Теория вероятностей и статистика в Машинном Обучении

Описательная статистика, визуализация, предварительная обработка данных

Данные — это совокупность сведений, зафиксированных на определенных носителях в форме, пригодной для постоянного хранения, передачи и обработки.

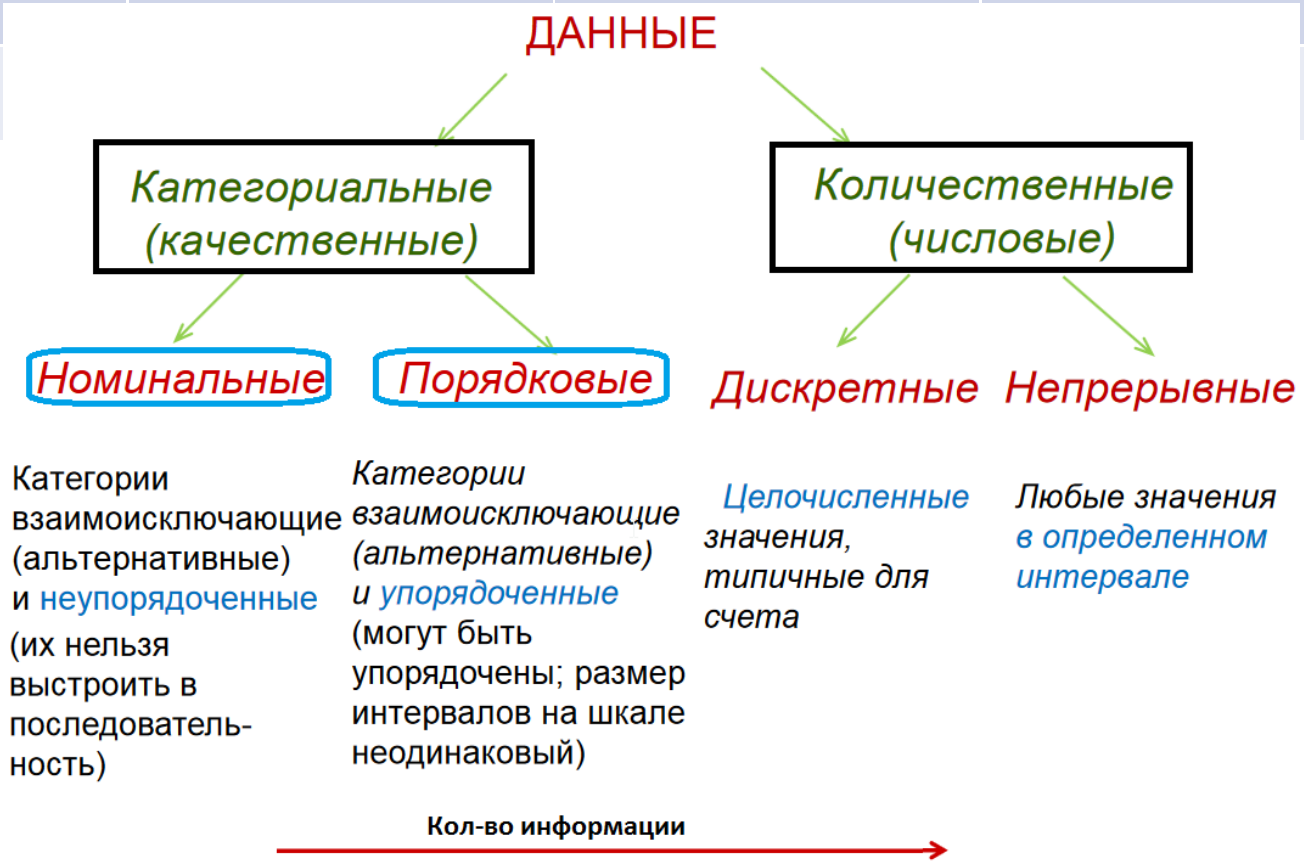
Статистические данные - данные, полученные в результате обследования большого числа объектов или явлений

Математическая статистика имеет дело с массовыми явлениями

Описательная статистика - методы описания статистических данных, представления их в форме таблиц, распределений и т.п.

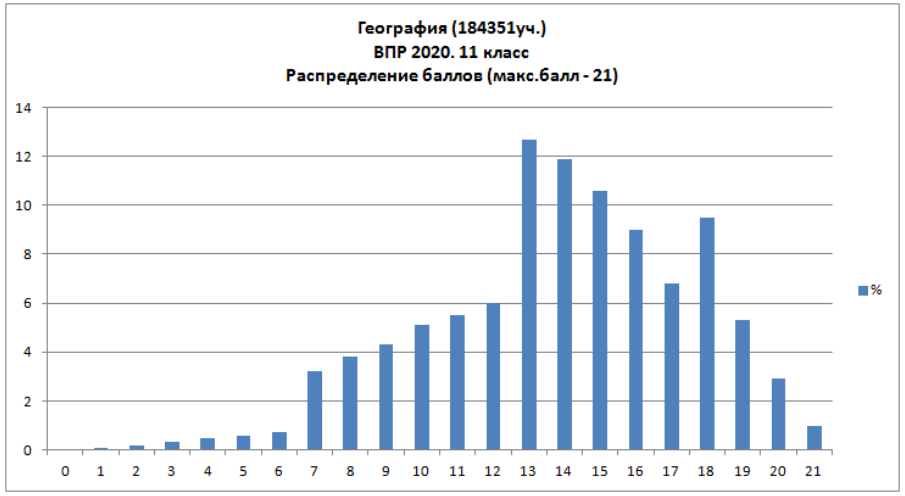
Структурированные данные
Типы данных

	Признак 1	Признак 2	Признак 2
Объект 1			
Объект 2			

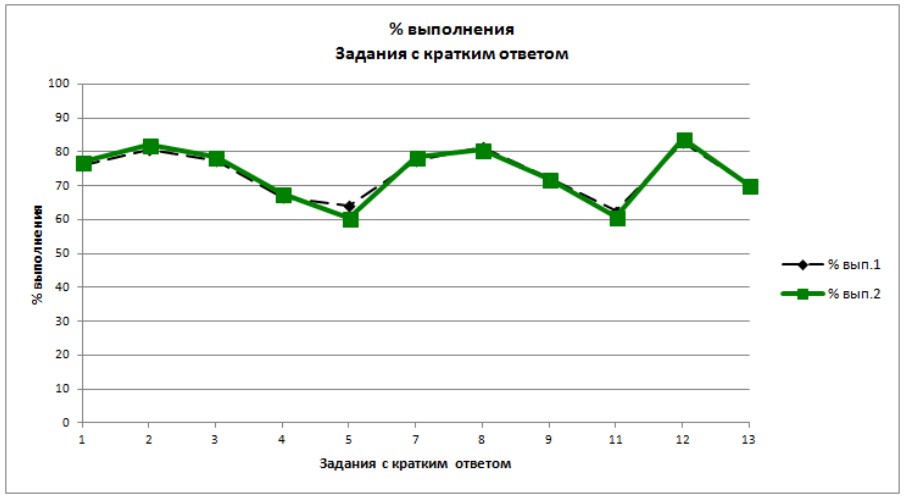


Визуализация данных

Гистограммы



Графики



Круговые диаграммы

Представления обучающихся 8 классов о стремлении учителей и администрации школы к тому, чтобы ситуации буллинга случались в школе реже

Связь переменных

Соотношение особенностей реакции на сообщения в мессенджерах и регулярности собственных публикаций в социальных сетях

Условное форматирование

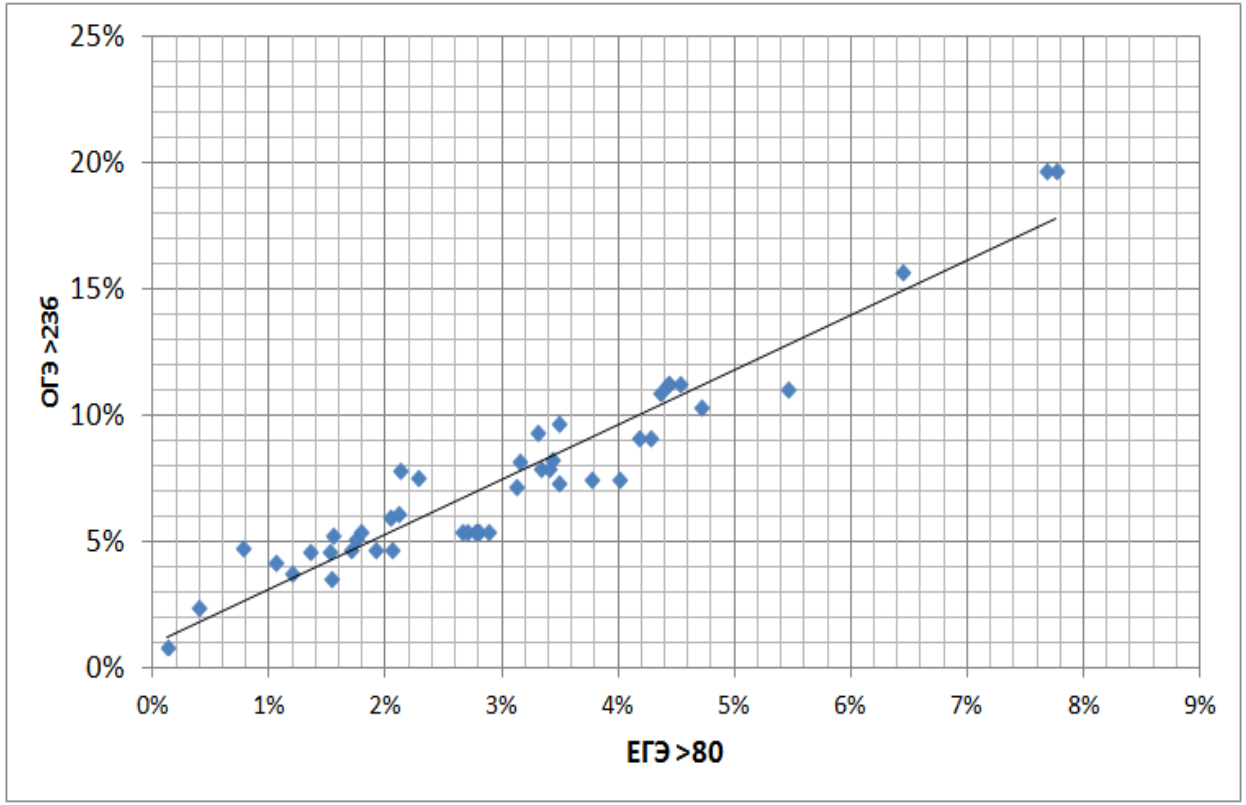
Таблица 16. Процент выполнения оцениваемых заданий в зависимости от предпочтений при выборе будущей профессии

Профессиональная сфера	Процент ответов		Процент выполнения оцениваемых заданий	
	6 класс	8 класс	6 класс	8 класс
медицина	18%	20%	52%	53%
полиция/прокуратура/суд	18%	20%	51%	52%
военная служба/служба в МЧС	17%	21%	48%	49%
дизайн, ремесло, народные промыслы	18%	17%	54%	54%
информационные технологии и программирование	13%	15%	54%	55%
государственное управление	9%	11%	51%	54%
СМИ, журналистика	8%	10%	55%	58%
наука/исследования	8%	9%	53%	55%
торговля, сфера услуг, сервис	6%	10%	54%	56%
транспорт	7%	8%	48%	48%
машино- и приборостроение	5%	6%	49%	49%
преподавание в школе/вузе	6%	5%	53%	56%
добыча полезных ископаемых	4%	3%	48%	49%
строительство	4%	4%	49%	49%
производство продуктов питания	3%	3%	50%	50%
производство тканей, пошив одежды	3%	2%	50%	53%
рабочие профессии (столяр, слесарь)	3%	2%	46%	48%

Условное форматирование изменяет вид диапазона ячеек на основе определенных условий — правил форматирования.

Диаграмма рассеивания

Доля учащихся высокобалльников в ЕГЭ и ОГЭ по регионам



Корреляция >0.95

Описательная статистика

1. Среднее: $$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n};$$

2. Мода: *наиболее часто встречающееся значение; мода может быть не единственна (если несколько значений встречаются одинаковое количество раз, а другие значения встречаются меньшее число раз) и может не существовать (если все элементы встречаются одинаковое количество раз);*

3. Медиана: *серединный элемент в ранжированных данных;*

*если в наборе данных **нечетное** число элементов, то в качестве медианы выбирается такой элемент, что половина всех оставшихся элементов не больше него, а другая половина — не меньше,*

*если в наборе данных **четное** число элементов, то в качестве медианы выбирается среднее из таких двух элементов, что половина оставшихся чисел не больше них, а другая половина — не меньше.*

Пример описательной статистики

Баллы за аттестационную работу:

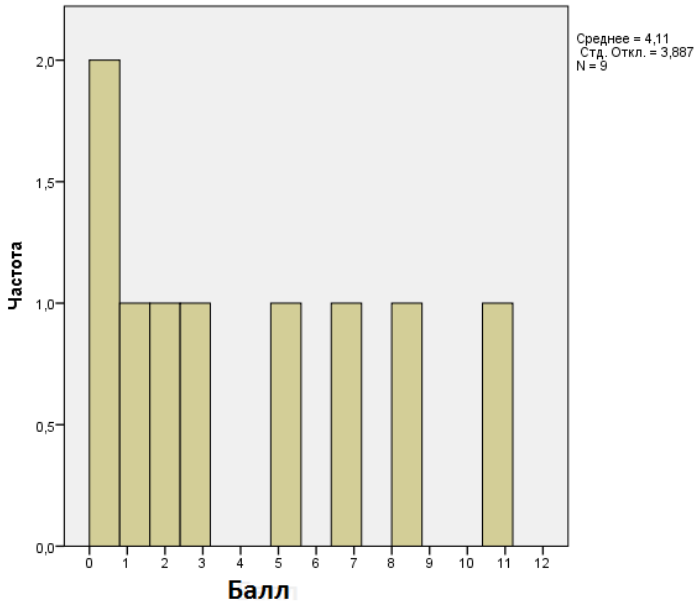
0, 3, 2, 7, 8, 11, 5, 0, 1

Среднее: 4,11 = (0+3+2+7+8+11+5+0+1):9

Ранжированный ряд: 0, 0, 1, 2, **3**, 5, 7, 8, 11

Мода: 0

Медиана: 3



Теория вероятностей и статистика в Машинном Обучении

Лабораторная работа №1

Описательная статистика, визуализация, предварительная обработка данных

Часть 1

1. Найти на сайте РосСтат https://rosstat.gov.ru/free_doc/new_site/zdor22/PublishSite_2022/index.html данные, распределение которых было бы близко к
   a) нормальному (визуально выглядит как кривая Гаусса - «колокол»),
   b) равномерному.

2. Найти распределение со смещенной медианой относительно среднего (~15% размаха) и несмещенной.

3. Посчитать описательные характеристики распределения выбранных данных (среднее, мода, медиана), дать визуальное представление данных (условное форматирование Excel, построить график рассеивания/ гистограмму/ круговую диаграмму).

В Excel2016 описательные статистики есть в блоке «Анализ данных»:
https://support.microsoft.com/ru-ru/office

Примеры Л.р. №1

Таблица 22

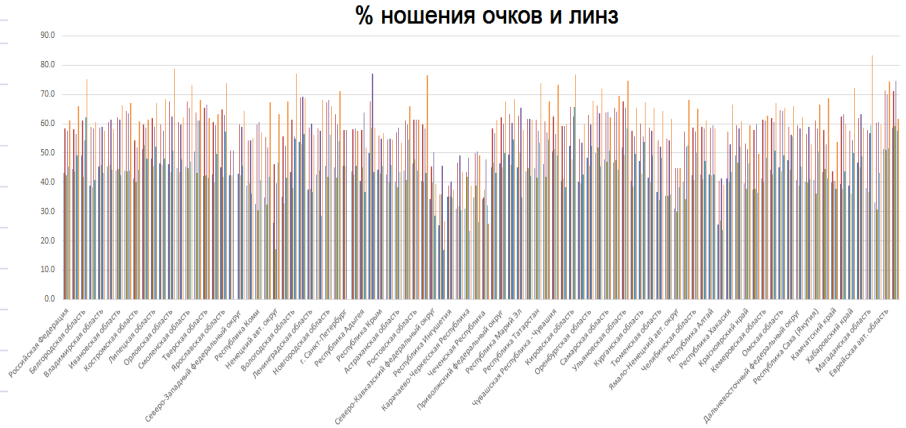
Курите ли вы в настоящее время

(в процентах)

Возраст	Да, ежедневно	Не каждый день (периодически)	Нет, совсем не курю и не курил ранее	Бросил курить	Отказ от ответа
	1	2	3	4	5
Всего	16.4	2.8	68.6	11.6	0.5
15-19 лет	3.2	1.6	93.3	1.2	0.8
из них 15-17 лет	1.3	0.8	96.8	0.3	0.8
20-24 лет	10.7	3.6	80.3	4.7	0.7
25-29 лет	15.4	4.6	70.9	8.6	0.6
30-34 лет	18.7	4.2	66.6	9.9	0.6
35-39 лет	21.8	4.4	61.7	11.3	0.8
40-44 лет	25.1	4.1	58.9	11.1	0.9
45-49 лет	25.5	3.8	58.5	11.5	0.7
50-54 лет	22.1	3.5	61.8	11.9	0.6
55-59 лет	20.6	2.2	65.0	11.9	0.3
60-64 лет	16.2	1.6	67.6	14.2	0.3
65-69 лет	12.4	1.6	70.7	14.8	0.3
70-74 лет	8.3	1.1	74.3	16.1	0.2
75-79 лет	5.0	1.0	77.9	15.8	0.4
80 лет и более	2.0	0.3	86.3	10.8	0.6

нормальное распределение

равномерное распределение



Ежедневно курят			
медиана	среднее знач	размах	отклонение
15.4	13.9	24.2	6%

Не каждый день			
медиана	среднее з	размах	отклонение
2.2	2.6	4.3	7%

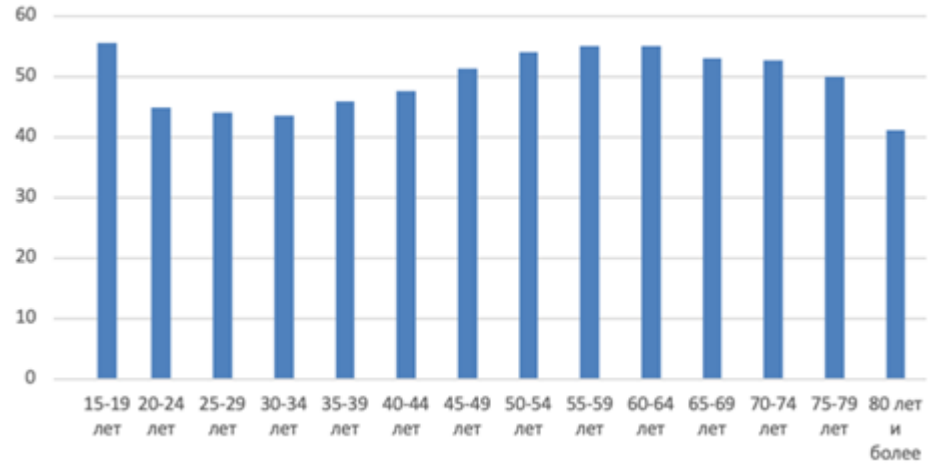
Не курят			
медиана	среднее знач	размах	отклонение
70.7	72.7	38.4	5%

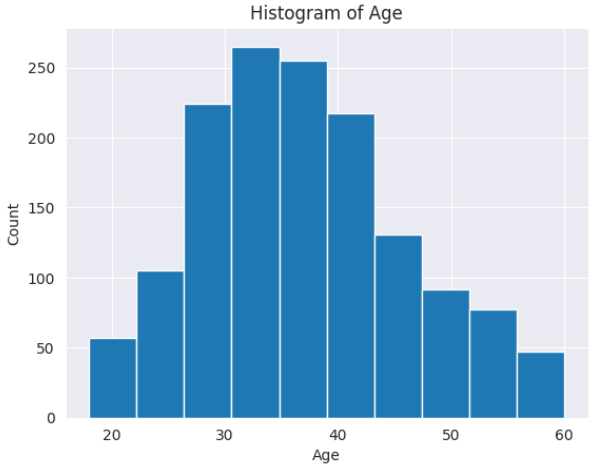
Бросили			
медиана	среднее з	размах	отклонение
11.3	10.3	15.8	7%

Примеры и типичные ошибки при анализе данных (Л.р.1)

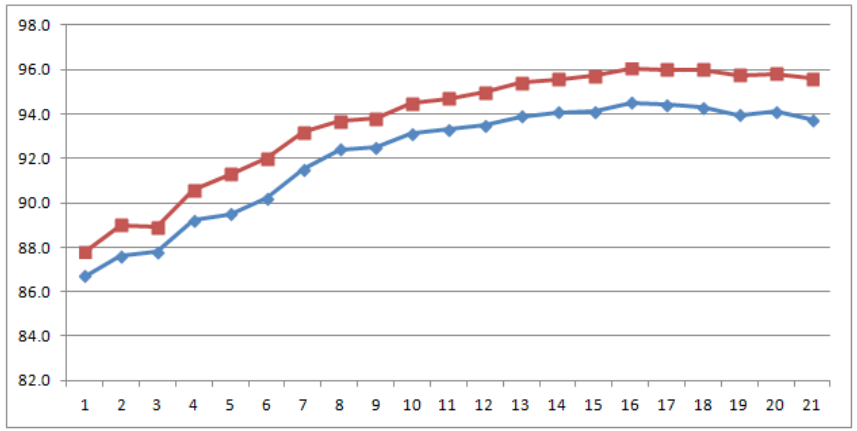
оценка респондентами своего здоровья

Прохождение диспансеризации





Доля лиц среди населения в возрасте 25-64 года, окончивших среднюю школу, % - мужчины-женщины – корр. 0,997



При подсчете среднего берутся среднее по мужчинам и женщинам – д.б. средневзвешенное