

# Описание данных

Н.В. Артамонов

11 апреля 2023 г.

## Содержание

<b>1 Кросс-секционные данные</b>	<b>1</b>
<b>2 Временные ряды</b>	<b>8</b>
<b>3 Панельные данные</b>	<b>9</b>

В этом разделе приведены описания основных наборов данных, используемых для решения задач.

Для работы с нужным датасетом необходимо загрузить его командой

```
data(dataset, package)
```

## 1 Кросс-секционные данные

Таблица 1: Набор данных `Labour` из пакета `Ecdat` (569 наблюдений) с данными о бельгийских фирмах за 1996 г.

<code>capital</code>	капитал (в млн евро)
<code>labour</code>	число сотрудников
<code>output</code>	выпуск (в млн евро)
<code>wage</code>	зарплата на одного сотрудника (в тыс евро)

Таблица 2: Набор данных **Electricity** из пакета **Ecdat** (158 наблюдений) с данными о производителях электроэнергии в US за 1970 г.

cost	общие издержки за год
q	общий выпуск электроэнергии
pl	уровень зарплата (wage rate)
pk	цена привлечения капитала (capital price index)
pf	цена на топливо (fuel price)

Таблица 3: Набор данных **sleep75** из пакета **wooldridge** (706 наблюдений). Основные переменные. Источник данных [5]

age	возраст (в годах)
educ	уровень образования (в года)
inlf	бинарная, 1 если участник рынка труда
leis1	нерабочее время, sleep-totwrk
smsa	бинарная, 1 если живёт в мегаполисе
male	гендерный фактор (бинарная, 1 если мужчина)
marr	семейный статус (бинарная, 1 если женат/замужем)
prot	бинарная, 1 если протестант
selfe	бинарная, 1 если самозанятый
sleep	продолжительность сна (мин/нед)
south	географический фактор (бинарная, 1 если живёт на юге)
spsepay	доход супруга/супруги
spwrk75	бинарная, 1 если супруг(а) работает
totwrk	занятость (мин/нед)
union	бинарная, 1 если член профсоюза
yngkid	бинарная, 1 если есть дети младше 3 лет
ytsmarr	сколько лет женат/замужем
hrwage	почасовая оплата

Таблица 4: Набор данных `wage2` из пакета `wooldridge` (935 наблюдений).  
Основные переменные. Источник данных [6]

<code>wage</code>	месячная зарплата
<code>hours</code>	недельная занятость в часах
<code>IQ</code>	результаты теста IQ
<code>KWW</code>	результаты теста knowledge of world work
<code>educ</code>	уровень образования (в годах)
<code>exper</code>	опыт работы в годах
<code>tenure</code>	стаж работы на текущем месте
<code>age</code>	возраст (в годах)
<code>married</code>	семейный статус (бинарная, 1 если женат/замужем)
<code>south</code>	географический фактор (бинарная, 1 если живёт на юге)
<code>urban</code>	место жительства (1 если живет в городе)
<code>sibs</code>	число братьев/сестёр
<code>brthord</code>	какой по счёту ребёнок в семье
<code>meduc</code>	уровень образования матери (в годах)
<code>feduc</code>	уровень образования отца (в годах)

Таблица 5: Набор данных `wage1` из пакета `wooldridge` (526 наблюдений).  
Основные переменные.

<code>wage</code>	средняя почасовая оплата
<code>educ</code>	уровень образования (в годах)
<code>female</code>	гендерный фактор
<code>exper</code>	опыт работы
<code>tenure</code>	стаж на текущем месте работы
<code>married</code>	семейный статус
<code>smsa</code>	живёт ли в мегаполисе (бинарная)
<code>south</code>	географический фактор (бинарная)
<code>west</code>	географический фактор (бинарная)
<code>northcen</code>	географический фактор (бинарная)

Таблица 6: Набор данных `loanapp` из пакета `wooldridge` (1989 наблюдений). Основные переменные. Источник данных [9]

<code>approve</code>	бинарная, 1 если кредитная заявка одобрена
<code>appinc</code>	доход заявителя (в \$1000)
<code>mortno</code>	бинарная, 1 если нет ипотечной кредитной истории
<code>unem</code>	уровень безработицы в отрасли в %
<code>dep</code>	количество иждивенцев
<code>male</code>	гендерный фактор
<code>married</code>	семейный статус
<code>yjob</code>	стаж на текущей работе
<code>self</code>	бинарная, 1 если самозанятый

Таблица 7: Набор данных `SwissLabor` о рынке труда Швейцарии из пакета `AER` (872 наблюдений). Источник данных [7]

<code>participation</code>	Является ли участником рынка труда? (фактор, "yes"/"no")
<code>income</code>	логарифм дополнительного дохода (nonlabor income)
<code>age</code>	возраст (в десятилетиях)
<code>education</code>	уровень образования
<code>youngkids</code>	число маленьких детей (младше 7 лет)
<code>oldkids</code>	число старших детей (старше 7 лет)
<code>foreign</code>	является ли иностранцем? (фактор, "yes"/"no")

Таблица 8: Набор данных из файла **Mroz** из пакета **Ecdat** содержит данные о рынке труда замужних женщин. Основные переменные. Источник данных [10]

LFP	бинарная, 1 женщина работает
WHRS	уровень занятости (в часах)
KL6	число детей моложе 6 лет в семье
K618	число детей от 6 до 18 лет в семье
WA	возраст
WE	уровень образования (в годах)
WW	средняя почасовая оплата
HHRS	занятость мужа
HA	возраст мужа
HE	уровень образования мужа (а годах)
HW	зарплата мужа
FAMINC	доход домашнего хозяйства
WMED	уровень образования матери
WFED	уровень образования отца
UN	уровень безработицы в стране проживания
CIT	бинарная, 1 если живет в мегаполисе
AX	предыдущий опыт работы (в годах)

Таблица 9: Набор данных **diamonds** из пакета **ggplot2** с данными о бриллиантах (53940 наблюдений). Основные переменные.

price	цена бриллианта
carat	вес бриллианта (в каратах)
cut	качество огранки (упорядоченный фактор с уровнями Fair<Good< Very Good<Premium<Ideal)
color	цвет (упорядоченный фактор с уровнями J<I<H<G<F<E<D)
clarity	прозрачность (упорядоченный фактор с уровнями I1<SI2<SI1<VS2<VS1<VVS2<VVS1<IF)
x	длина (в мм)
y	ширина (в мм)
z	глубина (в мм)

Таблица 10: Набор данных `Diamond` из пакета `Ecdat` с данными о бриллиантах (308 наблюдений). Основные переменные. Источник данных [4]

carat	вес бриллианта (в каратах)
colour	цвет (фактор с уровнями D,E,F,G,H,I)
clarity	прозрачность (фактор с уровнями IF,VVS1,VVS2,VS1,VS2)
certification	орган по сертификации (фактор с уровнями GIA,IGI,HRD)
price	цена в Сингапуре

Таблица 11: Набор данных `nlsw88` из пакета `Counterfactual` о занятости женщин в США (2246 наблюдений). Основные переменные. Источник данных: сайт Stata <http://www.stata-press.com/data/r10/g.html>

hours	недельная занятость (в часах)
married	семейный статус
ttl_exp	общий стаж работы
smsa	бинарная, 1 если живёт в мегаполисе
south	географический фактор (бинарная, 1 если живёт на юге)
wage	почасовая оплата (в \$)
age	возраст (в годах)
grade	уровень образования (в годах)

Таблица 12: Набор данных из файла `default.csv` с сайта <http://meit.mgimo.ru/node/237> о банкротствах по студенческим займам (6778 наблюдений)

default	бинарная переменная, равная 1 если индивид признал себя банкротом по студенческому займу
age	возраст
adepcnt	количество иждивенцев у индивида плюс 1
acadmos	количество месяцев, которые индивид прожил по текущему адресу
majordrg	количество зарегистрированных серьёзных правонарушений у этого индивида
minordrg	количество зарегистрированных мелких правонарушений у этого индивида
ownrent	1, если индивид живёт в собственном доме, и 0, если снимает
income	месячный доход в \$
spending	среднемесячный расход по кредитной карте
inc_rep	income, делённая на количество иждивенцев
exp_inc	доля месячных расходов по кредитной карте в годовой заработной плате
selfempl	1, если индивид самозанятый, и 0 иначе

Таблица 13: Набор данных `stockton3` из пакета `PoEdata` с данными о стоимости домов (2610 наблюдений). Основные переменные. Источник данных <https://github.com/ccolonescu/PoEdata>

sprice	цена продажи дома (в \$)
livarea	жилая площадь (кв.футы)
pool	наличие бассейна (бинарная)
lgelot	размер участка (бинарный фактор, 1 если участок больше 5 акров)
age	возраст (в годах)
beds	число спален

Таблица 14: Набор данных из файла `applications.csv` о поступивших на магистерские и PhD-программы (400 наблюдений)

<code>admit</code>	бинарный фактор, 1 если заявка одобрена
<code>GPA</code>	средняя оценка за время обучения
<code>GRE</code>	балл за экзамен graduate record exam
<code>rank</code>	категориальная переменная, обозначающая престиж университета (1 – высший престиж, 4 – низший престиж)

Таблица 15: Набор данных `Consumption` из пакета `Ecdat` об индивидуальных доходах и расходах в Канаде (квартальные данные с 1947Q1 по 1996Q4).

<code>yd</code>	индивидуальный располагаемый доход в ценах 1986
<code>se</code>	индивидуальные расходы на потребление в ценах 1986

## 2 Временные ряды

Таблица 16: Набор данных `Icescream` из пакета `Ecdat` о потреблении мороженого в США (недельные данные с 1951-03-18 по 1953-07-11, всего 30 наблюдений). Источник данных [8]

<code>cons</code>	потребление мороженого (в пинтах)
<code>income</code>	средний недельный доход семьи (в \$)
<code>price</code>	цена мороженого (за пинту)
<code>temp</code>	средняя температура (по Фаренгейту)



Таблица 17: Панель **Guns** из пакета **AER** с данными по 51 штату США с 1977 по 1999 гг. (всего 1173 наблюдения). Основные переменные. Источник данных [1]

state	factor indicating state
year	factor indicating year
violent	violent crime rate (incidents per 100,000 members of the population)
murder	murder rate (incidents per 100,000).
robbery	robbery rate (incidents per 100,000)
prisoners	incarceration rate in the state in the previous year (sentenced prisoners per 100,000 residents; value for the previous year)
afam	percent of state population that is African-American, ages 10 to 64
cauc	percent of state population that is Caucasian, ages 10 to 64
male	percent of state population that is male, ages 10 to 29
population	state population, in millions of people.
income	real per capita personal income in the state (US dollars)
density	population per square mile of land area, divided by 1,000
law	factor. Does the state have a shall carry law in effect in that year?

### 3 Панельные данные

Таблица 18: Панель **LaborSupply** из пакета **plm**, **Ecdat** с данными по 532 индивидуумам с 1979 по 1988 гг. (всего 5320 наблюдений). Основные переменные. Источник данных [\[11\]](#)

lnhr	логарифм годовой занятости в часах
lnwge	логарифм почасовой оплаты
kids	число детей
age	возраст
disab	бинарная, 1 если плохое здоровье

Таблица 19: Панель **Cigar** из пакета **plm** с данными по 46 штатам США с 1963 по 1992 гг. (всего 1380 наблюдений). Основные переменные. Источник данных [\[2\]](#)

state	state abbreviation
year	the year
price	price per pack of cigarettes
pop	population
pop16	population above the age of 16
cpi	consumer price index (1983=100)
ndi	per capita disposable income
sales	cigarette sales in packs per capita
pimin	minimum price in adjoining states per pack of cigarettes

Таблица 20: Панель **Gasoline** из пакета **plm**, **Ecdat** с данными о потреблении бензина по 18 странам OECD с 1960 по 1978 гг. (всего 342 наблюдений). Основные переменные. Источник данных [\[3\]](#)

lgaspcar	логарифм потребления бензина
lincomer	логарифм реального дохода на душу населения
lrgmg	логарифм реальной цены на бензин
lcarpcar	логарифм объёма рынка машин

Таблица 21: Панель Loan Dream Housing Finance company deals in all home loans. They have a presence across all urban, semi-urban, and rural areas. Customer-first applies for a home loan after that company validates the customer eligibility for a loan.(всего 342 наблюдений). Основные переменные. Источник данных [Loan](#)

LoanID	Unique Loan ID
Gender	Male/ Female
Married	applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Under Graduate)
Self Employed	Self-employed (Y/N)
ApplicantIncome	ApplicantIncome
CoapplicantIncome	CoapplicantIncome
LoanAmount	LoanAmount
Loan Amount Term	Loan Amount Term
Credit History	credit history meets guidelines
Property Area	Urban/ Semi-Urban/ Rural
Loan Status	Loan approved

Таблица 22: 1Customers.csv Shop Customer Data is a detailed analysis of a imaginative shop’s ideal customers. It helps a business to better understand its customers. The owner of a shop gets information about Customers through membership cards. Dataset consists of 2000 records and 8 columns.

Customer ID	—
Gender	—
Age	—
Annual Income	—
Spending Score	Score assigned by the shop, based on customer behaviour and spending nature
Profession	—
Work Experience	in years
Family Size	—

Таблица 23: `2laptops.csv` This is a dataset about laptops scrap from Flipkart which contains information on various aspects of laptops such as their price, discount, specifications, and warranty. The dataset contains a total of 920 entries, each representing a single laptop. It is important to note that there are some missing values in the dataset for some of the columns, including "discount "In build sw and "warranty". This should be taken into consideration when analyzing the data. Additionally, all the columns have a data type of "object which may require further processing to convert the data into a usable format.

title	provides a brief description of the laptop
price	includes the cost of the laptop
discount	mentions any applicable discounts on the laptop's price
Processor	specifies the type of processor used in the laptop
RAM	mentions the amount of RAM the laptop has
OS	lists the operating system installed on the laptop
SSD	indicates the size of the solid-state drive
Display	mentions the screen size and display specifications of the laptop
In build sw	lists any software pre-installed on the laptop
warranty	provides information on the warranty offered for the laptop

Таблица 24: `3insurance.csv` Machine Learning with R by Brett Lantz is a book that provides an introduction to machine learning using R. As far as I can tell, Packt Publishing does not make its datasets available online unless you buy the book and create a user account which can be a problem if you are checking the book out from the library or borrowing the book from a friend. All of these datasets are in the public domain but simply needed some cleaning up and recoding to match the format in the book.

age	age of primary beneficiary
sex	insurance contractor gender, female, male
bmi	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height objective index of body weight ( $kg/m^2$ ) using the ratio of height to weight, ideally 18.5 to 24.9
children	Number of children covered by health insurance
	Number of dependents
smoker	Smoking
region	the beneficiary's residential area in the US, northeast, southeast southwest, northwest.
charges	Individual medical costs billed by health insurance

Таблица 25: `4winequality-red.csv` The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones). This dataset is also available from the UCI machine learning repository, [winequality](#) Content

fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
Output variable (based on sensory data)
quality (score between 0 and 10)

Таблица 26: 5marketing-campaign.csv Customer Personality Analysis is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors and concerns of different types of customers. Customer personality analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company's database, a company can analyze which customer segment is most likely to buy the product and then market the product only on that particular segment. Content:

#### People

ID	Customer's unique identifier
YearBirth	Customer's birth year
Education	Customer's education level
Marital Status	Customer's marital status
Income	Customer's yearly household income
Kidhome	Number of children in customer's household
Teenhome	Number of teenagers in customer's household
Dt Customer	Date of customer's enrollment with the company
Recency	Number of days since customer's last purchase
Complain	1 if the customer complained in the last 2 years, 0 otherwise

#### Products

MntWines	Amount spent on wine in last 2 years
MntFruits	Amount spent on fruits in last 2 years
MntMeatProducts	Amount spent on meat in last 2 years
MntFishProducts	Amount spent on fish in last 2 years
MntSweetProducts	Amount spent on sweets in last 2 years
MntGoldProds	Amount spent on gold in last 2 years

#### Promotion

NumDealsPurchases	Number of purchases made with a discount
AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise
Response	1 if customer accepted the offer in the last campaign, 0 otherwise

#### Place

NumWebPurchases	Number of purchases made through the company's website
NumCatalogPurchases	Number of purchases made using a catalogue
NumStorePurchases	Number of purchases made directly in stores
NumWebVisitsMonth	Number of visits to company's website in the last month

Таблица 27: Панель 6 `PlacementDataFullClass.csv` This data set consists of Placement data of students in a XYZ campus. It includes secondary and higher secondary school percentage and specialization. It also includes degree specialization, type and Work experience and salary offers to the placed students

Таблица 28: `californiaHousingPrices.csv` Context This is the dataset used in the second chapter of Aurélien Géron’s recent book ‘Hands-On Machine learning with Scikit-Learn and TensorFlow’. It serves as an excellent introduction to implementing machine learning algorithms because it requires rudimentary data cleaning, has an easily understandable list of variables and sits at an optimal size between being too toyish and too cumbersome. The data contains information from the 1990 California census. So although it may not help you with predicting current housing prices like the Zillow Zestimate dataset, it does provide an accessible introductory dataset for teaching people about the basics of machine learning. [california](#) Content

1. longitude	A measure of how far west a house is; a higher value is farther west
2. latitude	A measure of how far north a house is; a higher value is farther north
3. housingMedianAge	Median age of a house within a block; a lower number is a newer building
4. totalRooms	Total number of rooms within a block
5. totalBedrooms	Total number of bedrooms within a block
6. population	Total number of people residing within a block
7. households	Total number of households, a group of people residing within a home unit, for a block
8. medianIncome	Median income for households within a block of houses (measured in tens of thousands of US Dollars)
9. medianHouseValue	Median house value for households within a block (measured in US Dollars)
10. oceanProximity	Location of the house w.r.t ocean/sea



Таблица 29: SpotifyYoutube.csv Context Dataset of songs of various artist in the world and for each song is present: Several statistics of the music version on spotify, including the number of streams; Number of views of the official music video of the song on youtube. [youtubespotify](#) Content

Track	name of the song, as visible on the Spotify platform.
Artist	name of the artist.
Urlspotify	the Url of the song.
Album	the album in wich the song is contained on Spotify.
Album type	indicates if the song is relesead on Spotify as a single or contained in an album.
Uri	spotify link used to find the song through the API.”
Danceability	describes how suitable a track is for dancing based on a beat strength, andcombination of musical elements including tempo, rhythm stability, overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
Energy	is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
Key	the key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C $\sharp$ /D $\flat$ , 2 = D, and so on. If no key was detected, the value is -1.
Loudness	the overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.
Speechiness	detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
Acousticness	a confidence measure <sup>17</sup> from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.

Instrumentalness	<p>predicts whether a track contains no vocals.</p> <p>"Ooh" and "aah" sounds are treated as instrumental in this context.</p> <p>Rap or spoken word tracks are clearly "vocal".</p> <p>The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.</p> <p>Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.</p>
Liveness	<p>detects the presence of an audience in the recording.</p> <p>Higher liveness values represent an increased probability that the track was performed live.</p>
Valence	<p>A value above 0.8 provides strong likelihood that the track is live.</p> <p>a measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.</p> <p>Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).</p>
Tempo	<p>the overall estimated tempo of a track in beats per minute (BPM).</p> <p>In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.</p>
Duration ms	the duration of the track in milliseconds.
Stream	number of streams of the song on Spotify.
Url youtube	url of the video linked to the song on Youtube, if it have any.
Title	title of the videoclip on youtube.
Channel	name of the channel that have published the video.
Views	number of views.
Likes	number of likes.
Comments	number of comments.
Description	description of the video on Youtube.
Licensed	Indicates whether the video represents licensed content, which means that the content was uploaded to a channel linked to a YouTube content partner and then claimed by that partner.
official video	boolean value that indicates if the video found is the official video of the song.

Таблица 30: `MedicalCostPersonalDatasets.csv` Context Machine Learning with R by Brett Lantz is a book that provides an introduction to machine learning using R. As far as I can tell, Packt Publishing does not make its datasets available online unless you buy the book and create a user account which can be a problem if you are checking the book out from the library or borrowing the book from a friend. All of these datasets are in the public domain but simply needed some cleaning up and recoding to match the format in the book. **MedCostPersData** Content

age	age of primary beneficiary
sex	insurance contractor gender, female, male
bmi	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight ( $kg/m^2$ ) using the ratio of height to weight, ideally 18.5 to 24.9
children	Number of children covered by health insurance / Number of dependents
smoker	Smoking
region	the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
charges	Individual medical costs billed by health insurance

Таблица 31: `Studentsgamificationgrades.csv` Context This dataset was created as a result of students who used a gamified learning application with interactive flashcards and badges to engage them in learning about statistics and multidimensional statistical analysis. The dataset includes data on various variables, including the practice exam grades before using the educational platform, the final exam grades after using the platform, whether or not the student is a user of the platform, the average grade for each of the six quizzes, and the number of times each quiz was submitted. **Gamification in education** Content

Таблица 32: `SocialNetworkAds.csv` Context The dataset is simple, you have information on gender, age and whether the user purchased the product shown to them or not. You have to determine if the user will buy. The user is identified by a unique identifier in the ID column. The unique identifier doesn't need to be used for the model, but you should keep it for testing purposes! [Gamification in education](#) Content

Таблица 33: `madridHousesClean.csv` Context Clean dataset for machine learning. [Madrid houses-clean](#) Content

id	identifier
sq mt built	size of the house in square meter built.
n rooms	number of rooms.
n bathrooms	number of bathrooms.
sq mt allotment	for detached houses, full size including house, pool, garden, etc.
floor	height.
is renewal needed	boolean.
has lift	boolean.
is exterior	boolean.
energy certificate	higher values mean mor efficient.
has parking	boolean.
neighborhood	Madrid's counties.
district	Madrid's districts.
house type	flat, house, study, duplex, top floor.
buy rice	target.

Таблица 34: `SocialNetworkAds.csv` Context The dataset is simple, you have information on gender, age and whether the user purchased the product shown to them or not. You have to determine if the user will buy. The user is identified by a unique identifier in the ID column. The unique identifier doesn't need to be used for the model, but you should keep it for testing purposes! [Gamification in education](#) Content

Таблица 35: **BankPersonalLoanModelling.csv** Data Description: The file Bank.xls contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6) accepted the personal loan that was offered to them in the earlier campaign. Domain:Banking Context: This case is about a bank (Thera Bank) whose management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9 success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget.

**BankPersonalLoanModelling** Content

ID	Customer ID
Age	Customer's age in completed years
Experience	years of professional experience
Income	Annual income of the customer (000)
ZIP Code	Home Address ZIP code.
Family	Family size of the customer
CCAvg	Avg. spending on credit cards per month (000)
Education	Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
Mortgage	Value of house mortgage if any. (000)
10.Personal Loan	Did this customer accept the personal loan offered in the last campaign?
11.Securities Account	Does the customer have a securities account with the bank?
12.CD Account	Does the customer have a certificate of deposit (CD) account with the bank?
13.Online	Does the customer use internet banking facilities?
14.Credit card	Does the customer use a credit card issued by

Таблица 36: **udemycourses.csv** **Udemy Courses** Content

Таблица 37: `exams.csv` Context Description: This dataset contains information on the performance of high school students in mathematics, including their grades and demographic information. The data was collected from three high schools in the United States. This dataset was created for educational purposes and was generated, not collected from actual data sources. `exams` Content

Gender	The gender of the student (male/female)
Race/ethnicity	The student's racial or ethnic background (Asian, African-American, Hispanic, etc.)
Parental level of education	The highest level of education attained by the student's parent(s) or guardian(s)
Lunch	Whether the student receives free or reduced-price lunch (yes/no)
Test preparation course	Whether the student completed a test preparation course (yes/no)
Math score	The student's score on a standardized mathematics test
Reading score	The student's score on a standardized reading test
Writing score	The student's score on a standardized writing test

Таблица 38: `madridHousesClean.csv` Context Clean dataset for machine learning. **Madrid houses-clean** Content

id	identifier
sq mt built	size of the house in square meter built.
n rooms	number of rooms.
n bathrooms	number of bathrooms.
sq mt allotment	for detached houses, full size including house, pool, garden, etc.
floor	height.
is renewal needed	boolean.
has lift	boolean.
is exterior	boolean.
energy certificate	higher values mean mor efficient.
has parking	boolean.
neighborhood	Madrid's counties.
district	Madrid's districts.
house type	flat, house, study, duplex, top floor.
buy rice	target.

## Список литературы

- [1] Ayres, I., and Donohue, J.J. (2003). Shooting Down the ‘More Guns Less Crime’ Hypothesis. *Stanford Law Review*, 55, 1193–1312
- [2] Baltagi B, Levin D (1992). Cigarette taxation: Raising revenues and reducing consumption. *Structural Change and Economic Dynamics*, 3(2), 321-335
- [3] Baltagi, B.H. and Y.J. Griggin (1983) “Gasoline demand in the OECD: an application of pooling and testing procedures”, *European Economic Review*, 22.
- [4] Chu, Singfat (2001) “Pricing the C’s of Diamond Stones”, *Journal of Statistics Education*, 9(2).
- [5] J.E. Biddle and D.S. Hamermesh (1990), “Sleep and the Allocation of Time,” *Journal of Political Economy* 98, 922-943.

Таблица 39: CarPriceAssignment.csv Context A Chinese automobile company Geely Auto aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts. They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market. The company wants to know: Which variables are significant in predicting the price of a car How well those variables describe the price of a car Based on various market surveys, the consulting firm has gathered a large data set of different types of cars across the America market.

#### Car Price Prediction Multiple Linear Regression Content

CarID	Unique id of each observation (Integer)
Symboling	Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.(Categorical)
carCompany	Name of car company (Categorical)
fueltype	Car fuel type i.e gas or diesel (Categorical)
aspiration	Aspiration used in a car (Categorical)
doornumber	Number of doors in a car (Categorical)
carbody	body of car (Categorical)
drivewheel	type of drive wheel (Categorical)
enginelocation	Location of car engine (Categorical)
wheelbase	Weelbase of car (Numeric)
carlength	Length of car (Numeric)
carwidth	Width of car (Numeric)
carheight	height of car (Numeric)
curbweight	The weight of a car without occupants or baggage.
enginetype	Type of engine. (Categorical)
cylindernumber	cylinder placed in the car (Categorical)
enginesize	Size of car (Numeric)
fuelsystem	Fuel system of car (Categorical)
boreratio	Boreratio of car (Numeric)
stroke	Stroke or volume inside the engine (Numeric)
compressionratio	compression ratio of car (Numeric)
horsepower	Horsepower (Numeric)
peakrpm	car peak rpm (Numeric)
citympg	Mileage in city (Numeric)
highwaympg	Mileage on highway (Numeric)
price	(Dependent variable) Price of car (Numeric)



- [6] M. Blackburn and D. Neumark (1992), “Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials”, *Quarterly Journal of Economics* 107, 1421-1436.
- [7] Gerfin, M. (1996). Parametric and Semi-Parametric Estimation of the Binary Response Model of Labour Market Participation. *Journal of Applied Econometrics*, 11, 321–339.
- [8] Hildreth, C. and J. Lu (1960) Demand relations with autocorrelated disturbances, Technical Bulletin No 2765, Michigan State University.
- [9] W.C. Hunter and M.B. Walker (1996), “The Cultural Affinity Hypothesis and Mortgage Lending Decisions”, *Journal of Real Estate Finance and Economics* 13, 57-70
- [10] Mroz, T. (1987) “The sensitivity of an empirical model of married women’s hours of work to economic and statistical assumptions”, *Econometrica*, 55, 765-799.
- [11] Ziliak, Jim (1997) “Efficient Estimation With Panel Data when Instruments are Predetermined: An Empirical Comparison of Moment-Condition Estimators”, *Journal of Business and Economic Statistics*, 419-431