

27. Използване на големи обеми от данни за решаване на сложни и комплексни проблеми

1) Какво означава „големи обеми от данни“?

[ресурсни файлове урок 27](#)

„Големи обеми от данни“ означава, че имаме **много записи** (редове) и/или **много показатели** (колони). Например:

- списък с 50 000 продажби
- данни за 30 години валежи по месеци
- резултати от анкета с хиляди отговори

Колкото повече данни има, толкова по-трудно е да ги разбираме „на око“. Затова използваме **анализ**, който обобщава данните.

2) Какво е „анализ на данни“?

Анализ на данни е процес, при който:

1. събираме данните
2. проверяваме ги (дали няма грешки)
3. подреждаме ги и ги обработваме
4. изчисляваме показатели (например средно, минимум, максимум)
5. правим заключение (какви показват резултатите)

3) Какво е Big Data?

Big Data (големи данни) е понятие, което се използва, когато данните са:

- **много големи по обем** (огромно количество записи)
- **разнообразни** (текст, числа, снимки, местоположение и т.н.)
- **постъпват бързо** (на живо – например данни от устройства, сайтове, системи)

В училище няма да работим с истинска Big Data платформа, но ще използваме Excel, за да разберем **идеята**: когато данните са много, трябва да ги обобщим и анализираме.

4) За какво се използват големите данни (примери)

4.1) В транспорта

- анализ на трафик, задръствания
- оптимизиране на маршрути

4.2) В енергетиката

- прогнозиране на потребление
- откриване на загуби, аварии

4.3) В бизнеса

- кои продукти се купуват най-много
- кога са най-силните продажби (месеци/дни)

4.4) В здравеопазването

- анализ на заболявания по региони
- прогнози за натоварване на болници

5) Инструменти в Excel за анализ

Excel има два вида подходи:

1. **Функции и формули** (SUM, AVERAGE, COUNT, IF, SUMIF и др.)
2. **Готови инструменти за анализ** (Analysis ToolPak)

Днес фокусът е върху **Analysis ToolPak → Descriptive Statistics** (Описателна статистика).

6) Какво е Analysis ToolPak?

Analysis ToolPak е допълнителен модул (добавка) в Excel, който дава готови инструменти за статистика и анализ.

Примерни инструменти вътре:

- Descriptive Statistics (описателна статистика)
- Histogram (хистограма)
- Regression (регресия) и други.

Най-важното: не пишем формули ръчно за всяко нещо – Excel ни прави таблица с резултати.

7) Активиране на Analysis ToolPak

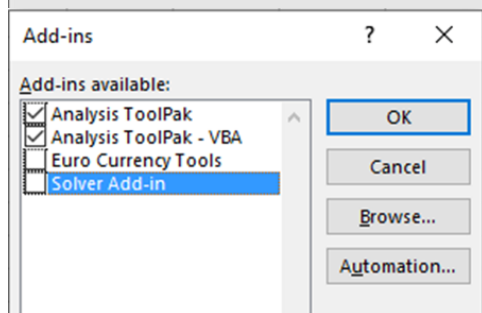
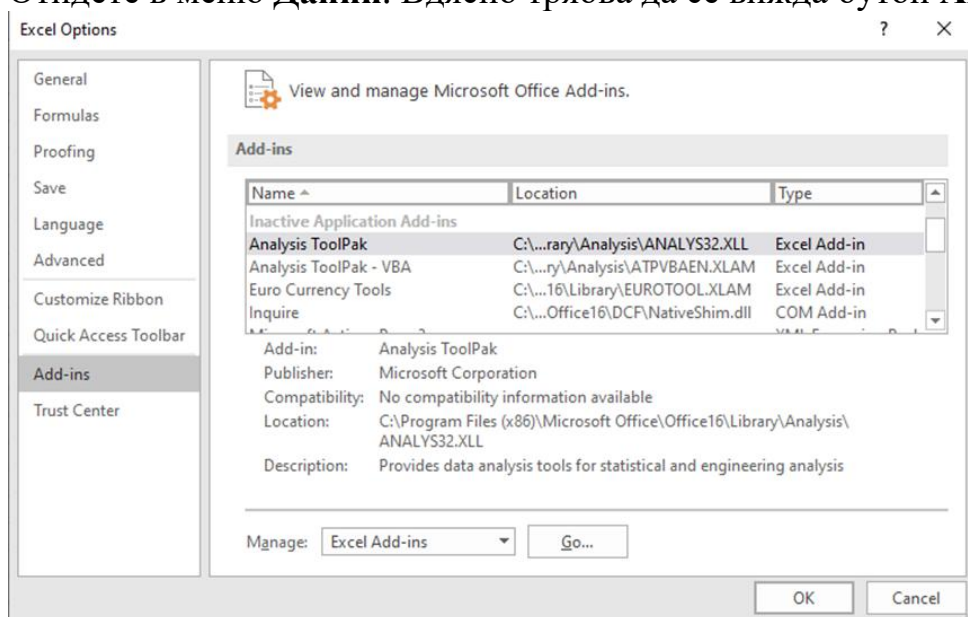
Ако в Excel **не виждате** бутон „Анализ на данни“ (Data Analysis), трябва да включите ToolPak.

Стъпки

1. Отворете Excel.
2. Меню **Файл** → **Опции**
3. Вляво: **Добавки**
4. Долу при „Управление“ изберете: **Добавки на Excel**
5. Натиснете **Отиди...**
6. Сложете отметка на: **Analysis ToolPak**
7. Натиснете **ОК**

Проверка

Отидете в меню **Данни**. Вдясно трябва да се вижда бутон **Анализ на данни**.



8) Какво е Descriptive Statistics?

Descriptive Statistics (Описателна статистика) е инструмент, който прави **обобщение** на избрани данни:

- средно
- минимум, максимум
- стандартно отклонение
- брой стойности и др.

Идеята е да разберем **как изглеждат данните като цяло**, без да гледаме всеки ред.

9) Задача 1: Извършете анализ на данни за средногодишните валежи през последните 30 години.

Отворете файла **valezhi.xlsx** (папка урок 27) и използвайте **Descriptive Statistics**. Създайте таблица с основните тенденции.

9.1) Подготовка на данните

1. Отворете файла **valezhi.xlsx**.
2. Намерете колоната с данните за валежите (например „Валежи“, „Количество“, „mm“).
3. Проверете дали данните са **числа** (не текст).
 - Ако са подравнени вдясно обикновено са числа.
 - Ако са текст (вляво), може да има запетайки/точки или интервали.

Важно: Descriptive Statistics работи правилно само с числови данни.

9.2) Стартиране на анализа

1. Меню **Данни**
2. Натиснете **Анализ на данни**
3. Изберете **Описателна статистика (Descriptive Statistics)**
4. Натиснете **ОК**

9.3) Обяснение на полетата в прозореца

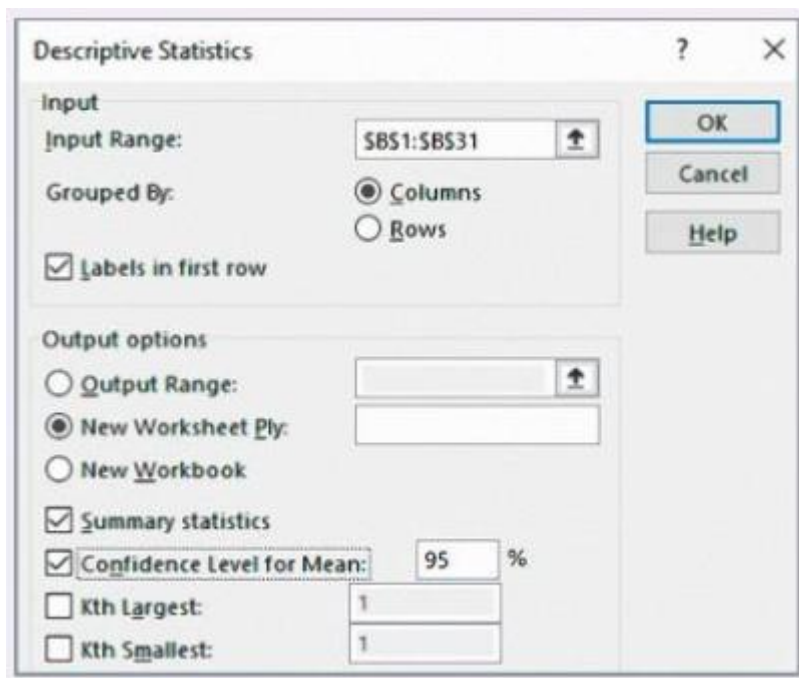
Input Range (Входен диапазон)

Тук посочвате клетките с данните.

Пример: B1:B31 (ако имате 30 години + заглавие)

Grouped By (Групиране по)

- **Columns (Колони)** → ако данните са в една колона (най-често)
- **Rows (Редове)** → ако данните са в един ред



Labels in first row (Етикети в първия ред)

Ако първата клетка е заглавие (например „Валежи“), сложете отметка.

Output Options (Къде да се покажат резултатите)

Имате избор:

- **Output Range** → резултатите да се покажат на избрано място в текущия лист
- **New Worksheet Ply** → резултатите да са в нов лист
- **New Workbook** → в нов файл

Summary statistics (Обобщаващи показатели)

Сложете отметка – това е най-важното.

Confidence Level for Mean (Доверителен интервал за средната стойност)

Това е по-напреднала част.

Ако е включено, Excel добавя статистически интервал (примерно 95%).

Може да оставите по подразбиране (95%), ако учебникът го иска, или да не го включвате.

Kth Largest / Kth Smallest (к-та най-голяма/най-малка)

Ако включите, Excel ще изведе например 3-тата най-голяма стойност. За нашата задача не е задължително.

9.4) Натиснете OK

Excel генерира таблица с резултати.

10) Какво означават резултатите

Ето най-честите показатели:

10.1) Measures of Central Tendency (мерки за централна тенденция)

- **Mean (Средно аритметично)**
„Средната стойност“ – получава се като сбор/брой.
Пример: средни валежи за периода.
- **Median (Медиана)**
Стойността „в средата“, ако подредим данните по големина.
 - добра е когато има много крайни стойности (много високи/много ниски)
- **Mode (Мода)**
Най-често срещаната стойност (ако има повторения).

	A	B
1	Средногодишно количество валежи (мм)	
2	Mean	602,8
3	Standard Error	8,880742436
4	Median	592,5
5	Mode	584
6	Standard Deviation	48,64182959
7	Sample Variance	2366,027586
8	Kurtosis	-0,915205338
9	Skewness	0,214786265
10	Range	173
11	Minimum	521
12	Maximum	694
13	Sum	18084
14	Count	30
15	Confidence Level(95,0%)	18,16315767

10.2) Measures of Variability (мерки за разсейване)

- **Standard Deviation (Стандартно отклонение)**

Показва колко „разпръснати“ са данните около средната стойност.

- малко отклонение → стойностите са близки една до друга
- голямо отклонение → стойностите варират много

- **Variance (Дисперсия)**

Подобно на стандартното отклонение, но в квадратни единици (по-рядко се тълкува директно).

- **Range (Размах)**

Максимум – Минимум

Показва „колко широко“ са разположени данните.

10.3) Minimum / Maximum

- Minimum = най-малка стойност
- Maximum = най-голяма стойност

10.4) Sum / Count

- Sum = сбор
- Count = брой стойности (колко години има)

10.5) Skewness (Асиметрия)

Показва дали данните „наклоняват“ повече към ниски или към високи стойности.

- положителна → има по-големи крайни стойности вдясно
- отрицателна → има по-малки крайни стойности вляво

Важно е да се знае, че **показва изкривяване** на разпределението.

10.6) Kurtosis (Ексцес)

Показва дали разпределението е „по-плоско“ или „по-остро“ спрямо нормално разпределение.

Достатъчно е да се знае, че е показател за „форма“ на разпределението.

11) Изводи:

- Какви са средните валежи за периода (Mean)
- Колко се променят валежите от година до година (Std Dev / Range)
- Кои са най-сухата и най-влажната година (Min/Max)
- Има ли силни отклонения (Skewness)

ОНЛАЙН СОФТУЕР ЗА СТАТИСТИЧЕСКА ОБРАБОТКА НА ДАННИ:

<https://www.mathportal.org/calculators/statistics-calculator/>

<https://www.socscistatistics.com/>