Предобработка данных

Елена Эльзессер,
Наставник DA, Яндекс.Практикум

Яндекс Практикум

как проект?

0
не понимаю

5
мне надо подумать

10
у меня вагон идей

План

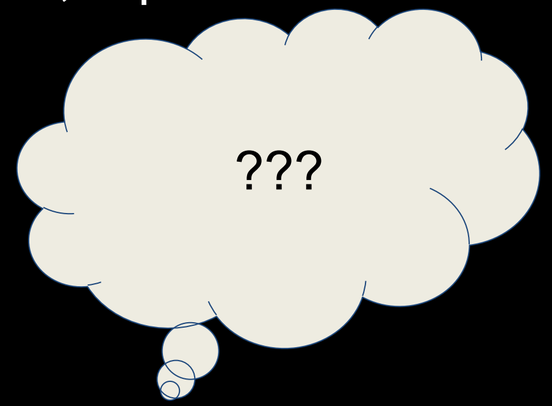
for feature in DataFrame.columns:

- посмотрим на данные

- поругаем все, что нам не нравится

- поймем, как исправить

придумаем для себя “догмы” предобработки данных (дубликаты, пропуски, скрытые пропуски, фиктивные значения, аномалии, некорректные форматы)



DataFrame.info()

RangeIndex: 57429 entries, 0 to 57428
Data columns (total 21 columns):

#	Column	Non-Null Count	Dtype
0	Advertisement age restriction	5395 non-null	object
1	Advertisement display type	12207 non-null	object
2	Advertisement duration	57429 non-null	int64
3	Advertisement expected duration	57429 non-null	int64
4	Advertisement file type	57429 non-null	object
5	Advertisement format	34505 non-null	object
6	Advertisement ID	57429 non-null	int64
7	Advertisement player	40766 non-null	object
8	Advertisement skip	37475 non-null	object
9	Advertisement TV Clip ID	57429 non-null	int64
10	Advertiser	57429 non-null	object
11	Article level1	57429 non-null	object
12	Article level4	57429 non-null	object
13	Banner Network	57427 non-null	object
14	Brands list	57429 non-null	object
15	Day	57429 non-null	datetime64[ns]
16	Day type	57429 non-null	object
17	First issue date	57429 non-null	datetime64[ns]
18	Holding	57429 non-null	object
19	Site	57429 non-null	object
20	Week day	57429 non-null	object

Duplicate rows	11204
Duplicate rows (%)	19.5%

DataFrame.info()

RangeIndex: 57429 entries, 0 to 57428
Data columns (total 21 columns):

#	Column	Non-Null Count	Dtype
0	Advertisement age restriction	5395 non-null	object
1	Advertisement display type	12207 non-null	object
2	Advertisement duration	57429 non-null	int64
3	Advertisement expected duration	57429 non-null	int64
4	Advertisement file type	57429 non-null	object
5	Advertisement format	34505 non-null	object
6	Advertisement ID	57429 non-null	int64
7	Advertisement player	40766 non-null	object
8	Advertisement skip	37475 non-null	object
9	Advertisement TV Clip ID	57429 non-null	int64
10	Advertiser	57429 non-null	object
11	Article level1	57429 non-null	object
12	Article level4	57429 non-null	object
13	Banner Network	57427 non-null	object
14	Brands list	57429 non-null	object
15	Day	57429 non-null	datetime64[ns]
16	Day type	57429 non-null	object
17	First issue date	57429 non-null	datetime64[ns]
18	Holding	57429 non-null	object
19	Site	57429 non-null	object
20	Week day	57429 non-null	object

Duplicate rows	11204
Duplicate rows (%)	19.5%

1. названия столбцов
2. много пропусков
3. мб не тот формат
4. много категорийных переменных
5. избыточность
6.

смотрим на данные

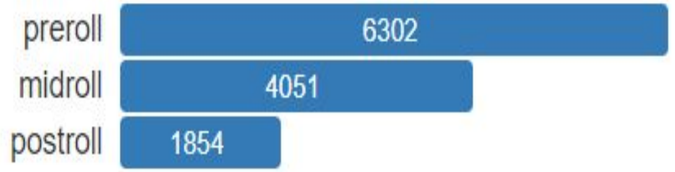
Advertisement display type

Categorical

HIGH CORRELATION

MISSING

Distinct	3
Distinct (%)	< 0.1%
Missing	45222
Missing (%)	78.7%



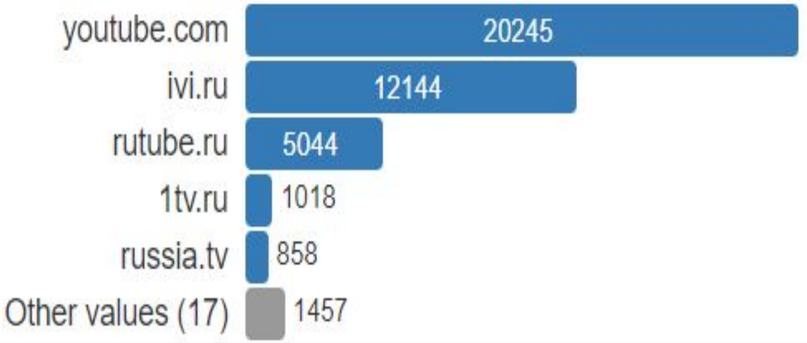
Advertisement player

Categorical

HIGH CORRELATION

MISSING

Distinct	22
Distinct (%)	0.1%
Missing	16663
Missing (%)	29.0%

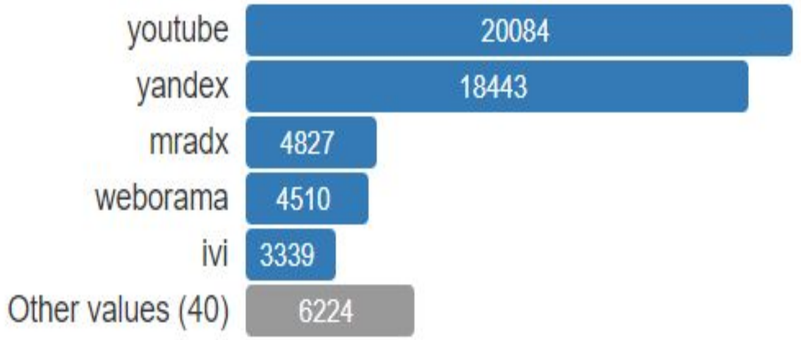


смотрим на данные

Banner Network

Categorical

Distinct	45
Distinct (%)	0.1%
Missing	2
Missing (%)	< 0.1%



df[df['Banner Network'].isna()]

Advertiser	Article level1	Article level4	Banner Network	Brands list	Day	Day type	First issue date	Holding	Site	Week day
HILLSIDE ASSETS	УСЛУГИ	WEB-САЙТ	NaN	ТОРРЕНТ-ТВ	2011-08-07	Выходной	2011-08-07	OTHERSITES.RU	tvforsite.ru	Воскресенье
HILLSIDE ASSETS	УСЛУГИ	WEB-САЙТ	NaN	ТОРРЕНТ-ТВ	2011-08-23	Рабочий	2011-08-07	OTHERSITES.RU	tvspectr.ru	Вторник

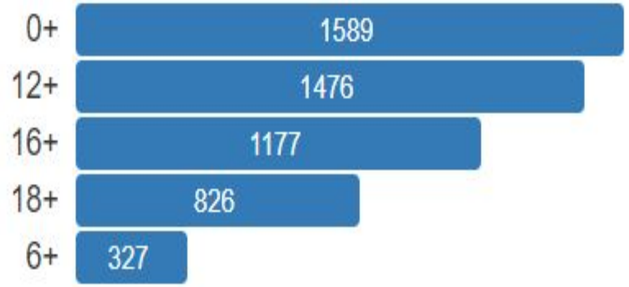
смотрим на данные

Advertisement age restriction

Categorical

MISSING

Distinct	5
Distinct (%)	0.1%
Missing	52034
Missing (%)	90.6%



df[df['Advertisement age restriction'].isna()].sample(2, random_state=42)

Advertisement skip	Advertisement TV Clip ID	Advertiser	Article level1	Article level4	Banner Network	Brands list	Day	Day type	First issue date	Holding	Site
11	0	TOYOTA	ТОВАРЫ	ВНЕДОРОЖНИКИ	mradx	LEXUS	2011-08-21	Выходной	2011-08-15	RUTUBE	rutube.ru
NaN	0	BEIERSDORF AG (BDF)	ТОВАРЫ	ДЕЗОДОРАНТ ДЛЯ МУЖЧИН	youtube	NIVEA	2011-08-30	Рабочий	2011-08-26	GOOGLE PROJECTS	youtube.com

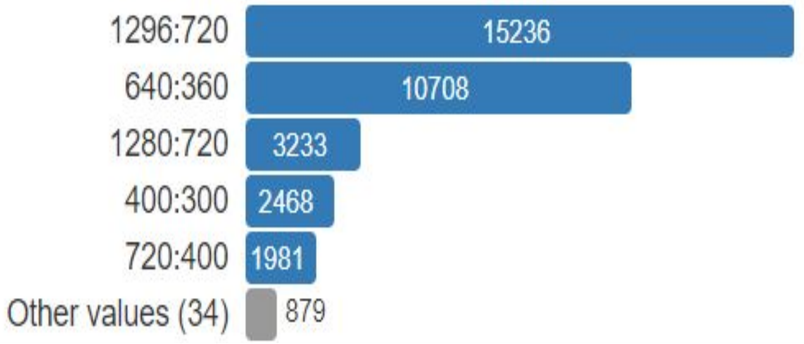
смотрим на данные

Advertisement format

Categorical

MISSING

Distinct	39
Distinct (%)	0.1%
Missing	22924
Missing (%)	39.9%



Advertisement file type

Categorical

Distinct	6
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%

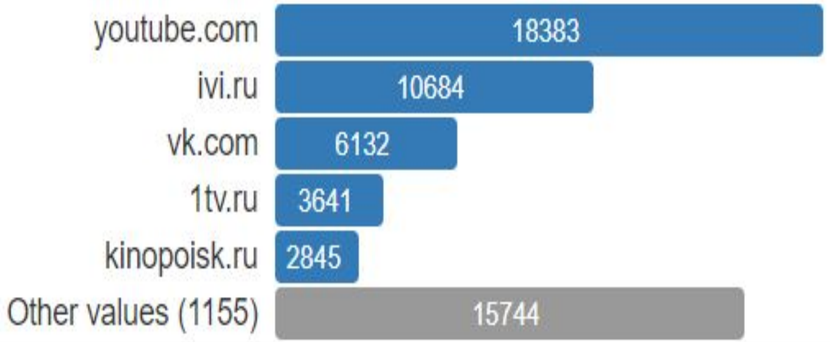
смотрим на данные

Site

Categorical

HIGH CARDINALITY

Distinct	1160
Distinct (%)	2.0%
Missing	0
Missing (%)	0.0%

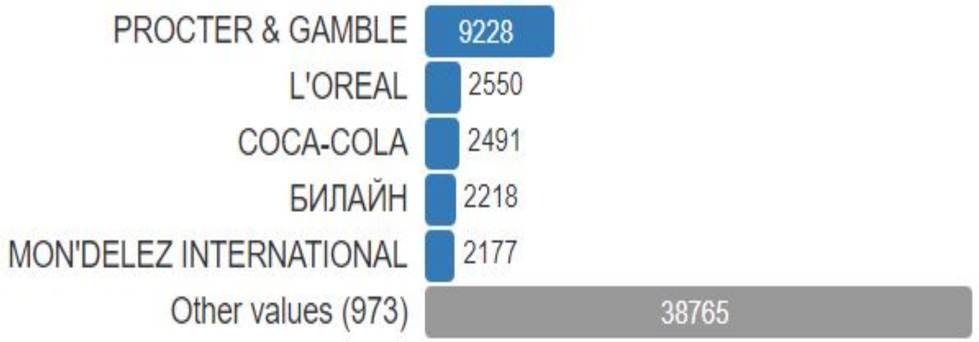


Advertiser

Categorical

HIGH CARDINALITY

Distinct	978
Distinct (%)	1.7%
Missing	0
Missing (%)	0.0%



смотрим на данные

Article level1

Categorical

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%

Article level4

Categorical

HIGH CARDINALITY

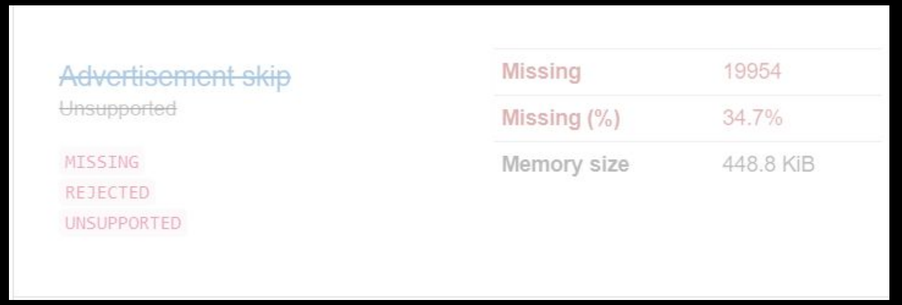
Distinct	517
Distinct (%)	0.9%
Missing	0
Missing (%)	0.0%

DataFrame.describe()

	Advertisement duration	Advertisement expected duration
count	57429.000000	57429.000000
mean	56.026467	36.762333
std	734.322391	89.857140
min	0.000000	0.000000
25%	15.000000	15.000000
50%	20.000000	20.000000
75%	26.000000	30.000000
max	32767.000000	1761.000000

DataFrame['Advertisement skip'].value_counts()

5	23811
NS	10317
11	2894
6	186
10	85
1	53
7	27
12	24
20	24
0	20
30	13
15	12
8	9

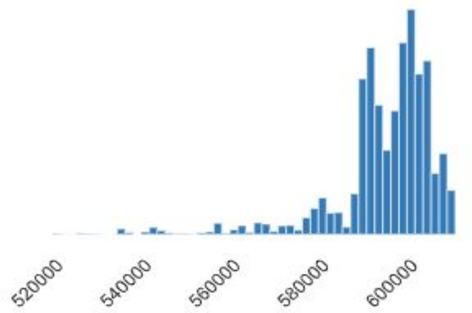


смотрим на данные

Advertisement ID

Real number ($\mathbb{R}_{\geq 0}$)

Distinct	6541	**Mean**	598345.2405
Distinct (%)	11.4%	**Minimum**	523270
Missing	0	**Maximum**	614056
Missing (%)	0.0%	**Zeros**	0
Infinite	0	**Zeros (%)**	0.0%
Infinite (%)	0.0%	**Memory size**	448.8 KiB



Advertisement TV Clip ID

Real number ($\mathbb{R}_{\geq 0}$)

ZEROS

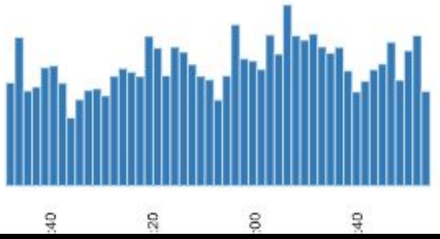
Distinct	271	**Mean**	893227.7394
Distinct (%)	0.5%	**Minimum**	0
Missing	0	**Maximum**	2948016
Missing (%)	0.0%	**Zeros**	39636
Infinite	0	**Zeros (%)**	69.0%
Infinite (%)	0.0%	**Memory size**	448.8 KiB

смотрим на данные

Day
Date

Distinct	49
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%

Minimum	2011-07-25 00:00:00
Maximum	2011-09-11 00:00:00



Day type
Categorical

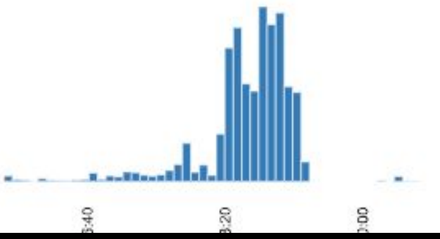
HIGH_CORRELATION

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%

First issue date
Date

Distinct	191
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%

Minimum	2011-01-01 00:00:00
Maximum	2011-12-24 00:00:00



“догмы” предобработки данных

спрашивает

Алексей

данные представлены флагами (1 и 0), при этом существует вероятность, что

пропущенные значения тоже равны 0. Как отличить пропуски от реальных данных?

Целевая переменная (!) яд/не яд

противоречия явные, неявные, асессоры, особенности набора, baseline

Дана

правило, когда возможно заполнить пропуски нулями?

Здравый смысл: когда меры центральной тенденции - плохо(см. пример с TV ID) или невозможно (поток данных, незаконченная сессия)

Ксения

критерии, которые помогут определиться “в зависимости от задачи”.

еще критерии

1. Сроки
2. Постоянство
3. Впервые?
4. Откуда (строгая отчетность,
   наши собирали сами, наши собирали с помощью,
   извне: с обратной связью, без обратной связи, полностью анонимизированные
   UGC: доски, опросы, соцсети)

“догмы” предобработки данных

1. Здравый смысл
2. Делаем обзор данных:
   смотрим всё, что мы можем узнать,
   фиксируем свои наблюдения,
   копаем, чтобы понять масштаб и причину (помощники в принятии решения)
3. Составляем план работ:
   группировка по типу “нехорошести” (пропуски, дубли и т.д.),
   группировка по признакам (по типу данных, по смыслу, по каналу получения и т.п.)
4. Опционально. Думаем на будущее и автоматизируем.

что думаете?

0
не понимаю

5
мне надо подумать

10
у меня вагон идей

вопросы?

Спасибо!