Анализ данных в R

23 сентября 2020 г.

Pipeline анализа данных

Получение и импорт данных

Первичный анализ и визуализация

Очистка данных

Трансформация данных

Построение модели

Импорт

Откуда приходят данные

Какими бывают данные

- Cross-sectional — матрица объекты-признаки за фиксированный период времени
- Time series — данные о каком-то объкте в разные периоды времени
- Panel data — матрица объекты-признаки в разные периоды времени

Какими бывают данные

Имя	Возраст	Пол
Андрей	22	М
Вика	19	Ж
Кирилл	34	М
Александр	24	М
Мария	26	Ж

Компания	Год	Выручка	Число работников
Газпром	2006	15,6	1500
Газпром	2007	18,1	2500
Газпром	2008	20,3	3001
Сбербанк	2006	8	1100
Сбербанк	2007	9,1	1200
Сбербанк	2008	10,4	2300
ВТБ	2006	7	600
ВТБ	2007	6,5	800
ВТБ	2008	8	1200

Год	1996	1997	1998	1999	2000	2001	2002	2003
ВВП	195,4	201,2	180,1	242,2	250,5	301,12	300,28	302,9

Немного про обозначения

	mpg	cyl	disp	hp	drat	wt	qsec	vs
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0

- Вся таблица — матрица (или так называемый DataFrame)
- Один столбец в DataFrame — массив, имеющий данные одного типа

Типы данных

- Числовые (int/dbl) — целые или вещественные числа
- Строковые — текст
- Бинарные — состоящие из двух категорий (1/0; Да/Нет; TRUE/FALSE)
- Категориальные — состоящие из нескольких категорий (1/2/3; Москва/Лондон/Париж)
- DateTime — специальный формат, связанный со временем и датой

Первичный анализ

Статистики

Хотим посмотреть на закономерности в данных!

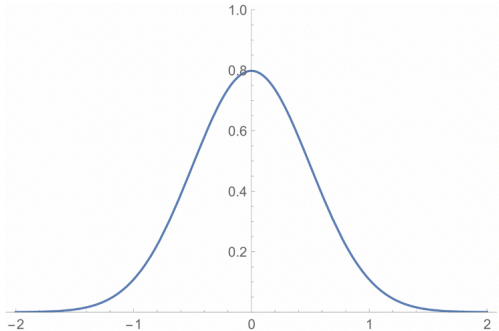
- Среднее — простое среднее арифметическое $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Дисперсия — степень разброса значений около среднего $Var = \sum_{i=1}^{n} (x_i - \bar{x})^2$
- Среднеквадратическое отклонение — корень из дисперсии $\sqrt{Var} = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}$
- Квантиль — значение, которое случайная величина не превышает с фиксированной вероятностью
- Медиана — значение, для которого 50% значений в выборке находится ниже и 50% выше. По сути это 50% квантиль
- Максимум и минимум

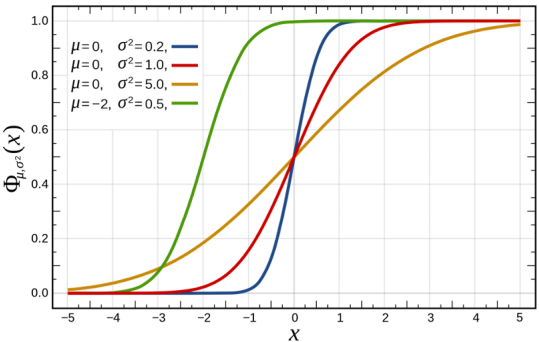
Немного о распределениях

- Функция распределения случайной величины X — это вероятность того, что случайная величина X примет значение, меньшее или равное x, где x — произвольное действительное число

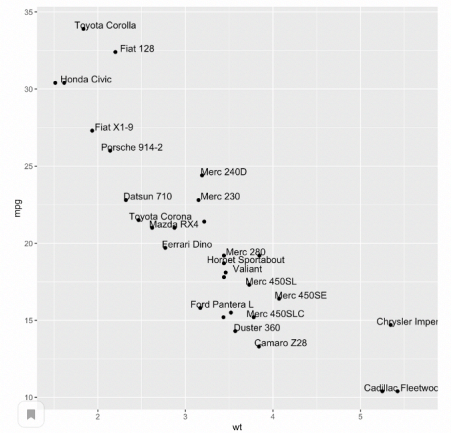
- Плотность распределения случайной величины X — производная от функции распепределения. По сути она показывает какую-то среднюю вероятность, приходящую на бесконечно малый отрезок

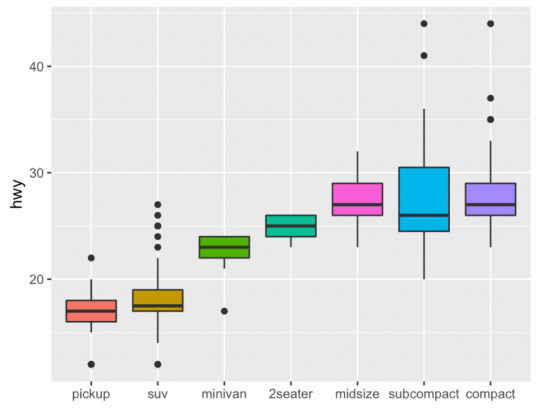
Немного о распределениях



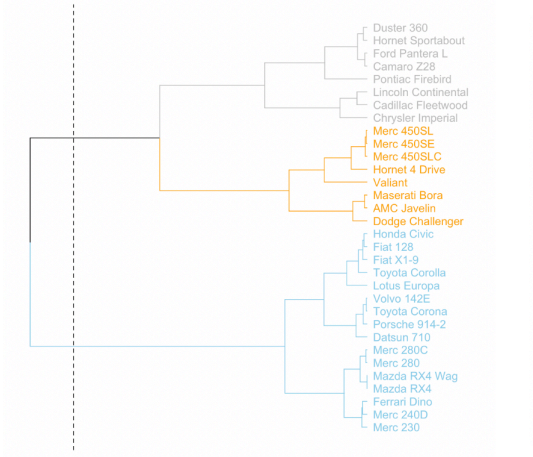


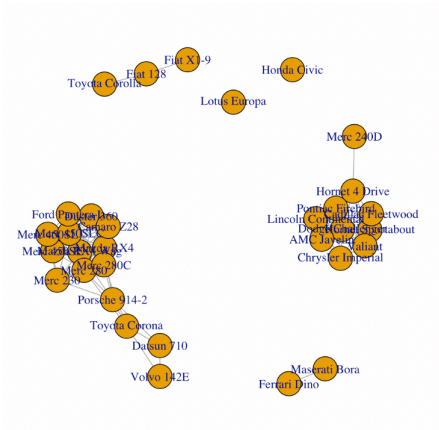
Возможности ggplot



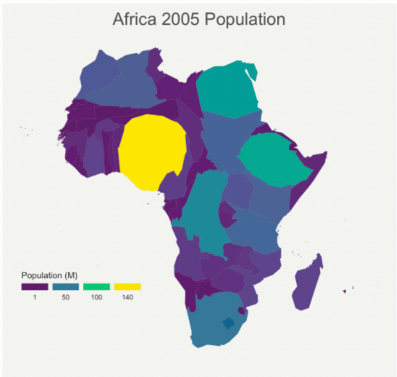


Возможности ggplot





Возможности ggplot



Очистка данных

Проблемы опросников

Респондент	Пол	Возраст	Зарплата
Максим	М	25	
Вика	Ж	19	60000
Андрей	М	219	75000
Ахмед	М	23	30000
Кирилл	М	22	55000
Александра	Ж		120000

- Пропущенные значения — кто-то из респондентов не захотел отвечать на вопросы

- Аномальные значения (выбросы) — при сборе допустилась ошибка, в следствие чего одно значение сильно отличается от остальных и не попадает в наше распределение

Некоторые способы решения

- Построение графиков для обнаружения выбросов (плотности распределения / barplot / ...)
- Нахождение основных статистик для обнаружения выбросов (среднее / максимальное значение / минимальное / ...)
- Удаление пропущенных значений
- Разумное заполнение пропущенных значений (нулями / средним / ...)

Трансформация данных

Получение подтаблиц

Например, мы хотим вывести тех людей, которые зарабатывают меньше 80000 рублей в месяц:

Респондент	Пол	Возраст	Зарплата
Павел	М	25	90000
Вика	Ж	19	60000
Андрей	М	22	75000
Миша	М	23	110000
Кирилл	М	22	55000
Александра	Ж	25	120000

Респондент	Пол	Возраст	Зарплата
Вика	Ж	19	60000
Андрей	М	22	75000
Кирилл	М	22	55000

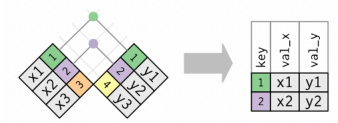
Агрегирование информации

Отдельно для мужчин и женщин хотим посчитать среднюю з/п:

Респондент	Пол	Возраст	Зарплата
Павел	М	25	90000
Вика	Ж	19	60000
Андрей	М	22	75000
Миша	М	23	110000
Кирилл	М	22	55000
Александра	Ж	25	120000

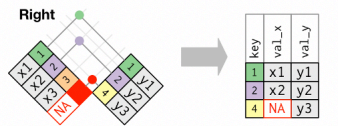
Пол	Средняя з/п
М	82500
Ж	90000

Объединение табличек

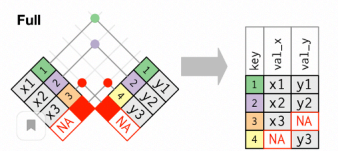


Left

Right



Full



Модели

Основные виды задач

- Регрессия — прогнозирование непрерывного значения (как правило это вещественное число) для конкретного наблюдения
- Классификация — отнесение наблюдения к одному из нескольких классов (частный случай: бинарная классификация)
- Кластеризация — отнесение наблюдения к одному из кластеров (как правило, не знаем сколько их и как устроены)

Примеры задач регрессии

- Бизнес: какая выручка магазина будет в следующем месяце?
- Экономика: какой спрос будет на товар в следующем году?
- Анализ изображений: сколько лет человеку на фотографии?
- Социология: сколько человек сэмигрирует в город N?

Примеры задач классификации

- Кредитный скоринг: вернет ли клиент кредит?
- Рекомендации: понравится ли пользователю фильм?
- Медицина: болен ли пациент?
- Биология: к какому виду цветков относится растение?
- Социология: зарабатывают ли женщины меньше мужчин?
- Баловство: Выживет ли пассажир на Титанике?

Примеры задач кластеризации

- Тексты: определение темы текста
- Маркетинг: поиск схожих пользователей в социальных сетях
- Социология: выделять группы схожих анкет
- Социология: выявлять типы людей и формировать поведенческие паттерны

Подробнее о моделях мы поговорим в следующей лекции!