

Статистические данные: как анализировать?

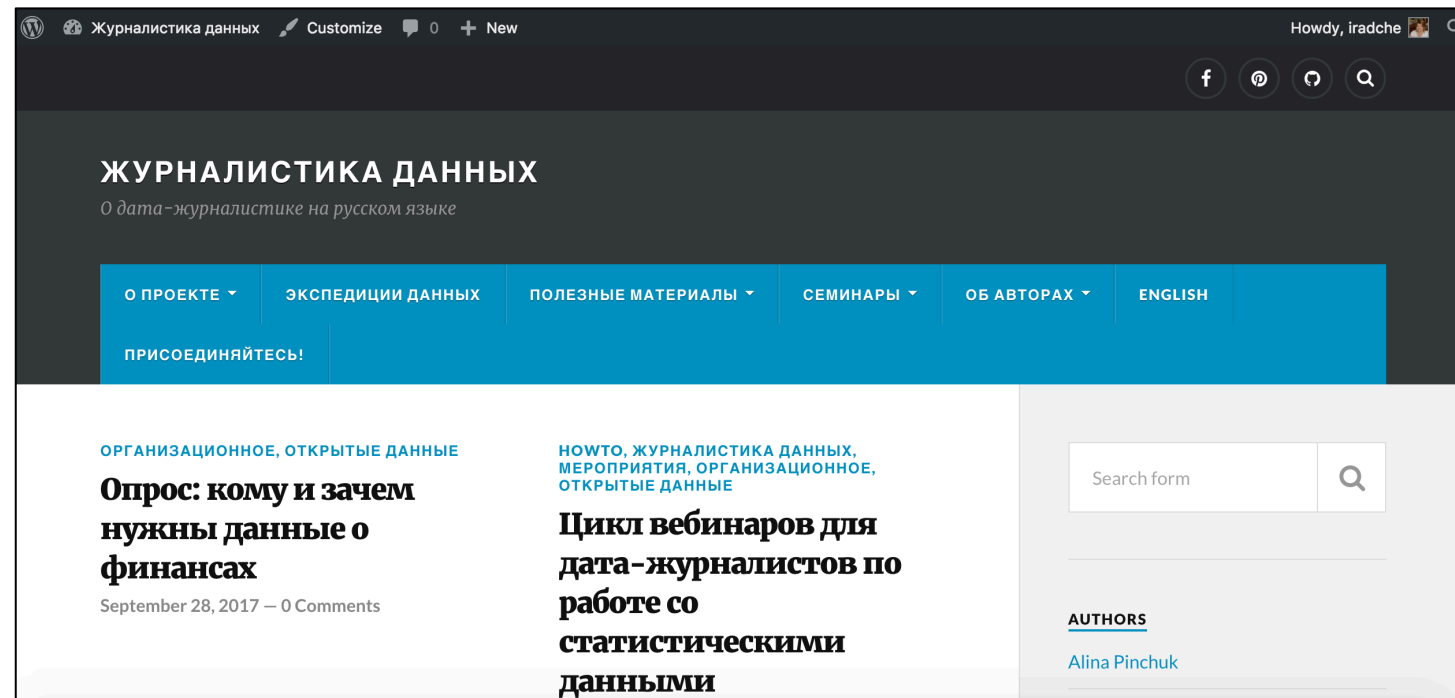
ВЕБИНАР 29 СЕНТЯБРЯ 2017 Г.

ИРИНА РАДЧЕНКО (IRADCHE@GMAIL.COM)

Анализ данных

Поиск неких шаблонов в данных, моделирование и тд.

Примеры и инструкции:



Очень легко сделать математические ошибки.

Пример ошибок в статье:

<http://inosmi.ru/social/20140905/222814116.html>

Вычисление процентов

$$X_{\text{нов}} - X_{\text{стар}}$$

$$X_{\text{стар}}$$

Пример:

ДТП в этом году 60, а в прошлом – 40. Увеличилось количество ДТП на $20/40 = \frac{1}{2}$

ДТП в этом году 40, а в прошлом – 60. Уменьшилось количество ДТП на $-20/60 = -1/3$

Нормировка (для сравнения)

Значение (событие)

$$\frac{\text{Значение (событие)}}{\text{Население}} \times \text{Количество населения}$$

Пример:

60 ДТП		40 ДТП
$\frac{60 \text{ ДТП}}{1\,000} \times 1\,000 = 60 \text{ ДТП на } 1\,000 \text{ людей}$		$\frac{40 \text{ ДТП}}{500} \times 1\,000 = 80 \text{ ДТП на } 1\,000 \text{ людей}$

Теория четырех Россий

«Согласно центрo-периферийной теории, любое заселенное людьми пространство иерархично. Оно делится на центр, полупериферию и периферию. Центр в масштабах страны — крупные и крупнейшие города (Россия-1). Полупериферию, второй иерархический уровень, образуют менее крупные и средние города (Россия-2). Наконец, есть периферия — самая обширная часть пространства, сельские поселения и малые города (Россия-3). Эти три типа пространства, которые соединены на территории страны и присутствуют в каждом регионе, имеют разный социум и разные ресурсы развития. Как следствие, различается их скорость модернизации.

А Россия-4 — это Северный Кавказ. Там модернизационные процессы начались позже, и центрo-периферийная модель пока не очень работает. Но лет через 50 на Северном Кавказе будет, как сегодня в России».

Источник: <https://www.novayagazeta.ru/articles/2013/11/18/57242-171-chetyre-rossii-187-na-odnoy-territorii> (Из интервью с Натальей Васильевной Зубаревич), а также см. https://www.vedomosti.ru/opinion/articles/2011/12/30/chetyre_rossii?

Среднее, медиана, мода и выброс

Медиана – число выборки: ровно половина из элементов выборки больше него, а другая половина меньше него.

Среднее арифметическое – сумма всех чисел, разделенная на их количество.

Мода – значение, которое встречается наиболее часто.

Выброс – результат измерения, выделяющийся из общей выборки.

LibreOffice Vanilla Файл Правка Вид Вставка Формат Лист **Данные** Сервис Окно Справка

Сортировка...
Сортировать по возрастанию
Сортировать по убыванию

Автофильтр
Ещё фильтры

Задать диапазон...
Выбрать диапазон...
Обновить диапазон

Сводная таблица

Содержимое ячейки
Проверка...
Промежуточные итоги...
Форма...

Потоки...
Источник XML...

Совмещённые операции...
Текст по столбцам...
Объединить...
Группа и структура

Статистика

Выборка...
Описательная статистика...
Дисперсионный анализ (ANOVA)...
Корреляция...
Ковариация...
Экспоненциальное сглаживание...
Скользящее среднее...
Регрессия...
t-критерий...
F-критерий...
z-критерий...
Критерий хи-квадрат...

	A	B	C	D	E	F
1	Должность	Зарплата				
2	Ректор	1000000				
3	Бухгалтер	800000				
4	Профессор	60000				
5	Доцент	40000				
6	Доцент	40000				
7	Преподаватель	20000				
8	Преподаватель	20000				
9	Преподаватель	20000				
10	Преподаватель	20000				
11	Преподаватель	20000				
12	Доцент	40000				
13	Доцент	40000				
14	Преподаватель	20000				
15	Преподаватель	20000				
16	Преподаватель	20000				
17	Преподаватель	20000				
18	Профессор	60000				
19	Доцент	40000				
20	Доцент	40000				
21	Преподаватель	20000				
22						
23						
24						

Корреляция

Корреляция (от лат. *correlatio* «соотношение, взаимосвязь») или **корреляционная зависимость** — статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.

Источник: <https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D1%80%D1%80%D0%B5%D0%BB%D1%8F%D1%86%D0%B8%D1%8F>

Общая теория статистики: Учебник / Под ред. Р. А. Шмойловой. — 3-е издание, переработанное. — Москва: Финансы и Статистика, 2002. — 560 с. — [ISBN 5-279-01951-8](#).

Корреляция

Понятие коэффициента корреляции в статистическом анализе является единицей измерения того, насколько хорошо спрогнозированное значение соотносится с реальными данными. Оно дает нам понимание, насколько хорошо прогностика продемонстрировала свою "пригодность" при работе с реальными данными.

Коэффициент корреляции это число между 0 и 1. Если соотношений между спрогнозированными значениями и реальными данными не обнаружено, коэффициент корреляции будет равен 0 или очень близко к этому. Чем выше соотношение между спрогнозированными значениями и реальными данными, тем лучше и коэффициент корреляции. Абсолютно приемлемый результат дает коэффициент 1.0. Таким образом, чем выше коэффициент корреляции, тем лучше.

Источник: http://www.timingsolution.ru/WebHelp/scr/ts_r.htm

Корреляция

Отождествление корреляции и причинности — ошибка, состоящая в убеждении, что наличие корреляции означает причинно-следственную связь.

Список логических ошибок:

https://ru.rationalwiki.org/wiki/%D0%9B%D0%BE%D0%B3%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%BE%D1%88%D0%B8%D0%B1%D0%BA%D0%B0

Линейная регрессия

Линейная регрессия ([англ. Linear regression](#)) — используемая в [статистике регрессионная модель](#) зависимости одной (объясняемой, зависимой) переменной от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) с [линейной функцией](#) зависимости.

Линейная регрессия используется для предсказания зависимых переменных на основе значения независимых переменных, с которыми они связаны.

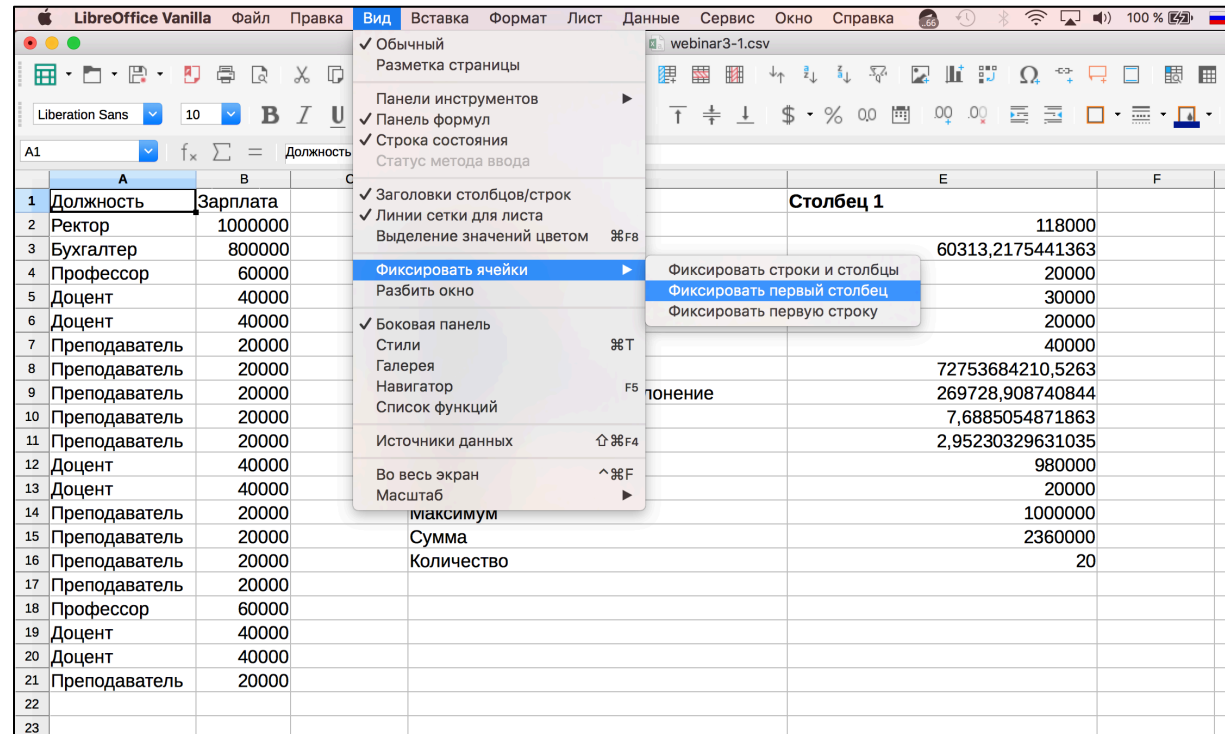
Пример:

Оценки школьников в школах и доходы их семей.

Источник: https://ru.wikipedia.org/wiki/Линейная_регрессия

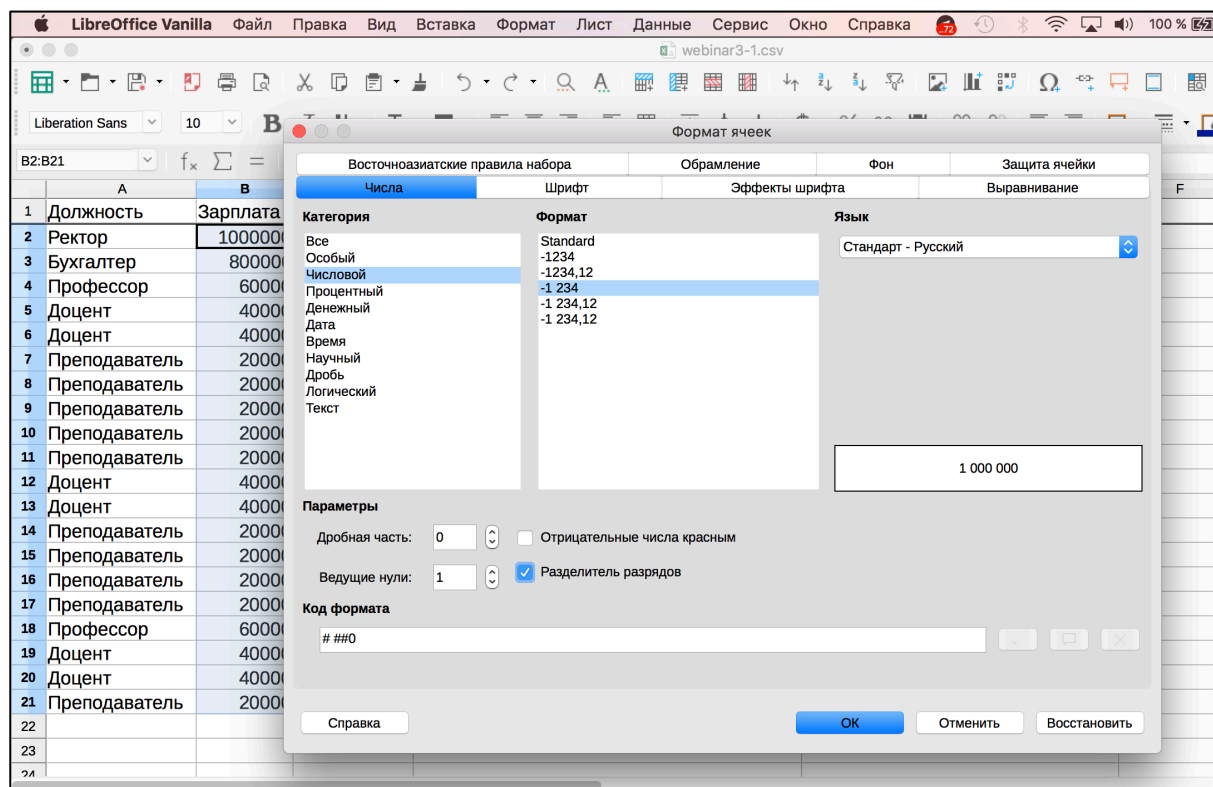
Полезное.

Закрепление строк/столбцов

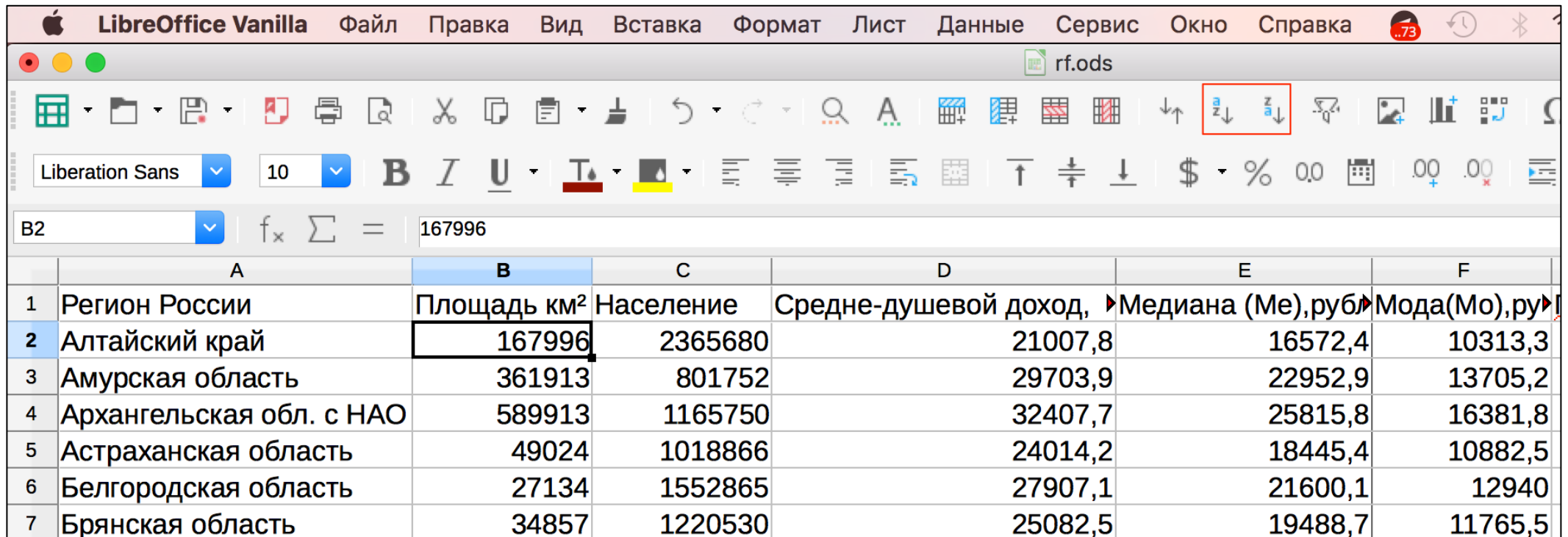


Полезное.

Форматирование ячеек с числами



Полезное. Сортировка



The screenshot shows the LibreOffice Vanilla interface. The menu bar includes: Apple icon, LibreOffice Vanilla, Файл, Правка, Вид, Вставка, Формат, Лист, Данные, Сервис, Окно, Справка. The toolbar contains various icons, with the sort icons (a box with 'a' and a downward arrow, and a box with 'z' and a downward arrow) highlighted with a red rectangle. The spreadsheet below has columns A through F. Column A lists Russian regions, B shows area in km², C shows population, D shows average per capita income, E shows median income, and F shows mode. Row 2 is selected, and the formula bar shows the value 167996 for cell B2.

	A	B	C	D	E	F
1	Регион России	Площадь км²	Население	Средне-душевой доход,	Медиана (Me),руб	Мода(Mo),руб
2	Алтайский край	167996	2365680	21007,8	16572,4	10313,3
3	Амурская область	361913	801752	29703,9	22952,9	13705,2
4	Архангельская обл. с НАО	589913	1165750	32407,7	25815,8	16381,8
5	Астраханская область	49024	1018866	24014,2	18445,4	10882,5
6	Белгородская область	27134	1552865	27907,1	21600,1	12940
7	Брянская область	34857	1220530	25082,5	19488,7	11765,5

Спасибо за внимание!



<http://iRadche.ru>

<http://about.me/Irina.Radchenko>



 [@iRadche](#)

 <http://iRadche.livejournal.com/>

 <https://www.facebook.com/iRadche>

 <http://www.slideshare.net/iRadche>