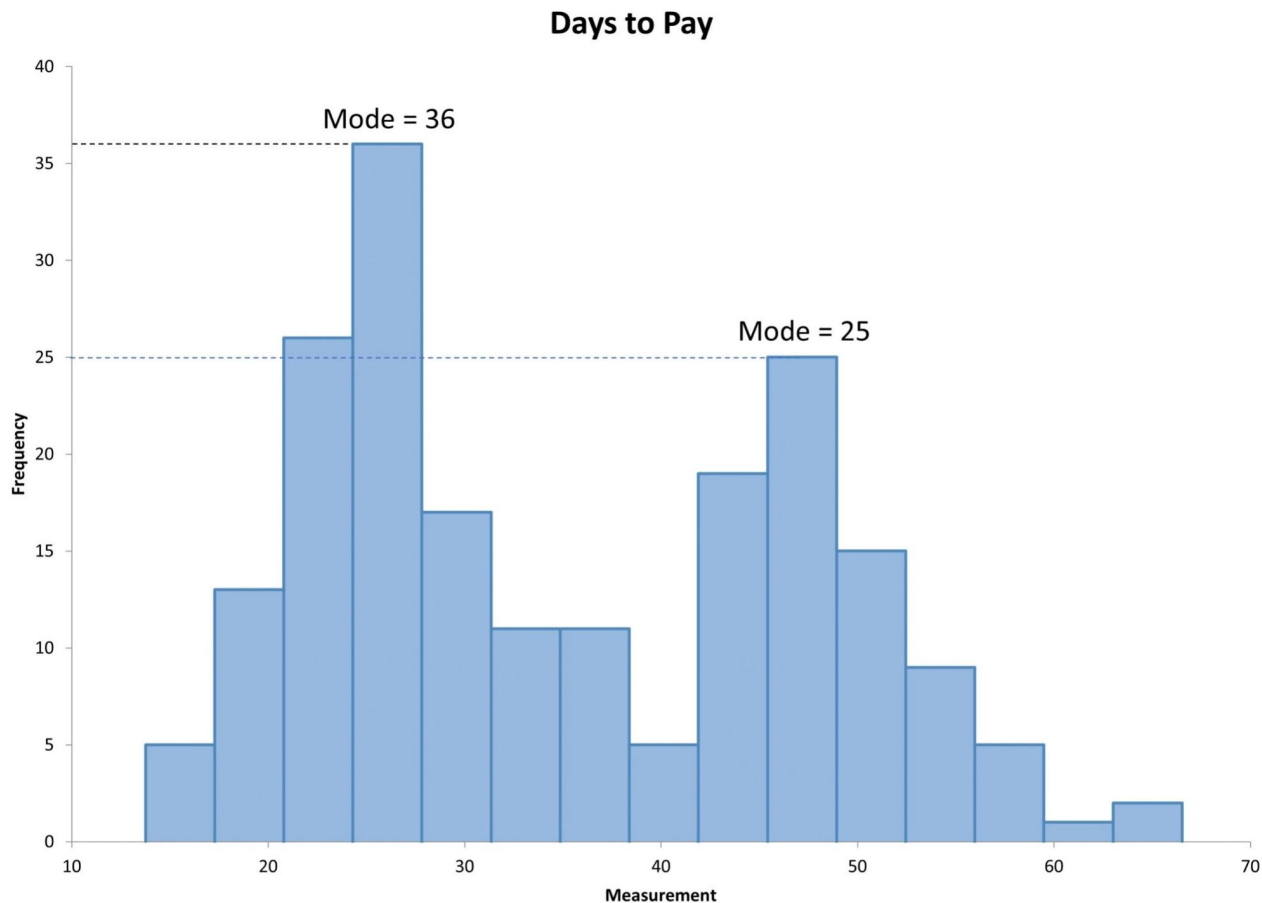
- вид статистического графика, показывающего распределение величины
- используется для проверки на нормальность распределения

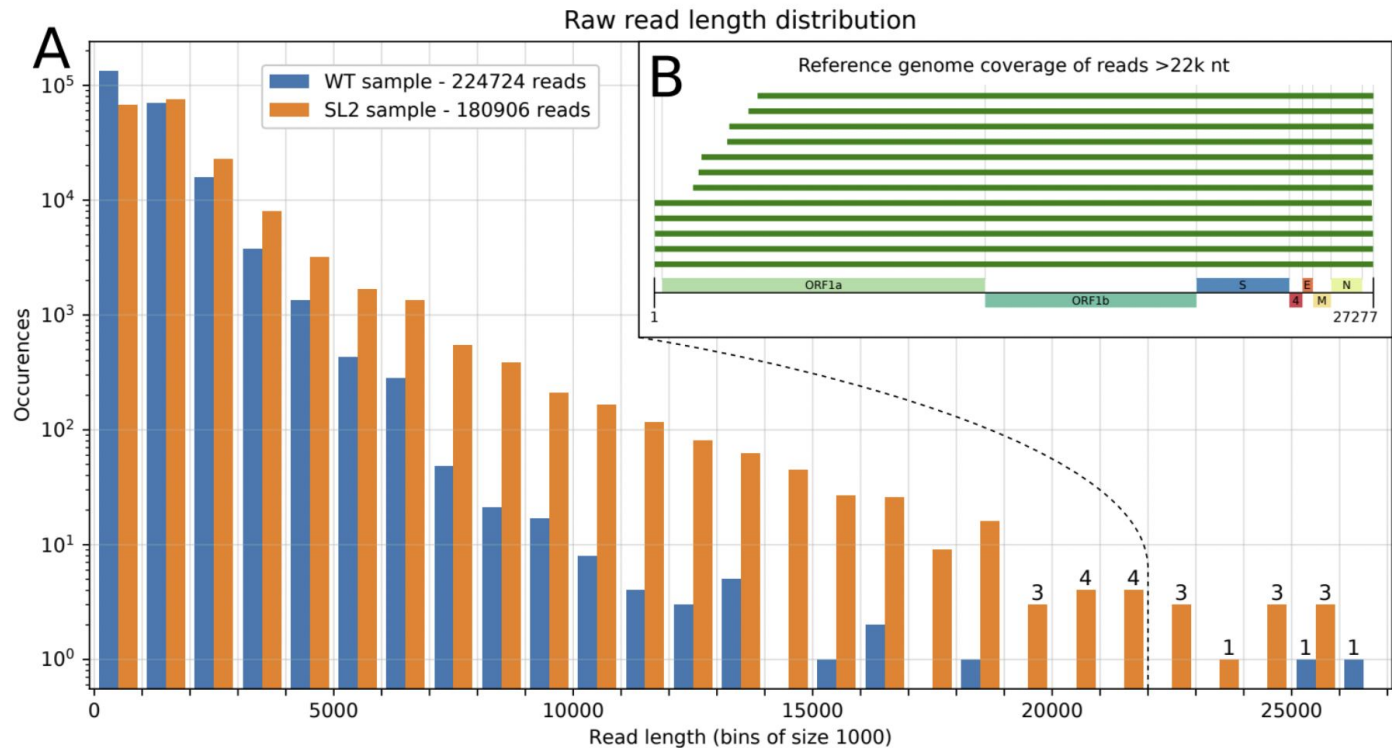


Мода — одно или несколько значений во множестве наблюдений, которое встречается наиболее часто (мода = типичность).

Гистограмма

Бимодальное
распределение





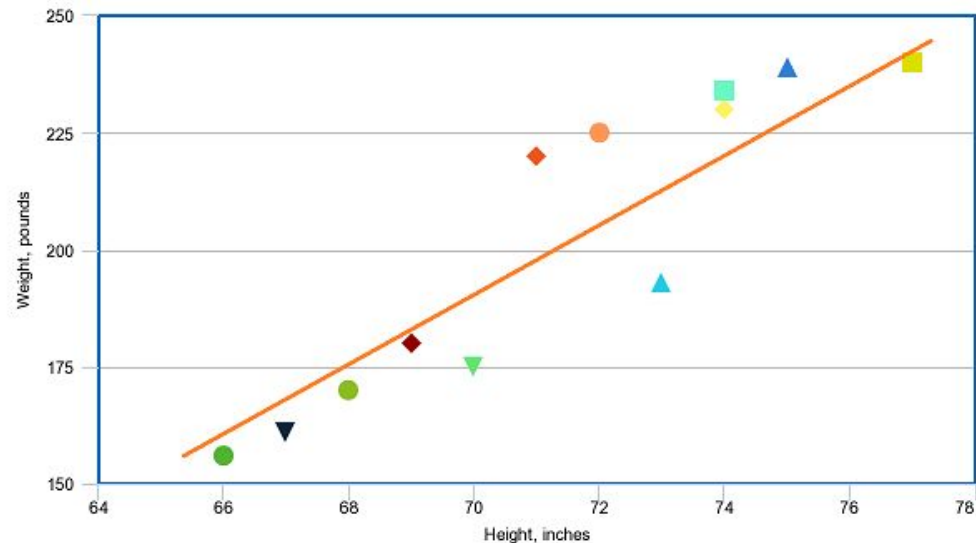
A - LENGTH DISTRIBUTION HISTOGRAM OF OXFORD NANOPORE RAW READS

Да, она похожа на положительно скошенную, но поскольку данные резко обрываются с одной стороны (в начале их просто нет) - такой вид называется усеченным видом диаграммы.

1. Какой тип (вид) гистограммы?
2. Какой образец характеризуется большим числом ридов в целом?
3. Какой образец характеризуется большим числом именно длинных ридов в целом?

4. Scatter plot / Диаграмма рассеяния

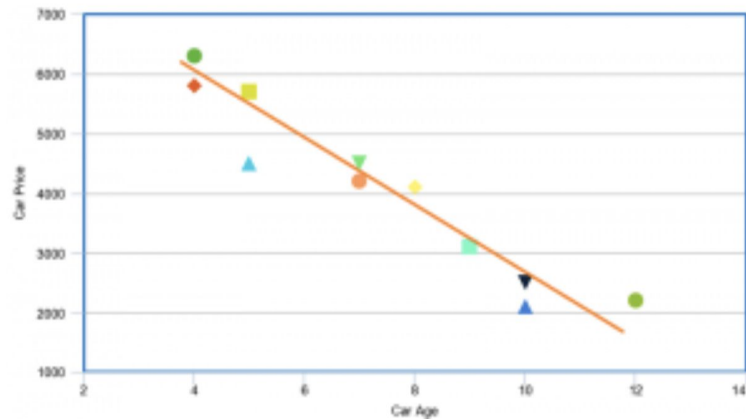
- способ представления двумерных данных в графическом виде (когда данные характеризуются двумя признаками, например, пациенты определяются 1) по возрасту и 2) наличию/отсутствию той или иной болезни)
- хорошо подходит для показа, как одна переменная (признак) зависит от другой переменной (другого признака)
- следовательно, можно попытаться предсказать поведение переменных по построению линии тренда (тип корреляции двух переменных)



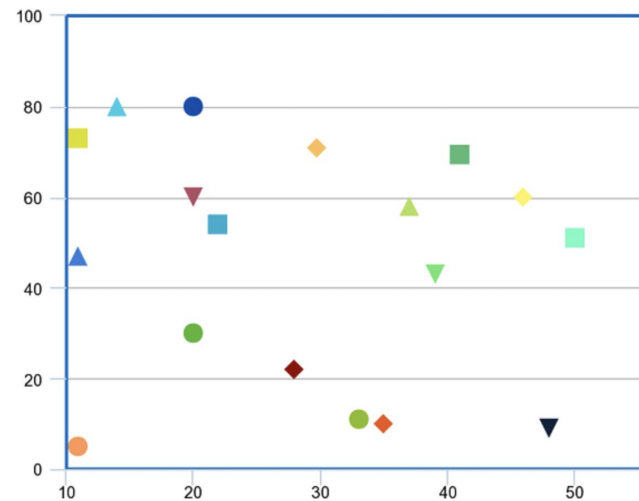
Корреляция переменных
бывает 3 типов:

- 1) Диаграмма рассеяния с положительной корреляцией между двумя переменными

Scatter plot



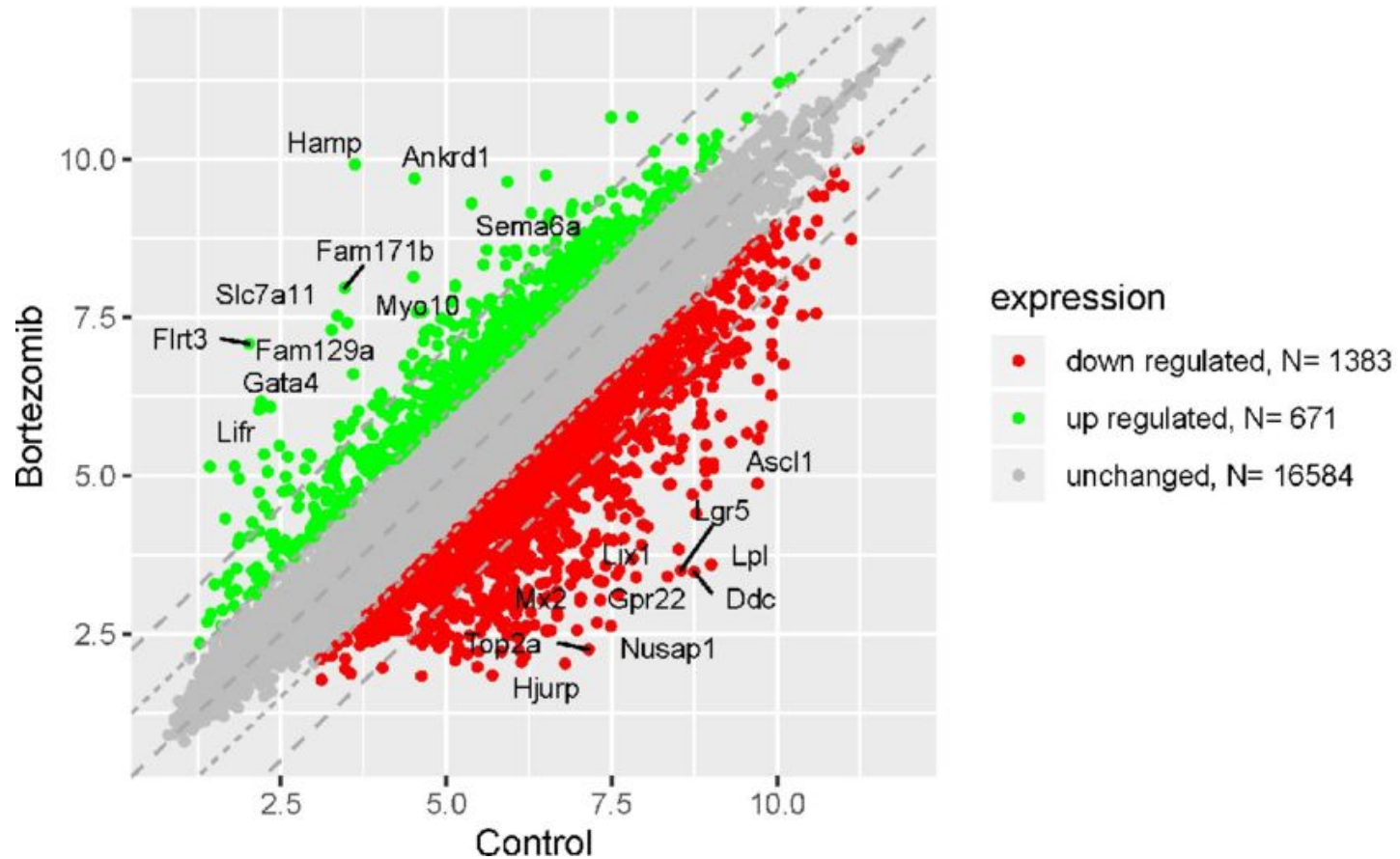
2) Диаграмма рассеяния с отрицательной корреляцией между двумя переменными



3) Корреляция между двумя рассматриваемыми переменными отсутствует

The scatter plot of global gene expression in bortezomib-treated PC12-derived nerve cells compared to the control cells.

Genes are represented by dots (red colour: Downregulation; green colour: Upregulation).



1. Какой тип корреляции между данными?
2. Как можно охарактеризовать данные? i.e., где сильнее выражена экспрессия upregulated / downregulated genes, если сравнить два вида клеток друг с другом?

Exploratory Data Analysis (EDA)

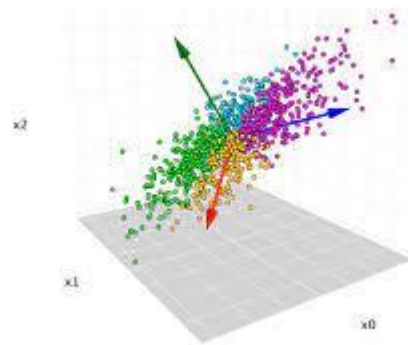
II. Снижение размерности данных

1. Principal component analysis (PCA) / Метод главных компонент

- один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации
- подходит для анализа многомерных данных (характеризуются несколькими признаками, в отличие от одномерных данных)

Основной задачей метода главных компонент является *замена исходных данных на некие агрегированные значения в новом пространстве*, решая при этом две задачи - первая из которых состоит в 1) объединении наиболее важных (с точки зрения минимизации среднеквадратичной ошибки) значений в меньшее количество параметров, но более информативных (уменьшение размерности пространства данных), а вторая - 2) уменьшить шум в данных.

Применяется во многих областях, таких как распознавание образов, компьютерное зрение, биоинформатика, etc.



Пример того, как выглядит PCA

PCA: step by step

1. Стандартизация

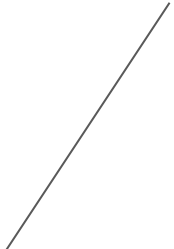
Зачем нужна стандартизация? - PCA очень чувствителен к значениям исходных переменных.

То есть, если есть большие различия между значениями переменных, то переменные с большими диапазонами будут преобладать. Например, x_1 , которая находится в диапазоне от 0 до 100, будет преобладать над x_2 , которая находится в диапазоне от 0 до 1. Это приведет к необъективным результатам. Преобразование данных к одному масштабу может предотвратить эту проблему.

Цель первого шага (стандартизации) - стандартизировать диапазоны исходных переменных, чтобы избавиться от большого разброса значений.

После завершения стандартизации все переменные будут преобразованы в единый масштаб.

Математически это можно сделать путем вычитания среднего (mean) и деления полученной разности на стандартное отклонение (sd) для каждой переменной (value).


$$z = \frac{value - mean}{standard\ deviation}$$

PCA: step by step

2. РАСЧЕТ МАТРИЦЫ КОВАРИАЦИИ

В теории вероятностей и статистике **ковариация** является мерой совместной изменчивости двух случайных величин.
Ковариация — это способ показать то, насколько два массива данных линейно зависимы между собой.

Цель второго шага - увидеть, есть ли какая-либо связь между переменными во входных данных. Иногда переменные сильно коррелированы (имеют зависимость). А это значит, что данные содержат избыточную информацию.

Итак, чтобы идентифицировать взаимосвязи, нужно вычислить *ковариационную матрицу*.

Ковариационная матрица представляет собой симметричную матрицу размера $p \times p$ (где p - количество измерений (признаков)), элементами которой являются ковариации всех возможных пар исходных переменных. Например, для 3-мерного набора данных с 3 переменными x , y и z ковариационная матрица представляет собой матрицу 3×3 , состоящую из следующих элементов:

$$\begin{bmatrix} \text{Cov}(x, x) & \text{Cov}(x, y) & \text{Cov}(x, z) \\ \text{Cov}(y, x) & \text{Cov}(y, y) & \text{Cov}(y, z) \\ \text{Cov}(z, x) & \text{Cov}(z, y) & \text{Cov}(z, z) \end{bmatrix}$$

Матрица ковариаций для трехмерного случая

Что можно понять по матрице ковариации?
Имеет значение знак ковариации между двумя переменными:

- + переменные коррелированы (увеличиваются/уменьшаются вместе)
- переменные обратно коррелированы (одна увеличивается, другая уменьшается)

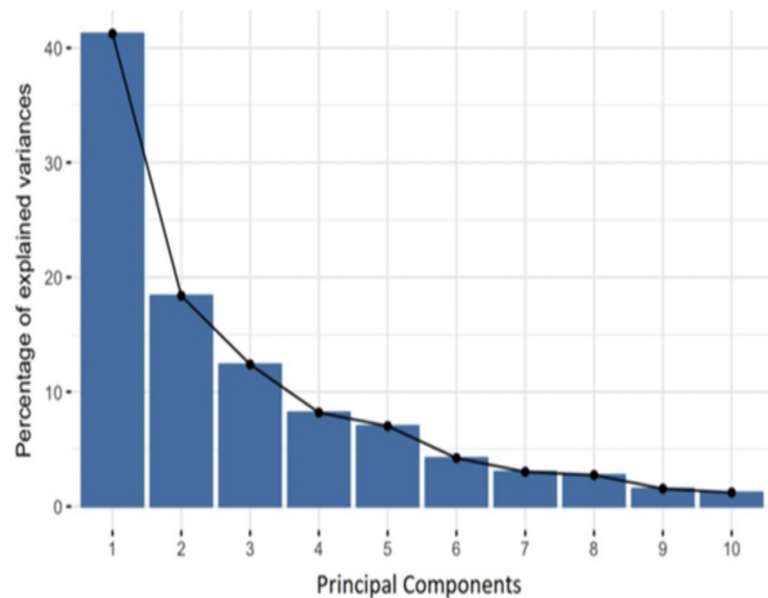
PCA: step by step

Из шага 2, на основании матрицы ковариации, мы извлекаем и получаем новые переменные, которые не коррелируют между собой. Они будут называться компонентами (линейные комбинации исходных переменных)*.

3. Определение главных компонент

Главные компоненты - это новые переменные, с помощью которых мы и будем описывать наши объекты. 'Ненужные' в данном случае отбрасываются

Например, 10-мерные данные дают 10 главных компонент, но PCA попытается поместить максимум возможной информации в первый компонент, затем максимум оставшейся информации во второй и так далее, пока не получится следующая диаграмма:

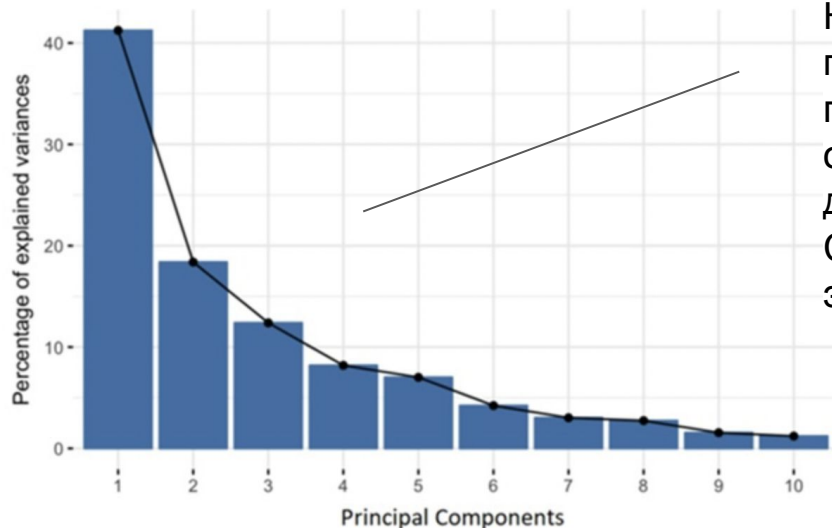


*при этом будет меняться система измерения. Но после всегда можно вернуться в исходную систему

PCA: step by step

3. Определение главных компонент

! В данной лекции мы упускаем подсчеты 1) собственных векторов и 2) собственных значений компонент с целью упрощения понимания материала

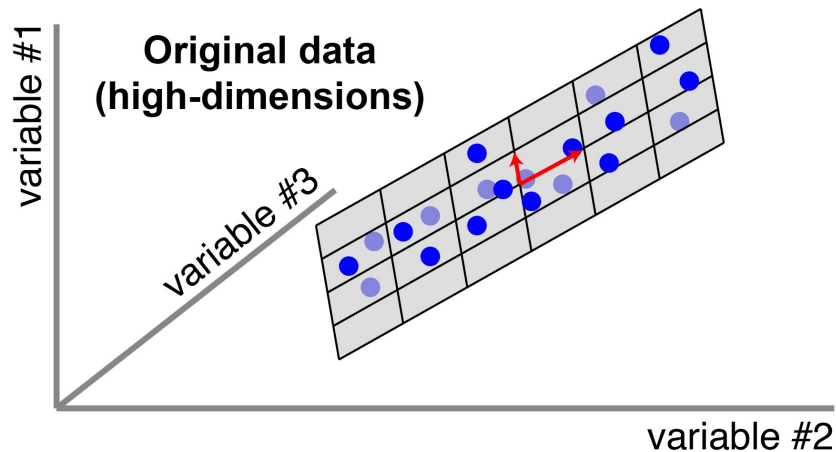


Например, исходя из этого графика, выберем столько главных компонент, которые бы описывали $M\%$ дисперсии* данных (например, 80-90%). Сколько компонент нам для этого достаточно взять?

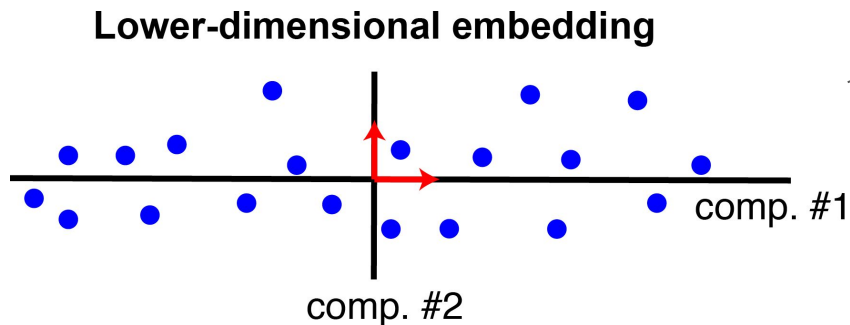
Таким образом, на этом шаге мы можем отбросить компоненты с маленькой информацией, тем самым уменьшая размерность без особой потери данных.

*дисперсия показывает охват информации

PCA: пример



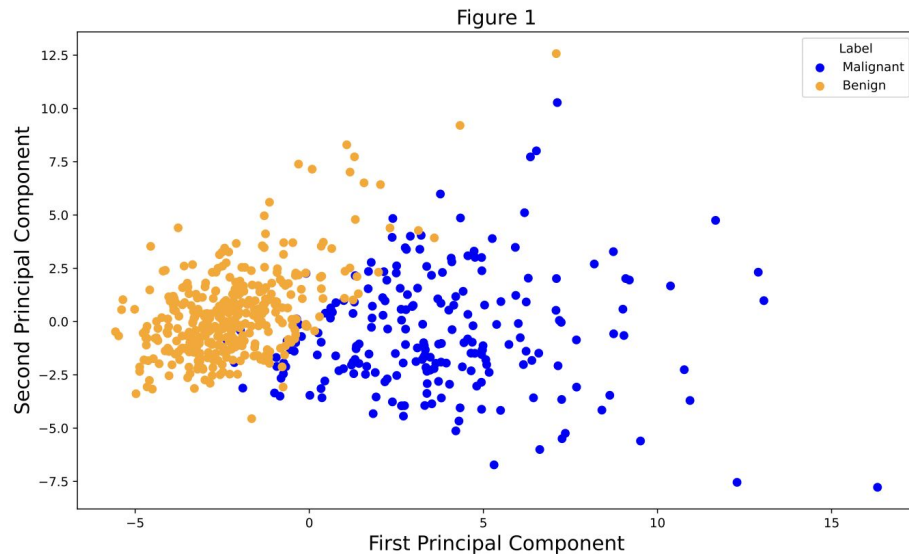
На входе: несложный сет данных из 3 переменных. Но что, если данные будут характеризоваться, например, 100 переменными (признаками)?



Результат PCA по первым двум компонентам

PCA: breast cancer example


<https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data> Данные



Результатом PCA является диаграмма рассеяния (scatter plot of PCA), которая отображает некую кластеризацию 2 типов рака груди по первым двум компонентам проведенного анализа

Figure 1 shows a scatterplot colored by the type of breast cancer using the Matplotlib package in python.

Датасет для семинара

 R ARUN · UPDATED 4 YEARS AGO

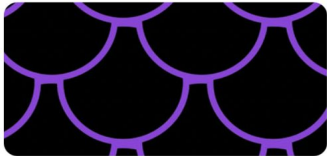
▲ 2

New Notebook

Download (890 kB)

⋮

Mice Protein Expression Data Set



[Data Card](#) [Code \(2\)](#) [Discussion \(0\)](#)

<https://www.kaggle.com/datasets/rarunk4495/mice-protein-expression-data-set>