**Лабораторная работа 2**

**Изучение методов и алгоритмов предобработки, анализа и визуализации данных.**

Numpy - это библиотека Python для вычислительно эффективных операций с многомерными массивами, предназначенная в основном для научных вычислений.

import numpy as np

Pandas - это библиотека Python, предоставляющая широкие возможности для анализа данных. С ее помощью очень удобно загружать, обрабатывать и анализировать табличные данные с помощью SQL-подобных запросов.

import pandas as pd

Основными структурами данных в Pandas являются классы Series и DataFrame. Первый из них представляет собой одномерный индексированный массив данных некоторого фиксированного типа. Второй - это двумерная структура данных, представляющая собой таблицу, каждый столбец которой содержит данные одного типа. Можно представлять её как словарь объектов типа Series.

**Описательные статистики**

**Меры центральной тенденции**

**Среднее значение**

Данная характеристика описывает среднее значение в наборе данных. Вычислить её довольно просто: сложите все значения и разделите полученную сумму на количество значений.

В случае со средним значением «серединой» датасета будет среднее арифметическое его значений. Среднее значение отражает типичный показатель в наборе данных. Если мы случайно выберем один из показателей, то, скорее всего, получим значение, близкое к среднему.

**Медиана**

Следующая мера центральной тенденции, о которой пойдёт речь, — медиана. Медиана, как и среднее значение, нужна для определения типичного значения в наборе данных, но при этом не требует вычислений.

Чтобы найти медиану, данные нужно расположить в порядке возрастания. Медианой будет значение, которое совпадает с серединой набора данных. Если количество значений чётное, то берётся среднее двух значений, которые «окружают» середину.

Если в данных есть выбросы — значения, которые гораздо выше или ниже остальных, — это может негативно повлиять на среднее значение. Таким образом, среднее значение не робастно, а медиана — напротив, выбросоустойчива.

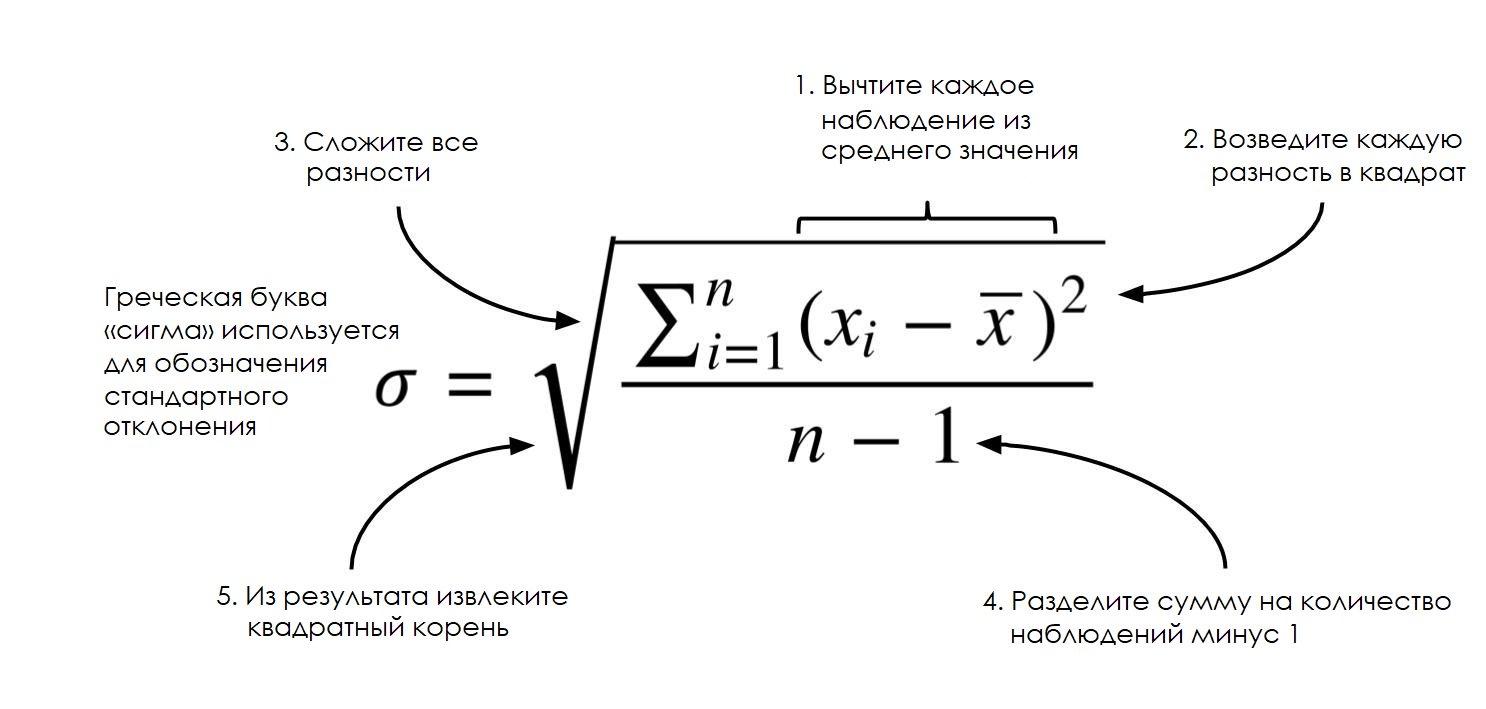
**Меры разброса данных**

**Размах**

Наша первая мера разброса — размах. Из всех измерений, которые мы рассмотрим далее, его вычислить проще всего. Для этого нужно просто вычесть из наибольшего значения в наборе данных наименьшее.

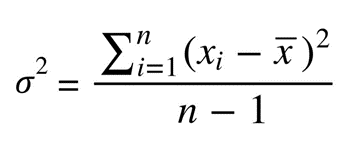
**Стандартное отклонение**

Стандартное отклонение тоже является мерой разброса данных. Оно помогает узнать, как сильно данные отличаются от типичного значения. Иными словами, оно говорит о том, как сильно данные отличаются от среднего арифметического. Отношение к среднему арифметическому хорошо видно при расчёте отклонения:



**Дисперсия**

Часто стандартное отклонение и дисперсию связывают вместе и делают это не без причины. Вот уравнение дисперсии, ничего не напоминает?



Дисперсия и стандартное отклонение — почти одно и то же! Дисперсия — просто квадрат стандартного отклонения. Более того, обе величины отражают одну и ту же вещь — меру разброса, хотя стоит отметить, что единицы измерения разные. В каких бы единицах ни измерялись ваши данные, единицы измерения отклонения будут такими же, а у дисперсии они будут возведены в квадрат.

**Выполнение лабораторной работы**

1. Загрузите данные о выживаемости пассажиров Титаника из файла titanic\_train.csv
2. Представьте данные в виде таблицы. Посмотрите на первые и последние 5 строк.
3. Выведите на экран основную информацию (info) о наборе данных и признаках.
4. Выведите описательные статистики (медиану, среднее, квартили, минимальное, максимальное значения, дисперсию) массива данных в виде таблицы (describe).
5. Избавьтесь от пустых значений, заменив их на медианные значения для количественных признаков и на последнее встречающееся значения с помощью метода “ffill” в аргументах функции fillna для категориальных.
6. Выведите на экран сколько мужчин и сколько женщин находилось на борту?
7. Выведите распределение переменной  Pclass по всем классам (социально-экономический статус) и это же распределение, только для мужчин / женщин по отдельности. Сколько было мужчин 2-го класса?
8. Каковы медиана и стандартное отклонение платежей (Fare)? Округлите до 2 десятичных знаков.
9. Правда ли, что люди моложе 30 лет выживали чаще, чем люди старше 60 лет? Каковы доли выживших в обеих группах?
10. Правда ли, что женщины выживали чаще мужчин? Каковы доли выживших в обеих группах?
11. Найдите самое популярное имя среди пассажиров Титаника мужского пола?
12. Сравните графически распределение (гистограммы или ящики с усами) стоимости билетов и возраста у спасенных и у погибших. Средний возраст погибших выше, верно?
13. Как отличается средний возраст мужчин / женщин в зависимости от класса обслуживания? Выберите верные утверждения:

* В среднем мужчины 1-го класса старше 40 лет
* В среднем женщины 1-го класса старше 40 лет
* Мужчины всех классов в среднем старше женщин того же класса
* В среднем люди в 1 классе старше, чем во 2-ом, а те старше представителей 3-го класса

1. Постройте попарные зависимости признаков Age, Fare, Pclass, Sex, SibSp, Parch, Embarked и Survived. (метод scatter\_matrix Pandas или pairplot Seaborn).
2. Постройте гистограммы распределения каждого признака (столбца данных).
3. Как плата за билет (Fare) зависит от класса каюты (Pclass)? Постройте boxplot.
4. Каково соотношение погибших и выживших в зависимости от пола? Отобразите c помощью Seaborn.countplot c аргументом hue.
5. Каково соотношение погибших и выживших в зависимости от класса каюты? Отобразите c помощью Seaborn.countplot c аргументом hue.
6. Как факт выживания зависит от возраста пассажира? Проверьте (графически) предположение, что молодые чаще выживали. Пусть, условно, молодые - младше 30 лет, пожилые – старше 60 лет.
7. Постройте график рассеяния на осях Age и Fare.  Cиним отметьте пассажиров, которые не выжили (Survived = 0) и красным — выживших (Survived = 1).