

# Статистические данные: как анализировать?

---

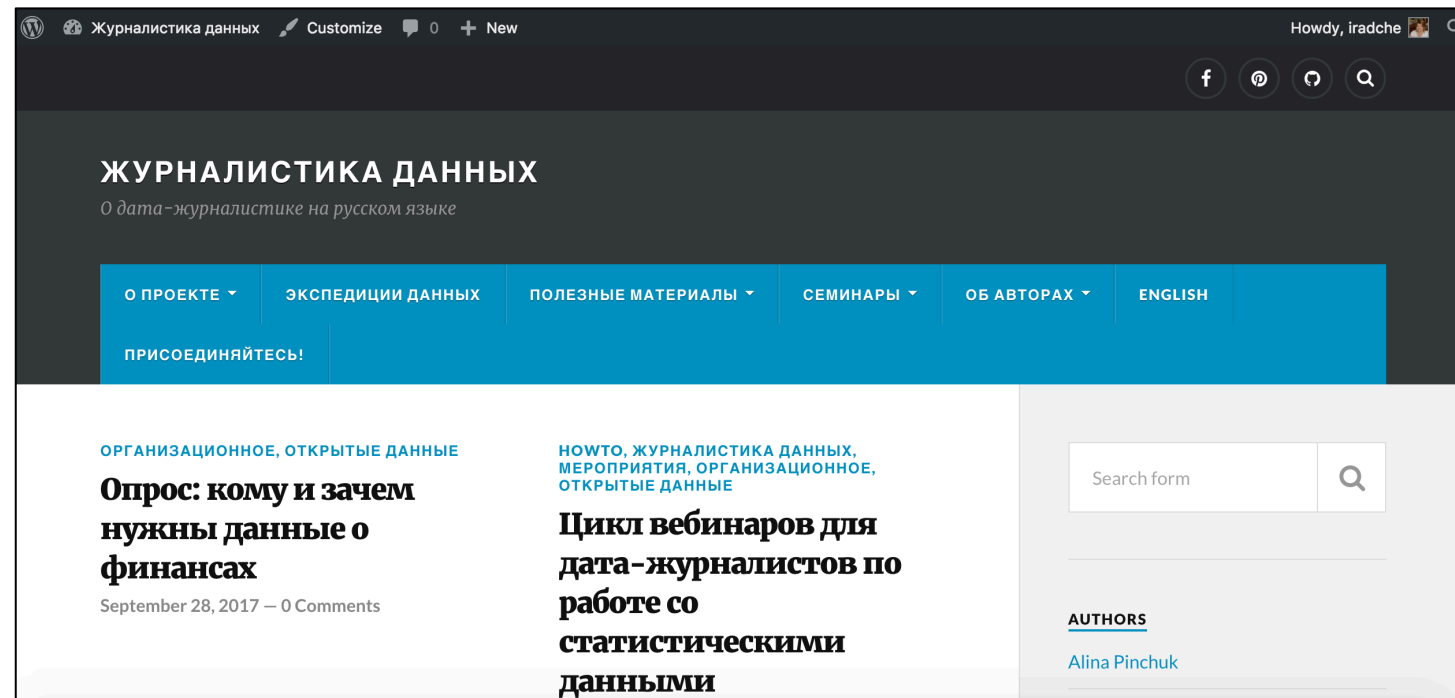
ВЕБИНАР 29 СЕНТЯБРЯ 2017 Г.

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?](https://www.vedomosti.ru/opinion/articles/2011/12/30/chetyre_rossii?)

# Среднее, медиана, мода и выброс

---

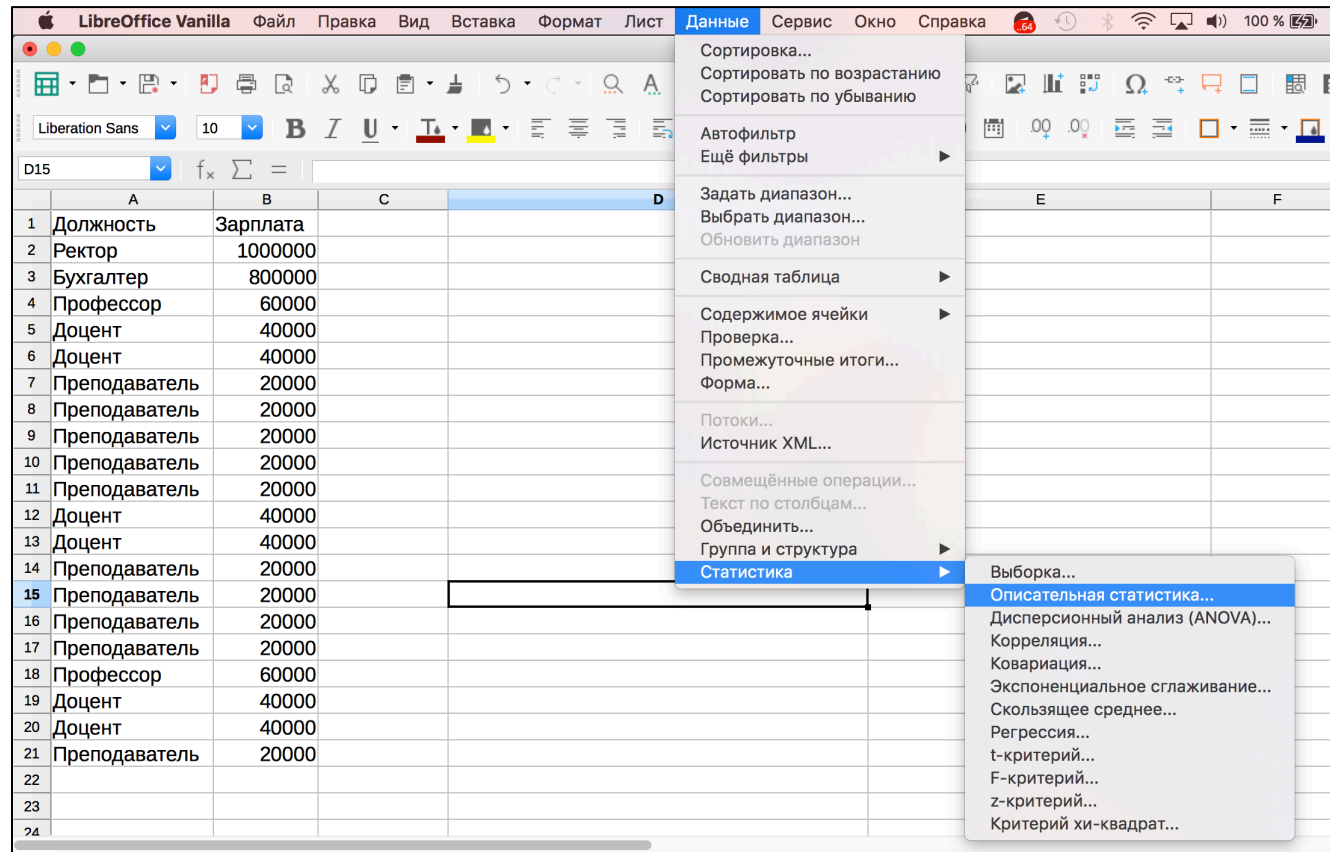
**Медиана** – число выборки: ровно половина из элементов выборки больше него, а другая половина меньше него.

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

**Мода** – значение, которое встречается наиболее часто.

**Выброс** – результат измерения, выделяющийся из общей выборки.

# Автоматическая описательная статистика в LibreOffice





# Автоматическая описательная статистика в LibreOffice

The screenshot shows the LibreOffice Vanilla interface with a spreadsheet titled 'webinar3-1.csv'. The spreadsheet contains data for job positions and salaries. Column D shows the results of automatic descriptive statistics for the salary data in column B. The statistics include the mean, standard deviation, mode, median, quartiles, dispersion, standard deviation, skewness, range, minimum, maximum, sum, and count.

	A	B	C	D	E	F
1	Должность	Зарплата			<b>Столбец 1</b>	
2	Ректор	1000000		Среднее	118000	
3	Бухгалтер	800000		Стандартная ошибка	60313,2175441363	
4	Профессор	60000		Мода	20000	
5	Доцент	40000		Медиана	30000	
6	Доцент	40000		Первый квартиль	20000	
7	Преподаватель	20000		Третий квартиль	40000	
8	Преподаватель	20000		Дисперсия	72753684210,5263	
9	Преподаватель	20000		Среднеквадратическое отклонение	269728,908740844	
10	Преподаватель	20000		Эксцесс	7,6885054871863	
11	Преподаватель	20000		Асимметрия	2,95230329631035	
12	Доцент	40000		Диапазон	980000	
13	Доцент	40000		Минимум	20000	
14	Преподаватель	20000		Максимум	1000000	
15	Преподаватель	20000		Сумма	2360000	
16	Преподаватель	20000		Количество	20	
17	Преподаватель	20000				
18	Профессор	60000				
19	Доцент	40000				
20	Доцент	40000				
21	Преподаватель	20000				
22						

# 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](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](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)

# Вычисление корреляции в LibreOffice

The screenshot shows the LibreOffice Calc interface. The 'Данные' (Data) menu is open, and the 'Статистика' (Statistics) submenu is also open, with 'Корреляция...' (Correlation...) selected. The main spreadsheet displays data for various Russian regions, including area and population.

	A	B	C	D
1	Регион России	Площадь км²	Население	Средне-дух
2	Алтайский край	167996	2365680	
3	Амурская область	361913	801752	
4	Архангельская обл. с НАО	589913	1165750	
5	Астраханская область	49024	1018866	
6	Белгородская область	27134	1552865	
7	Брянская область	34857	1220530	
8	Владимирская область	29084	1389599	
9	Волгоградская область	112877	2535202	
10	Вологодская область	144527	1183860	
11	Воронежская область	52216	2335408	
12	Город Москва	2511	12380664	
13	Город Санкт-Петербург	1439	5281579	
14	Еврейская авт. область	36266	164217	
15	Забайкальский край	431892	1078983	
16	Ивановская область	21437	1023170	
17	Иркутская область	774846	2408901	
18	Кабардино-Балкарская Реп	12470	864454	
19	Калининградская область	15125	986261	
20	Калужская область	29777	1014570	
21	Камчатский край	464275	314729	
22	Карачаево-Черкесская Реп	14277	466432	
23	Кемеровская область	95725	2708844	
24	Кировская область	120274	1201694	

# Вычисление корреляции в LibreOffice

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

webinar3-2.csv

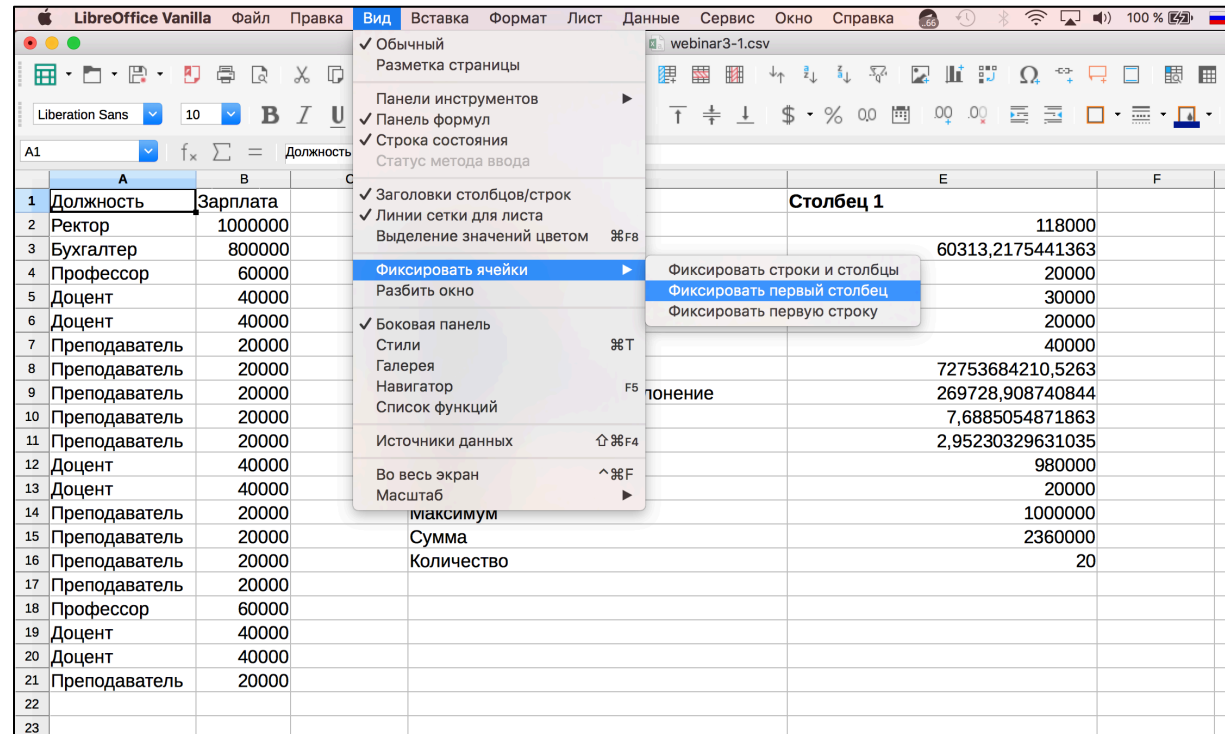
Liberation Sans 10 B I U T A

C89 =CORREL(\$B\$2:\$B\$85; \$C\$2:\$C\$85)

	A	B	C	D	E	F
72	Томская область	314391	1078891	23786,9	19103,7	12321
73	Тульская область	25679	1499417	26425,7	20946,7	13161
74	Тюменская область без АО	160122	1477903	29130,5	22112,3	127
75	Тюменская область с АО	1464173	3660030	41116,5	29859,1	157
76	Удмуртская Республика	42061	1516826	24931,1	19697,2	122
77	Ульяновская область	37181	1252887	22770,4	18060,3	11361
78	Хабаровский край	787633	1333294	37677,3	29388,9	17880
79	Ханты-Мансийский АО	534801	1646078	44519,7	33395,5	18791
80	Челябинская область	88529	1646078	24464	19474,8	12341
81	Чеченская Республика	15647	3502323	22457,3	17065,8	9855
82	Чувашская Республика	18343	1414865	18507,8	14999	9850
83	Чукотский АО	721481	49822	57333,4	43322,9	24736
84	Ямало-Ненецкий АО	769250	536049	63035,9	46687,2	25610
85	Ярославская область	36177	1271912	27006,4	20950,8	12608
86						
87		Корреляции	Столбец 1	Столбец 2		
88		Столбец 1	1			
89		Столбец 2	-0,028250572	1		
90						
91						
92						
93						

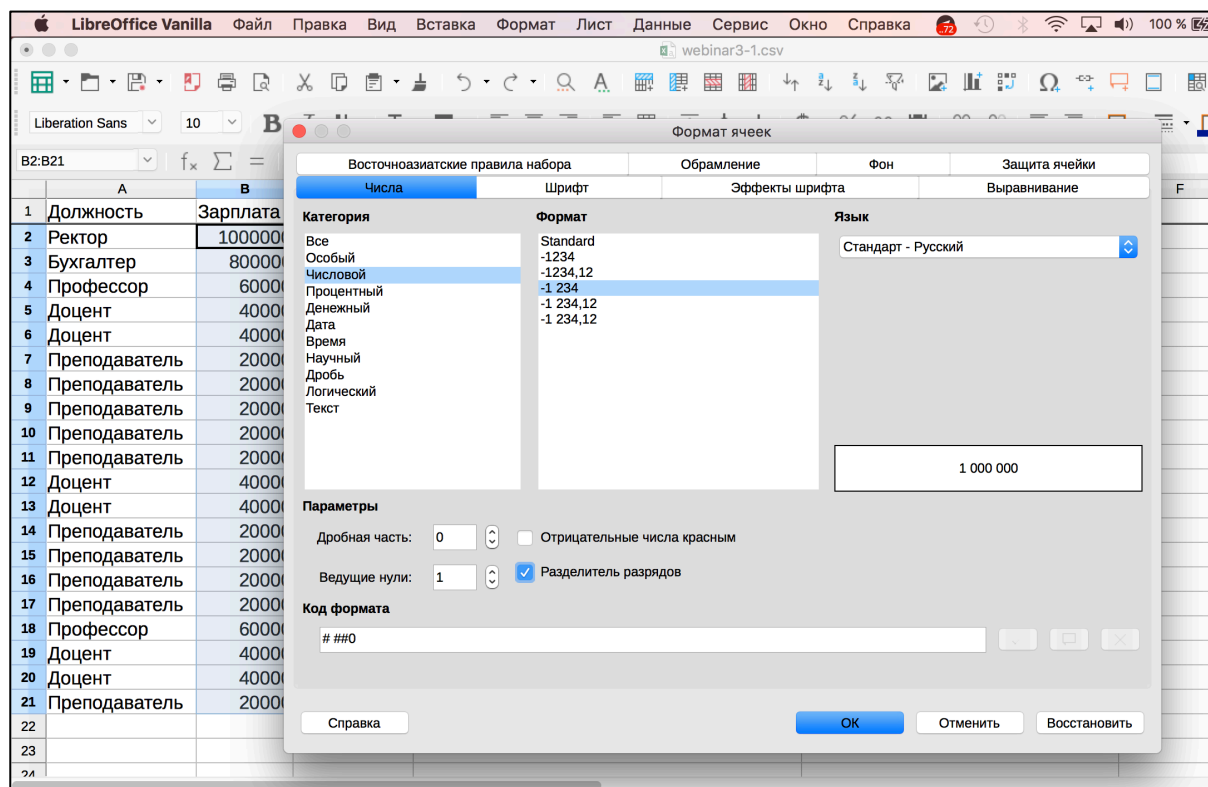
# Полезное.

## Закрепление строк/столбцов



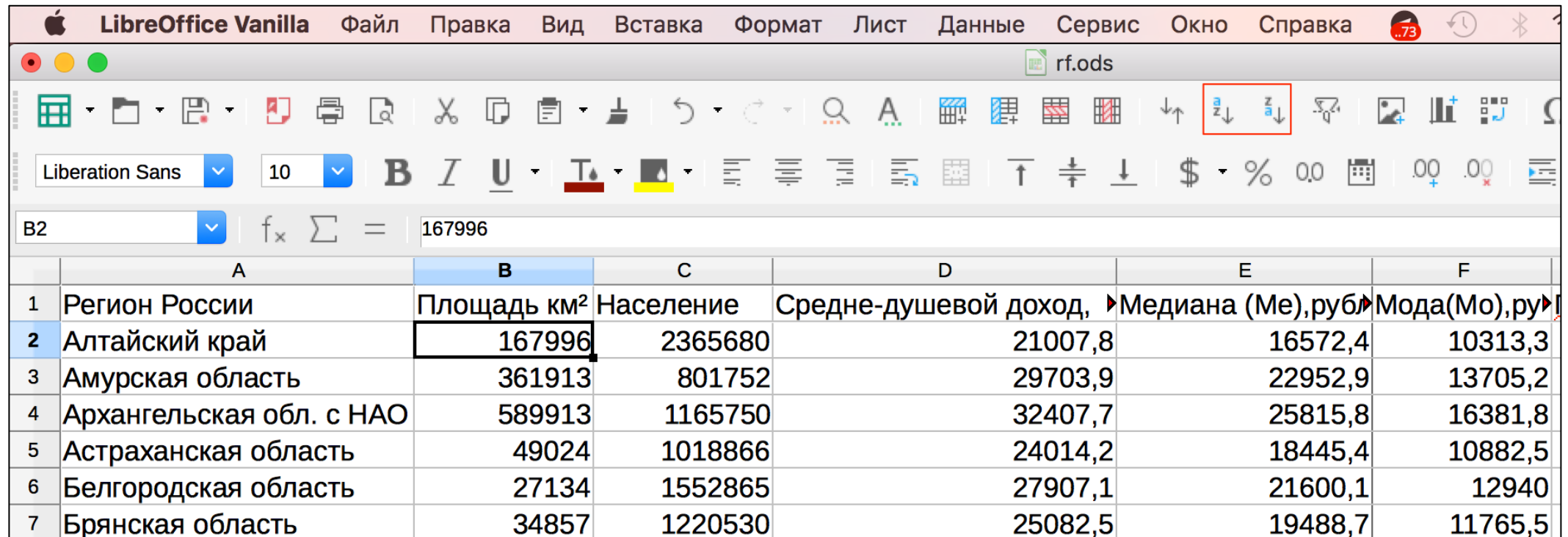
# Полезное.

## Форматирование ячеек с числами





# Полезное. Сортировка



The screenshot shows the LibreOffice Vanilla interface. The menu bar includes: Файл, Правка, Вид, Вставка, Формат, Лист, Данные, Сервис, Окно, Справка. The toolbar contains various icons, with the sort icons (A-Z and Z-A) highlighted by a red box. The spreadsheet shows a table with 7 rows and 6 columns. The first row is the header, and the subsequent rows contain data for different regions of Russia. The cell B2 is selected, and its value, 167996, is displayed in the formula bar.

	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>