УКРАЇНА

ЦЕНТРАЛЬНА СПІЛКА СПОЖИВЧИХ ТОВАРИСТВ УКООПСПІЛКА

Новомосковський кооперативний коледж економіки та права ім. С.В.Литвиненка
Дніпропетровської облспоживспілки

Розглянуто та схвалено на засіданні циклової комісії
комерційних та економічних дисциплін
Протокол № 1від «31»серпня 2018р.
Голова циклової комісії_____Н.І.Пластун

Спеціальності: 081 Право, 071 Облік і оподаткування, 076 Підприємництво, торгівля та біржова діяльність, 241 Готельно-ресторанна справа Курс І
Дисципліна «ІНФОРМАТИКА»

Лекція 3

Лекція-презентація **Тема: Аналіз та візуалізація даних**

Навчальна мета: ознайомитися з поняттям *вибірка даних*, вивчити статистичні характеристики рядів даних, з'ясувати що таке візуалізація даних, тренди та інфографіка.

Виховна мета: формувати пізнавальний інтерес студентів та показати значення теми для подальшого вивчення дисципліни, підвищувати інформаційну культуру.

Розвивальна мета: спонукати до пізнавальної, творчої діяльності; розвивати самостійність та творче мислення.

Методична мета: використання інтерактивних методів навчання при викладанні дисципліни «Інформатика».

План

1. Вибірка і ряди даних. Деякі статистичні характеристики рядів даних
2. Візуалізація рядів даних
3. Тренди
4. Інфографіка

Технічні засоби навчання:

- інтерактивна дошка, мультимедійний проектор, ноутбук.

Наочність:

Презентація: «Аналіз та візуалізація даних».

Міждисциплінарні зв'язки:

Забезпечувані: Математика

Забезпечуючи: Інформатика та комп'ютерна техніка

Література

Базова

1. *Інформатика 10 клас.: Й.*Я. Ривкінд, Т.І.Лисенко, Л.А.Черникова, В.В. Шакотько – К.: "Генеза ", 2018 с.49-69

1. Вибірка і ряди даних. Деякі статистичні характеристики рядів даних

В багатьох дослідженнях для аналізу даних, установлення певних закономірностей, формулювання висновку, надання рекомендацій, прогнозування тощо потрібно використати багато даних. Методи отримання, опрацювання й аналізу даних, які характеризують масові явища, вивчає наука **статистика**. Так, наприклад, для аналізу тенденцій змінення маси учнів 10-х класів України за останні роки, ризику серцевих захворювань людей певного віку на планеті, популярності продуктів харчування серед населення певного регіону потрібно проаналізувати сотні тисяч або навіть мільйони даних.

Зрозуміло, що провести зважування, вивчити історії хвороб, провести анкетування сотень тисяч або навіть мільйонів людей практично неможливо. Тому для аналізу створюють певну вибірку об'єктів дослідження, тобто з усієї множини об'єктів дослідження відбирають певну кількість і на ній проводять дослідження.

Що більше така вибірка, то точніше буде проведено аналіз і зроблено відповідні висновки. Тобто вибірка повинна бути **масовою**.

Але не тільки кількість даних у вибірці визначає рівень точності аналізу і висновків.

Так, у першому і другому з наведених вище прикладів доцільно вибирати людей різних регіонів і різної статі, а у третьому - людей різного віку. Кажуть, що вибірка даних має бути **репрезентативною** (показовою, характерною, типовою).

Дані, отримані з дослідженої вибірки заносять у таблицю. Така форма подання даних з вибірки зручна для їх аналізу та прогнозів. Дані з кожного рядка і стовпця такої таблиці утворюють **ряди даних**.

Наведемо кілька прикладів вибірок і рядів даних.

Синоптична служба збирає і зберігає дані про температуру, опади, атмосферний тиск вже понад 160 років. Для прогнозування температури та ймовірності опадів у Львові в першій декаді червня наступного року потрібно вибрати відповідні дані, наприклад за останні 10-15 років саме про Львів і саме про першу декаду червня, проаналізувати отримані два ряди даних (про температуру і кількість опадів) і зробити відповідний прогноз погоди.

Команда учнів (4 особи) України бере участь у міжнародних олімпіадах з інформатики починаючи з 1992 року. У таблиці подано результати її виступів з 2005 по2017 рік. Тут вибіркою є вказані в таблиці роки, а рядами даних - загальна кількість медалей у ці роки, а також кількість золотих, срібних і бронзових медалей

Результати виступів команди учнівства України у міжнародних олімпіадах з інформатики

Рік	Кількість медалей	золоті	срібні	бронзові
2005	4	2	1	1
2006	4	1	2	1
2007	4	1	2	1
2008	3	0	1	2
2009	4	1	1	2
2010	3	0	1	2
2011	3	0	1	2
2012	4	1	1	2
2013	4	0	1	3
2014	3	0	1	2
2015	4	0	3	1
2017	3	1	2	0

За цими рядами даних або за деякими з них можна побудувати графіки або діаграми і візуалізувати їх, використовуючи, наприклад, табличний процесор.

Розглянемо деякі статистичні характеристики ряду даних: **середнє арифметичне, стандартне відхилення, мода і медіана.**

Ви знаєте, що середнім арифметичним n чисел називається сума цих чисел, поділена на число n.

	A	B	C
1			
2		Рік	Урожайність т/га
3		2006	1,34
4		2007	1,16
5		2008	1,52
6		2009	1,5
7		2010	1,59
8		2011	1,66
9		2012	1,65
10		2013	2,17
11		2014	1,95
12		2015	2,16
13		Середнє	=СРЗНАЧ(C3:C12)

Так, можна знайти середнє арифметичне врожайності соняшнику в Україні за 2006-2015 роки, використовуючи табличний процесор.

Для обчислення середнього арифметичного в табличному процесорі можна використати відому вам функцію **AVERAGE** або **СРЗНАЧ**. Нагадаємо, що аргументами цієї функції може бути діапазон клітинок, список клітинок, а також їх комбінації, наприклад **СРЗНАЧ(B2:C5; F4;E7).** Приклад обчислення середньої врожайності соняшнику за 2006-2015 роки і формулу для її обчислення **=СРЗНАЧ(C3:C12)** - див.малюнок.

Обчислене в наведеному прикладі середнє арифметичне визначає, яка б була врожайність кожного року (1,6 т/га), якщо вона щороку була б однаковою. Аналогічно **середнє арифметичне будь-якого ряду даних визначає, які б були значення в цьому ряді, якщо б вони були всі однакові.**

Зазначимо, що не для всіх рядів даних середнє арифметичне є показовою характеристикою самого цього ряду.

Наприклад, для ряду даних 2,5; 2,8; 2,3; 2,55; 2,47, у якому дані незначно відрізняються одне від одного, середнє арифметичне дорівнює 2,524, що незначно відрізняється від усіх членів цього ряду, а значить, достатньо показово характеризує весь цей ряд даних. А для ряду 4,7; 6,2; 5,1; 12,4; 14,1, у якому дані значно відрізняються одне від одного, середнє арифметичне дорівнює 8,5, що значно відрізняється від усіх членів цього ряду, а значить, недостатньо показово характеризує весь цей ряд даних.

Для визначення, наскільки показово середнє арифметичне ряду даних характеризує весь ряд даних, можна використати таку характеристику ряду даних, як **стандартне відхилення**. Стандартне відхилення характеризує, наскільки широко розташовані значення ряду даних відносно їх середнього арифметичного.

Стандартне відхилення обчислюється за формулою:

$$S = \sqrt{\frac{(x_1 - x_0)^2 + (x_2 - x_0)^2 + (x_3 - x_0)^2 + \ldots + (x_n - x_0)^2}{n}},$$

де $x_1, x_2 \ldots x_n$ члени ряду даних, а x_0 - середнє арифметичне цього ряду даних.

Для першого з вищенаведених двох прикладів рядів даних стандартне відхилення дорівнює:

$$S = \sqrt{\frac{(2,5-2,524)^2 + (2,8-2,524)^2 + (2,3-2,524)^2 + (2,55-2,524)^2 + (2,47-2,524)^2}{5}} \approx 0,16,$$

а для другого:

$$S = \sqrt{\frac{(4,7-8,5)^2 + (6,2-8,5)^2 + (5,1-8,5)^2 + (12,4-8,5)^2 + (14,1-8,5)^2}{5}} \approx 3,95.$$

D8 =СТАНДОТКЛ

	A	B	C	D
1				
2		x1	2,5	4,7
3		x2	2,8	6,2
4		x3	2,3	5,1
5		x4	2,55	12,4
6		x5	2,47	14,1
7		Середнє	2,524	8,5
8		Стандартне відхилення	0,161567	3,946137

Очевидно, що середнє арифметичне першого ряду даних менше відрізняється від усіх членів ряду даних, а значить, більш показово характеризує весь цей ряд даних. А середнє арифметичне другого ряду даних більше відрізняється від усіх членів ряду даних, а значить, менш показово характеризує весь цей ряд даних.

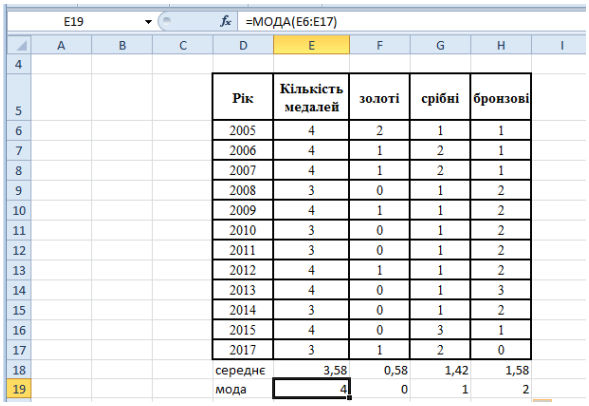
Автоматизувати обчислення **стандартного відхилення** в табличному процесорі можна, використавши функцію **STDEVP** (**СТАНДОТКЛОН**).

Ще однією характеристикою ряду даних є мода.

Мода - це значення в ряді даних, яке повторюється найчастіше. Таке значення є показовим, наприклад, під час дослідження цін на ринку (ціна, яка трапляється найчастіше), під час дослідження попиту взуття, одягу (розміри, які купують найбільше) та ін.

У розглянутому вище прикладі мода кількостей медалей, які вибороло учнівство України на міжнародних олімпіадах з інформатики за 2005-2017 роки, дорівнює 4 (тому що найчастіше в ці роки команда нашої країни завойовувала 4 медалі), мода кількостей золотих медалей - 0, мода кількостей срібних медалей - 1, мода кількостей бронзових медалей - 2.

Якщо в ряді даних два або більше значень повторюються найбільшу кількість разів, то кожне з них вважається модою ряду даних. Так, наприклад, у ряді даних 2, 3, 3, 2, 1 модою є і число 2, і число 3.



	Рік	Кількість медалей	золоті	срібні	бронзові
6	2005	4	2	1	1
7	2006	4	1	2	1
8	2007	4	1	2	1
9	2008	3	0	1	2
10	2009	4	1	1	2
11	2010	3	0	1	2
12	2011	3	0	1	2
13	2012	4	1	1	2
14	2013	4	0	1	3
15	2014	3	0	1	2
16	2015	4	0	3	1
17	2017	3	1	2	0
18	середнє	3,58	0,58	1,42	1,58
19	мода	4	0	1	2

У табличному процесорі є спеціальна функція для обчислення моди ряду даних **МОДА**. Аргументами цієї функції може бути діапазон клітинок, список клітинок, а також їх комбінації, наприклад: **МОДА(B2:C5; F4;E7)**.

На малюнку наведено приклад обчислення моди для кількостей завойованих медалей і формула для її обчислення: **=МОДА(E6:E17)**

Розглянемо ще одну характеристику ряду даних - медіану.

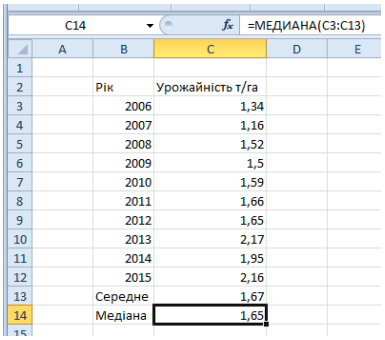
Медіаною впорядкованого ряду даних називається значення, яке поділяє ряд даних на дві рівні частини, тобто зліва і справа від цього значення знаходиться однакова кількість членів упорядкованого ряду даних.

Якщо у впорядкованому ряді даних непарна кількість членів, то медіана такого ряду даних дорівнює значенню його середнього члена, а якщо в такому ряді даних парна кількість членів, то його медіана обчислюється як середнє арифметичне значень двох середніх членів.

Наприклад, для ряду даних 2; 3; 5; 6; 7 медіана дорівнює 5, для ряду даних 2; 3; 5; 6; 7; 9 медіана дорівнює (5 + 6) / 2 = 5,5, а для ряду даних 2; 2; 4; 4; 4; 5; 6 медіана дорівнює 4.

Медіана використовується, наприклад, для визначення місця побудови шкіл, дитячих садочків, магазинів, підприємств побуту тощо. Потрібно визначити ряд відстаней, які мають подолати мешканці певної місцевості до цього закладу, і побудувати його в точці, яка визначається медіаною цього ряду.

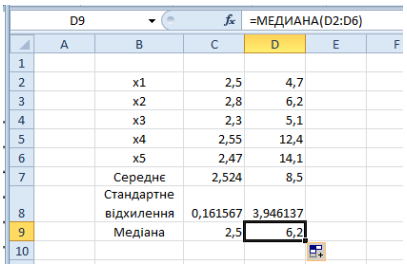
У табличному процесорі є спеціальна функція для обчислення медіани ряду даних: **МЕДІАНА**. Аргументами цієї функції може бути діапазон клітинок, список клітинок, а також їх комбінації, наприклад **МЕДІАНА(B2:C5; F4;E7)**.



На малюнку наведено приклад обчислення медіани ряду даних урожайності соняшнику з використанням табличного процесора за формулою **=МЕДИАНА(C3:C13).**

Звертаємо вашу увагу, що в електронній таблиці для знаходження медіани ряд даних не обов'язково має бути впорядкований. Табличний процесор спочатку впорядковує ряд даних, а потім визначає його медіану.

Зазначимо, що коли члени ряду даних незначно відрізняються одне від одного, то і середнє арифметичне, і медіана більш показово характеризують весь цей ряд.



А якщо члени ряду даних значно відрізняються одне від одного, то медіана більш показово характеризує весь цей ряд даних, ніж середнє арифметичне.

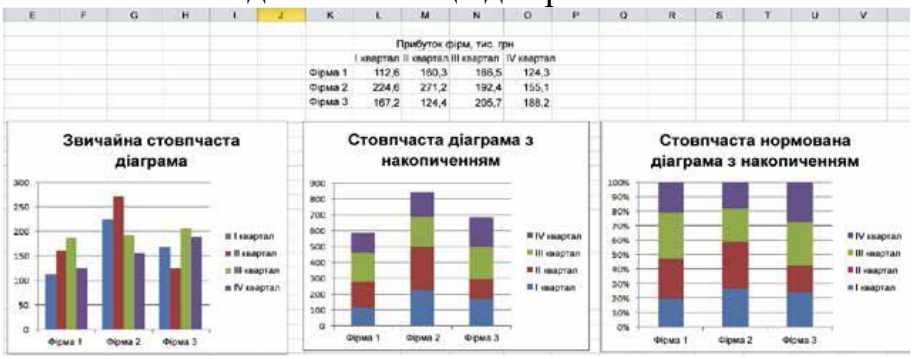
2. Візуалізація рядів даних

Для наочного подання й аналізу рядів даних використовують діаграми. На діаграмах числові дані подаються геометричними фігурами, точками, відрізками, прямокутниками, секторами круга та ін. Розміри цих фігур пропорційні числовим даним, за якими побудовано діаграму. Це дає можливість візуально оцінити співвідношення між числами в одному або в кількох рядах даних.

На уроках інформатики ви будували стовпчасті діаграми, гістограми, секторні й точкові діаграми. Нагадаємо, що стовпчасту діаграму доцільно будувати тоді, коли потрібно порівняти значення кількох рядів даних. Але, крім звичайної стовпчастої діаграми можна побудувати **стовпчасту діаграму з накопиченням** і **нормовану стовпчасту діаграму з накопиченням**.

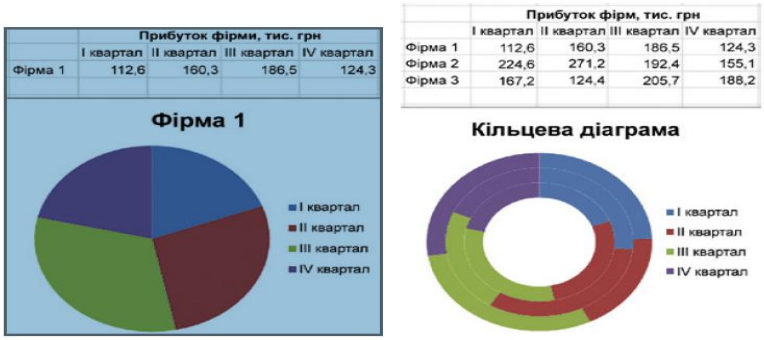
Стовпчаста діаграма з накопиченням відображає частини цілого в усьому цілому (прибутки фірми в кожному із чотирьох кварталів, що в сумі дають прибуток фірми за рік) або кожний із доданків у сумі їх значень для кількох рядів даних.

Нормована стовпчаста діаграма з накопиченням також відображає частини цілого в усьому цілому, але у відсотках. Усе ціле приймається за 100%, визначаються відсотки кожної частини від цих 100%, і всі ці відсотки - частини відображаються частинами одного стовпця діаграми.



Секторні діаграми будуються для одного ряду даних, якщо потрібно відобразити частку кожного окремого даного в загальній сумі. На малюнку наведено секторну діаграму, на якій зображено прибутки однієї фірми за кожний із чотирьох кварталів року.

Подібну діаграму можна побудувати для кількох рядів даних – **кільцеву діаграму**. На малюнку наведено кільцеву діаграму, на якій зображено прибутки трьох фірм за кожний із чотирьох кварталів року. Внутрішнє кільце відповідає першій фірмі в таблиці, зовнішнє - третій. На такій діаграмі зручно візуально порівнювати прибутки різних фірм в одних і тих самих або в різних кварталах.



Аналогічно, якщо побудувати кільцеву діаграму за даними таблиці, у якій наведено прибутки однієї фірми за кожний із чотирьох кварталів кількох років, то буде зручно візуально порівнювати, наприклад, прибутки фірми в одному й тому самому кварталі, але в різні роки.

3. Тренди

Тренд - основна тенденція змінення певного процесу. Ряди даних можна використовувати для прогнозування певного явища, процесу. Марічка займається фітнесом, стежить за своїм харчуванням і спостерігає за зміненням маси свого тіла за півроку.

Марічку цікавить прогноз, як змінюватиметься її маса протягом наступних місяців, якщо вона буде харчуватися як і останні півроку і займатиметься фітнесом з такою самою інтенсивністю.

Цей прогноз можна отримати, побудувавши лінію тренда в Excel.

Лінія тренда - це лінія, уздовж якої розташовуються на діаграмі точки, що зображають дані з певного ряду даних.

Побудуємо лінію тренду за даними таблиці.

Для цього потрібно за даними цієї таблиці побудувати точкову діаграму (**Вставка→Точечная→Точечная с маркерами**) і виконати **Макет→Анализ →Линия тренду→Прогнозируемая прямая с трендом.**

Отримаємо точкову діаграму за даними таблиці і лінію тренду, яка задається лінійною функцією. За цією лінією тренду можна зробити прогноз, що маса Марічки в жовтні (точка на прямій, що відповідає числу 6 на горизонтальній осі) становитиме приблизно 59,7 кг, а в листопаді (точка на прямій, що відповідає числу 8 на горизонтальній осі) - 59,3 кг.

Якщо вибрати діаграму і виконати **Макет→ Анализ →Линия тренду→Дополнительные параметры линии тренда**, то відкриється вікно **Формат линии тренду**, у якому можна:

- установити інший період прогнозування (поля **Вперед на** і **Назад на);**
- задати іншу назву лінії тренду в легенді (перемикач **другое** і поле справа від нього);
- вибрати іншу функцію, яка задаватиме лінію тренду:
  - якщо значення в ряді даних зростають або спадають, то доцільно вибрати **Линейная;**
  - якщо значення ряду даних спочатку зростають, а потім спадають або навпаки, то доцільно вибрати **Полиноминальная** зі степенем 2;
  - якщо значення в ряді даних зростають, потім спадають, потім знову зростають, то доцільно вибрати **Полиноминальная** зі степенем 3;
  - якщо розташування точок на діаграмі відповідає більш складним закономірностям, то можна вибирати іншу функцію із запропонованих.

Точність прогнозу, для якого й будується лінія тренду, залежить від того, наскільки щільно точки розташовані уздовж лінії тренду;

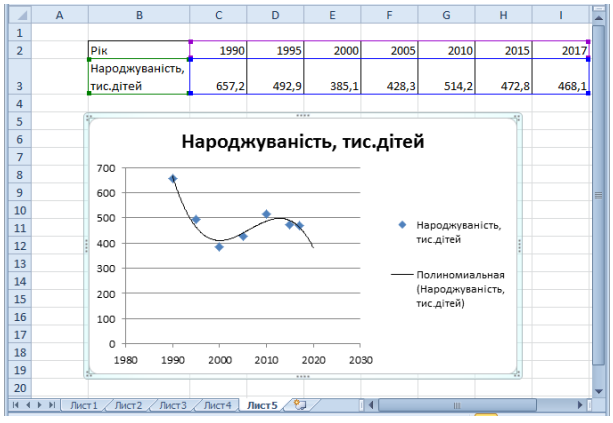
- відформатувати лінію тренду (вкладки **Тип линии, Цвет линии** та ін.).

Спрогнозуємо народжуваність в Україні в наступні роки, використавши дані таблиці народжуваності в Україні за попередні роки, починаючи з 1990 року.

Народжуваність в Україні

Рік	1990	1995	2000	2005	2010	2015	2017
Народжуваність, тис.дітей	657,2	492,9	385,1	428,3	514,2	472,8	468,1

Оскільки, за даними таблиці, за вказані роки народжуваність в Україні спочатку спадала, потім зростала, потім знову спадала, то виконаємо **(Вставка→Точечная→Точечная с маркерами)** і далі **Макет→Анализ→Линия тренду→Дополнительные параметры линии тренда** і у вікні **Формат линии тренду** виберемо **Полиноминальная** зі степенем 3 і встановимо **Прогноз→ Вперед** на 3 періоди.

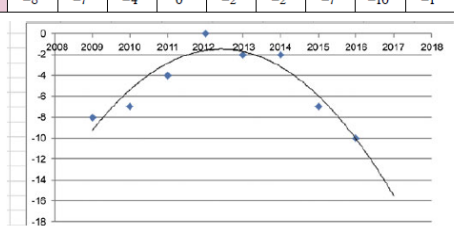


Отримана лінія тренду визначає, що у 2020 році народжуваність в Україні становитиме приблизно 380 тис. дітей.

Зазначимо, що прогнозування з використанням лінії тренду не завжди є правильним. Так, якщо за даними 2009-2016 років, наведеними в таблиці, побудувати лінію тренду і спрогнозувати новорічну температуру в Києві у 2017році, то вона мала б дорівнювати -15^{0}С, хоча реальні дані зовсім інші.

Таблиця температур у Києві 1 січня у 2009–2018 роках

Рік	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
t, °C	−8°	−7°	−4°	0°	−2°	−2°	−7°	−10°	−1°	3°



Цей приклад демонструє, що не до всіх процесів можна застосувати прогнозування з використанням лінії тренду.

4. Інфографіка

Вивчаючи у школі різні предмети, ви досить часто використовували наочність для кращого сприйняття та аналізу відомостей. Це малюнки, графіки, діаграми, схеми, таблиці. Ви також використовували їх у своїх рефератах, комп'ютерних презентаціях.

Їх часто можна побачити на екранах телевізорів, різних сайтах, рекламних щитах (білбордах), під час презентацій нових товарів тощо. Графічне подання відомостей, даних різних видів називають *інформаційною графікою*, або *інфографікою***.**

Психологи твердять, що людина значно краще сприймає відомості, якщо їх подано з використанням графічних зображень, комбінацією графіки, тексту, чисел. Інфографіку широко використовують перш за все для покращення сприйняття великого обсягу відомостей, а також відомостей, що мають досить складну структуру.

Інфографіку часто створюють у графічному редакторі. Нескладну інфографіку можна створити в текстовому процесорі, у редакторі презентацій.

Також існує багато спеціальних онлайн-ресурсів для створення інфографіки. Наприклад,

-Easel.ly (www.easel.ly) - пропонує набір безкоштовних шаблонів для створення інфографіки. Усі структурні елементи майбутньої інфографіки можна редагувати і налаштувати на свій смак. У цьому сервісі є також бібліотека готових форм, стрілок, показчиків і ліній для створення блок-схем, легке налаштування колірних палітр і шрифтів. Також можна додавати зображення з носіїв даних;

- **Infogr.am** (infogram.com) частково безкоштовний ресурс для створення схем, графіків і географічних карт з можливістю завантаження відео та фото для створення інтерактивної інфографіки. Усі дані для майбутньої інфографіки заносяться в таблицю. Їх можна редагувати в будь-який момент, а вбудований генератор автоматично оновить готову інфографіку. Після завершення всіх правок результат можна опублікувати на сайті **Infogram**, вбудувати створену інфографіку у свій сайт або блог, а також поділитися посиланням із друзями, використовуючи соціальні мережі;

- **Venngage.com** (venngage.com) - частково безкоштовний ресурс для створення і публікації інфографіки з досить простим у використанні набором можливостей. Для користувачів доступні готові схеми, теми оформлення, графіки та піктограми, також можна завантажити авторські зображення і тло. Серед додаткових можливостей є можливість створювати анімацію та ін.

У всіх наведених прикладах ресурсів для створення інфографіки інтерфейс англомовний. На жаль, на даний час відсутні ресурси з україномовним інтерфейсом для створення інфографіки.