НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Data Analysis National Olympiad - DANO

КАТЕГОРИАЛЬНЫЕ ПЕРЕМЕННЫЕ

Елена Семерикова,
Валерия Стефаненко

Москва, 2021

КОДИРОВКА ДАННЫХ

в повседневной жизни

КОДИРОВКА ДАННЫХ

— это такое представление переменных, которое разбивает их на категории с общими признаками или идентифицирует уникальность объекта.

РОССТАТ:

Средняя начисленная зарплата за октябрь 2019 г., руб.

Отношение заработной платы женщин к заработной плате мужчин, в процентах

Добыча полезных ископаемых	61 246	80 663	75,9%
Деятельность в области информации и связи	53 887	78 980	68,2%
Деятельность профессиональная, научная и техническая	55 187	77 719	71,0%
Строительство	53 296	61 263	87,0%
Транспортировка и хранение	41 431	55 943	74,1%
Обрабатывающие производства	40 669	52 500	77,5%
Торговля оптовая и розничная; ремонт автотранспортных средств и мотоциклов	41 160	52 122	79,0%
Деятельность в области культуры, спорта, организации досуга и развлечений	35 079	47 265	74,2%

Источник: https://rosstat.gov.ru/storage/mediabank/BPL9rLwU/Sbornik.pdf

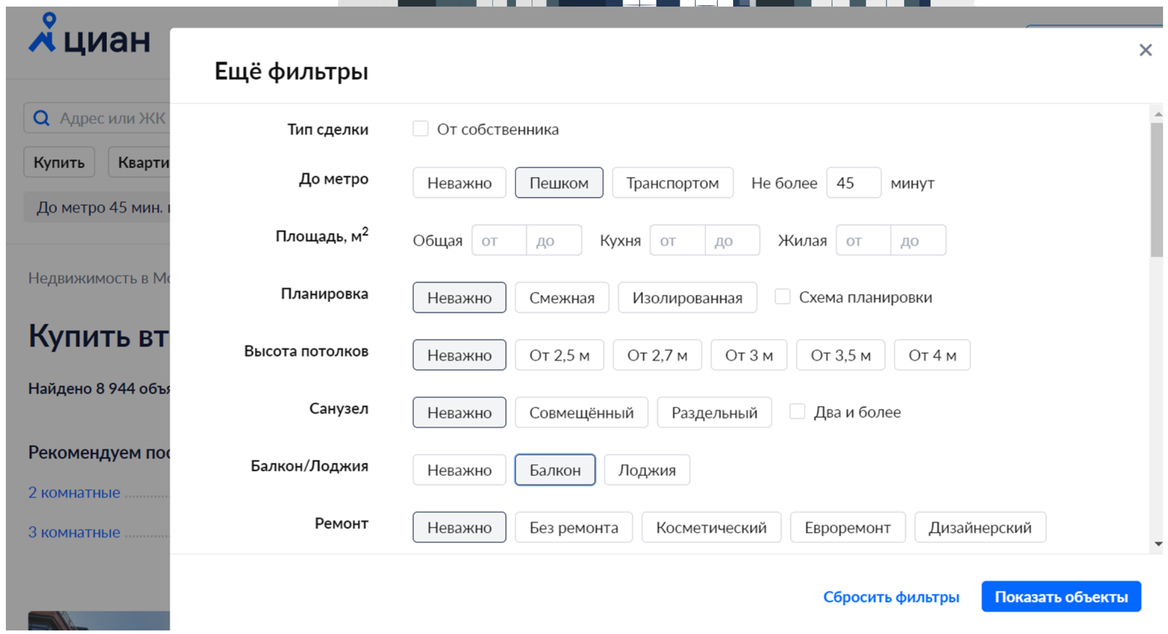
ЯНДЕКС.GO:
данные по поездкам в такси

Час пик	Дамми переменная
Нет	0
Да	1
Нет	0
Нет	0
Нет	0
Нет	0
Да	1
Нет	0
Нет	0
Нет	0
Нет	0
Да	1
Нет	0
Нет	0
Нет	0
Да	1

ДАММИ-ПЕРЕМЕННАЯ, или ФИКТИВНАЯ ПЕРЕМЕННАЯ

разбивает переменные
на 2 категории «0» и «1»
с общими признаками
для удобства их сравнения
и исследования уникальных
для таких категорий признаков

ЦИАН:
Поиск жилья



№	Доступность метро	On_foot	Наличие балкона	Balcony
1	пешком	1	Да	1
2	транспортом	0	Да	1
3	транспортом	0	Нет	0
4	пешком	1	Да	1
5	пешком	1	Нет	0
6	транспортом	0	Нет	0
7	транспортом	0	Да	1

ЯНДЕКС.GO:

данные по поездкам в такси

УТРО = «1»
ДЕНЬ = «2»
ВЕЧЕР = «3»

Время суток	Period	Period_1	Period_2	Period_3
Вечер	3	0	0	1
день	2	0	1	0
утро	1	1	0	0
утро	1	1	0	0
утро	1	1	0	0
день	2	0	1	0
день	2	0	1	0
утро	1	1	0	0
утро	1	1	0	0
день	2	0	1	0
день	2	0	1	0
утро	1	1	0	0
день	2	0	1	0
вечер	3	0	0	1
вечер	3	0	0	1
день	2	0	1	0

РАЗМЕР ПРЕДПРИЯТИЙ

в рамках государственной программы поддержки

от количественных к качественным переменным

Размер предприятия	Обороты, руб	Категория	Business_size
до 15 человек	120 млн	микропредприятия	1
до 100 человек	800 млн	малые	2
до 250 человек	2 млрд	средние	3

Источник: http://base.garant.ru/12154854/

ВВП СТРАНЫ 2019

перекодировка данных

Данные РОССТАТ, млрд ₽ → Данные Worldbank, Млрд $

Сопоставимость данных

110 046,1 — 1 687,4

Данные Worldbank, $ → Данные Worldbank, млрд $

Удобство представления данных

1,68745E+12 → 1 687,4

ЗАДАНИЕ 1

для самопроверки

Допустим, вы исследуете стоимость овощей. Для этого приходите на рынок каждый месяц и записываете их среднюю стоимость. Собранные вами данные представлены в таблице справа.

1. Поясните, как закодировать данные на основании времени года?

2. Во сколько раз картофель стоит дороже весной, чем осенью?

Месяц	Стоимость картофеля за кг, руб
Декабрь	20
Январь	21,5
Февраль	21,5
Март	22
Апрель	35
Май	40
Июнь	40
Июль	25
Август	17
Сентябрь	18
Октябрь	18
Ноябрь	18,5

ЗАДАНИЕ 2

для самопроверки

В Москве можно увидеть довольно много многоэтажек. Однако в последнее время также пользуется популярностью и малоэтажное строительство. Вы располагаете данными о количестве этажей в нескольких домах, а также о технологии их строительства.

1. Подумайте, сколько категорий можно придумать для группировки данных, по какому принципу?

2. Закодируйте данные по принципу технологии.

3. Сформируйте дамми-переменную, исходя из классификации:
   a. малоэтажный- 1-4 этажа (с учетом мансарды);
   b. многоэтажный - 5 этажей и выше

Дом	Кол-во этажей	Технология
1	12	Панельный
2	4	Монолитный
3	7	Панельный
4	3	Кирпичный
5	9	Панельный
6	5	Монолитный
7	3	Кирпичный
8	2	Монолитный

ЗАДАНИЕ 3

для самопроверки

Справа представлены данные прогноза погоды для США и Канады. Однако в странах применяются разные метрические системы для измерения температур: в США - градус Фаренгейта, а в Канаде – градус Цельсия.

Ответьте на вопрос, на сколько средняя температура за неделю в Цельсиях в США отличается от средней температуры за неделю в Канаде?

Дата	Температура в США, °F	Температура в Канаде, °C
26/07/2021	87	27
27/07/2021	90	26
28/07/2021	86	24
29/07/2021	87	25
30/07/2021	86	25
31/07/2021	86	25
01/08/2021	84	23

ПОЛЕЗНЫЕ ССЫЛКИ

для изучения

Борис Демешев. Дамми (фиктивные) переменные. Разные зависимости для подвыборок
https://www.youtube.com/watch?v=fxIVaU9OMn8

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ