Введение Таблицы, кодировки, шкалы

Аббакумов
Вадим Леонардович

- **Анализ данных на Python в примерах и задачах. Часть 1**
- https://www.youtube.com/watch?v=enpPFqcIFj8&list=PLlb7e2G7aSpRb95_Wi7lZ-zA6fOjV3_l7

- **Анализ данных на Python в примерах и задачах. Часть 2**
- https://www.youtube.com/watch?v=5l0e_Q0gpnc&list=PLlb7e2G7aSpT1ntsozWmWJ4kGUsUs141Y

Технические детали-1

- Питон или Пайтон?
- Python или R.

- Anaconda + Python 3
- Jupiter Notebook (входит в дистрибутив)

- Позднее будем добавлять библиотеки

FB Propher, TensorFlow (или PyTorch)…

Технические детали - 2

- это не курс по питону!
- код не оптимальный

- вопросы сразу во время лекции

- Экзамена не будет, оценка по результатам лабораторных работ

Технические детали - 3

- Анализ данных на R в примерах и задачах. Часть 1-2, (видеозапись курса на сайте CSC и на Youtube)

- Курс на R перевели на Python вместе с Владимиром Владимировичем Кукушкиным

Технические детали - 4

- Команда, метод, функция, процедура
- Pandas, Numpy, Matplotlib

Литература

- **Geron** Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 2 edition Есть русский перевод (все за деньги)
- **Hastie, Tibshirani, Friedman** The Elements of Statistical Learning: Data Mining, Inference, and Prediction 2 edition Бесплатно, есть русский перевод (за деньги)
- **Goodfellow, Bengio, Courville** Deep Learning Бесплатно, есть русский перевод (за деньги)

Литература

- Курпатов Четвертая мировая война. Будущее уже рядом (1-я глава)

- Ray Kurzweil (Курцвейл)
- *Google* Технический директор в области машинного обучения и обработки естественного языка

Про секс, бабло и внутреннее самосовершенствование

- ***The Sexiest Job of the 21st Century***
- Нашлось 12 млн результатов…
- Harvard Business Review

- ***highest salary in IT***
- Enterprise Architect ($122,585), Software Engineering Manager ($114,163) and Software Development Manager ($109,809) Cloud Engineer ($98,626), **Data Scientist ($97,027)** and Analytics Manager ($95,238)
- **https://www.forbes.com/sites/louiscolumbus/2019/09/18/glassdoors-highest-paying-tech-jobs-companies-of-2019/#2f9ca9062a3f**

Внутреннее самосовершенствование

- Анализ данных облагораживает человека, ибо полон разочарований (С Тартаковер, цитата неточная)

Анализ данных. Примеры - 1

- скидки на Каско для Мицубиши Лансер

- Марк Цукерберг (Facebook)
- Наказанием за плохое поведение будет блокировка показа рекламы

- видео, на котором распознаем носит ли рабочий каску
- любое анормальное поведение

Анализ данных. Примеры - 2

- Cambridge Analytica
- 300 лайков, и мы предскажем Ваше поведение лучше Вашей мамы

- теперь компания Emerdata
- Erik Prince, who is best known for founding private military group Blackwater USA.

Анализ данных. Примеры - 3

- Джефф Бе́зос
- Амазон
- 200 рекомендаций - 30 рекомендаций
- Генерирует 30+% продаж

- продажи старых книг иногда даже превосходят продажи новых
- книжки перестали утилизовать

Своя игра (Jeopardy!)
IBM Watson

- «Самопроизвольно закипает и без внешних причин охлаждается, хорошо взаимодействует с металлами одиннадцатой группы таблицы Менделеева, помогает снять стресс и является эффективным чистящим и моющим средством. Что за создание описано в одном журнале?»

Анализ данных. Примеры - 4

- IBM Watson в медицине
- RTB (Real-Time Bidding)

Газпром-Нефть
Данные – новая нефть

- По видео распознаем, носит ли рабочий каску (в чем выгода?)
- Управление скидками
- Чат бот
- Прогноз продаж

2045

- Курцвейл
- «Технологическая сингулярность» — это состояние техники, когда она перестаёт быть нашим подручным инструментом, а мы лишаемся всякой возможности управлять ею. Она как бы берёт управление нашим и не нашим и вообще всем бытием на себя.
- Программирование умрет как человеческая профессия

- Генеральная совокупность и выборка
- По части о целом

Модель

- Модель - абстрактное представление реальности в какой-либо форме (например, в математической, физической, символической, графической или дескриптивной), предназначенное для представления определённых аспектов этой реальности и позволяющее получить ответы на изучаемые вопросы.

- В нашем вводном курсе моделью всегда будет функция.

Таблица данных

- Анализируемые данные организованы в виде таблицы.
- Объекты и характеристики объектов

Данные - наблюдения

Например,

данные о покупателях: об их возрасте, составе семьи, покупательных возможностях, образовании;

данные о фирмах-конкурентах, характеристиках товаров и т.д.

- Изучаемые объекты будем называть наблюдениями.
- В примерах выше объектами будут покупатели или фирмы-конкуренты.

Данные-переменные

- Характеристики изучаемых объектов будут называться переменными.
- Например, пол опрашиваемых покупателей, их возраст, оценка товара (отличная, удовлетворительная, плохая), сумма, которую готов заплатить клиент за товар и т.д.

Терминология

- В курсе будут синонимами
- объекты = наблюдения = строки таблицы
- Позднее: наблюдение = точка
- характеристики объектов = переменные = столбцы таблицы.

Например,

VIEWTABLE: Abbat.Albuquerque

	PRICE	площадь	возраст	FEATS	NE	CUST
1	2050	2650	13	7	1	1
2	2080	2600	.	4	1	1
3	2150	2664	6	5	1	1
4	2150	2921	3	6	1	1
5	1999	2580	4	4	1	1
6	1900	2580	4	4	1	0
7	1800	2774	2	4	1	0
8	1560	1920	1	5	1	1
9	1450	2150	.	4	1	0
10	1449	1710	1	3	1	1
11	1375	1837	4	5	1	0
12	1270	1880	8	6	1	0
13	1250	2150	15	3	1	0
14	1235	1894	14	5	1	1
15	1170	1928	18	8	1	1
16	1180	1830	.	3	1	0

Если данные не представлены в виде таблицы?

- В курс не рассматриваются такие случаи.
- Такое случается редко.
- Обращаться к специалисту (например, к знающему коллеге или на специализированный форум).
- **Примеры**: медицинская карта пациента, когда покупатель посещал магазин

Ames housing dataset

- ценах домов в Айове
- 2930 домов
- характеристики дома и его цена
- какая площадь дома, какой подъезд к дому, состояние дома, есть ли гараж и так далее
- цель - построение модели для определения цены дома

Терминология

- Экспериментальное значение
- Модельное, теоретическое значение
- Критерий качества измеряет, насколько хорошо соответствуют друг другу

Зачем нужна модель?

- Zillow
- https://en.wikipedia.org/wiki/Zillow
- прибыль в 2019 году была 305 миллионов долларов

- Модель для определения цены дома
- подсчитана цена для 110 миллионов домов в америке,
- Бесплатно выложена на сайте

Кодирование данных

- Кодирование – это сопоставление каждому значению переменной некоторого числа, это число называется кодом.

- Например, при анализе результатов опроса кодируем пол:
- 0 – мужской,
- 1 – женский.
- При анализе продаж на АЗС можно кодировать:
- АИ-**92** –1,
- АИ-**95** – 2,
- АИ-**98** – 3,
- Дизельное топливо – 4,
- ... еще 20000 видов сопутствующих товаров

Например,

- Ответы на вопрос анкеты о количестве членов семьи (включая опрошенного), то допустимыми ответами были «отказ отвечать», «1», «2», «3», «4 или больше», то возможным вариантом кодировки будет -9999 – отказ отвечать,
- 0 – ноль членов семьи, невозможное значение, ошибка анкетера,
- 1 – семья из одного человека, то есть респондент живет один,
- 2 – семья из двух человек,
- 3 – семья из трех человек,
- 4 – семья из четырех или большего числа людей.
- Если в таблице данных встретятся текстовые значения, например «три» вместо числа «3», то аналитик сам себе осложнил задачу.

Пример перекодировки

возраст	Код
Моложе 20 лет	1
От 20 до 29 лет	2
От 30 до 39 лет	3
От 40 до 49 лет	4
От 50 до 59 лет	5
От 60 до 69 лет	6
От 70 до 79 лет	7
80 лет и старше	8
Нет ответа	-9999

Не выбирайте случайную кодировку

- Кодировка может помочь, а может помешать
- Задача определения возраста (80-80)
- «Ввести код года»

Шкалирование

- Встречаются ситуации, когда некоторые числа не могут использоваться, «запрещены».
- Часто бывает, что «запрещены» (так как не имеют смысла) некоторые операции, например сложение!

Иногда числа нельзя складывать

- Как это – запрещено число 3/2!?
- Как это я не могу сложить 1+3? Всегда могу!
- А вот не всегда, причем все об этом знают.

Запрет на числа

- Если в школьной математической задаче получен ответ «Для выполнения работы в течение часа требуется полтора землекопа», то ясно, что с решением что-то не так.
- С другой стороны, ответ «Работа будет выполнена за полтора часа» не вызывает вопросов.
- Получается, что иногда число 1,5 допустимо, разрешено, а иногда запрещено.

При опросе покупателей регистрировалось место жительства респондента

- «Живу в Тихвине» кодировался числом 1,
- «Живу в Выборге» – числом 2,
- «Живу в Петербурге» - числом 3.
- В городах было опрошено одинаковое количество респондентов.
- Чему равно среднее арифметическое ответов на вопрос «В каком городе Вы живете?»

Запрет на сложение

- Среднее арифметическое будет равно двум.
- Как Вы относитесь к заявлению, что в среднем все опрошенные живут в Выборге?

Три вида анализа данных.

Чтобы предупредить возможные ошибки с «запрещенными» числами и операциями, введено понятие шкалы, в которой измерена переменная.

Для данных в разных шкалах разработаны свои методы анализа.

Номинальная шкала

- Номинальная шкала, если значения переменной являются условными именами.
- Обычно эти имена пронумерованы (или заменены хешами), и номера используются в качестве кодов. Значение номера не имеет смысла.
- Иногда говорят «шкала наименований».

Примеры переменных, измеренных в номинальной шкале

- имя, (непрактично)
- фамилия, (непрактично)
- пол,
- национальность,
- цвет,
- город,
- код товара
- и т.д.

Номинальная шкала

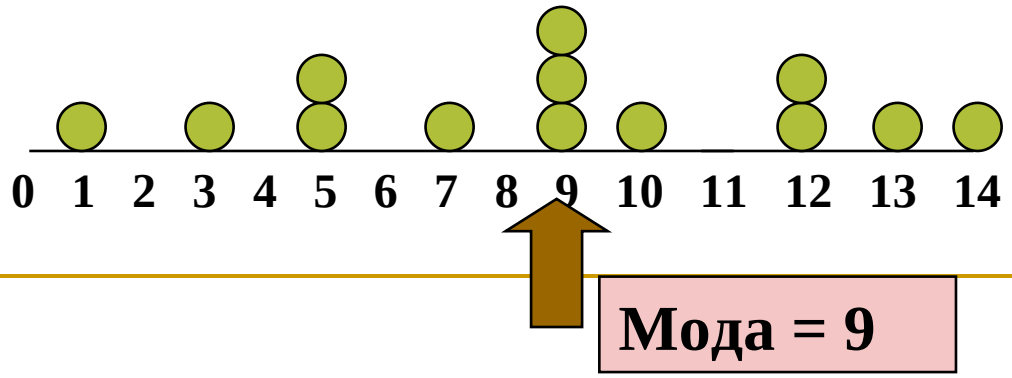
- Очевидно, что сравнивая коды товаров, невозможно ничего выяснить.
- Бессмысленно также искать среднее значение кодов товаров.

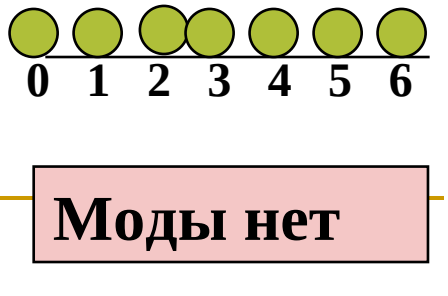
Номинальная шкала

- Единственная доступная операция - это подсчет. Сколько раз переменная принимала значение «а»
- Типичное значение **мода**, и только она

Мода

- Значение, которое встречается чаще всего
- Мод может быть несколько (мультимодальность)
- Мода нестабильна





Номинальная шкала

- Например, рассмотрим выборку из 60 мужчин и 40 женщин, для мужчин выбран код 1, а женщин код 2.

- Неправильно: среднее значение пола равно 1,4.
- Правильно: 60% выборки составляют мужчины.

Порядковая (ранговая) шкала

- Переменная измерена в *порядковой* (ранговой) *шкале*, если значения переменной можно сравнивать между собой, но разность значений не имеет смысла.

Примеры переменных, измеренных в порядковой шкале

- сорт товара (высший сорт, первый сорт, второй сорт);
- ранги предпочтений покупателей (1 – наиболее привлекательный товар, 2 – мало привлекательный, 3 – непривлекательный товар).
- рейтинг, например место компании в списке 100 лучших компаний
- оценка качества услуг в баллах.

Количественная шкала

- *Количественная шкала* позволяет отражать абсолютные количественные характеристики исследуемых объектов.
- Почти всегда это характеристики, измеренные в рублях, метрах, секундах, килограммах.
- Выделяют несколько вариантов количественной шкалы. Нам они не важны.

Количественная шкала: примеры

- Средний чек в универмаге,
- Количество литров бензина, проданных на автозаправке за день,
- заработная плата респондентов,
- систолическое давление пациента.

Задача. Определить шкалу.

- три переменные –
- стартовый номер спортсмена-бегуна,
- место, которое спортсмен занял в результате соревнований
- время, за которое он пробежал дистанцию.

Задача. Укажите шкалу, в которой измерены переменные

- Тестовые баллы:
- Раса:
- Плохо, Удовлетворительно, Хорошо, Отлично:
- Высота:
- Коэффициент интеллекта IQ:
- Страна, в которой родился
- Доход

Задача. Укажите шкалу, в которой измерены переменные

- Образование

О смене шкалы

- Часто от исходных числовых данных, измеренных в количественной шкале, переходят к порядковой шкале. Пример: группировка по возрасту
- От ранговой или количественной шкал можно перейти к номинальной шкале.
- В обратном направлении переход трудный, но часто неизбежный.
- Например переменную «город» заменяем на переменную «средний доход в городе»
- Например One-Hot encoding

Наблюдения заменяем на ранги.

- Ранг – номер наблюдения в упорядоченном ряду наблюдений
- Выборка состоит из чисел 7, 5, 12, 2, 8, 16.
- Упорядочение проводим от меньшего к большему.
- Тогда первый ранг будет иметь число 2 (самое маленькое), второй ранг - 5, третий - 7, четвертый - 8, пятый ранг – 16 (самое большое число).

Типичное значение